Object recognition and computer vision 2024

Learning visual representations for robotics

Ivan Laptev

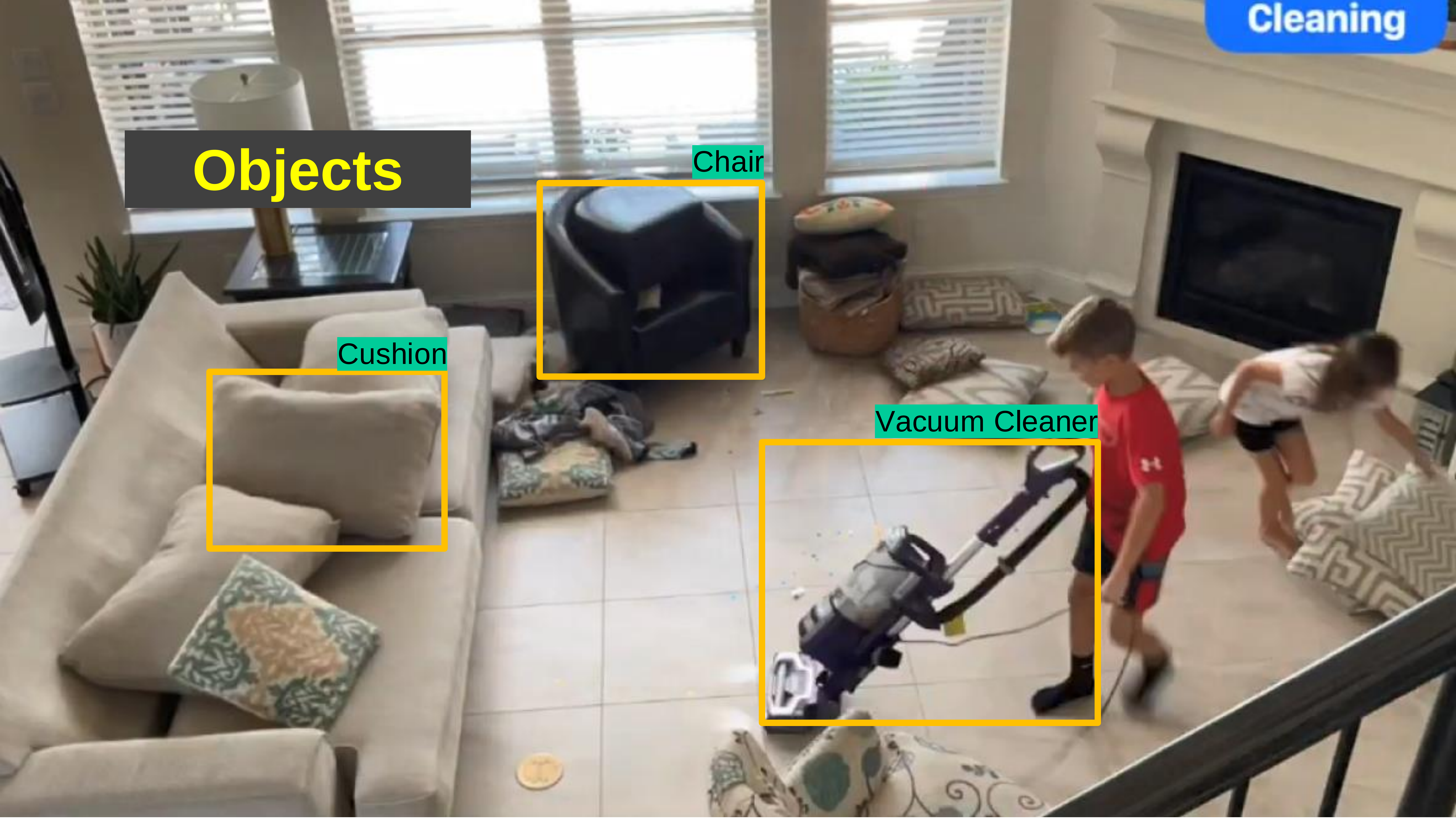
Ivan.Laptev@mbzuai.ac.ae
https://www.di.ens.fr/~laptev
Professor, MBZUAI, United Arab Emirates
(on leave from Inria/Willow, DI ENS)

