# Diffusion Models: A Generative AI Big Bang

Arash Vahdat, Karsten Kreis

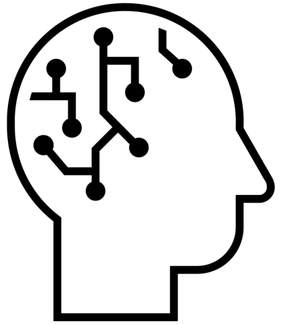
GTC, March 18th 2024

# Deep Generative Learning

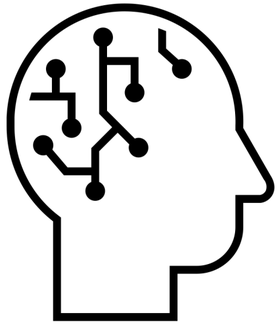
## Learning to generate data

Samples from a Data Distribution

Train

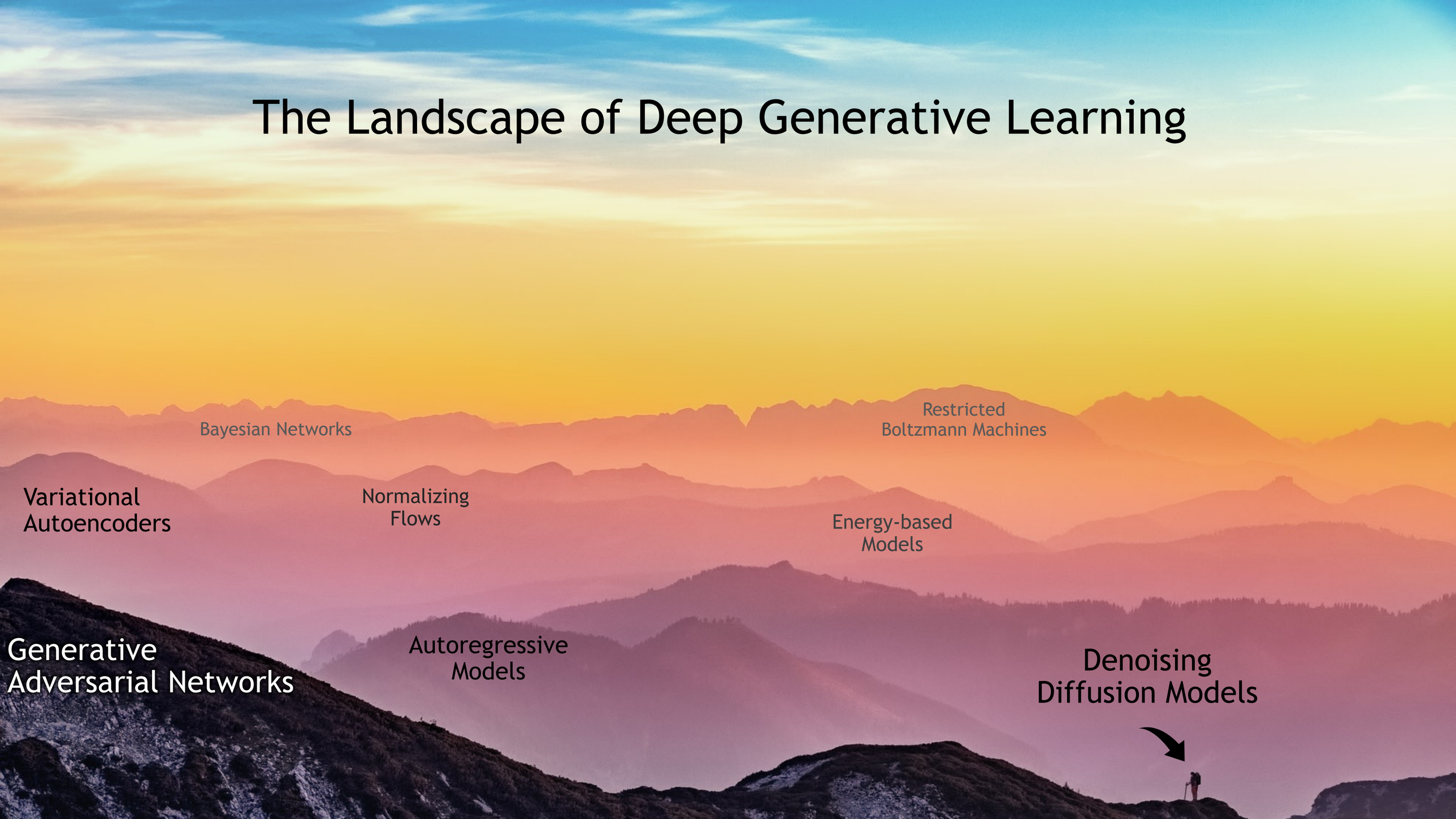


Neural Network



Sample

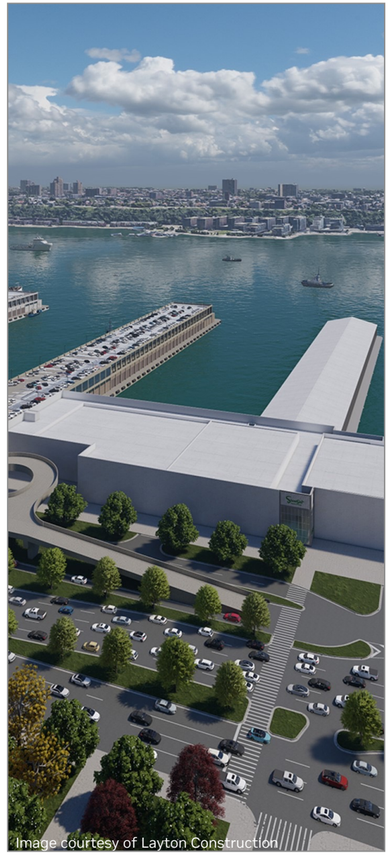
# The Landscape of Deep Generative Learning




DALL-E-3


DALL-E-3

# Generative AI Applications




Architecture / Design


Feature Learning


Film / Video

3D FX / Game Dev

Marketing


Photography

Slide: Courtesy of Ashlee Martino-Tarr and Daniela Flamm Jackson

# Agenda

- An Introduction to Diffusion Models
- Acceleration
- Conditioning & Guidance
- Personalization
- Latent Diffusion Models
- Video Diffusion Models
- 3D and 4D Generation

# Agenda

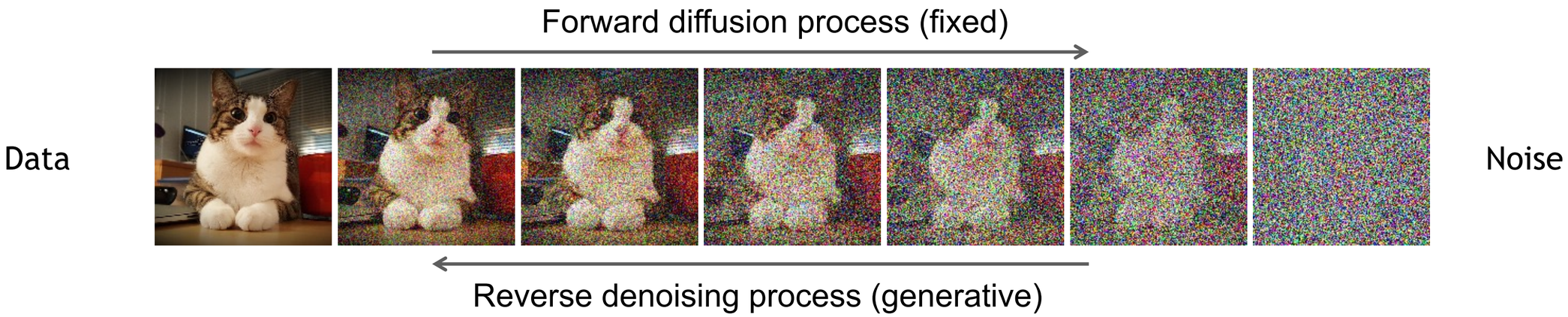
- An Introduction to Diffusion Models
- Acceleration
- Conditioning & Guidance
- Personalization
- Latent Diffusion Models
- Video Diffusion Models
- 3D and 4D Generation

# Denoising Diffusion Models

## Learning to generate by denoising

Denoising diffusion models (a.k.a. score-based generative models) consist of two processes:

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

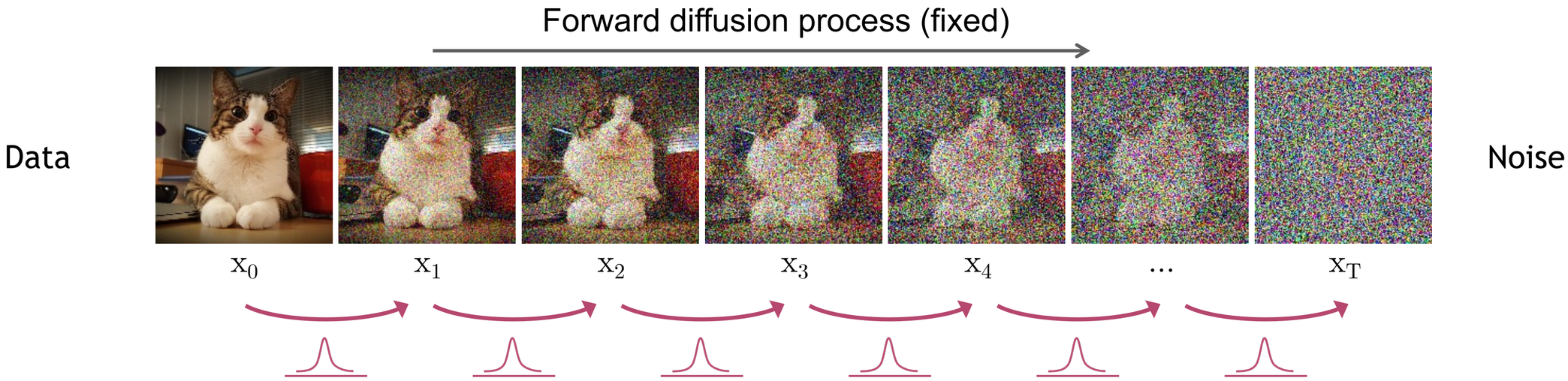




Sohl-Dickstein et al., Deep Unsupervised Learning using Nonequilibrium Thermodynamics, ICML 2015
Ho et al., Denoising Diffusion Probabilistic Models, NeurIPS 2020
Song et al., Score-Based Generative Modeling through Stochastic Differential Equations, ICLR 2021

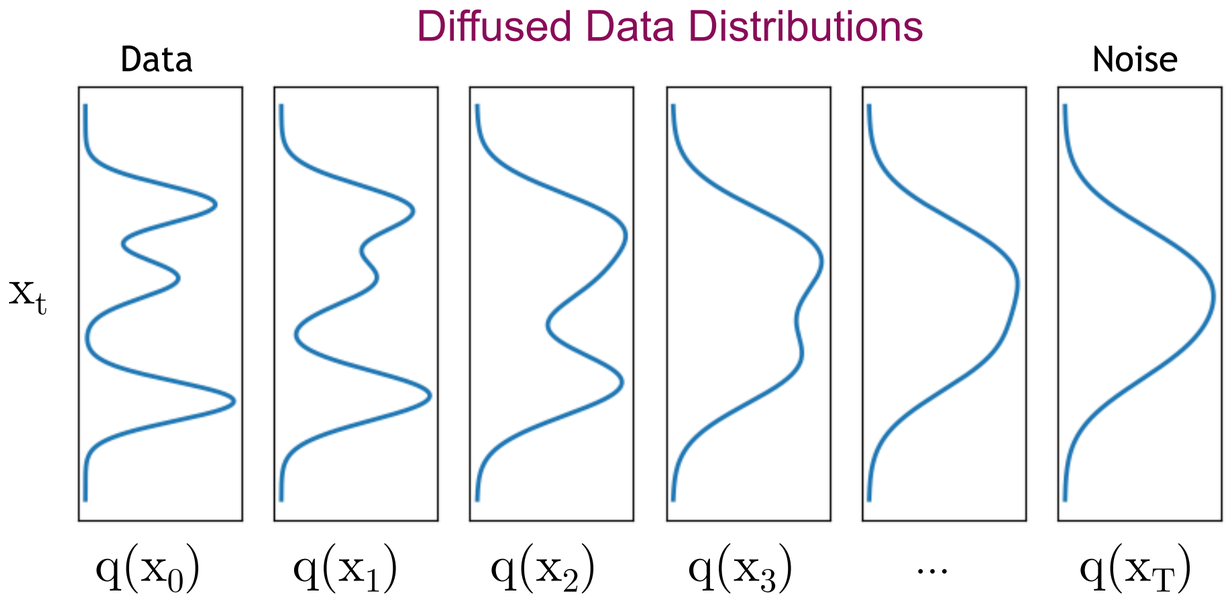
# Forward Diffusion Process

The formal definition of the forward process in T steps:




$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x_t}; \sqrt{1-\beta_t}\mathbf{x_{t-1}}, \beta_t\mathbf{I})$$

# What happens to a distribution in the forward diffusion?

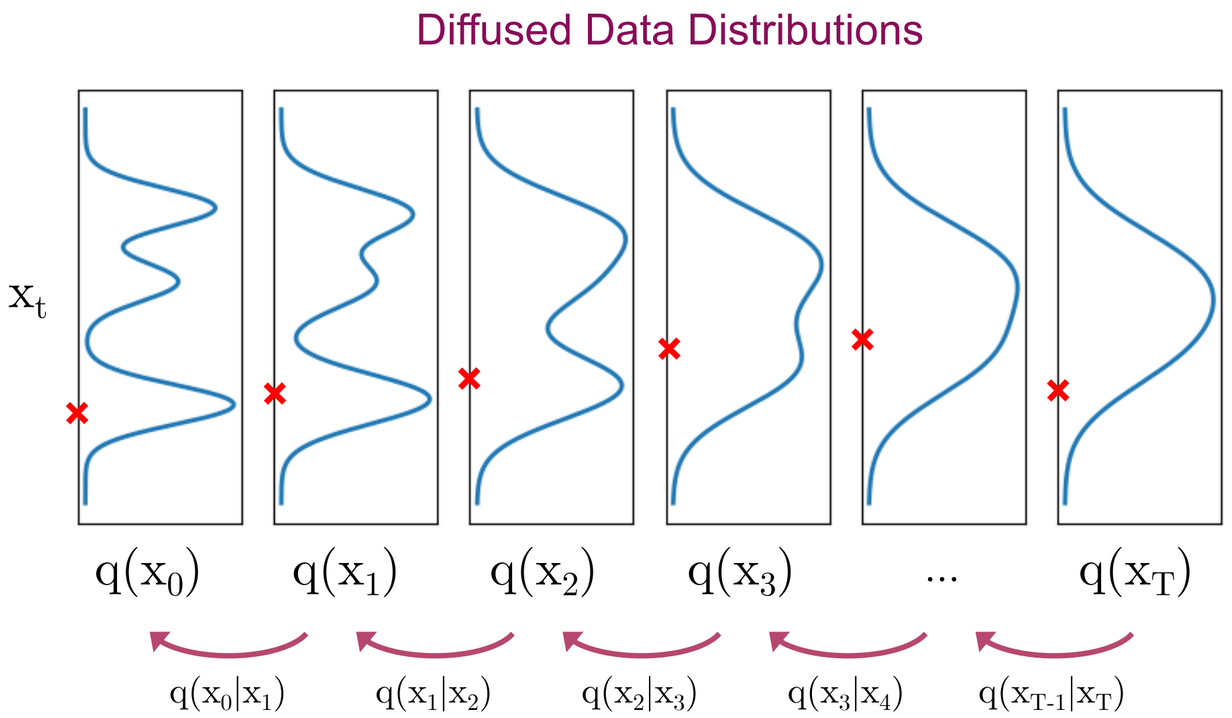



Diffusion kernel as a Gaussian convolution.

We can sample $\mathbf{x}_t \sim q(\mathbf{x}_t)$ by diffusing training data samples.

# Generative Learning by Denoising

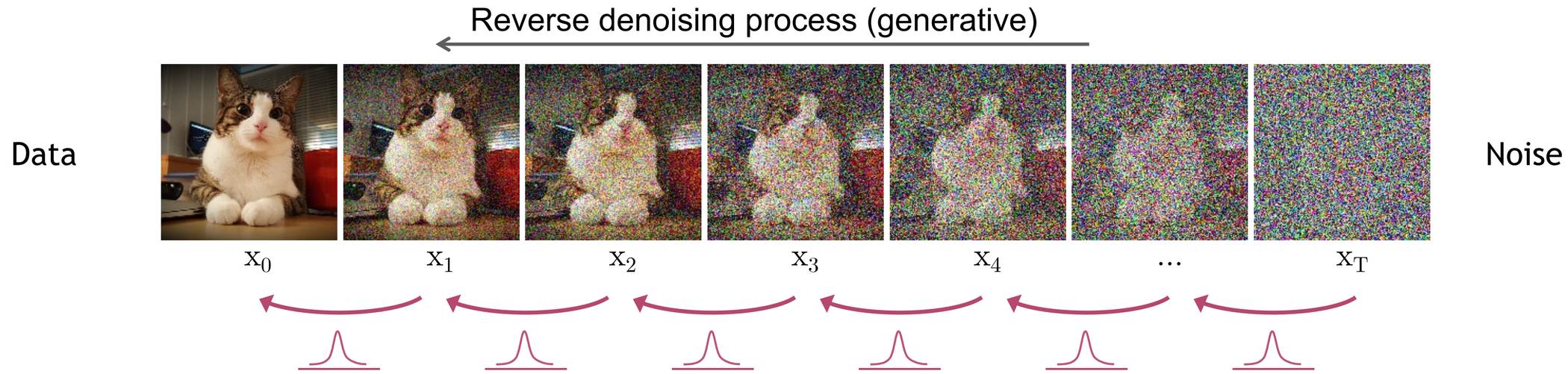
Recall, that the diffusion process is designed such that $q(\mathbf{x}_T) \approx \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I}))$




**True Denoising Distribution**

# Reverse Denoising Process

Formal definition of forward and reverse processes in T steps:

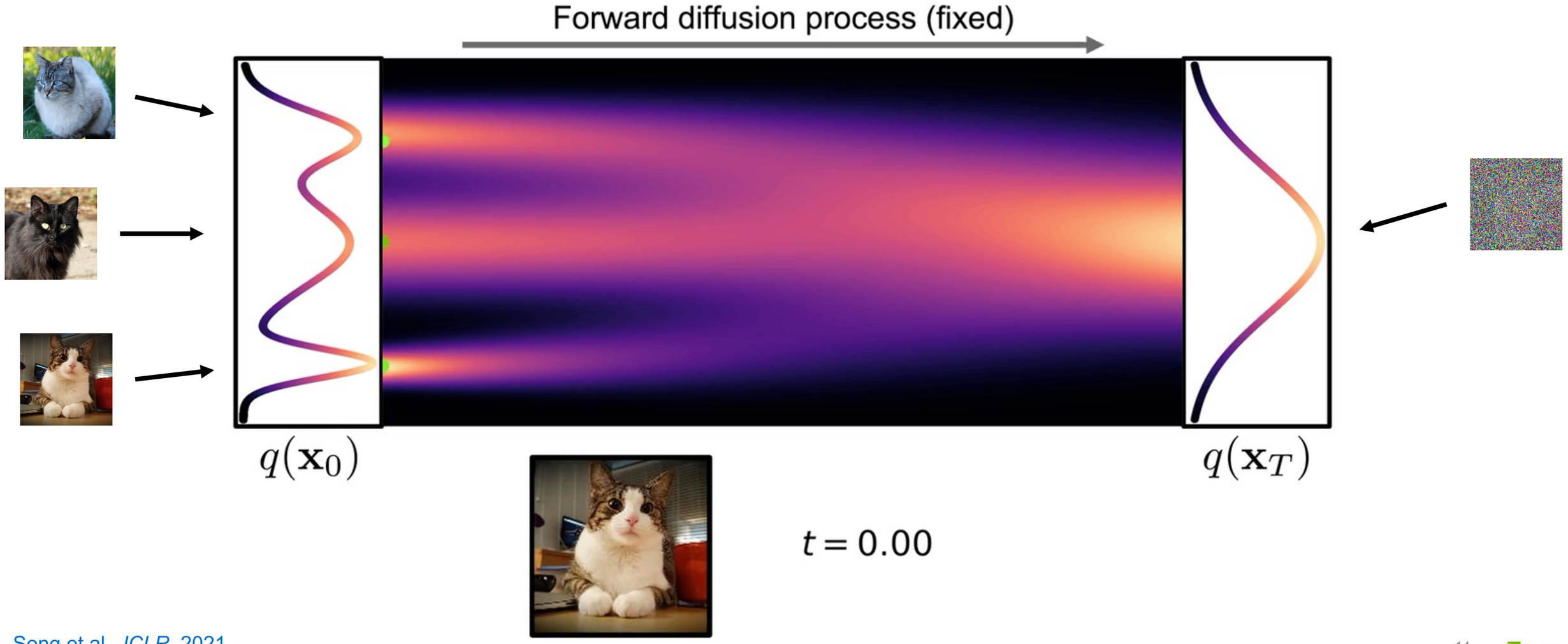



$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \underbrace{\mu_\theta(\mathbf{x}_t, t)}, \sigma_t^2\mathbf{I})$$

Trainable network
(U-net, Denoising Autoencoder)

# Forward Diffusion Process as Stochastic Differential Equation

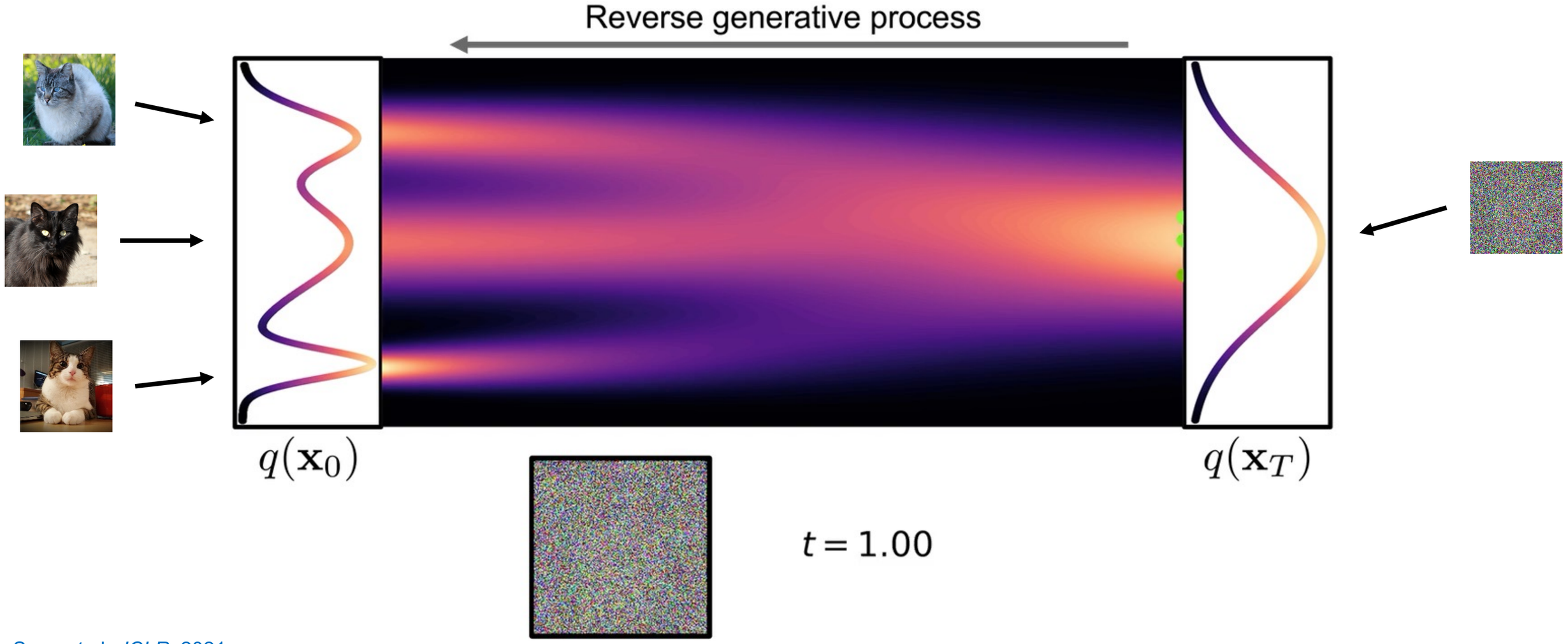
## Diffusion models in infinite steps!




