Lightricks

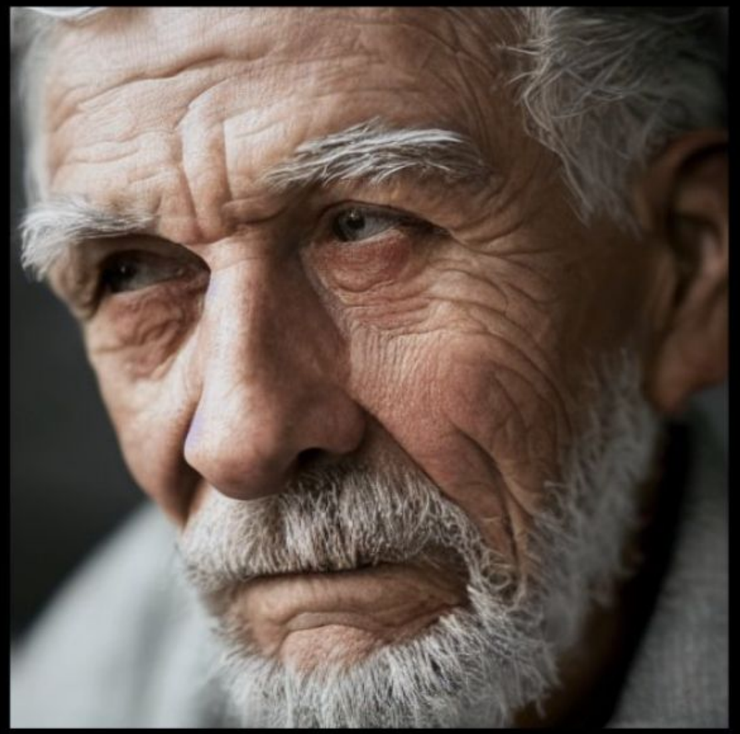
The Text2Image Magic

Naomi Ken Korem
Researcher

MIDJOURNEY

DALL-E 2

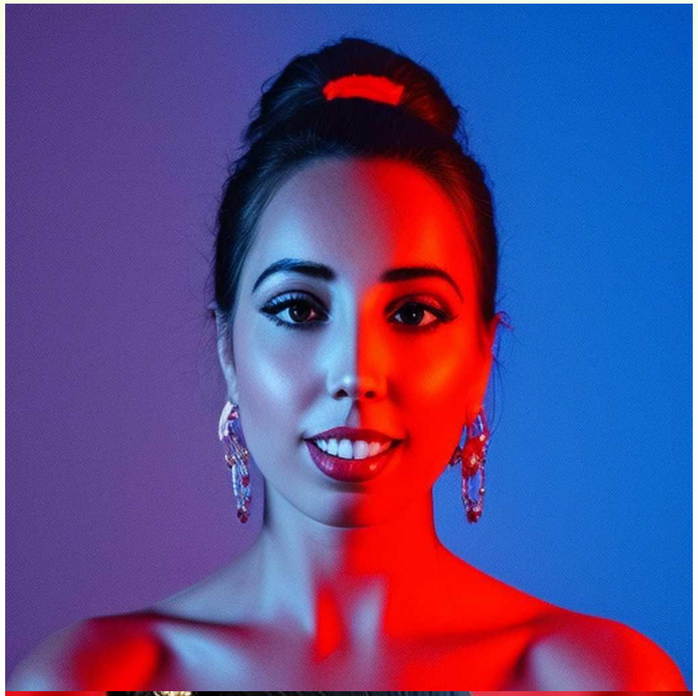
STABLEDIFFUSION

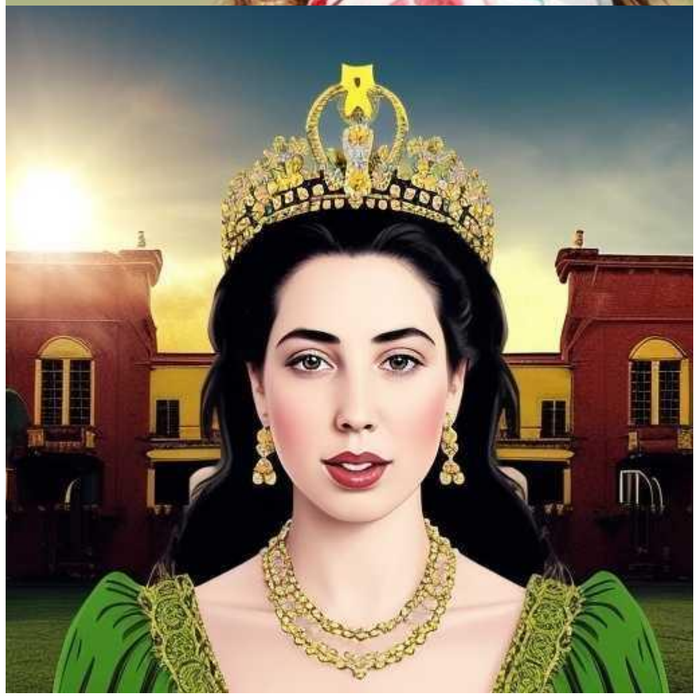


film still, portrait of an old man, wrinkles, dignified look, grey silver hair, peculiar nose, wise, eternal wisdom and beauty, incredible lighting and camera work, depth of field, bokeh, screenshot from a hollywood movie

LAION
@laion_ai

Guiding Stable Diffusion with our CLIP H:
"Professional HDR photo of a polar bear holding a lollipop on a rooftop in Hong Kong looking up at a UFO in the night sky. A UFO flies above the polar bear. The polar bear holds a lollipop on a rooftop. The background shows Hong Kong."





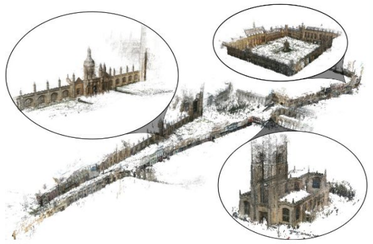
“Add fireworks to the sky”

“Replace the fruits with cake”

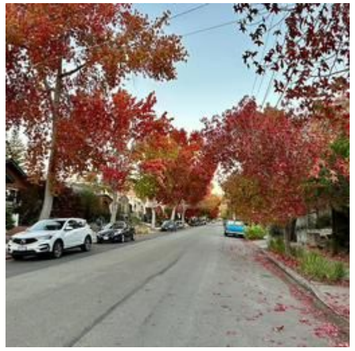
“Turn it into a still from a western”

“Make his jacket out of leather”

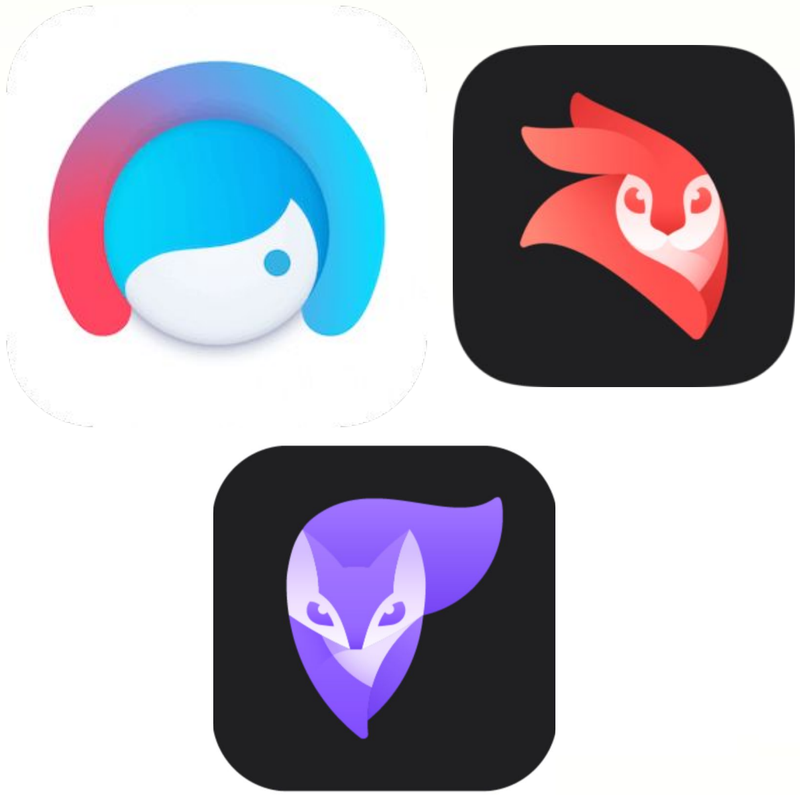
Lightricks



Naomi Ken Korem



Lightricks



Lightricks

Timeline

OpenAI publish DALL-E paper and open source CLIP model

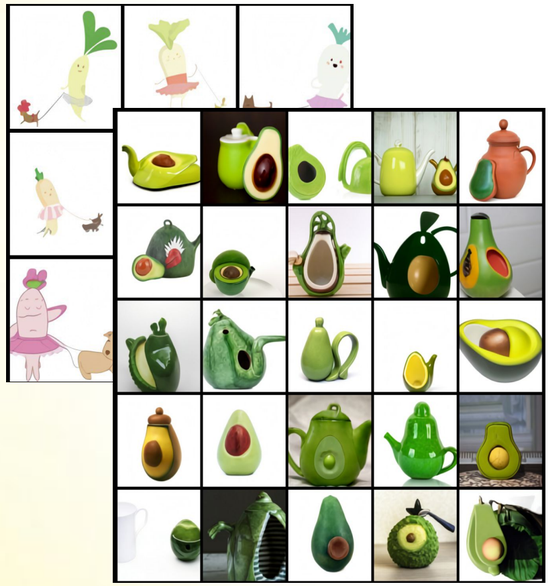
DALLE-2 is published by OpenAI and hype hits mainstream

Google Brain releases Imagen

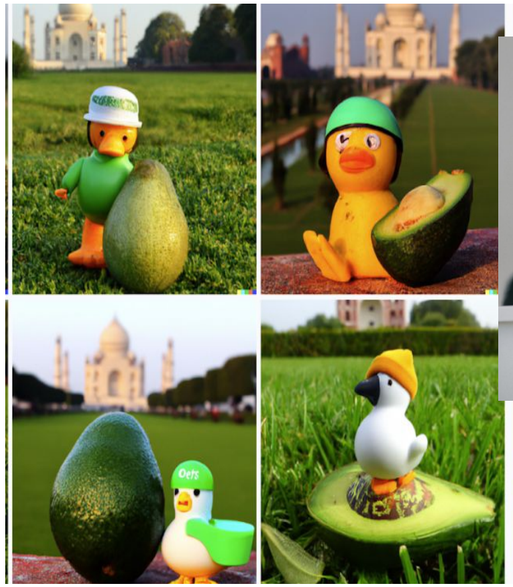
Stability.ai releases Stable Diffusion, open license

A LOT of works based on latest released models

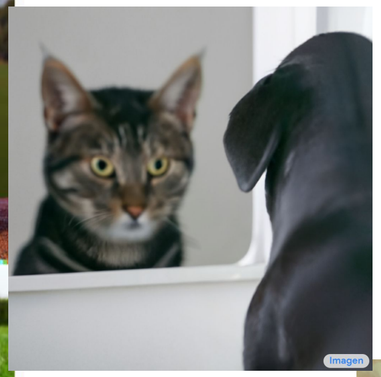
1-2021



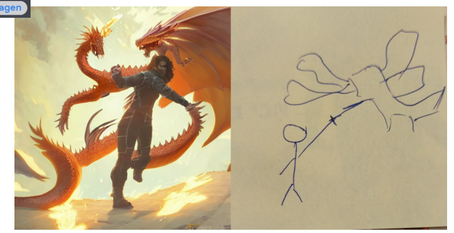
04-2022



5-2022



8-2022



"Add fireworks to the sky"

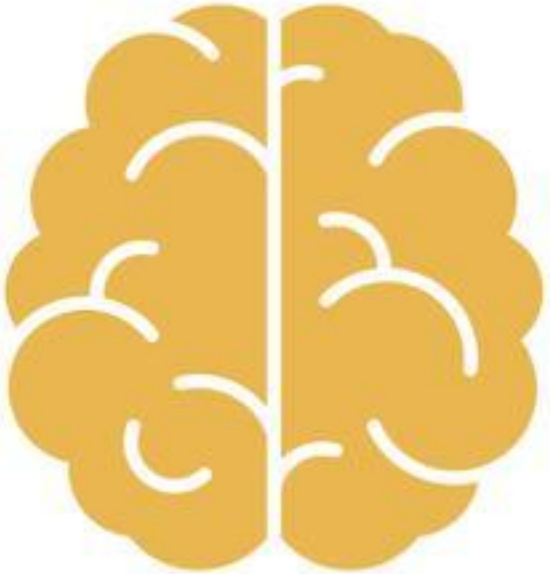
Lightricks

DeepFloyd (just came out)

capybara holding a neon sign with text that reads "capybara podcast", a professional photo of a capybara podcasting, capybara chimera animorph, transformer animal, anamorphic, 8k, 4k, 85 mm, f2. ... by rr.disc0

4 bottles of wine next to each other labeled

System 1



Fast, intuitive and emotional

System 2

Slow, conscious and effortful

What cause this fast progress?

