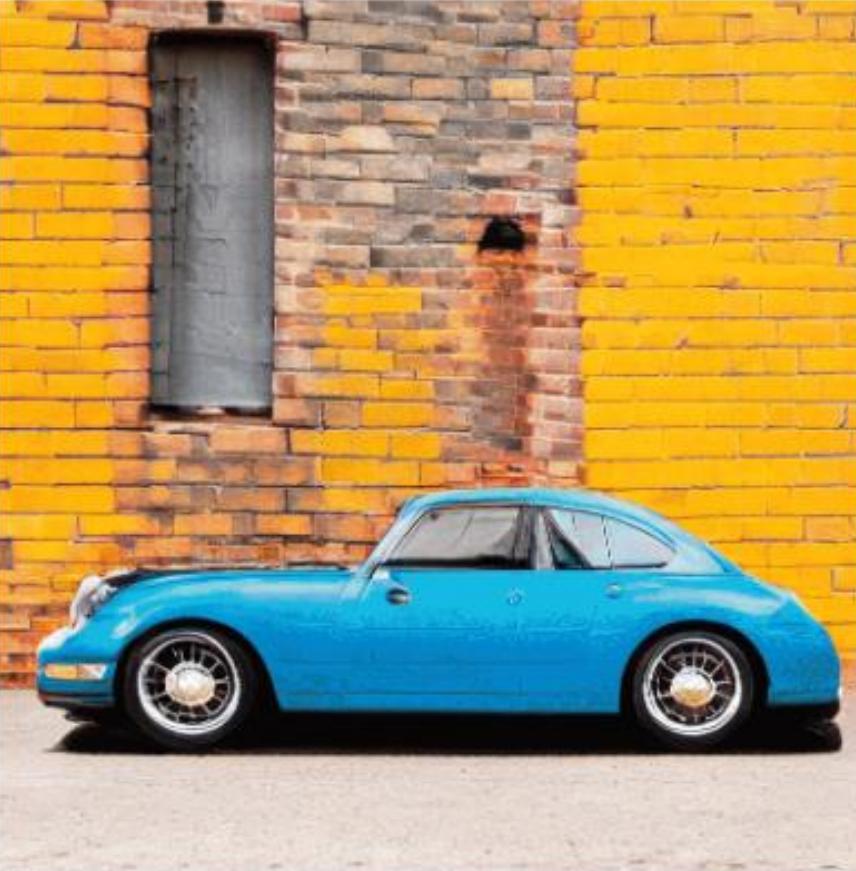
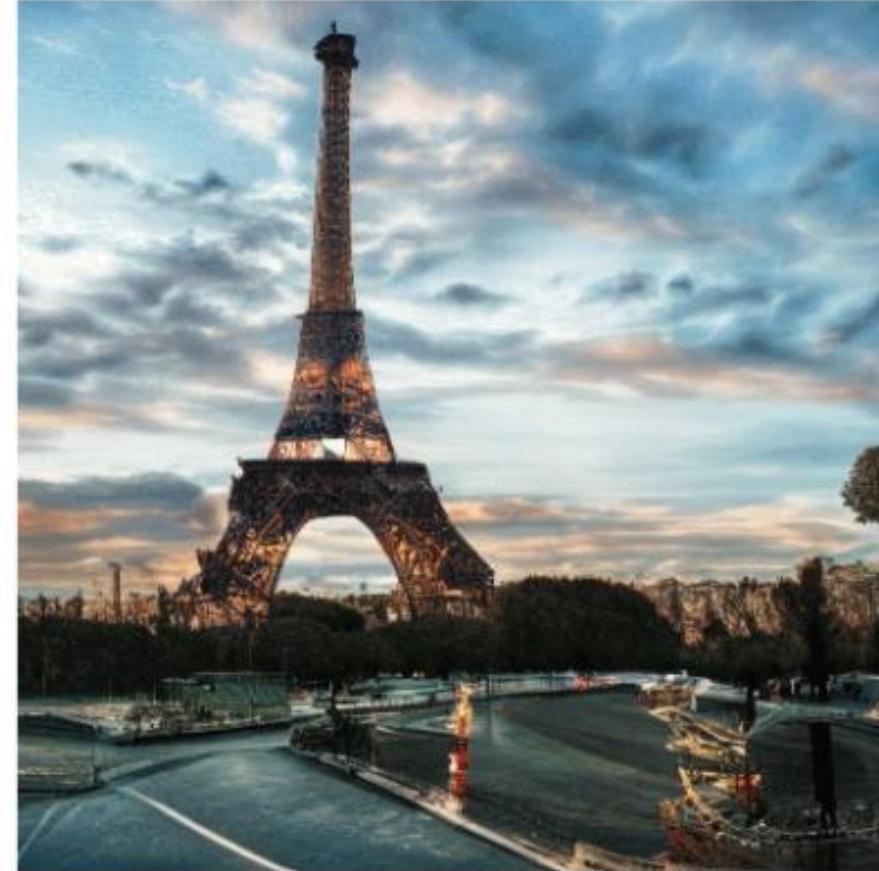




A living room with a fireplace at a wood cabin. Interior design.



a blue Porsche 356 parked in front of a yellow brick wall.



Eiffel Tower, landscape photography



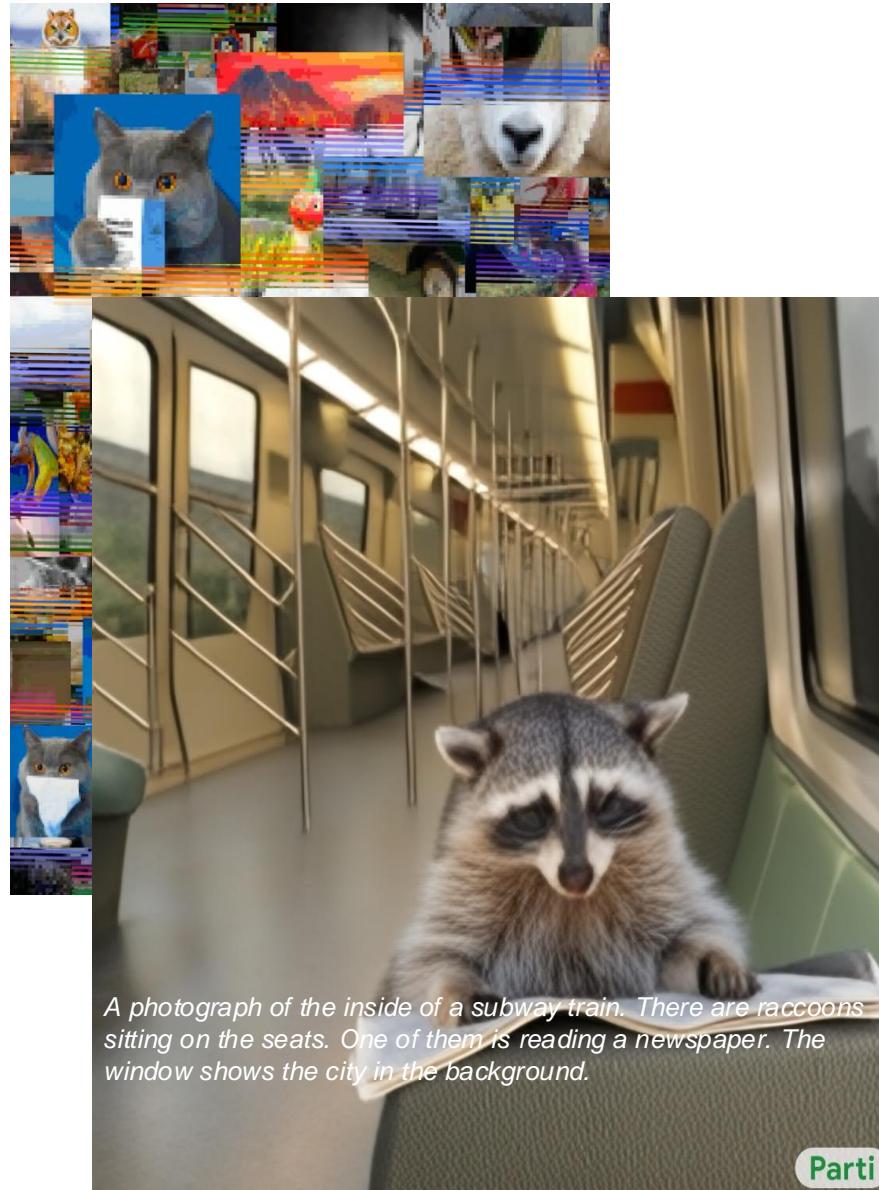
A painting of a majestic royal tall ship in Age of Discovery.

# Lecture 14: Text-to-Image Synthesis

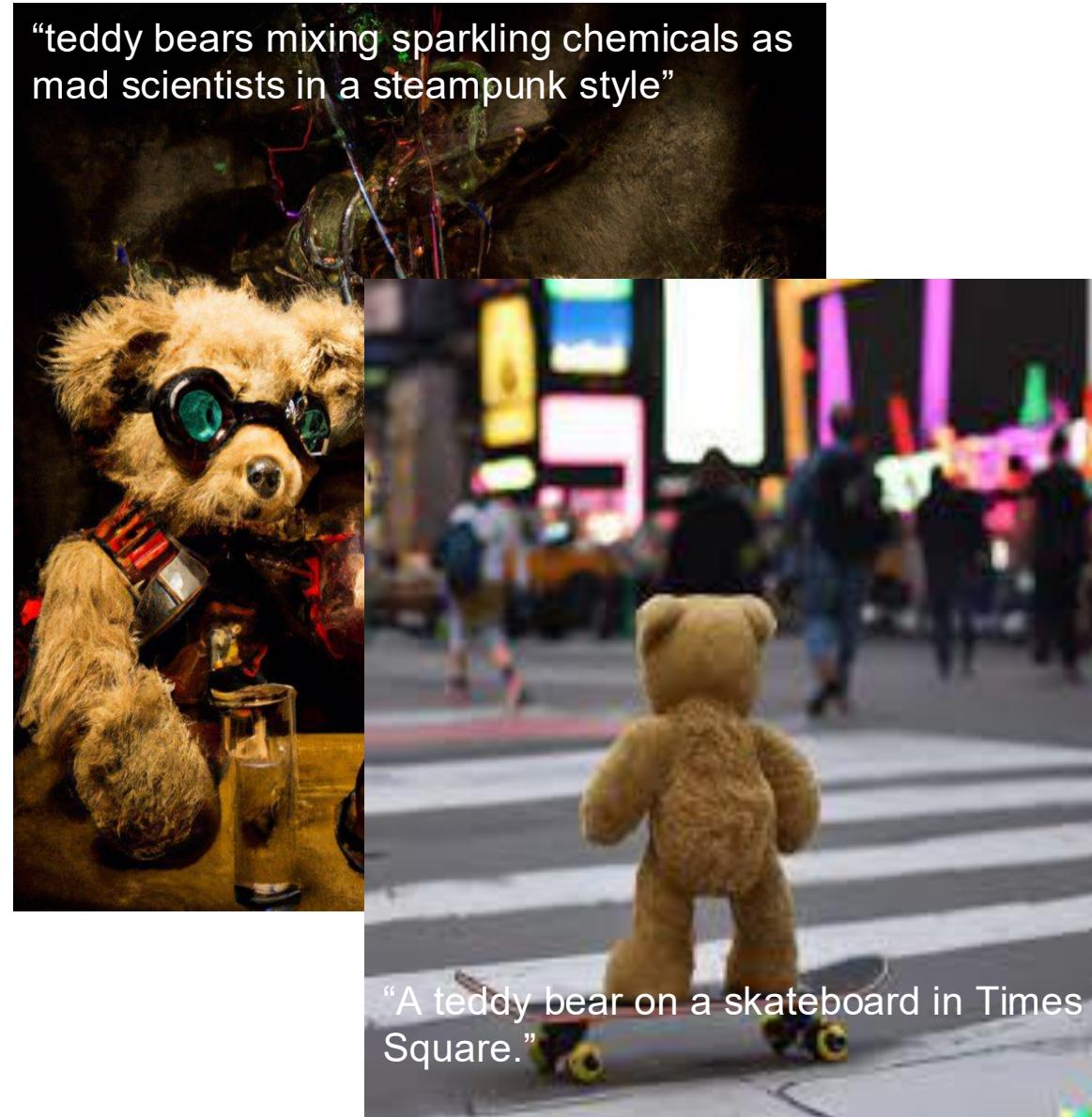
Jun-Yan Zhu

16-726 Spring 2025

# Text-to-Image Everywhere



Autoregressive models  
(Image GPT, Parti)



Diffusion models  
(DALL-E 2, Imagen)



GANs, Masked GIT  
(GigaGAN, MUSE)

# Text-to-Image Everywhere

The collage illustrates the widespread application of AI-generated imagery across different platforms and domains:

- Top Left:** A post by Scott Lighthiser (@LighthiserScott) from Sep 18, featuring a dark, skeletal creature. The caption includes hashtags #stablediffusion and #Alart.
- Top Middle:** A post by Orcton (@OrctonAI) from Sep 15, showing a portrait of a man with a beard in a city at night. The caption discusses creating personalized digital assistants and includes hashtags #midjourney, #Midjourneyai, #Alart, #Digitalart, and #animated.
- Top Right:** A post by Matt DesLauriers (@mattdesl) from Sep 13, demonstrating color palettes generated from text prompts like "tokyo neon" and "green garden, blue sky". It also includes a link to a thread about AI tools for generating color palettes.
- Middle Left:** A post by Scott Lighthiser (@LighthiserScott) from Sep 18, showing a woman's face with a complex, organic texture. The caption includes hashtags #stablediffusion and #Alart.
- Middle Center:** A post by Matt Reed (@mcreed) from Sep 9, featuring pixelated versions of Mario and Luigi. The caption reads "I am at a loss for everything" and includes hashtags #stablediffusion and #Alart.
- Middle Right:** A post by Replicate (@replicatino) from Sep 9, showcasing a new open-source model for producing seamless tiling images. It includes a grid of various textures like stone, leaves, and fruit.
- Bottom Left:** A post by Replicate (@replicatino) from Sep 9, featuring a close-up of a metallic, ornate mask with glowing yellow eyes.
- Bottom Right:** A post by Stable Diffusion (@StableDiffusion), Pics & DreamStudio (@DiffusionPics), and Jeffon Zuckergates (@JeffonZuckergates) from Sep 2, showing a portrait of Elon Musk.

Where/when did it start?

# First Text-to-Image System

First the farmer gives hay to the goat. Then the farmer gets milk from the cow.



Step 1: Image Selection.

Step 2: Layout Optimization (Minimum overlap, Centrality, Closeness)

A Text-to-Picture Synthesis System for Augmenting Communication

Xiaojin Zhu, Andrew Goldberg, Mohamed Eldawy, Charles Dyer, and Bradley Strock. AAAI 2007

# First Text-to-Image System



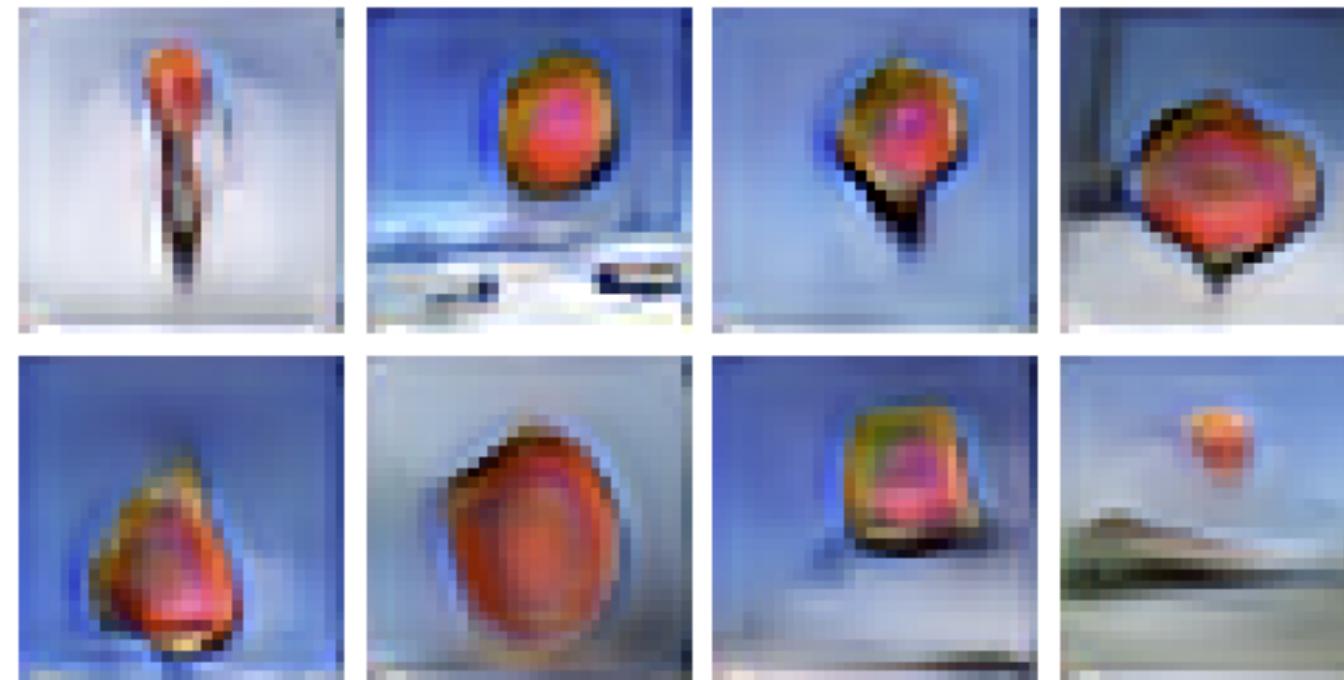
Therapy for people  
with communicative disorders



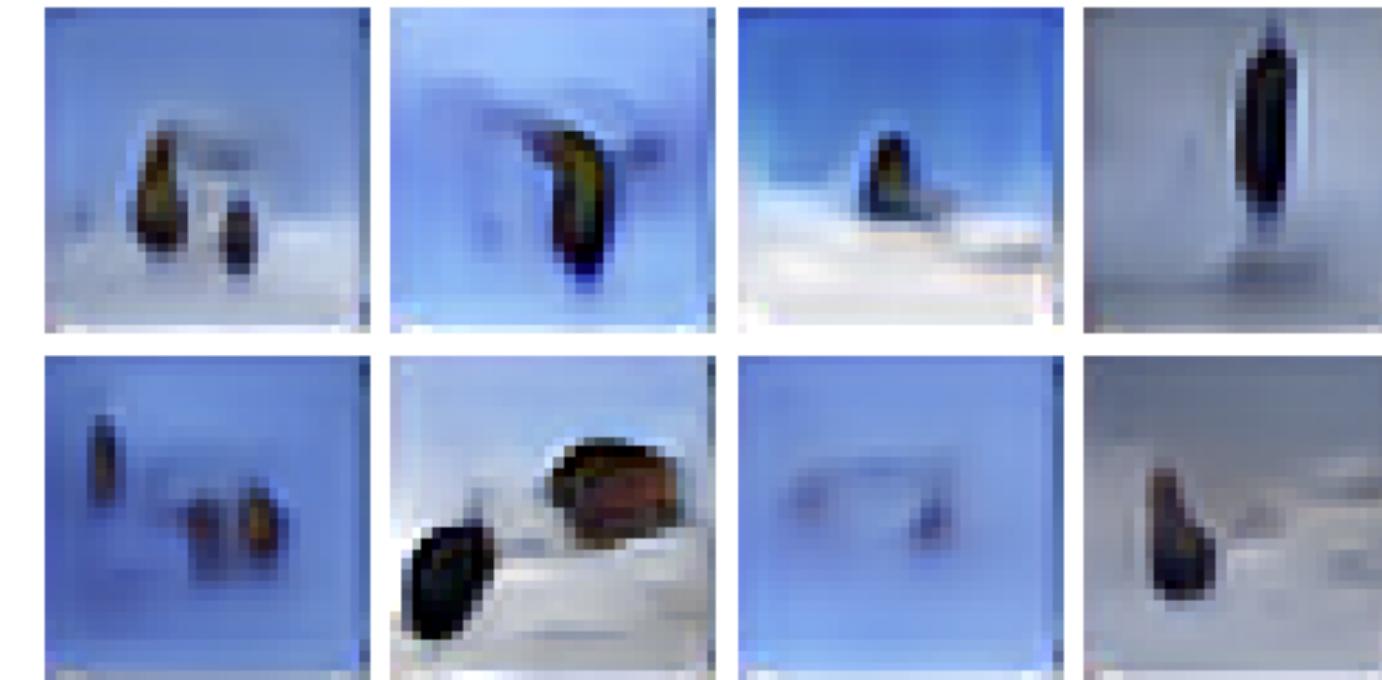
Math learning and reading comprehension  
for young children

A Text-to-Picture Synthesis System for Augmenting Communication  
Xiaojin Zhu, Andrew Goldberg, Mohamed Eldawy, Charles Dyer, and Bradley Strock. AAAI 2007

# First Deep Learning Work



A stop sign is flying in  
blue skies.



A herd of elephants fly-  
ing in the blue skies.

Generating Images from Captions with Attention.

Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, Ruslan Salakhutdinov. ICLR 2016.

# First Deep Learning Work



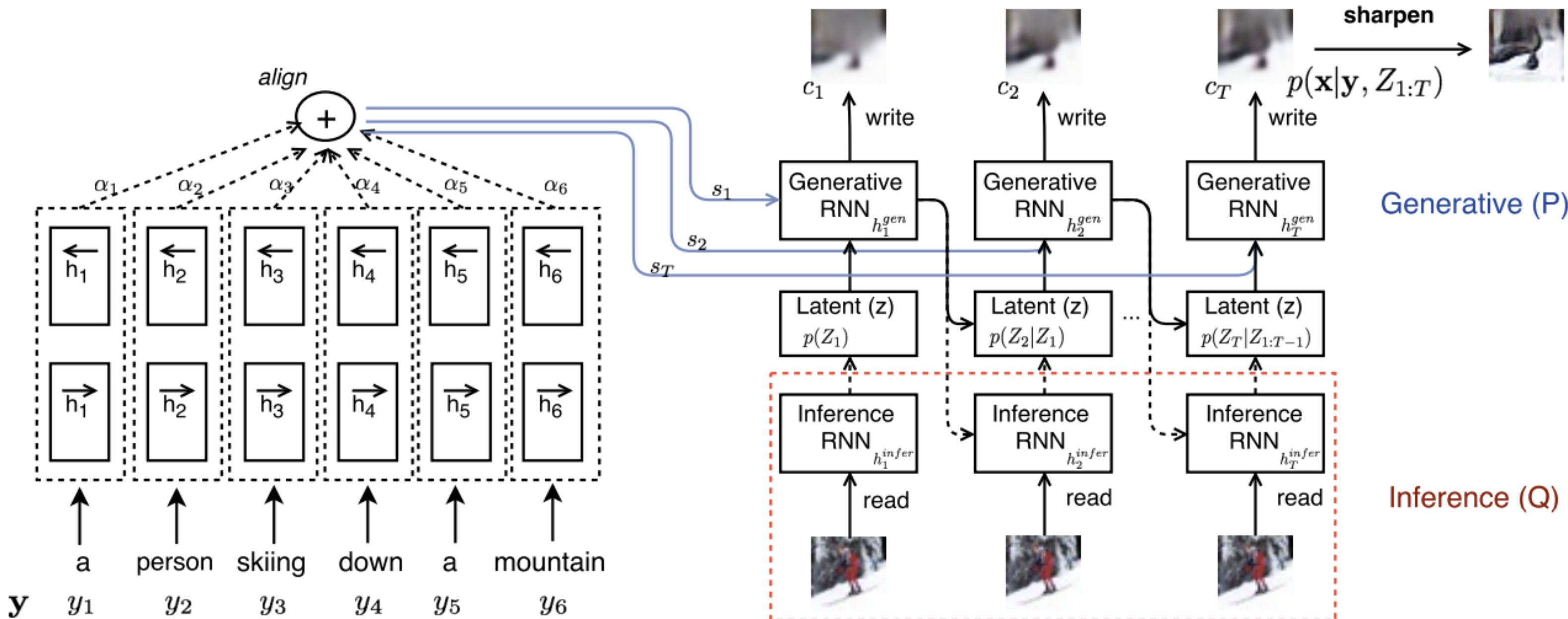
A toilet seat sits open in  
the grass field.

A person skiing on sand  
clad vast desert.

Generating Images from Captions with Attention.

Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, Ruslan Salakhutdinov. ICLR 2016.

