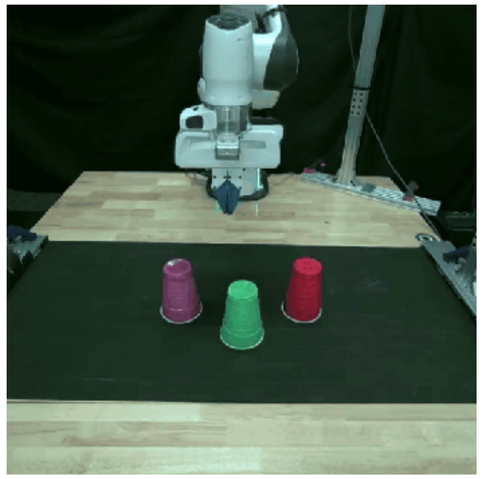
Force-Conditioned Video Models as Physical World Models for Agents

Presented by: Chen Sun

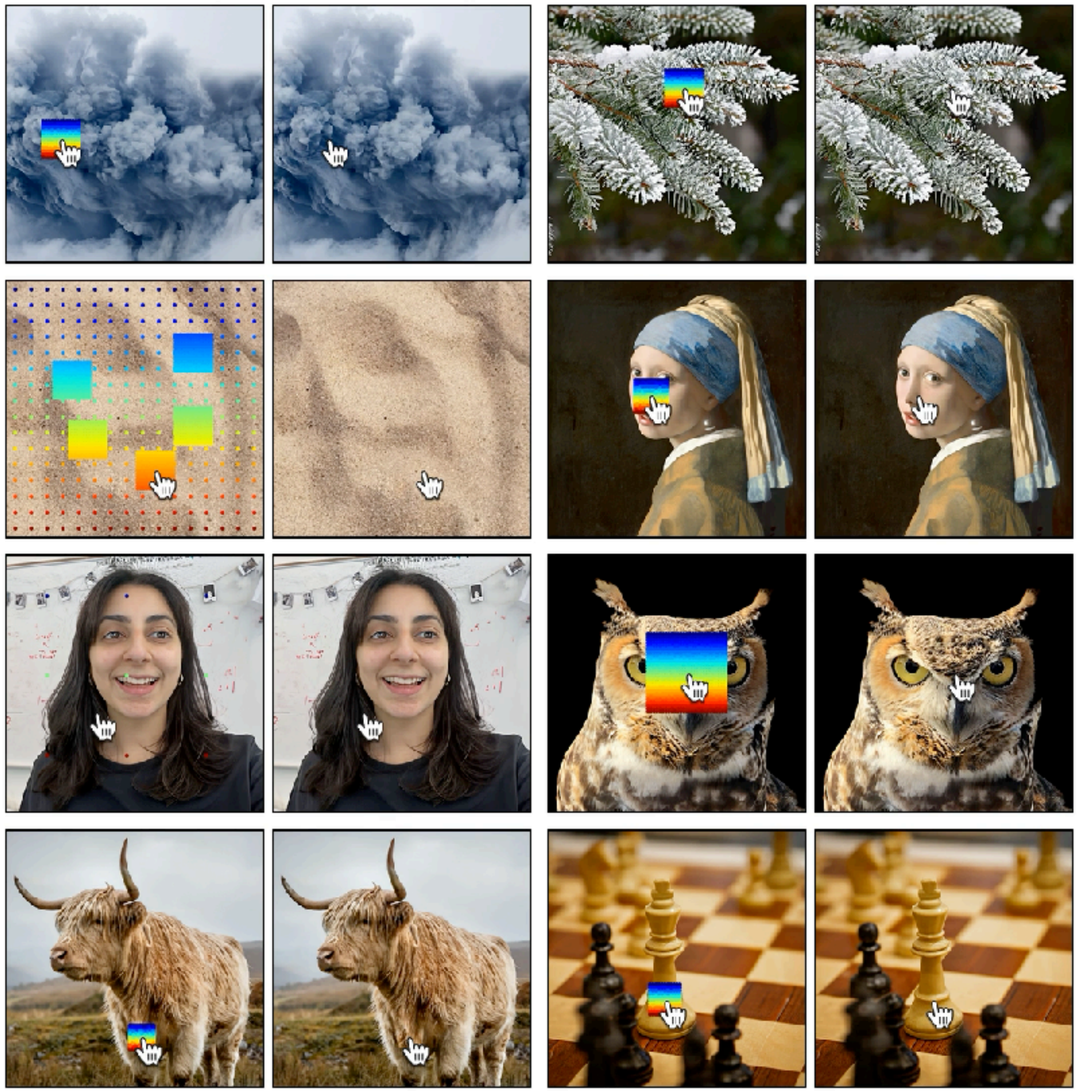
Projects led by (and slides credit to): Nate Gillman, Calvin Luo, Zilai Zeng
Collaborators: Arjan Chakravarthy, Charles Herrmann, Daksh Aggarwal, Deqing Sun, Evan Luo, Michael Freeman, Mingxi Jia, Yilun Du, Yinghua Zhou, Zitian Tang



You use pretty pictures to illustrate ideas, I help generate them!

FLUID: SCALING AUTOREGRESSIVE TEXT-TO-IMAGE GENERATIVE MODELS WITH CONTINUOUS TOKENS

Lijie Fan[1,*] **Tianhong Li**[2,*] **Siyang Qin**[1,†] **Yuanzhen Li**[1] **Chen Sun**[1]
Michael Rubinstein[1] **Deqing Sun**[1] **Kaiming He**[2] **Yonglong Tian**[1,*]
[1]Google DeepMind [2]MIT [*] equal contribution, project lead [†] equal contribution

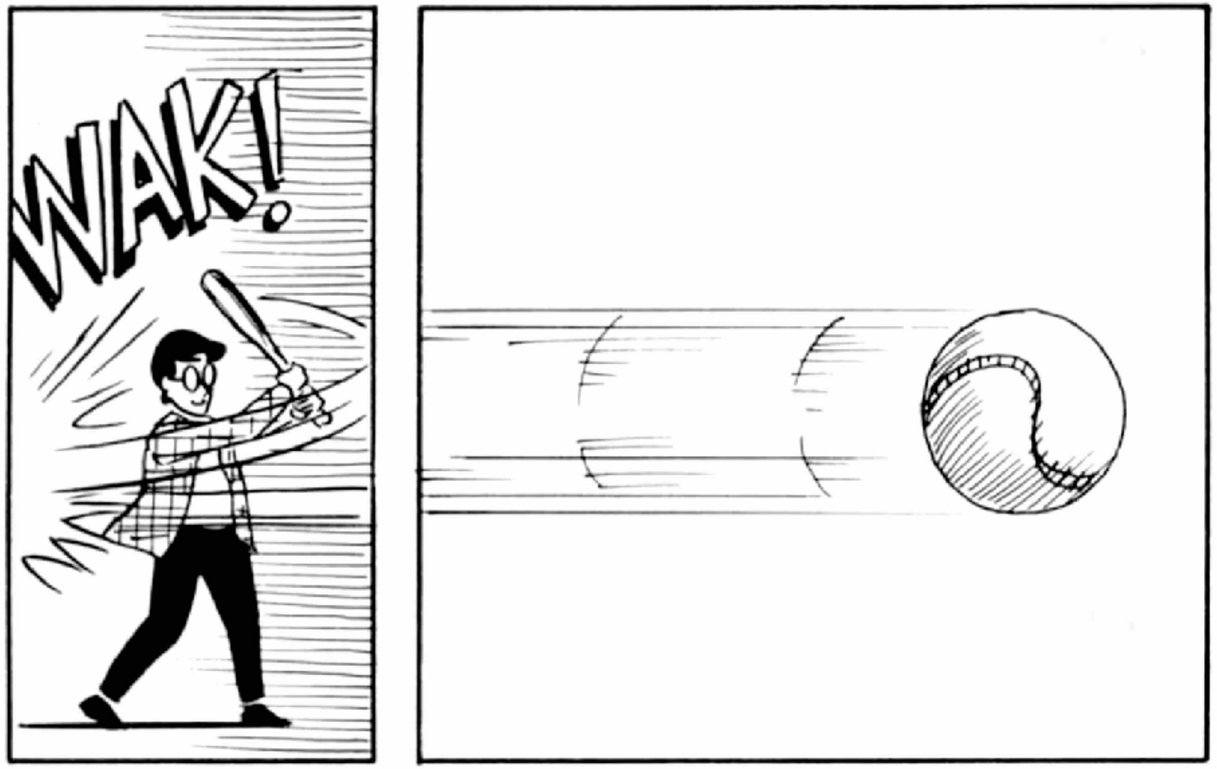


From Pretty Pictures / Videos to “World Models”

Action: “Hit”

Ha and Schmidhuber, World Models, NeurIPS 2018.

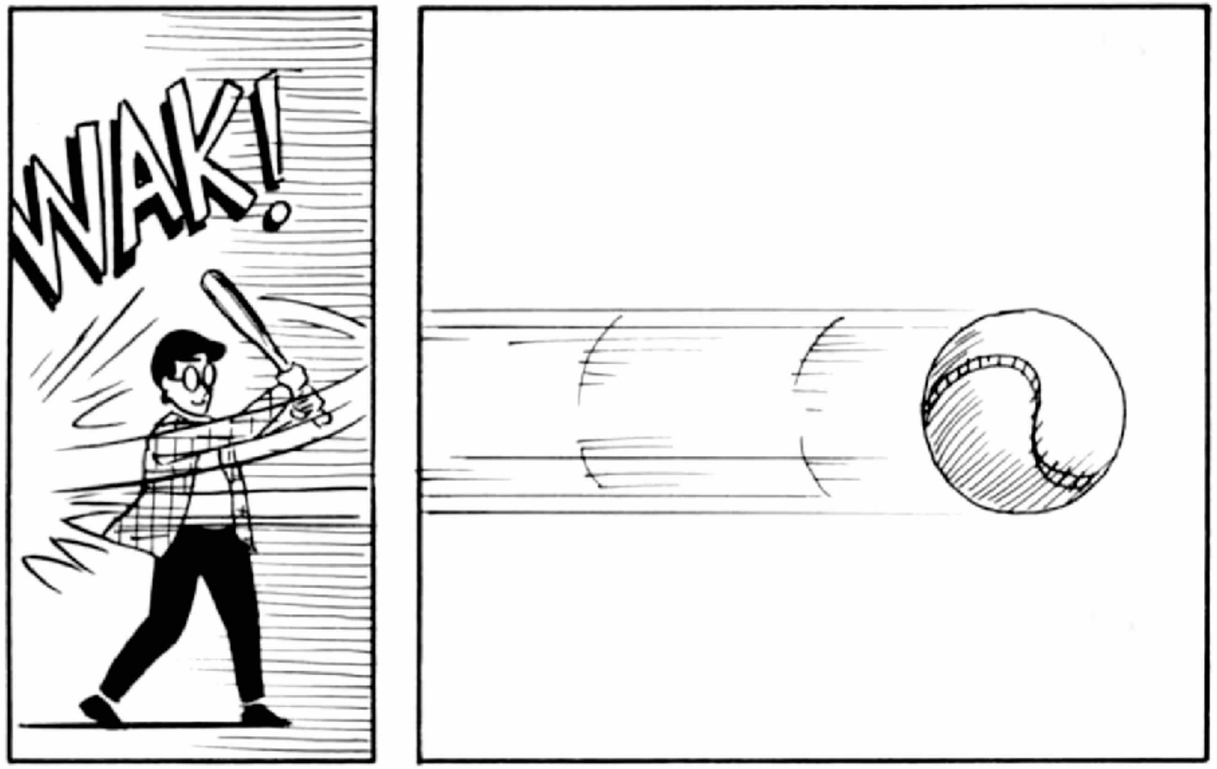
From Pretty Pictures / Videos to “World Models”



$$s_{t+1} \sim p_\theta(s_{t+1} \mid s_{t-H:t}, a_{t-H:t})$$

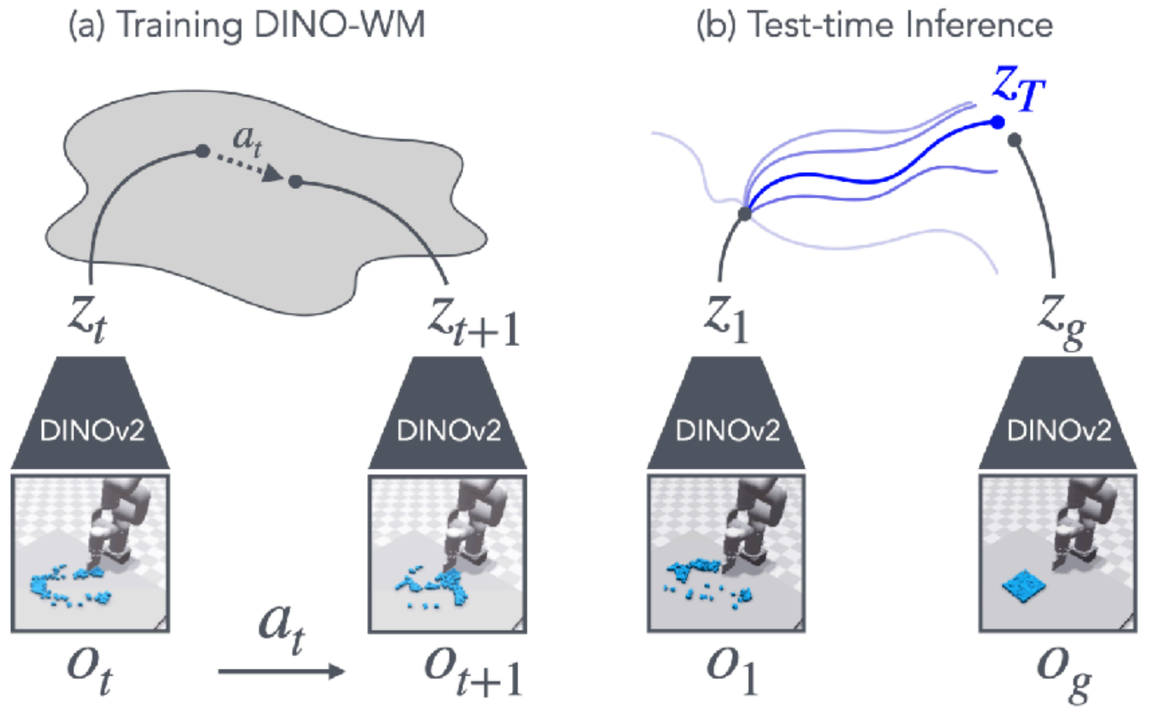
Ha and Schmidhuber, World Models, NeurIPS 2018.

From Pretty Pictures / Videos to “World Models”



$$s_{t+1} \sim p_\theta(s_{t+1} \mid s_{t-H:t}, a_{t-H:t})$$

Ha and Schmidhuber, World Models, NeurIPS 2018.



Zhou et al., DINO-WM, arXiv:2411.04983

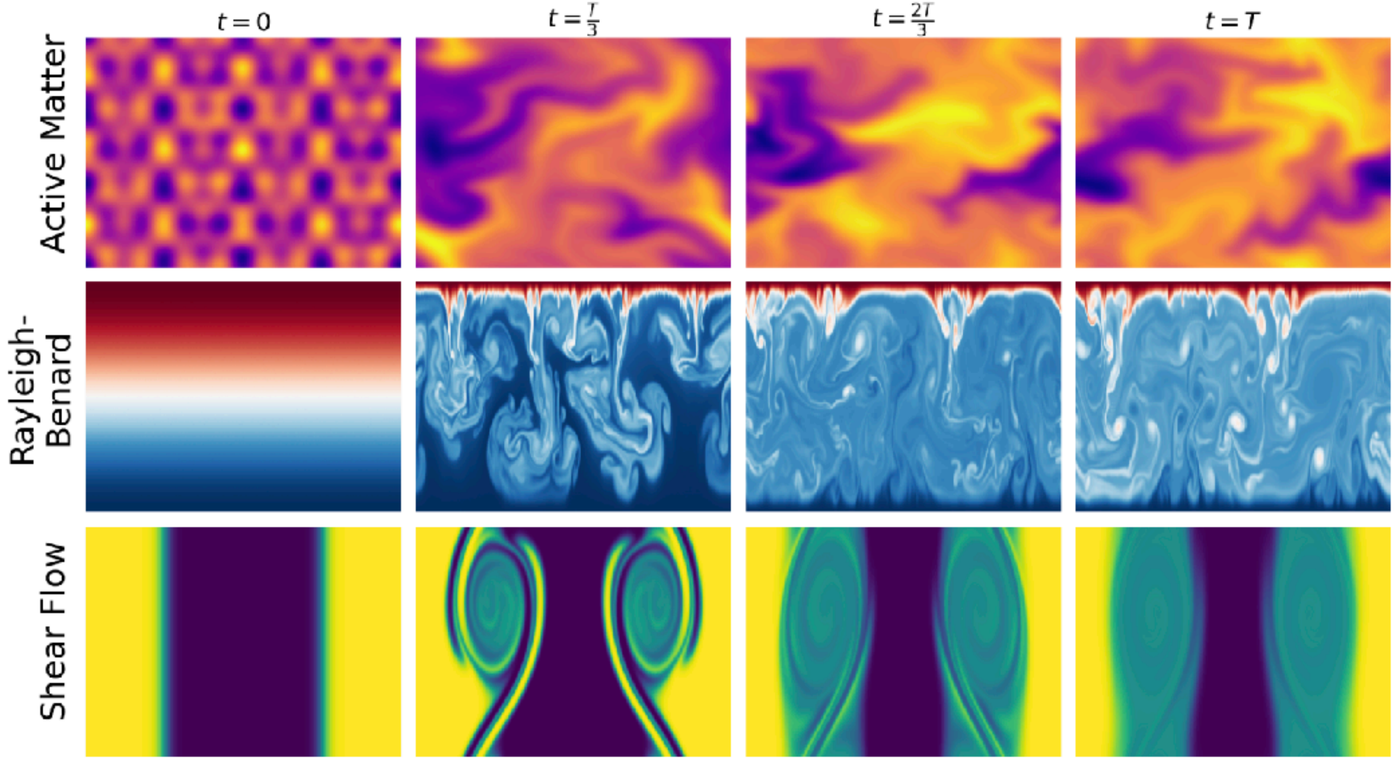
“World Models” Might be Relevant to (Some of) You

REPRESENTATION LEARNING FOR SPATIOTEMPORAL PHYSICAL SYSTEMS

Helen Qu[1†] **Rudy Morel**[1] **Michael McCabe**[1,4] **Alberto Bietti**[1] **François Lanusse**[2]
Shirley Ho[1,3,4] **Yann LeCun**[3]

The Polymathic AI Collaboration

[1]Flatiron Institute
[2]Université Paris-Saclay, Université Paris Cité, CEA, CNRS, AIM
[3]New York University
[4]Princeton University



“Showing up at the first conference for world models”

Why are we interested in video models?

