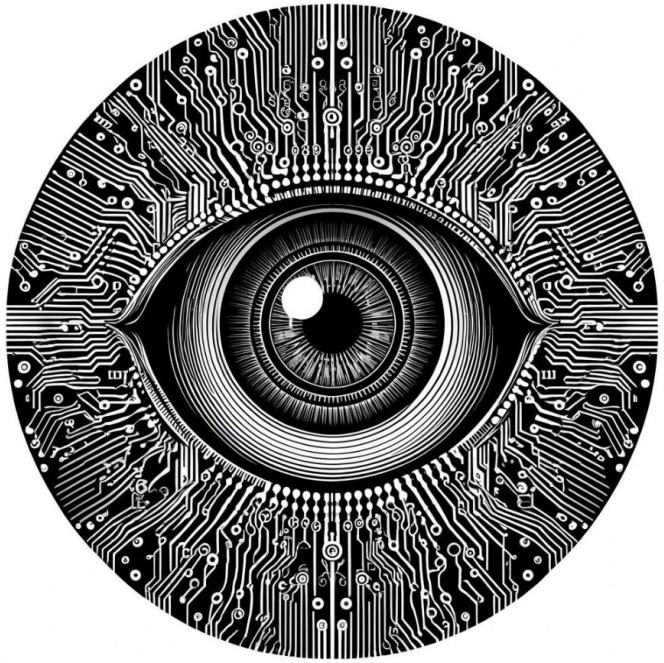
KI Service Zentrum by Hasso-Plattner-Institut HPI

Image Generation with Diffusion Models



Antonio Rueda-Toicen

SPONSORED BY THE
Federal Ministry of Education and Research

Learning goals

- Gain an overview of the denoising diffusion process
- Recognize the use of CLIP models to guide image generation with text prompts