# First Deep Learning Work

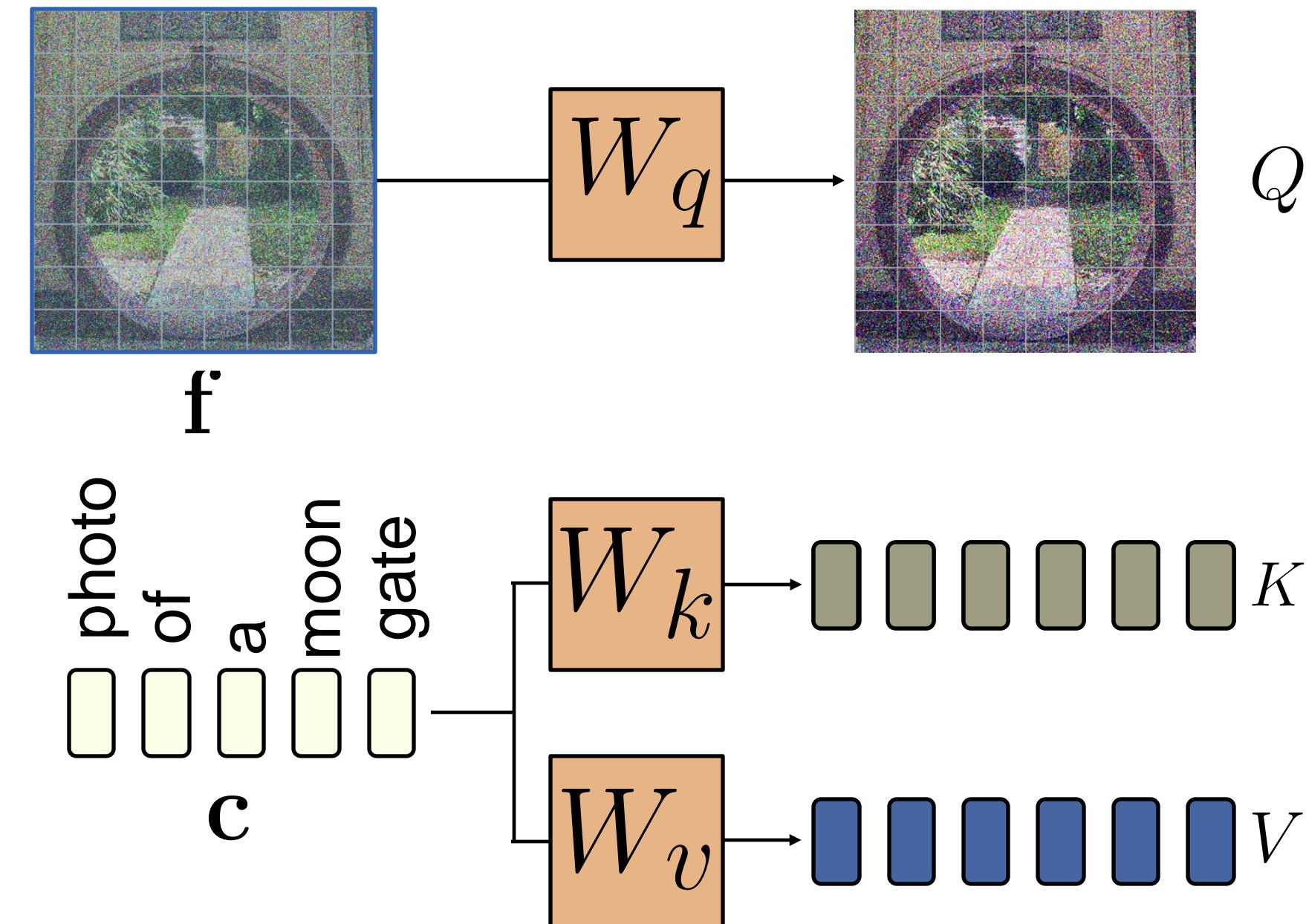


VAES + RNN+ cross-attention

Generating Images from Captions with Attention.

Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, Ruslan Salakhutdinov. ICLR 2016.

# Text-Image Cross-Attention



$$Q \cdot \text{Softmax}(\mathbf{*})$$

$$= (\square, \square, \square, \square, \square, \square)$$

$$= \sum (\square, \square, \square, \square, \square, \square) * (\square, \square, \square, \square, \square, \square)$$

i.e.

$$\text{Output} = \text{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{d'}}\right)V$$

# How could we improve it?

# How could we improve it?

- Better generative modeling techniques.
- Better text encoders.
- Better generator architectures.
- Better ways to connect text and image.
- Bigger data + more GPU/TPU computing.
- Bigger model sizes.

# GANs-based Text-to-Image

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.



# GANs-based Text-to-Image

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

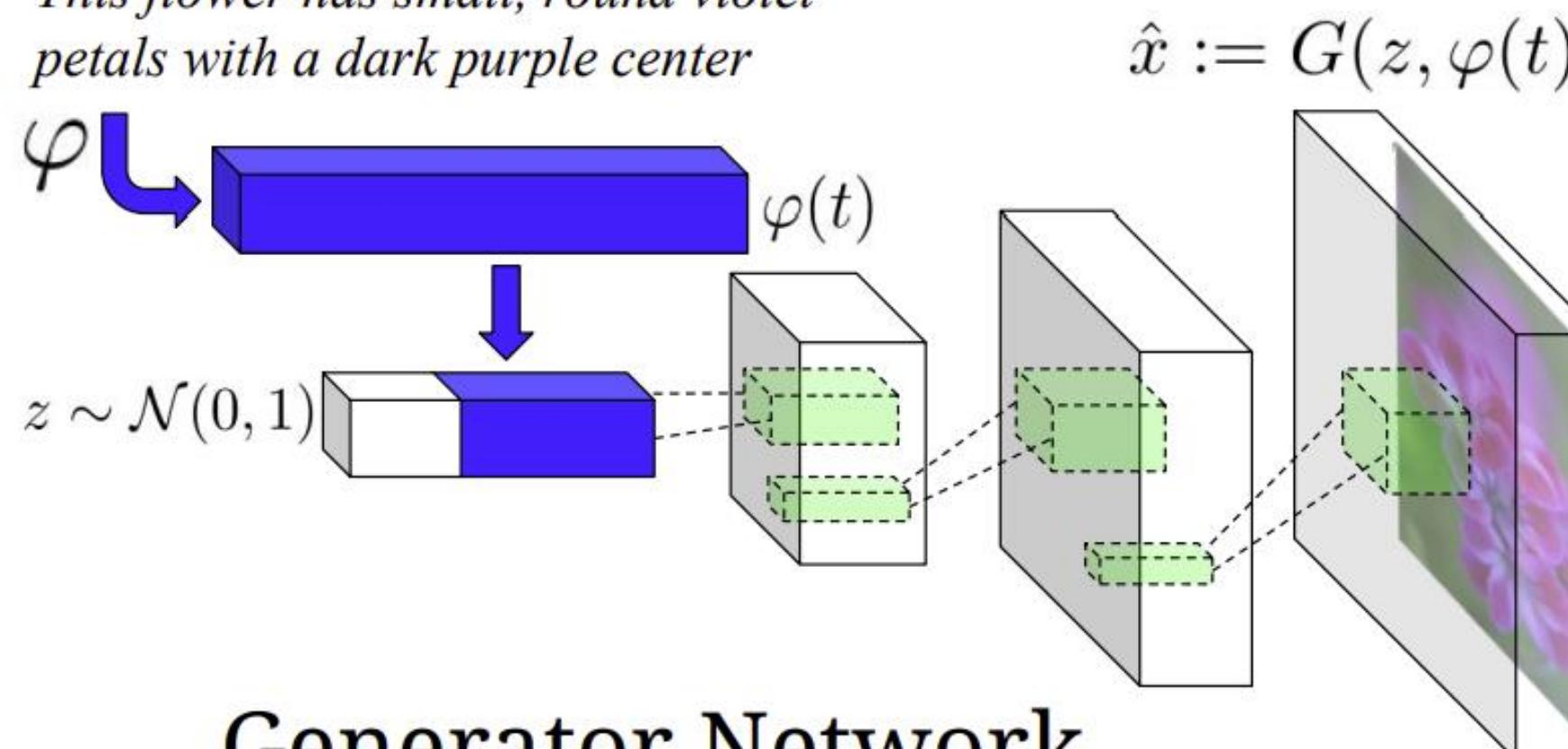


this white and yellow flower  
have thin white petals and a  
round yellow stamen



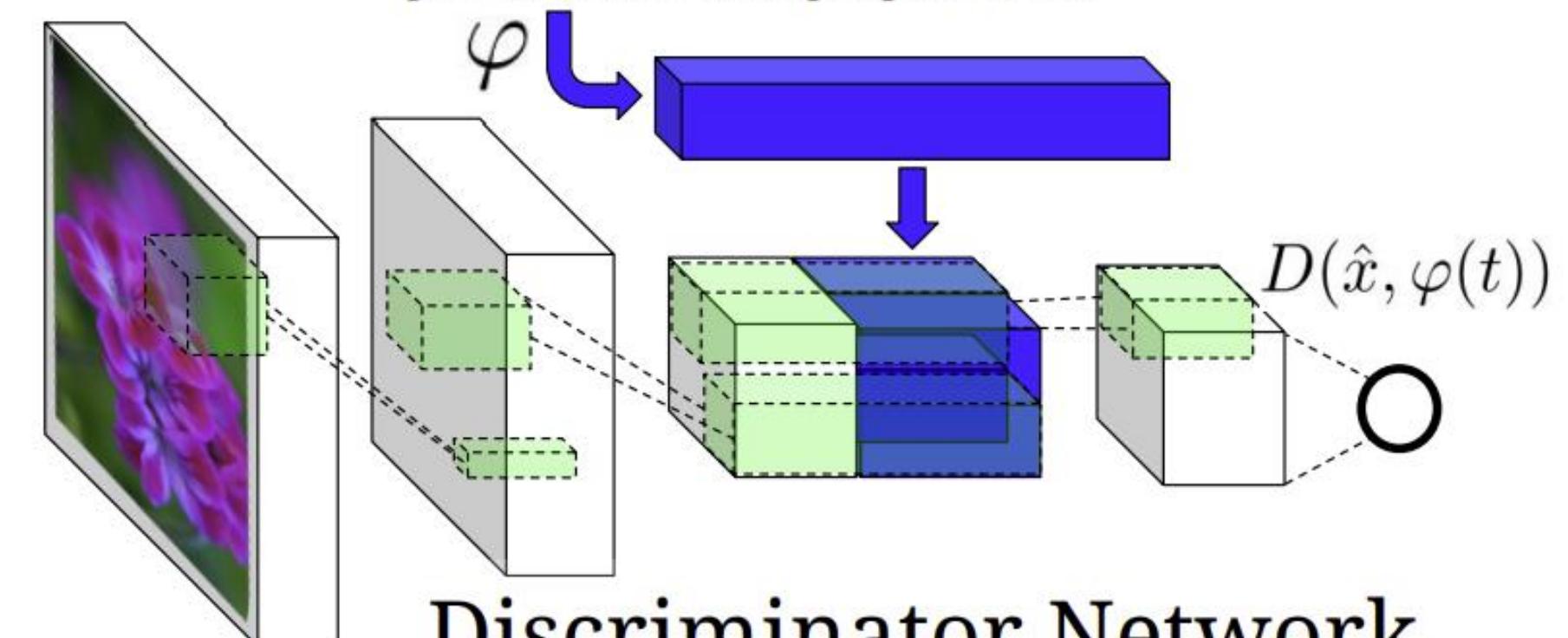
# GANs-based Text-to-Image

*This flower has small, round violet petals with a dark purple center*



Generator Network

*This flower has small, round violet petals with a dark purple center*



Discriminator Network

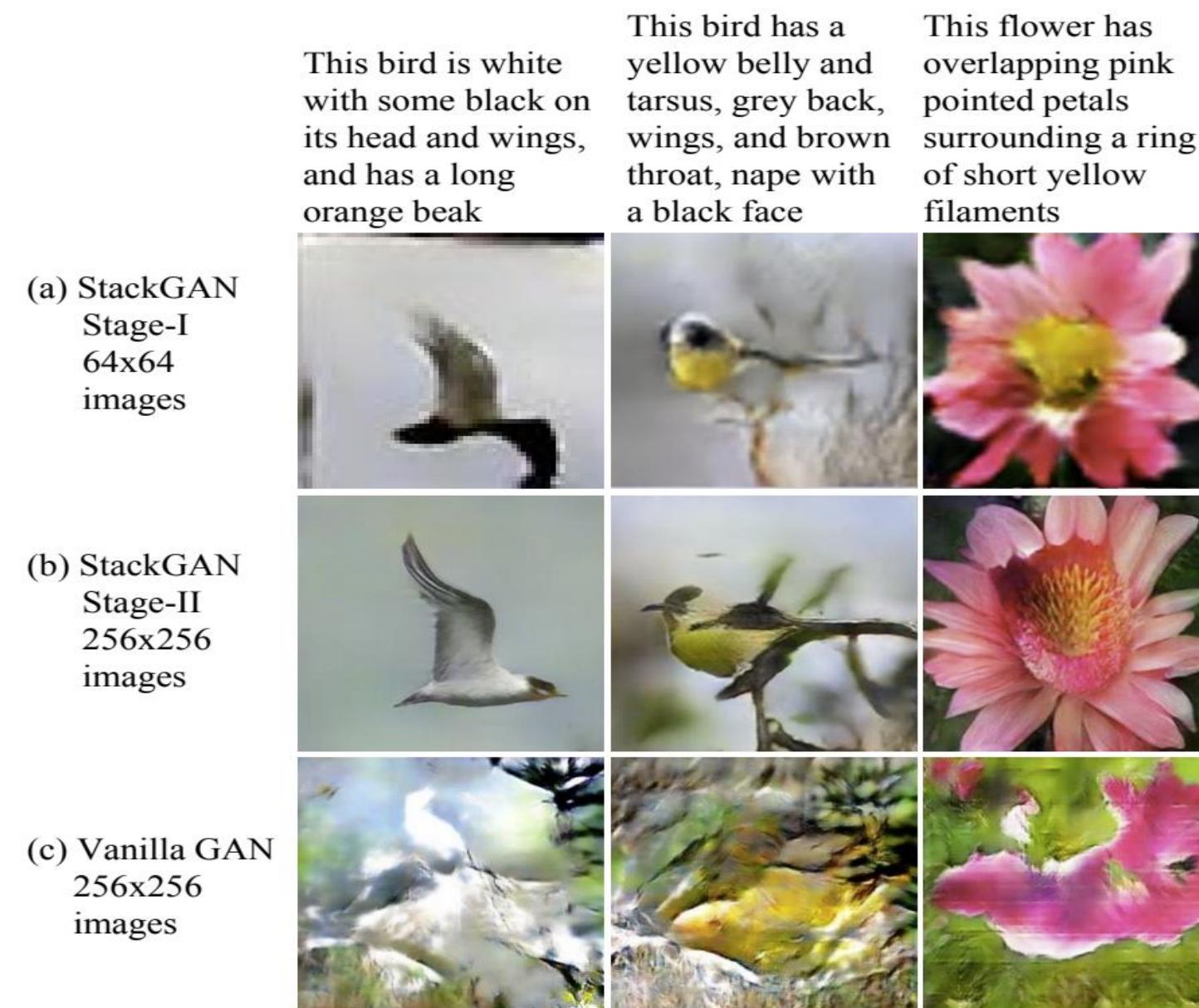
Conditional GAN + CNN + concatenation

Generative Adversarial Text to Image Synthesis

Scott Reed et al., ICML 2016

# How to increase resolution?

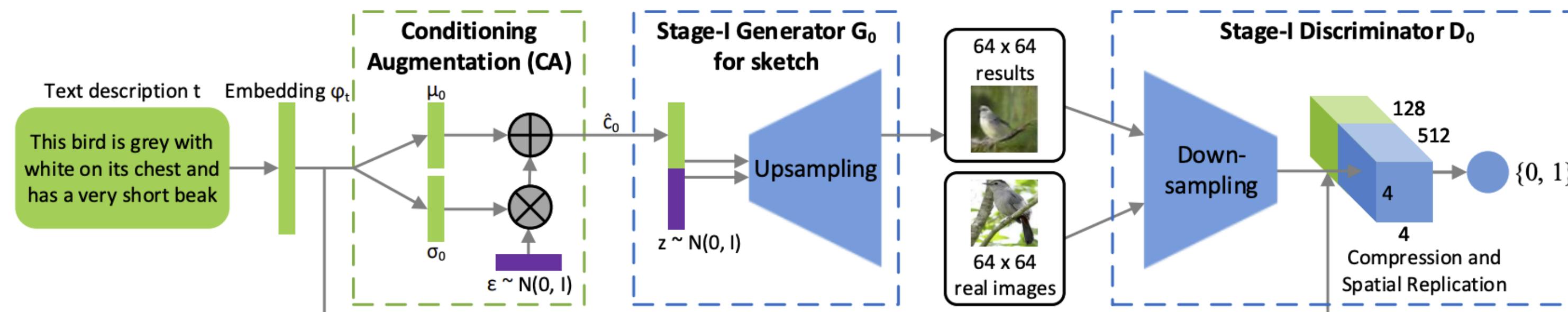
# +Two-stage Models



## Two-stage Conditional GAN + CNN + concatenation

StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks  
Han Zhang et al., ICCV 2017

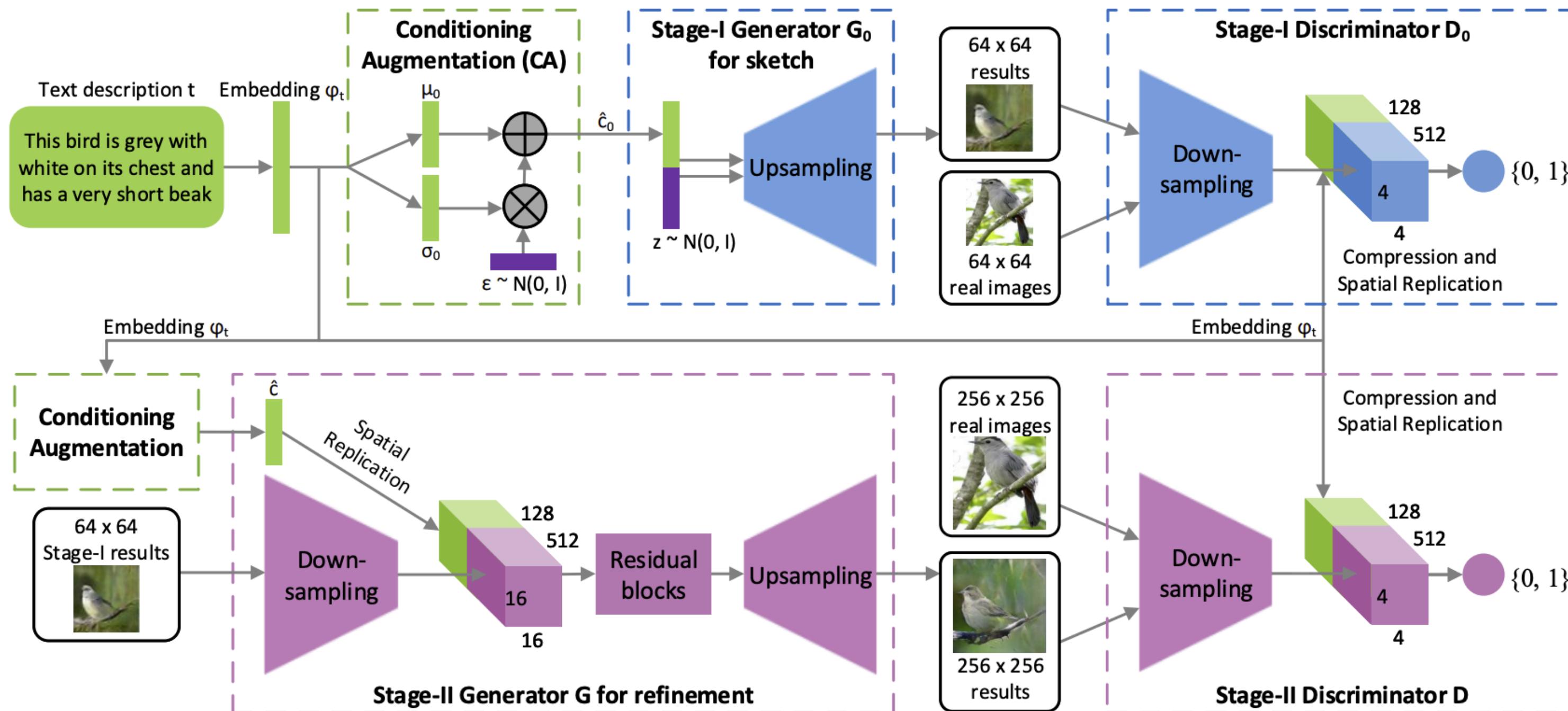
# +Two-stage Models



## Two-stage Conditional GAN + CNN + concatenation

StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks  
Han Zhang et al., ICCV 2017

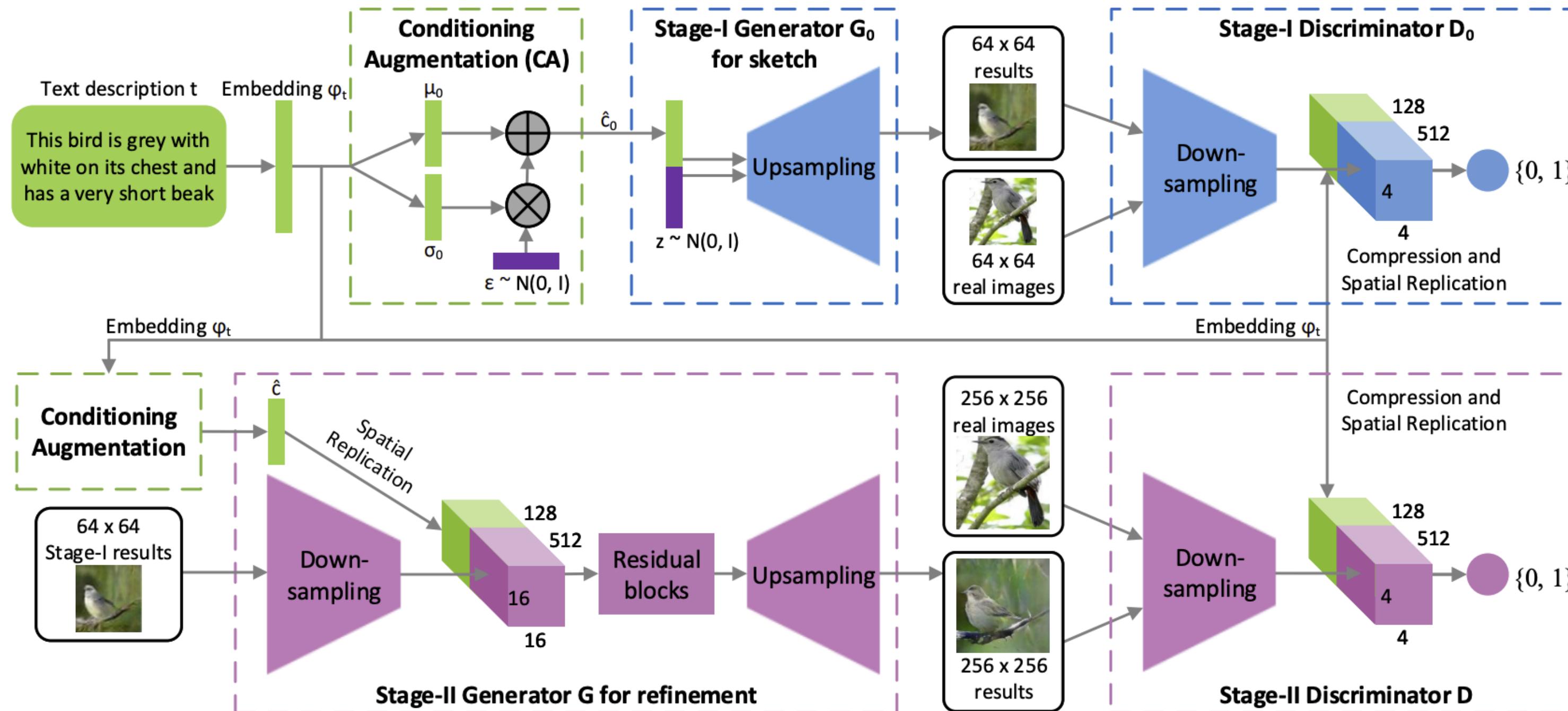
# +Two-stage Models



Two-stage Conditional GAN + CNN + concatenation

StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks  
Han Zhang et al., ICCV 2017

# +Two-stage Models



Two-stage Conditional GAN + CNN + concatenation

StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks  
Han Zhang et al., ICCV 2017

# +Two-stage Models

Text  
description

This flower has  
a lot of small  
purple petals in  
a dome-like  
configuration

This flower is  
pink, white,  
and yellow in  
color, and has  
petals that are  
striped

This flower has  
petals that are  
dark pink with  
white edges  
and pink  
stamen

This flower is  
white and  
yellow in color,  
with petals that  
are wavy and  
smooth

64x64  
GAN-INT-CLS



256x256  
StackGAN



# +Two-stage Models

Text  
description

64x64  
GAN-INT-CLS



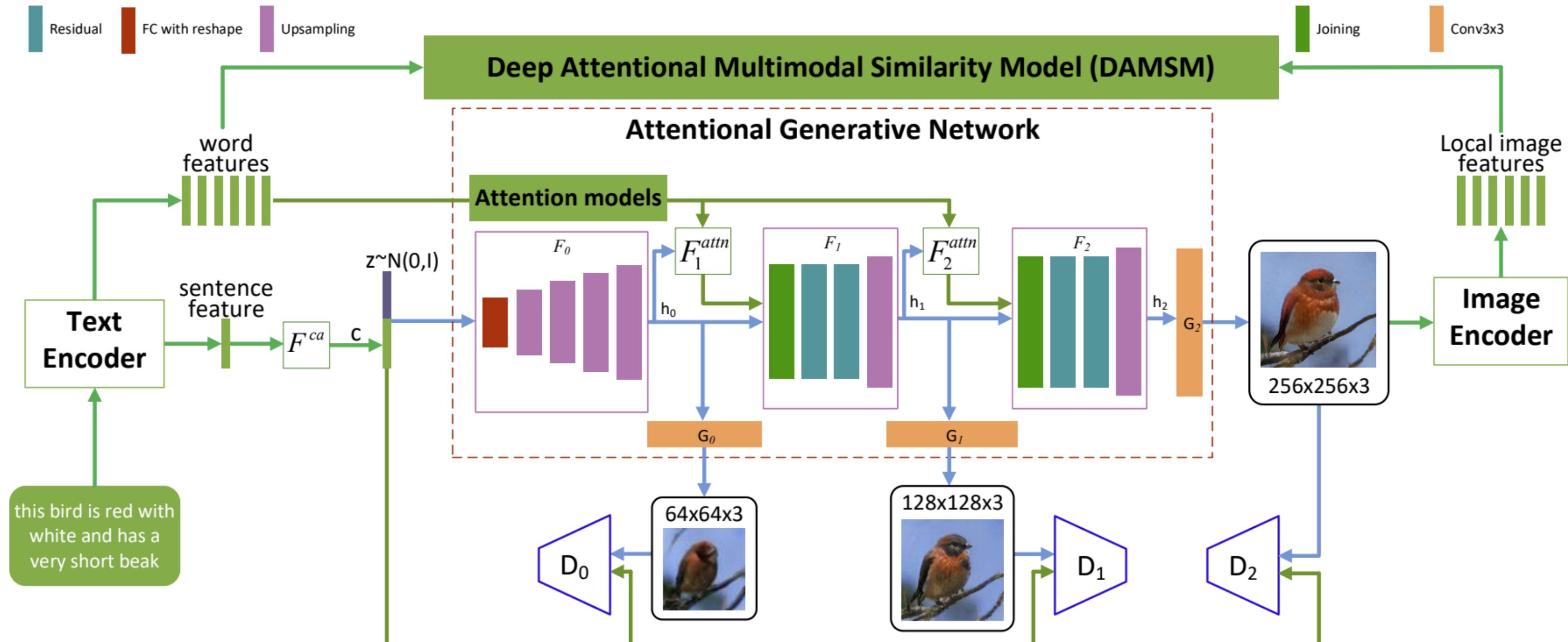
256x256  
StackGAN



# + Cross-attention to connect Text and Image



# + Cross-attention to connect Text and Image



AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks  
Tao Xu et al., CVPR 2018

Got Stuck in 2018-2020  
(Birds, MS COCO)

# Who shall we blame?

- **Better generative modeling techniques: VAEs, GANs?**
- **Better text encoders: LSTM/RNN?**
- **Better generator architectures: CNNs?**
- Better ways to connect text and image.
- Bigger data + more GPU/TPU computing.
- Bigger model sizes.

