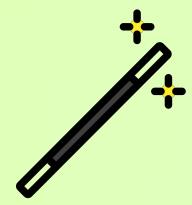




Controlling Generative AI



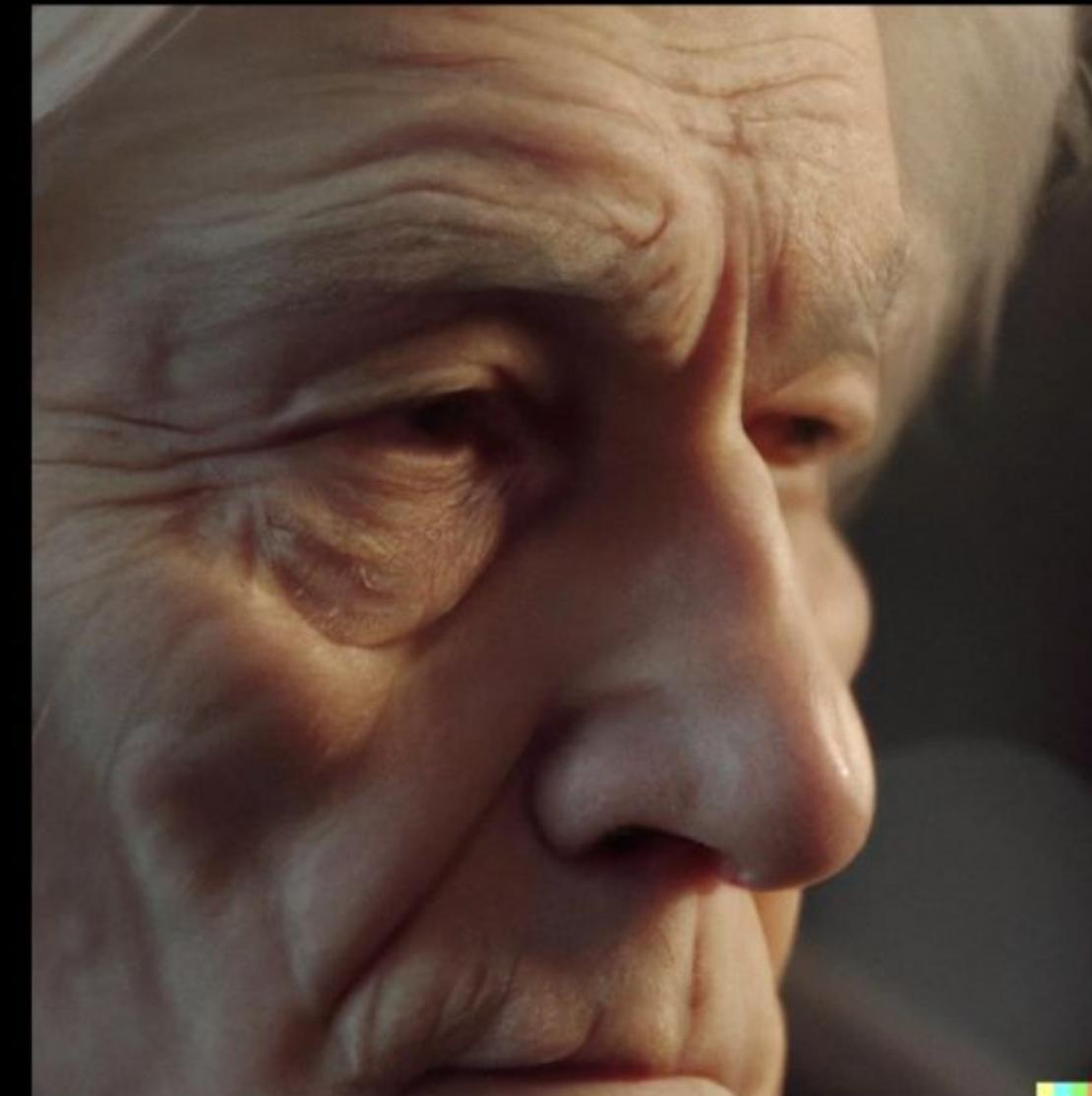
Naomi Ken Korem
Researcher



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



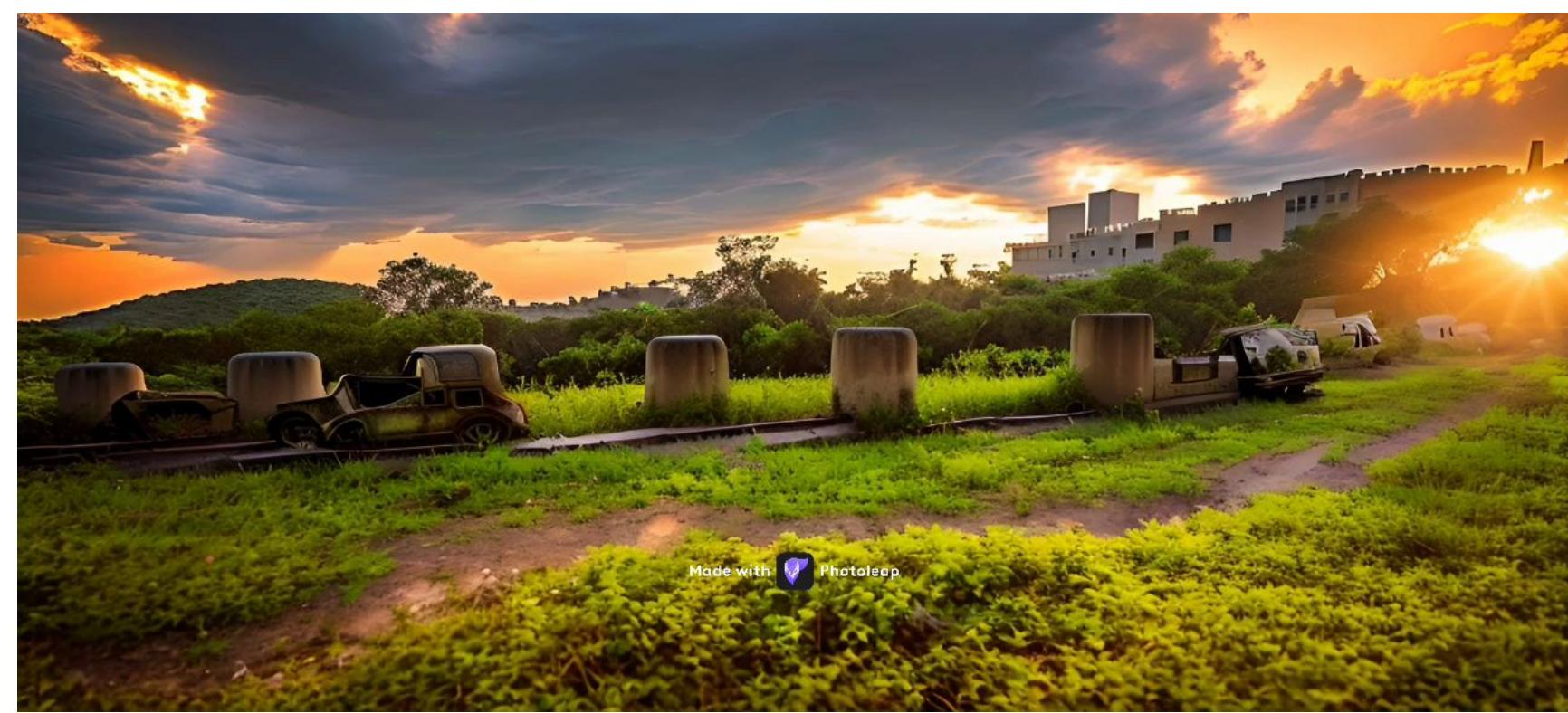
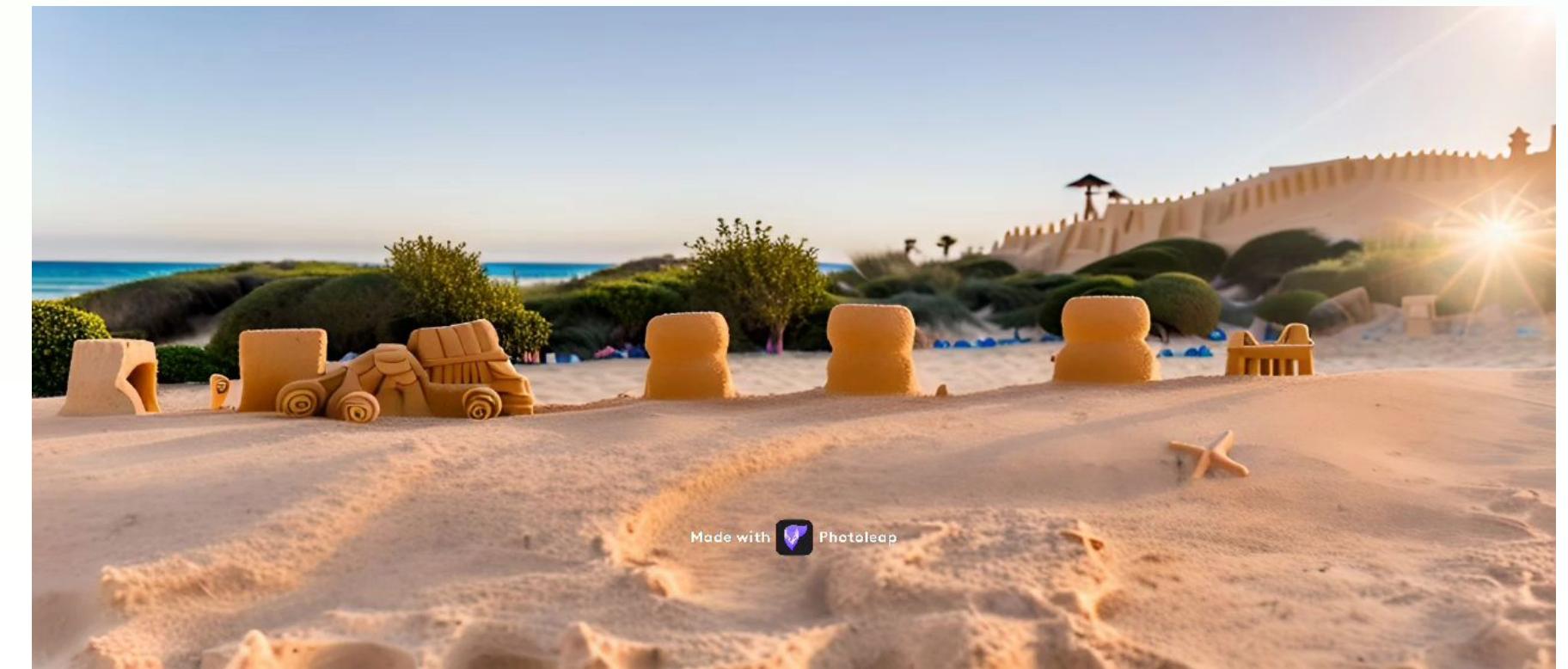
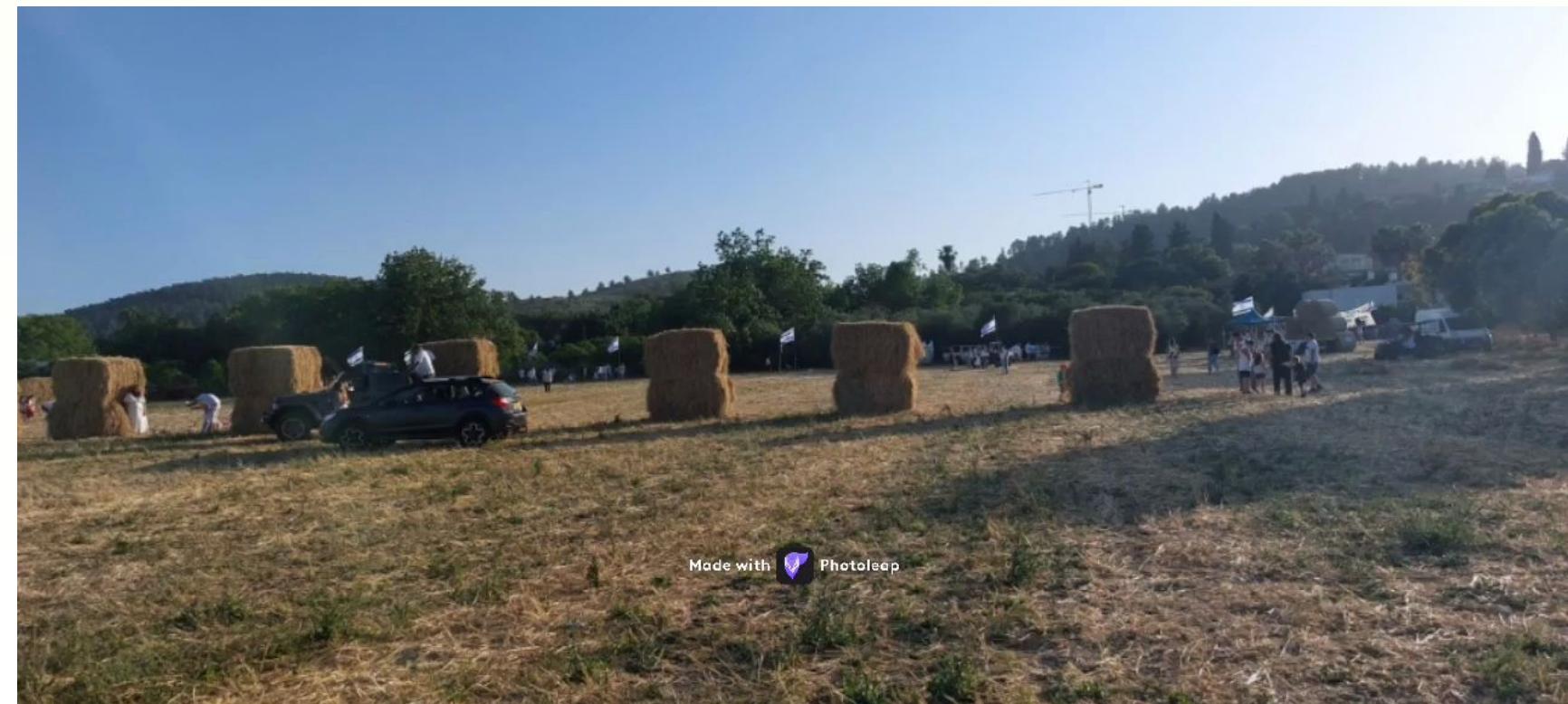
LAION
@laion_ai

...

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."



It is all about the control



"Add fireworks to the sky"



"Replace the fruits with cake"

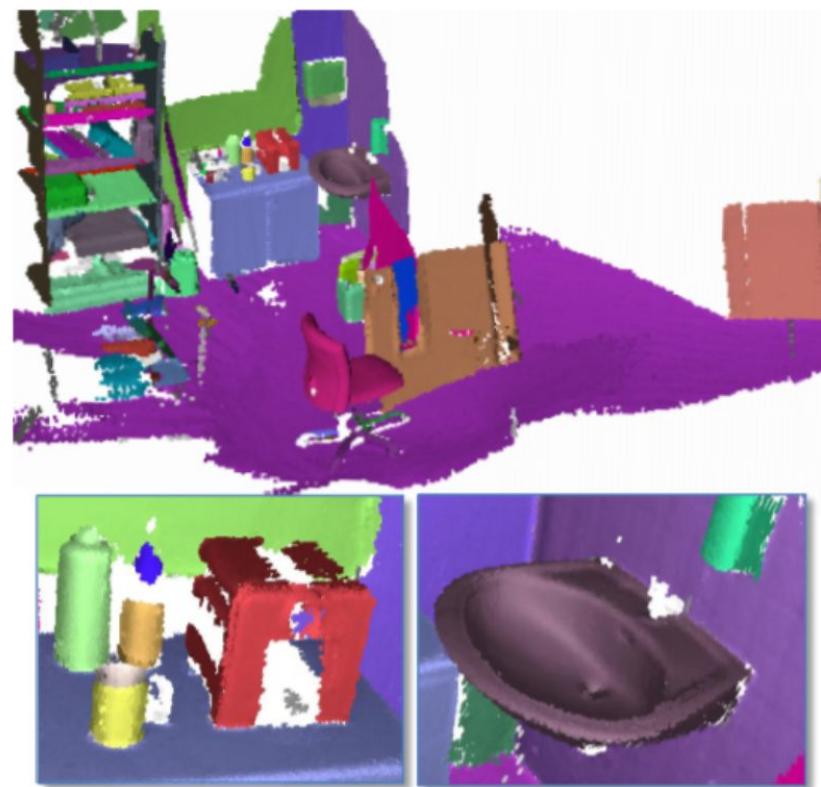


"Turn it into a still from a western"

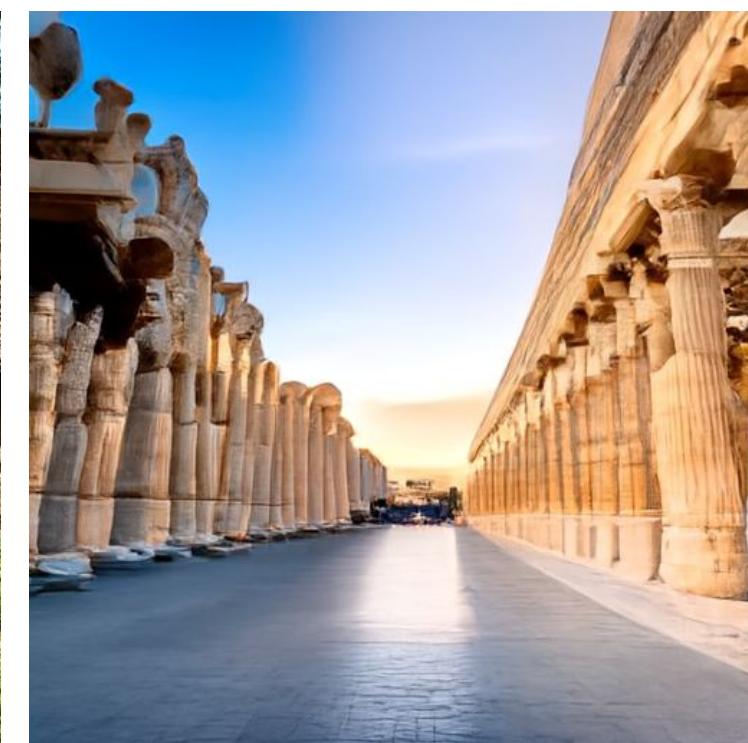


"Make his jacket out of leather"





Naomi Ken Korem



Lighttricks

Bridge the gap between imagination and creation



AGENDA

- Unet, Gans, StyleGAN and Gan Control
- Diffusion models, Text to Image models
- Controlling text 2 Image models



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.



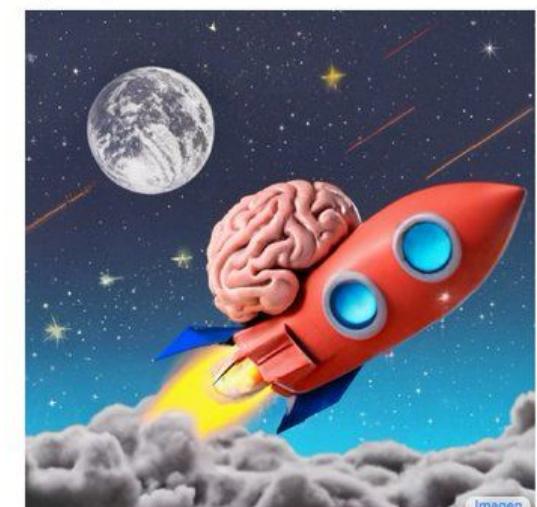
Teddy bears swimming at the Olympics 400m Butterfly event.



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



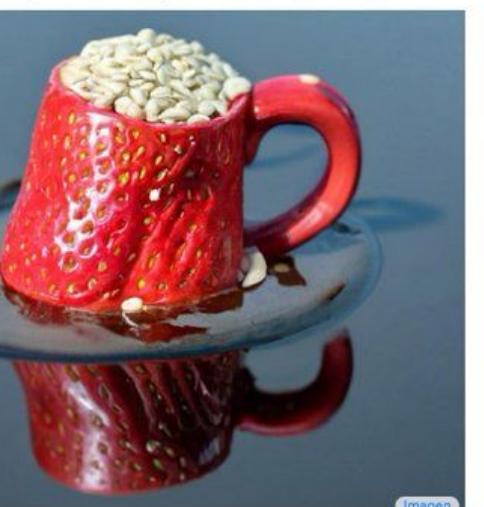
A cute sloth holding a small treasure chest. A bright golden glow is coming from the chest.



A brain riding a rocketship heading towards the moon.



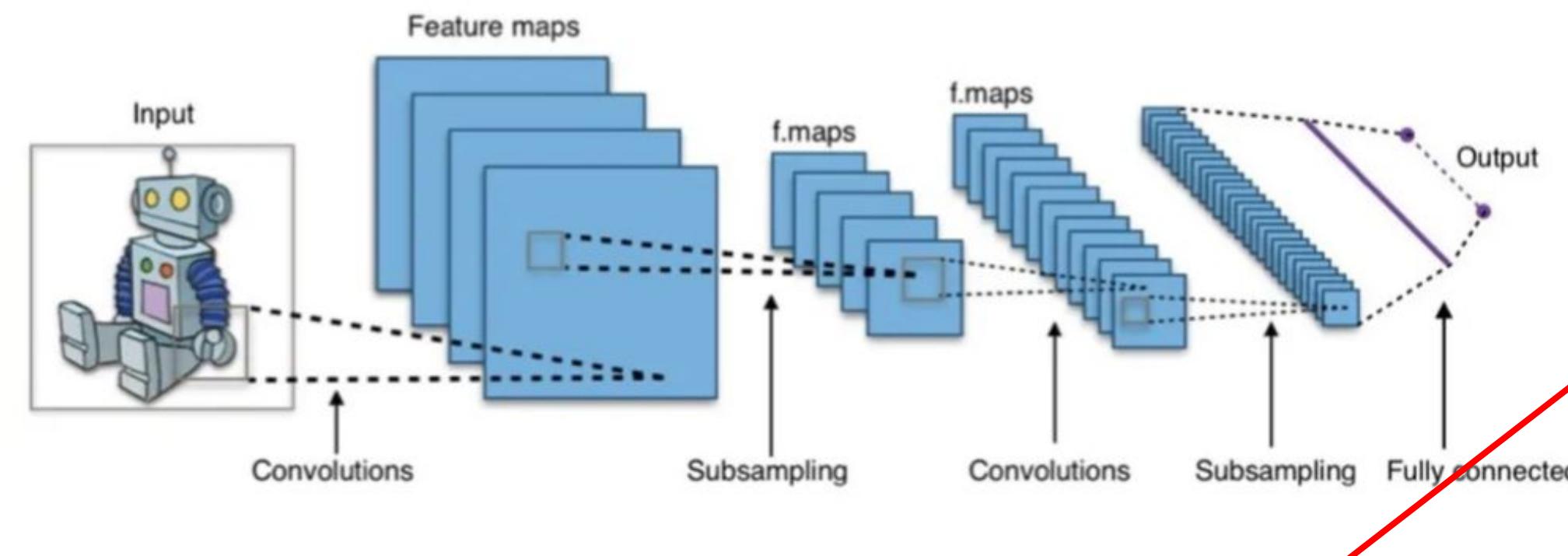
A dragon fruit wearing karate belt in the snow.



A strawberry mug filled with white sesame seeds. The mug is floating in a dark chocolate sea.

Back to 2015

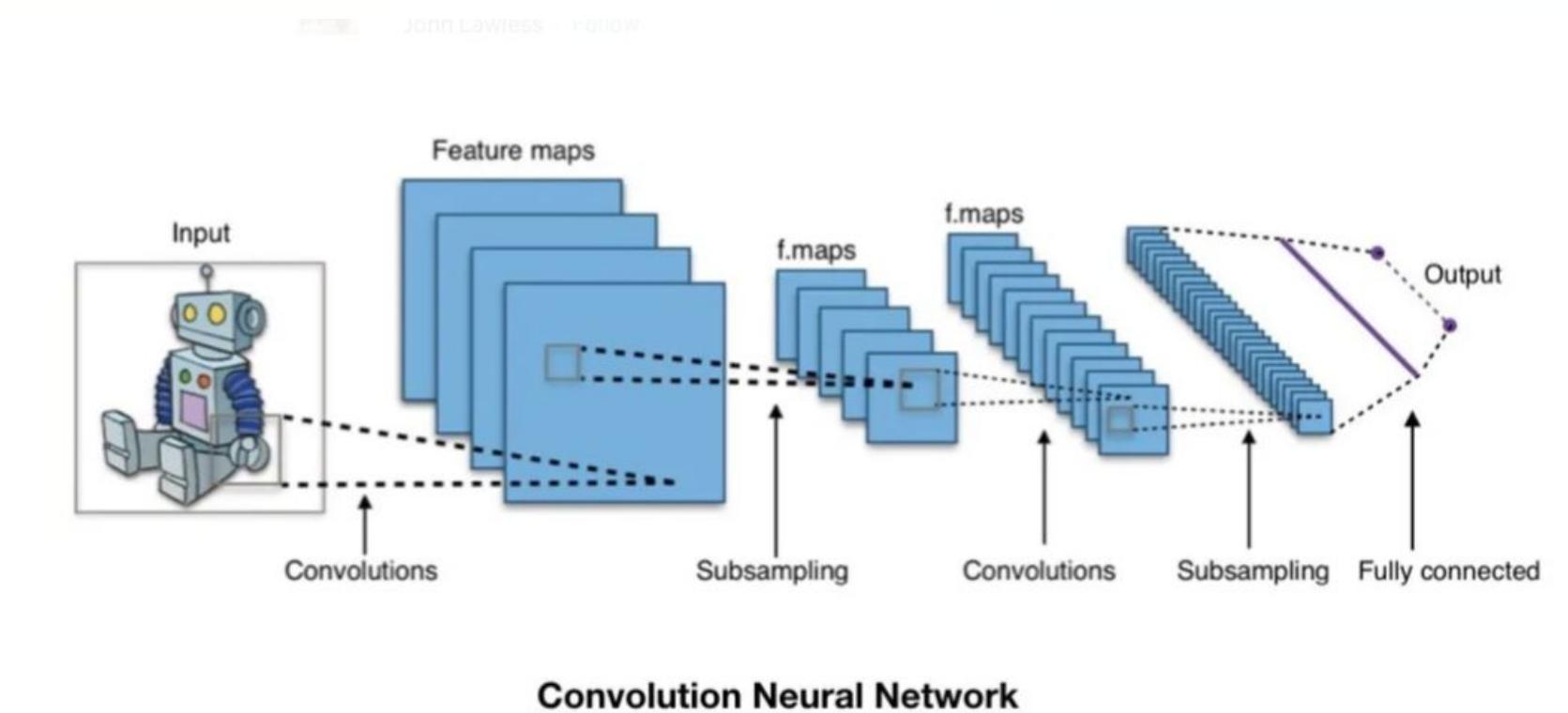
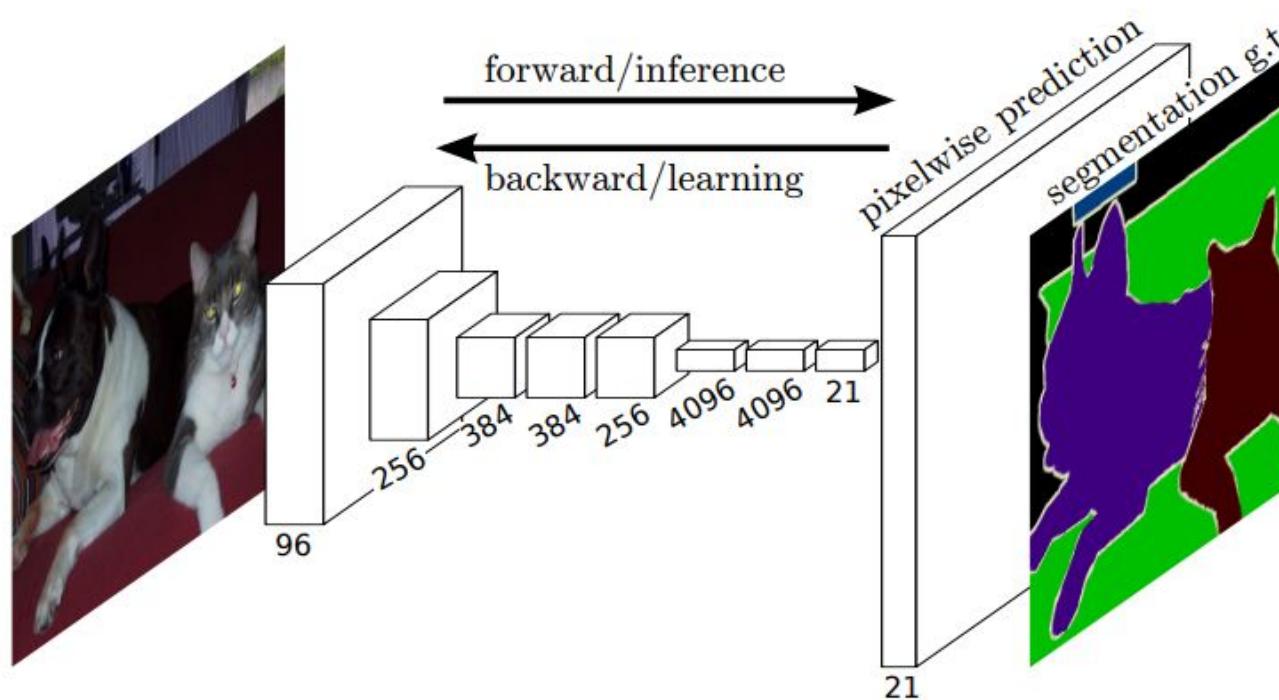
John Lawless / FOLLOW



Convolution Neural Network

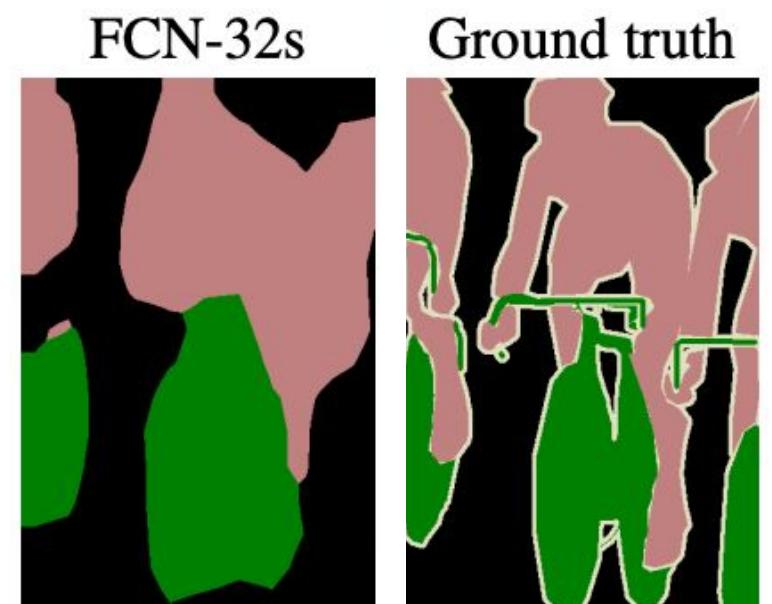
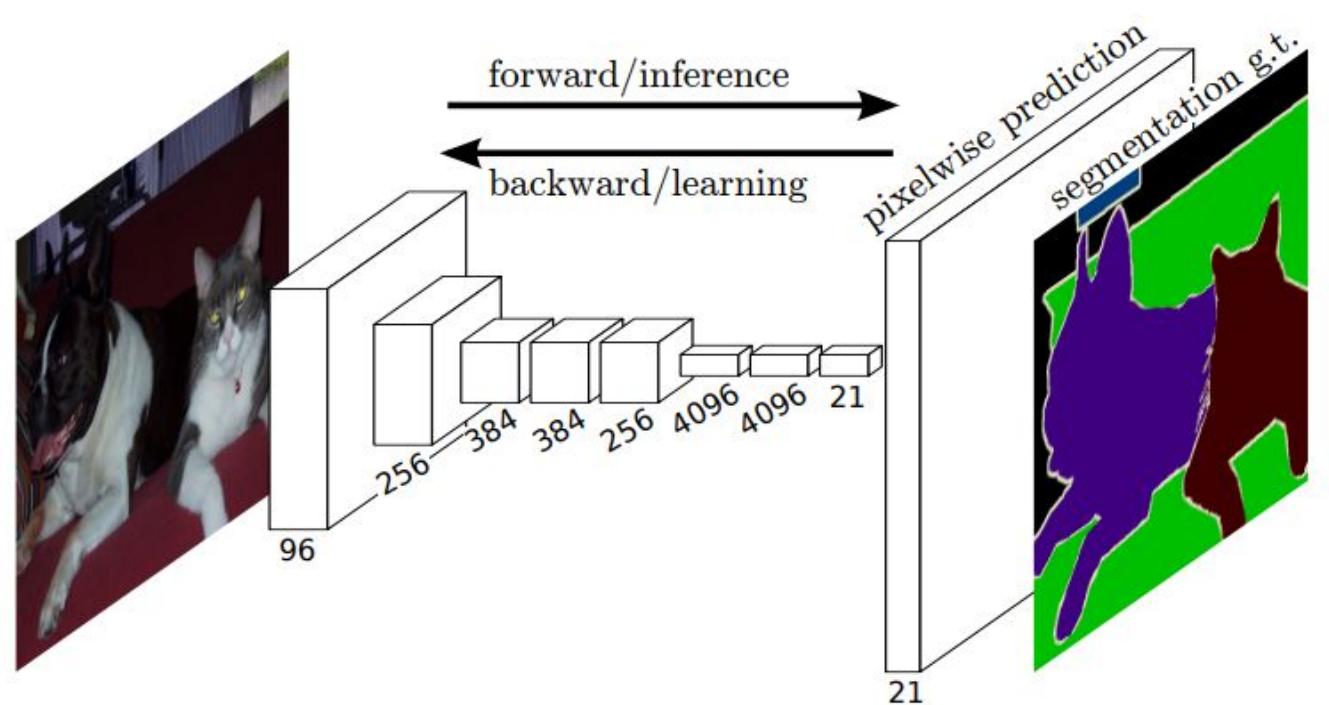
Source: Wikipedia

Image 2 Image? Fully Convolutional Models



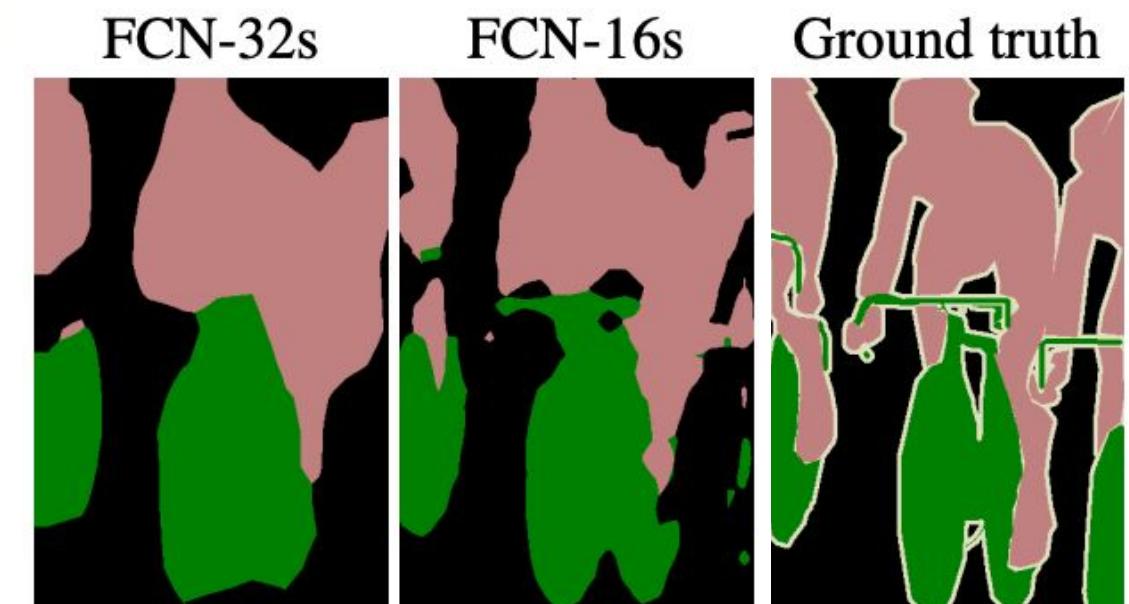
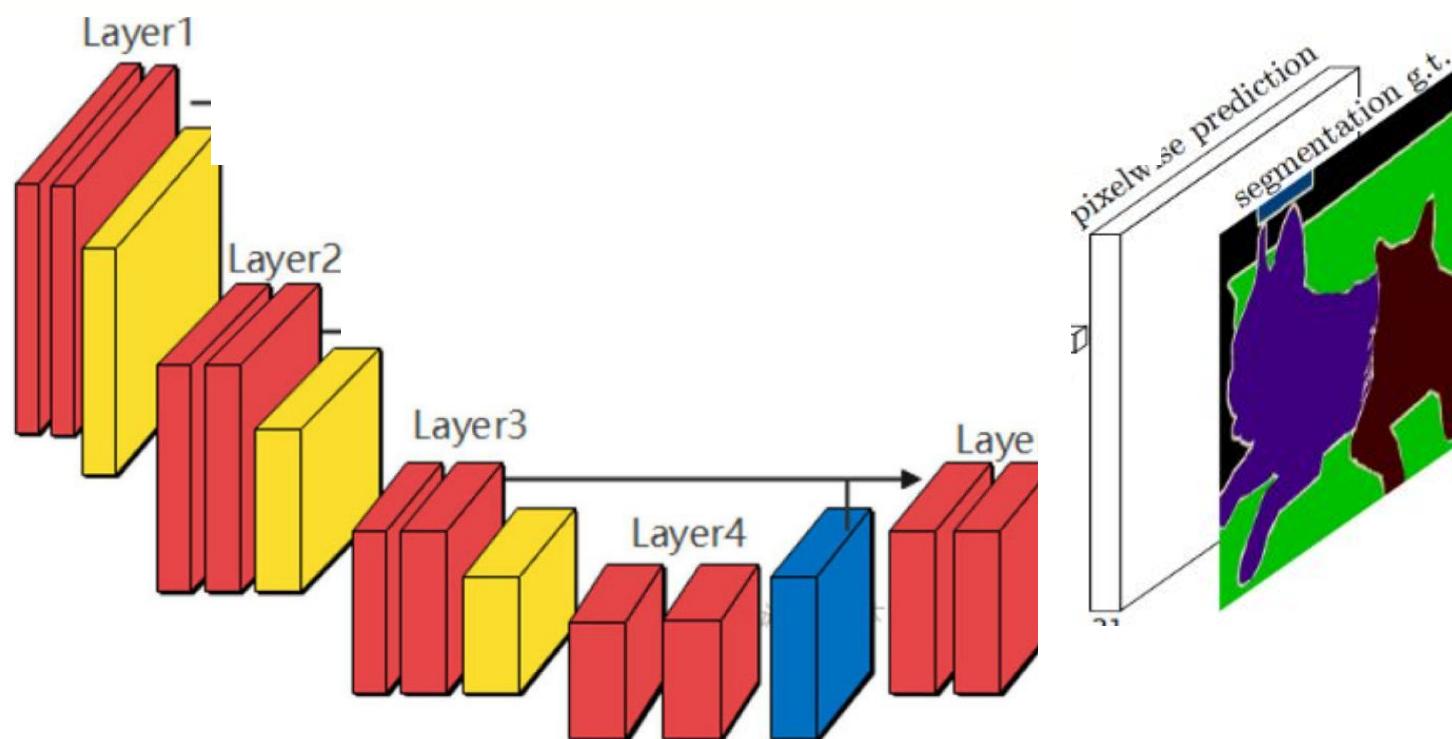
[Fully Convolutional Networks for Semantic Segmentation](#)

Back to 2015 – Fully Convolutional Models



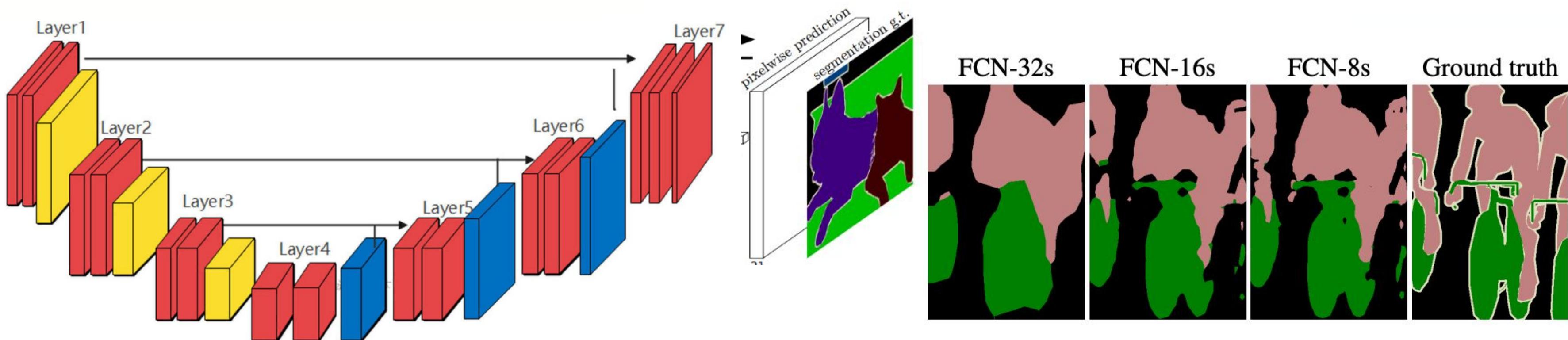
[Fully Convolutional Networks for Semantic Segmentation](#)

Back to 2015 – Fully Convolutional Models



[Fully Convolutional Networks for Semantic Segmentation](#)

Back to 2015 – Fully Convolutional Models



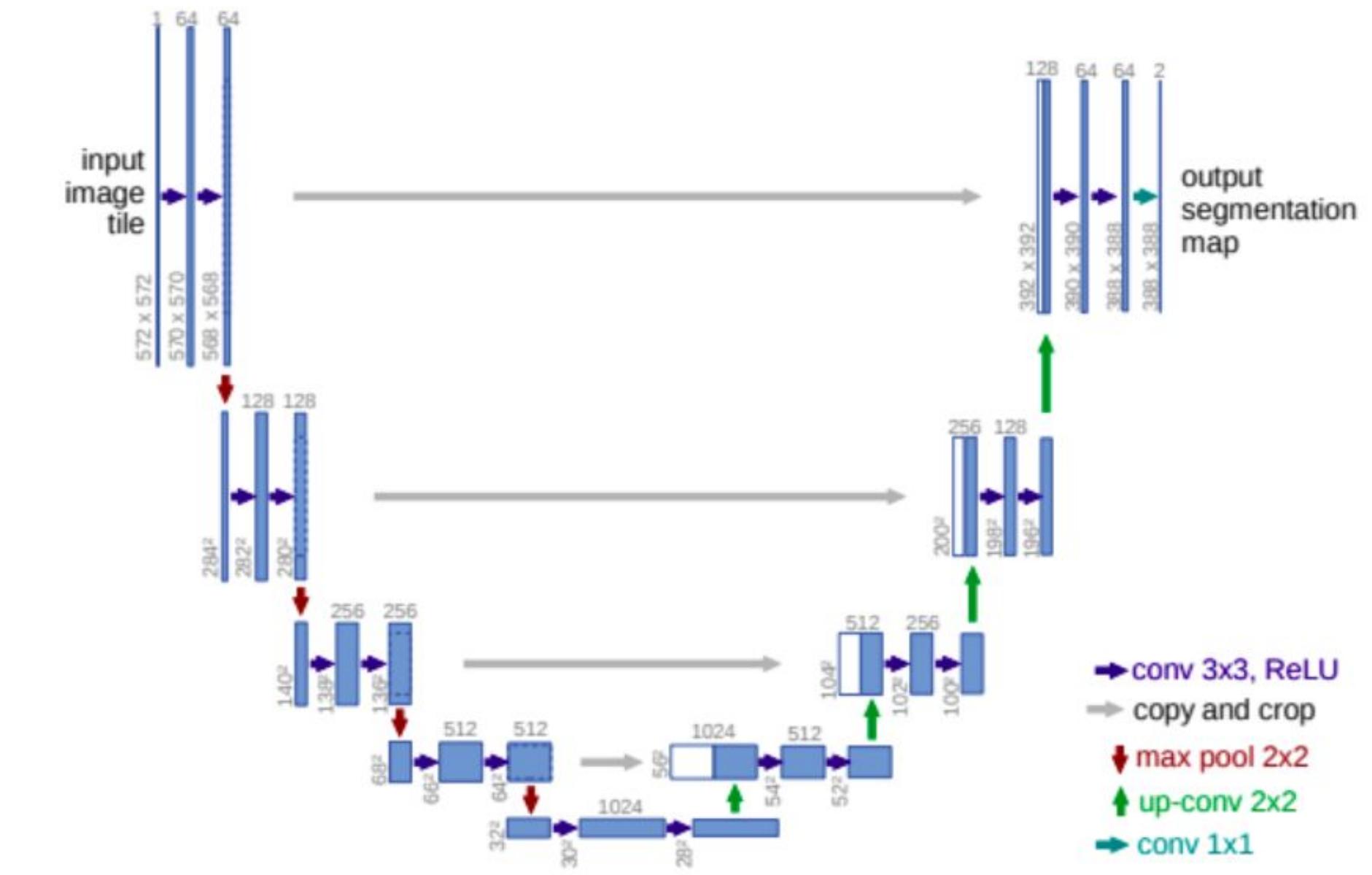
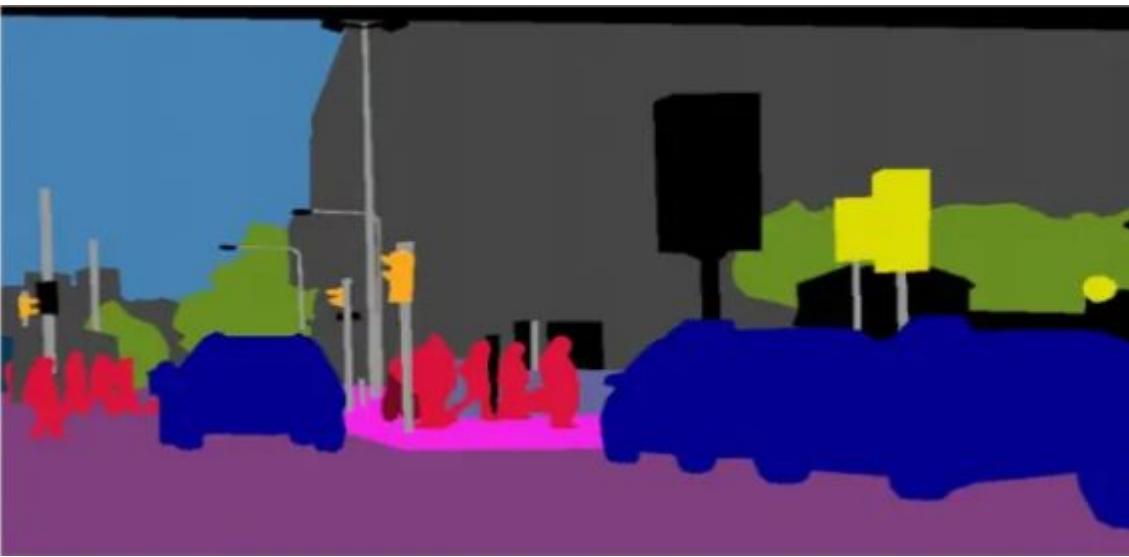
[Fully Convolutional Networks for Semantic Segmentation](#)