Image generation with diffusion models

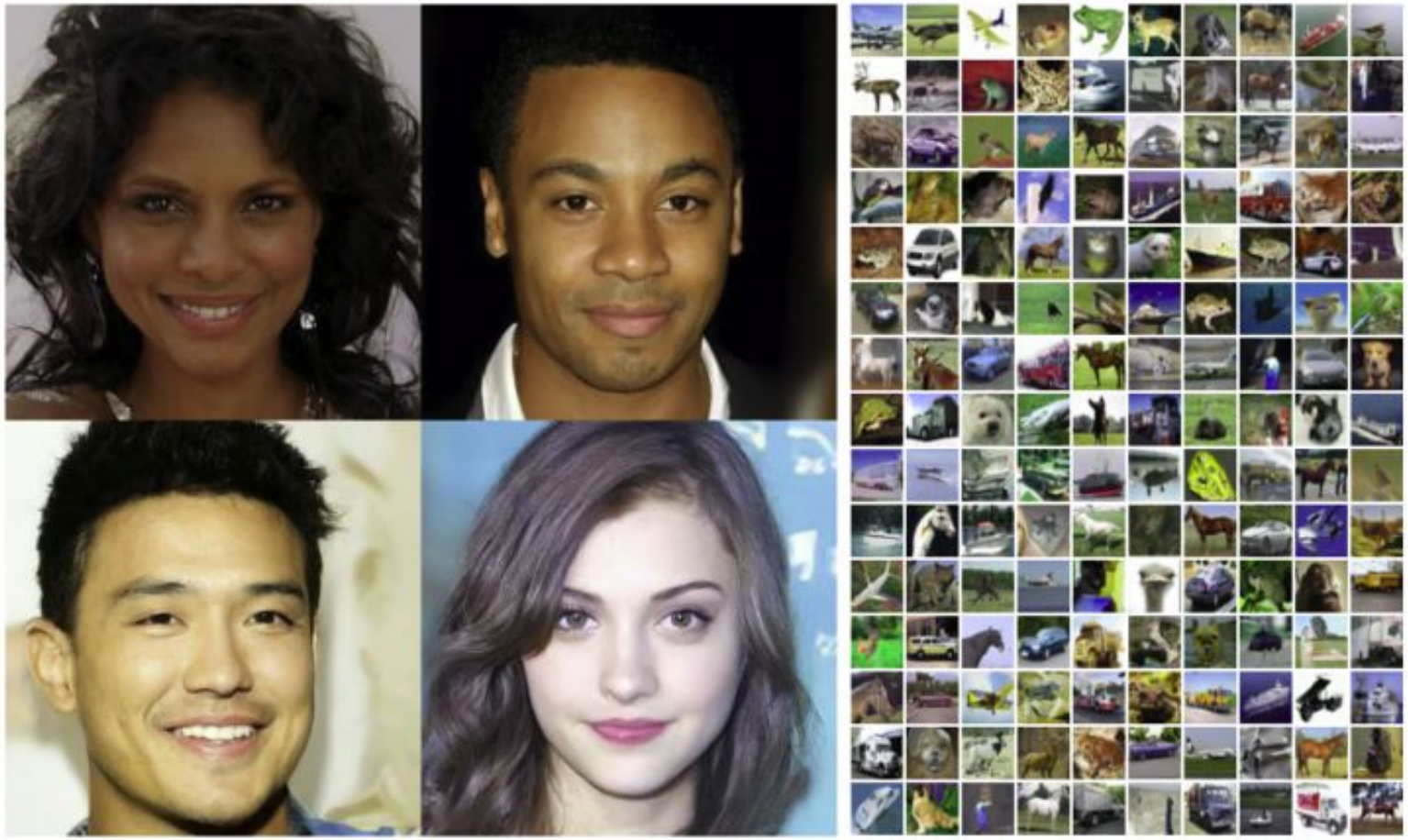
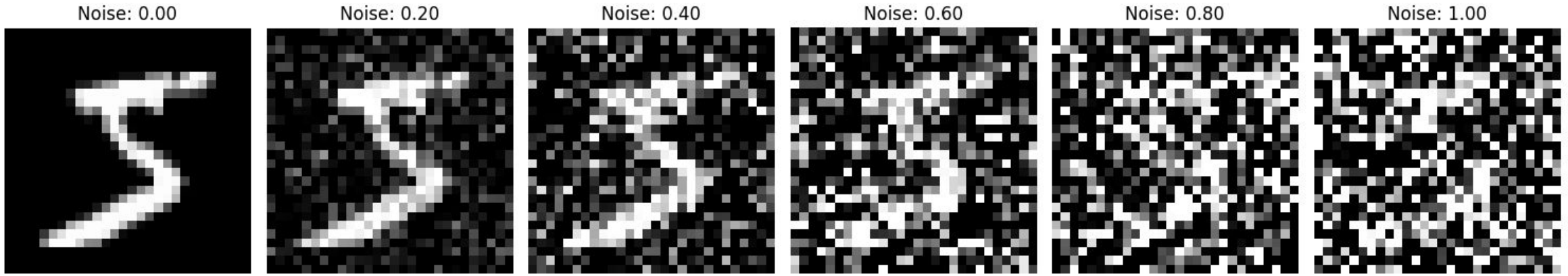


Figure 1: Generated samples on CelebA-HQ 256 × 256 (left) and unconditional CIFAR10 (right)

