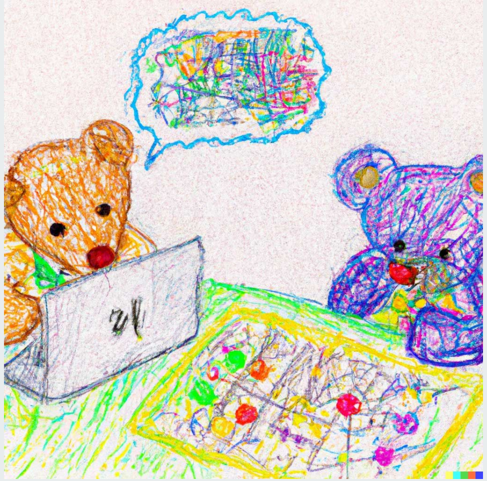
# Text-2-Image

*“Teddy bears working on new AI research as kids’ crayon art”* –>

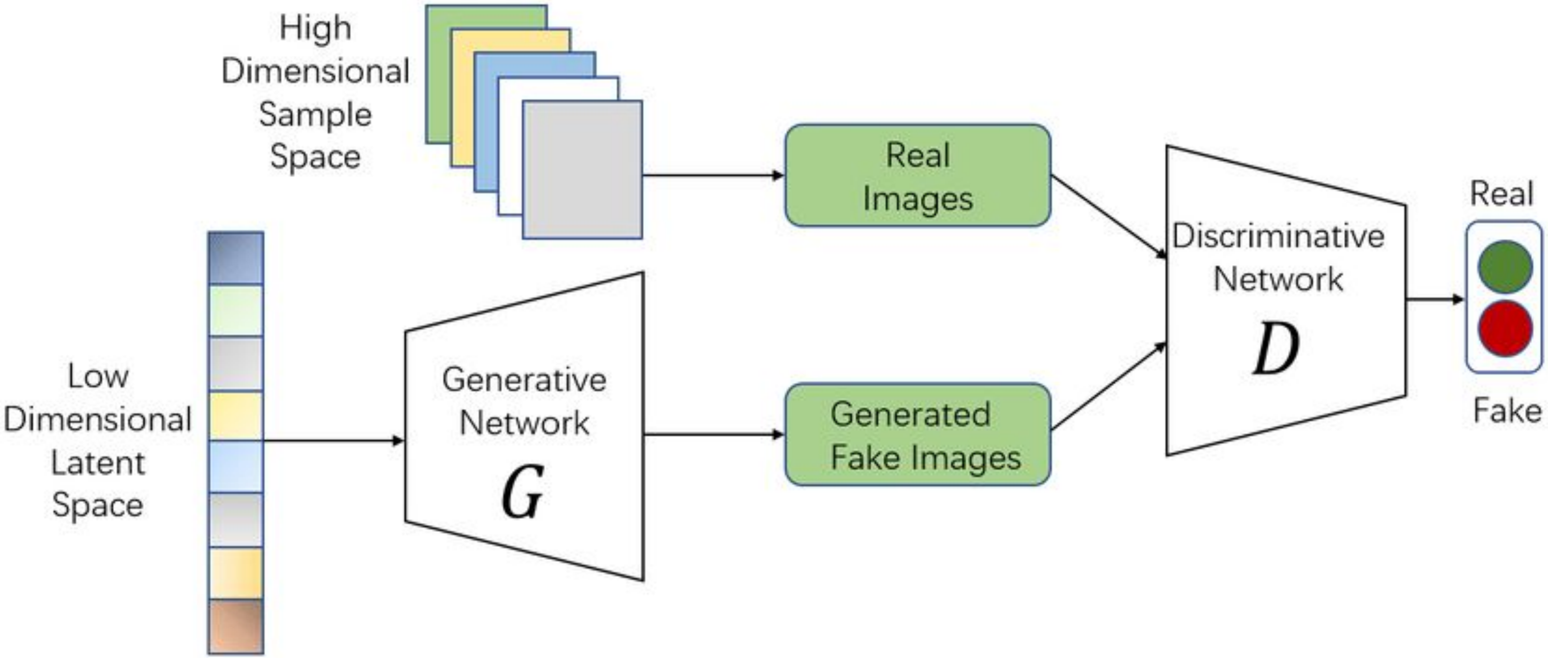


Karol Urbańczyk

# Agenda

1. How it all started - **initial attempts** and background knowledge (GANs, VAEs)
2. **DALL-E** - first non-domain-specific approach
3. **GLIDE** - introducing Diffusion
4. **DALL-E 2**
5. Open-source approaches & community
6. Google comes into play (**Imagen**, **Parti**)

# Vanilla GAN for image generation (2014)



# GAN generating image conditioned on Text (2016) [1]

What if we do not generate from the noise, but concatenate textual description to it instead?

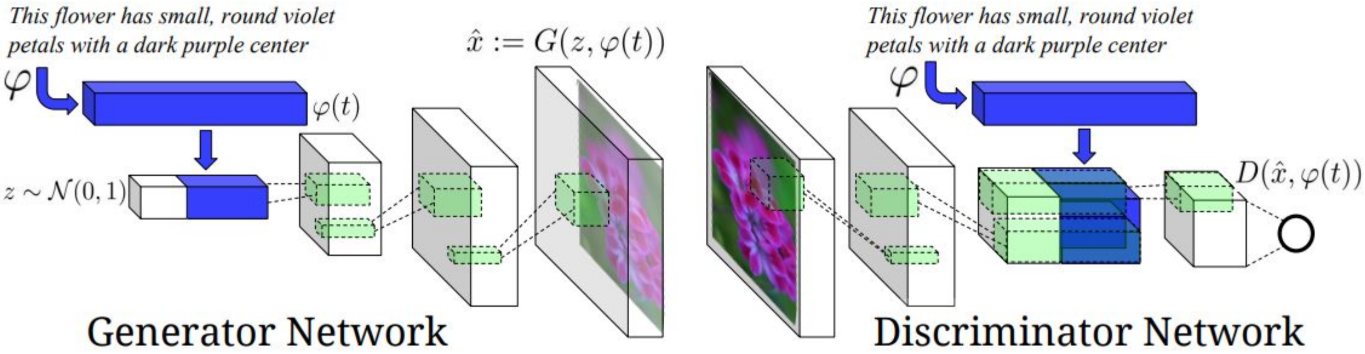
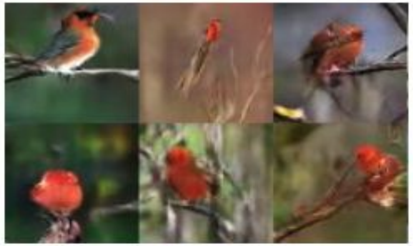

This flower has small, round violet petals with a dark purple center
$\varphi$
$\varphi(t)$
$z \sim \mathcal{N}(0,1)$
$\hat{x} := G(z, \varphi(t))$
Generator Network
This flower has small, round violet petals with a dark purple center
$\varphi$
$D(\hat{x}, \varphi(t))$
Discriminator Network

*Figure 2.* Our text-conditional convolutional GAN architecture. Text encoding $\varphi(t)$ is used by both generator and discriminator. It is projected to a lower-dimensions and depth concatenated with image feature maps for further stages of convolutional processing.

