

Building Physical Agents via Structured Representations

Jiajun Wu



Stanford University

We want robots in the real world with humans.

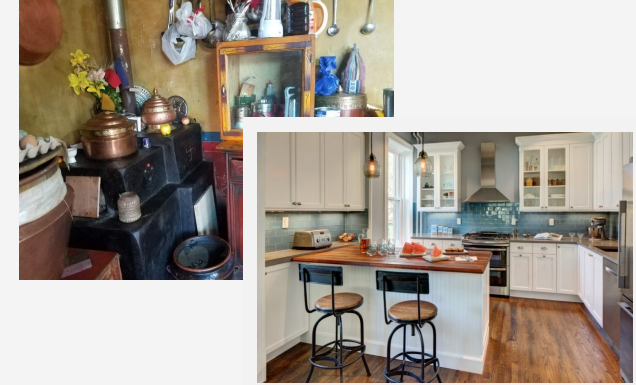
Ecologically complex



Dynamic & Uncertain



Diverse



Interactive & Social



Multitask



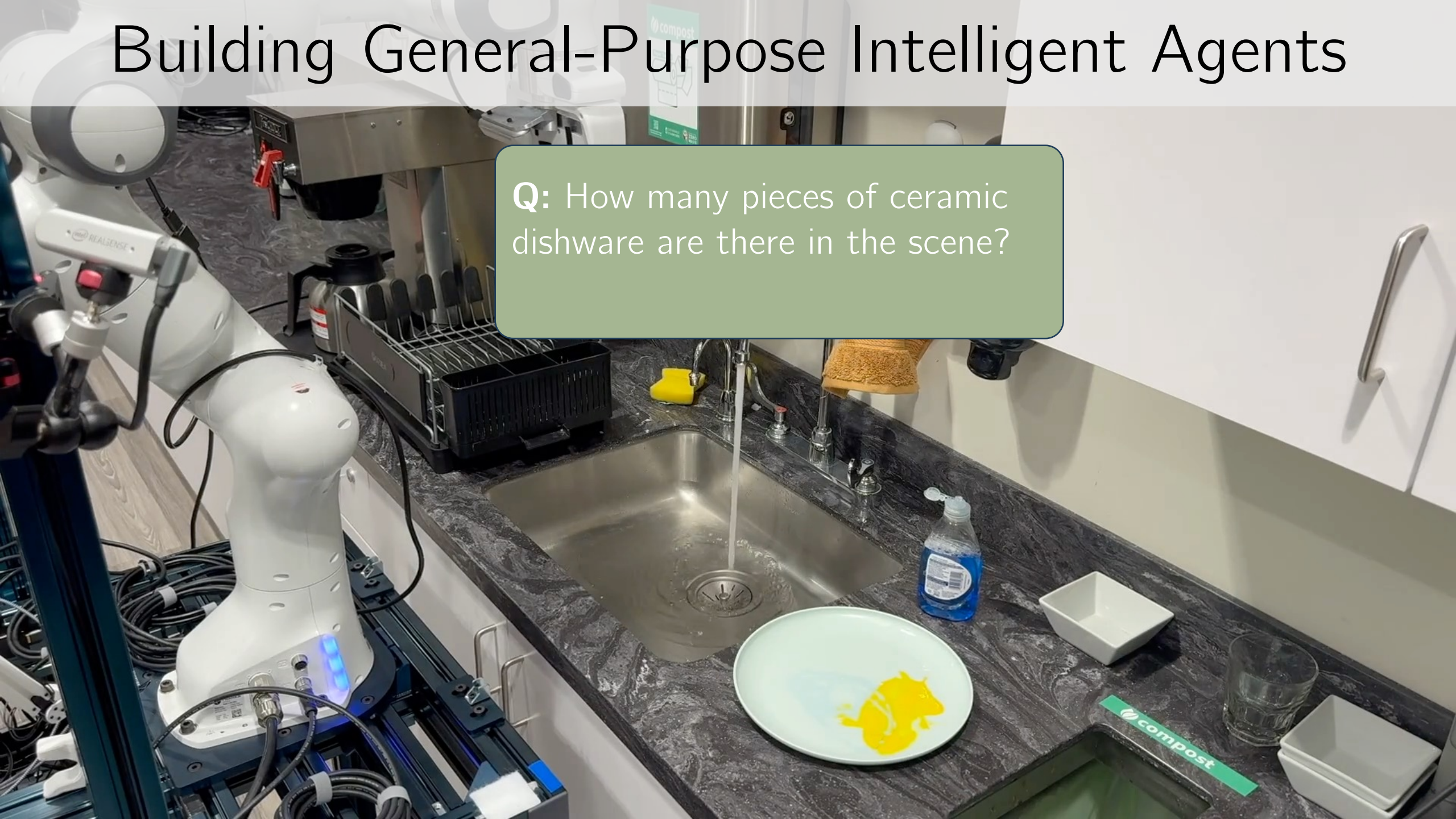
What do we need beyond foundation models?

Building General-Purpose Intelligent Agents



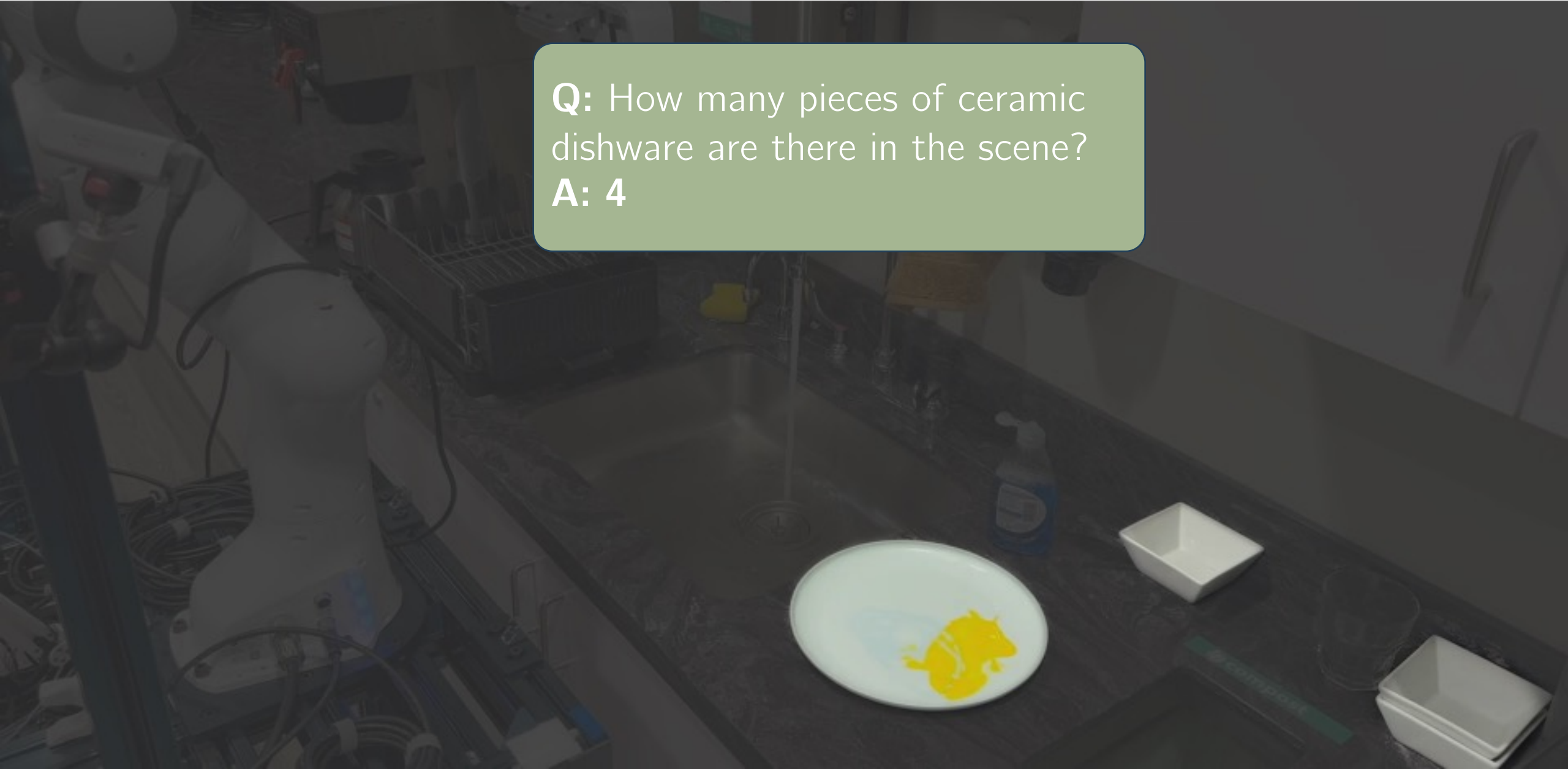
Building General-Purpose Intelligent Agents

Q: How many pieces of ceramic dishware are there in the scene?



Building General-Purpose Intelligent Agents

Q: How many pieces of ceramic dishware are there in the scene?
A: 4



Building General-Purpose Intelligent Agents

Instruction: Can you help me wash that please?



Building General-Purpose Intelligent Agents

A robotic arm is positioned in a kitchen sink. On the counter next to the sink is a white plate with a yellow and blue food item on it. A green text box is overlaid on the right side of the image.

Instruction: Can you help me wash that please?

Building General-Purpose Intelligent Agents



Generalize across objects, environments, and goals



(Autonomous 2x)

Adaptive to Human Actions



(Autonomous 4x)

Generalizes to Different Initial States



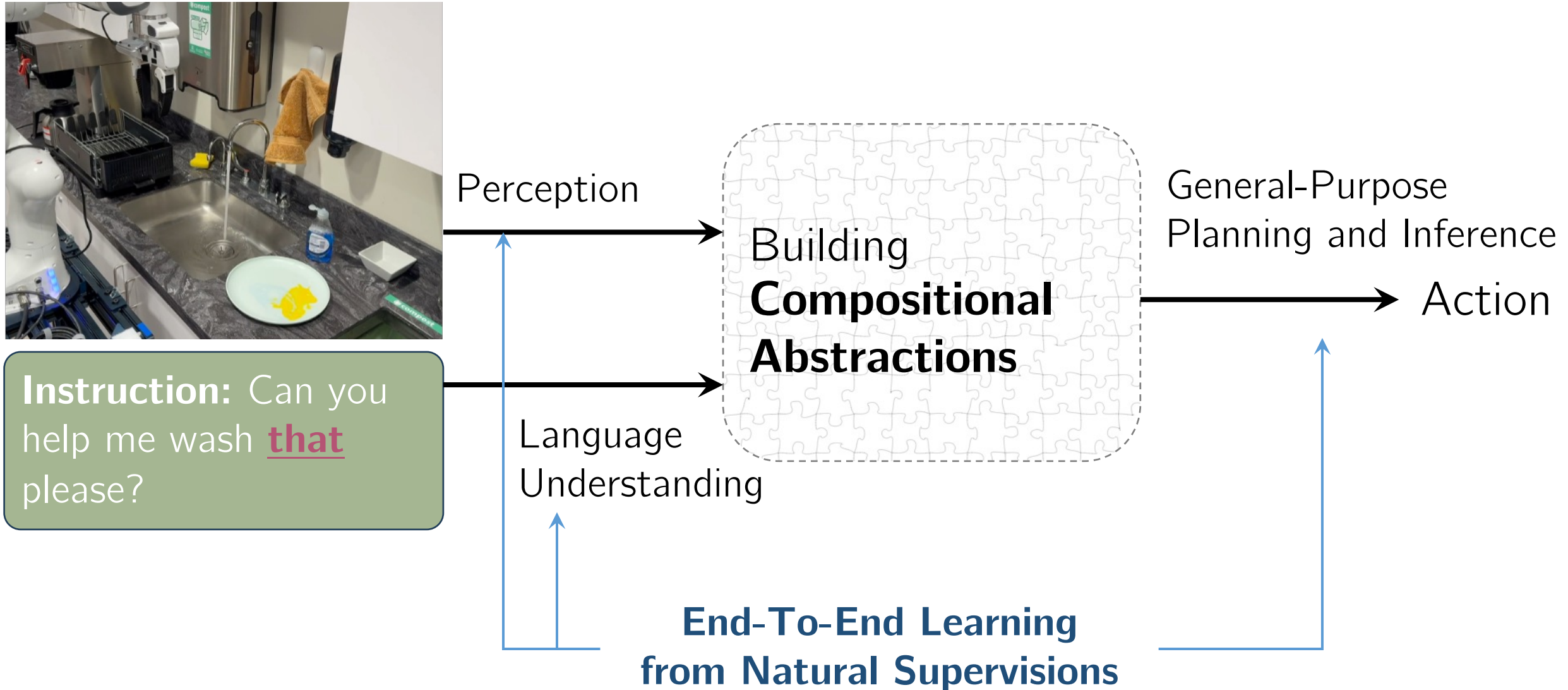
(Autonomous 4x)

Adaptive to Updated Observations



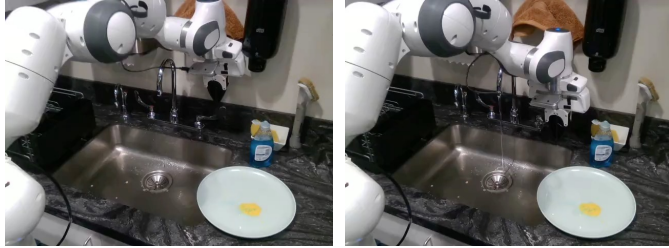
(Autonomous 4x)

Learning Compositional Abstractions for Inference



Learning Compositional World Models from Language

“turn on faucet”



“grasp plate and wash and place on dish rack”

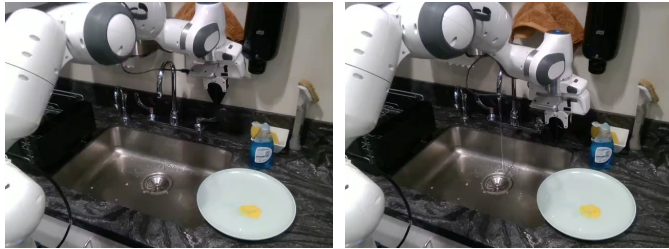


“wipe ...”