Song et al., *ICLR*, 2021

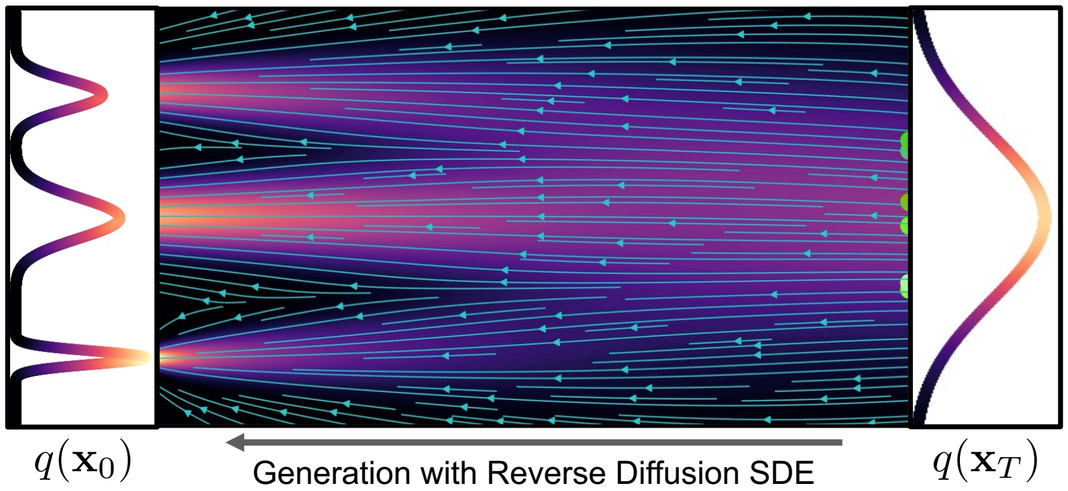
# The Generative Reverse Stochastic Differential Equation

## Diffusion models in infinite steps!

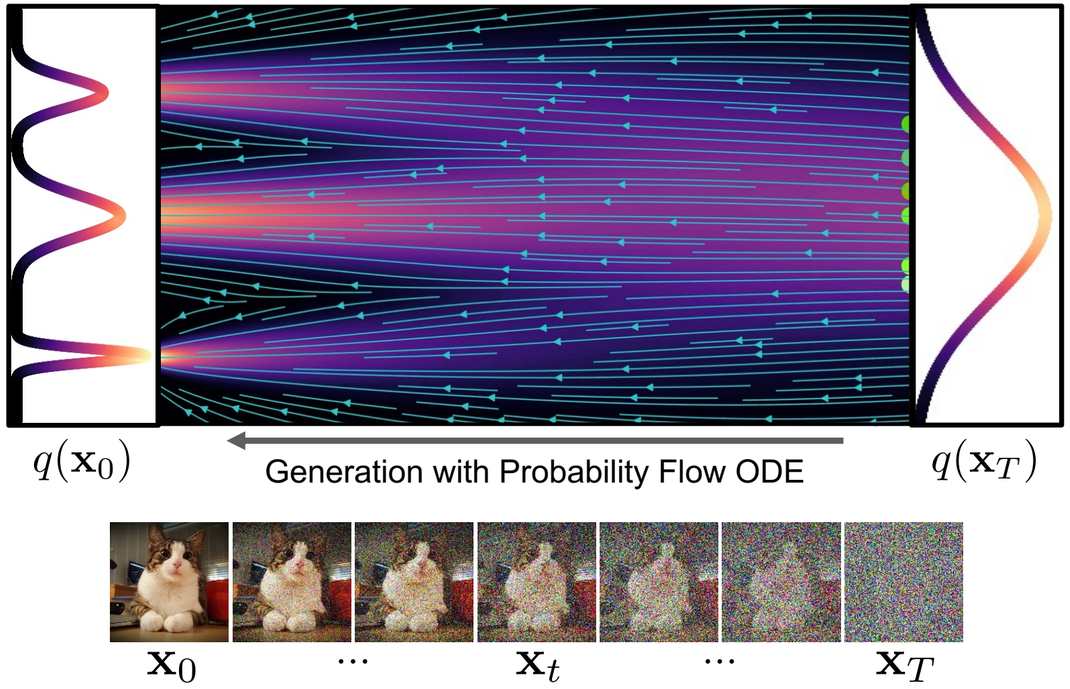



Song et al., *ICLR*, 2021

# Synthesis with SDE vs. ODE




**Generative Reverse Diffusion SDE (stochastic)**



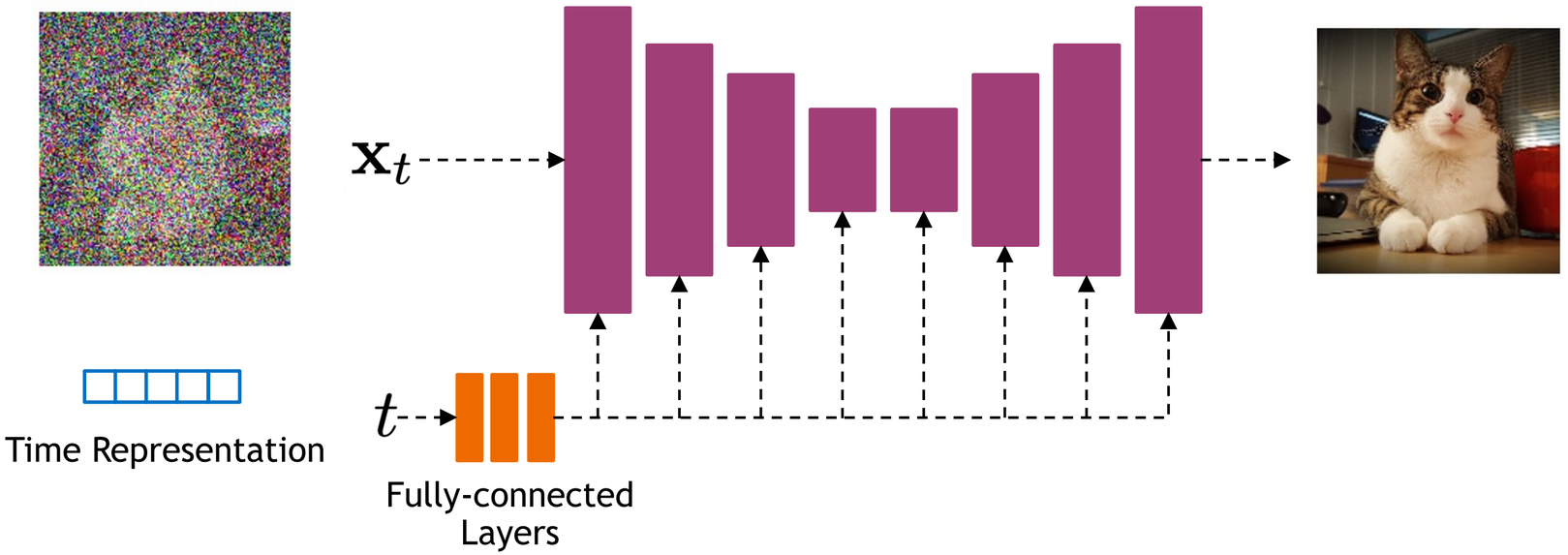

**Generative Probability Flow ODE (deterministic)**

Song et al., *ICLR*, 2021

# Training Diffusion Models

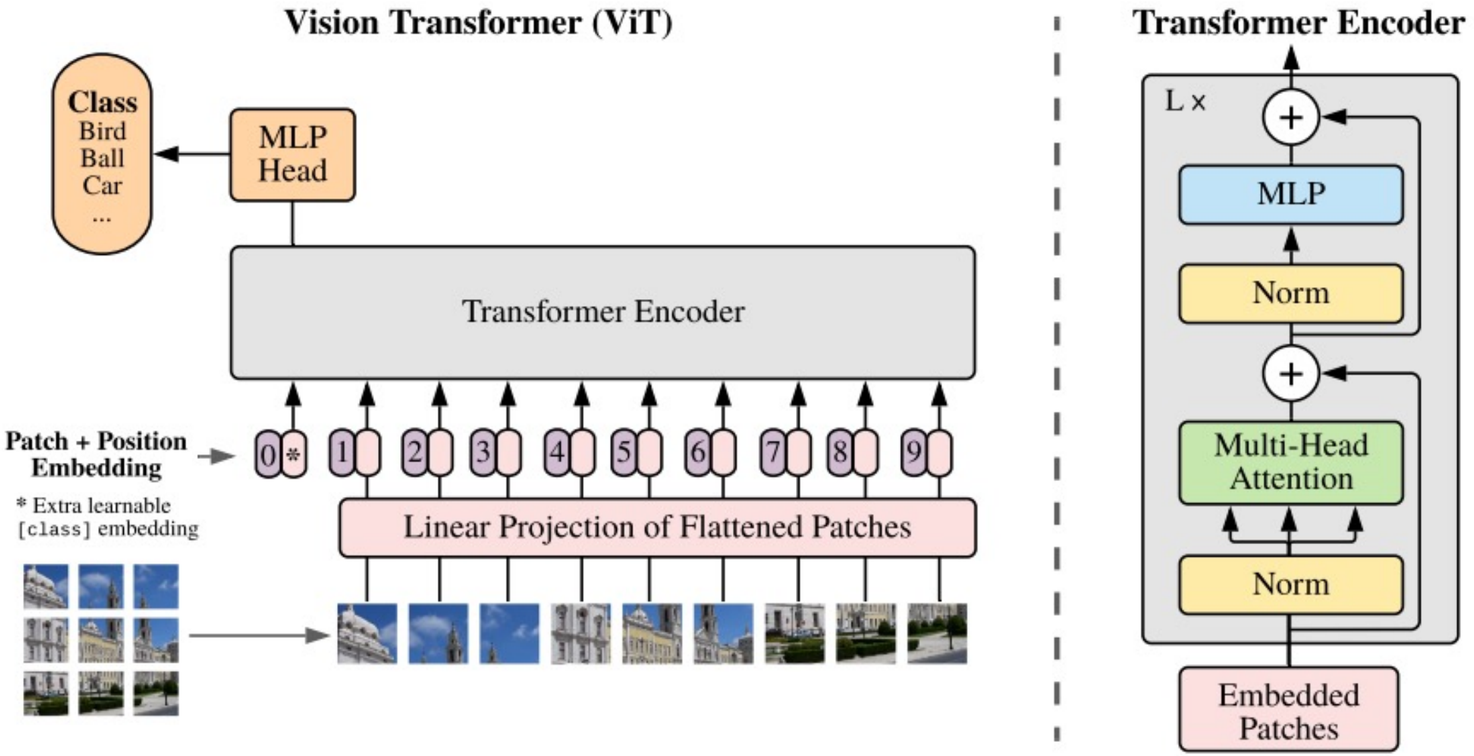
## Network Architectures

Diffusion models often use U-Net architectures with residual convolutional neural nets to represent the denoising model



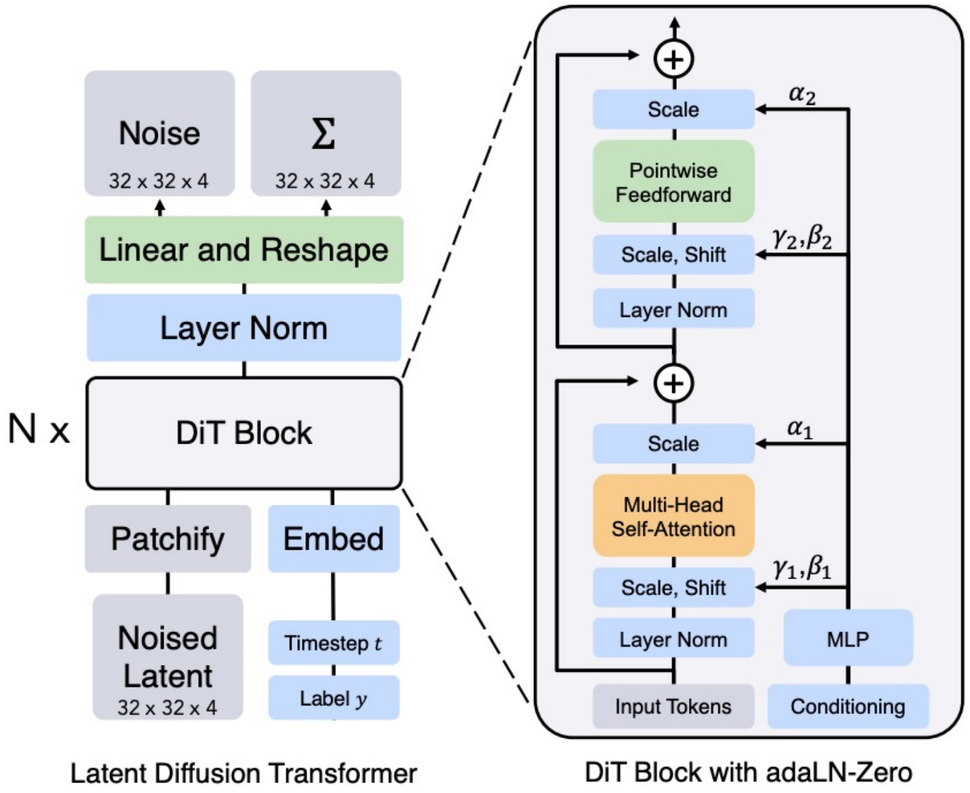

Time representation: sinusoidal positional embeddings.

# Transformers for Diffusion Models

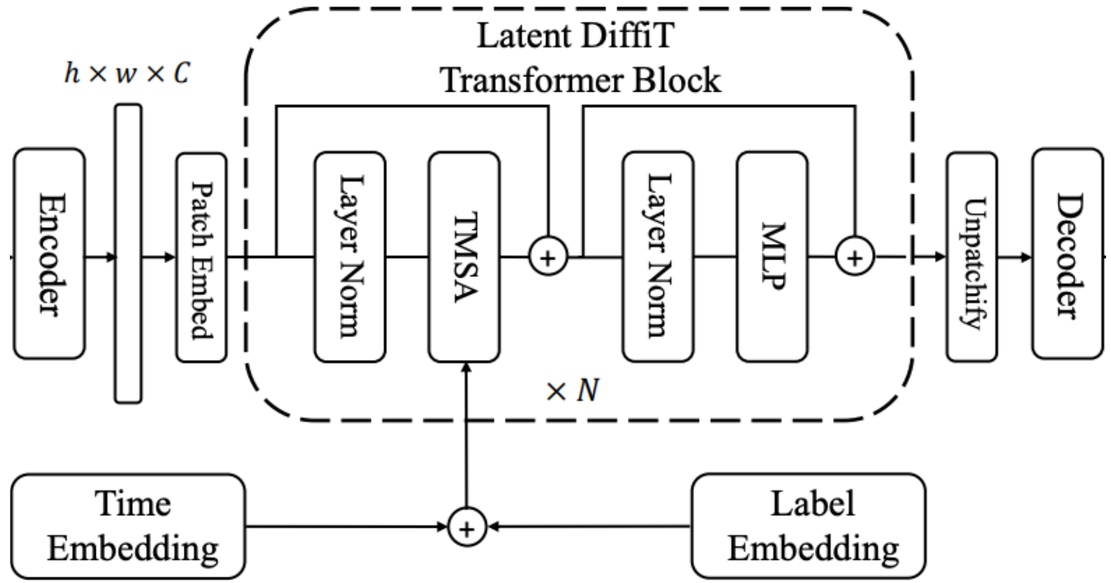



Dosovitskiy et al. ICLR 2021

# Transformers for Diffusion Models



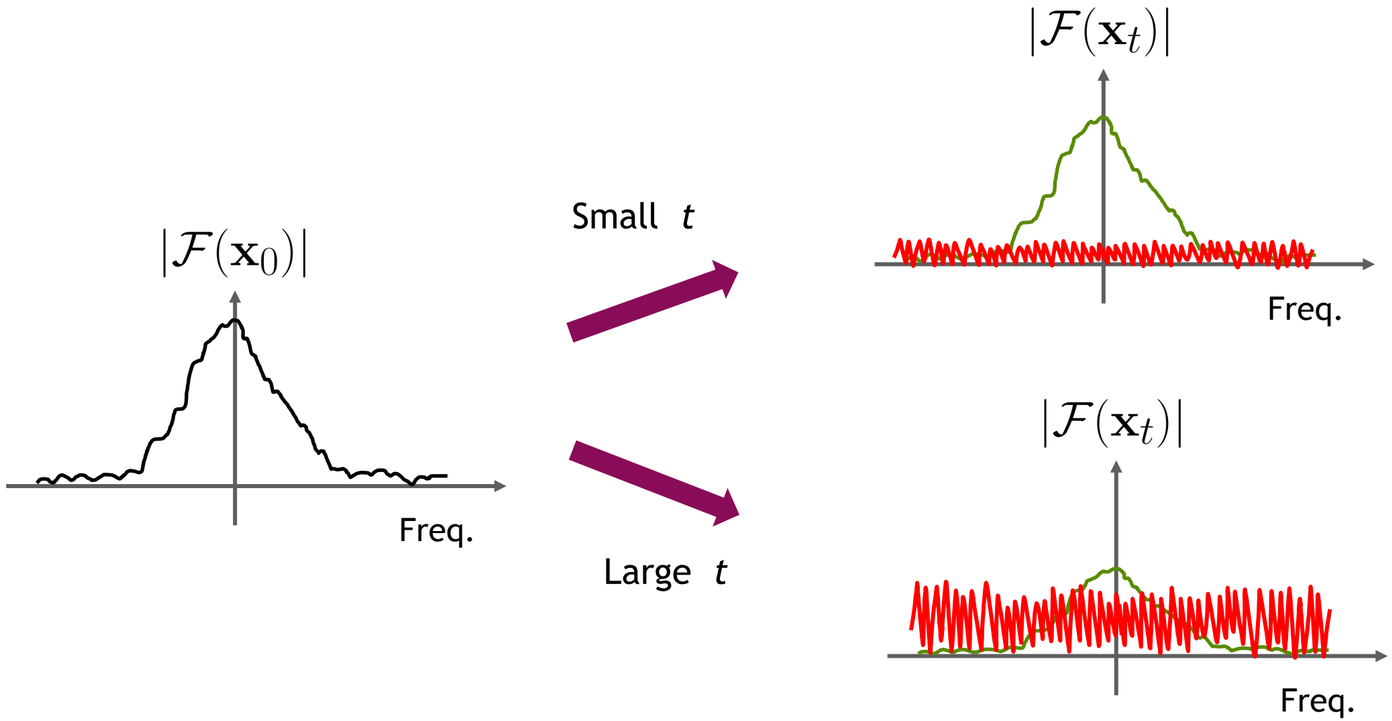

DiT by Peebles and Xie




DiffiT by Hatamizadeh et al.

Sora

# What happens to an image in the forward diffusion process?




Data

Noise

**In the forward diffusion, the high frequency content is perturbed faster.**

# Content-Detail Tradeoff

Reverse denoising process (generative)

Data


Noise

The denoising model is specialized for generating the high-frequency content (i.e., low-level details)

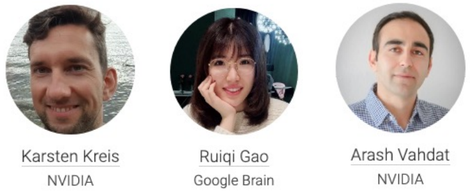
The denoising model is specialized for generating the low-frequency content (i.e., coarse content)

The weighting of the training objective for different timesteps is important!

# Tutorials on Diffusion Models

**CVPR 2022:** *Denoising Diffusion-based Generative Modeling: Foundations and Applications*

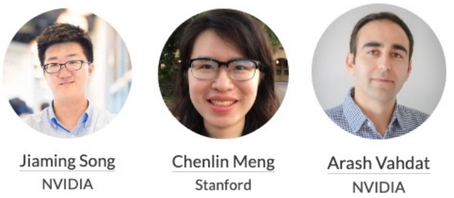
Website (~4 hours long, over 100,000 views on Youtube):
https://cvpr2022-tutorial-diffusion-models.github.io/




---

**CVPR 2023:** *Denoising Diffusion Models: A Generative Learning Big Bang*

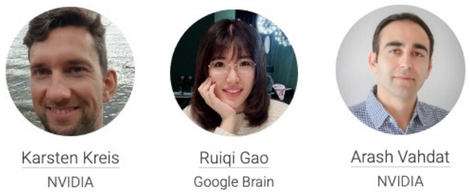
Website:
https://cvpr2023-tutorial-diffusion-models.github.io/




---

**NeurIPS 2023:** *Latent Diffusion Models: Is the Generative AI Revolution Happening in Latent Space?*

Website:
https://neurips2023-ldm-tutorial.github.io/

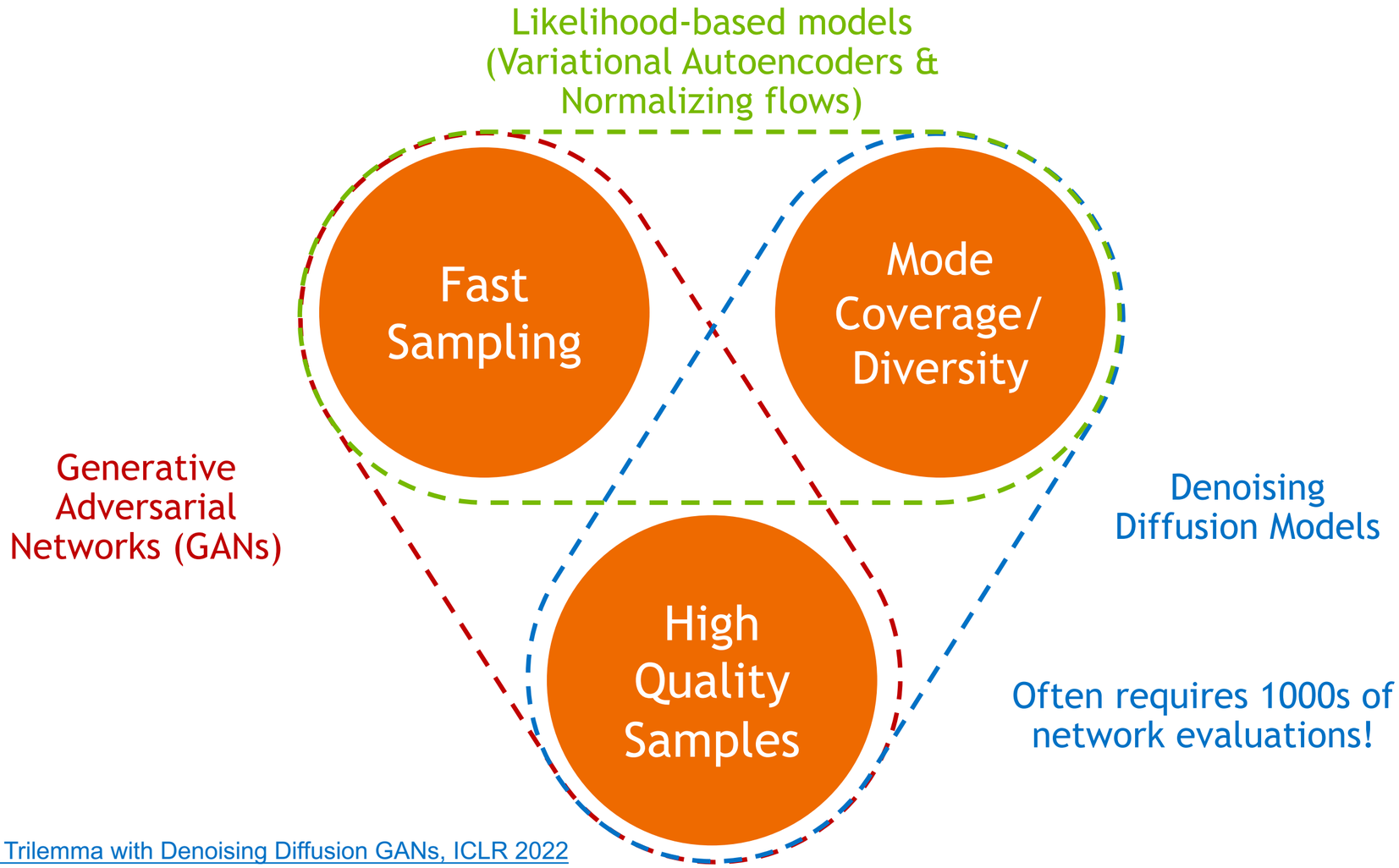


# Agenda

- An Introduction to Diffusion Models
- **Acceleration**
- Conditioning & Guidance
- Personalization
- Latent Diffusion Models
- Video Diffusion Models
- 3D and 4D Generation

