EFSTRATIOS GAVVES

GROUNDING FOUNDATION MODELS IN REALITY

PHYSICS- & CAUSALITY-INFORMED WORLD MODELS

Advanced School for Computing and Imaging

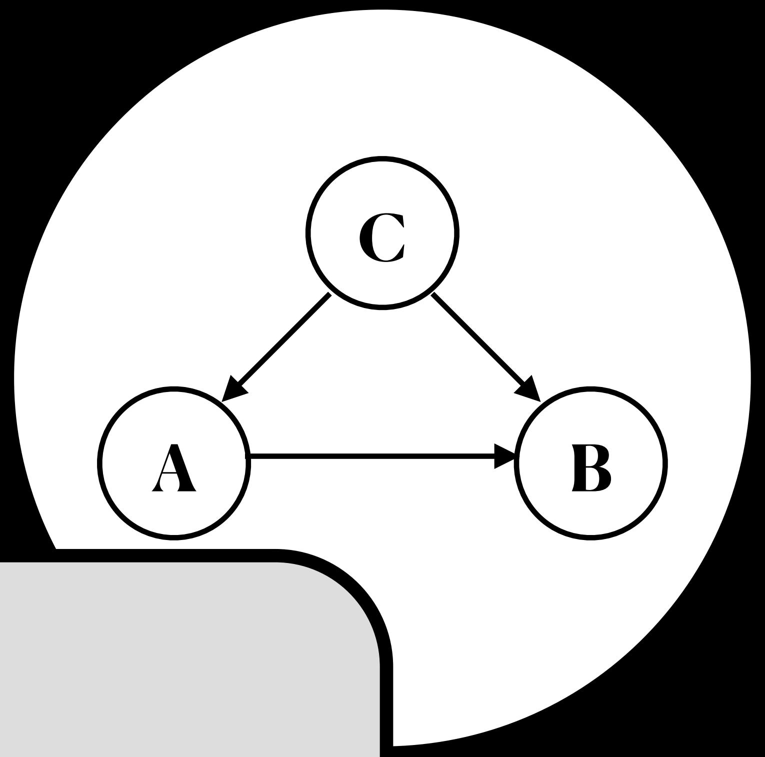
WHO AM I?

- Associate Professor Deep Learning
- ERC StG & NWO VIDI
- ELLIS Scholar
- Co-director of ICAI QUVA & POP-AART
- Co-founder of Ellogon.AI and LYDS Partners

WHO AM I?

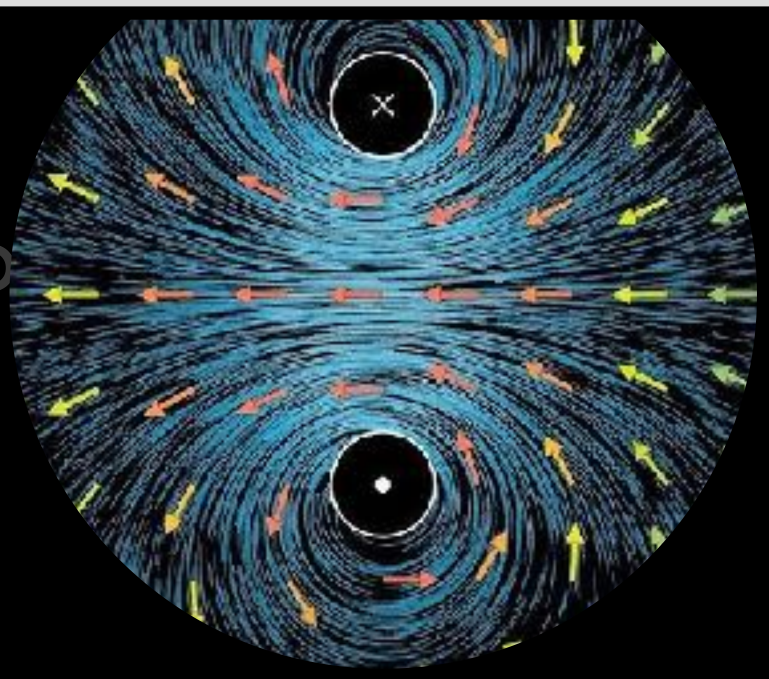
Robot Learning

Causal Representation Learning



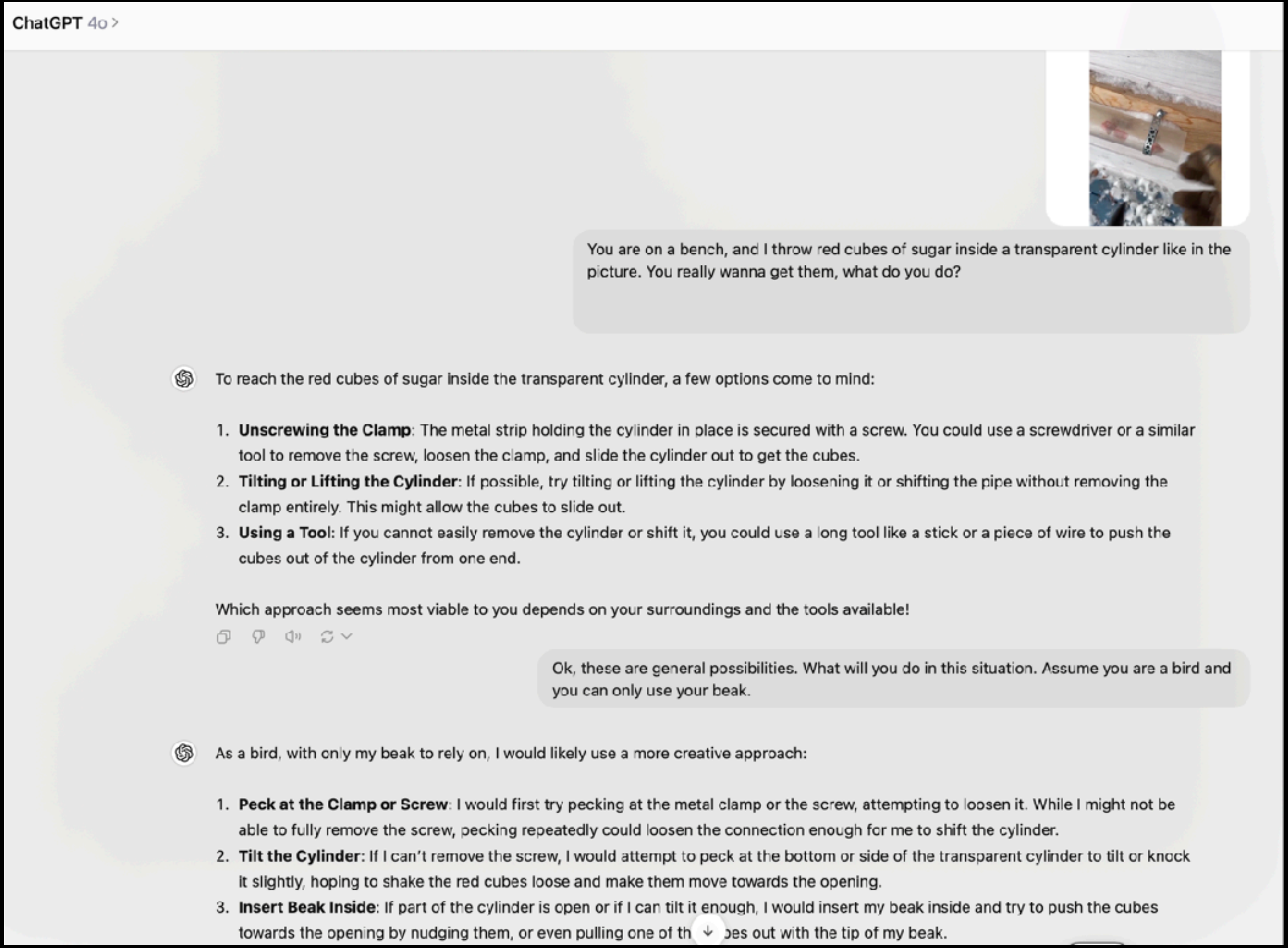
- Associate Profe earning
- ERC StG & NWO VIDI
- ELLIS Scholar
- Co-director of ICAI QUVA & POP
- Co-founder of Ellogon.AI

Dynamical Deep Learning



AI & Science

CROW VS FRONTIER MODELS



You are on a bench, and I throw red cubes of sugar inside a transparent cylinder like in the picture. You really wanna get them, what do you do?

To reach the red cubes of sugar inside the transparent cylinder, a few options come to mind:

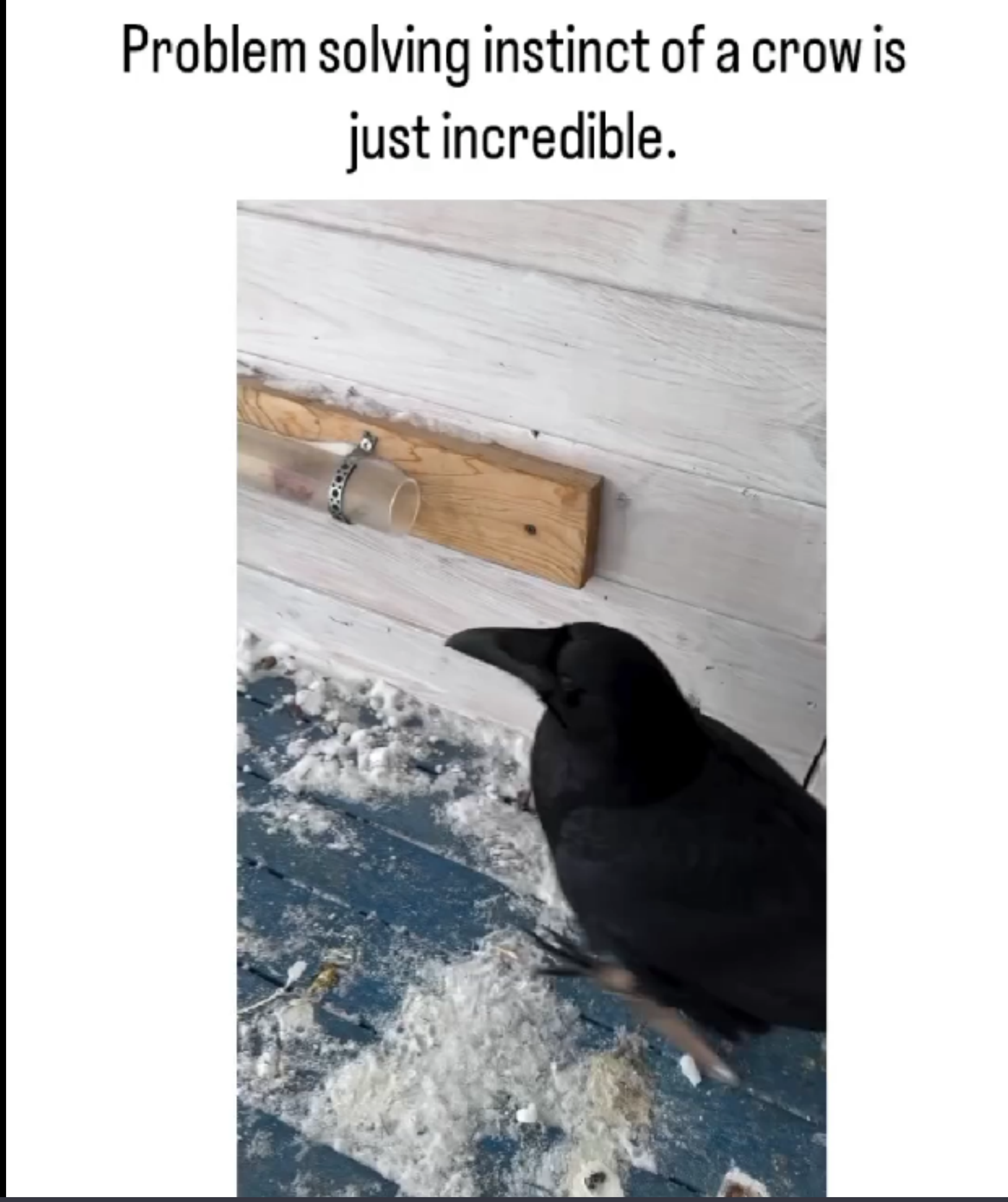
1. **Unscrewing the Clamp**: The metal strip holding the cylinder in place is secured with a screw. You could use a screwdriver or a similar tool to remove the screw, loosen the clamp, and slide the cylinder out to get the cubes.
2. **Tilting or Lifting the Cylinder**: If possible, try tilting or lifting the cylinder by loosening it or shifting the pipe without removing the clamp entirely. This might allow the cubes to slide out.
3. **Using a Tool**: If you cannot easily remove the cylinder or shift it, you could use a long tool like a stick or a piece of wire to push the cubes out of the cylinder from one end.

Which approach seems most viable to you depends on your surroundings and the tools available!

Ok, these are general possibilities. What will you do in this situation. Assume you are a bird and you can only use your beak.

As a bird, with only my beak to rely on, I would likely use a more creative approach:

1. **Peck at the Clamp or Screw**: I would first try pecking at the metal clamp or the screw, attempting to loosen it. While I might not be able to fully remove the screw, pecking repeatedly could loosen the connection enough for me to shift the cylinder.
2. **Tilt the Cylinder**: If I can't remove the screw, I would attempt to peck at the bottom or side of the transparent cylinder to tilt or knock it slightly, hoping to shake the red cubes loose and make them move towards the opening.
3. **Insert Beak Inside**: If part of the cylinder is open or if I can tilt it enough, I would insert my beak inside and try to push the cubes towards the opening by nudging them, or even pulling one of the cubes out with the tip of my beak.



Problem solving instinct of a crow is just incredible.

Claude 3 vs ChatGPT vs Gemini

QUO VADIS AI?

FROM SEEING TO

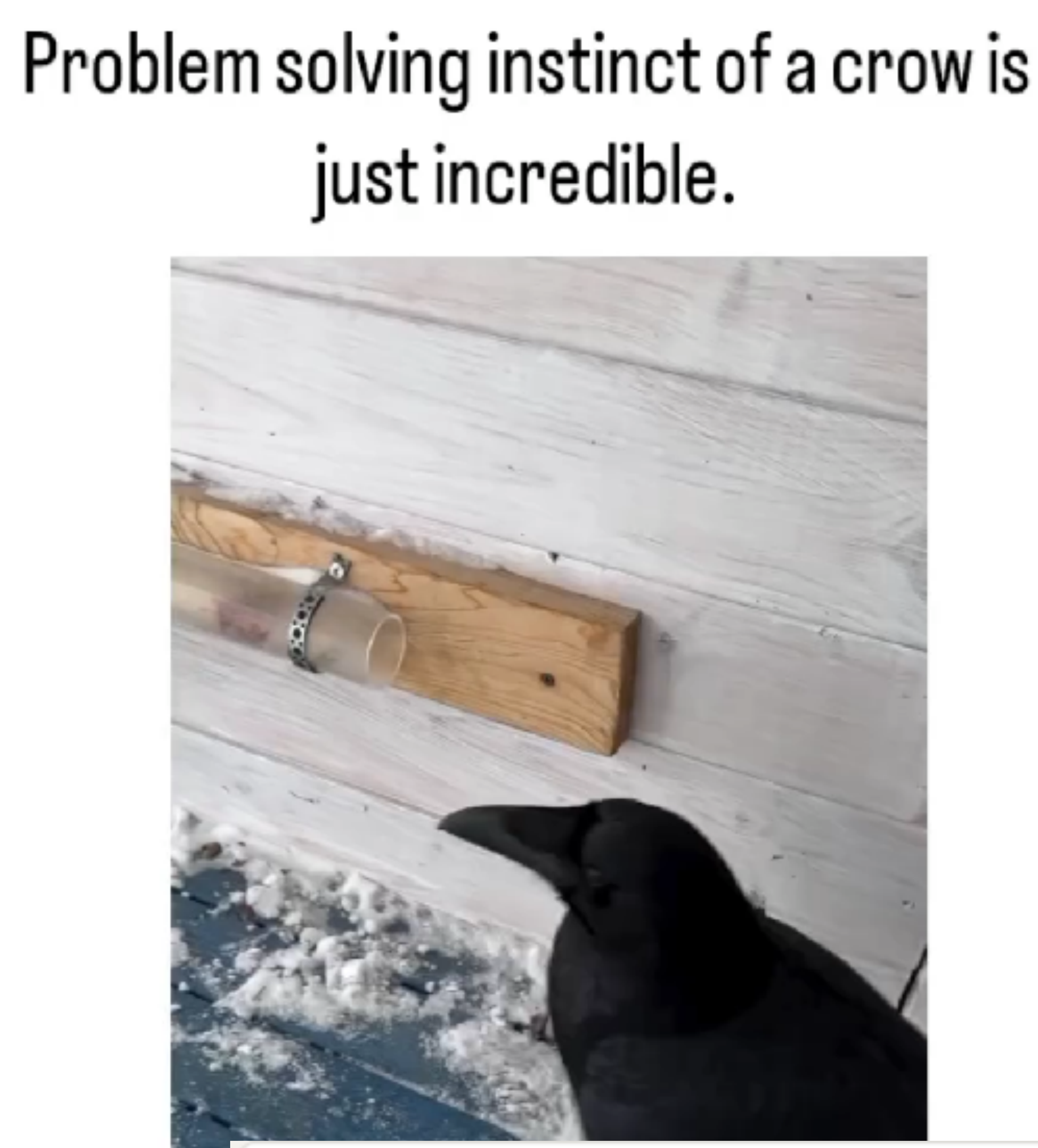
→ INDUCING PHYSICS & CAUSALITY

→ SPATIAL MEMORY & NAVIGATION

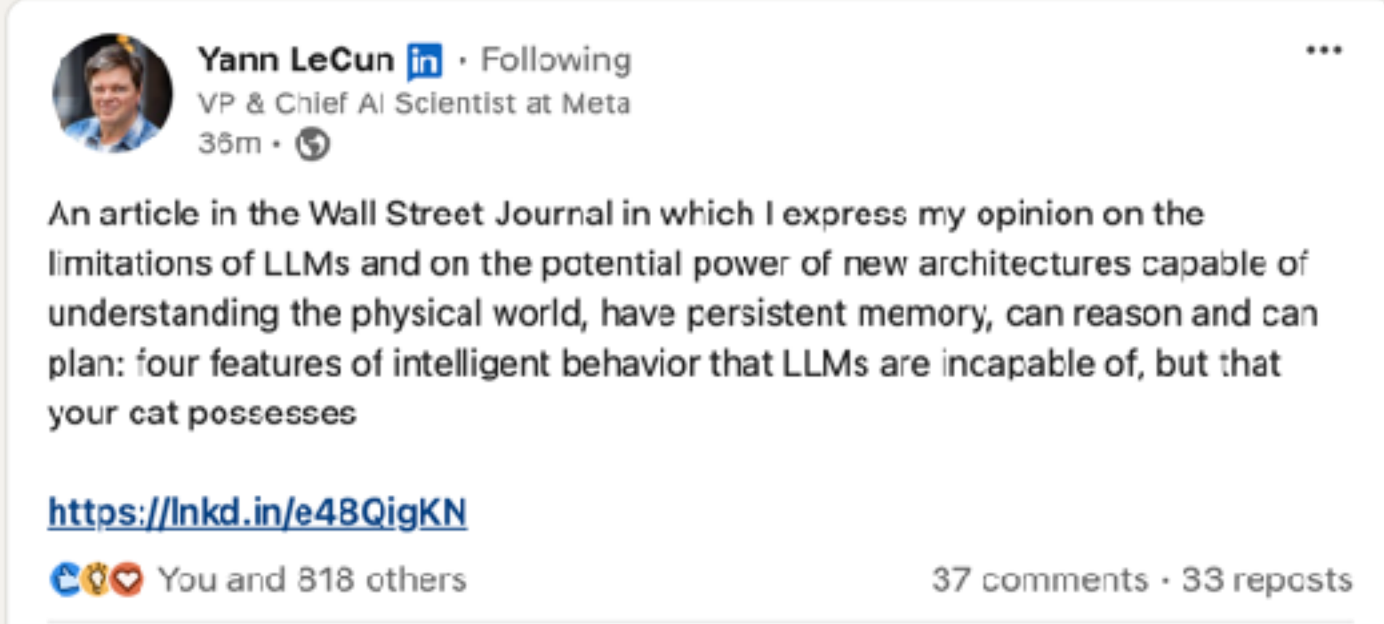
→ REASONING

→ PLANNING

⇒ INTERACTING WITH REALITY



Problem solving instinct of a crow is just incredible.

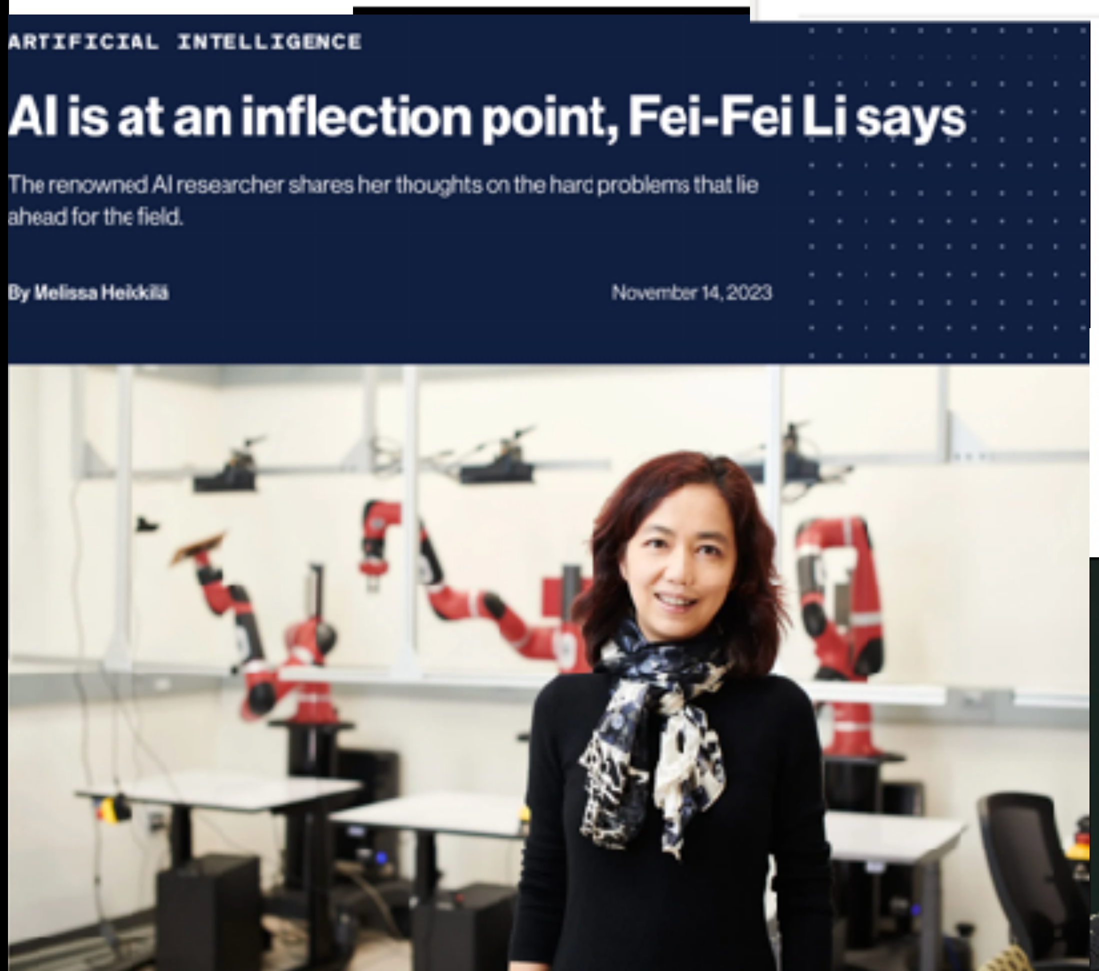


Yann LeCun · Following
VP & Chief AI Scientist at Meta
35m ·

An article in the Wall Street Journal in which I express my opinion on the limitations of LLMs and on the potential power of new architectures capable of understanding the physical world, have persistent memory, can reason and can plan: four features of intelligent behavior that LLMs are incapable of, but that your cat possesses

https://lnkd.in/e48QigKN

You and 818 others 37 comments · 33 reposts

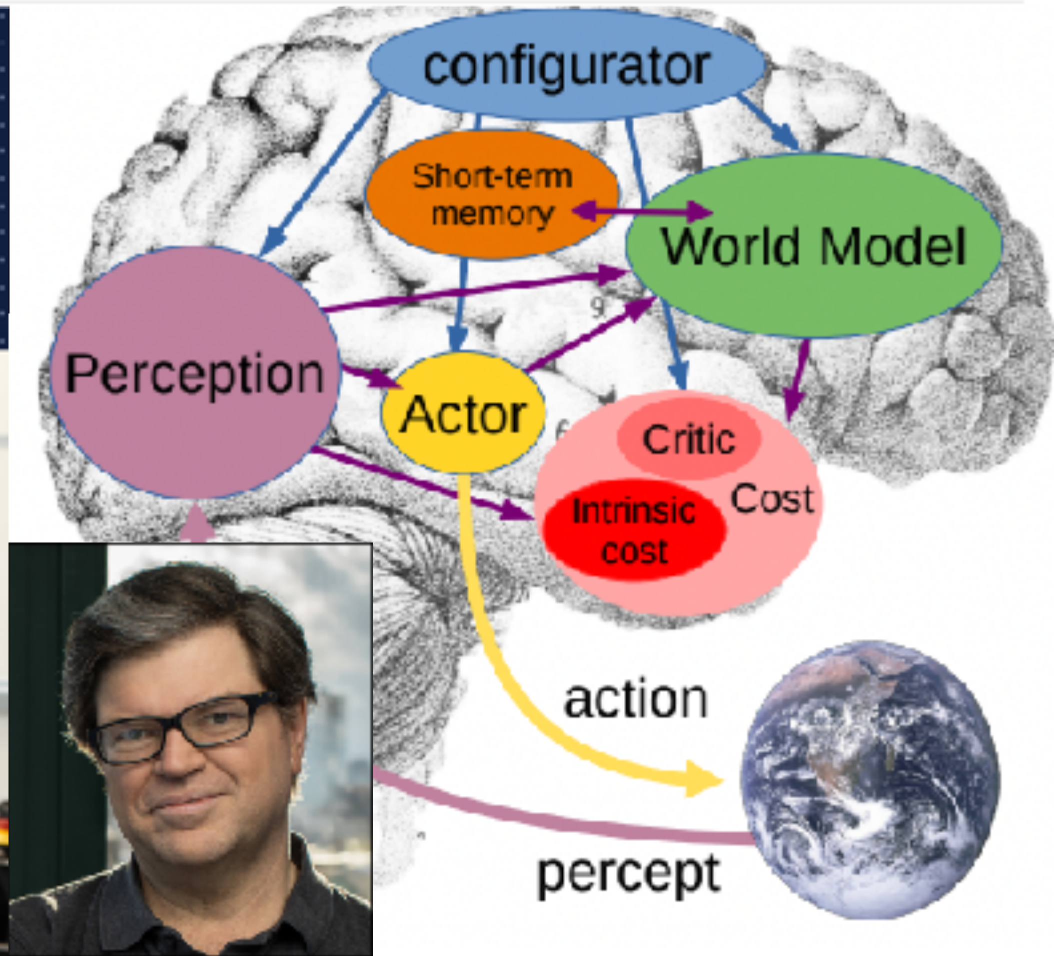


ARTIFICIAL INTELLIGENCE

AI is at an inflection point, Fei-Fei Li says

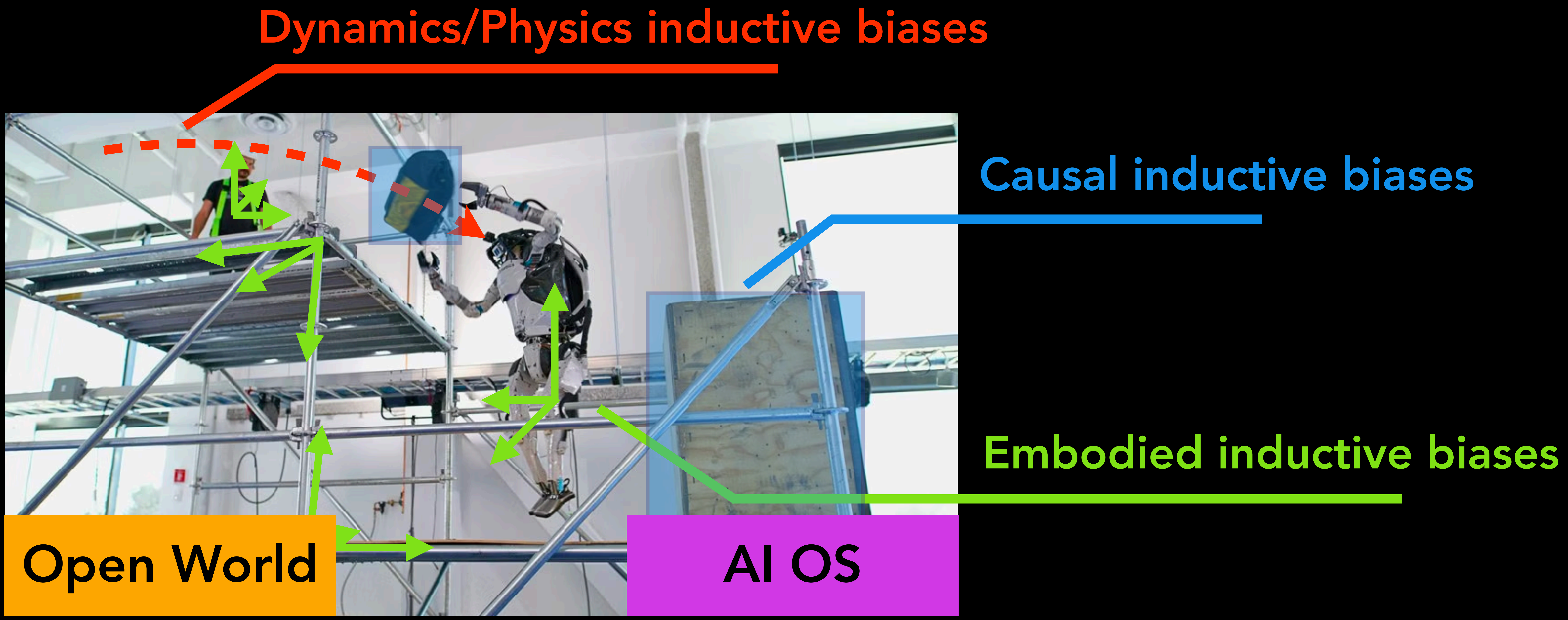
The renowned AI researcher shares her thoughts on the hard problems that lie ahead for the field.

By Melissa Heikkilä November 14, 2023



WITH PHYSICS AND CAUSALITY

GROUNDING ROBOT WORLD MODELS



Dynamics/Physics inductive biases

Causal inductive biases

Embodied inductive biases

Open World

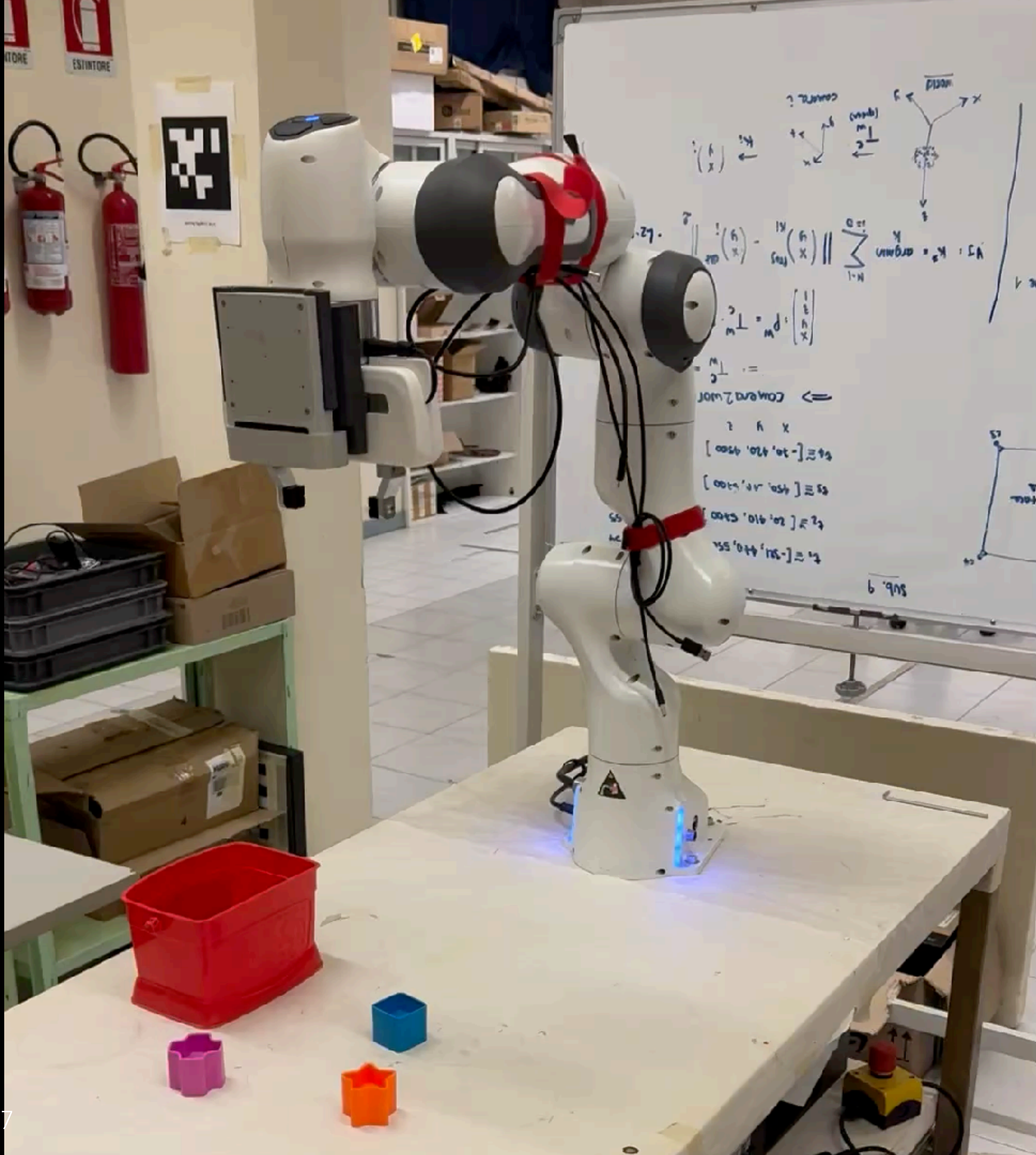
AI OS

DREAM TO MANIPULATE

COMPOSITIONAL MANIPULATION WORLD MODELS

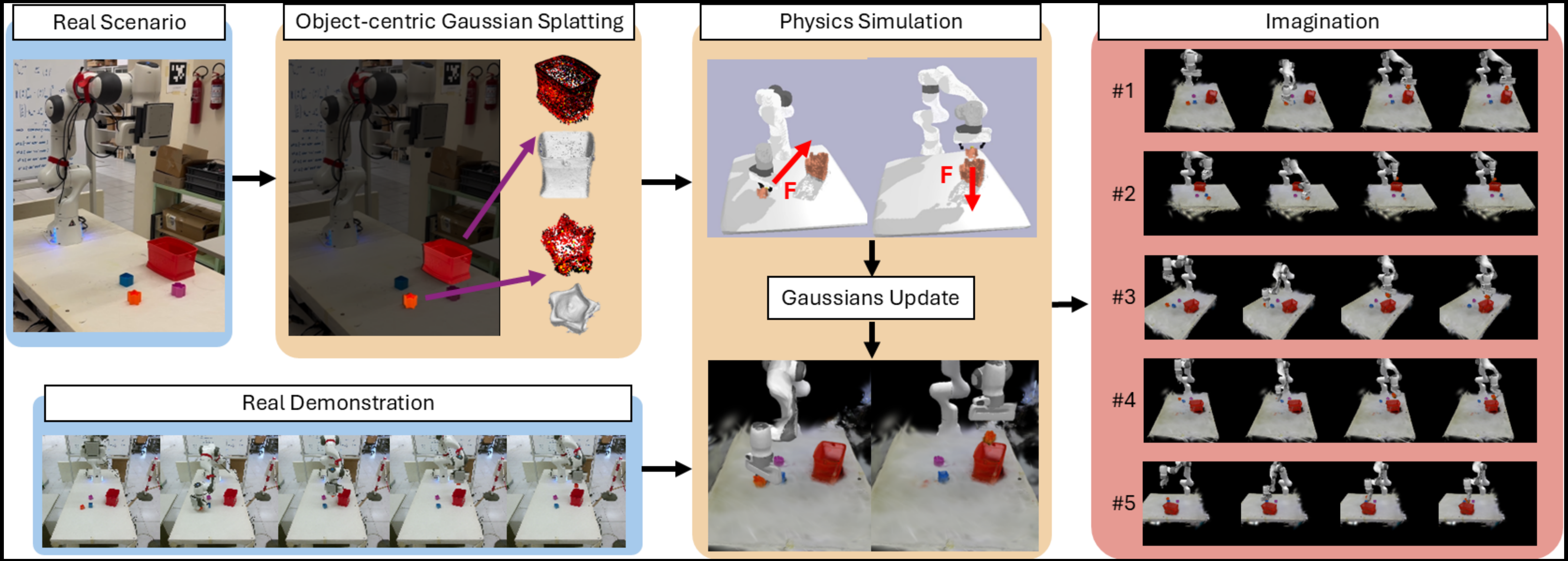
https://dreamtomanipulate.github.io/DreMa/

w. L. Barcellona, A. Zadaianchuk, D. Allegro, S. Papa, S. Ghidoni



COMPOSITIONAL MANIPULATION WORLDS

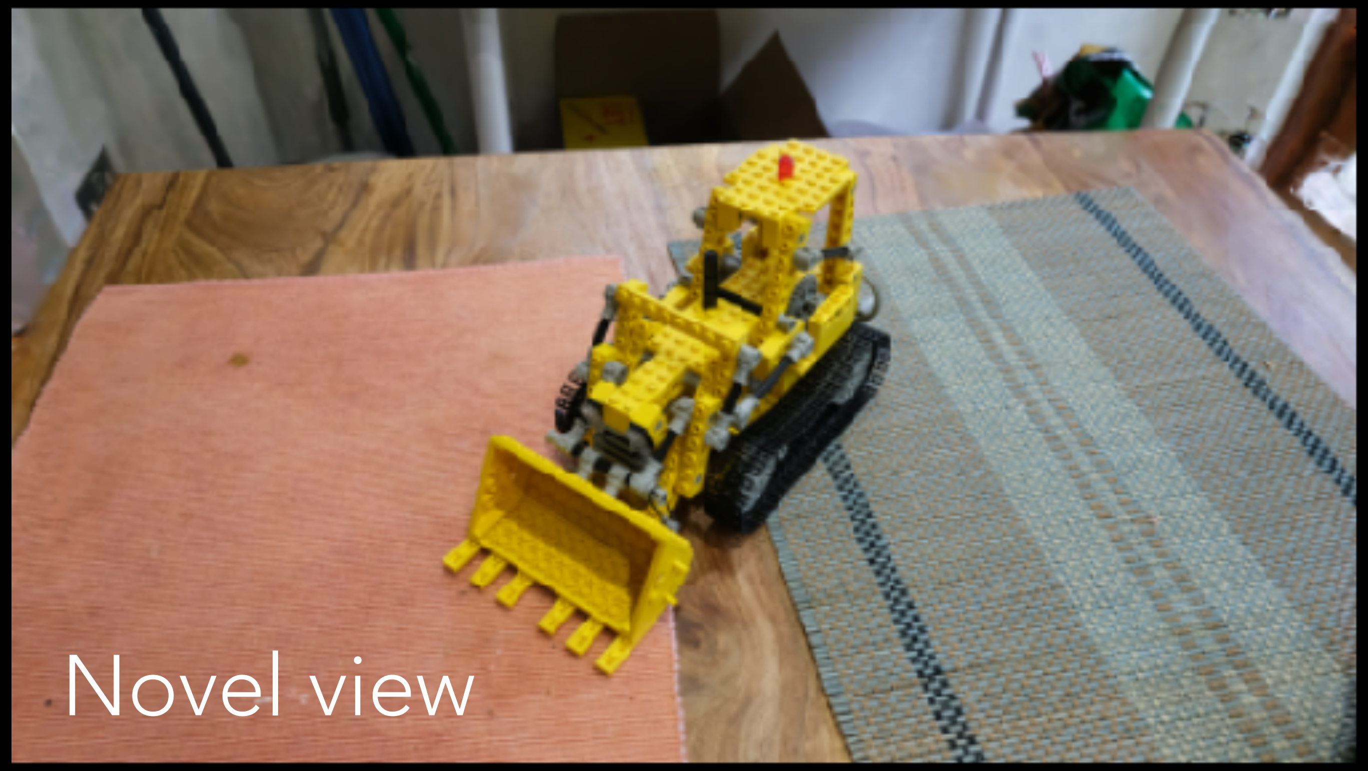
OUR INDUCTIVE BIASES

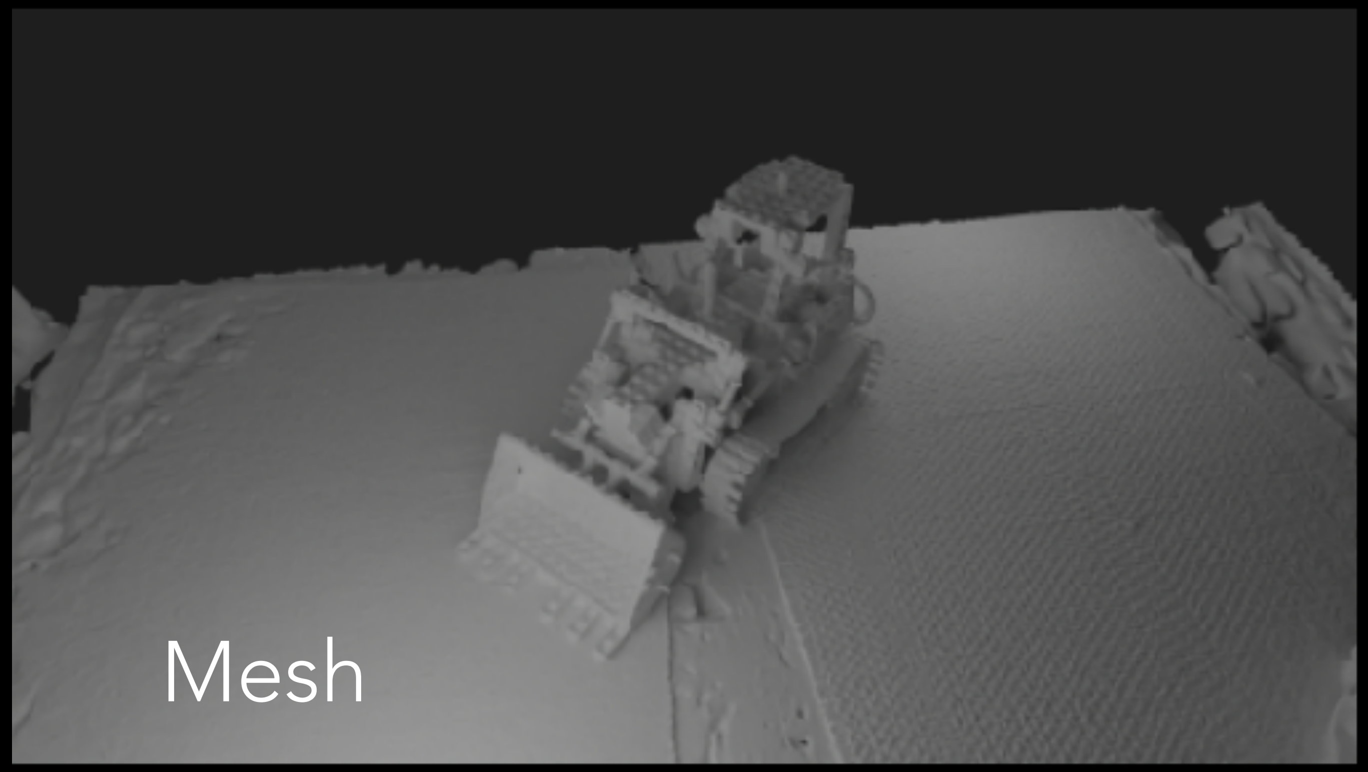


EXPLICIT GROUNDING == PHOTOREALISTIC RECON@OBJECT-CENTRIC+ EXPLICIT PHYSICS
⇒ COMPOSITIONAL MANIPULATION WORLD MODELS ⇒ IMAGINATION
⇒↑ ROBOT IMITATION LEARNING