Image from Denoising Diffusion Probabilistic Models

Corrupting an image with Gaussian noise



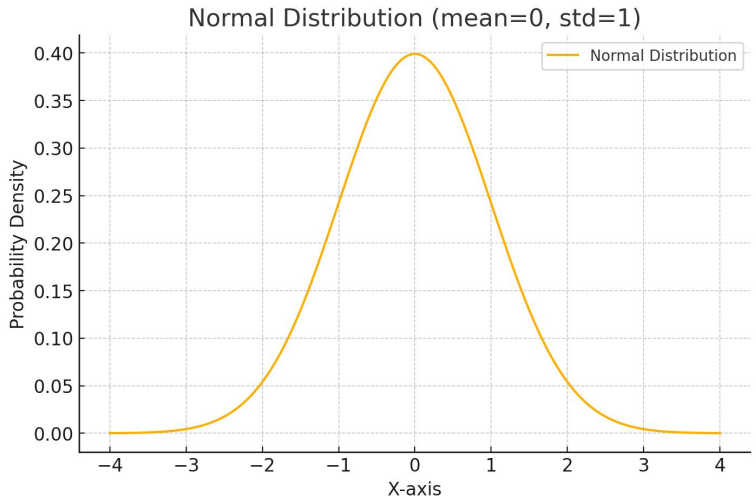
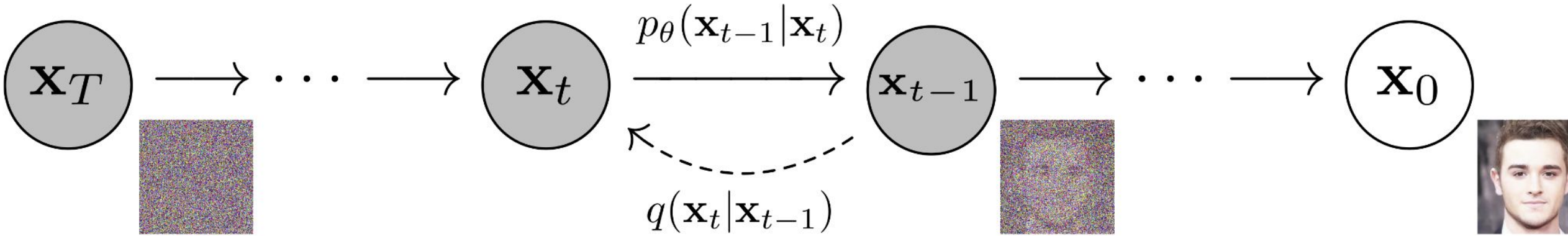
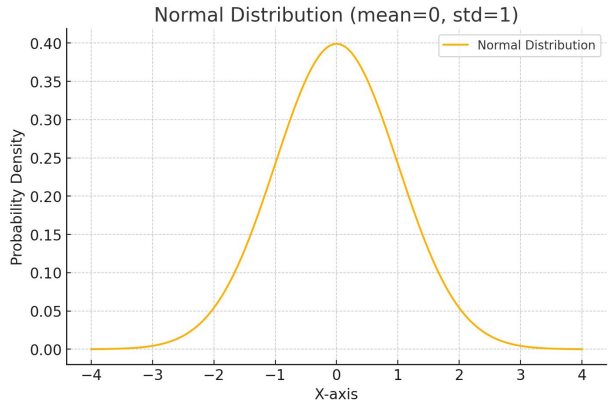


Image from source

Iterative denoising process





Forward diffusion

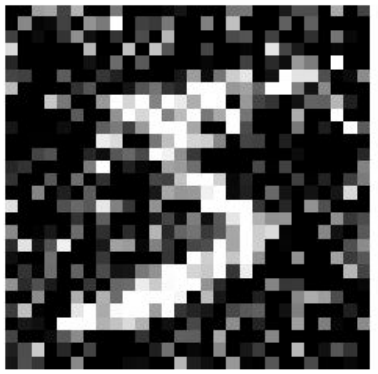
Beta controls how much noise is added on each time step, it is increased gradually. This increase is called the "noise schedule"

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = N(\mathbf{x}_t; \sqrt{1-\beta_t} \cdot \mathbf{x}_{t-1}, \beta_t \cdot \mathbf{I})$$

```python
import numpy as np

def step_forward(x_prev, beta_t):
    # Scale down the previous position
    mean = np.sqrt(1 - beta_t) * x_prev

    # Add random noise
    std = np.sqrt(beta_t)
    noise = np.random.normal(0, std, size=x_prev.shape)

    x_t = mean + noise
    return x_t
```