1. **DATA**

   5 Billion image, text pairs!

2. **Parallel training in scale** -

   2.5M training steps of batch size 2048 in very short time. (256 TPU-v4 chips)

3. **Multi domain connections**

   New architectures, layers (Bert, GPT3, StyleGAN, ViT)

Lightricks

The building blocks that got us there -

1. Transformers
2. Diffusion models
3. Text conditioned diffusion models (Dalle2, glide)
4. Stable Diffusion

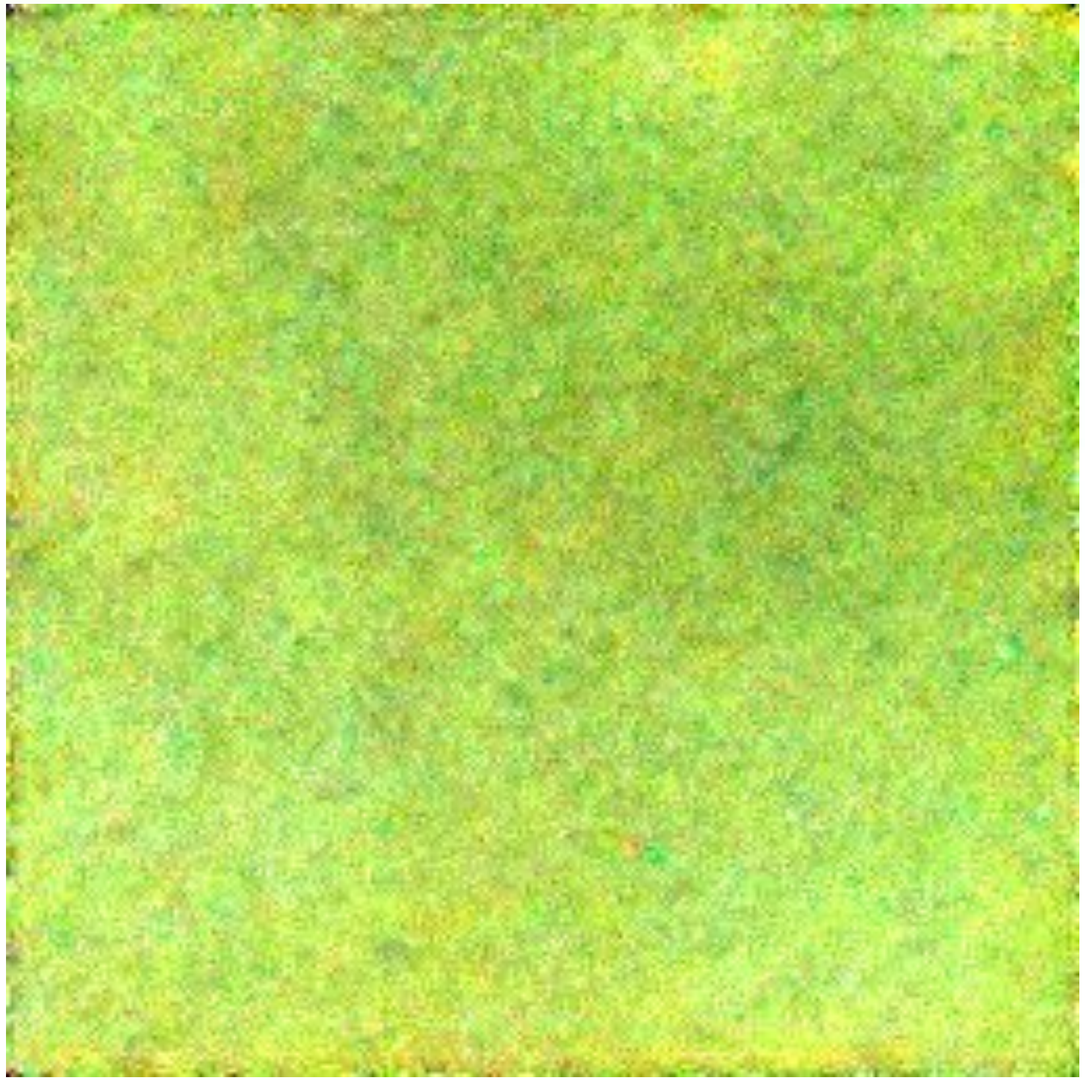
Lightricks

Agenda

1. Diffusion models
2. Hands on #1 training a diffusion model
3. Text 2 Image
4. Hands on #2 Stable diffusion pipelines

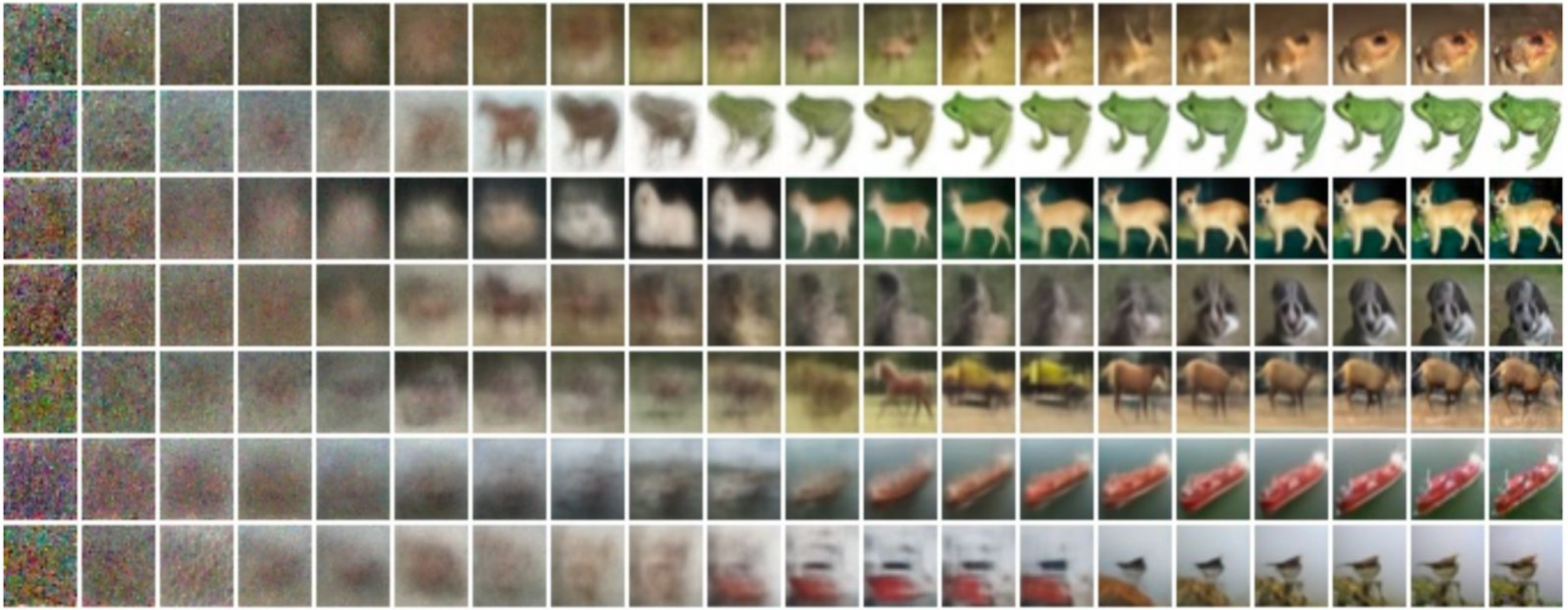
Diffusion models

- Model that learns to **Iteratively de-noise** until image is formed.
-
- Can be **conditioned** on a class of images, text embedding, low-res image, etc.
- **Slow** to run and train, but inference becoming faster and faster these days.

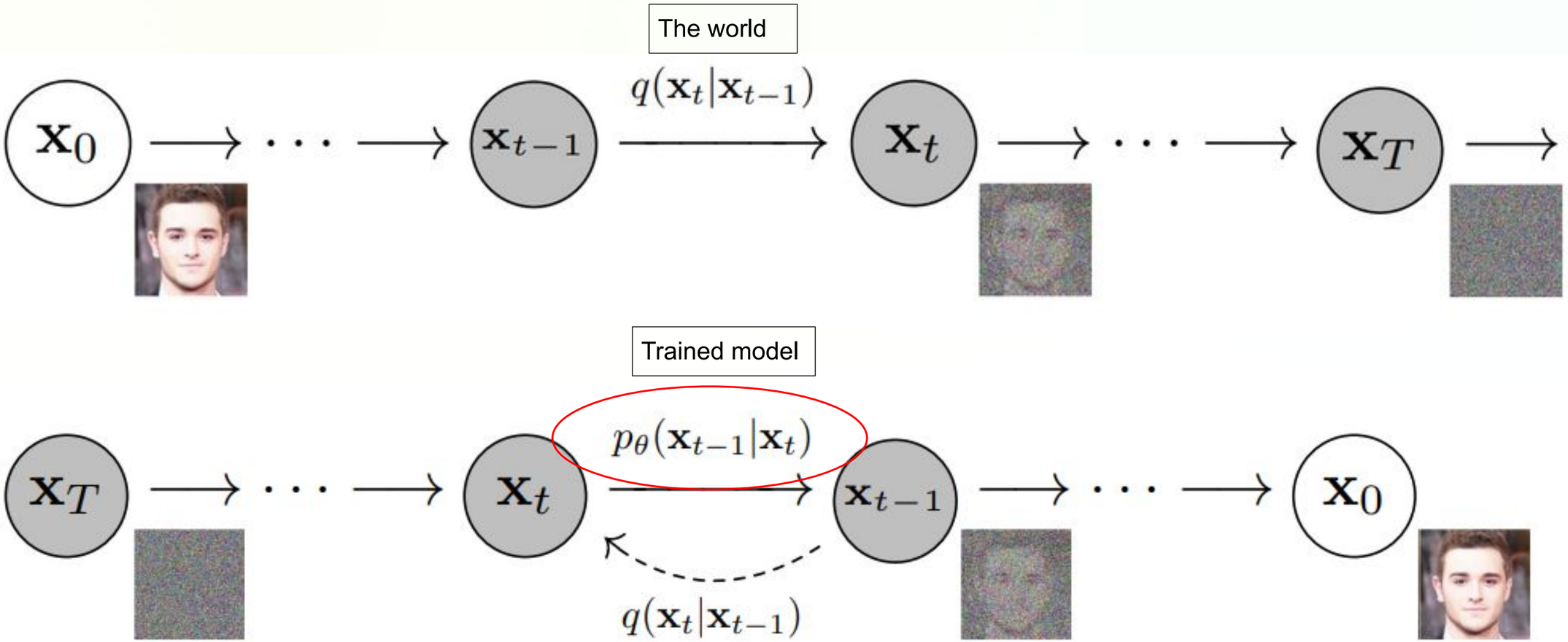


DiscoDiffusion v5, “Cheese and grapes next to a glass of wine”