How could we synthesize images  
beyond single or a few categories

# Taming Transformers for High-Resolution Image Synthesis

Patrick Esser\*    Robin Rombach\*    Björn Ommer

Heidelberg Collaboratory for Image Processing, IWR, Heidelberg University, Germany  
\*Both authors contributed equally to this work



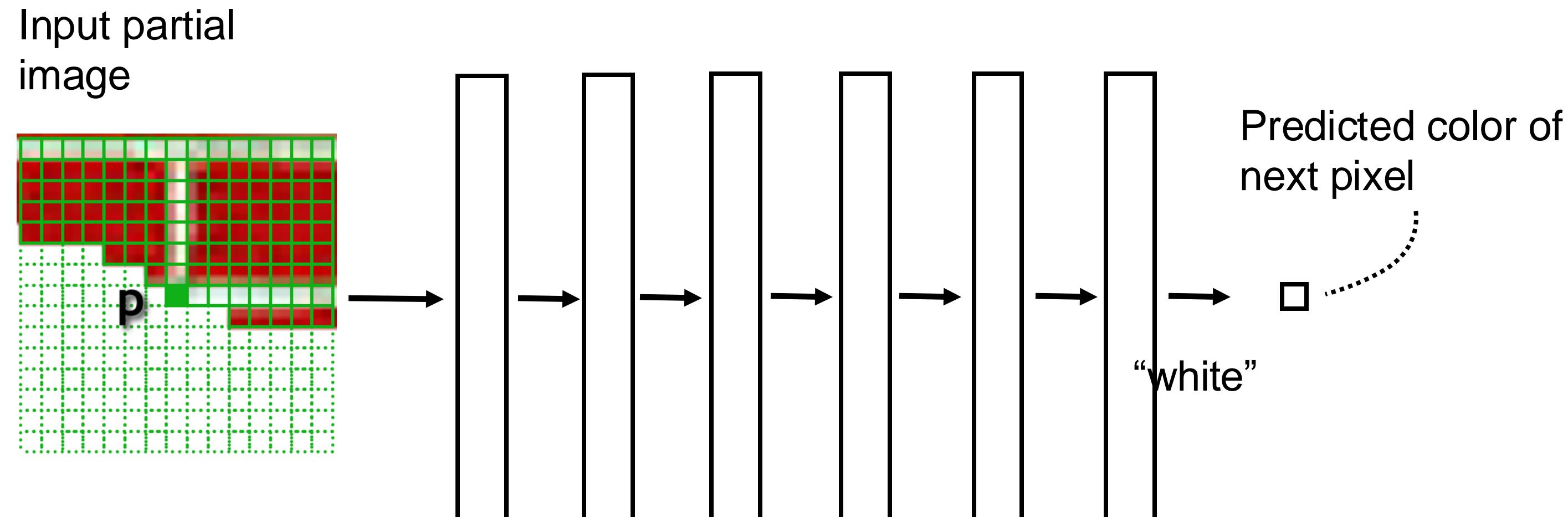
Figure 1. Our approach enables transformers to synthesize high-resolution images like this one, which contains 1280x460 pixels.

## Abstract

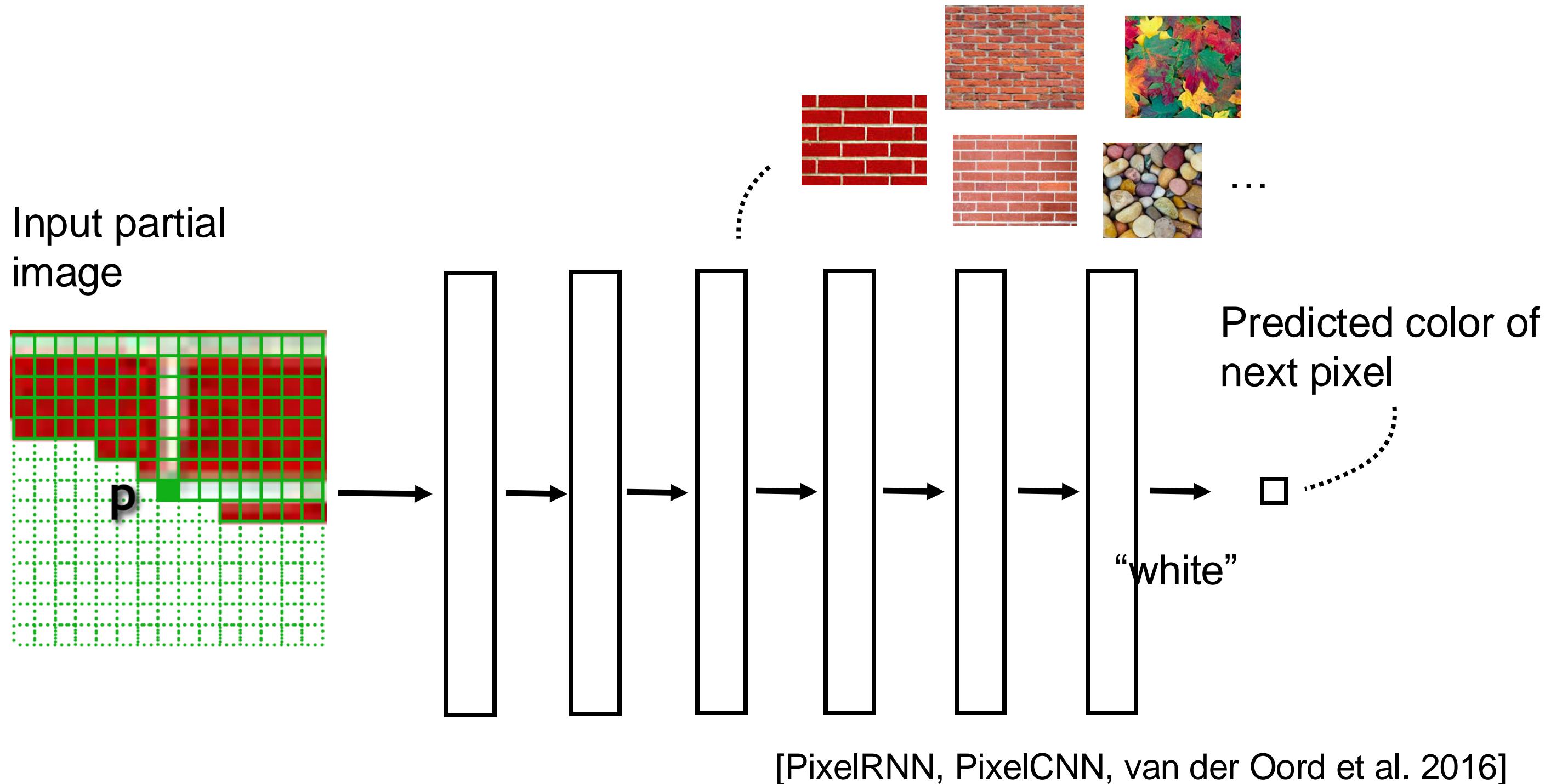
Designed to learn long-range interactions on sequential inputs, convolutional neural networks (CNNs) have become the state-of-the-art results

and are increasingly adapted in other areas such as audio [12] and vision [8, 16]. In contrast to the predominant vision architecture, convolutional neural networks (CNNs), the transformer architecture contains no built-in inductive bias with respect to the locality of interactions and is therefore free from the need to explicitly define receptive fields for inputs. However,

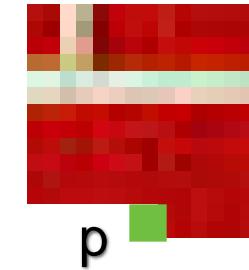
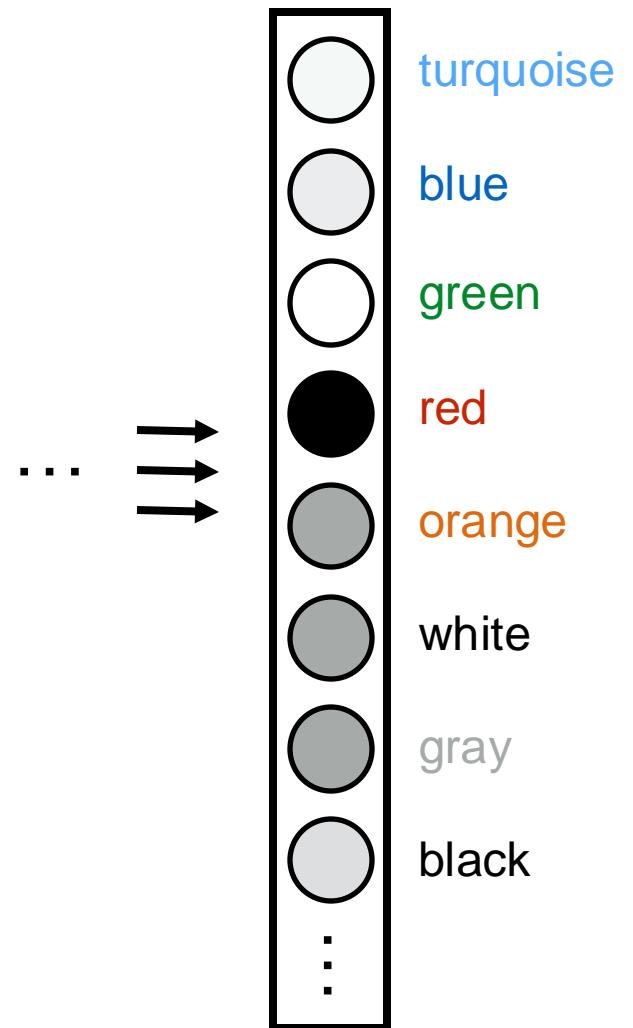
# Autoregressive (AR) image synthesis



[PixelRNN, PixelCNN, van der Oord et al. 2016]

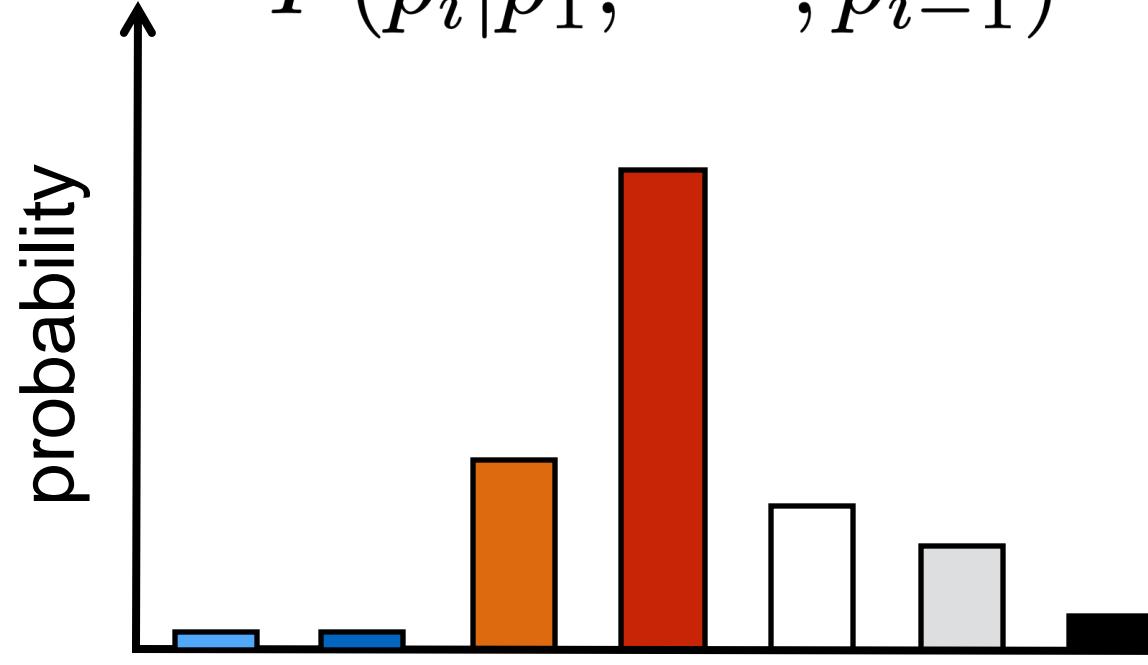


## Network output

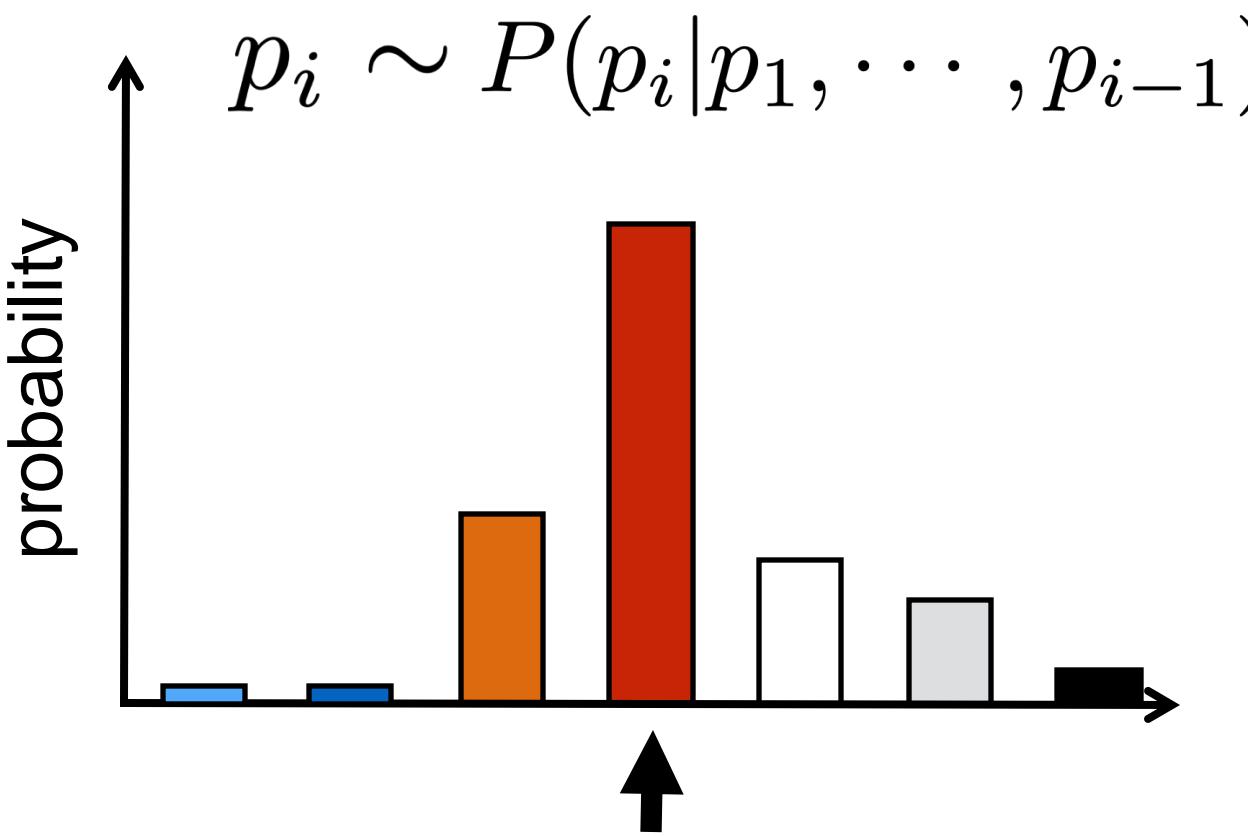
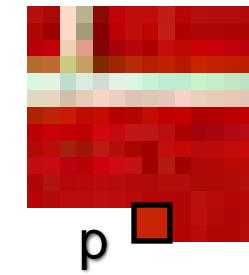
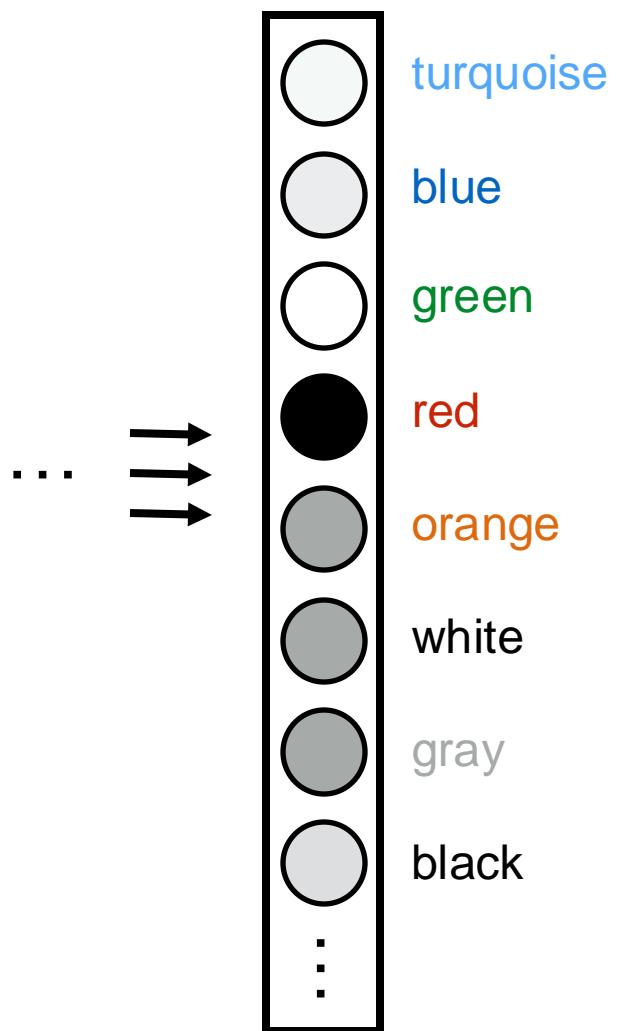


$P(\text{next pixel} \mid \text{previous pixels})$

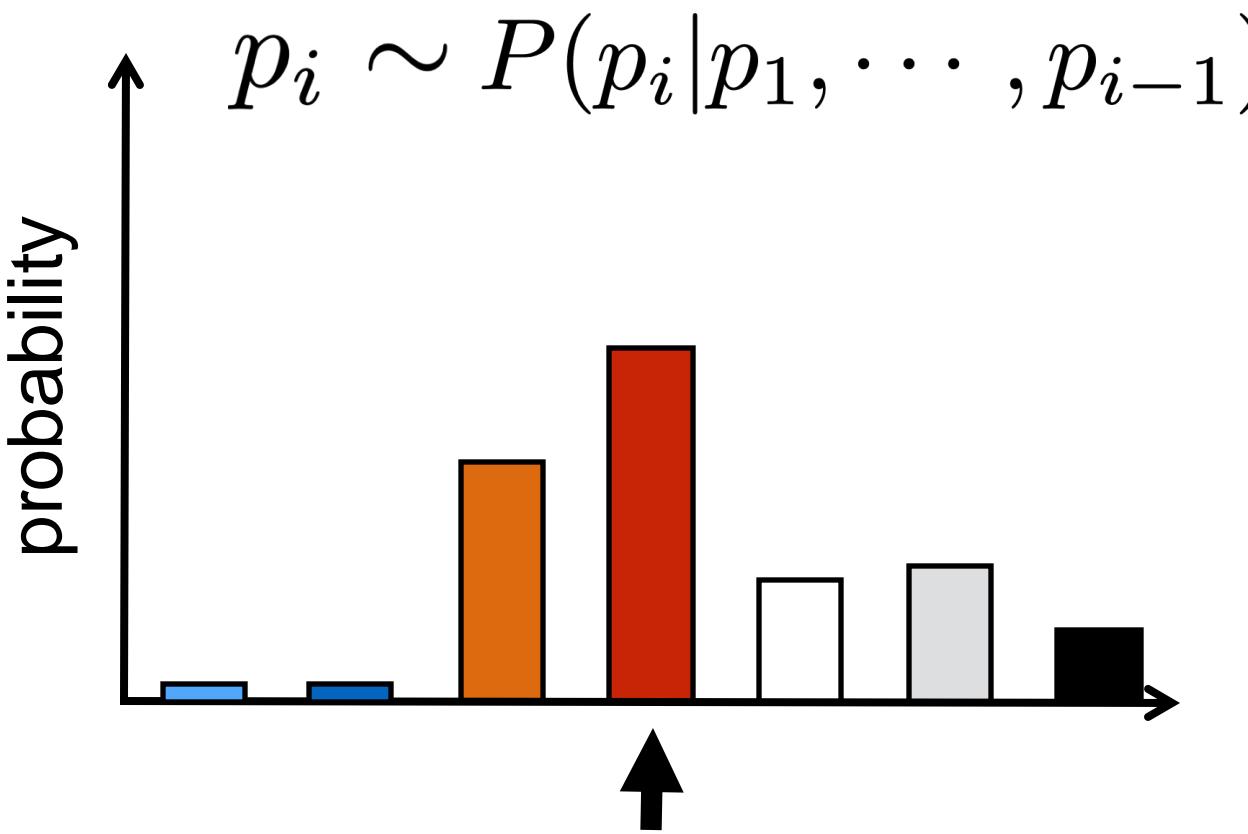
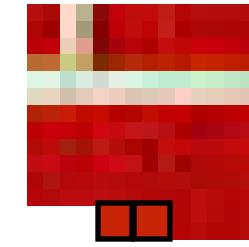
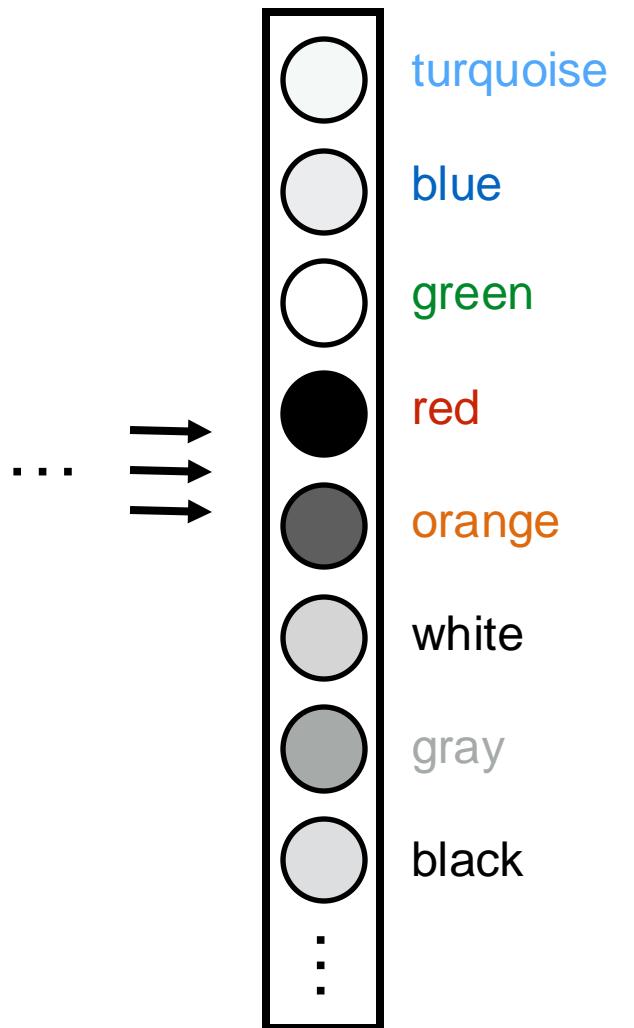
$$P(p_i | p_1, \dots, p_{i-1})$$