# What makes a good generative model?

## The generative learning trilemma

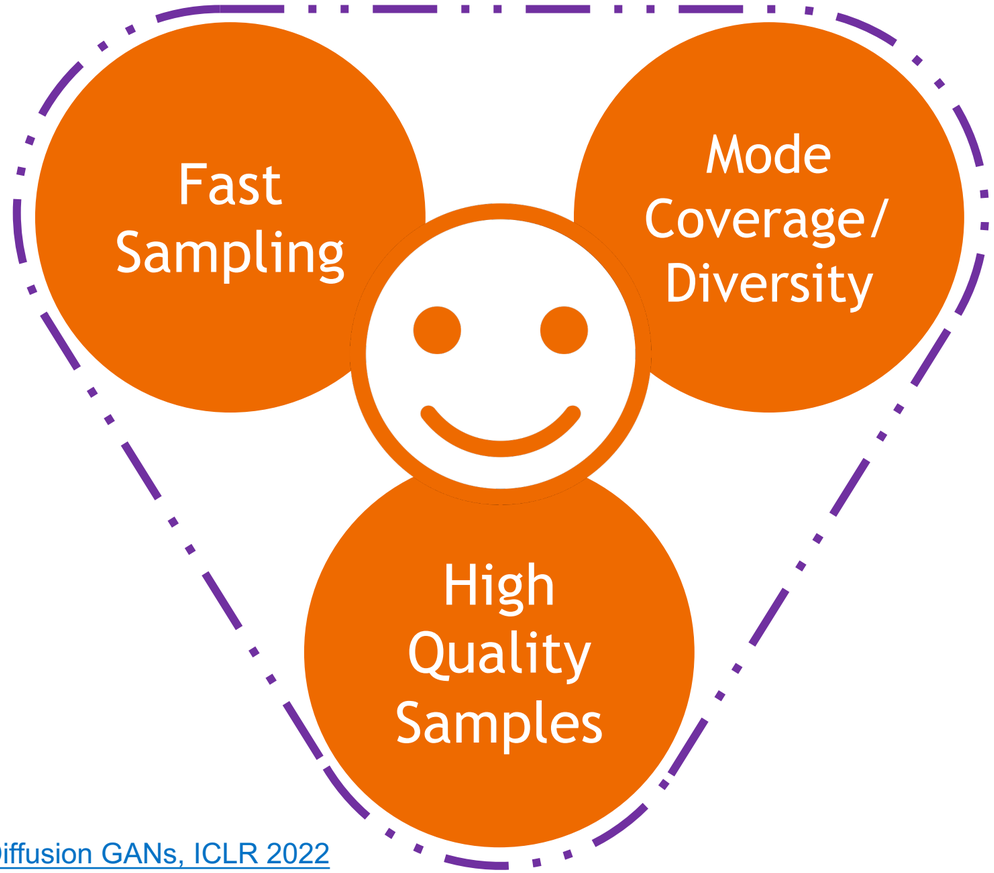



Tackling the Generative Learning Trilemma with Denoising Diffusion GANs, ICLR 2022

# What makes a good generative model?

## The generative learning trilemma

Tackle the trilemma by accelerating diffusion models



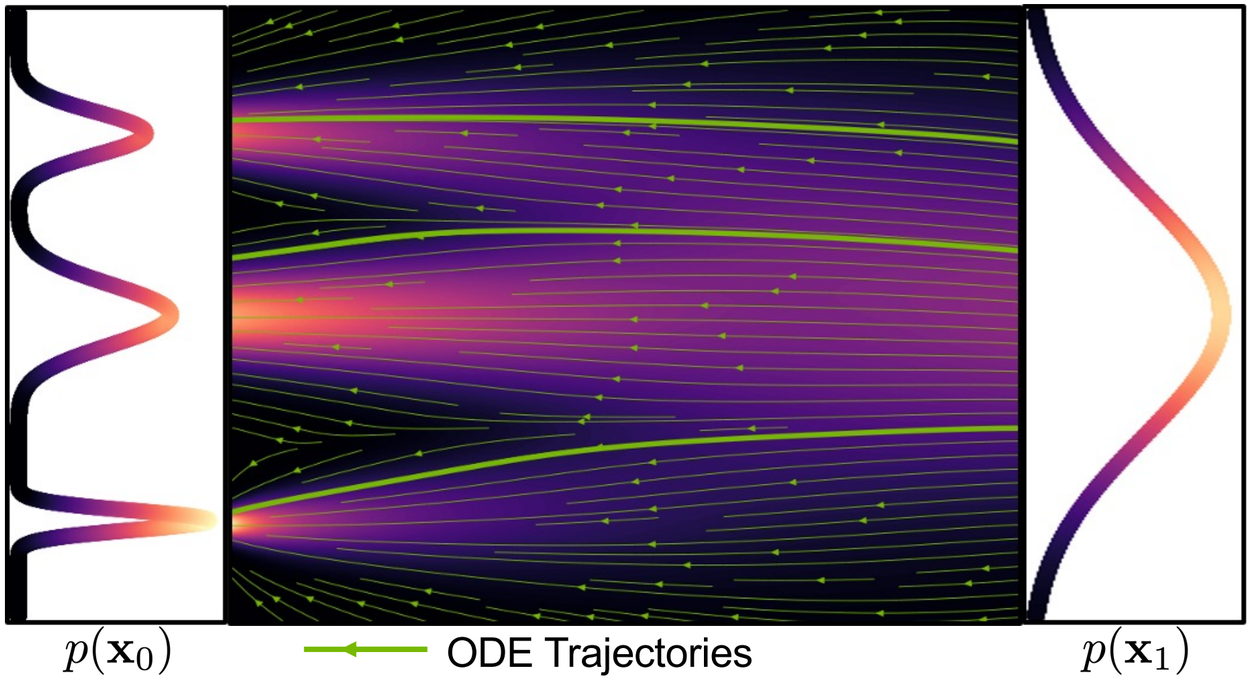

Tackling the Generative Learning Trilemma with Denoising Diffusion GANs, ICLR 2022

# Acceleration Techniques

# Acceleration Techniques

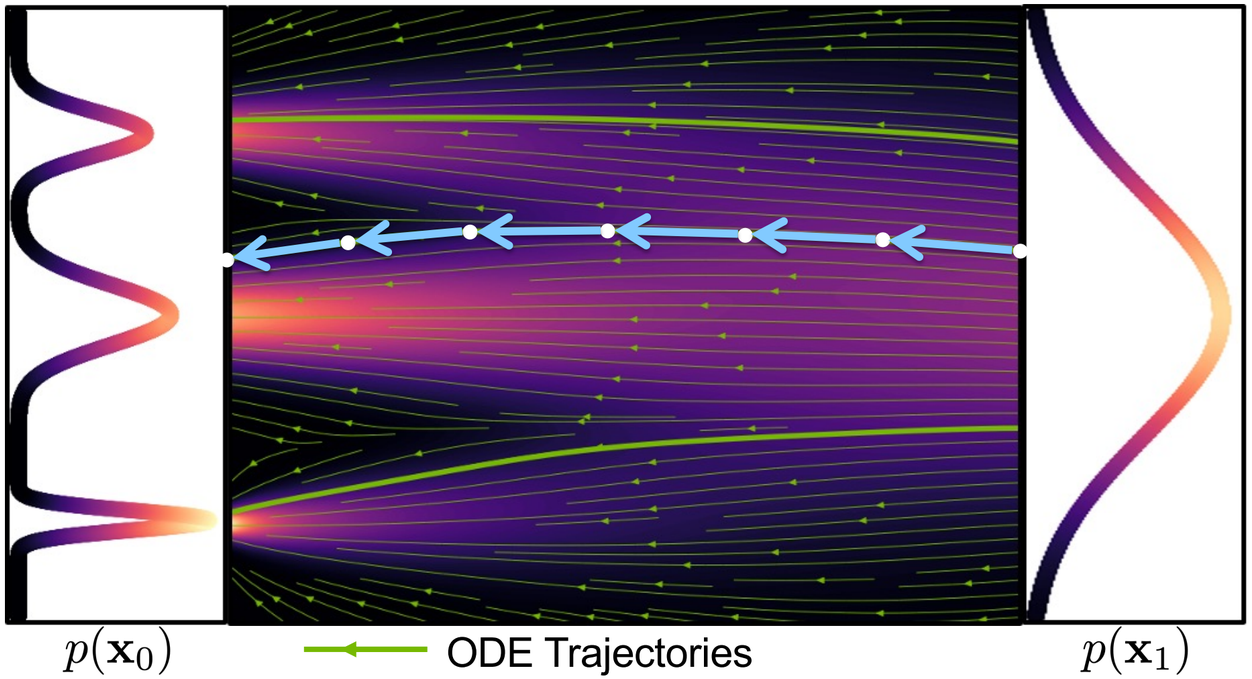
# Generative ODEs

## Solve ODEs with as little function evaluations as possible



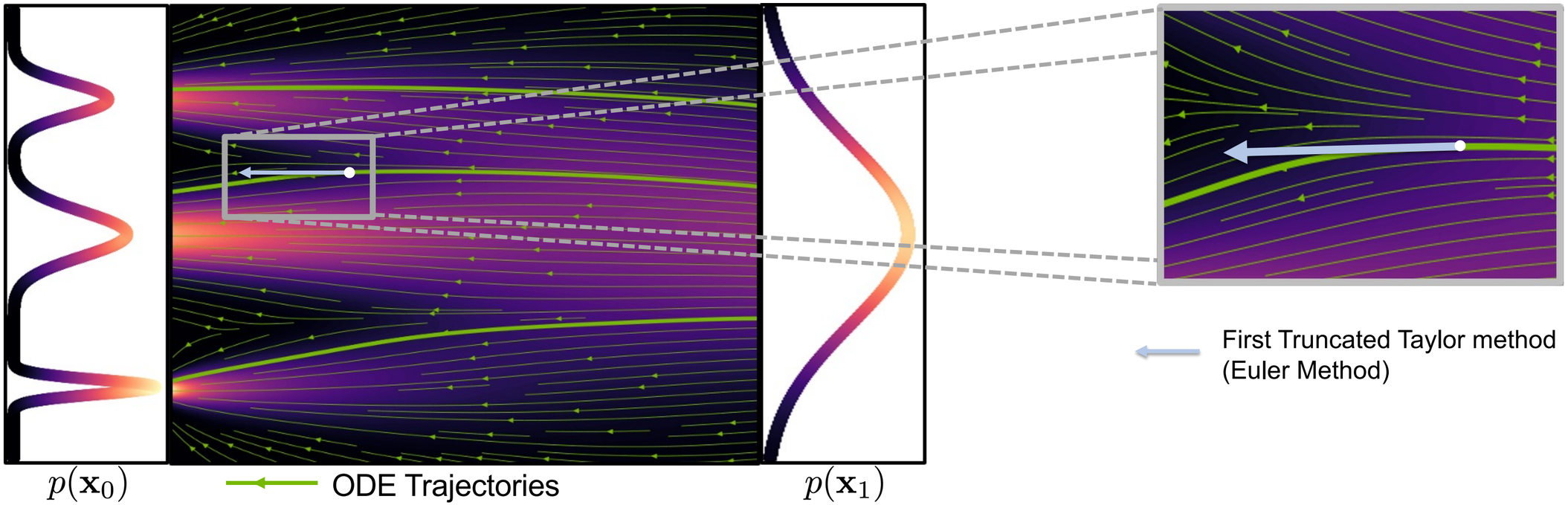
# Generative ODEs

## Solve ODEs with as little function evaluations as possible



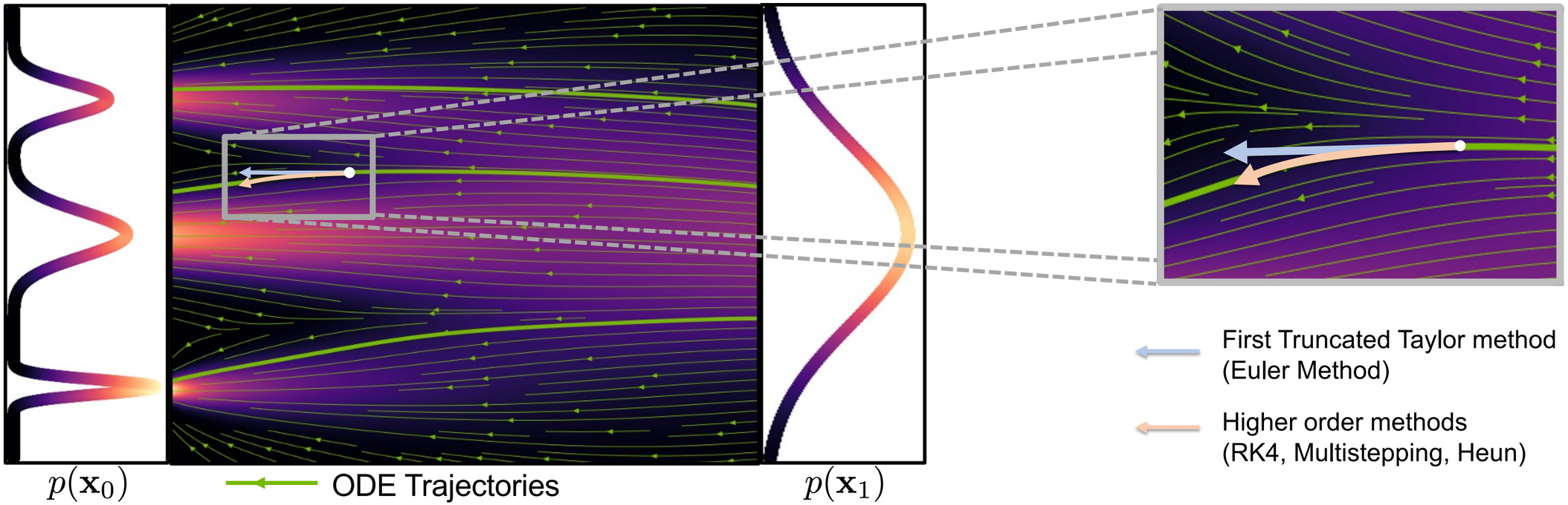
# Generative ODEs

## Solve ODEs with as little function evaluations as possible



# Generative ODEs

## Solve ODEs with as little function evaluations as possible



# Higher-Order ODE Solvers

DPM-Solver++(2M)
$(N = 15)$

DPM-Solver++(2M)
$(N = 20)$

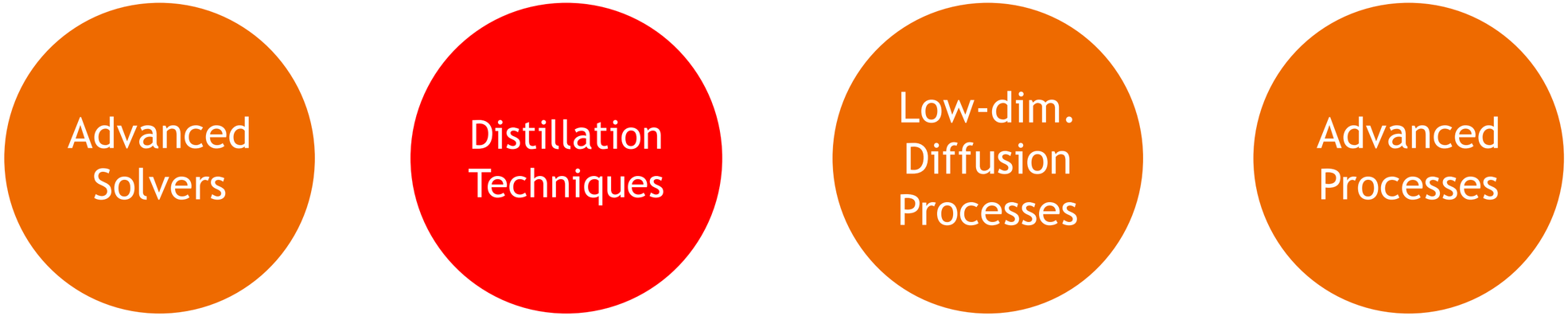
DPM-Solver++(2M)
$(N = 50)$

Image from Lu et al. 2023

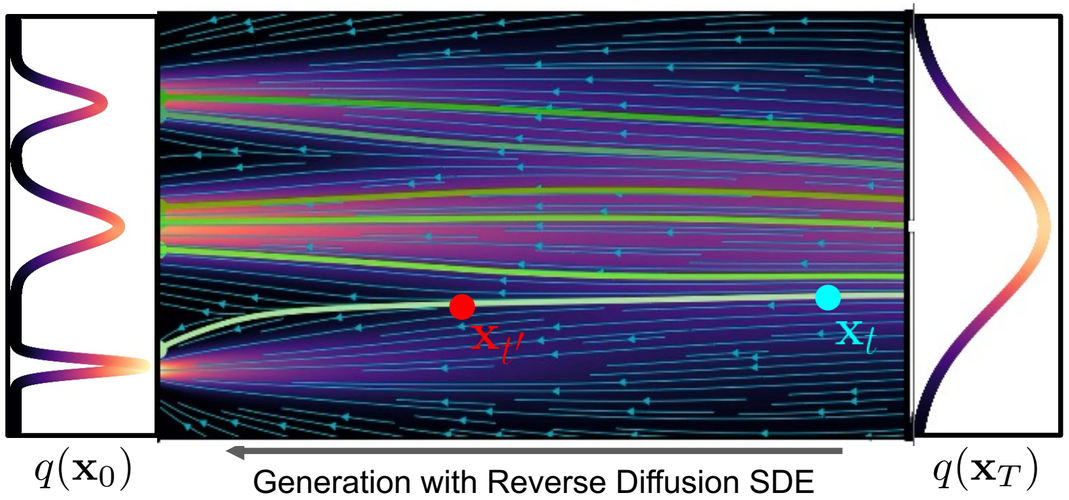
# A Rich Body of Work on ODE/SDE Solvers for Diffusion Models

- Runge-Kutta adaptive step-size ODE solver:
  - Song et al., “Score-Based Generative Modeling through Stochastic Differential Equations”, *ICLR*, 2021
- Higher-Order adaptive step-size SDE solver:
  - Jolicoeur-Martineau et al., “Gotta Go Fast When Generating Data with Score-Based Models”, *arXiv*, 2021
- Reparametrized, smoother ODE:
  - Song et al., “Denoising Diffusion Implicit Models”, *ICLR*, 2021
  - Zhang et al., "gDDIM: Generalized denoising diffusion implicit models", arXiv 2022
- Higher-Order ODE solver with linear multistepping:
  - Liu et al., “Pseudo Numerical Methods for Diffusion Models on Manifolds”, *ICLR*, 2022
- Exponential ODE Integrators:
  - Zhang and Chen, “Fast Sampling of Diffusion Models with Exponential Integrator”, *arXiv*, 2022
  - Lu et al., “DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps”, *NeurIPS*, 2022
  - Lu et al., "DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models", NeurIPS 2022
- Higher-Order ODE solver with Heun’s Method:
  - Karras et al., “Elucidating the Design Space of Diffusion-Based Generative Models”, *NeurIPS*, 2022
- Many more:
  - Zhao et al., "UniPC: A Unified Predictor-Corrector Framework for Fast Sampling of Diffusion Models", arXiv 2023
  - Shih et al., "Parallel Sampling of Diffusion Models", arxiv 2023
  - Chen et al., "A Geometric Perspective on Diffusion Models", arXiv 2023

# Acceleration Techniques



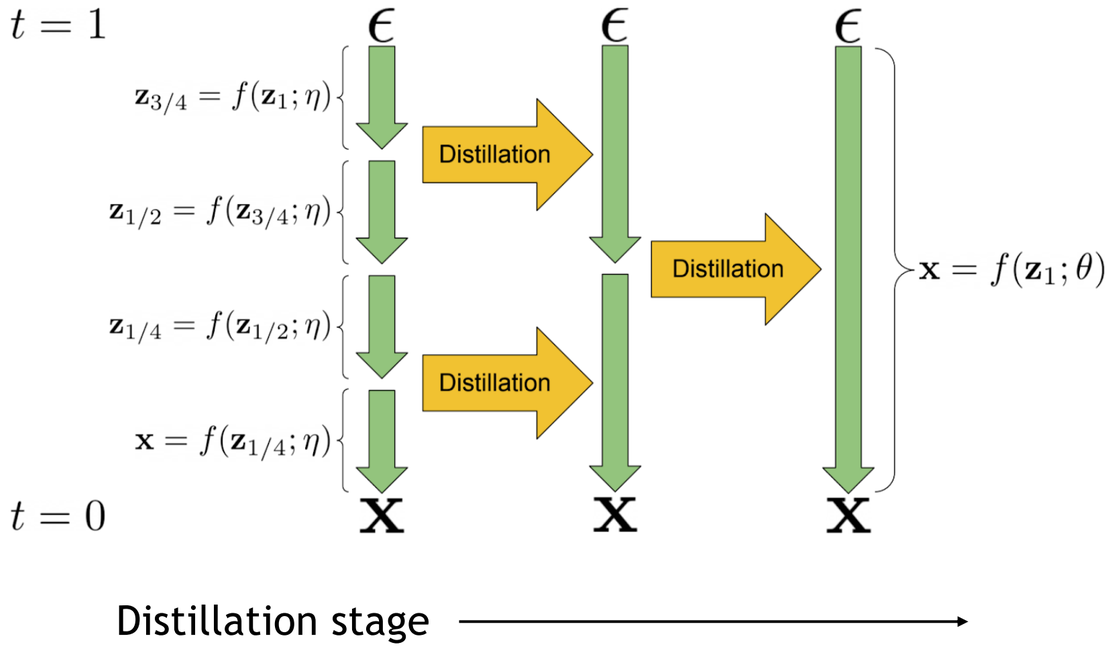
# ODE Distillation




Can we train a neural network to directly predict $\mathbf{x}_{t'}$ given $\mathbf{x}_t$ ?

# Progressive Distillation

- Distill a deterministic ODE sampler to the same model architecture.
- At each stage, a “student” model is learned to distill two adjacent sampling steps of the “teacher” model to one sampling step.
- At next stage, the “student” model from previous stage will serve as the new “teacher” model.



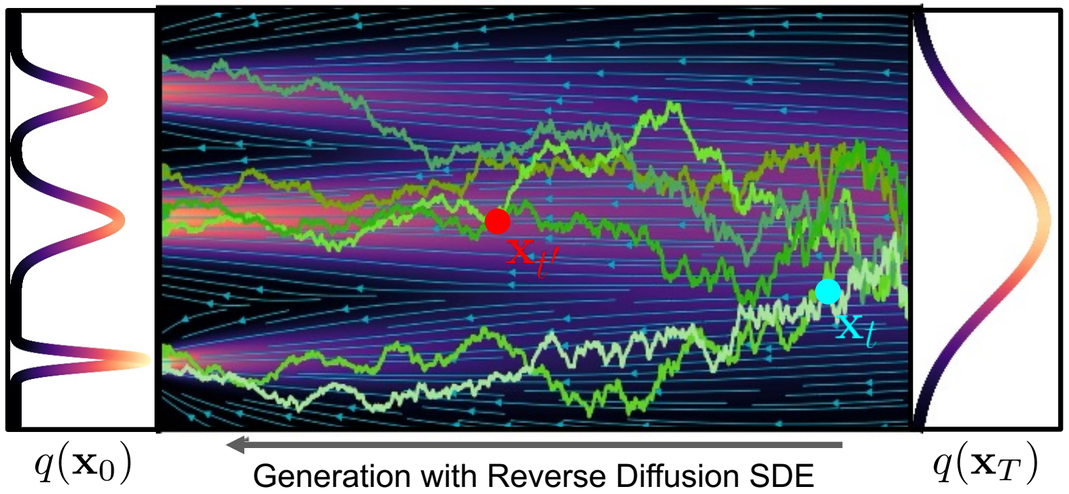

Salimans & Ho, “Progressive distillation for fast sampling of diffusion models”, ICLR 2022.

# Progressive Distillation in Latent Space

(a) 2 denoising steps (b) 4 denoising steps (c) 8 denoising steps

Meng et al., "On Distillation of Guided Diffusion Models", CVPR 2023 (Award Candidate)

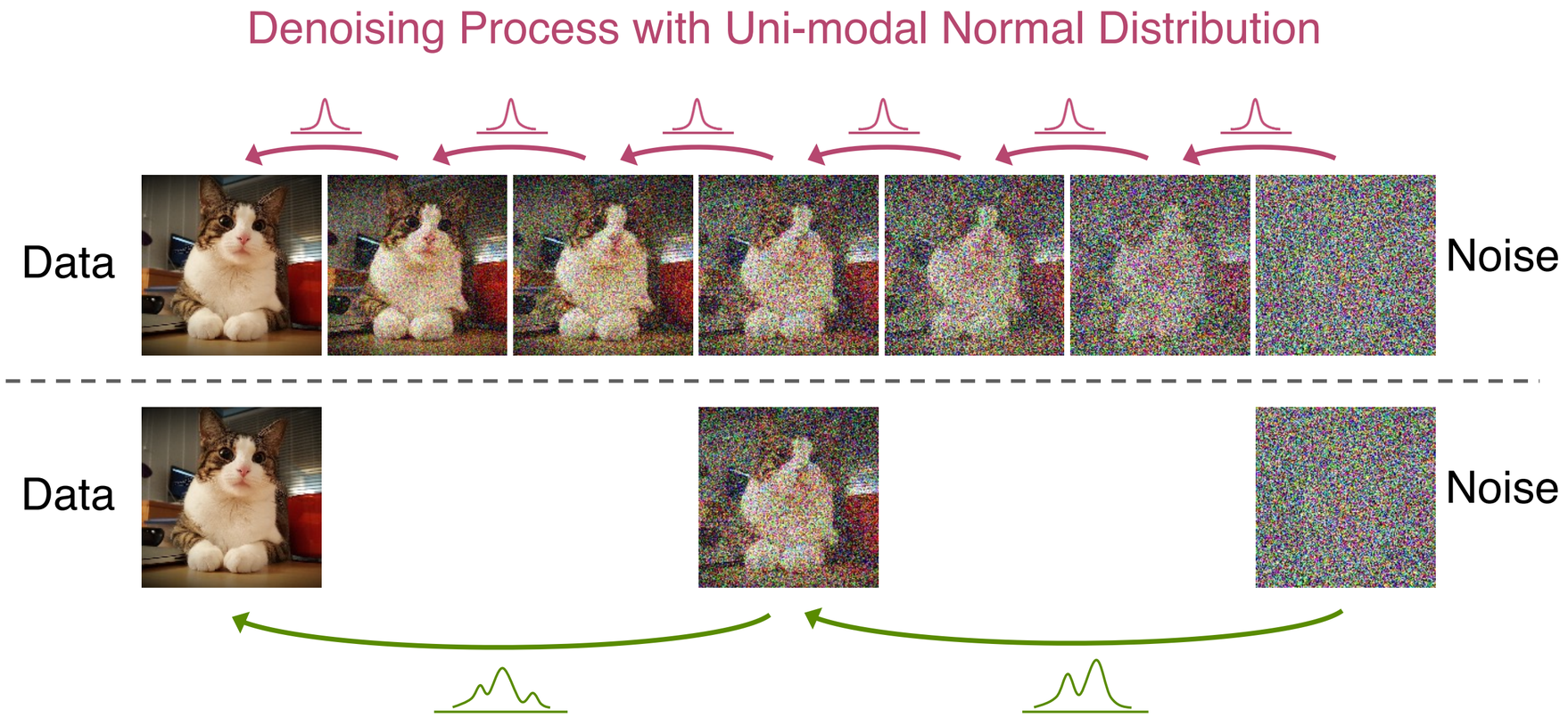
# SDE Distillation




Can we train a neural network to directly predict distribution of $\mathbf{x}_{t'}$ given $\mathbf{x}_t$ ?