Noise: 0.20

Noise: 0.40



Reverse diffusion with U-net

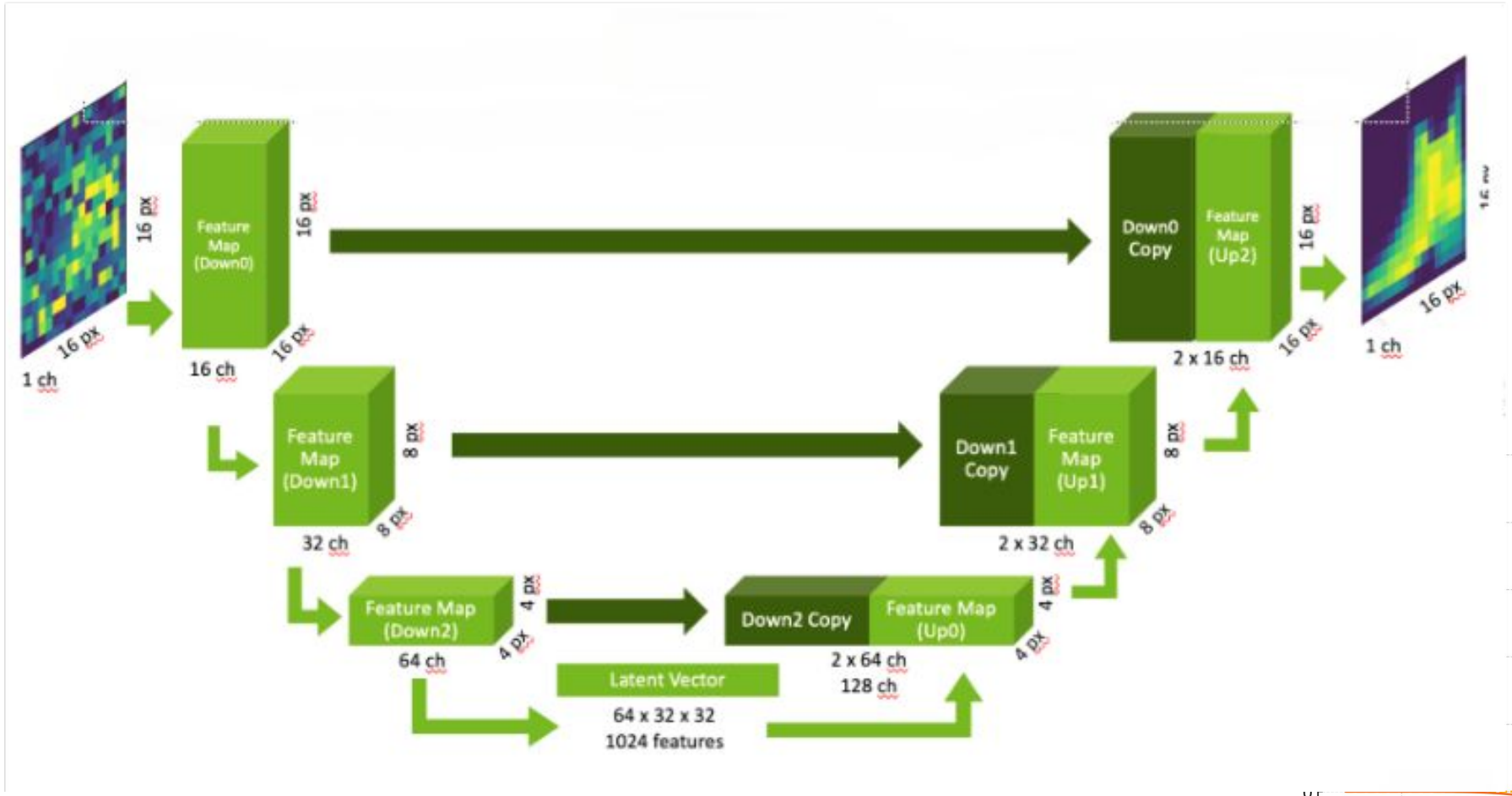