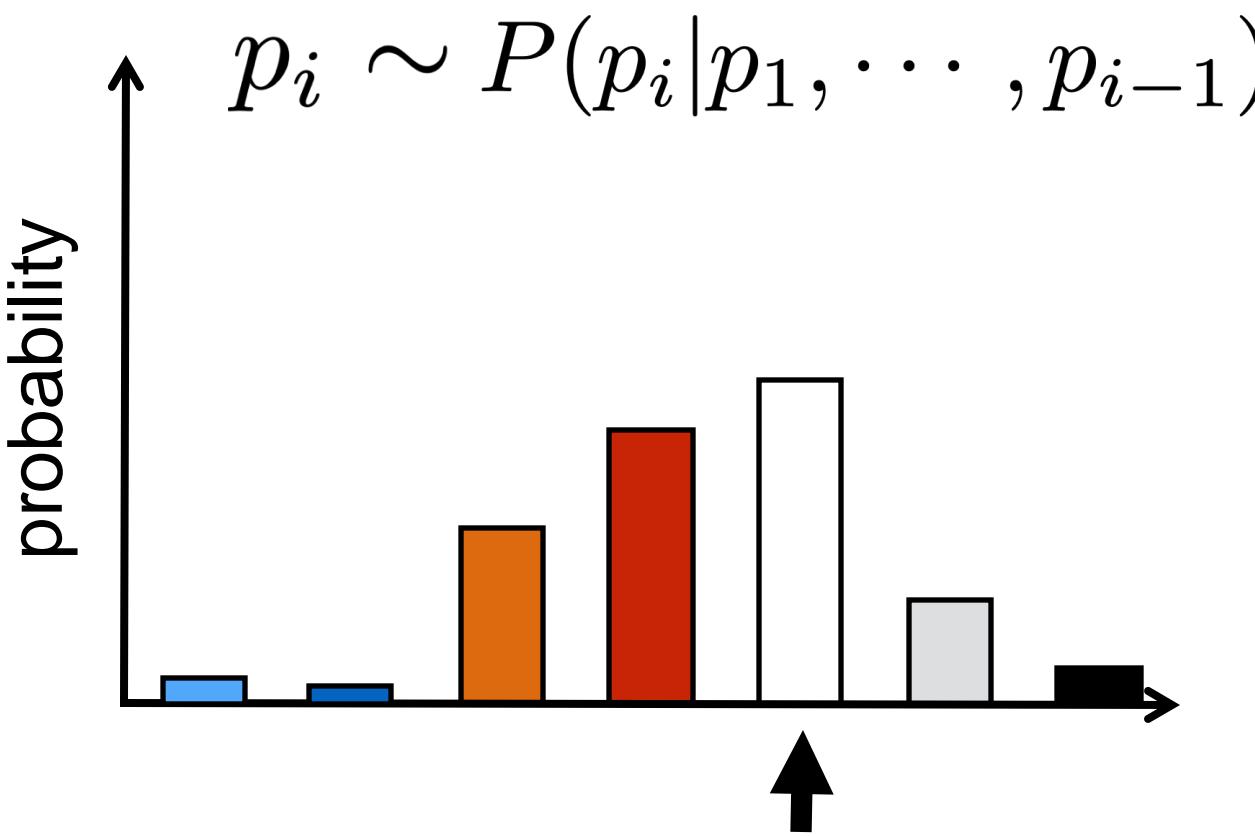
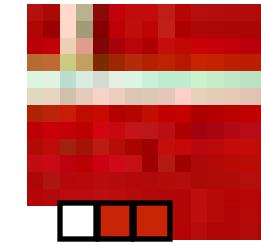
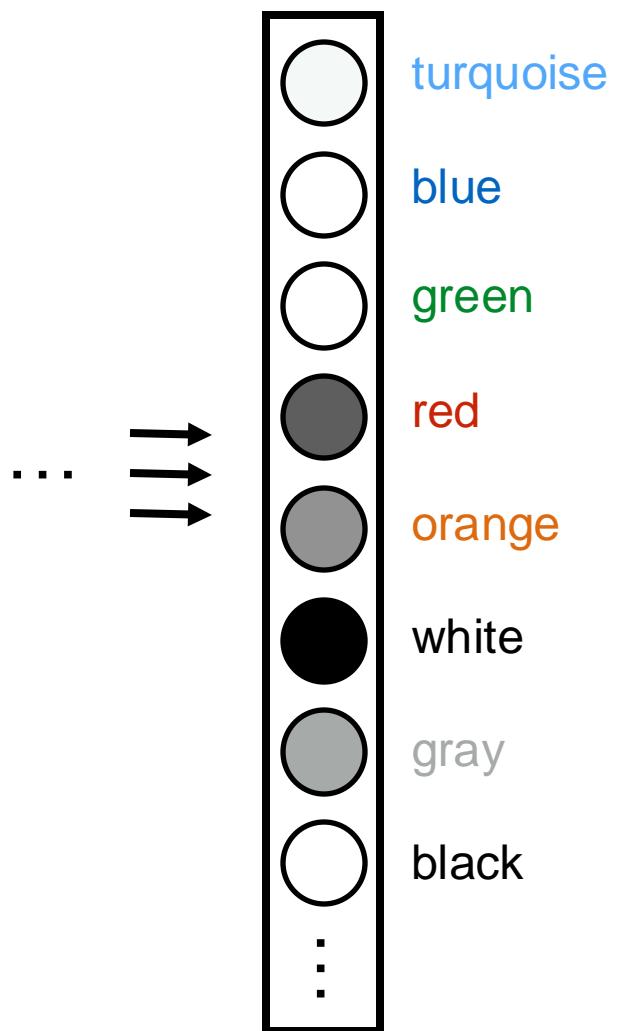
## Network output



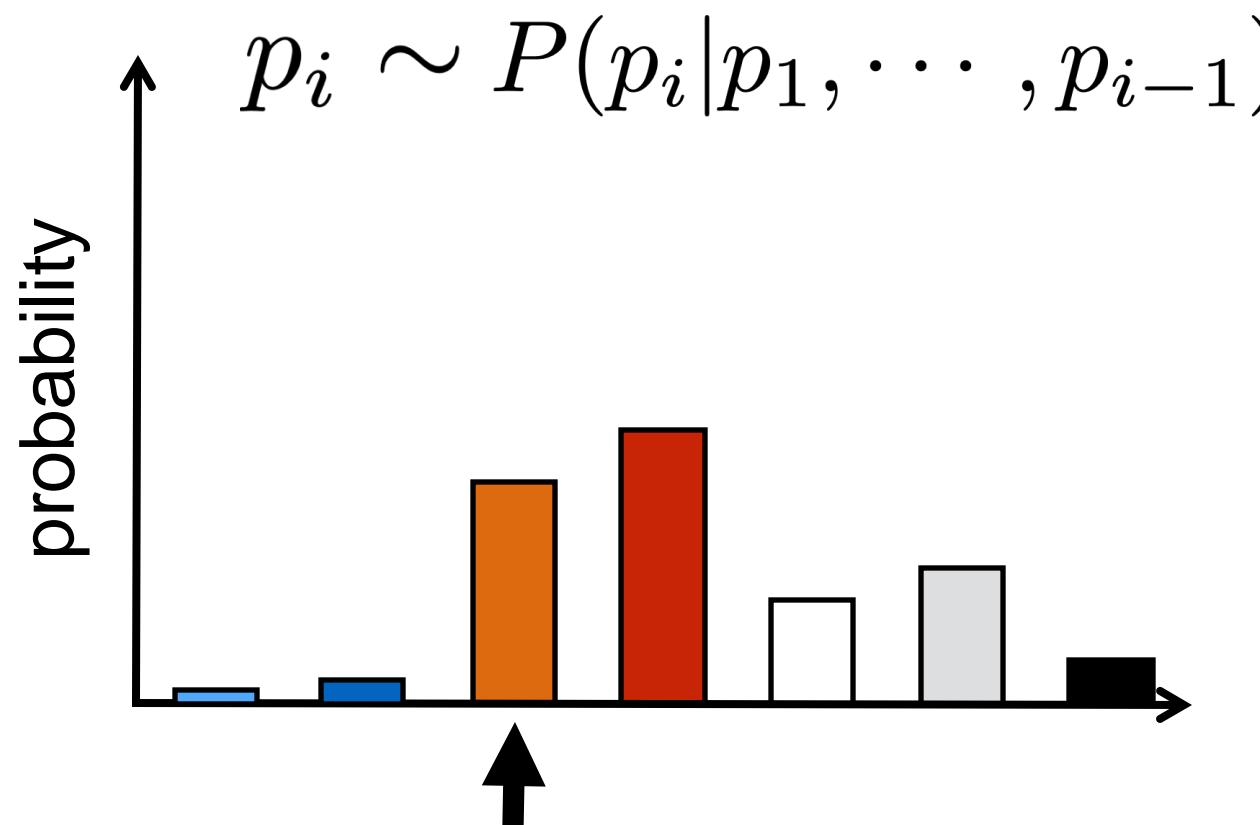
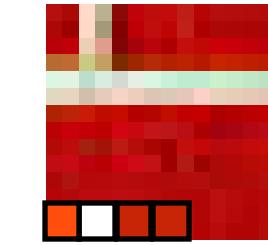
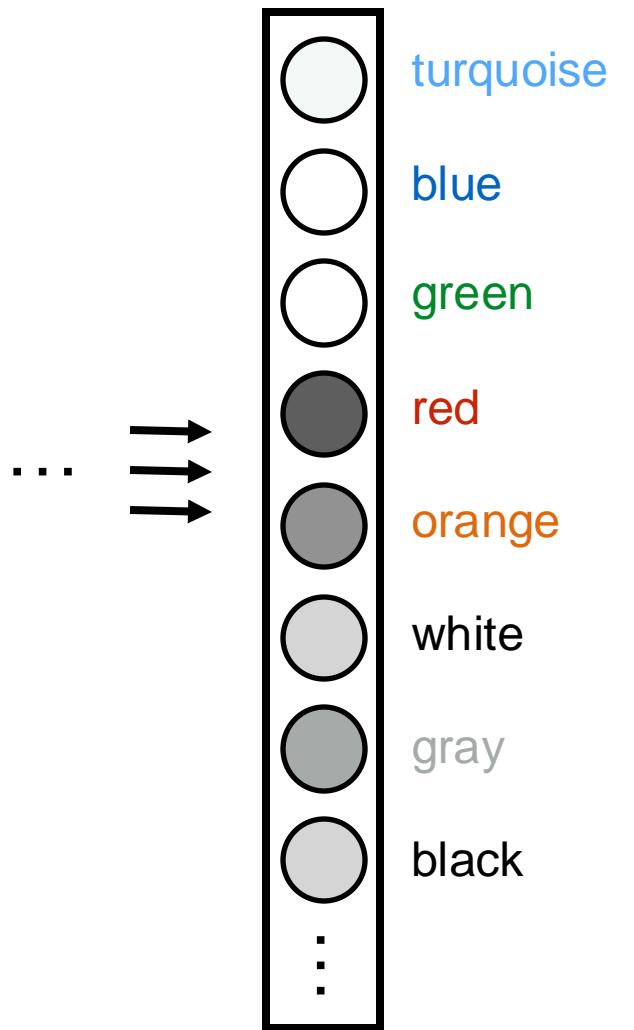
## Network output



## Network output

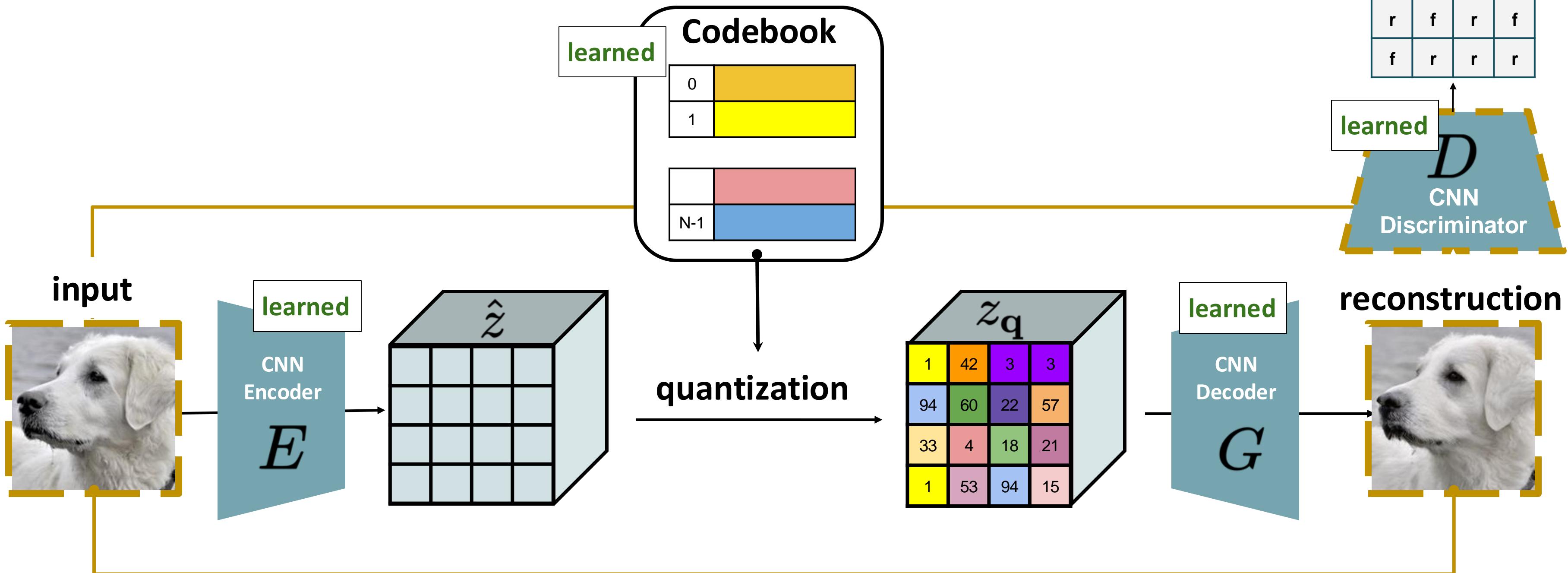


## Network output



# From VQ-VAE<sup>1</sup> to VQGAN

<sup>1</sup>: Neural Discrete Representation Learning, v.d.Oord et al, <https://arxiv.org/abs/1711.00937>

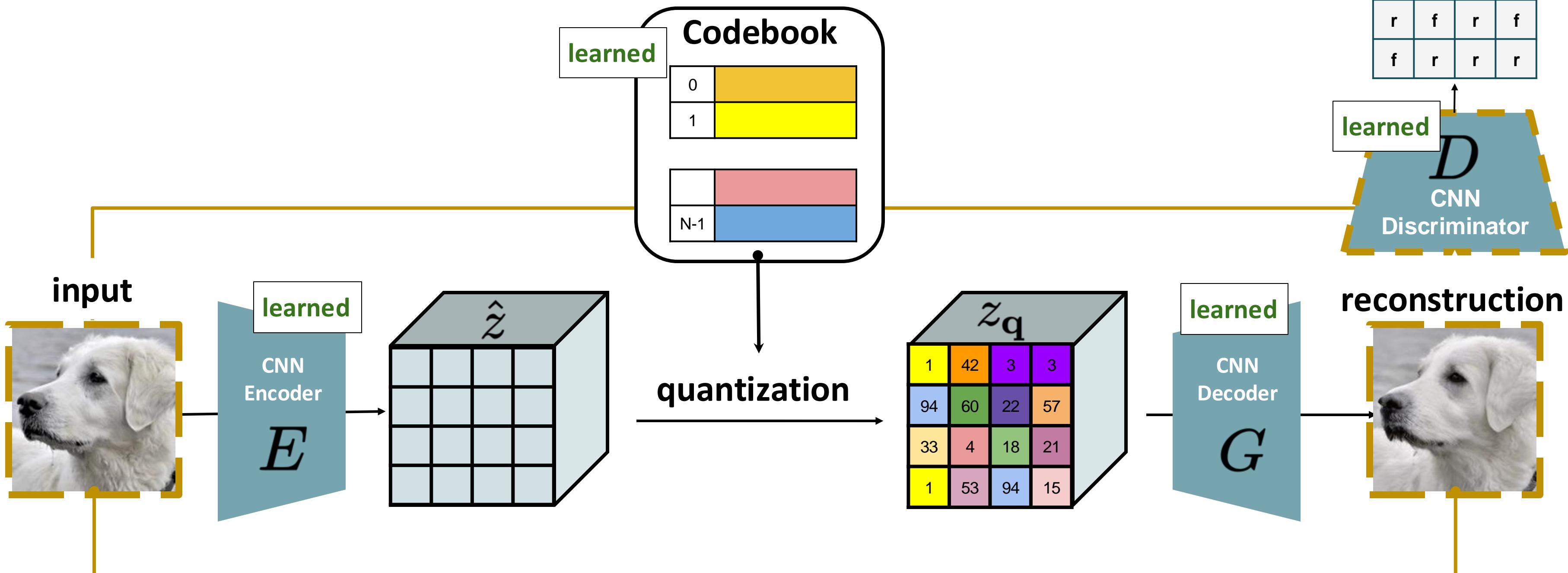


i) replace L2/L1 rec. loss with Perceptual loss (includes pixel-level)

ii) add (patch-wise) Discriminator to favor realism over perfect reconstruction

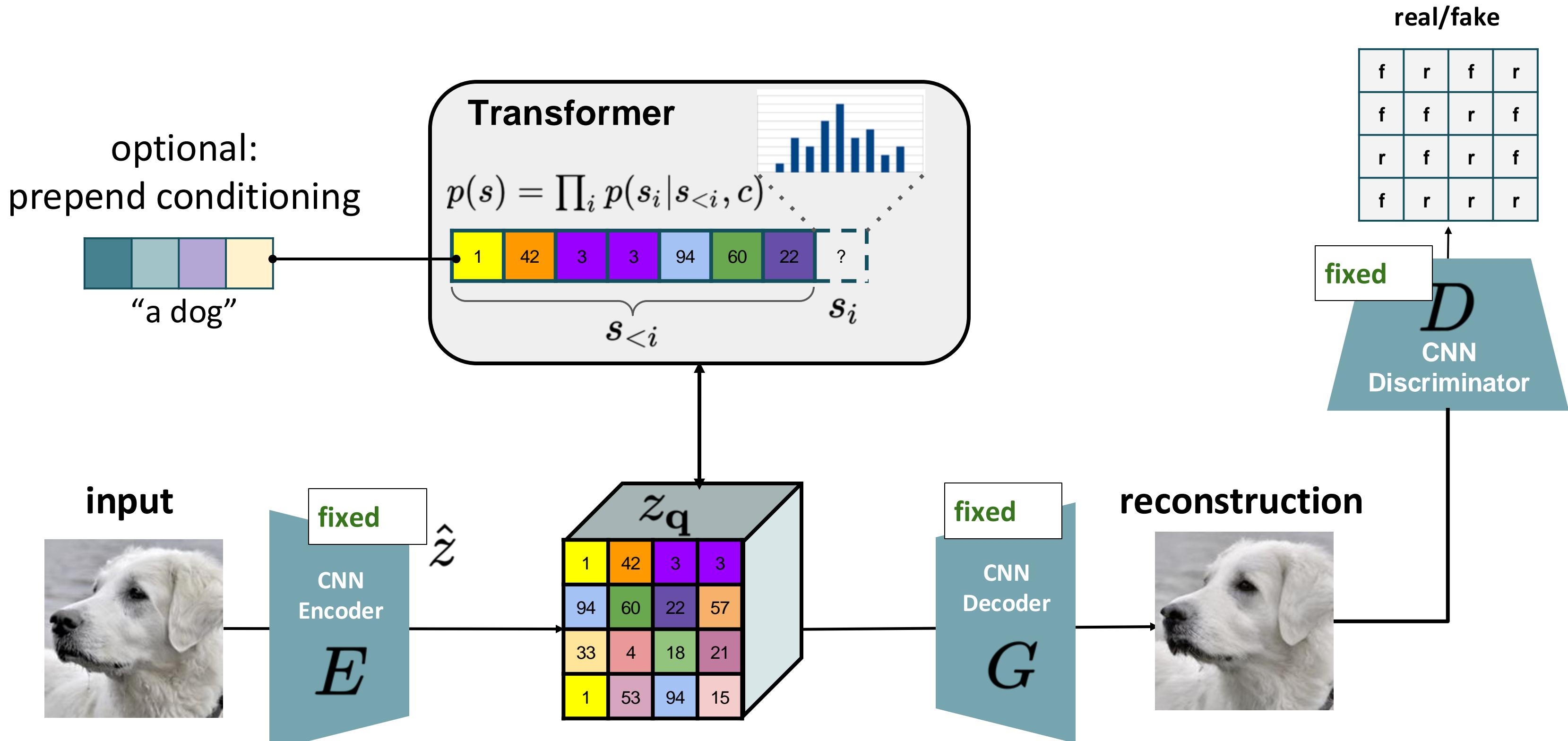
# From VQ-VAE<sup>1</sup> to VQGAN

<sup>1</sup>: Neural Discrete Representation Learning, v.d.Oord et al, <https://arxiv.org/abs/1711.00937>



$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{VQ}} + \lambda \mathcal{L}_{\text{GAN}} \text{ where } \lambda = \frac{\nabla_{G_L} [\mathcal{L}_{\text{rec}}]}{\nabla_{G_L} [\mathcal{L}_{\text{GAN}}] + \delta}$$

# Transformer Training





Slide credit: Robin Rombach

# Scaling VQGAN for Text-to-Image!

- see recently released “Parti” paper by Google (text-to-image model)
  - <https://parti.research.google/>

