CS-461

Foundation Models and Generative AI

Generative Models II:
Diffusion Models and Beyond

Karsten Kreis and Ruiqi Gao, Guest Lecture, Fall Semester 2025/26

Agenda

What makes diffusion great?

Video diffusion models

Go beyond video gen: world models

- Controllable video generation
- Multimodal models

3D/4D generation

Agenda

What makes diffusion great?

Video diffusion models

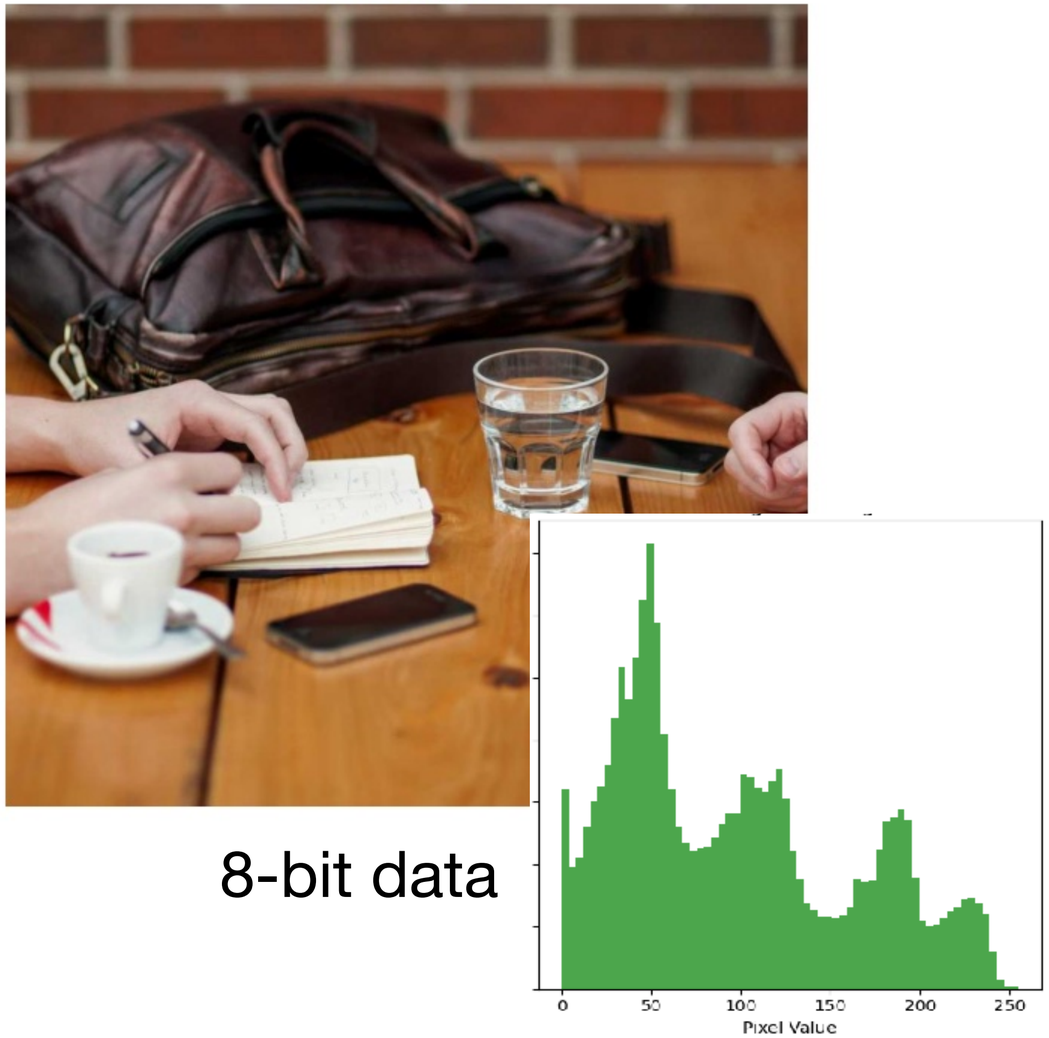
Go beyond video gen: world models

- Controllable video generation
- Multimodal models

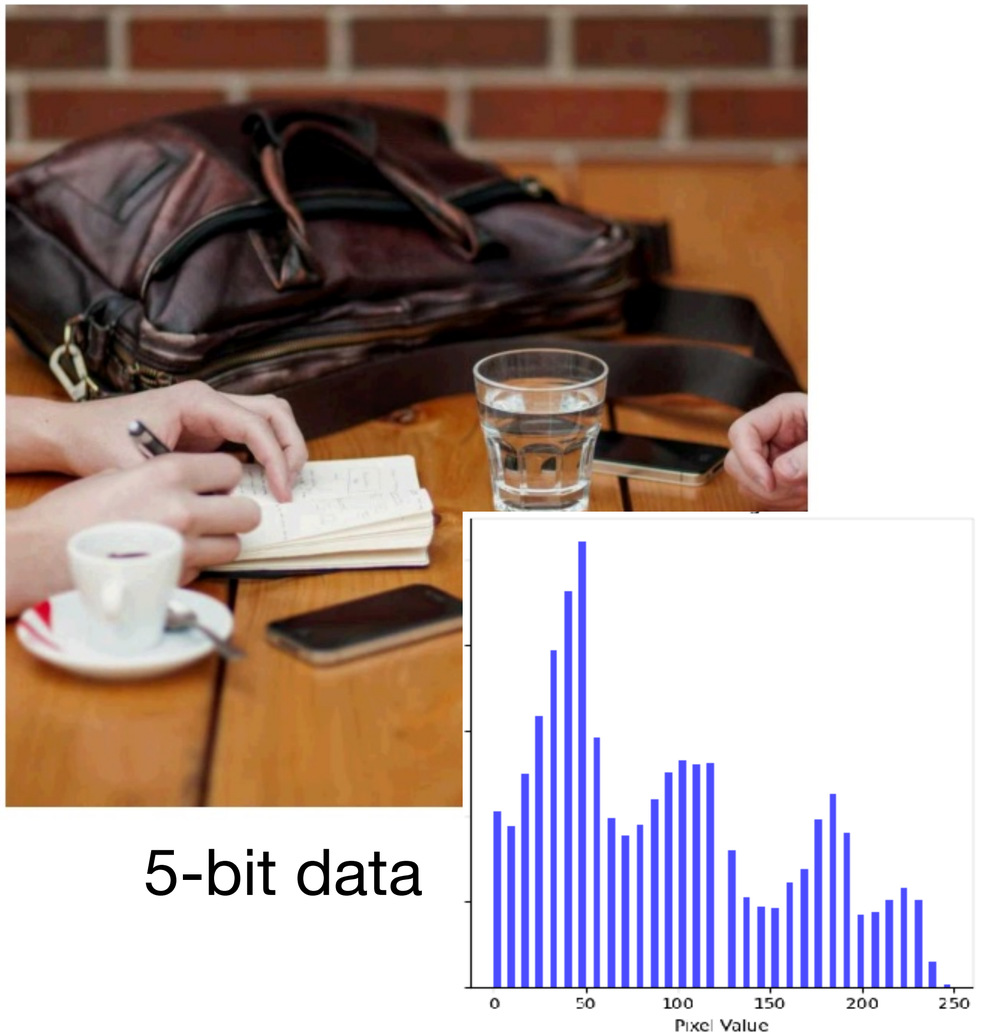
3D/4D generation

Blind spots of human perception

Can you tell the difference between the two images?



8-bit data

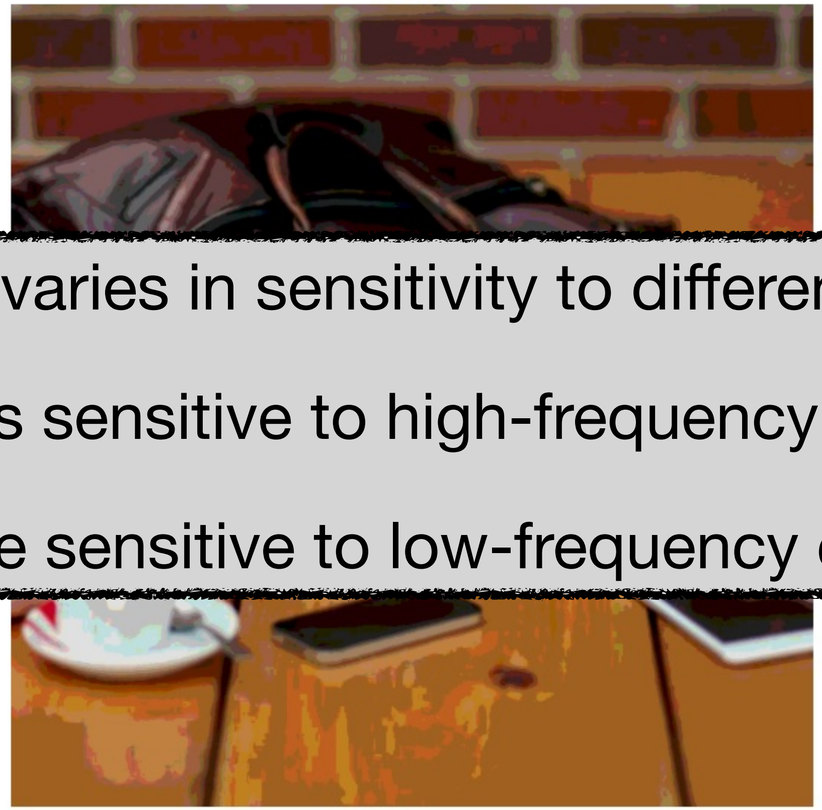


5-bit data

Continue reducing number of bits…

Can you tell the difference of the images?

5-bit data



3-bit data

1-bit data

Human perception varies in sensitivity to different bits of a visual signal

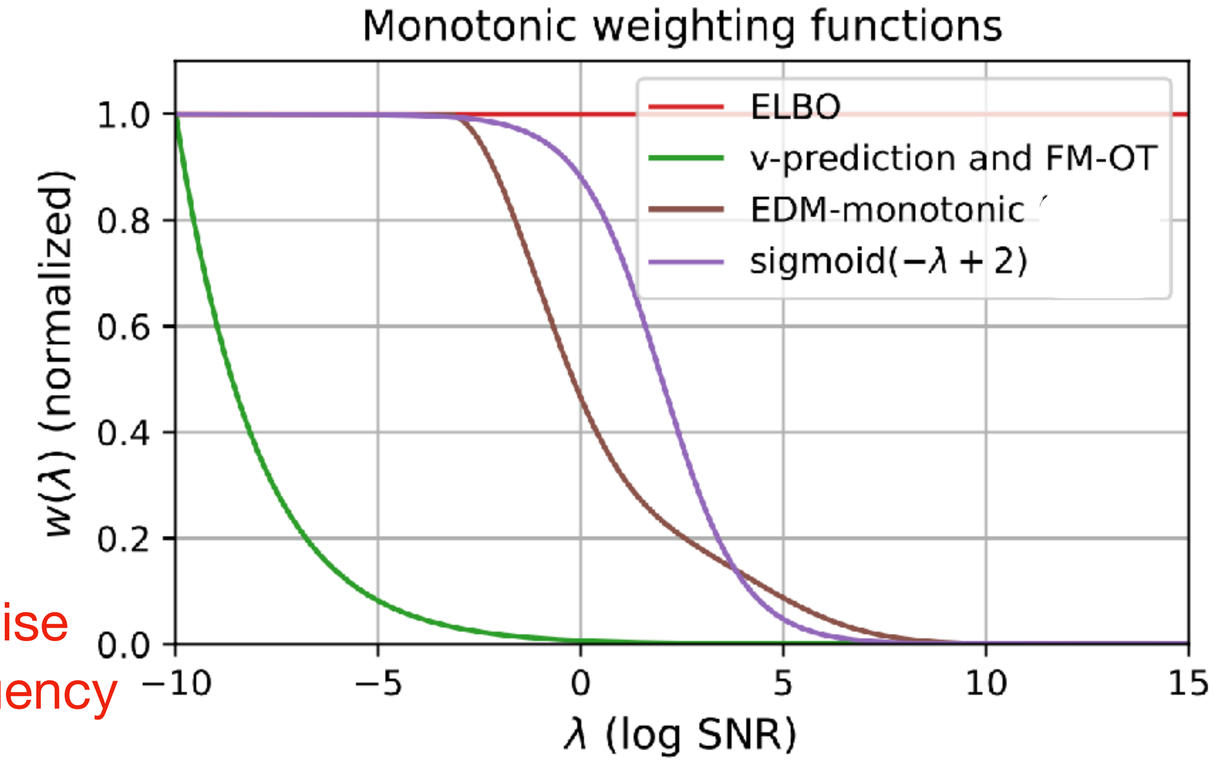
Less sensitive to high-frequency details.

More sensitive to low-frequency content.

Diffusion training objective: reweighed ELBO

Rebalancing the importance of different frequency components

$$\mathcal{L}_w(\mathbf{x}) = \frac{1}{2}\mathbb{E}_{t\sim\mathcal{U}(0,1),\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I})}\left[w(\lambda_t)\cdot -\frac{d\lambda}{dt}\cdot ||\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{z}_t;\lambda_t)-\boldsymbol{\epsilon}||_2^2\right]$$



High noise
Low frequency

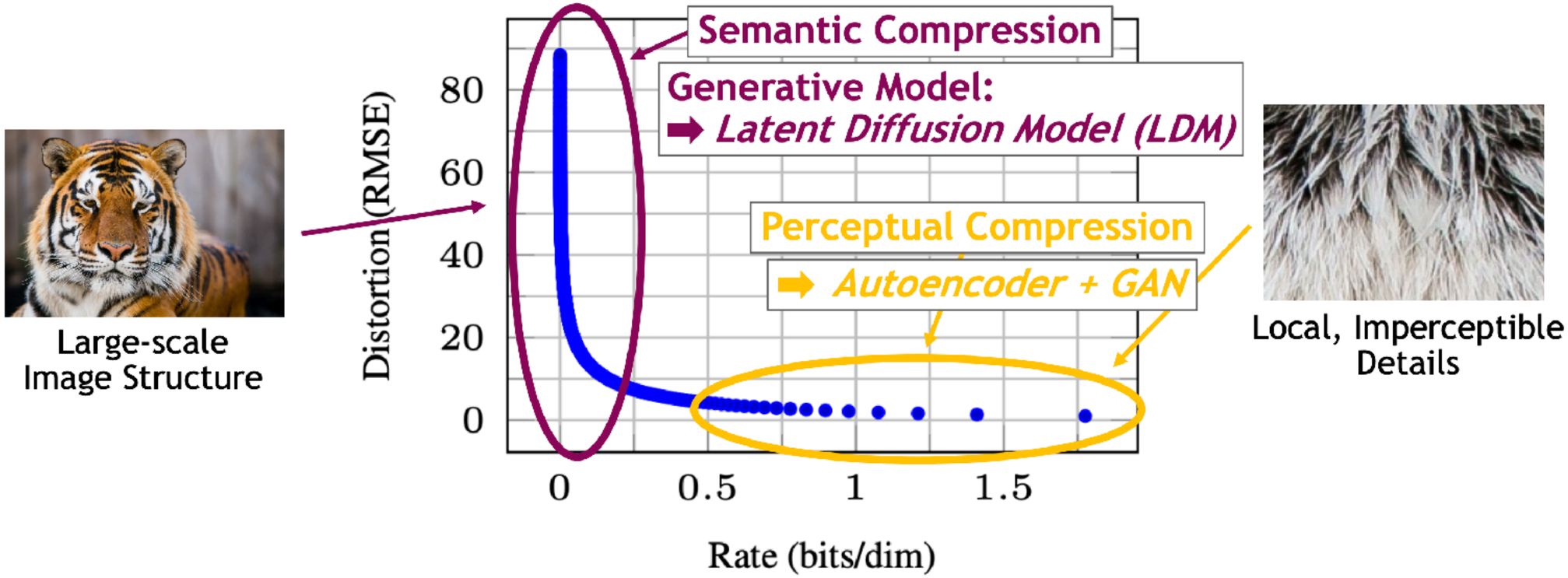
Low noise
High frequency

Key feature that makes diffusion great!

