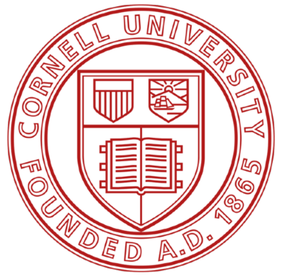
Lecture 25: Evaluation + robotics

CS 5788: Introduction to Generative Models



Many slides from (or inspired by) Volodymyr Kuleshov's generative modeling class

Today

Mix of topics (one by request!):

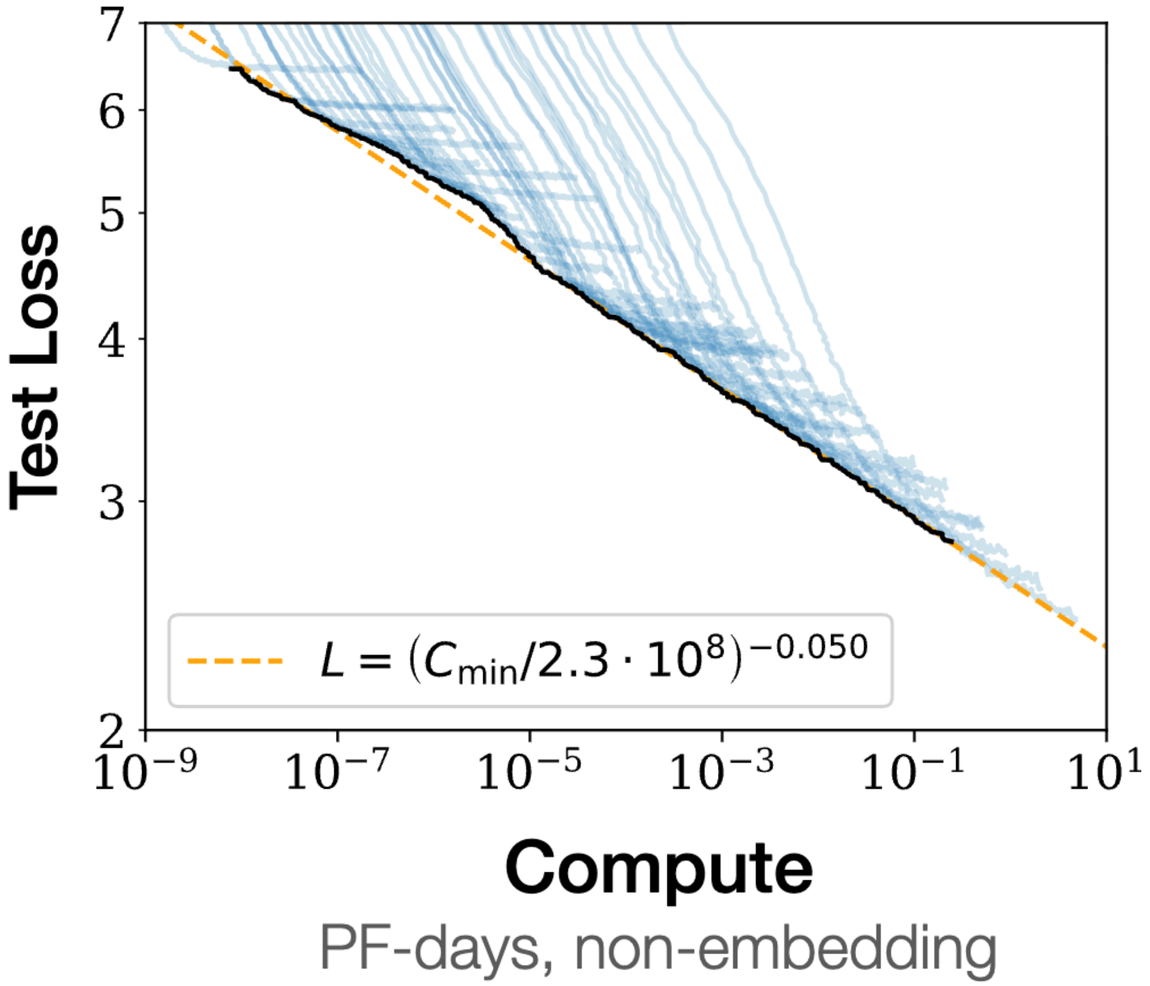
- Evaluating generative models
- Generative models for robotics

Reminder: presentations

- Short, 3-minute talks from each group.
- Give an overview of your project and current progress.
- Motivate the problem and your solution.
- Please sign up for Thursday's class. We have not heard of any group that is unable to attend one of the two sessions (please tell us by tonight at 11:59pm if you can't make it).

How can we evaluate our generative models?

Log likelihood



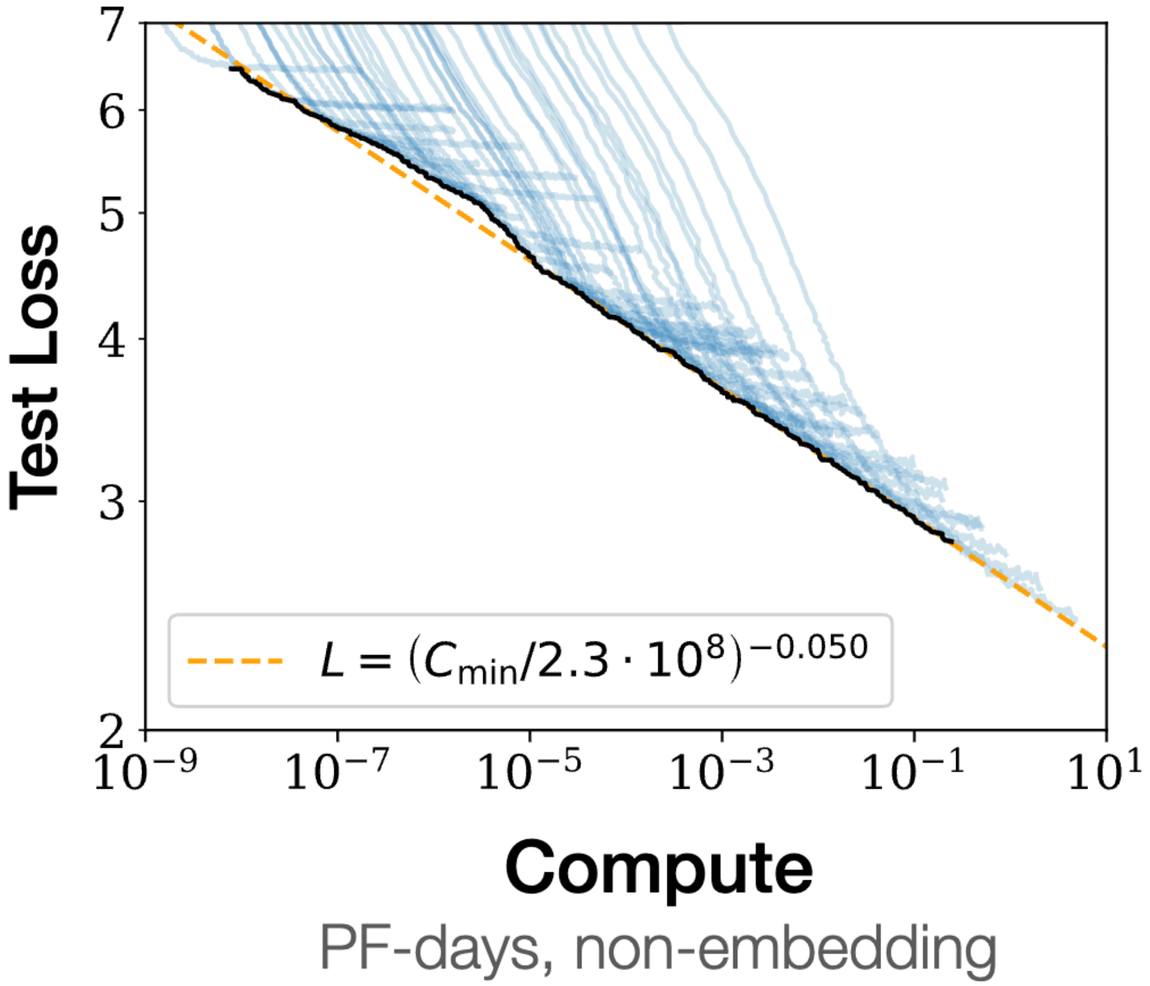
$$\text{NLL} = \sum_{i=1}^{N} -\log(p_\theta(\mathbf{x}_i))$$

Strengths:

- The loss that autoregressive (and other) models are directly optimizing.
- Highly correlated with other metrics.
- Not easily “gameable”.
- Statistically well motivated, e.g., as compression.

Figure source: [Kaplan et al., 2020]

Log likelihood



$$\text{NLL} = \sum_{i=1}^{N} -\log(p_\theta(\mathbf{x}_i))$$

Shortcomings:

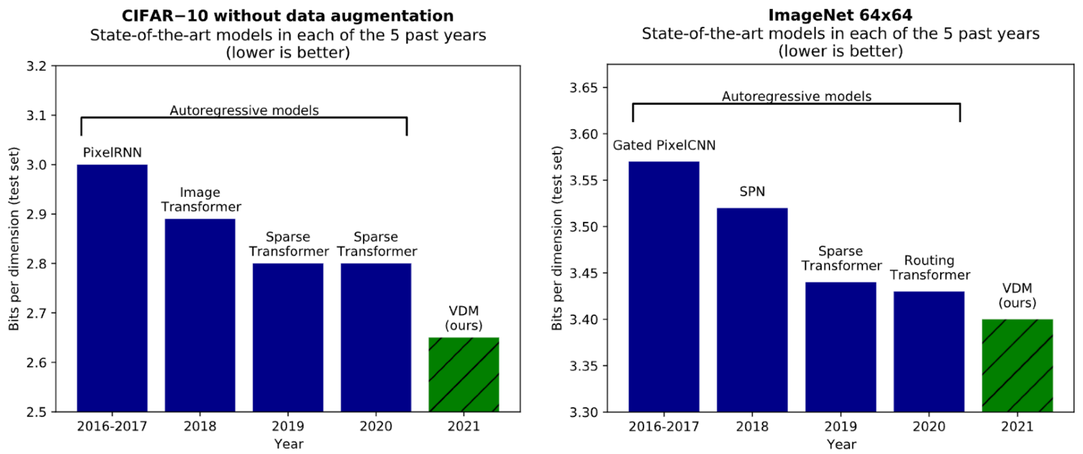
- Can't directly compute it for many models.
- Correlated with what we want (e.g., image generation fidelity) but might want to optimize that more directly.
- Motivation stronger for very large N, like internet-scale language datasets.

