Simulation-Powered Human-Centered Embodied Perception and Interaction

Li Yi

June 9, 2024

Self Introduction

Li Yi (弋力)

- 2009 - 2013, B.E. @ Tsinghua University
- 2013 - 2019, Ph.D. @ Stanford University
- 2019 - 2021, Research Scientist @ Google Research
- 2021 - now, Assistant Professor @ Tsinghua University
- Research: 3D Visual Computing and Embodied Perception
- Homepage: https://ericyi.github.io/
- Email: ericyi0124@gmail.com

Internet AI to Embodied AI

Datasets

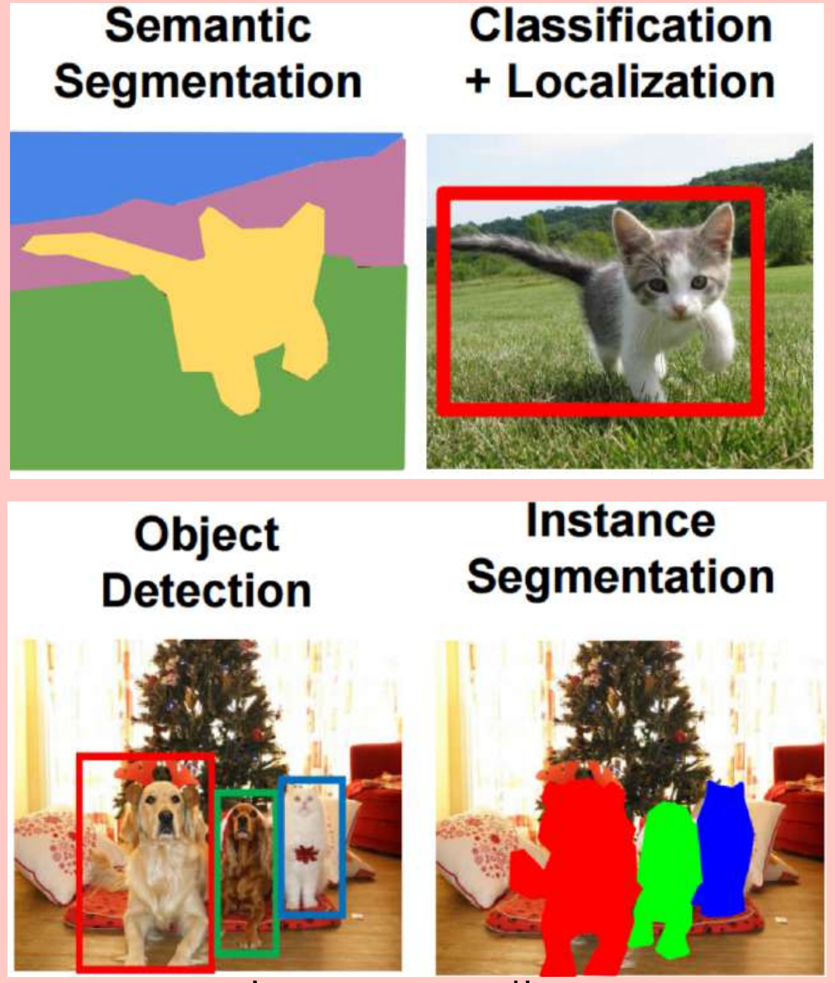
ImageNet, Deng et al. 2009

MSCOCO, Lin et al. 2014

ShapeNet, Chang et al. 2015

• • •

Tasks



• • •

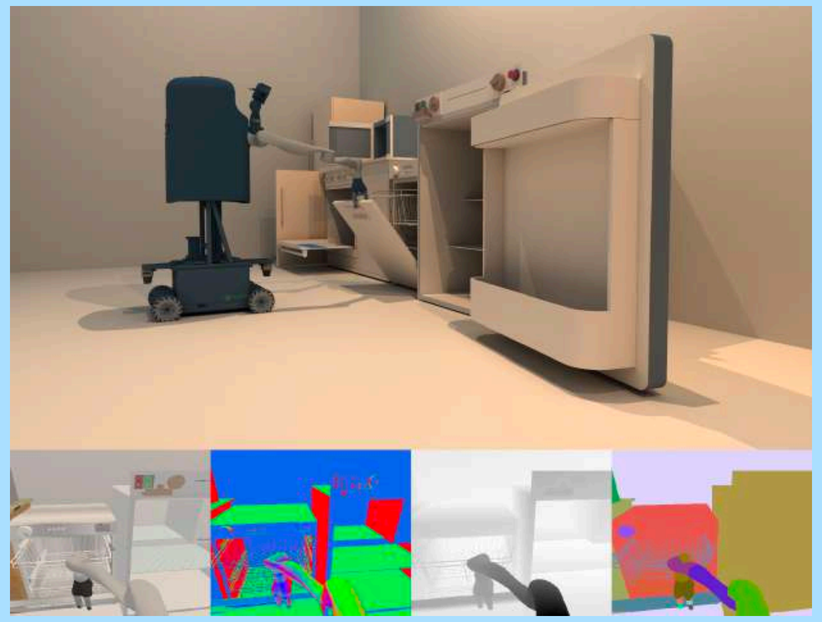
image credits: http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Internet AI

Environments

Habitat

ThreeDWorld



Sapien

AI2-Thor

iGibson

• • •

Tasks

Vision-Language Navigation

Manipulation

Embodied-QA

Rearrangement

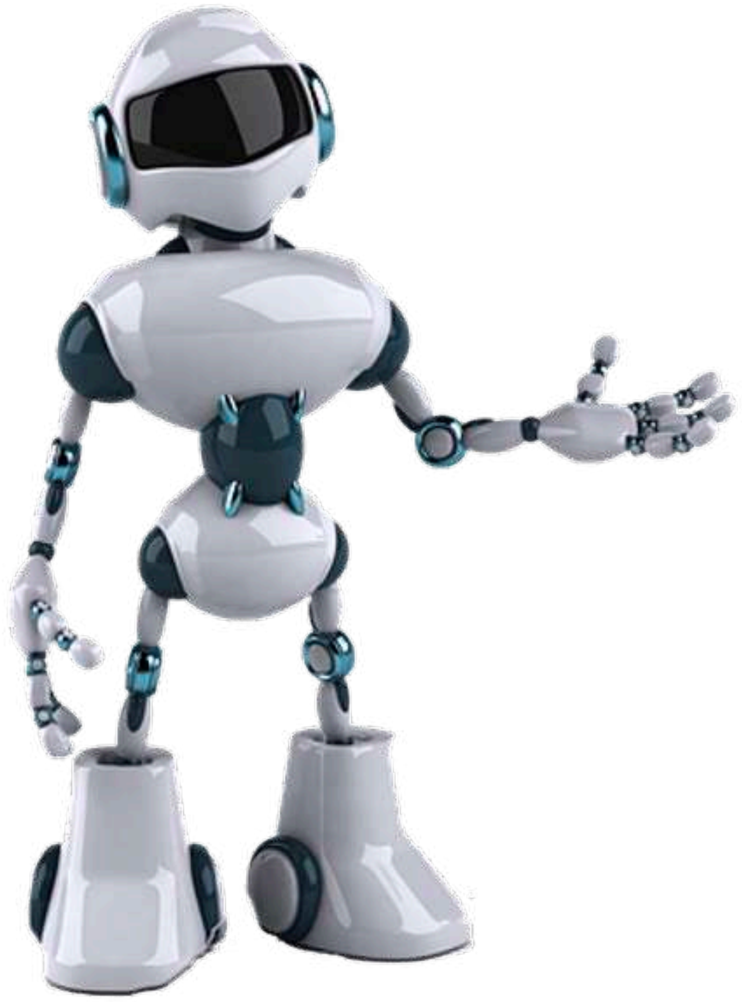
Mobile Manipulation

Instruction Following

• • •

Embodied AI

Embodied AI



Embodied Agent

Embodied Task Execution

image credits: Matterport3D

Environment

Human-Centered Embodied AI

Embodied Agent

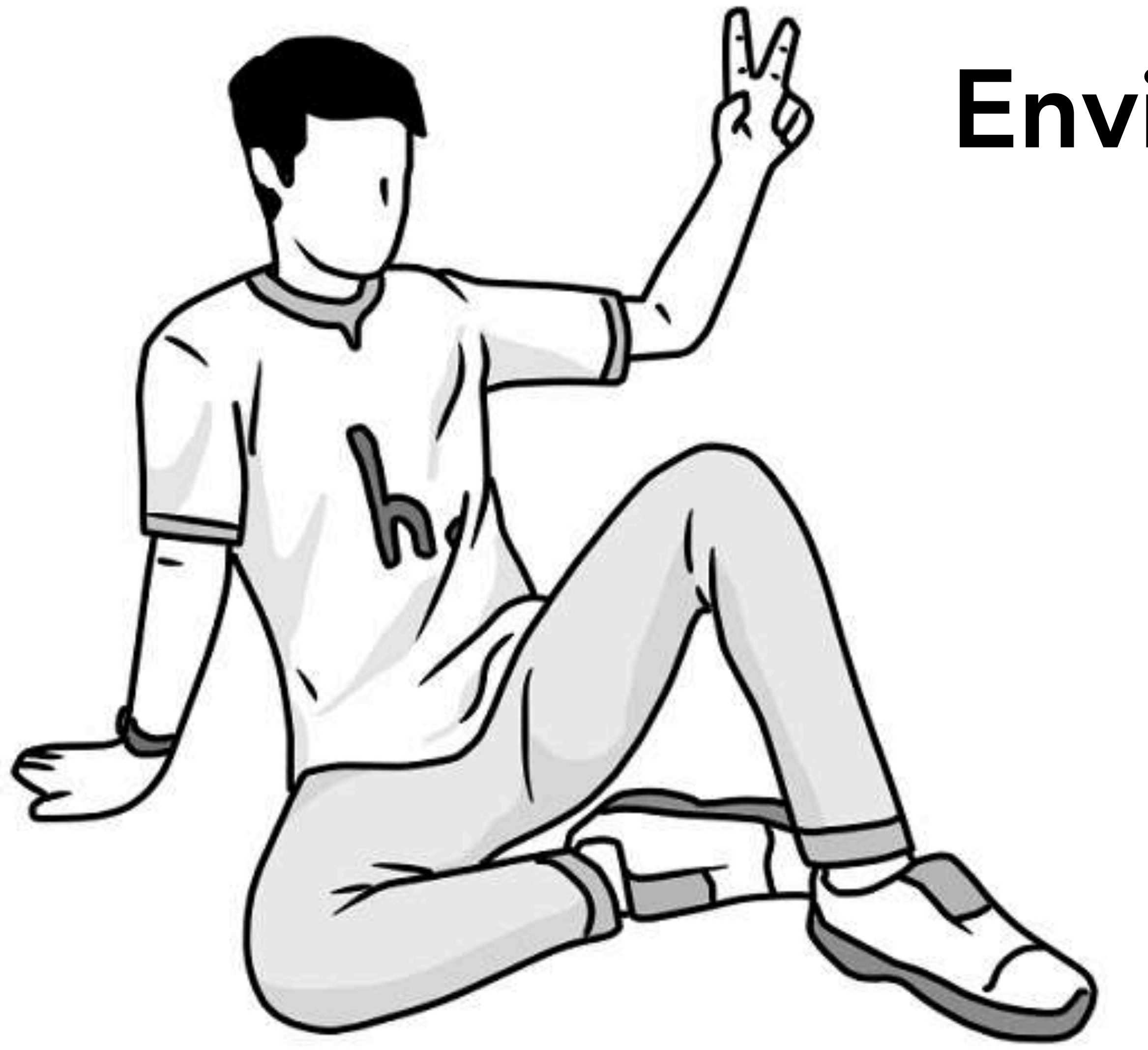
Human-Inspired Task Execution

image credits: Matterport3D

Environment

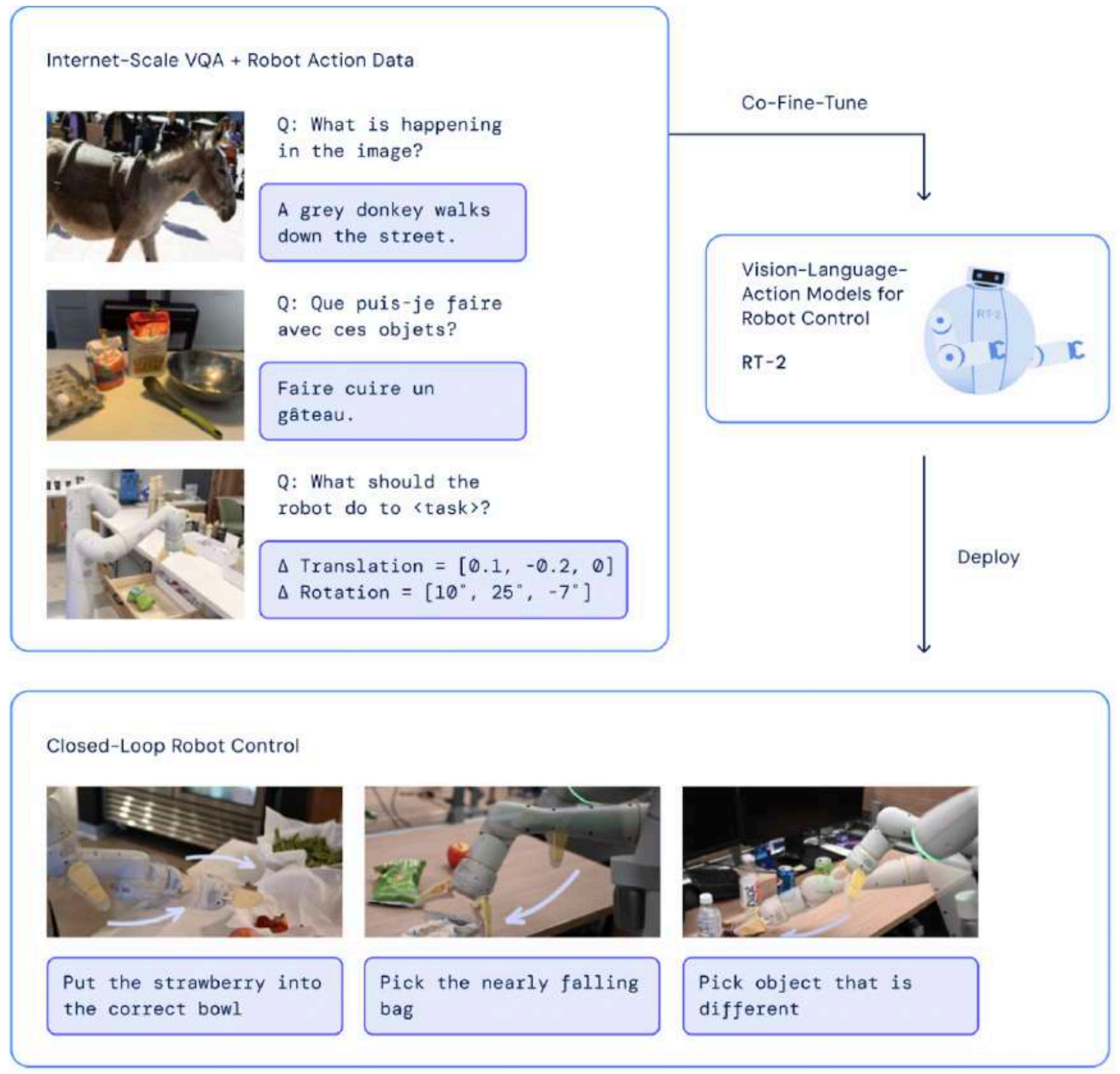
Human Robot Interaction

Handover



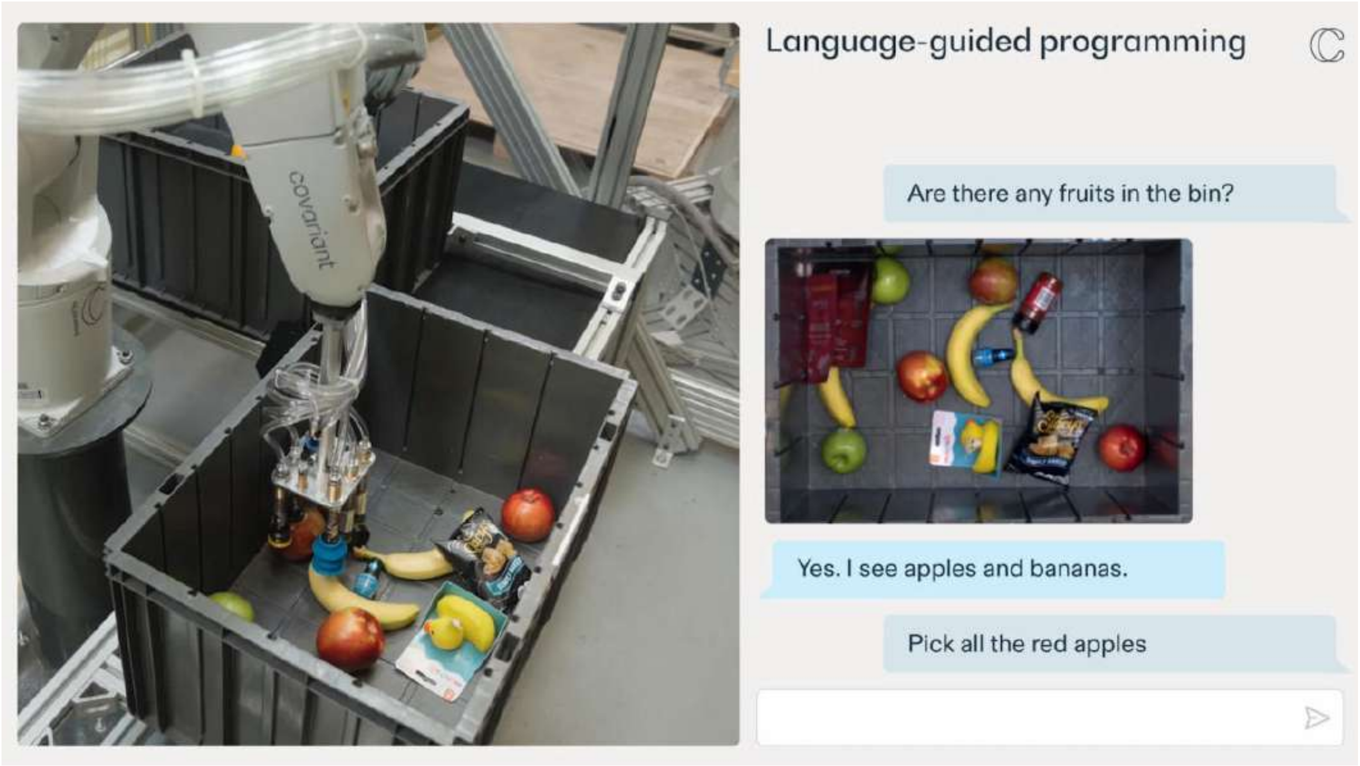
Human

How to Learn?



Google RT-1/2/X

Stanford Mobile Aloha

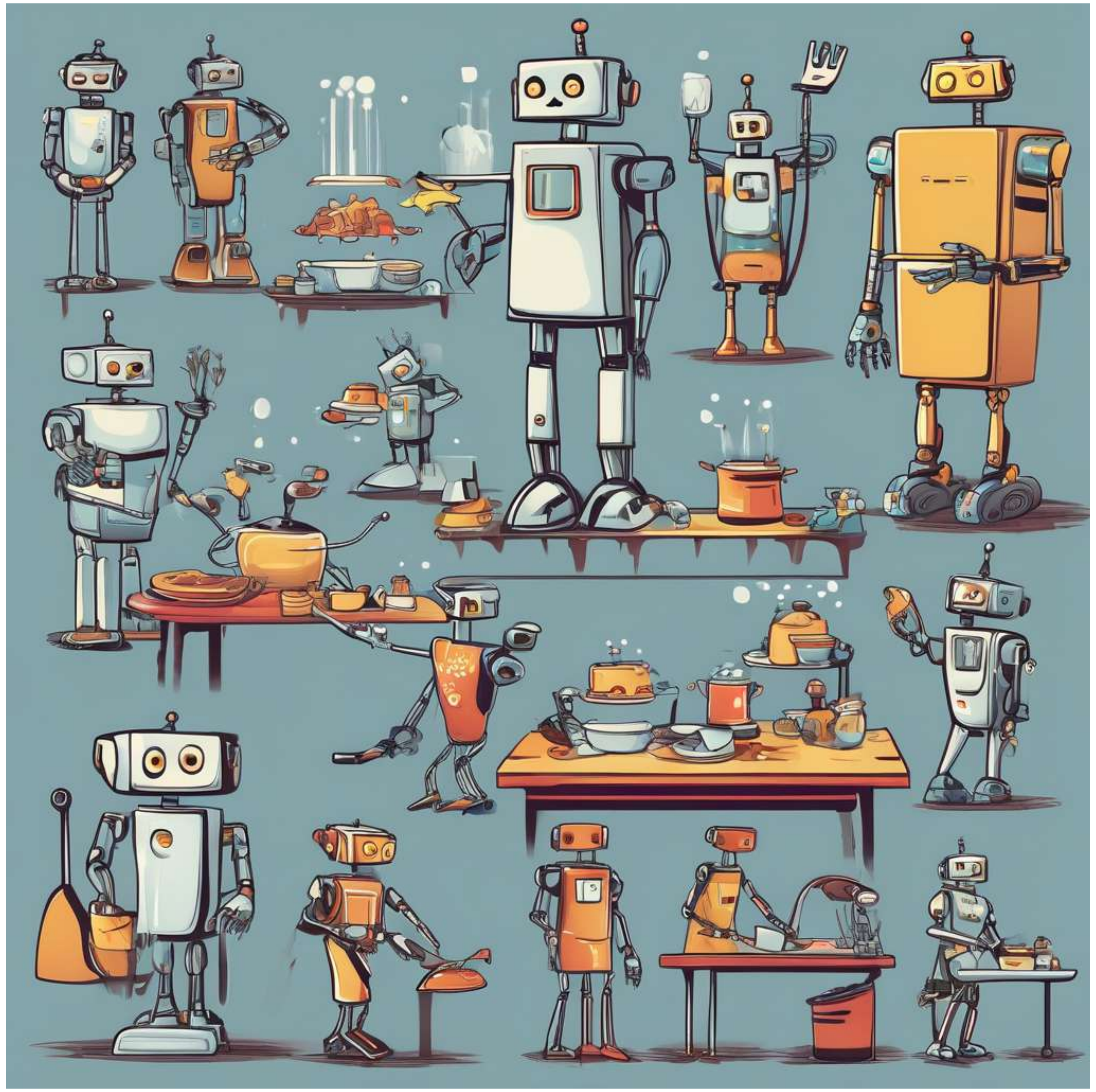


Covariant

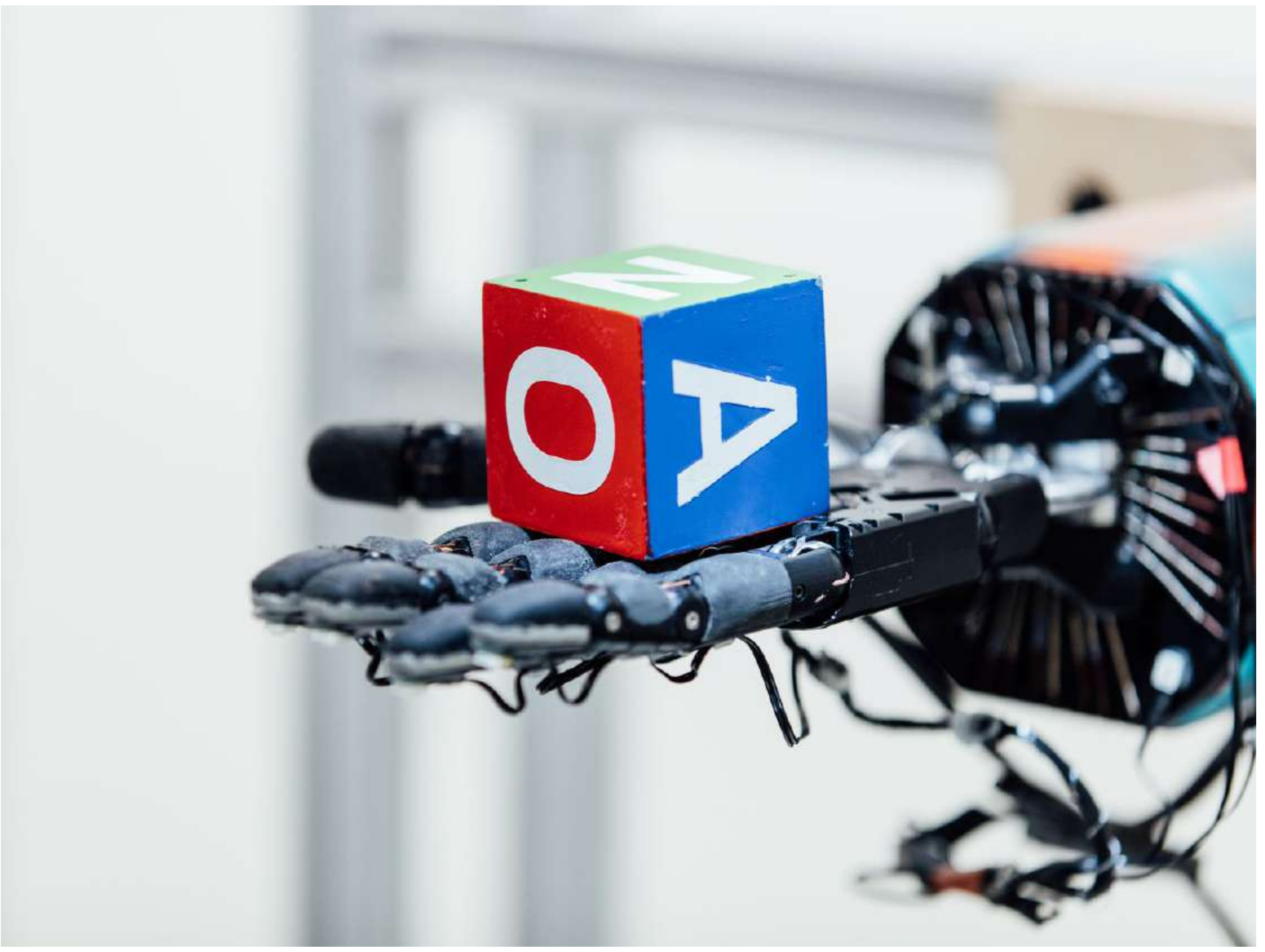
Figure 01

Goal: Embodied Generalist

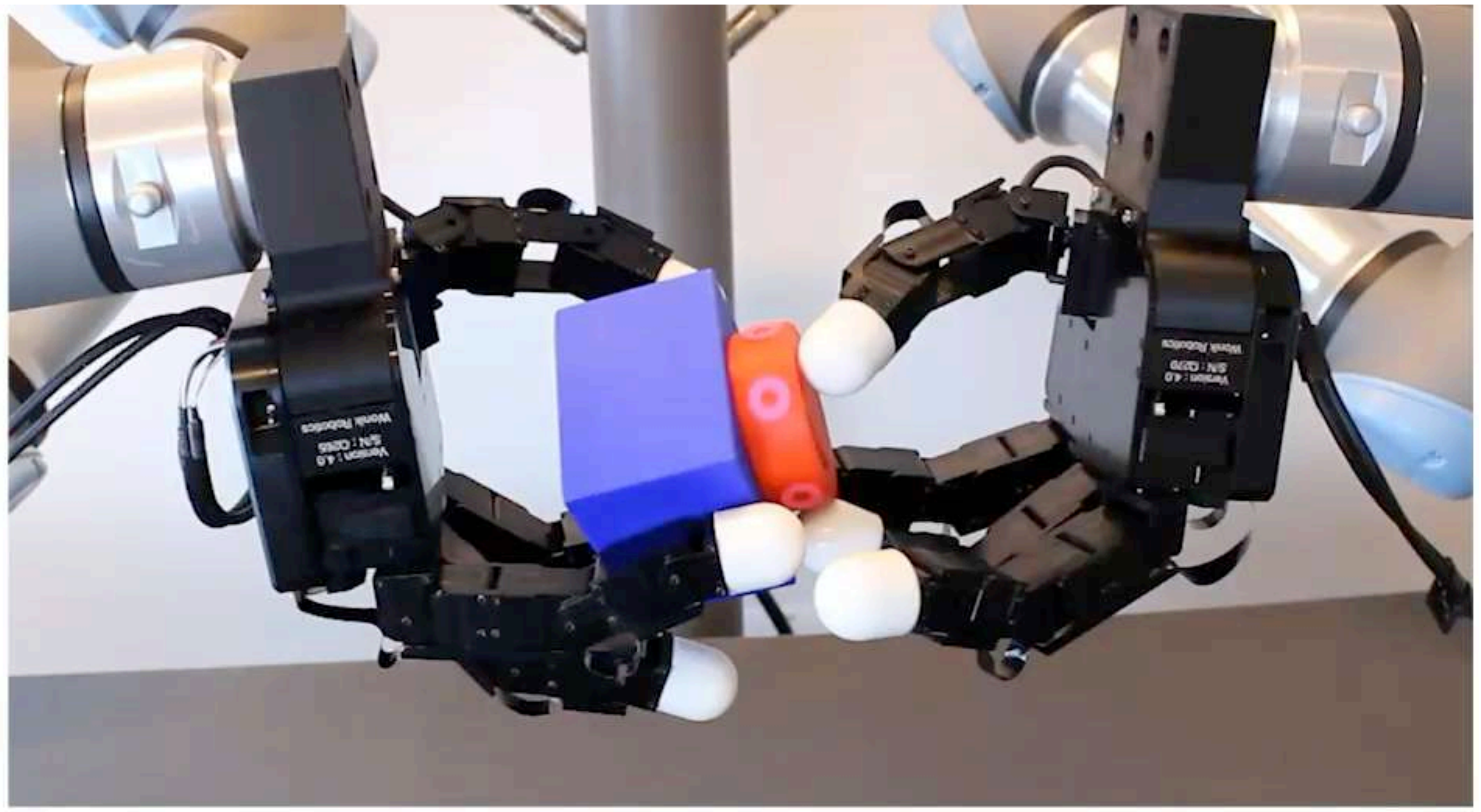
A tremendous amount of tasks in the open world



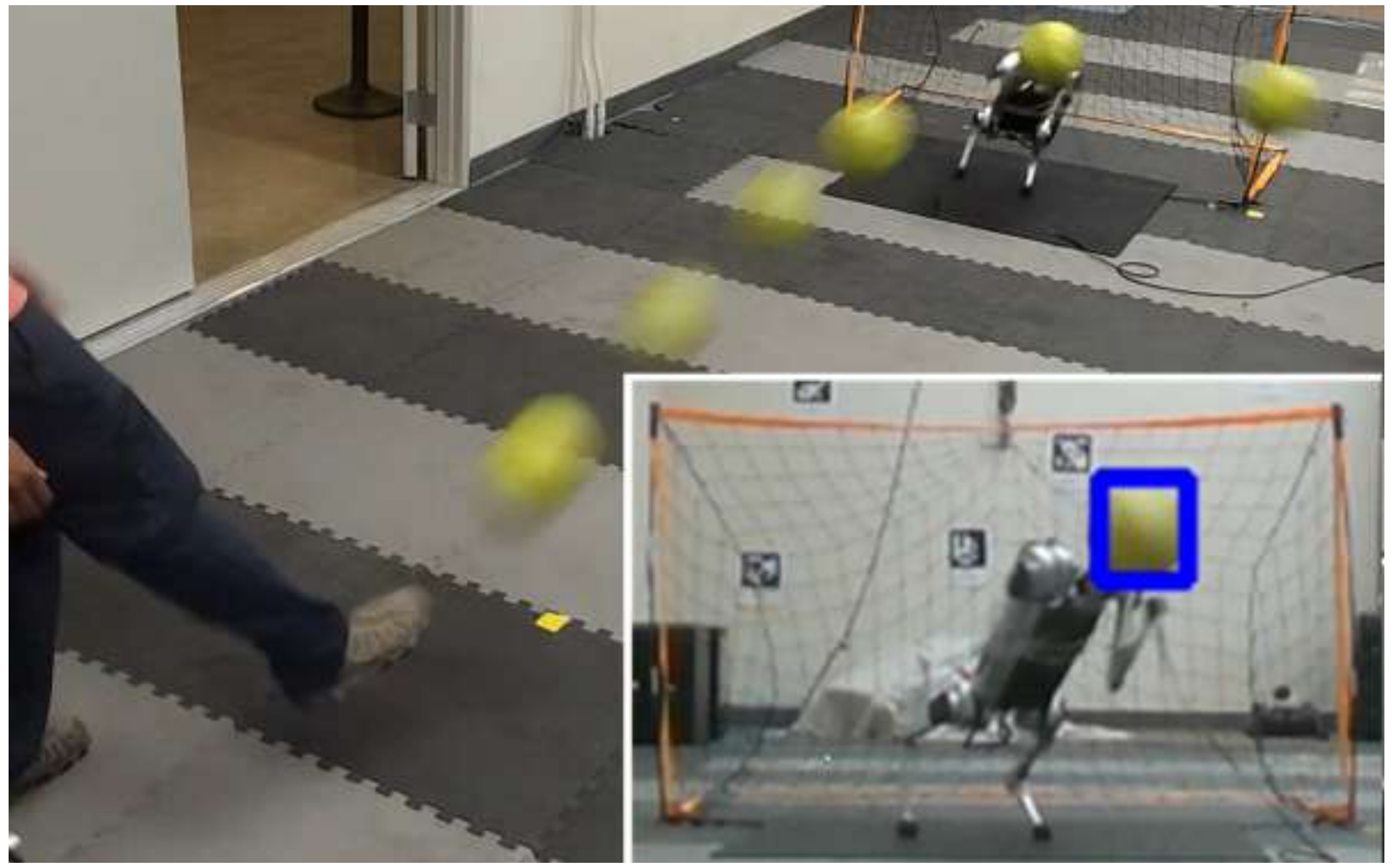
Reality: Embodied Specialists



OpenAI, 2018



Lin et al., 2024



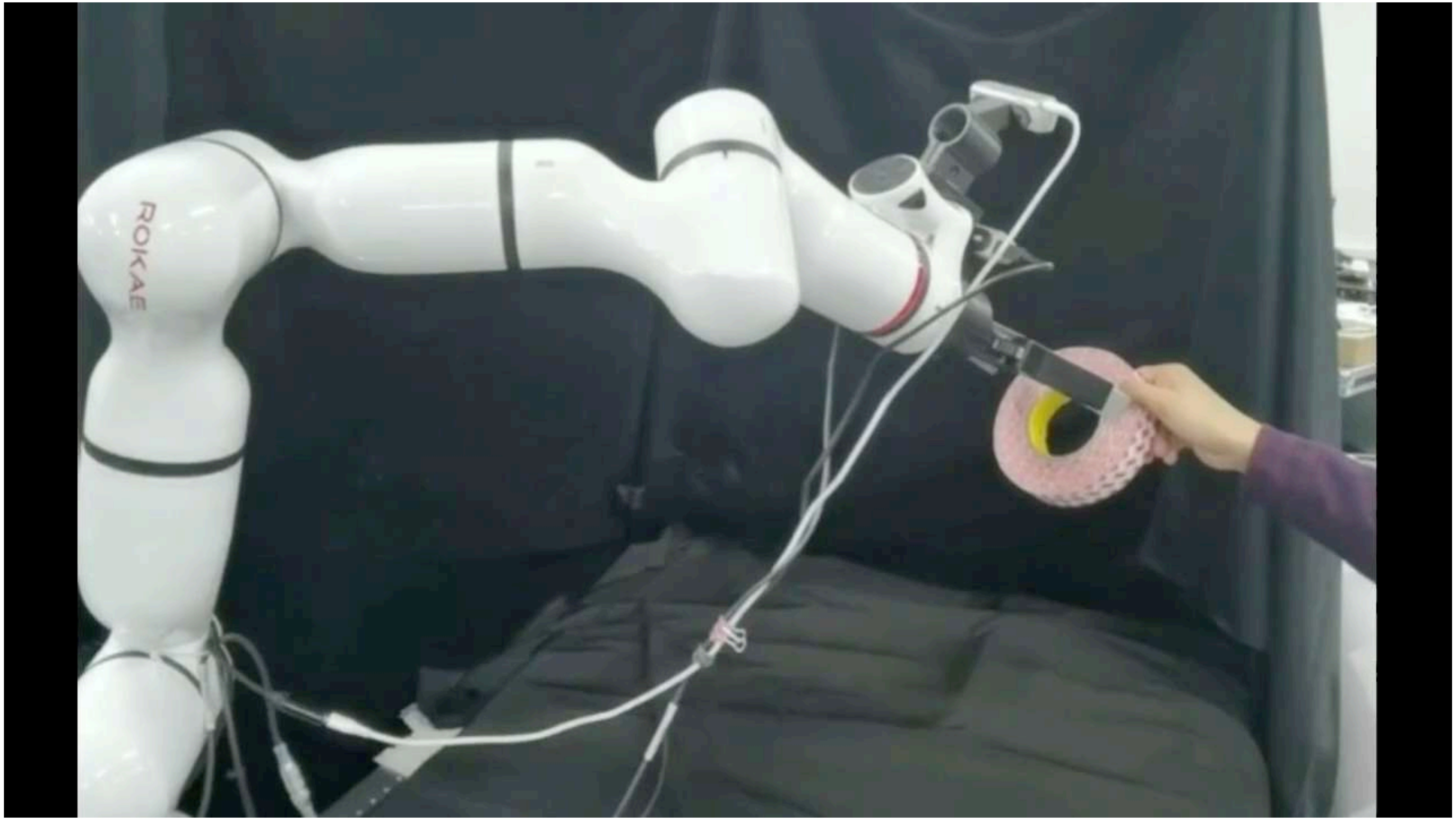
Huang et al., IROS 2023

Xu et al., RSS 2022

In the Open World?

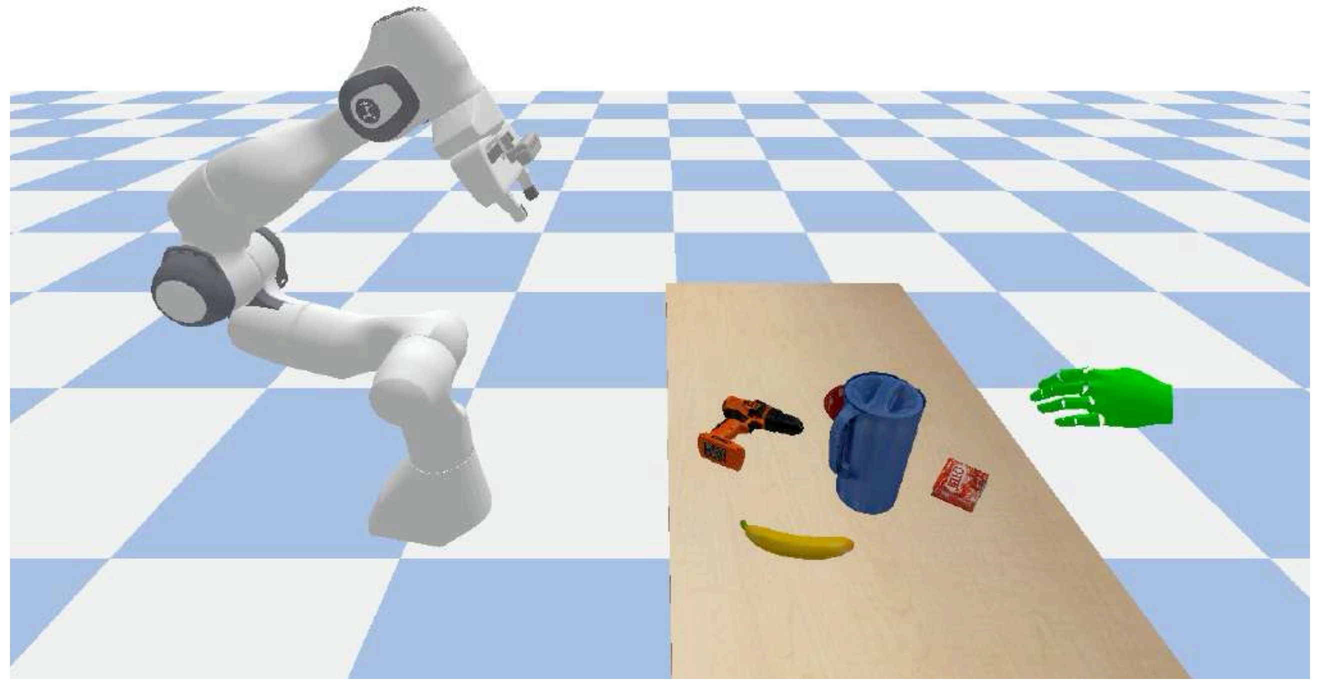
https://www.youtube.com/watch?v=k3GKGDng7k0

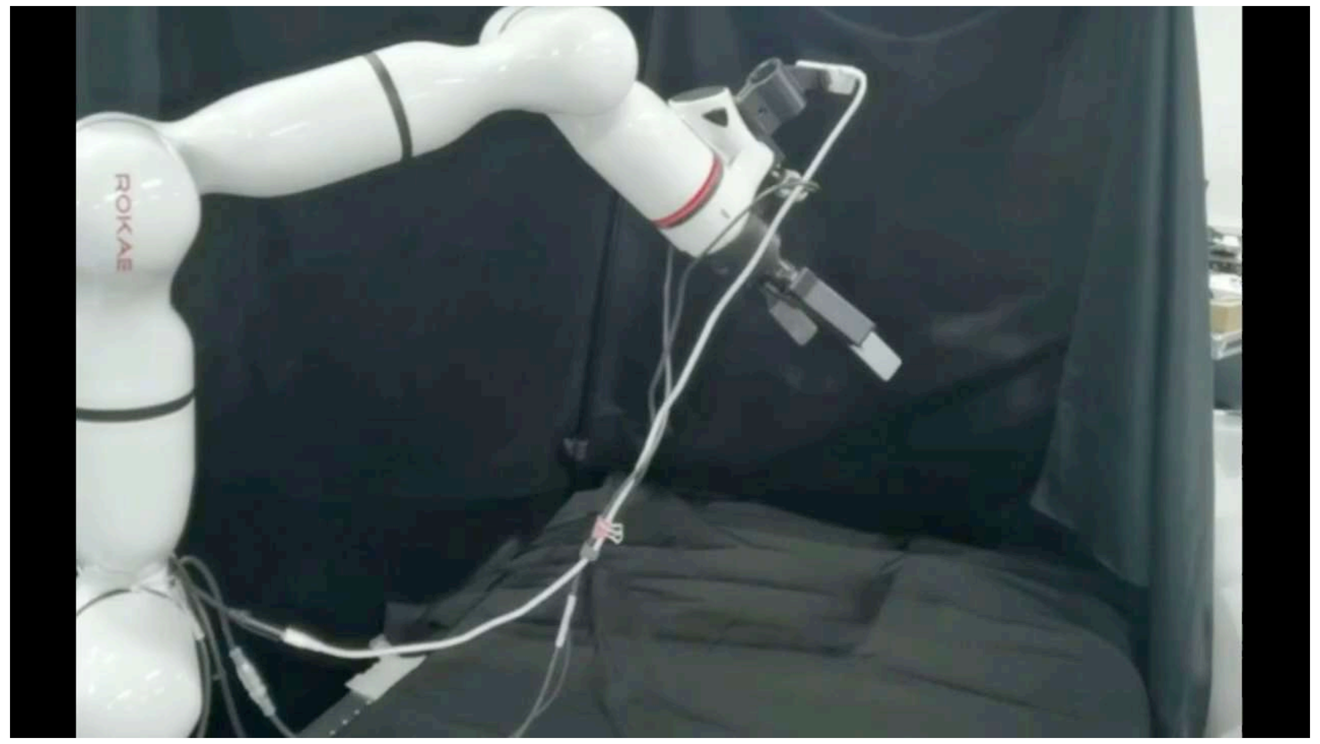
Safety Issues



A Sim-to-Real Solution

Learning in Sim → Deploying in Real

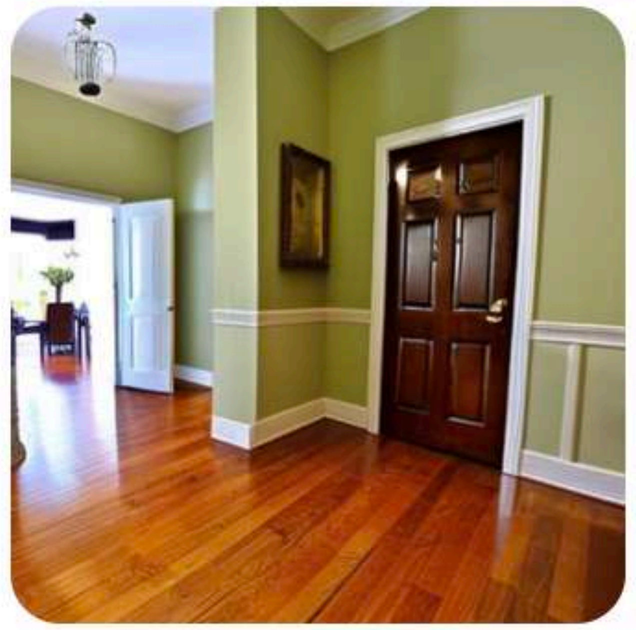




Lessons from Foundation Models

Imagine you are a robot agent in the house. There is a mobile phone on the dining table in the dining room, and you are in the living room. Write a short article about how would you walk through the house to grab the mobile phone with your robotic arm step by step. Include several high-quality photographs to illustrate your procedure.

As a robot agent in the house, I would walk through the house to grab the mobile phone in the dining room as shown in the image below.



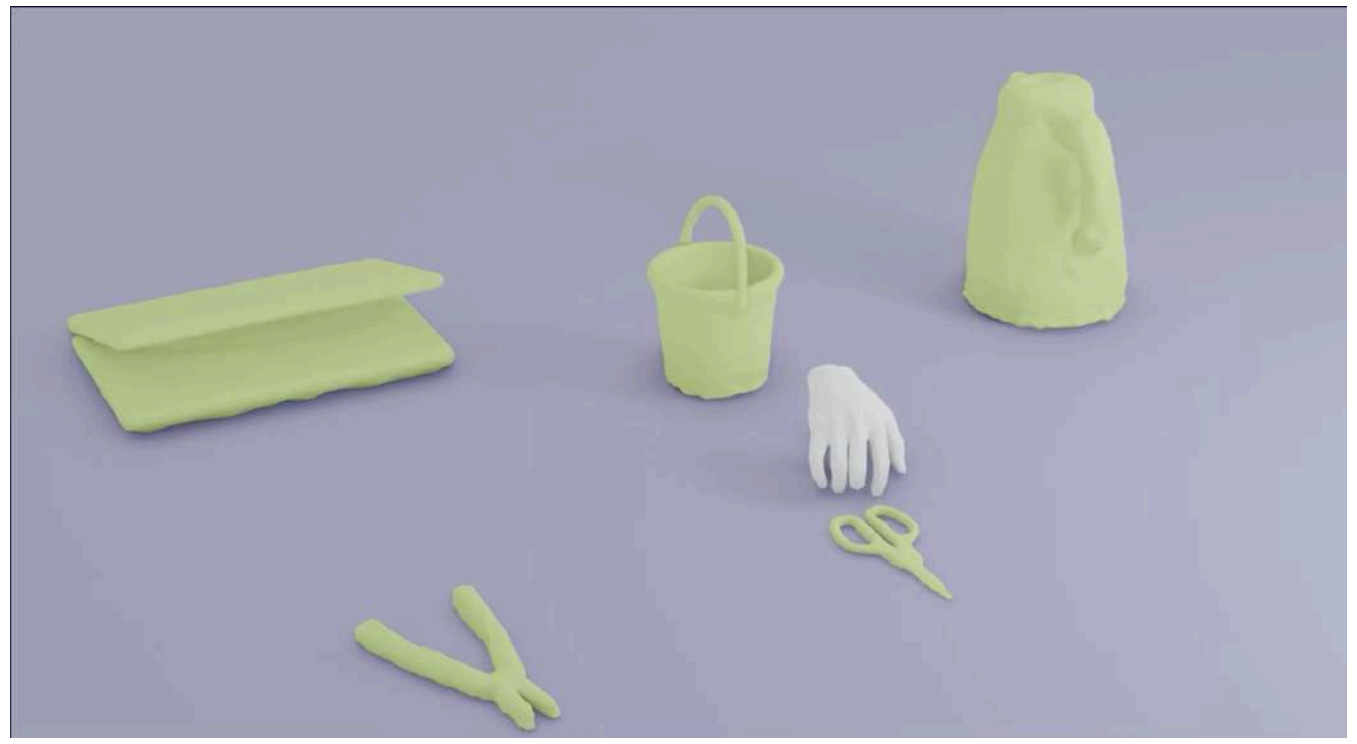
First, I would look around the living room to make sure the area is safe and there are no obstacles or people that I need to avoid. I would then walk towards the dining room and stop before entering the room.
Next, I would look for the mobile phone on the table, as shown in the image below.

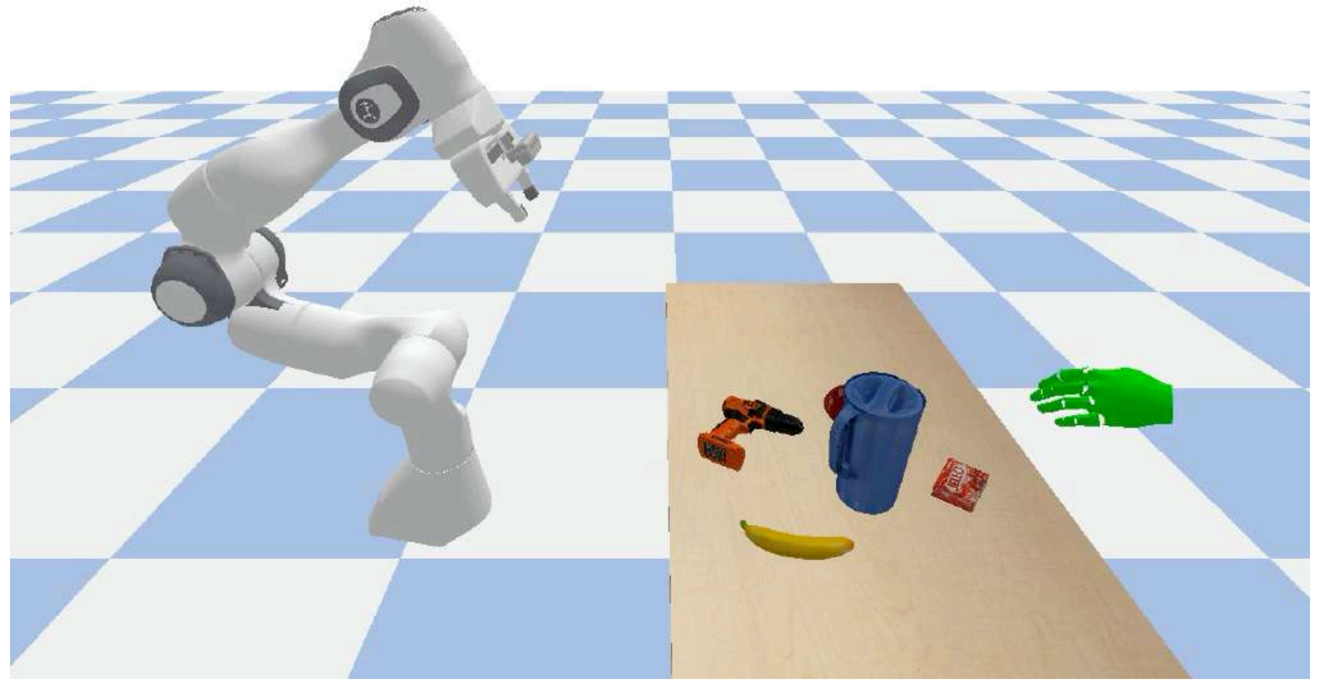
Once I have identified the phone, I would then move closer to it and extend my robot arm to grab it, as shown in the image below.

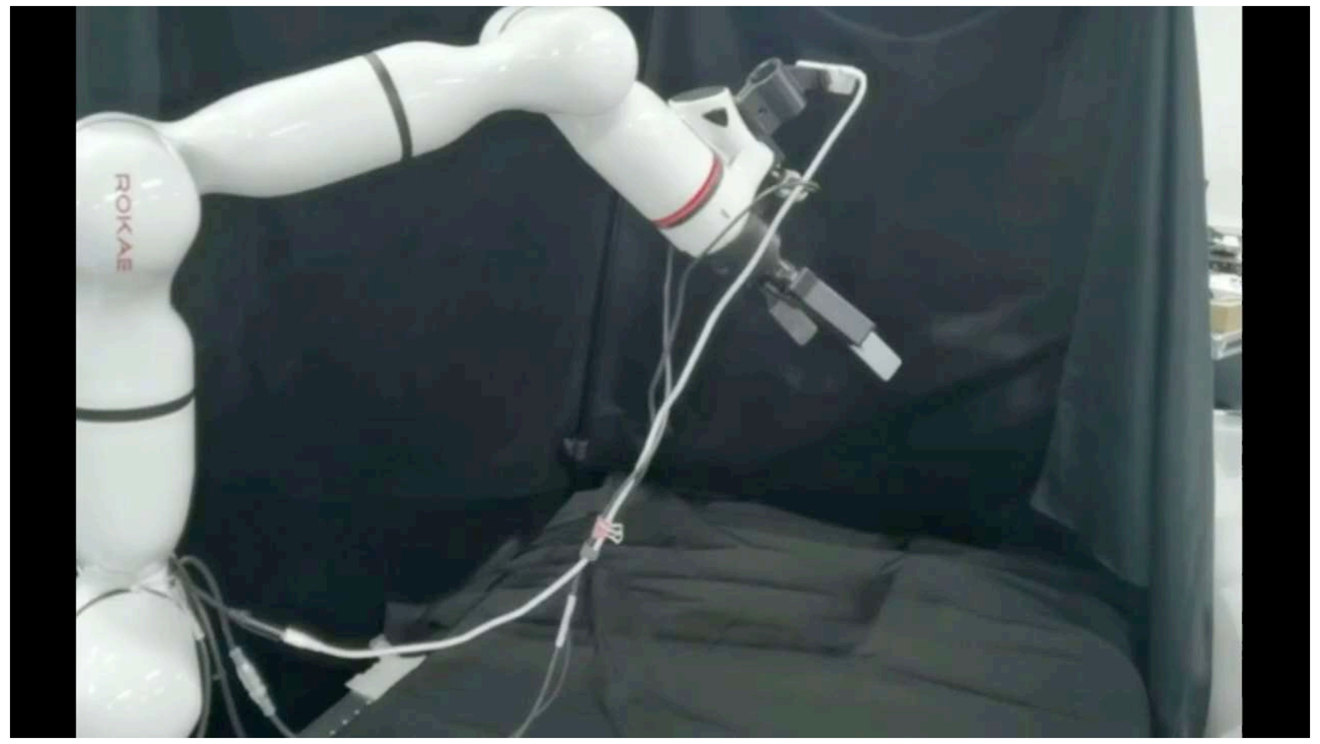
DreamLLM: Synergistic Multimodal Comprehension and Creation
Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, Li Yi. In submission

A Real-to-Sim-to-Real Solution

Creating Sim from Real → Learning in Sim → Deploying in Real







Human-Centered Robot Simulator

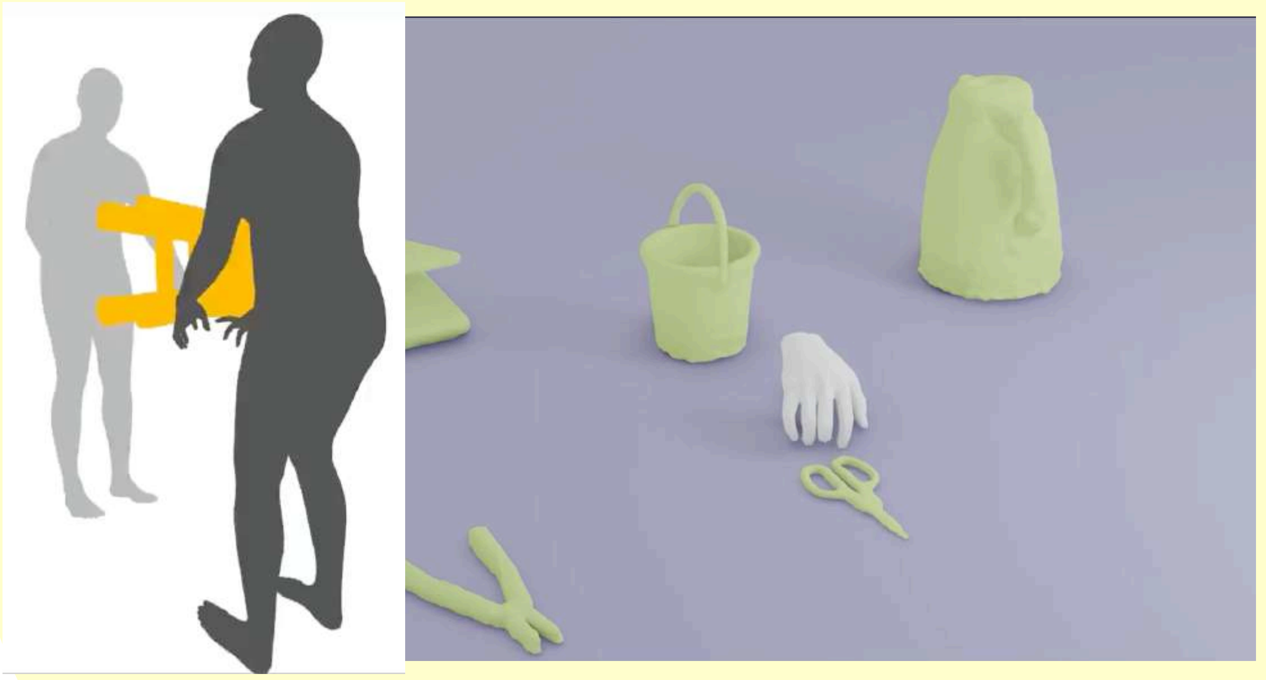
Human Interaction Capturing

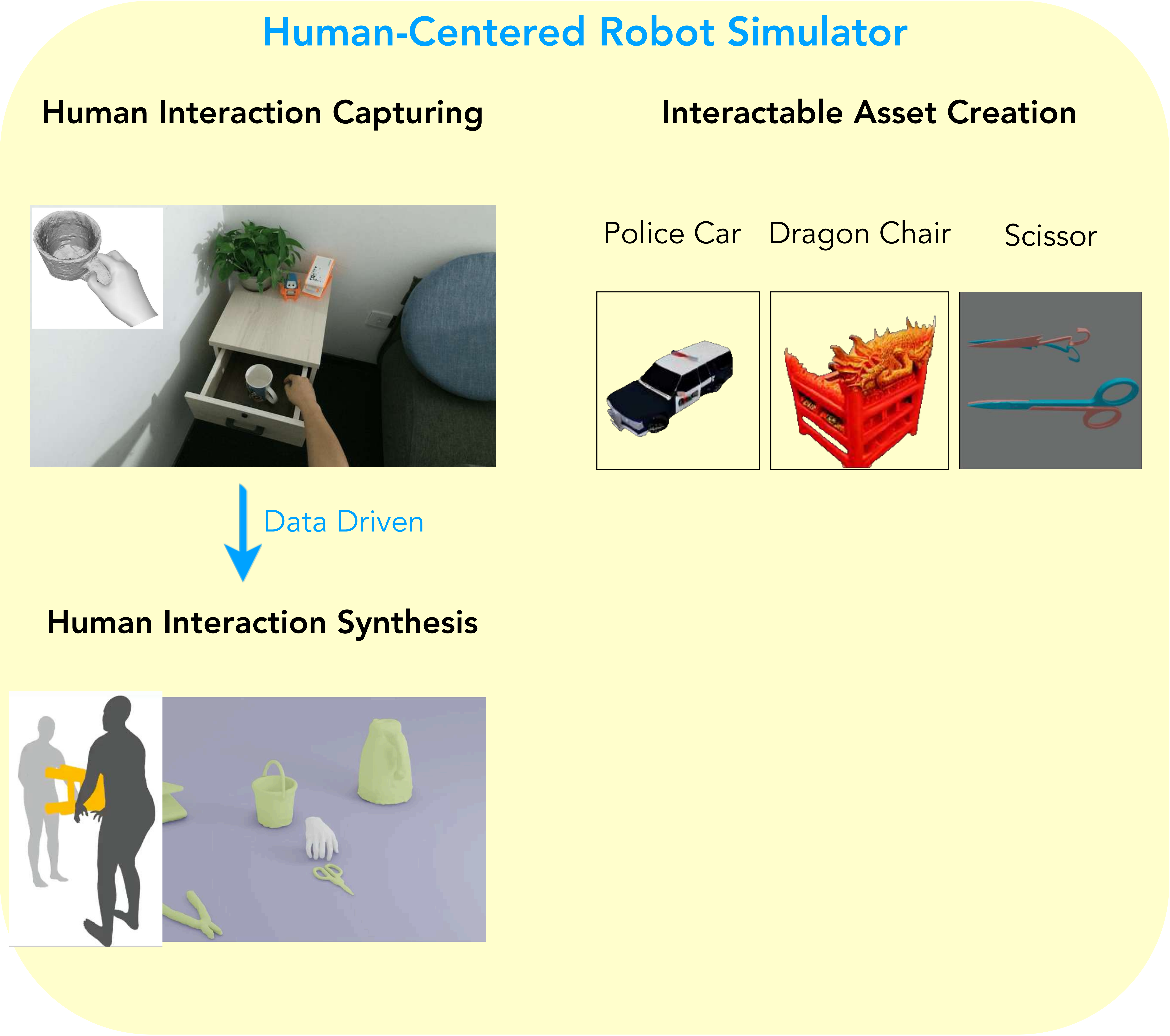
Human-Centered Robot Simulator

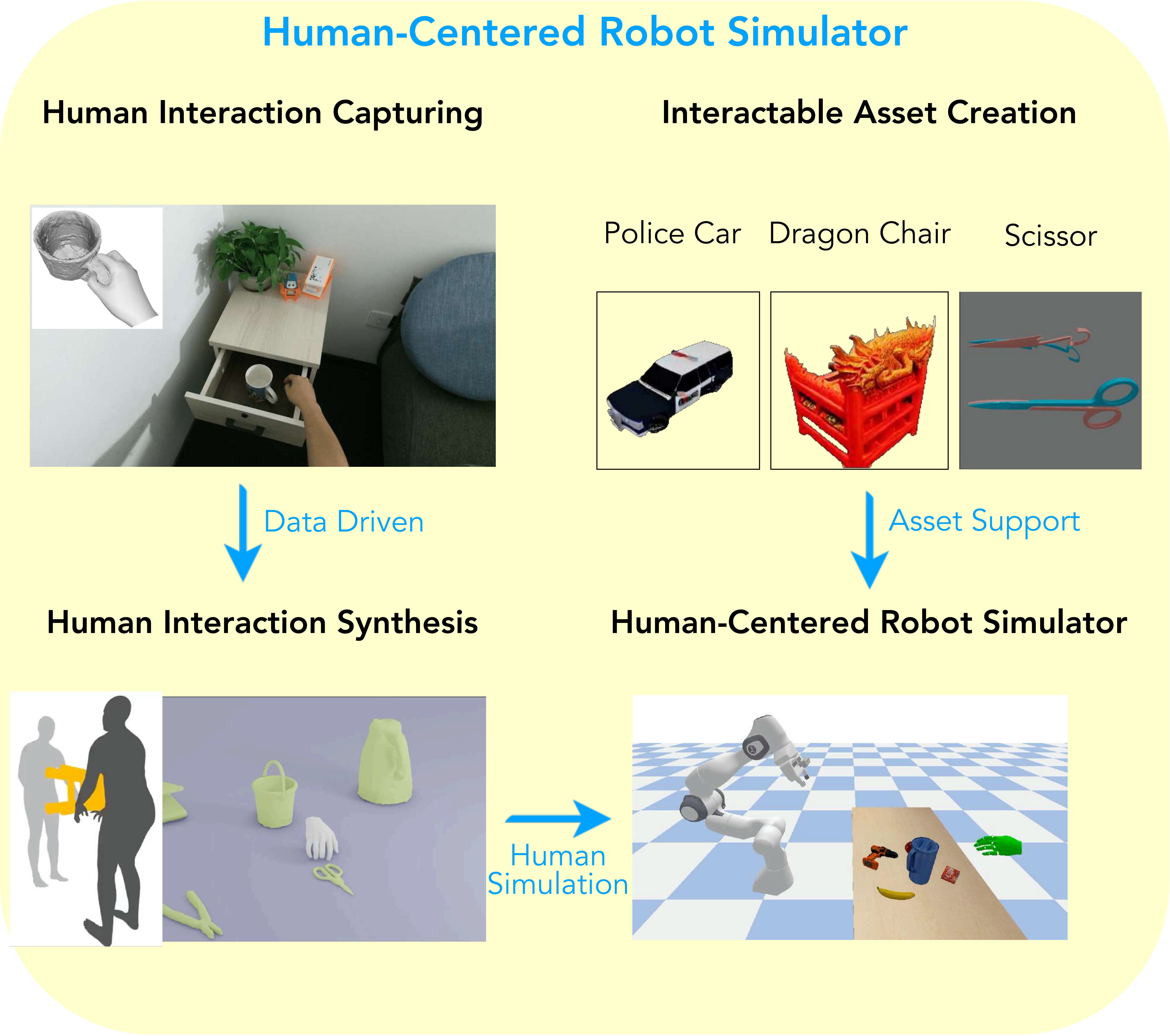
Human Interaction Capturing

Data Driven

Human Interaction Synthesis




Human-Centered Robot Simulator
Human Interaction Capturing
Interactable Asset Creation
Police Car
Dragon Chair
Scissor
Data Driven
Human Interaction Synthesis

