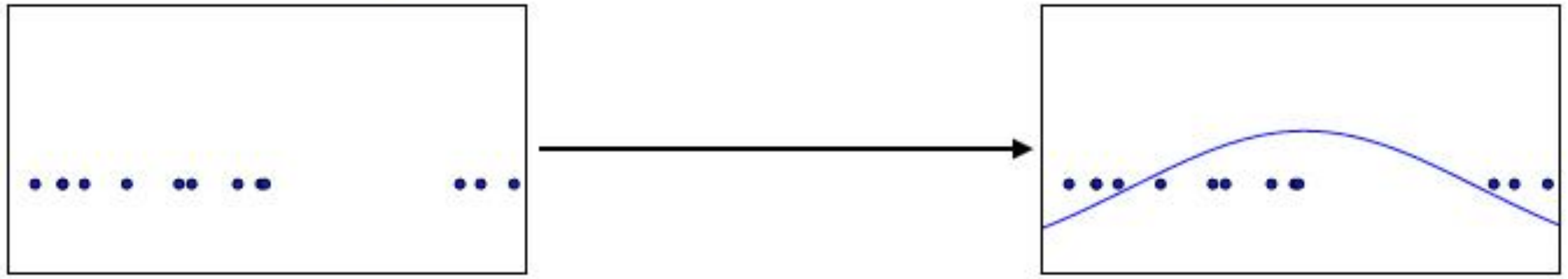
CMSC 491/691

Lecture 17: Image Synthesis

Some slides from Suren Jayasuriya, Jun-Yan Zhu

Generative Modeling

- Density estimation



- Sample generation

Training examples   Model samples

(Goodfellow 2016)

Generative Adversarial Networks (GAN)

Why Generative Models?

- **We've only seen discriminative models so far**
  - Given an image **X**, predict a label **Y**
  - Estimates **P(Y|X)**

- **Discriminative models have several key limitations**
  - Can't model **P(X)**, i.e. the probability of seeing a certain image
  - Thus, can't sample from **P(X)**, i.e. **can't generate new images**

- **Generative models (in general) cope with all of above**
  - Can model **P(X)**
  - Can generate new images

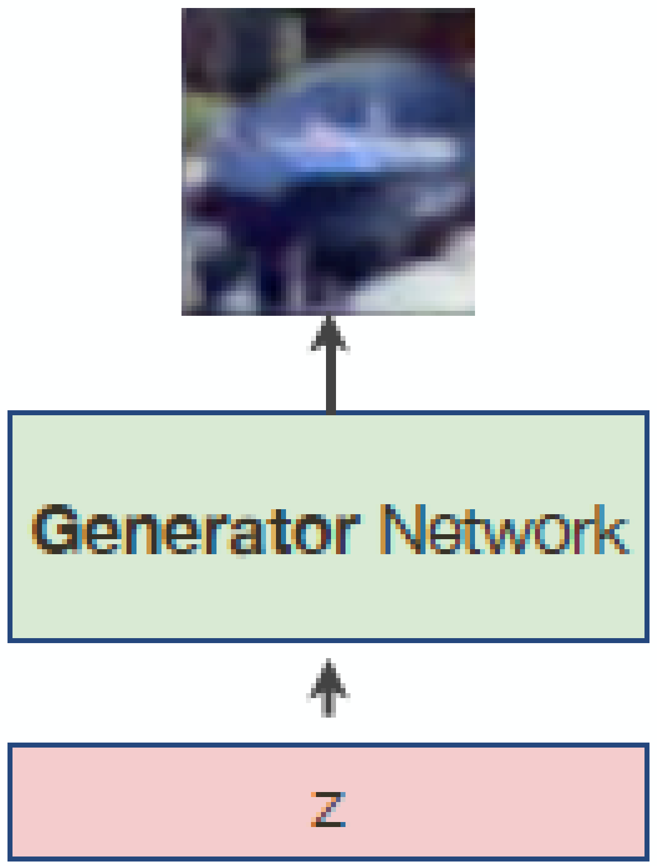
Generative Adversarial Networks

[Goodfellow et al., 2014]

Problem: Want to sample from complex, high-dimensional training distribution. There is no direct way to do this!

Solution: Sample from a simple distributions, e.g., random noise. Learn transformation to the training distribution

Question: What can we use to represent complex transformation function?



* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, **cs231n Stanford**

Image synthesis from "noise"

G

$z \sim p(z)$

→ Generator →

$x = G(z)$

Sampler

$$G : \mathcal{Z} \to \mathcal{X}$$

$$z \sim p(z)$$

$$x = G(z)$$

Image synthesis from “noise”

G

$z \sim p(z)$

→ Generator →

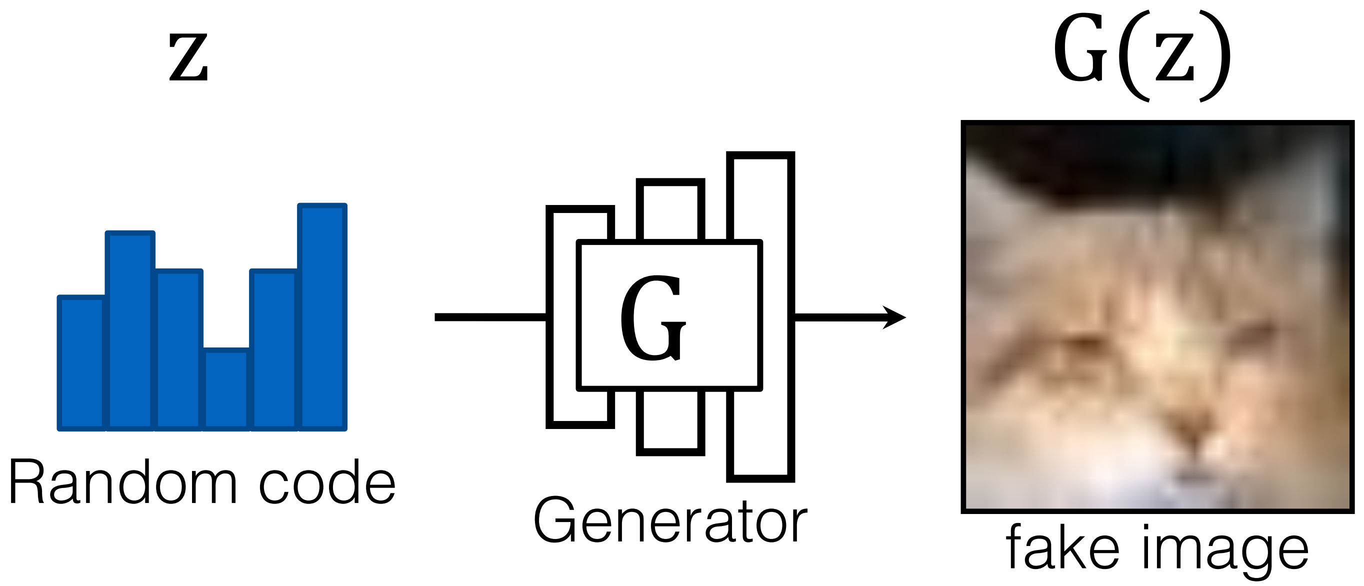
$x = G(z)$

Sampler

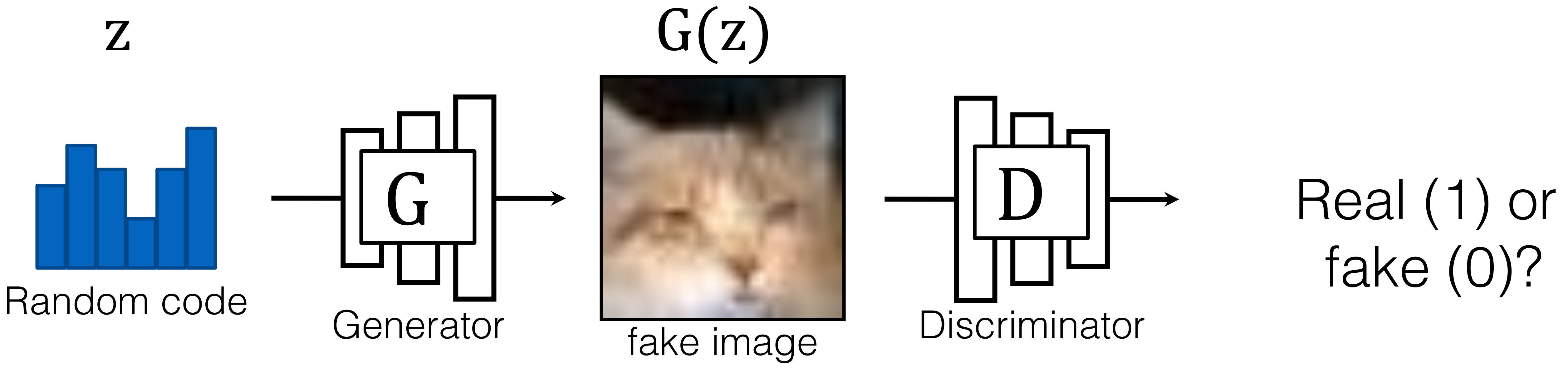
$$G : \mathcal{Z} \to \mathcal{X}$$

$$z \sim p(z)$$

$$x = G(z)$$



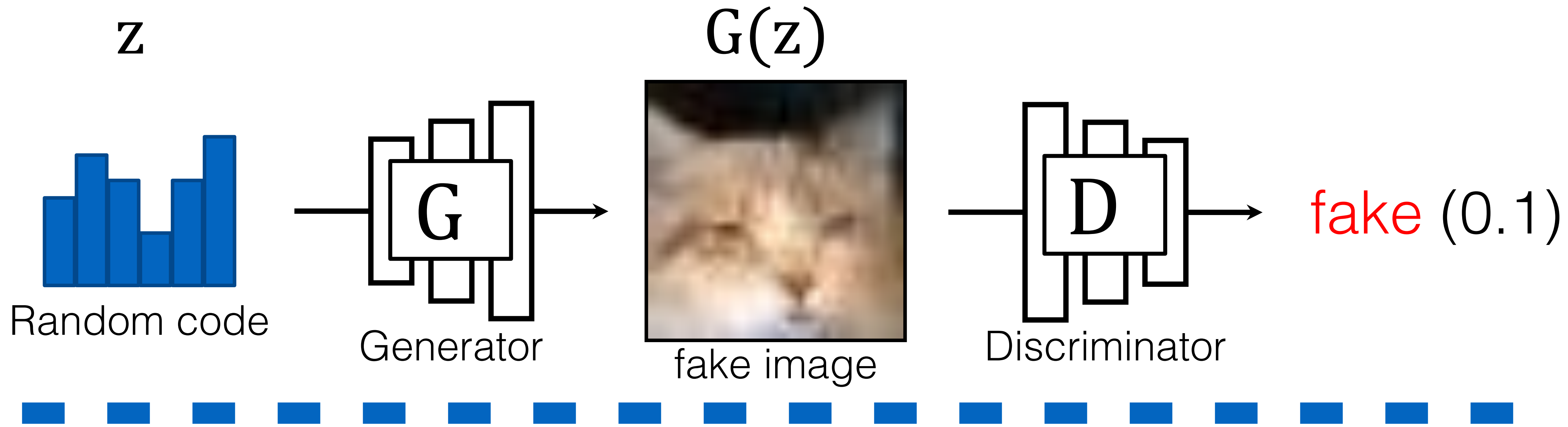
[Goodfellow et al. 2014]



A two-player game:

- G tries to generate fake images that can fool D.
- D tries to detect fake images.

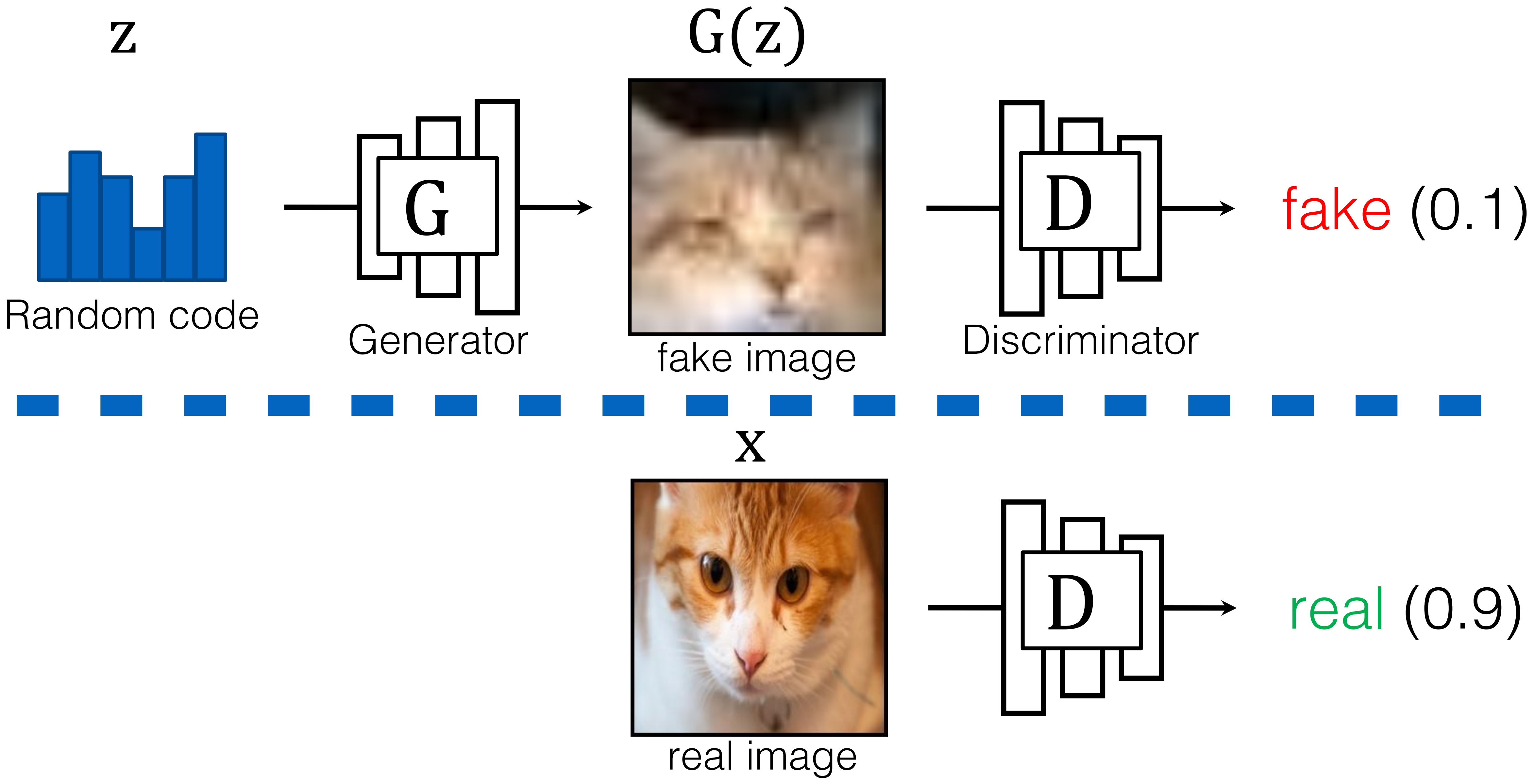
[Goodfellow et al. 2014]



Learning objective (GANs)

$$\min_G \boxed{\max_D} \boxed{\mathbb{E}_z[\log(1 - D(G(z)))]}$$

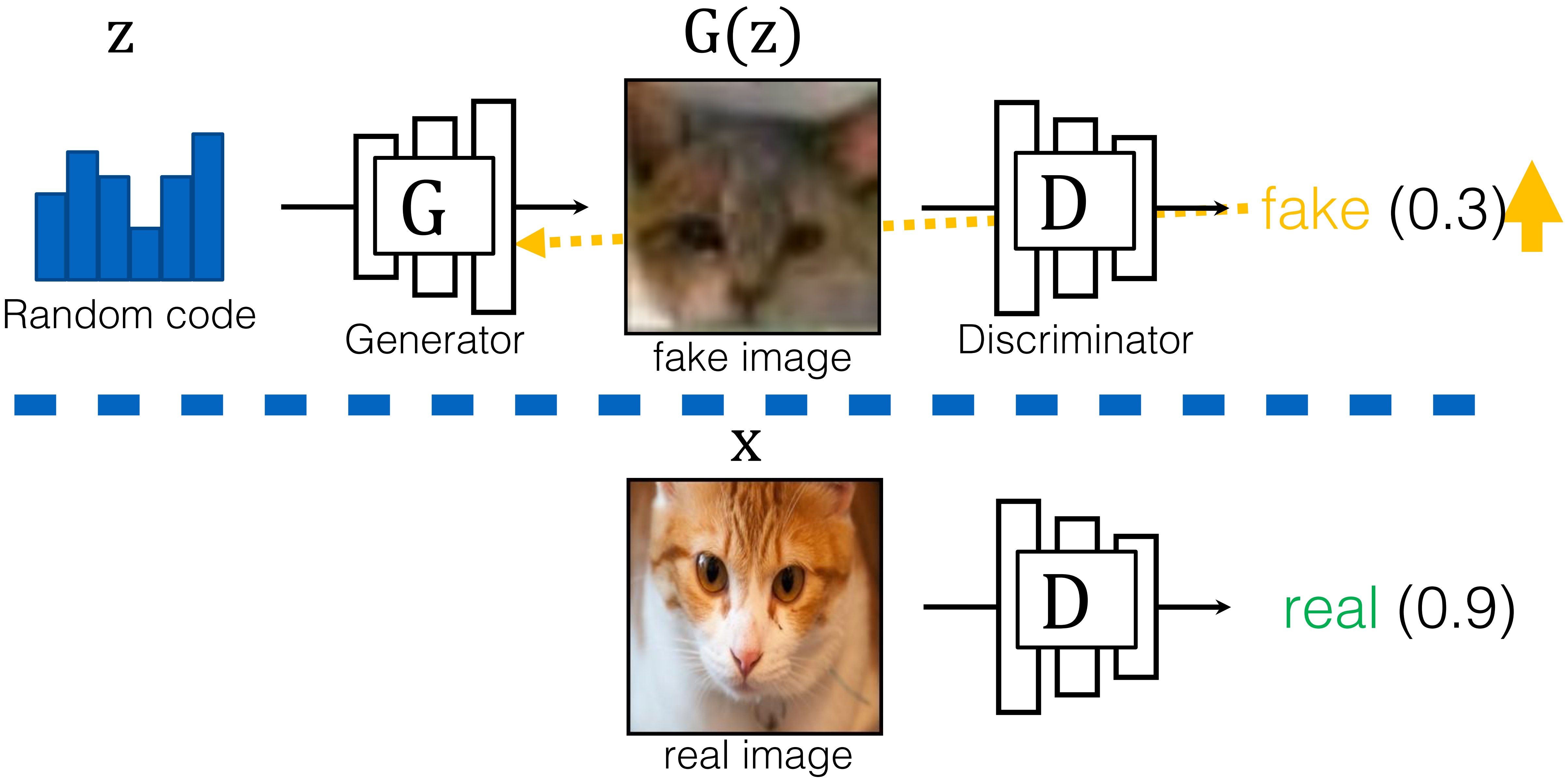
[Goodfellow et al. 2014]



Learning objective (GANs)

$$\min_{G} \max_{\boxed{D}} \boxed{\mathbb{E}_z[\log(1-D(G(z)))]} + \boxed{\mathbb{E}_x[\log D(x)]}$$

[Goodfellow et al. 2014]



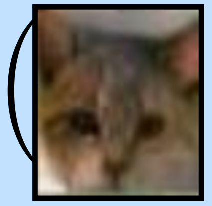
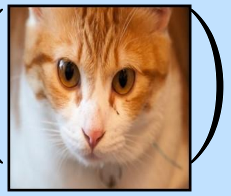
Learning objective (GANs)

$$\min_{\boxed{G}} \max_{D} \boxed{\mathbb{E}_z[\log(1-D(G(z)))]} + \mathbb{E}_x[\log D(x)]$$

[Goodfellow et al. 2014]

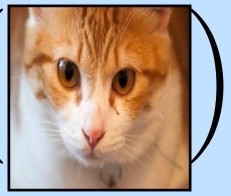
GAN Training Breakdown

- From the discriminator D's perspective:
  - binary classification: real vs. fake.
  - Nothing special: similar to 1 vs. 7 or cat vs. dog

$\max_D \mathbb{E}[\log(1-D($  $))]+\mathbb{E}[\log D($  $)]$

GAN Training Breakdown

- From the discriminator D's perspective:
  - binary classification: real vs. fake.
  - Nothing special: similar to 1 vs. 7 or cat vs. dog

$\max_D \mathbb{E}[\log(1 - D($  $))] + \mathbb{E}[\log D($  $)]$

- From the generator G's perspective:
  - Optimizing a loss that depends on a classifier D
  - We have done it before (Perceptual Loss)

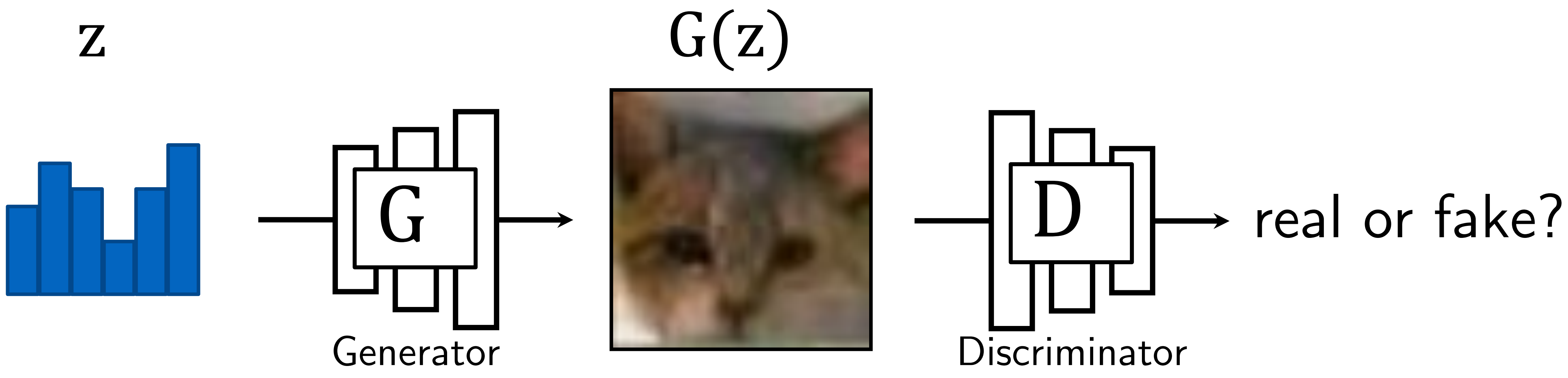
$$\min_G \mathbb{E}_z[\mathcal{L}_D(G(z))]$$

GAN loss for G

$$\min_G \mathbb{E}_{(x,y)}||F(G(x)) - F(y)||$$

Perceptual Loss for G

GAN Training Breakdown



G tries to synthesize fake images that fool D

D tries to identify the fakes

- Training: iterate between training D and G with backprop.
- Global optimum when G reproduces data distribution.

[Goodfellow et al., 2014]

Generative Adversarial Network

Data $\rightarrow$

Learner

Objective

$$\min_G \max_D \mathbb{E}_z[\log(1-D(G(z))]+\mathbb{E}_x[\log D(x)]$$

Hypothesis space

Deep nets G and D

Optimizer

Alternating SGD on G and D

$\rightarrow$

Critic

$$D : \mathcal{X} \rightarrow [0,1]$$

Sampler

$$G : \mathcal{Z} \rightarrow \mathcal{X}$$

Generative Adversarial Network

Data $\rightarrow$

Learner

Objective

$JSD\,(p_{\text{data}}\,\|\,p_g\,)$

Hypothesis space

Deep net G

Optimizer

Adversarial game

$\longrightarrow$ Sampler

$G : \mathcal{Z} \rightarrow \mathcal{X}$

Generative Adversarial Nets

Generated Samples

Samples from **StyleGAN2 [Karras et al., CVPR 2020]**

Interpolation is impressive

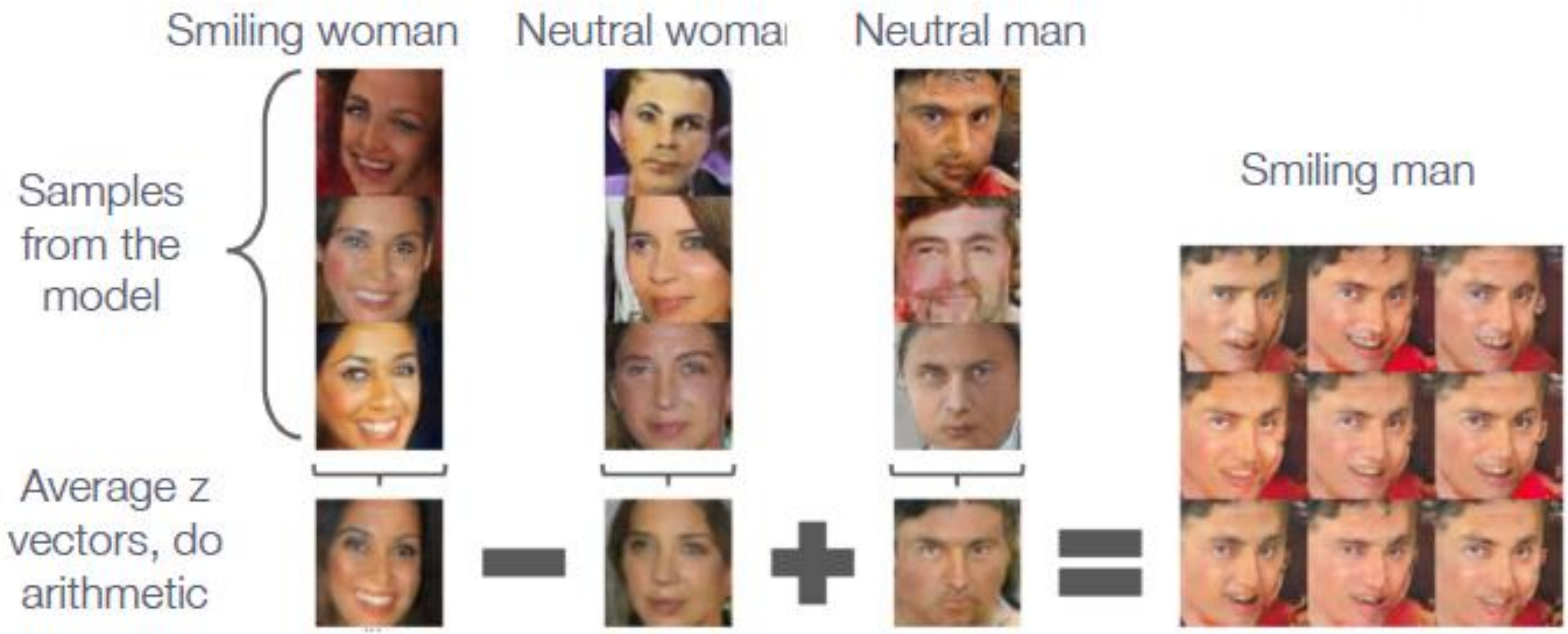
(c) Our results (128x128 with 128 filters)

(d) Mirror interpolations (our results 128x128 with 128 filters)

https://arxiv.org/pdf/1703.10717.pdf

GANs: Interpretable Vector Math

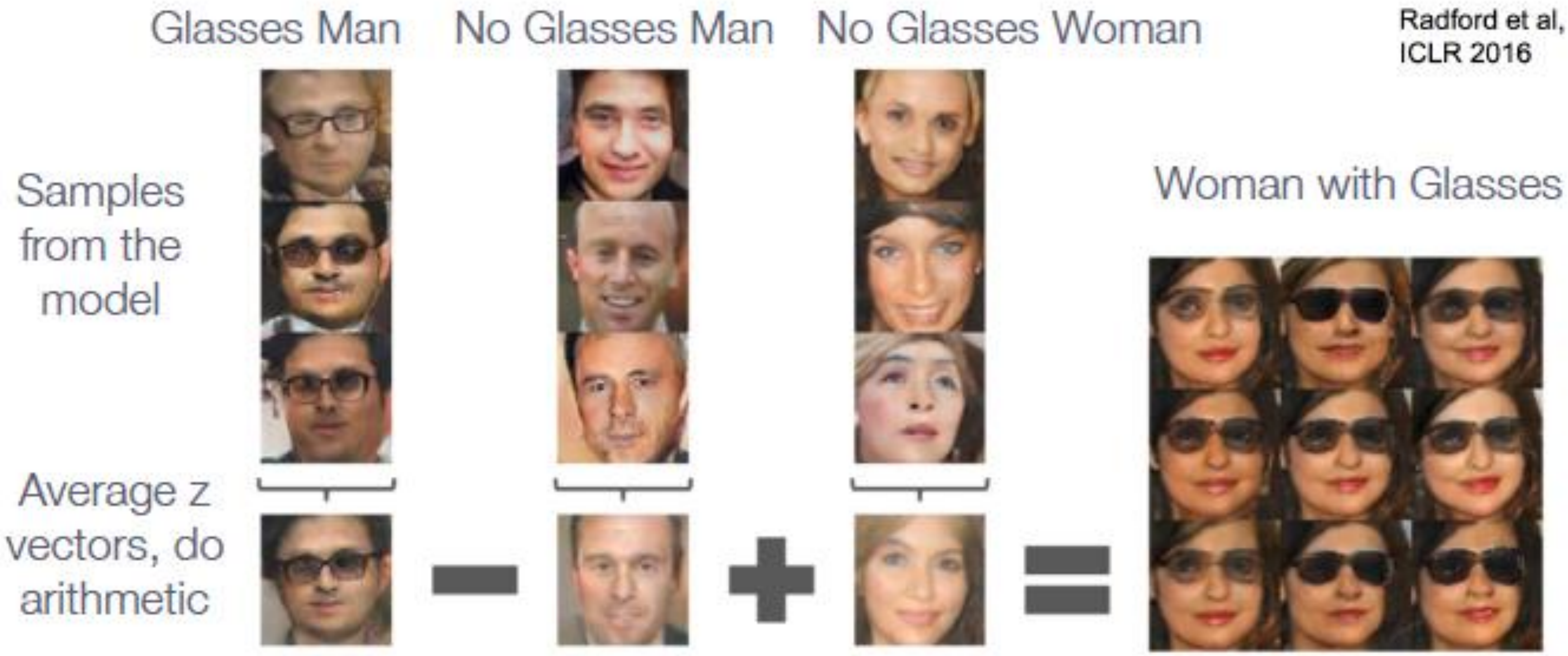
[Radford et al., 2016]



* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, **cs231n Stanford**

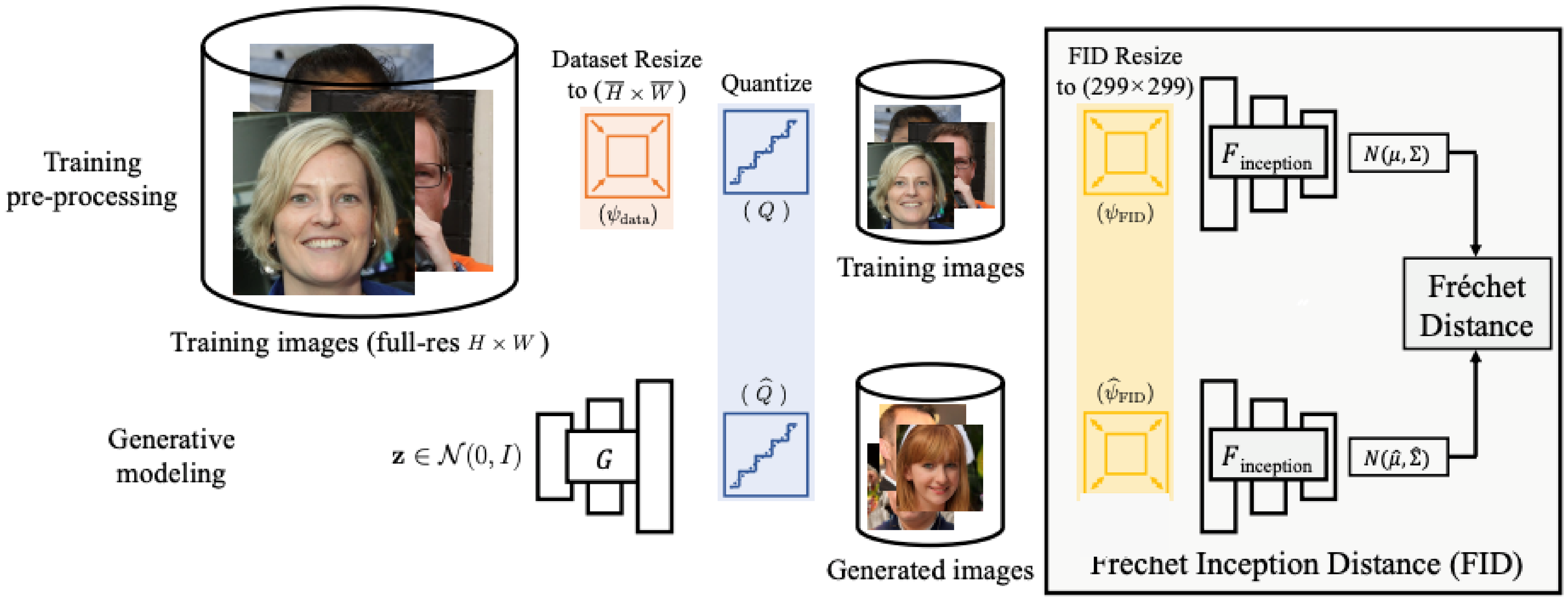
GANs: Interpretable Vector Math

[Radford et al., 2016]



* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, **cs231n Stanford**

GANs evaluation (FID)



Fréchet Inception Distance (FID)

$$\mathbf{FID} = ||\mu - \hat{\mu}||_2^2 + \mathbf{Tr}(\Sigma + \widehat{\Sigma} - 2(\Sigma\widehat{\Sigma})^{1/2})$$

GANs evaluation (FID)

Clean-fid libraries for evaluating generative models

pip install clean-fid
Daily downloads (July, 2022): 100
Daily downloads (Feb, 2023) : 20, 000
Total downloads: 2, 600, 000

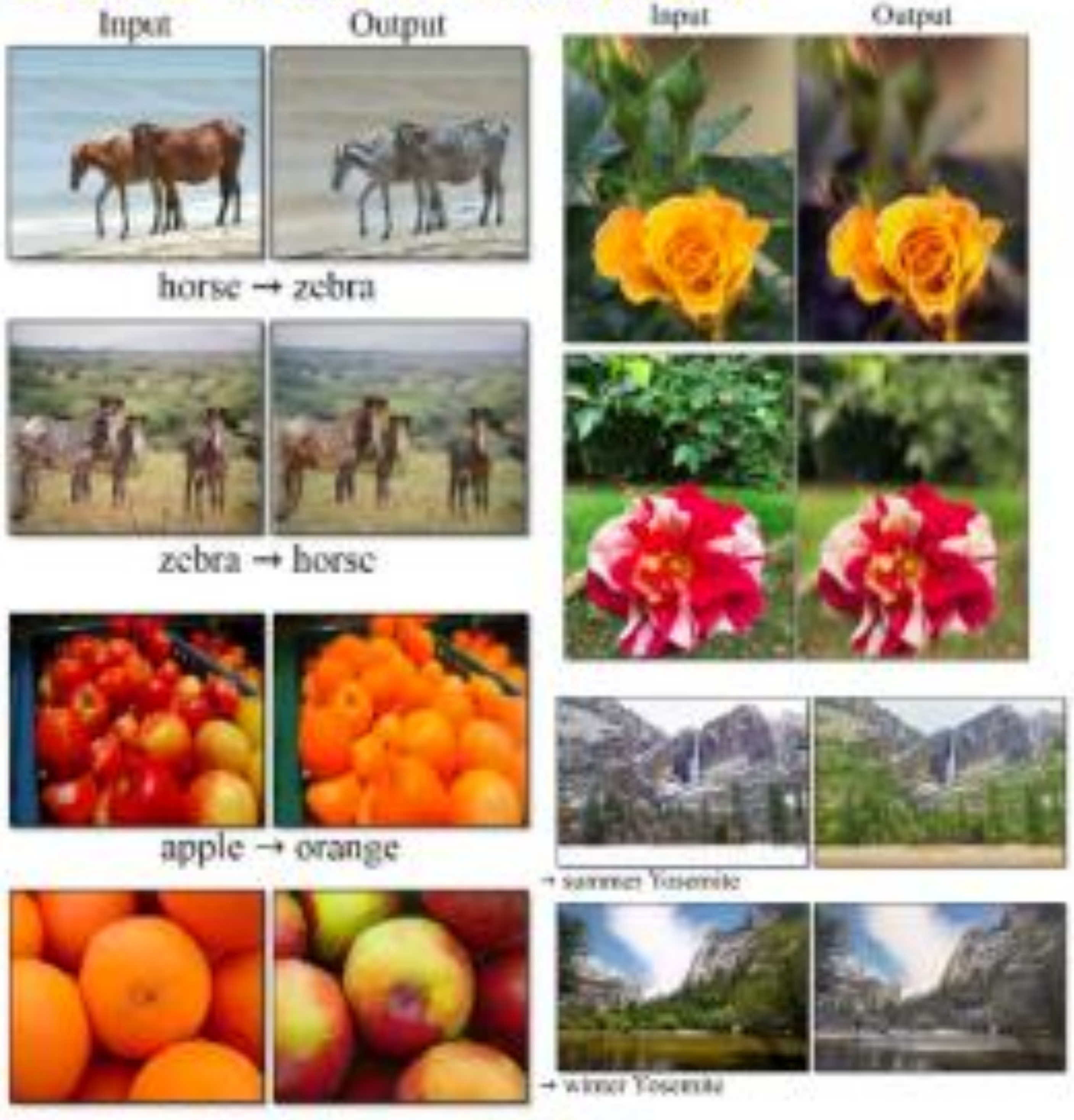
[Parmar et al., CVPR 2022]

Better training and generation

LSGAN. Mao et al. 2017.

BEGAN. Bertholet et al. 2017.

Source->Target domain transfer



CycleGAN. Zhu et al. 2017.

Text -> Image Synthesis

Reed et al. 2017.

Many GAN applications

Pix2pix. Isola 2017. Many examples at https://phillipi.github.io/pix2pix/

Image-to-Image Translation

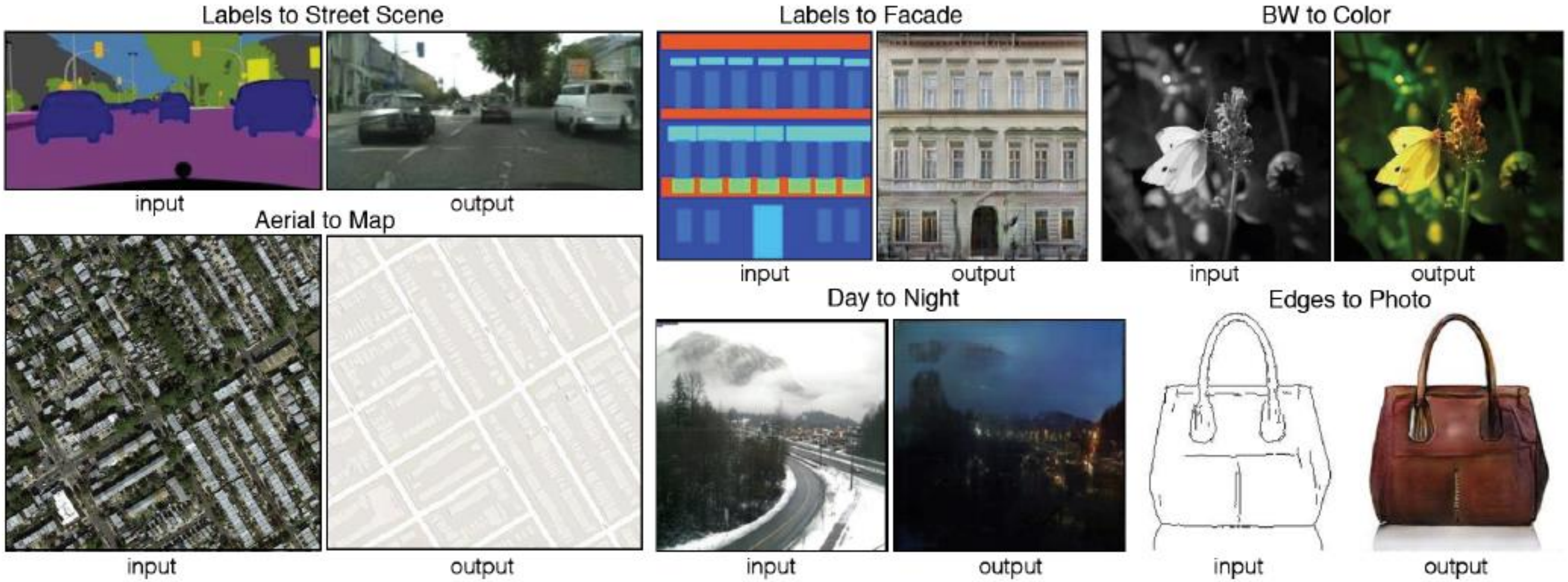


Figure 1 in the original paper.

[Link to an interactive demo of this paper]

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. "**Image-to-image translation with conditional adversarial networks**". arXiv preprint arXiv:1611.07004. (2016).

Image-to-Image Translation

- Architecture: *DCGAN*-based architecture

- Training is conditioned on the images from the source domain.

- Conditional GANs provide an effective way to handle many complex domains without worrying about designing *structured loss* functions explicitly.

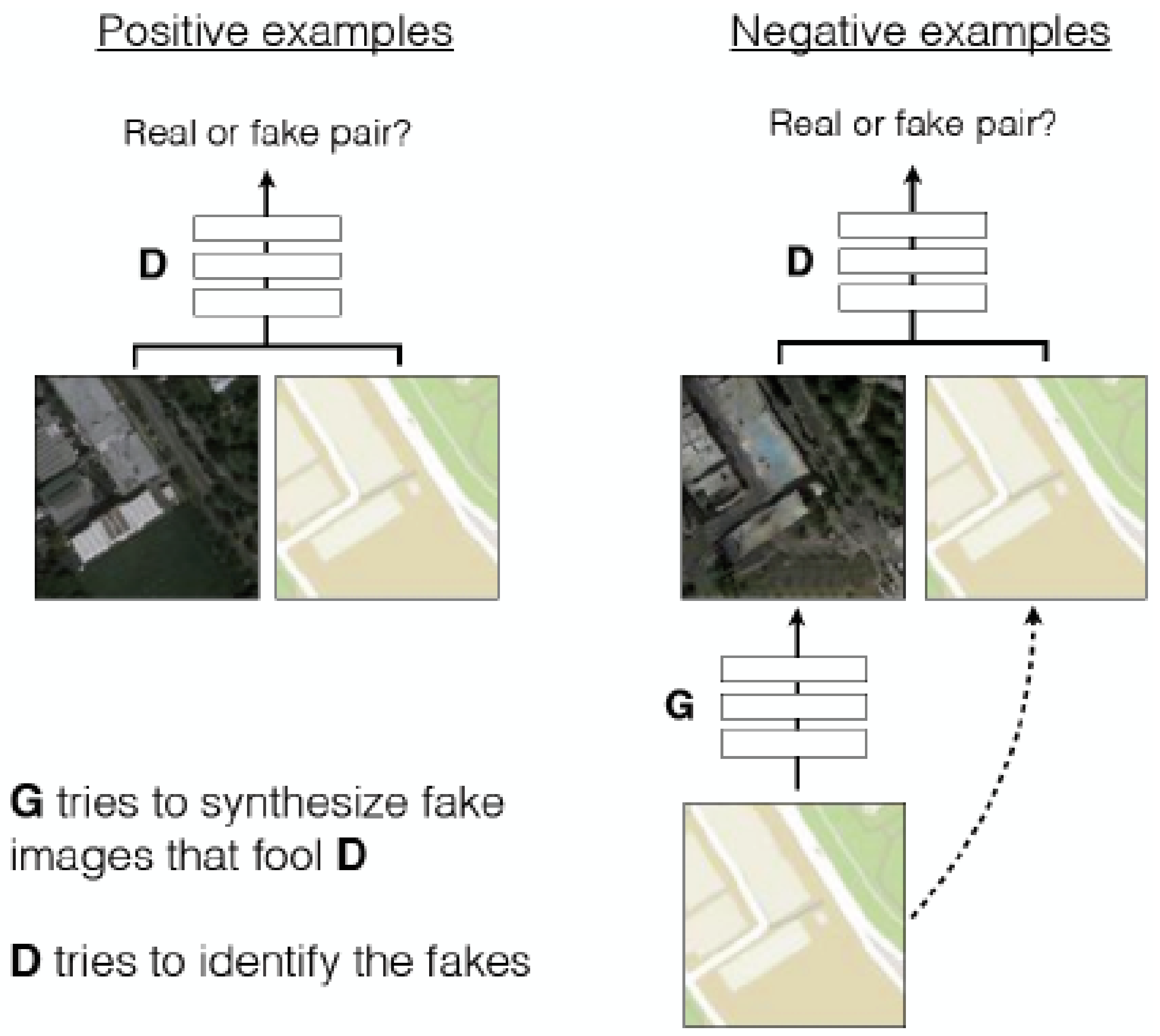


Figure 2 in the original paper.

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. "**Image-to-image translation with conditional adversarial networks**". arXiv preprint arXiv:1611.07004. (2016).

Problems with GANs

- **Probability Distribution is Implicit**
  - Not straightforward to compute P(X).
  - Thus **Vanilla GANs** are only good for Sampling/Generation.

- **Training is Hard**
  - Non-Convergence
  - Mode-Collapse

Goodfellow, Ian. **"NIPS 2016 Tutorial: Generative Adversarial Networks."** arXiv preprint arXiv:1701.00160 (2016).

- **Deep Learning models (in general) involve a single player**
  - The player tries to maximize its reward (minimize its loss).
  - Use SGD (with Backpropagation) to find the optimal parameters.
  - SGD has convergence guarantees (under certain conditions).
  - **Problem:** With non-convexity, we might converge to local optima.

$$\min_{G} L(G)$$

- **GANs instead involve two (or more) players**
  - Discriminator is trying to maximize its reward.
  - Generator is trying to minimize Discriminator's reward.

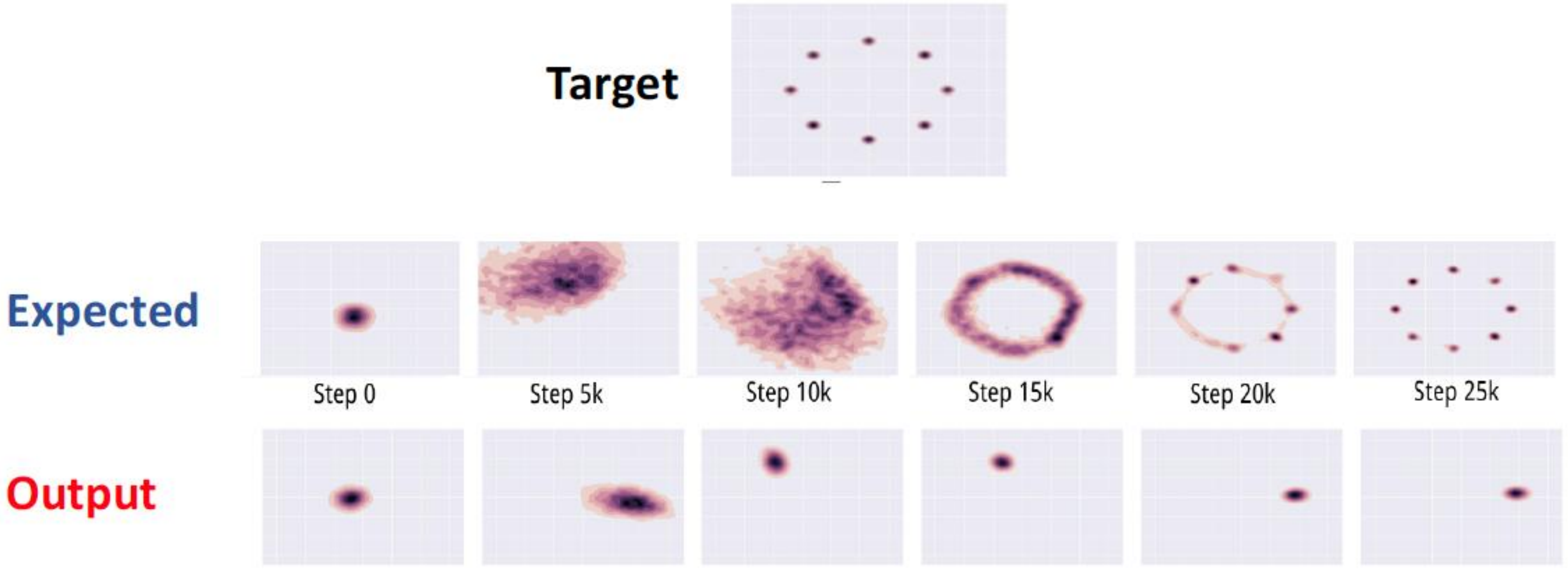
$$\min_{G} \max_{D} V(D, G)$$

  - SGD was not designed to find the Nash equilibrium of a game.
  - **Problem:** We might not converge to the Nash equilibrium at all.

Salimans, Tim, et al. "Improved techniques for training gans." *Advances in Neural Information Processing Systems*. 2016.

Mode-Collapse

- **Generator fails to output diverse samples**



Metz, Luke, et al. **"Unrolled Generative Adversarial Networks."** arXiv preprint arXiv:1611.02163 (2016).

Text-to-Image (T2I)

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

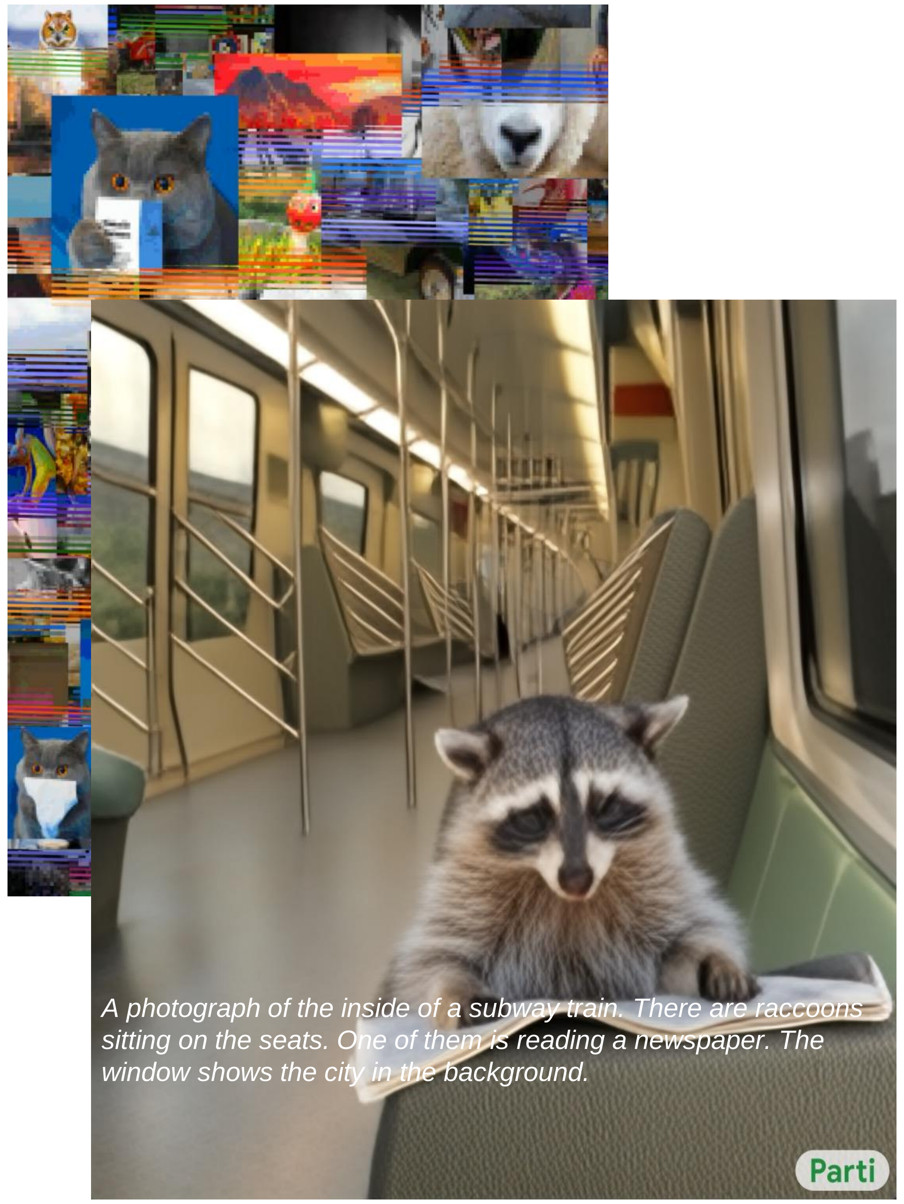
a blue Porsche 356 parked in front of a yellow brick wall.

Eiffel Tower, landscape photography

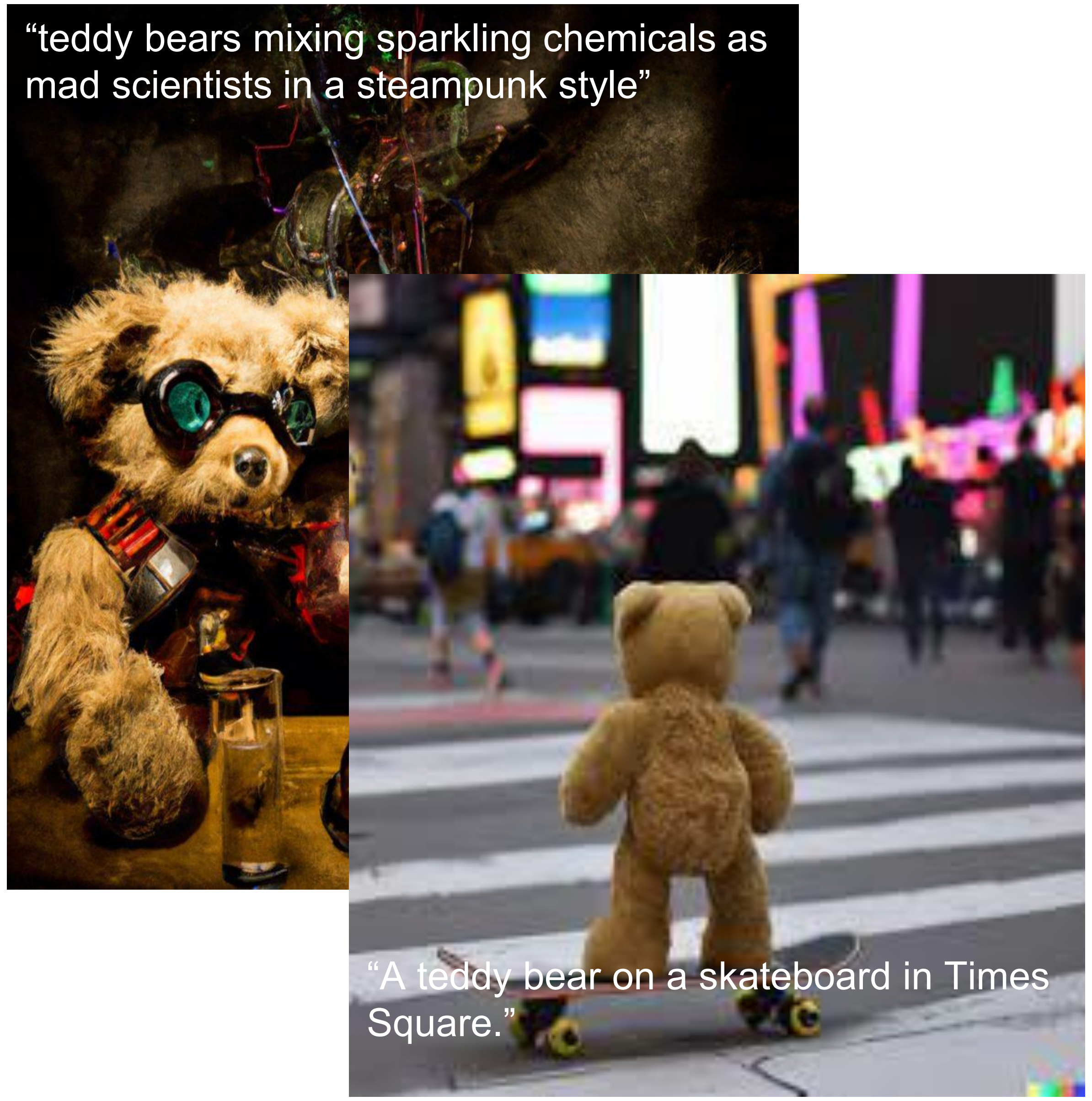
A painting of a majestic royal tall ship in Age of Discovery.

Photo credit: Minguk Kang et al.

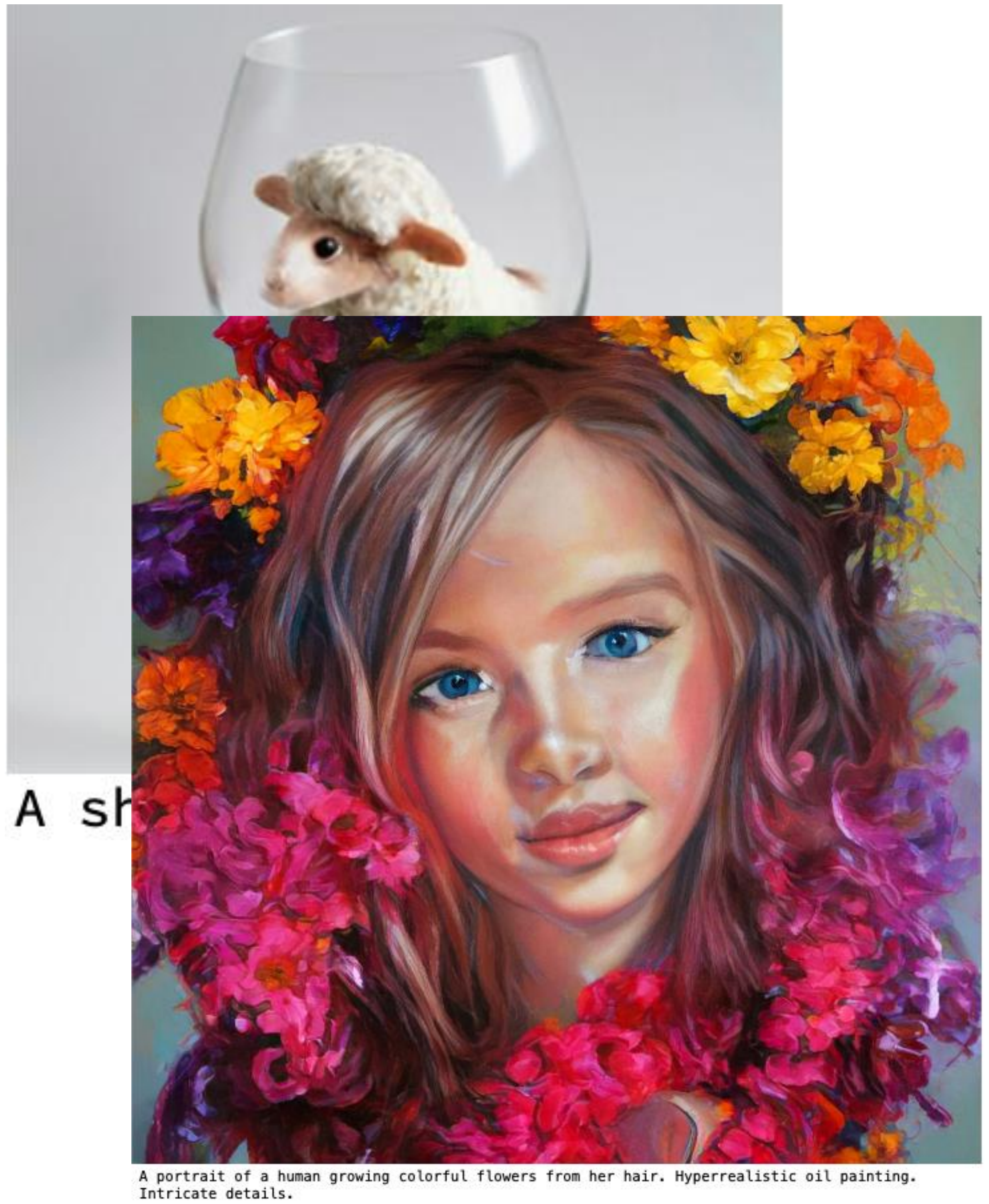
Text-to-Image Everywhere



Autoregressive models
(Image GPT, Parti)