Figure source: [Kaplan et al., 2020]

Variational lower bounds?

Many models (like diffusion and VAEs) optimize the evidence lower bound. Does this actually lead to a good model in the log likelihood sense?

Comparing diffusion to autoregressive models



(a) CIFAR-10 without data augmentation

(b) ImageNet 64x64

Figure 1: Autoregressive generative models were long dominant in standard image density estimation benchmarks. In contrast, we propose a family of diffusion-based generative models, *Variational Diffusion Models* (VDMs), that outperforms contemporary autoregressive models in these benchmarks. See Table 1 for more results and comparisons.

- In some cases, can obtain variational lower bounds that, when optimized, outperform models that are based on exact likelihoods.
- Example: variational diffusion models. Get details right to improve the variance of the variational lower bound (e.g., learn the noise schedule).
- Outperformed autoregressive models when measured in bits-per-dimension.

[Kingma et al., “Variational diffusion models”, 2021]

Directly optimizing image fidelity

- If you care mostly about whether generated images “look real”, is this the best metric?
- What if you don’t necessarily have a large model or data?

Evaluation using an external classifier

- Suppose that we have a classifier $p(y \mid \mathbf{x})$ for our dataset and a model to evaluate, $p_\theta(\mathbf{x})$.
- Generate a large batch of images.
- Test whether these images follow the categorical distribution that you expect, by computing the **inception score** (named after Inception neural net):
  1. Do the generated images clearly belong to one category?
  2. Do they also cover the space of categories well?

Slide adapted from Volodymyr Kuleshov

Desiderata for classifier-based evaluation metric

high classifier entropy

low classifier entropy

- We want low entropy: each images should only be assigned one distinct category.
- For example, we want to minimize:

$$H(\mathbf{x}) = -\int_y p_\psi(y \mid \mathbf{x}) \log p_\psi(y \mid \mathbf{x}) dy$$

Slide and figure adapted from Volodymyr Kuleshov

Desiderata for classifier-based evaluation metric

low diversity

high diversity

- We want to cover all of the categories, in their proper ratios.
- So, $p_\psi(y \mid \mathbf{x})$ should have high entropy when we consider the full distribution of images that it generates.

Slide and figure adapted from Volodymyr Kuleshov

Combining both criteria together

- Inception score [Salimans et al., 2016] captures both of these criteria:

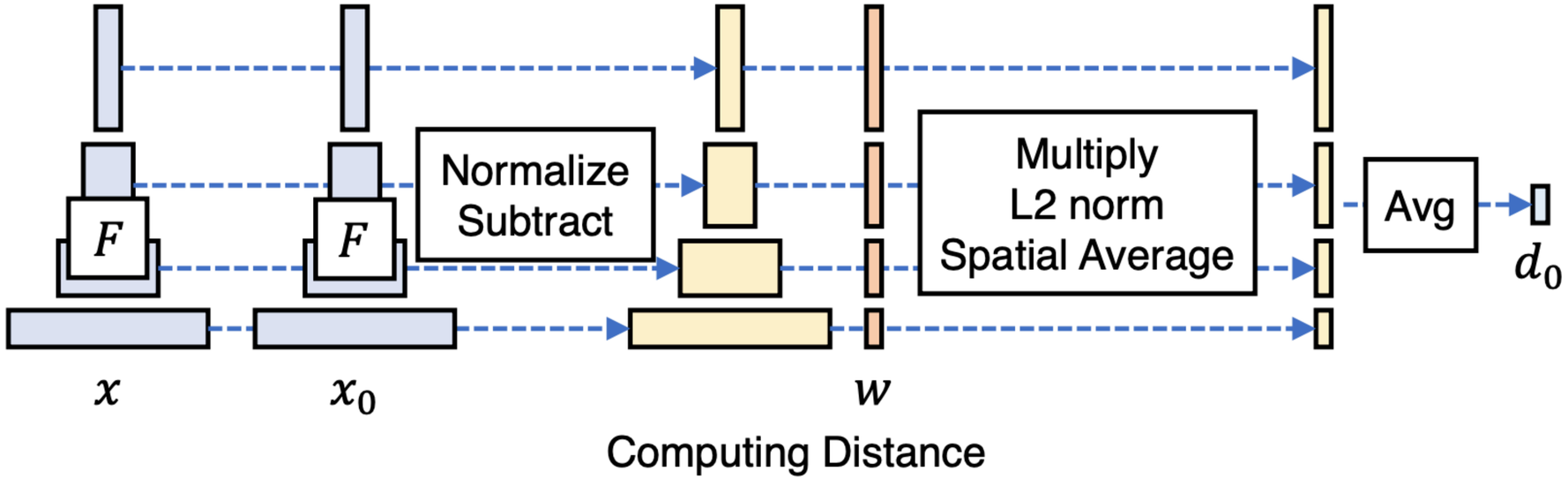
$$\text{IS} = \exp\left(\mathbb{E}_{\mathbf{x}}\left[D_{\text{KL}}(p_{\psi}(y \mid \mathbf{x}) \parallel p(y))\right]\right)$$

Slide and figure adapted from Volodymyr Kuleshov

What about conditional prediction tasks?

Measuring quality of predictions using perceptual loss

Suppose that we predict an image $\hat{\mathbf{x}}$ when the true image was $\mathbf{x}$. How can we measure the prediction quality?



Perceptual loss computes their similarity between a pair of images in the feature space of a neural net.

Source: [Zhang et al., "LPIPS", 2018]. See also: [Gatys et al., 2015, Johnson et al., 2016]

Perceptual loss

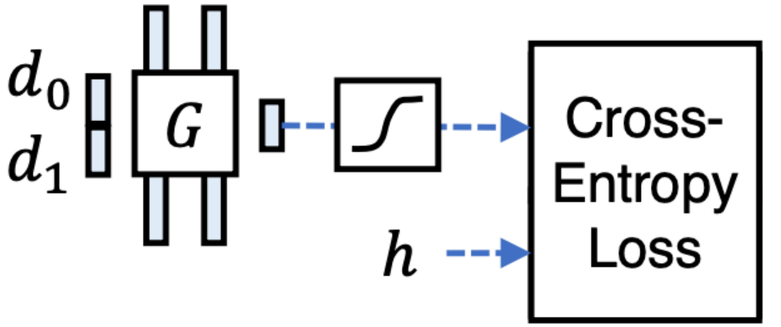
Ground Truth **Bicubic** **Ours** (ℓ_{pixel}) **Ours** (ℓ_{feat})

Super-resolution with perceptual loss

Source: [Johnson et al., 2016]

Learning human perceptual similarity

Which patch is more similar?



Predicting Perceptual Judgement

Directly learn image similarity functions using human feedback. The learned model (LPIPS) can be used either as an evaluation metric or loss.

Source: [Zhang et al., "LPIPS", 2018]

Can we use network features for unconditional modeling?

Fréchet inception distance

- Inception score is not general purpose.
- It requires a classifier for specific category-based generation task.
- What if we use network features instead?
- Fréchet inception distance (FID): measure similarity between a batch of image features (e.g., penultimate layer of Inception network) and those of the ground truth
- Widely used in other domains (e.g., FAD and FVD for audio and video) using modality-appropriate networks.

Fréchet inception distance

- Assume that the real and generated data are from a Gaussian distribution.
- Fréchet distance between Gaussian distributions:

$$d_F(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \mathcal{N}(\boldsymbol{\mu}', \boldsymbol{\Sigma}')) = \|\boldsymbol{\mu} - \boldsymbol{\mu}'\|^2 + \operatorname{tr}\left(\boldsymbol{\Sigma} + \boldsymbol{\Sigma}' - 2(\boldsymbol{\Sigma}\boldsymbol{\Sigma}')^{\frac{1}{2}}\right)$$

- Caveat: Requires lots of samples to estimate accurately and (often computed using 50k samples!), since the dimensionality of the features is typically high.
- What's still missing for conditional generation?

