GenAI in 2024

Overview of recent advances in generative AI for vision

Presented by Kanchana Ranasinghe | June 2024

Acknowledgements

- Content borrowed from:
  https://cvpr2023-tutorial-diffusion-models.github.io
  https://cvpr2022-tutorial-diffusion-models.github.io

- Lazebnik Slides:
  https://slazebni.cs.illinois.edu/spring17/lec11_gan.pdf

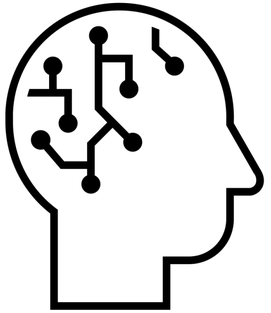
Overview

Deep Generative Learning

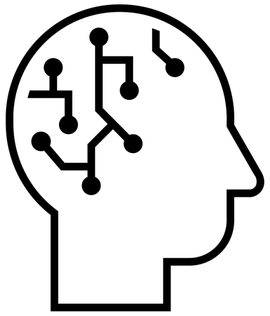
Learning to generate data

Samples from a Data Distribution

Train



Neural Network



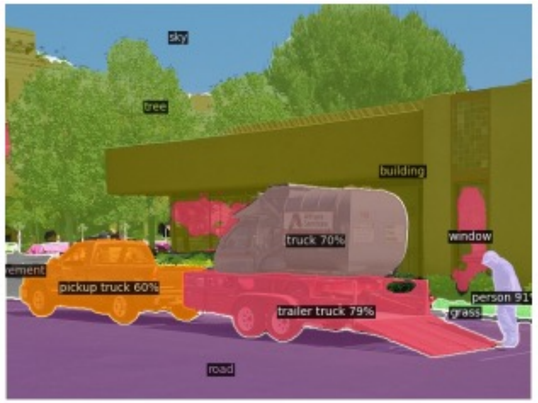
Sample

Generative Learning Applications

Art & Design

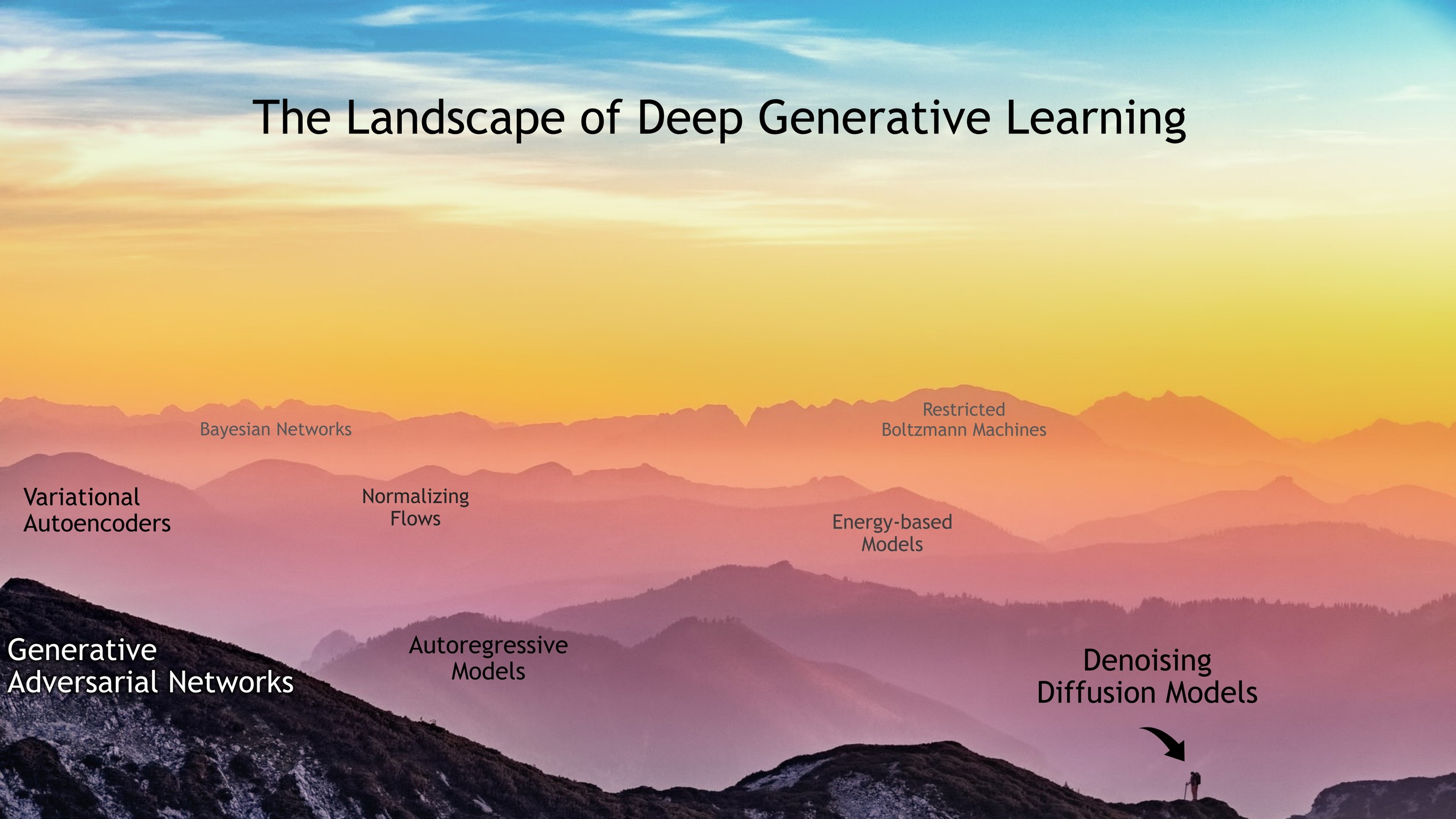
Content Generation

Representation Learning

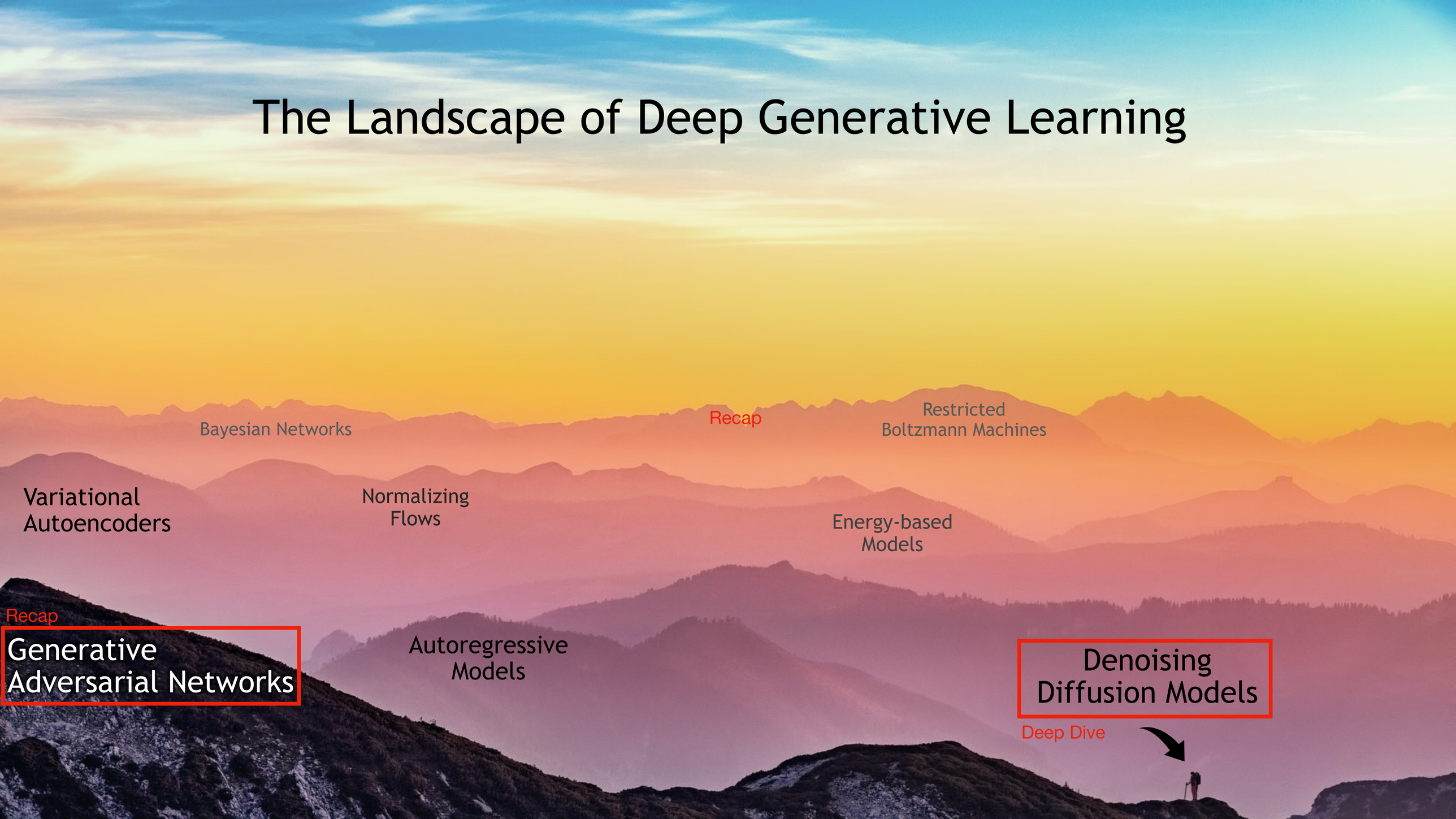


Entertainment

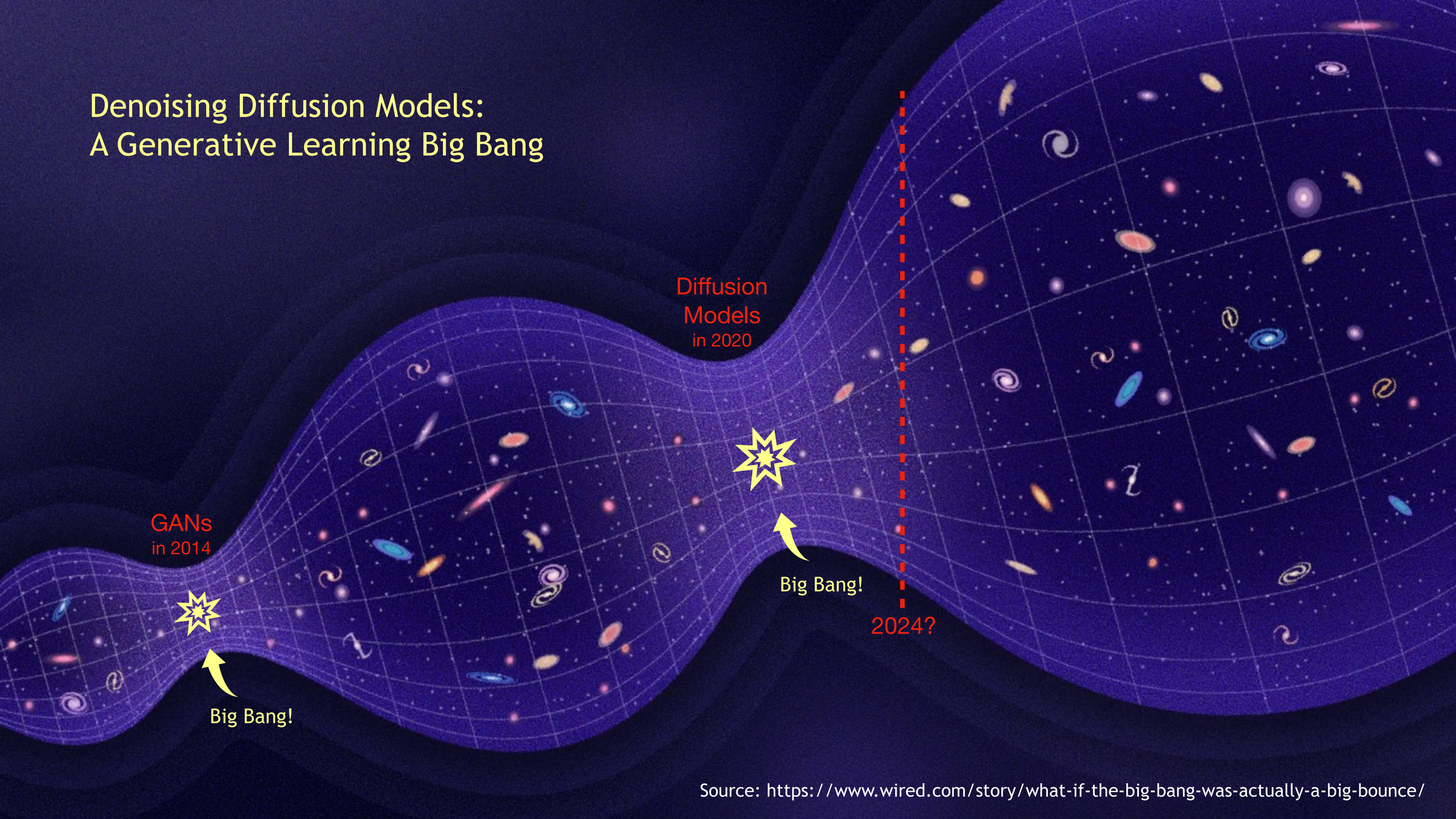
The Landscape of Deep Generative Learning