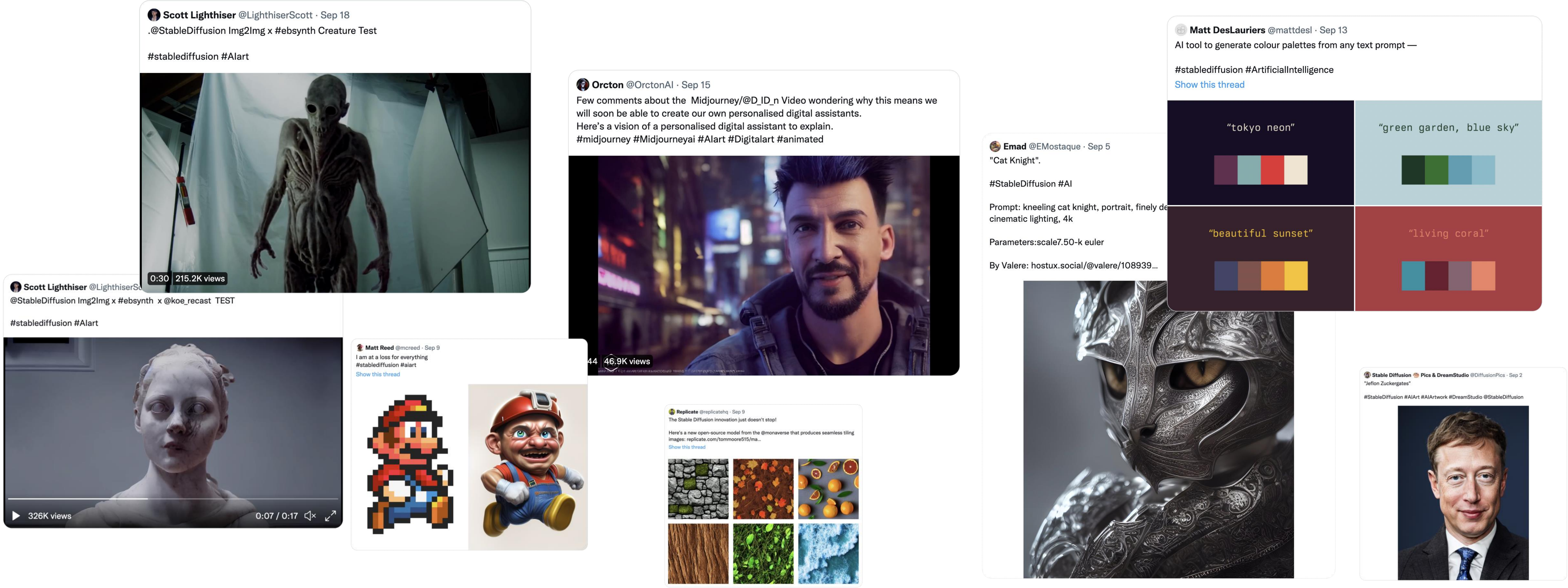


Diffusion models
(DALL-E 2, Imagen)



GANs, Masked GIT
(GigaGAN, MUSE)

Text-to-Image Everywhere



Slides credit: Robin Rombach

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

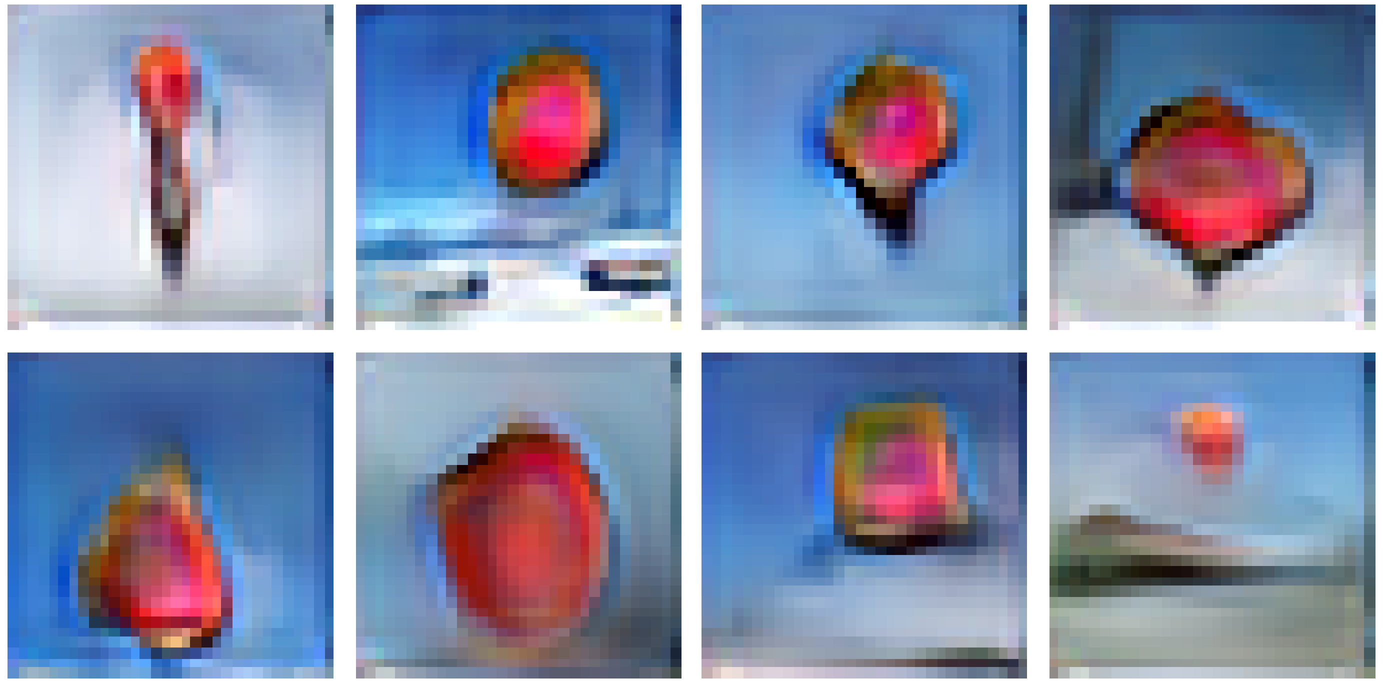
First Text-to-Image System

Therapy for people
with communicative disorders

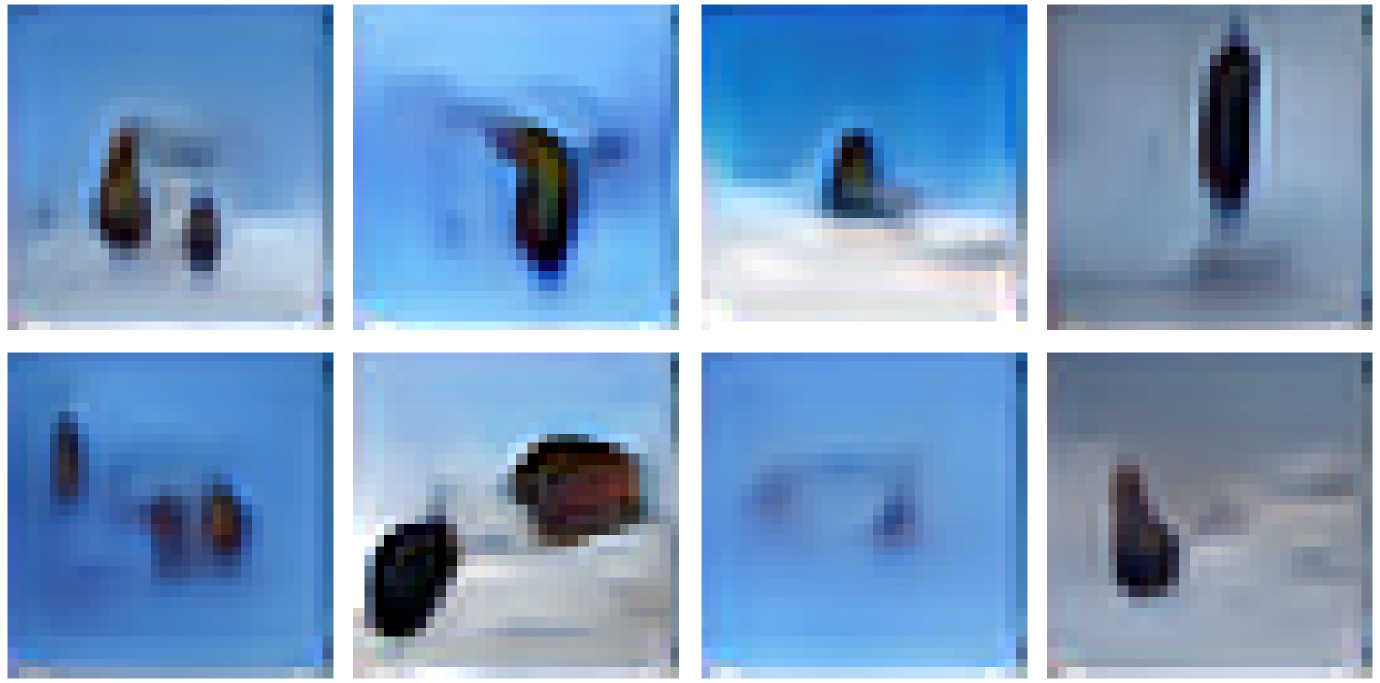
Math learning and reading comprehension
for young children

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.

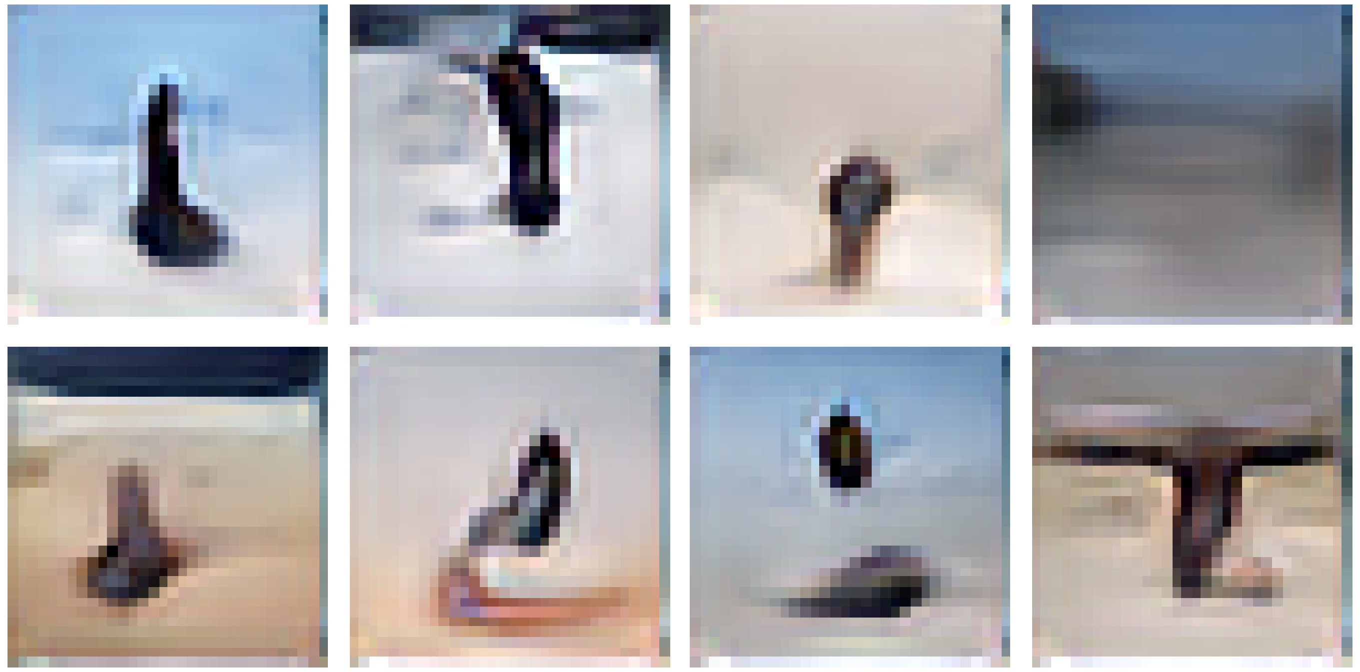


A herd of elephants flying in the blue skies.

Generating Images from Captions with Attention.
Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, Ruslan Salakhutdinov. ICLR 2016.

First Deep Learning Work

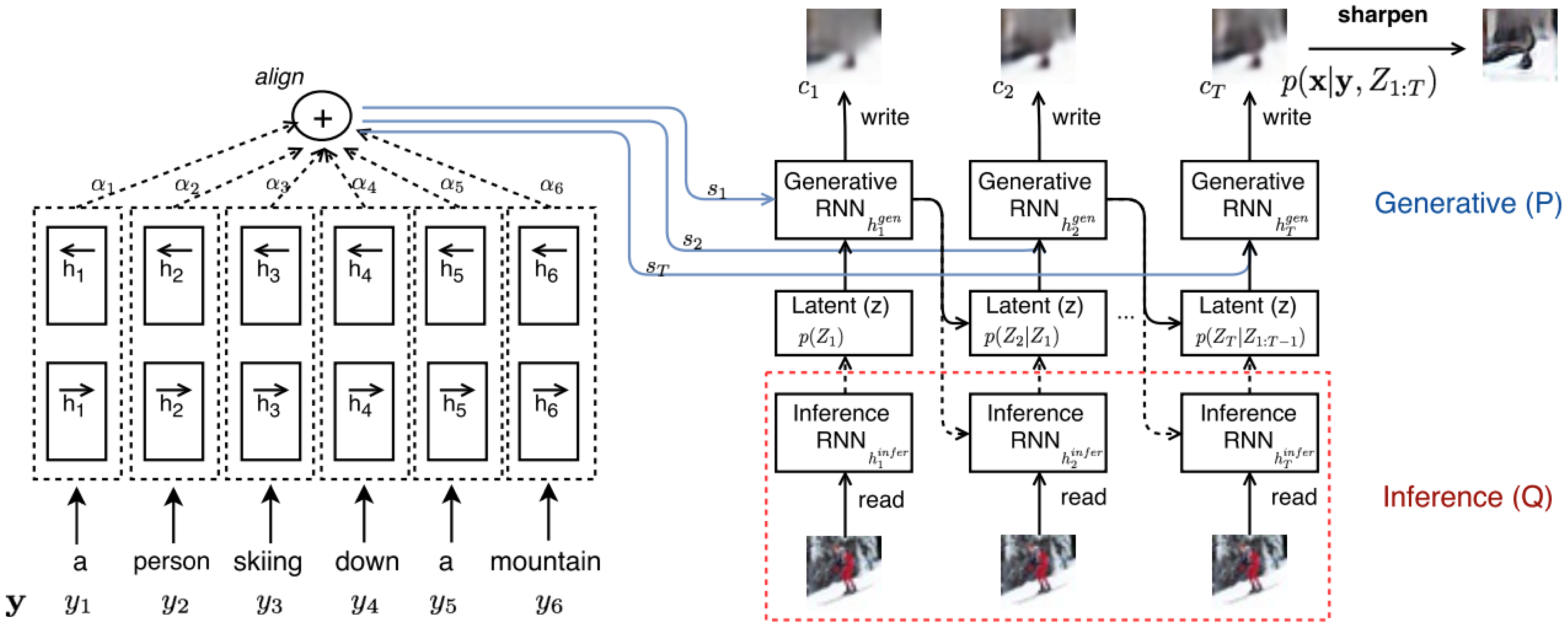
A toilet seat sits open in the grass field.



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.

Can we improve it?

How can 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.

GAN-based Text-to-Image

Generative Adversarial Text to Image Synthesis
Scott Reed et al., ICML 2016

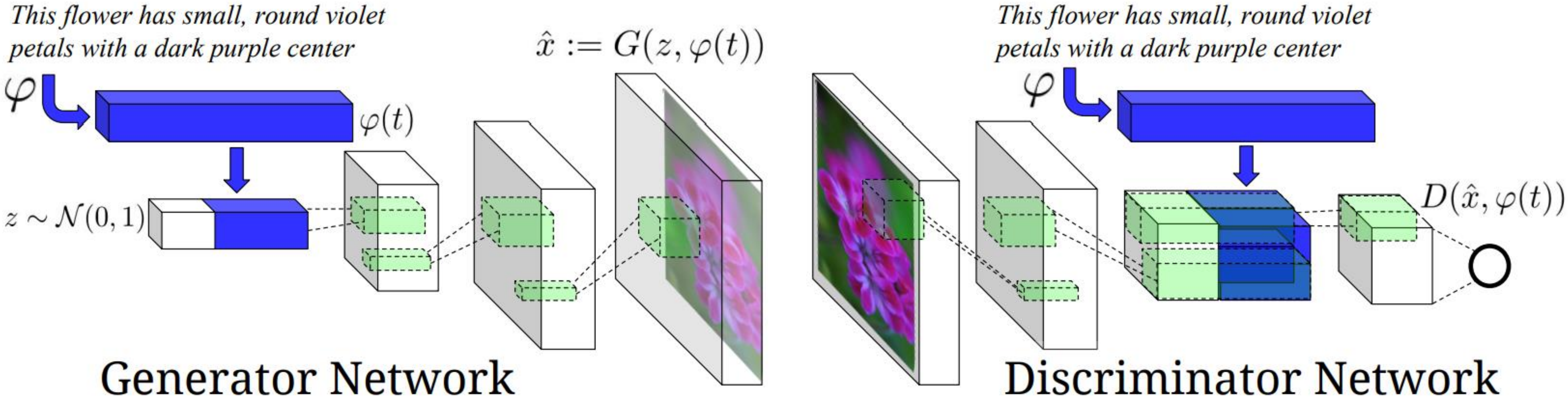
GAN-based Text-to-Image

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

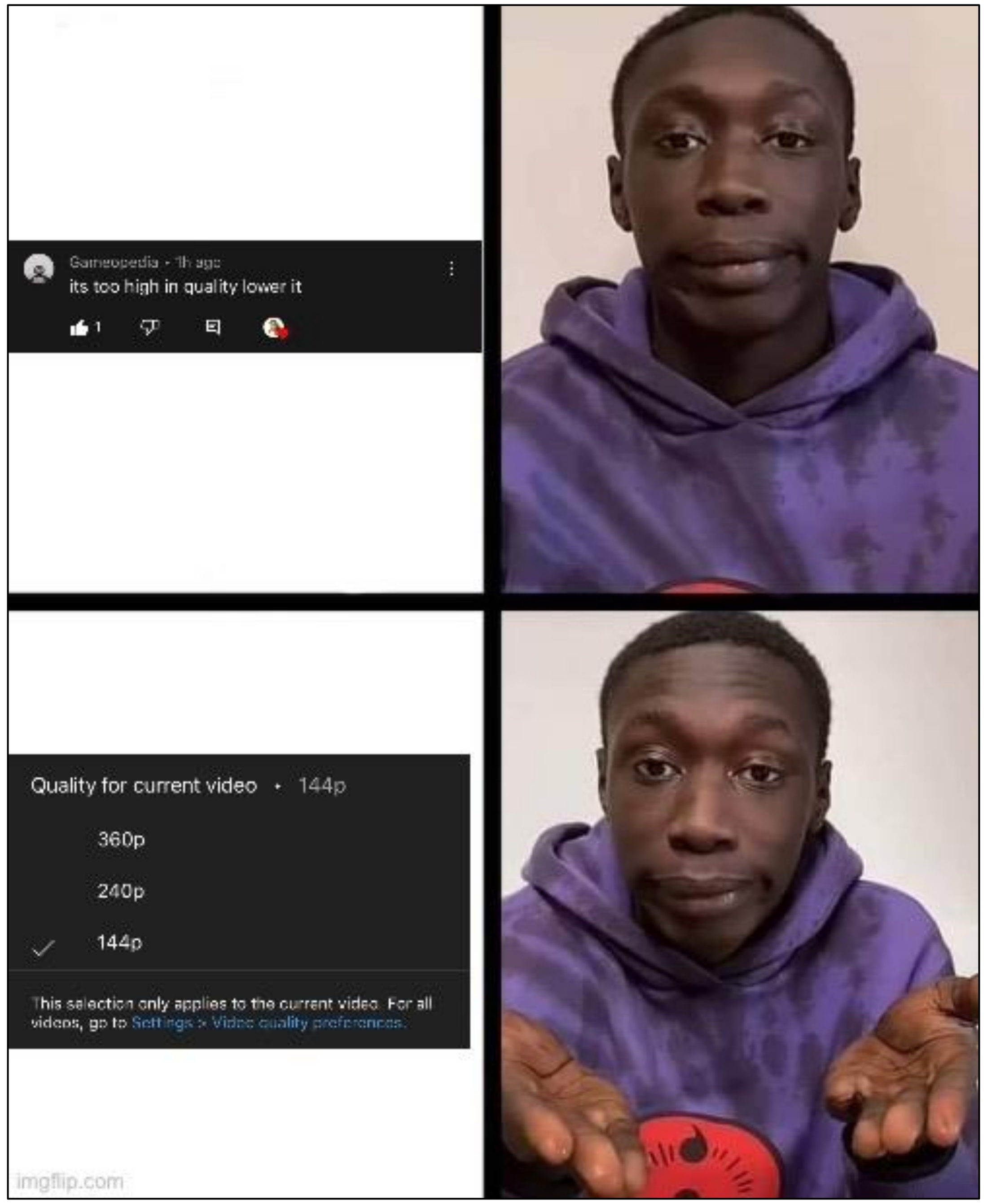
Generative Adversarial Text to Image Synthesis
Scott Reed et al., ICML 2016

GAN-based Text-to-Image



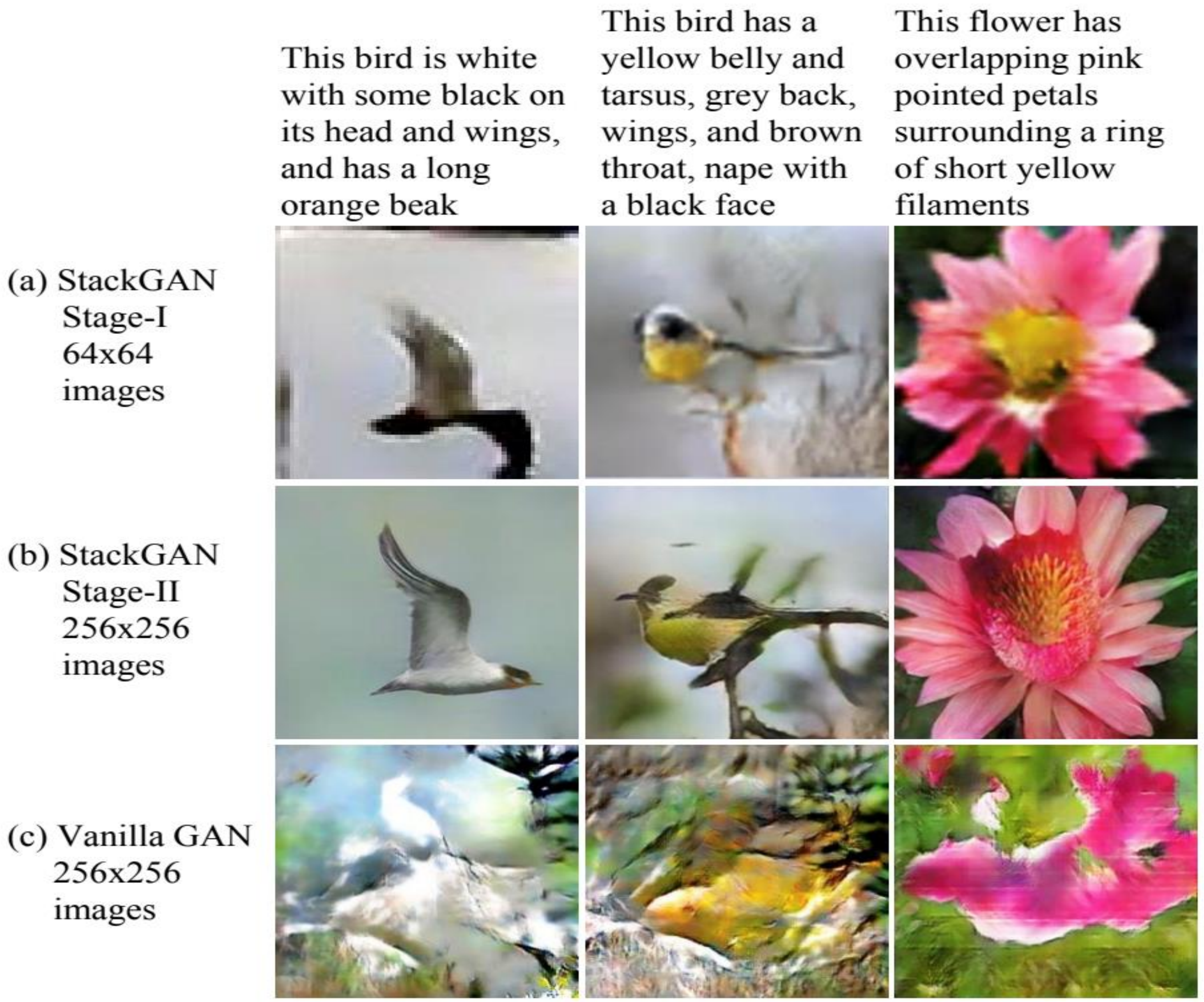
Conditional GAN + CNN + concatenation

Generative Adversarial Text to Image Synthesis
Scott Reed et al., ICML 2016



But these images are tiny ...
How can we make them HD?

