Diffusion Generative AI for Computer Vision and Science

—— Lecture 1: Introduction

Professor: Siyu Zhu

AI3 institute, Fudan University

Contents

- Biography
- Course Description
- Overview of Key Concepts

Biography

- ## Education Background
  - **Undergraduate: Zhejiang University; PhD: HKUST**
    - Superviser: Professor Long Quan（IEEE Fellow，Chair of ICCV 2011 and CVPR 2023）
  - **Research interests：3D reconstruction and Generative model for video and 3D**
    - Published over 60 papers in international conferences and journals(CVPR/ICCV/ECCV/TPAMI).
    - Achieved first place in over 40 international competitions and leaderboards related to 3D vision and graphics.
- ## Work Experience
  - **2013-2017 Co-founder, Altizure（acquired by Apple Inc）**
    - The city-scale 3D reconstruction technology broke the monopoly Google and Bentley in this field.
    - Apple's Object Capture provides 3D content creation tools to developers worldwide.
  - **2017-2023 Director, Alibaba Cloud/AI Lab**
    - Contribute to various aspects of computer vision and machine learning-related products.
  - **2023- Professor, AI institute, Fudan University**
    - Open-source projects in the fields of video and 3D generation, including the face video generation model Hallo (8.5K GitHub stars, ranked 2nd in the field of face video generation) and the human body video generation model Champ (3.6K GitHub stars, ranked 3rd in the field of human body video generation).

Contents

- Biography
- Course Description
- Overview of Key Concepts

Course Description

- **Objective**
  - Master the principles and theories of diffusion model-based generative AI.
  - Build programming skills and project development experience.
  - Apply diffusion-based generative AI to real-world problems.
- **Coursework**
  - Attendance 5%
  - Participation 5%
  - Literature Review 45%
  - Course Project 45%
  - **Students are encouraged to develop innovative projects that combine generative models with their academic disciplines.**

Course Schedule

Lecture #	Topic	Content
1	Introduction	Overview of this Course
2	Basics of Generative AI (1)	VAE, GAN, Flow-based models, and applications
3	Basics of Generative AI (2)	Probability distributions, random variables
4	Fundamentals of Diffusion Models (1)	Diffusion model principles, DDPMs, DDIMs, SGMs, Score SDEs, VDMs
5	Fundamentals of Diffusion Models (2)	Model structures for generation processes: Unet, DiT, Latent space
6	Machine Learning Frameworks and Tools	Machine learning frameworks and tools: training, inference, optimization
7	Images + Diffusion Models	LDM、Stable Diffusion、DALL-E、GigaGAN
8	Audio + Diffusion Models	Audio Diffusion Models、VideoDiffusion、Sora
9	3D + Diffusion Models	NeRF、3D-VAE、DreamFusion
10	Biology (Structure) + Diffusion Models	AlphaFold3、ESMFold、RFdiffusion、SE(3) Diffusion
11	Physics and Meteorology + Diffusion Models	GraphCast、Pangu-Meteorology、NowcastNet、Fuxi
12	Advanced Topics in Diffusion Models (1)	External expert talk
13	Advanced Topics in Diffusion Models (2)	External expert talk
14	Paper and Project Presentations	
15	Paper and Project Presentations	
16	Paper and Project Presentations	

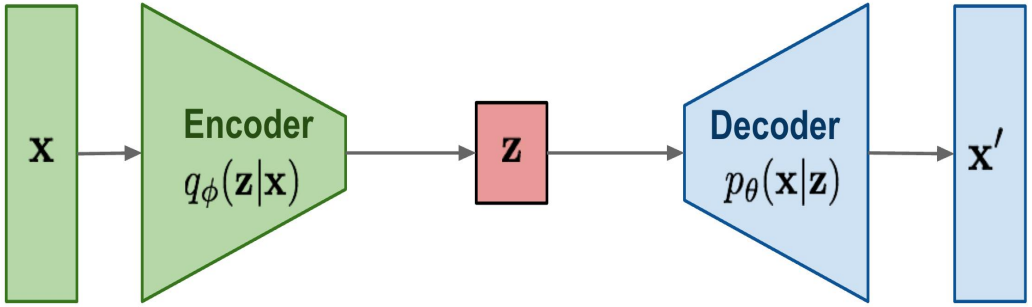
Contents

- Biography
- Course Description
- Overview of Key Concepts

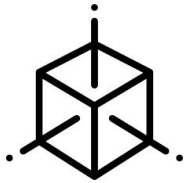
Overview of Generative AI

Input

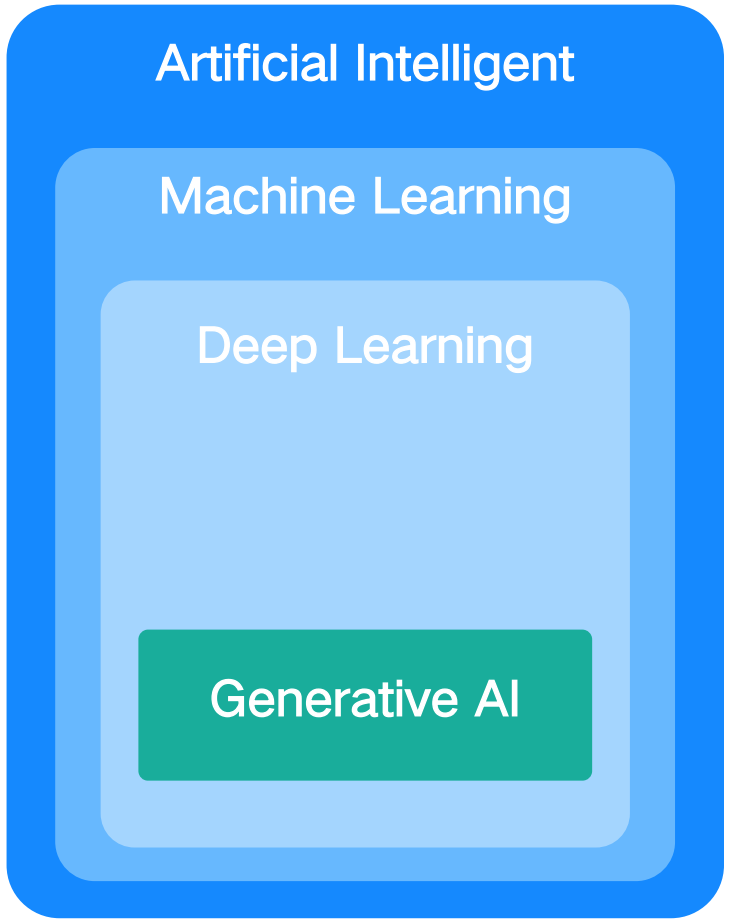
VAE: maximize variational lower bound



Output



Conception of AI



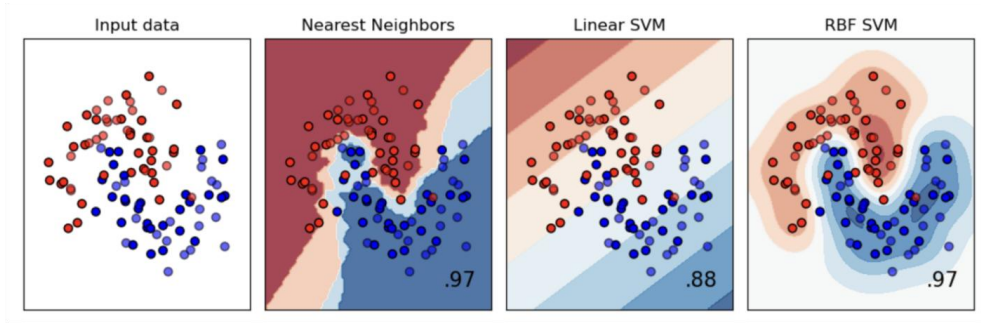
- **Artificial Intelligence (1956)**
  - Refers to computational intelligence designed to replace or surpass human intelligence.
- **Machine Learning (1990)**
  - A subfield of AI focused on analyzing large datasets to identify patterns, enabling systems to automatically learn, improve, and make accurate decisions and predictions.
- **Deep Learning (2012)**
  - A branch of machine learning that utilizes multi-layered neural networks to process data and make decisions.
- **Generative AI Models (2021)**
  - Based on specific prompts, data, or conditions, generative AI can produce new content, such as text, images, audio, video, or other material structures.

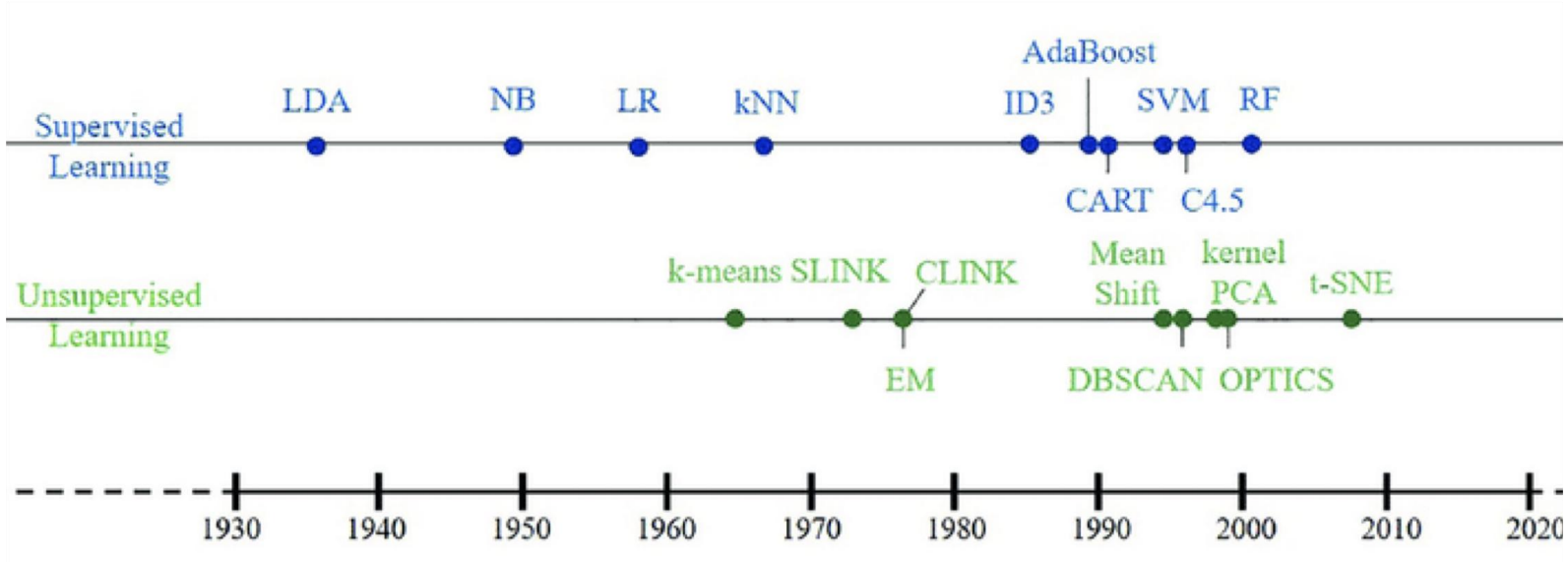
Artificial Intelligence (1956)

- The 1956 Dartmouth workshop was a landmark event that established AI as an academic discipline.
- AI received its name, defined its mission, achieved its first major success, and brought together its key players.