With reweighed objective, the model spends more capacity on modeling low frequency component, which is more important to human perception

Latent diffusion models

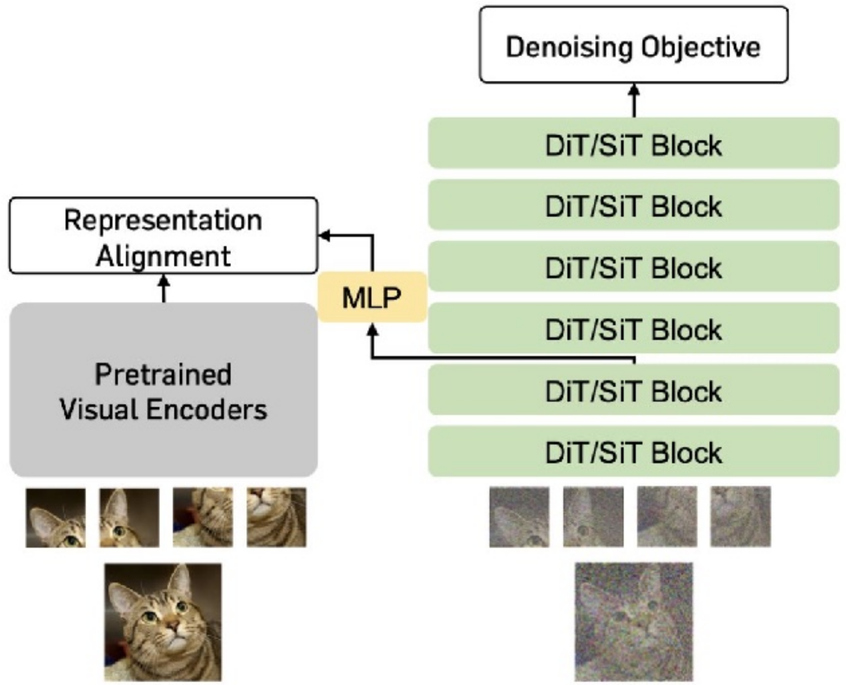
Compress the visually less important bits with an auto encoder

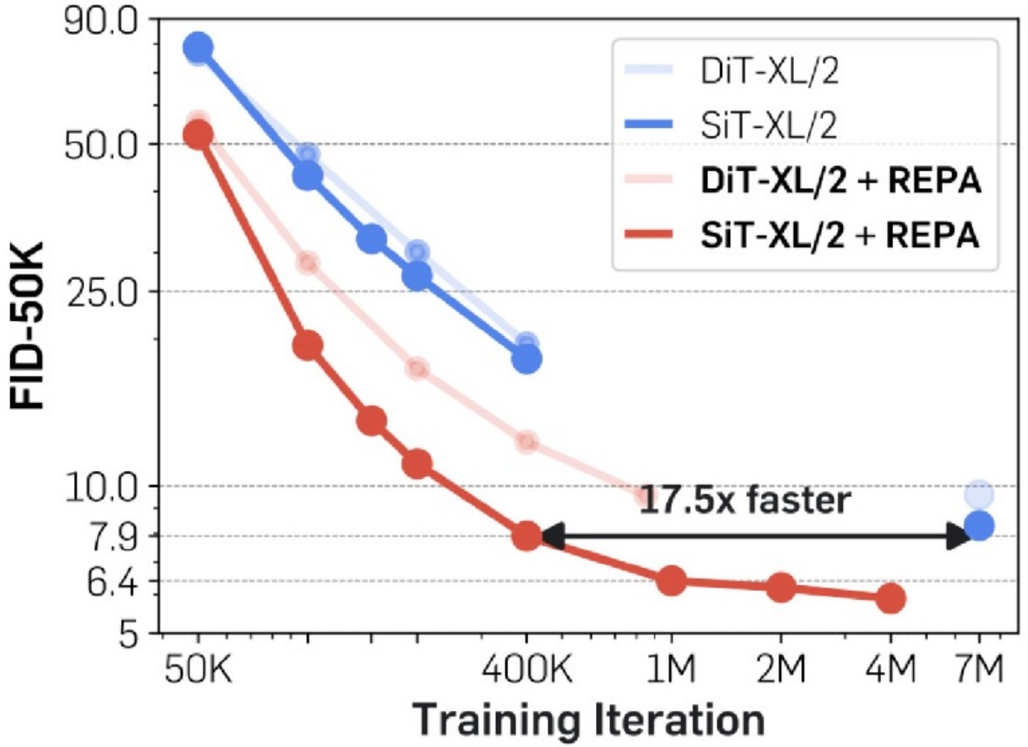


Slide from: https://neurips2023-ldm-tutorial.github.io/

Open question: other ways to leverage perceptual inductive bias?

- REPA: align diffusion model intermediate features with pretrained perceptual features → greatly increase training efficiency

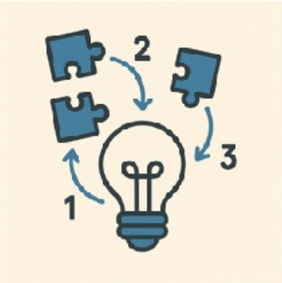




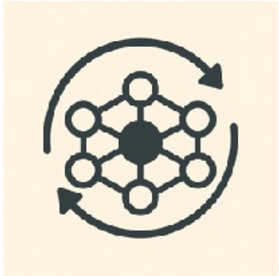
Devil in the details

More things diffusion models have done right…

Simple “regression” objective: stable for scaling up



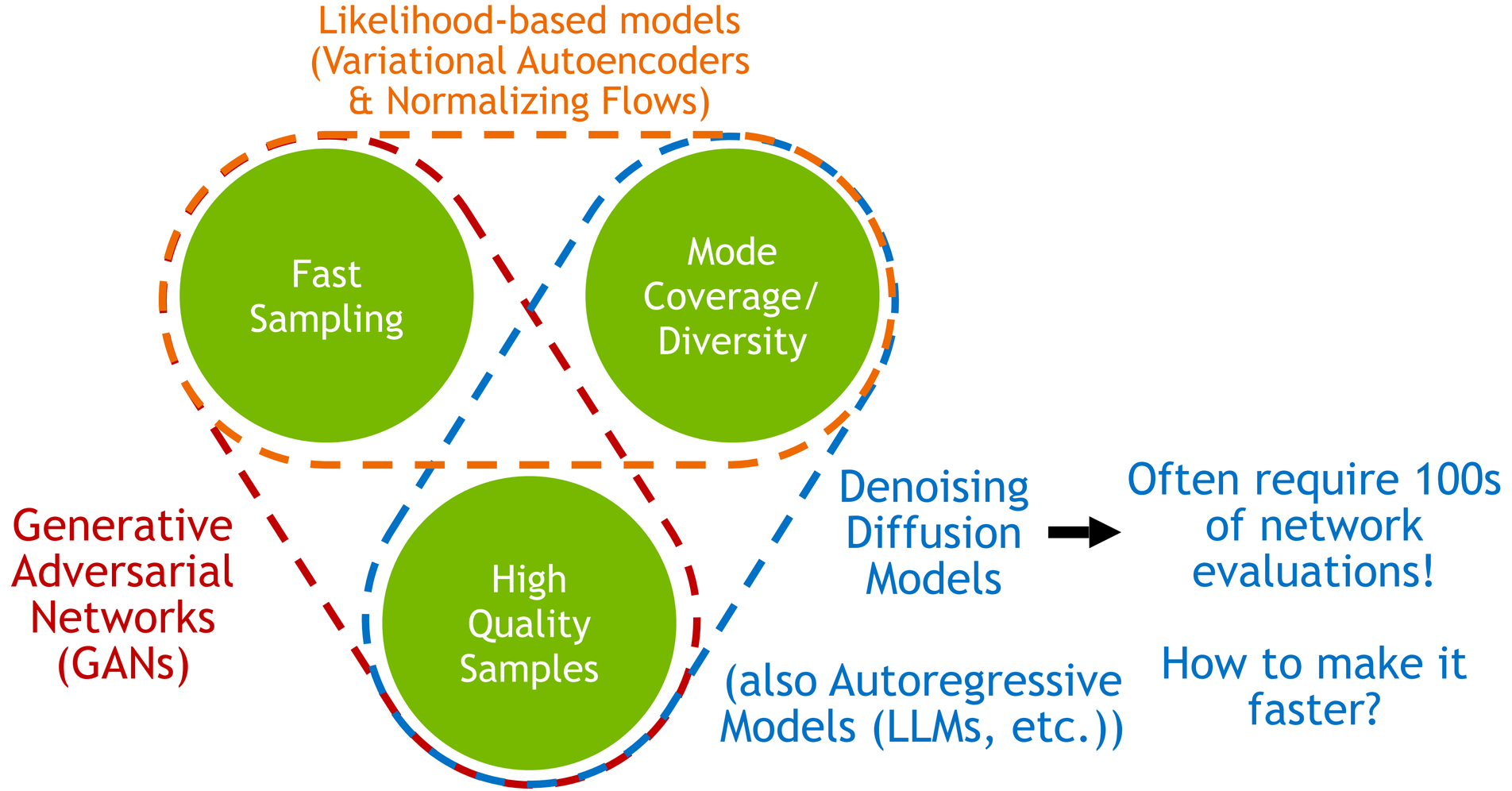
Divide and conquer: break the generation problem into small steps



Weight sharing across steps: borrowing strength from each other, data augmentation

What makes a Good Generative Model?

Generative learning trilemma

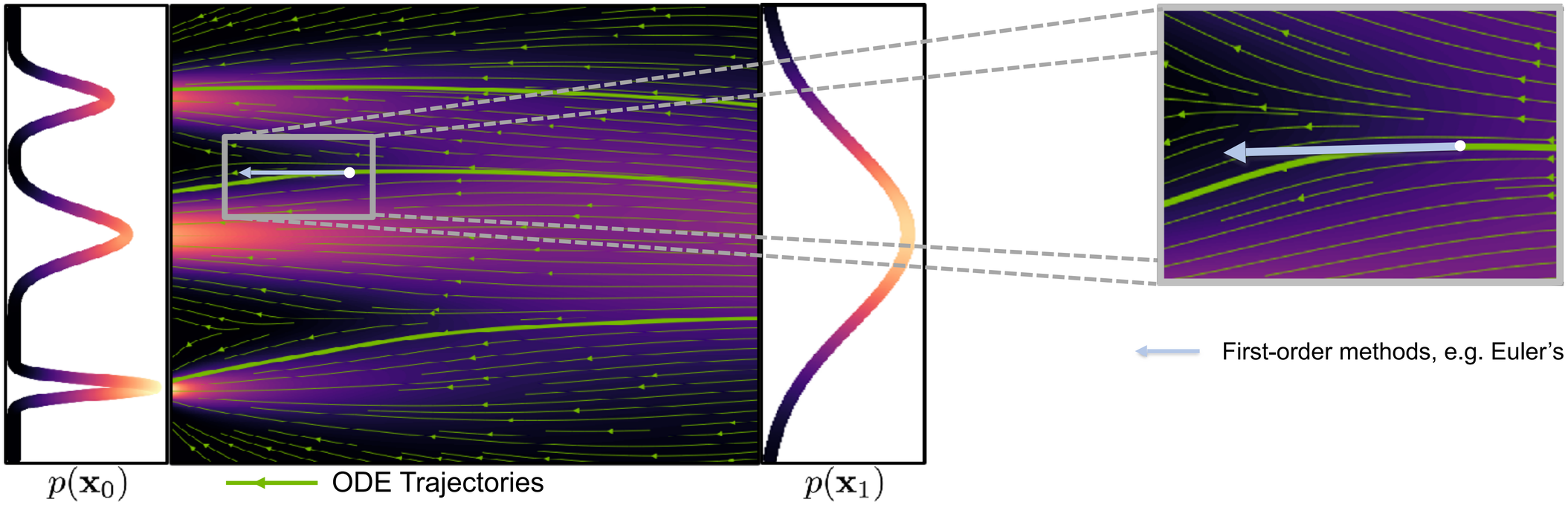


Xiao et al., ”Tackling the Generative Learning Trilemma with Denoising Diffusion GANs”, ICLR 2022

Approach 1: fast ODE/SDE solvers

Diffusion Models are slow due to their iterative sampling process!

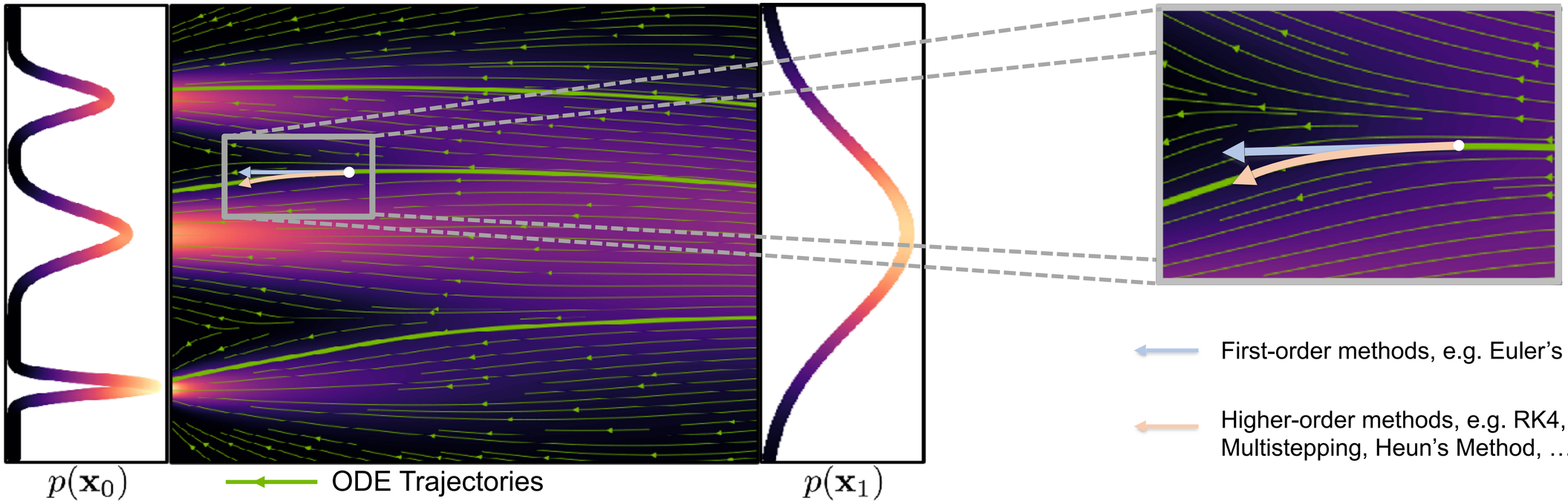
Fast samplers and solvers based on ODE/SDE literature.