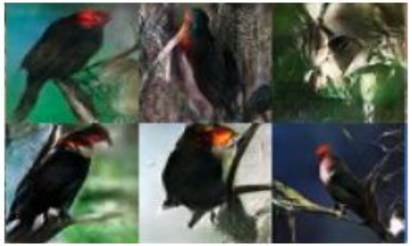
1. Pair of (Real Image, Real Caption) as input and target variable is set to 1
2. Pair of (Wrong Image, Real Caption) as input and target variable is set to 0
3. Pair of (Fake Image, Real Caption) as input and target variable is set to 0

# GAN generating image conditioned on Text (2016) [2]

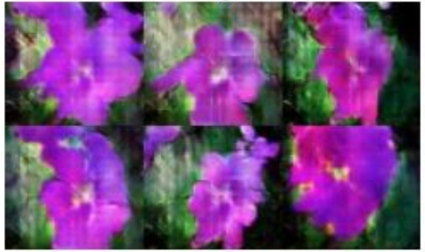
this small bird has a pink breast and crown, and black primaries and secondaries.



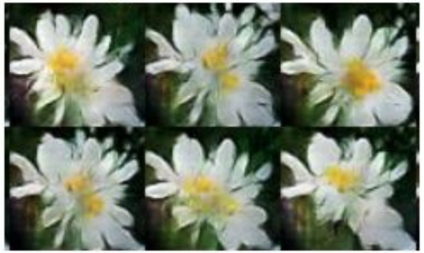
this magnificent fellow is almost all black with a red crest, and white cheek patch.



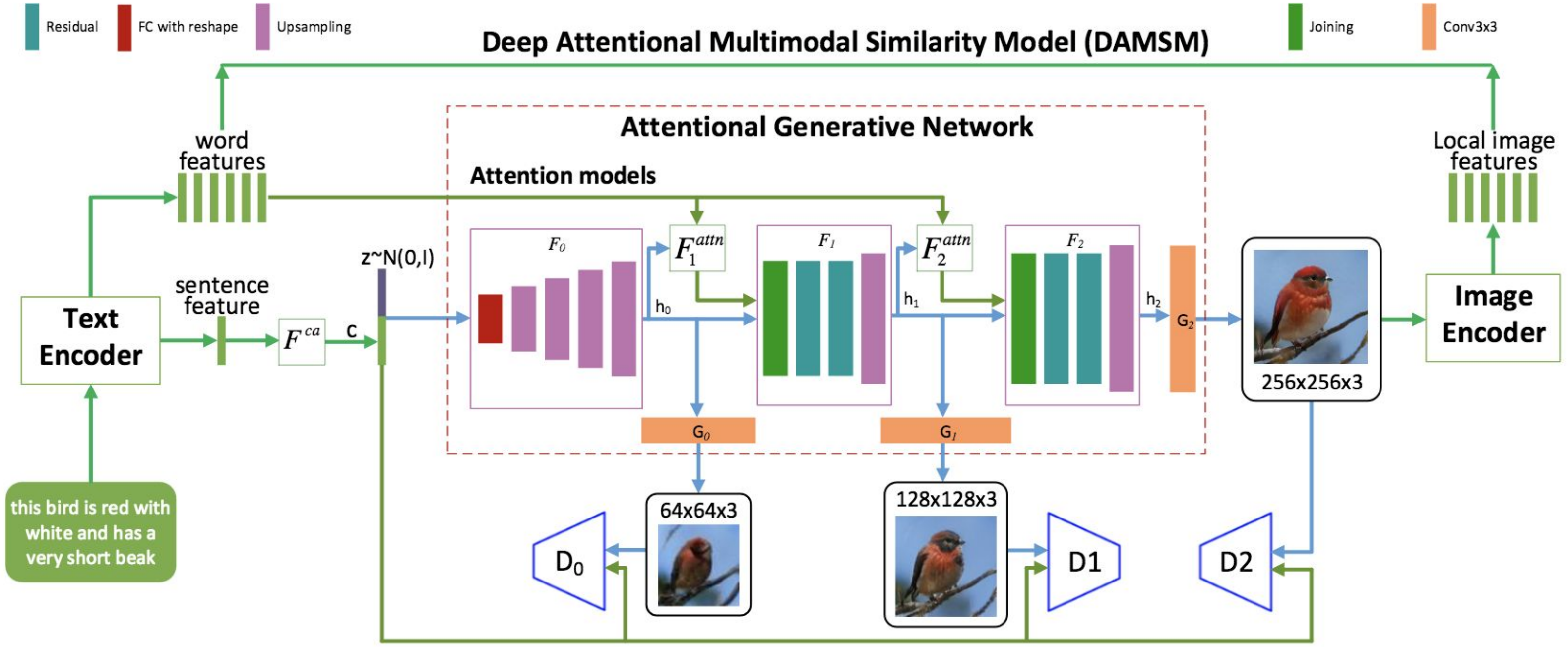
the flower has petals that are bright pinkish purple with white stigma



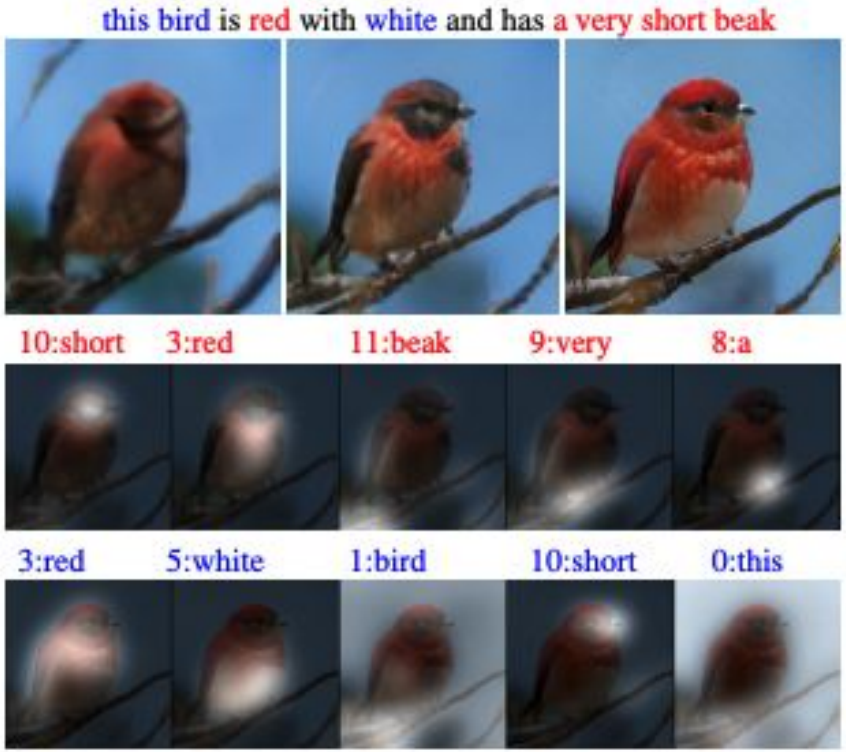
this white and yellow flower have thin white petals and a round yellow stamen