Unet Architecture

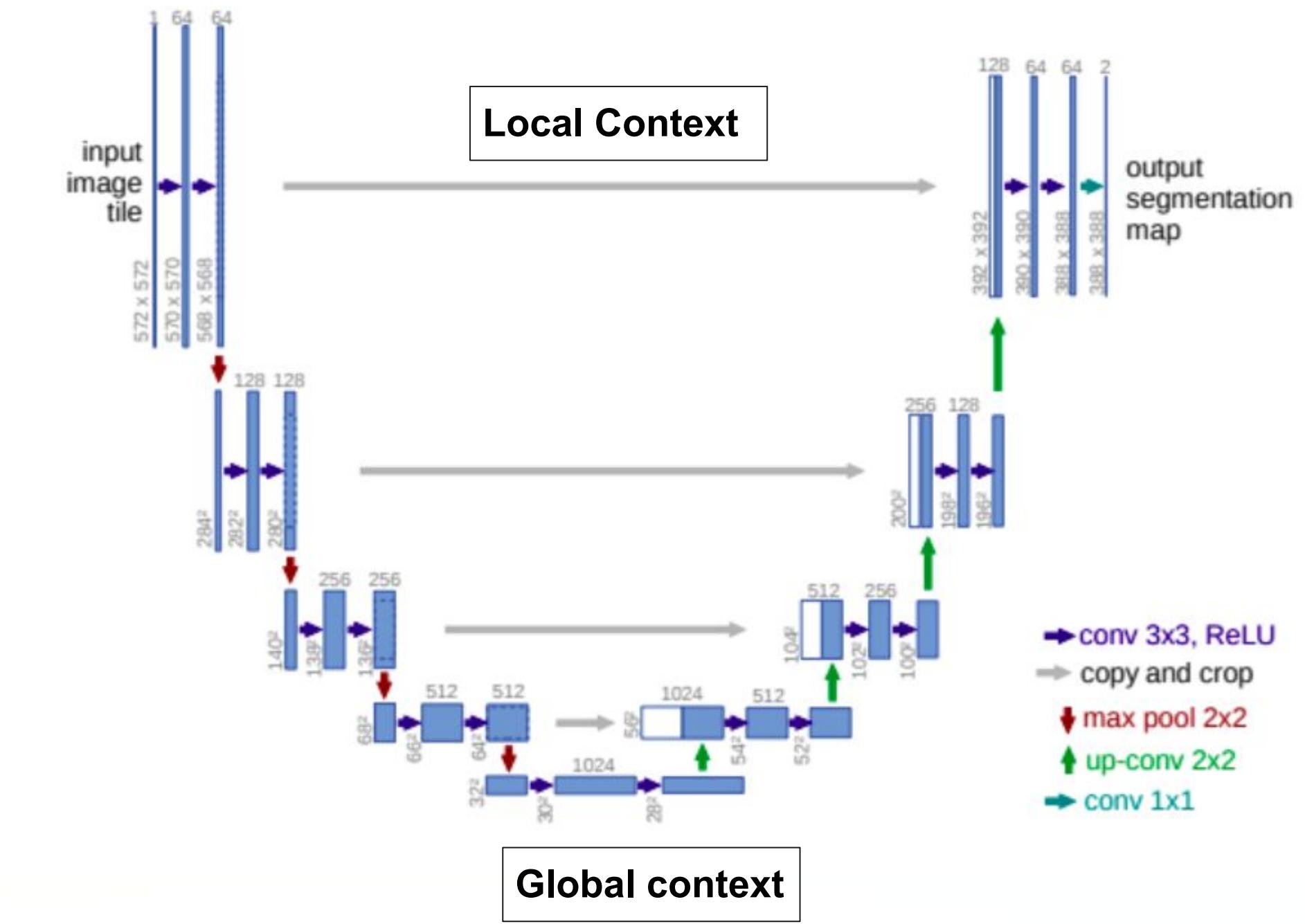
Input:



Output:



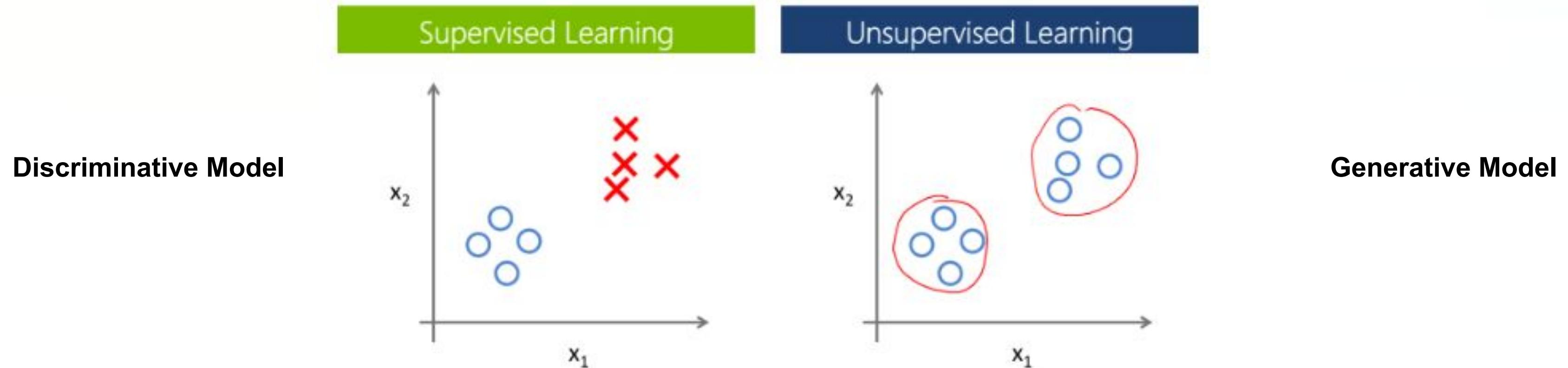
Unet Architecture



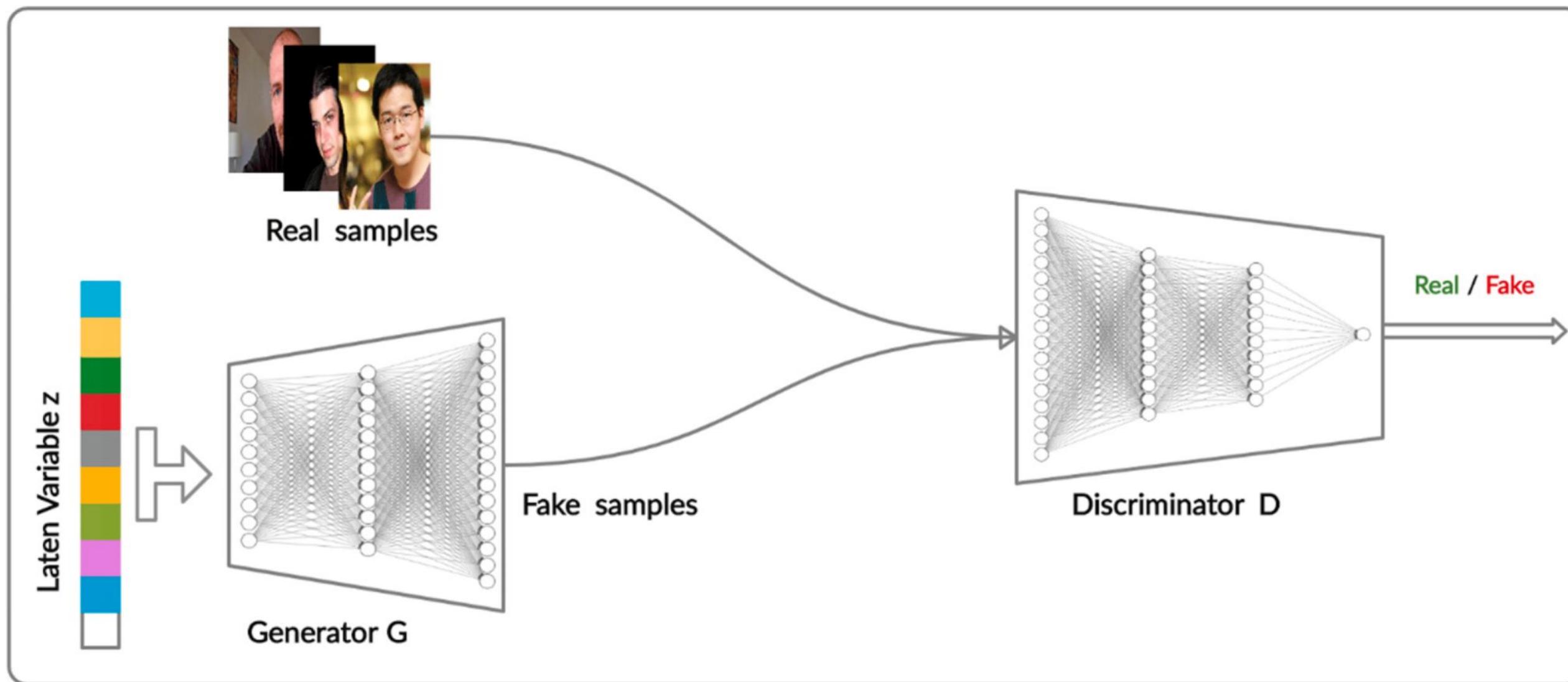
Generative Adversarial Network



Supervised/ Unsupervised Learning



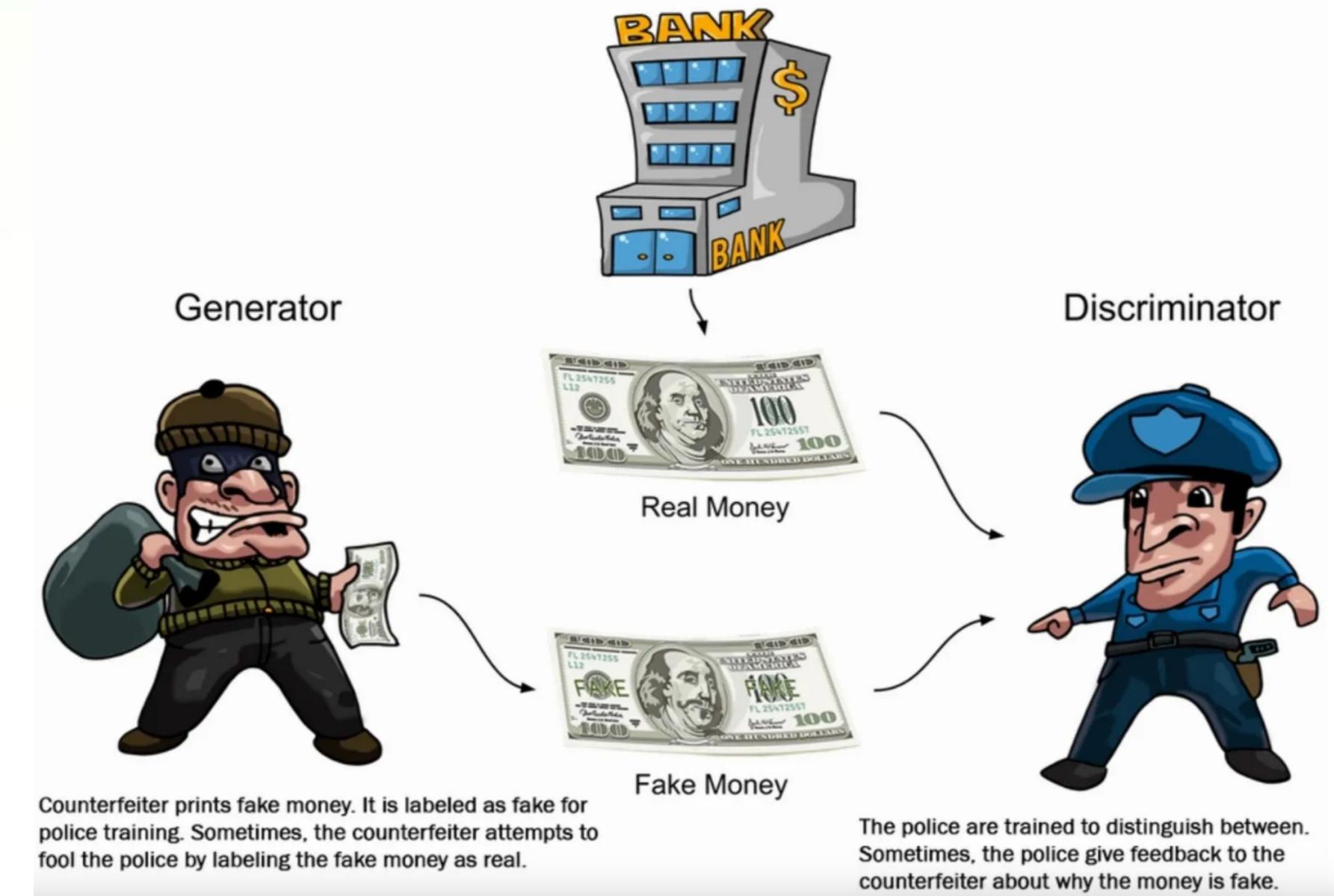
Generative Adversarial Network (GAN)



[Generative Adversarial Networks](#), Ian Goodfellow, Yoshua Bengio

Input: random noise vector (dimensions - 1xd)
Output: Image (dimensions wxhx3)

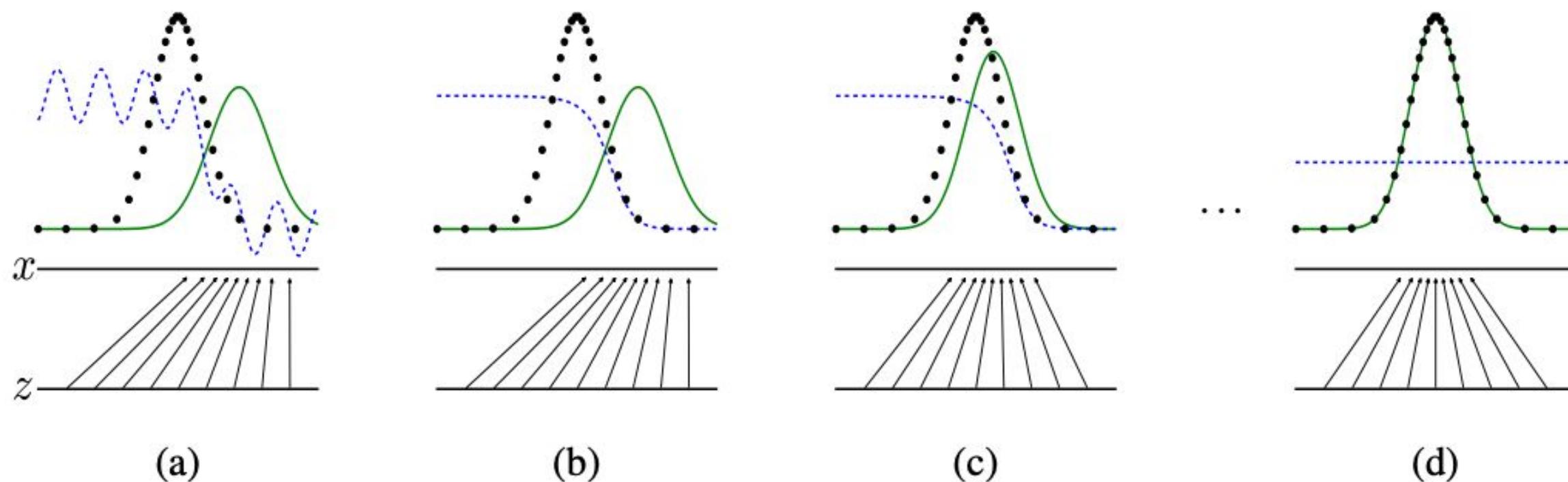
Generative Adversarial Network (GAN)



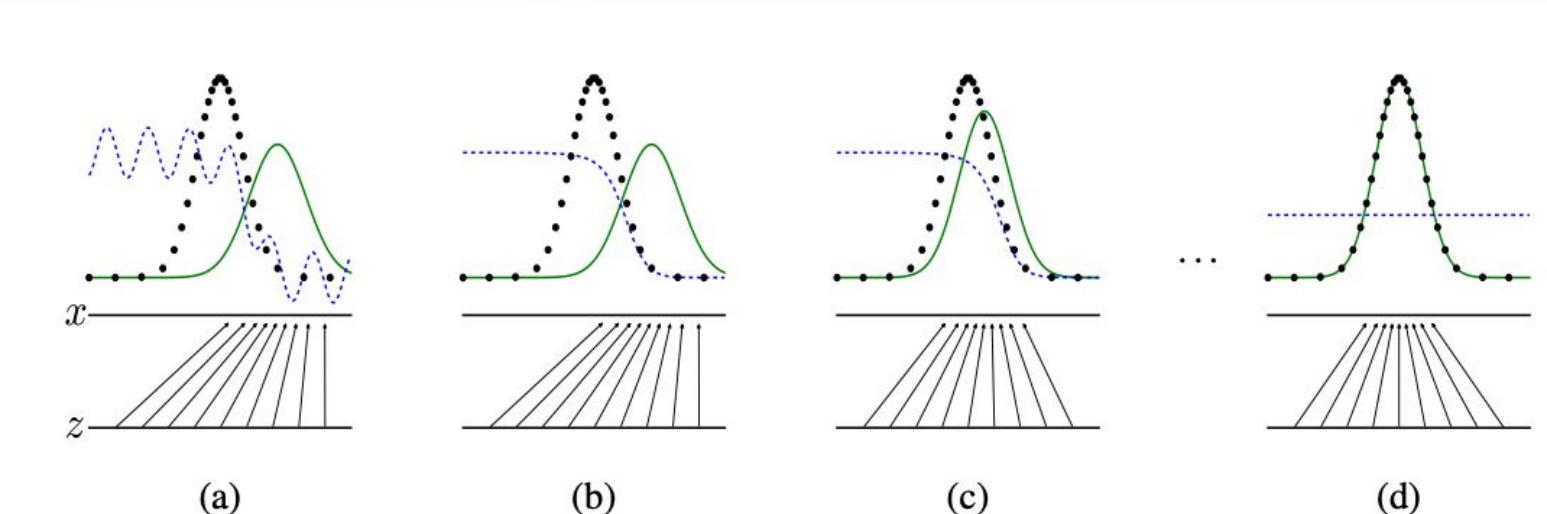
[Generative Adversarial Networks](#), Ian Goodfellow, Yoshua Bengio

Generative Adversarial Network (GAN)

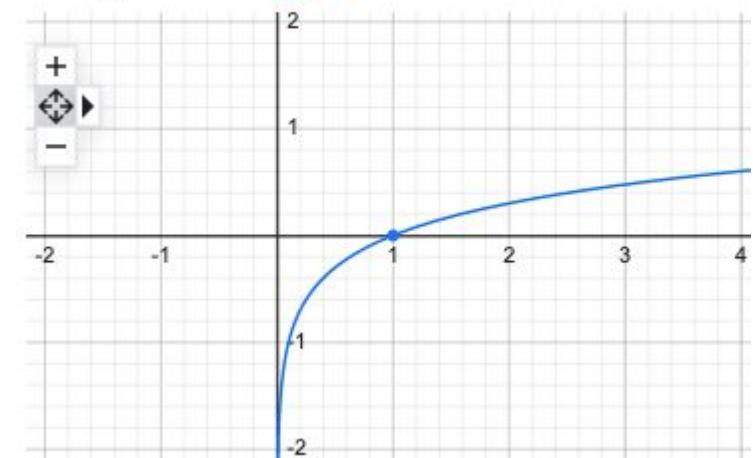
Black - Original distribution
Green - Generated distribution
Blue - Discriminator



The Objective:



Graph for $\log(x)$



Discriminator Maximizes:

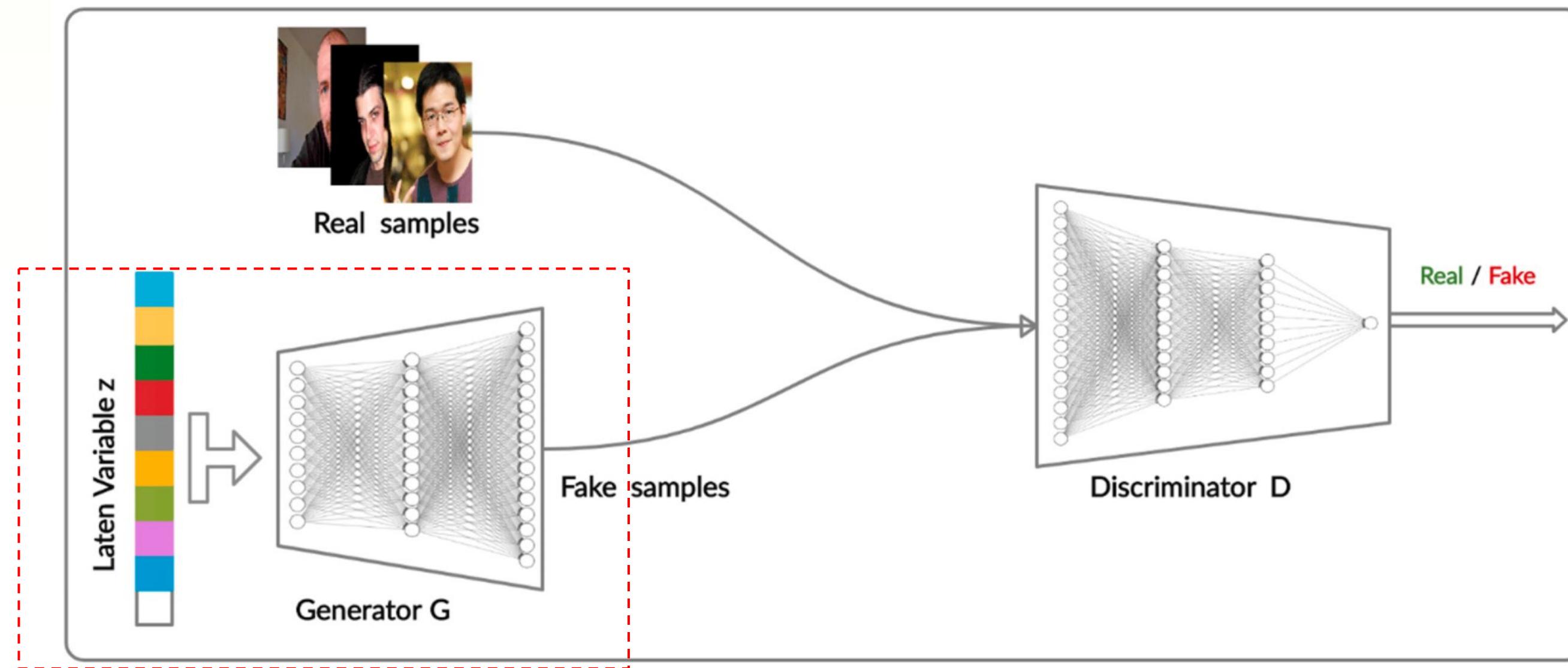
$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^m \left[\log D(\mathbf{x}^{(i)}) + \log (1 - D(G(\mathbf{z}^{(i)}))) \right]$$

Generator Minimizes:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^m \log \left(1 - D \left(G \left(\mathbf{z}^{(i)} \right) \right) \right)$$

Generative Adversarial Networks, Ian Goodfellow, Yoshua Bengio

Generative Adversarial Network (GAN)



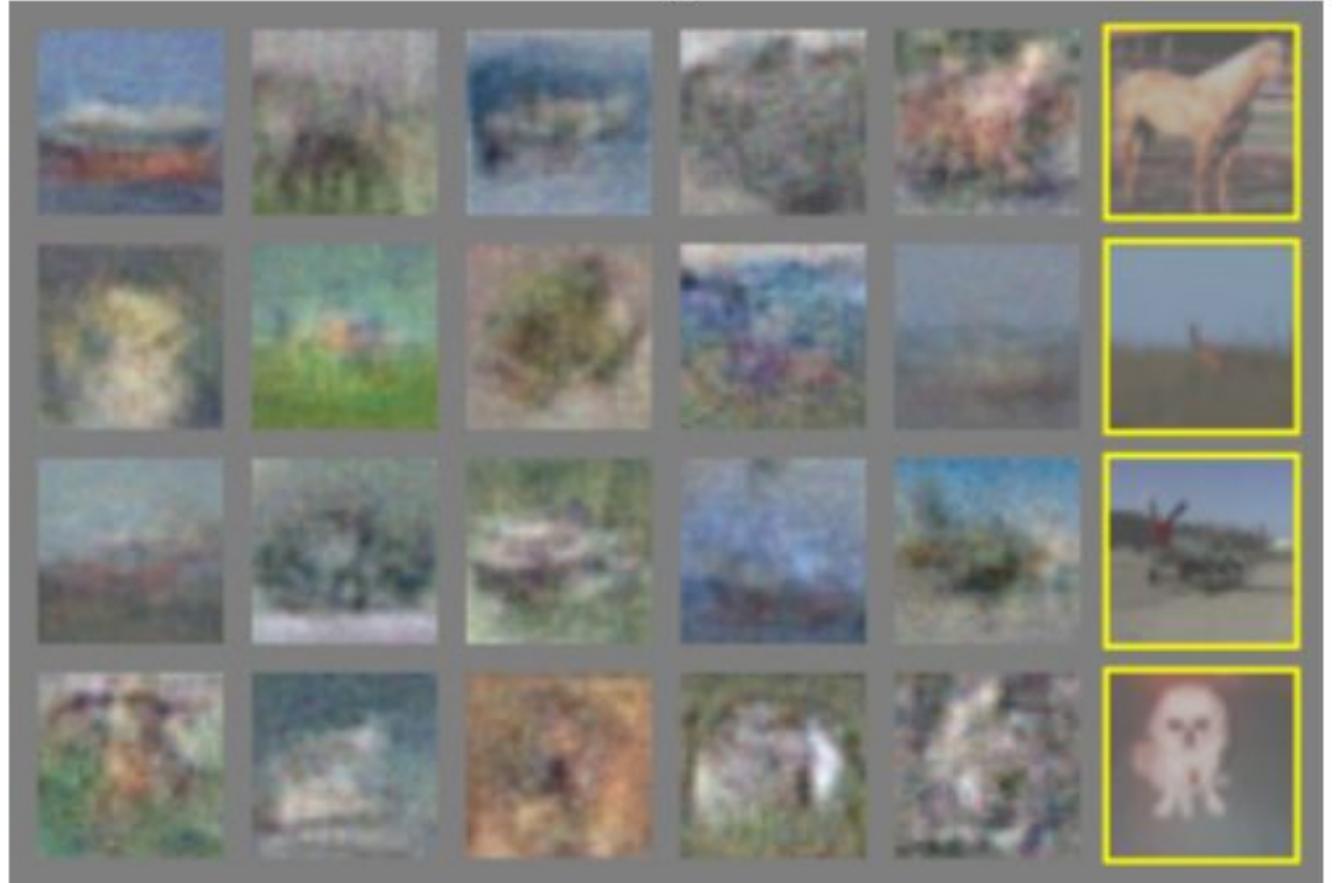
[Generative Adversarial Networks](#)



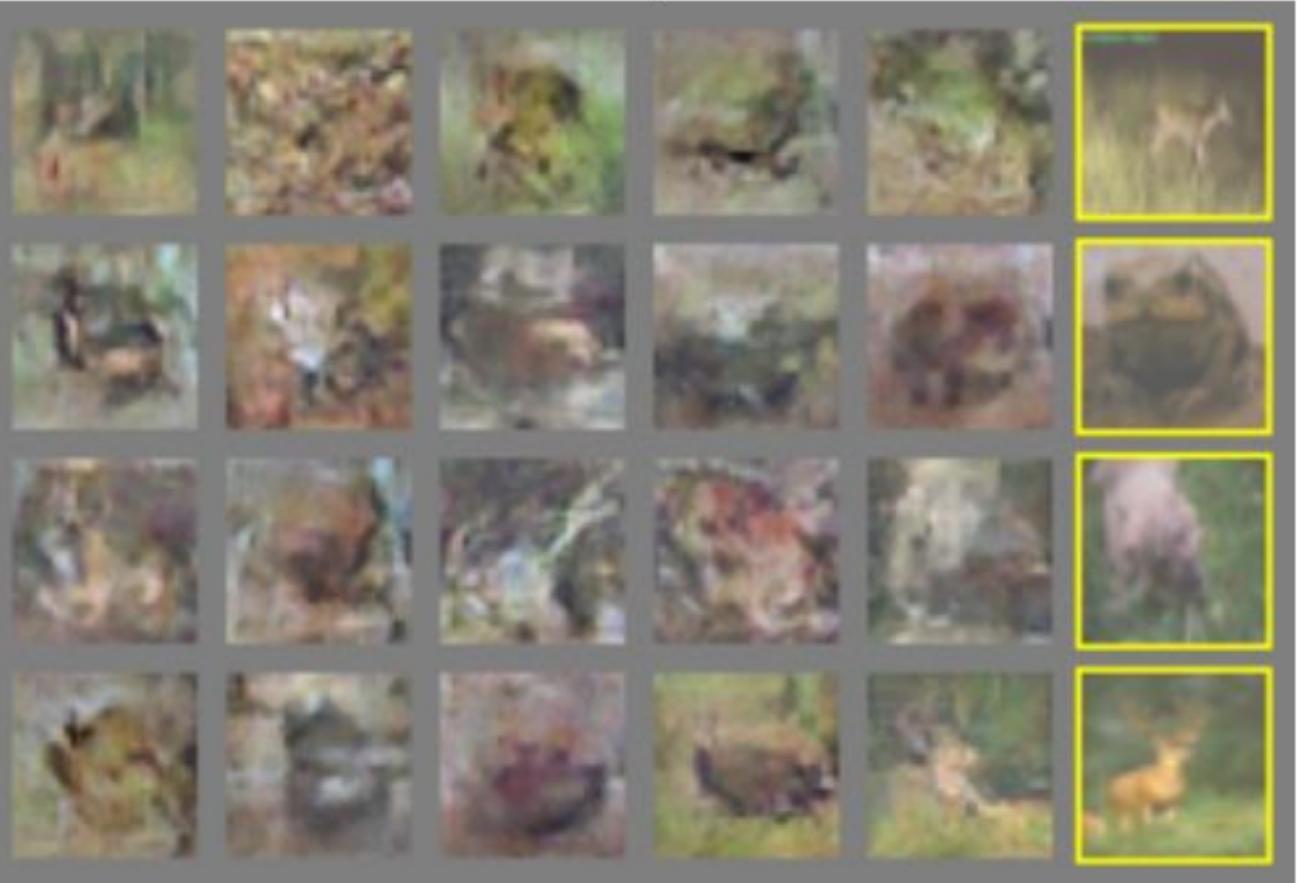
a)



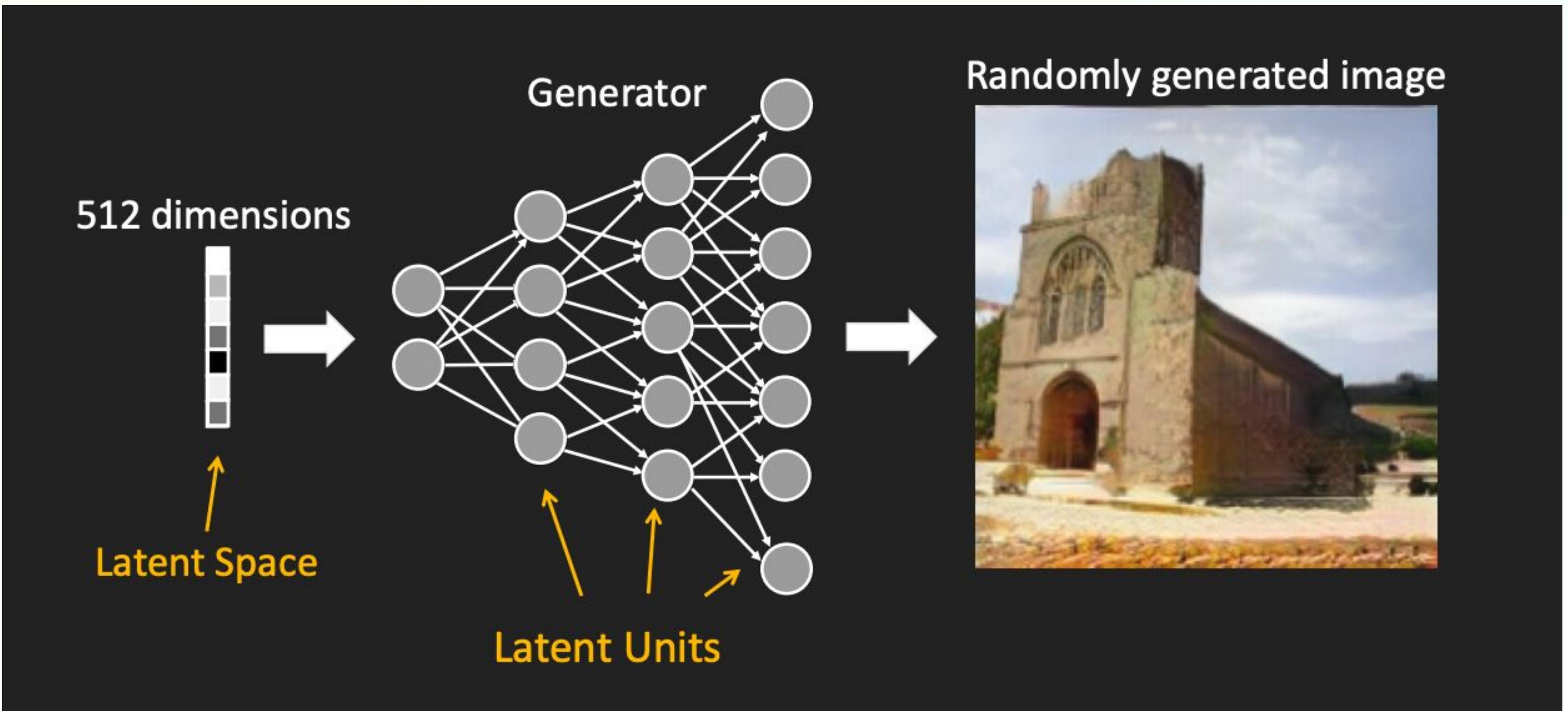
b)



c)



d)



*Input and output

GANs Resolution Progress



2014



2015



2016



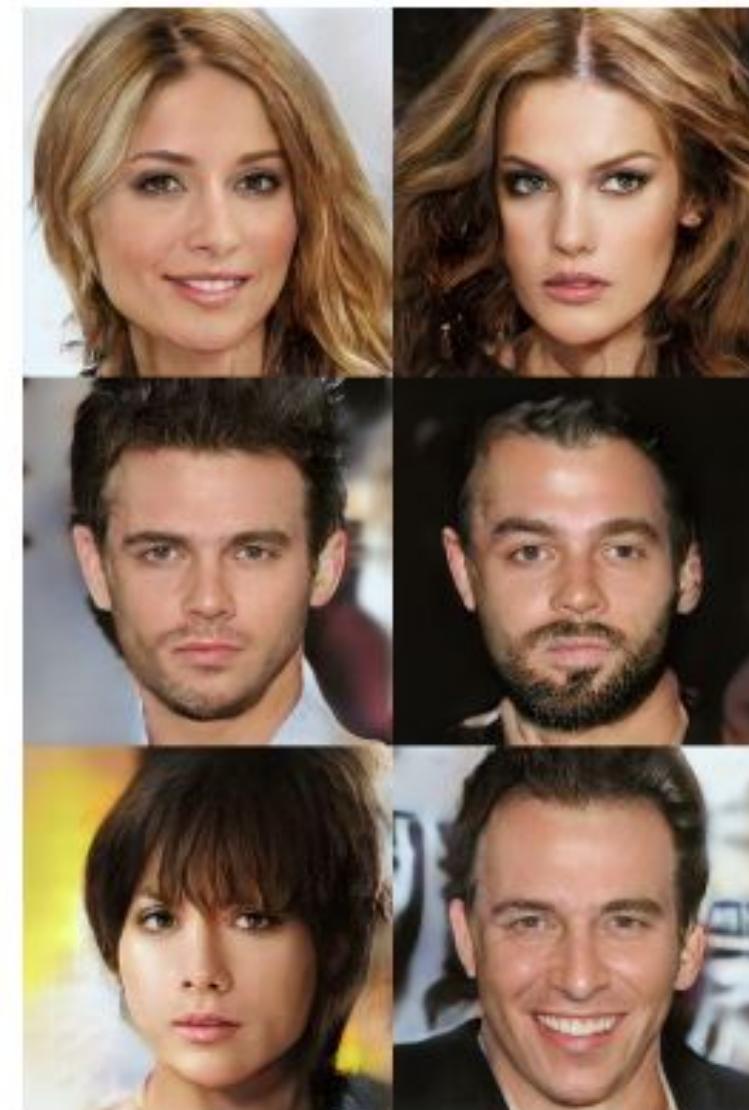
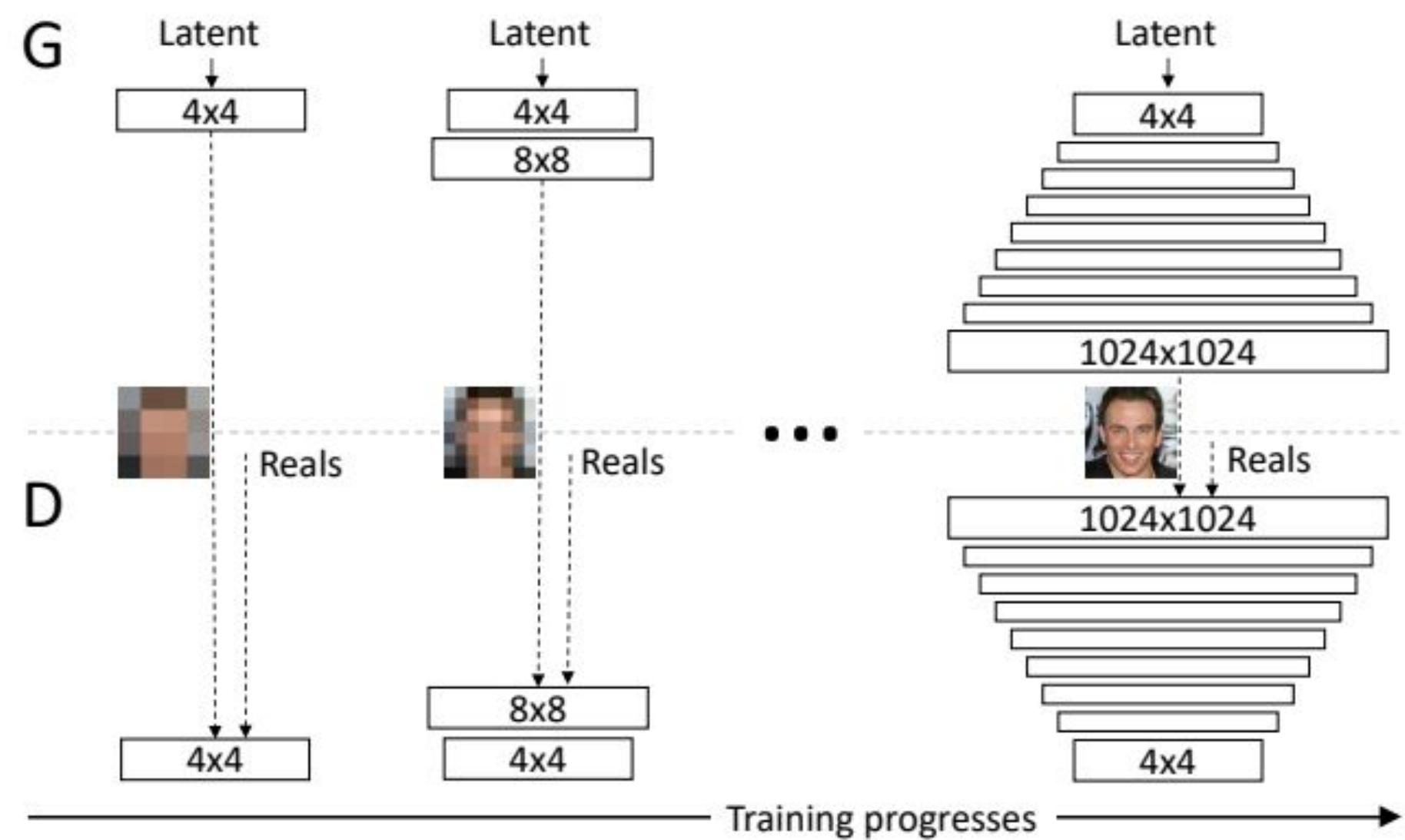
2017



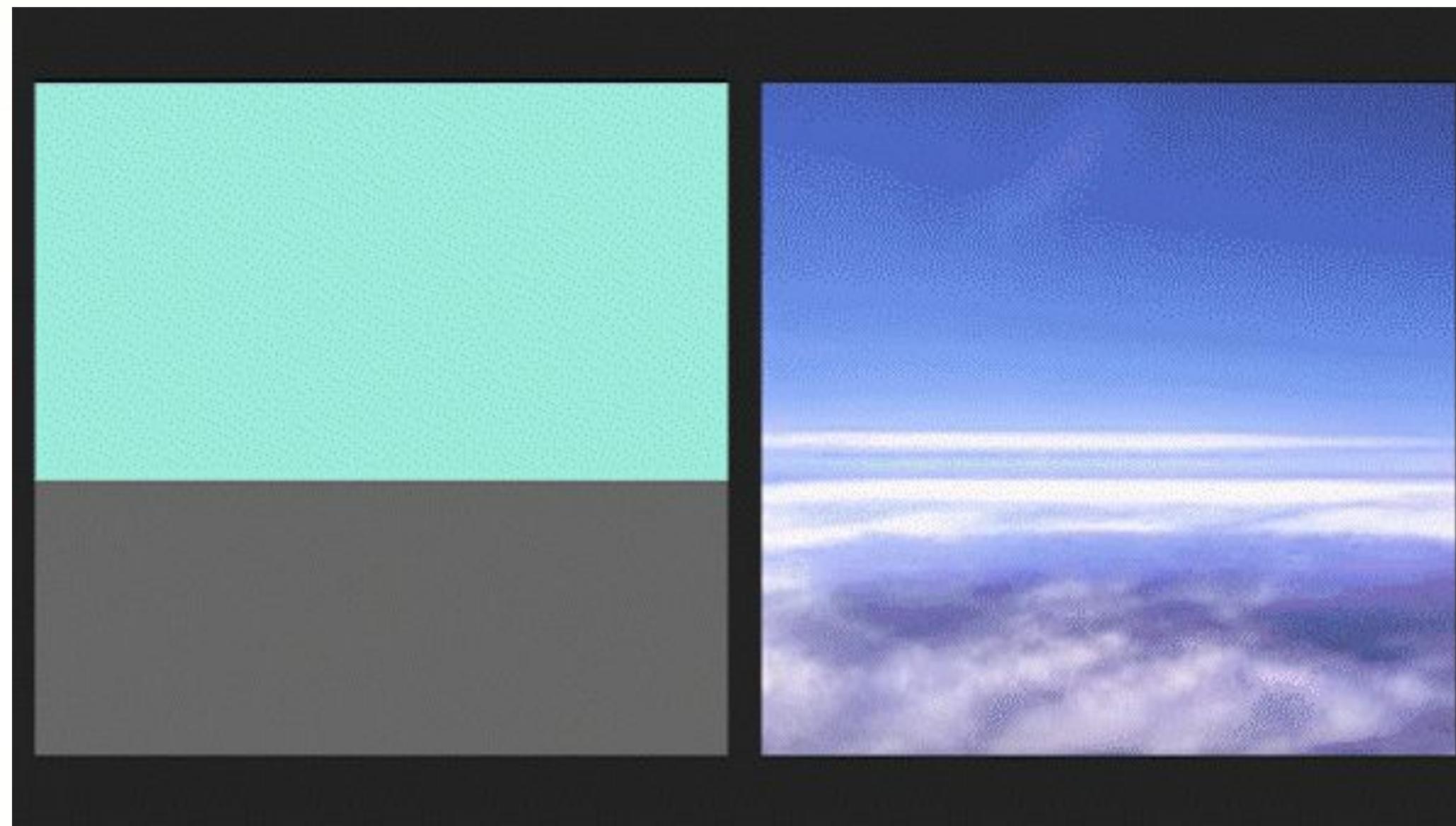
2018



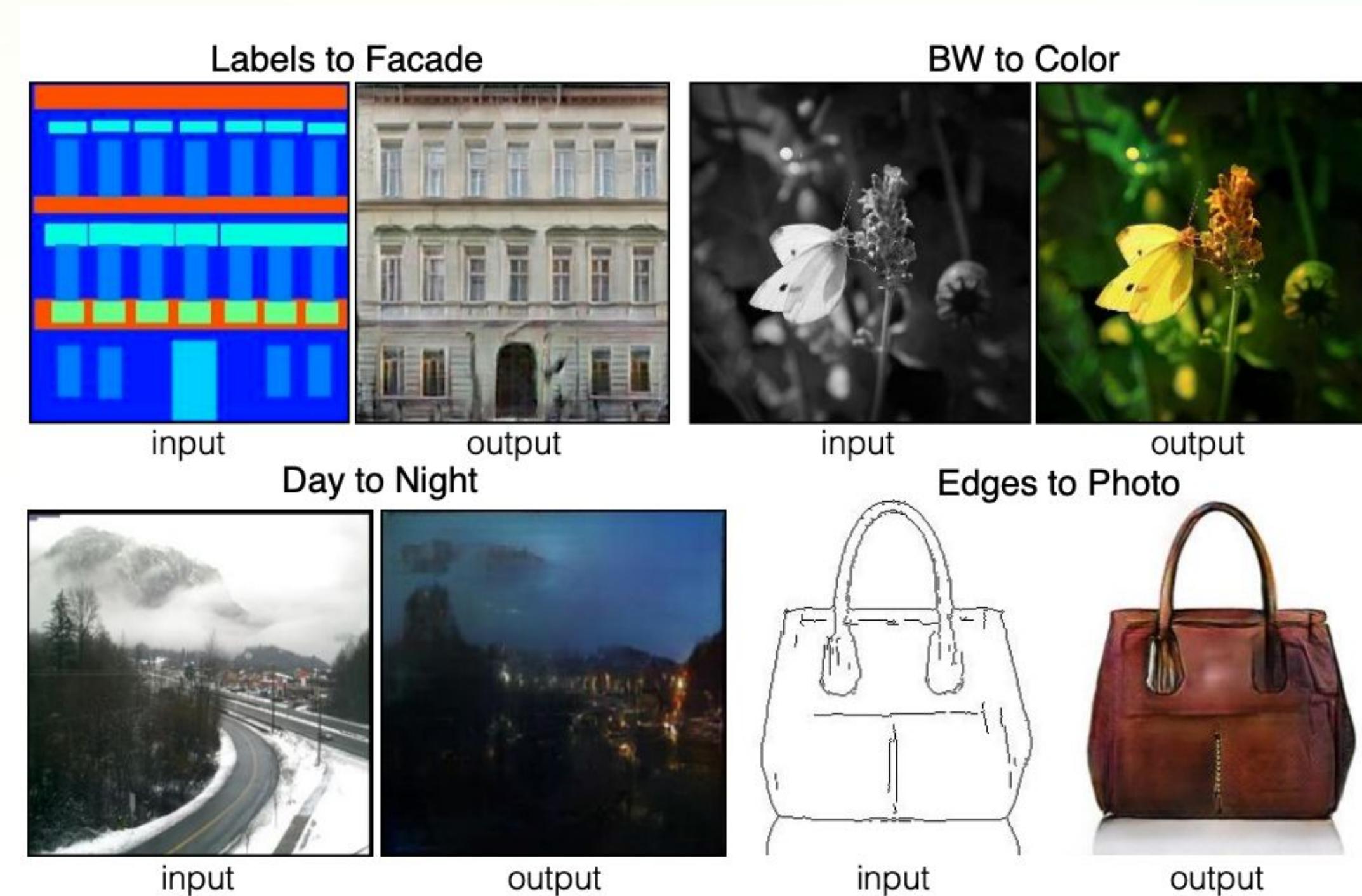
Progressive Growing



Now Lets Try to Add Control



Pix2Pix



Why not Train Image 2 Image Unet?



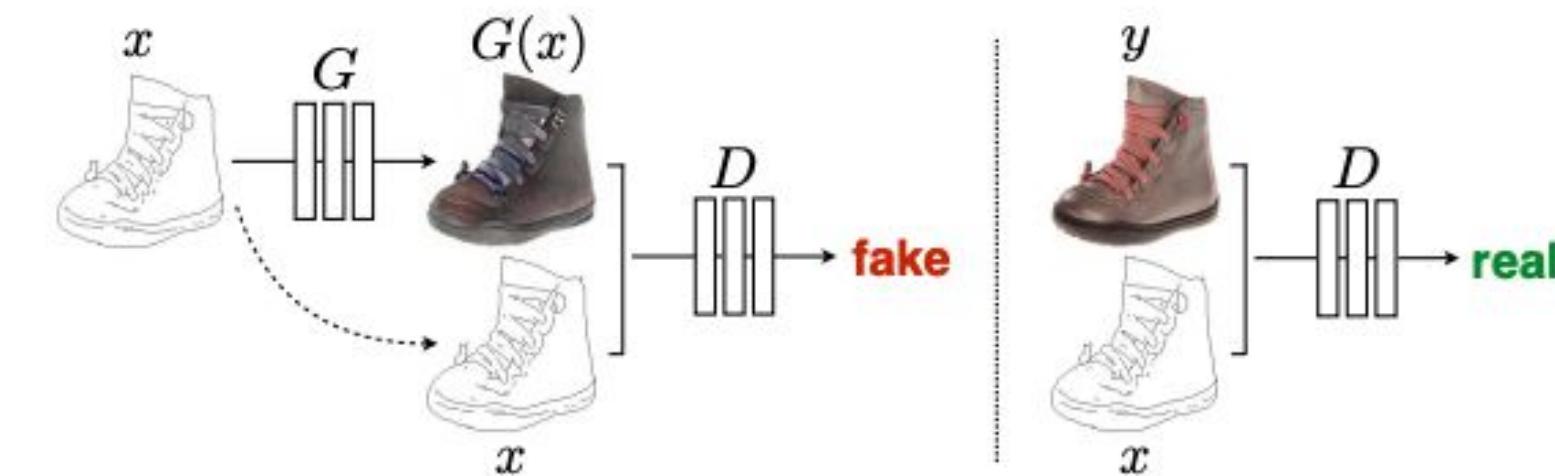
Pix2Pix

The objective: Generator minimizes
Discriminator maximizes

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z)))] ,$$

$$G^* = \arg \min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G)$$

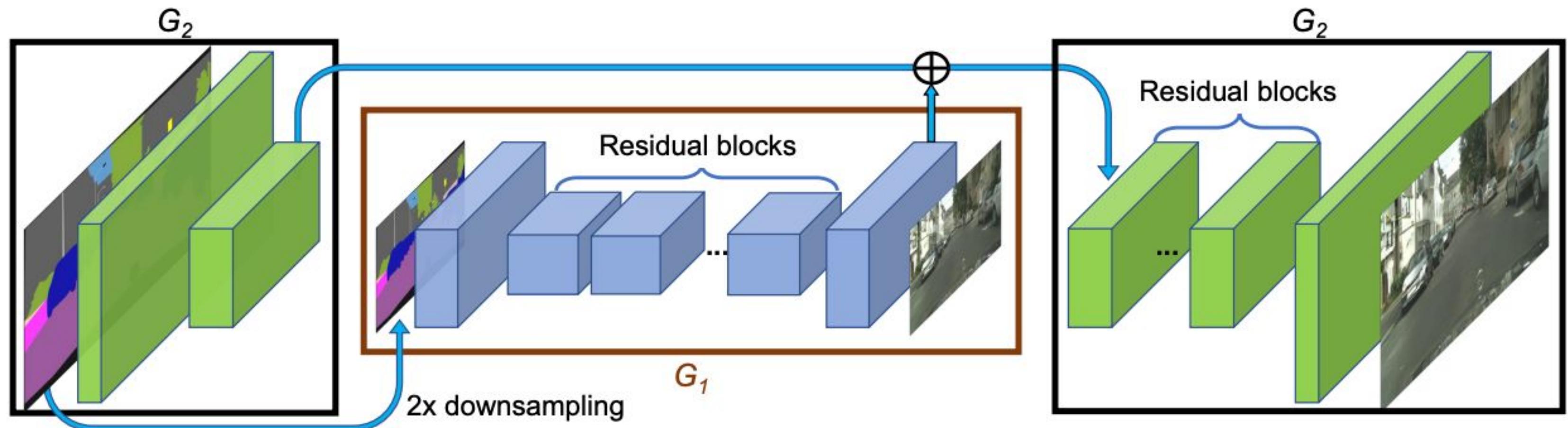
Architecture



PatchGan



Pix2PixHD - Can we improve the resolution?