Approach 1: fast ODE/SDE solvers

Diffusion Models are slow due to their iterative sampling process!

Fast samplers and solvers based on ODE/SDE literature.



Approach 1: fast ODE/SDE solvers

Diffusion Models are slow due to their iterative sampling process!

Fast samplers and solvers based on ODE/SDE literature.

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

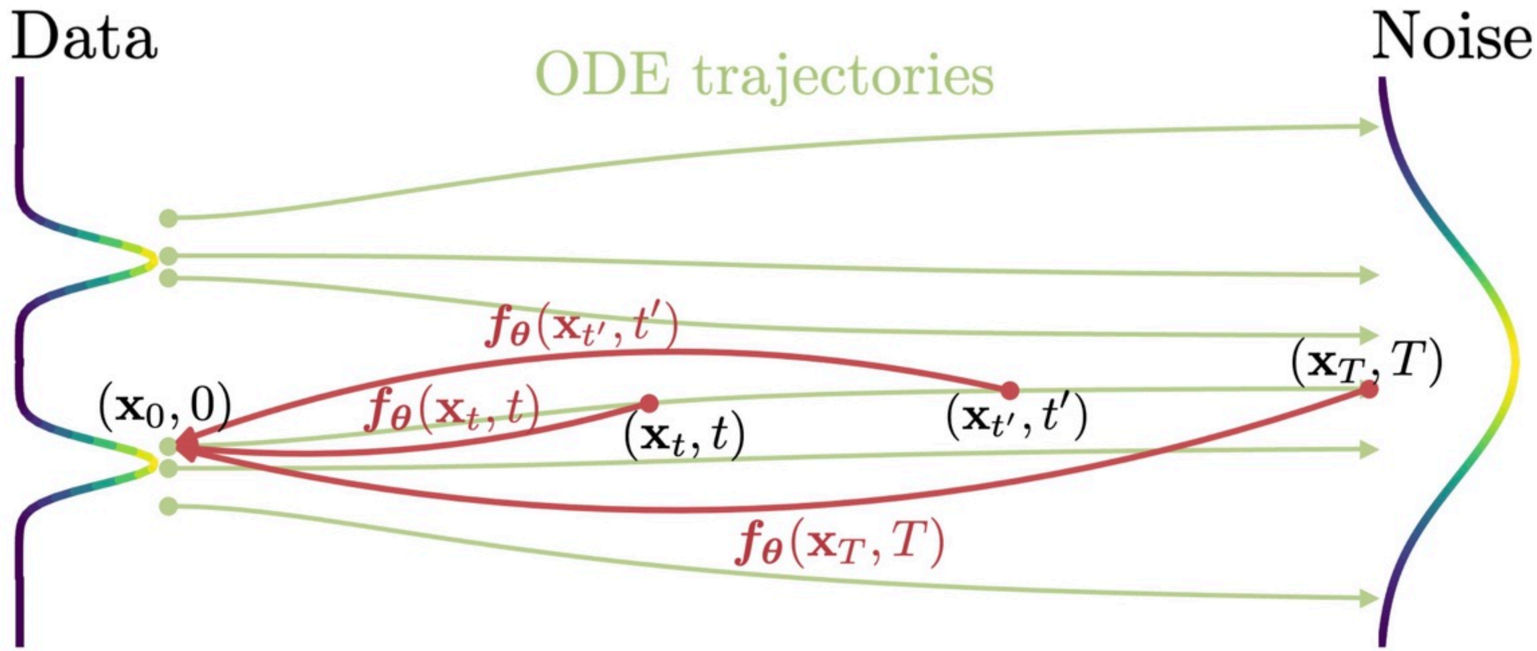
DPM-Solver++(2M)
$(N = 20)$

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

Lu et al., ”DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models”, 2022

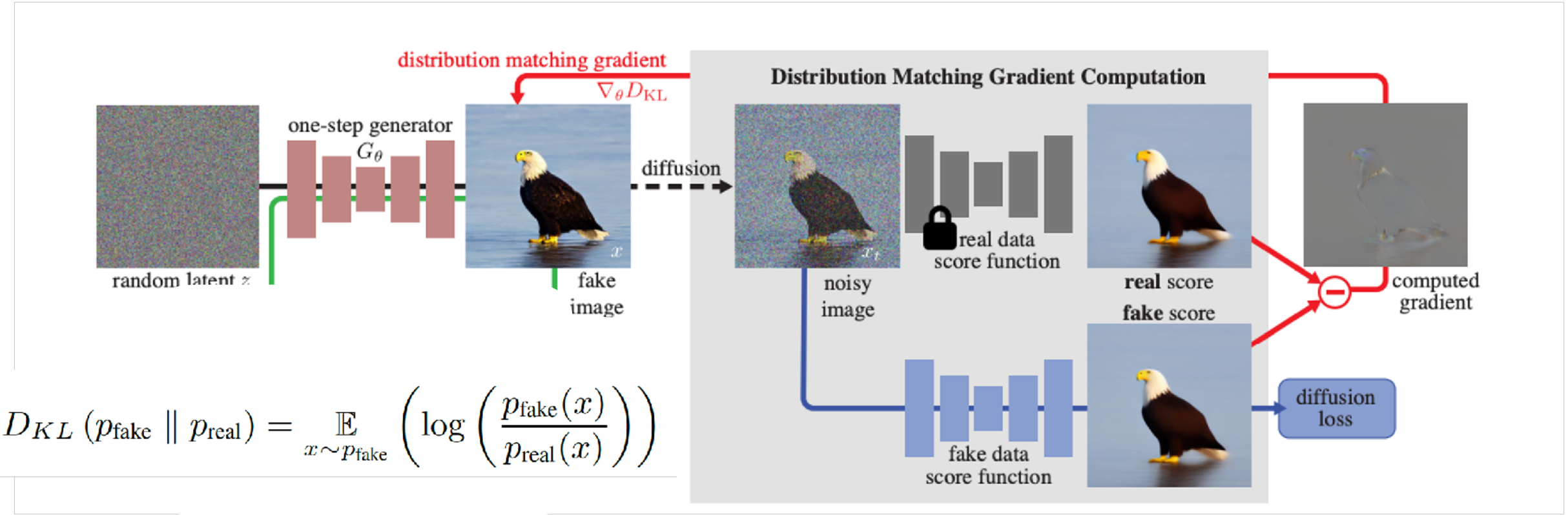
Approach 2: trajectory distillation

- Learn a function that reproduces multi-step mapping along the sampling trajectory with one-step. E.g. progressive distillation, consistency model.
- Pros: one-to-one mapping to the teacher model; cons: hard to reduce to very few sampling steps.



Approach 3: variational distillation

- Matching the distribution of the student model with the teacher model via a variational objective.
- Pros: one-step generation, very fast. Cons: mode collapse in distribution matching. Extra compute cost.



$$D_{KL}\left(p_{\text{fake}} \parallel p_{\text{real}}\right) = \mathop{\mathbb{E}}_{x \sim p_{\text{fake}}} \left(\log \left(\frac{p_{\text{fake}}(x)}{p_{\text{real}}(x)} \right) \right)$$

Agenda

What makes diffusion great?

Video diffusion models

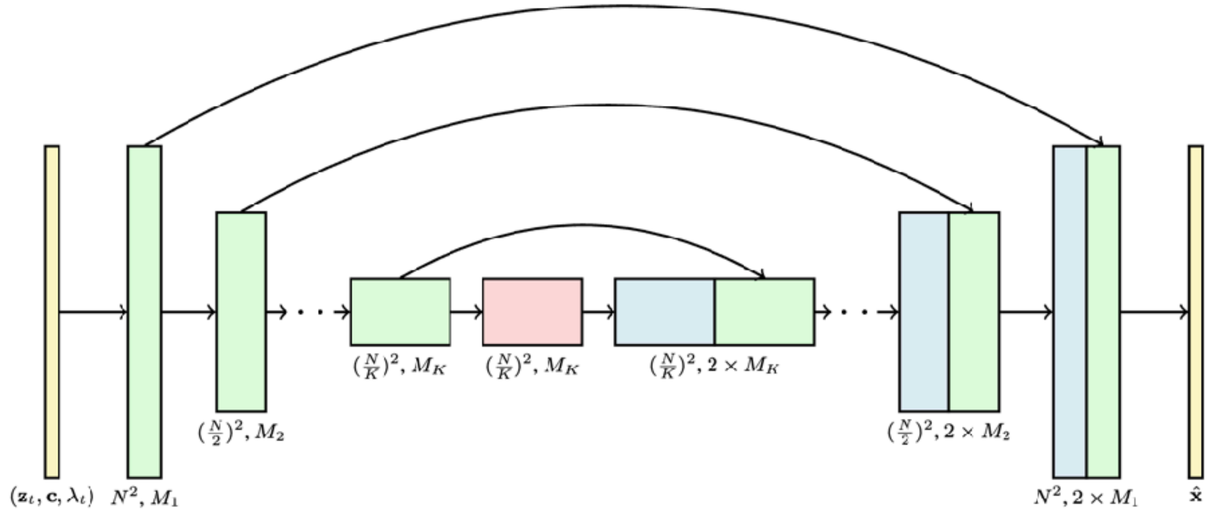
Go beyond video gen: world models

- Controllable video generation
- Multimodal models

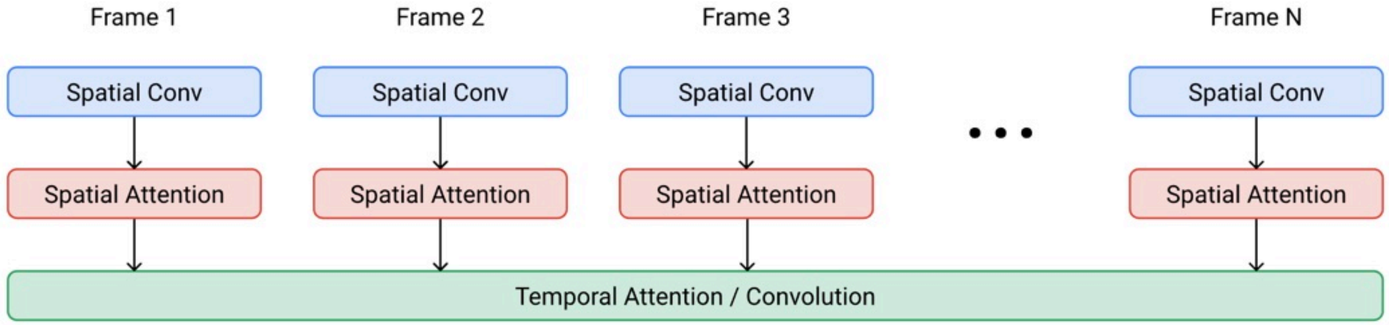
3D/4D generation

Video diffusion models

Ho et al. 2020, U-Net w/ factorized spatial-temporal attention



Samples conditioned on a single word *fireworks*.



Reference: https://arxiv.org/abs/2204.03458

Imagen Video, 2022

Cascaded framework: one base model + a few super-resolution models

Sprouts in the shape of text 'Imagen' coming out of a fairytale book.

Base → 

TSR → 

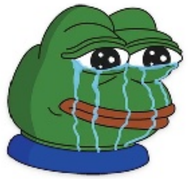
SSR → 

SSR → 

TSR → 

TSR → 

SSR → 



Too many models to train/tune

Error accumulation

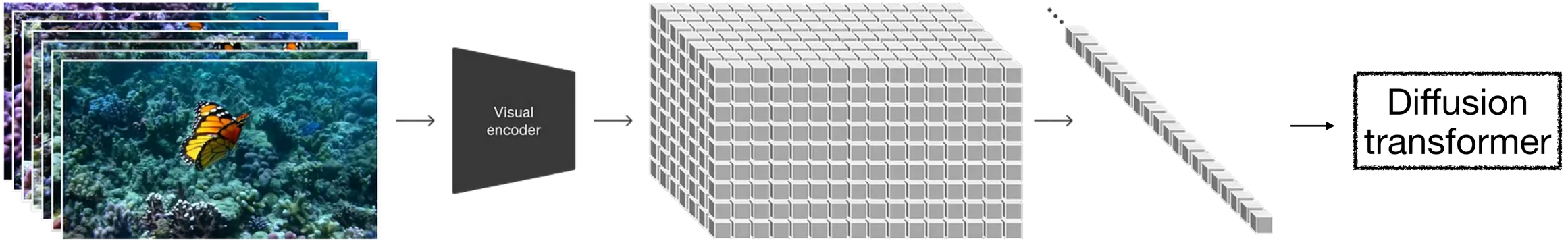
Serving models is expensive

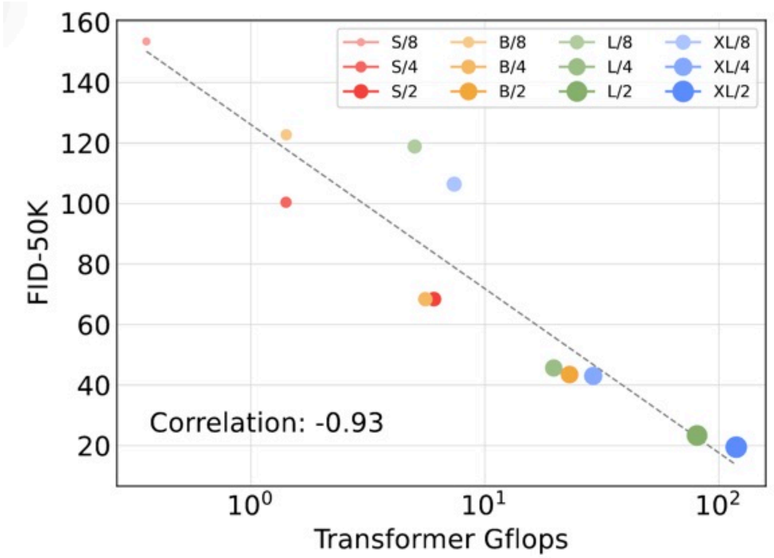
Hard to do scaling law study

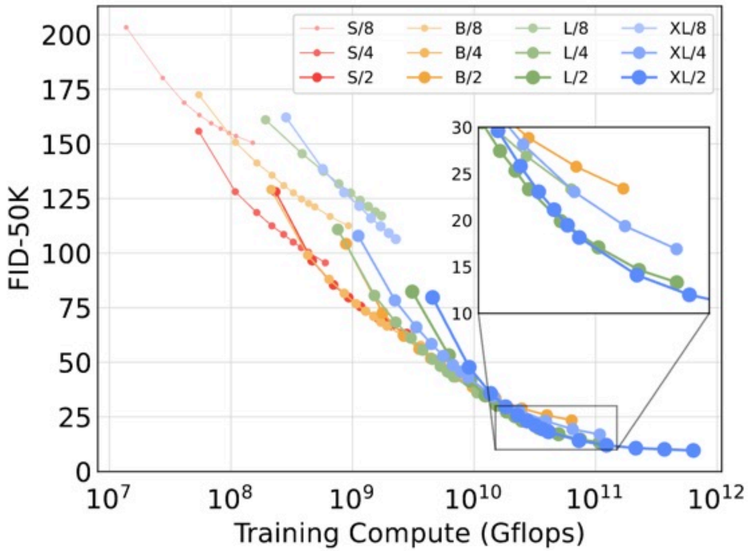
Reference: https://arxiv.org/abs/2210.02303