Machine Learning (1990)

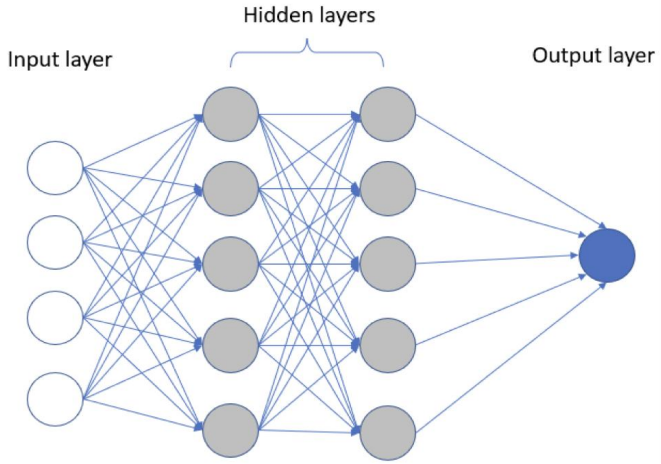
- Machine learning empowers systems to automatically analyze data, identify patterns, and make predictions or decisions.

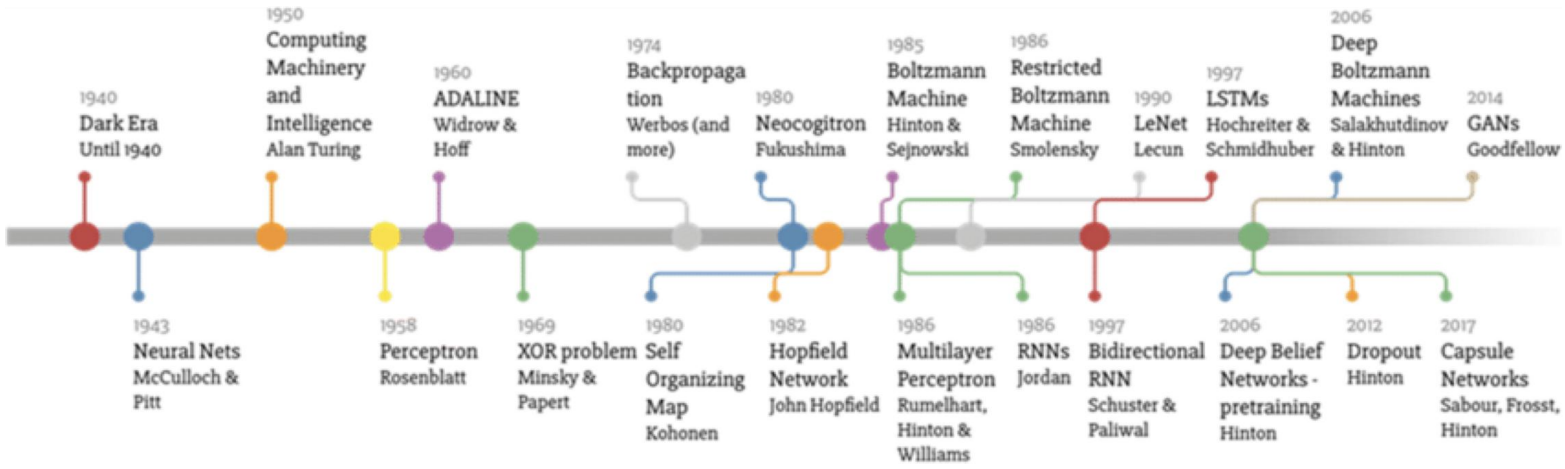




Deep Learning（2012）

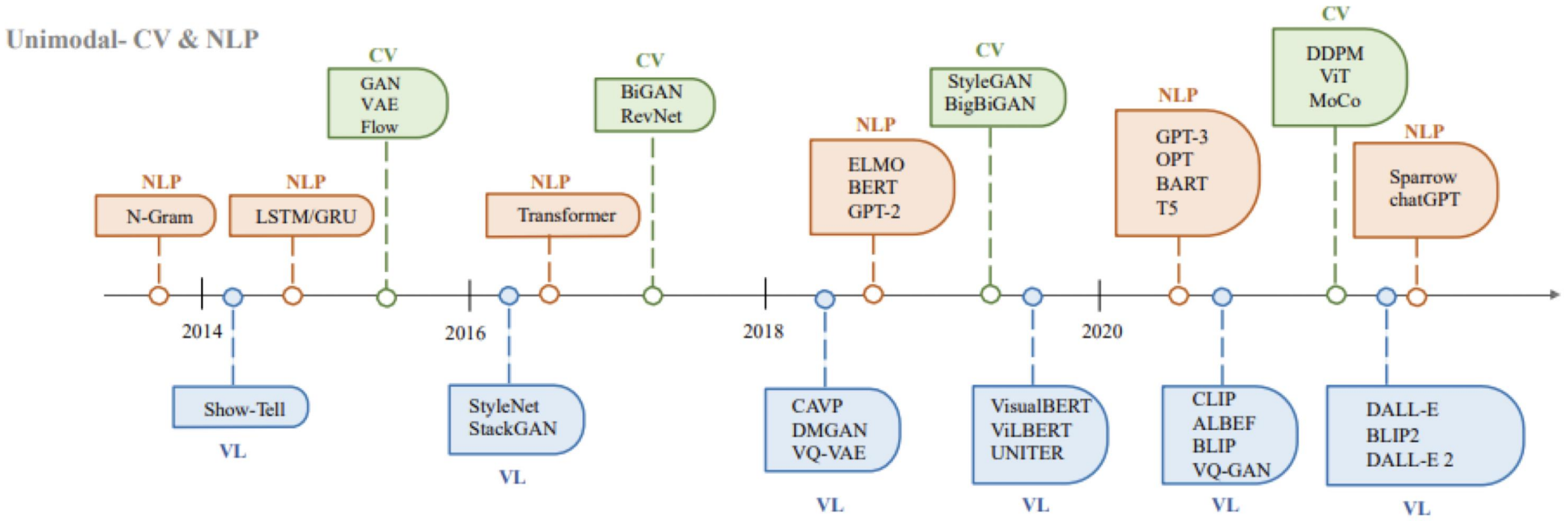
- Deep learning, a branch of machine learning, uses neural networks with multiple layers to model and interpret complex data.





Category of Generative Models

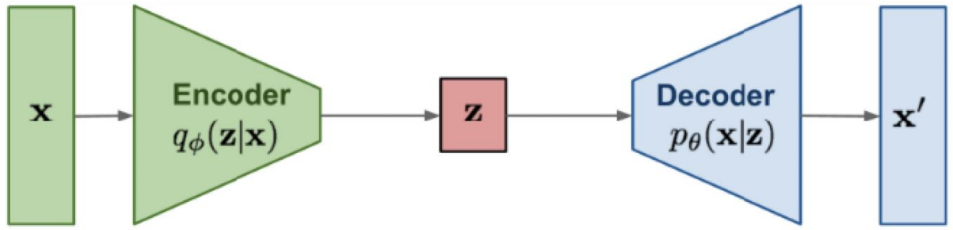
- Generative AI learns from existing data and generate entirely new content in various forms, such as text, images, music, and videos, with similar characteristics to the original data.
- Scaling Law + Self-supervised Learning
- Autoregressive Generative Models + Diffusion Generative Models



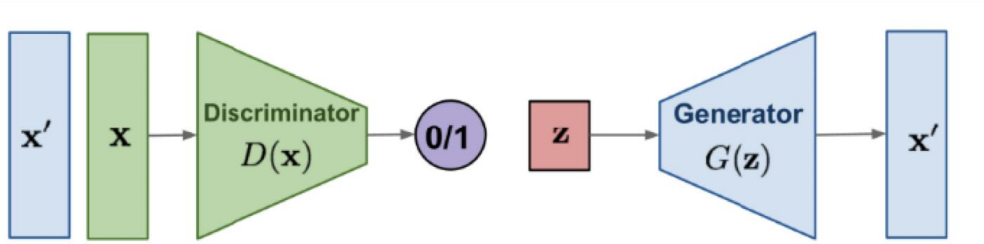
Category of Generative Models

- Large Vision Models, LVM

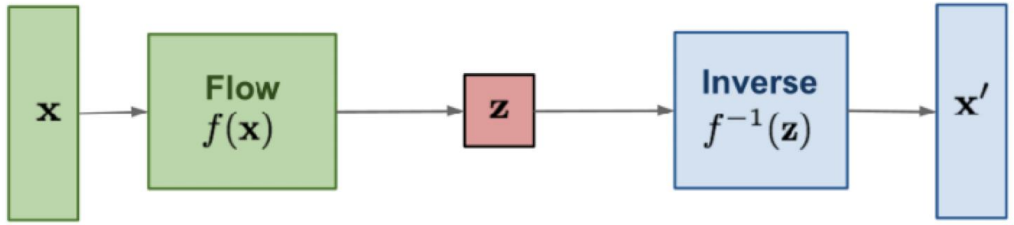
VAE: maximize variational lower bound



GAN: Adversarial training



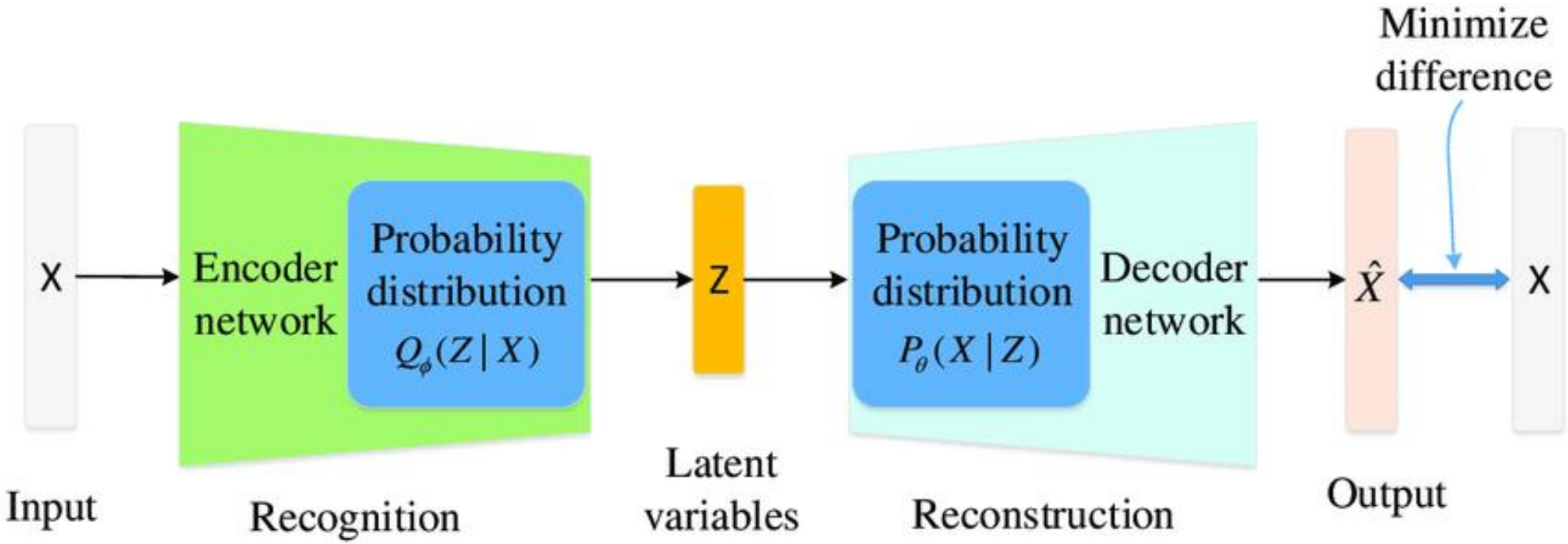
Flow-based models: Invertible transform of distributions