Inria informatics mathematics

ENS

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

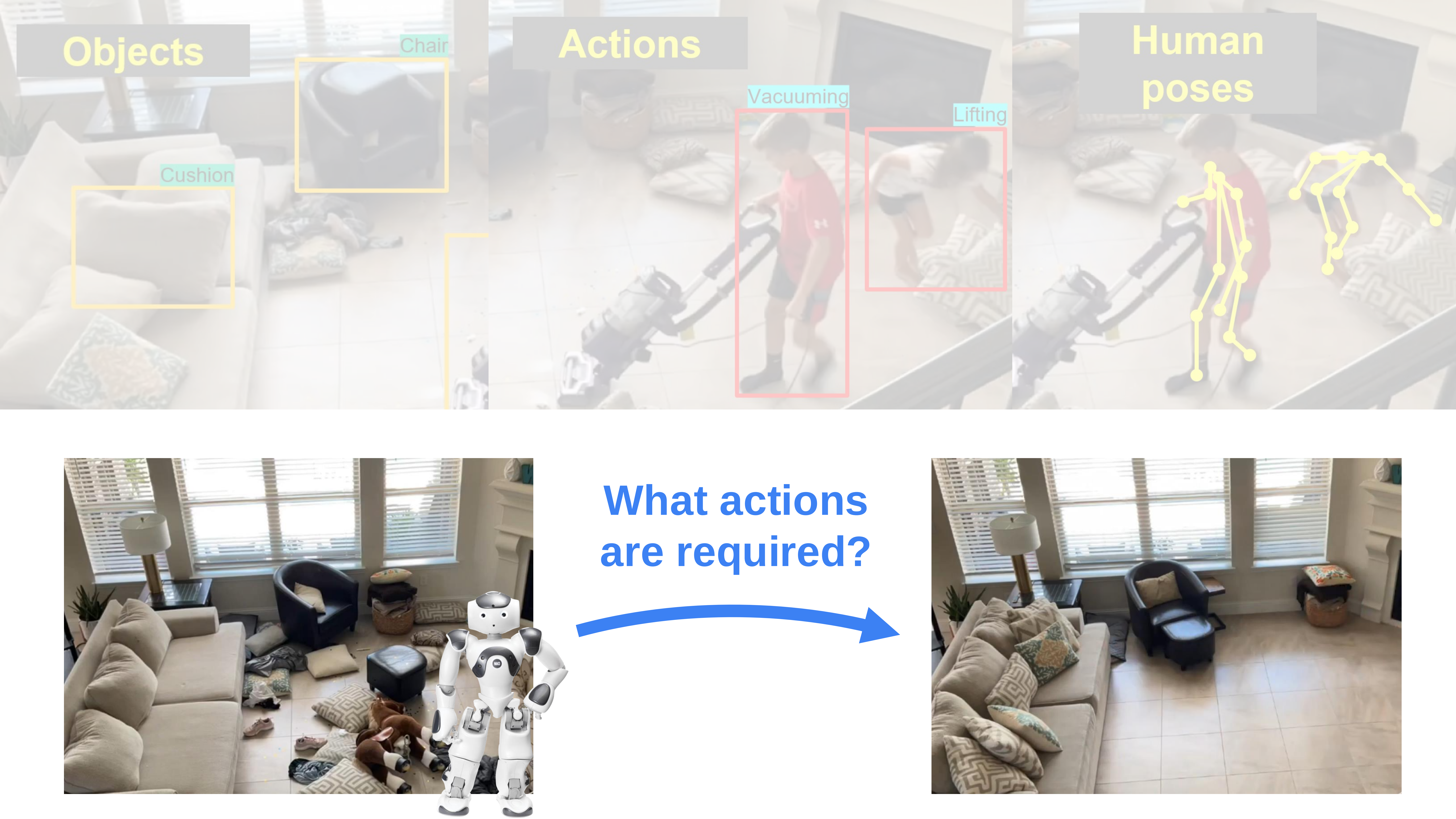

NASA

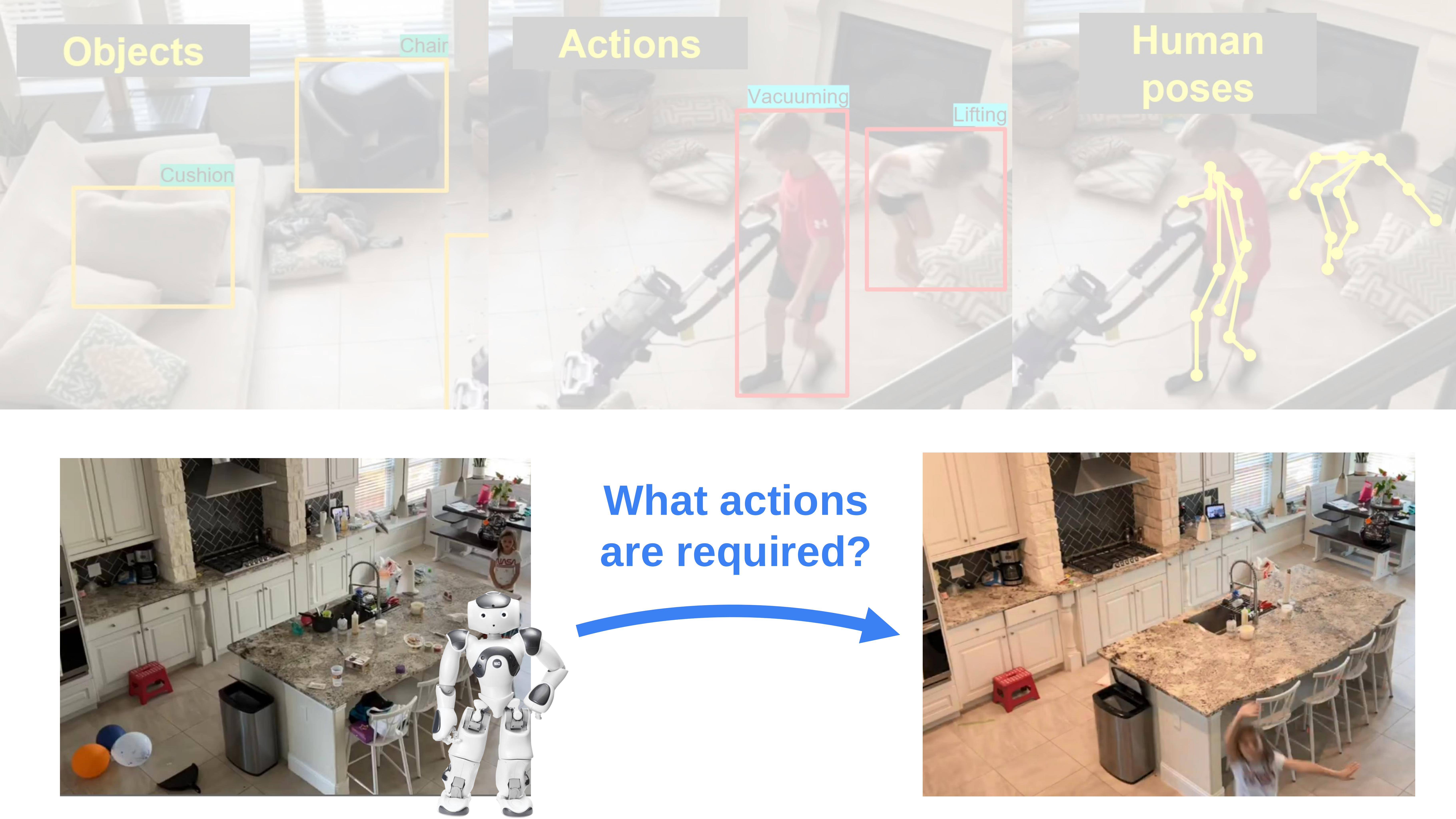

Cleaning
Objects
Chair
Cushion
Vacuum Cleaner


Cleaning
Actions
Vacuuming
Lifting

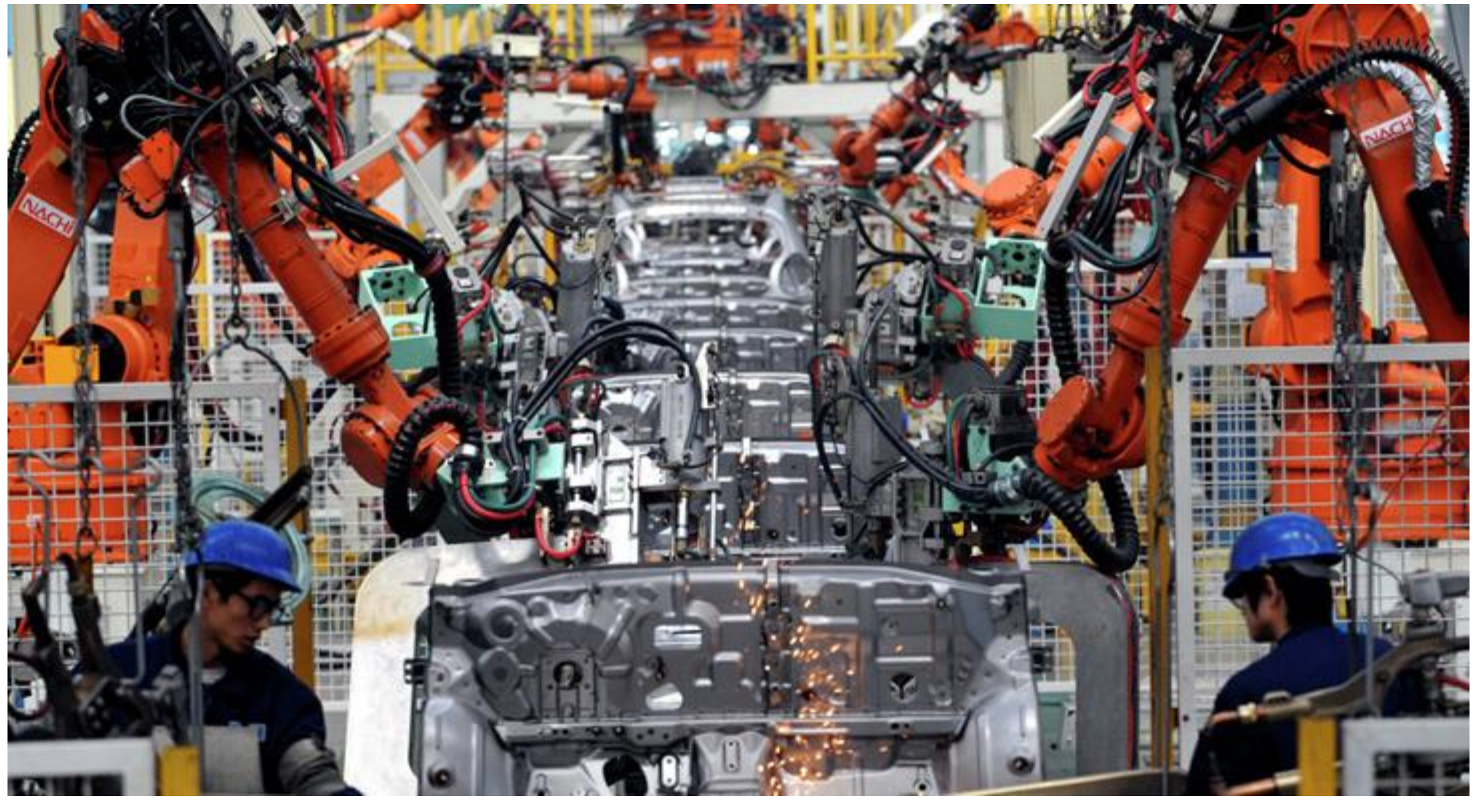

Cleaning
Human poses


Objects
Chair
Cushion
Actions
Vacuuming
Lifting
Human poses


Objects
Chair
Cushion
Actions
Vacuuming
Lifting
Human poses
What actions are required?


Objects
Chair
Cushion
Actions
Vacuuming
Lifting
Human poses
What actions are required?

Robotics



Structured → **Perception** → **Unstructured**

Factory Robots:

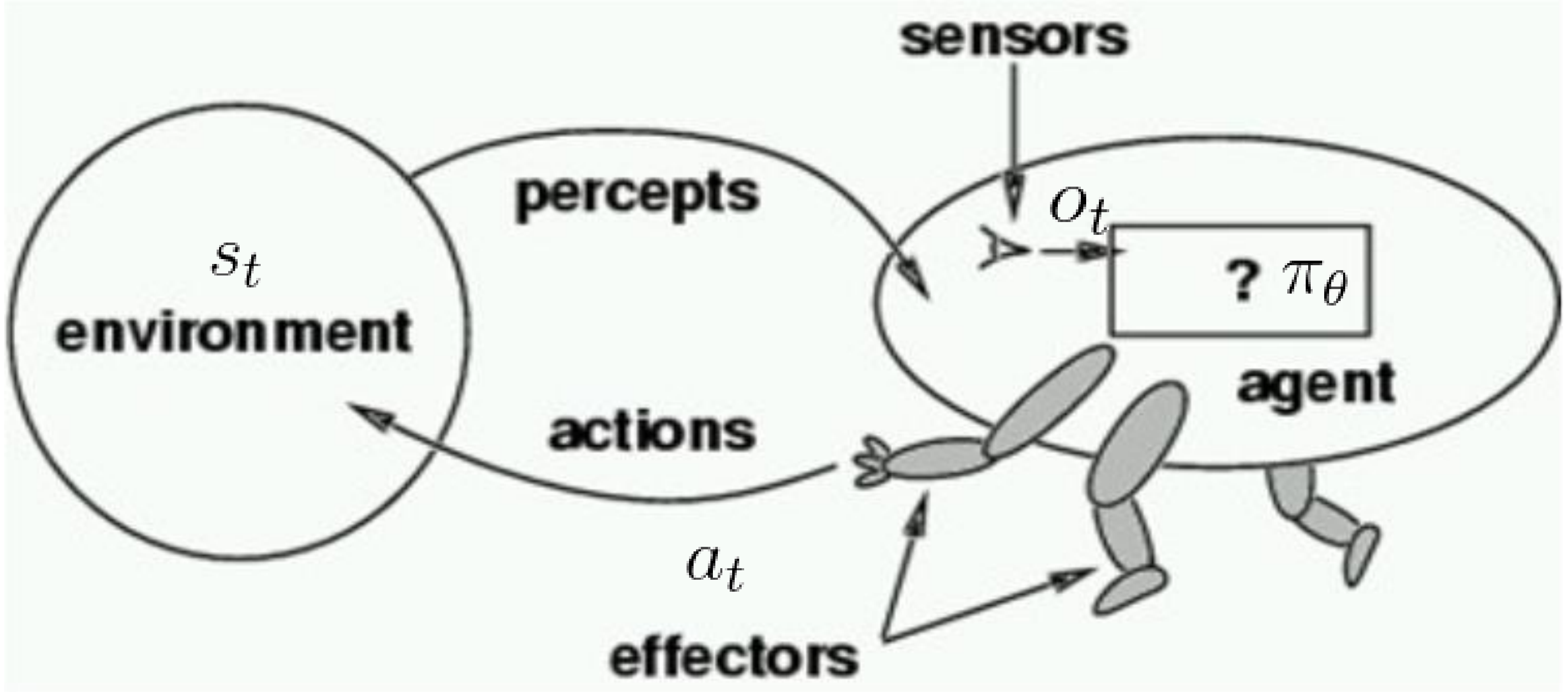
- Specialized, Task specific
- Very constrained factory environment where everything is predefined

Collaborative Robots:

- Open environment with varying conditions
- Needs to be generalist, handle multiple tasks, collaborate with people

How to learn actions given raw sensory input?