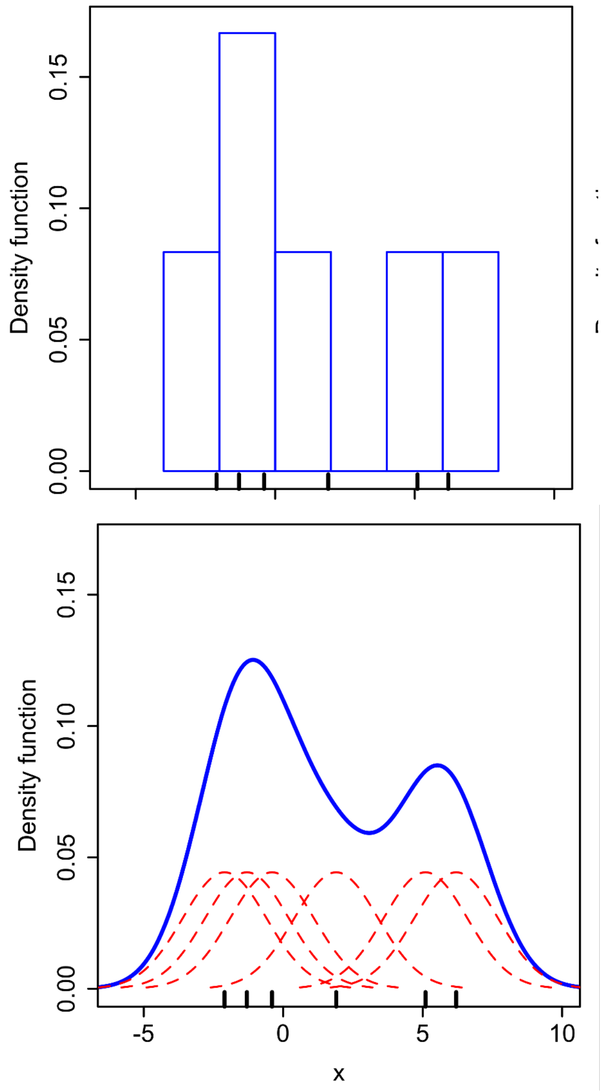
Example: evaluation for diffusion transformers (DiT)

Class-Conditional ImageNet 256×256					
Model	FID↓	sFID↓	IS↑	Precision↑	Recall↑
BigGAN-deep [2]	6.95	7.36	171.4	0.87	0.28
StyleGAN-XL [53]	2.30	4.02	265.12	0.78	0.53
ADM [9]	10.94	6.02	100.98	0.69	0.63
ADM-U	7.49	5.13	127.49	0.72	0.63
ADM-G	4.59	5.25	186.70	0.82	0.52
ADM-G, ADM-U	3.94	6.14	215.84	0.83	0.53
CDM [20]	4.88	-	158.71	-	-
LDM-8 [48]	15.51	-	79.03	0.65	0.63
LDM-8-G	7.76	-	209.52	0.84	0.35
LDM-4	10.56	-	103.49	0.71	0.62
LDM-4-G (cfg=1.25)	3.95	-	178.22	0.81	0.55
LDM-4-G (cfg=1.50)	3.60	-	247.67	**0.87**	0.48
DiT-XL/2	9.62	6.85	121.50	0.67	**0.67**
DiT-XL/2-G (cfg=1.25)	3.22	5.28	201.77	0.76	0.62
DiT-XL/2-G (cfg=1.50)	**2.27**	**4.60**	**278.24**	0.83	0.57

- Latent diffusion model (so hard to compare with other models unless they use same codebook),
- Main goal is maximizing generated image quality

Source: [Peebles & Xie, 2023]

What if you can only sample?



- What do you do if you can only sample from the model? For example, if you use a GAN
- For discrete distributions you can histogram: sample N points, then place them into buckets, thus fitting a simpler distribution.
  - What do you with distributions over vectors?
- You can use **Kernel density estimation**.
- Sample N points from your model and center a kernel
  - When the kernel is a Gaussian, you get a GMM!
- Not very accurate for high dimensional data.

Slide adapted from Volodymyr Kuleshov, figure source: Wikipedia

Generative models for robotics

Imitation learning

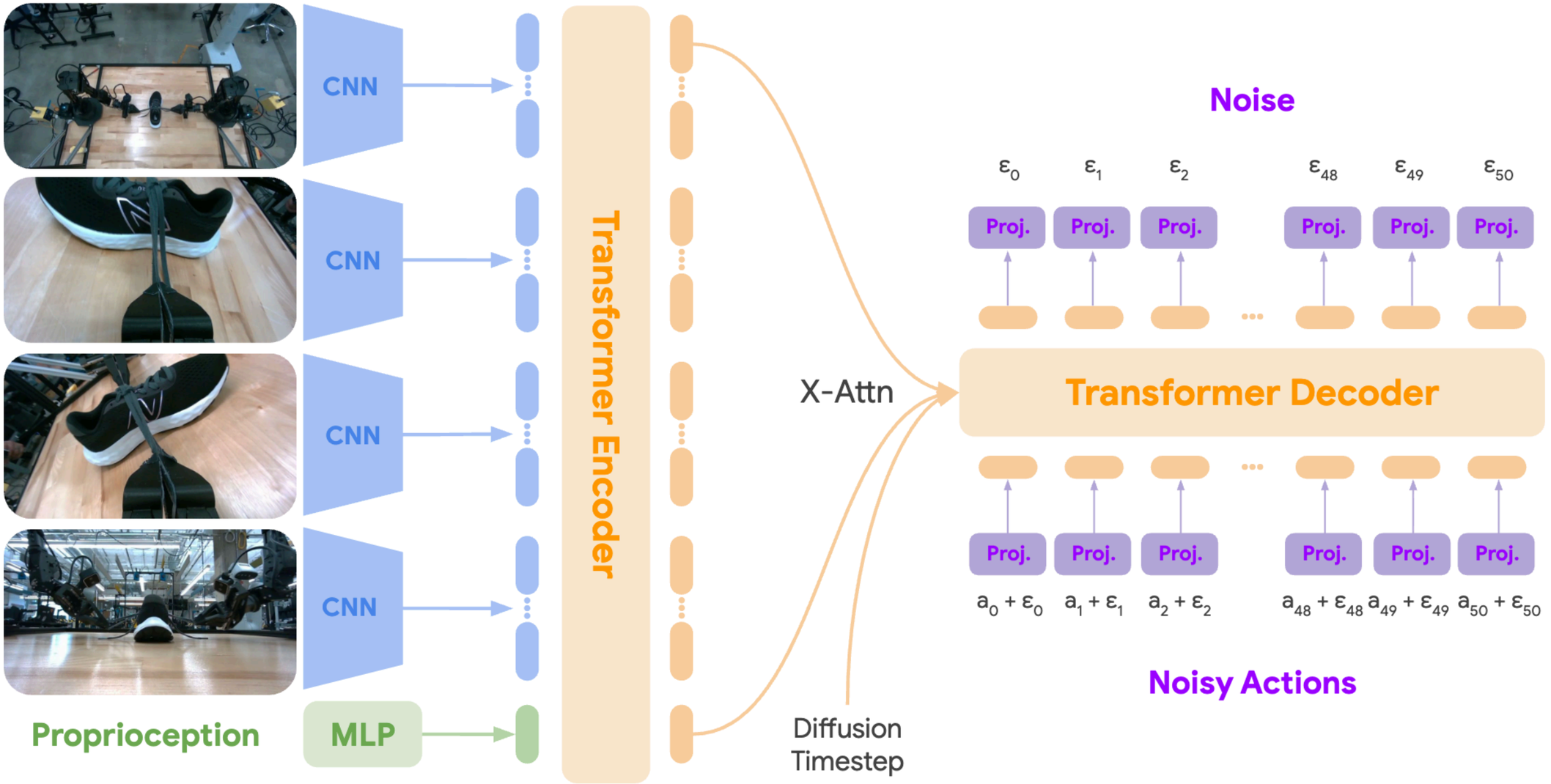
- Language models learn to imitate human writing.
- Analogously, can we train robots to imitate human motor commands?

Teleoperation

[ALOHA 2: An Enhanced Low-Cost Hardware for Bimanual Teleoperation, 2024]

Teleoperation for manipulation tasks

Predicting actions from video

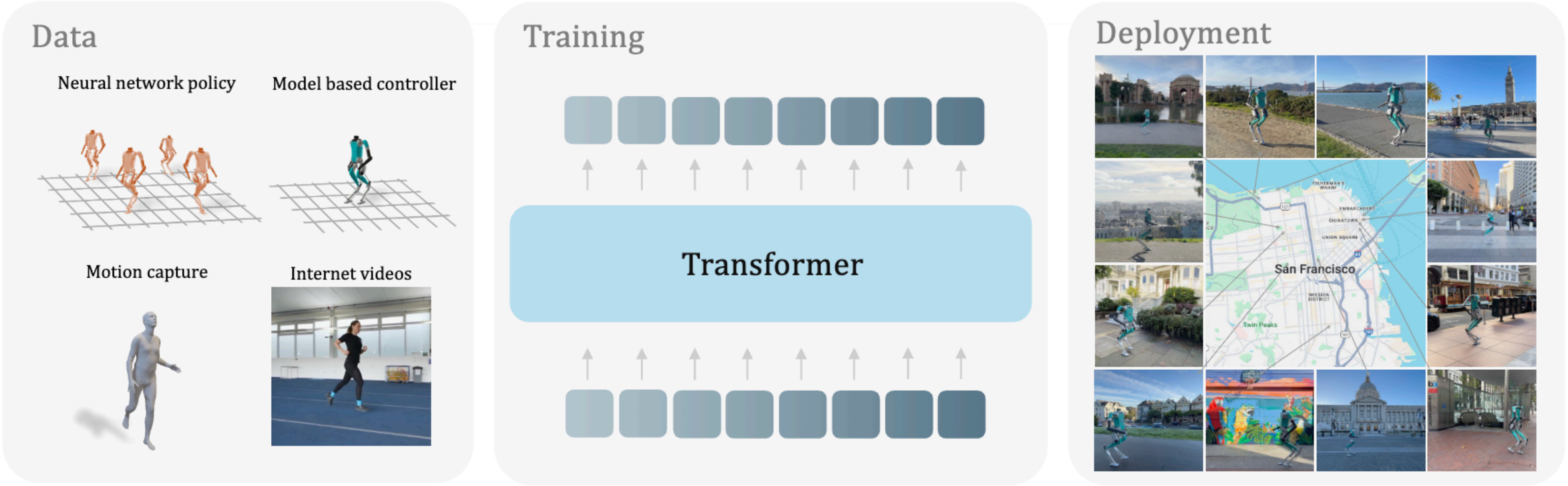


[Zhao et al., "ALOHA Unleashed: A Simple Recipe for Robot Dexterity", CoRL 2024]