Diffusion models: Gradually add Gaussian noise and then reverse

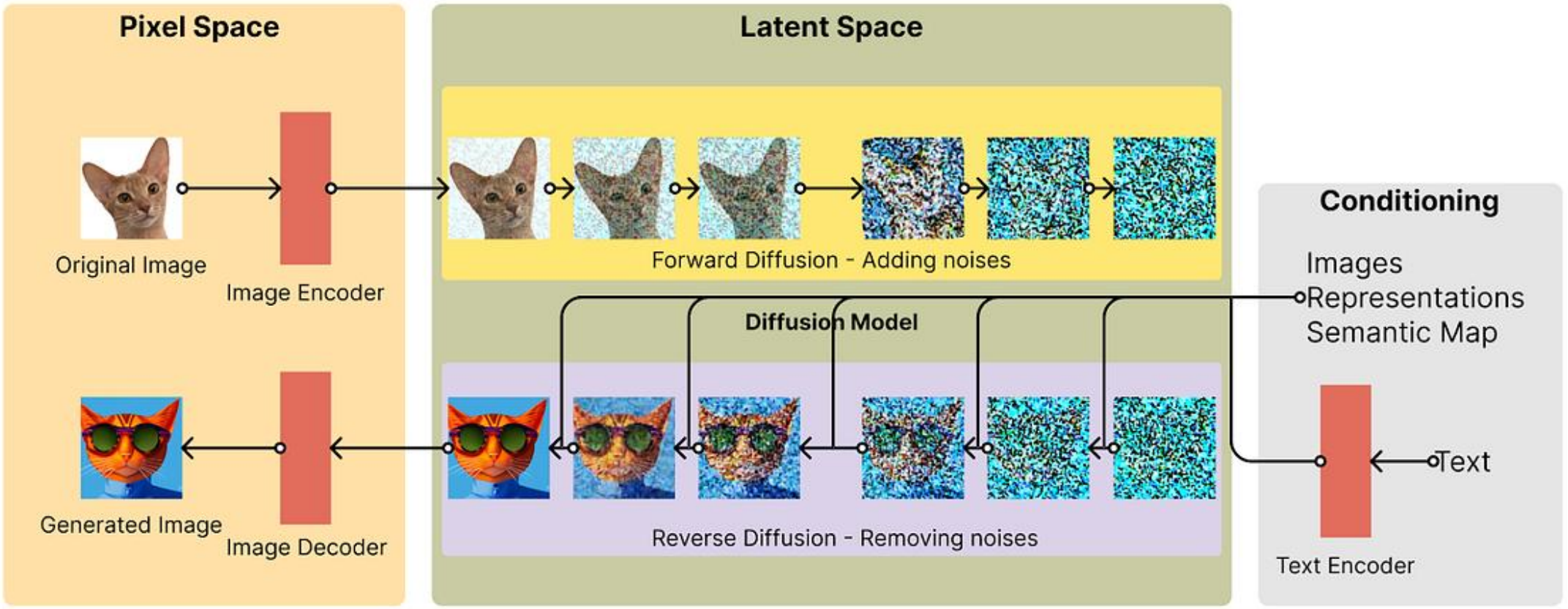
Unifying perception and generation

- Ilya Sutskever: compression is generalization.
  - The best lossless compression for a dataset is the best generalization for data outside the dataset.



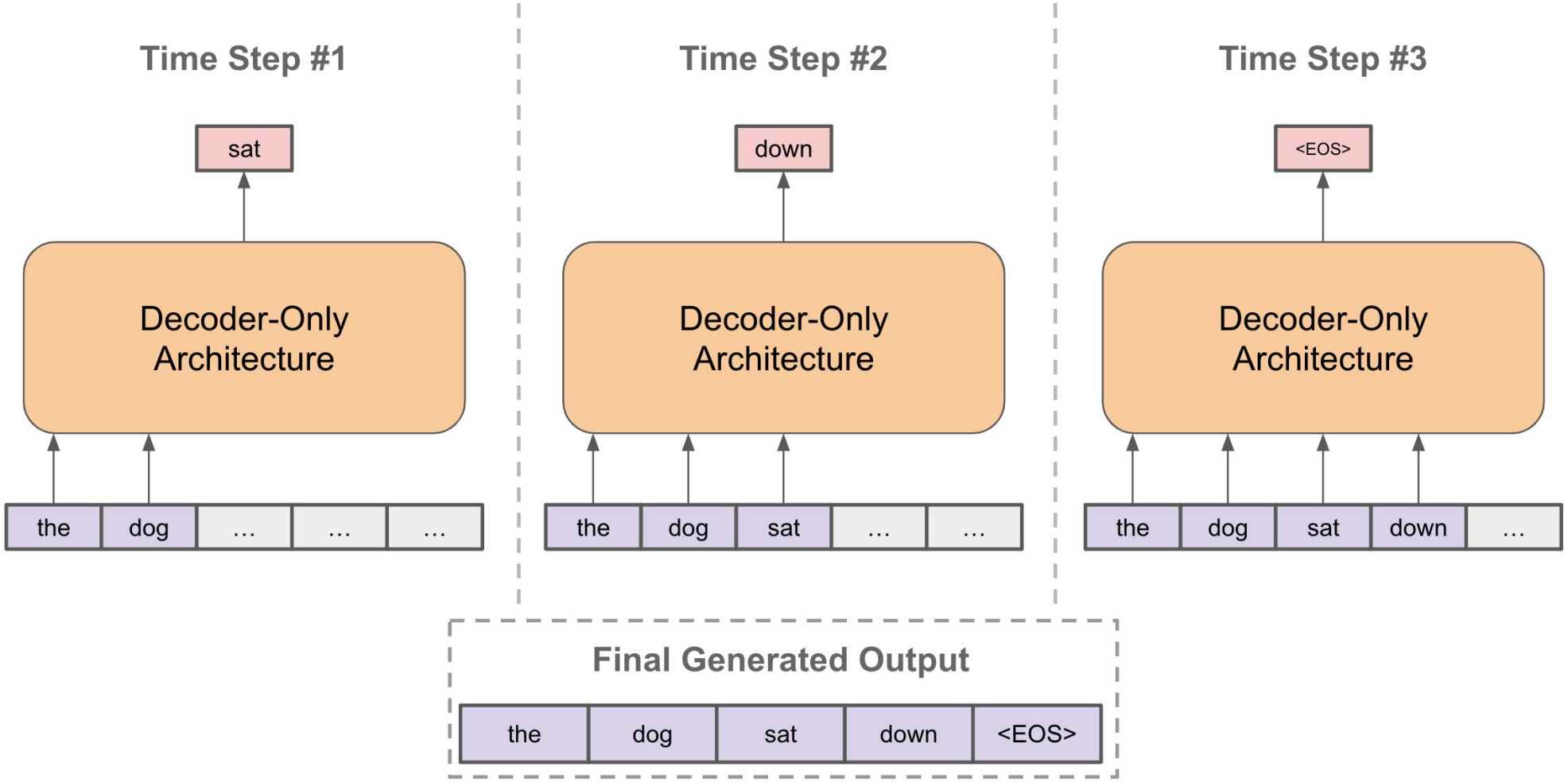
Large Models: Autoregression v.s. Diffusion

- Diffusion Model



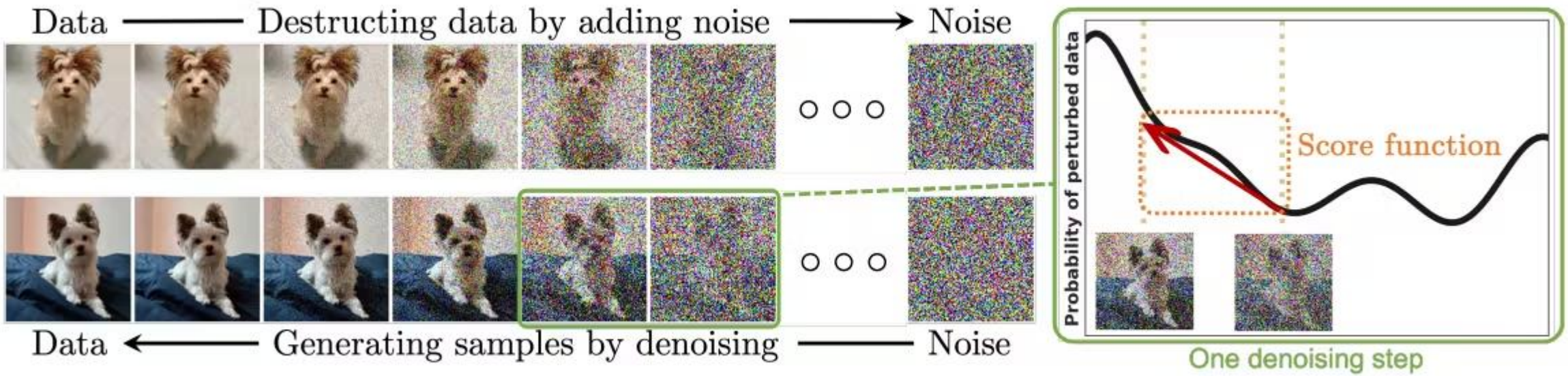
Large Models: Autoregression v.s. Diffusion

- Autoregressive Language Model



Diffusion Models：DDPMs

- Denoising Diffusion Probabilistic Models (DDPMs)



$$p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$$

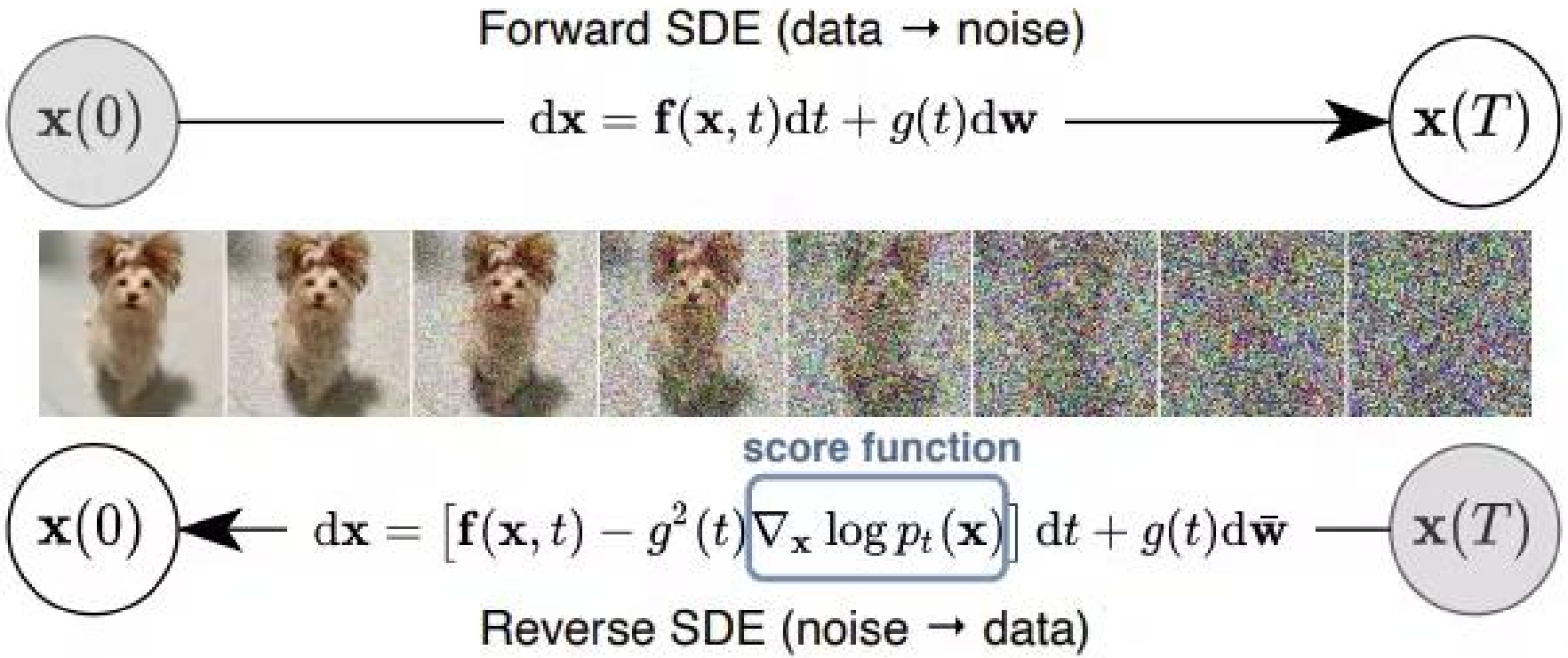
$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \underbrace{\mu_\theta(\mathbf{x}_t, t), \sigma_t^2}_{\text{Trainable network}}\mathbf{I})$$