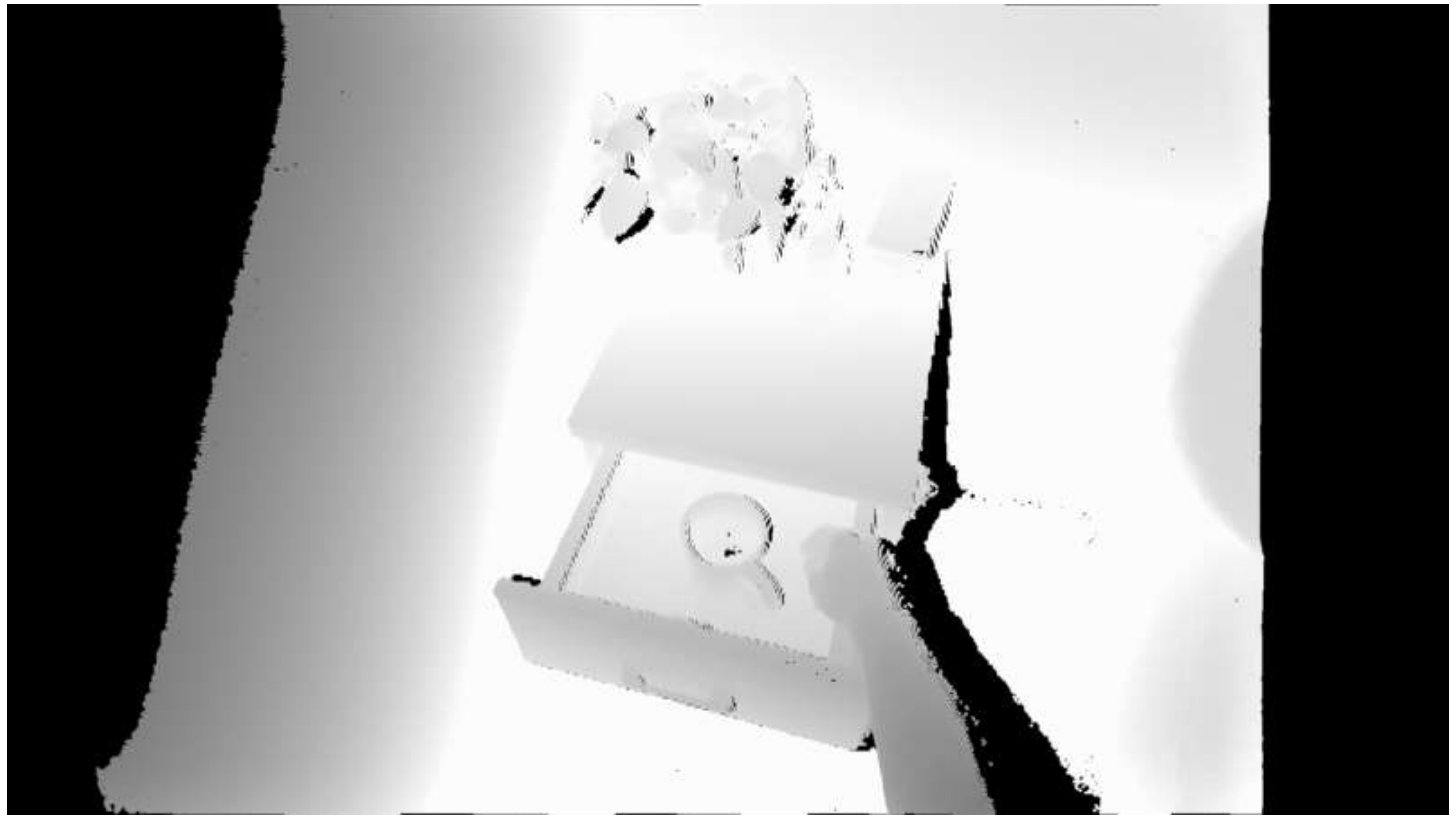

Human-Centered Robot Simulator
Human Interaction Capturing
Interactable Asset Creation
Police Car
Dragon Chair
Scissor
Data Driven
Asset Support
Human Interaction Synthesis
Human-Centered Robot Simulator
Human Simulation

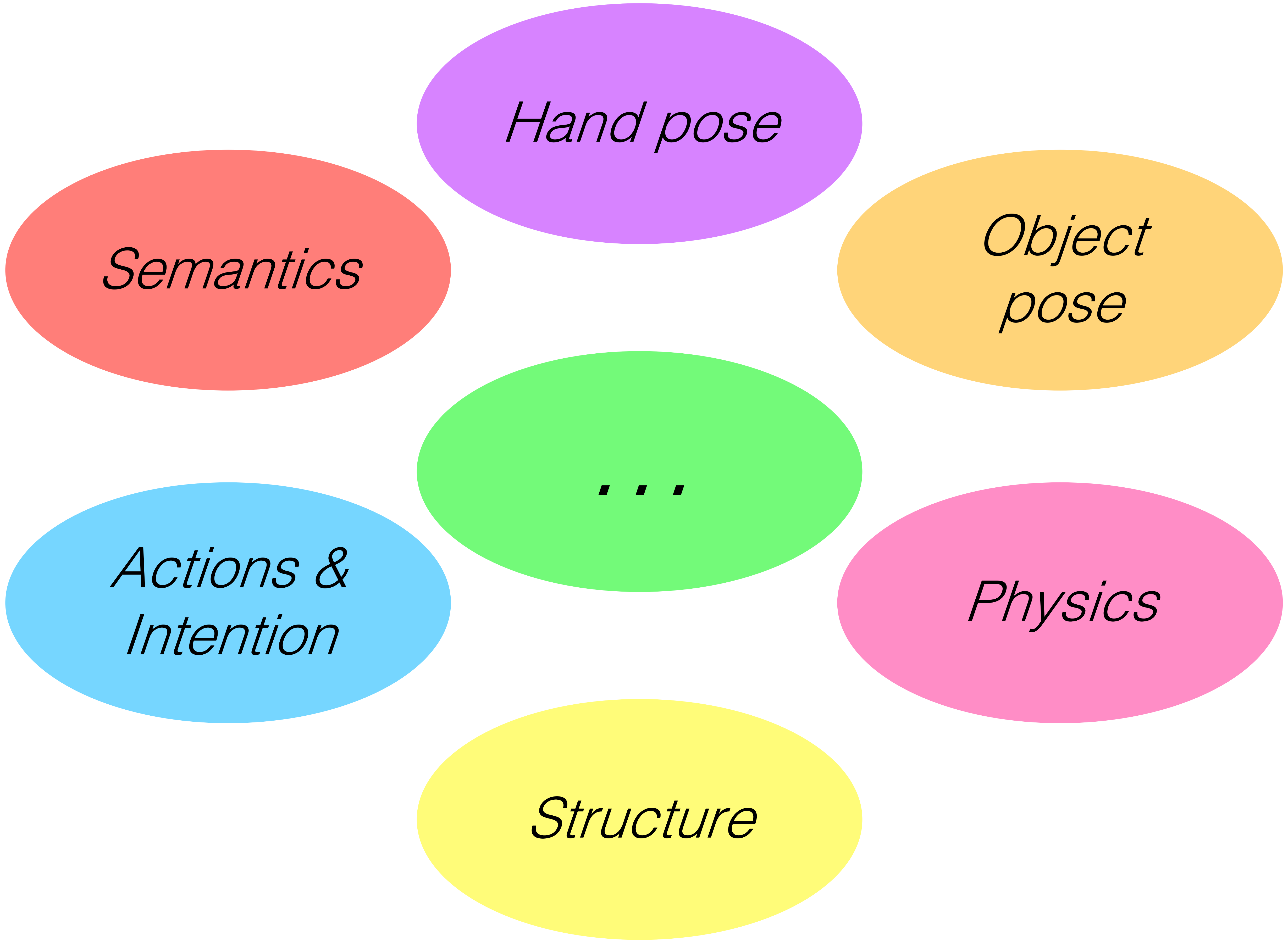
Human-Centered Robot Simulator
Human Interaction Capturing
Interactable Asset Creation
Police Car
Dragon Chair
Scissor
Data Driven
Asset Support
Human Interaction Synthesis
Human-Centered Robot Simulator
Human
Simulation
Human-Centered EAI
Open-World Perception


Human-Centered Robot Simulator
Human-Centered EAI
Human Interaction Capturing
Interactable Asset Creation
Open-World Perception
Police Car
Dragon Chair
Scissor
Data Driven
Asset Support
Visual Perception
Human Interaction Synthesis
Human-Centered Robot Simulator
Human-Centered Robotics
Human Simulation
Simulation Support
Collaborative Transport
Human-to-Robot Handover

Human-Centered Robot Simulator
Human-Centered EAI
Human Interaction Capturing
Interactable Asset Creation
Open-World Perception
Police Car
Dragon Chair
Scissor
Data Driven
Asset Support
Visual Perception
Human Interaction Synthesis
Human-Centered Robot Simulator
Human-Centered Robotics
Human Simulation
Simulation Support
Collaborative Transport
Human-to-Robot Handover

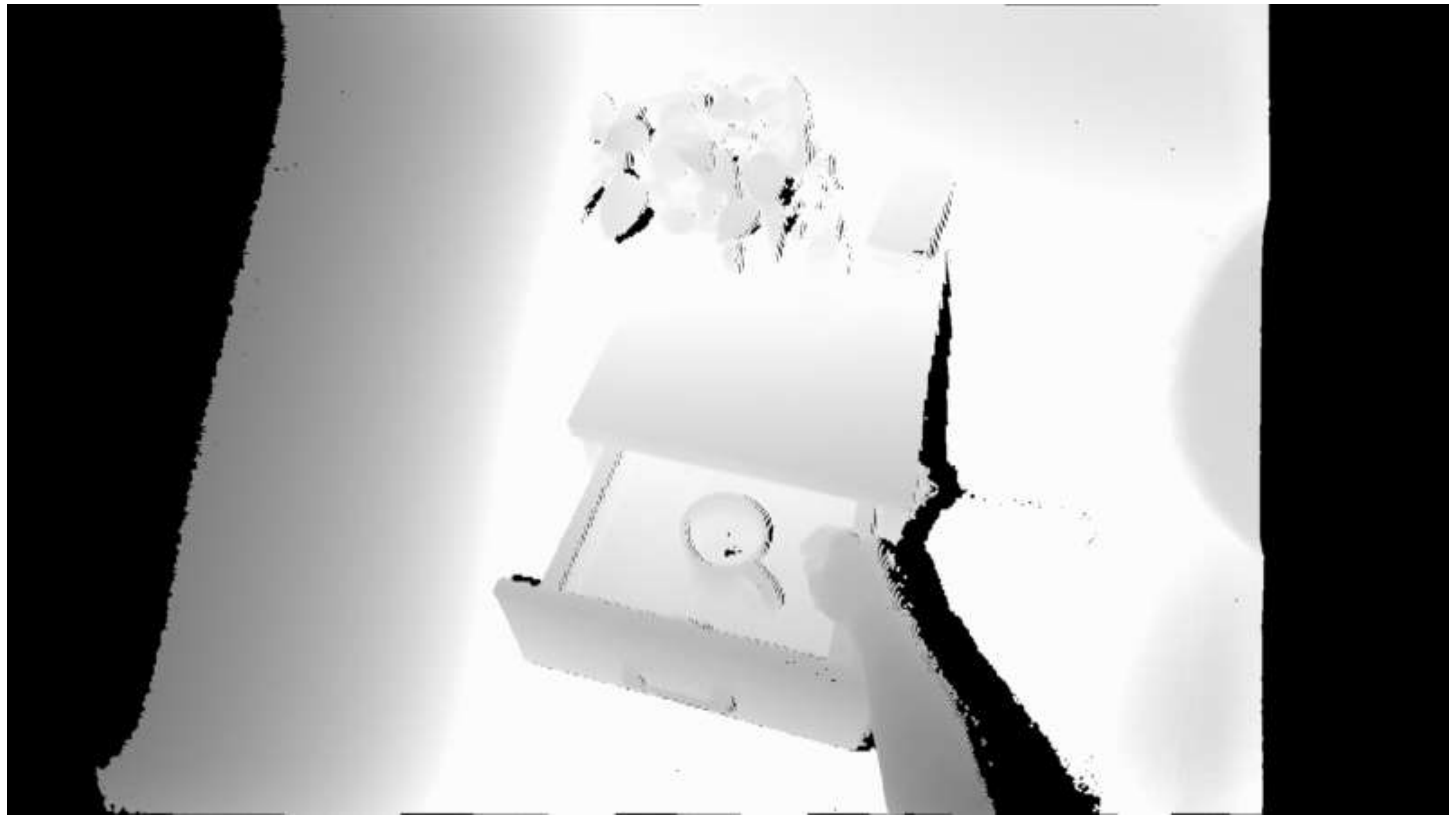
Egocentric Perception in HOI





HOI4D Dataset

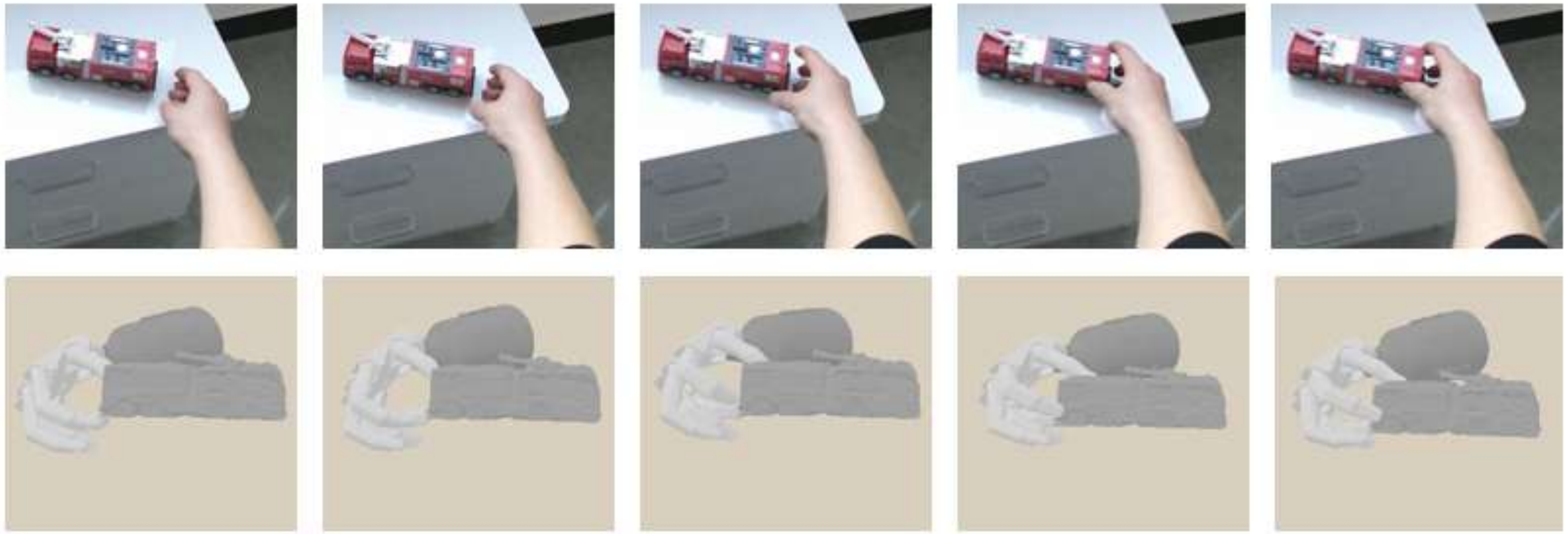
- The first dataset for 4D egocentric category-level human-object interaction



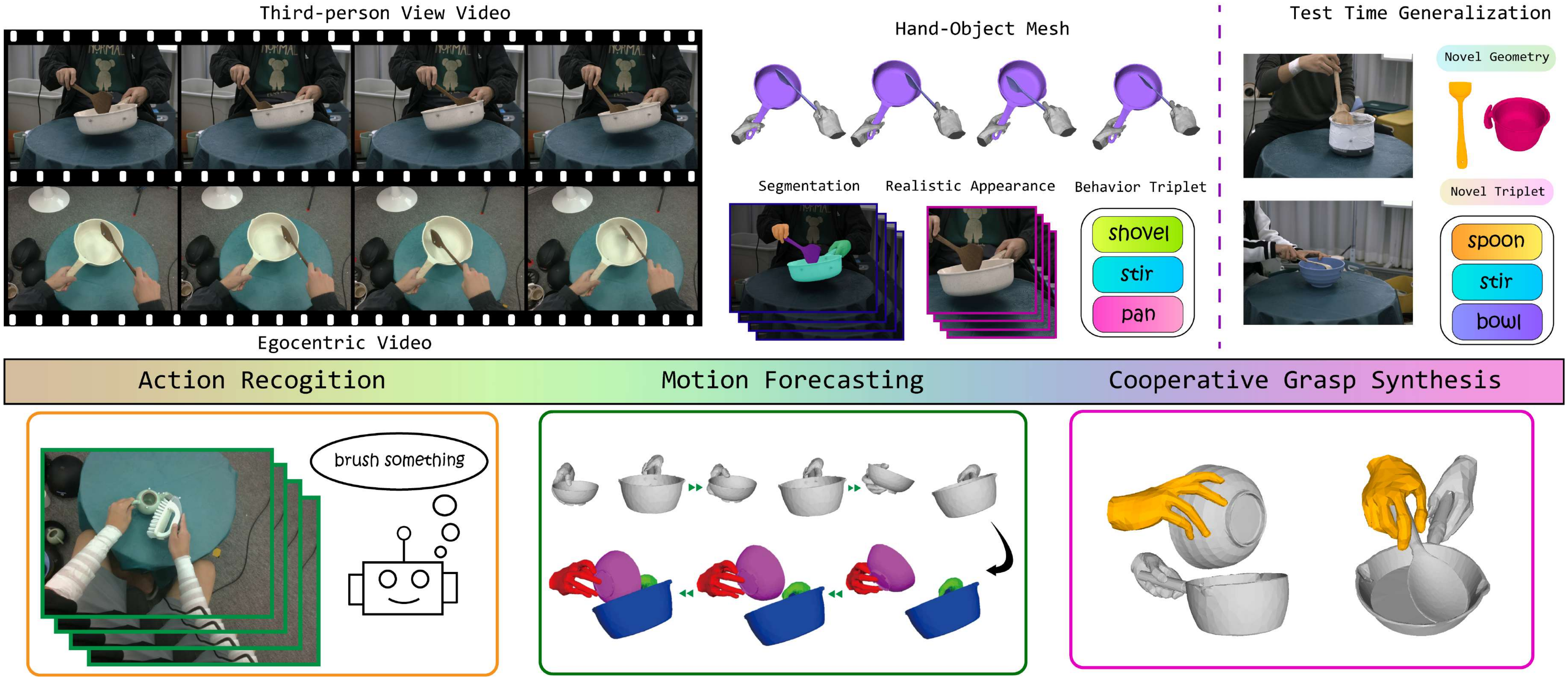
HOI4D: A 4D Egocentric Dataset for Category-Level Human-Object Interaction
Yunze Liu, Yun Liu*, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, Li Yi. CVPR 2022*

Application - Robot Learning from Human Demonstration

- Learning robotic dexterous manipulation from human demonstration



TACO: Bimanual Tool-ACtion-Object Understanding



TACO: Benchmarking Generalizable Bimanual Tool-ACtion-Object Understanding

Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, Li Yi. CVPR 2024

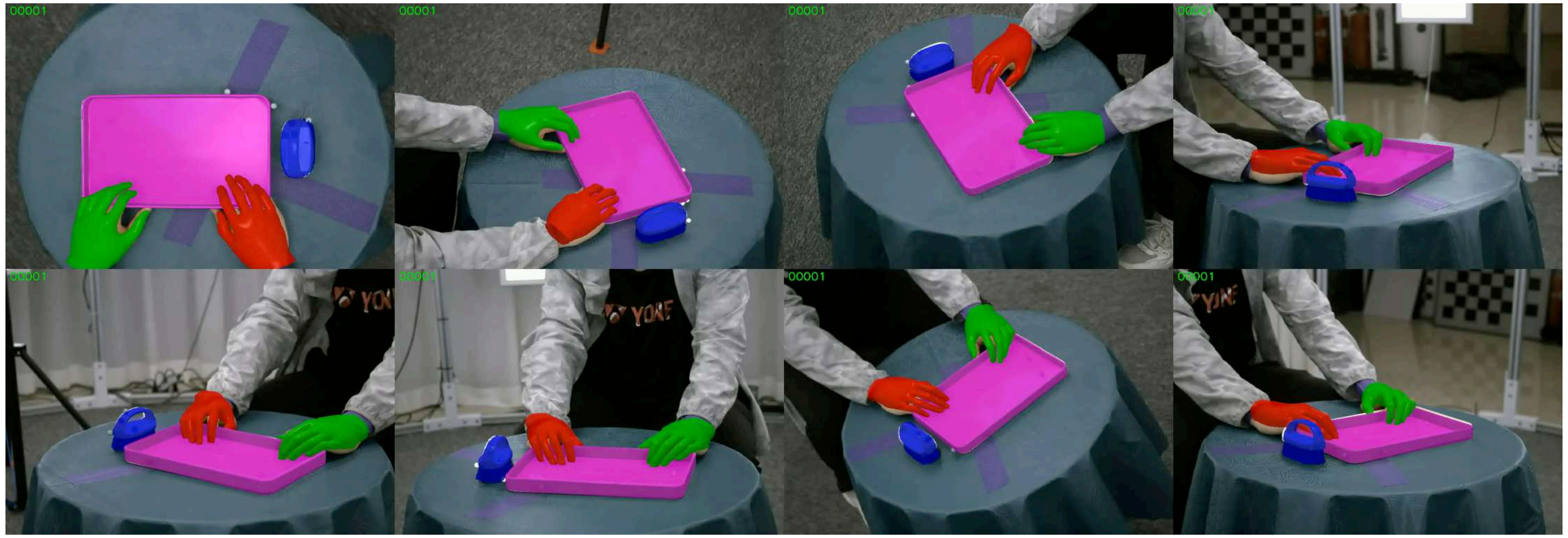
TACO: Bimanual Tool-ACtion-Object Understanding

- **2,500 Hand-object Manipulation Sequences**
- **131 Tool-ACtion-Object Compositions**
- **5,200,000 RGB Frames**
- **20 Object Categories**
- **196 Object Models**
- **15 Actions**
- **14 Actors**
- **Automatic Annotation**

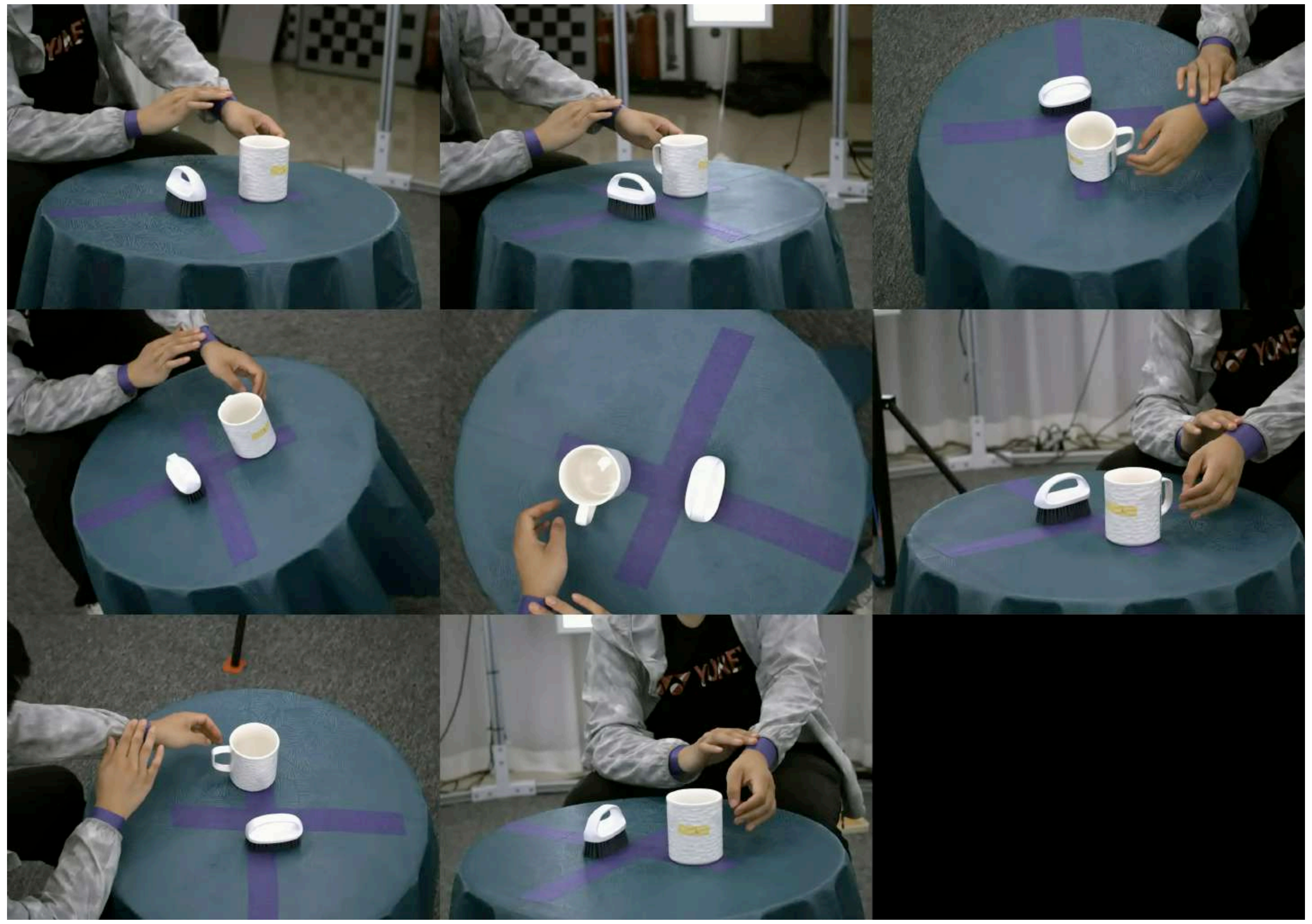
TACO: Bimanual Tool-ACtion-Object Understanding

automatic data aquizition pipeline

TACO: Bimanual Tool-ACtion-Object Understanding



4D hand-object mesh sequences



realistic hand-object appearances

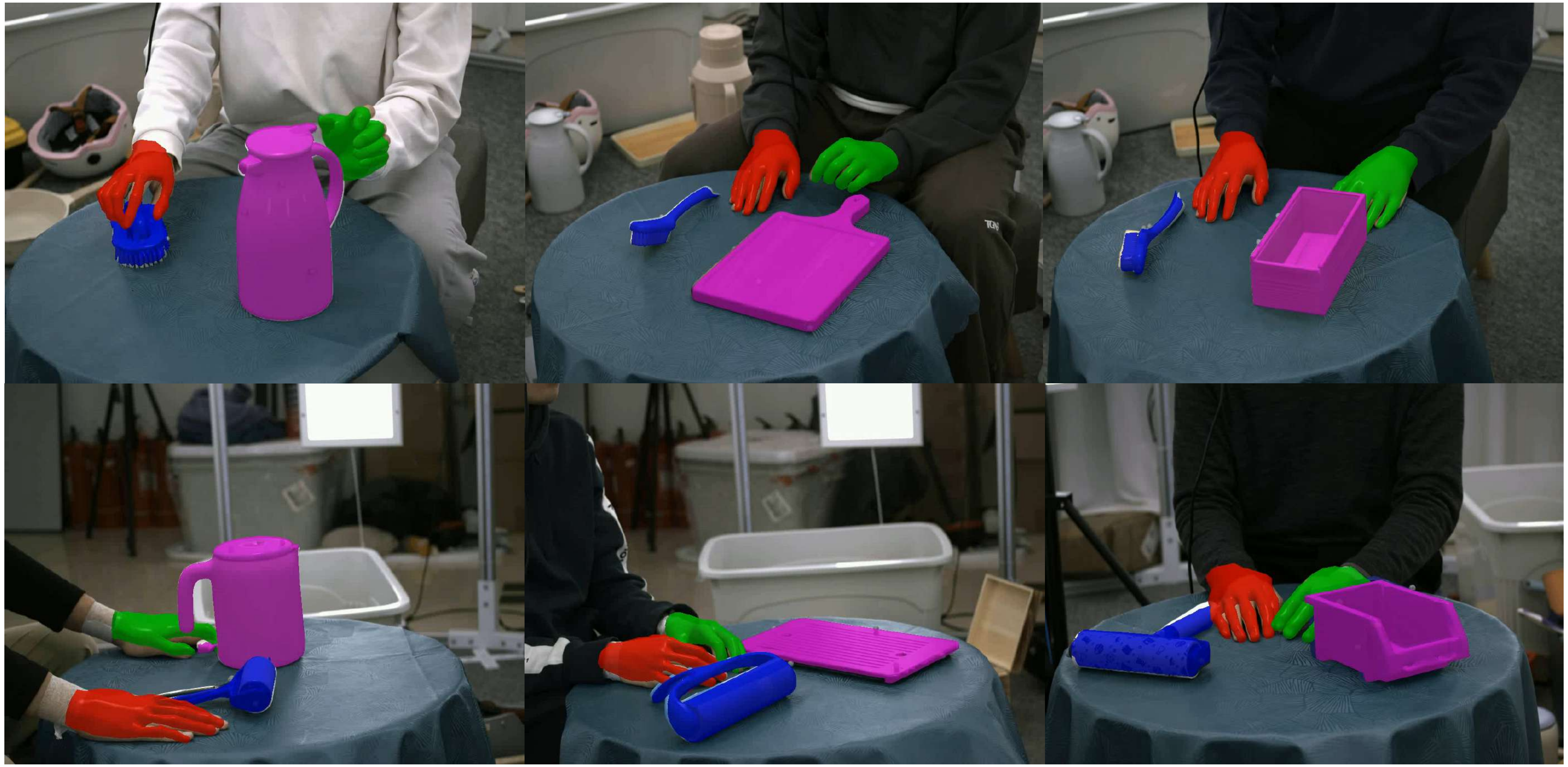
TACO Diversity

dust

kettle | plate | box

brush

roller



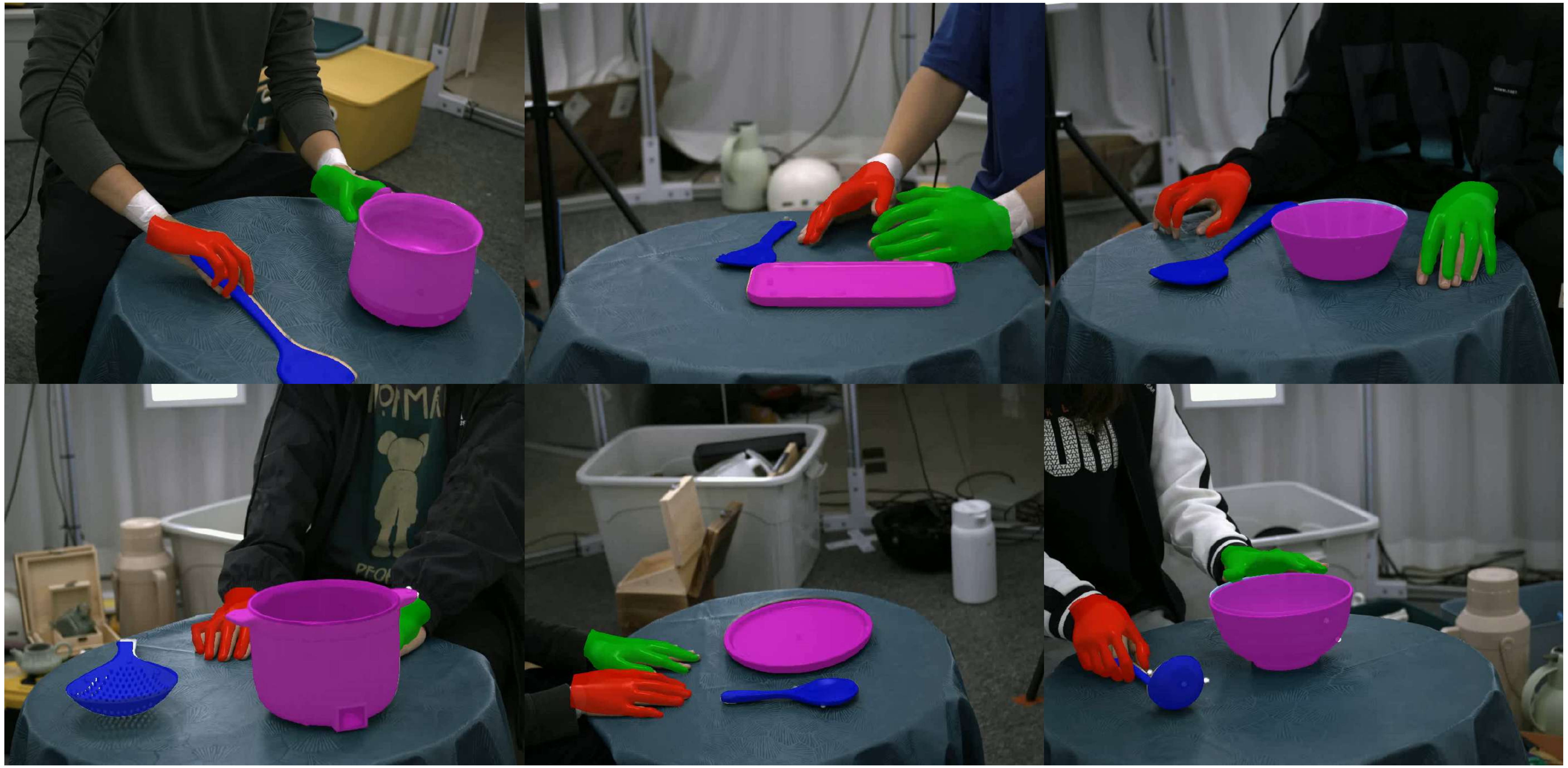
TACO Diversity

Stir

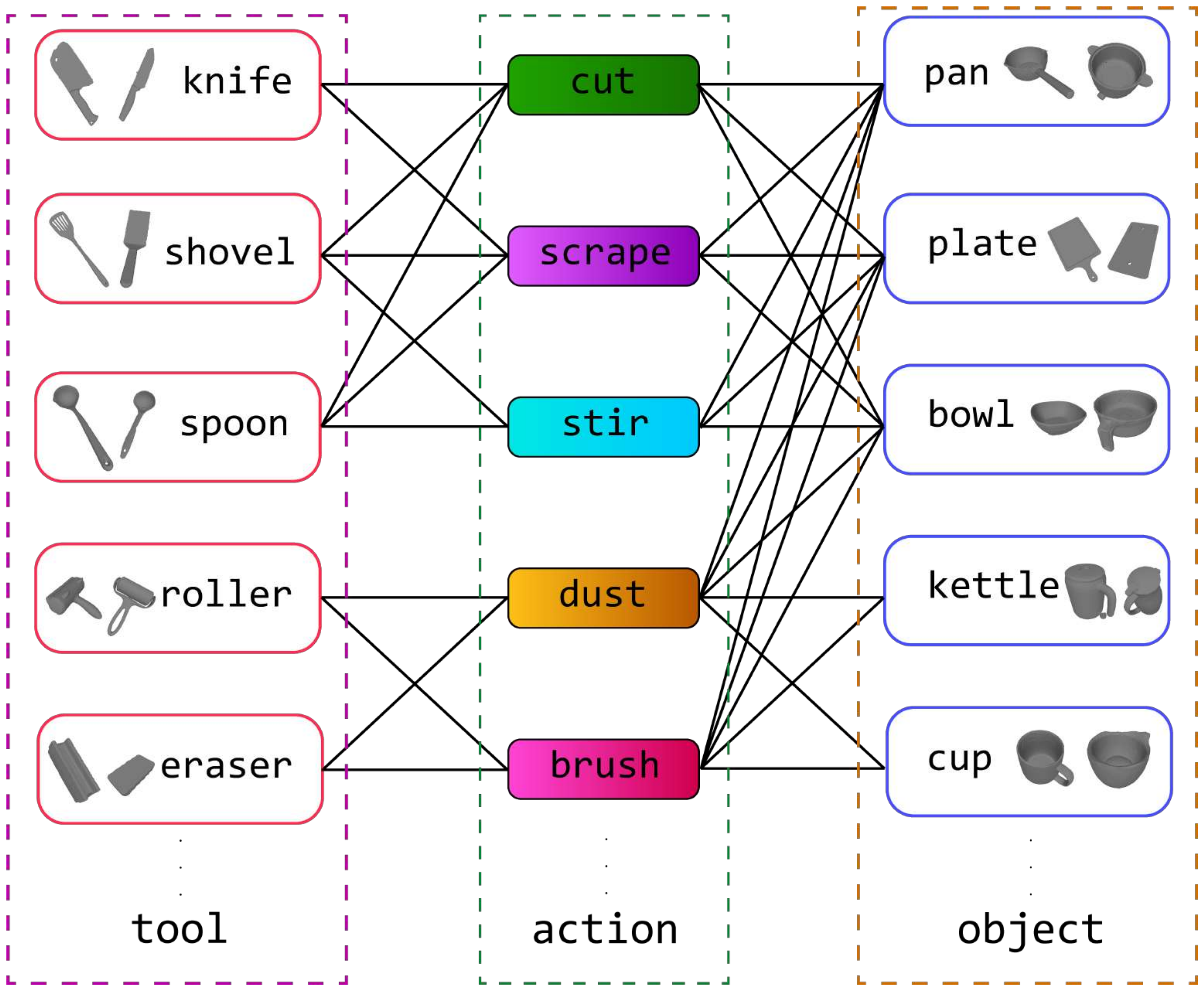
pan plate bowl

shovel

spoon



TACO: Bimanual Tool-ACtion-Object Understanding

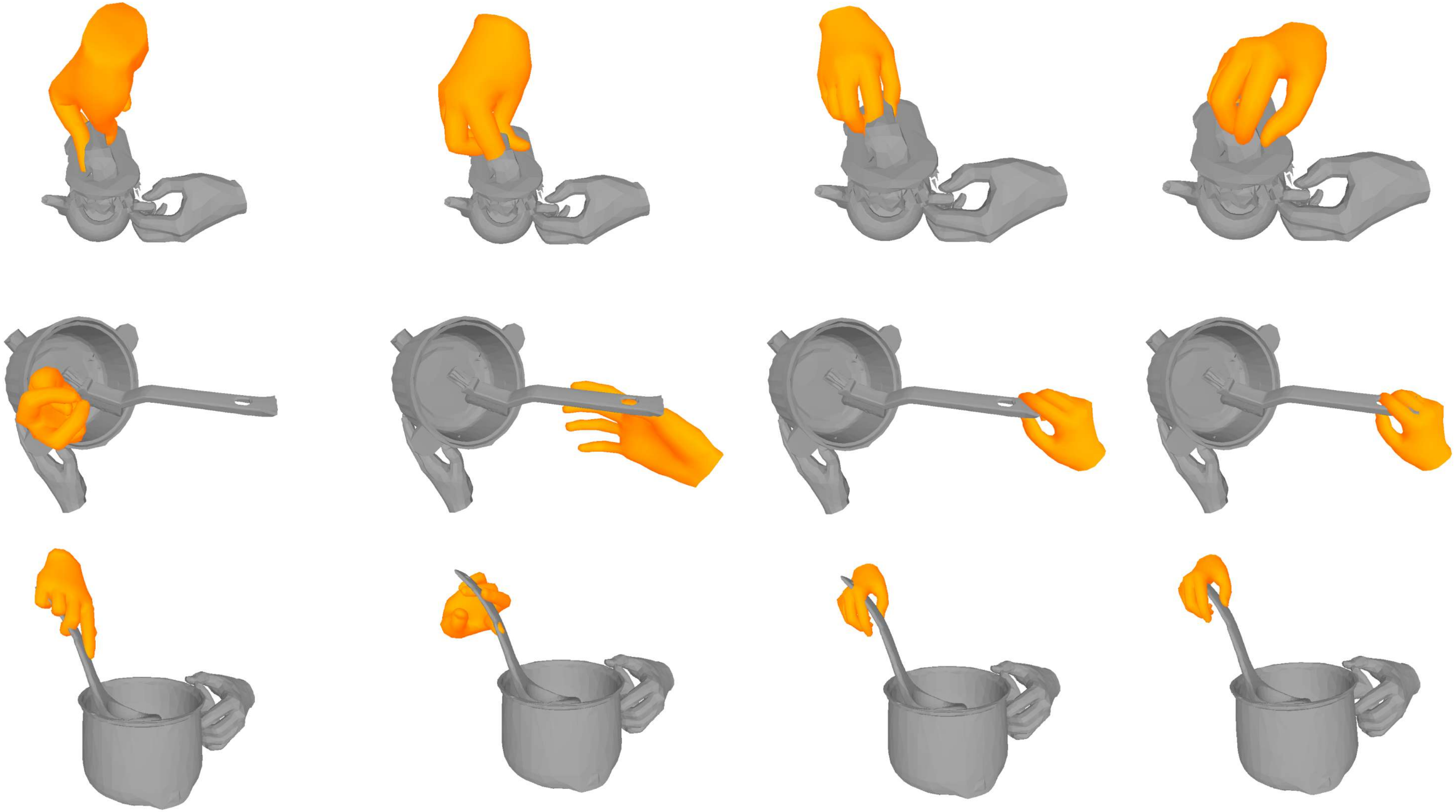


correlated interaction triplets

Supporting different generalization purposes:

- test set 1: no generalization
- test set 2: geometry generalization
- test set 3: triplet generalization
- test set 4: compound generalization: novel tool category

Application – Cooperative Grasp Synthesis



Social Behavior Capturing

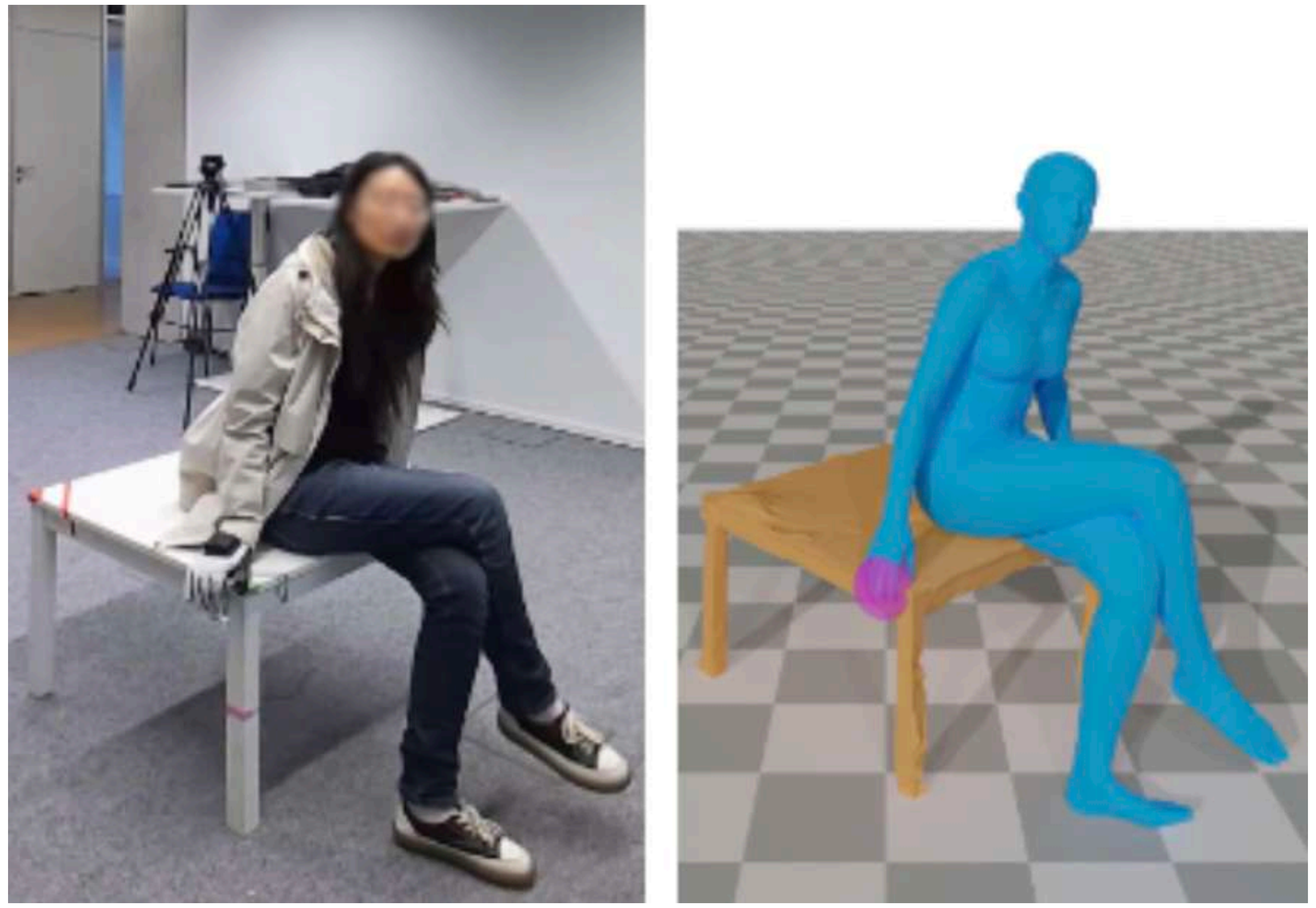
(a) allocentric camera views (b) human-object mesh seqnences (c) corresponding 2D masks

CORE4D: A 4D Human-Object-Human Interaction Dataset for Collaborative Object REarrangement

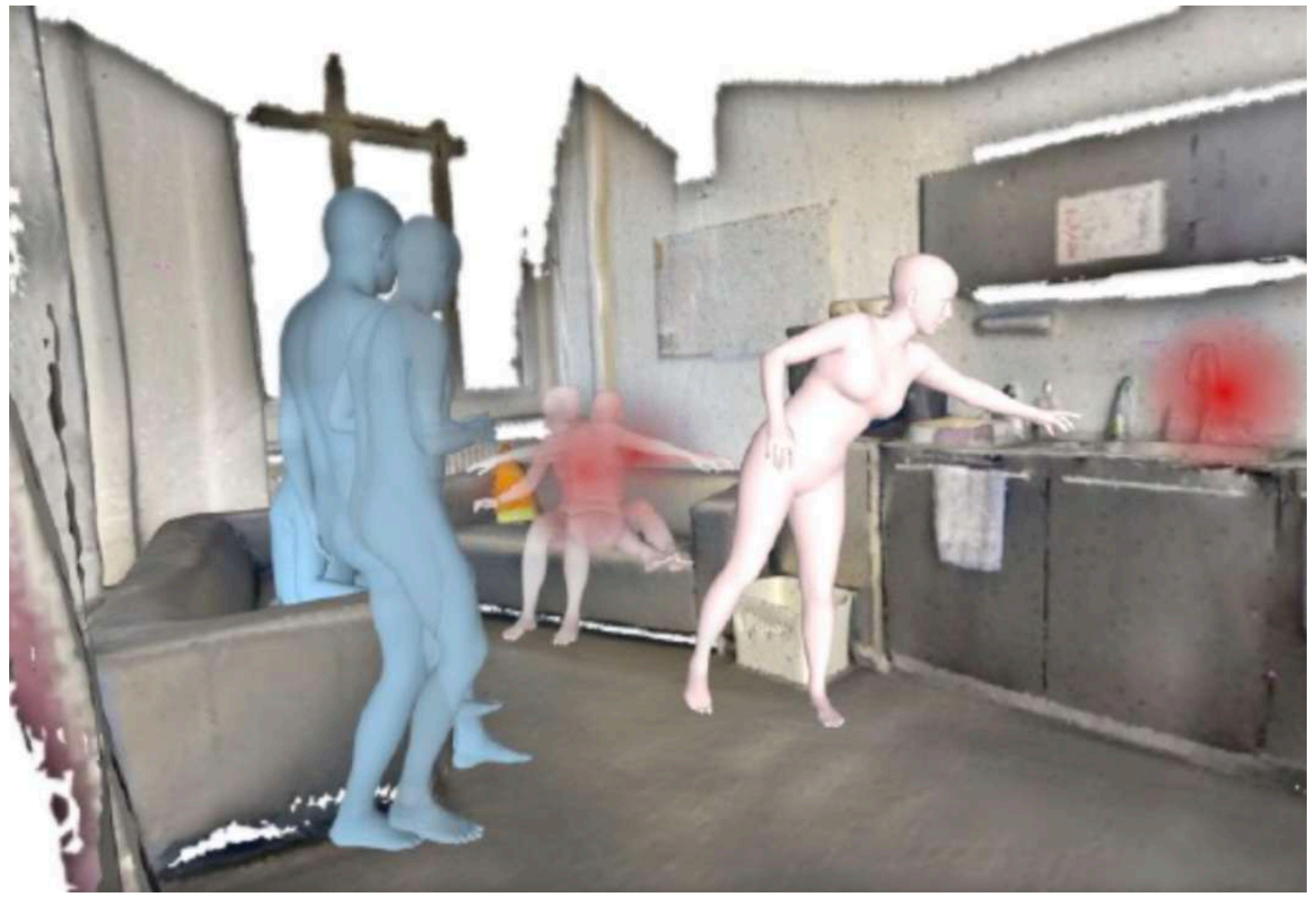
Chengwen Zhang, Yun Liu, Ruofan Xing, Bingda Tang, Li Yi. In submission


Human-Centered Robot Simulator
Human-Centered EAI
Human Interaction Capturing
Interactable Asset Creation
Open-World Perception
Police Car
Dragon Chair
Scissor
Data Driven
Asset Support
Visual Perception
Human Interaction Synthesis
Human-Centered Robot Simulator
Human-Centered Robotics
Human Simulation
Simulation Support
Collaborative Transport
Human-to-Robot Handover

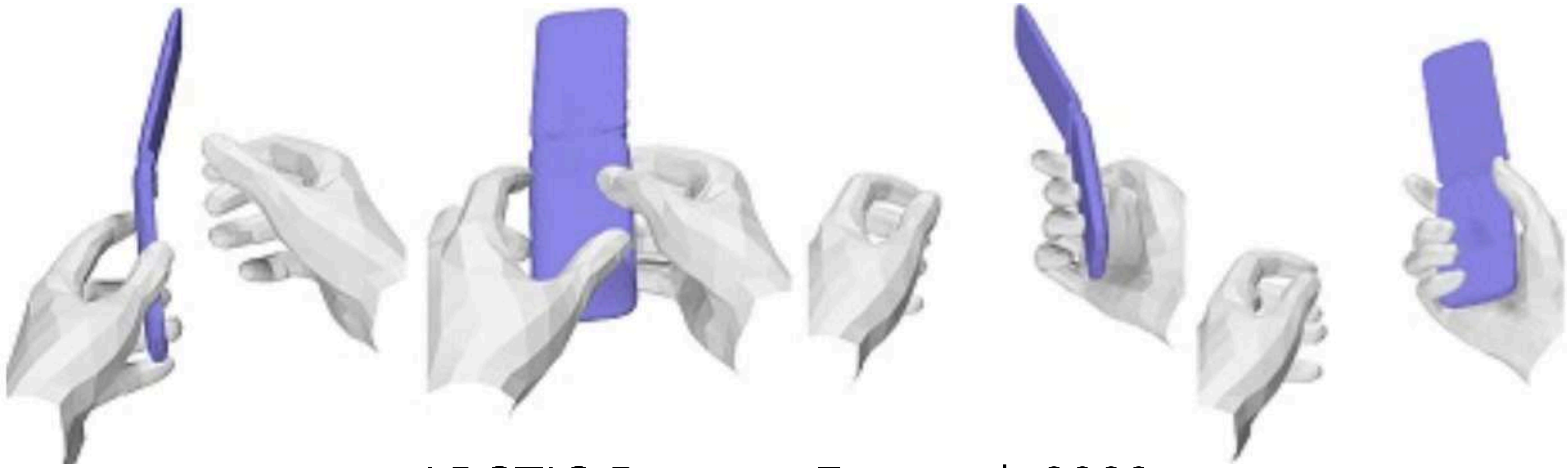
Interaction Synthesis on Different Levels



BEHAVE Dataset, Bhatnagar et al. 2022



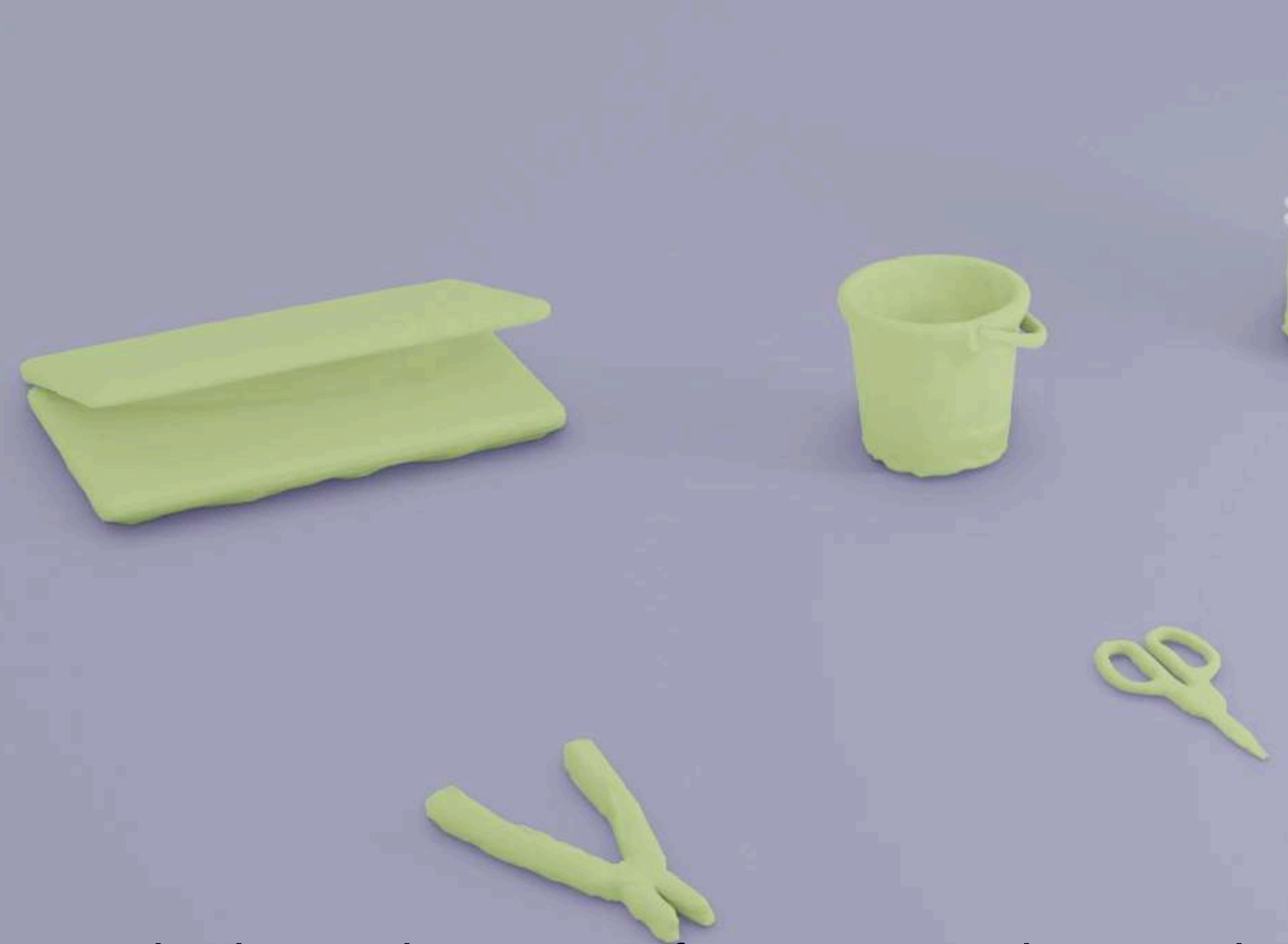
EgoBody Dataset, Zhang et al. 2022



ARCTIC Dataset, Fan et al. 2023

HOI4D, Liu et al. 2022

Synthesized result of proposed CAMS framework

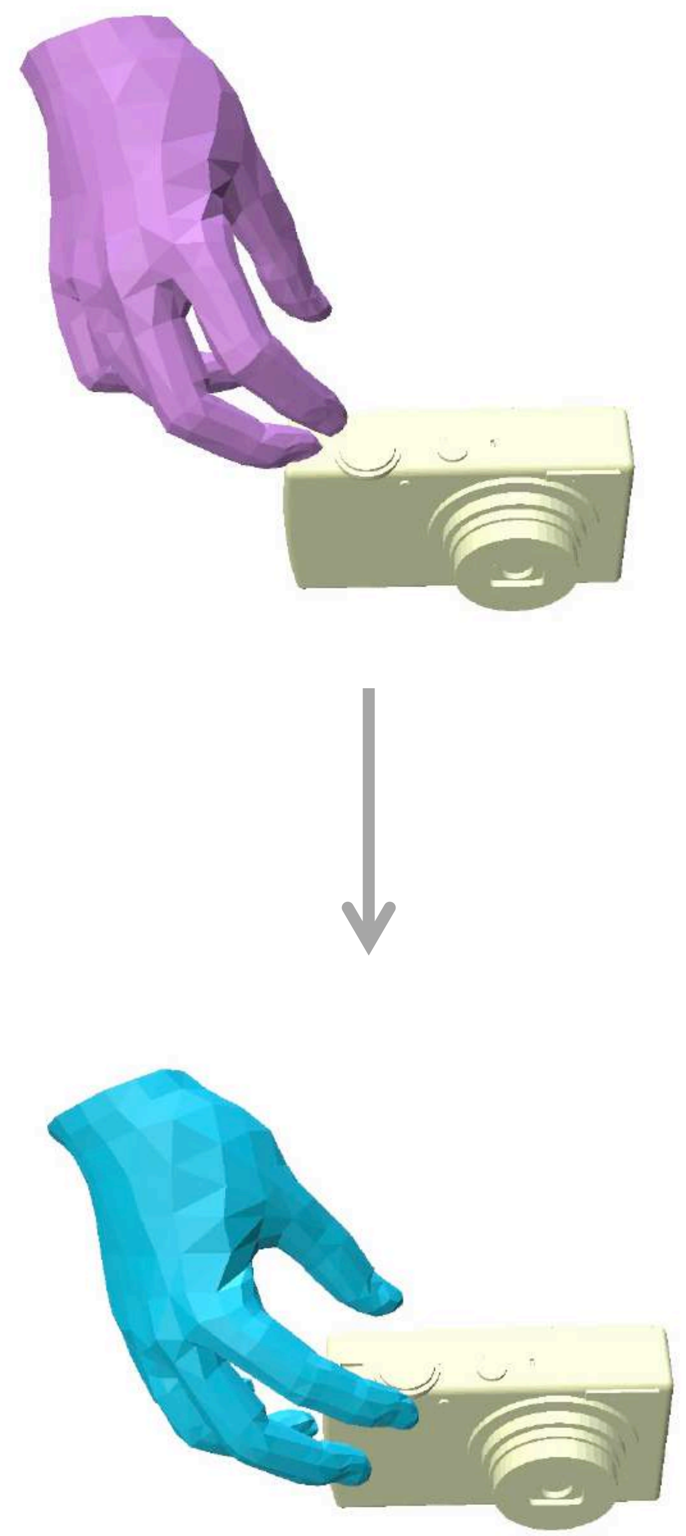


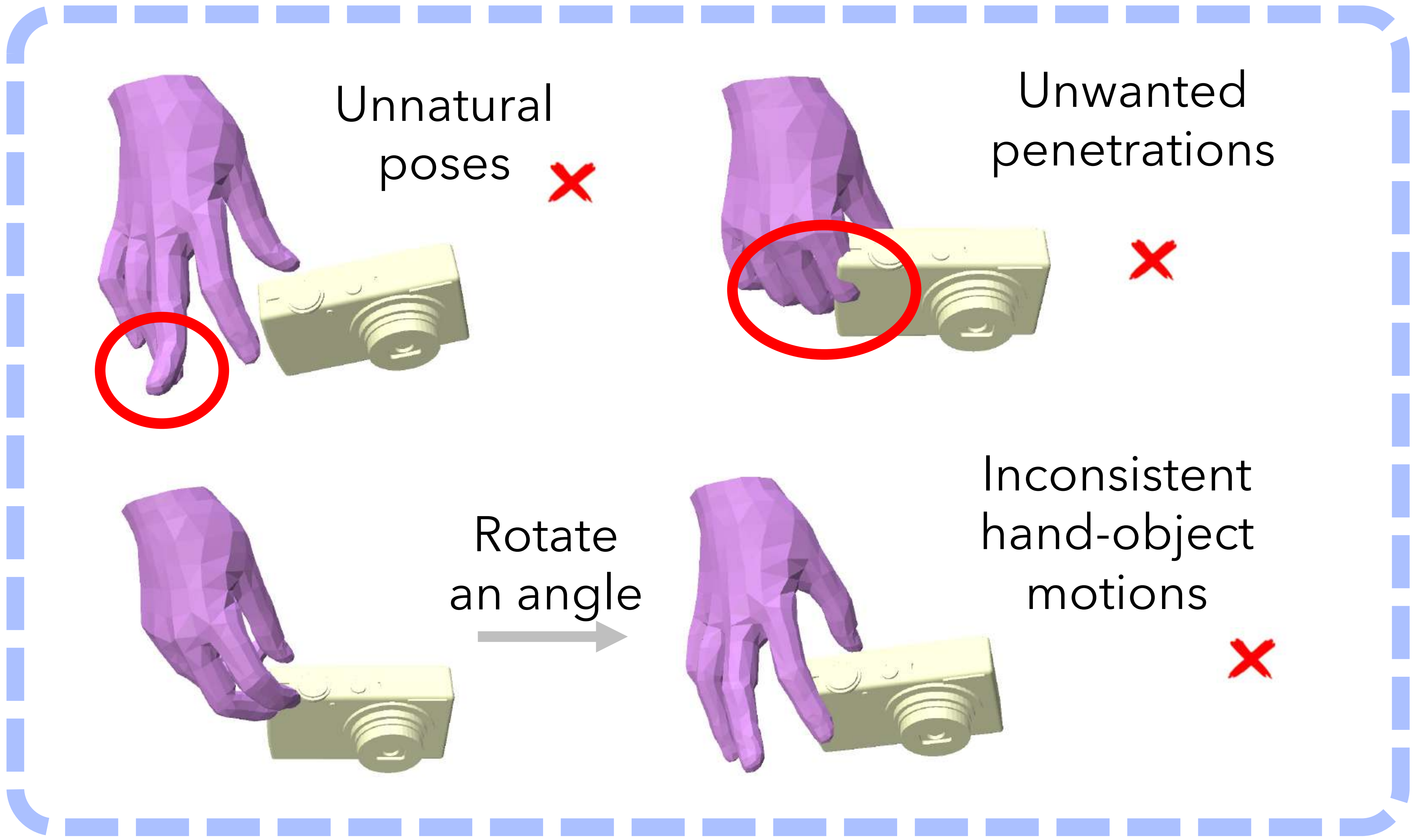
CAMS: CAnonicalized Manipulation Spaces for Category-Level Functional Hand-Object Manipulation Synthesis