Behavior cloning for object manipulation

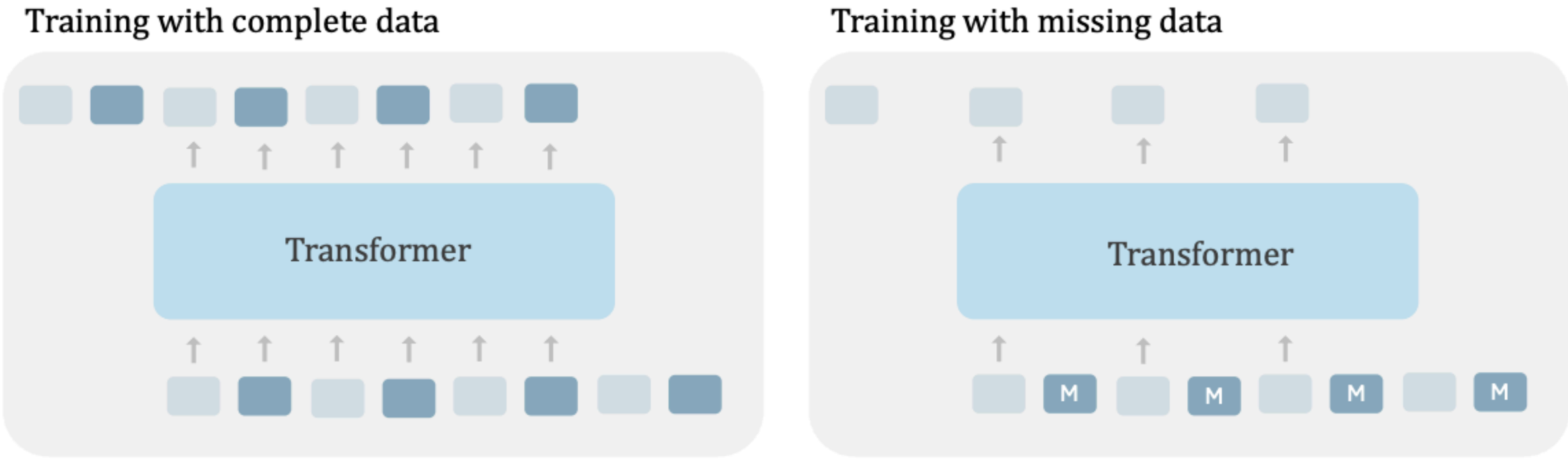
Behavior cloning for object manipulation

Generative modeling for humanoid locomotion



[Radosavovic et al., "Humanoid Locomotion as Next Token Prediction", 2024]

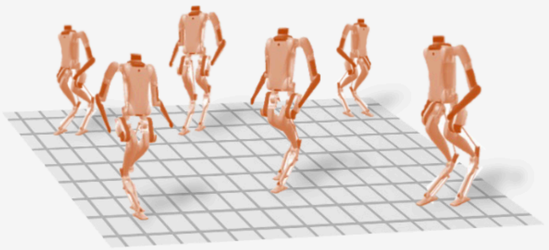
Generative modeling for humanoid locomotion



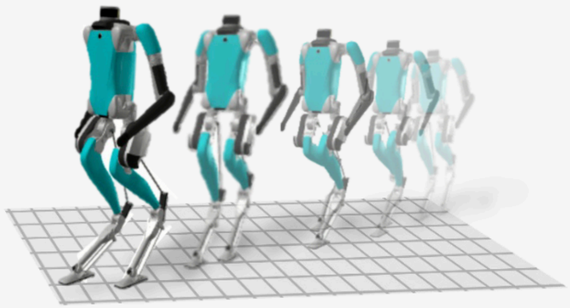
[Radosavovic et al., "Humanoid Locomotion as Next Token Prediction", 2024]

Next-token prediction from many sources

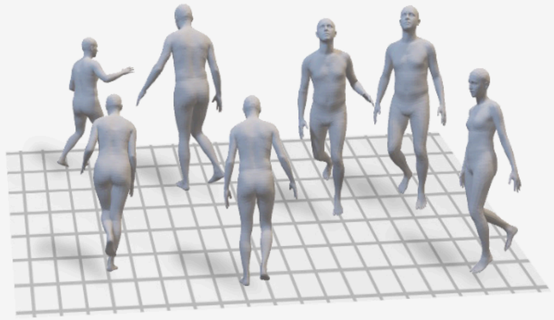
Neural Net Controller



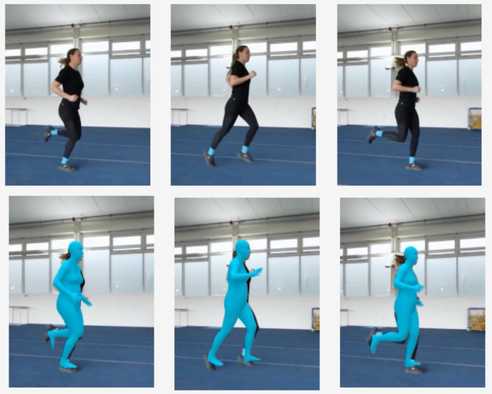
Model based Controller



MoCap



Internet Videos



[Radosavovic et al., “Humanoid Locomotion as Next Token Prediction”, 2024]

What about video generators?

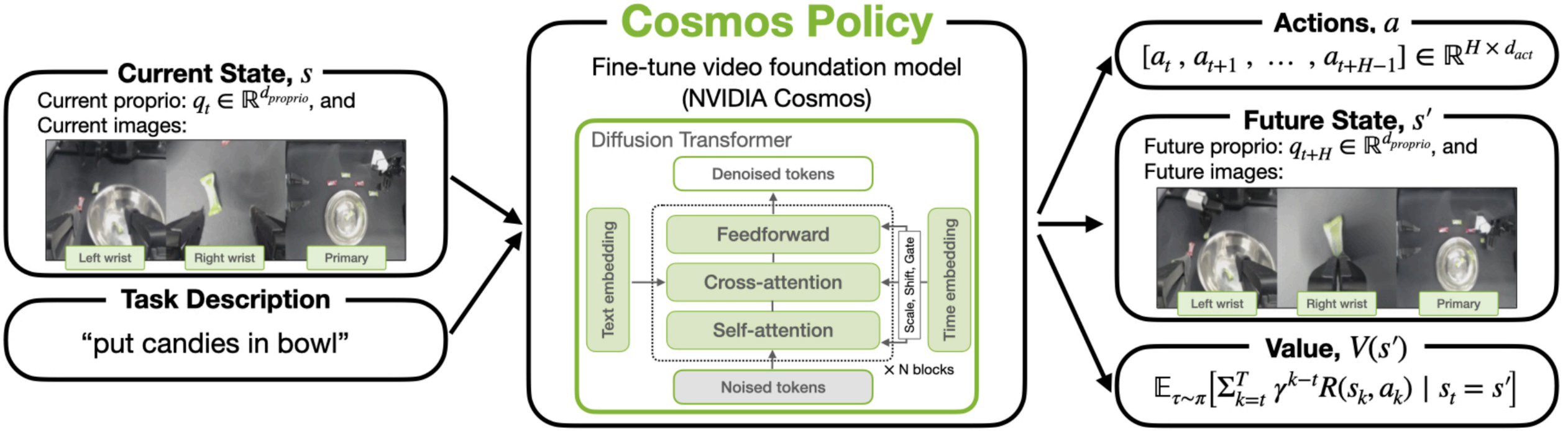
"A robotic arm on a table in a robotics lab folds a T-shirt”

Source: Veo 3

What about video generators?

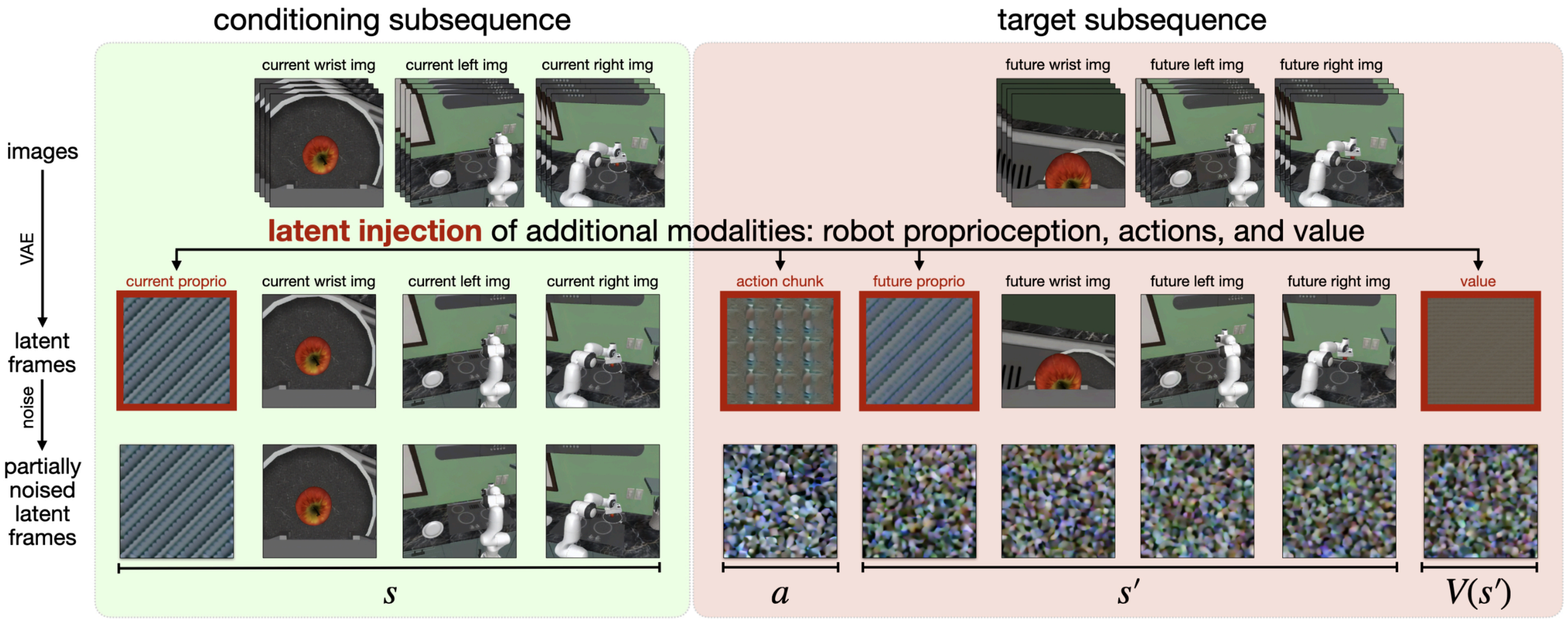
- These models produce good-looking videos. Can they simulate the physical world, too?
- How do you adapt them to robotics tasks?
- Where do *actions* fit in?