(U-net, Denoising Autoencoder)

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

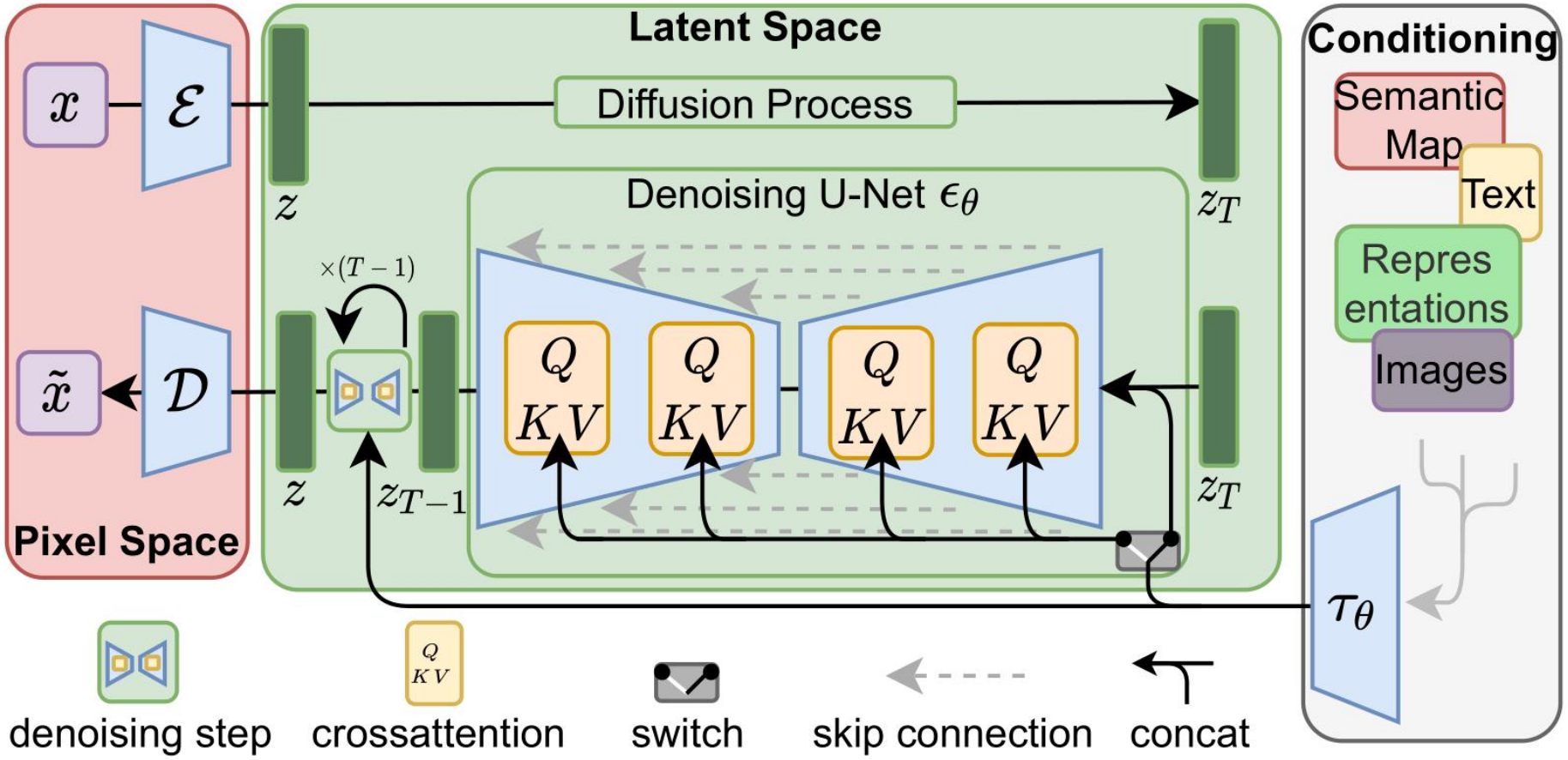
Diffusion Models：Score SDEs

- Stochastic Differential Equations (Score SDEs)



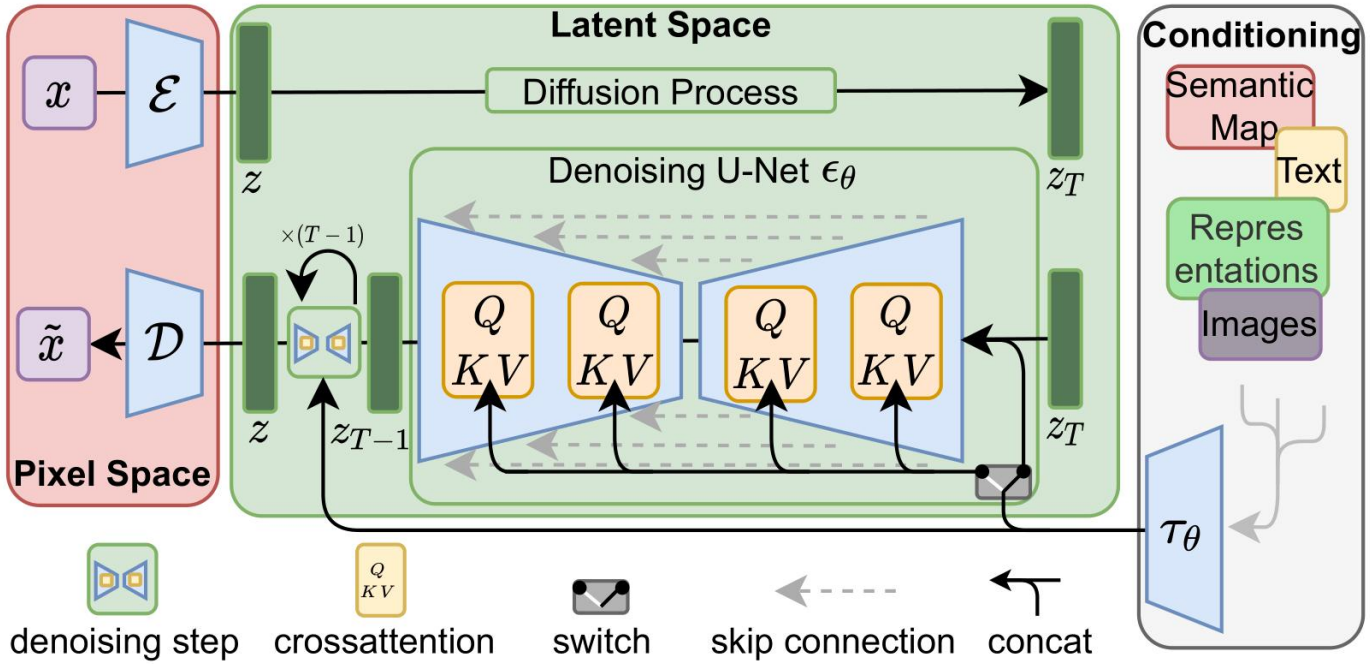
Architecture of Diffusion Models

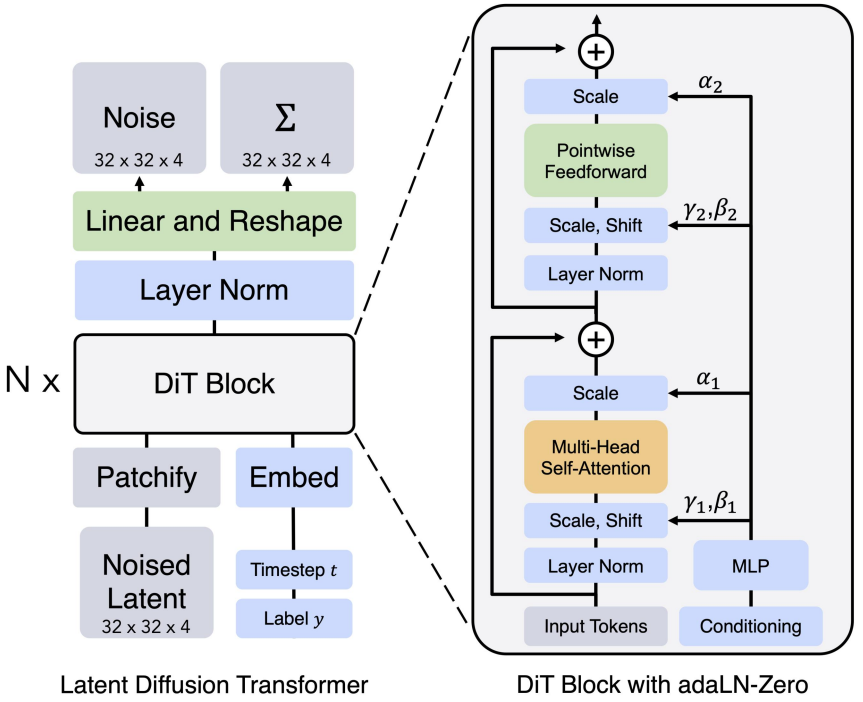
- Denoising Space：Pixel diffusion (original input) v.s. Latent space diffusion