Learning Compositional World Models from Language

"turn on faucet"



"grasp plate and wash and place on dish rack"



"wipe ..."



LLM-Based
Code Generation

Learning Algorithms

Symbolic
Verification

Neuro-
Symbolic
World
Models

```
feature dirtiness-degree(o: object)
  extractor: 
feature on(a: object, b: object)
  extractor: 
```

Symbolic Types
Define Compositionality

```
behavior wash (obj: Object, faucet: Object):
  body:
    achieve holding(obj)
    achieve faucet-on(faucet)
    do  sensorimotor-policy...
  eff:
    if has-soap(obj): clean(obj) = True
```

Neural Modules
Handle Perception and Control

Learning State Features from Demonstrations



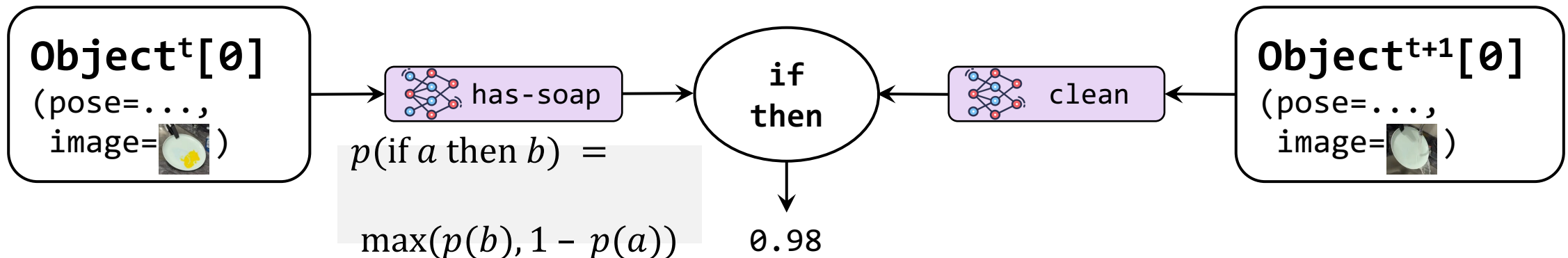
.....



(Before)

```
if has-soap(obj):  
  clean(obj) = True
```

(After)



Learning State Features from Demonstrations



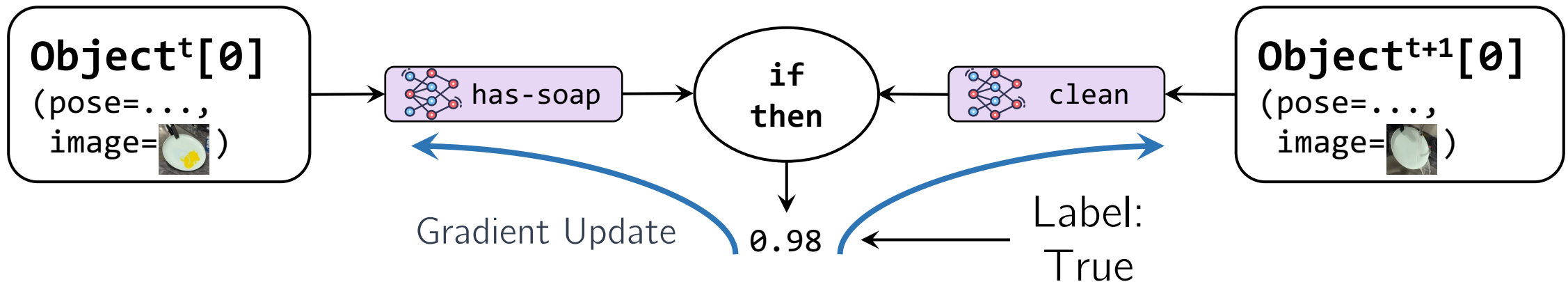
.....



(Before)

```
if has-soap(obj):  
  clean(obj) = True
```

(After)



Compositional Abstractions Enable Generalization

Compositional abstractions for *states, actions, and plans* enable data-efficient learning, faster planning, and better generalization.

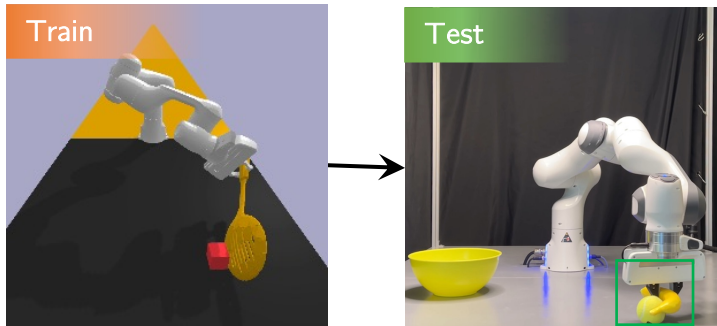


Weiyu Liu

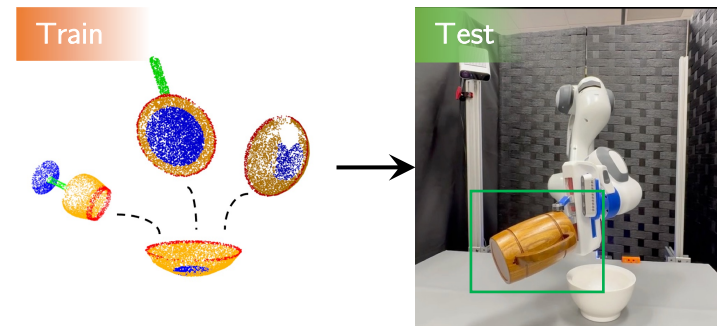
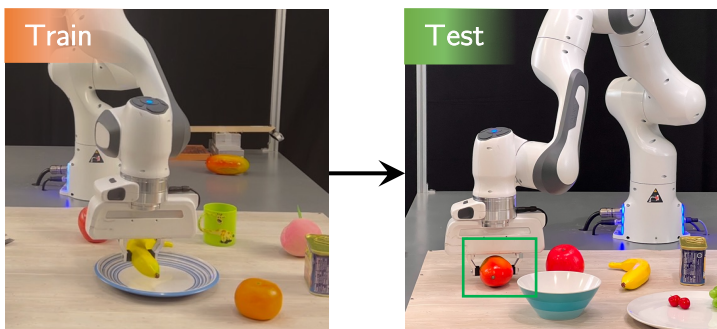


Jiayuan Mao

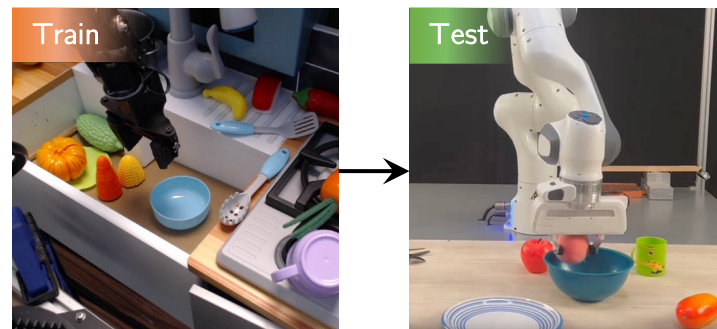
Generalization to Novel Objects



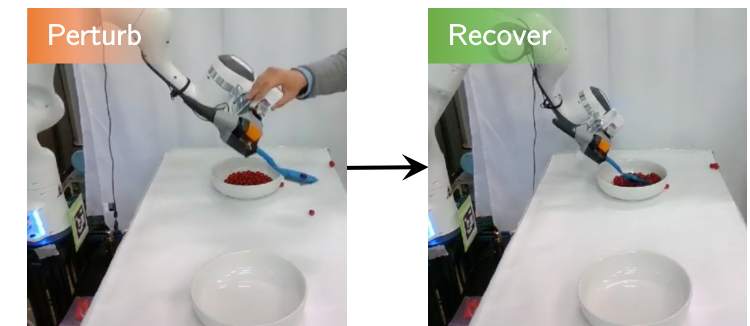
Generalization to Novel Words



Generalization to Novel Embodiments



Generalization to Novel States



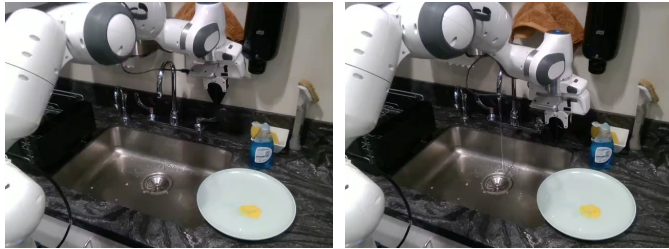
Interpretation of Under-Specified Goals

Set up a table for my breakfast.



Limitation: BLADE discovers and grounds skills in 2D.

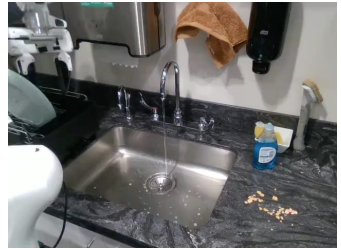
"turn on faucet"



"grasp plate and wash and place on dish rack"



"wipe ..."



LLM-Based
Code Generation

Learning Algorithms

Symbolic
Verification

Neuro-
Symbolic
World
Models

```
feature dirtiness-degree(o: object)
  extractor: 
feature on(a: object, b: object)
  extractor: 
```

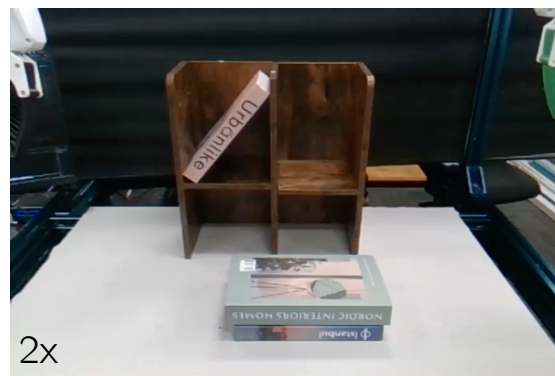
Symbolic Types
Define Compositionality

```
behavior wash (obj: Object, faucet: Object):
  body:
    achieve holding(obj)
    achieve faucet-on(faucet)
    do  sensorimotor-policy...
  eff:
    if has-soap(obj): clean(obj) = True
```

Neural Modules
Handle Perception and Control

Discovering Spatial-Temporal Skills Using FMs

Language Goal:
"Place book on shelf"



Video
Language
Model



Move Grasp: Green Book



Adjust: Pink Book



Transit Place: Green Book, Shelf

Detection & Segmentation

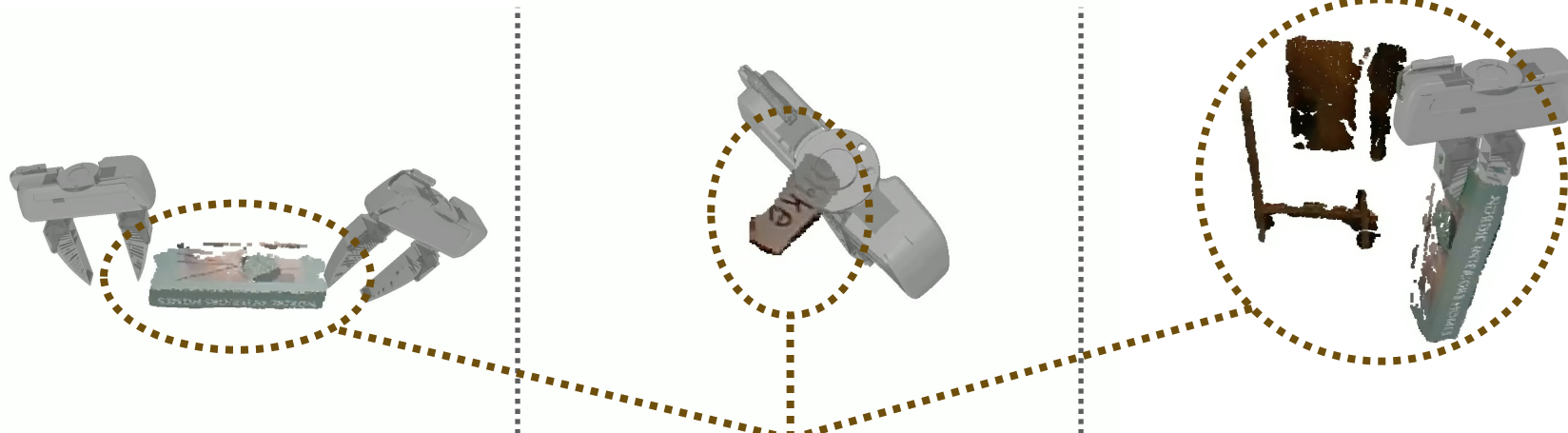
Relevant Entity Point Cloud & Demonstrated Actions

Bookshelf



Book

Book



Spatial Structure

Discovering Spatial-Temporal Skills Using FMs



Demo 1



Demo 2



Demo 3

...