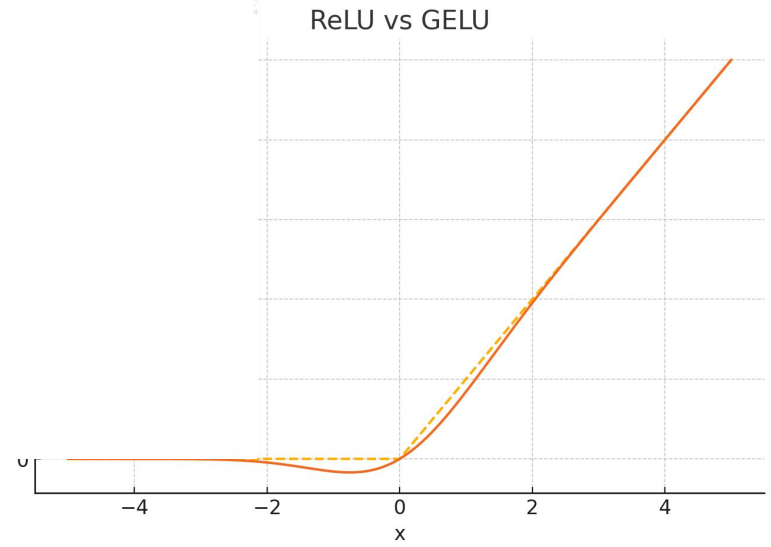
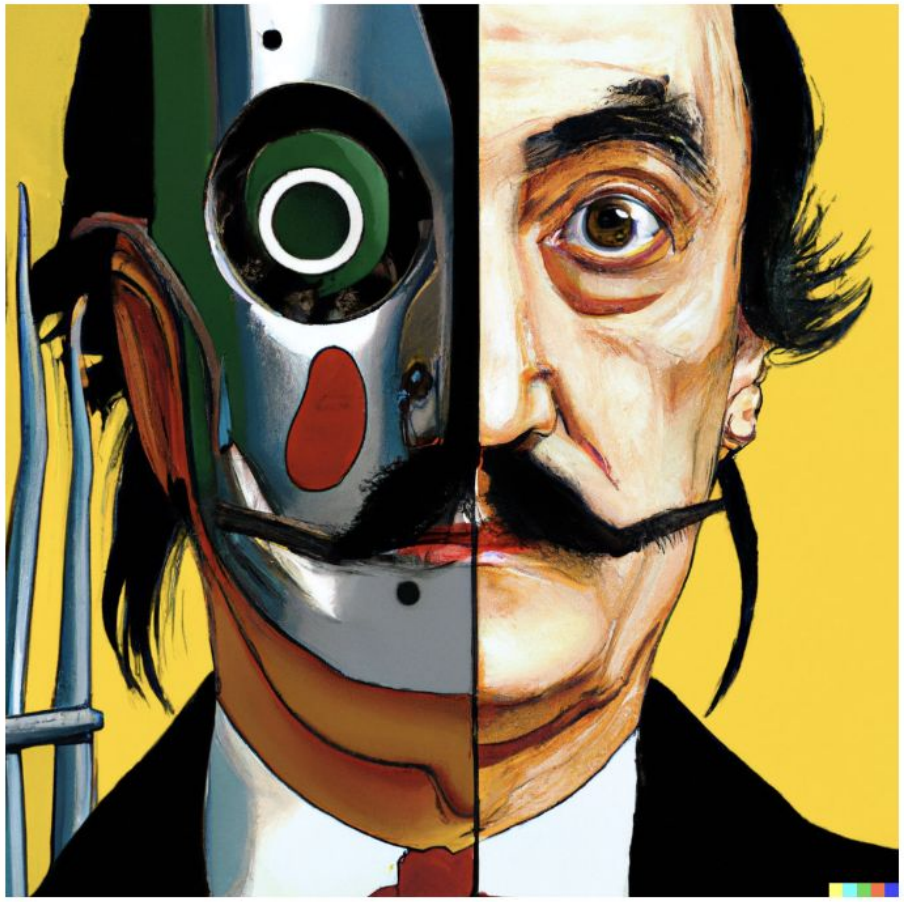