The Landscape of Deep Generative Learning

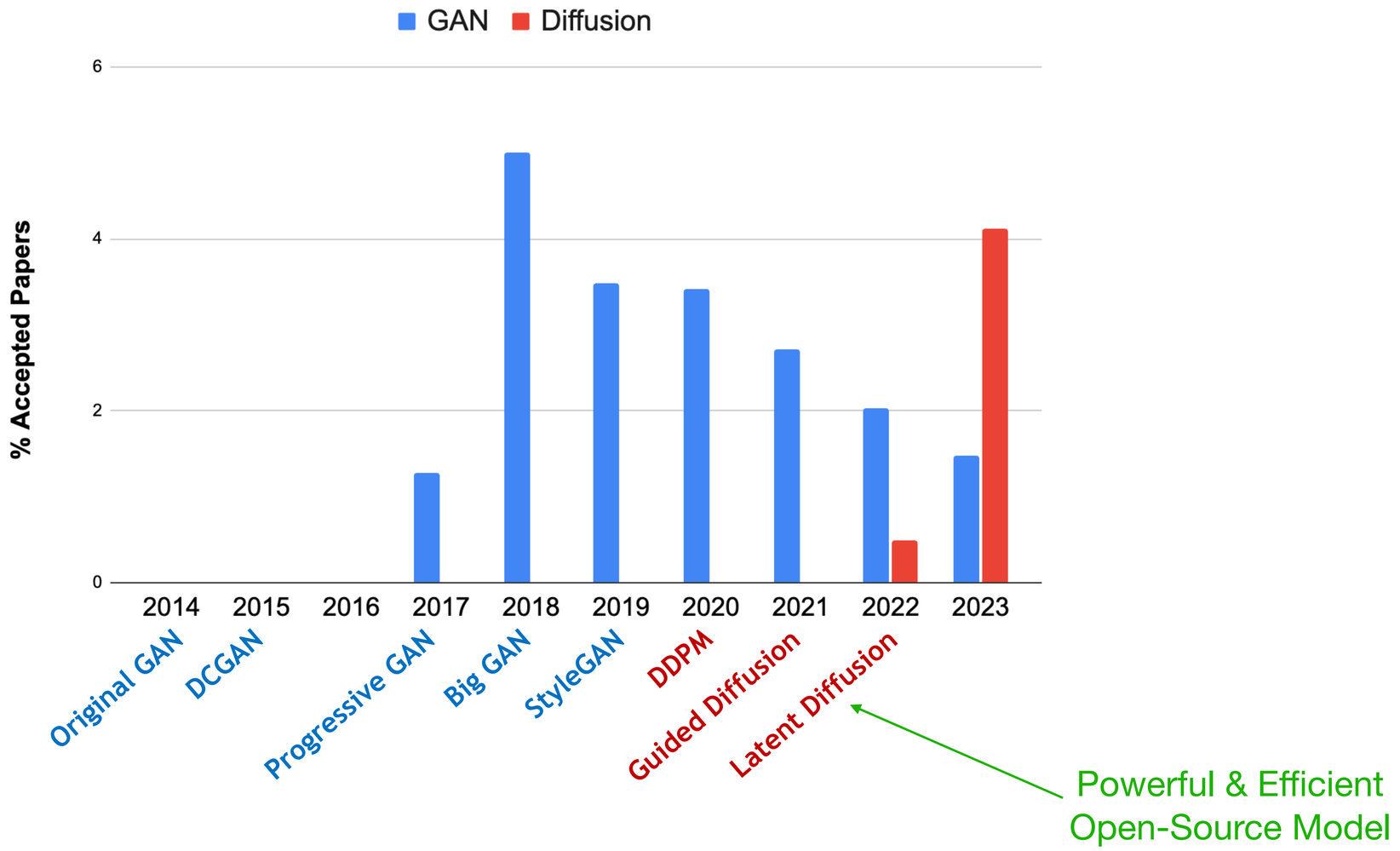


Denoising Diffusion Models: A Generative Learning Big Bang



Source: https://www.wired.com/story/what-if-the-big-bang-was-actually-a-big-bounce/

We May Not Know Cosmology, But We Know CVPR

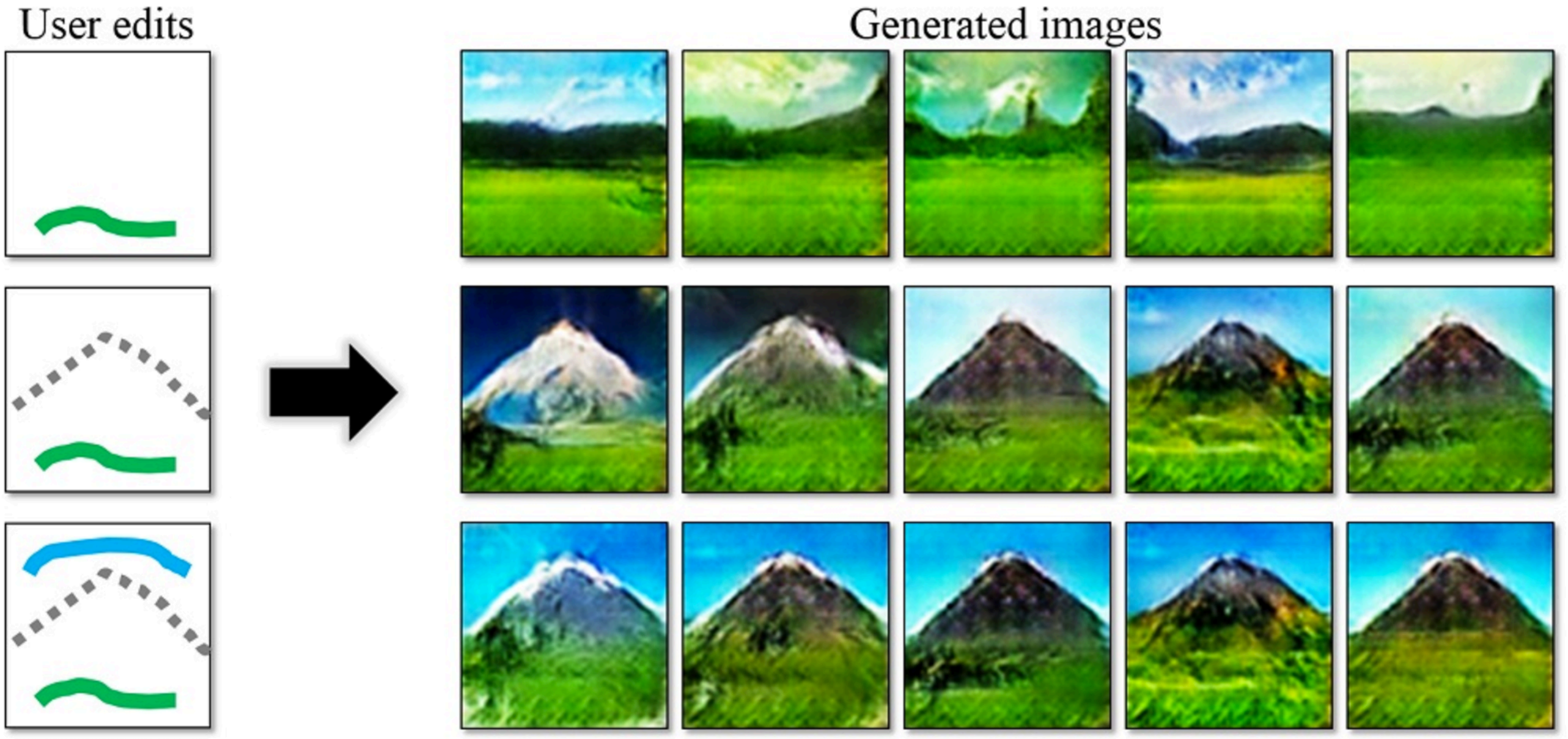


Recap: GANs

Which one is Computer generated?

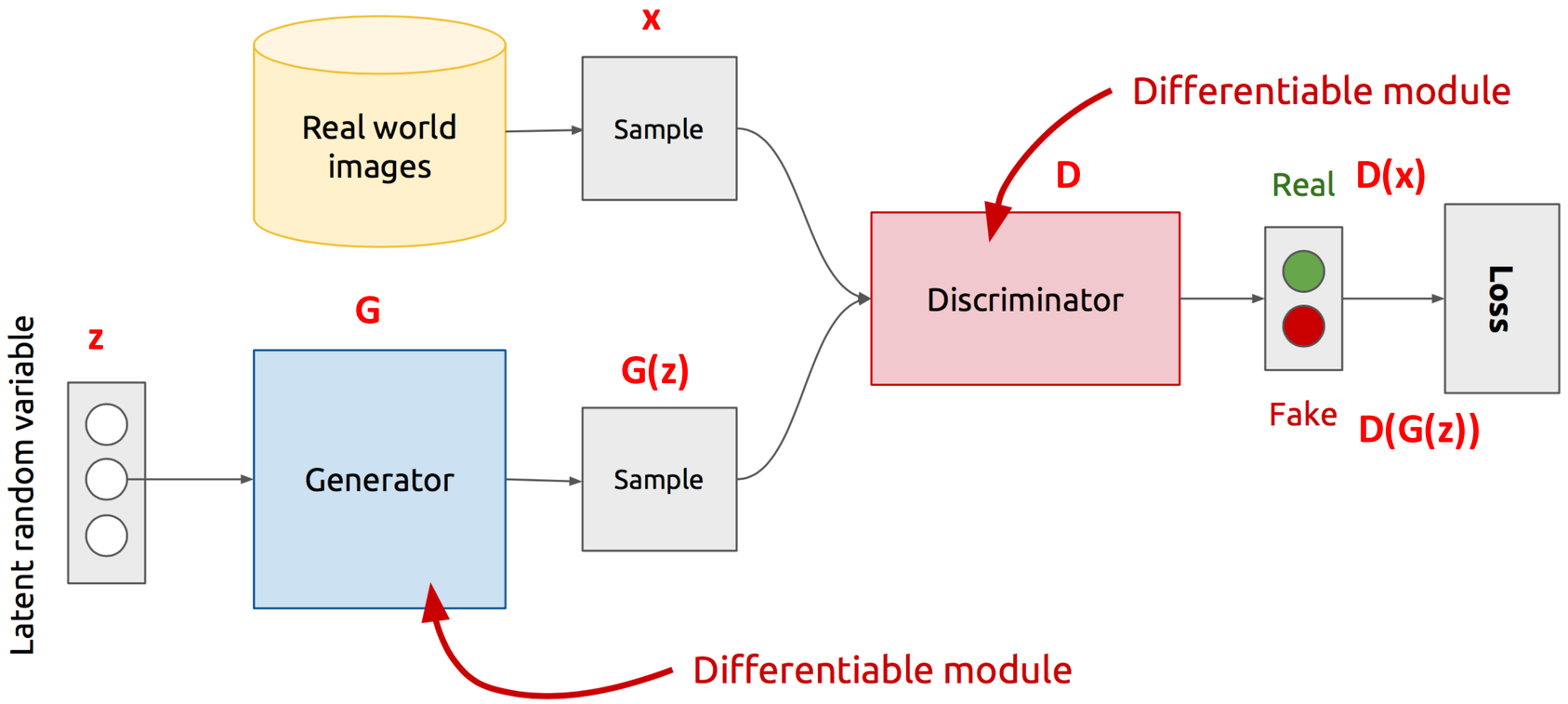
Ledig, Christian, et al. "Photo-realistic single image super-resolution using a generative adversarial network." *arXiv preprint arXiv:1609.04802* (2016).

Magic of GANs…



http://people.eecs.berkeley.edu/~junyanz/projects/gvm/

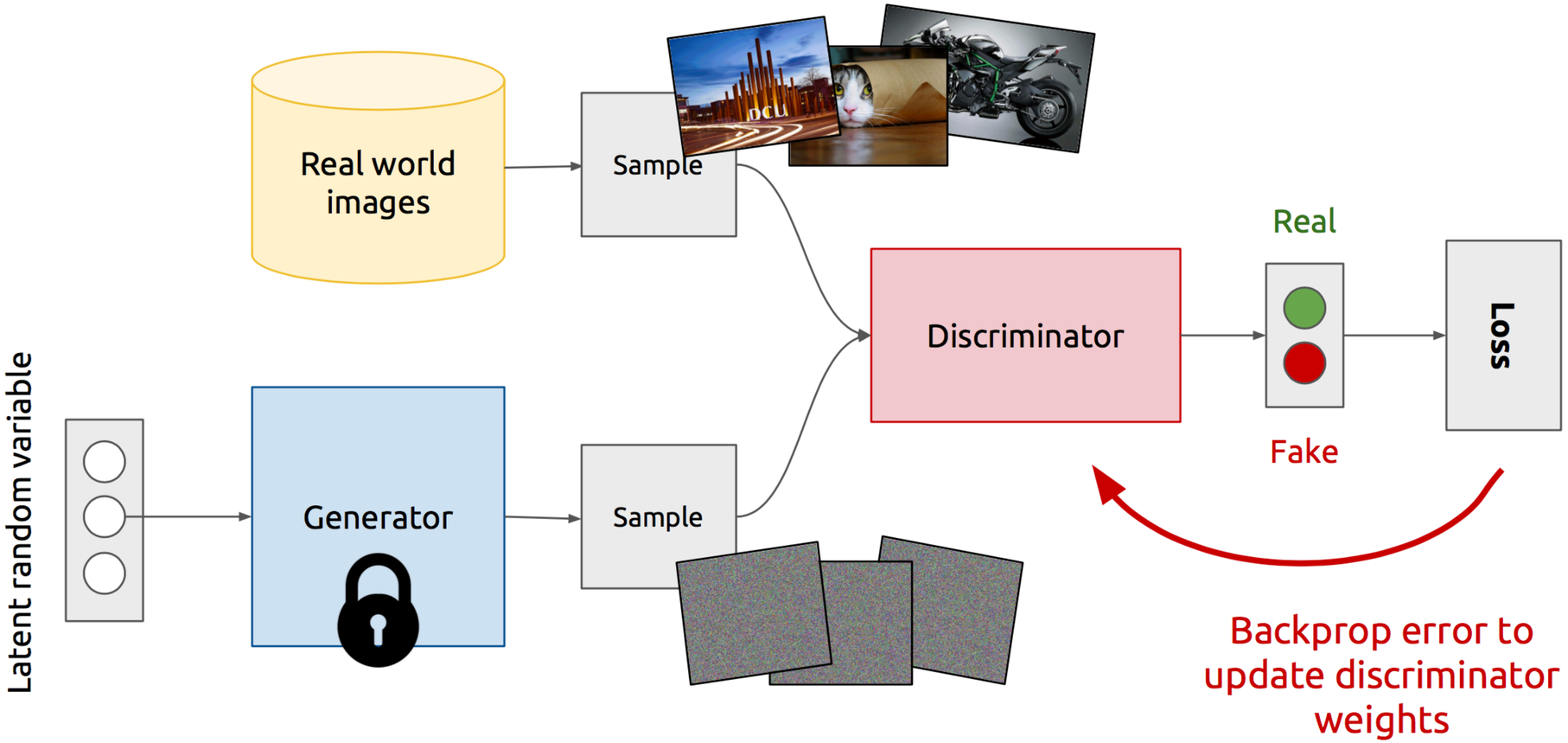
GAN's Architecture



- **Z** is some random noise (Gaussian/Uniform).
- **Z** can be thought as the latent representation of the image.

