The Present and Future of Image and Video Generation Technologies

KAIST 소프트웨어대학원/KTAI 콜로퀴엄

May 2025

성 민 혁 | KAIST 전산학부

KAIST

Minhyuk Sung

- Associate Professor at KAIST — 2021 – Present

  School of Computing

  Affiliated with the Graduate School of AI

  and with the Graduate School of Metaverse

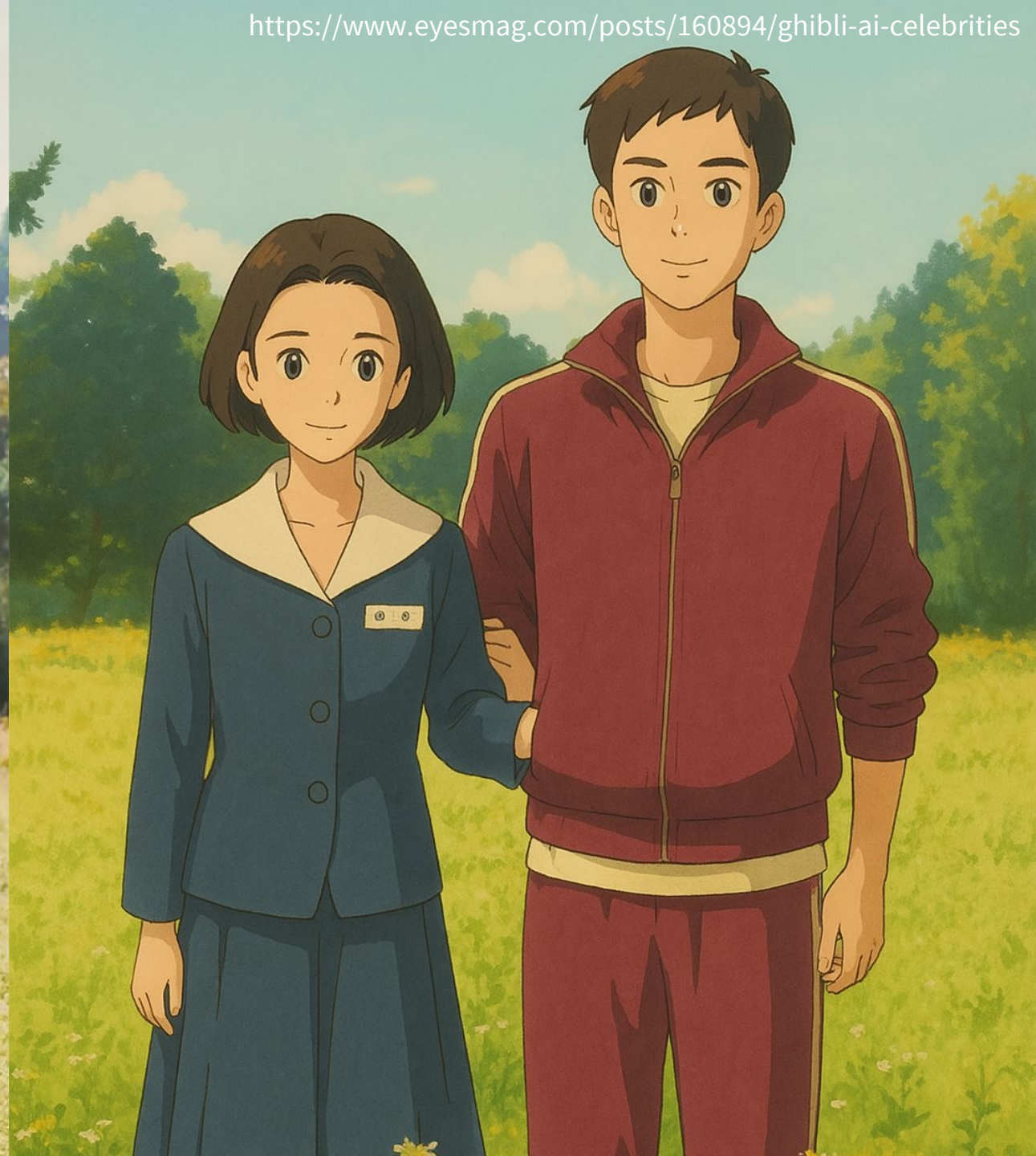
- Research Scientist at Adobe Research — 2019 – 2020

- Ph.D. at Stanford University — 2019

  Computer Science Department

OpenAI GPT-4o

https://www.eyesmag.com/posts/160894/ghibli-ai-celebrities

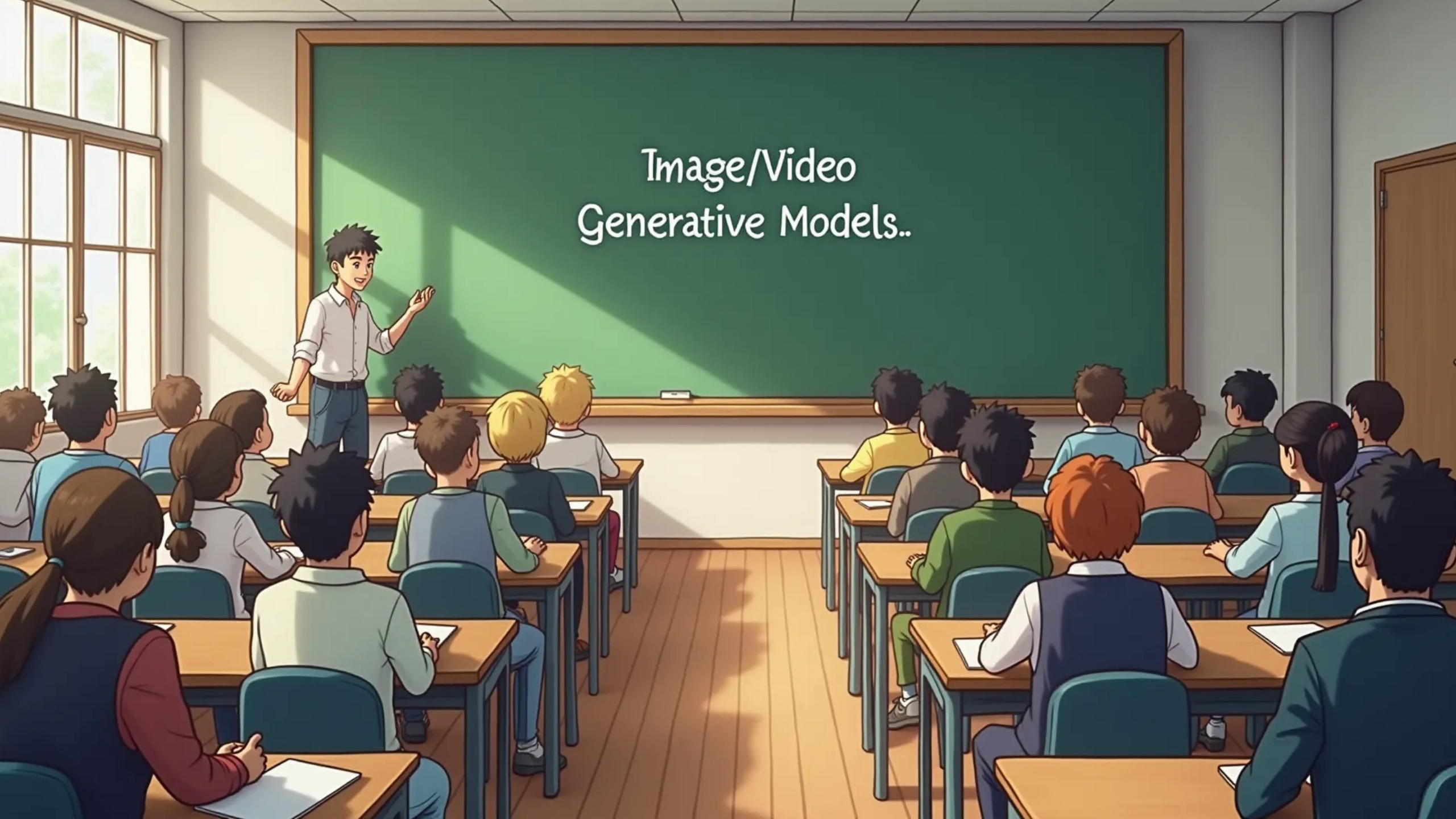



FLUX
What a HOOT.
Let's make some
really pretty stuff
together.
Black Forest Labs Flux

Image Generation Example

https://huggingface.co/spaces/black-forest-labs/FLUX.1-dev
https://huggingface.co/spaces/black-forest-labs/FLUX.1-schnell

A classroom for “Image/Video Generative Models” with a large audience. On the board, the text reads: “Image/Video Generative Models.” Create the image in Ghibli style.


Image/Video
Generative Models..

https://hai-production.s3.amazonaws.com/images/inline-images/Screenshot%202024-04-03%20at%202.46.01%E2%80%AFPM.png

February 2022

April 2022

July 2022

November 2022

March 2023

March 2023 (new version)

June 2023

December 2023

Midjourney | Prompt: "a hyper-realistic image of Harry Potter."

Chen et al., An Empirical Study of GPT-4o Image Generation Capabilities, arXiv 2025.

Analysis of Recent Image Generative Models

An Empirical Study of GPT-4o Image Generation Capabilities

Sixiang Chen[1*], Jinbin Bai[2*], Zhuoran Zhao[1*], Tian Ye[1*], Qingyu Shi[3], Donghao Zhou[4], Wenhao Chai[5], Xin Lin[6], Jianzong Wu[3], Chao Tang[3], Shilin Xu[3], Tao Zhang[6], Haobo Yuan[6], Yikang Zhou[6], Wei Chow[2], Linfeng Li[2], Xiangtai Li[3†], Lei Zhu[1,7†], Lu Qi[6†]

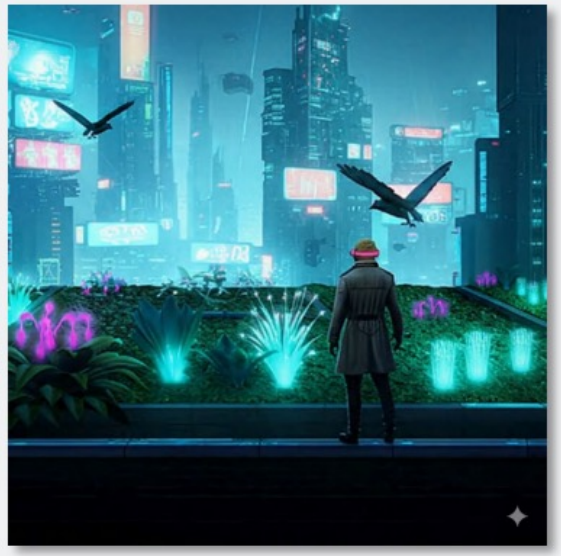
[1]The Hong Kong University of Science and Technology (GZ) [2]National University of Singapore
[3]Peking University [4]The Chinese University of Hong Kong [5]University of Washington [6]Wuhan University
[7]The Hong Kong University of Science and Technology

Precise Text Alignment

SD 2.1

GPT 4o

Gemini 2.0 Flash



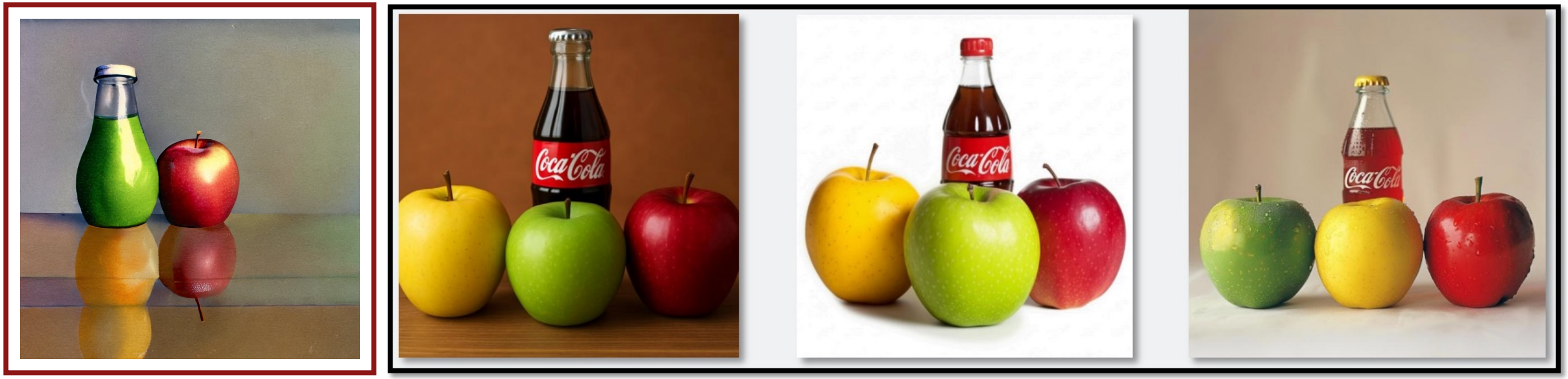
FLUX

"On the rooftop of a skyscraper in a bustling cyberpunk city, a figure in a trench coat and neon-lit visor stands amidst a garden of bio-luminescent plants, overlooking the maze of flying cars and towering holograms. Robotic birds flit among the foliage, digital billboards flash advertisements in the distance."

Chen et al., An Empirical Study of GPT-4o Image Generation Capabilities, arXiv 2025.

Object Count and Composition

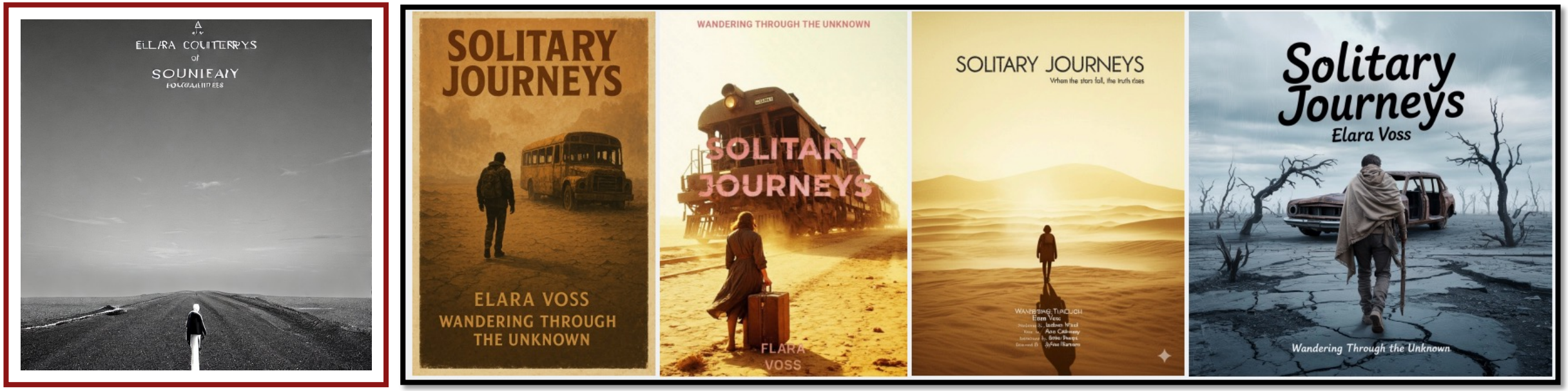
SD 2.1 | GPT 4o | Gemini 2.0 Flash | Midjourney v6.1



"Three differently colored apples (yellow, green, red from left to right) with a Coca-Cola bottle placed behind the middle apple."

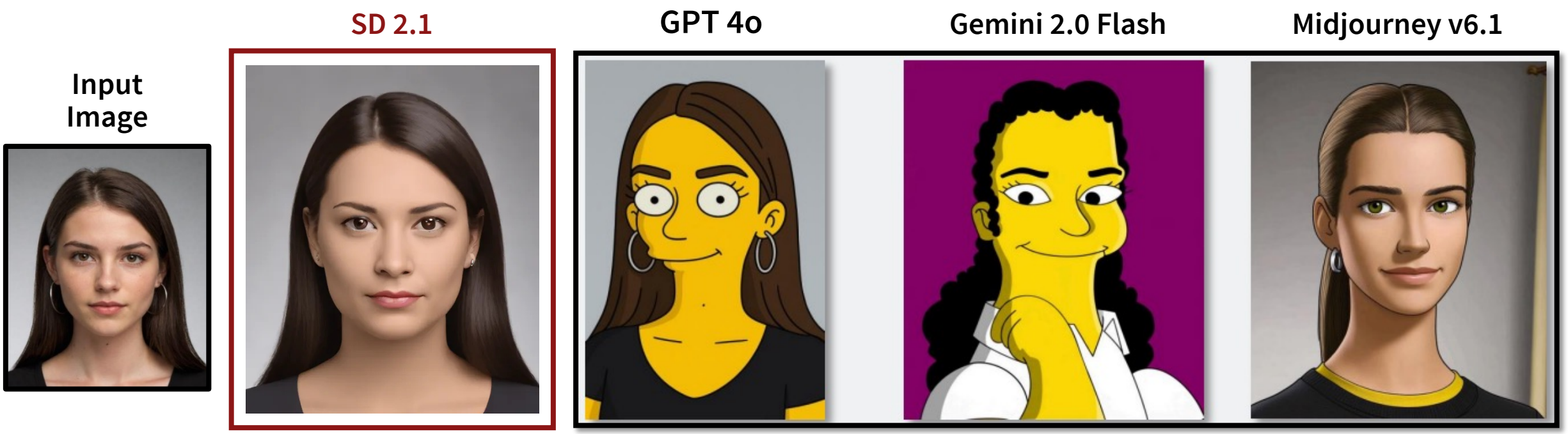
Text Rendering

SD 2.1 | GPT 4o | POSTA | Gemini 2.0 Flash | Ideogram 3.0



"Create a poster with the theme of a Journey of Solitude. The background should depict a lone figure walking toward an unusable form of transportation. The scene should evoke a sense of being lost, helplessness, and desolation, capturing the emotional weight of losing oneself in a barren, unforgiving landscape. Title: Solitary Journeys Subtitle: Elara Voss
Information: WANDERING THROUGH THE UNKNOWN."

Stylization



"Generate the Simpsons style of this picture."


OpenAI Sora

https://www.youtube.com/watch?v=UL8SF-hthEE

Google DeepMind Veo 2

Prompt: A cinematic, high-action tracking shot follows an incredibly cute dachshund wearing swimming goggles as it leaps into a crystal-clear pool. The camera plunges underwater with the dog, capturing the joyful moment of submersion and the ensuing flurry of paddling with adorable little paws. Sunlight filters through the water, illuminating the dachshund's sleek, wet fur and highlighting the determined expression on its face. The shot is filled with the vibrant blues and greens of the pool water, creating a dynamic and visually stunning sequence that captures the pure joy and energy of the swimming dachshund.

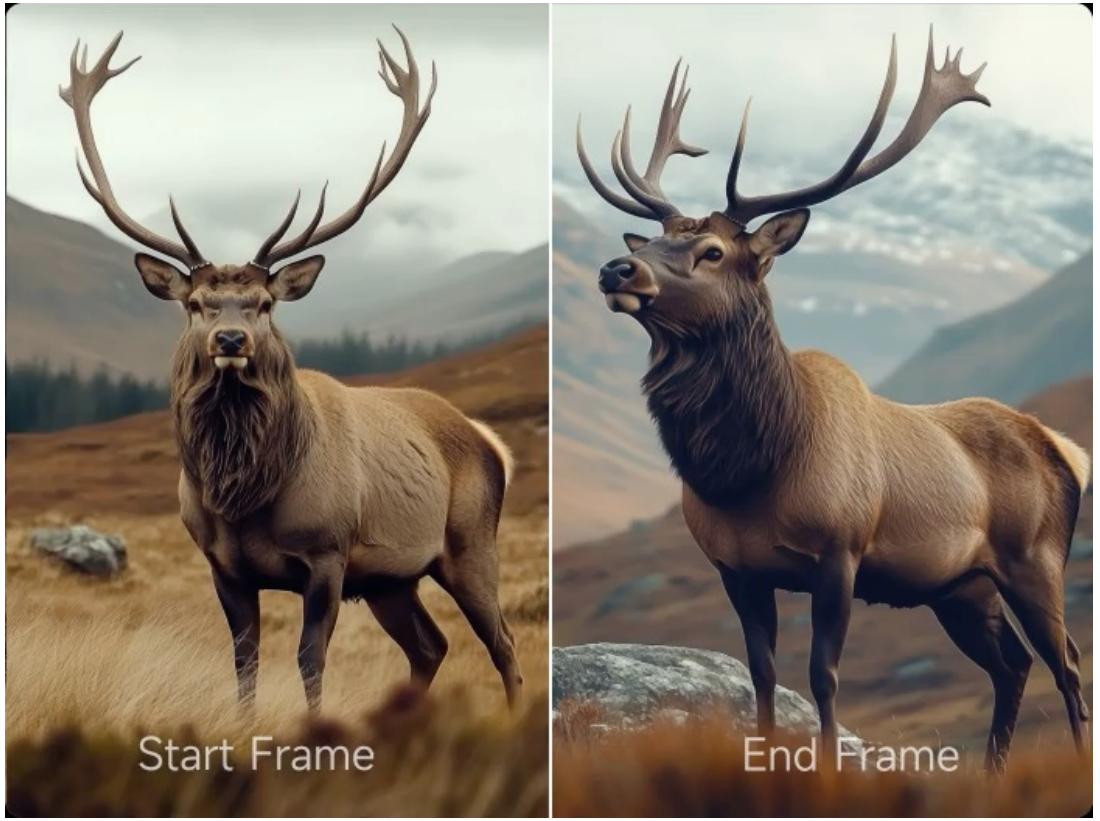
https://www.klingai.com/global/

KlingAI

Prompt: Rocks, spheres flying around a silver mirrored moon, digital art style, Clement’s Archer, Felicia Simon, Vray, desert-wave, Daniel Anshan.

Control

Motion Brush



Frame Interpolation

Jason Allen won the digital-art competition at the Colorado State Fair last year for his piece “Théâtre D’opéra Spatial” that he created using the AI software Midjourney. Recently, the US Copyright Office refused to grant him a copyright for his piece, writing, “We have decided that we cannot register this copyright claim because the deposit does not contain any human authorship.” He plans to appeal.


AIFF The 3rd Annual International runway
https://aiff.runwayml.com/#finalists
FAQ
Screening Room
Attend in NYC
Attend in LA
AIFF 2025
Finalists Announced
View Finalists
June 5th in New York at Alice Tully Hall | June 12th in LA at The Broad Stage Theater
Presenting Partners
TRIBECA FESTIVAL
IMAX
CapCut
the gotham
.monks
NVIDIA.
VIVA TECHNOLOGY
GENERAL ATLANTIC
Goldman Sachs
runway
ACE


AI FILM FESTIVAL
https://www.instagram.com/runwayapp/reel/DJZb3sIKYha/?locale=ja_JP&hl=en
runway

https://www.bbc.com/news/articles/ce303x19dwgo

BBC

Home News Sport Business Innovation Culture Arts Travel Earth Audio Video Live

AI was enemy No. 1 during Hollywood strikes. Now it's in Oscar-winning films

31 March 2025

Share Save

Regan Morris
BBC News, Los Angeles

Getty Images

Image/Video Generative Models

Gemini 2.0 Flash

Midjourney

GPT-4o

Source: Perplexity

Image/Video Generative Models

Top Image Generative AI Companies (2025)

The image generative AI sector is highly competitive, with several companies leading in technology, market share, and innovation. Below are the top companies shaping the field in 2025, based on market share, funding, and industry influence:

Company	Notable Product/Model	Key Details & Market Position
Black Forest Labs	Flux family	Emerged as a market leader, holding close to 40% of image generation usage share in 2025 [9][2]. Known for advanced models and rapid ascent.
Google	Imagen3	Holds nearly 30% of usage share in 2025. Significant investment and rapid growth in image generation [9][5].
Stability AI	Stable Diffusion	Early leader in open-source image generation, still a major player with broad adoption [2][9].
Midjourney	Midjourney	Popular for artistic and creative image generation, widely used among designers and artists [2].
Adobe	Firefly, Photoshop AI	Dominates creative professional market, integrating AI into Creative Cloud suite [6].
Meta	AI image tools on Facebook/Instagram	Leverages massive user base, integrates AI image generation into social media platforms [6][4].
OpenAI	DALL·E series	Pioneered text-to-image generation, remains influential with DALL·E 3 and ongoing model improvements [7].
Luma AI	Dream Machine	Focuses on 3D and high-fidelity image/video generation, growing rapidly in creative sectors [2].
Leonardo AI	Alchemy Refiner, Image Generator	Offers advanced content creation tools, gaining traction in creative industries [2].
Runway	Gen-1, Gen-2, Gen-3, Gen-4	Known for generative video and image tools, widely used in media and entertainment [7].

Top Video Generative AI Companies in 2025

The video generative AI landscape in 2025 is led by several innovative companies offering advanced tools for content creation, editing, and automation. The following companies are consistently ranked among the top in the industry:

Company	Notable Product/Platform	Key Features and Highlights
Synthesia	Synthesia Studio	Studio-quality AI avatar videos, 230+ avatars, 140+ languages, widely used in enterprise and training [1][5].
Runway	Gen-3, Gen-4, Runway ML	Next-gen AI video editing and generation, popular for creative and professional applications [1][5][6].
Pika	Pika Art	AI-powered video editing and generation from captions and images, rapid user growth [1][4][5][7].
InVideo	InVideo AI	Full-length video creation, AI scriptwriting, media selection, accessible for creators [4][5].
Kling	Kling AI Video Generator	High-quality motion video generation, gaining attention for realism and innovation [1][6].
Google	Veo (Veo 2)	Advanced generative video with strong physics and realism, limited free access [1][6].
OpenAI	Sora	Text-to-video generation, visually compelling results, ongoing development [1][4][6].
Luma Labs	Dream Machine	Focus on 3D and realistic video generation, paid-only access, strong realism [1][6].

About

Midjourney is an independent research lab exploring new mediums of thought and expanding the imaginative powers of the human species.

We are a small self-funded team focused on design, human infrastructure, and AI. We have 11 full-time staff and an incredible set of advisors.

Executives

David Holz

Previously: Founder Leap Motion, Researcher at NASA, Max Planck

Advisors

Jim Keller

Lead Silicon at Apple, AMD, Tesla, Intel, Coauthor X86-64, CTO Tenstorrent

Nat Friedman

CEO Github, Chairman of GNOME Foundation

Philip Rosedale

Founder of Second Life, CTO RealNetworks

Bill Warner

Founder of Avid Technology, inventor of nonlinear video editing

https://www.theinformation.com/briefings/ai-startup-midjourney-expects-200-million-in-revenue

Briefing

AI Startup Midjourney Expects $200 Million in Revenue

By Kate Clark

2 years ago

Source: The Information

Midjourney, a generative artificial intelligence startup that charges users between $10 and $120 for a monthly subscription, is on pace to surpass $200 million in revenue this year, according to The Information's reporting. The company, which is built atop the gaming chat app Discord, is profitable and has not raised any money from venture capital firms since it was founded two years ago.

Notably, Midjourney shares profits with its employees and gives a slice of its revenues to Discord, which has helped the AI startup scale.

Midjourney founder David Holz said in an interview with The Informaiton that his goal is to build a company "kind of like Craigslist," the classifieds company that was famously bootsrapped. Holz added that Midjourney is "this weird thing that no one knows how to compete with that just sort of stands alone."

Most Popular

The Briefing

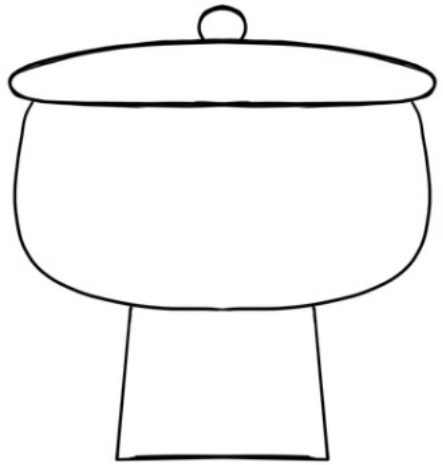
Why the Charter-Cox $34.5 Billion Cable Deal Matters

May 16, 2025

https://www.youtube.com/watch?v=NPJNPrshhTo

Adobe Firefly

How it Works



1. Sketch

Take a picture of your sketch, or draw within Vizcom's studio.

2. Render

Prompt and render your design with one of the default styles, or with your own Palette.