GauGAN - Spade

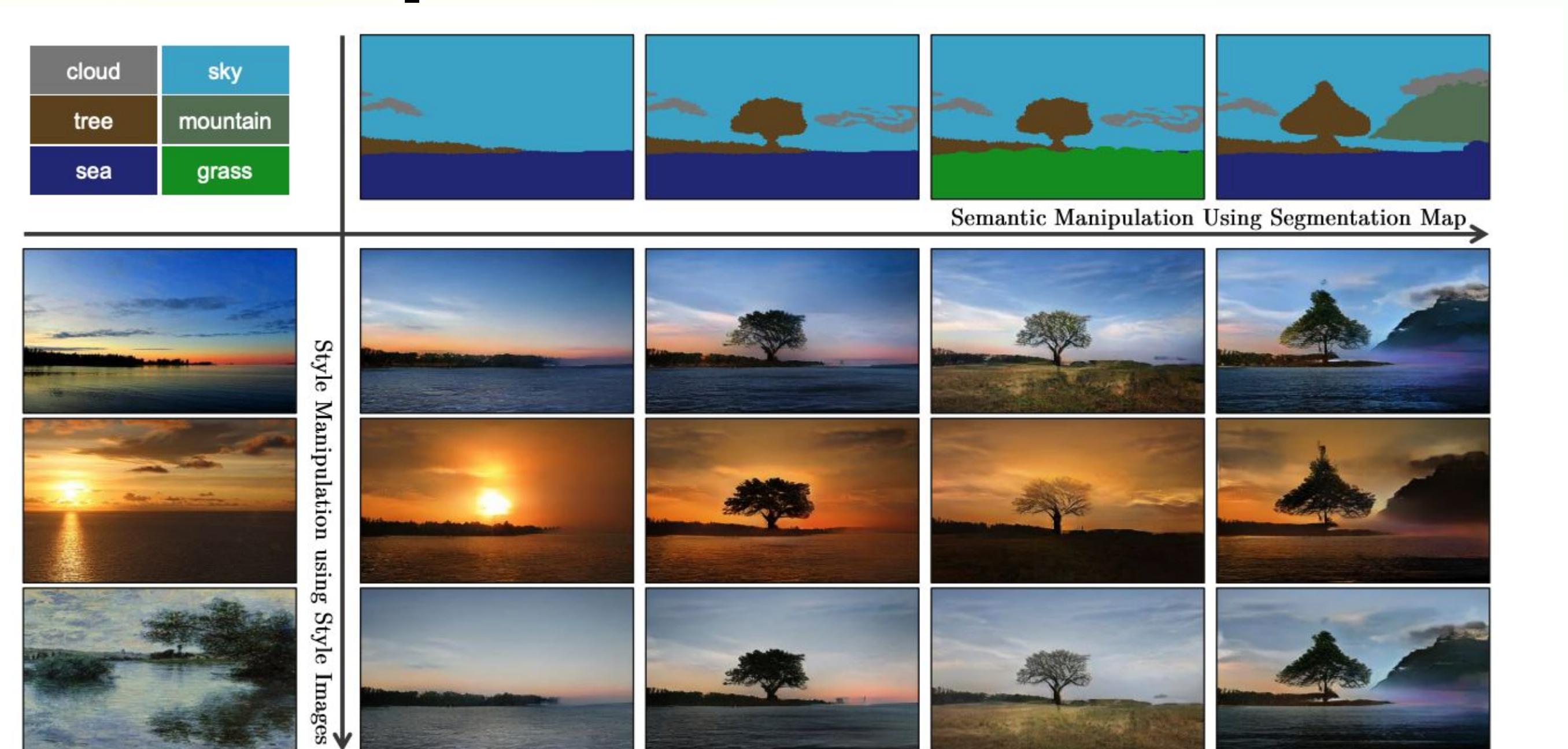
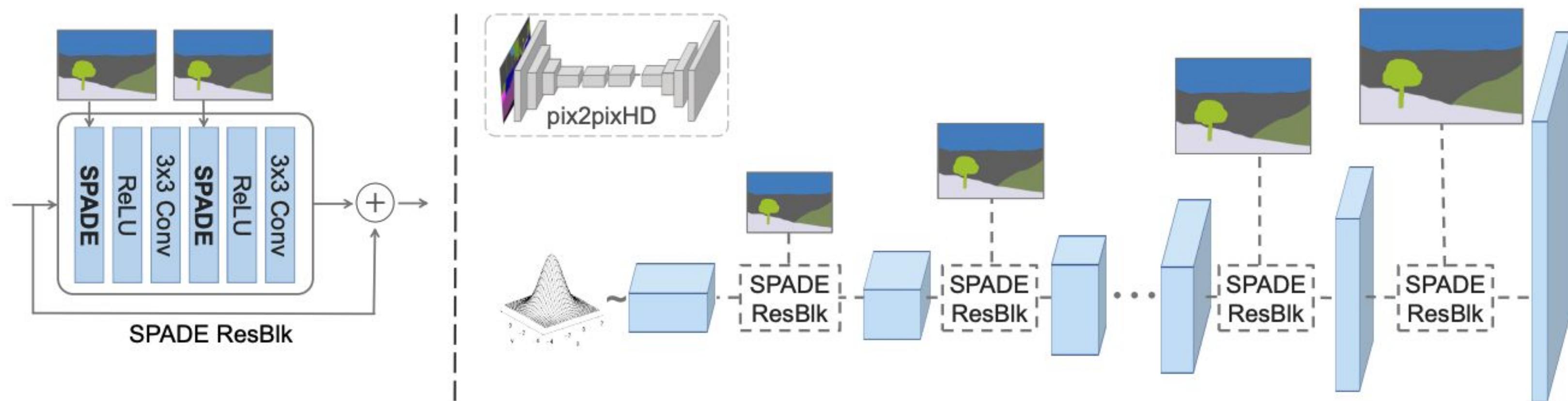


Figure 1: Our model allows user control over both semantic and style as synthesizing an image. The semantic (e.g., the existence of a tree) is controlled via a label map (the top row), while the style is controlled via the reference style image (the leftmost column). Please visit our [website](#) for interactive image synthesis demos.

GauGAN - Spade



GauGAN - Spade

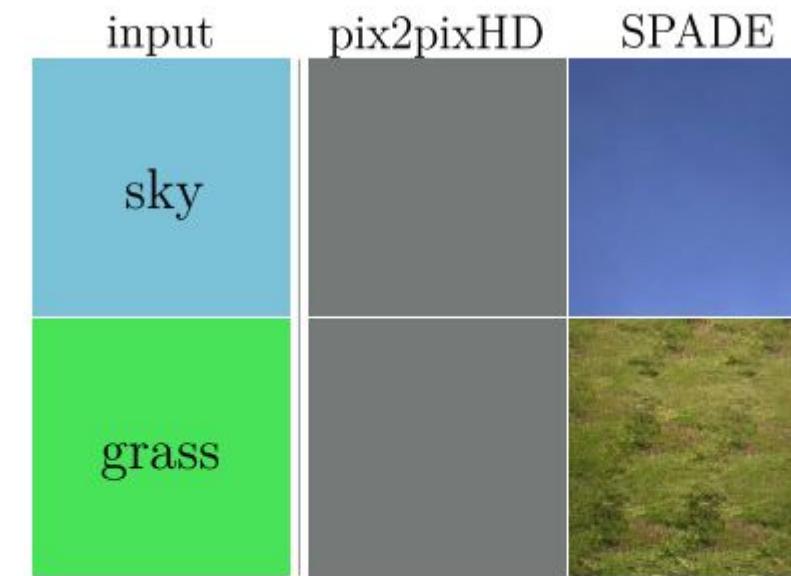
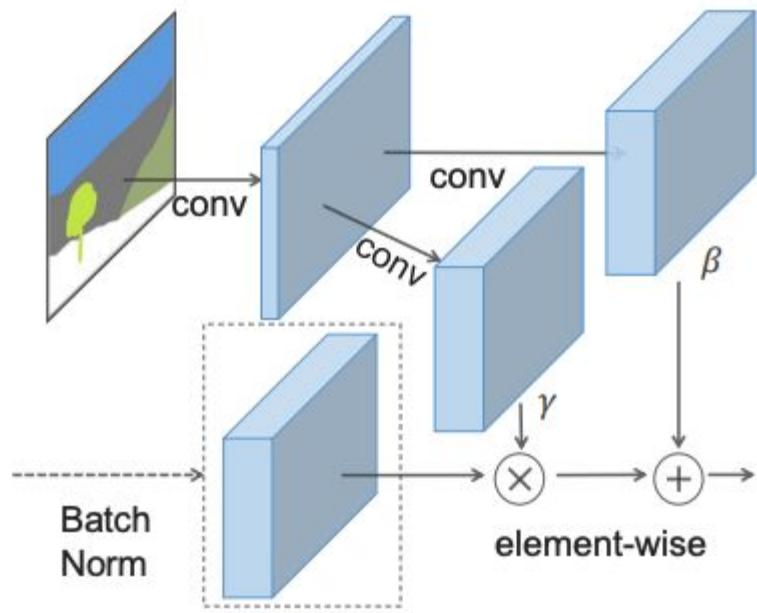
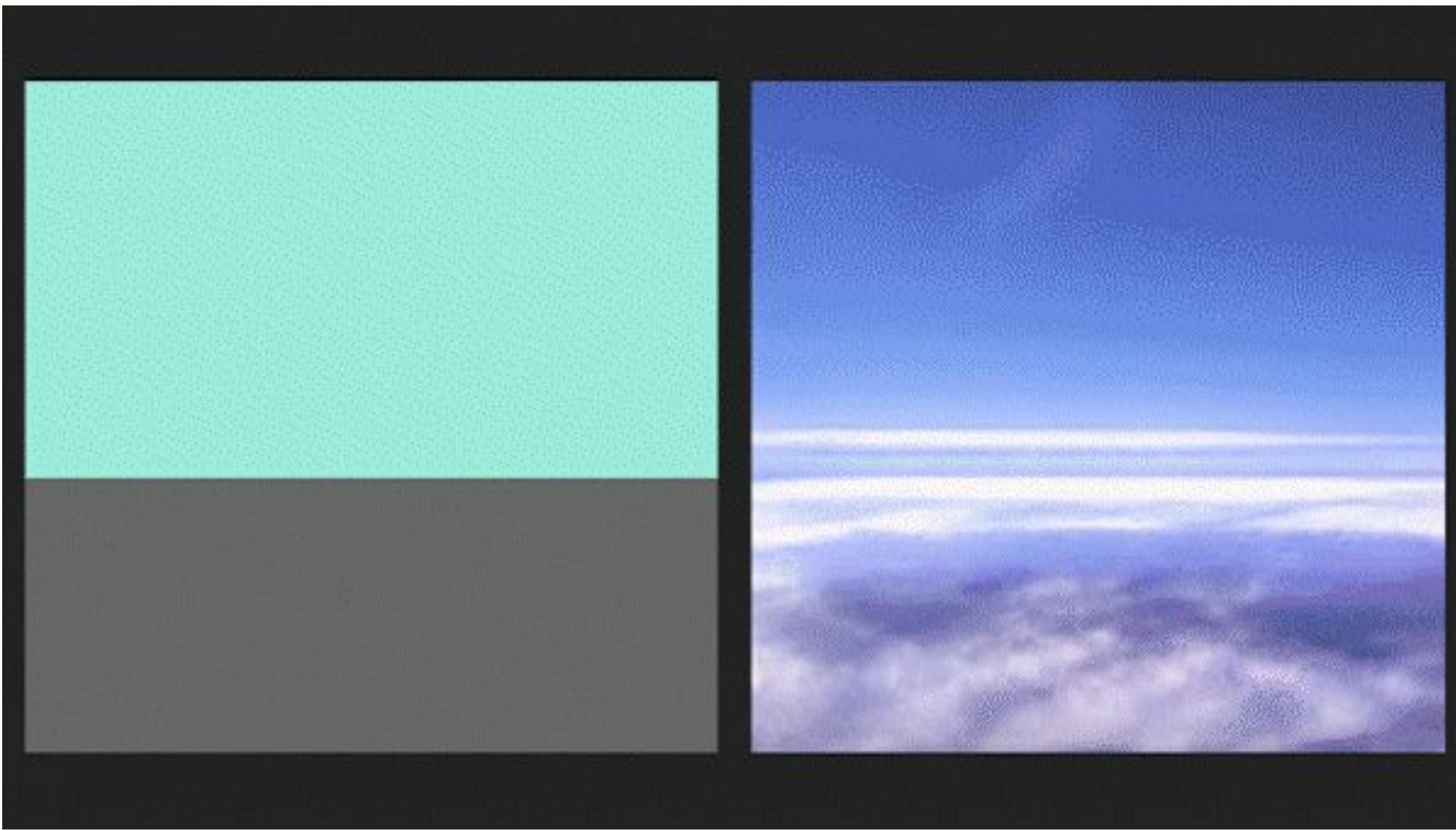
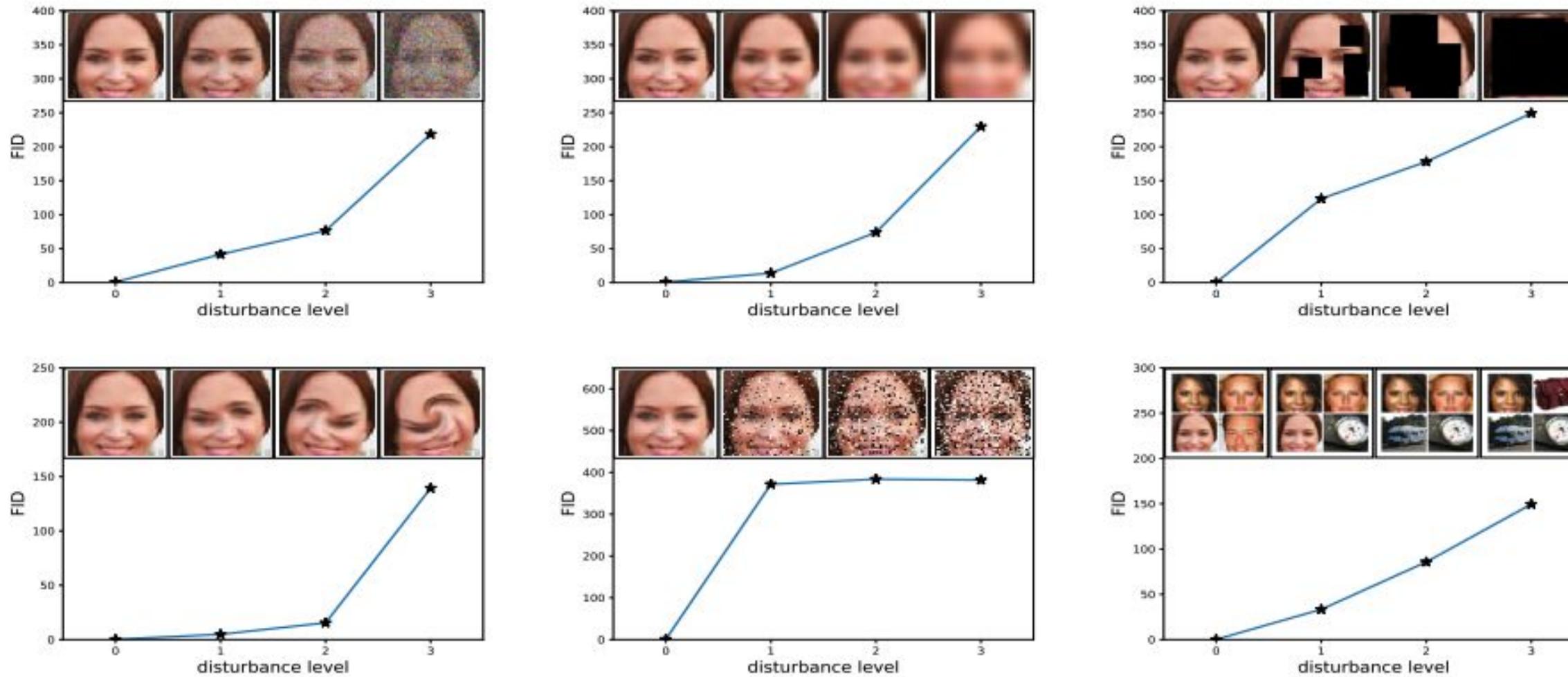


Figure 3: Comparing results given uniform segmentation maps: while the SPADE generator produces plausible textures, the pix2pixHD generator [48] produces two identical outputs due to the loss of the semantic information after the normalization layer.

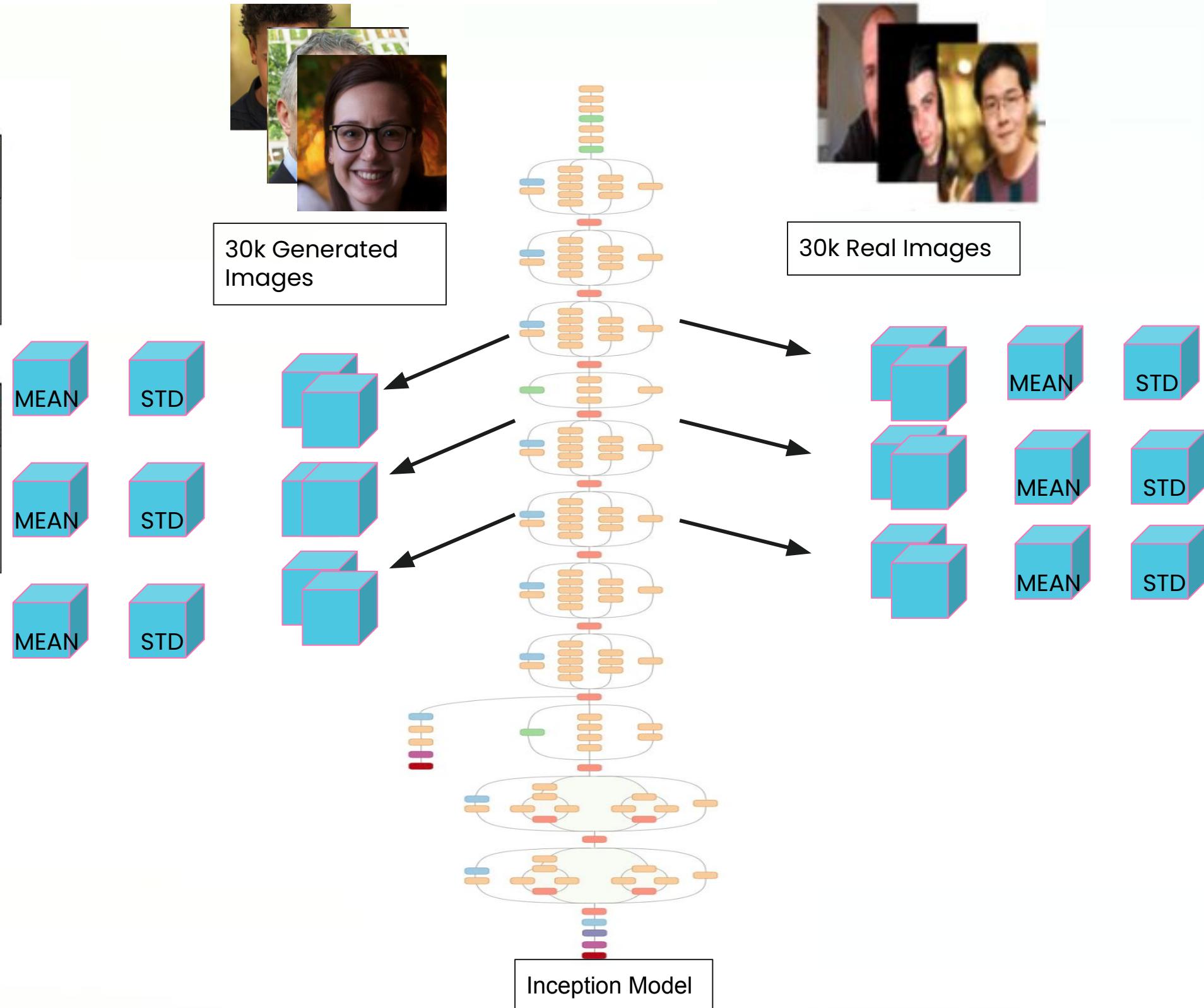
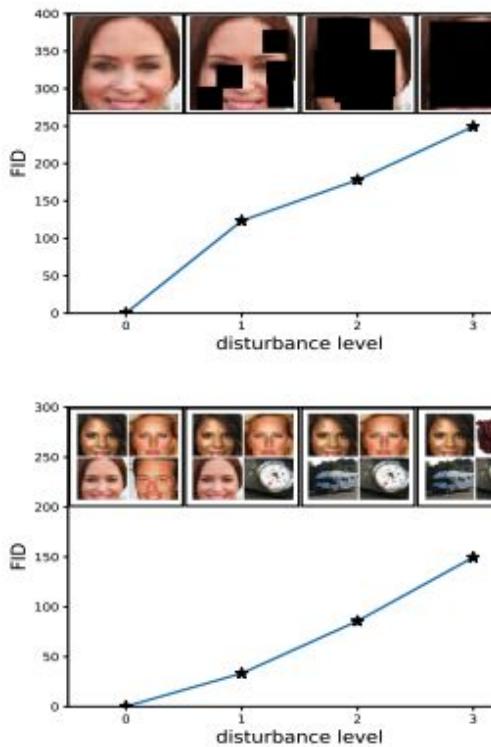
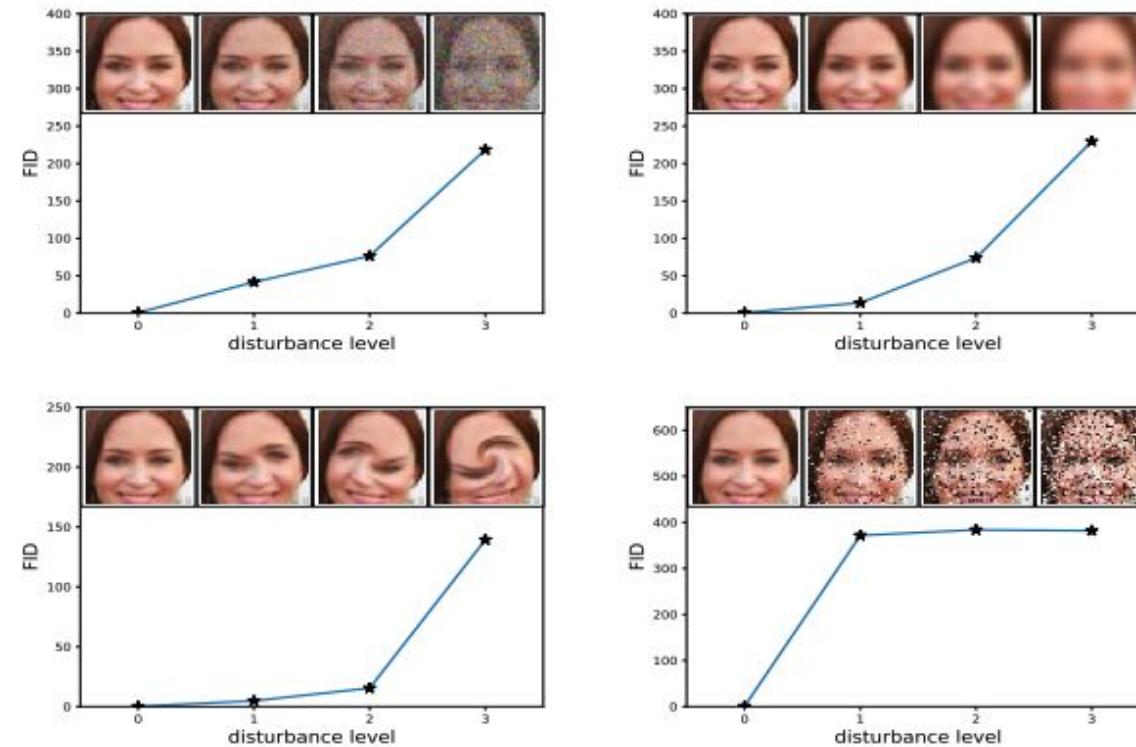


Generative Models Evaluation: FID - Fréchet Inception Distance



[GANs Trained by a Two Time-Scale Update Rule](#)
Converge to a Local Nash Equilibrium

Generative Models Evaluation: FID - Fréchet Inception Distance



[GANs Trained by a Two Time-Scale Update Rule](#)
Converge to a Local Nash Equilibrium

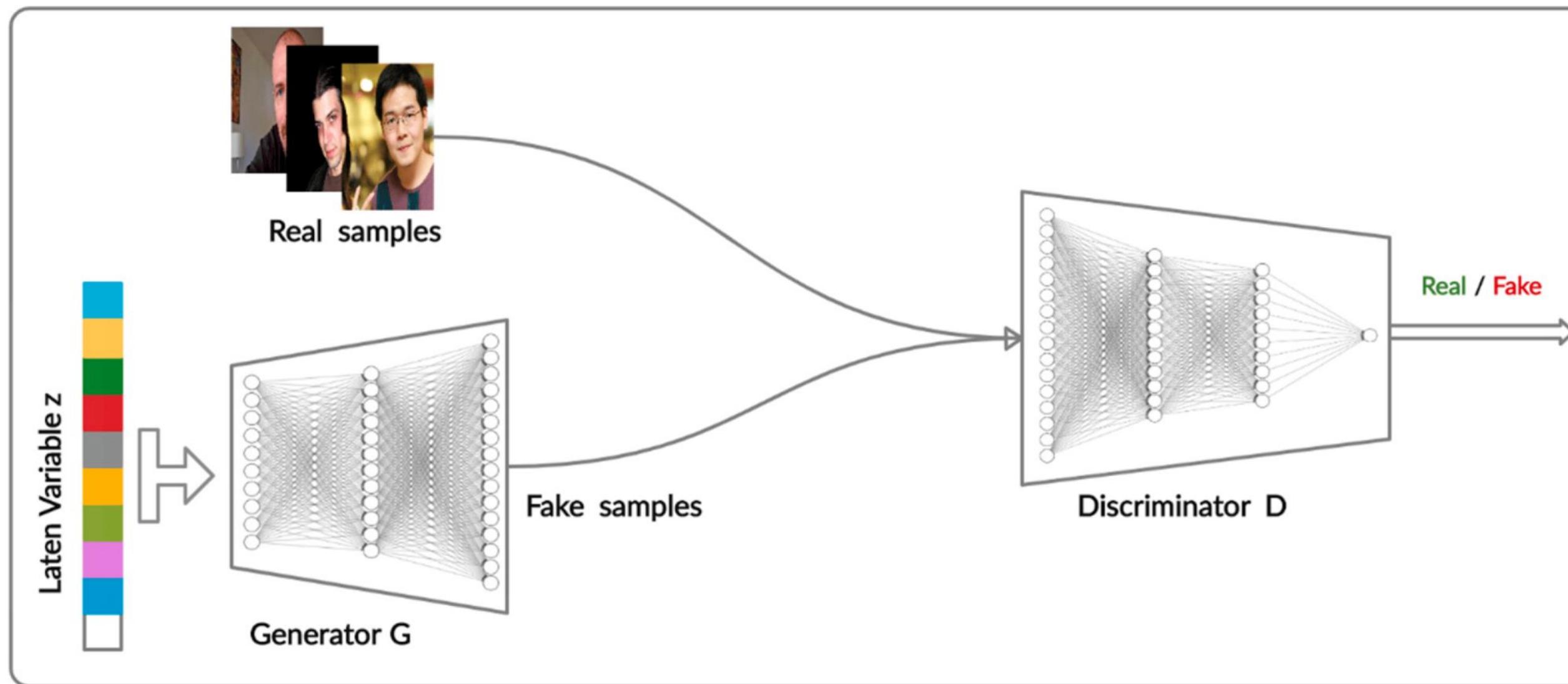
StyleGAN

Nvidia, 2018



Analyzing and Improving the Image Quality of
StyleGAN

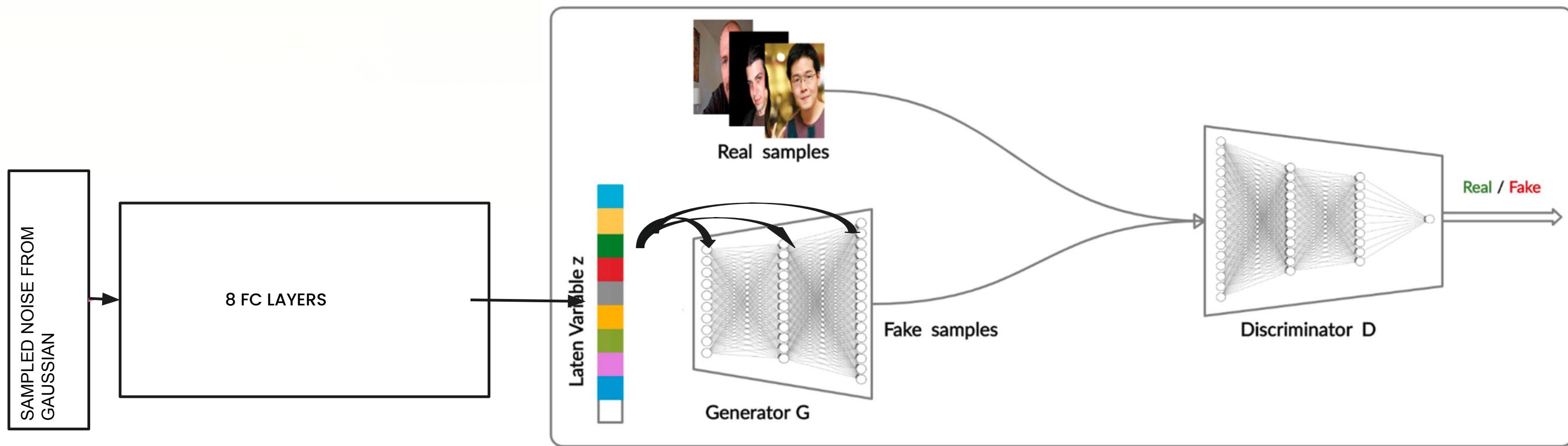
GAN



[**Generative Adversarial Networks**](#), Ian Goodfellow, Yoshua Bengio

StyleGAN

Nvidia, 2018

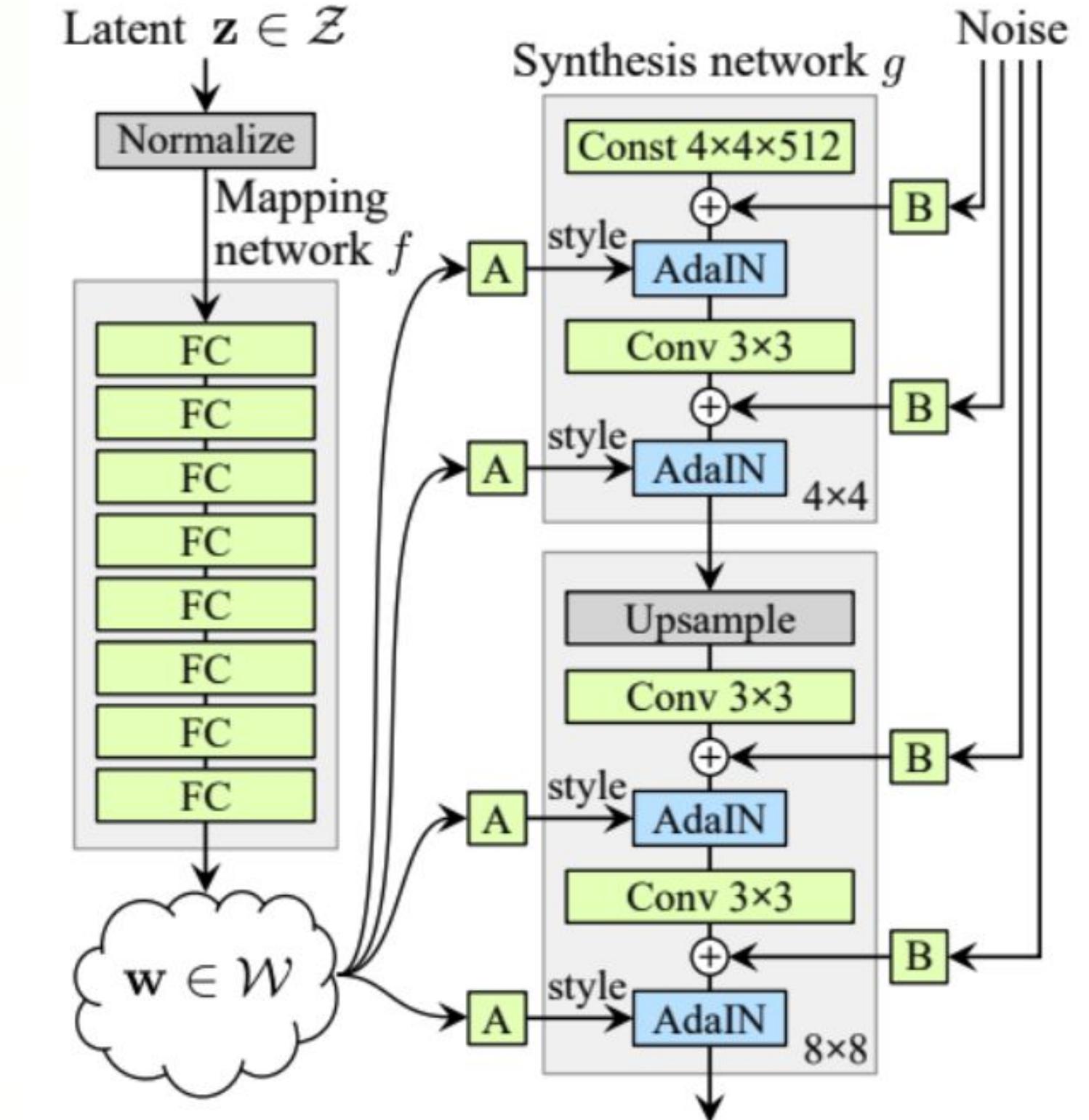


Analyzing and Improving the Image Quality of
StyleGAN

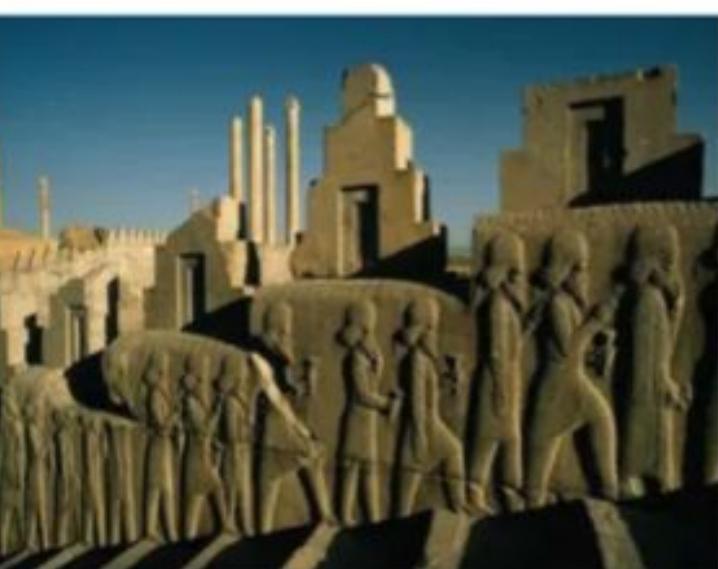
StyleGAN

Nvidia, 2018

Analyzing and Improving the Image Quality of
StyleGAN



AdaIN Block - Style Transfer



+



=

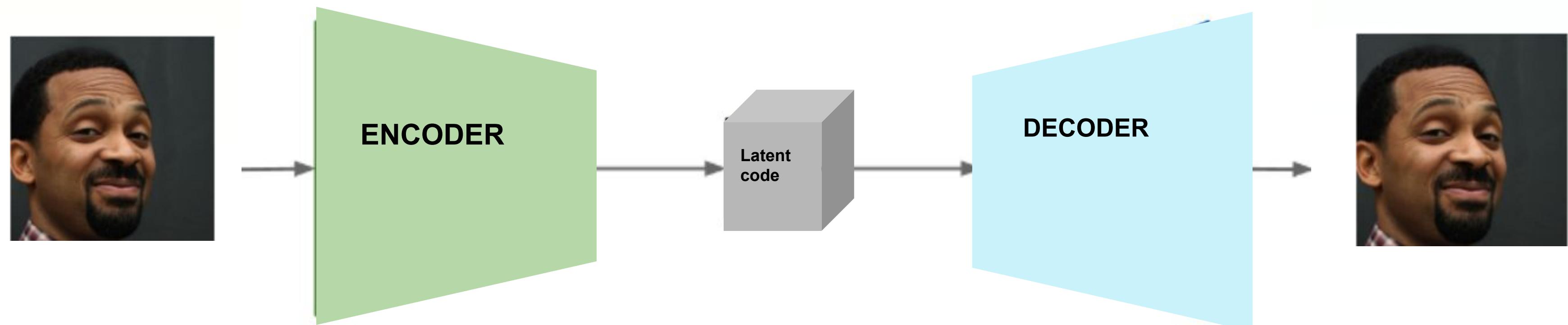


Ancient city of Persepolis

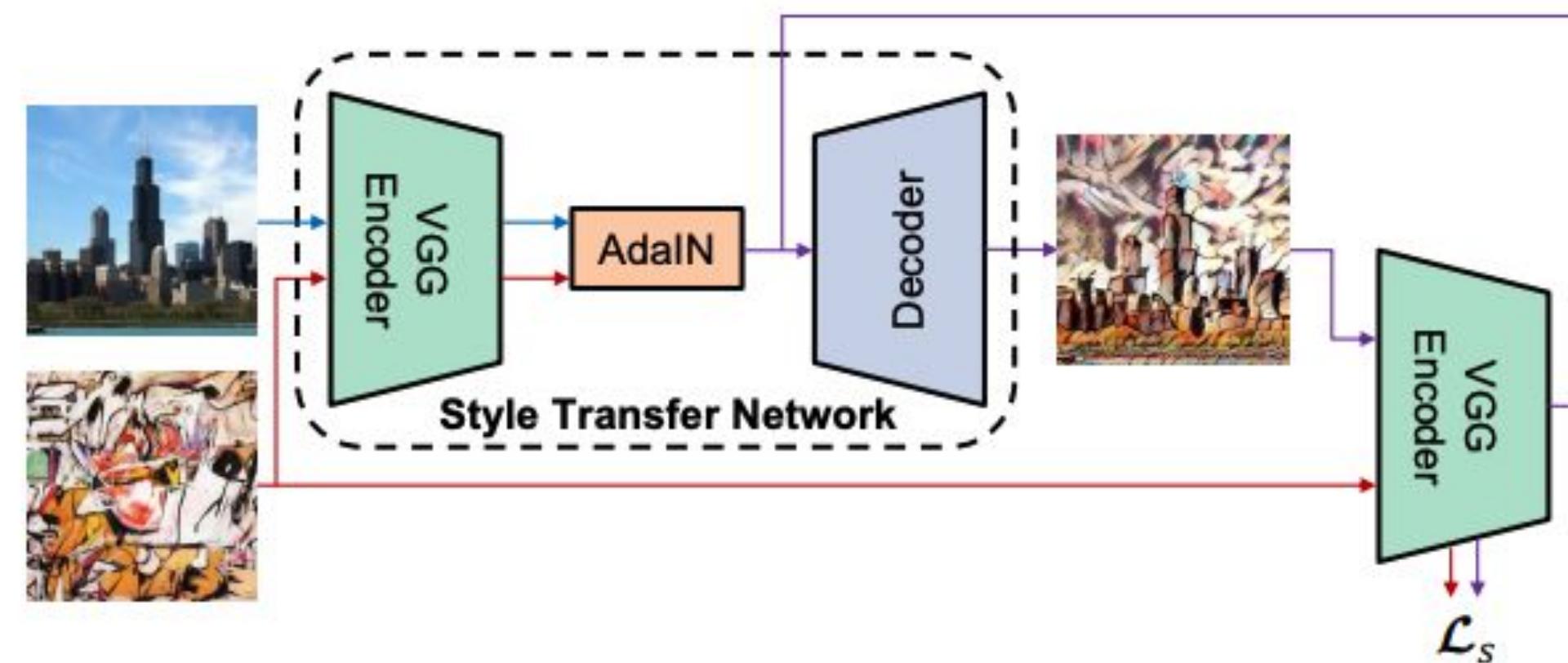
The Starry Night (Van Gogh)

Persepolis
in Van Gogh style

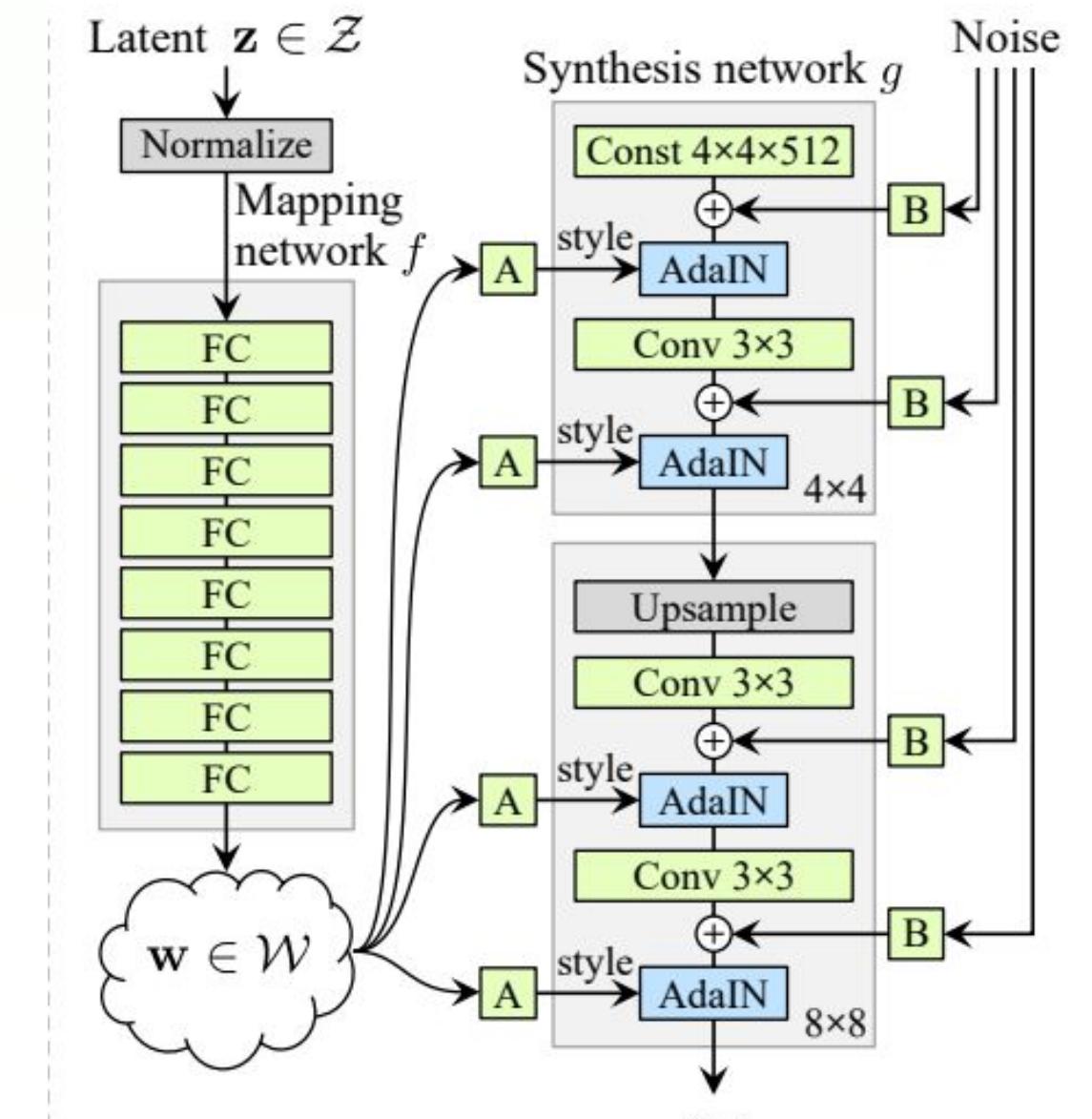
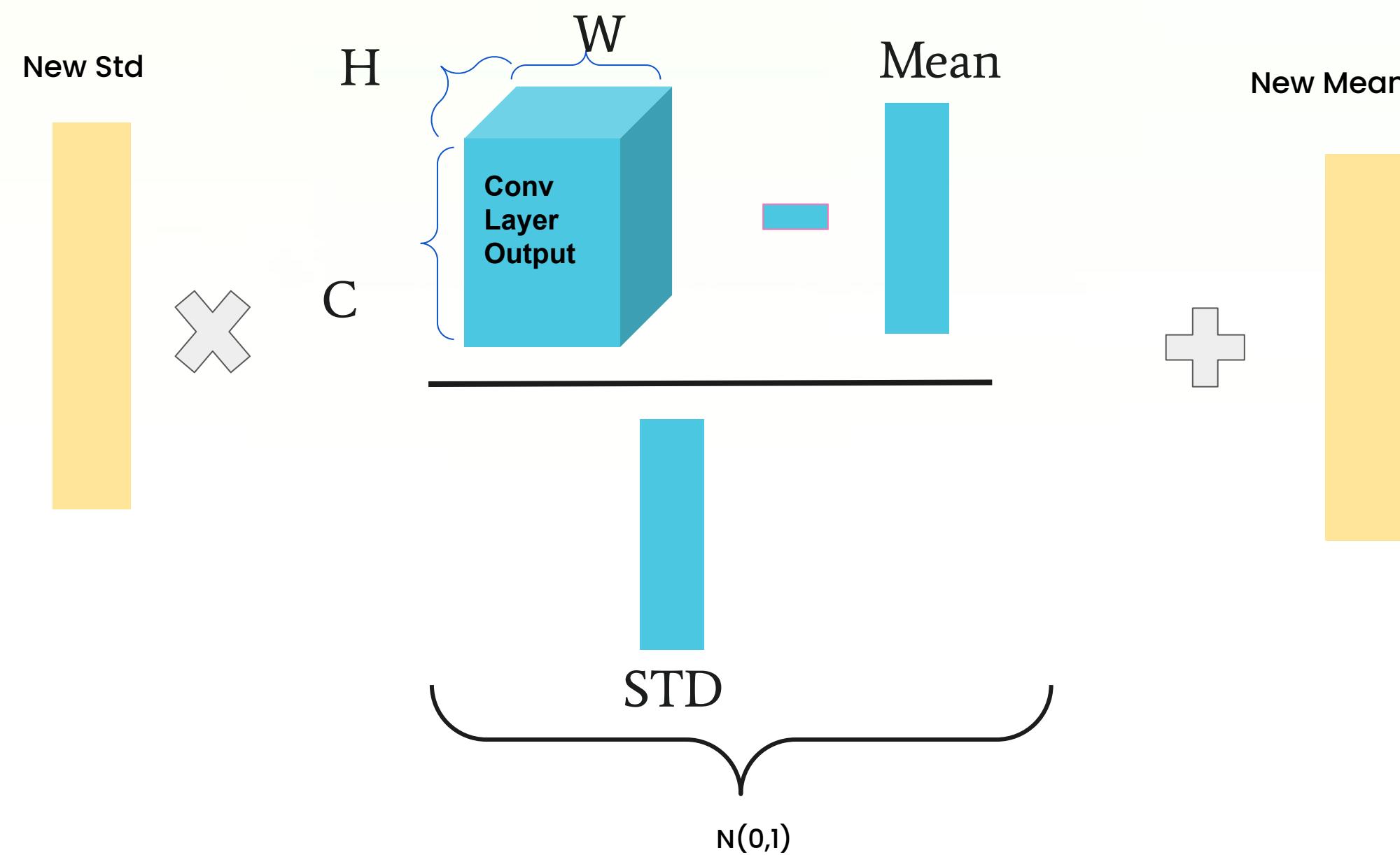
AutoEncoder