Image source

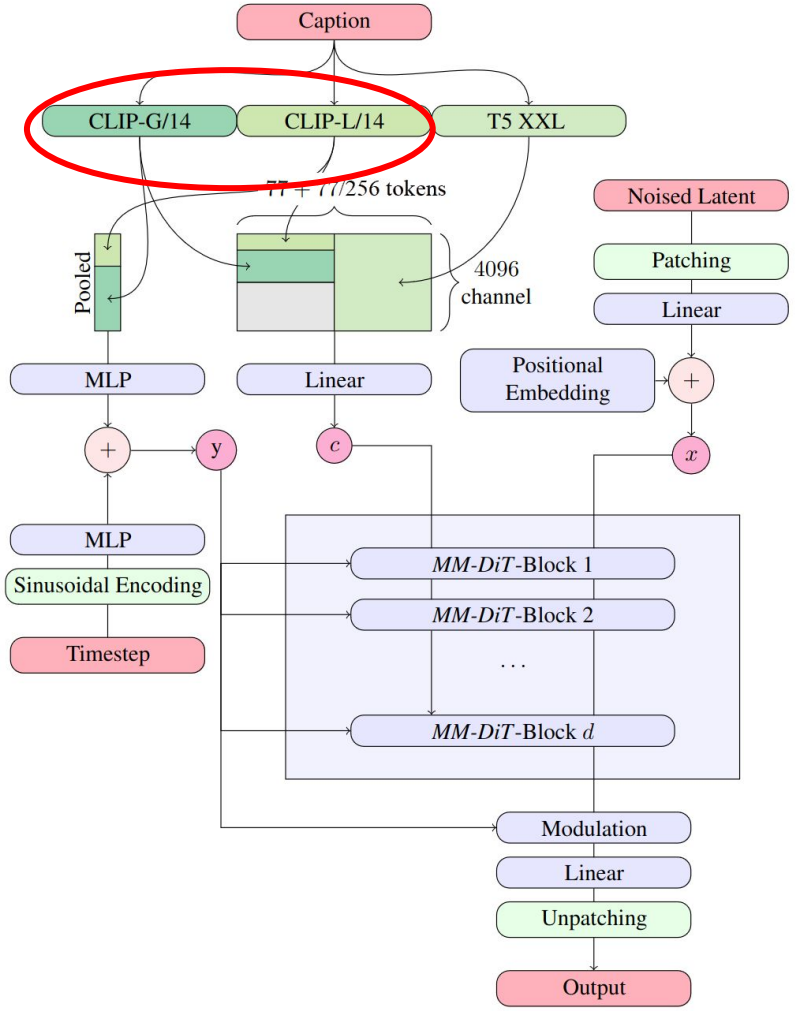


CLIP to guide text to image generation



vibrant portrait painting of Salvador Dalí with a robotic half face

Image from [Hierarchical Text-Conditional Image Generation with CLIP Latents]



Architecture diagram from [Stable Diffusion 3.5]