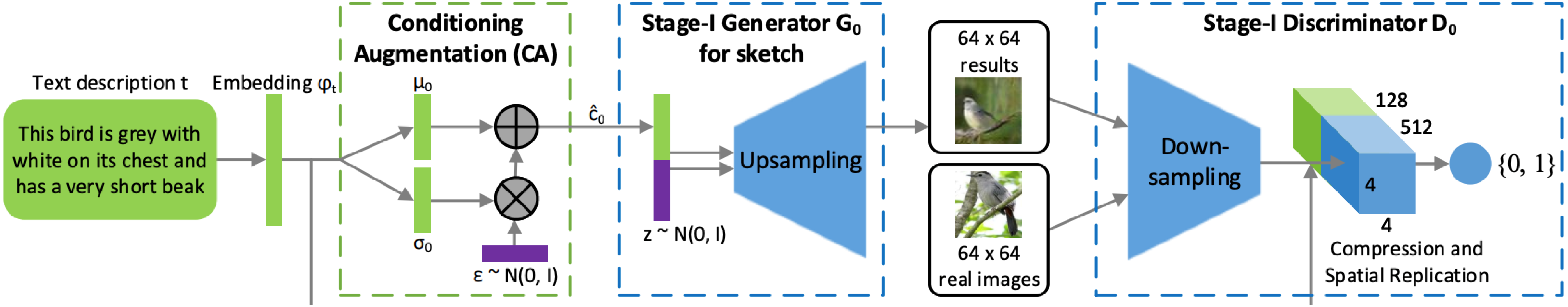
+Two-stage Models



Two-stage Conditional GAN + CNN + concatenation

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

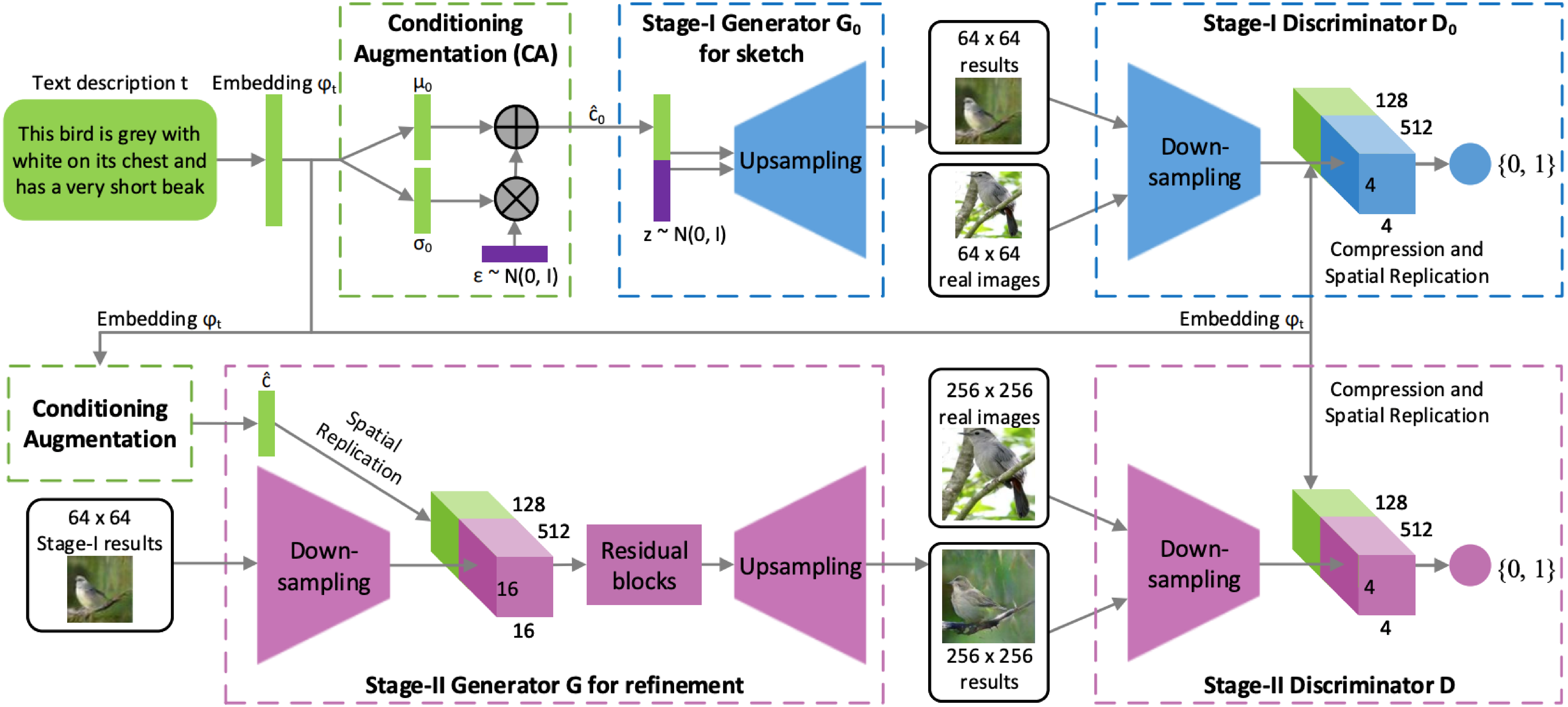
+Two-stage Models



Two-stage Conditional GAN + CNN + concatenation

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

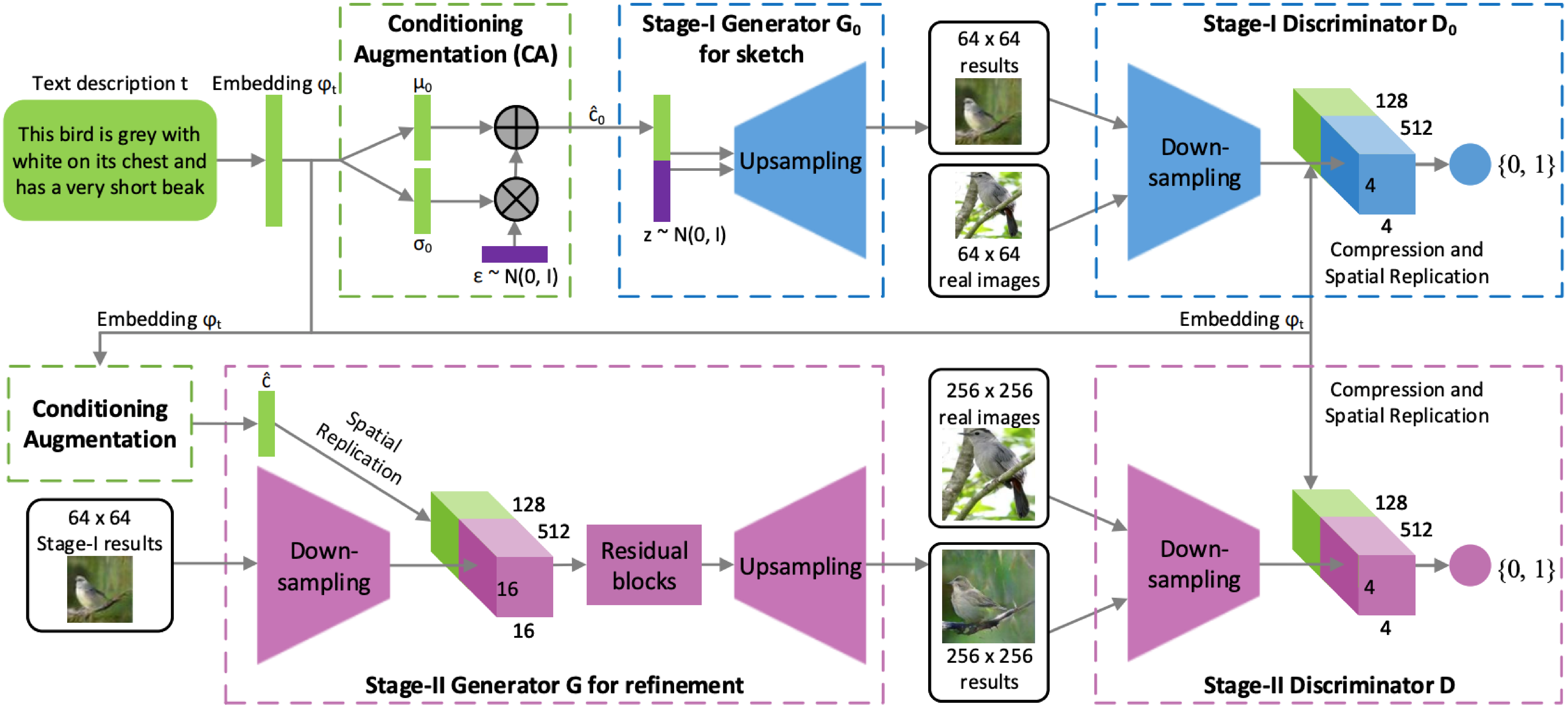
+Two-stage Models



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

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

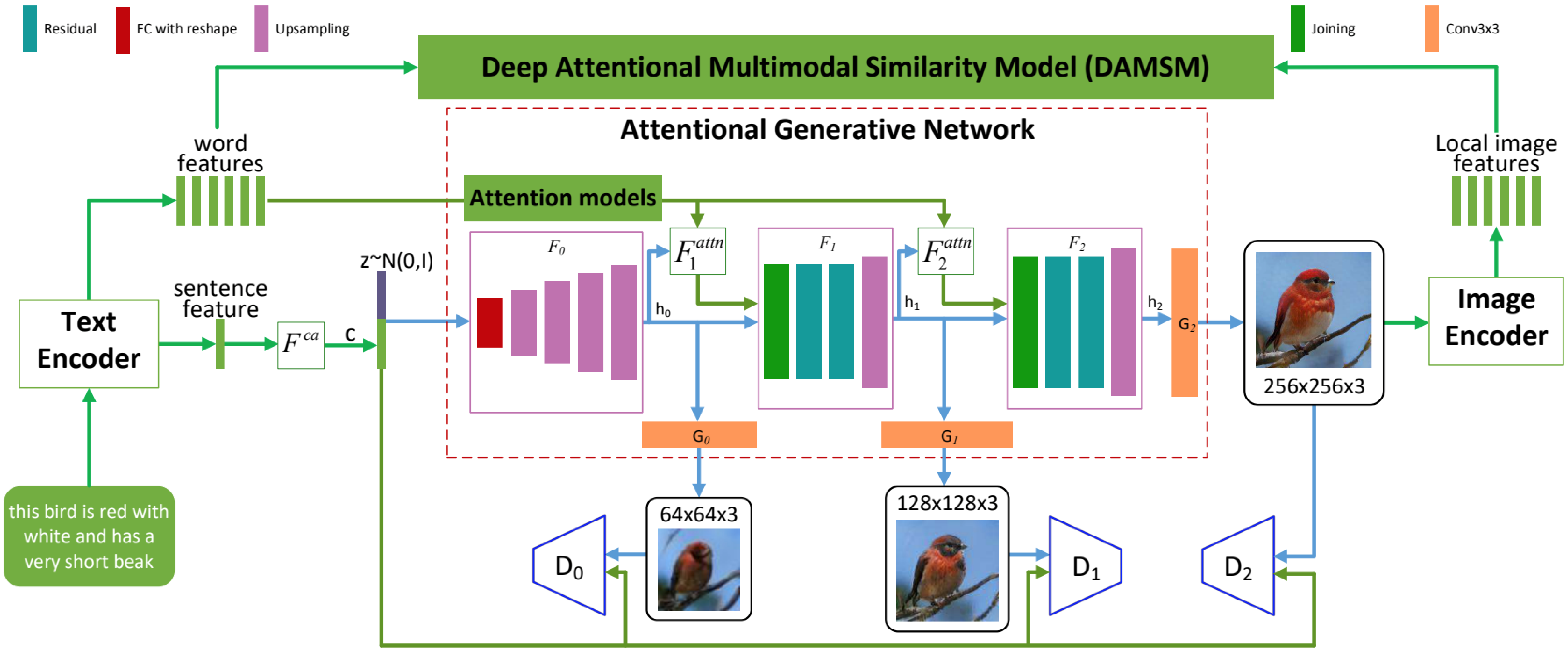
+Two-stage Models

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

+ 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

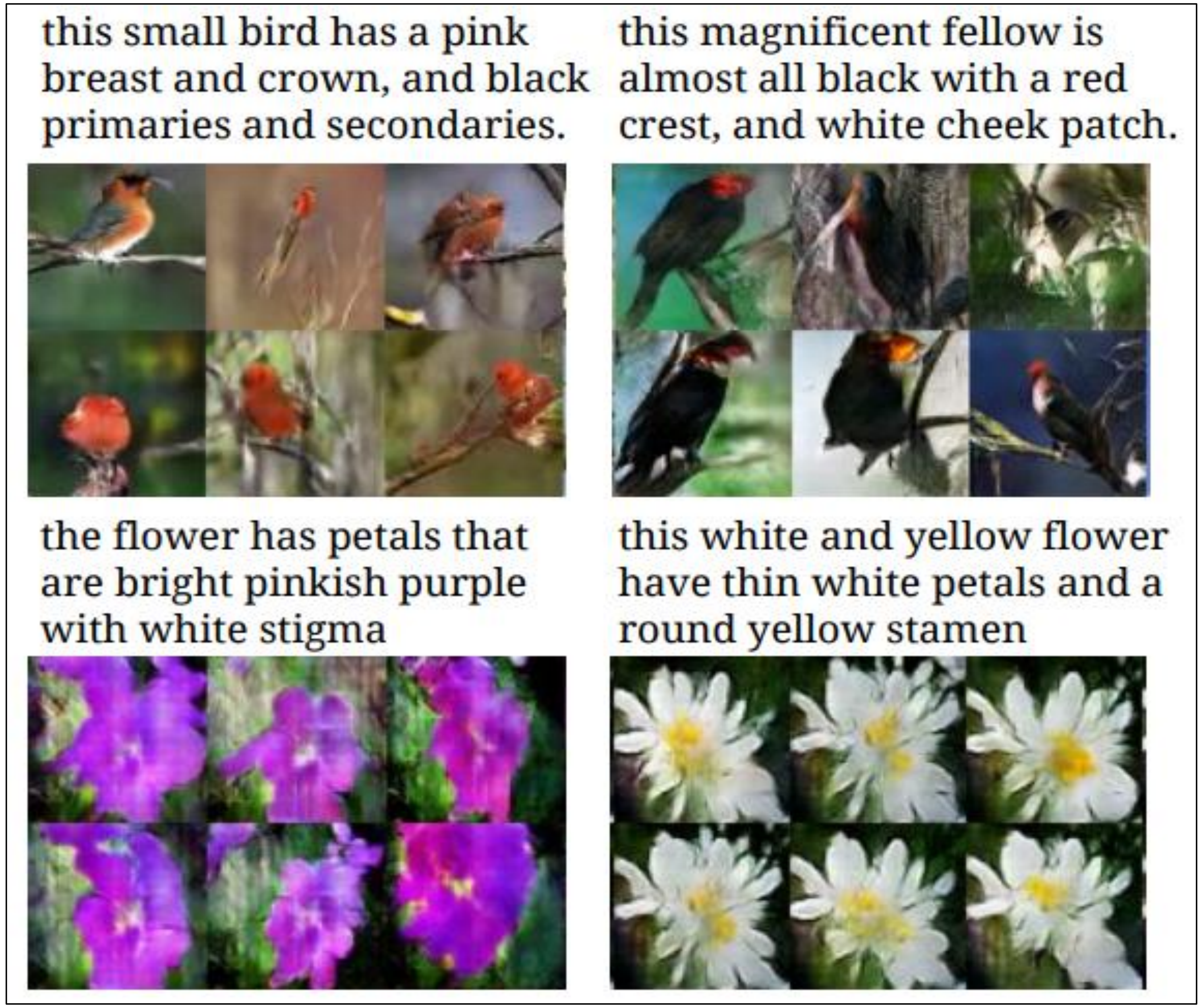
+ Cross-attention to connect Text and Image

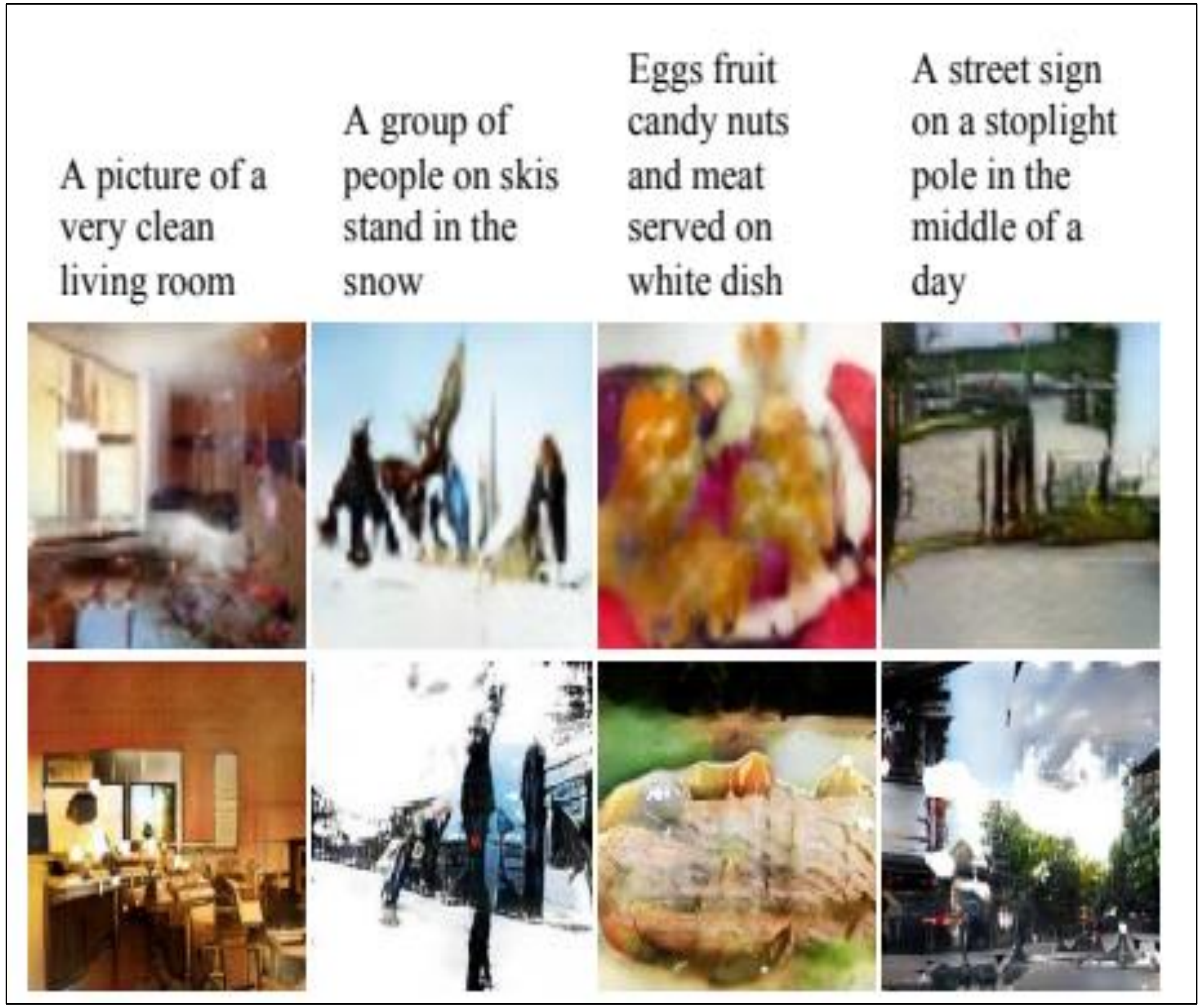


AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks
Tao Xu et al., CVPR 2018

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

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





"GAN_INT_CLS" Reed et al. ICML 2016 https://arxiv.org/abs/1605.05396

"StackGAN" Zhang et al. ICCV 2017 https://arxiv.org/abs/1612.03242

Can 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

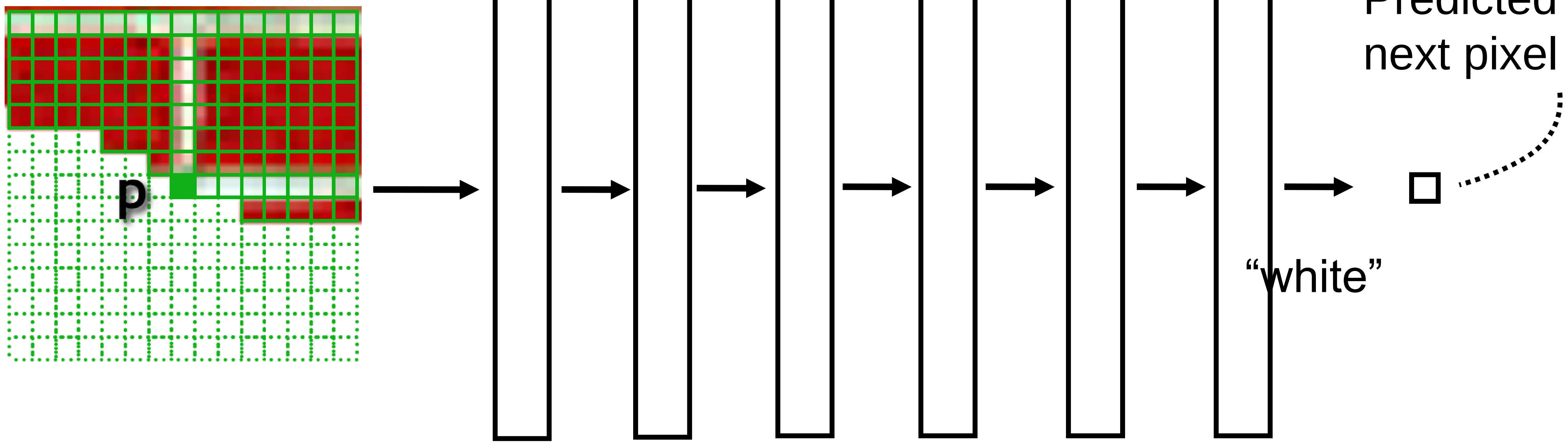
Designed to learn long-range interactions on sequential ... state-of-the-art results

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 ... the locality of interactions and is therefore free ... inputs. However,

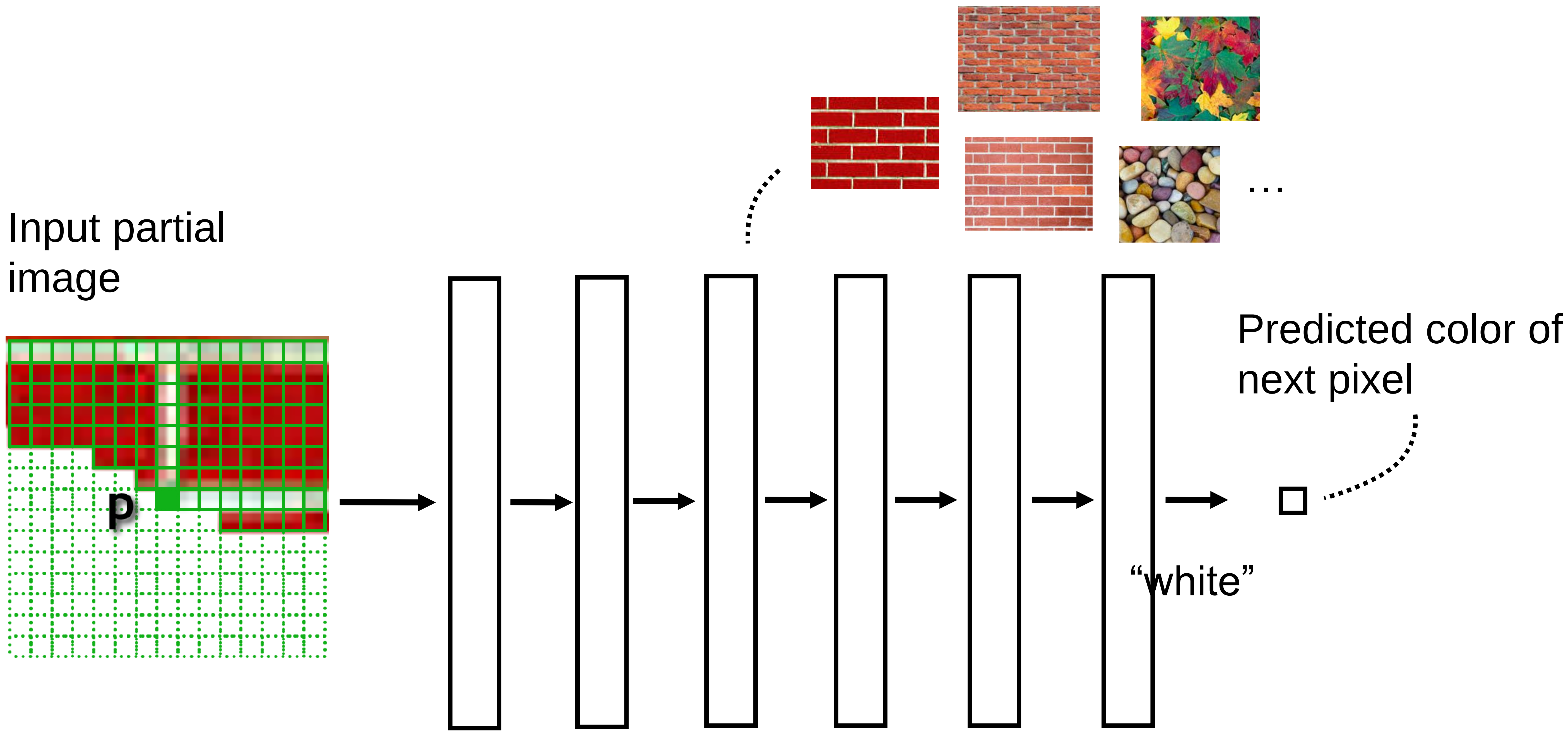
Autoregressive (AR) image synthesis

Input partial image

Predicted color of next pixel

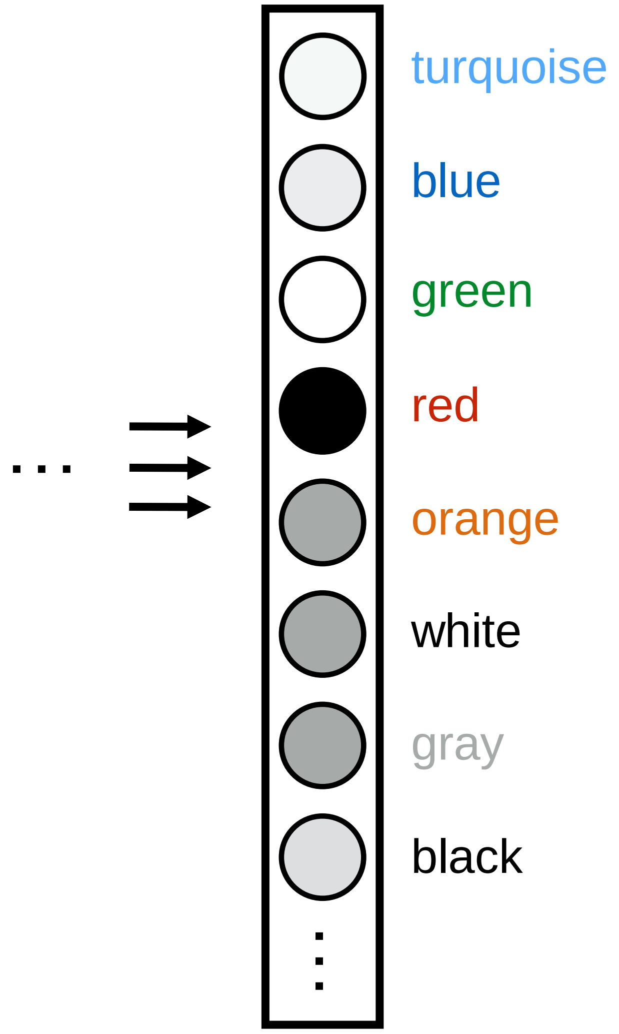


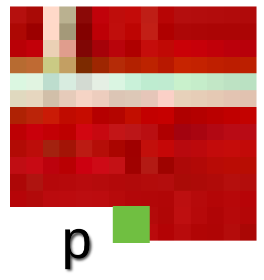
[PixelRNN, PixelCNN, van der Oord et al. 2016]



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

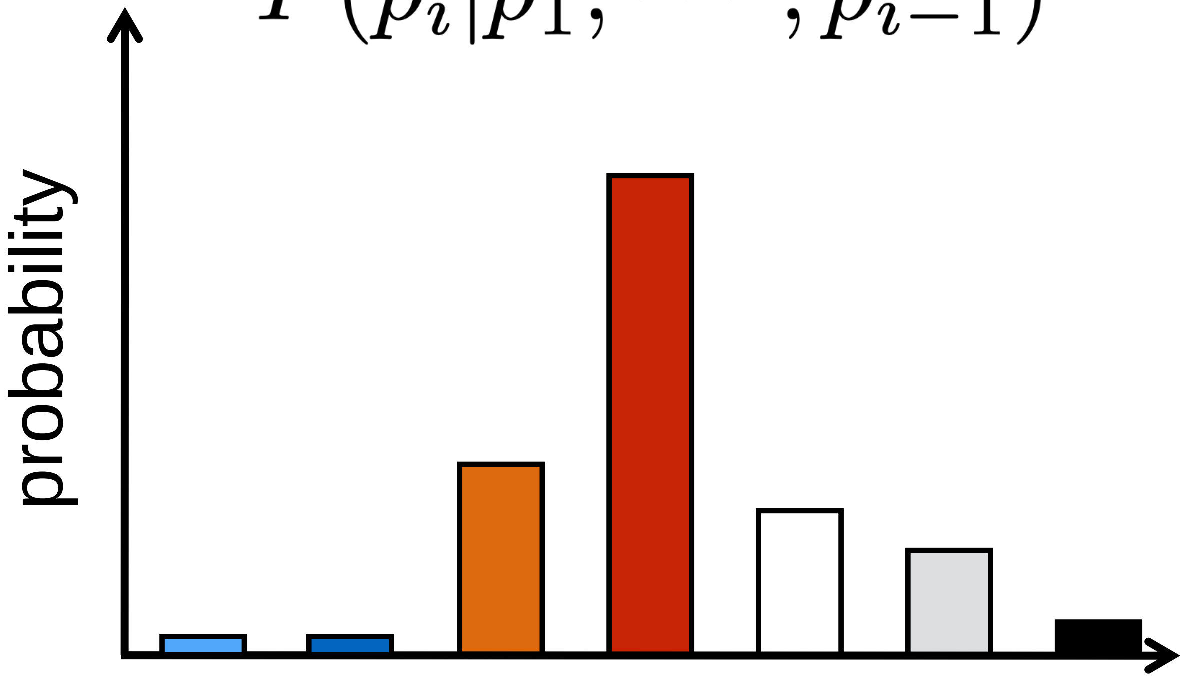
Network output



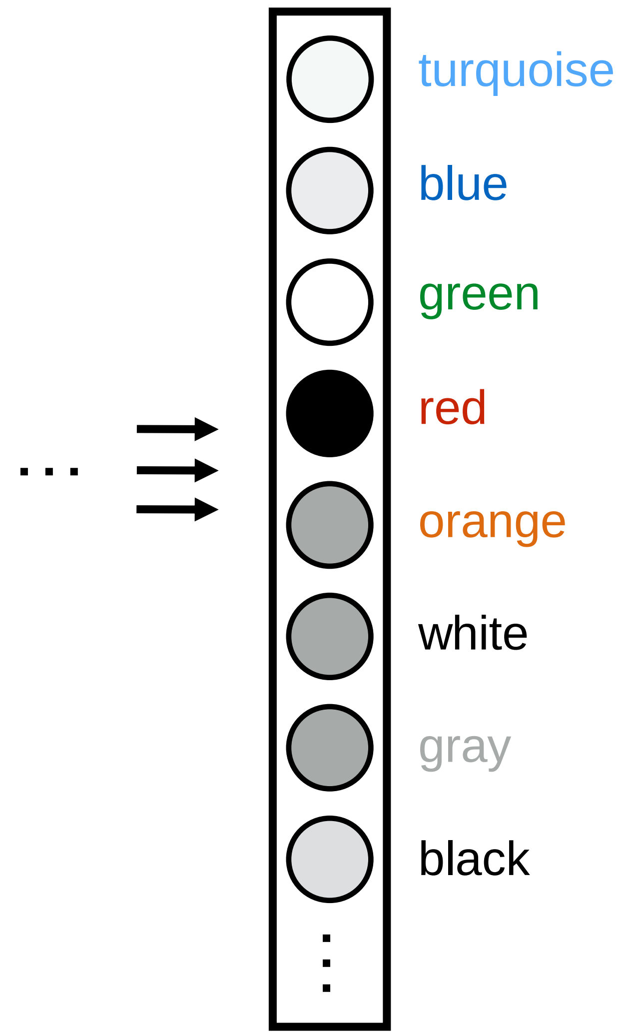


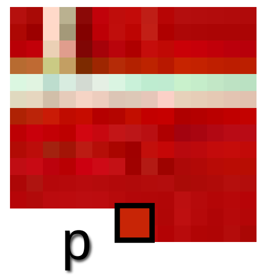
P(next pixel | previous pixels)