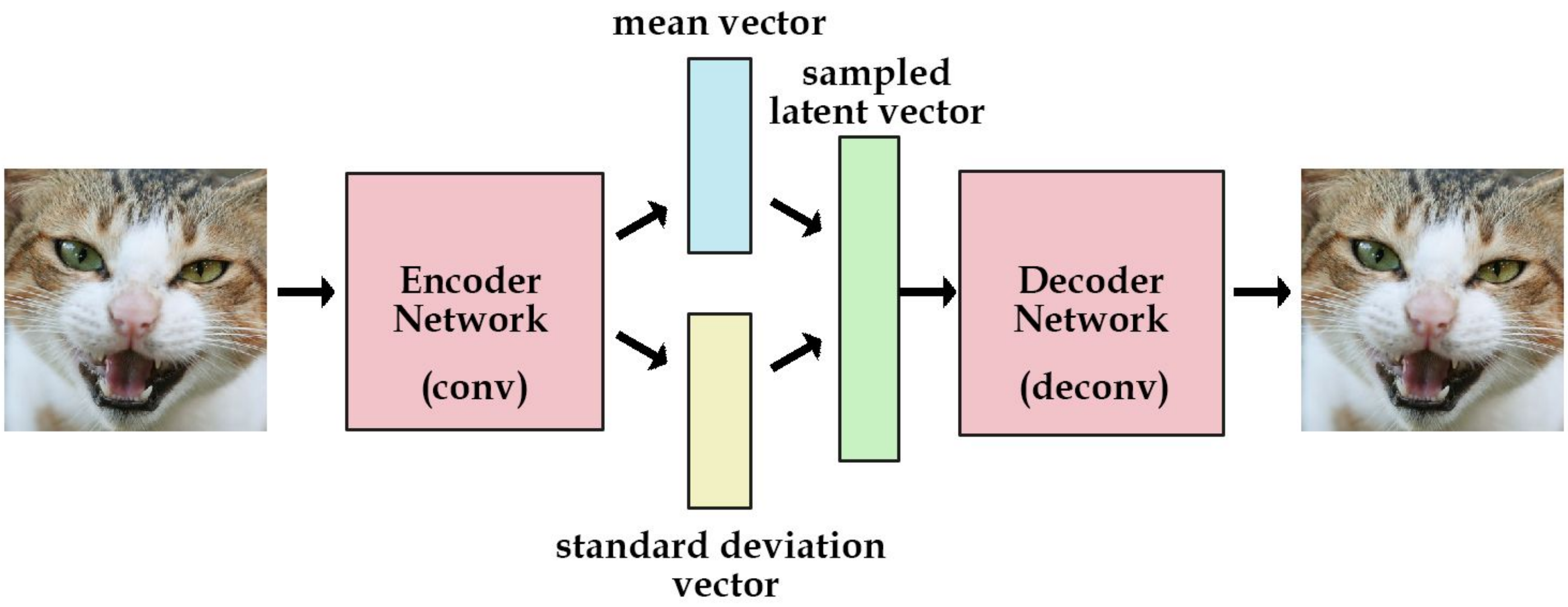
# AttnGAN - first use of attention mechanism (2017) [1]



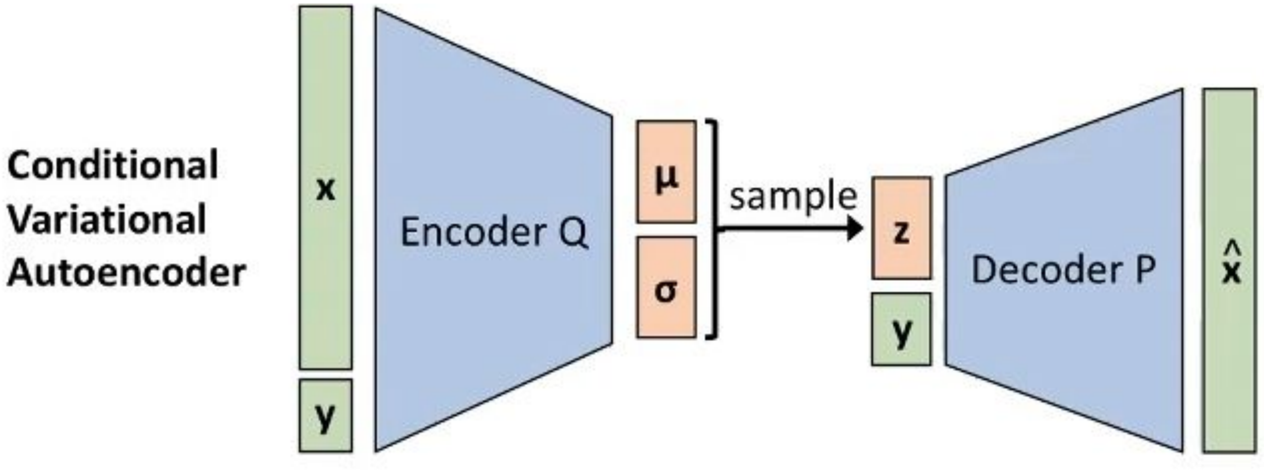
# AttnGAN - first use of attention mechanism (2017) [2]