Perception-Action cycle

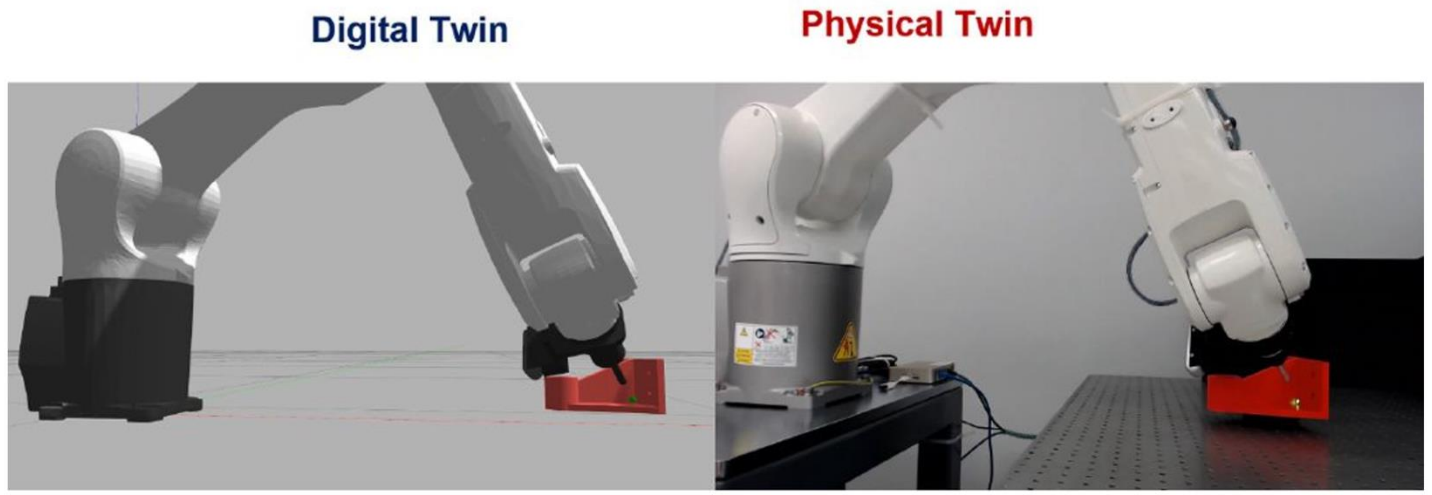


s_t state
o_t observation
a_t action
π_θ policy

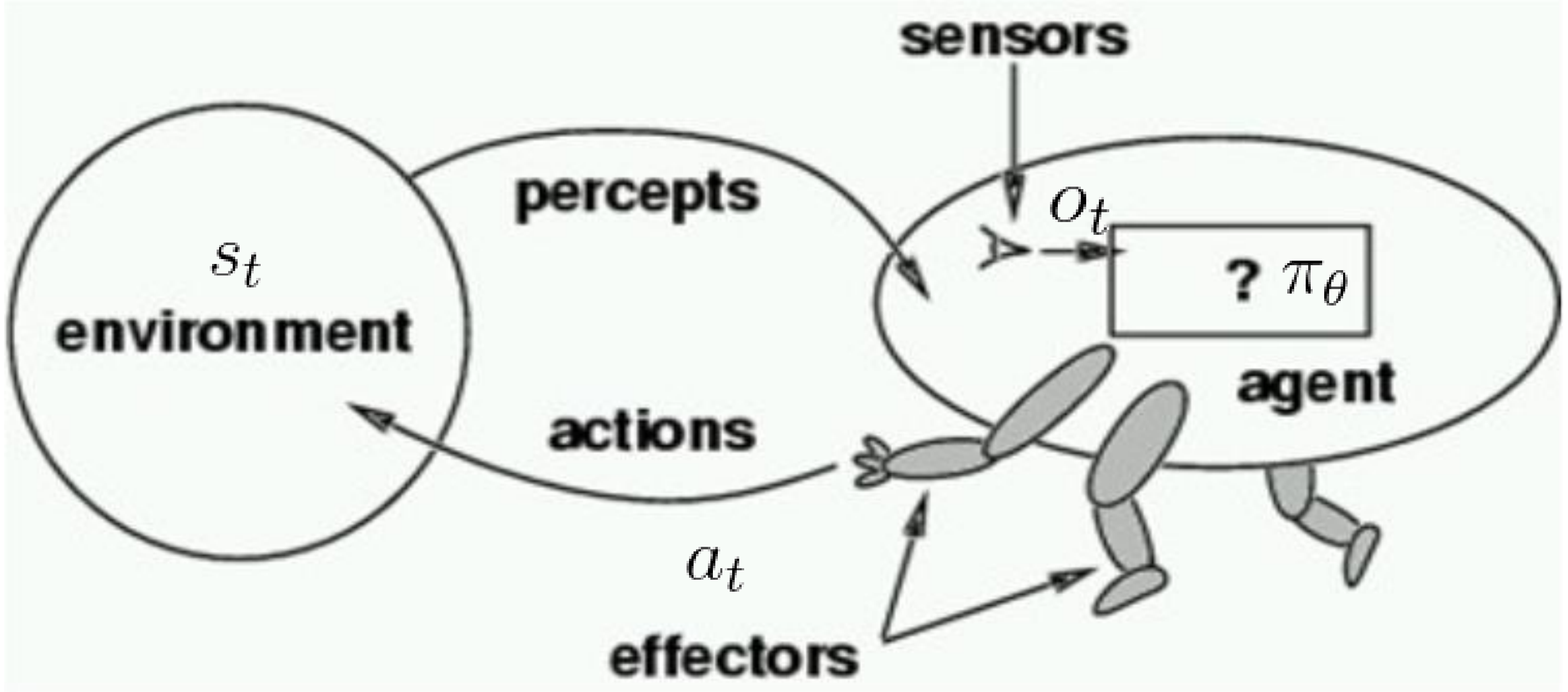
How to obtain $\pi_\theta(a_t|o_t)$?

Strategy 1: State-based

- estimate $\tilde{s}_t$ from o_t
- use Newtonian physics and explicit 3D geometry to derive a_t



Perception-Action cycle



s_t state
o_t observation
a_t action
π_θ policy

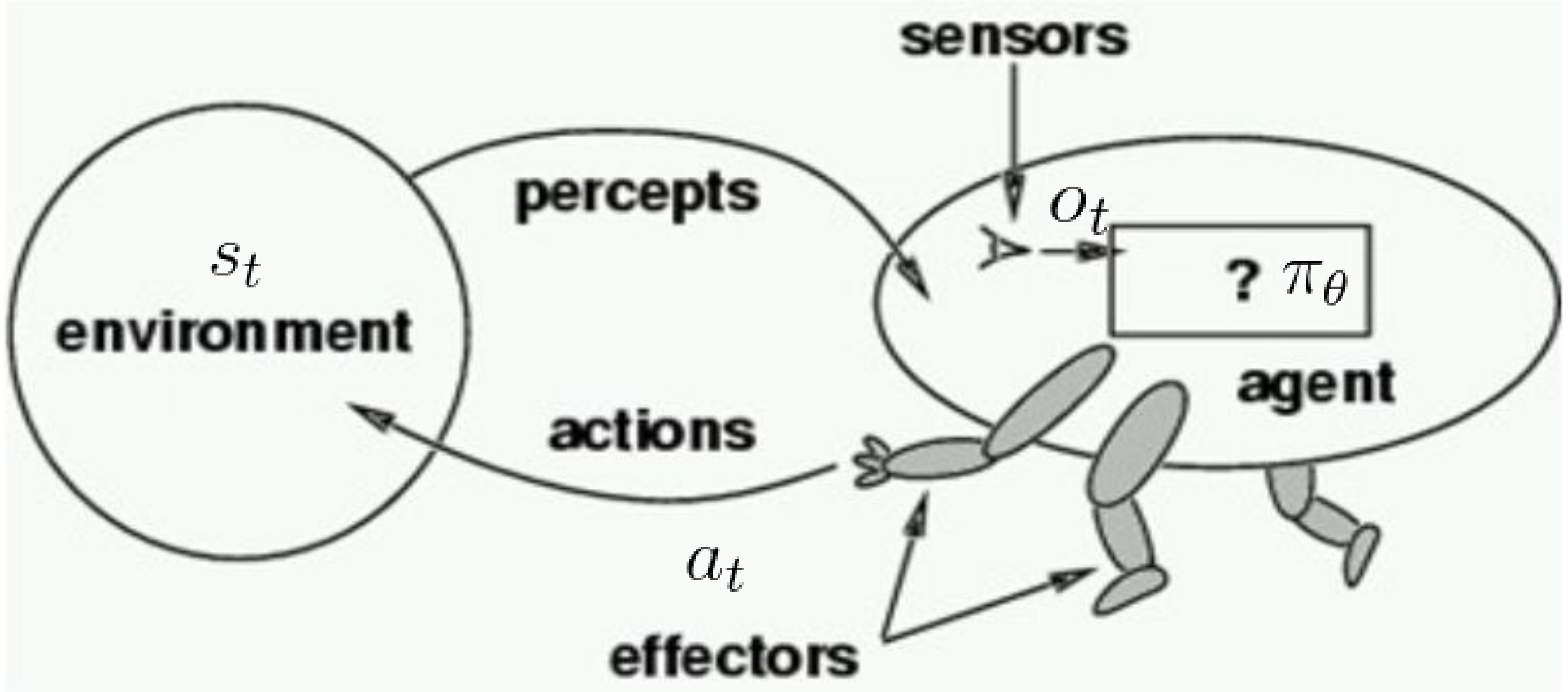
How to obtain $\pi_\theta(a_t|o_t)$?