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

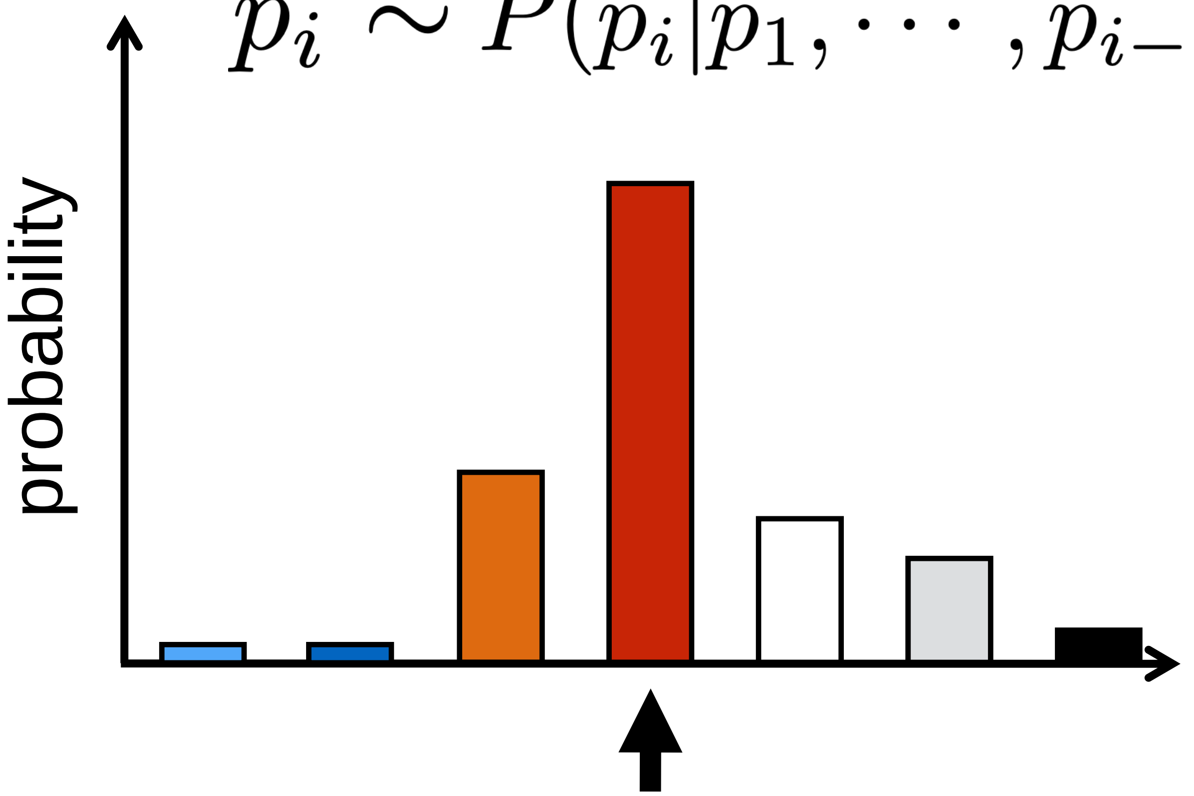


Network output

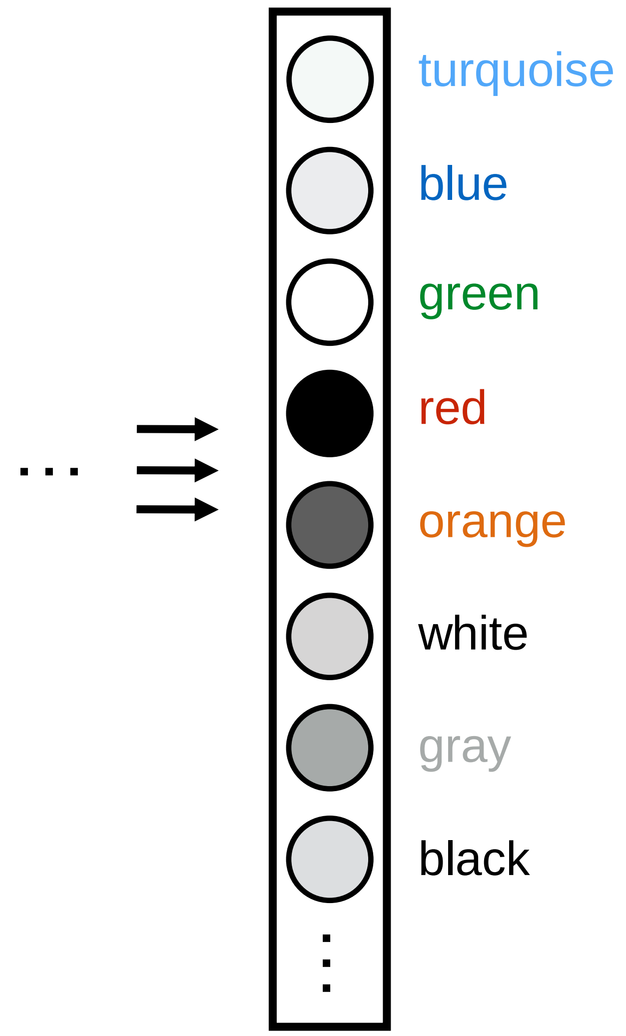


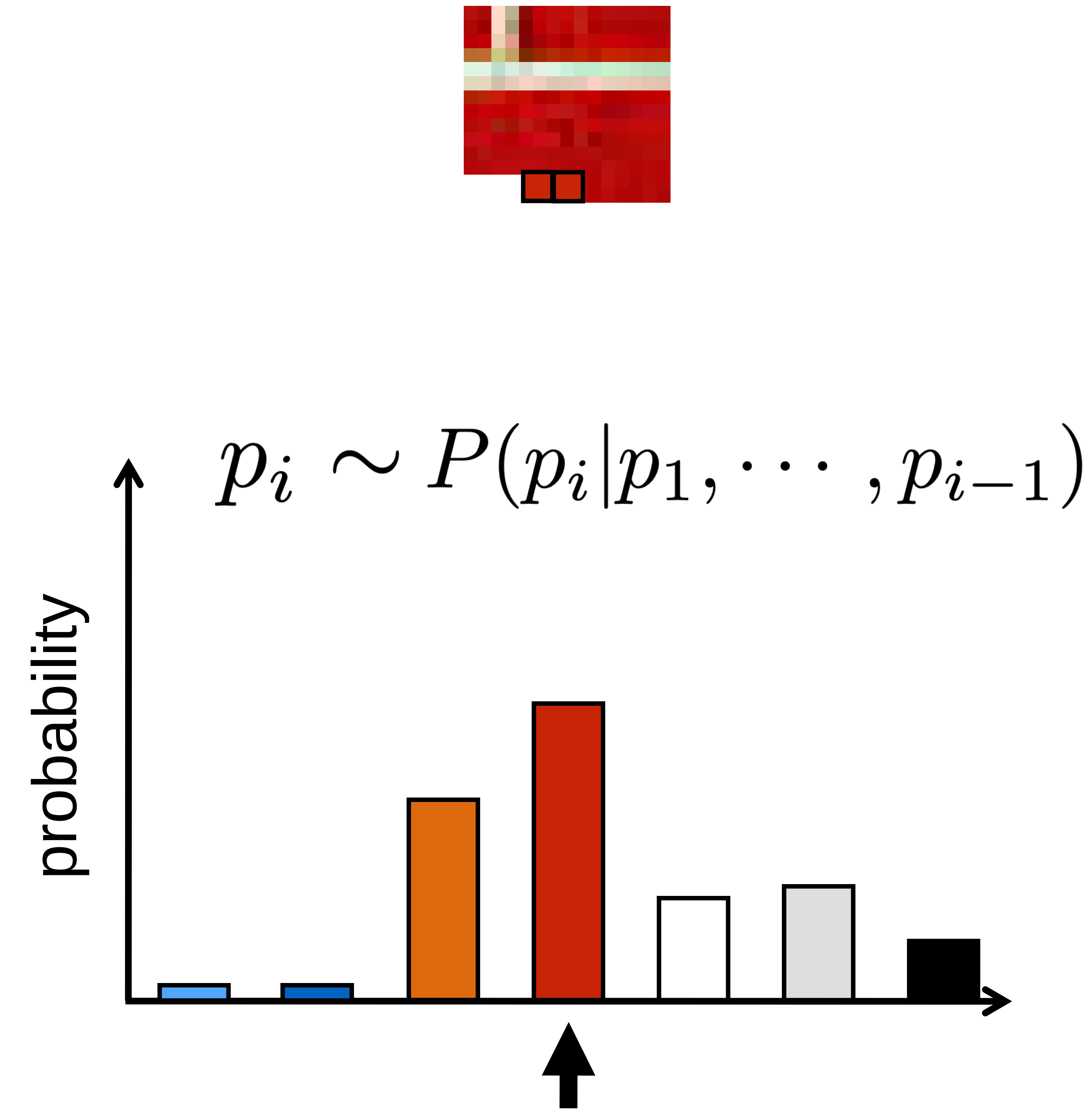


$$p_i \sim P(p_i | p_1, \cdots, p_{i-1})$$

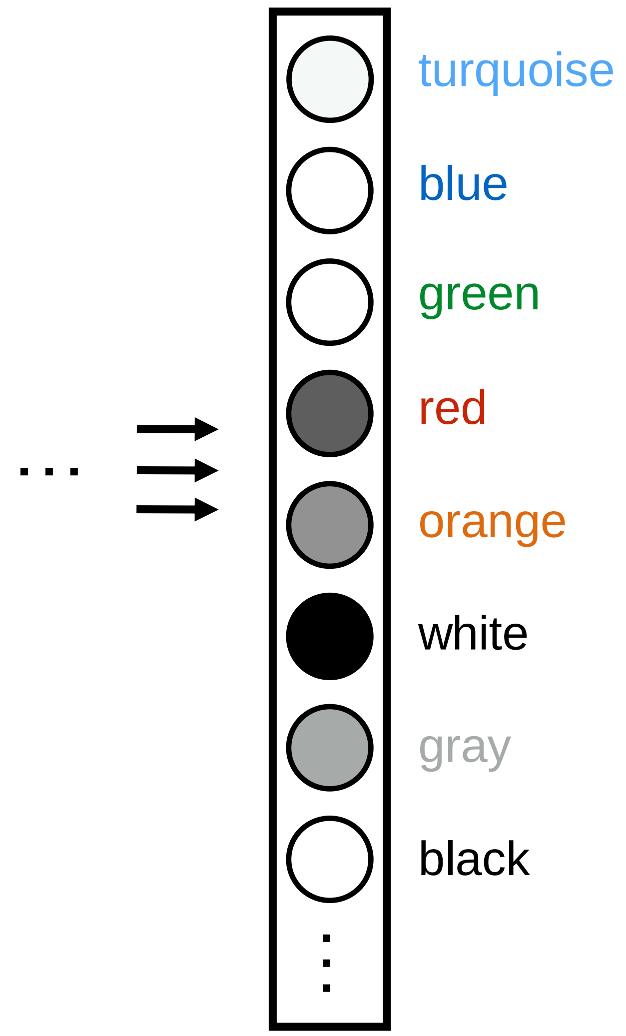


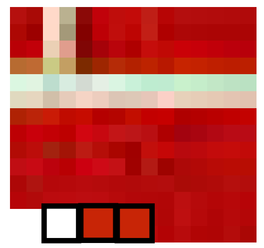
Network output



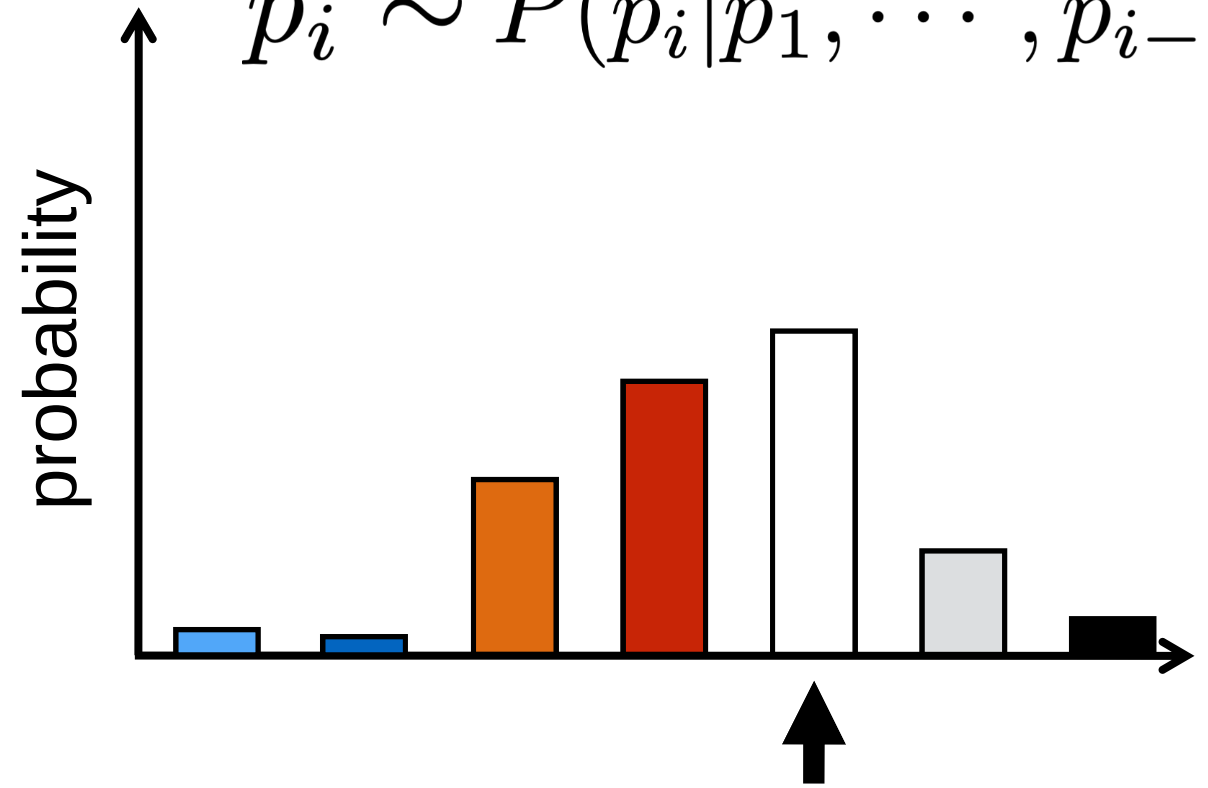


Network output

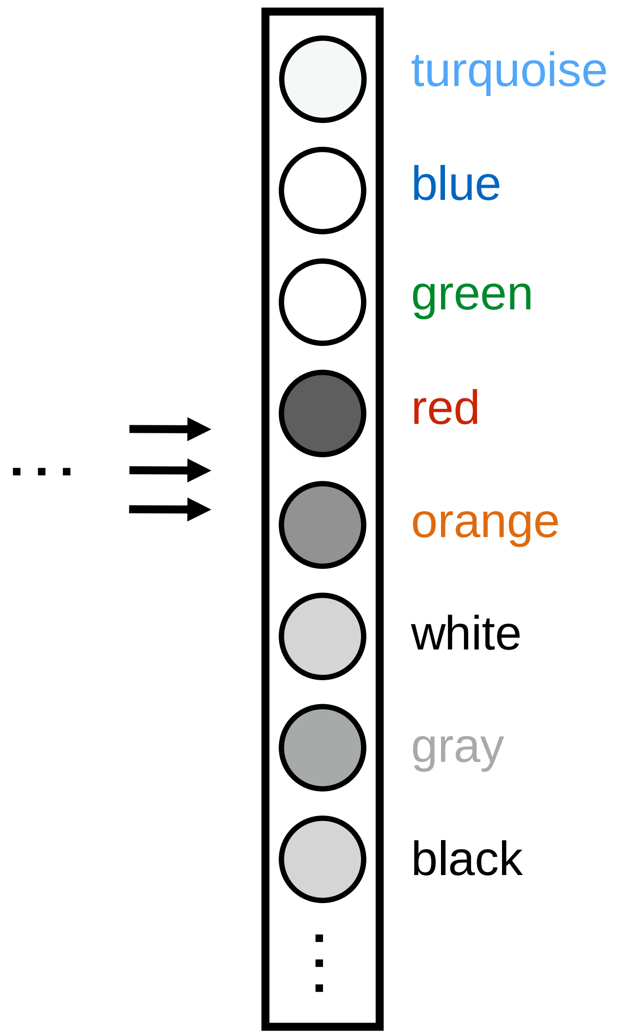


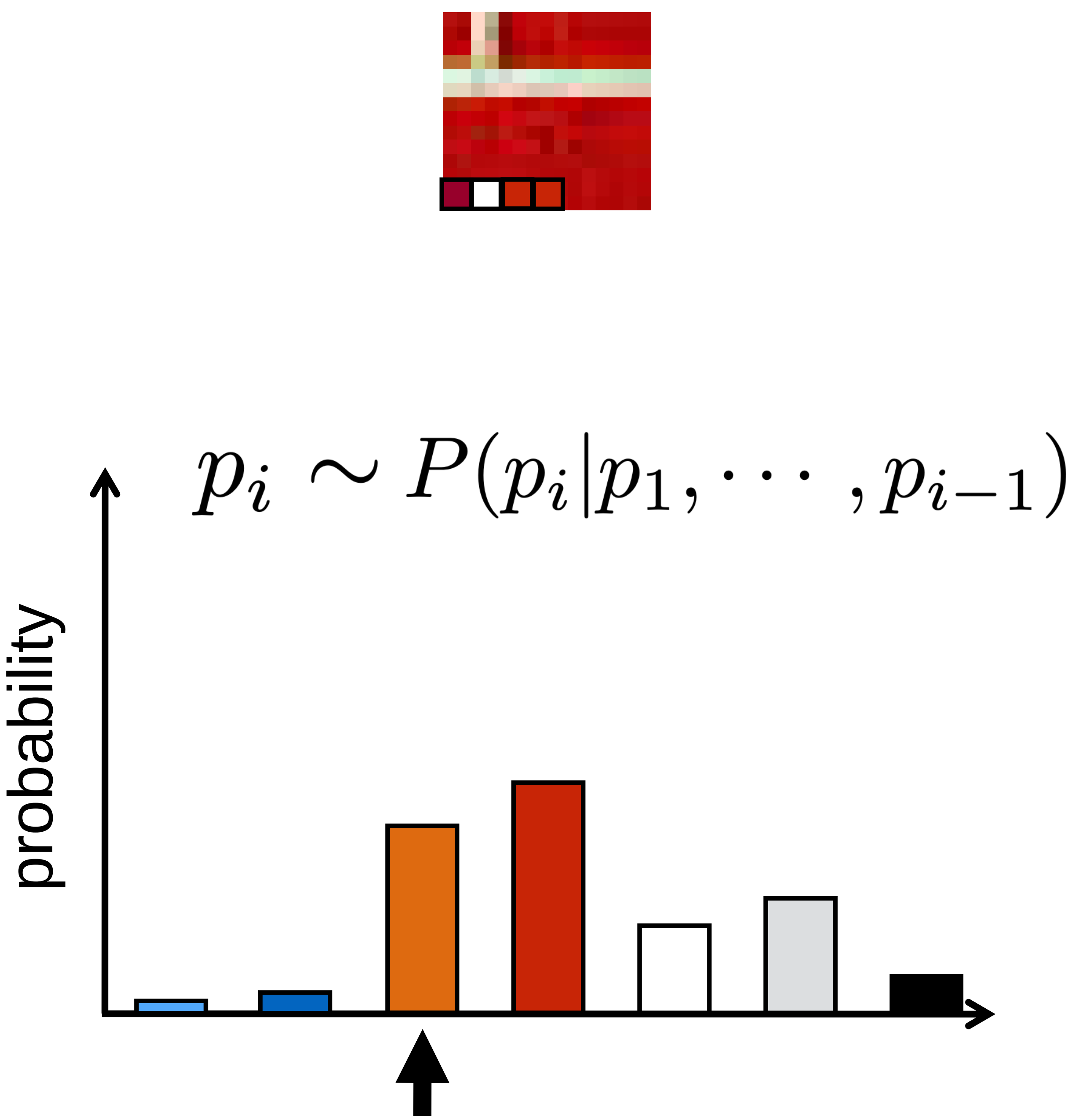


$$p_i \sim P(p_i | p_1, \cdots, p_{i-1})$$



Network output

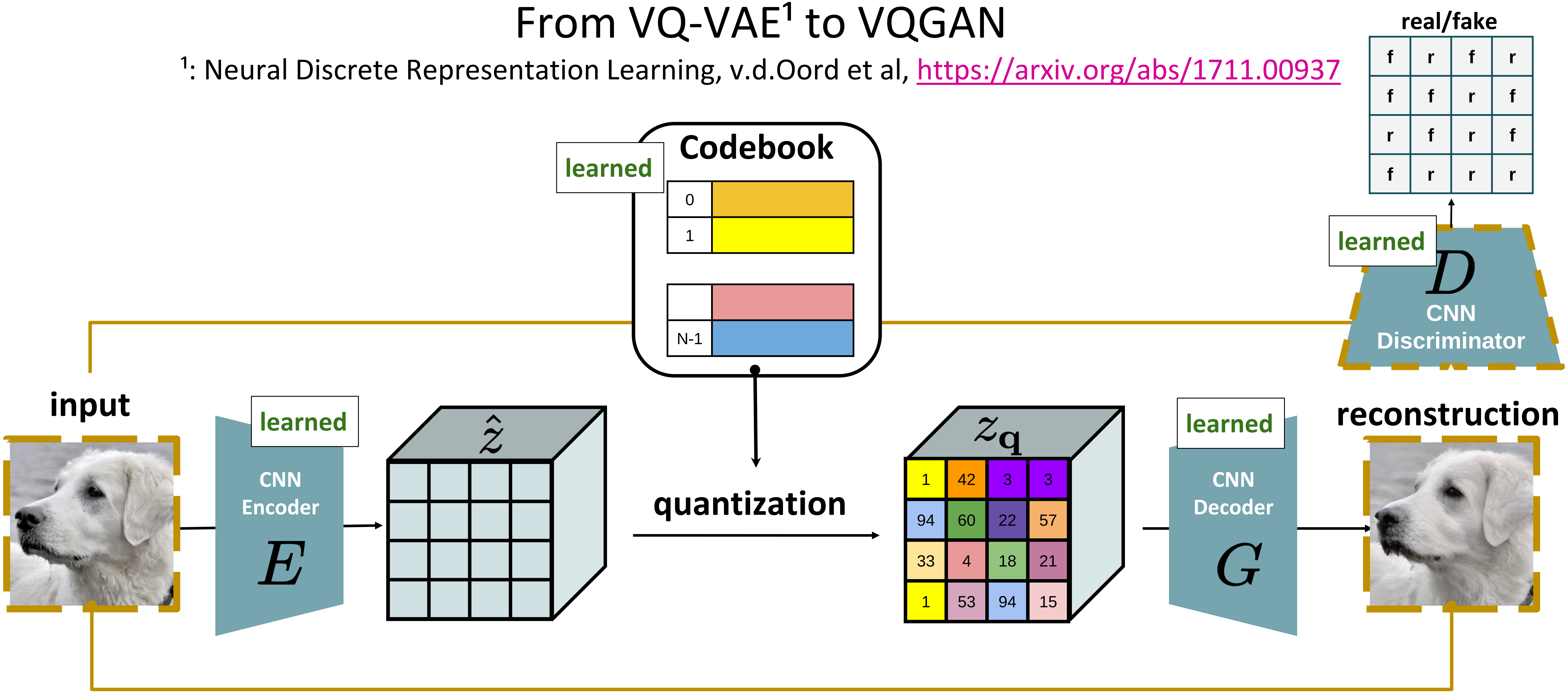




Generation is super slow?
What should we do?

From VQ-VAE[1] to VQGAN

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



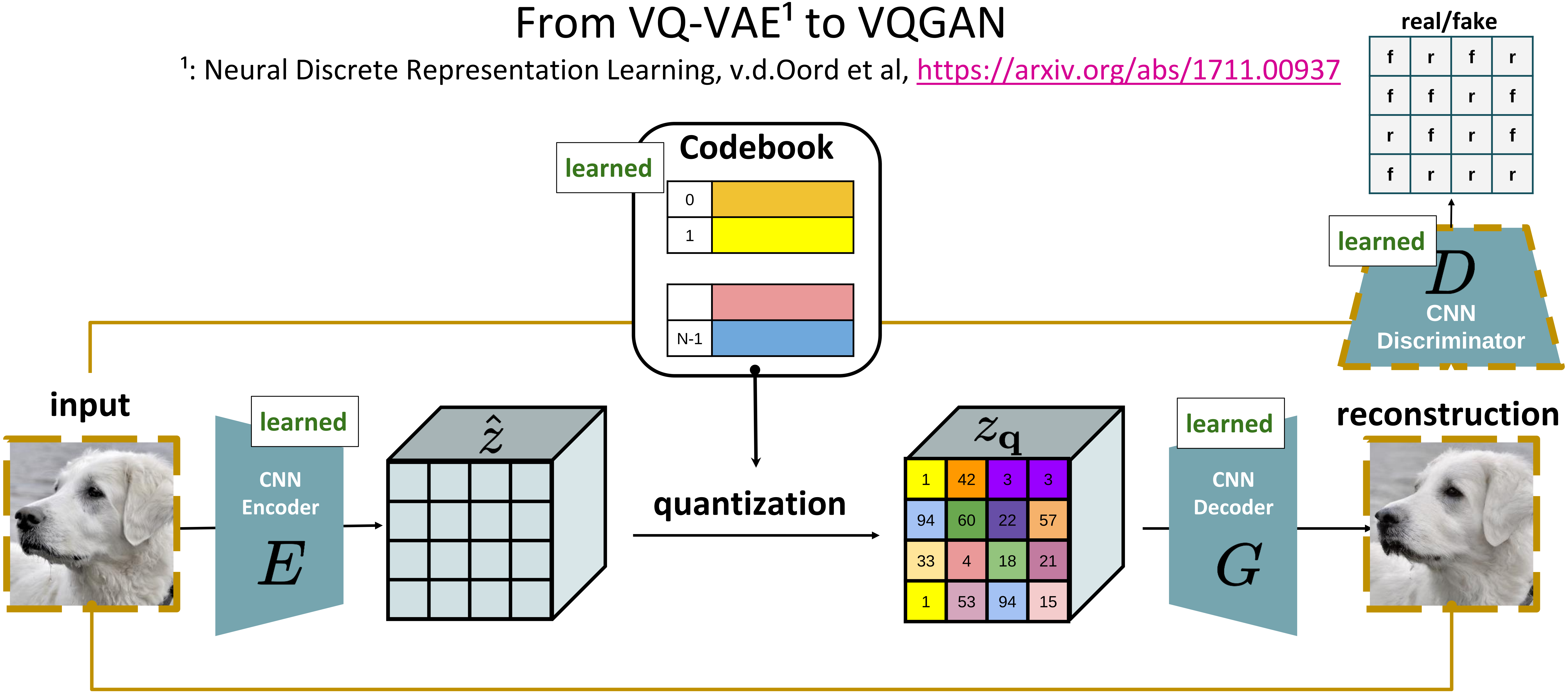
i) replace L2/L1 rec. loss with Perceptual loss (includes pixel-level)

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

Slide credit: Robin Rombach

From VQ-VAE[1] to VQGAN

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



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

Slide credit: Robin Rombach

Slide credit: Robin Rombach

Scaling VQGAN for Text-to-Image!

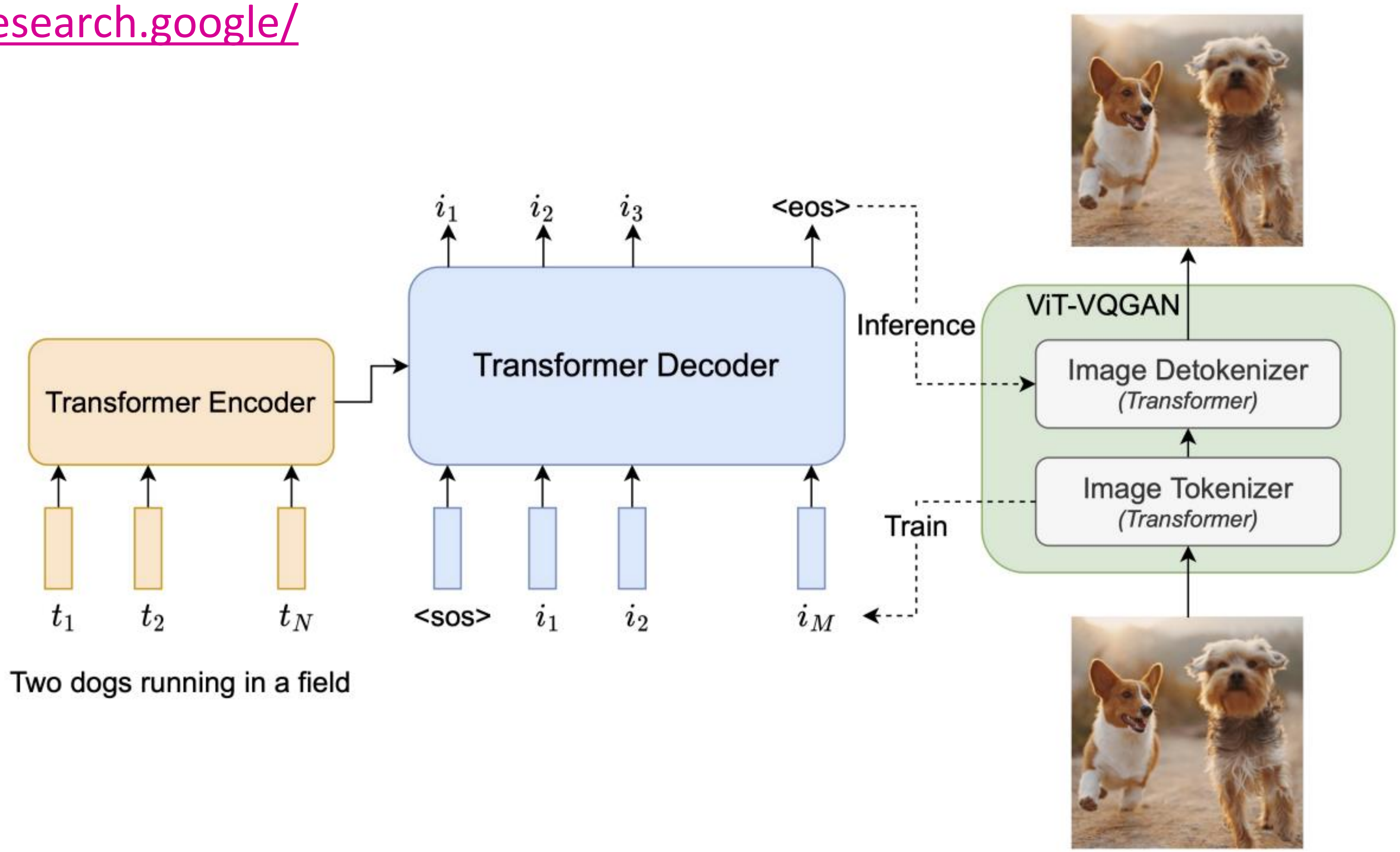
- see recently released “Parti” paper by Google (text-to-image model)
  - https://parti.research.google/

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!

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/



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

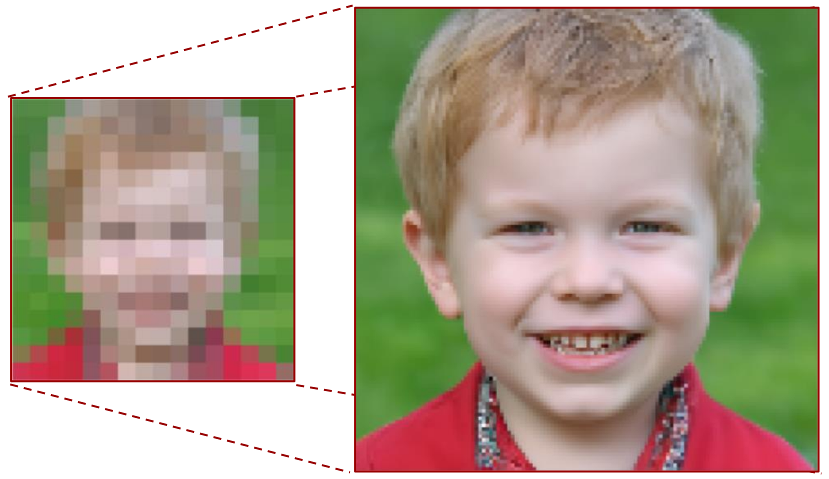


Image Super-Resolution via Iterative Refinement

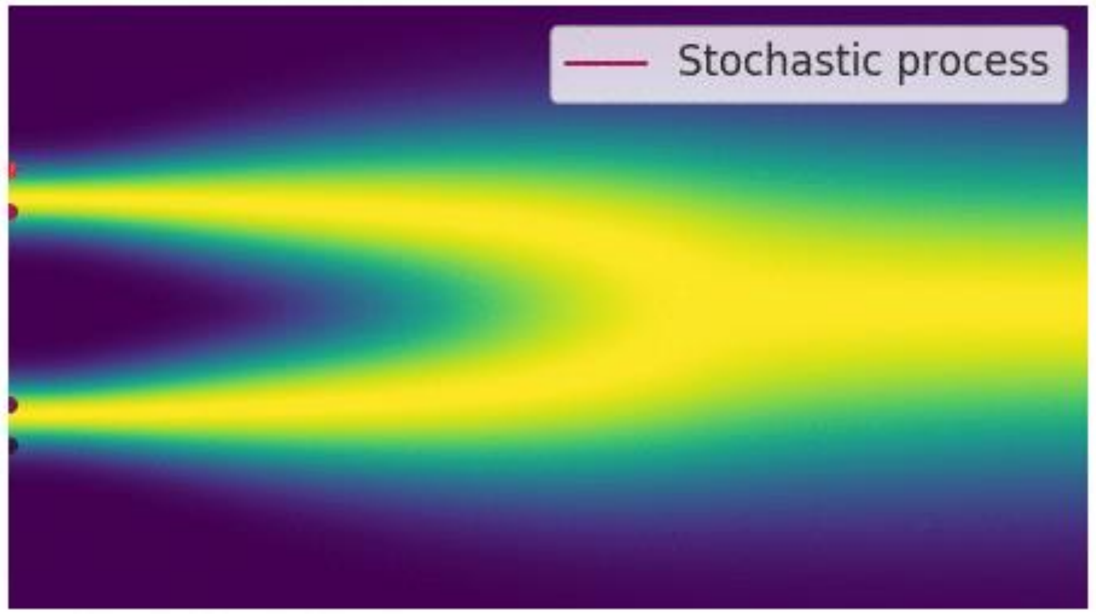
Chitwan Saharia, et al

https://arxiv.org/abs/2104.07636

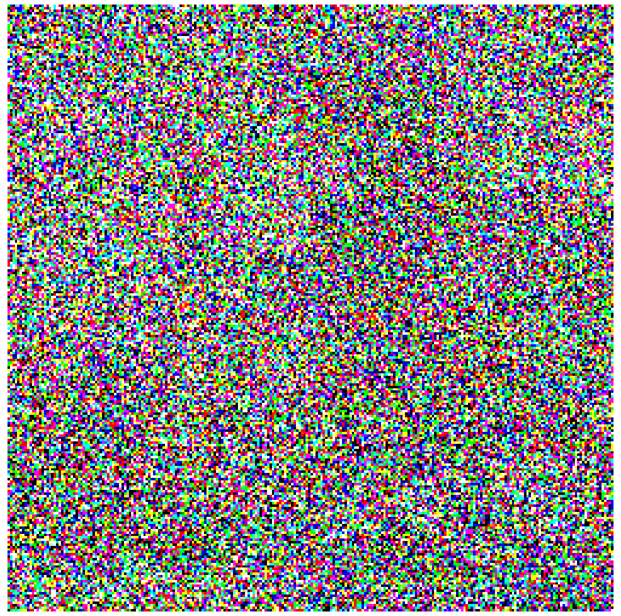
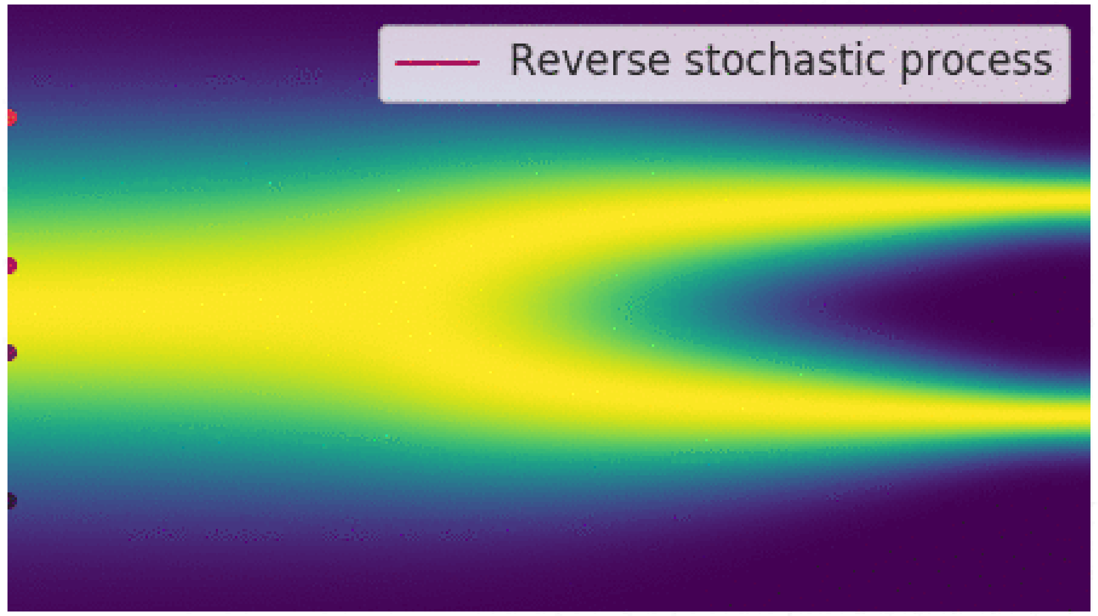
… but very expensive :(

Slide credit: Robin Rombach

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

Animations from https://yang-song.github.io/blog/2021/score/

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

Forward diffusion process (fixed)

Data

Noise

Reverse denoising process (generative)

[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]

Slide credit: Karsten Kreis et al.