Architecture of Diffusion Models

- Denoising Network：U-Net v.s. Transformer



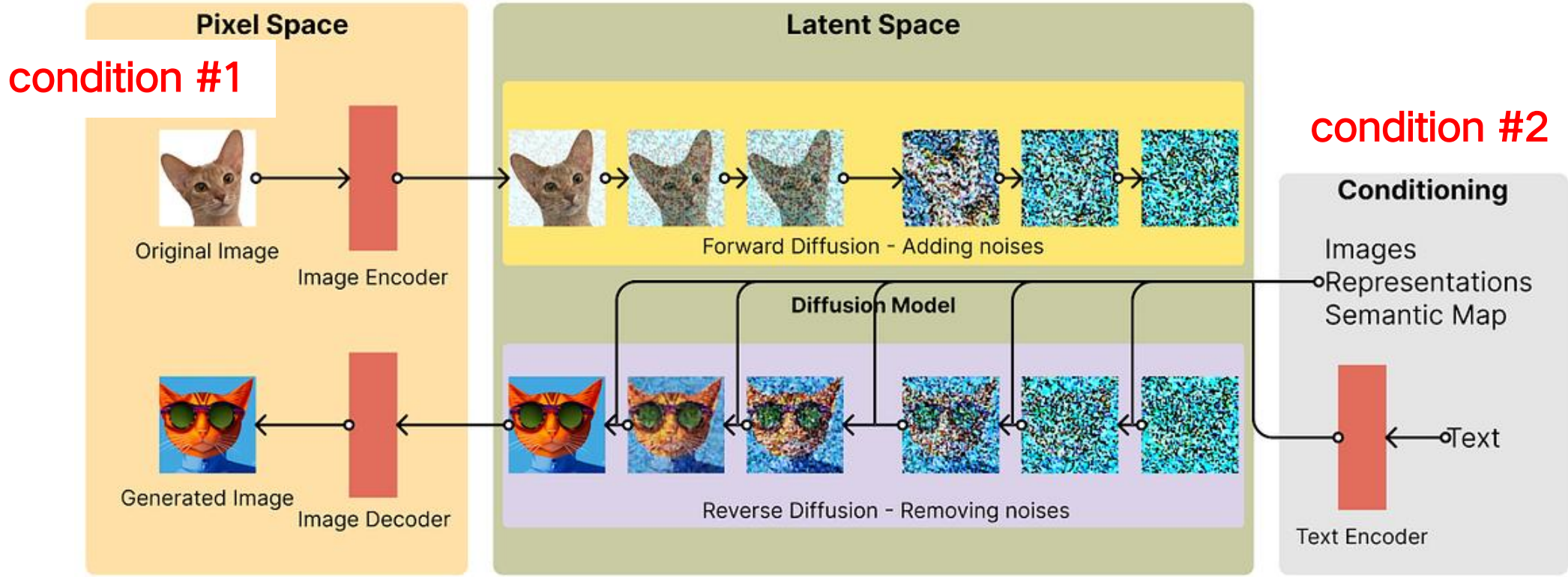


Latent Diffusion Transformer

DiT Block with adaLN-Zero

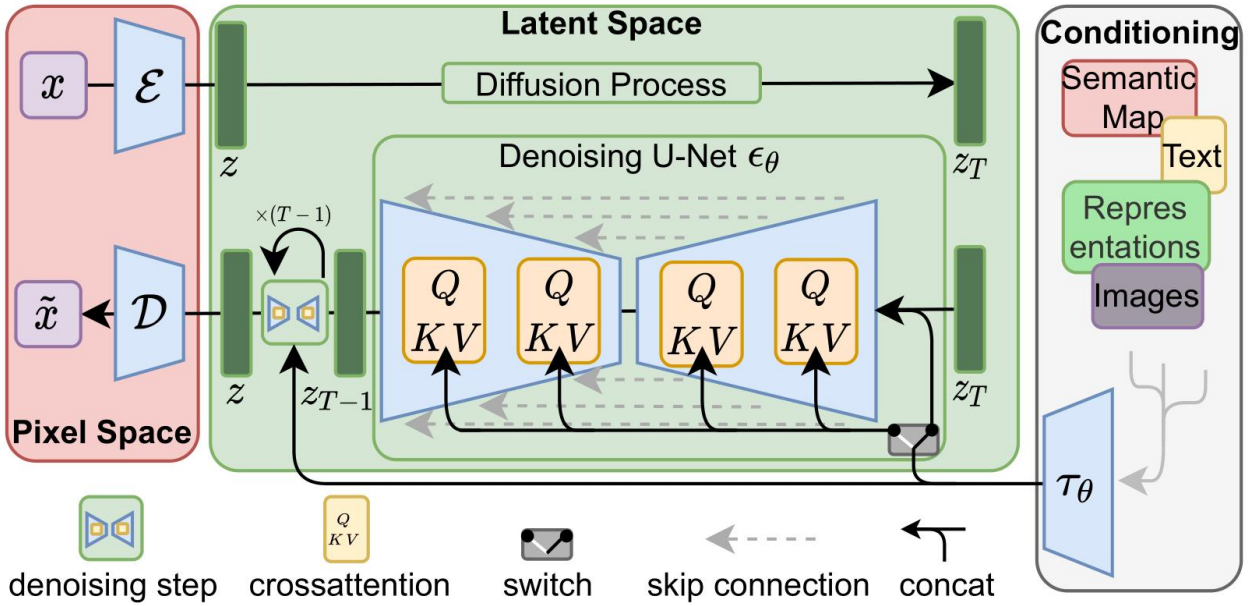
Architecture of Diffusion Models

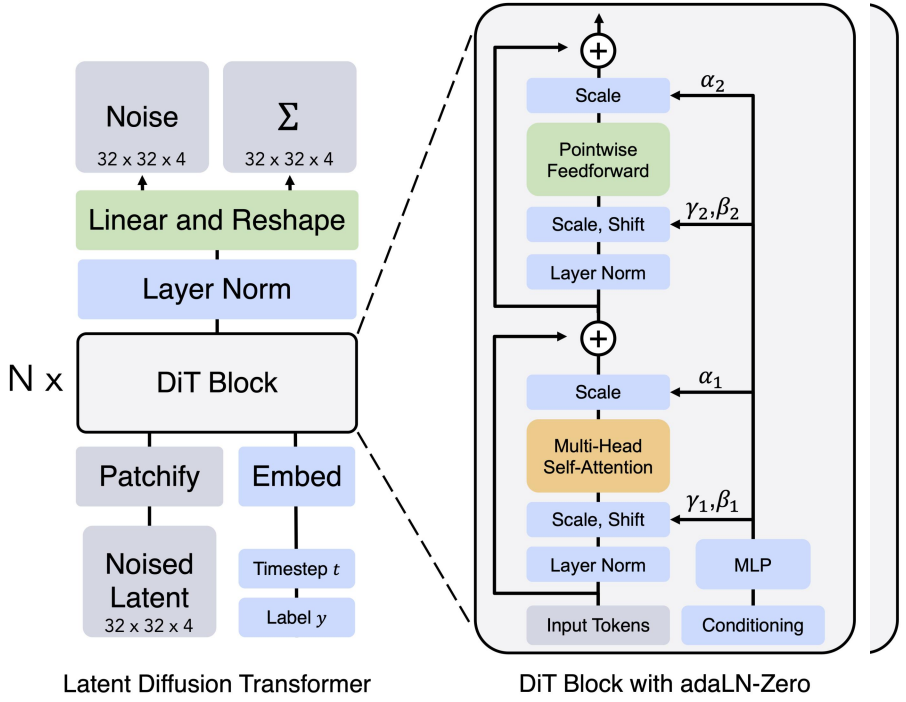
- Conditioning
  - Text, Appearance, Geometry, Motion



Datasets of Diffusion Models:

- U-Net (CNN): Structured matrix-like inputs, such as 2D images and 3D video.
- DiT (Transformer): Tokenized structured inputs .

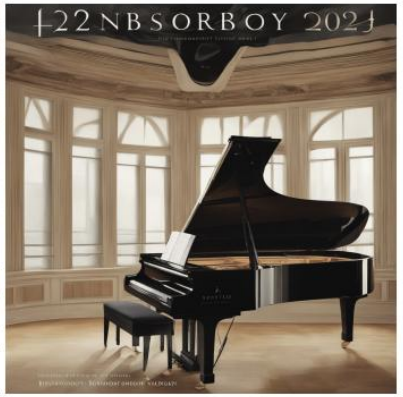




Image+ Diffusion Models

- LDM, Stable Diffusion, DALL-E, GigaGAN
- High-Resolution Image Synthesis with Latent Diffusion Models

Last Year, e.g. SDXL

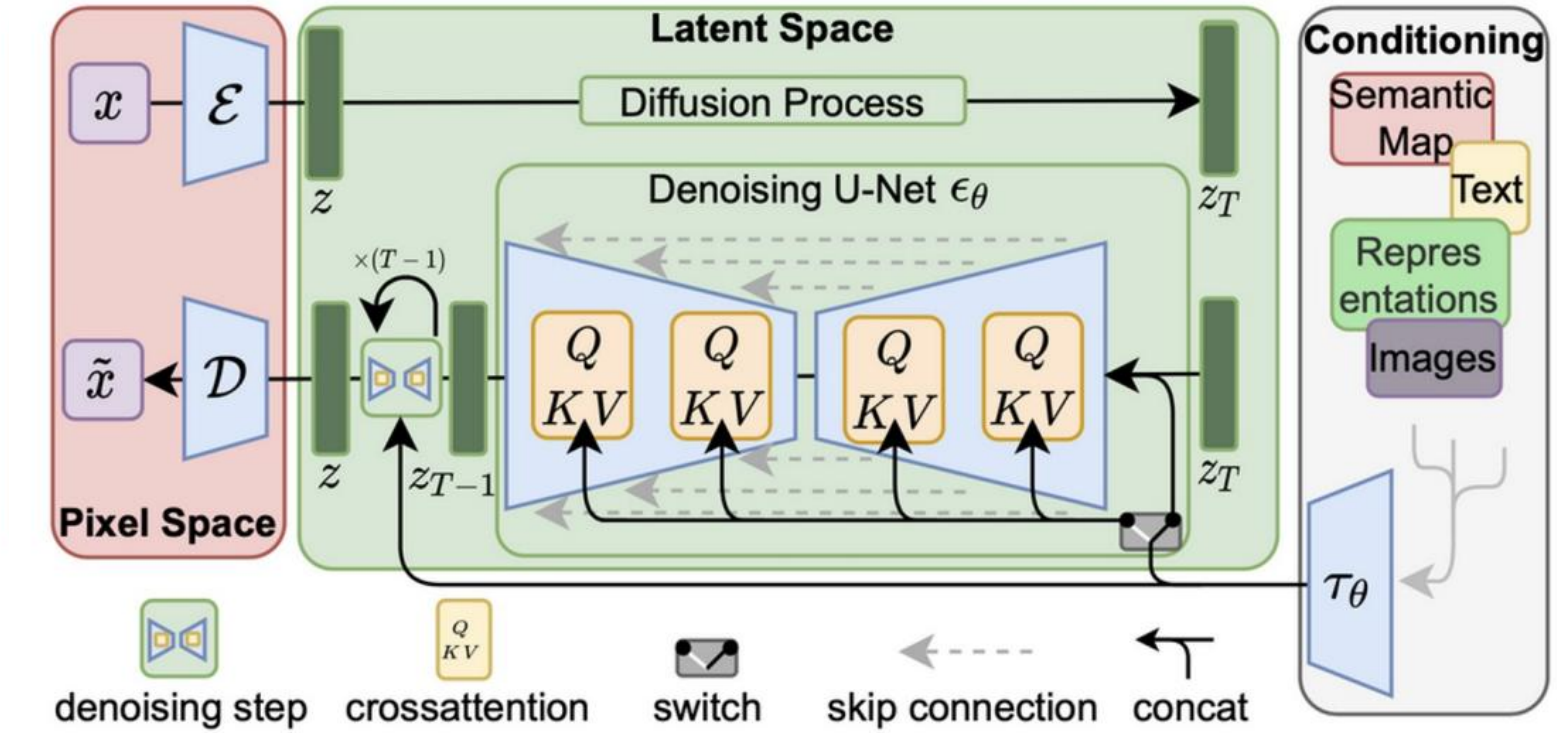



Today



Image+ Diffusion Models

- LDM: High-Resolution Image Synthesis with Latent Diffusion
  - Variational autoencoder (VAE)
  - Condition (text) encoder
  - Conditional denoising U-Net



Image+ Diffusion Models

- Recaption (DALL-E 3)
  - Data: the importance of re-caption and text encoder (T5)
  - Less noisy; more detailed

Image			
Alt Text	now at victorian plumbing.co.uk	is he finished...just about!	23 (19 of 30) 1200
SSC	a white modern bathtub sits on a wooden floor.	a quilt with an iron on it.	a jar of rhubarb liqueur sitting on a pebble background.
DSC	this luxurious bathroom features a modern freestanding bathtub in a crisp white finish. the tub sits against a wooden accent wall with glass-like panels, creating a serene and relaxing ambiance. three pendant light fixtures hang above the tub, adding a touch of sophistication. a large window with a wooden panel provides natural light, while a potted plant adds a touch of greenery. the freestanding bathtub stands out as a statement piece in this contemporary bathroom.	a quilt is laid out on a ironing board with an iron resting on top. the quilt has a patchwork design with pastel-colored strips of fabric and floral patterns. the iron is turned on and the tip is resting on top of one of the strips. the quilt appears to be in the process of being pressed, as the steam from the iron is visible on the surface. the quilt has a vintage feel and the colors are yellow, blue, and white, giving it an antique look.	rhubarb pieces in a glass jar, waiting to be pickled. the colors of the rhubarb range from bright red to pale green, creating a beautiful contrast. the jar is sitting on a gravel background, giving a rustic feel to the image.