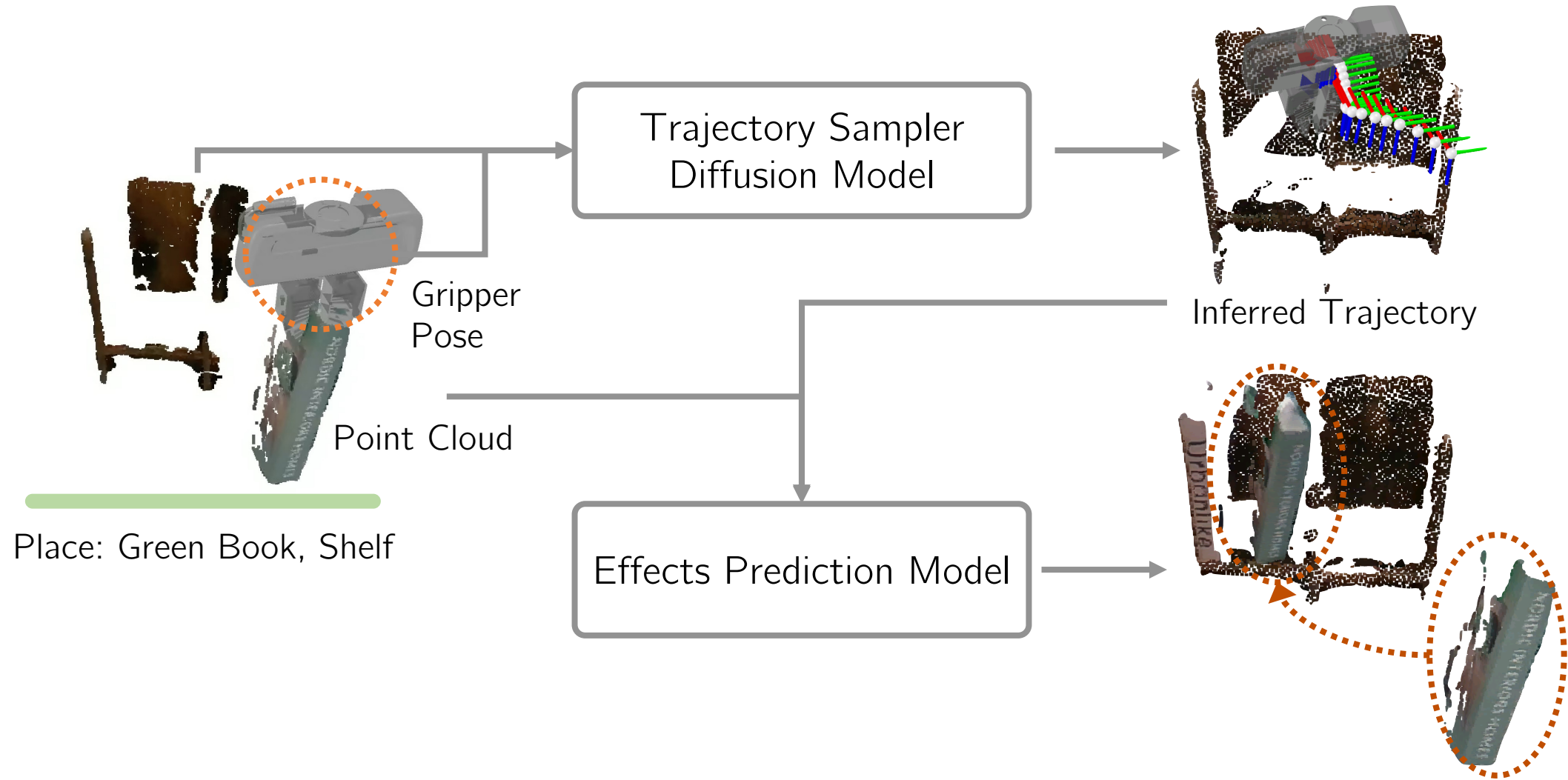


Demonstrations for Skills

...  ...

Place: Green Book, Shelf

Learning Composable Skills: Two Models Per Skill



Planning with Learned Skills



Planning with Learned Skills

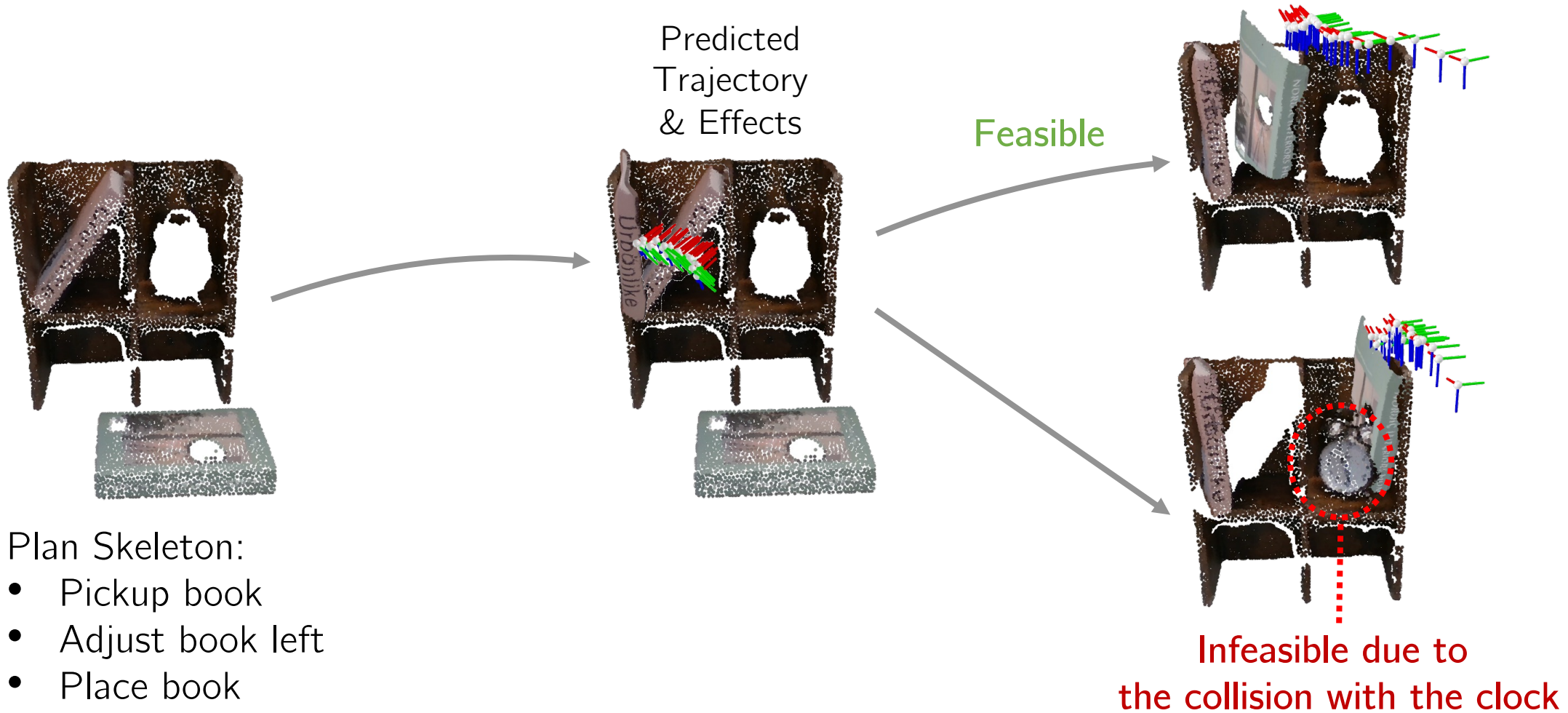


Planning with Learned Skills



Neil Nie

Weiyu Liu



Planning with Learned Skills



Generalization: New Geometric Constraints

Training Data: A Common Set of 10 Demos

3 Steps: Grasp; Adjust; Place



Generalization: Longer Task Horizon

Training Data: A Common Set of 10 Demos

3 Steps: Grasp; Adjust; Place



5-Step Task: “Place two books on shelf”



6-Step Task: “Place two books on shelf”

Generalization: Longer Task Horizon

Training Data: A Common Set of 10 Demos Per Peg

2 Steps: Grasp Mug; Hang Mug



8-Step Task: “Hang four mugs on mug tree”



6-Step Task: “Hang three mugs on mug tree”

Generalization: Longer Task Horizon

Training Data: A Common Set of 10 Demos

5 Steps: Turn on Faucet; Grasp Plate; Wash Plate; Place Plate; Turn off Faucet



8-Step Task: "Wash, and put away two plates"

8x

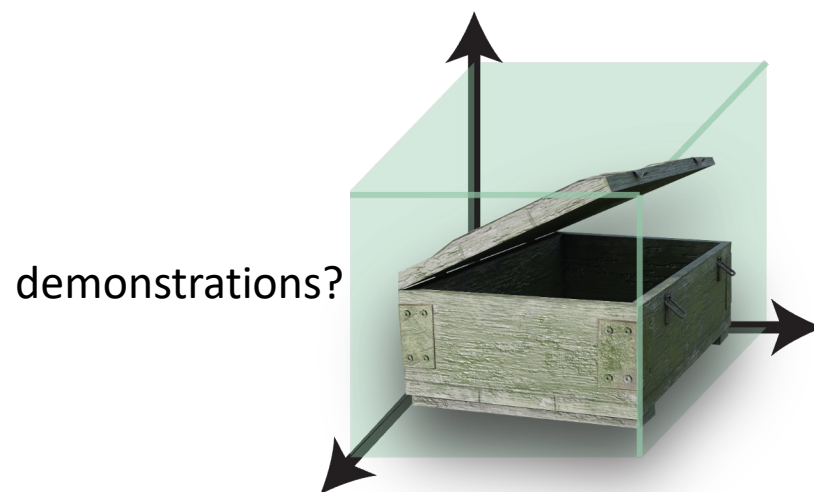
What if we don't have video demonstrations?



Neural Object Kinematics (NeuROK)



Chen Geng

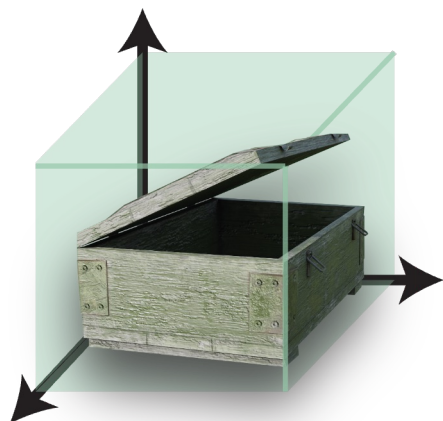


Input: Static 3D Object

Neural Object Kinematics (NeuROK)