Introducing AdalN Block



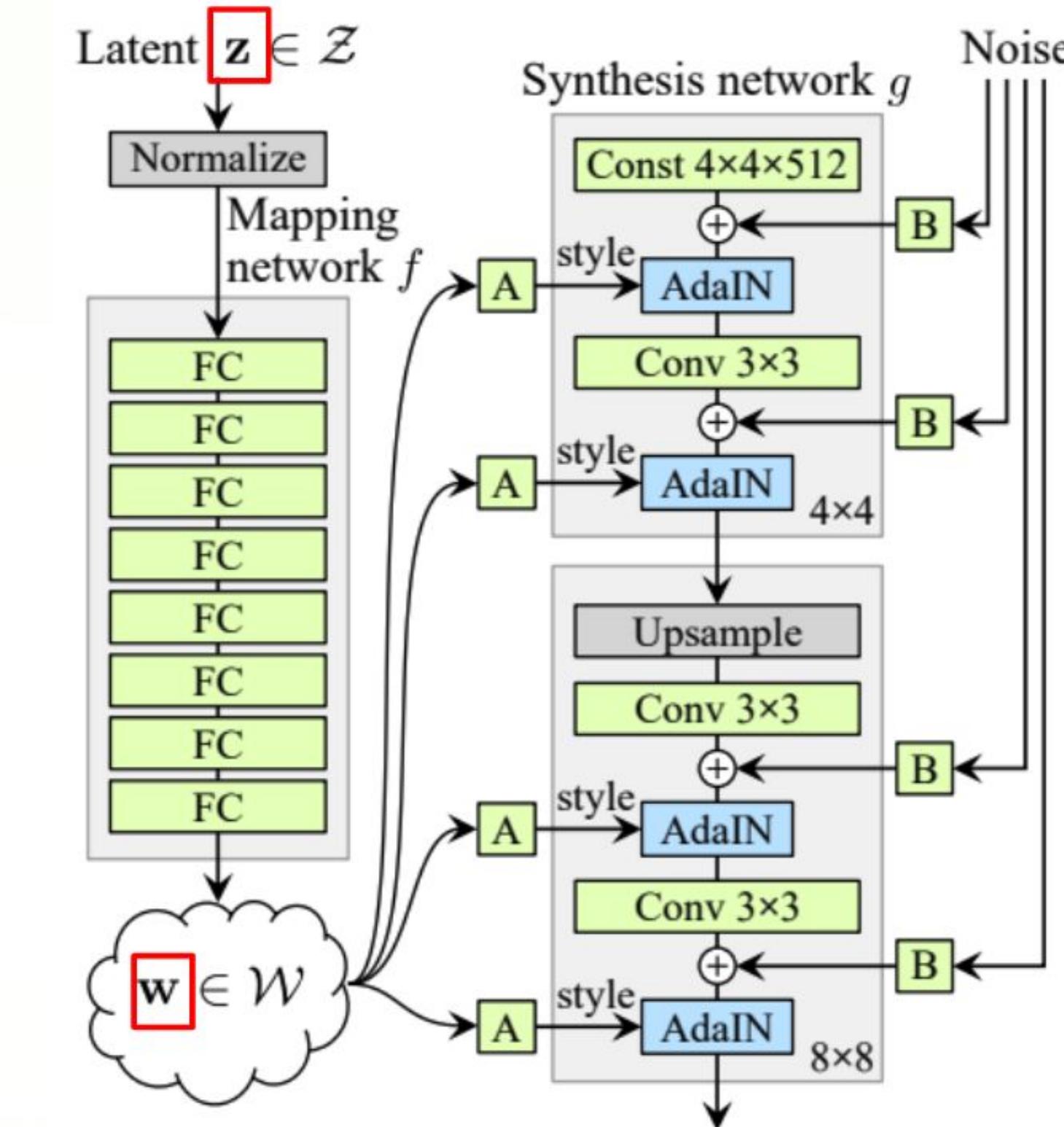
[X. Huang, S. Belongie, Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. ICCV 2017](#)



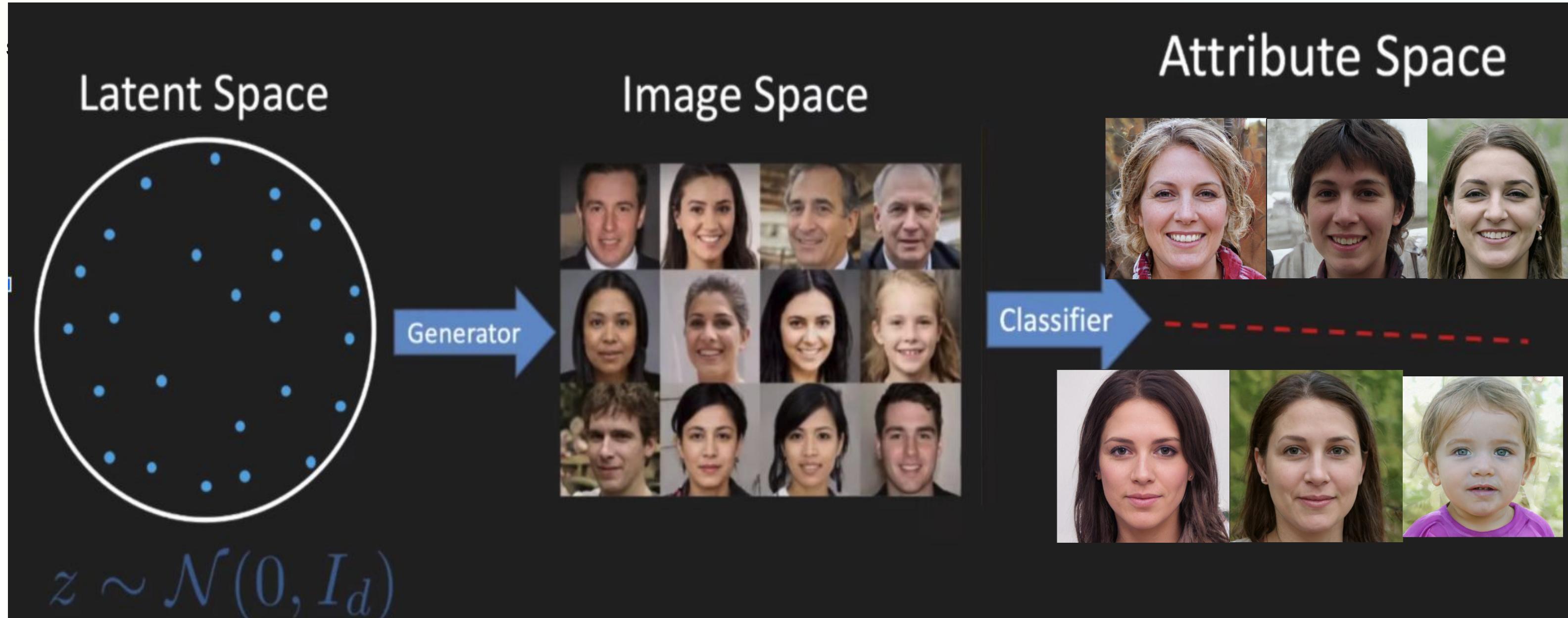
(b) Style-based generator

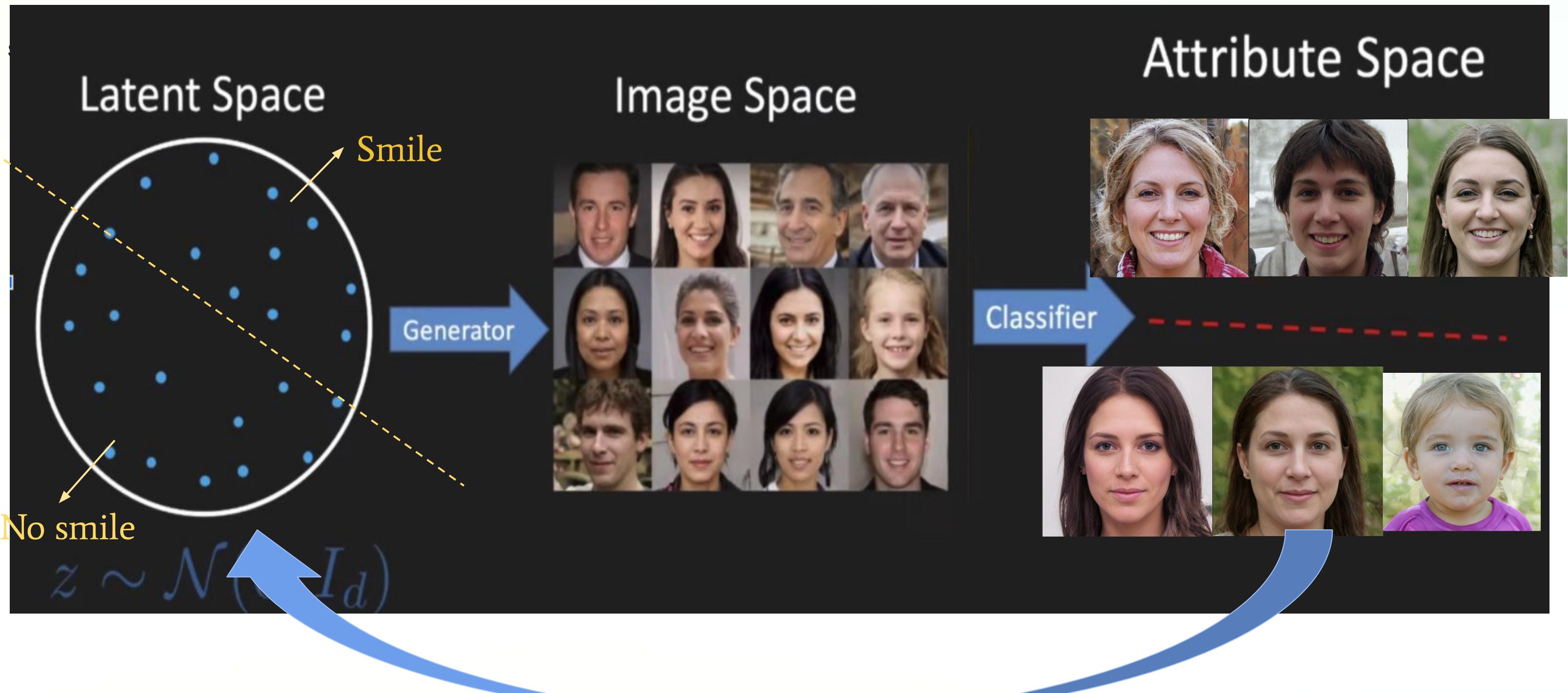
$$\text{AdaIN}(\mathbf{x}_i, \mathbf{y}) = \mathbf{y}_{s,i} \frac{\mathbf{x}_i - \mu(\mathbf{x}_i)}{\sigma(\mathbf{x}_i)} + \mathbf{y}_{b,i},$$

StyleGAN Control











StyleGAN Inversion Direct / Encoder

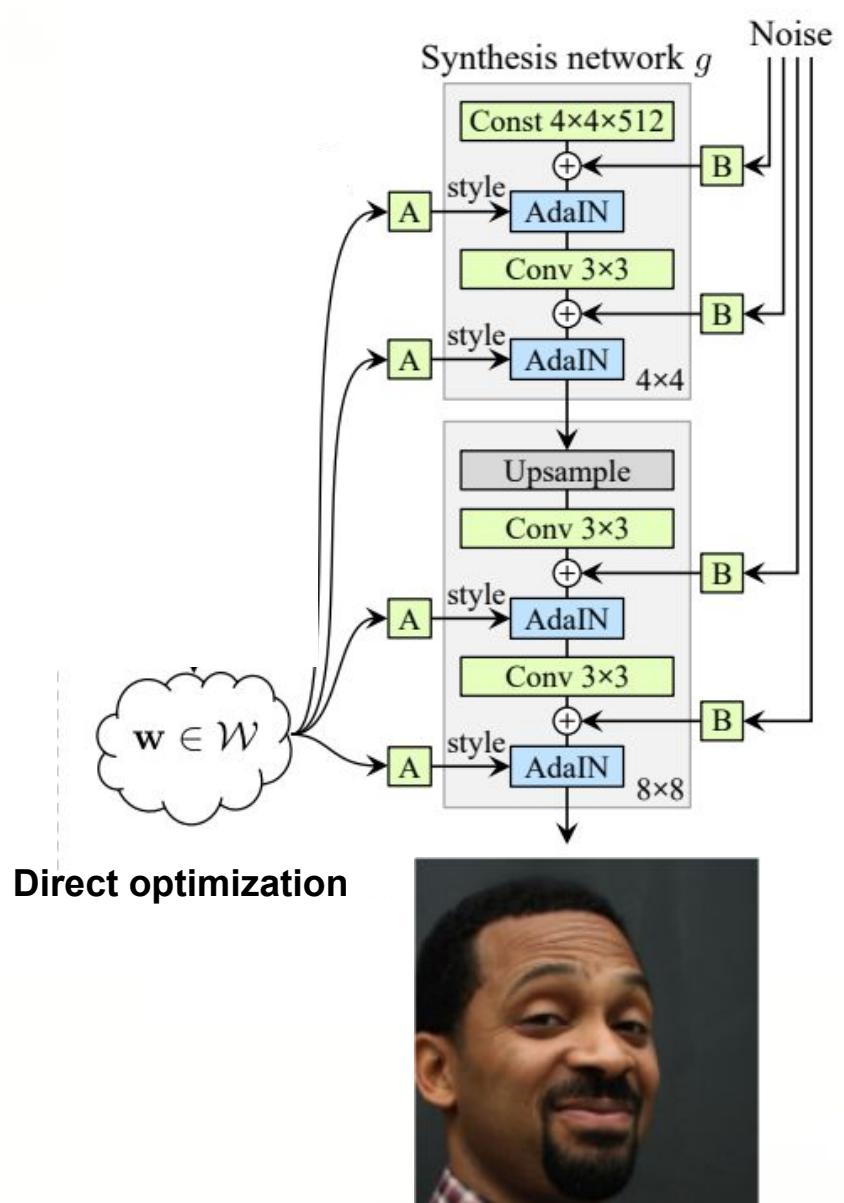


Input

Direct optimization

Encoder

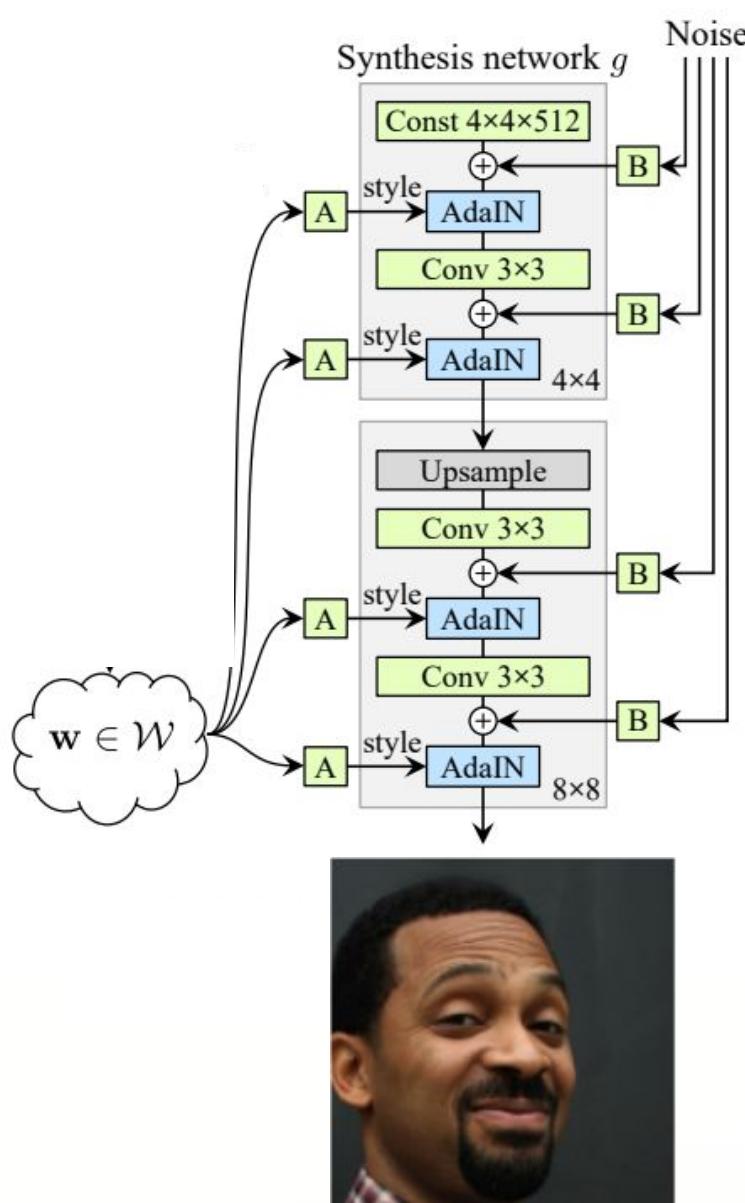
StyleGAN Inversion Direct / Encoder



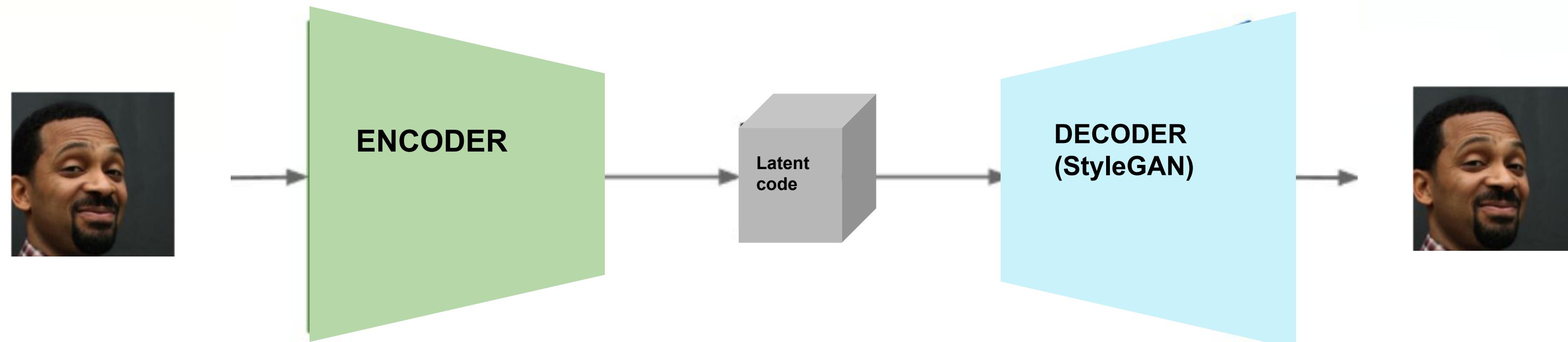
StyleGAN Inversion Direct / Encoder



Encoder

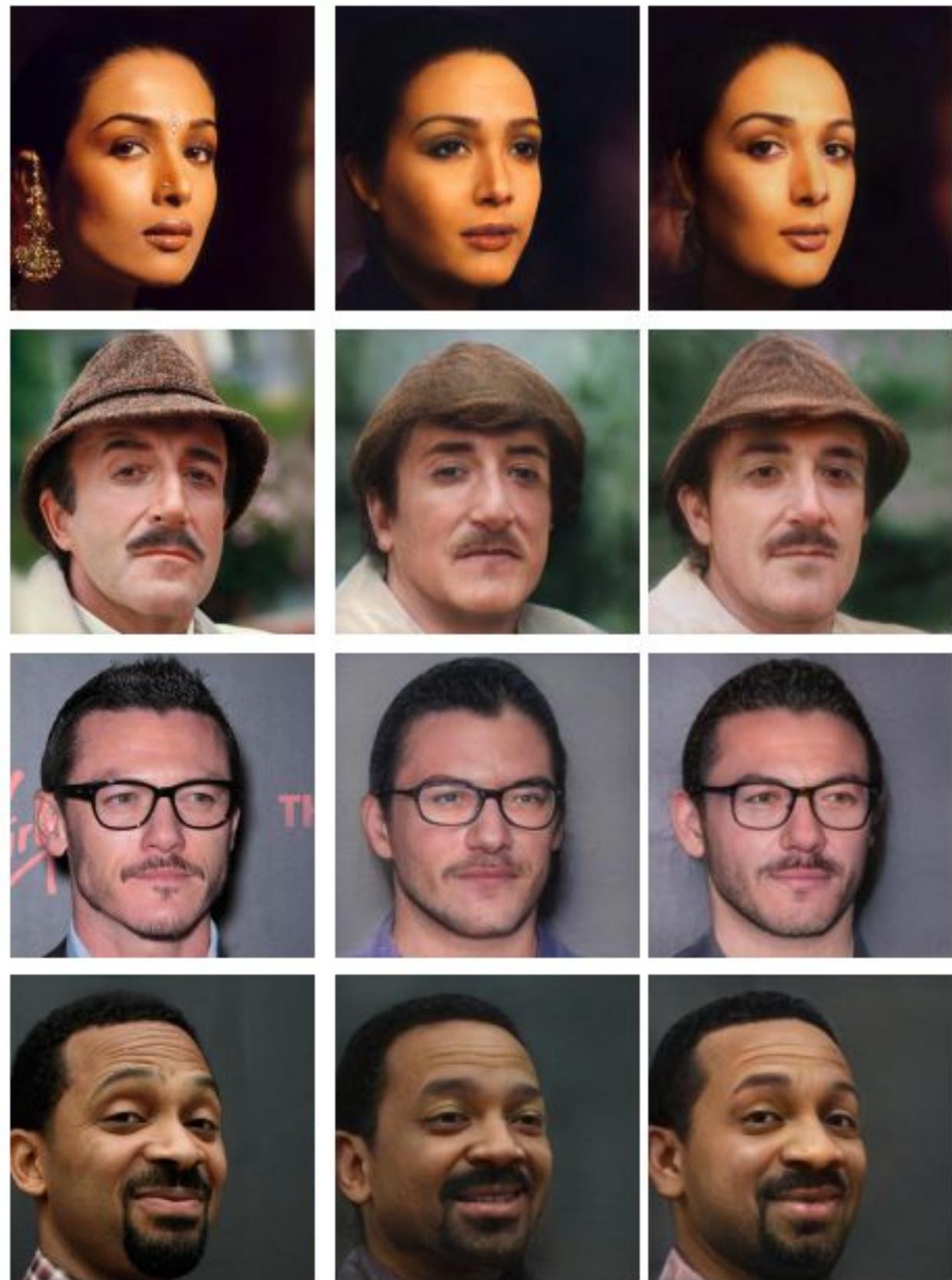


AutoEncoder



StyleGAN Inversion Direct / Encoder

which one is better?



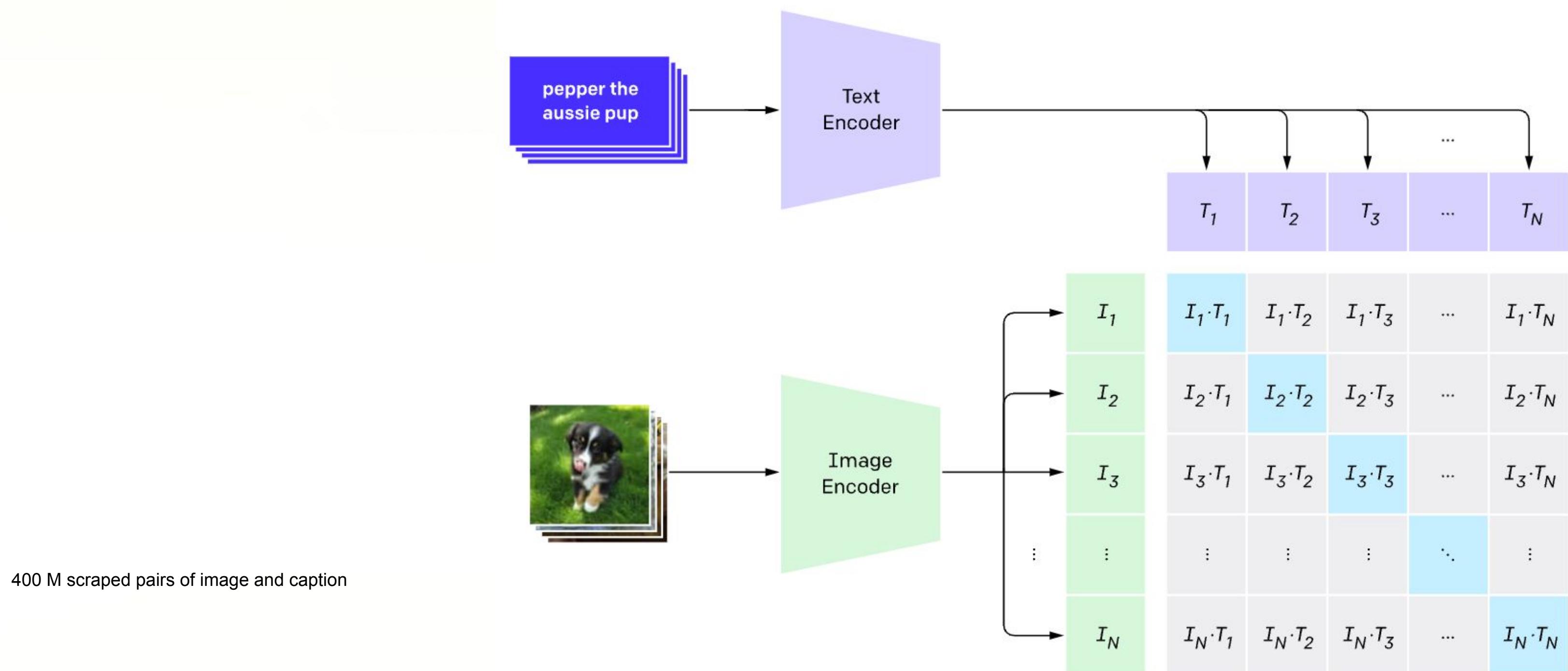
Input

Direct optimization

Encoder

CLIP- Text and Image encoder

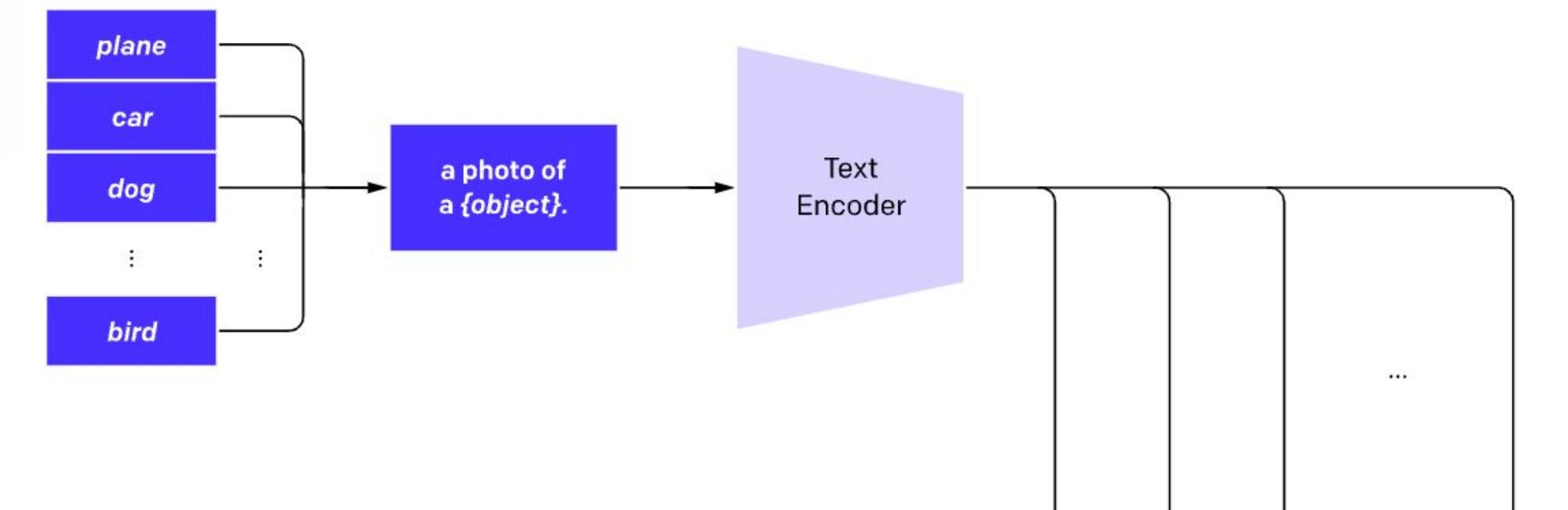
OpenAI, Jan 2021



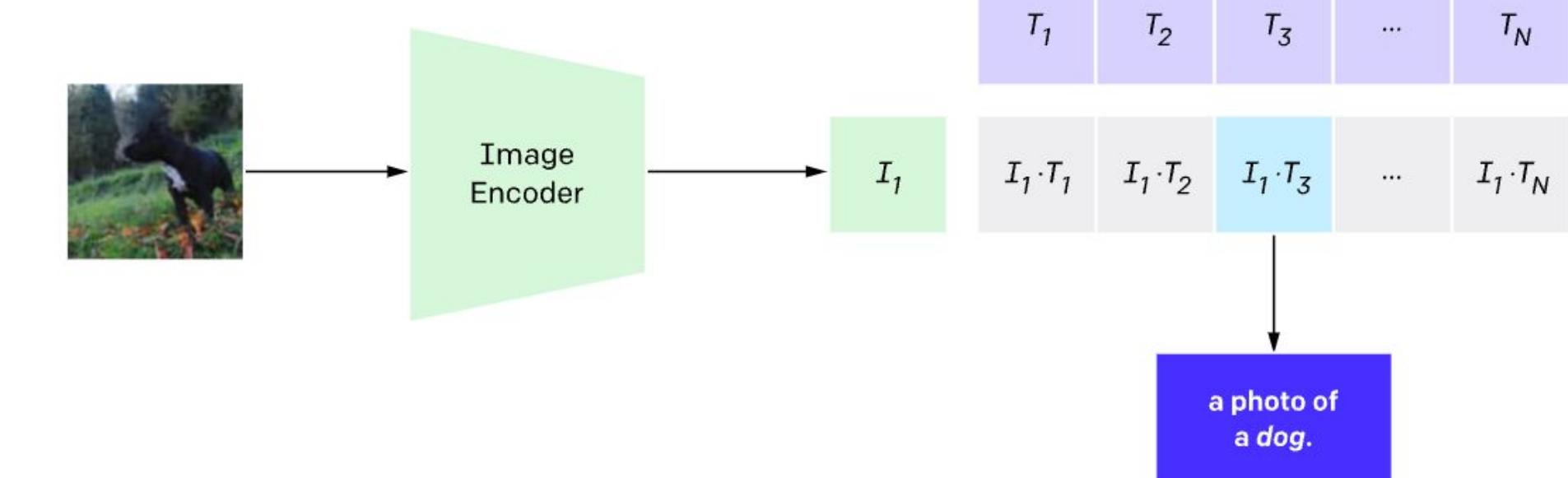
CLIP- Text and Image encoder

OpenAI, Jan 2021

2. Create dataset classifier from label text



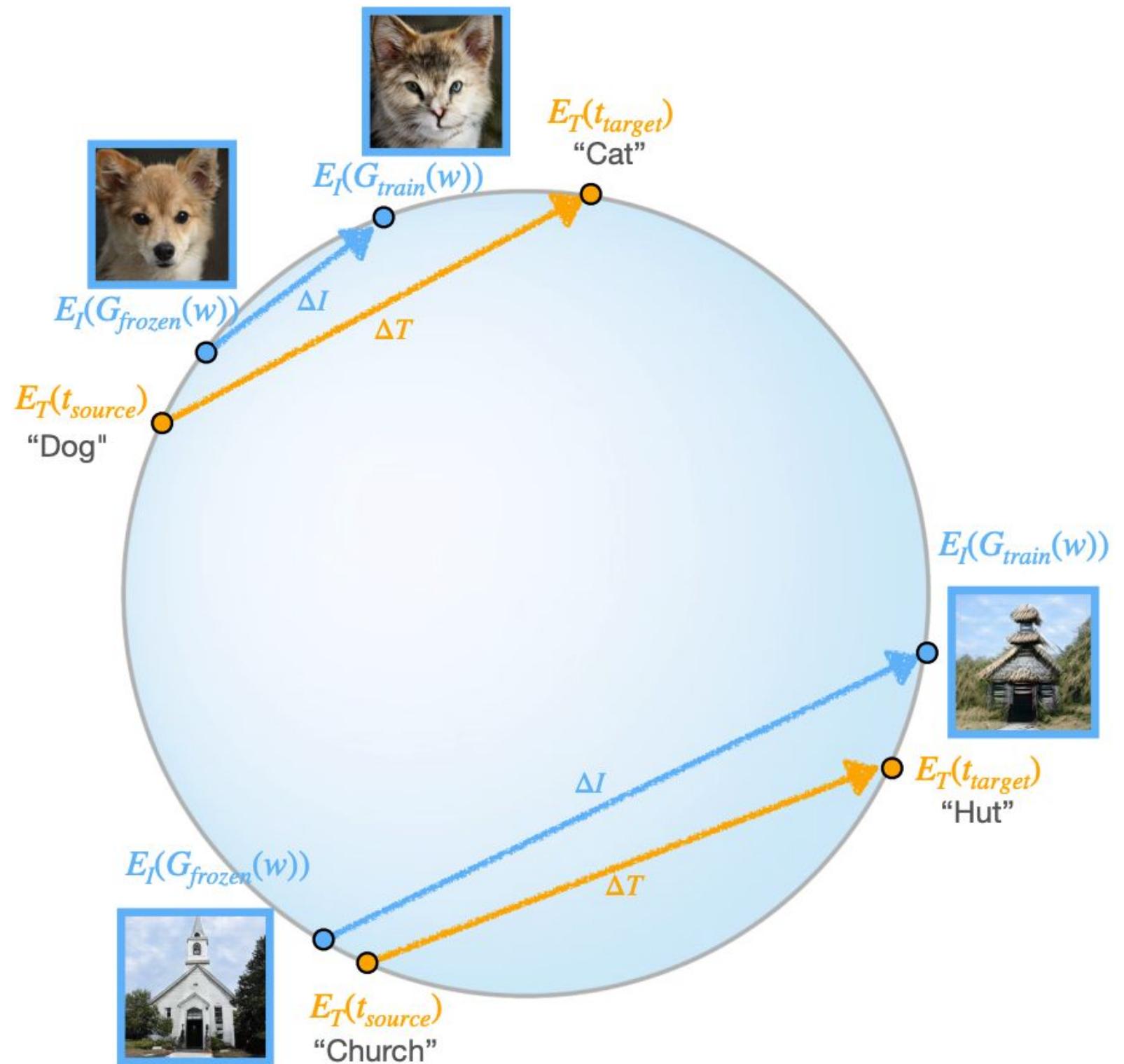
3. Use for zero-shot prediction



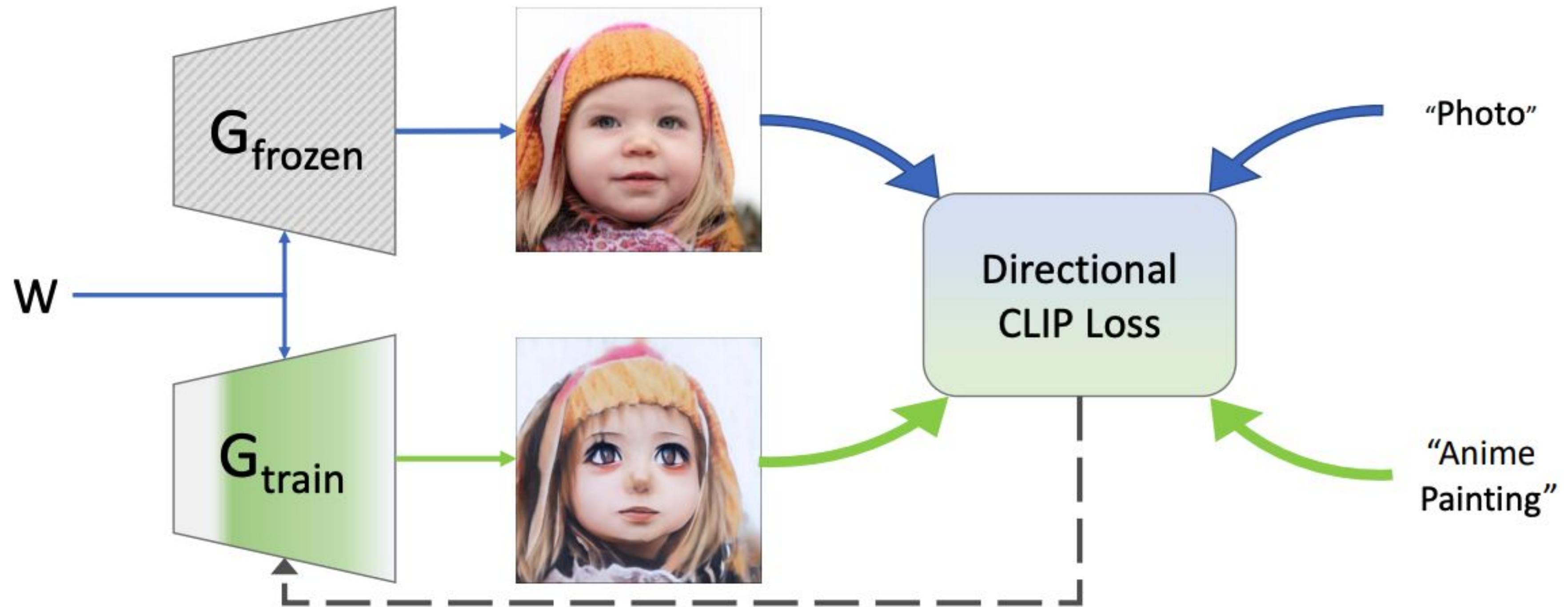
StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery

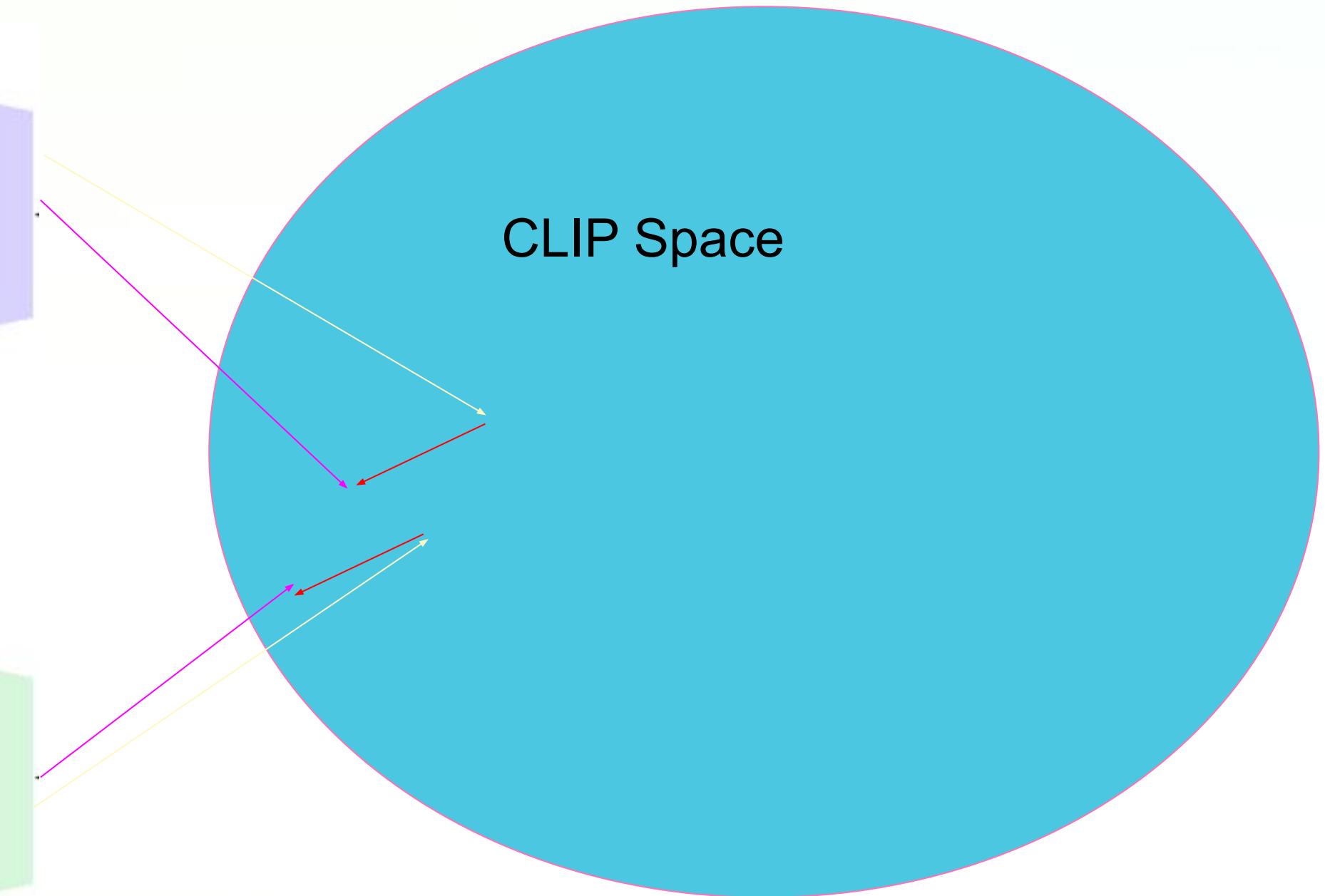
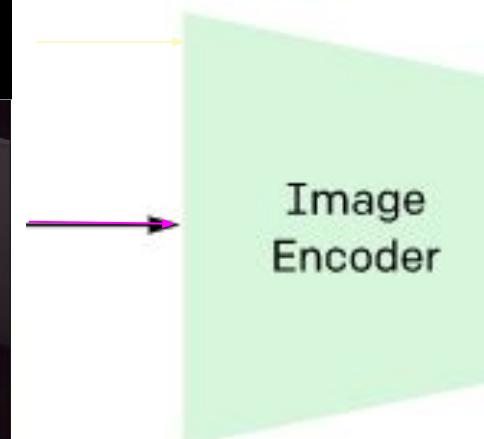
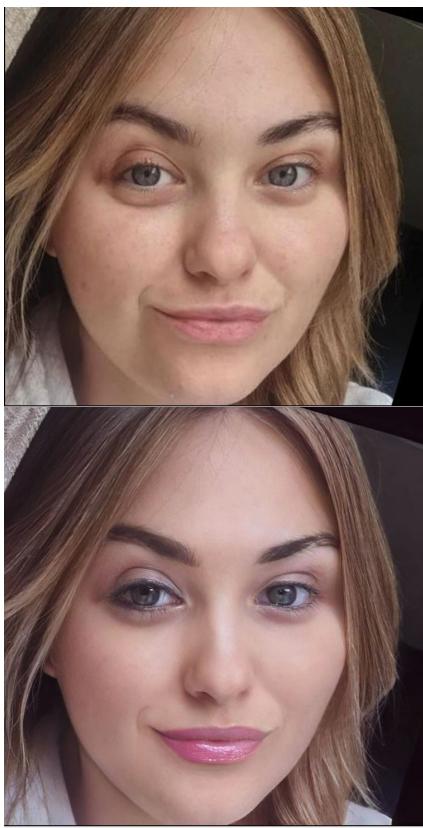
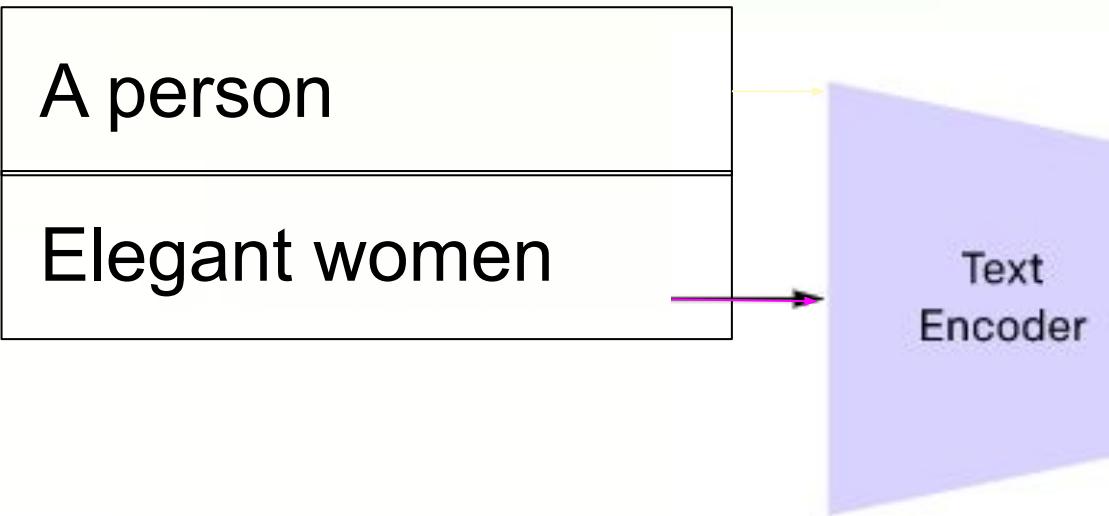
StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators



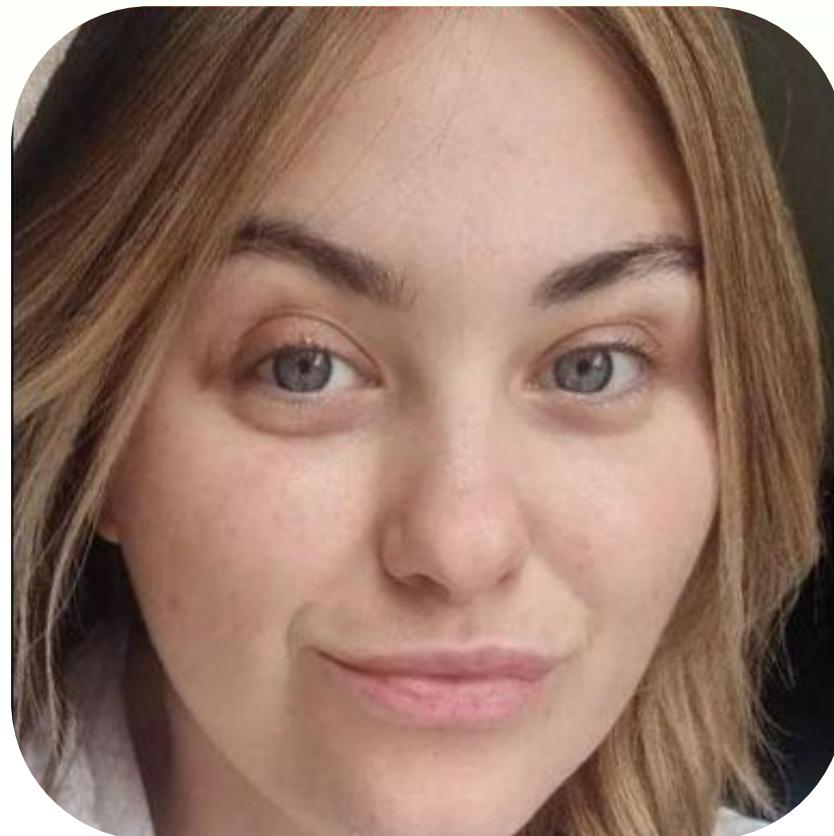


$$\begin{aligned}\Delta T &= E_T(t_{target}) - E_T(t_{source}), \\ \Delta I &= E_I(G_{train}(w)) - E_I(G_{frozen}(w)), \\ \mathcal{L}_{direction} &= 1 - \frac{\Delta I \cdot \Delta T}{|\Delta I| |\Delta T|},\end{aligned}$$