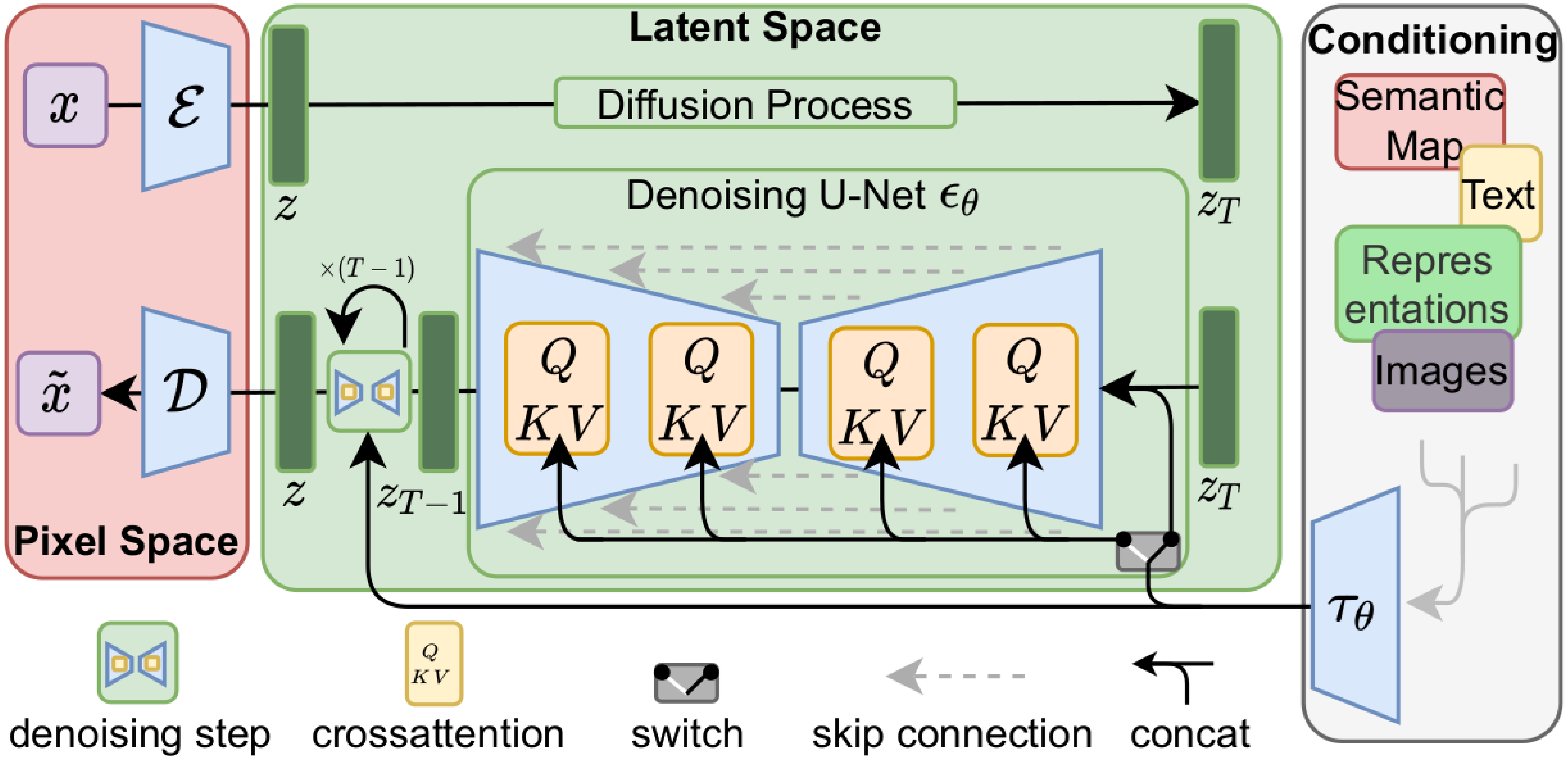
Latent Diffusion Modeling: Architecture

Autoencoder with KL or VQ regularization.

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

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

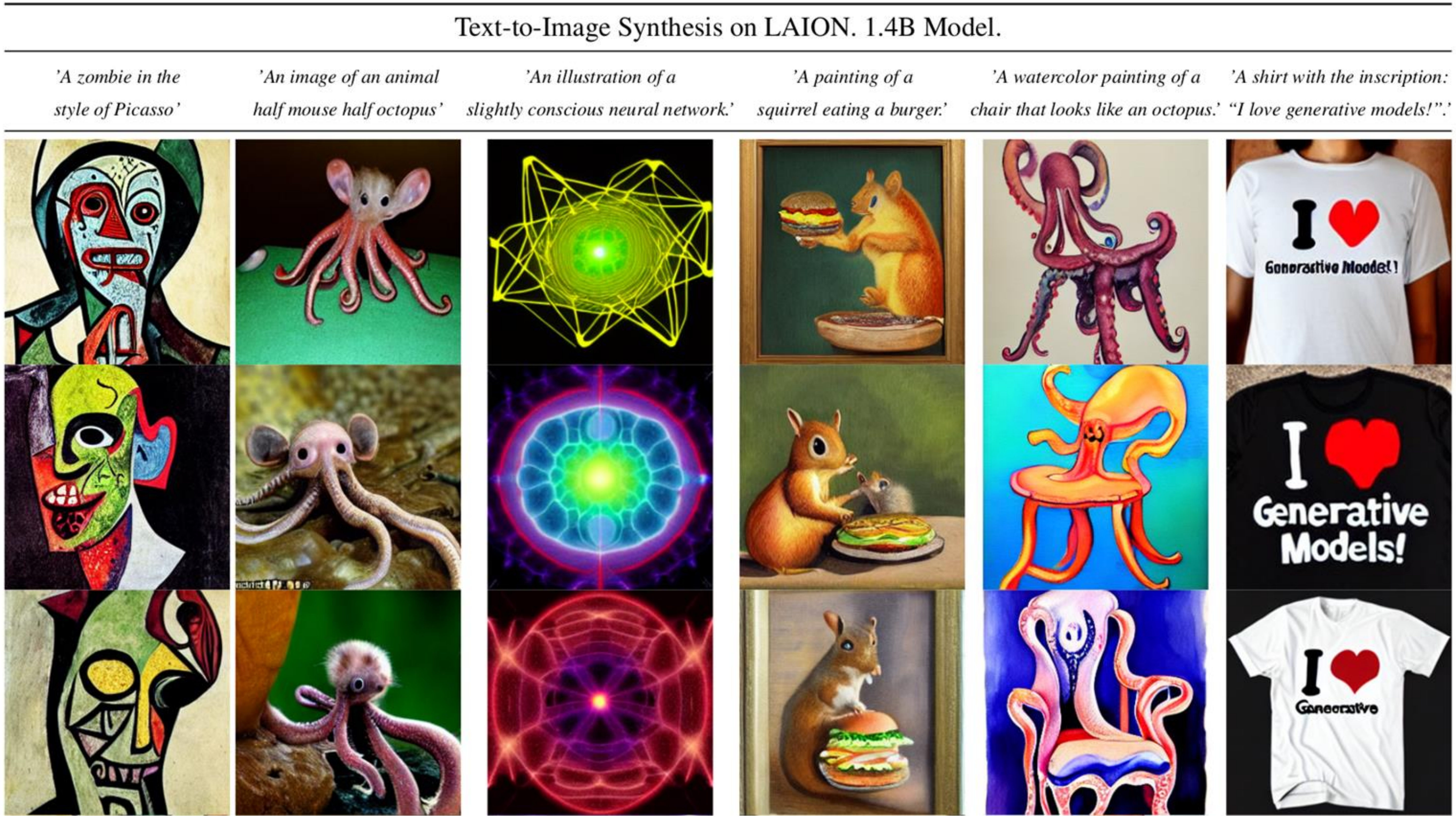
where $\lambda = \frac{\nabla_{G_L}[\mathcal{L}_{rec}]}{\nabla_{G_L}[\mathcal{L}_{GAN}]+\delta}$



Slide credit: Robin Rombach

LDMs for Text-to-Image Synthesis

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

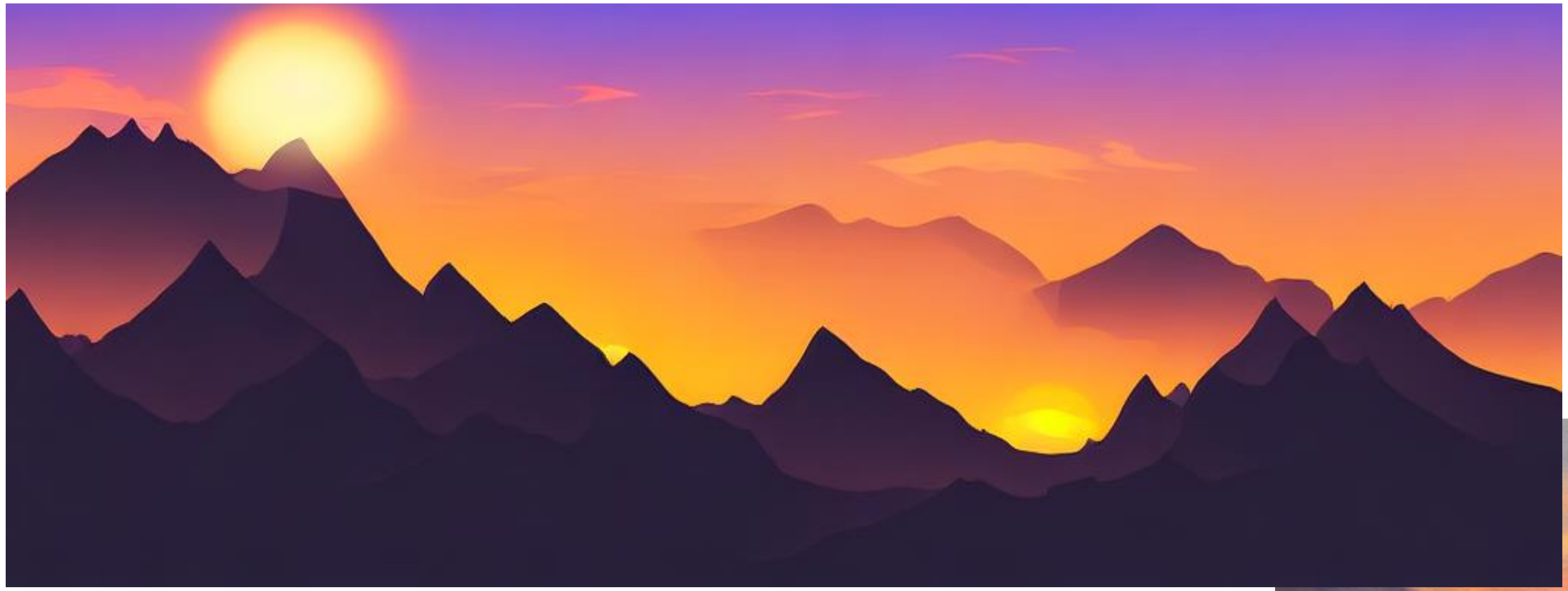


Slide credit: Robin Rombach

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"

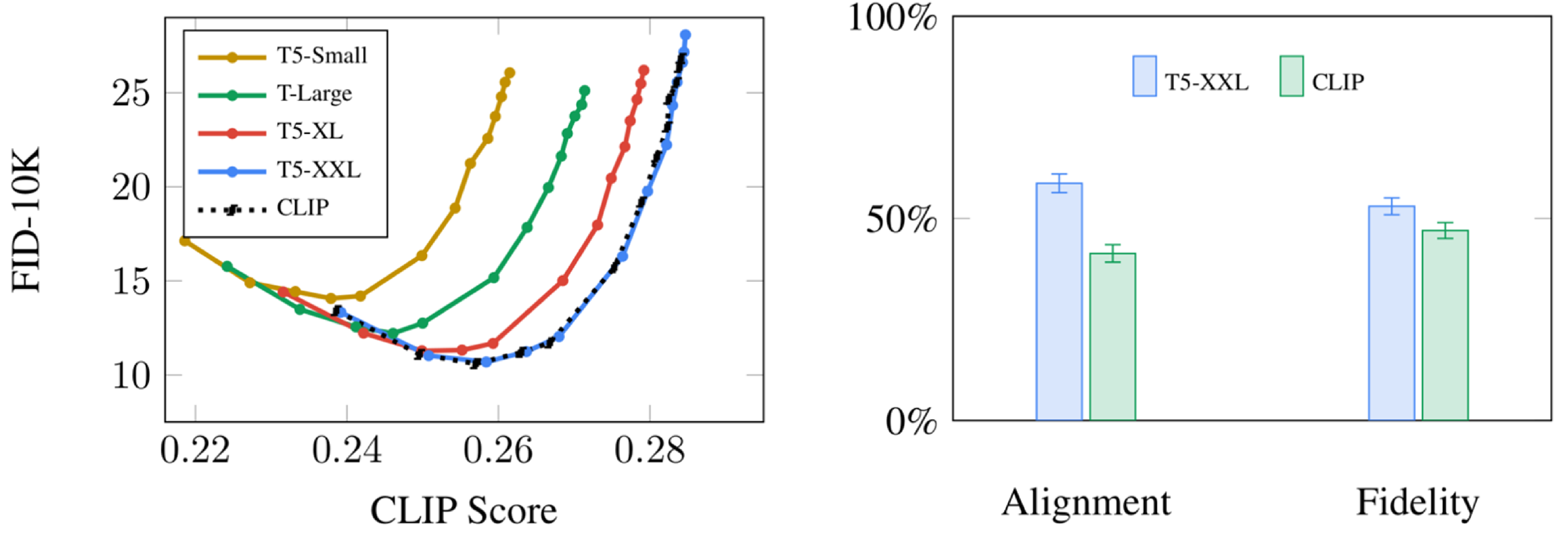
Slide credit: Robin Rombach

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



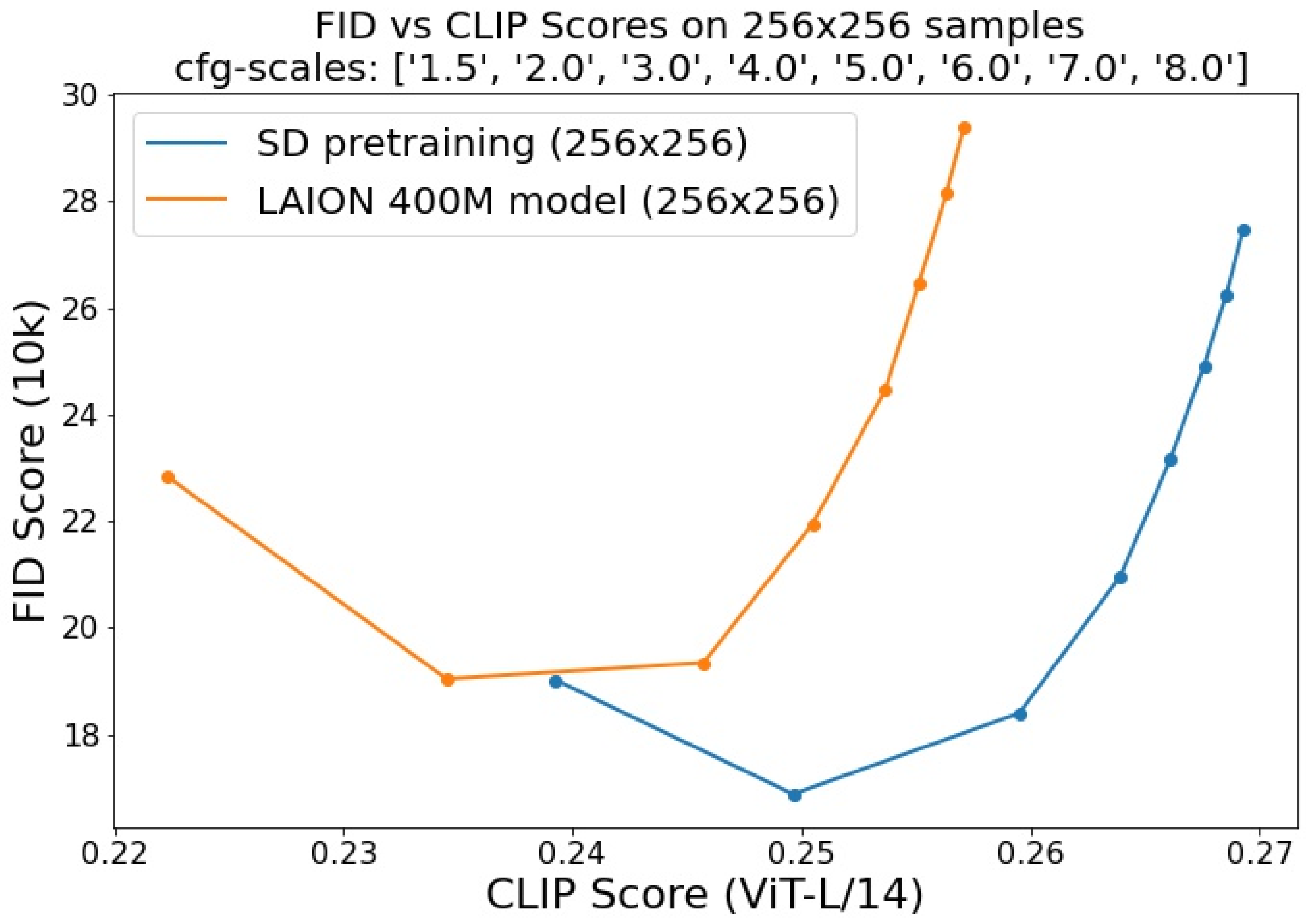
(a) Pareto curves comparing various text encoders. (b) Comparing T5-XXL and CLIP on DrawBench.

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):

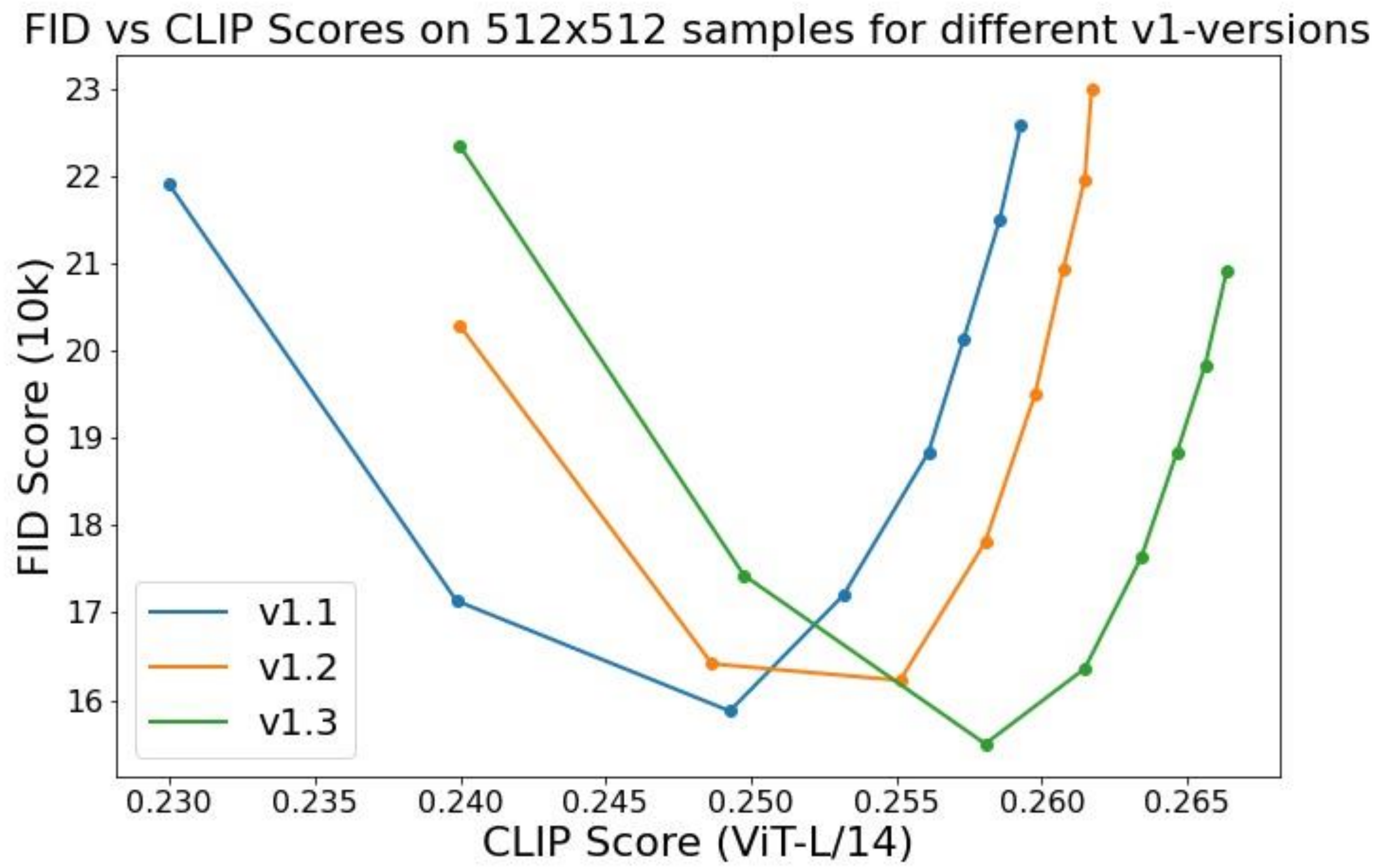
- 194k steps at resolution 512x512 on laion-high-resolution (170M examples from LAION-5B with resolution >= 1024x1024).

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 >= 512x512, 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

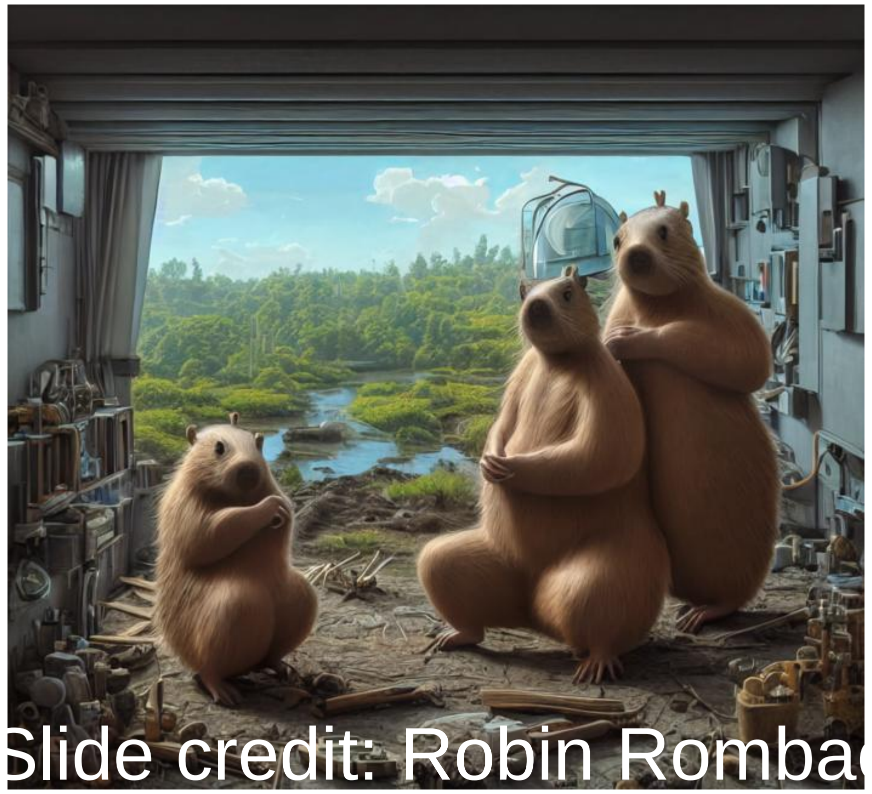


10k random COCO val captions / 50 decoding steps

→ 4.2 GB checkpoint (EMA only, fp32)



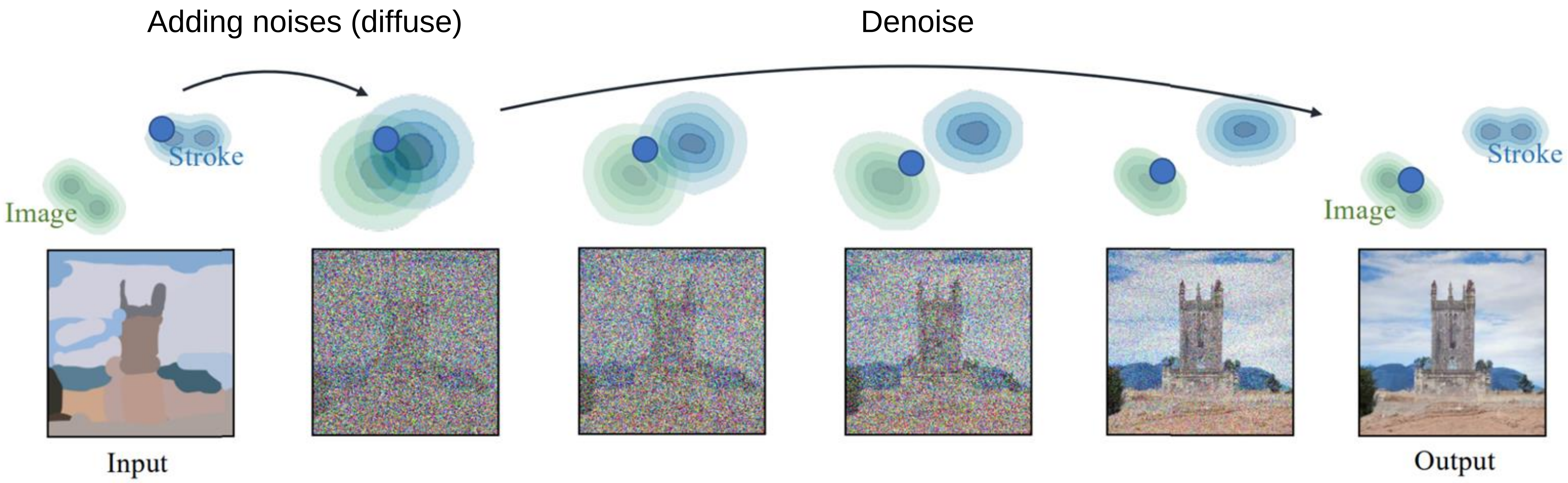





Slide credit: Robin Rombach

Text-Guided Image-to-Image

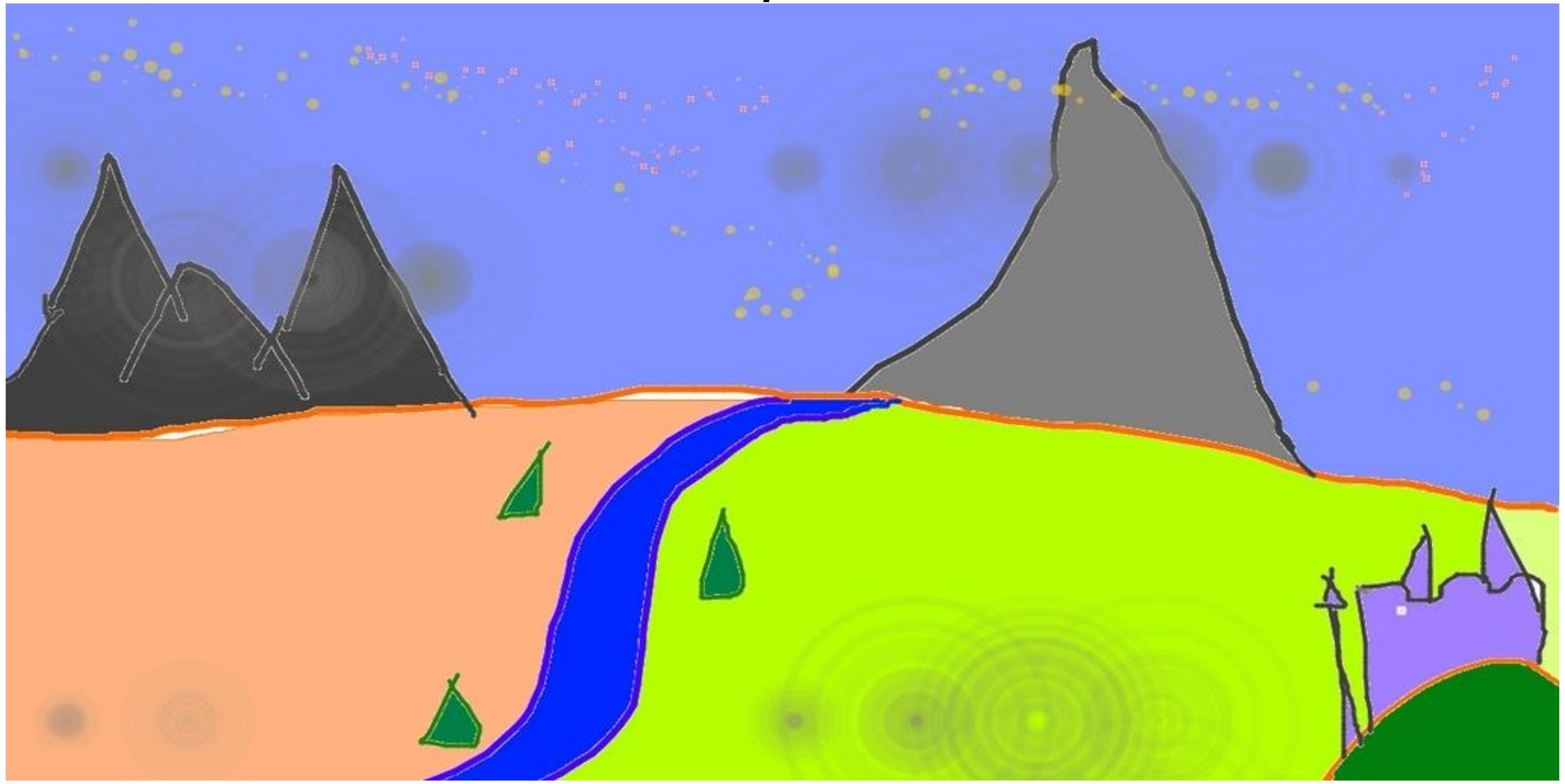
SDEdit (https://arxiv.org/abs/2108.01073) recipe: diffuse → denoise