Diffusion models- from noise to images in steps

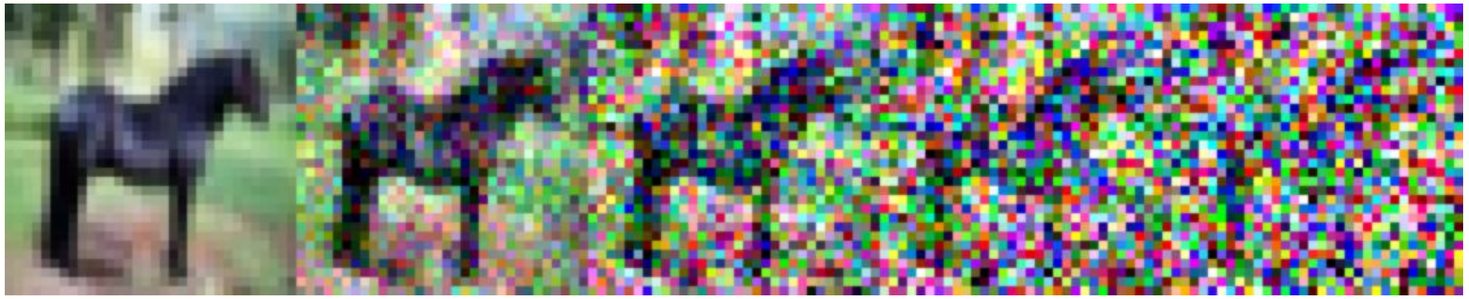


Training a diffusion model



Creating data for training

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})$$



sample an arbitrary step of the noised latents directly conditioned on the input x0

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \text{ where } \bar{\alpha}_t = \prod_{i=1}^{T} \alpha_i \qquad \alpha_t := 1-\beta_t$$

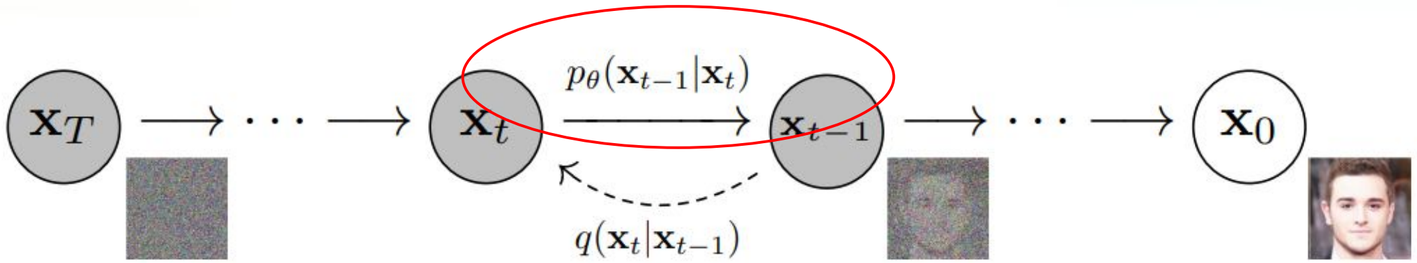
$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$$

Lightricks

Diffusion model - The reverse step

Diffusion model can be trained to predict $\mathbf{x}_{t-1}$ given $\mathbf{x}_t$, t

But, In practice those models are trained to predict the added noise



$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$$

Loss: L2($noise_\theta(\mathbf{x}_t), \epsilon$)

Iteration step: $\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_t}(\mathbf{x}_t - noise_\theta(\mathbf{x}_t)) + \sqrt{(1 - \bar{\alpha}_t)} noise_\theta(\mathbf{x}_t)$

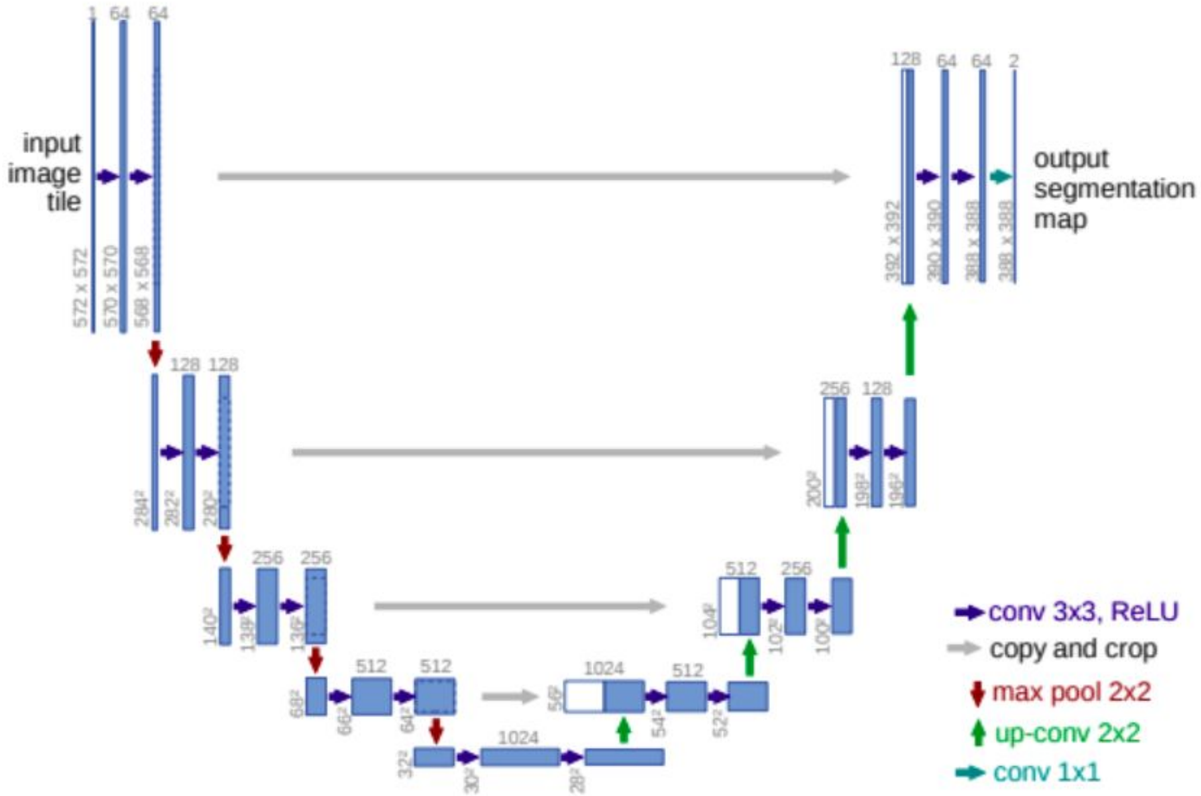
Dimensions:

$\bar{\alpha}_t$ - scalar

$\mathbf{x}_t$ Tensor 64x64x3

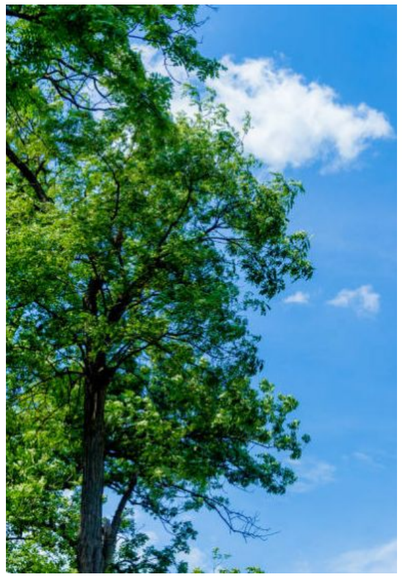
Lightricks

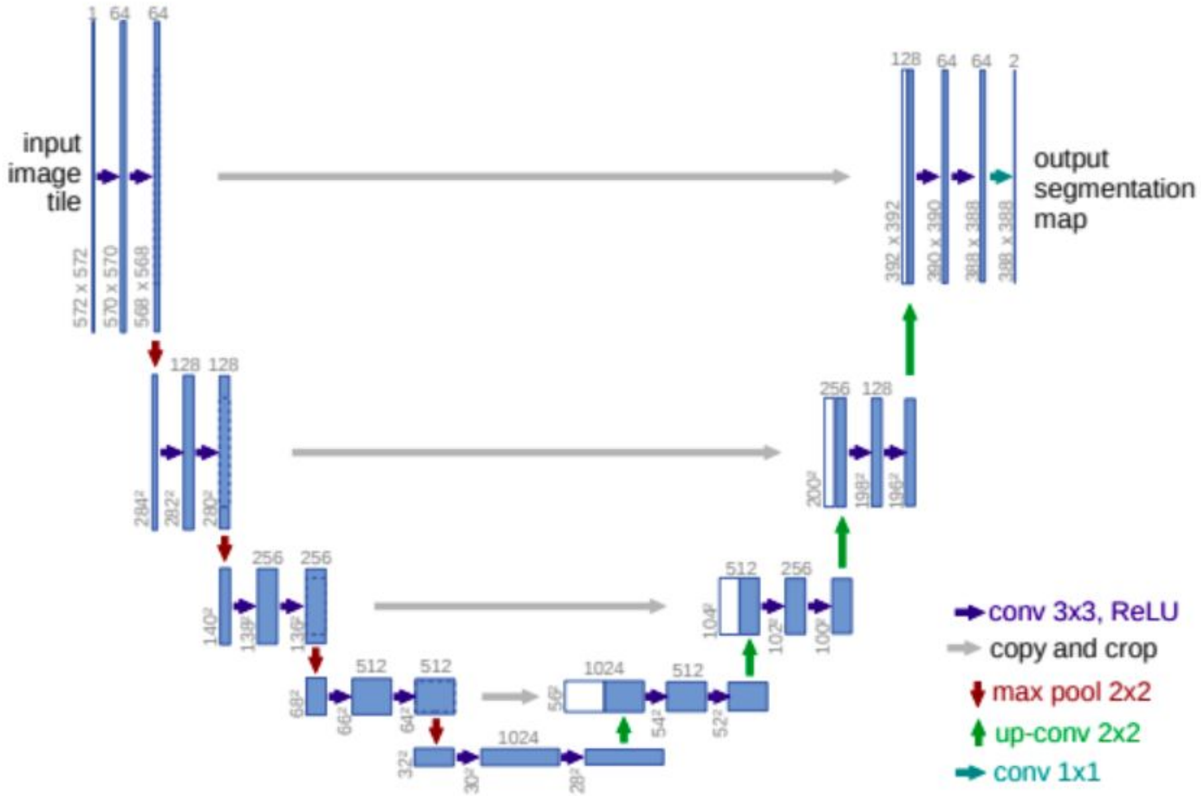
Unet Architecture



Unet 2015

Unet Architecture

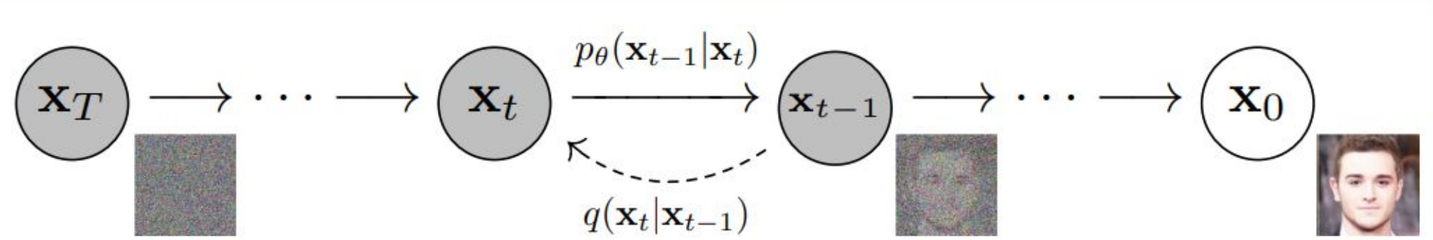


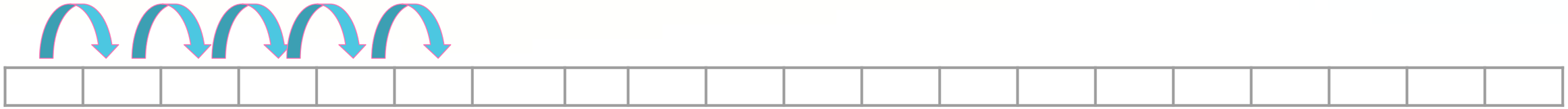


Unet 2015

Inference

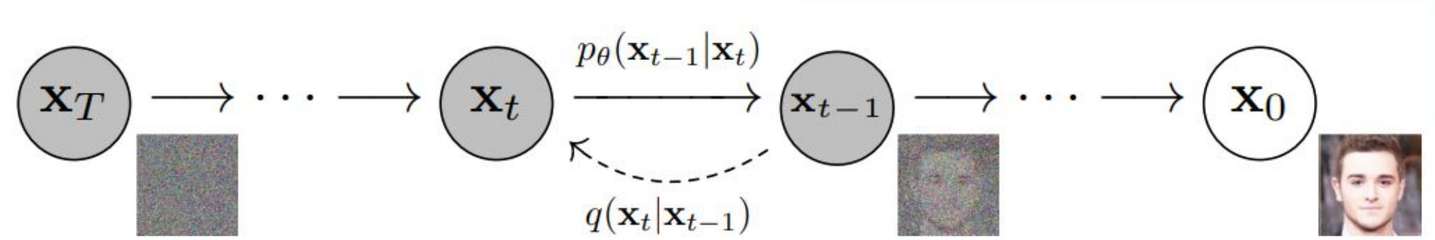
1000 steps??