# Advanced Approximation of Reverse Process

## Normal assumption in denoising distribution holds only for small step

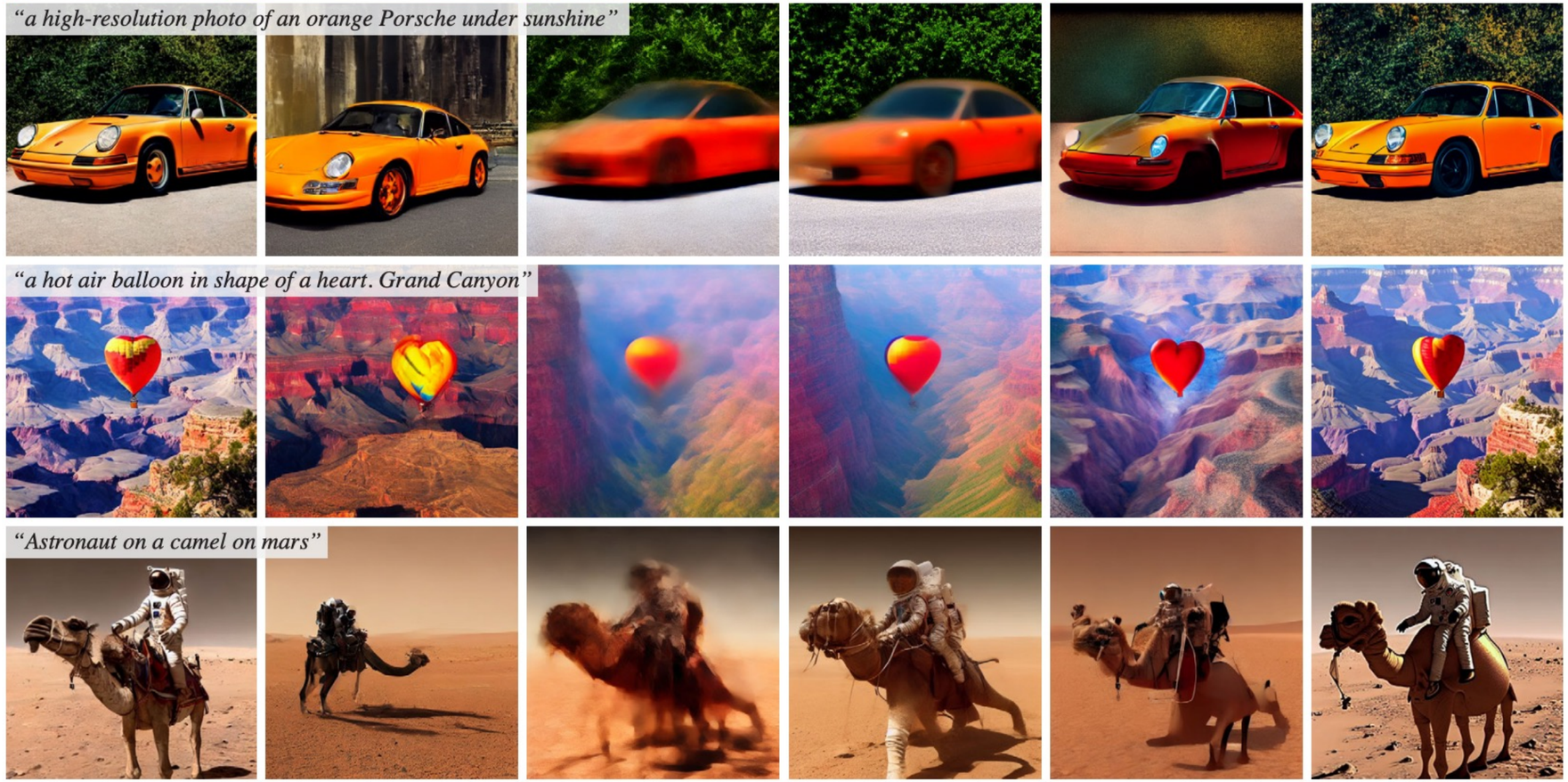



Requires more complicated functional approximators!

Xiao et al., “Tackling the Generative Learning Trilemma with Denoising Diffusion GANs”, ICLR 2022.
Gao et al., “Learning energy-based models by diffusion recovery likelihood”, ICLR 2021.

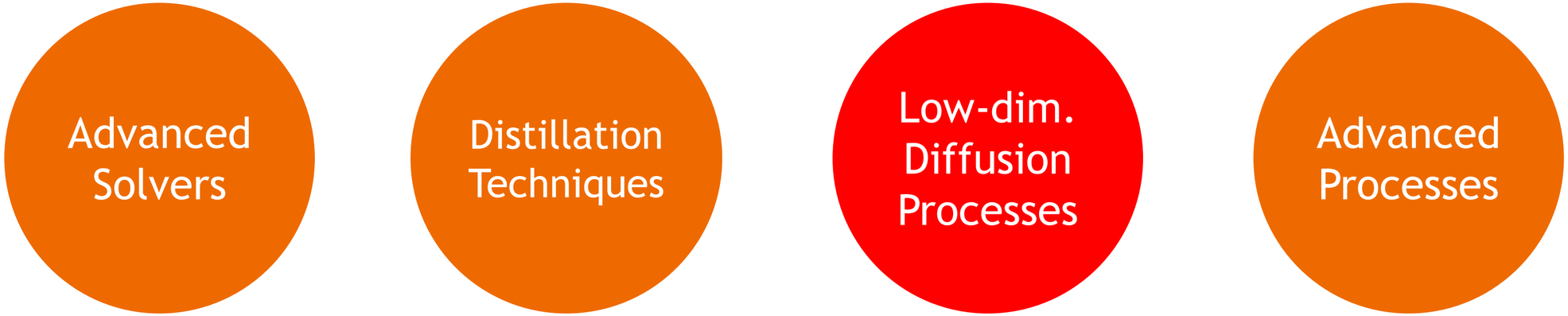
# Distribution Matching Distillation

## One-step Sample Generation



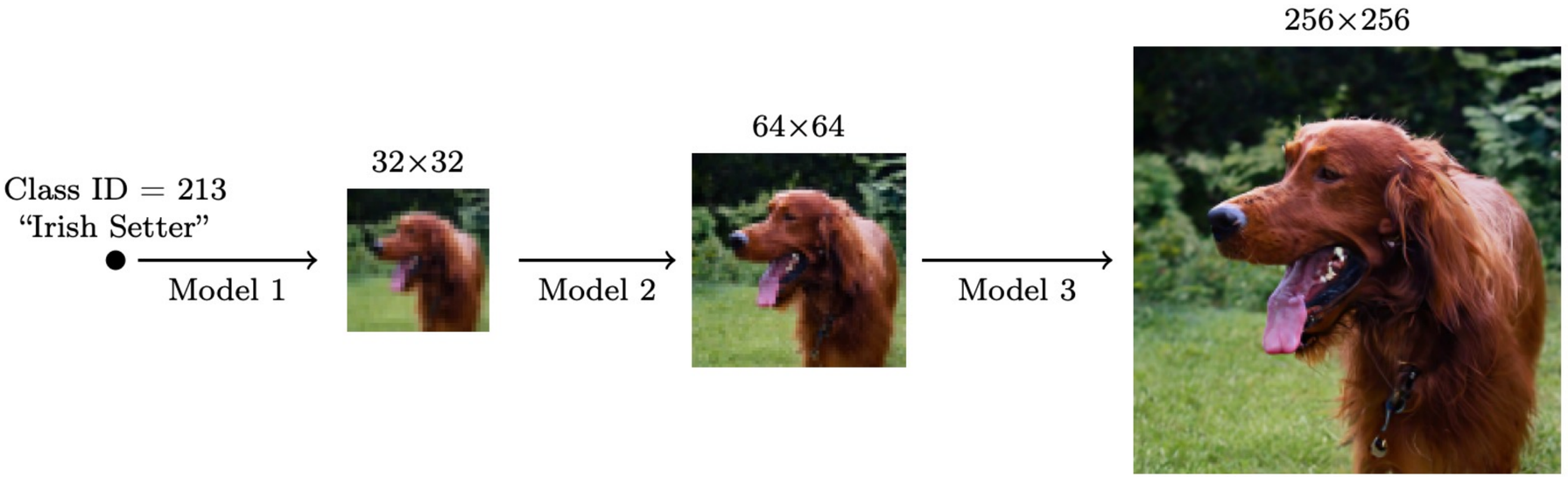

Yin et al., "One-step Diffusion with Distribution Matching Distillation", arXiv 2023

# Acceleration Techniques



# Cascaded Generation

## Pipeline

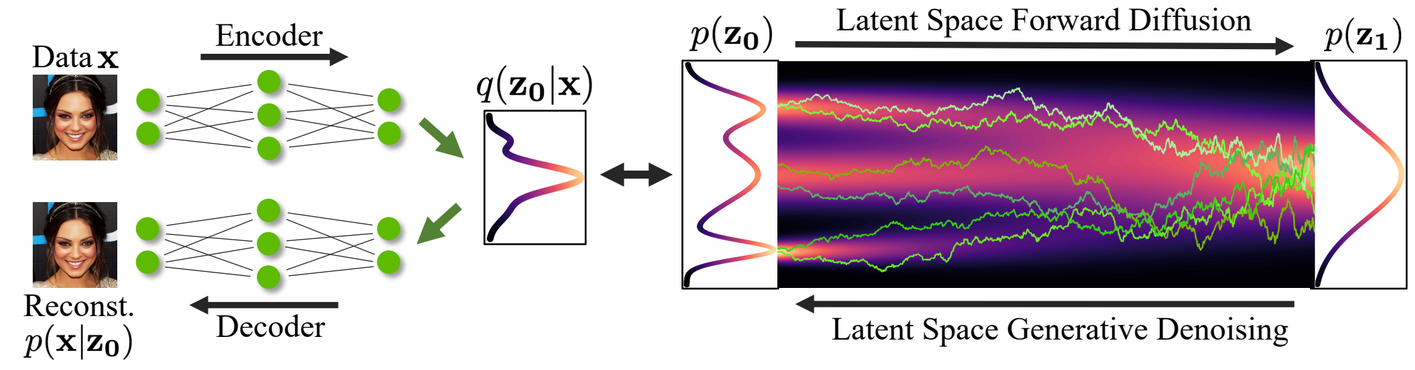



A popular approach for both image and video generation!


Ho et al., "Cascaded Diffusion Models for High Fidelity Image Generation", 2021.
Ramesh et al., "Hierarchical Text-Conditional Image Generation with CLIP Latents", arXiv 2022.
Saharia et al., "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding", arXiv 2022.
Ho et al., Imagen Video: High Definition Video Generation with Diffusion Models

# Latent Diffusion Models

## Variational autoencoder + score-based prior




**Main Idea**

Encoder maps the input data to an embedding space

Denoising diffusion models are applied in the latent space

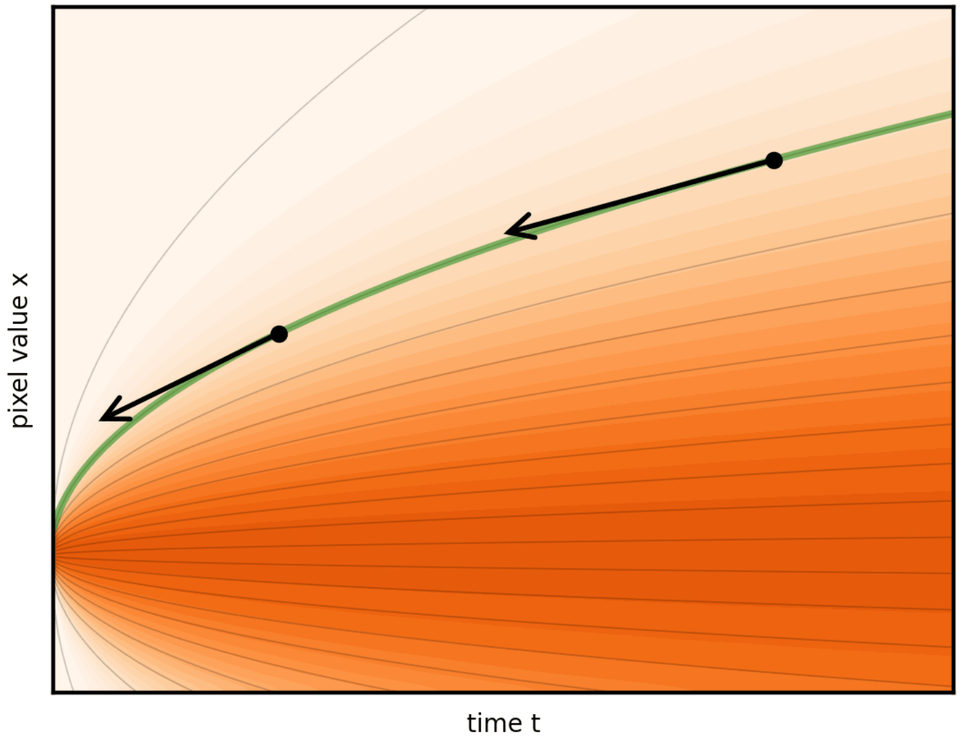

Vahdat et al., “Score-based Generative Modeling in Latent Space”, *NeurIPS*, 2021
Rombach et al., “High-Resolution Image Synthesis with Latent Diffusion Models”, *CVPR*, 2022
Sinha et al., “D2C: Diffusion-Denoising Models for Few-shot Conditional Generation”, *NeurIPS*, 2021
Mittal et al., “Symbolic Music Generation with Diffusion Models”, *ISMIR*, 2021

# Acceleration Techniques

# ODE interpretation

## Deterministic generative process

Different time and data scaling result in different ODE curvatures.




Song et al., “Denoising Diffusion Implicit Models”, ICLR 2021.
Karras et al., “Elucidating the Design Space of Diffusion-Based Generative Models”, arXiv 2022.

# Agenda

- An Introduction to Diffusion Models
- Acceleration
- **Conditioning & Guidance**
- Personalization
- Latent Diffusion Models
- Video Diffusion Models
- 3D and 4D Generation

# Impressive Conditional Diffusion Models

## Text-to-image generation

DALL·E 2

*“a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese”*

IMAGEN

*“A photo of a raccoon wearing an astronaut helmet, looking out of the window at night.”*

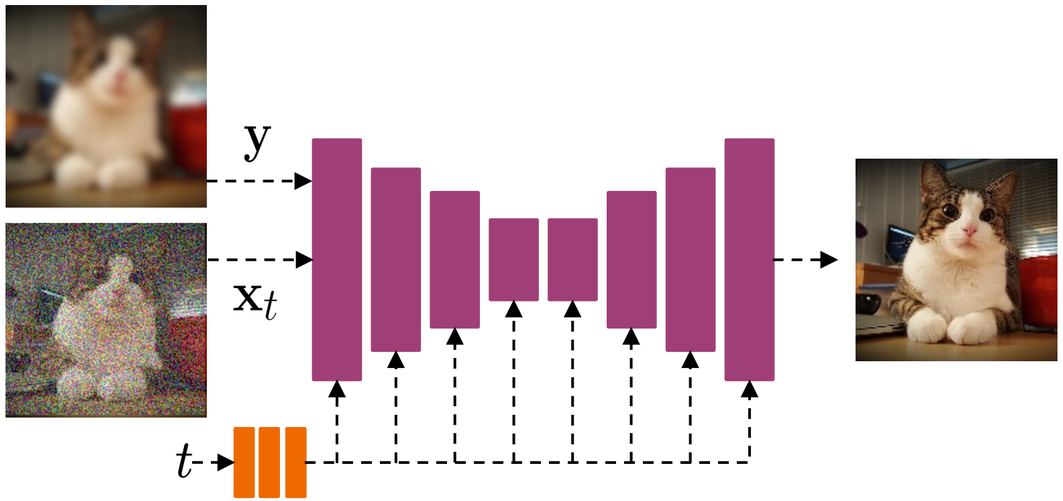

Ramesh et al., “Hierarchical Text-Conditional Image Generation with CLIP Latents”, arXiv 2022.
Saharia et al., “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding”, arXiv 2022.

# Explicit Conditional Training

Conditional training can be considered as training $p(\mathbf{x}|\mathbf{y})$:

- $\mathbf{y}$ is the input conditioning (e.g., blurry image)
- $\mathbf{x}$ is generated output (e.g., sharp image)

The conditional score is simply a U-Net with $\mathbf{x_t}$ and $\mathbf{y}$ together in the input.



# Attention Layers for Text Conditioning

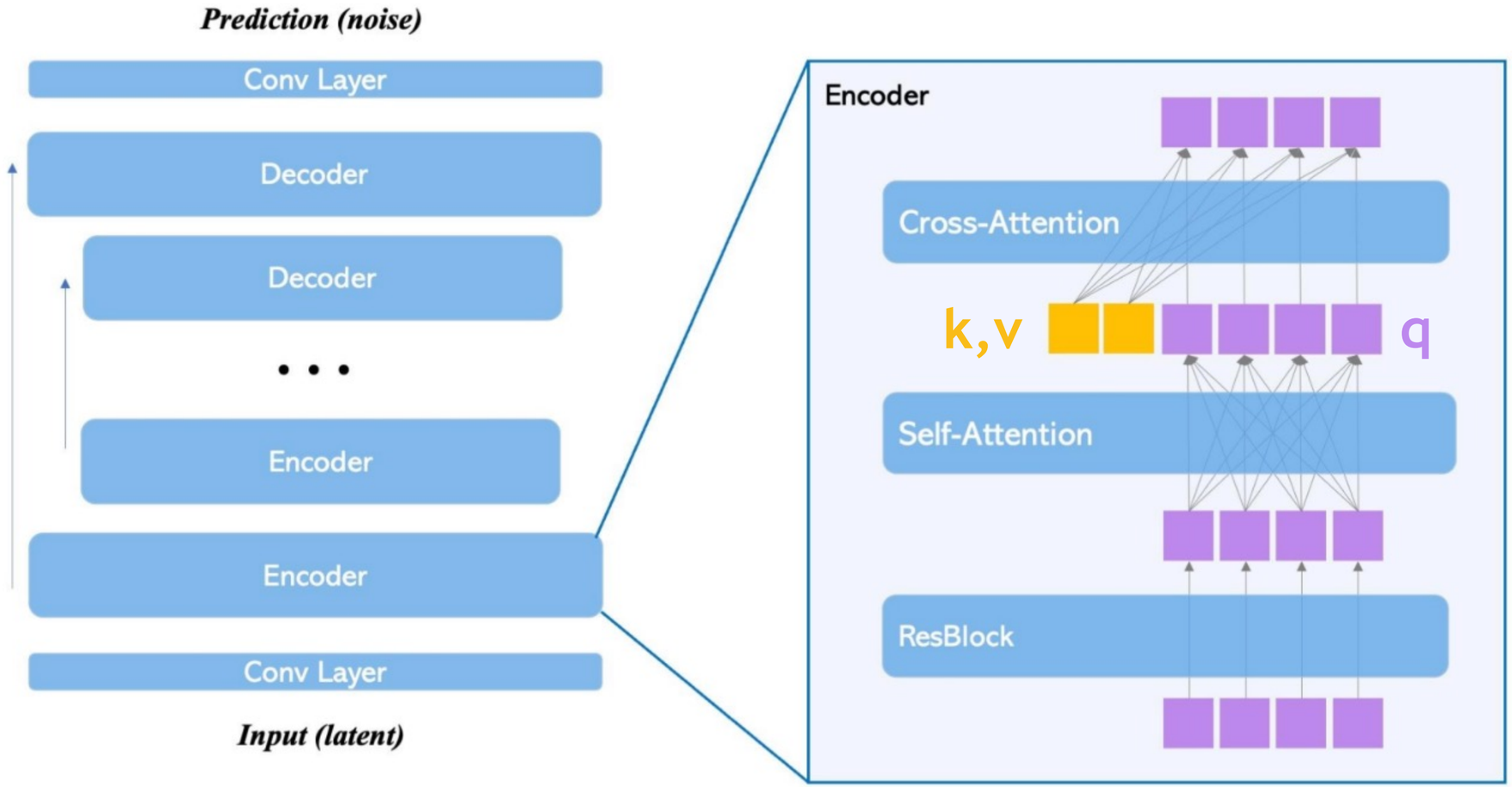



Image from: GLIGEN, Li et al., CVPR 2023

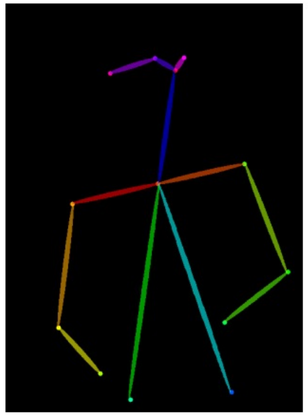
# ControlNet

Input Canny edge

Default

“masterpiece of fairy tale, giant deer, golden antlers”

“..., quaint city Galic”



Input human pose

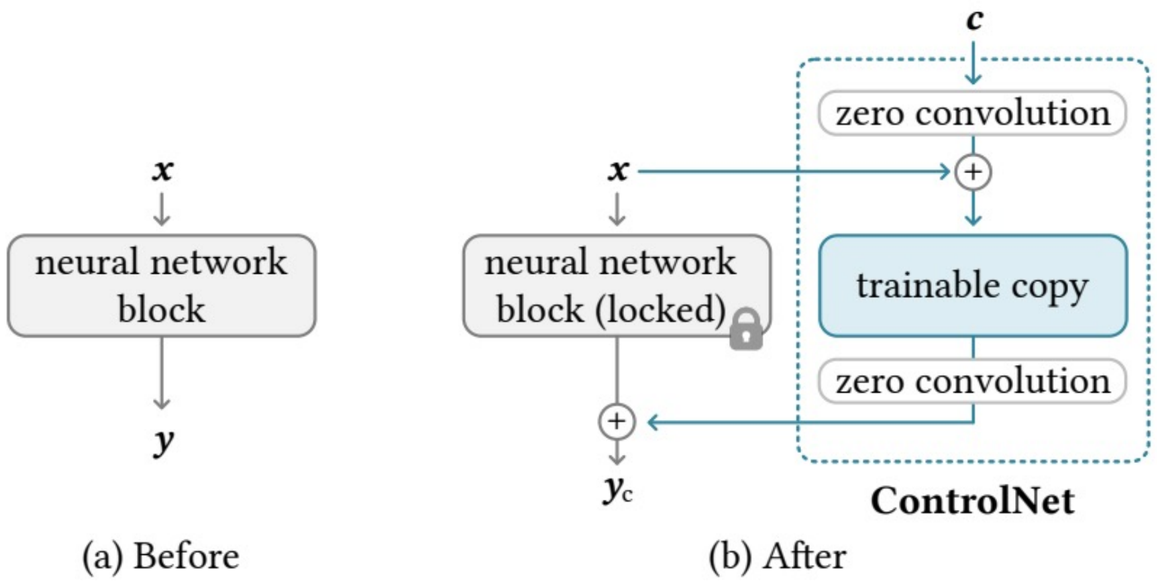
Default

“chef in kitchen”

“Lincoln statue”