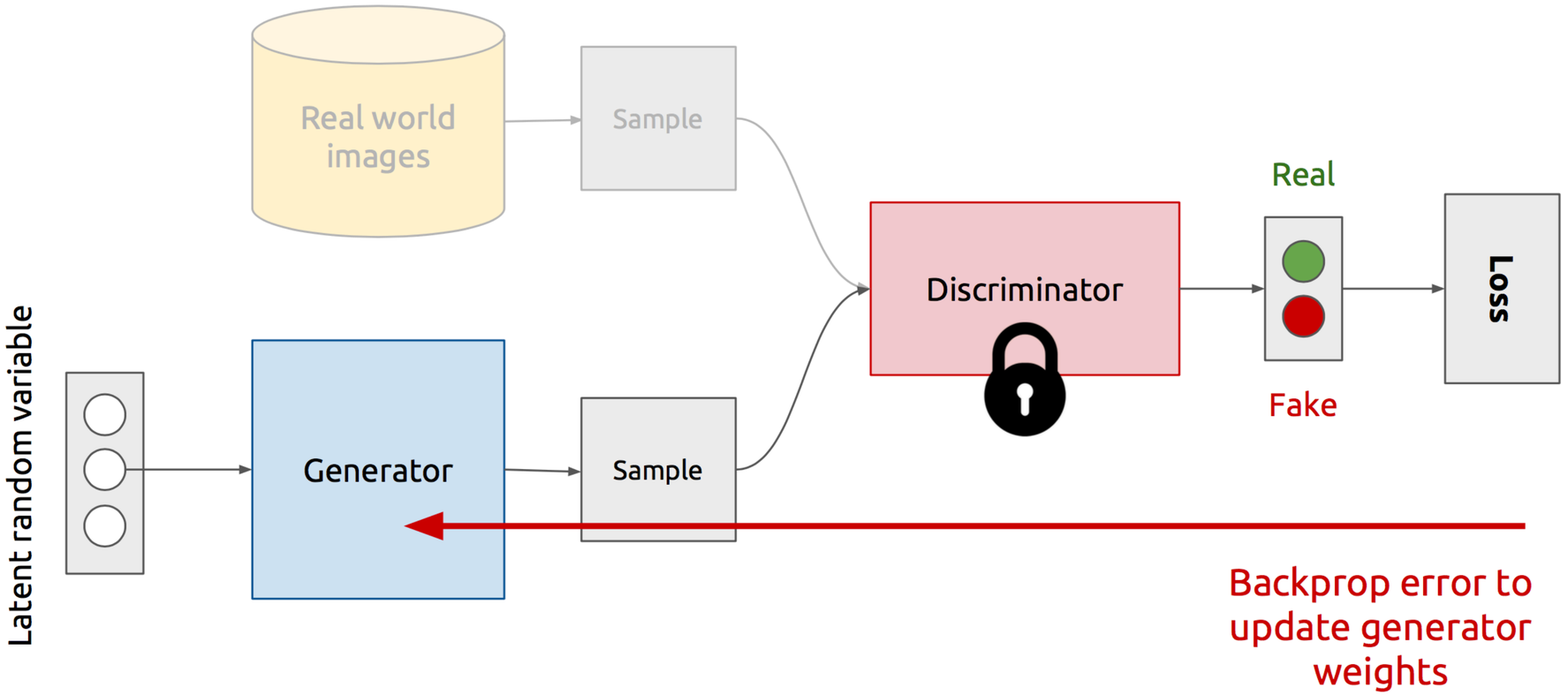
https://www.slideshare.net/xavigiro/deep-learning-for-computer-vision-generative-models-and-adversarial-training-upc-2016

Training Discriminator



https://www.slideshare.net/xavigiro/deep-learning-for-computer-vision-generative-models-and-adversarial-training-upc-2016

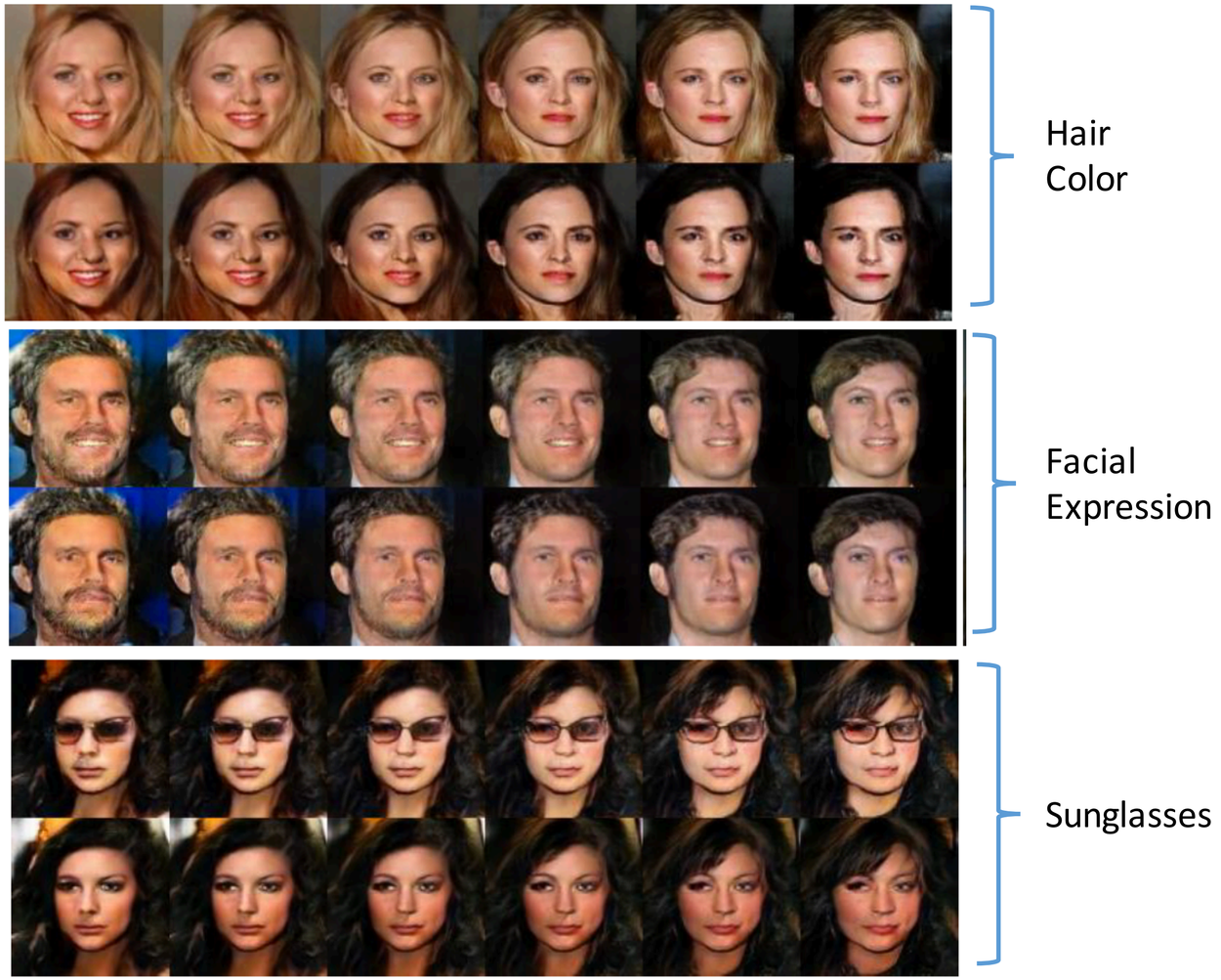
Training Generator



https://www.slideshare.net/xavigiro/deep-learning-for-computer-vision-generative-models-and-adversarial-training-upc-2016

GANs in Action

Generating Custom Images



GANs in Action

Image-to-Image Translation

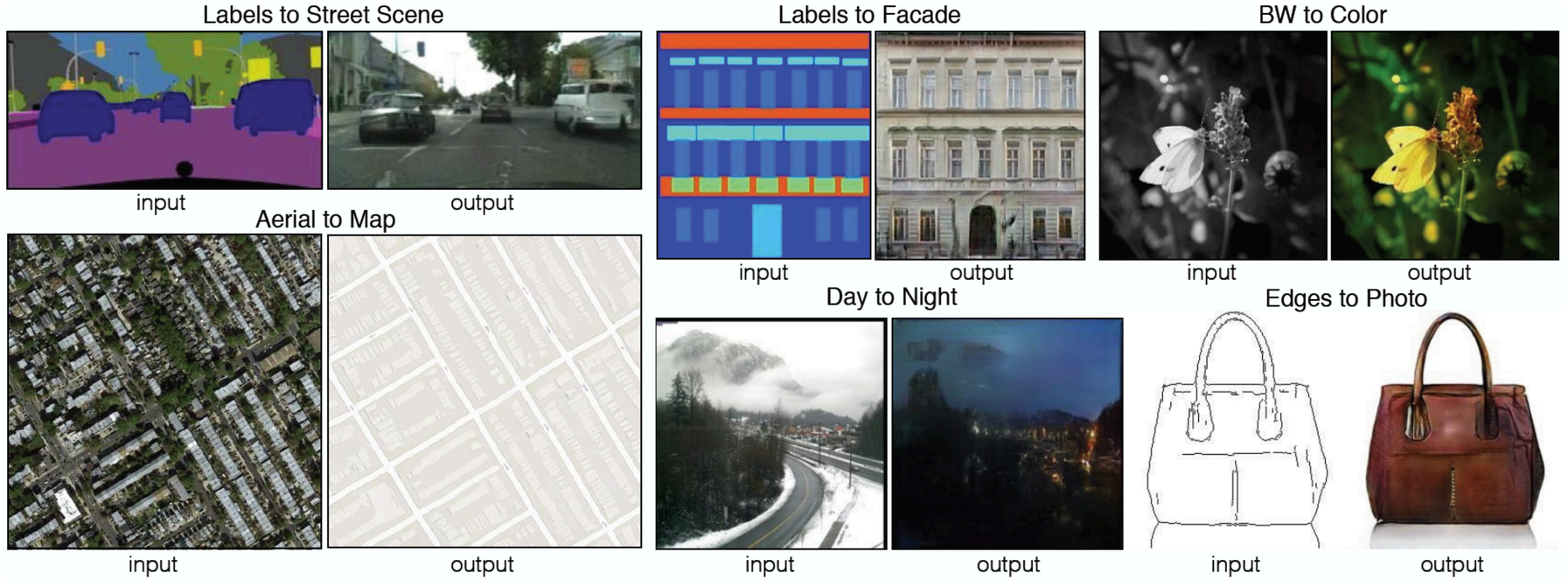


Figure 1 in the original paper.

[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).

GANs in Action

Text-to-Image Synthesis

Motivation

Given a text description, generate images closely associated.

Uses a conditional GAN with the generator and discriminator being condition on “dense” text embedding.

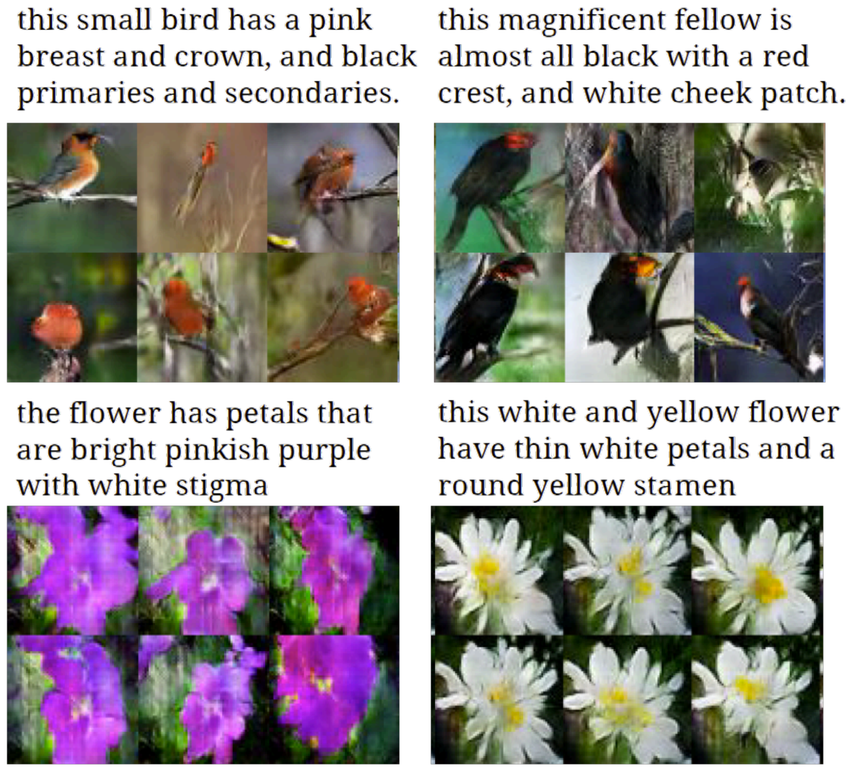


Figure 1 in the original paper.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. “**Generative adversarial text to image synthesis**”. ICML (2016).