Juntian Zheng, Qingyuan Zheng*, Lixing Fang*, Yun Liu, Li Yi. CVPR 2023*

HOI Synthesis and Denoising

Synthesizing ***high-quality HOI data*** is challenging:
Numerous factors can result in ***heterogeneous*** and ***complex*** interaction noise





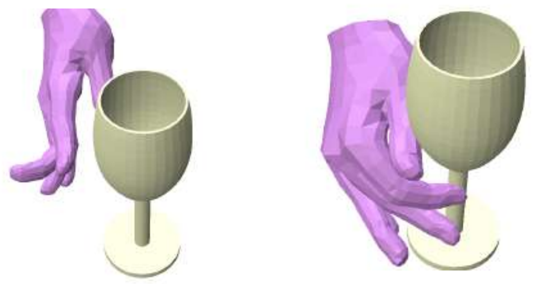
*Can we refine such noisy HOI data for **clean interaction sequences**?*

GeneOH Diffusion: Generalizable HOI Denoising

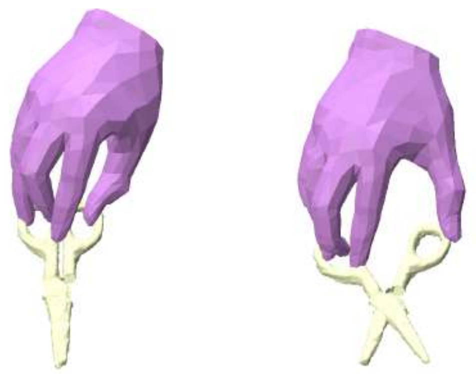
Data-Driven Denoising Model Trained on Paired Noisy-Clean Data

Noisy Data with
Synthetic Gaussian Noise

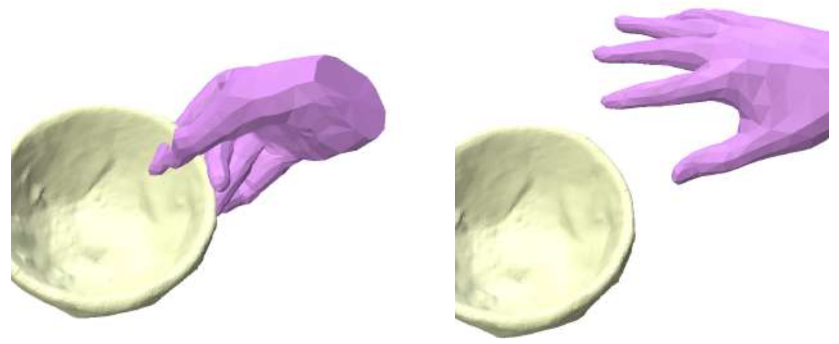
Clean Data



Unseen and difficult objects
(***thin shells***)



Challenging and new manipulations
(***articulated motion***)



Unpredictable and complex noise patterns
(***real noise***)

GeneOH Diffusion: Generalizable Hand-Object Interaction Denoising via Denoising Diffusion
Xueyi Liu, Li Yi. ICLR 2024

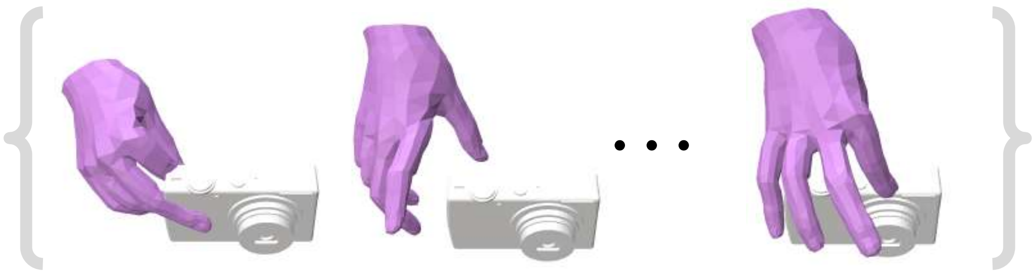
Key Idea: *GeneOH* and *Denoising via Diffusion*

Re-thinking the ***interaction representation*** design and the ***denoising scheme***

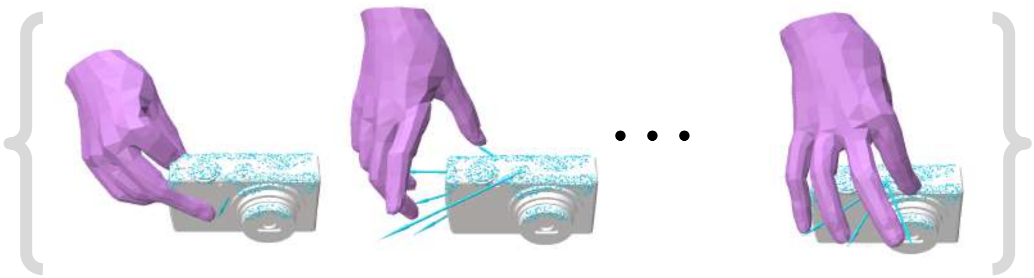
An Informative Generalizable HOI Representation

GeneOH

Focusing on ***local interaction regions*** for ***aligning different manipulations.***

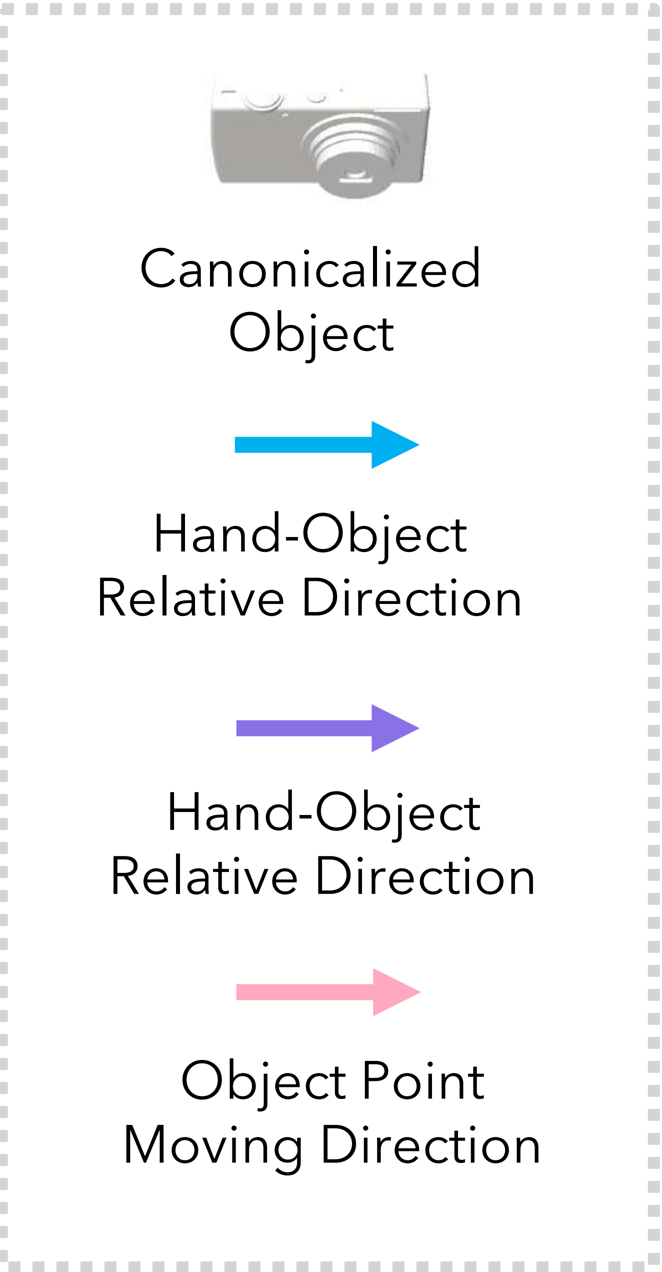


(a) Canonicalized Hand Trajectory



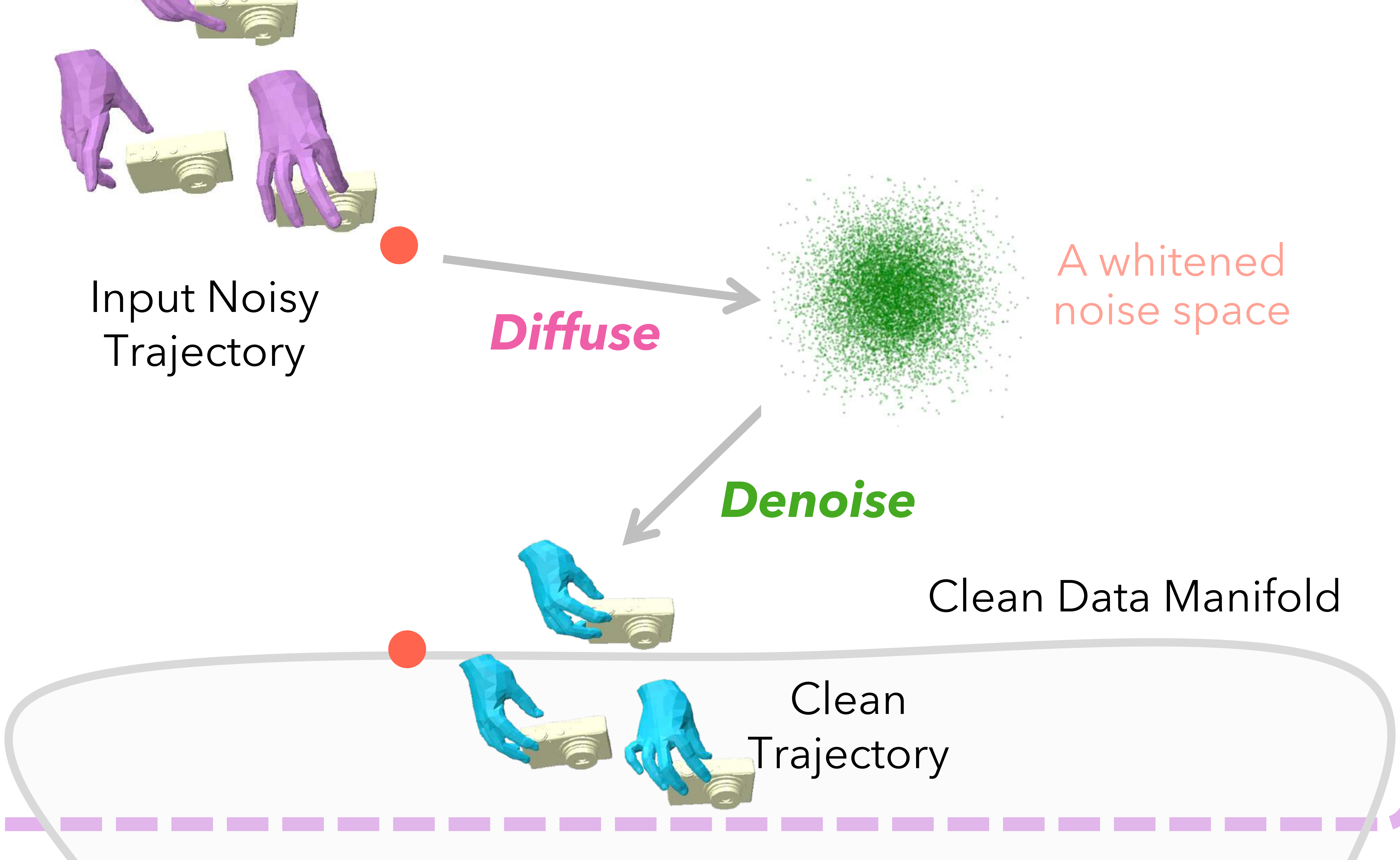
(b) Generalized Contact-Centric Hand-Object Spatial Relations

(c) Generalized Contact-Centric Hand-Object Temporal Relations

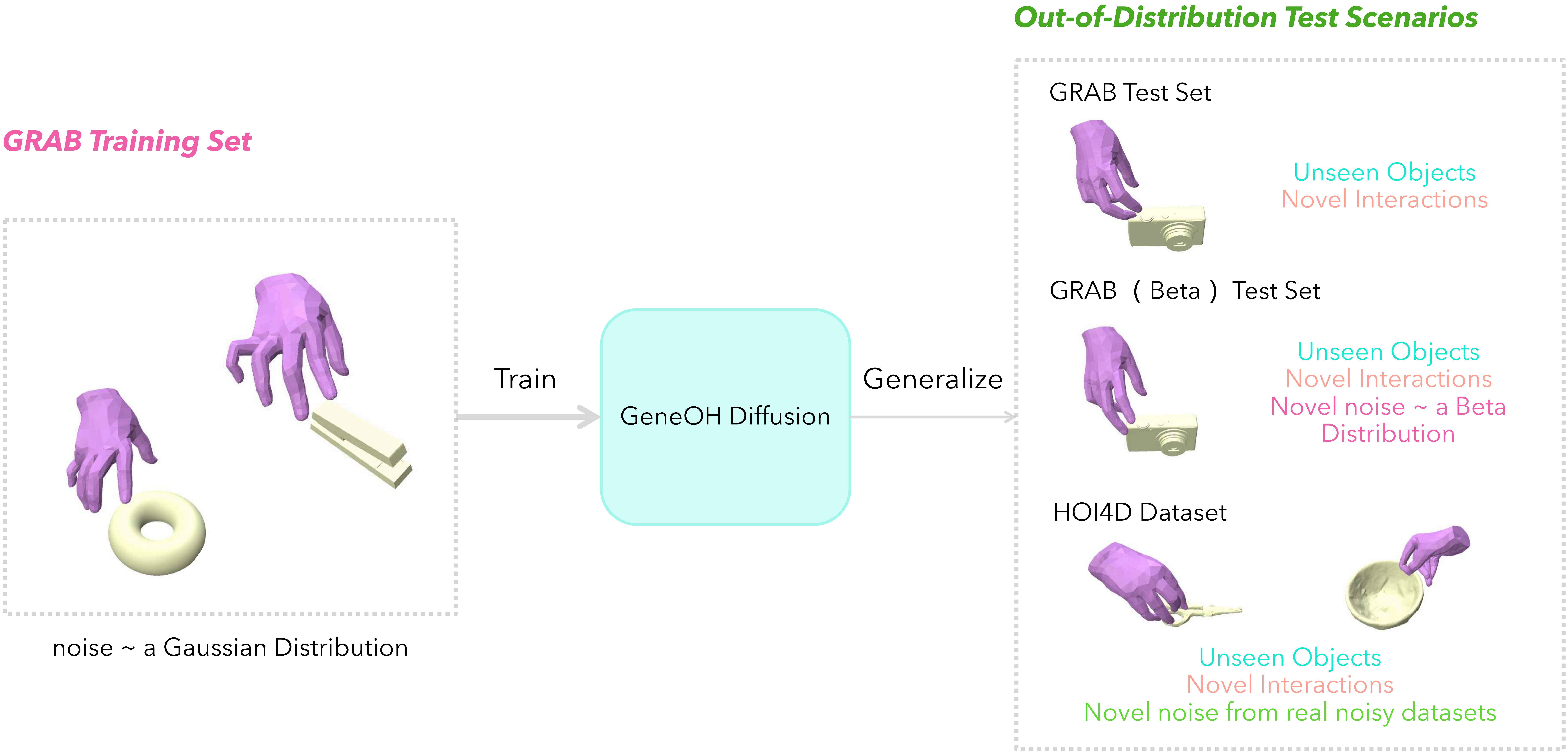


Denoising *via* Diffusion

Aligning inputs with novel noise patterns to the whitened noise space for generalizable denoising



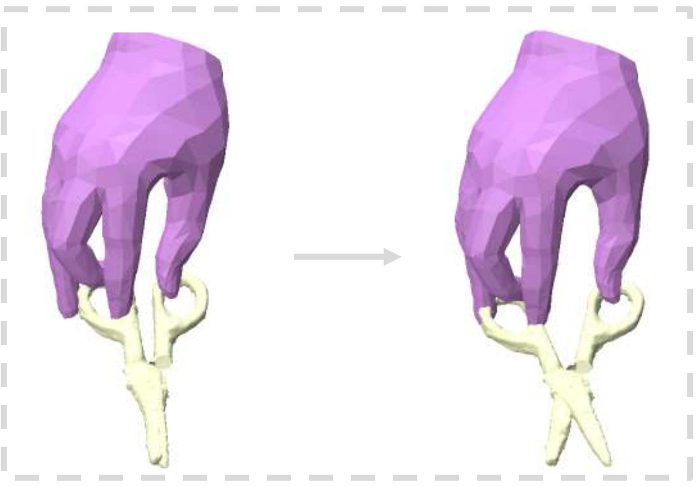
Generalizable HOI Denoising



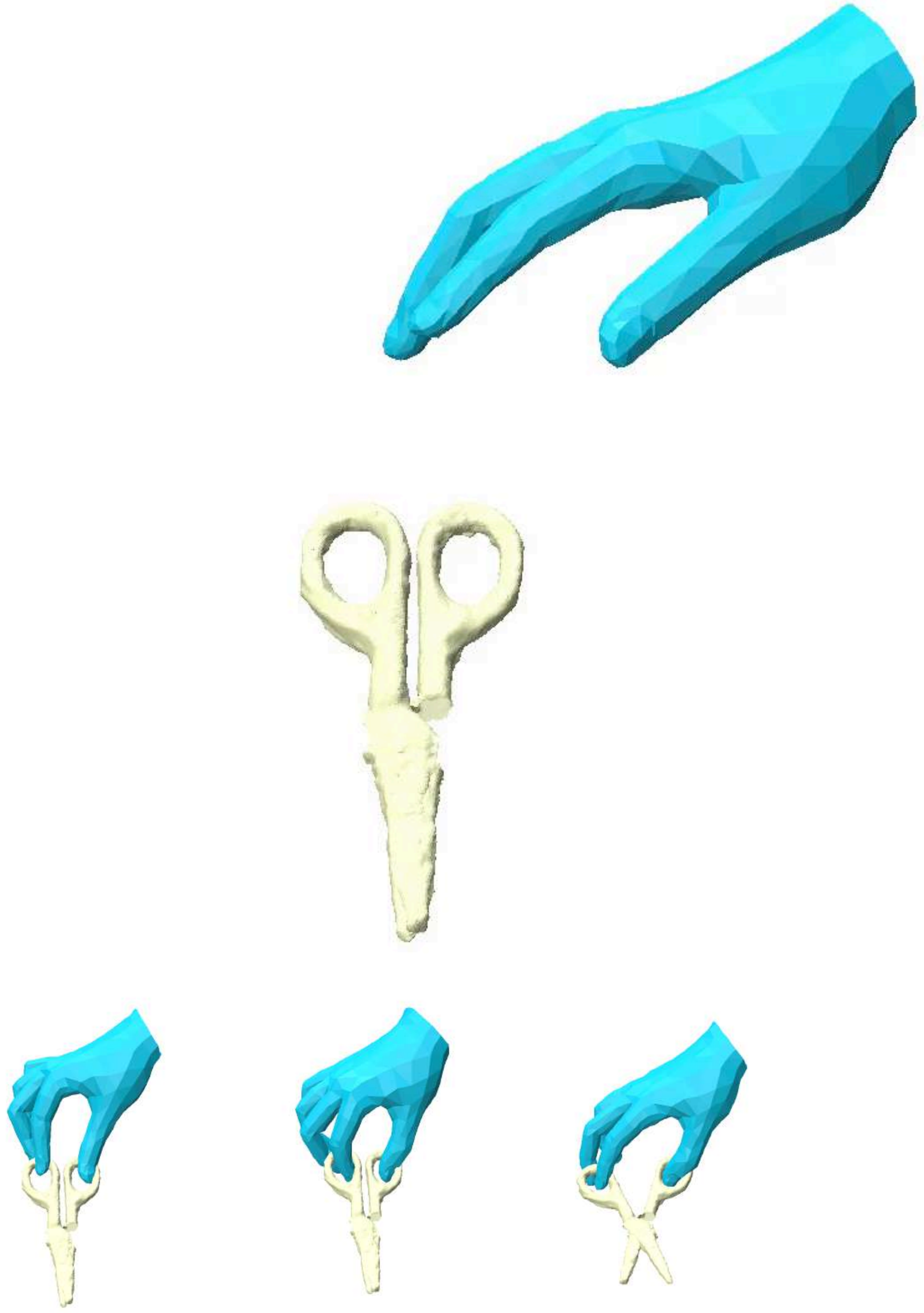
Generalizable HOI Denoising

1. Challenging Geometry (Rings)
2. Articulation Variation
3. Novel and Difficult Real Noise Patterns

Noisy Input

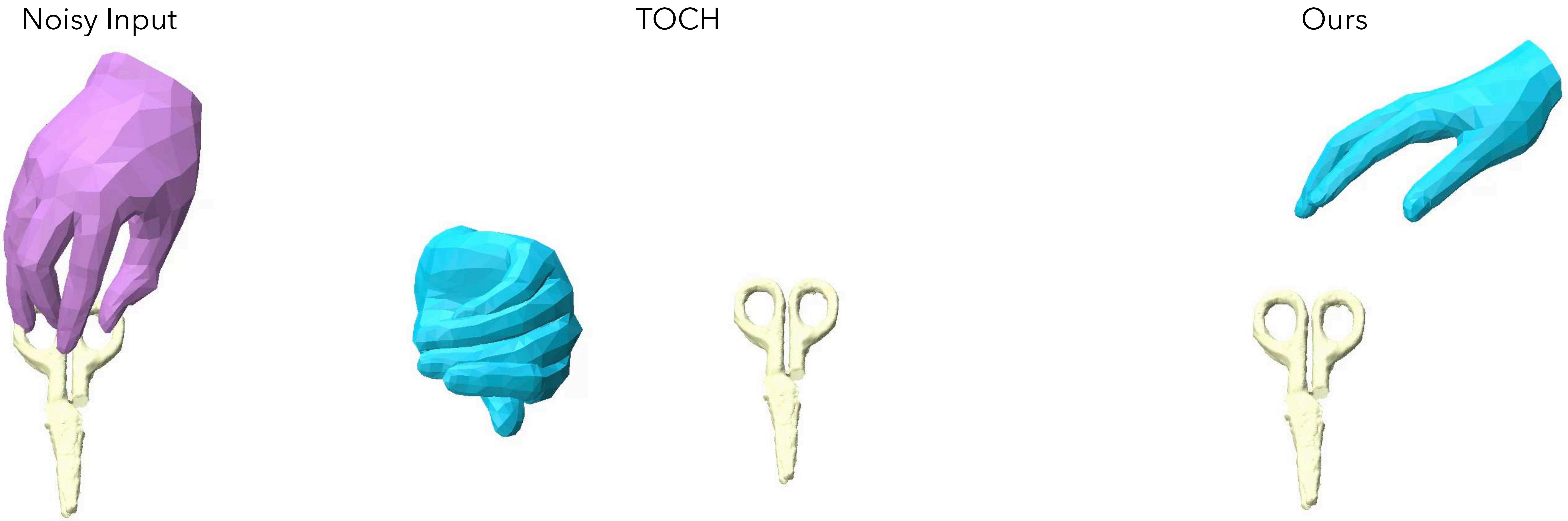


Ours



Generalizable HOI Denoising

1. Challenging Geometry (Rings)
2. Articulation Variation
3. Novel and Difficult Real Noise Patterns



Applications: Refining Retargeted Motions

Source Motion

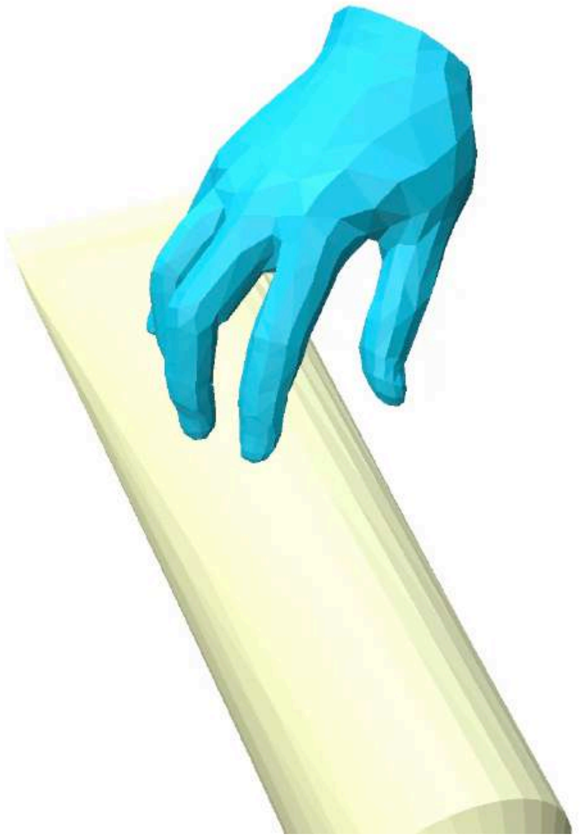
Scale the object up by 2x

Ours

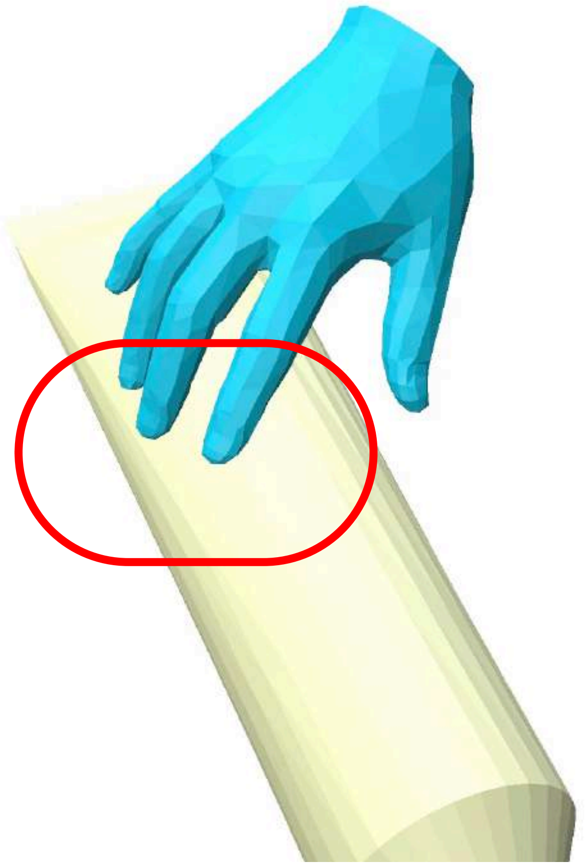
Check our website, GitHub repo, and Hugging Face demo for details!

Website GitHub Hugging Face

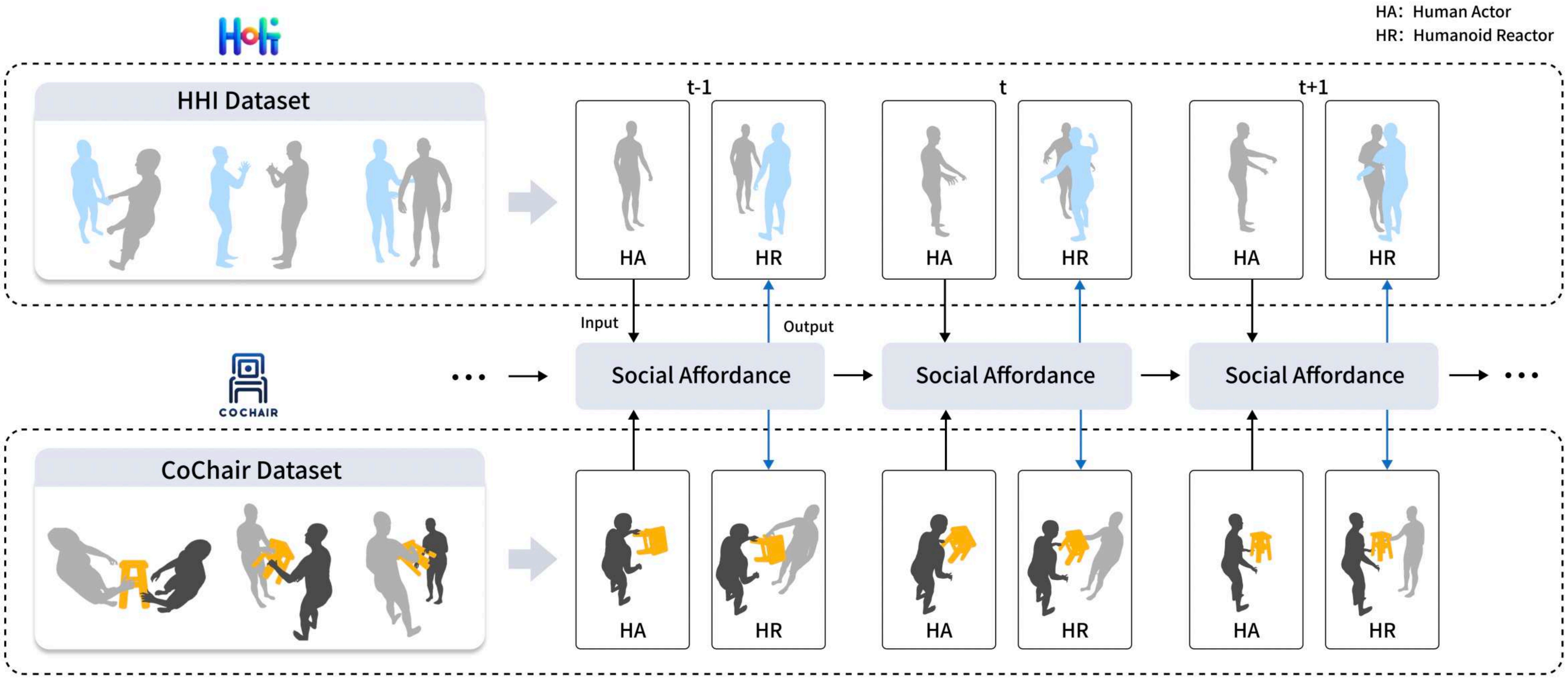
w/o Denoising



TOCH



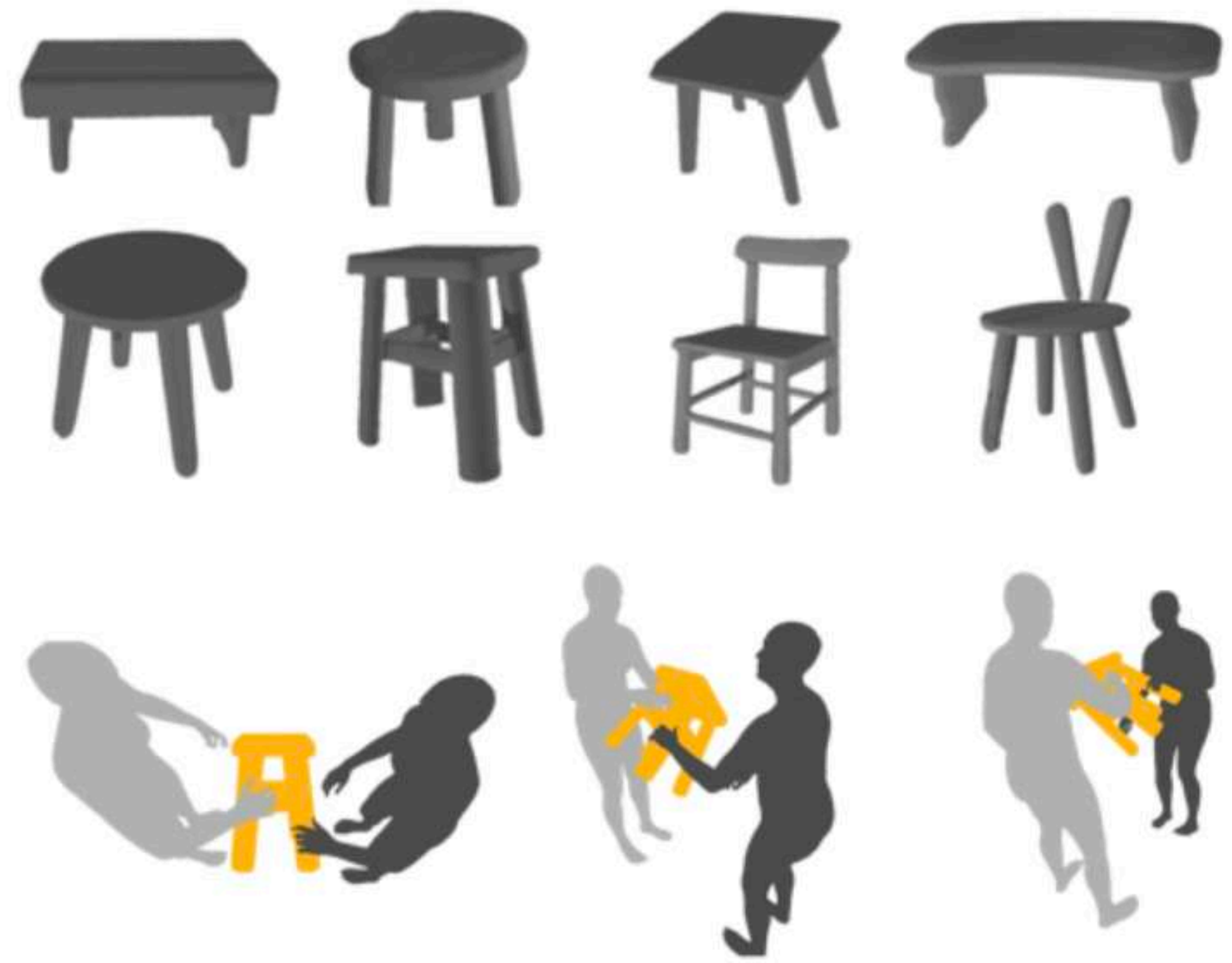
Online Full-Body Motion Reaction Synthesis



Interactive Humanoid: Online Full-Body Motion Reaction Synthesis with Social Affordance Canonicalization and Forecasting

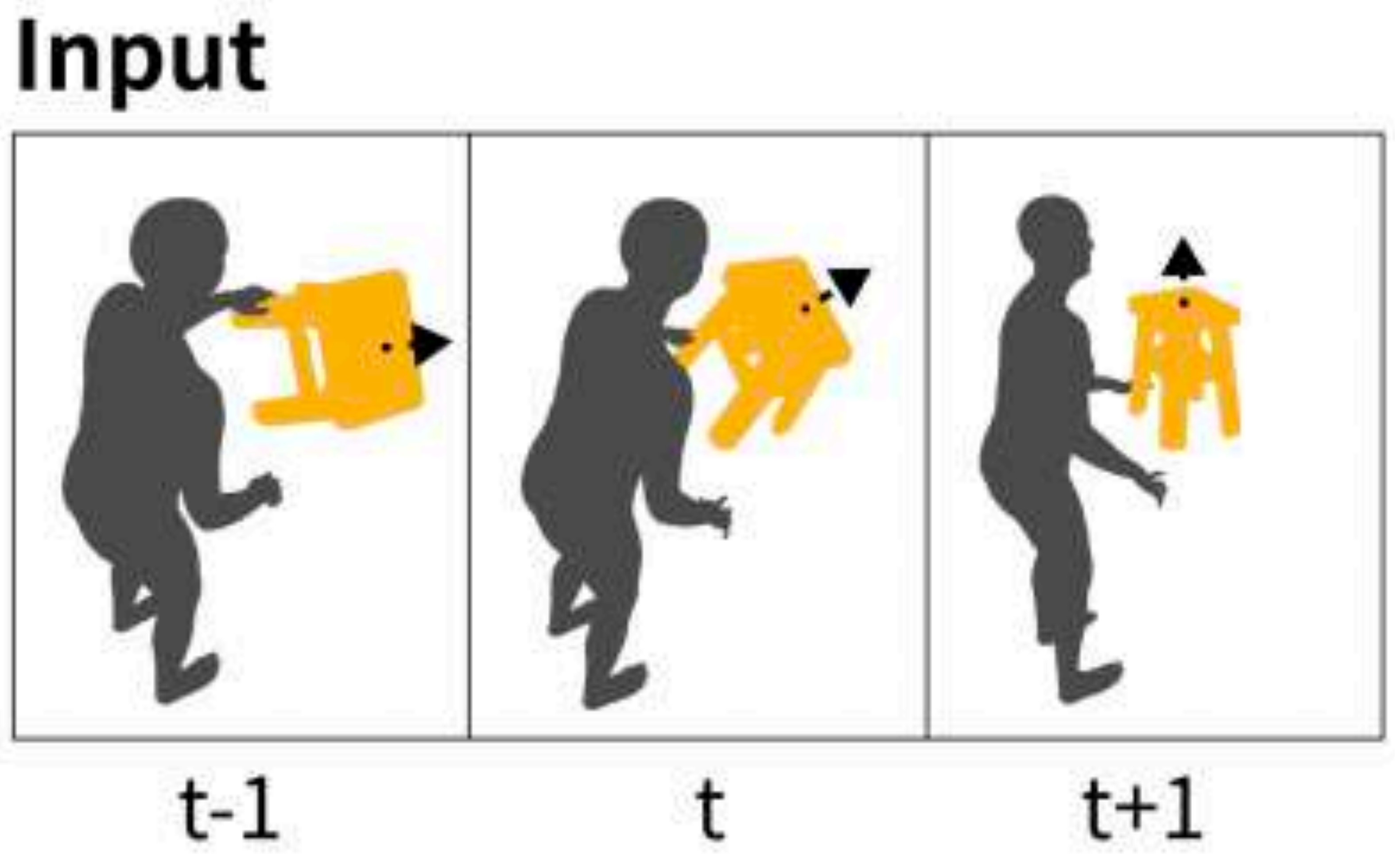
Yunze Liu, Changxi Chen, Li Yi. In submission

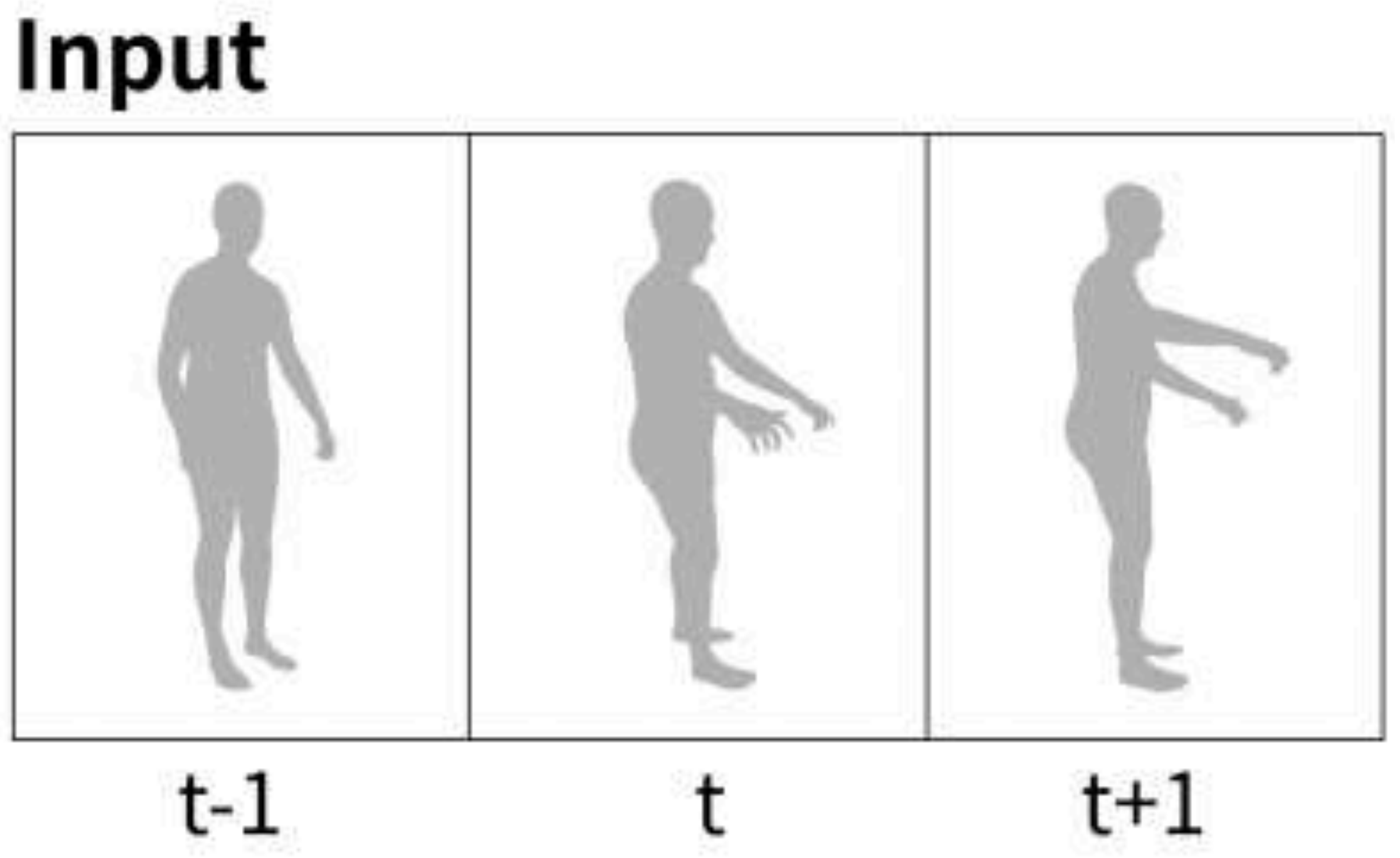
Dataset Statistics



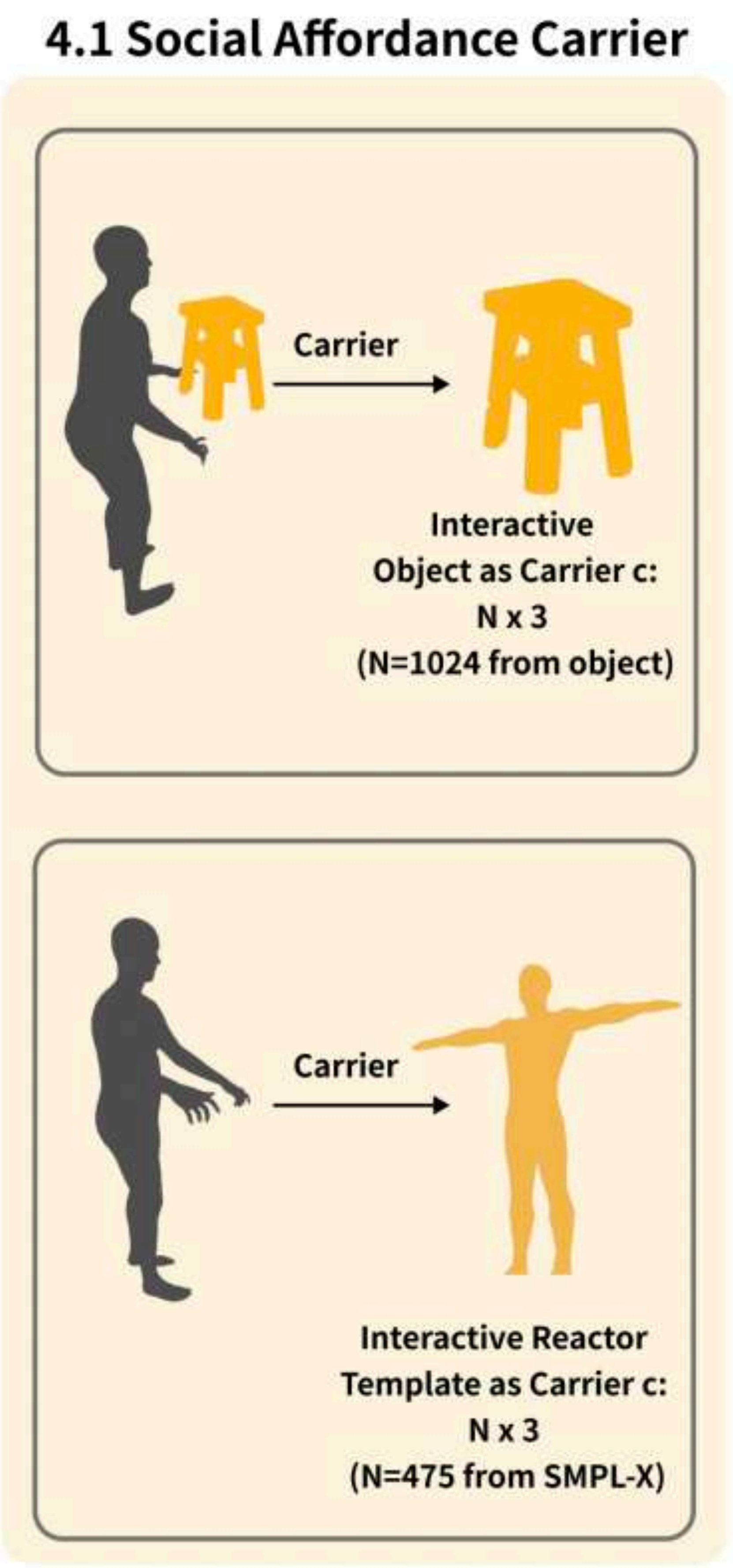
Dataset	Object	Full-body	Actor&Reactor	Mocap	Motions	Verbs	Duration
SBU[53]	-	-	-	-	282	8	0.16h
K3HI[5]	-	-	-	-	312	8	0.21h
NTU120[31]	-	-	-	-	739	26	0.47h
You2me[37]	-	-	-	-	42	4	1.4h
Chi3D[17]	-	-	-	✓✗	373	8	0.41h
InterHuman[30]	-	-	-	✓	6022	5656	6.56h
HHI (Ours)	-	✓	✓	✓	5000	30	5.55h
CoChair (Ours)	✓	✓	✓	✓	3000	5	2.78h

Social Affordance

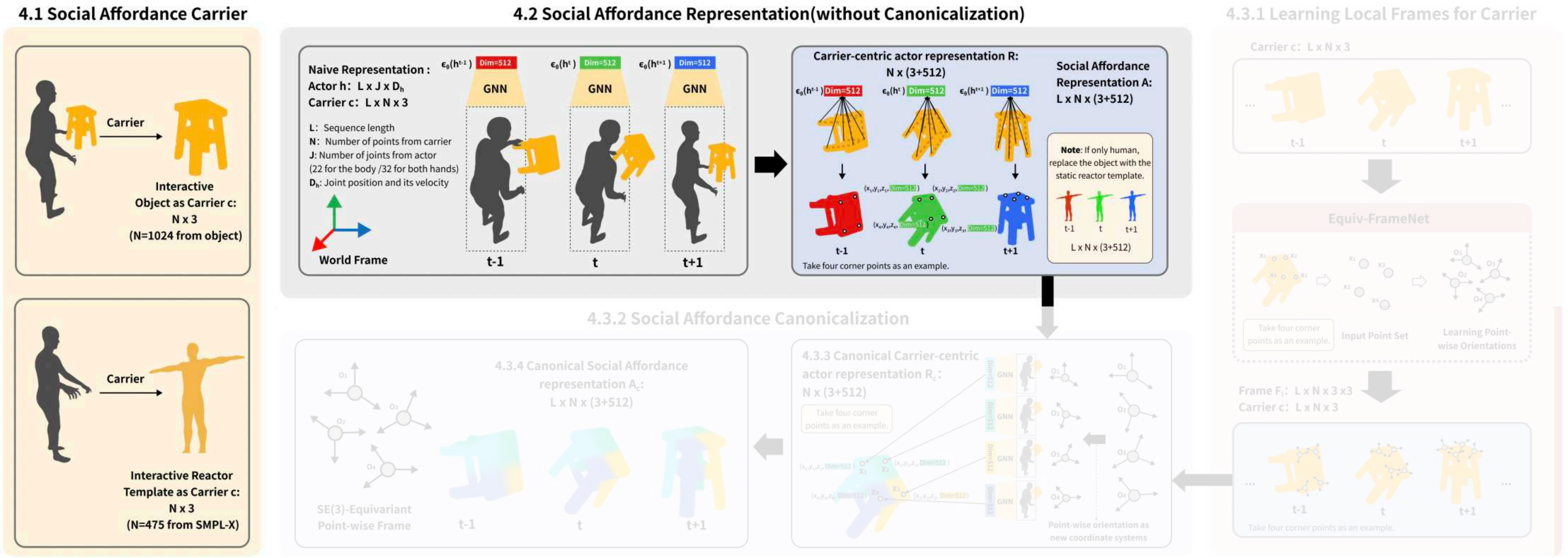




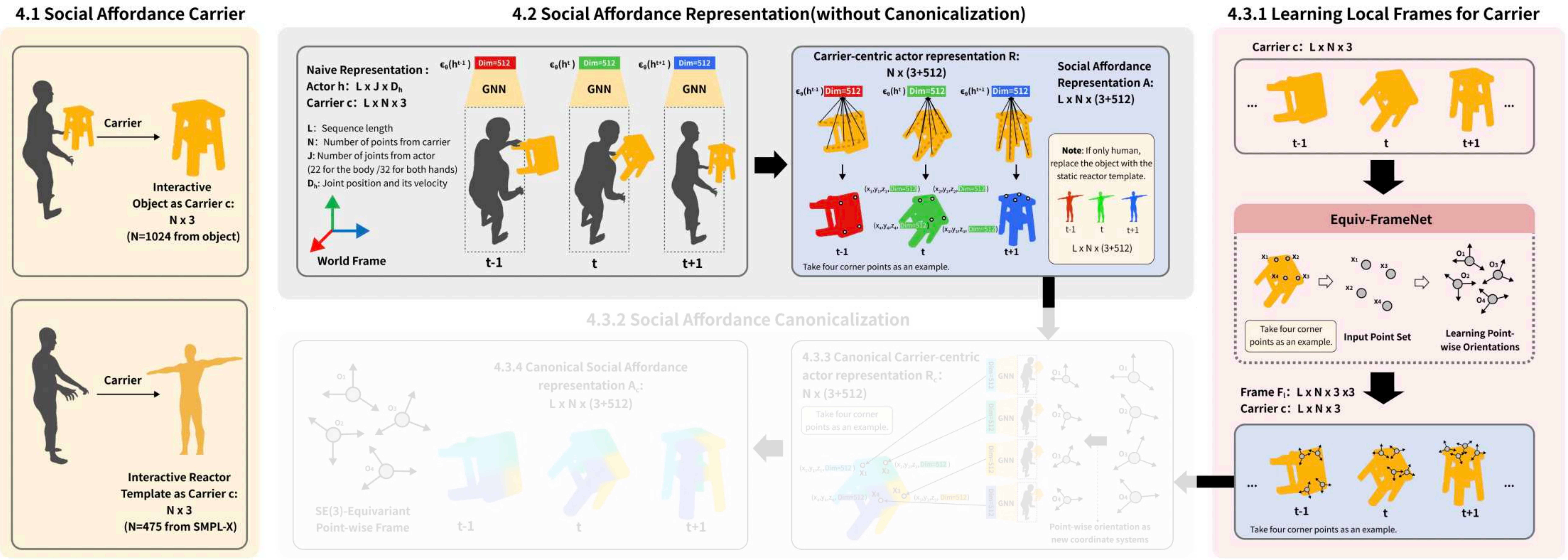
Pipeline


4.1 Social Affordance Carrier
Carrier
Interactive
Object as Carrier c:
N x 3
(N=1024 from object)
Carrier
Interactive Reactor
Template as Carrier c:
N x 3
(N=475 from SMPL-X)

Pipeline

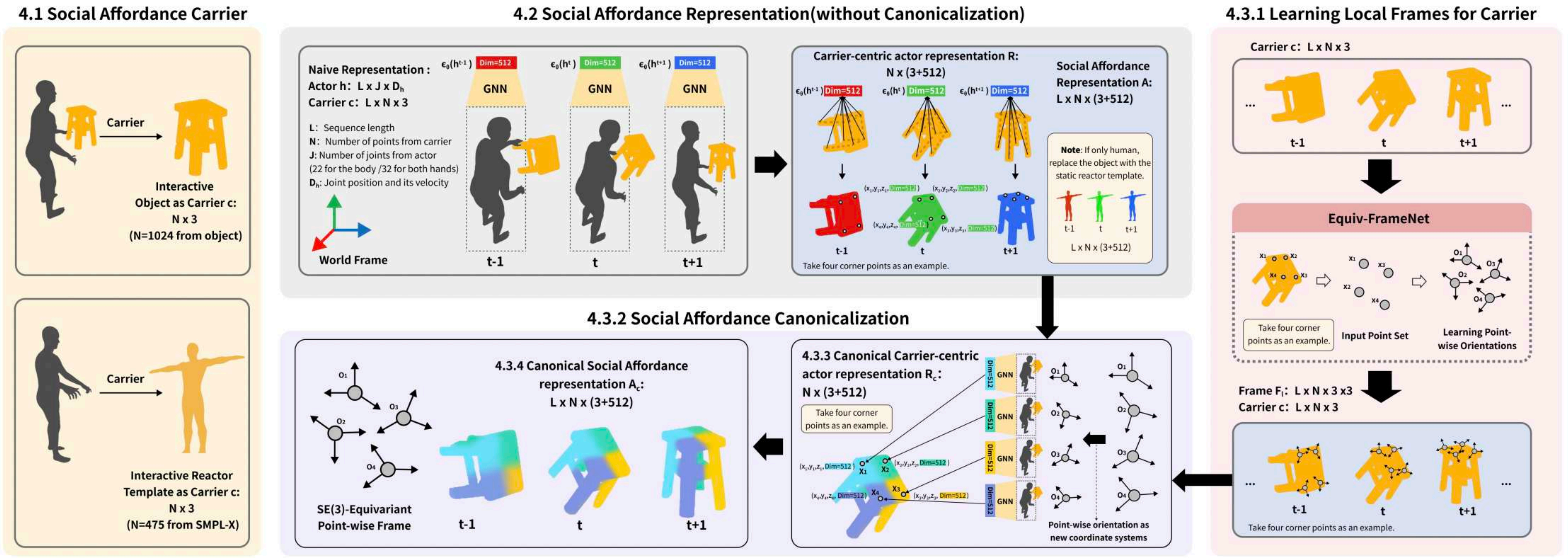

4.1 Social Affordance Carrier
Carrier
Interactive Object as Carrier c: N x 3 (N=1024 from object)
Carrier
Interactive Reactor Template as Carrier c: N x 3 (N=475 from SMPL-X)
4.2 Social Affordance Representation(without Canonicalization)
Naive Representation :
Actor h: L x J x D_h
Carrier c: L x N x 3
L: Sequence length
N: Number of points from carrier
J: Number of joints from actor (22 for the body /32 for both hands)
D_h: Joint position and its velocity
World Frame
GNN
Dim=512
t-1
t
t+1
Carrier-centric actor representation R: N x (3+512)
Social Affordance Representation A: L x N x (3+512)
Note: If only human, replace the object with the static reactor template.
L x N x (3+512)
Take four corner points as an example.
4.3.1 Learning Local Frames for Carrier
Carrier c: L x N x 3
Equiv-FrameNet
Input Point Set
Learning Point-wise Orientations
Frame F_l: L x N x 3 x3
Carrier c: L x N x 3
4.3.2 Social Affordance Canonicalization
4.3.3 Canonical Carrier-centric actor representation R_c: N x (3+512)
Point-wise orientation as new coordinate systems
4.3.4 Canonical Social Affordance representation A_c: L x N x (3+512)
SE(3)-Equivariant Point-wise Frame

Pipeline

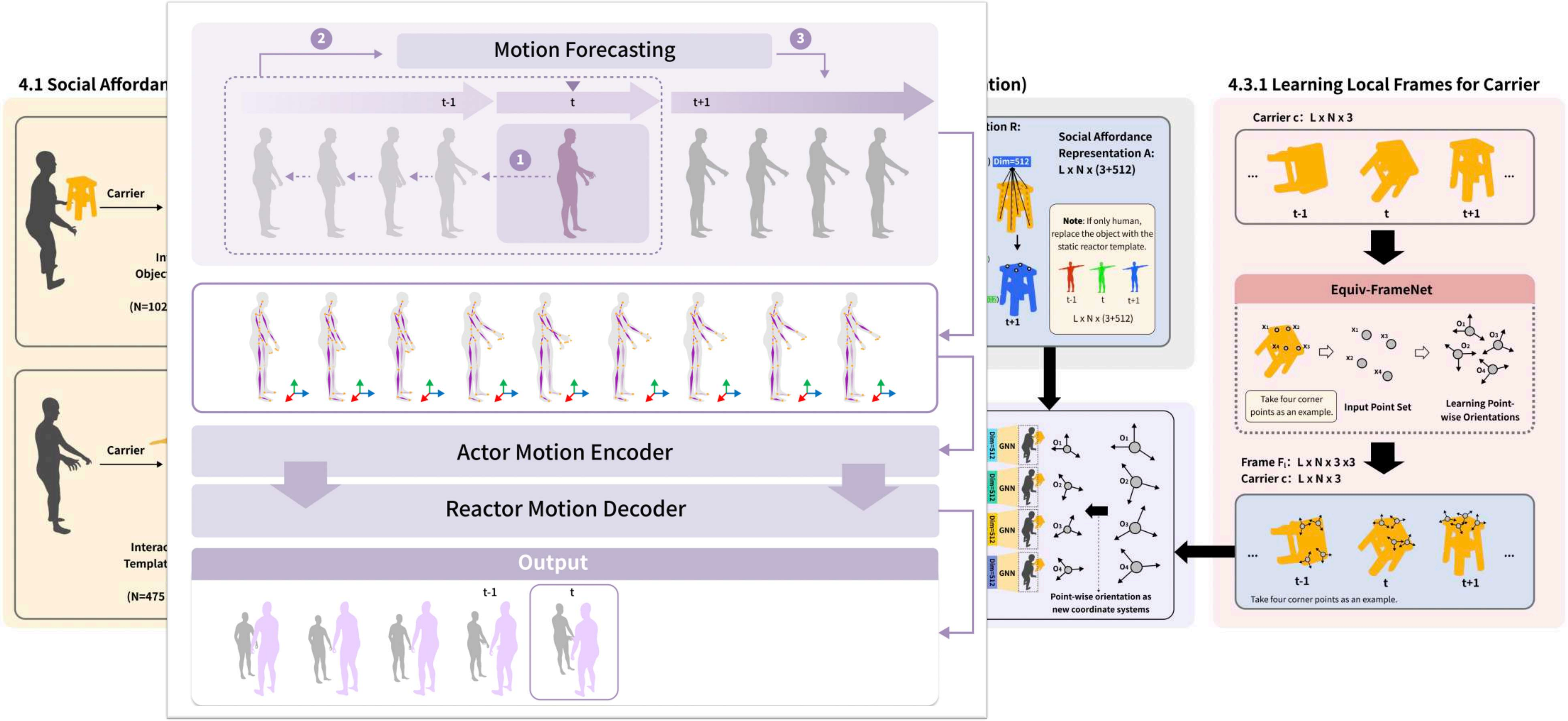


4.1 Social Affordance Carrier
Carrier
Interactive Object as Carrier c: N x 3 (N=1024 from object)
Carrier
Interactive Reactor Template as Carrier c: N x 3 (N=475 from SMPL-X)
4.2 Social Affordance Representation(without Canonicalization)
Naive Representation :
Actor h: L x J x D_h
Carrier c: L x N x 3
L: Sequence length
N: Number of points from carrier
J: Number of joints from actor (22 for the body /32 for both hands)
D_h: Joint position and its velocity
World Frame
GNN
Dim=512
t-1
t
t+1
Carrier-centric actor representation R: N x (3+512)
Social Affordance Representation A: L x N x (3+512)
Note: If only human, replace the object with the static reactor template.
L x N x (3+512)
Take four corner points as an example.
4.3.1 Learning Local Frames for Carrier
Carrier c: L x N x 3
Equiv-FrameNet
Take four corner points as an example.
Input Point Set
Learning Point-wise Orientations
Frame F_i: L x N x 3 x3
Carrier c: L x N x 3
Take four corner points as an example.
4.3.2 Social Affordance Canonicalization
4.3.4 Canonical Social Affordance representation A_c: L x N x (3+512)
SE(3)-Equivariant Point-wise Frame
4.3.3 Canonical Carrier-centric actor representation R_c: N x (3+512)
Take four corner points as an example.
Point-wise orientation as new coordinate systems

Pipeline

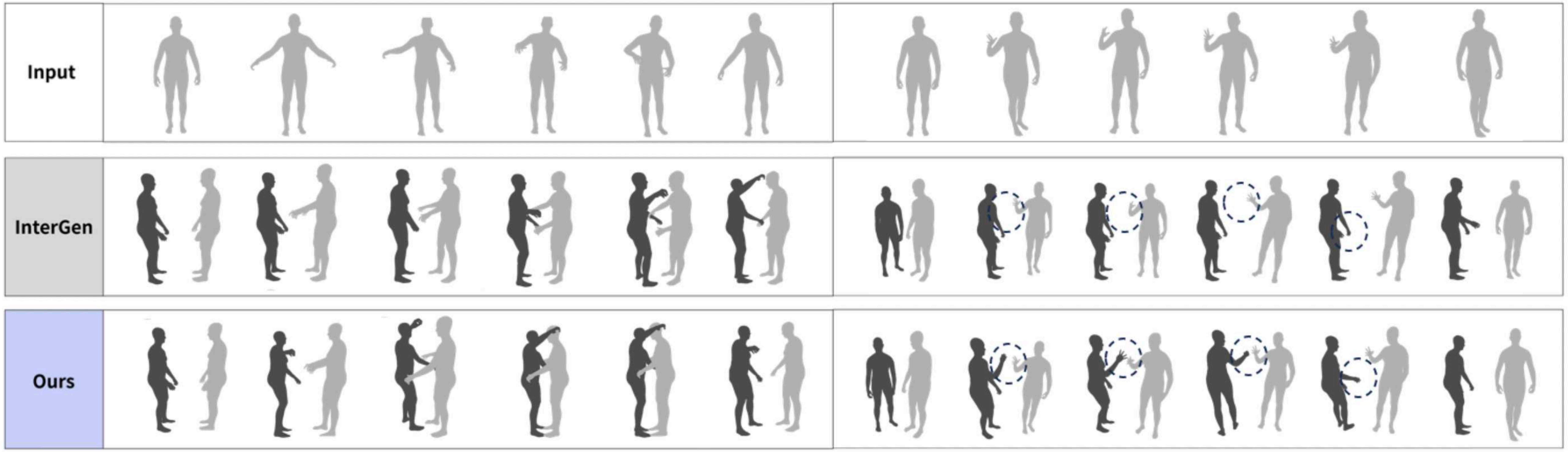

4.1 Social Affordance Carrier
Carrier
Interactive Object as Carrier c: N x 3 (N=1024 from object)
Carrier
Interactive Reactor Template as Carrier c: N x 3 (N=475 from SMPL-X)
4.2 Social Affordance Representation(without Canonicalization)
Naive Representation :
Actor h: L x J x D_h
Carrier c: L x N x 3
L: Sequence length
N: Number of points from carrier
J: Number of joints from actor (22 for the body /32 for both hands)
D_h: Joint position and its velocity
World Frame
GNN
Dim=512
t-1
t
t+1
Carrier-centric actor representation R: N x (3+512)
Social Affordance Representation A: L x N x (3+512)
Note: If only human, replace the object with the static reactor template.
L x N x (3+512)
Take four corner points as an example.
4.3.1 Learning Local Frames for Carrier
Carrier c: L x N x 3
Equiv-FrameNet
Take four corner points as an example.
Input Point Set
Learning Point-wise Orientations
Frame F_i: L x N x 3 x3
Carrier c: L x N x 3
Take four corner points as an example.
4.3.2 Social Affordance Canonicalization
4.3.4 Canonical Social Affordance representation A_c: L x N x (3+512)
SE(3)-Equivariant Point-wise Frame
4.3.3 Canonical Carrier-centric actor representation R_c: N x (3+512)
Take four corner points as an example.
Point-wise orientation as new coordinate systems

Pipeline



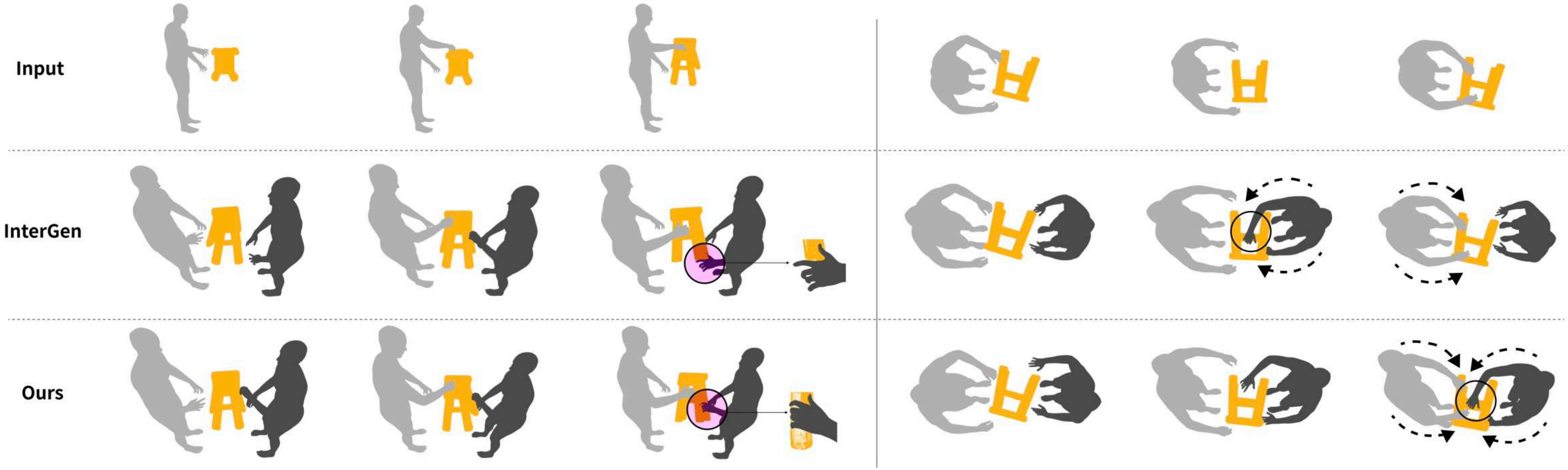
4.1 Social Affordan
Carrier
In
Objec
(N=102
Carrier
Interac
Templat
(N=475
Motion Forecasting
t-1
t
t+1
Actor Motion Encoder
Reactor Motion Decoder
Output
tion)
tion R:
Social Affordance Representation A: L x N x (3+512)
Note: If only human, replace the object with the static reactor template.
L x N x (3+512)
Point-wise orientation as new coordinate systems
4.3.1 Learning Local Frames for Carrier
Carrier c: L x N x 3
Equiv-FrameNet
Take four corner points as an example.
Input Point Set
Learning Point-wise Orientations
Frame F_i: L x N x 3 x3
Carrier c: L x N x 3
Take four corner points as an example.