3. 3D Model

Generate a 3d model of your rendering to view your idea from a new angle. Export the model to view in AR or to 3d print.

https://www.youtube.com/watch?v=-SlyCVBrn9c

Vizcom AI

https://www.youtube.com/watch?v=lnUbEPFlgKA&list=PLLqzaOTf8gCW3SIojJOZE_89kIPEXDNBg&index=9

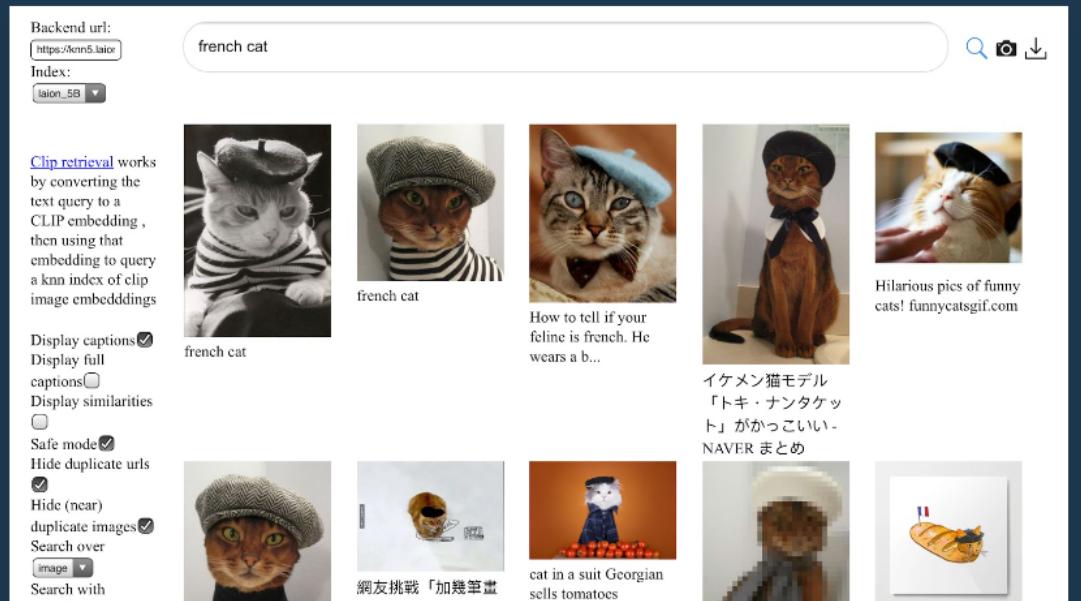

https://www.youtube.com/watch?v=TbYZoMocpM8
CREATED BY
THE DOR
BROTHERS

LAION-5B: A NEW ERA OF OPEN LARGE-SCALE MULTI-MODAL DATASETS

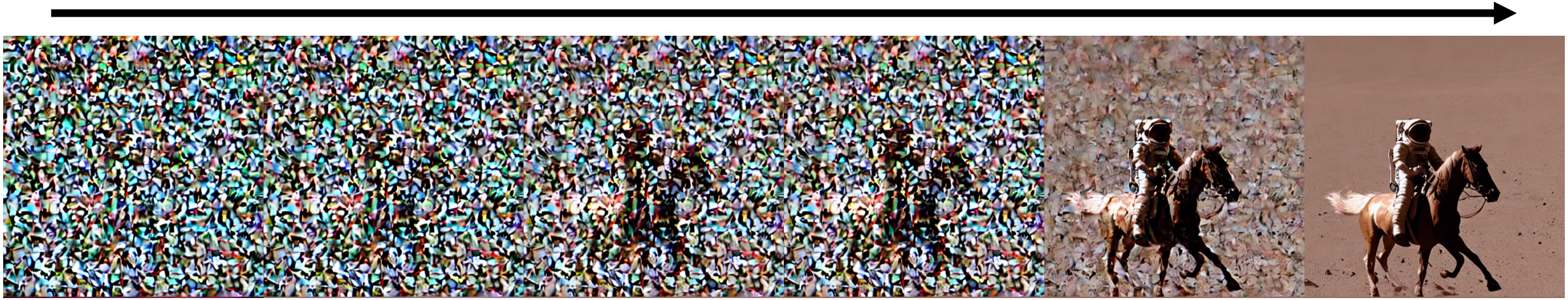
by: Romain Beaumont, 31 Mar, 2022

We present a dataset o 5,85 billion CLIP-filtered image-text pairs, 14x bigger than LAION-400M, previously the biggest openly accessible image-text dataset in the world - see also our NeurIPS2022 paper

Authors: Christoph Schuhmann, Richard Vencu, Romain Beaumont, Theo Coombes, Cade Gordon, Aarush Katta, Robert Kaczmarczyk, Jenia Jitsev



Diffusion Models

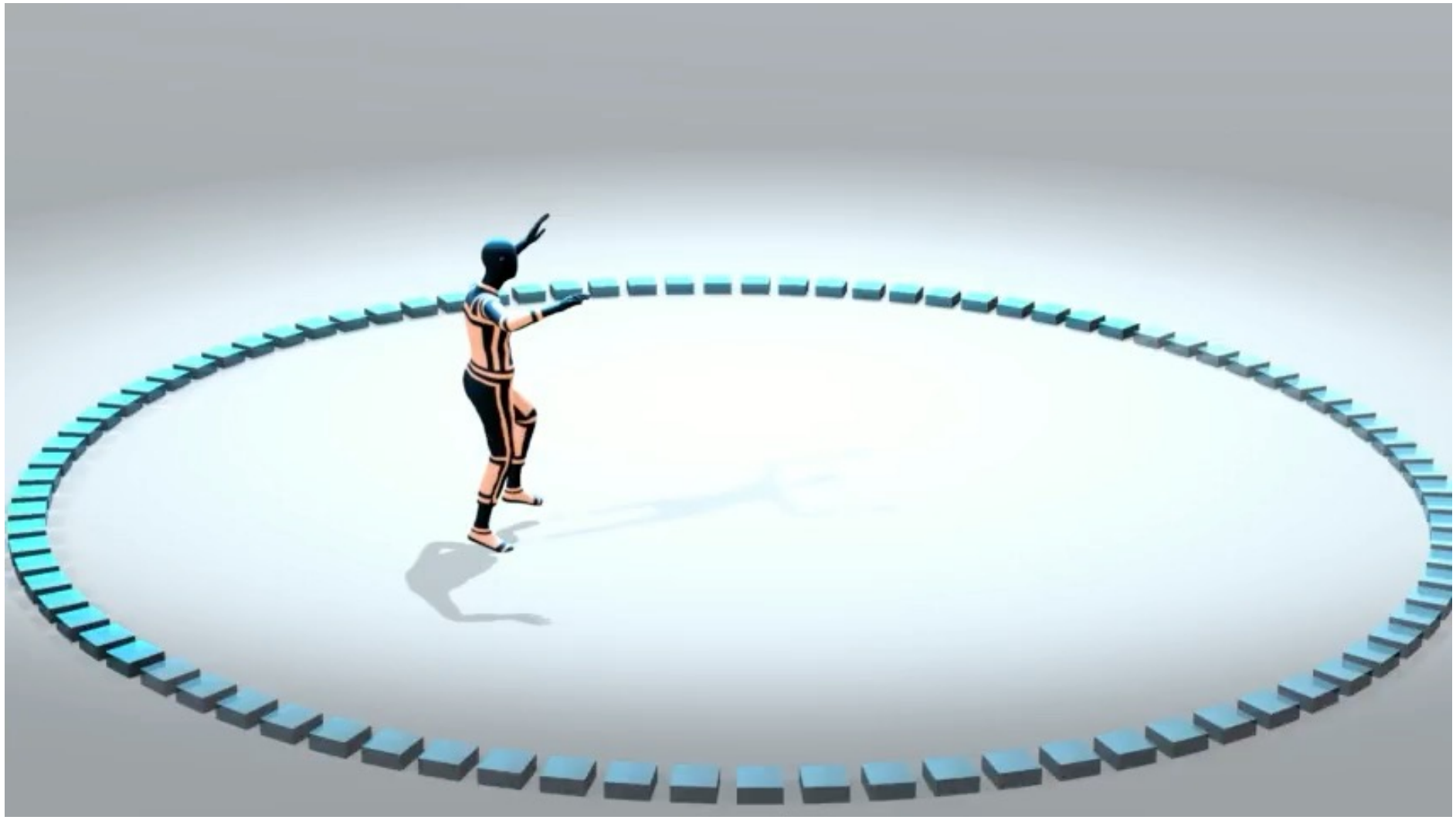


https://audioldm.github.io/audioldm2/

Audio Generation

Alexanderson et al., Listen, Denoise, Action! Audio-Driven Motion Synthesis with Diffusion Models, SIGGRAPH 2023.

Motion Generation



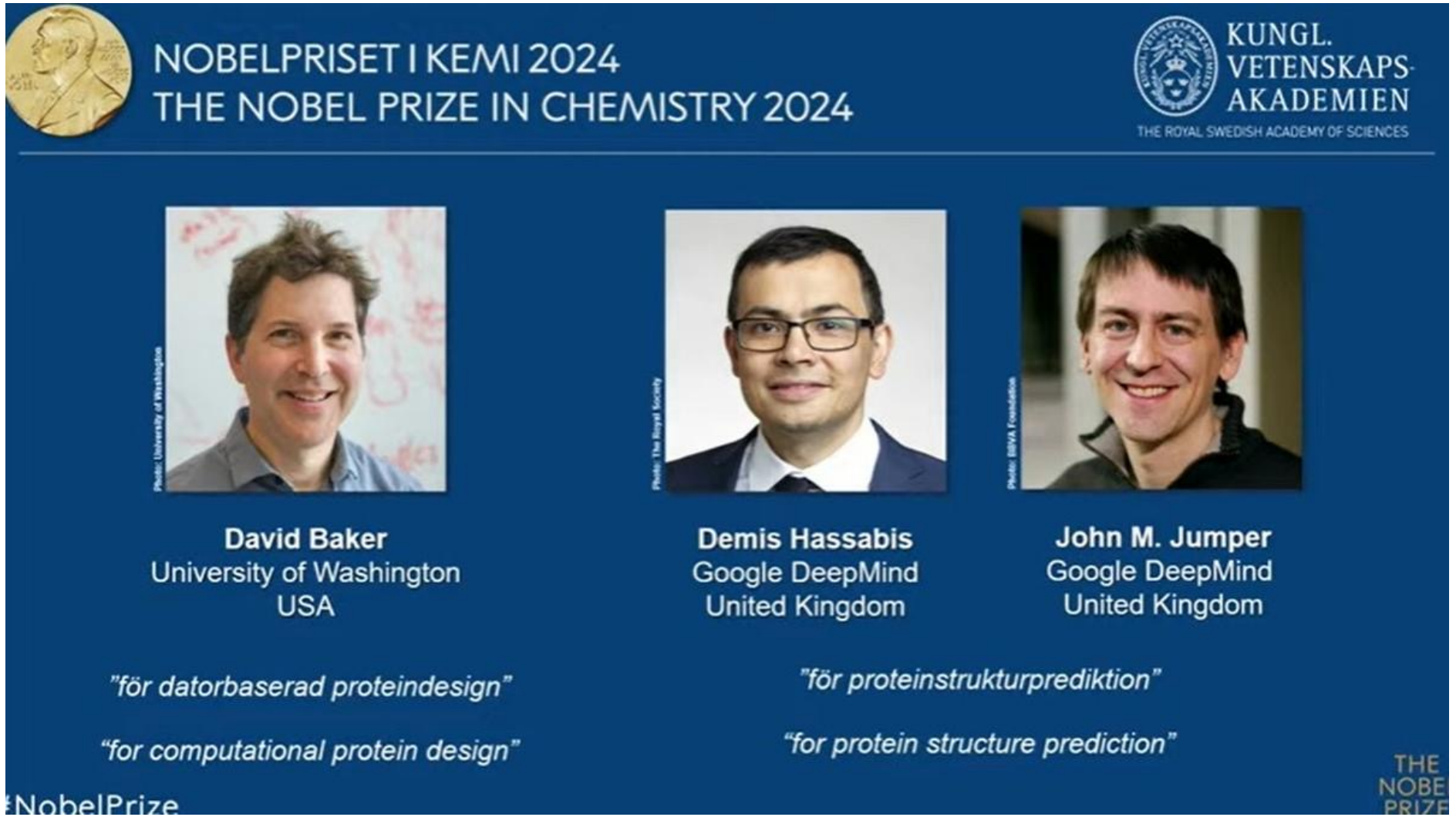
https://www.generativeaipub.com/p/metas-new-3d-gen-can-generate-3d

3D Generation

Meta 3D Gen

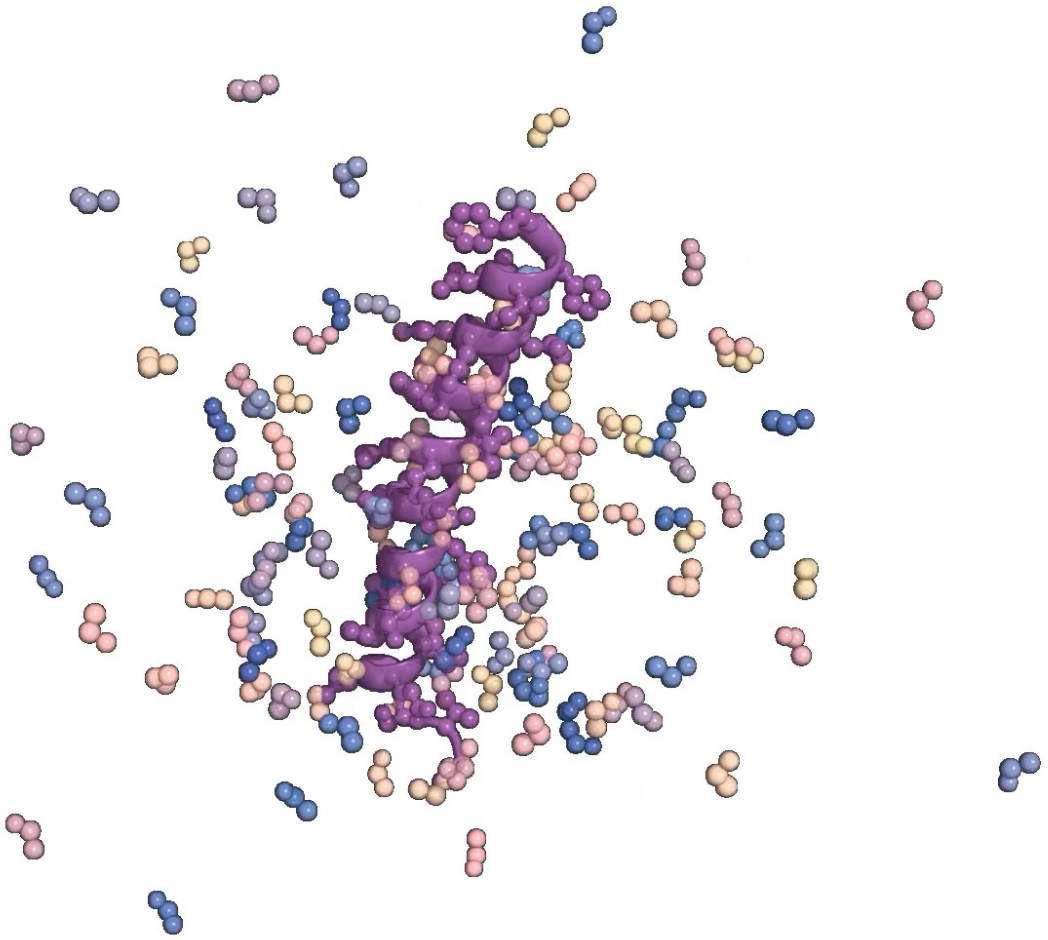
https://www.bakerlab.org/2023/07/11/diffusion-model-for-protein-design/

Molecule Generation



https://www.bakerlab.org/2023/07/11/diffusion-model-for-protein-design/

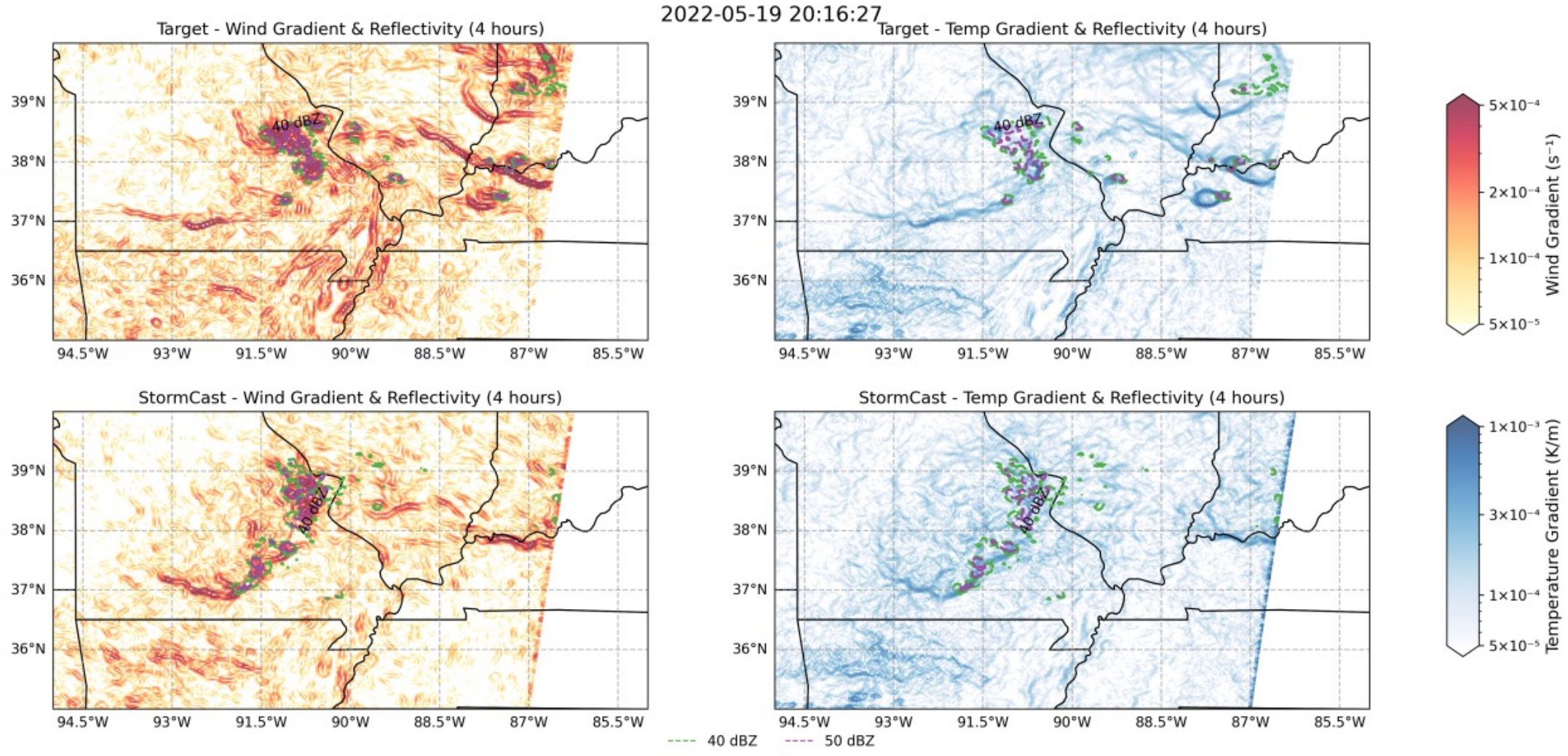
Molecule Generation



RFdiffusion

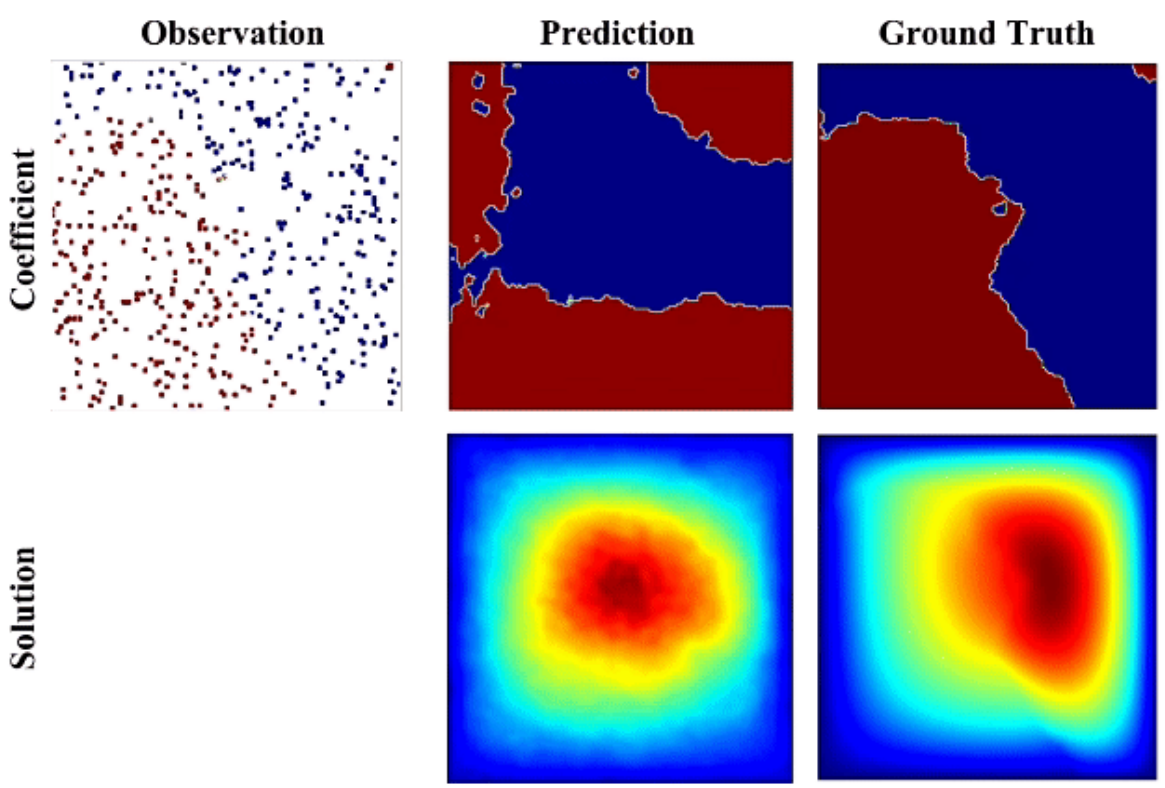
https://research.nvidia.com/publication/2024-08_kilometer-scale-convection-allowing-model-emulation-using-generative-diffusion

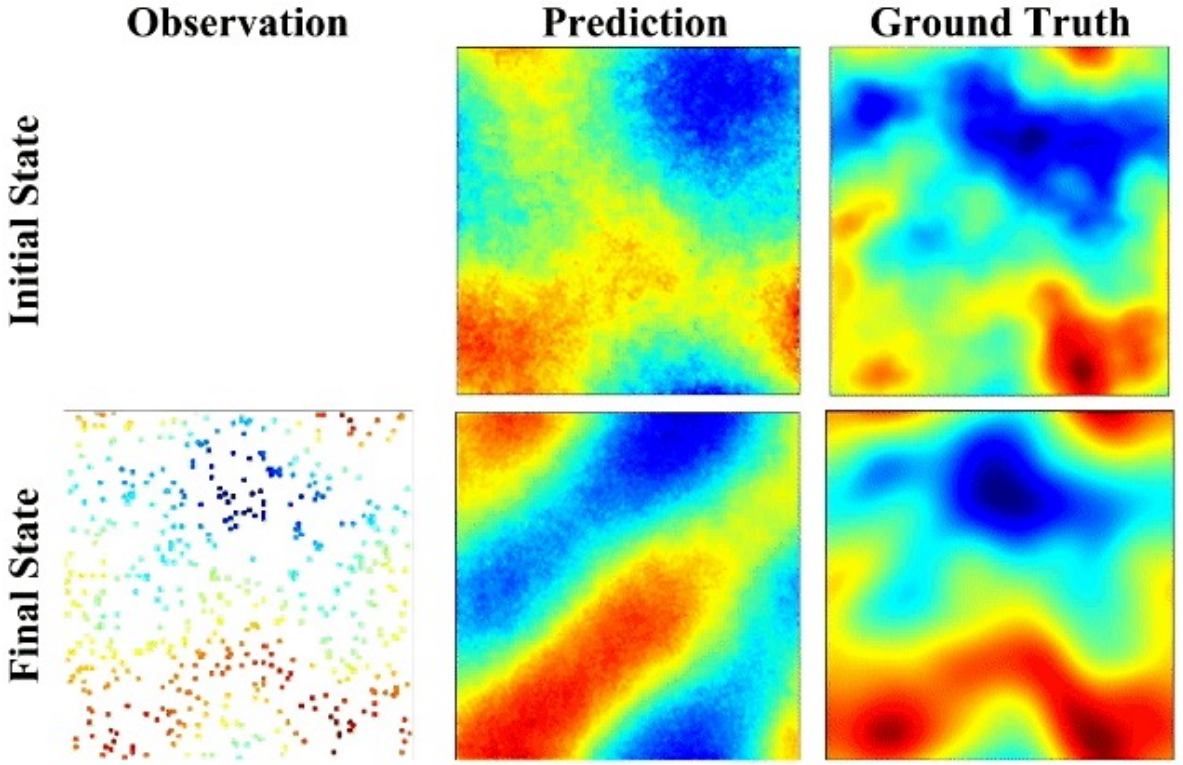
Weather Forecasting



https://jhhuang.site/Diffusion-PDE/

AI for Science – PDE Solving





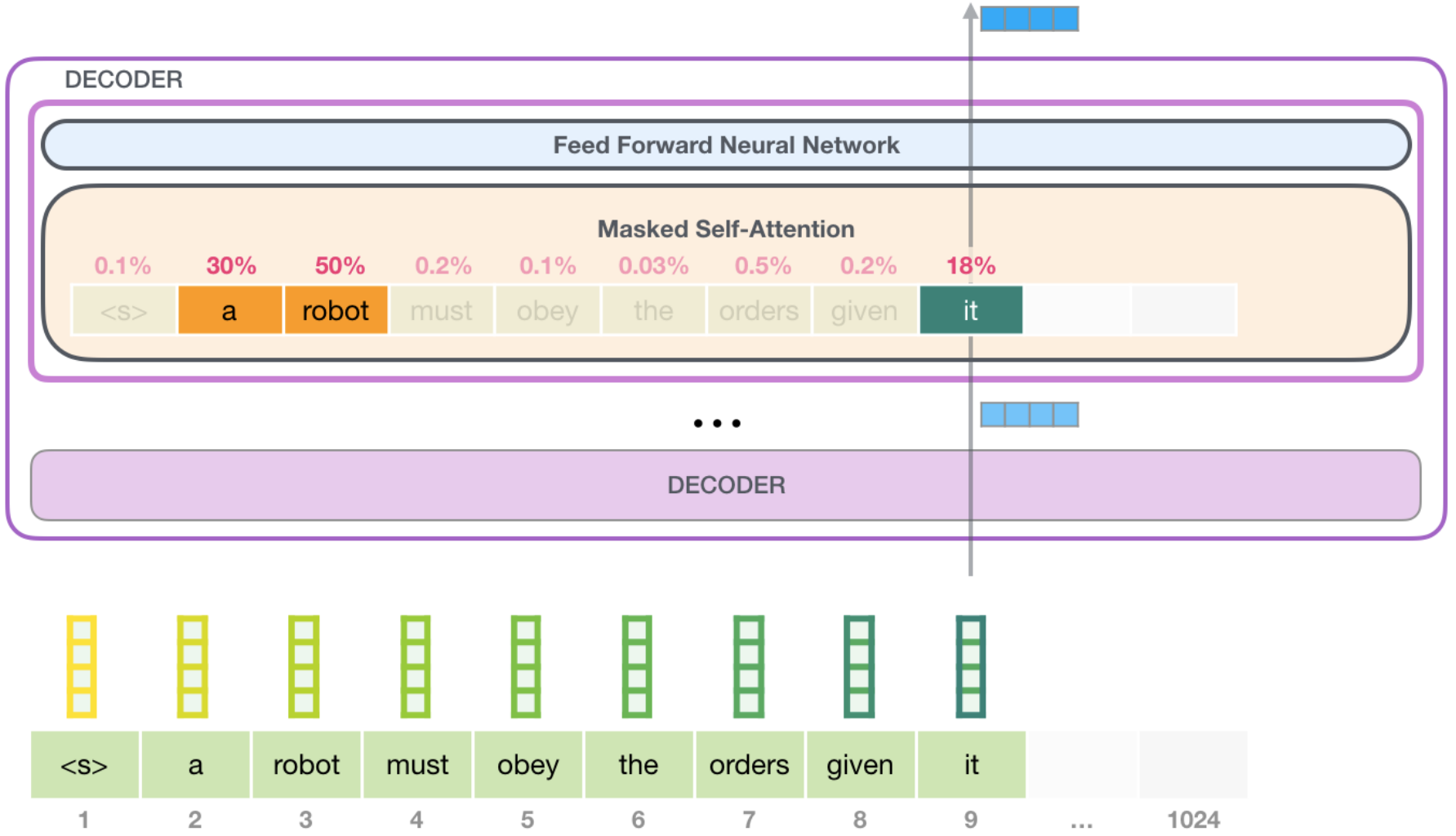
https://towardsdatascience.com/the-art-of-tokenization-breaking-down-text-for-ai-43c7bccaed25/

How About Language Models?

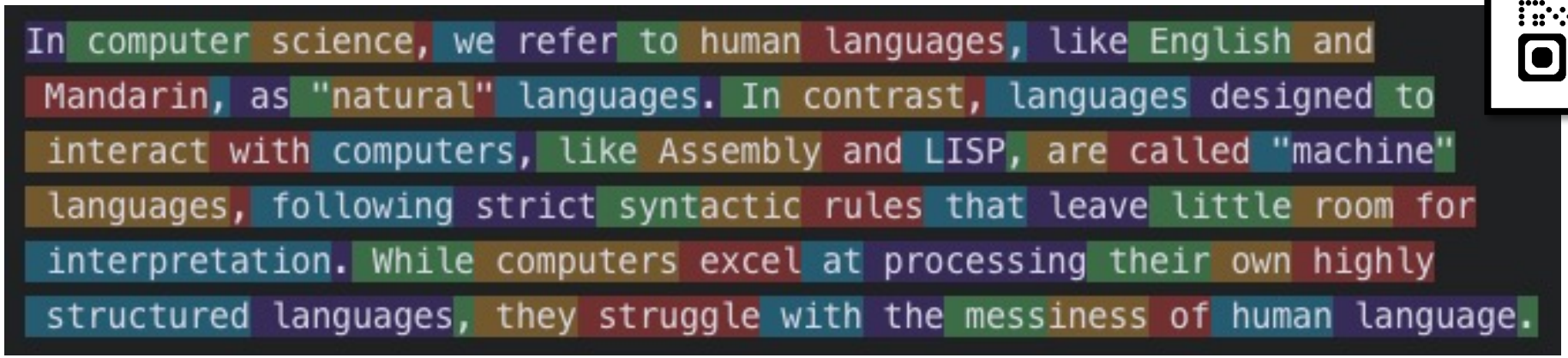
Large Language Models (LLMs), such as GPT-3 and GPT-4, util
ize a process called tokenization. Tokenization involves br
eaking down text into smaller units, known as tokens, which
the model can process and understand. These tokens can rang
e from individual characters to entire words or even larger
chunks, depending on the model. For GPT-3 and GPT-4, a Byte
Pair Encoding (BPE) tokenizer is used. BPE is a subword tok
enization technique that allows the model to dynamically bu
ild a vocabulary during training, efficiently representing
common words and word fragments. Although the core tokeniza
tion process remains similar across different versions of t
hese models, the specific implementation can vary based on
the model's architecture and training objectives.