350M



750M



3B



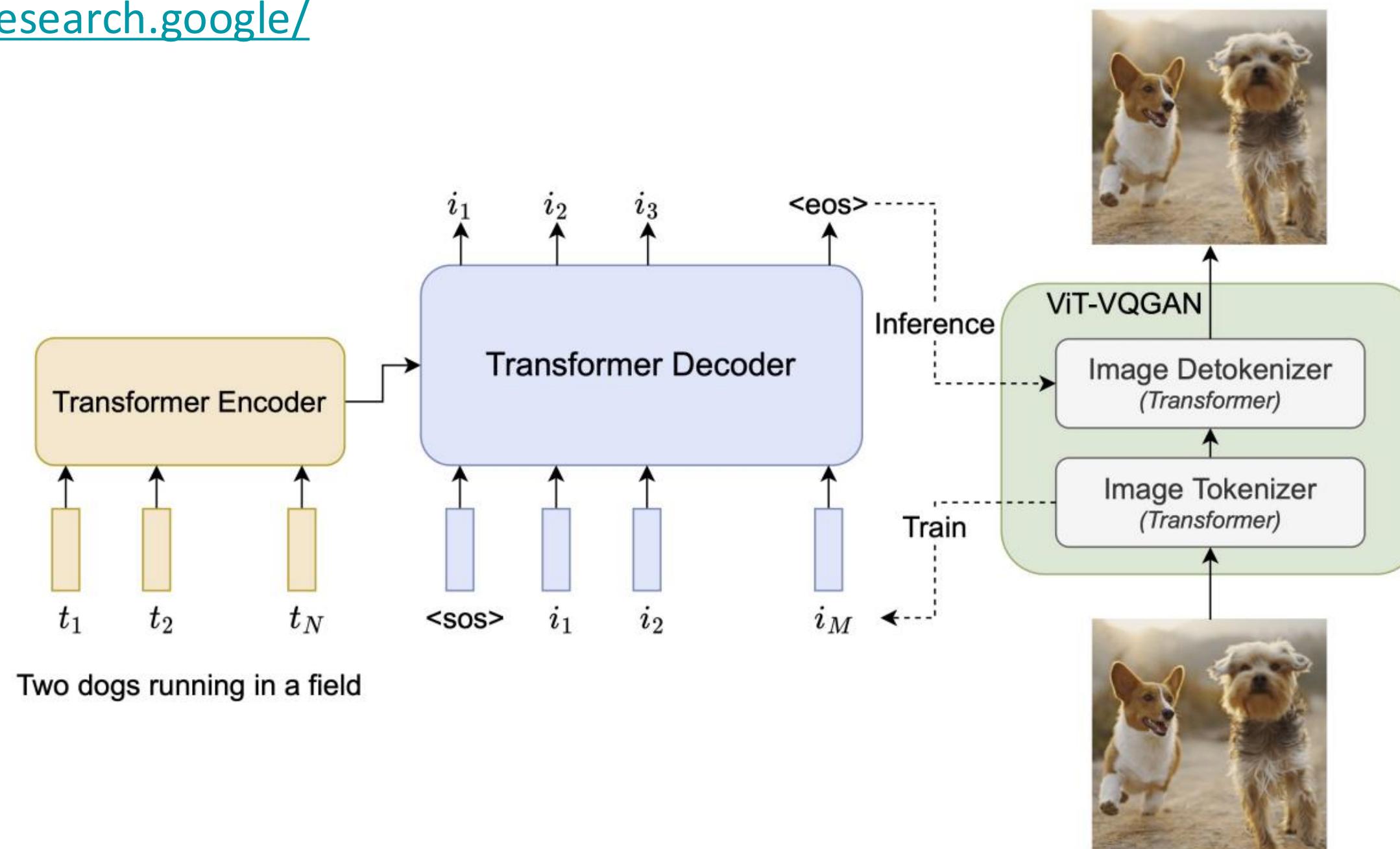
20B



*A portrait photo of a kangaroo wearing an orange hoodie and blue sunglasses standing on the grass in front of the Sydney Opera House holding a sign on the chest that says Welcome Friends!*

# Scaling VQGAN for Text-to-Image!

- see recently released “Parti” paper by Google (text-to-image model)
  - <https://parti.research.google/>



Transformer-based Encoder/Decoder + Transformer-based Autoregressive models

# Another Approach: Diffusion Models!

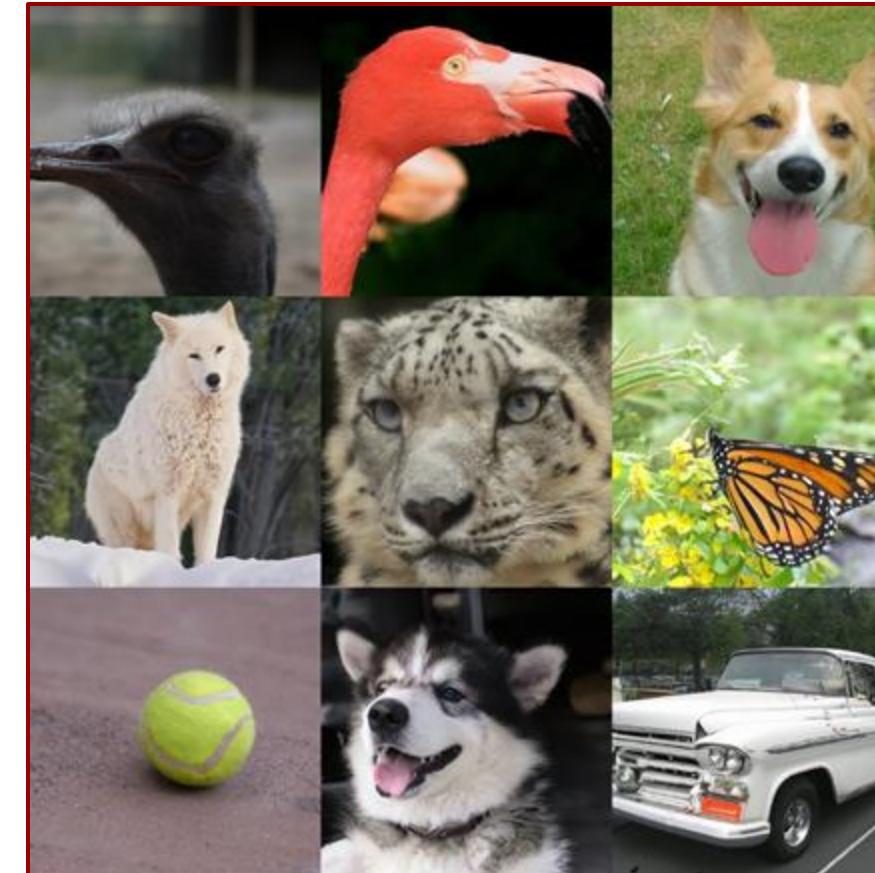
great results for image synthesis



*Denoising Diffusion Probabilistic Models*

*Jonathan Ho, Ajay Jain, et al*

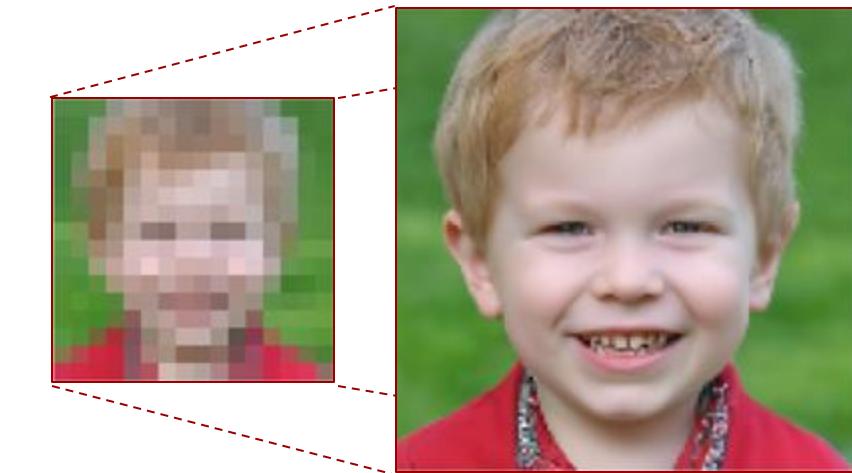
<https://arxiv.org/abs/2006.11239>



*Diffusion Models beat GANs on Image Synthesis*

*Prafulla Dhariwal, Alex Nichol*

<https://arxiv.org/abs/2105.05233>



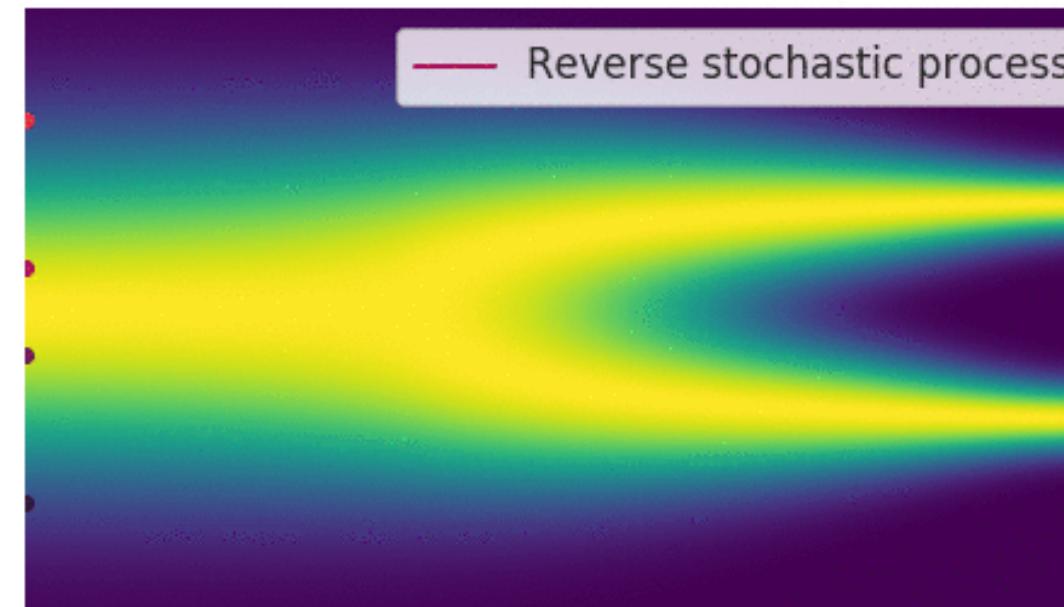
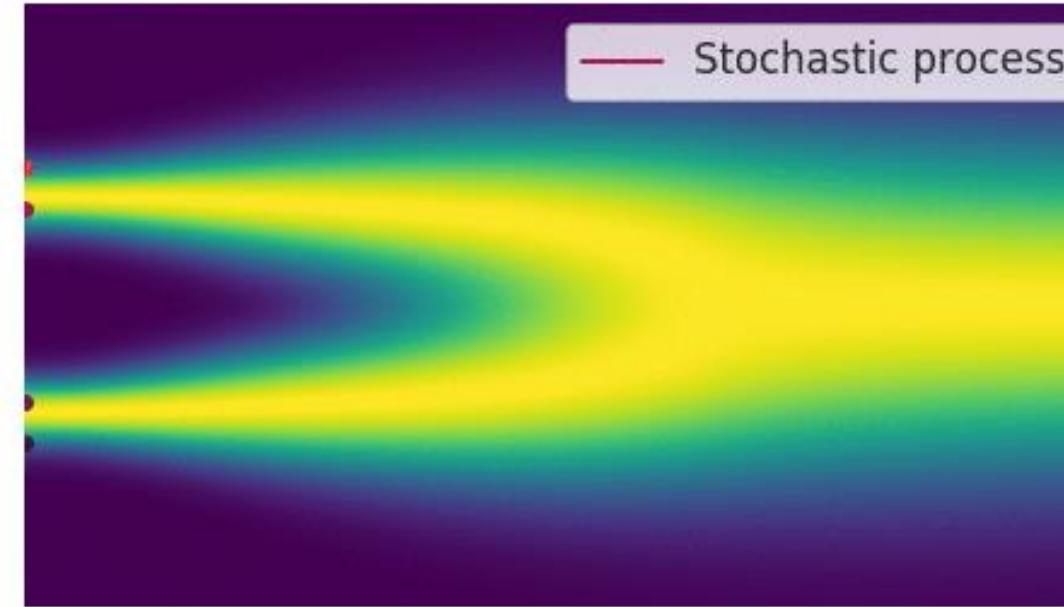
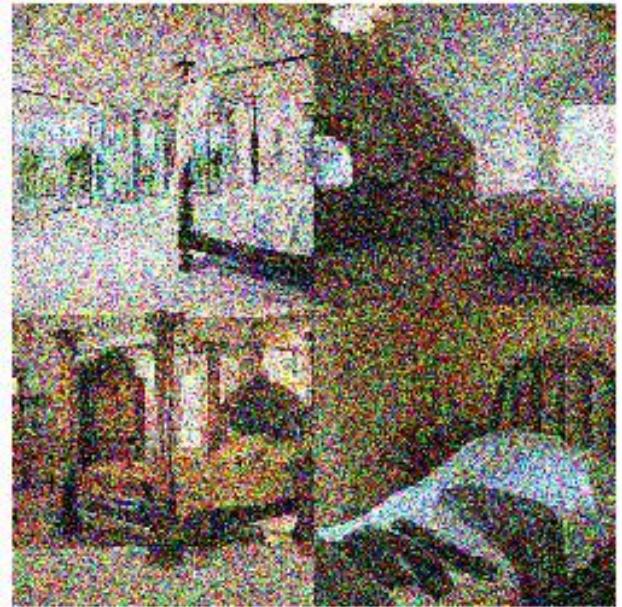
*Image Super-Resolution via Iterative Refinement*

*Chitwan Saharia, et al*

<https://arxiv.org/abs/2104.07636>

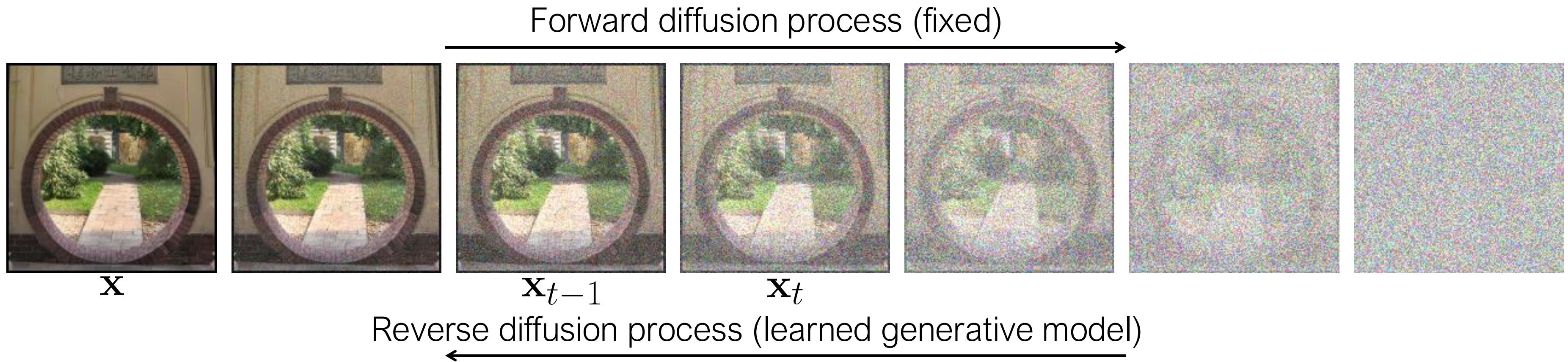
... but very expensive :(

# Brief Overview of Diffusion Models

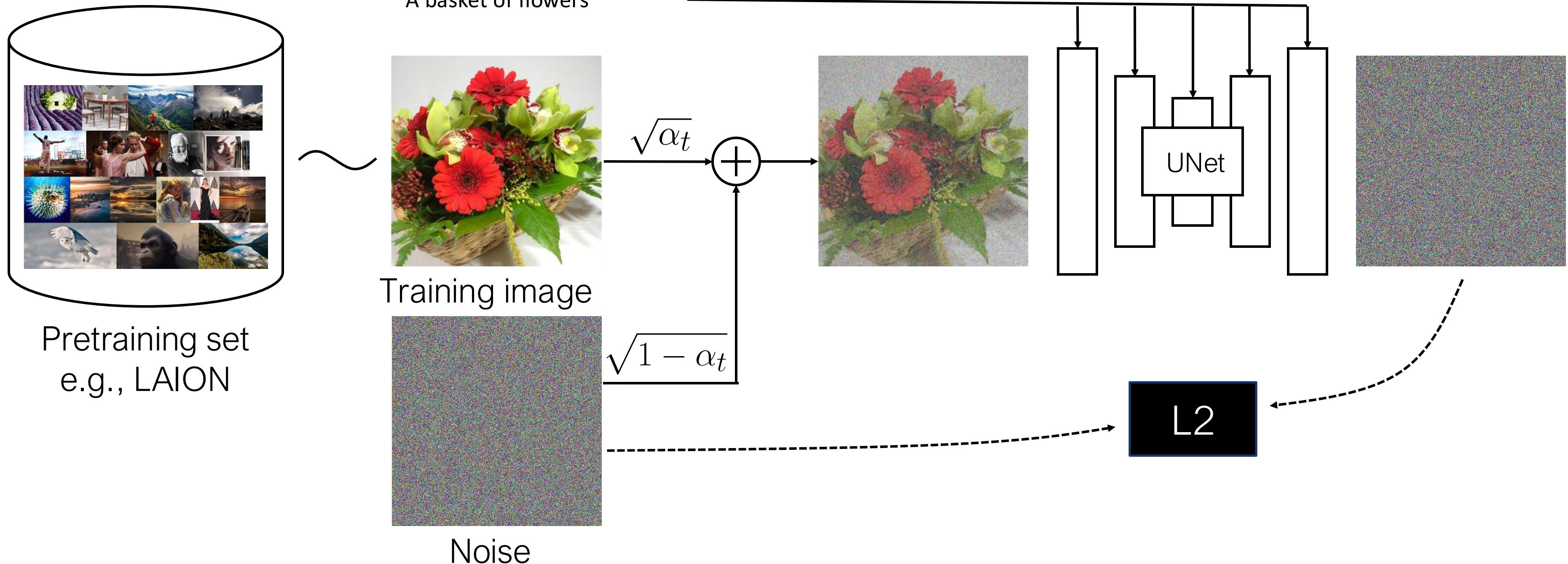


- “destroy” the data by gradually adding small amounts of gaussian noise
- “create” data by gradually denoising a noisy code from a stationary distribution

# Diffusion model inference



# Diffusion model training



\*slides credit: from custom-diffusion

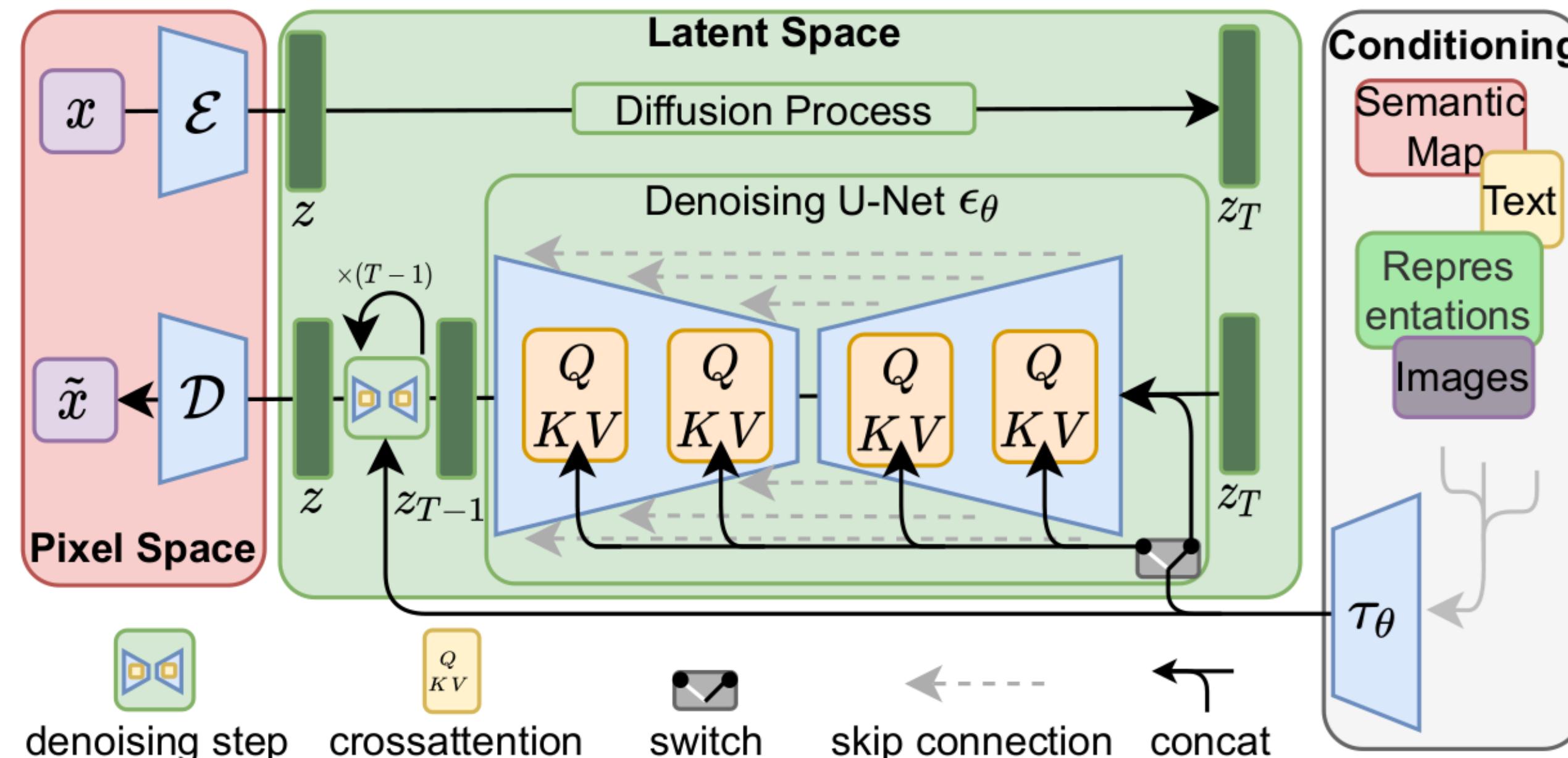
# Latent Diffusion Modeling: Architecture

Autoencoder with KL or VQ regularization.

$$\text{VQ-reg.: } \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{VQ}} + \lambda \mathcal{L}_{\text{GAN}}$$

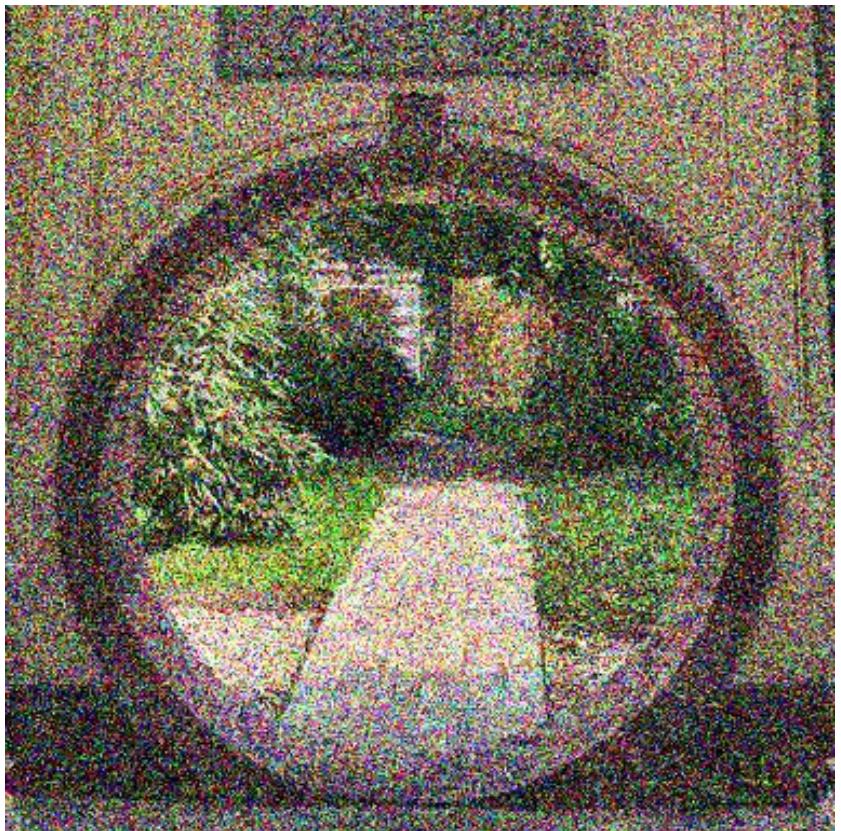
$$\text{where } \lambda = \frac{\nabla_{G_L}[\mathcal{L}_{\text{rec}}]}{\nabla_{G_L}[\mathcal{L}_{\text{GAN}}] + \delta}$$

$$\text{KL-reg.: } \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \beta \mathcal{L}_{\text{KL}} + \lambda \mathcal{L}_{\text{GAN}}$$