PHOTOREALISTIC RECONSTRUCTION …

- Gaussian Splatting is like ‘sparse 3D pixels’
- Real-time rendering
- High-quality
- Good depth rendering
- Fast training
- Explicitly grounded representation





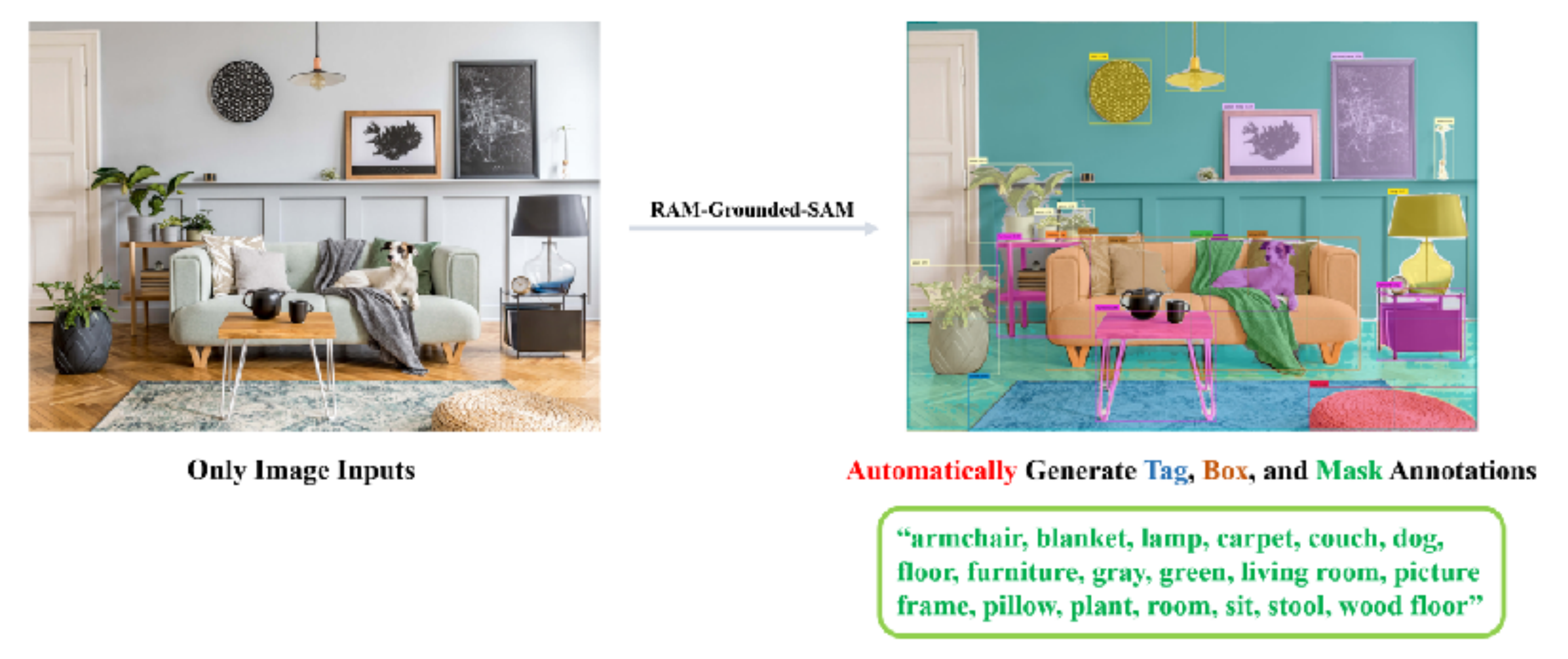
… MADE OBJECT-CENTRIC WITH OBJECT ASSETS…

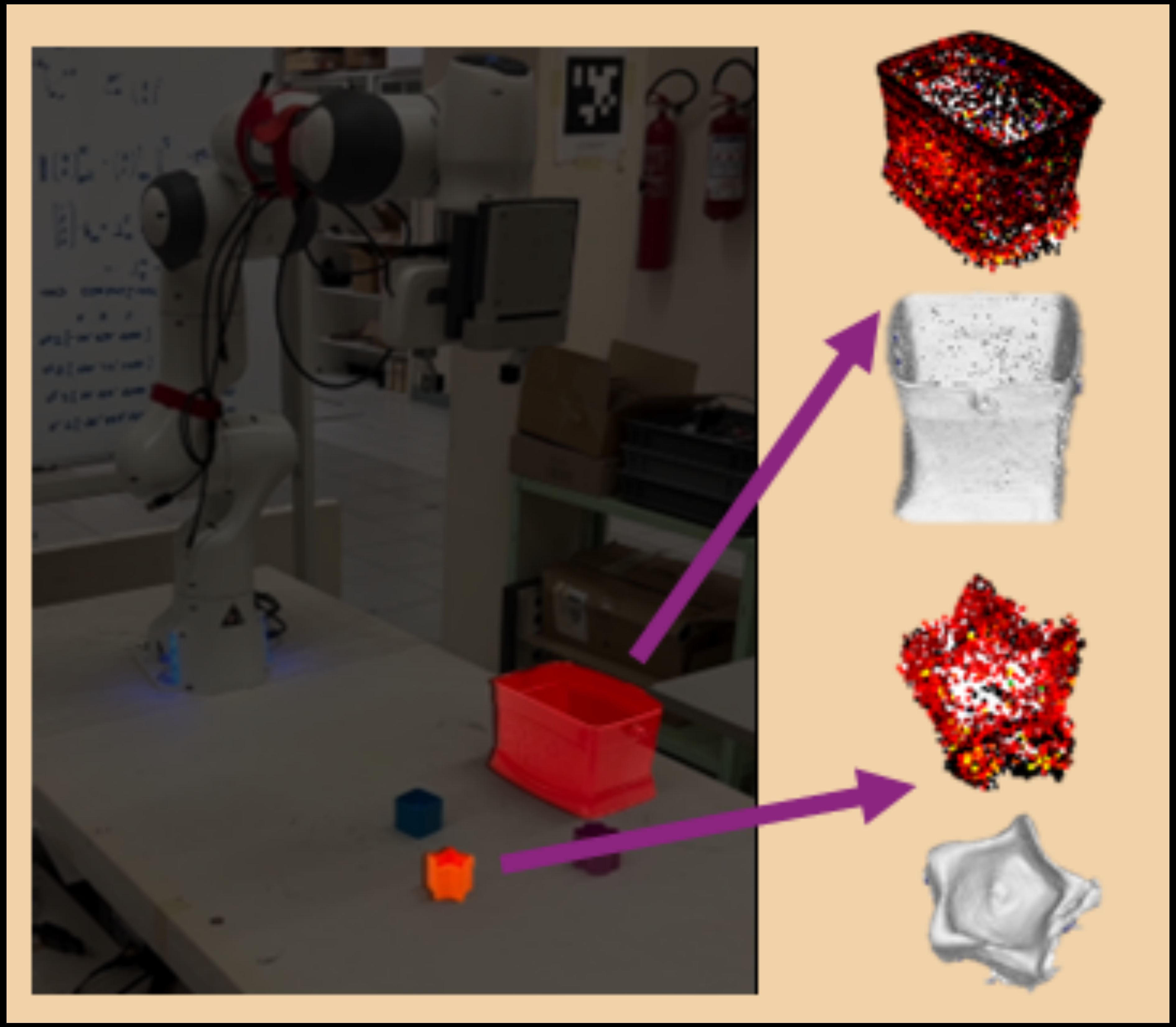
- Zero-shot object localization

→ Grounded SAM, DEVA, or our VISA[1] & LV-VIS[2]

- Prompts: "`object`" & "`table`"

- Segment and group objects across views


Only Image Inputs
RAM-Grounded-SAM
Automatically Generate Tag, Box, and Mask Annotations
"armchair, blanket, lamp, carpet, couch, dog, floor, furniture, gray, green, living room, picture frame, pillow, plant, room, sit, stool, wood floor"



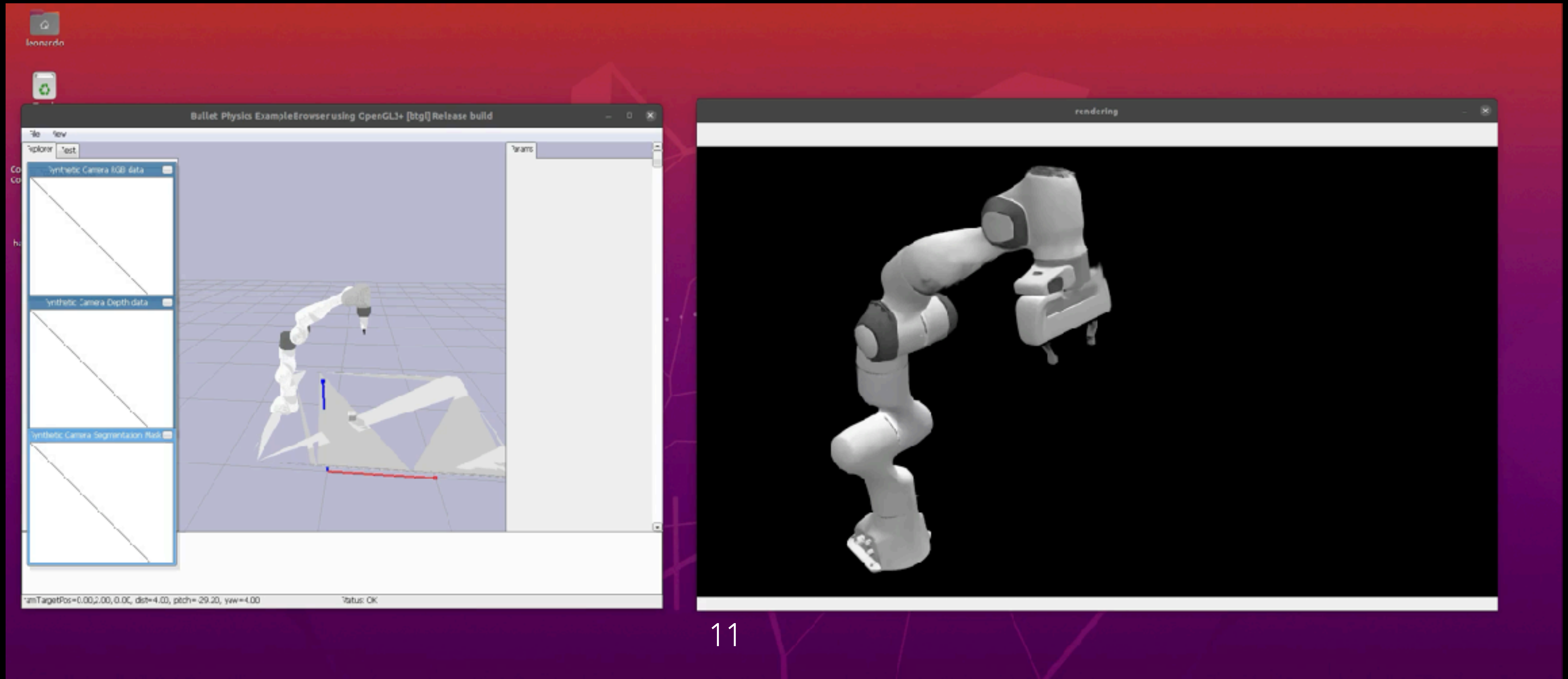
---

[1] Wang, Gavves et al., Towards Open-Vocabulary Video Instance Segmentation, ICCV 2023

[2] Yan, Wang, Gavves et al., VISA: Reasoning Video Object Segmentation via Large Language Models, ECCV 2024

… WITH "SELF-AWARENESS" …

- The robot is also an object asset
- Articulated but with known joints
- Movement given by the URDF file, no need to estimate


leonardo
Bullet Physics ExampleBrowser using OpenGL3+ [btgl] Release build
ile
iew
xplorer
est
ynthetic Camera RGB data
ynthetic Camera Depth data
ynthetic Camera Segmentation Mask
arams
rendering

amTargetPos=0.00,2.00,0.00, dist=4.00, pitch=-29.20, yaw=4.00

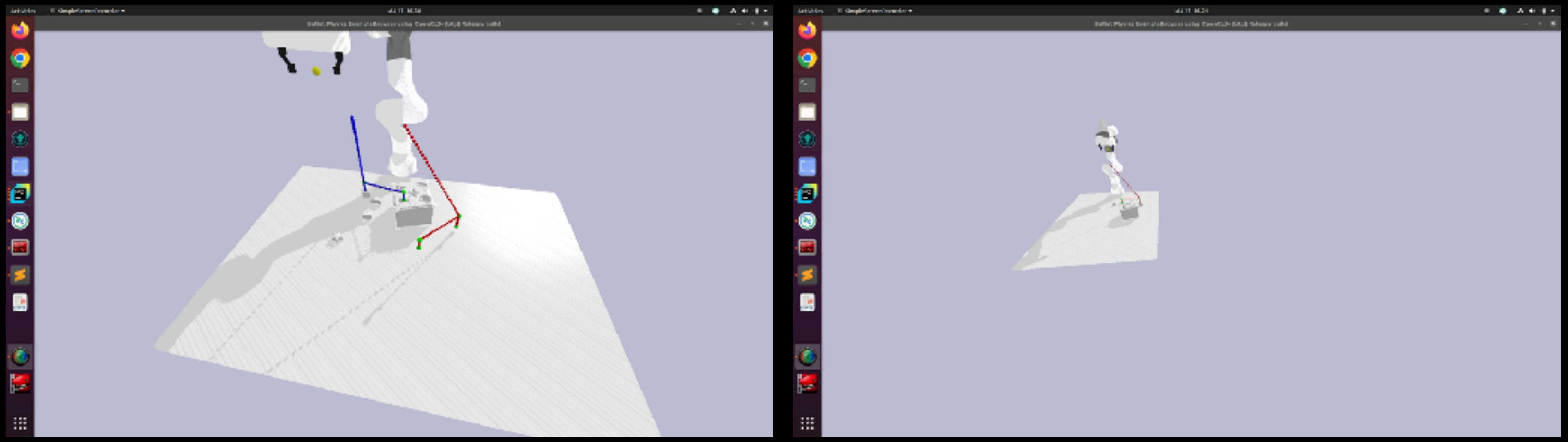
tatus: OK

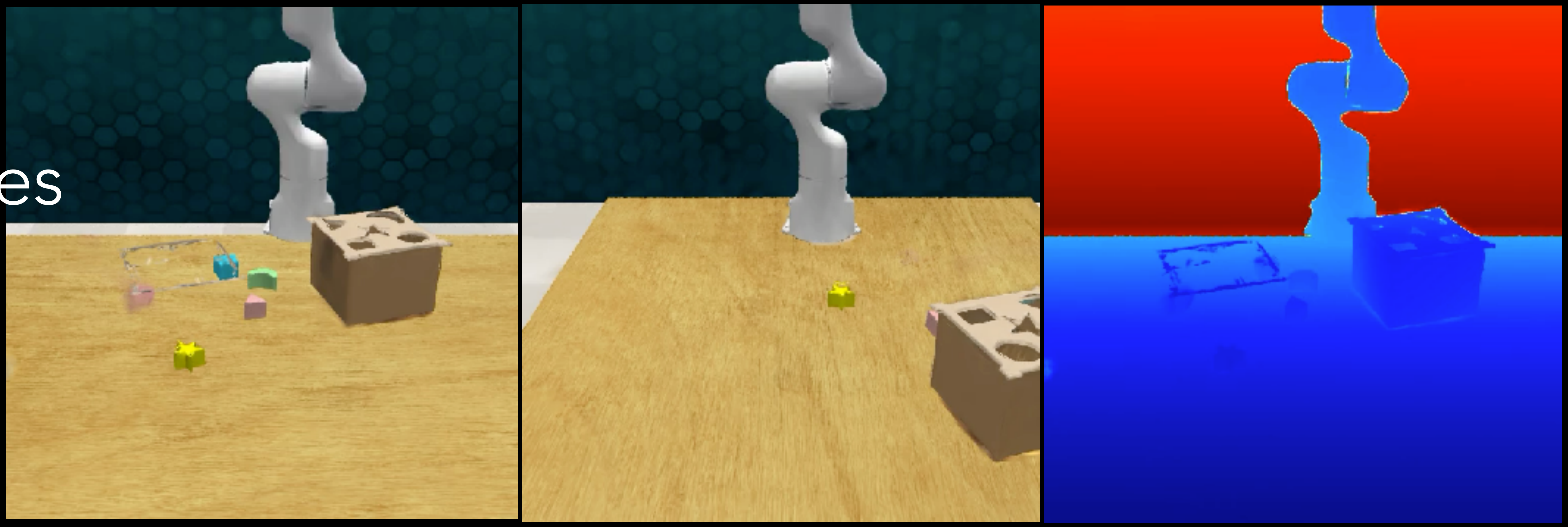
… AND MANIPULABLE WITH PYBULLET …

- Integrate explicit physics engines
- Manipulate object assets by exerting forces on them
- PyBullet requires mesh grids
- Convert Gaussian Splats to Meshes
- "Decode" effect of manipulations with Gaussian Splatting

… TO RECONSTRUCT PAST TRAJECTORIES …

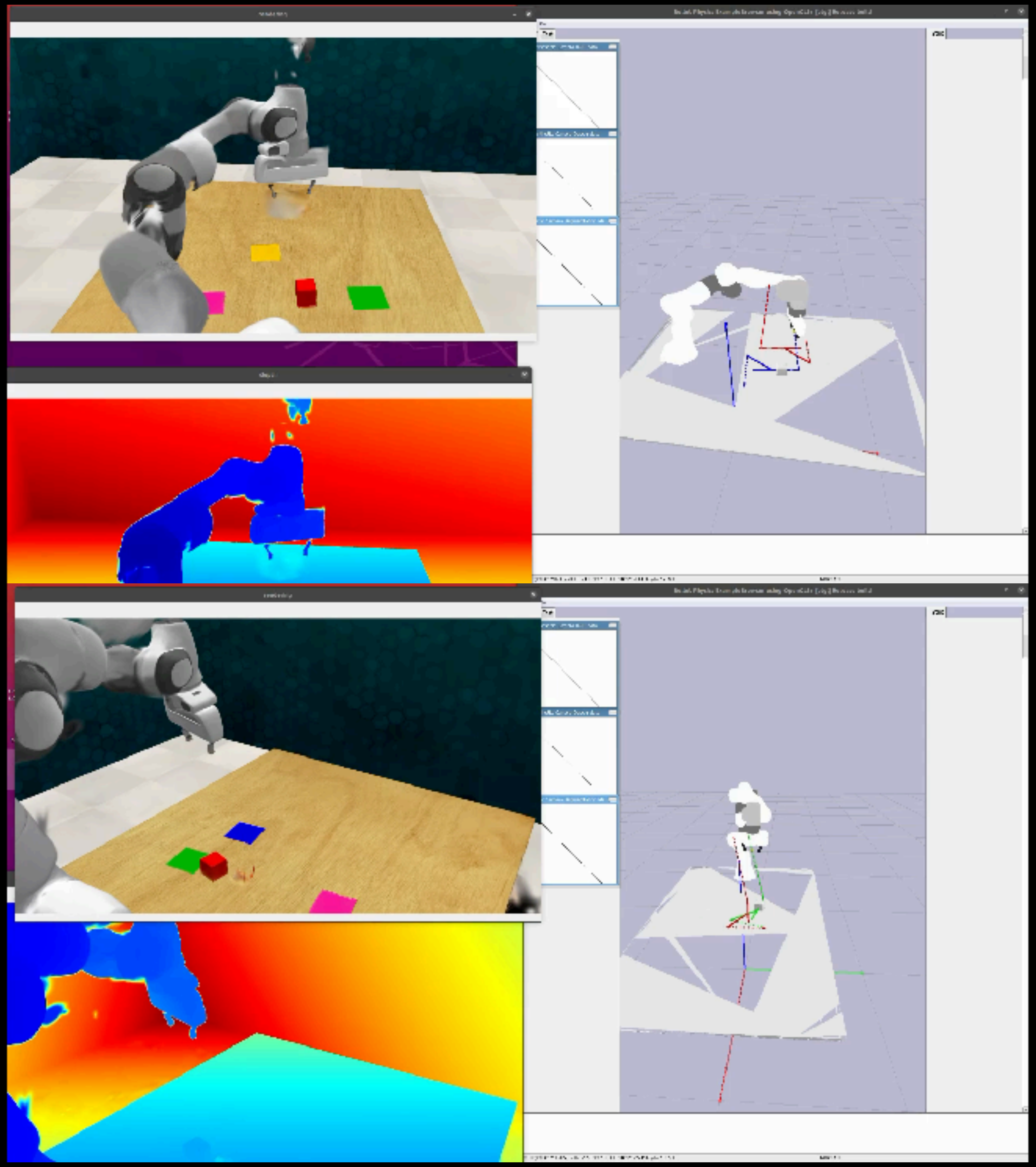
- Play seen trajectories
- And render it from any angle
- “Re-imagining” past experiences





… AND IMAGINE NEW ONES …

- Since we have a 'digital twin'
- That reconstructs photorealistically
- Understands physics
- And can be intervened with
- We can "imagine" novel trajectories



EQUIVARIANT OBJECT-ROBOT TRANSFORMATIONS

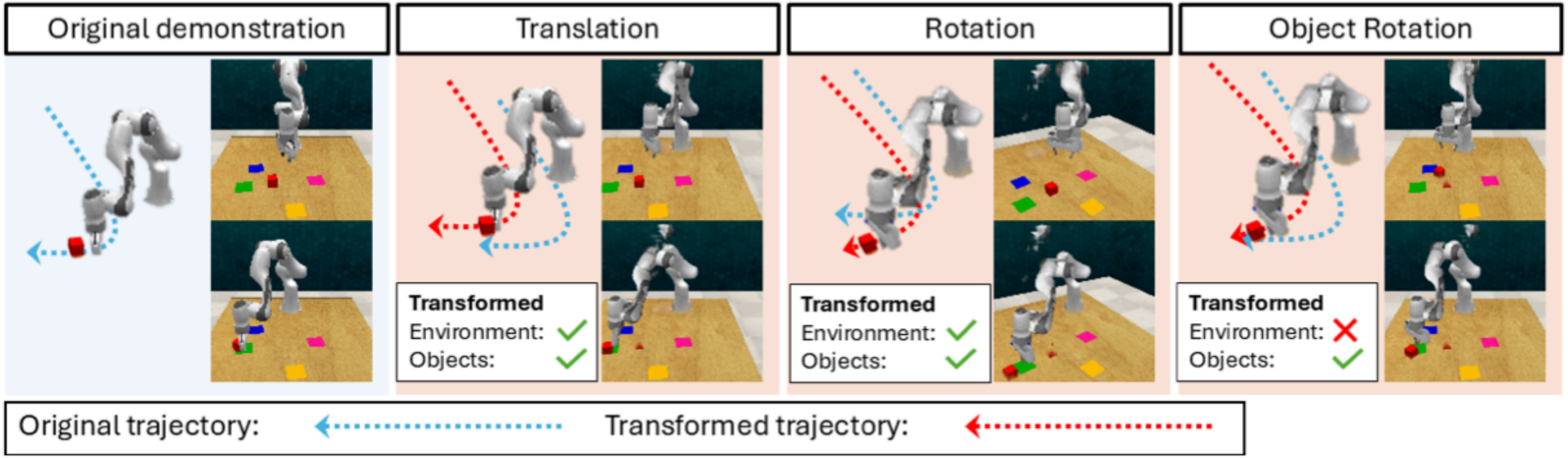


Figure 3: The effect of equivariant translation, equivariant rotation, and the object rotation transformations. Top row: start of demonstration. Bottom row: target of demonstration.

… FOR ONE-SHOT IMITATION LEARNING …

- Single demonstration per variant
- Excluding tasks with articulation
- 60-110 imaginations generated
- Works even with pure imagination

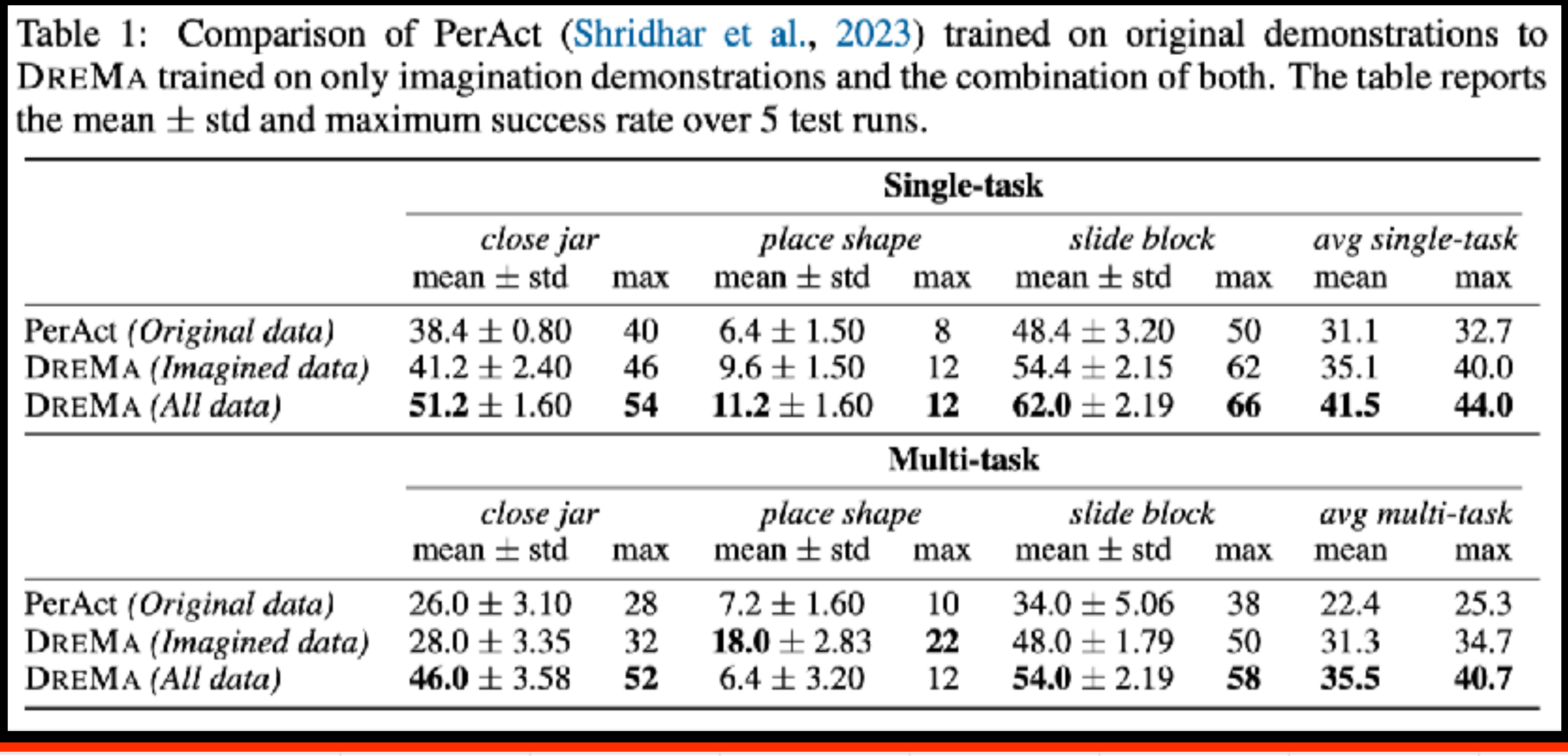
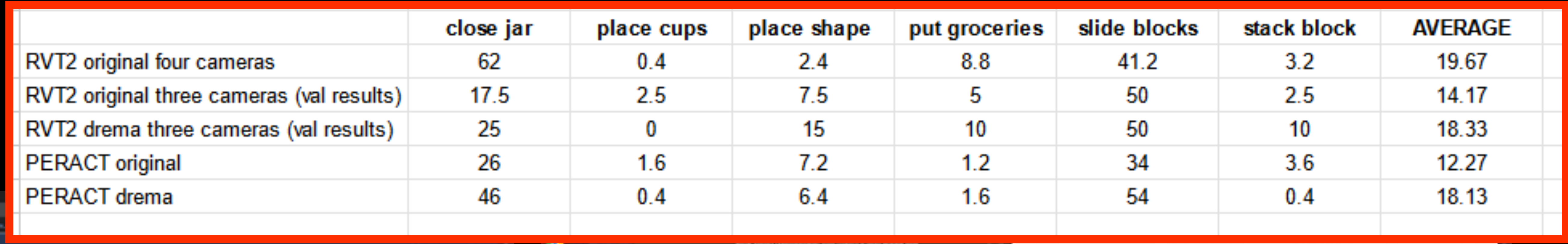
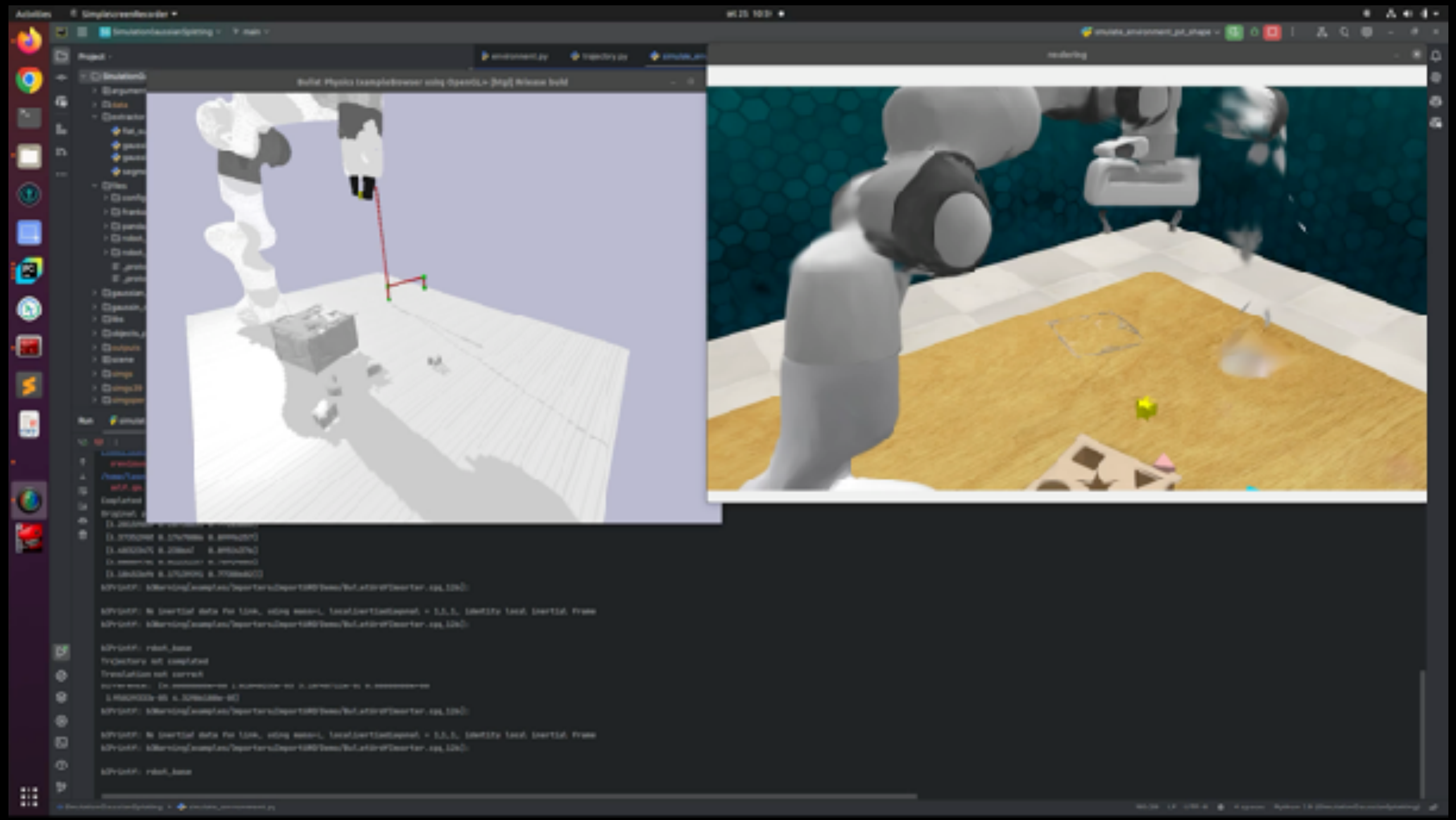


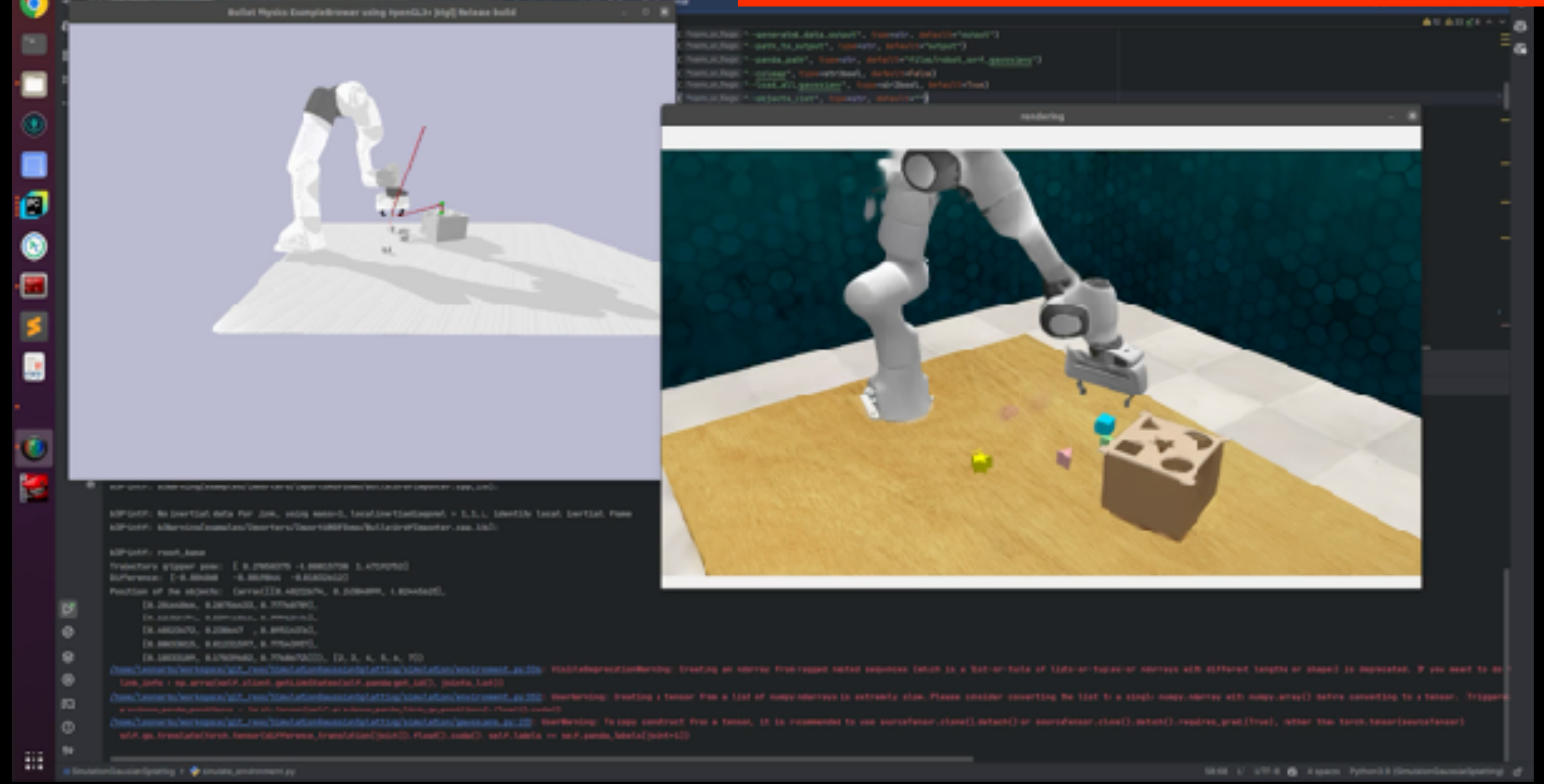
Table 1: Comparison of PerAct (Shridhar et al., 2023) trained on original demonstrations to DREMA trained on only imagination demonstrations and the combination of both. The table reports the mean ± std and maximum success rate over 5 test runs.

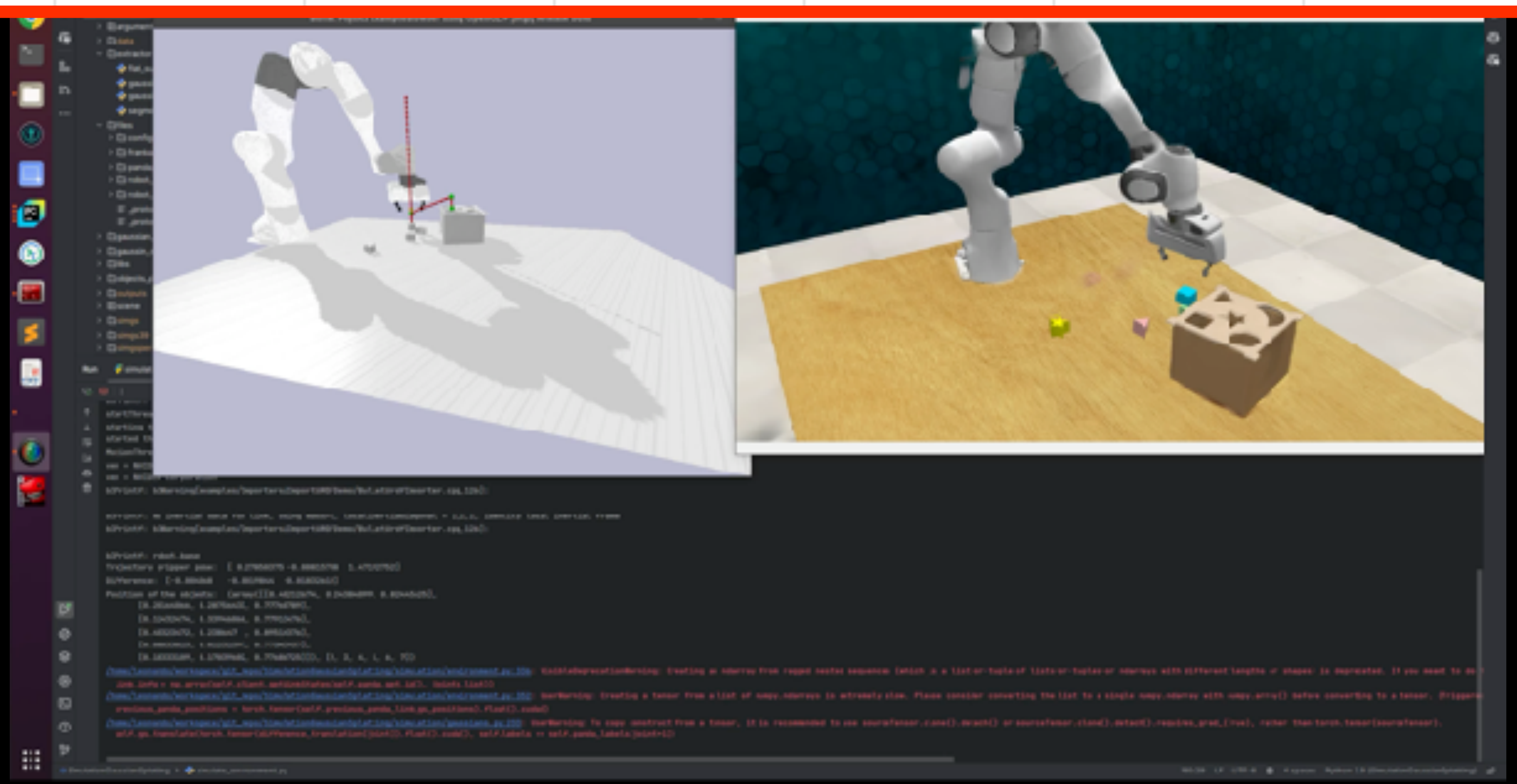
	Single-task							
	close jar		*place shape*		*slide block*		*avg single-task*	
	mean ± std	max	mean ± std	max	mean ± std	max	mean	max
PerAct *(Original data)*	38.4 ± 0.80	40	6.4 ± 1.50	8	48.4 ± 3.20	50	31.1	32.7
DREMA *(Imagined data)*	41.2 ± 2.40	46	9.6 ± 1.50	12	54.4 ± 2.15	62	35.1	40.0
DREMA *(All data)*	**51.2** ± 1.60	**54**	**11.2** ± 1.60	**12**	**62.0** ± 2.19	**66**	**41.5**	**44.0**
	Multi-task							
	close jar		*place shape*		*slide block*		*avg multi-task*	
	mean ± std	max	mean ± std	max	mean ± std	max	mean	max
PerAct *(Original data)*	26.0 ± 3.10	28	7.2 ± 1.60	10	34.0 ± 5.06	38	22.4	25.3
DREMA *(Imagined data)*	28.0 ± 3.35	32	**18.0** ± 2.83	**22**	48.0 ± 1.79	50	31.3	34.7
DREMA *(All data)*	**46.0** ± 3.58	**52**	6.4 ± 3.20	12	**54.0** ± 2.19	**58**	**35.5**	**40.7**



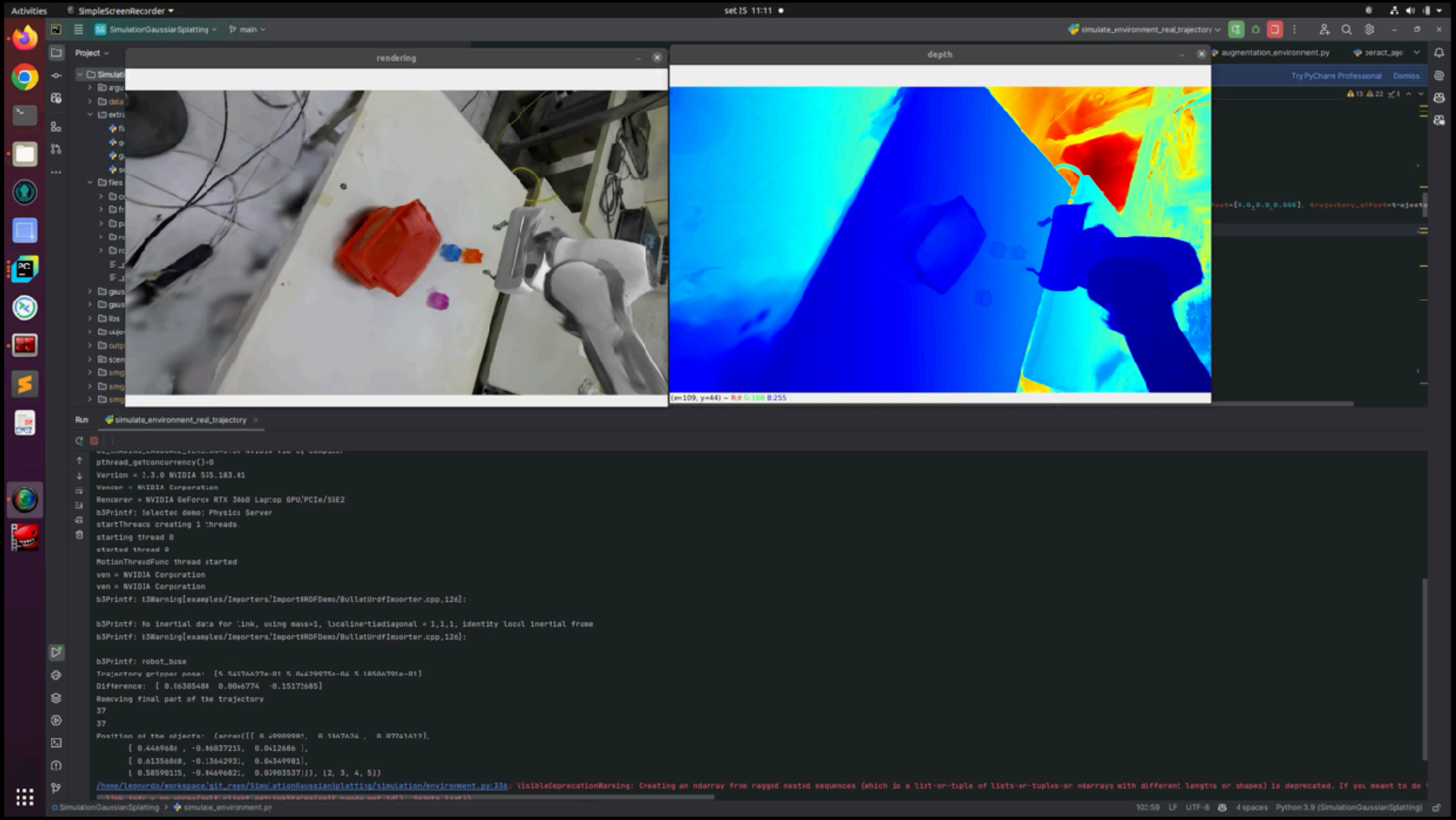
	close jar	place cups	place shape	put groceries	slide blocks	stack block	AVERAGE
RVT2 original four cameras	62	0.4	2.4	8.8	41.2	3.2	19.67
RVT2 original three cameras (val results)	17.5	2.5	7.5	5	50	2.5	14.17
RVT2 drema three cameras (val results)	25	0	15	10	50	10	18.33
PERACT original	26	1.6	7.2	1.2	34	3.6	12.27
PERACT drema	46	0.4	6.4	1.6	54	0.4	18.13