X/KnightNemo

Lessons from LLMs

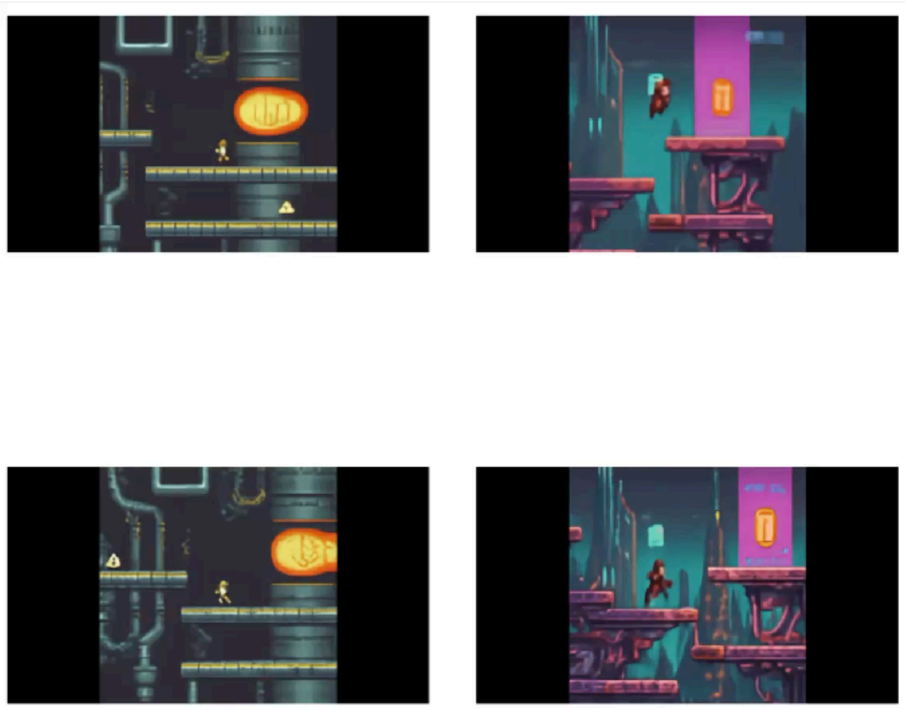
Wayve Gaia-1 (Hu et al.) 2023

Lessons from LLMs: The Merit of Data

Wayve Gaia-1 (Hu et al.) 2023

Waymo World Model (2026)

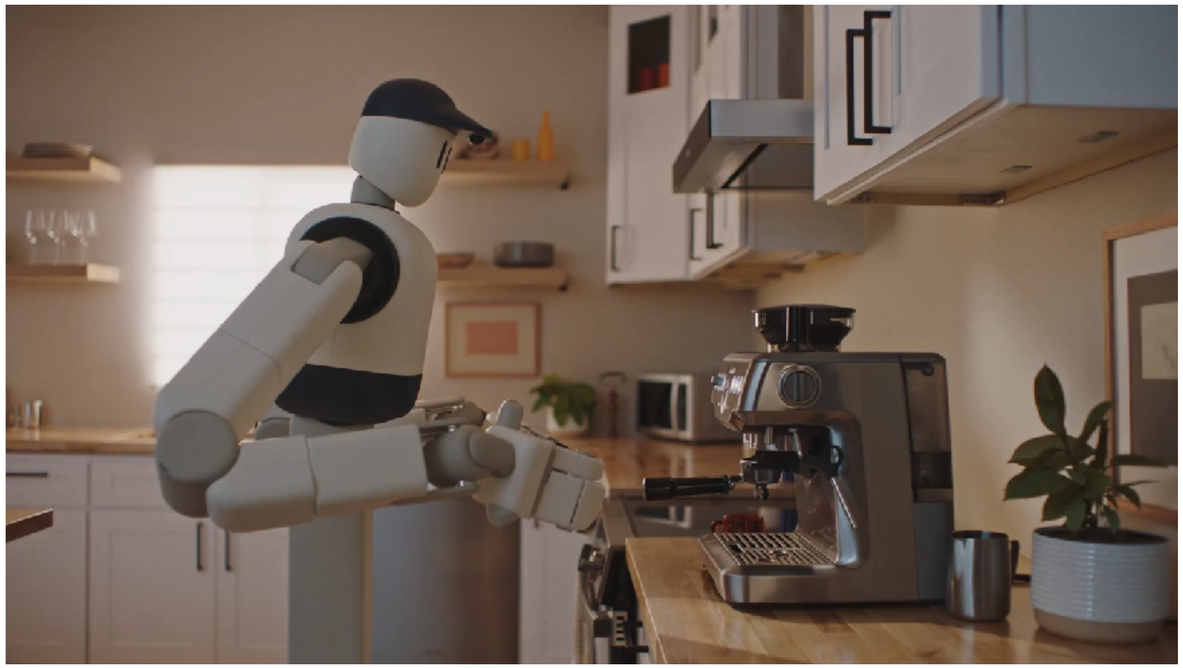
Lessons from LLMs: The Merit of Data



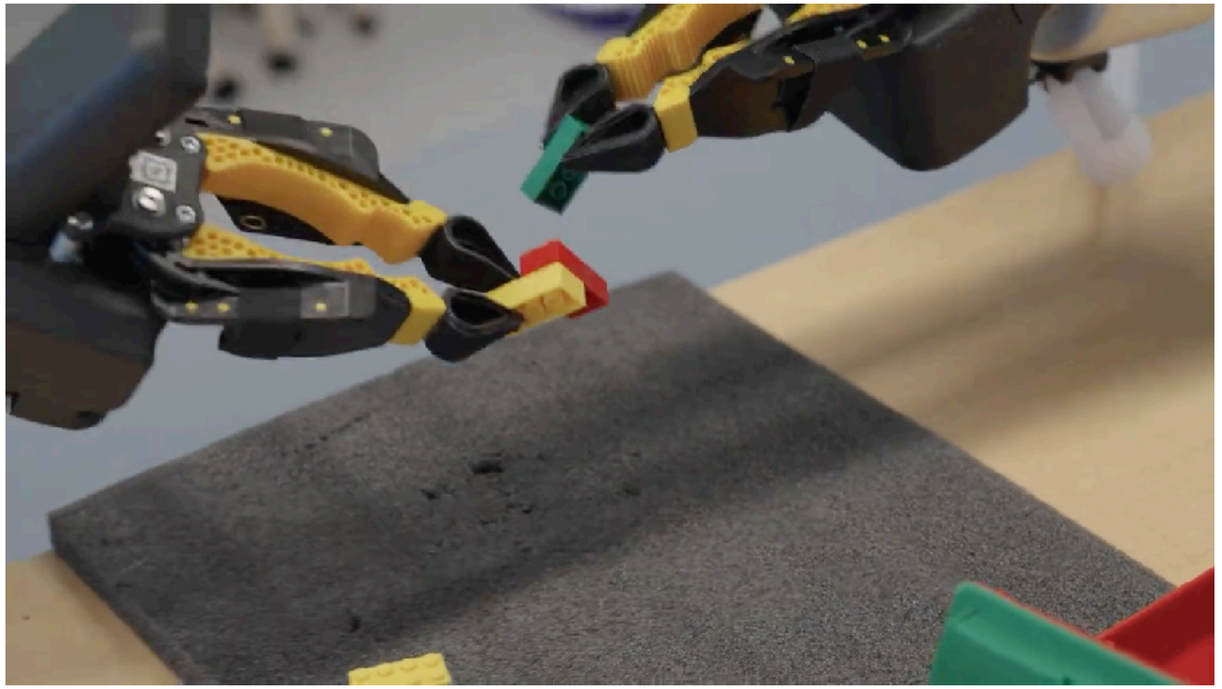
Genie: Generative Interactive Environments (2024)

Genie-3

Our Dream: Generalizable Decision Making in Physical Agents



Sunday Robotics

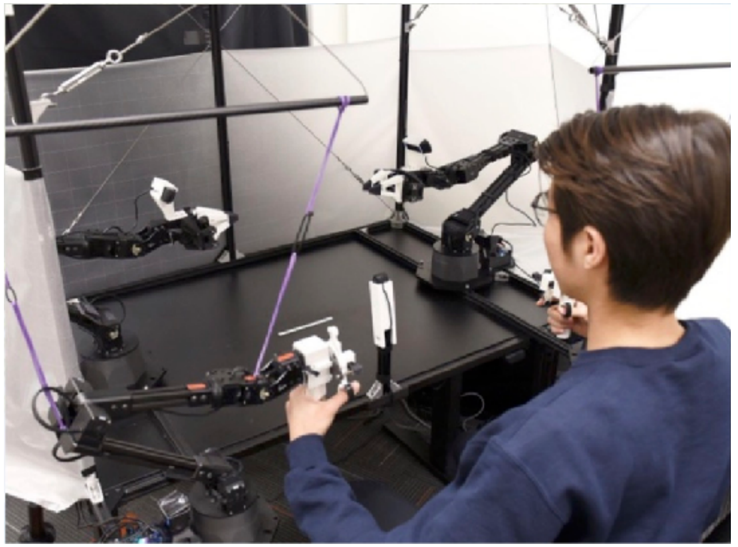


Generalist AI

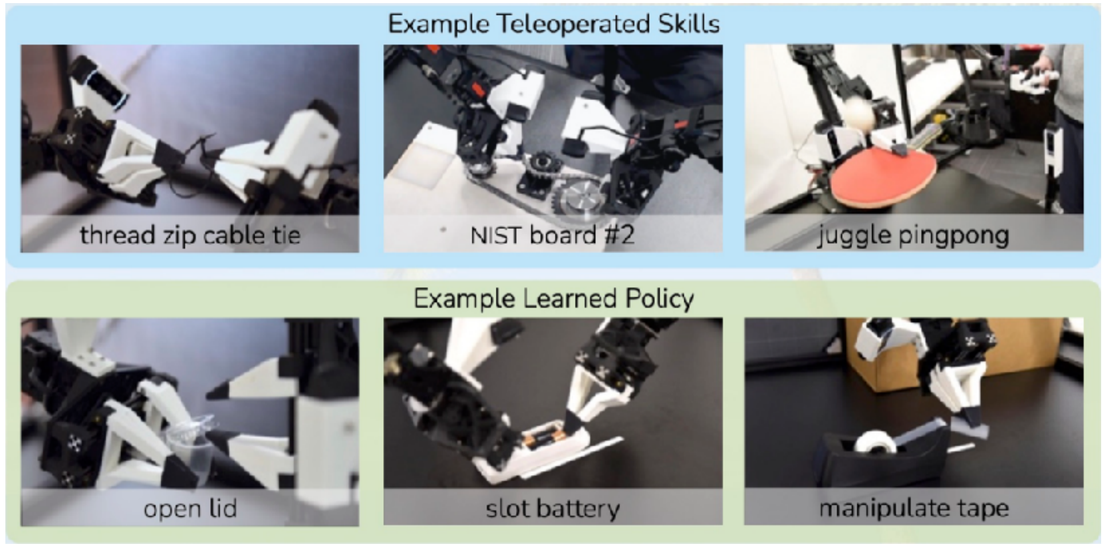
Questions: How to Leverage Human Data and Video Models?

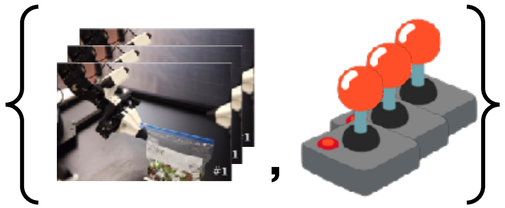
Learning by Imitating Expert Behaviors

Teleoperation



Deployment



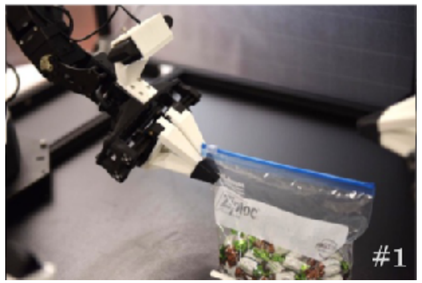


Behavior Cloning

$$\pi_{\boldsymbol{\theta}}(a \mid s)$$

Zhao et al. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. 2023

Policy Parameterization



$$\pi_{\boldsymbol{\theta}}(a_t \mid o_t)$$

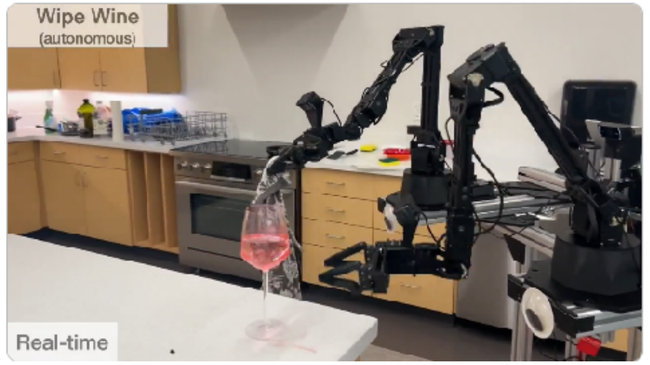
Visual Observation

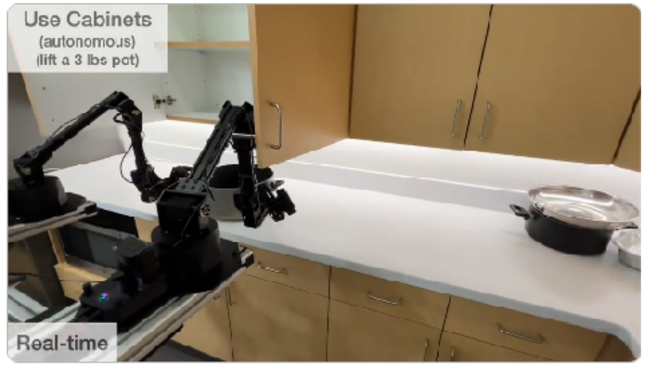
Action Sequence

Action-Predictive Policy

Learning by Imitating Expert Behaviors

Novel Tasks





?

?

?

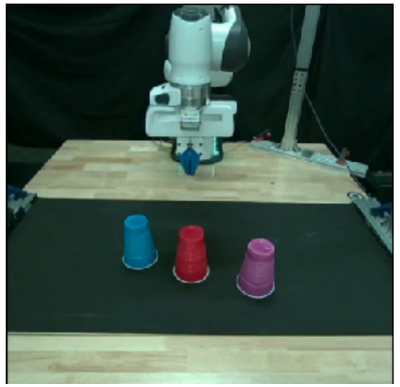
Direct Deployment

Limited, Static Skill Set

$$\pi_{\boldsymbol{\theta}}(a \mid s)$$

Fu et al. Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low-Cost Whole-Body Teleoperation. 2023

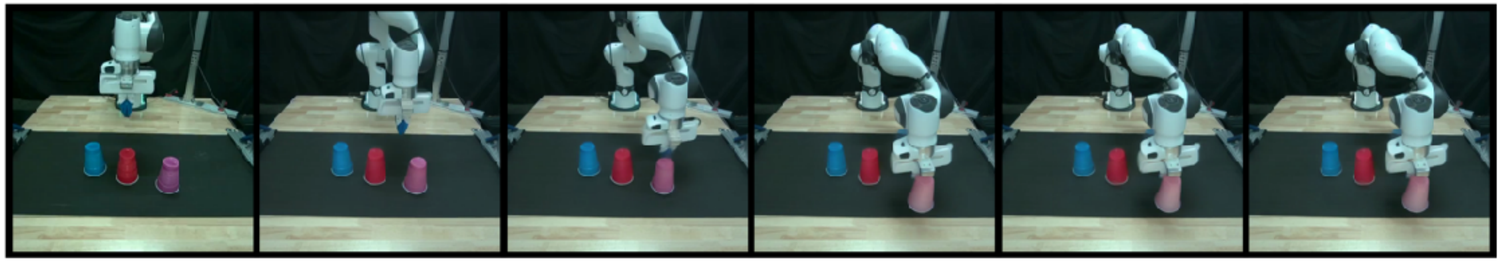
Policy Parameterization with Video Models



Task: Push Pink Cup

ϵ_θ

Video Generative Model



$f_{\text{inv}}(o_t, o_{t+1}) = a_t$

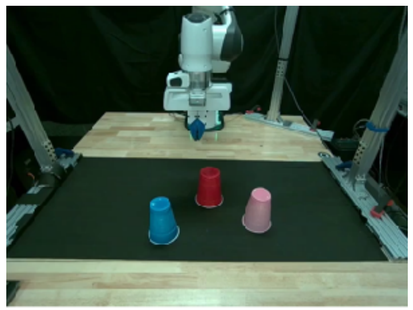
Inverse Dynamics Model

Action Sequence

“Visual Planner”

Integrating Web Video Prior

"Push the red cup"



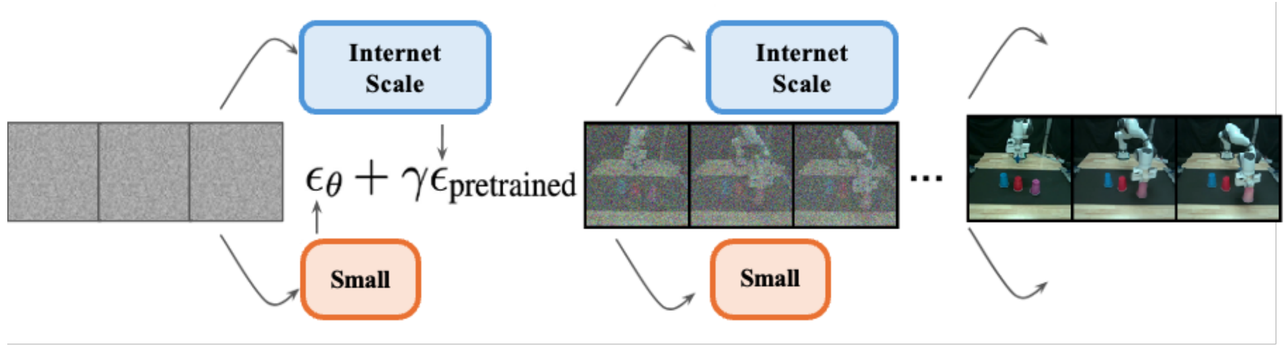
In-Domain Pretraining

ϵ_θ

"Push the purple mug to the left"

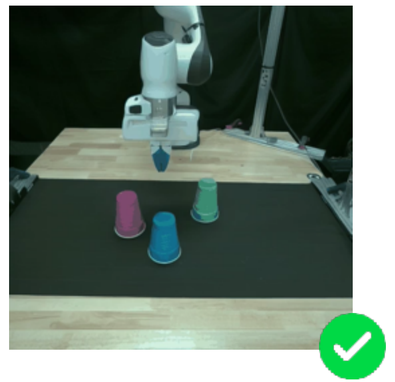
Internet-Scale Pretraining

$\epsilon_{\text{general}}$



"Push the purple cup"

Integrating Video Prior

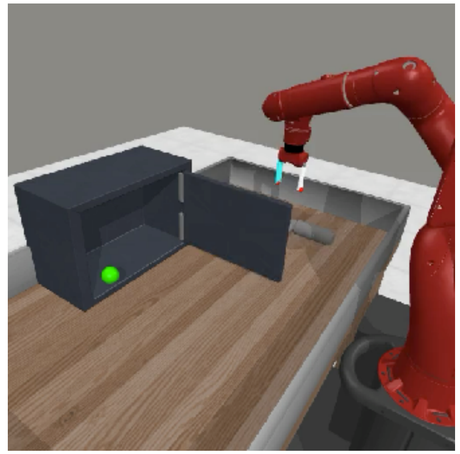


"Push the purple cup"

Luo et al. Solving New Tasks by Adapting Internet Video Knowledge. 2025
Yang et al. Probabilistic Adaptation of Text-to-Video Models. 2023

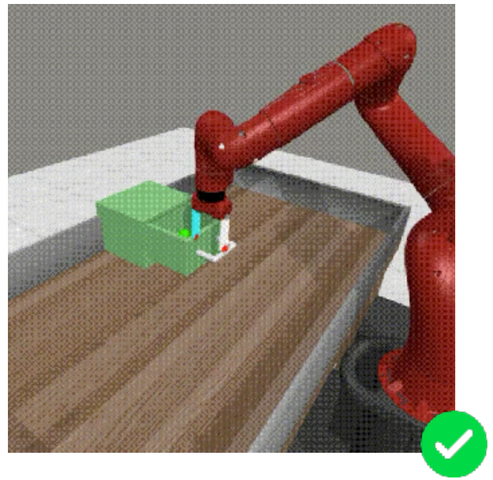
Limited Generalization Capability

Training Task

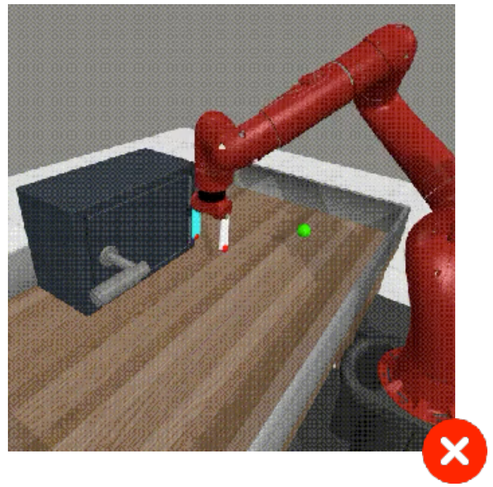


Door Close

Novel Tasks



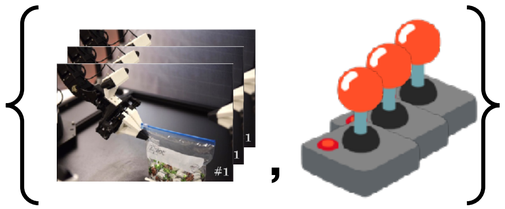
Drawer Close



Door Open

How To Tackle Novel Tasks?