# Diffusion Model Architecture

photo  
of  
a  
moon  
gate



→

Text  
transformer

ResNet

Self Cross

Attention

⋮ ⋮ ⋮

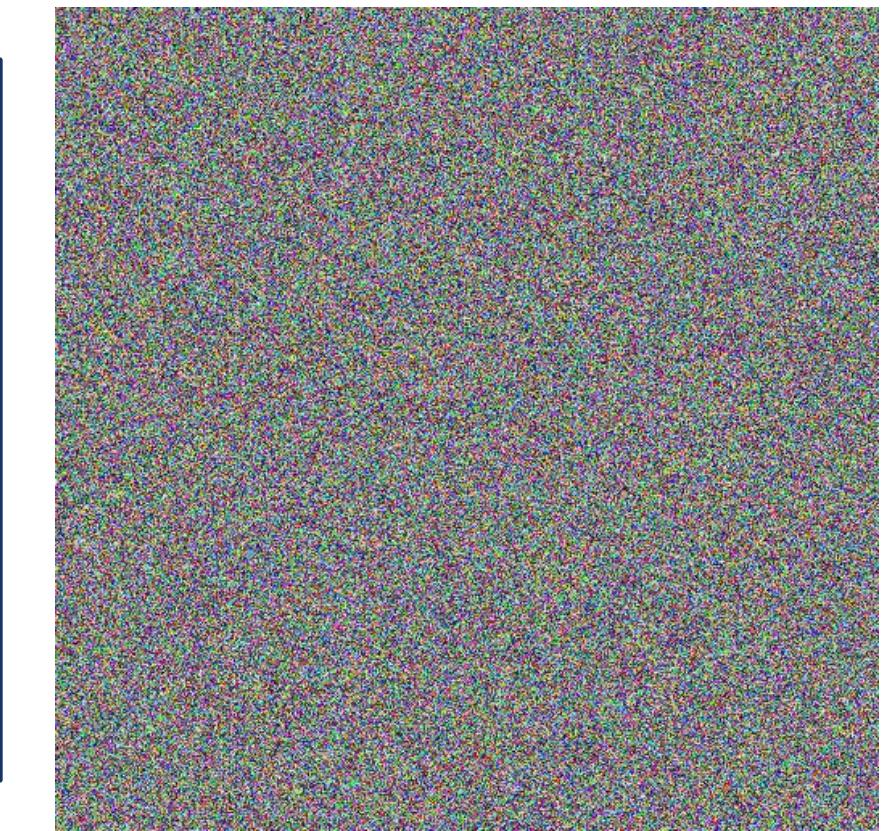
ResNet

Self Cross

Attention

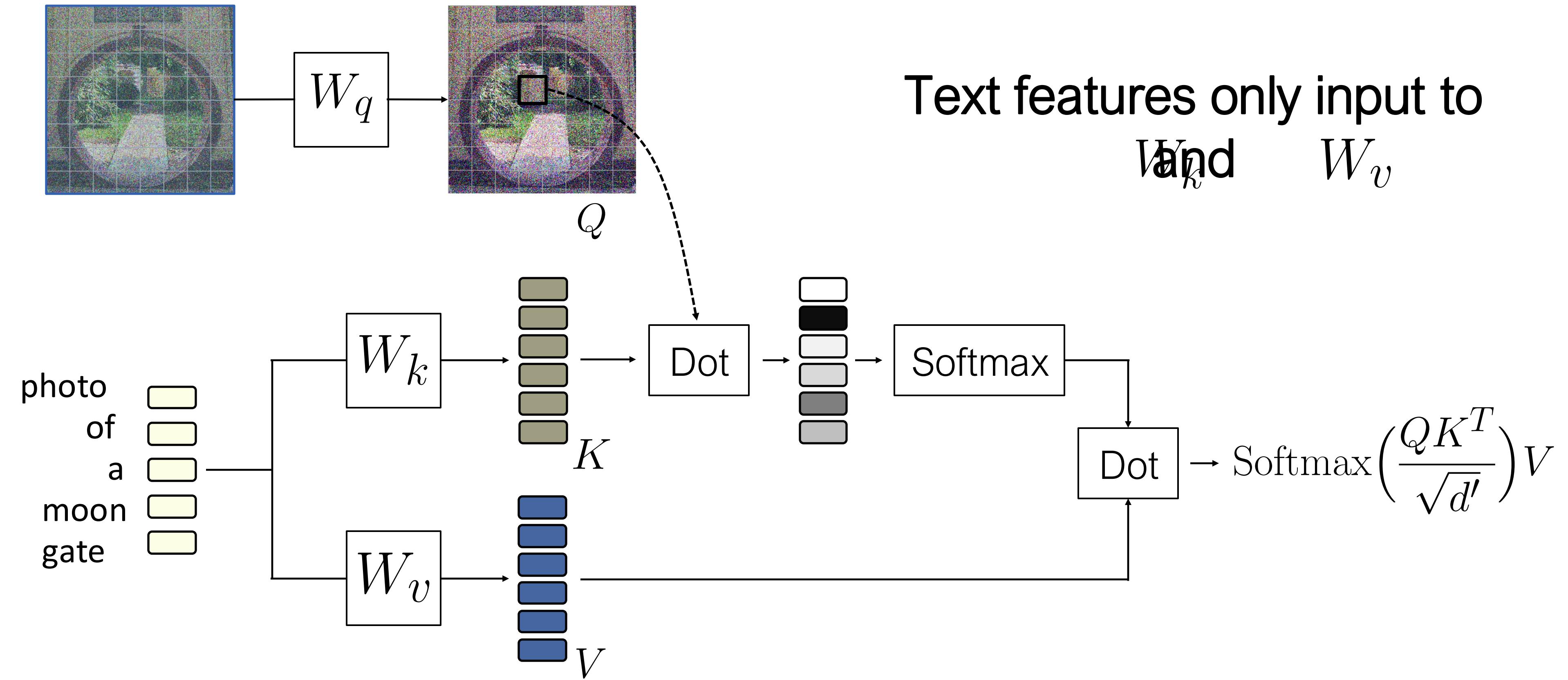
Diffusion Model U-Net

$x_t$



$\epsilon_t$

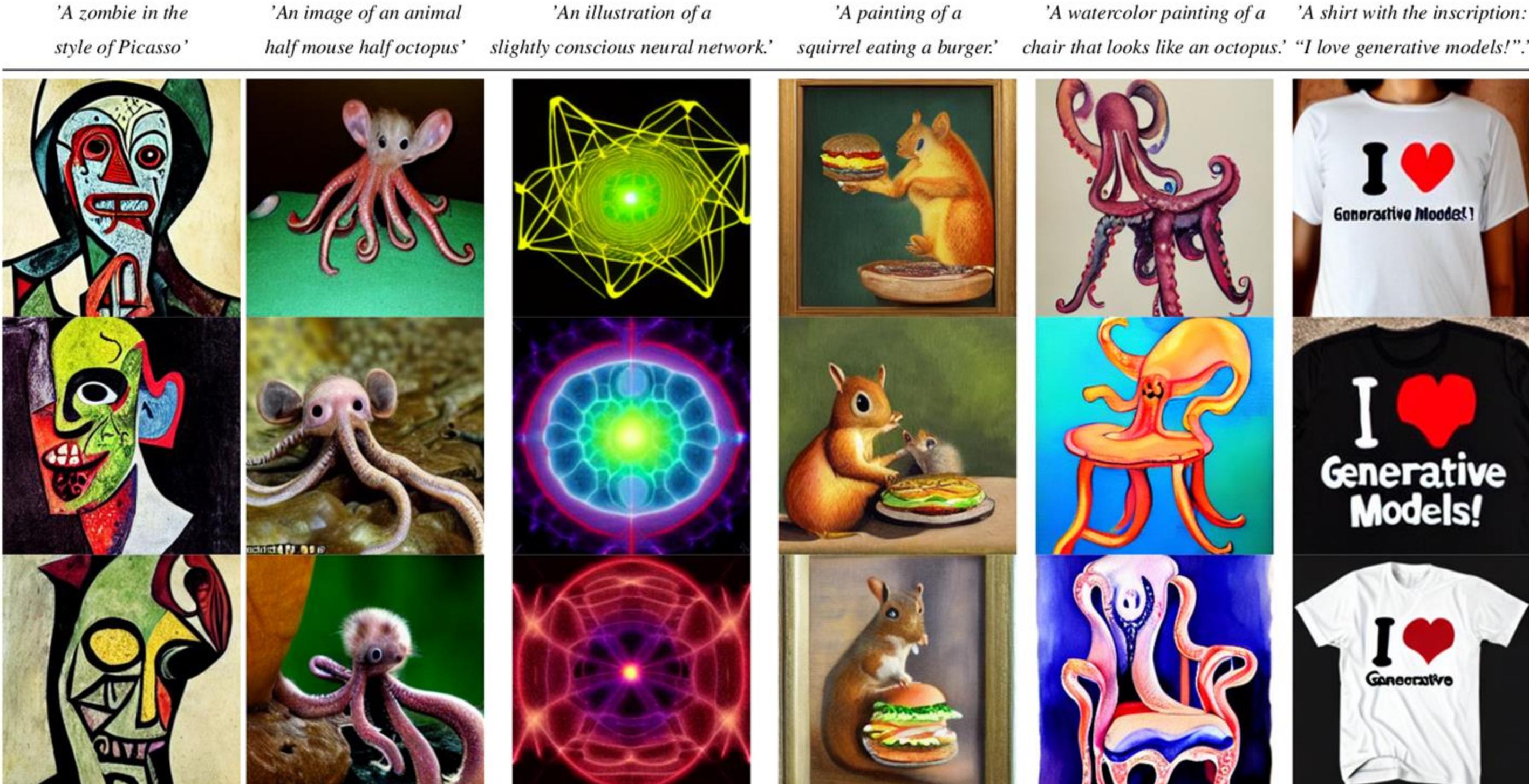
# Text-to-Image Cross-Attention



# LDMs for Text-to-Image Synthesis

- 32x32 cont. space
- 600M Transformer
- 800M UNet
- 400M Image/Text Pairs

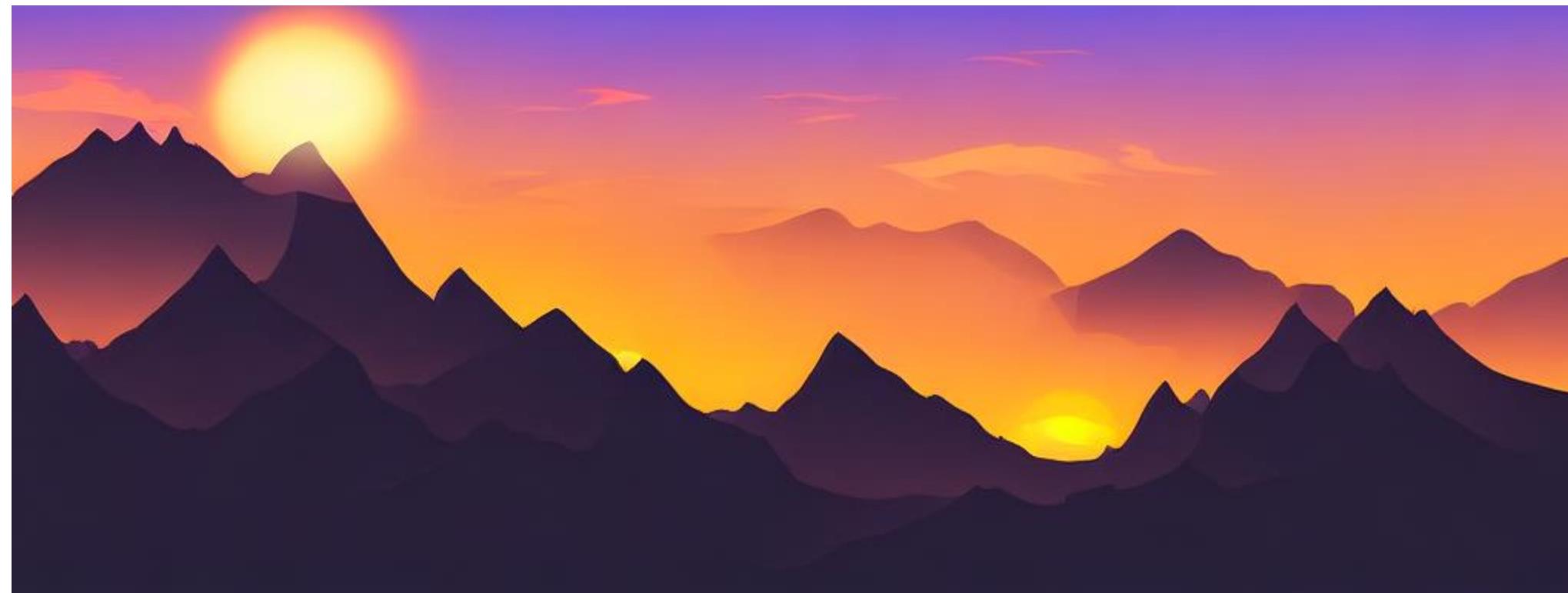
Text-to-Image Synthesis on LAION. 1.4B Model.



# LDMs for Text-to-Image Synthesis

convolutional sampling (train on  $256^2$ , generate on  $>256^2$ )

*"A sunset over a mountain range, vector image"*



*"A sunset over a mountain range, oil on canvas"*



# “Cheat Code”: Classifier-Free Diffusion Guidance

Jonathan Ho, Tim Salimans

- see <https://arxiv.org/abs/2207.12598>

- works very well for conditional image generation:

$$\hat{\epsilon}_\theta(x_t; y, t) \leftarrow \epsilon_\theta(x_t; \emptyset, t) + s \cdot (\epsilon_\theta(x_t; y, t) - \epsilon_\theta(x_t; \emptyset, t)), \quad s \geq 1.0$$

$s = 1.0$

Text Prompt

$s = 7.5$

Constant Embedding





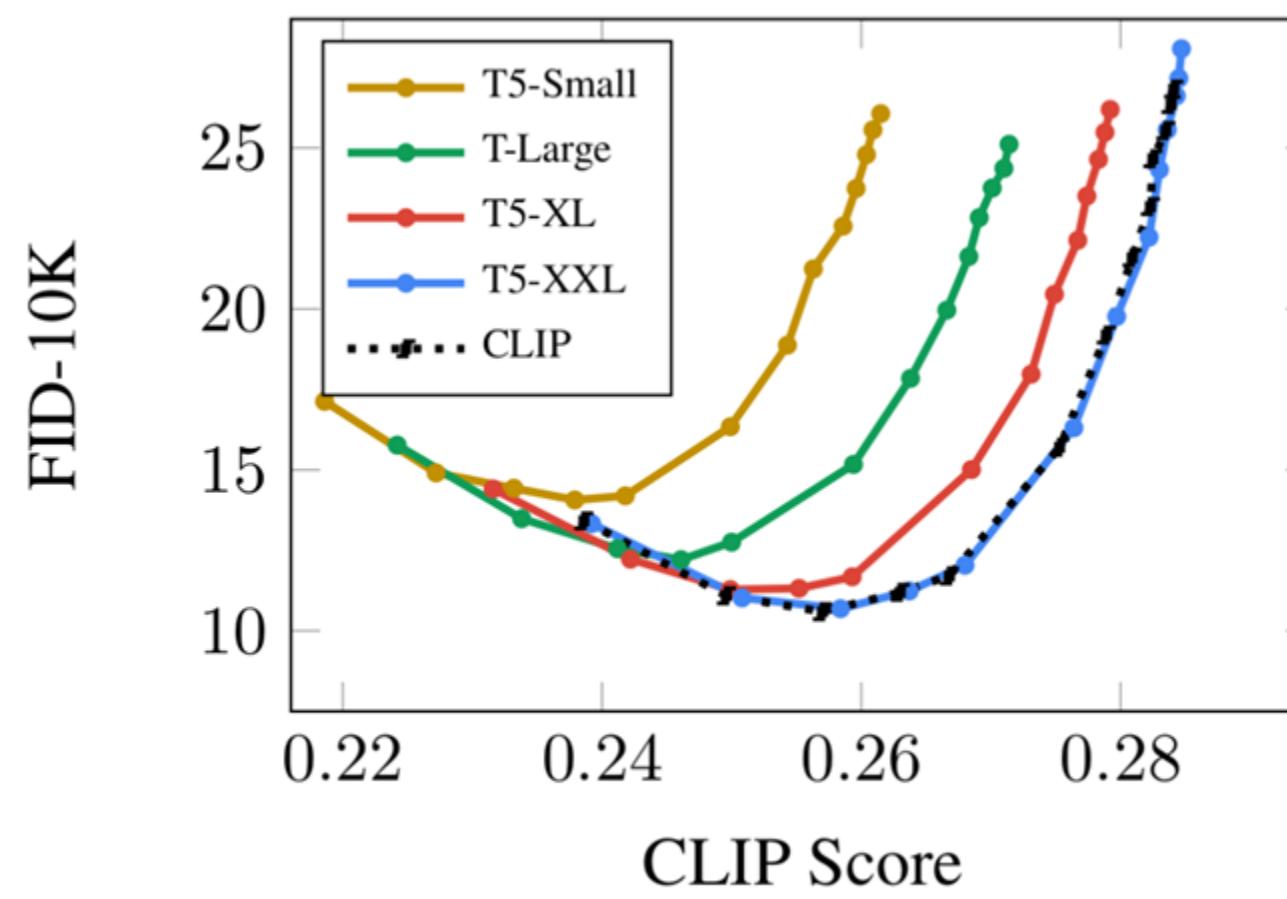
# Stable Diffusion

Latent Diffusion ++

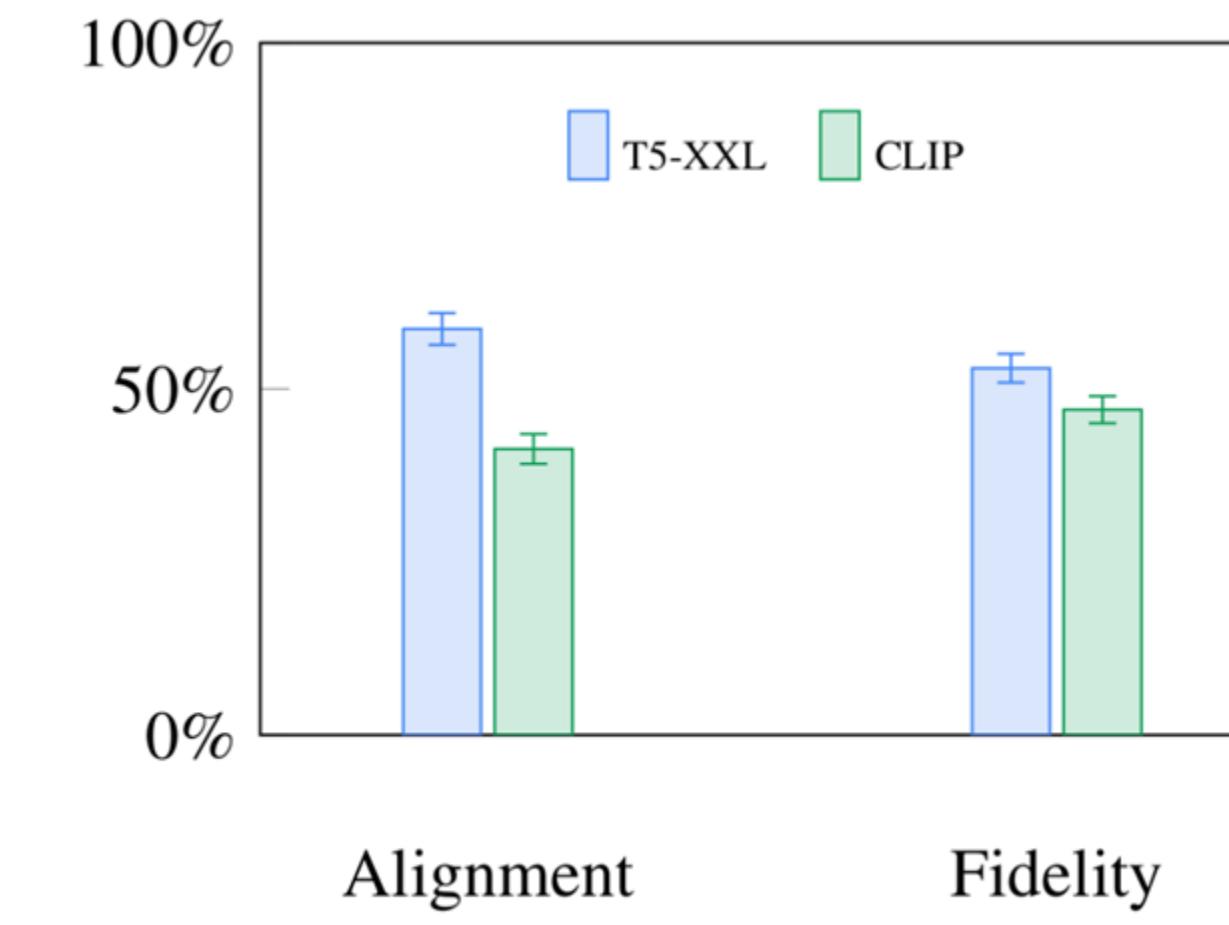


# From Latent to Stable Diffusion

- goal: achieve a small model that people can actually run locally on “small” GPUs (~10GB VRAM)
- progressive training: pretrain on 256x256, then continue on 512x512
- fix text encoder (as in Imagen)
- → choose CLIP (ViT-L/14) since performance/size tradeoff seems significant



(a) Pareto curves comparing various text encoders.



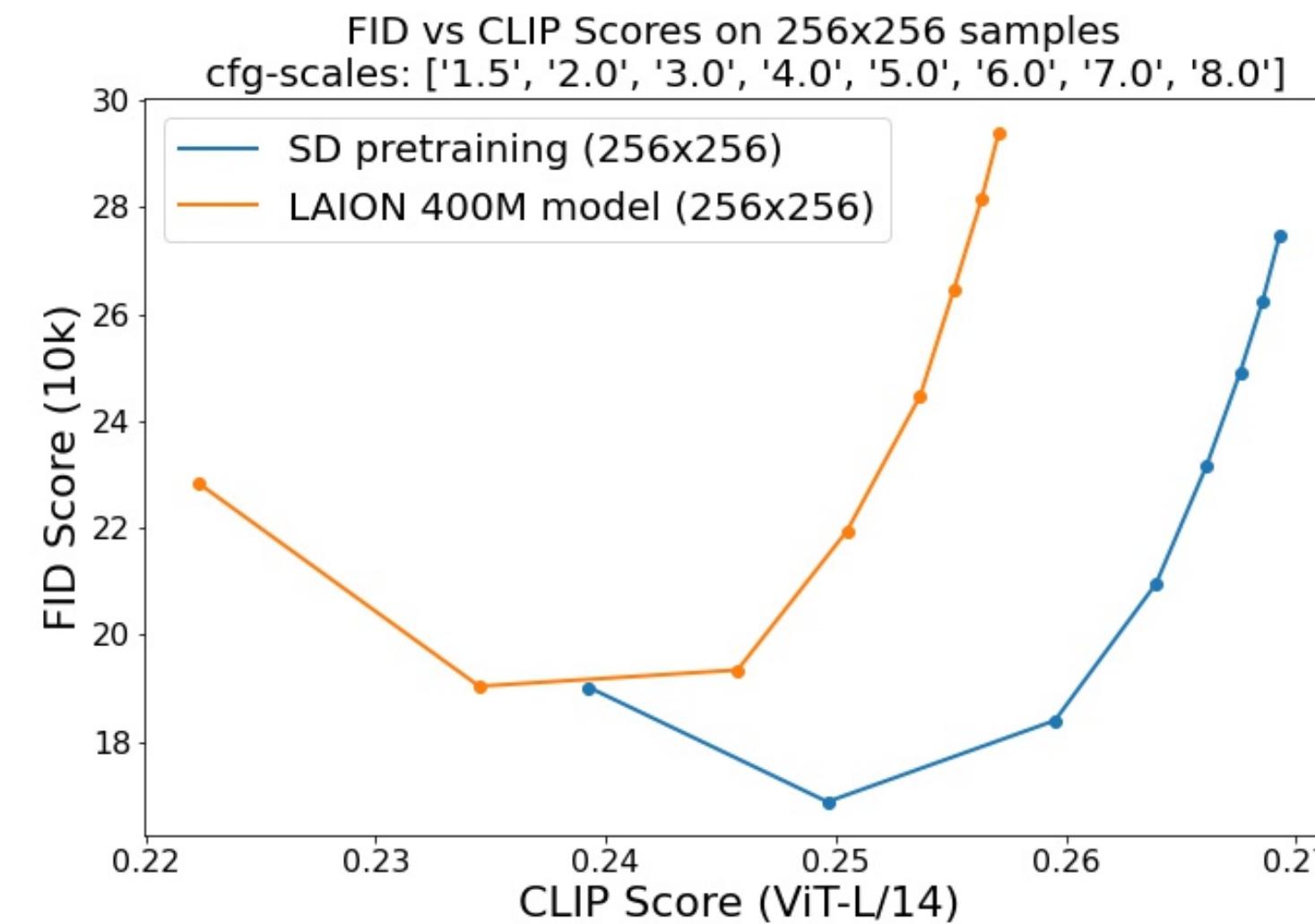
(b) Comparing T5-XXL and CLIP on DrawBench.