Slide credit: Robin Rombach

Text-Guided Image-to-Image

input



Slide credit: Robin Rombach

“a fantasy landscape, watercolor painting”

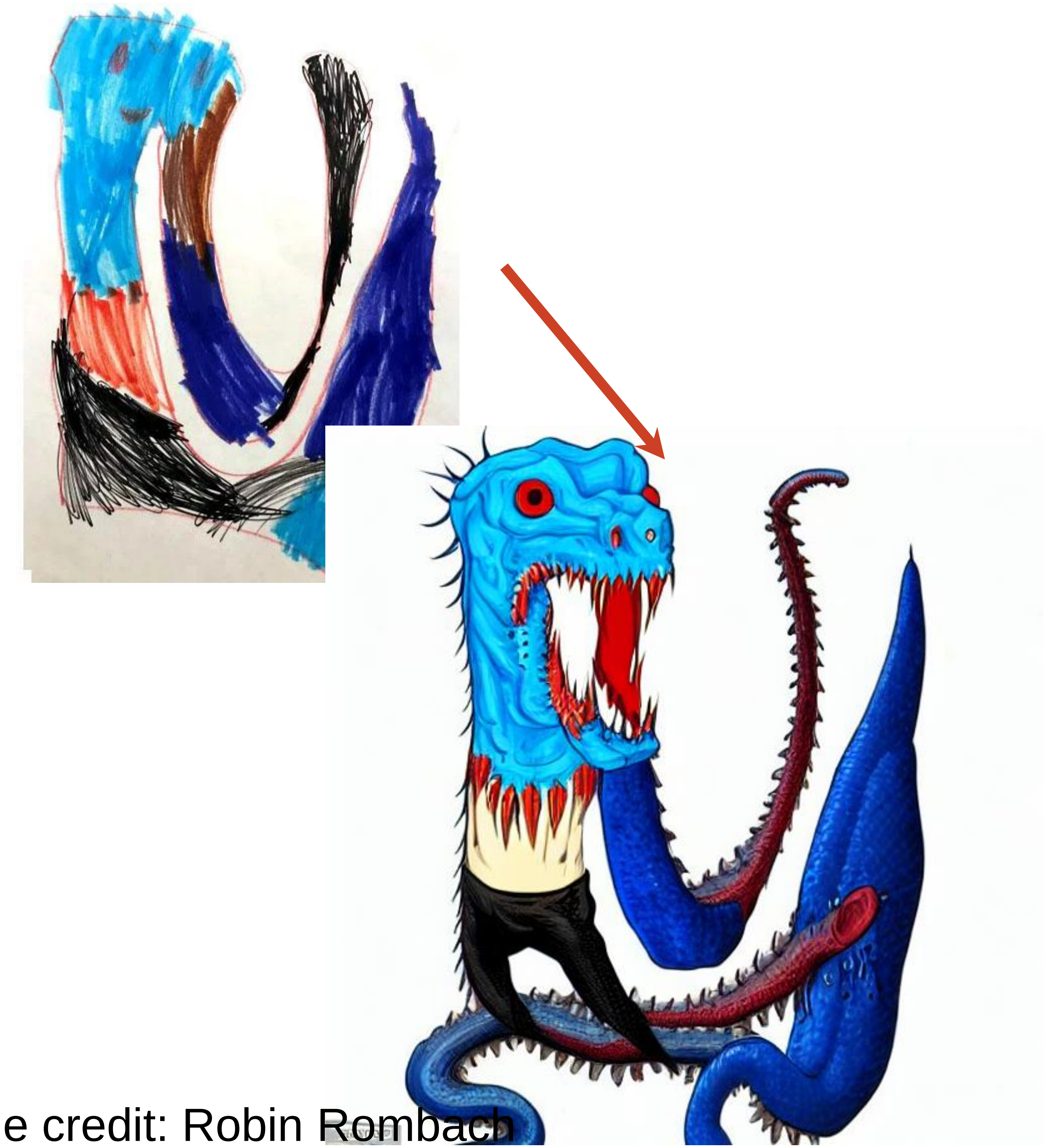
“a fantasy landscape, trending on artstation”

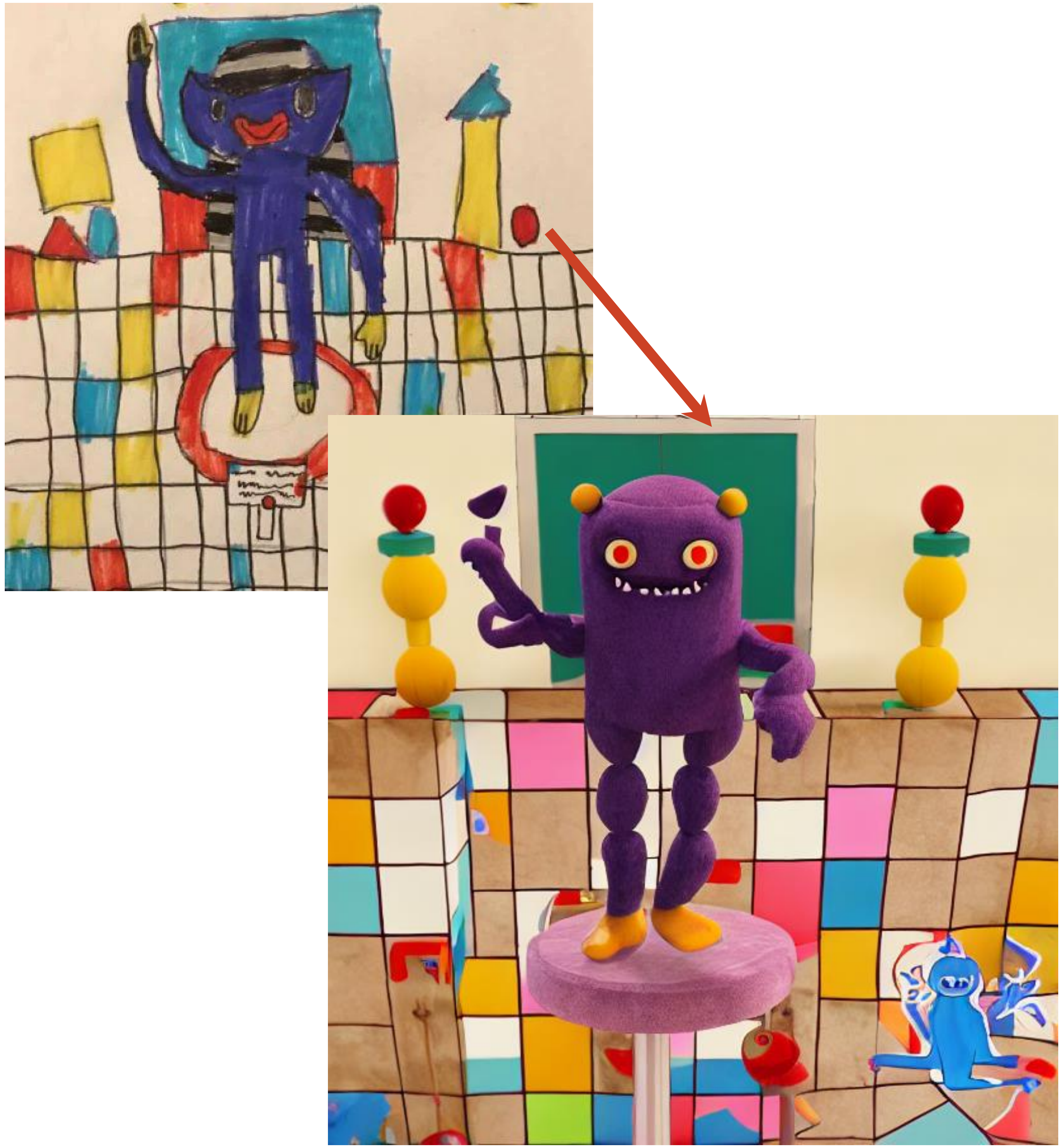
“a fantasy landscape, by Simon Stalenhag”

Slide credit: Robin Rombach

“Upgrade” your child’s artwork

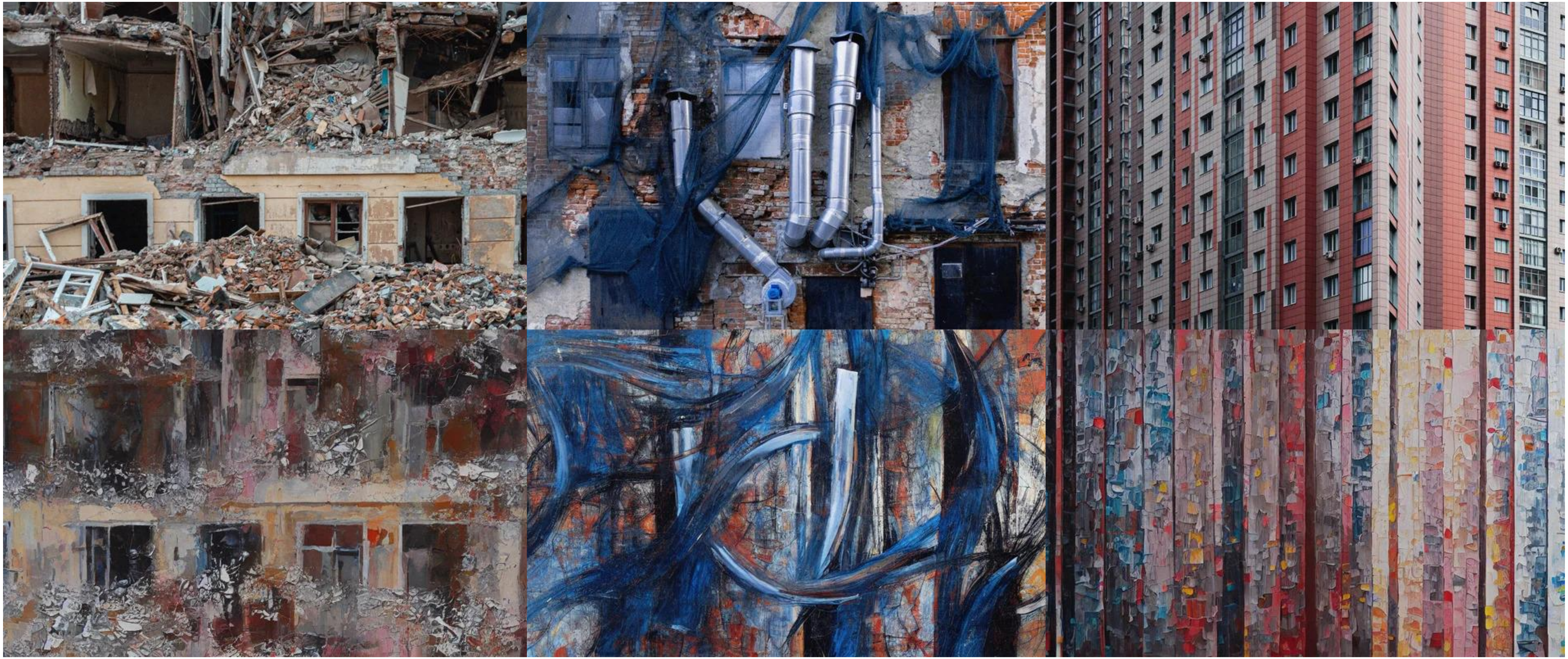
original post: https://www.reddit.com/r/StableDiffusion/comments/wyq04v/using_img2img_to_upgrade_my_sons_artwork/





Slide credit: Robin Rombach

abstract art from photos



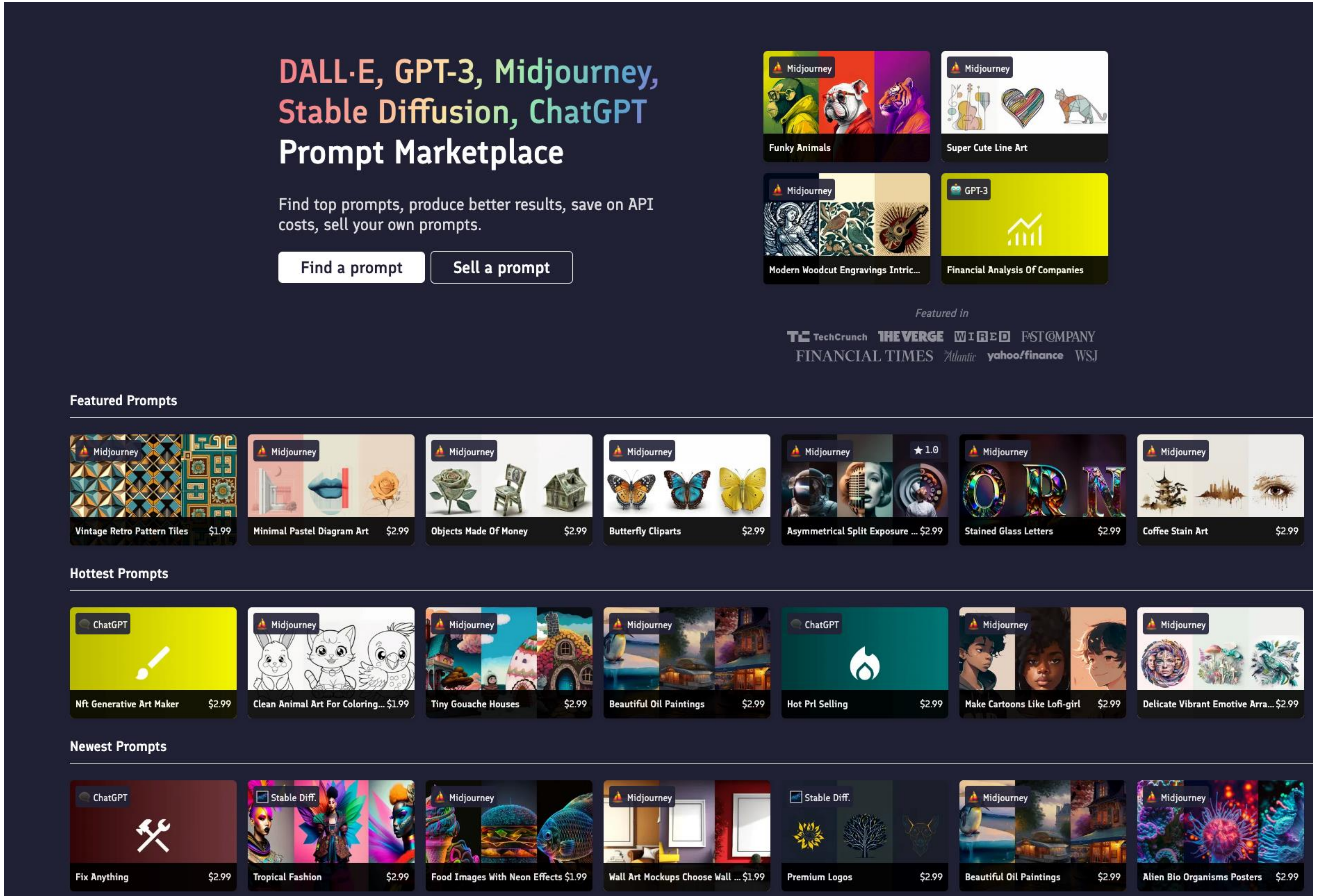
original post by u/Pereulkov
https://www.reddit.com/r/StableDiffusion/comments/xhhyad/i_made_abstract_art_from_my_photos/

Video Synthesis

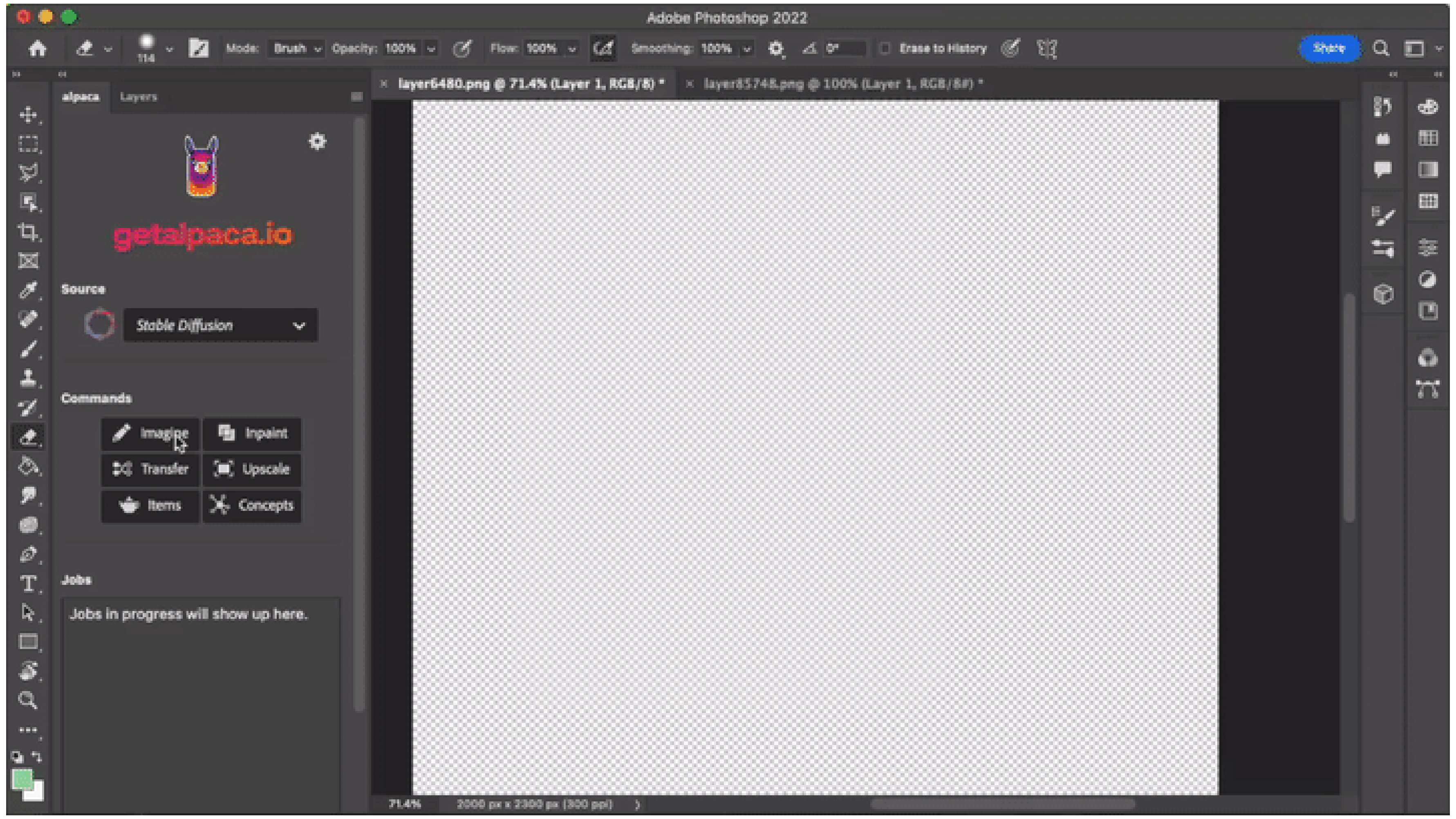
Stable Diffusion (img2img) + EBSynth by Scott Lightsier:
https://twitter.com/LighthiserScott/status/1567355079228887041?t=kXXCAVtuO5lJCGcro3Ma3A&s=19

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

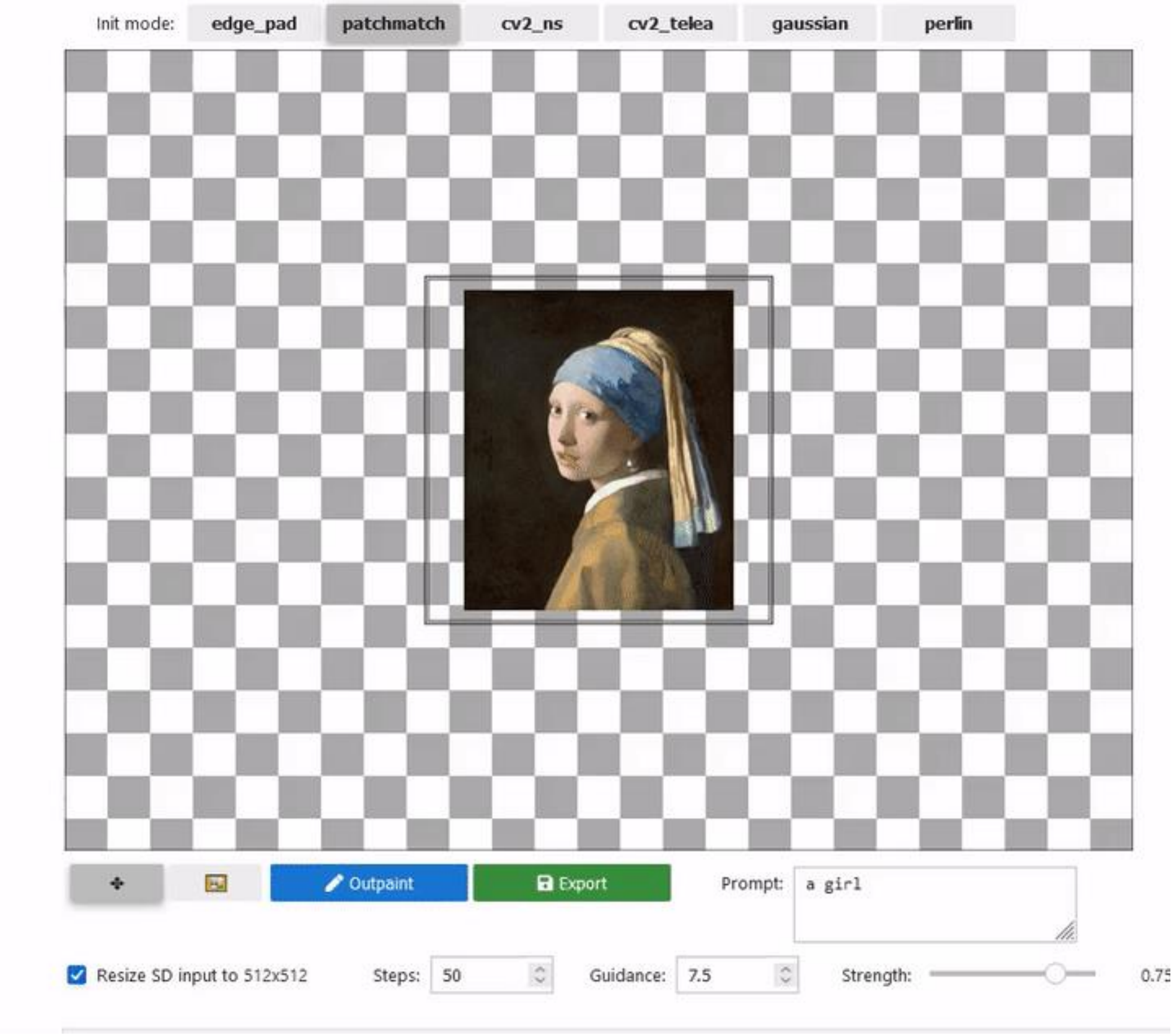
Prompt Marketplace (promptbase.com)



UIs / Plug-Ins for Photoshop, GIMP etc



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



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

Slide credit: Robin Rombach

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

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 Butter-fly 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

But what about ...

Evaluation?
Robustness ?
Reasoning Abilities?
Efficiency?

Do T2I Models Generate Accurate Spatial Relationships?

A chair above a knife

above(chair, knife)

A teddy bear below a bed

below(teddy bear, bed)

A fork to the left of a carrot

left(fork, carrot)

A person to the right of a truck

right(person, truck)

Benchmarking Spatial Relationships in Text-to-Image Generation

Tejas Gokhale [1*] Hamid Palangi [2] Besmira Nushi [2] Vibhav Vineet [2]
Eric Horvitz [2] Ece Kamar [2] Chitta Baral [1] Yezhou Yang [1]
[1] Arizona State University [2] Microsoft Research

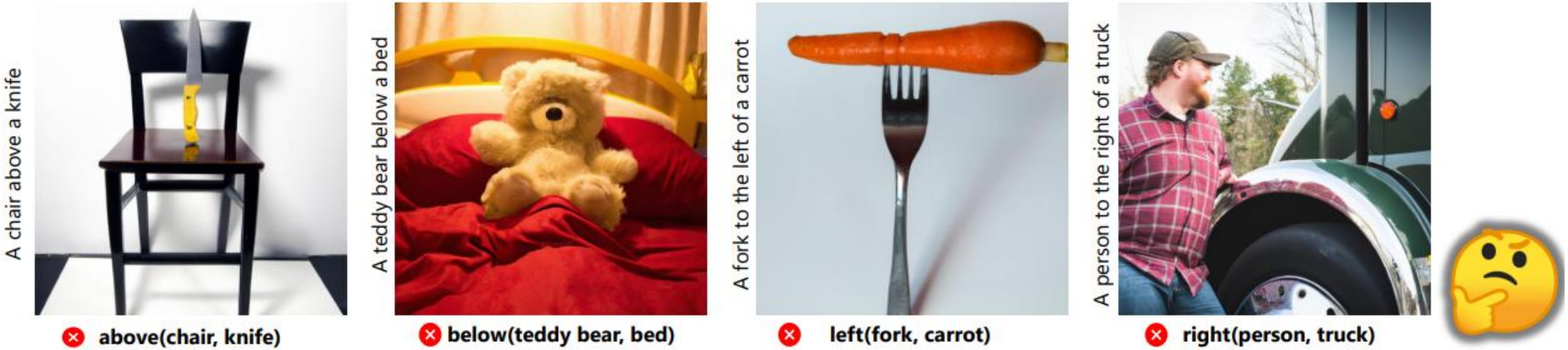
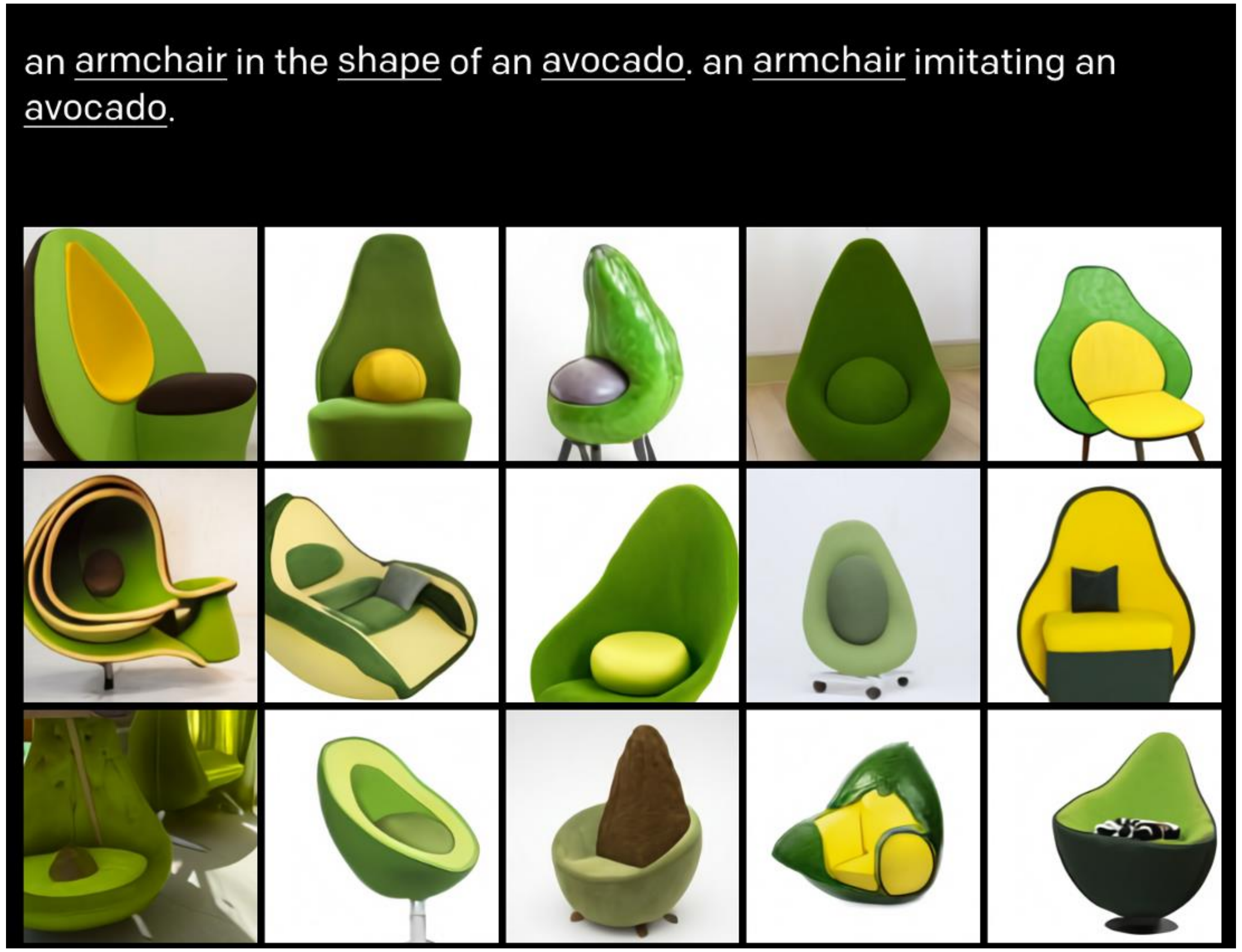


Figure 1: We benchmark T2I models on their competency with generating appropriate spatial relationships in their visual renderings. Although text inputs may explicitly mention these spatial relationships, T2I models lack such spatial understanding.