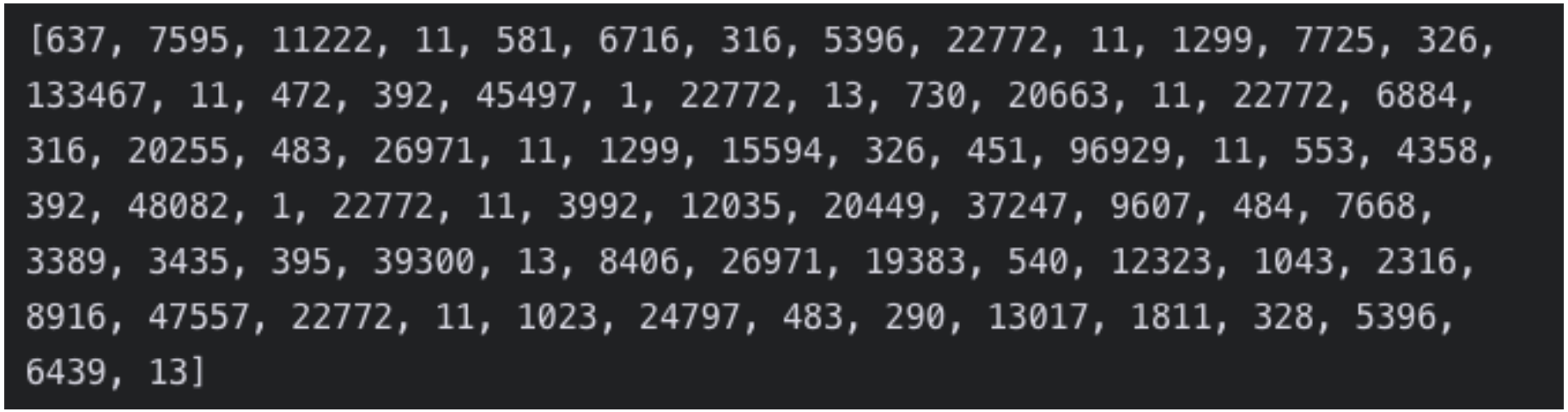
https://jalammar.github.io/illustrated-gpt2/

Autoregressive Models



Tokenization





Try yourself: https://platform.openai.com/tokenizer

https://jalammar.github.io/illustrated-gpt2/

Learning Categorical Distributions

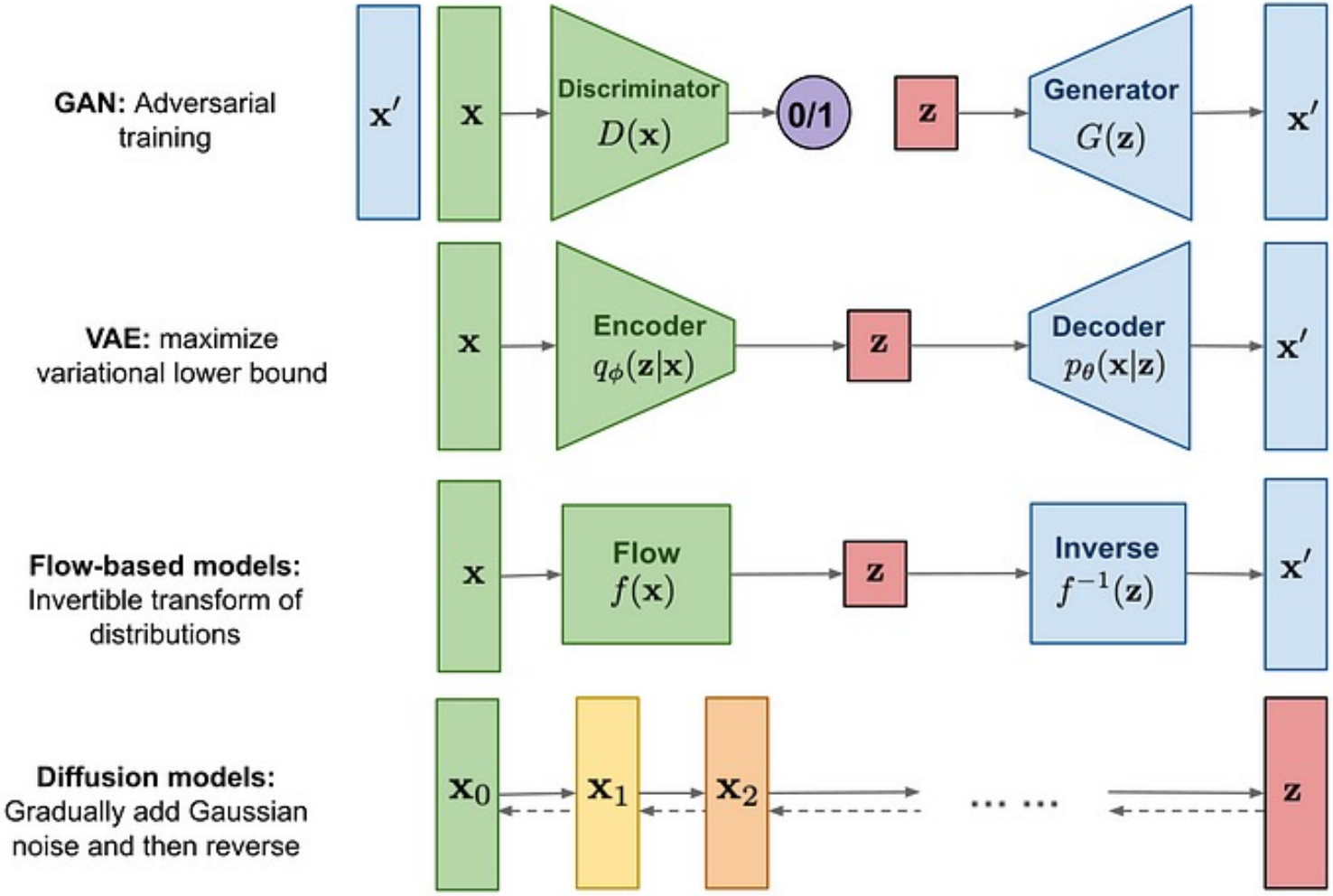
Word	Value vector	Score	Value X Score
<s>		0.001	
a		0.3	
robot		0.5	
must		0.002	
obey		0.001	
the		0.0003	
orders		0.005	
given		0.002	
it		0.19	
		Sum:	

Autoressive Models vs. Diffusion Models

- **Autoregressive models** are suitable for sequential discrete data.
- **Diffusion models** are suitable for continuous data in high-dimensional spaces.

https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Generative Models

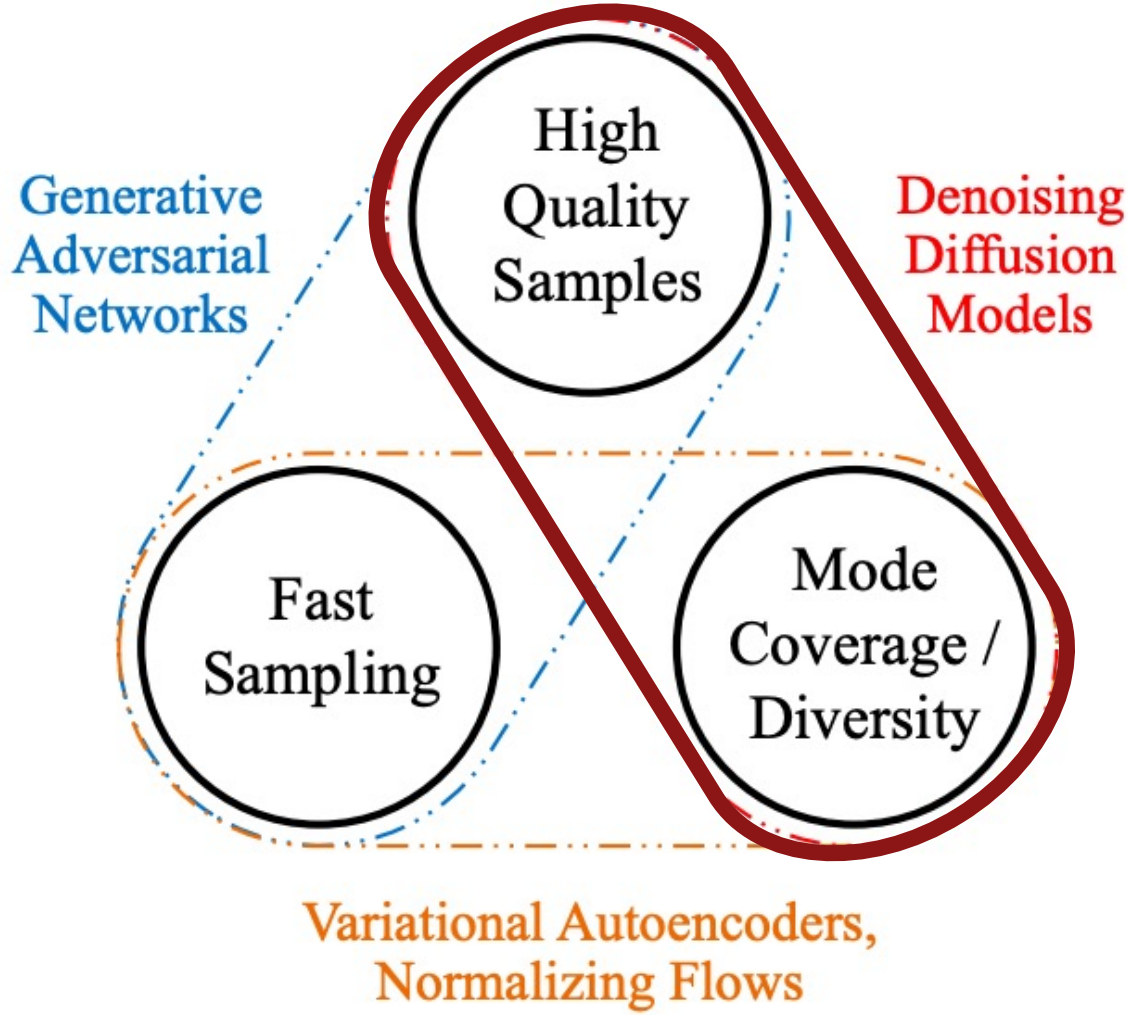


Kerras et al., Analyzing and Improving the Image Quality of StyleGAN.

StyleGAN2

https://developer.nvidia.com/blog/improving-diffusion-models-as-an-alternative-to-gans-part-1/

Generative Models – Comparison



Higher Quality & Diversity

Diffusion Models Beat GANs on Image Synthesis

Prafulla Dhariwal*
OpenAI
prafulla@openai.com

Alex Nichol*
OpenAI
alex@openai.com

Abstract

We show that diffusion models can achieve image sample quality superior to the current state-of-the-art generative models. We achieve this on unconditional image synthesis by finding a better architecture through a series of ablations. For conditional image synthesis, we further improve sample quality with classifier guidance: a simple, compute-efficient method for trading off diversity for fidelity using gradients from a classifier. We achieve an FID of 2.97 on ImageNet 128×128, 4.59 on ImageNet 256×256, and 7.72 on ImageNet 512×512, and we match BigGAN-deep even with as few as 25 forward passes per sample, all while maintaining better coverage of the distribution. Finally, we find that classifier guidance combines well with upsampling diffusion models, further improving FID to 3.94 on ImageNet 256×256 and 3.85 on ImageNet 512×512. We release our code at https://github.com/openai/guided-diffusion.

Diffusion Models

- (+) High quality
- (+) Diversity
- (–) Slow

Diffusion Models

The generative process of a diffusion model is an iterative denoising process.

CS492(D): Diffusion Models and Their Applications

Minhyuk Sung, **KAIST**, Fall 2024

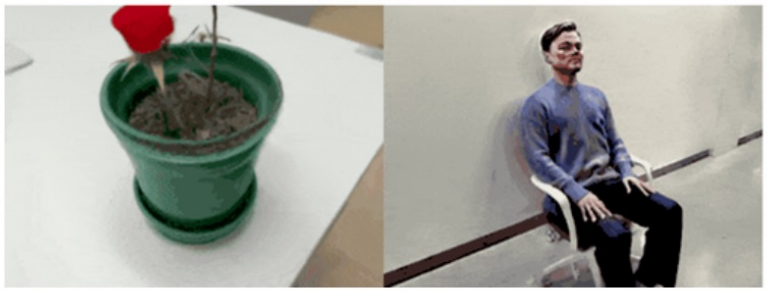


Table of contents

https://pixabay.com/photos/pictures-photos-photo-collection-382009/

Let’s Consider a Collection of Real Photos.

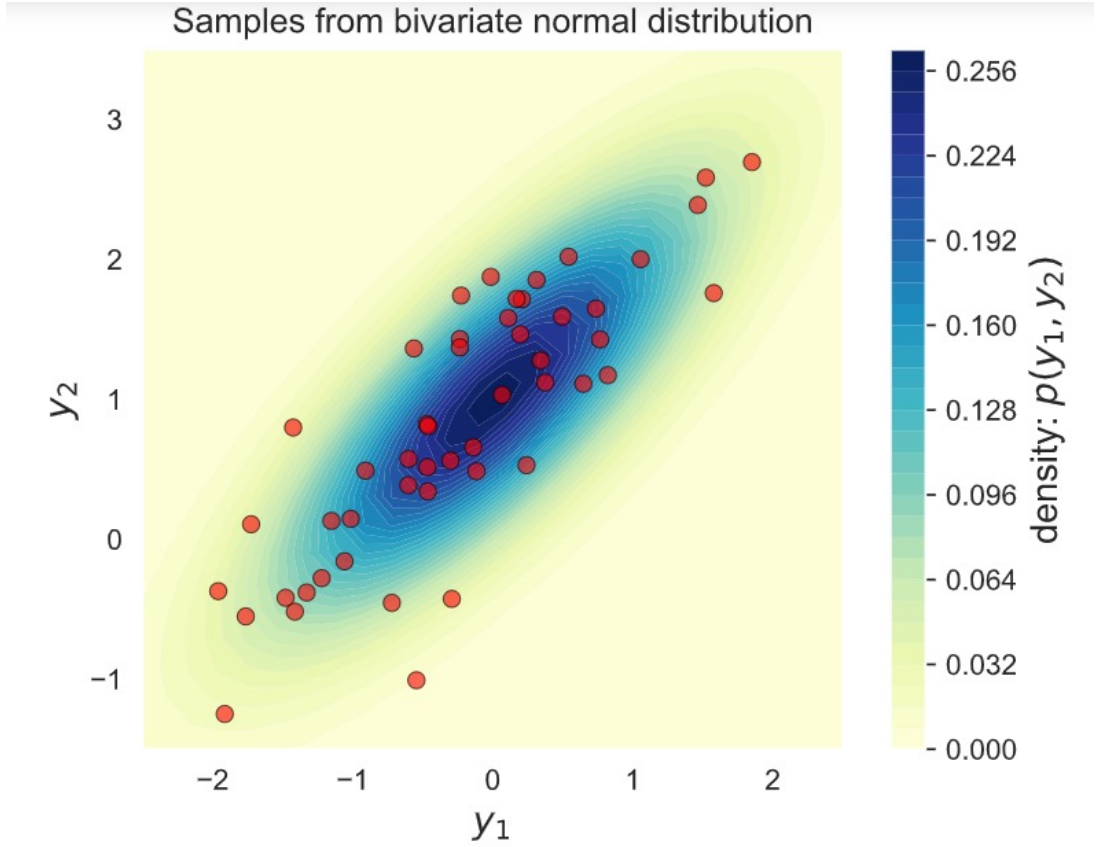
https://pixabay.com/photos/pictures-photos-photo-collection-382009/

How to Generate a New Photo?

A Simpler Example

Let's consider a collection of 2D points.

How can we sample a new 2D point?



A Simpler Example

The 2D points are samples from a specific probability distribution.

If the distribution has a specific form (e.g., Gaussian), we can sample from it directly.

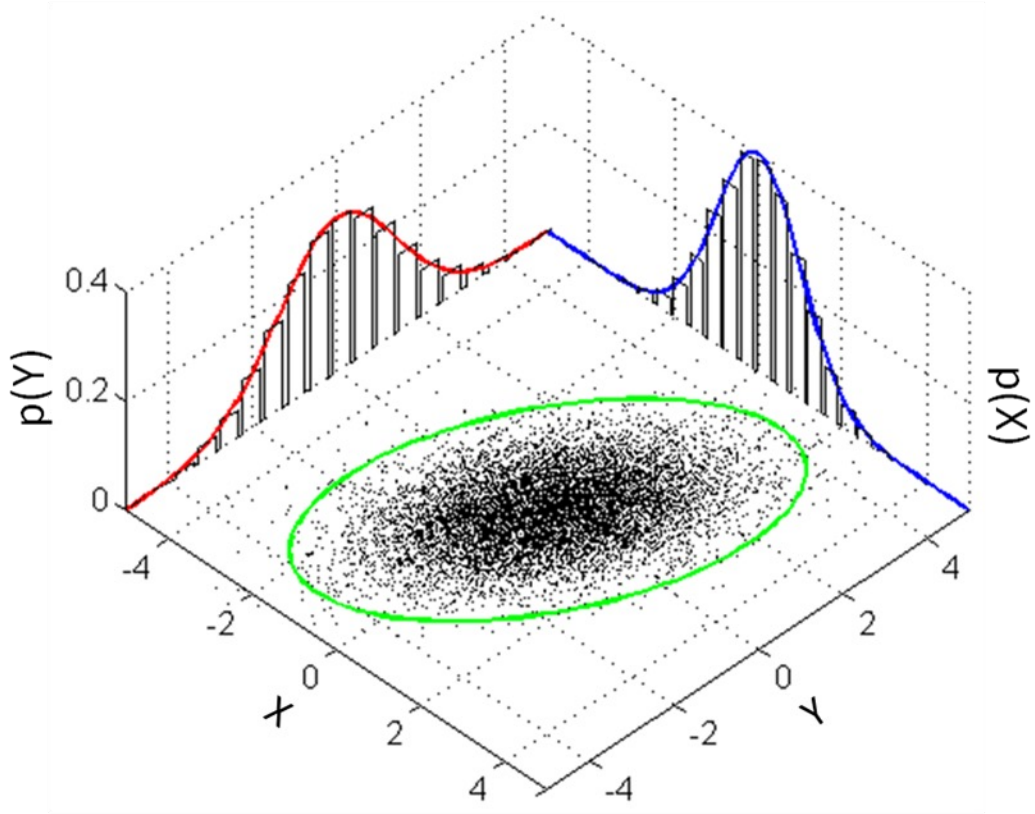
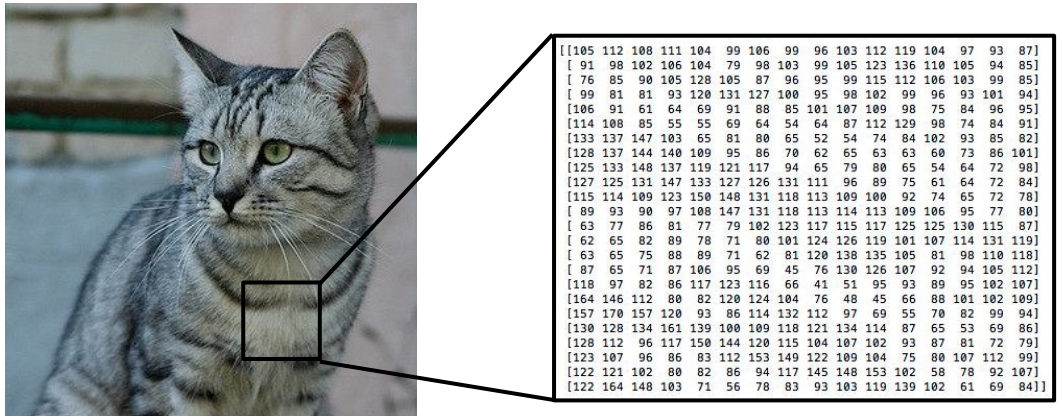


Image from Stanford CS231N

Statistical Perspective for Real Images

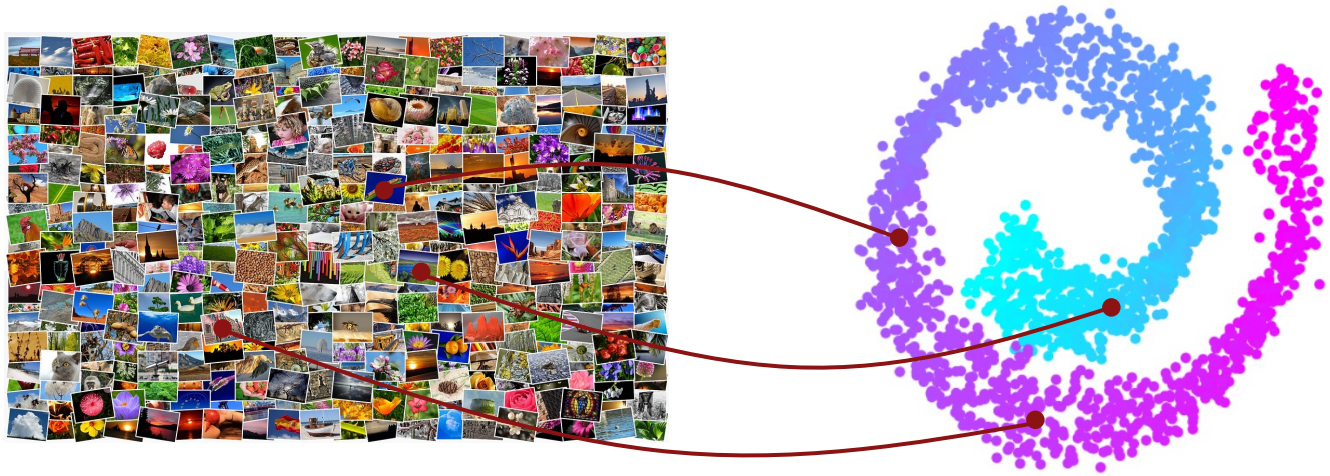
- Let's consider RGB images with a resolution of 256×256.
- An image can be represented by a $256 \times 256 \times 3$ vector.
- This means that an image is a point in a $256 \times 256 \times 3$-dimensional space.



Statistical Perspective for Real Images

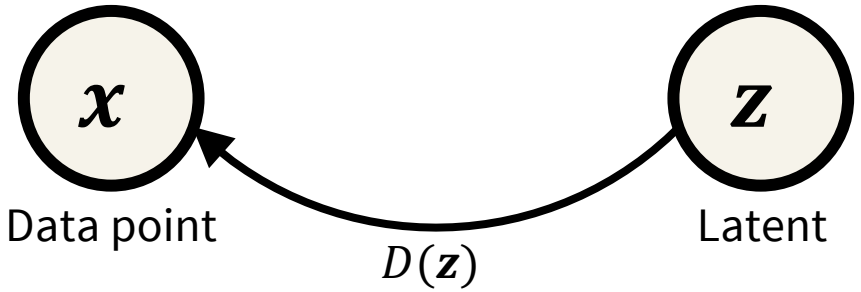
- Images are sample points in a high-dimensional space.
- Can we derive the PDF of the data distribution from the samples?

  No, we don't know the distribution; we only have samples.



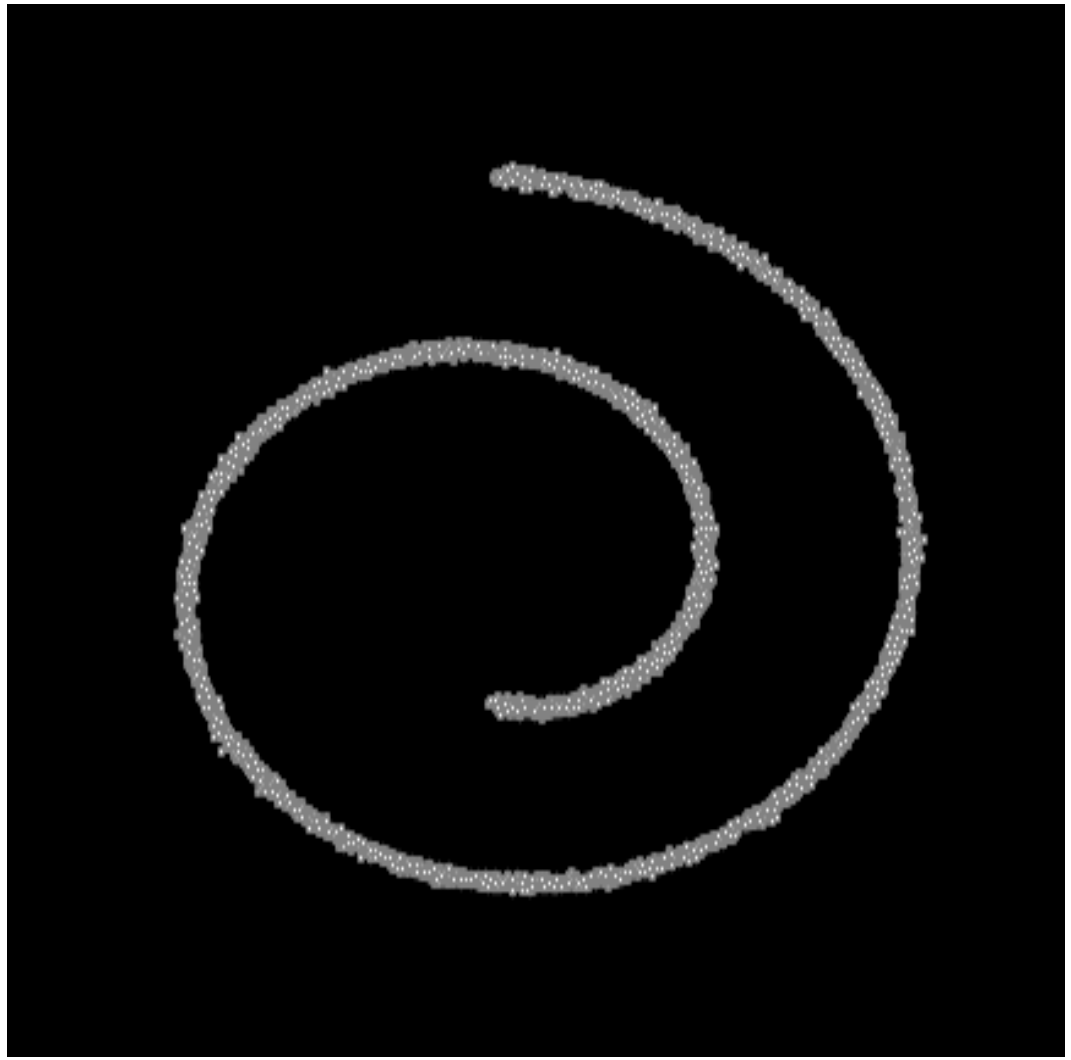
The Basic Idea

- Map a simple distribution which PDF is known $p(\boldsymbol{z})$ (e.g., a standard Gaussian distribution $\mathcal{N}(\boldsymbol{x}; \mathbf{0}, \mathbf{I})$) to the data distribution $p(\boldsymbol{x})$.
- Sample from $p(\boldsymbol{z})$ and map it to a data point.



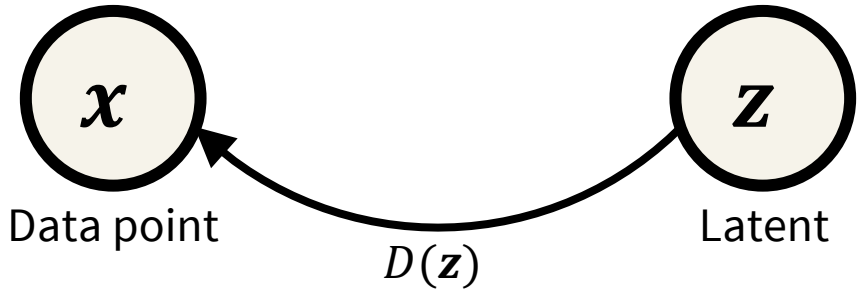
https://github.com/azad-academy/denoising-diffusion-model

The Basic Idea



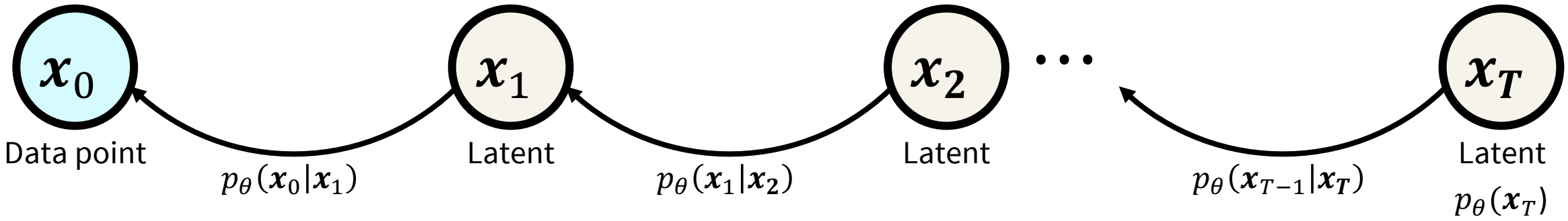
GANs / VAEs

- GANs and VAEs map a latent distribution to the data distribution directly.



Diffusion Models

- Diffusion models map a latent distribution to the data distribution in a sequential, iterative manner.



Denoising Process

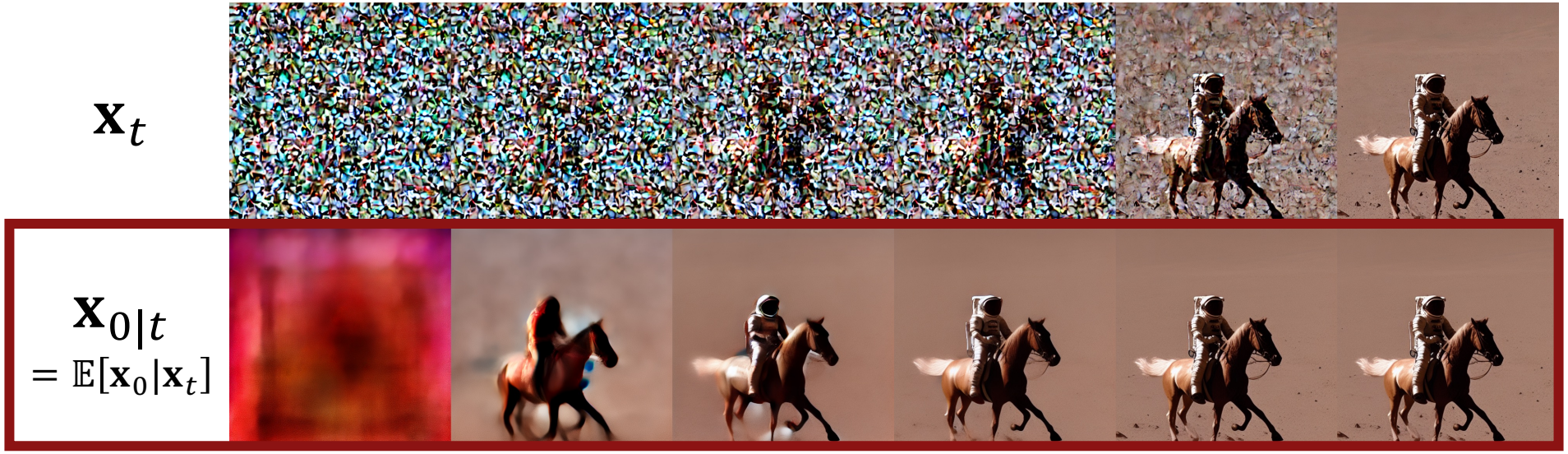
- Diffusion models map a latent distribution to the data distribution in a sequential, iterative manner.
- This iterative process can be seen as a denoising process.

$\mathbf{x}_t$

Bradley Efron, Tweedie's Formula and Selection Bias, 2011.

Refinement Process

- At each step, the expected final output can be estimated easily.
- When viewed from this perspective, the denoising process can also be interpreted as a refinement process.

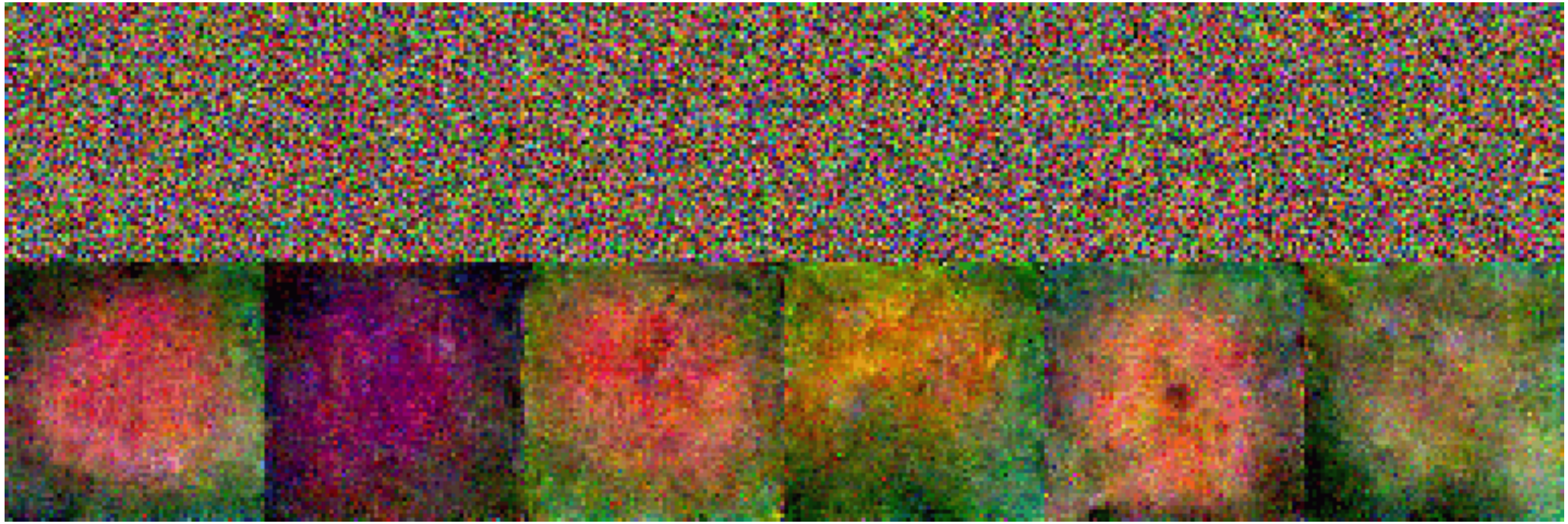


https://keras.io/examples/generative/ddim/

Diffusion Models

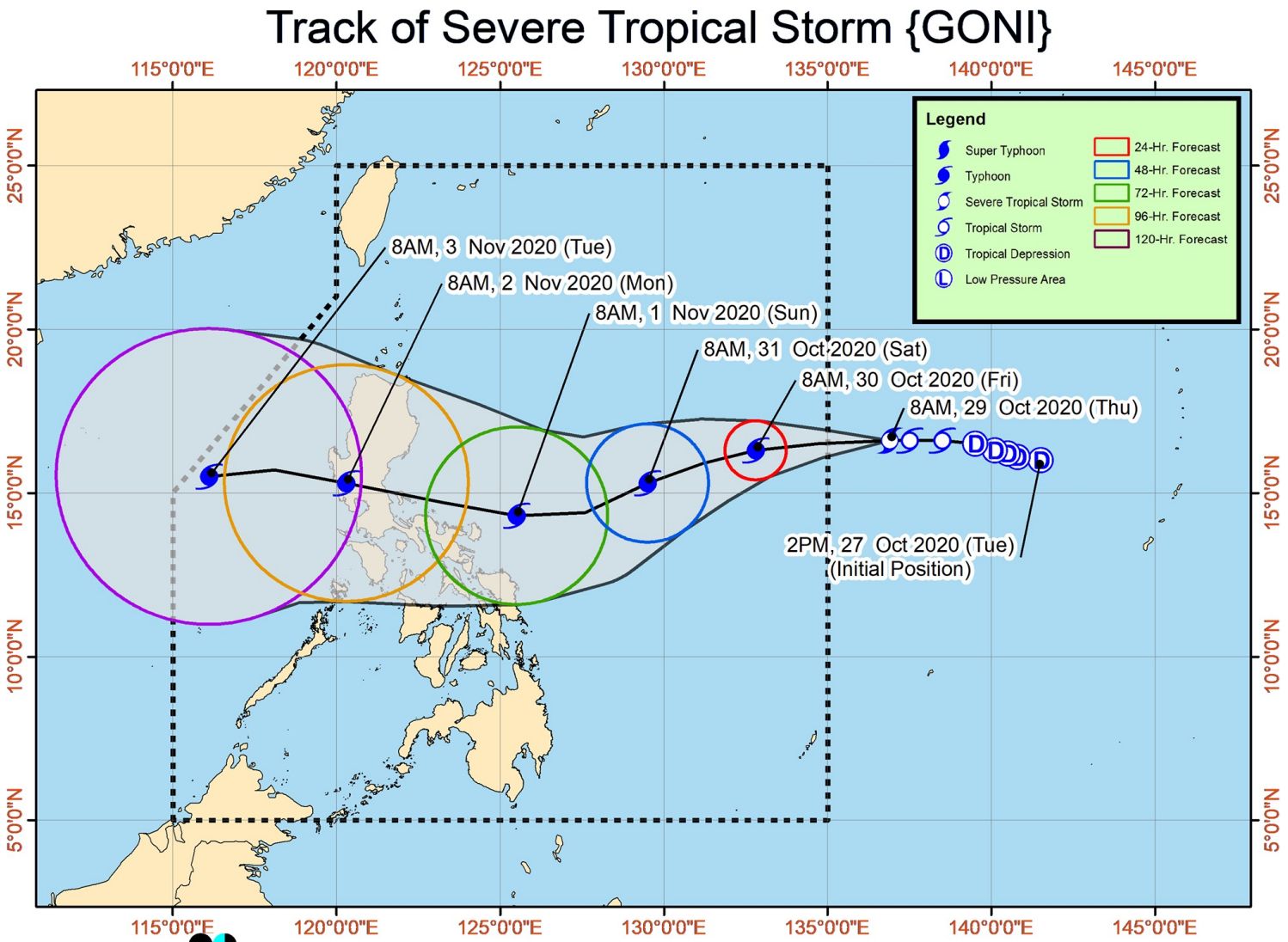
$\mathbf{x}_t$

$\boldsymbol{x}_0$ Prediction



https://www.facebook.com/photo.php?fbid=3273573949419158&id=302746419835274&set=a.302759263167323

Path Prediction



Philippine Atmospheric Geophysical and Astronomical Services Administration

https://www.theladders.com/career-advice/career-path-dream-job

https://www.theladders.com/career-advice/career-path-dream-job

“Begin with the end in mind"

What’s Your Desired Future?