Figure from Imagen, <https://arxiv.org/abs/2205.11487>

# From Latent Diffusion to Stable Diffusion

## Stage 1: Pretraining @256x256

- 237k steps at resolution 256x256 on LAION 2B(en)
- batch-size = 2048
- ~ 64 A100 GPUs



10k random COCO val captions / 50 decoding steps

# From Latent Diffusion to Stable Diffusion

Stage 2: Training @512x512. batch-size=2048, #gpus=256

*part 1 (v1.1):*

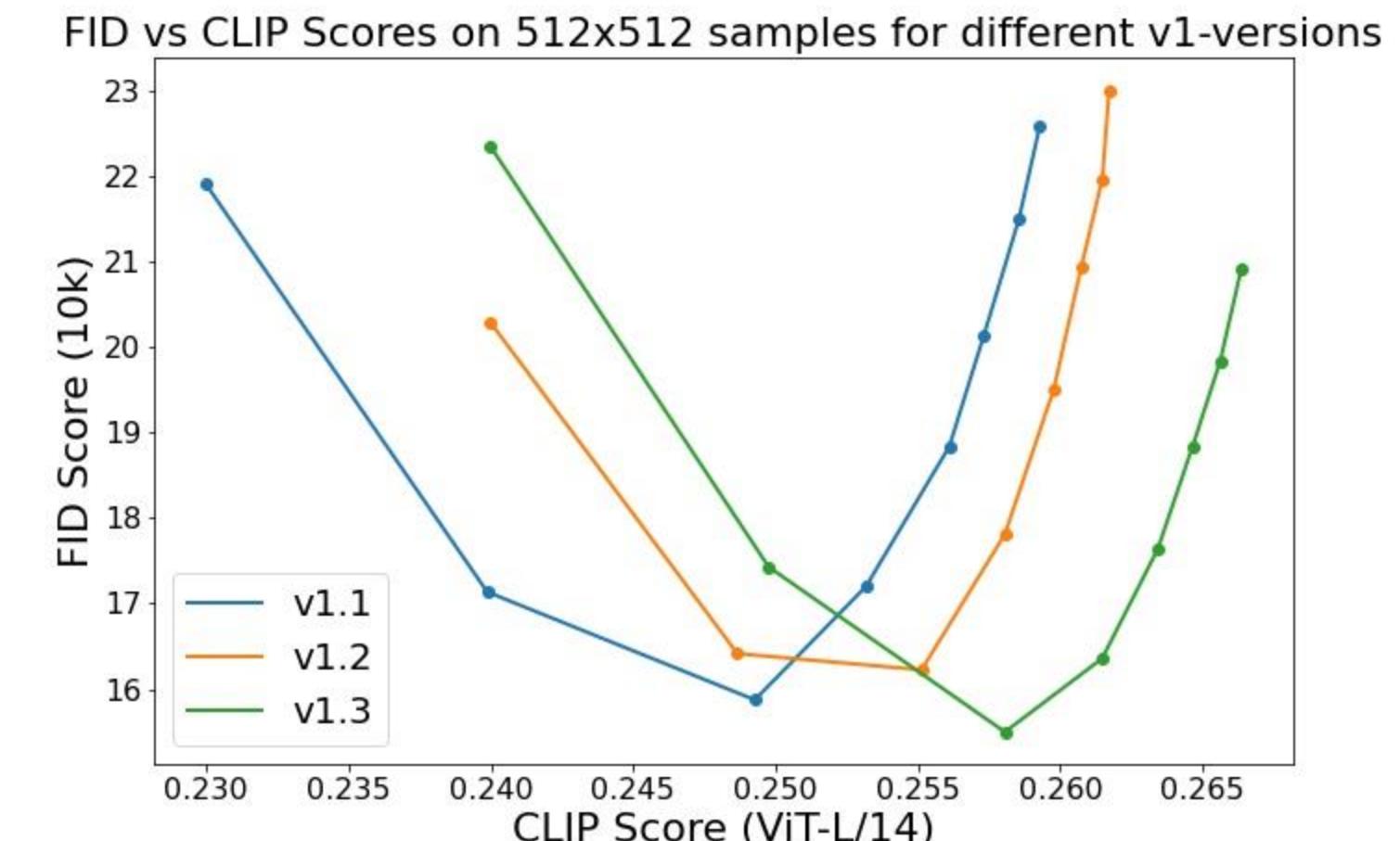
- 194k steps at resolution 512x512 on laion-high-resolution (170M examples from LAION-5B with resolution  $\geq 1024 \times 1024$ ).

*part 2 (v1.2):*

- 515k steps at resolution 512x512 on "laion-improved-aesthetics" (a subset of laion2B-en, filtered to images with an original size  $\geq 512 \times 512$ , estimated aesthetics score  $> 5.0$ , and an estimated watermark probability  $< 0.5$ )

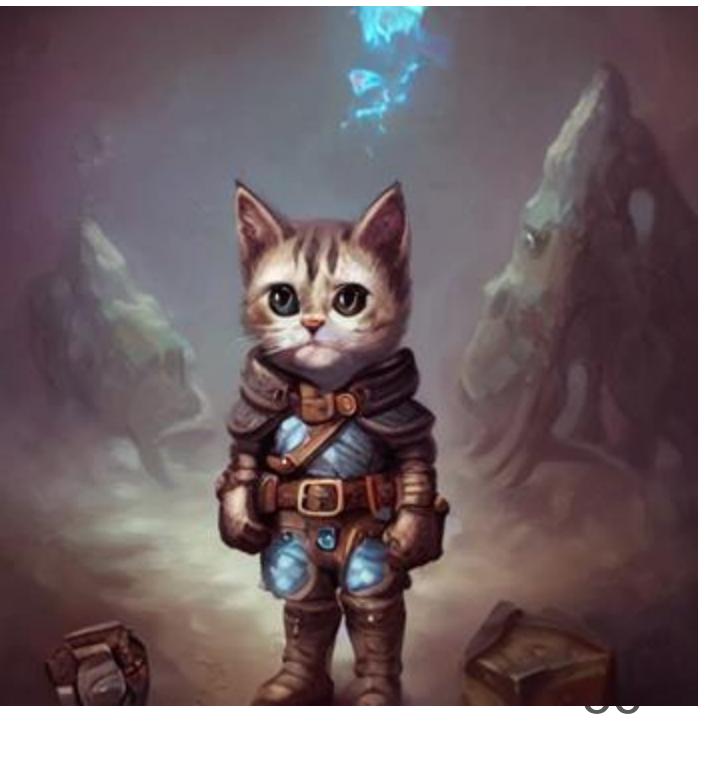
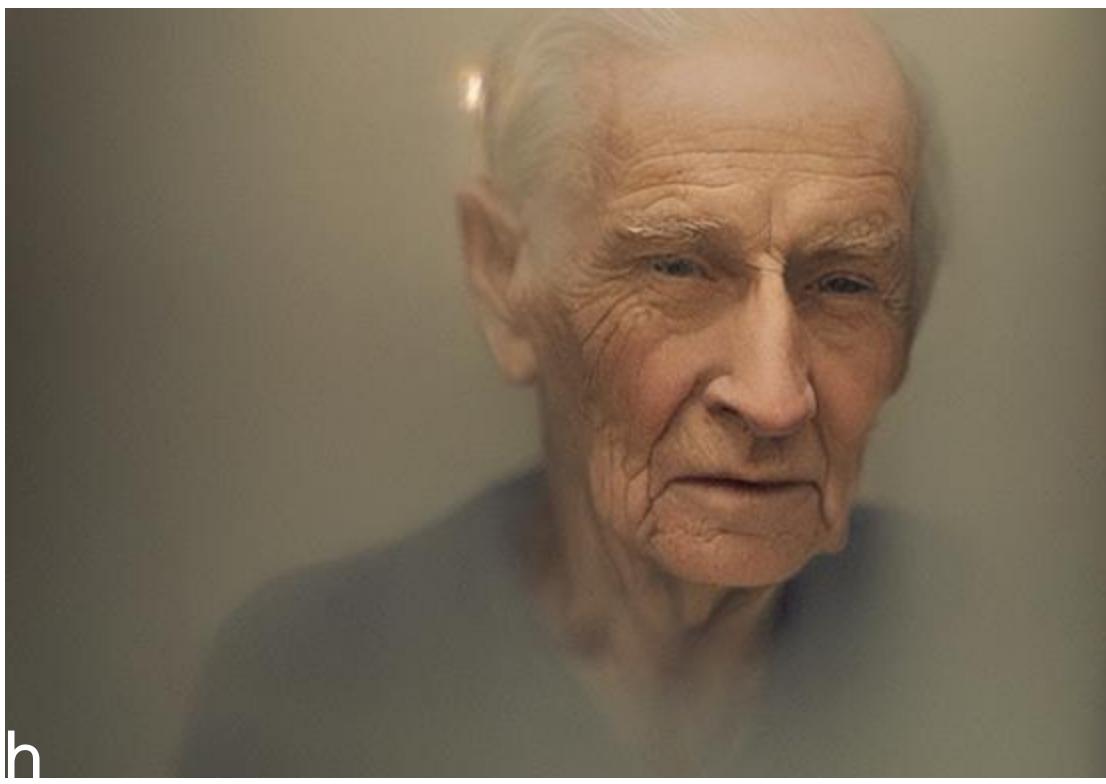
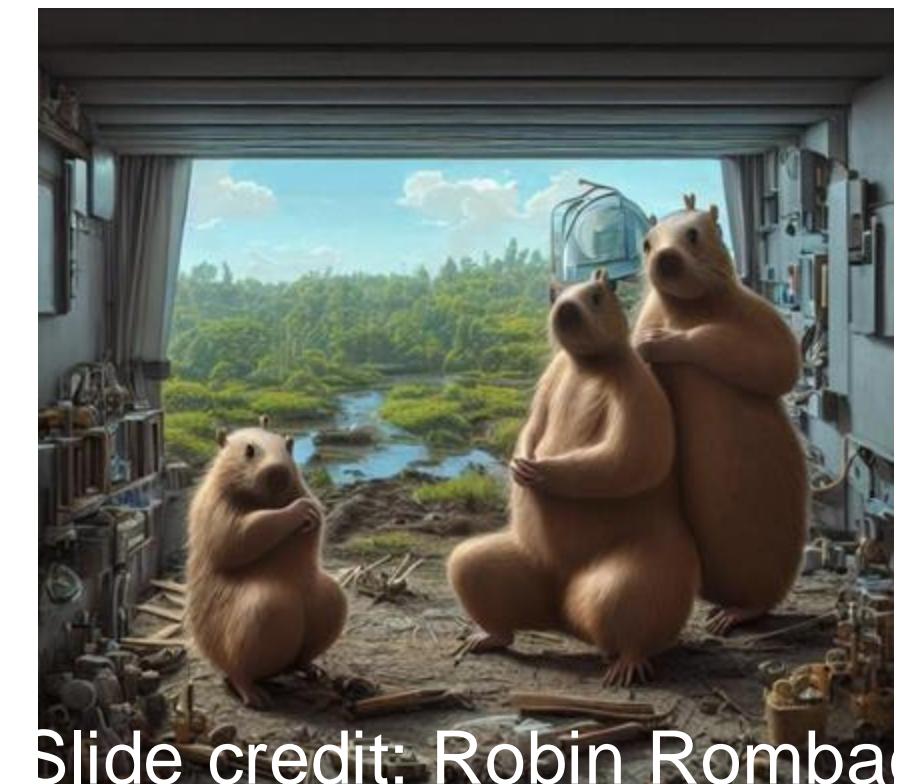
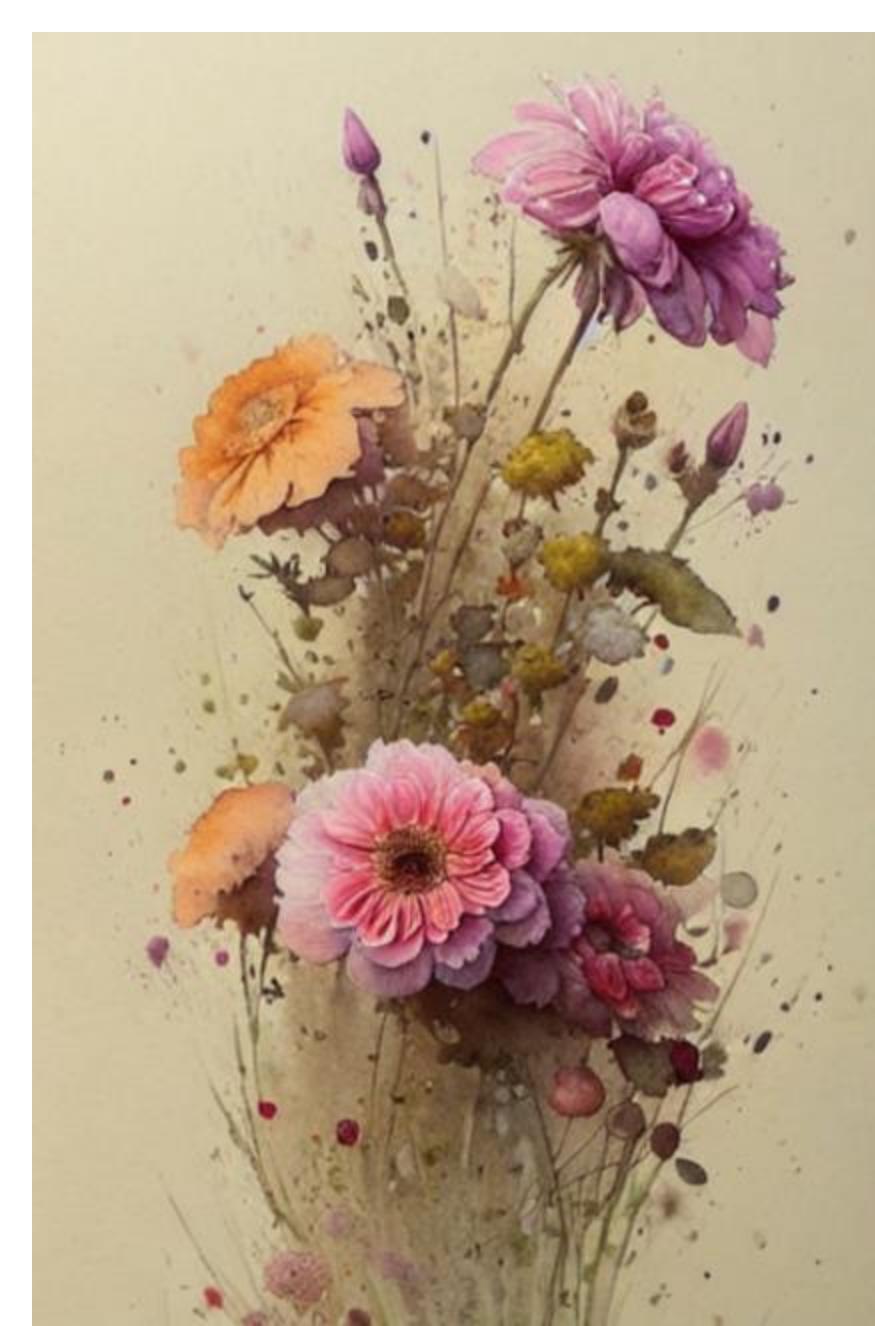
*part 3/4 (v1.3/v1.4):*

- 195k/225k steps at resolution 512x512 on "laion-improved-aesthetics" and 10% dropping of the text-conditioning



10k random COCO val captions / 50 decoding steps

→ 4.2 GB checkpoint (EMA only, fp32)



Slide credit: Robin Rombach

# Video Synthesis

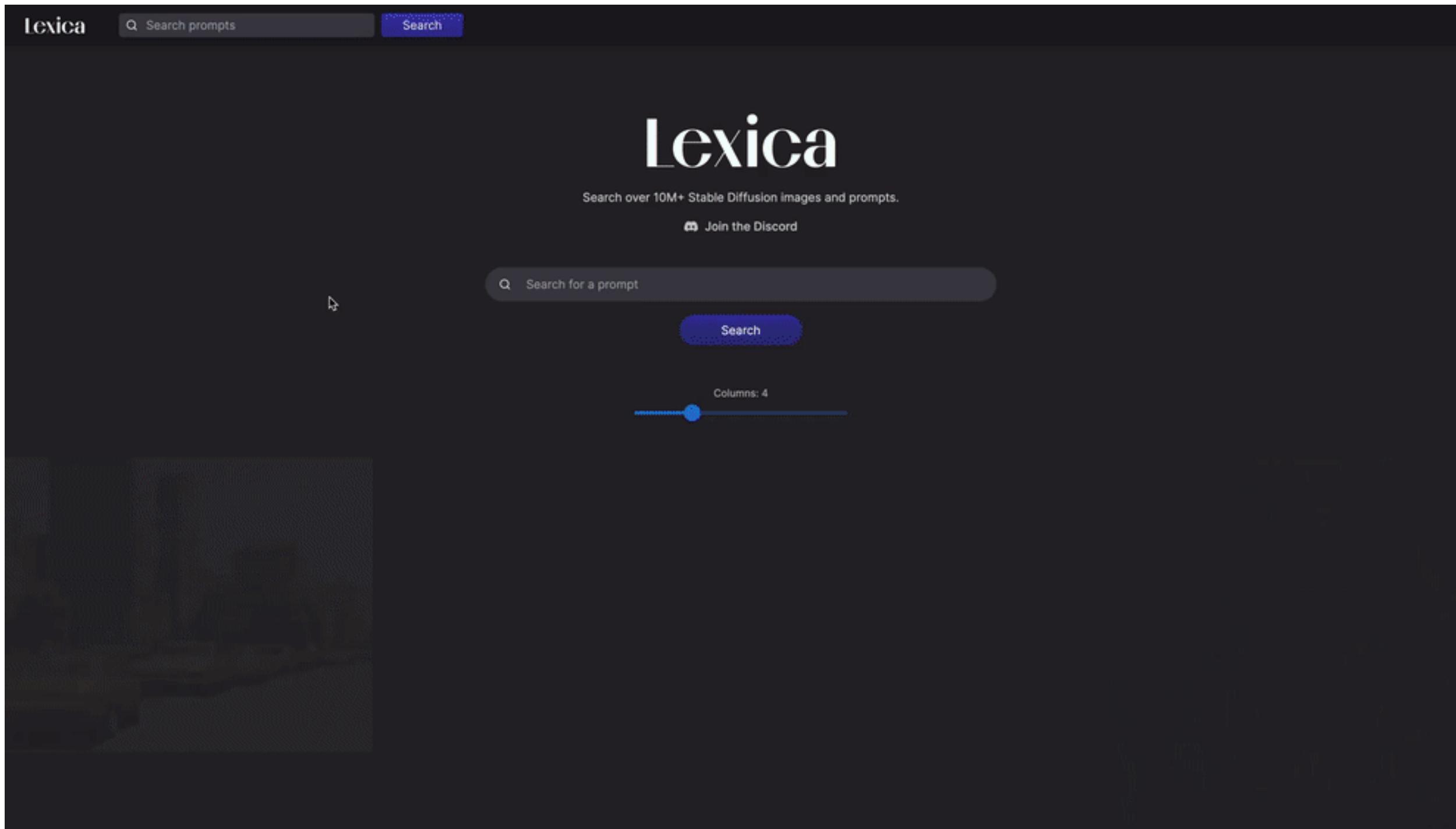


Stable Diffusion (img2img) + EBSynth by Scott Lightsier:

<https://twitter.com/LighthiserScott/status/1567355079228887041?t=kXXCAVuO5IJCGcro3Ma3A&s=19>

EBSynth: single-frame video stylization app: <https://ebsynth.com/>

# Prompt Search Engine (lexica.art)



# Prompt Marketplace ([promptbase.com](https://promptbase.com))

**DALL·E, GPT-3, Midjourney, Stable Diffusion, ChatGPT Prompt Marketplace**

Find top prompts, produce better results, save on API costs, sell your own prompts.

[Find a prompt](#) [Sell a prompt](#)



Funky Animals



Super Cute Line Art



Modern Woodcut Engravings Intric...



Financial Analysis Of Companies

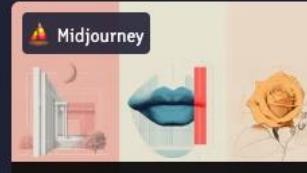
*Featured in*

TechCrunch THE VERGE WIRED FAST COMPANY FINANCIAL TIMES Atlantic yahoo/finance WSJ