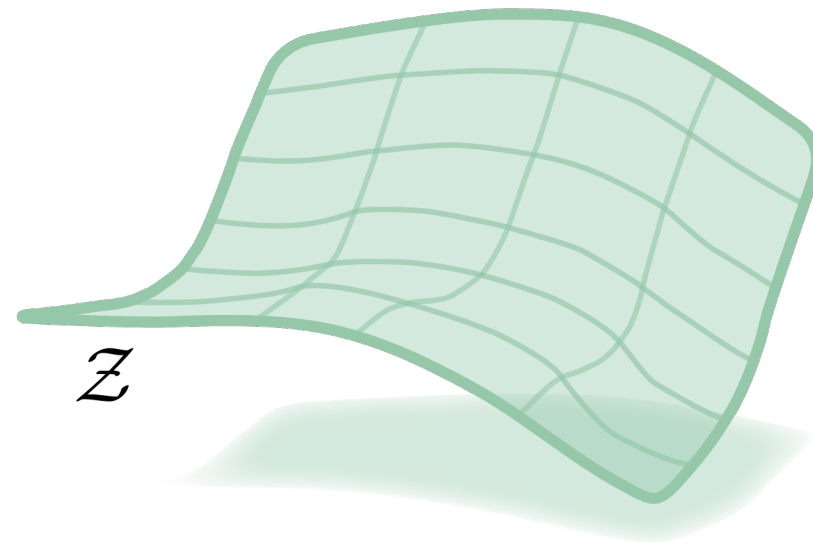
Chen Geng



Input: Static 3D Object



Feed-Forward
Prediction

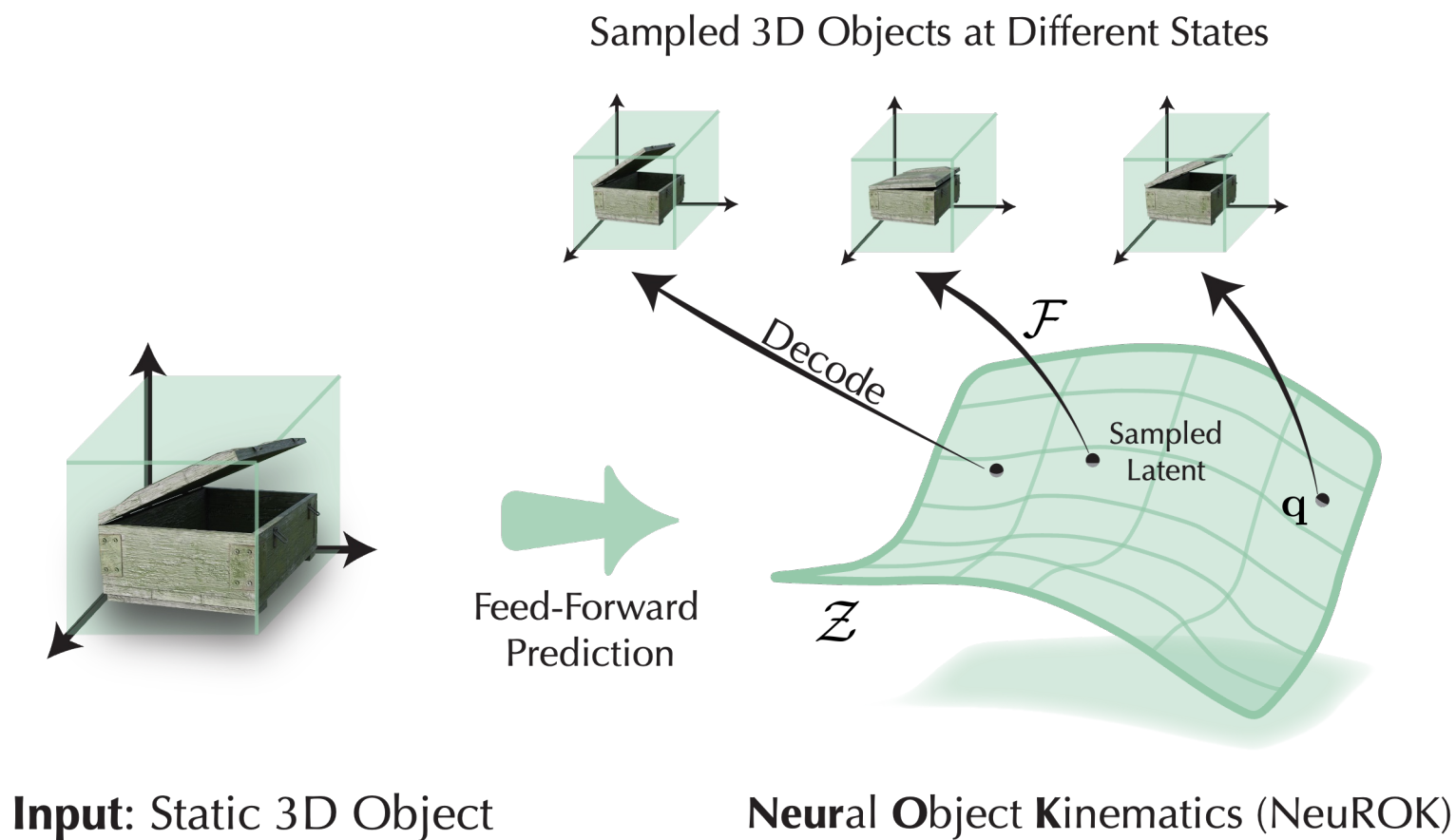


Neural Object Kinematics (NeuROK)

Neural Object Kinematics (NeuROK)



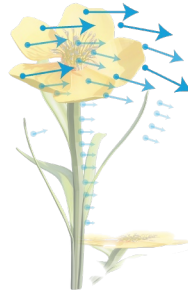
Chen Geng



Encoder-Decoder Training Framework

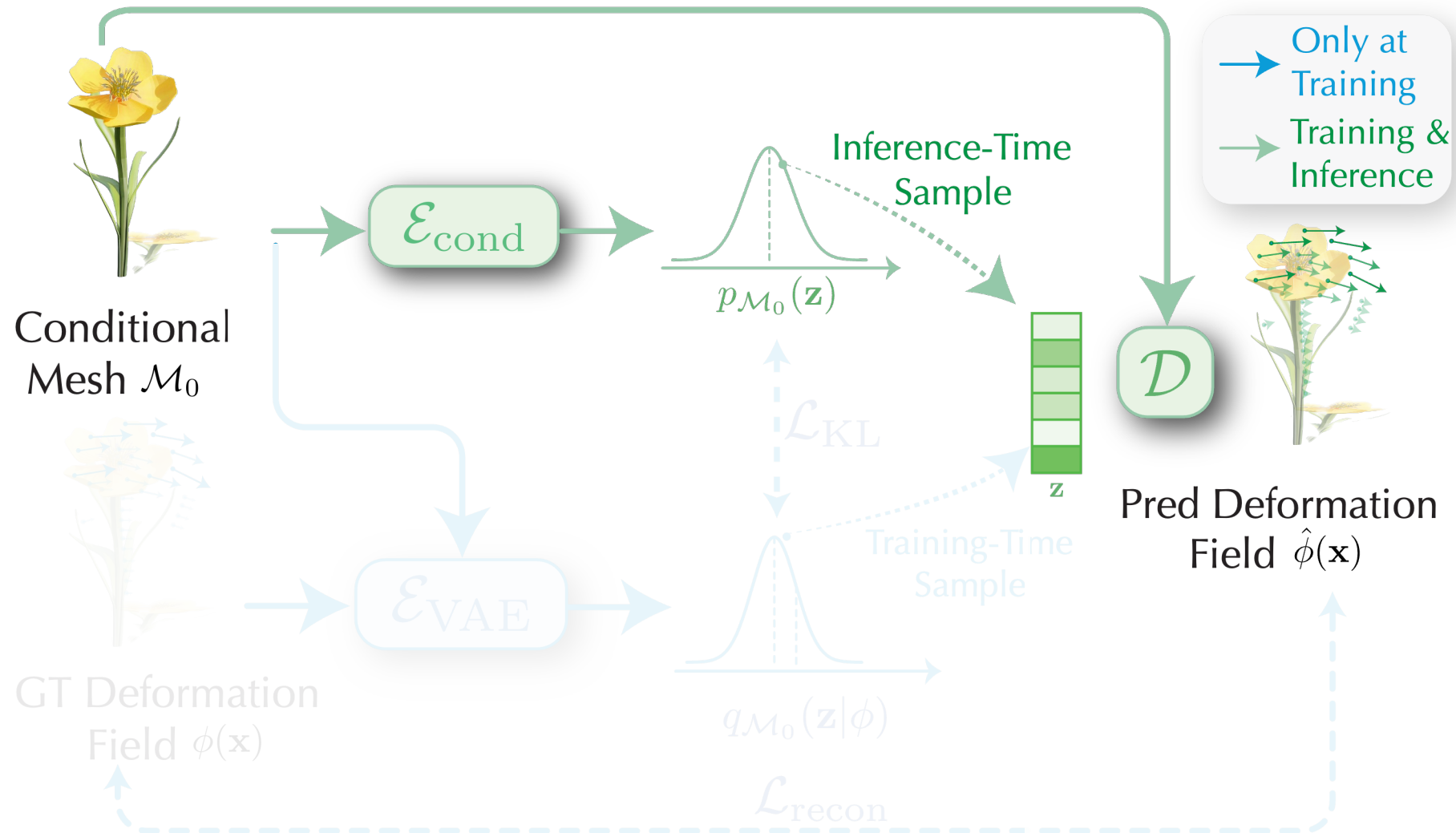


Conditional
Mesh $\mathcal{M}_0$



GT Deformation
Field $\phi(\mathbf{x})$

Encoder-Decoder Training Framework









Scan your kitchen with an iPhone...

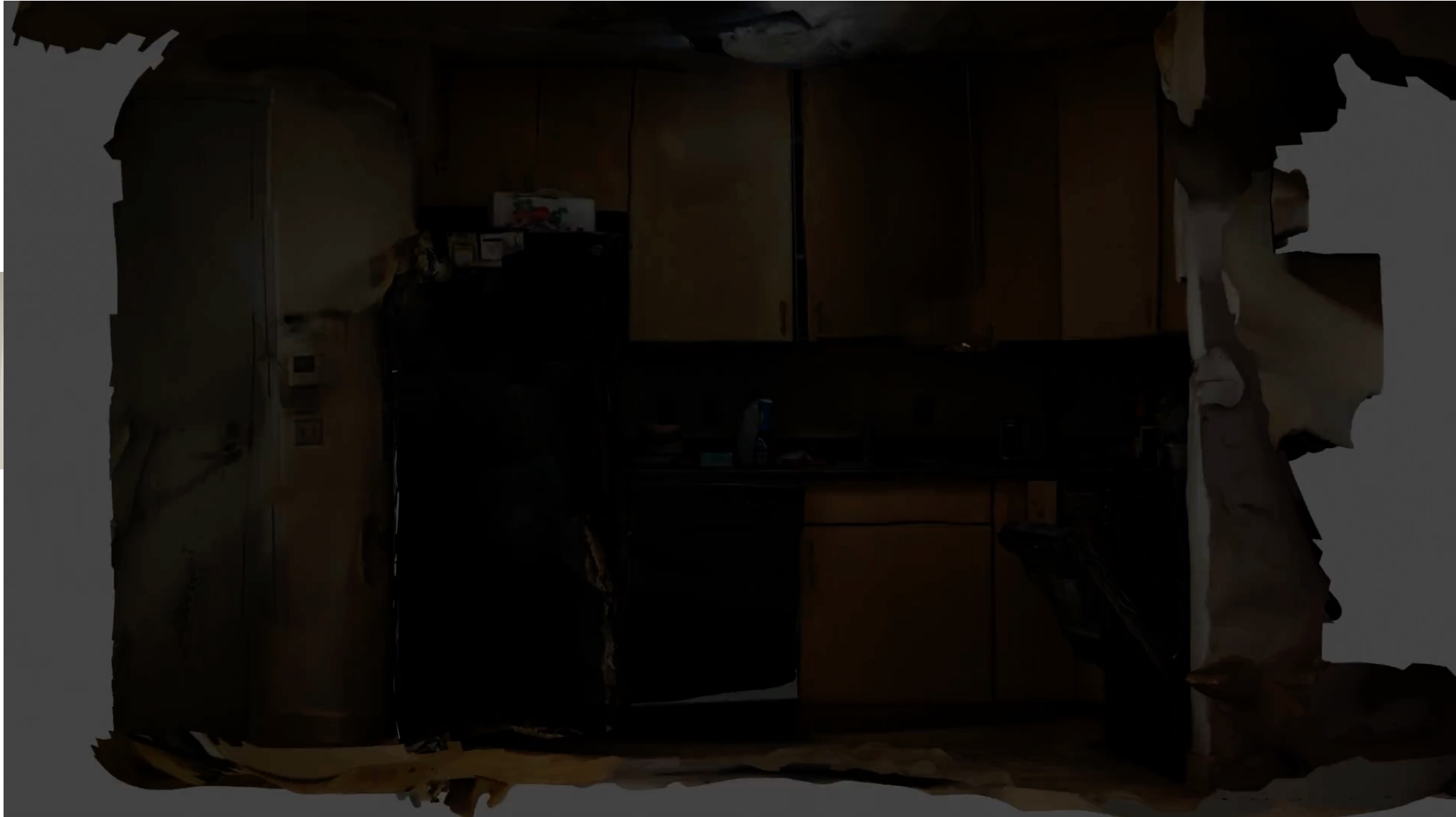


Scanning process

... and interact with it!



Scanning process



Scan your office with an iPhone...



Scanning process

Incorporate Language Conditioning

- Input: 3D scene + text prompt
- Output: an animated 3D scene that follows the text prompt



Static 3D scene

Prompt:

“A child jumps on a seesaw, sending a brick flying.”