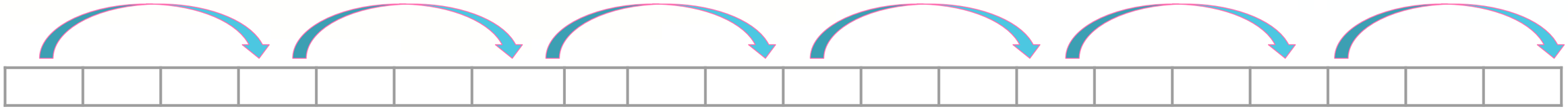




Inference

1000 steps??

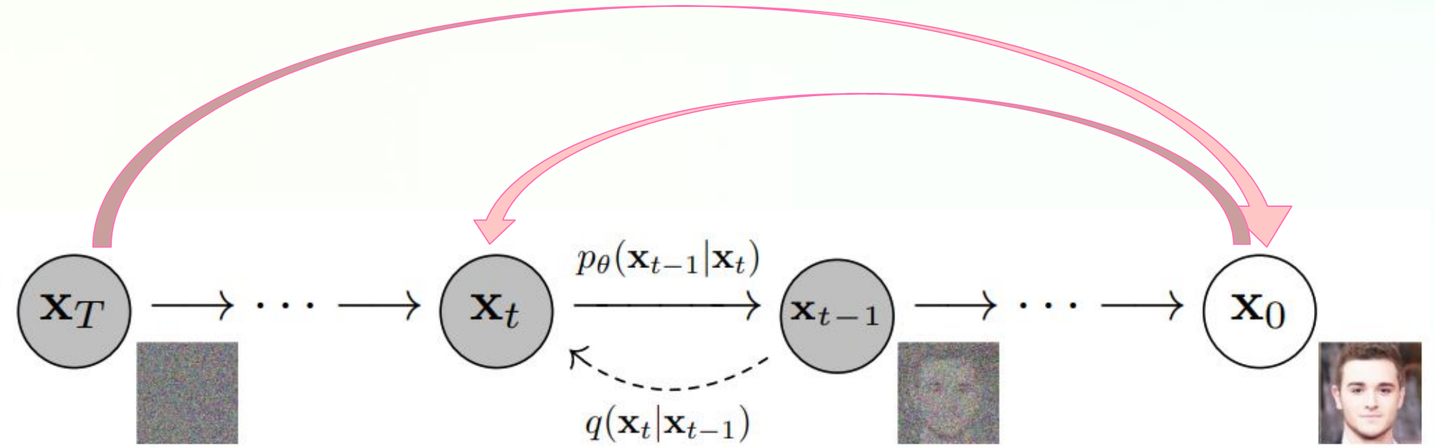


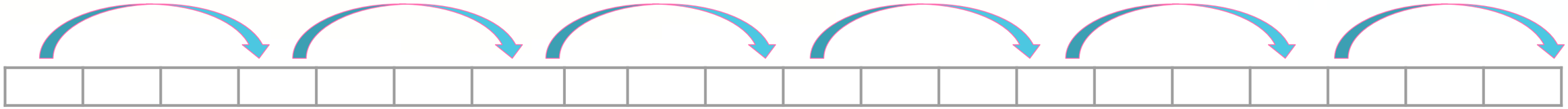


$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$$

Inference

1000 steps??

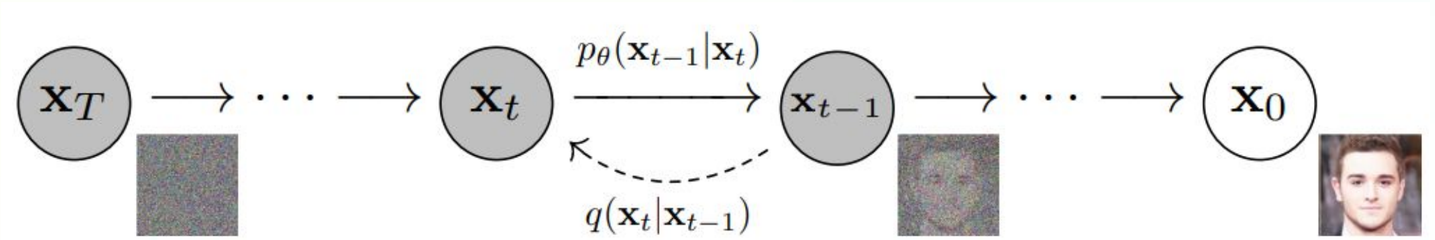


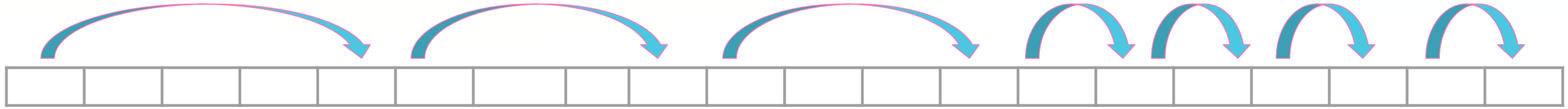


$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$$

Inference

1000 steps??





$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$$

Lightricks

Questions?

Lightricks

Hands on session

Lightricks

Hands on session (30 minutes)

Training a denoising model

1. Go to
https://github.com/Naomi-Ken-Korem/text-to-image-presentation

2. Open Diffusion models part 1.ipynb
(solution:
Diffusion models part 1_solution.ipynb)

Lightricks

Part 2 - Text condition Diffusion models

A dragon fruit wearing karate belt in the snow.

An art gallery displaying Monet paintings. The art gallery is flooded. Robots are going around the art gallery using paddle boards.

A marble statue of a Koala DJ in front of a marble statue of a turntable. The Koala is wearing large marble headphones.

GLIDE

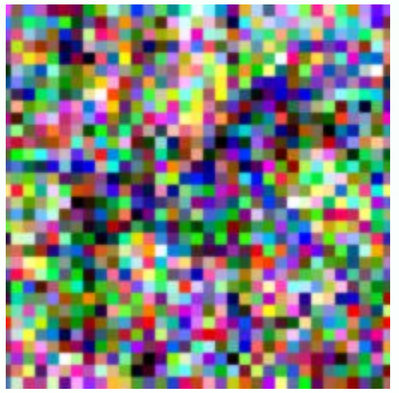
OpenAI, Dec 2021

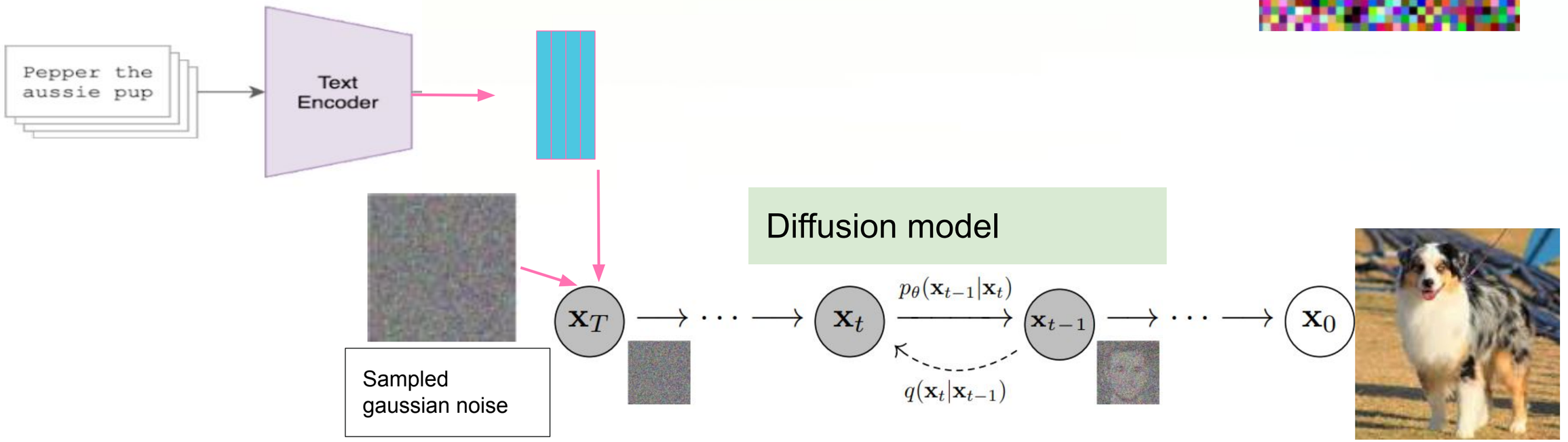
Given a text prompt + random noise image, generates a related image using a text transformer model based on the model of “Attention is all you need” (Bert) and a conditioned diffusion model.

“a boat in the canals of venice”

“a painting of a fox in the style of starry night”