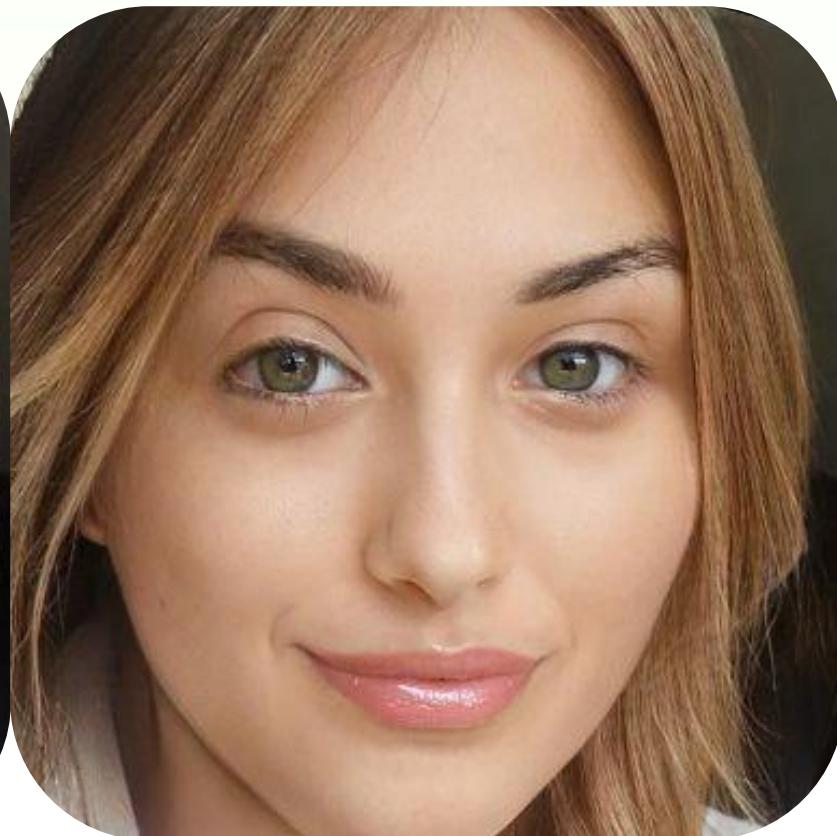




Face GANs

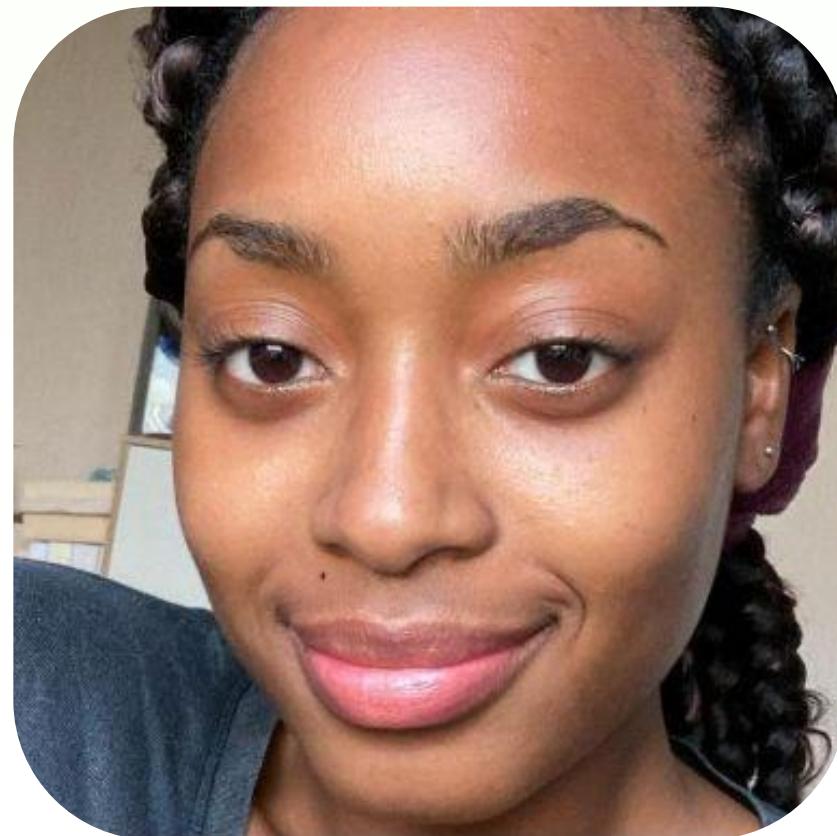


Before

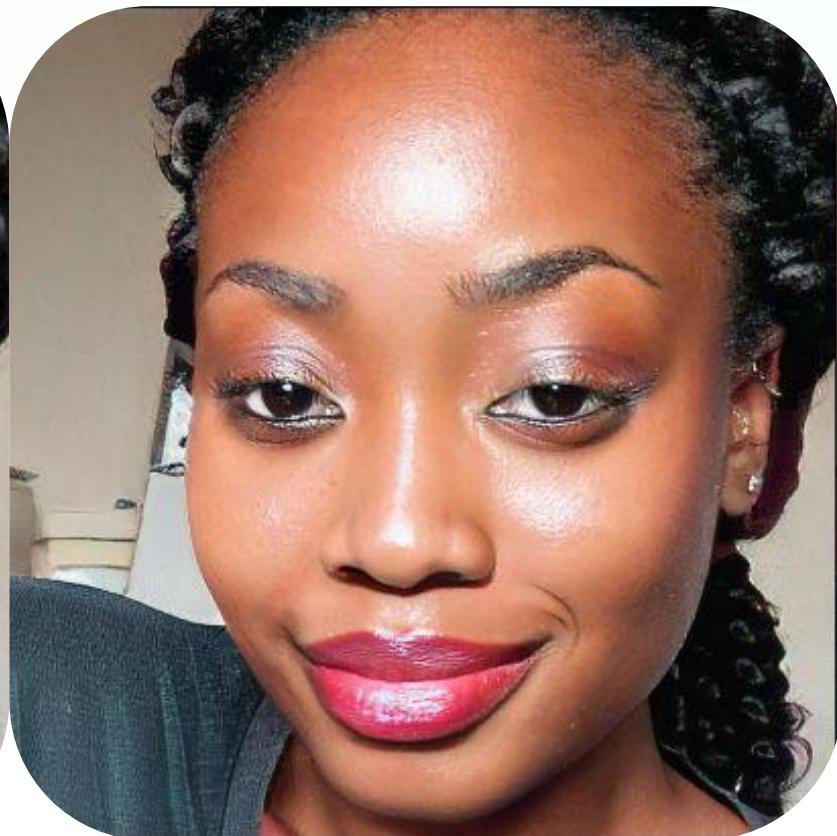


After

"Elegant business"



Before



After

"Bold Lips"

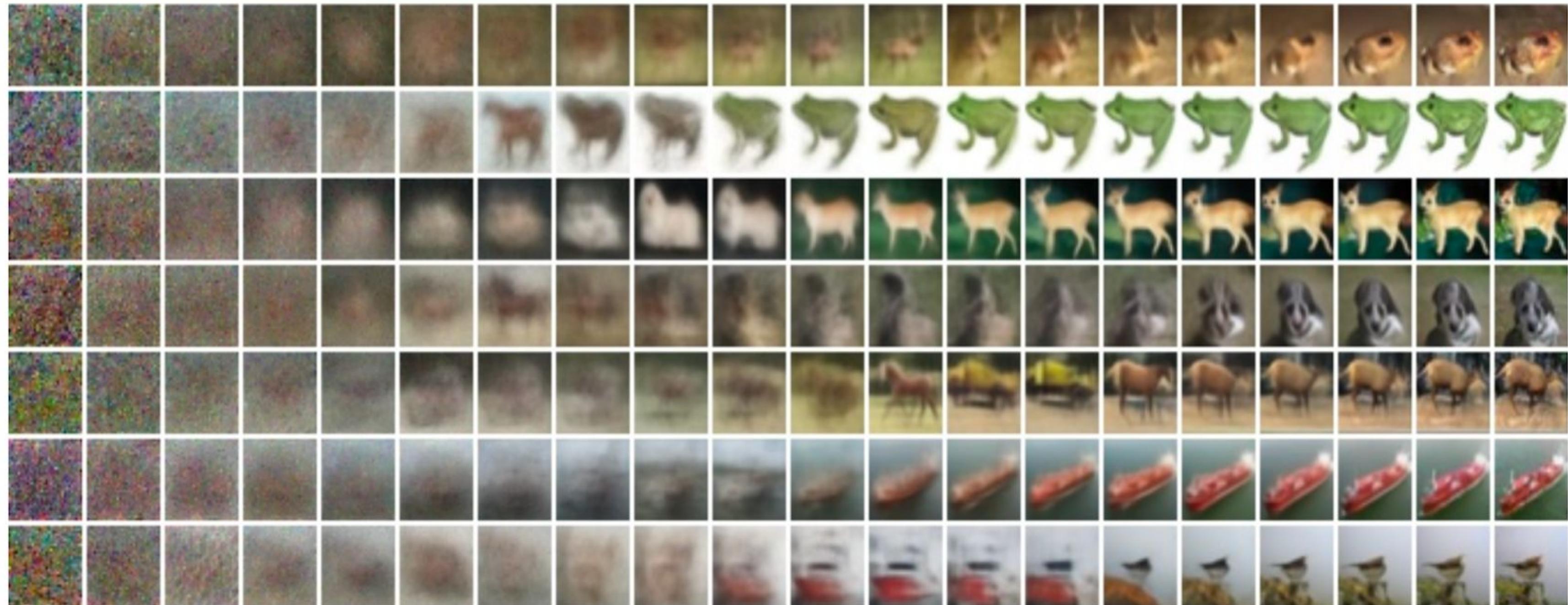
Part 2

GANs Drawbacks

- Unstable Training
- Mode collapse

Diffusion models

Guided image generation, UC Berkeley, June 2020, OpenAI, June 2021, Google Brain, July 2021

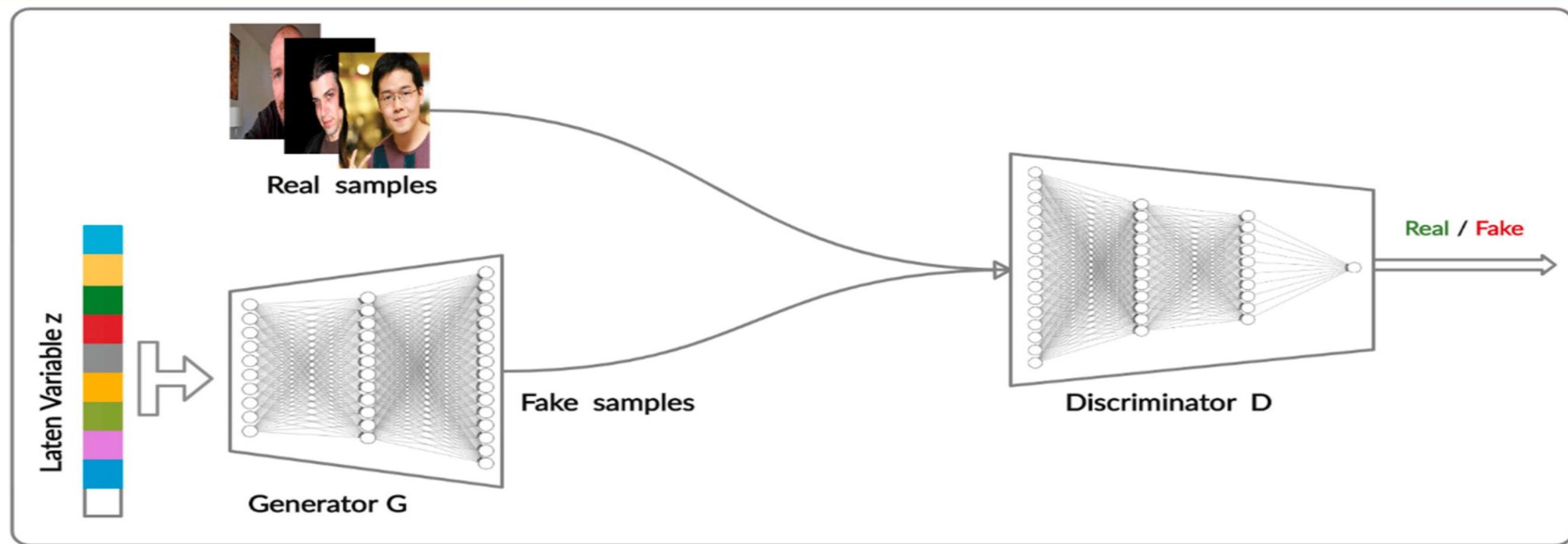


Diffusion model

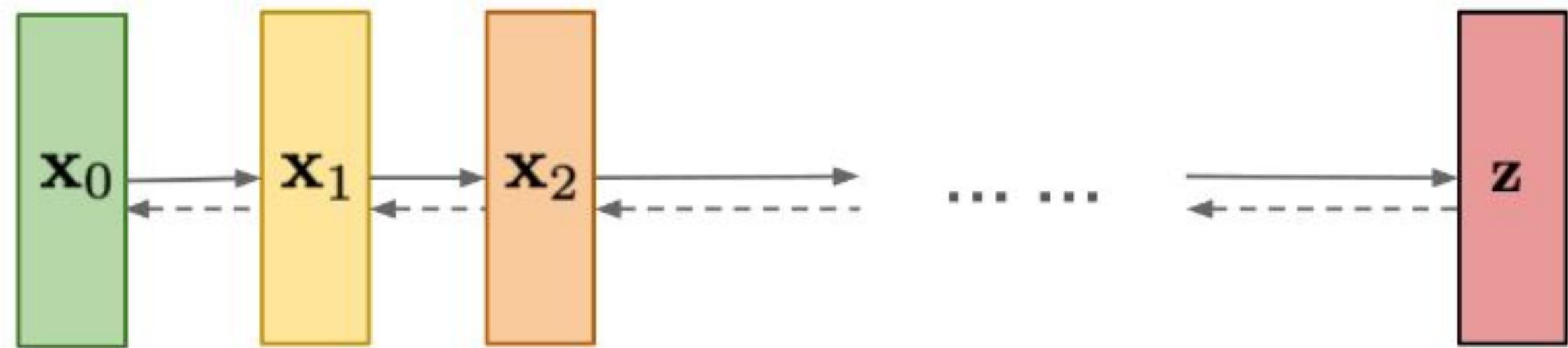
Input: random noise tensor (dimensions -
64x64x6)

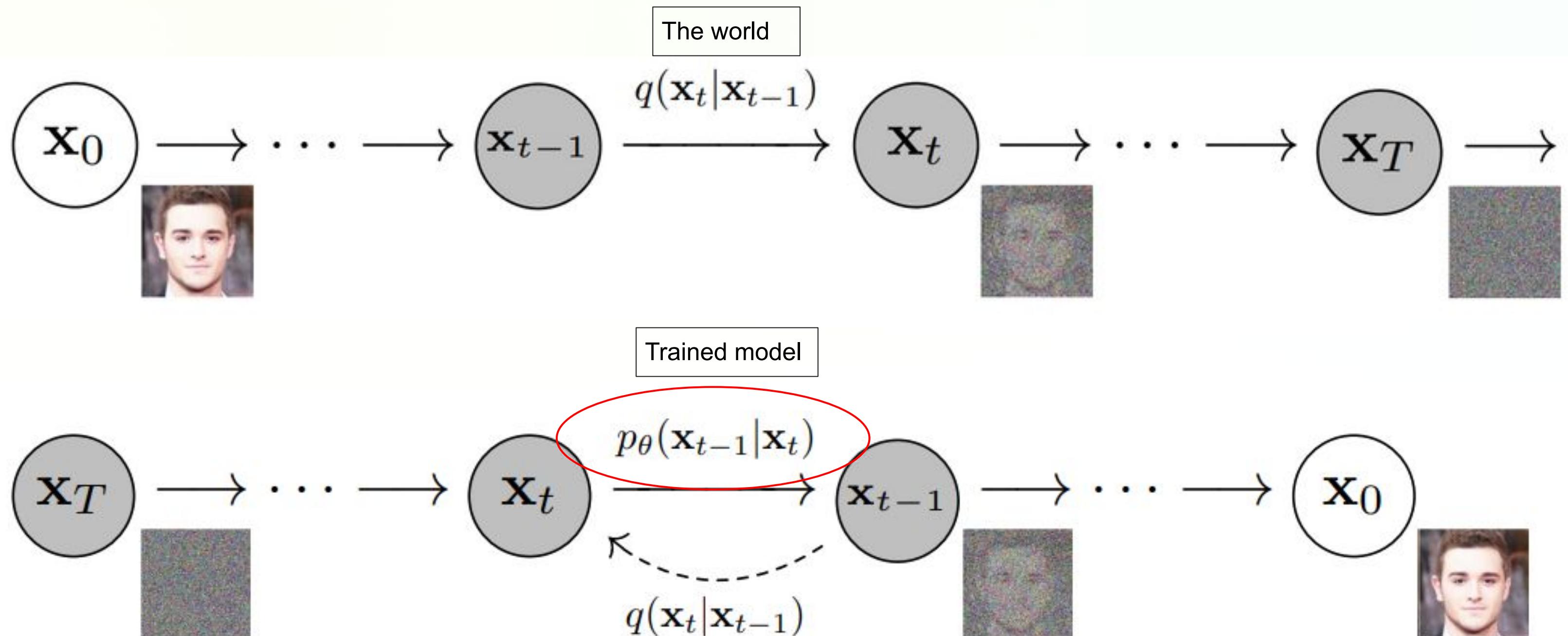
Output: tensor (dimensions 64x64x3)

GAN: Adversarial training



Diffusion models:
Gradually add Gaussian noise and then reverse





Creating Noised Images

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



sample an arbitrary step of the noised latents directly conditioned on the input \mathbf{x}_0

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

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}$$

Training a Diffusion Model - the Reverse Step

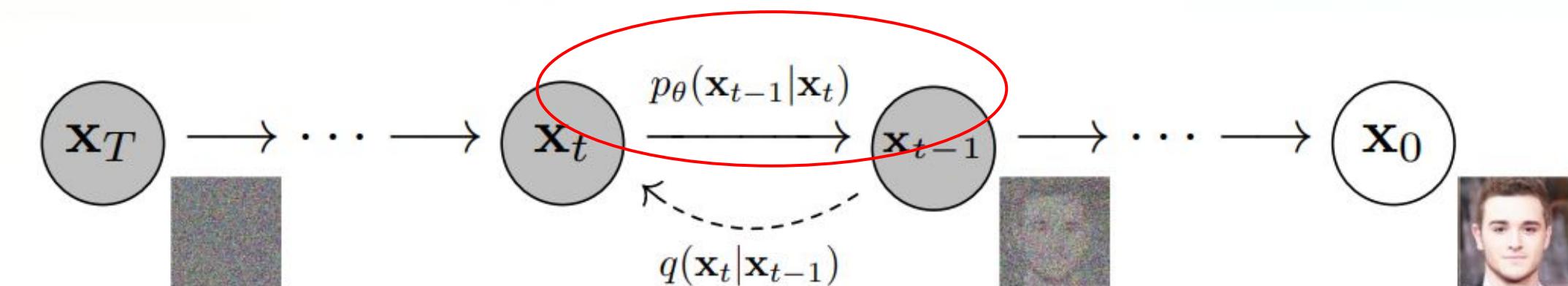
Diffusion model can be trained to predict \mathbf{x}_{t-1} given \mathbf{x}_t, t

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

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$$

Loss: L2($\text{noise}_\theta(\mathbf{x}_t), \epsilon$)

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



Dimensions:

$\bar{\alpha}_t$ - scalar

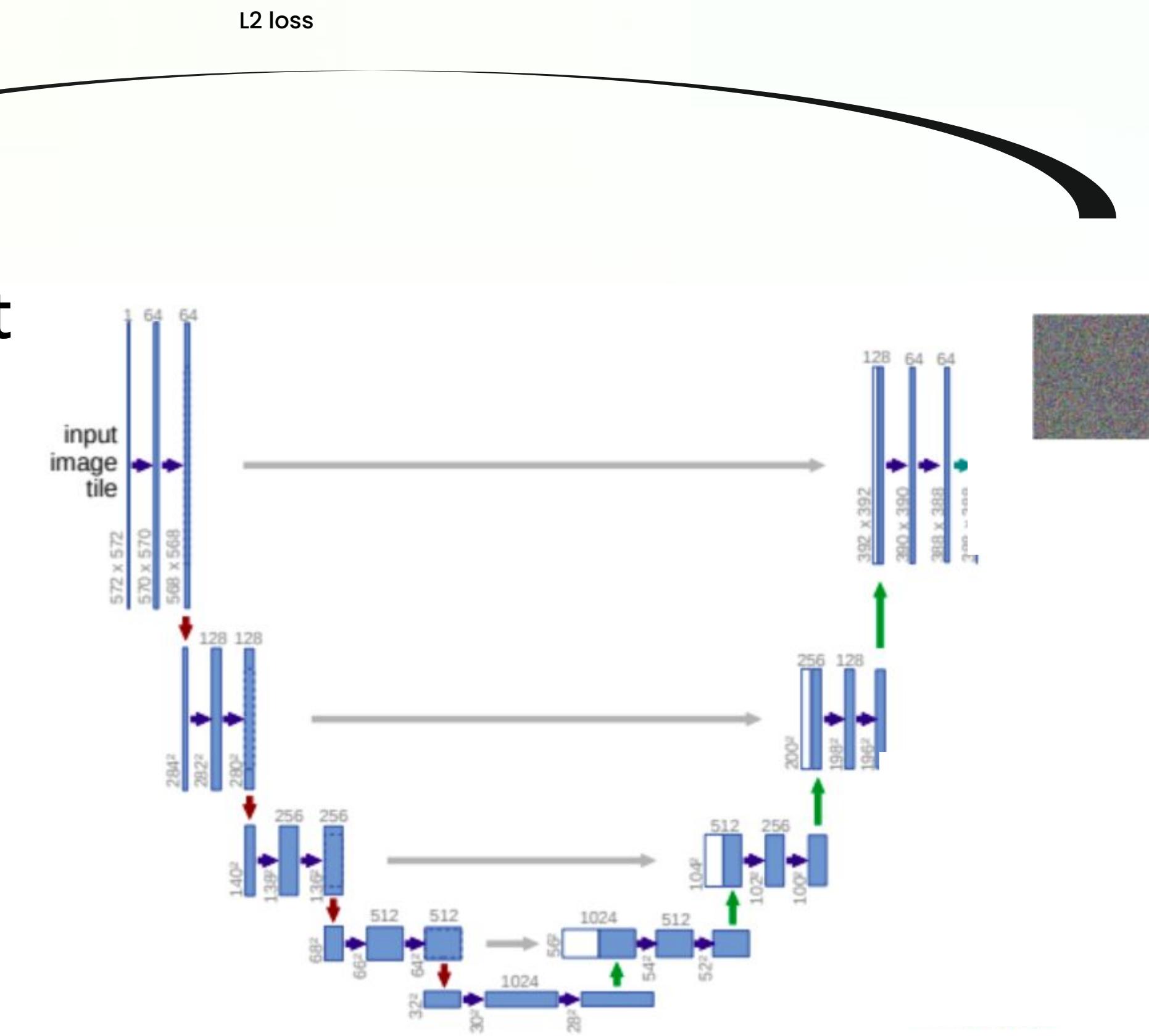
\mathbf{x}_t Tensor 64x64x3

Training process

$$\sqrt{\bar{\alpha}_t} \quad \begin{matrix} + \\ \text{pink plus sign} \end{matrix} \quad \sqrt{1 - \bar{\alpha}_t} \quad \begin{matrix} = \\ \text{pink equals sign} \end{matrix} \quad ,$$

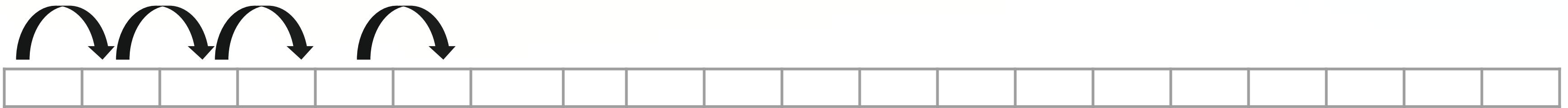
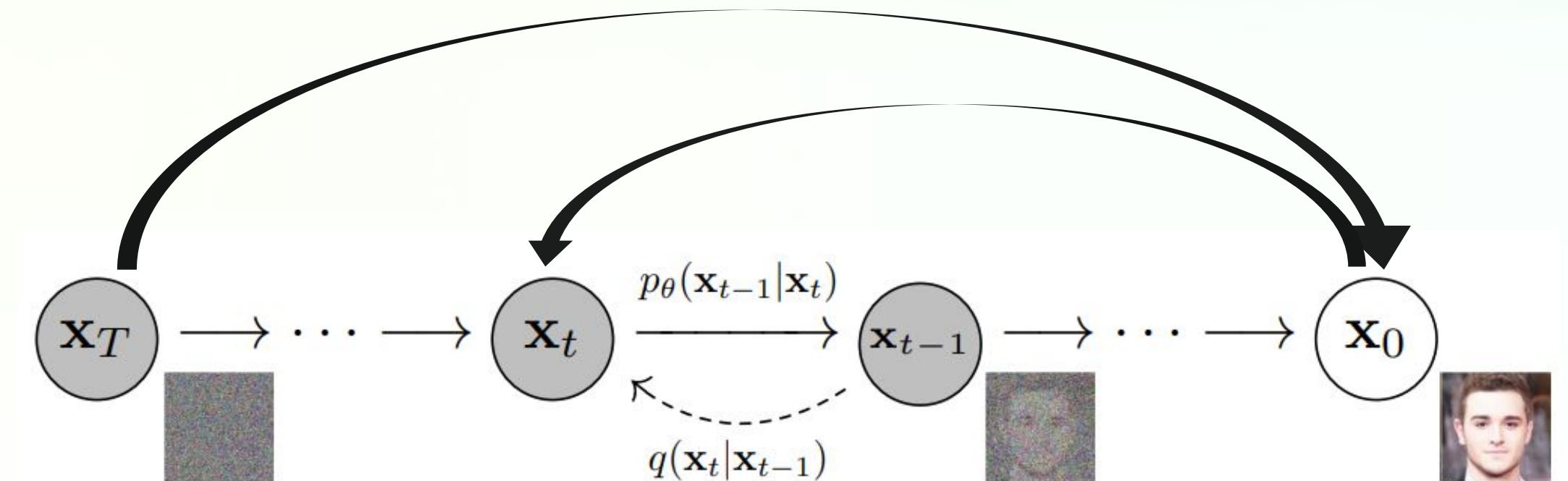
$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$$

t in (0,1000)



Inference

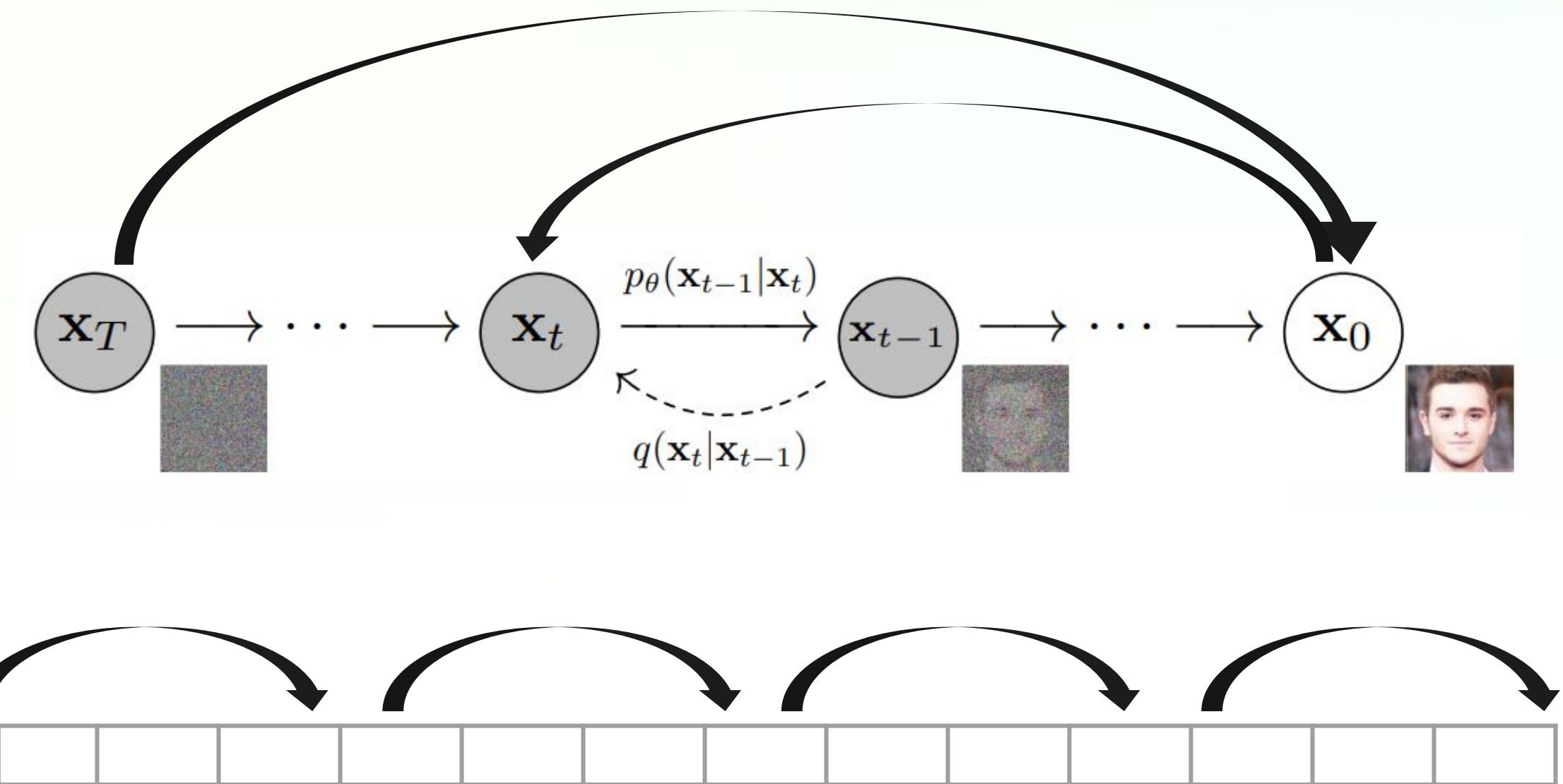
1000 steps??



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

Inference

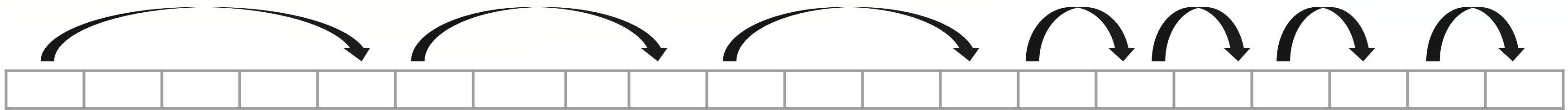
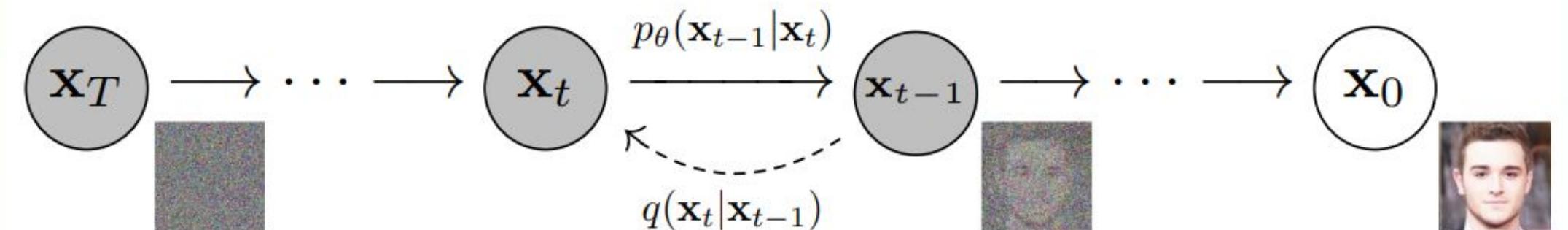
1000 steps??



$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$$

Inference

1000 steps??



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

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.

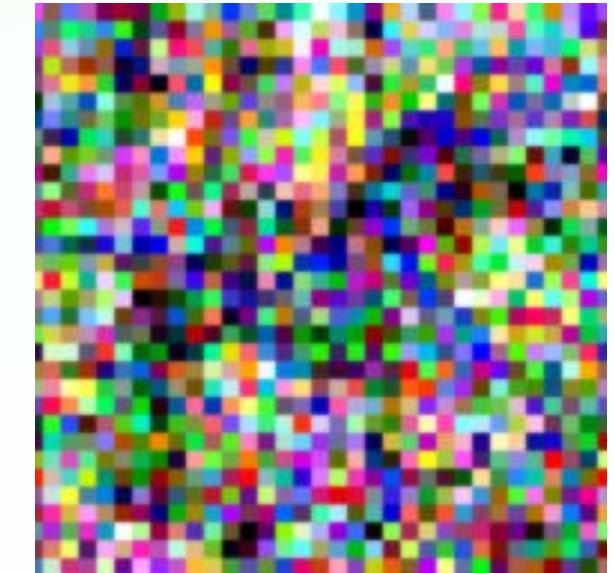
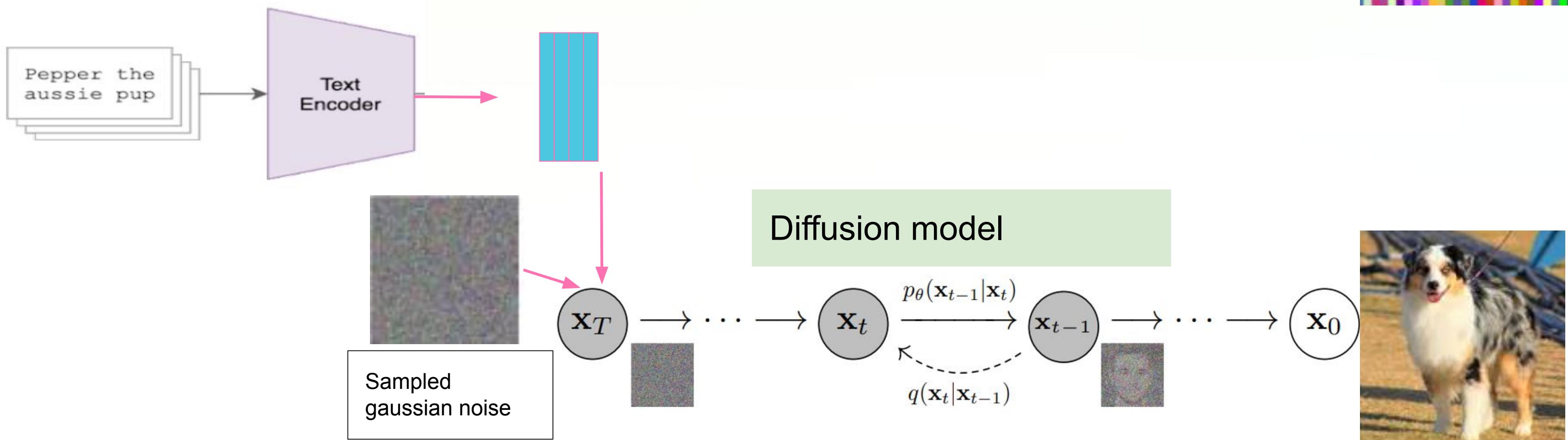


"a boat in the canals of venice"



"a painting of a fox in the style of starry night"

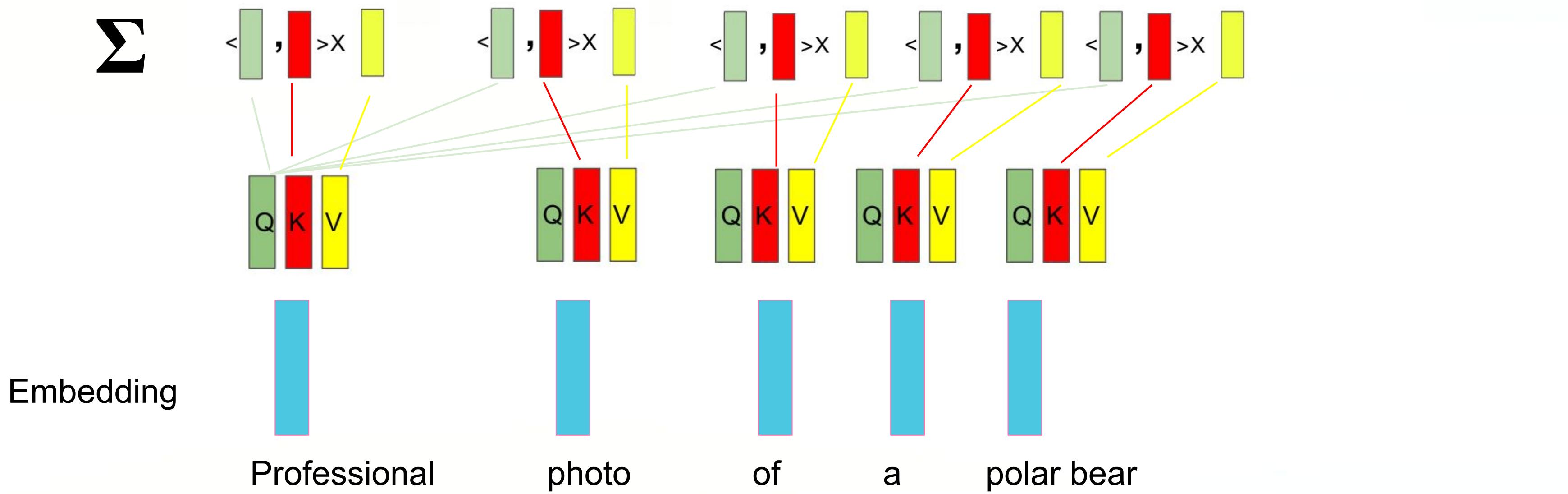
Glide: Text 2 Image Architecture



Transformers

Google, Dec 2017

Attention Layer



Transformers Training

“Deep learning is a subfield of **artificial** intelligence (AI) that focuses on training computer systems to learn and make intelligent decisions by analyzing **large** amounts of data. It is inspired by the structure and function of the **human** brain.”

Transformers Training

“Deep learning is a subfield of artificial intelligence (AI) that focuses on training computer systems to learn and make intelligent decisions by analyzing large amounts of data. It is inspired by the structure and function of the human brain.”

Positional Embedding

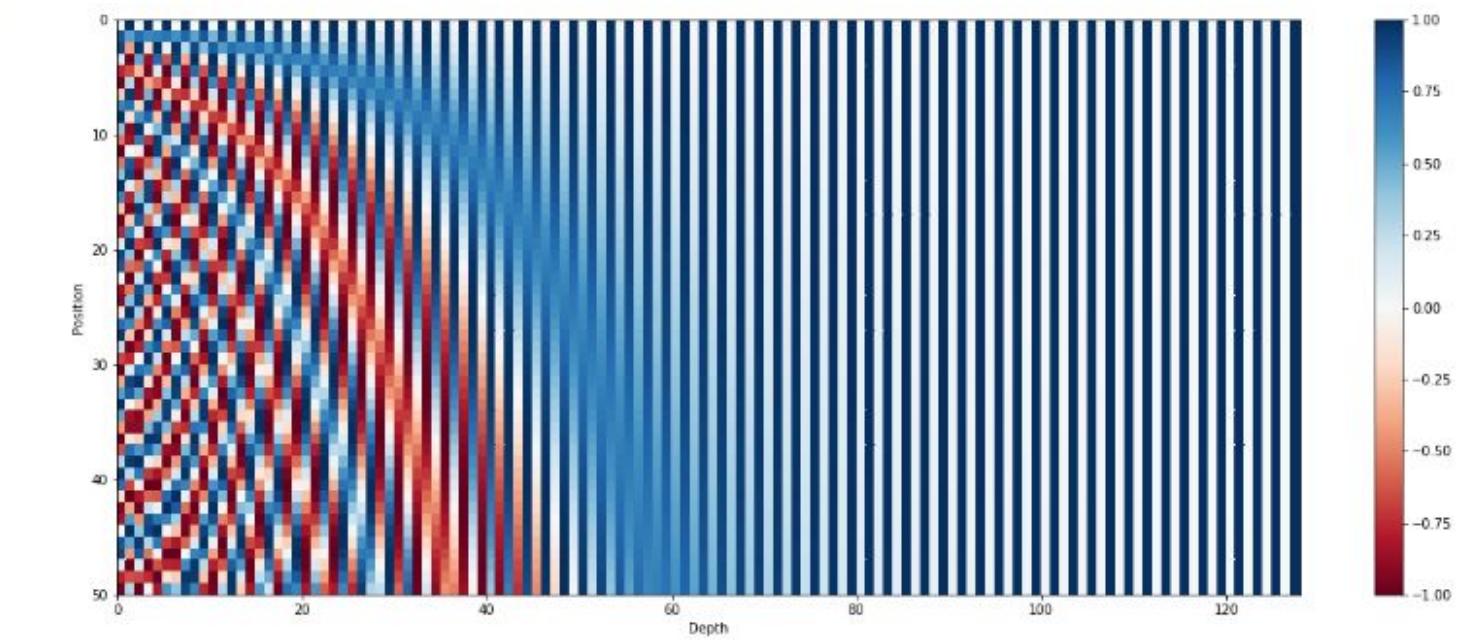
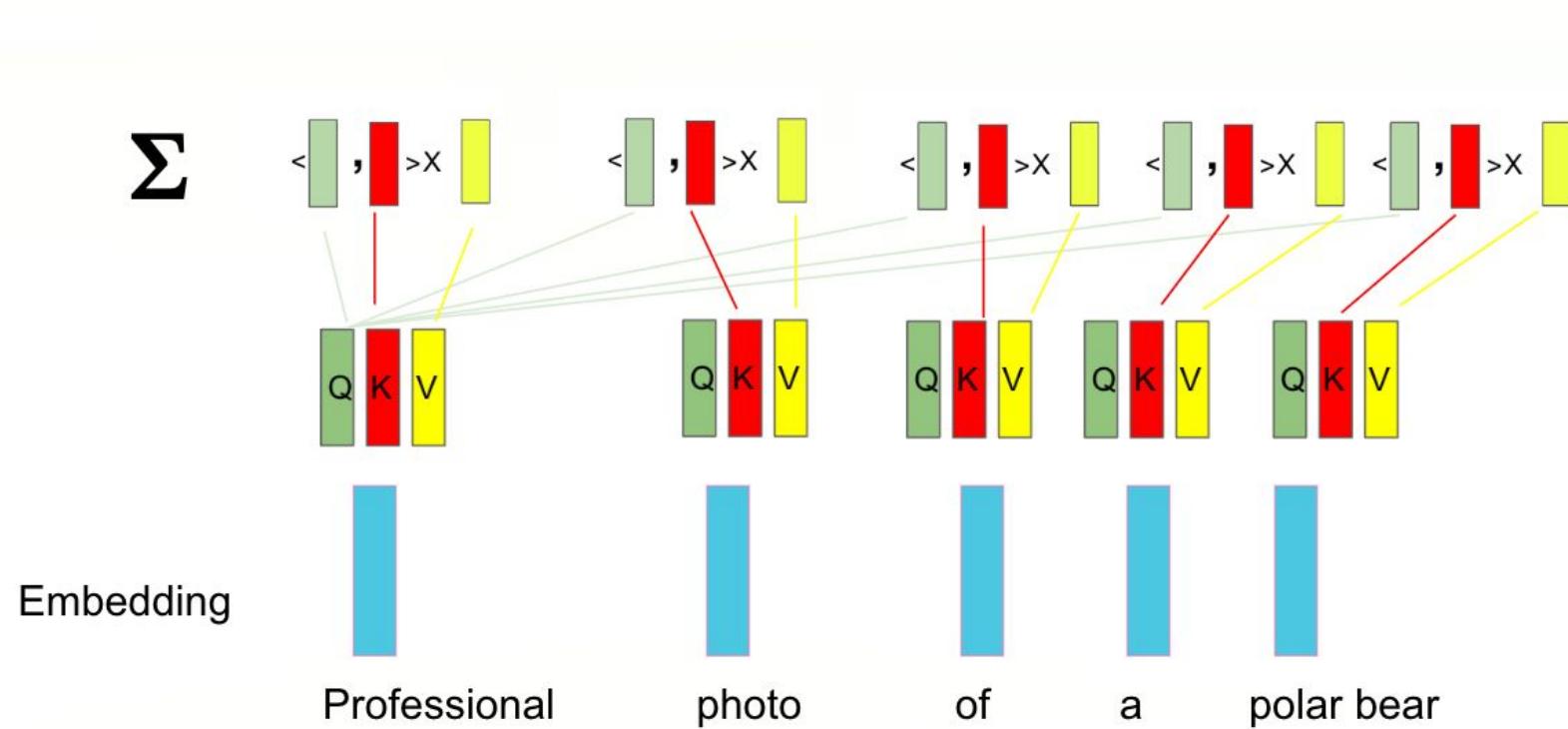
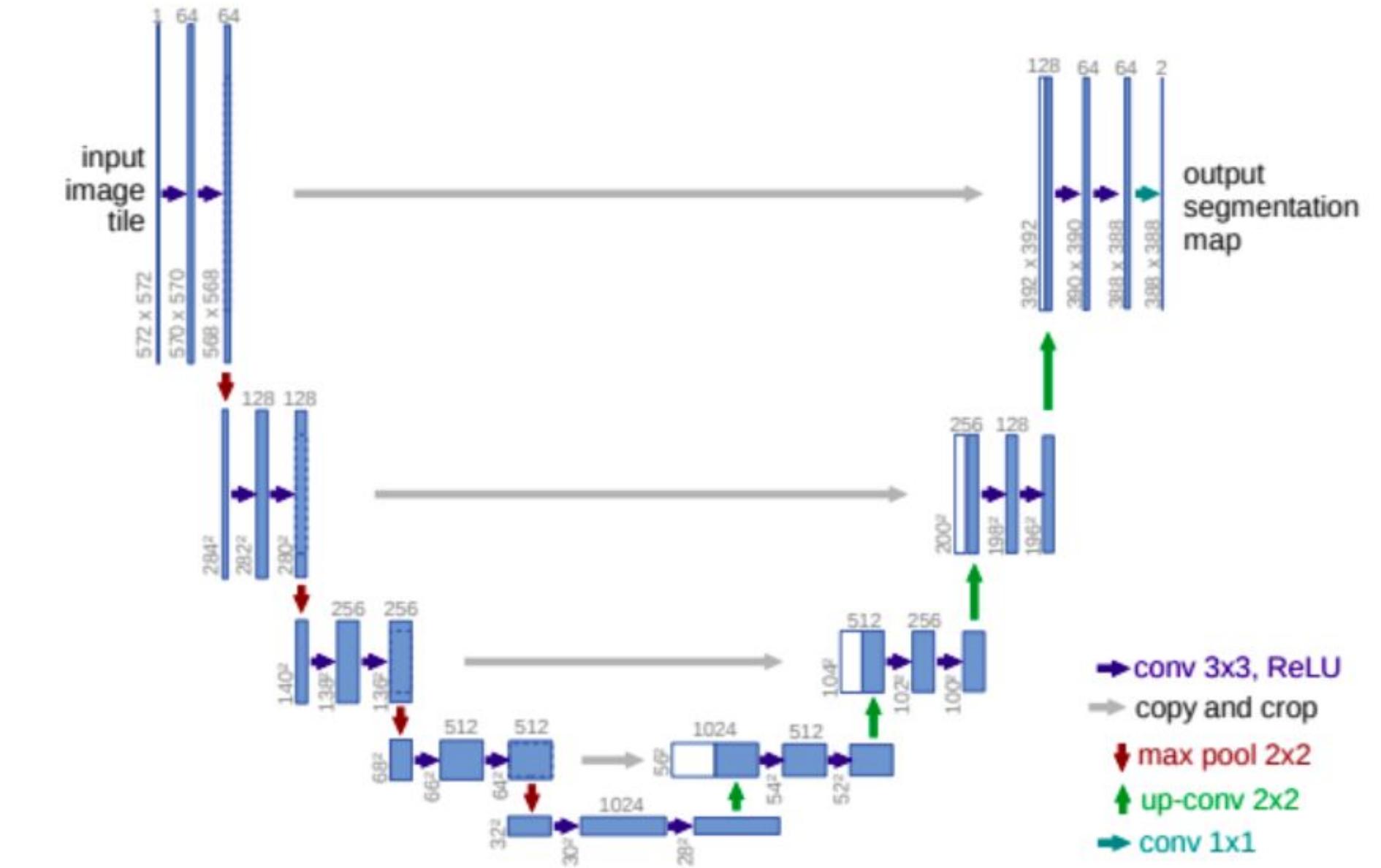