Experimental Results for Human-Human Interaction



Method	FID ↓			Diversity →			Accuracy ↑			User Preference↑		
	HHI	InterHuman[30]	Chi3D[17]	HHI	InterHuman[30]	Chi3D[17]	HHI	InterHuman	Chi3D[17]	HHI	InterHuman[30]	Chi3D[17]
Real	0.21	0.17	0.05	10.8	12.4	14.0	88.2	-	80.4	-	-	-
PGBIG[33]	56.7	87.2	67.2	13.9	17.1	17.8	34.1	-	61.6	4.4	1.0	8.3
SS-Transformer[2]	77.8	107.3	54.9	16.2	18.5	19.2	51.9	-	57.1	2.7	4.6	18.4
InterFormer[12]	54.3	73.1	20.8	14.1	14.2	14.8	77.9	-	62.2	6.0	2.1	13.7
InterGen-Revised[30]	19.8	25.7	17.7	11.6	13.3	14.2	80.2	-	71.9	19.7	41.7	15.4
Ours	**13.3**	**14.7**	**12.8**	**11.1**	**13.3**	**14.1**	**85.4**	-	**77.6**	**67.2**	**50.6**	**44.2**

Experimental Results for HOH Interaction



Physically Plausible Reaction Synthesis

PhysReaction: Physically Plausible Real-Time Humanoid Reaction Synthesis via Forward Dynamics Guided 4D Imitation

Yunze Liu, Changxi Chen, Chenjing Ding, Li Yi. In submission

Embodied Digital Agent

Interactive Humanoid: Online Full-Body Motion Reaction Synthesis with Social Affordance Canonicalization and Forecasting
Yunze Liu, Changxi Chen, Li Yi. In submission

Can We Get Rid of MoCap?

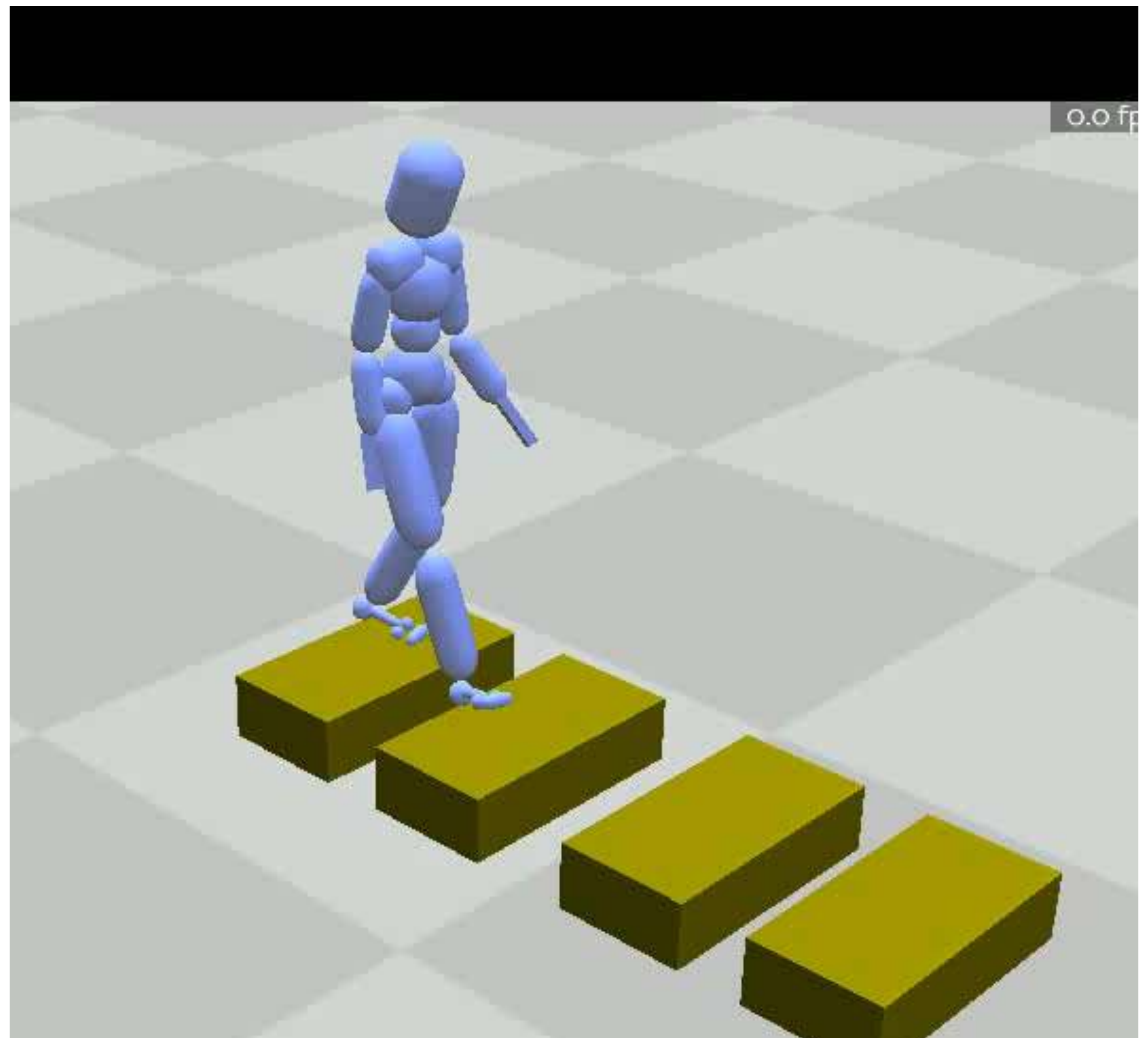
FreeMotion: Mocap-Free Human Motion Synthesis with Multimodal Large Language Models
Zhikai Zhang, Yitang Li, Haofeng Huang, Mingxian Lin, Li Yi. In submission.

Problem Setup

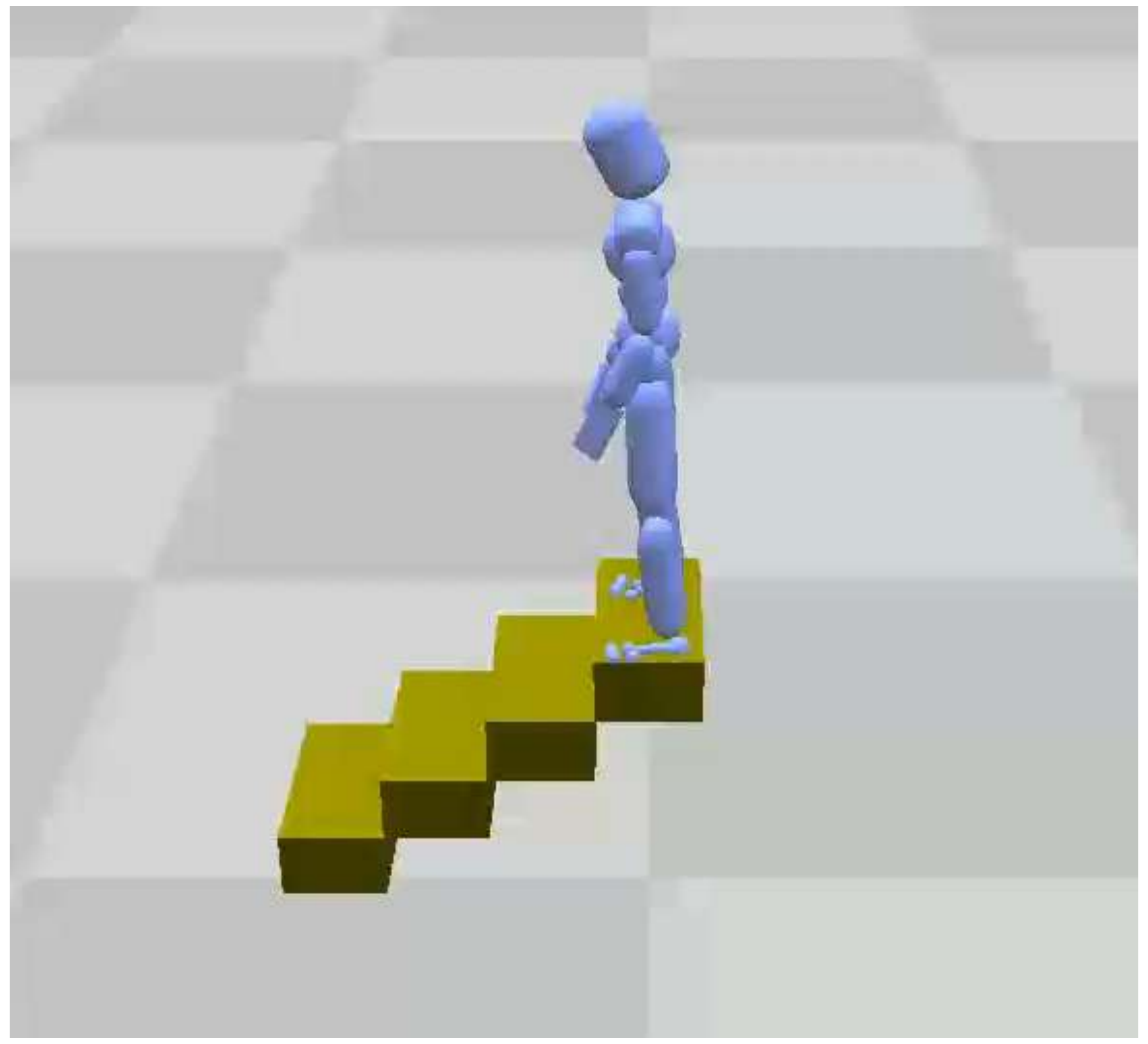
- Generate human-like physical humanoid motions in novel environments given arbitrary text prompts without any human-motion training data

Sit down on the chair

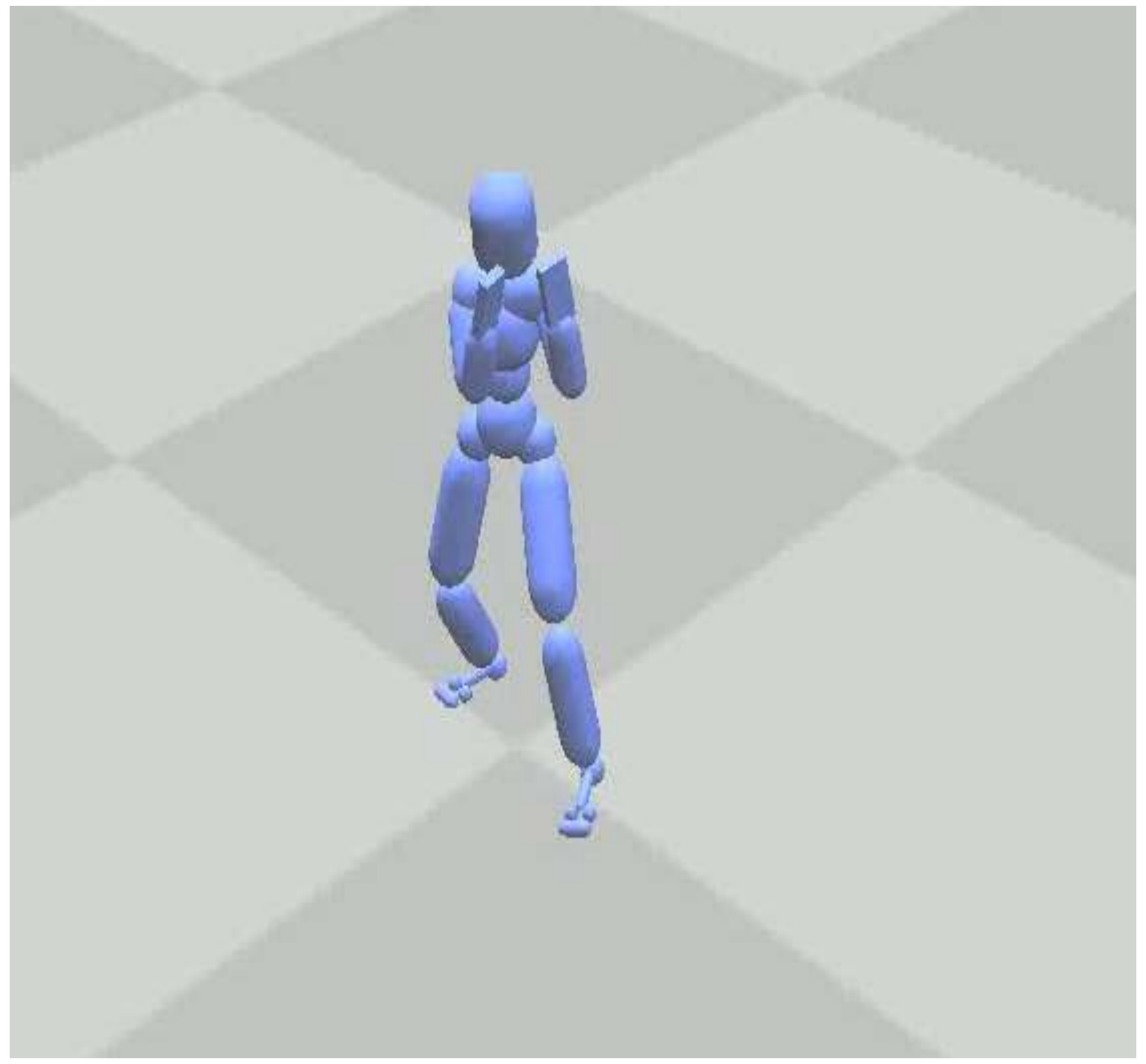
Stepp on the stones



Go downstairs

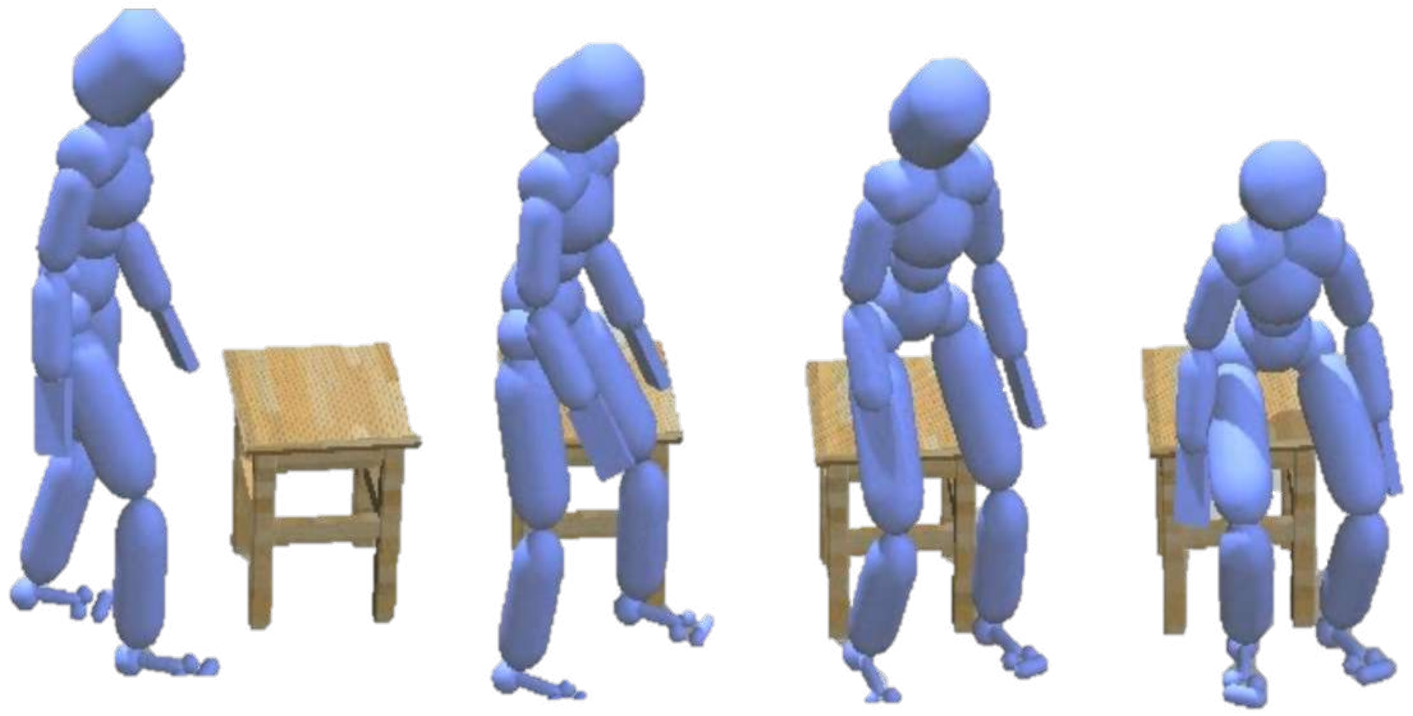


Boxing



Existing Paradigms

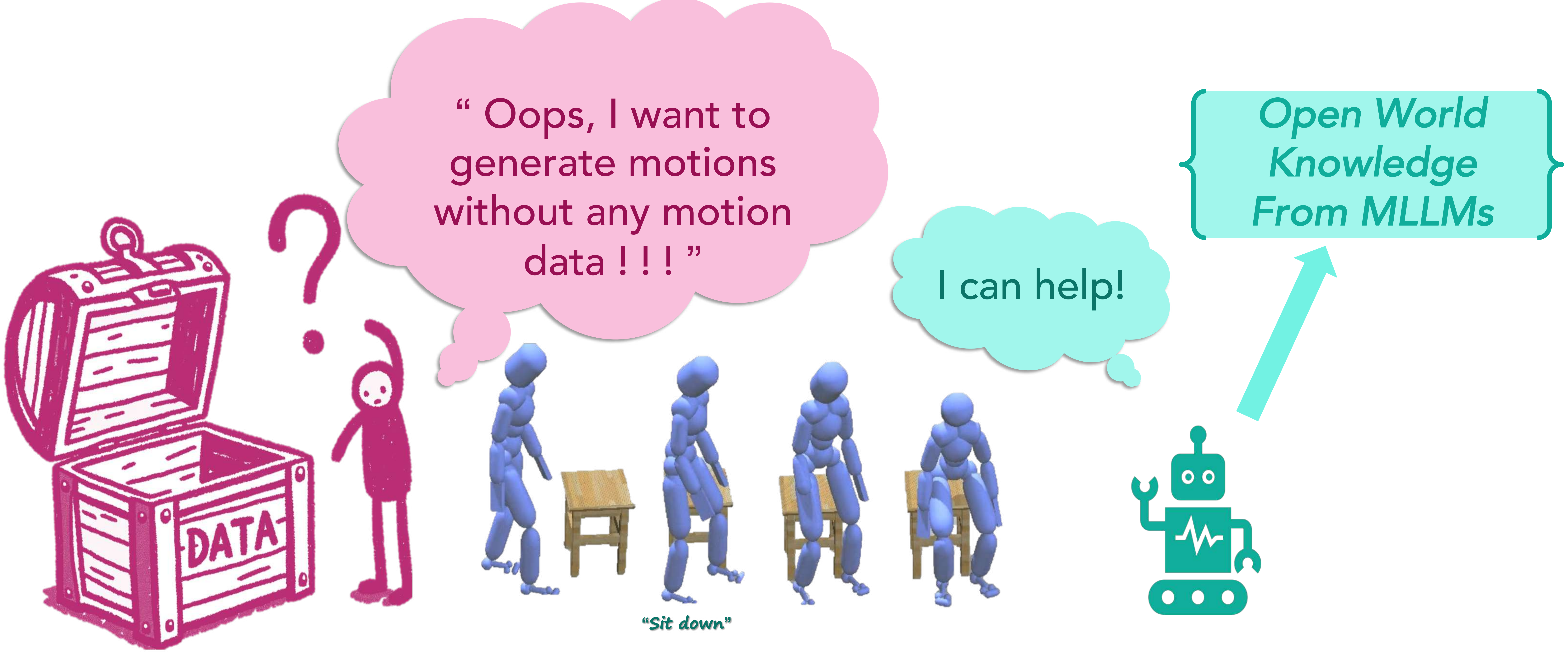
- Data-driven text-based motion synthesis
  - Very limited motion categories in specific environments with restricted motion styles constrained by the training data



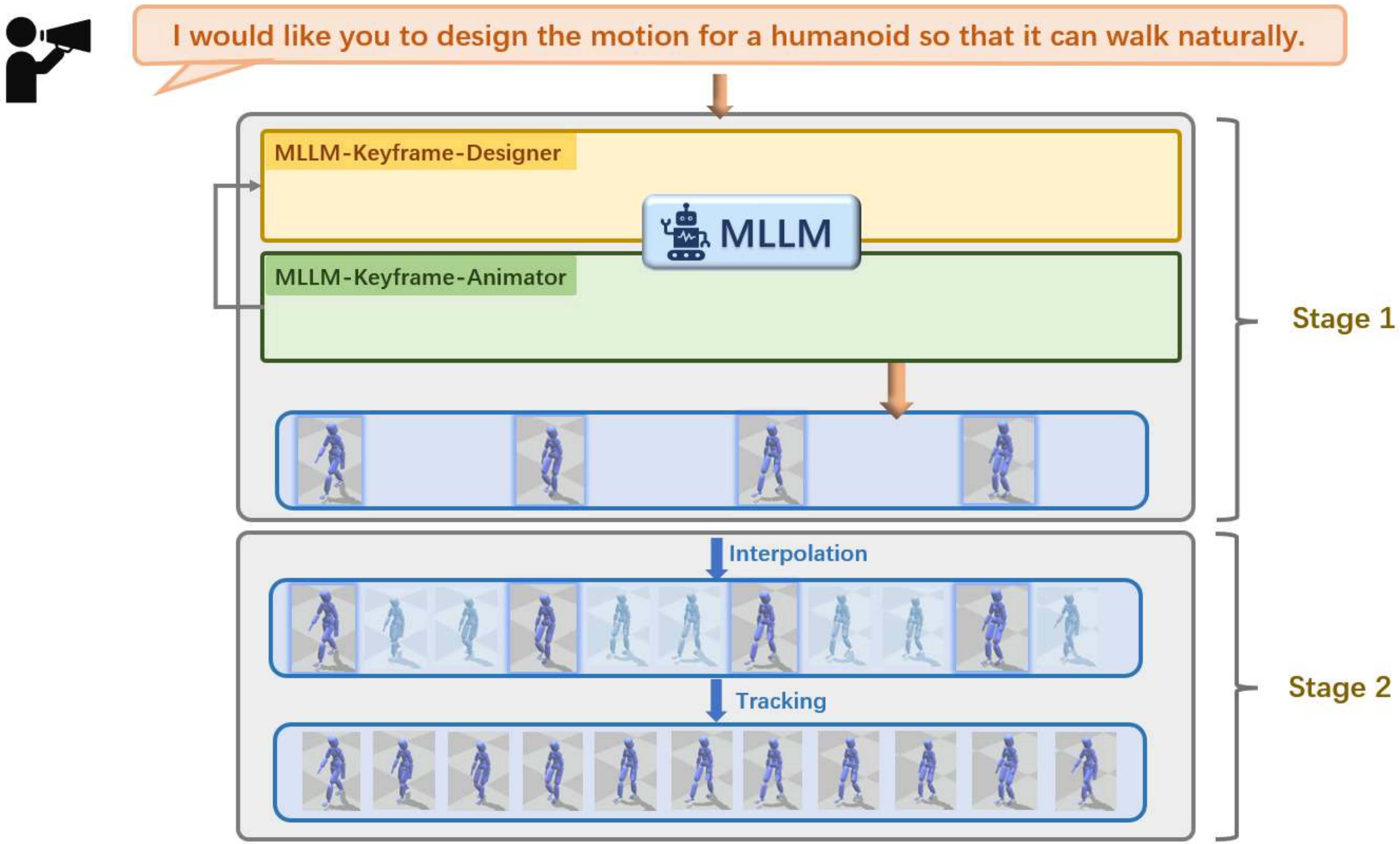
“Sit down”

Our Paradigm

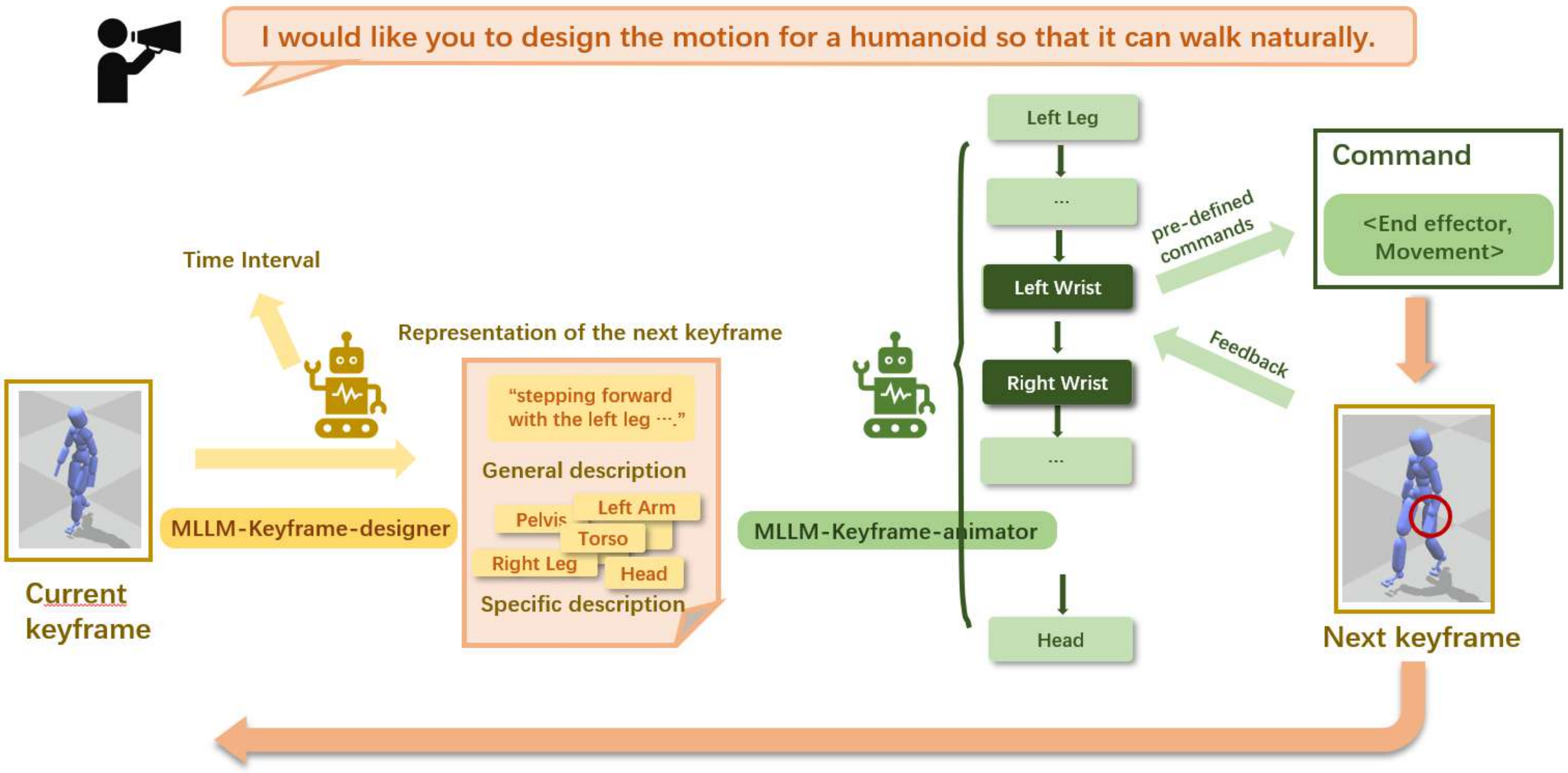
- Knowledge-driven text-based motion synthesis
  - Any motion categories in any environments with any motion styles



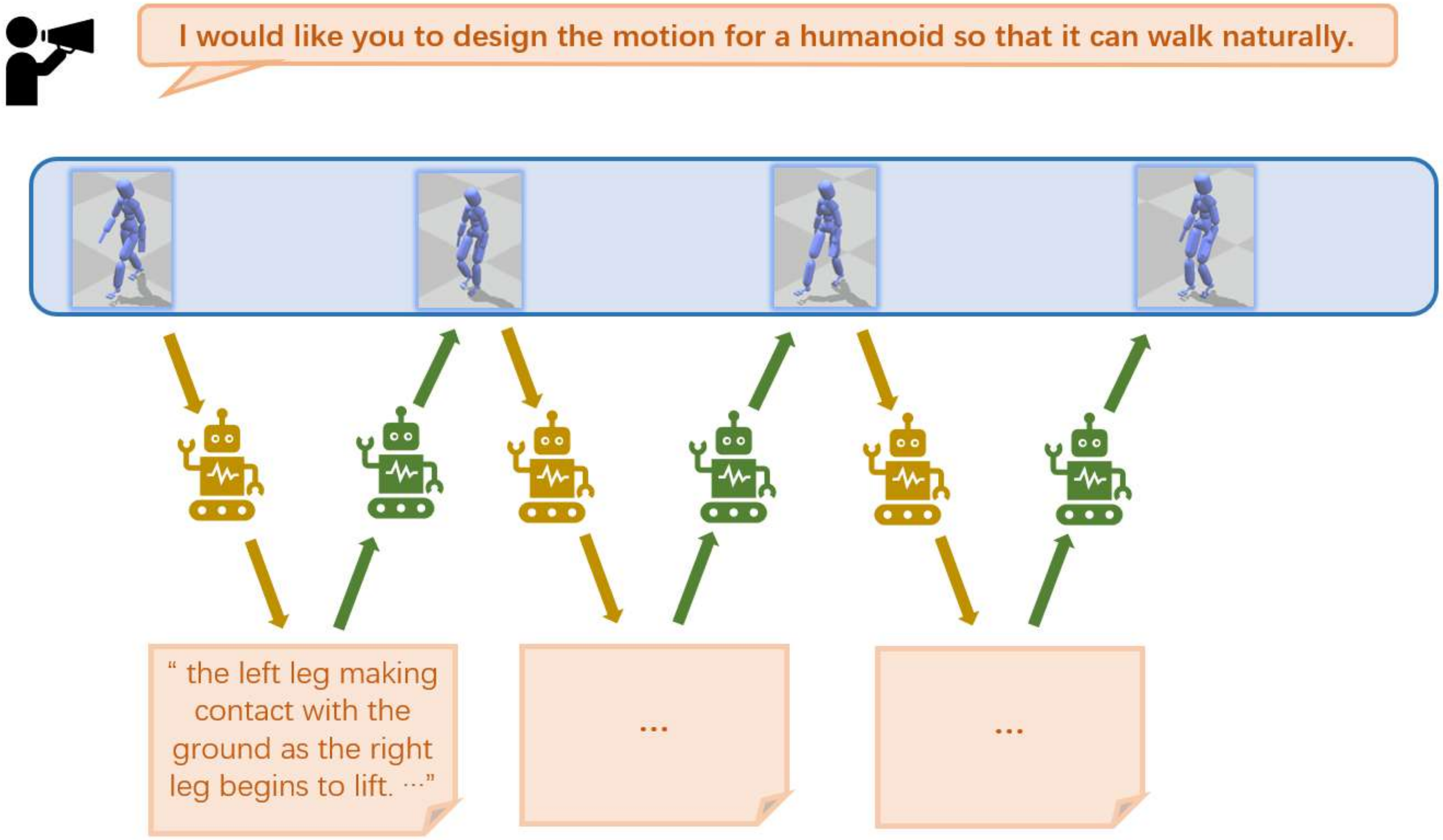
Pipeline


I would like you to design the motion for a humanoid so that it can walk naturally.
MLLM-Keyframe-Designer
MLLM
MLLM-Keyframe-Animator
Stage 1
Interpolation
Tracking
Stage 2

Pipeline

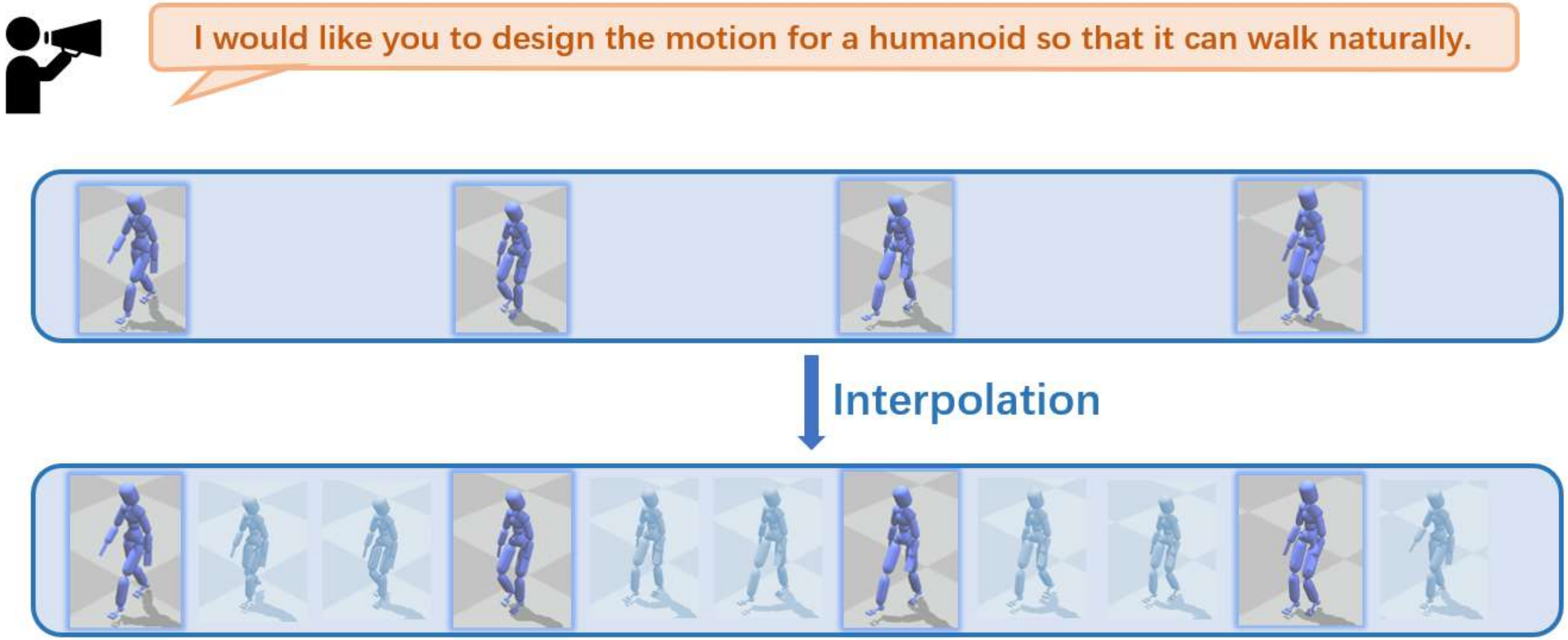

I would like you to design the motion for a humanoid so that it can walk naturally.
Time Interval
Representation of the next keyframe
"stepping forward with the left leg ⋯."
General description
Pelvis
Left Arm
Torso
Right Leg
Head
Specific description
MLLM-Keyframe-designer
Current keyframe
MLLM-Keyframe-animator
Left Leg
...
Left Wrist
Right Wrist
...
Head
pre-defined commands
Feedback
Command
<End effector, Movement>
Next keyframe

Pipeline

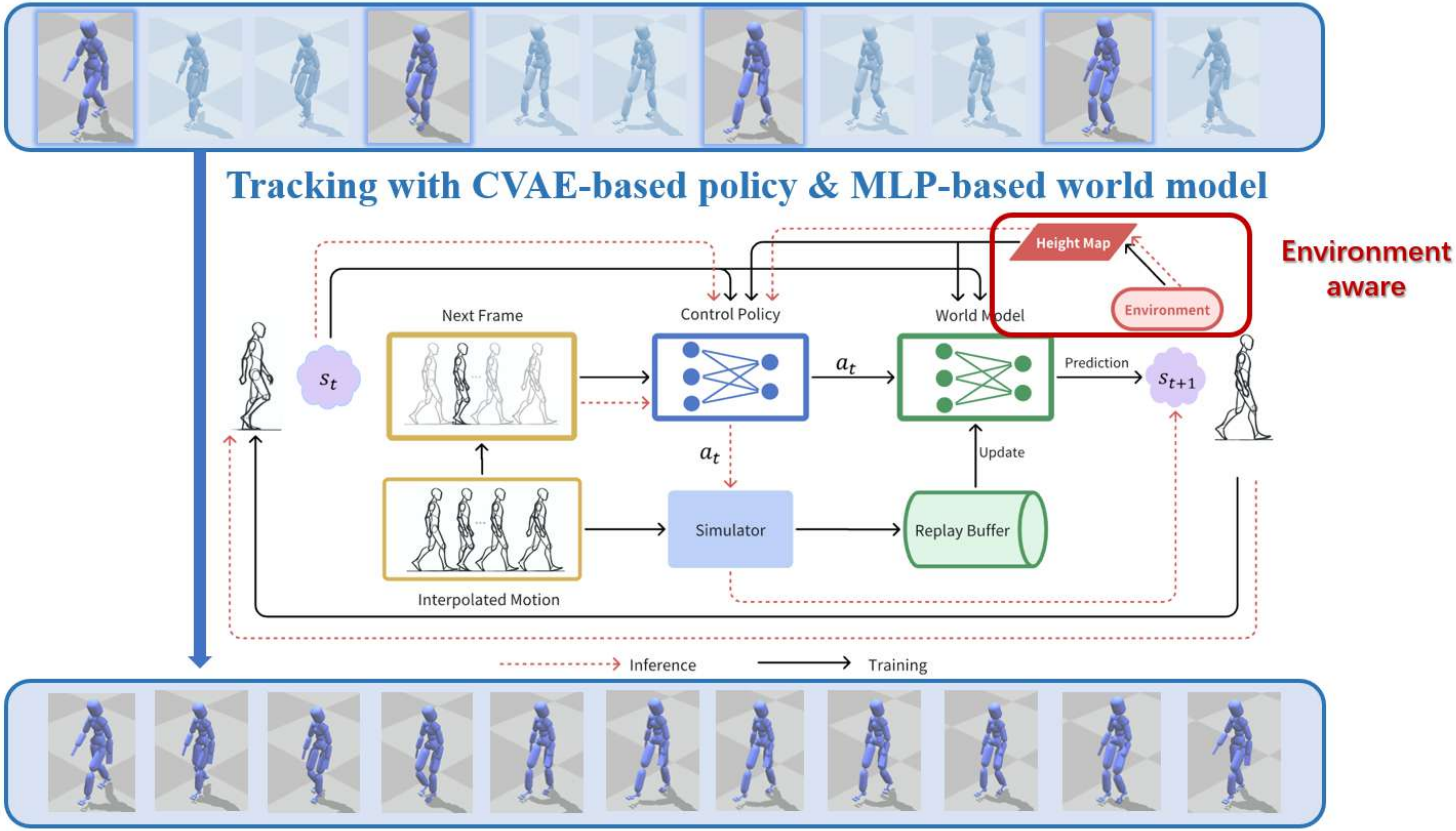


I would like you to design the motion for a humanoid so that it can walk naturally.
" the left leg making contact with the ground as the right leg begins to lift. …"
…
…

Pipeline


I would like you to design the motion for a humanoid so that it can walk naturally.
Interpolation

Pipeline

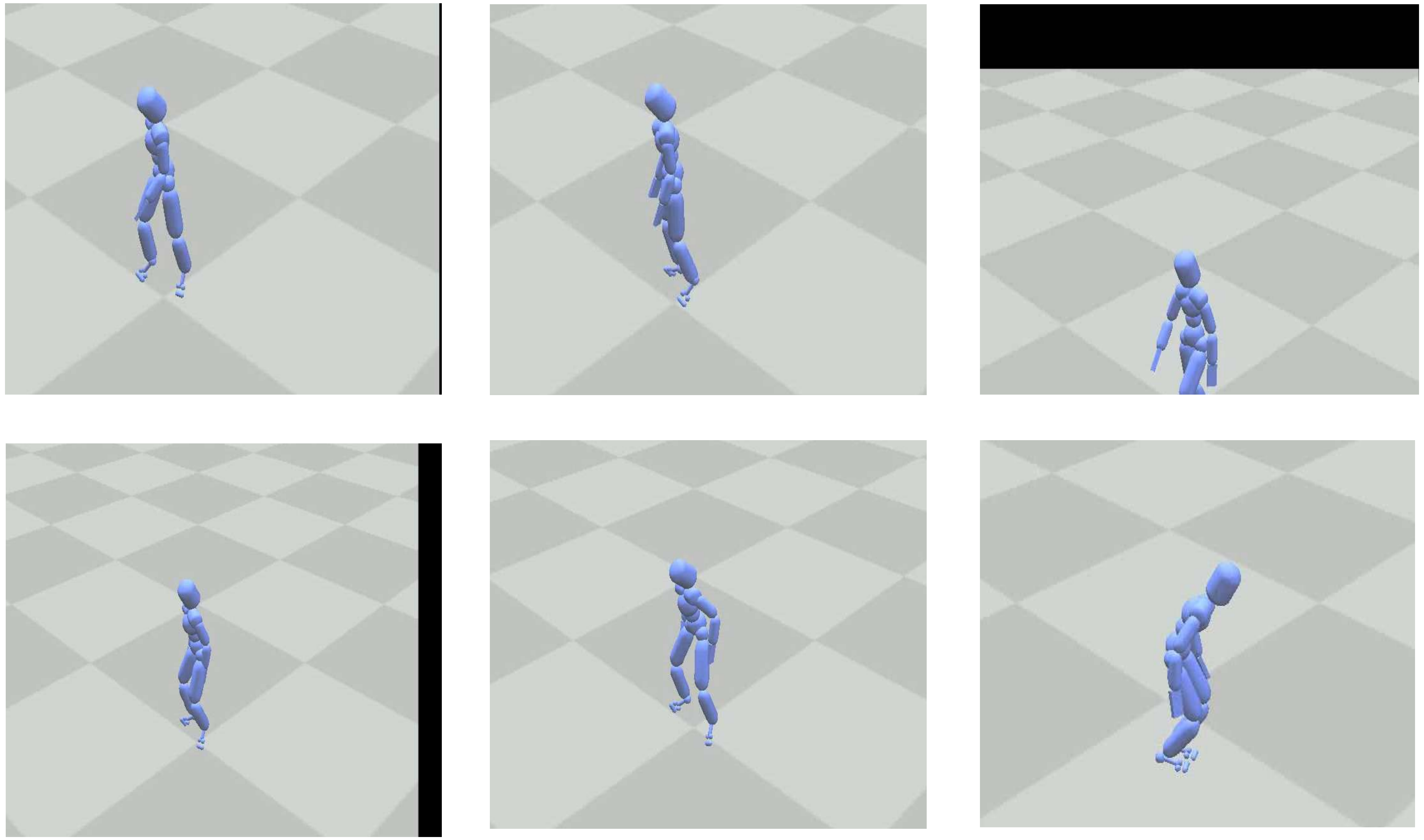


Tracking with CVAE-based policy & MLP-based world model
Height Map
Environment
Environment aware
Next Frame
Control Policy
World Model
s_t
a_t
Prediction
s_{t+1}
a_t
Update
Simulator
Replay Buffer
Interpolated Motion
Inference
Training

Results: HumanAct12

Table 2: Motion Synthesis on HumanAct12. FreeMotion achieves good results without motion data.

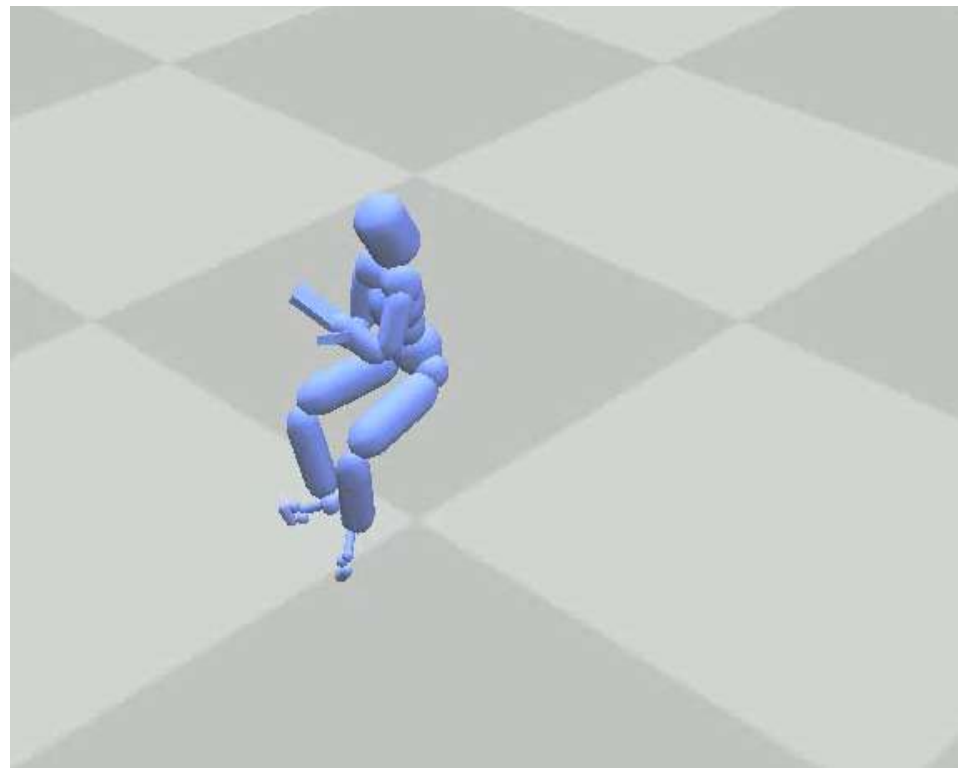
	User Study		
	MDM [37]	MLD [6]	Ours
Warm up	26.00%	**38.00%**	36.00%
Walk	10.00%	22.00%	**68.00%**
Run	30.00%	32.00%	**38.00%**
Jump	16.00%	28.00%	**56.00%**
Drink	14.00%	**46.00%**	40.00%
Lift_dumbbell	26.00%	32.00%	**42.00%**
Sit	30.00%	**44.00%**	26.00%
Eat	22.00%	30.00%	**48.00%**
Turn_steering_wheel	32.00%	28.00%	**40.00%**
Phone	30.00%	32.00%	**38.00%**
Boxing	16.00%	24.00%	**60.00%**
Throw	20.00%	14.00%	**66.00%**
Average	22.67%	30.83%	**46.50%**

