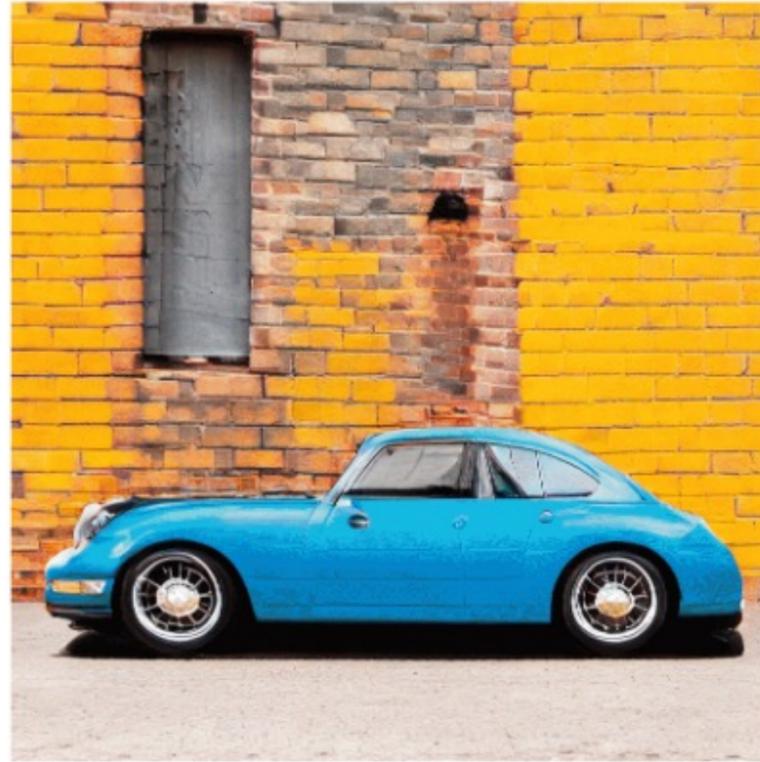
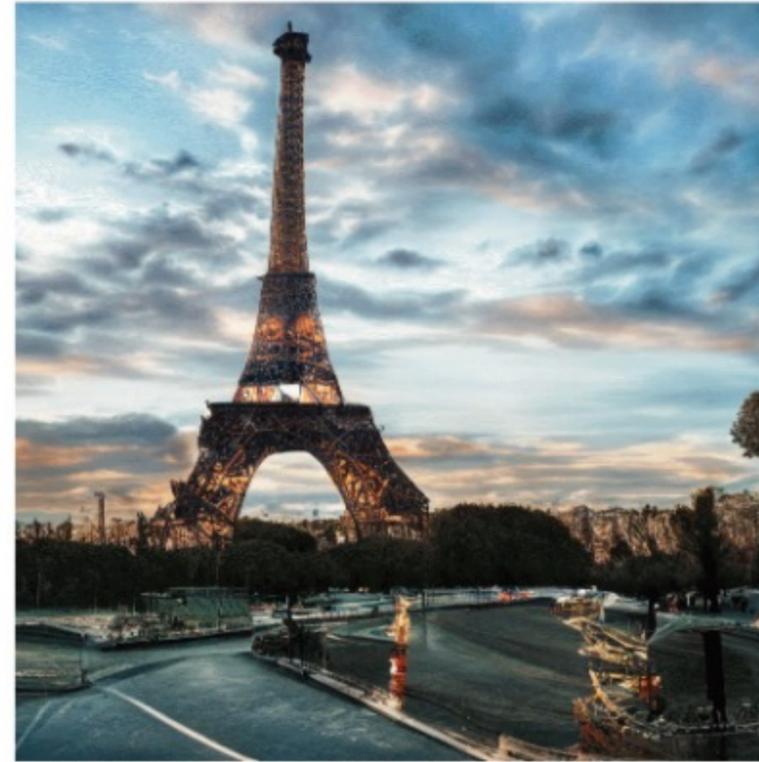




A living room with a fireplace at a wood cabin. Interior design.



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



Eiffel Tower, landscape photography



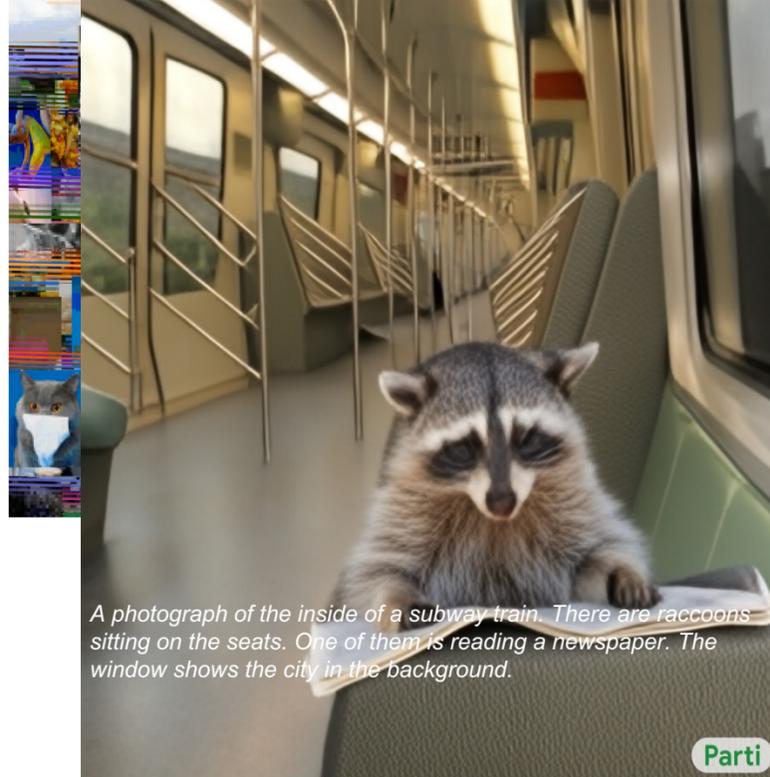
A painting of a majestic royal tall ship in Age of Discovery.

Lecture 13: Text-to-Image Synthesis

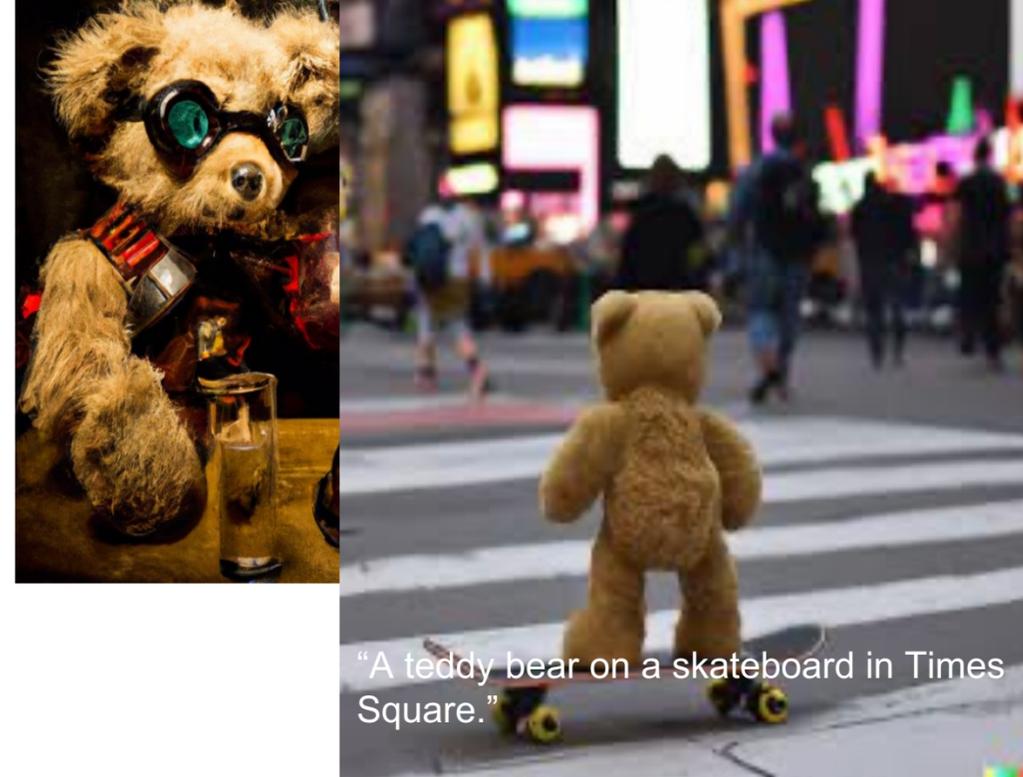
Jun-Yan Zhu

16-726 Spring 2023

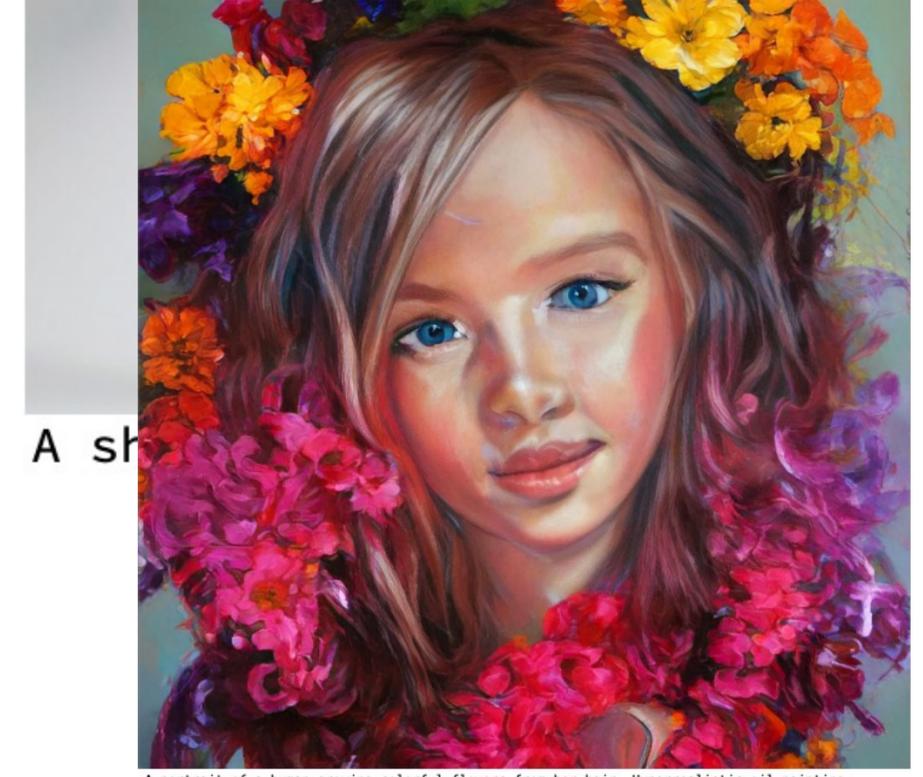
Text-to-Image Everywhere



Autoregressive models
(Image GPT, Parti)



Diffusion models
(DALL-E 2, Imagen)



GANs, Masked GIT
(GigaGAN, MUSE)

Text-to-Image Everywhere

Scott Lighthiser @LighthiserScott · Sep 18
.@StableDiffusion Img2Img x #ebsynth Creature Test
#stablediffusion #AIart



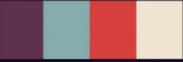
0:30 215.2K views

Orcton @OrctonAI · Sep 15
Few comments about the Midjourney/@D_ID_n Video wondering why this means we will soon be able to create our own personalised digital assistants. Here's a vision of a personalised digital assistant to explain.
#midjourney #Midjourneyai #AIart #Digitalart #animated

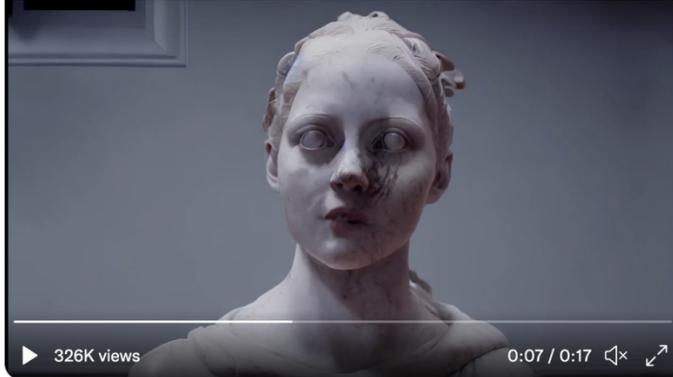


44 46.9K views

Matt DesLauriers @mattdesl · Sep 13
AI tool to generate colour palettes from any text prompt —
#stablediffusion #ArtificialIntelligence
[Show this thread](#)

"tokyo neon" 	"green garden, blue sky" 
"beautiful sunset" 	"living coral" 

Scott Lighthiser @LighthiserScott · Sep 18
.@StableDiffusion Img2Img x #ebsynth x @koe_recast TEST
#stablediffusion #AIart



326K views 0:07 / 0:17

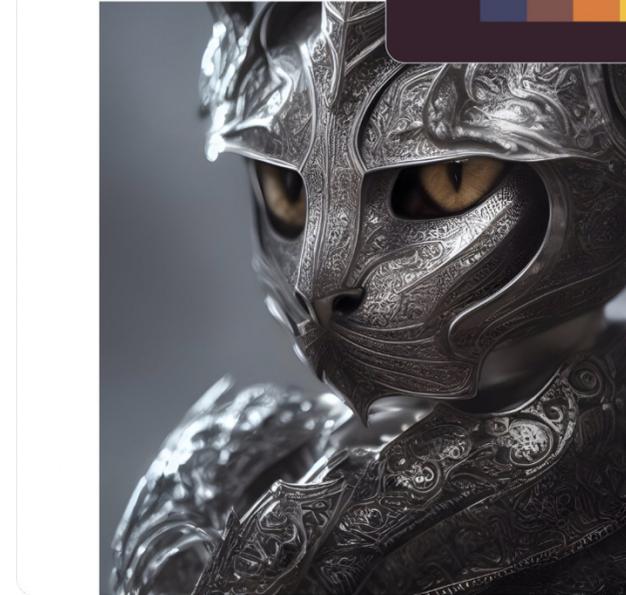
Matt Reed @mcreed · Sep 9
I am at a loss for everything
#stablediffusion #aiart
[Show this thread](#)



Replicate @replicatehq · Sep 9
The Stable Diffusion innovation just doesn't stop!
Here's a new open-source model from the @monaverse that produces seamless tiling images: replicate.com/tommoore515/ma...
[Show this thread](#)



Emad @EMostaque · Sep 5
"Cat Knight".
#StableDiffusion #AI
Prompt: kneeling cat knight, portrait, finely detailed, cinematic lighting, 4k
Parameters:scale7.50-k euler
By Valere: hostux.social/@valere/108939...



Stable Diffusion Pics & DreamStudio @DiffusionPics · Sep 2
"Jeffon Zuckergates"
#StableDiffusion #AIart #AIartwork #DreamStudio @StableDiffusion



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.



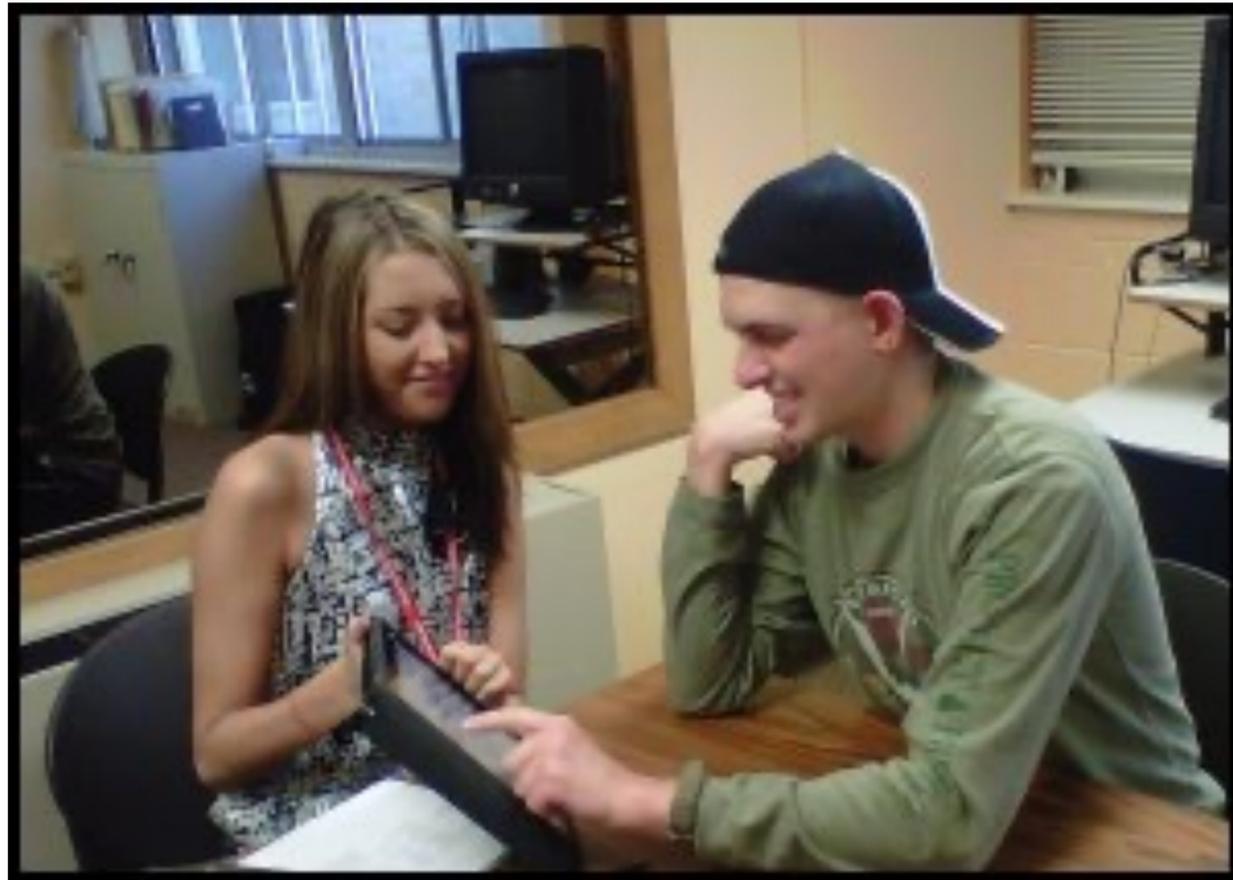
Step 1: Image Selection.

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

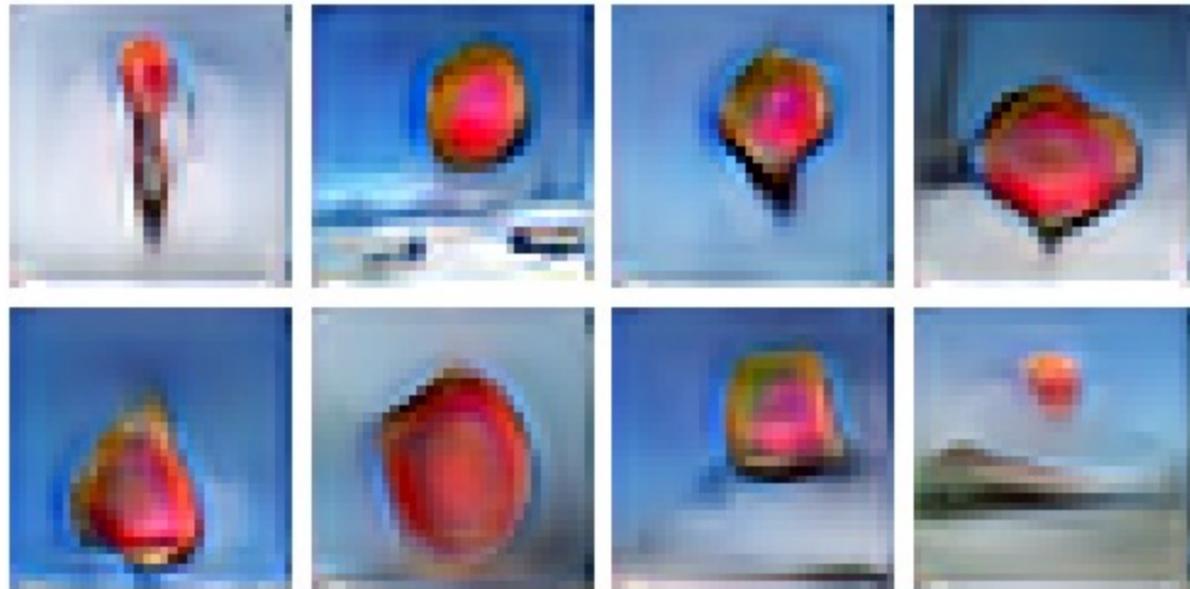


Math learning and reading comprehension
for young children

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 flying 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.

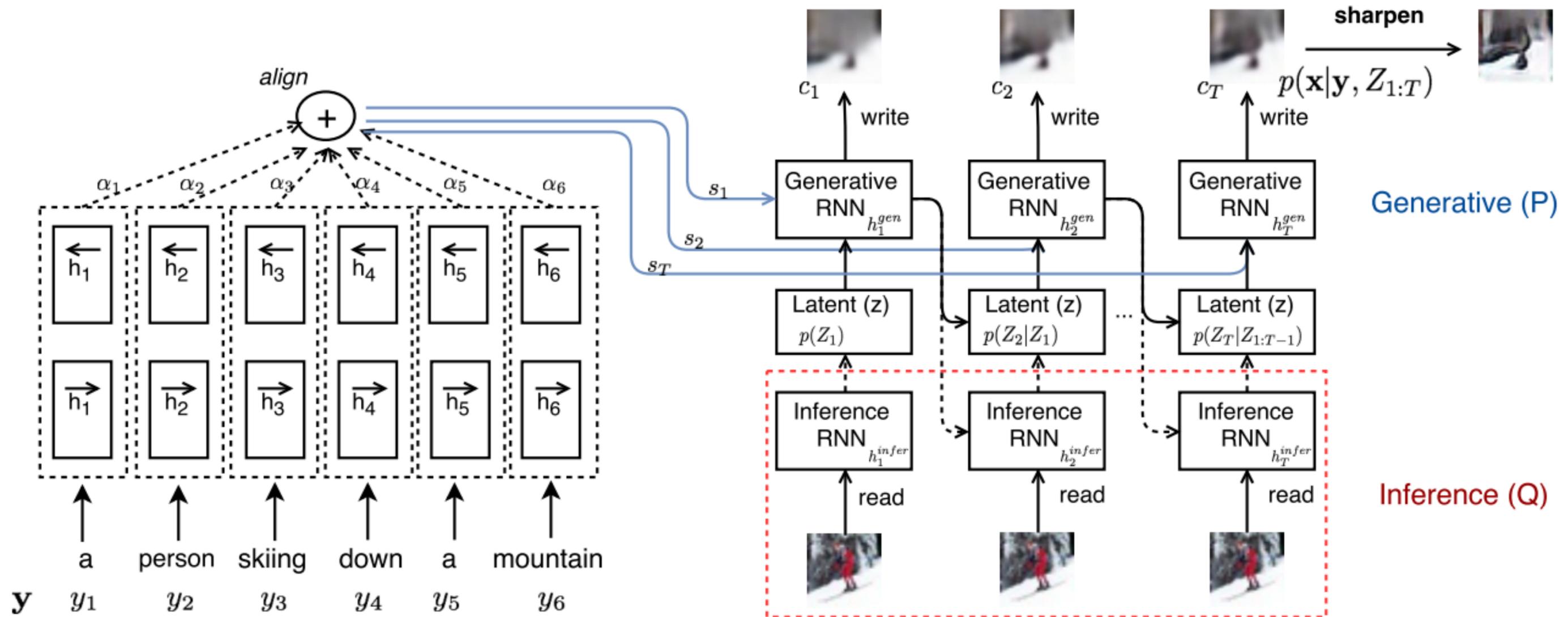


A person skiing on sand clad vast desert.