CLIP as input to decoders

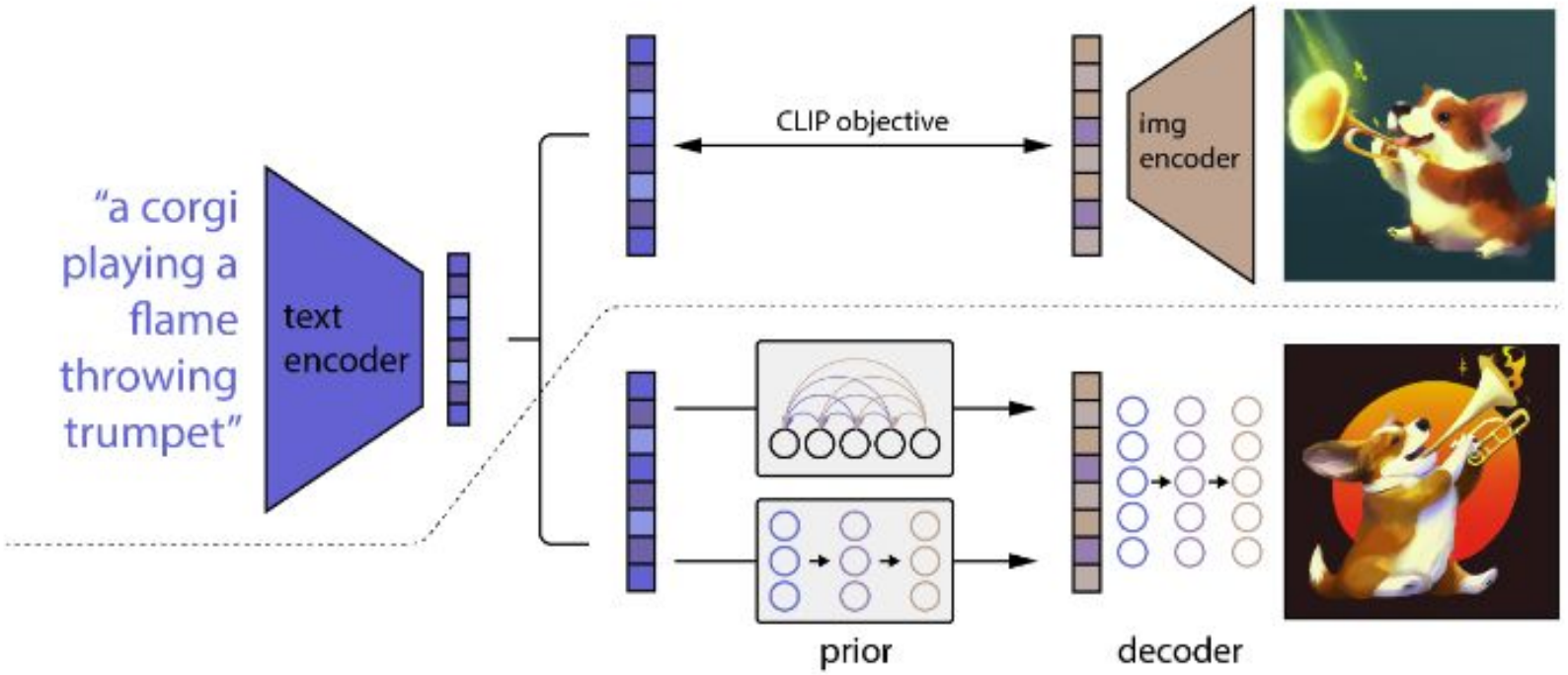


Figure 2: A high-level overview of unCLIP. Above the dotted line, we depict the CLIP training process, through which we learn a joint representation space for text and images. Below the dotted line, we depict our text-to-image generation process: a CLIP text embedding is first fed to an autoregressive or diffusion prior to produce an image embedding, and then this embedding is used to condition a diffusion decoder which produces a final image. Note that the CLIP model is frozen during training of the prior and decoder.

Image from Hierarchical Text-Conditional Image Generation with CLIP Latents

Summary

Diffusion models generate high quality images by reversing a noise addition process.

- They iteratively denoise from pure noise to generate images.

The forward diffusion process corrupts images

- Gradually adds Gaussian noise to images following a schedule (beta parameter)

The reverse diffusion process is about learning to predict the noise

- Uses a U-net architecture to estimate what noise was added at each step
- We predict the noise component to subtract it from the corrupted image
- The network is trained to minimize the difference between predicted and actual noise

CLIP enables text-guided image generation

- CLIP text embeddings help us control the reverse diffusion process.

SPONSORED BY THE
Federal Ministry of Education and Research

Further reading and references

Denoising Diffusion Probabilistic Models

- https://arxiv.org/abs/2006.11239

Hierarchical Text-Conditional Image Generation with CLIP Latents

- https://arxiv.org/abs/2204.06125

The Annotated Diffusion Model

- https://huggingface.co/blog/annotated-diffusion

The Physics Principle That Inspired Modern AI Art

- https://www.quantamagazine.org/the-physics-principle-that-inspired-modern-ai-art-20230105/

SPONSORED BY THE
Federal Ministry of Education and Research