Glide text 2 image architecture



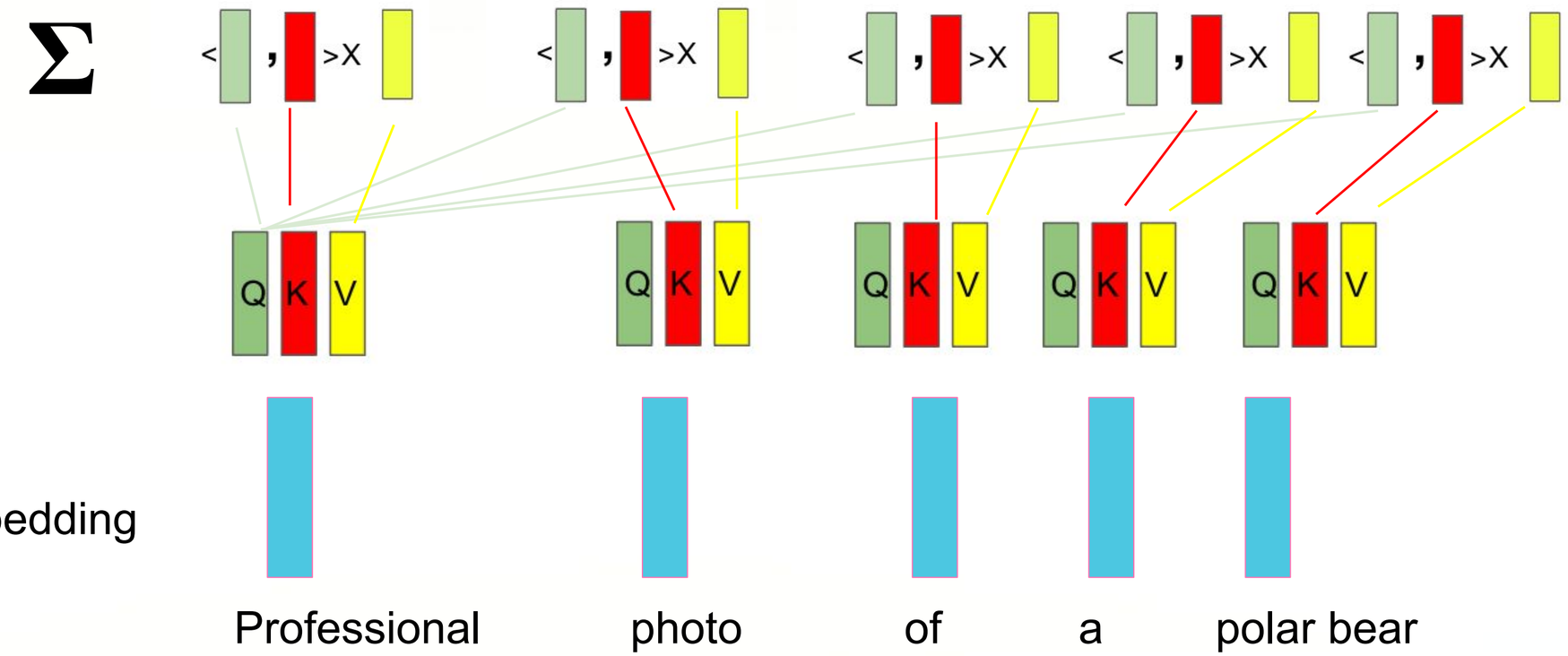


Lightricks

Transformers

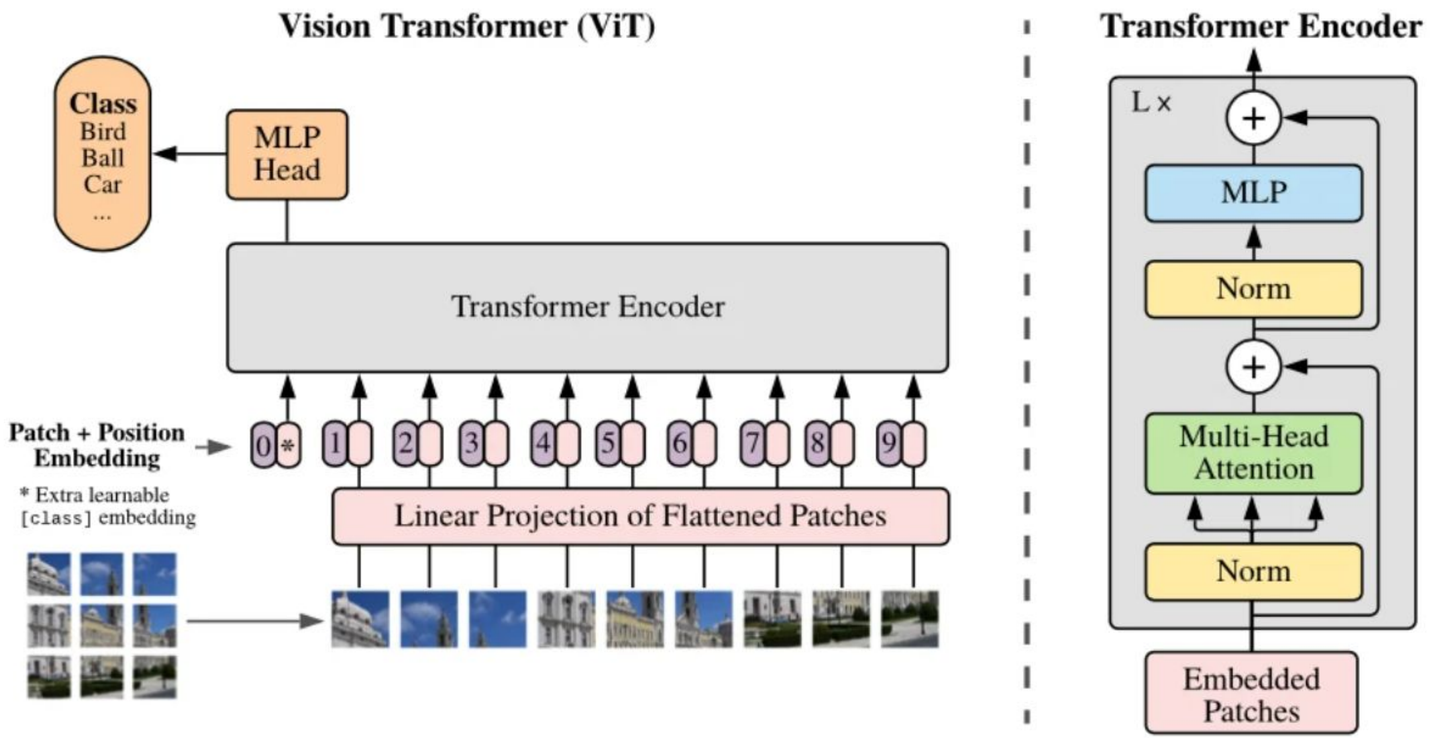
Google, Dec 2017

Attention Layer



[Attention is all you need]

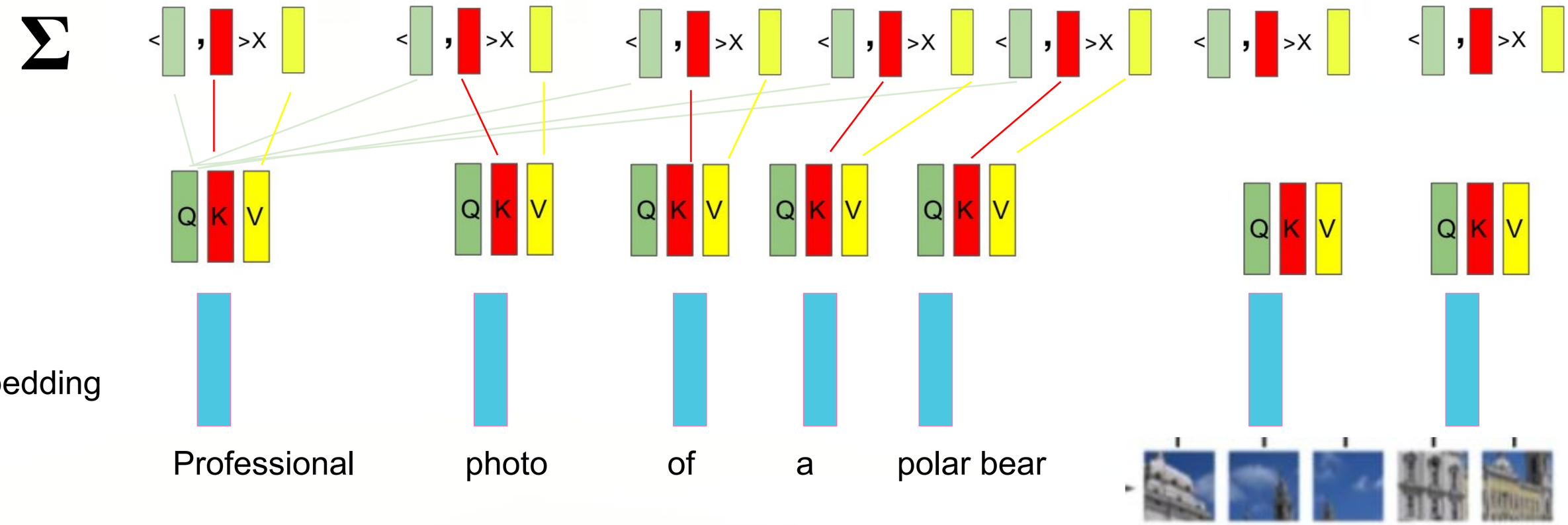
Visual Transformers



[An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale]

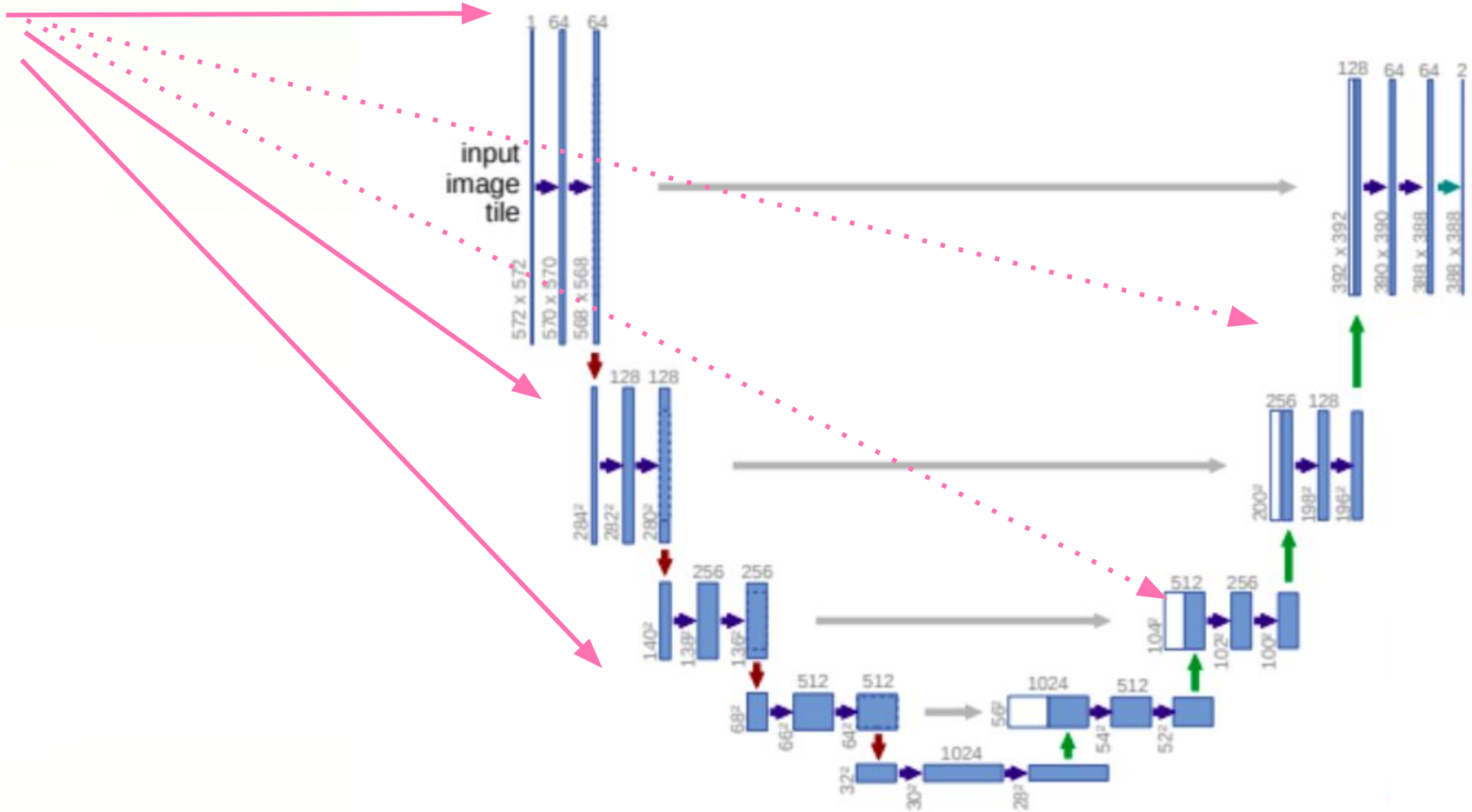
Lightricks

Cross Attention Layer



Lightricks

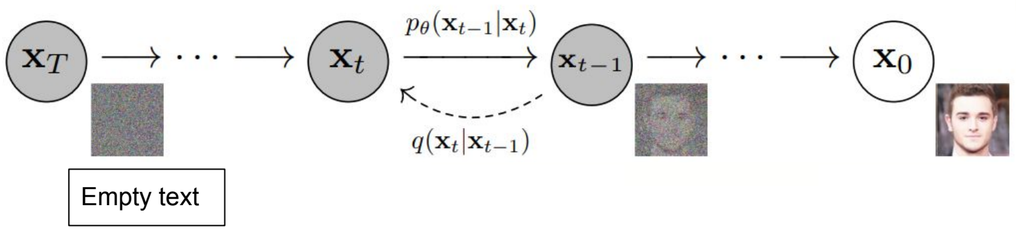
Adding textual input into the diffusion model