User Input: A horse riding an astronaut
Expanded Prompt: Render of a humorous setting where a white horse, looking a bit puzzled, is sitting on top of an astronaut's back. The astronaut tries to balance the horse while surrounded by asteroids and space debris.

User Input: A woman stands wearing a black and white coat over a dark vest, orange shirt and pink tie.
Expanded Prompt: Photo of a confident woman with short brunette hair standing against a neutral background. She is wearing a black and white striped coat that reaches her knees. Underneath the coat, she has a dark vest and a bright orange shirt. Around her neck, she has tied a vibrant pink tie which stands out against the other colors.

Audio/Video + Diffusion Models： Sora

- Audio Diffusion Models, VideoDiffusion, Sora

- Consistency: consistency in 3D rendering, long-range coherence, and object permanence.
- High fidelity.
- Surprising length: extended video length capability (Sora: 1 minute vs. previous systems: seconds).
- Flexible resolution: generation of videos across various durations, aspect ratios, and resolutions.

Audio/Video + Diffusion Models： Sora

- Audio Diffusion Models, VideoDiffusion, Sora

- The **DiT** framework by Meta (2022.12) is designed for video processing.
- Google's **MAGViT** (2022.12) focuses on Video Tokenization.
- Google DeepMind introduced **NaViT** (2023.07) to support various resolutions and aspect ratios.
- OpenAI's **DALL-E 3** (2023.09) enhances Video Caption generation for improved conditioned video creation.

Audio/Video + Diffusion Models： Sora

- Audio Diffusion Models, VideoDiffusion, Sora

- Given a Sora demo (the walking woman in the Tokyo street), the key elements of a physical world, in the graphical way...

- Appearance
- Geometry
- Lighting
- Motion & Animation
- Audio

Audio/Video + Diffusion Models： Sora

- Sora failure case in geometry and appearance.

Audio/Video + Diffusion Models： Sora

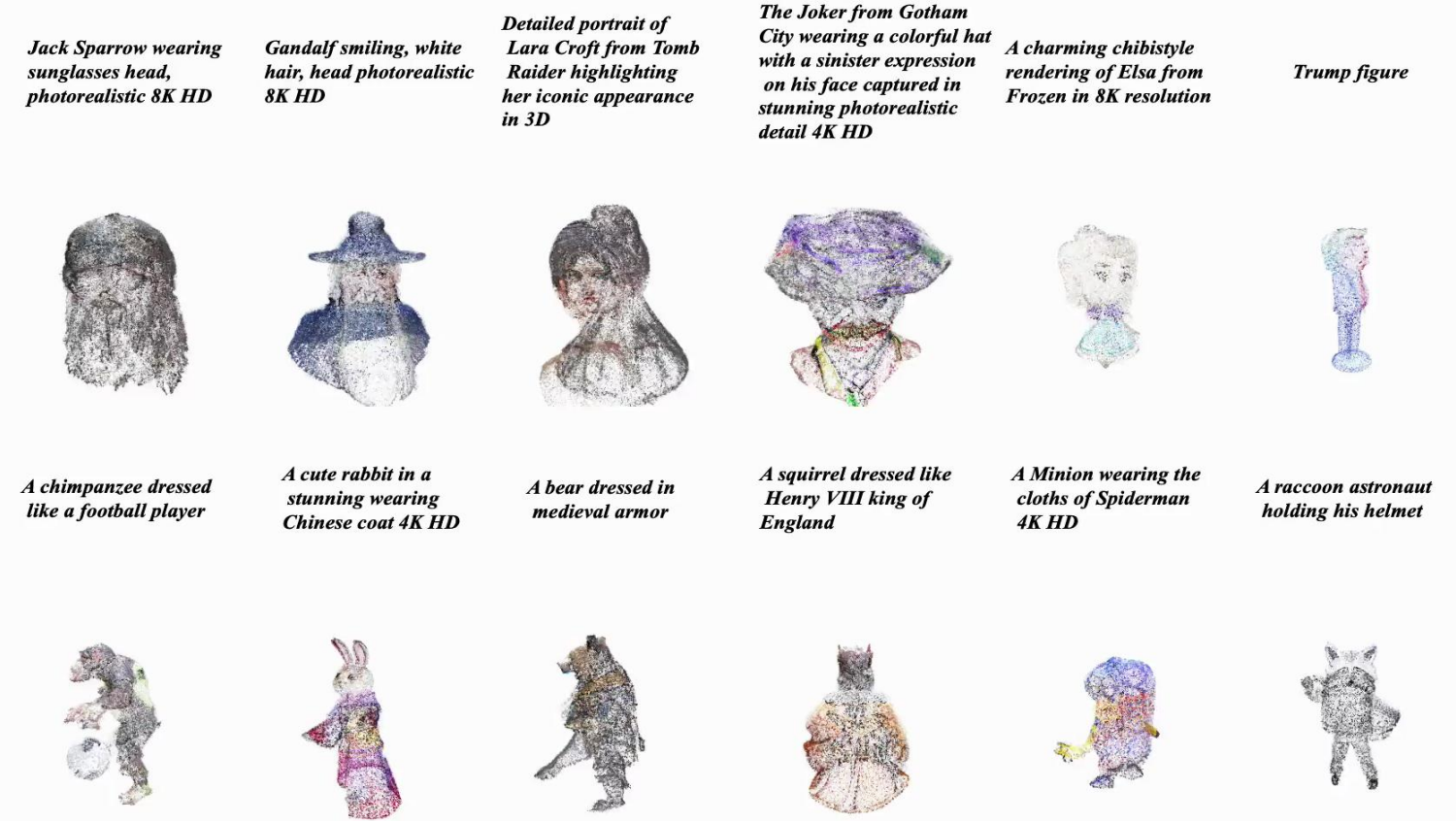
- Sora failure case in lighting.

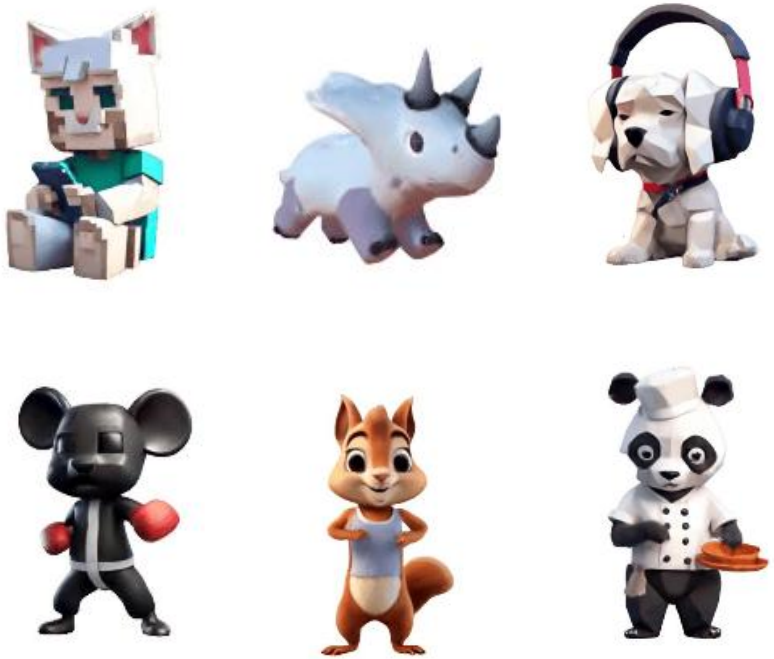
Audio/Video + Diffusion Models： Sora

- Sora failure case in motion and animation.

3D + Diffusion Models

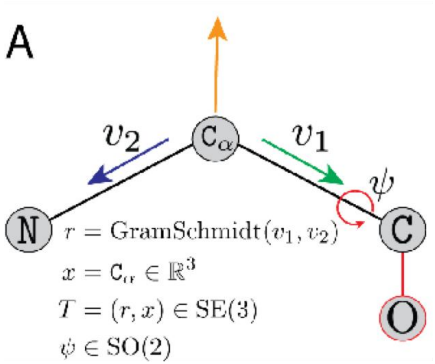
- NeRF, 3D-VAE, DreamFusion, Stag4D

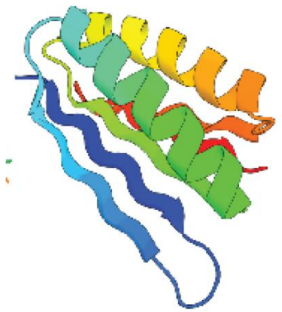




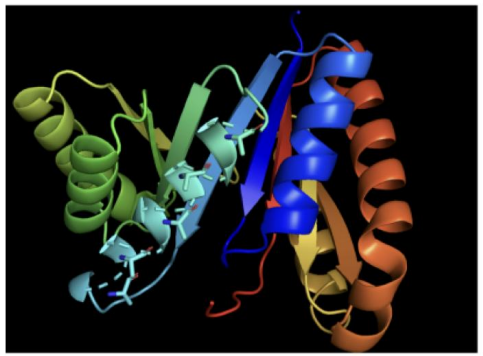
Biology (Structure) + Diffusion Models

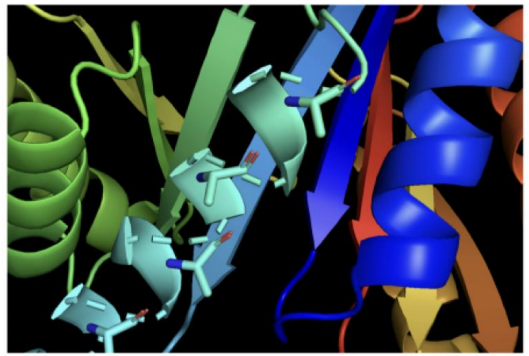
- AlphaFold3, ESMFold, RFdiffusion, SE(3) Diffusion