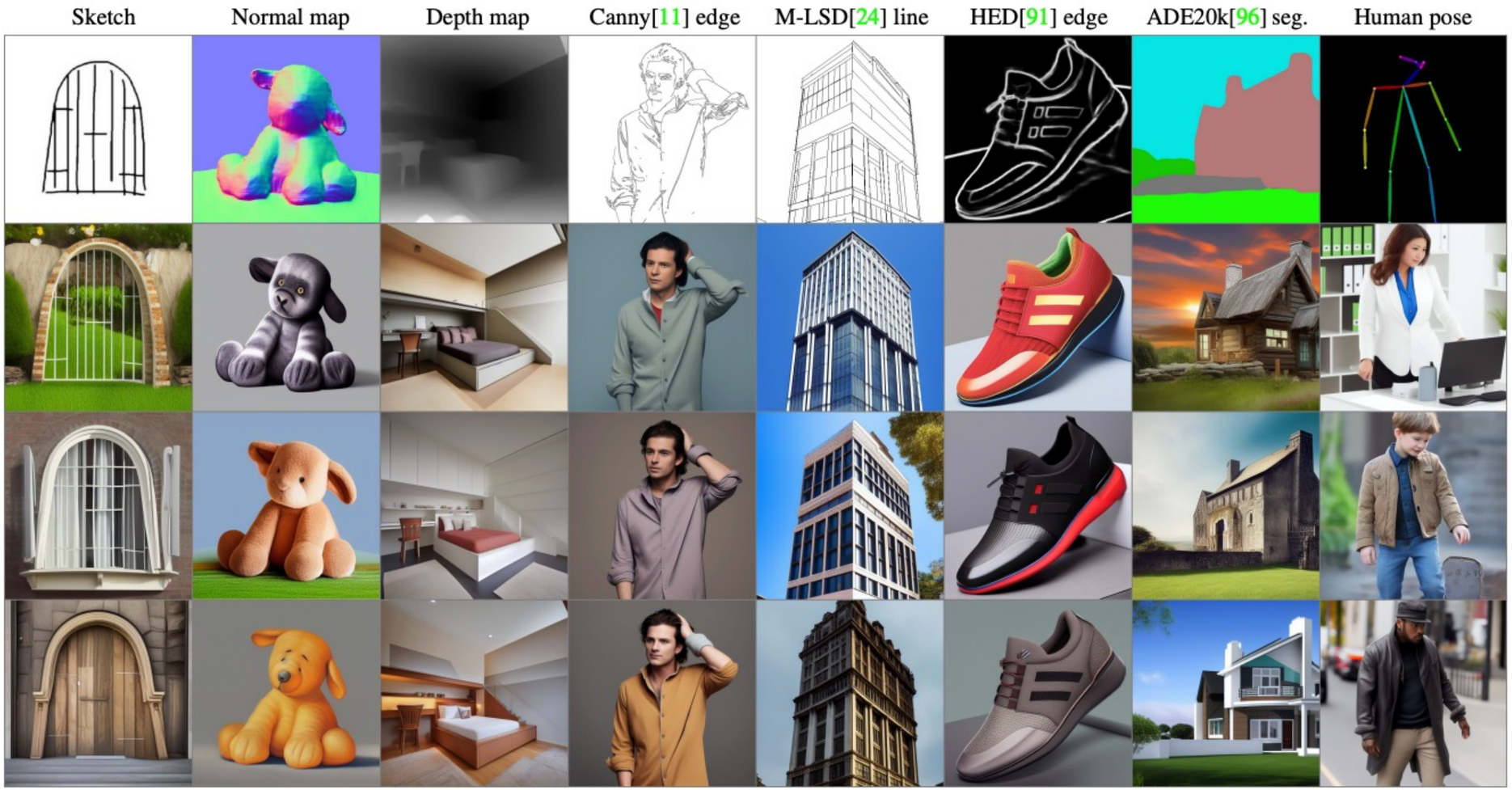
Zhang et al., ICCV 2023

# ControlNet




Zhang et al., ICCV 2023

# ControlNet

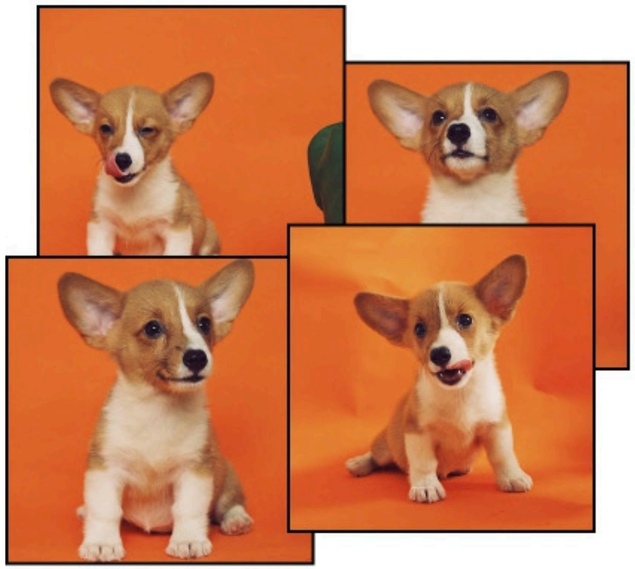



Zhang et al., ICCV 2023

# Agenda

- An Introduction to Diffusion Models
- Acceleration
- Conditioning & Guidance
- **Personalization**
- Latent Diffusion Models
- Video Diffusion Models
- 3D and 4D Generation

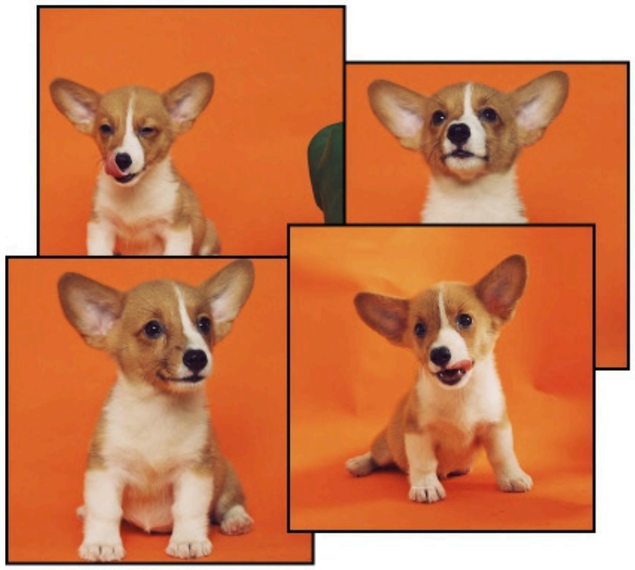
# Diffusion Model Personalization



(Real) images of your dog

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

# Diffusion Model Personalization



(Real) images of your dog

in the Acropolis

swimming

in a doghouse

sleeping

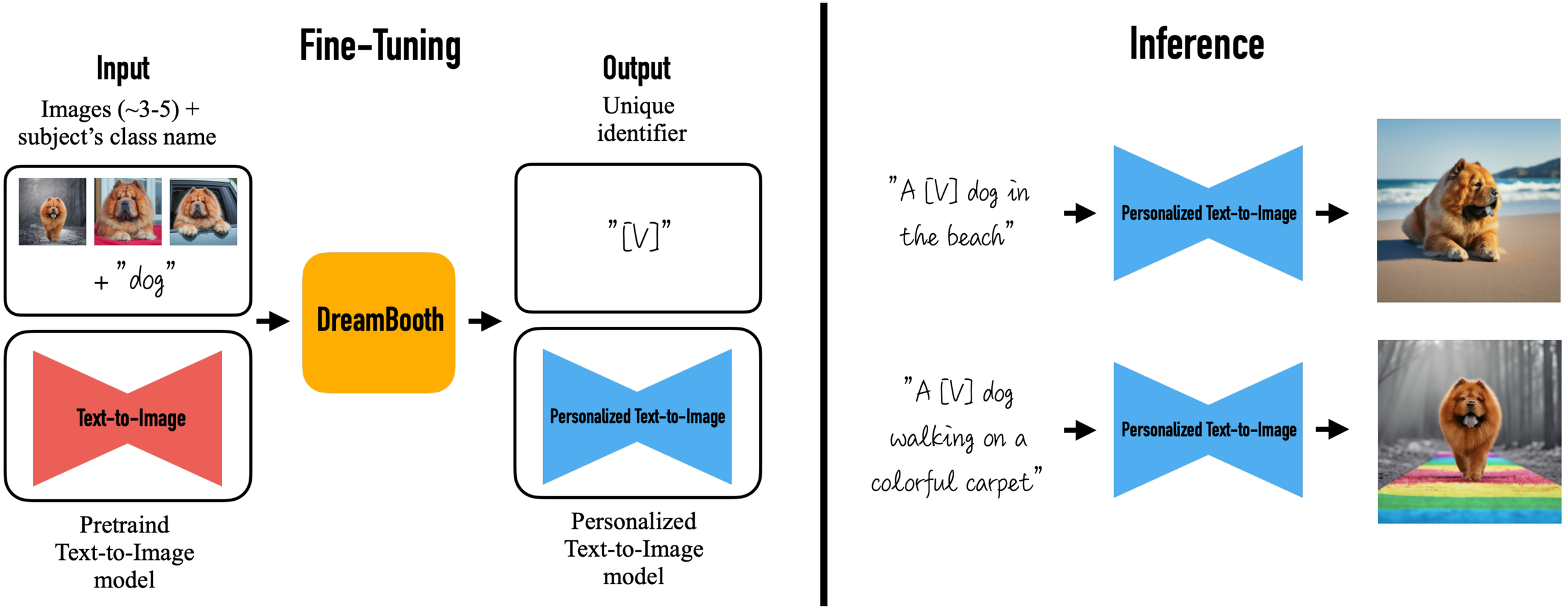
in a bucket

getting a haircut

Generated images by *personalized* diffusion model

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

# Diffusion Model Personalization – “DreamBooth”




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

# Diffusion Model Personalization – “DreamBooth”


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

# Diffusion Model Personalization – “DreamBooth”


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

# Diffusion Model Personalization – “DreamBooth”


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

# Diffusion Model Personalization – “DreamBooth”


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

# Diffusion Model Personalization – “LoRA”

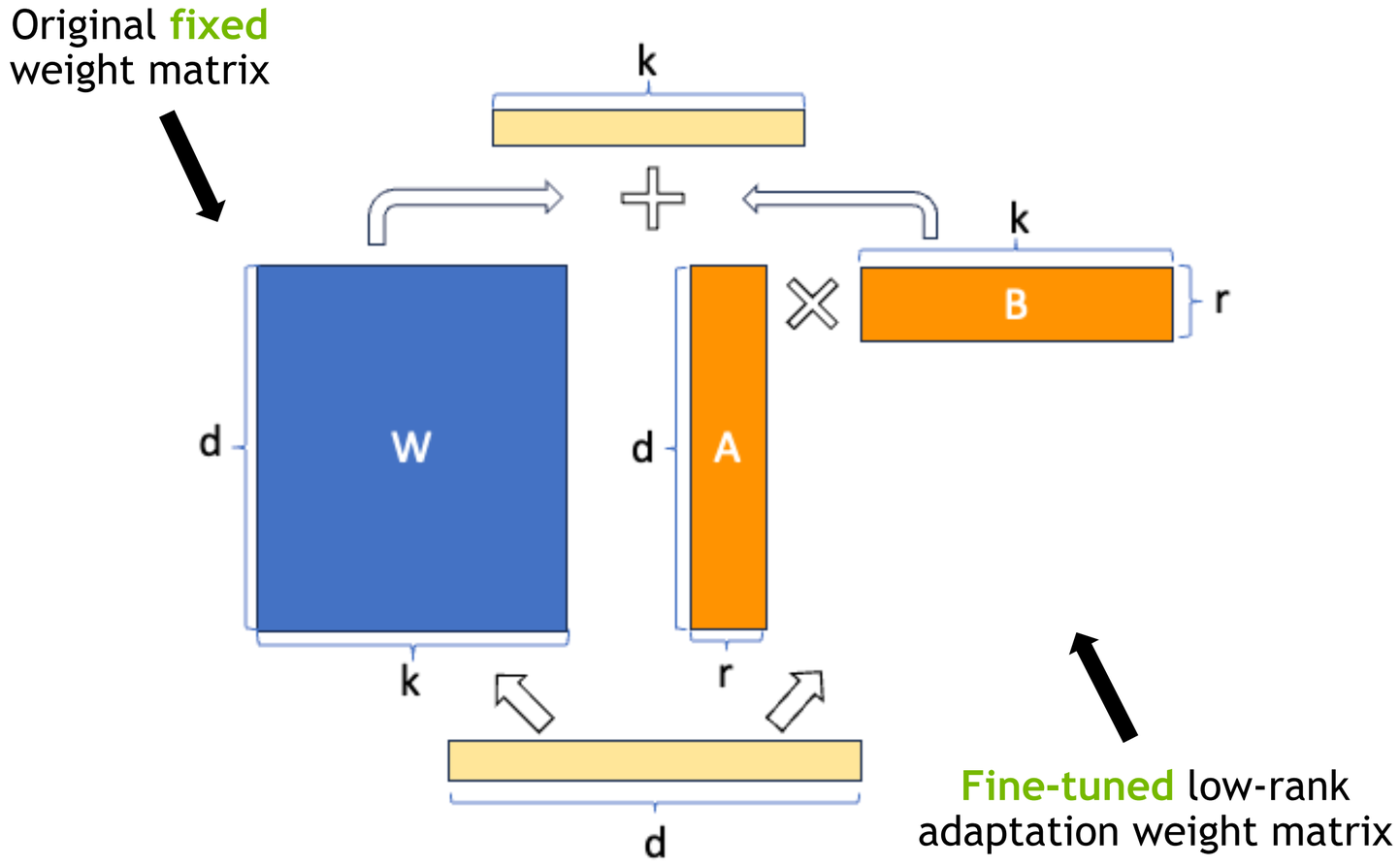
*Issue:*

DreamBooth needs to fine-tune entire large diffusion model!

*Low-Rank Adaption (LoRA):*

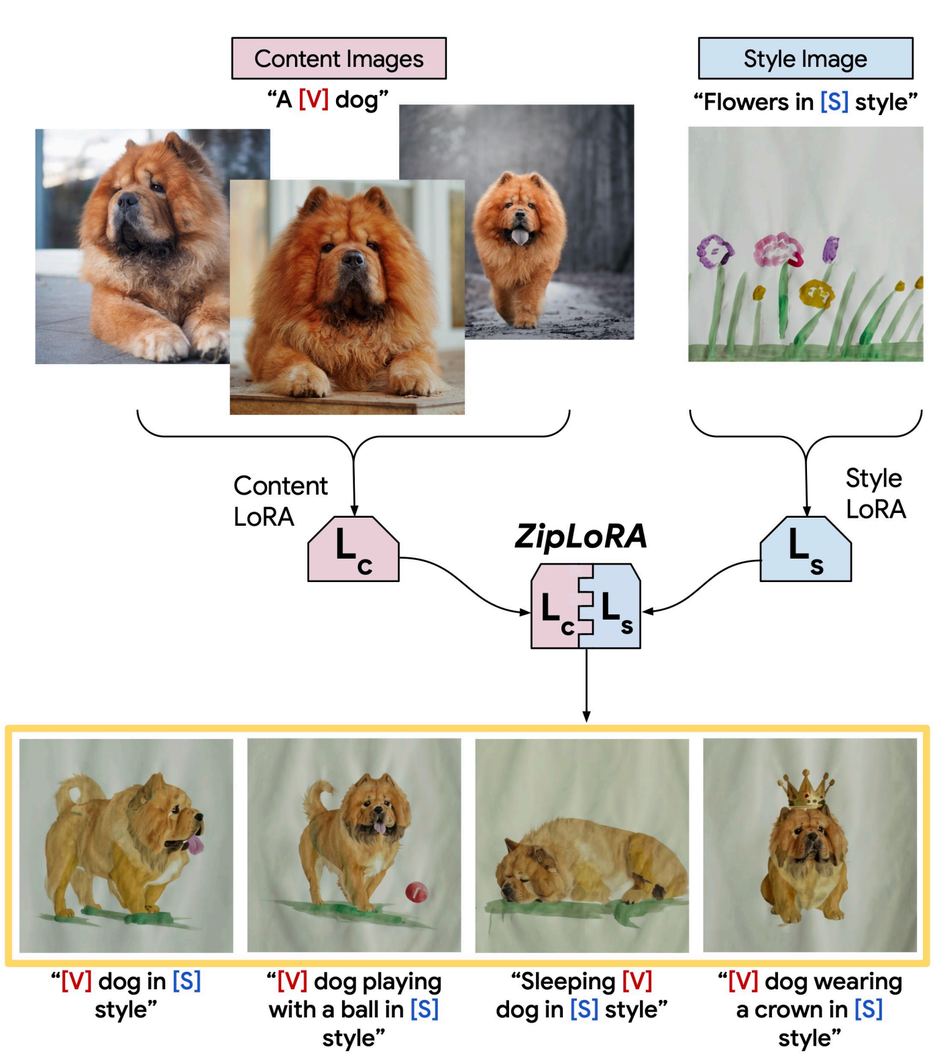
Train only **low-rank adaptation weight matrices**.

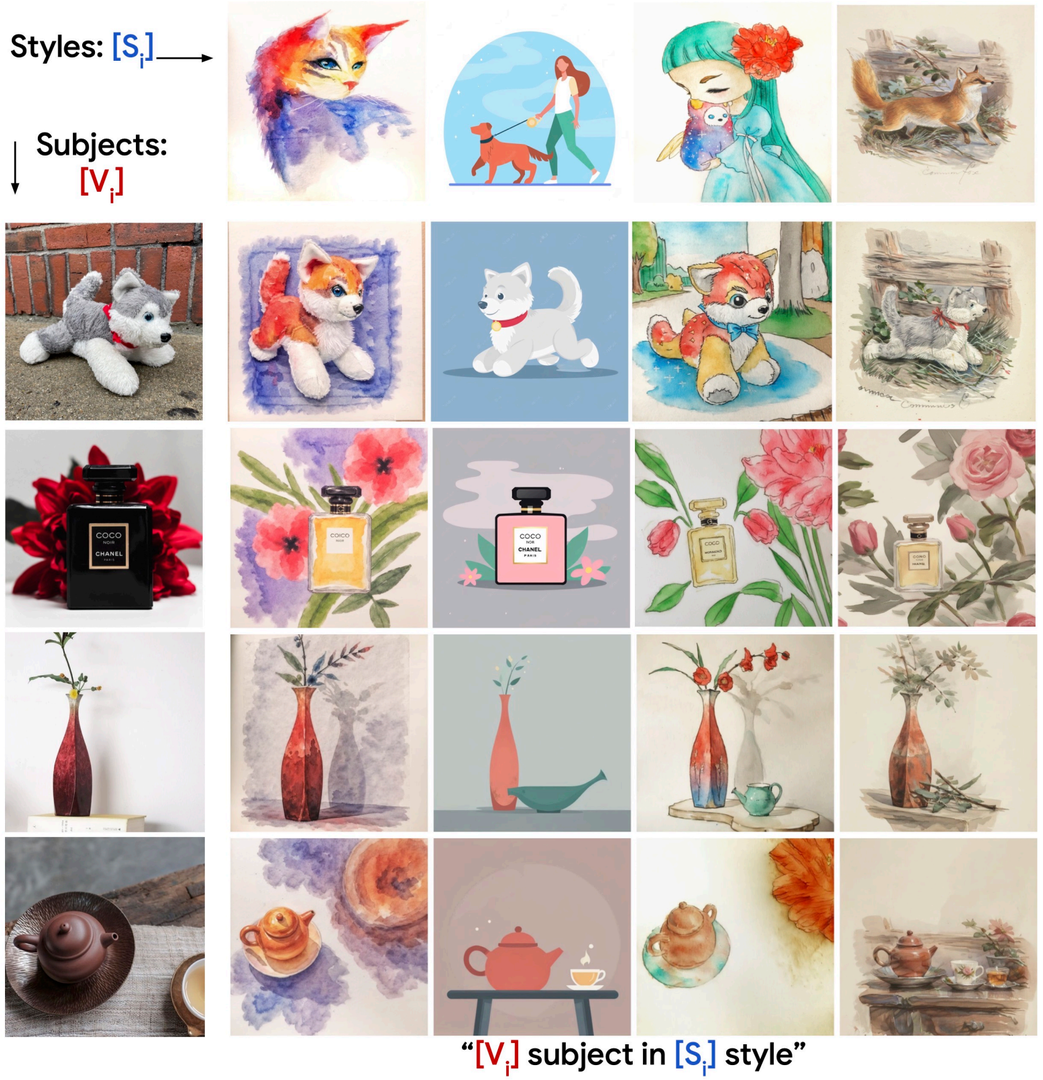
Low rank: only few fine-tuning parameters! Efficient! Low storage!




Hu et al, “LoRA: Low-Rank Adaptation of Large Language Models”, *ICLR*, 2022

# Diffusion Model Personalization – “ZipLoRA”




Shah et al, “ZipLoRA: Any Subject in Any Style by Effectively Merging LoRAs”, *arXiv*, 2023

# Diffusion Model Personalization – with Encoders


Predict model updates with encoder + fast fine-tuning:

→ Personalization within seconds (instead of minutes!)

Gal et al, “Encoder-based Domain Tuning for Fast Personalization of Text-to-Image Models”, *SIGGRAPH*, 2023

# Agenda

- An Introduction to Diffusion Models
- Acceleration
- Conditioning & Guidance
- Personalization
- **Latent Diffusion Models**
- Video Diffusion Models
- 3D and 4D Generation

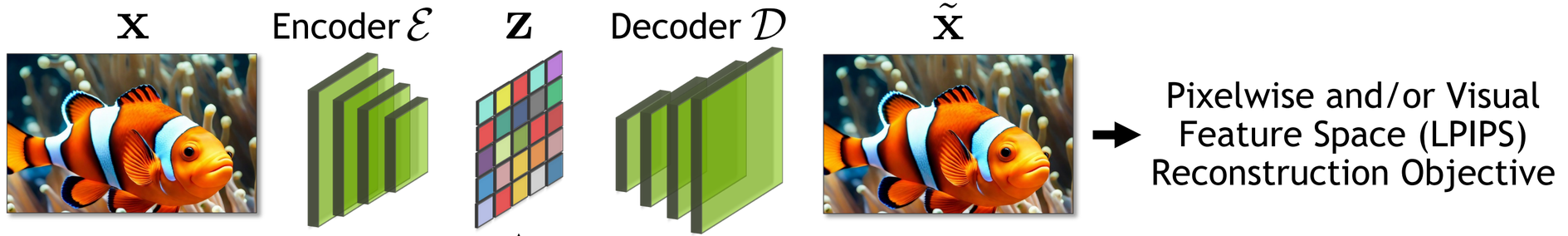
# Latent Diffusion Models

## Map Data into Compressed Latent Space. Train Diffusion Model efficiently in Latent Space.

- Stage 1:

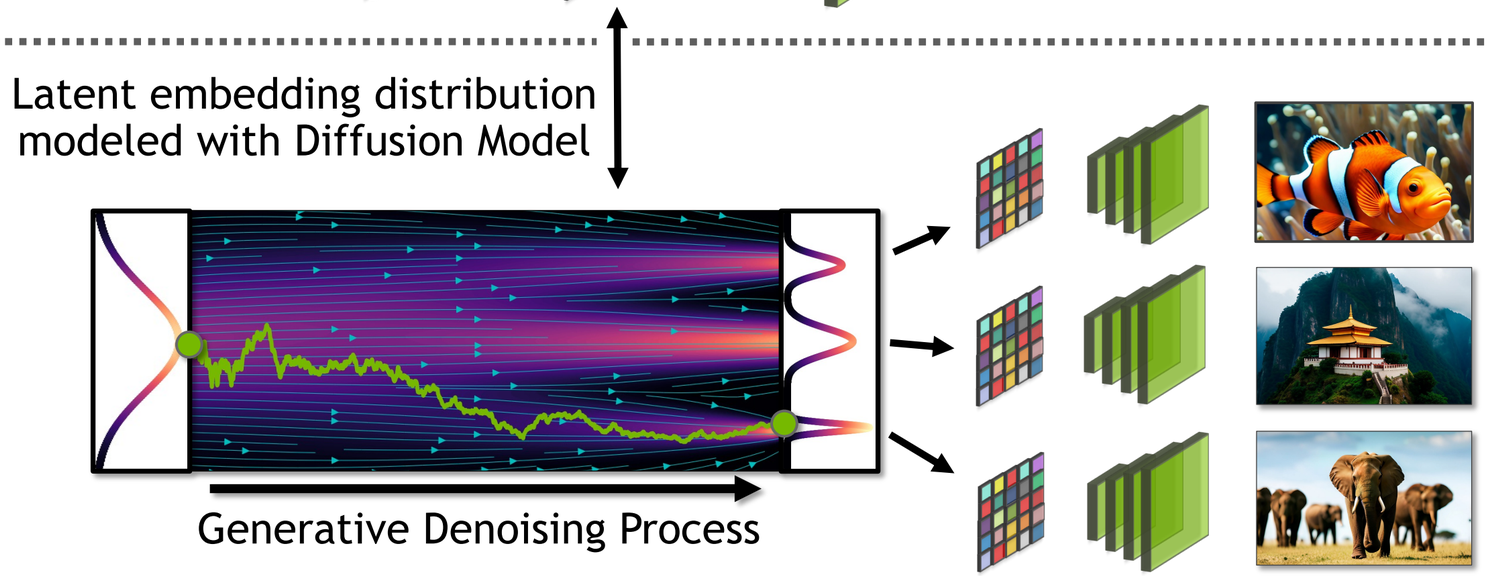
Train Autoencoder

$$\tilde{\mathbf{x}} = \mathcal{D}(\mathcal{E}(\mathbf{x}))$$




- Stage 2:

Train **Latent** Diffusion Model

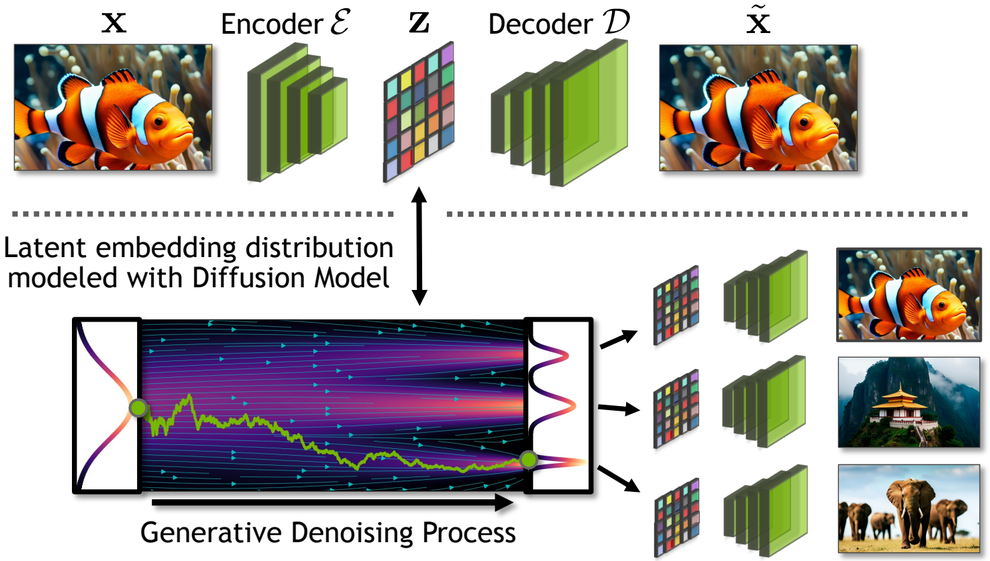




Vahdat et al., "Score-based Generative Modeling in Latent Space", *NeurIPS*, 2021
Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models", *CVPR*, 2022
Sinha et al., "D2C: Diffusion-Denoising Models for Few-shot Conditional Generation", *NeurIPS*, 2021
Mittal et al., "Symbolic Music Generation with Diffusion Models", *ISMIR*, 2021

# Latent Diffusion Models

## Map Data into Compressed Latent Space. Train Diffusion Model efficiently in Latent Space.




**Advantages:**

1. *Compressed latent space:* Train diffusion model in **lower resolution** latent space ➡ **computationally more efficiently**
2. *Regularized smooth/compressed latent space:* **Easier task** for diffusion model and **faster sampling**
3. *Flexibility*: **Autoencoder can be tailored to data** (images, video, text, graphs, 3D point clouds, meshes, etc.)