Generating Images from Captions with Attention.

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

First Deep Learning Work

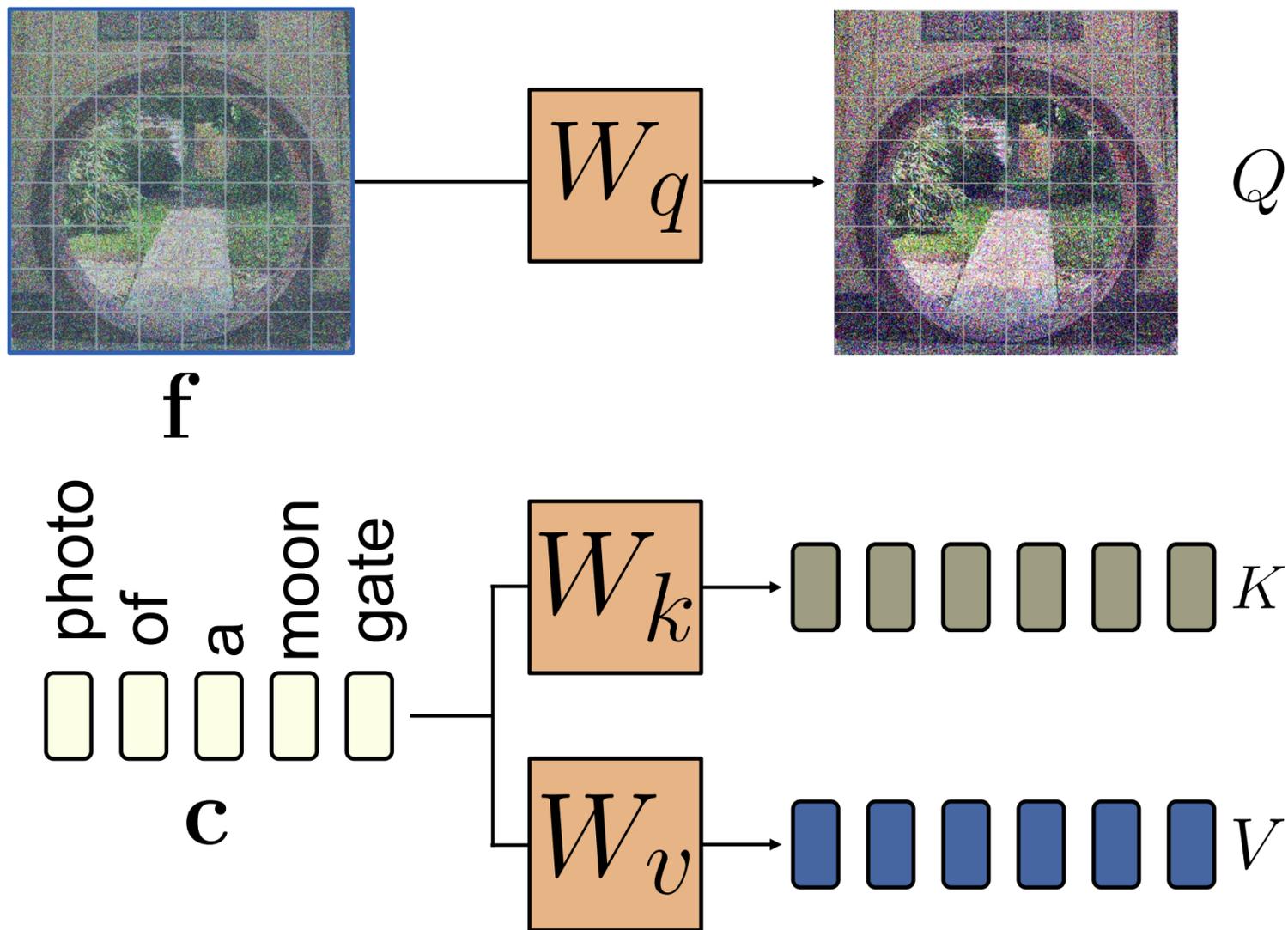


VAES + RNN+ cross-attention

Generating Images from Captions with Attention.

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

Text-Image Cross-Attention



$$Q \text{ Softmax} \left(* \right) = \text{[Six boxes: light gray, dark gray, light gray, white, black, white]}$$

$$= \sum \left(\text{[Six boxes: light gray, dark gray, light gray, white, black, white]} * \text{[Six boxes: blue, blue, blue, blue, blue, blue]} \right)$$

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

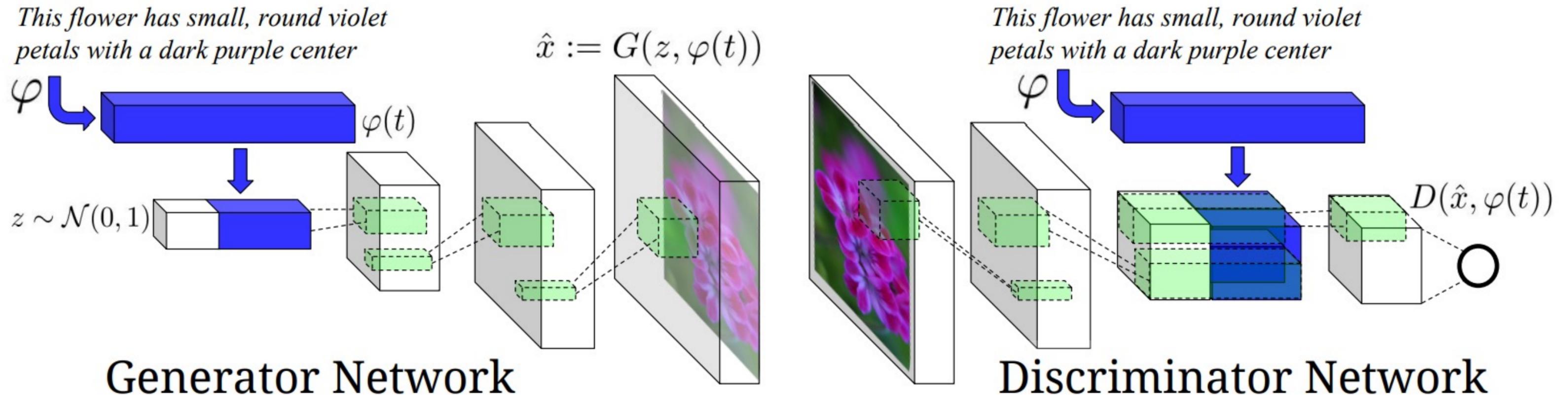
the flower has petals that are bright pinkish purple with white stigma



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



GANs-based Text-to-Image



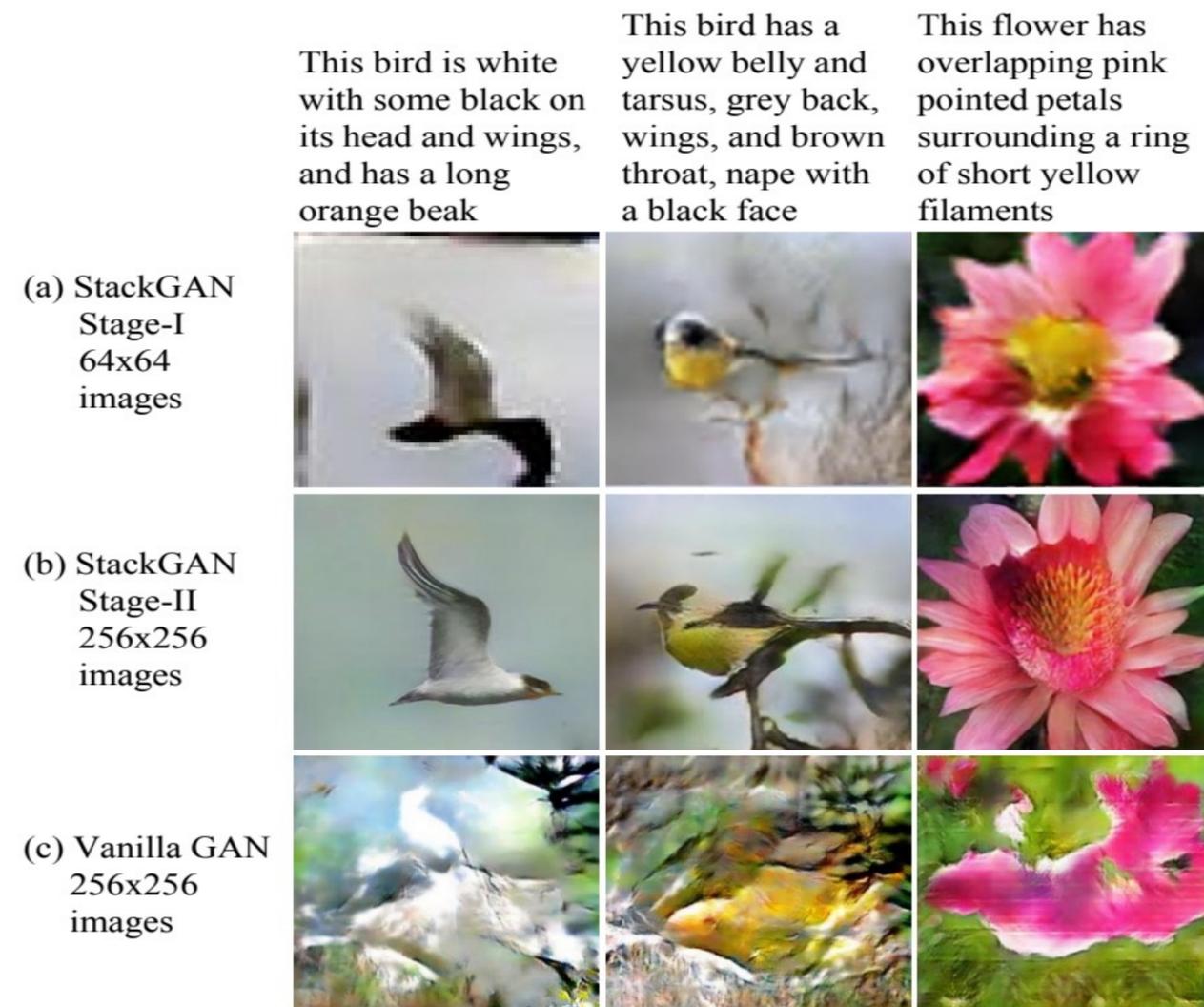
Conditional GAN + CNN + concatenation

Generative Adversarial Text to Image Synthesis

Scott Reed et al., ICML 2016

How to increase resolution?