GANs: Shortcomings

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

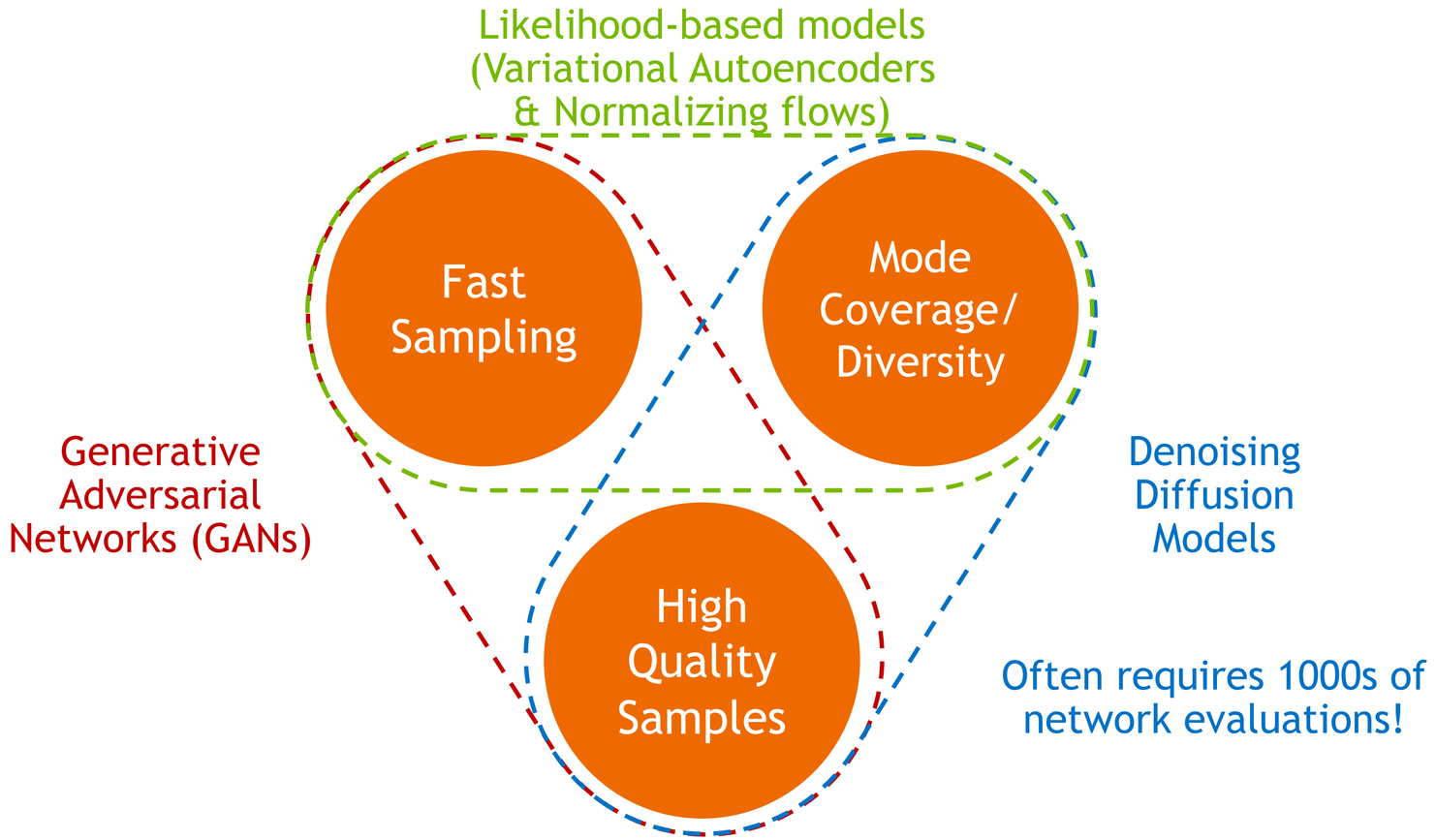
Goodfellow, Ian. **"NIPS 2016 Tutorial: Generative Adversarial Networks."** arXiv preprint arXiv:1701.00160 (2016).

GANs: Shortcomings

- Training GANs is hard
  - Mode Collapse: all the images look similar
  - Non-convergence: does not learn anything good
- They are fast during inference though!

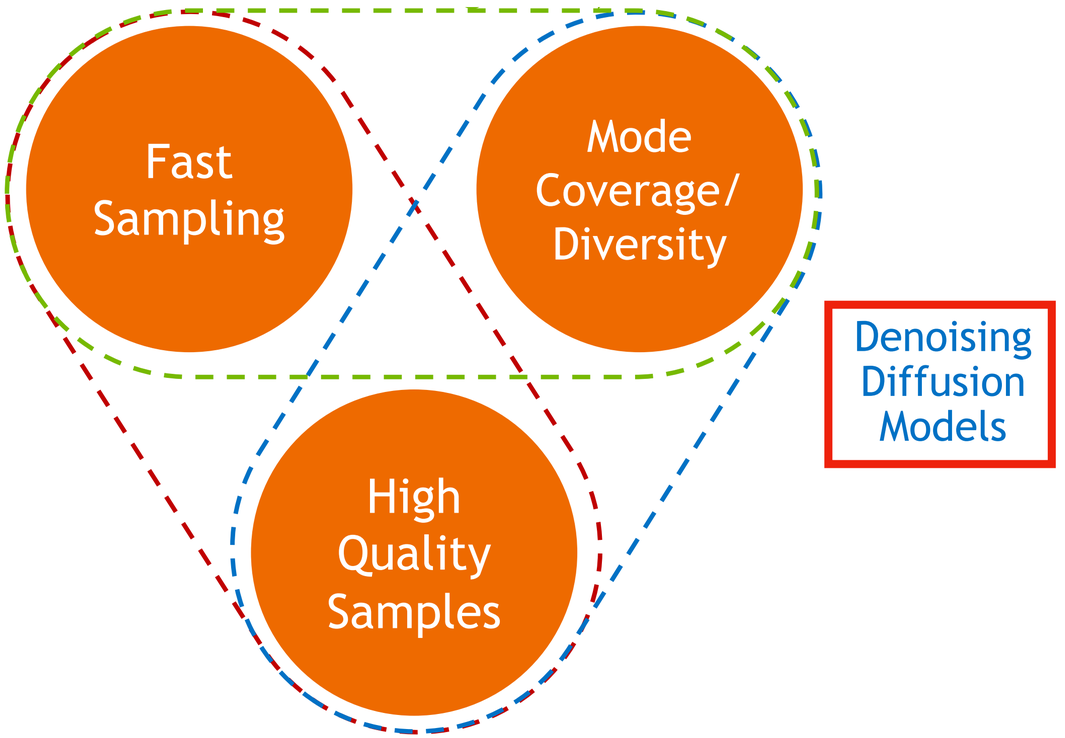
What makes a good generative model?

The generative learning trilemma



Tackling the Generative Learning Trilemma with Denoising Diffusion GANs, ICLR 2022

Deep Dive: Diffusion Models



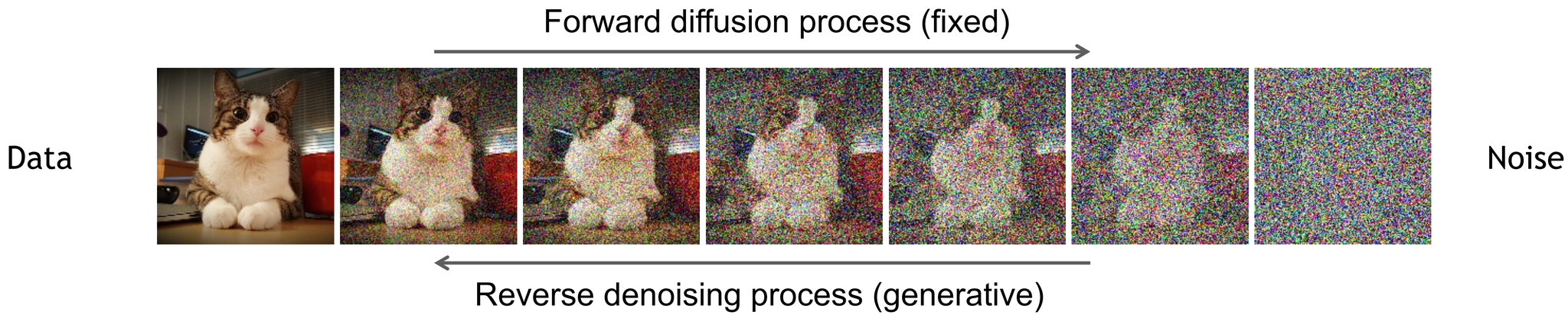
Tackling the Generative Learning Trilemma with Denoising Diffusion GANs, ICLR 2022

Denoising Diffusion Models

Learning to generate by denoising

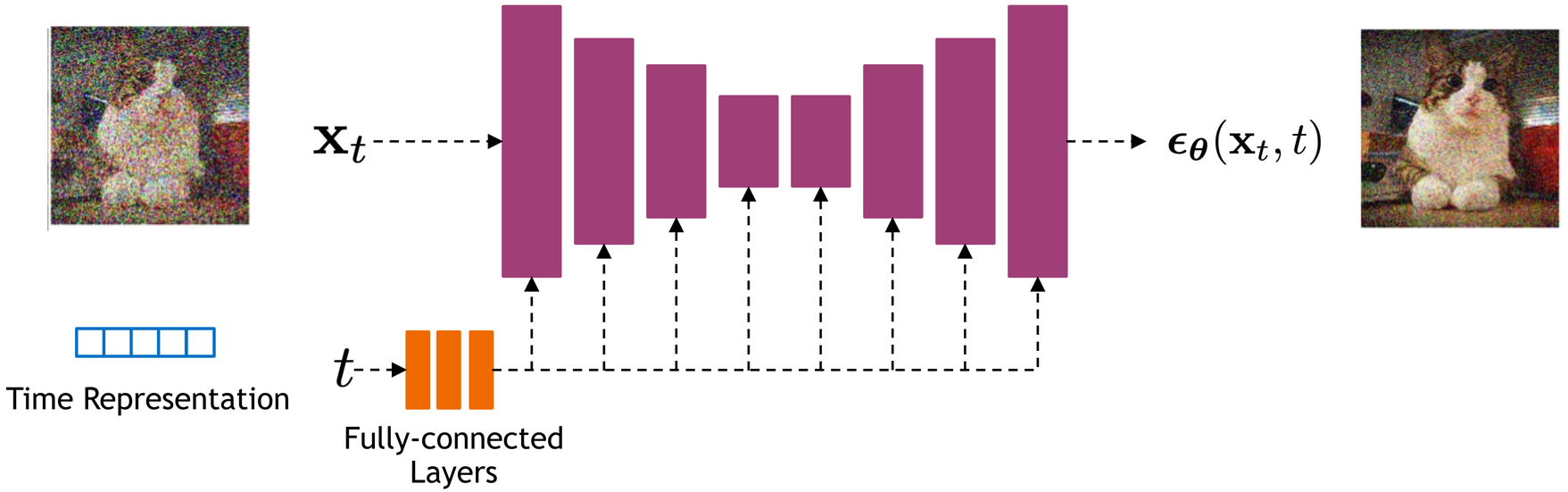
Denoising diffusion models consist of two processes:

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



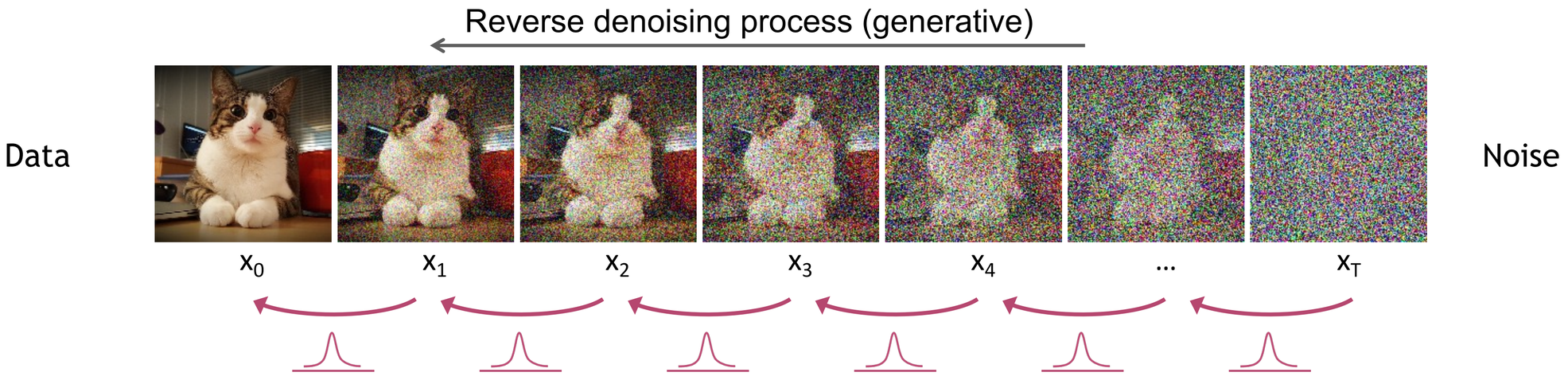
Sohl-Dickstein et al., Deep Unsupervised Learning using Nonequilibrium Thermodynamics, ICML 2015
Ho et al., Denoising Diffusion Probabilistic Models, NeurIPS 2020
Song et al., Score-Based Generative Modeling through Stochastic Differential Equations, ICLR 2021

Reverse Denoising Process



Reverse Denoising Process

Formal definition of forward and reverse processes in T steps:



$$p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$$
$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \underbrace{\mu_\theta(\mathbf{x}_t, t)}, \sigma_t^2\mathbf{I})$$

Trainable network
(U-net, Denoising Autoencoder)

$$\Rightarrow \quad p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$$

Training and Sample Generation

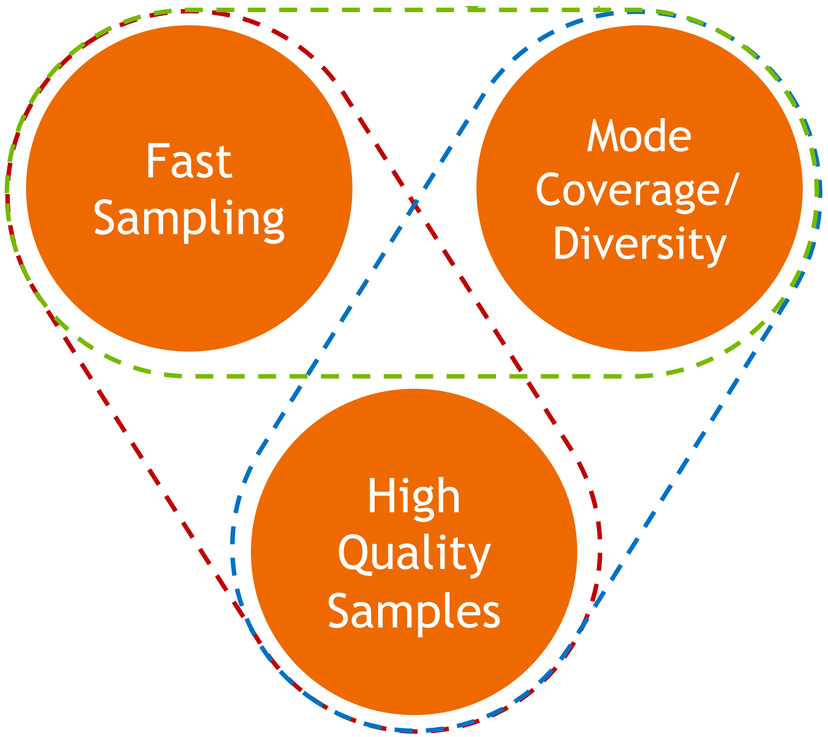
Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
6: **until** converged

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

Diffusion Model Inference is relative slow
(T network iterations for each generation)



Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
4: $\quad \mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

Diffusion Models: The Fancy

What can you do with Diffusion Models?