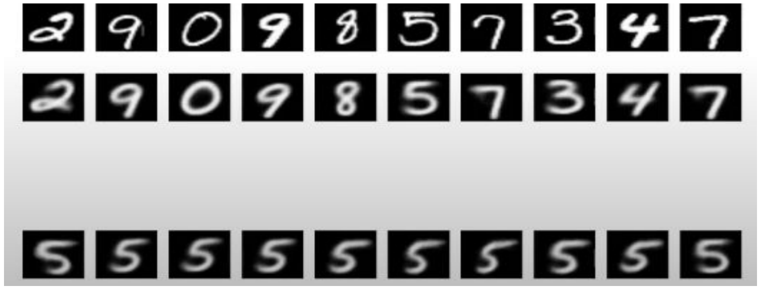


# Variational Autoencoder (VAE)



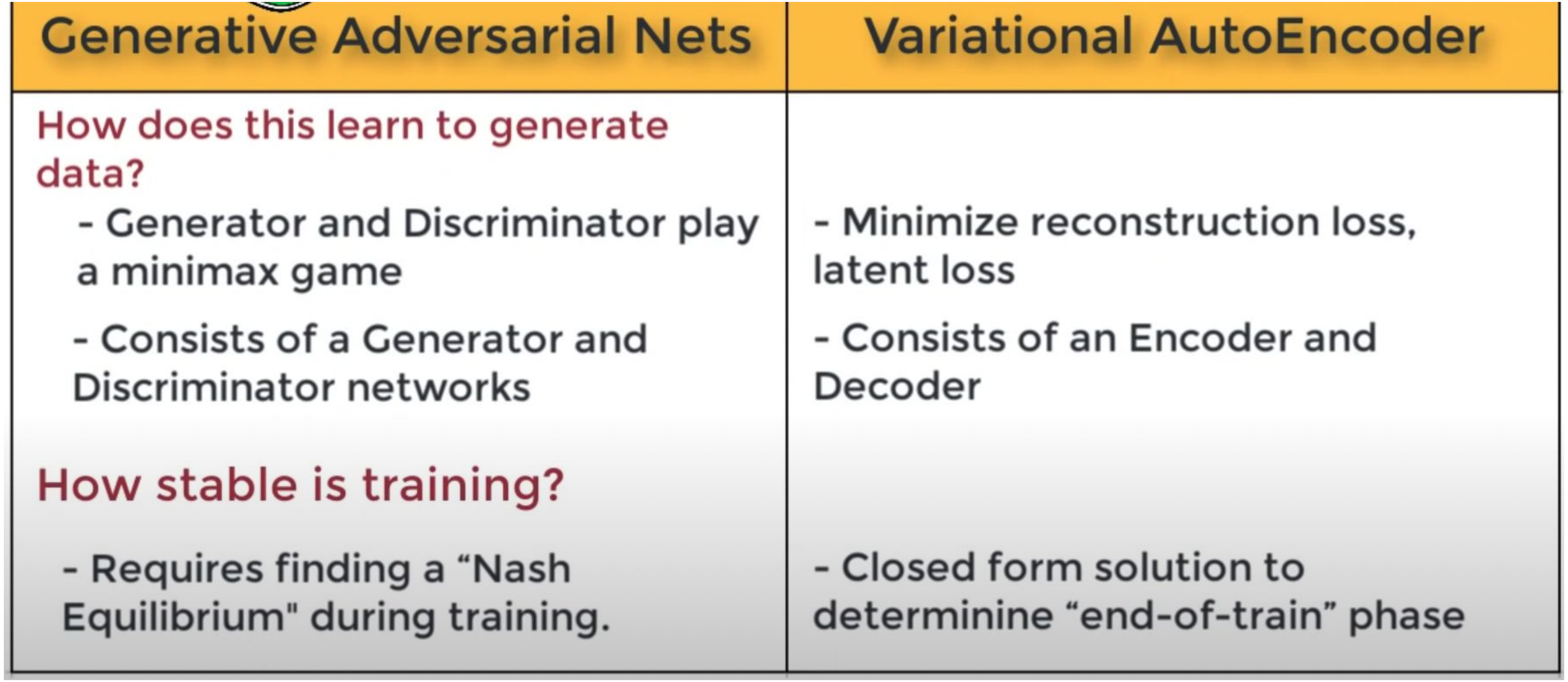
# Conditional VAE





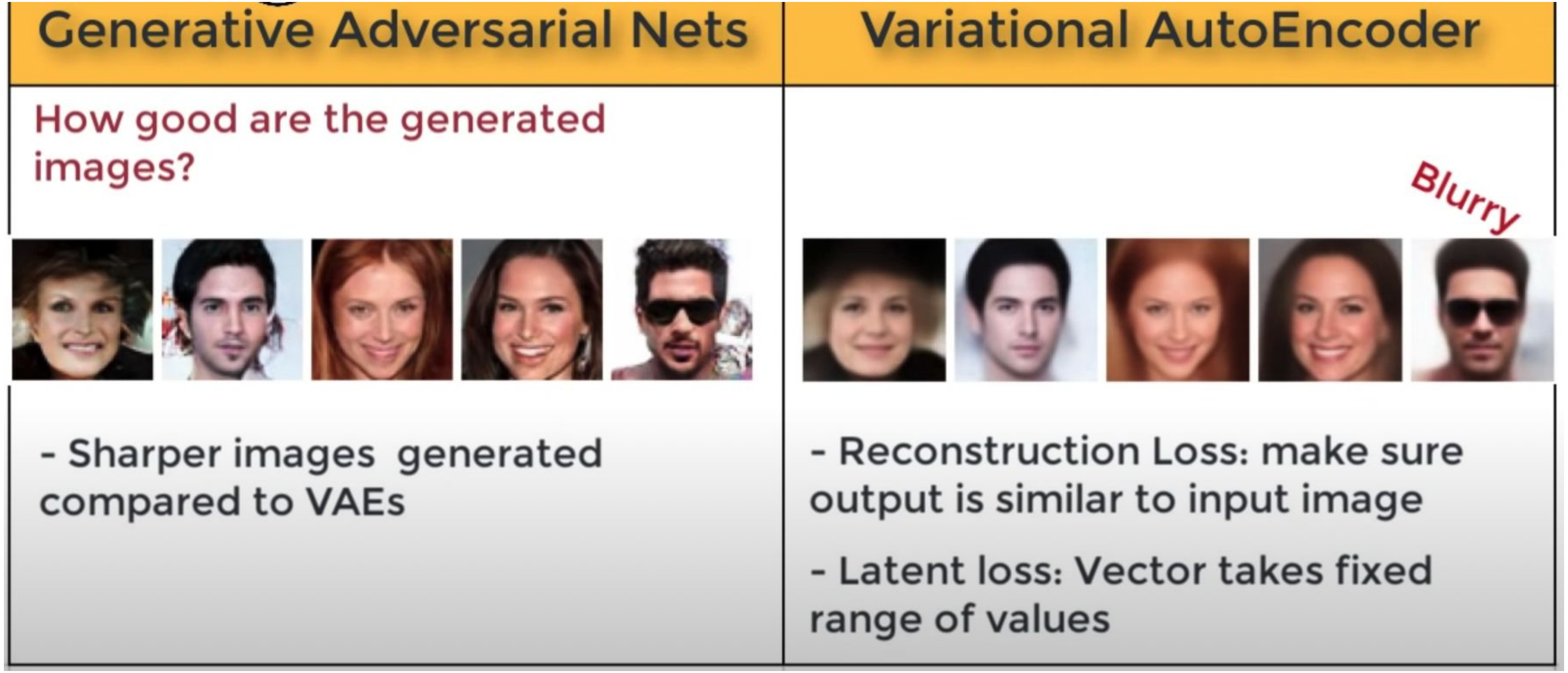
1

# GANs vs VAEs



| Generative Adversarial Nets | Variational AutoEncoder |
|---|---|
| How does this learn to generate data? | |
| - Generator and Discriminator play a minimax game | - Minimize reconstruction loss, latent loss |
| - Consists of a Generator and Discriminator networks | - Consists of an Encoder and Decoder |
| How stable is training? | |
| - Requires finding a “Nash Equilibrium" during training. | - Closed form solution to determinine “end-of-train” phase |

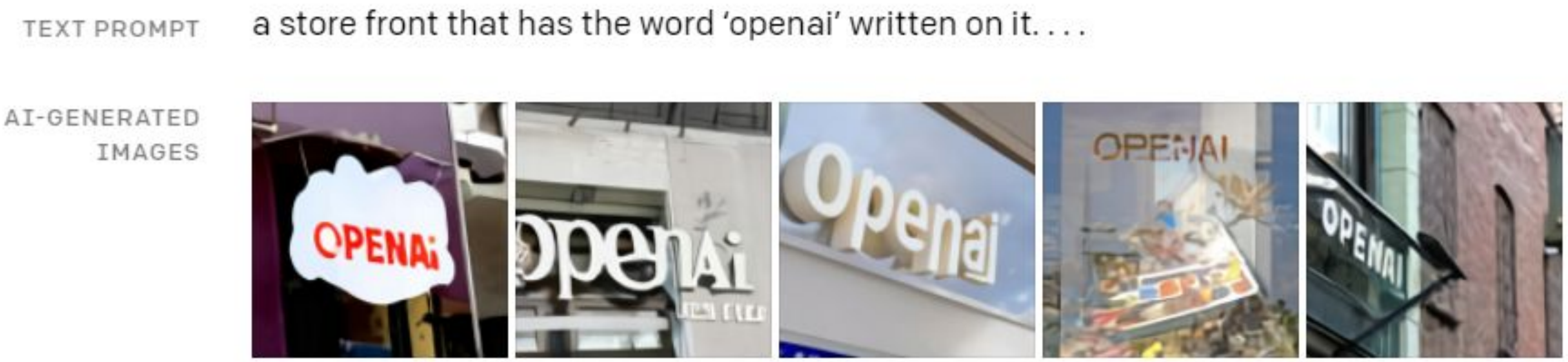
# GANs vs VAEs



| Generative Adversarial Nets | Variational AutoEncoder |
|---|---|
| How good are the generated images? | Blurry |
| - Sharper images generated compared to VAEs | - Reconstruction Loss: make sure output is similar to input image<br>- Latent loss: Vector takes fixed range of values |

# What is famous Dall-E (January 2021)

- First serious attempt on **zero-shot** text-2-image generation. Dall-E is **not domain-specific**
- This comes from **huge dataset** and lots of resources invested into **engineering**
- **In principle it consists of already known concepts, but scaled significantly…**



TEXT PROMPT a store front that has the word 'openai' written on it. . . .