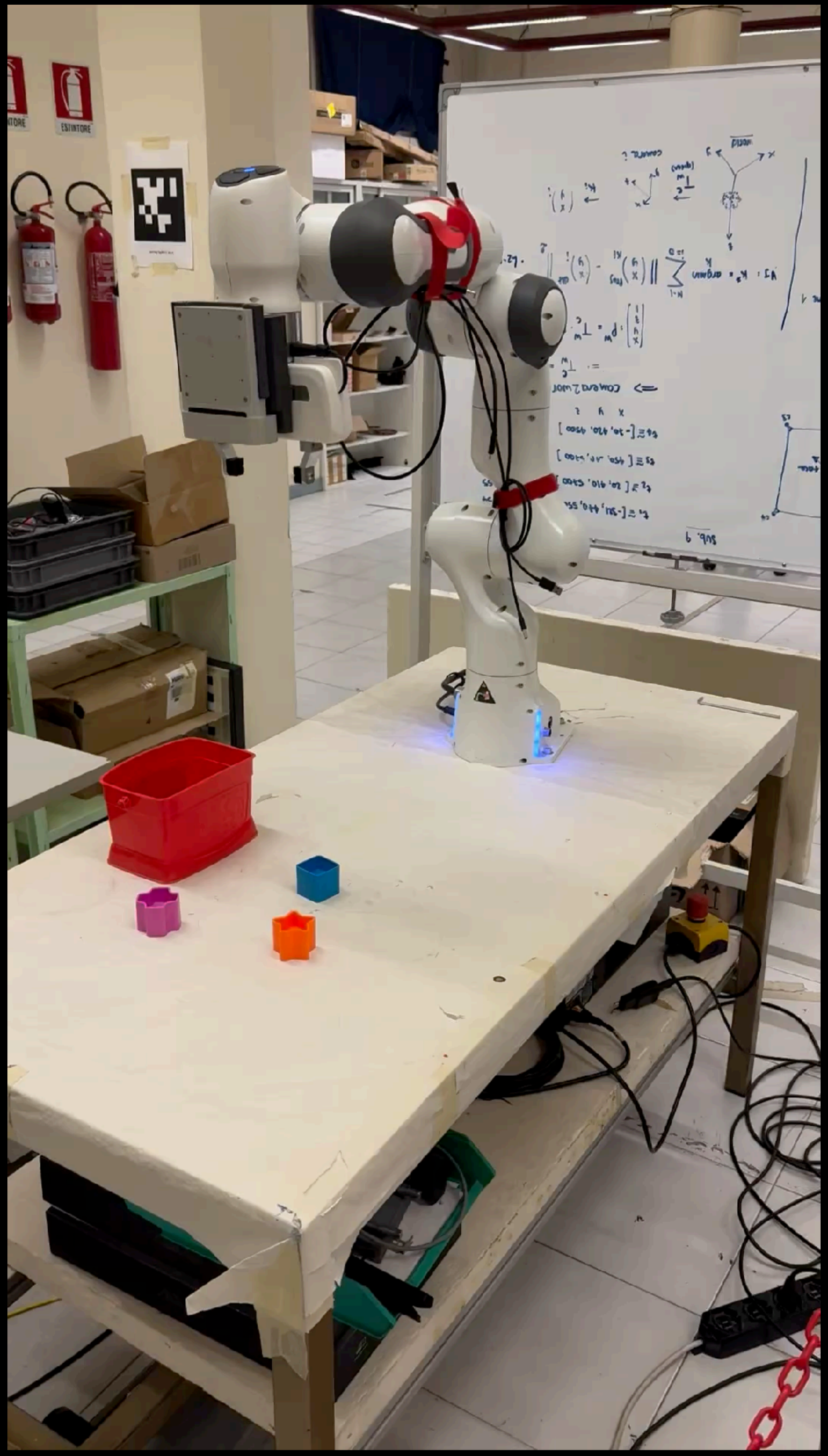
… THAT WORKS EVEN WITH REAL ROBOTS


rendering
depth
simulate_environment_real_trajectory
SimulationGaussianSplatting
main
Run
(x=109, y=44) ~ R:0 G:168 B:255

```
pthread_getconcurrency()=0
Version = 3.3.0 NVIDIA 535.183.01
Vendor = NVIDIA Corporation
Renderer = NVIDIA GeForce RTX 3060 Laptop GPU/PCIe/SSE2
b3Printf: Selected demo: Physics Server
startThreads creating 1 threads.
starting thread 0
started thread 0
MotionThreadFunc thread started
ven = NVIDIA Corporation
ven = NVIDIA Corporation
b3Printf: b3Warning[examples/Importers/ImportURDFDemo/BulletUrdfImporter.cpp,126]:

b3Printf: No inertial data for link, using mass=1, localinertiadiagonal = 1,1,1, identity local inertial frame
b3Printf: b3Warning[examples/Importers/ImportURDFDemo/BulletUrdfImporter.cpp,126]:

b3Printf: robot_base
Trajectory gripper pose:  [5.54576627e-01 5.84429975e-04 5.18586791e-01]
Difference:  [ 0.06305488  0.0066774  -0.15172685]
Removing final part of the trajectory
37
37
Position of the objects:  (array([[ 0.49999982,  0.1367626 ,  0.07741612],
       [ 0.4469686 , -0.06037215,  0.0412686 ],
       [ 0.61356068, -0.13642931,  0.04349981],
       [ 0.58590115, -0.04696821,  0.03903537]]), [2, 3, 4, 5])
/home/leonardo/workspace/git_repo/SimulationGaussianSplatting/simulation/environment.py:336: VisibleDeprecationWarning: Creating an ndarray from ragged nested sequences (which is a list-or-tuple of lists-or-tuples-or ndarrays with different lengths or shapes) is deprecated. If you meant to do
```



… THAT WORKS EVEN WITH REAL ROBOTS

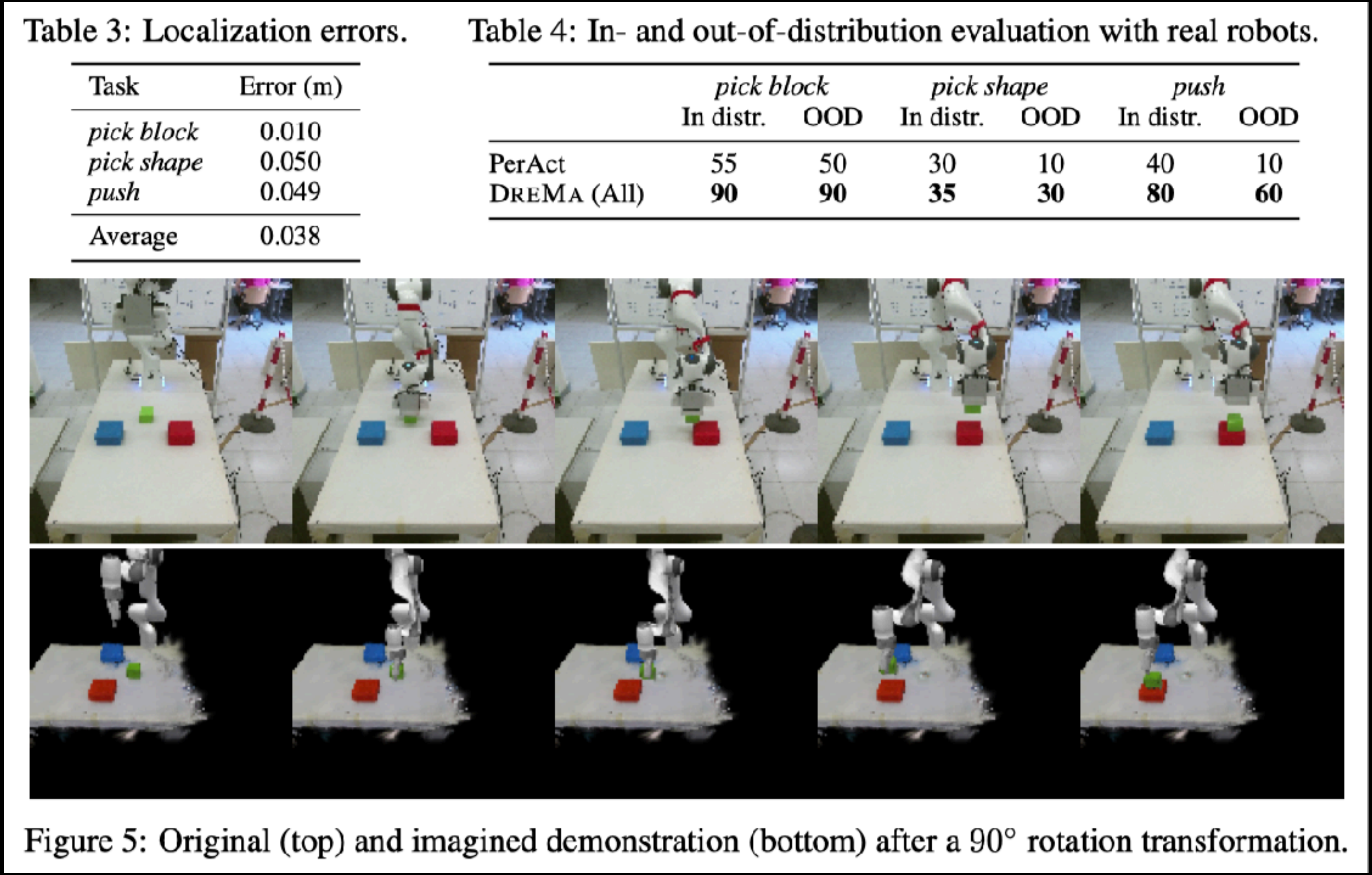


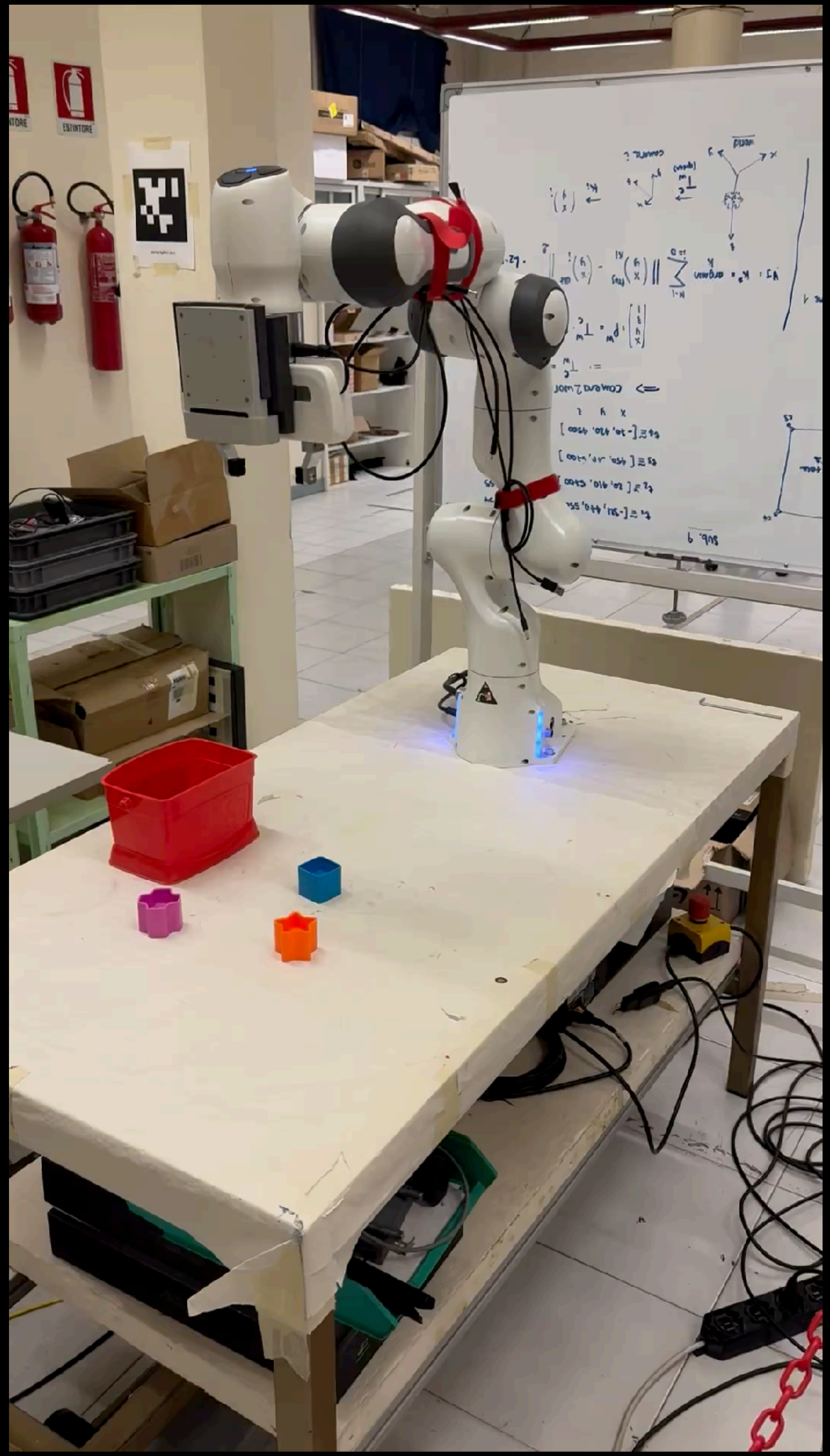
Table 3: Localization errors.

Task	Error (m)
pick block	0.010
pick shape	0.050
push	0.049
Average	0.038

Table 4: In- and out-of-distribution evaluation with real robots.

	pick block		*pick shape*		*push*	
	In distr.	OOD	In distr.	OOD	In distr.	OOD
PerAct	55	50	30	10	40	10
DREMA (All)	**90**	**90**	**35**	**30**	**80**	**60**

Figure 5: Original (top) and imagined demonstration (bottom) after a 90° rotation transformation.



MORE IMAGINED AUGMENTATIONS HELP

- Simple transformations for now
- Main challenge: make sure demonstrations are valid
- Lower efficacy but better scalability

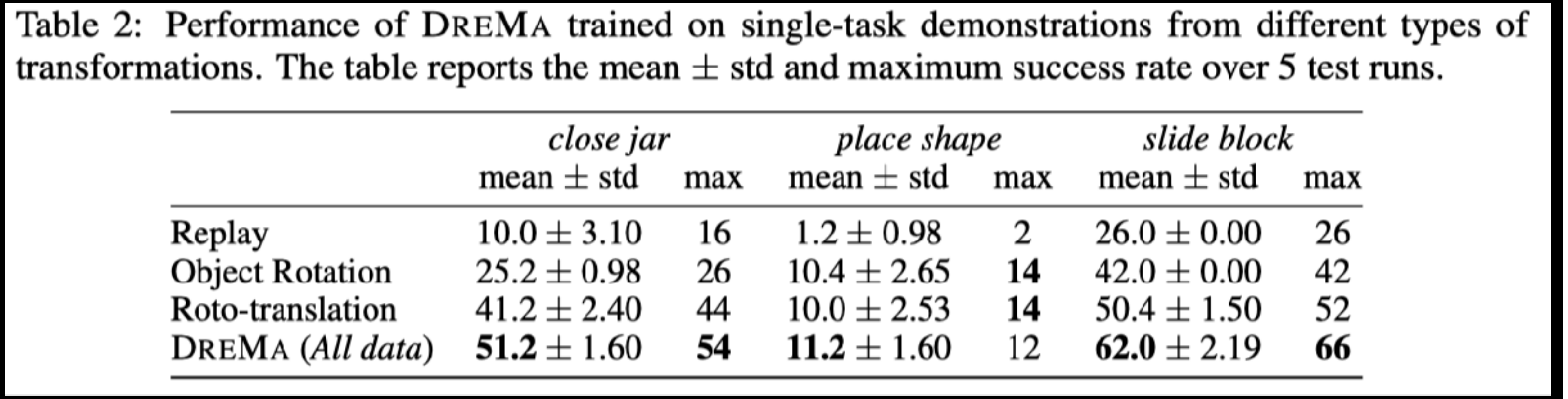


Table 2: Performance of DREMA trained on single-task demonstrations from different types of transformations. The table reports the mean ± std and maximum success rate over 5 test runs.

	close jar		*place shape*		*slide block*	
	mean ± std	max	mean ± std	max	mean ± std	max
Replay	10.0 ± 3.10	16	1.2 ± 0.98	2	26.0 ± 0.00	26
Object Rotation	25.2 ± 0.98	26	10.4 ± 2.65	**14**	42.0 ± 0.00	42
Roto-translation	41.2 ± 2.40	44	10.0 ± 2.53	**14**	50.4 ± 1.50	52
DREMA (*All data*)	**51.2** ± 1.60	**54**	**11.2** ± 1.60	12	**62.0** ± 2.19	**66**

Equivariant transformations are complementary …

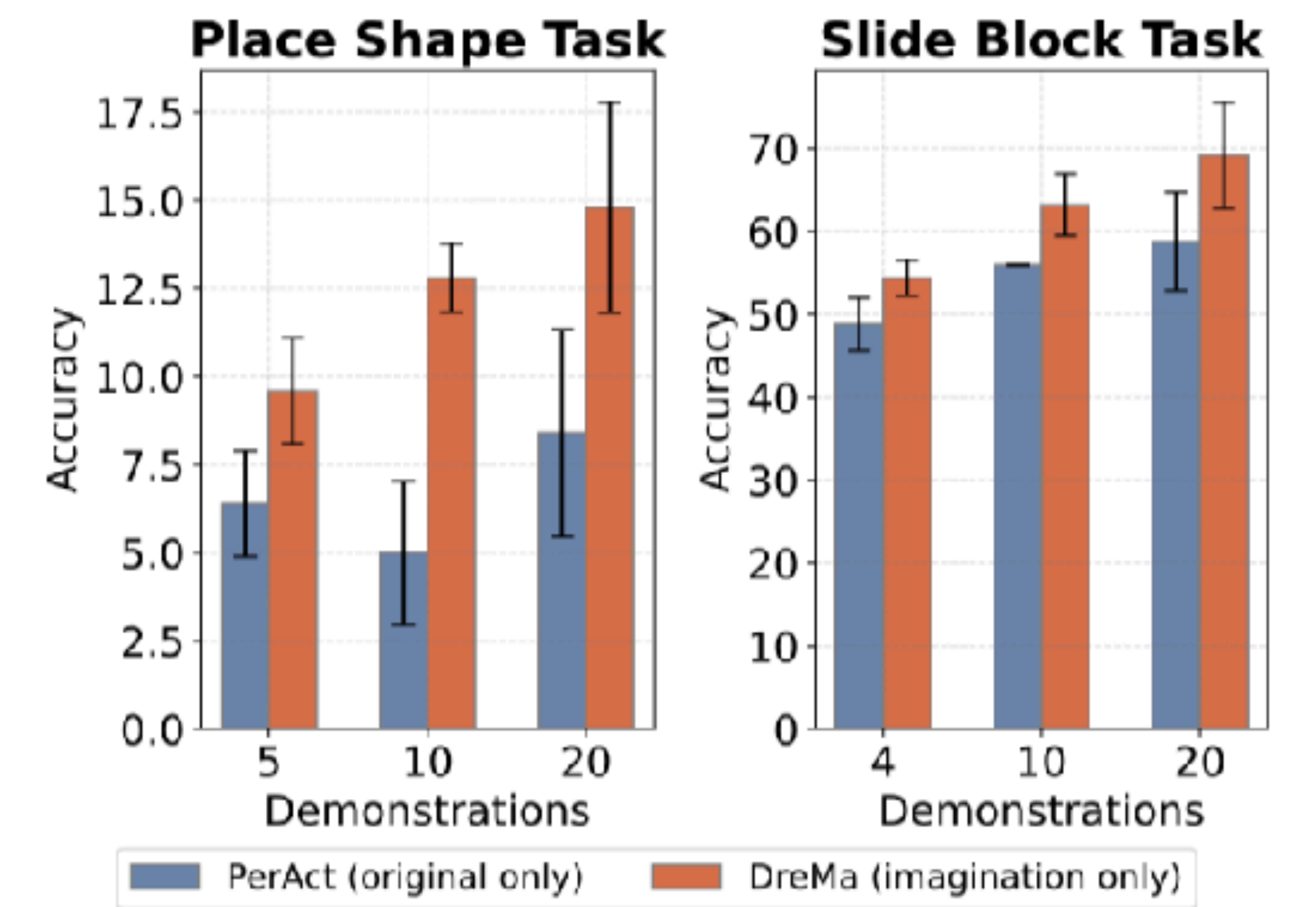
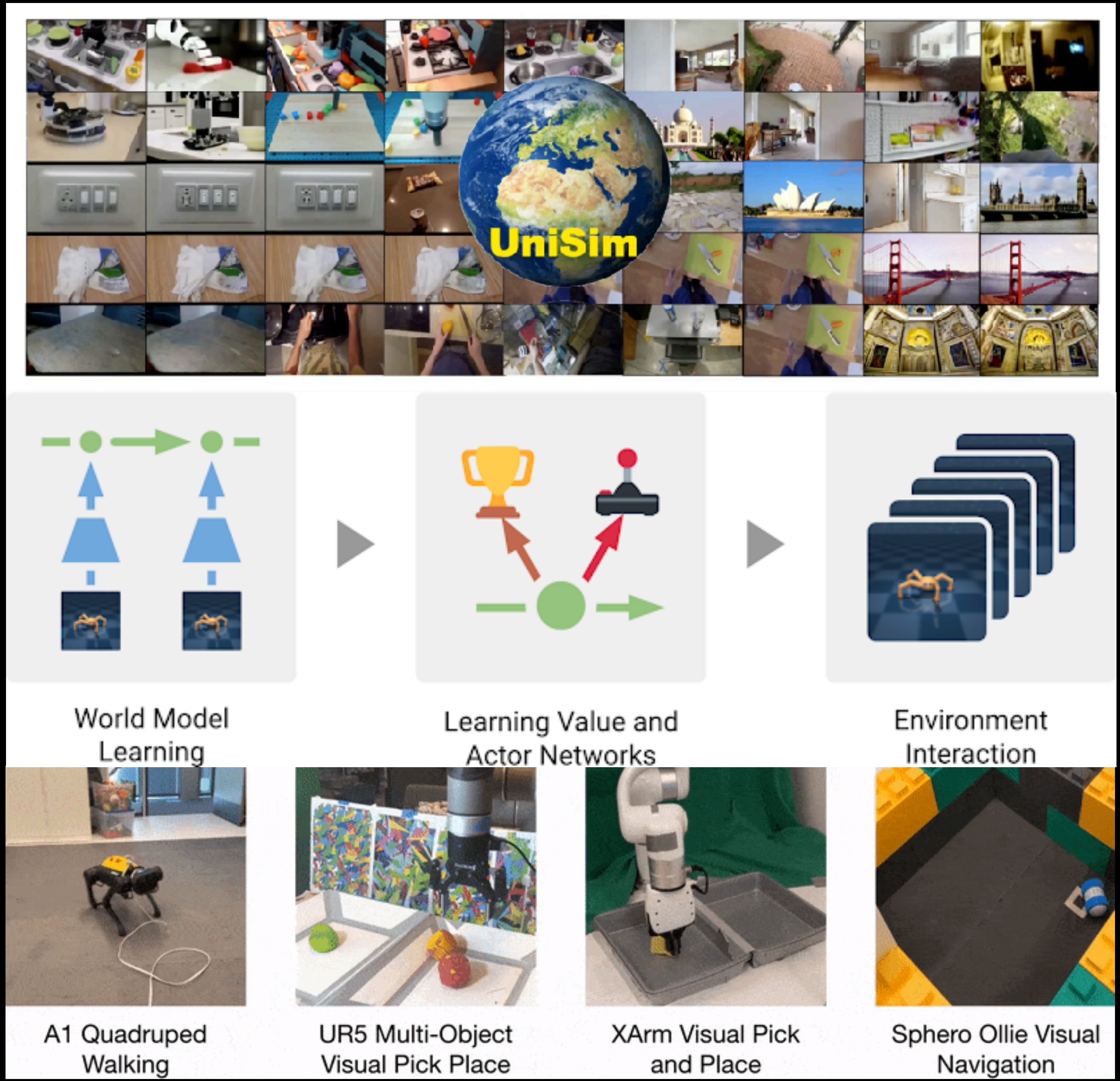


Figure 4: Imagined demonstrations keep improving imitation learning even with increasing number of original data.

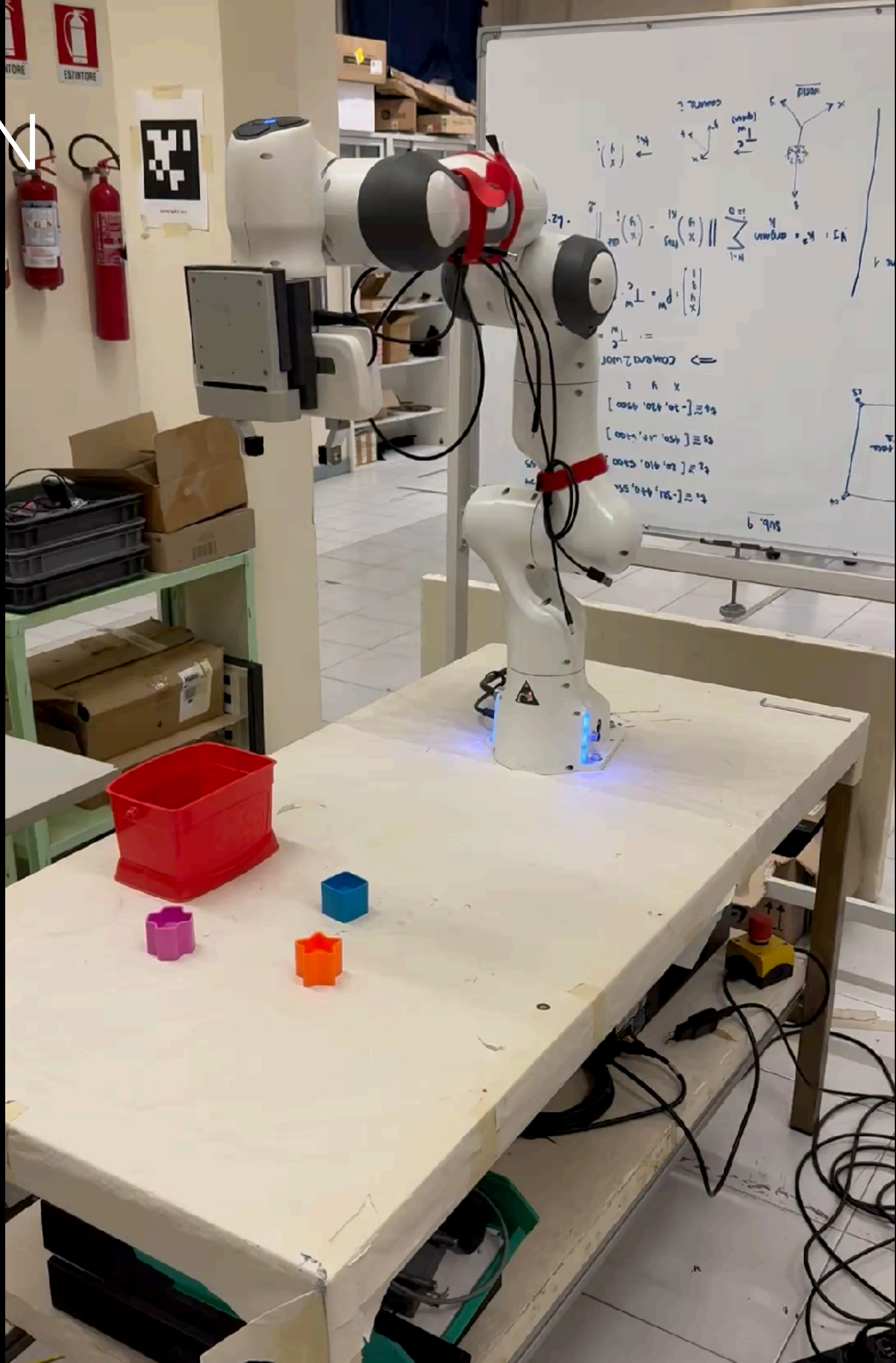
… and scale

WHAT MAKES IT NOT A WORLD MODEL?

- Not end-to-end yet
- Dynamics are not learned
- In fact, not much learning (from us) yet
- BUT, the idea is not to return to feature engineering
- Start with grounding explicitly
- Then everything else neural


UniSim
World Model Learning
Learning Value and Actor Networks
Environment Interaction
A1 Quadruped Walking
UR5 Multi-Object Visual Pick Place
XArm Visual Pick and Place
Sphero Ollie Visual Navigation

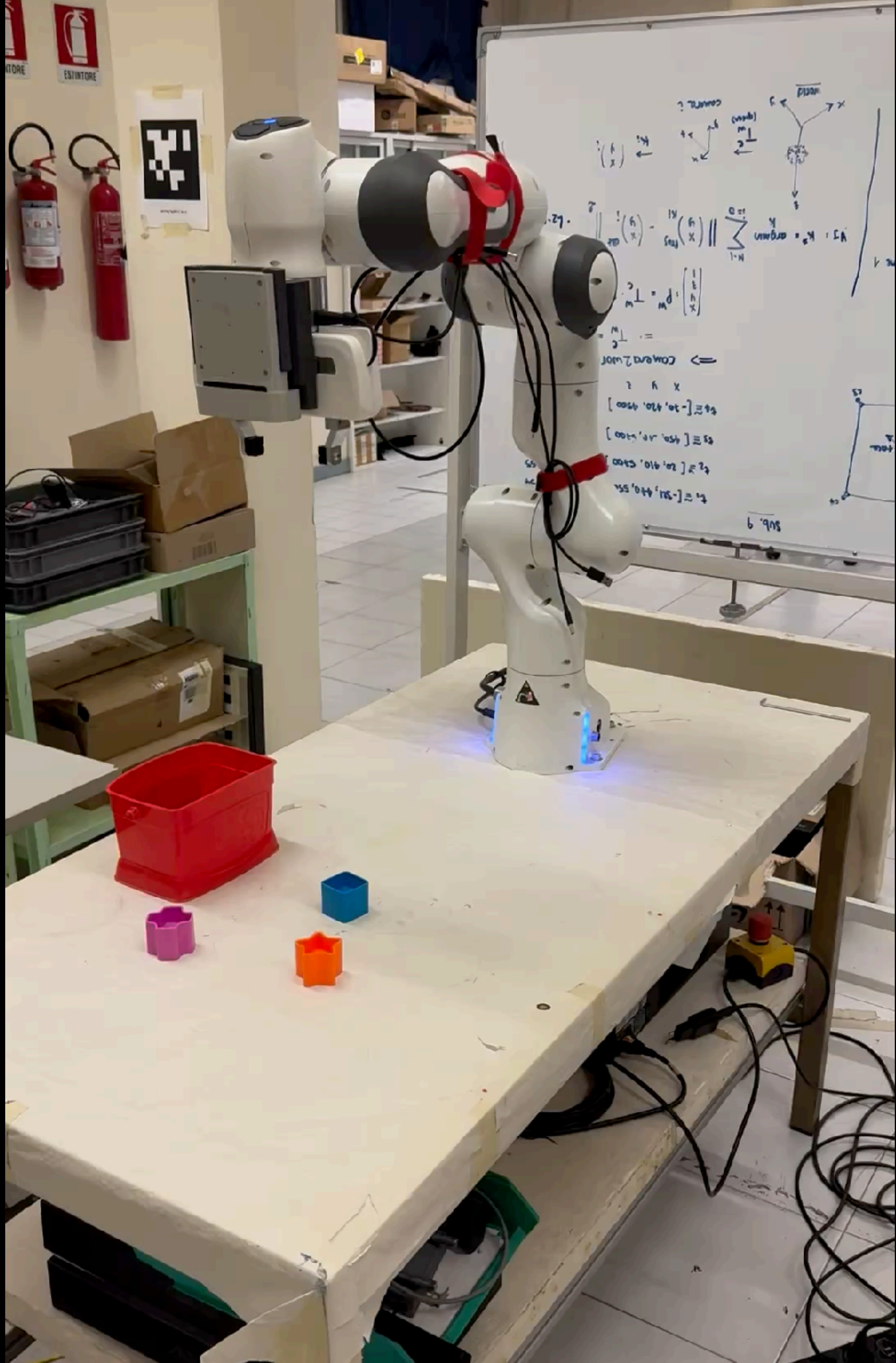
LEARNING & GENERALIZATION



- For one, add representations to object assets
- Learn (fine-tune) on observed trajectories
- Learn with differentiable physics
- Learn with partial observability
- Learn with manipulations
- Learn dynamics beyond physics (eg, causality, third-person actors, theory of mind, …)
- Learn with dynamic scenes

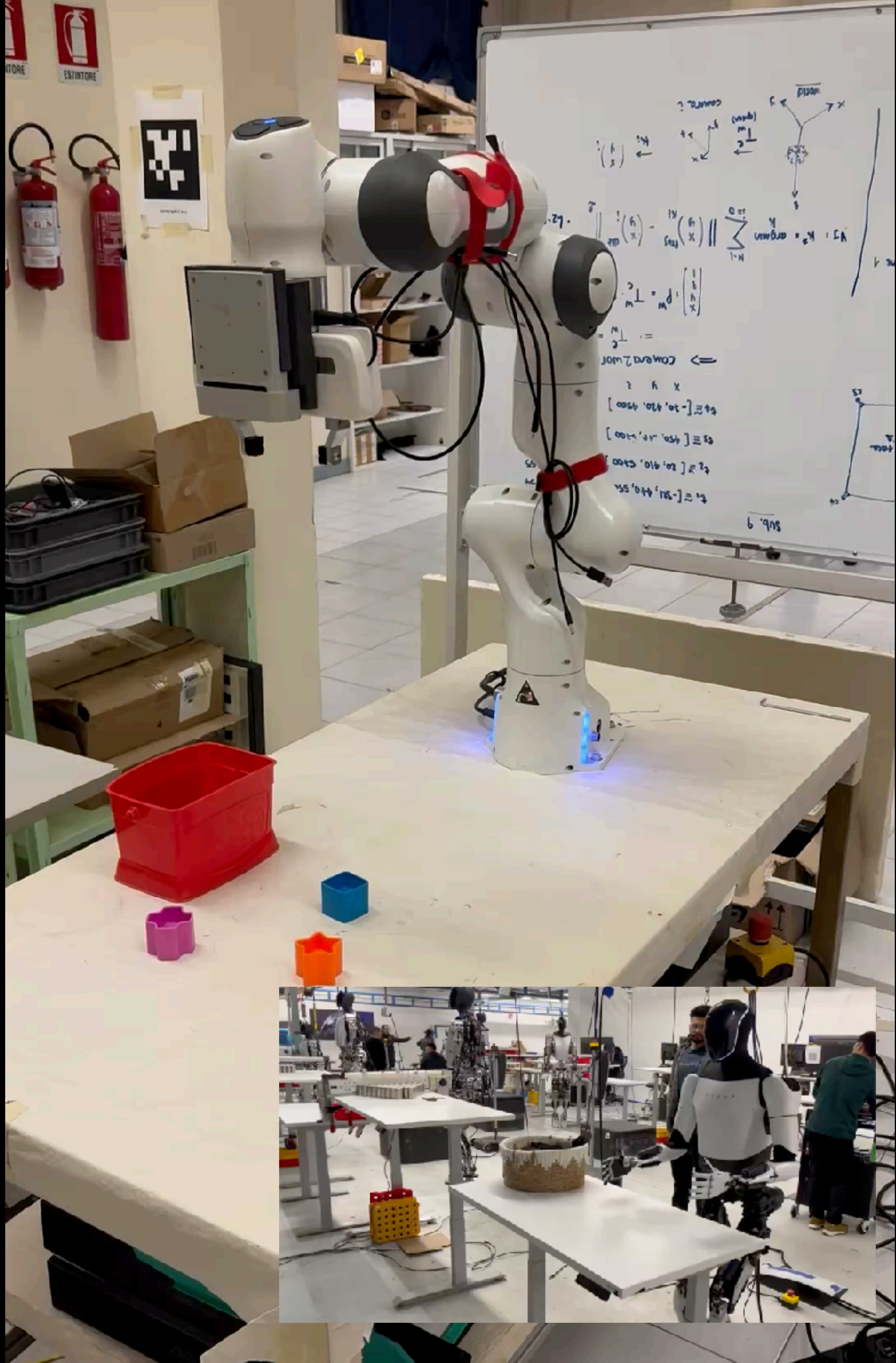
BEYOND IMITATION: LEARNING RL POLICIES

- Why stop with imitation learning?
- Grounded imagination & reasoning opens up lots of exciting possibilities
- Maximizing future reward in future imagination?
- …



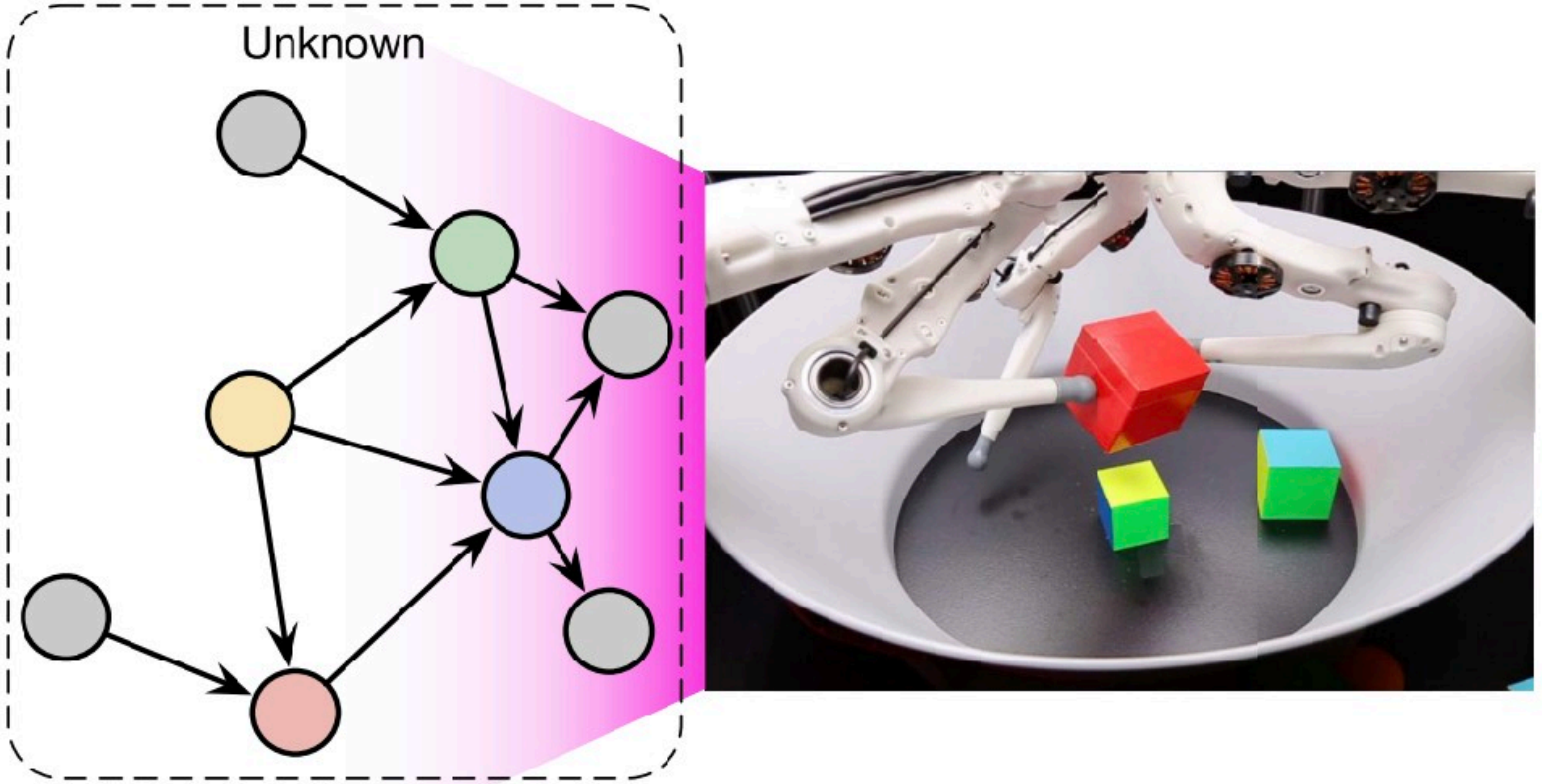
SKY IS THE LIMIT

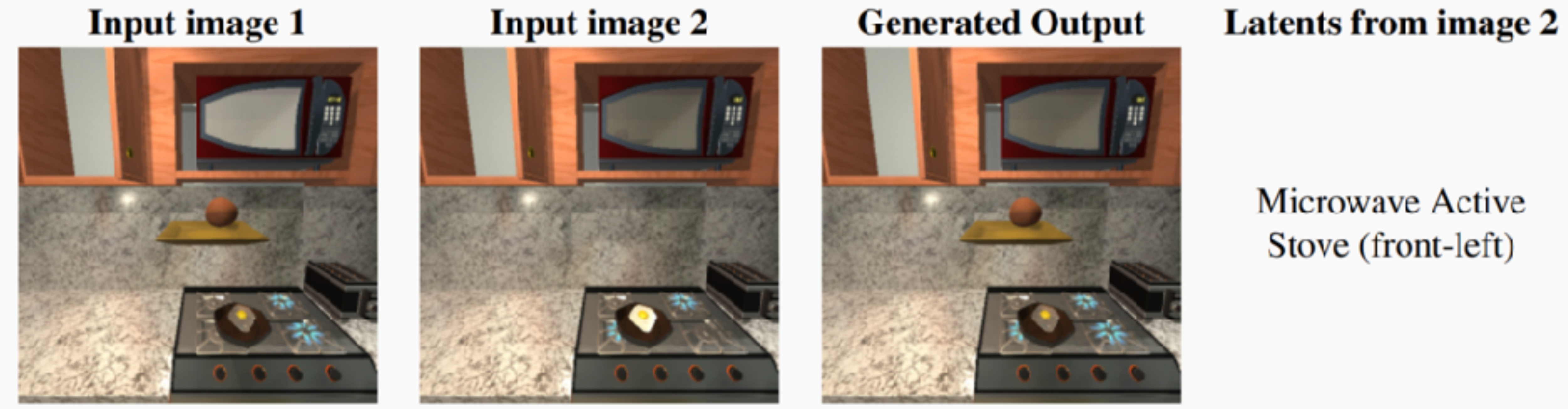
- Articulated objects (identifying joints)
- Deformable objects (with complex physics engines)
- Closed feedback loop (‘eye-hand coordination’)
- Physical parameter identification (friction, etc)
- Safety and interpretability
- Stochasticity (many possible futures)



LIPPE ET AL, ICLR, ICML, UAI, 2022-23

CAUSAL INDUCTIVE BIASES

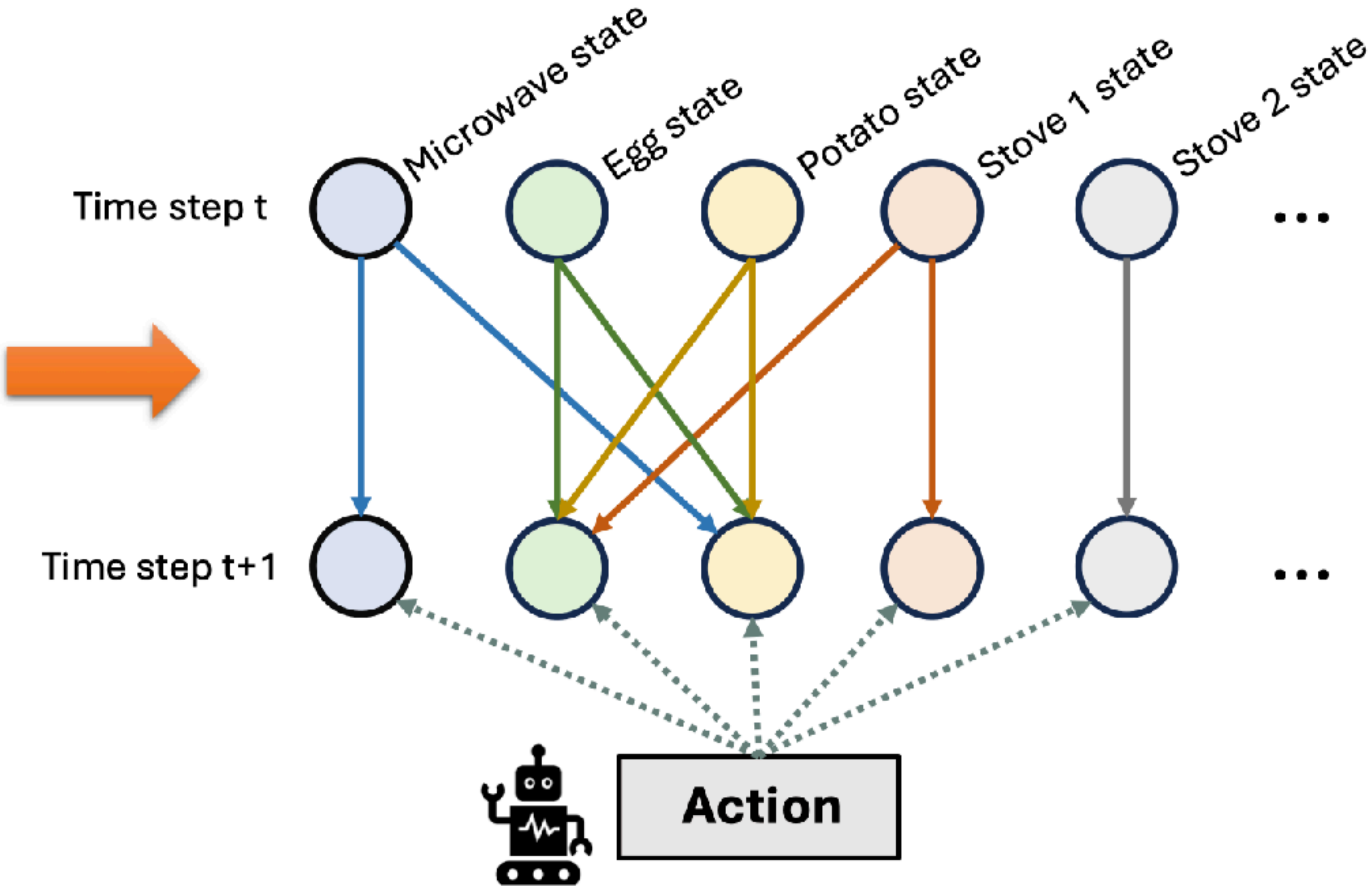




Manipulating Image 1 by turning on the Microwave and the front-left Stove. Note the egg staying uncooked despite the stove being turned on, which the model has never seen in training and shows BISCUIT's ability to perform novel interventions.