Vahdat et al., "Score-based Generative Modeling in Latent Space", *NeurIPS*, 2021
Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models", *CVPR*, 2022
Sinha et al., "D2C: Diffusion-Denoising Models for Few-shot Conditional Generation", *NeurIPS*, 2021
Mittal et al., "Symbolic Music Generation with Diffusion Models", *ISMIR*, 2021

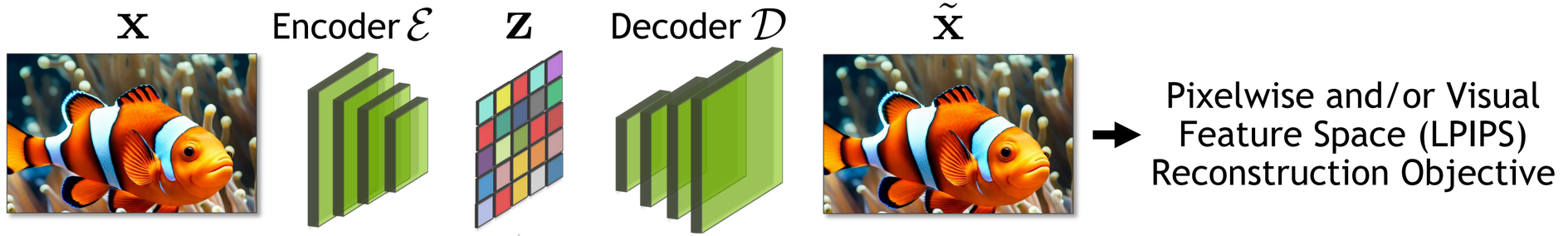
# Latent Diffusion Models

## Add Adversarial Patch-based Discriminator on top of Reconstruction Loss for Perceptual Compression

- Stage 1:

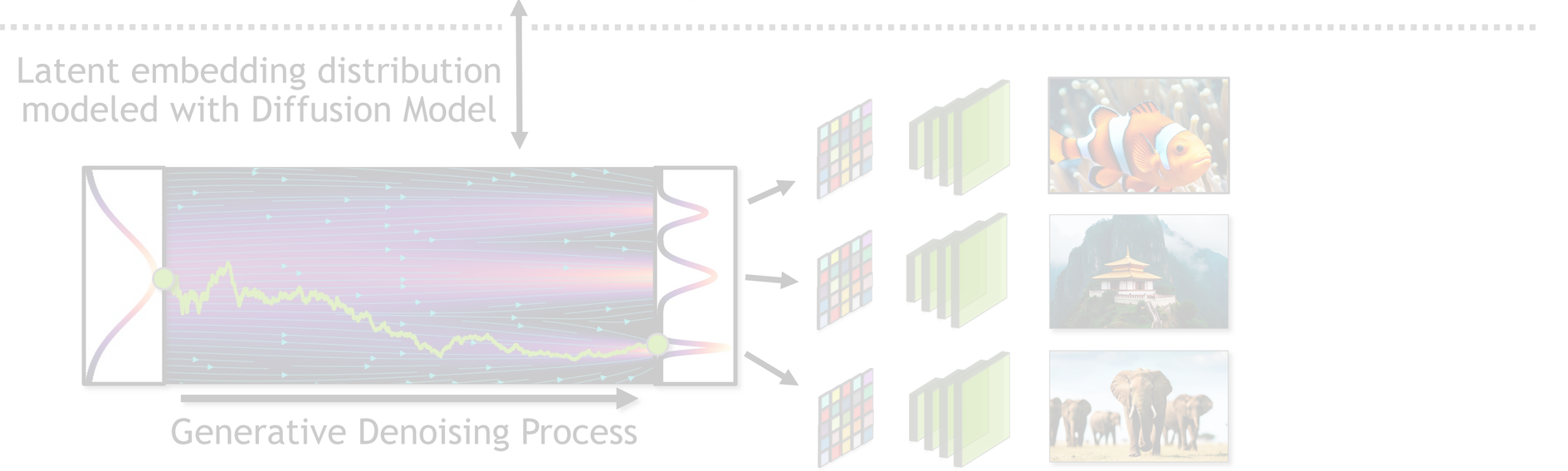
Train Autoencoder

$$\tilde{\mathbf{x}} = \mathcal{D}(\mathcal{E}(\mathbf{x}))$$




- Stage 2:

Train Latent Diffusion Model




Rombach et al., “High-Resolution Image Synthesis with Latent Diffusion Models”, *CVPR*, 2022

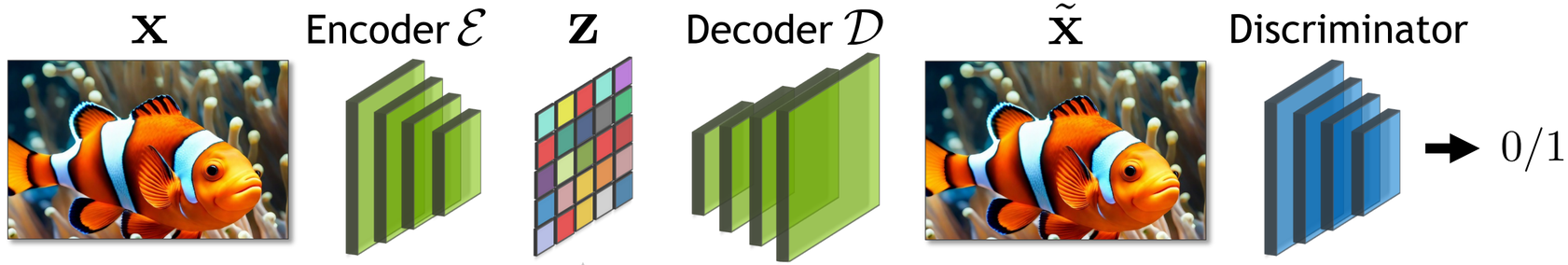
# Latent Diffusion Models

## Add Adversarial Patch-based Discriminator on top of Reconstruction Loss for Perceptual Compression

- Stage 1:

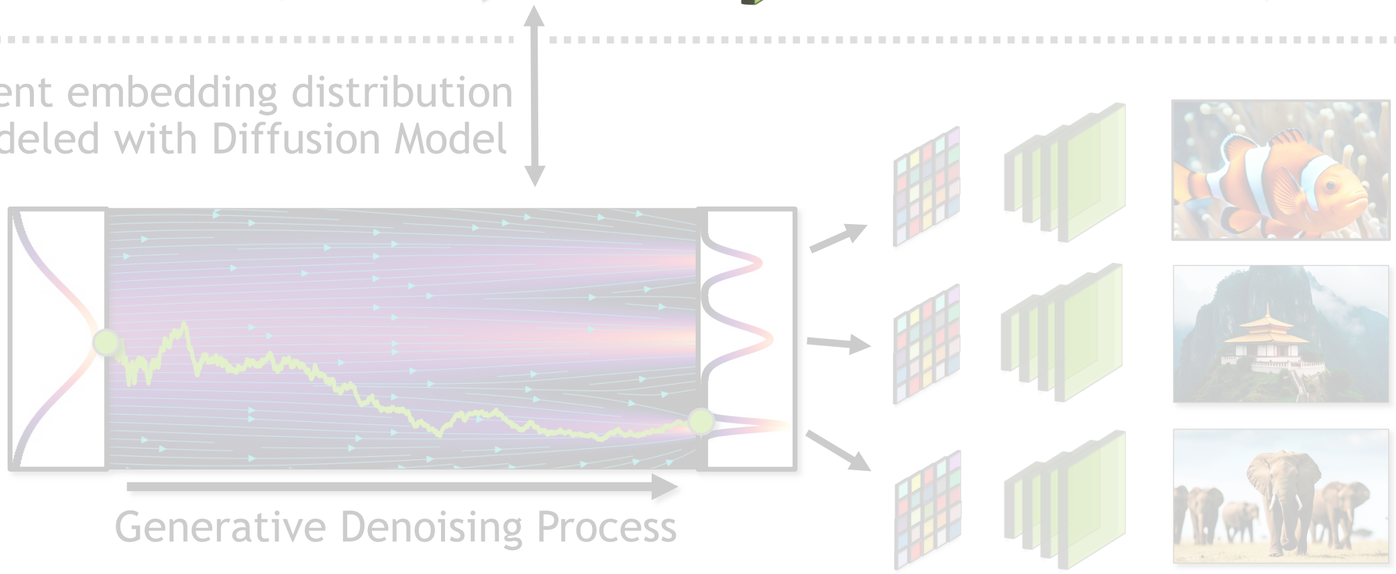
Train Autoencoder

$$\tilde{\mathbf{x}} = \mathcal{D}(\mathcal{E}(\mathbf{x}))$$




- Stage 2:

Train Latent Diffusion Model




Rombach et al., “High-Resolution Image Synthesis with Latent Diffusion Models”, *CVPR*, 2022

Input

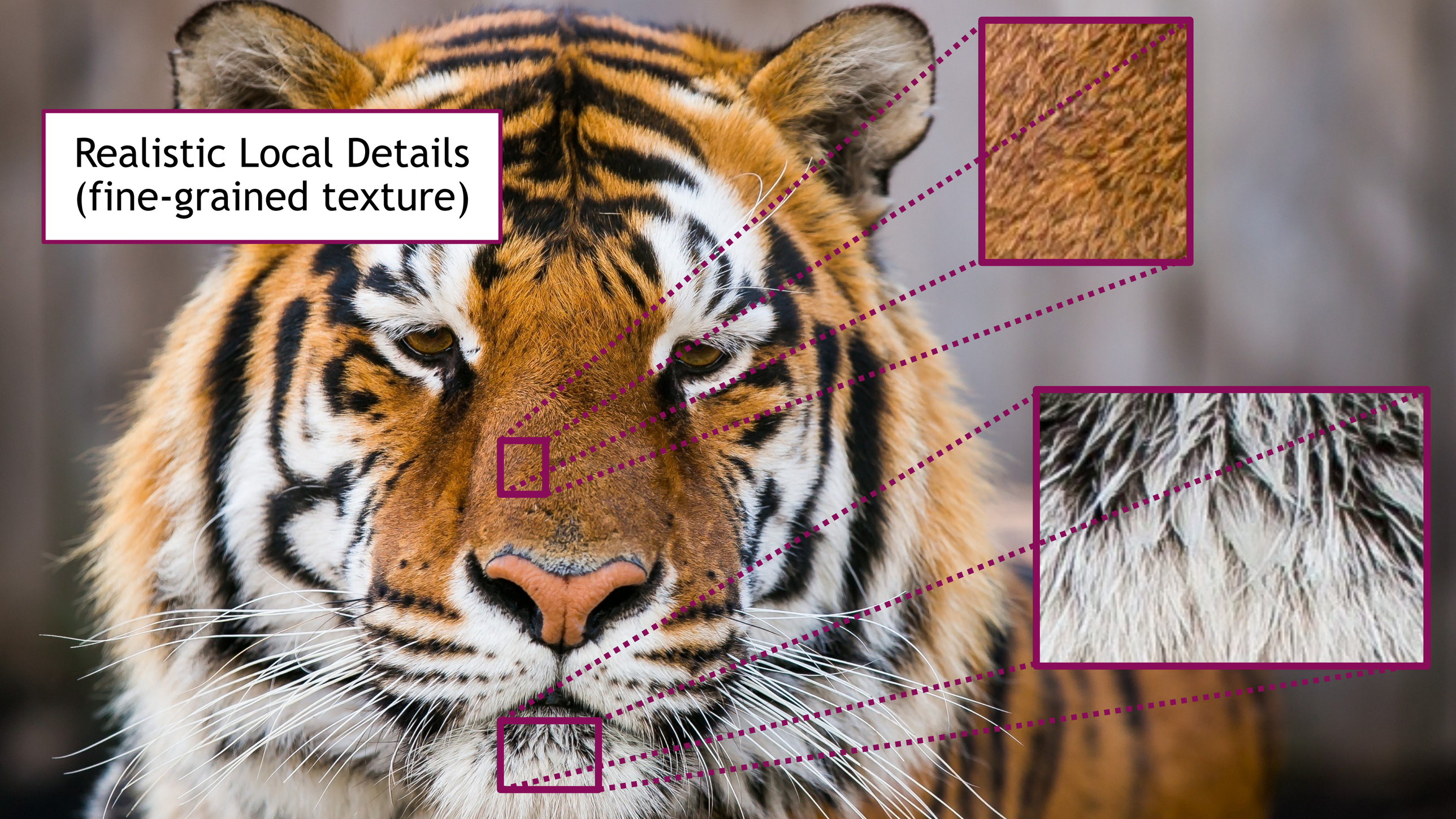
Reconstruction without Discriminator

Reconstruction with Discriminator

Rombach et al., “High-Resolution Image Synthesis with Latent Diffusion Models”, *CVPR*, 2022

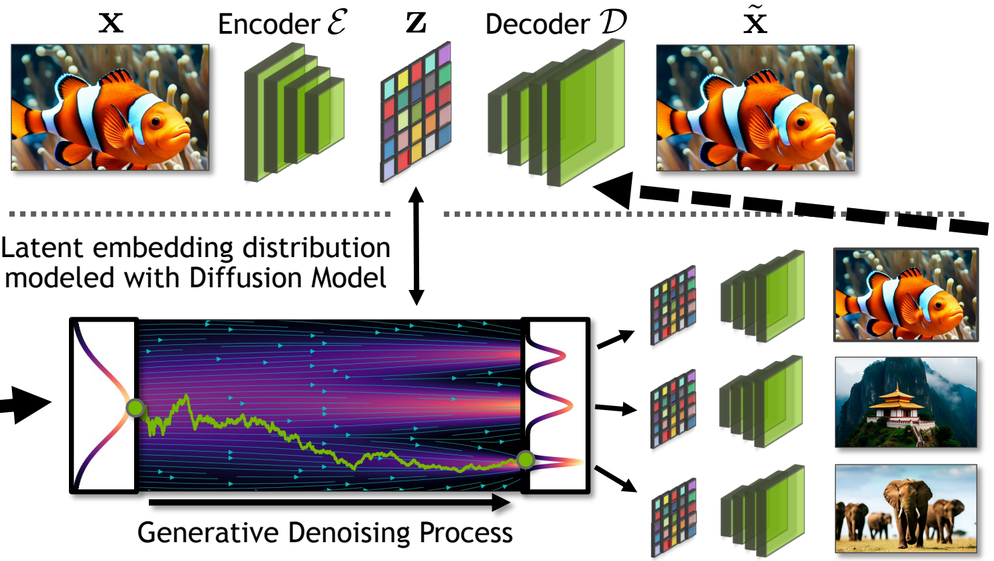

What makes an image look realistic and high-quality?


Realistic Global Structure
(correct placement of ears,
eyes, fur pattern, etc.)

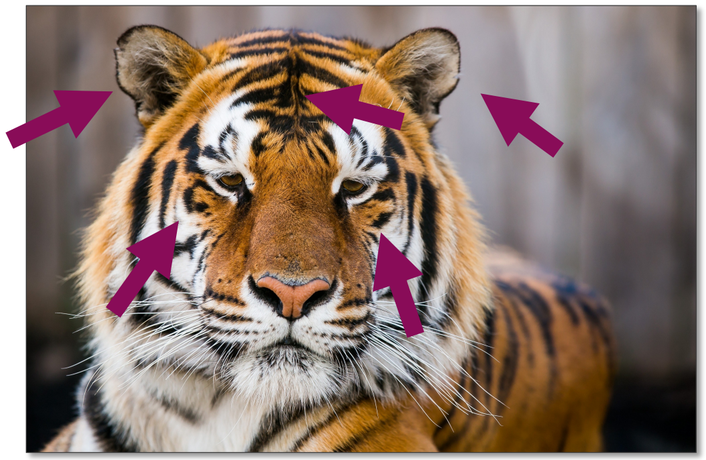

Realistic Local Details
(fine-grained texture)

# Latent Diffusion Models

Map Data into Compressed Latent Space. Train Diffusion Model efficiently in Latent Space.




**Global semantic structure** generated by latent **diffusion model.**



**Local "imperceptible" details** generated by **decoder** (with adversarial objective).

Latent space compression needs to be tuned carefully!

Balance between what content the latent DM needs to model and what is generated by decoder!

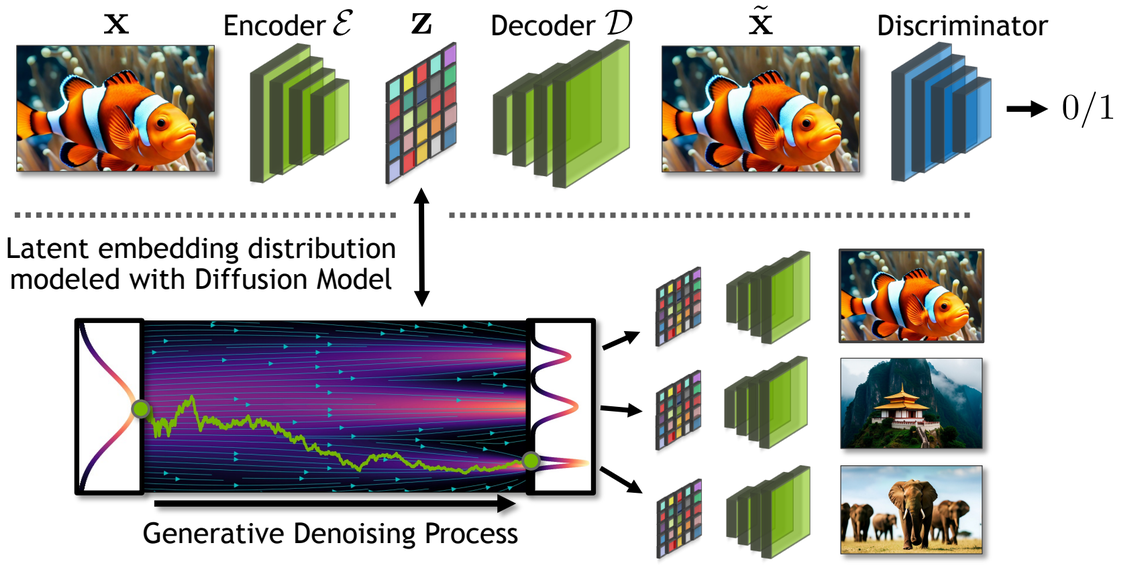
Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models", *CVPR*, 2022

# Latent Diffusion Models

## Latent Diffusion Models offer Excellent Trade-off between Performance and Compute Demands

*LDM "Recipe"*:

1. Train strong autoencoder
    - Compress…
      (downsampling factor / latent space regularization)
    - …while ensuring high visual quality on reconstructions
      ("upper bound" on synthesis quality)



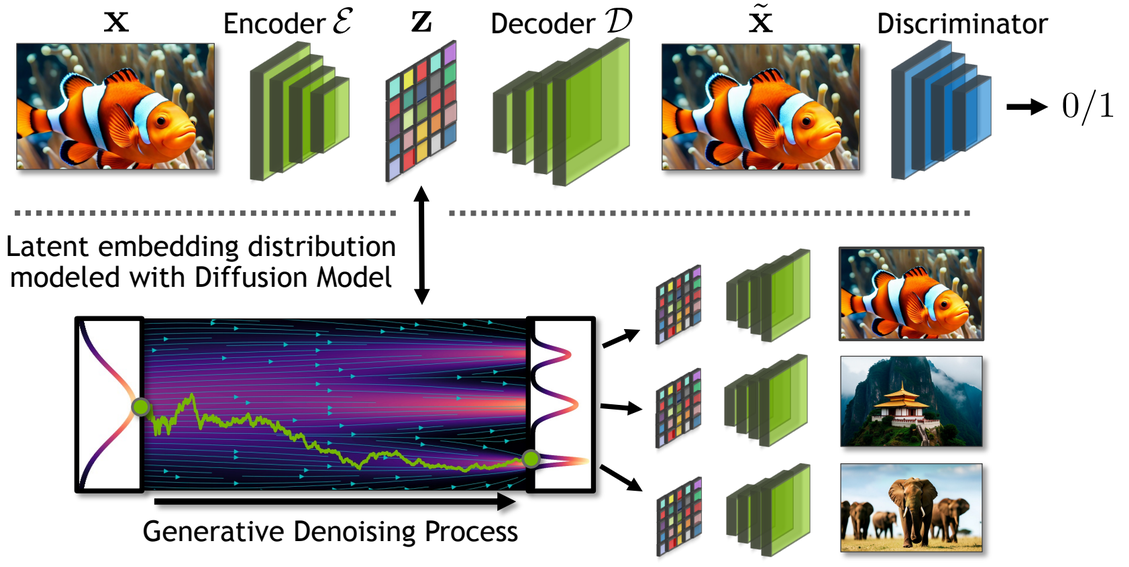

Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models", *CVPR*, 2022

# Latent Diffusion Models

## Latent Diffusion Models offer Excellent Trade-off between Performance and Compute Demands

*LDM "Recipe"*:

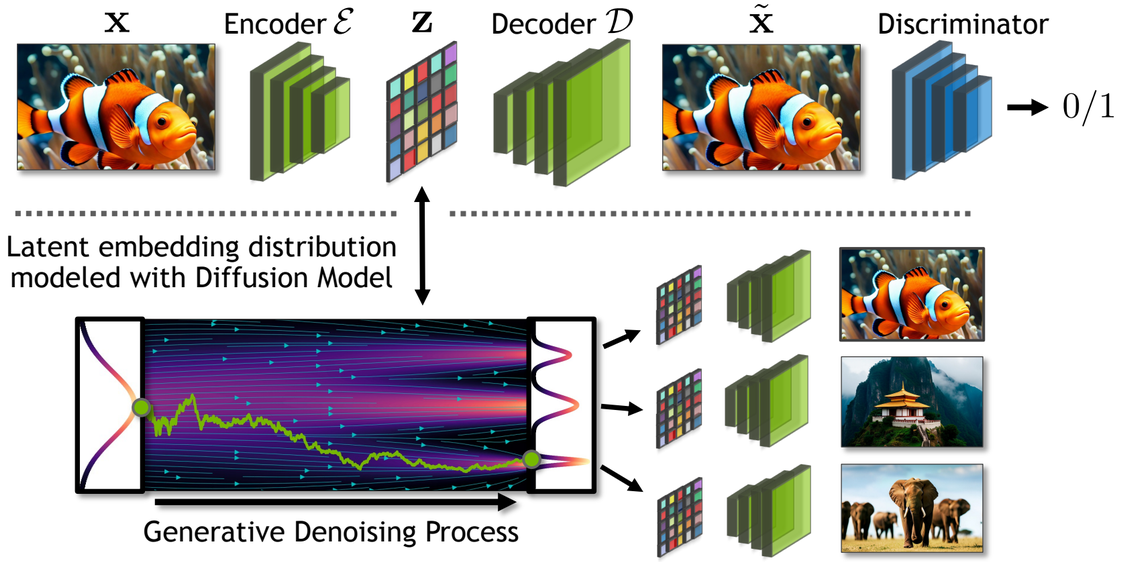
1. Train strong autoencoder
   - Compress…
     (downsampling factor / latent space regularization)
   - …while ensuring high visual quality on reconstructions
     ("upper bound" on synthesis quality)
2. Train efficient latent diffusion model
   - Latent space compression/regularization makes diffusion model training easier ➡ but trade-off with respect quality? Not really…
   - …because discriminator ➡ high quality despite compression (re-generate details, not encode)!




Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models", *CVPR*, 2022

# Latent Diffusion Models

## Latent Diffusion Models offer Excellent Trade-off between Performance and Compute Demands




➡ LDM with appropriate regularization, compression, downsampling ratio and strong autoencoder reconstruction:

- **Computationally efficient** diffusion model in latent space (compression & lower resolution).
- Yet **very high-performance** (latent diffusion + autoencoder + discriminator = ❤️).
- **Highly flexible** (can adjust autoencoder for different tasks and data).

# Image Generation with Latent Diffusion Models

Many state-of-the-art large-scale text-to-image models are latent diffusion models:

- Stability AI's **Stable Diffusion**
- Meta's **Emu**
- OpenAI's **DALL-E 3** and **Sora**

Common observation:

- (Latent) diffusion model **technology is mature** for practical image generation.
- The above models achieve high-performance mostly by **large data training** and **sophisticated data captioning**, **filtering** and **fine-tuning** strategies.

Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models", *CVPR*, 2022
Dai et al., "Emu: Enhancing Image Generation Models Using Photogenic Needles in a Haystack", *arXiv*, 2023
Betker et al., "Improving Image Generation with Better Captions" *(DALL-E 3), 2023*

https://stability.ai/news/stable-diffusion-sdxl-1-announcement

https://stability.ai/news/stable-diffusion-sdxl-1-announcement



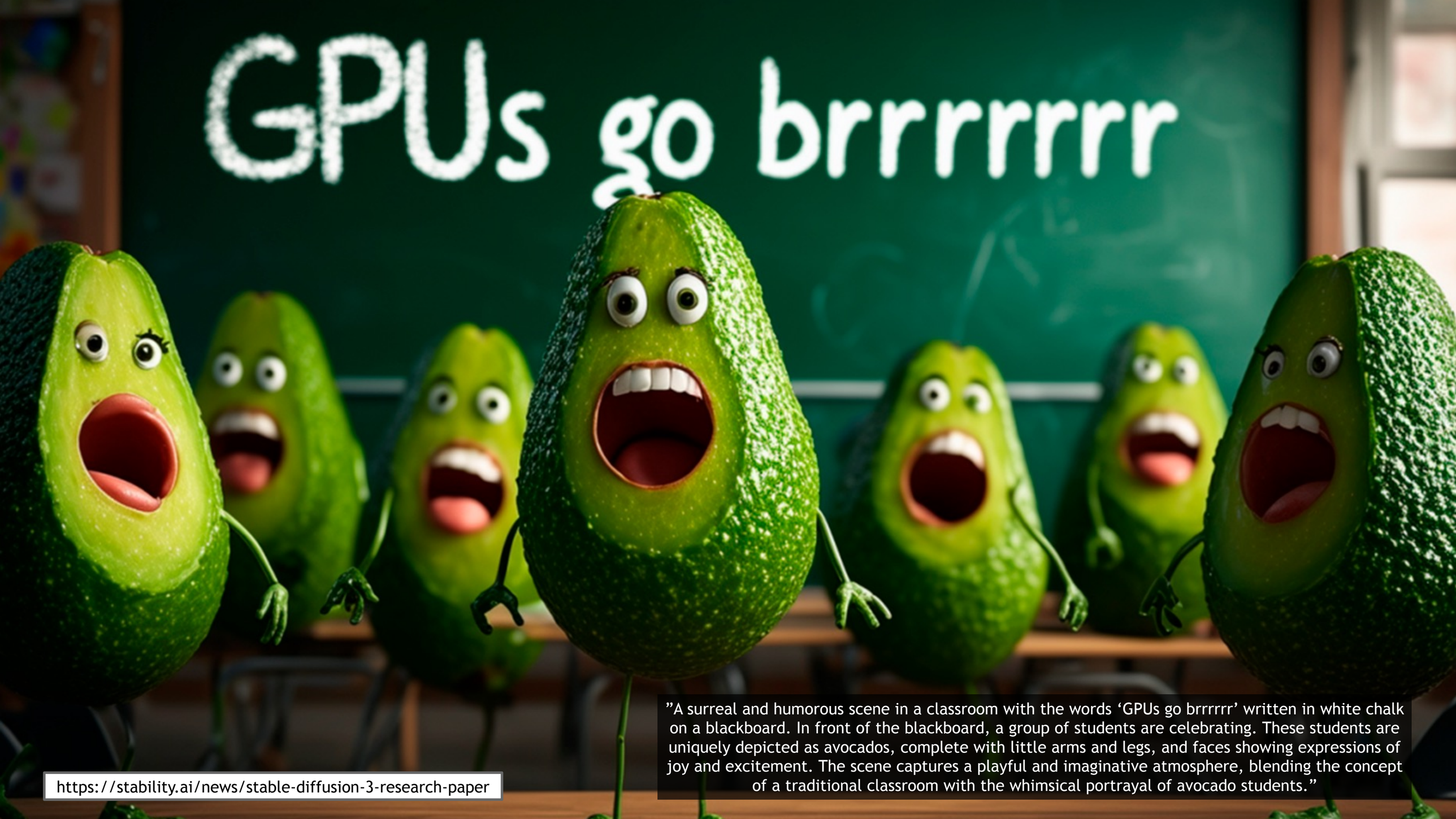




Dai et al., “Emu: Enhancing Image Generation Models Using Photogenic Needles in a Haystack”, *arXiv*, 2023


”Frog sitting in a 1950s diner wearing a leather jacket and a top hat. On the table is a giant burger and a small sign that says ”Froggy Fridays””

https://stability.ai/news/stable-diffusion-3-research-paper




”A surreal and humorous scene in a classroom with the words ‘GPUs go brrrrrr’ written in white chalk on a blackboard. In front of the blackboard, a group of students are celebrating. These students are uniquely depicted as avocados, complete with little arms and legs, and faces showing expressions of joy and excitement. The scene captures a playful and imaginative atmosphere, blending the concept of a traditional classroom with the whimsical portrayal of avocado students.”

https://stability.ai/news/stable-diffusion-3-research-paper

”In a fantastical setting, a highly detailed furry humanoid skunk with piercing eyes confidently poses in a medium shot, wearing an animal hide jacket. The artist has masterfully rendered the character in digital art, capturing the intricate details of fur and clothing texture.”

Betker et al., “Improving Image Generation with Better Captions” *(DALL-E 3), 2023*

# Agenda

- An Introduction to Diffusion Models
- Acceleration
- Conditioning & Guidance
- Personalization
- Latent Diffusion Models
- **Video Diffusion Models**
- 3D and 4D Generation



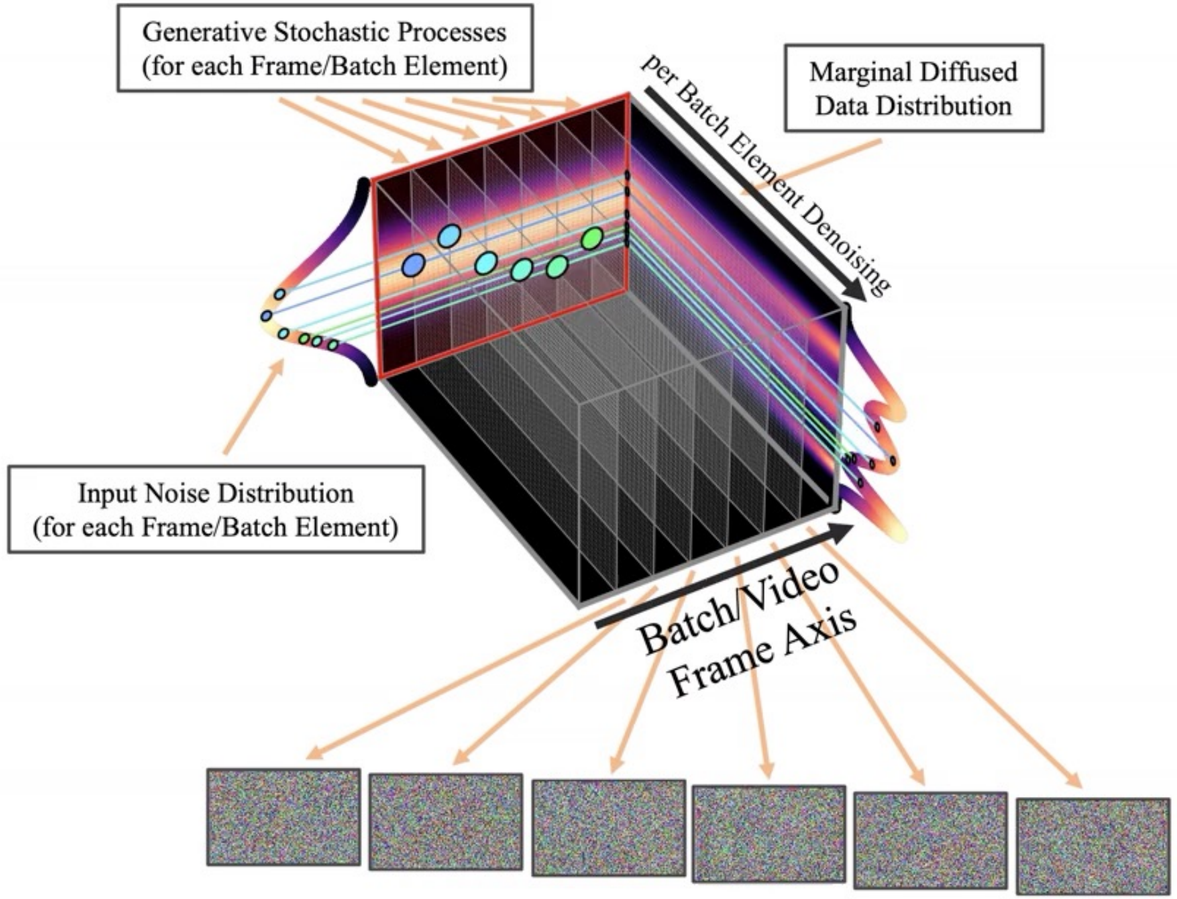
Blattmann et al., “Align Your Latents: High-Resolution Video Synthesis with Latent Diffusion Models”, *CVPR*, 2023
Emu Video, https://emu-video.metademolab.com/

”A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.”

*Brooks, Peebles, et al. https://openai.com/sora*

# From Image to Video Diffusion Models

## Common Idea: Temporally Align an Image Diffusion Model via Video Fine-tuning




Before temporal video fine-tuning,
different batch samples are independent.

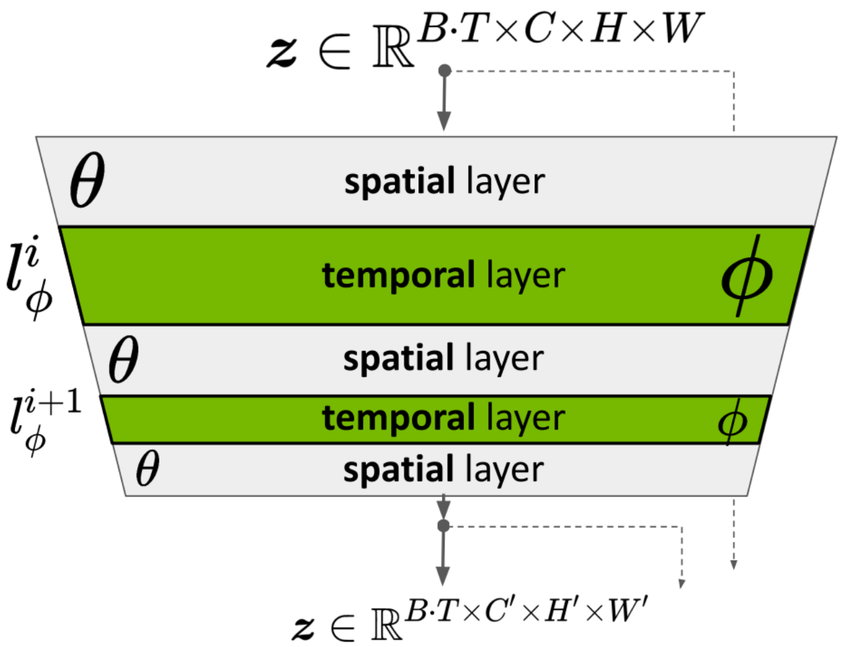
# From Image to Video Diffusion Models

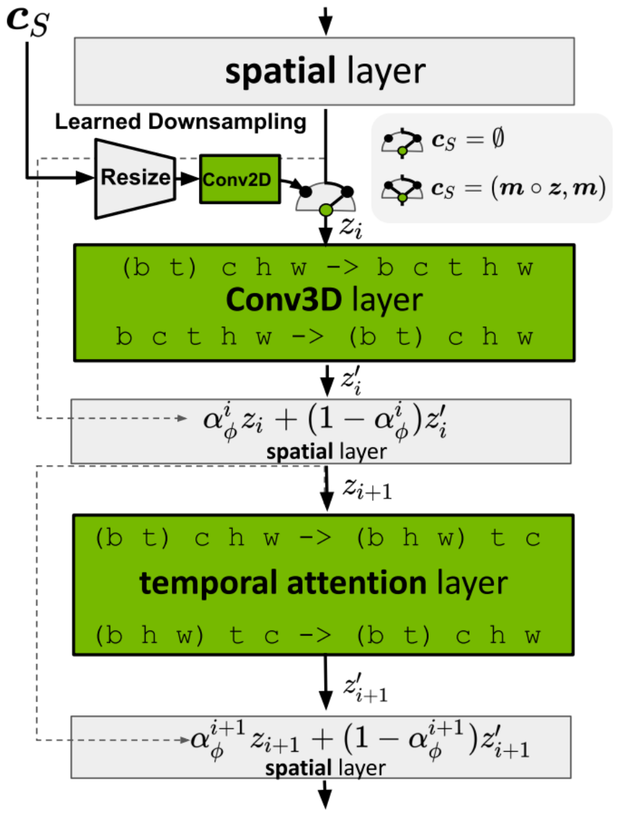
## Common Idea: Temporally Align an Image Diffusion Model via Video Fine-tuning

- **Spatial layers** interpret frames as independent images.
- **Temporal layers** interpret frames as sequence and model temporal dynamics.
- Implemented by shifting temporal axis into batch dimension for spatial layers.

1. Can train both spatial and temporal layers jointly, using both image and video datasets.

2. Or can fix pre-trained spatial layers and only train temporal layers with video data.

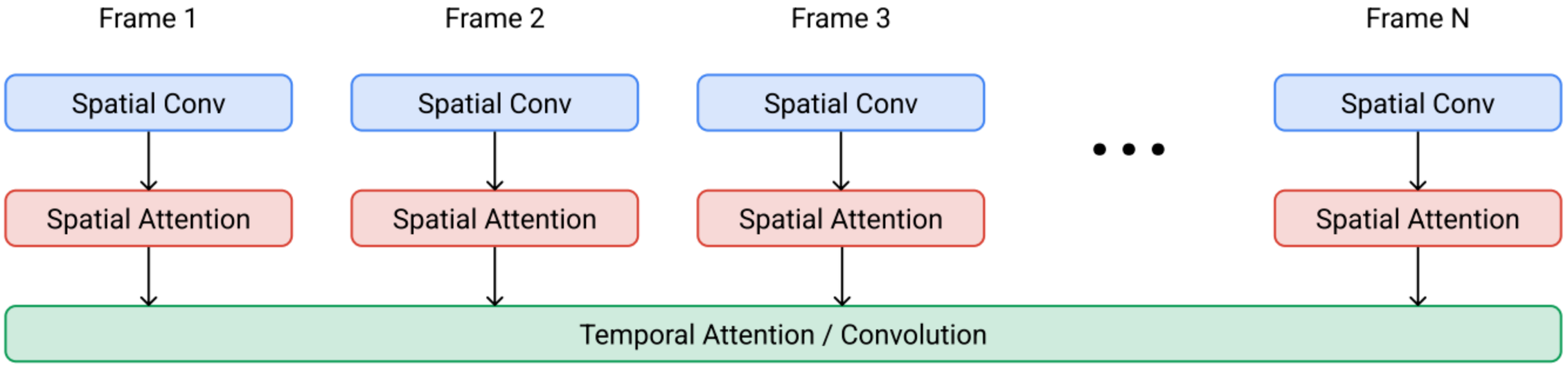



Blattmann et al., “Align Your Latents: High-Resolution Video Synthesis with Latent Diffusion Models”, *CVPR, 2023*

# From Image to Video Diffusion Models

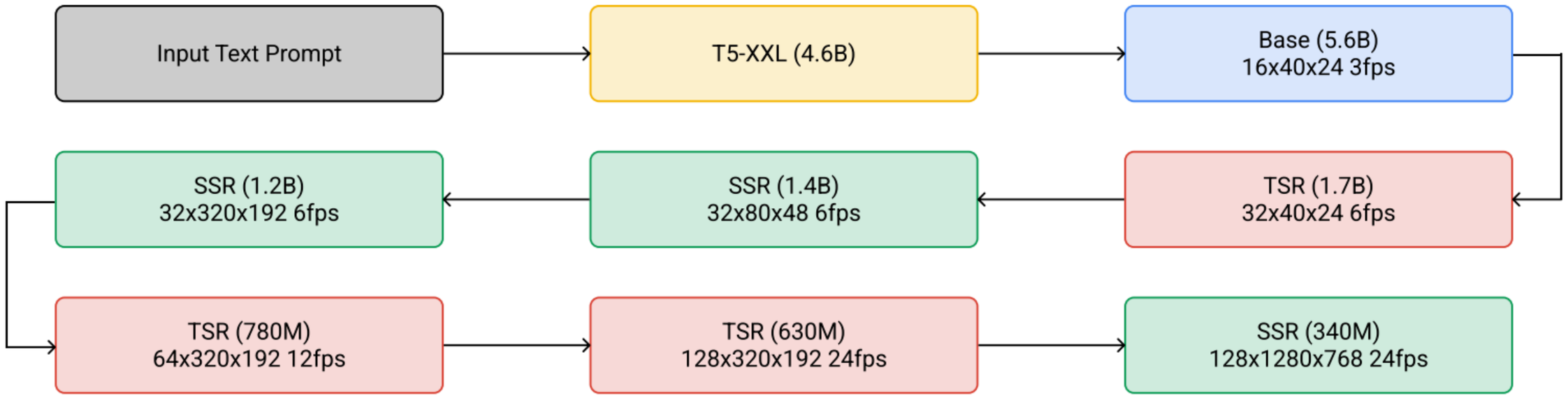
## Common Idea: Temporally Align an Image Diffusion Model via Video Fine-tuning




Ho et al., “Imagen Video: High Definition Video Generation with Diffusion Models”, 2022

# From Image to Video Diffusion Models

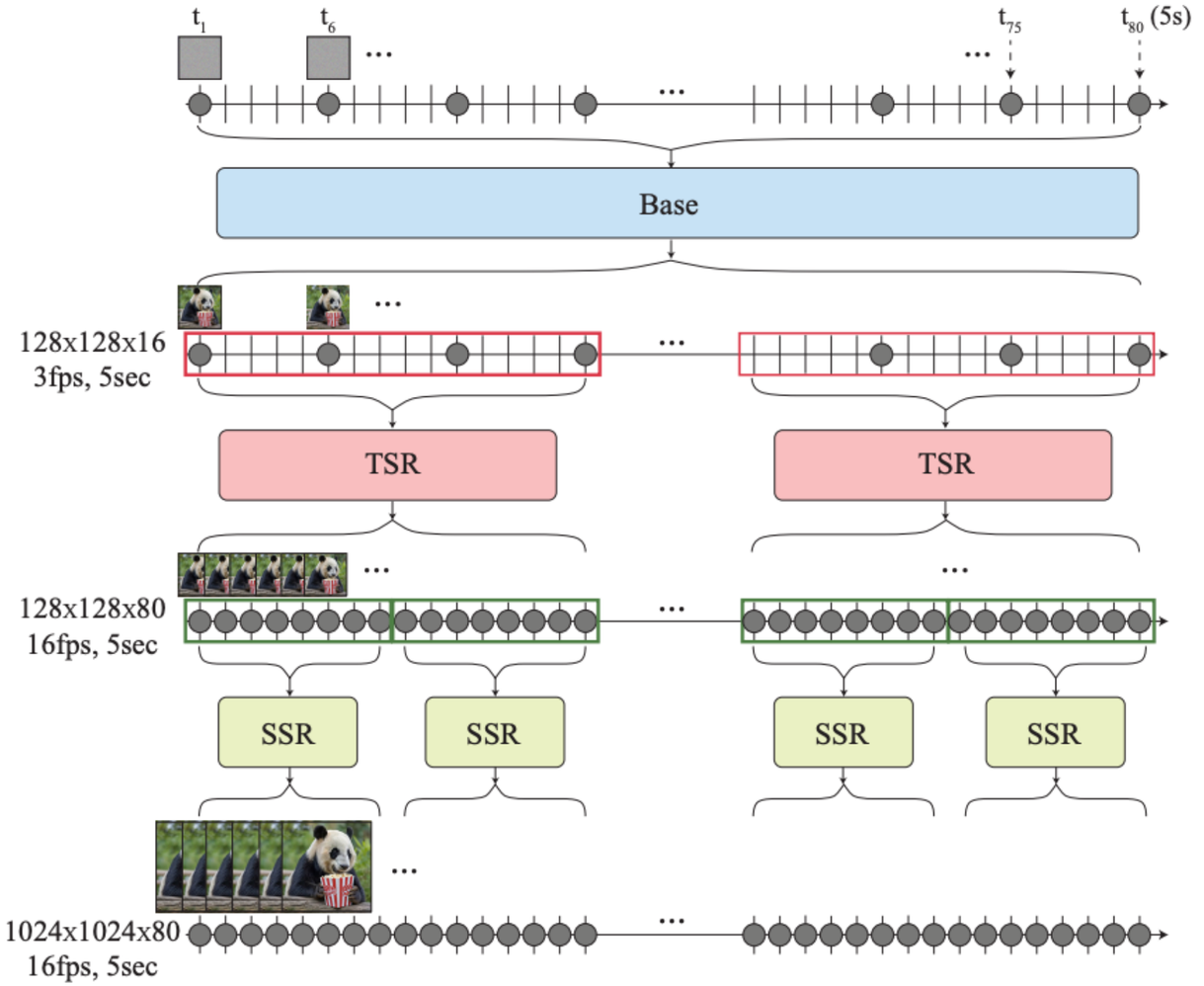
## Spatial and Temporal Cascaded Diffusion Models




Ho et al., “Imagen Video: High Definition Video Generation with Diffusion Models”, 2022

# From Image to Video Diffusion Models

## Spatial and Temporal Cascaded Diffusion Models




*Bar-Tal et al.*, “Lumiere: A Space-Time Diffusion Model for Video Generation”, 2024

# Video Latent Diffusion Models

*“A storm trooper vacuuming the beach.”*

*“Two pandas discussing an academic paper.”*

Blattmann et al., “Align Your Latents: High-Resolution Video Synthesis with Latent Diffusion Models”, *CVPR, 2023*

# Video Latent Diffusion Models

*“Close up of grapes on a rotating table. High definition.”*

*“Sunset time lapse at the beach with moving clouds and colors in the sky, 4k, high resolution.”*

Blattmann et al., “Align Your Latents: High-Resolution Video Synthesis with Latent Diffusion Models”, *CVPR, 2023*

# Stable Video Diffusion

- Larger and carefully curated dataset and careful final fine-tuning on high-quality data
- End-to-end training including image backbone

Blattmann et al., “Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets”, *2023*

The camera directly faces colorful buildings in Burano Italy. An adorable dalmation looks through a window on a building on the ground floor. Many people are walking and cycling along the canal streets in front of the buildings.”

*Brooks, Peebles, et al. https://openai.com/sora*

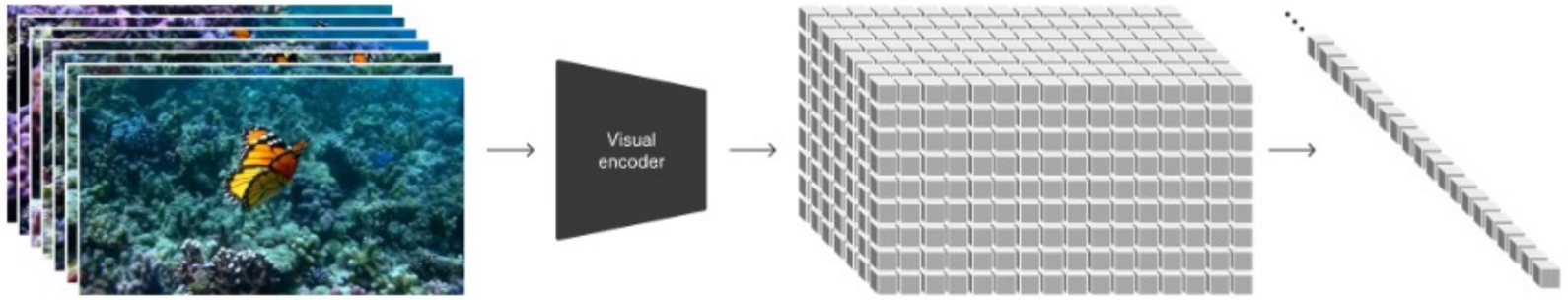
”An extreme close-up of an gray-haired man with a beard in his 60s, he is deep in thought pondering the history of the universe as he sits at a cafe in Paris, his eyes focus on people offscreen as they walk as he sits mostly motionless, he is dressed in a wool coat suit coat with a button-down shirt , he wears a brown beret and glasses and has a very professorial appearance, and the end he offers a subtle closed-mouth smile as if he found the answer to the mystery of life, the lighting is very cinematic with the golden light and the Parisian streets and city in the background, depth of field, cinematic 35mm film.”