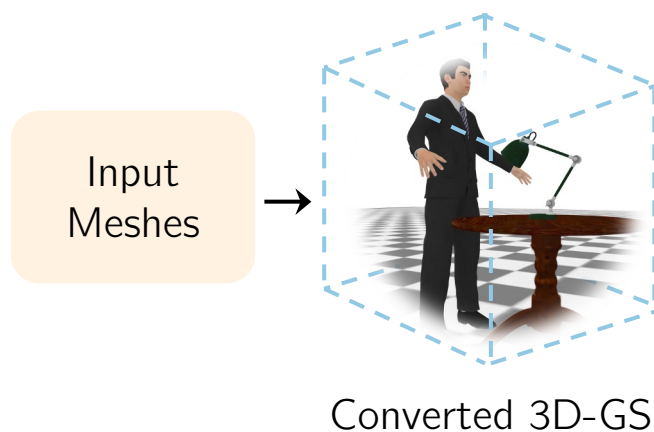
Motion
generation



4D scene

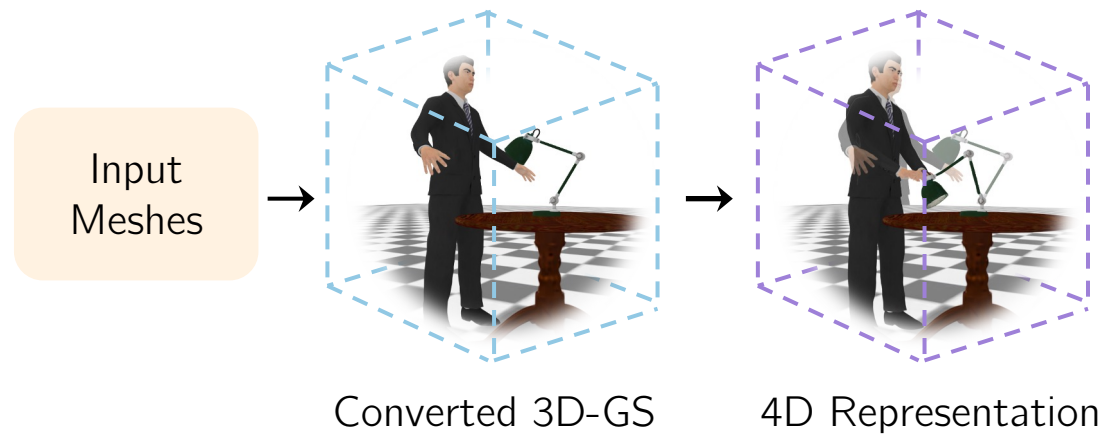
Distilling from Video Models

- Our pipeline begins with a **static 3D scene** of multiple objects.



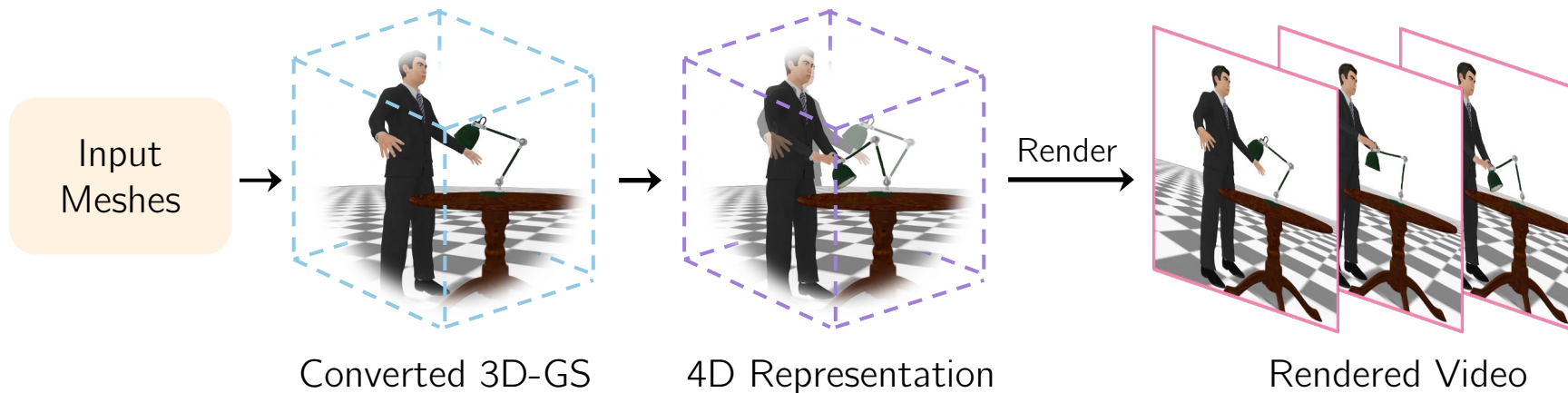
Distilling from Video Models

- The static 3D scene is used to construct a 4D representation.



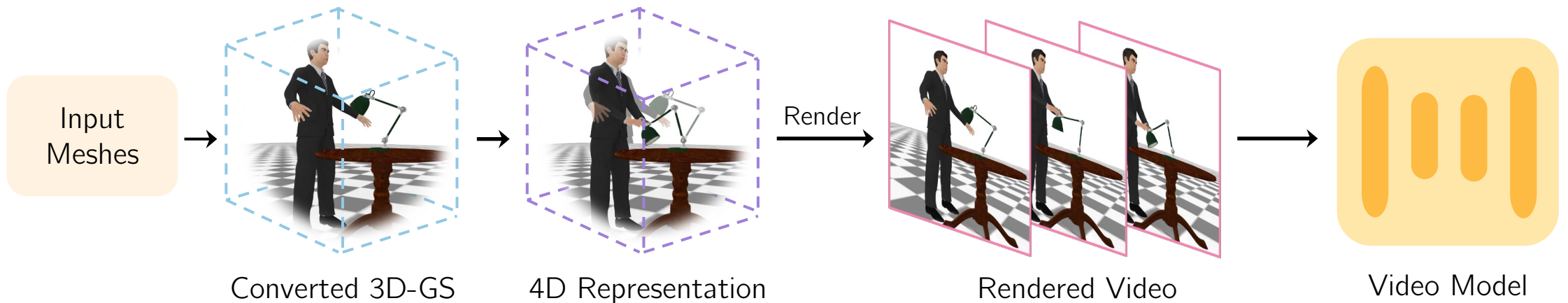
Distilling from Video Models

- At each iteration, a **video** is rendered from the **4D representation**.



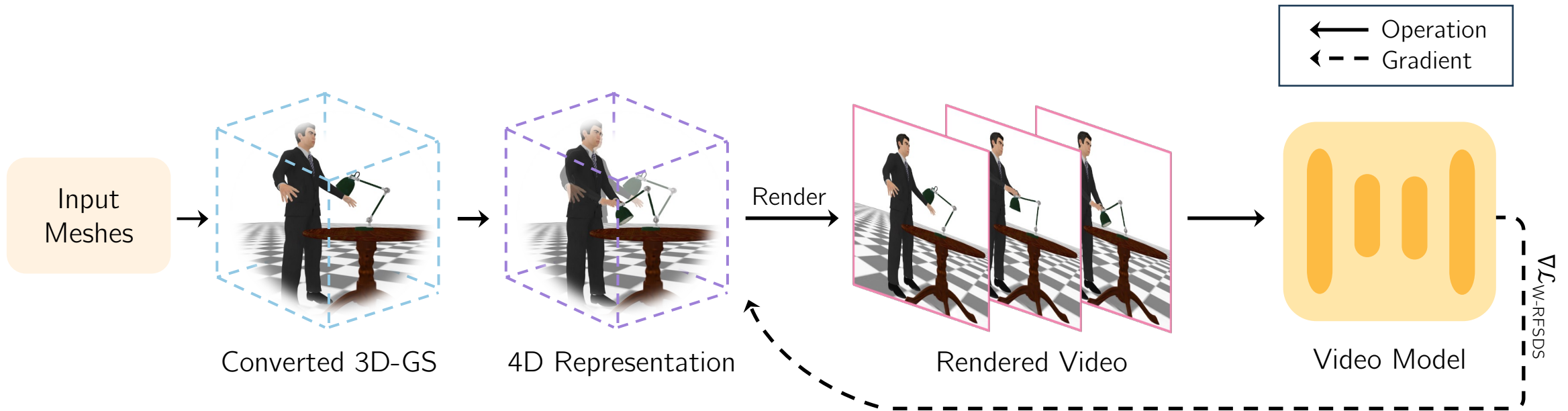
Distilling from Video Models

- The **video** is passed to the **video generation model**.



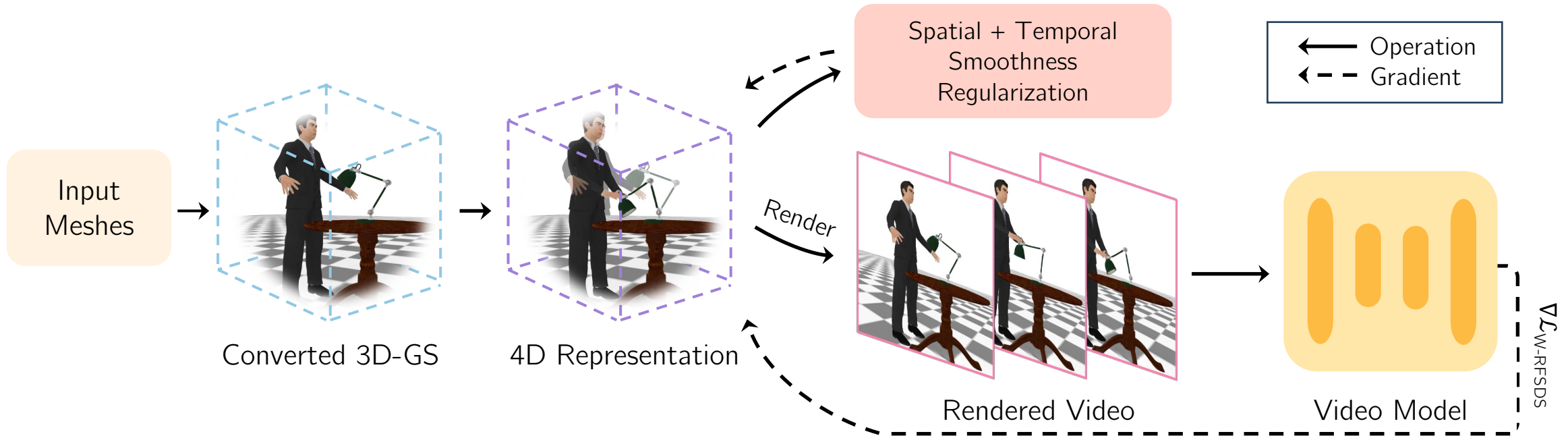
Distilling from Video Models

- Updated gradients are computed to refine the 4D representation.



Distilling from Video Models

- We additionally compute **regularization terms** to enforce spatial and temporal smoothness in the **4D representation**.





Input meshes

Text prompt: "A robot arm picking up a brick."



Input meshes

Text prompt: "A basketball hits the rim and bounces away."



Input meshes



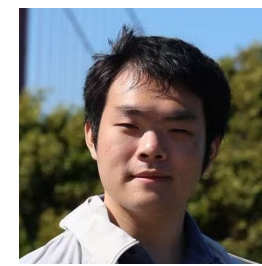
Text prompt: "A cat jumps on a cushion."

Robotic Manipulation

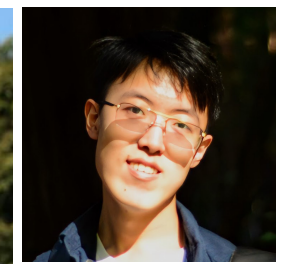
- The generated **dense object flow** can be used for planning in robotic manipulation.



Robotic Manipulation



Yanzhe Lyu



Chen Geng



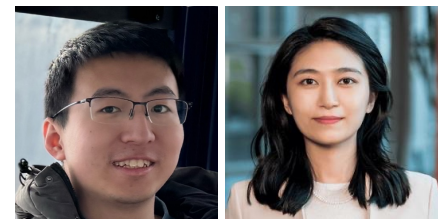
Karthik
Dharmarajan

- The generated **dense object flow** can be used for planning in robotic manipulation.