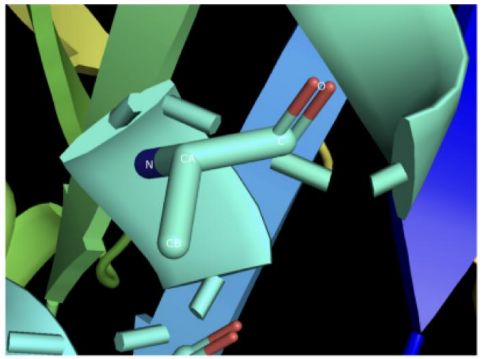




Sequence ⇨ Structure ⇨ Function



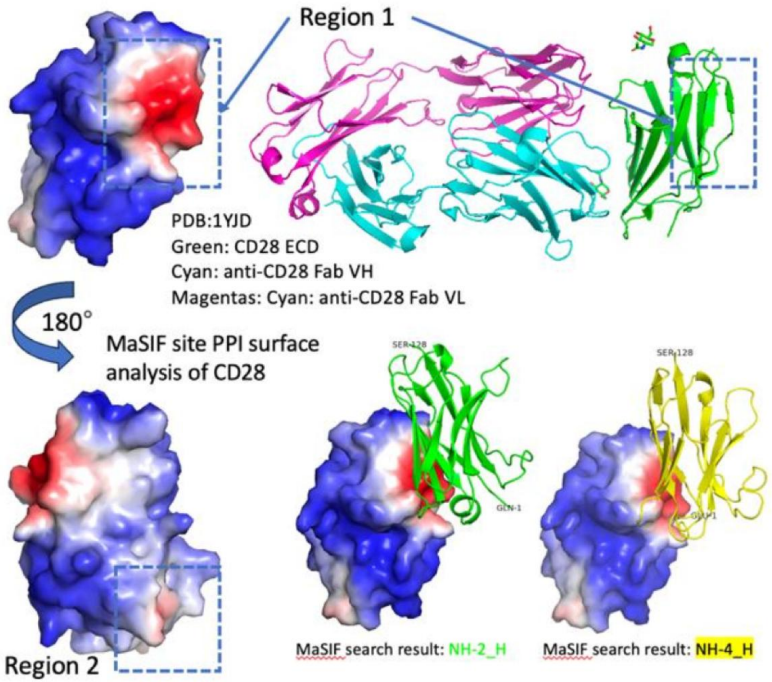




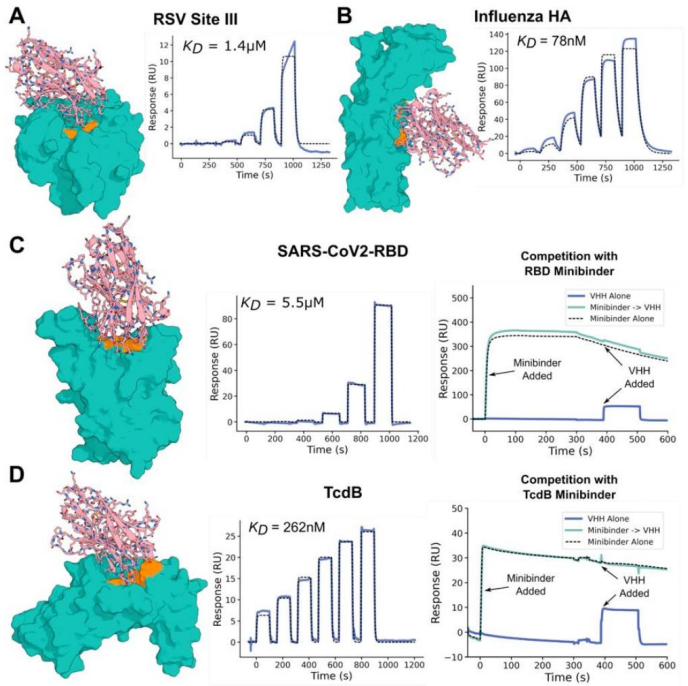
Biology (Structure) + Diffusion Models

- AlphaFold3, ESMFold, RFdiffusion, SE(3) Diffusion

CD28 antigen-antibody generation
MaSIF

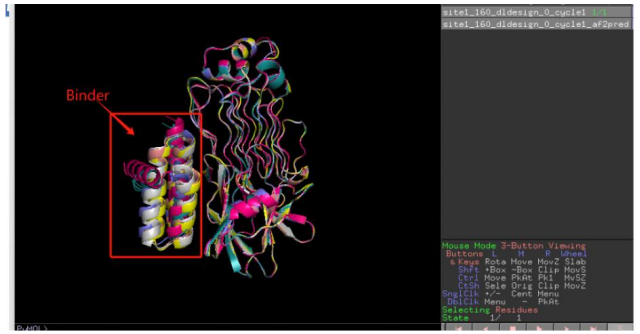


De Nove Binder
By David Baker teams

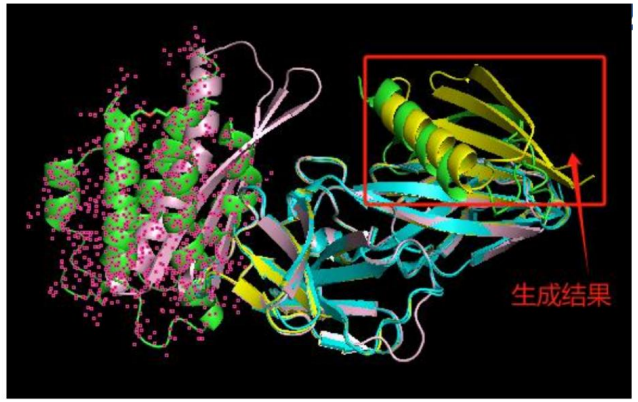


Bennett NR, et al. Atomically accurate de novo design of single-domain antibodies. bioRxiv [Preprint]. 2024

De Nove Binder IL-7A

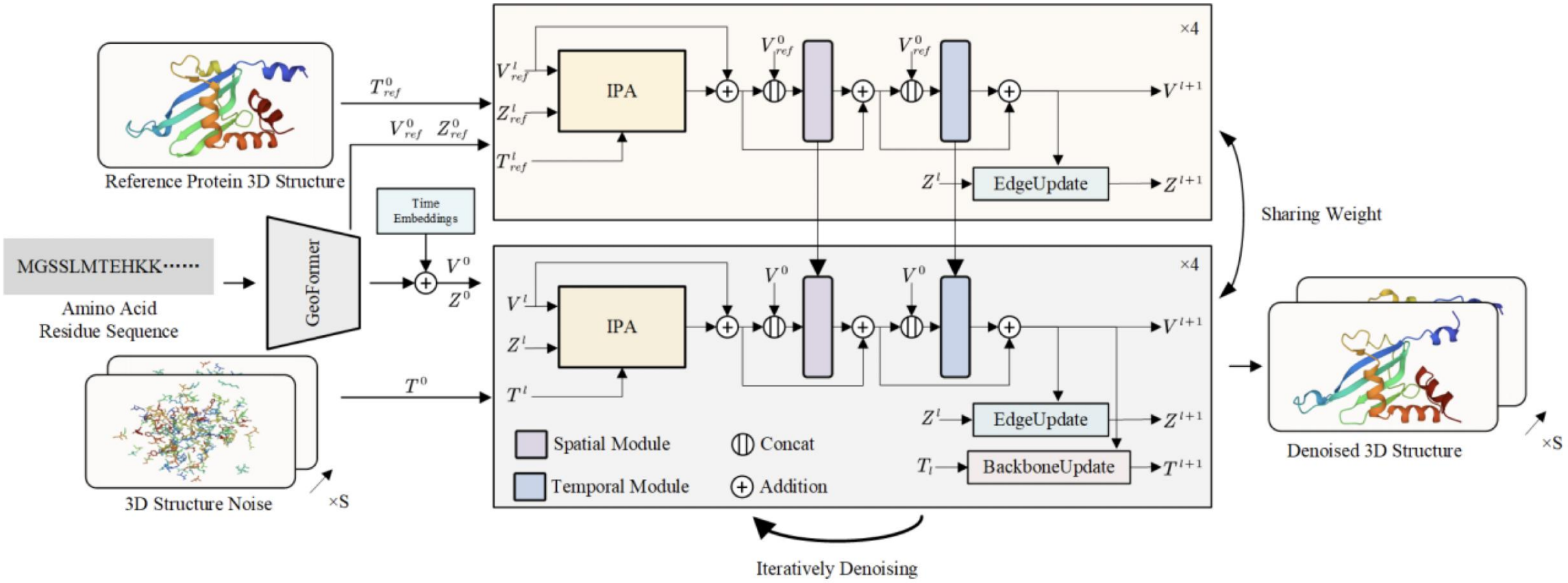


De Nove Binder InsulinR



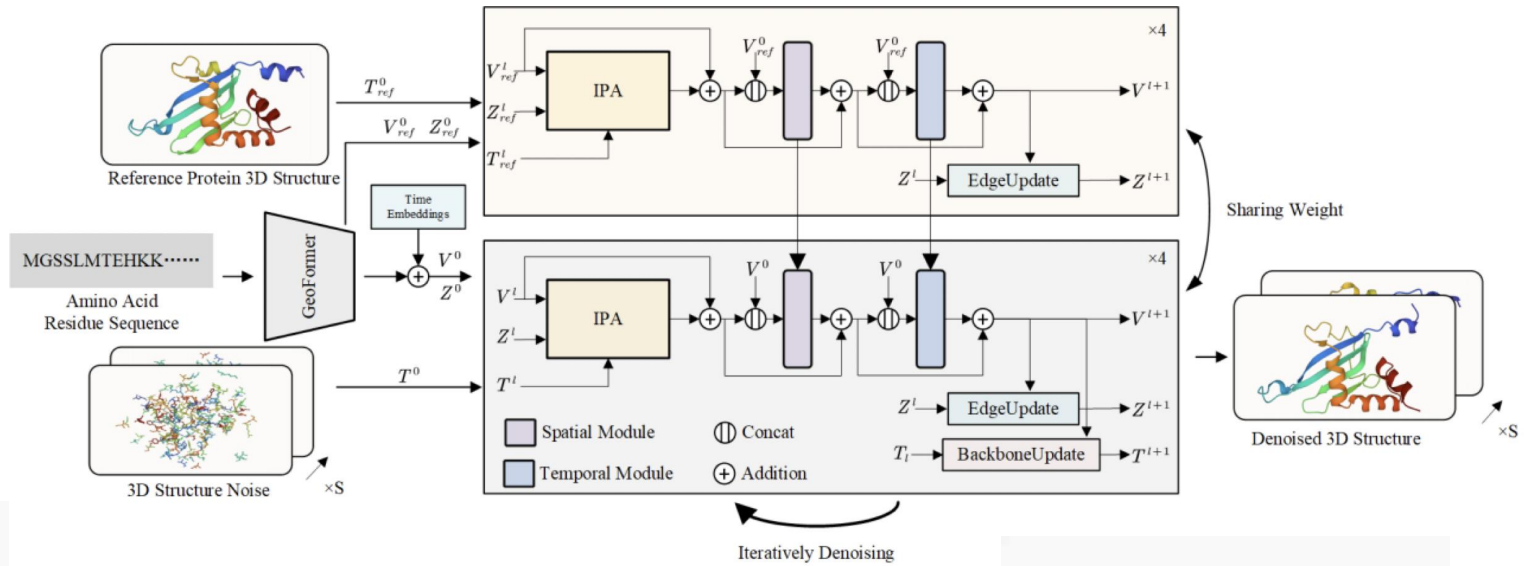
Biology (Structure) + Diffusion Models

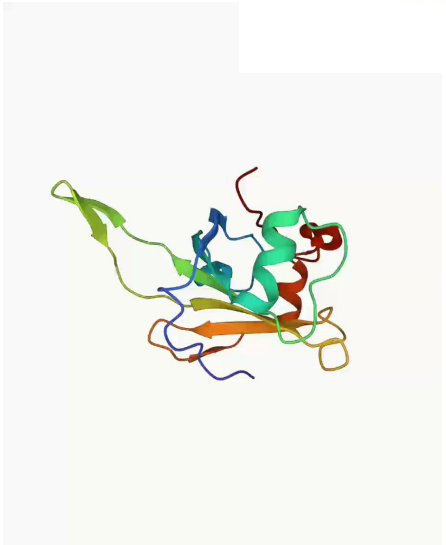
- AlphaFold3, ESMFold, RFdiffusion, SE(3) Diffusion