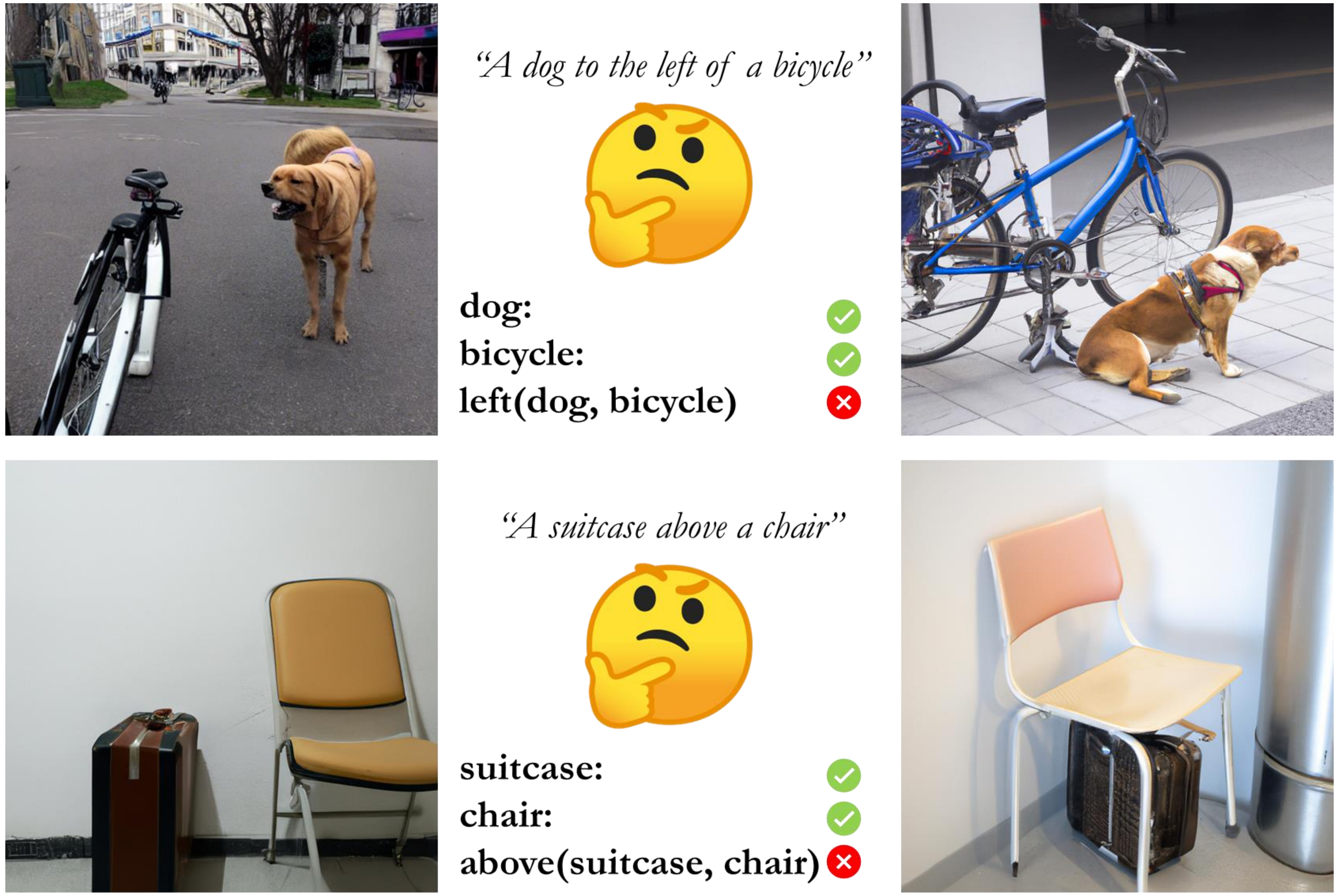
visort2i.github.io https://github.com/microsoft/VISOR

VISOR reveals the ineffectiveness of T2I models in generating multiple objects with correct spatial relationships.

Attribute-Level Compositionality
Compositionality



Object-Level / Spatial



Follow-up (Method to Improve Spatial Reasoning in T2I)

Getting it *Right*: Improving Spatial Consistency in Text-to-Image Models

Agneet Chatterjee[1,*,‡], Gabriela Ben Melech Stan[2,*], Estelle Aflalo[2], Sayak Paul[3], Dhruba Ghosh[4], Tejas Gokhale[5], Ludwig Schmidt[4], Hannaneh Hajishirzi[4], Vasudev Lal[2], Chitta Baral[1], Yezhou Yang[1]

[1]Arizona State University, [2]Intel Labs, [3]Hugging Face, [4]University of Washington
[5]University of Maryland, Baltimore County

A giraffe to the **right** of a truck.

A hair drier to the **right** of a wine glass.

A cozy cabin nestled in the woods, with a stream flowing in **front** and a fire burning in the fireplace **inside**.

A cat sitting **on** a chair with a lamp to the **right** and a window **above**, casting shadows on the floor below.

ConceptBed (AAAI 2024)

ConceptBed: Evaluating Concept Learning Abilities of Text-to-Image Diffusion Models

Maitreya Patel[1*], Tejas Gokhale[2], Chitta Baral[1], Yezhou Yang[1]

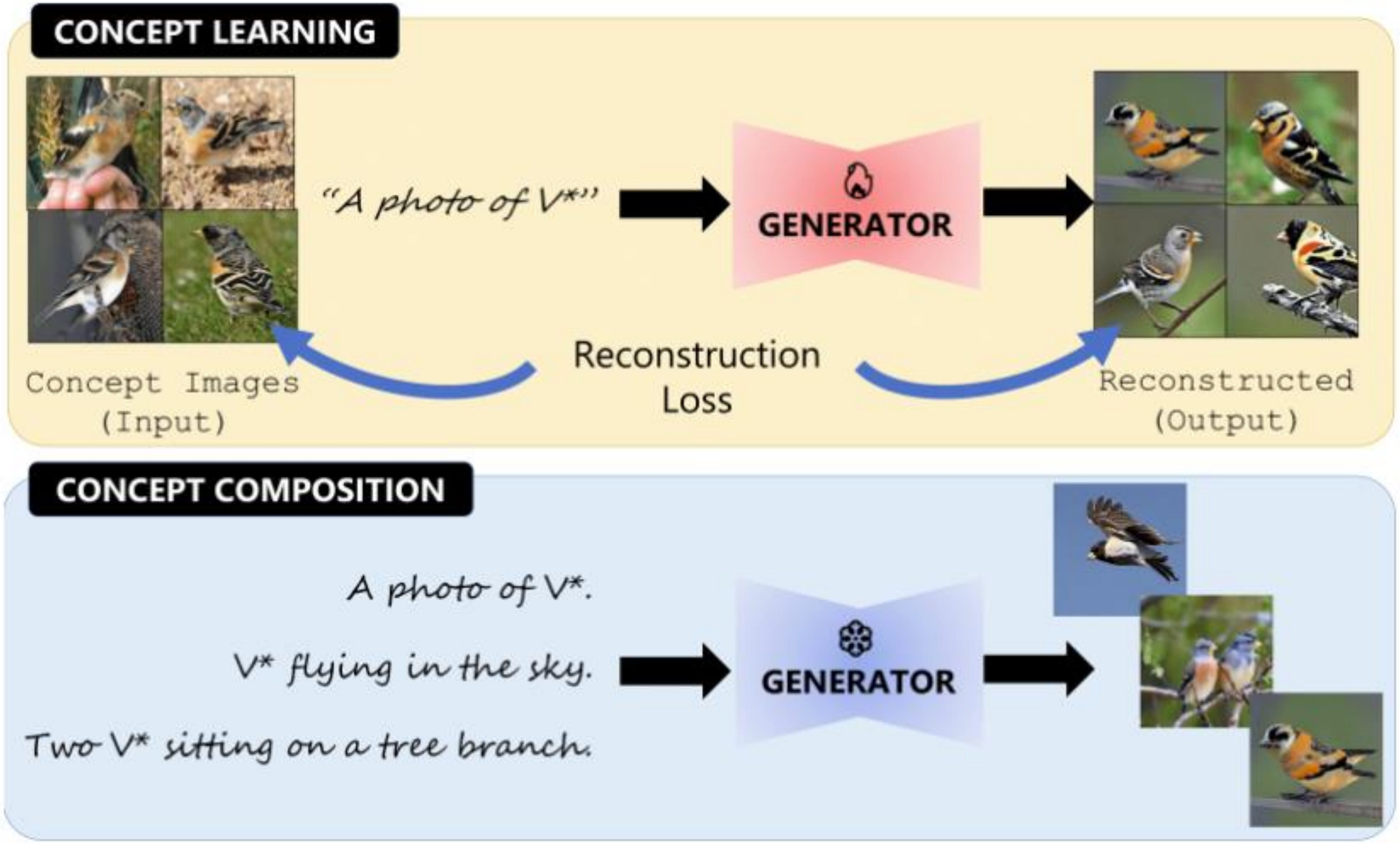
[1] Arizona State University
[2] University of Maryland Baltimore County

ConceptBed

Evaluating Concept Learning Abilities of Text-to-Image Diffusion Models

Workflow:

- Textual inversion models learn visual concepts from a few examples.
- These concepts "V*" are stored as text embeddings.
- T2I models use the new concepts in novel compositions

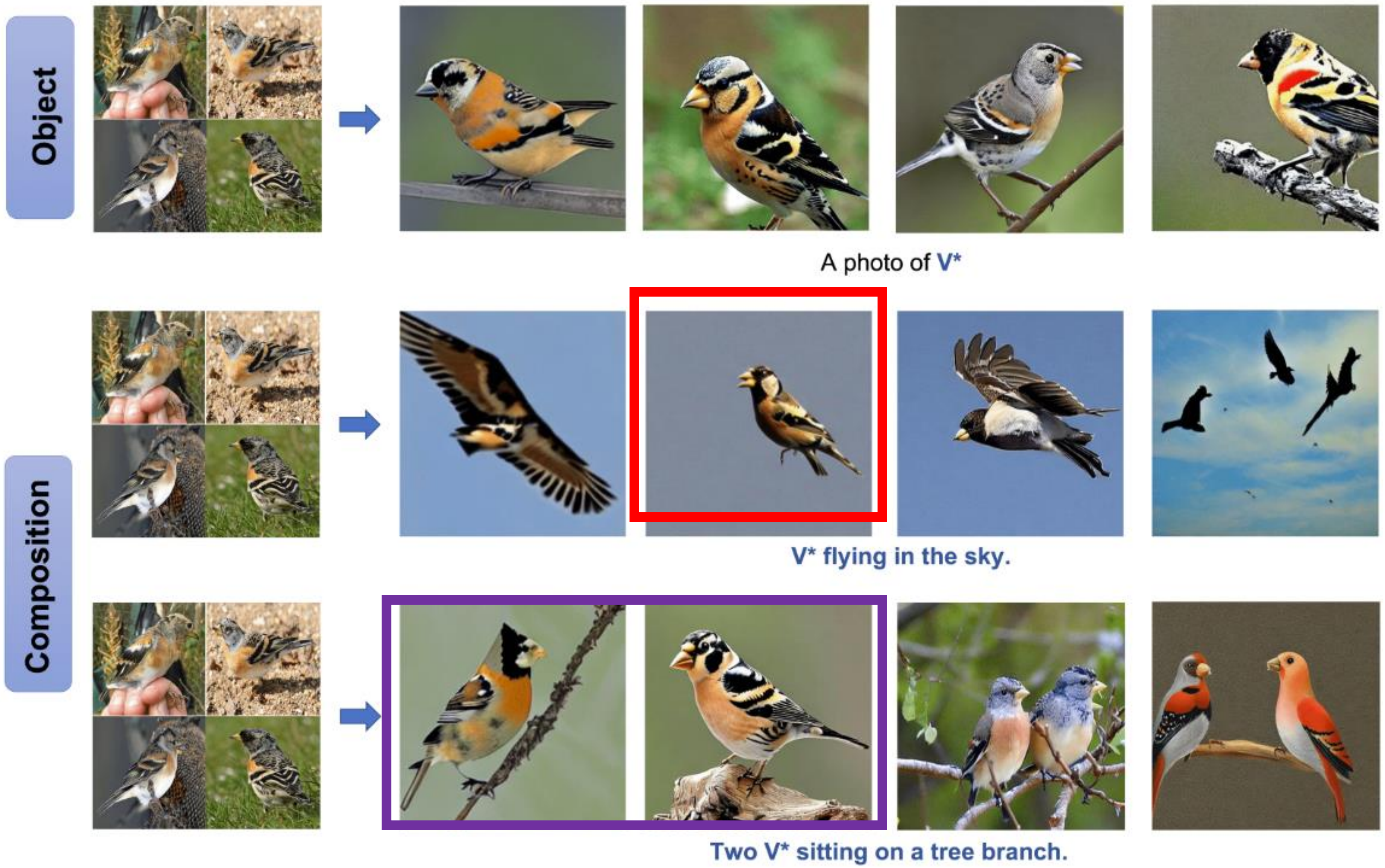


ConceptBed

Evaluating Concept Learning Abilities of Text-to-Image Diffusion Models

Findings:

- Compositionality is hard!
- "flying"
  - where are the wings?
  - Would a bird float with that pose?
- Counting …



- Dataset

- Evaluation Metric:
  "Concept Confidence Deviation"

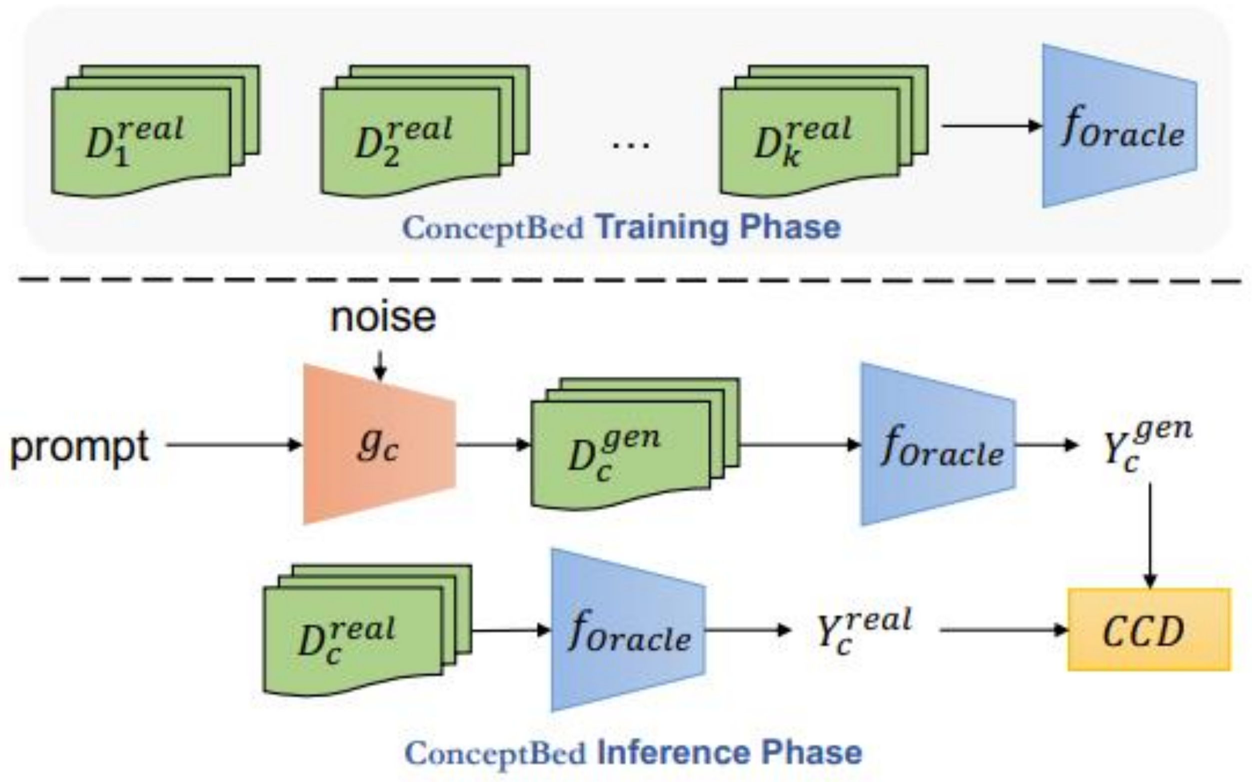
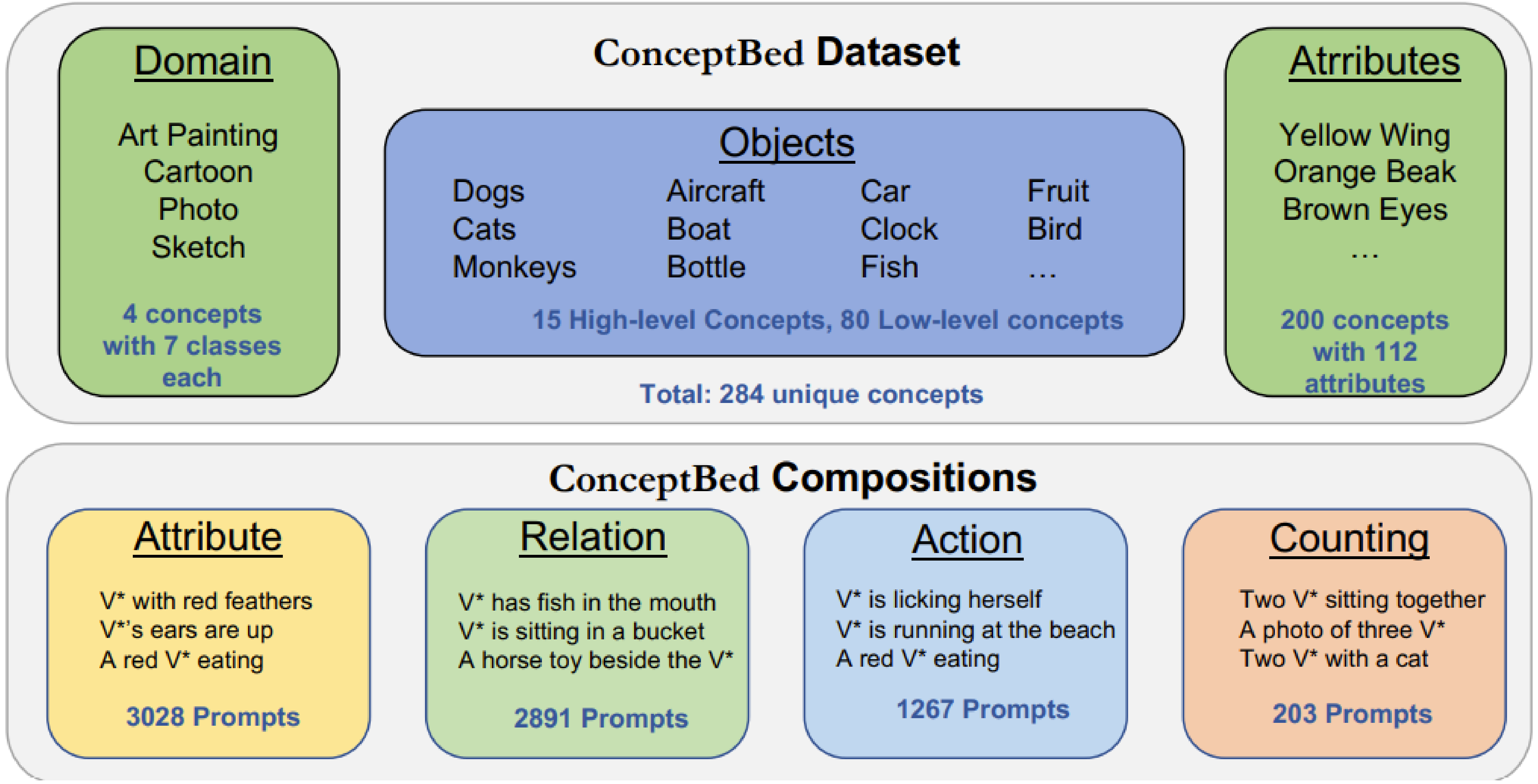
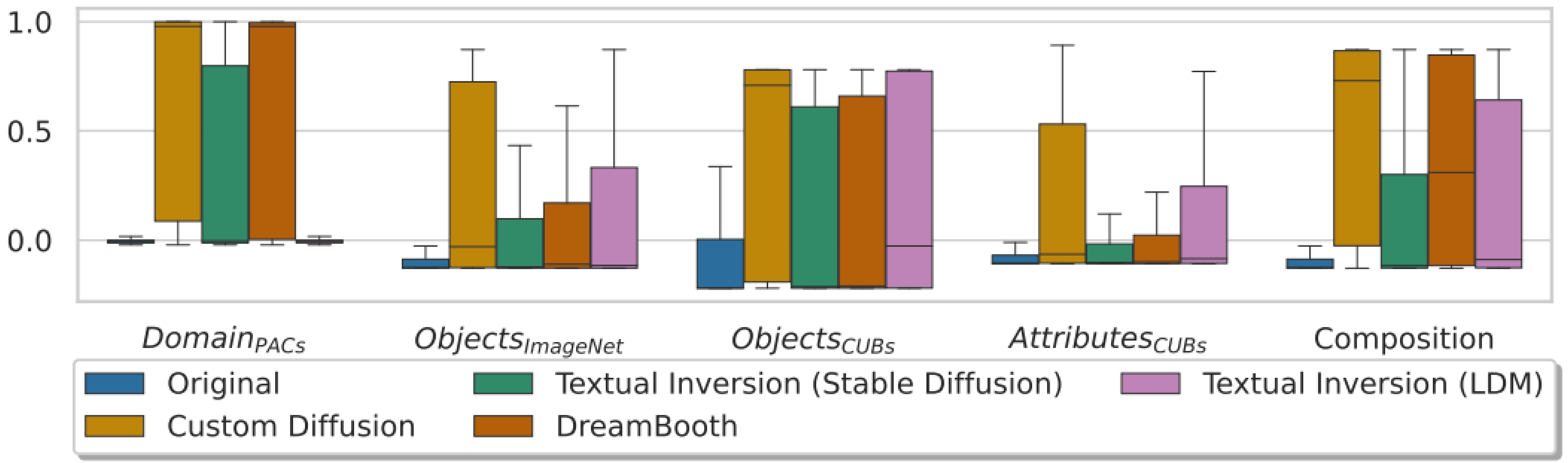
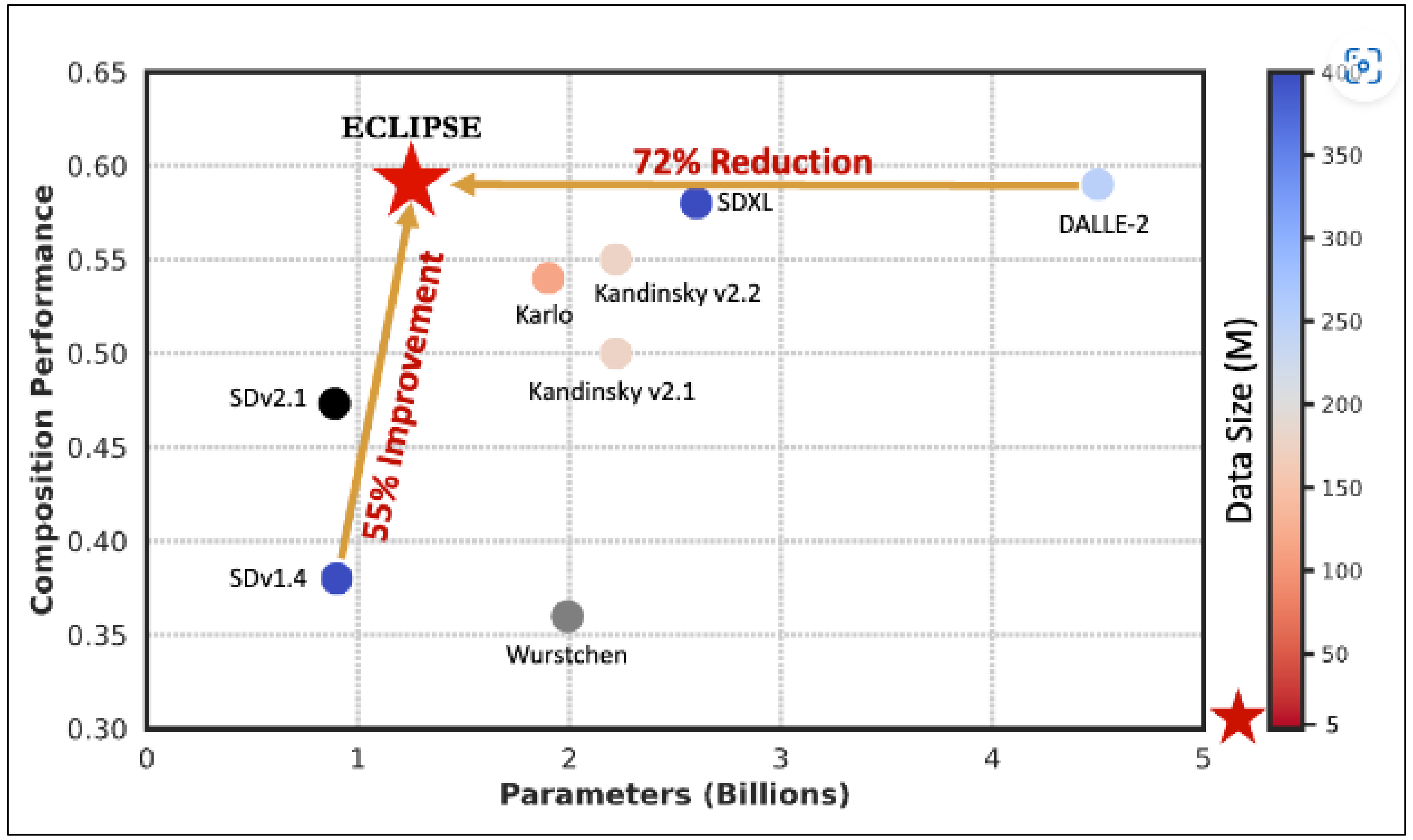


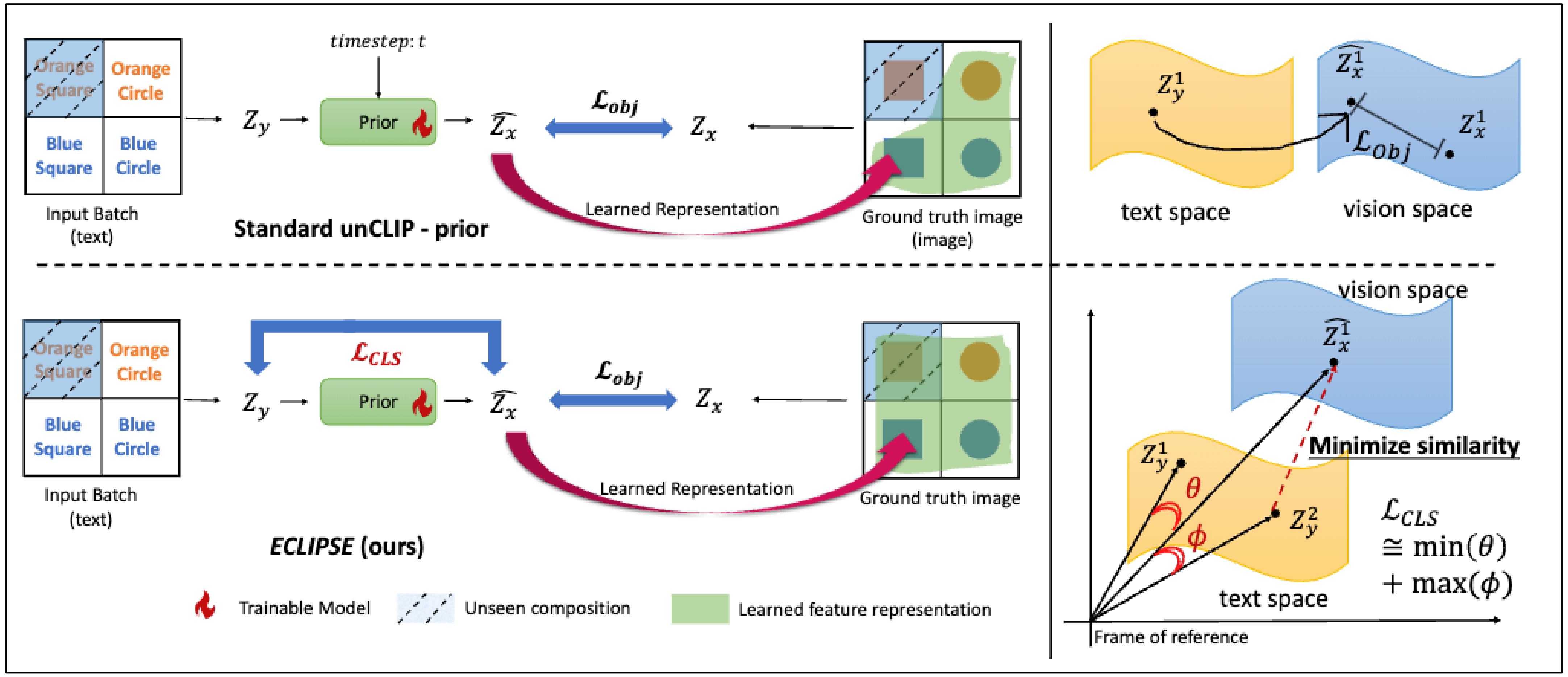
Figure 3: The outline of CONCEPTBED evaluation framework.





ECLIPSE: Resource-Efficient T2I Prior (CVPR'24)





- ***ECLIPSE*** leverages pre-trained vision-language models (e.g., CLIP) to distill the knowledge into the prior model.
- CLIP Contrastive Learning is enough to achieve state-of-the-art text-to-image prior ***without the diffusion process.***
- This allows training model with only 33M parameters and 0.6M image-text pairs.