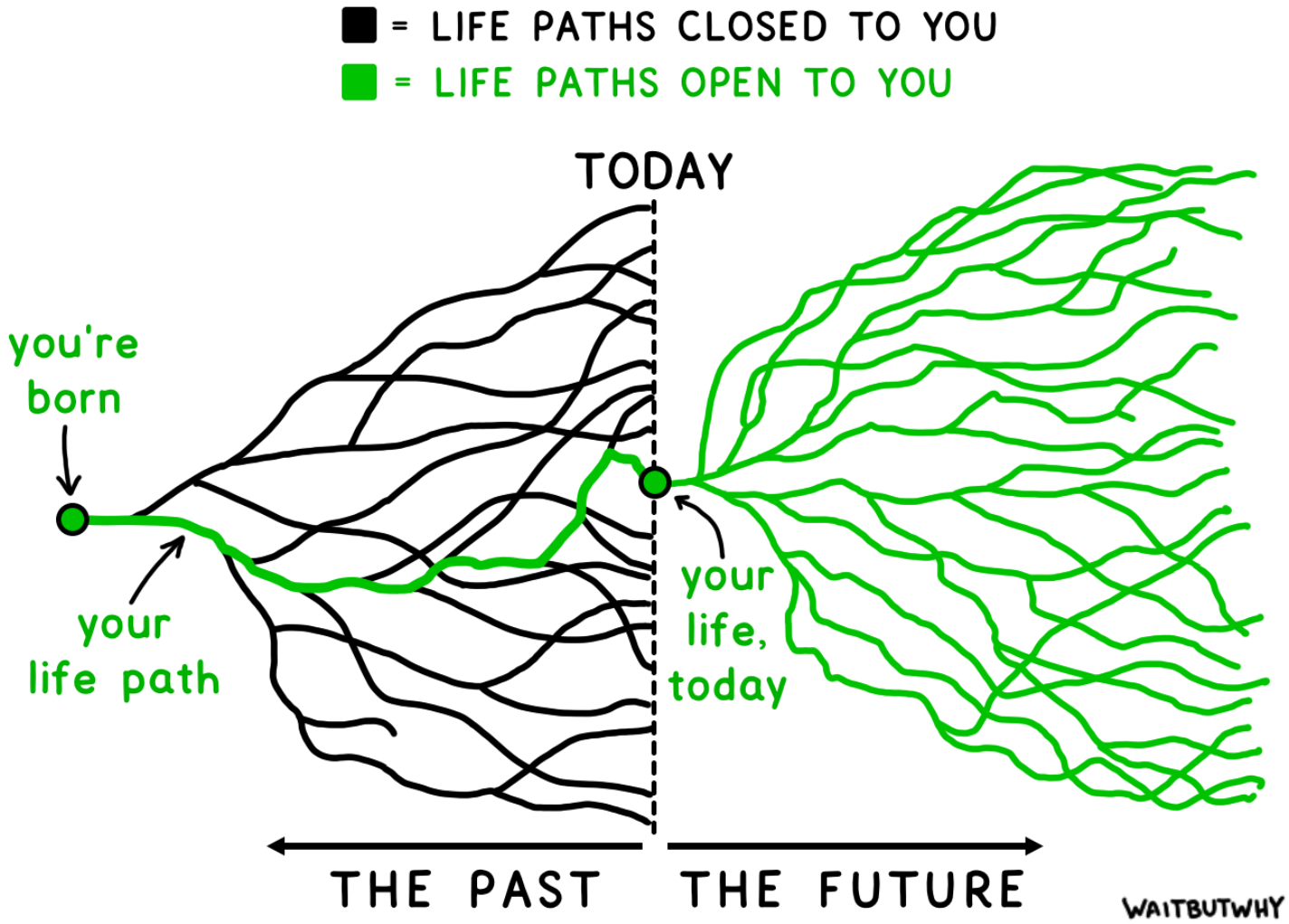
Begin With the End in Mind means to look out across the coming hours, days, weeks, months, and years with a clear vision of your desired destination—and then to proactively pursue that future.

Even if your vision for the future changes, having an end in mind for a meeting, project, or a lifetime gives you more clarity about where to start and how to adapt.

There’s no better time than today to decide where you want to be tomorrow.

Tim Urban. https://waitbutwhy.com/.

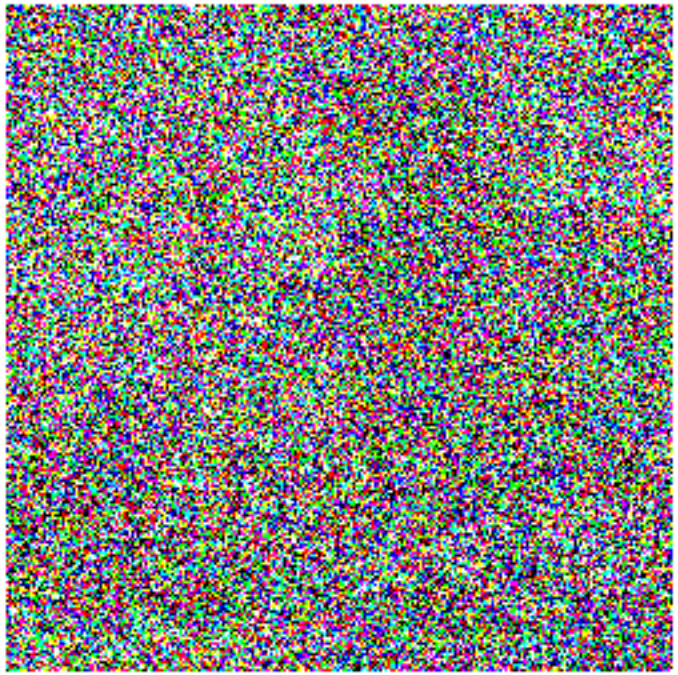
Life Paths

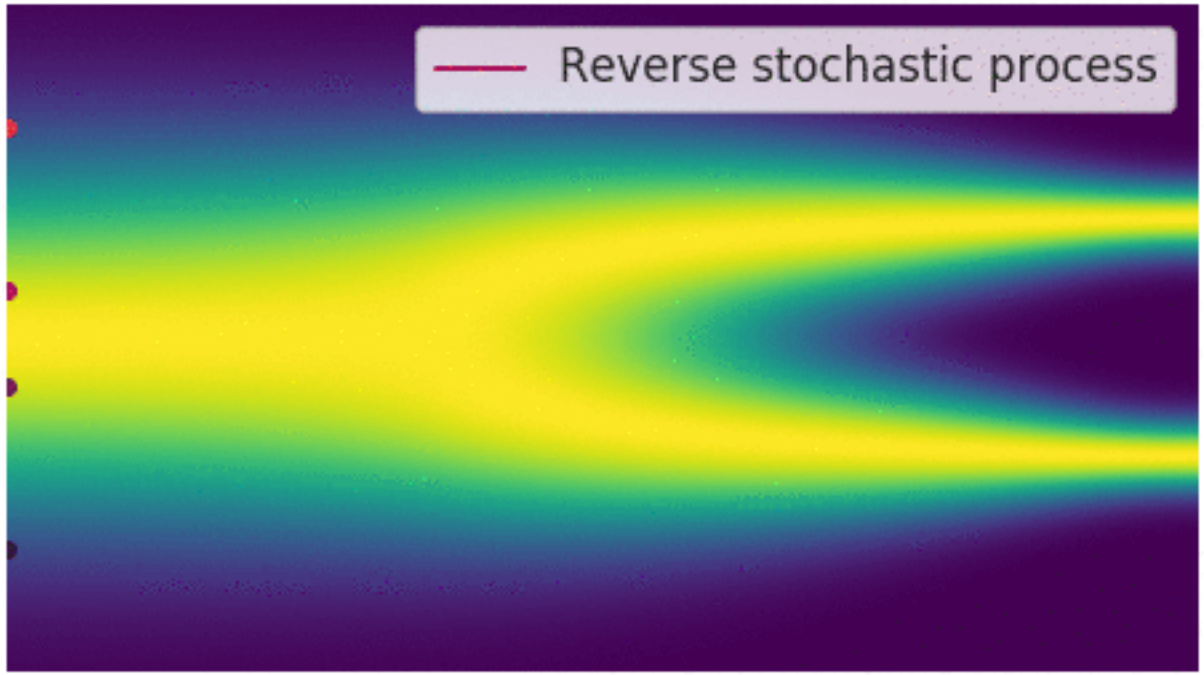


https://yang-song.net/blog/2021/score/

Stochastic Differential Equations

In a continuous-time domain, the mappings are formulated as Stochastic Differential Equations (SDEs).

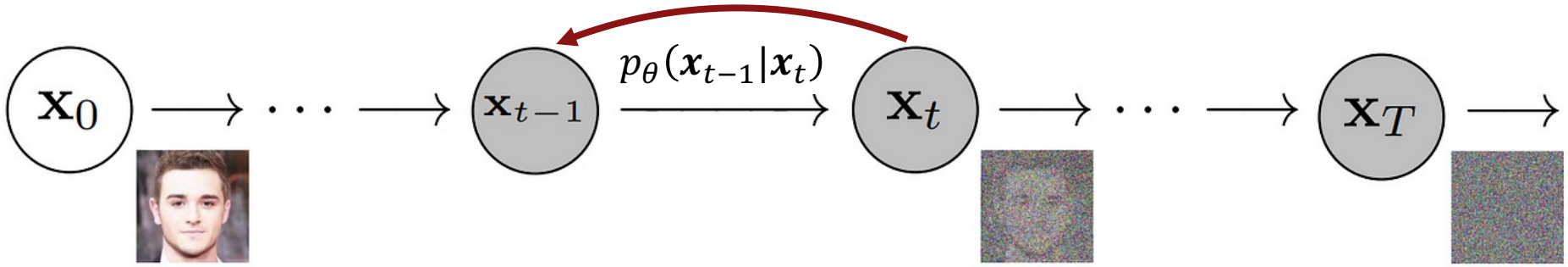




Neural Network as a Mean Predictor

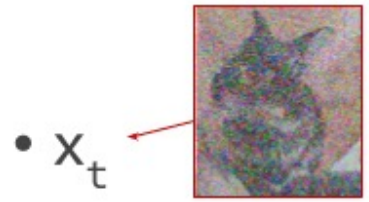
At each step, a neural network predicts the mean of the distribution for the next denoised data point:

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\mu_\theta(\boldsymbol{x}_t, t), \tilde{\sigma}_t^2 \mathbf{I})$$



https://sander.ai/2023/08/28/geometry.html

Stochastic Denoising Process

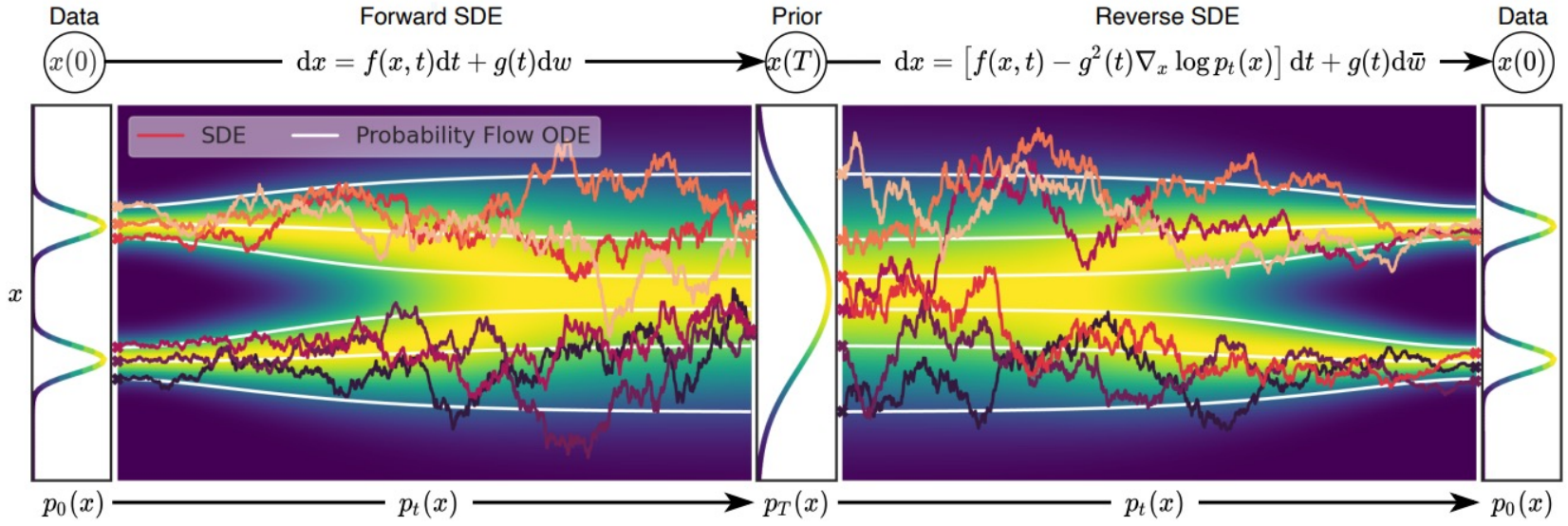


Song et al., Score-Based Generative Modeling through Stochastic Differential Equations, ICLR 2021.

Probability Flow ODE

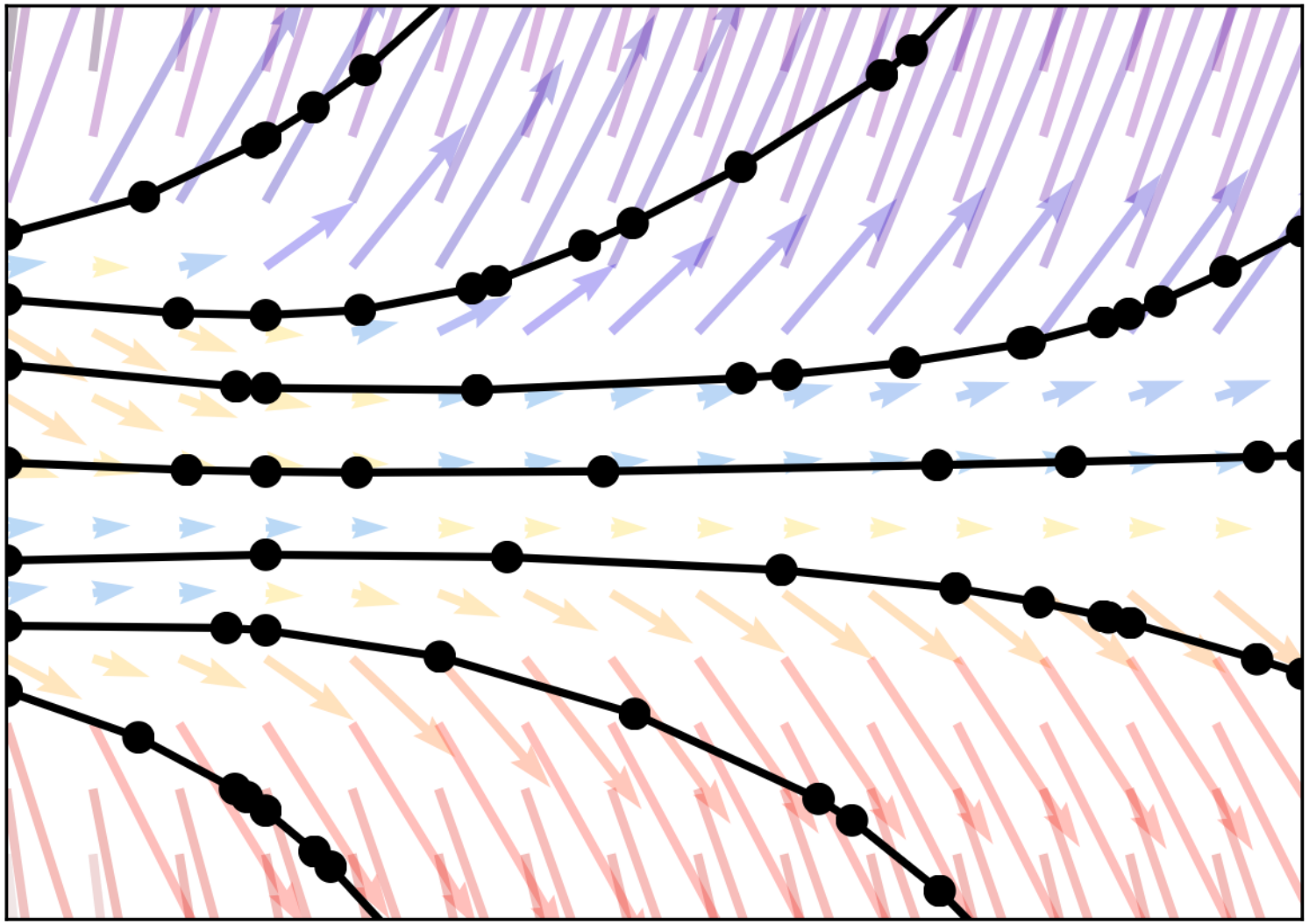
A SDE has a corresponding ODE (no stochatisity) that has same marginal distribution at every time step.

- Red curves: SDE
- White curves: ODE



Chen et al., Neural Ordinary Differential Equations, NeurIPS 2018.

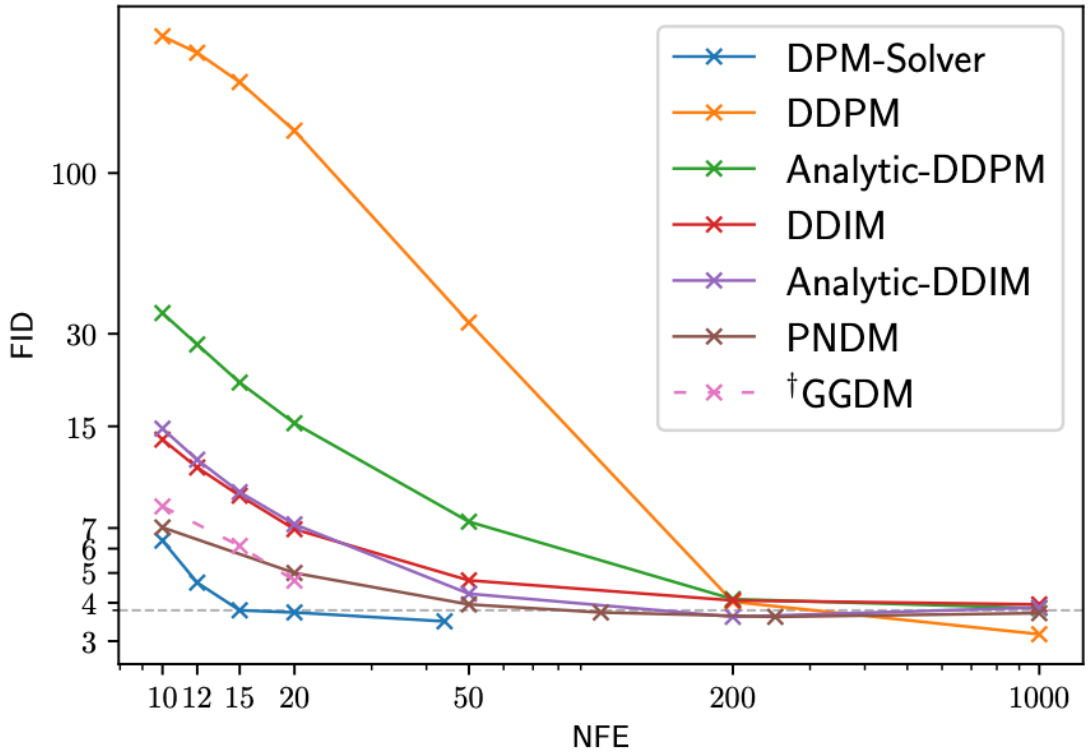
ODE Solvers



Lu et al., DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps, NeurIPS 2022.

ODE Solvers

The ODE can be solved much faster using a different ODE solver at test time, without requiring any fine-tuning of the network.



Diffusion Models

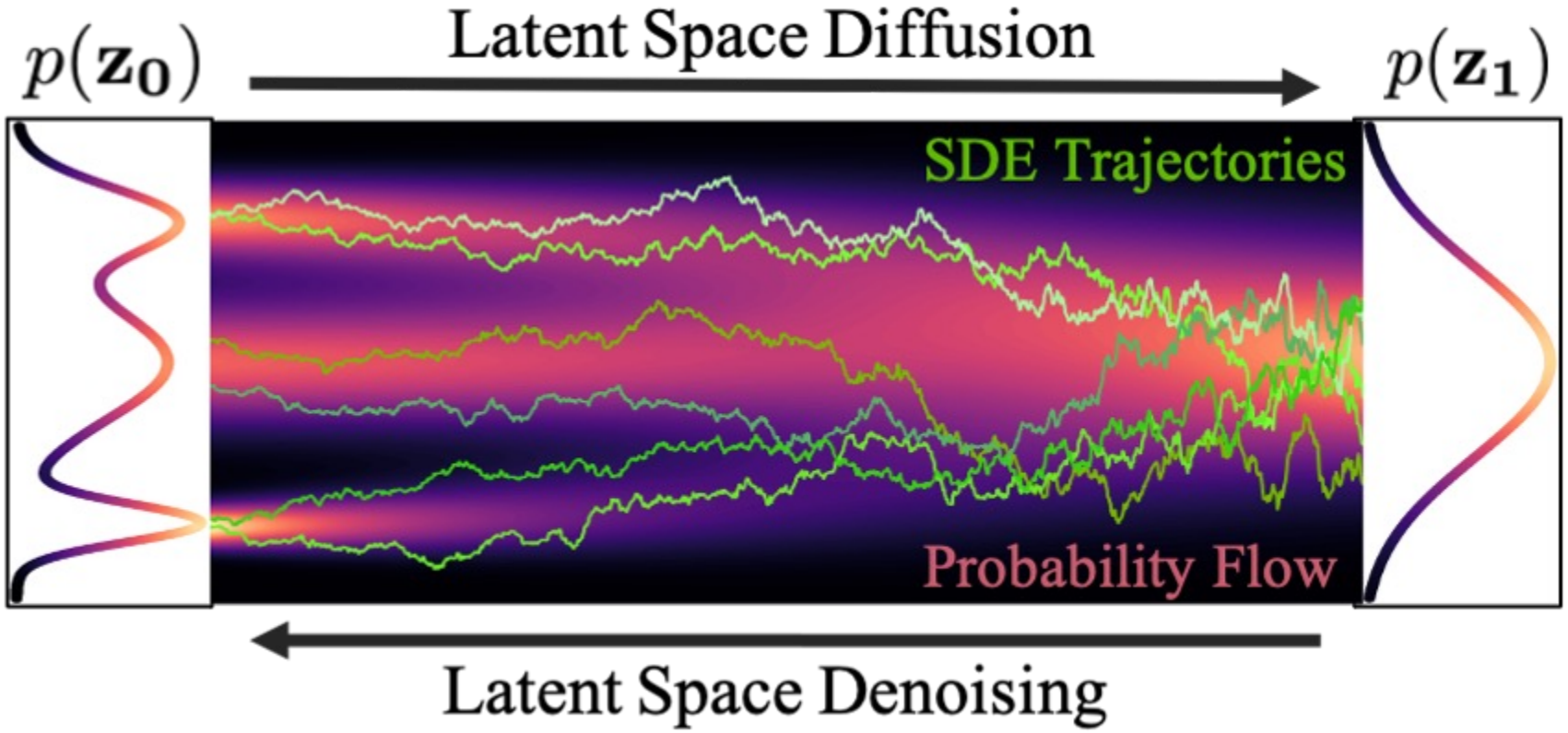
- (+) High quality
- (+) Diversity
- (−) Slow

- (+)(Relatively) easy to implement and train
- (+) Easy to convert a conditional model
- (+) Easy to personalize
- (+) Easy to align to the given reward model
- (+) Easy to distill knowledge
- …

https://yang-song.net/blog/2021/score/

Guided Generation

The denoising process can be guided in multiple different ways.



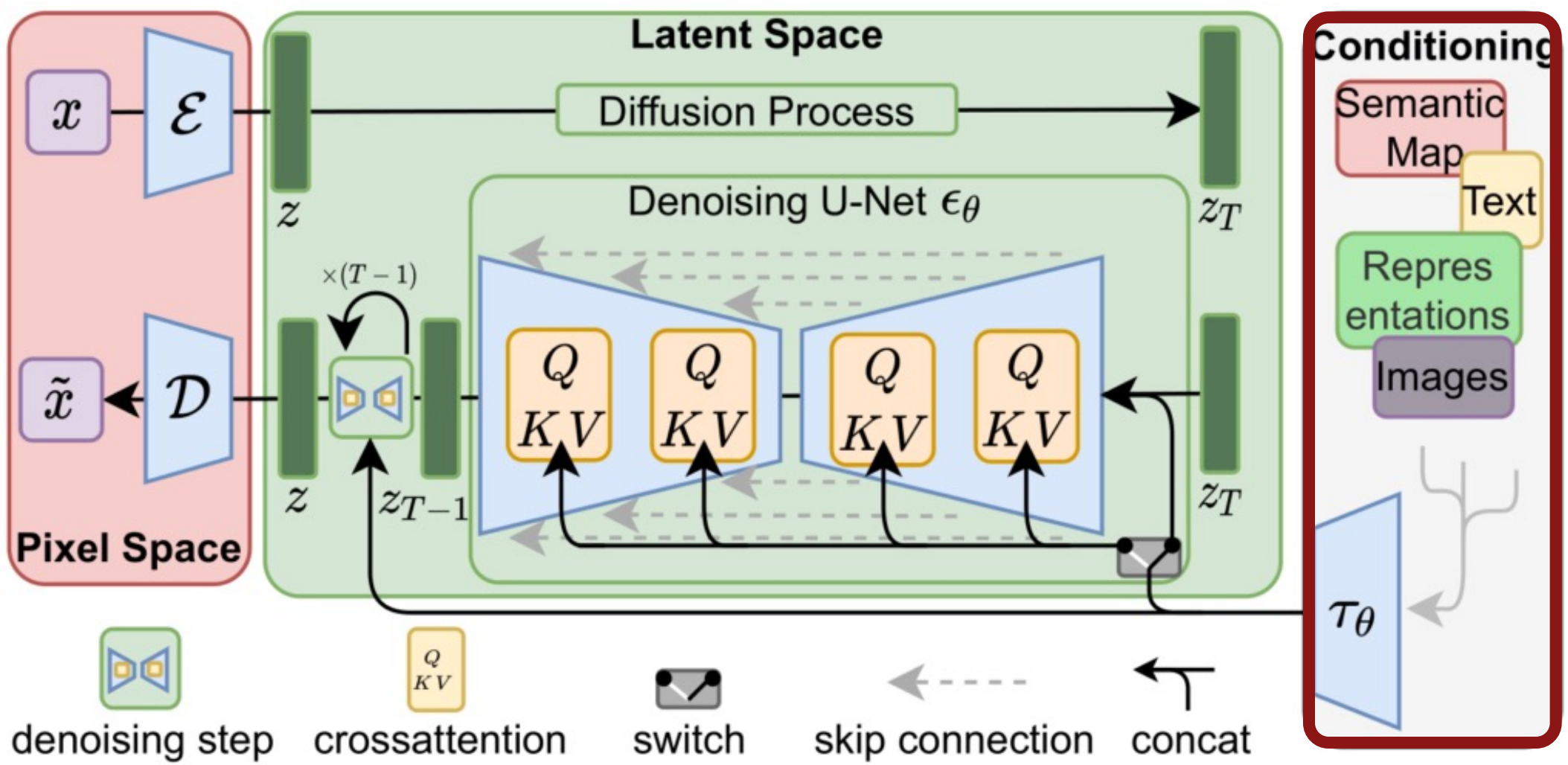
Guided Generation

Three Categories:

1. Manipulation in Neural Network Prediction
2. Guidance Based on the Expected Final Output
3. Particle Filtering

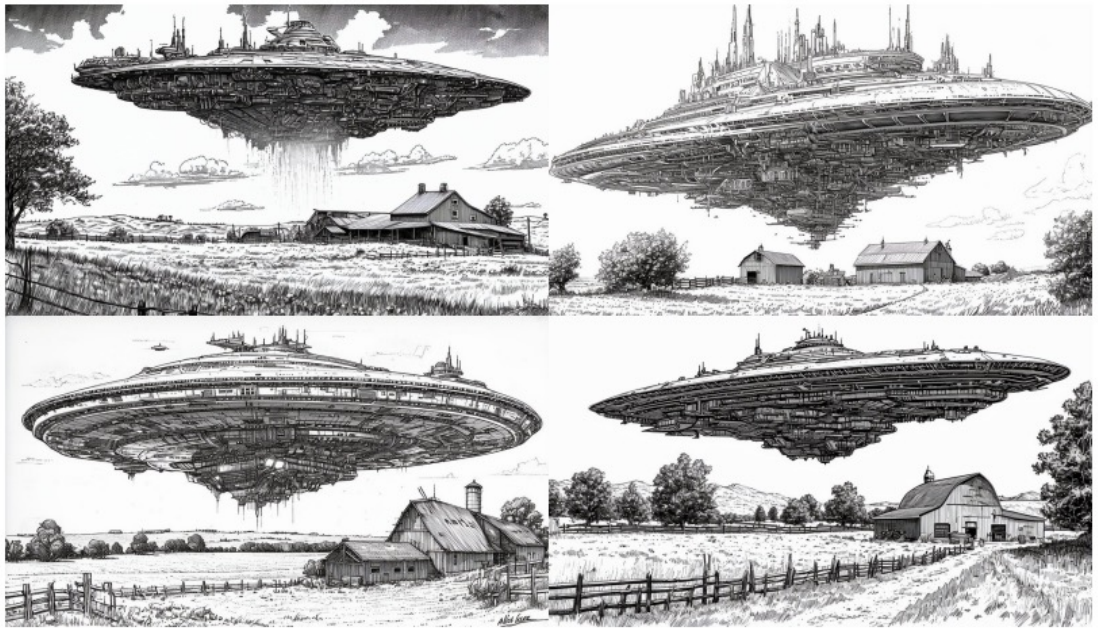
1. Manipulation in Neural Network Prediction

Network Architecture



Stable Diffusion 3. Esser et al., Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. ICML 2024.

Example: Text-to-Image (T2I)



detailed pen and ink drawing of a massive complex alien space ship above a farm in the middle of nowhere.

photo of a bear wearing a suit and tophat in a river in the middle of a forest holding a sign that says "I cant bear it".