AI-GENERATED IMAGES

# Crucial component of Dall-E is dVAE

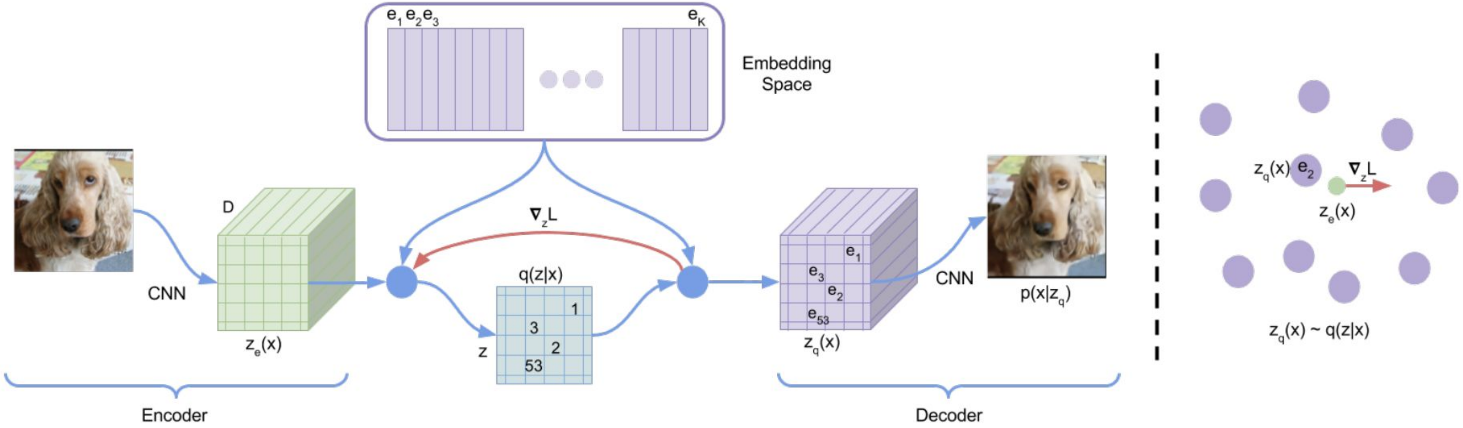
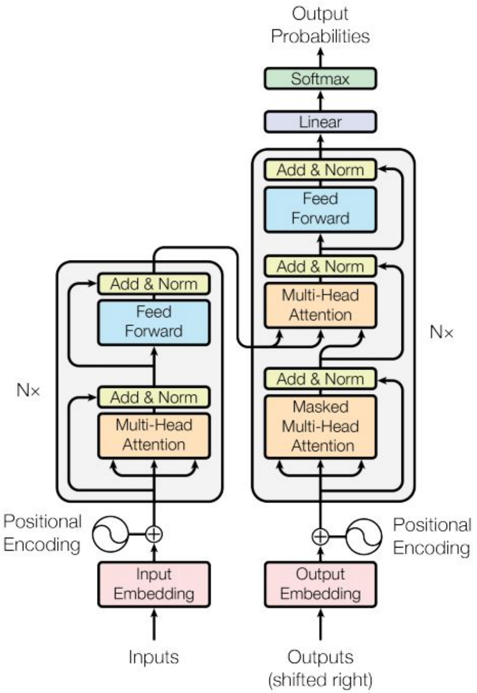
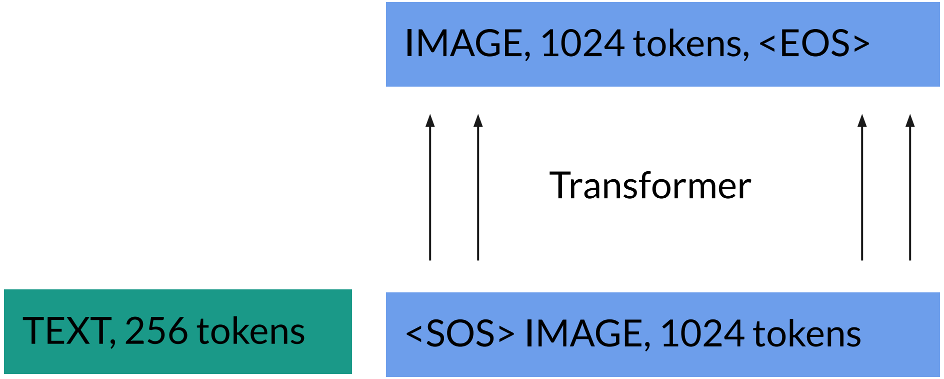


Figure 1: Left: A figure describing the VQ-VAE. Right: Visualisation of the embedding space. The output of the encoder $z(x)$ is mapped to the nearest point $e_2$. The gradient $\nabla_z L$ (in red) will push the encoder to change its output, which could alter the configuration in the next forward pass.

# Dall-E - training with Transformer



- Autoregressive transformer
- Next token prediction
- Inference: pass vector through a VQ-VAE decoder, rank with CLIP, “cherrypick” ;) and voilà!