Finetuning video generators for robotics tasks



Kim et al., Cosmos Policy: Fine-Tuning Video Models for Visuomotor Control and Planning, 2026

Finetuning video generators for robotics tasks

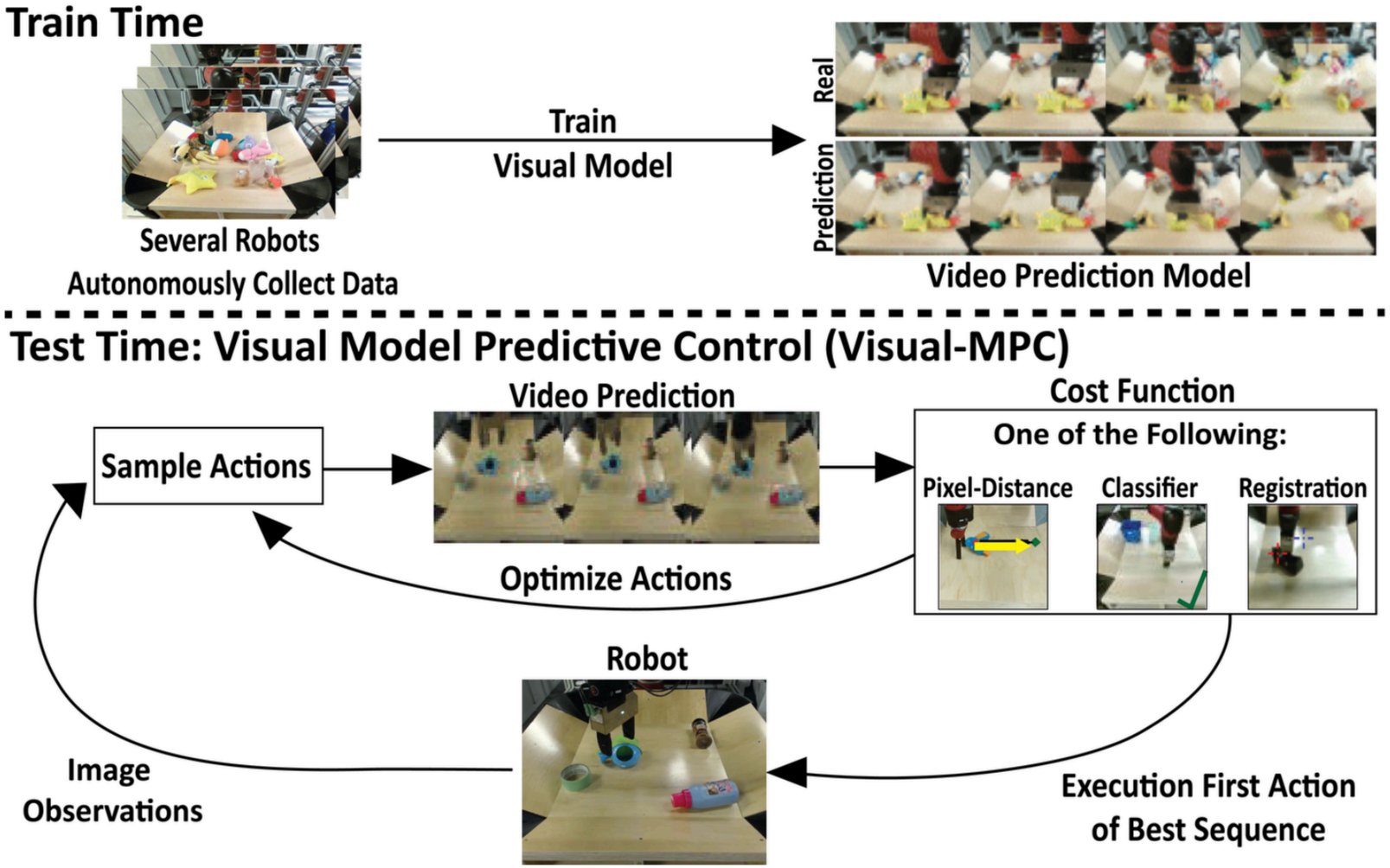


Kim et al., Cosmos Policy: Fine-Tuning Video Models for Visuomotor Control and Planning, 2026

World modeling

- Can we do it without finetuning? This would be more scalable.
- What if we could train an action-conditioned future predictor: $p_\theta(\mathbf{x}_{t+1} \mid \mathbf{x}_t, \mathbf{a}_t)$:
  - Given my current visual observation $\mathbf{x}_t$ and I take action $\mathbf{a}_t$, what will the future image image $\mathbf{x}_{t+1}$ be?
- This is sometimes known as a *world model* [Ha & Schmidhuber, 2018], though often this term is used imprecisely in the field.

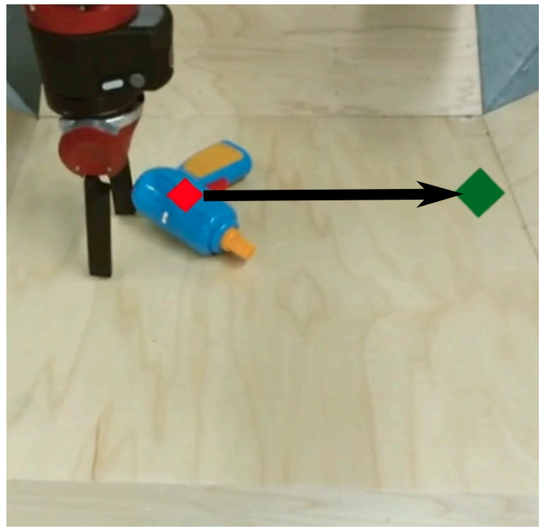
Visual model-based control



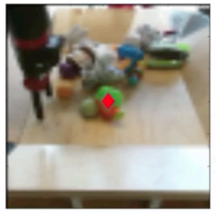
[Ebert et al., "Visual Foresight: Model-Based Deep Reinforcement Learning for Vision-Based Robotic Control", 2018]

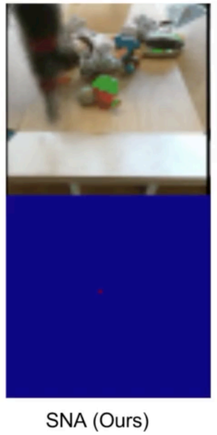
Visual model-based control (Visual MPC)

- If I know $p_\theta(\mathbf{x}_{t+1} \mid \mathbf{x}_t, \mathbf{a}_t)$, can I use it to directly control a robot?
- Search for sequences of actions that achieve our goal.
- Plug in different actions, observe many possible futures. Execute the actions that succeed.