SORA, Feb 2024

Less is more: generic transformer architecture, a single latent diffusion model







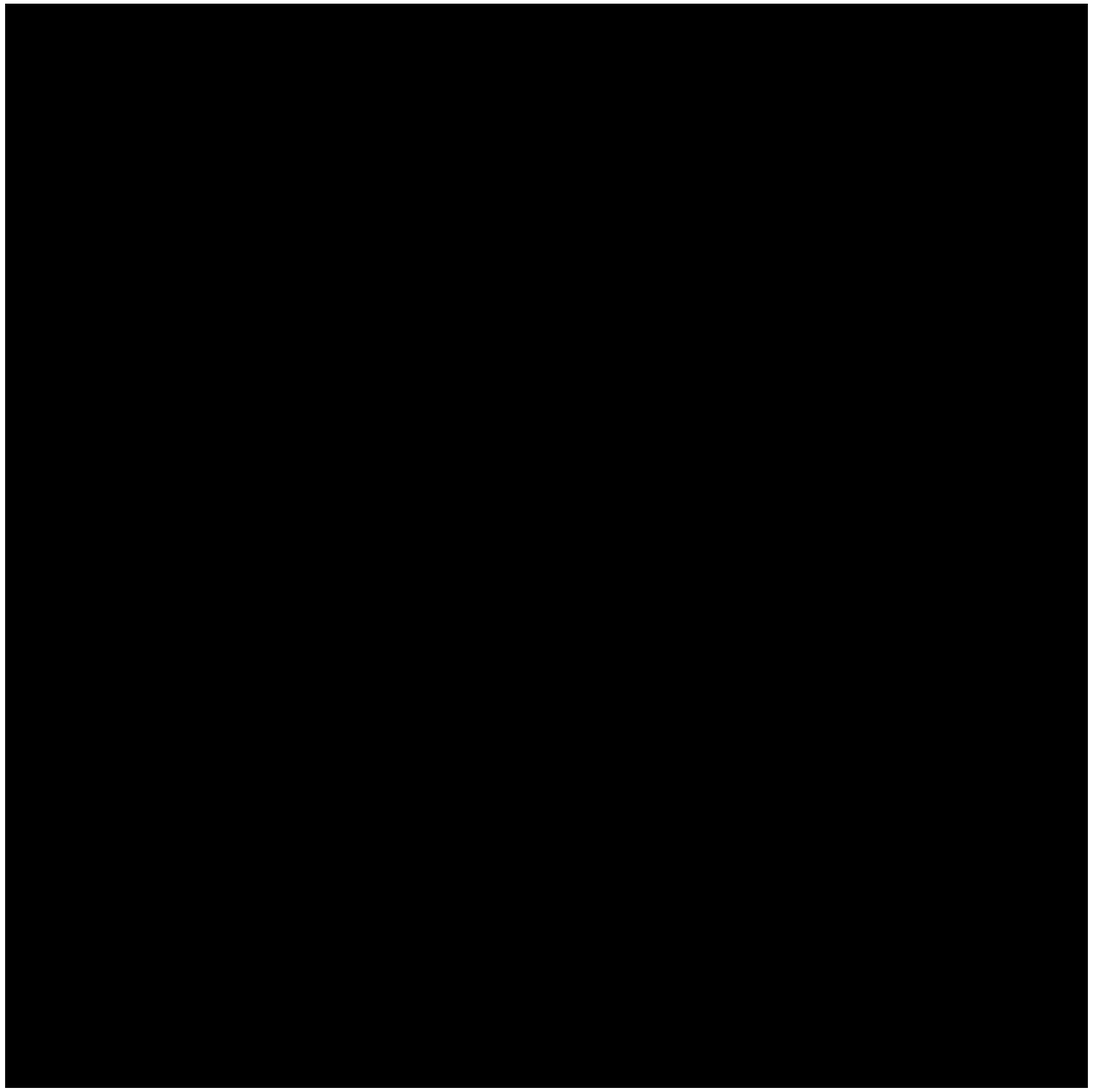
✅ Diffusion transformer scales well with increasing compute and model size.

✅ Support arbitrary aspect ratio generation.

References: https://openai.com/sora/ https://arxiv.org/abs/2212.09748

Veo 1: May 2024

Veo series: video models by Google



😔 slow in motion

😔 unrealistic physics / hallucination

😔 not great on text following

Veo 2: Dec 2024

- Create videos at resolutions up to 4k
- Understand camera controls in prompts, e.g. wide shot, POV and drone shot
- Better recreate real-world physics and realistic human expression

Veo 3: May 2025

Veo 3: sound on for videos, generating all audio natively.

Sound effects, ambient noise, dialogue, …

Veo 3: excelling in physics and realism

Veo 3: designed for greater control

- improved prompt adherence, following a series of actions and scenes with greater accuracy

“A delicate feature rests on a fence post. A gust of wind lifts it, sending it dancing over rooftops. It floats and spins, finally caught in a spiderweb on a high balcony.”

Agenda

What makes diffusion great?

Video diffusion models

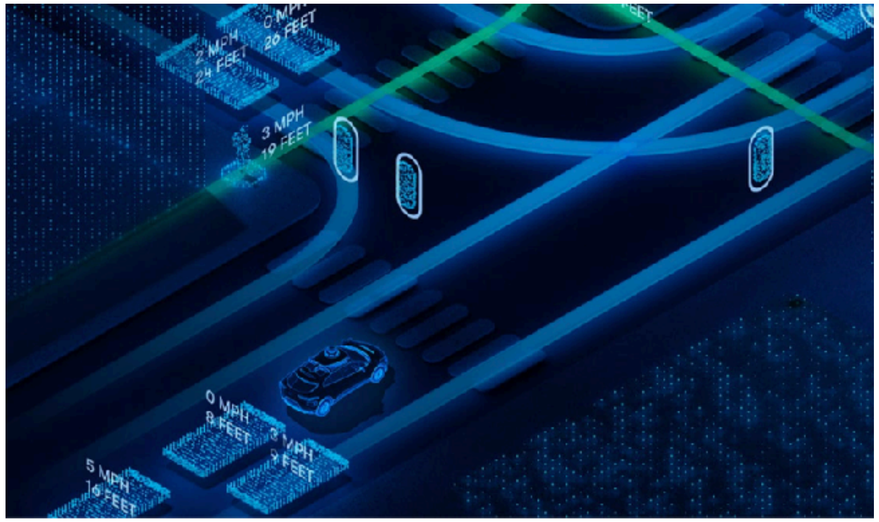
Go beyond video gen: world models

- Controllable video generation
- Multimodal models

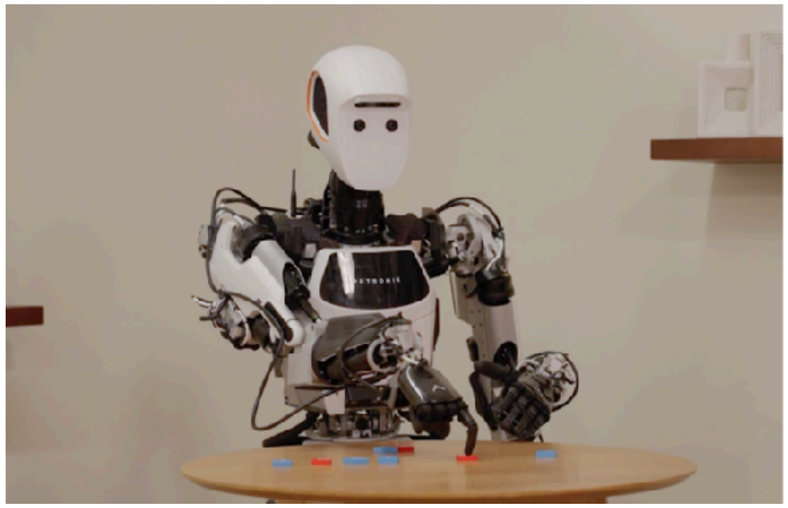
3D/4D generation

Embodied AGI

Building agents that are capable of understanding, reasoning and interacting with the real world.



Autonomous driving



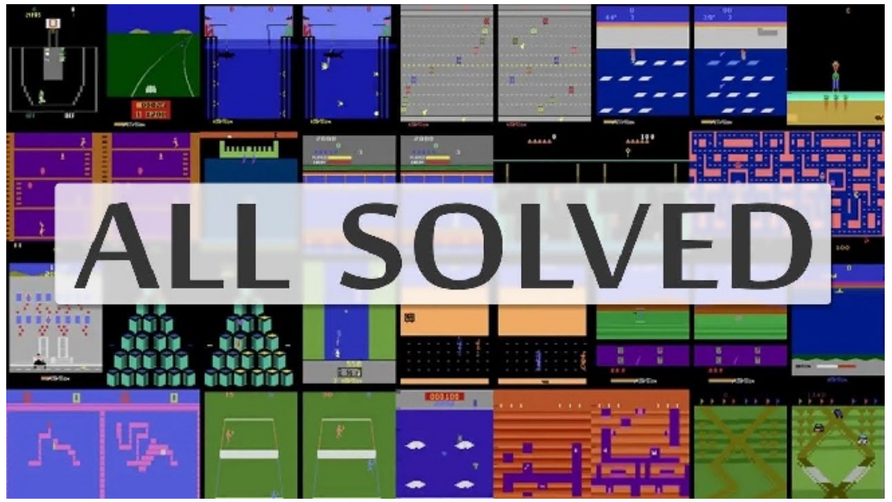
Humanoid robots

Multi-agent interaction

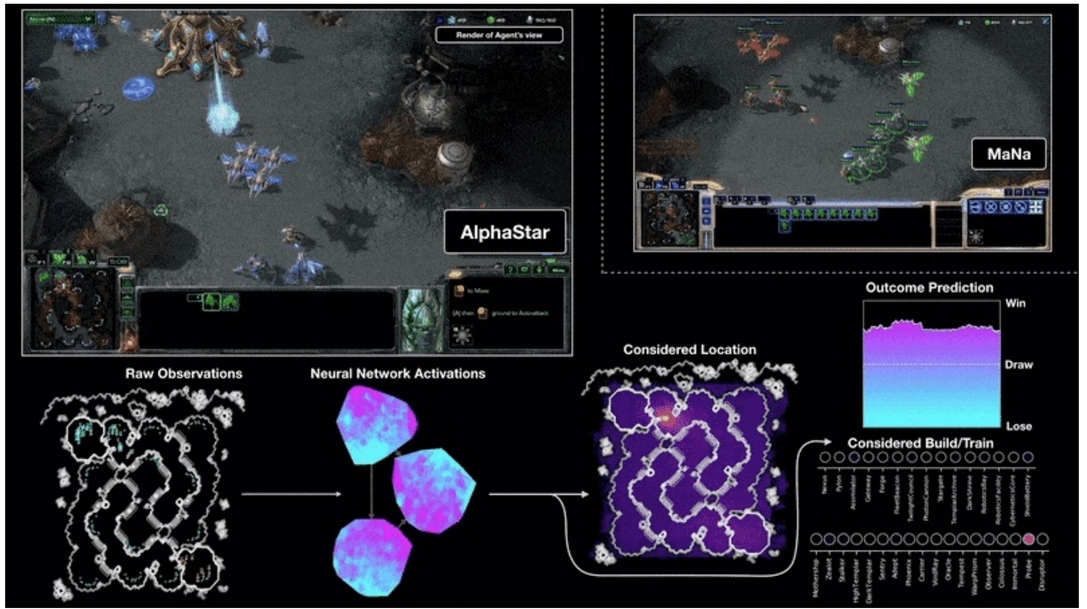
Training an AI agent

Works really well in domain specific or simulated environments

AlphaGo, 2016



Agent 57
2020



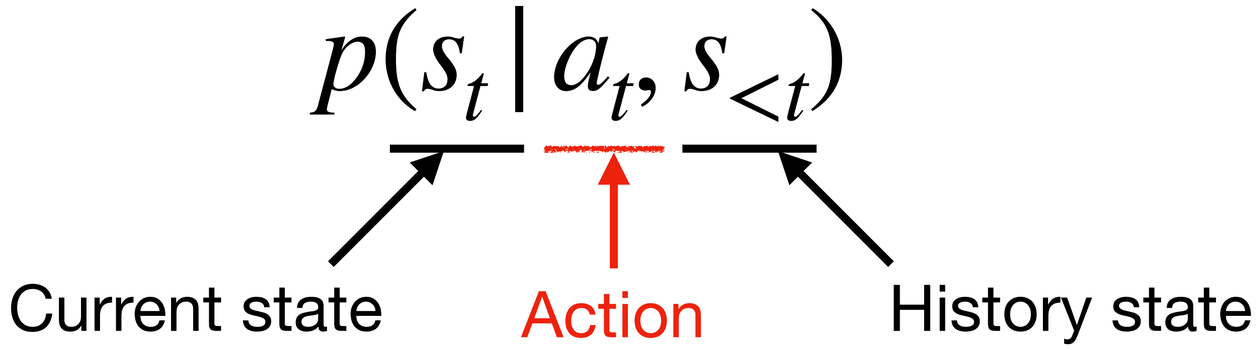
AlphaStar
2019

But … real world is complicated

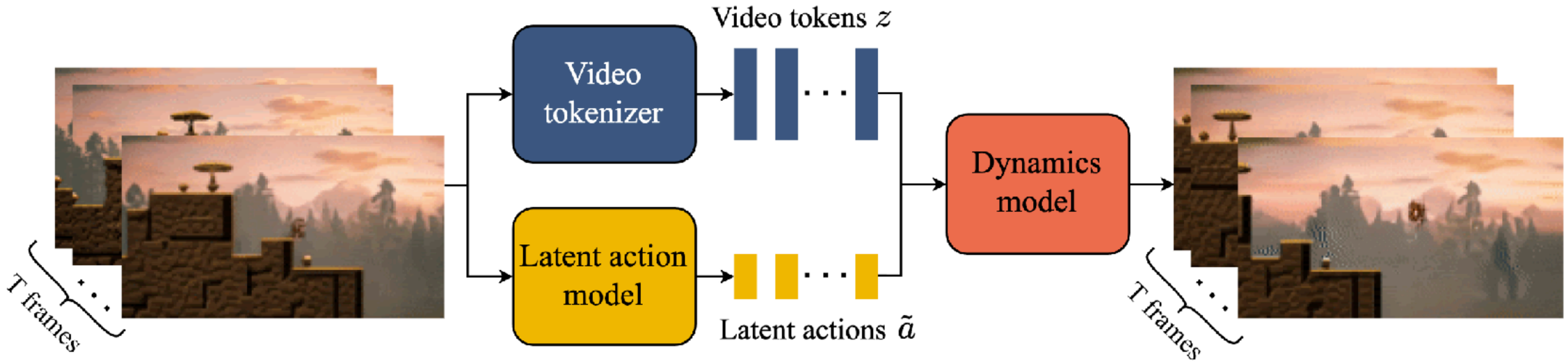
Domain specific agents cannot generalize well across various tasks or environments.