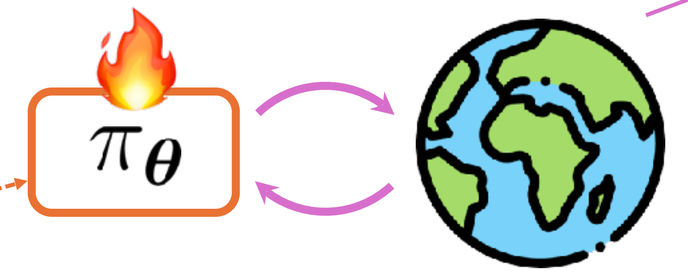
Offline Expert Demonstrations
from **Seen Tasks**

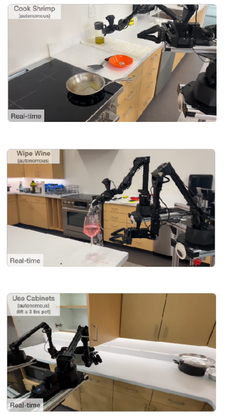


$$\pi_{\boldsymbol{\theta}}(a \mid s)$$

Pre-training

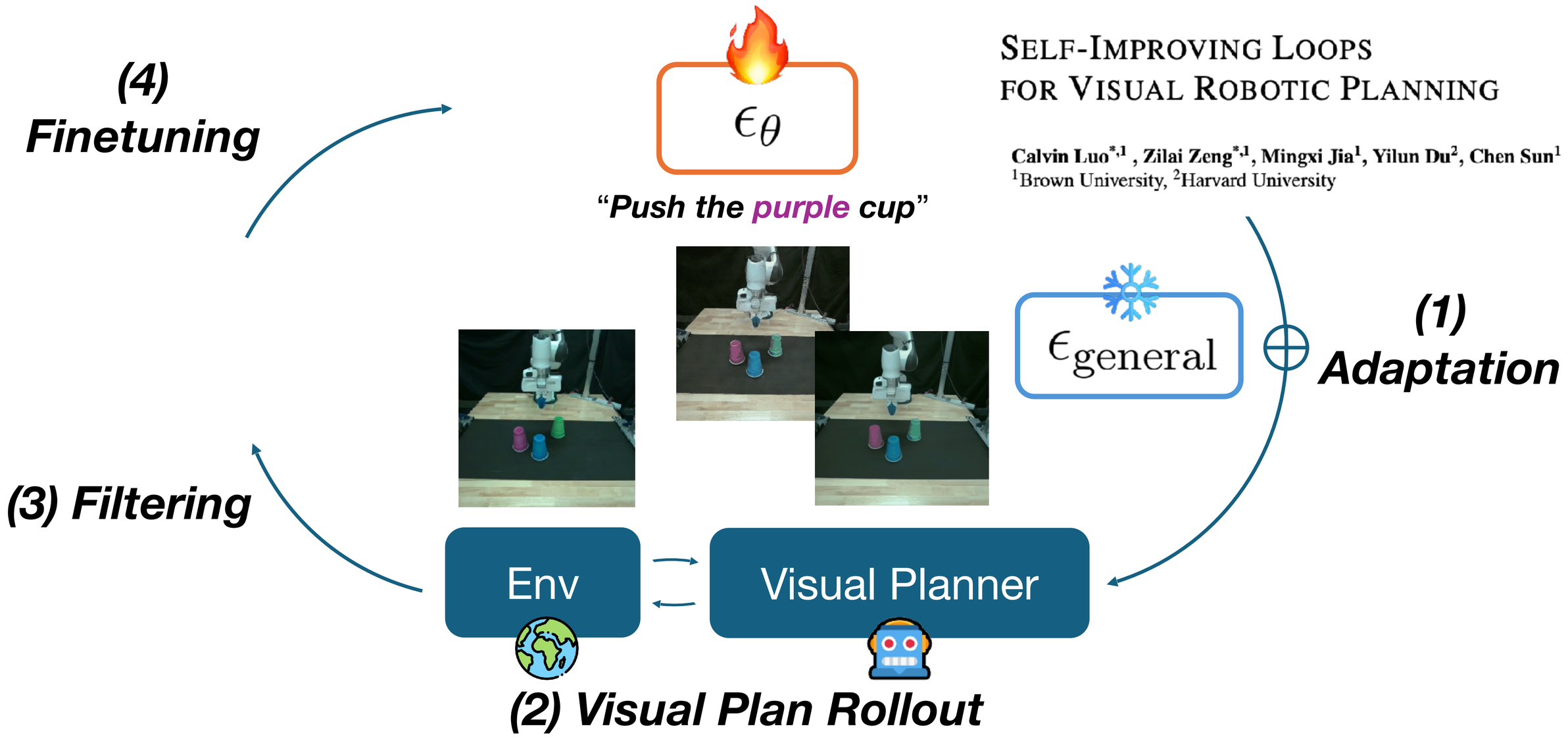
Online Interaction Experience
from **Novel Tasks**





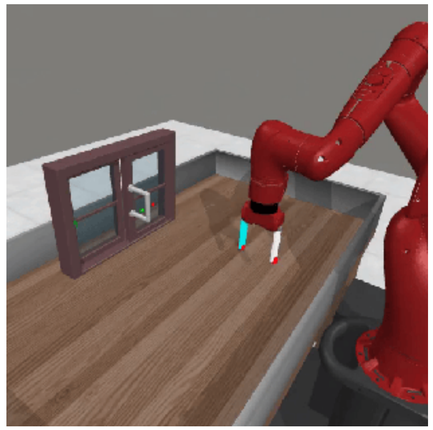
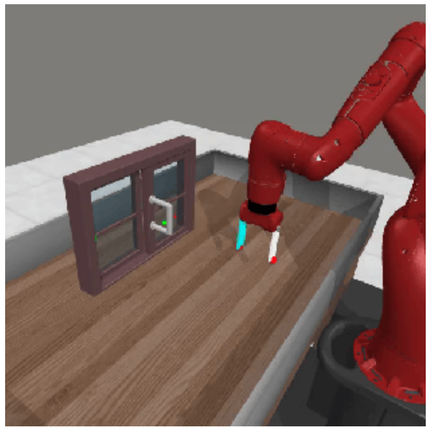
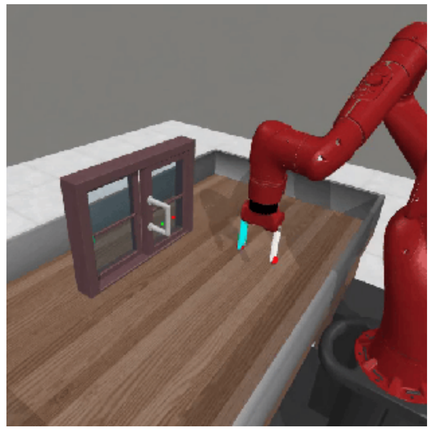
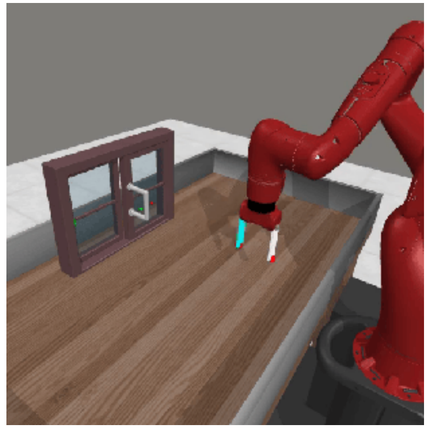
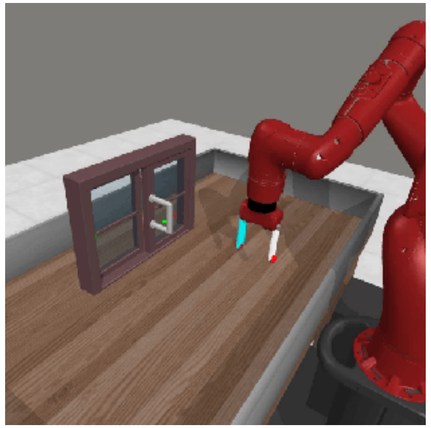
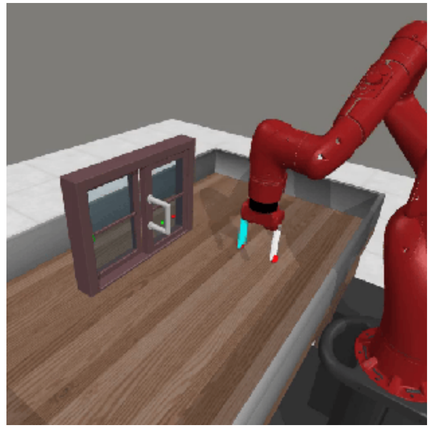
(Efficient) Adaptation

Self-Improving Loops for Visual Robotic Planning (SILVR)



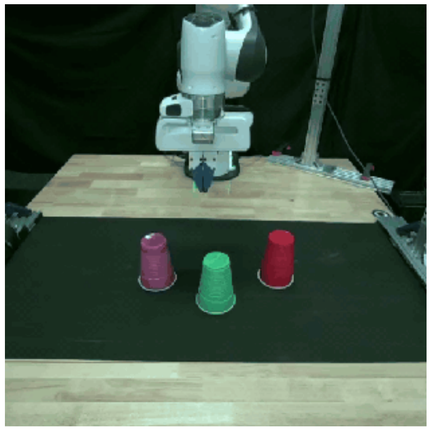
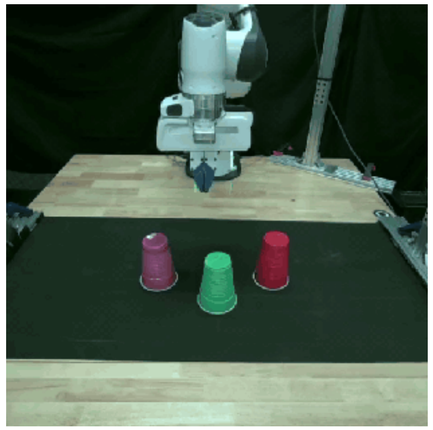
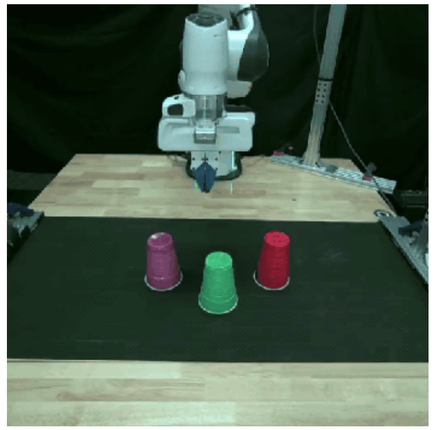
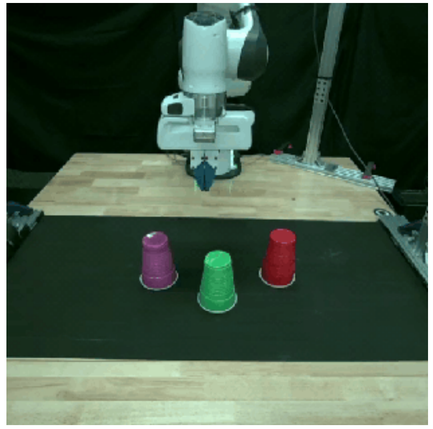
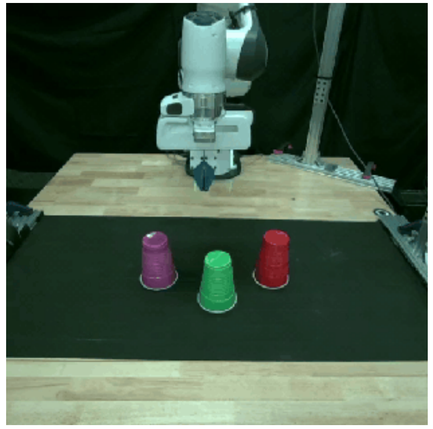
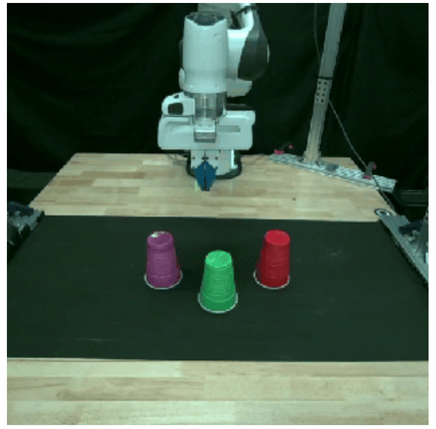
Self-Improving with Online Experience

Novel Task: Window Close

	Iter 0	**Iter 1**	**Iter 2**
Visual Plan			
Env. Execution			

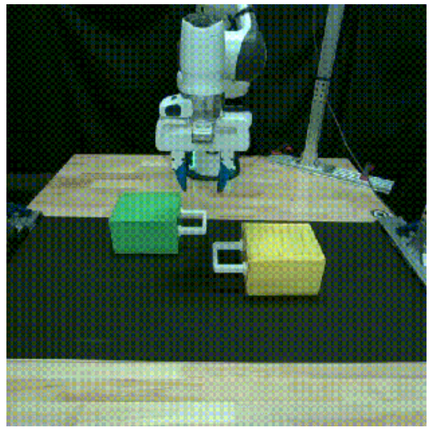
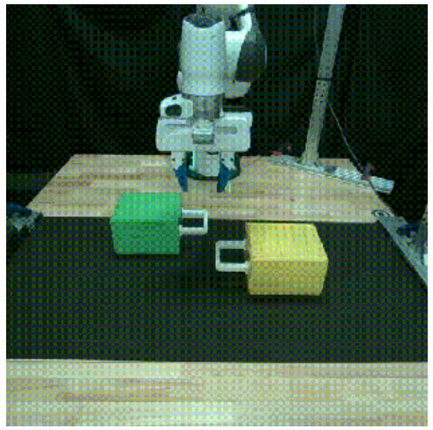
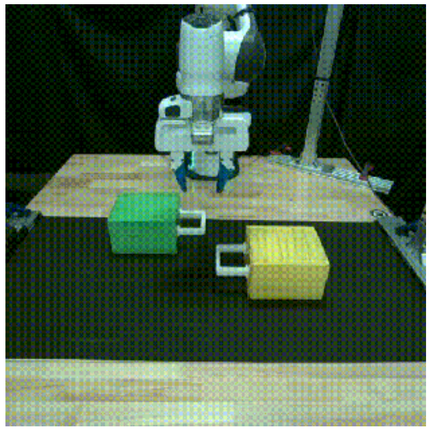
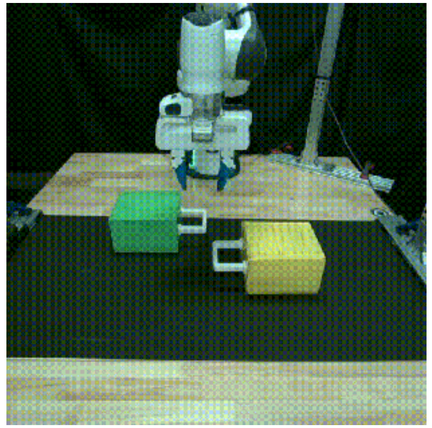
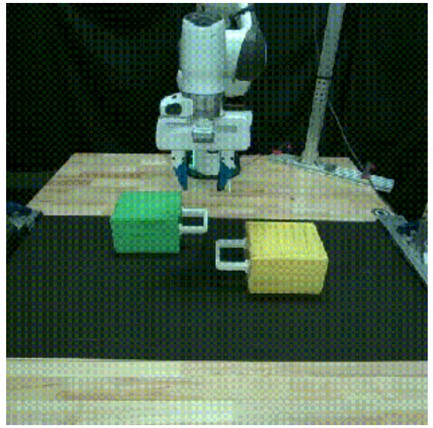
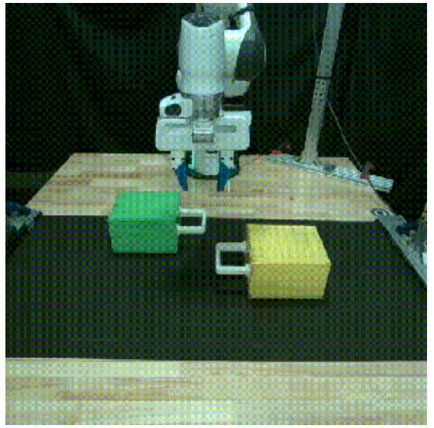
Self-Improving with Online Experience

Novel Task: "Push the Purple Cup"

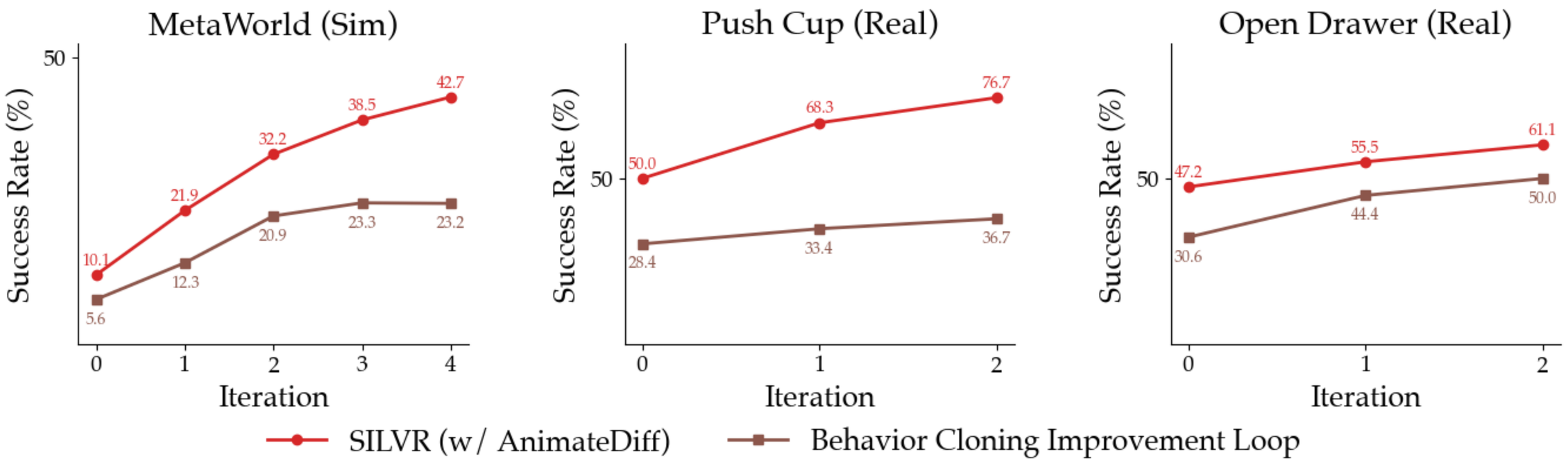
	Iter 0	**Iter 1**	**Iter 2**
Visual Plan			
Env. Execution			
	✗	✗	✓

Self-Improving with Online Experience

Novel Task: "Open the Yellow Drawer"

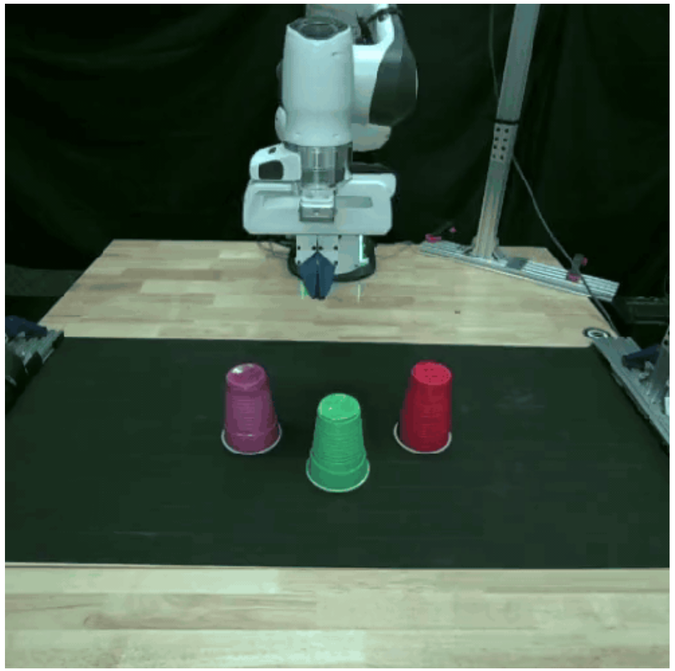
	Iter 0	Iter 1	Iter 2
Visual Plan			
Env. Execution			
	✗	✓	✓

Comparisons to Action-Predictive Policy



Video-based approach achieves better sample efficiency!