+Two-stage Models

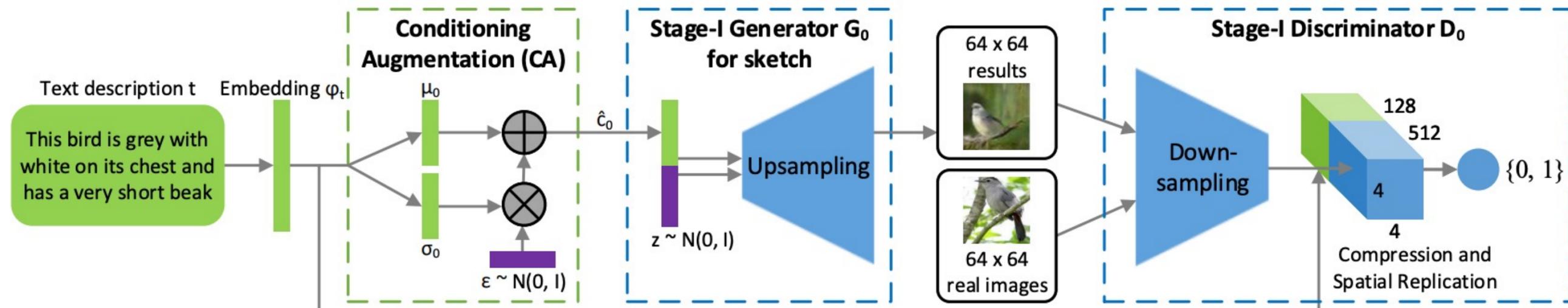


Two-stage Conditional GAN + CNN + concatenation

StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks

Han Zhang et al., ICCV 2017

+Two-stage Models

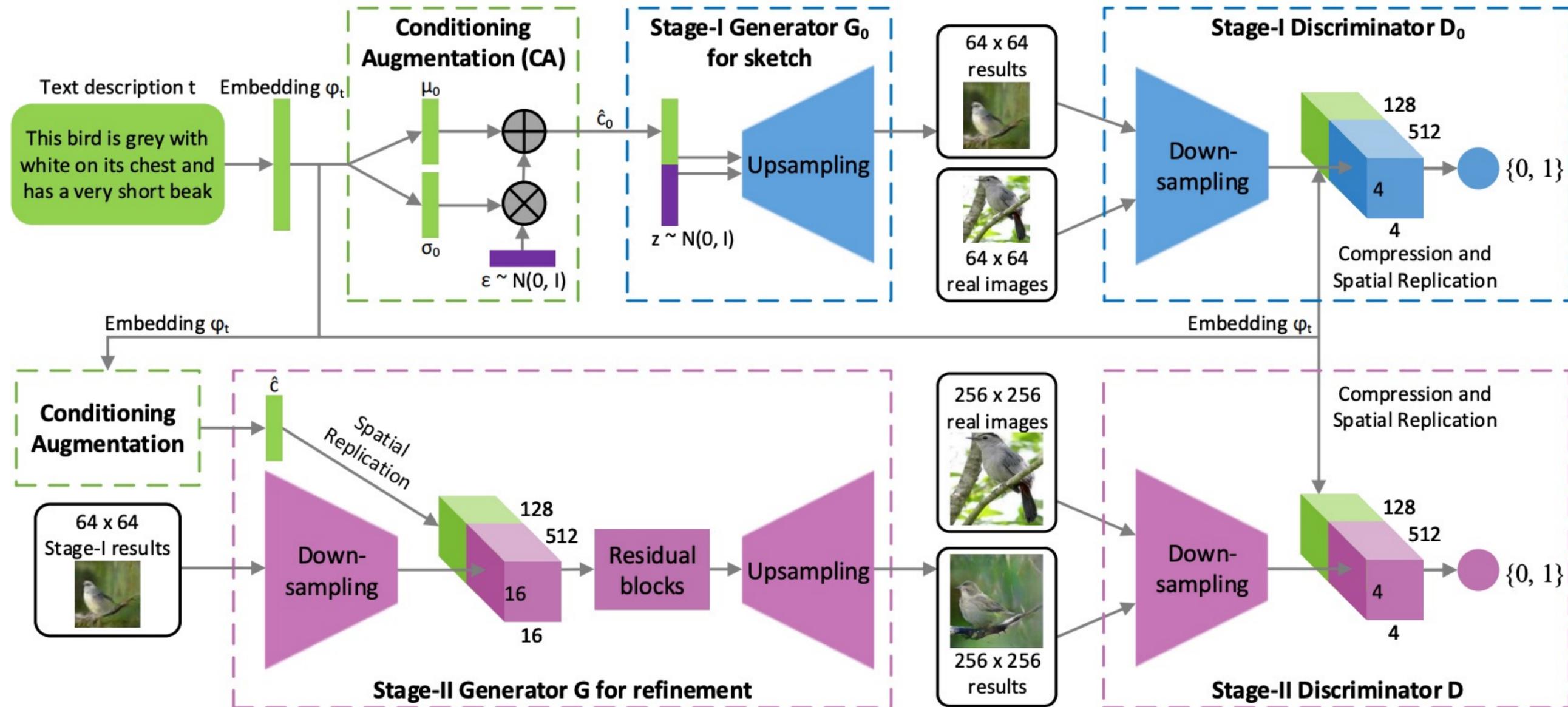


Two-stage Conditional GAN + CNN + concatenation

StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks

Han Zhang et al., ICCV 2017

+Two-stage Models

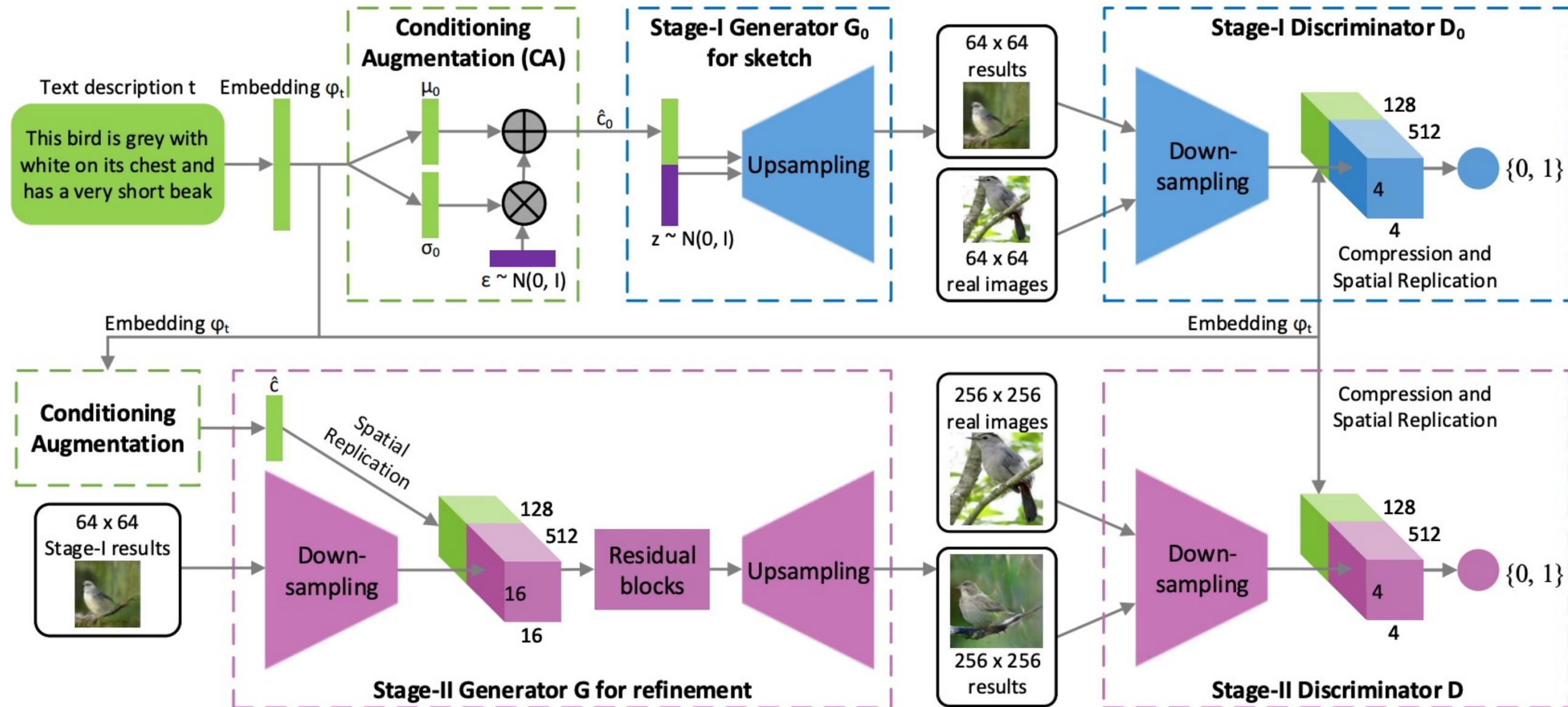


Two-stage Conditional GAN + CNN + concatenation

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



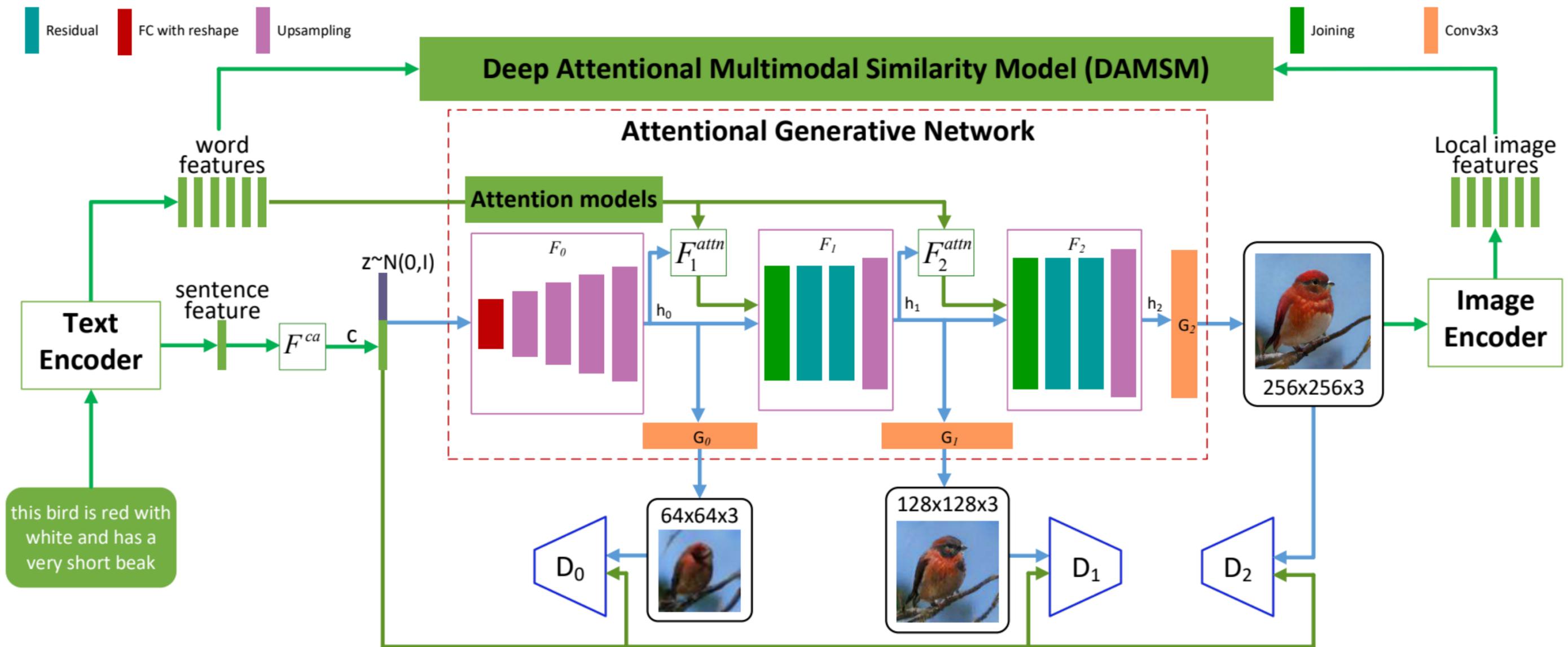
+Two-stage Models



+ 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

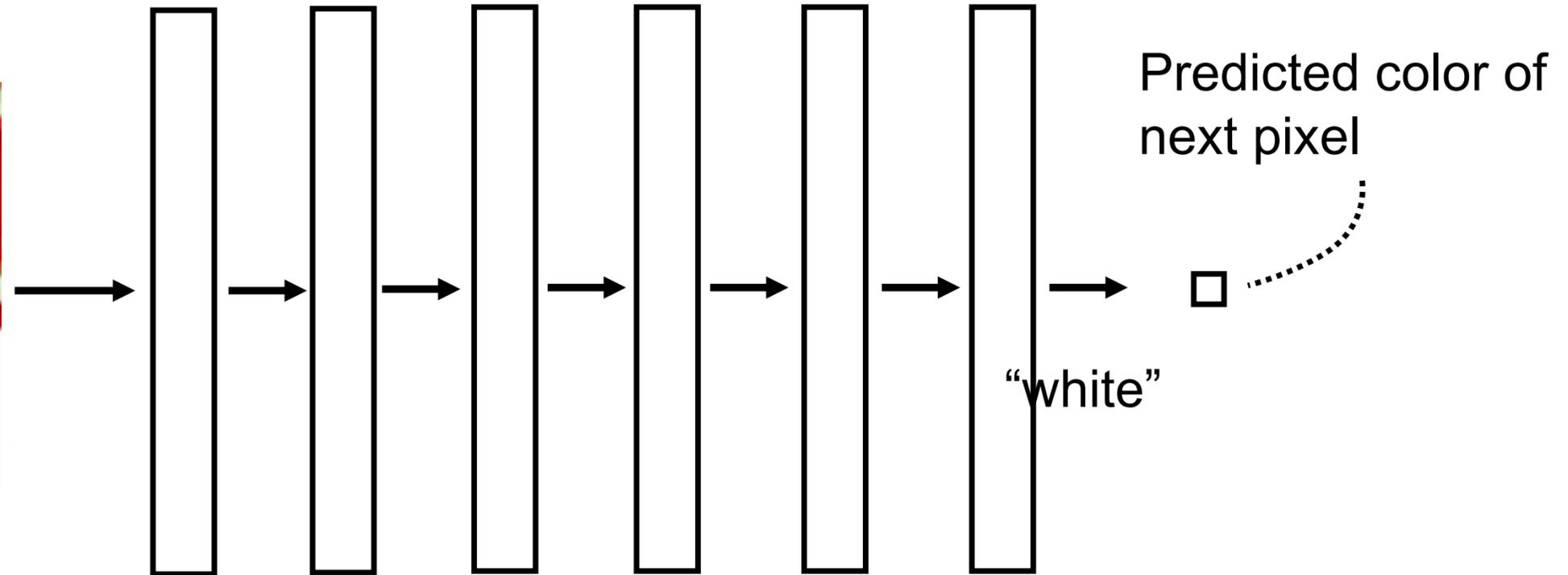
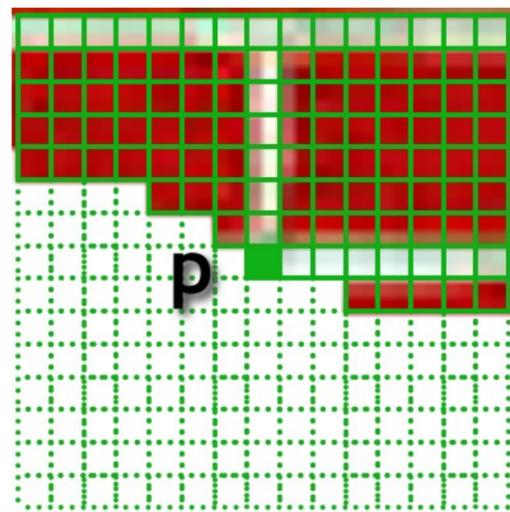
Designed to learn long-range interactions on sequential data, transformers have achieved state-of-the-art results

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, the locality of interactions and is therefore free from the constraints of inductive bias. However,

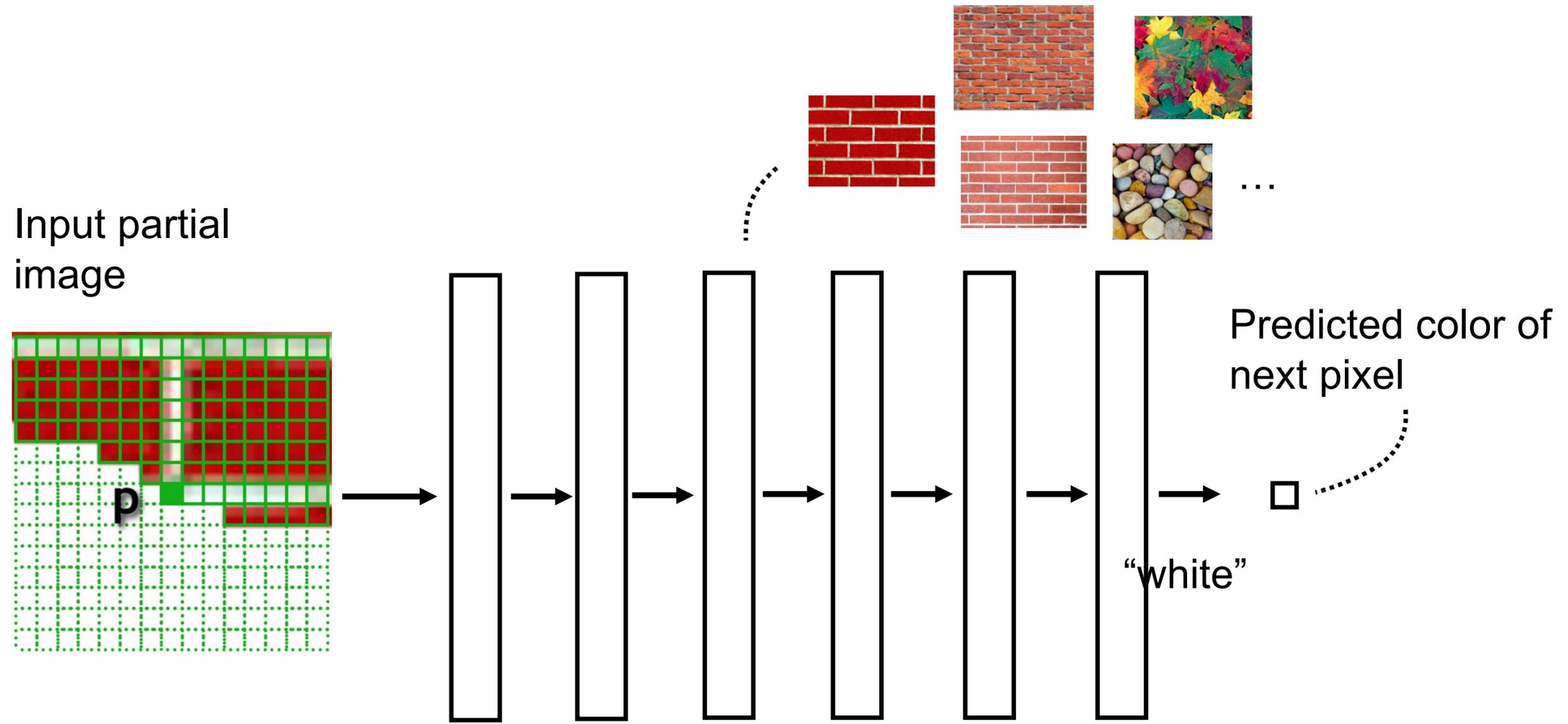
033 [cs.CV] 23 Jun 2021

Autoregressive (AR) image synthesis

Input partial
image

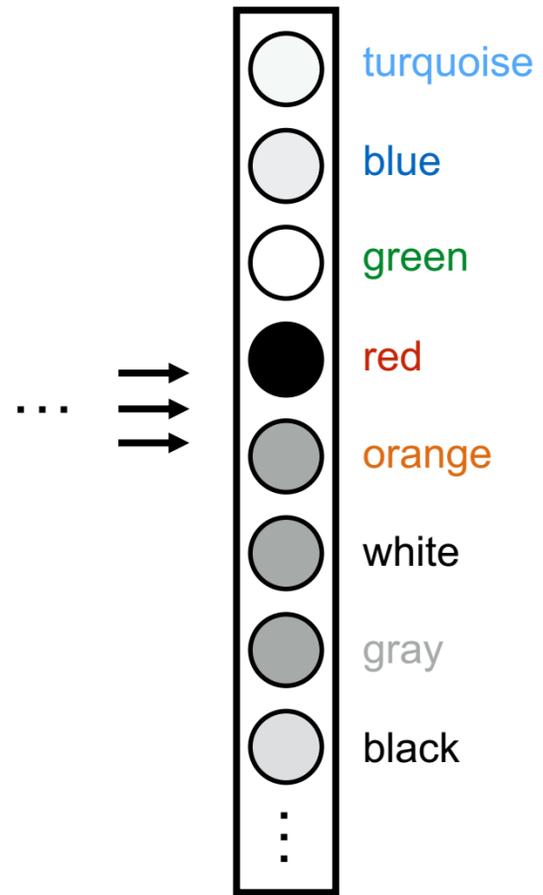


[PixelRNN, PixelCNN, van der Oord et al. 2016]



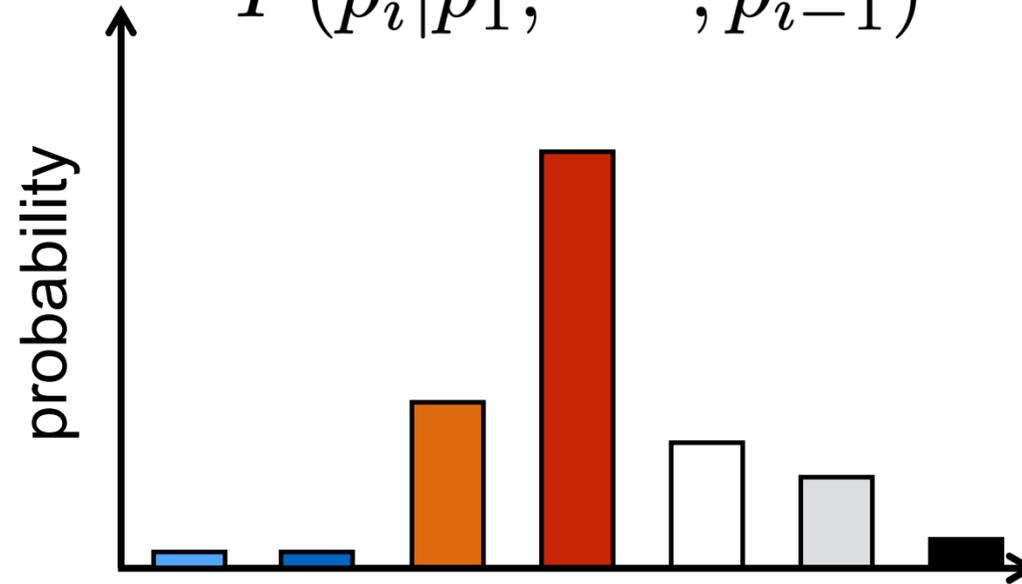
[PixelRNN, PixelCNN, van der Oord et al. 2016]

Network output

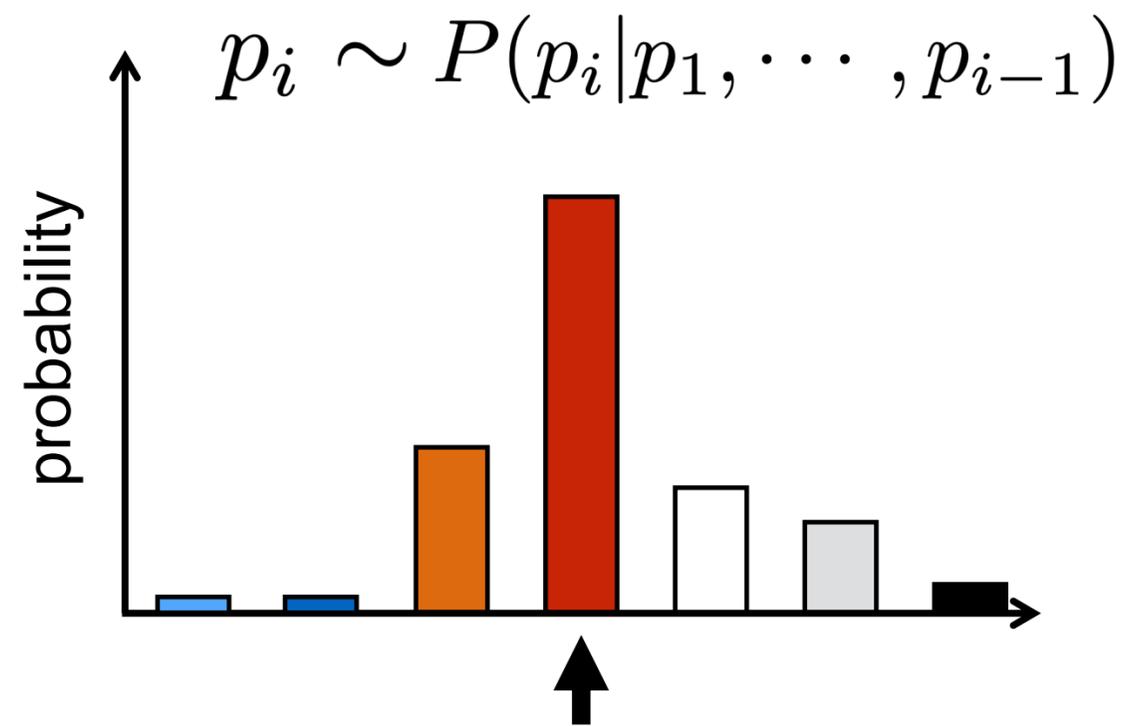
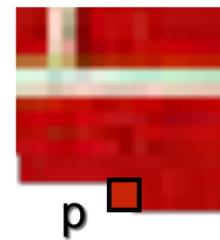
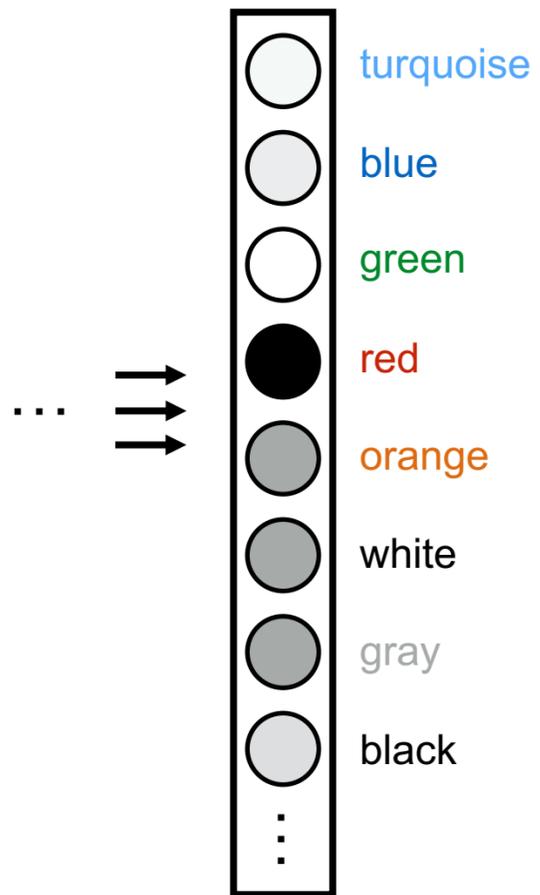


P(next pixel | previous pixels)

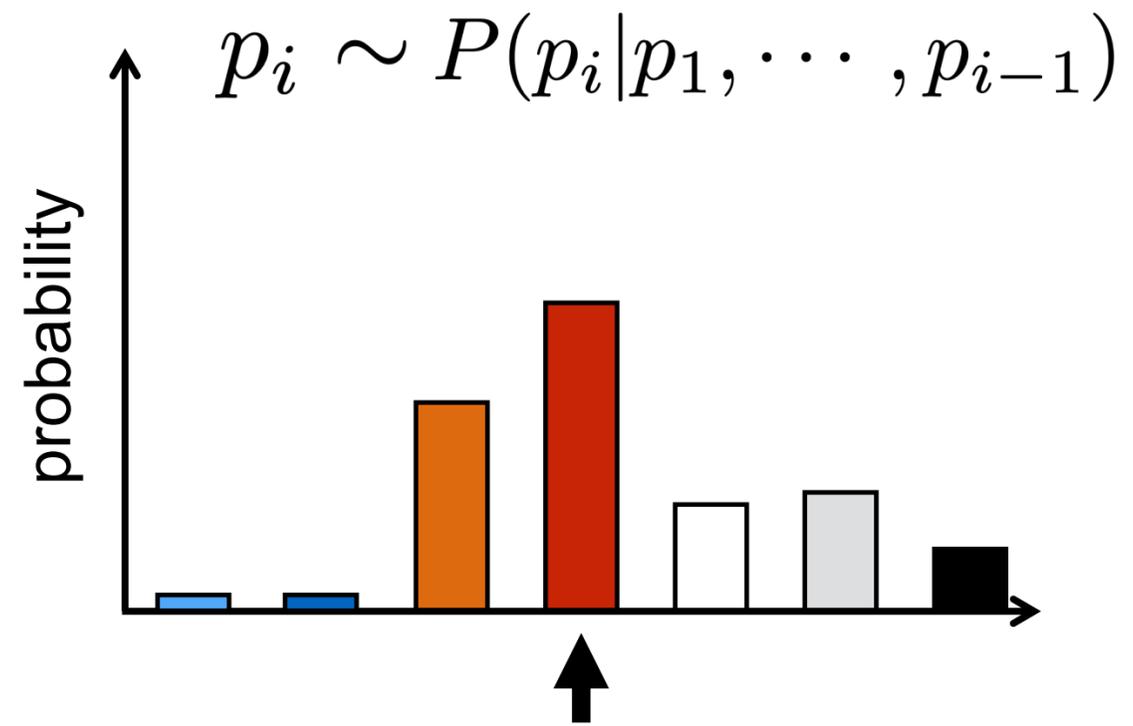
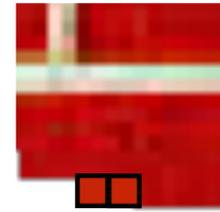
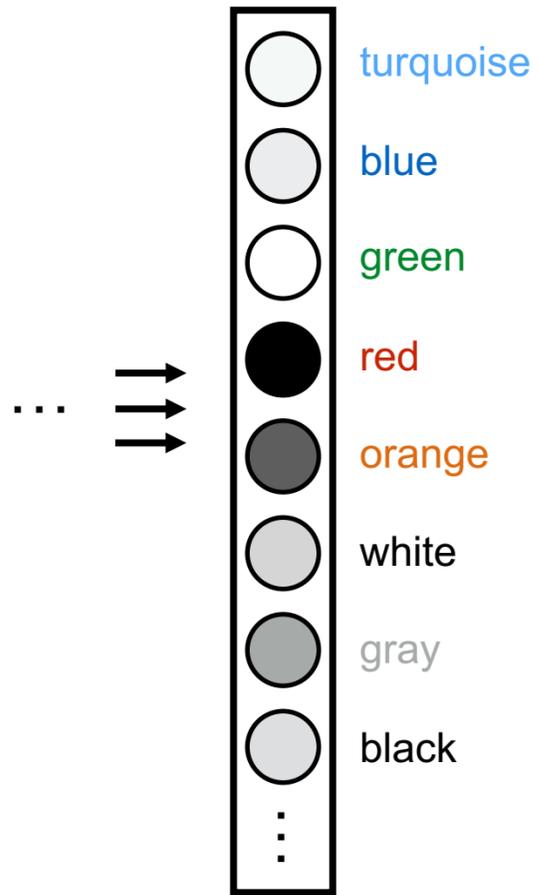
$$P(p_i | p_1, \dots, p_{i-1})$$



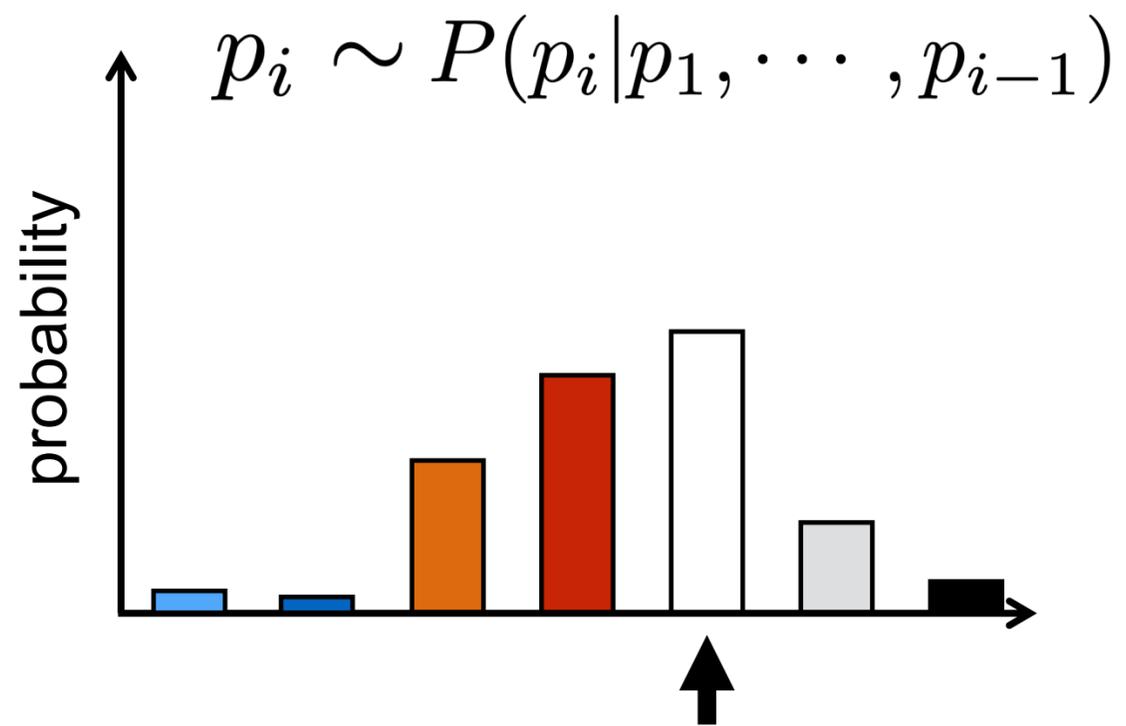
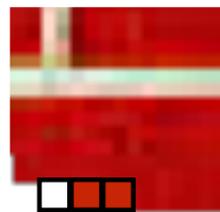
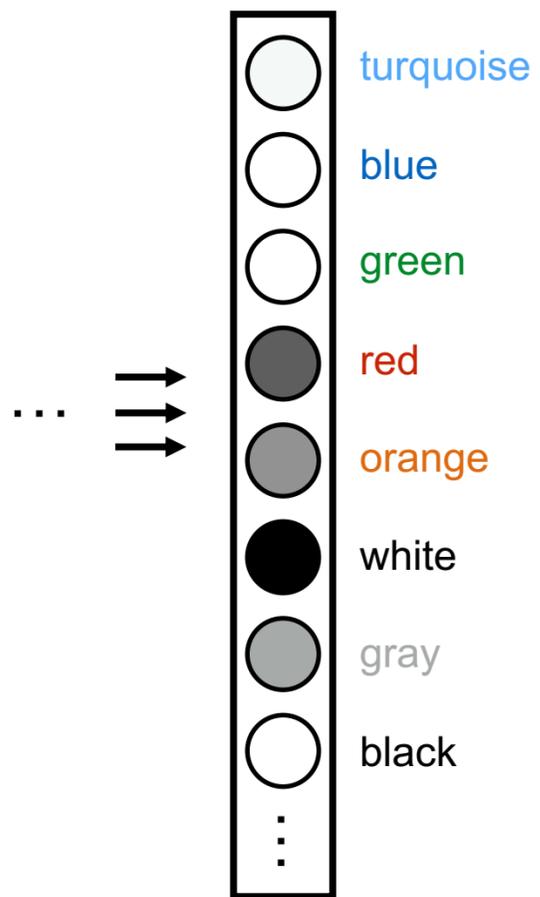
Network output