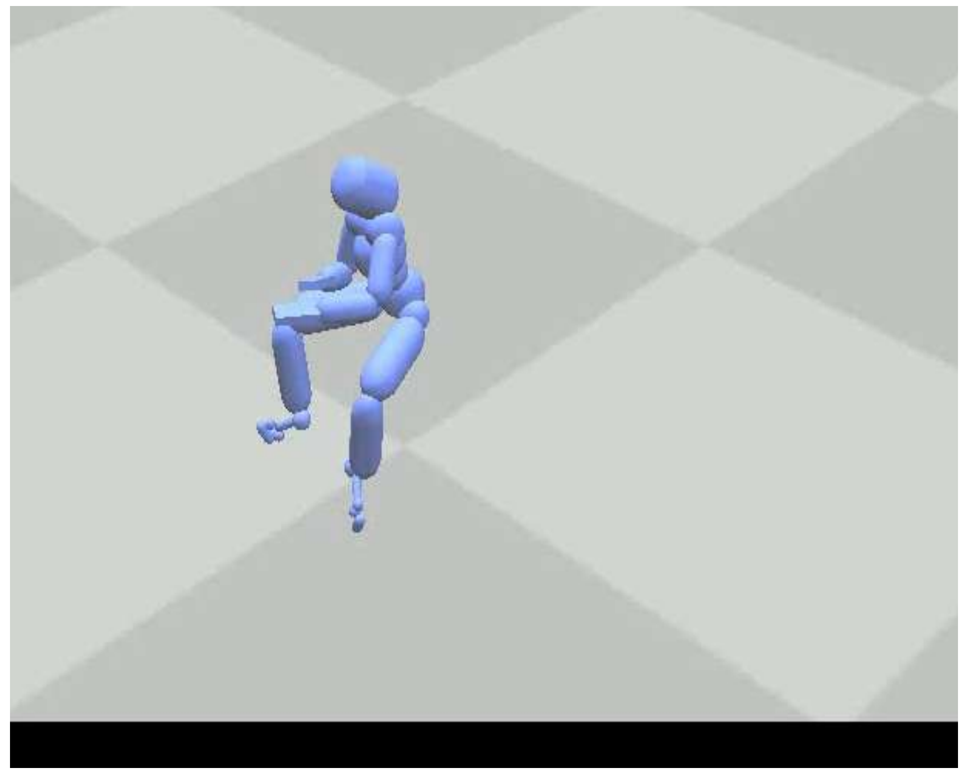


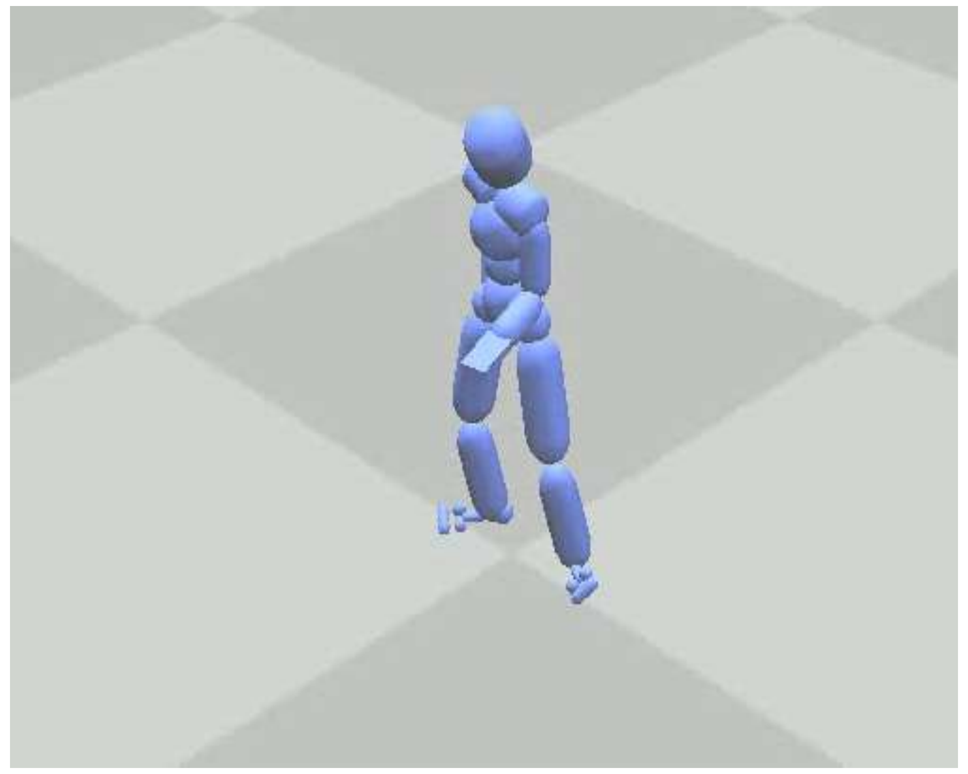
Results: HumanAct12

MLD | MDM | Ours

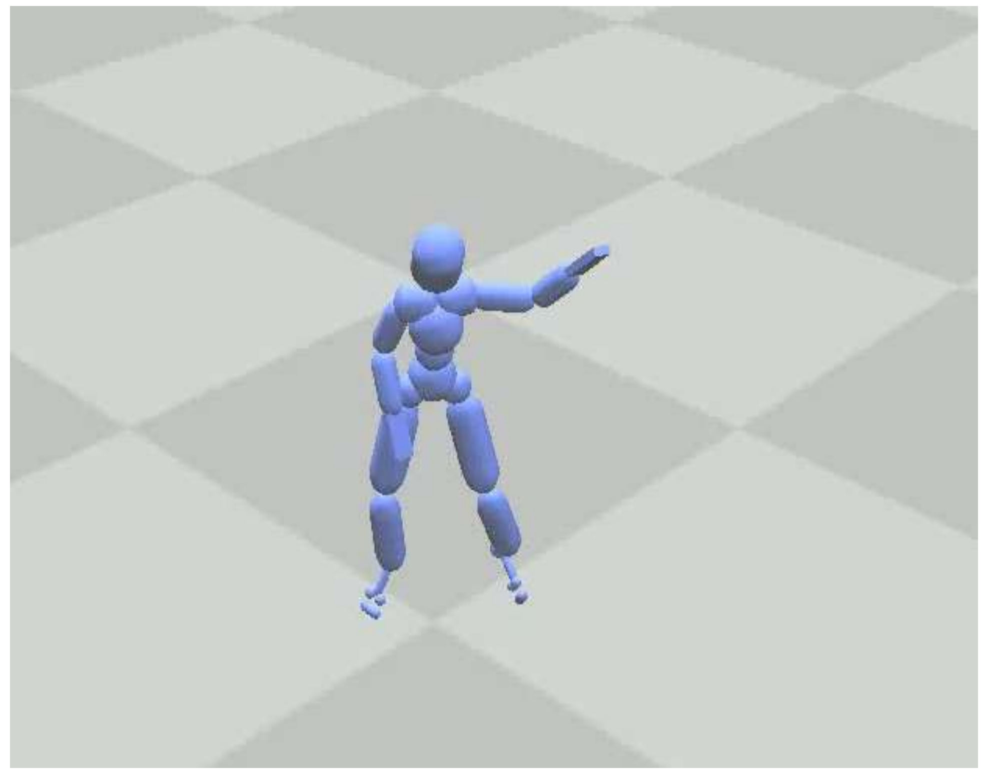
Eat

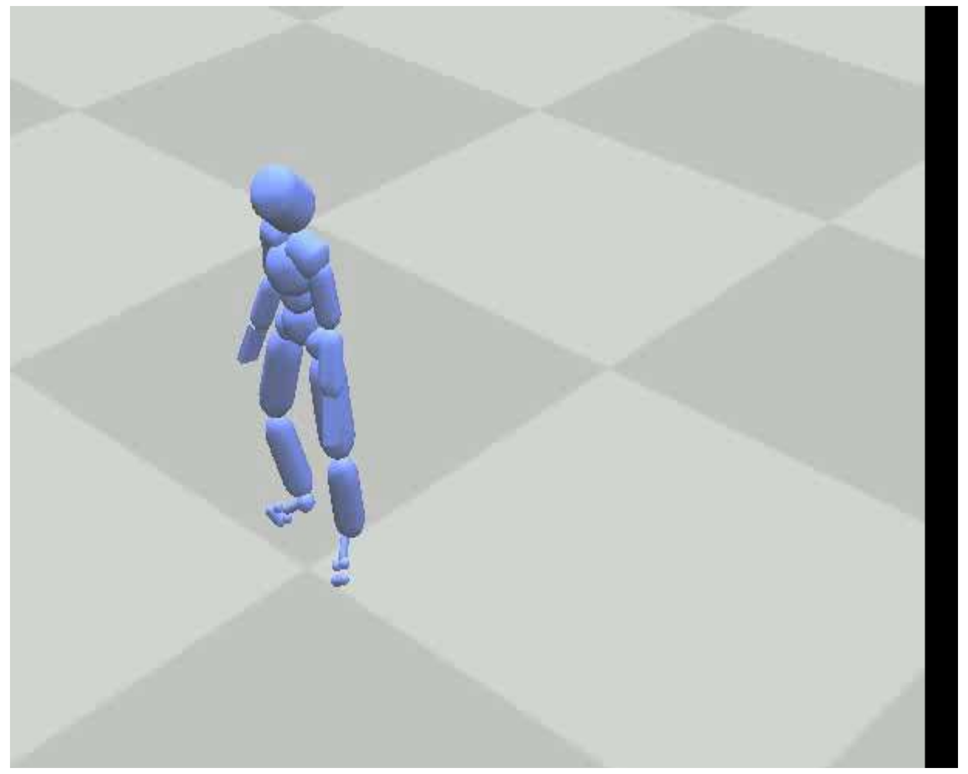


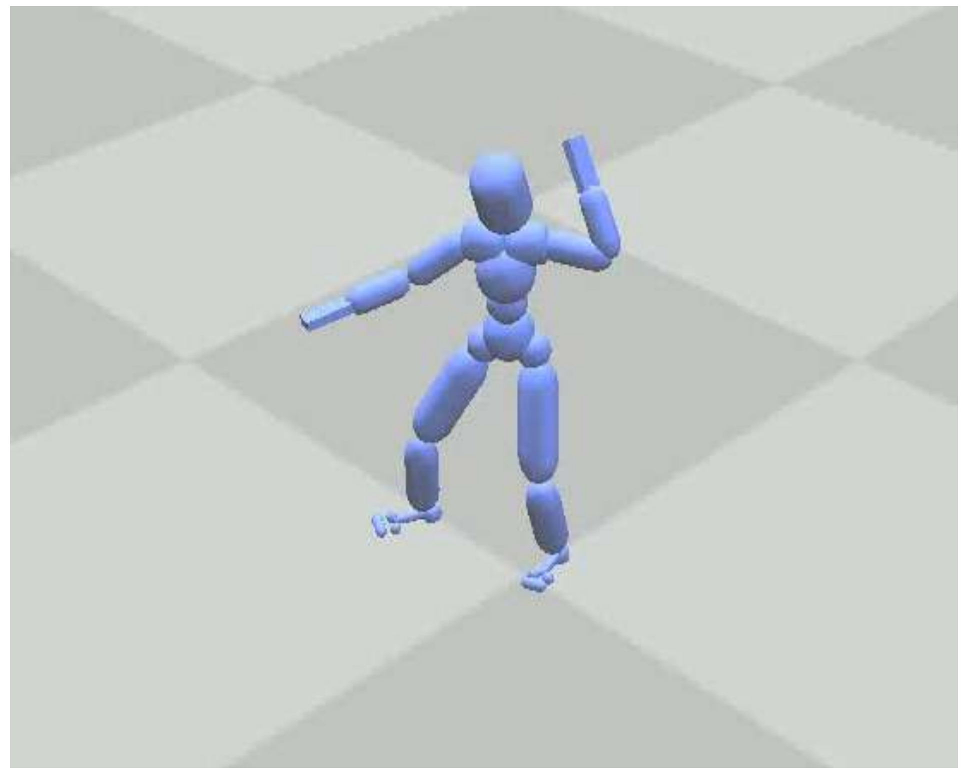




Throw



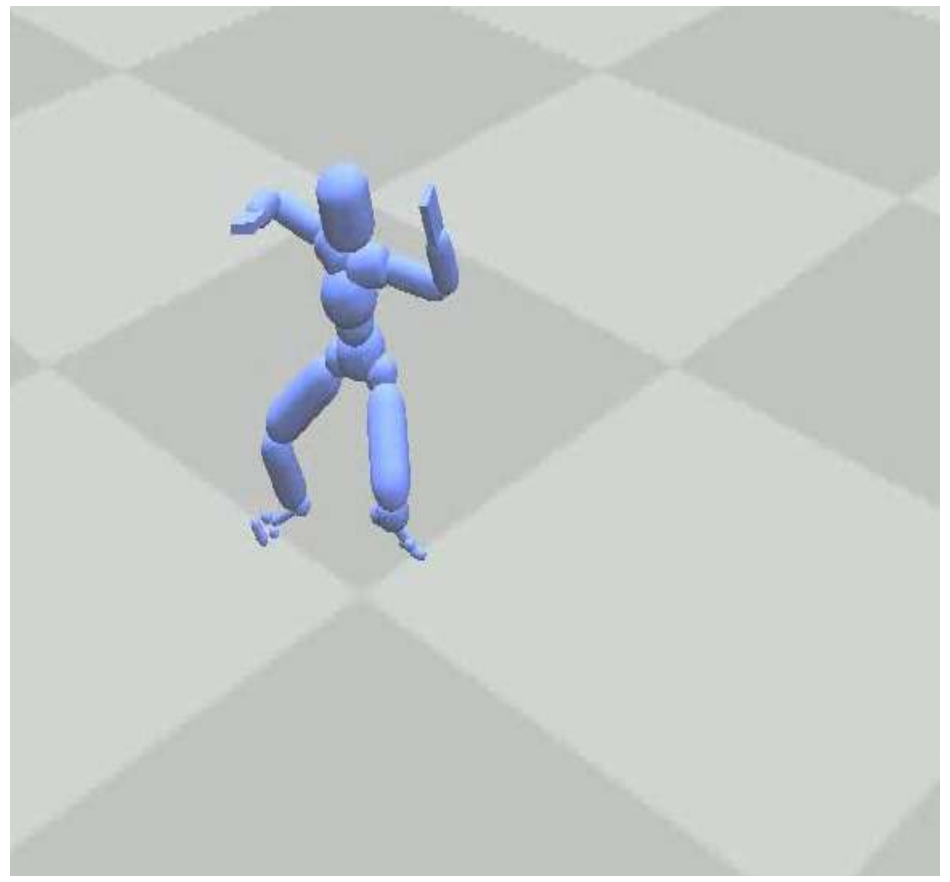


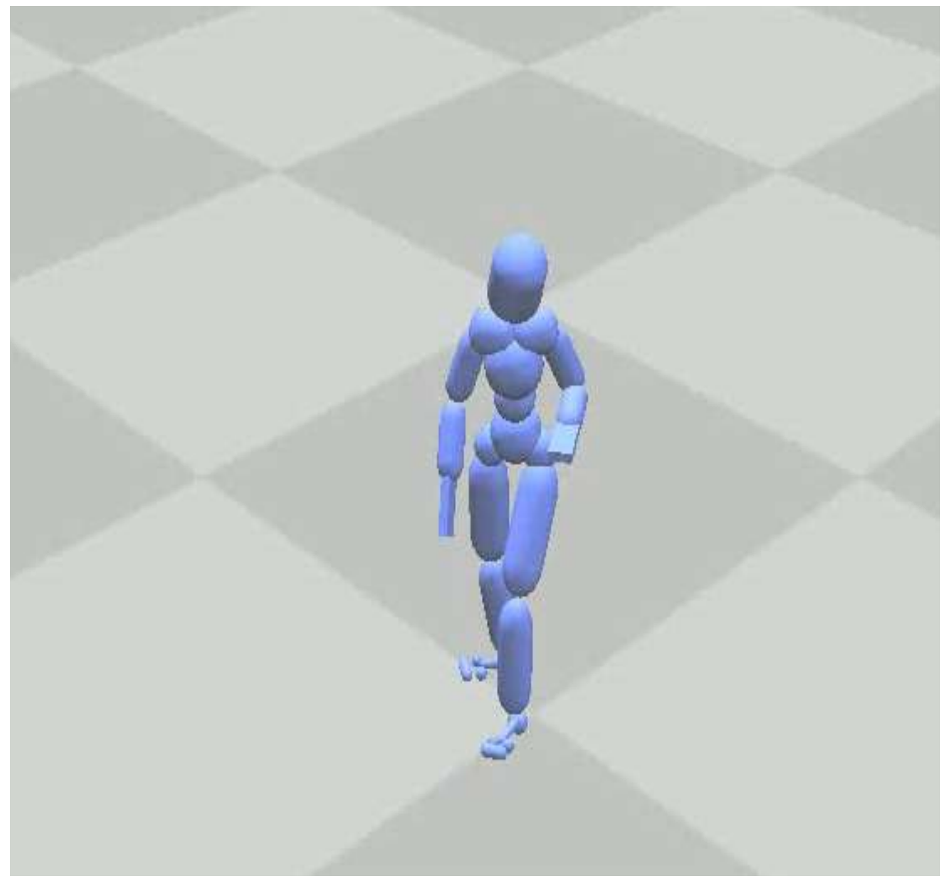


Results: Olympic Sports

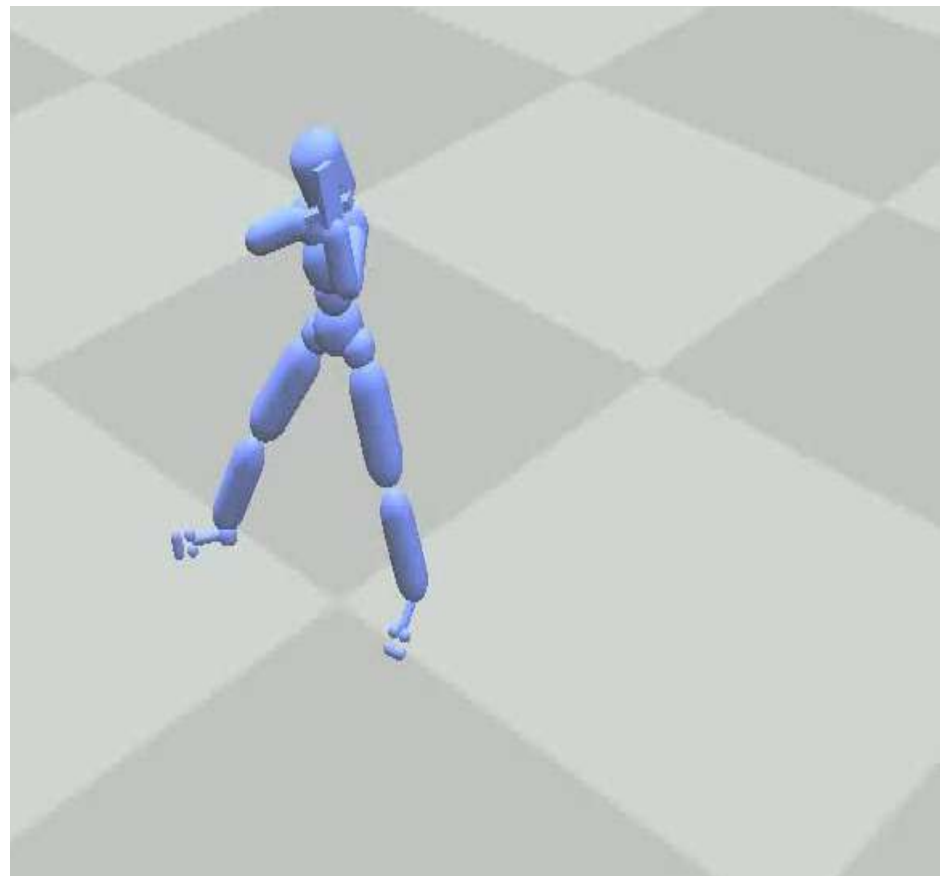
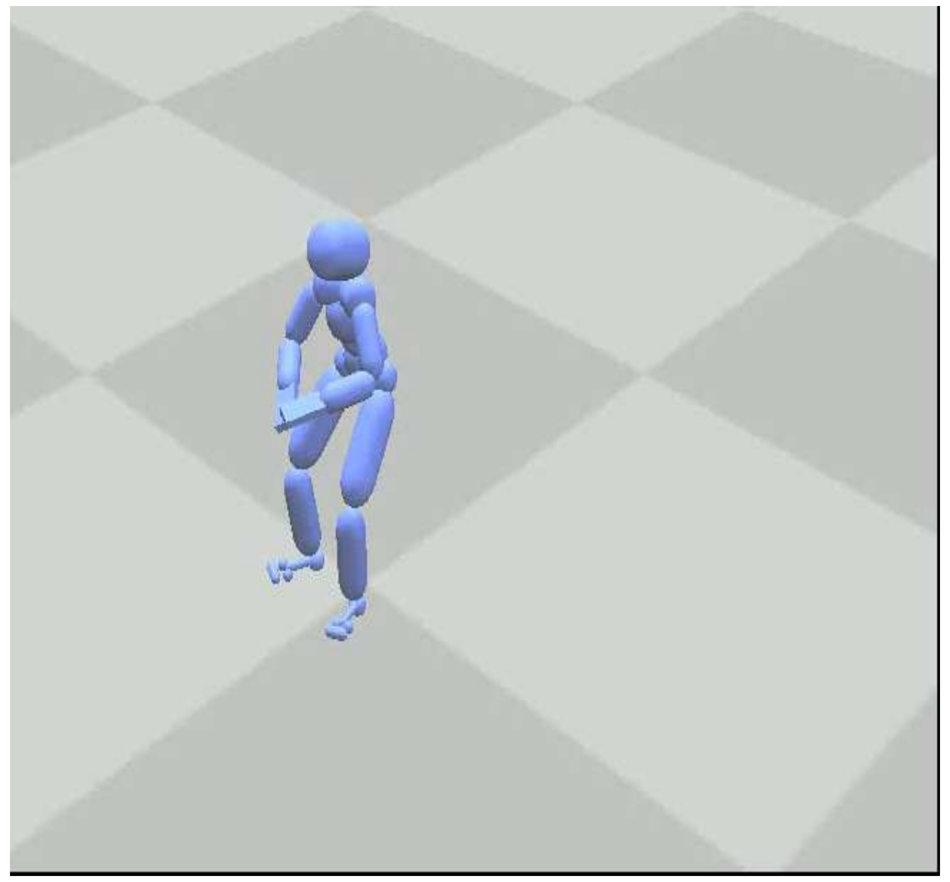
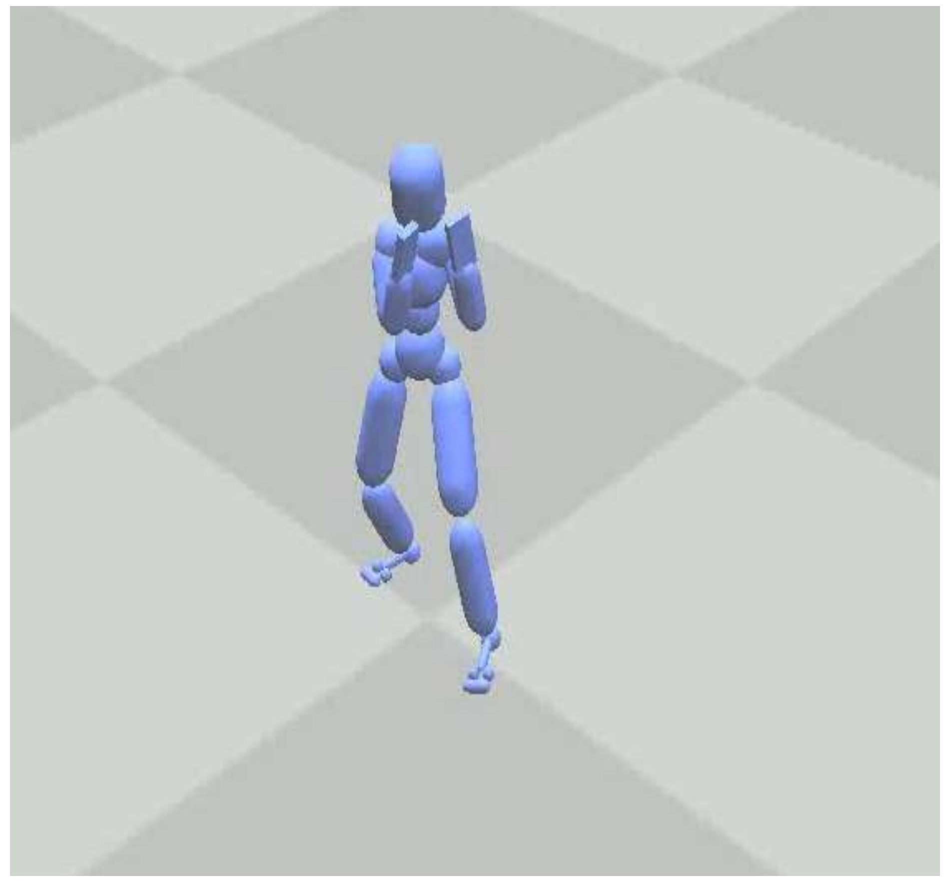
MotionCLIP | AvatarCLIP | Ours

Tae kwon do

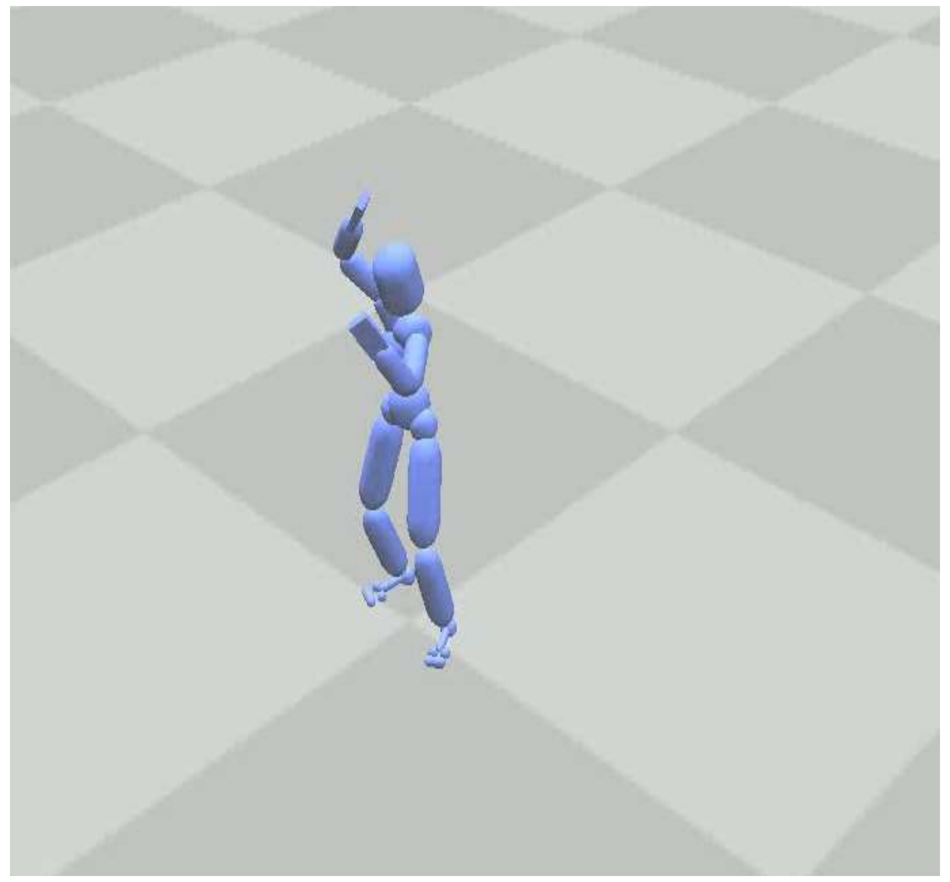
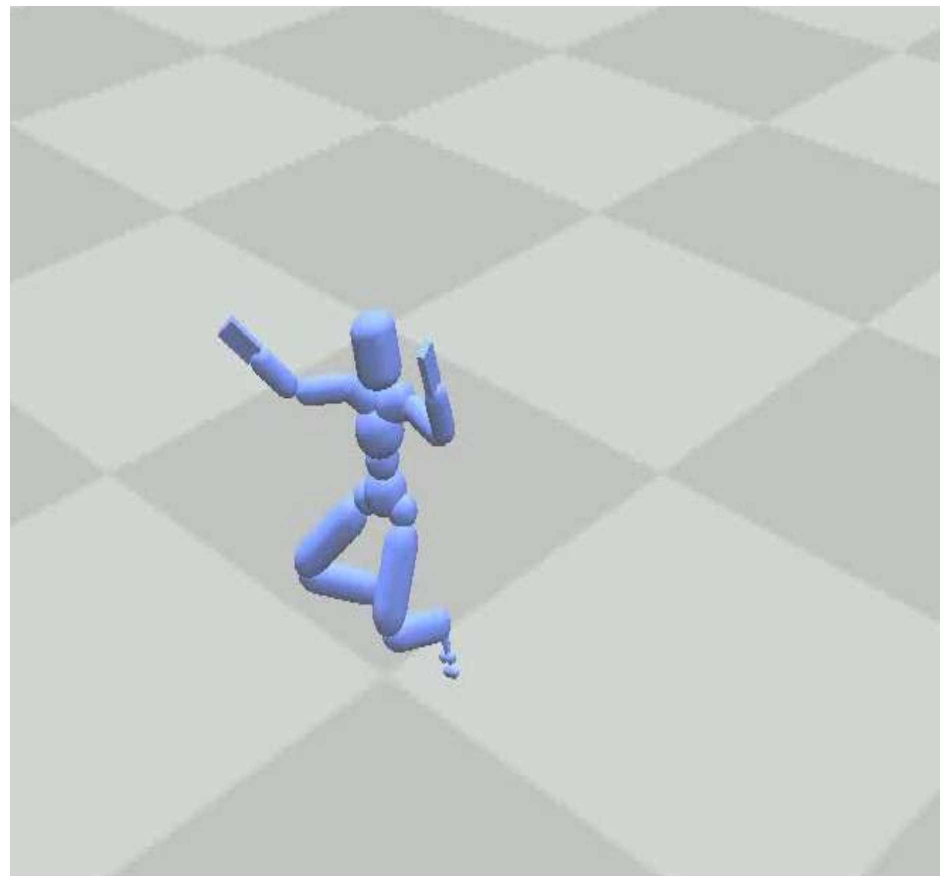
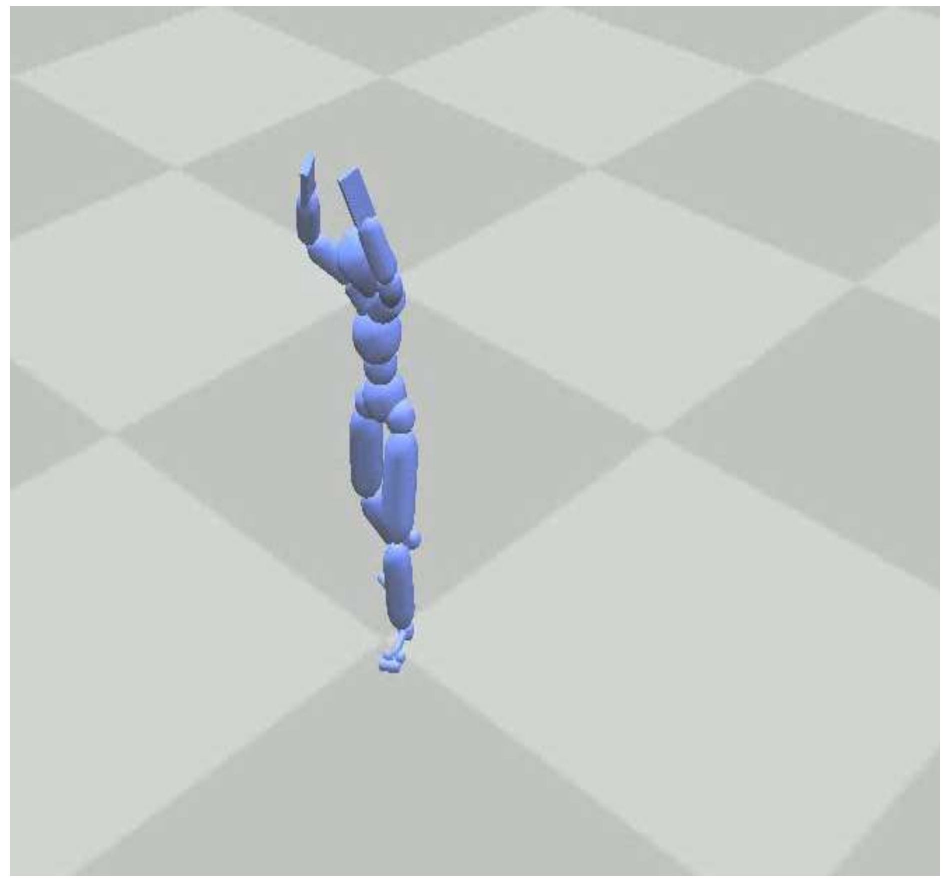
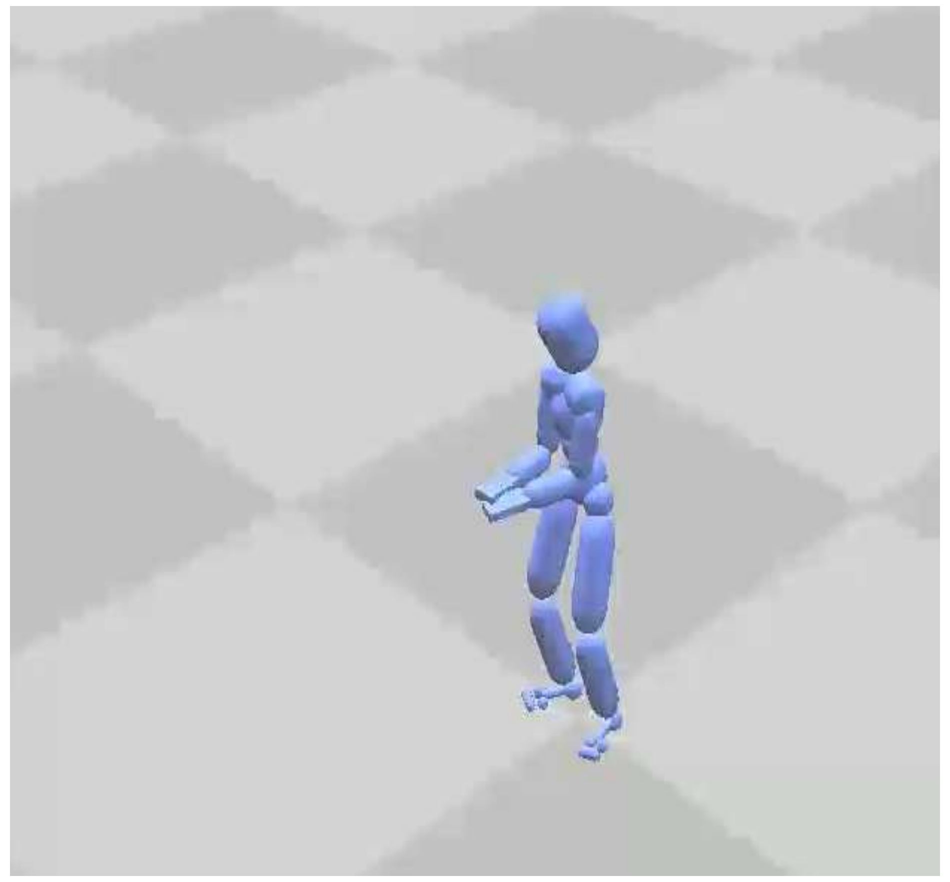




Boxing

Results: Olympic Sports

	MotionCLIP	AvatarCLIP	Ours
Hand ball			
Jump shot			

Results: Human-Scene Interaction

Sit down

Table 5: Human-Scene Interaction. FreeMotion achieves good results on three interaction tasks.

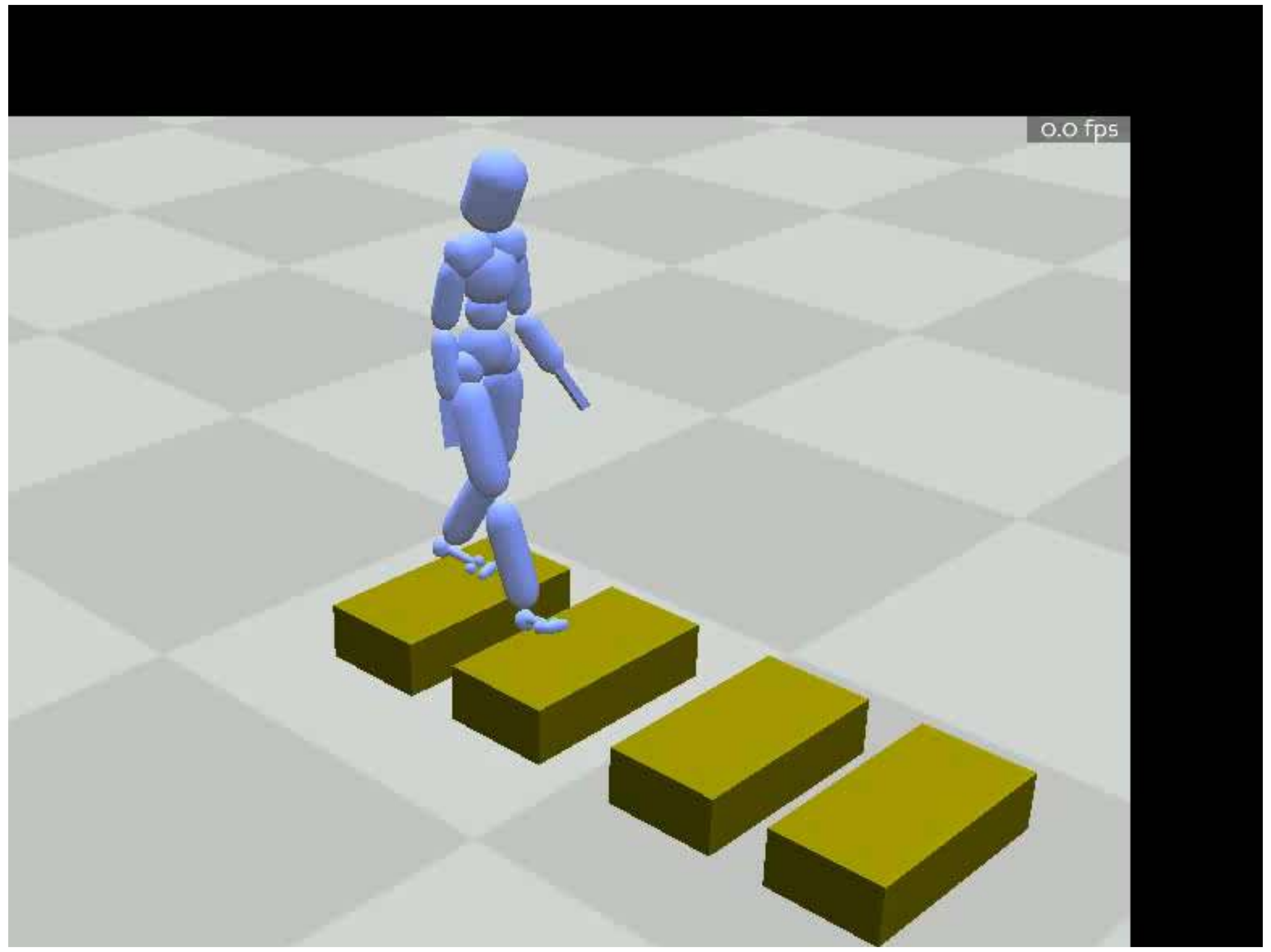
Methods	Success Rate (%) ↑			Contact Error ↓		
	Sit	Lie Down	Reach	Sit	Lie Down	Reach
InterPhys - Sit [14]	93.7	-	-	0.09	-	-
InterPhys - Lie Down [14]	-	80.0	-	-	0.30	-
UniHSI [41]	94.3	**81.5**	**97.5**	**0.032**	**0.061**	0.016
AMP-Sit [28]	83.6	-	-	0.074	-	-
AMP-Lie Down [28]	-	28.3	-	-	0.334	-
AMP-Reach [28]	-	-	96.6	-	-	0.041
Ours	**95**	60	95	0.066	0.224	**0.012**

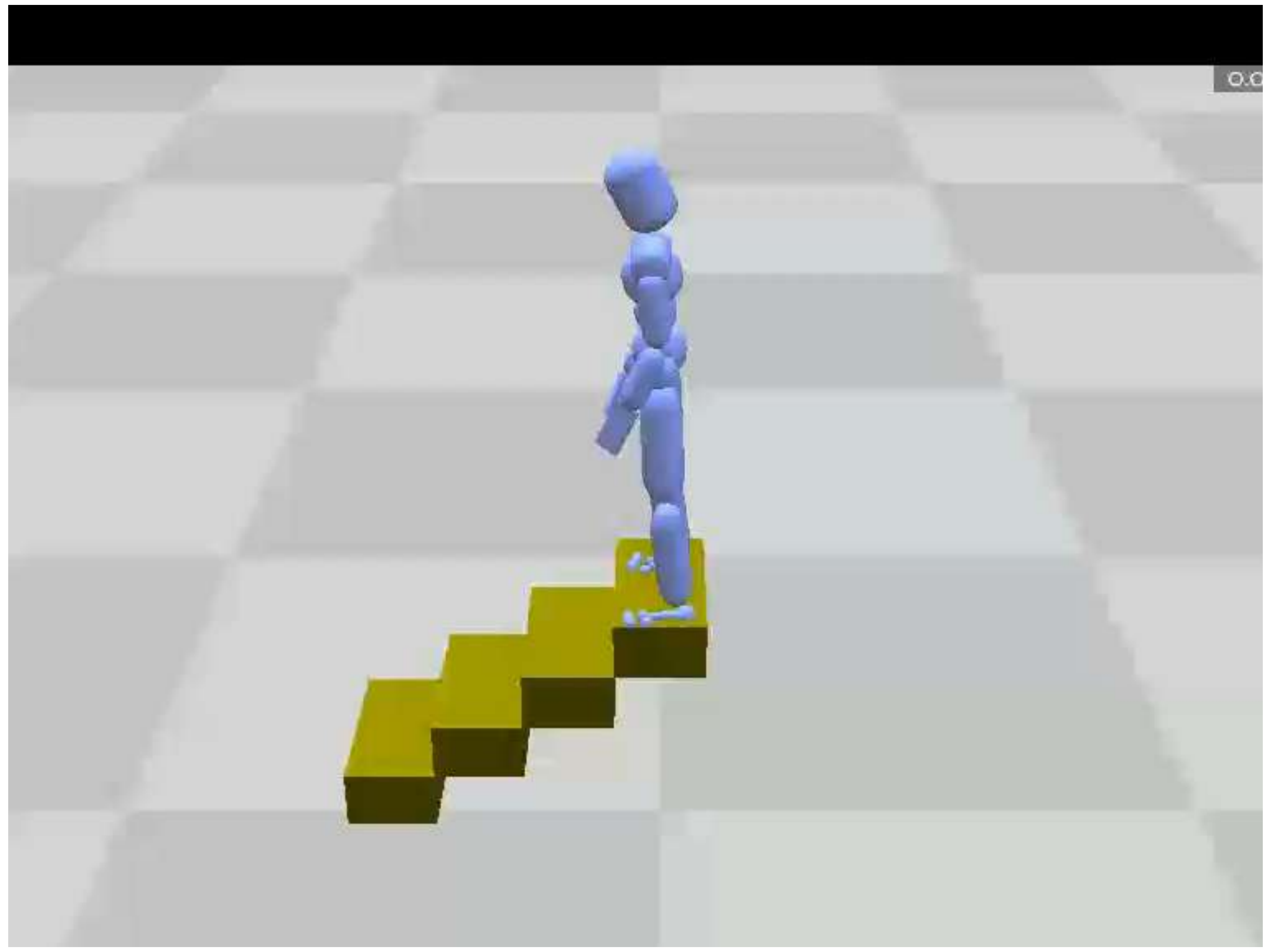
Results: Stepping Stones

Different terrains

Table 6: Stepping Stones. Please see the text for a detailed explanation of the numbers.

Task Parameter	ALLSTEPS [42]	Ours
Flat ($\Theta = 0$)		
$\Phi = 0$	**1.45, 1.50**	1.40, 1.45
$\Phi = 20$	1.35, 1.40	**1.40, 1.40**
Single-step ($\Phi = 0$)		
$\Theta = 50$	**0.80, 0.80**	0.60, 0.75
$\Theta = -50$	0.90, 0.95	**1.00, 1.10**
Continuous-step ($\Phi = 0$)		
$\Theta = 50$	—, 0.65	**0.50, 0.65**
$\Theta = -50$	0.65, 0.70	**0.75, 0.85**
Spiral ($\Phi = 20$)		
$\Theta = 30$	**0.80, 0.85**	0.40, 0.80
$\Theta = -30$	1.00, 1.10	**1.10, 1.30**





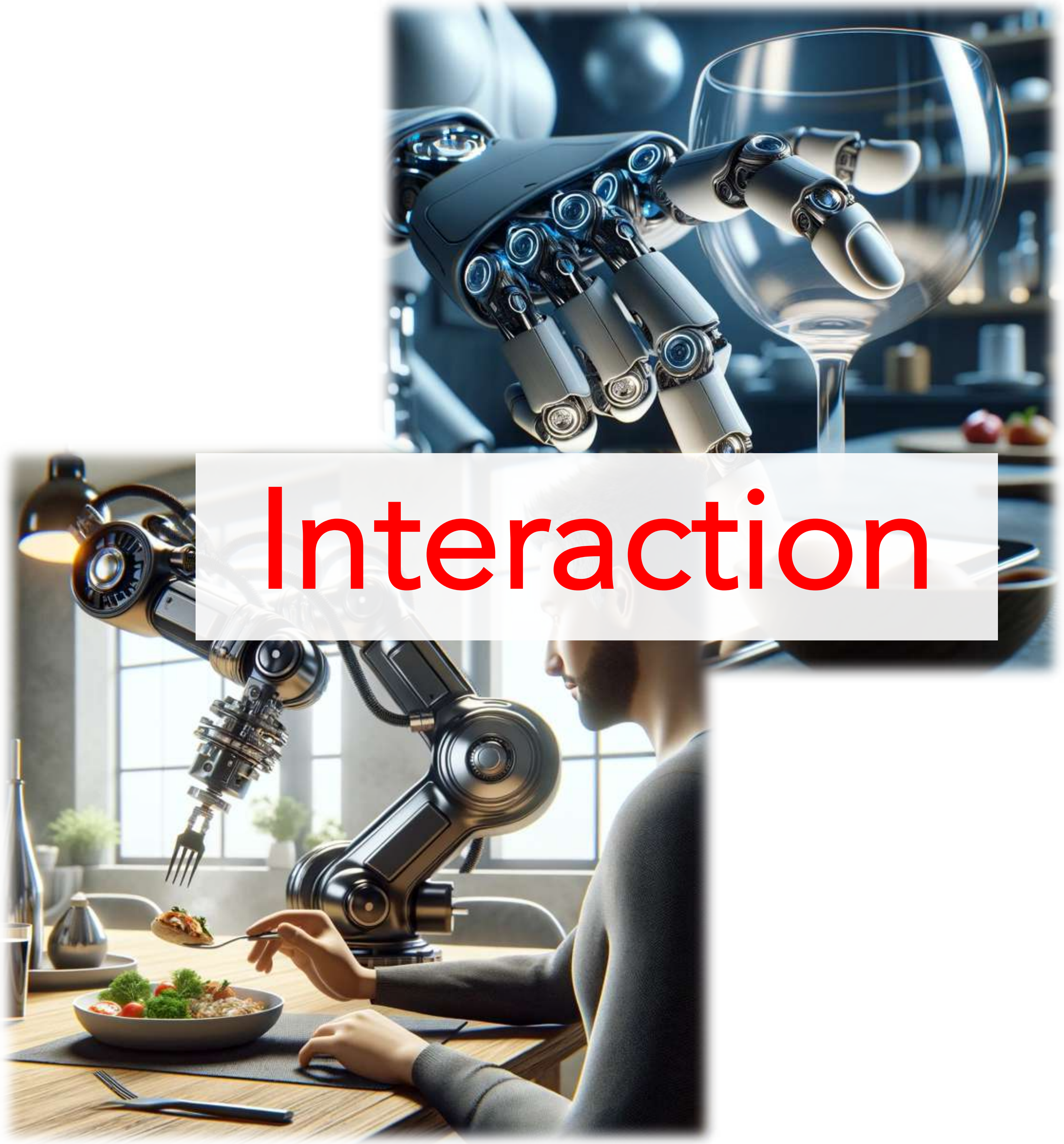
Take Home Messages

- MLLMs contain strong common sense knowledge useful for physical interaction synthesis.
- A digital physical humanoid can interact with the open world solely driven by MLLMs.
- Real humanoid robots can potentially learn from the generated dynamic motion trajectories.

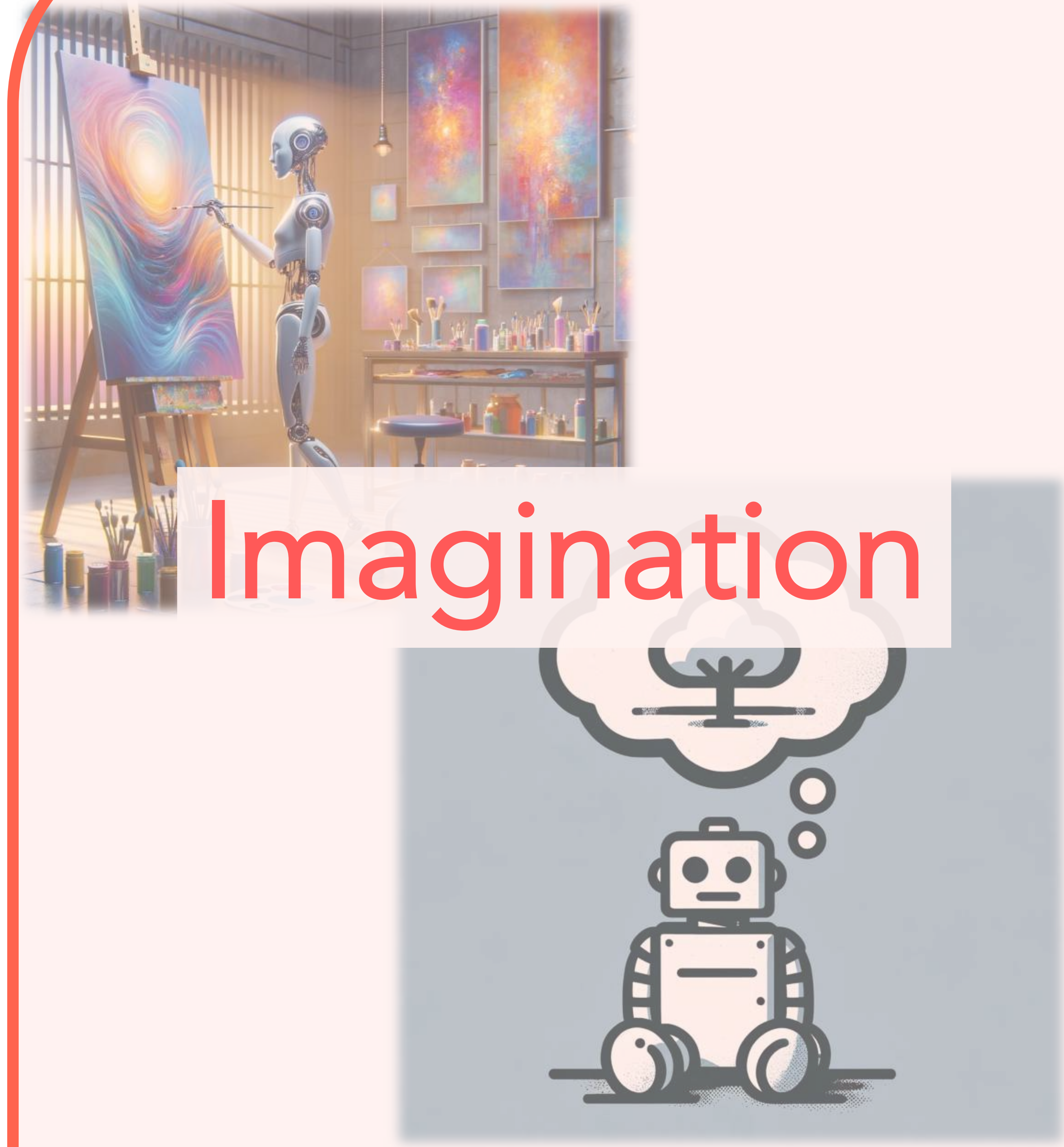
Human-Centered Robot Simulator
Human Interaction Capturing
Interactable Asset Creation
Police Car
Dragon Chair
Scissor
Data Driven
Asset Support
Human Interaction Synthesis
Human-Centered Robot Simulator
Human Simulation
Human-Centered EAI
Open-World Perception
Visual Perception
Human-Centered Robotics
Simulation Support
Collaborative Transport
Human-to-Robot Handover

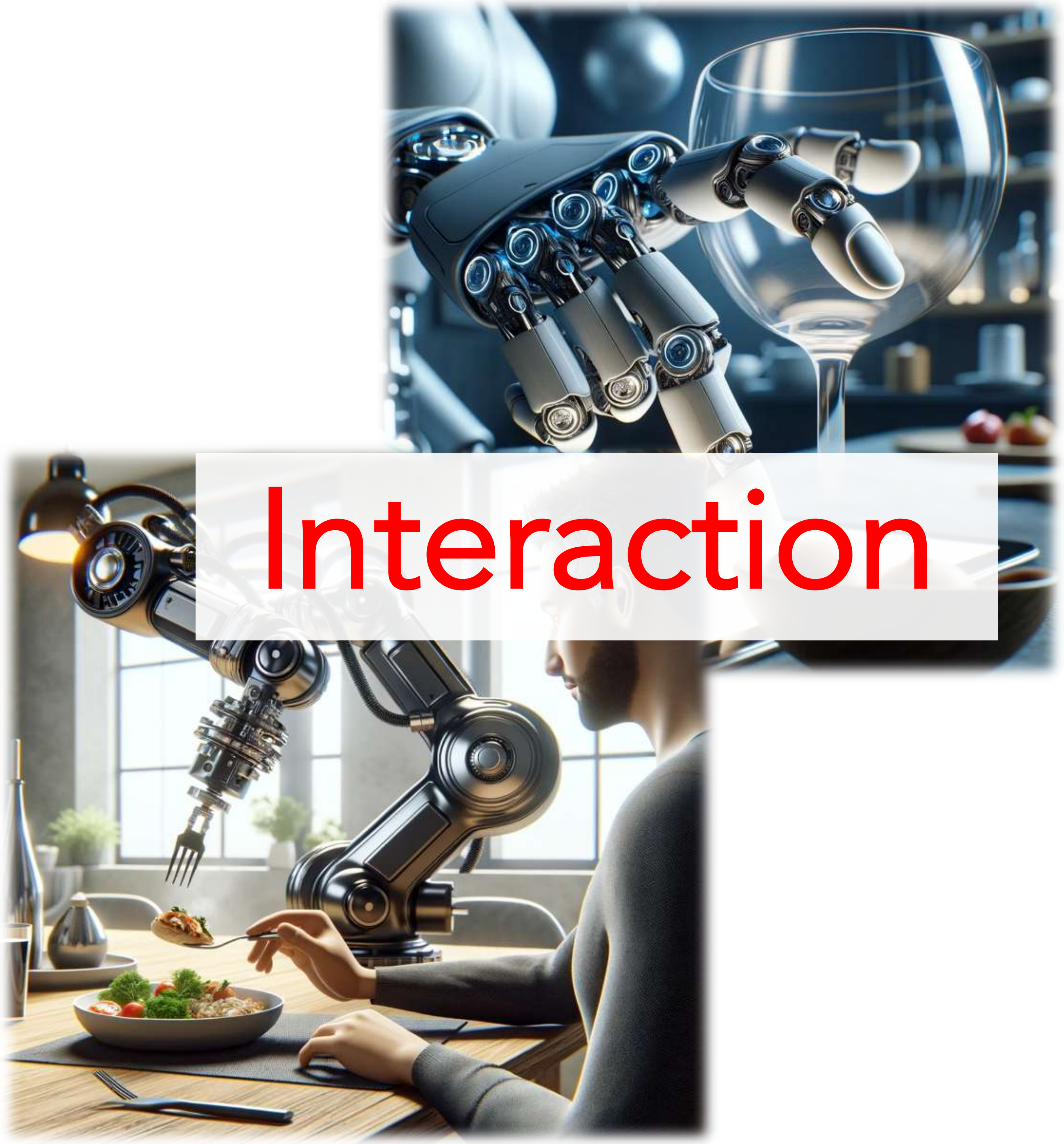

Human-Centered Robot Simulator
Human-Centered EAI
Human Interaction Capturing
Interactable Asset Creation
Open-World Perception
Police Car
Dragon Chair
Scissor
Data Driven
Asset Support
Visual Perception
Human Interaction Synthesis
Human-Centered Robot Simulator
Human-Centered Robotics
Human Simulation
Simulation Support
Collaborative Transport
Human-to-Robot Handover

Embodied Intelligence



Embodied Intelligence



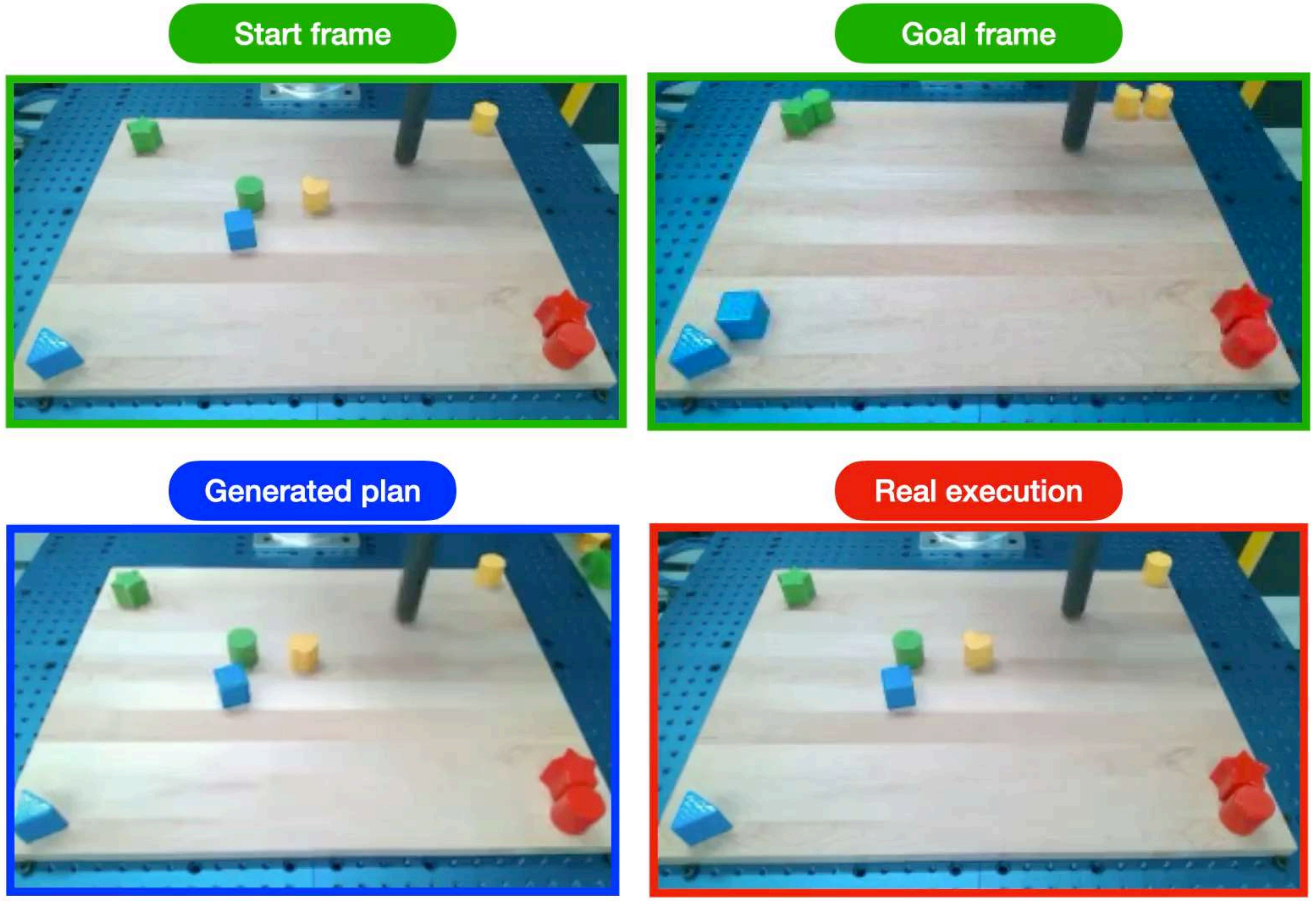


Imagination

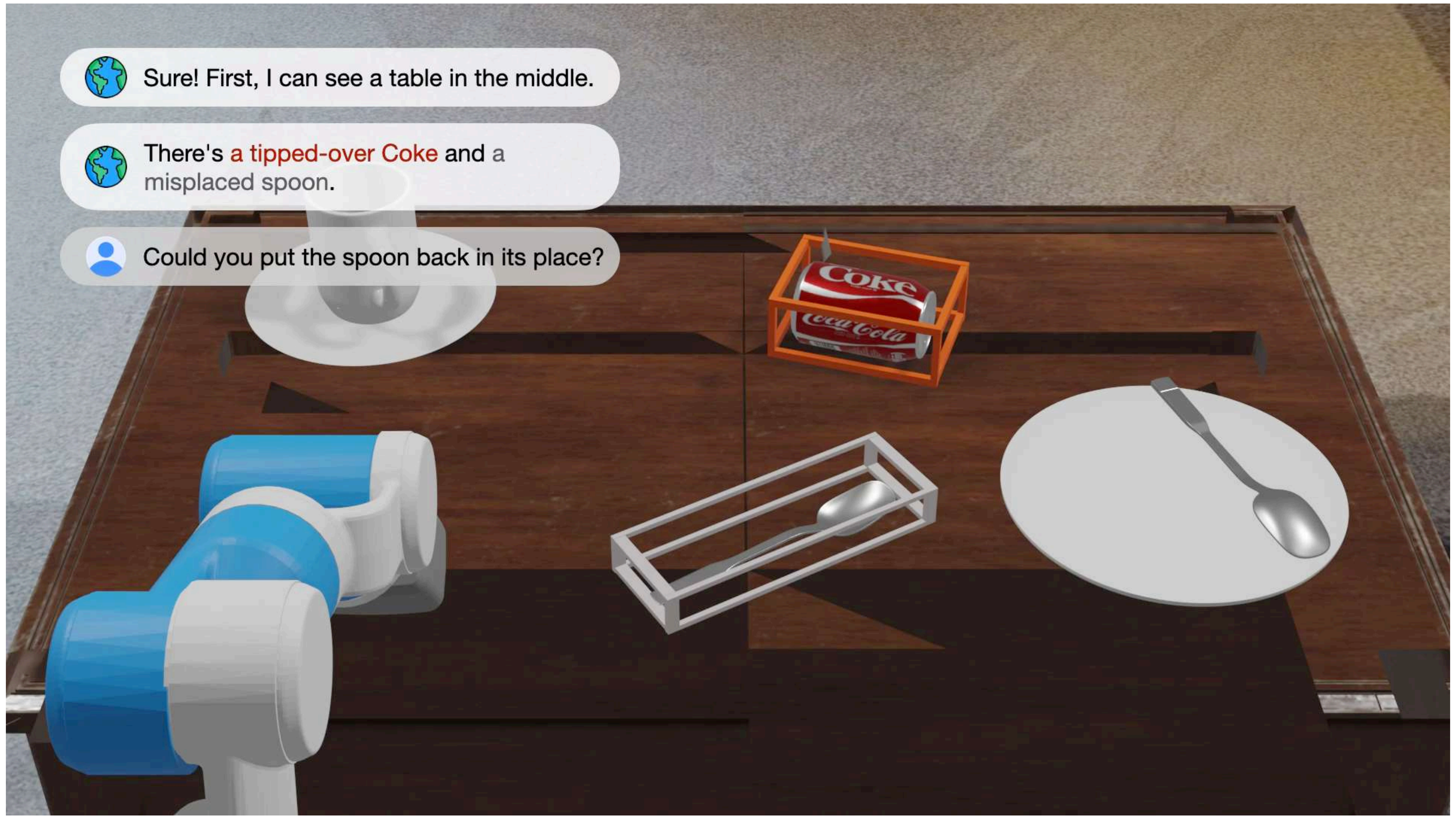
Reasoning

Interaction

Imagination: Generative Intelligence Empowered World Models



UniSim



3D-VLA

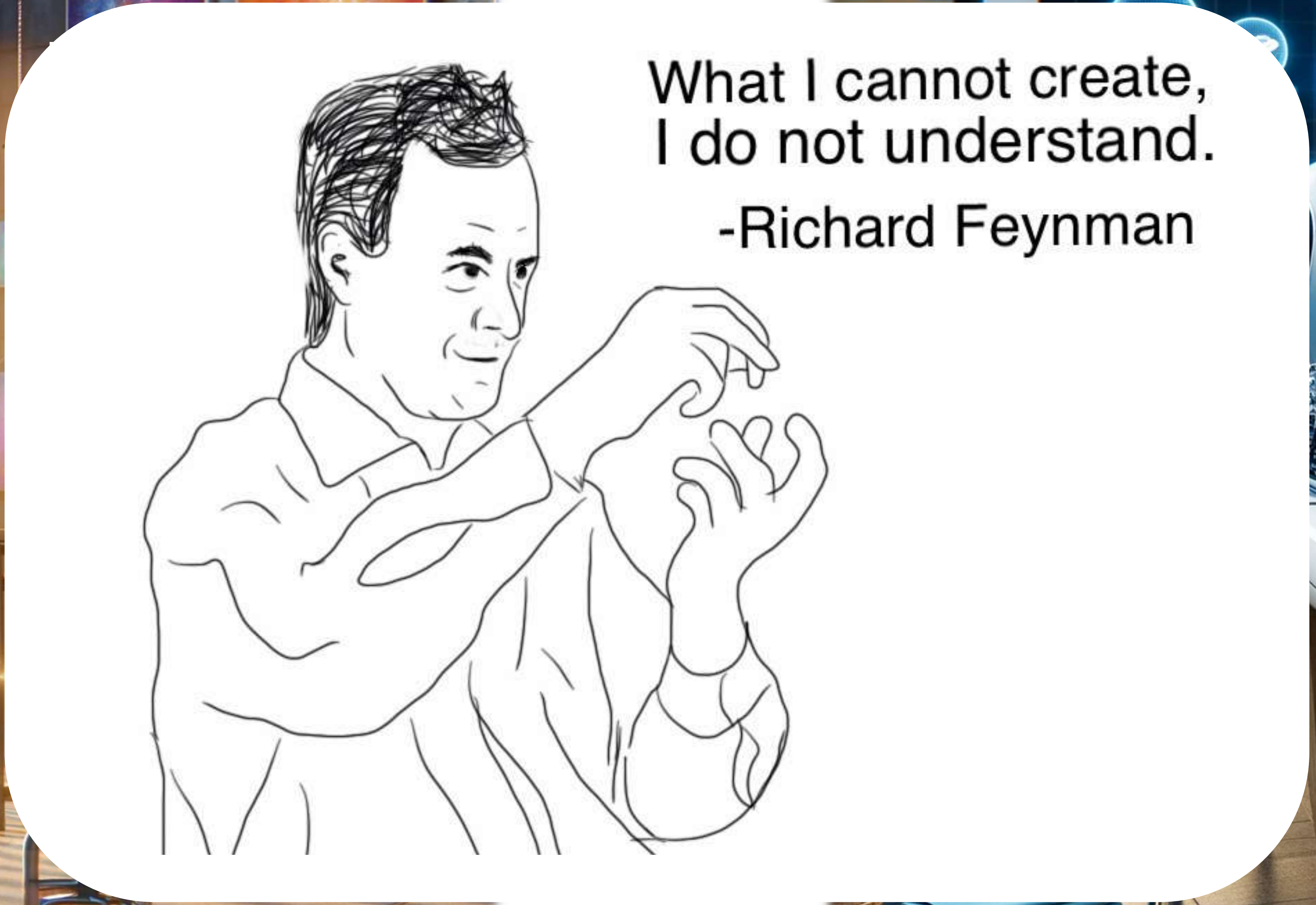
Yang et al., 2024. "Learning Interactive Real-World Simulators".
Zen et al., 2024. "3D-VLA: A 3D Vision-Language-Action Generative World Model".

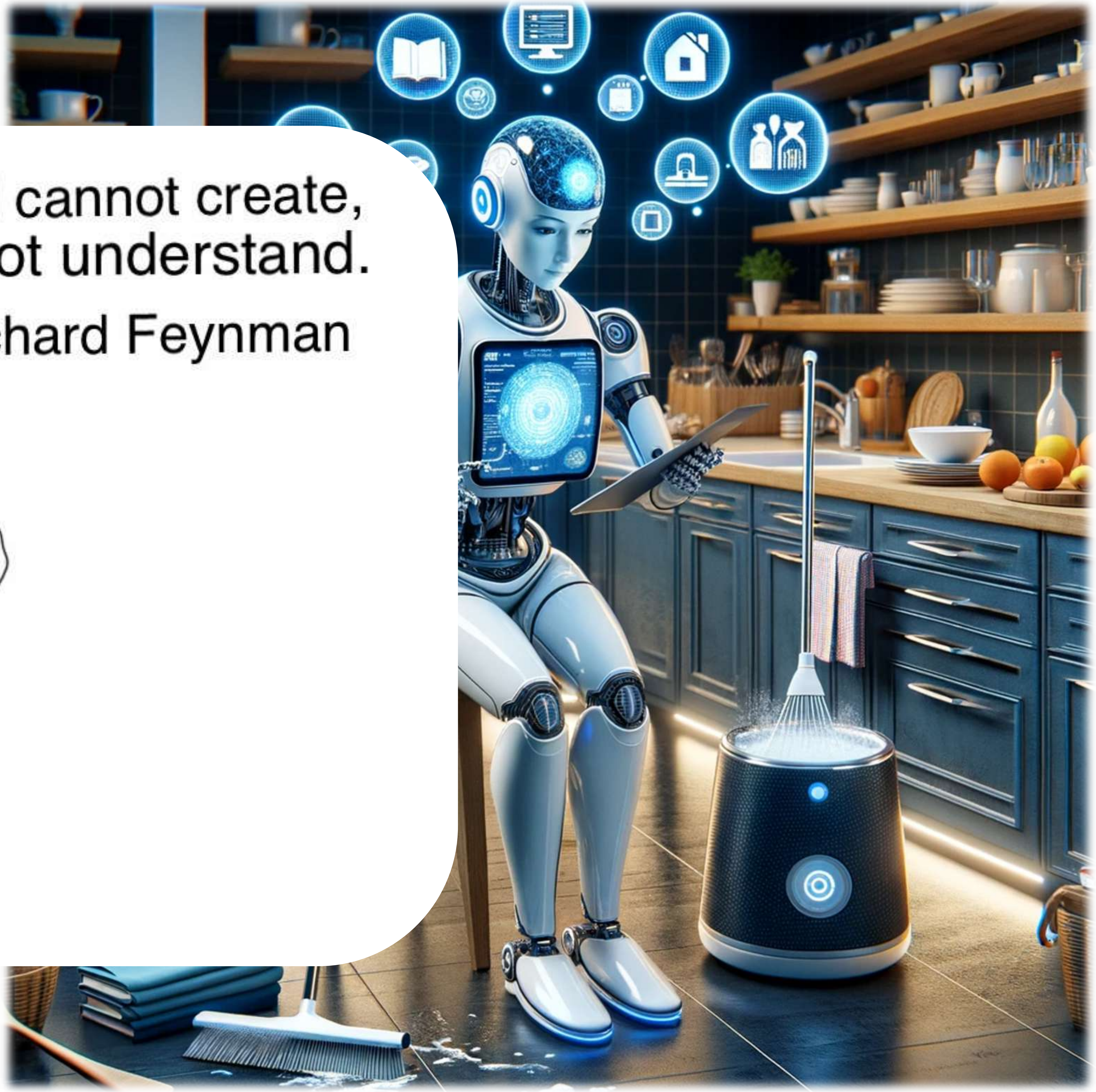
Key Observation: Synergy Between Reasoning and Imagination

Imagination

Reasoning

Key Observation: Synergy Between Reasoning and Imagination





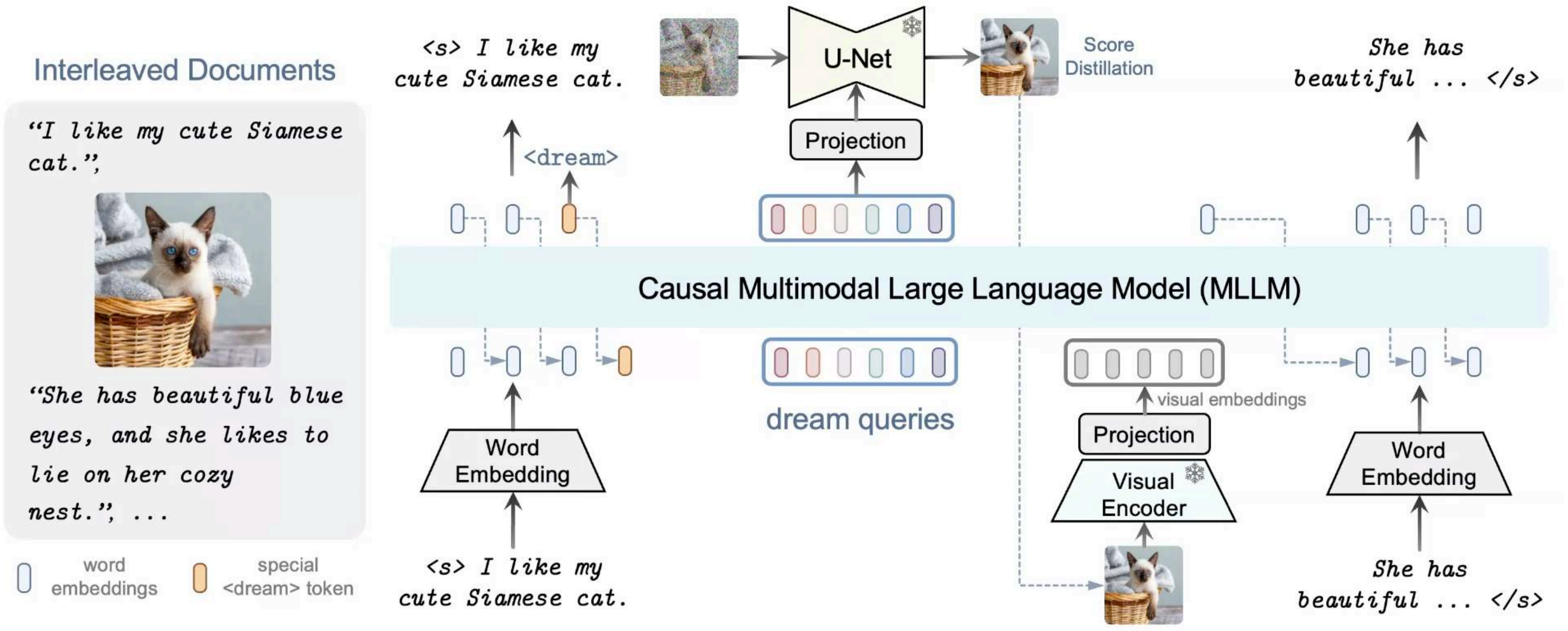
Imagination

Reasoning

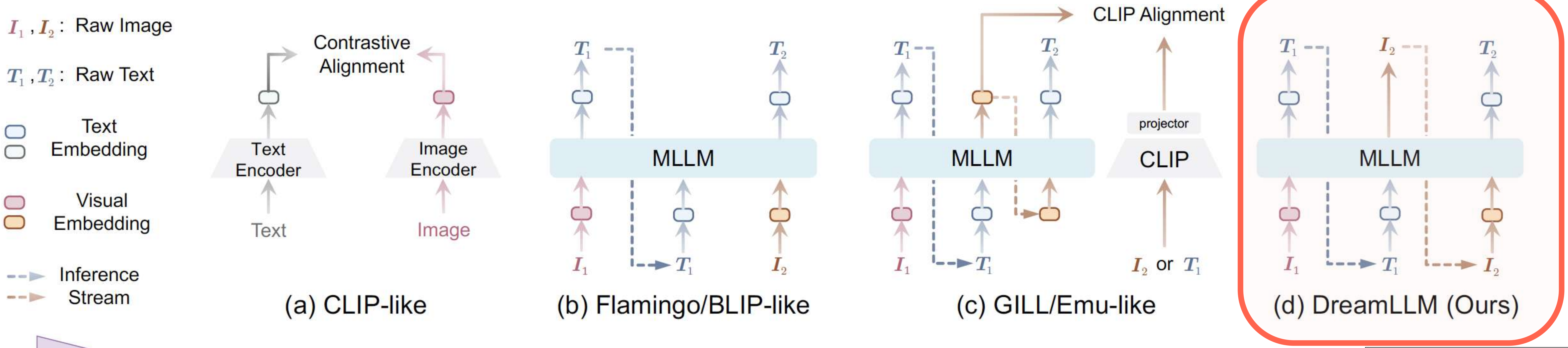
DreamLLM

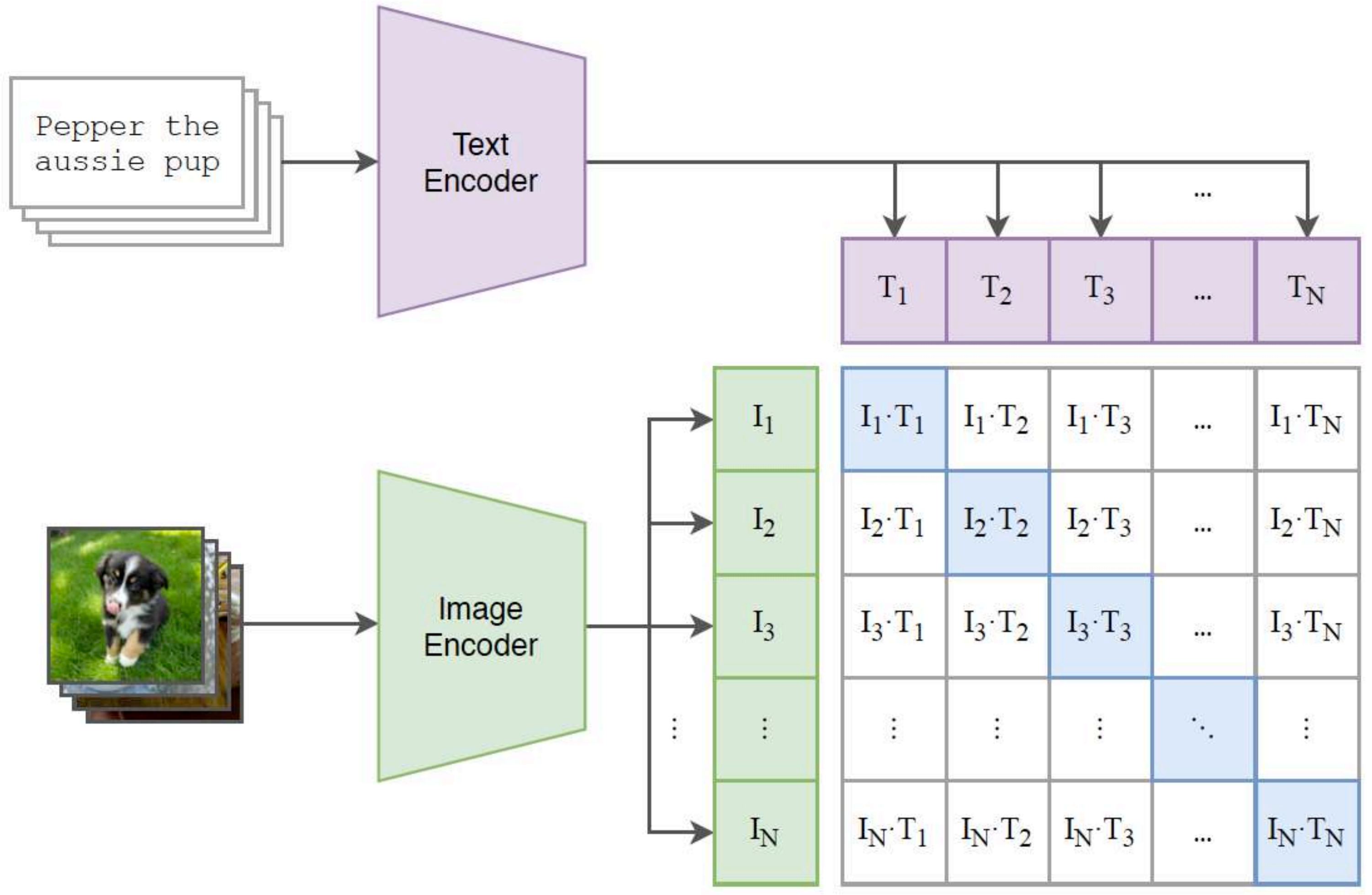
DreamLLM: Synergistic Multimodal Comprehension and Creation
Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma†, Li Yi†. ICLR 2024 (spotlight).