#### Featured Prompts



Vintage Retro Pattern Tiles \$1.99



Minimal Pastel Diagram Art \$2.99



Objects Made Of Money \$2.99



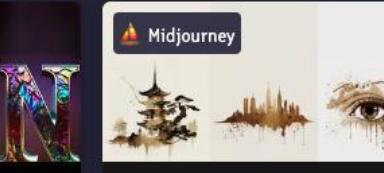
Butterfly Cliparts \$2.99



Asymmetrical Split Exposure ... \$2.99



Stained Glass Letters \$2.99

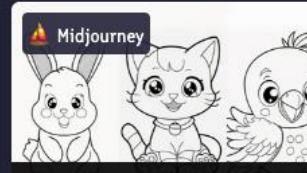


Coffee Stain Art \$2.99

#### Hottest Prompts



Nft Generative Art Maker \$2.99



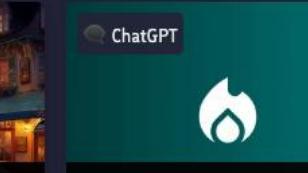
Clean Animal Art For Coloring... \$1.99



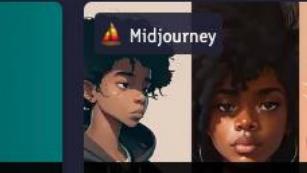
Tiny Gouache Houses \$2.99



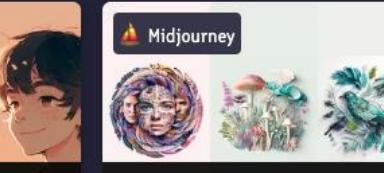
Beautiful Oil Paintings \$2.99



Hot Prl Selling \$2.99



Make Cartoons Like Lofi-girl \$2.99



Delicate Vibrant Emotive Arra... \$2.99

#### Newest Prompts



Fix Anything \$2.99



Tropical Fashion \$2.99



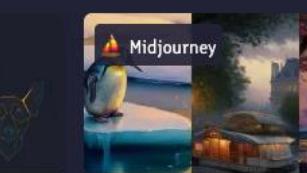
Food Images With Neon Effects \$1.99



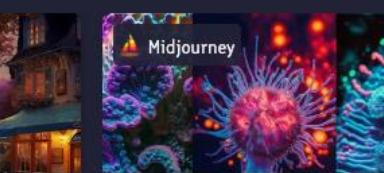
Wall Art Mockups Choose Wall ... \$1.99



Premium Logos \$2.99

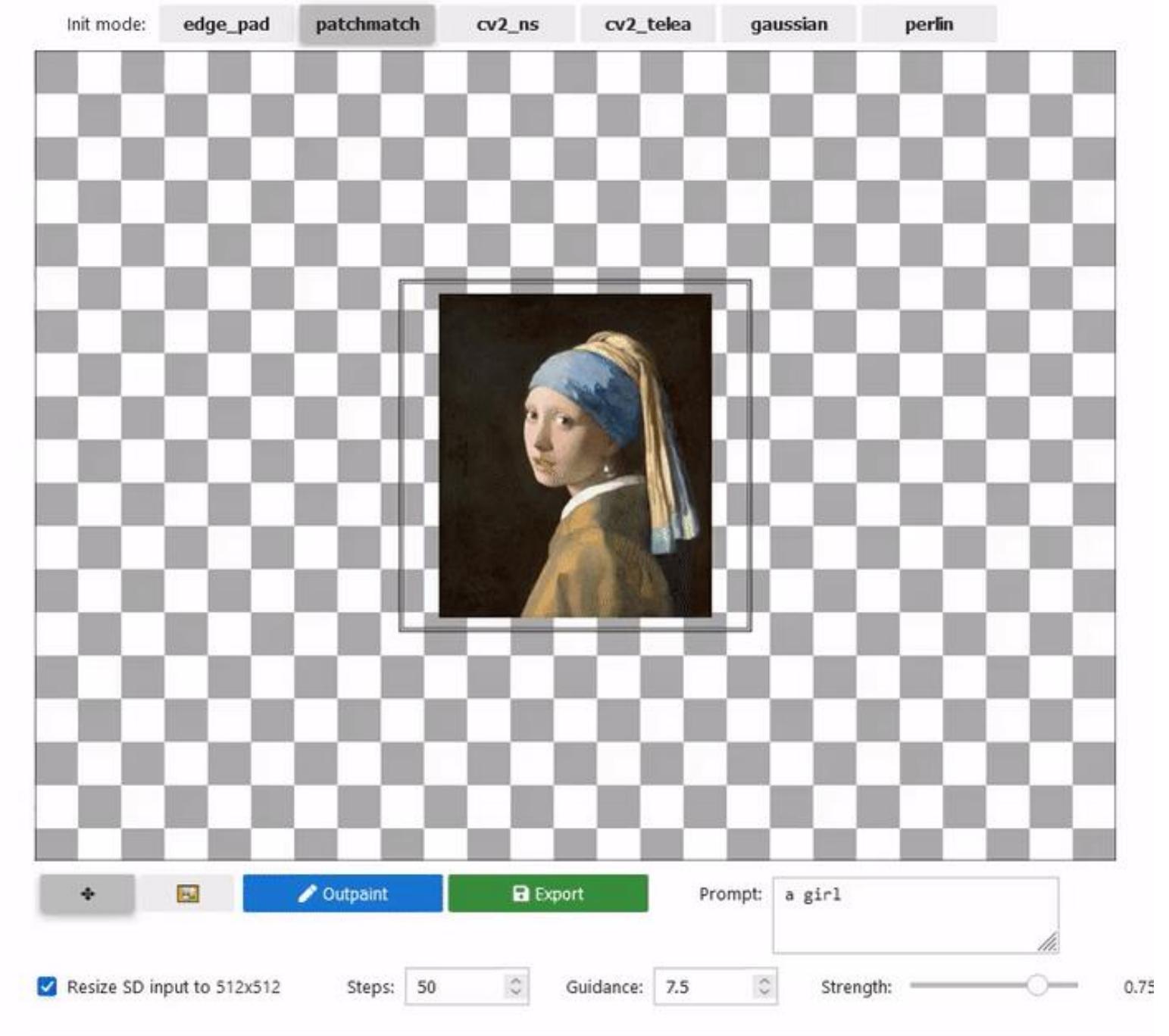
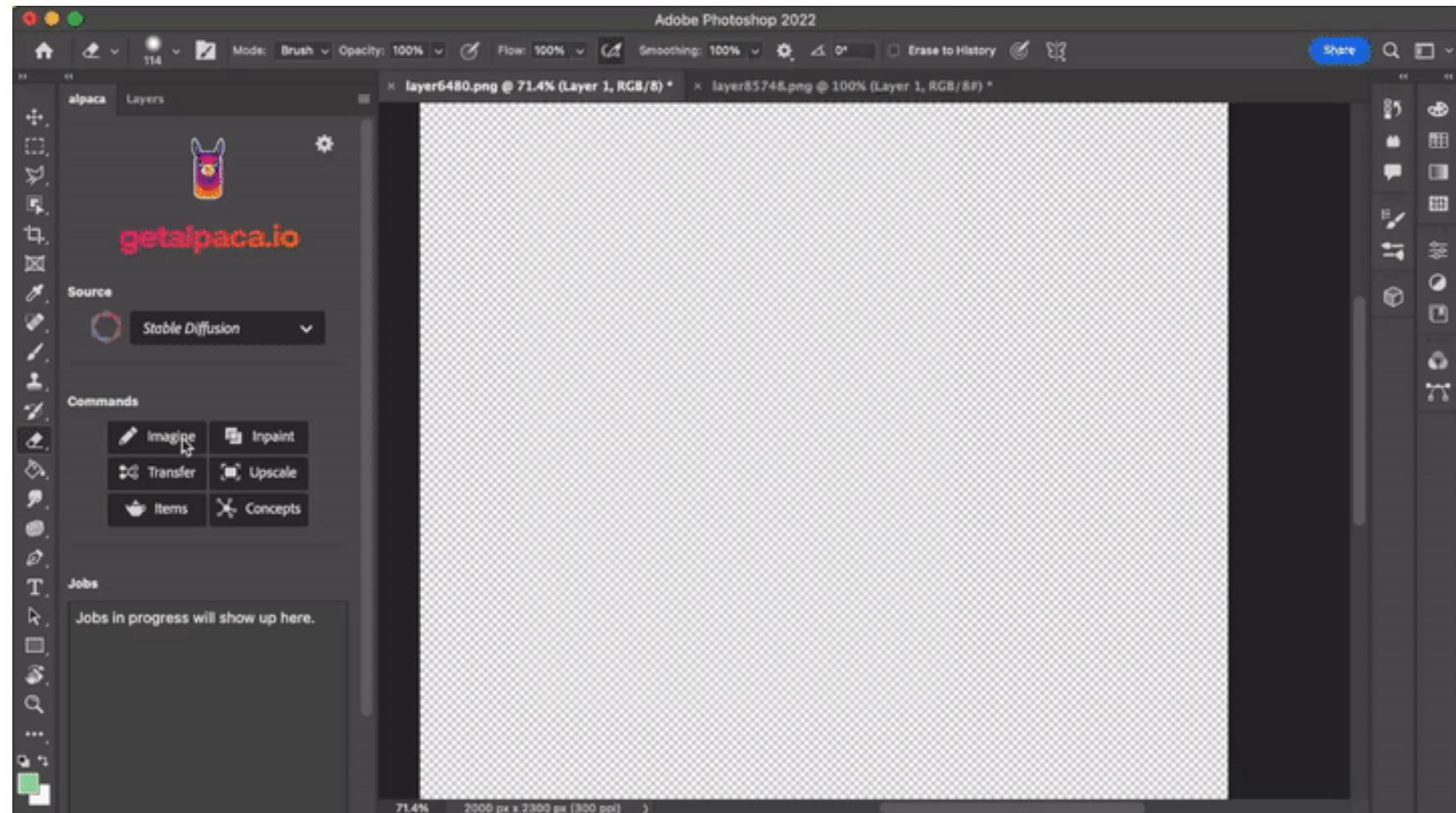


Beautiful Oil Paintings \$2.99



Alien Bio Organisms Posters \$2.99

# UIs / Plug-Ins for Photoshop, GIMP etc

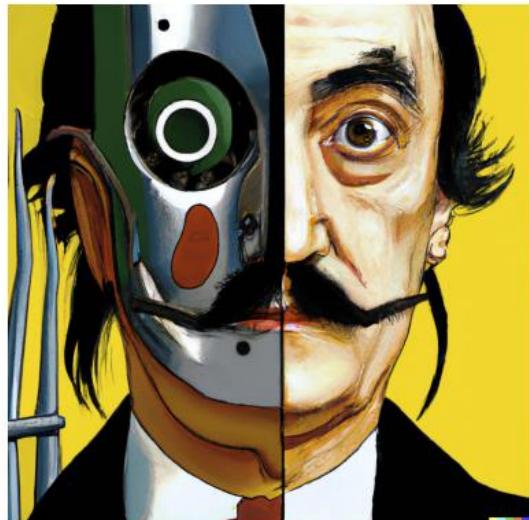


<https://twitter.com/wbuchw/status/1563162131024920576>

<https://github.com/lkwq007/stablediffusion-infinity>

**What if you have 1,000+ GPUs/TPUs**

# DALL-E 2, Imagen



vibrant portrait painting of Salvador Dalí with a robotic half face



a shiba inu wearing a beret and black turtleneck



a close up of a hand palm with leaves growing from it



an espresso machine that makes coffee from human souls, artstation



panda mad scientist mixing sparkling chemicals, artstation



a corgi's head depicted as an explosion of a nebula



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.

- Pixel-based Diffusion (No encoder-decoder)
- pre-trained text encoder (CLIP, t5)
- Diffusion model + classifier-free guidance
- Cascaded models: 64->128->512

<https://cdn.openai.com/papers/dall-e-2.pdf>  
<https://arxiv.org/abs/2205.11487>

# Diffusion vs. Autoregressive vs. GANs

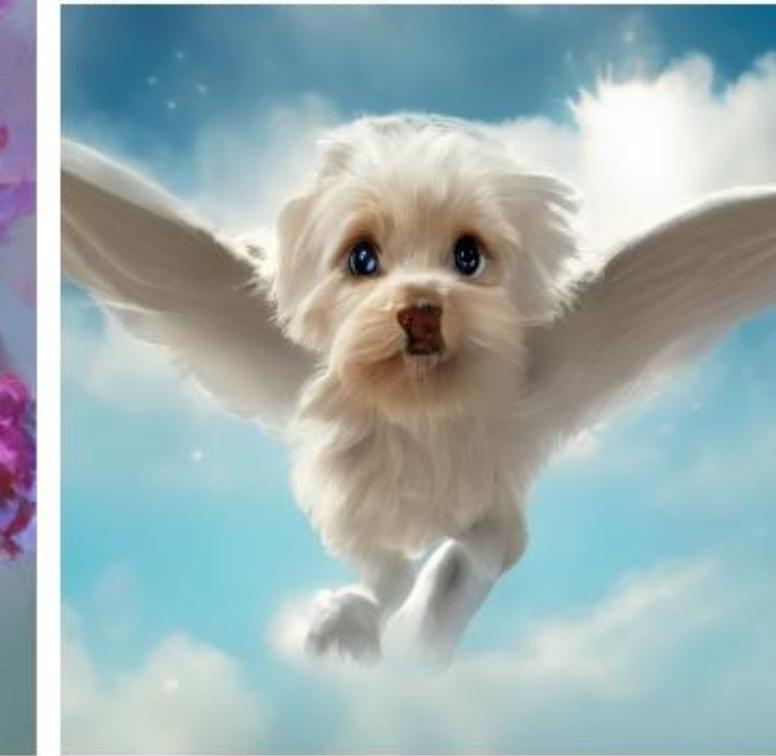
# GigaGAN: Scaling up GANs



A portrait of a human growing colorful flowers from her hair. Hyperrealistic oil painting. Intricate details.

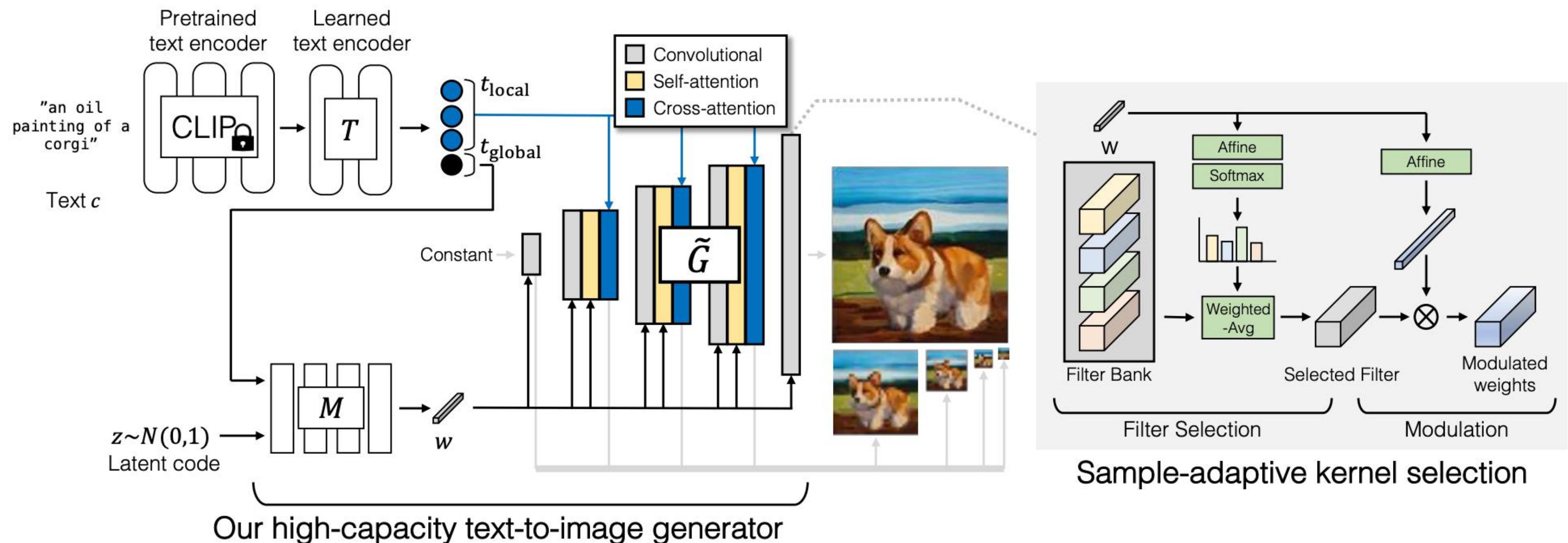


A golden luxury motorcycle parked at the King's palace. 35mm f/4.5.

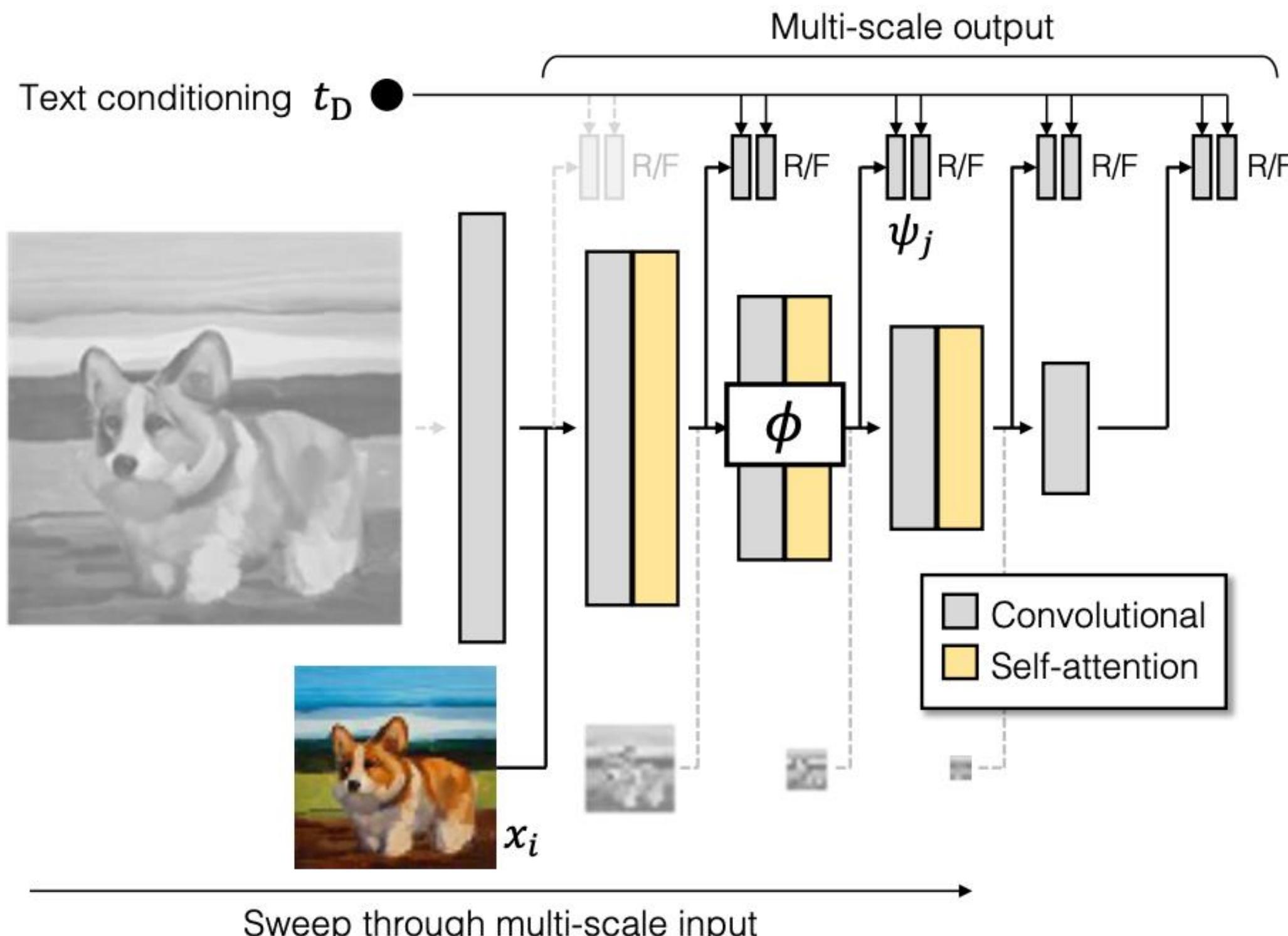


a cute magical flying maltipoo at light speed, fantasy concept art, bokeh, wide sky

# GigaGAN Generator



# GigaGAN Discriminator



# Style Mixing

"A Toy sport sedan, CG art."

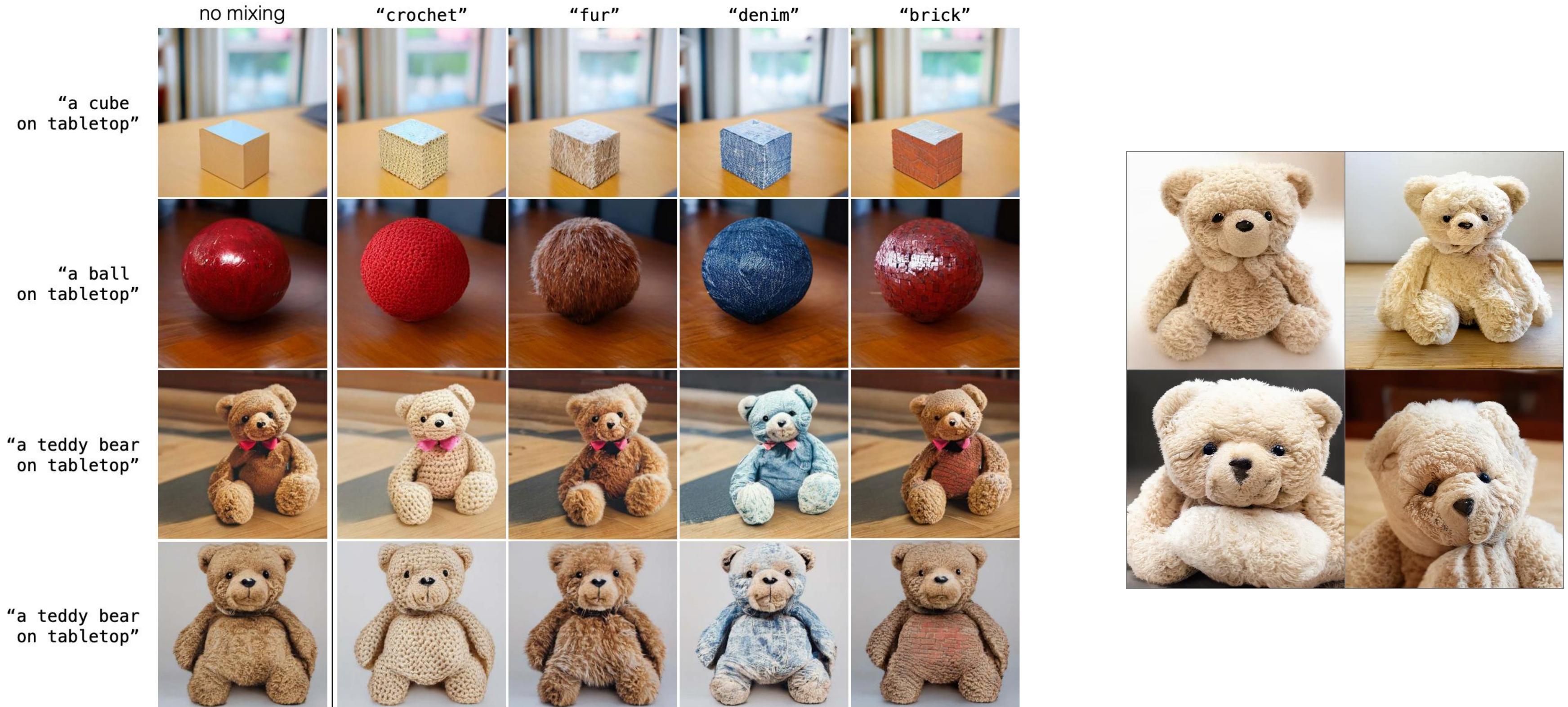
Fine styles



Coarse styles



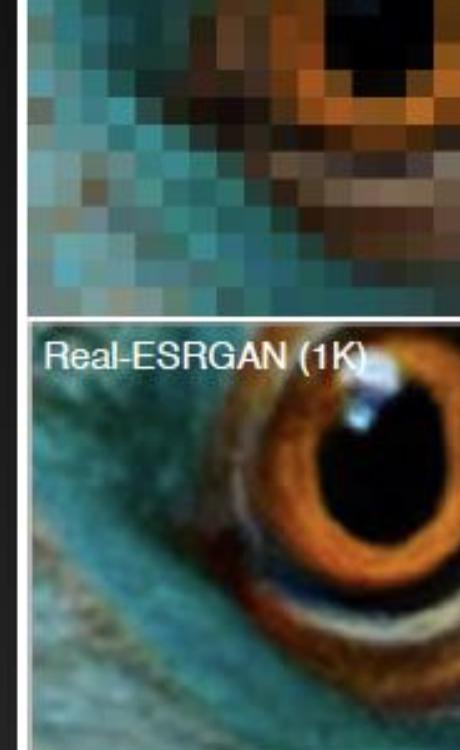
# Prompt Mixing





GigaGAN Upsampler (4096px, 16Mpix, 3.66s)

Input



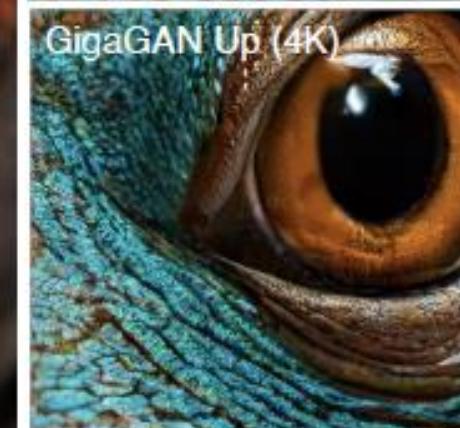
Real-ESRGAN (1K)



SD Upscaler (1K)



GigaGAN Up (1K)



GigaGAN Up (4K)

# Comparison between Different Models



Ours (512px, 0.14s / img, truncation  $\psi = 0.8$ )



Ours (512px, 0.14s / img, truncation  $\psi = 0.8$ )



LDM (256px, 9.4s / img, 250 steps, guidance=6.0)



LDM (256px, 9.4s / img, 250 steps, guidance=6.0)



Stable Diffusion v1.5 (512px, 2.9s / img, 50 steps, guidance=7.5)



Stable Diffusion v1.5 (512px, 2.9s / img, 50 steps, guidance=7.5)

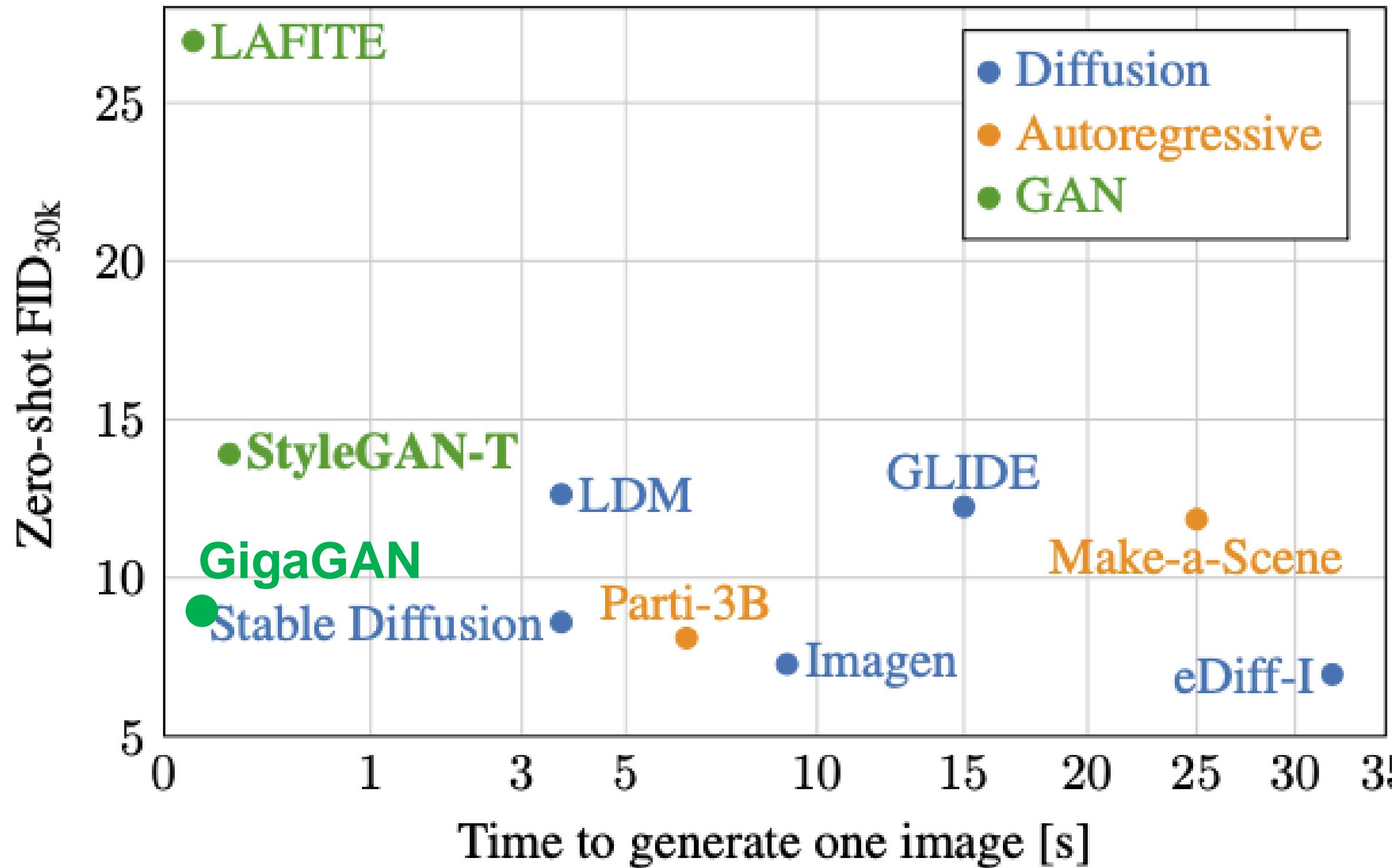


DALL-E 2 (1024px)



DALL-E 2 (1024px)

# StyleGAN-T



# How could we improve it?

- Better generative modeling techniques: VAEs, GANs, diffusion, AR, Hybrid
- Better text encoders: RNN/LSTM -> Transformers (CLIP, T5)
- Better generator architectures: RNN/LSTM -> CNN -> CNN + Transformer
- Better ways to connect text and image: concatenation -> AdaIN -> cross-attention
- More data + GPU/TPU computing: a few hundred A100.
- Bigger model sizes: 1B-20B.