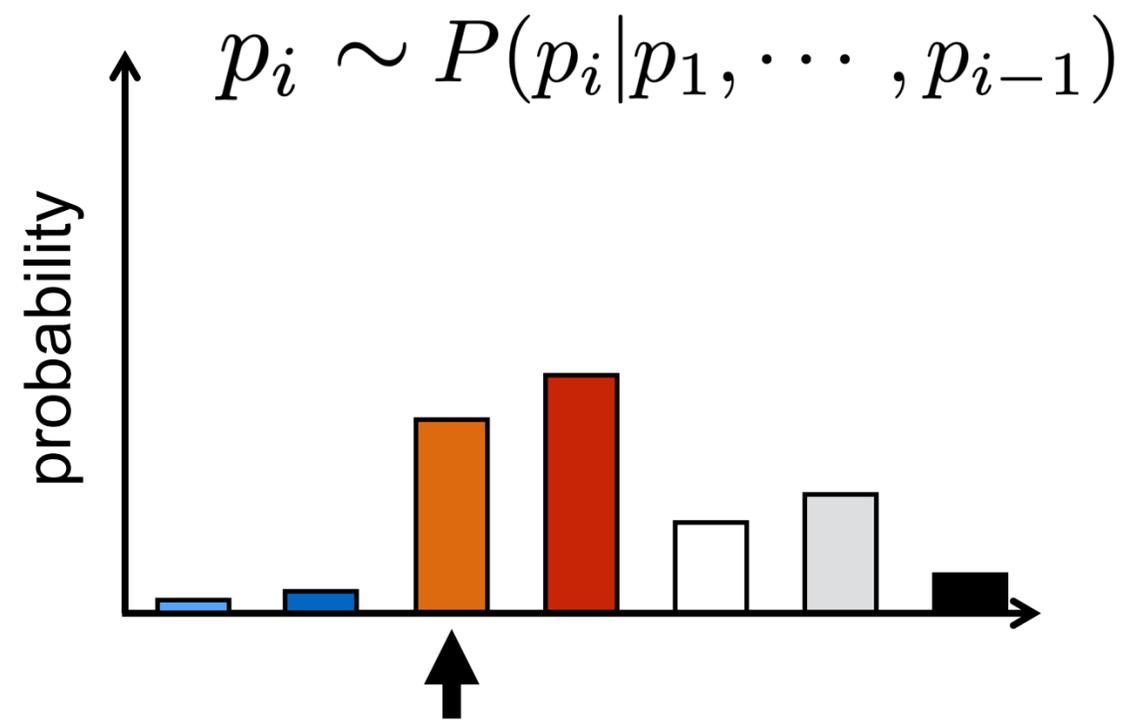
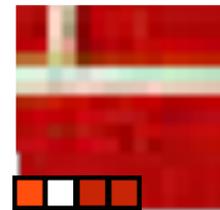
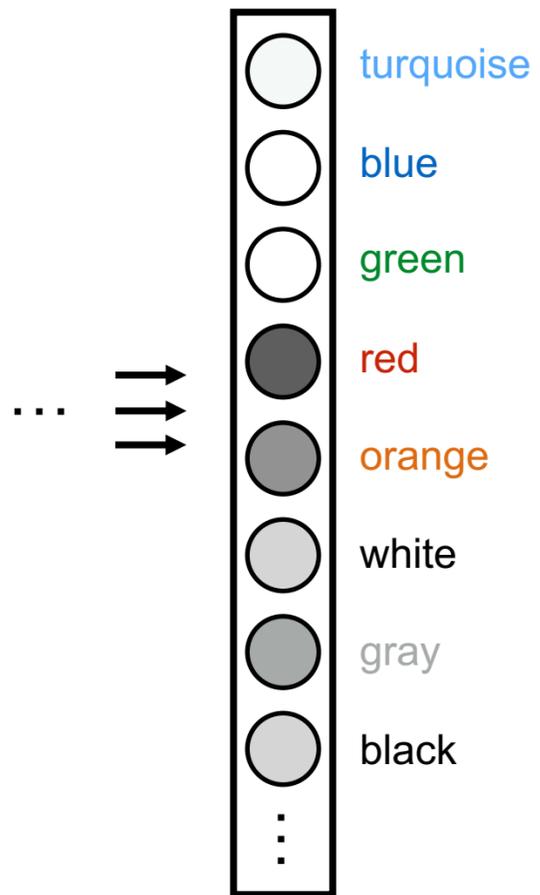
Network output



Network output



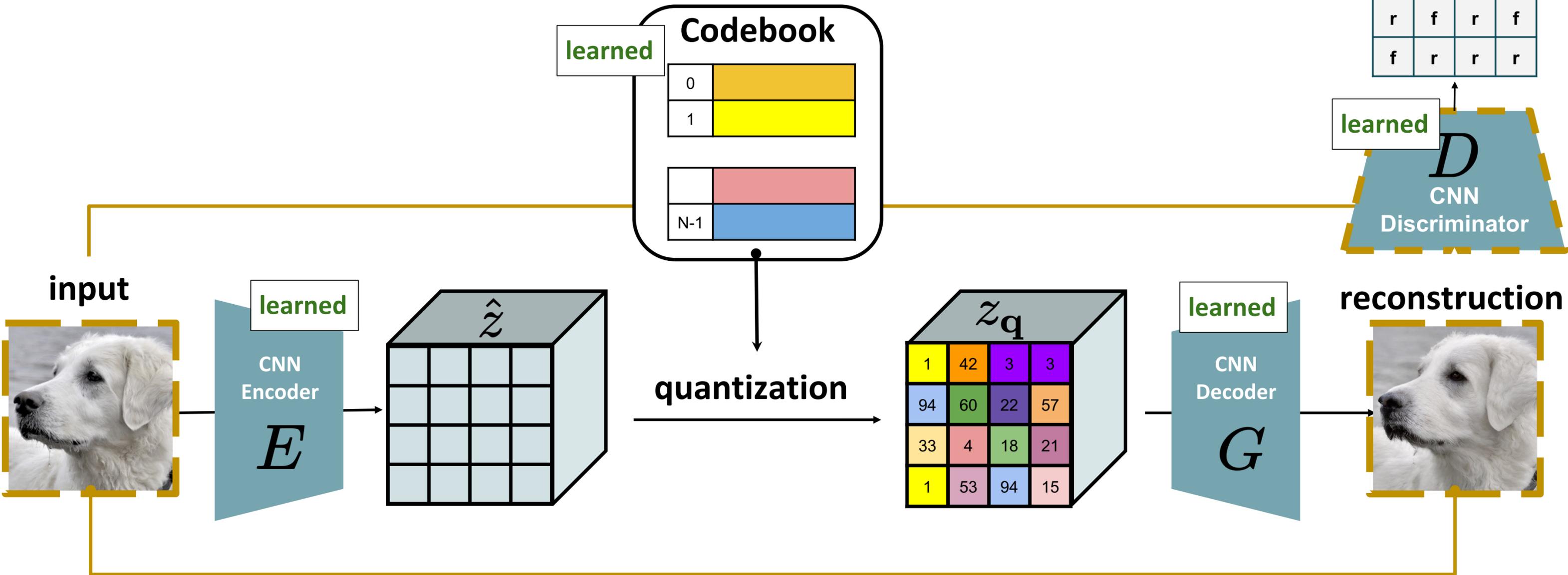
Network output



Generation is super slow?
What should we do?

From VQ-VAE¹ to VQGAN

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

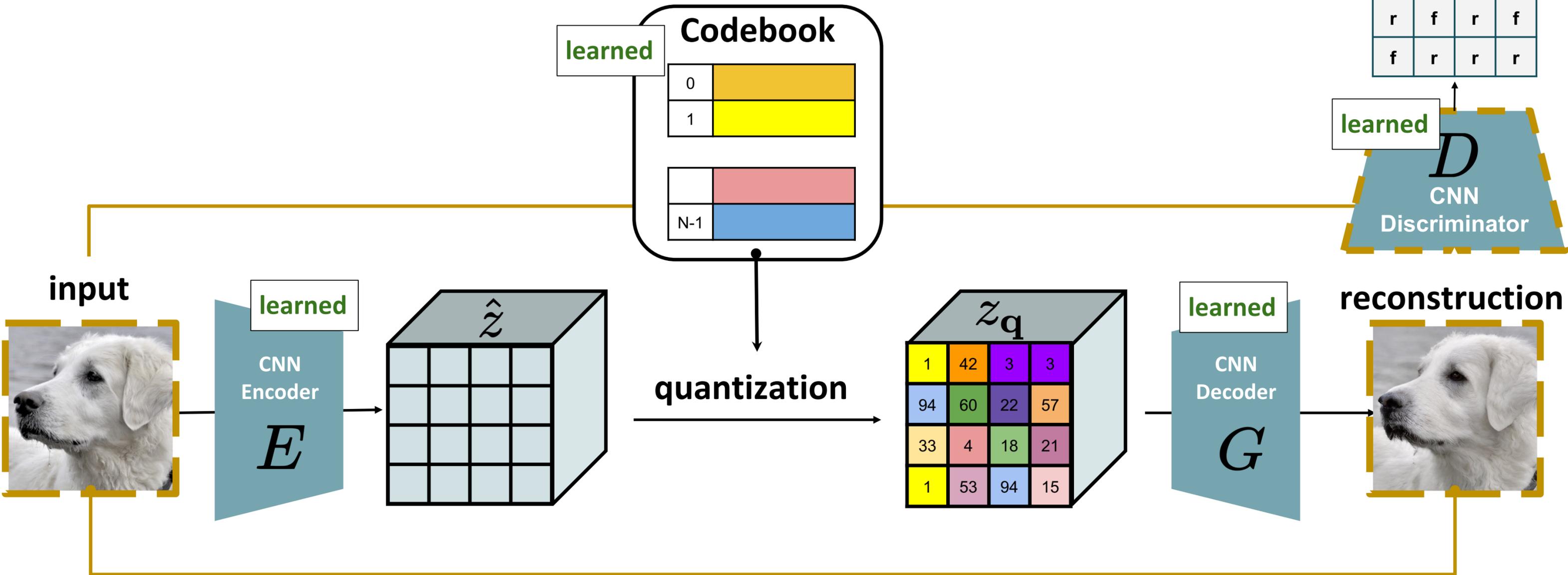


i) replace L2/L1 rec. loss with Perceptual loss (includes pixel-level)

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

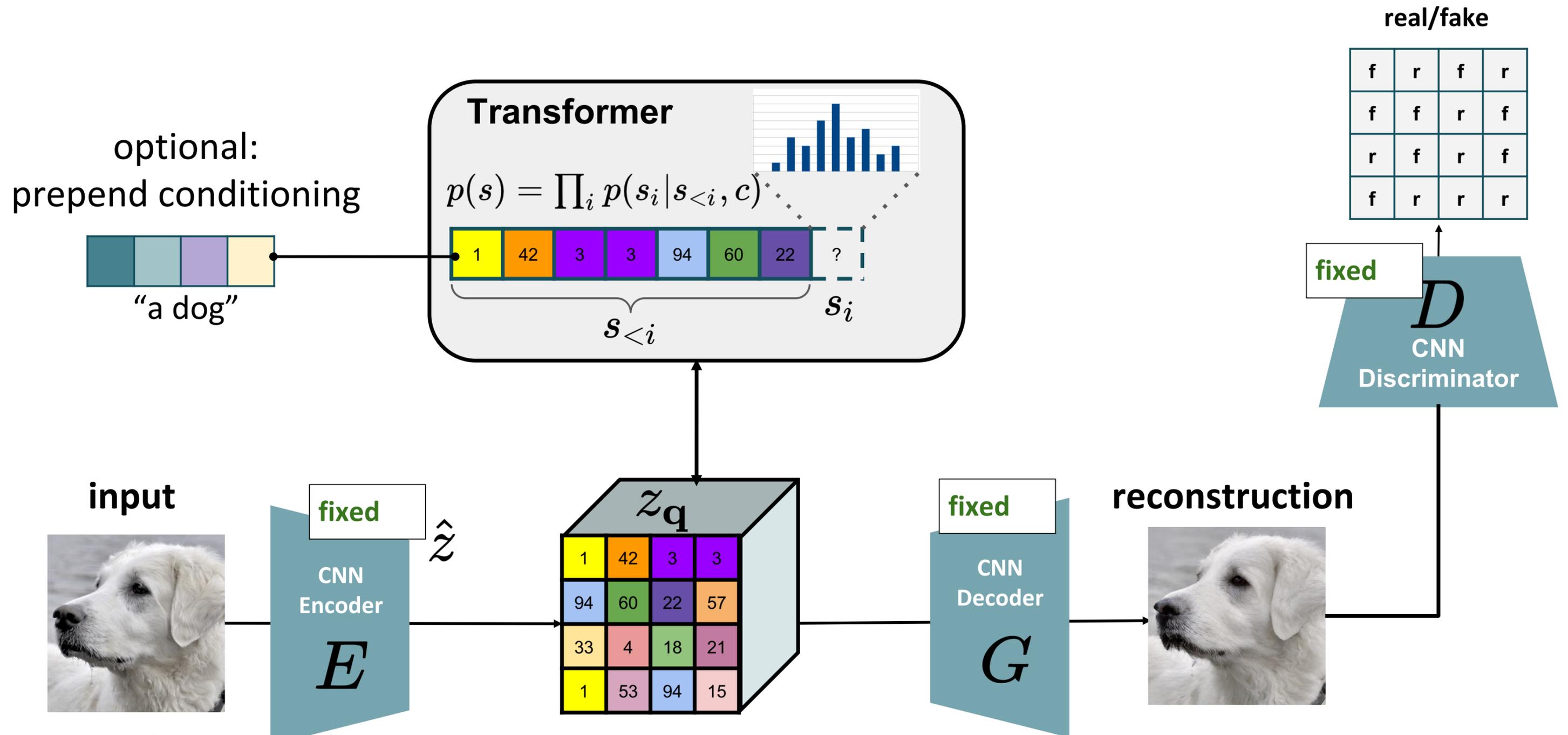
From VQ-VAE¹ to VQGAN

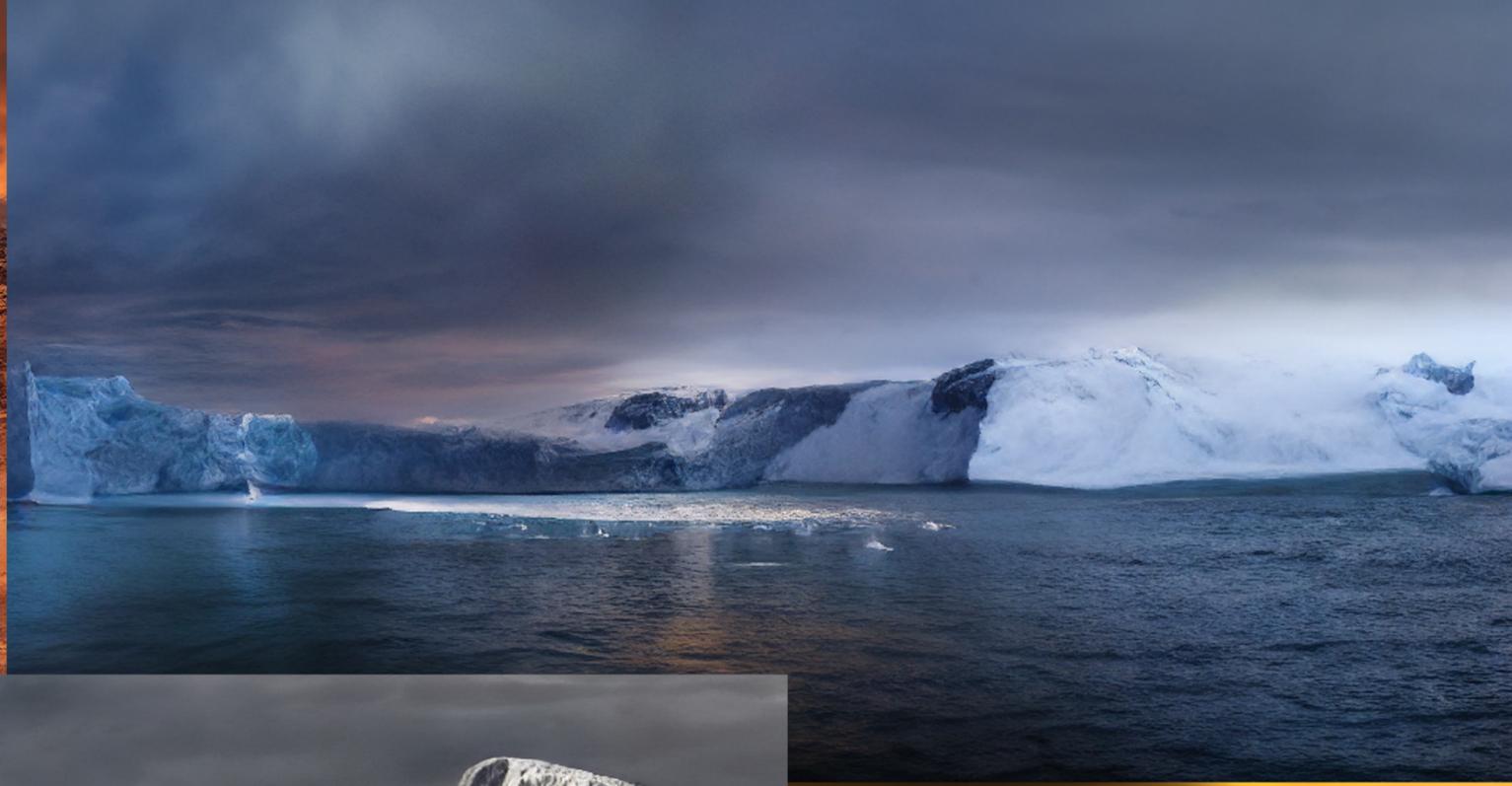
¹: Neural Discrete Representation Learning, v.d.Oord et al, <https://arxiv.org/abs/1711.00937>



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

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

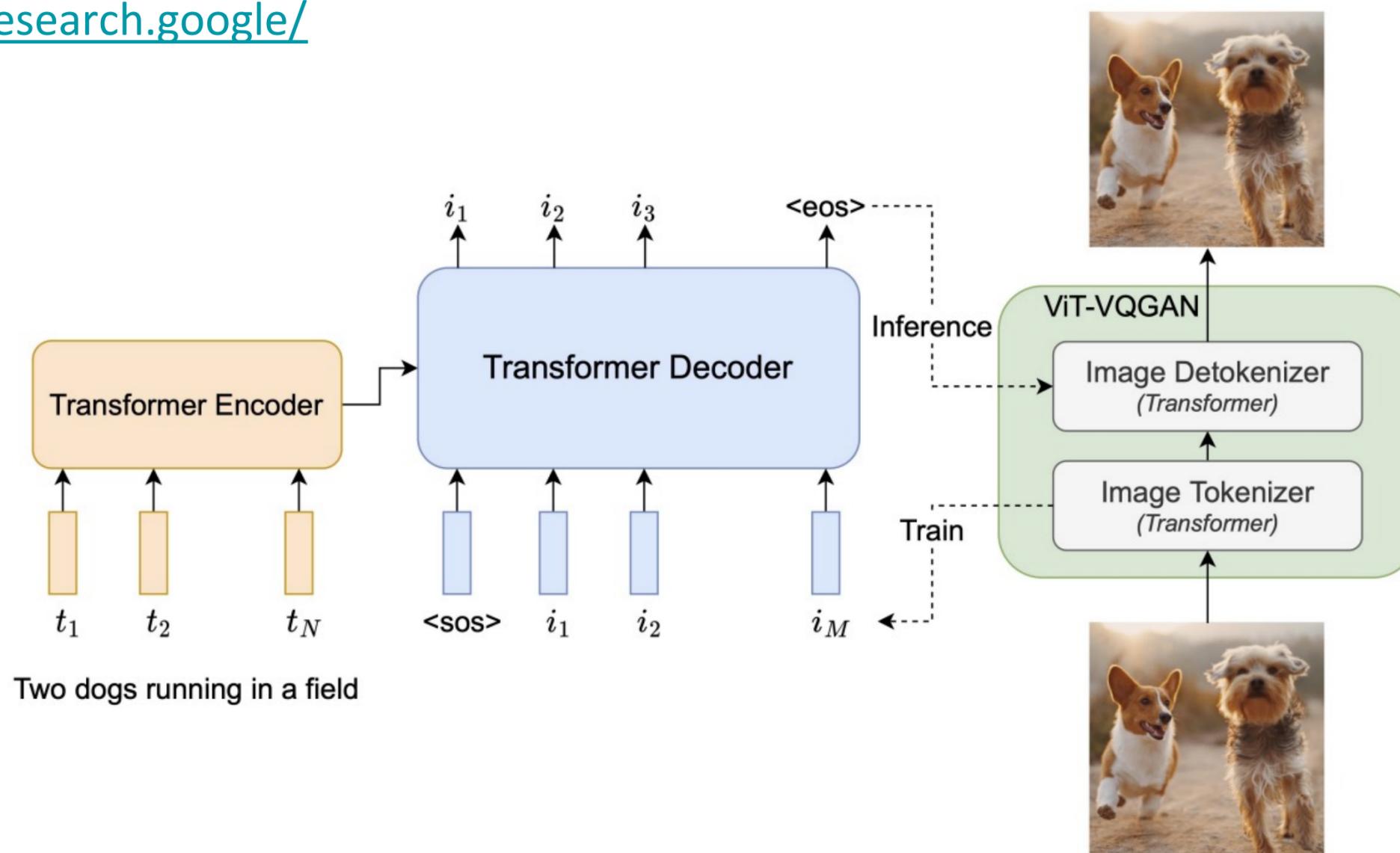
20B



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/>



Another Approach: Diffusion Models!