BISCUIT: LIPPE, MAGLIACANE, COHEN, GAVVES, ET AL, ICLR, ICML 2022-2024

TEMPORAL CAUSAL REPRESENTATION LEARNING

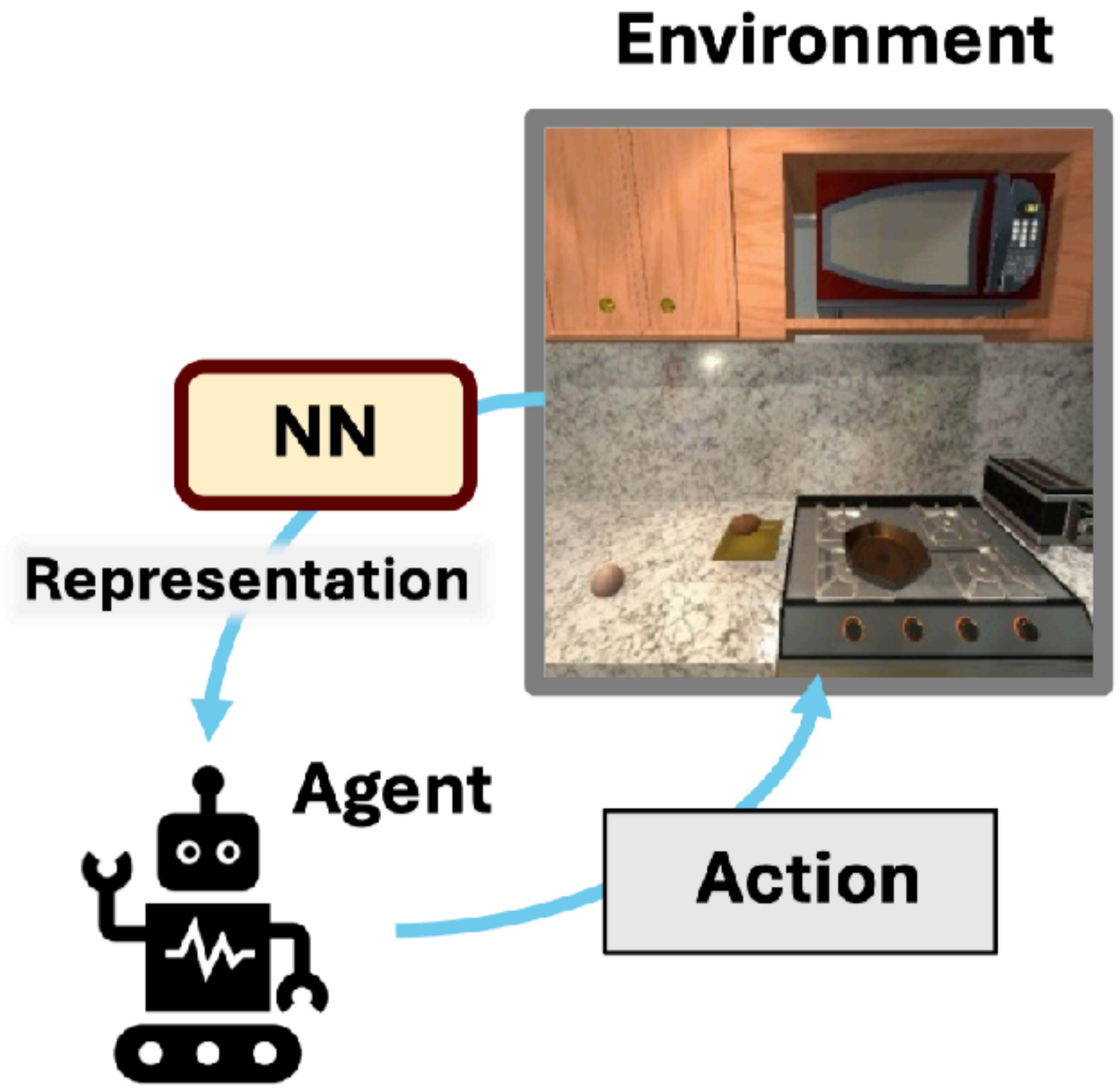


https://phlippe.github.io/

[1] B Schölkopf, F Locatello, S Bauer, N R Ke, N Kalchbrenner, A Goyal, Y Bengio, Towards Causal Representation Learning, Proceedings of the IEEE, 2021

BISCUIT: LIPPE, MAGLIACANE, COHEN, GAVVES, ET AL, ICLR, ICML 2022-2024

TEMPORAL CAUSAL REPRESENTATION LEARNING



Representation Learning Tasks

What are the causal variables of the environment?

How do they interact with each other?

How can the robot agent intervene on causal variables?

https://phlippe.github.io/

[1] B Schölkopf, F Locatello, S Bauer, N R Ke, N Kalchbrenner, A Goyal, Y Bengio, Towards Causal Representation Learning, Proceedings of the IEEE, 2021

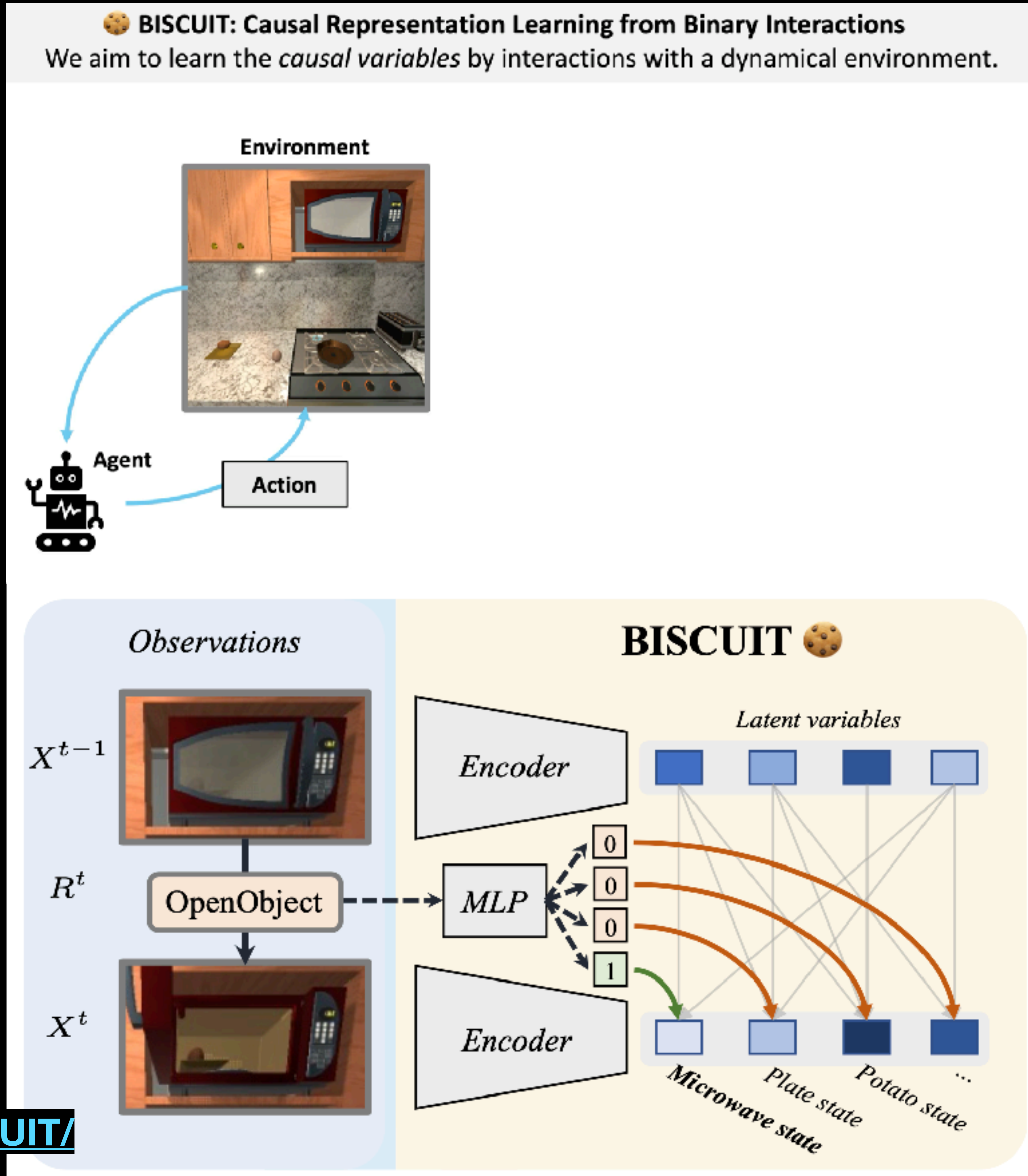
BISCUIT: LIPPE, MAGLIACANE, ASANO, LOWE, COHEN, GAVVES, UAI 23

CAUSAL LEARNING FROM BINARY INTERACTIONS

- Many interactions are binary

(Turn lights on/off, open/close door …)

- Learn latents to reflect change
- Provable & manipulable causal factors
- By integrating probabilistic causal priors



BISCUIT: Causal Representation Learning from Binary Interactions

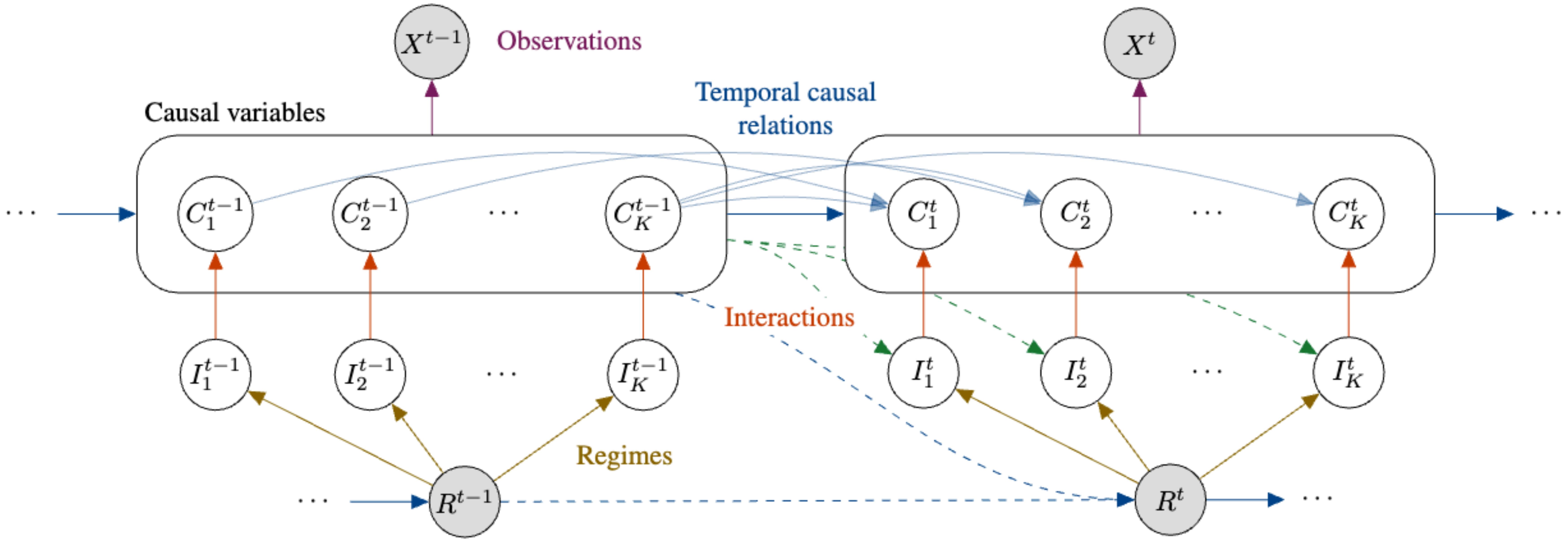
We aim to learn the *causal variables* by interactions with a dynamical environment.

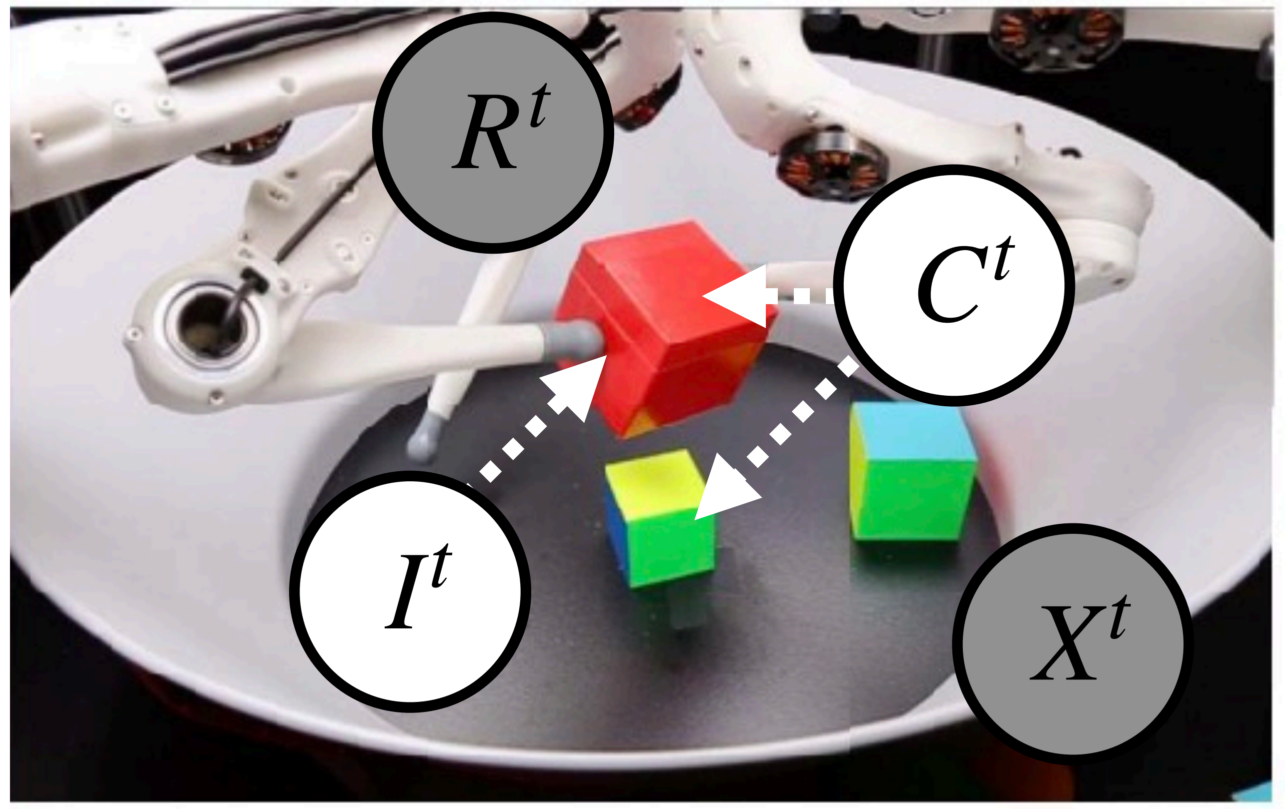
https://phlippe.github.io/BISCUIT/

BISCUIT: LIPPE, MAGLIACANE, ASANO, LOWE, COHEN, GAVVES, UAI 23

CAUSAL LEARNING FROM BINARY INTERACTIONS

- Causal model: a temporal DAG
- Observed: images X^t and "regime" variables R^t
- Latents: causal C^{t-1} and interactions $I^t = f(R^t, C^{t-1})$



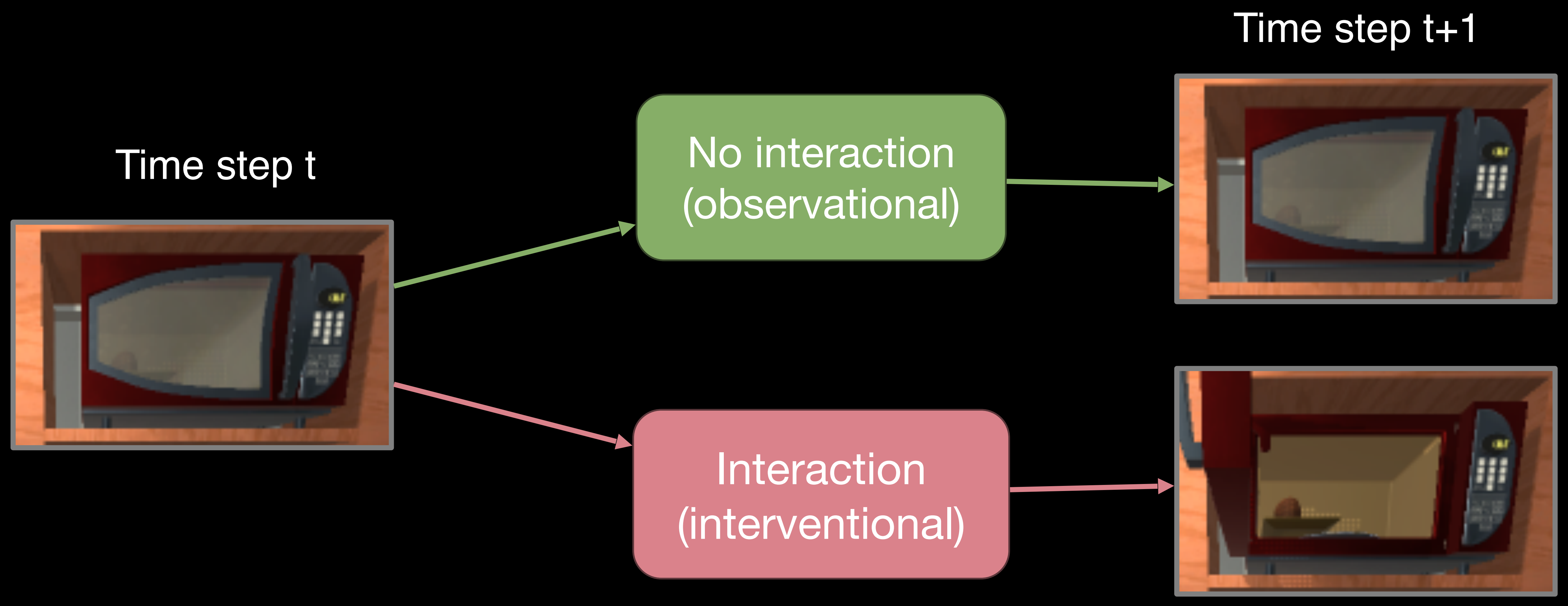


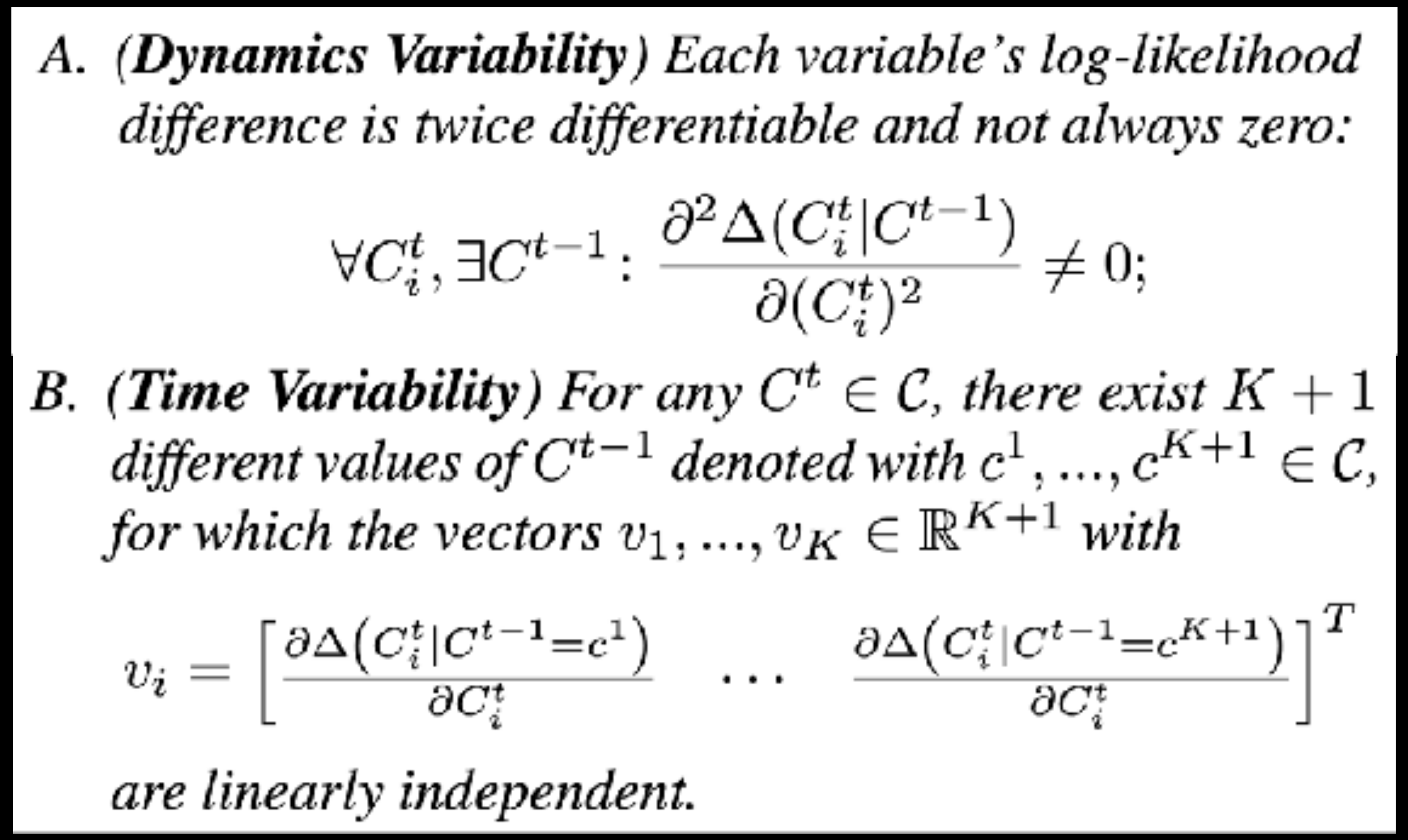
https://phlippe.github.io/BISCUIT/

UP TO PERMUTATION AND COMPONENT-WISE TRANSFORMATIONS

BINARY INTERACTIONS FOR IDENTIFIABILITY

- Assumption #1: interactions described by a binary variable
- Assumption #2: distinct interaction patterns → interactions not functions of other interactions
- Assumption #3: mechanisms vary sufficiently with interactions or time





A. (**Dynamics Variability**) *Each variable's log-likelihood difference is twice differentiable and not always zero:*

$$\forall C_i^t, \exists C^{t-1}: \frac{\partial^2 \Delta(C_i^t | C^{t-1})}{\partial (C_i^t)^2} \neq 0;$$

B. (**Time Variability**) *For any $C^t \in \mathcal{C}$, there exist $K+1$ different values of C^{t-1} denoted with $c^1, ..., c^{K+1} \in \mathcal{C}$, for which the vectors $v_1, ..., v_K \in \mathbb{R}^{K+1}$ with*

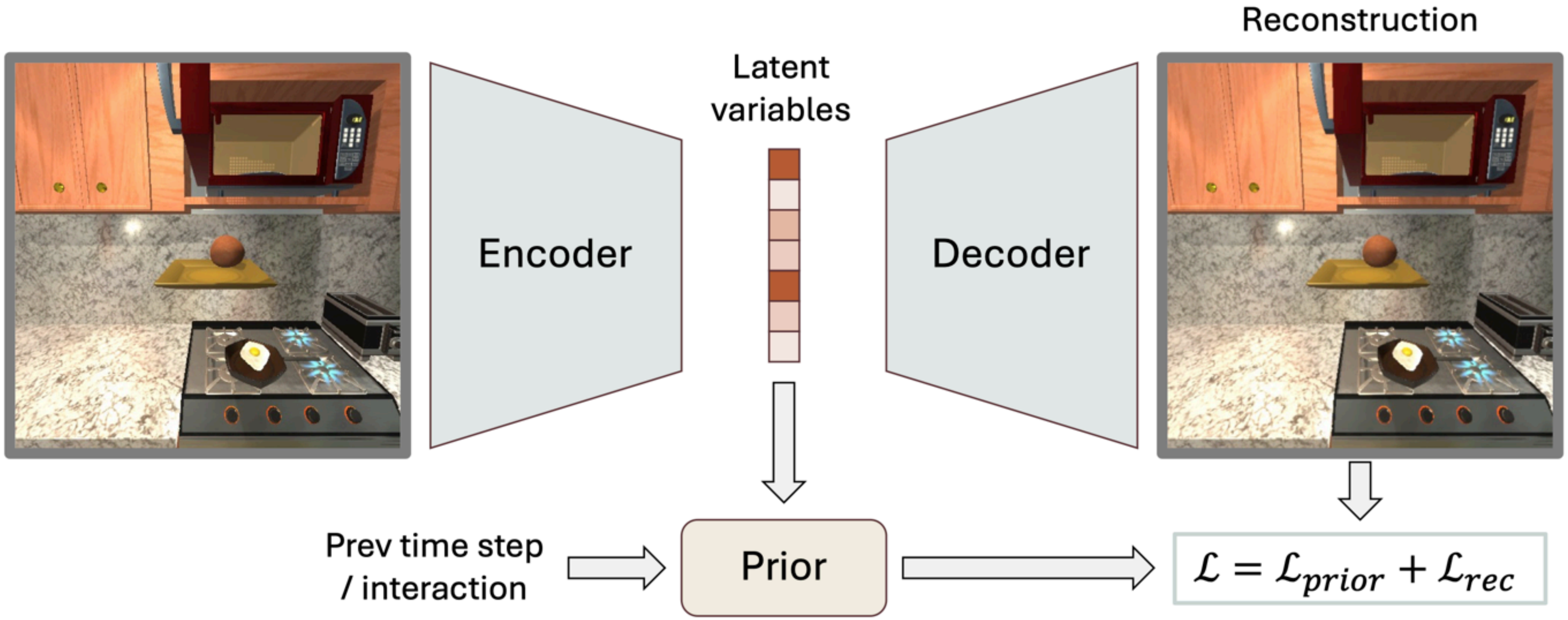
$$v_i = \left[\frac{\partial \Delta\left(C_i^t | C^{t-1}=c^1\right)}{\partial C_i^t} \quad \cdots \quad \frac{\partial \Delta\left(C_i^t | C^{t-1}=c^{K+1}\right)}{\partial C_i^t} \right]^T$$

are linearly independent.

https://phlippe.github.io/BISCUIT/

BISCUIT: LIPPE, MAGLIACANE, ASANO, LOWE, COHEN, GAVVES, UAI 23

OPTIMIZING WITH VARIATIONAL INFERENCE

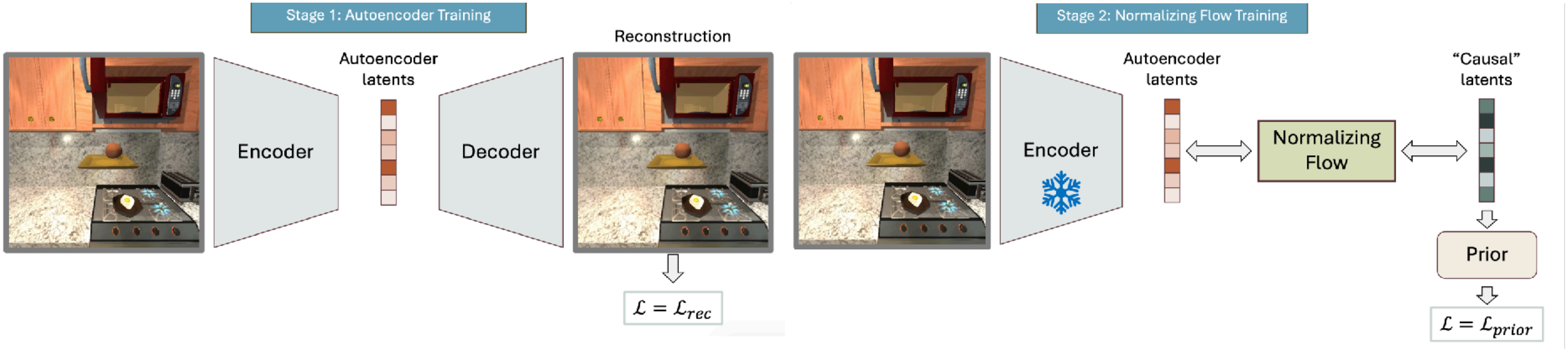


$$\mathcal{L} = \mathcal{L}_{prior} + \mathcal{L}_{rec}$$

https://phlippe.github.io/BISCUIT/

BISCUIT: LIPPE, MAGLIACANE, ASANO, LOWE, COHEN, GAVVES, UAI 23

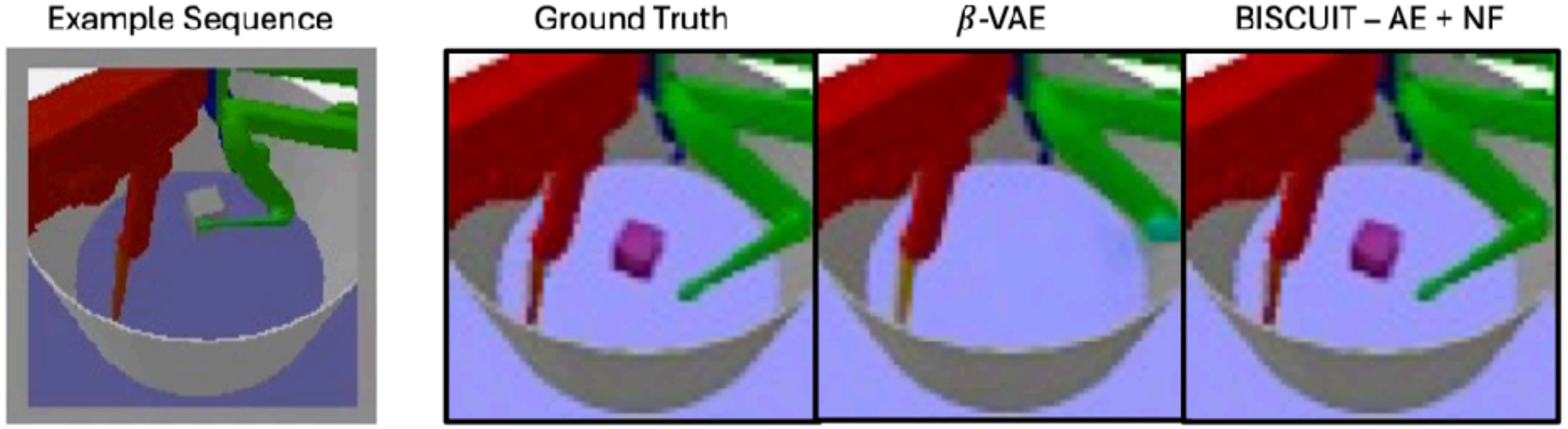
TWO-STAGE LEARNING IN COMPLEX SETTINGS


Stage 1: Autoencoder Training
Reconstruction
Autoencoder latents
Encoder
Decoder
$\mathcal{L} = \mathcal{L}_{rec}$
Stage 2: Normalizing Flow Training
Autoencoder latents
"Causal" latents
Encoder
Normalizing Flow
Prior
$\mathcal{L} = \mathcal{L}_{prior}$

https://phlippe.github.io/BISCUIT/

CAUSALWORLD TRI-FINGER

DISCOVERING INTERACTIONS



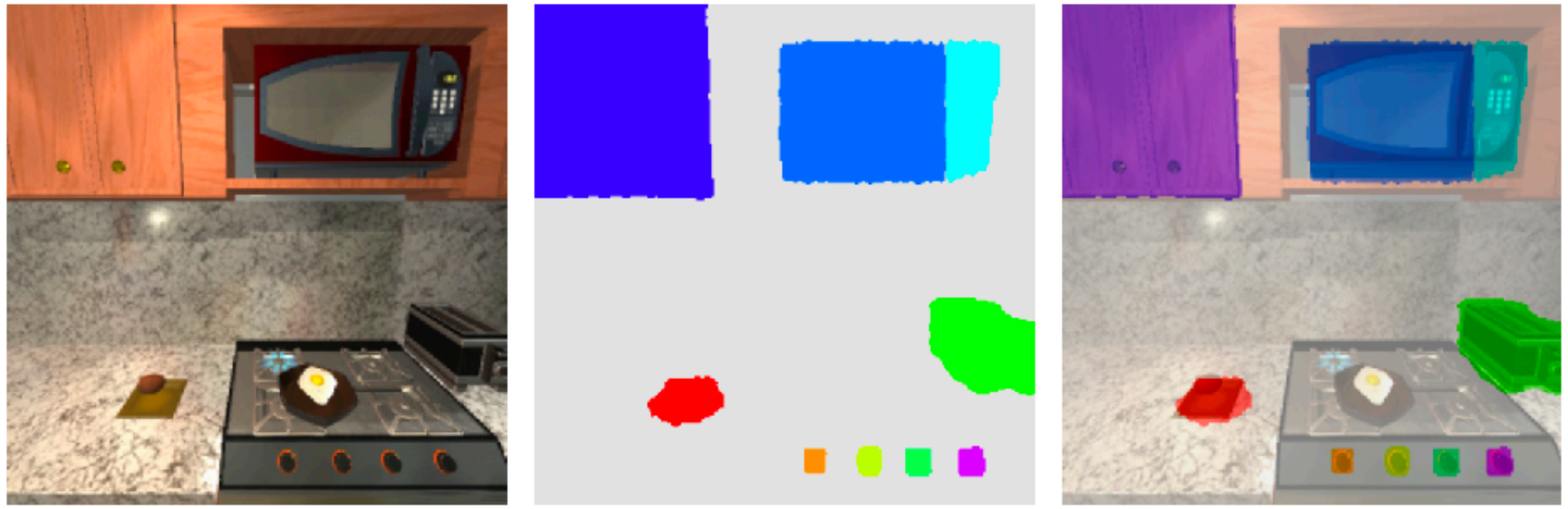
R^2 scores (diag ↑/sep ↓)

Models	**CausalWorld**
iVAE (Khemakhem et al., 2020a)	0.28 / 0.00
LEAP (Yao et al., 2022b)	0.30 / 0.00
DMS (Lachapelle et al., 2022b)	0.32 / 0.00
BISCUIT-NF (Ours)	**0.97** / 0.01

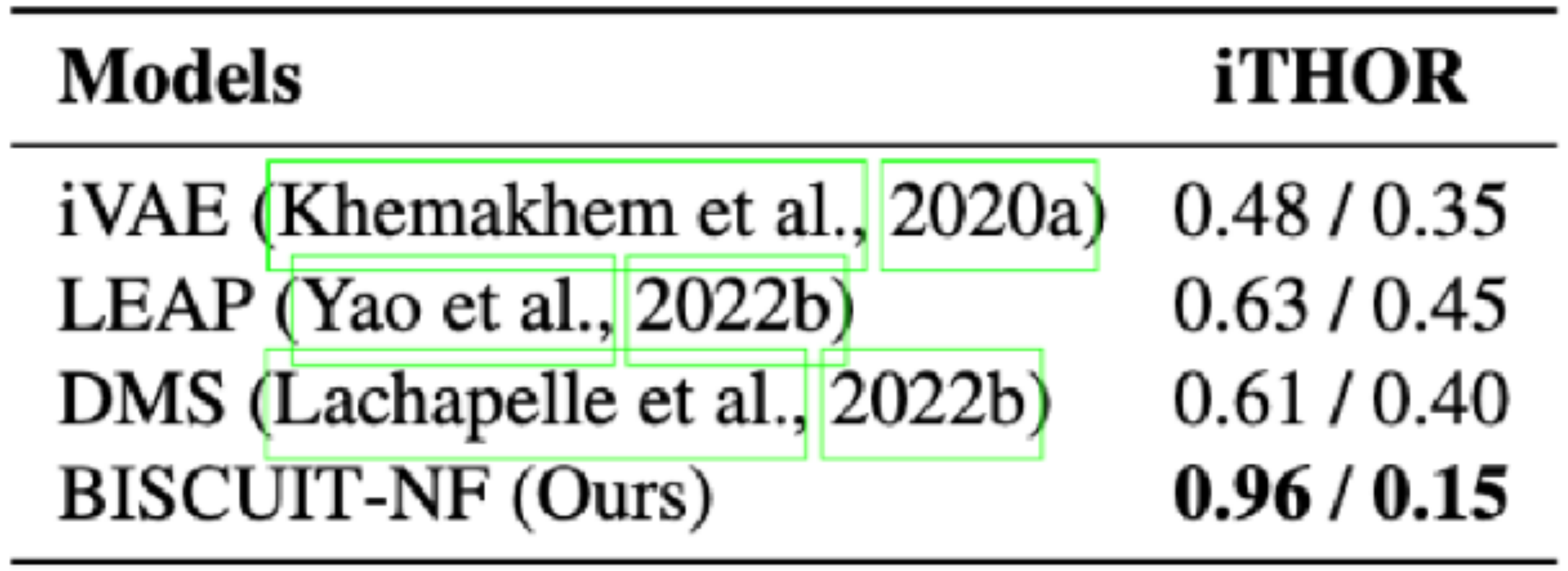
https://phlippe.github.io/BISCUIT/

ITHOR OBJECTS AS CAUSAL VARIABLES

CAUSE-AND-EFFECT DRIVES SEMANTICS (?)