To train a general embodied agent …

We need world models to simulate diverse environments and interactions

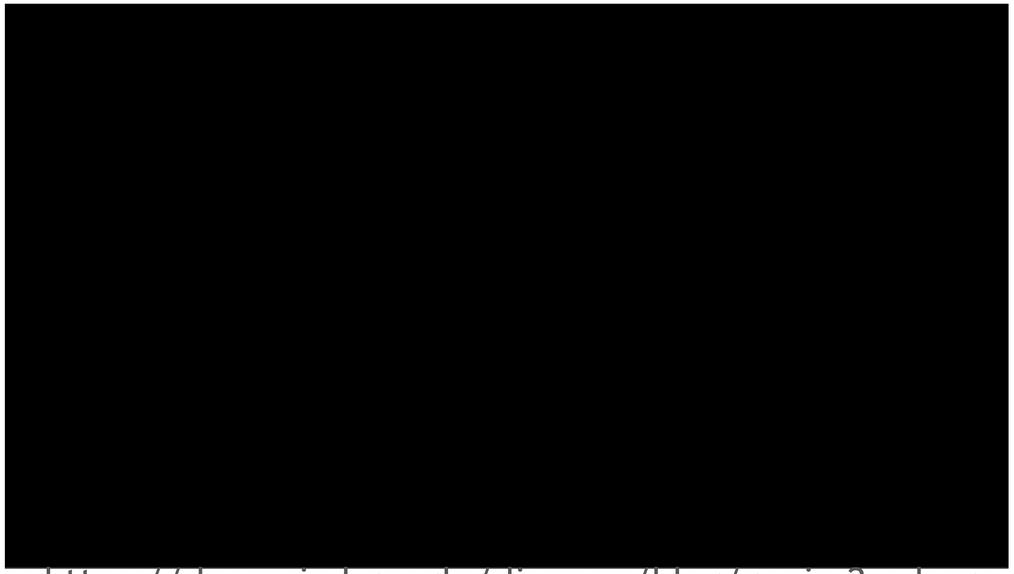


Genie: latent action control



Reference: https://arxiv.org/abs/2402.15391

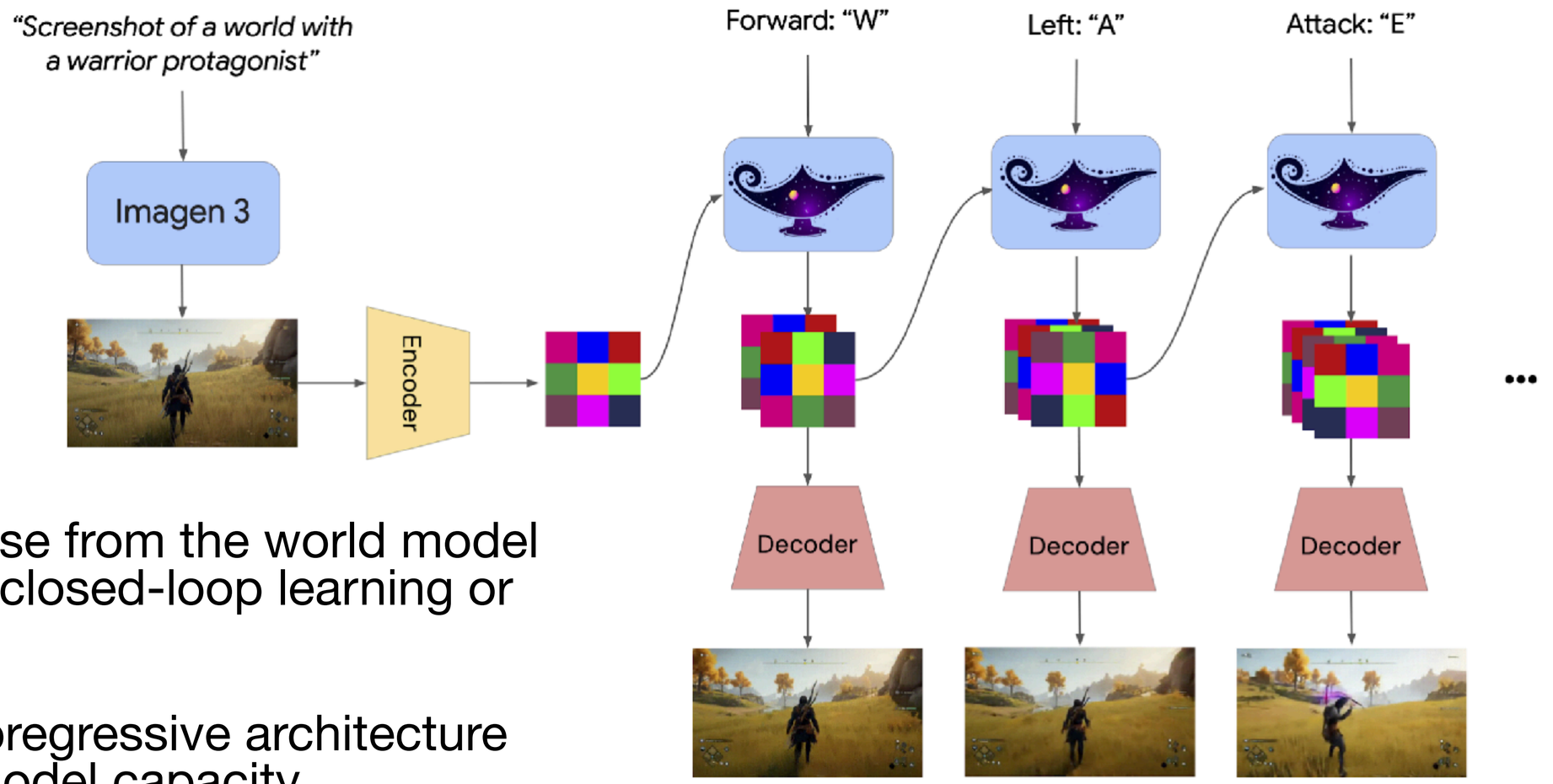
Genie 2: keyboard mouse action control



Reference: https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/

Genie 2: autoregressive video diffusion models

Generate one latent frame at each time

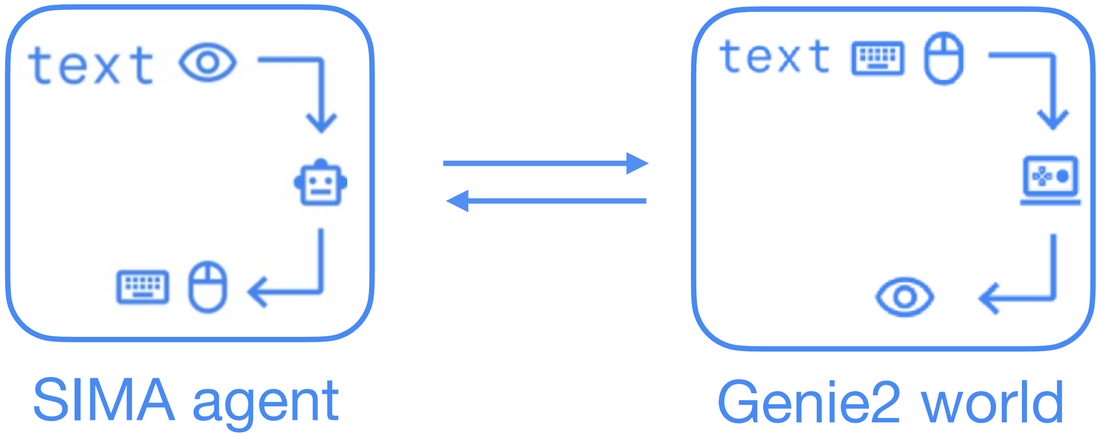


- Pros: get response from the world model faster, better for closed-loop learning or online RL.
- Challenges: autoregressive architecture constrains the model capacity

Reference: https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/

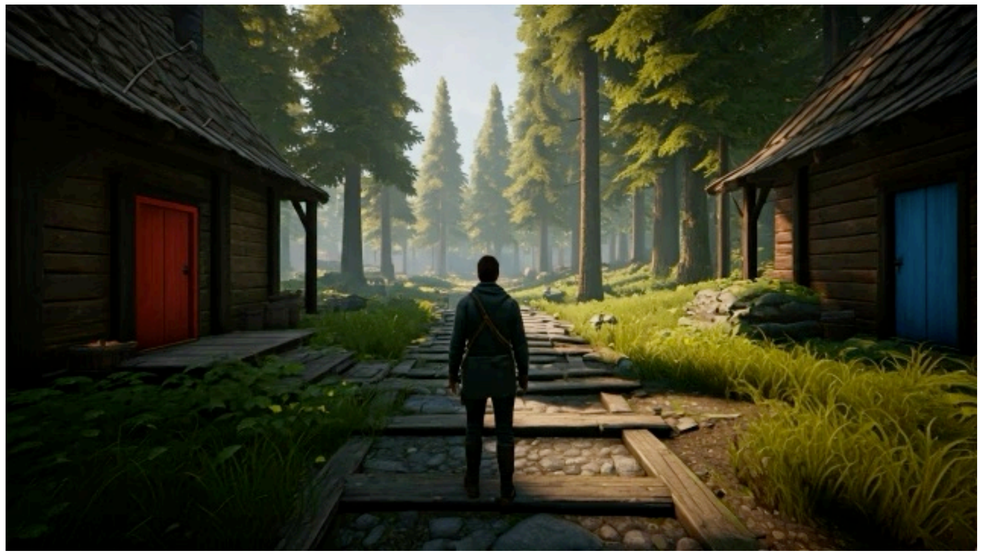
Genie 2: interact with an agent

Genie 2 (world model) simulates the outcome of SIMA (agent) given text and visual input

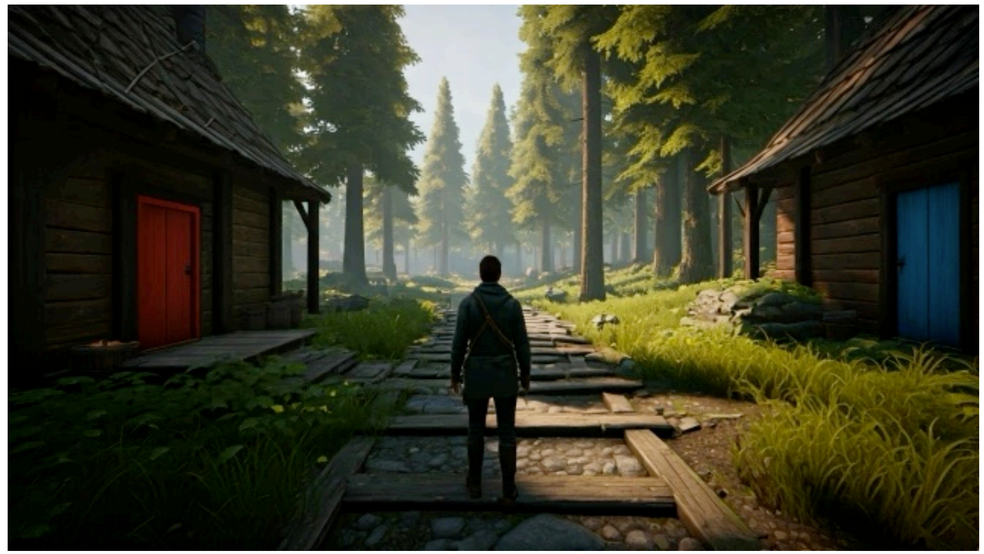


SIMA agent

Genie2 world



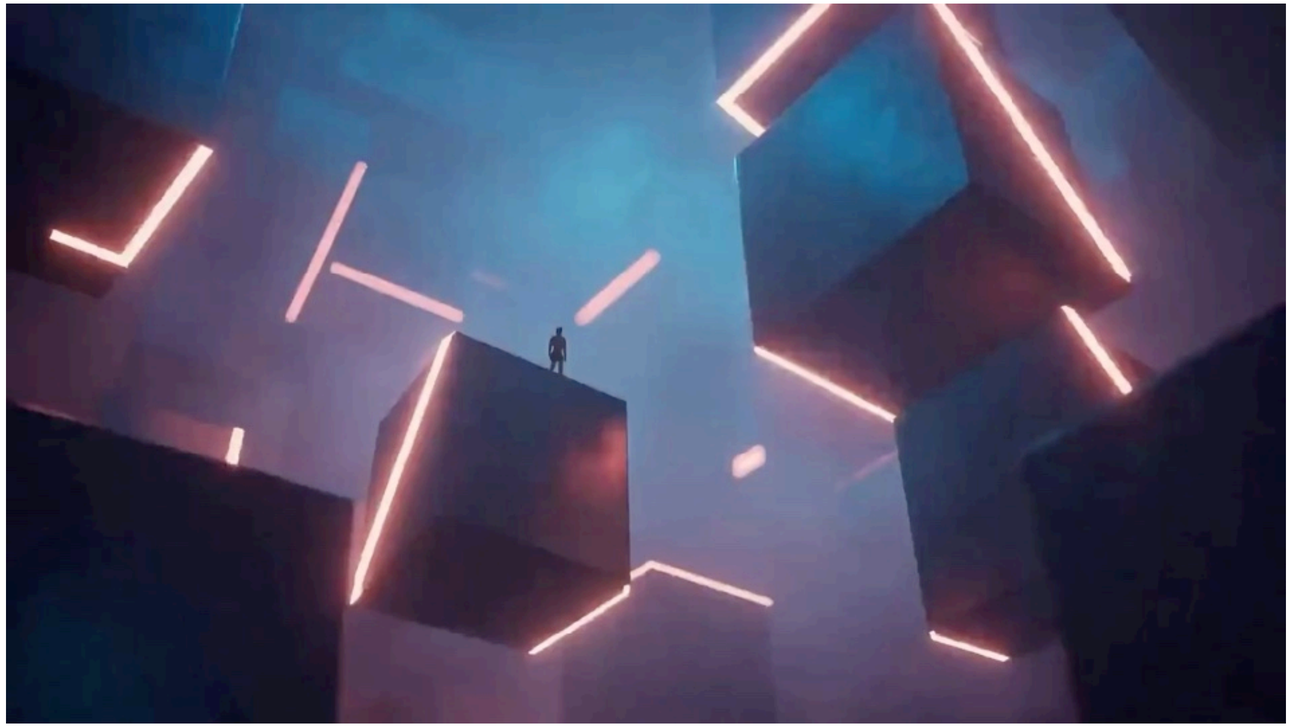
“Open the blue door”



“Open the red door”