Video “World Models” Facilitate Generalizable Decision Making



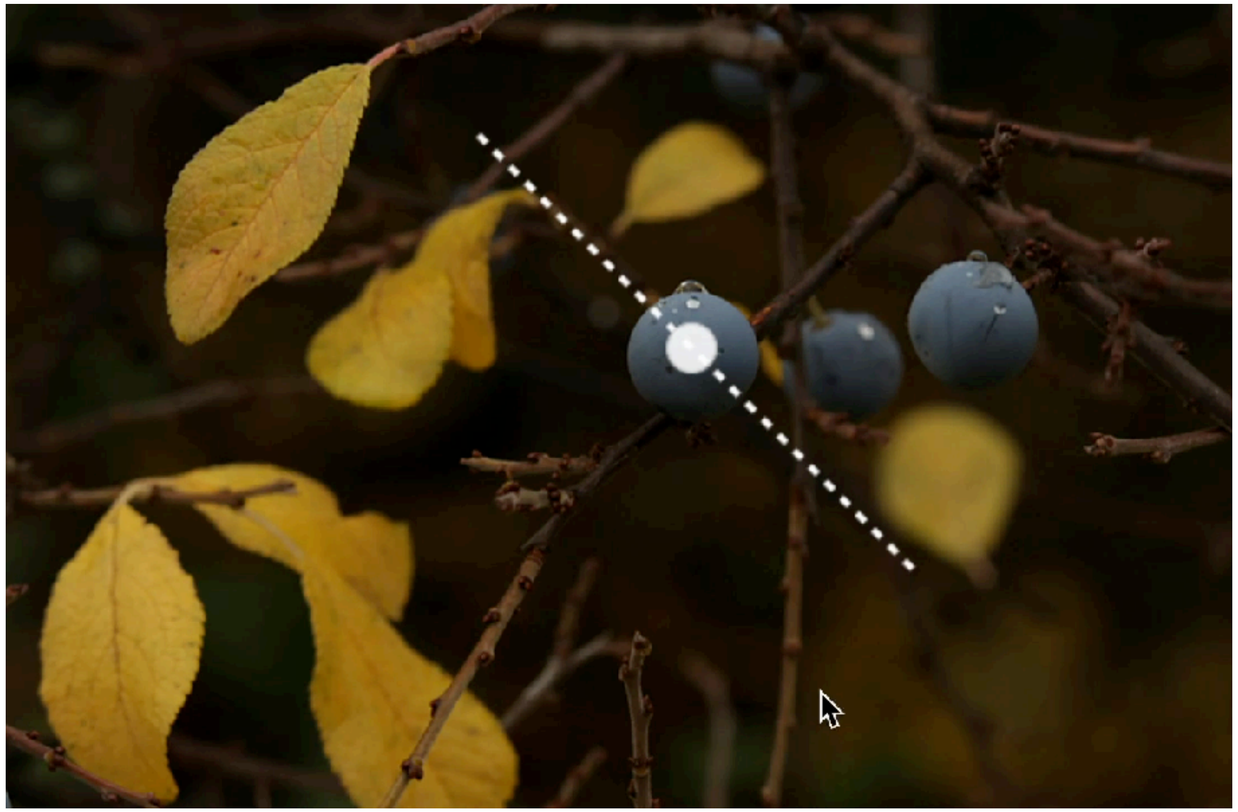
Input:
Current Observation + Goal

Input:
Current Observation + Navigation Direction

Question: Can Users Specify the Physical Attributes, Causes, and Effects?

Missing piece—Control / interact / simulate!

Local point force (“poke”)



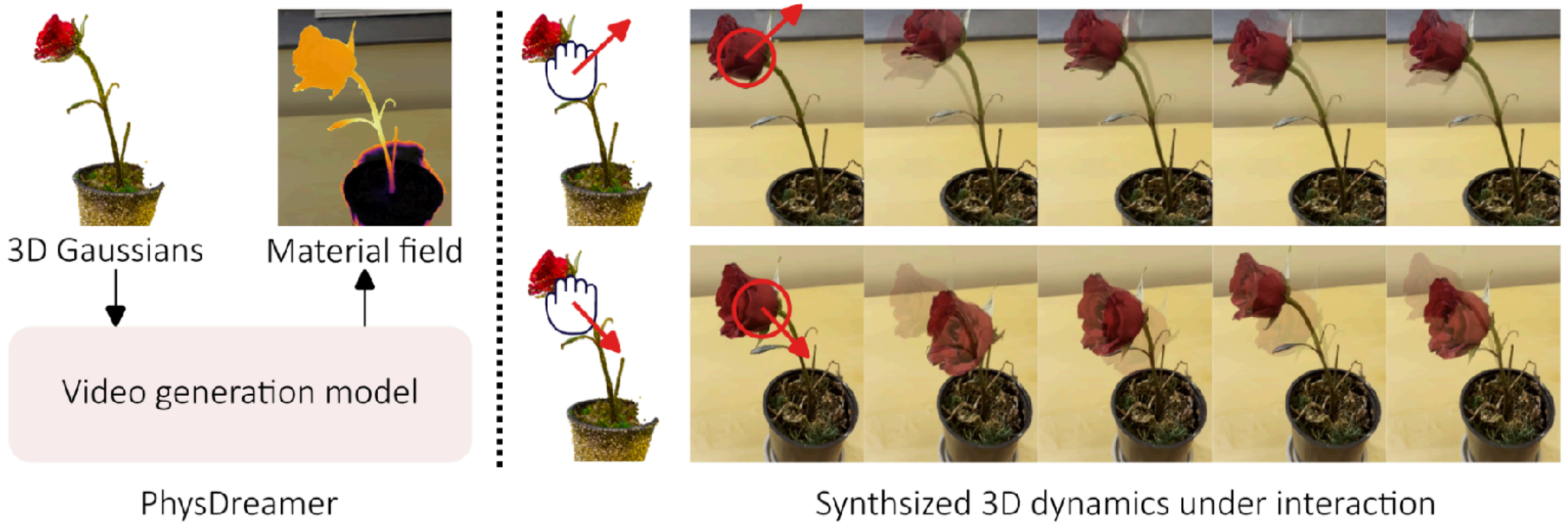
Missing piece—Control / interact / simulate!

Global wind force field

How to build video model as “neural simulators”?

Approach 1:
Rules-based physics simulator + “neural re-rendering”

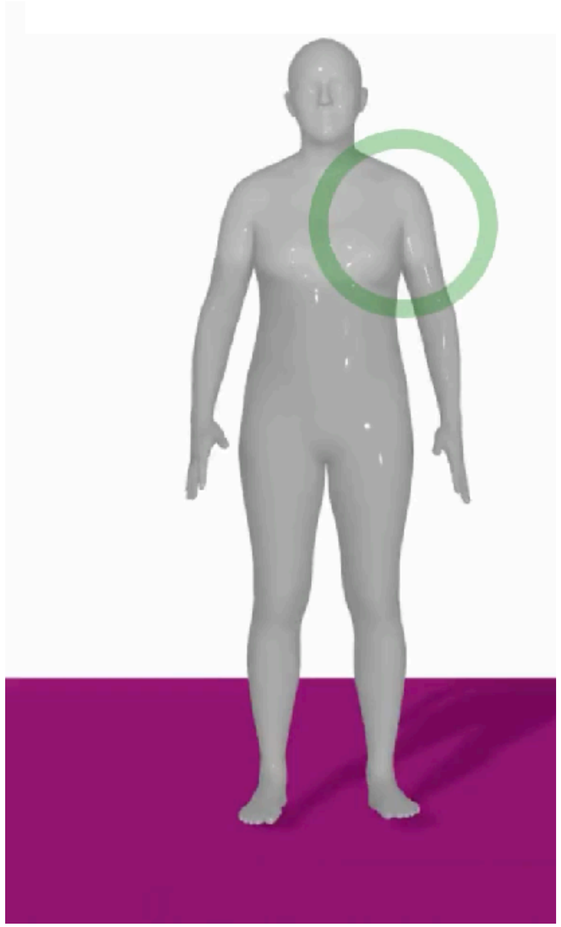
Phys Dreamer = Physics-Based Simulation + Video Diffusion Prior



PhysDreamer (ECCV 2024, Zhang et al.)

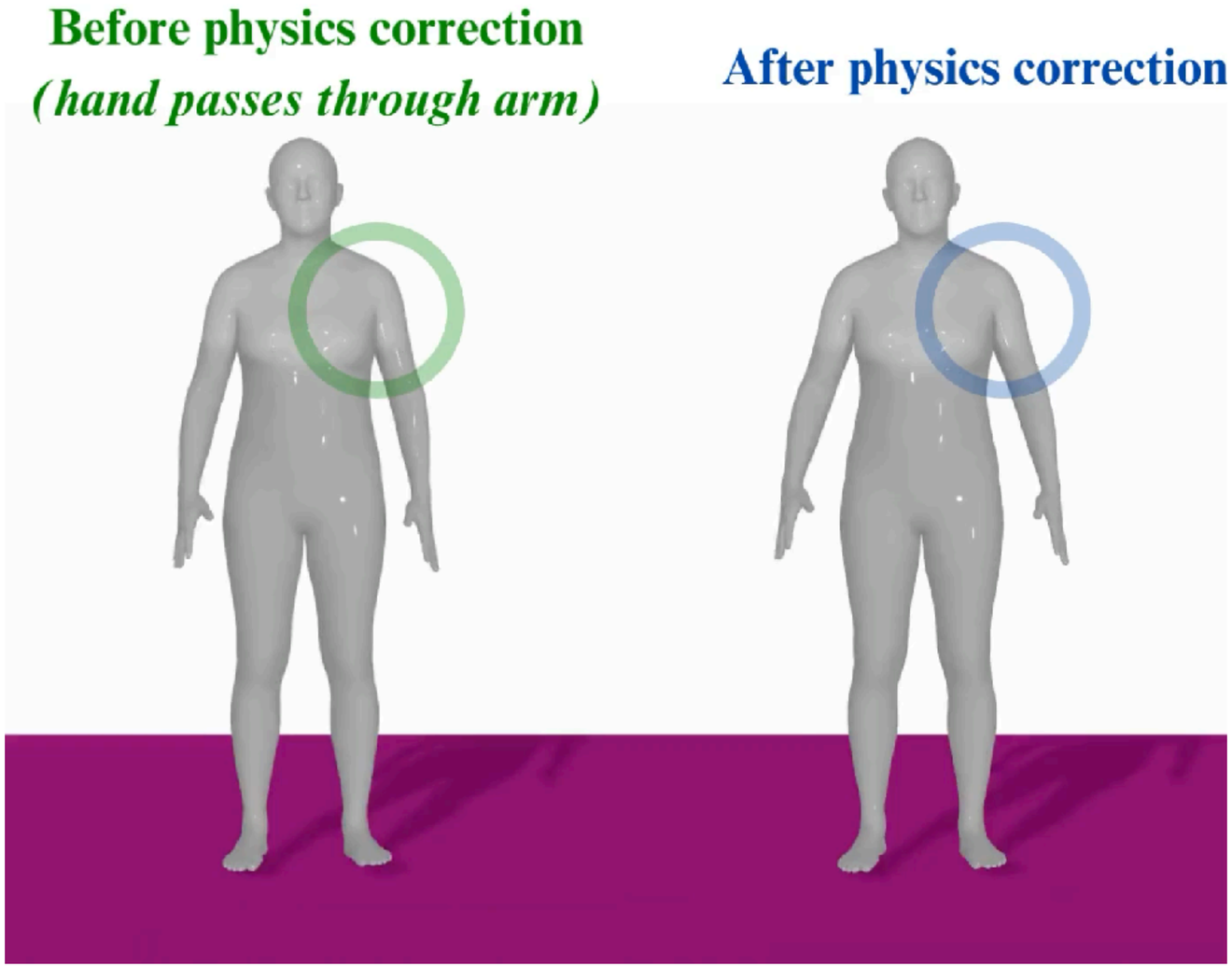
Approach 1: rules-based physics simulator + neural re-rendering

One idea: use physics simulator to “correct” the generative model’s output!



Approach 1: rules-based physics simulator + neural re-rendering

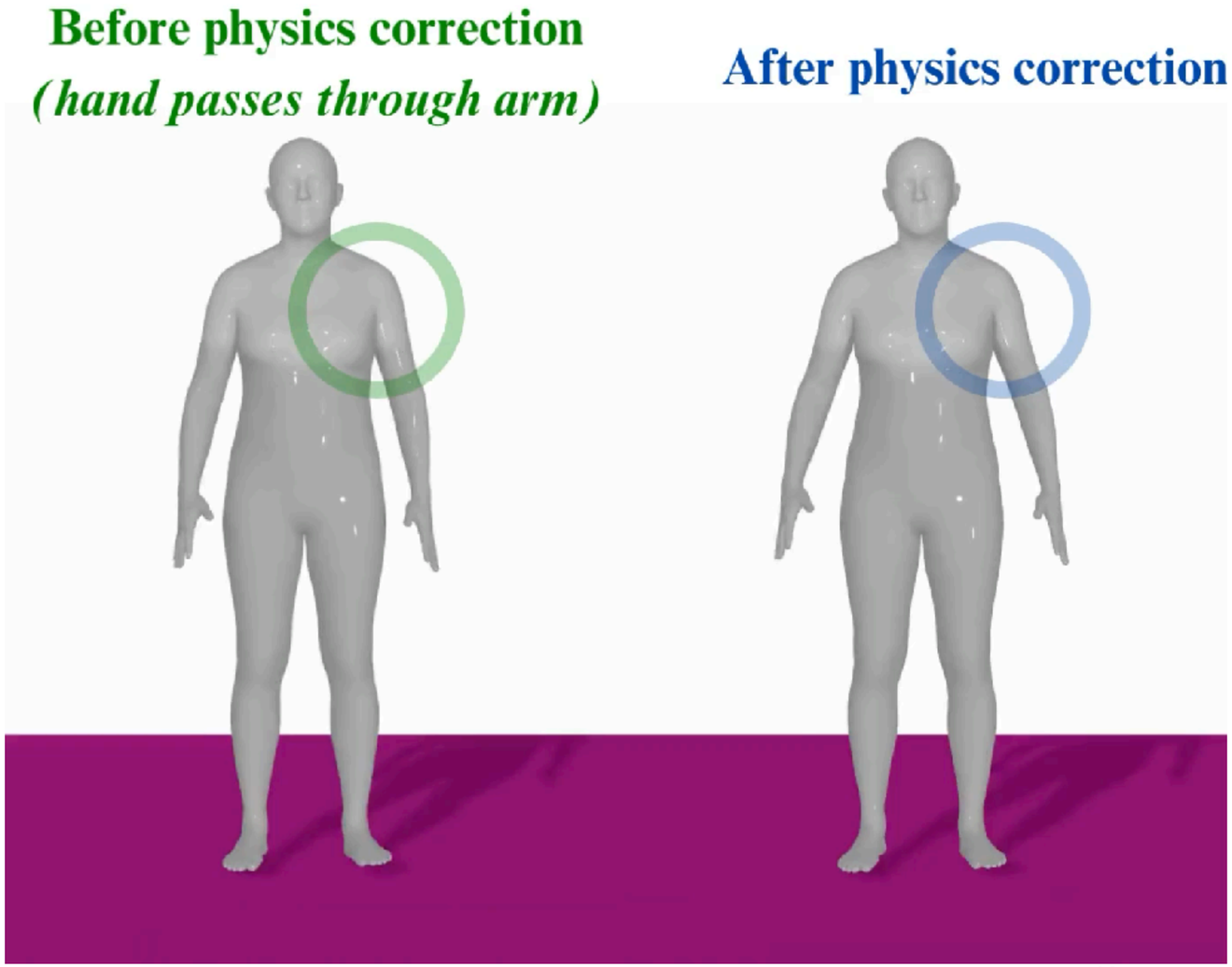
One idea: use physics simulator to “correct” the generative model’s output!



Self-Correcting Self-Consuming Loops for Generative Model Training (ICML 2024, Gillman et al.)

Approach 1: rules-based physics simulator + neural re-rendering

One idea: use physics simulator to “correct” the generative model’s output!



To make this work, need access to:

- Frozen, pretrained policy $\pi(a_t; s_t, \hat{q}_{t+1})$
- Goal: imitate the generated motion sequence $\hat{q}_{1:T}$

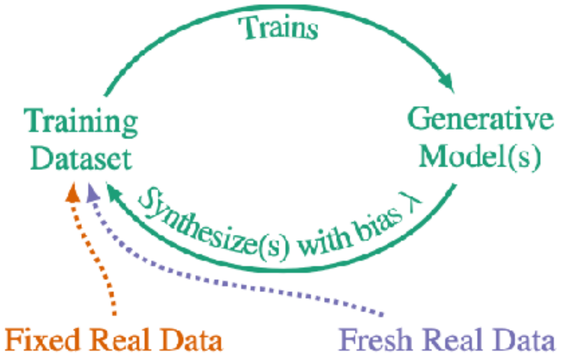
Self-Correcting Self-Consuming Loops for Generative Model Training (ICML 2024, Gillman et al.)

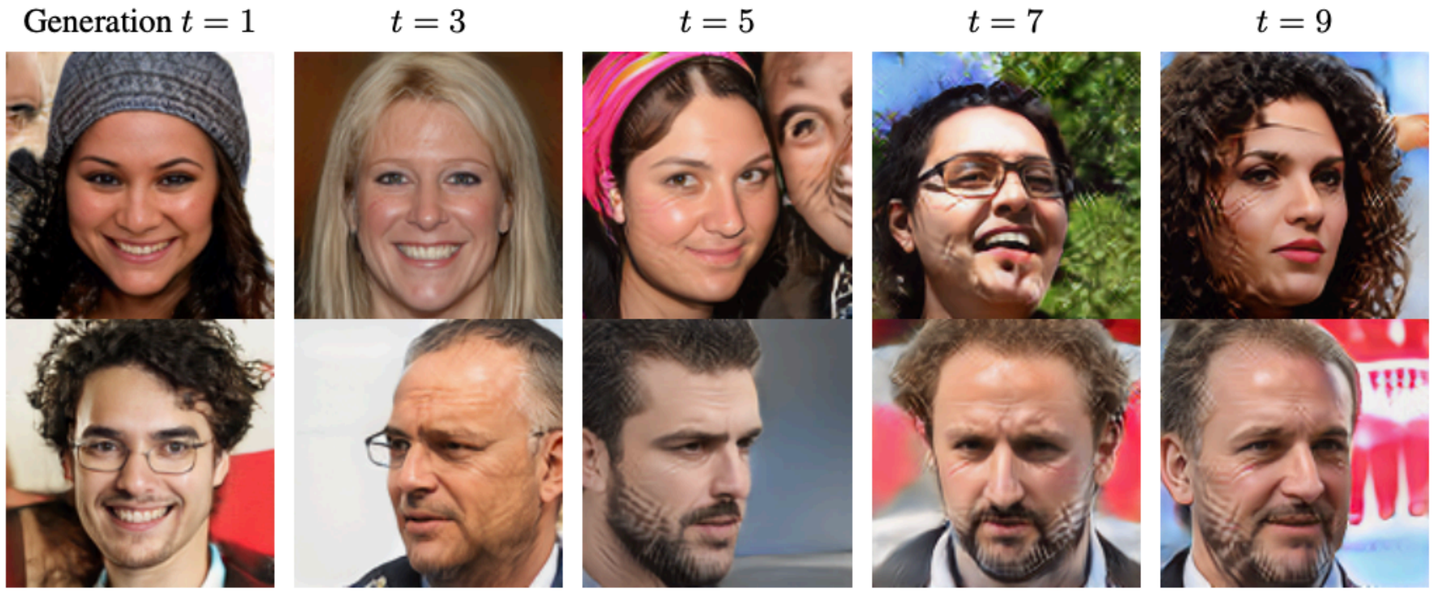
Approach 1: rules-based physics simulator + neural re-rendering

Motivation: "self consuming generative models go MAD"