*Brooks, Peebles, et al. https://openai.com/sora*

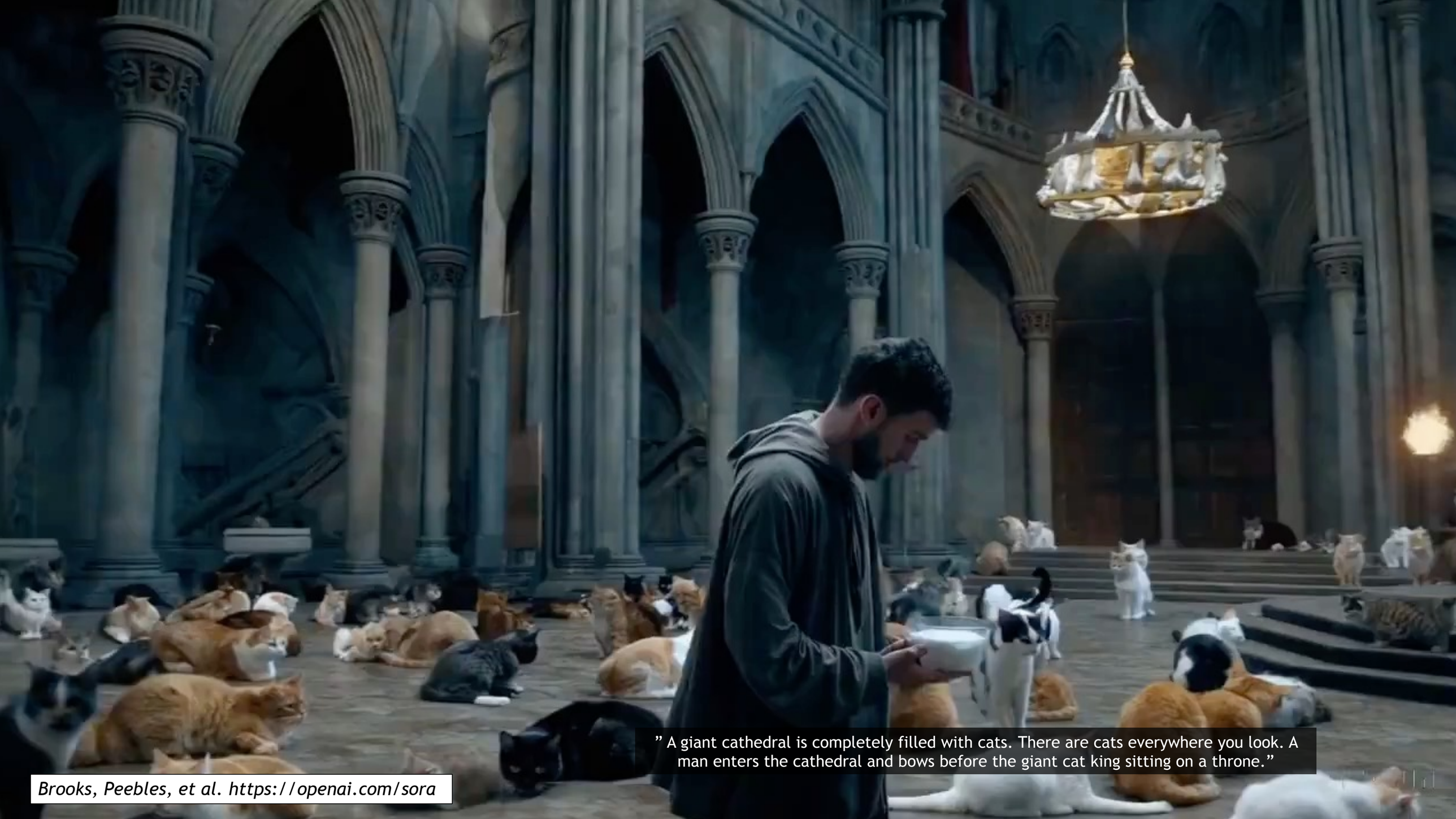
# Sora

## State-of-The-Art Text-to-Video Diffusion Model

- Spatio-temporal autoencoder
- Patch embeddings in latent space (like in vision transformers, but in latent space; similar to tokens in LLMs)
- Diffusion model over patches
- Neural Network likely based on large transformer (similar technology as in LLMs)




*Brooks, Peebles, et al. https://openai.com/research/video-generation-models-as-world-simulators*



” A giant cathedral is completely filled with cats. There are cats everywhere you look. A man enters the cathedral and bows before the giant cat king sitting on a throne.”

*Brooks, Peebles, et al. https://openai.com/sora*


*Brooks, Peebles, et al. https://openai.com/sora*

”A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.”

*Brooks, Peebles, et al. https://openai.com/sora*

# Agenda

- An Introduction to Diffusion Models
- Acceleration
- Conditioning & Guidance
- Personalization
- Latent Diffusion Models
- Video Diffusion Models
- 3D and 4D Generation

# 3D Diffusion Models?

- Large-scale image diffusion models require 100’s million to billions of image-caption pairs for training.
  - → Internet

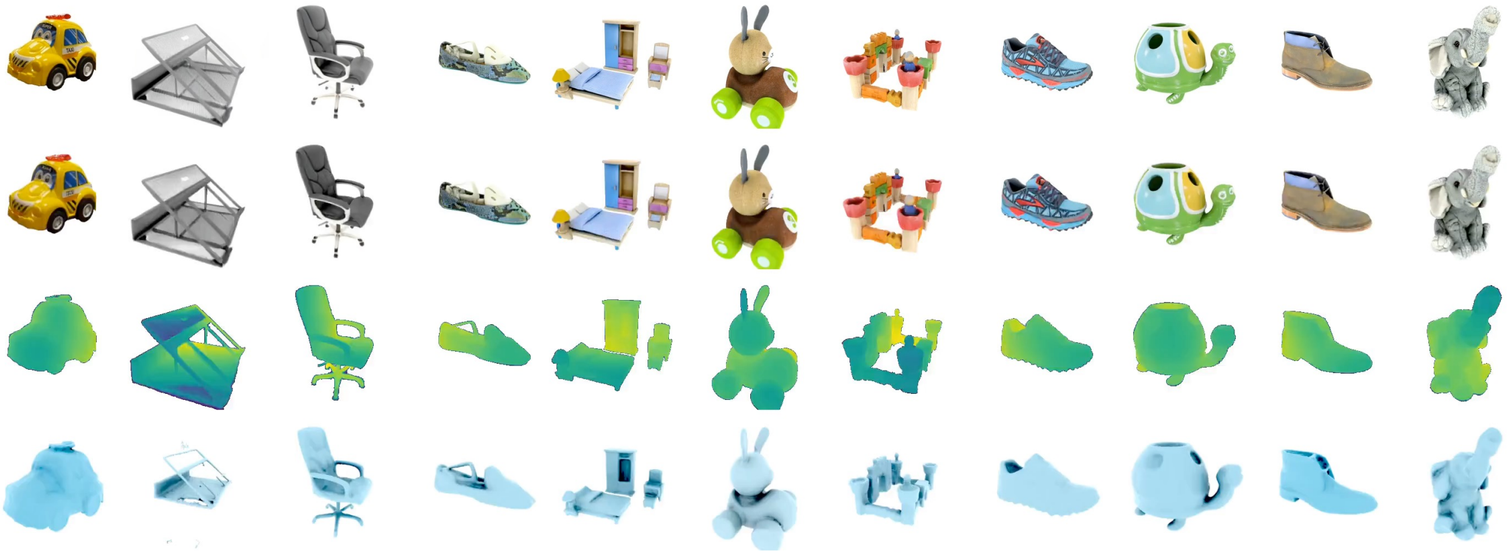
- 3D objects? Scarce.
  - → Objaverse-XL ~ 10 million objects only.

*But can we also use large-scale text-to-image models for 3D generation?*

↓

**Score Distillation Sampling**

We can train 3D generative models on this data:



*Xu et al.*, “DMV3D: Denoising Multi-View Diffusion using 3D Large Reconstruction Model”, 2023

# Text-to-3D with Score Distillation Sampling

1. 3D object, parametrized through learnable 3D representation (radiance field, mesh, 3D Gaussians, etc.)

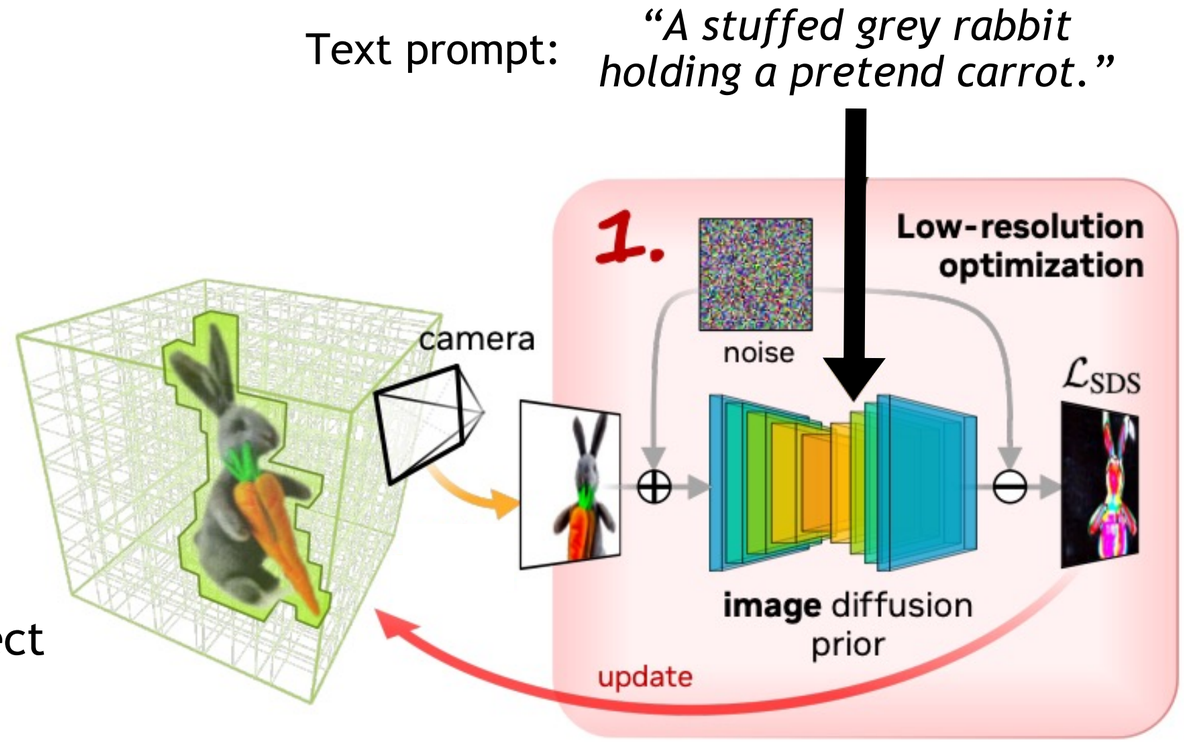
2. (Differentiably) render 3D object from different camera perspectives.

3. Provide 2D renderings to pre-trained large-scale text-to-image diffusion model.

4. Diffusion model has learnt what a good image of the object looks like for the text prompt ➡ feedback/gradient.

5. Backprop diffusion model gradient back into 3D representation to make it look realistic from all camera directions.

6. If it looks good from all directions, it is likely 3D consistent, too!




Poole et al., "DreamFusion: Text-to-3D using 2D Diffusion", ICLR 2023
Lin et al., "Magic3D: High-Resolution Text-to-3D Content Creation", CVPR 2023

# Text-to-3D with Score Distillation Sampling

Score Distillation Sampling with **video diffusion models**?

Generate **moving** & **dynamic 3D objects**?

**Text-to-4D** generation?

*a beautiful dress made out of garbage bags*

*an imperial state crown of england*

*a blue poison-dart frog sitting on a water lily*

*neuschwanstein castle, aerial view*

Lin et al., “Magic3D: High-Resolution Text-to-3D Content Creation”, CVPR 2023

# Text-to-4D

“A corgi is running fast.”

“A llama running fast.”

“A bee fluttering its wings fast.”

“A panda dancing.”

“Clown fish swimming.”

“A turtle swimming.”

Ling et al., “Align Your Gaussians”, CVPR, 2024

# Text-to-4D

“A monkey is playing bass.”

“A dog wearing a Superhero outfit with red cape flying through the sky.”

“A squirrel playing on a swing set.”

“A panda surfing a wave.”

“A cat singing.”

Ling et al., “Align Your Gaussians”, CVPR, 2024

# Text-to-4D

“Volcano eruption.”

“Assassin with sword running fast.”

“Waves crashing against a lighthouse.”

“Beer pouring into a glass.”

“Wood on fire.”

Ling et al., “Align Your Gaussians”, CVPR, 2024

# Autoregressively Extended Generation with Changing Prompts

“Running.” ➡ “Walking.” ➡ “Dancing.”

Ling et al., “Align Your Gaussians”, CVPR, 2024

# Composing Dynamic 4D Assets in Scenes

Ling et al., “Align Your Gaussians”, CVPR, 2024

# Composing Dynamic 4D Assets in Scenes

Ling et al., “Align Your Gaussians”, CVPR, 2024

# Diffusion Models - Summary

- **Generation by denoising:** Diffusion models provide a powerful generation framework based on iterative denoising, inverting a diffusion process.
- **Acceleration:** Diffusion models sampling can be slow, but many frameworks for acceleration exist.
- **Conditioning and guidance:** Diffusion models offer powerful guidance, control and editing capabilities.
- **Personalization:** We can fine-tune diffusion models on a few images for personalized generation.
- **Latent diffusion models:** Powerful framework combining efficiency, flexibility and high expressivity.
- **Video generation:** Diffusion models can be used for state-of-the-art video generation.
- **3D and 4D generation:** We can distill static 3D as well as dynamic 4D objects from pre-trained text-to-image and text-to-video diffusion models.

**Many more applications beyond visual domain!** Audio and music, biology and chemistry, climate modeling, …


@ArashVahdat

http://arashvahdat.com/

@karsten_kreis

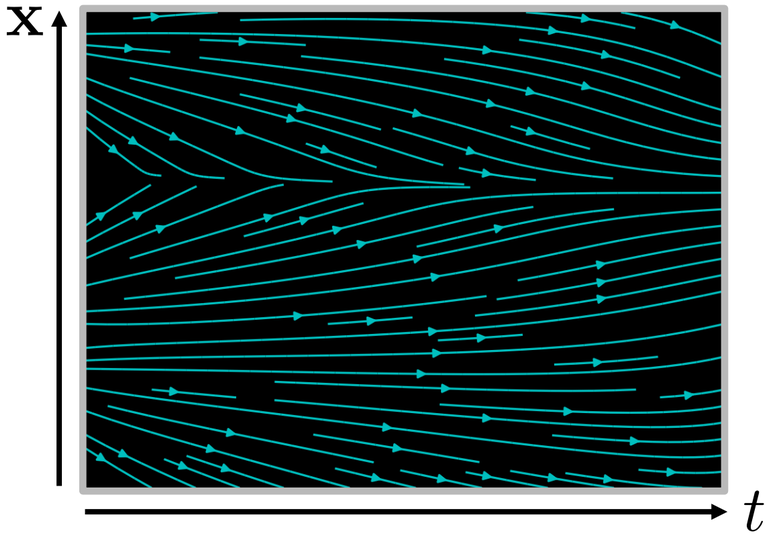
https://karstenkreis.github.io/

# Backup slides

# Crash Course in Differential Equations

**Ordinary Differential Equation (ODE):**

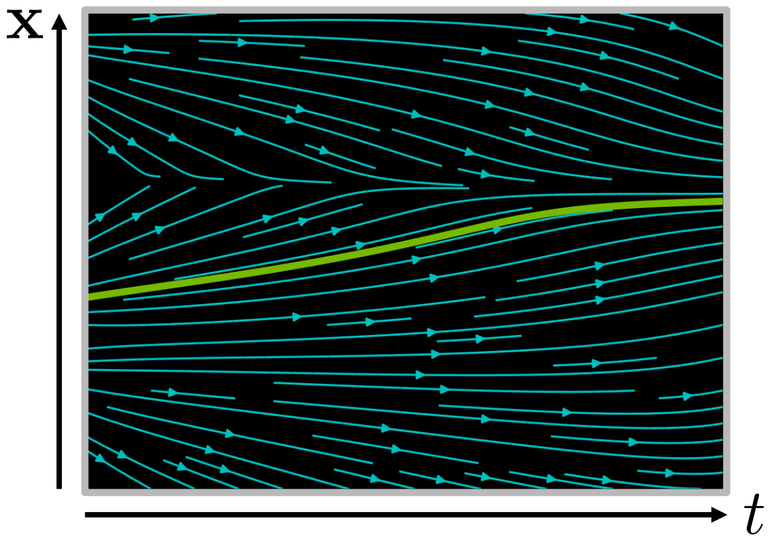
$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}, t) \quad \text{or} \quad \mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t$$



# Crash Course in Differential Equations

**Ordinary Differential Equation (ODE):**

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}, t) \quad \text{or} \quad \mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t$$




Analytical Solution:

$$\mathbf{x}(t) = \mathbf{x}(0) + \int_0^t \mathbf{f}(\mathbf{x}, \tau)\mathrm{d}\tau$$

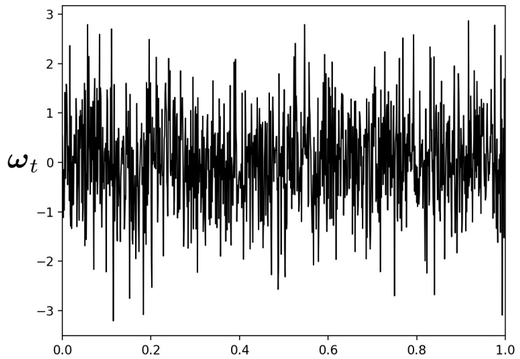
Iterative Numerical Solution:

$$\mathbf{x}(t + \Delta t) \approx \mathbf{x}(t) + \mathbf{f}(\mathbf{x}(t), t)\Delta t$$

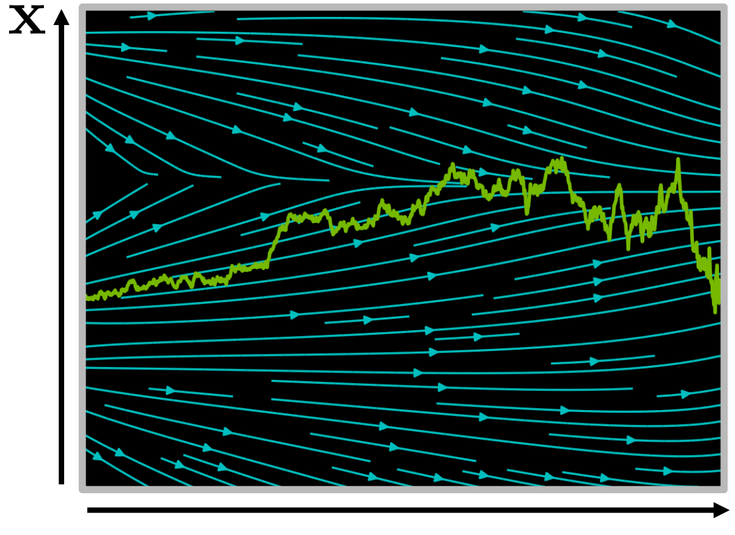
**Stochastic Differential Equation (SDE):**

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \underbrace{\mathbf{f}(\mathbf{x}, t)}_{\text{drift coefficient}} + \underbrace{\sigma(\mathbf{x}, t)\boldsymbol{\omega}_t}_{\text{diffusion coefficient}}$$

Wiener Process (Gaussian White Noise)




$$\left( \mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + \sigma(\mathbf{x}, t)\mathrm{d}\boldsymbol{\omega}_t \right)$$

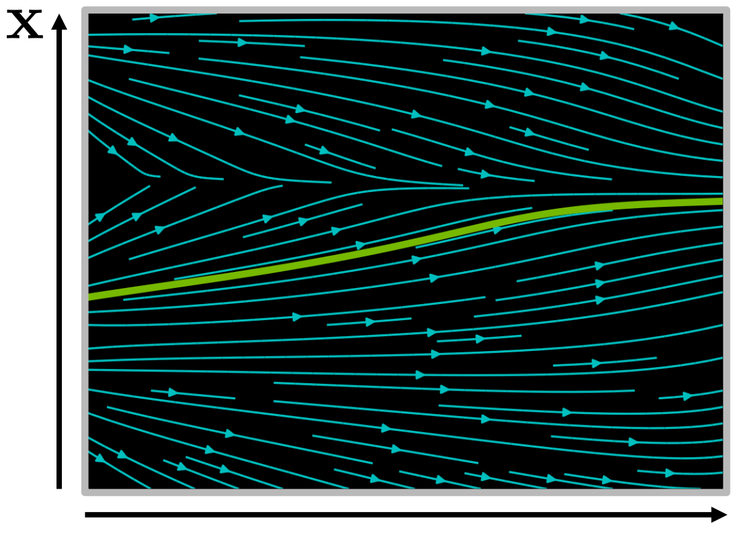



$$\mathbf{x}(t + \Delta t) \approx \mathbf{x}(t) + \mathbf{f}(\mathbf{x}(t), t)\Delta t + \sigma(\mathbf{x}(t), t)\sqrt{\Delta t}\,\mathcal{N}(\mathbf{0}, \mathbf{I})$$

# Crash Course in Differential Equations

**Ordinary Differential Equation (ODE):**

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}, t) \quad \text{or} \quad \mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t$$




Analytical Solution:

$$\mathbf{x}(t) = \mathbf{x}(0) + \int_0^t \mathbf{f}(\mathbf{x}, \tau)\mathrm{d}\tau$$

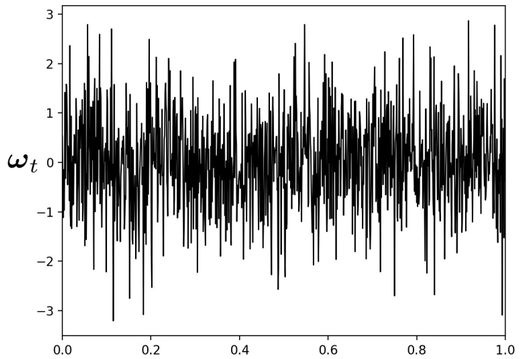
Iterative Numerical Solution:

$$\mathbf{x}(t + \Delta t) \approx \mathbf{x}(t) + \mathbf{f}(\mathbf{x}(t), t)\Delta t$$

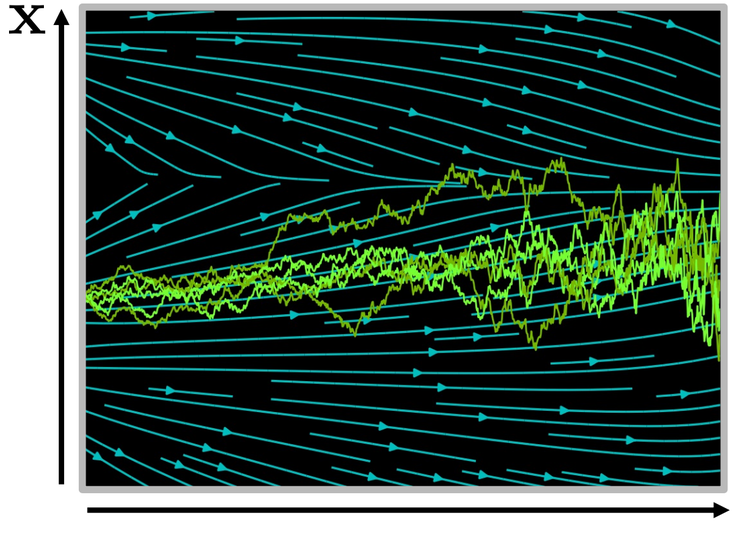
**Stochastic Differential Equation (SDE):**

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \underbrace{\mathbf{f}(\mathbf{x}, t)}_{\text{drift coefficient}} + \underbrace{\sigma(\mathbf{x}, t)\boldsymbol{\omega}_t}_{\text{diffusion coefficient}}$$

Wiener Process (Gaussian White Noise)




$$\left( \mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + \sigma(\mathbf{x}, t)\mathrm{d}\boldsymbol{\omega}_t \right)$$




$$\mathbf{x}(t + \Delta t) \approx \mathbf{x}(t) + \mathbf{f}(\mathbf{x}(t), t)\Delta t + \sigma(\mathbf{x}(t), t)\sqrt{\Delta t}\,\mathcal{N}(\mathbf{0}, \mathbf{I})$$