Strategy 1: State-based

- estimate $\tilde{s}_t$ from o_t
- use Newtonian physics and explicit 3D geometry to derive a_t

estimating $\tilde{s}_t$ from o_t can be very hard

Perception-Action cycle



s_t state

o_t observation

a_t action

π_θ policy

How to obtain $\pi_\theta(a_t|o_t)$?

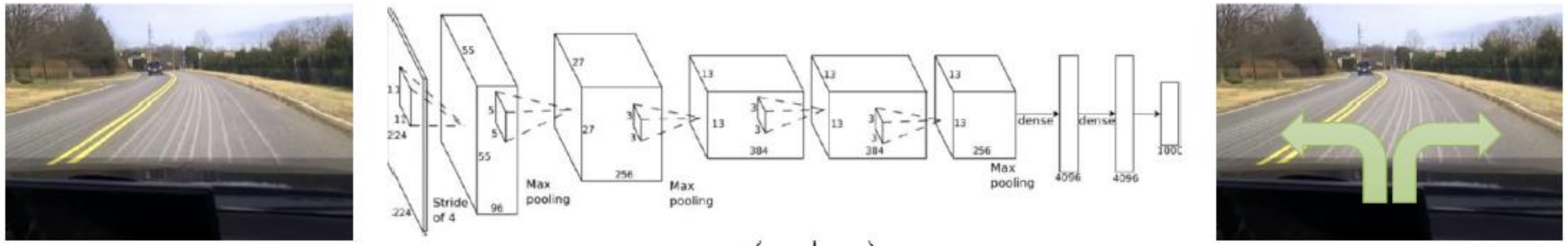
Strategy 1: State-based

- estimate $\tilde{s}_t$ from o_t
- use Newtonian physics and explicit 3D geometry to derive a_t

Strategy 2: sensor-based

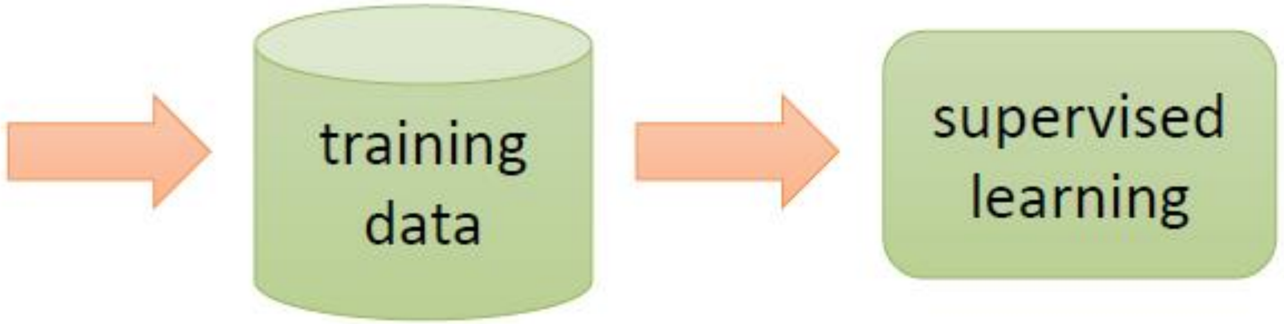
- learn $\pi_\theta(a_t|o_t)$ from the data

Imitation Learning



$\mathbf{o}_t$ $\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$ $\mathbf{a}_t$