Pushing task

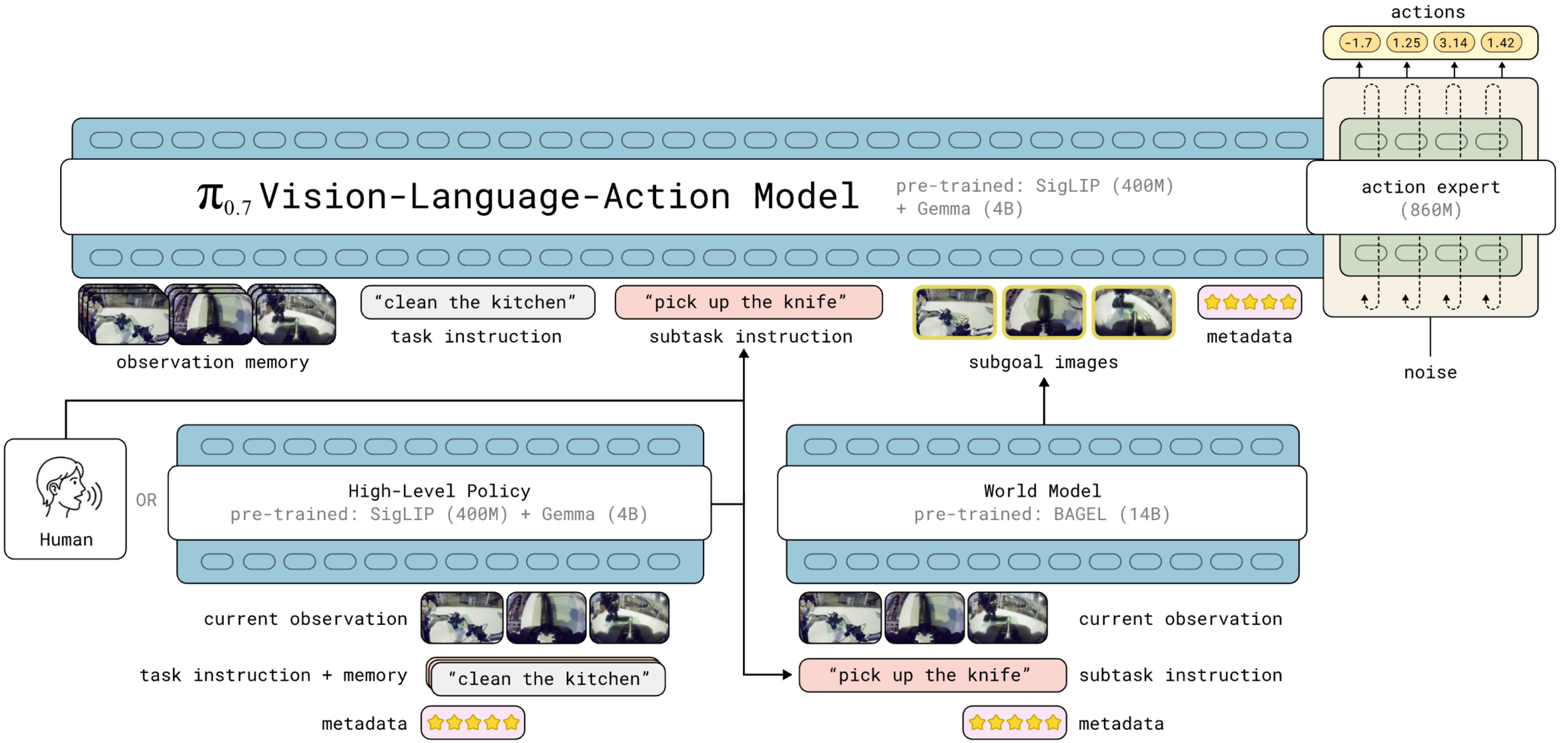




Future prediction

From [Ebert et al., "Self-Supervised Visual Planning with Temporal Skip Connections" 2017]

World models as part of behavior cloning models

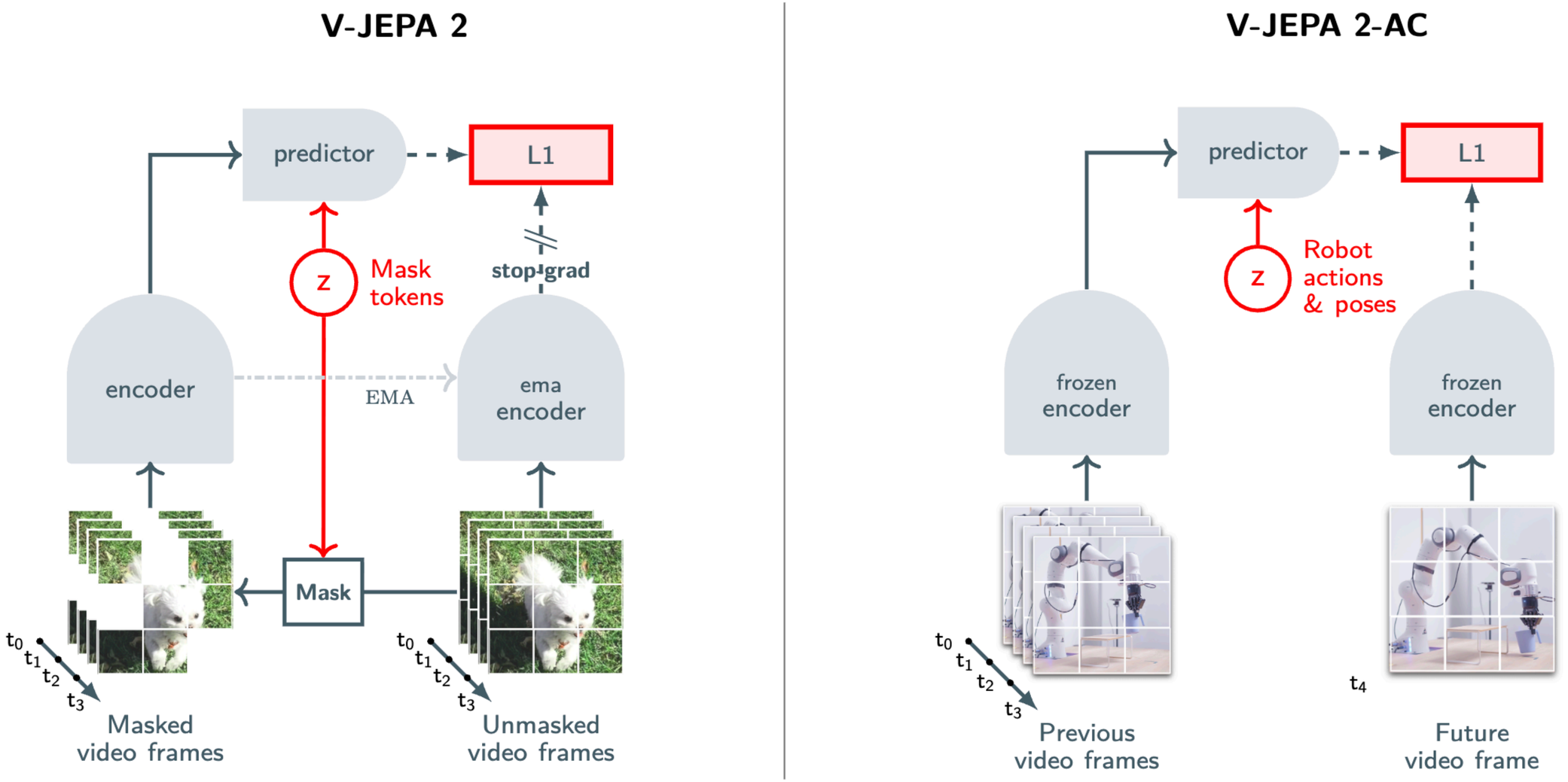


Source: [Physical Intelligence, $\pi_{0.7}$: a Steerable Generalist Robotic Foundation Model with Emergent Capabilities]

World models in latent space

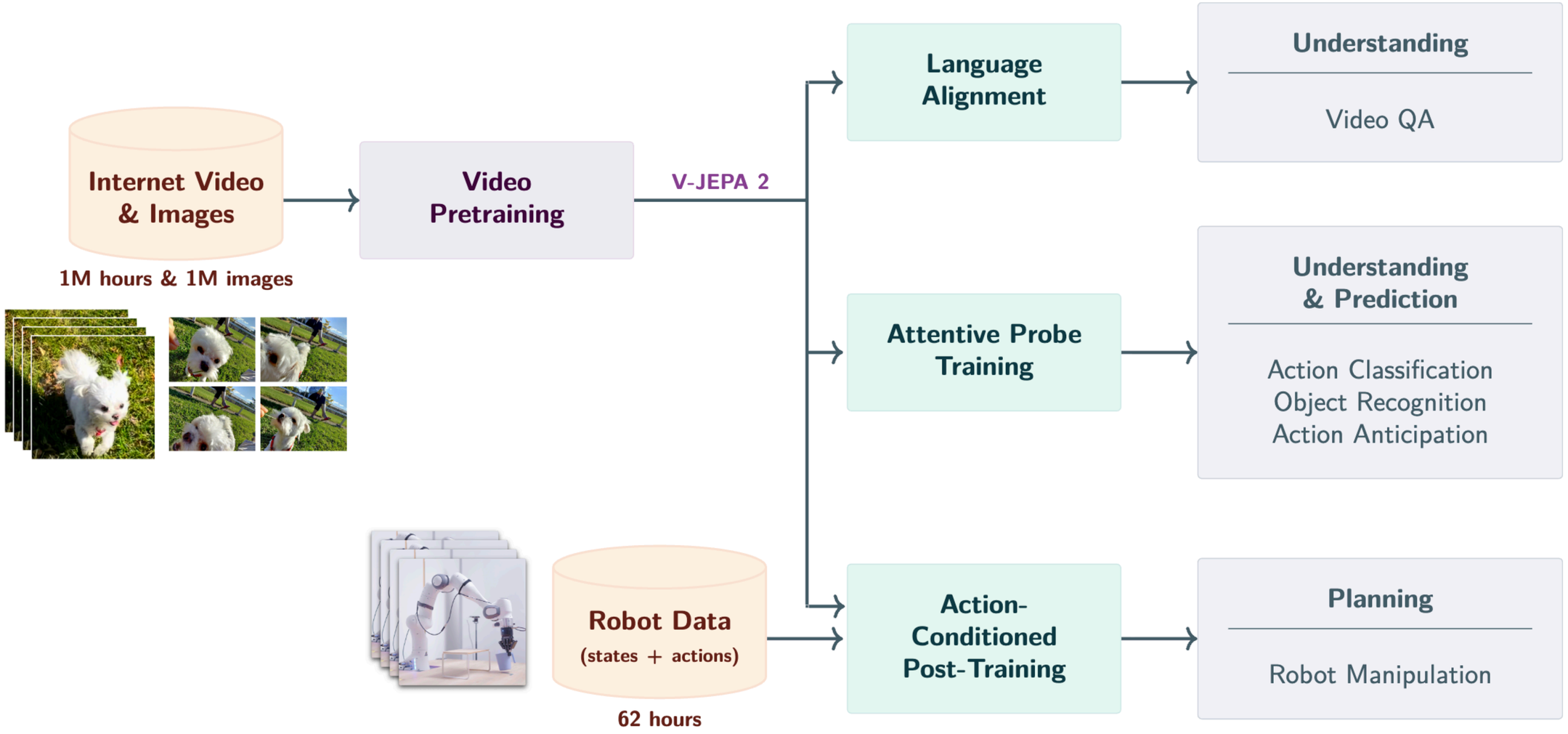
- Predicting raw sensory data is very expensive (e.g., state-of-art open source video generators take 30+ mins to generate a 10 sec. video on one GPU).
- These raw signals might not capture task-relevant information.
- Idea: predict the future in a feature space instead.