Other types of actions: motion control

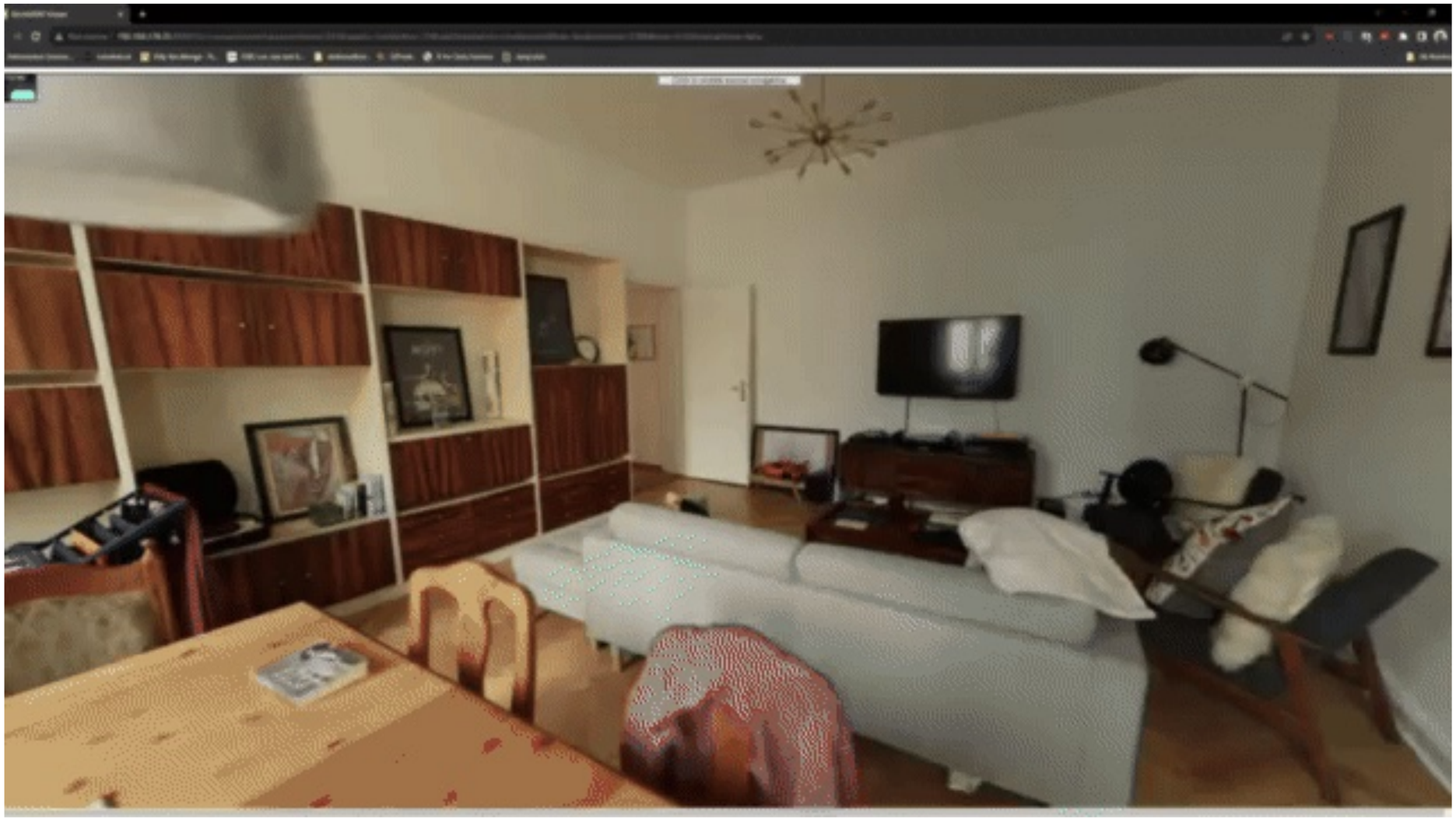
Define the exact movement of objects by selecting an object and defining their path



Media source: https://deepmind.google/models/veo/

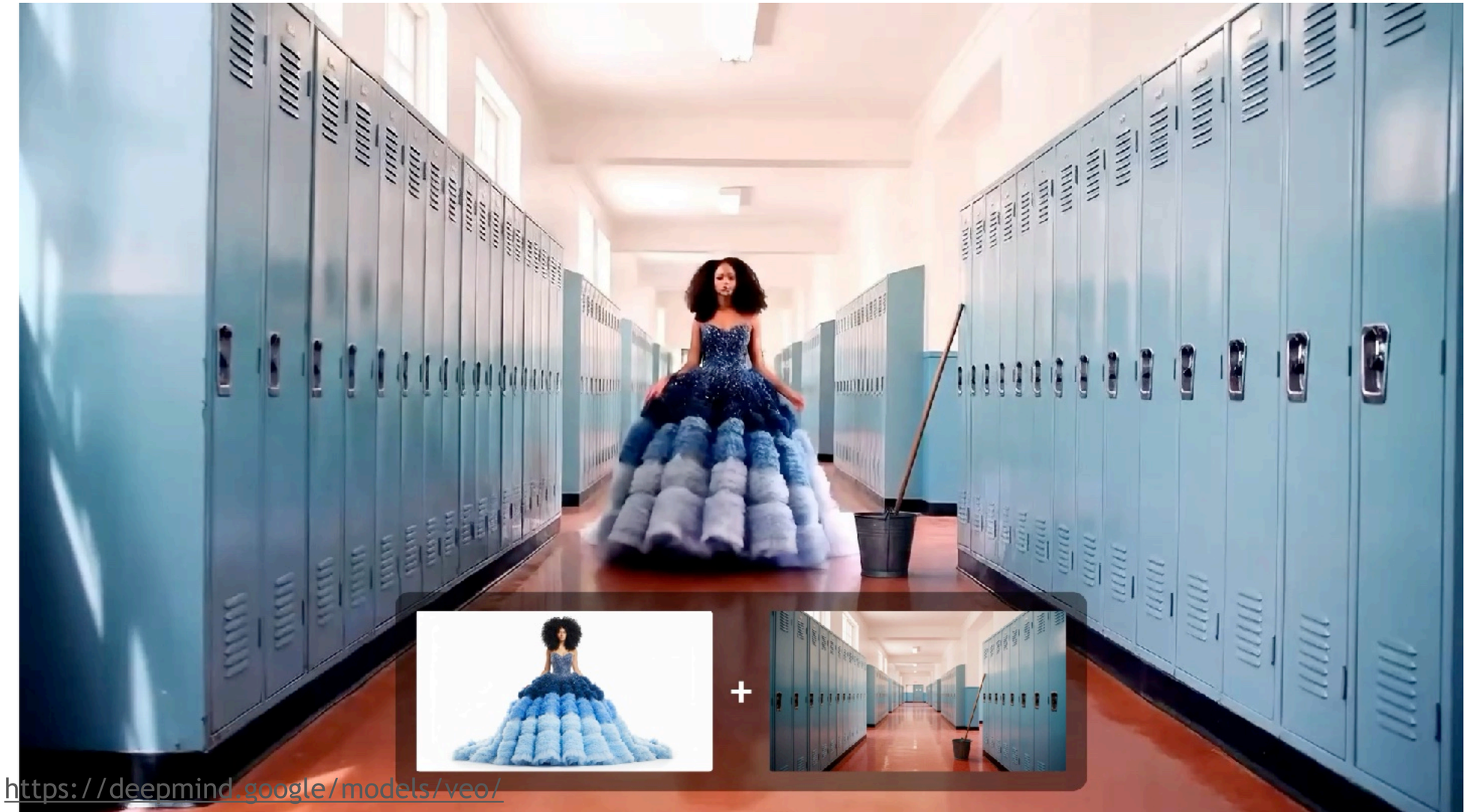
Other types of actions: camera control

Define the location and viewing direction of the agent



Scene + character control

Control what agent to put into what world



Media source: https://deepmind.google/models/veo/

Open questions for video controls

- A unified formulation of: text? Multimodal?
- A unified conditioning mechanism.
- Combining different controls together.

Agenda

Video diffusion models

- What makes diffusion great
- Journey of video diffusion models

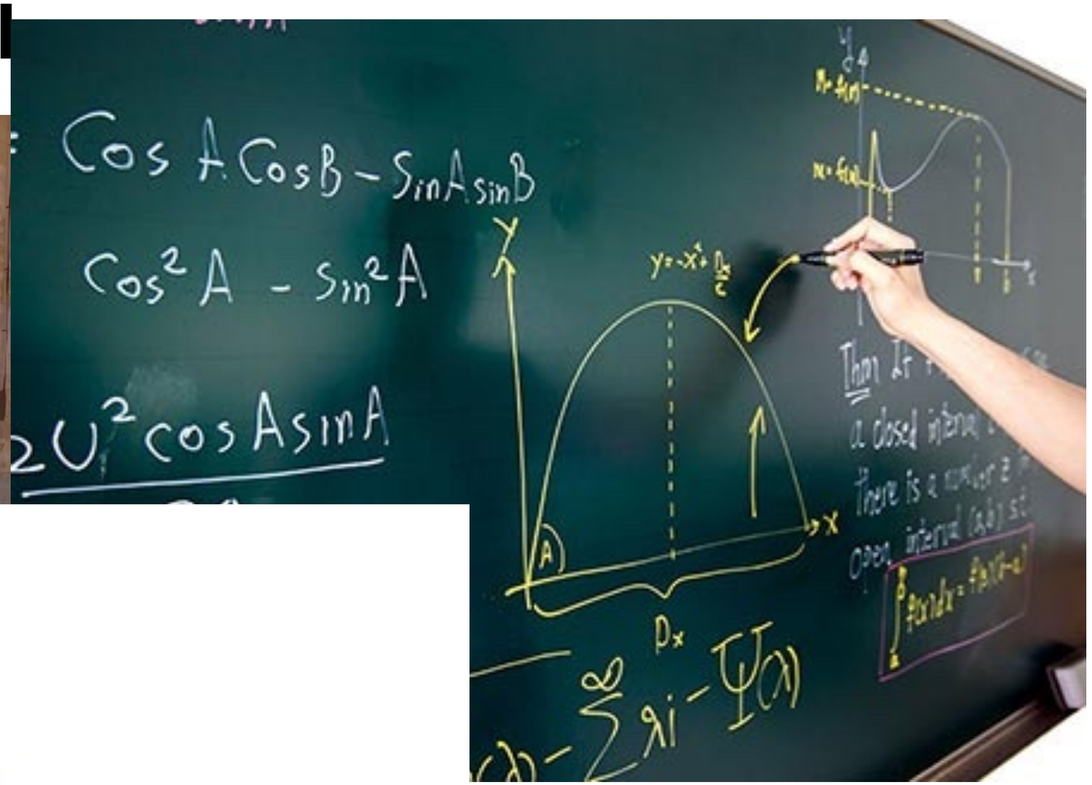
Go beyond video gen: video model as world models

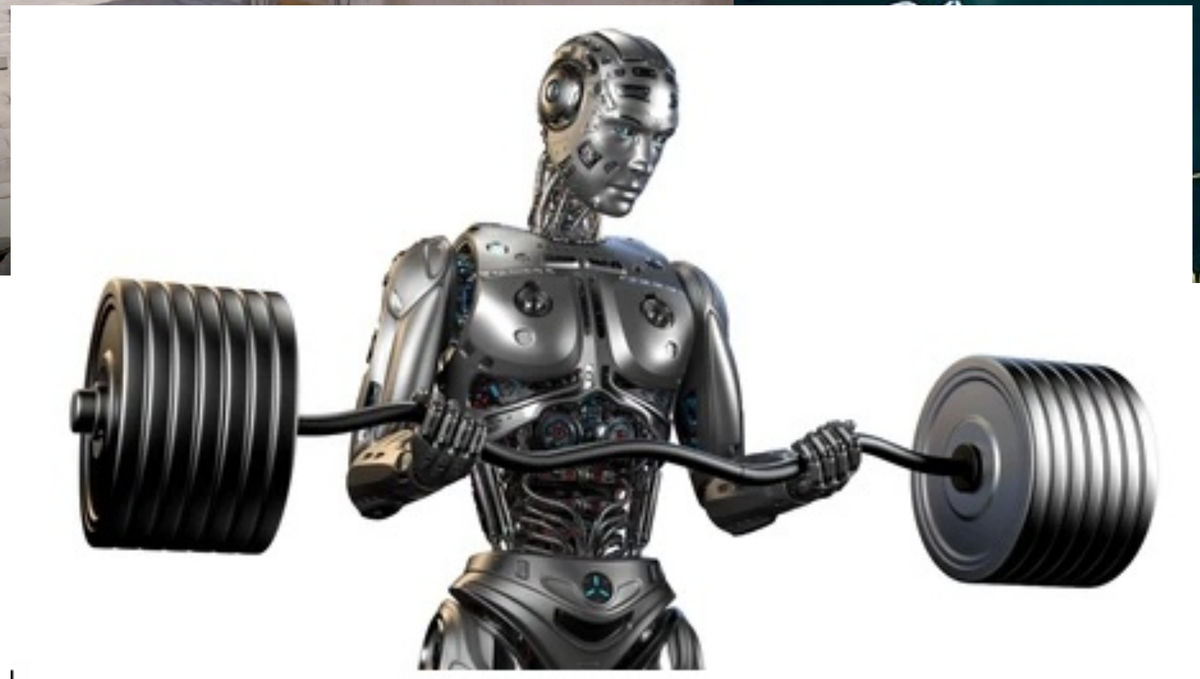
- Control video generation with actions
- Multimodal models

Remaining challenges

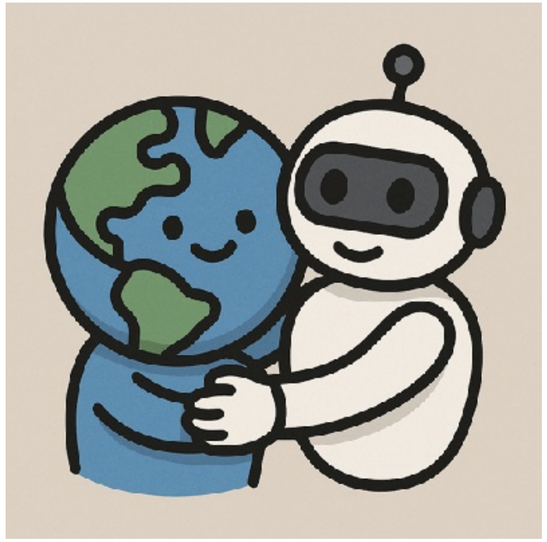
Why is multimodal important?

The world is generically multi-modal





Unifying world model + agent: let the model output action

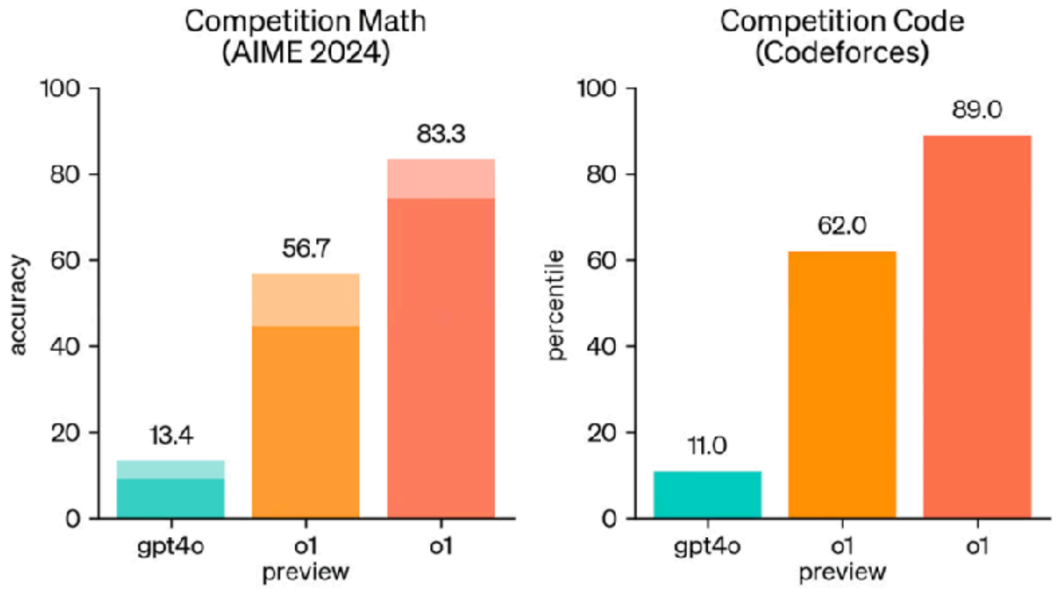


- Native generation of actions + observations

Unifying world model + agent: let the model output action

“human reasoning is based on the construction and evaluation of mental models of the world around us”