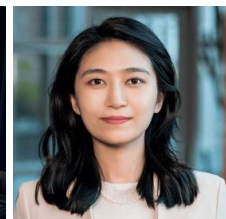


Lyu*, Geng*, Dharmarajan, Zhang, Alzayer, Wu, Wu. CVPR 2026
Dharmarajan, Huang, Wu, Fei-Fei, Zhang. ICRA 2026

Casting Embodied Cognition as POMDP World Modeling

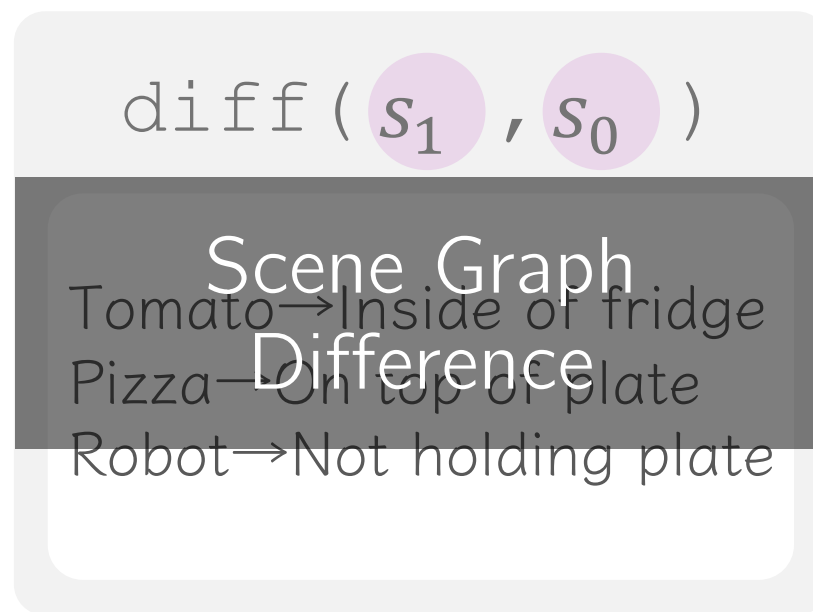
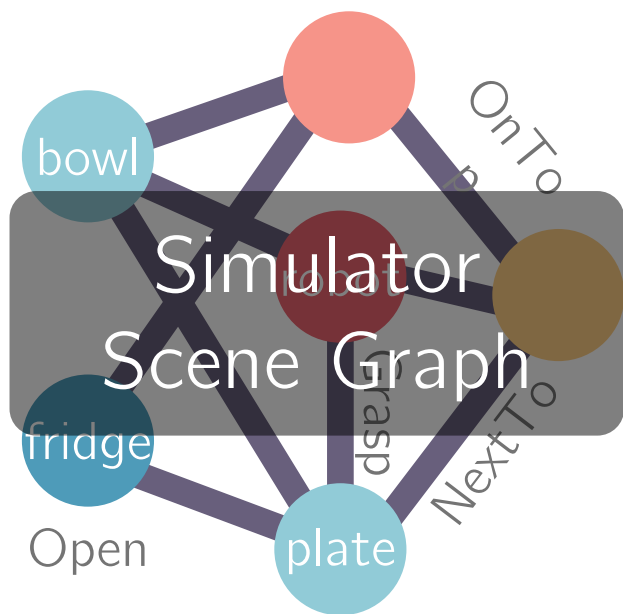


Qineng



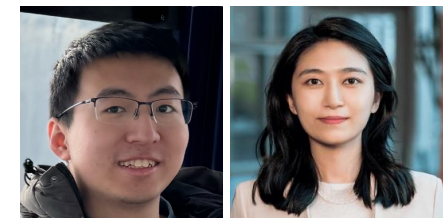
Manling

ENACT evaluates embodied cognition through POMDP world modeling via egocentric interaction.



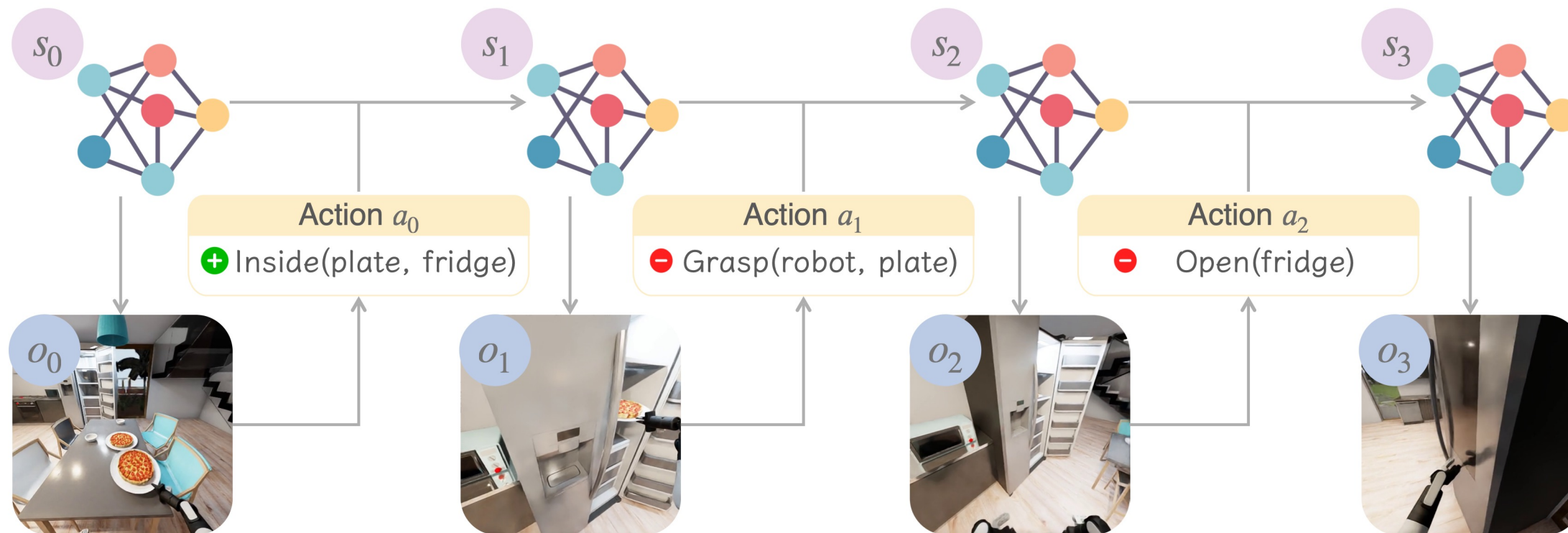
State Space s Observation Space o Action Space a

Casting Embodied Cognition as POMDP World Modeling



Qineng

Manling



State s



Observation o



Action a

$s_1 - s_0$

Given simulated manipulation dataset...

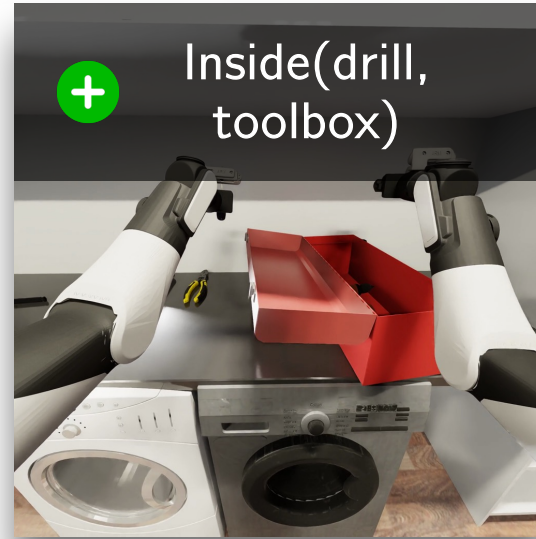
Extract keyframes and scene graph changes



Extract keyframes and scene graph changes



Extract keyframes and scene graph changes



Forward world modeling as reordering frames

↻ Shuffling

Current State



OnTop(drill,
table)



Grasping(driver)



OnTop(driver,
table)



Inside(drill,
toolbox)

Forward world modeling as reordering frames

↻ Shuffling

Current State



OnTop(drill,
table)



Grasping(driver)



OnTop(driver,
table)



Inside(drill,
toolbox)

Forward world modeling as reordering frames

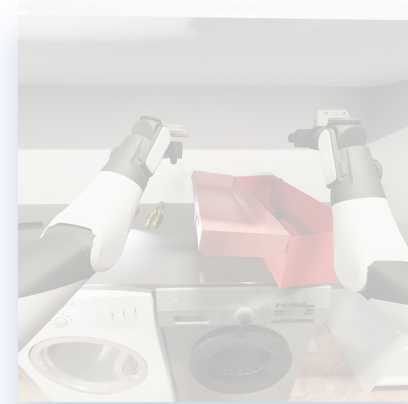
Current State



OnTop(drill,
table)



Grasping(driver)



OnTop(driver,
table)



Inside(drill,
toolbox)

MLLM
Output

Forward world modeling as reordering frames

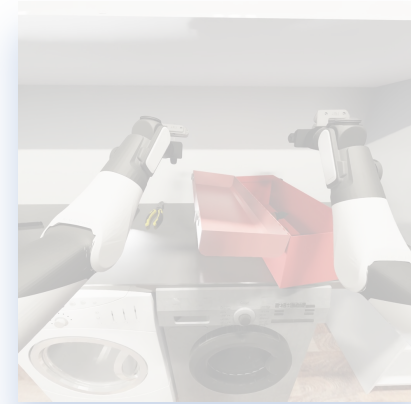
Current State



− OnTop(drill, table)



+ Grasping(driver)



− OnTop(driver, table)



+ Inside(drill, toolbox)

MLLM
Output



Forward world modeling as reordering frames

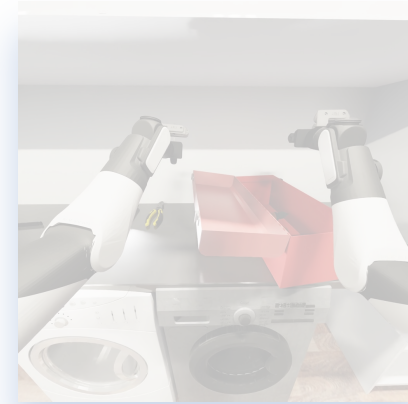
Current State



− OnTop(drill, table)



+ Grasping(driver)



− OnTop(driver, table)



+ Inside(drill, toolbox)

MLLM
Output



Forward world modeling as reordering frames

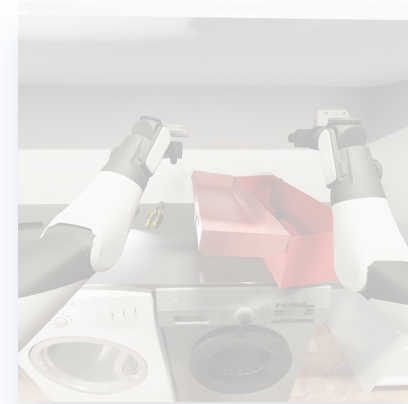
Current State



− OnTop(drill, table)



+ Grasping(driver)



− OnTop(driver, table)



+ Inside(drill, toolbox)

MLLM
Output



Forward world modeling as reordering frames

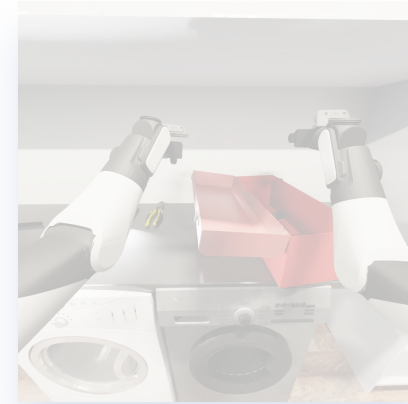
Current State



OnTop(drill,
table)



Grasping(driver)



OnTop(driver,
table)



Inside(drill,
toolbox)

MLLM
Output



Inverse world modeling as reordering actions

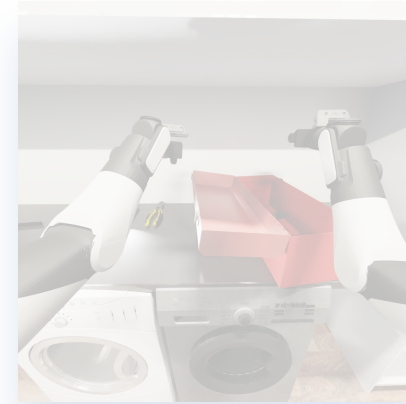
Current State



− OnTop(drill,
table)



+ Grasping(driver)



− OnTop(driver,
table)



+ Inside(drill,
toolbox)

MLLM
Output



Inverse world modeling as reordering actions

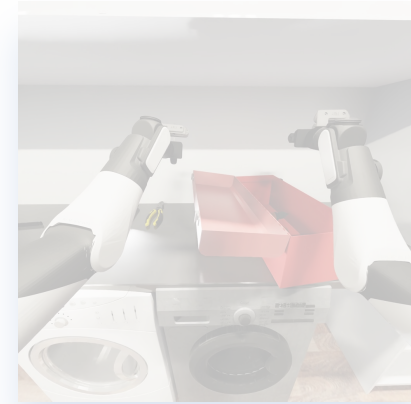
Current State



− OnTop(drill,
table)



+ Grasping(driver)



− OnTop(driver,
table)



+ Inside(drill,
toolbox)

MLLM
Output



Inverse world modeling as reordering actions

↻ Shuffling

Current State



OnTop(drill,
table)



Grasping(driver)



OnTop(driver,
table)



Inside(drill,
toolbox)

Inverse world modeling as reordering actions

↻ Shuffling

Current State



Inside(drill,
toolbox)



OnTop(drill,
table)



OnTop(driver,
table)



Grasping(driver)

MLLM
Output

Inverse world modeling as reordering actions

Current State



Inside(drill,
toolbox)



OnTop(drill,
table)



OnTop(driver,
table)



Grasping(driver)

MLLM
Output

Inverse world modeling as reordering actions

Current State



Inside(drill,
toolbox)



OnTop(drill,
table)



OnTop(driver,
table)



Grasping(driver)

MLLM
Output



Inside(drill,
toolbox)

Inverse world modeling as reordering actions

Current State



Inside(drill,
toolbox)



OnTop(drill,
table)



OnTop(driver,
table)



Grasping(driver)

MLLM
Output



OnTop(drill,
table)



Inside(drill,
toolbox)

Inverse world modeling as reordering actions

Current State



Inside(drill,
toolbox)



OnTop(drill,
table)



OnTop(driver,
table)



Grasping(driver)

MLLM
Output



OnTop(drill,
table)



OnTop(driver,
table)



Inside(drill,
toolbox)