World models in latent space



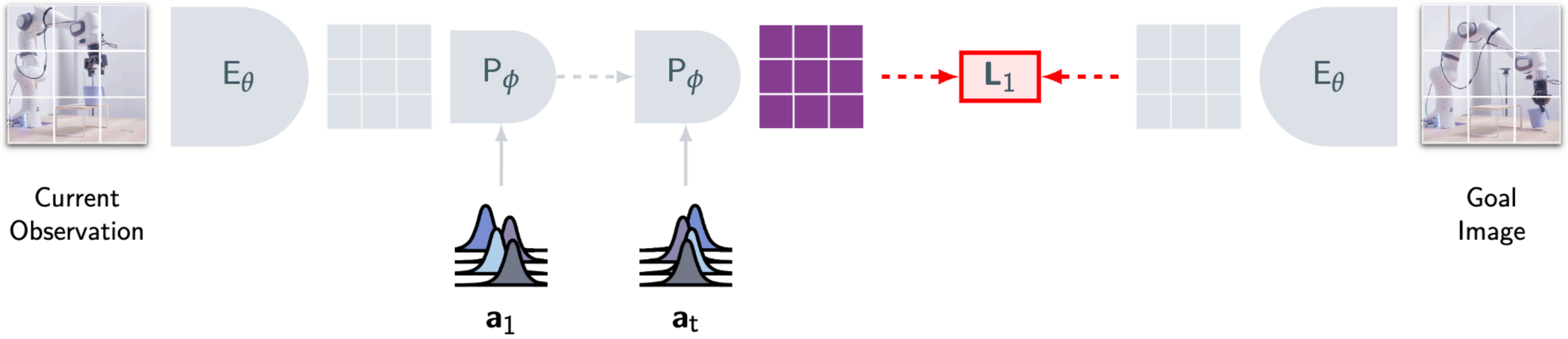
Source: [Assran et al. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning]

World models in latent space



Source: [Assran et al. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning]

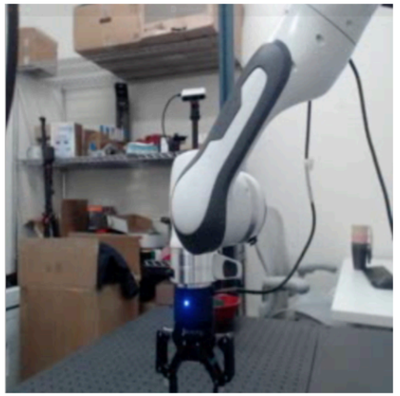
Model predictive control in latent space



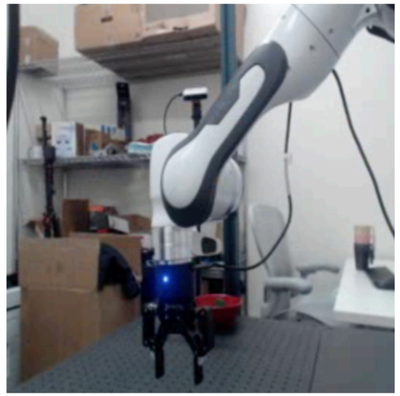
Source: [Assran et al. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning]