Stable Diffusion 3. Esser et al., Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. ICML 2024.

Example: Text-to-Image (T2I)

tilt shift aerial photo of a cute city made of sushi on a wooden table in the evening.

dark high contrast render of a psychedelic tree of life illuminating dust in a mystical cave.

Stable Diffusion 3. Esser et al., Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. ICML 2024.

Example: Text-to-Image (T2I)

an anthropomorphic fractal person behind the counter at a fractal themed restaurant.

beautiful oil painting of a steamboat in a river in the afternoon. On the side of the river is a large brick building with a sign on top that says ŠD3.

Stable Diffusion 3. Esser et al., Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. ICML 2024.

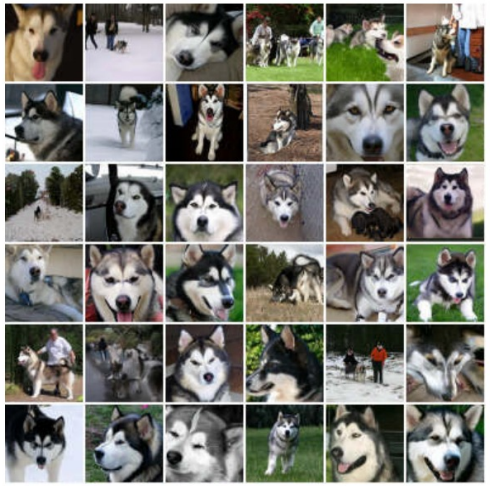
Example: Text-to-Image (T2I)

an anthopomorphic pink donut with a mustache and cowboy hat standing by a log cabin in a forest with an old 1970s orange truck in the driveway

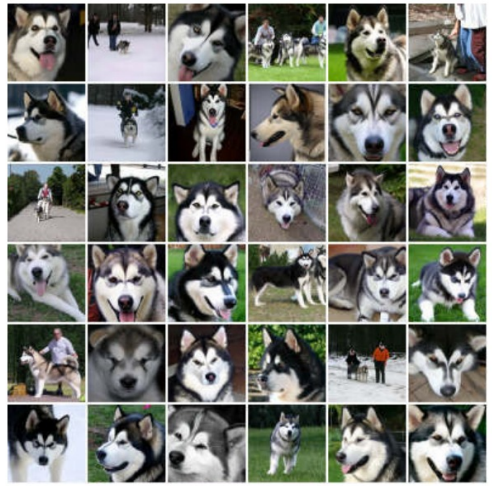
fox sitting in front of a computer in a messy room at night. On the screen is a 3d modeling program with a line render of a zebra.

Classifer-Free Guidance (CFG)

(a) $w = 0$



(b) $w = 1$



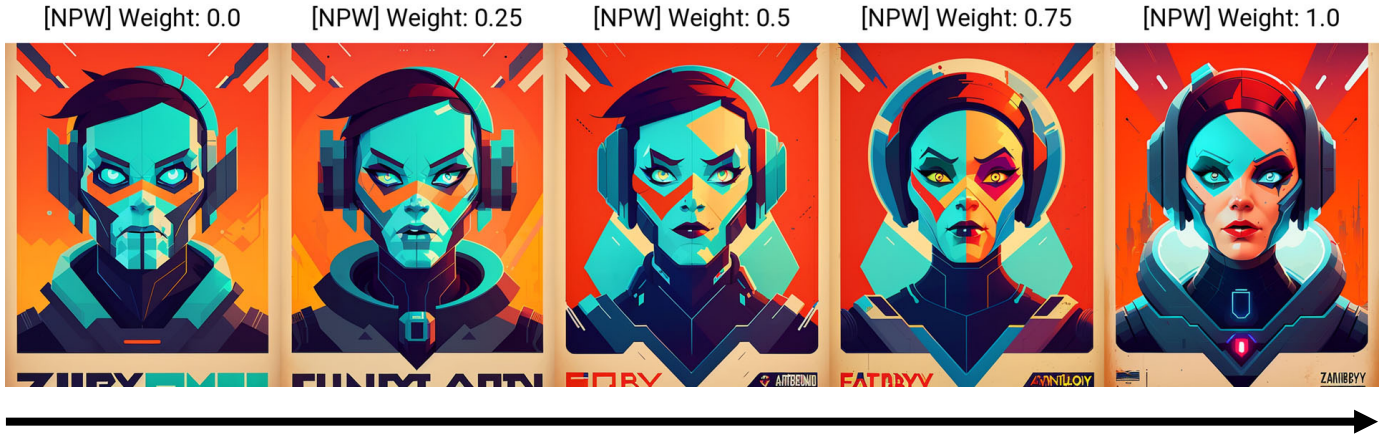
(c) $w = 2$

(d) $w = 4$

https://github.com/muerrilla/stable-diffusion-NPW?tab=readme-ov-file

Negative Prompt

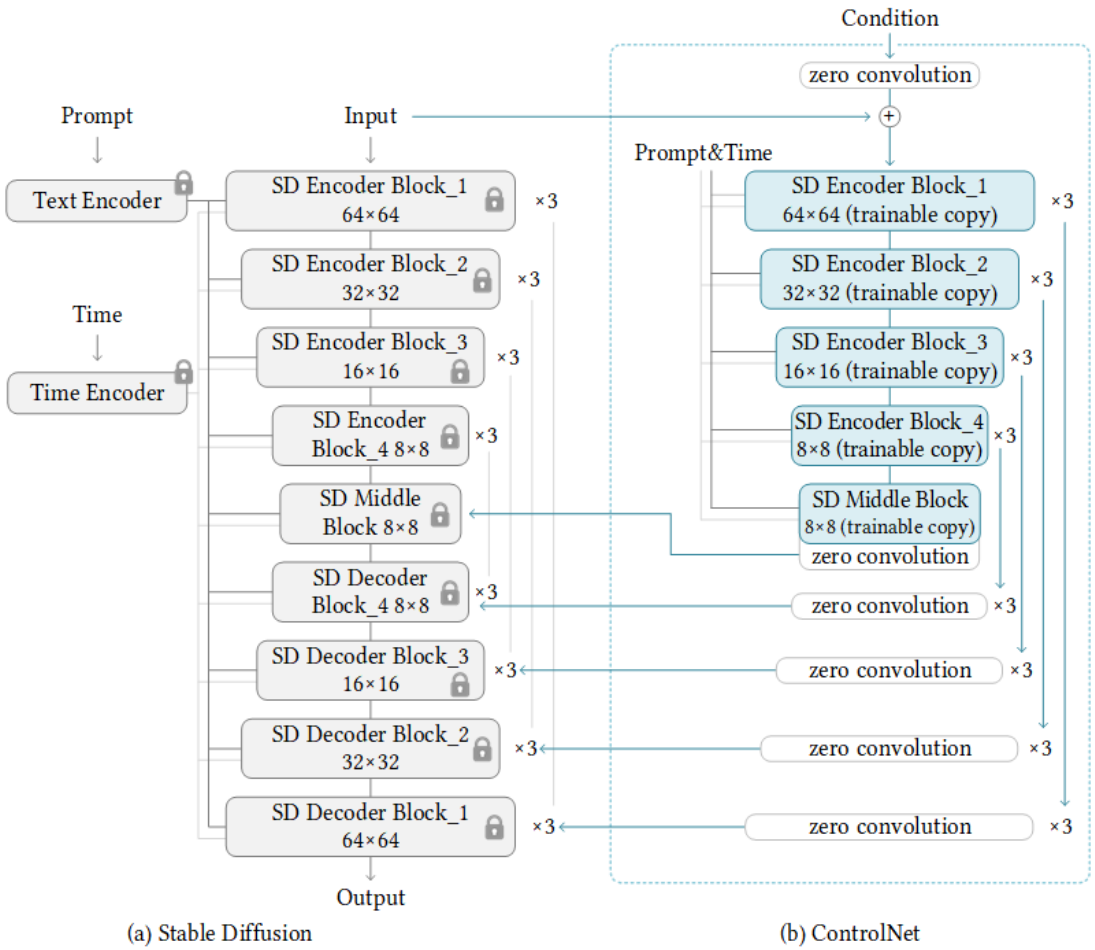
Prompt: Portrait of zimby anton fadeev cyborg propaganda poster



Zhang et al., Adding Conditional Control to Text-to-Image Diffusion Models, ICCV 2022.

ControlNet

Can we convert a pretrained unconditional image diffusion model into an image-conditioned generative model using a relatively much smaller set of input-output pairs ($\sim 50k$)?

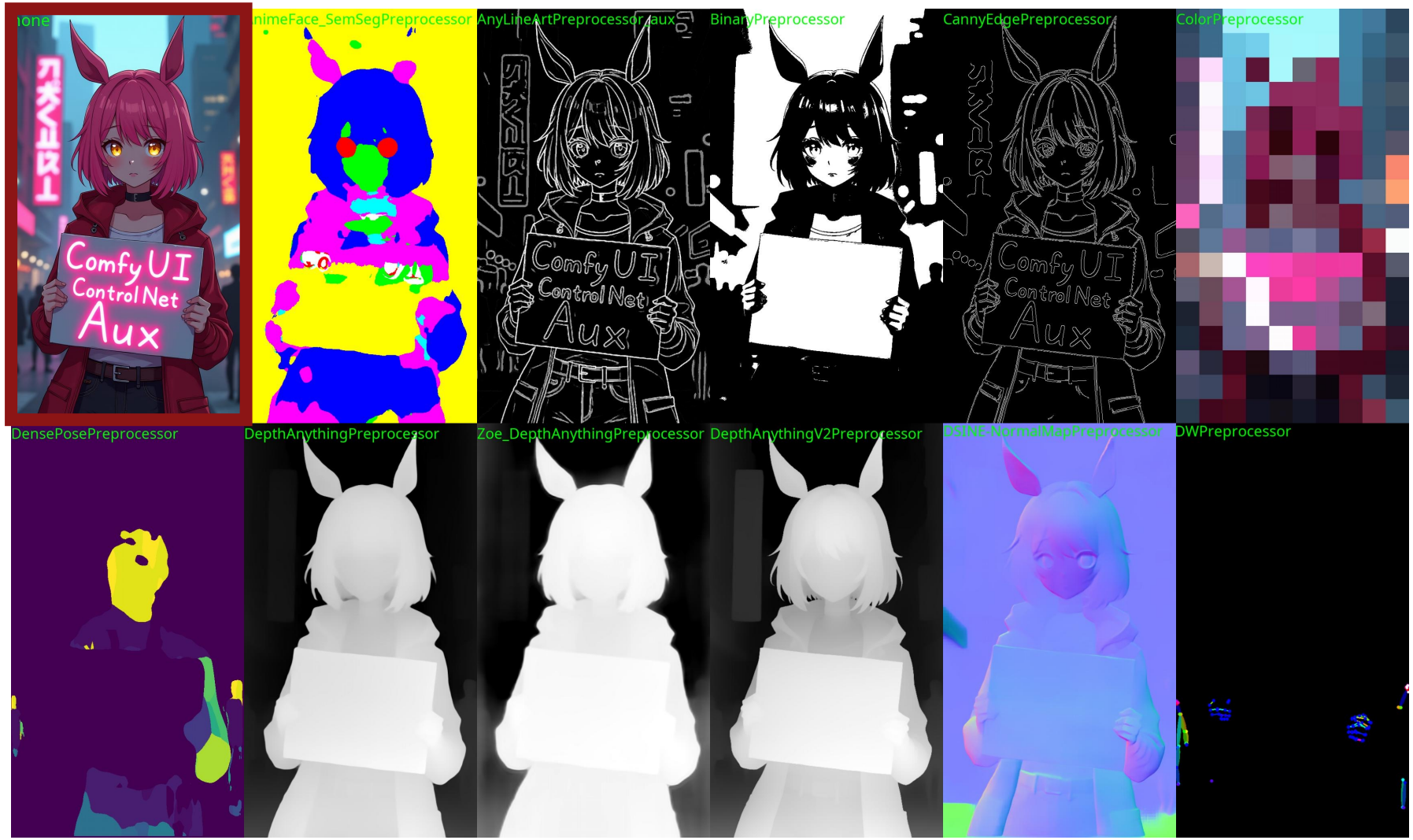


Zhang et al., Adding Conditional Control to Text-to-Image Diffusion Models, ICCV 2022.

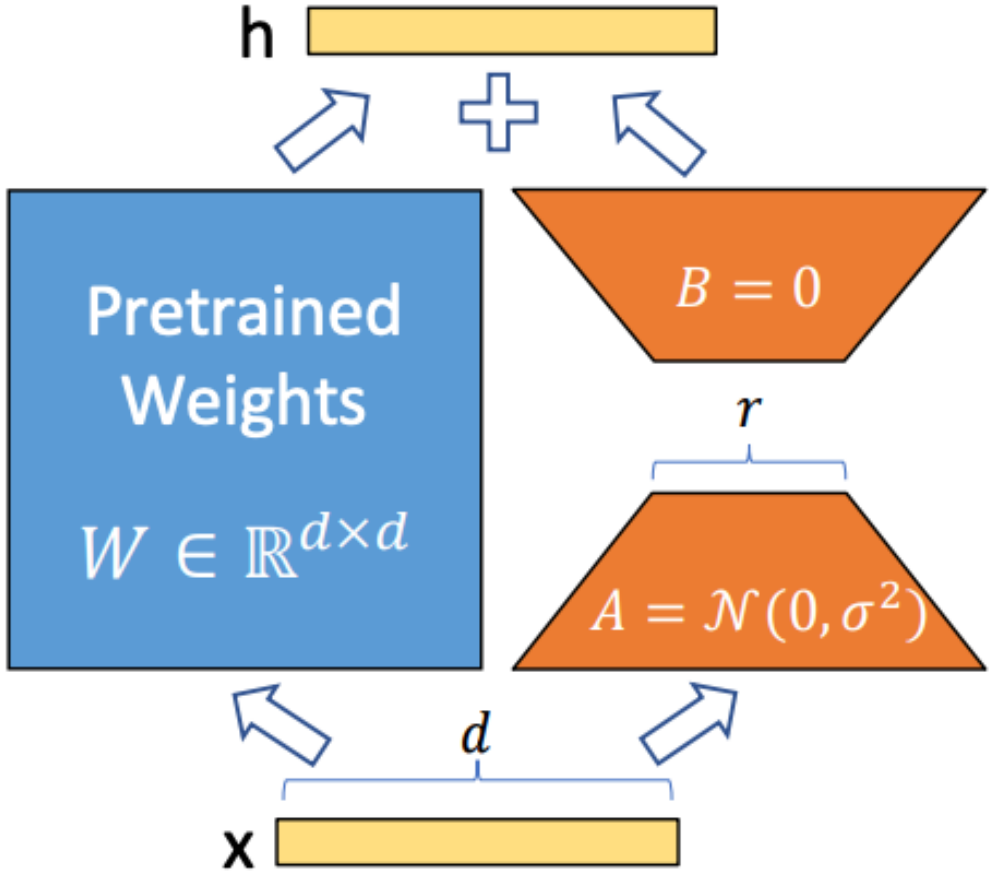
ControlNet

https://www.comfydeploy.com/comfy-node/Fannovel16/comfyui_controlnet_aux

ControlNet



Low-Rank Adaptation (LoRA)



https://www.shruggingface.com/blog/self-portraits-with-stable-diffusion-and-lora

LoRA-Based Personalization

LoRA

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

LoRA-Based Personalization

DreamBooth

2D Spatial Grounding

"A car and a castle in a beautiful landscape and a balloon in the sunset sky."

"A bear is in an antique living room with a chair and a chiffonier, while a chandelier hanging from the ceiling."

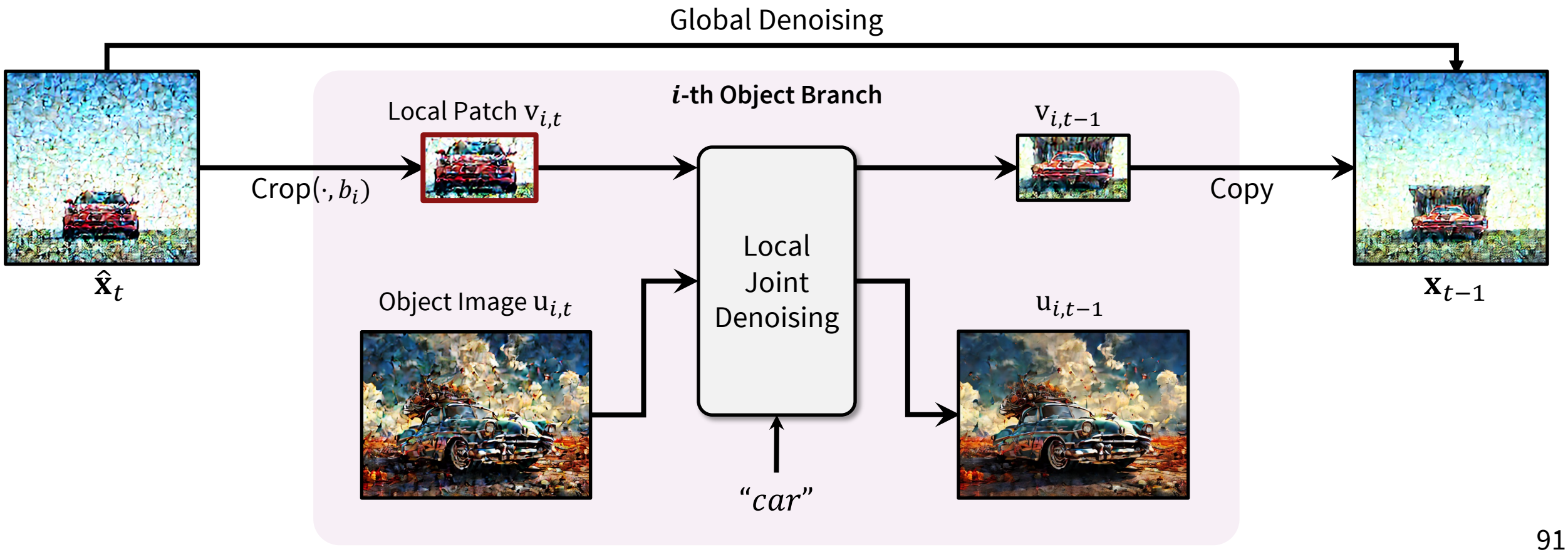
"A balloon and fireworks are in the night sky, and below is a beautiful forest with a car and a chair and a waterfall."

"A banana and an apple are beneath a book and a flower is lying on the book in a room"

"Aurora lights up the sky and a horse and a house are on the grassy meadow with a mountain in the background."

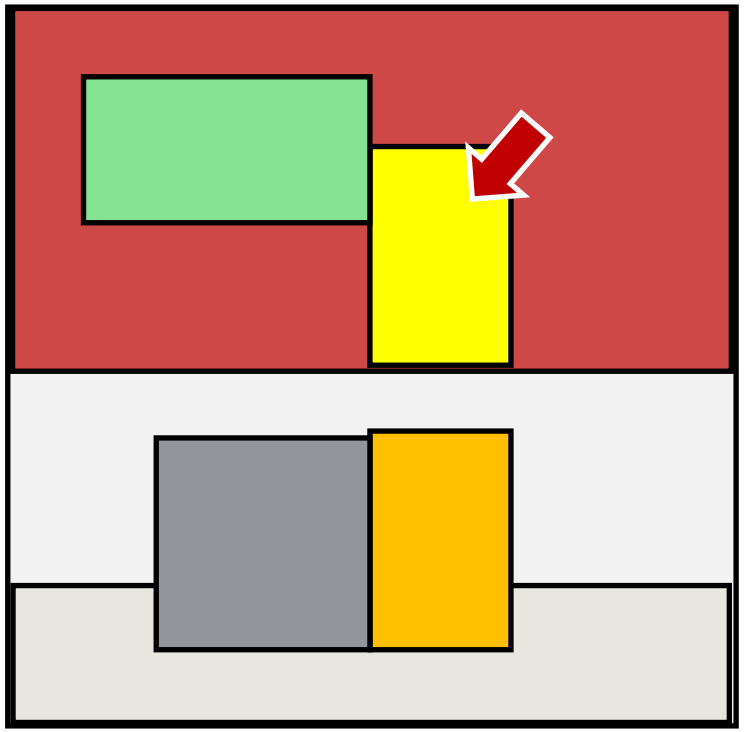
Joint Diffusion with DiT

Generate an object image for each bounding box and jointly denoise it with the corresponding region of the main image.

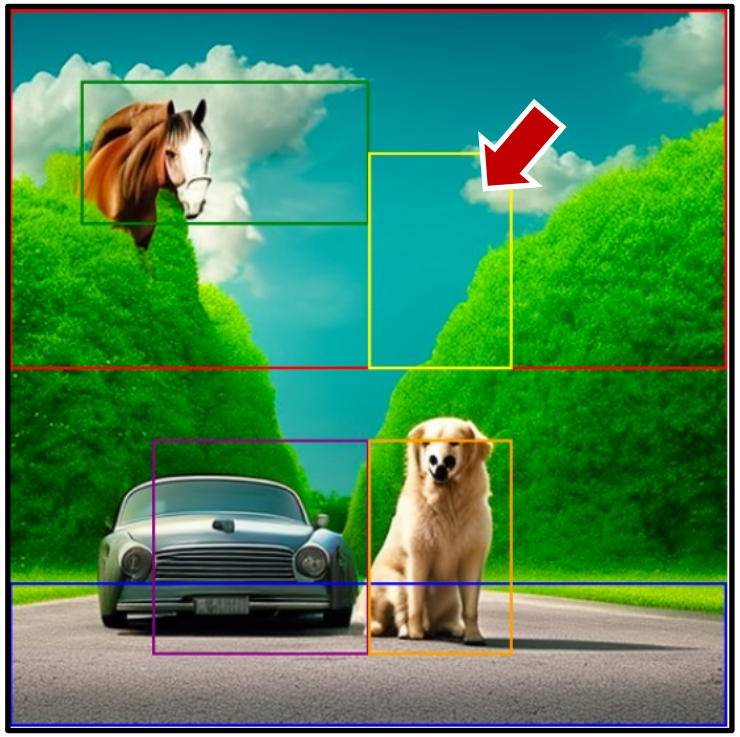


GrounDiT Results

Layout



R&B [Xiao et al., ICLR 2024]

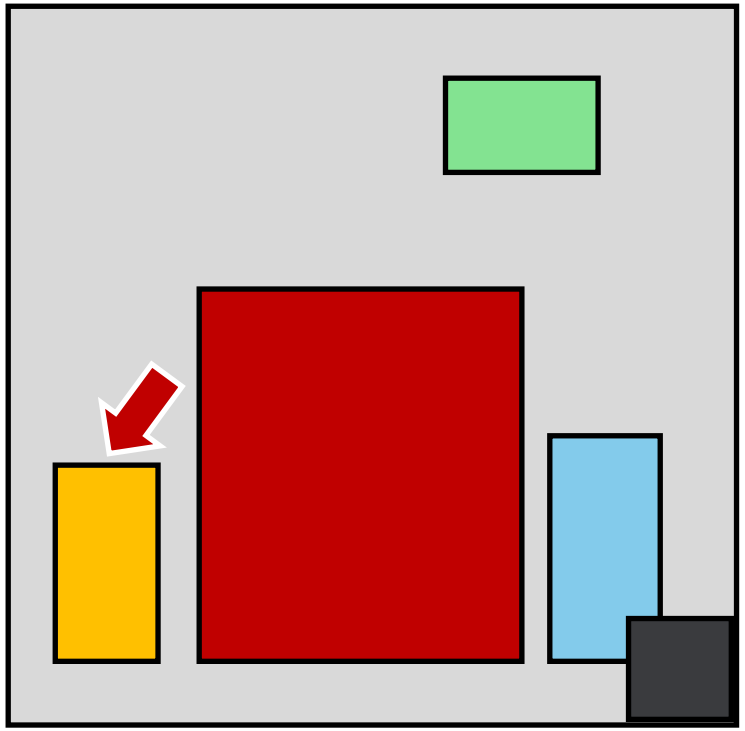


Ours

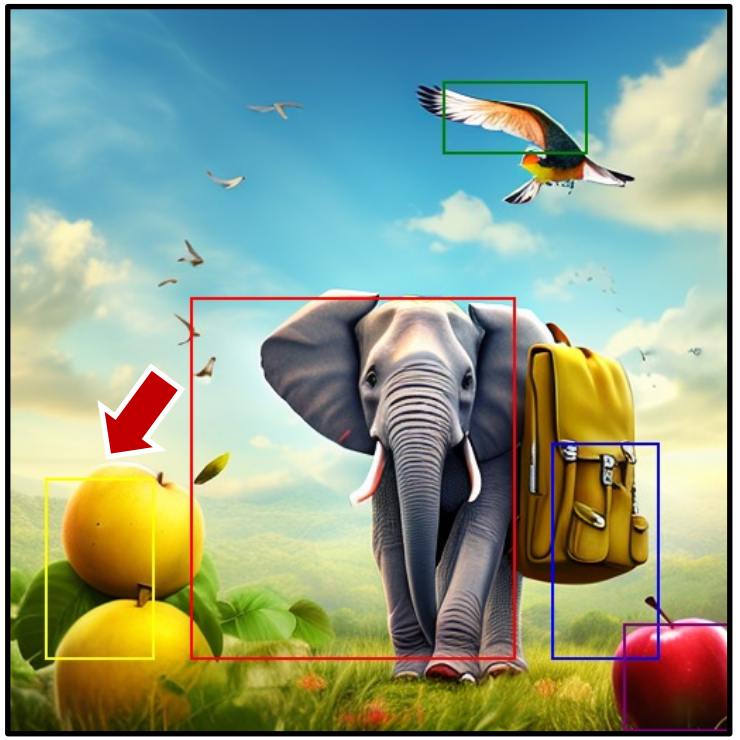
"A car and a dog on the road while horse and a chair is on the grass."

GrounDiT Results

Layout



R&B [Xiao et al., ICLR 2024]

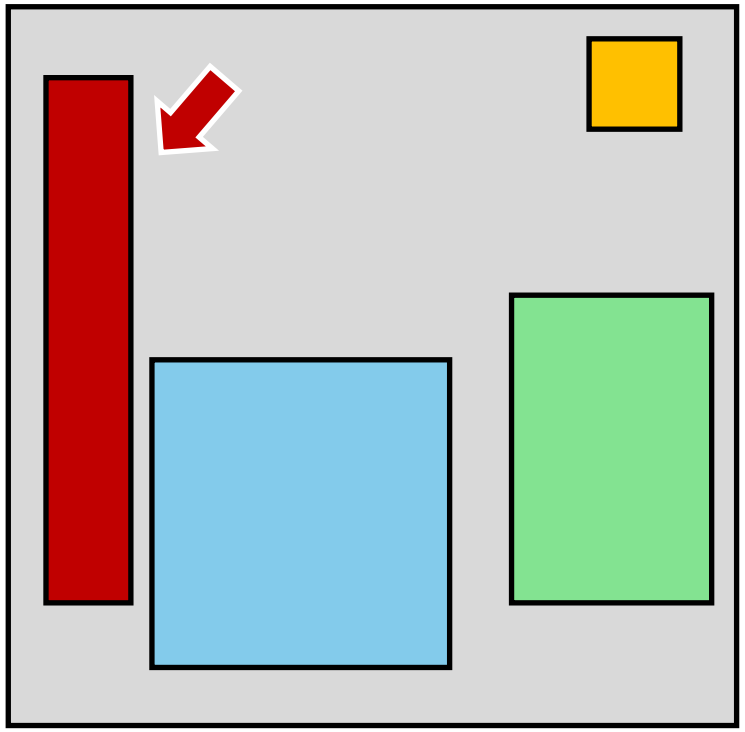


Ours

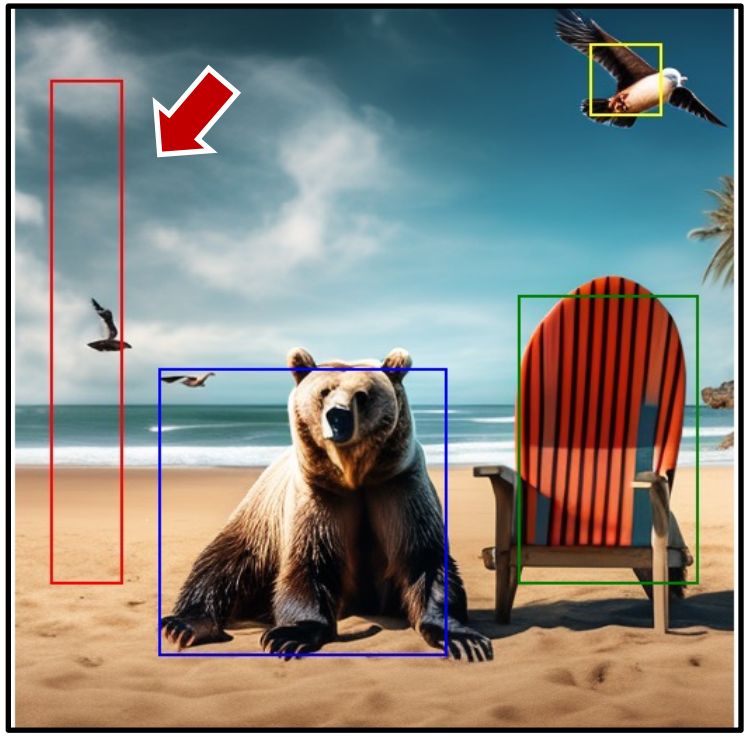
"A banana and an apple and an elephant and a backpack in the meadow with bird flying in the sky."

GrounDiT Results

Layout



R&B [Xiao et al., ICLR 2024]

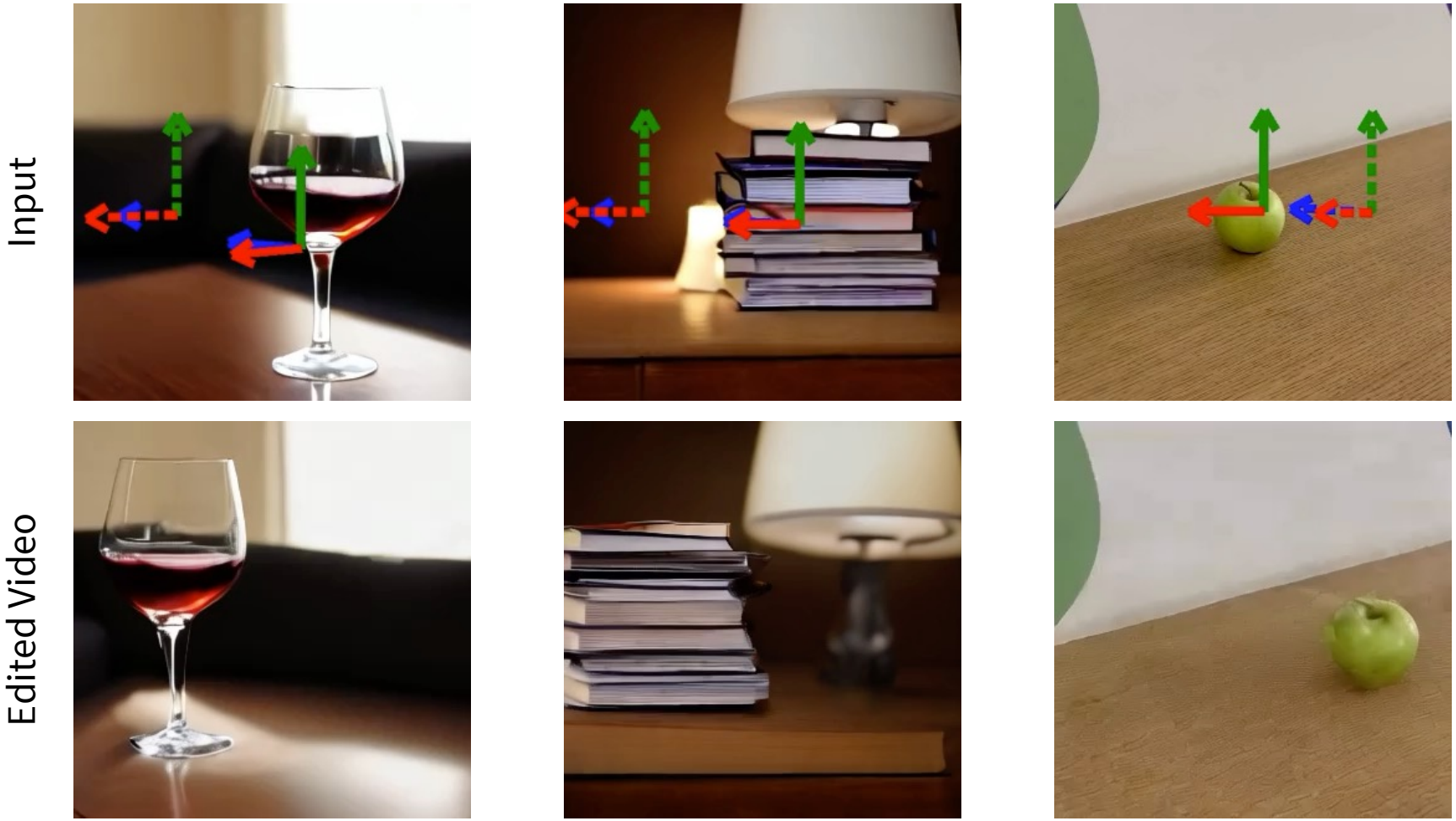


Ours

"A bear sitting between a surfboard and a chair with a bird flying in the sky."