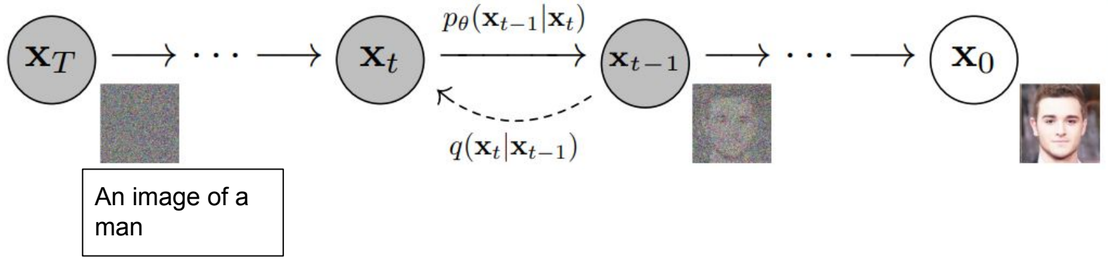


Lightricks

Guidance scale trick-

extrapolation in the direction of the textual concept





$$x = \mathbf{x}_{empty} + guidancescale * (\mathbf{x}_{prompt} - \mathbf{x}_{empty})$$

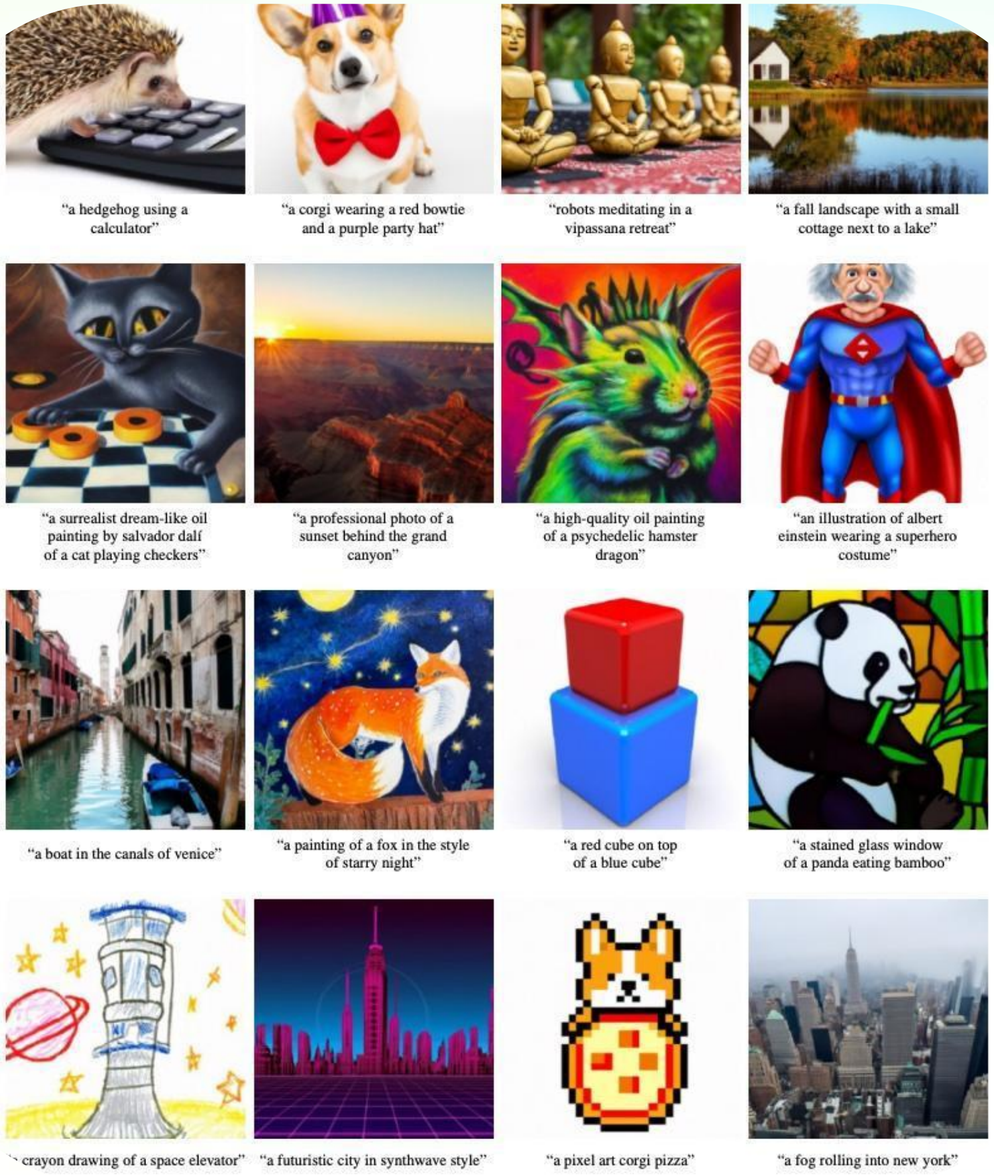
$$\mathbf{x}_{empty:t-1} = P(x_t, \text{""})$$

$$\mathbf{x}_{prompt:t-1} = P(x_t, \text{" prompt "})$$

Lightricks

Seminal publications

Now we can understand Imagen and Dalle-2



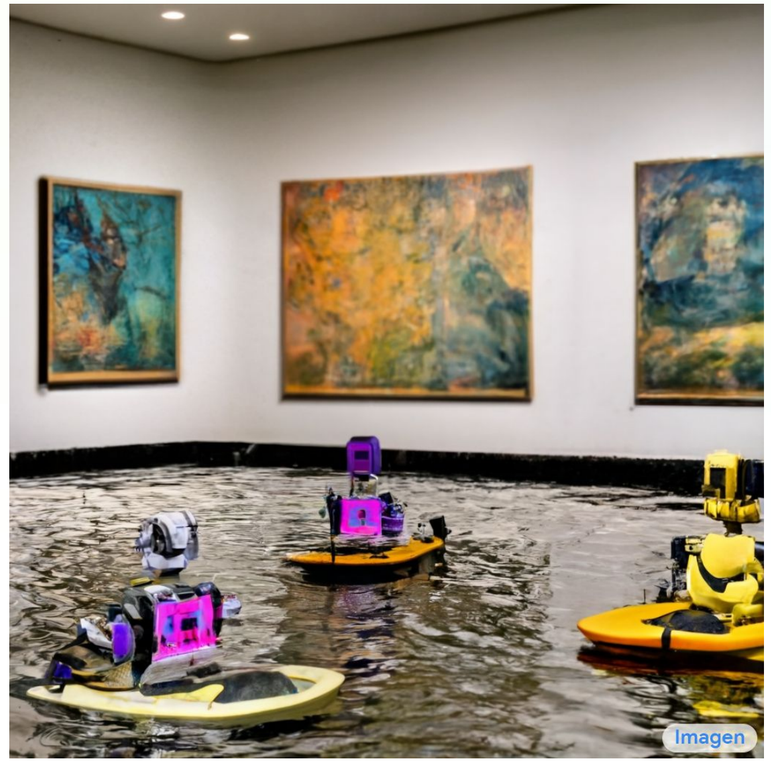
Imagen

Google Brain, May 2022

Huge text model (T5-XXL)

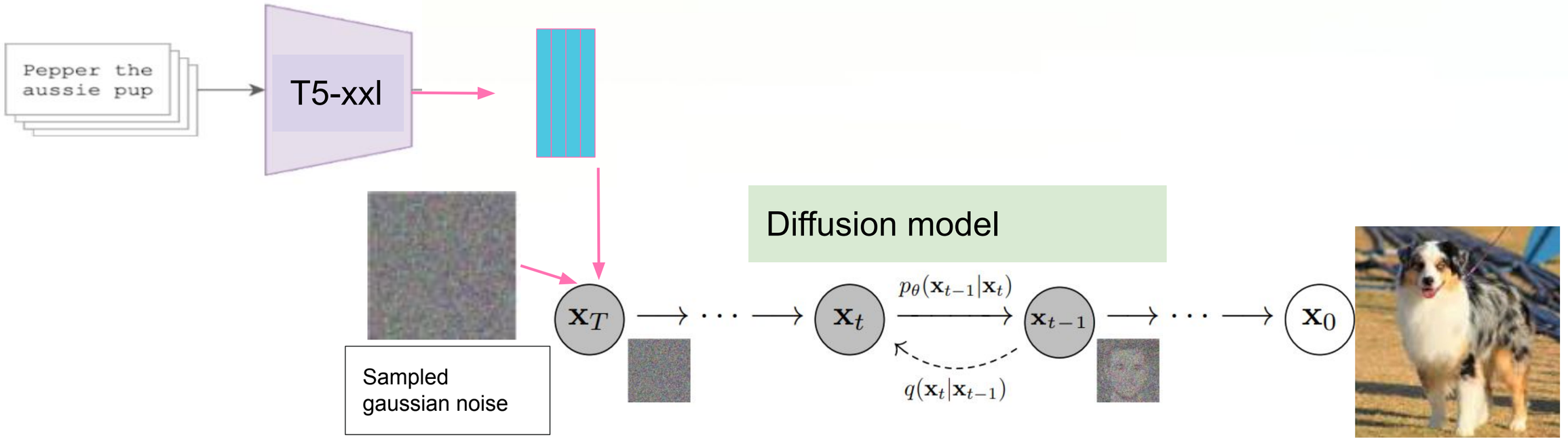
Given text embedding generates an image using a guided diffusion model.

Learn two 4x super-resolution steps (16x in total).



An art gallery displaying Monet paintings. The art gallery is flooded. Robots are going around the art gallery using paddle boards.