$\mathbf{o}_t$
$\mathbf{a}_t$

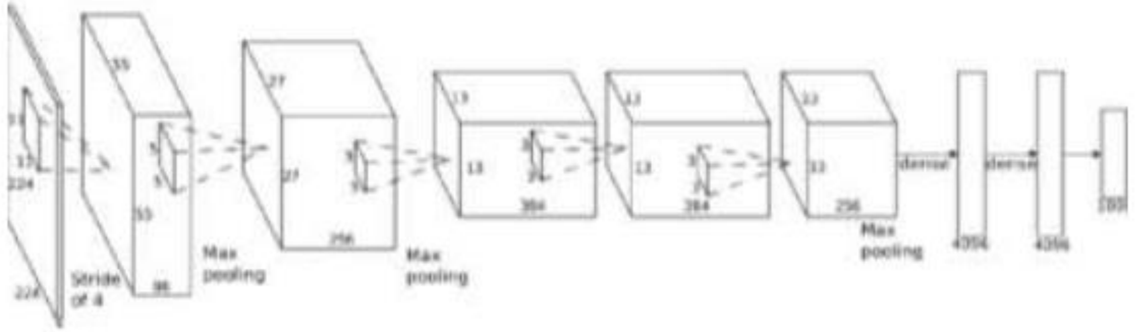


$\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$

behavior cloning

Slide credit: S. Levine

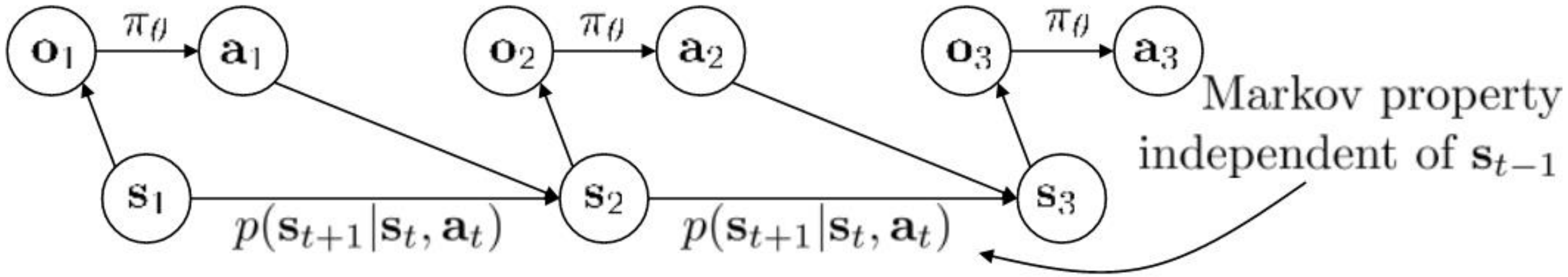
Imitation Learning



$\mathbf{o}_t$ $\quad \pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$ $\quad \mathbf{a}_t$

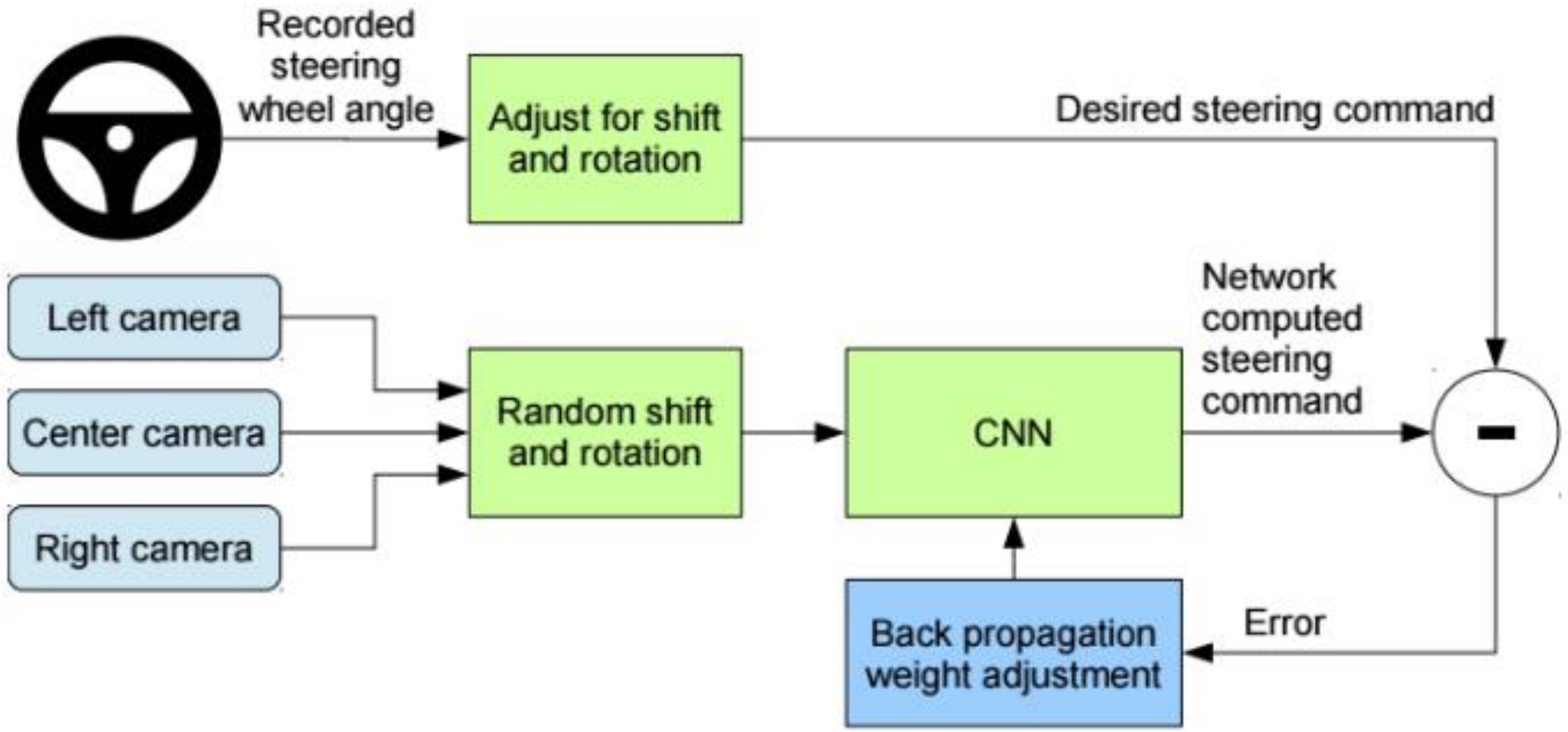
$\mathbf{s}_t$ – state
$\mathbf{o}_t$ – observation
$\mathbf{a}_t$ – action

$\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$ – policy
$\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$ – policy (fully observed)



Slide credit: S. Levine

Imitation Learning



1. Collect a set of training data $\mathcal{D} = \{(o_i^*, a_i^*)\}_{i=1\ldots n}$ where actions are performed by an expert agent.
2. Train a model π_θ to minimize

$$\mathcal{L}(\pi_\theta) = l(\pi_\theta(o_i^*), a_i^*)$$

where l is any loss function.
For example : $\mathcal{L}(\pi_\theta) = \|\pi_\theta(o_i^*) - a_i^*\|_2^2$

Slide credit: S. Levine

End-to-end Driving via Conditional Imitation Learning

Felipe Codevilla, Antonio López - Computer Vision Center (CVC)
Matthias Müller - King Abdullah University of Science and Technology (KAUST)
Vladlen Koltun, Alexey Dosovitskiy - Intel Visual Computing Lab

We propose conditional imitation learning which allows an autonomous vehicle trained end-to-end to be directed by high-level commands.