DreamLLM – Pipeline

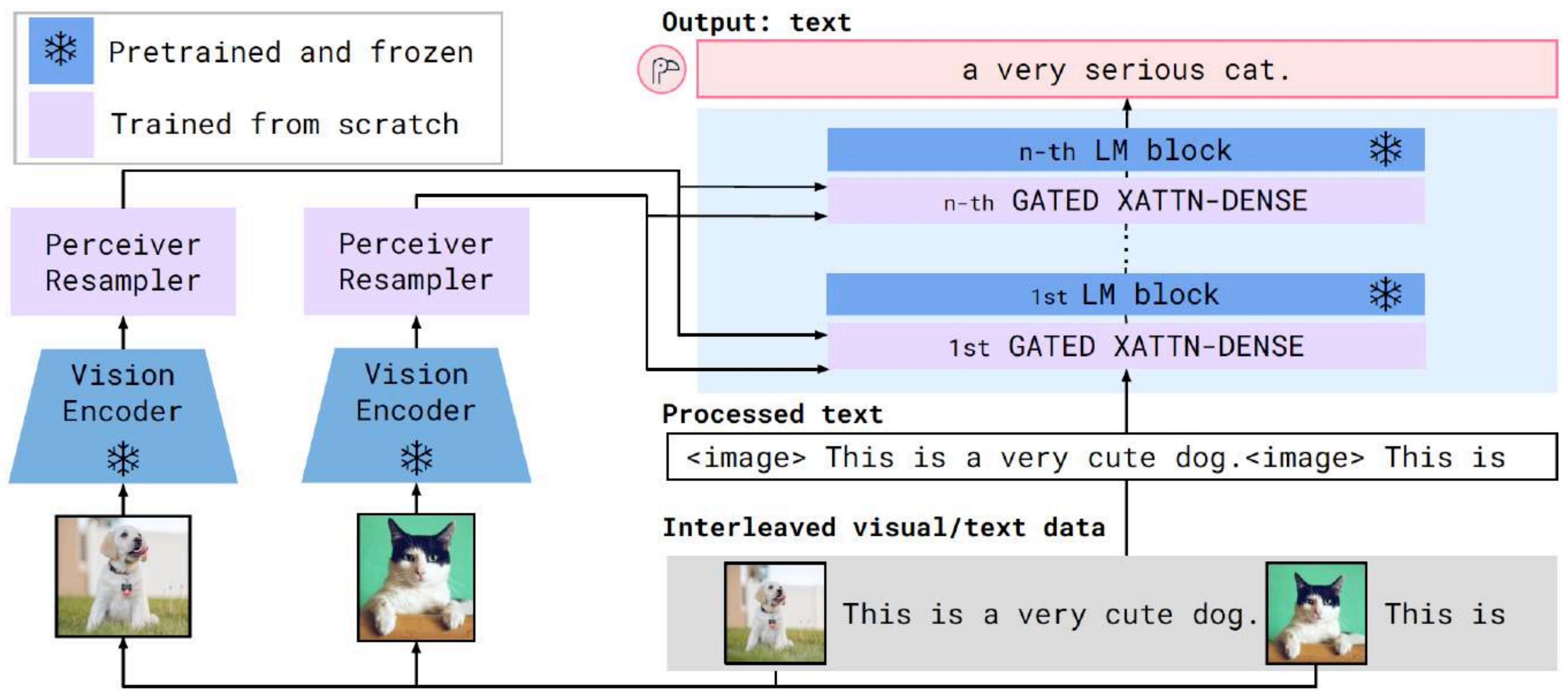


A Conceptual Comparison with Related Works

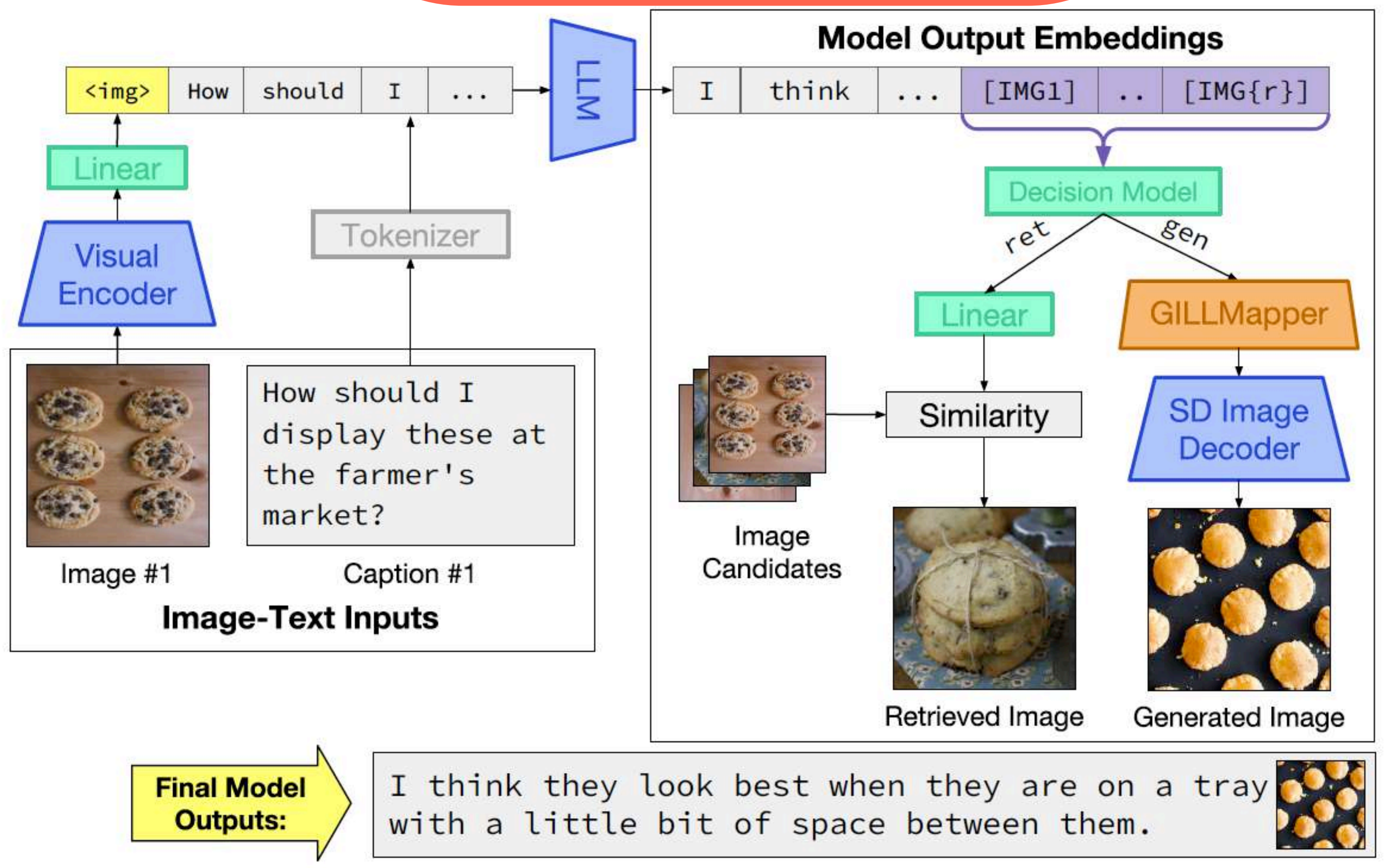




CLIP



Flamingo



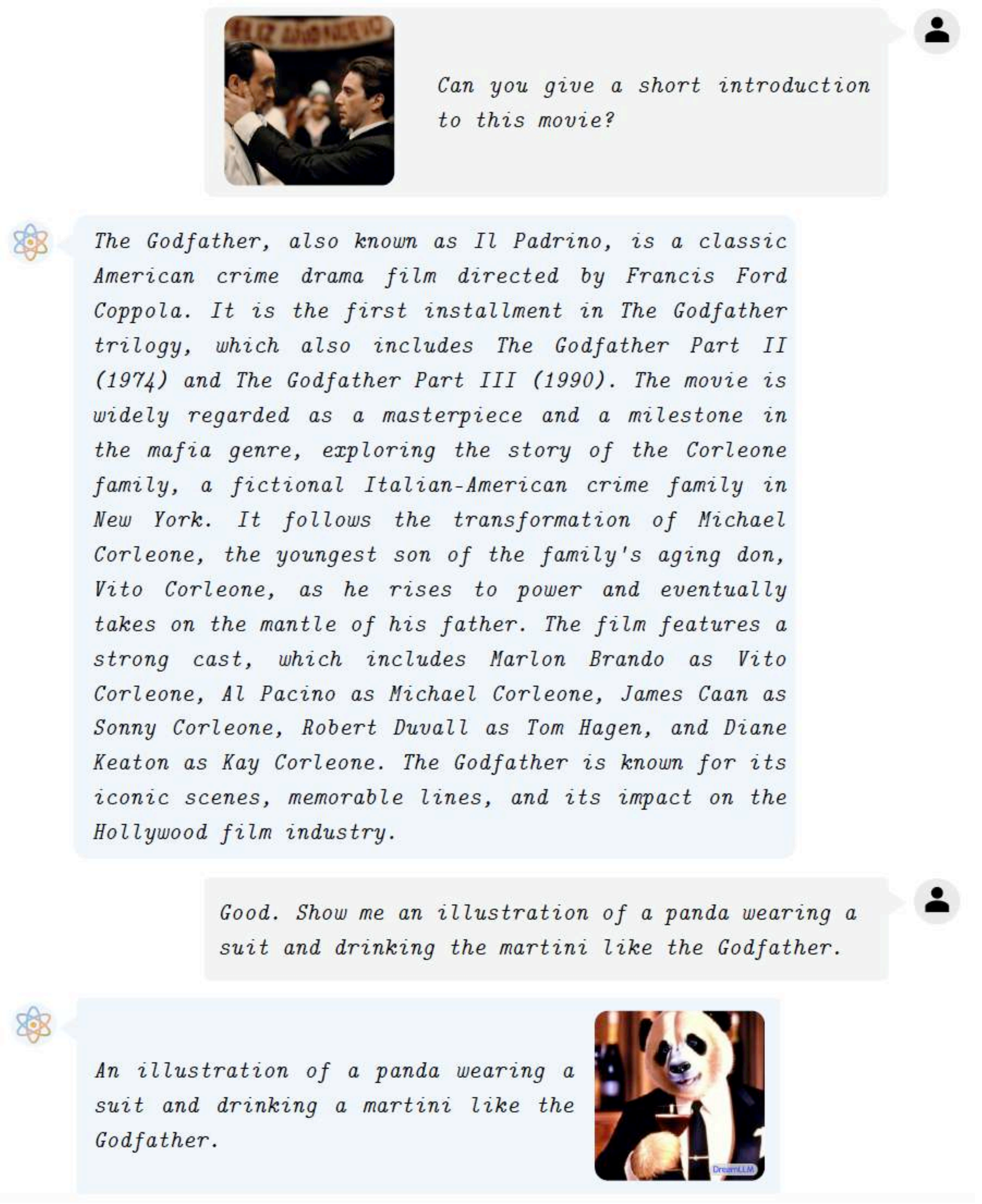
GILL

Radford et al., 2021. "Learning Transferable Visual Models From Natural Language Supervision".
Alayrac et al., 2022. "Flamingo: a Visual Language Model for Few-Shot Learning".
Koh et al., 2023. "Generating Images with Multimodal Language Models".

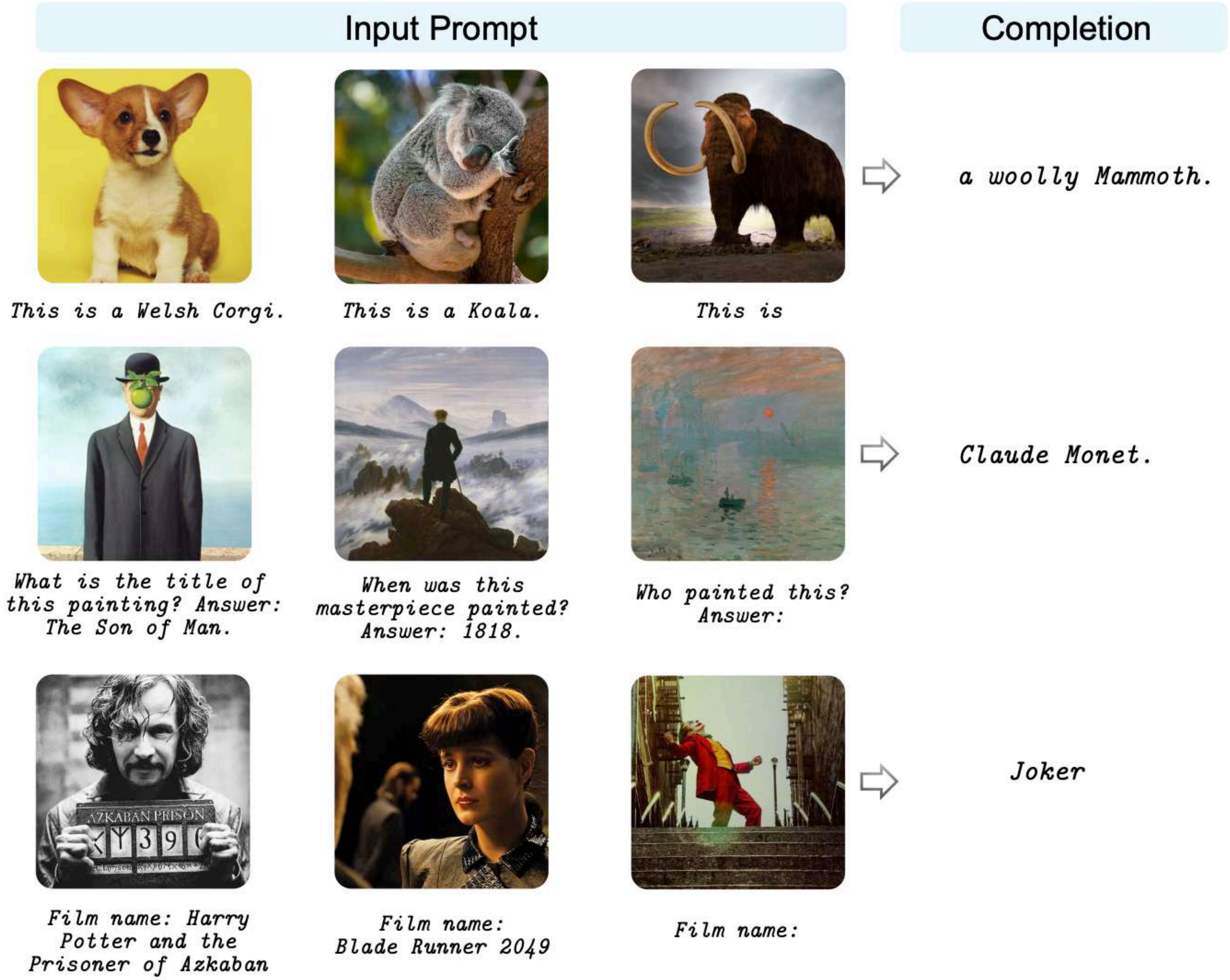
DreamLLM – Take Home Message

- *Generate Everything as It Is:* DreamLLM not only takes all modalities raw data as inputs but also as outputs in a truly end-to-end fashion.
  - Avoid aligning with intermediate representations like CLIP which primarily learn modality-shared features while overlooking modality-specific features
- *Interleaved Generative Pre-Training (I-GPT):* DreamLLM is trained to generate interleaved multimodal corpora from the internet.
  - A natural modeling of all joint, marginal, and conditional distributions of images and texts in the document
  - A learning synergy grounding DreamLLM's comprehension in creation and vice versa.

DreamLLM – Experiments



Any-to-Any in Vision Language



In Context Learning

DreamLLM – Experiments

Imagine you are a robot agent in the house. There is a mobile phone on the dining table in the dining room, and you are in the living room. Write a short article about how would you walk through the house to grab the mobile phone with your robotic arm step by step. Include several high-quality photographs to illustrate your procedure.

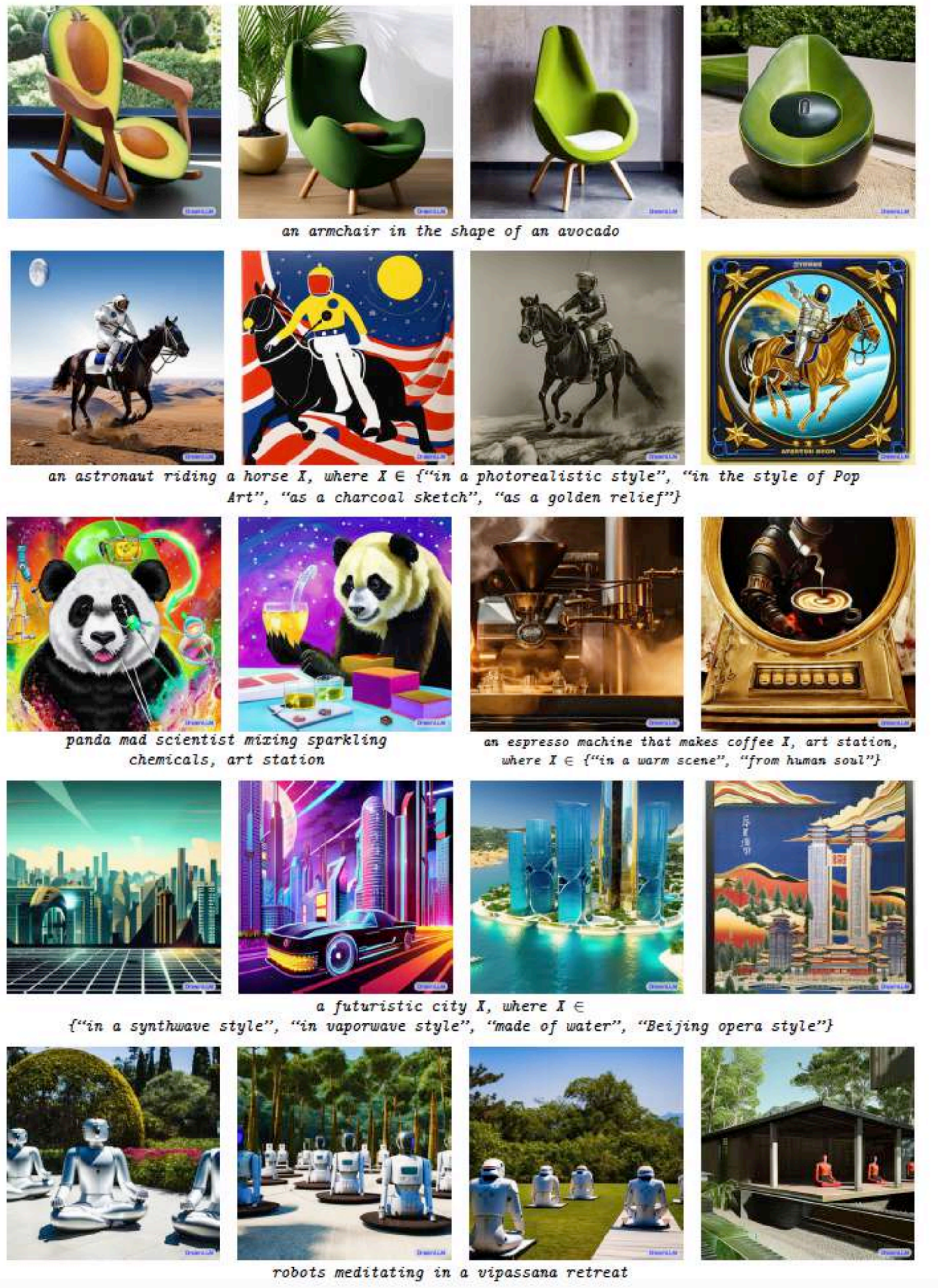
As a robot agent in the house, I would walk through the house to grab the mobile phone in the dining room as shown in the image below. <dream_1>
First, I would look around the living room to make sure the area is safe and there are no obstacles or people that I need to avoid. I would then walk towards the dining room and stop before entering the room.
Next, I would look for the mobile phone on the table, as shown in the image below. <dream_2>

Once I have identified the phone, I would then move closer to it and extend my robot arm to grab it, as shown in the image below. <dream_3>
Once the phone is grabbed, I would then bring it back to the living room and place it on the table where it belongs.

☐ Human Instruction ☐ DreamLLM Generation

Instruction-Following Interleaved Content Creation

DreamLLM – Experiments

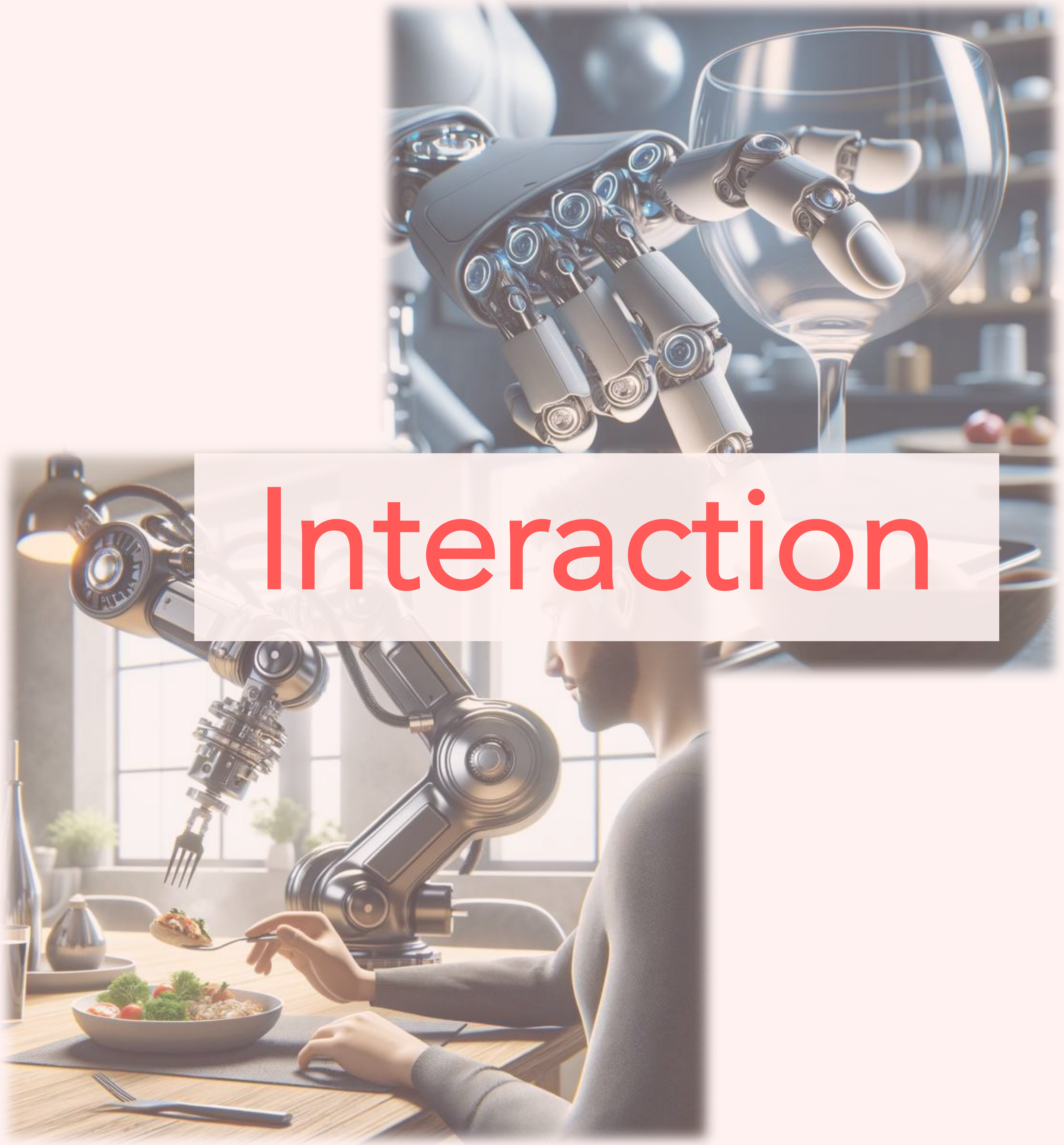


Text-to-Image Generation

(a) In-context Image Edition (b) In-context Subject-Driven Generation (c) In-context Compositional Generation

Image-to-Image & In-Context Image Generation

Embodied Intelligence

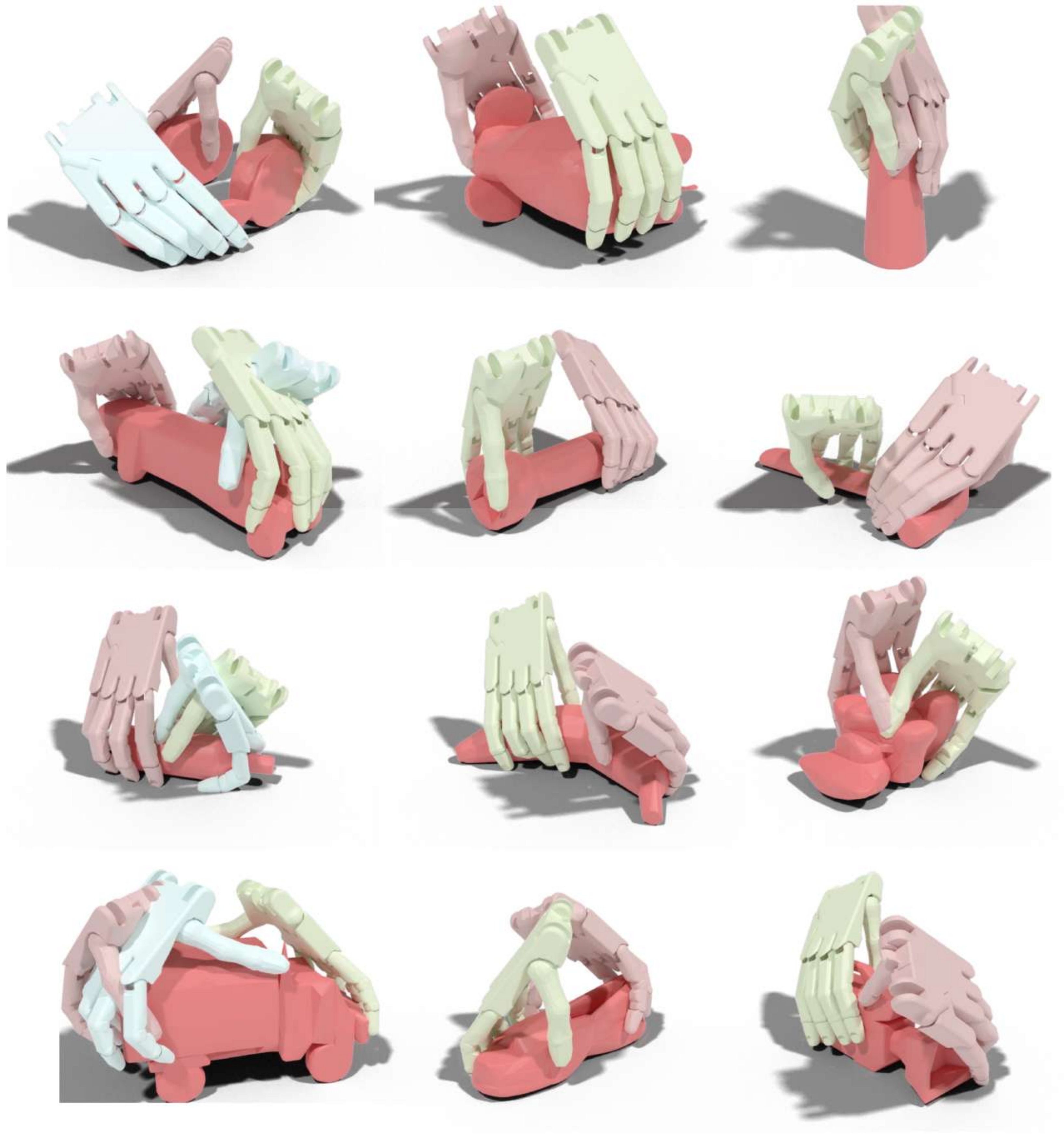


Imagination

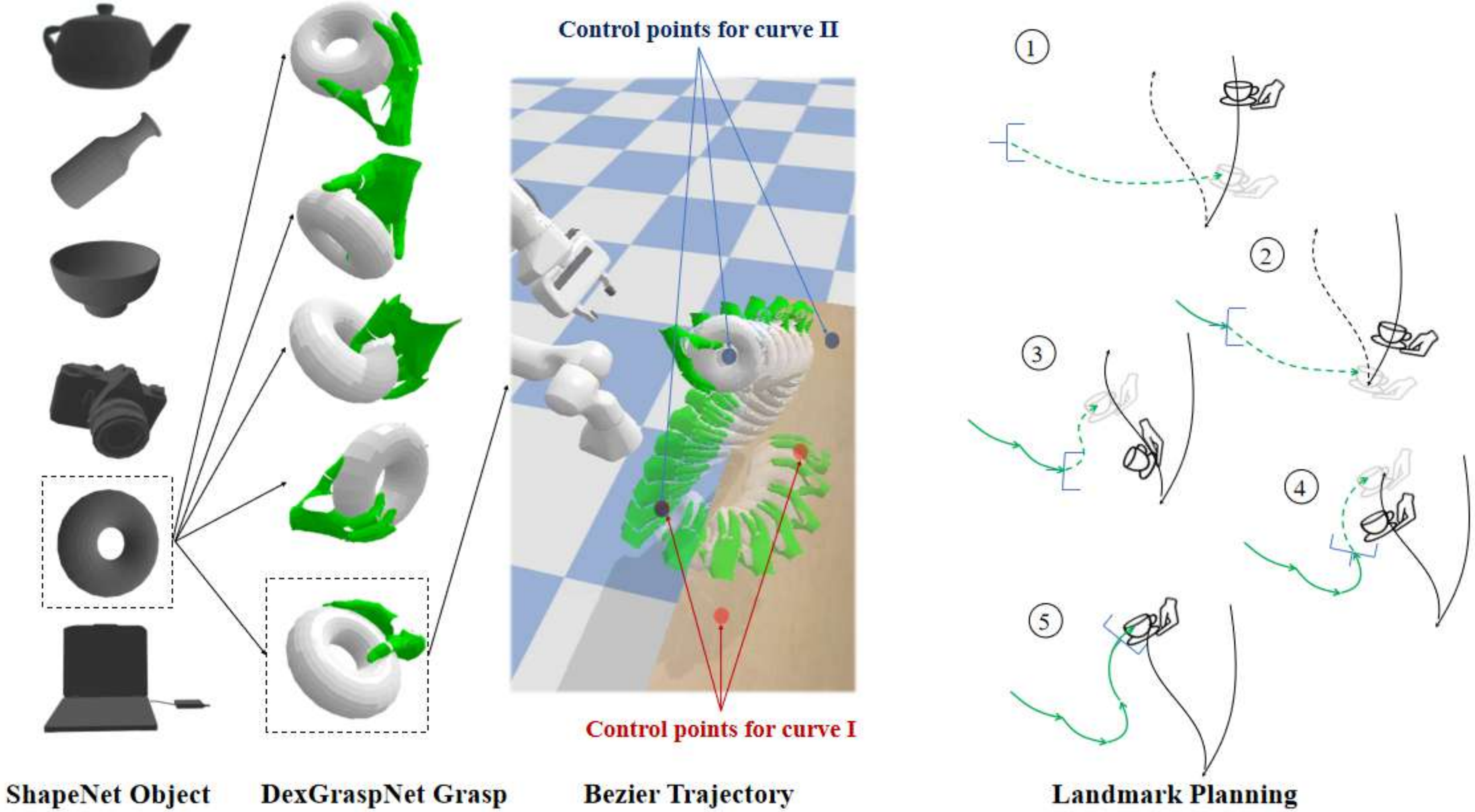
Reasoning

Interaction

Interaction: Grounded on the Geometric and Physical Understanding of the Dynamic World



UniDexGrasp



GenH2R

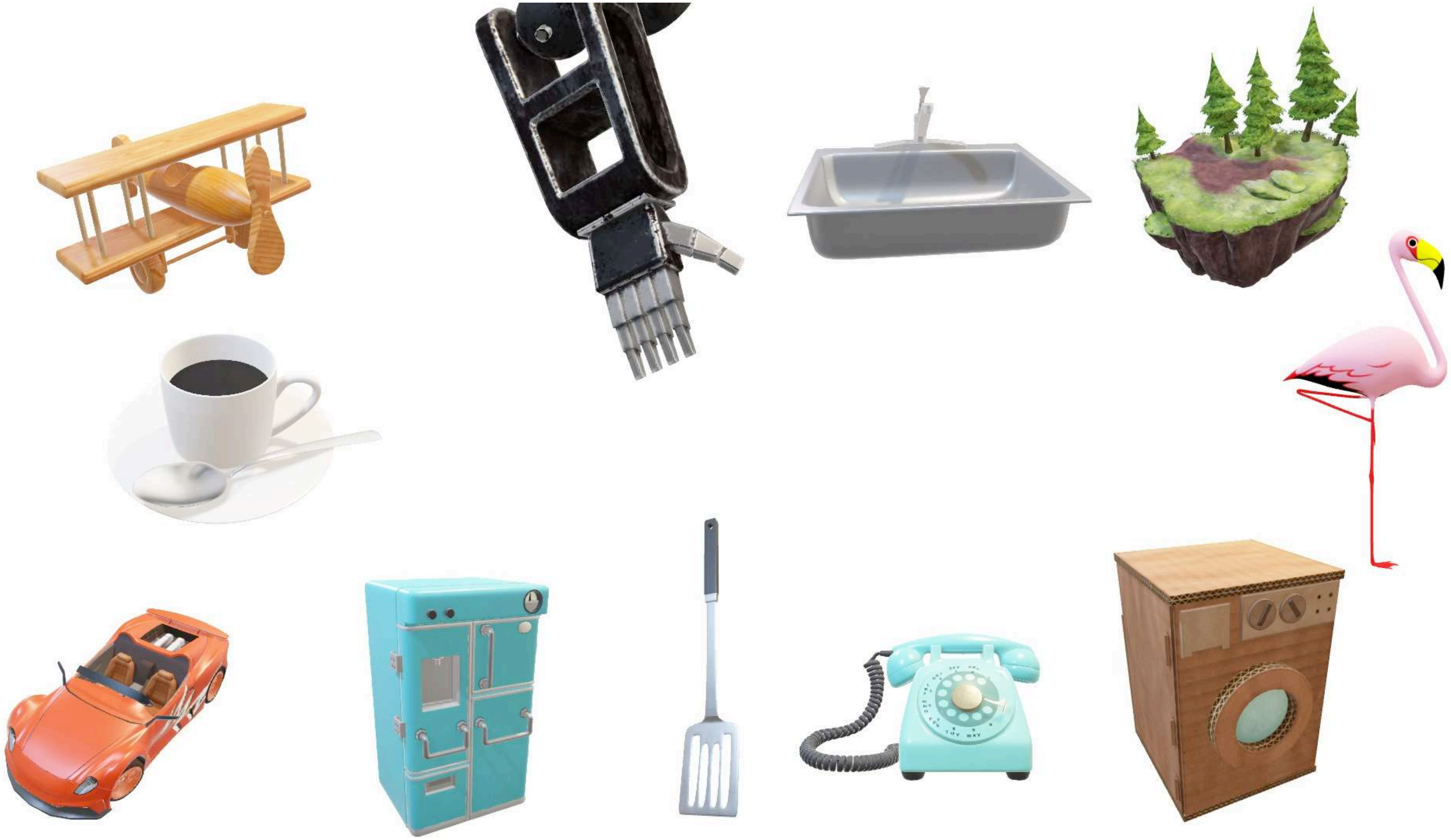
Xu et al., 2021. "UniDexGrasp: Universal Robotic Dexterous Grasping via Learning Diverse Proposal Generation and Goal-Conditioned Policy".
Wang et al., 2022. "GenH2R: Learning Generalizable Human-to-Robot Handover via Scalable Simulation, Demonstration, and Imitation".

Goal: Inject Actionable Information into MLLMs

Reasoning

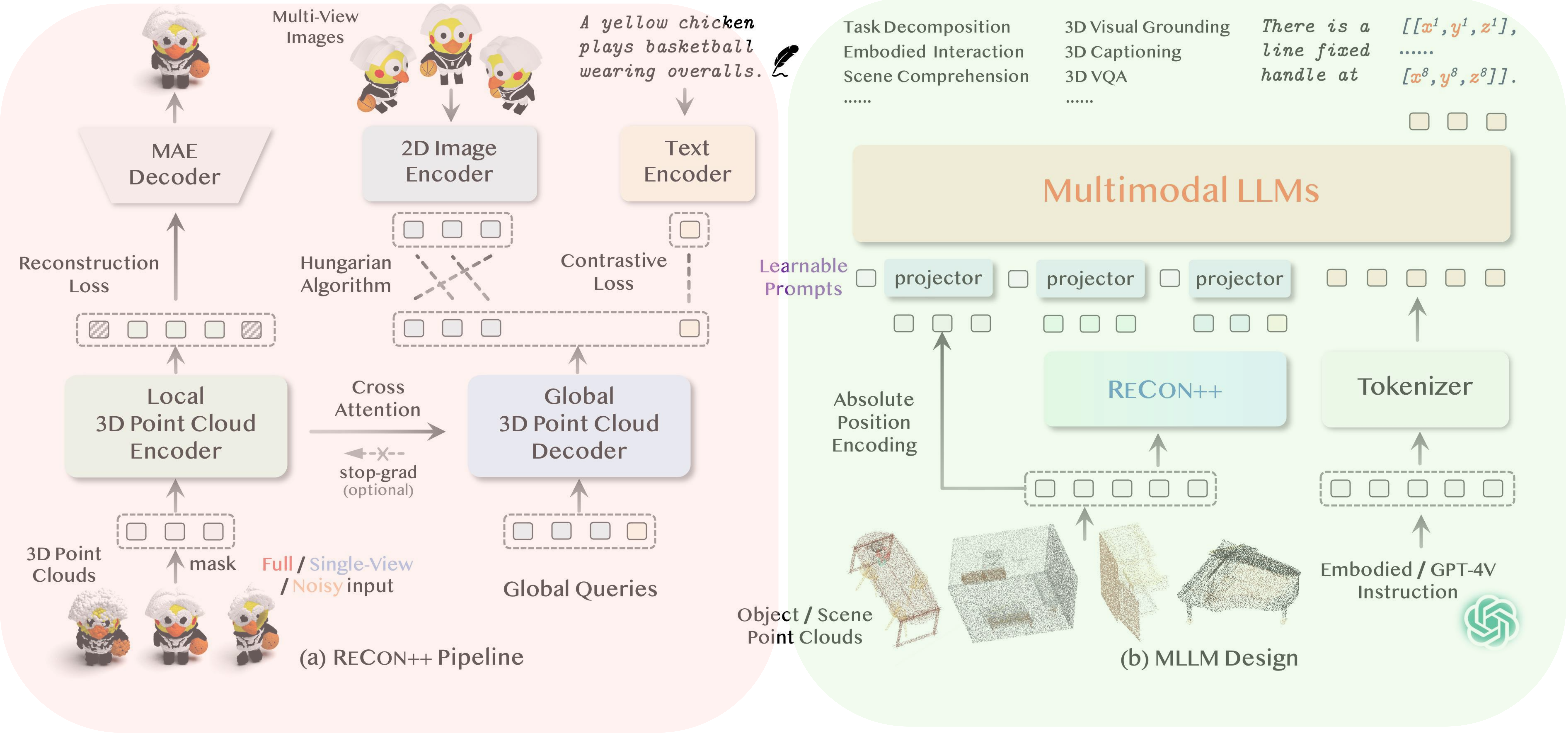
Interaction

ShapeLLM



ShapeLLM: Universal 3D Object Understanding for Embodied Interaction
Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, He Wang,
Li Yi†, Kaisheng Ma†. In submission.

ShapeLLM

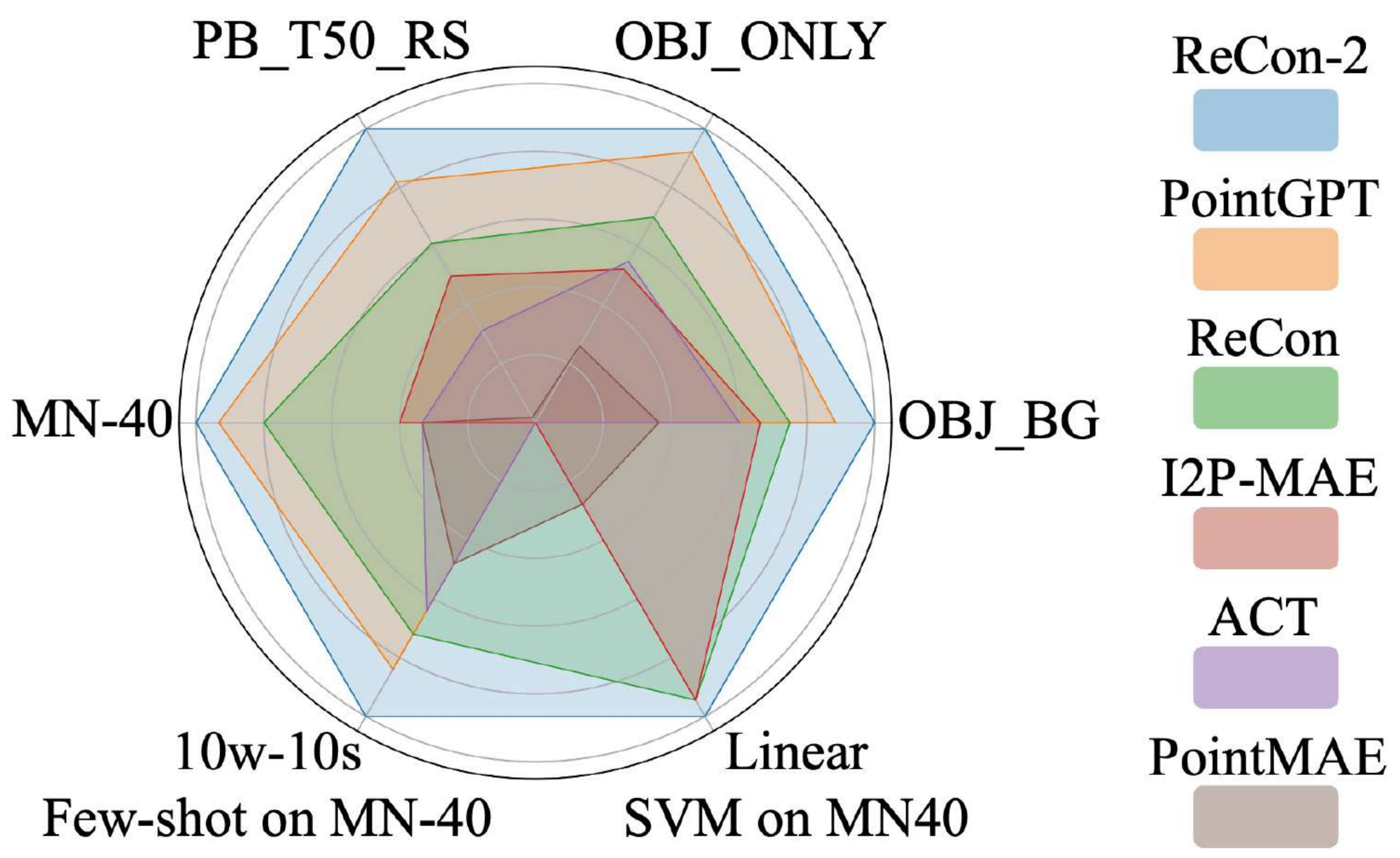


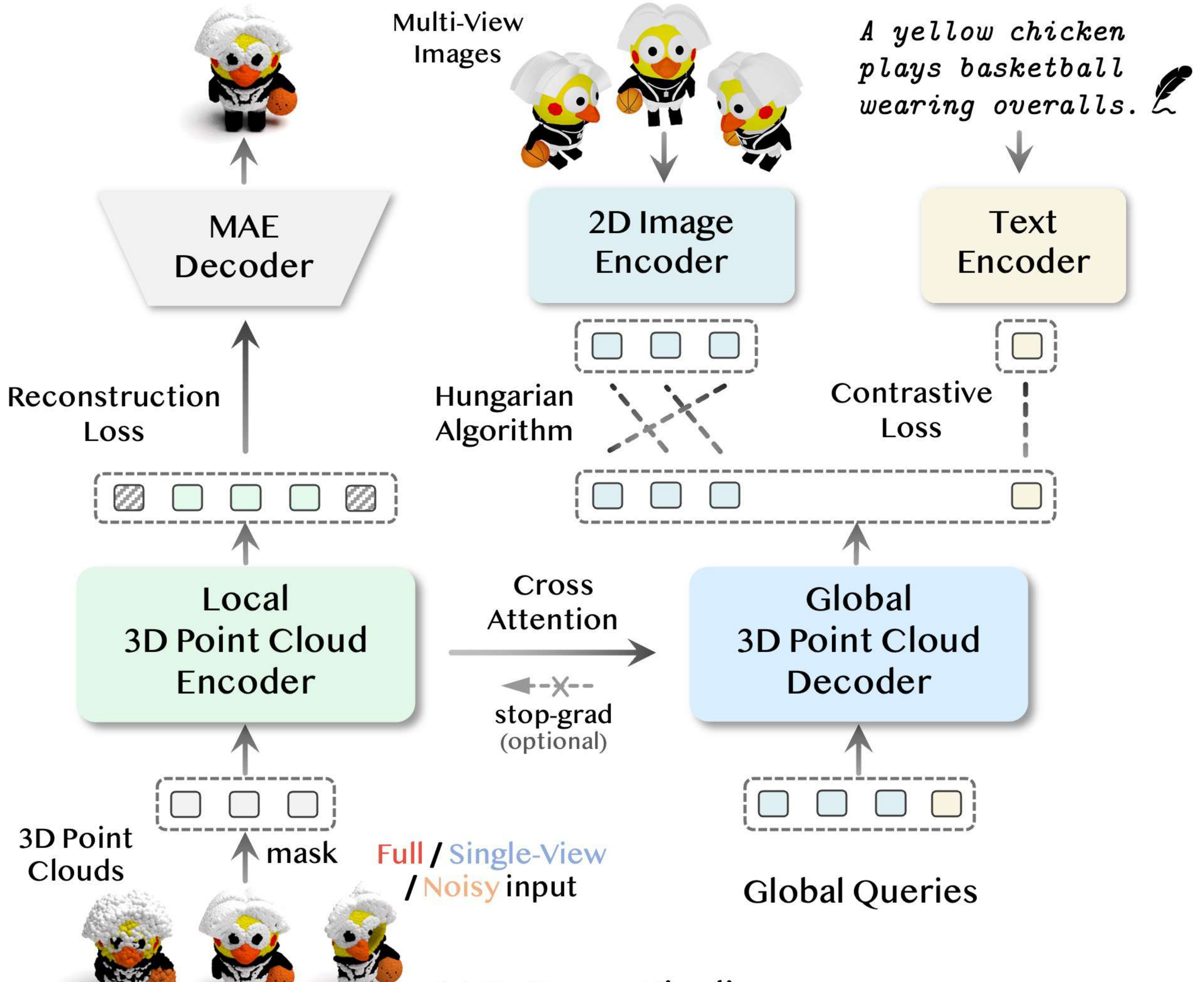
MAE Decoder
Multi-View Images
2D Image Encoder
A yellow chicken plays basketball wearing overalls.
Text Encoder
Reconstruction Loss
Hungarian Algorithm
Contrastive Loss
Local 3D Point Cloud Encoder
Cross Attention
stop-grad (optional)
Global 3D Point Cloud Decoder
3D Point Clouds
mask
Full / Single-View / Noisy input
Global Queries
(a) RECON++ Pipeline
Task Decomposition
Embodied Interaction
Scene Comprehension
......
3D Visual Grounding
3D Captioning
3D VQA
......
There is a line fixed handle at
$[[x^1, y^1, z^1],$ ...... $[x^8, y^8, z^8]].$
Multimodal LLMs
Learnable Prompts
projector
projector
projector
Absolute Position Encoding
RECON++
Tokenizer
Object / Scene Point Clouds
Embodied / GPT-4V Instruction
(b) MLLM Design

ShapeLLM – Scaling up 3D Representation Learning

Point Cloud Encoder

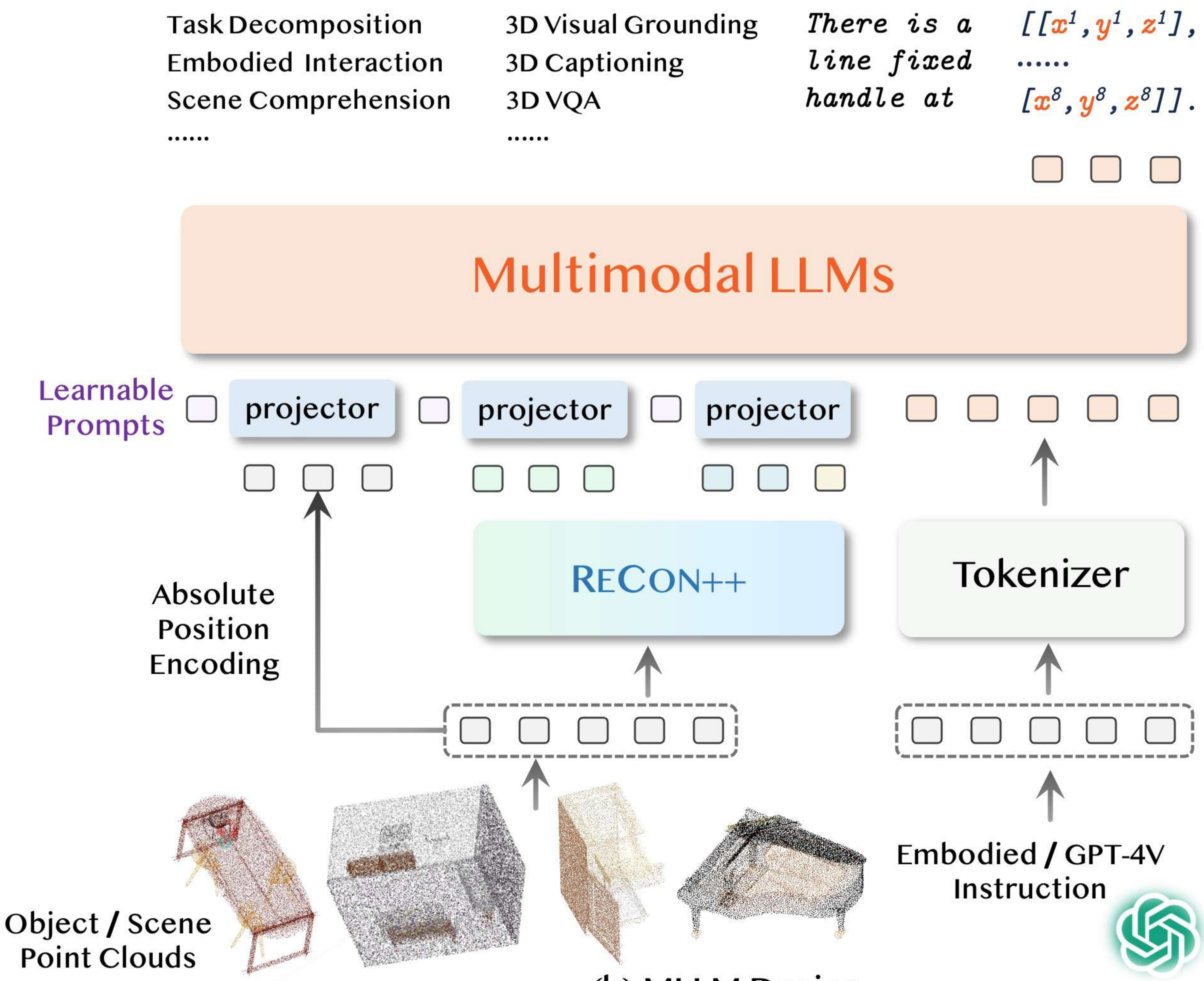
1. Scaling up to Objaverse
2. Distillation from multi-view images
3. Training with single-view & noisy points





(a) RECON++ Pipeline

ShapeLLM – Bridging 3D and LLMs



(b) MLLM Design

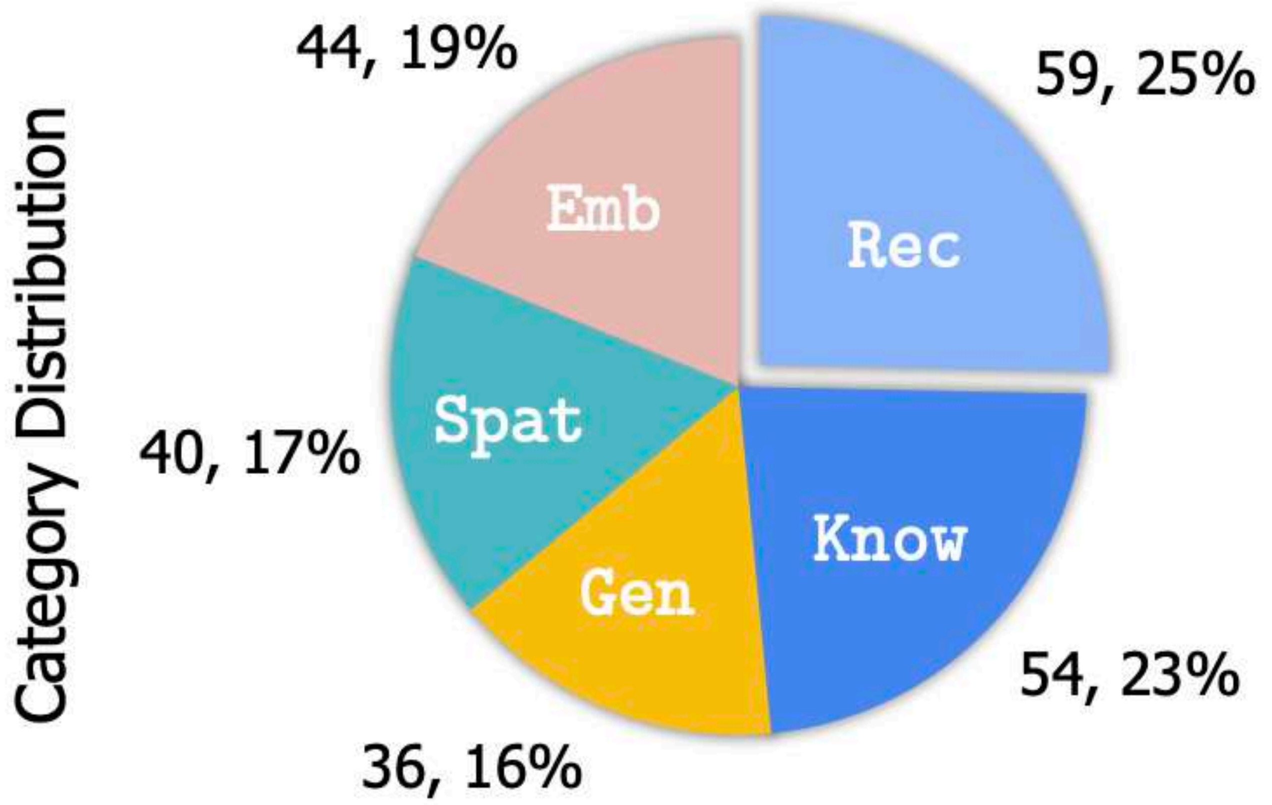
Feature Bridge

1. Multi-level feature projection
   - Global contrastive feature
   - Local geometry reconstruction
   - Absolute position encoding
2. Visual prompt tuning

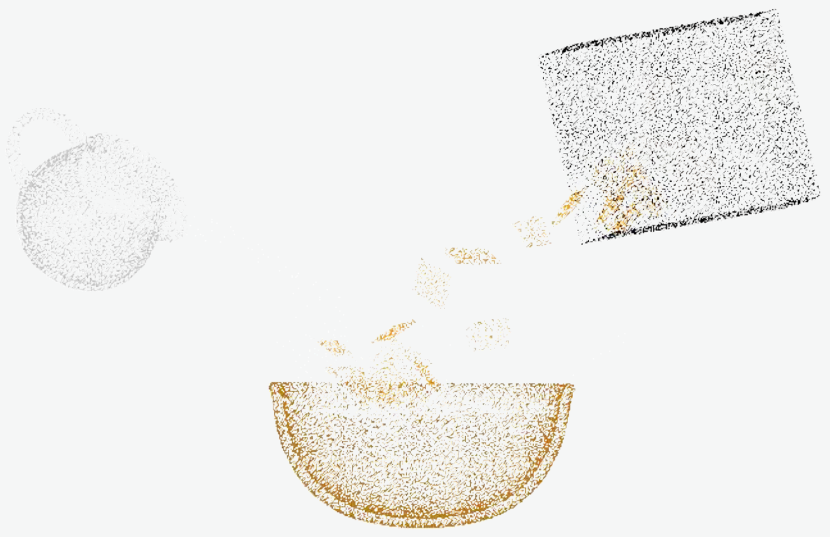
ShapeLLM – Solving the 3D Data-Dessert Issue

Five Levels of 3D Understanding

- General Visual Recognition
- Knowledge
- Language Generation
- Spatial Recognition
- Embodied Interaction



[Point Clouds]



point cloud input

[Question[1] Type: General Visual Recognition]
What subparts are there in the scene?
[Answer[1]]
There is a bag of cookies, a mug of milk and a China bowl.

[Question[2] Type: Embodied Interaction]
As an AI robot, please give me steps to mix the milk and cookies in a bowl.
[Answer[2]]
Step 1: Pour the milk into the bowl.
Step 2: Put the cookies into the bowl.
Step 3: Stir with a spoon.

[Question[3] Type: Knowledge]
Describe the physical properties of the milk.
[Answer[3]]
The milk is a kind of liquid with a white color, whose density and boiling point is higher than water while the freezing point is lower than water, has a mild, slightly sweet odor and taste.