Imagen (and deepFloyd) architecture

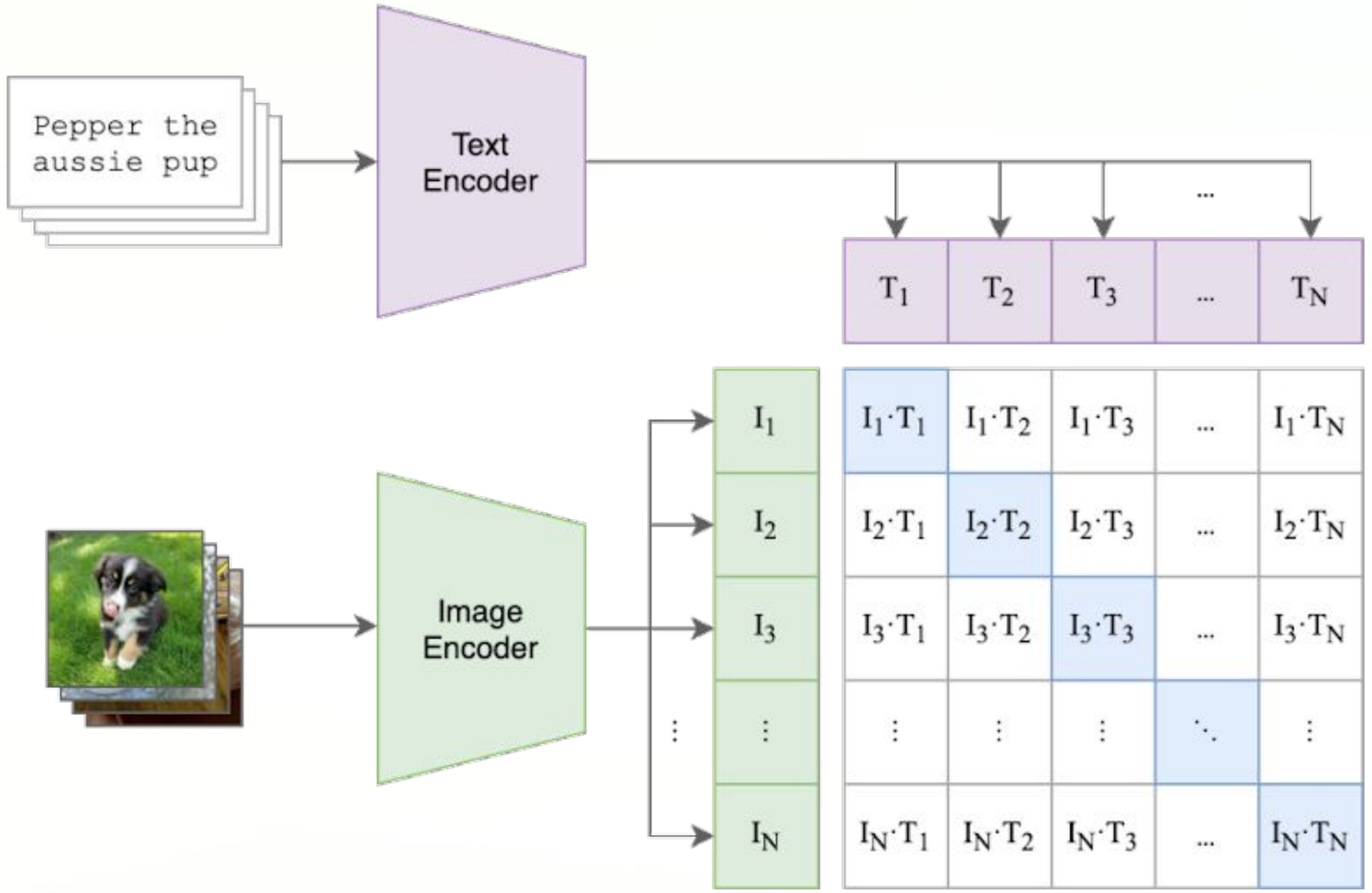


Lightricks

Imagen: Super Resolution models

CLIP- Text and Image encoder

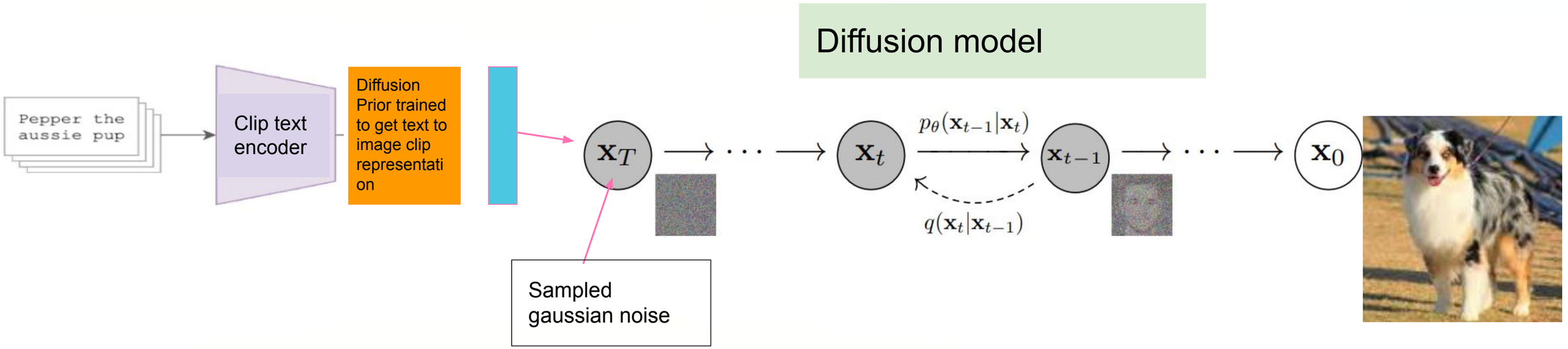
OpenAI, Jan 2021



Lightricks

Dalle-2
openAI april 2022

Unclip - from clip image vector back into image



...

Emad
@EMostaque

Delighted to announce the public open source release of #StableDiffusion!

Please see our release post and retweet!
stability.ai/blog/stable-di...

Proud of everyone involved in releasing this tech that is the first of a series of models to activate the creative potential of humanity

תרגם את הציוץ

stability.ai
Stable Diffusion Public Release — Stability.Ai
We are delighted to announce the public release of Stable Diffusion and the launch of DreamStudio Lite.

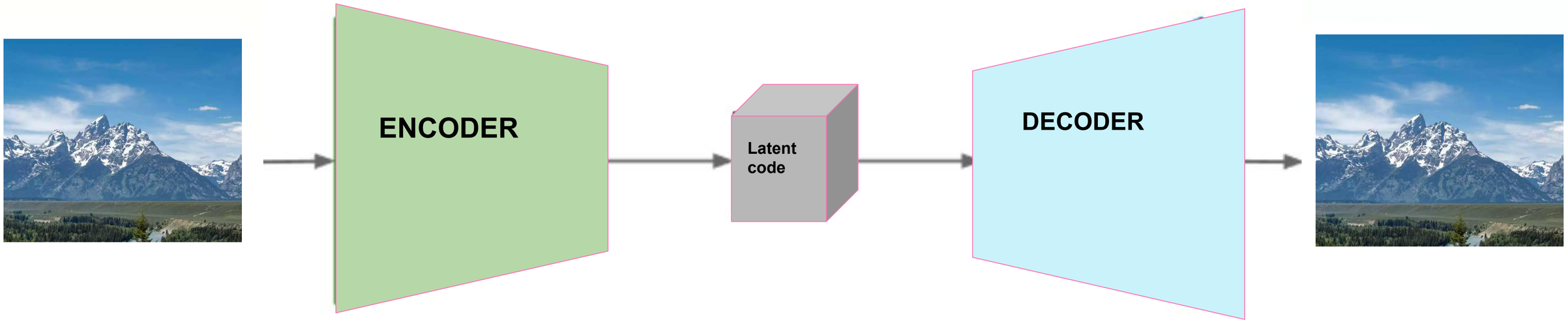
9:07 אחה"צ · 22 באוג' 2022 · Twitter Web App

Stable Diffusion
(Image to Image pipe)

High-Resolution Image Synthesis with Latent Diffusion Models
Robin Rombach*, Andreas Blattmann*, Dominik Lorenz, Patrick Esser, Björn Ommer

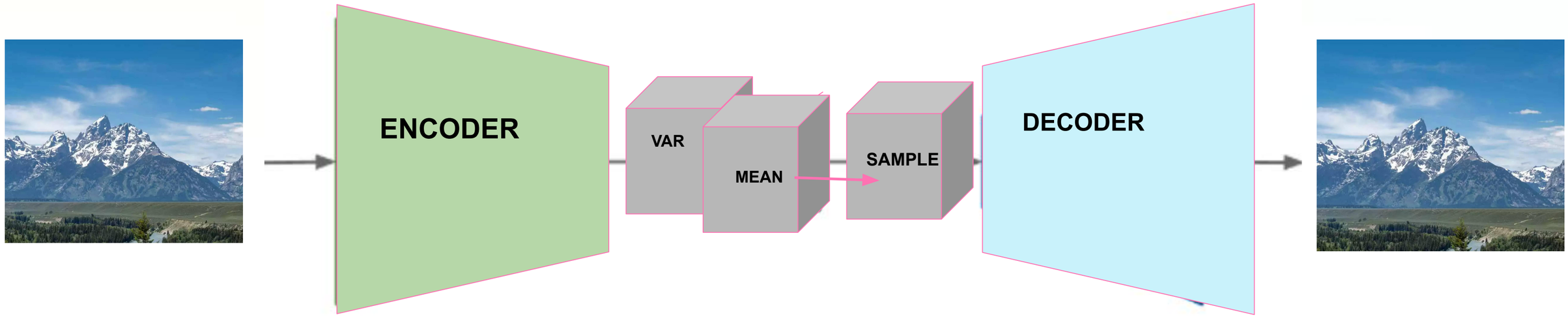
Lightricks

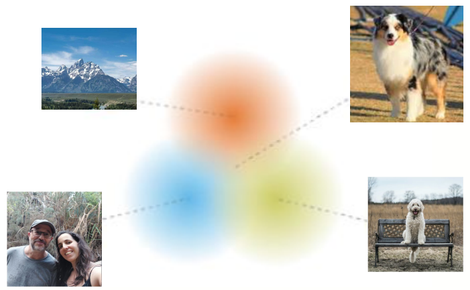
AutoEncoder



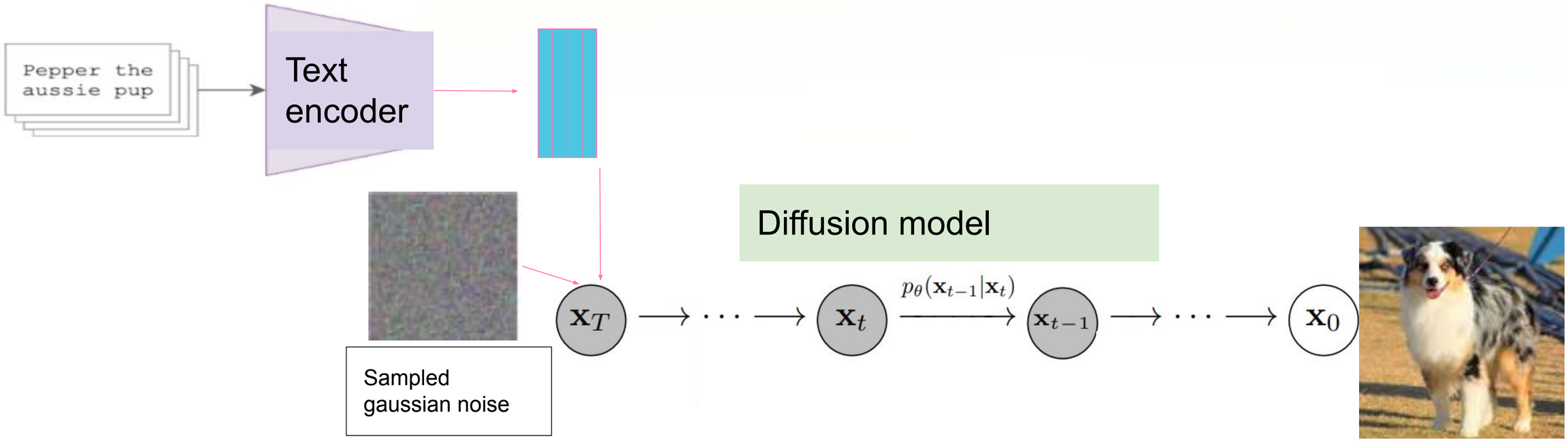
Lightricks

Variational AutoEncoder



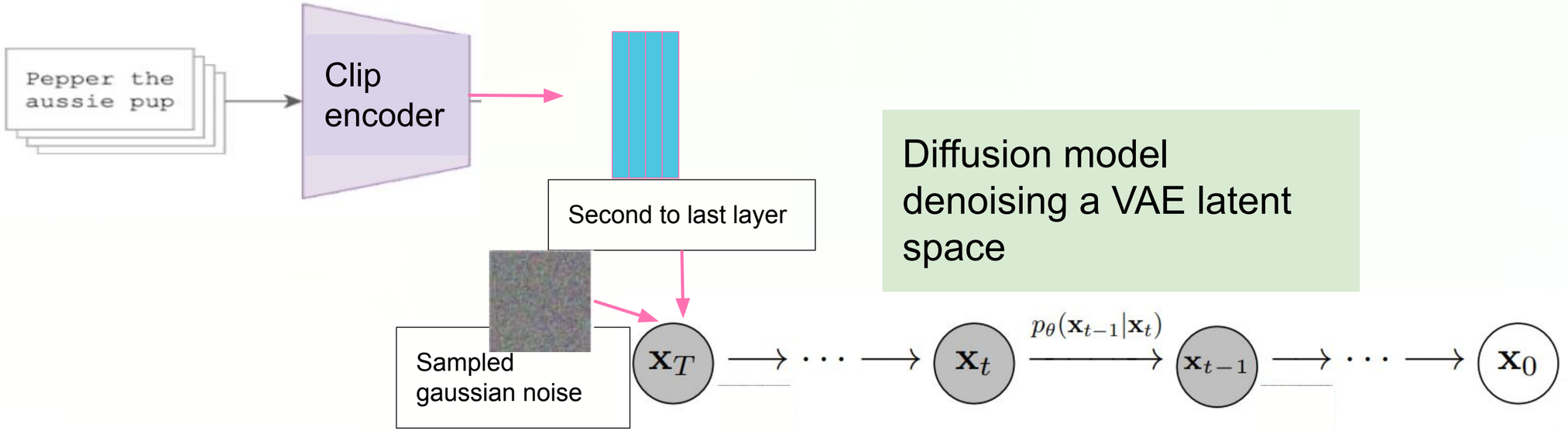


Original Text to Image pipe



Lightricks

Stable Diffusion Text to Image pipe



Diffusion model denoising a VAE latent space

VAE Decoder

[High-Resolution Image Synthesis with Latent Diffusion Models]