Inverse world modeling as reordering actions

Current State



Inside(drill,
toolbox)



OnTop(drill,
table)



OnTop(driver,
table)



Grasping(driver)

MLLM
Output



OnTop(drill,
table)



Grasping(driver)

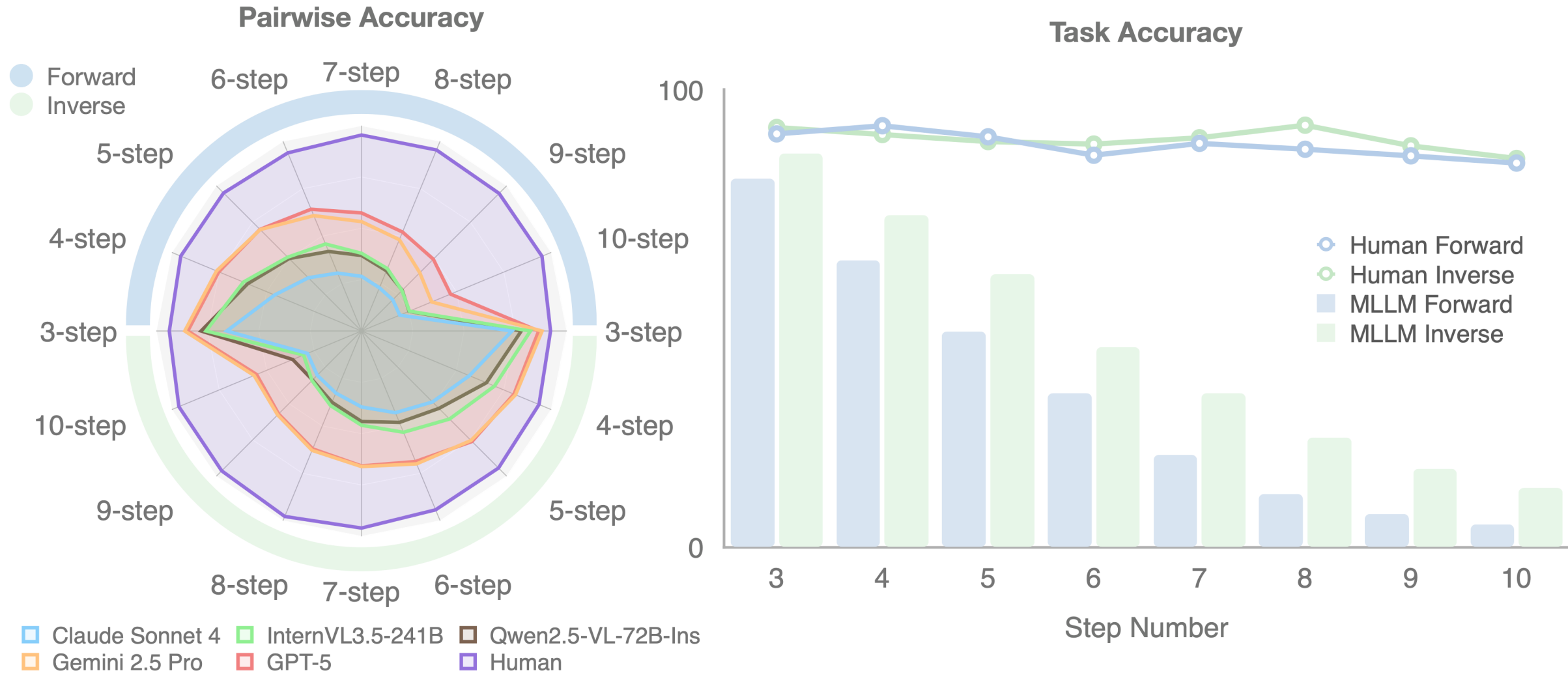


OnTop(driver,
table)



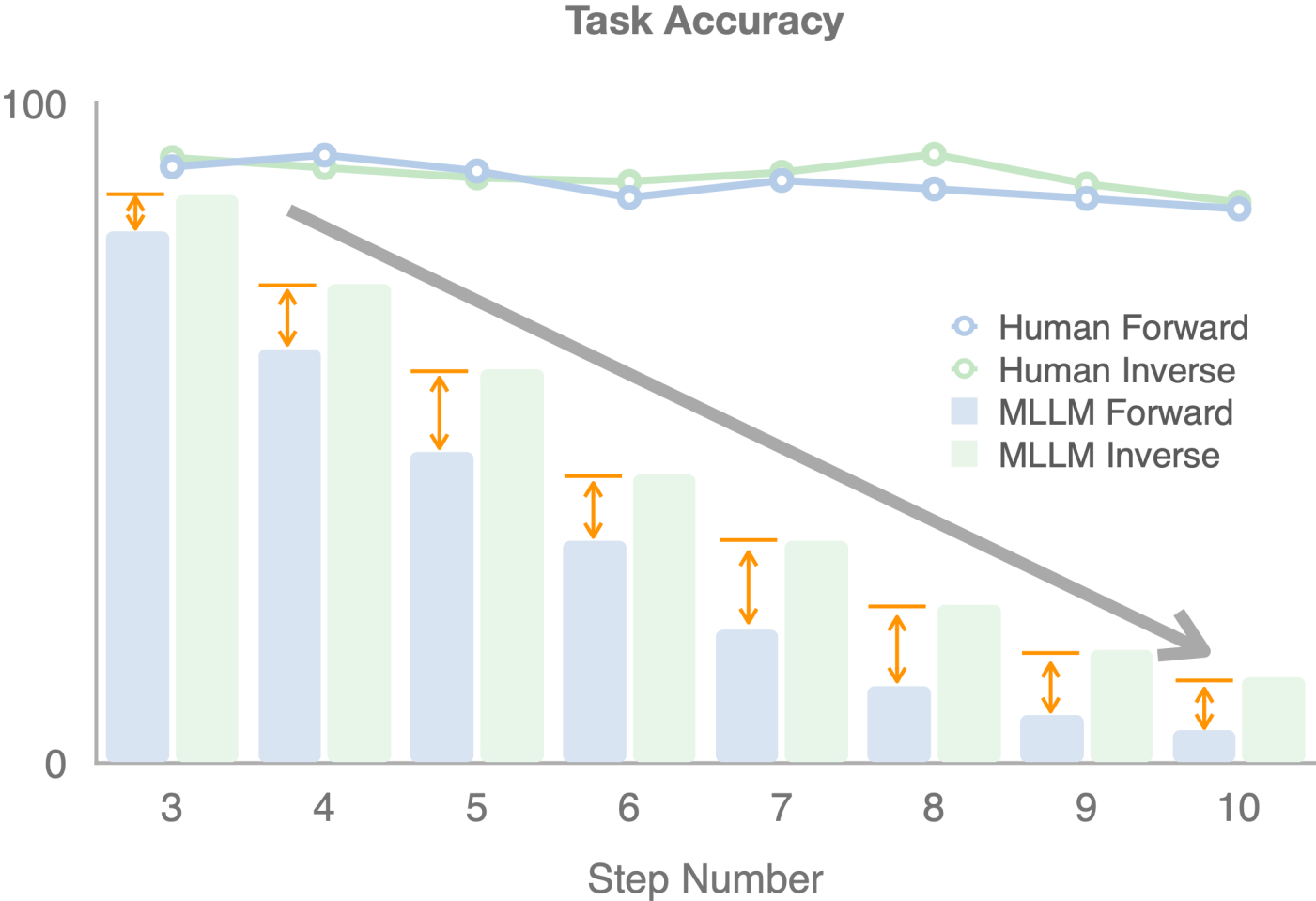
Inside(drill,
toolbox)

MLLMs performance on ENACT across varied step numbers



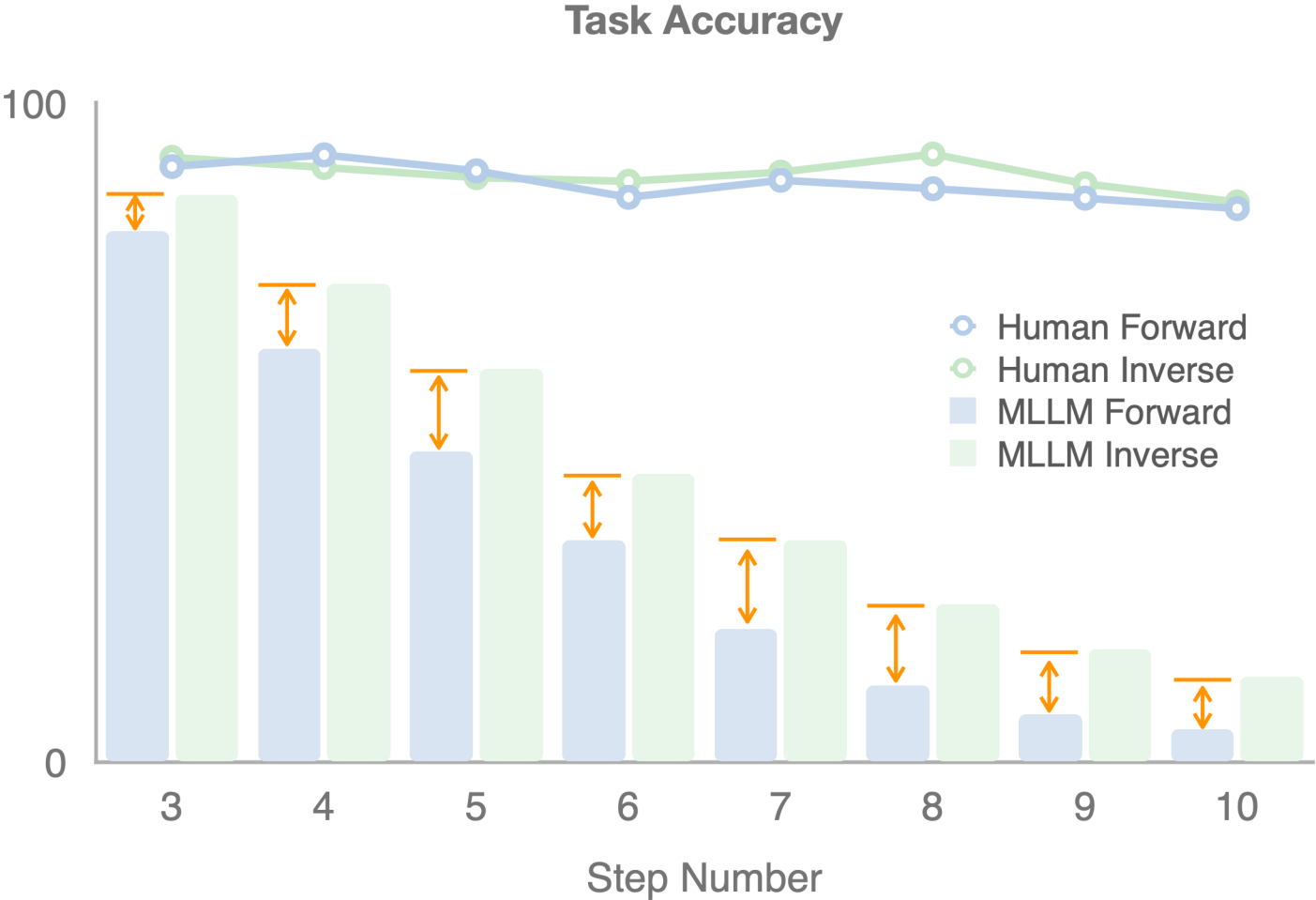
MLLMs performance on ENACT across varied step numbers