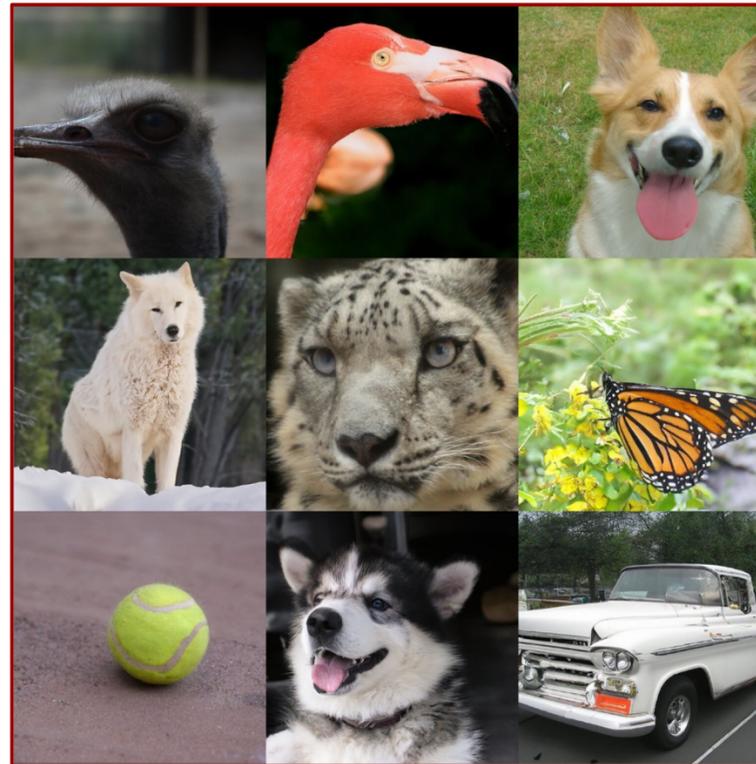
great results for image synthesis



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>

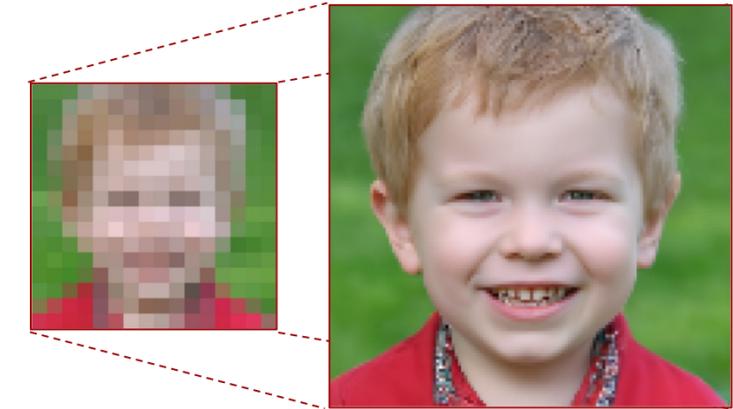


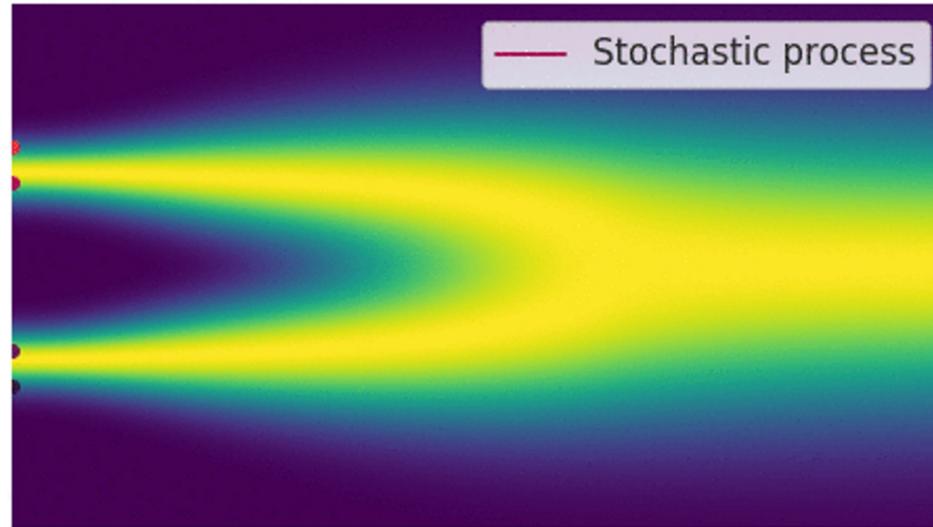
Image Super-Resolution via Iterative Refinement

Chitwan Saharia, et al

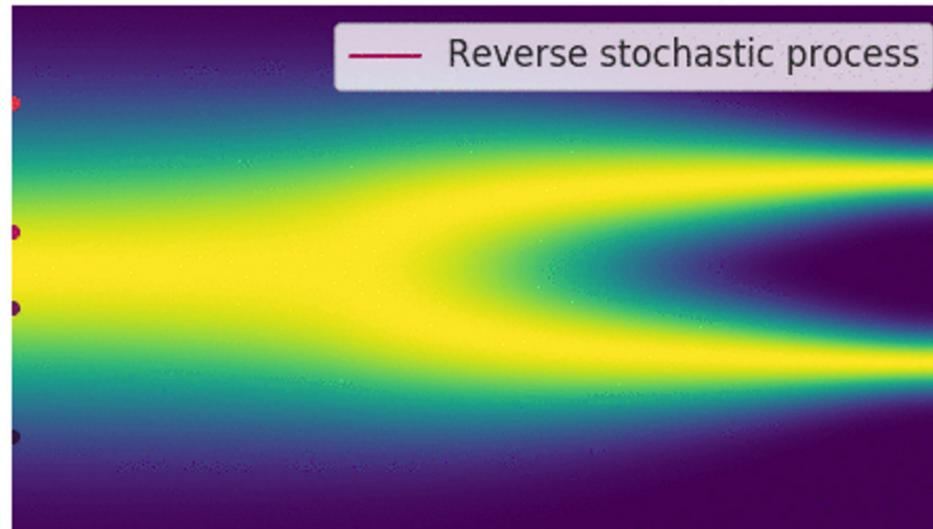
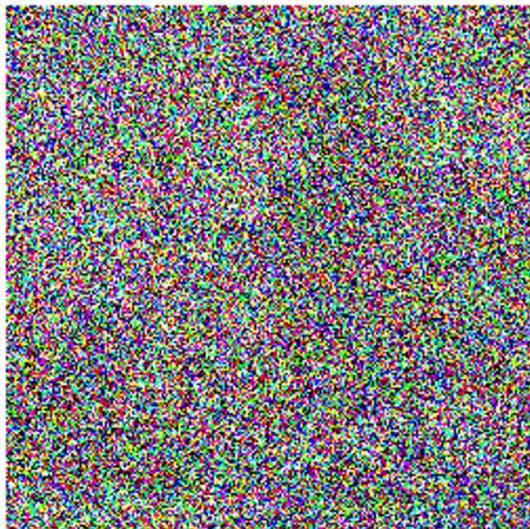
<https://arxiv.org/abs/2104.07636>

... but very expensive :(

Brief Overview of Diffusion Models



- “destroy” the data by gradually adding small amounts of gaussian noise



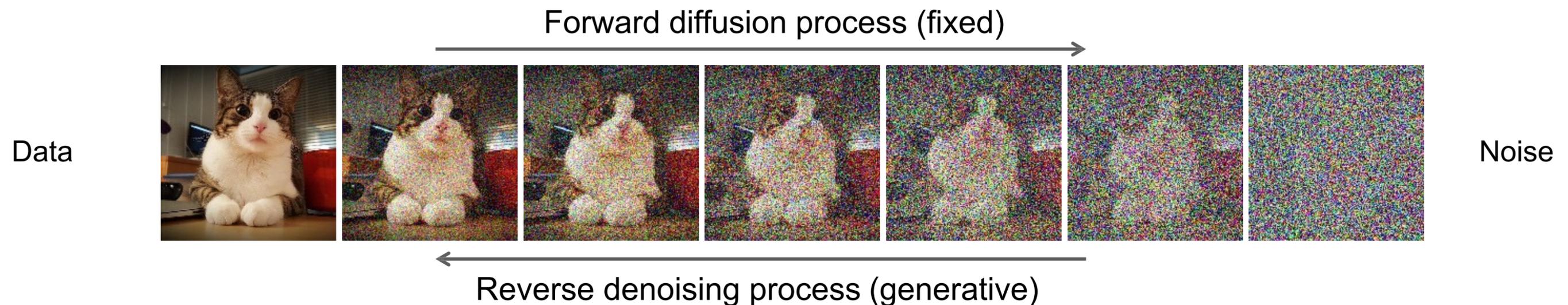
- “create” data by gradually denoising a noisy code from a stationary distribution

Denoising Diffusion Models

Learning to generate by denoising

Denoising diffusion models consist of two processes:

- Forward diffusion process that gradually adds noise to input
- Reverse denoising process that learns to generate data by denoising



[Sohl-Dickstein et al., Deep Unsupervised Learning using Nonequilibrium Thermodynamics, ICML 2015](#)

[Ho et al., Denoising Diffusion Probabilistic Models, NeurIPS 2020](#)

[Song et al., Score-Based Generative Modeling through Stochastic Differential Equations, ICLR 2021](#)

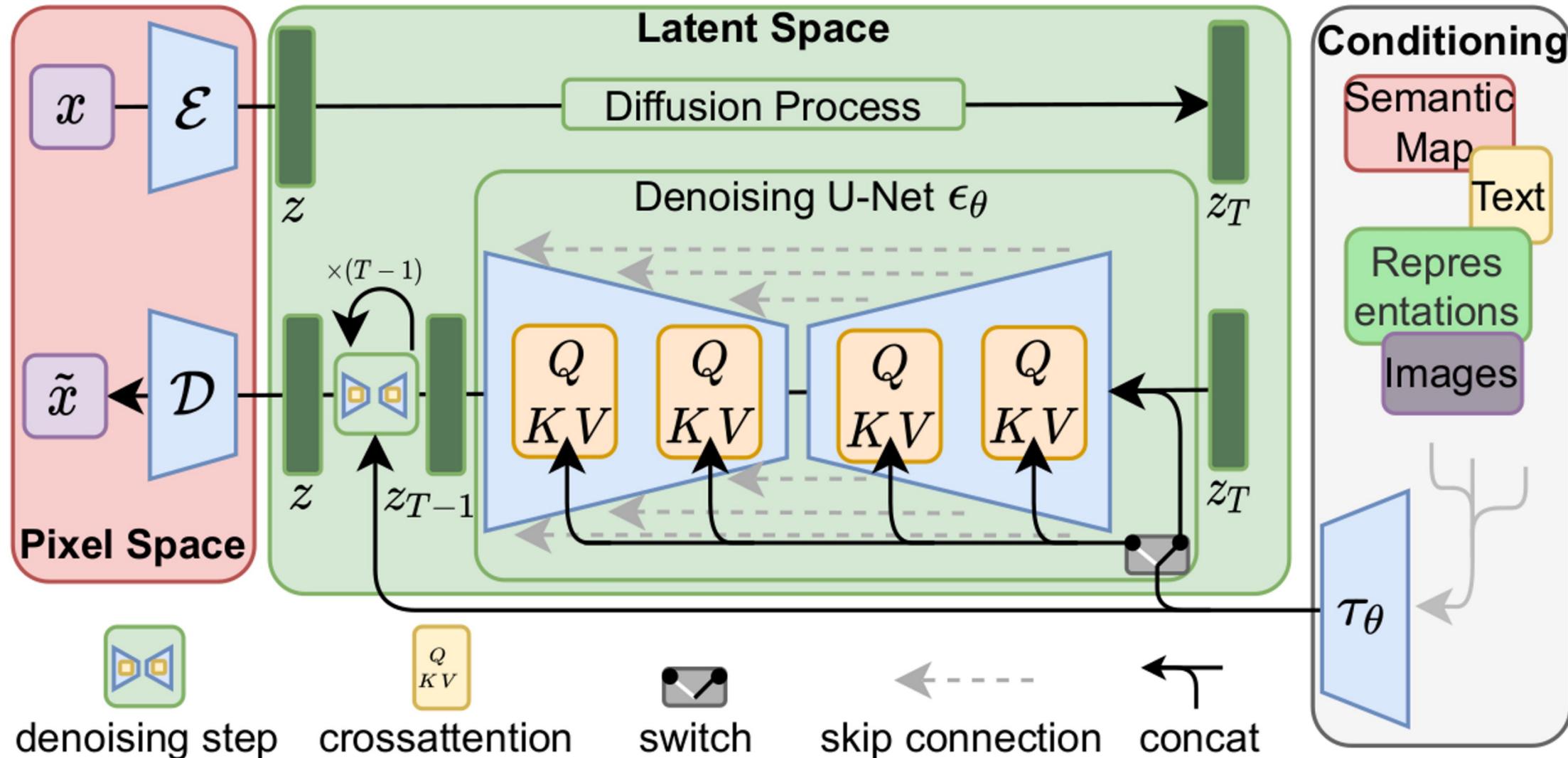
Latent Diffusion Modeling: Architecture

Autoencoder with KL or VQ regularization.

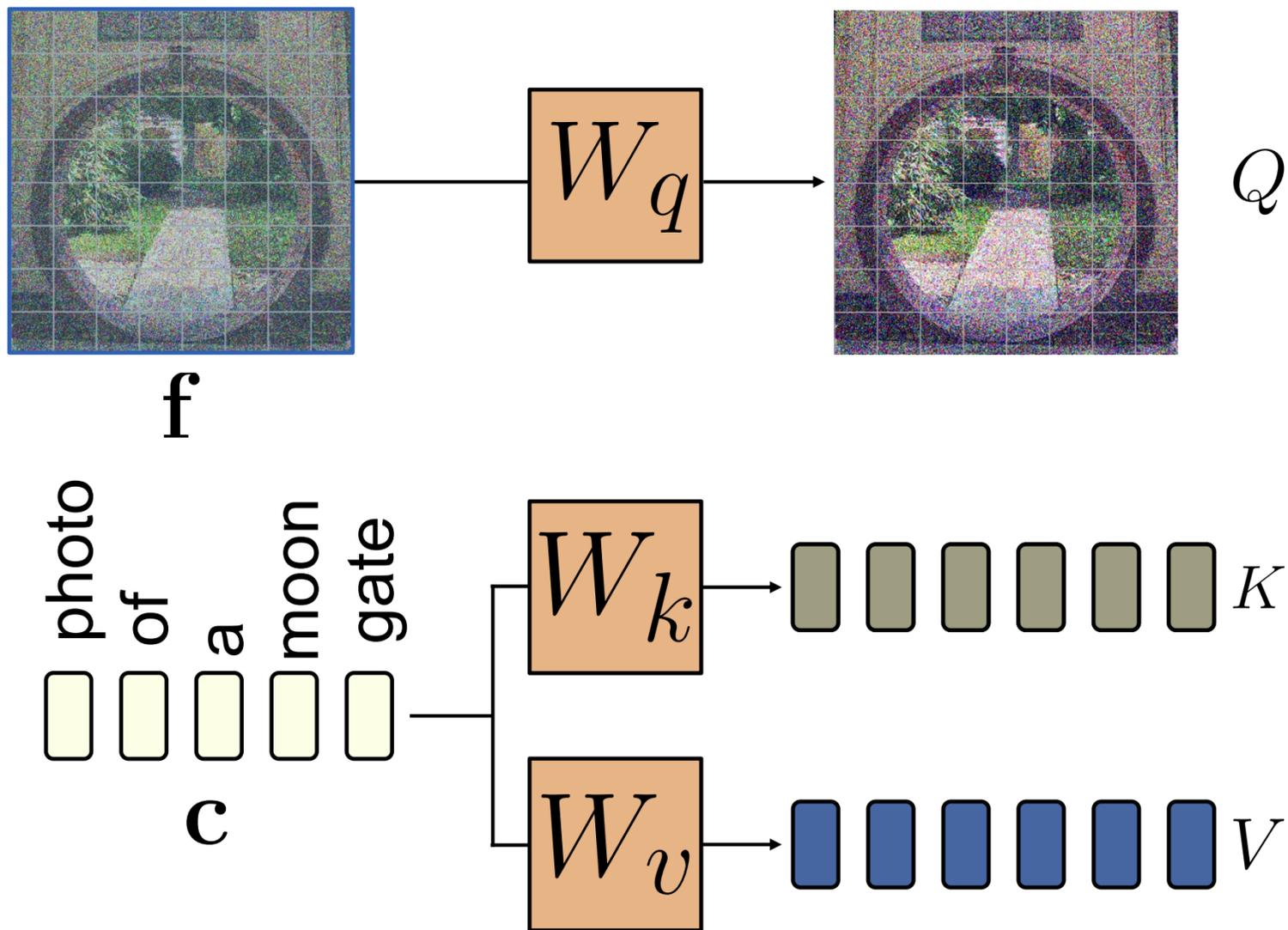
VQ-reg.: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{VQ}} + \lambda \mathcal{L}_{\text{GAN}}$

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

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



Text-Image Cross-Attention



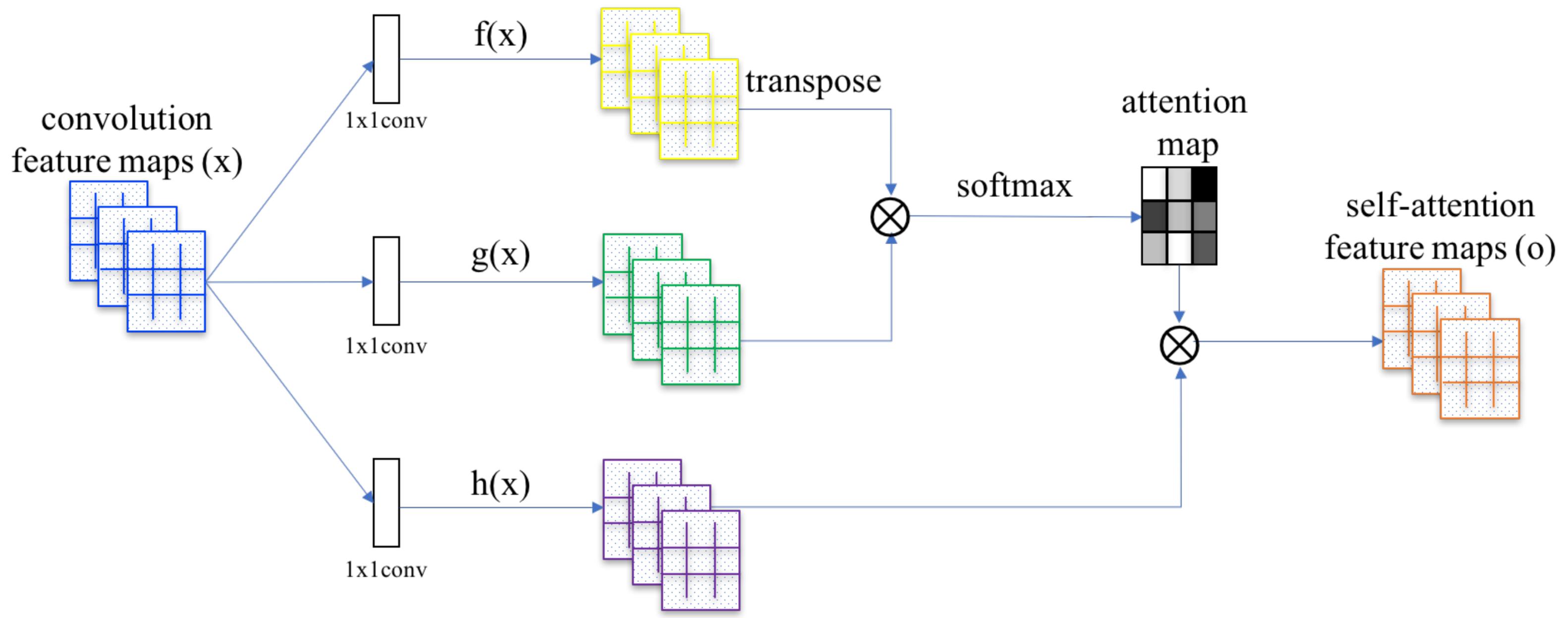
$$Q \text{ Softmax} \left(* \right) = \text{[6 grayscale boxes]}$$

$$= \sum \left(\text{[6 grayscale boxes]} * \text{[6 blue boxes]} \right)$$

i.e.