— *Kenneth Craik "The Nature of Explanation"*



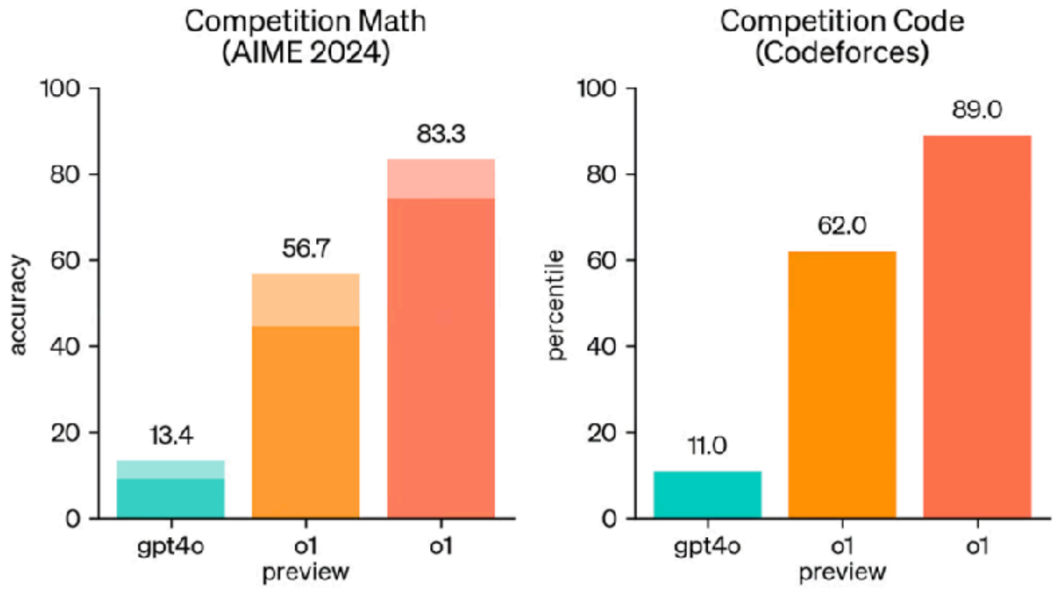
- Native generation of actions + observations
- Visual thinking and reasoning

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

Unifying world model + agent: let the model output action

“human reasoning is based on the construction and evaluation of mental models of the world around us”

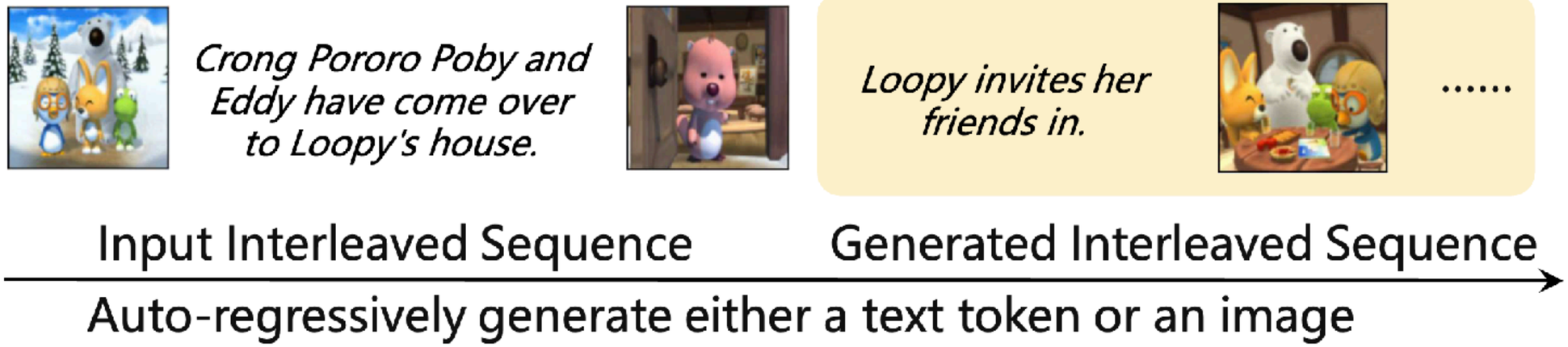
— *Kenneth Craik "The Nature of Explanation"*



- Native generation of actions + observations
- Visual thinking and reasoning
- Better inductive bias to compress the visual data, than “what is important to perception”. Can compress more aggressively

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

Hype in interleaved text+image generation in the past year



- Consider text as action, image as state, then it models all the following:

  p(state | action, history), p(action | state, history), p(state, action | history)

  World model | Agent | Agent w/ mental world model

Image source: https://arxiv.org/pdf/2401.10208

Hype in interleaved text+image generation in the past year

For image: discrete or continuous tokens? AR or diffusion?

Chameleon: Mixed-Modal Early-Fusion Foundation Models

Chameleon Team[1,*]

[1]FAIR at Meta

Discrete auto-regressive

SHOW-O: ONE SINGLE TRANSFORMER TO UNIFY MULTIMODAL UNDERSTANDING AND GENERATION

Jinheng Xie[1†] Weijia Mao[1†] Zechen Bai[1†] David Junhao Zhang[1†] Weihao Wang[2]
Kevin Qinghong Lin[1] Yuchao Gu[1] Zhijie Chen[2] Zhenheng Yang[2] Mike Zheng Shou[1*]

Discrete diffusion

Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model

Chunting Zhou[μ*] Lili Yu[μ*] Arun Babu[δ†] Kushal Tirumala[μ]
Michihiro Yasunaga[μ] Leonid Shamis[μ] Jacob Kahn[μ] Xuezhe Ma[σ]
Luke Zettlemoyer[μ] Omer Levy[†]

Emerging Properties in Unified Multimodal Pretraining

Chaorui Deng[*1], Deyao Zhu[*1], Kunchang Li[*2‡], Chenhui Gou[*3‡], Feng Li[*4‡]
Zeyu Wang[5‡], Shu Zhong[1], Weihao Yu[1],Xiaonan Nie[1], Ziang Song[1], Guang Shi[1§]
Haoqi Fan[*†]

Continuous diffusion

Interleaved generation

Should we use autoregressive or continuous diffusion for image / video?

Discrete autoregressive

✅ Unified training objective for vision and text

✅ Simple hps design space

❌ Lossy in compression

❌ slow and not flexible in sampling

Continuous diffusion

❌ Distinct training objectives, mixing clean and noisy tokens

❌ A lot of domain specific hps to consider: noise schedule, weighting, guidance, …

✅ Almost lossless compression

✅ flexibility in sampling: distillation, guidance, advanced ODE/SDE samplers

Challenge 1: Long term consistency

How to ensure the world remains consistent as the agent navigates in it?

- Static parts of the environment should remain the same over time.

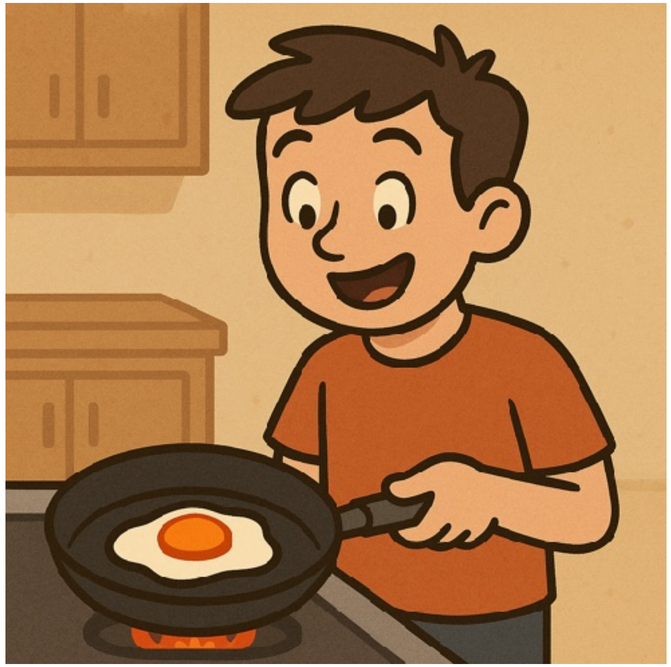
A person went outside of a room and came back later. The room should remain the same.

Challenge 1: Long term consistency

How to ensure the world remains consistent as the agent navigates in it?

- Dynamic parts of the scene should progress over time reasonably, even if it is outside of the field of view of the video.





A person pan fried an egg, did something else and later came back. The egg should be cooked with time passing by.

🤔 How about betting on long context transformer + scale?

- Probably fine for video generation only, but not for world models.
- Not efficient: we'll hit the context length limit with videos very soon. Unlikely to fit a 1h video into the context, with the current hardware.
- No such data at scale: long-term exploration videos are rare and not diverse enough.

Challenge 2: Real-time interaction

Let the agent navigate and interact in real time

- For many use cases, getting the feedback from the world model as soon as possible is critical for training a good agent.

Agenda

What makes diffusion great?

Video diffusion models

Go beyond video gen: world models

- Controllable video generation
- Multimodal models

3D/4D generation

What is behind a video?

An entire 3D world!

“More than 90% of the pixels are static in the 3D world”

- A large portion of pixel changes in video can be explained away by camera movement.
- Can we disentangle camera motion and scene dynamics?
- Can we maintain a memory that does not grow linearly as the number of frames (maybe 3D aware)?

3D grounded generation

Avoid memory growing linearly as the number of frames

- Retrieval-based methods: selecting most relevant history
- Stateful models: maintain a “state” that progresses over time.
- Conditional on some 3D representation of the world

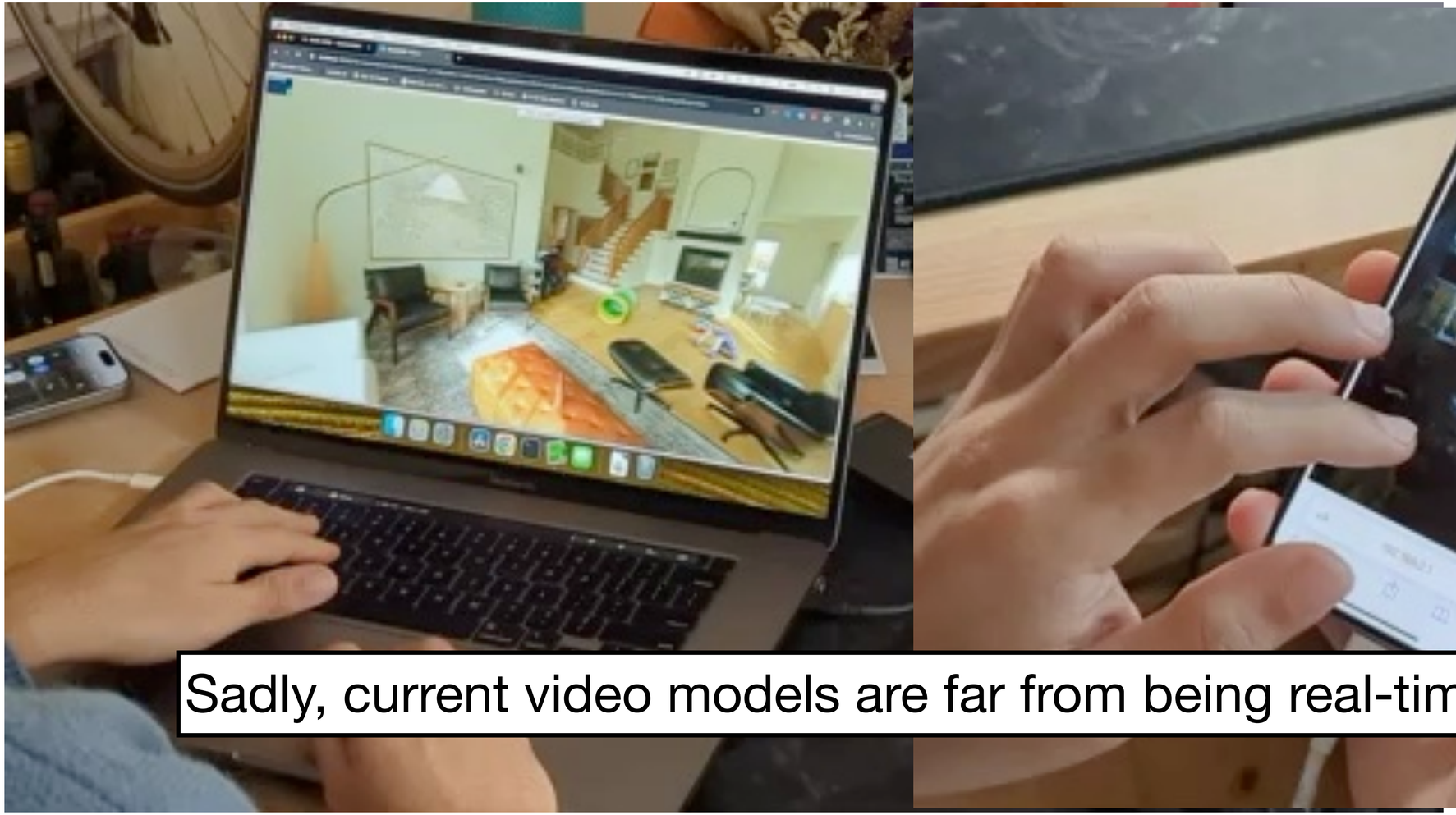
Video World Models with Long-term Spatial Memory

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, Gordon Wetzstein

Video source: https://arxiv.org/abs/2506.05284

Real-time interaction

Navigate and interact in real time



Sadly, current video models are far from being real-time 💔

Zip-NeRF:

Anti-Aliased Grid-Based Neural Radiance Fields

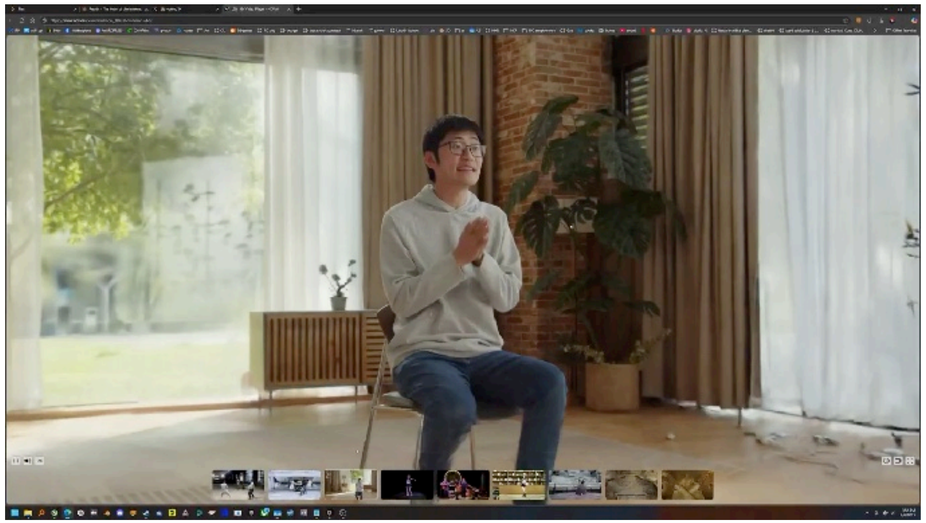
ICCV 2023

3D models enable real-time interactive experiences cheaply

Interactive 4D scenes—baked out experiences

Video source: https://www.4dv.ai/

Baked out 4D scenes



- Explore-only, can’t interact
+ **One-time optimization cost, then “free” to explore**

Fast video model

+ **Interact with diverse actions!**
- Expansive per-action model call

Can we find a way to combine the best of two worlds?