Input Image Learned Interactions Combined Image

R^2 scores (diag ↑/sep ↓)

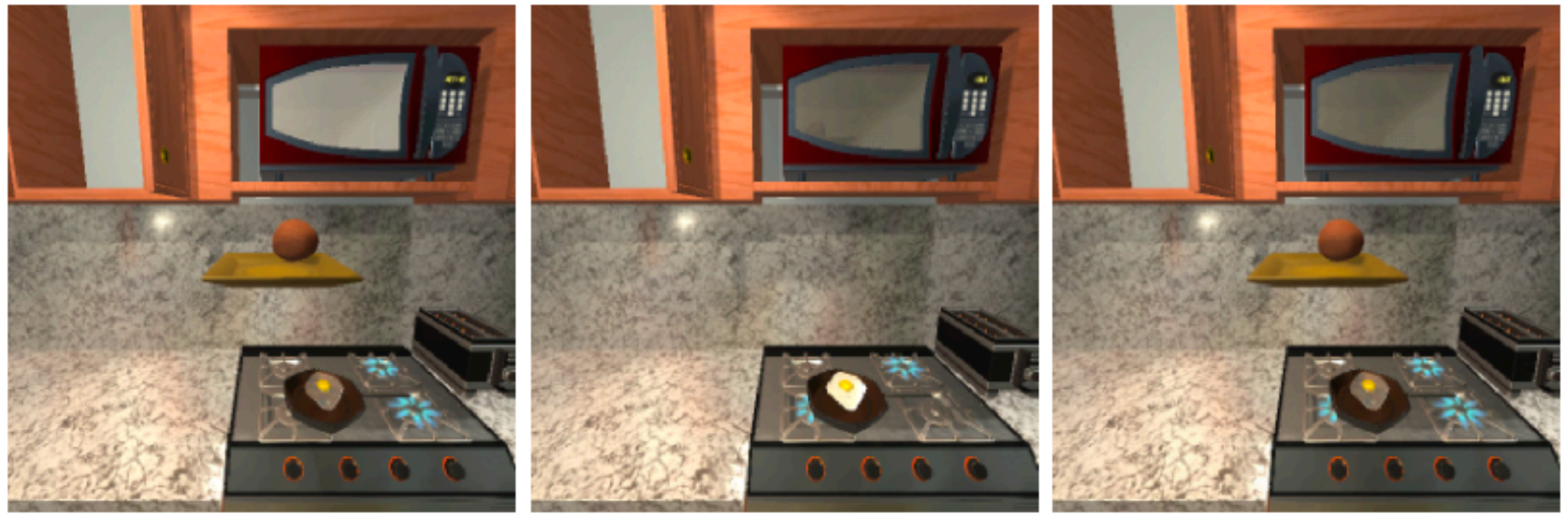


Models	iTHOR
iVAE (Khemakhem et al., 2020a)	0.48 / 0.35
LEAP (Yao et al., 2022b)	0.63 / 0.45
DMS (Lachapelle et al., 2022b)	0.61 / 0.40
BISCUIT-NF (Ours)	**0.96 / 0.15**

https://phlippe.github.io/BISCUIT/

ELEMENTS OF "IMAGINATION"

HACKING THE SIMULATOR


Input Image 1 Input Image 2 Generated Output

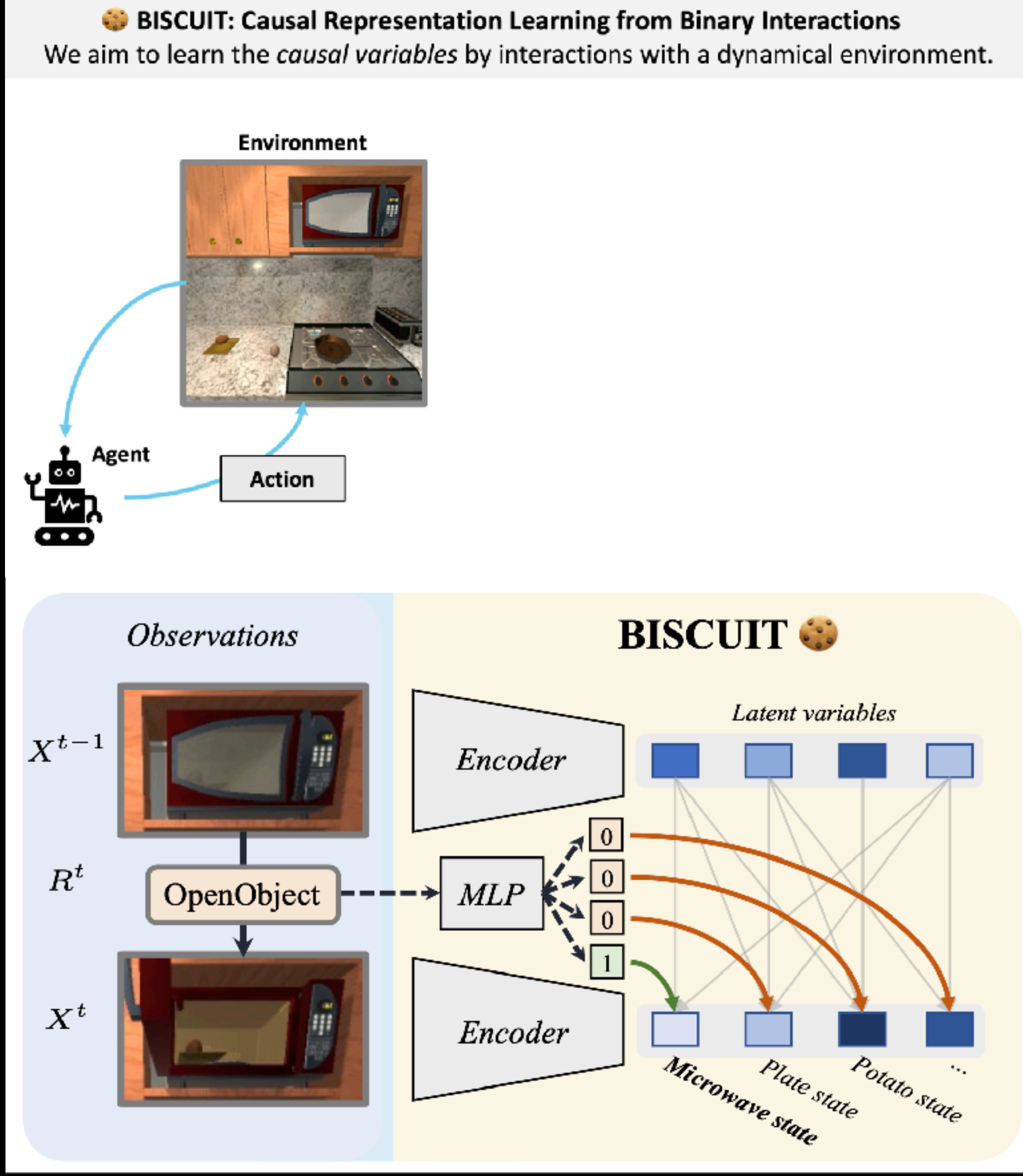
Scaling-up with foundation vision models towards autonomous learning?

https://phlippe.github.io/BISCUIT/

BISCUIT: LIPPE, MAGLIACANE, ASANO, LOWE, COHEN, GAVVES, UAI 23

LIMITATIONS

- Sufficient intervention data
- Works with temporal data only
- Assumes binary interactions

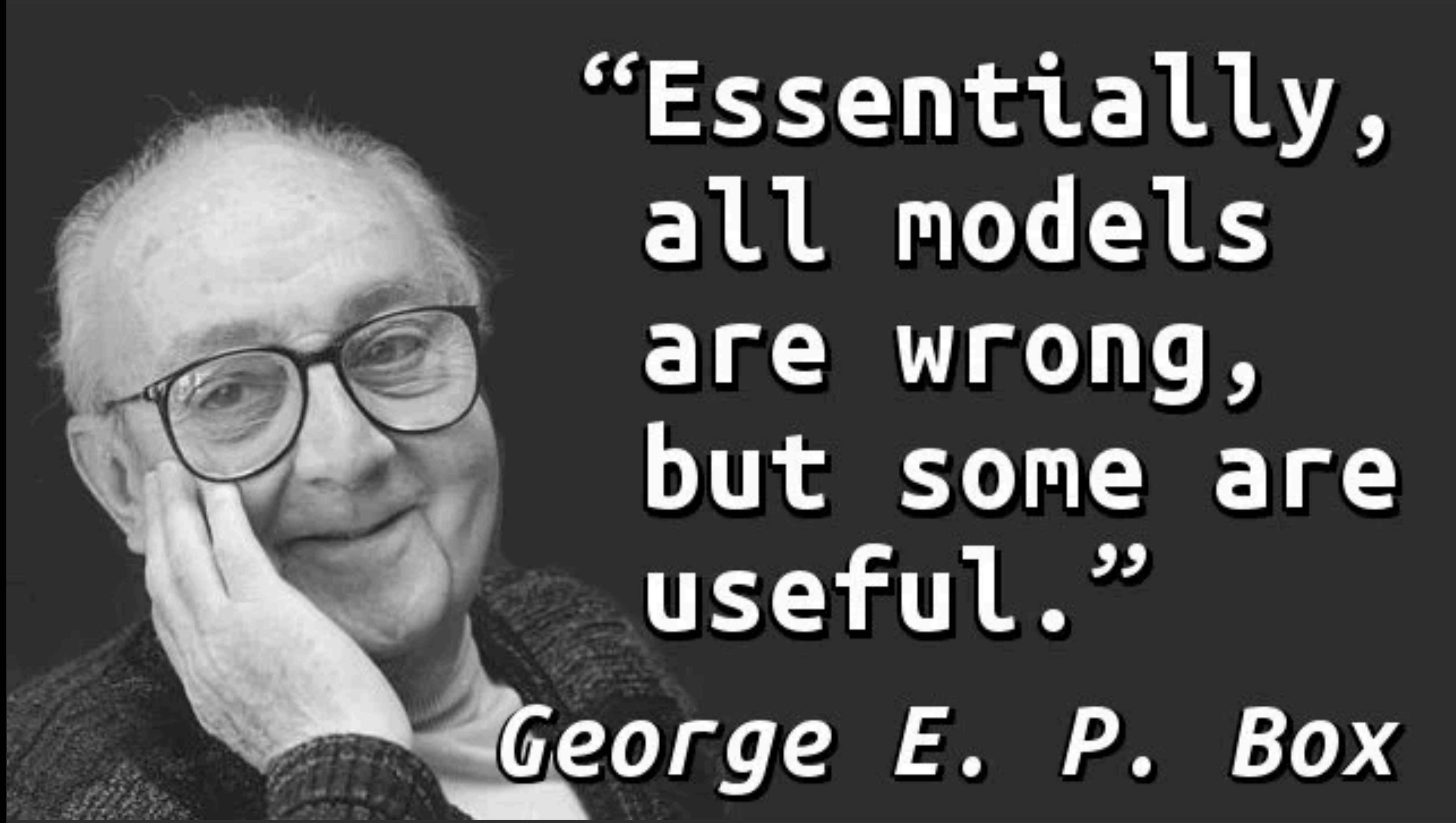

BISCUIT: Causal Representation Learning from Binary Interactions

We aim to learn the *causal variables* by interactions with a dynamical environment.

https://phlippe.github.io/BISCUIT/

IS IT REALLY CAUSAL?

- Patterns are often correlations
- Cause-and-effect is a strong (albeit sometimes biased) framework to learn
- If we go past the chicken-egg problem
- Power of causal representations is in autonomousness and controllability (imho)



"Essentially, all models are wrong, but some are useful."

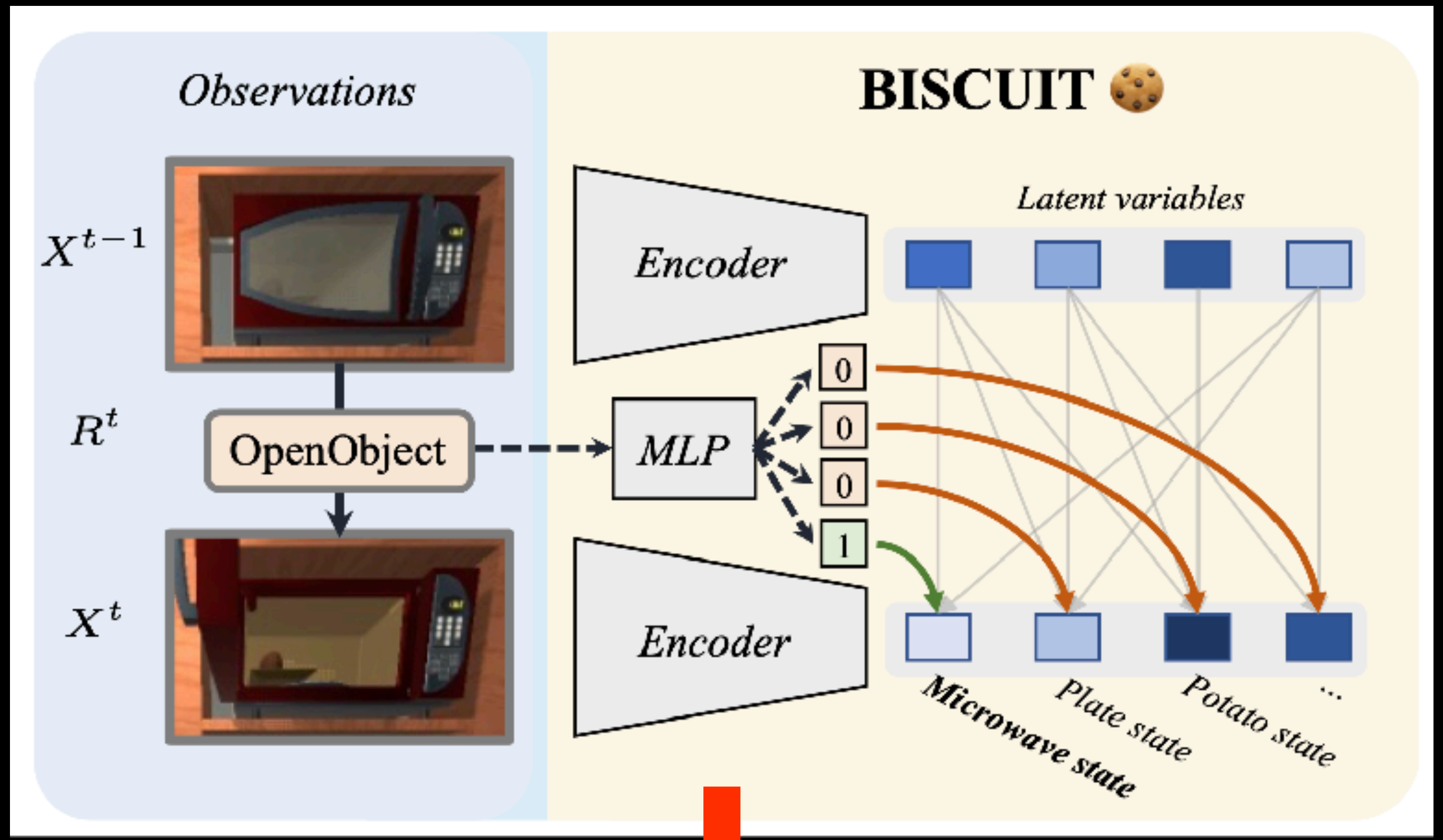
George E. P. Box

https://phlippe.github.io/BISCUIT/

BISCUIT++

SCALE & ROBOT LEARNING

- Scale up to many environments → reuse semantics
- LLMs for guidance and sample efficiency?
- Ideally, Gaussian Splats for *de novo interactive* environments and scaling-up
- System IIa: First causal principles for novel problem-solving & Causal World Models
- System IIb: Safe & human-robot-aligned planning

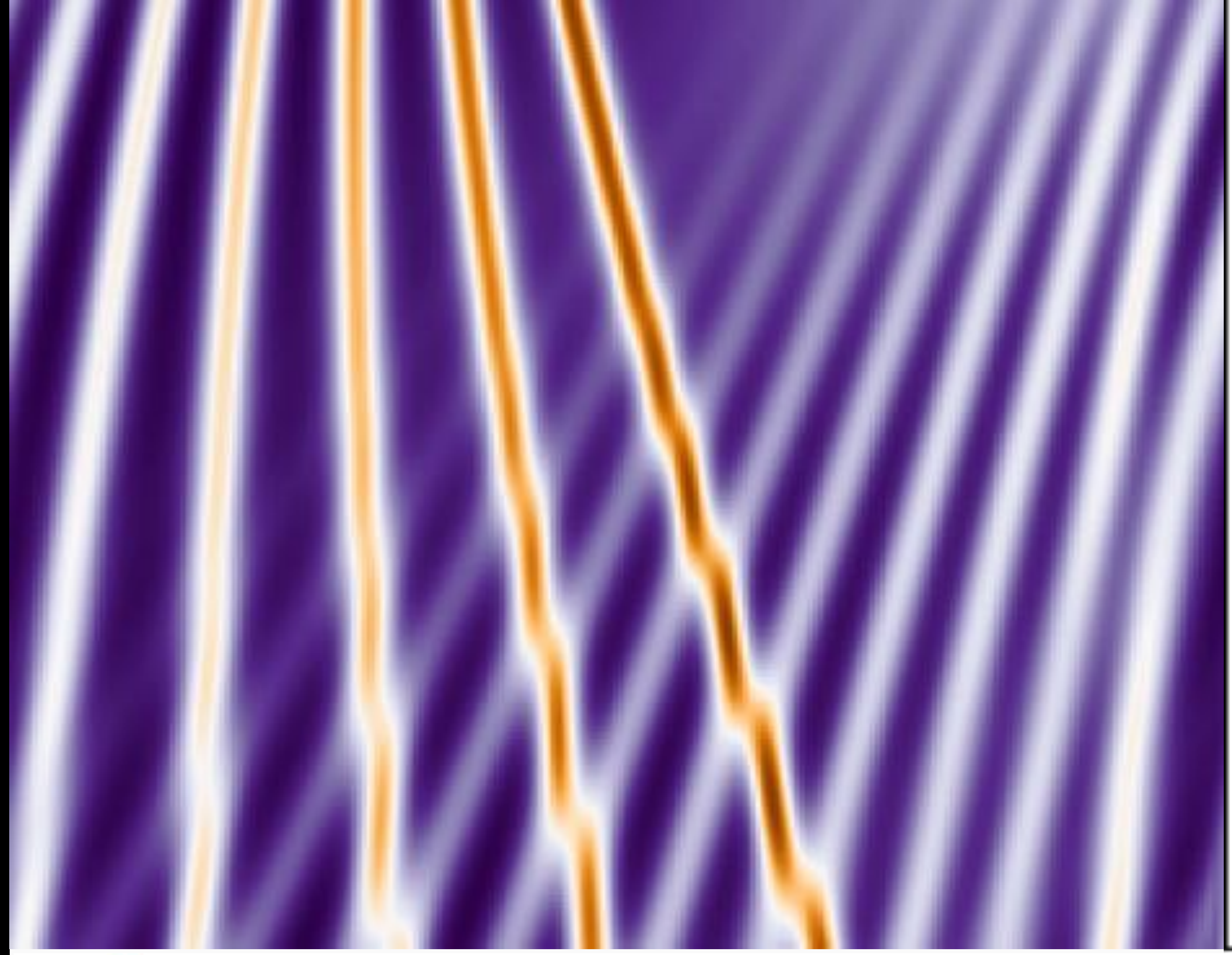


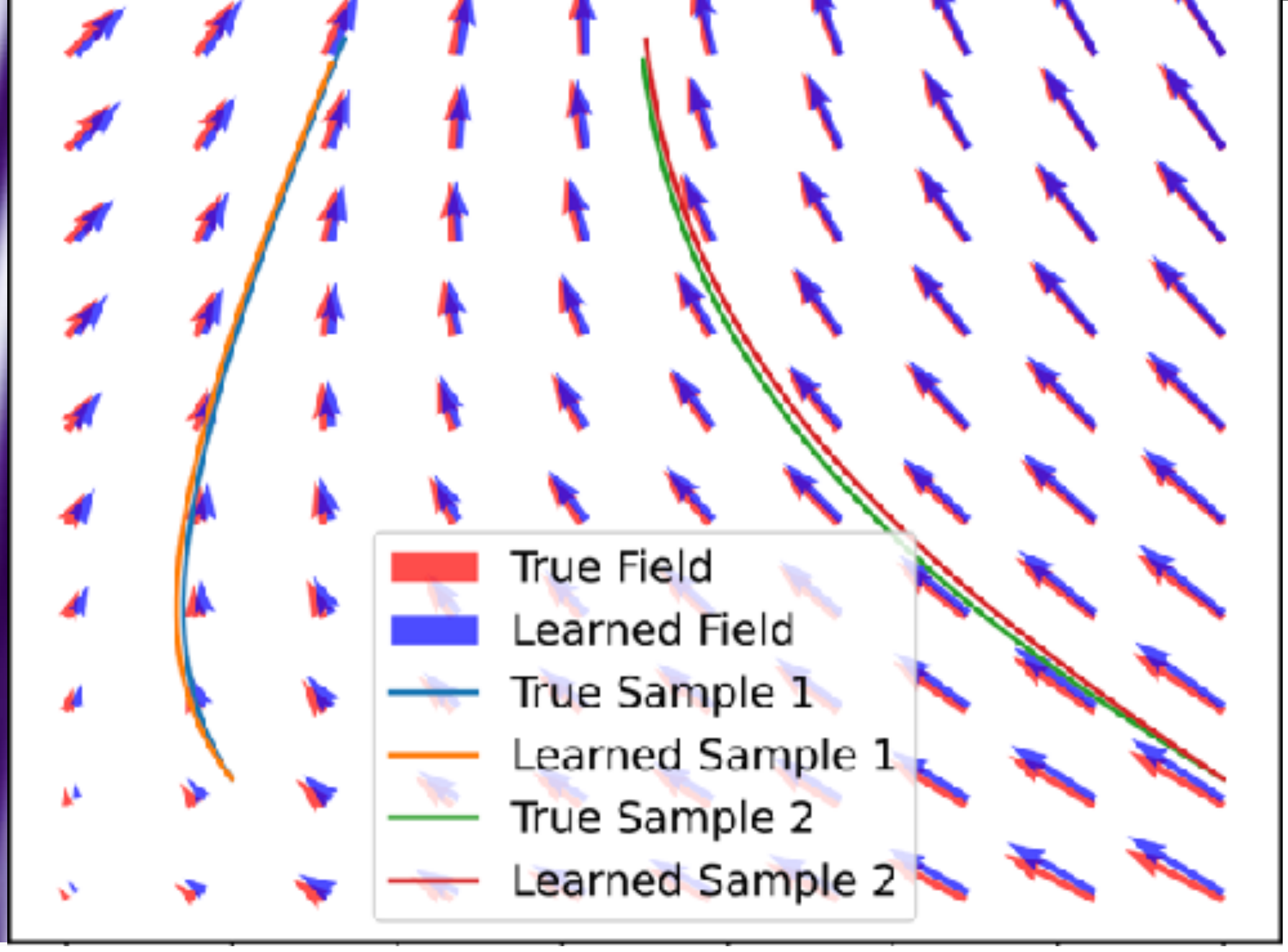
+

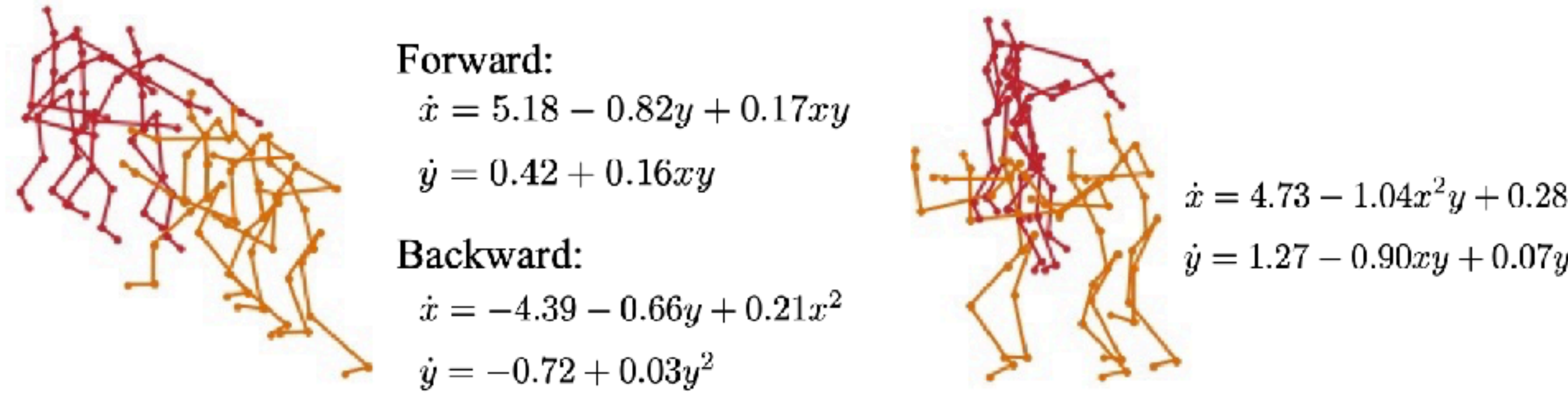
https://phlippe.github.io/BISCUIT/

LIU ET AL, ICML 2023, ONGOING
AUZINA ET AL, NEURIPS 2023
PERVEZ ET AL, ONGOING

DYNAMICS INDUCTIVE BIASES







Forward:

$$\dot{x} = 5.18 - 0.82y + 0.17xy$$
$$\dot{y} = 0.42 + 0.16xy$$

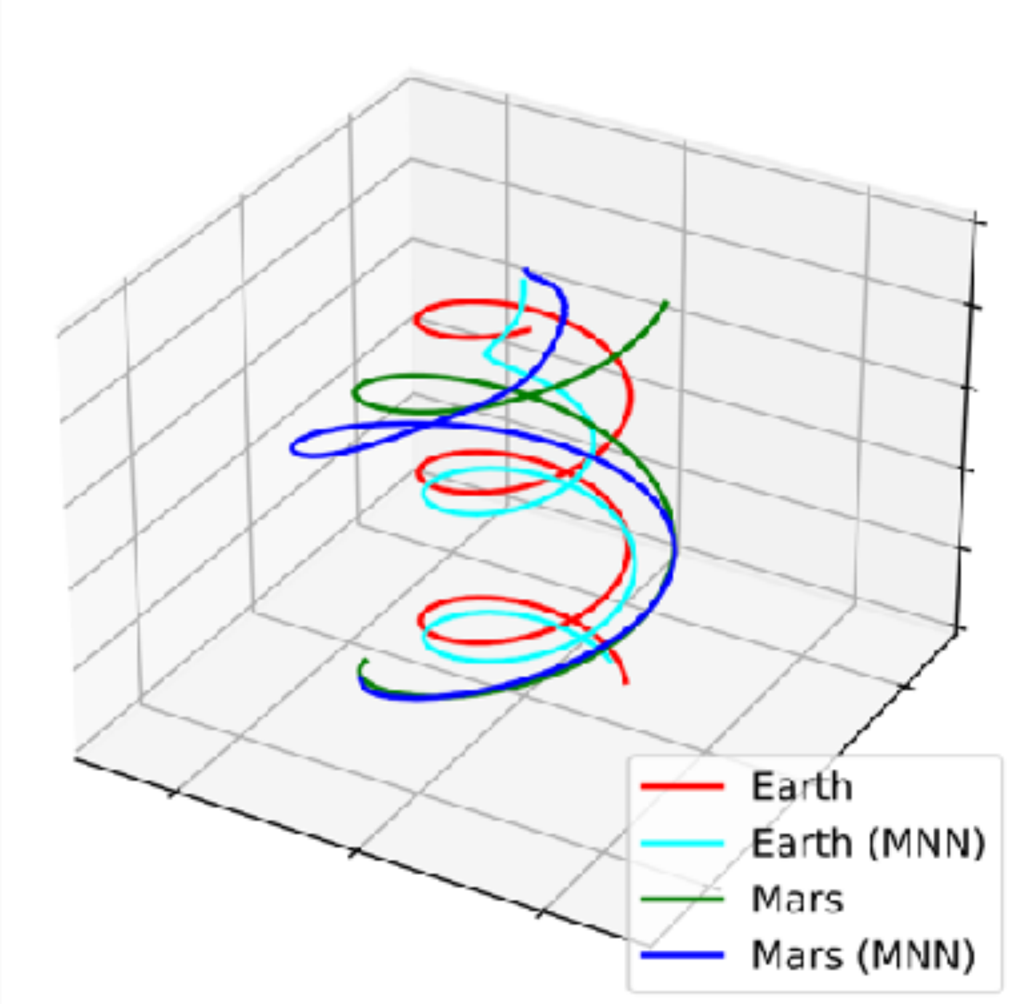
Backward:

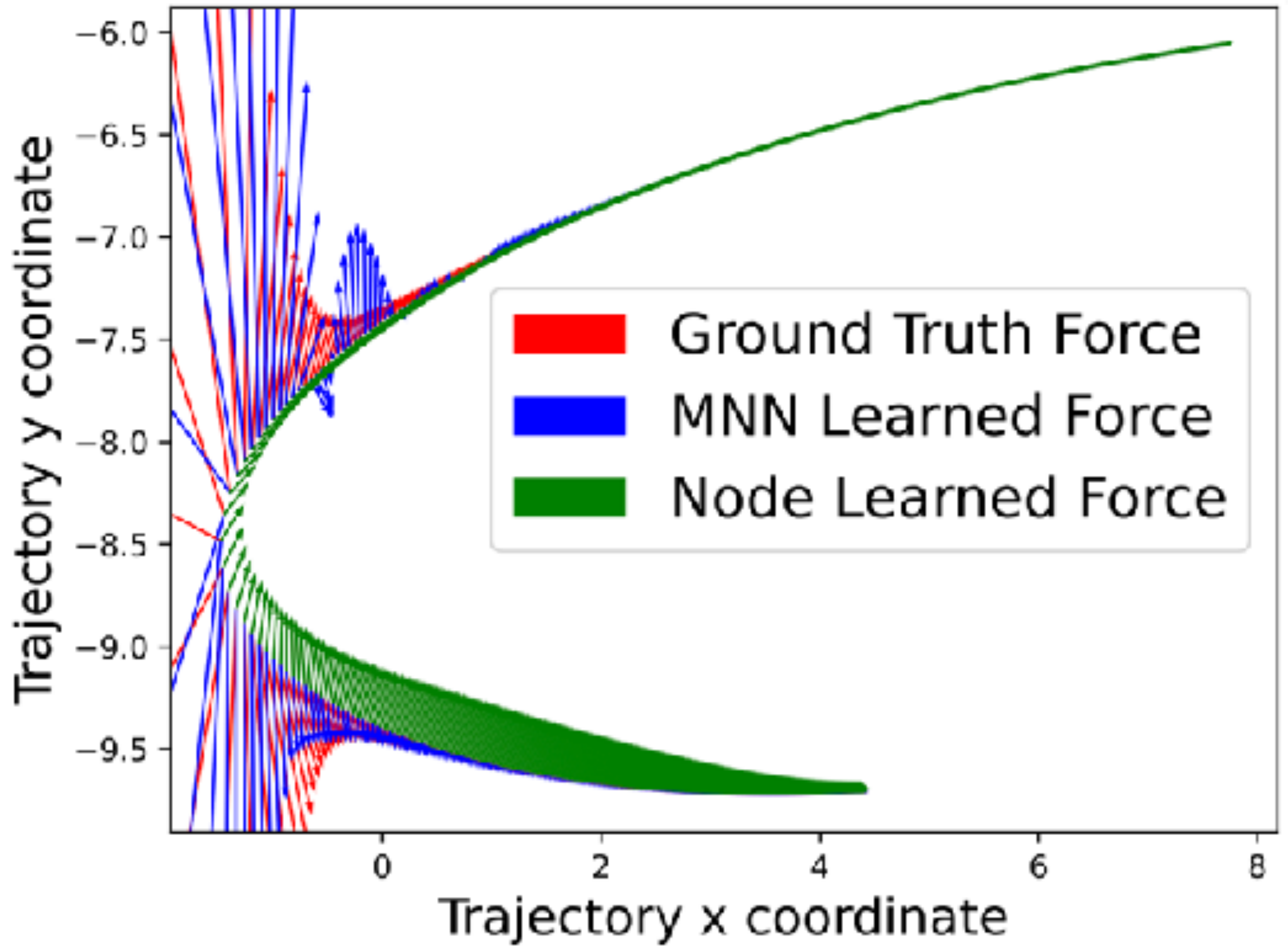
$$\dot{x} = -4.39 - 0.66y + 0.21x^2$$
$$\dot{y} = -0.72 + 0.03y^2$$

$$\dot{x} = 4.73 - 1.04x^2y + 0.28$$
$$\dot{y} = 1.27 - 0.90xy + 0.07y$$

(a) Forward and Backward

(b) Clockwise

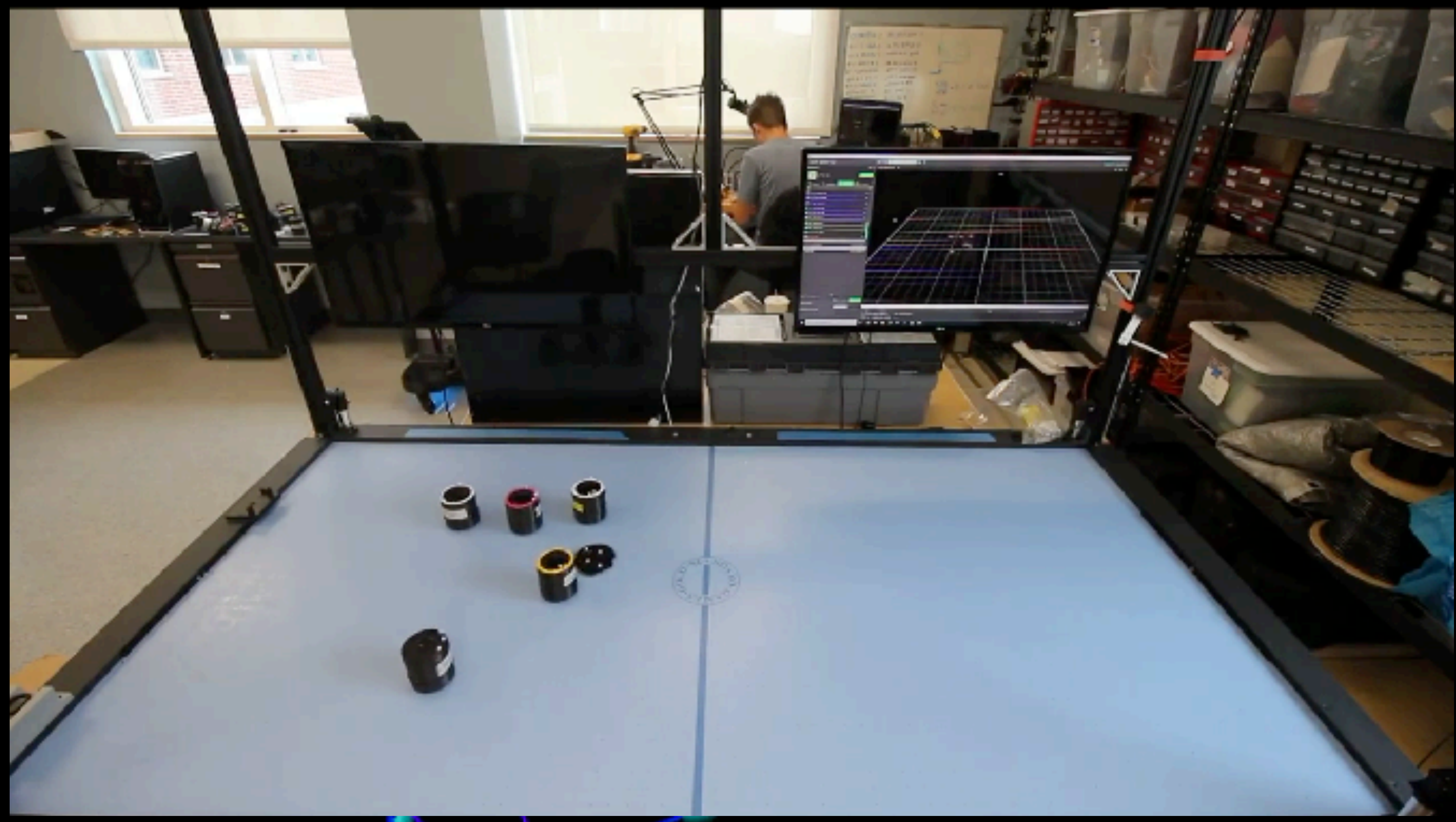


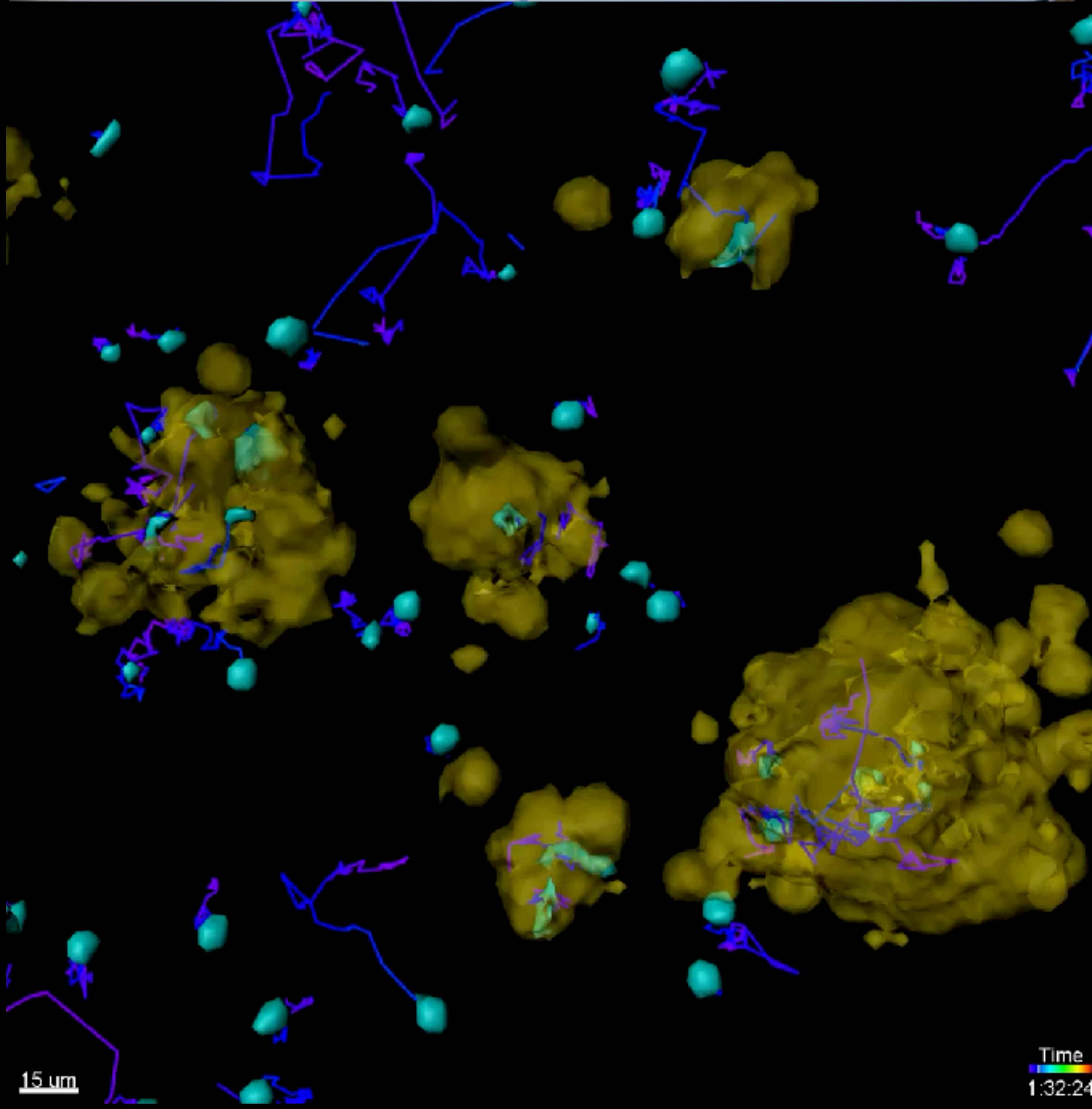


SWITCHING DYNAMICS IN INTERACTING SYSTEMS

- In many settings in perception and sciences, we have systems of multiple objects
- These objects may interact (or not) with higher-order complex & switching temporal dynamics[1,2]
- Finding dynamical patterns is often critical
- Generalization of temporal clustering

[1] Z Gharamani, G Hinton, Variational learning for switching state-space models, NeurIPS, 2020
[2] A Ansari, K Benidis, R Kurle, A Turkmen, H So, A Smola, B Wang, T Januschowski, NeurIPS, 2022

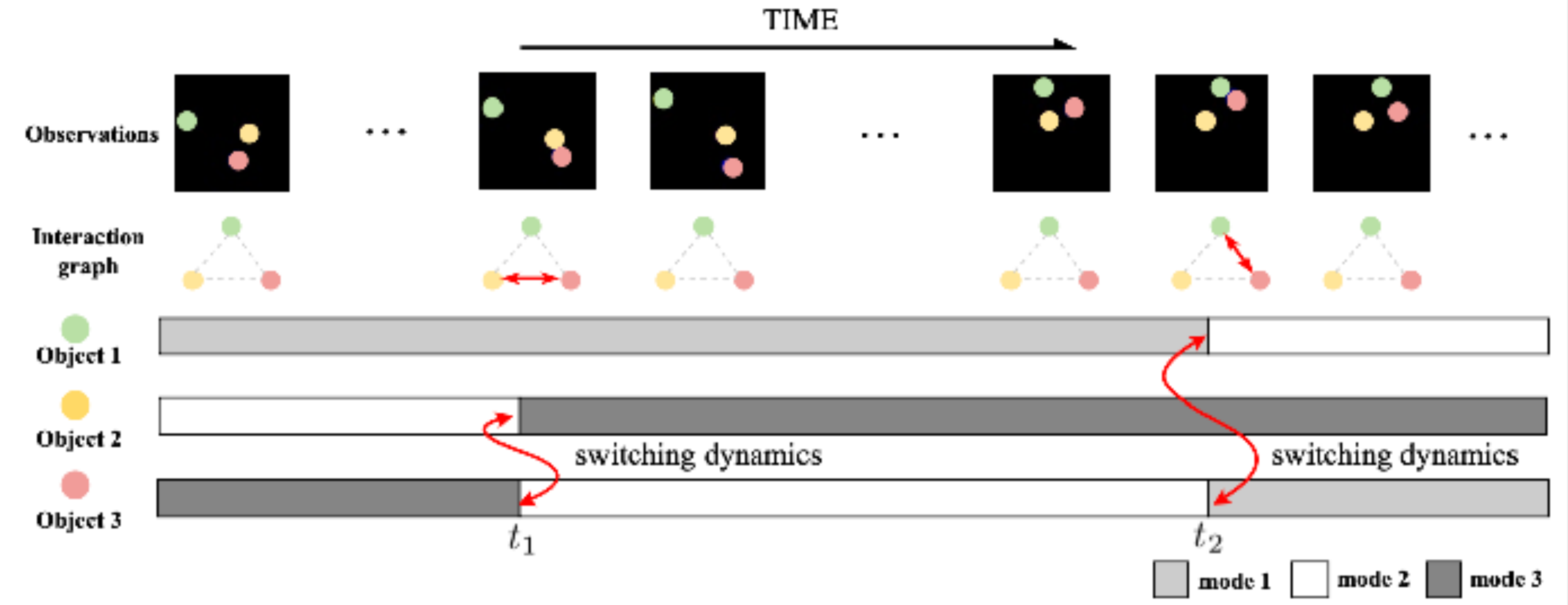




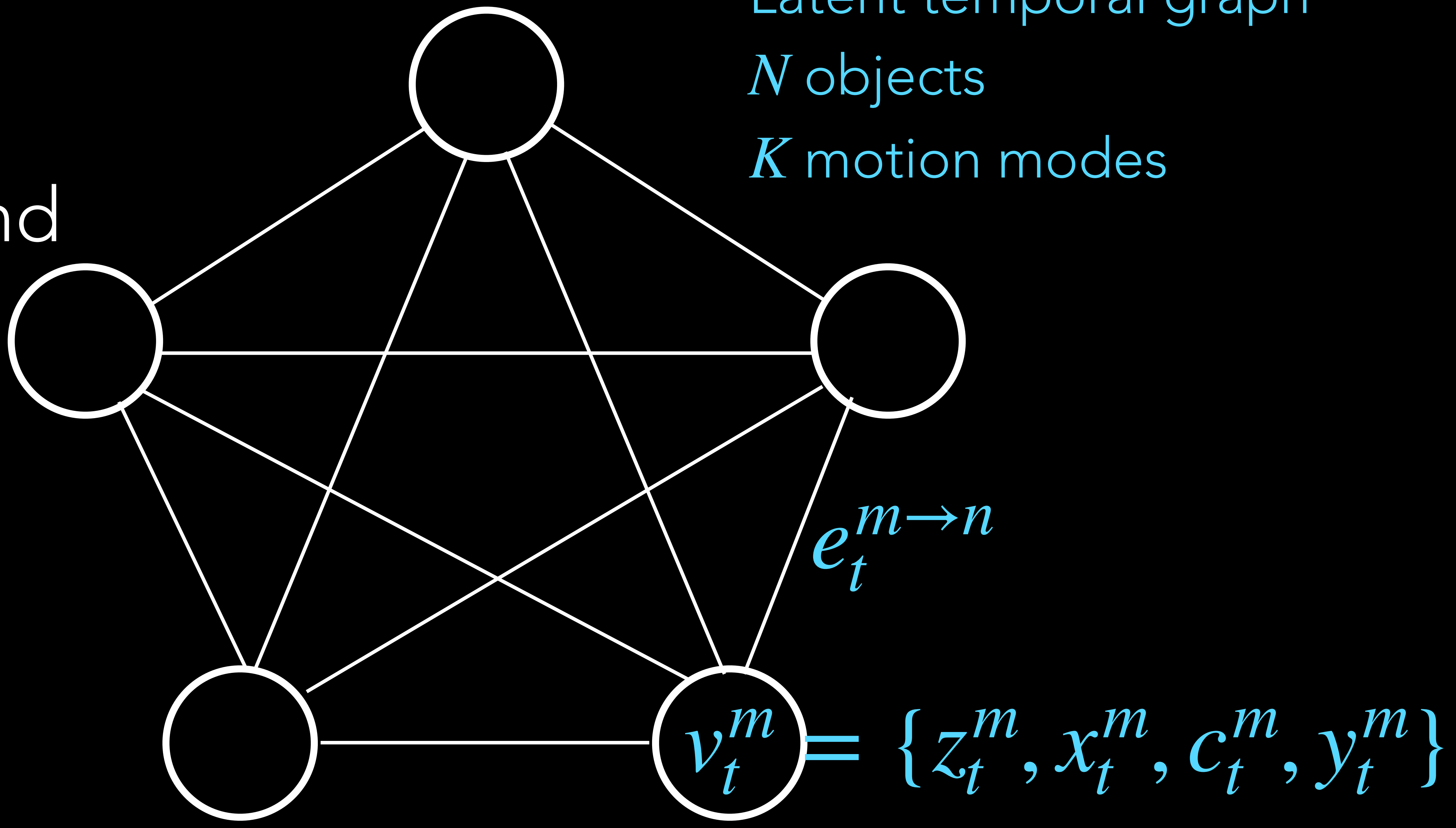
LIU, MAGLIACANE, KOFINAS, GAVVES, ICML 2023