Self-Consuming Generative Models Go MAD

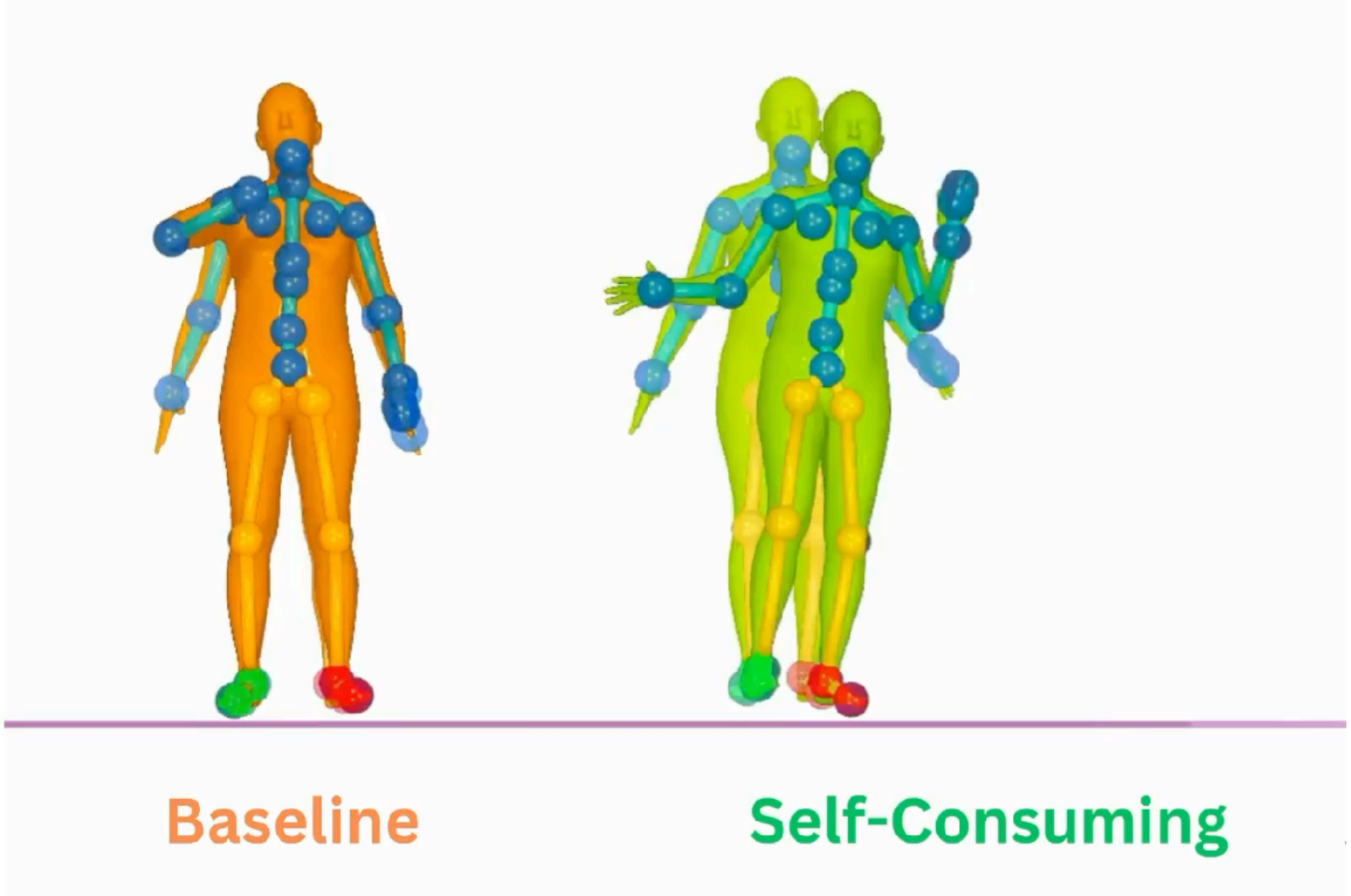
Sina Alemohammad,[†] Josue Casco-Rodriguez,[†] Lorenzo Luzi,[†] Ahmed Imtiaz Humayun,[†]
Hossein Babaei,[†] Daniel LeJeune,[‡] Ali Siahkoohi,[§] Richard G. Baraniuk[†]
[†]Department of Electrical and Computer Engineering, Rice University
[‡]Department of Statistics, Stanford University
[§]Department of Computational Applied Mathematics and Operations Research, Rice University





Approach 1: rules-based physics simulator + neural re-rendering

Motivation: “self consuming generative models go MAD”



Self-Correcting Self-Consuming Loops for Generative Model Training (ICML 2024, Gillman et al.)

Approach 1: rules-based physics simulator + neural re-rendering

Theorem (Stability of Iterative Fine-Tuning with Correction)**.** *Suppose that we have a* sufficiently nice *self-consuming loop weight update procedure* $\theta_0 \to \theta_1 \to \theta_2 \to \dots$ *with:*

1. *Synthetic augmentation percentage* $\lambda \geq 0$*, correction strength* $\gamma \geq 0$*,*
2. λ *and* γ *both satisfying* $\lambda \cdot C < \frac{1+\gamma}{2+\gamma}$*.*

Then with high likelihood*, the self-consuming loop with self-correction strength* γ *satisfies the following stability estimate for all* $t > 0$*:*

$$\|\theta_t - \theta^\star\| \leq c \cdot \sum_{i=0}^{t} \left(\frac{\rho}{1+\gamma}\right)^i + \left(\frac{\rho}{1+\gamma}\right)^t \|\theta_0 - \theta^\star\|.$$

"if correction function is good enough, then the resulting generative model will be excellent"

Self-Correcting Self-Consuming Loops for Generative Model Training (ICML 2024, Gillman et al.)

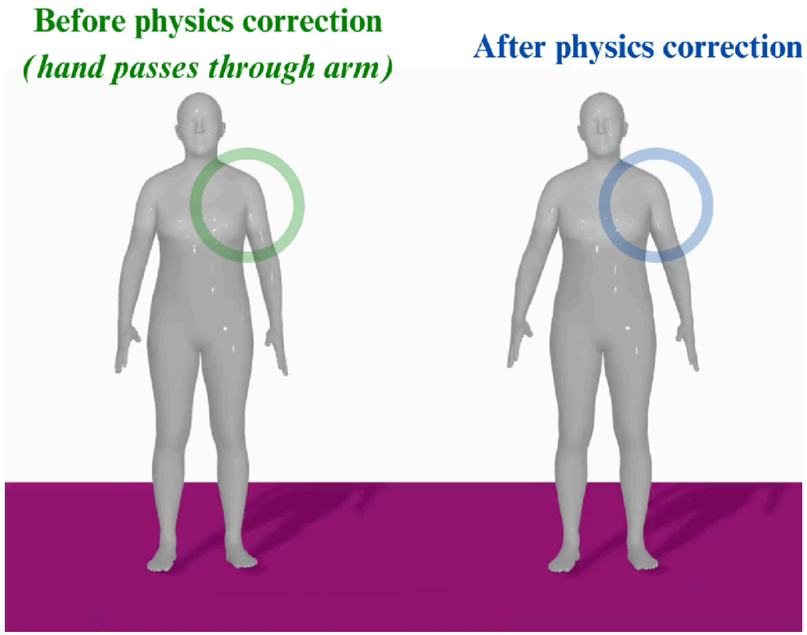
Approach 1: rules-based physics simulator + neural re-rendering

Physics correction successfully avoids MAD-ness

prompt: ***"a person stands with feet wide, stretches both hands up over his head and then swings down by the waist and hangs arms down before standing up"***

Self-Correcting Self-Consuming Loops for Generative Model Training (ICML 2024, Gillman et al.)

Approach 1: rules-based physics simulator + neural re-rendering



Generalization beyond human motion?

To make this work, need access to:

- Frozen, pretrained policy $\pi(a_t; s_t, \hat{q}_{t+1})$
- Goal: imitate the generated motion sequence $\hat{q}_{1:T}$

We **also** need:

- An “inverse renderer”
- A physics simulator

Both of which *only* exist for specific domains

Our research vision

Physics-aware video generative modeling, without 3D assets or physics simulator at inference.

Why? Base video models are improving at astronomical pace; our purely neural approach scales on top of that!

Our research vision

Physics-aware video generative modeling, without 3D assets or physics simulator at inference.

CogVideoX
Sept 2024

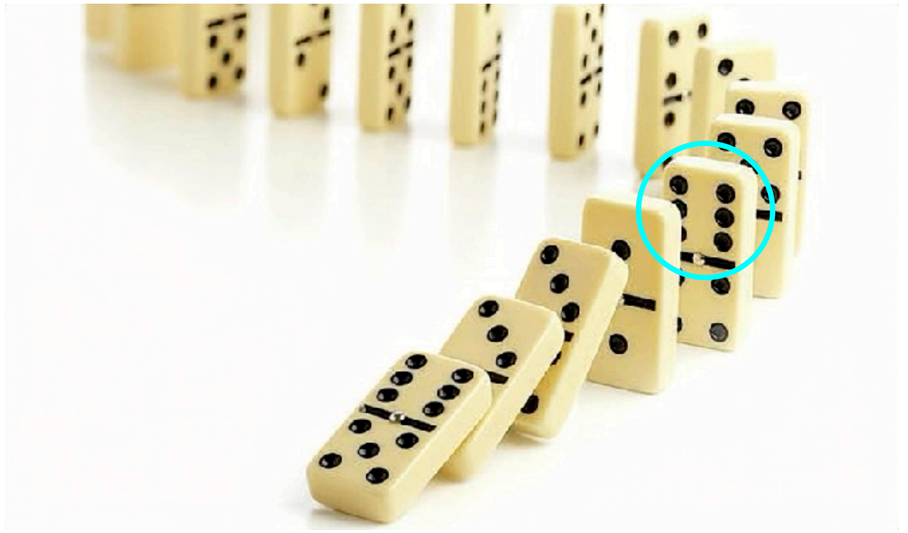
Why? Base video models are improving at astronomical pace; our purely neural approach scales on top of that!

Our research vision

Physics-aware video generative modeling, without 3D assets or physics simulator at inference.

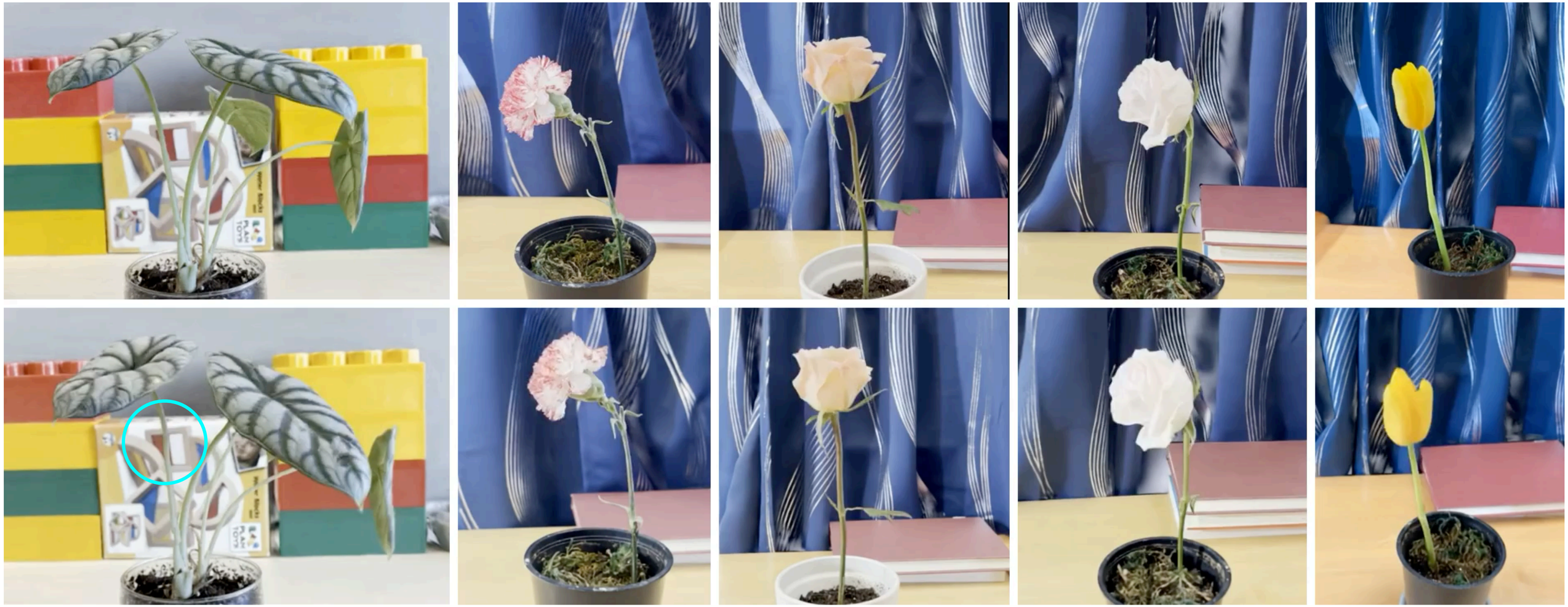
CogVideoX
Sept 2024

Why? Base video models are improving at astronomical pace; our purely neural approach scales on top of that!

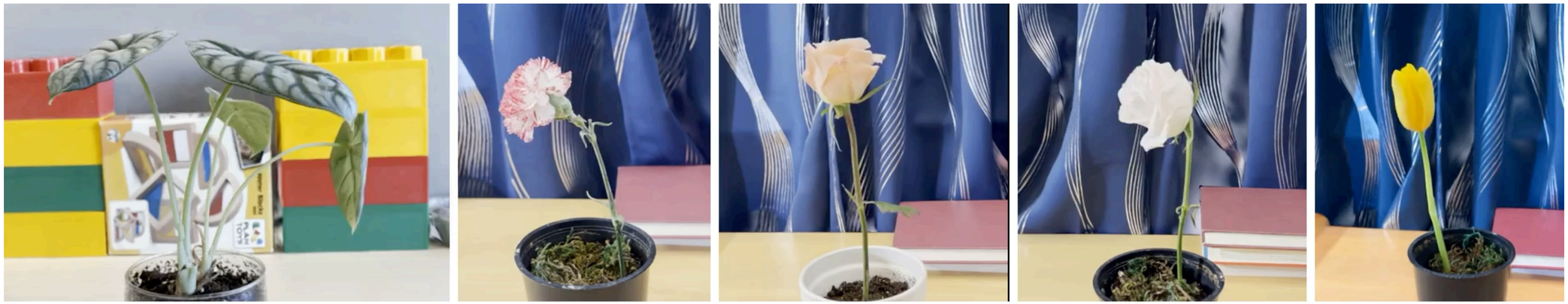


Wan2.2
July 2025

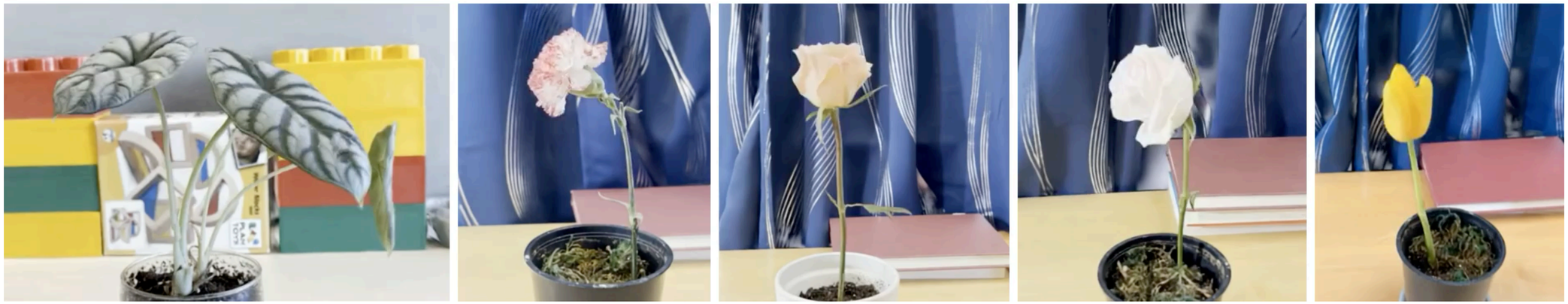
Which row is simulated, and which is generated purely neurally?



Which row is simulated, and which is generated purely neurally?



PhysDreamer (ECCV 2024, Zhang et al.)



Force Prompting

Our fundamental assumption:

Strong
motion prior

Prompt: “the windmill turns”

Our fundamental assumption:

Strong motion prior

+

Synthesized force control

Prompt: “the windmill turns”

Our fundamental assumption:

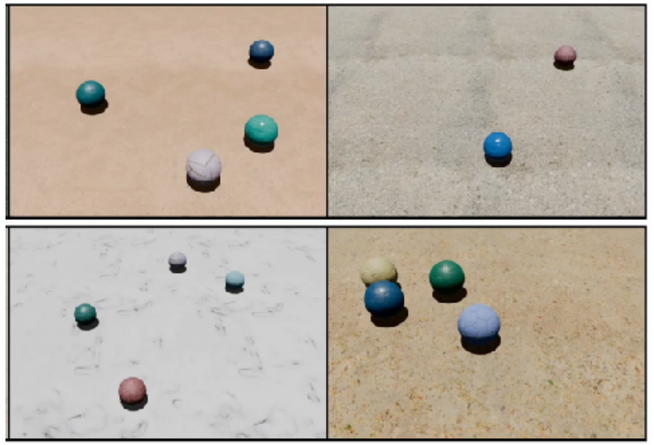
Strong motion prior

+

Synthesized force control

=

Generalized force control



Prompt: “the windmill turns”

Our fundamental assumption:

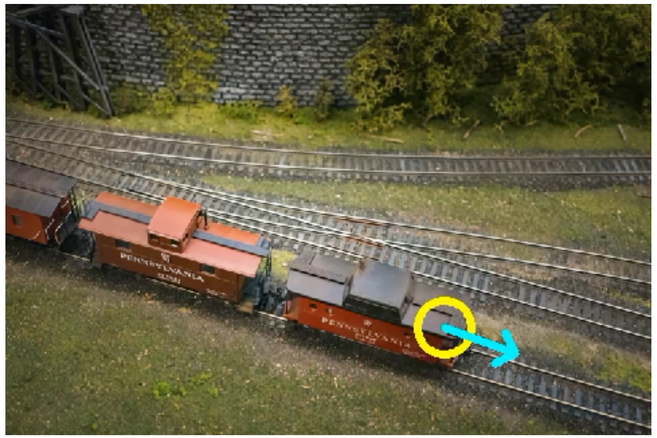
Strong motion prior

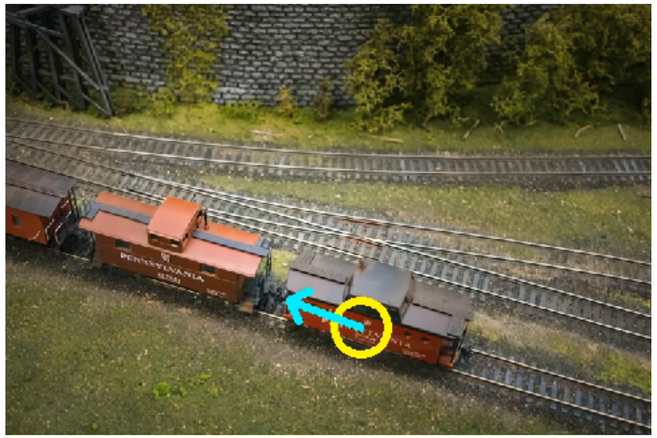
+

Synthesized force control

=

Generalized force control



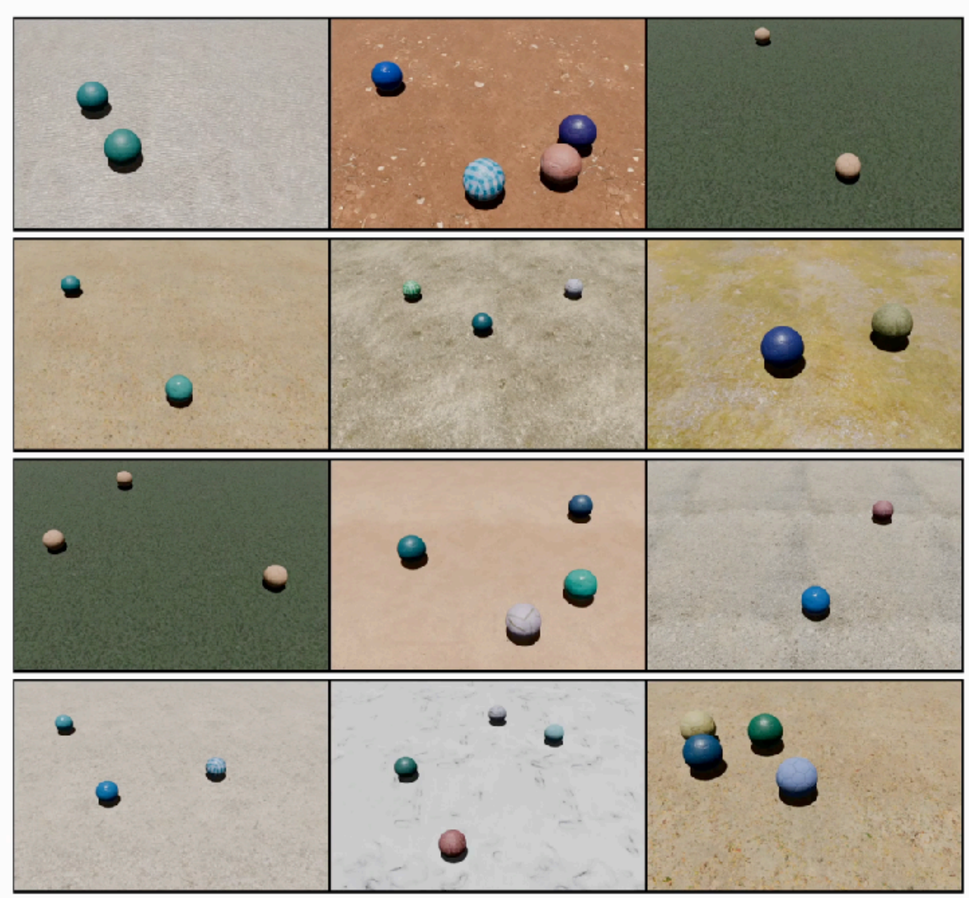


Prompt: “the train moves”

Training data: not diverse and not scaled

Global force (wind)

Point force (poke)



+ a single PhysDreamer flower

The biggest question:
How well does this generalize?