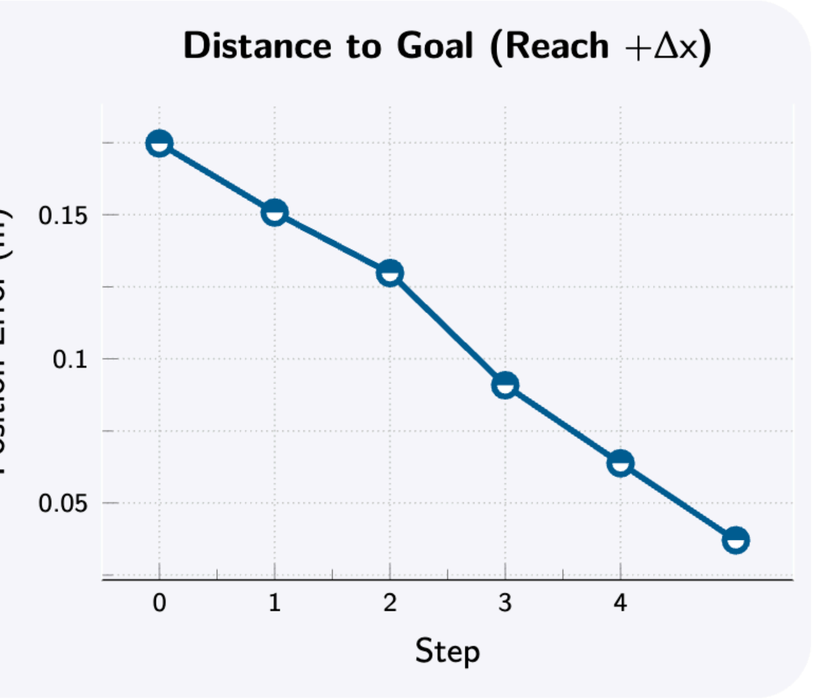
Start Frame Goal Frame

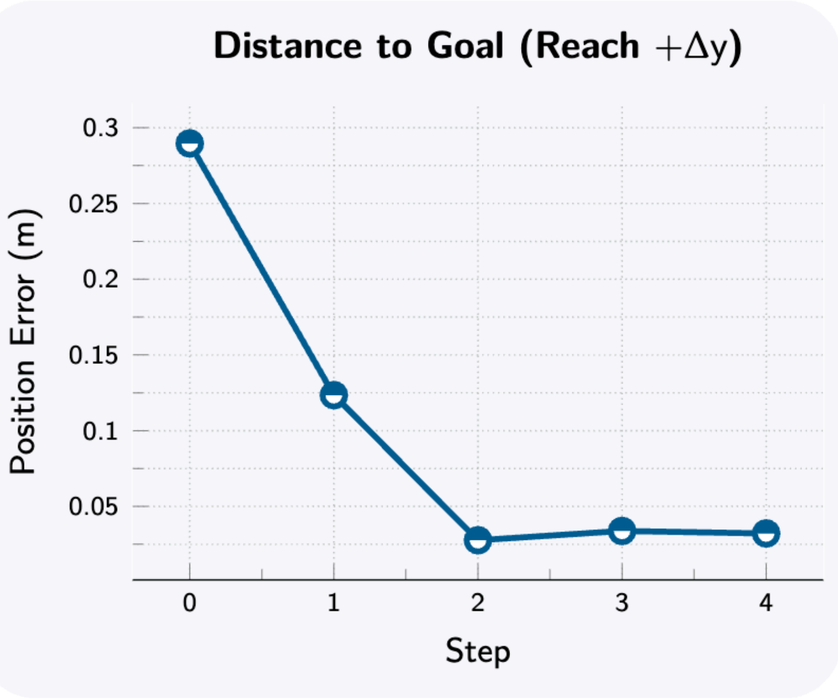
Start Frame Goal Frame

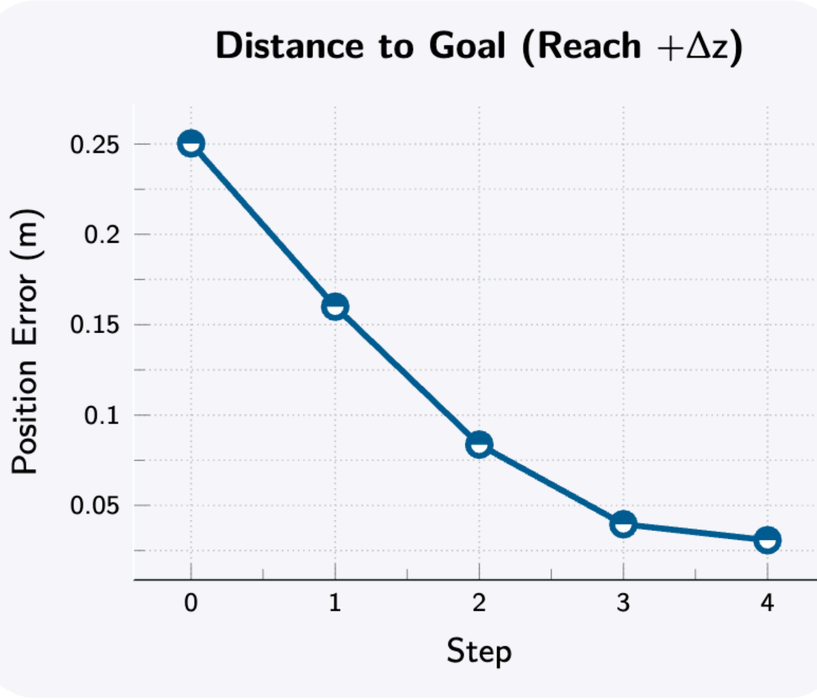


Start Frame Goal Frame

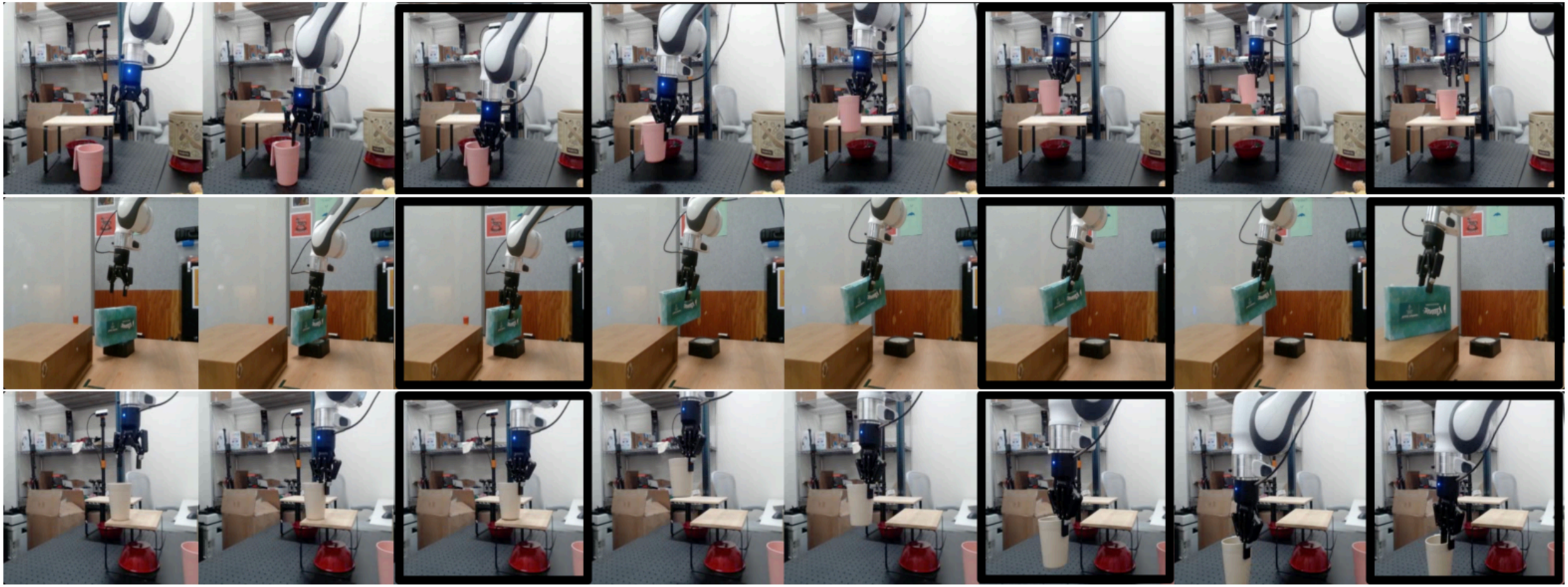








Source: [Assran et al. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning]

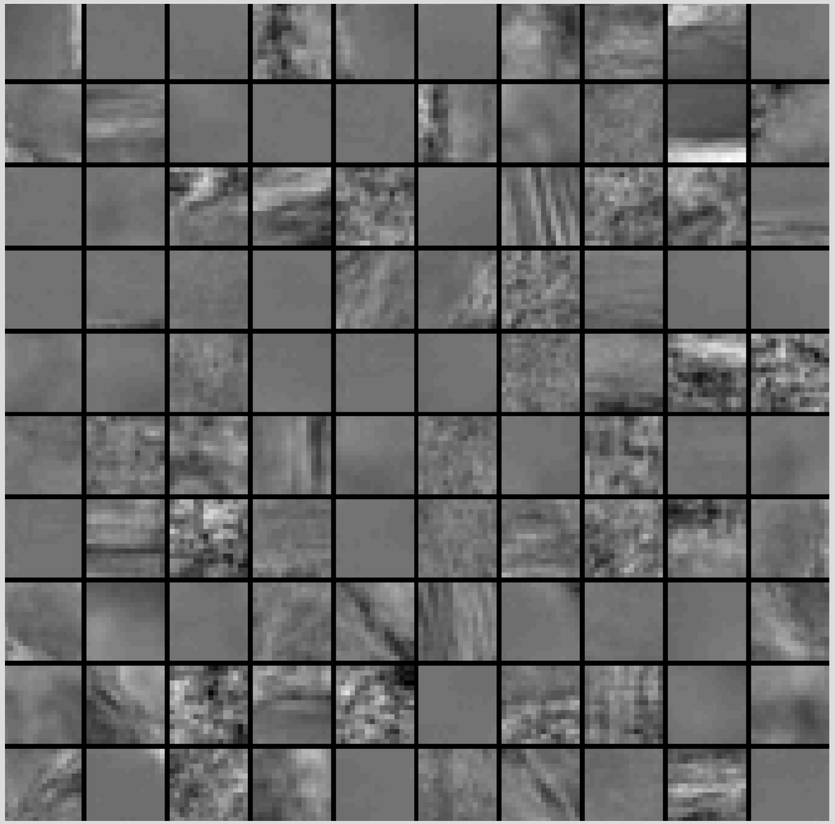


Source: [Assran et al. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning]

Quick course recap

Simple generative models and probability fundamentals let us fit complex distributions

Simple probabilistic models



Gaussian mixture model for image patches

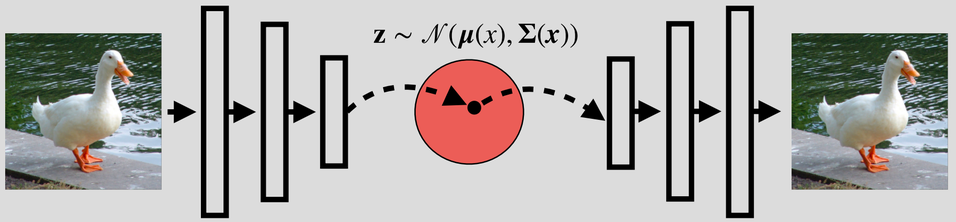
Source: [Zoran & Weiss, “Natural Images, Gaussian Mixtures and Dead Leaves”, 2012]

We can combine these ideas with neural networks to create powerful generative models.

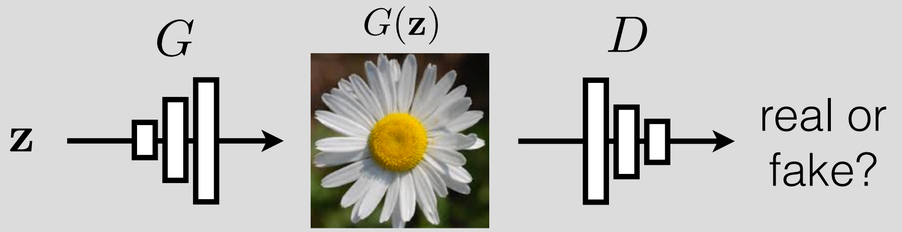
For example:

latent variable models

Latent variable models



VAEs



GANs

Figure source: Isola, Torralba, Freeman

We can combine these ideas with neural networks to create powerful generative models.

For example:

and diffusion models

Diffusion model from PS3

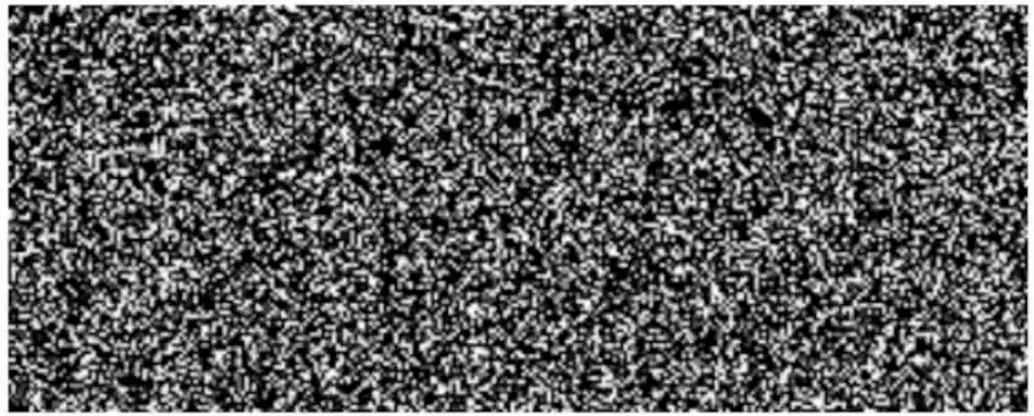


Figure source: Isola, Torralba, Freeman

You can adapt these models to downstream tasks through fine-tuning and RL.

And you can scale them up.

But there’s still a lot to do to make them safe, robust, fast, etc.

The City
BY C. P. CAVAFY
TRANSLATED BY EDMUND KEELEY
[Poem text omitted]

SOME TREES
John Ashbery
[Poem text omitted]

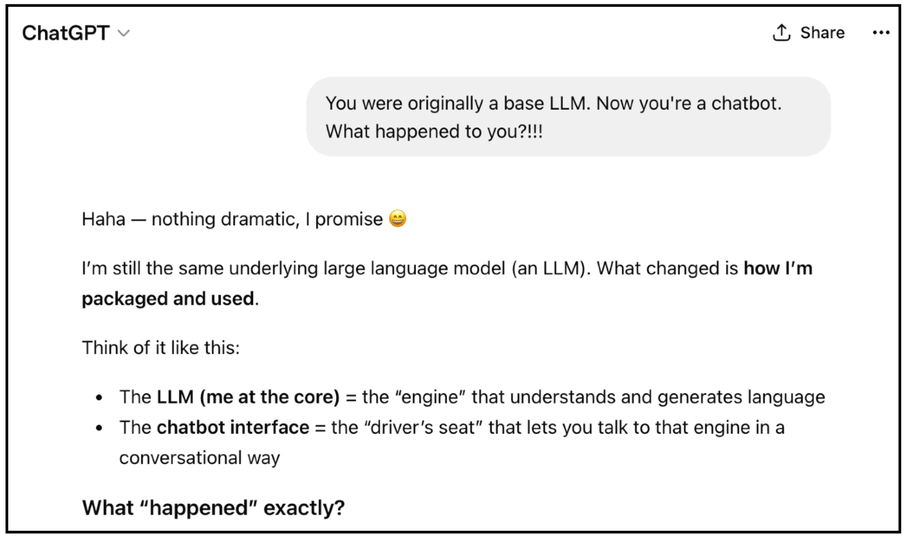
Shadows on the Way
Wallace Stevens

I must have shadows on the way
If I am to walk I must have
Each step taken slowly and alone
To have it ready made

And I must think in lines of grey
To have dim thoughts to be my guide
Must look on blue and green
And never let my eye forget
That color is my friend
And purple must surround me too

The yellow of the sun is no more
Intrusive than the bluish snow
That falls on all of us. I must have
Grey thoughts and blue thoughts walk with me
If I am to go away at all.

Base language model



… to Chatbot

Next class: project presentations