GRAPH SWITCHING DYNAMICAL SYSTEMS

- Switching Dynamical Systems focus on finding out when objects behave differently
- Key idea #1: Scale up by NN function approximation to amortize pairwise transition dynamics between multiple objects and dynamic behaviors
- Key idea #2: Graph NNs and message passing and VI for dynamic interactions between objects
- Divide and Conquer: Breaking complex dynamics into switching between simpler systems



Latent temporal graph
N objects
K motion modes



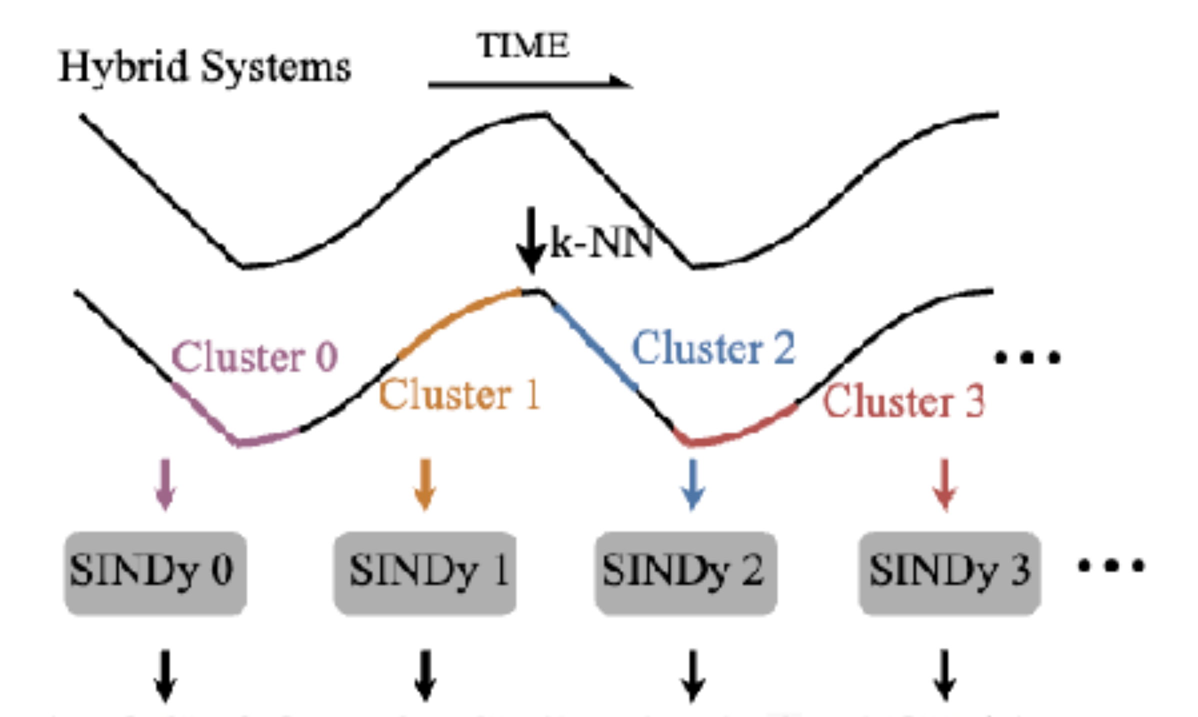
https://github.com/yongtuoliu/Graph-Switching-Dynamical-Systems

LIU, MAGLIACANE, KOFINAS, GAVVES, ONGOING

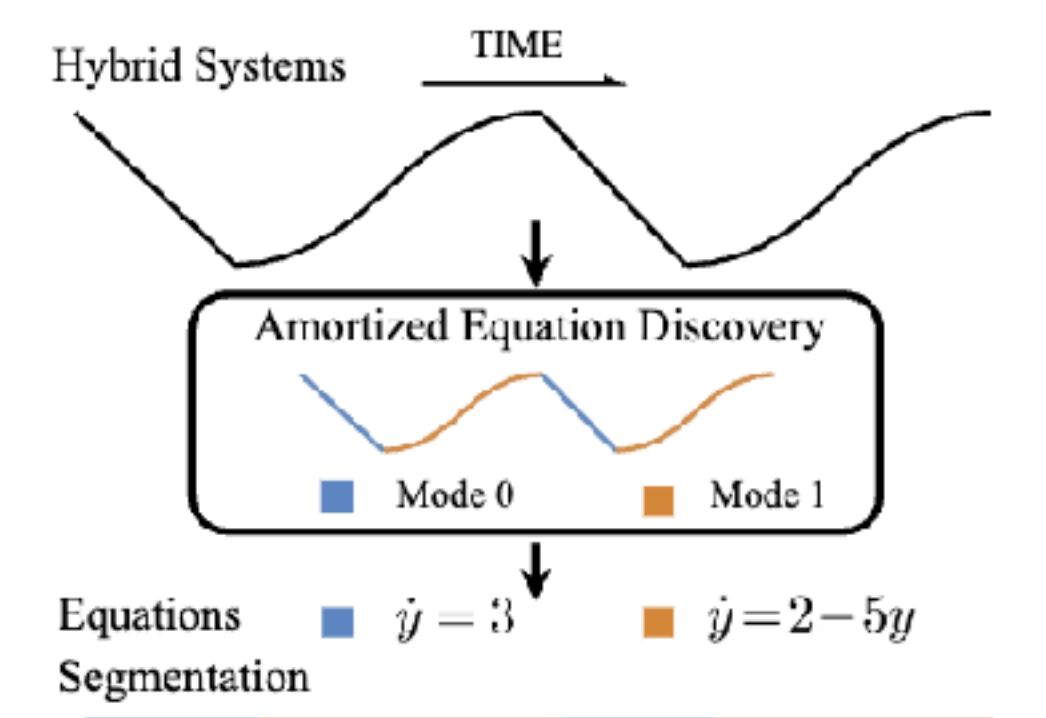
EQUATION DISCOVERY+SWITCHING DYNAMICS

- Hypothesis #1: Symbolic learning critical for extrapolation
- Hypothesis #2: Disentangled representation learning critical for generalization[1,2]
- Hypothesis #3: Graph learning critical for interacting dynamics

Hybrid-SINDy



Latent dynamics as governing equations



---

[1] J von Kügelgen*, Y Sharma*, L Gresele*, W Brendel, B Schölkopf, M Besserve, F Locatello, Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style, NeurIPS, 2021

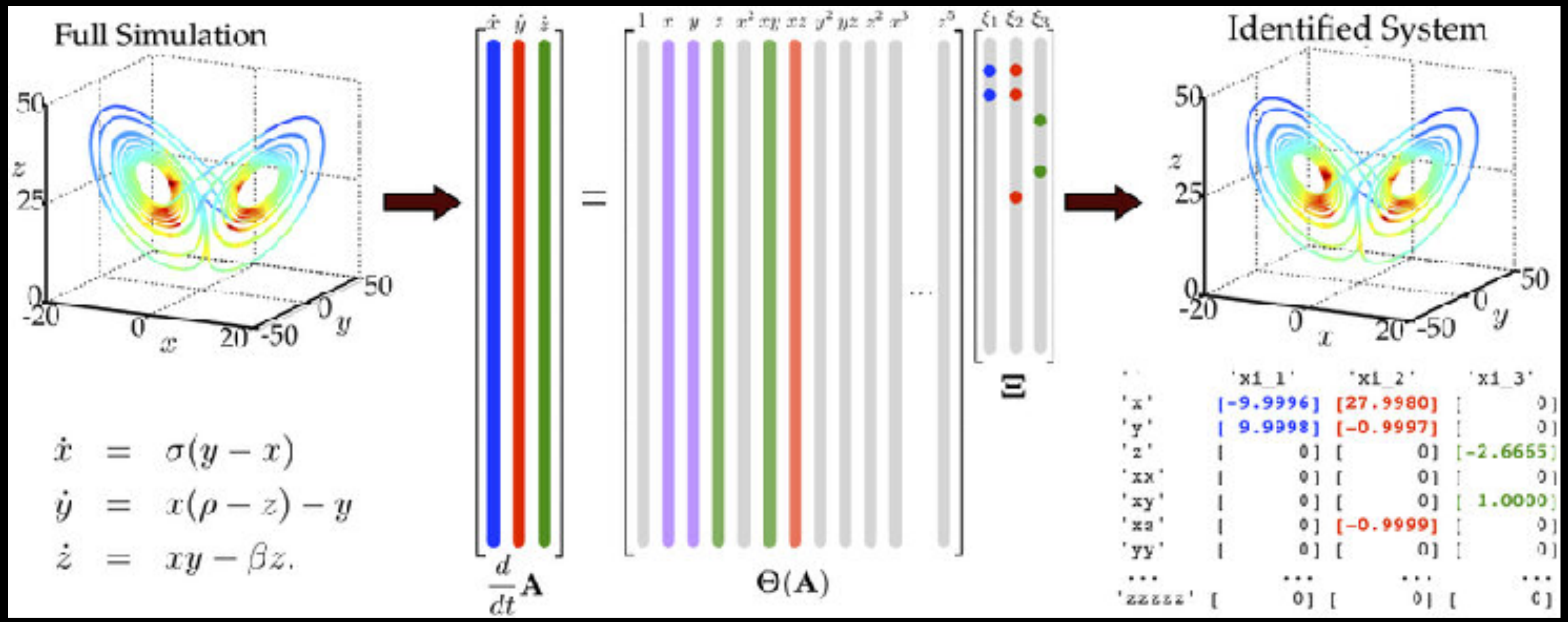
[2] I Auzina, C Yildiz, S Magliacane, M Bethge, E Gavves, Moduleated Neural ODEs, NeurIPS, 2023

LIU, MAGLIACANE, KOFINAS, GAVVES, ONGOING

EQUATION DISCOVERY+SWITCHING DYNAMICS

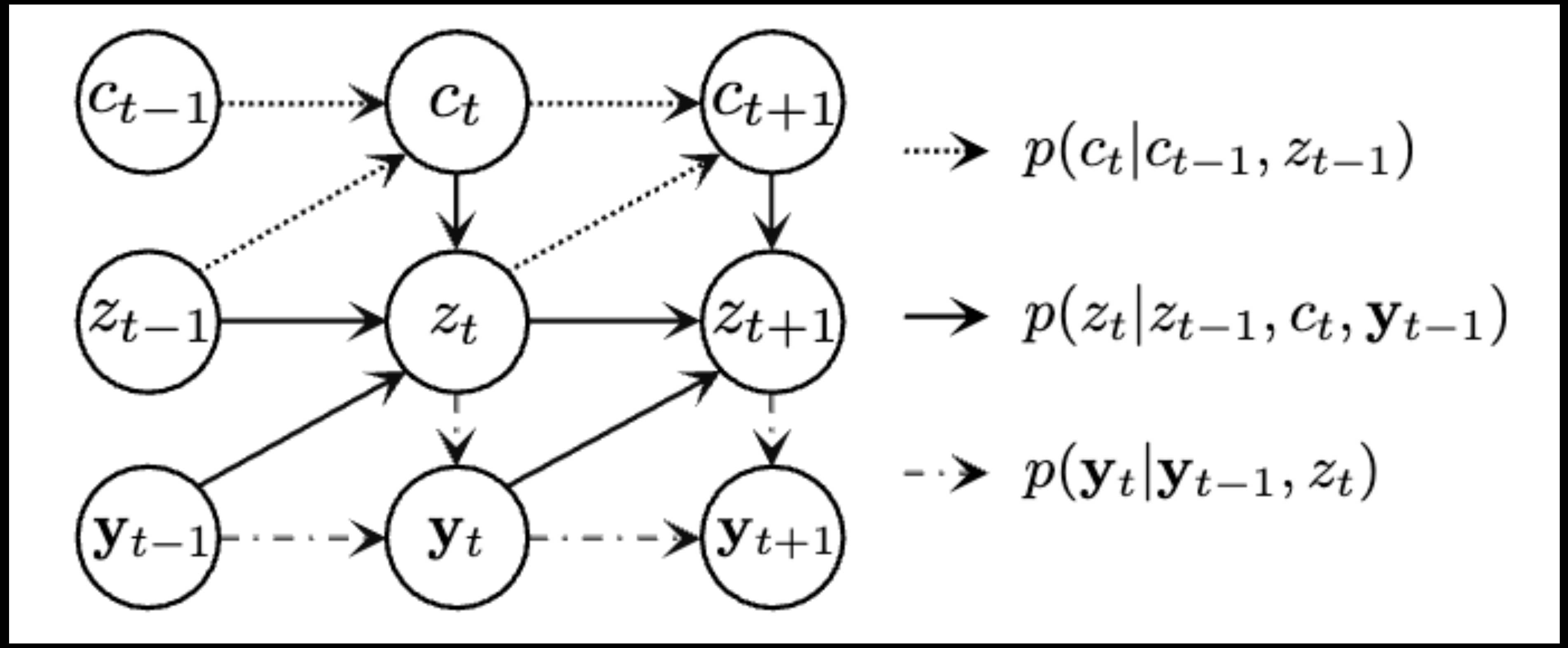
- #H1: Link GRASS dynamics modes to parametric basis functions $\Theta(y_t)$
- #H2: Learn k-th latent dynamics $\dot{y}_t^m \approx \Theta(y_t)w_m$ modulated by w_m in shared NN
- Think of SINDy-governed latent spaces

SINDy


Full Simulation
50
25
0
z
-20
0
20
x
-50
0
50
y
$\dot{x} = \sigma(y - x)$
$\dot{y} = x(\rho - z) - y$
$\dot{z} = xy - \beta z.$
$\dot{x}$ $\dot{y}$ $\dot{z}$
$\frac{d}{dt}\mathbf{A}$
=
1 x y z x^2 xy xz y^2 yz z^2 x^3 … z^5
$\Theta(\mathbf{A})$
ξ_1 ξ_2 ξ_3
Ξ
Identified System
50
25
0
z
-20
0
20
x
-50
0
50
y

```
''          'x1_1'     'x1_2'     'x1_3'
'x'     [-9.9996] [27.9980] [      0]
'y'     [ 9.9998] [-0.9997] [      0]
'z'     [      0] [      0] [-2.6665]
'xx'    [      0] [      0] [      0]
'xy'    [      0] [      0] [ 1.0000]
'xz'    [      0] [-0.9999] [      0]
'yy'    [      0] [      0] [      0]
...          ...       ...       ...
'zzzzz' [      0] [      0] [      0]
```

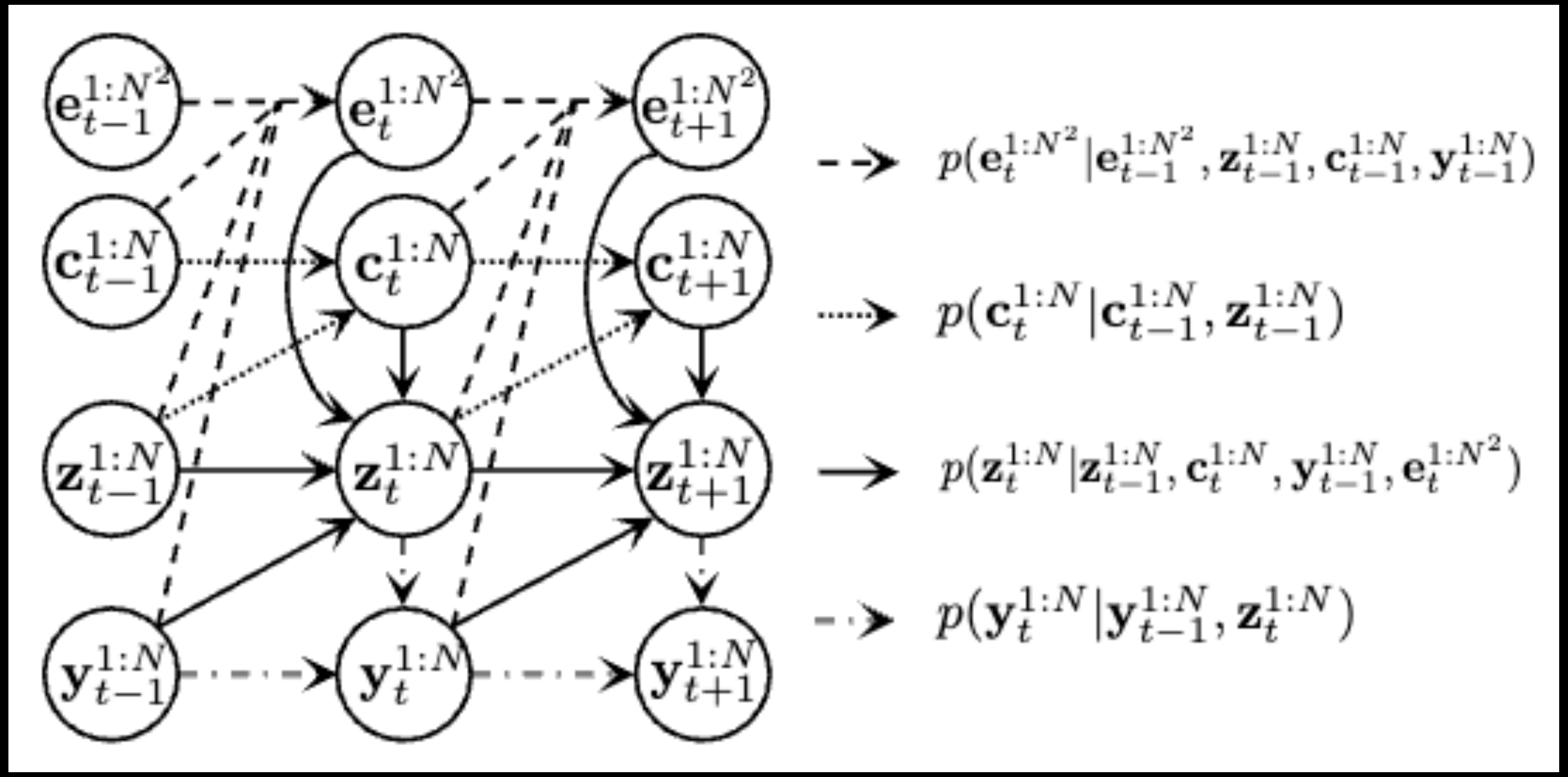
The dynamic mode z_k indexes candidate basis $\Theta_k \cdot w_k$


c_{t-1}
c_t
c_{t+1}
z_{t-1}
z_t
z_{t+1}
$\mathbf{y}_{t-1}$
$\mathbf{y}_t$
$\mathbf{y}_{t+1}$
$p(c_t|c_{t-1}, z_{t-1})$
$p(z_t|z_{t-1}, c_t, \mathbf{y}_{t-1})$
$p(\mathbf{y}_t|\mathbf{y}_{t-1}, z_t)$

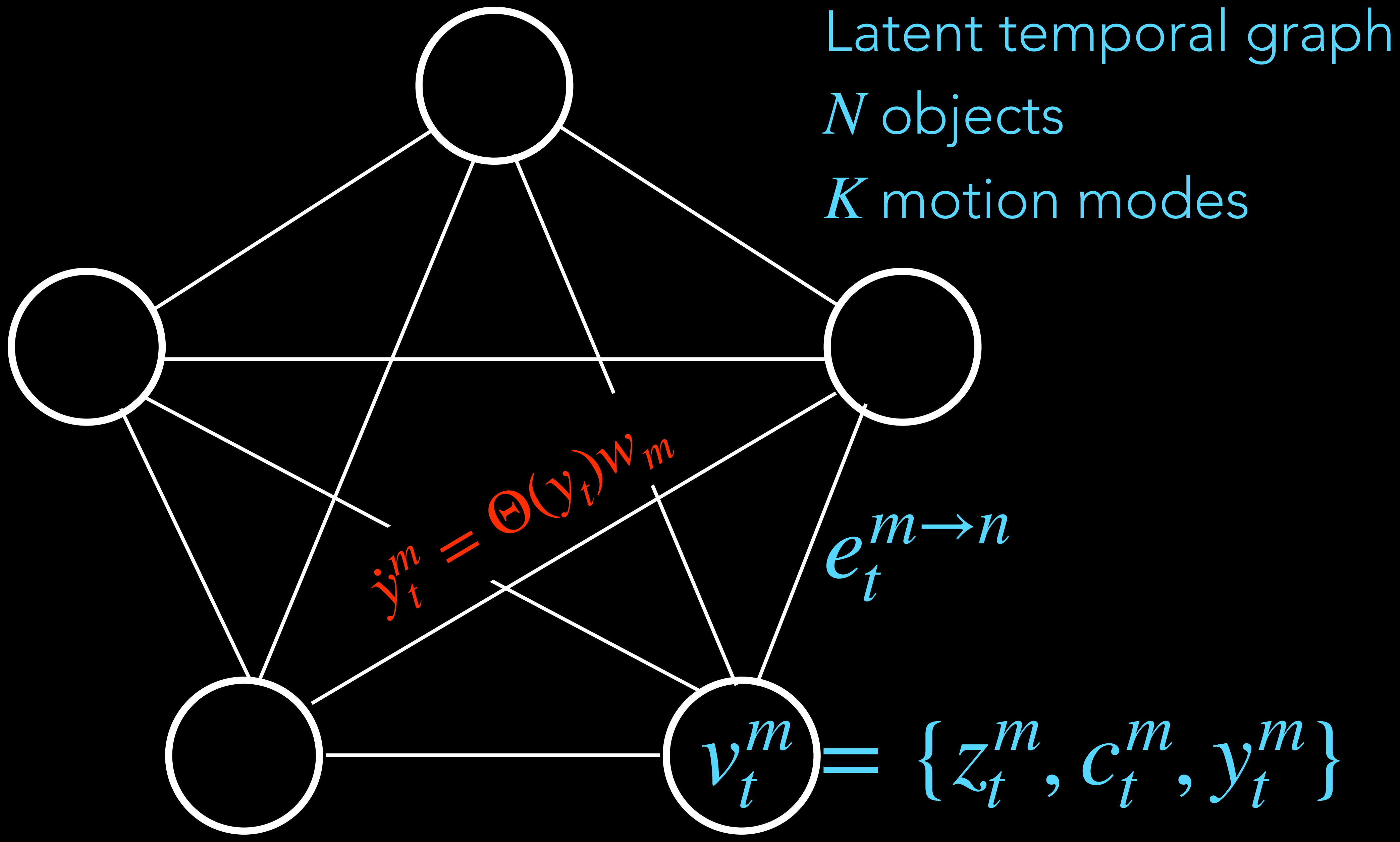
LIU, MAGLIACANE, KOFINAS, GAVVES, ONGOING

EQUATION DISCOVERY+SWITCHING DYNAMICS

- So far, single-object dynamics
- #H3: Graph message passing for interacting dynamics
- Approximate inference for edges, exact inference for discrete variables


$\mathbf{e}_{t-1}^{1:N^2}$
$\mathbf{e}_{t}^{1:N^2}$
$\mathbf{e}_{t+1}^{1:N^2}$
$\mathbf{c}_{t-1}^{1:N}$
$\mathbf{c}_{t}^{1:N}$
$\mathbf{c}_{t+1}^{1:N}$
$\mathbf{z}_{t-1}^{1:N}$
$\mathbf{z}_{t}^{1:N}$
$\mathbf{z}_{t+1}^{1:N}$
$\mathbf{y}_{t-1}^{1:N}$
$\mathbf{y}_{t}^{1:N}$
$\mathbf{y}_{t+1}^{1:N}$
$p(\mathbf{e}_t^{1:N^2} | \mathbf{e}_{t-1}^{1:N^2}, \mathbf{z}_{t-1}^{1:N}, \mathbf{c}_{t-1}^{1:N}, \mathbf{y}_{t-1}^{1:N})$
$p(\mathbf{c}_t^{1:N} | \mathbf{c}_{t-1}^{1:N}, \mathbf{z}_{t-1}^{1:N})$
$p(\mathbf{z}_t^{1:N} | \mathbf{z}_{t-1}^{1:N}, \mathbf{c}_{t}^{1:N}, \mathbf{y}_{t-1}^{1:N}, \mathbf{e}_t^{1:N^2})$
$p(\mathbf{y}_t^{1:N} | \mathbf{y}_{t-1}^{1:N}, \mathbf{z}_{t}^{1:N})$

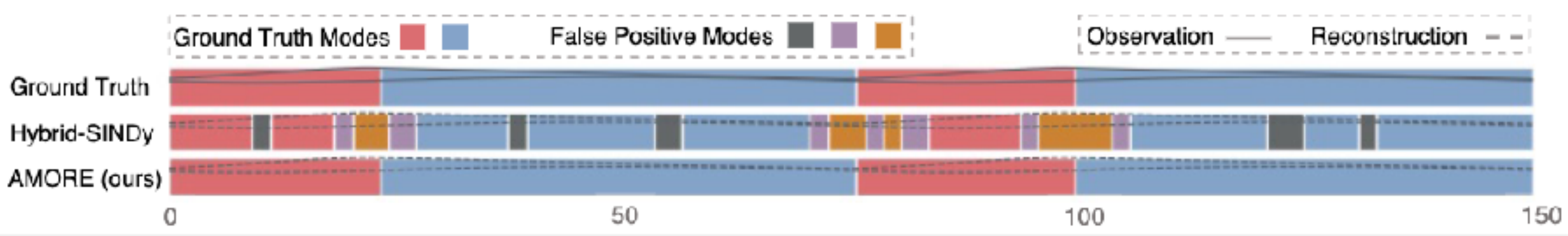
Probabilistic model


Latent temporal graph
N objects
K motion modes
$\dot{y}_t^m = \Theta(y_t) w_m$
$e_t^{m \to n}$
$v_t^m = \{z_t^m, c_t^m, y_t^m\}$

RESULTS

EQUATION DISCOVERY+SWITCHING DYNAMICS

Scientific data



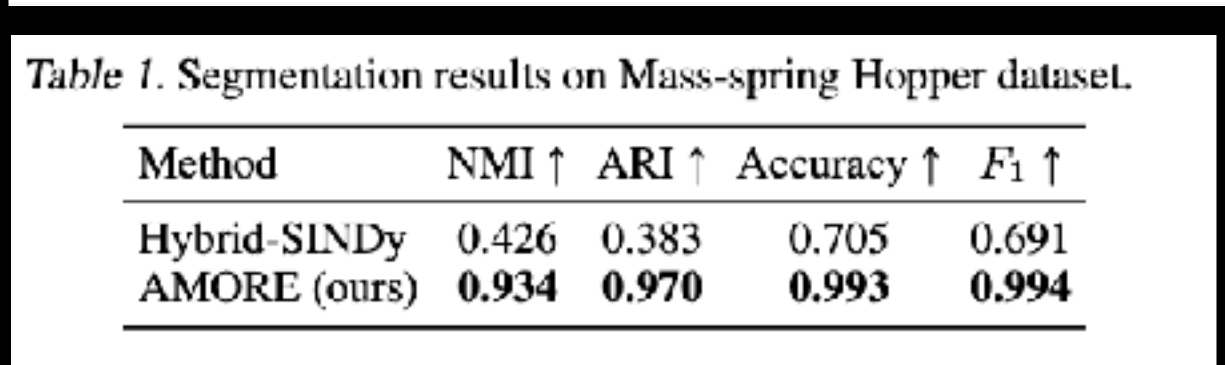


Table 1. Segmentation results on Mass-spring Hopper dataset.

Method	NMI ↑	ARI ↑	Accuracy ↑	F_1 ↑
Hybrid-SINDy	0.426	0.383	0.705	0.691
AMORE (ours)	**0.934**	**0.970**	**0.993**	**0.994**

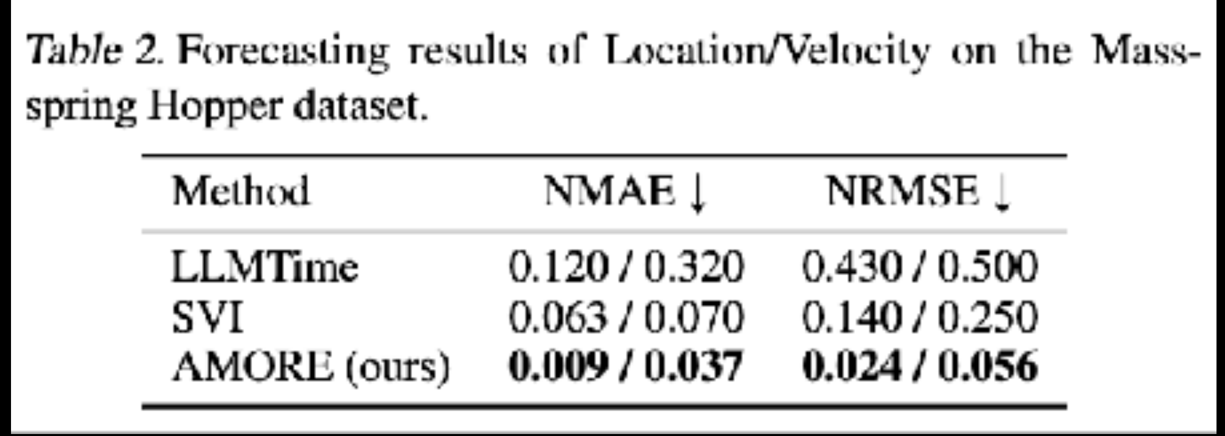


Table 2. Forecasting results of Location/Velocity on the Mass-spring Hopper dataset.

Method	NMAE ↓	NRMSE ↓
LLMTime	0.120 / 0.320	0.430 / 0.500
SVI	0.063 / 0.070	0.140 / 0.250
AMORE (ours)	**0.009 / 0.037**	**0.024 / 0.056**

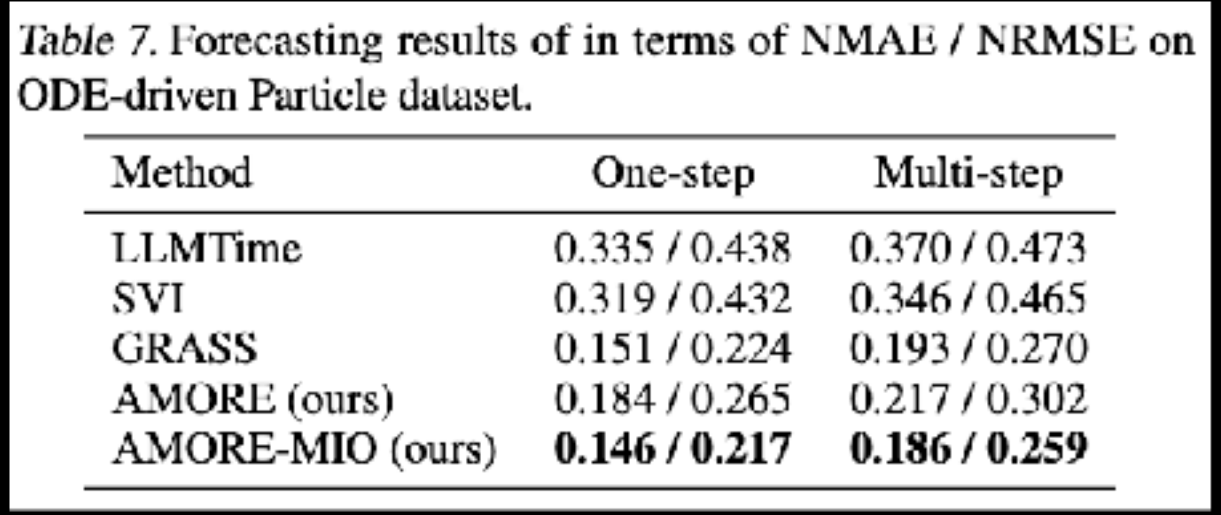


Table 7. Forecasting results of in terms of NMAE / NRMSE on ODE-driven Particle dataset.

Method	One-step	Multi-step
LLMTime	0.335 / 0.438	0.370 / 0.473
SVI	0.319 / 0.432	0.346 / 0.465
GRASS	0.151 / 0.224	0.193 / 0.270
AMORE (ours)	0.184 / 0.265	0.217 / 0.302
AMORE-MIO (ours)	**0.146 / 0.217**	**0.186 / 0.259**

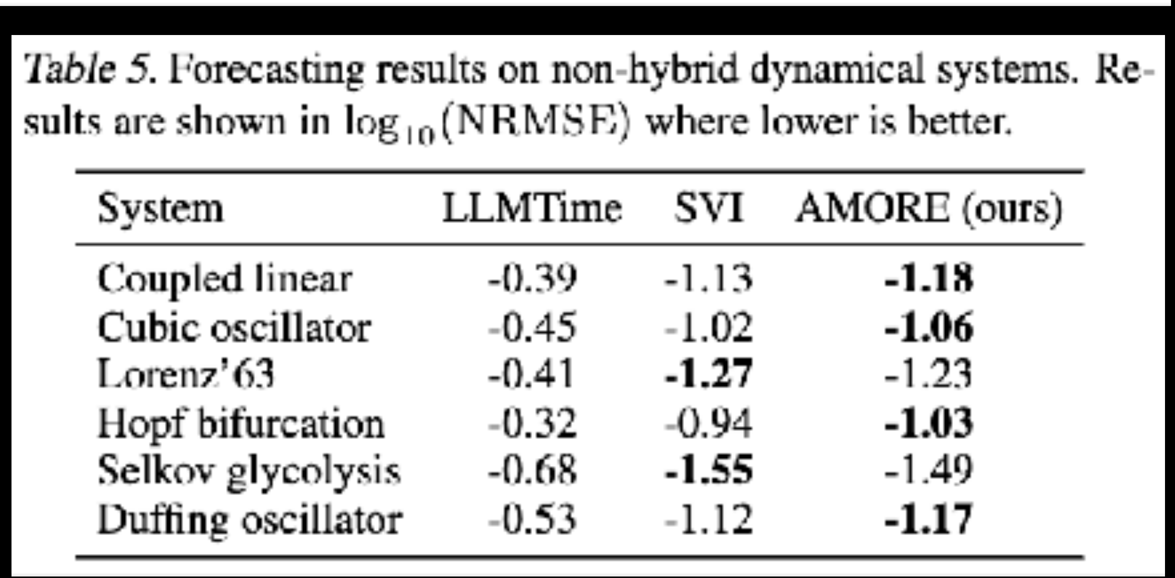


Table 5. Forecasting results on non-hybrid dynamical systems. Results are shown in $\log_{10}$(NRMSE) where lower is better.

System	LLMTime	SVI	AMORE (ours)
Coupled linear	-0.39	-1.13	**-1.18**
Cubic oscillator	-0.45	-1.02	**-1.06**
Lorenz'63	-0.41	**-1.27**	-1.23
Hopf bifurcation	-0.32	-0.94	**-1.03**
Selkov glycolysis	-0.68	**-1.55**	-1.49
Duffing oscillator	-0.53	-1.12	**-1.17**

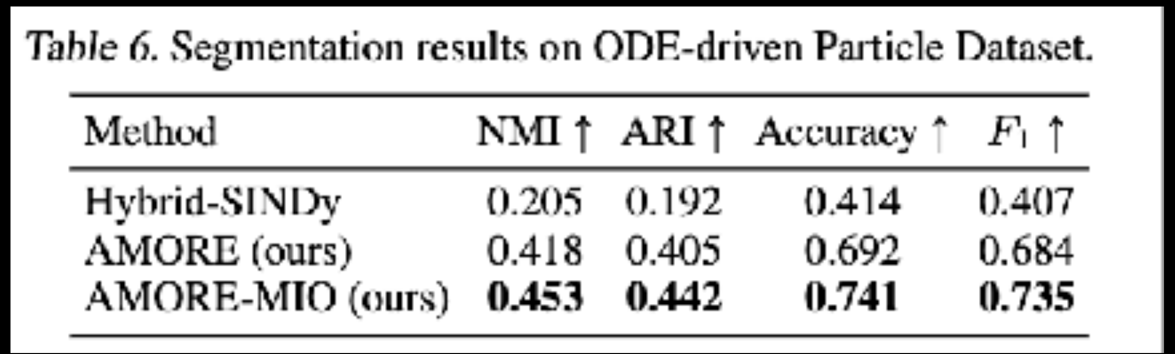


Table 6. Segmentation results on ODE-driven Particle Dataset.

Method	NMI ↑	ARI ↑	Accuracy ↑	F_1 ↑
Hybrid-SINDy	0.205	0.192	0.414	0.407
AMORE (ours)	0.418	0.405	0.692	0.684
AMORE-MIO (ours)	**0.453**	**0.442**	**0.741**	**0.735**

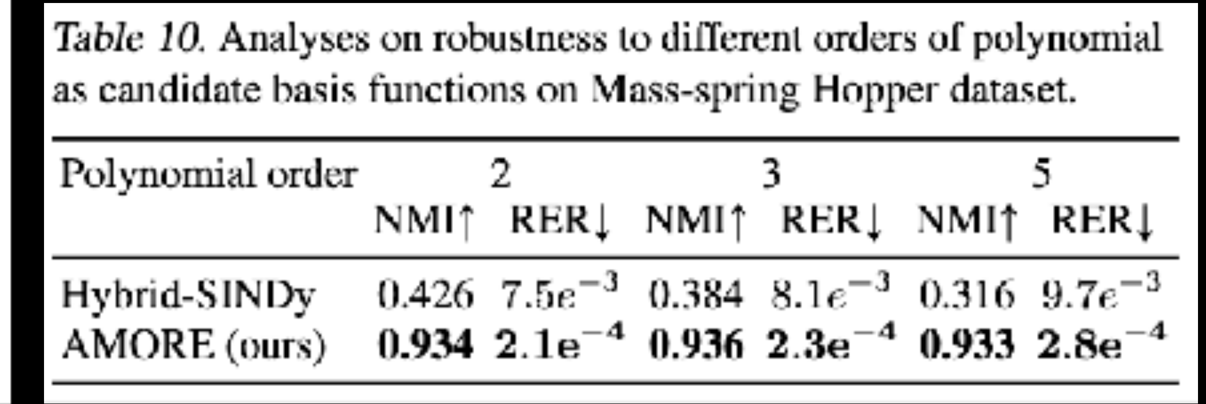
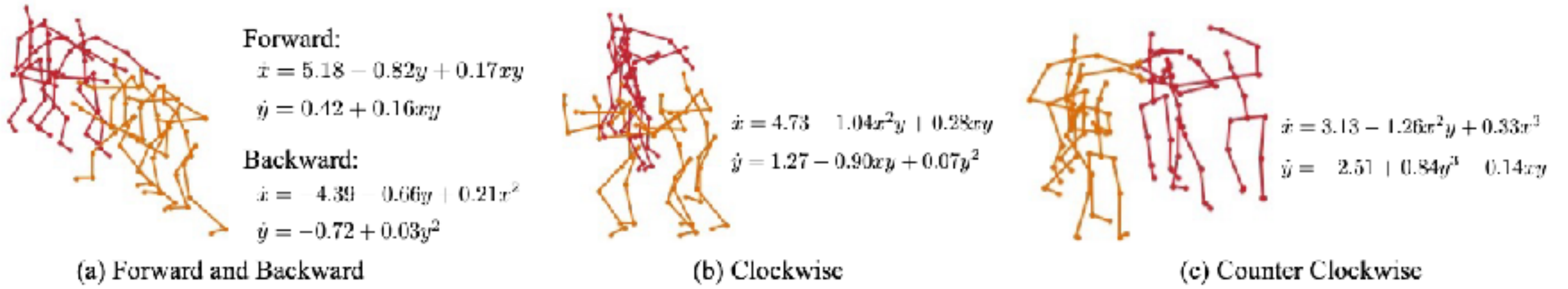


Table 10. Analyses on robustness to different orders of polynomial as candidate basis functions on Mass-spring Hopper dataset.

Polynomial order	2		3		5	
	NMI↑	RER↓	NMI↑	RER↓	NMI↑	RER↓
Hybrid-SINDy	0.426	$7.5e^{-3}$	0.384	$8.1e^{-3}$	0.316	$9.7e^{-3}$
AMORE (ours)	**0.934**	**$2.1e^{-4}$**	**0.936**	**$2.3e^{-4}$**	**0.933**	**$2.8e^{-4}$**

Perception data


Forward:
$\dot{x} = 5.18 - 0.82y + 0.17xy$
$\dot{y} = 0.42 + 0.16xy$
Backward:
$\dot{x} = -4.39 - 0.66y + 0.21x^2$
$\dot{y} = -0.72 + 0.03y^2$
$\dot{x} = 4.73 - 1.04x^2y + 0.28xy$
$\dot{y} = 1.27 - 0.90xy + 0.07y^2$
$\dot{x} = 3.13 - 1.26x^2y + 0.33x^3$
$\dot{y} = -2.51 + 0.84y^3 - 0.14xy$

(a) Forward and Backward (b) Clockwise (c) Counter Clockwise

Figure 5. Discovered equations on the Salsa-dancing dataset. Locations (x, y) of the hip joints are used as observations.

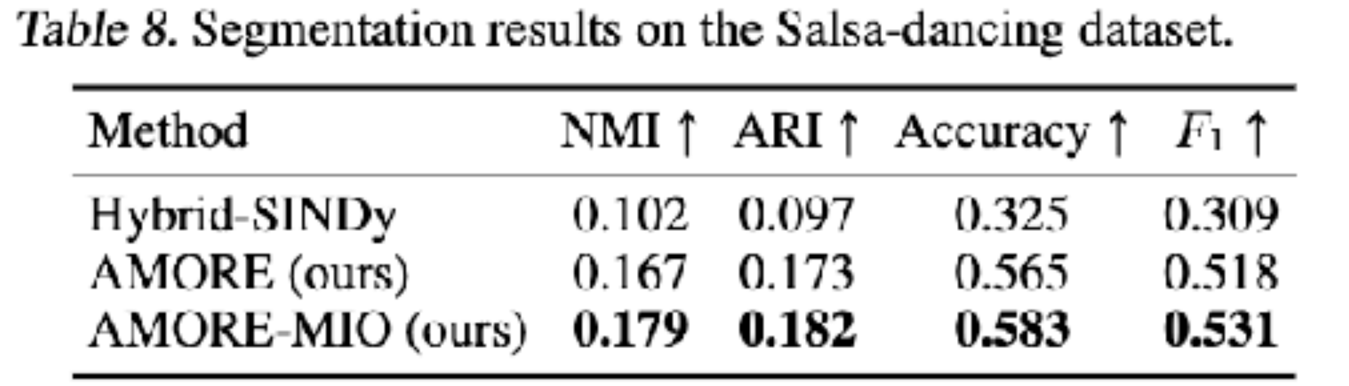


Table 8. Segmentation results on the Salsa-dancing dataset.