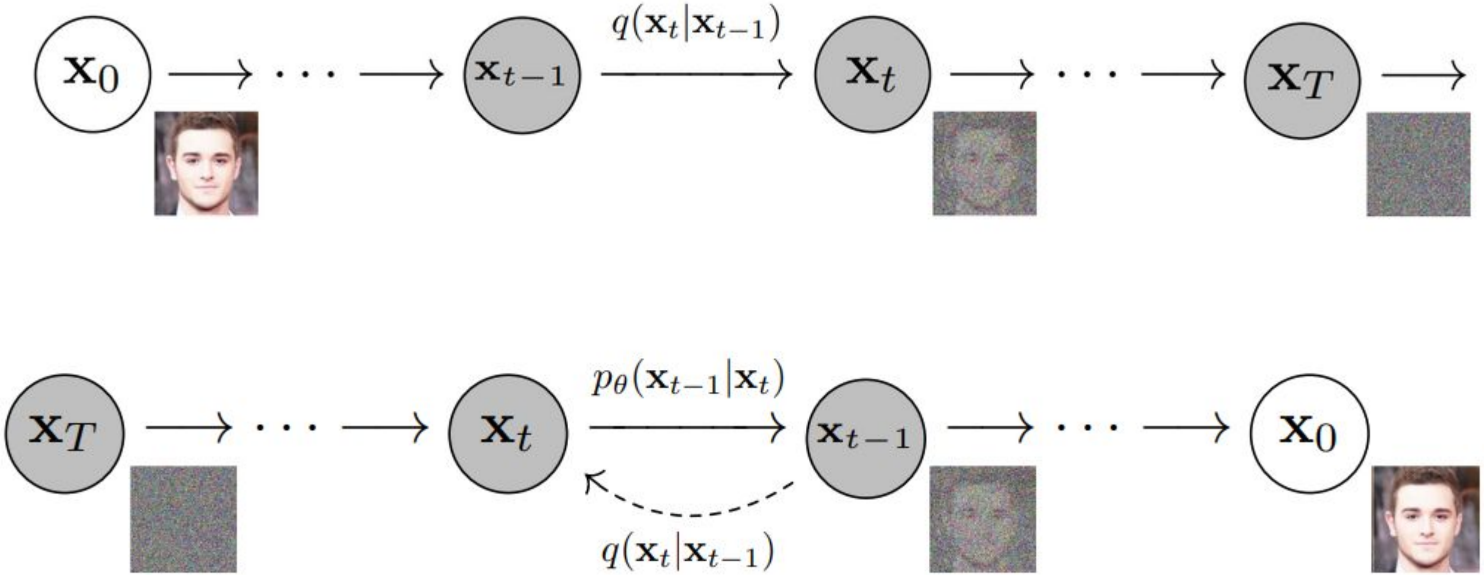
IMAGE, 1024 tokens, <EOS>

Transformer

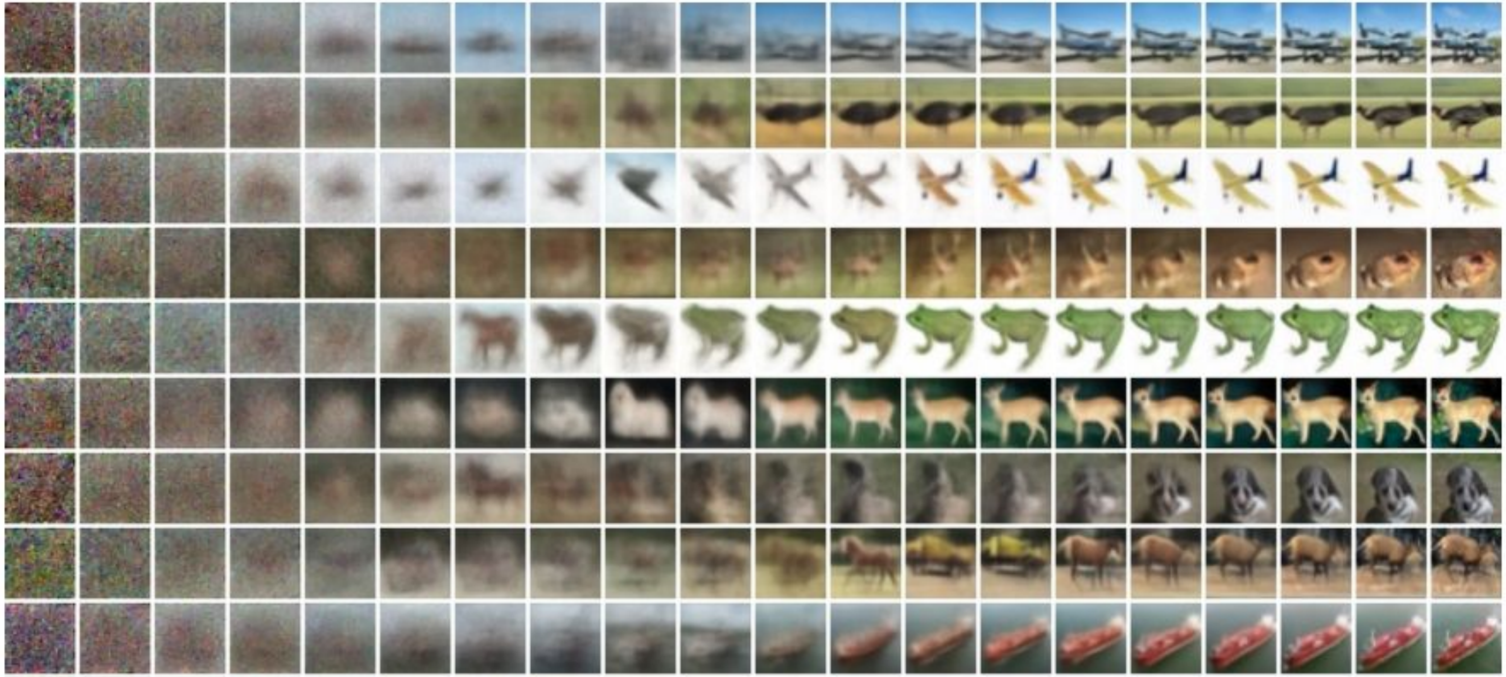
TEXT, 256 tokens

<SOS> IMAGE, 1024 tokens

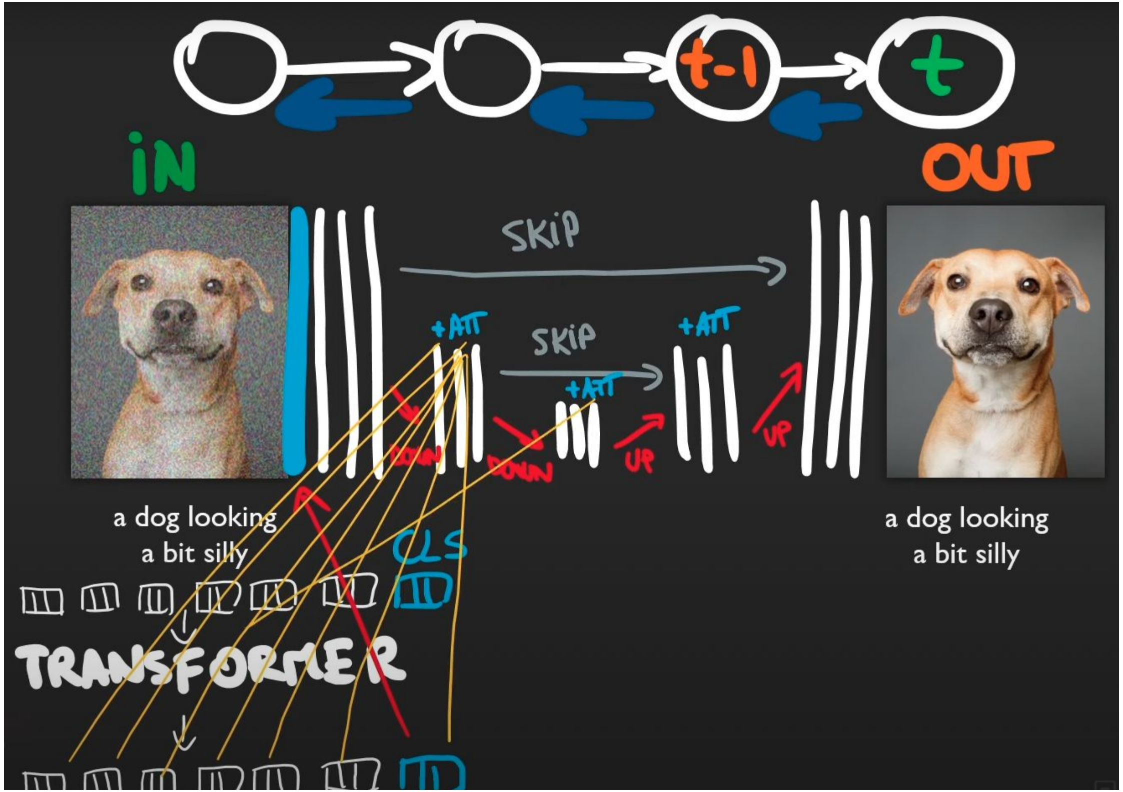
# Introducing diffusion concept



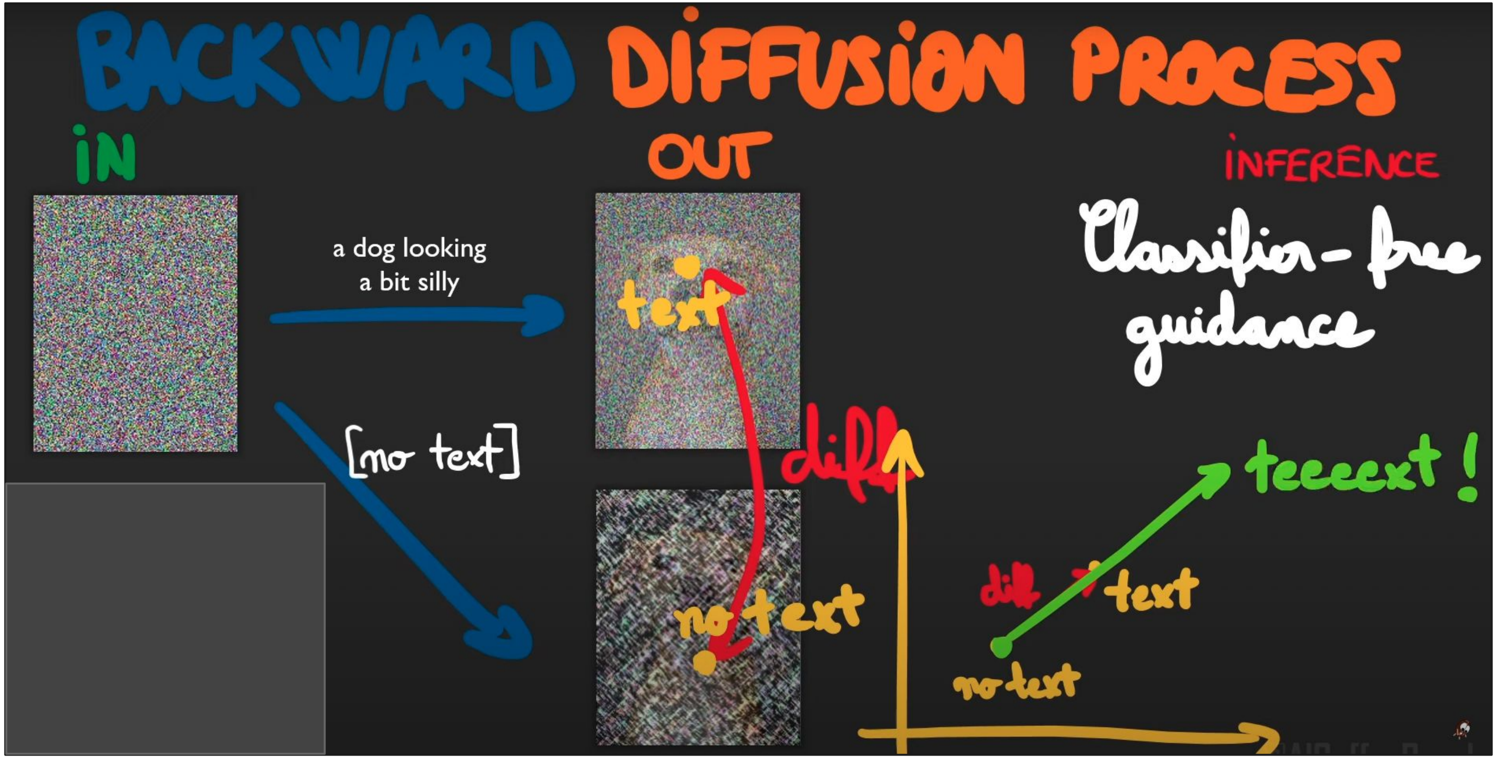
# Introducing diffusion concept



# GLIDE as a first text-2-image diffusion model (2022)

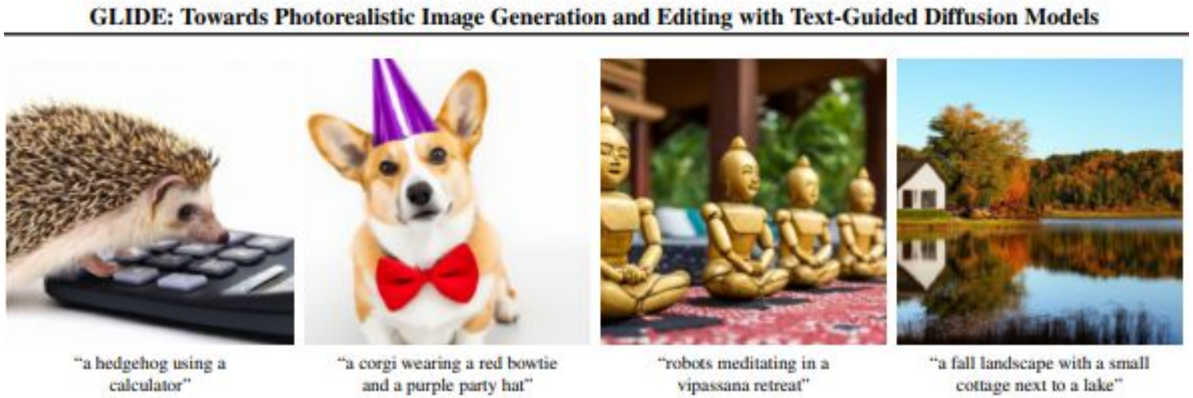


# GLIDE inference



# GLIDE characteristics (vs Dall-E)

- 3.5 B parameters vs 12 B in Dall-E
- more photorealistic, on the other hand not so wide domain