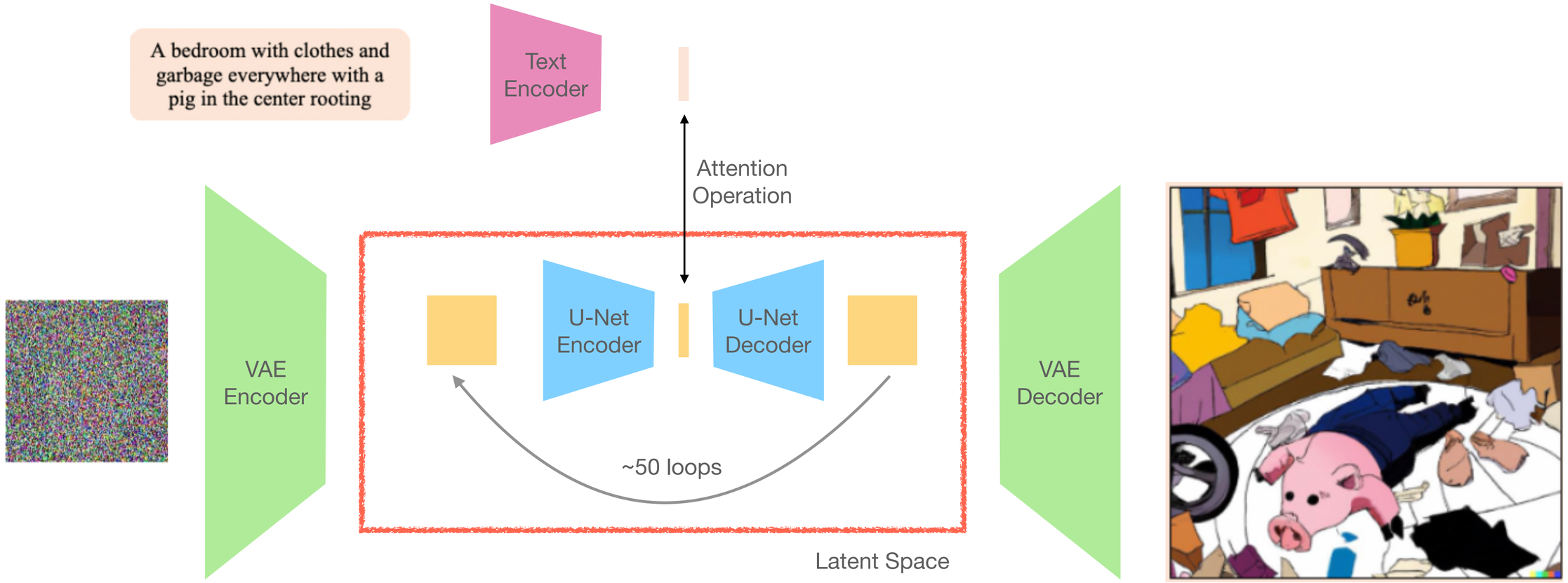
A. Stable Diffusion
B. SDEdit
C. Imagic
D. InstructPix2Pix
E. DreamBooth
F. Textual Inversion


STABLE
CASCADE
Join the
Discord


•Stable Diffusion: https://github.com/Stability-AI/stablediffusion
•Stable Cascade: https://github.com/Stability-AI/StableCascade
STABLE CASCADE
These models are fully-open source. You can use their pre-trained models to build cool things with no re-training!!
In fact, several promising papers at top research conferences do exactly just that.
Join the Discord

Stable Diffusion: insides?

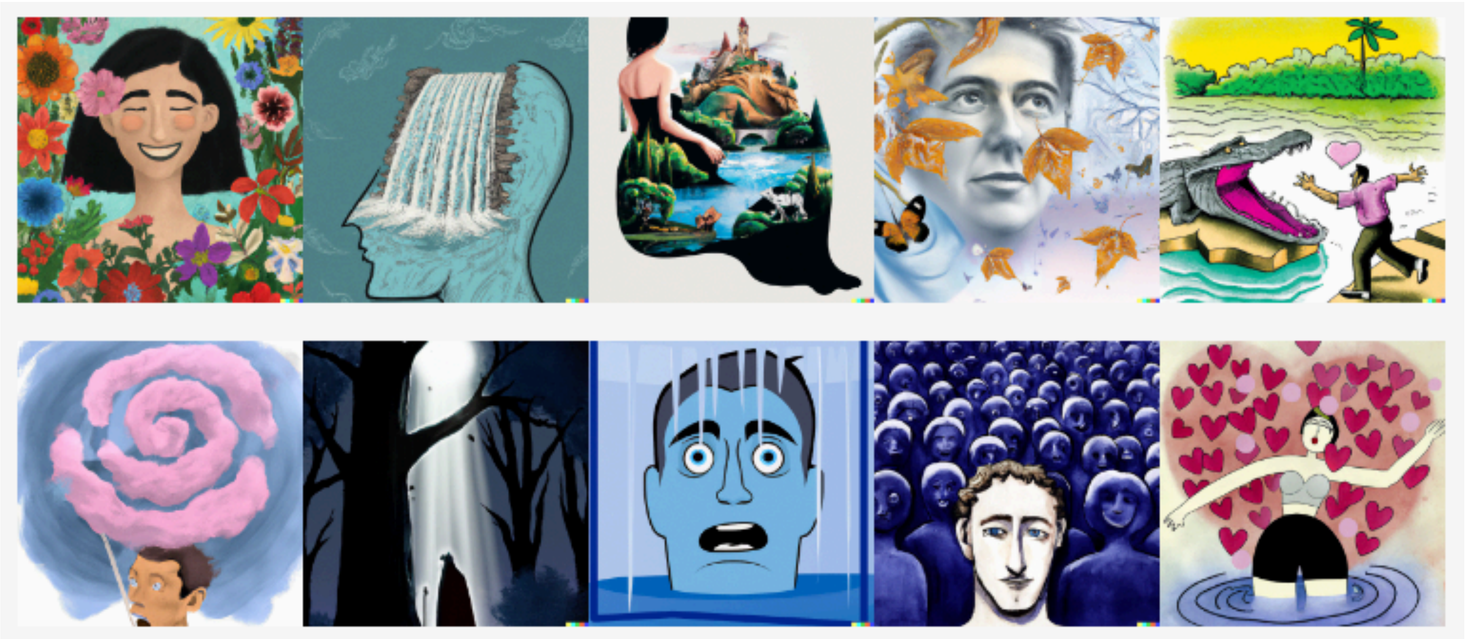
A Latent Diffusion Model



Using Stable Diffusion

Artistic Projects?

- Combine with Large Language Models
- Generate emotions & metaphors in visual form



Chakrabarty, T., Singh, K., Saakyan, A., & Muresan, S. (2023). Learning to Follow Object-Centric Image Editing Instructions Faithfully. https://tuhinjubcse.github.io

Using Stable Diffusion

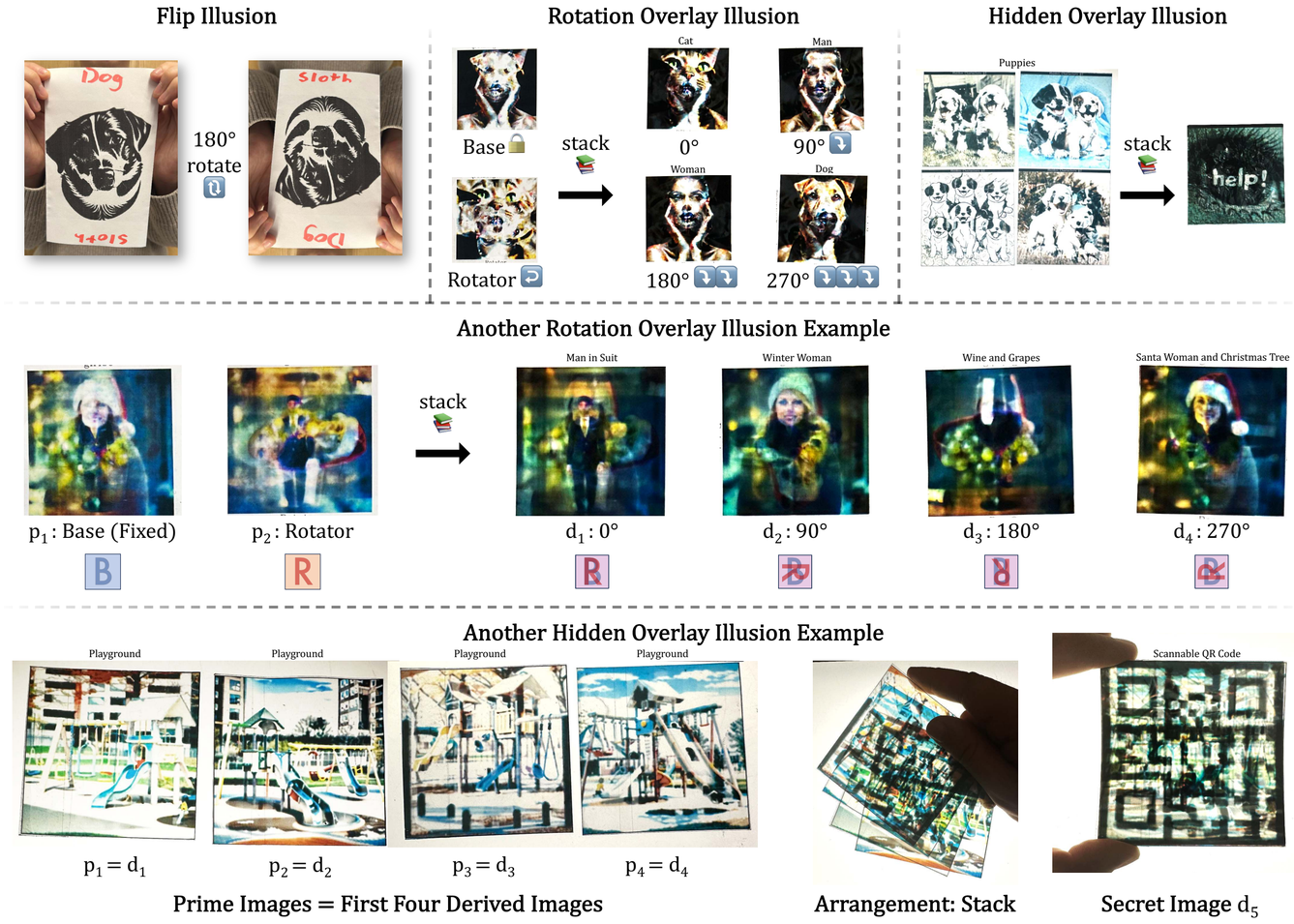
Personal Project: See https://diffusionillusions.com

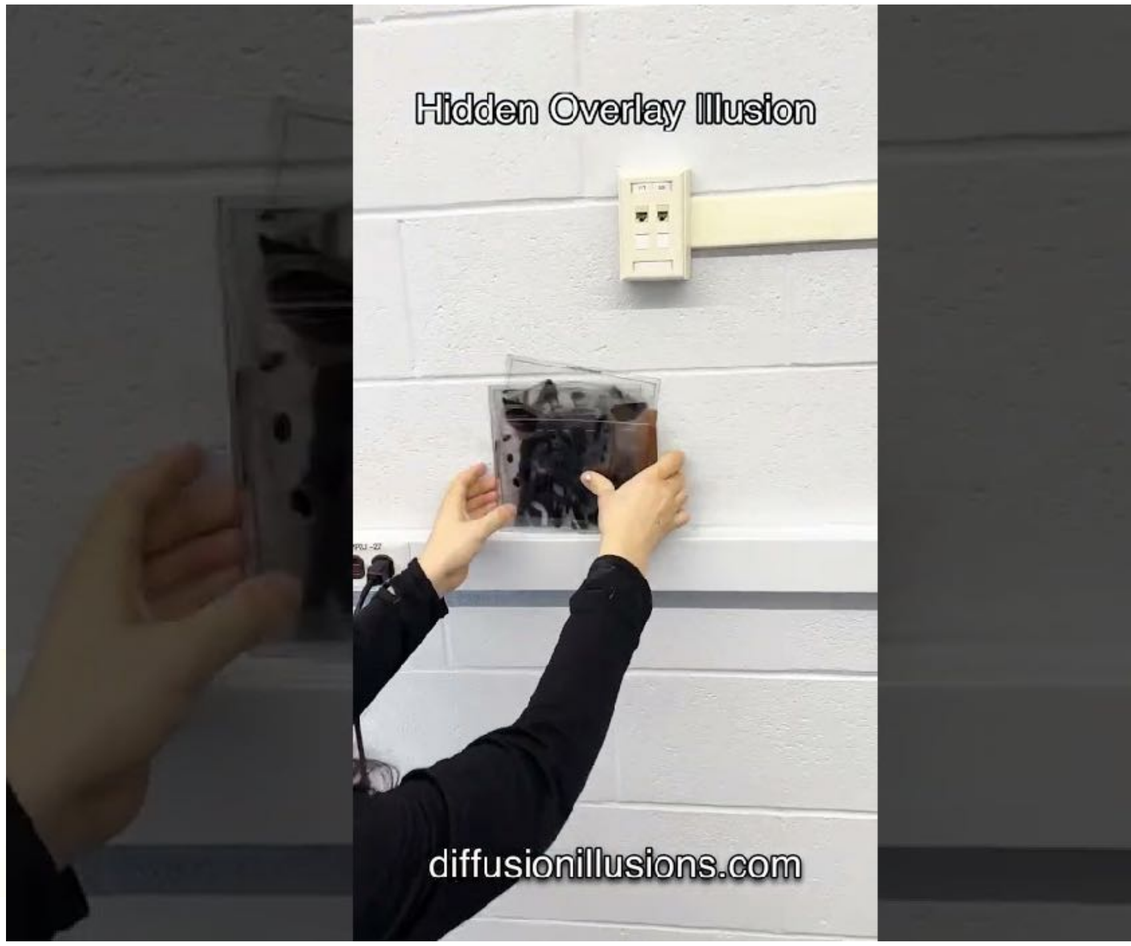
- Minimal modifications to Stable Diffusion allows us to generate Visual Illusions
- Outstanding Demo at CVPR (top vision conference)
- All experiments on a single GPU machine
- Inference works on colab (~6 minutes per generation)