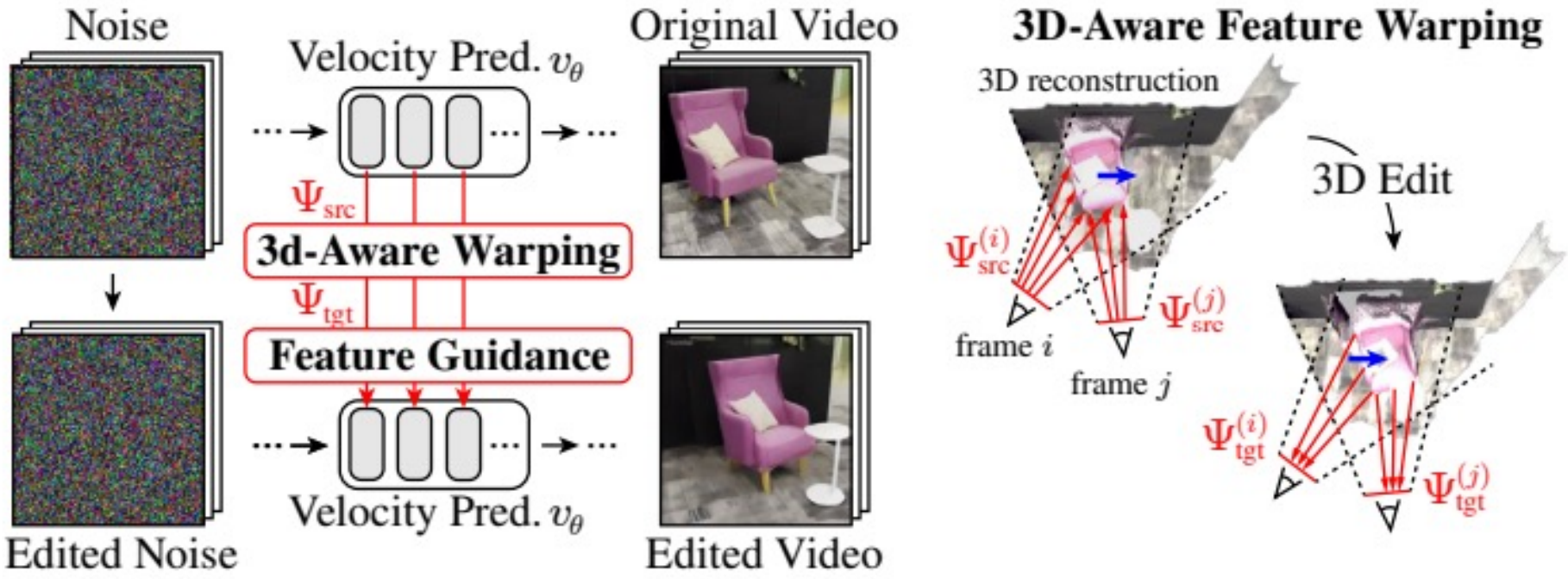
94

Compositional Video Editing



VideoHandles

Key Idea: Warp the attention maps of the score prediction network based on 3D reconstruction and the given 3D transformation.



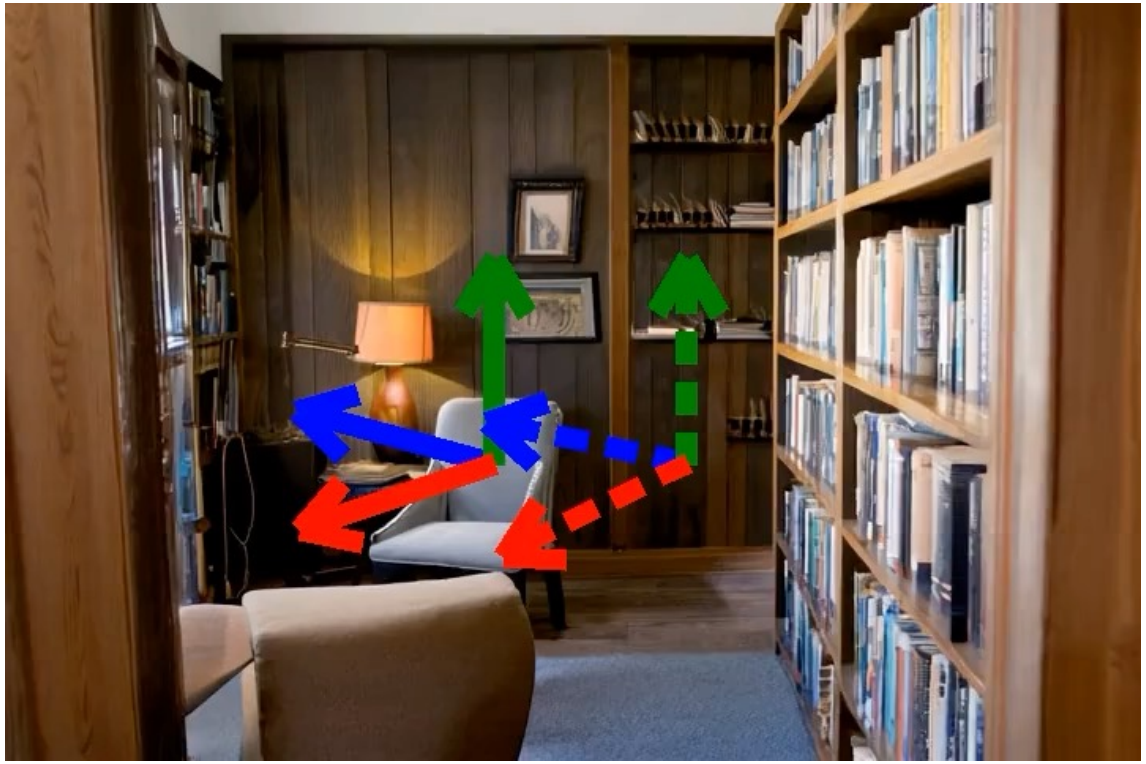
VideoHandles Results

Input

Output

VideoHandles Results

Input

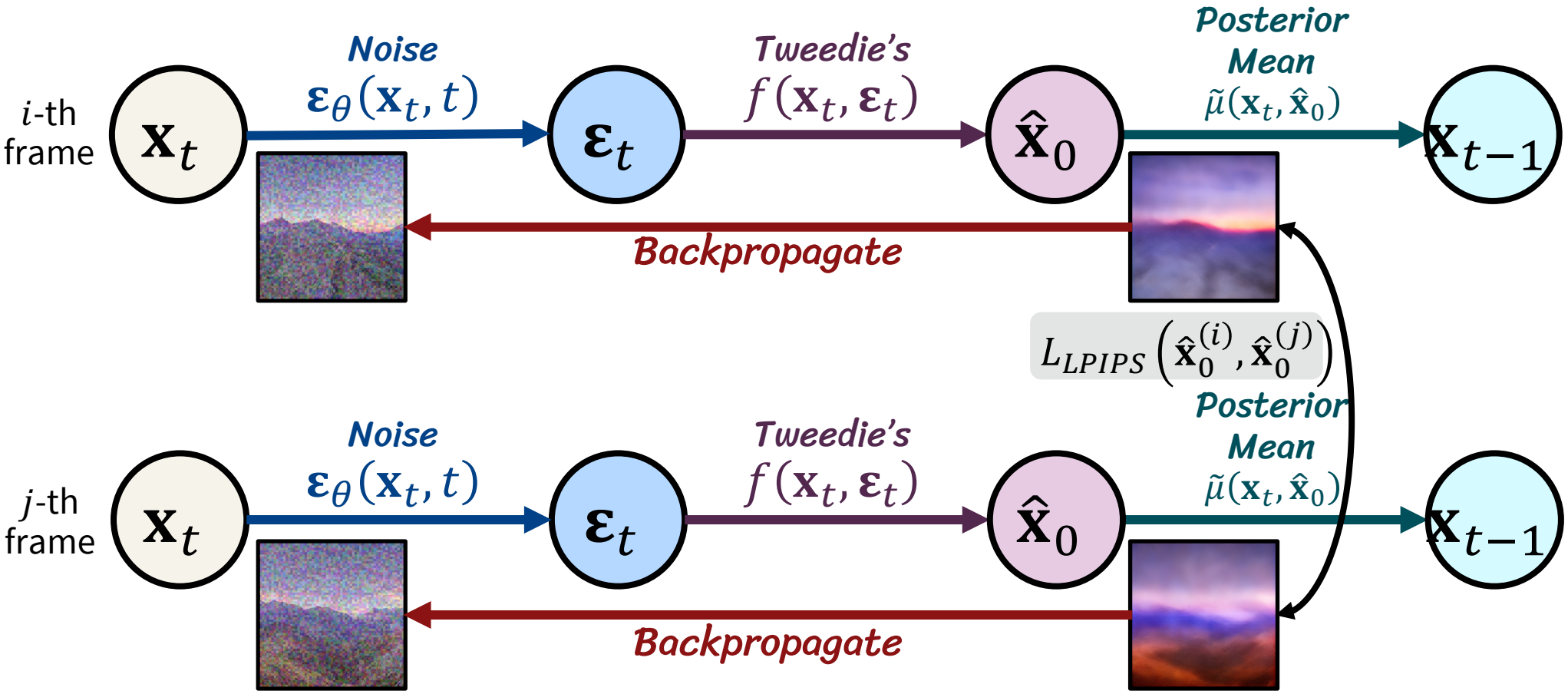


Output

2. Guidance Based on the Expected Final Output

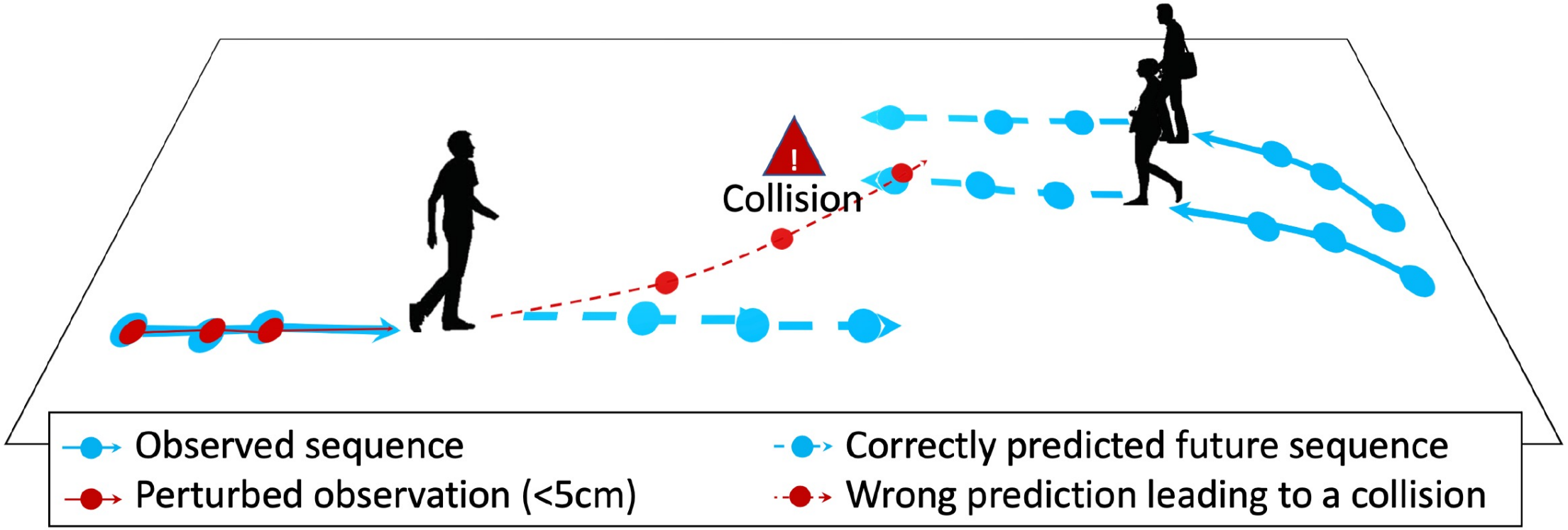
Guidance via Lookahead

Perform backpropagation from the expected final output to the noisy data point.



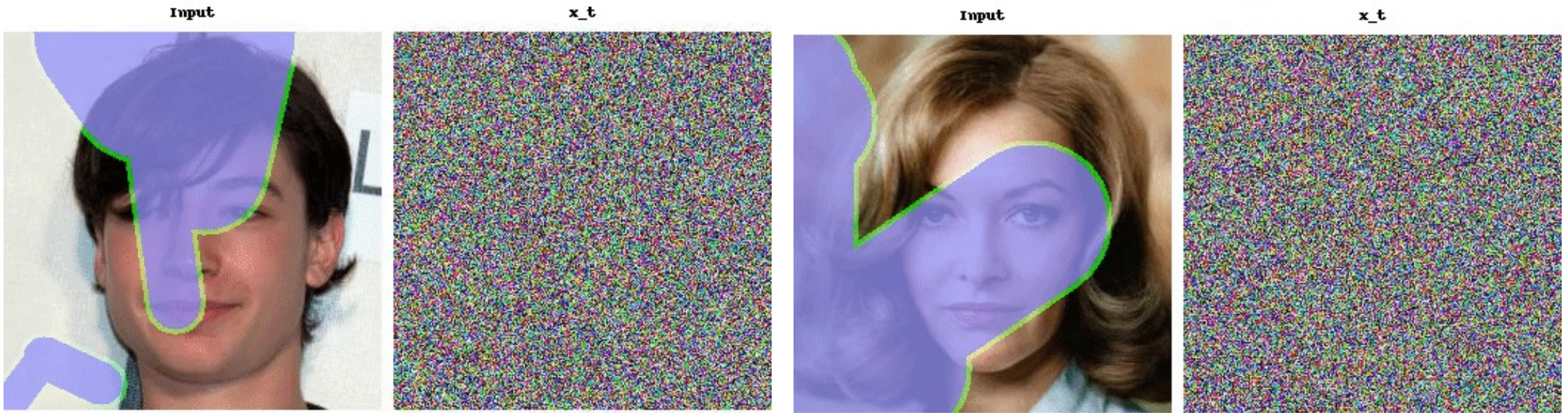
https://www.epfl.ch/labs/vita/research/prediction/analyzing-and-robustifying-trajectory-rediction-models/

Path Correction via Lookahead



Lugmayr et al., RePaint: Inpainting using Denoising Diffusion Probabilistic Models, CVPR 2022.

Application: Inpainting



Ye et al., TFG: Unified Training-Free Guidance for Diffusion Models, NeurIPS 2024.

Application: Style Transfer

"The Starry Night"

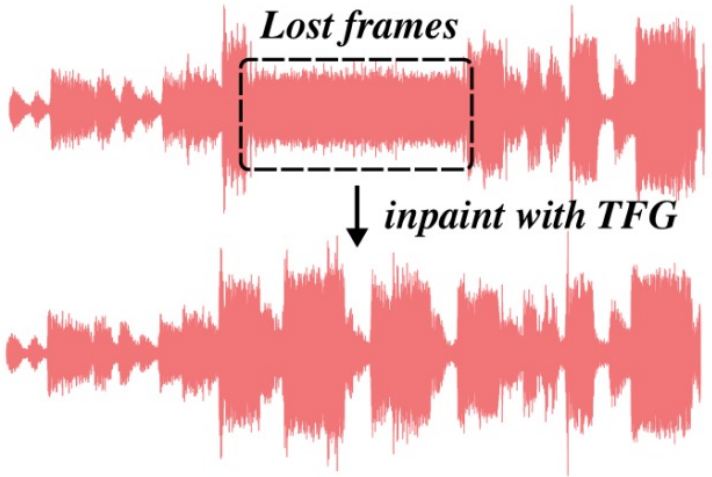
Style image

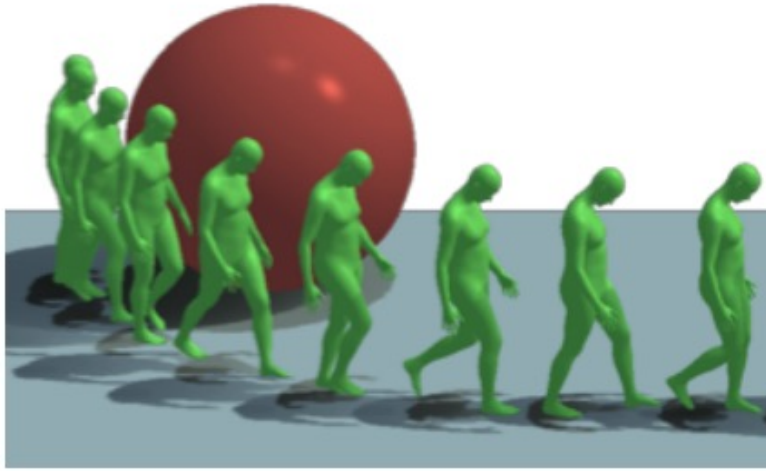
A pickup truck *A canal in Venice* *A white towel*

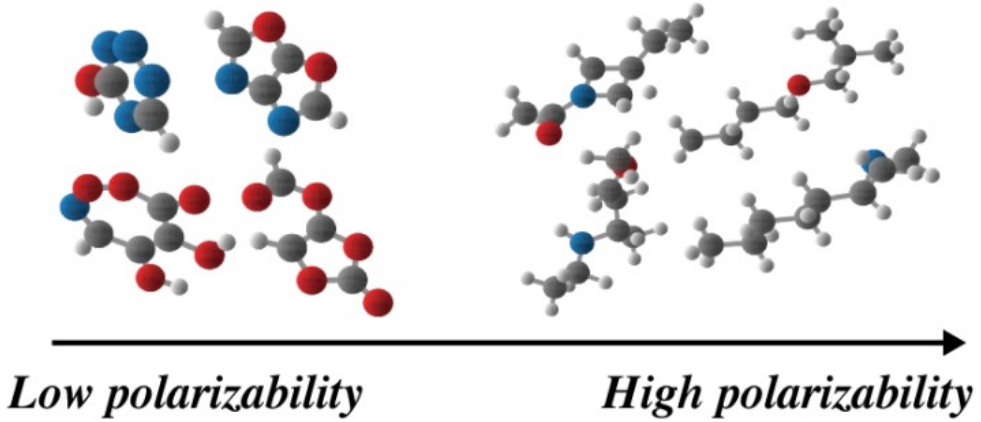
The Statue of Liberty in Minecraft *Concentric circle* *An Elephant*

Song et al., Loss-Guided Diffusion Models for Plug-and-Play Controllable Generation, ICML 2023.
Ye et al., TFG: Unified Training-Free Guidance for Diffusion Models, NeurIPS 2024.

Application in Audio/Motion/Molecule







Diffusion Synchronization

- Given a large image space, define multiple instance spaces with overlaps.
- Perform the generative processes for all instance frames jointly while synchronizing them.

Application: Wide Image Generation

MultiDiffusion (Bar-Tal et al.)

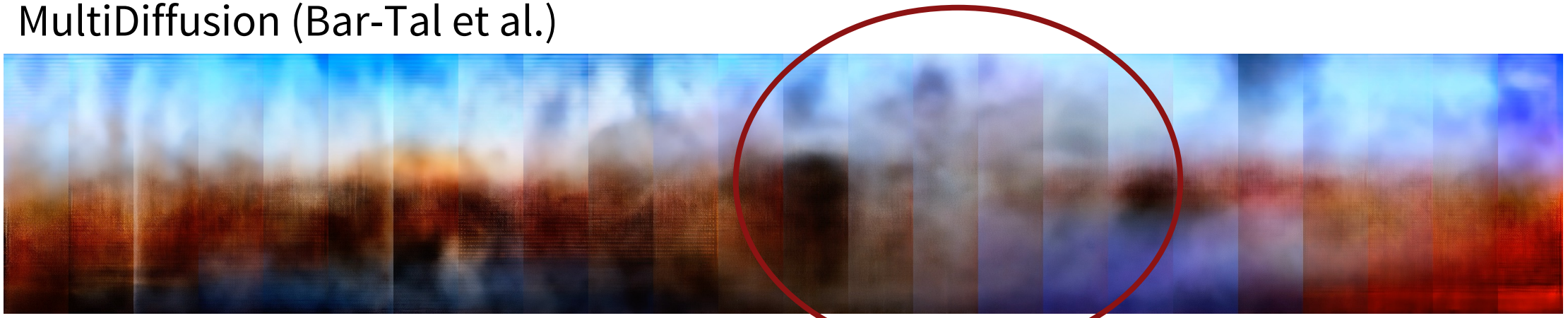
SyncDiffusion (Ours)

"A photo of a rock concert"

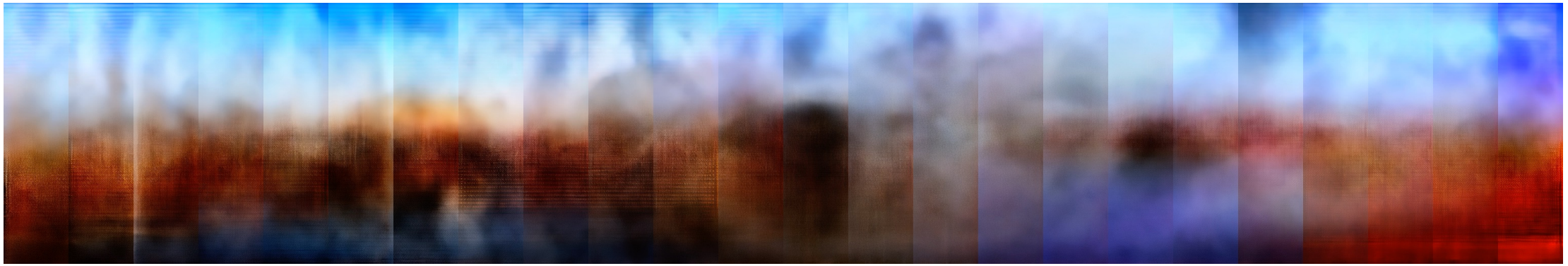
More results are available on https://syncdiffusion.github.io/.

Application: Wide Image Generation

MultiDiffusion (Bar-Tal et al.)



SyncDiffusion (Ours)



"Skyline of New York City"

More results are available on https://syncdiffusion.github.io/.

Application: Wide Image Generation

MultiDiffusion (Bar-Tal et al.)

SyncDiffusion (Ours)

"Silhouette wallpaper of a dreamy scene with shooting stars"

More results are available on https://syncdiffusion.github.io/.

Application: Wide Image Generation

MultiDiffusion (Bar-Tal et al.)

SyncDiffusion (Ours)

"An illustration of a beach in La La Land style"

More results are available on https://syncdiffusion.github.io/.

Application: Wide Image Generation

MultiDiffusion (Bar-Tal et al.)

SyncDiffusion (Ours)

"A cinematic view of a castle in the sunset"

More results are available on https://syncdiffusion.github.io/.

Diffusion Synchronization

The synchronization can be conducted in any canonical space (e.g., panorama space, texture space, etc).

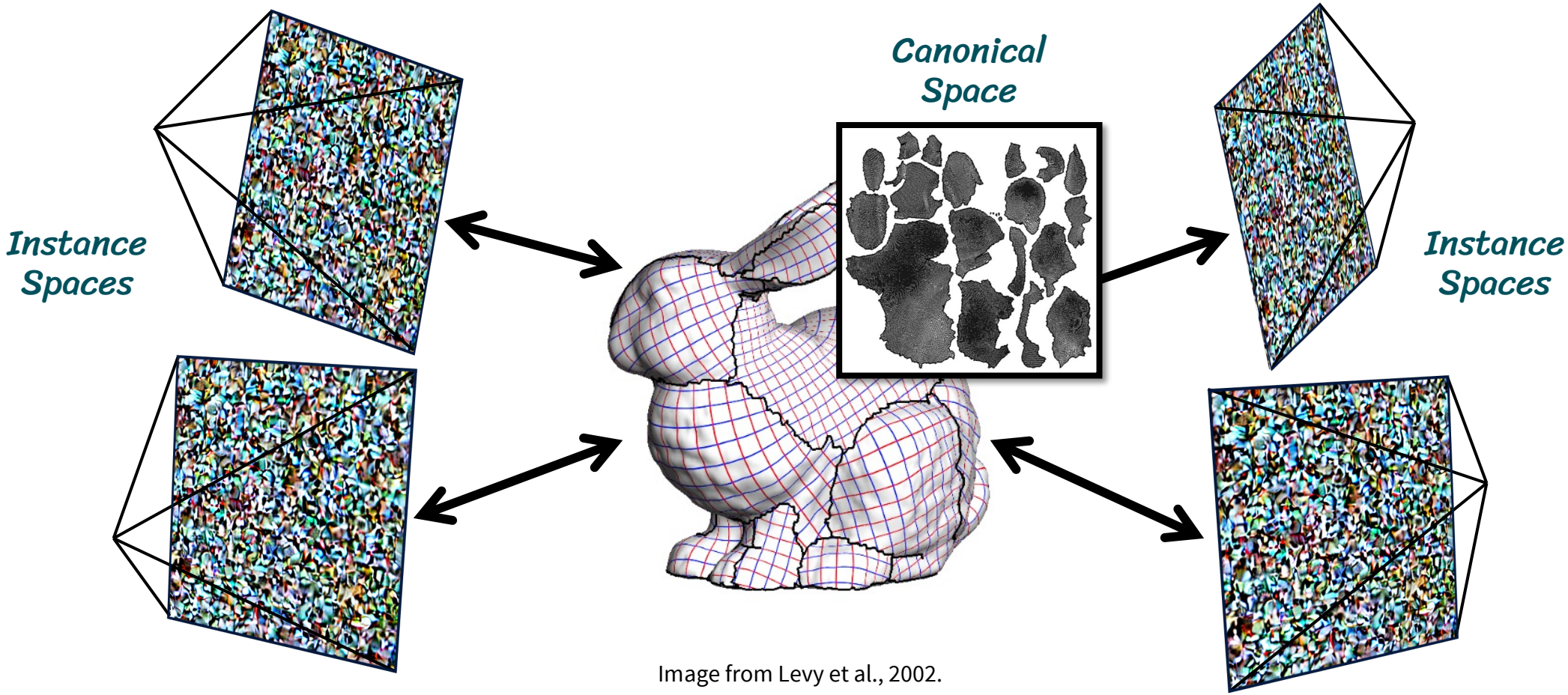


Image from Levy et al., 2002.

Diffusion Synchronization

1. Unproject Tweedie estimate of instance spaces into the canonical space.

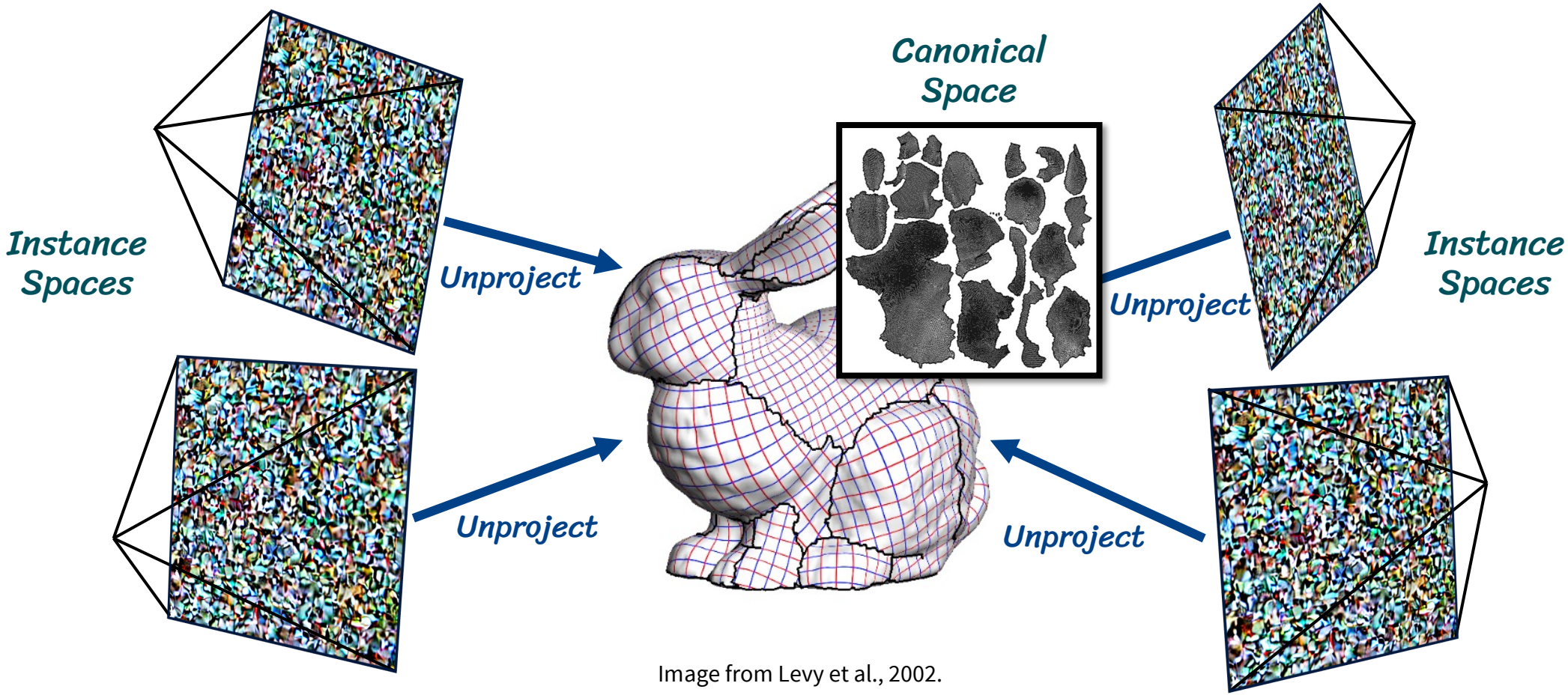


Image from Levy et al., 2002.