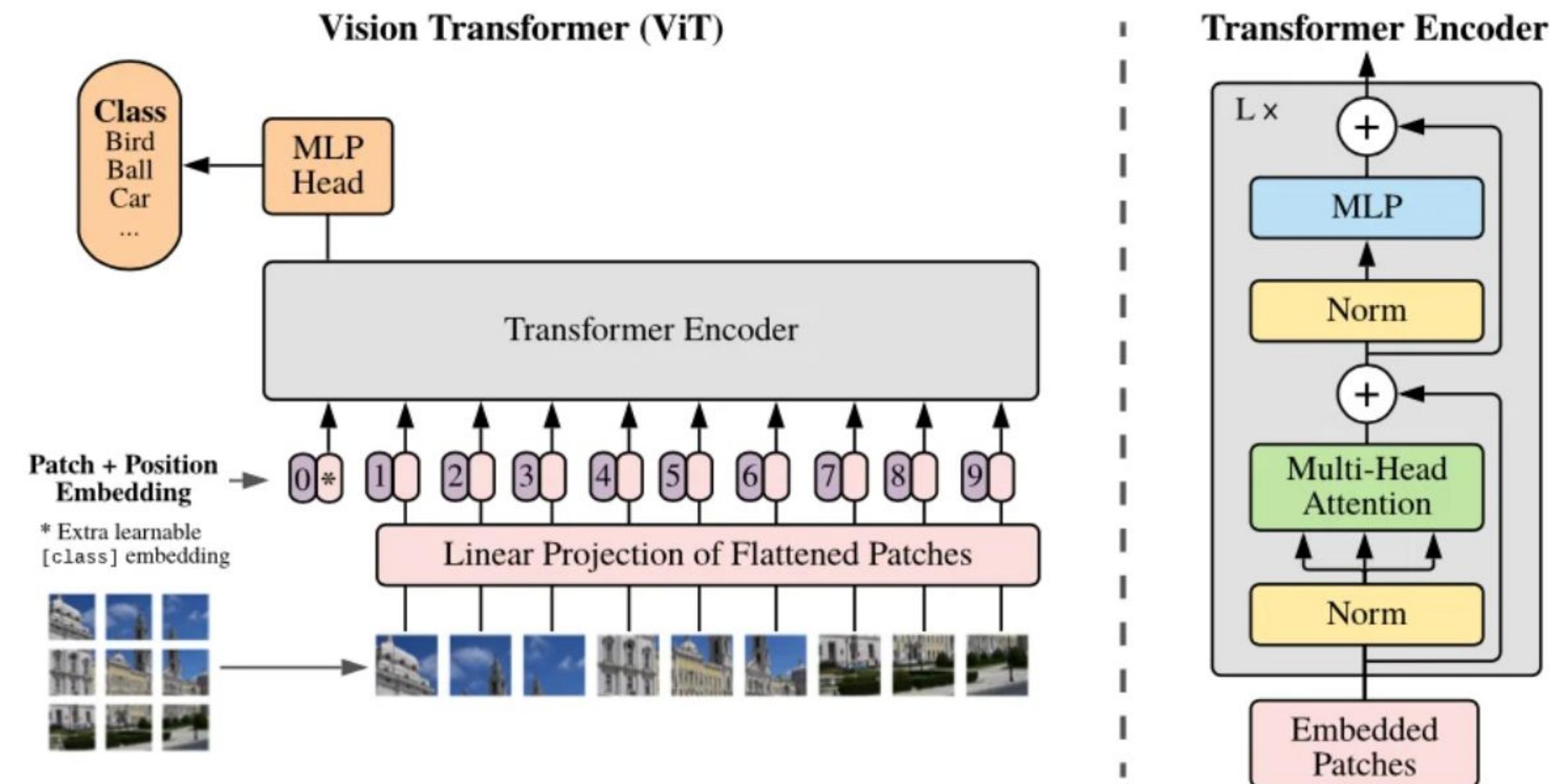
Figure 2 - The 128-dimensional positonal encoding for a sentence with the maximum lenght of 50. Each row represents the embedding vector \vec{p}_t

Global Context in the Vision Domain



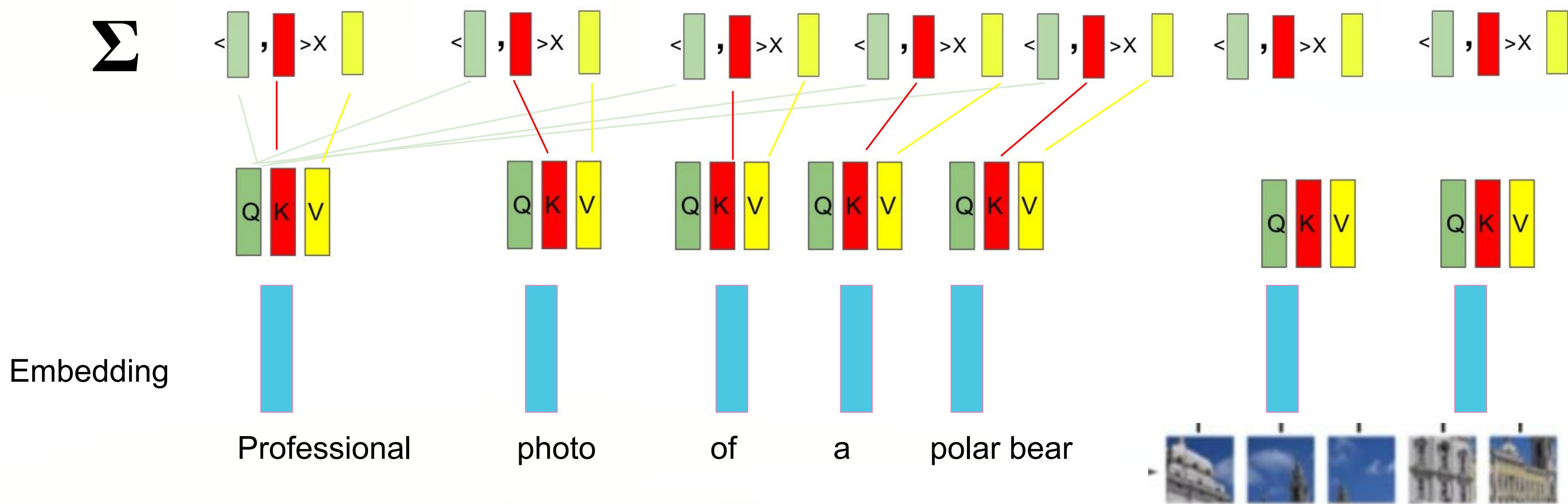
Global context

Visual Transformers

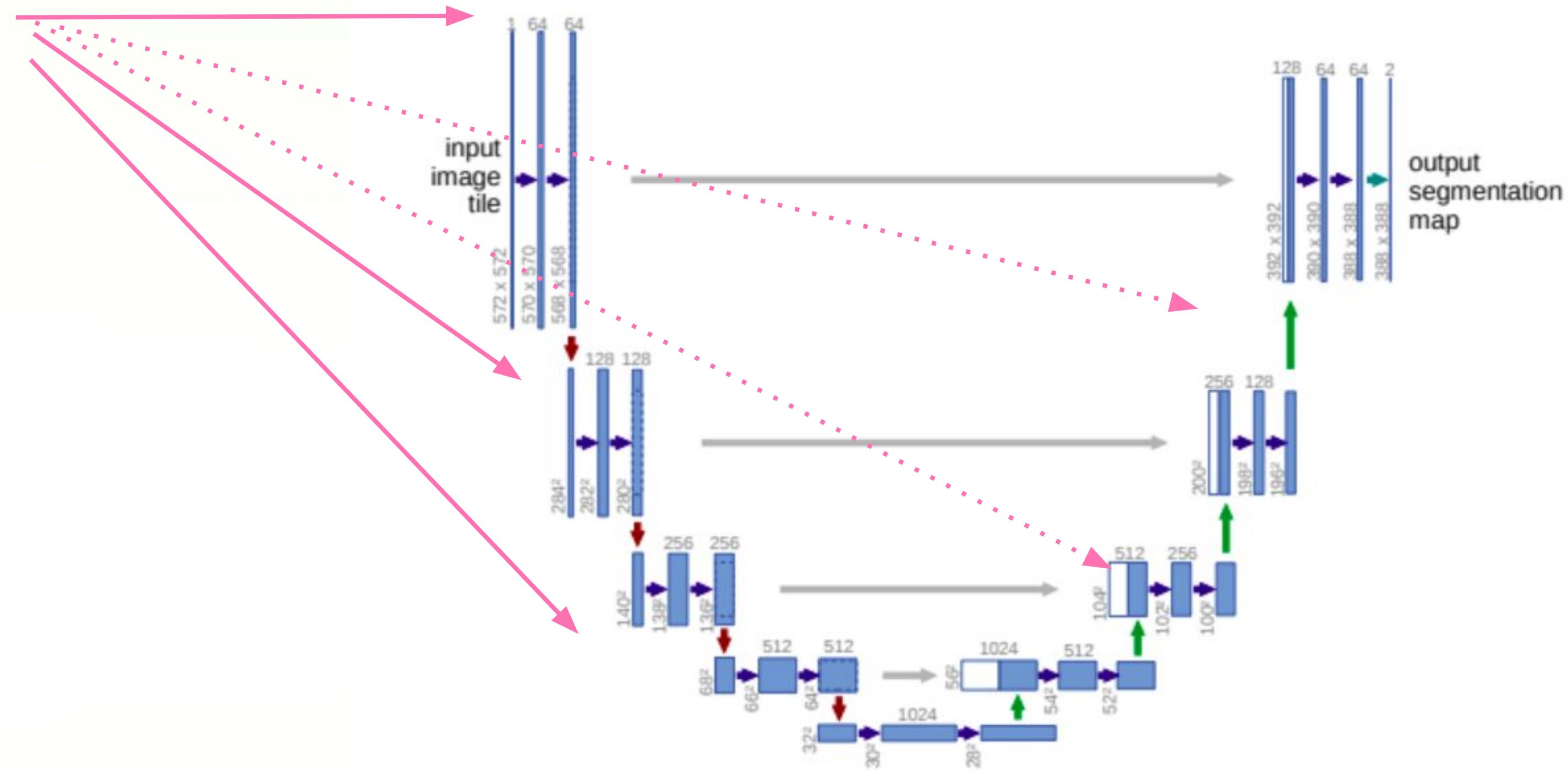


[An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale](#)

Cross Attention

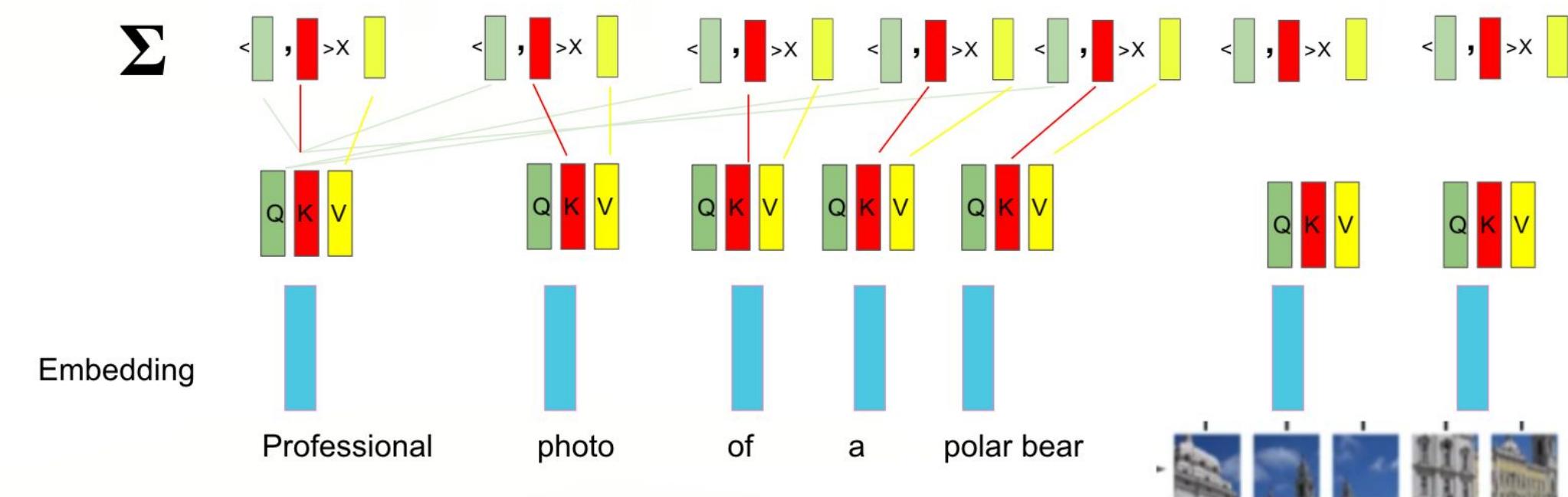
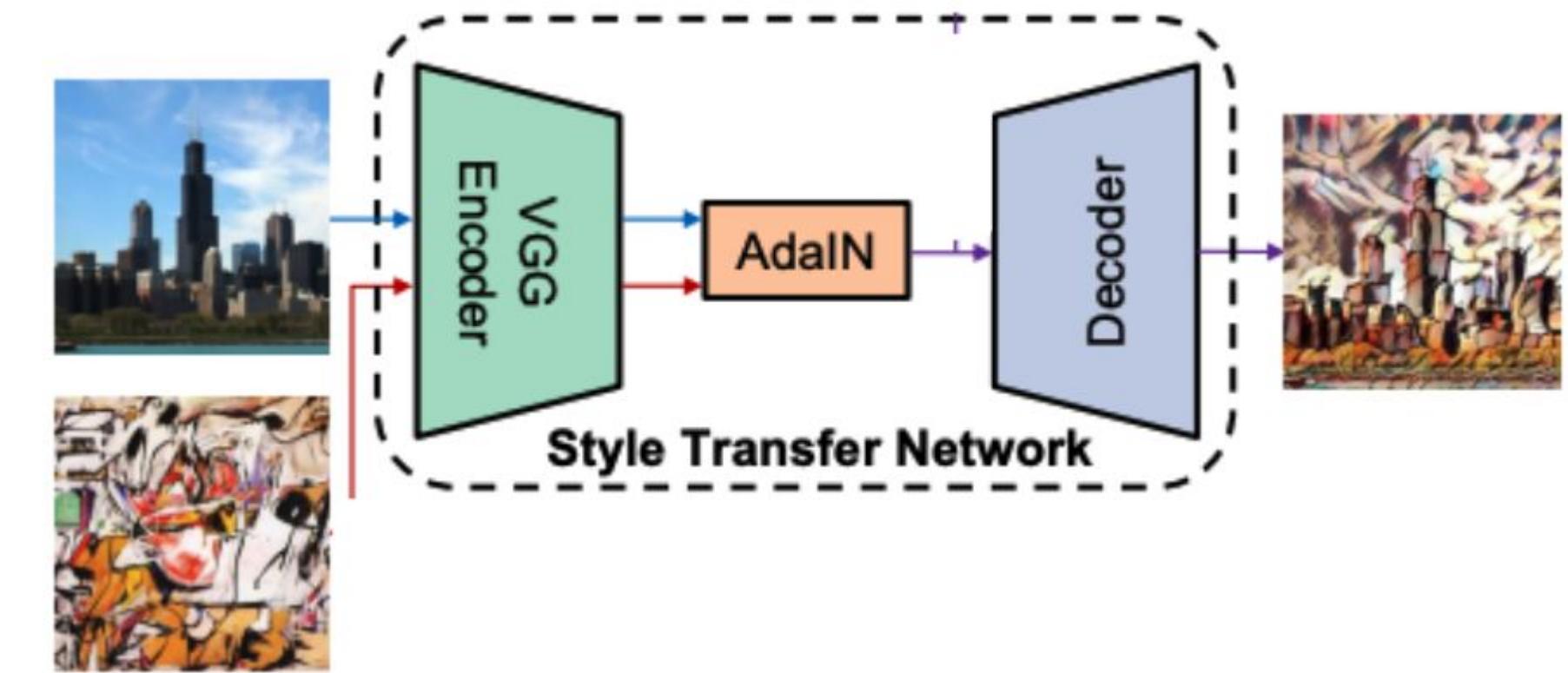


Adding Textual Input into the Diffusion Model



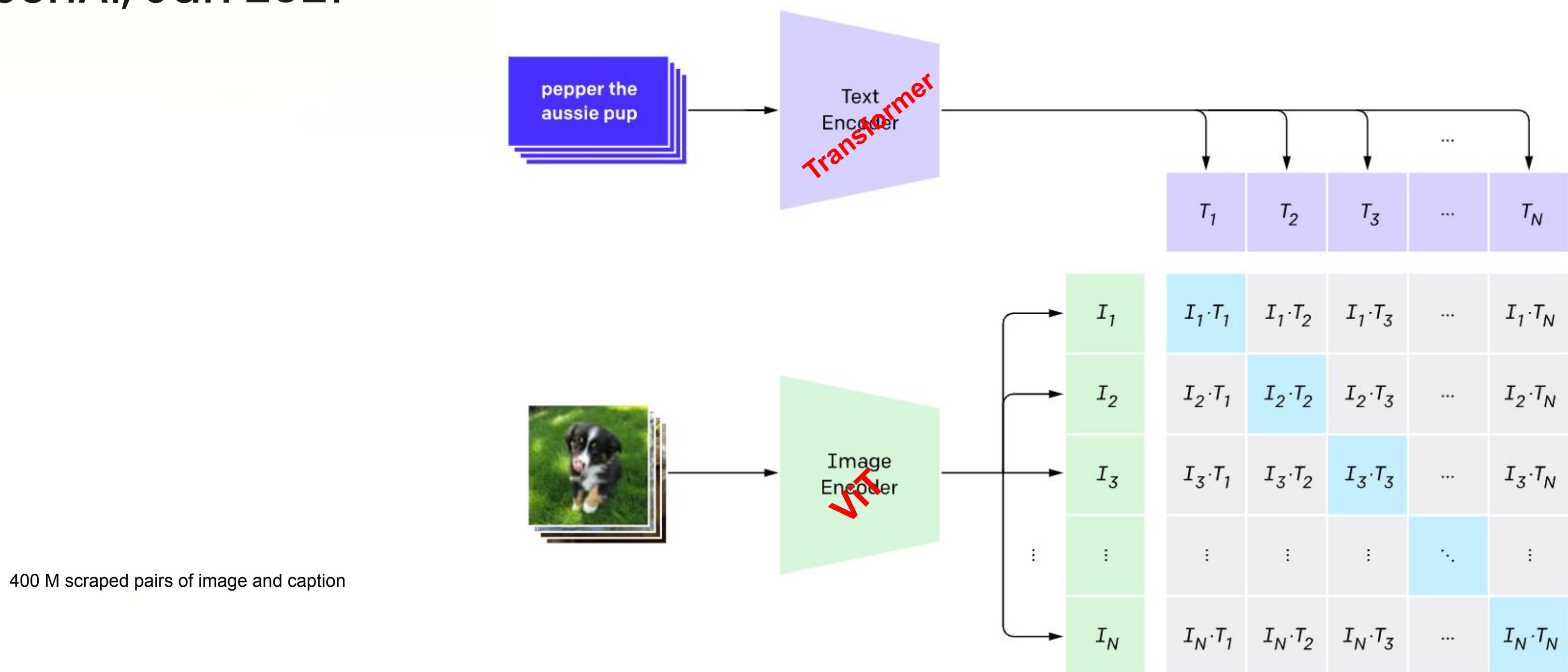
Multimodal Layers

1. Style/Text injection via channel statistics
2. Style/Text injection via Cross attention

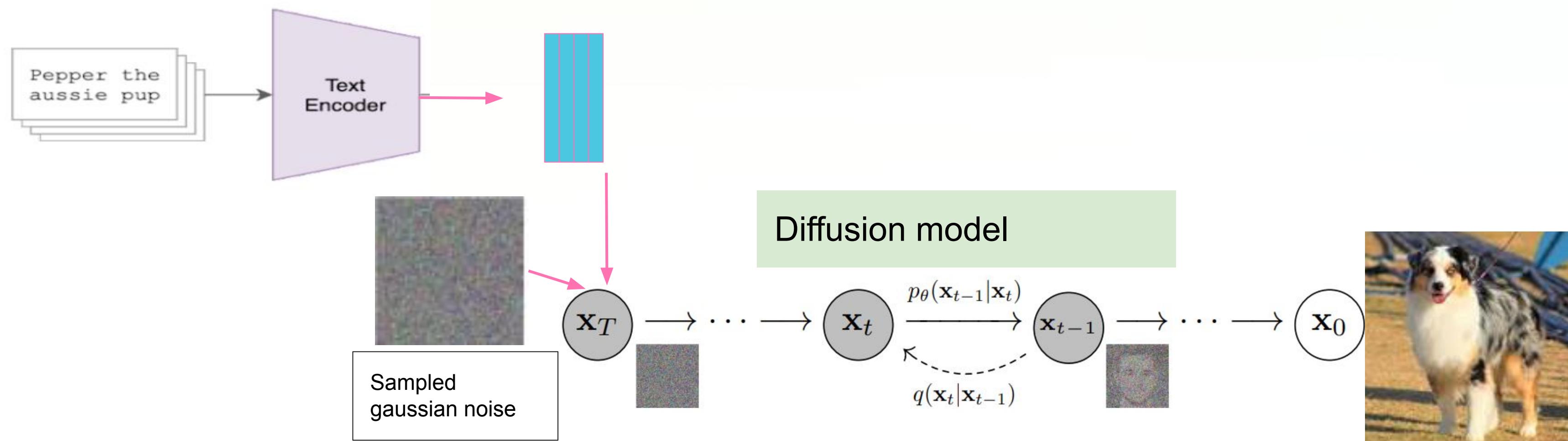


CLIP- Text and Image encoder

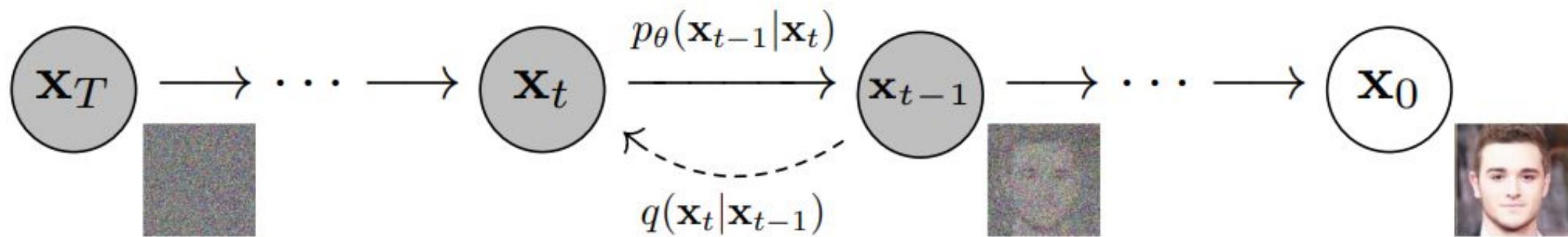
OpenAI, Jan 2021



Glide Text 2 Image Architecture



Class Free Guidance - extrapolation toward the textual concept



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

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

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

System 1



Fast, intuitive and
emotional

System 2



Slow, conscious
and effortful

What Caused 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. **New architectures**

Multi domain, layers (Bert, GPT3, StyleGAN, ViT)

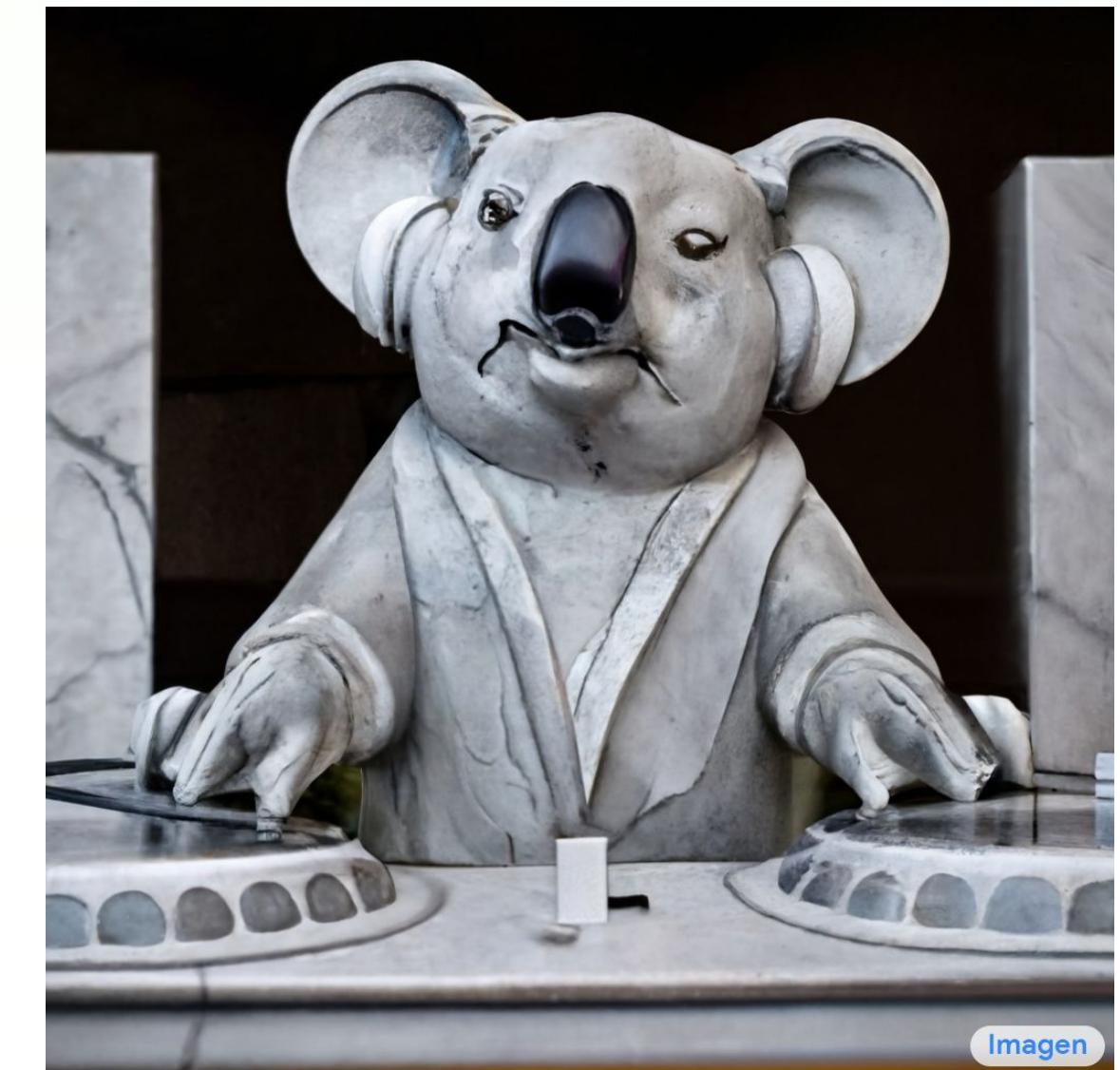
Imagen



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.

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 Architecture

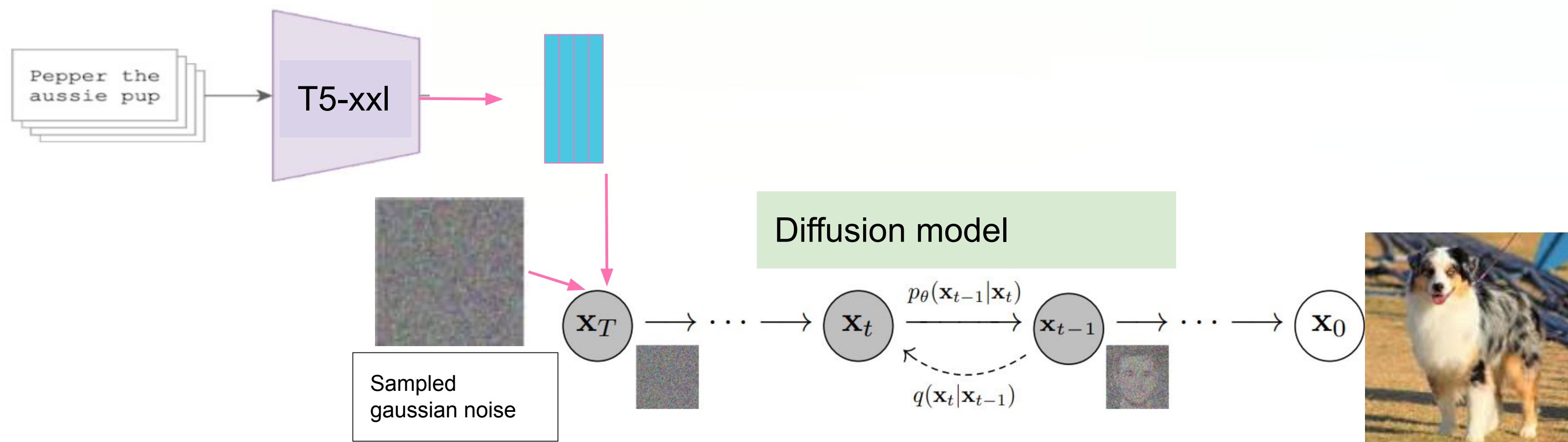
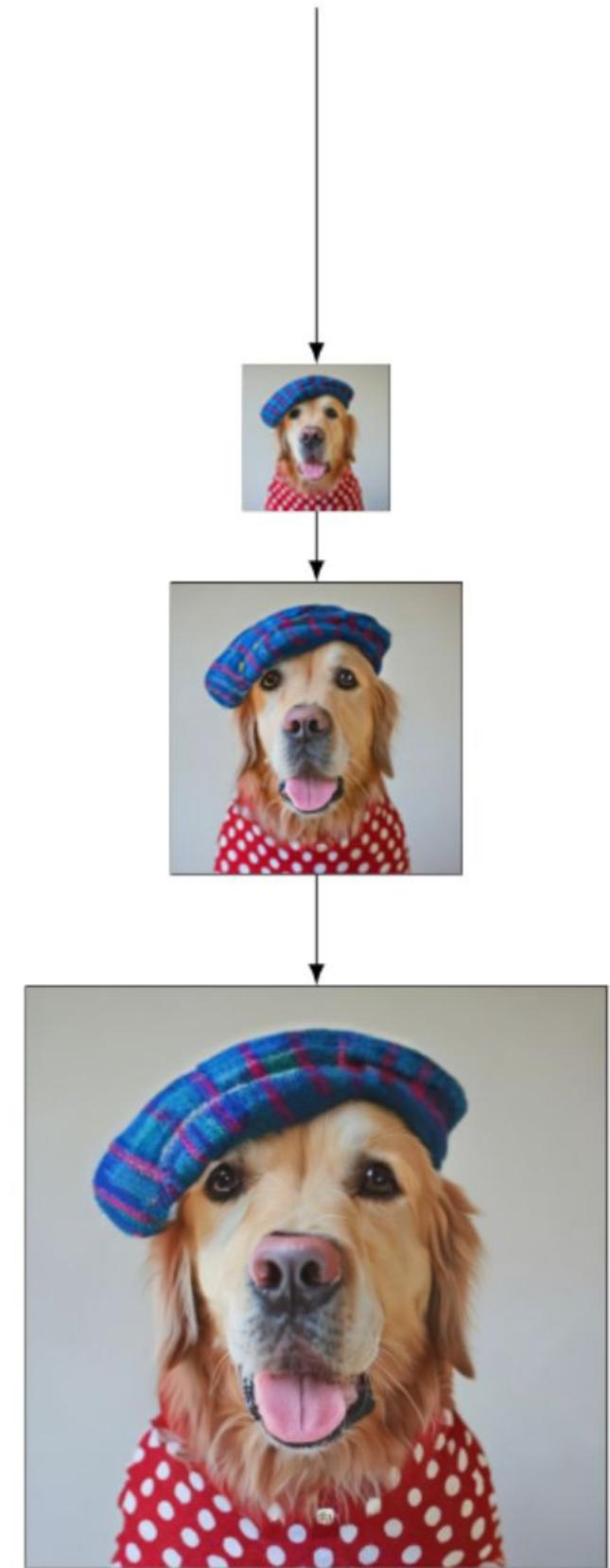


Imagen: Super Resolution models

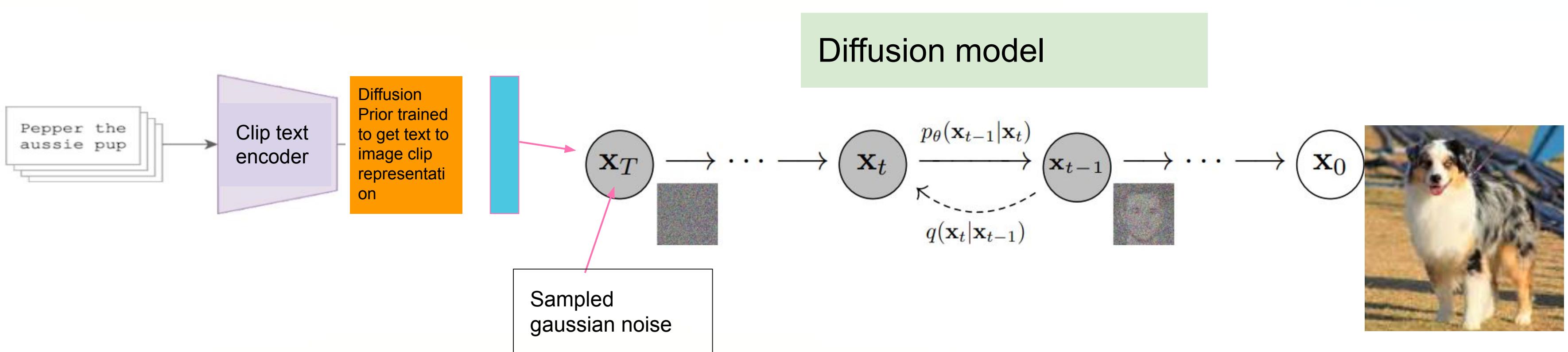
“A Golden Retriever dog wearing a blue checkered beret and red dotted turtleneck.”



Dalle-2

openAI april 2022

Unclip – from clip image vector back into image





Delighted to announce the public open source release of [#StableDiffusion!](#)

Please see our release post and retweet!
[stability.ai/blog/stable-di...](https://stability.ai/blog/stable-diffusion-public-release/)

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

[תרגום את הציג](#)

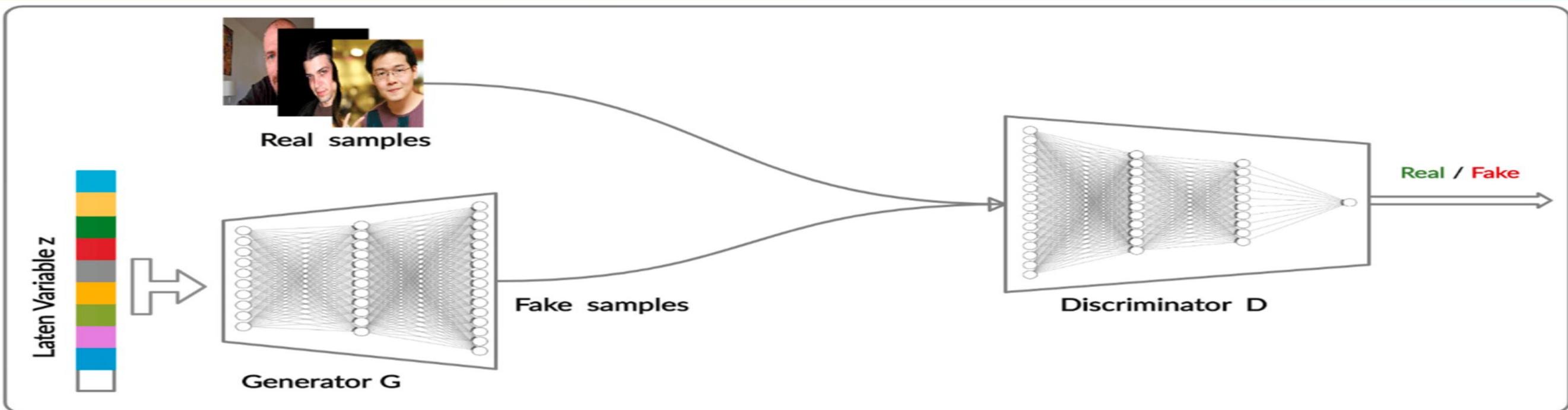
stability.ai

Stable Diffusion Public Release — Stability.Ai

We are delighted to announce the public release of Stable Diffusion and the launch of DreamStudio Lite.



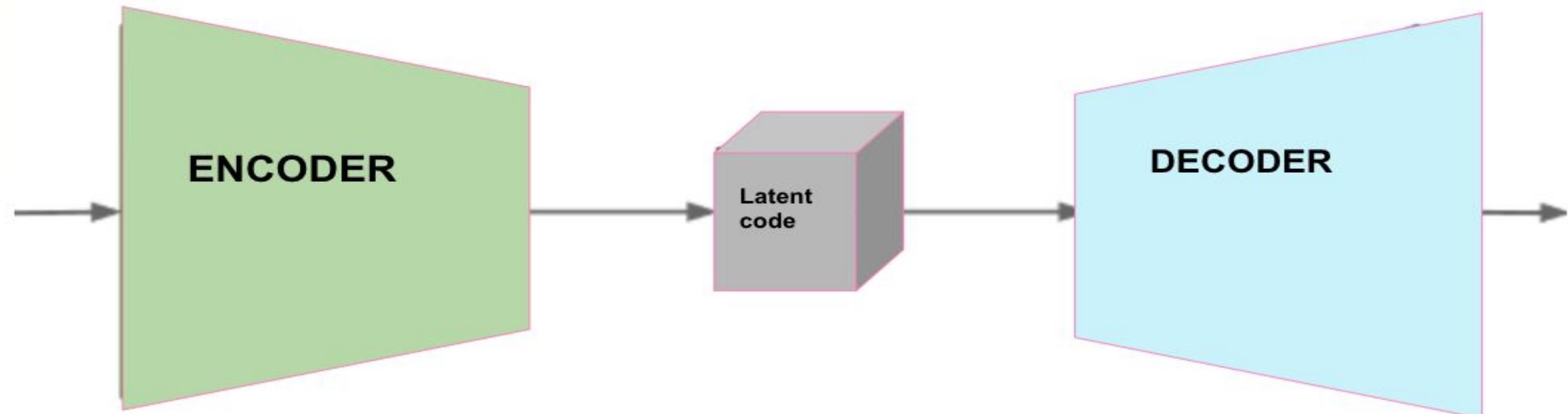
GAN: Adversarial training



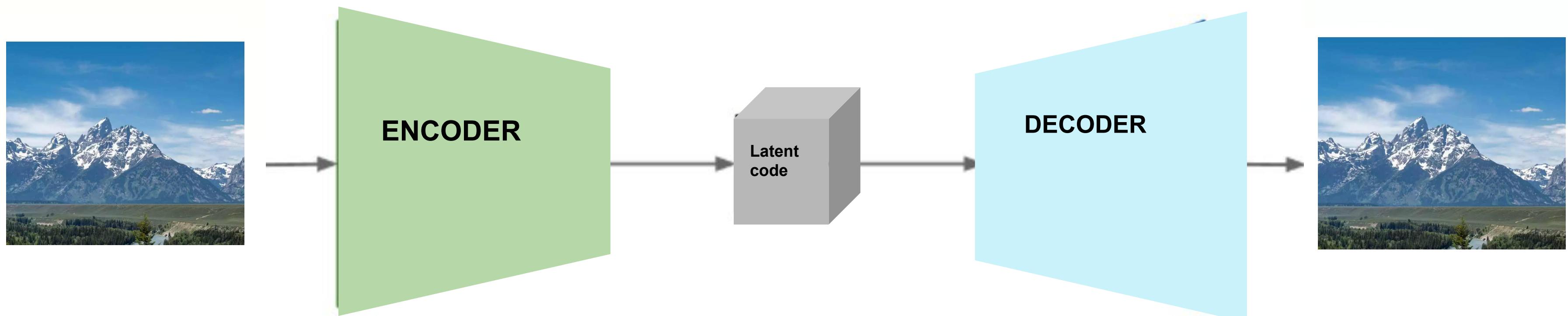
Diffusion models:
Gradually add Gaussian noise and then reverse



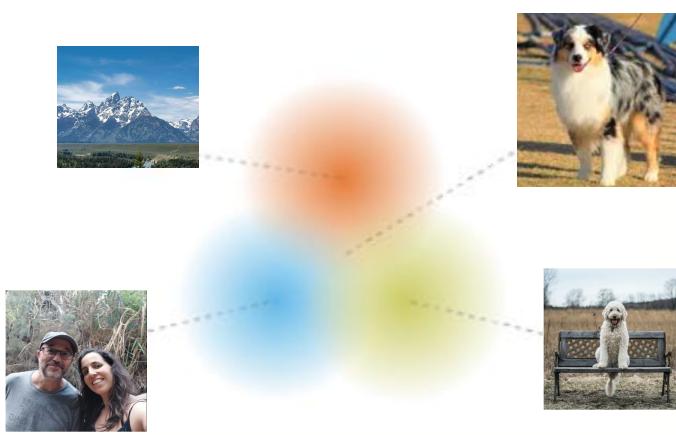
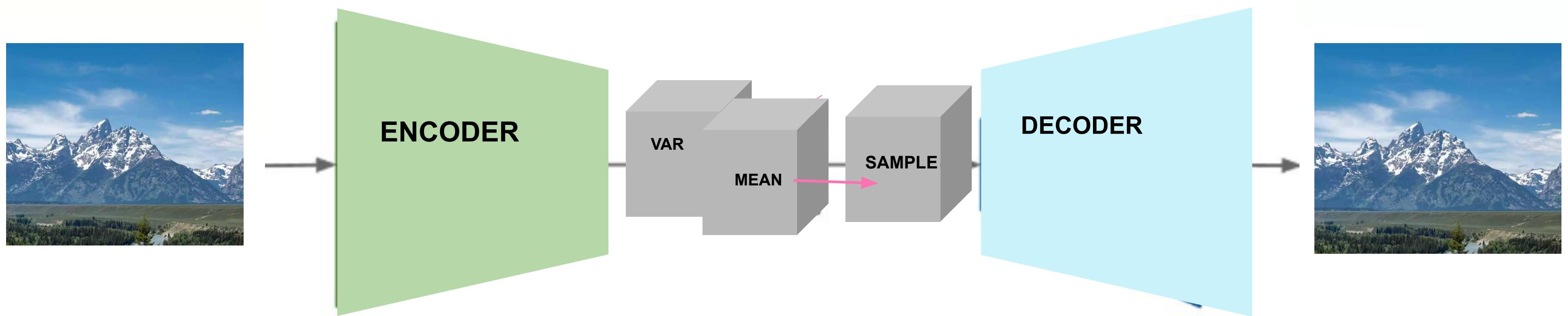
Variational Autoencoder:



AutoEncoder



Variational AutoEncoder



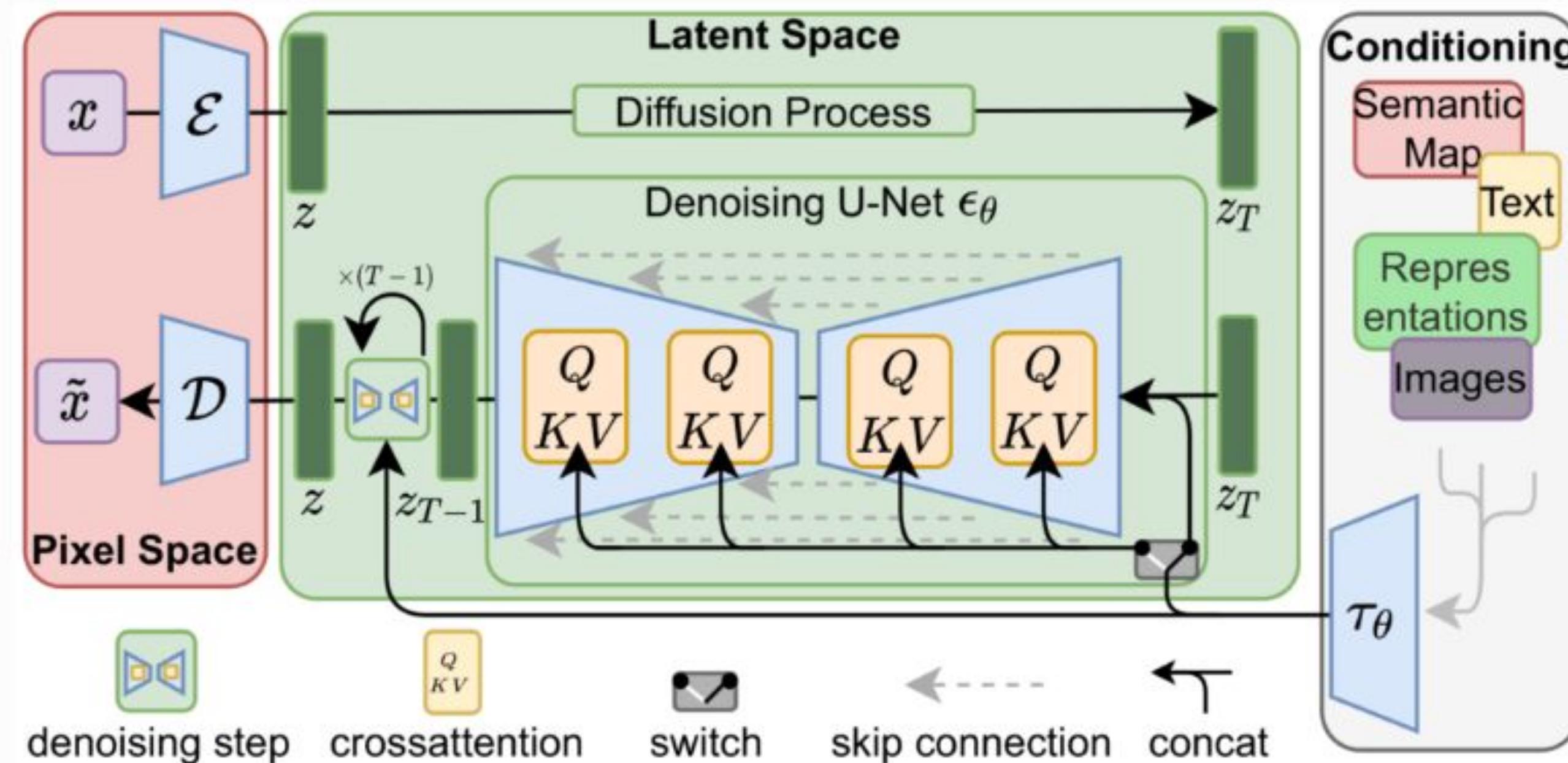
Taming Transformers for High-Resolution Image Synthesis