Lightricks

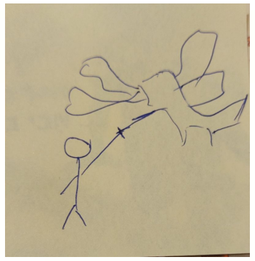
Editing pipelines

Lightricks

Stable Diffusion Image to Image pipe

A hero fighting a fire-breathing dragon

Clip encoder

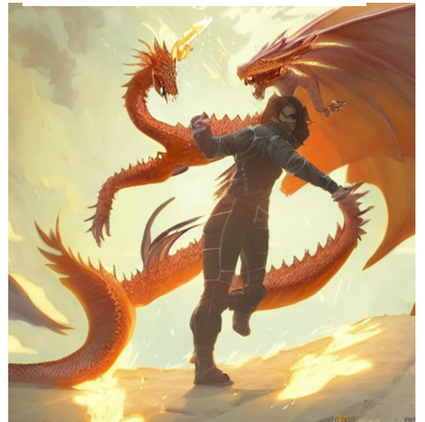


Encoder $q_\phi(\mathbf{z}|\mathbf{x})$

Diffusion model denoising a VAE latent space

$\mathbf{x}_T \longrightarrow \cdots \longrightarrow \mathbf{x}_t \xrightarrow{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} \mathbf{x}_{t-1} \longrightarrow \cdots \longrightarrow \mathbf{x}_0$

VAE Decoder



[High-Resolution Image Synthesis with Latent Diffusion Models]

Lightricks

Stable Diffusion Inpainting pipe (legacy)

A magic castle on the mountains

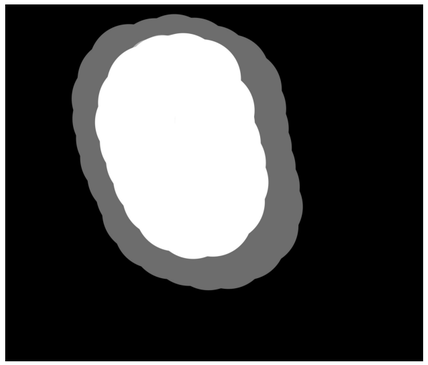
Clip encoder

Encoder $q_\phi(\mathbf{z}|\mathbf{x})$

Diffusion model denoising a VAE latent space

$\mathbf{x}_T \longrightarrow \cdots \longrightarrow \mathbf{x}_t \xrightarrow{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} \mathbf{x}_{t-1} \longrightarrow \cdots \longrightarrow \mathbf{x}_0$

VAE Decoder



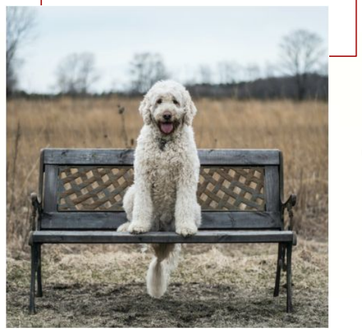
[High-Resolution Image Synthesis with Latent Diffusion Models]

Lightricks

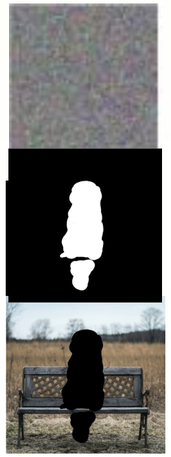
Stable Diffusion Inpainting pipe

Face of a yellow cat, high resolution, sitting on a park bench

Clip encoder



Encoder $q_\phi(\mathbf{z}|\mathbf{x})$



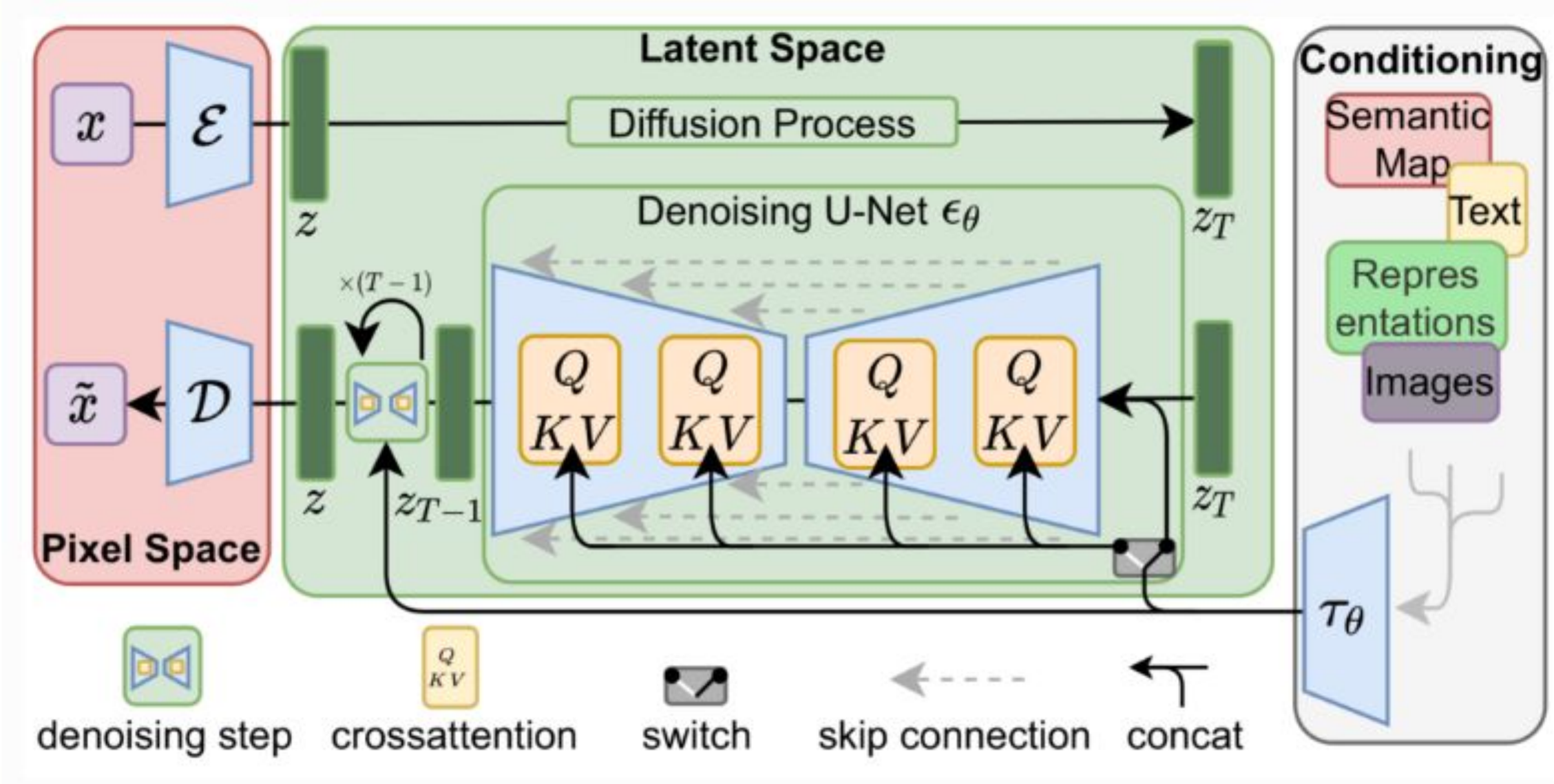
Inpainting finetuned diffusion model denoising a VAE latent space

$$\mathbf{x}_T \longrightarrow \cdots \longrightarrow \mathbf{x}_t \xrightarrow{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} \mathbf{x}_{t-1} \longrightarrow \cdots \longrightarrow \mathbf{x}_0$$

VAE Decoder

Lightricks

Inpainting


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

Useful gits to follow

https://github.com/huggingface/diffusers

https://github.com/AUTOMATIC1111/stable-diffusion-webui

Good image 2 image example:

https://colab.research.google.com/github/patil-suraj/Notebooks/blob/master/image_2_image_using_diffusers.ipynb

Lightricks

Questions?

Hands on session (40 minutes)

Stable_diffusion_editing_pipes.ipynb

Takeaway

1. You now know what is it diffusion models

2. We covered a lot of architectures, and technical ideas. Combining those existing ideas leads to great progress

3. It should be pretty easy for you to start playing with diffusers git

Lightricks

Thank you!

Lightricks

Instruct pix2pix

Tim Brooks* Aleksander Holynski* Alexei A.
Efros University of California, Berkeley
Jan, 2023

Lightricks

Instruct-pix2pix pipe

Turn her into a snake lady

Clip encoder

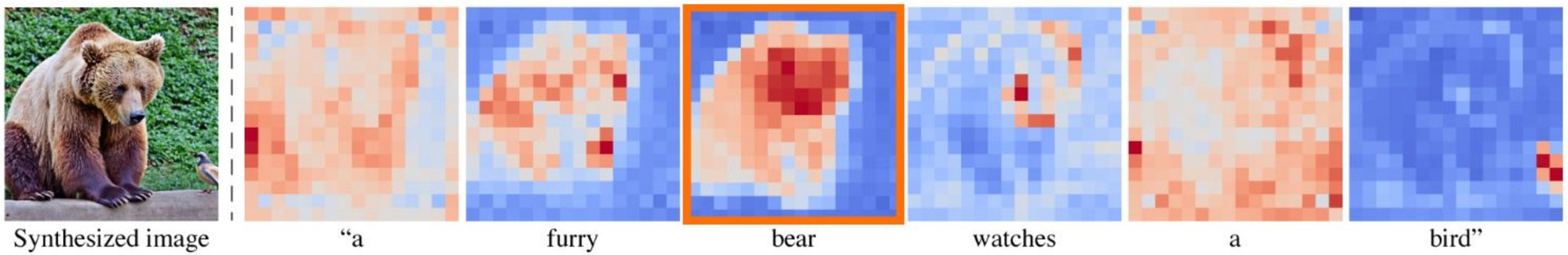
Instruct pix2pix model

Encoder $q_\phi(\mathbf{z}|\mathbf{x})$

$\mathbf{x}_T \longrightarrow \cdots \longrightarrow \mathbf{x}_t \xrightarrow{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} \mathbf{x}_{t-1} \longrightarrow \cdots \longrightarrow \mathbf{x}_0$

VAE Decoder

Cross attention visualization



Training Data Generation

(a) Generate text edits:

Input Caption: *"photograph of a girl riding a horse"* → GPT-3 → Instruction: *"have her ride a dragon"*

Edited Caption: *"photograph of a girl riding a dragon"*

700 sampels

(b) Generate paired images:

Input Caption: *"photograph of a girl riding a horse"*

Edited Caption: *"photograph of a girl riding a dragon"*

→ Stable Diffusion + Prompt2Prompt →

(c) Generated training examples:

"convert to brick"

"Color the cars pink"

"Make it lit by fireworks"

"have her ride a dragon"

...