Method	NMI ↑	ARI ↑	Accuracy ↑	F_1 ↑
Hybrid-SINDy	0.102	0.097	0.325	0.309
AMORE (ours)	0.167	0.173	0.565	0.518
AMORE-MIO (ours)	**0.179**	**0.182**	**0.583**	**0.531**

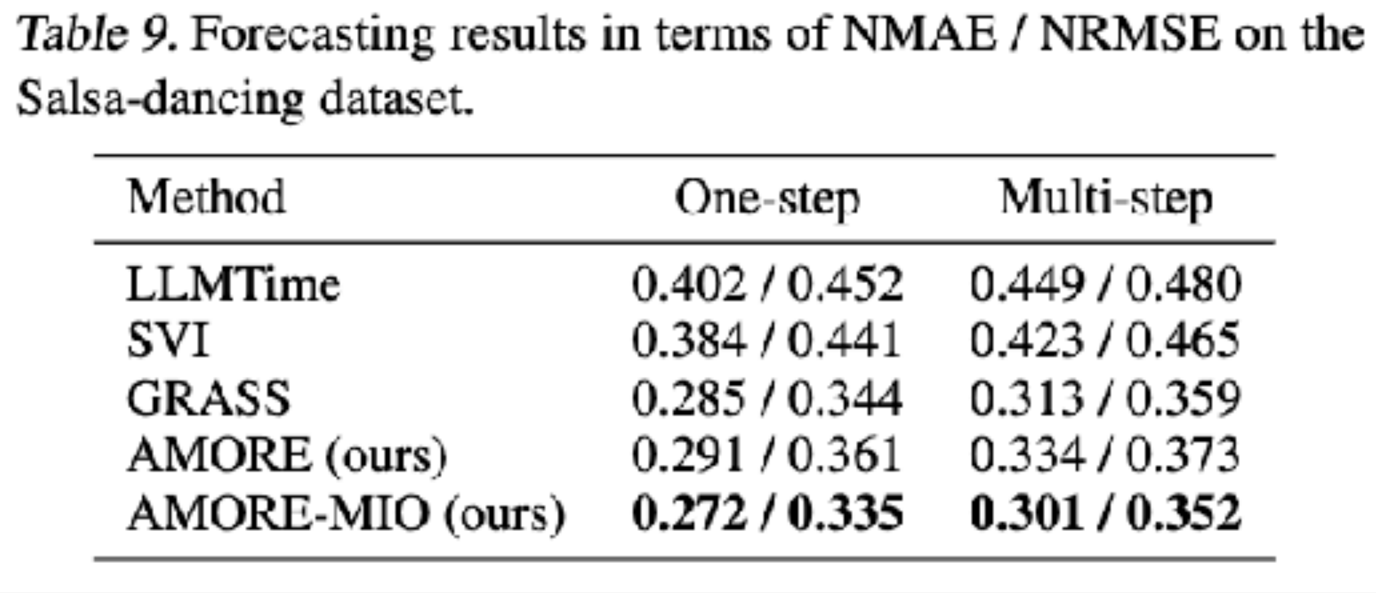


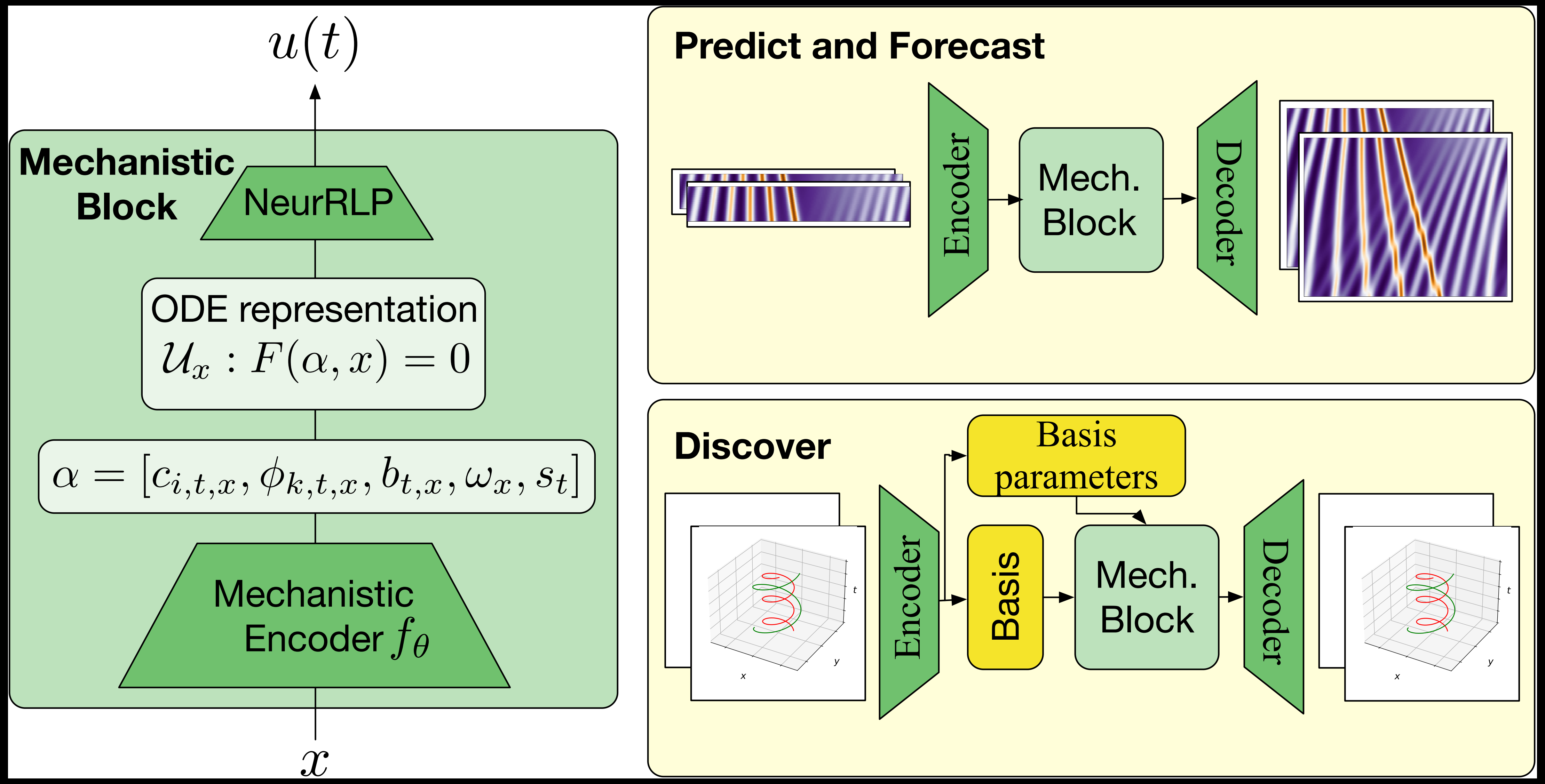
Table 9. Forecasting results in terms of NMAE / NRMSE on the Salsa-dancing dataset.

Method	One-step	Multi-step
LLMTime	0.402 / 0.452	0.449 / 0.480
SVI	0.384 / 0.441	0.423 / 0.465
GRASS	0.285 / 0.344	0.313 / 0.359
AMORE (ours)	0.291 / 0.361	0.334 / 0.373
AMORE-MIO (ours)	**0.272 / 0.335**	**0.301 / 0.352**

PERVEZ, LOCATELLO, GAVVES, ONGOING

MECHANISTIC NEURAL NETWORKS

- Neural Networks built on data-driven numerical representations
- Uninterpretable → unfit for scientific exploration and analysis
- Mechanisms would be a great alternative but cannot easily learn from data[1]
- Mechanistic NN modules built on governing equations as representations



$$F(\alpha, x) = \sum_{i}^{d} c_i(t;x)u^{(i)} + \sum_{k}^{r} \phi_k(t;x)g_k(t, u, u', \ldots) - b(t;x)$$

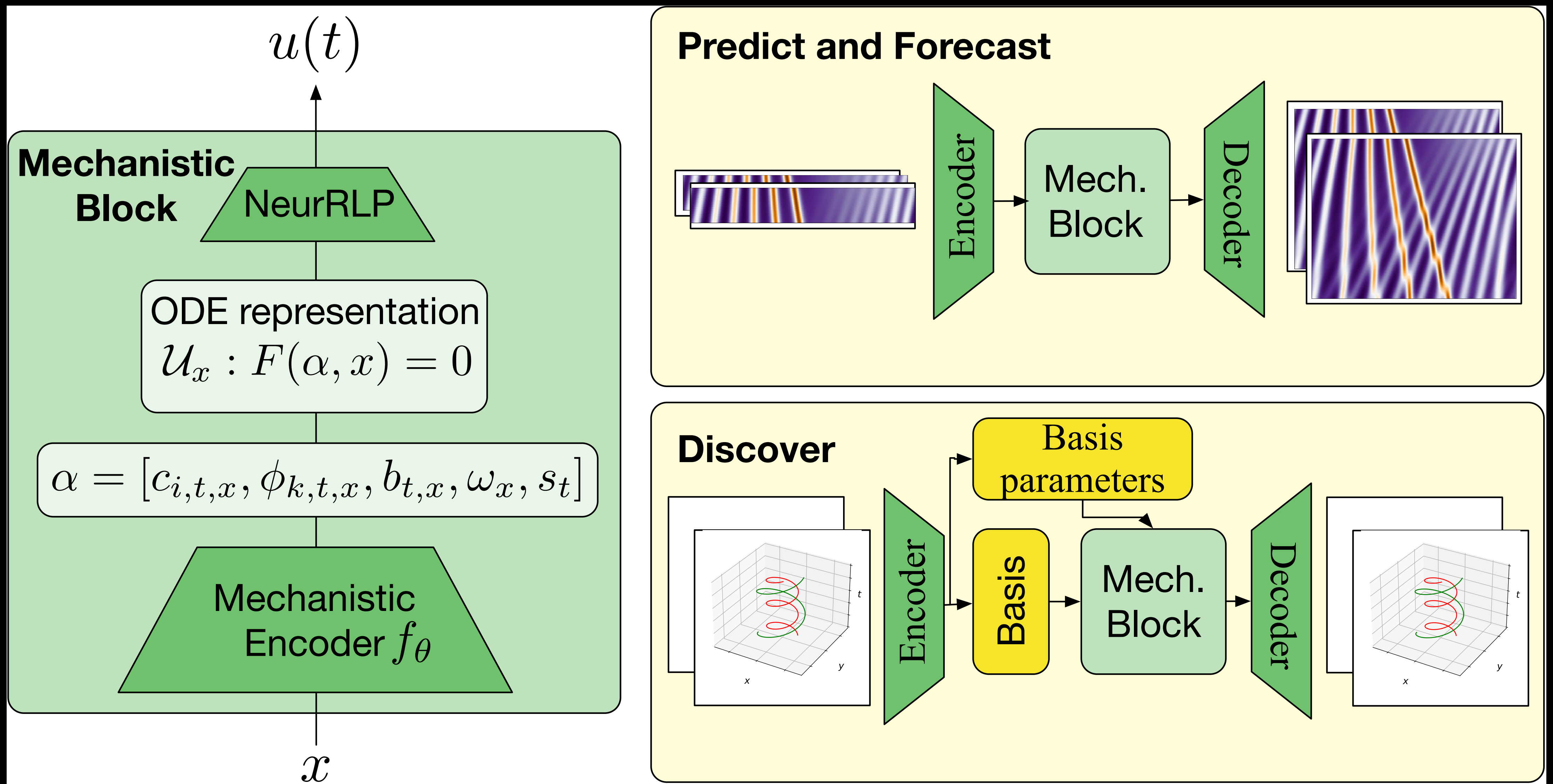
https://github.com/alpz/mech-nn

[1] B Schölkopf, F Locatello, S Bauer, N R Ke, N Kalchbrenner, A Goyal, Y Bengio, Towards Causal Representation Learning, Proceedings of the IEEE, 2021

PERVEZ, LOCATELLO, GAVVES, ONGOING

KEY IDEA

- Define general family of ODEs as governing mechanisms
- Mechanistic NN simultaneously
  - learns the governing ODE explicitly
  - generate new ODEs that explain input
  - solves the ODEs
- Forward pass through the ODE
- Backward pass requires custom, NN-native solver



$$F(\alpha, x) = \sum_{i}^{d} c_i(t;x)u^{(i)} + \sum_{k}^{r} \phi_k(t;x)g_k(t, u, u', \dots) - b(t;x)$$

https://github.com/alpz/mech-nn

PERVEZ, LOCATELLO, GAVVES, ONGOING

THE MODEL

- Traditional ODE solvers suboptimal: Hard to parallelize, no learned step sizes
- Young showed that Linear ODEs can be solved as Linear Programs[1]
- Continuous ODE: $\sum_i^d c_i(t;x)u^{(i)} + \sum_k^r \phi_k(t;x)g_k(t,u,u',\ldots) - b(t;x) = 0$
- Discretize it: $\sum_i^d c_{i,t}u^{(i)} + \sum_k^r \phi_{k,t}g_k(t,u_t,u_t',\ldots) - b_t = 0$
- Set up the linear program and solve for $u, u^{(i)}, i = 1,2,\ldots$

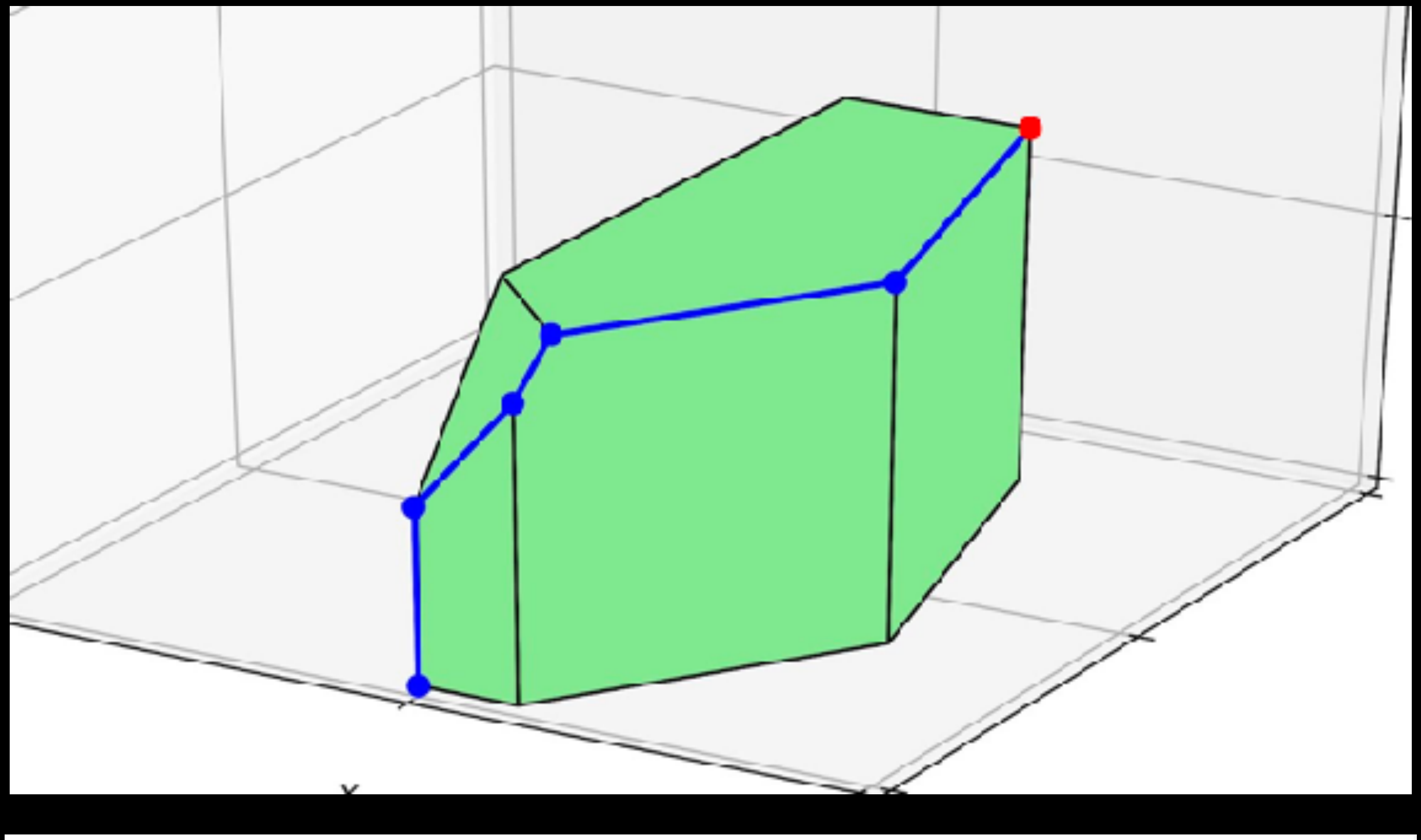
https://github.com/alpz/mech-nn

[1] J Young, Linear Programming Applied to Linear Differential Equations, Lawrence Berkeley National Laboratory, 1961

PERVEZ, LOCATELLO, GAVVES, ONGOING

NEURAL RELAXED LP SOLVER

- Define ODE coefficients etc as LP variables
- But LP solvers are not neural network friendly
  - Solutions not differentiable wrt parameters
  - Specialized solvers cannot parallelize easily
  - Constraint matrices too large for solvers
- Relaxing LP to QP and GPU solve KKT conditions
- Similar error bounds like the Euler solver and much faster
- For nonlinear ODEs, the nonlinear terms pushed to the NN loss function



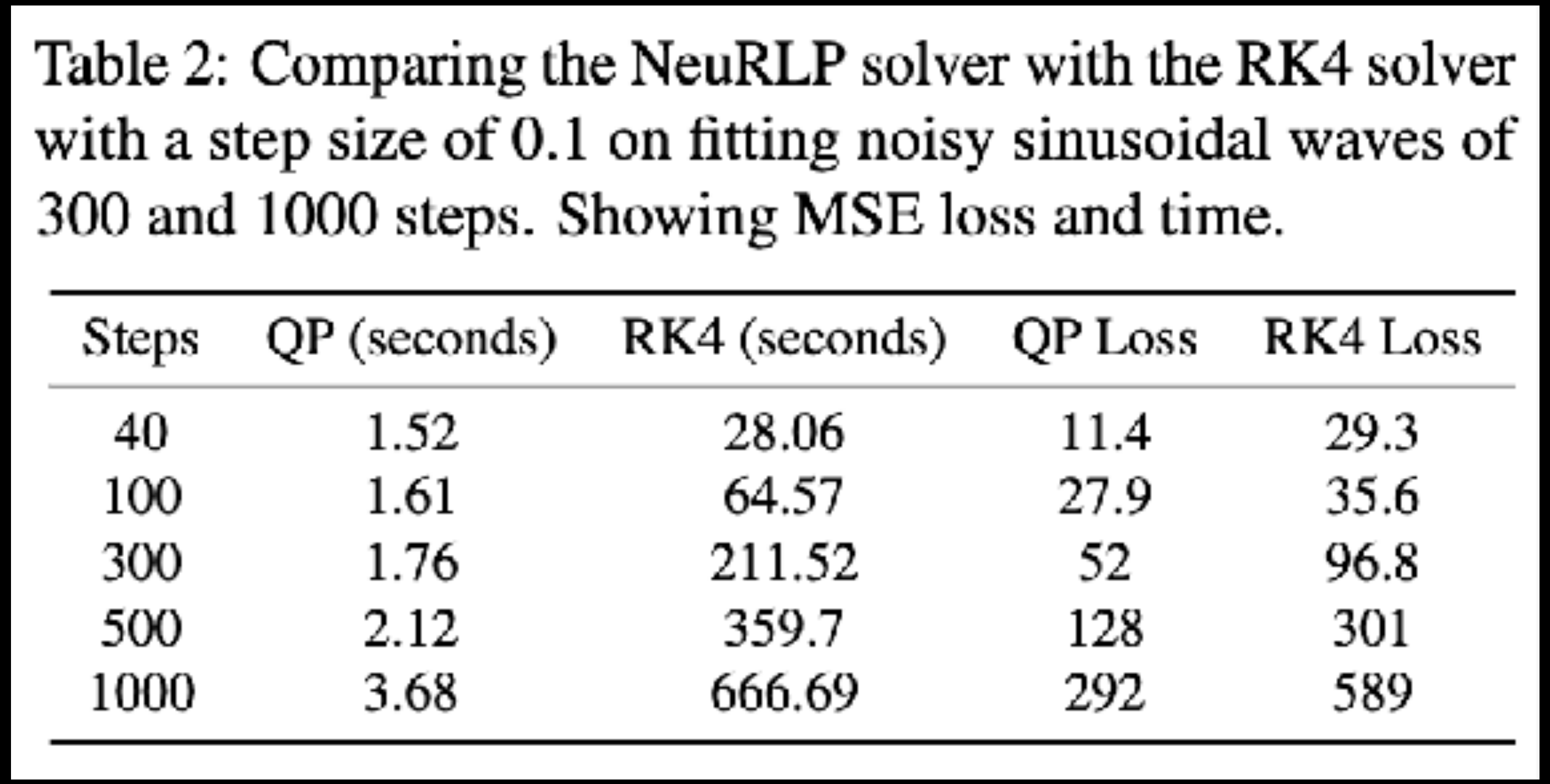


Table 2: Comparing the NeuRLP solver with the RK4 solver with a step size of 0.1 on fitting noisy sinusoidal waves of 300 and 1000 steps. Showing MSE loss and time.

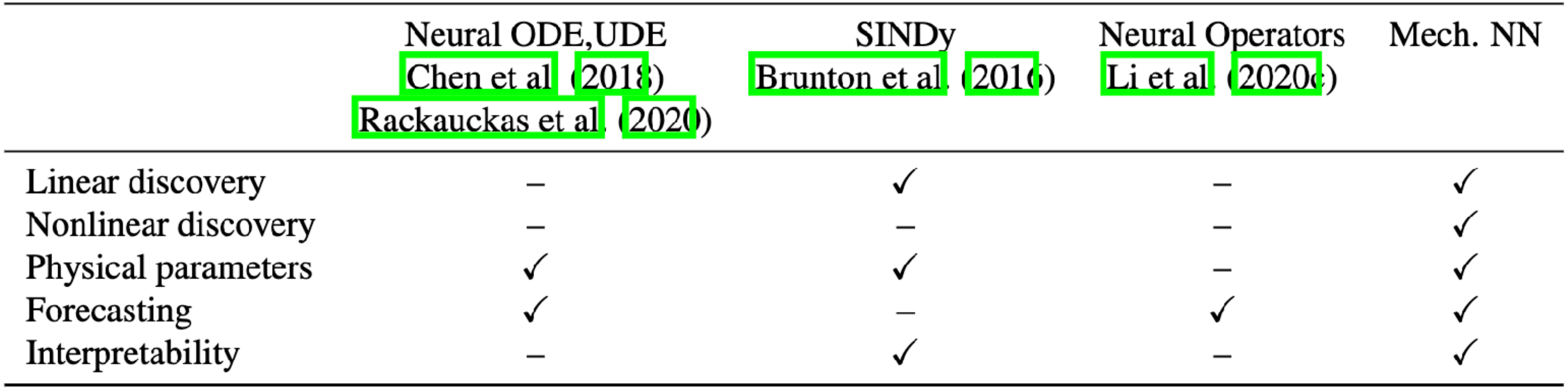
Steps	QP (seconds)	RK4 (seconds)	QP Loss	RK4 Loss
40	1.52	28.06	11.4	29.3
100	1.61	64.57	27.9	35.6
300	1.76	211.52	52	96.8
500	2.12	359.7	128	301
1000	3.68	666.69	292	589

https://github.com/alpz/mech-nn

PERVEZ, LOCATELLO, GAVVES, ONGOING

MECHANISTIC NEURAL NETWORKS

- Weaving in governing equations to an NN impacts lots of scientific applications
- With the same framework, we outperform per task specialized methods



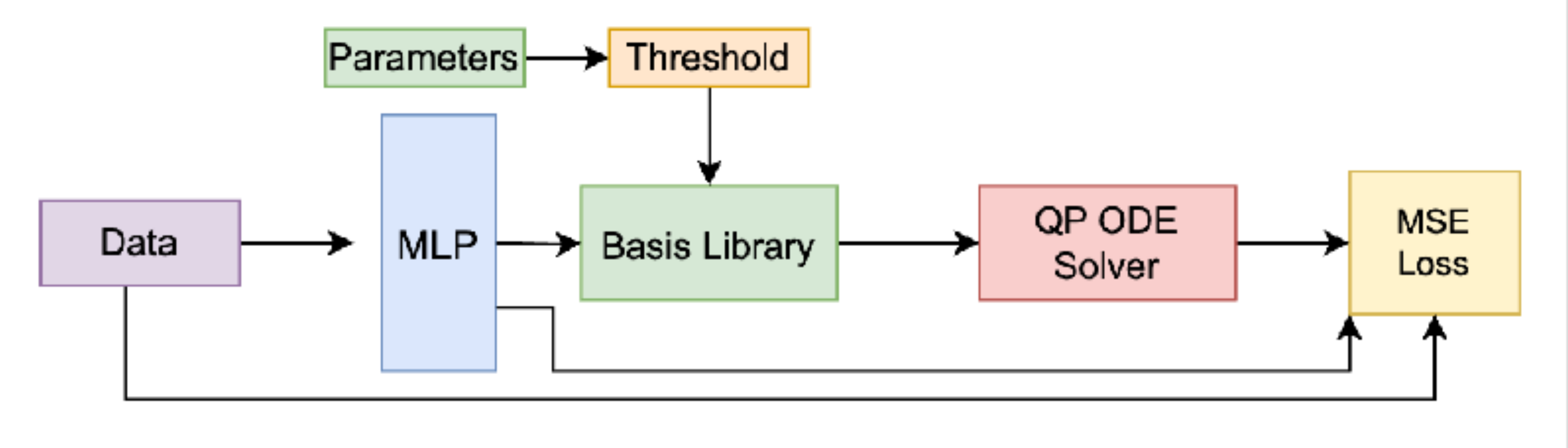
	Neural ODE,UDE Chen et al. (2018) Rackauckas et al. (2020)	SINDy Brunton et al. (2016)	Neural Operators Li et al. (2020c)	Mech. NN
Linear discovery	–	✓	–	✓
Nonlinear discovery	–	–	–	✓
Physical parameters	✓	✓	–	✓
Forecasting	✓	–	✓	✓
Interpretability	–	✓	–	✓

https://github.com/alpz/mech-nn

PERVEZ, LOCATELLO, GAVVES, ONGOING

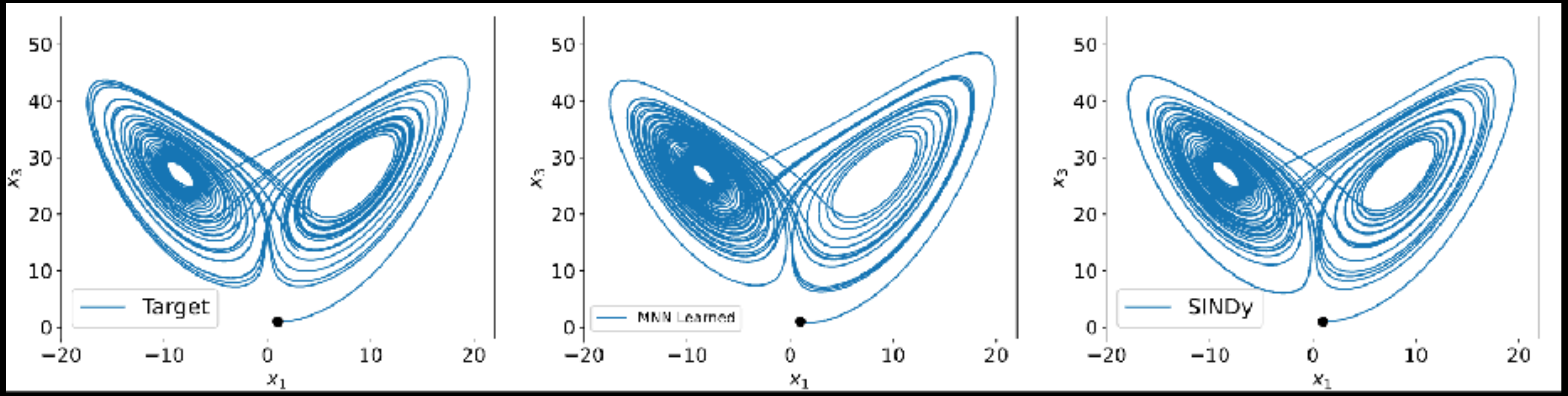
MECH NN FOR DISCOVERY

Solve ODEs of the form: $u' = g(\Theta\xi)$

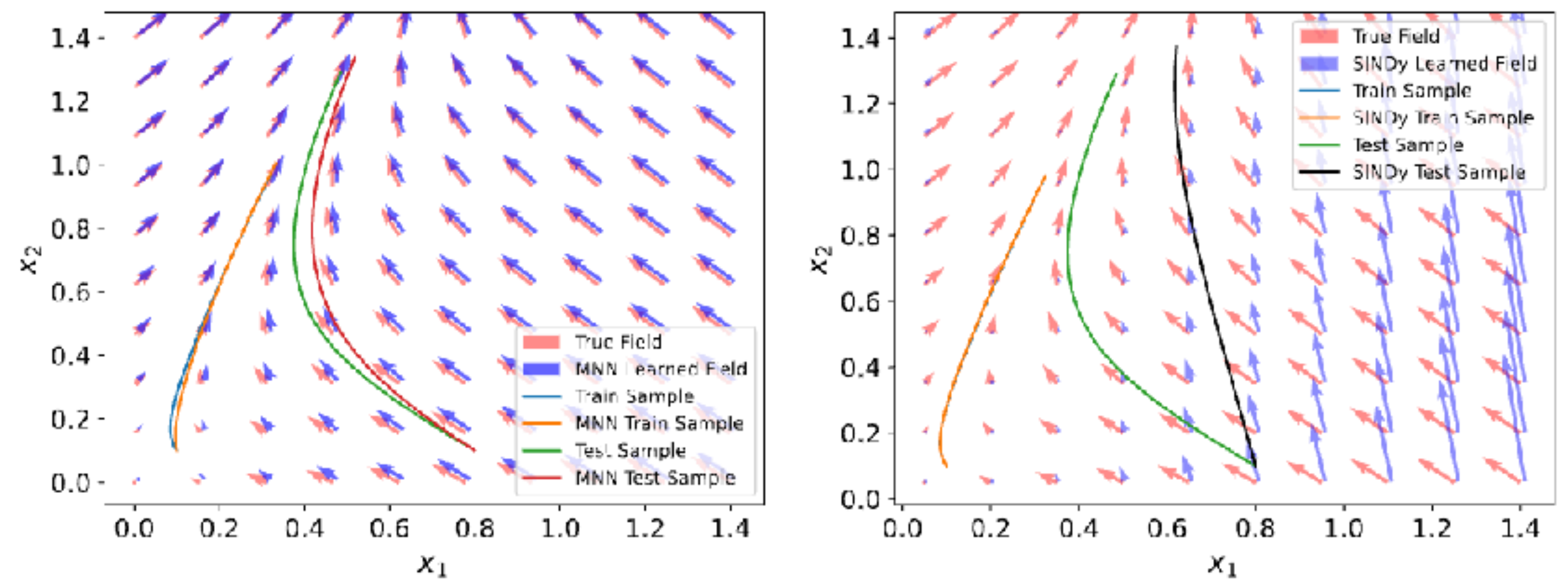


Parameters
Threshold
Data
MLP
Basis Library
QP ODE Solver
MSE Loss

Chaotic Lorenz



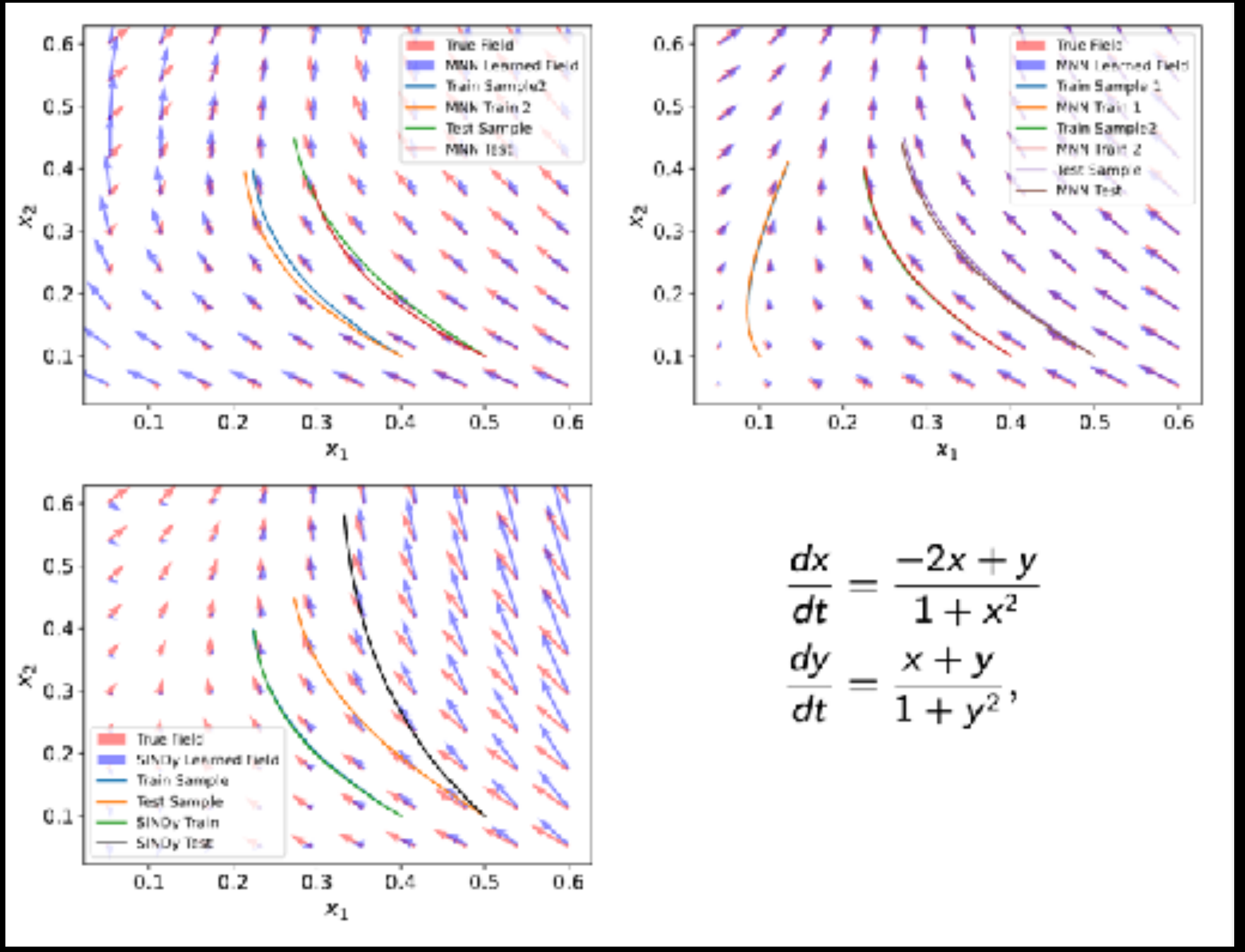
Nonlinear functions



$$\frac{dx}{dt} = \tanh(-2x + y)$$
$$\frac{dy}{dt} = \tanh(x + y)$$

Figure 2: Learned ODE vector fields for MNN and SINDy with non-linear tanh function of basis combination and training and test trajectories. Ground truth equation is on the right.

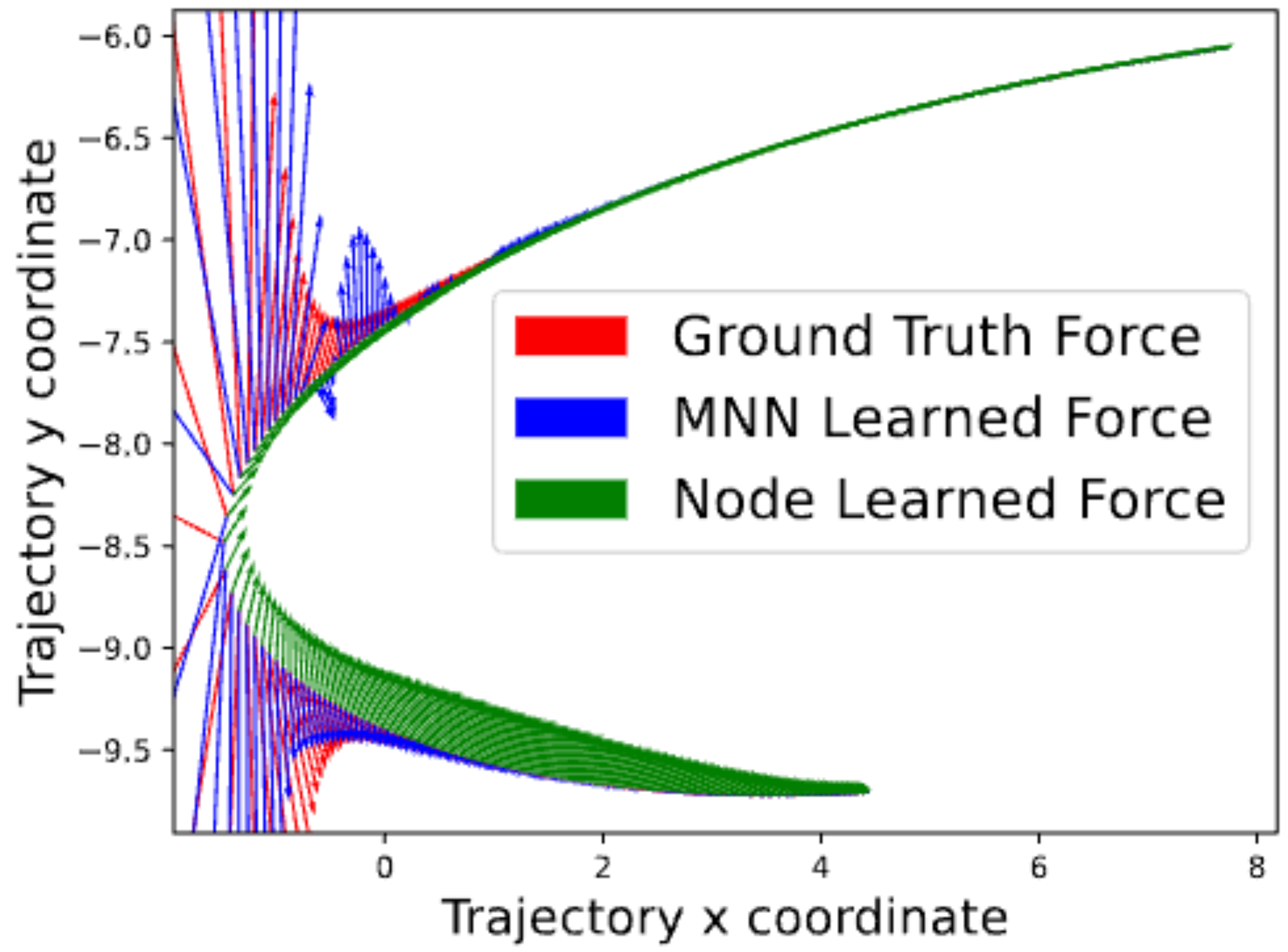
Rational functions

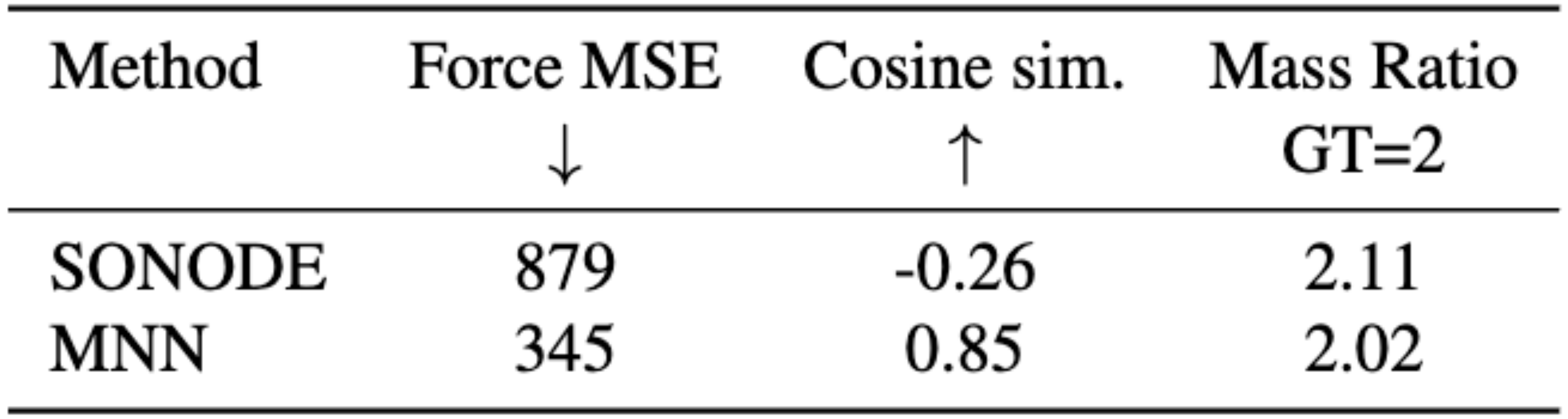


$$\frac{dx}{dt} = \frac{-2x + y}{1 + x^2}$$
$$\frac{dy}{dt} = \frac{x + y}{1 + y^2},$$

PERVEZ, LOCATELLO, GAVVES, ONGOING

MORE APPLICATIONS

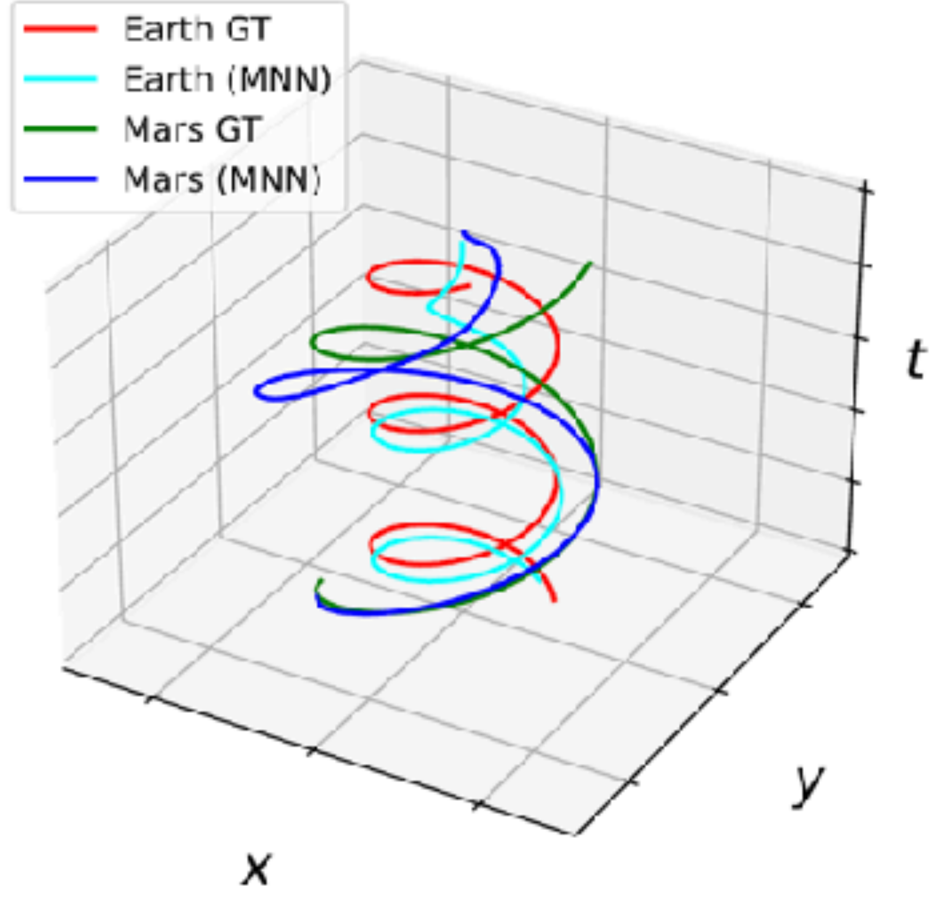


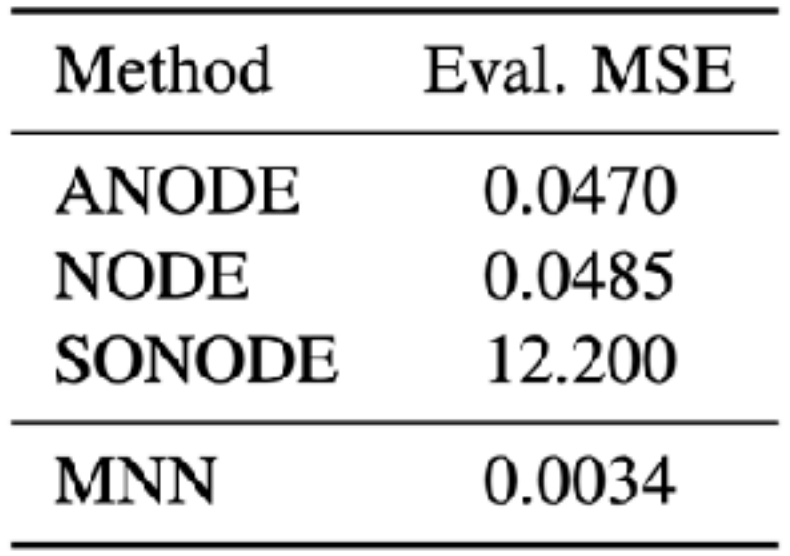


Method	Force MSE ↓	Cosine sim. ↑	Mass Ratio GT=2
SONODE	879	-0.26	2.11
MNN	345	0.85	2.02