Patrick Esser*, Robin Rombach*, Björn Ommer

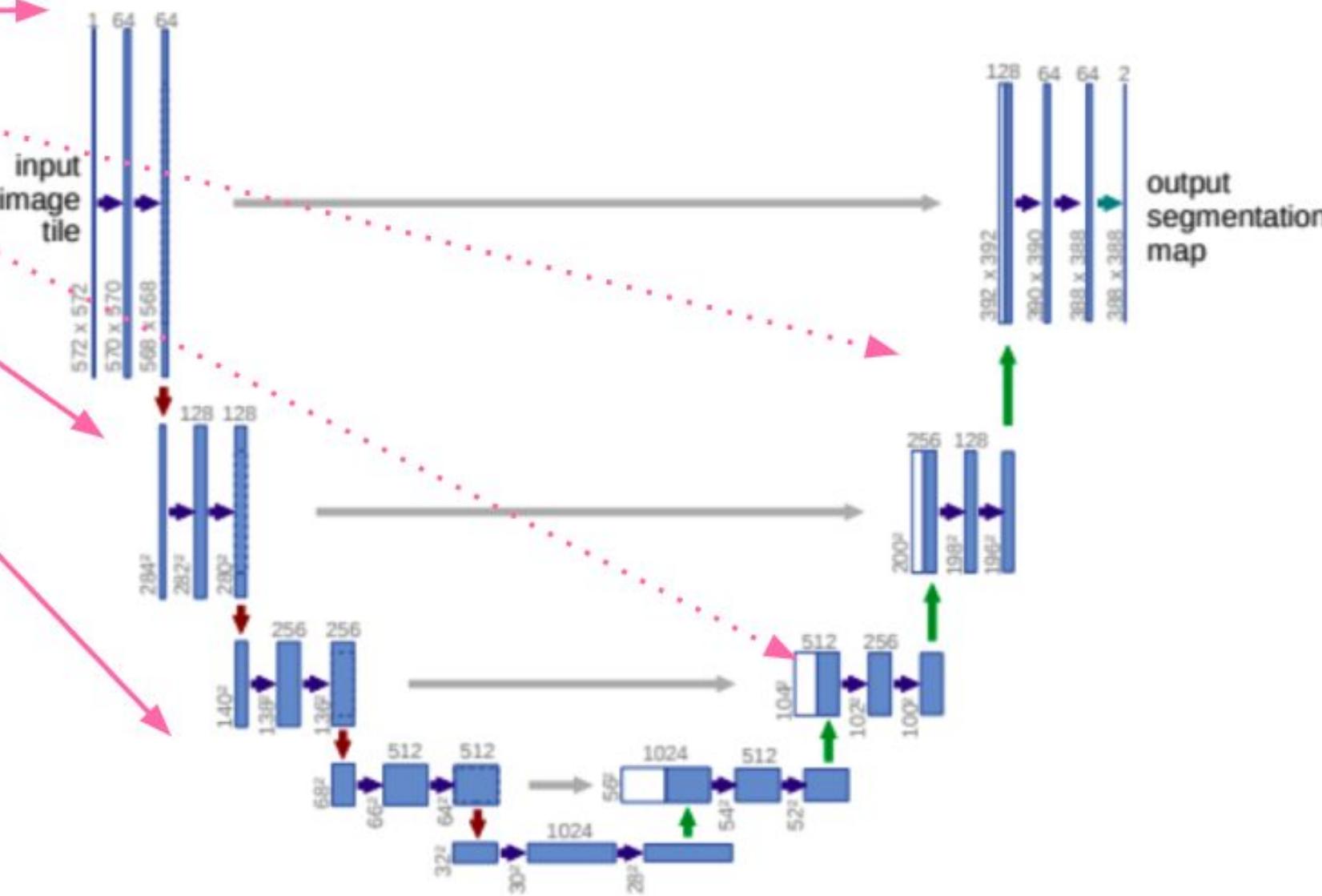


High-Resolution Image Synthesis with Latent Diffusion Models

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer



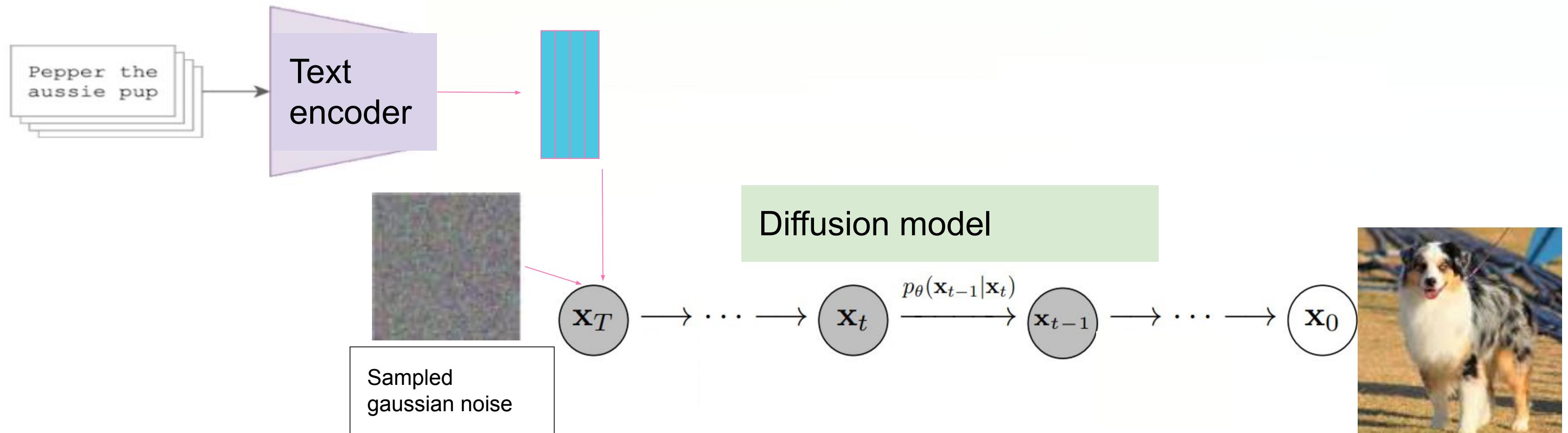
64x64x3 image



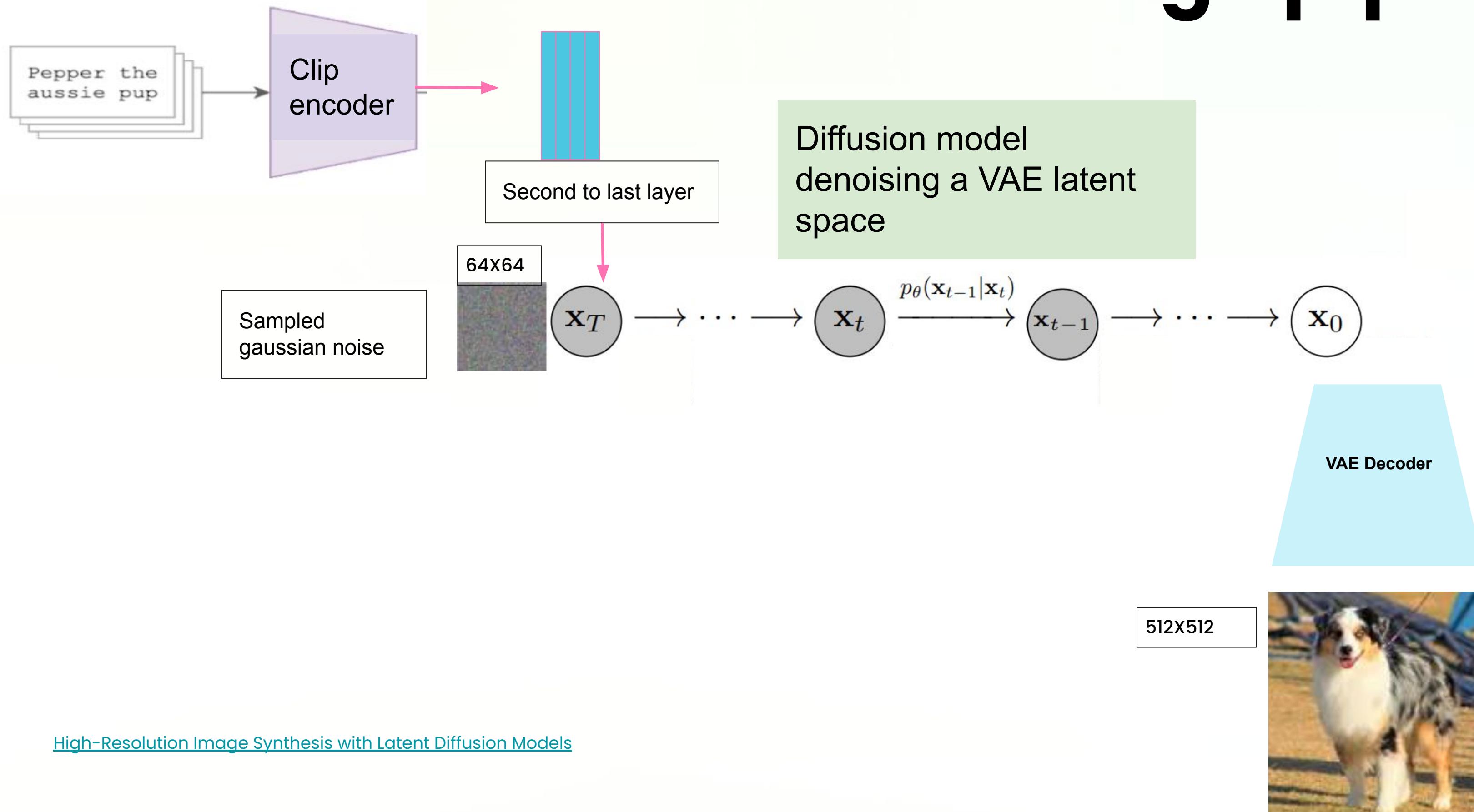
| 64x64x4 latent representation



Original Text to Image Pipe



Stable Diffusion Text to Image pipe



Part 3 - Controlling Text V=Condition Diffusion Models

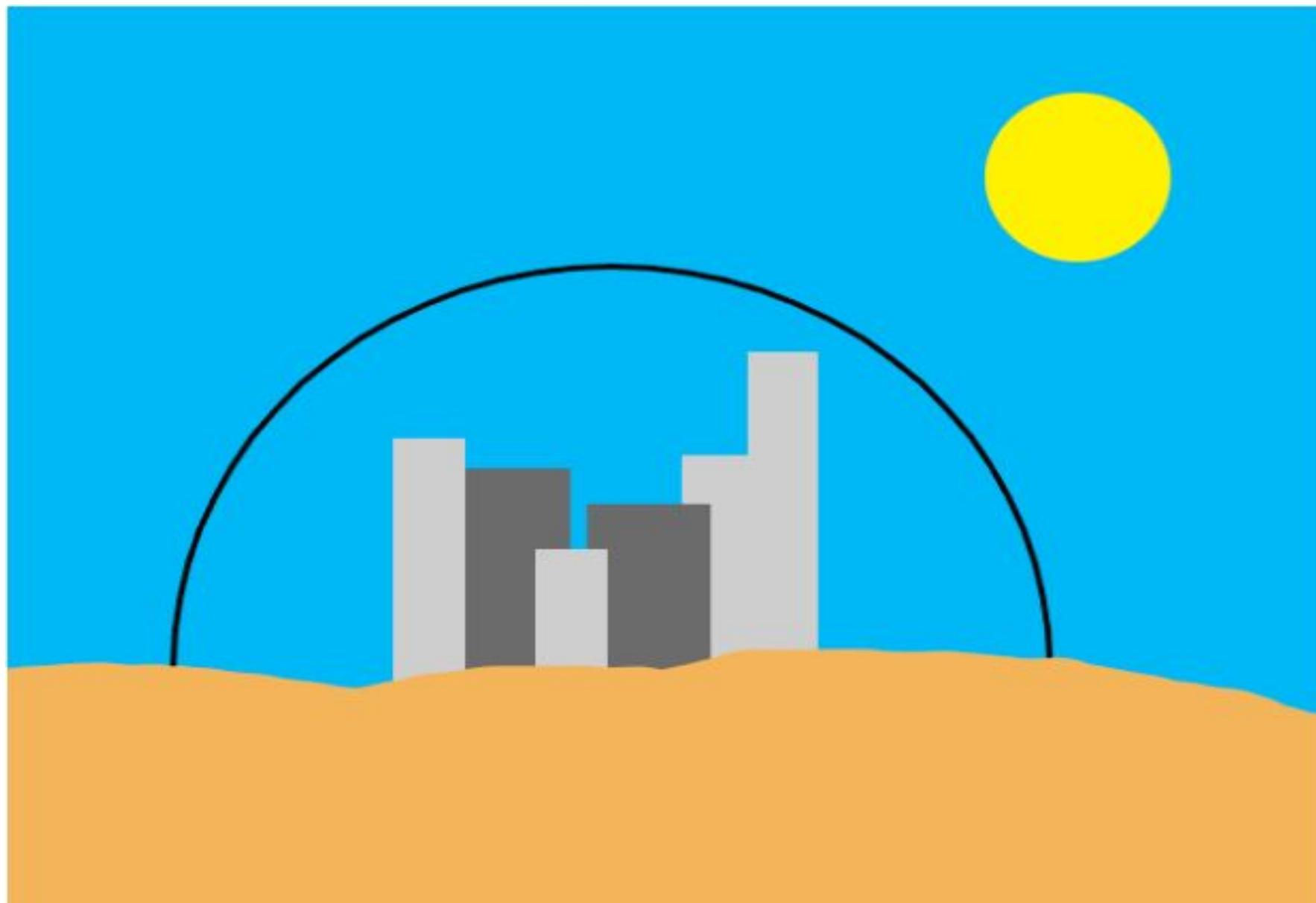
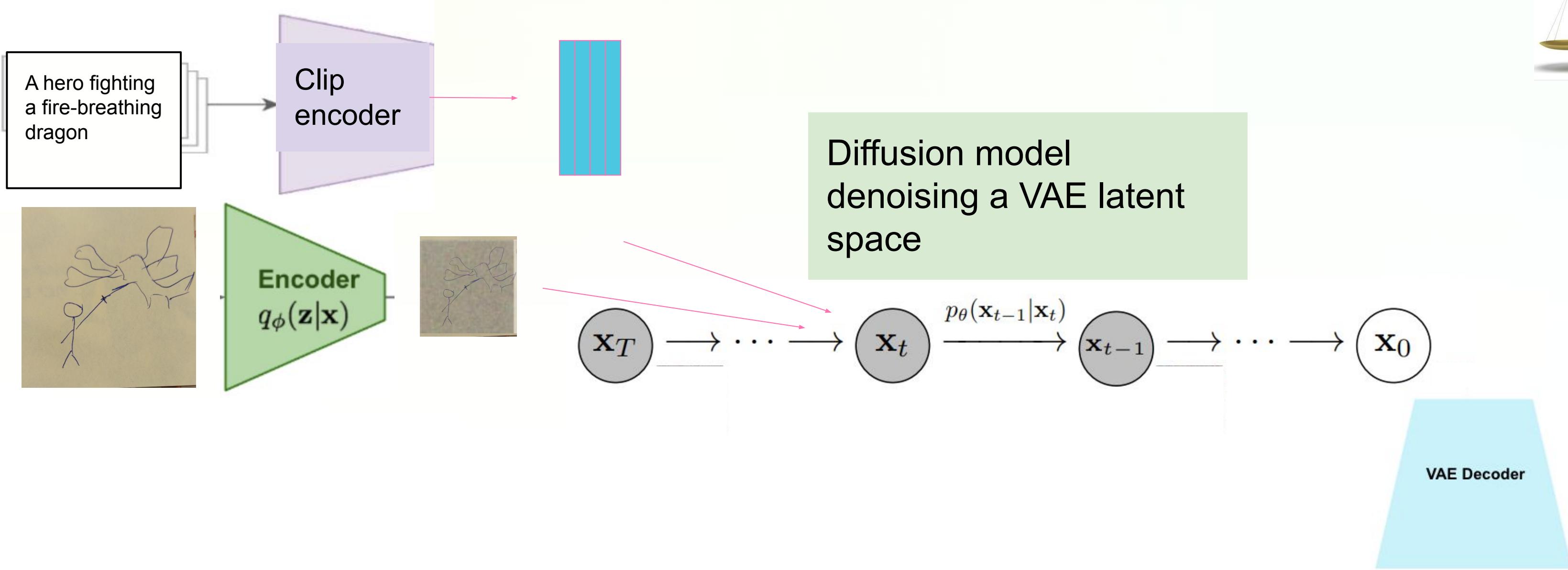
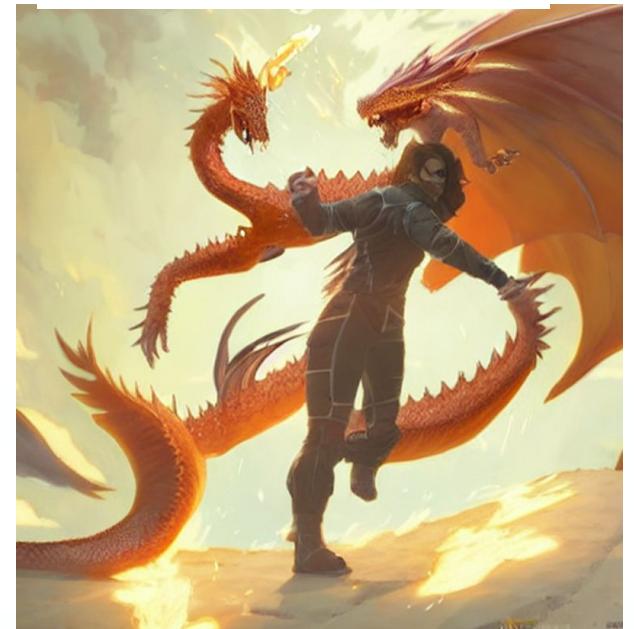


Image to Image pipe

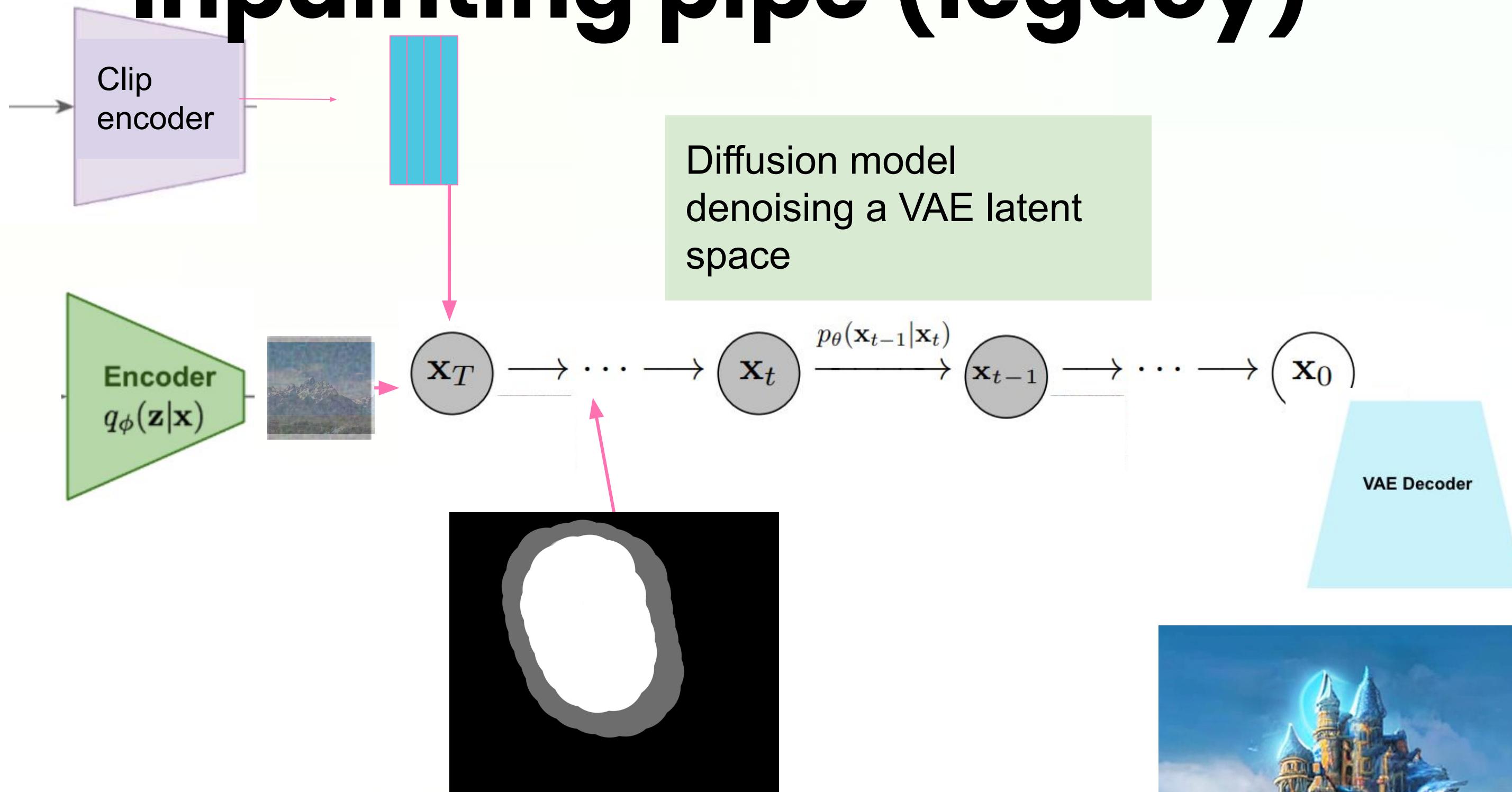


[High-Resolution Image Synthesis with Latent Diffusion Models](#)



Inpainting pipe (legacy)

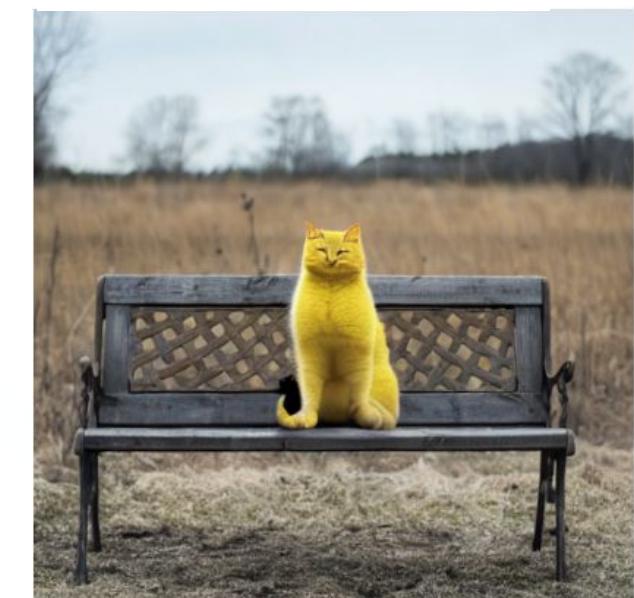
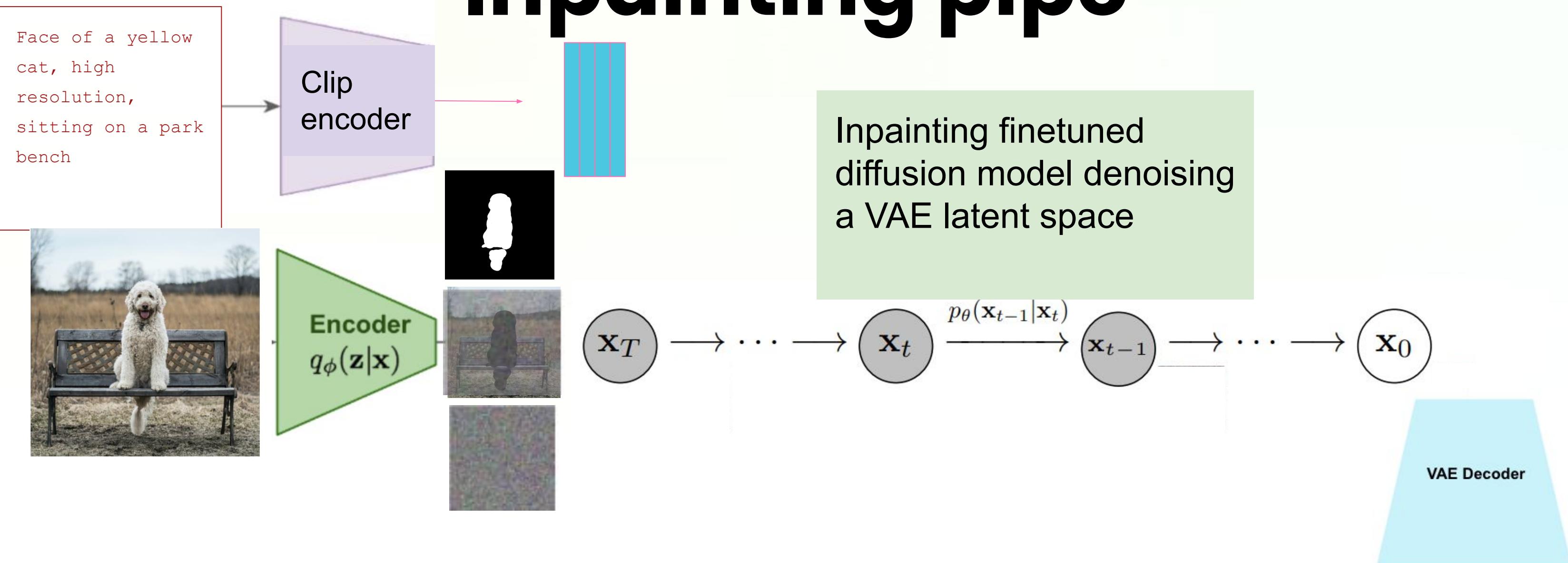
A magic castle on the mountains

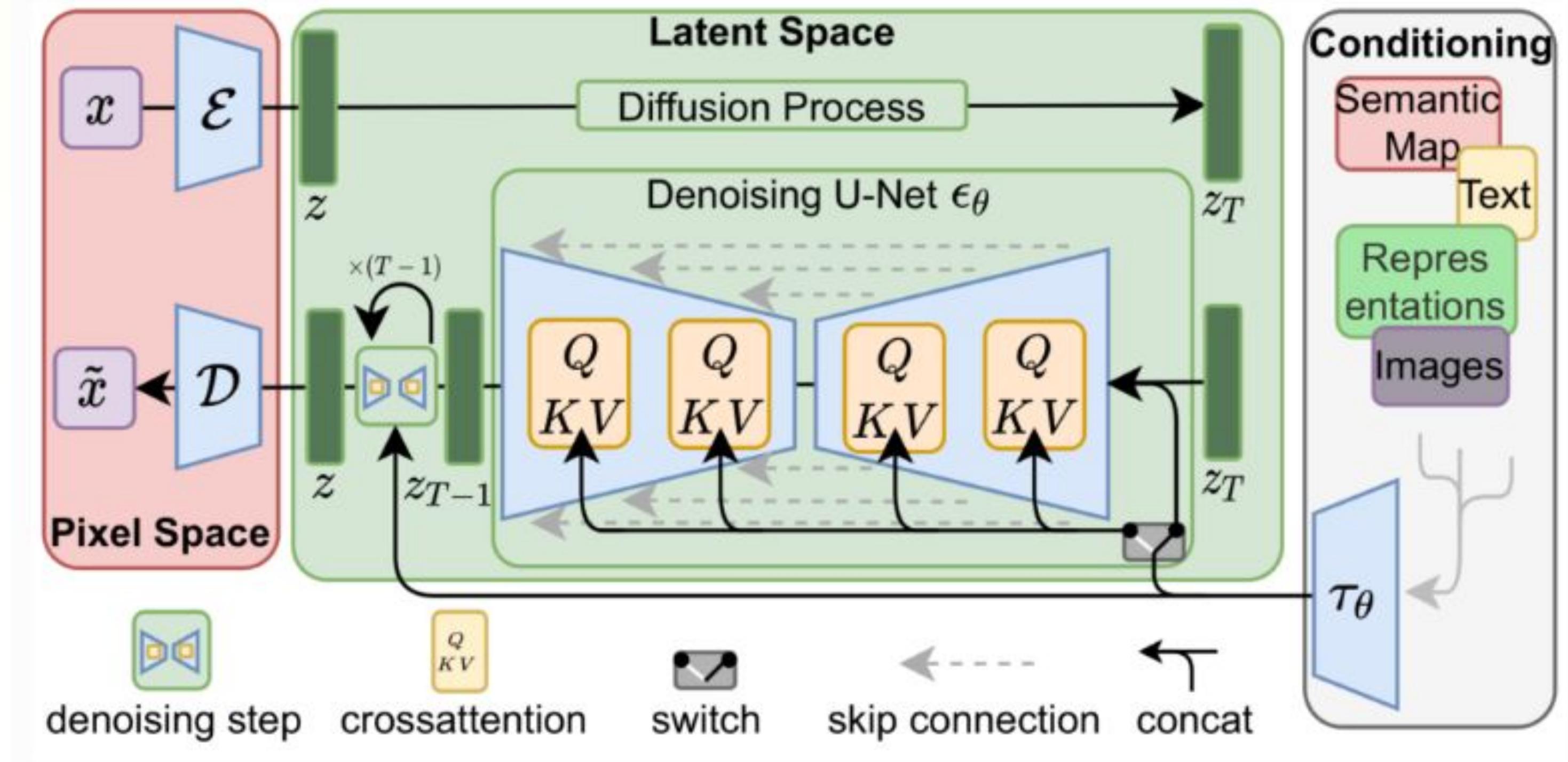


[High-Resolution Image Synthesis with Latent Diffusion Models](#)



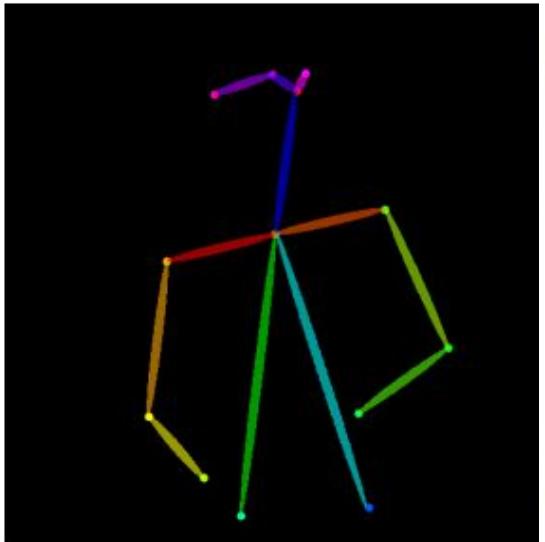
Inpainting pipe





ControlNet Enhance pretrained image diffusion models with task-specific conditions

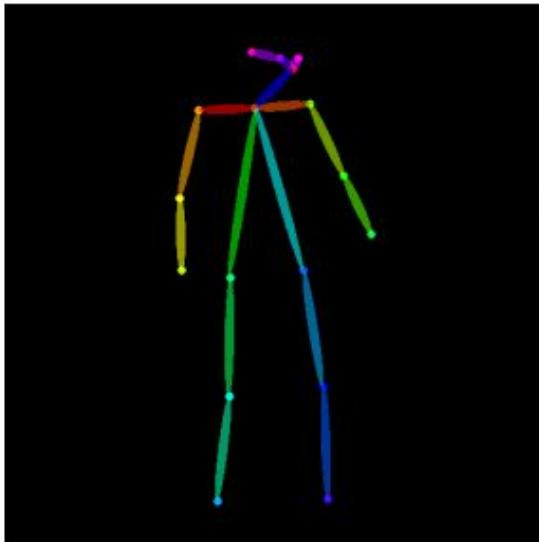
Input (openpose)



User Prompt



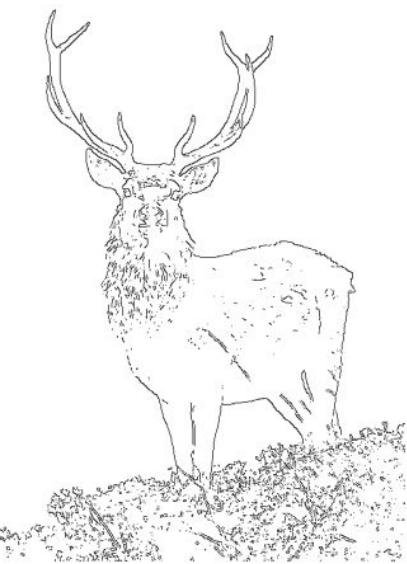
“chef in the kitchen”



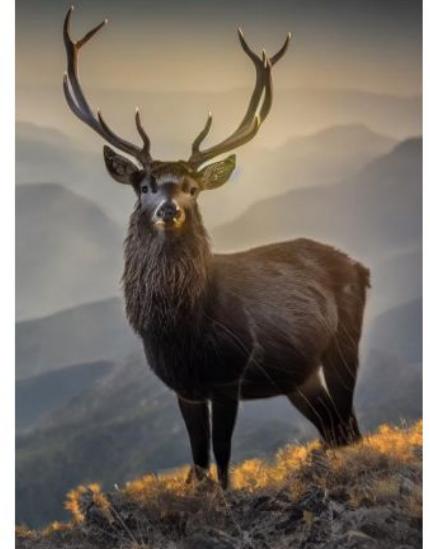
“astronaut”



Source image
(for canny edge detection)

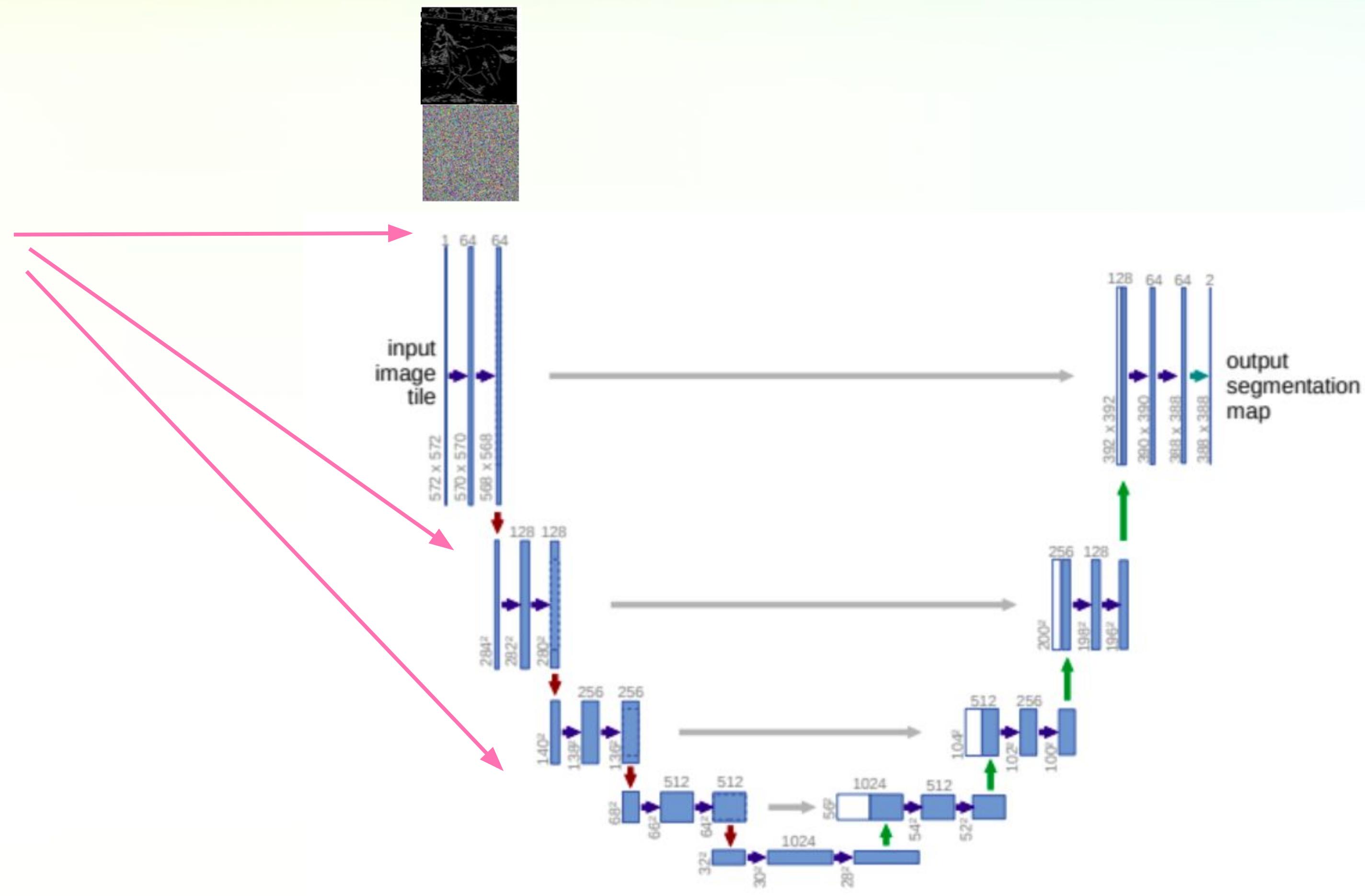


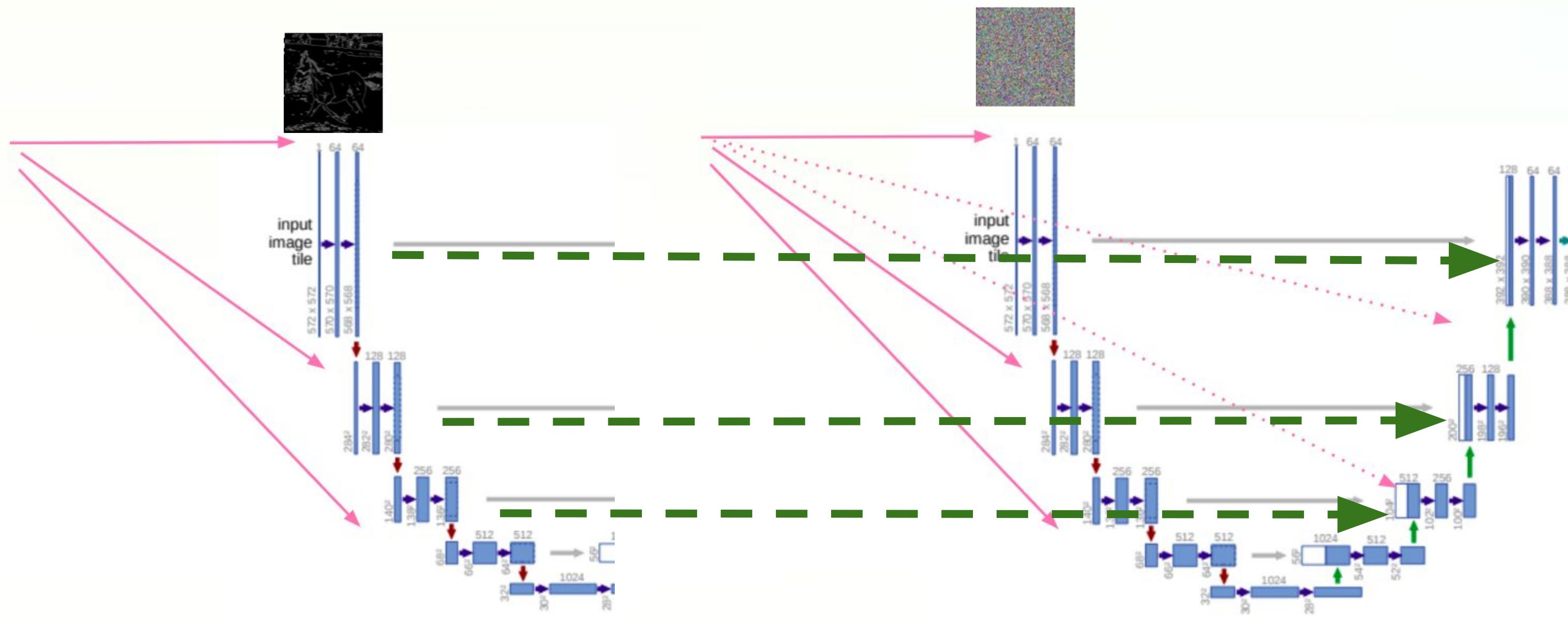
Canny edge (input)



Generated images (output)