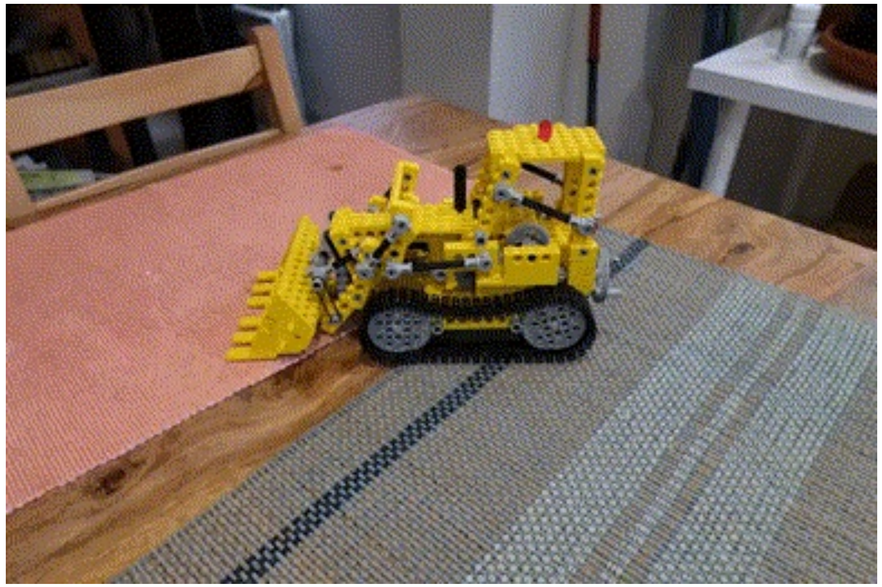
Video source: https://www.4dv.ai/

Traditional 3D reconstruction requires many captions images to work well

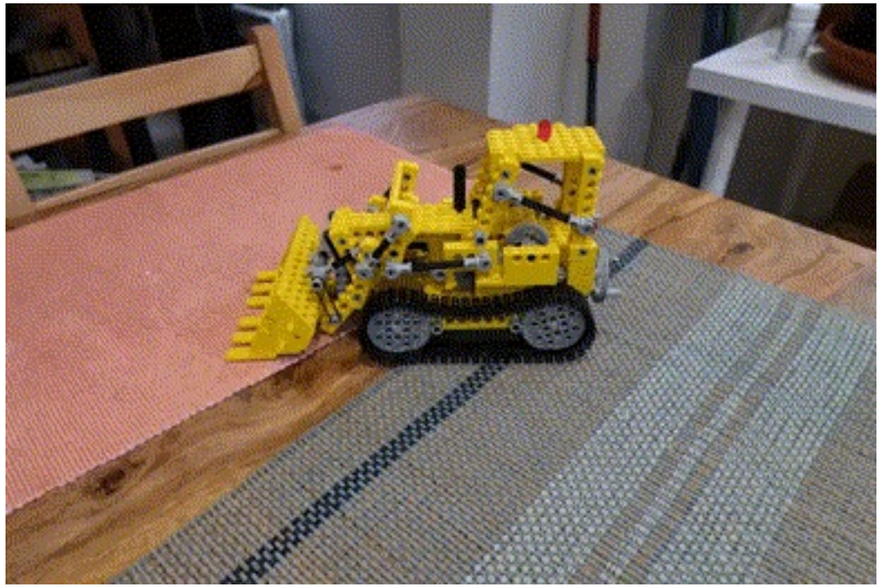
Reconstruction

Generation

100s of images + poses



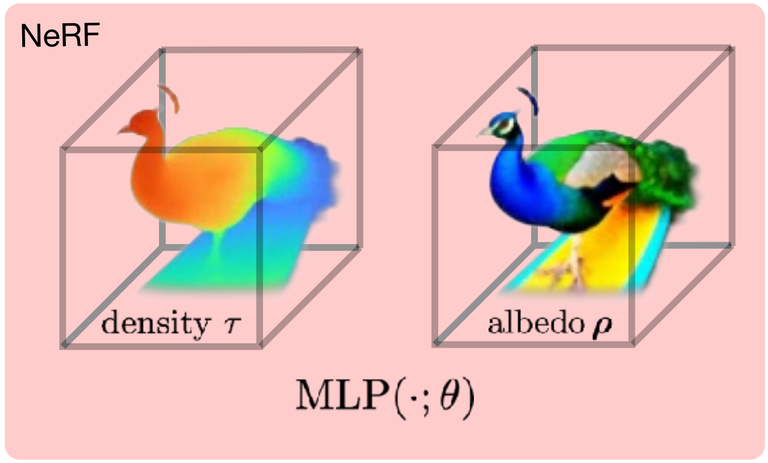
Single image

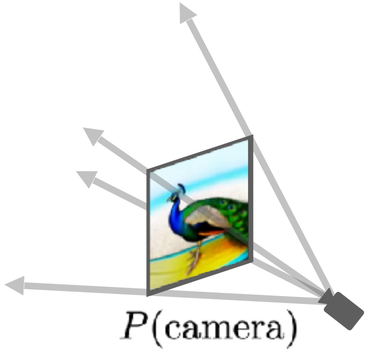


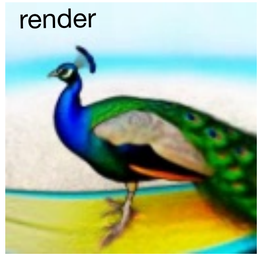
‘‘a bulldozer’’

Common Goal: Create a 3D scene consistent with any partial information

DreamFusion = NeRF + Score Distillation



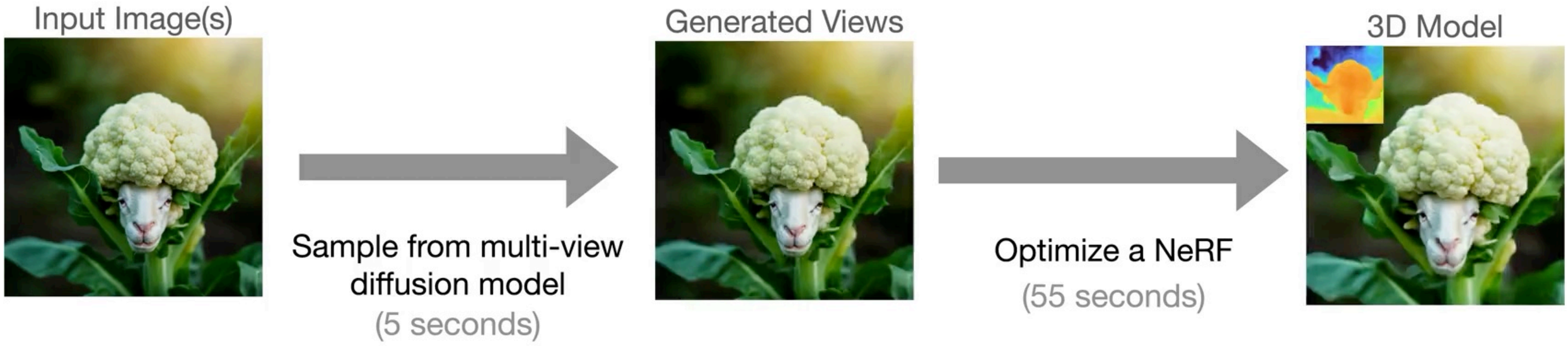


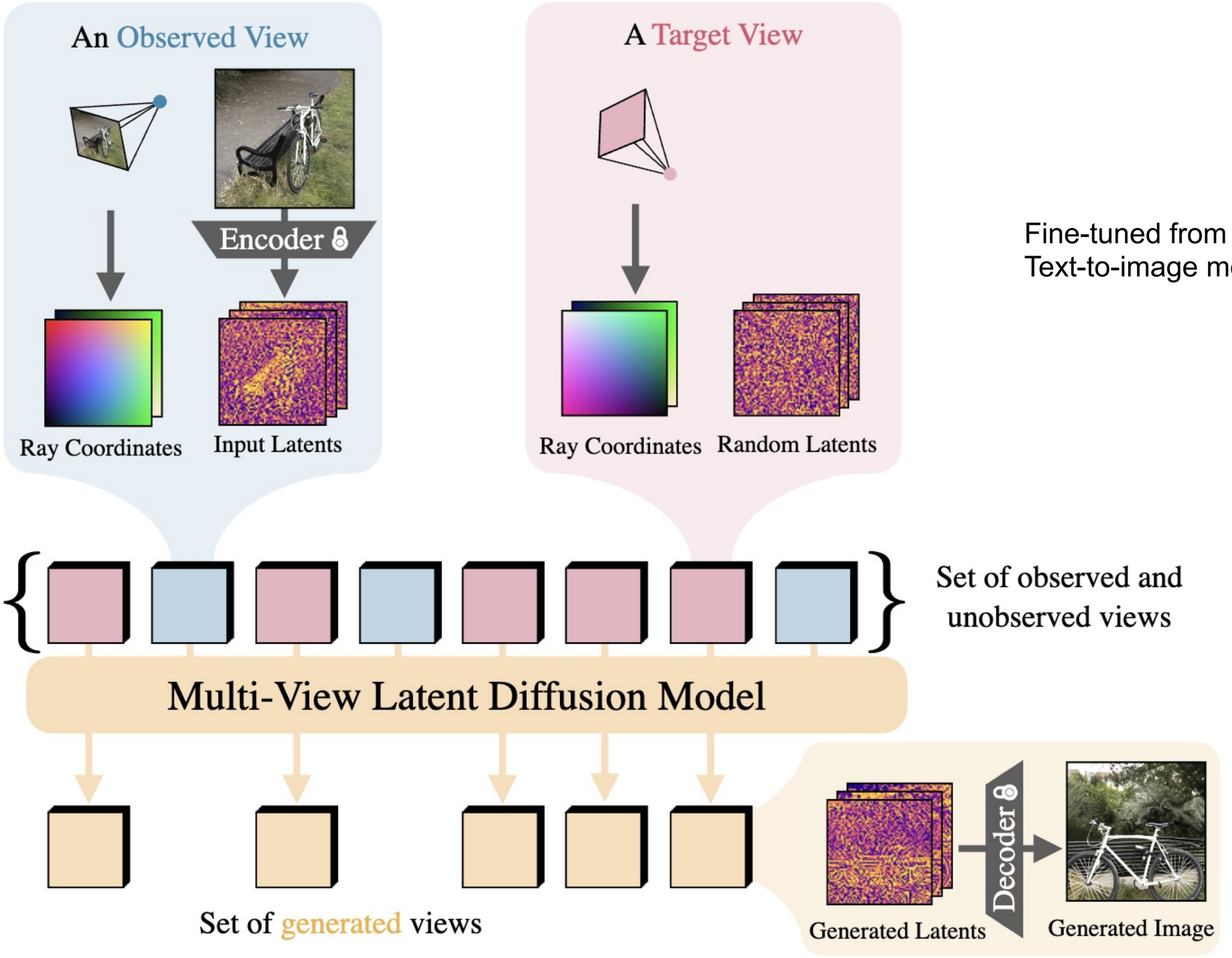


"a peacock on a surfboard"

Score Distillation

Generate, and then reconstruct

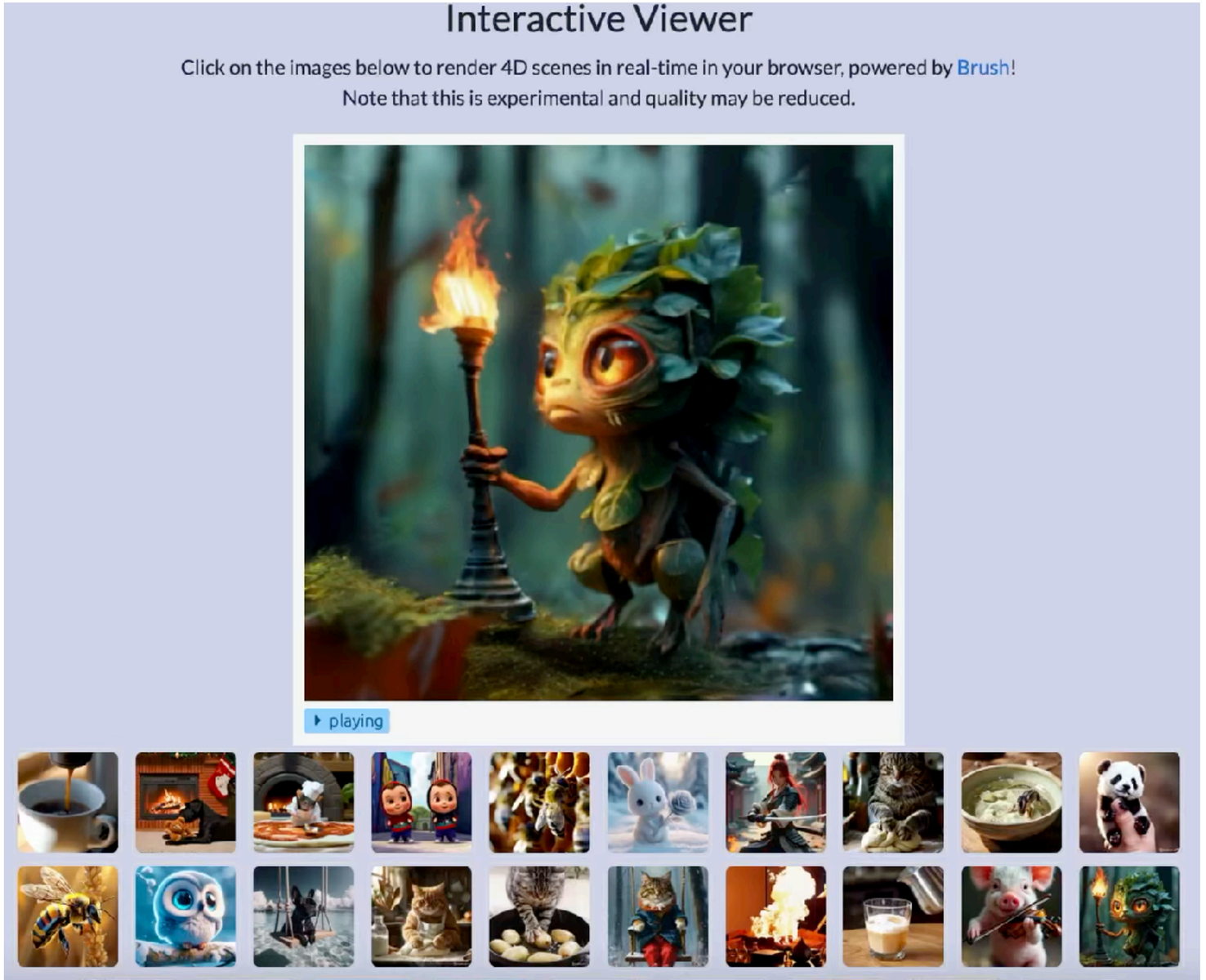




Fine-tuned from a Text-to-image model

Generate, and then reconstruct

CAT3D, CAT4D



Closing thoughts

- **Diffusion Models:**
  - Many design choices make it a robust and scalable generative modeling framework.
  - Sampling can be made faster by faster samplers / distillation.
- **Video Generation:** Developing rapidly. Initial usage in real-world content creation.
- **World models:** Excited active research area. Challenges remain.
- **3D and 4D Generation:** Not yet as robust and scalable. But promising and catching up quickly, powered by advances in image and video models.


Thank you!