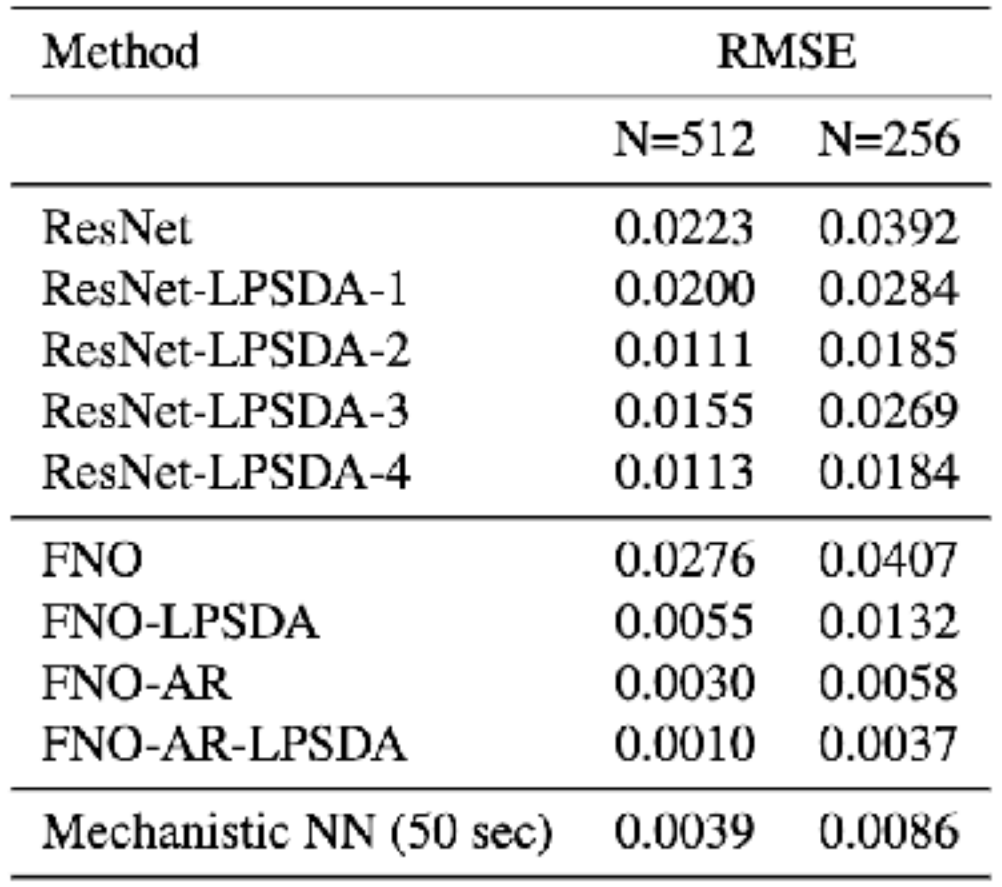
Discovering physical parameters
Mass ratio & force distribution 2-body problem

Forecasting
JPL Horizon planetary ephemerides

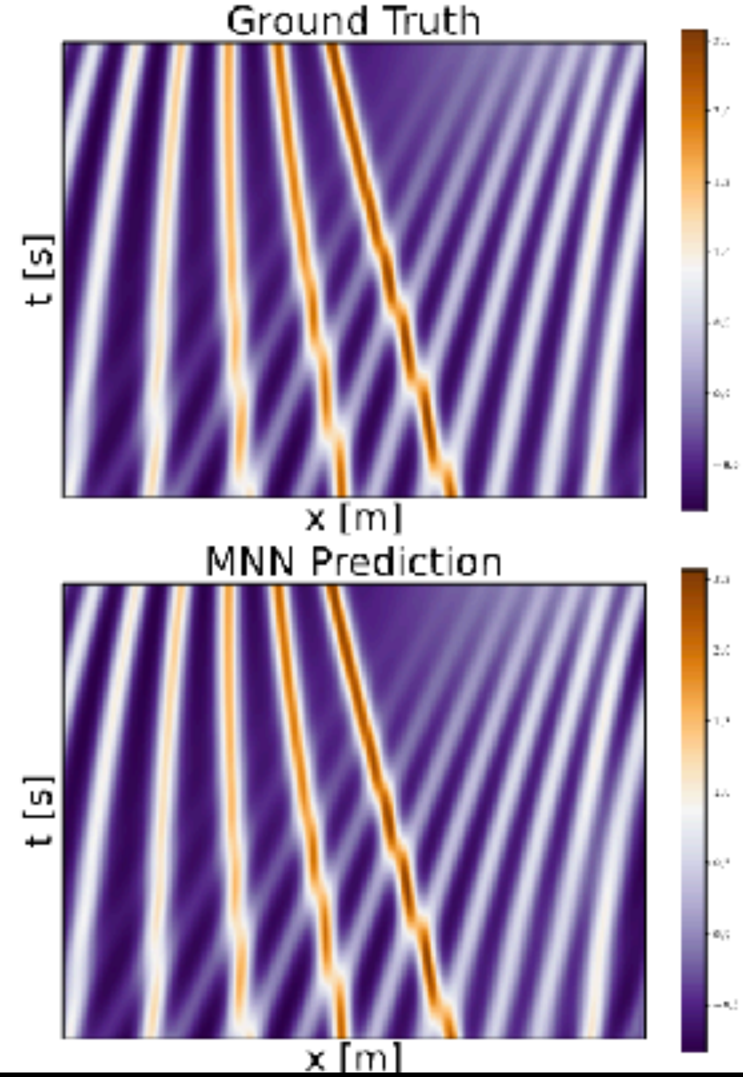




Method	Eval. MSE
ANODE	0.0470
NODE	0.0485
SONODE	12.200
MNN	0.0034



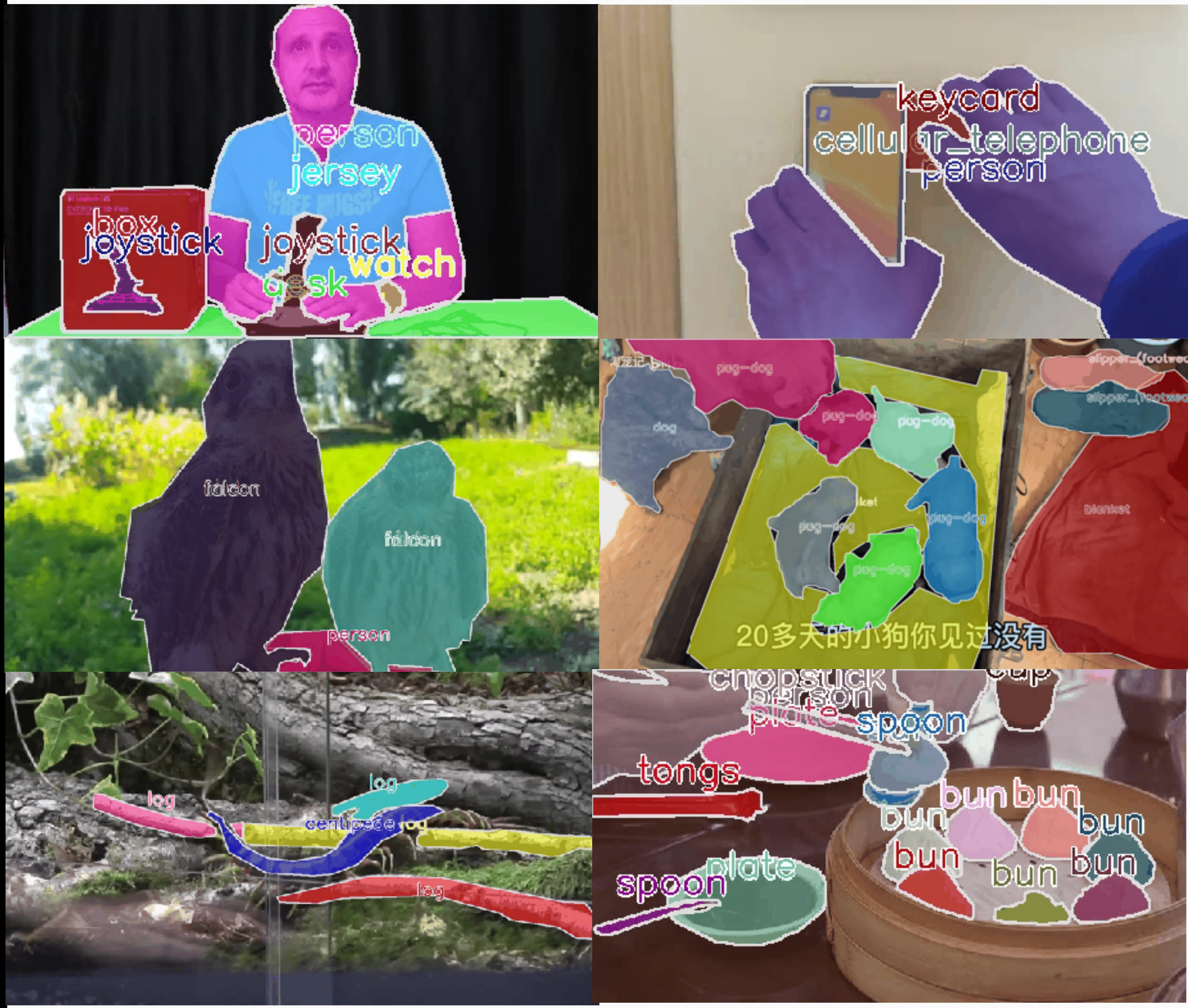
Method	RMSE	
	N=512	N=256
ResNet	0.0223	0.0392
ResNet-LPSDA-1	0.0200	0.0284
ResNet-LPSDA-2	0.0111	0.0185
ResNet-LPSDA-3	0.0155	0.0269
ResNet-LPSDA-4	0.0113	0.0184
FNO	0.0276	0.0407
FNO-LPSDA	0.0055	0.0132
FNO-AR	0.0030	0.0058
FNO-AR-LPSDA	0.0010	0.0037
Mechanistic NN (50 sec)	0.0039	0.0086



Neural PDE Solving
1d KdV and 2d Darcy Flow

WANG ET AL, CVPR, ICCV, 2022-24

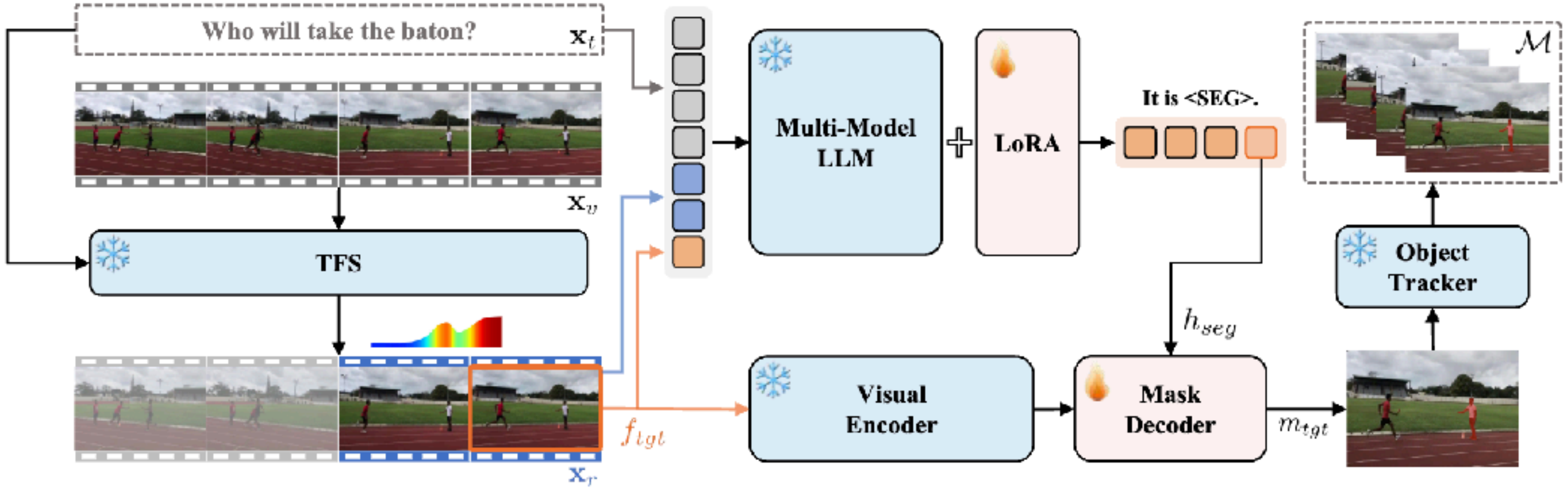
OPEN-WORLD INDUCTIVE BIASES

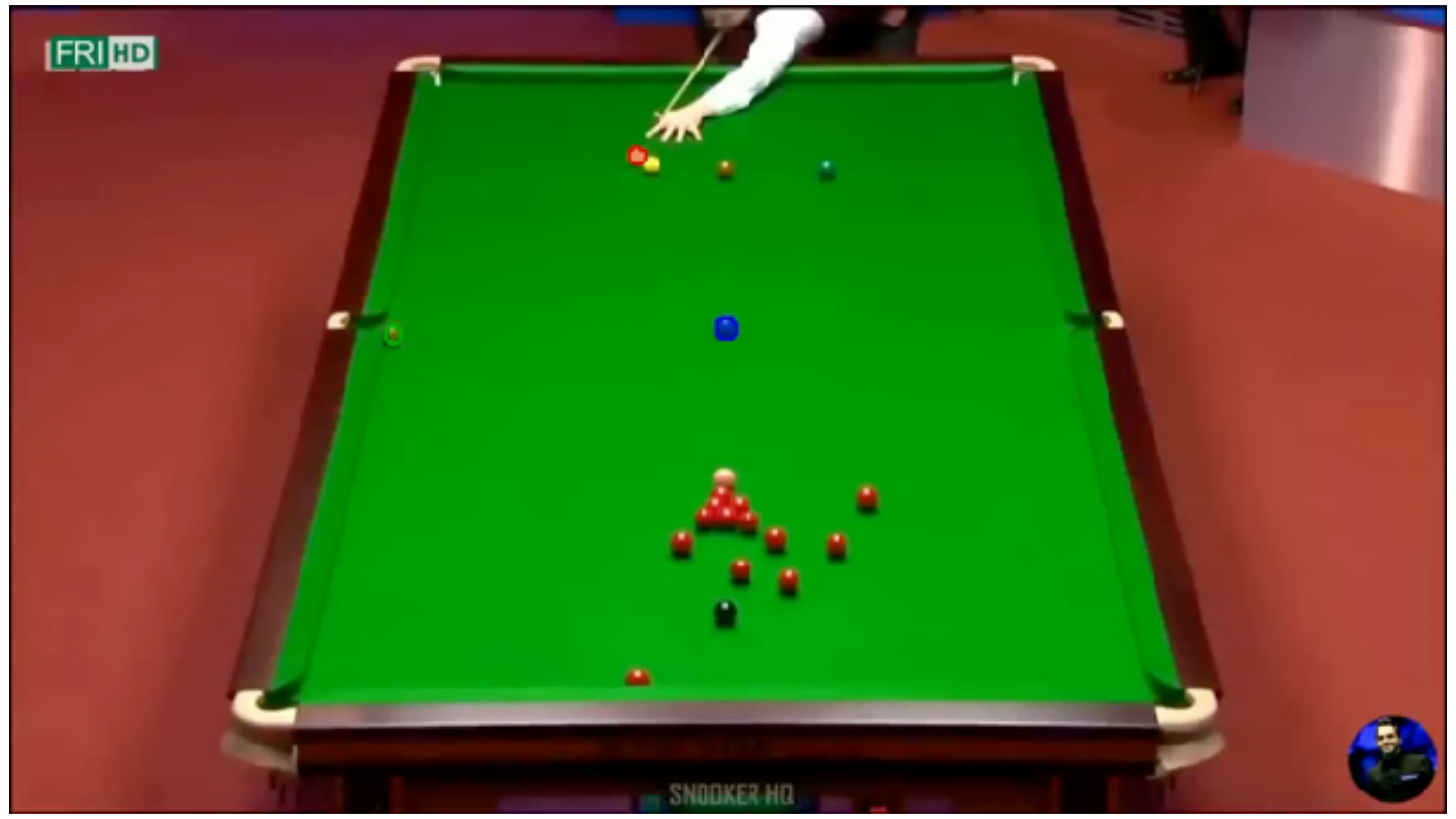


WANG, GAVVES ET AL, ONGOING

REASONING VIDEO INSTANCE SEGMENTATION

- Segment, classify, track by ‘commonsense reasoning’ w.r.t. the pixels and world knowledge
- Key idea: Scale-up sequence length in the input and relay to LLMs for ‘reasoning’
- Fantastic if we could pair this with formal reasoning





“Which ball should first be hit according to the rules?”

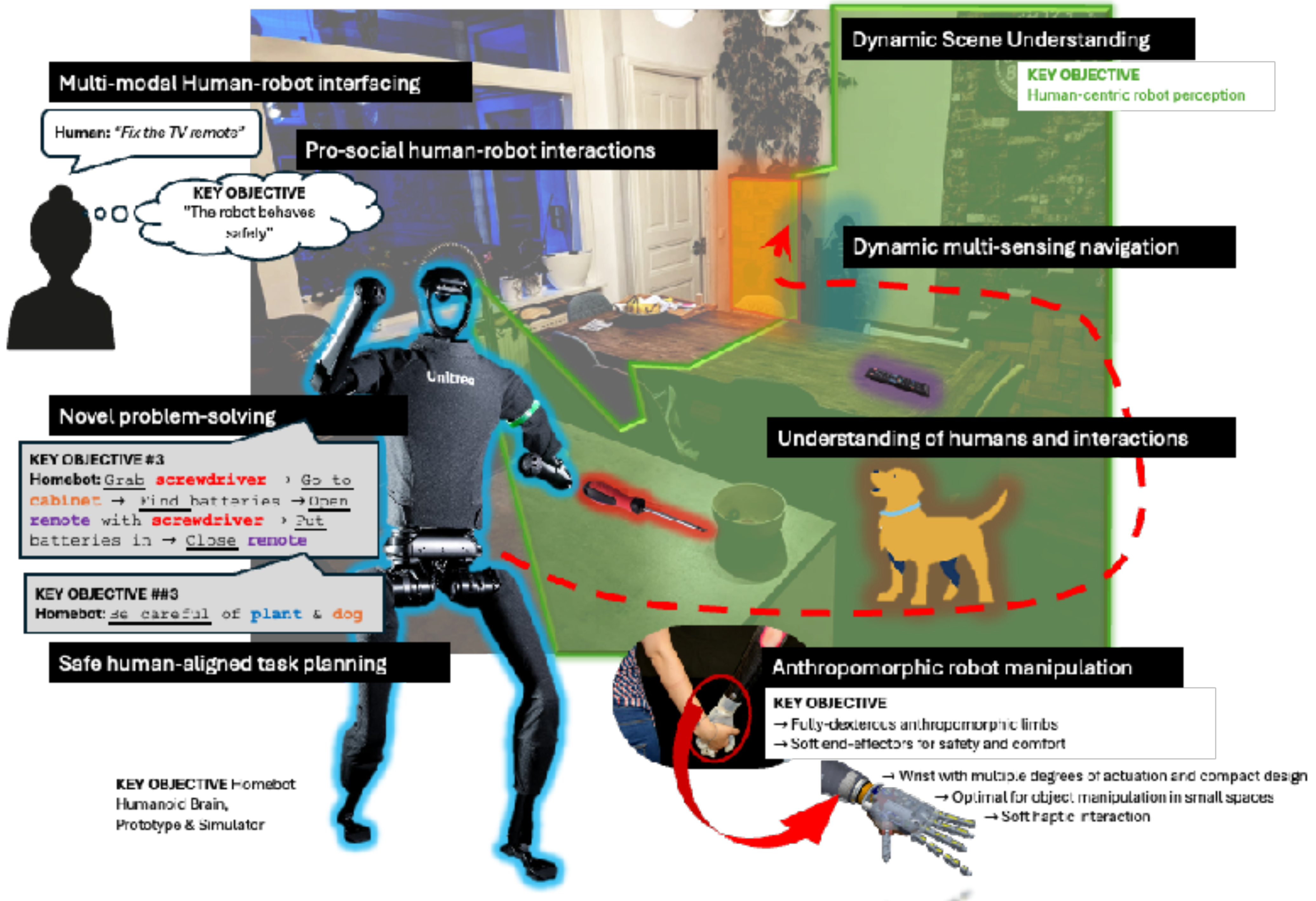
“Which ball is the target of this shot?”

“If this shot goes in, which ball is most likely to be hit next?”

PRO-SOCIAL & HUMAN-CENTRIC ROBOT AI

MY VISION & GOAL

System I Perception and Manipulation
System II Causal Learning for Robotics
'System III' Pro-social behavior


Multi-modal Human-robot interfacing
Human: "Fix the TV remote"
KEY OBJECTIVE "The robot behaves safely"
Pro-social human-robot interactions
Dynamic Scene Understanding
KEY OBJECTIVE Human-centric robot perception
Dynamic multi-sensing navigation
Unitree
Understanding of humans and interactions
Novel problem-solving
Safe human-aligned task planning
Anthropomorphic robot manipulation

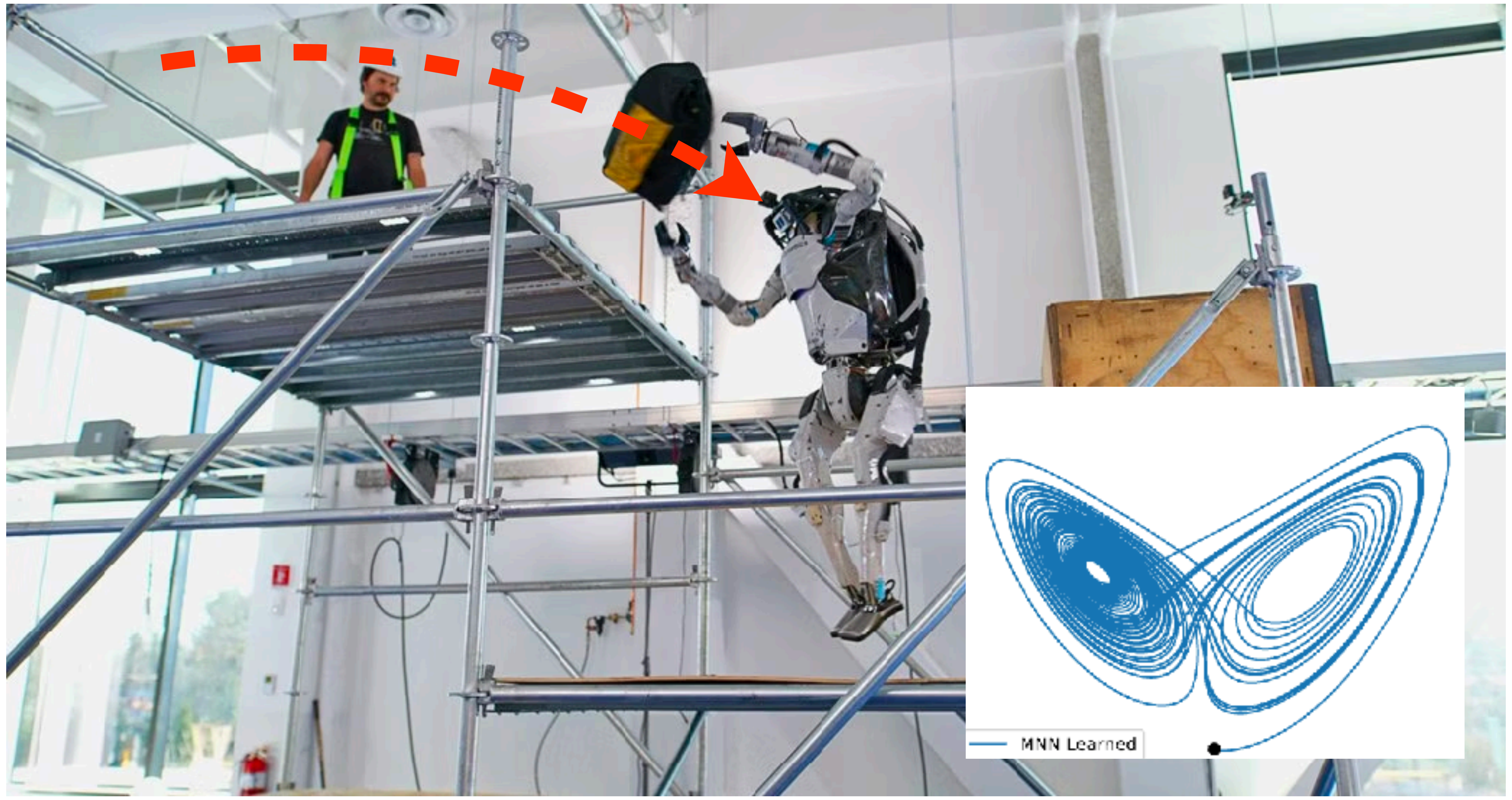
KEY OBJECTIVE #3
Homebot: Grab screwdriver → Go to cabinet → Find batteries → Open remote with screwdriver → Put batteries in → Close remote

KEY OBJECTIVE ##3
Homebot: Be careful of plant & dog

KEY OBJECTIVE
→ Fully-dexterous anthropomorphic limbs
→ Soft end-effectors for safety and comfort
→ Wrist with multiple degrees of actuation and compact design
→ Optimal for object manipulation in small spaces
→ Soft haptic interaction

KEY OBJECTIVE Homebot
Humanoid Brain,
Prototype & Simulator

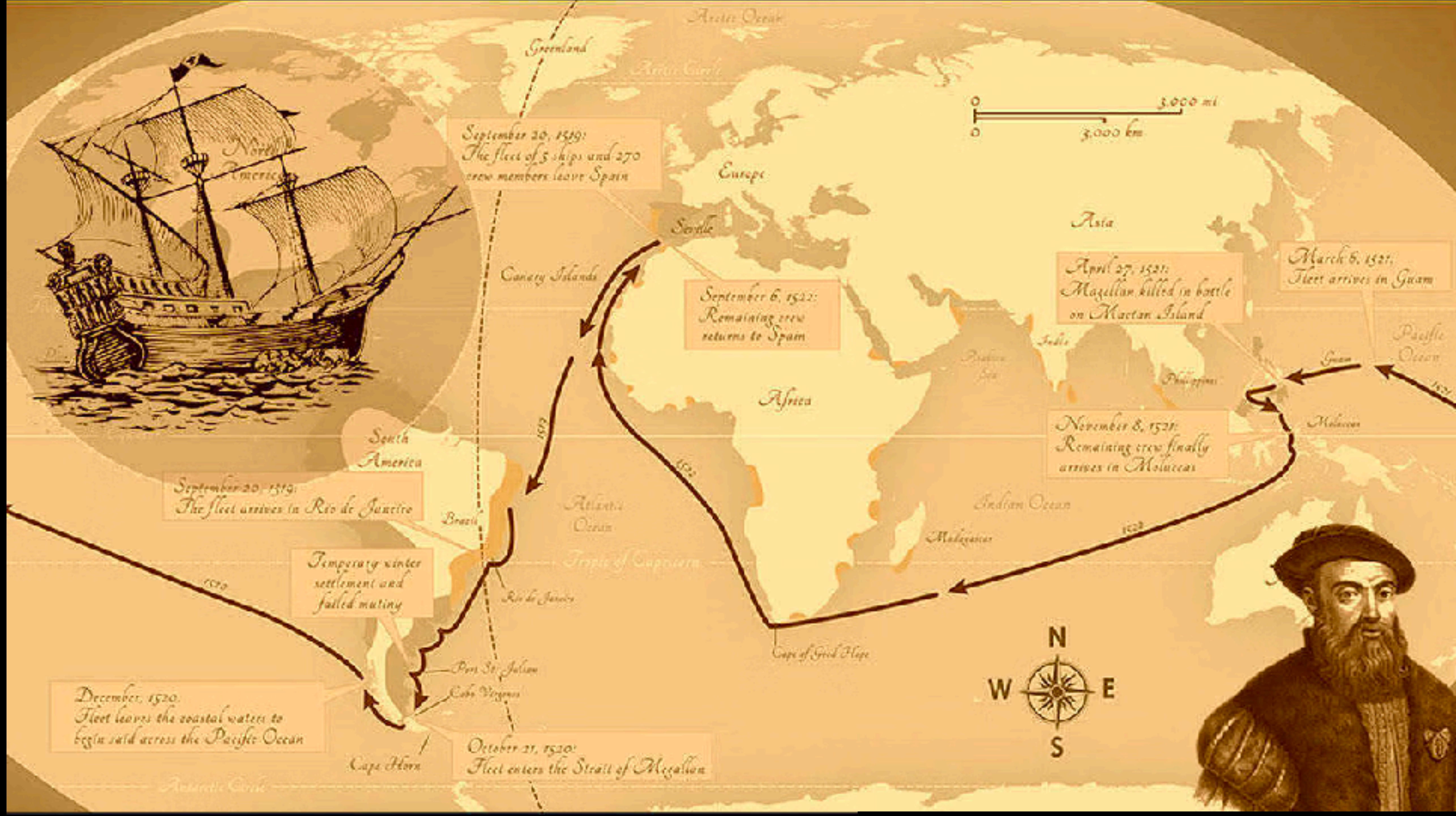
Reusable foundation 'world' dynamics from pixels
Neural+'Symbolic'+Equivariance Learning
Critical for true extrapolation



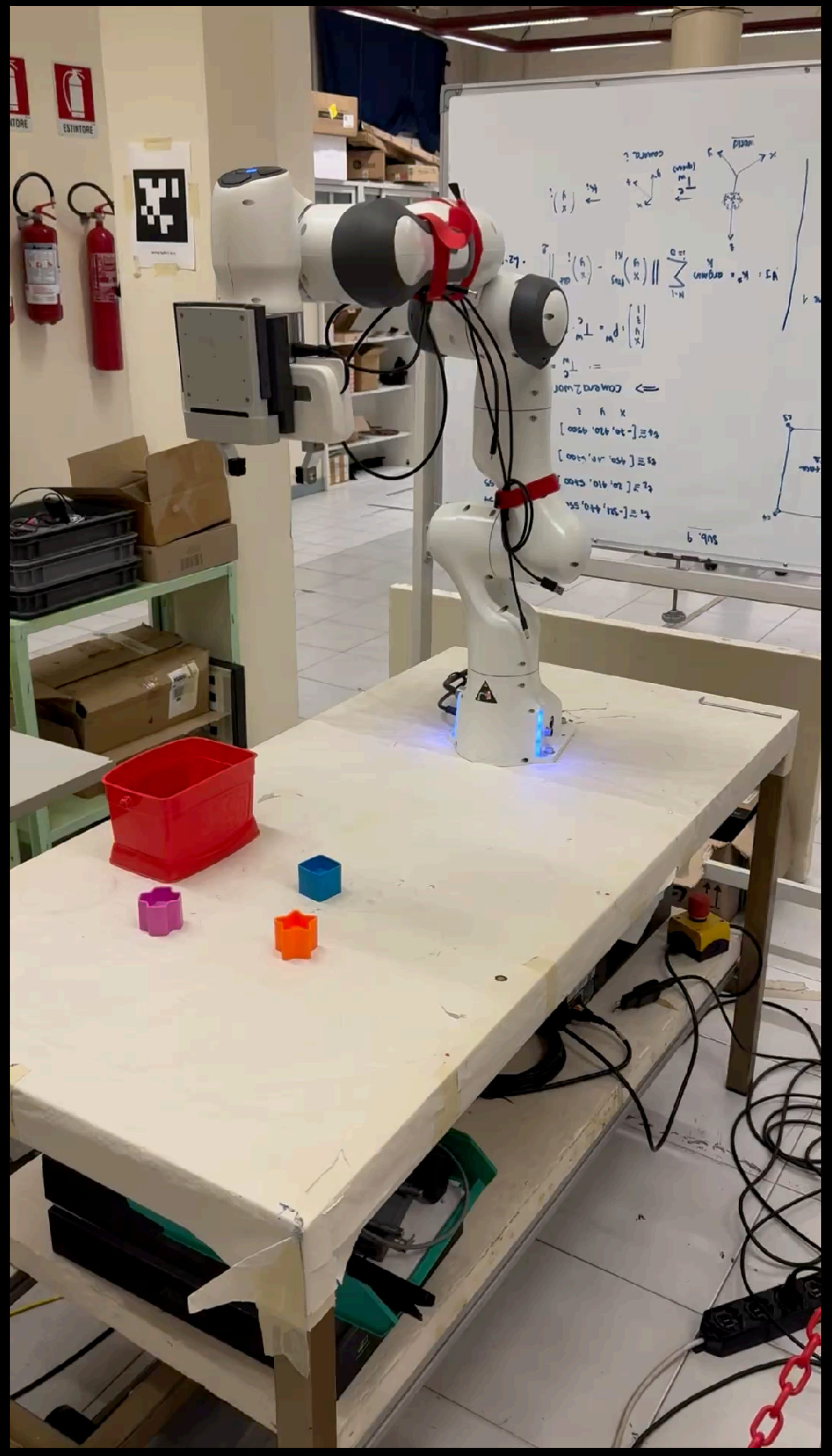
CVPR 2024 Workshop on Causal and Object-centric Representations for Robotics
Ideally repeated with NeurIPS/ICLR/ICML and CORL/ICRA/IROS

ACADEMIC AI IN THE YEARS OF CHATGPT

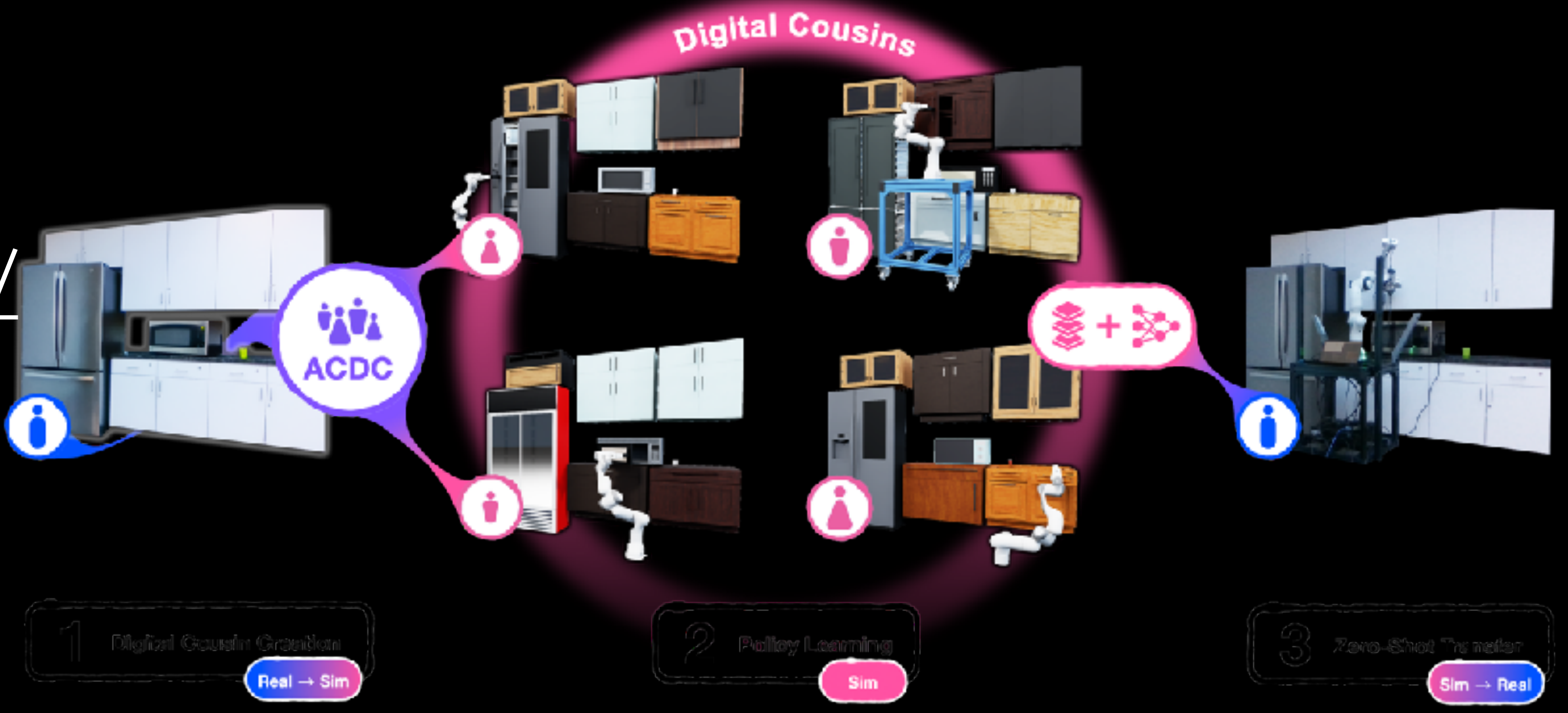
TO WRAP UP


September 20, 1519: The fleet of 5 ships and 270 crew members leave Spain
Europe
Asia
Seville
Canary Islands
September 6, 1522: Remaining crew returns to Spain
April 27, 1521: Magellan killed in battle on Mactan Island
March 6, 1521: Fleet arrives in Guam
Africa
South America
November 8, 1521: Remaining crew finally arrives in Moluccas
September 20, 1519: The fleet arrives in Rio de Janeiro
Atlantic Ocean
Indian Ocean
Temporary winter settlement and failed mutiny
December, 1520: Fleet leaves the coastal waters to begin sail across the Pacific Ocean
October 21, 1520: Fleet enters the Strait of Magellan
Cape Horn
Cape of Good Hope
N
W
E
S

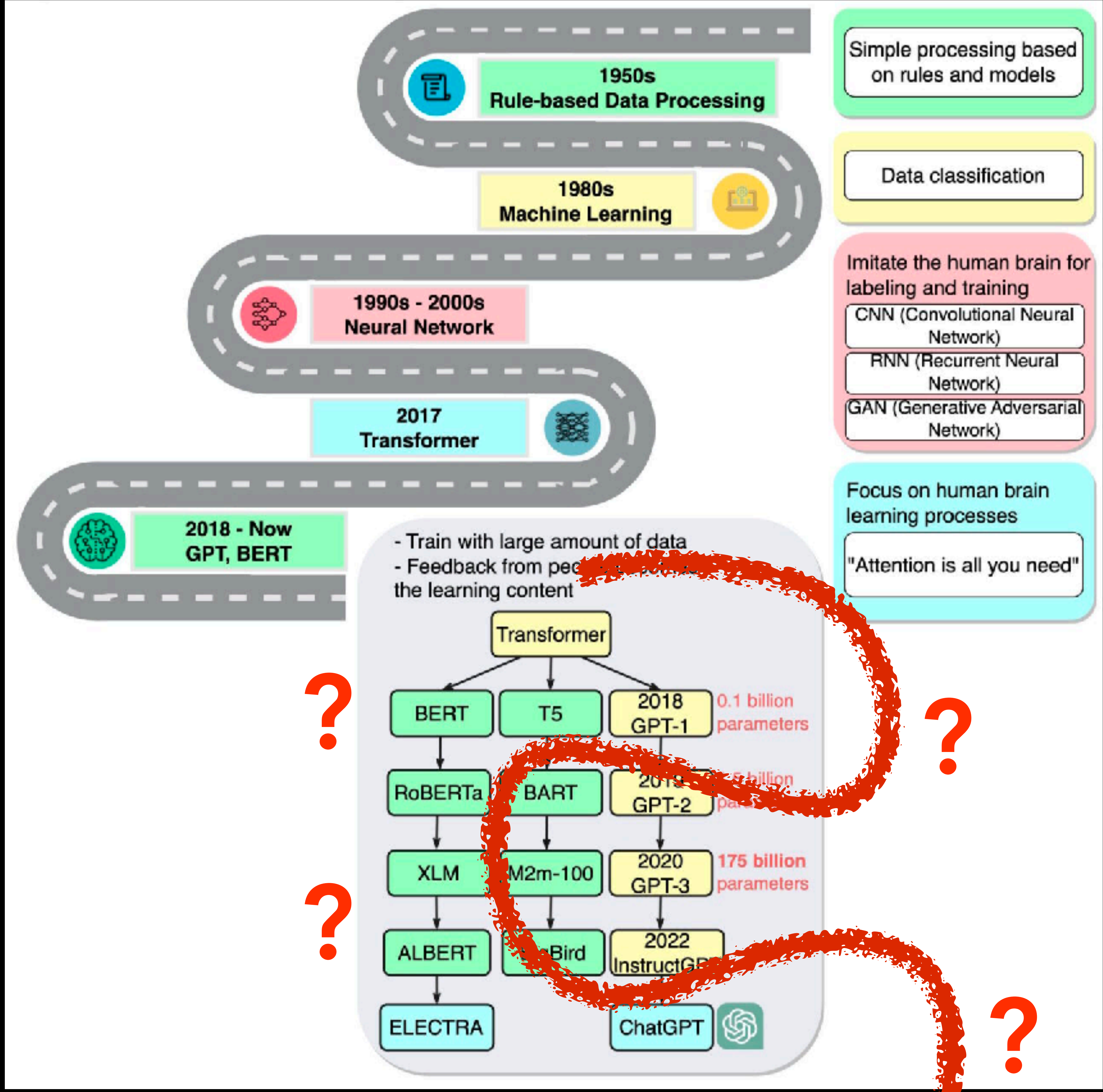
CONCLUSION



- Not yet end-to-end training, but all parts can be made differentiable
- “Physics & causal grounding”: great start toward fully neural robot world models
- And lots of exciting possibilities towards generalization and extrapolation
- We are in good company of Fei Fei Li: https://github.com/cremebrule/digital-cousins



CHEERS!


1950s
Rule-based Data Processing
Simple processing based on rules and models
1980s
Machine Learning
Data classification
1990s - 2000s
Neural Network
Imitate the human brain for labeling and training
CNN (Convolutional Neural Network)
RNN (Recurrent Neural Network)
GAN (Generative Adversarial Network)
2017
Transformer
Focus on human brain learning processes
"Attention is all you need"
2018 - Now
GPT, BERT
- Train with large amount of data
- Feedback from pee
the learning content
Transformer
BERT
T5
2018 GPT-1
0.1 billion parameters
RoBERTa
BART
GPT-2
billion
XLM
M2m-100
2020 GPT-3
175 billion parameters
ALBERT
Bird
2022 InstructG
ELECTRA
ChatGPT
?
?
?
?
?
?

Dynamics
Causal Learning
Open-World
Robots & Systems

egavves@uva.nl
@egavves