- longer inference time (probably because of sequential nature)



GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models

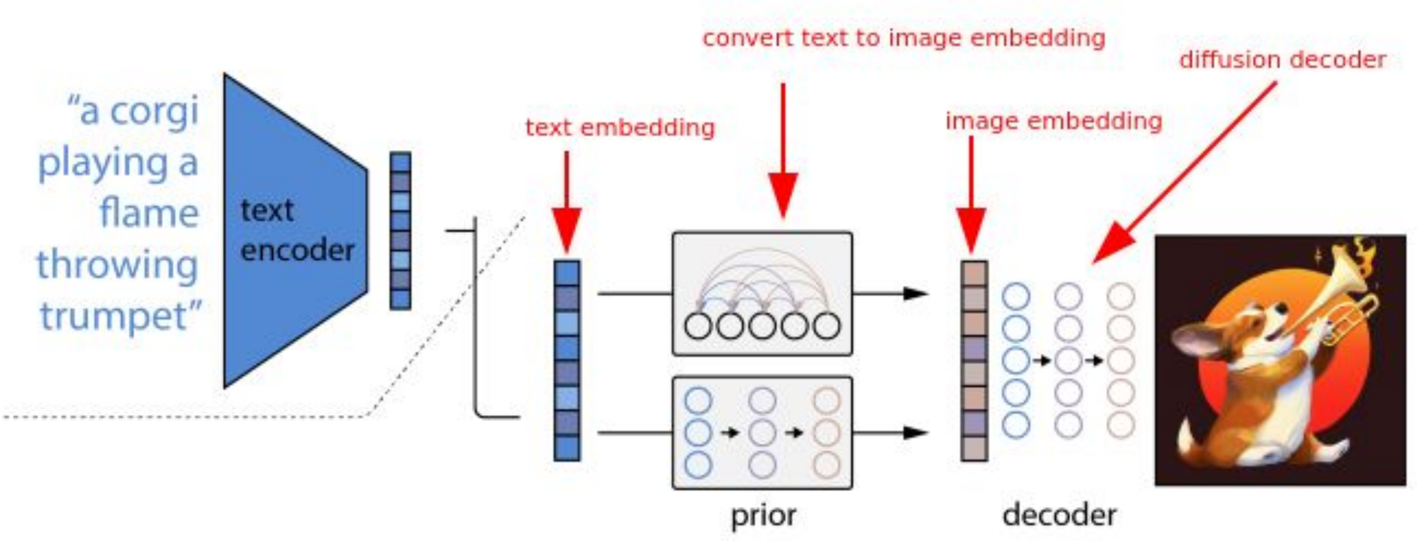
"a hedgehog using a calculator"

"a corgi wearing a red bowtie and a purple party hat"

"robots meditating in a vipassana retreat"

"a fall landscape with a small cottage next to a lake"

# DALL-E2 (combines diffusion, CLIP & GLIDE)


convert text to image embedding
diffusion decoder
text embedding
image embedding
"a corgi playing a flame throwing trumpet"
text encoder
prior
decoder