Flip Illusion

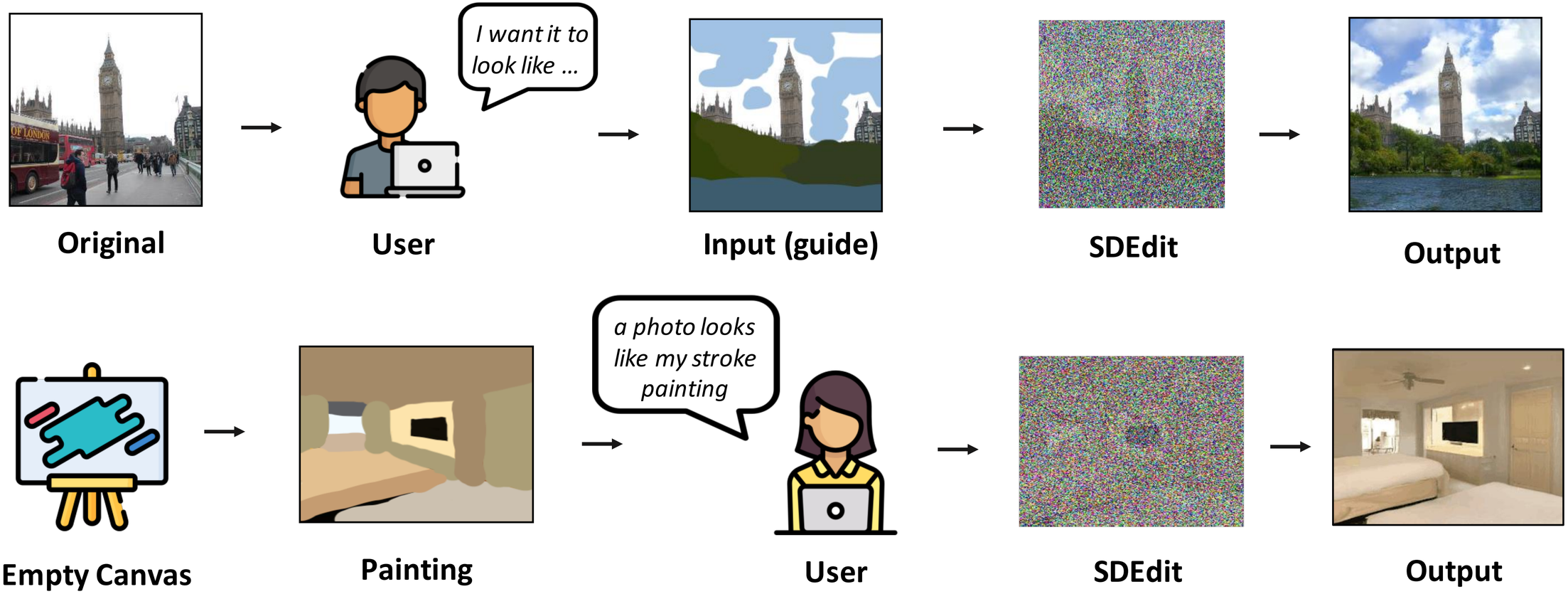
Using Stable Diffusion

Personal Projects? See https://diffusionillusions.com



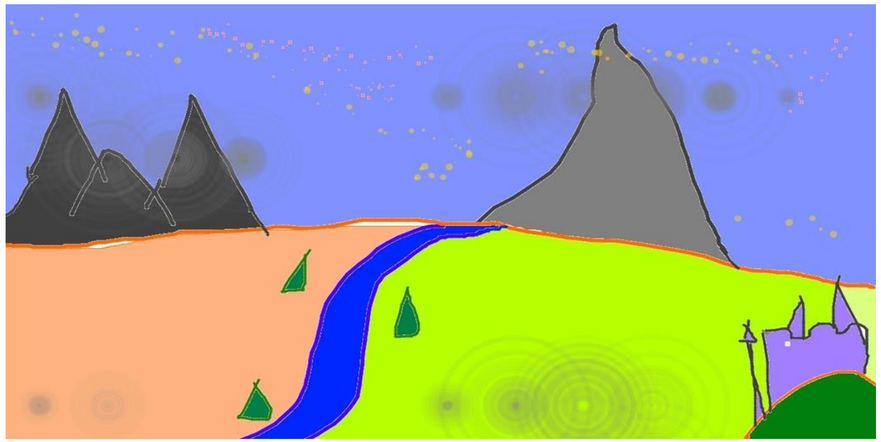


B. SDEdit



Meng et al., "SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations", ICLR 2022

Application to Stable Diffusion (img2img)



Input

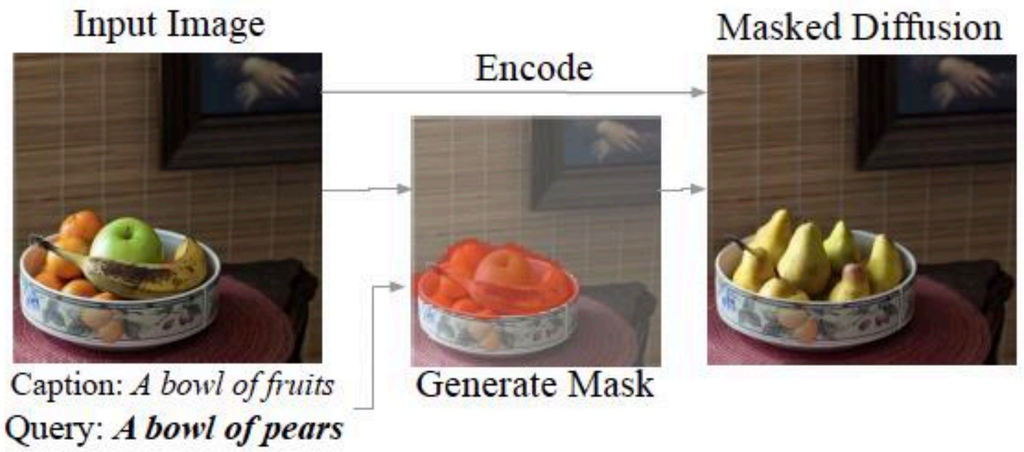
SDEdit/img2img →

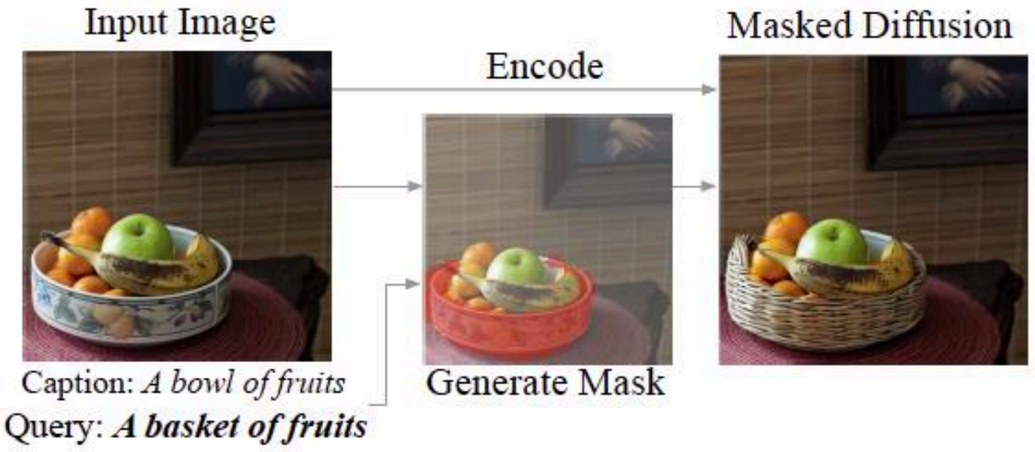
Outputs

Source: https://github.com/CompVis/stable-diffusion

DiffEdit: Diffusion-based semantic image editing with mask guidance

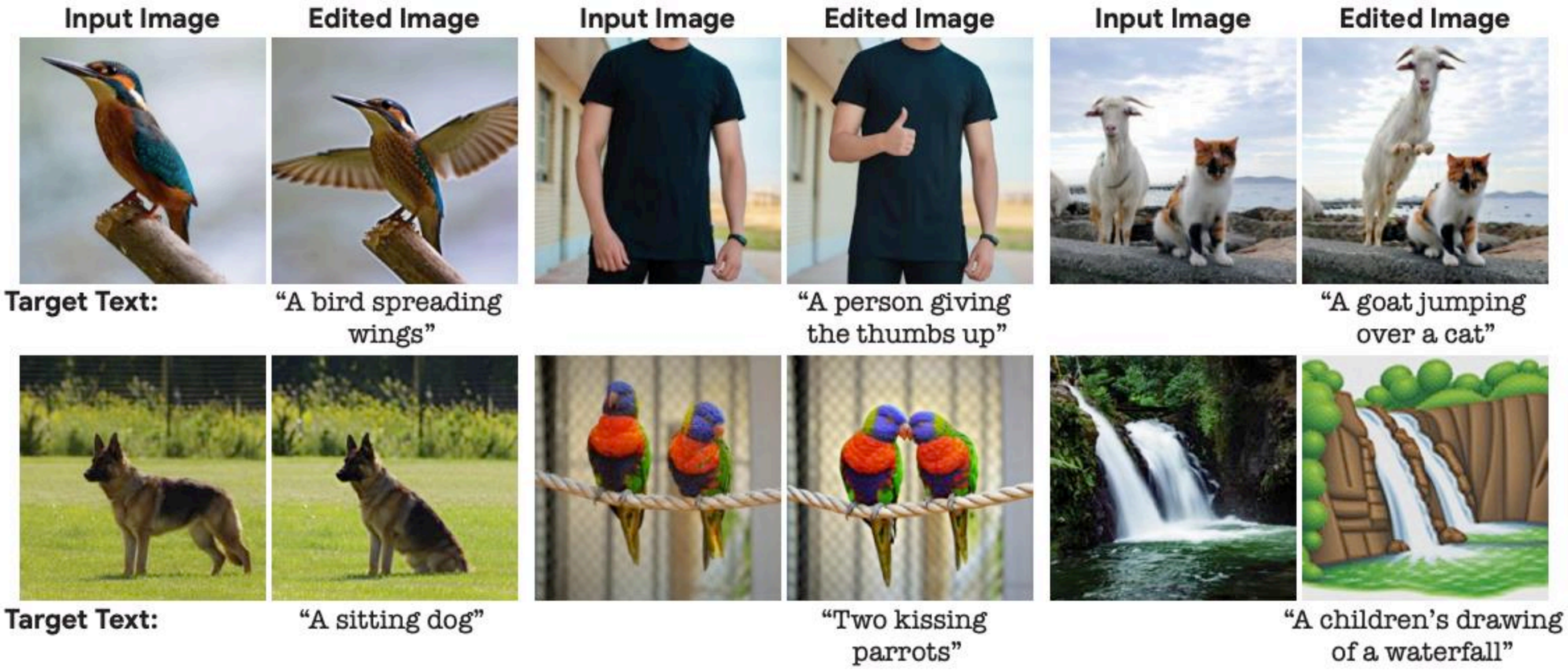
Instead of asking users to provide the mask, the model will generate the mask itself based on the caption and query.