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

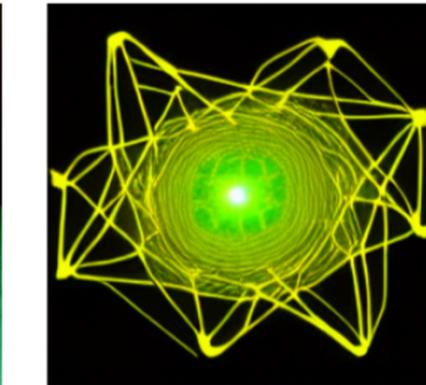
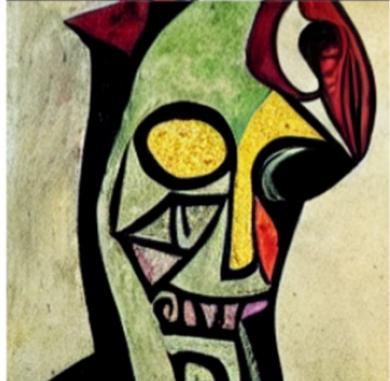
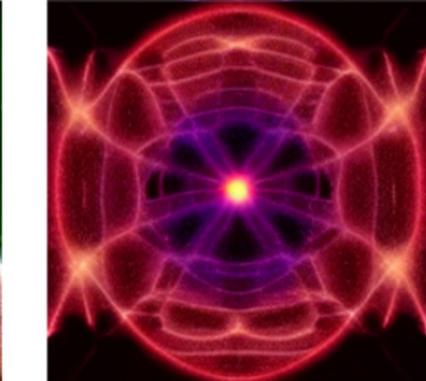
(Spatial) Self-attention Layer



LDMs for Text-to-Image Synthesis

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

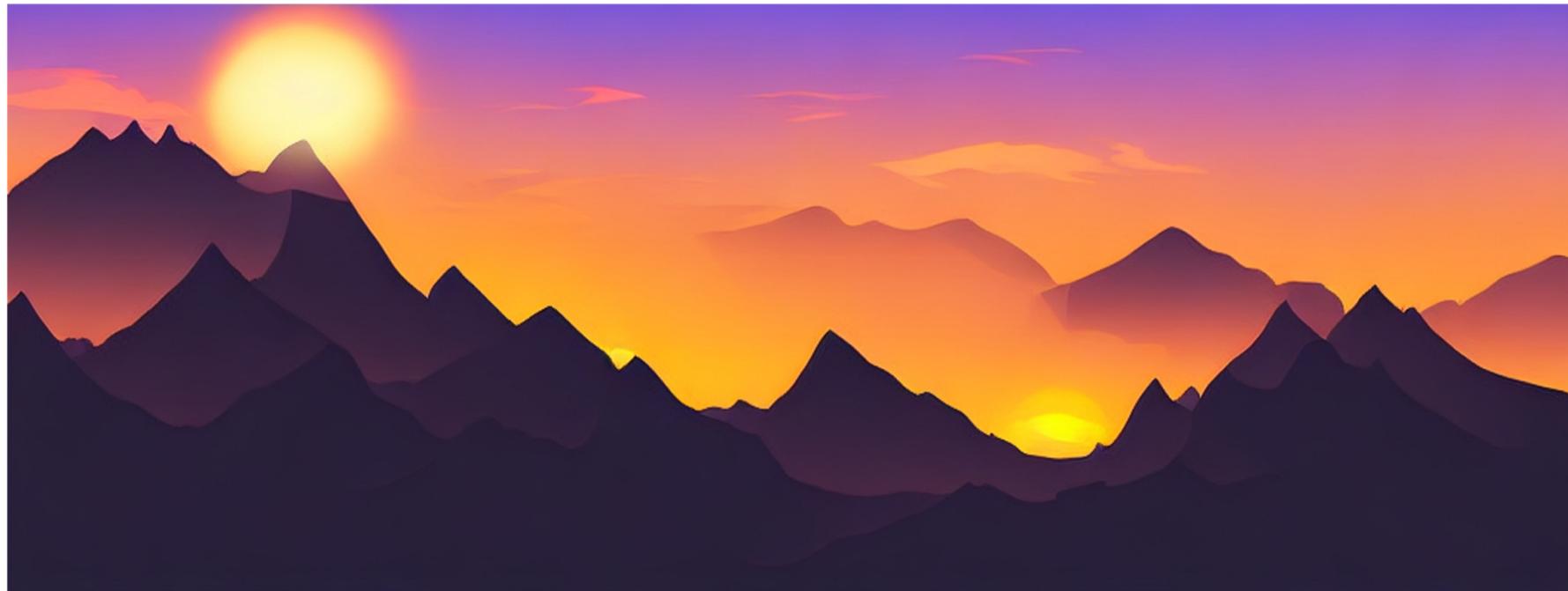
Text-to-Image Synthesis on LAION. 1.4B Model.

<i>'A zombie in the style of Picasso'</i>	<i>'An image of an animal half mouse half octopus'</i>	<i>'An illustration of a slightly conscious neural network.'</i>	<i>'A painting of a squirrel eating a burger.'</i>	<i>'A watercolor painting of a chair that looks like an octopus.'</i>	<i>'A shirt with the inscription: "I love generative models!"'</i>
					
					
					

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:

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

e.g. empty string

$s = 1.0$

$s = 7.5$





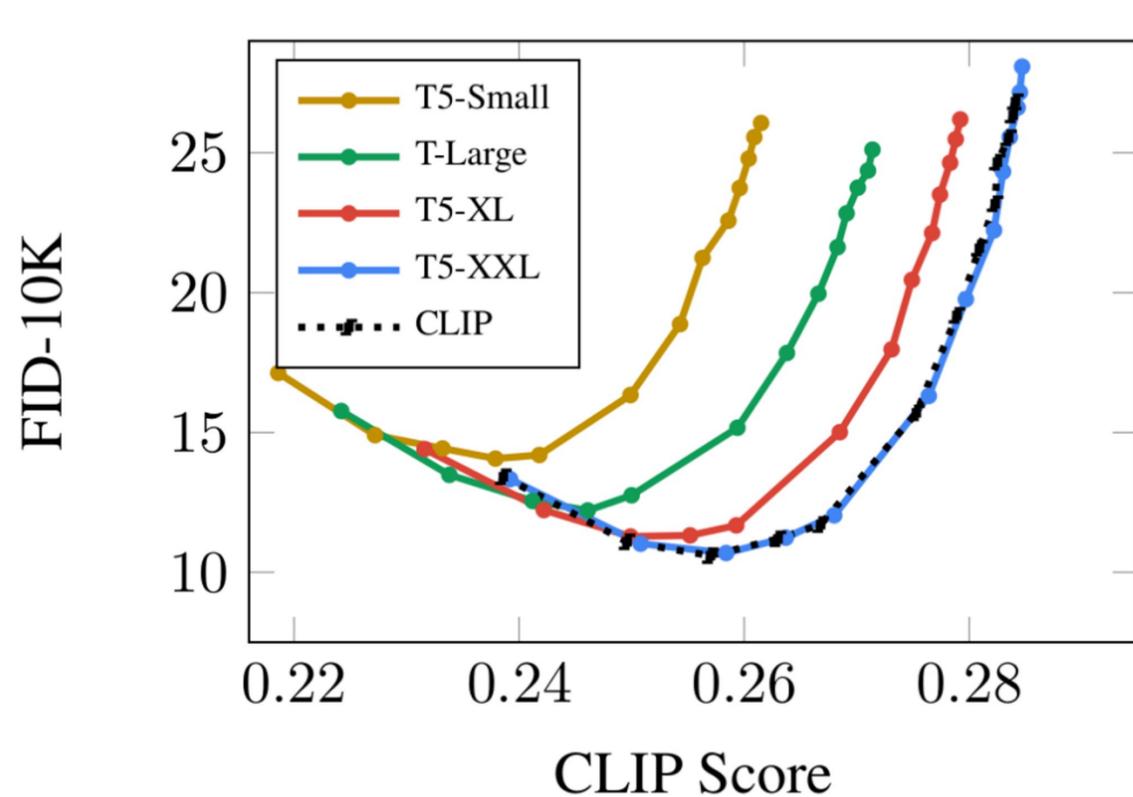
Stable Diffusion

Latent Diffusion ++

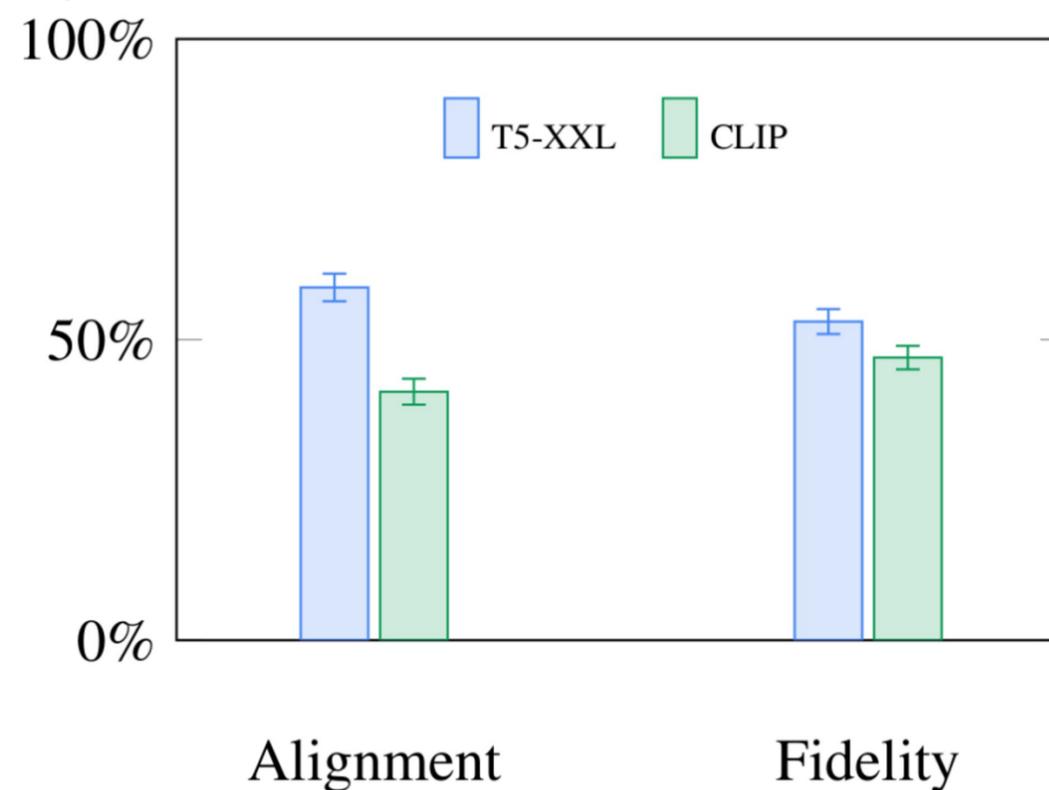


From Latent to Stable Diffusion

- 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.

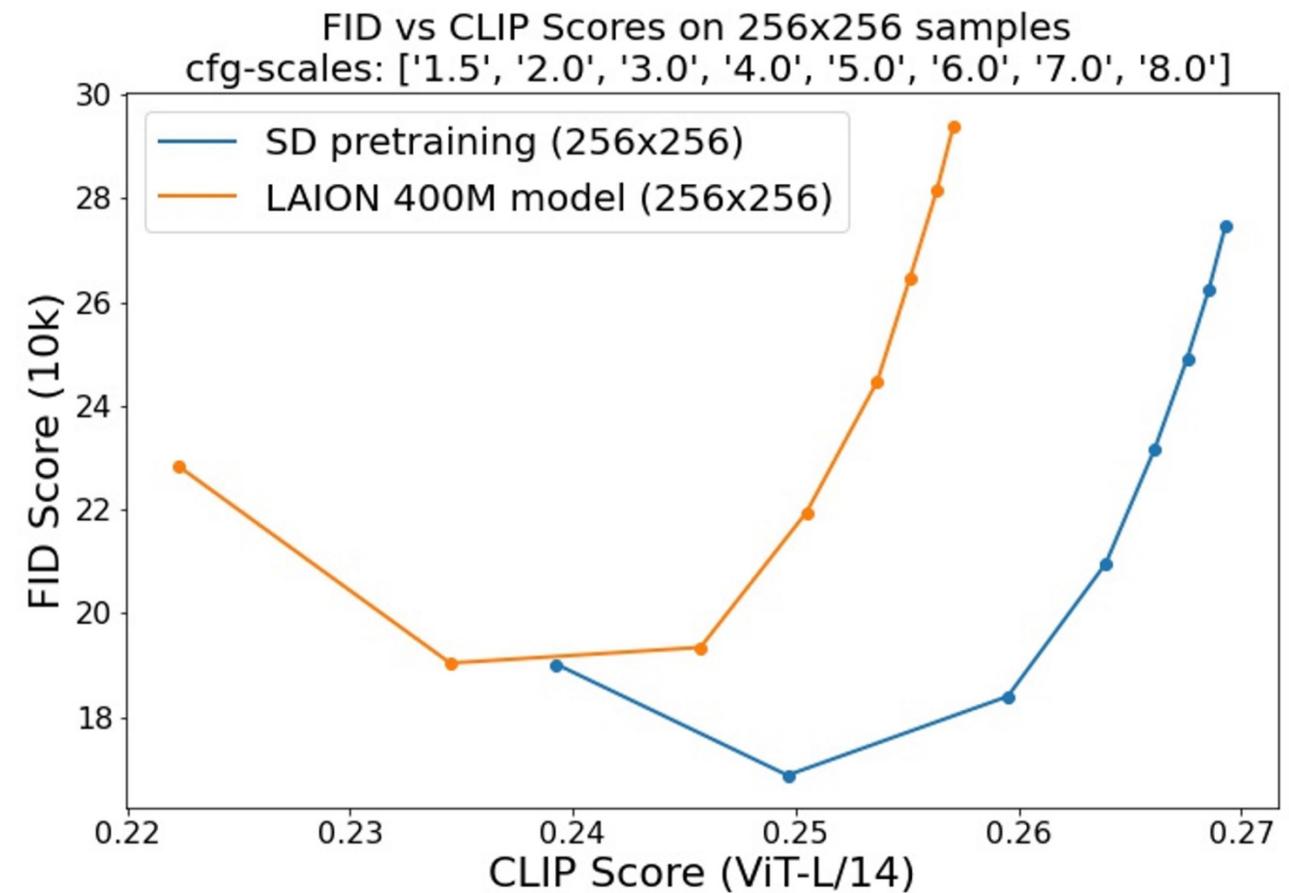


(b) Comparing T5-XXL and CLIP on DrawBench.

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 $\geq 1024 \times 1024$).

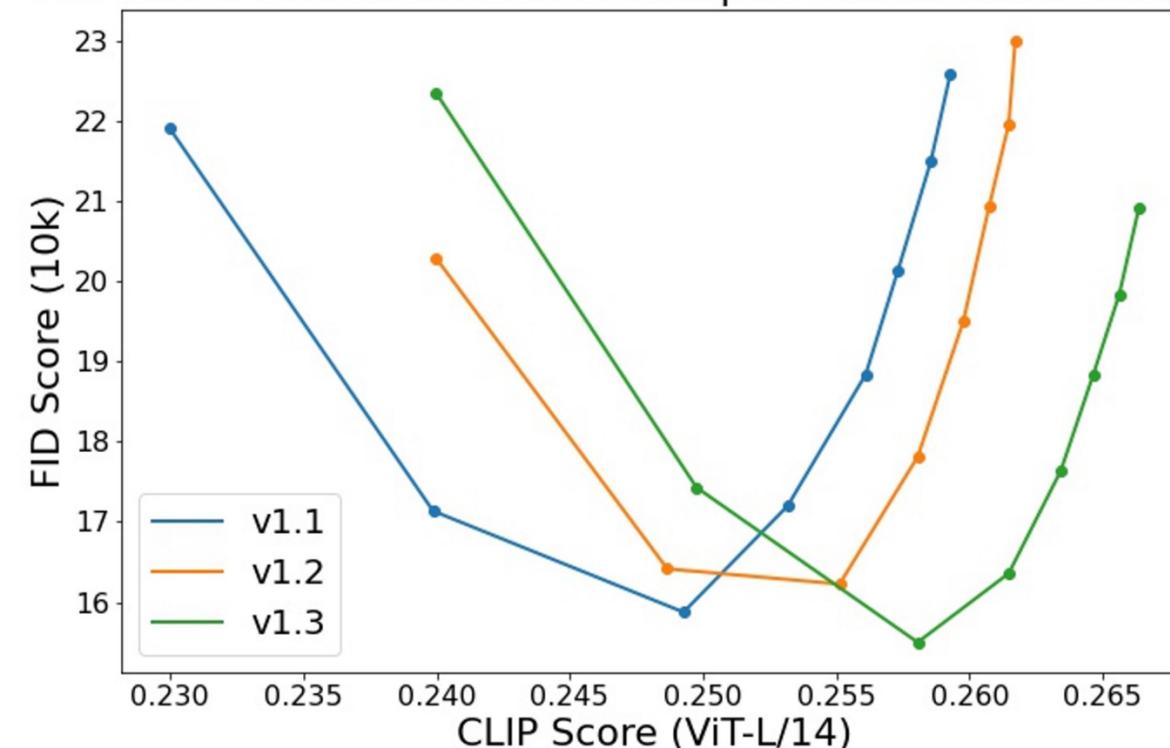
part 2 (v1.2):

- 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):

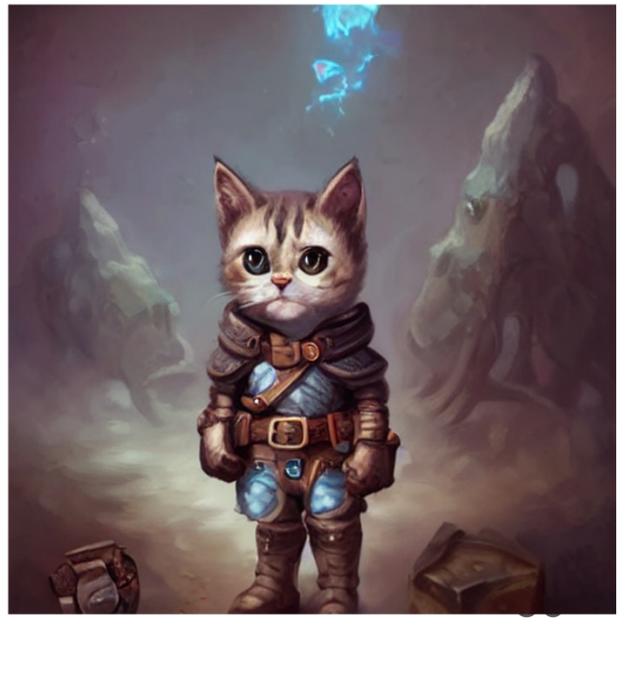
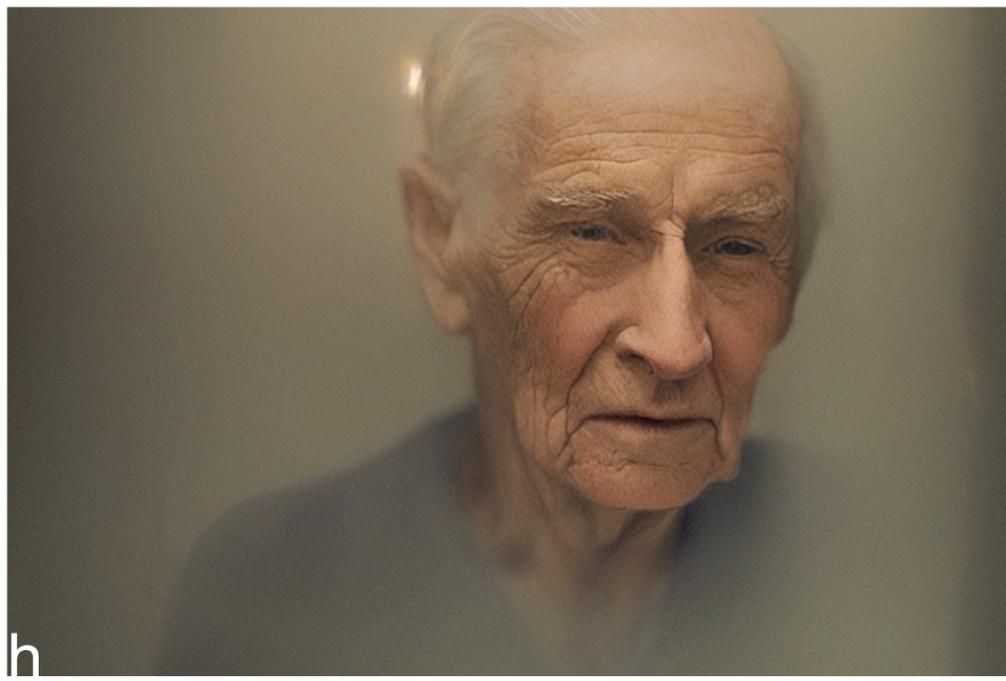
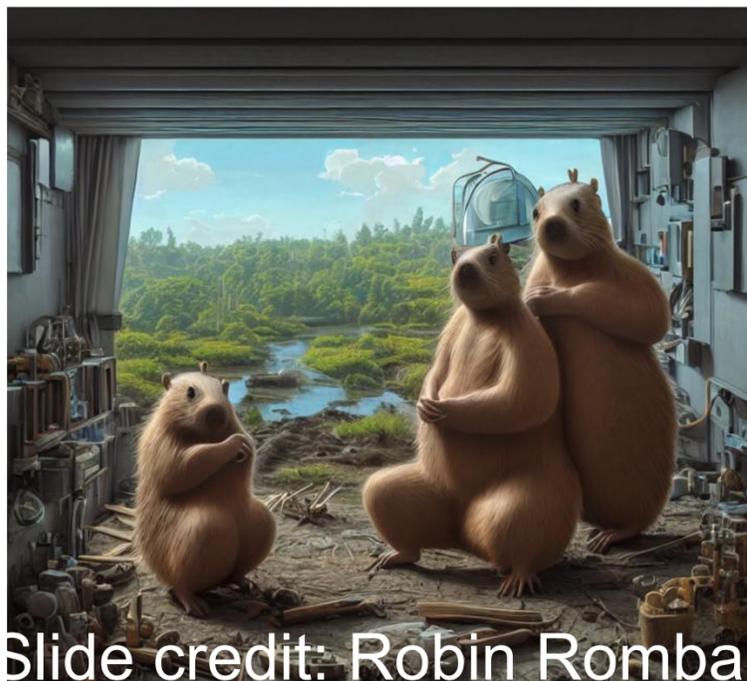
- 195k/225k steps at resolution 512x512 on "laion-improved-aesthetics" and 10% dropping of the text-conditioning

FID vs CLIP Scores on 512x512 samples for different v1-versions



10k random COCO val captions / 50 decoding steps

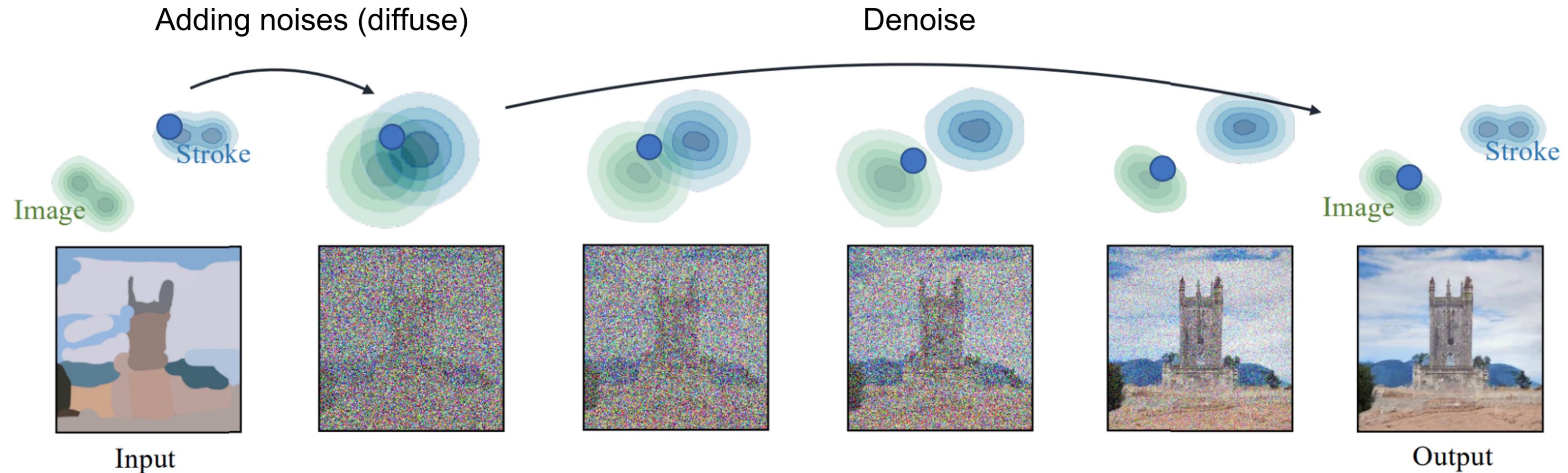
→ 4.2 GB checkpoint (EMA only, fp32)