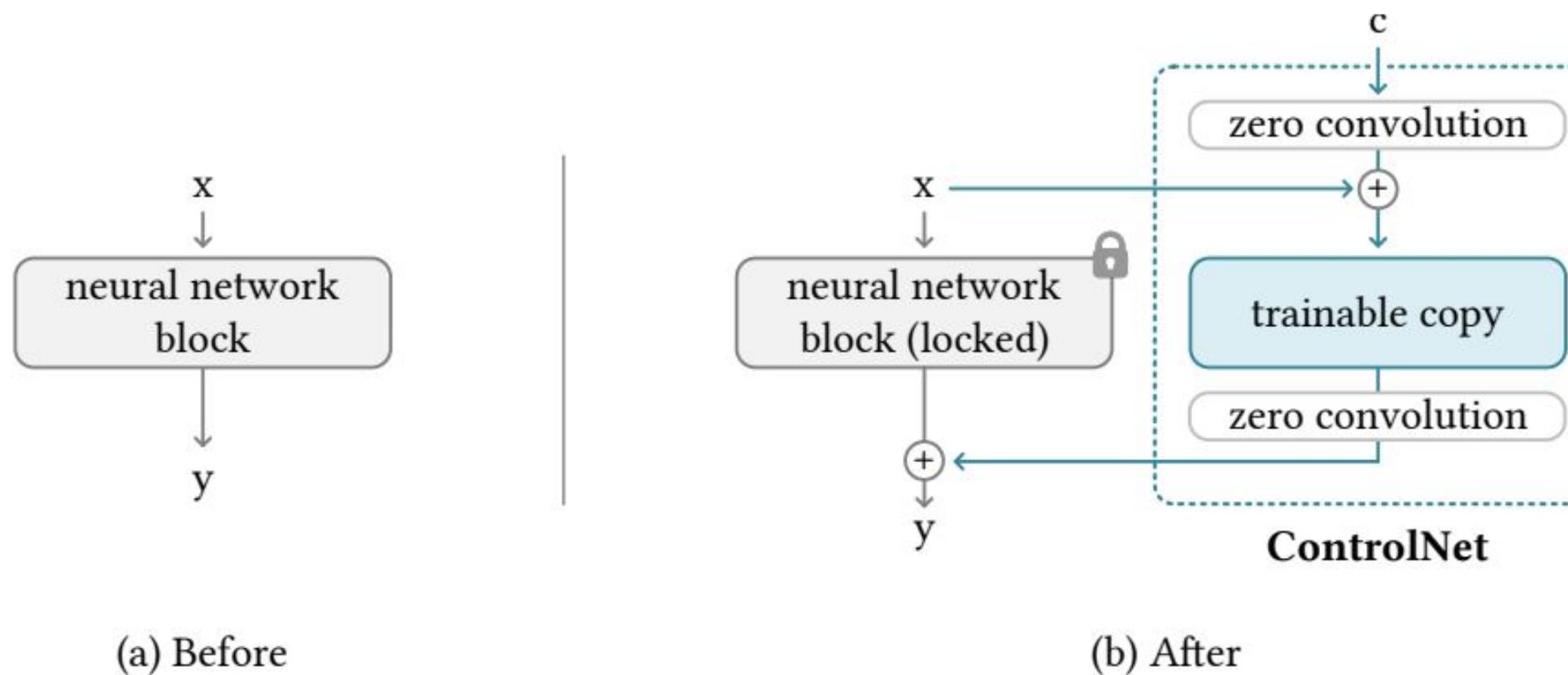


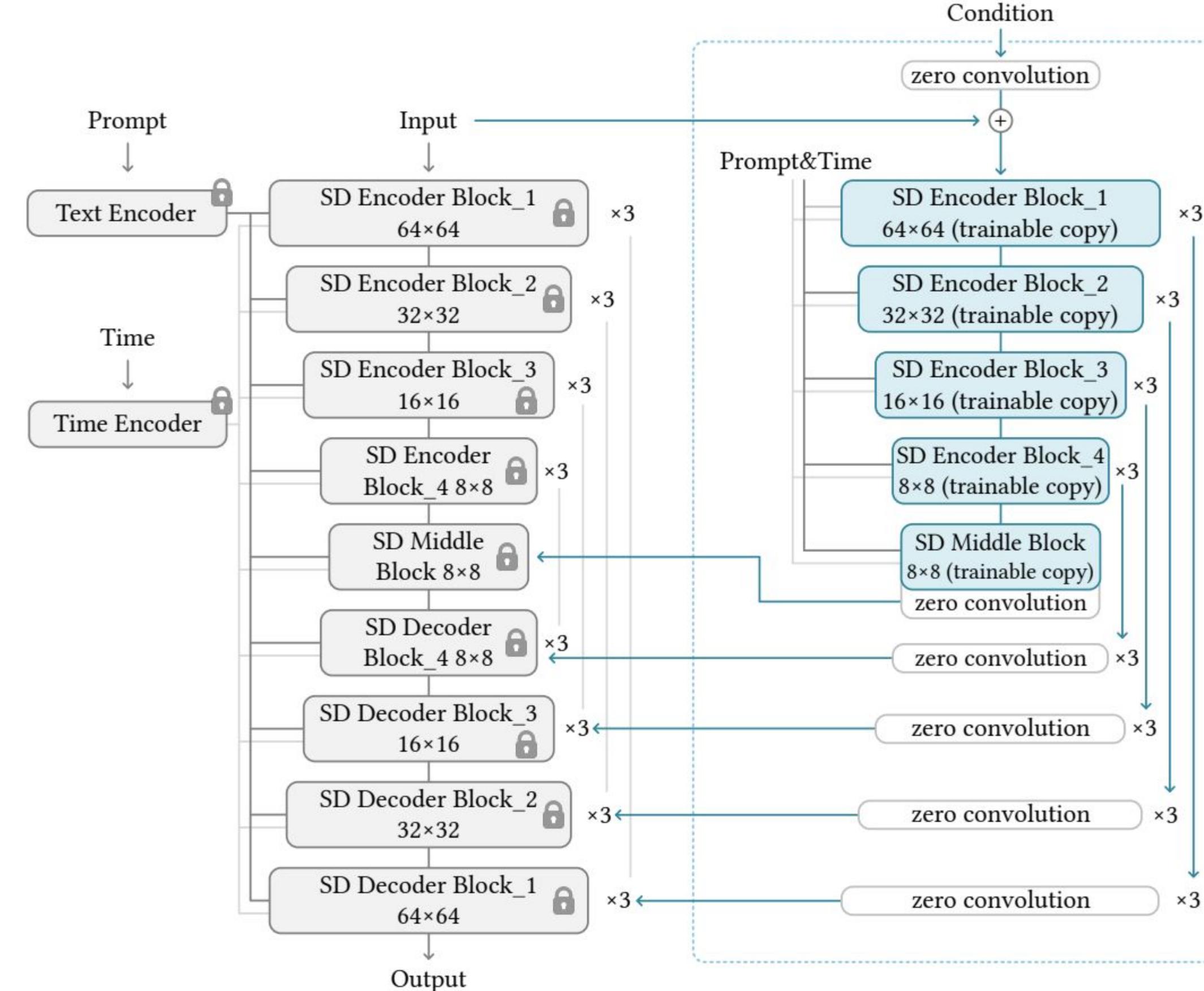
Trainable copy

Freezed copy

ControlNet

*zero convolution

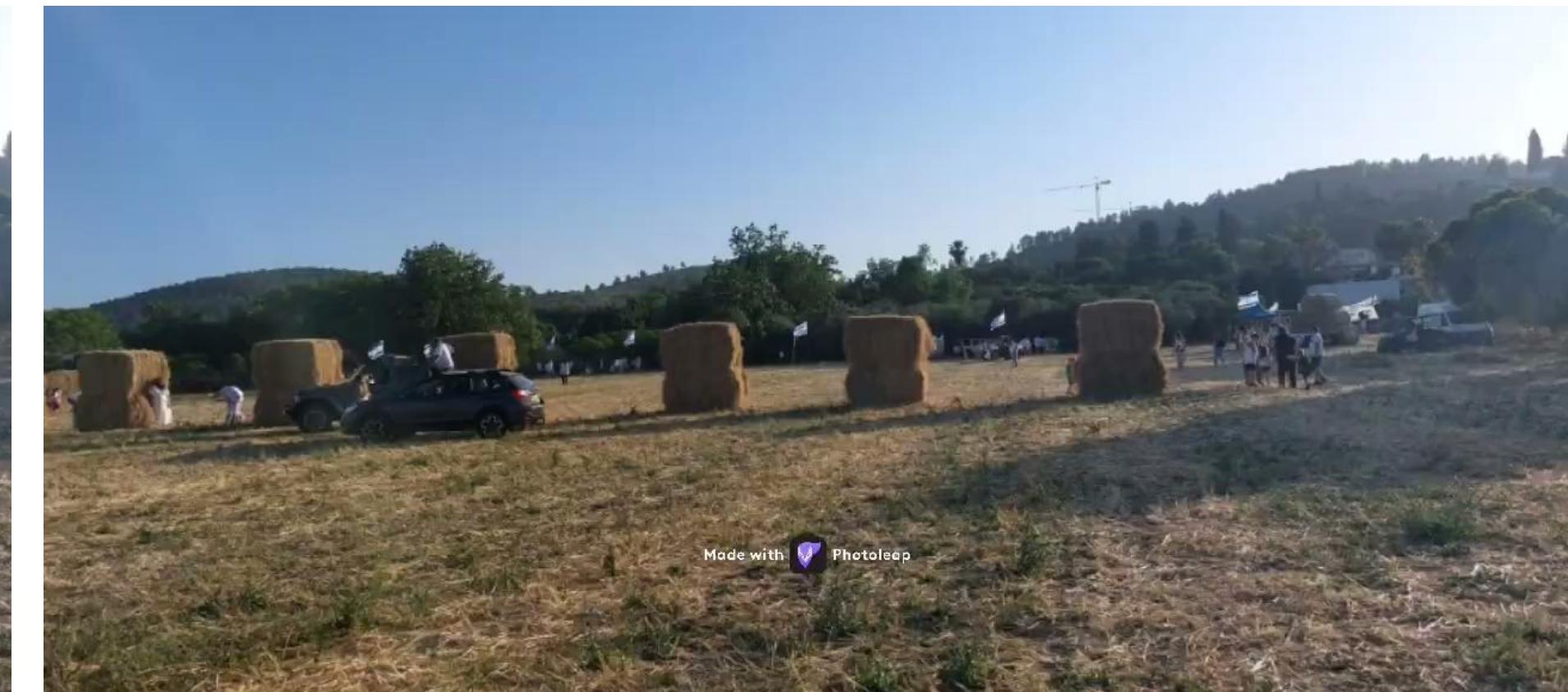




(a) Stable Diffusion

(b) ControlNet

It is all about the control

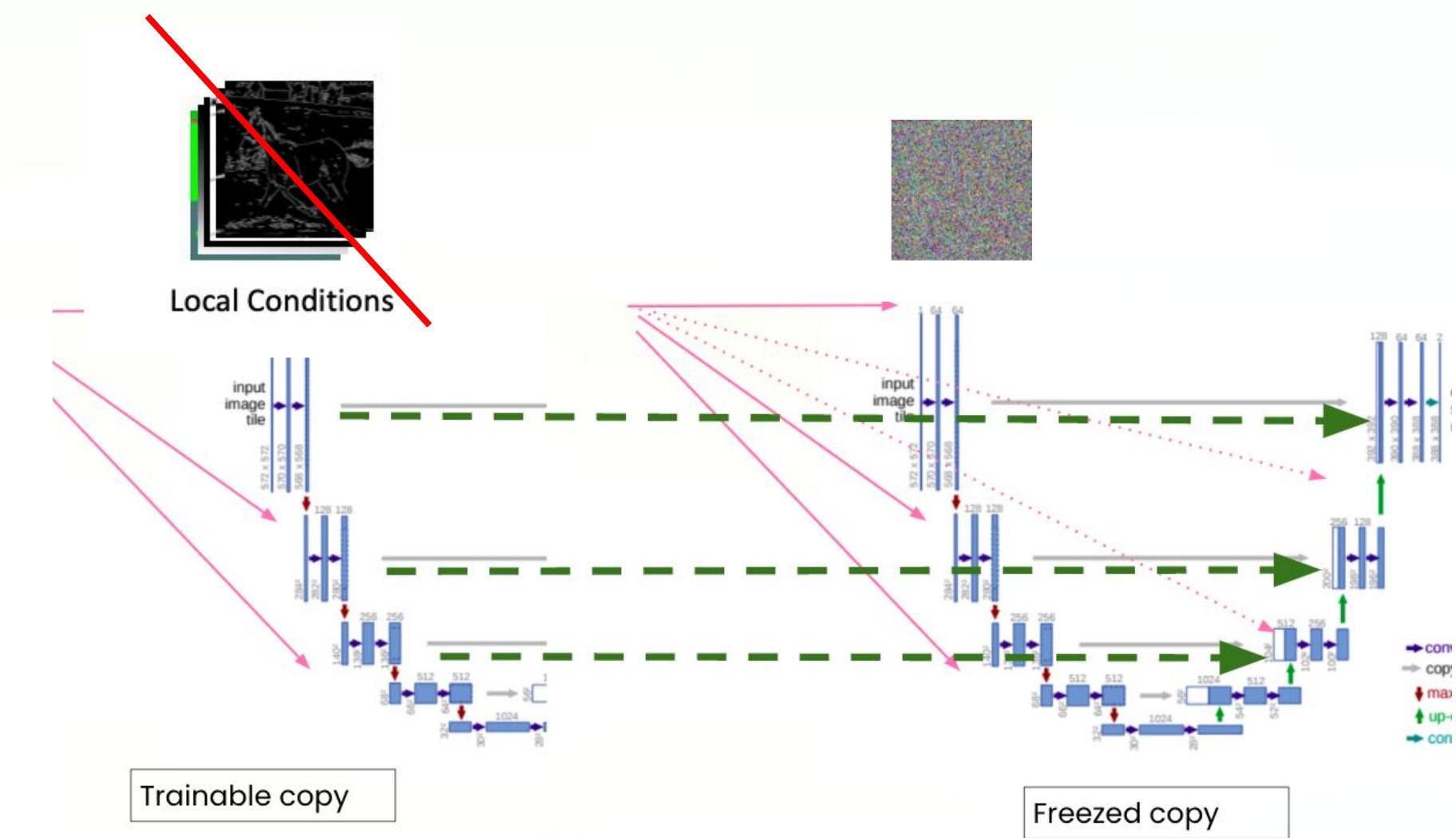


Uni-ControlNet: All-in-One Control

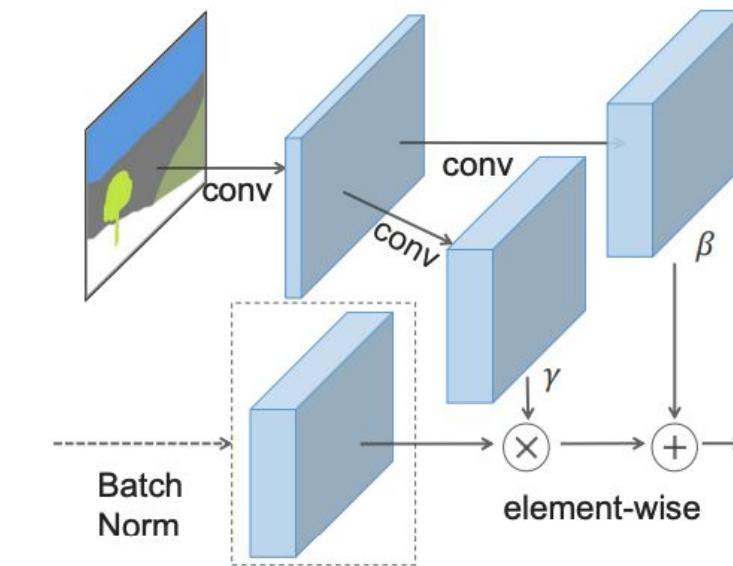
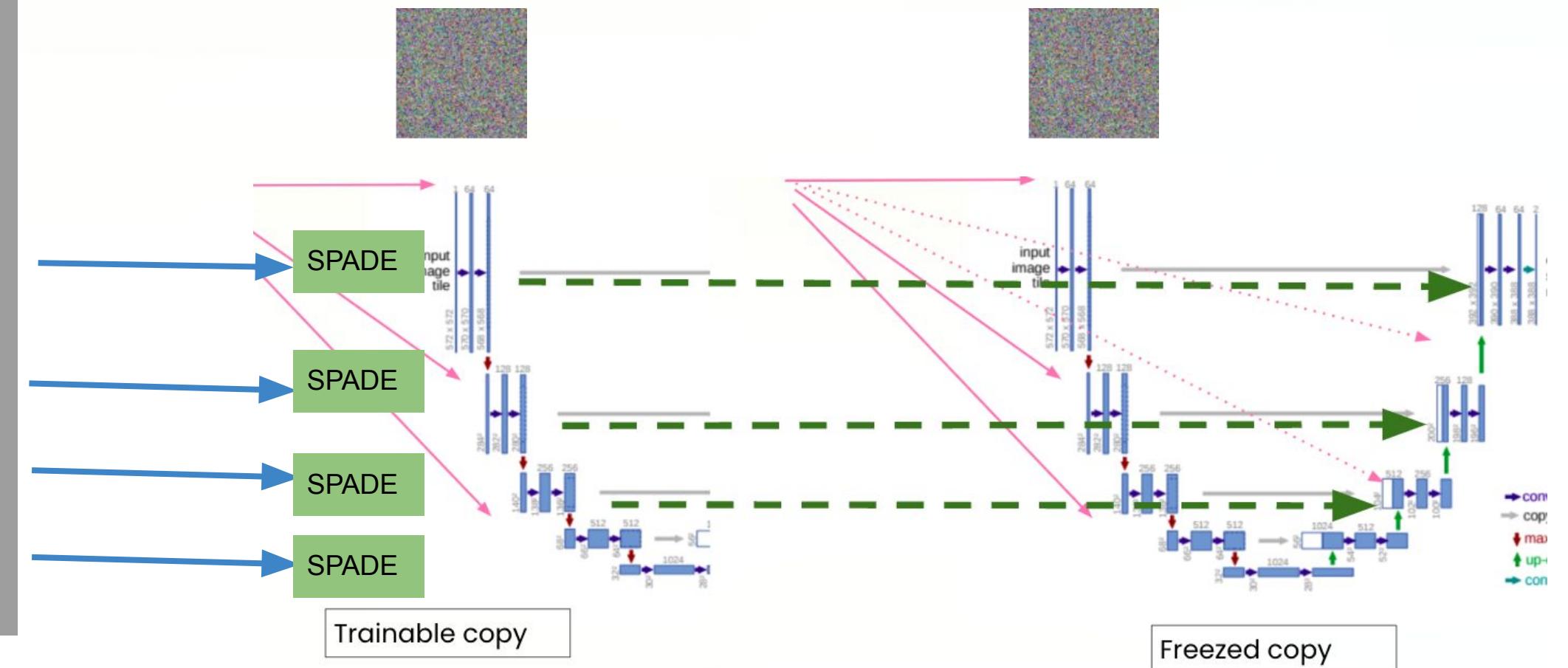
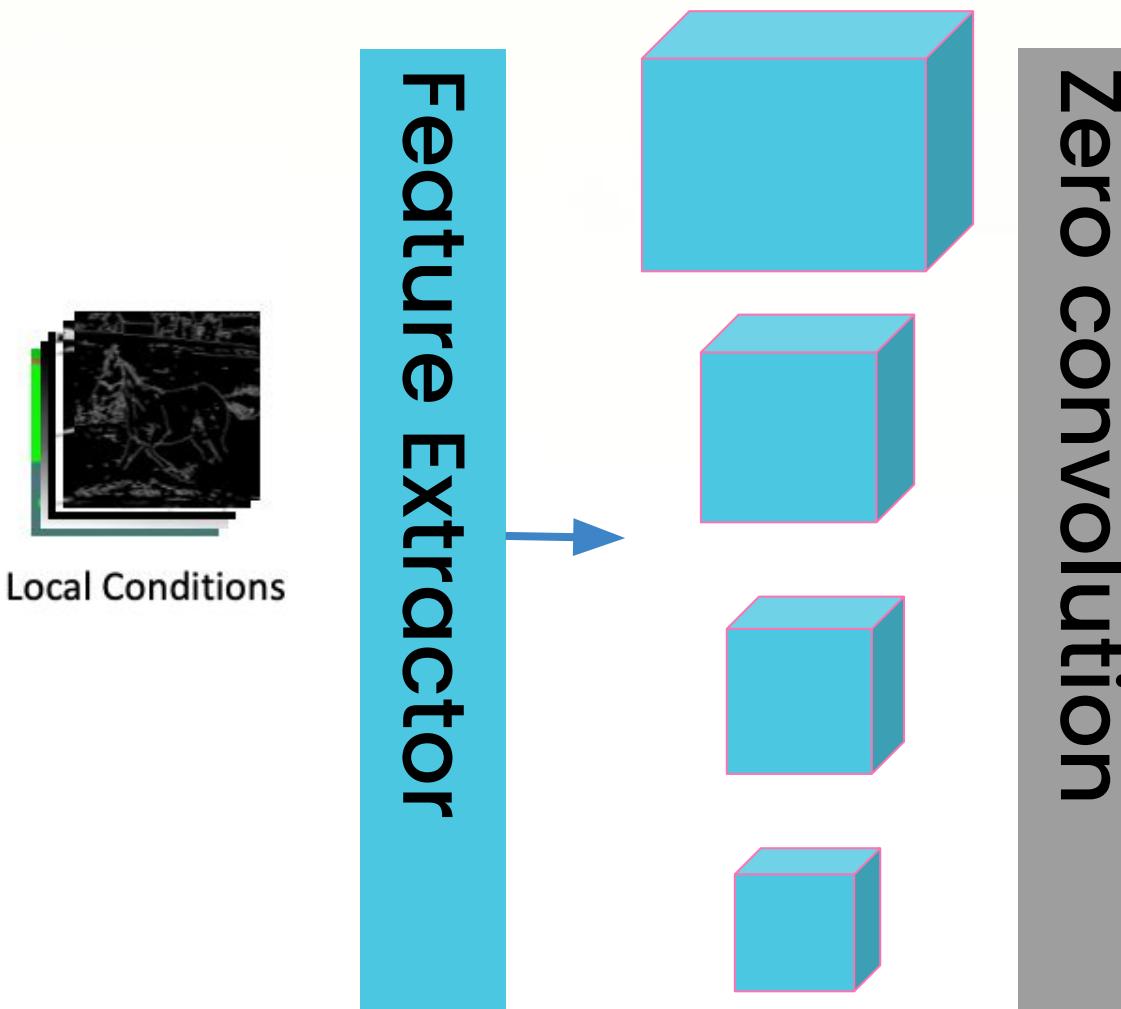
- Local and Global condition adapters
- Global adapter – via text tokens

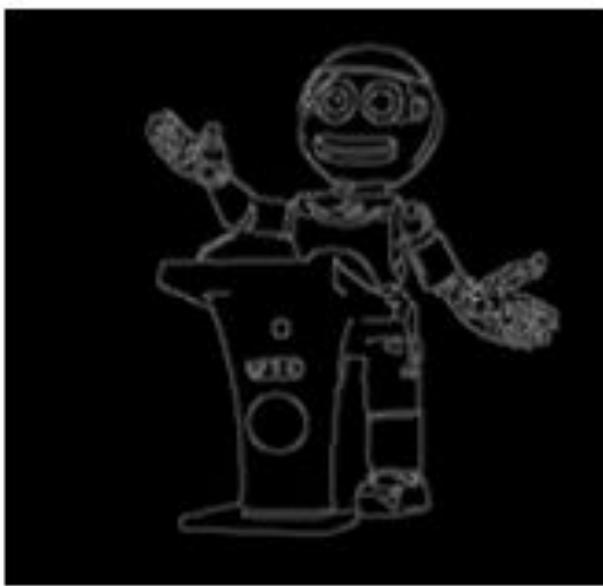
	Fine-tuning	Composable Control	Fine-tuning Cost	Adapter Number
ControlNet	✓	✗	N	N
Uni-ControlNet (Ours)	✓	✓	2	2

Local Control Adapter



Local Control Adapter





Condition-1

Condition-2

Ours

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/patal-suraj/Notebooks/blob/master/image_2_image_using_diffusers.ipynb

Questions?

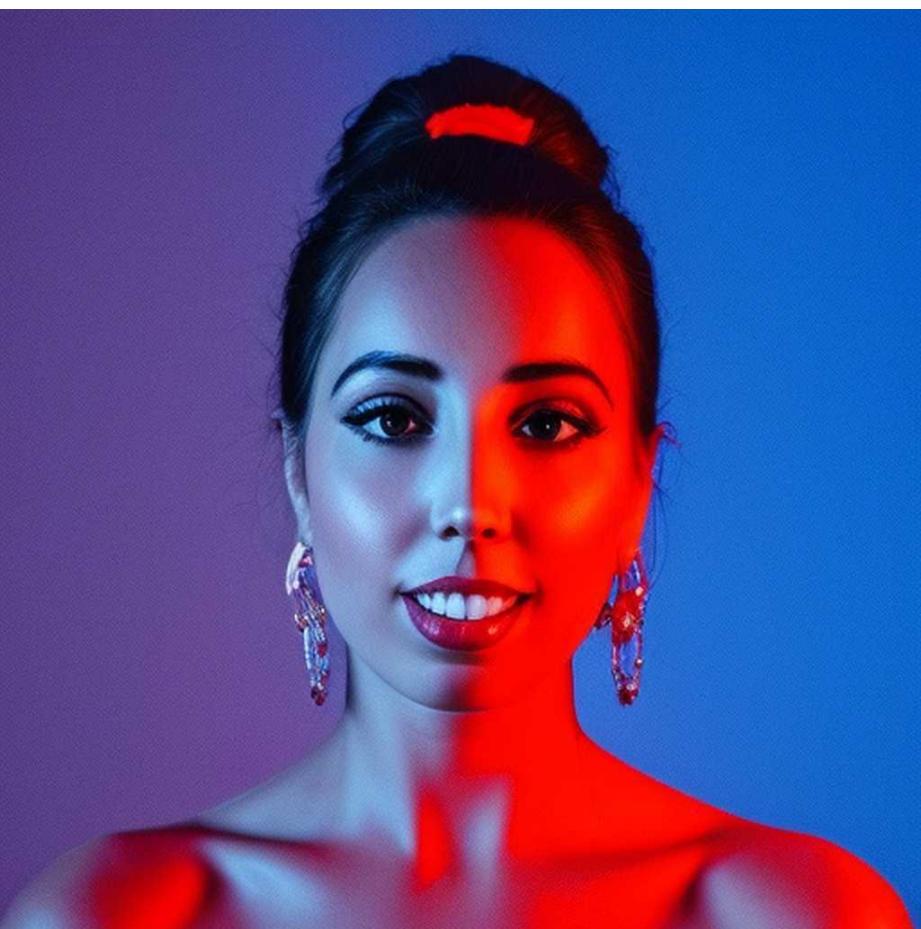
Takeaway

1. Industry will usually push the boundaries of images quality. A lot of expensive big experiments, academy will add the option to get control on the output.
2. Consolidity - today computer vision domain and NLP domain are not separated anymore, each one inherits ideas for the other
3. We covered a lot of architectures, and technical ideas. Combining those existing ideas leads to great progress



Thank you!

Lets Leverage the control



DreamBooth



Input images



in the Acropolis



swimming

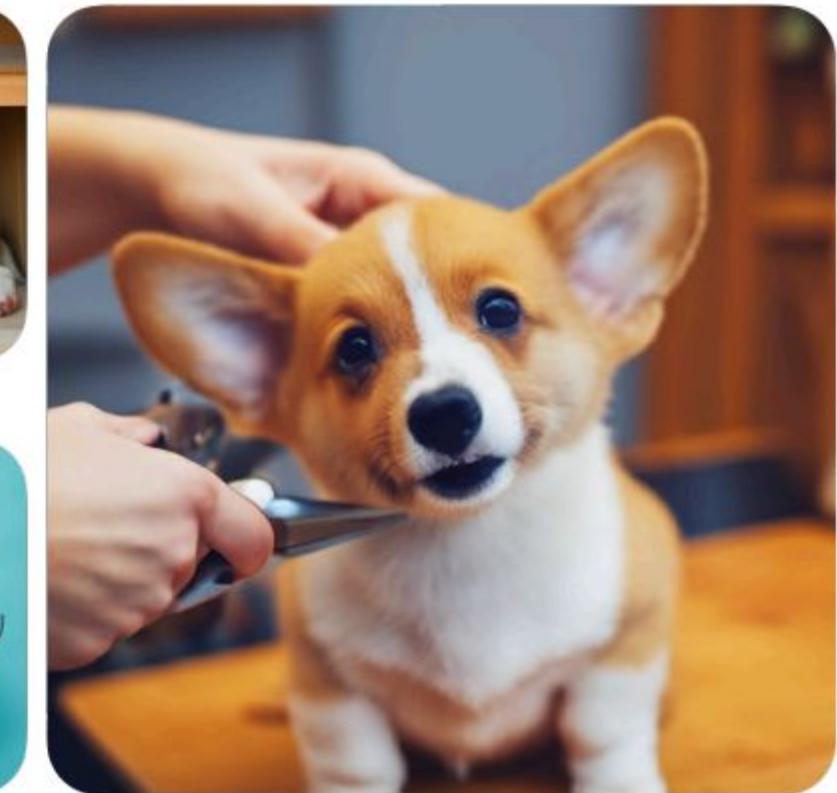
sleeping



in a doghouse

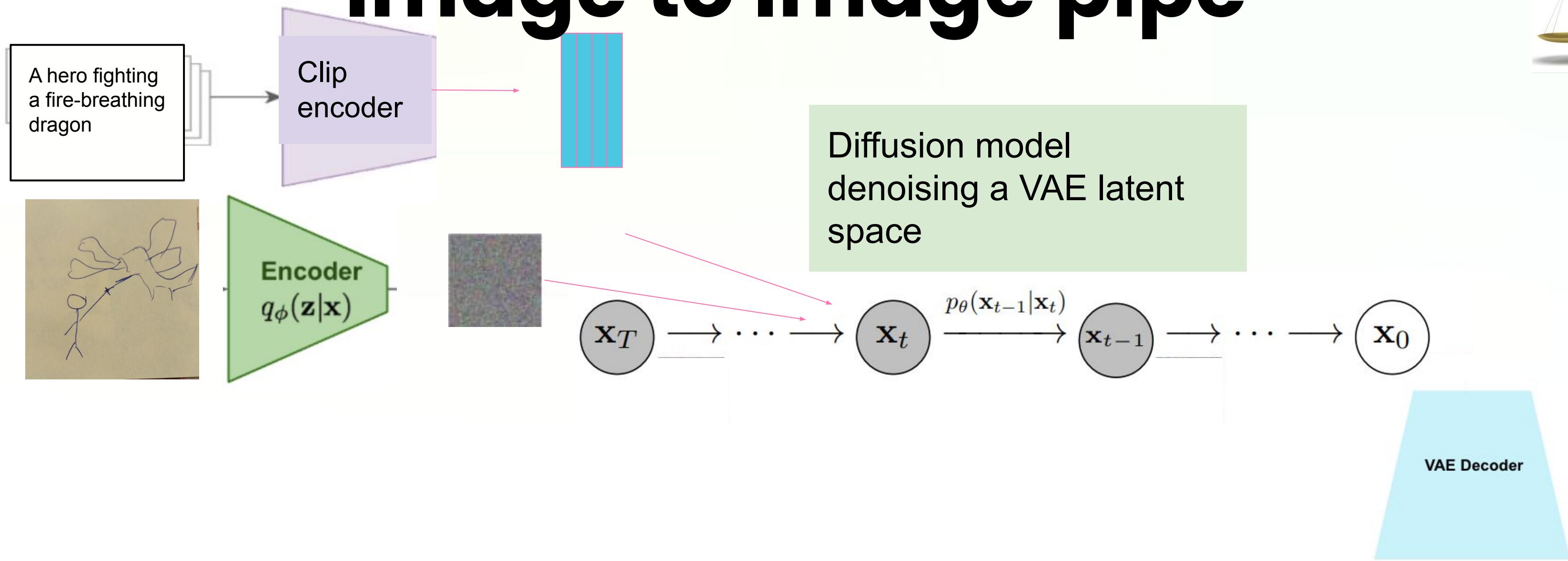


in a bucket

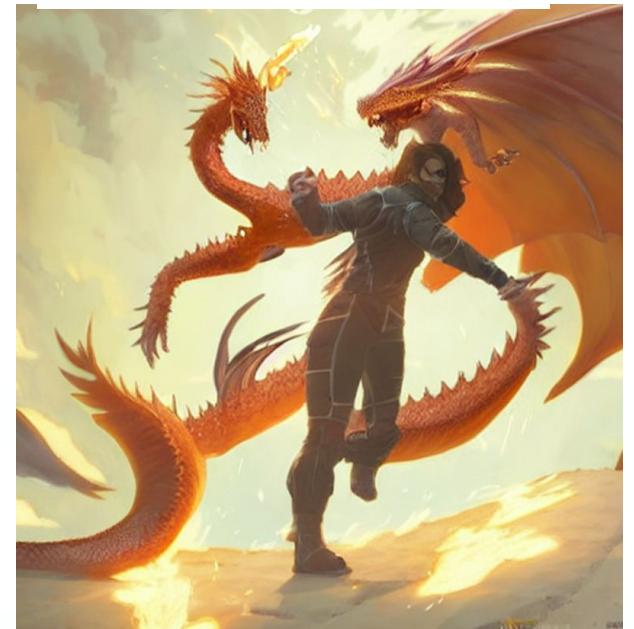


getting a haircut

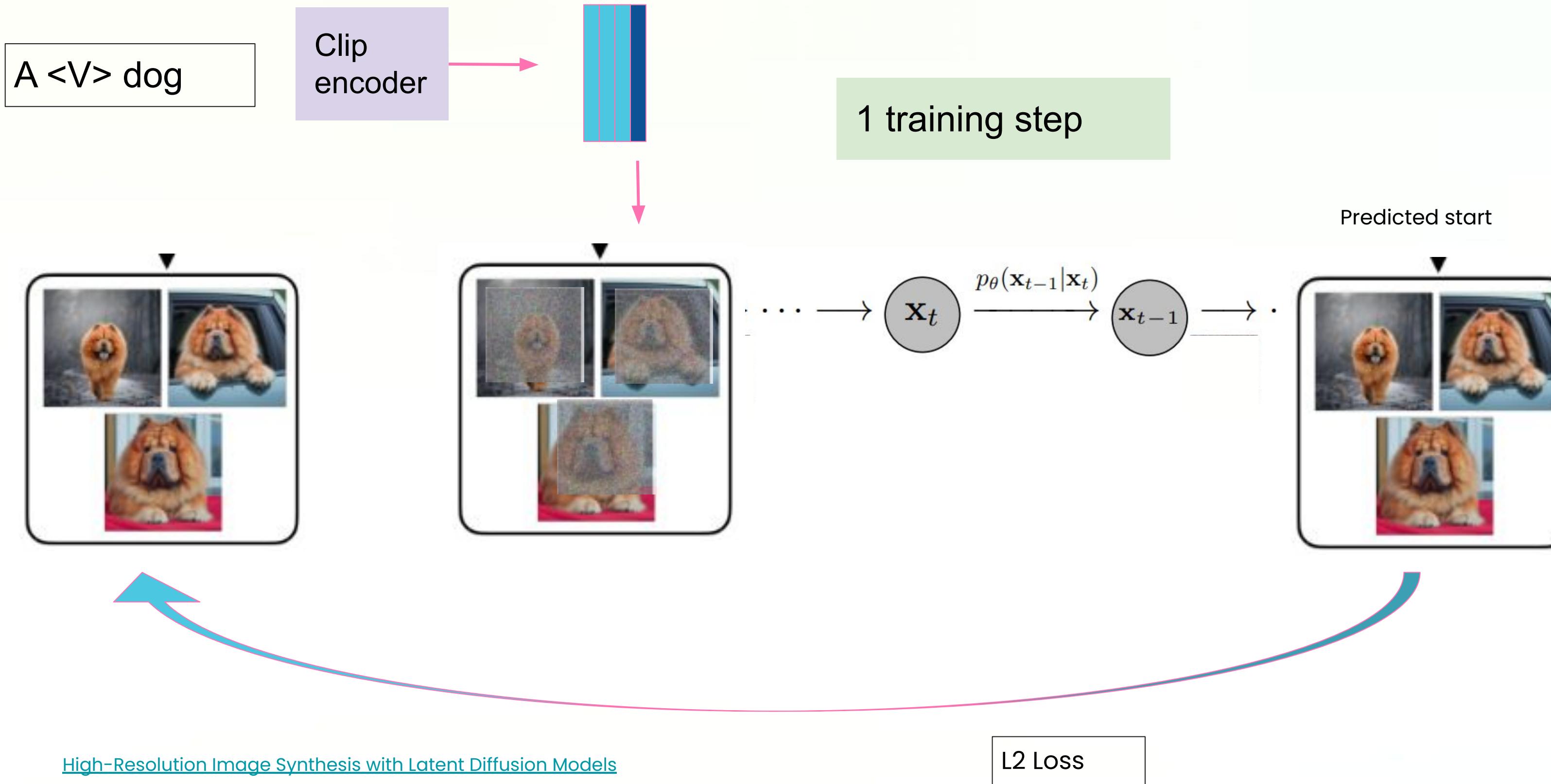
Image to Image pipe



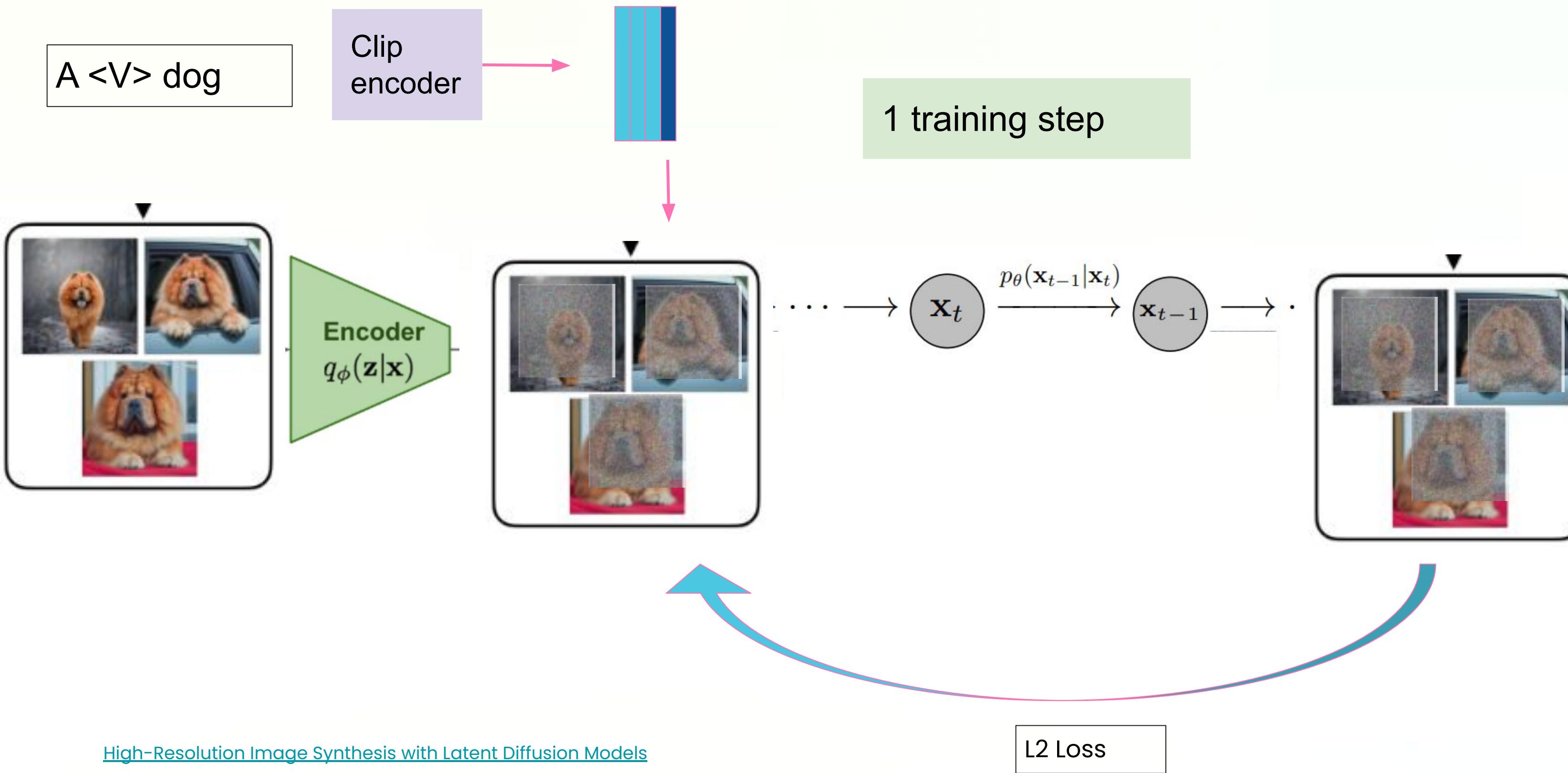
[High-Resolution Image Synthesis with Latent Diffusion Models](#)



Dreambooth Training

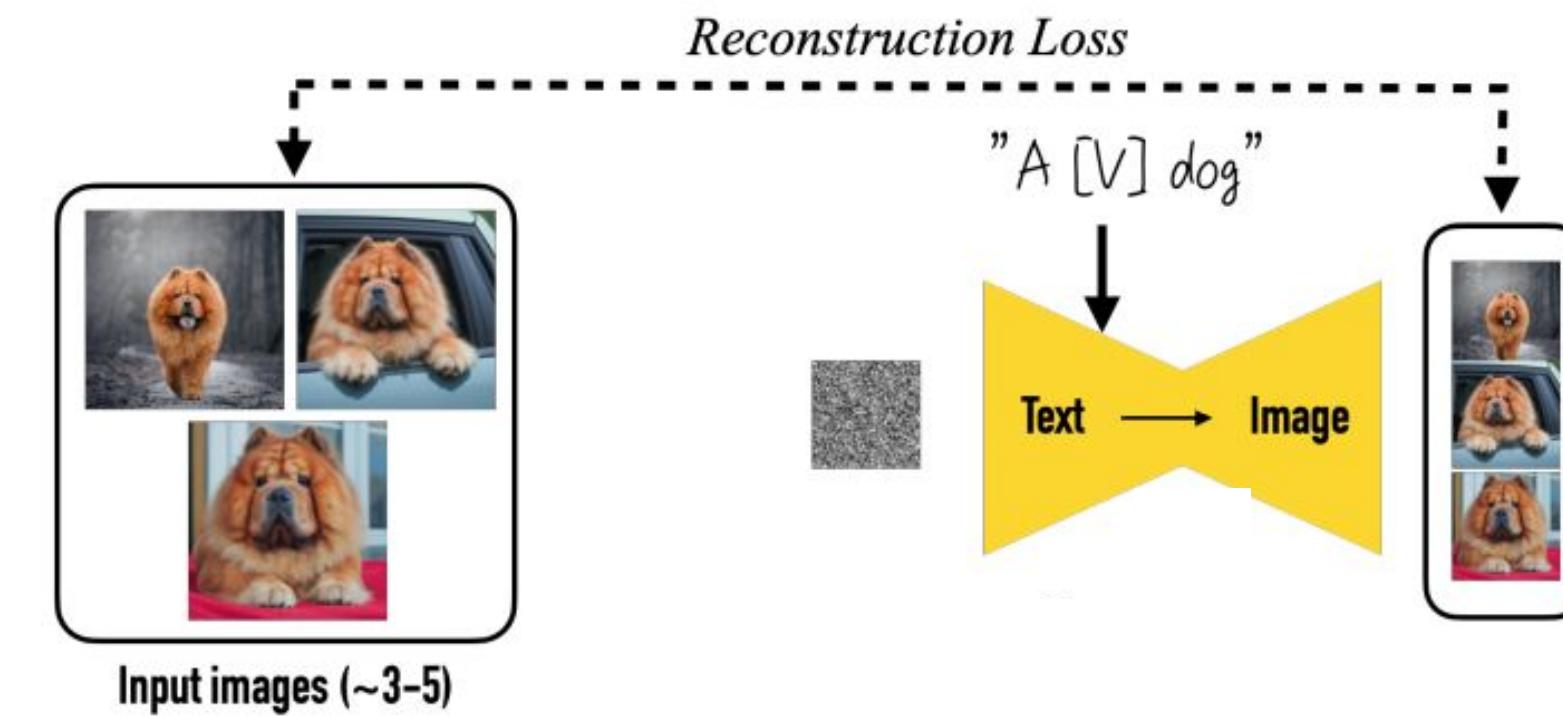


Dreambooth Training



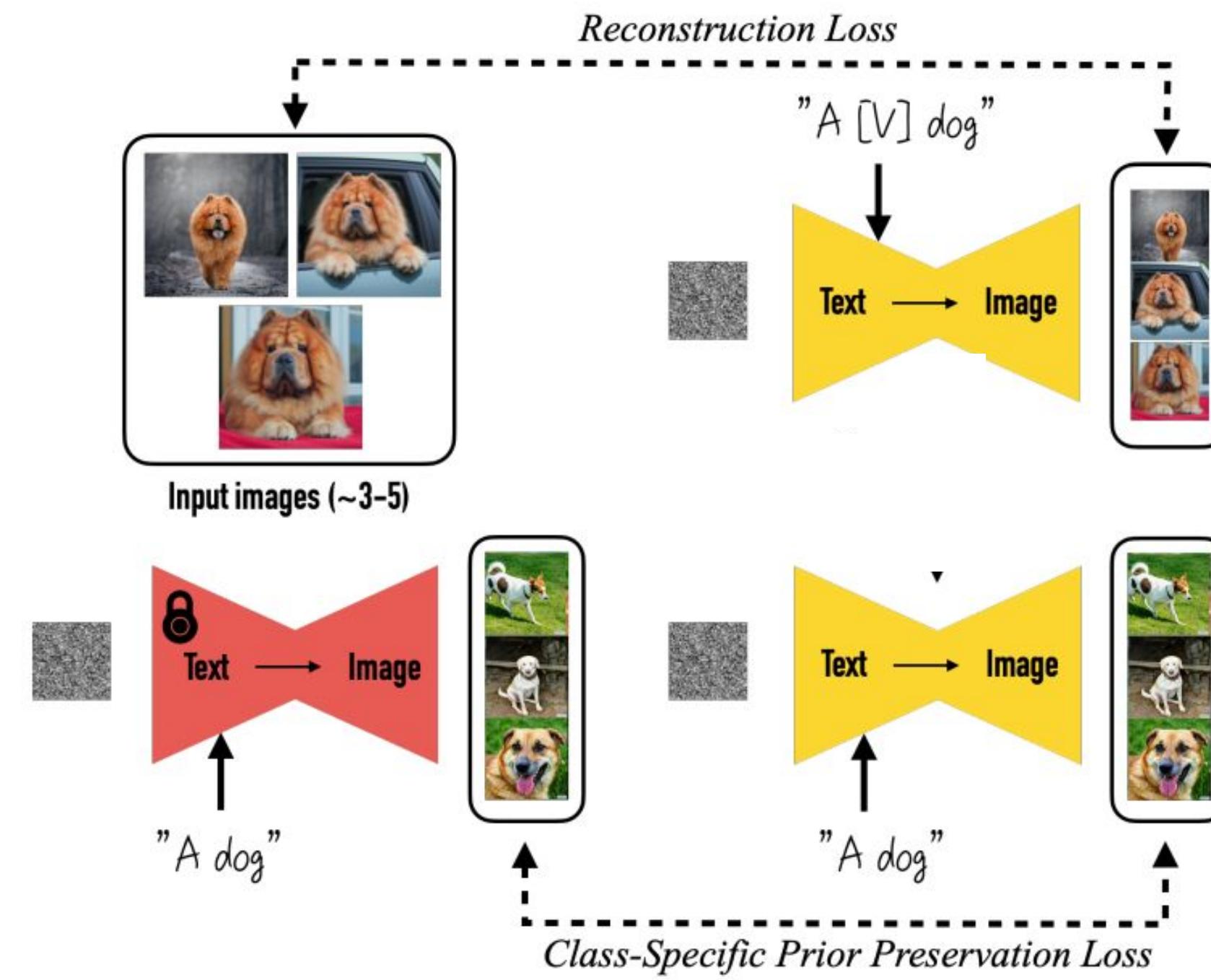
DreamBooth

- Rare text token
- Prior Preservation loss

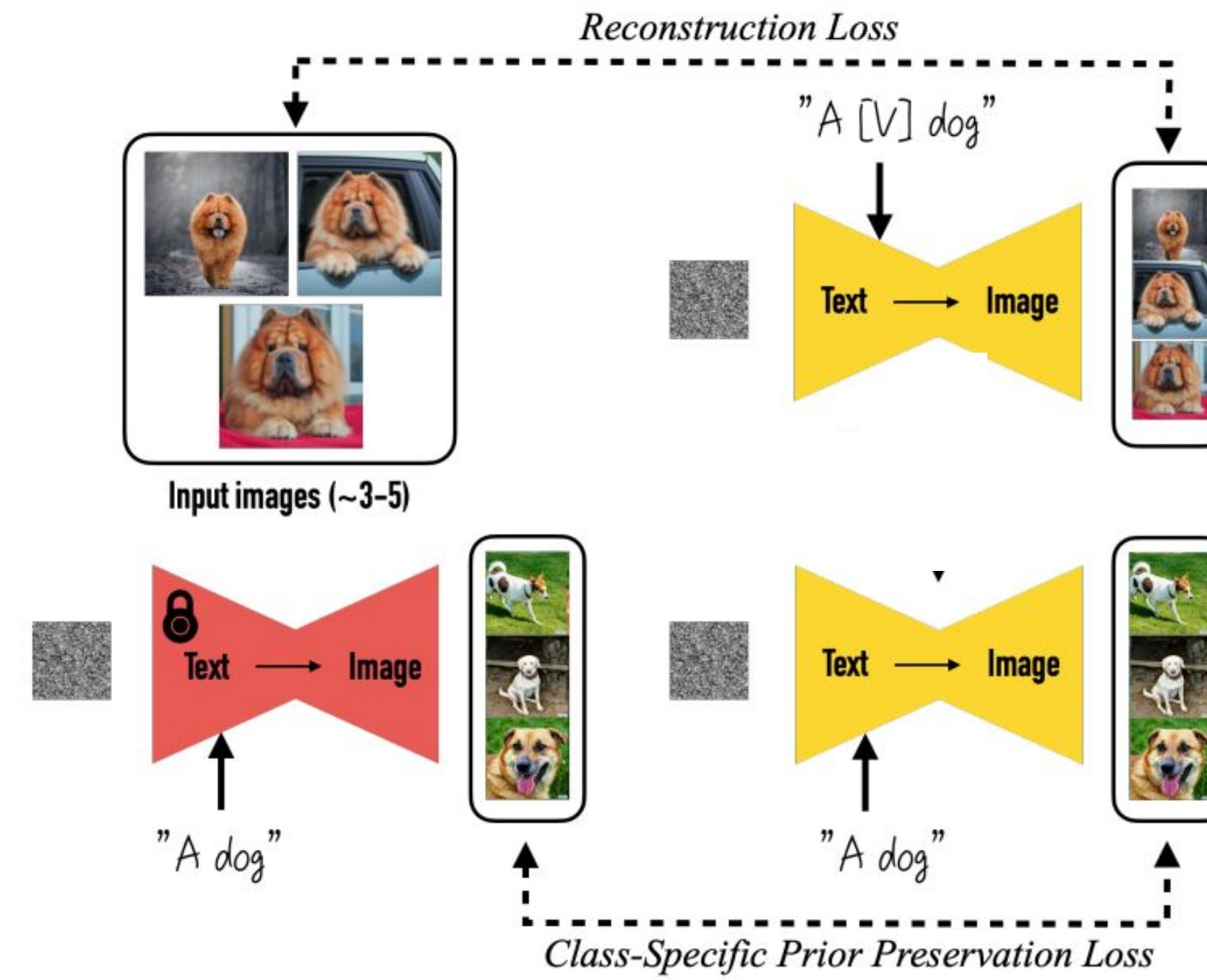
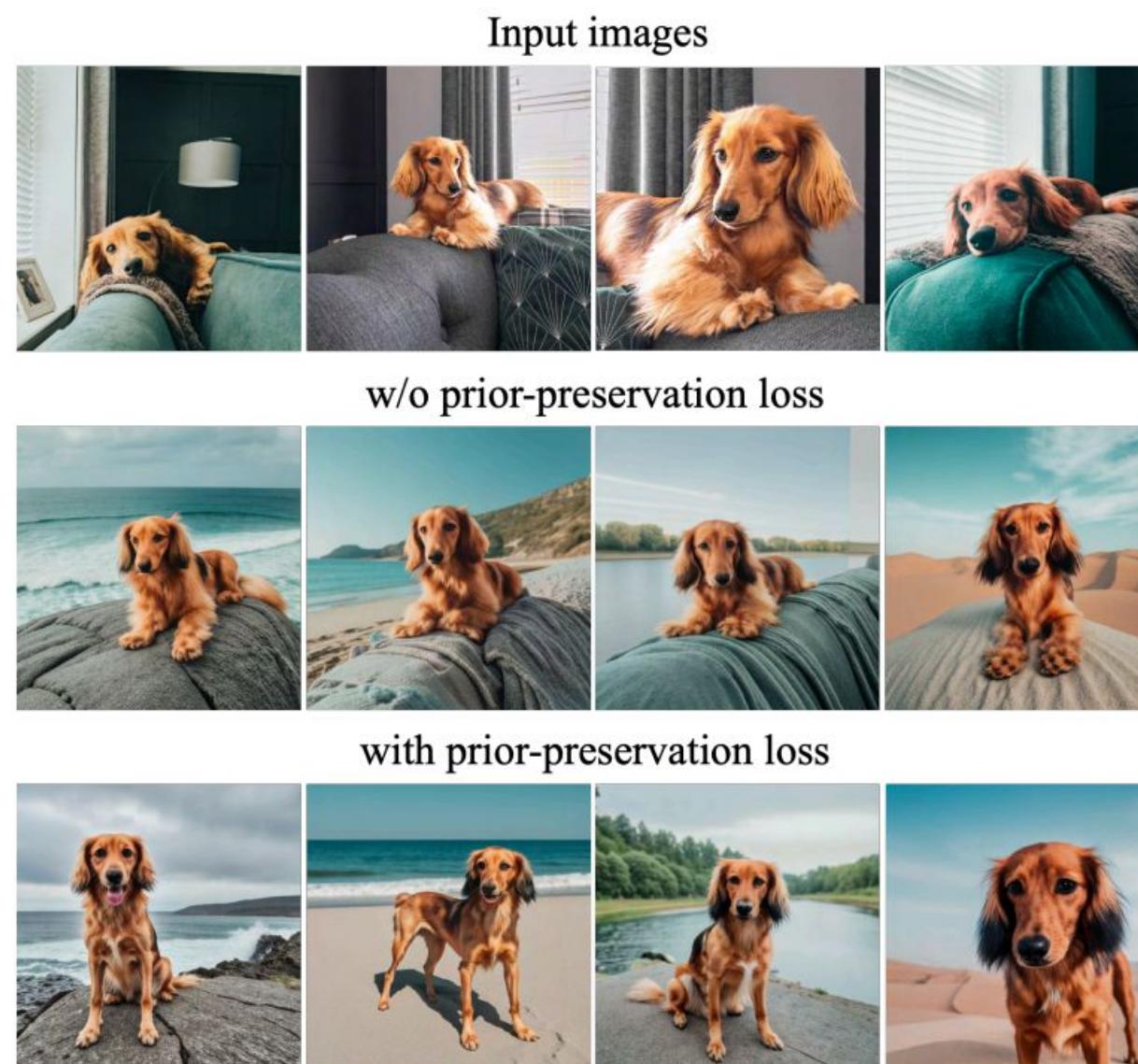


DreamBooth

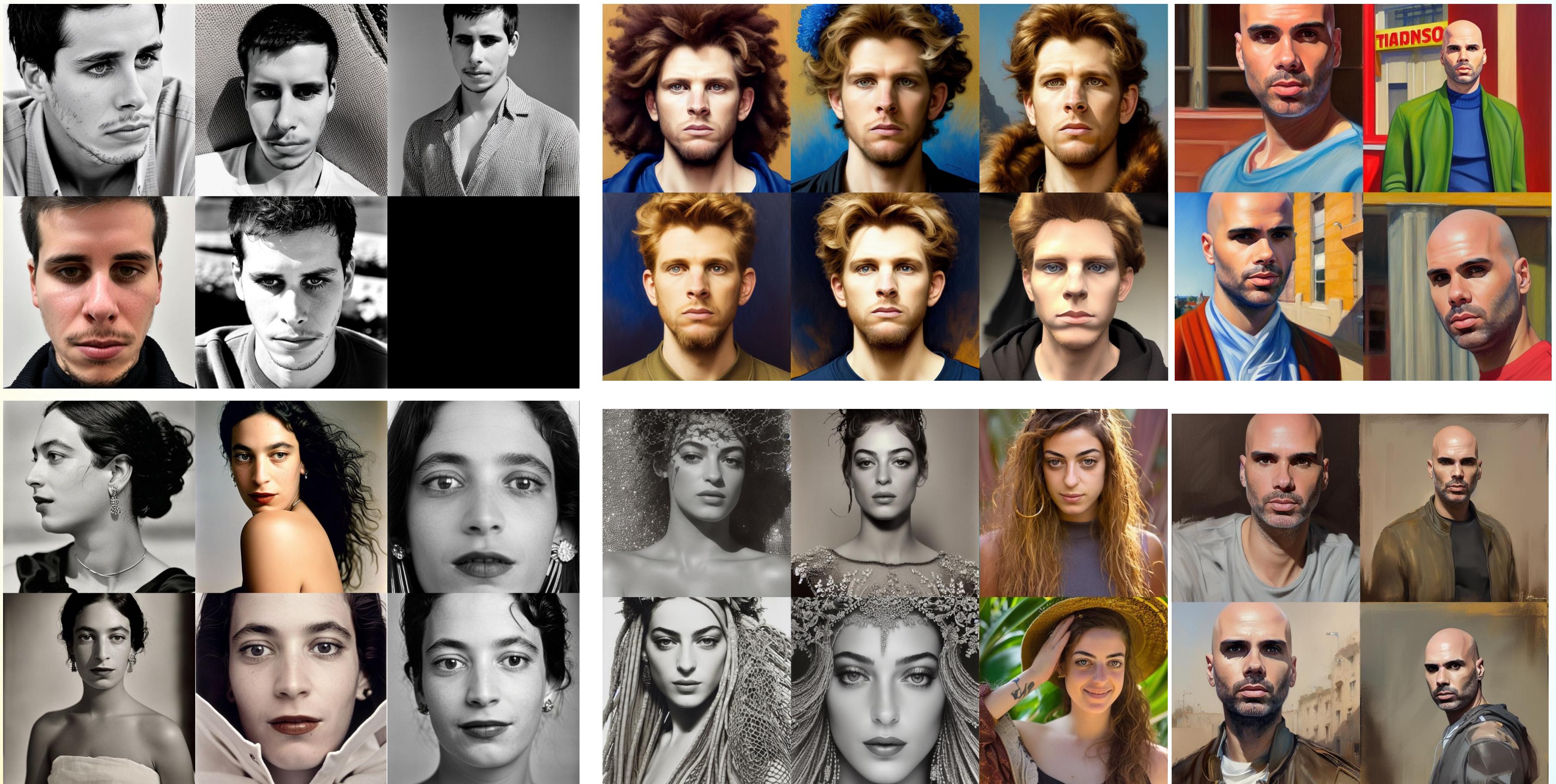
- Rare text token
- Prior Preservation loss



DreamBooth



DreamBooth In practice



Many Conditions in One Model

Uni-ControlNet 10M text-image pairs

Evaluation: fid: 5k

Table 2: **FID** on different controllable diffusion models. The best results are in **bold**.

	Canny	MLSD	HED	Sketch	Openpose	Depth	Segmentation	Style\Content
ControlNet	18.90	31.36	26.59	22.19	27.84	21.25	23.08	31.17
T2I-Adapter	18.98	-	-	18.83	29.57	21.35	23.84	28.86
Ours	17.79	26.18	17.86	20.11	26.61	21.20	23.40	23.98

Composer

64X64 (2B) batch size - 1024 1M steps pretrain
Finetune 200K steps on 60M samples
= 1B

256X256 (1.1B) batch size 512 1M steps

1024X1024 (1B) batch size 512 1M steps

Data of 1B

Evaluation:

Composer achieves a zero-shot FID of 9.2 in text-to-image synthesis on the COCO dataset