Couairon et al., "DiffEdit: Diffusion-based semantic image editing with mask guidance", ICLR 2023

C. Imagic

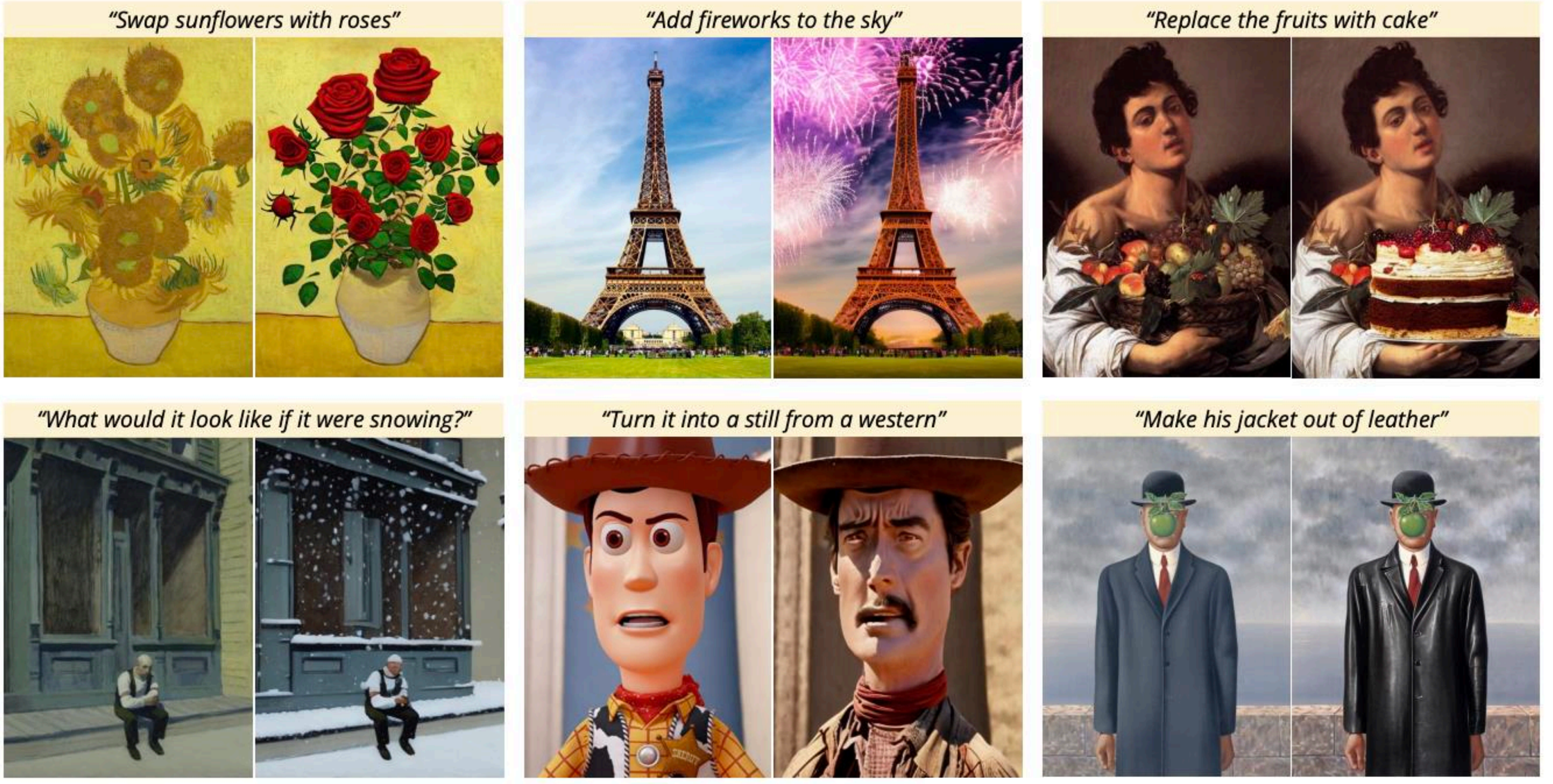


Kawar et al., "Imagic: Text-Based Real Image Editing with Diffusion Models", CVPR 2023

Imagic: Text-Based Real Image Editing with Diffusion Models

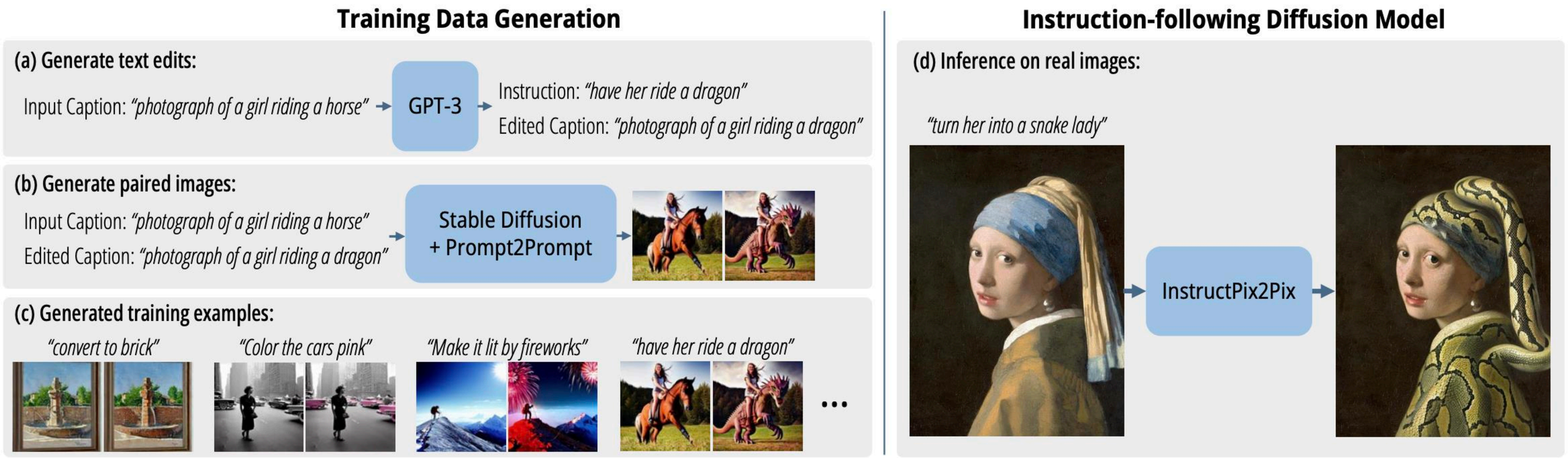
Kawar et al., "Imagic: Text-Based Real Image Editing with Diffusion Models", CVPR 2023

D. InstructPix2Pix



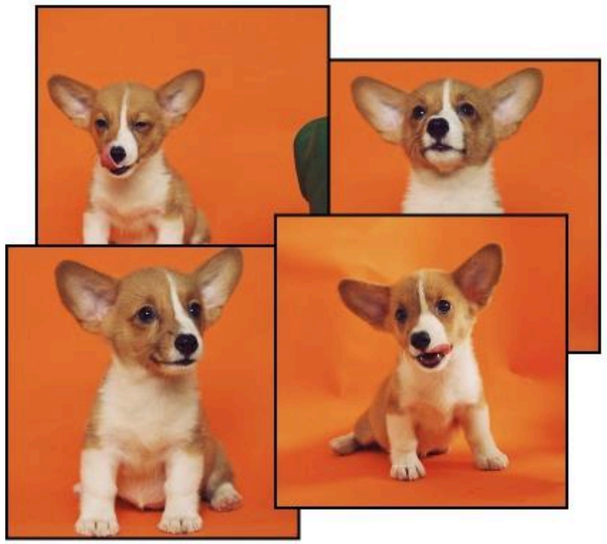
Brooks et al., "Instructpix2pix: Learning to follow image editing instructions", CVPR 2023

D. InstructPix2Pix



Brooks et al., "Instructpix2pix: Learning to follow image editing instructions", CVPR 2023

F. DreamBooth: Personalization!



Input images

Generated images

Make it generate only images of your dog!!

Ruiz et al., "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation", CVPR 2023

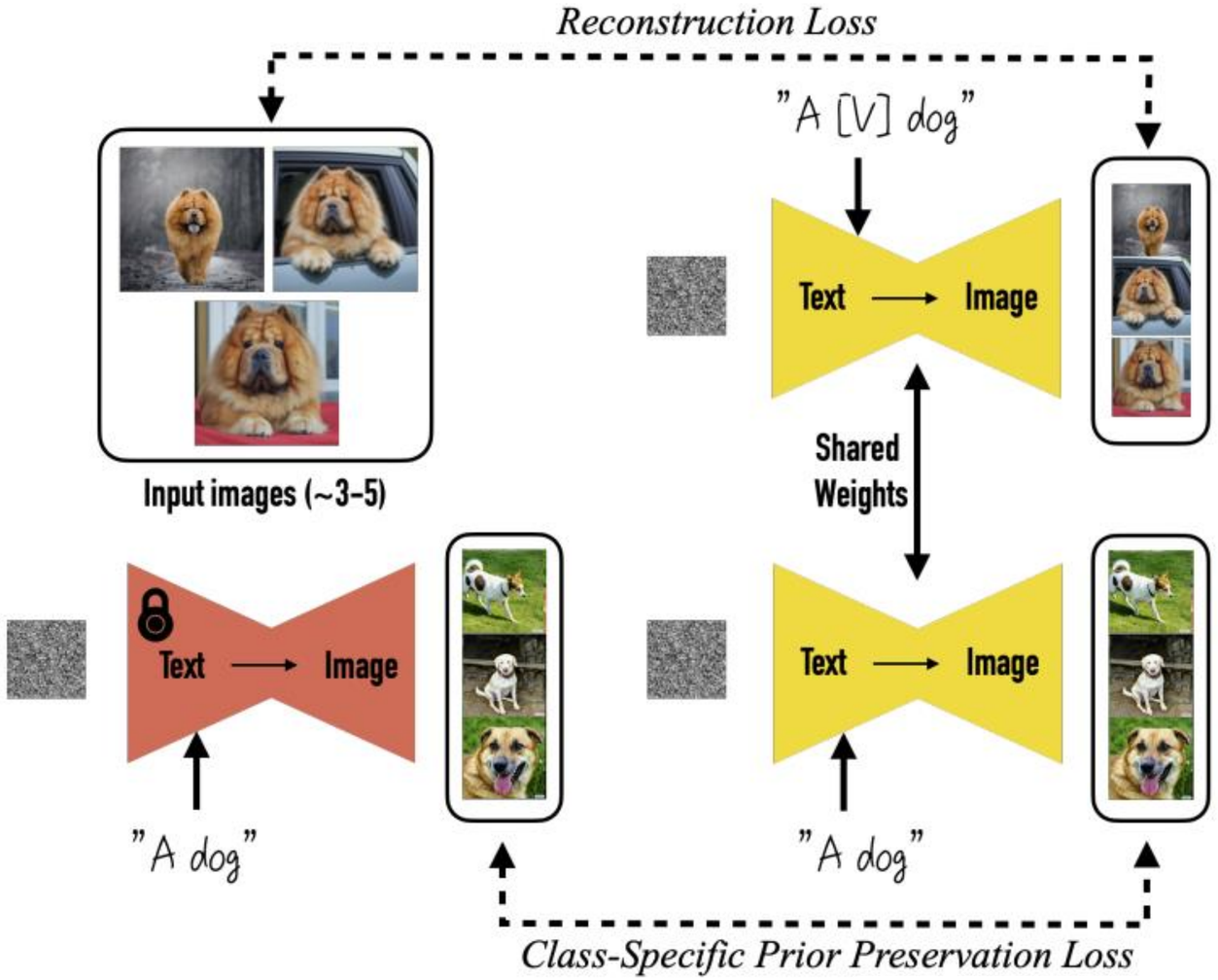
F. DreamBooth: Personalization!

Efficient Fine Tuning to Personalize a Diffusion Model

Ruiz et al., "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation", CVPR 2023

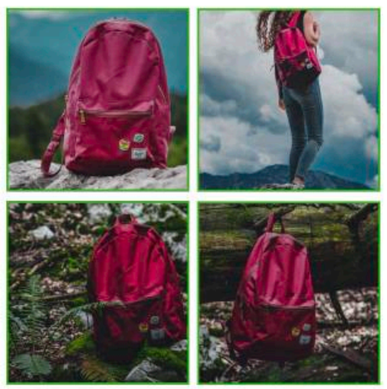
F. DreamBooth: Personalization!

Efficient Fine Tuning to Personalize a Diffusion Model



Ruiz et al., "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation", CVPR 2023

F. DreamBooth: Personalization!



Input images

A [V] backpack in the Grand Canyon

A wet [V] backpack in water

A [V] backpack in Boston

A [V] backpack with the night sky

Input images

A [V] teapot floating in milk

A transparent [V] teapot with milk inside

A [V] teapot pouring tea

A [V] teapot floating in the sea

Ruiz et al., "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation", CVPR 2023

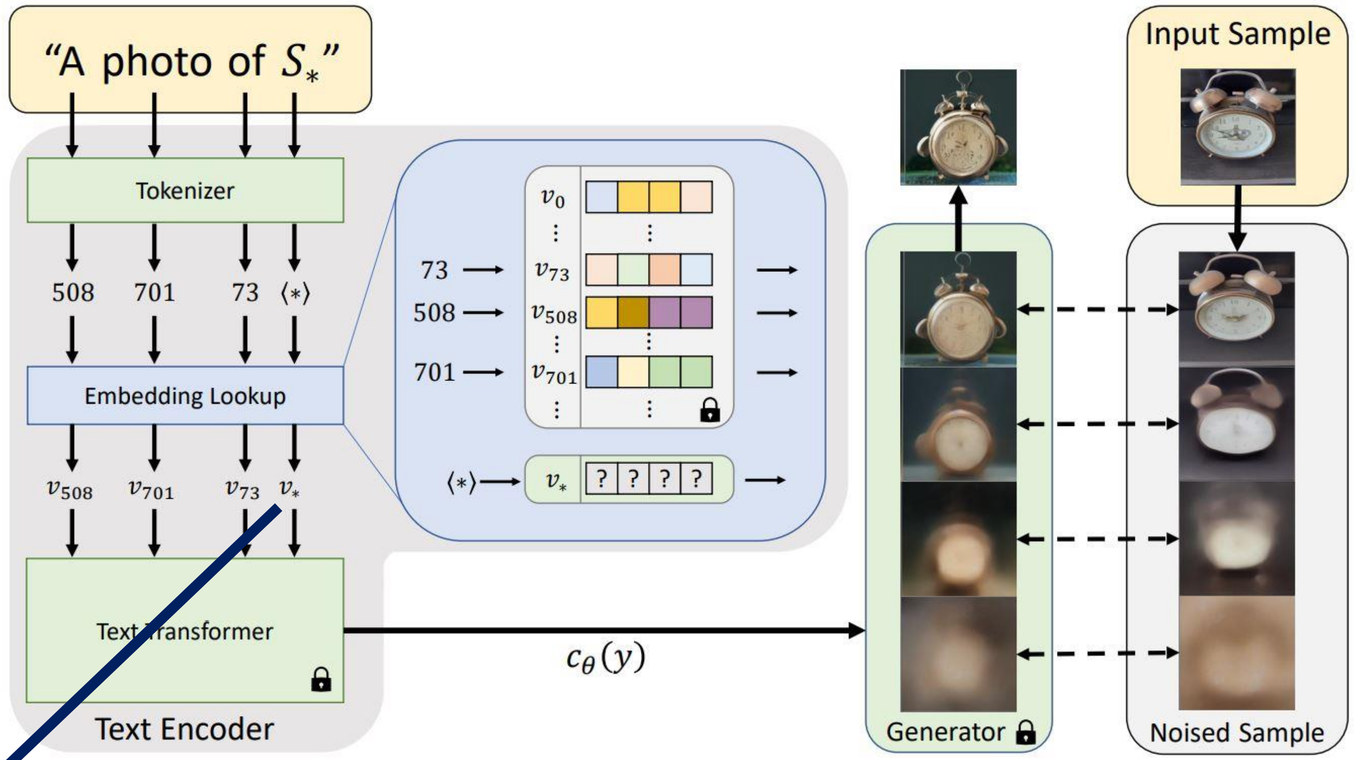
G. Textual Inversion: Personalization!