Slide credit: Robin Rombach

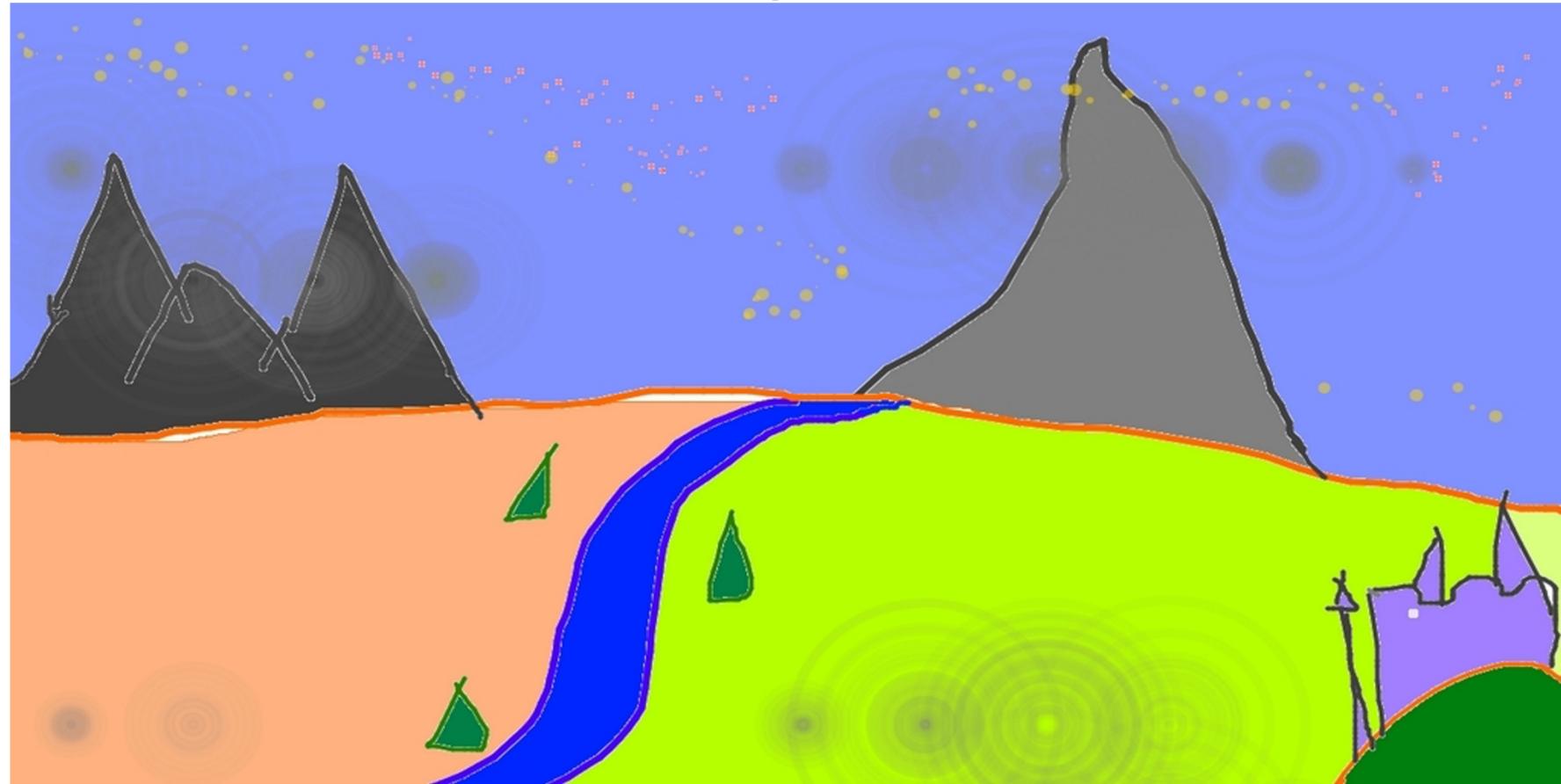
Text-Guided Image-to-Image

SDEdit (<https://arxiv.org/abs/2108.01073>) recipe: diffuse \rightarrow denoise



Text-Guided Image-to-Image

input



“a fantasy landscape, watercolor painting”



“a fantasy landscape, trending on artstation”

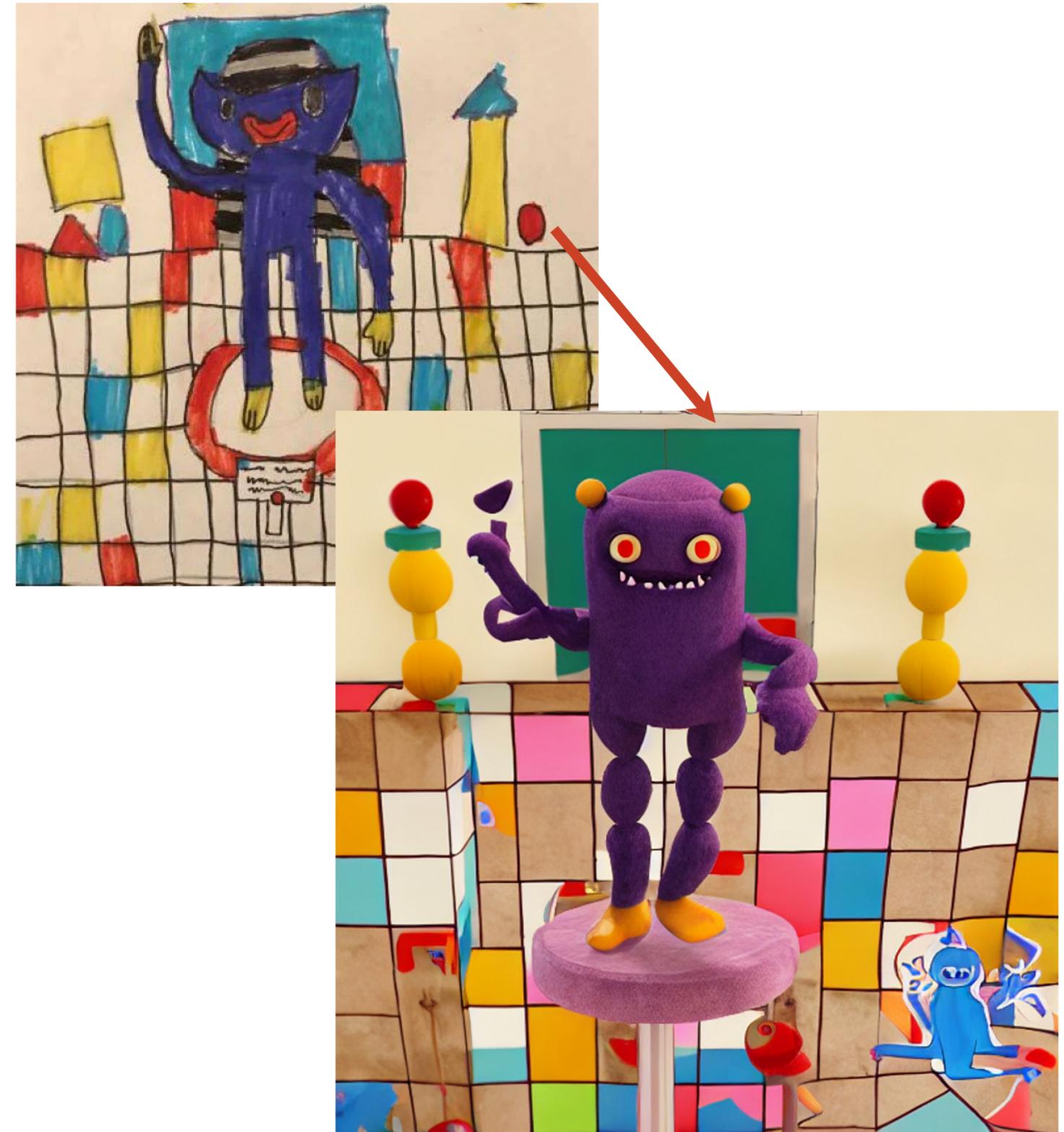


“a fantasy landscape, by Simon Stalenhag”

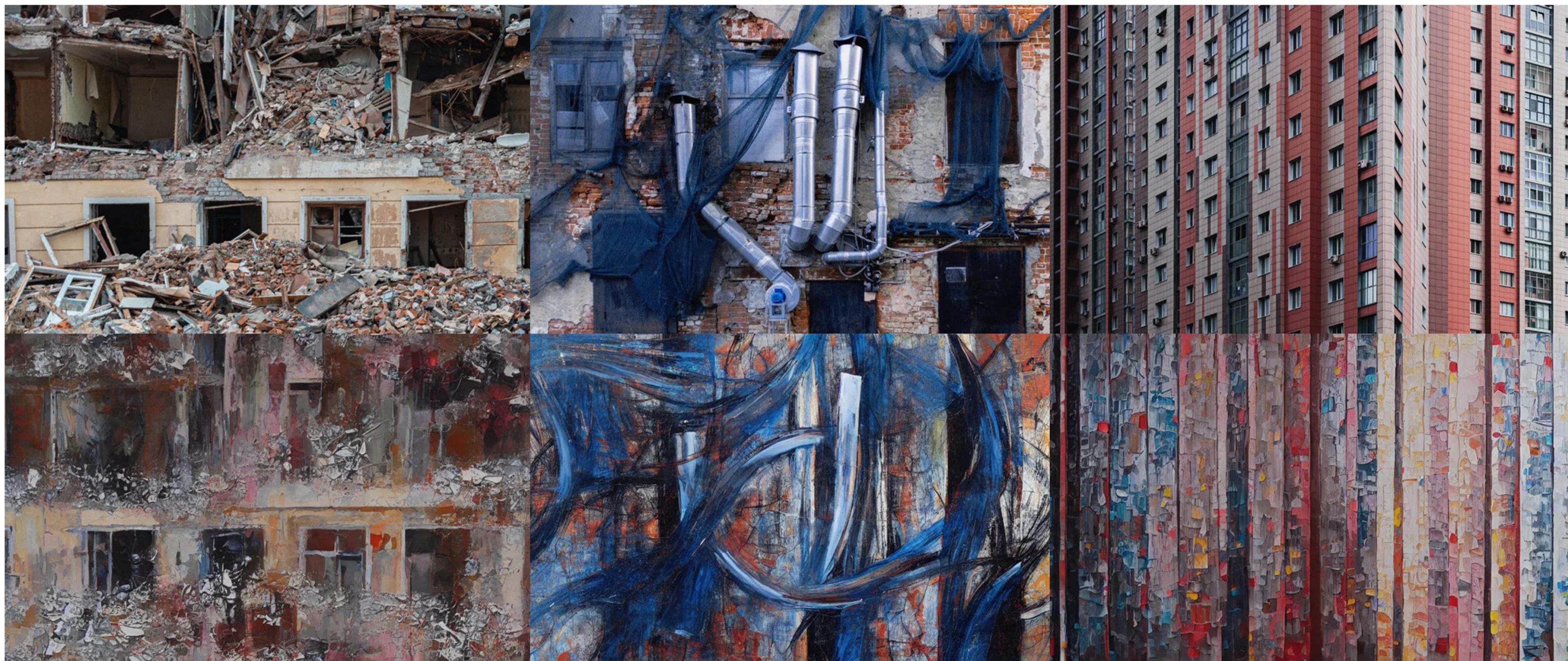


“Upgrade” your child’s artwork

original post: https://www.reddit.com/r/StableDiffusion/comments/wyq04v/using_img2img_to_upgrade_my_sons_artwork/



abstract art from photos



original post by [u/Pereulkov](https://www.reddit.com/user/Pereulkov)

https://www.reddit.com/r/StableDiffusion/comments/xhhyad/i_made_abstract_art_from_my_photos/

Video Synthesis

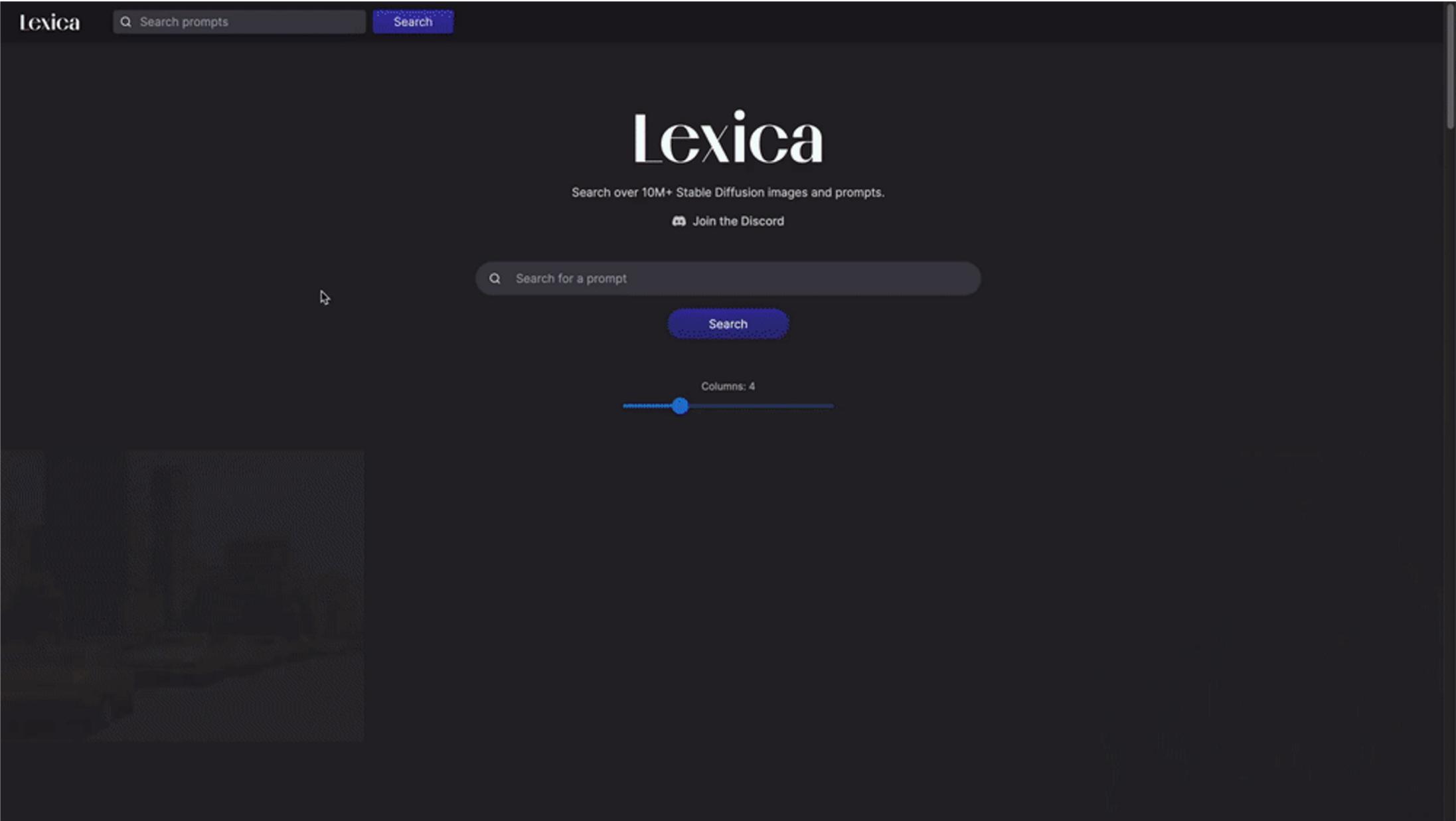


Stable Diffusion (img2img) + EBSynth by Scott Lightsier:

<https://twitter.com/LighthiserScott/status/1567355079228887041?t=kXXCAVtuO5IJCgro3Ma3A&s=19>

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

Prompt Search Engine (lexica.art)



Prompt Marketplace (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](#)

Featured in
TechCrunch THE VERGE WIRED FASTCOMPANY
FINANCIAL TIMES Atlantic yahoo!finance WSJ

Featured Prompts

- Midjourney: Vintage Retro Pattern Tiles \$1.99
- Midjourney: Minimal Pastel Diagram Art \$2.99
- Midjourney: Objects Made Of Money \$2.99
- Midjourney: Butterfly Cliparts \$2.99
- Midjourney: Asymmetrical Split Exposure ... \$2.99
- Midjourney: Stained Glass Letters \$2.99
- Midjourney: Coffee Stain Art \$2.99

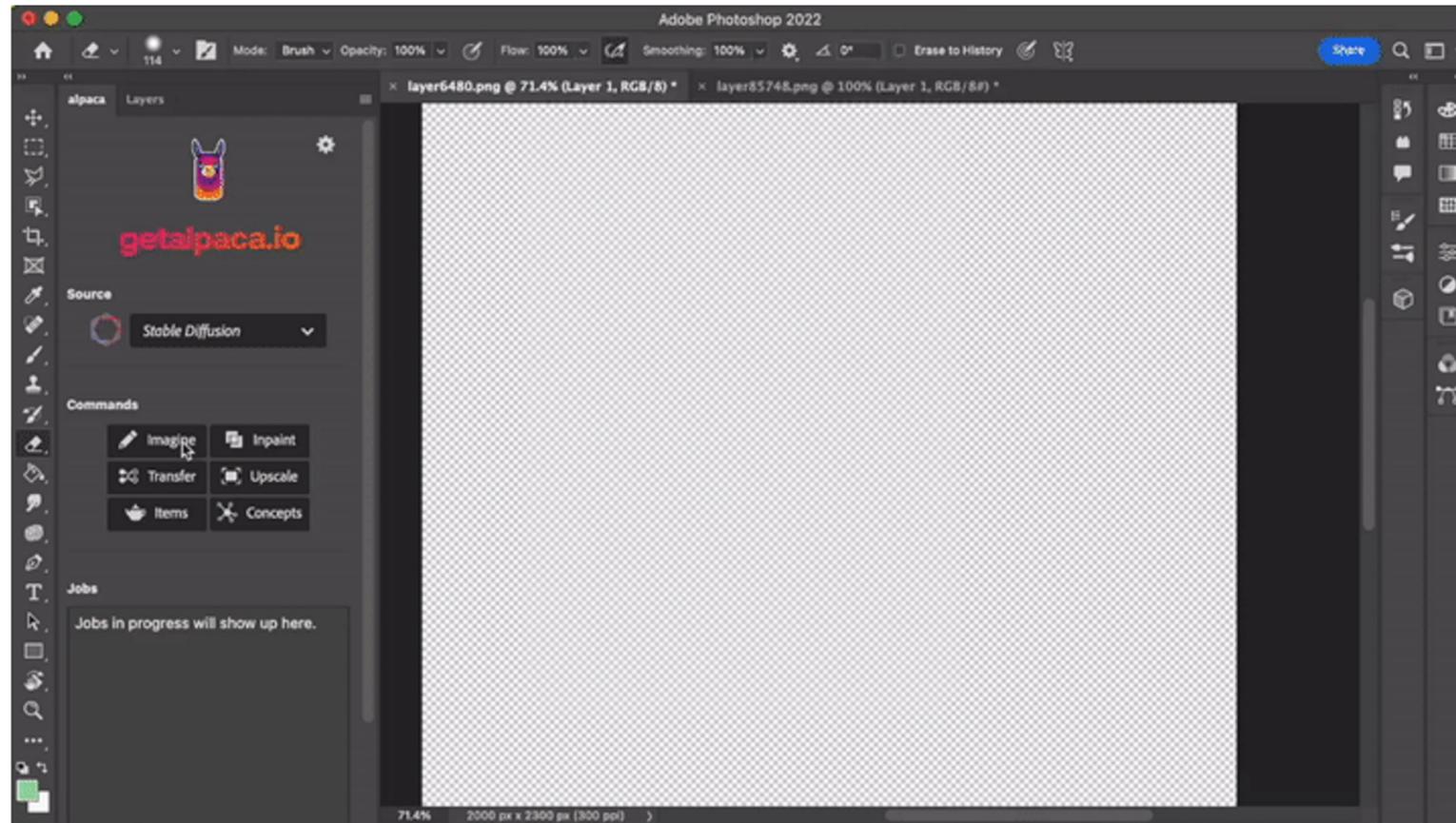
Hottest Prompts

- ChatGPT: Nft Generative Art Maker \$2.99
- Midjourney: Clean Animal Art For Coloring... \$1.99
- Midjourney: Tiny Gouache Houses \$2.99
- Midjourney: Beautiful Oil Paintings \$2.99
- ChatGPT: Hot Prl Selling \$2.99
- Midjourney: Make Cartoons Like Lofi-girl \$2.99
- Midjourney: Delicate Vibrant Emotive Arra... \$2.99

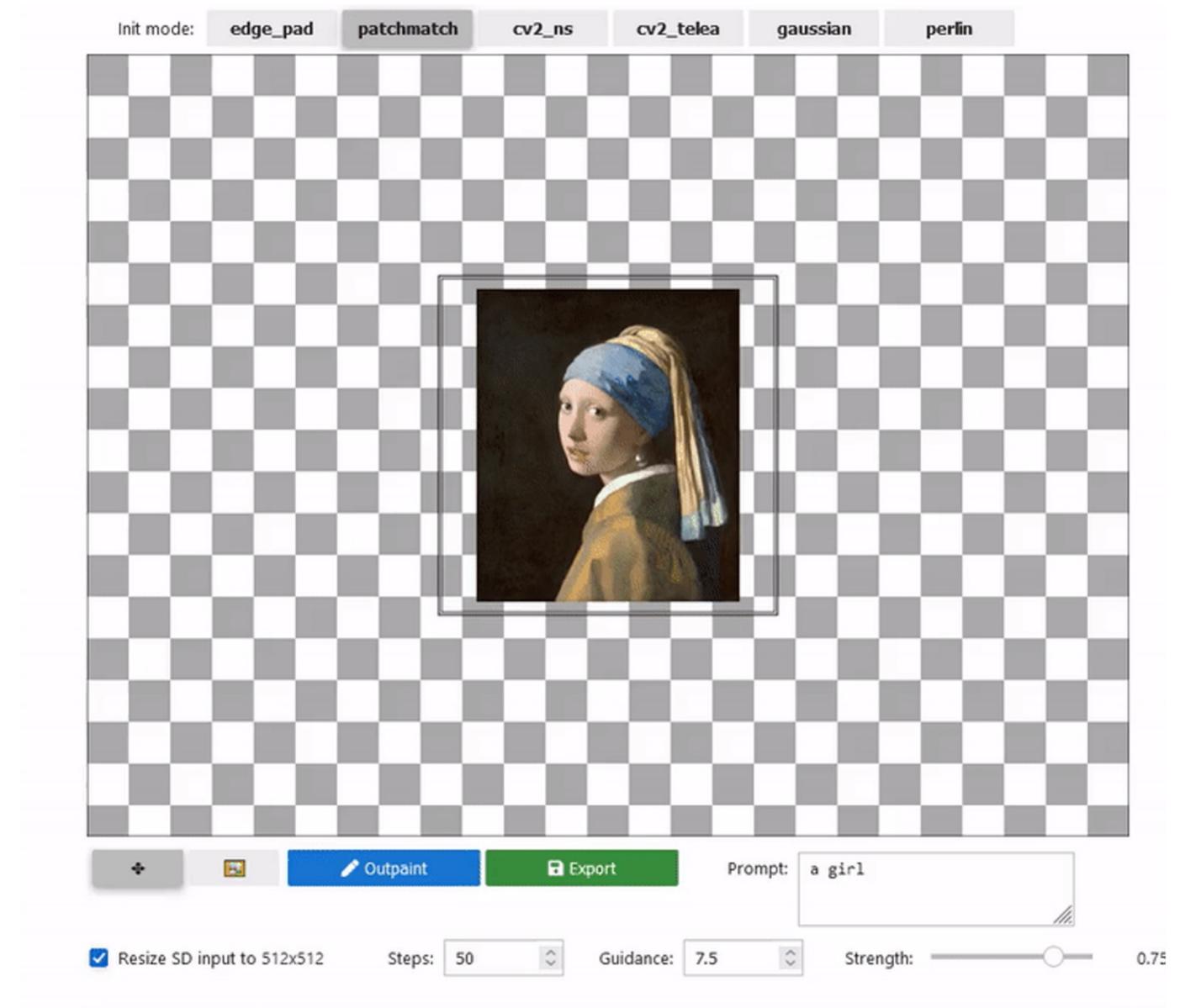
Newest Prompts

- ChatGPT: Fix Anything \$2.99
- Stable Diff.: Tropical Fashion \$2.99
- Midjourney: Food Images With Neon Effects \$1.99
- Midjourney: Wall Art Mockups Choose Wall ... \$1.99
- Stable Diff.: Premium Logos \$2.99
- Midjourney: Beautiful Oil Paintings \$2.99
- Midjourney: Alien Bio Organisms Posters \$2.99

UIs / Plug-Ins for Photoshop, GIMP etc



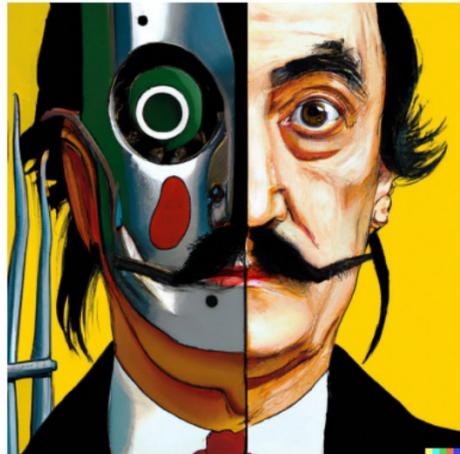
<https://twitter.com/wbuchw/status/1563162131024920576>



<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 handpalm with leaves growing from it



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.