The answer:
Pretty well!

Control signal generalizes: diverse geometries / materials

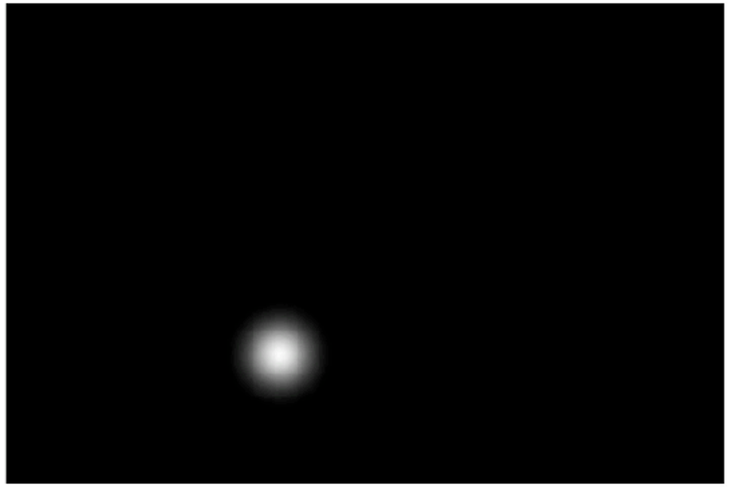
Control signal generalizes: diverse geometries / materials

Control over force ***magnitude***

Control signal generalizes: diverse geometries / materials

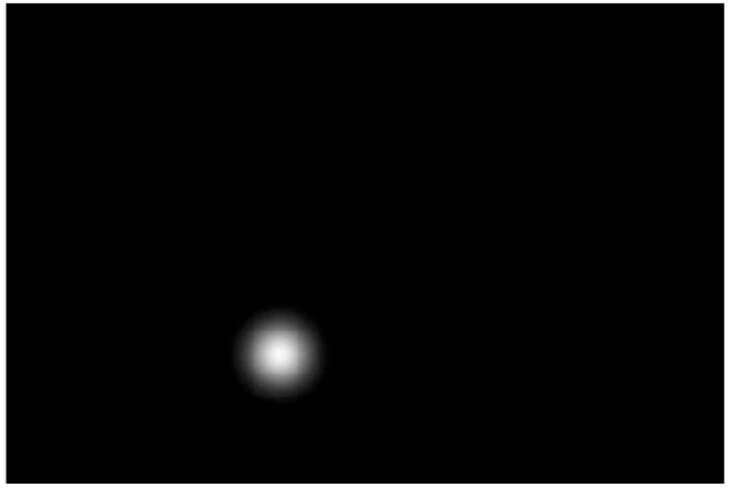
Control over force ***direction***

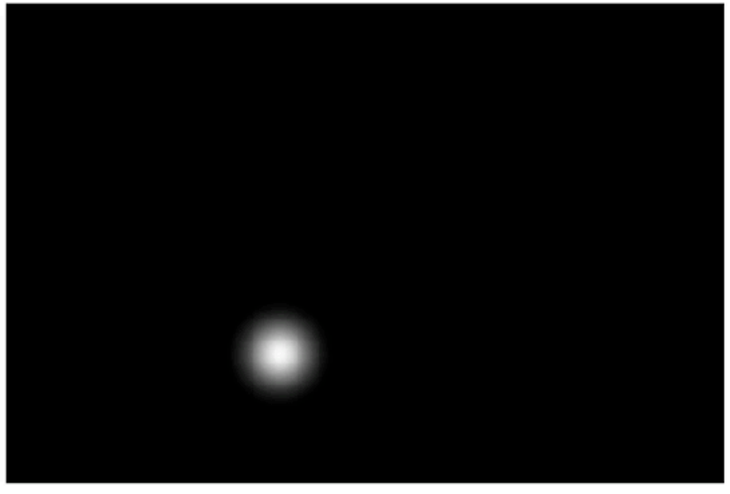
Control signal is Gaussian blob in ControlNet



Prompt: *"A cozy, woven swing gently sways back and forth under the shade of a palm tree"*

Control signal is Gaussian blob in ControlNet


Prompt: *"A cozy, woven swing gently sways back and forth under the shade of a palm tree"*

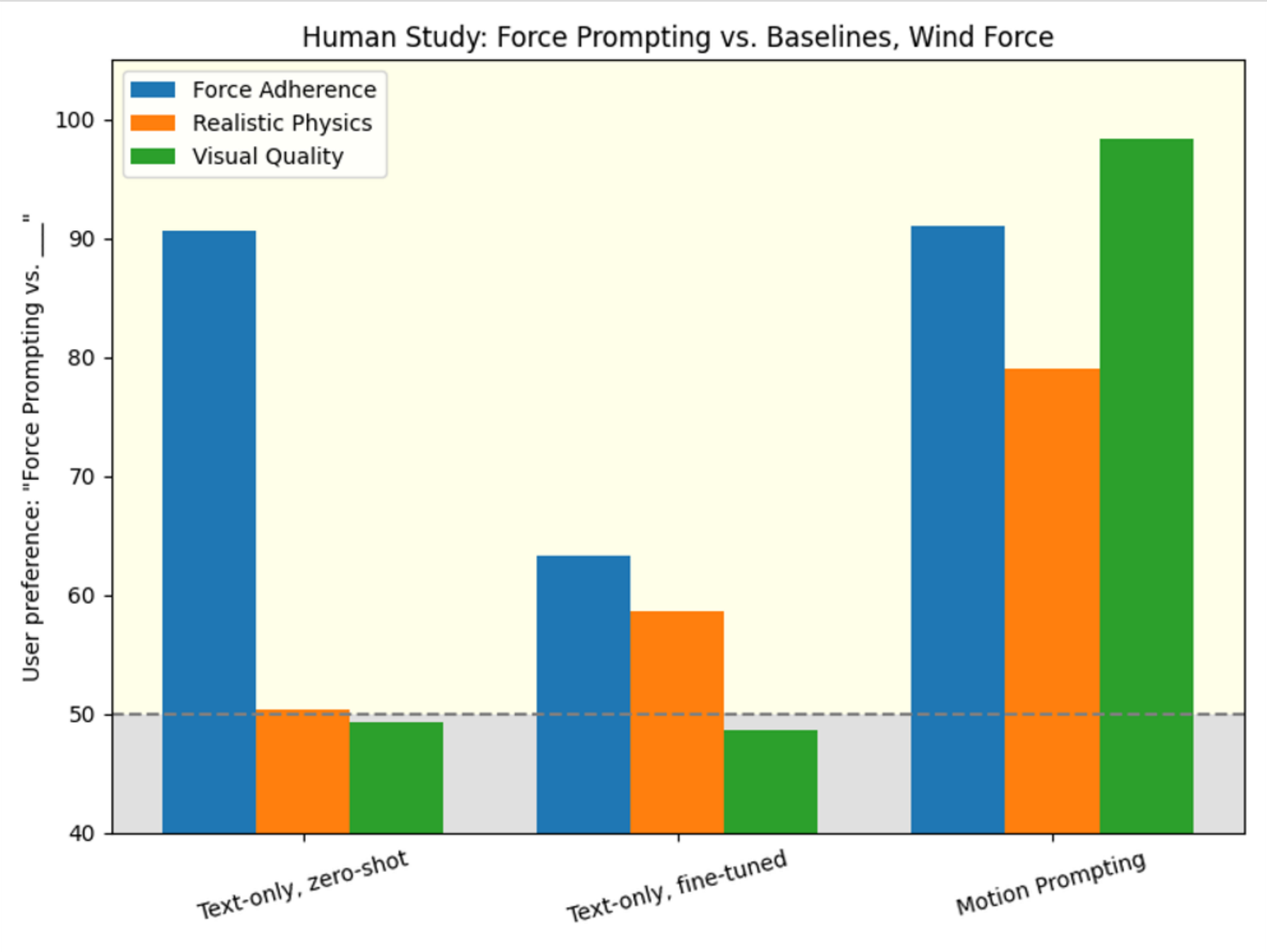
Force Prompting gives better physics control than baselines

Our metrics:

1. Force adherence
2. Realistic physics
3. Visual quality

Force Prompting gives better physics control than baselines

Force Prompting gives better physics control than baselines



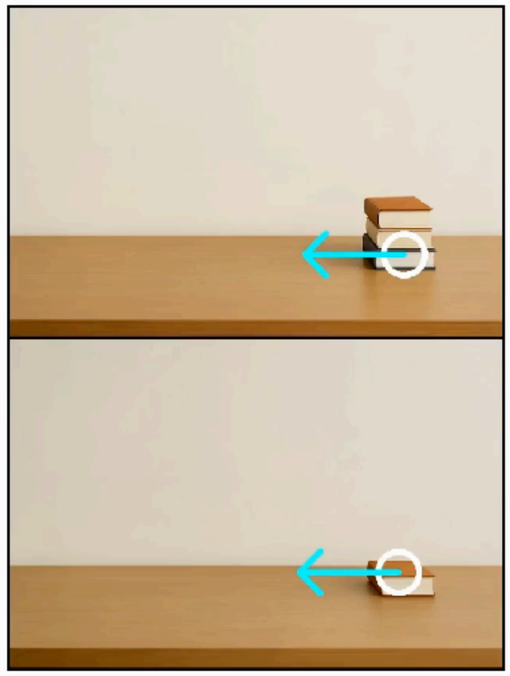
Emergent behaviors from Force Prompting

Emergent property 1: understanding force / mass relationship

Hints at Mass Understanding

The *same force* results in *different motion* depending on the object's *inferred mass*

Single book vs. stack of books

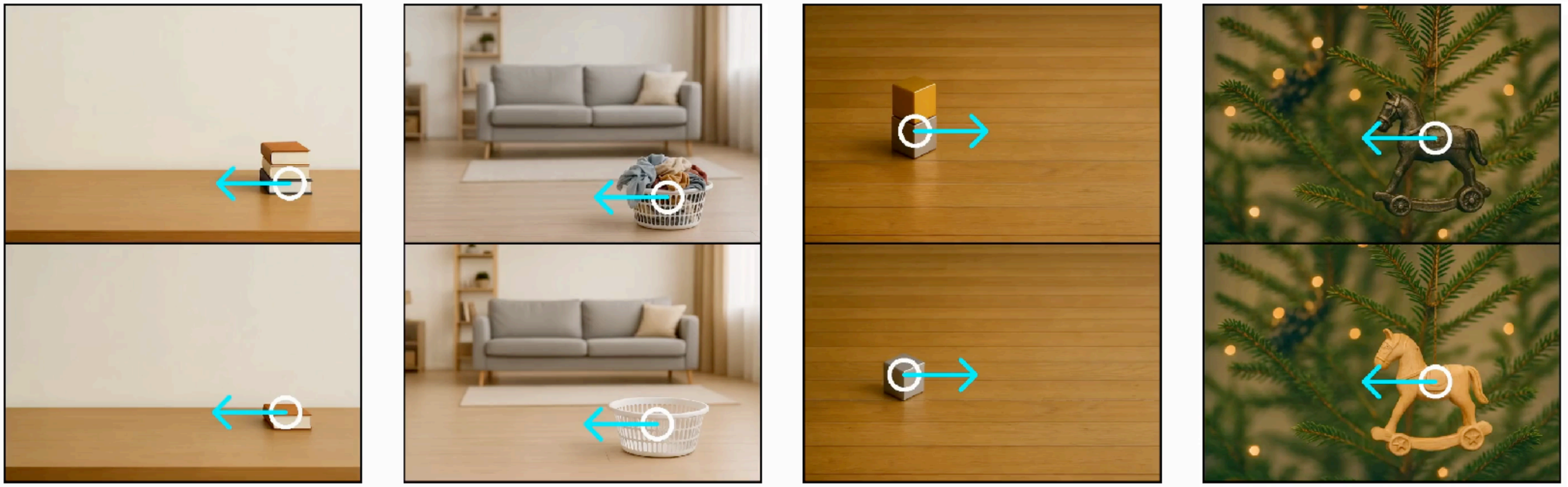


Emergent property 1: understanding force / mass relationship

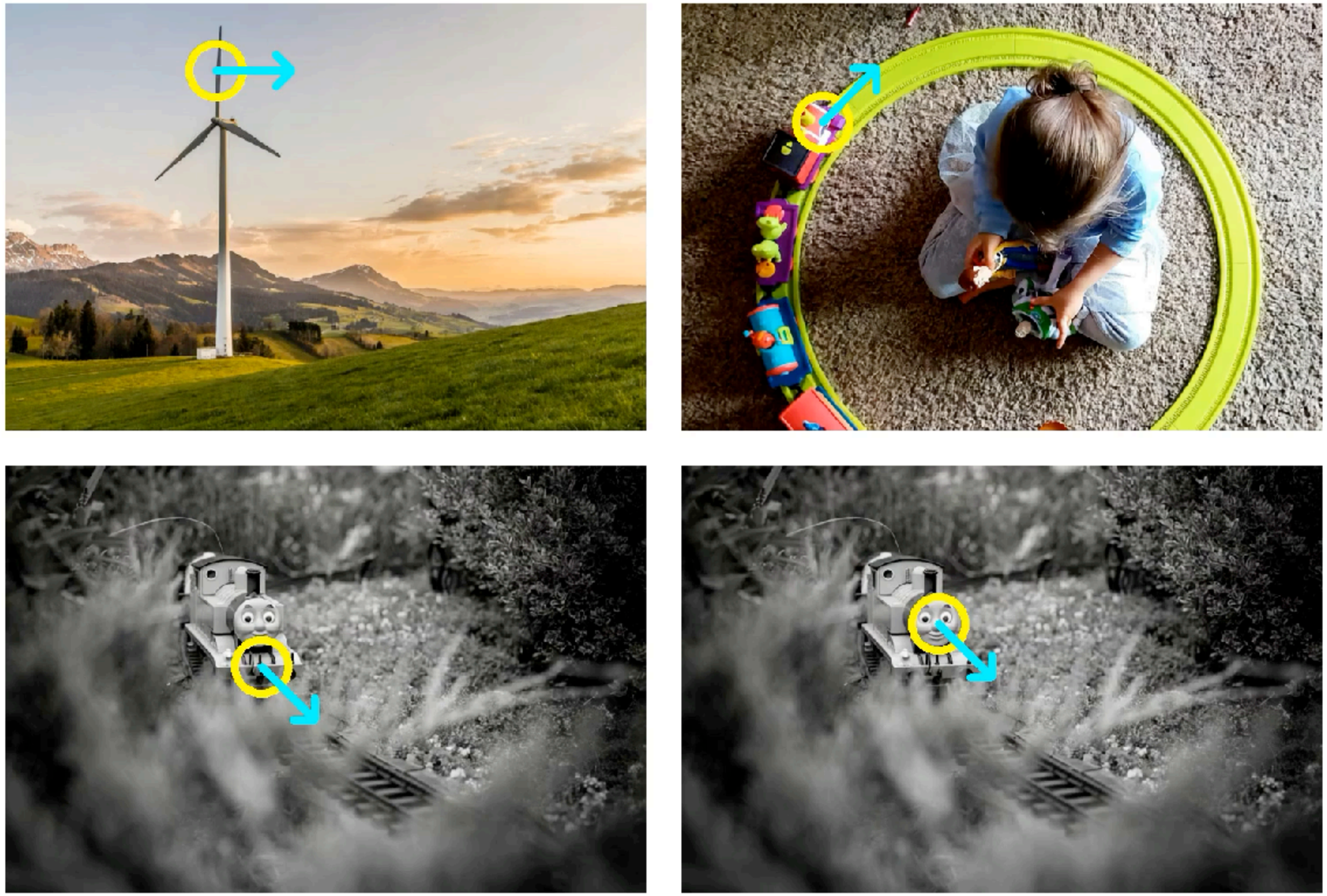
Hints at Mass Understanding

The *same force* results in *different motion* depending on the object's *inferred mass*

Single book vs. stack of books	*Empty laundry basket vs. full laundry basket*	*Single cube vs. stack of cubes*	*Wooden ornament vs. metal ornament*



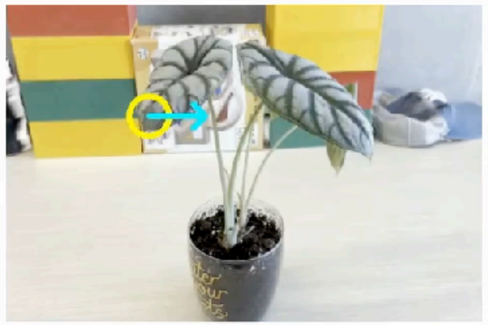
Emergent property 2: understanding affordances / atomicity



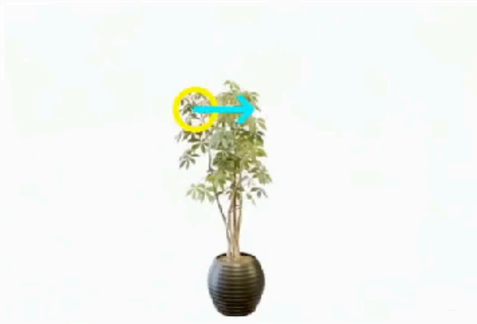
Emergent property 3: “multi-poke”

Single generative model can recreate demos that use simulators or 3D assets

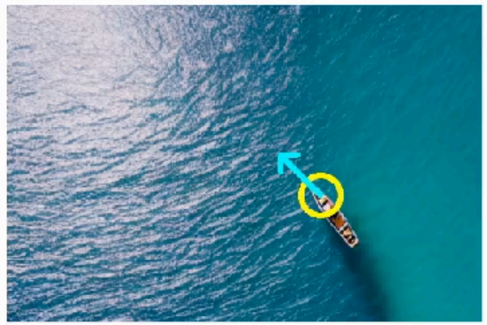
Recreating a PhysDreamer (ECCV 2024) demo



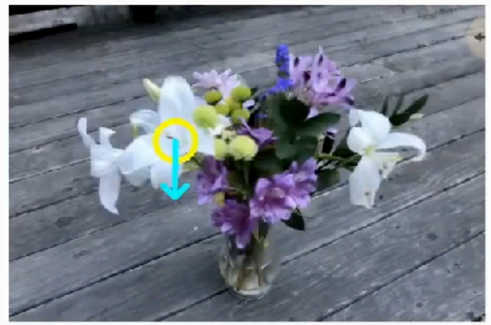
Recreating a DreamPhysics (AAAI 2025) demo



Recreating a MotionCraft (NeurIPS 2024) demo



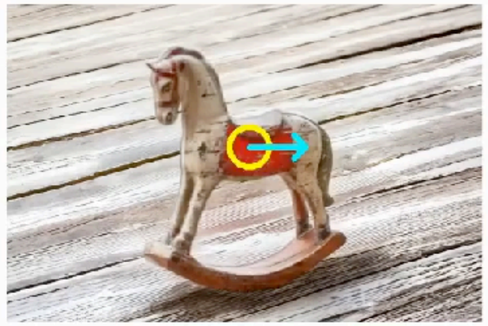
Recreating a PhysGaussian (CVPR 2024) demo