- MLLMs' performance decreases as step number increases



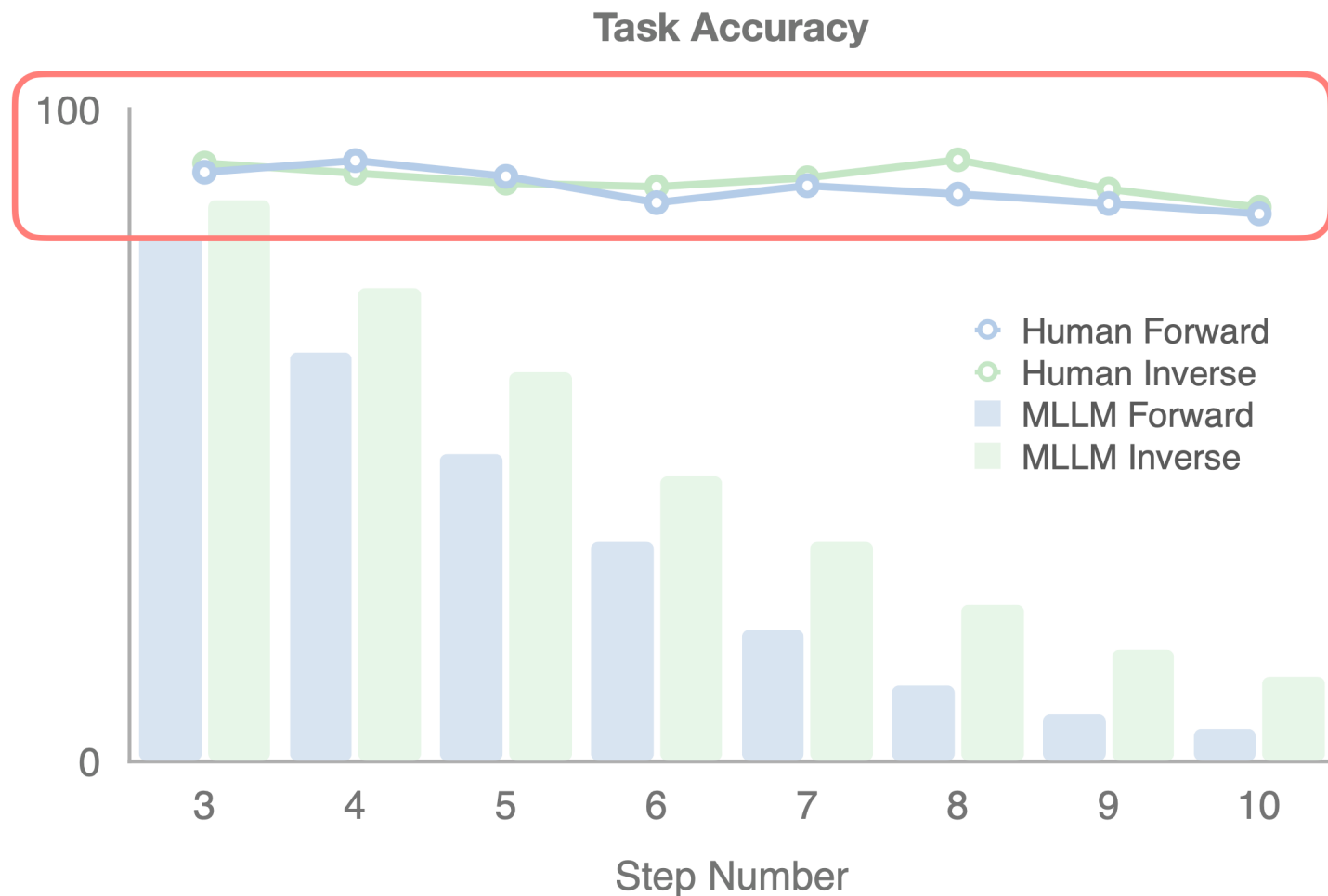
MLLMs performance on ENACT across varied step numbers

- MLLMs' performance decreases as step number increases
- MLLMs perform better on inverse setting than forward setting



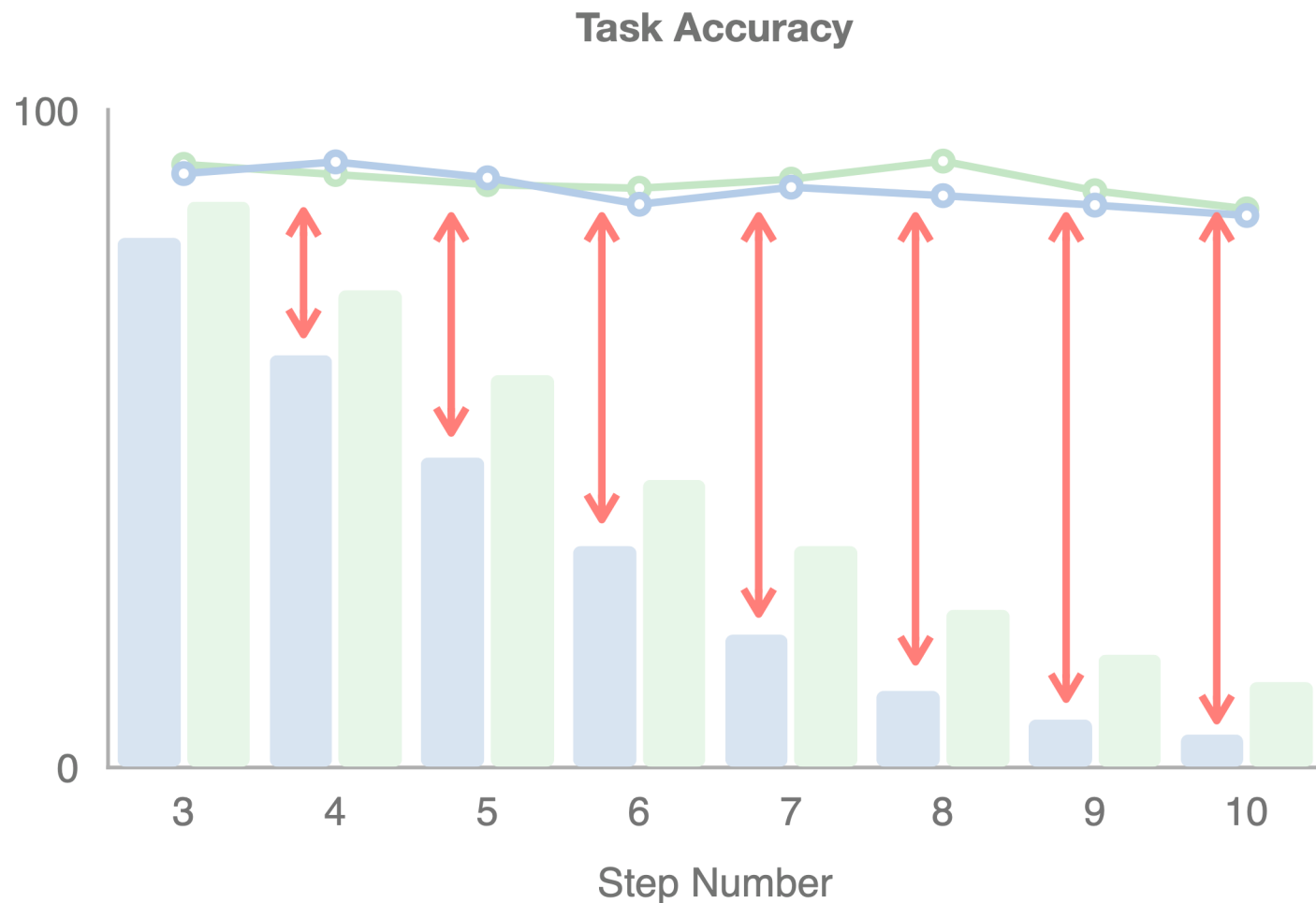
MLLMs performance on ENACT across varied step numbers

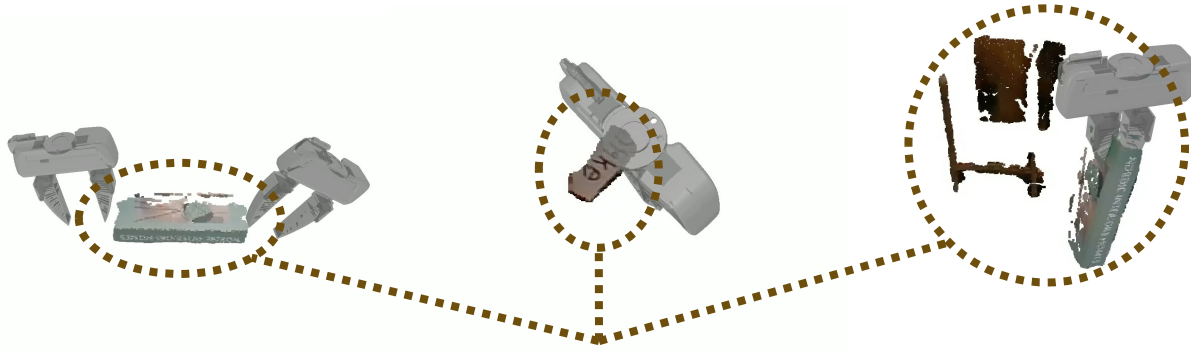
- MLLMs' performance decreases as step number increases
- MLLMs perform better on inverse setting than forward setting
- Humans achieve near-ceiling performance



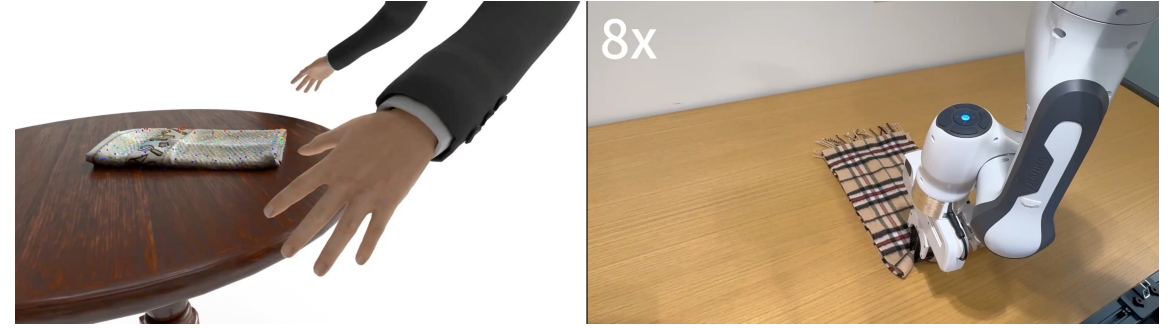
MLLMs performance on ENACT across varied step numbers

- MLLMs' performance decreases as step number increases
- MLLMs perform better on inverse setting than forward setting
- Humans achieve near-ceiling performance
- Remains a large gap between MLLMs and humans, especially at longer steps

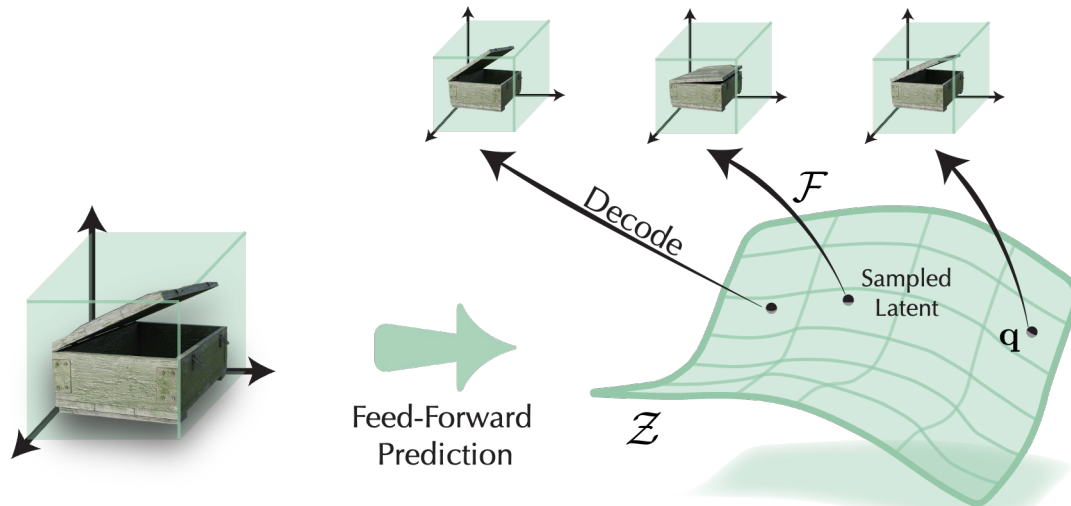




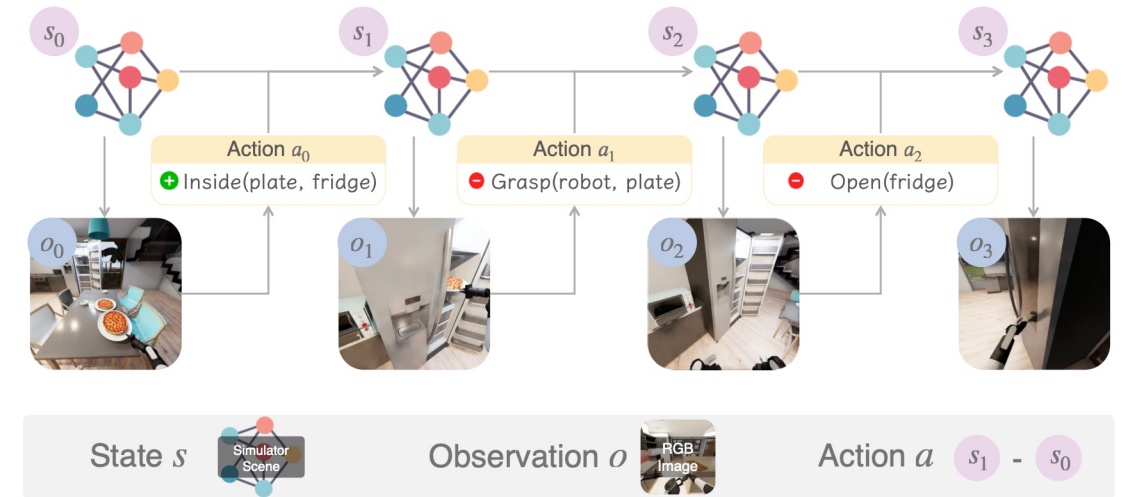
BLADE, CoRL 2024; STACK, ICRA 2026



CHORD, CVPR 2026; Dream2Flow, ICRA 2026



NeuROK, CVPR 2026



ENACT, ICLR 2026

- Structured representations (via FMs) from videos, 3D/4D data, and language
- From spatial reasoning to spatial-temporal-causal reasoning for physical agents