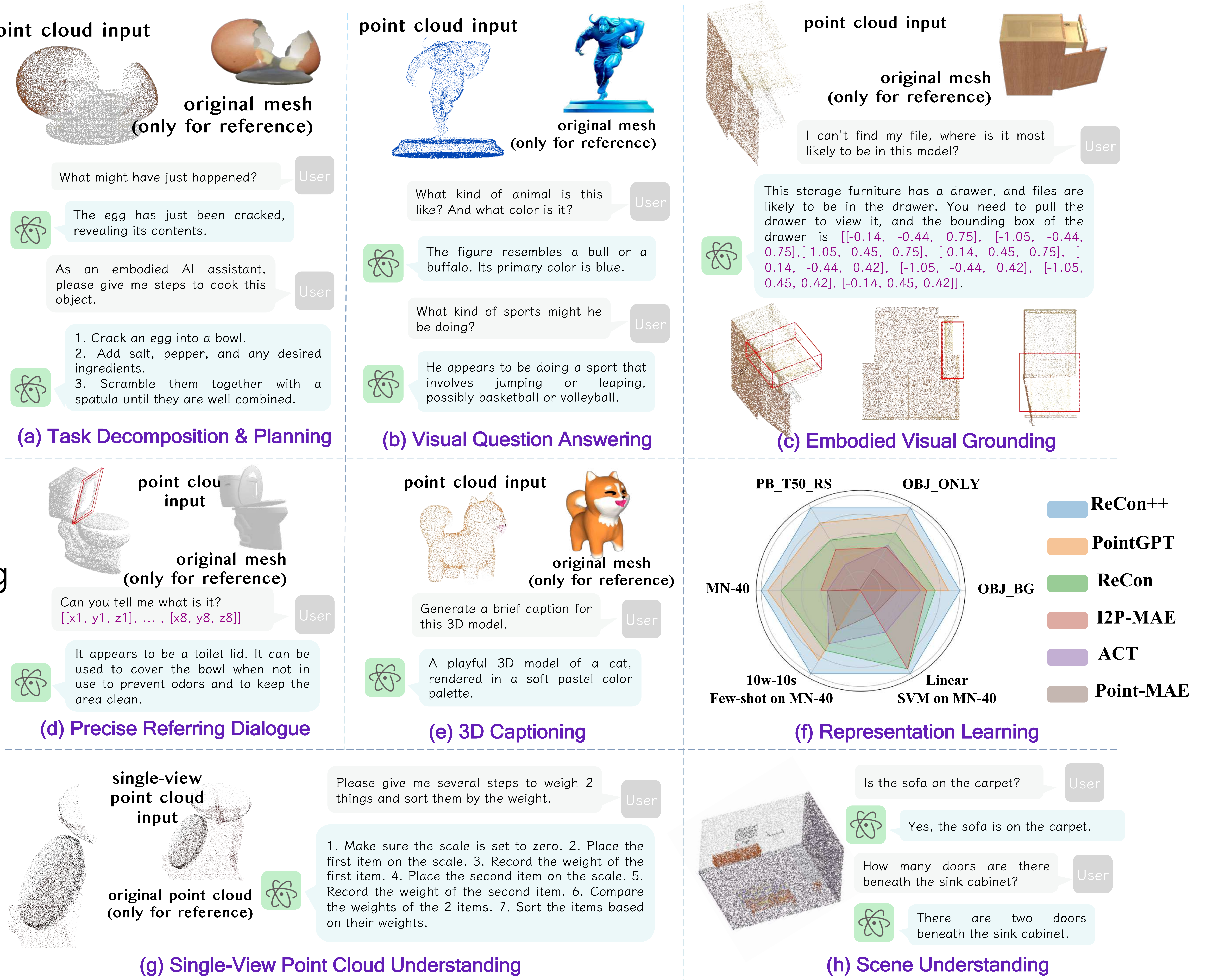
original mesh
(only for reference)

3D MM-Vet: 3D Multimodal Comprehension Evaluation Benchmark

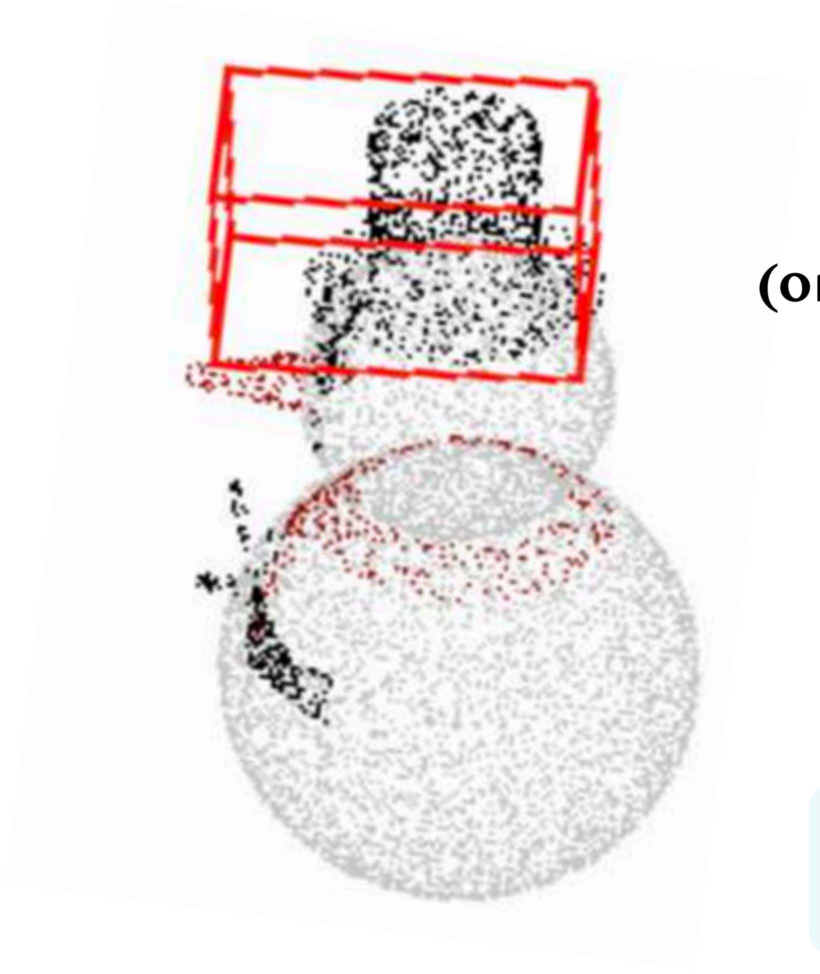
Open-World Embodied Reasoning Examples

3D Multimodal Generalist

1. Interaction-oriented open-world embodied reasoning
2. Physically grounded 3D object or part understanding
3. Generalizable 3D representation learning and understanding



Open-World Embodied Reasoning Examples

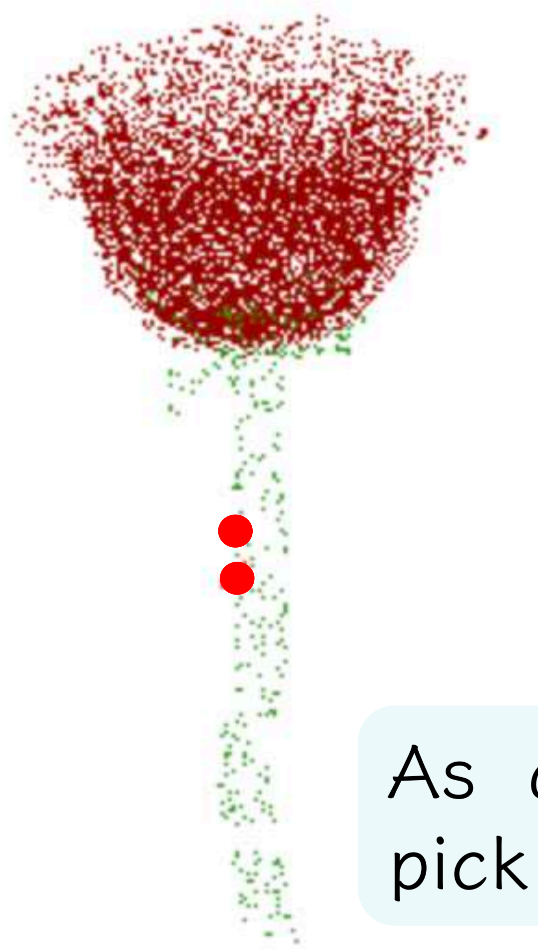


original mesh
(only for reference)

Where is its hat?

User

Its hat is located at [[-0.4, -0.45, 0.86], [-0.4, 0.28, 0.86], [0.08, 0.28, 0.86], [0.08, -0.45, 0.86], [-0.4, -0.45, 0.44], [-0.4, 0.28, 0.44], [0.08, 0.28, 0.44], [0.08, -0.45, 0.44]].



original mesh
(only for reference)

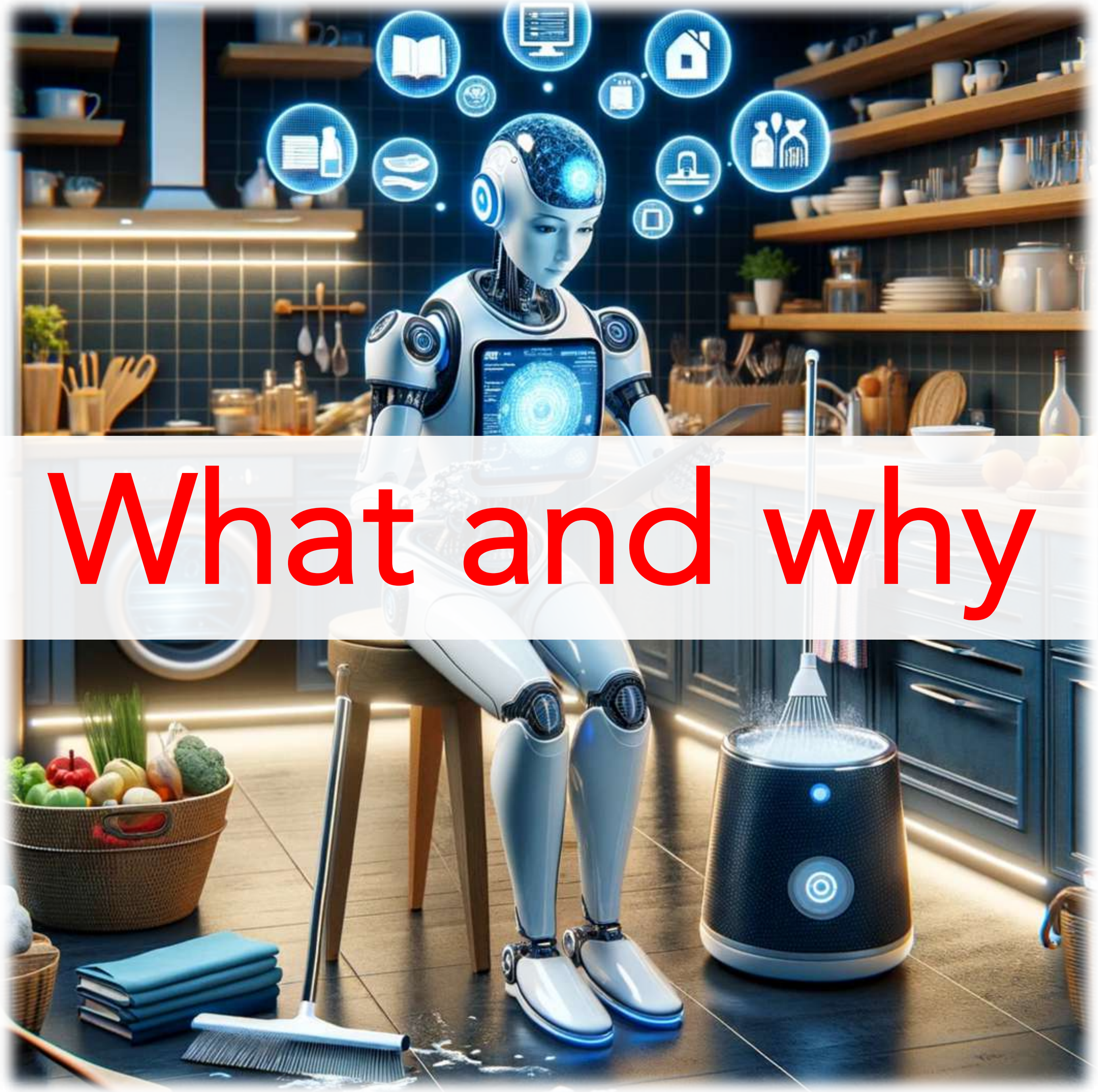
As a robot, how can you pick up this flower?

User

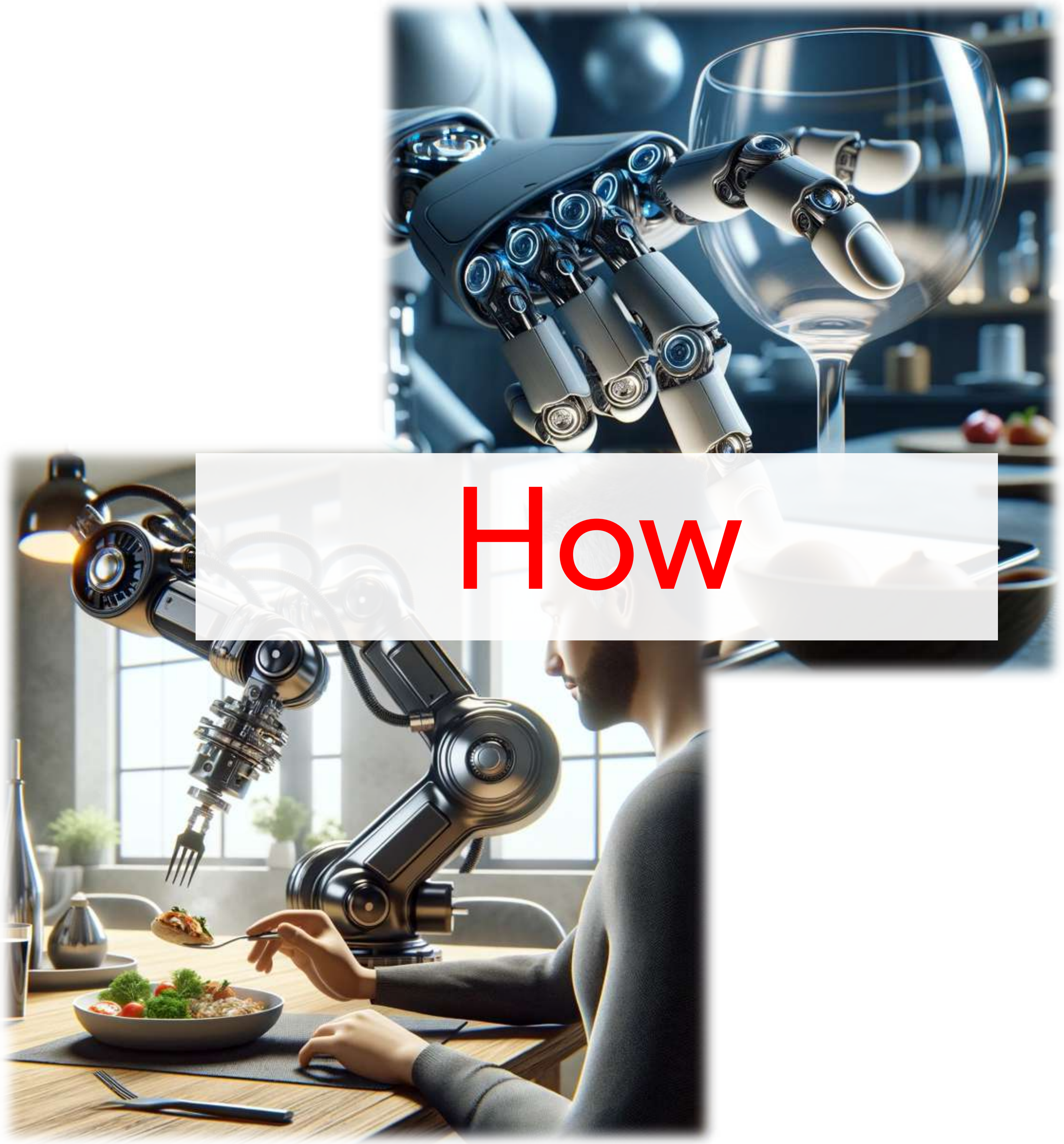
To pick up the flower, you would need to approach it from the front and gently grasp the stem at the point closest to the robot, such as [-0.01, 0.02, -0.49] or [-0.01, 0.02, -0.46].

Takeaway

What if



What and why



How

Imagination

Reasoning

Interaction

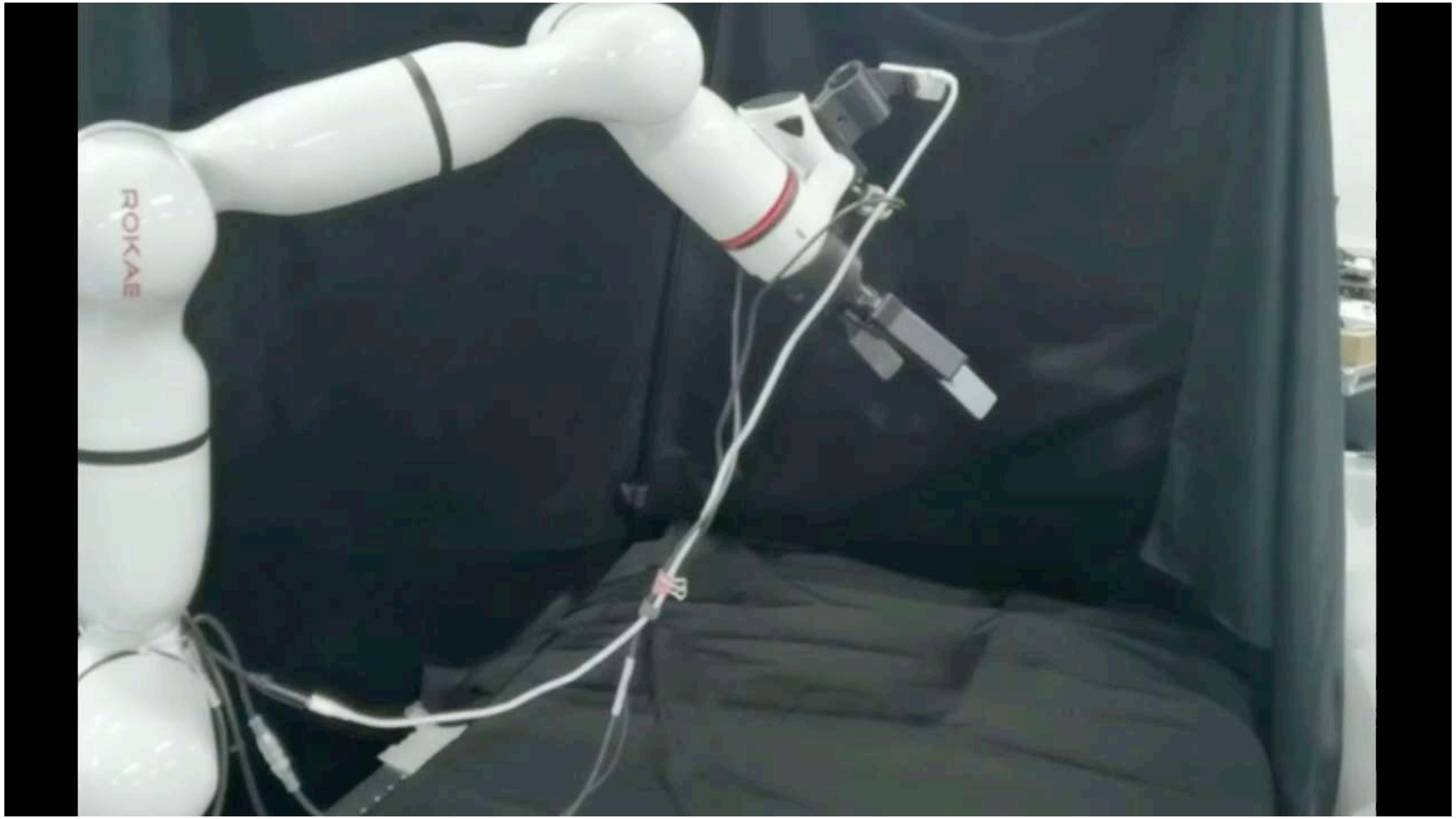

Human-Centered Robot Simulator
Human-Centered EAI
Human Interaction Capturing
Interactable Asset Creation
Open-World Perception
Police Car
Dragon Chair
Scissor
Data Driven
Asset Support
Visual Perception
Human Interaction Synthesis
Human-Centered Robot Simulator
Human-Centered Robotics
Human Simulation
Simulation Support
Collaborative Transport
Human-to-Robot Handover

GenH2R: Learning Generalizable Human-to-Robot Handover via Scalable Simulation, Demonstration, and Imitation

Zifan Wang*, Junyu Chen*, Ziqing Chen, Pengwei Xie, Rui Chen, Li Yi

CVPR 2024

Task Goal

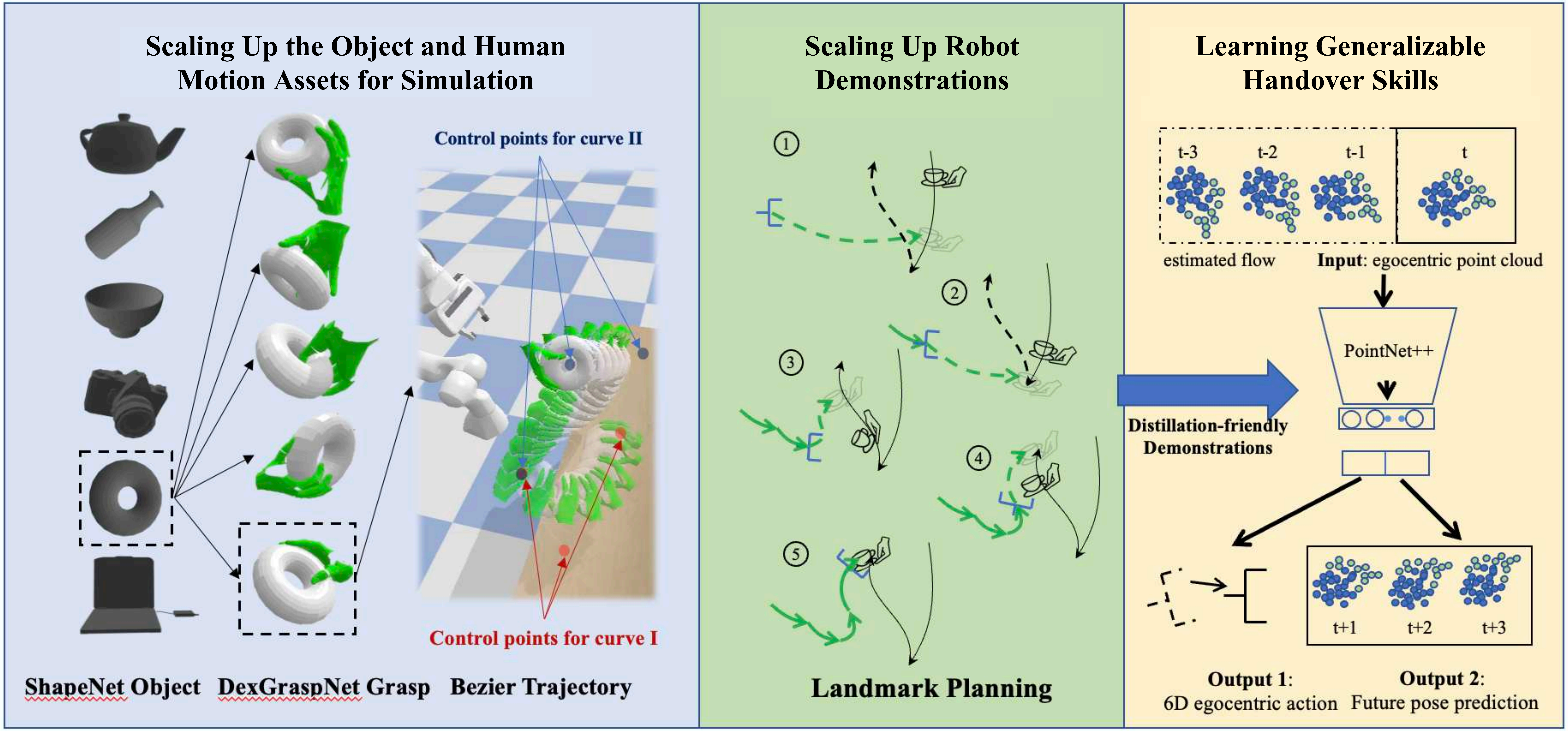


- Arbitrary objects
- Arbitrary human grasp
- Arbitrary human motion

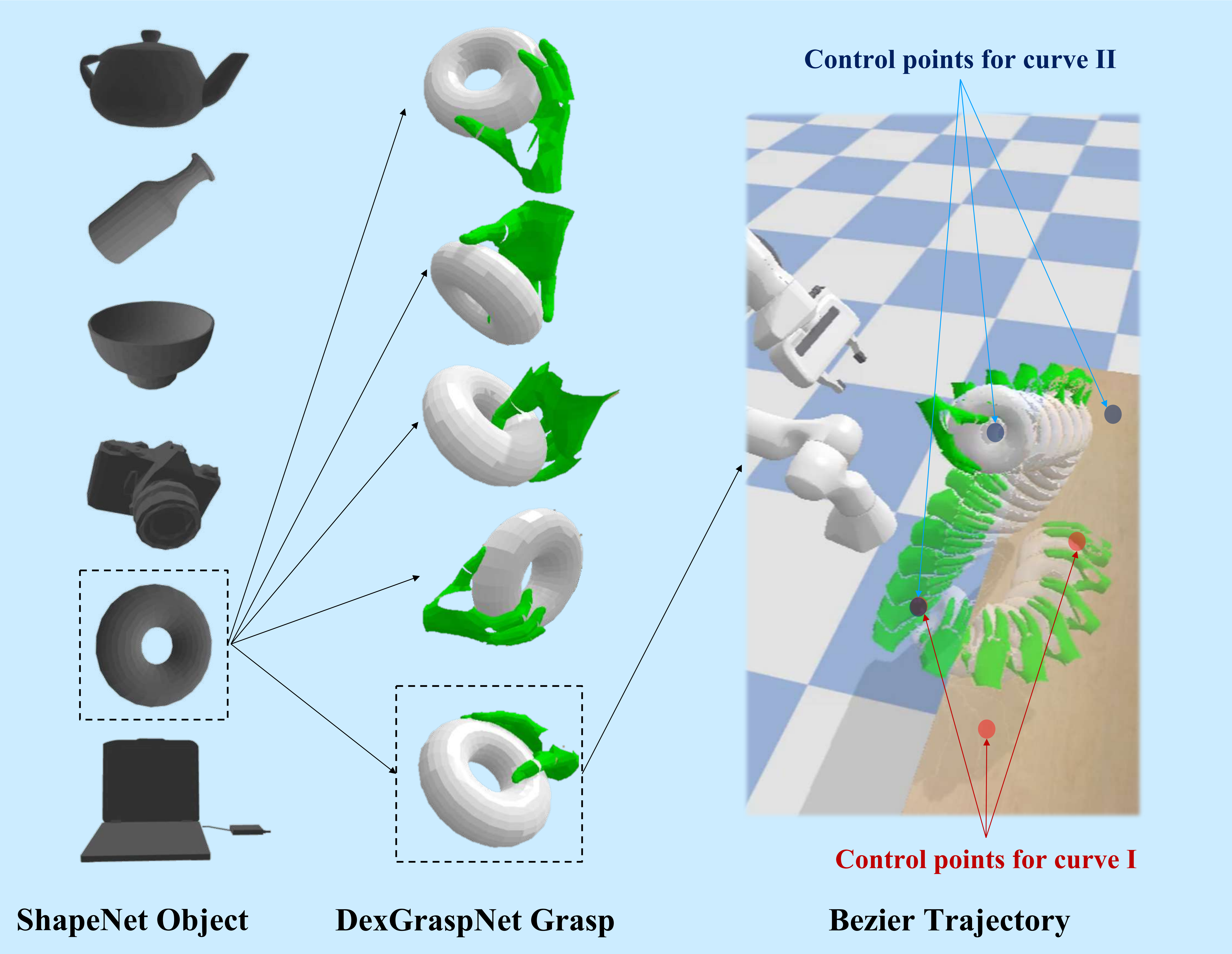
Challenges

- How to scale up the object and human motion assets for simulation?
- How to scale up robot demonstrations?
- How to learn generalizable handover skills?

Method Overview

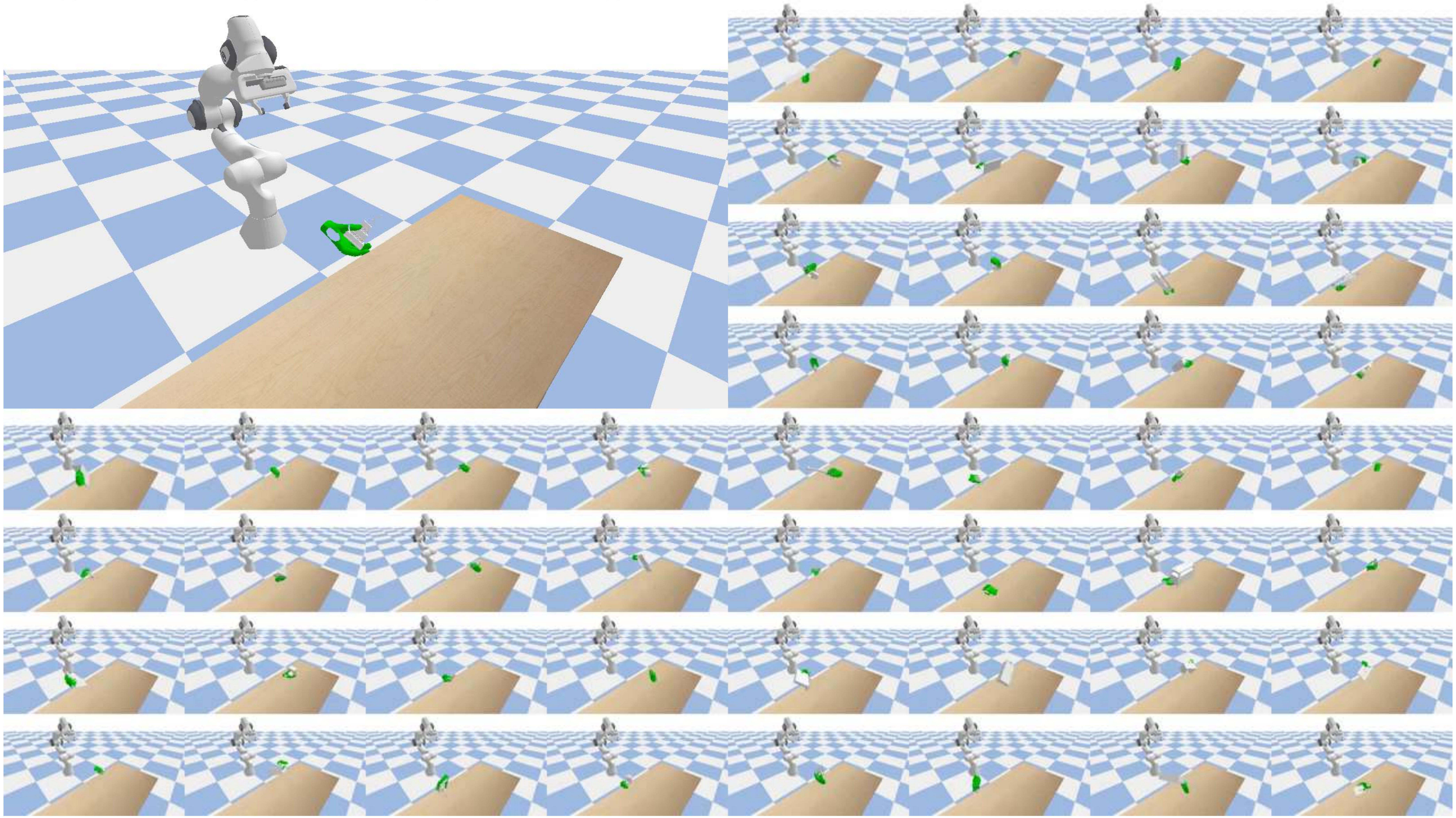


Human Simulation: Generating Hand-Object Trajectory

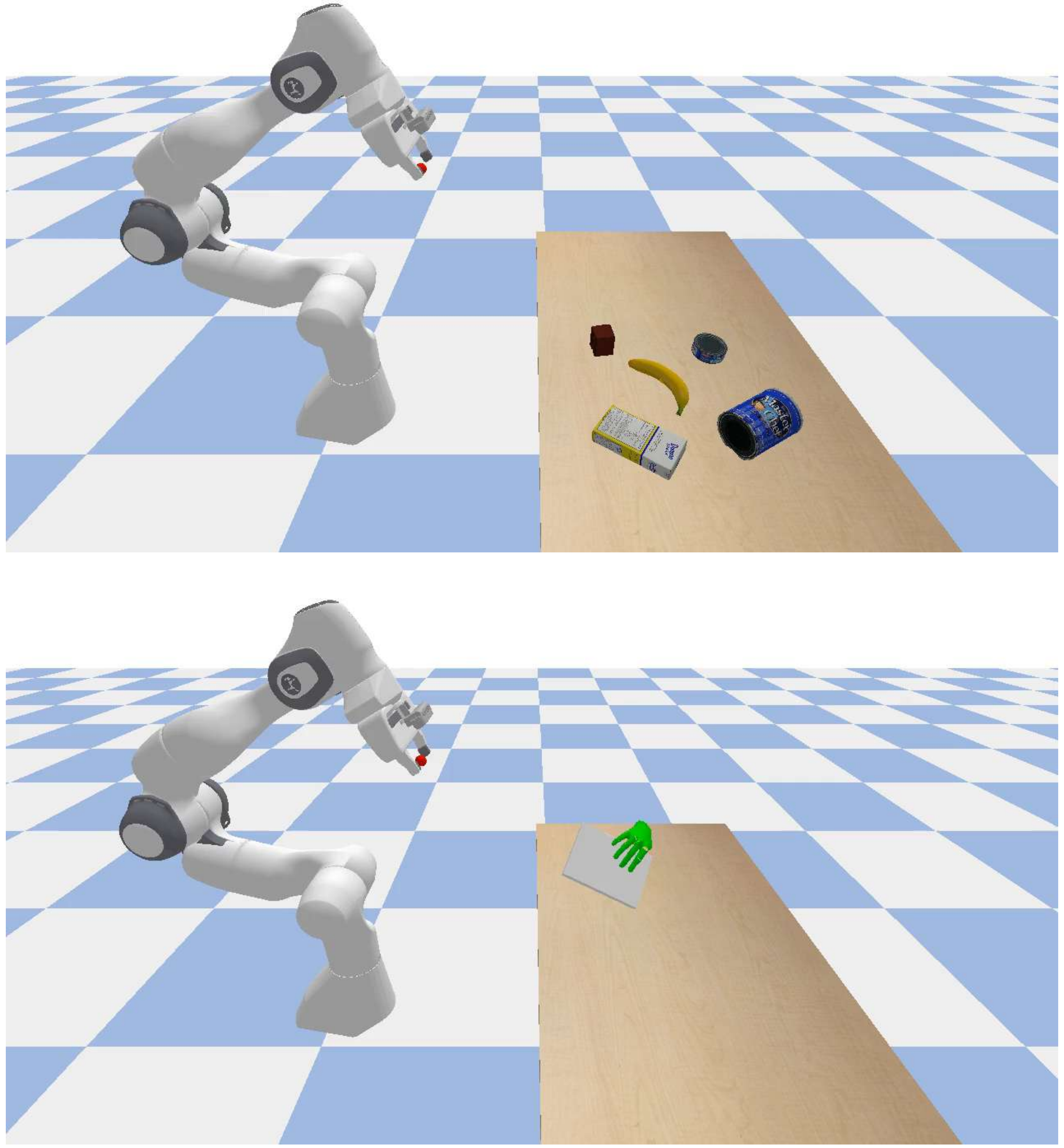


- 3,000+ objects from ShapeNet
- 250+ hand grasps for each object from DexGraspNet
- ~1,000,000 hand-object trajectories in total

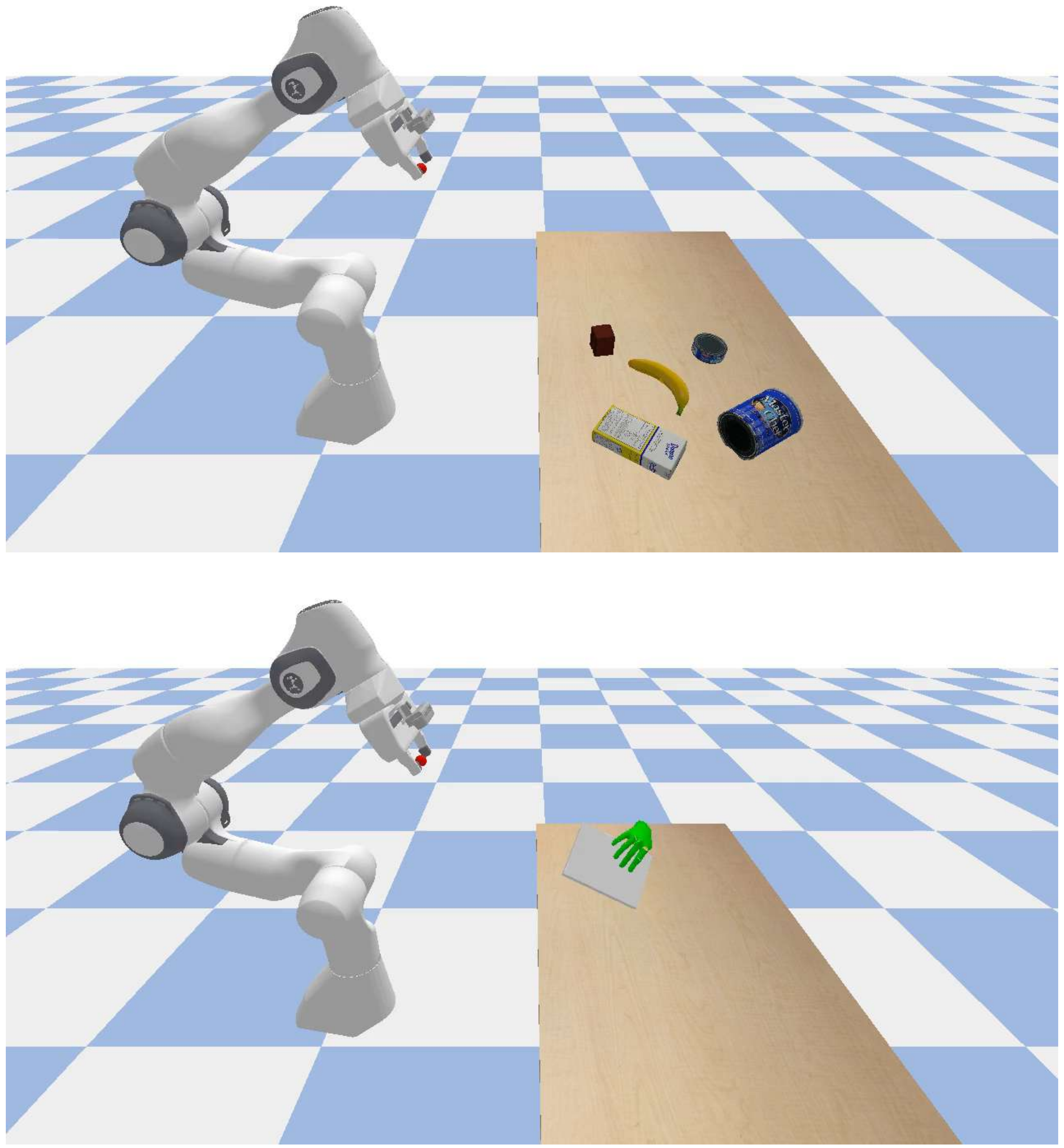
Human Simulation: Generating Hand-Object Trajectory



Robot Demonstration: Automatic Grasp and Motion Planning

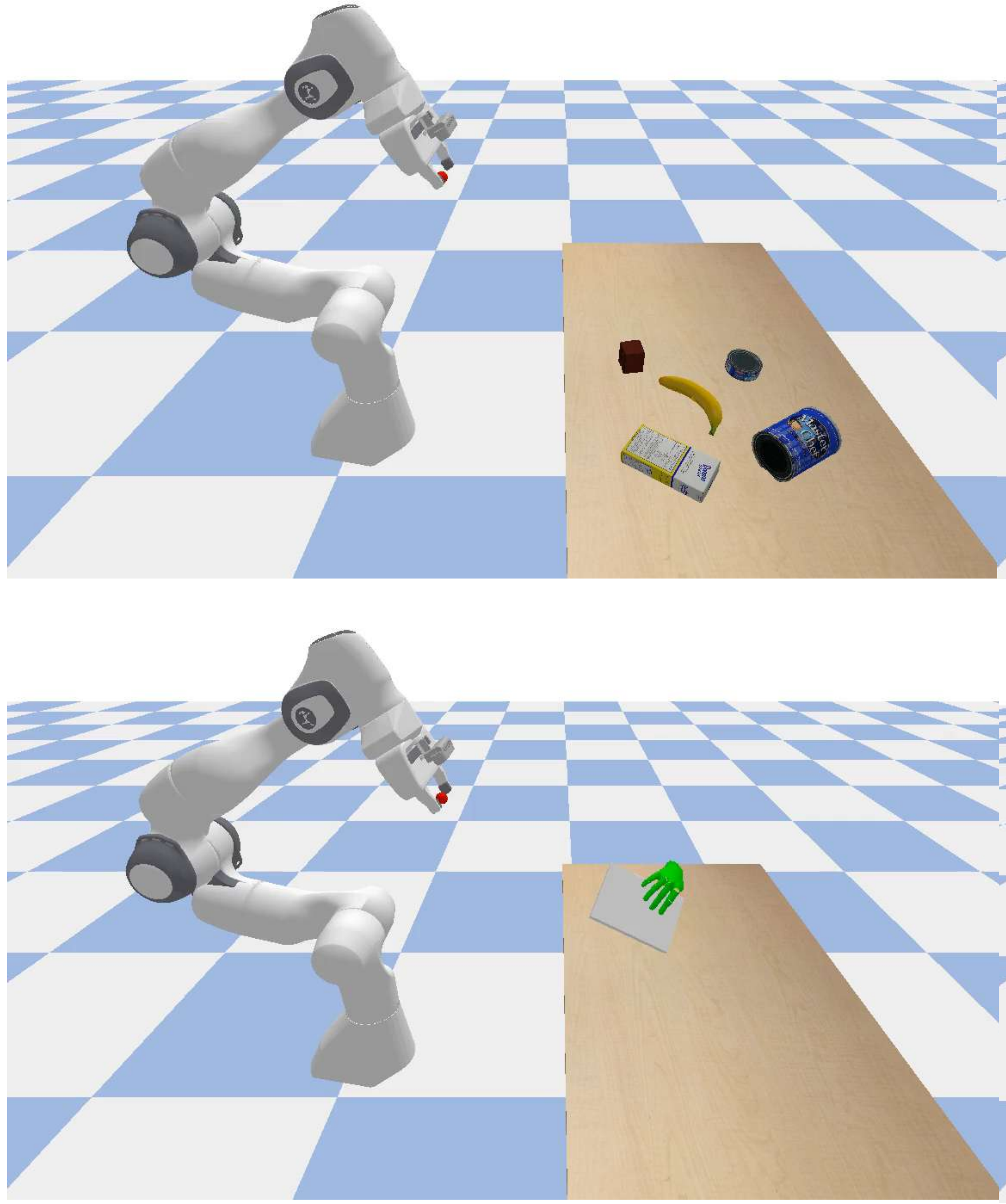


foresighted planner—planning once based on the privileged destination

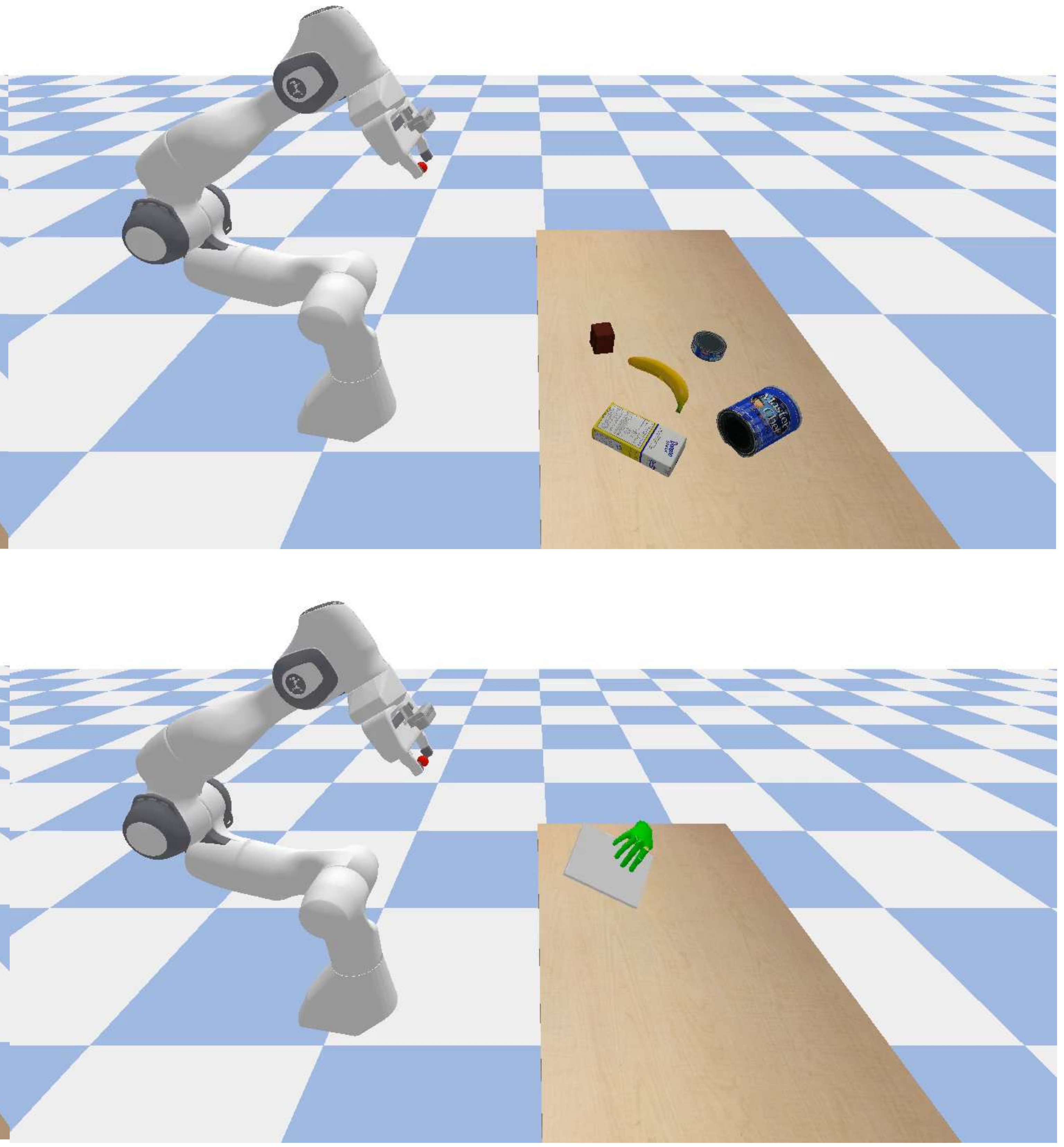


shortsighted planner—planning at each time step using privileged hand and object states

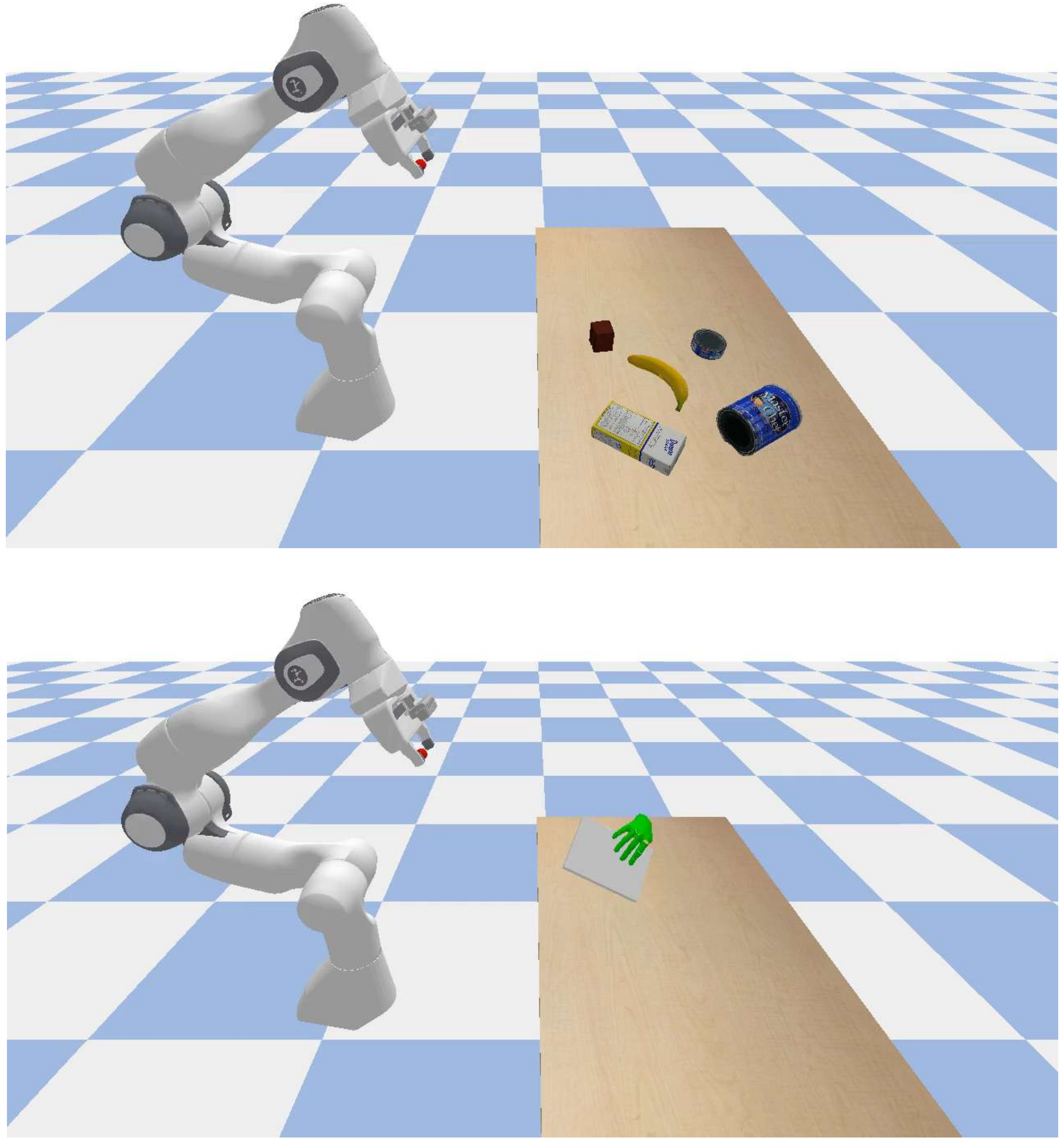
Robot Demonstration: Automatic Grasp and Motion Planning



foresighted planner—planning once based on the privileged destination

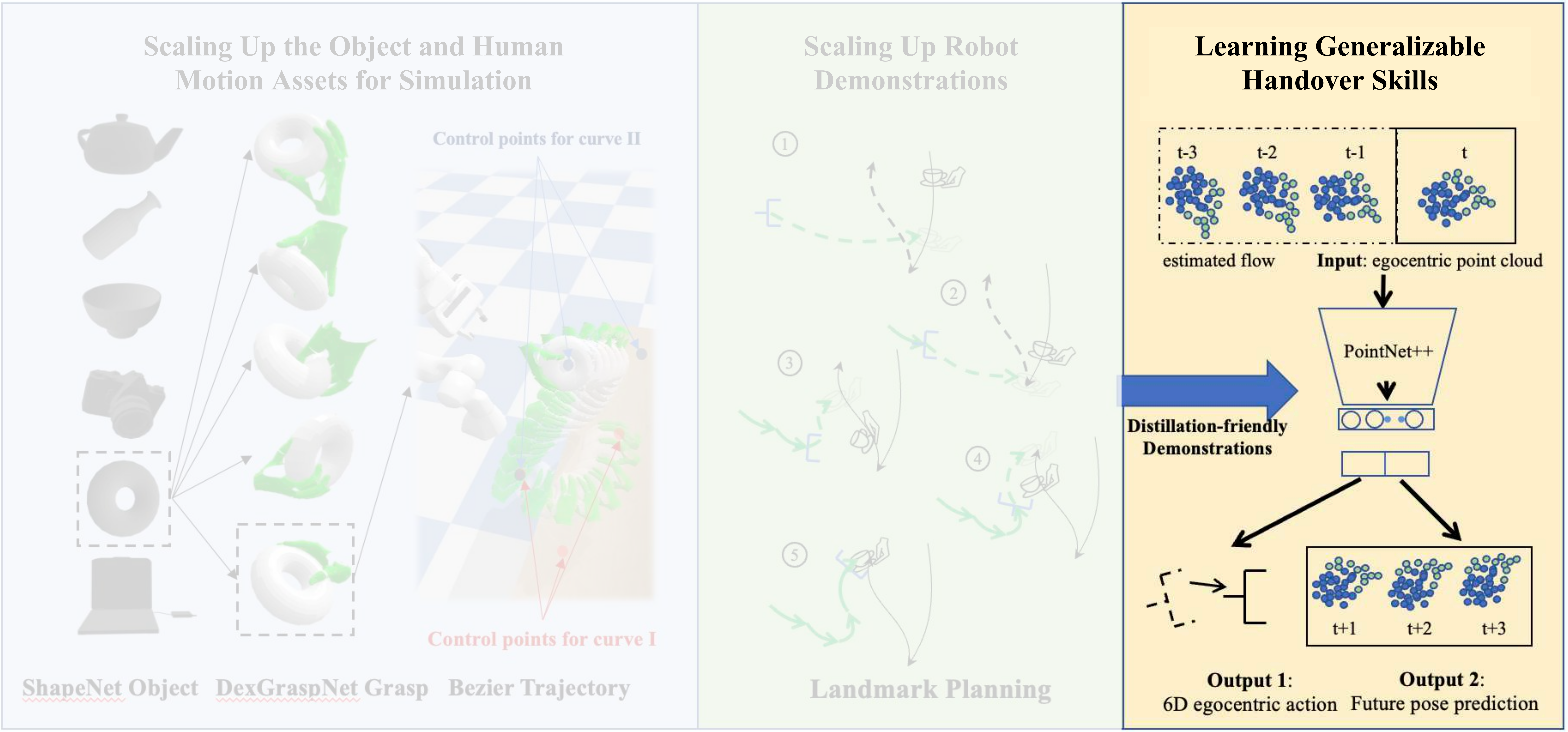


shortsighted planner—planning at each time step using privileged hand and object states



our planner—planning based upon adaptively selected future landmark

Demonstration Distillation: Forecast-Aided 4D Imitation Learning



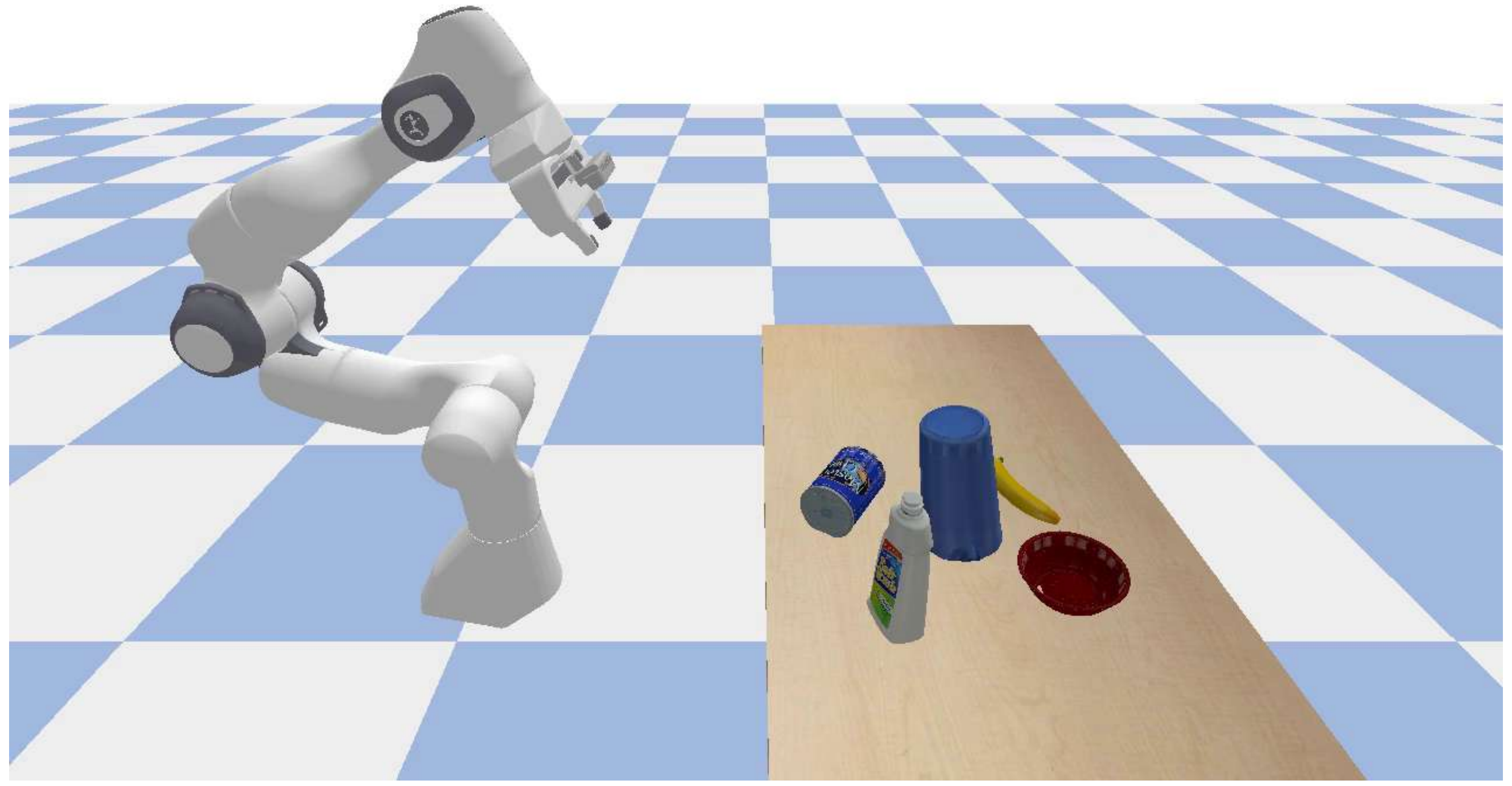
Quantitive Comparisons

		s0 (Sequential)			s0 (Simultaneous)			t0			t1		
		S	T	AS	S	T	AS	S	T	AS	S	T	AS
	OMG Planner [7] †	62.50	8.31	22.5	-	-	-	-	-	-	-	-	-
train on s0	GA-DDPG [8]	50.00	**7.14**	22.5	36.81	**4.66**	23.6	23.59	7.31	10.3	46.70	**5.50**	26.9
	Handover-Sim2real [3]	75.23	7.74	30.4	68.75	6.23	35.8	29.17	6.29	15.0	52.40	7.09	23.8
	Destination Planning	74.31	7.98	28.7	76.16	5.89	41.7	25.68	5.34	15.1	48.40	7.49	20.5
	Dense Planning	74.77	8.14	28.0	75.45	6.06	40.3	27.30	5.49	15.7	52.30	7.44	22.4
	Landmark Planning	77.78	8.15	29.0	79.17	6.06	42.0	29.63	5.22	17.7	54.20	7.41	23.3
train on t0	GA-DDPG [8]	54.86	7.29	24.1	50.69	5.86	27.8	24.05	4.70	15.3	25.50	5.86	14.1
	Handover-Sim2real [3]	65.97	7.18	29.5	62.50	6.04	33.5	33.71	5.91	18.4	47.10	6.35	24.1
	Destination Planning	0.93	11.76	0.1	6.48	11.22	0.9	5.96	7.57	2.5	1.60	11.38	0.2
	Dense Planning	80.79	8.59	27.4	**86.81**	6.41	44.1	39.04	5.81	21.6	64.20	7.24	28.4
	Landmark Planning	**87.27**	7.66	**35.8**	84.03	5.57	**48.0**	**40.43**	**4.85**	**25.4**	**62.40**	6.20	**32.8**

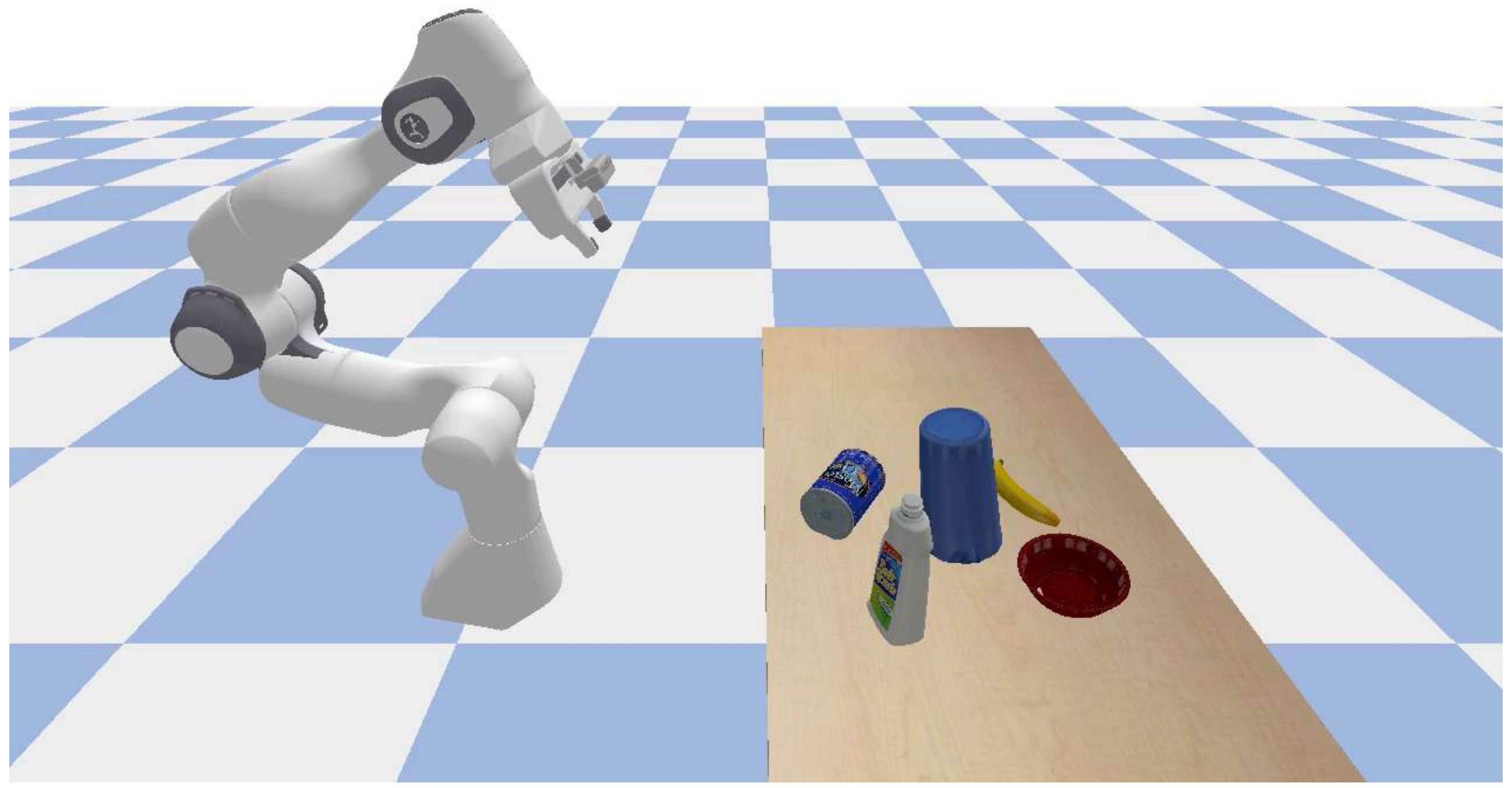
Qualitative Results: Simulation Experiments