# DALL-E2 samples

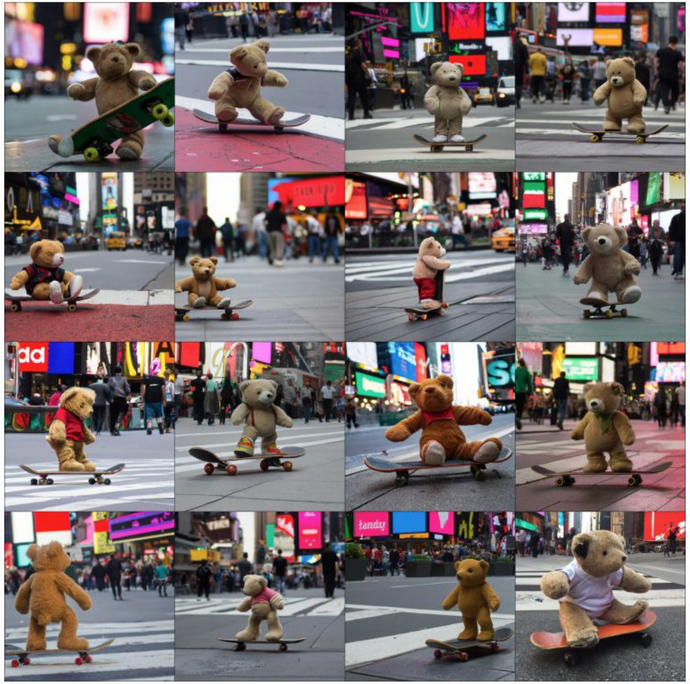


Figure 20: Random samples from unCLIP for prompt "A teddybear on a skateboard in Times Square."

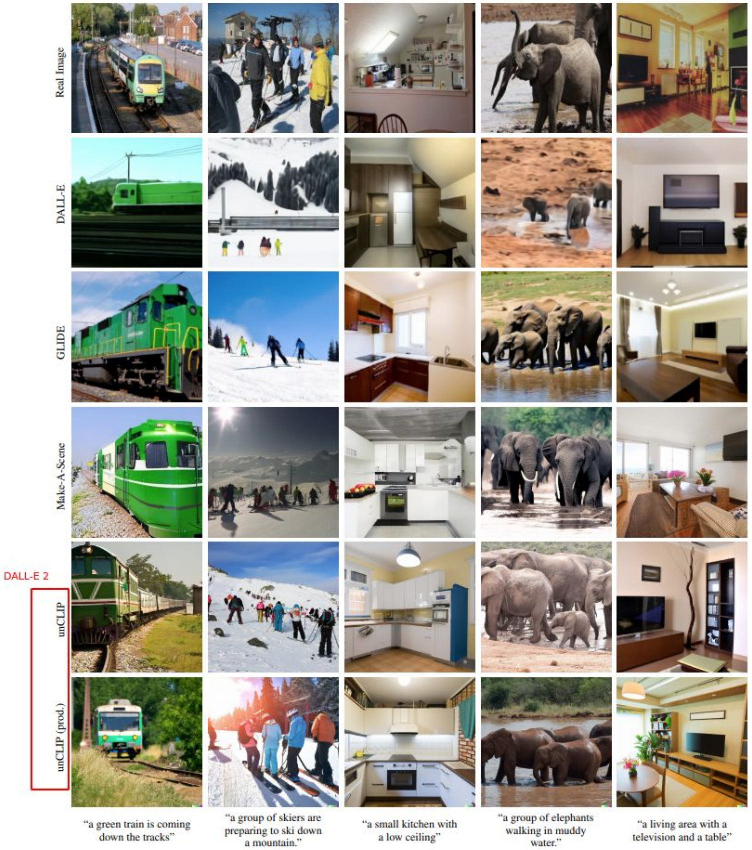

Real Image
DALL-E
GLIDE
Make-A-Scene
DALL-E 2
unCLIP
unCLIP (prod.)
"a green train is coming down the tracks"
"a group of skiers are preparing to ski down a mountain."
"a small kitchen with a low ceiling"
"a group of elephants walking in muddy water."
"a living area with a television and a table"

Figure 12: Random image samples on MS-COCO prompts.

# DALL-E2 limitations

- Confuses physical attributes (like colours and positions)
- Still confuses text generation
- Detailed scenes
- Still contains biases

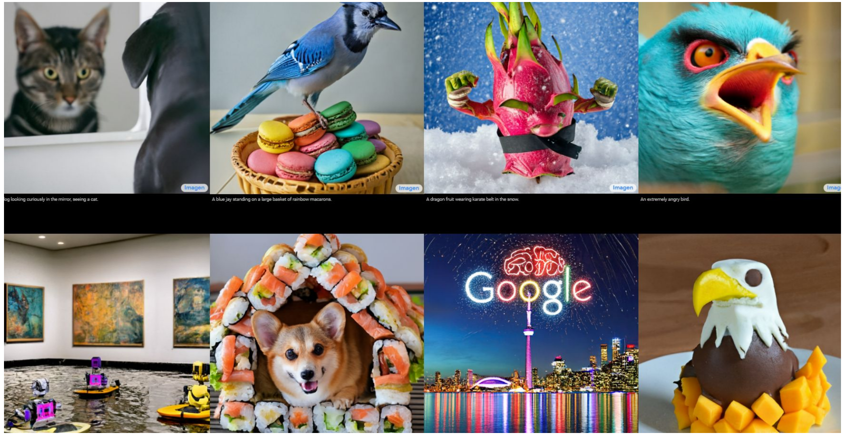
# Open Source models

- VQGAN + Clip
- Big GAN
- Disco diffusion
- Jax diffusion
- Dall-E Mini
- and probably many others…

# Google comes into play

May 2022: **Imagen** (diffusion model)

June 2022: **Parti** (autoregressive model)

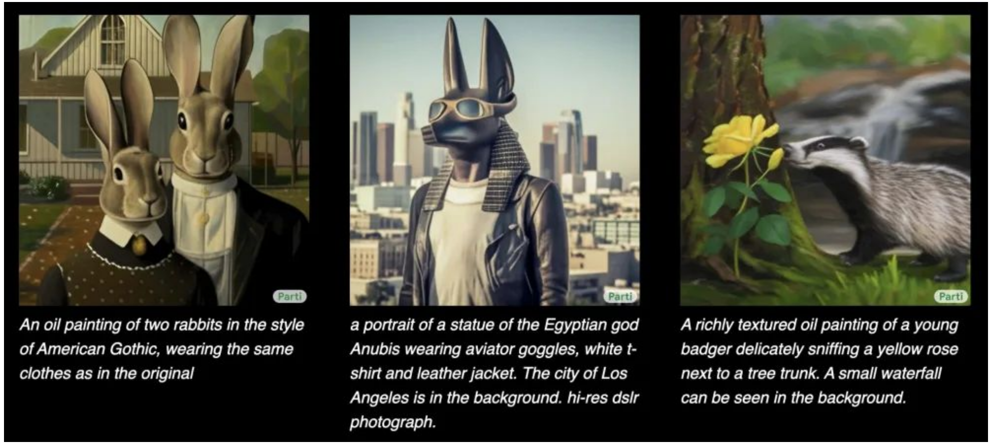


og looking curiously in the mirror, seeing a cat.

A blue jay standing on a large basket of rainbow macarons.

A dragon fruit wearing karate belt in the snow.

An extremely angry bird.



*An oil painting of two rabbits in the style of American Gothic, wearing the same clothes as in the original*

*a portrait of a statue of the Egyptian god Anubis wearing aviator goggles, white t-shirt and leather jacket. The city of Los Angeles is in the background. hi-res dslr photograph.*

*A richly textured oil painting of a young badger delicately sniffing a yellow rose next to a tree trunk. A small waterfall can be seen in the background.*

# Thank you!