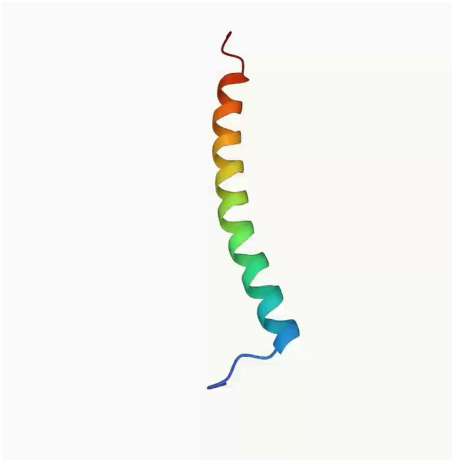


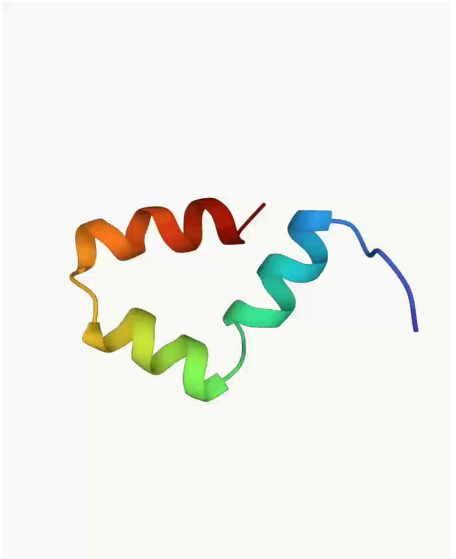
Biology (Structure) + Diffusion Models

- AlphaFold3, ESMFold, RFdiffusion, SE(3) Diffusion



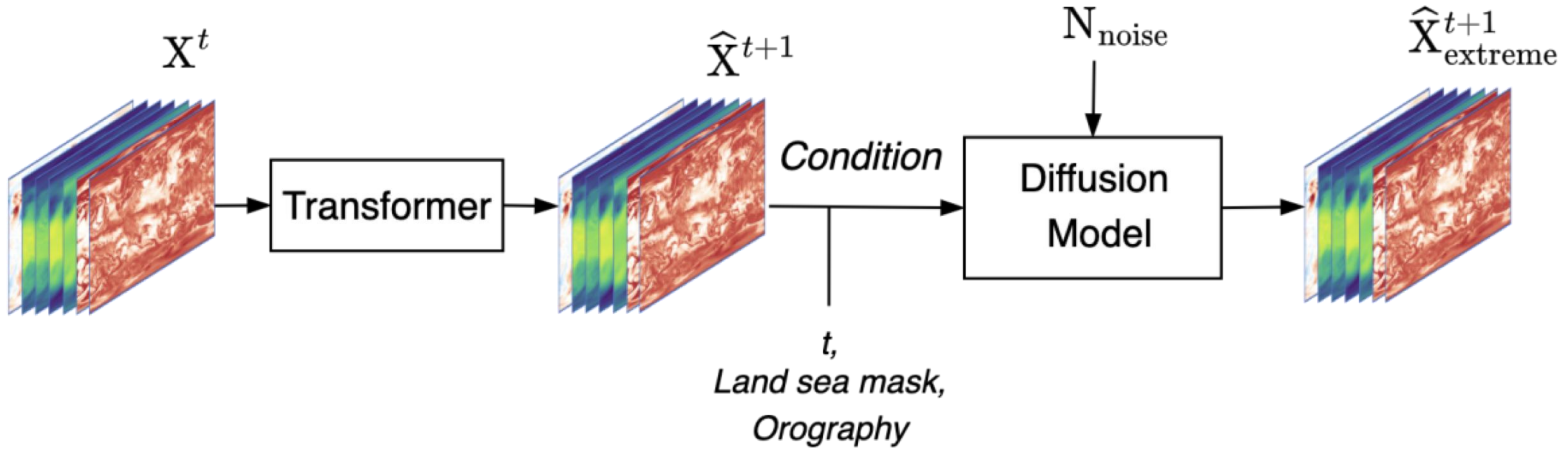


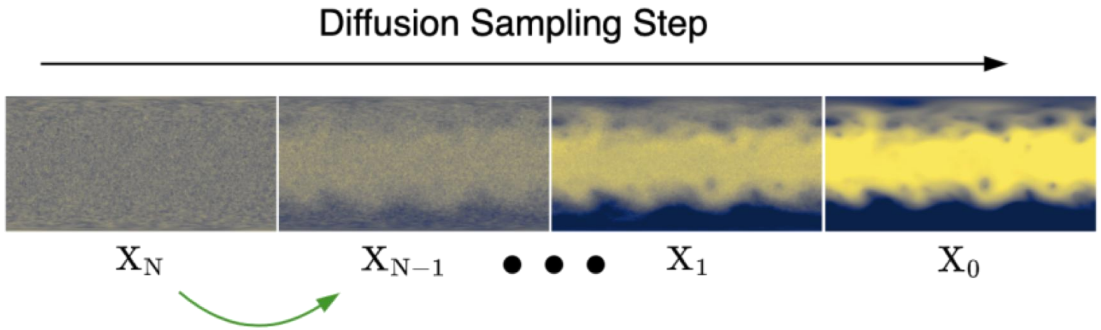


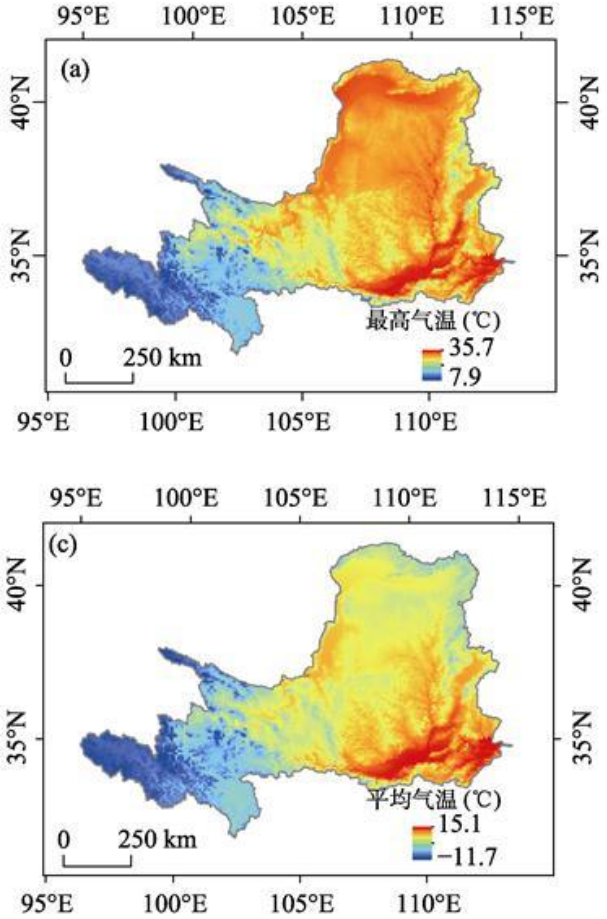


Physics and Meteorology + Diffusion Models

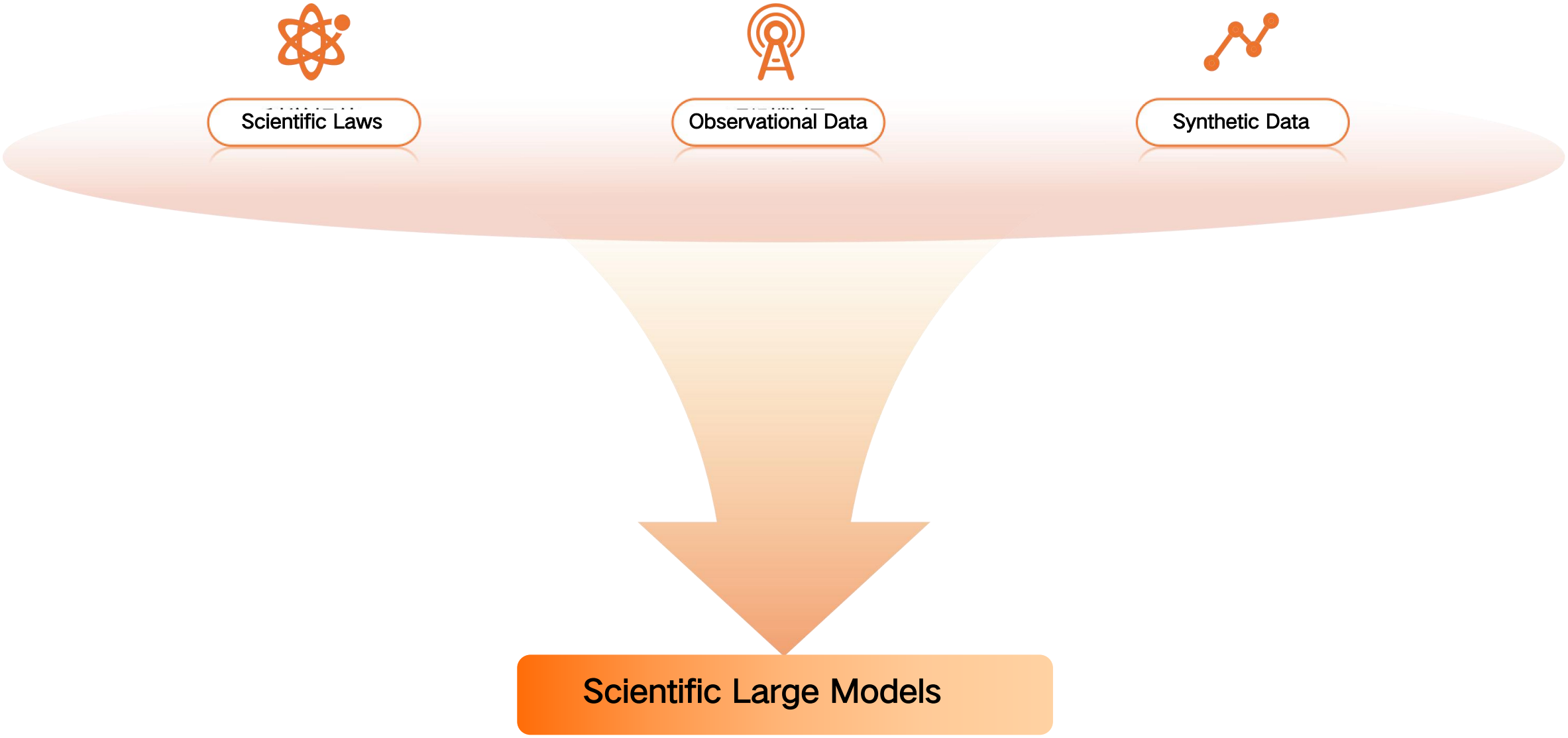
- GraphCast、Pangu-Meteorology、NowcastNet、Fuxi







Science + Diffusion Models



Thanks!