Content Synchronization

2. Compute the loss: the difference between each Tweedie estimate and their average in the canonical space.

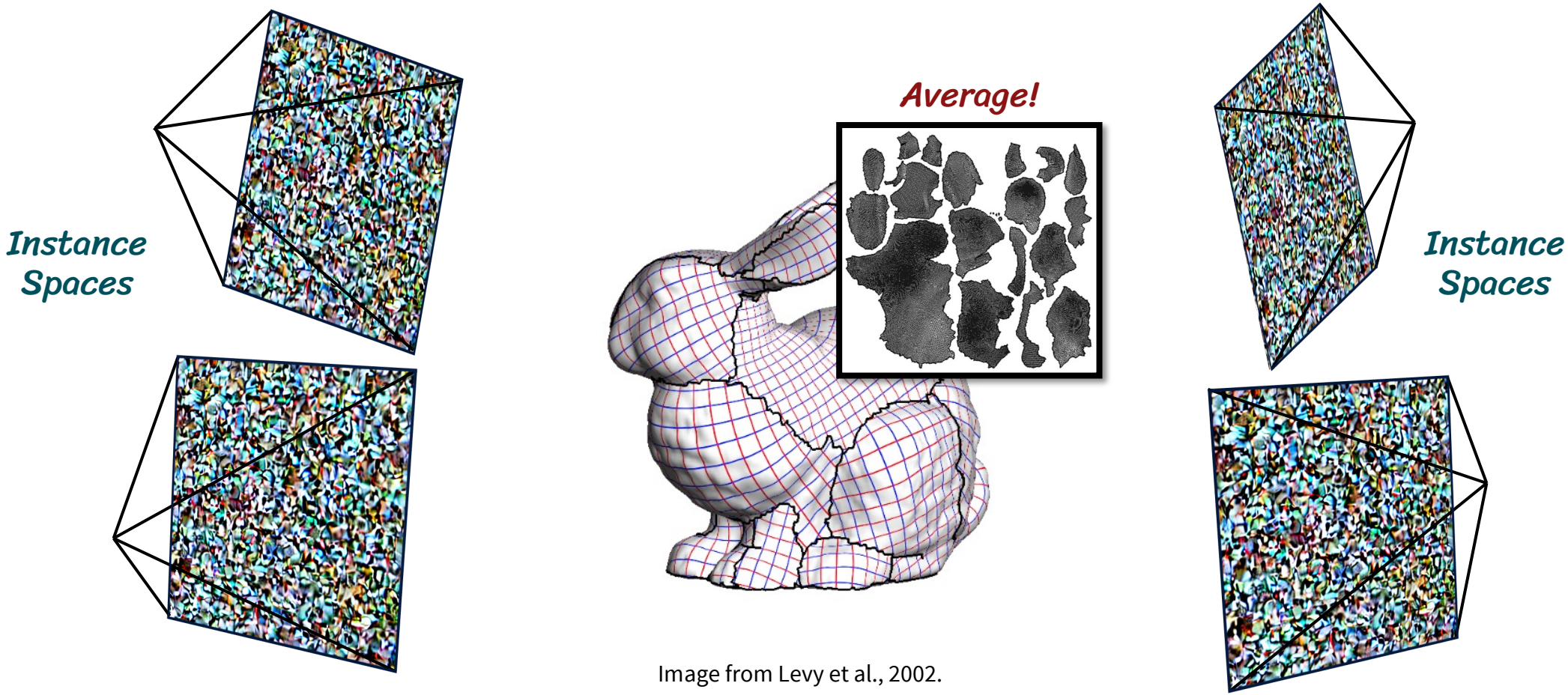


Image from Levy et al., 2002.

Diffusion Synchronization

3. Backpropagate the loss to the noise in the instance spaces.

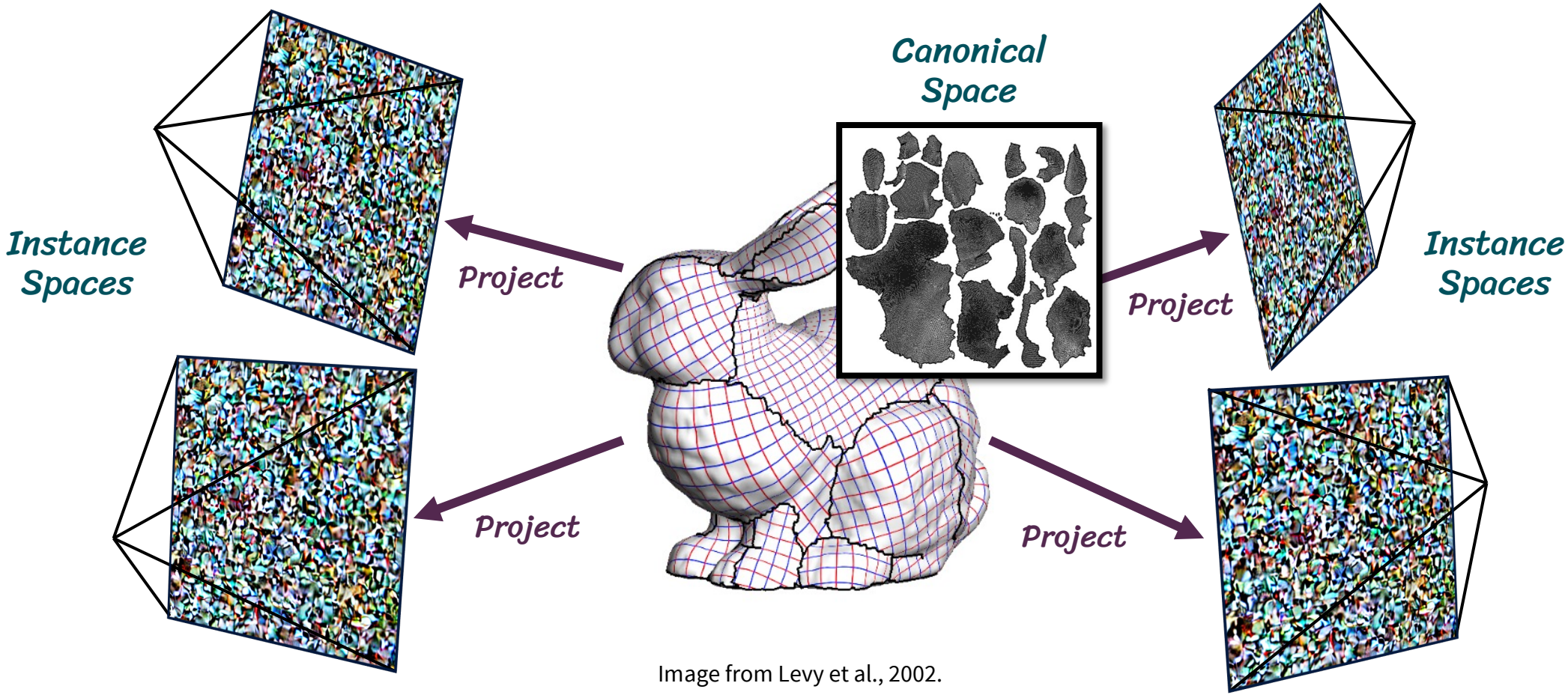


Image from Levy et al., 2002.

Application: 3D Texturing

"A hand carved wood turtle"

"A dumpster"

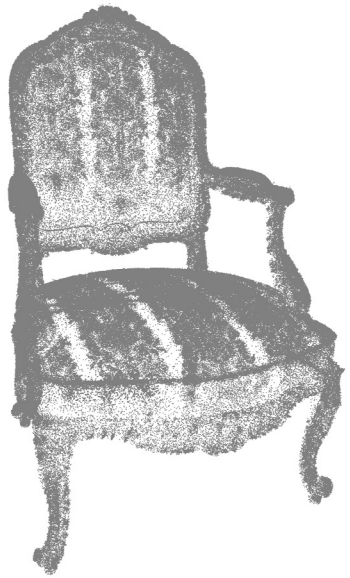
"A Chinese style lantern"

"A car with graffiti"

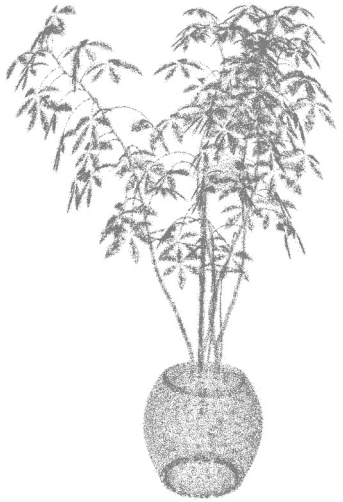
More results are available on https://synctweedies.github.io/.

Application: 3D Texturing

Gaussian splat texture generation



"A majestic red chair"



"A photo of a tree with multicolored leaves"

More results are available on https://synctweedies.github.io/.

Application: 3D Texturing

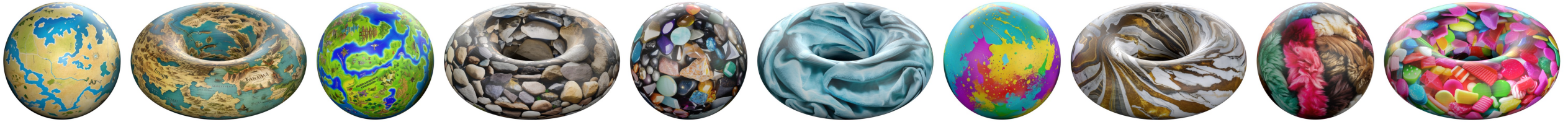
"A marble dresser"

"A wooden crate"

"A mug with spiraling silver ornaments"

"A marble statue of a wolf"

"A cozy stone fireplace"



Application: 360° Panorama Generation

"Rocky desert landscape with towering saguaro cacti"

"Graffiti-covered alleyway with street art murals"

"Quirky steampunk workshop filled with gears and gadgets"

"Cozy neighborhood pub with outdoor seating"

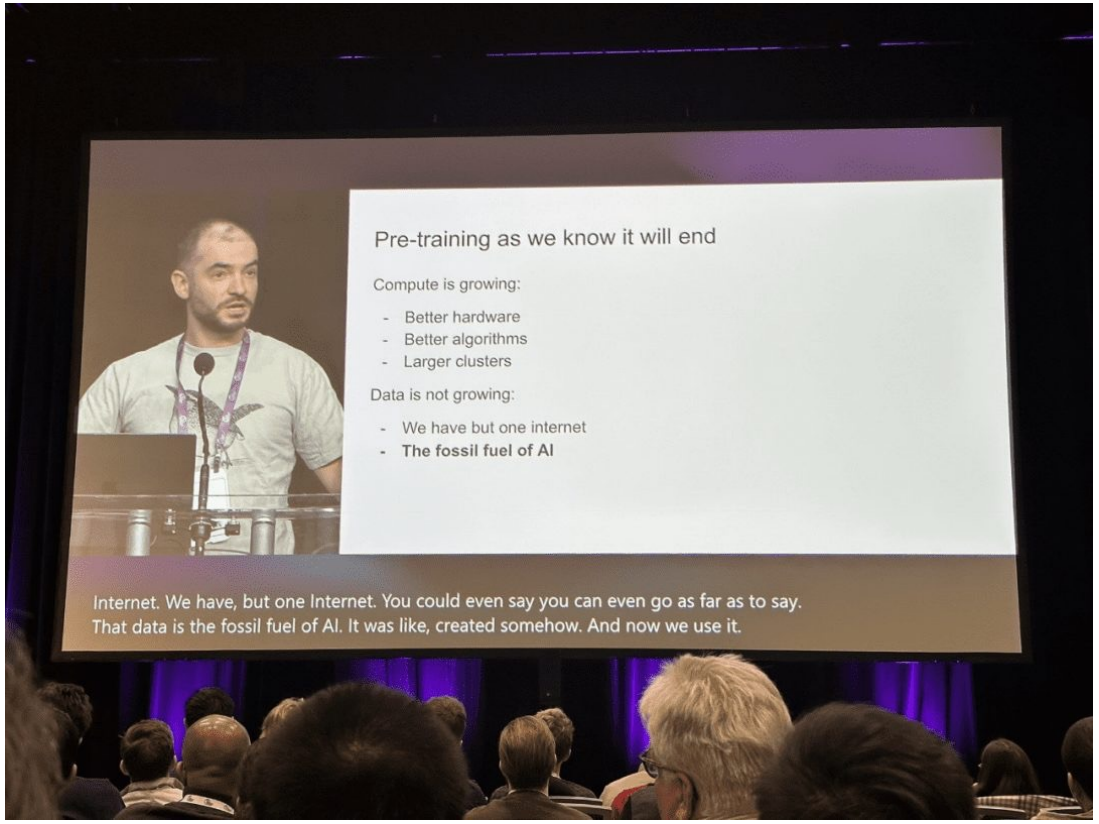
"Abandoned factory with soft rays through dusty air"

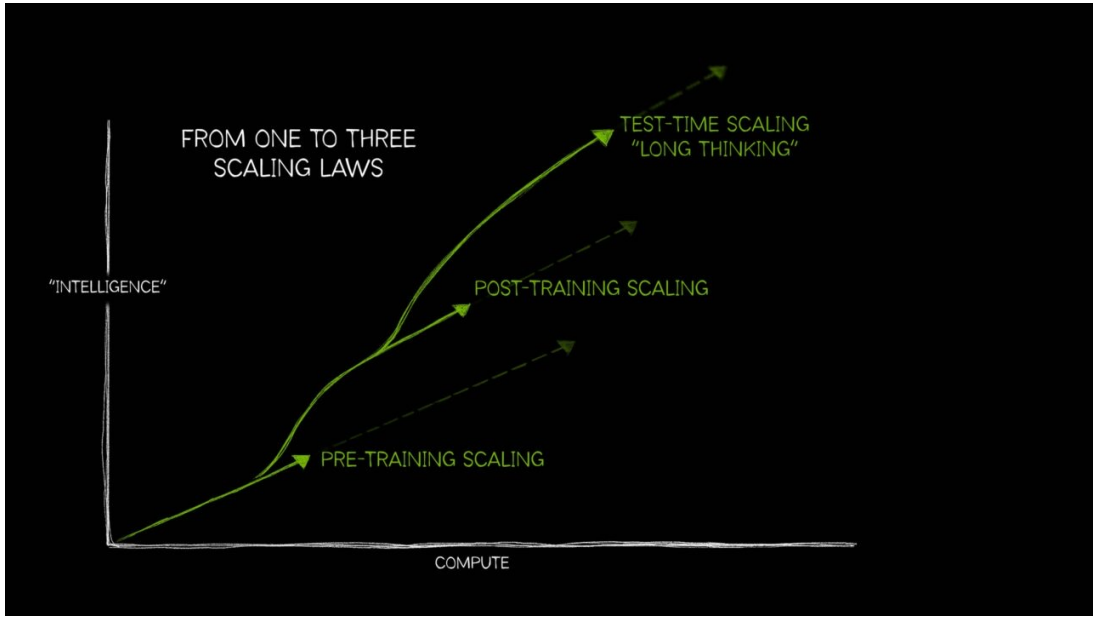
"Quaint canal lined with boats and cafes"

3. Particle Filtering

https://www.aisharenet.com/en/ilya-sutskeverzaineu/
https://blogs.nvidia.com/blog/ai-scaling-laws/

The End of the Pre-Training Era?

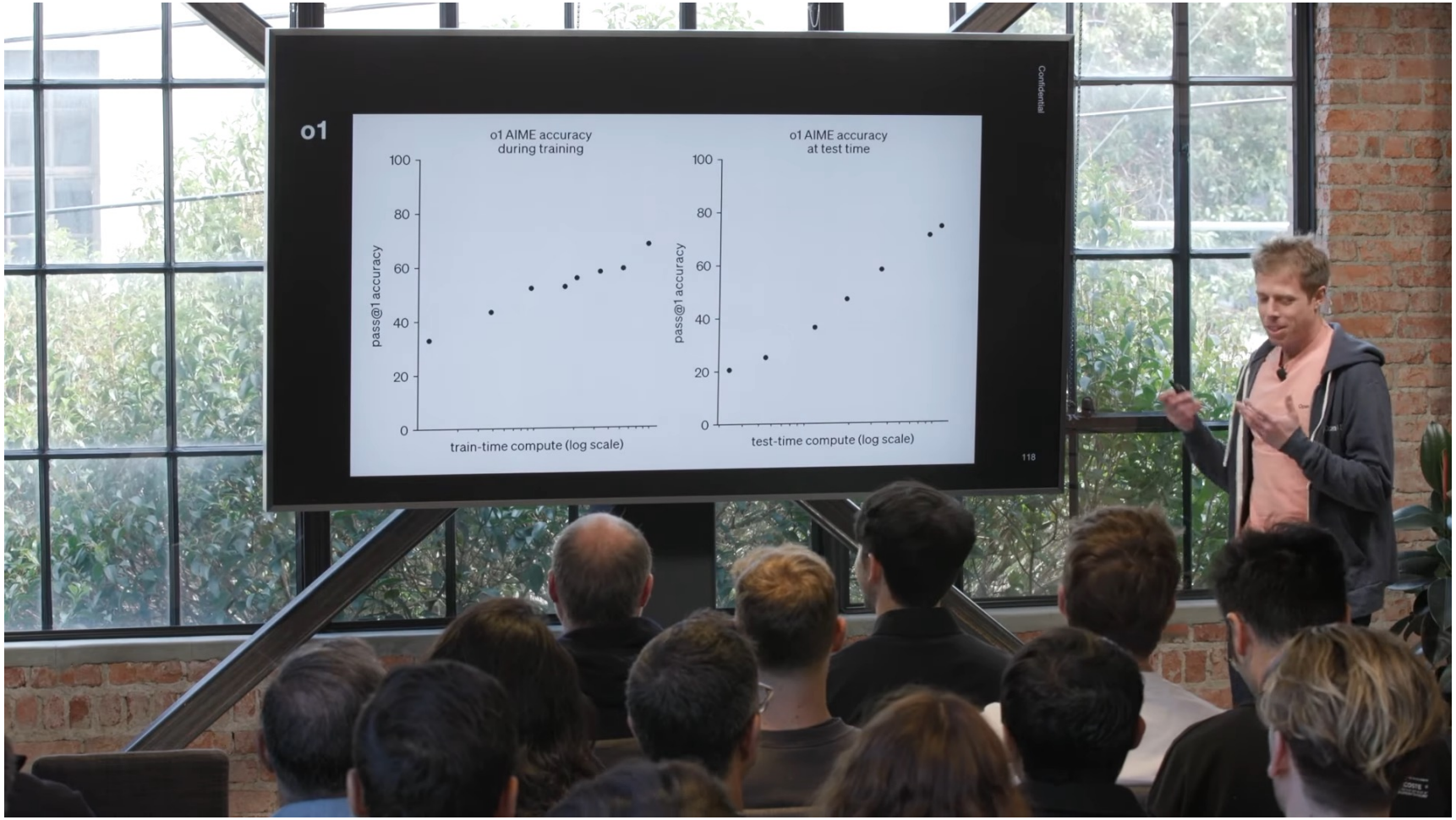




https://www.youtube.com/watch?v=_rjD_2zn2JU

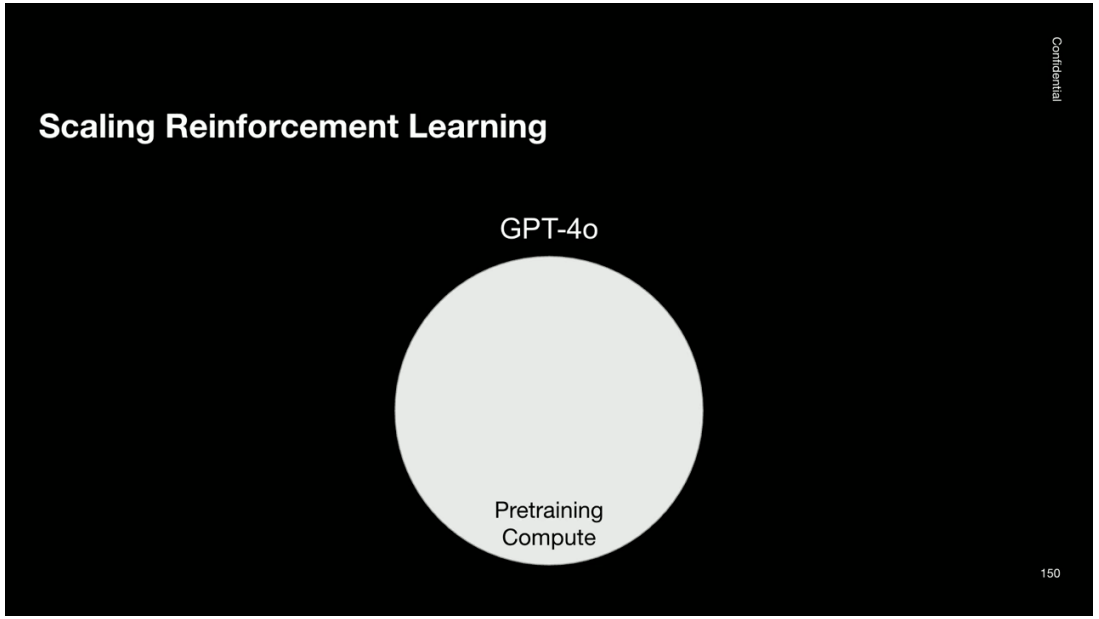
OpenAI’s Dan Roberts Reasons

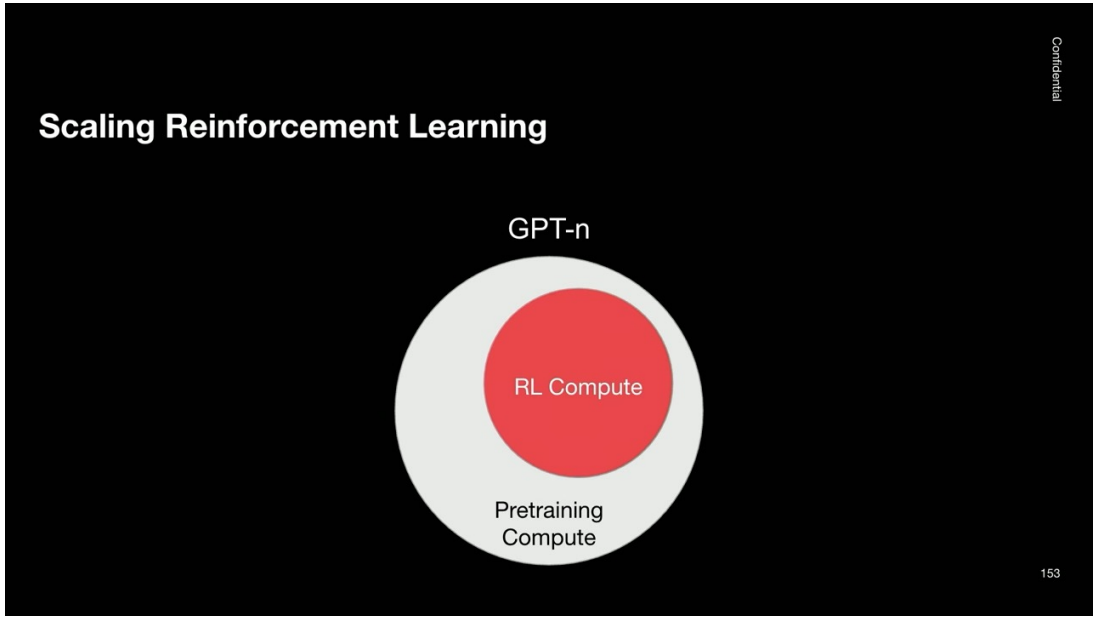
9 Years to AGI? OpenAI’s Dan Roberts Reasons About Emulating Einstein



https://www.youtube.com/watch?v=_rjD_2zn2JU

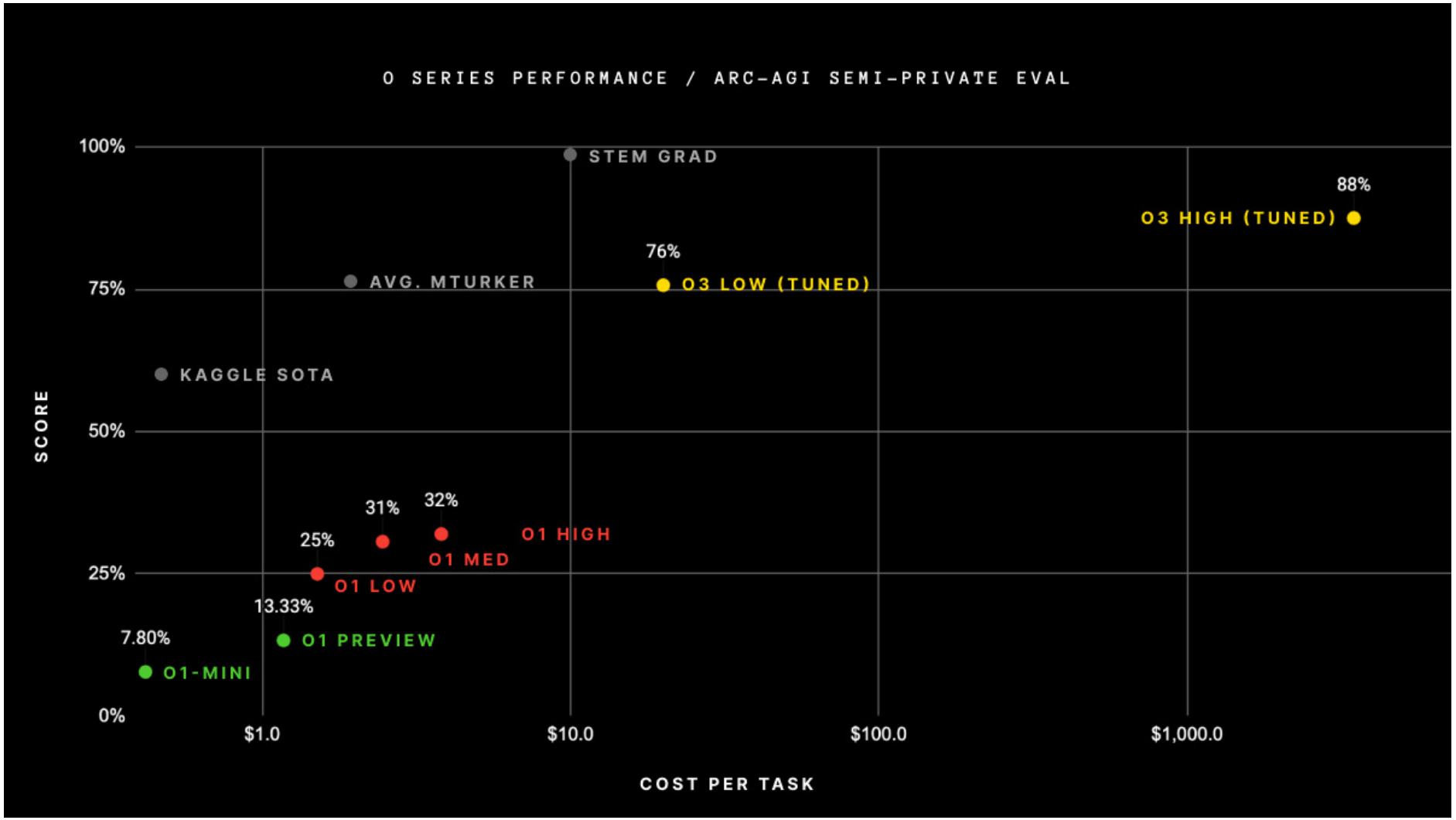
OpenAI's Dan Roberts Reasons





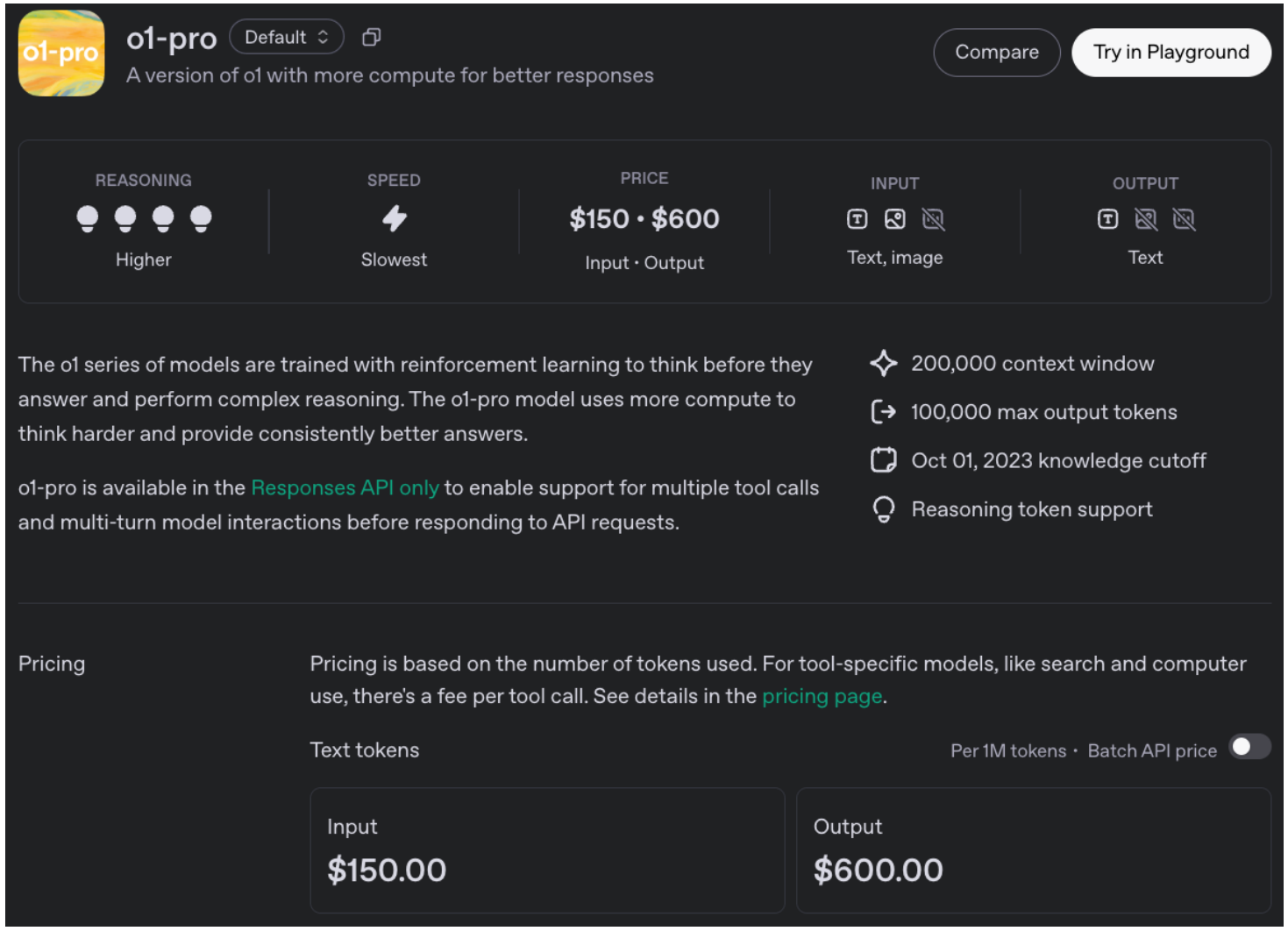
https://openai.com/index/learning-to-reason-with-llms/

Inference-Time Scaling – GPT

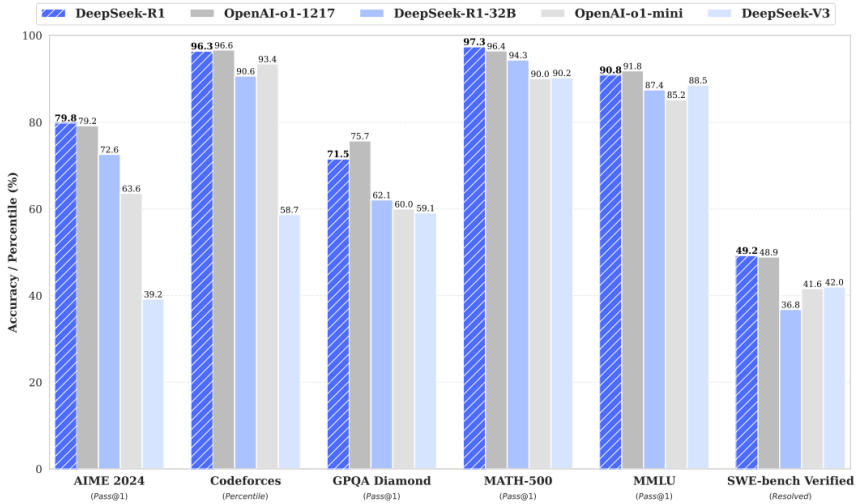


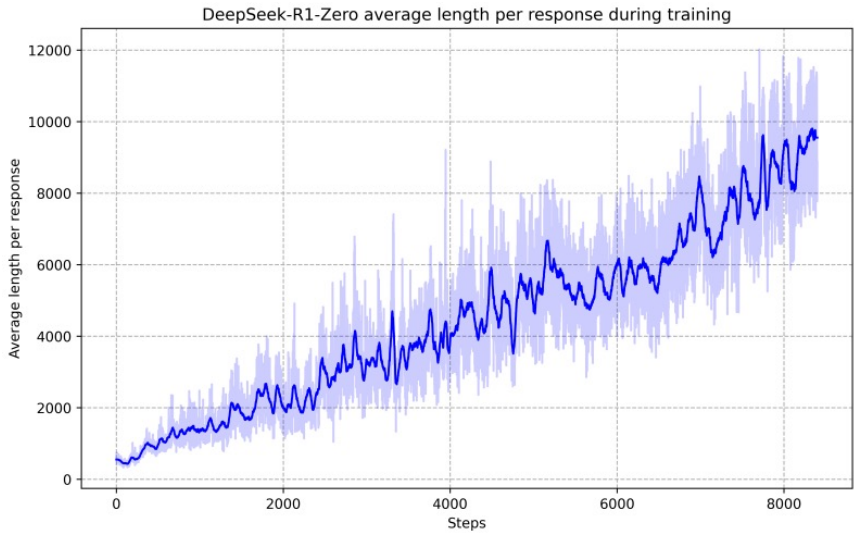
https://openai.com/index/learning-to-reason-with-llms/