Learn A Text Embedding to Personalize a Diffusion Model

Gal et al., "An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion", ICLR 2023

G. Textual Inversion: Personalization!

Learning (or optimizing) the Personalized Text Embedding



$$v_* = \arg\min_v \mathbb{E}_{z\sim\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c_\theta(y))\|_2^2\right]$$

Gal et al., "An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion", ICLR 2023

G. Textual Inversion: Some Results

Input samples

"S_* sports car"

"S_* made of lego"

"S_* onesie"

"da Vinci sketch of S_*"

Input samples

"Manga drawing of a steaming S_*"

"A S_* watering can"

"S_* Death Star"

"A poster for the movie 'The Teapot' starring S_*"

Gal et al., "An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion", ICLR 2023

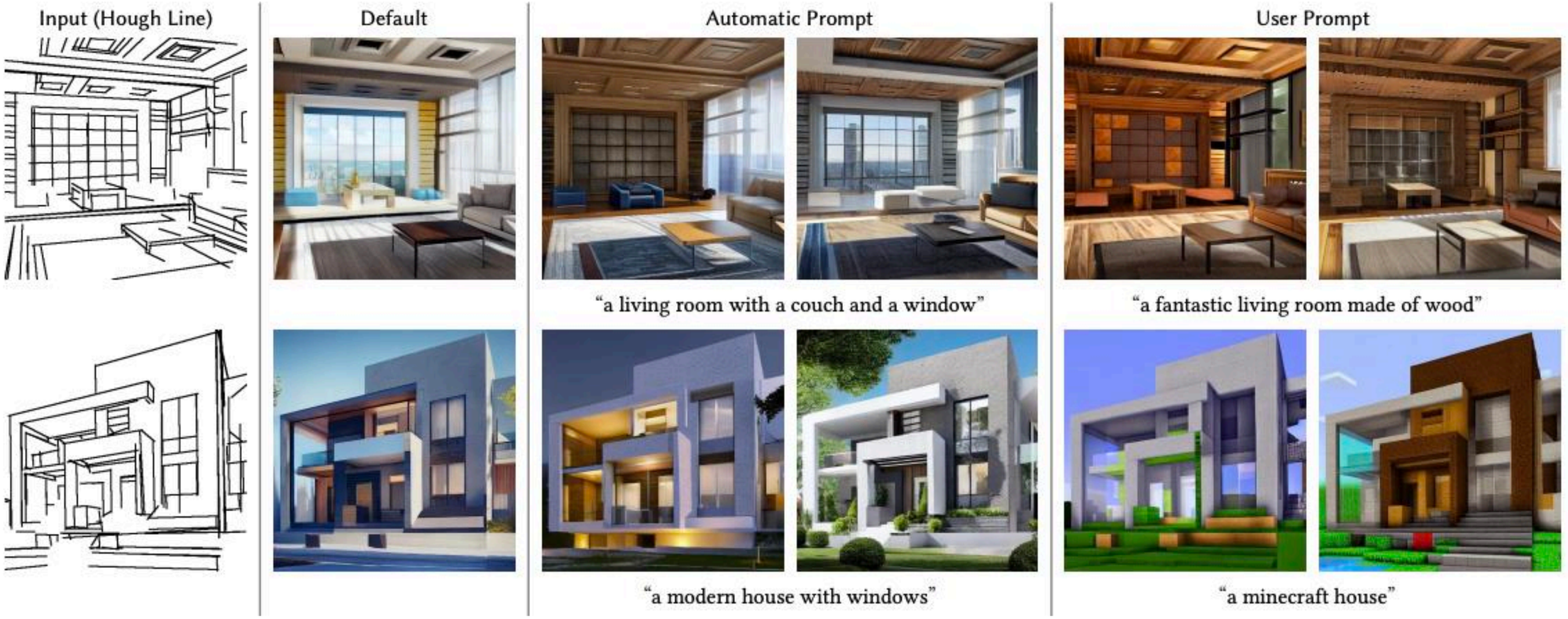
G. Textual Inversion: Personalization!

Can even learn an artistic style

Gal et al., "An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion", ICLR 2023

ControlNet

Controlling Diffusion Model Generation
ICCV ’23 Best Paper (Marr Prize)



Zhang, L., Rao, A., & Agrawala, M., *Adding Conditional Control to Text-to-Image Diffusion Models*, In ICCV 2023

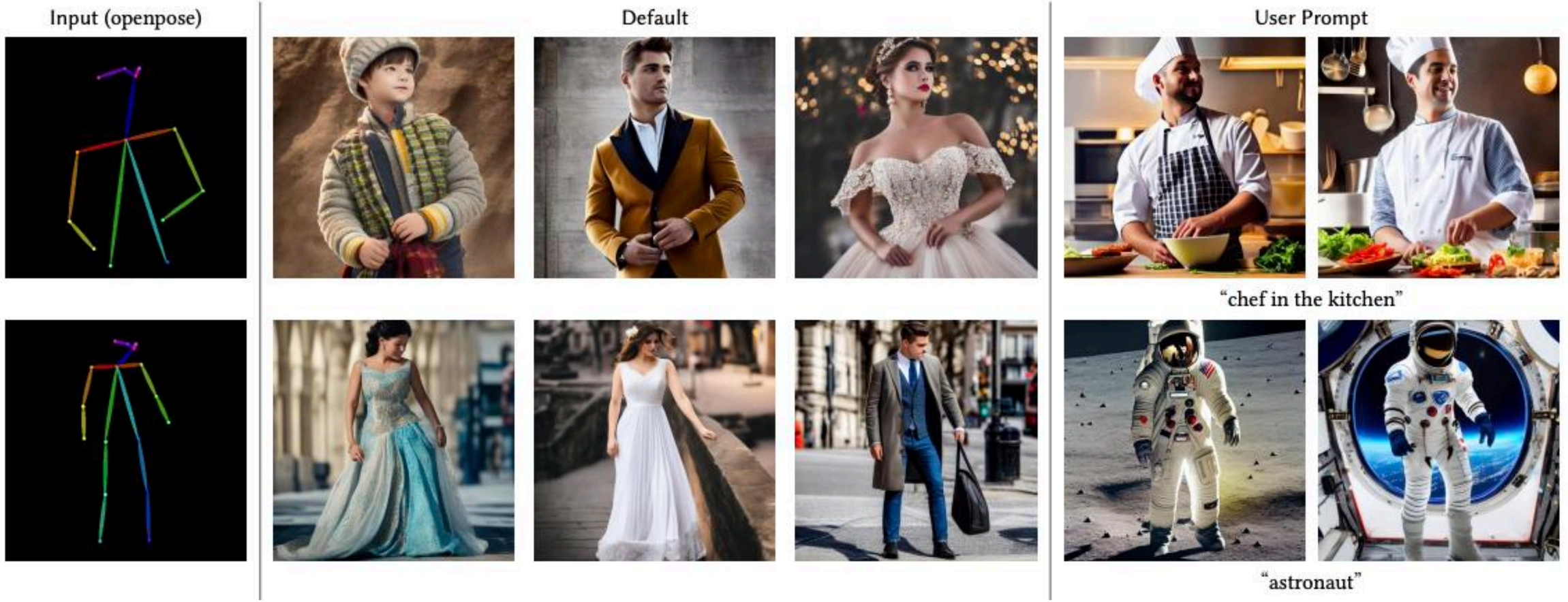
ControlNet

Controlling Diffusion Model Generation
ICCV ’23 Best Paper (Marr Prize)

Zhang, L., Rao, A., & Agrawala, M., *Adding Conditional Control to Text-to-Image Diffusion Models*, In ICCV 2023

ControlNet

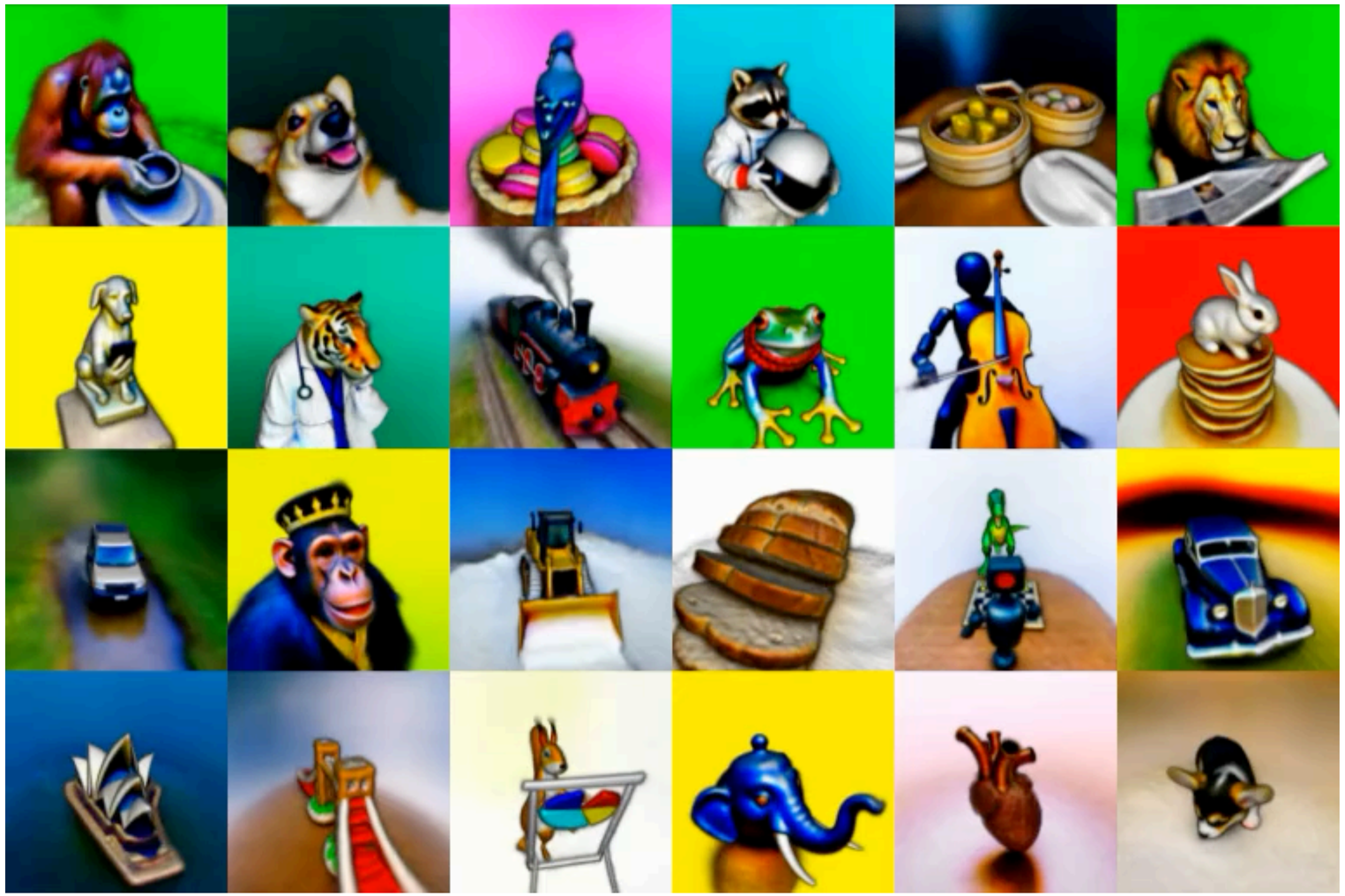
Artistic Applications in Fashion and Design?



Zhang, L., Rao, A., & Agrawala, M., *Adding Conditional Control to Text-to-Image Diffusion Models*, In ICCV 2023

Vision Transcends Beyond Images

Generating 3D objects, videos, and more!



DreamFusion

OpenSource Version:
https://github.com/ashawkey/stable-dreamfusion

Poole et al., "DreamFusion: Text-to-3D using 2D Diffusion", ICLR 2023

Vision Transcends Beyond Images

OpenSource Video Generation

Prompt: “*Darth Vader is surfing on waves.*”

Cerspense Models

OpenSource:
https://huggingface.co/cerspense/zeroscope_v2_576w

Poole et al., "DreamFusion: Text-to-3D using 2D Diffusion", ICLR 2023

Getting Started?

- Fast open-source diffusion models:
  1. https://github.com/Stability-AI/StableCascade
  2. https://huggingface.co/ByteDance/SDXL-Lightning

- Open-source video generation models:
  1. https://huggingface.co/cerspense/zeroscope_v2_576w
  2. https://makepixelsdance.github.io

- Animation Models:
  1. https://showlab.github.io/magicanimate/

Getting Started? HF Diffusers Library!

- Video Generation with Cerscape model: example code

```
import torch
from diffusers import DiffusionPipeline, DPMSolverMultistepScheduler
from diffusers.utils import export_to_video

pipe = DiffusionPipeline.from_pretrained("cerspense/zeroscope_v2_576w", torch_dtype=torch.float16)
pipe.scheduler = DPMSolverMultistepScheduler.from_config(pipe.scheduler.config)
pipe.enable_model_cpu_offload()

prompt = "Darth Vader is surfing on waves"
video_frames = pipe(prompt, num_inference_steps=40, height=320, width=576, num_frames=24).frames
video_path = export_to_video(video_frames)
```