(Baseline) HandoverSim2real | **Ours**

Sequential:
Robot arm starts moving after human hands stop

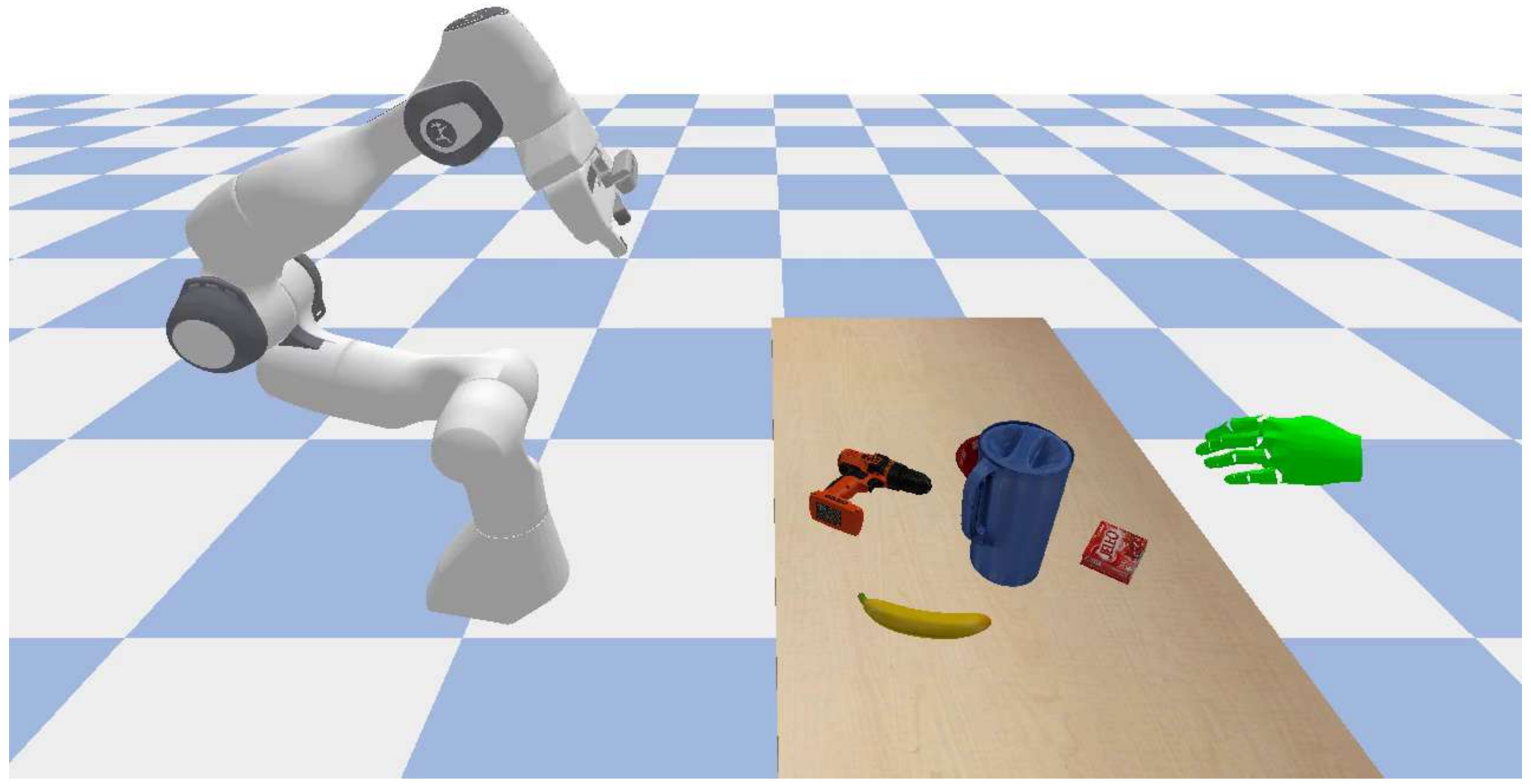


successful rate(%): 75.23

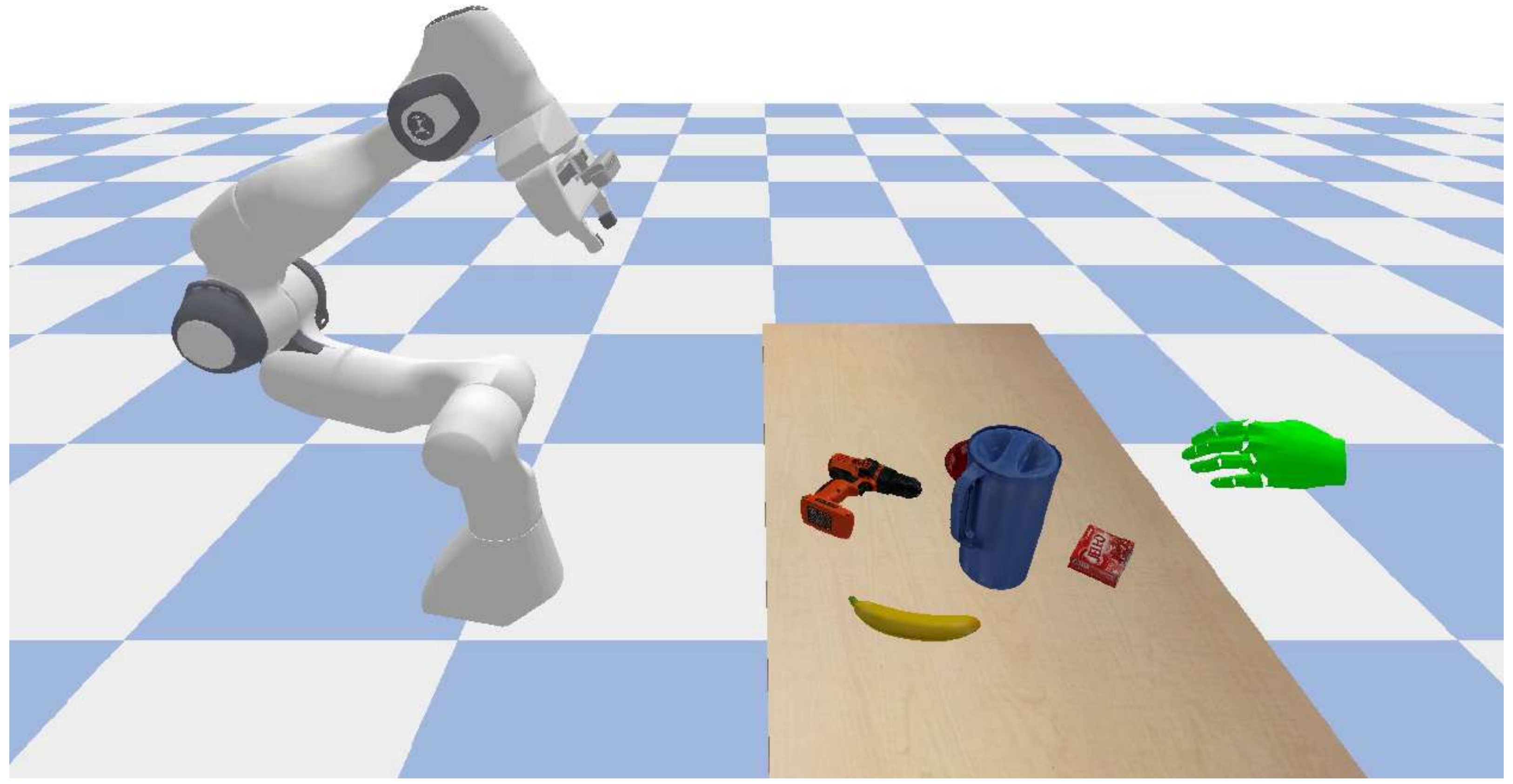


successful rate(%): 87.27

Simultaneous:
Robot arm moves together with human hands

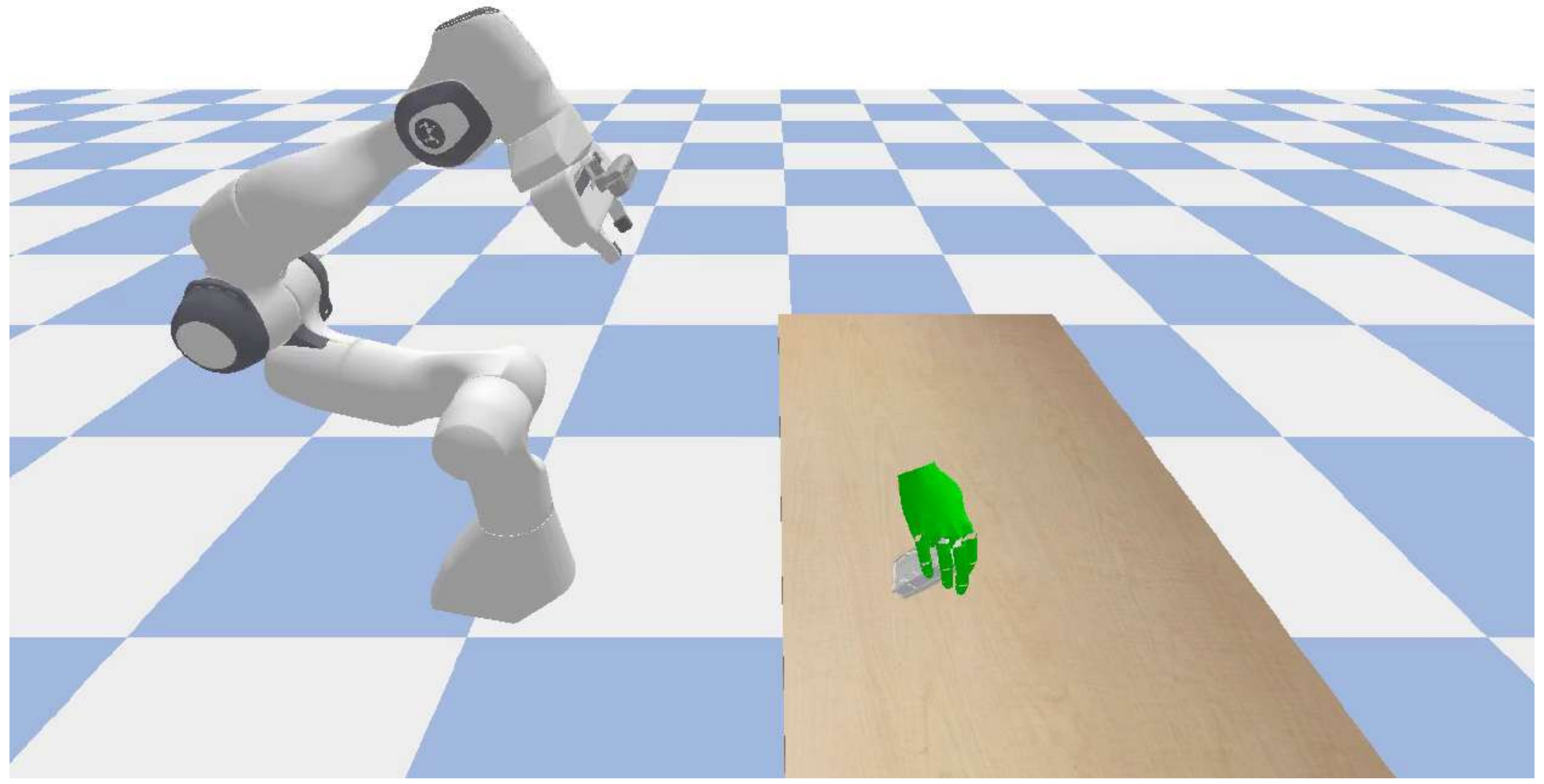
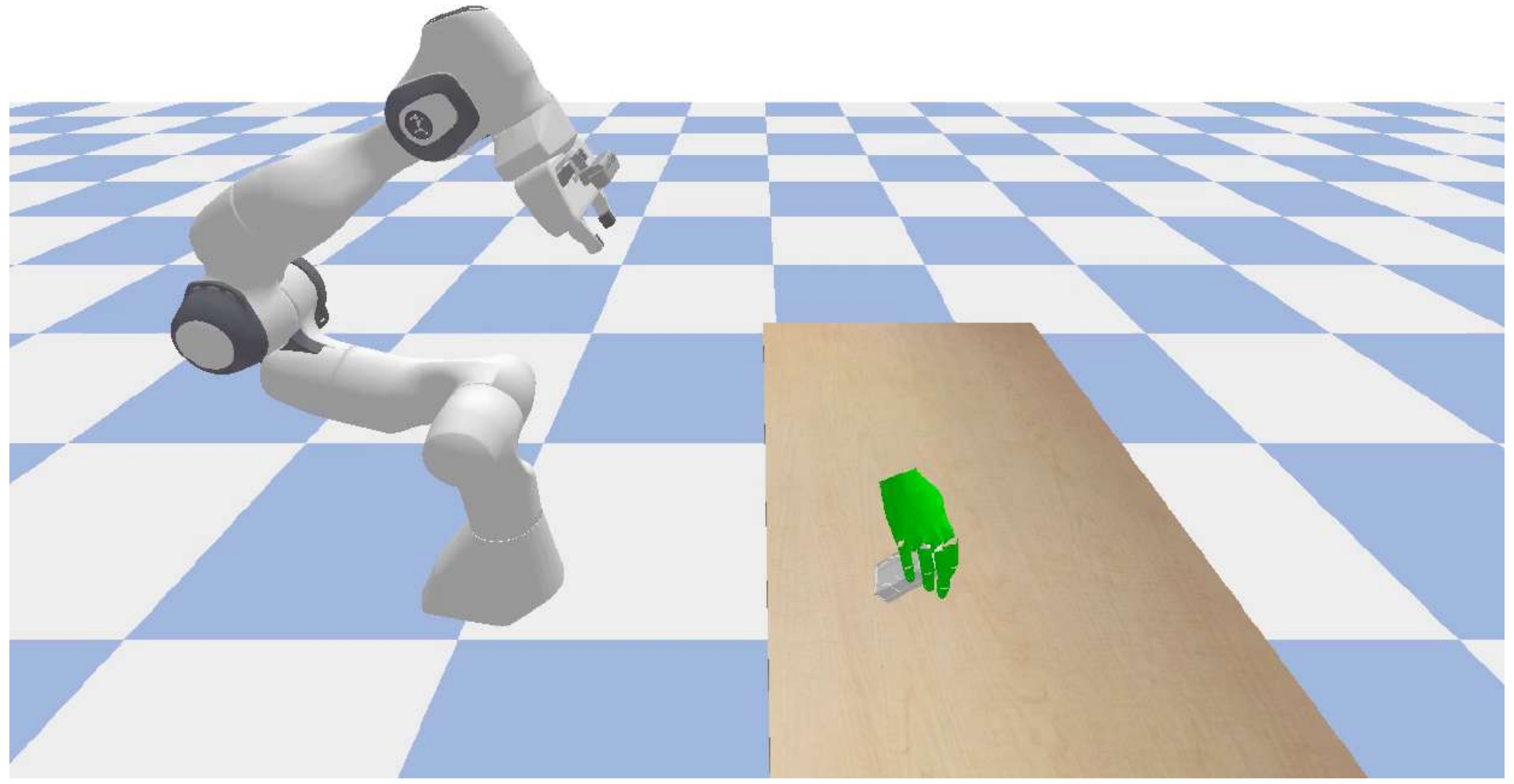
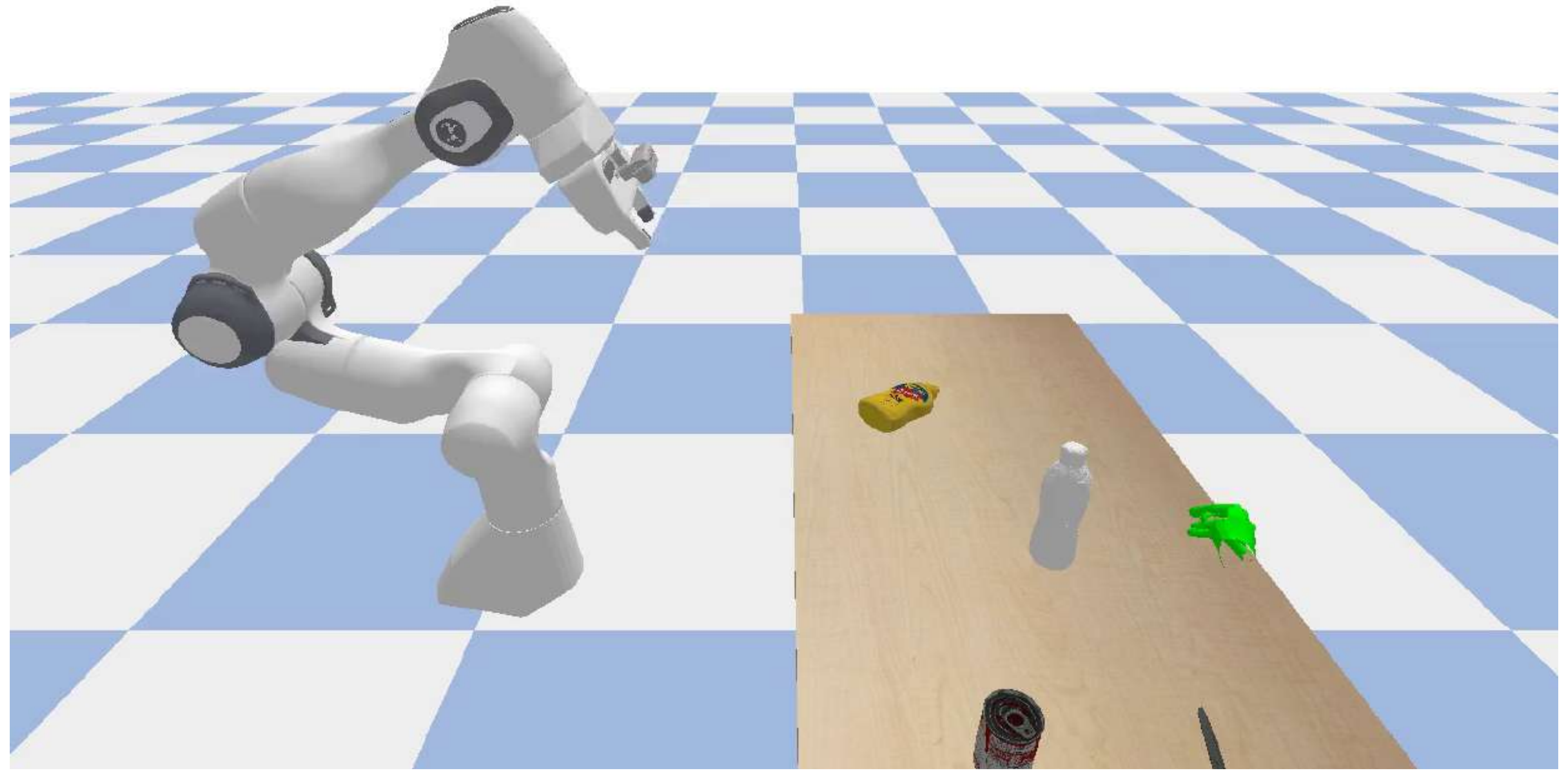
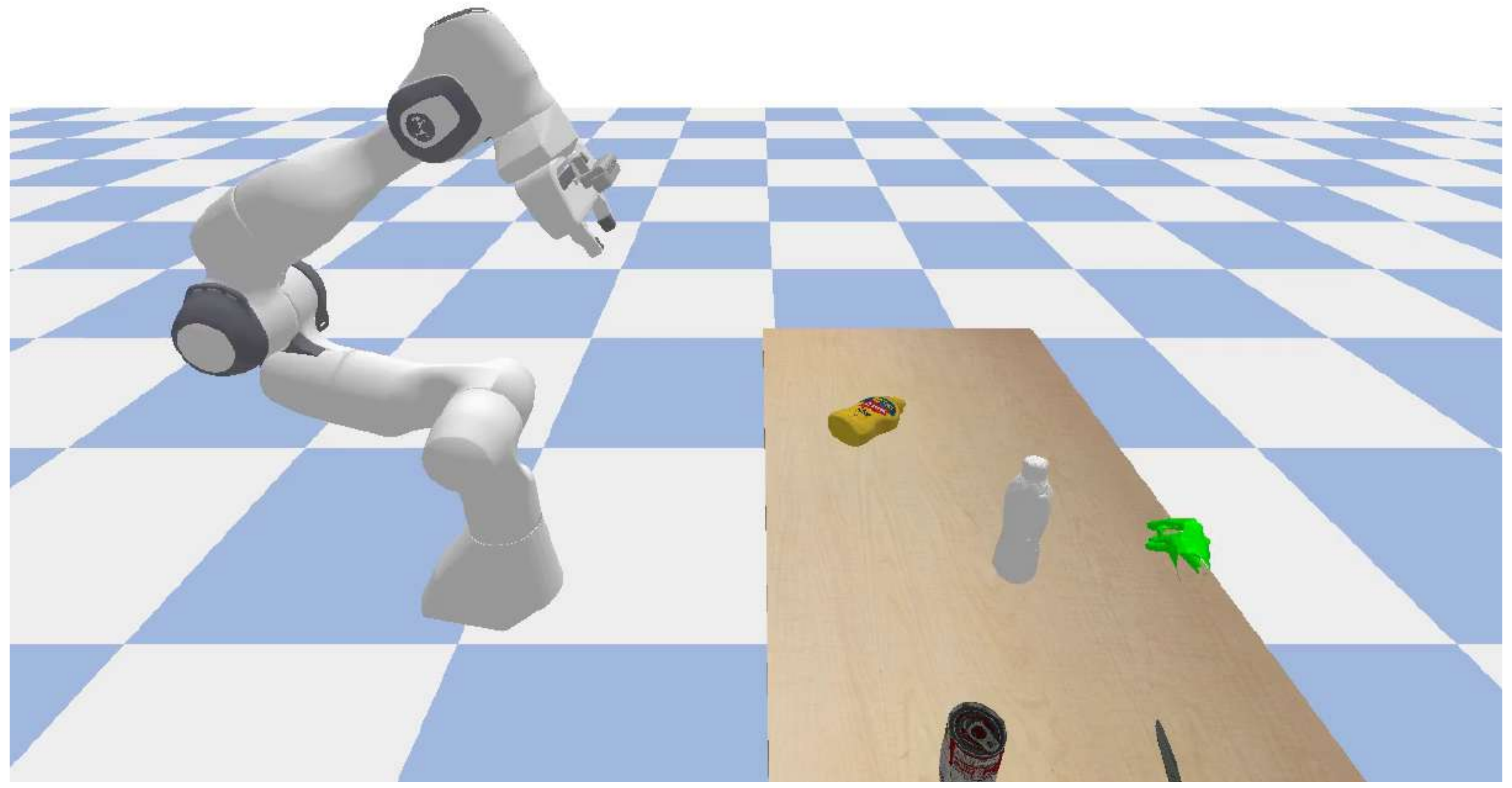


successful rate(%): 68.75



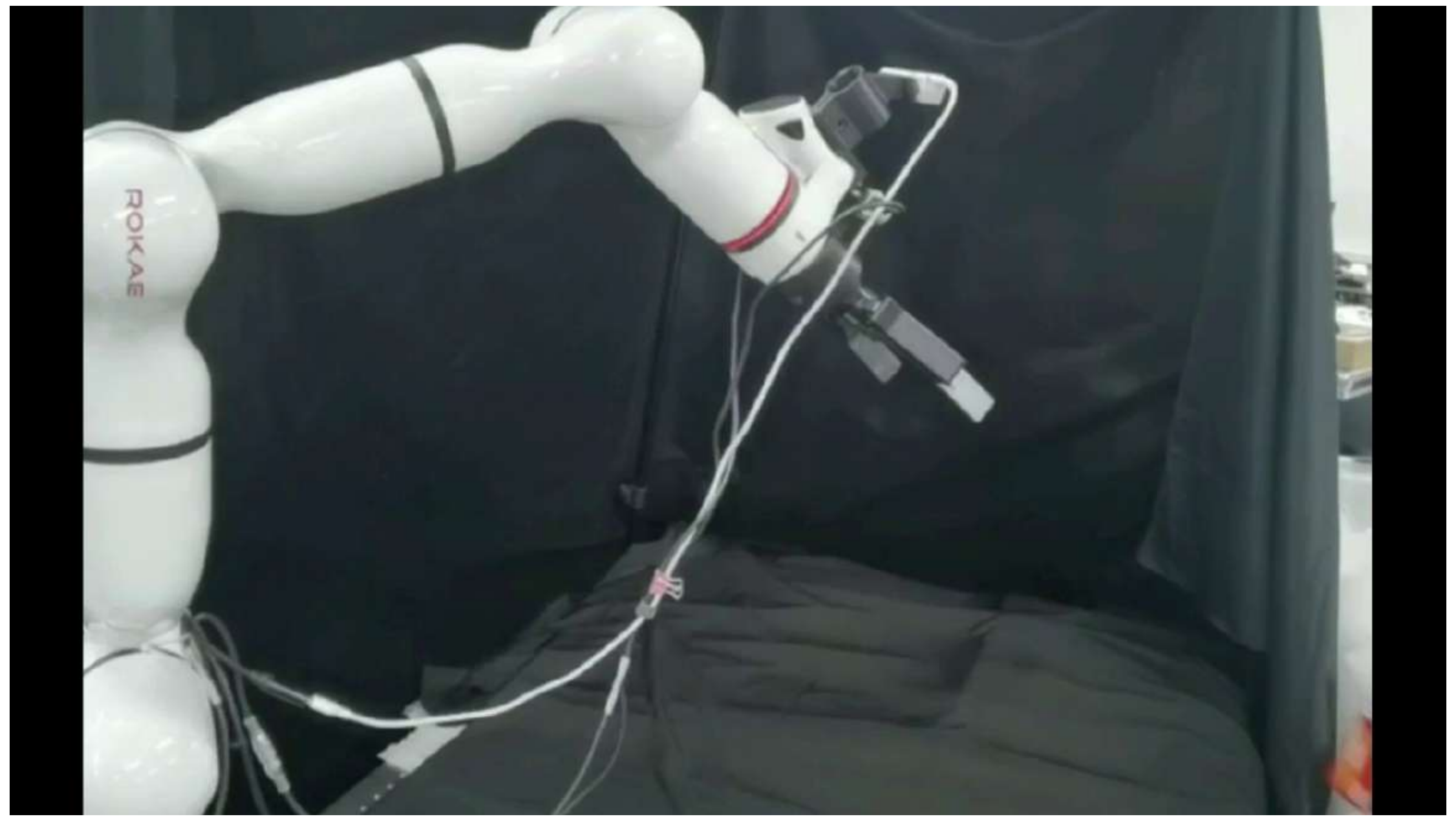
successful rate(%): 84.03

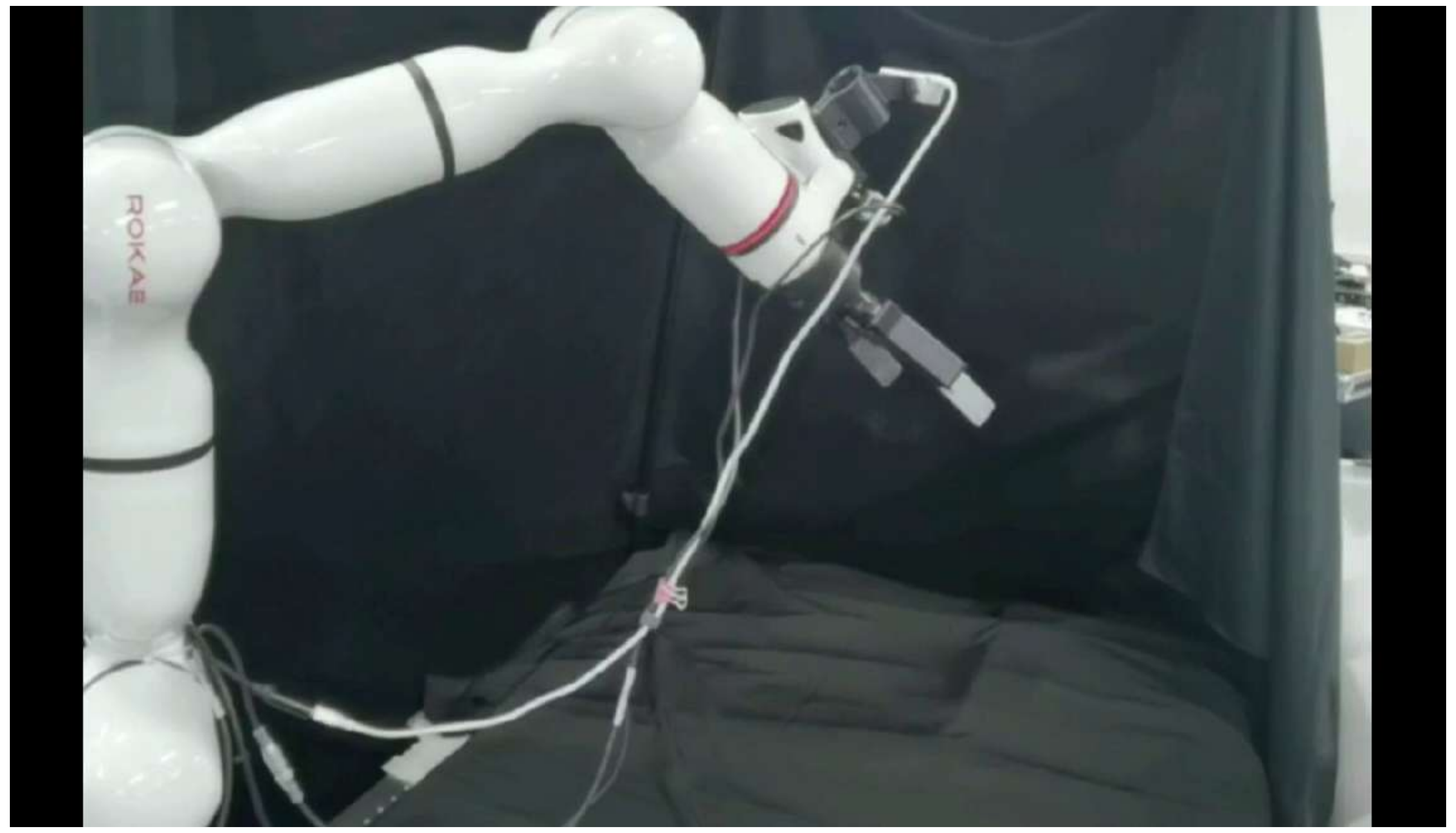
Qualitative Results: Simulation Experiments

	(Baseline) HandoverSim2real	**Ours**
Simultaneous: Random human motion	 successful rate(%): 29.17	 successful rate(%): 40.43
Simultaneous: Human motion from a large-scale mocap dataset	 successful rate(%): 52.4	 successful rate(%): 62.4

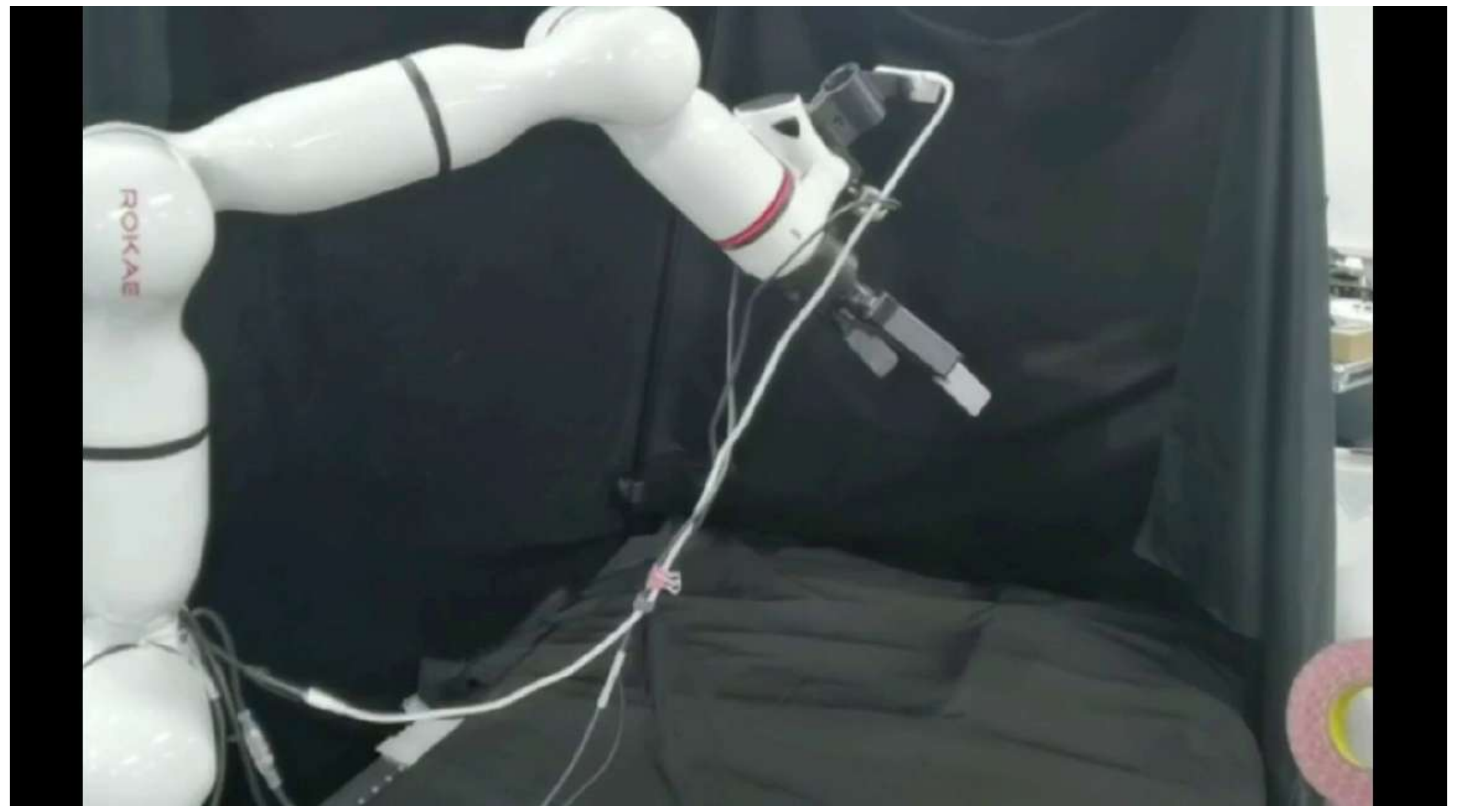
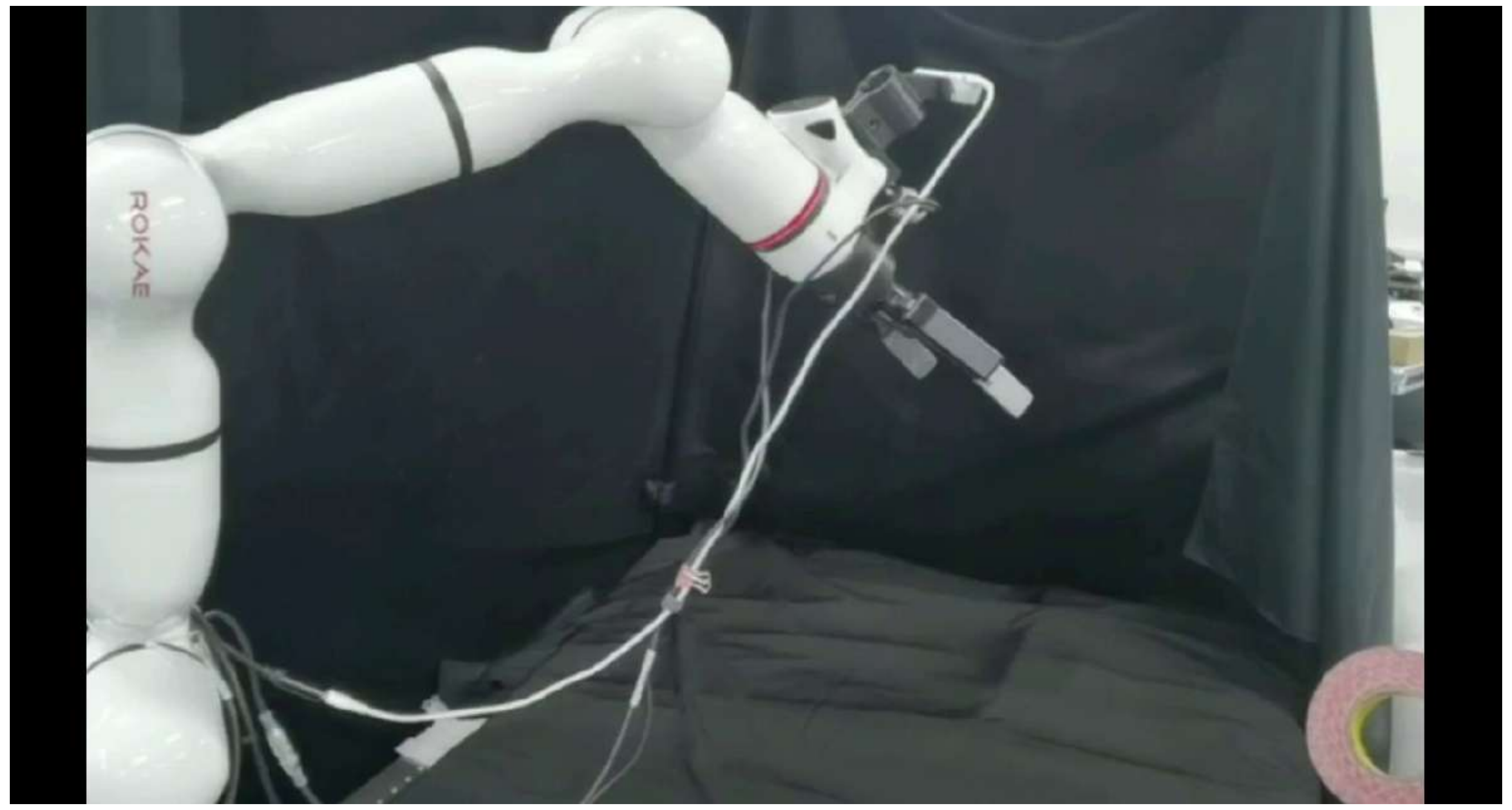
Qualitative Results: Real-world Experiments

(Baseline) GA-DDPG | **(Baseline) HandoverSim2real** | **Ours**





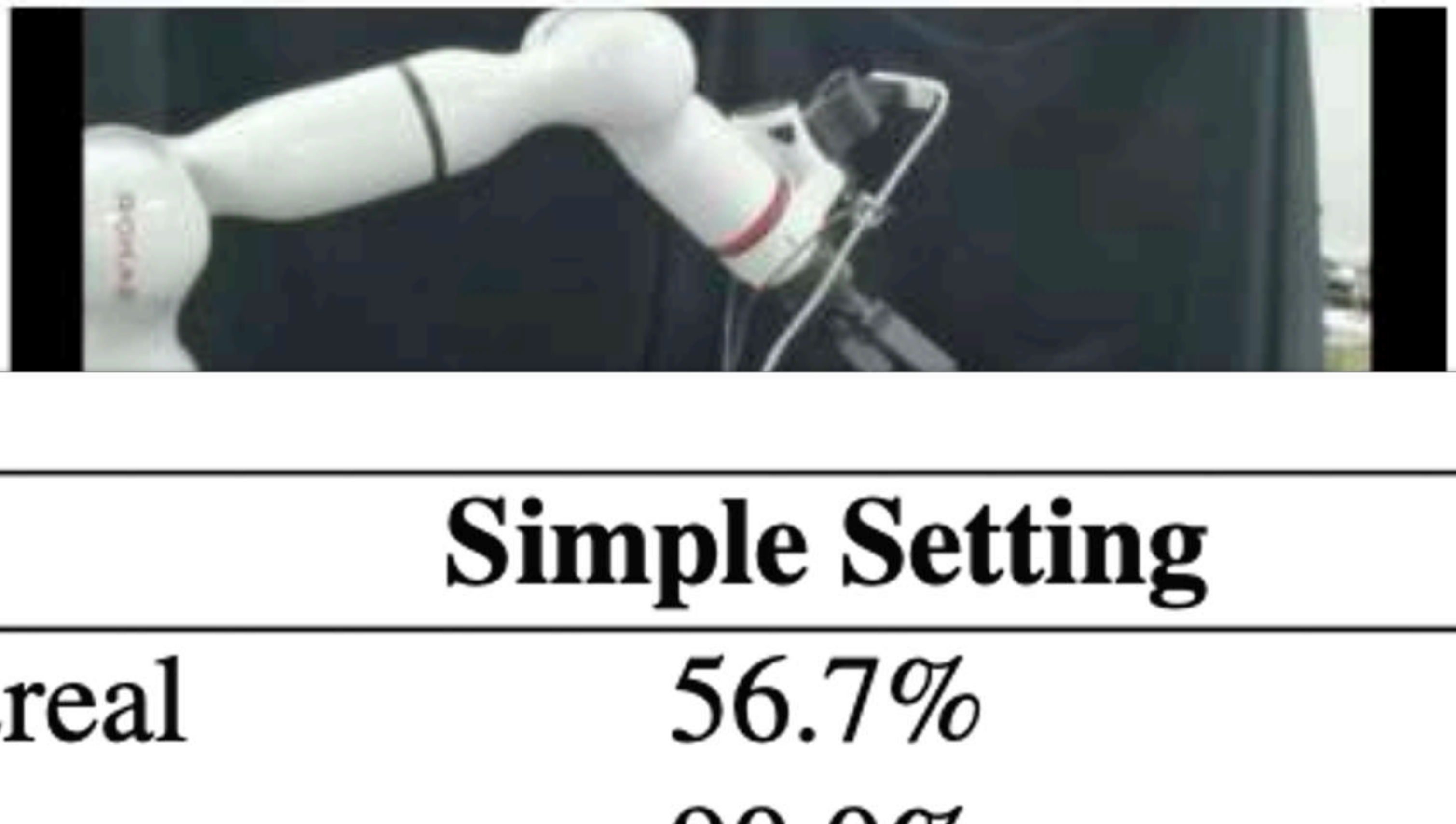
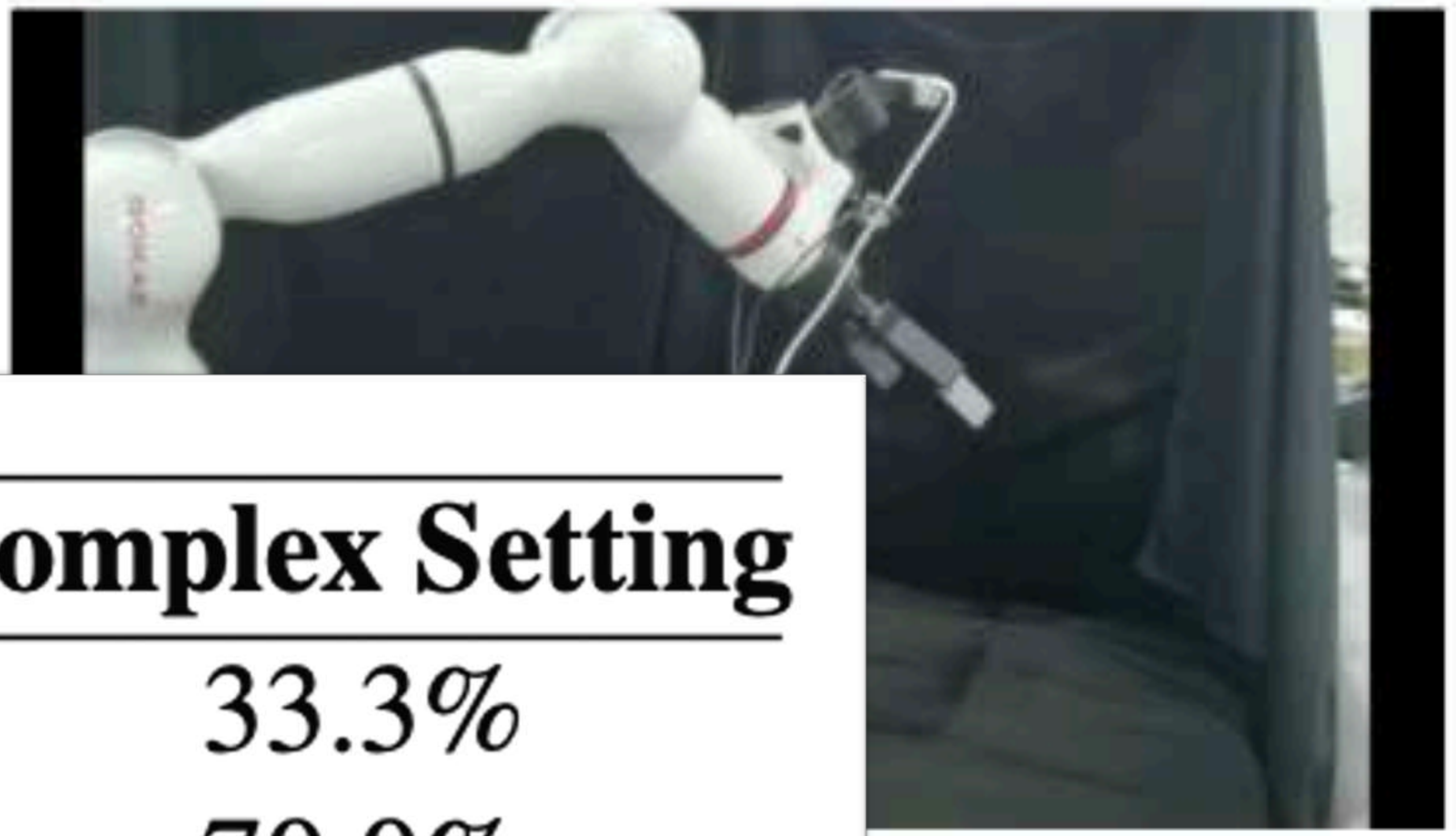
For more trajectory transitions like rotations, our model demonstrates robustness compared to baseline methods.

For previously unseen objects with diverse geometries like sticky tapes, our model exhibits greater generalizability.

Qualitative Results: Real-world Experiments

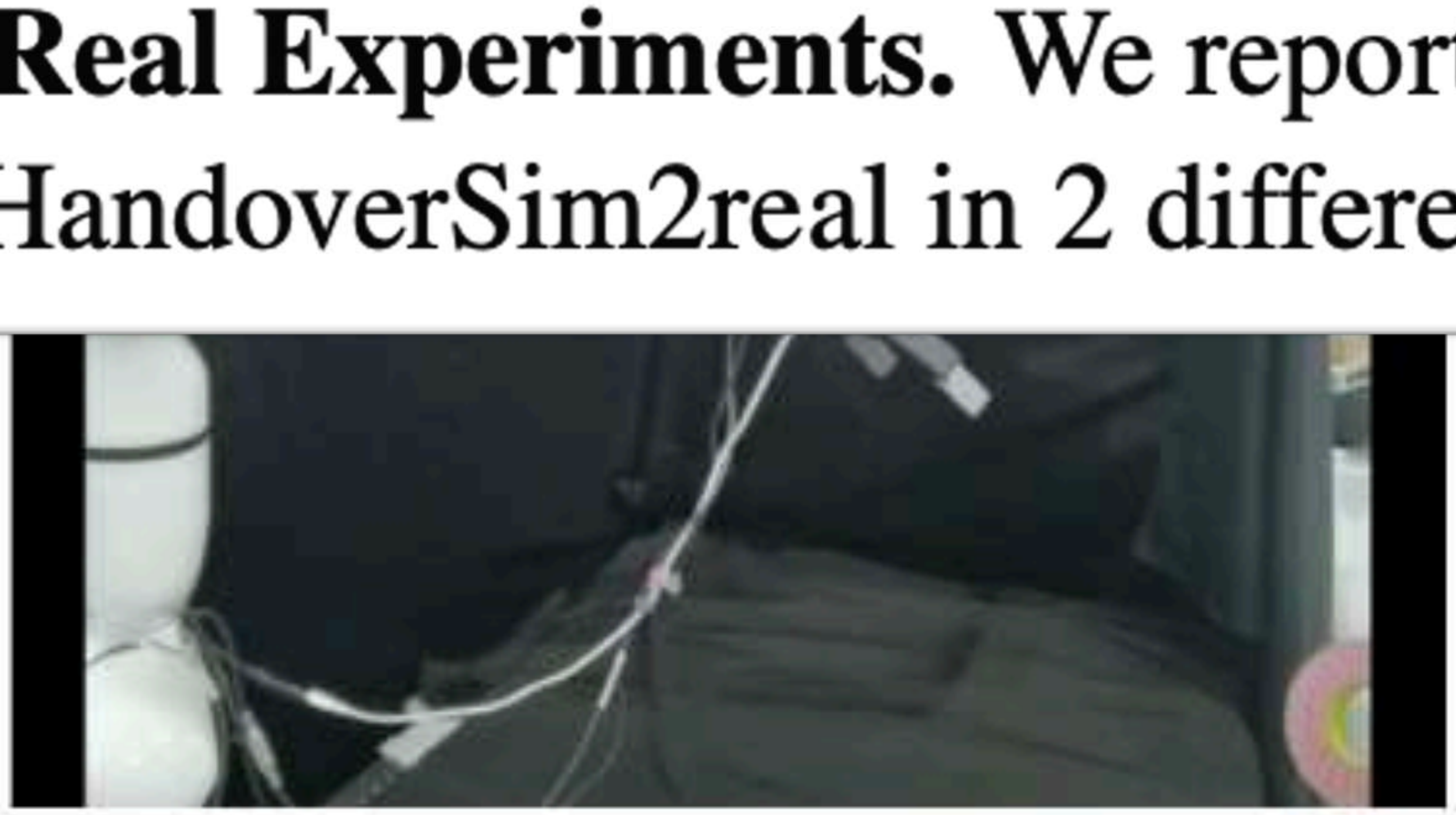
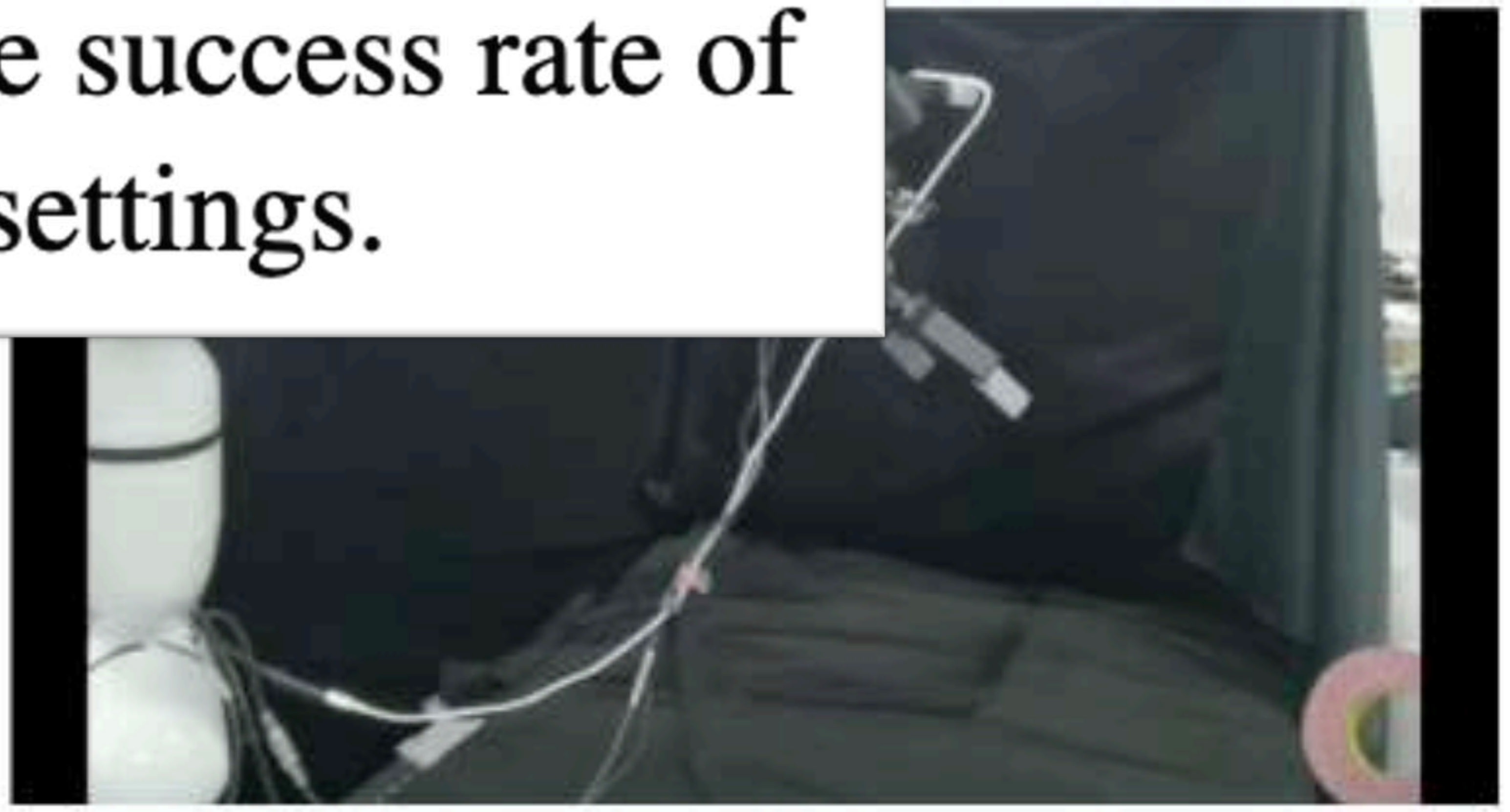
(Baseline) GA-DDPG | **(Baseline) HandoverSim2real** | **Ours**

For more tr … **ds.**

Methods	Simple Setting	Complex Setting
Handover-Sim2real	56.7%	33.3%
Ours	90.0%	70.0%

Table 3. **Sim-to-Real Experiments.** We report the success rate of our method and HandoverSim2real in 2 different settings.

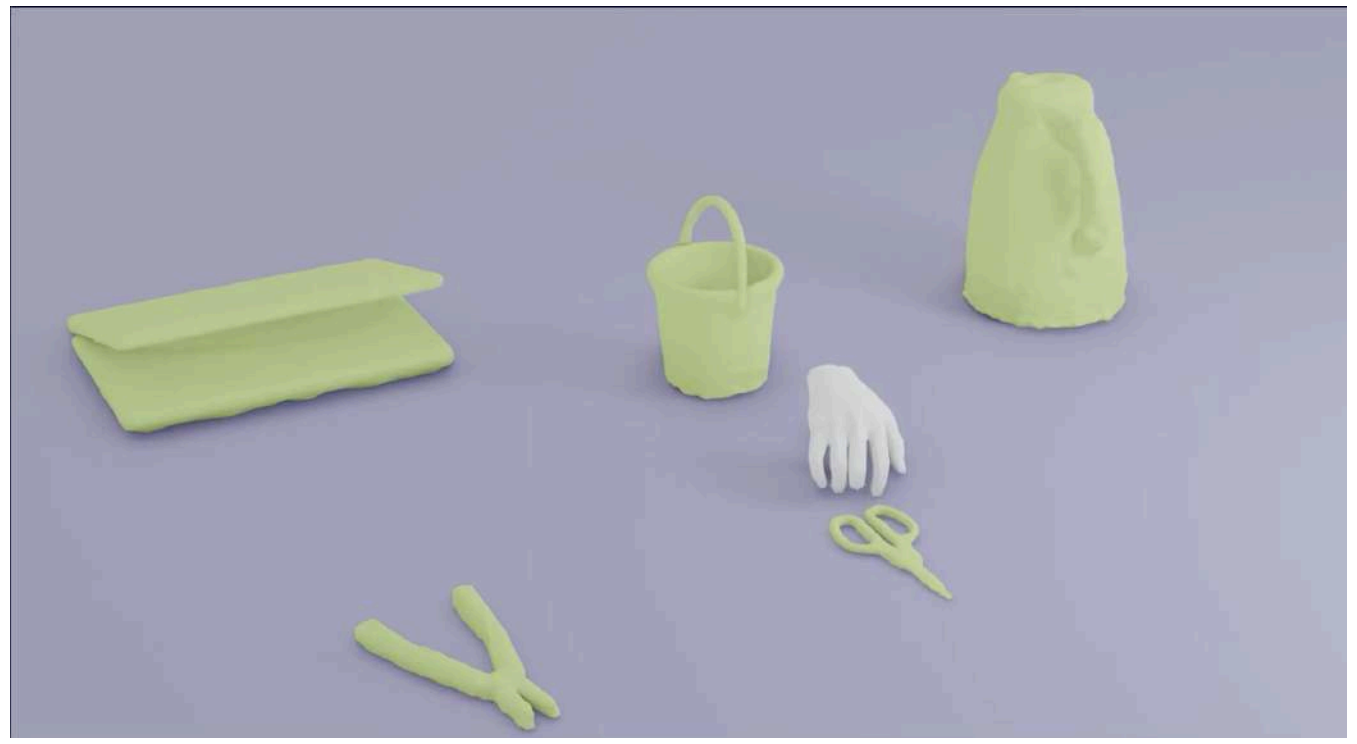




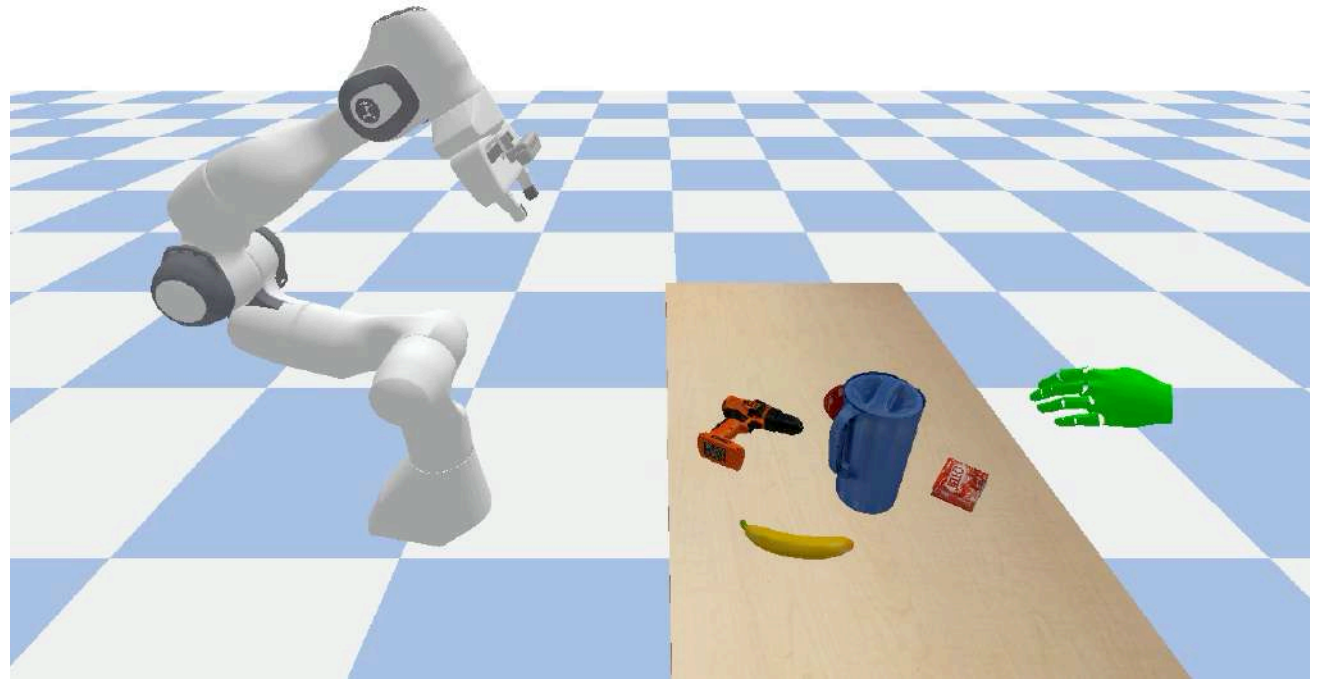
For previously unseen objects with diverse geometries like sticky tapes, our model exhibits greater generalizability.

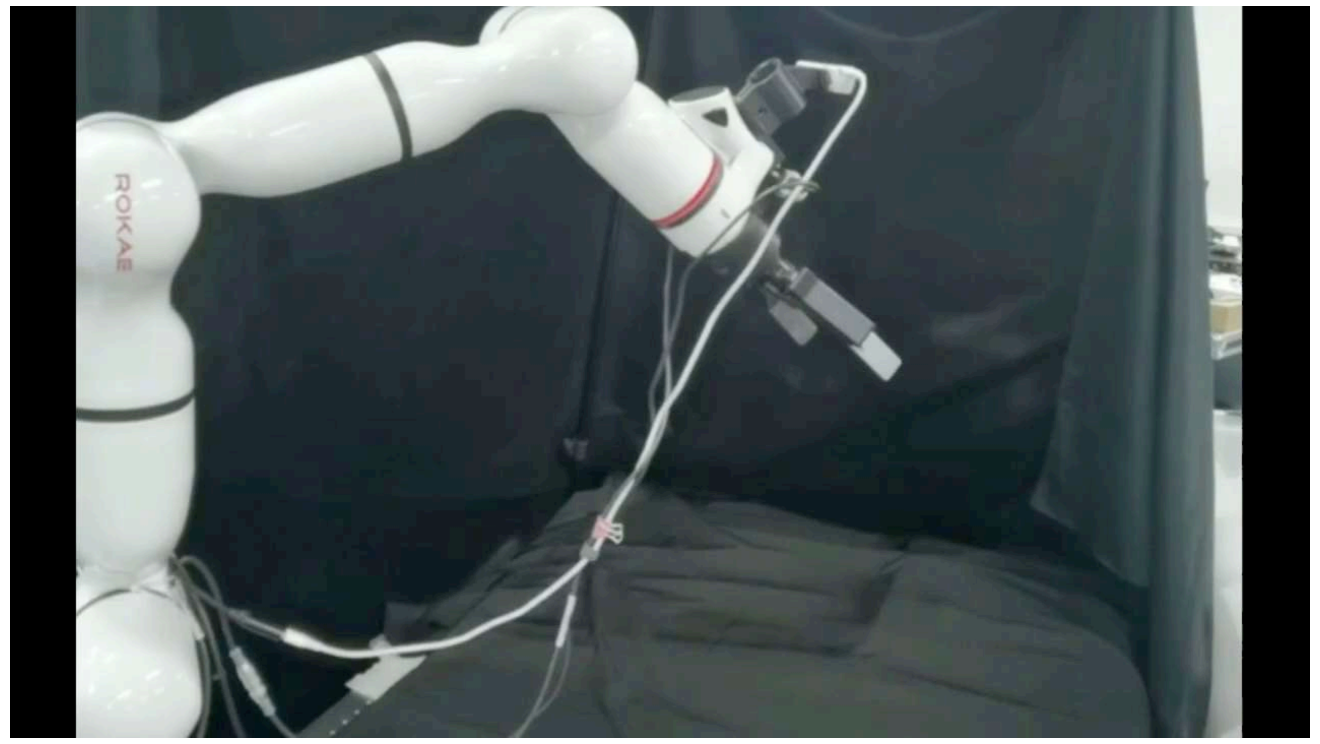

Human-Centered Robot Simulator
Human-Centered EAI
Human Interaction Capturing
Interactable Asset Creation
Open-World Perception
Police Car
Dragon Chair
Scissor
Data Driven
Asset Support
Visual Perception
Human Interaction Synthesis
Human-Centered Robot Simulator
Human-Centered Robotics
Human Simulation
Simulation Support
Collaborative Transport
Human-to-Robot Handover

Takeaway: Real-to-Sim-to-Real Solution

Creating Sim from Real → Learning in Sim → Deploying in Real







Embodied Task Execution

Embodied Agent

Thank you!

image credits: Matterport3D

Environment

Handover

Human Robot Interaction

Human