Recreating a PhysGen (ECCV 2024) demo

Recreating a Physics3D demo

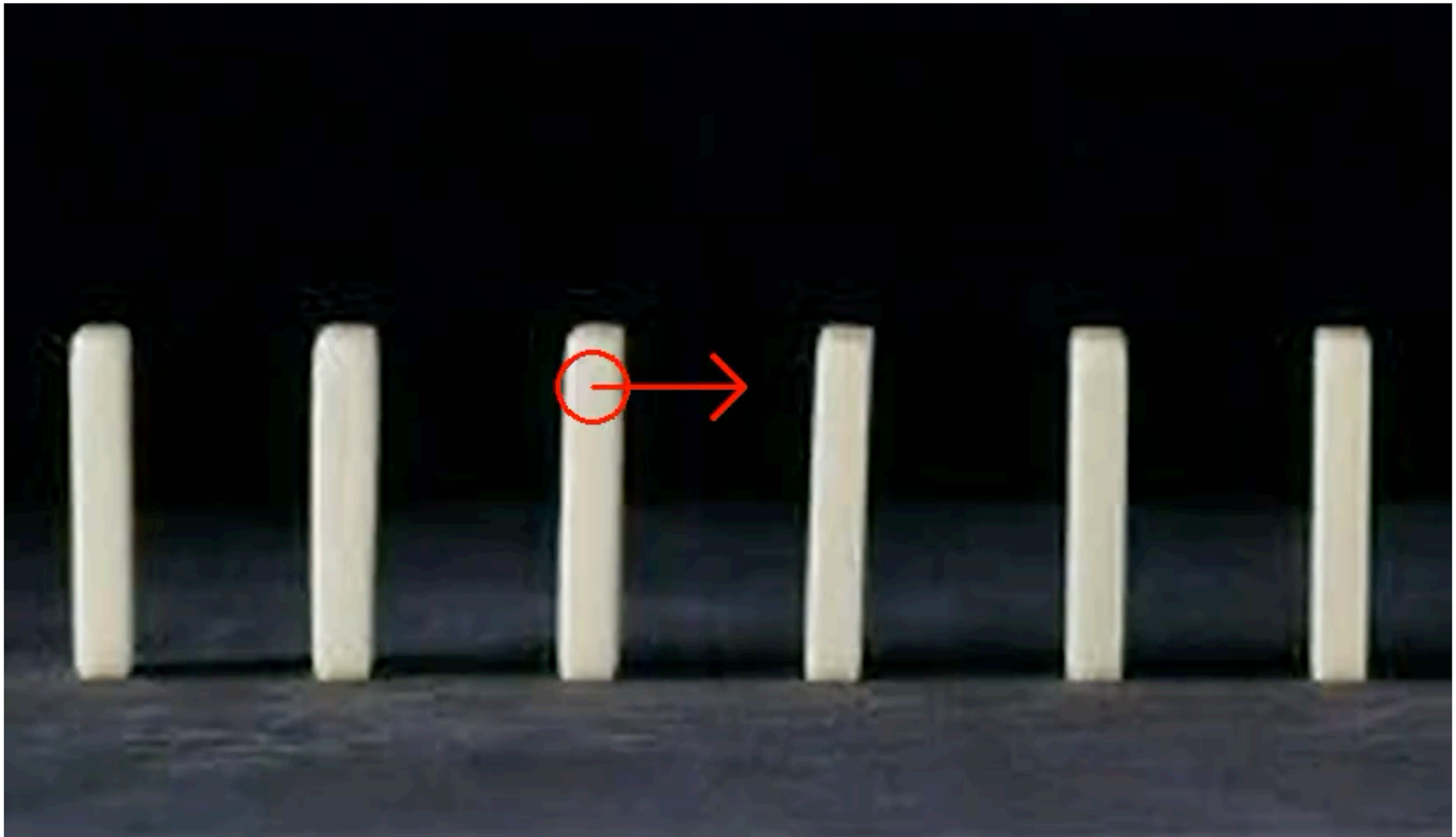
Recreating a PhysMotion demo



Recreating a PhysGen3D demo

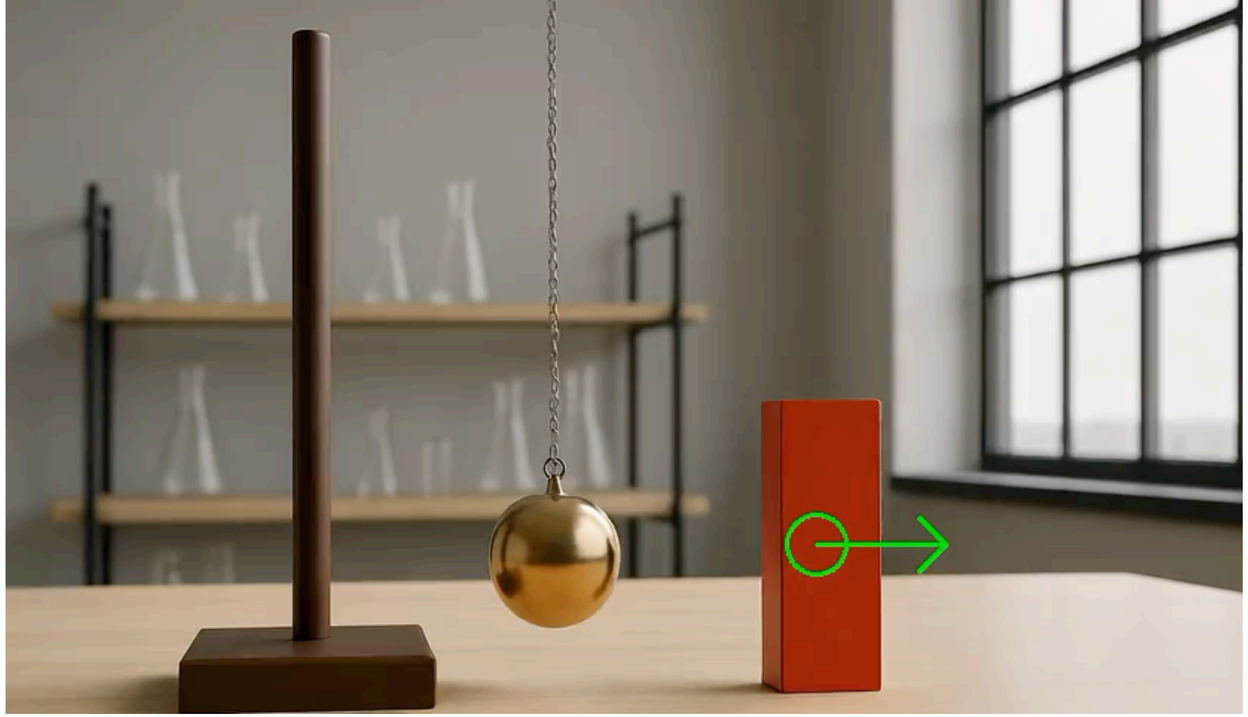
What if we do Force Prompting with a stronger base video model?

We can model more physical phenomena!



Applications?

How can we benefit from the huge progress of video generation models?

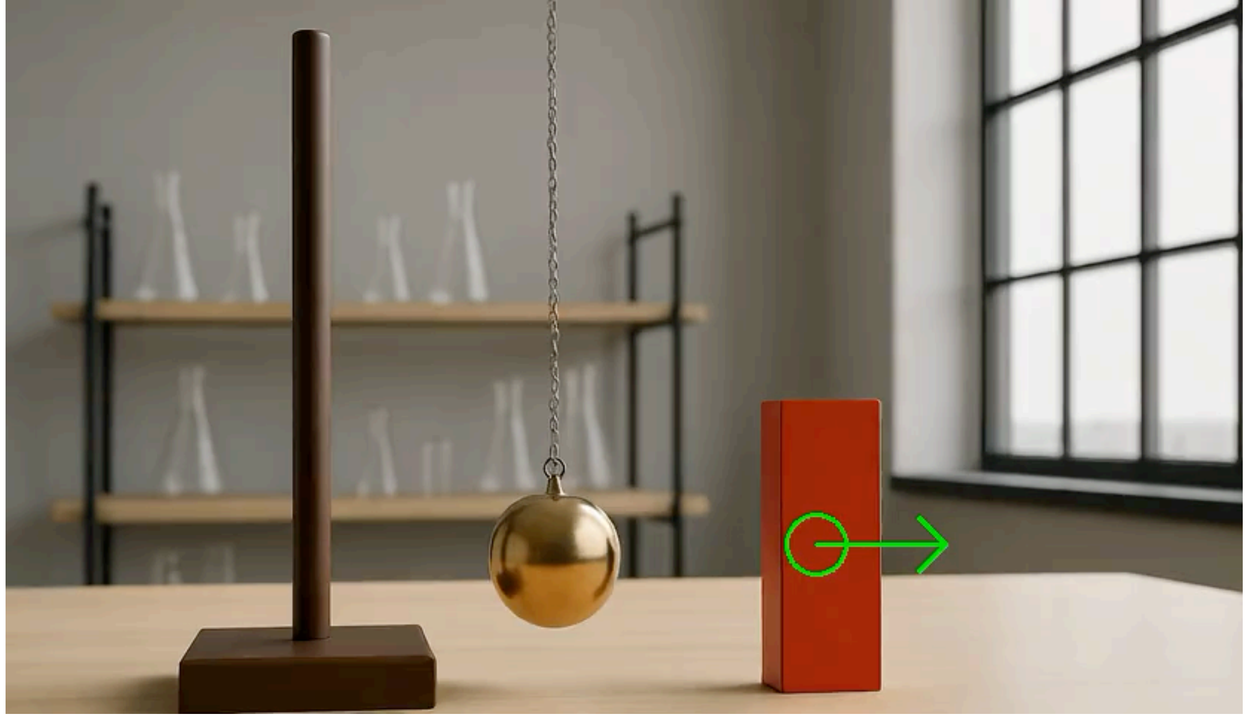


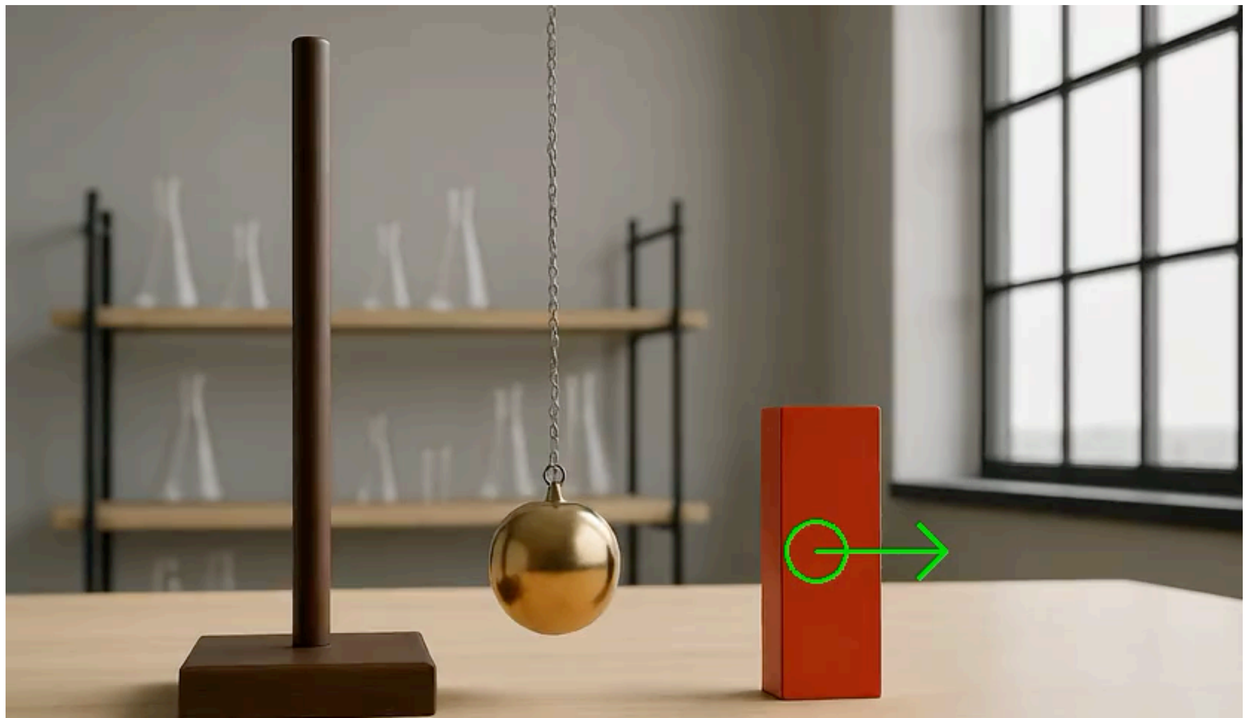
An idea: “chain of actions” towards a goal!

Applications?

How can we benefit from the huge progress of video generation models?

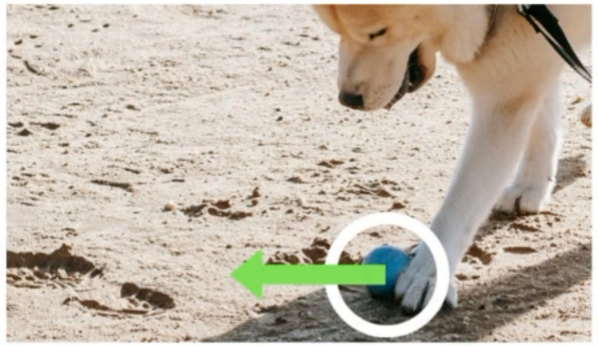
An idea: “chain of actions” towards a goal!





Teaching Video Models To Accomplish Physics-Conditioned Goals

Situation:
[initial frame]



Goal: *hit this specific ball in this specific direction*

Generate video using
Goal Force

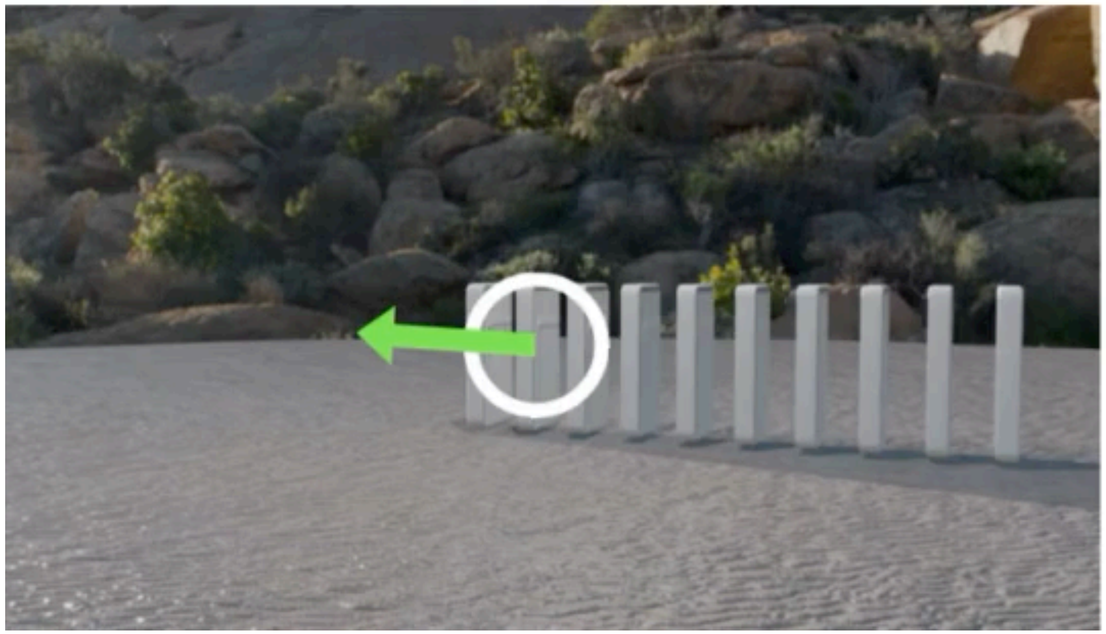
Output: *video with antecedent action (dog paw pushes ball) that ensures goal force happens (ball forced to the left)*

Goal Force (CVPR 2026, Gillman et al.)

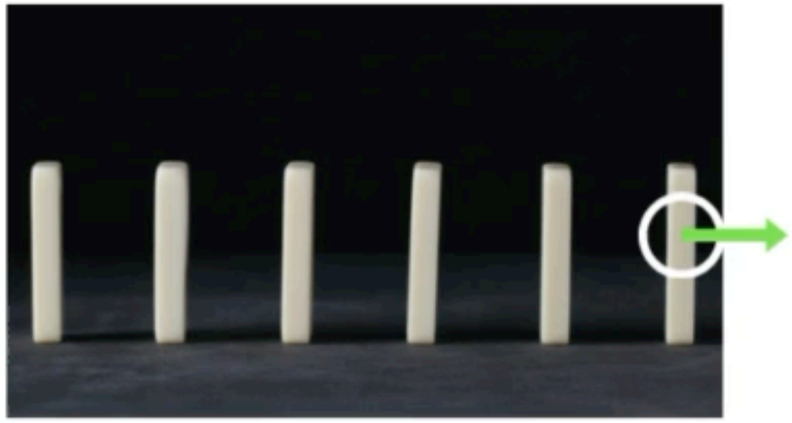
One scene, multiple outcomes via varying Goal Force vectors

Goal Force (CVPR 2026, Gillman et al.)

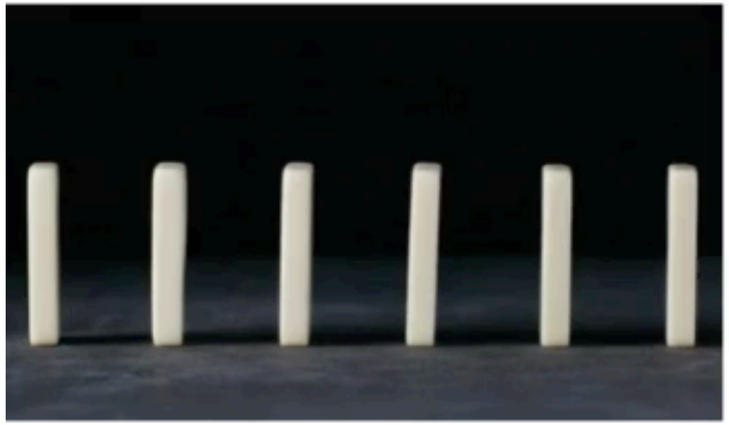
Synthetic training data recipe

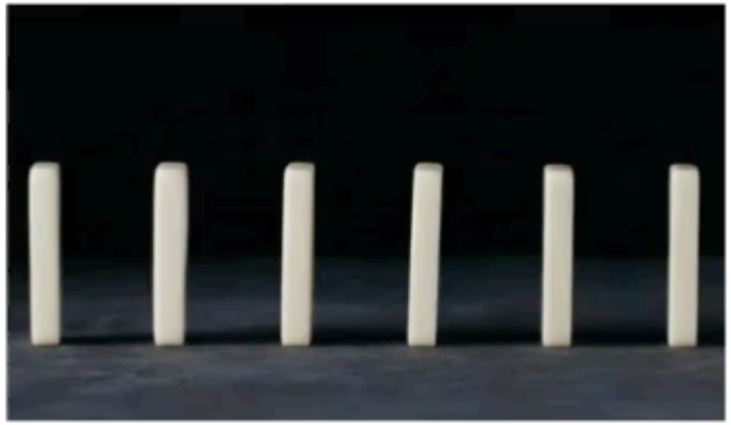


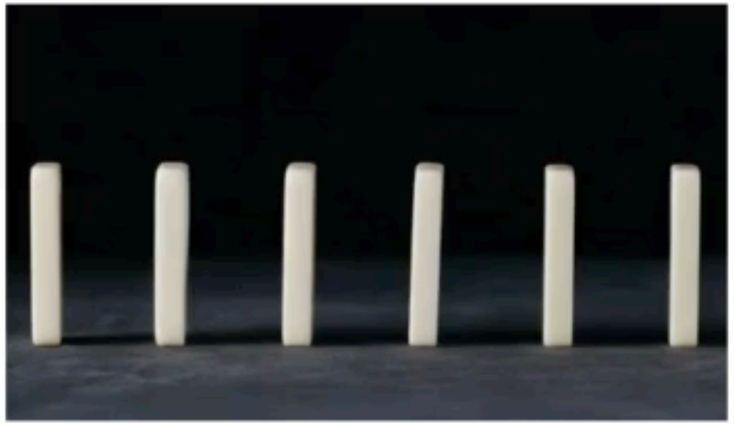
Goal Force prompting generates diverse probabilistic plans!