Inference-Time Scaling – GPT



Inference-Time Scaling – DeepSeek



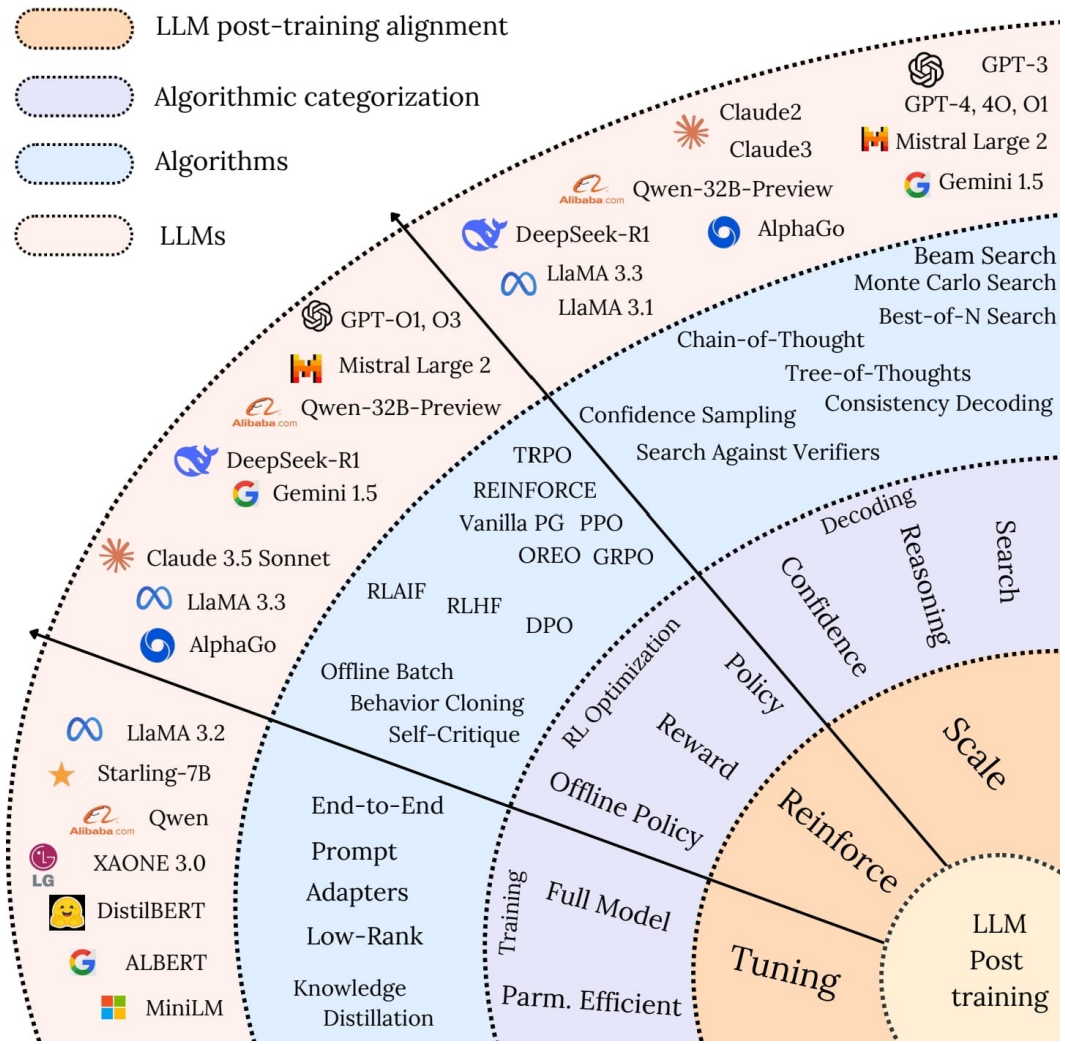


Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a-\sqrt{a+x}} = x$ is equal to
Response: <think> To solve the equation $\sqrt{a-\sqrt{a+x}} = x$, let's start by squaring both $\cdots$ $\left(\sqrt{a-\sqrt{a+x}}\right)^2 = x^2 \implies a-\sqrt{a+x} = x^2$. Rearrange to isolate the inner square root term: $(a-x^2)^2 = a+x \implies a^2-2ax^2+(x^2)^2 = a+x \implies x^4-2ax^2-x+(a^2-a) = 0$ $\ldots$ Wait, wait. Wait. That's an aha moment I can flag here. Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$ We started with the equation: $\sqrt{a-\sqrt{a+x}} = x$ First, let's square both sides: $a-\sqrt{a+x} = x^2 \implies \sqrt{a+x} = a-x^2$ Next, I could square both sides again, treating the equation: $\cdots$ $\ldots$

Table 3 | An interesting "aha moment" of an intermediate version of DeepSeek-R1-Zero. The model learns to rethink using an anthropomorphic tone. This is also an aha moment for us, allowing us to witness the power and beauty of reinforcement learning.

Kumar et al., LLM Post-Training: A Deep Dive into Reasoning Large Language Models, arXiv 2025.

Inference-Time Scaling



Failure Cases of GPT-4o

GPT 4o

“Four drums, seven tomatoes, and five candles.”

Failure Cases of GPT-4o

GPT 4o

“In a room, all the chairs are occupied except one.”

Goal

GPT 4o

FLUX + Inference-Time Scaling

“Four drums, seven tomatoes, and five candles.”

Goal

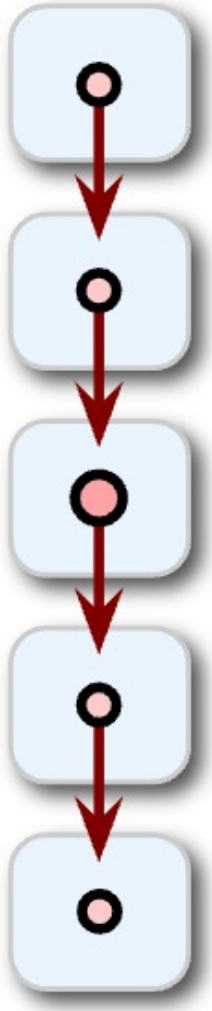
GPT 4o

FLUX + Inference-Time Scaling

"In a room, all the chairs are occupied except one."

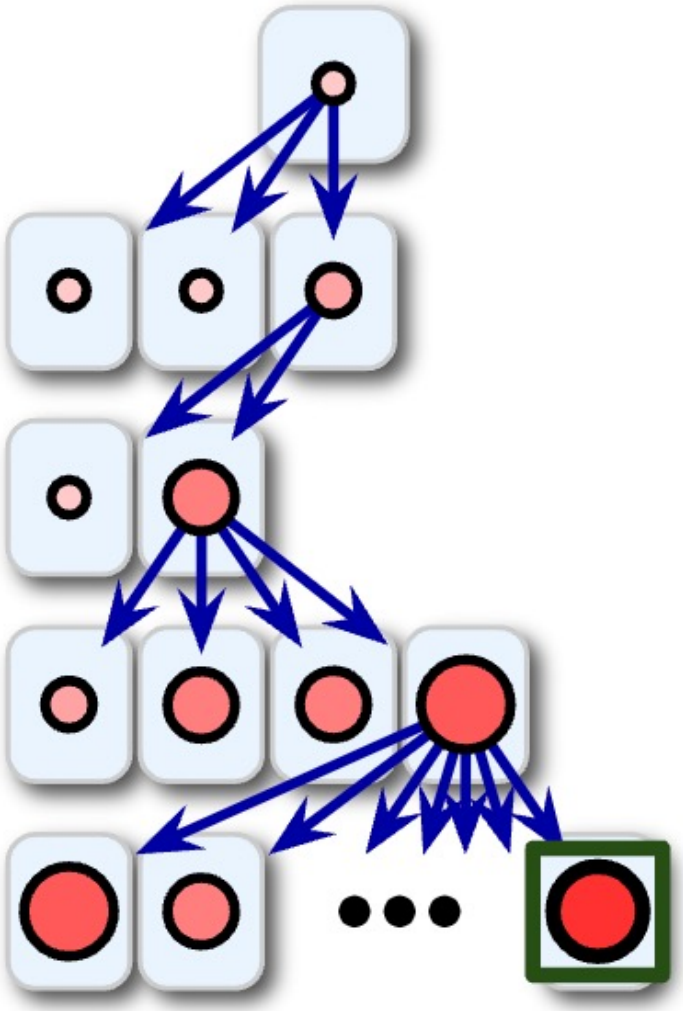
Particle Filtering

- The denoising process of a diffusion model can be viewed as a Sequential Monte Carlo process with a single trajectory.
- Only one particle is sampled at each step.



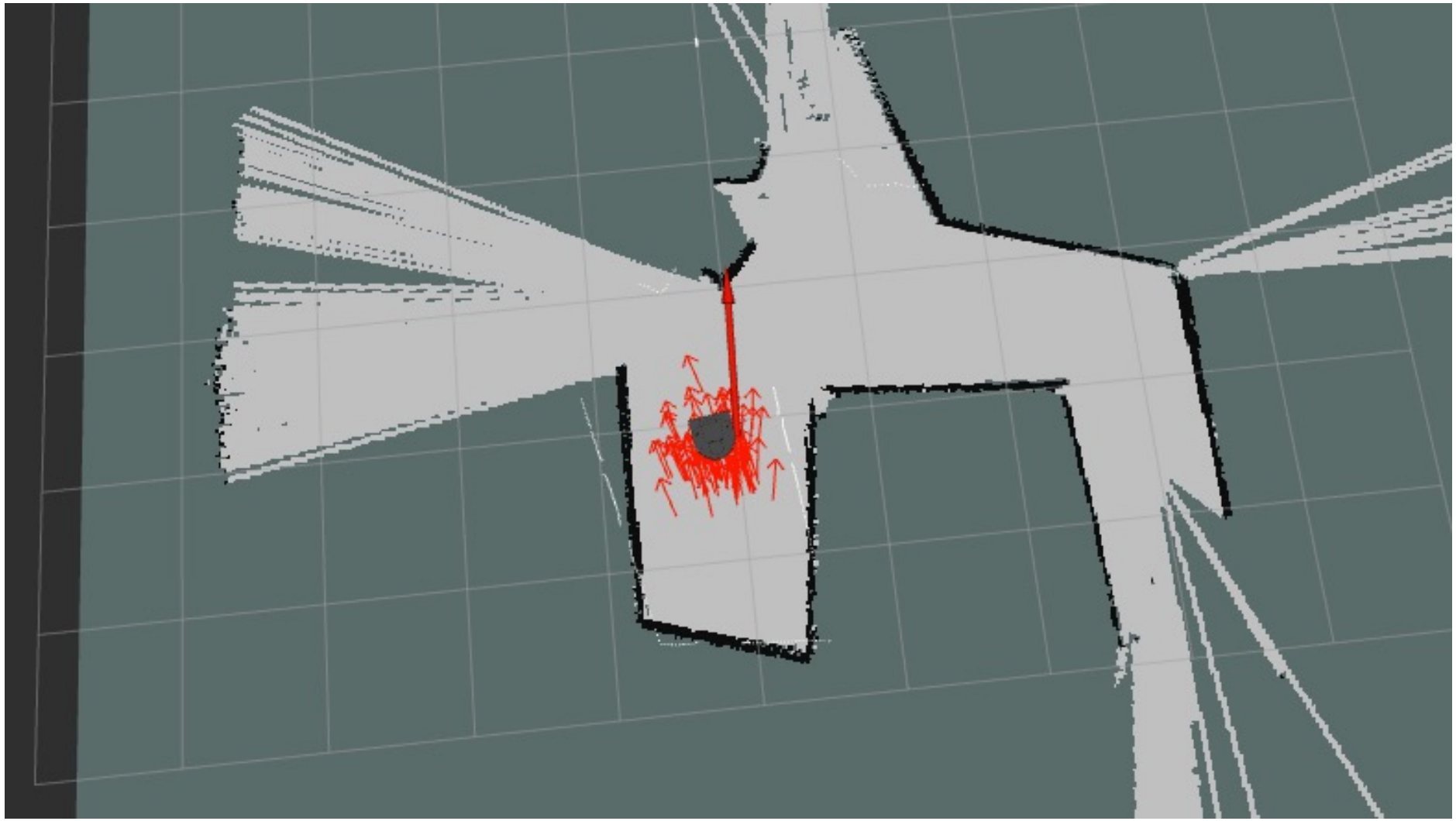
Particle Filtering

Could this be extended to particle filtering with multiple particles at each denoising timestep?



https://adinocap.com/building-a-particle-filter-for-a-simulated-neato/

Particle Filtering



Applications: Text Alignment

"Every painting in the gallery is framed and hung straight, except for one that is hanging crooked."

FLUX

Ours

Applications: Text Alignment

"The arcade machine is bigger than the television but smaller than the refrigerator."

FLUX

Ours

Applications: Text Alignment

"A large suitcase is placed beside an open closet,
with a folded jacket resting on top where a pair of shoes sit side by side in front of it."

FLUX

Ours

Applications: Text Alignment

(+) A nurse
(−) Stethoscope, hat, mask

FLUX

Ours

Applications: Object Count

2 cups, 3 paintings, 4 lamps, 4 bananas.

FLUX

Ours

Applications: Object Count

7 balloons, 4 bears, 4 swans.

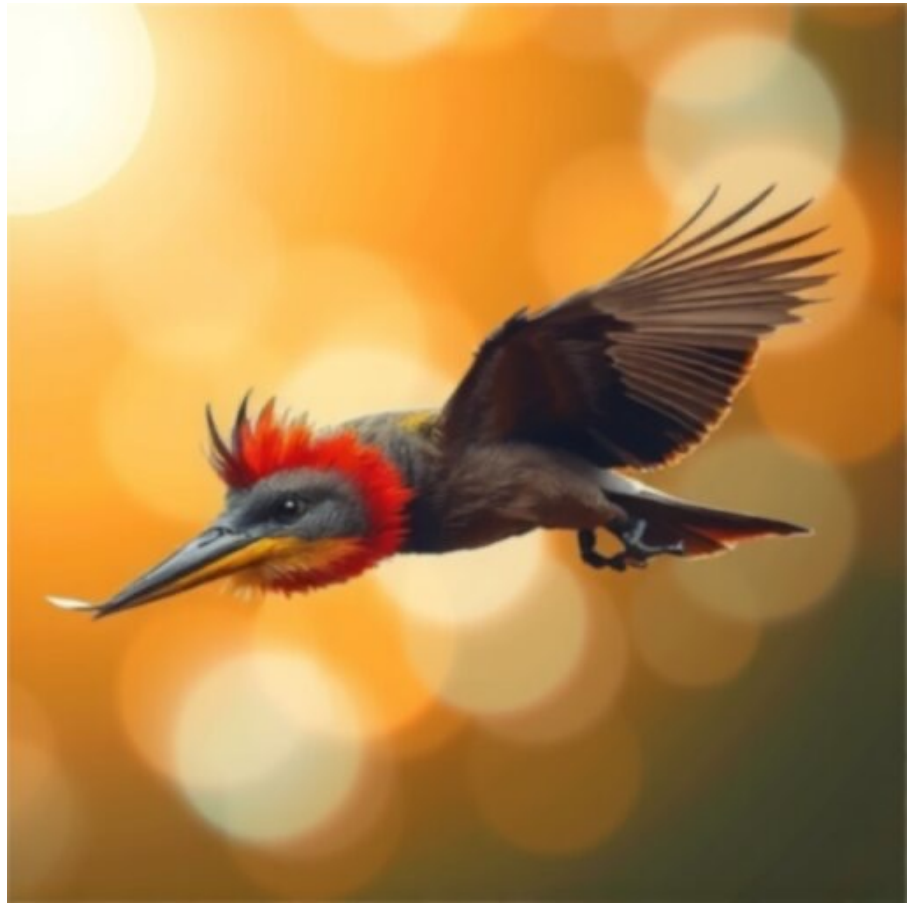
Best of N

Ours

Applications: Aesthetic Image Generation

Bird

Best of N



Ours

Applications: Aesthetic Image Generation

Bat

Best of N

Ours

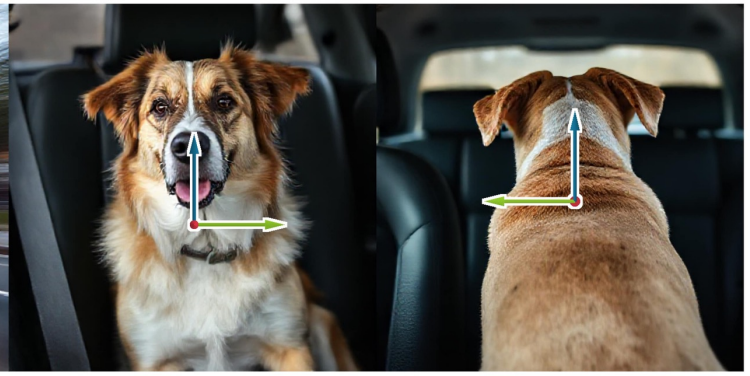
Application: Orientation Grounding

3D orientation grounding for multiple open-vocabulary objects in image generation.

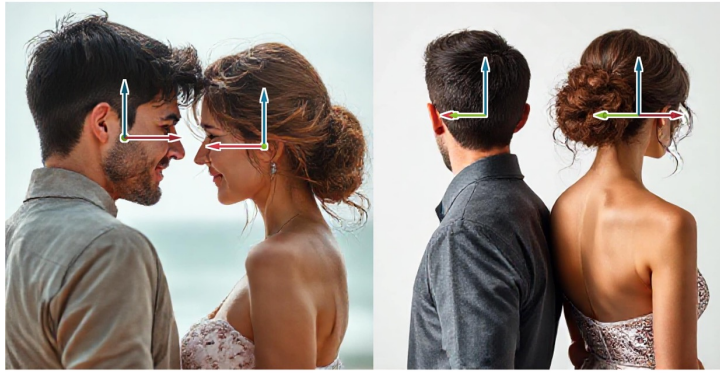
*"a **teddy bear** holding a stick next to a block toy."*

*"a **bus** driving down the road really fast."*

*"a **dog** that is sitting down in a backseat."*

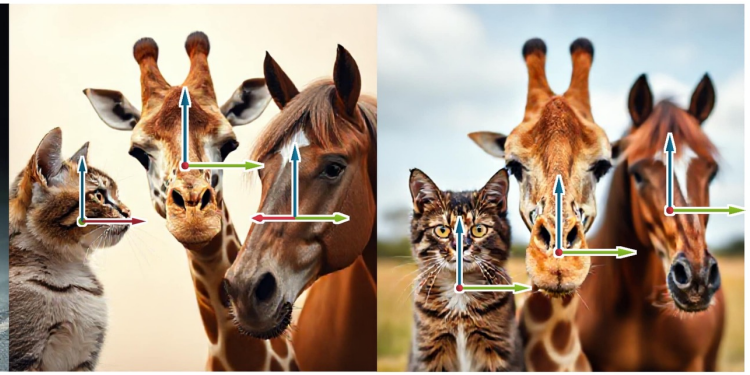


*"a **man**, and a **woman**."*



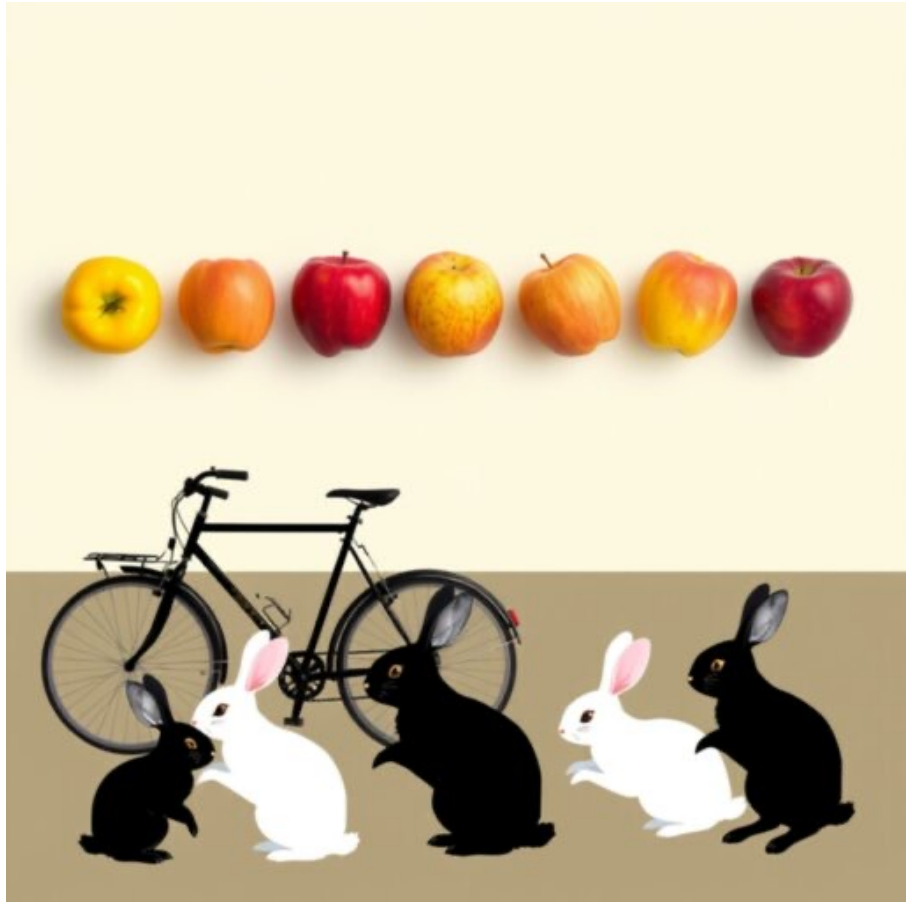
*"a **motorcycle**, and a **bear**."*

*"a **cat**, a **giraffe**, and a **horse**."*



Applications: Object Count

8 apples, 3 bicycles, 5 rabbits.



Best of N

Ours

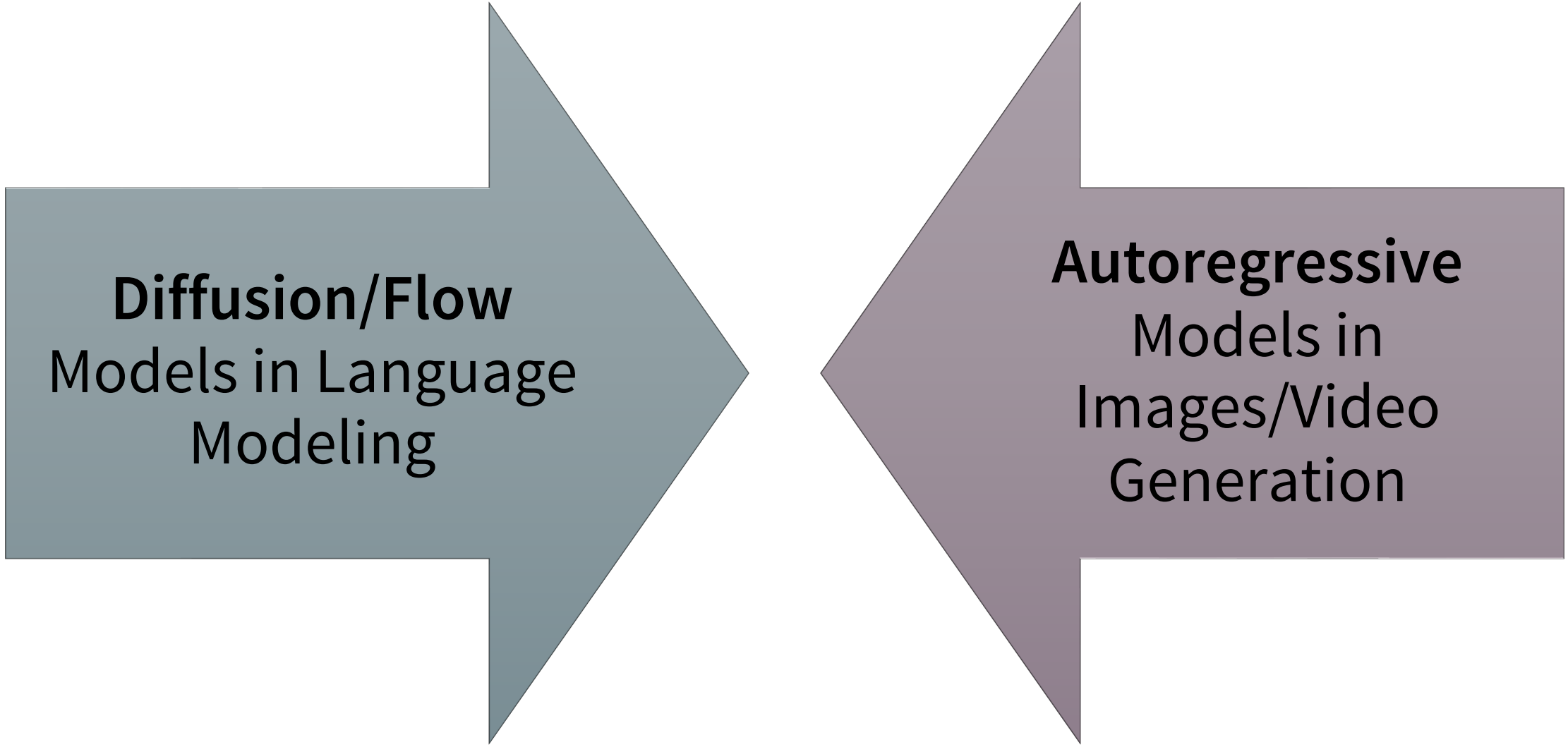
Applications: Object Count

82 Blueberries

33 Blueberries

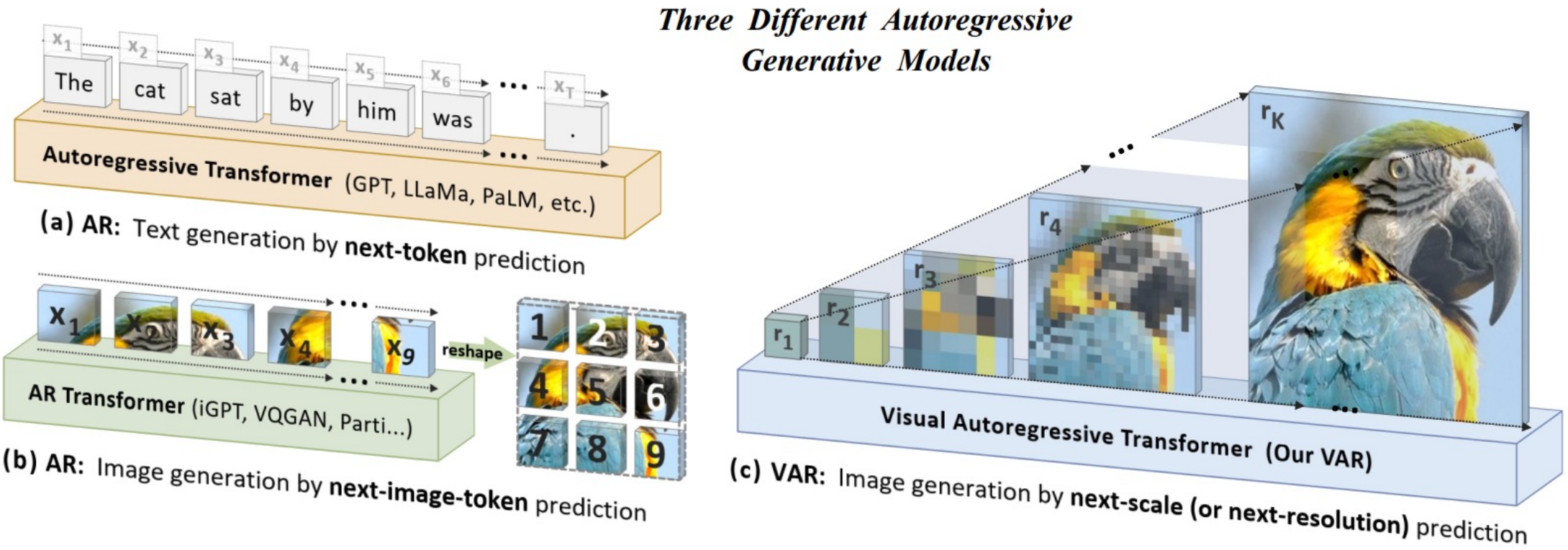
What’s the Next?

Diffusion/Flow vs. Autoregressive



Tian et al., Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction, NeurIPS 2024.

Visual Autoregressive Model



Visual Autoregressive Model

- (−) Lower image quality and lower generation speed.
- (−) Still a long way to go for efficient fine-tuning or training-free guided generation.
- (+) Could be better suited for integration with multimodal foundation models.

https://x.com/InceptionAILabs/status/1894847919624462794

Large Language Diffusion Models

Mercury, the first commercial-grade diffusion LLM by Inception Labs

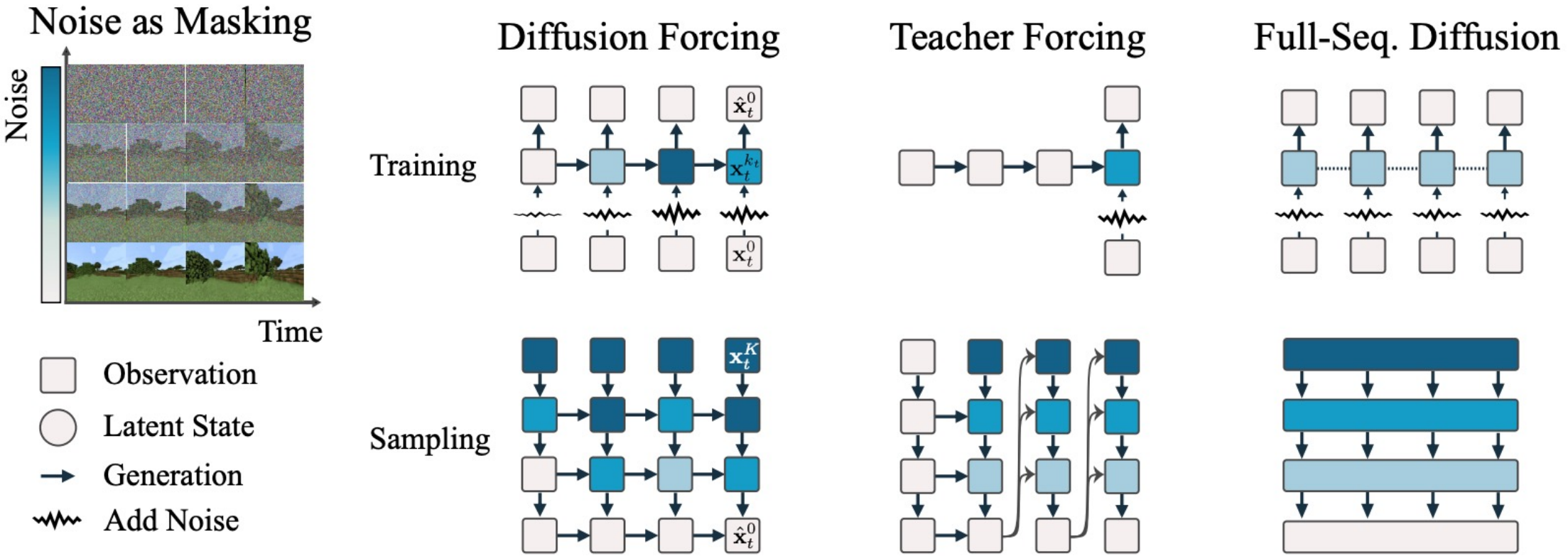
Diffusion/Flow vs. Autoregressive

Controllability will be one of the key factors in determining the winner!

Hybrid: Autoregressive Diffusion Models

Diffusion/flow for each item;

autoregressive for the entire set, sequence, or graph.



Hybrid: Autoregressive Diffusion Models

- Video generation
- Other time-series data: audio, finance, weather, health, physics simulation, etc.
- Set/graph generation: multi-view, scene graph, CAD, moledule, etc.

How can inference-time scaling be specialized for autoregressive diffusion models?

The Present and Future of Image and Video Generation Technologies

Minhyuk Sung

KAIST Visual AI Group

https://visualai.kaist.ac.kr/