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



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

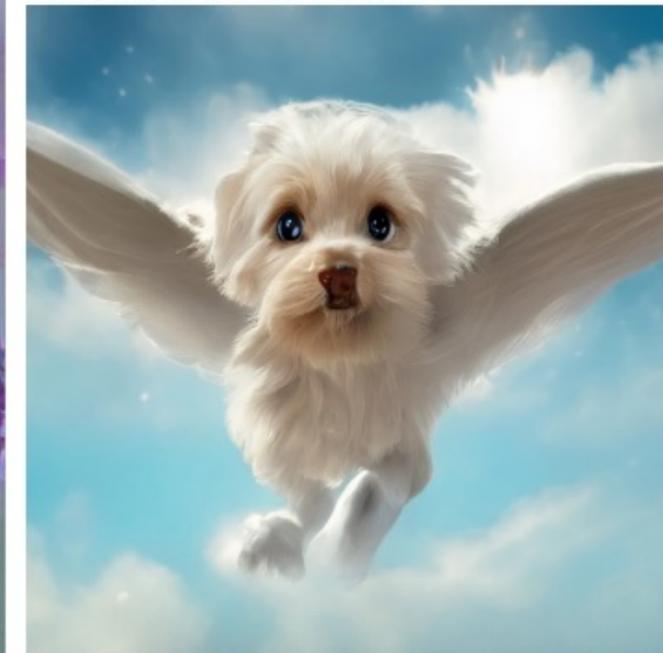
GigaGAN: Scaling up GANs



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

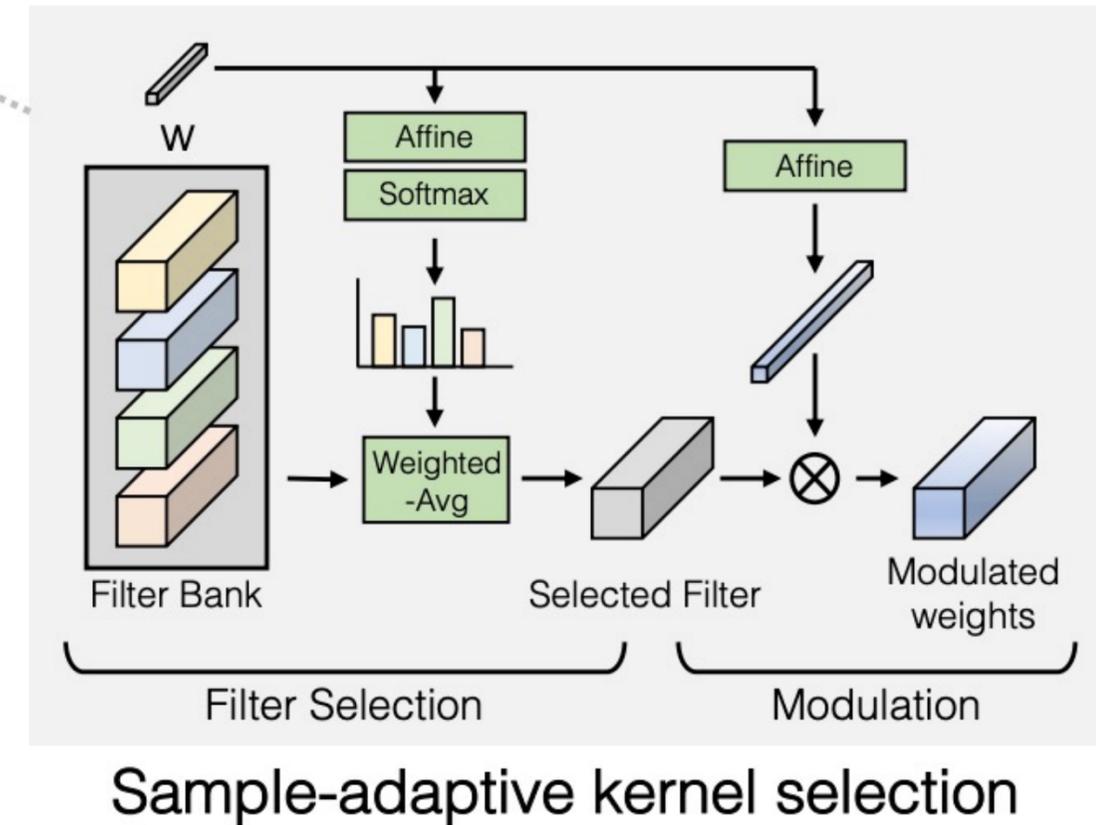
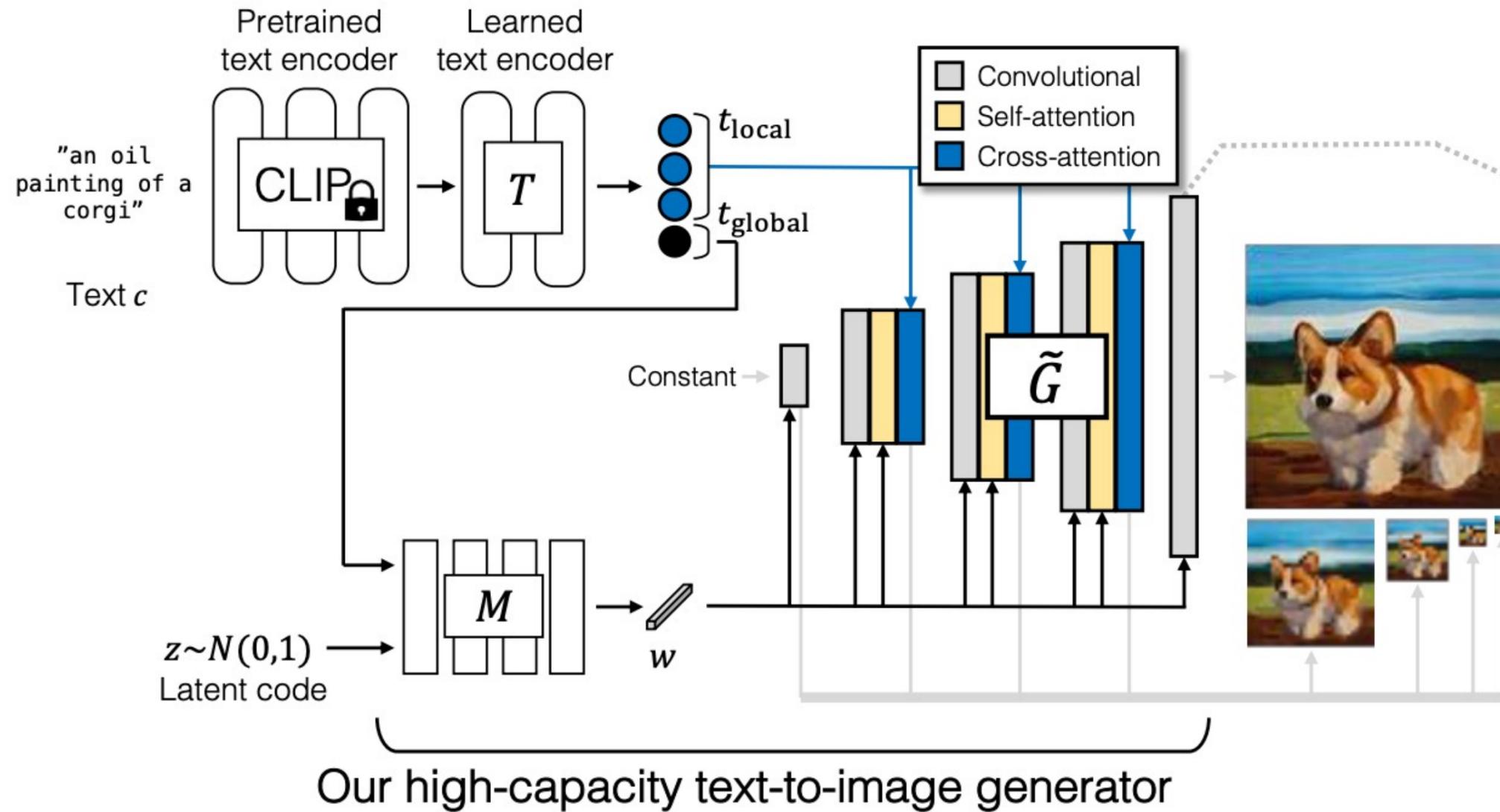


A golden luxury motorcycle parked at the King's palace. 35mm f/4.5.

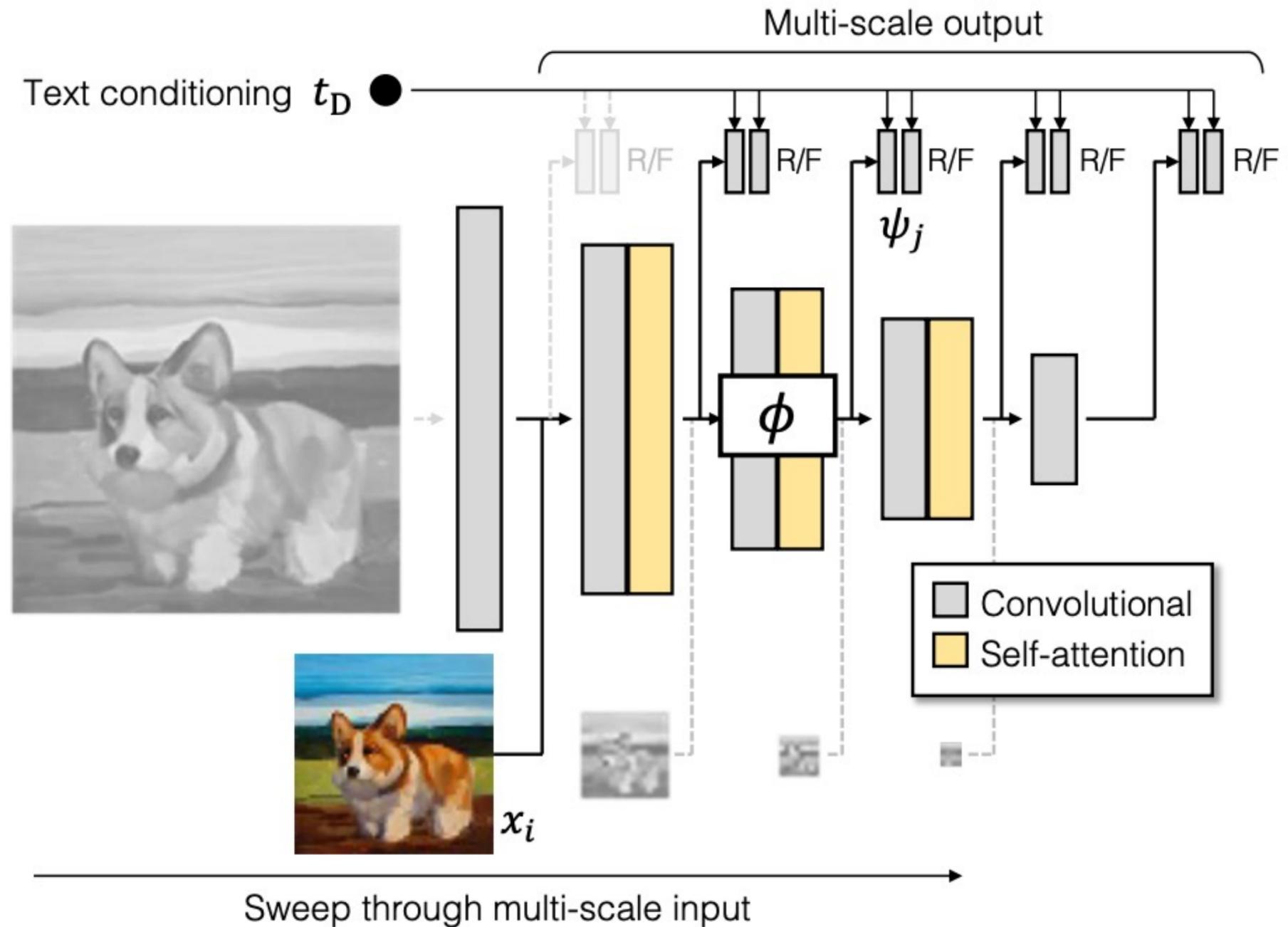


a cute magical flying maltipoo at light speed, fantasy concept art, bokeh, wide sky

GigaGAN Generator



GigaGAN Discriminator



Style Mixing

“A Toy sport sedan, CG art.”

Fine styles



Coarse styles



Prompt Mixing

no mixing

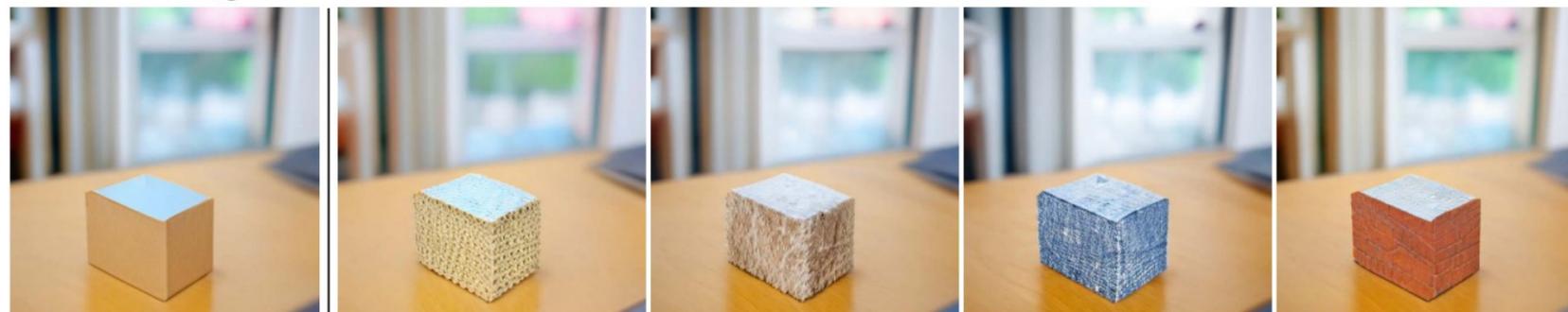
“crochet”

“fur”

“denim”

“brick”

“a cube on tabletop”



“a ball on tabletop”



“a teddy bear on tabletop”

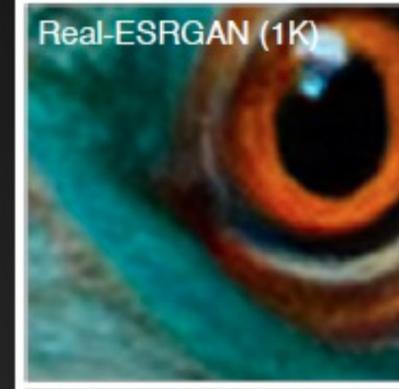


“a teddy bear on tabletop”





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



Comparison between Different Models

	Model	Type	# Param.	# Images	FID-30k ↓	Inf. time
256	DALL·E [75]	Diff	12.0B	1.54B	27.50	-
	GLIDE [63]	Diff	5.0B	5.94B	12.24	15.0s
	LDM [79]	Diff	1.5B	0.27B	12.63	9.4s
	DALL·E 2 [74]	Diff	5.5B	5.63B	10.39	-
	Imagen [80]	Diff	3.0B	15.36B	7.27	9.1s
	eDiff-I [5]	Diff	9.1B	11.47B	6.95	32.0s
	Parti-750M [101]	AR	750M	3.69B	10.71	-
	Parti-3B [101]	AR	3.0B	3.69B	8.10	6.4s
	Parti-20B [101]	AR	20.0B	3.69B	7.23	-
	LAFITE [108]	GAN	75M	-	26.94	0.02s
512	SD-v1.5* [78]	Diff	0.9B	3.16B	9.62	2.9s
	Muse-3B [10]	AR	3.0B	0.51B	7.88	1.3s
	GigaGAN	GAN	1.0B	0.98B	9.09	0.13s

Comparison between Different Models



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



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



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



DALL-E 2 (1024px)



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



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



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



DALL-E 2 (1024px)

StyleGAN-T



A painting of a fox in the style of starry night.

Beautiful landscape of an ocean. Mountain in the background. Sun is setting.



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



Surrealist dream-like oil painting by Salvador Dali of a cat playing checkers

Fall landscape with a small cottage next to a lake.



Panda mad scientist mixing sparkling chemicals, artstation

StyleGAN-T



A painting of a fox in the style of starry night.

Beautiful landscape of an ocean. Mountain in the background. Sun is setting.



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



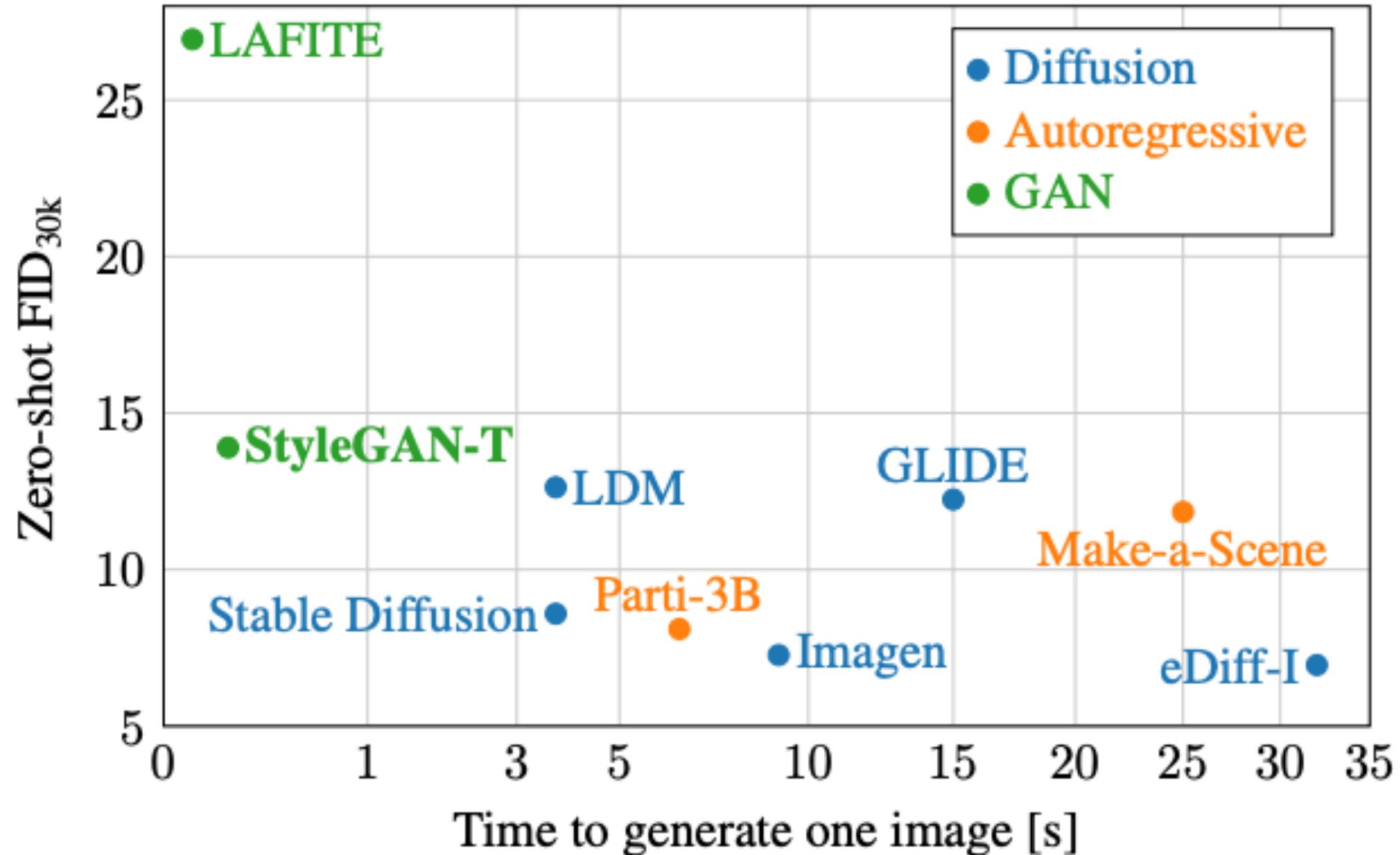
Surrealist dream-like oil painting by Salvador Dali of a cat playing checkers

Fall landscape with a small cottage next to a lake.



Panda mad scientist mixing sparkling chemicals, artstation

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.