$x_1 \sim p(x)$ $x_2 \sim p(x)$ $x_3 \sim p(x)$



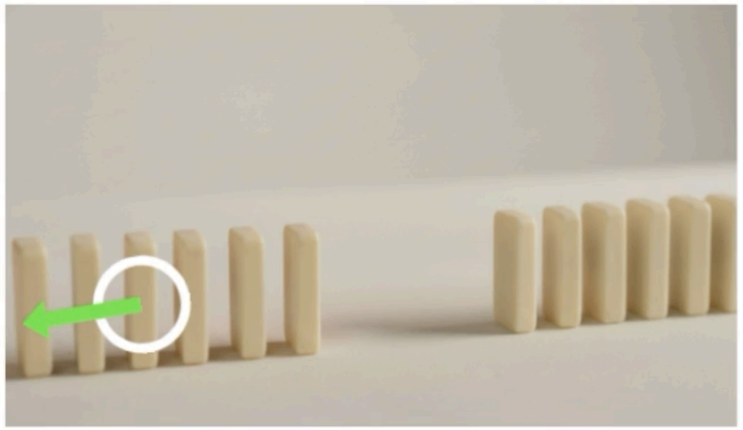


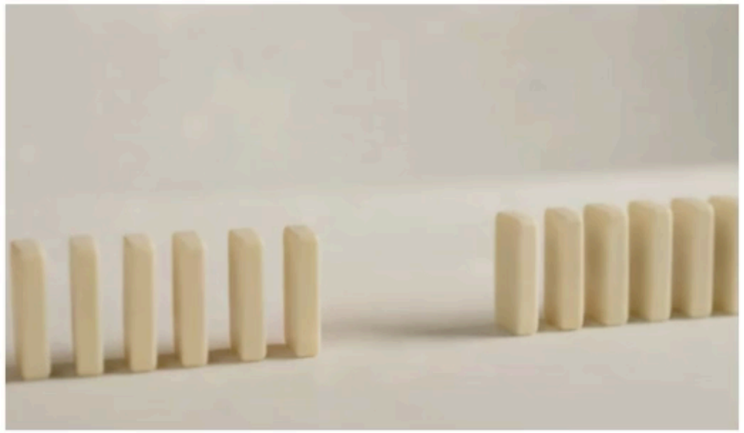


Goal Force prompting generates diverse probabilistic plans!

$$x_1 \sim p(x) \qquad x_2 \sim p(x)$$

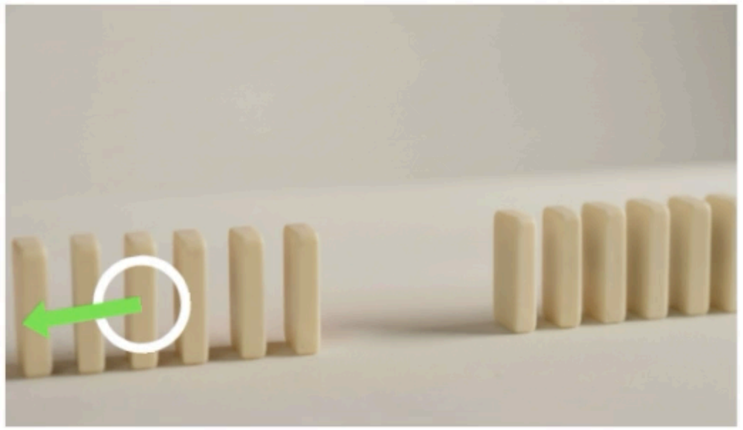
Goal Force prompting generates physically plausible plans

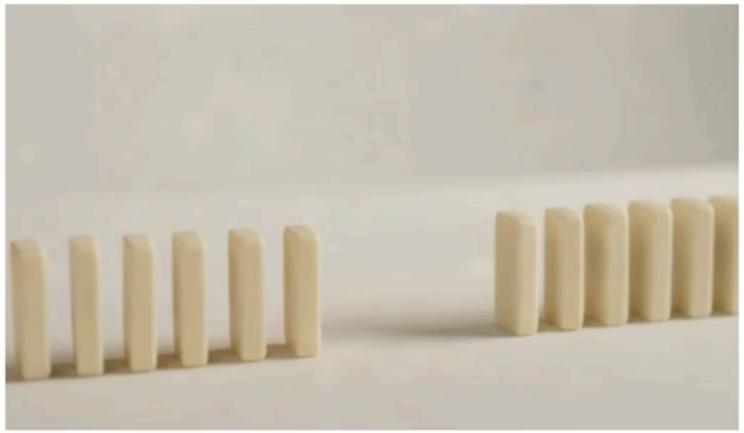




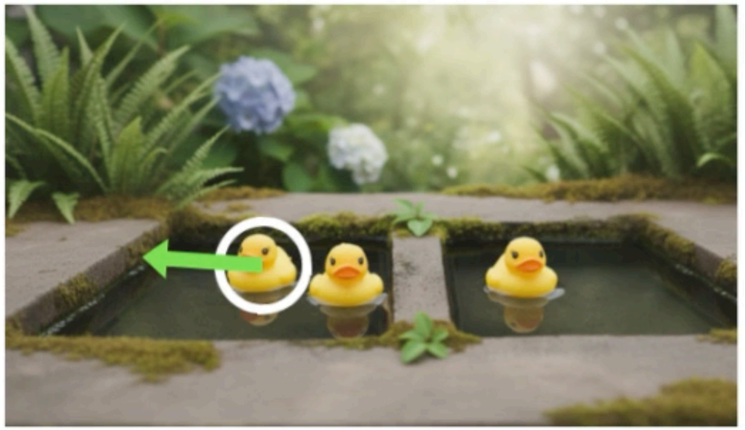
*In order to knock over the domino in the direction of the **goal force prompt**, it must be hit by one of the dominos in the line on the left, because the dominos on the right are separated by a gap.*

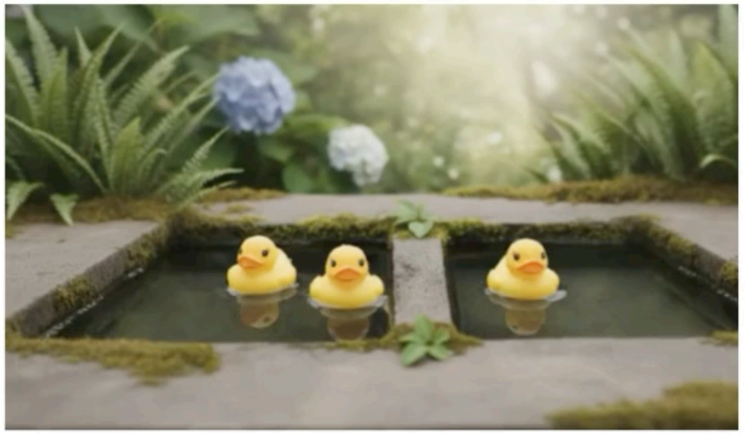
Goal Force prompting generates physically plausible plans





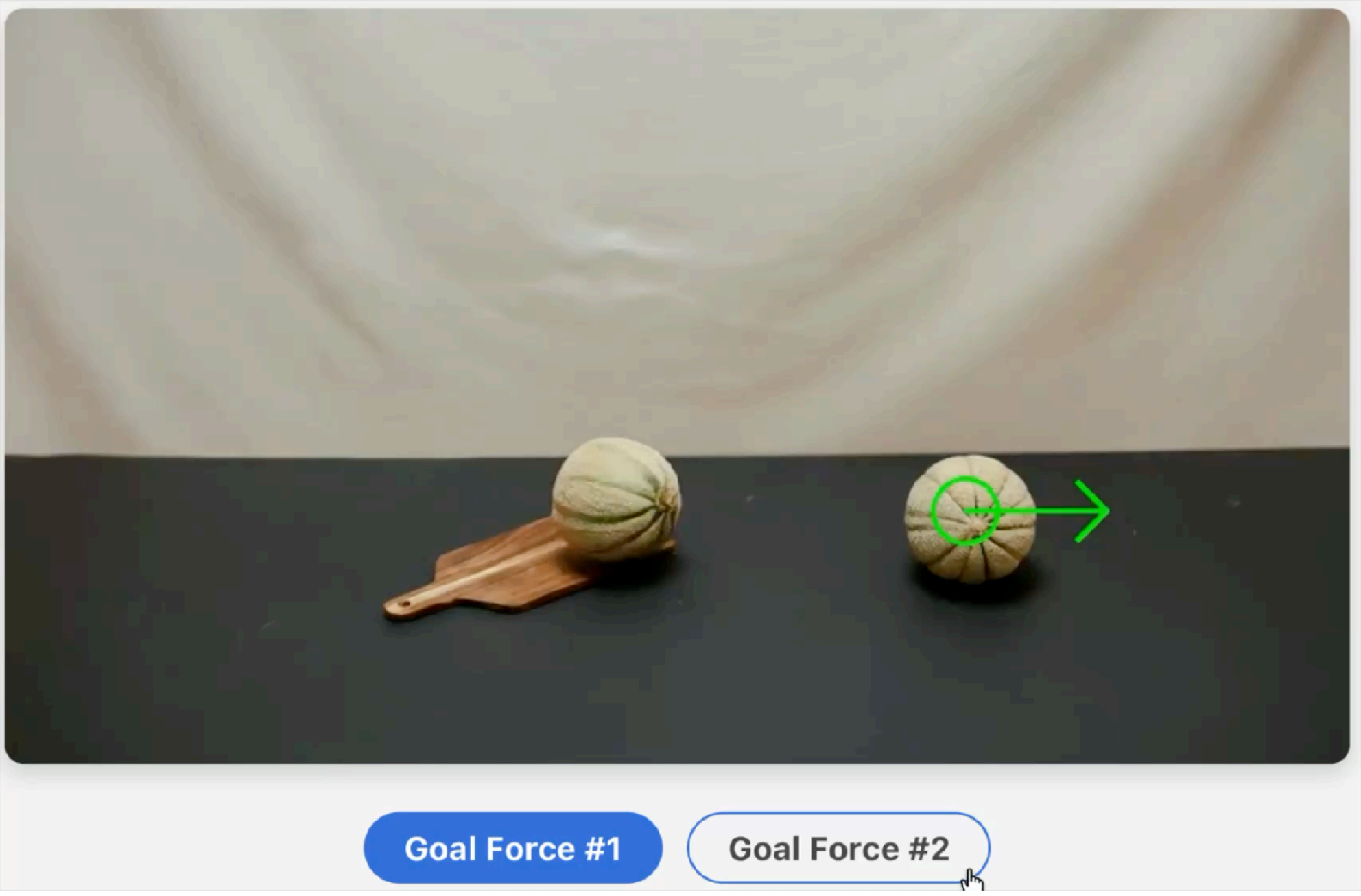
*In order to knock over the domino in the direction of the **goal force prompt**, it must be hit by one of the dominos in the line on the left, because the dominos on the right are separated by a gap.*





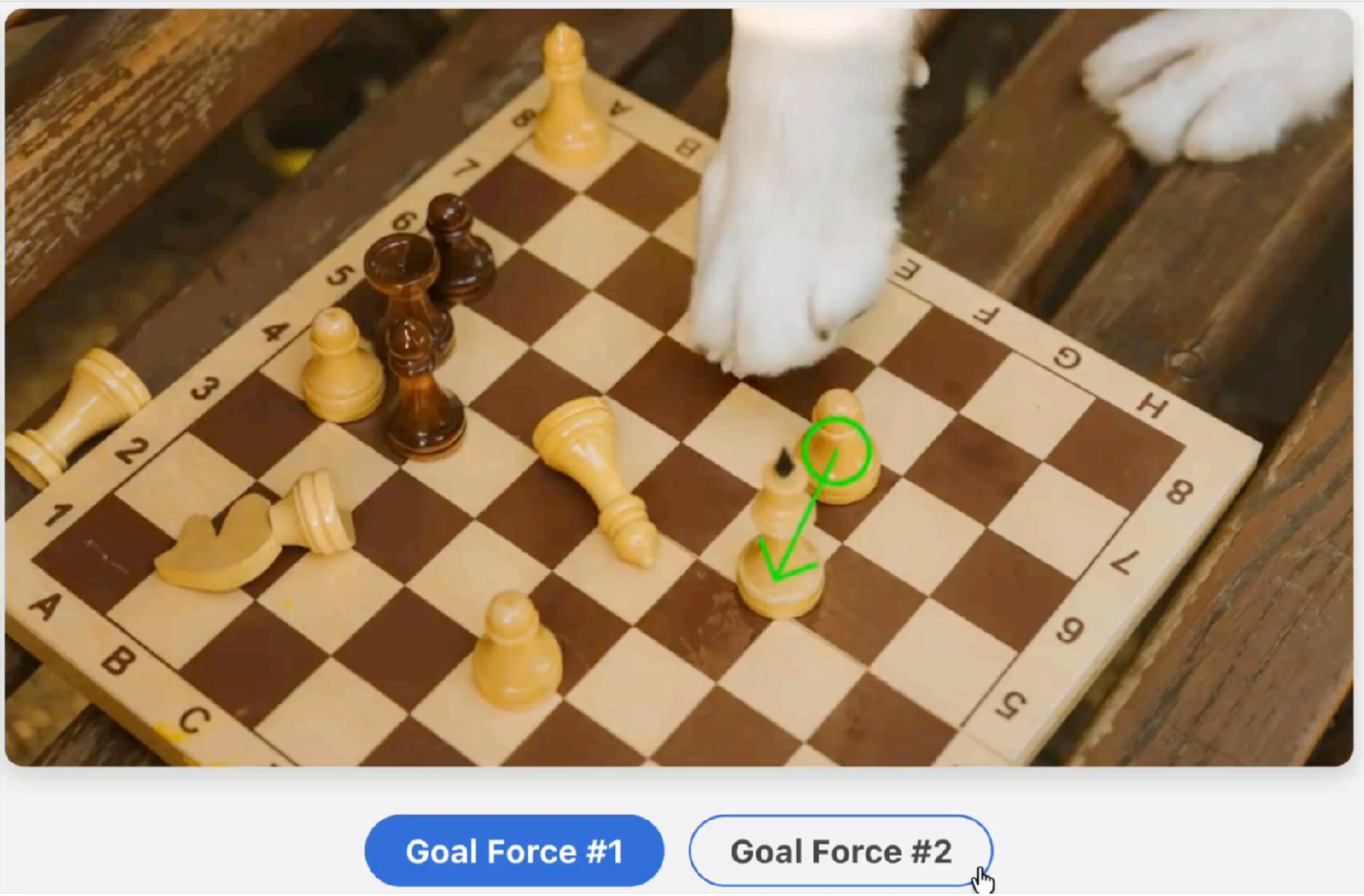
*In order to move the rubber duck in the direction of the **goal force prompt**, it must be hit by the center duck, because the path from the other duck is blocked by a concrete barrier.*

Goal Force prompting generates videos of multi-object collisions…



Goal Force prompting generates videos of human-object interactions…

Goal Force prompting generates videos of (non-human)-object interactions…



Limitations of Goal Force

Failure case: melting / distortion artifacts due to base video model's limitations

While the **goal force** is executed successfully, the base video model's limitations sometimes cause it to generate physically implausible collisions, where objects exhibit visual artifacts like "melting" or distortion.

Limitations of Video World Models

Do generative video models understand physical principles?

Saman Motamed[a,1], Laura Culp[b], Kevin Swersky[b], Priyank Jaini[b,†], and Robert Geirhos[b,†]

[a]INSAIT, Sofia University; work done while at Google DeepMind; [b]Google DeepMind; [†]Joint last authors.

*"We find that across a range of current models … physical understanding is **severely limited**, and **unrelated to visual realism**"*

Representation Learning for Spatiotemporal Physical Systems

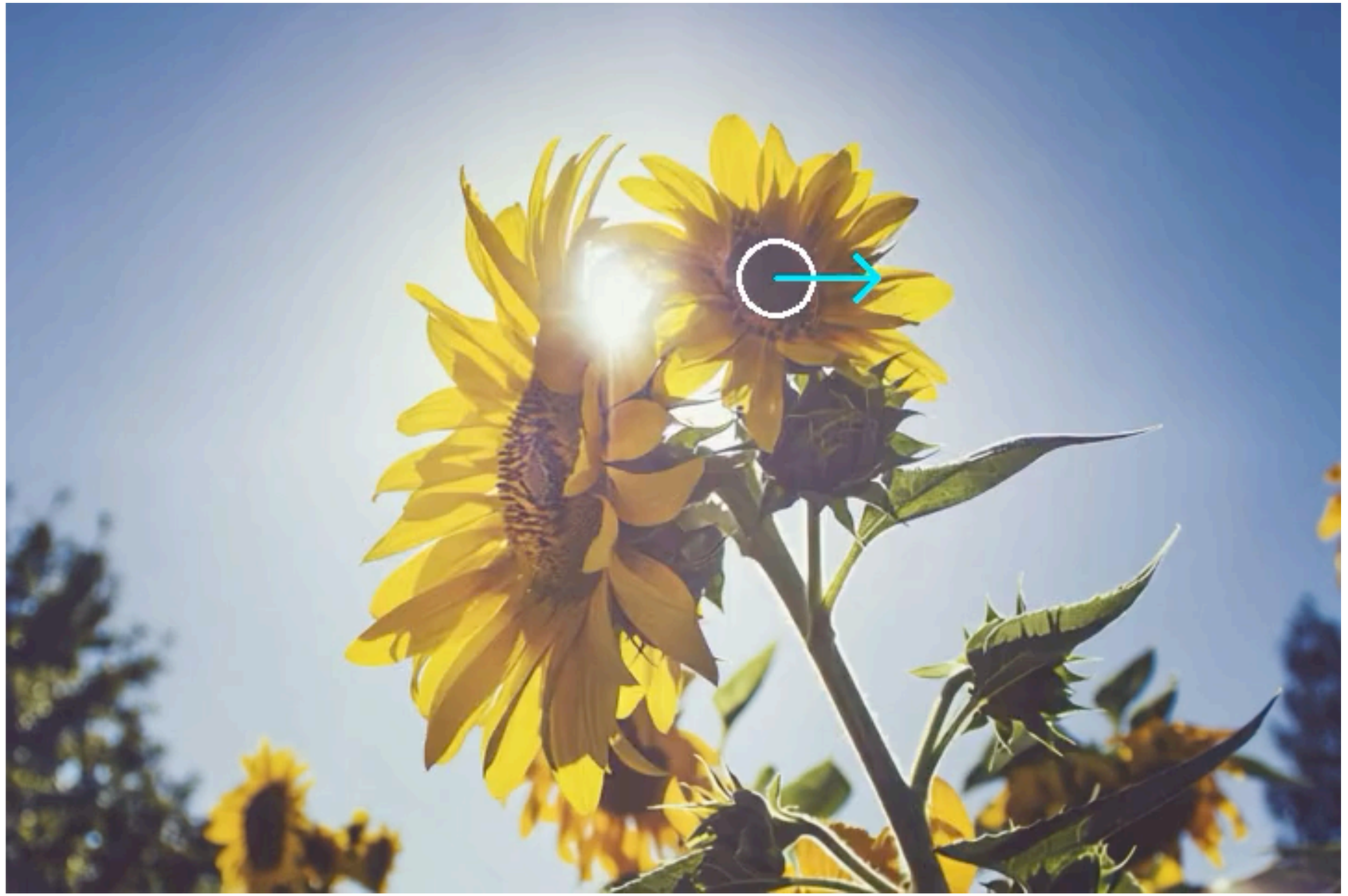
Helen Qu[1†] Rudy Morel[1] Michael McCabe[1,4] Alberto Bietti[1] François Lanusse[2] Shirley Ho[1,3,4] Yann LeCun[3]

The Polymathic AI Collaboration

[1]Flatiron Institute
[2]Université Paris-Saclay, Université Paris Cité, CEA, CNRS, AIM
[3]New York University
[4]Princeton University

*"Surprisingly, we find that … methods that learn in the latent space … outperform those optimizing **pixel-level** prediction objectives"*

Thanks for listening! Your Feedback Welcomed!



Project page with interactive demos, code, paper

force-prompting.github.io

goal-force.github.io

diffusion-supervision.github.io/silvr/