Experiments in simulation and on a physical vehicle show that the method allows for goal-directed navigation guided by a topological planner or a user.

End-to-end driving via conditional imitation learning. Codevilla et al., ICRA 2018

Learning to Fly by Crashing

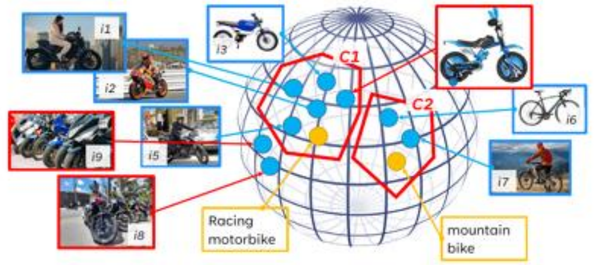
Dhiraj Gandhi, Lerrel Pinto, Abhinav Gupta

Carnegie Mellon University
The Robotic Institute

Learning to fly by crashing. Gandhi, Pinto and Gupta, IROS 2017

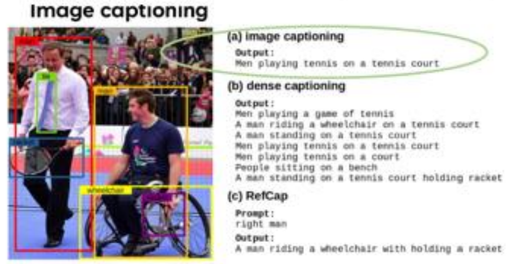
Passive vs. Active vision

Image-text retrieval



Visual grounding

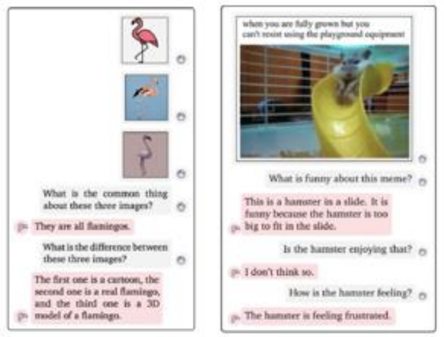
Image and video captioning



Visual Question answering

Image generation

Visual dialog

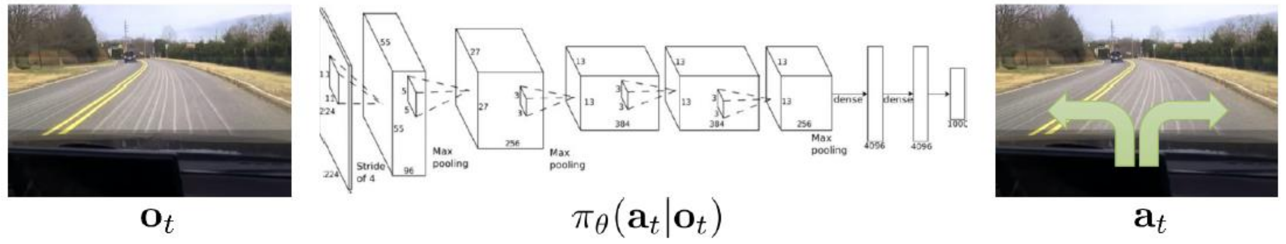


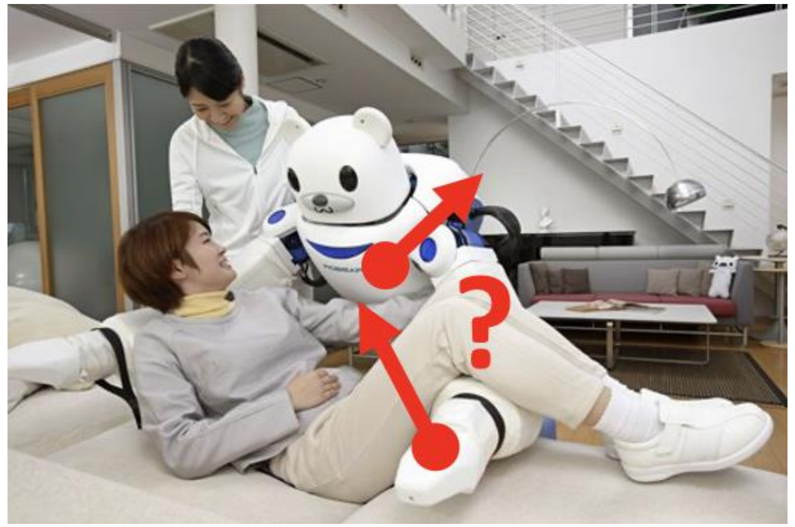
Machine translation

Passive:
Observations are pre-recorded

Active:
Observations depend on actions

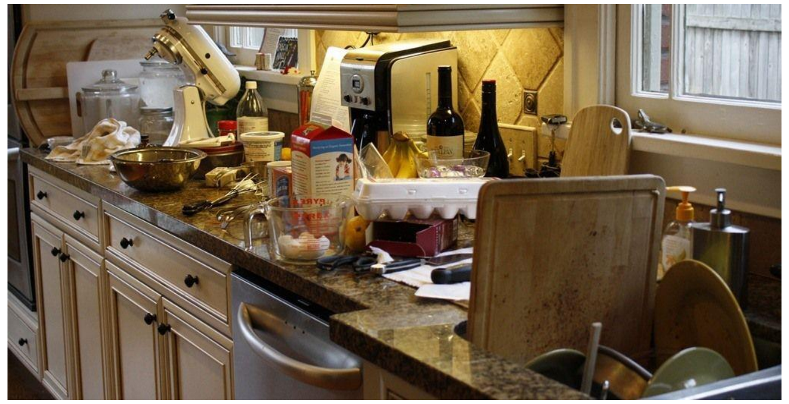
Challenges





- Supervision is costly or not unvaiable

Learn from human demonstrations



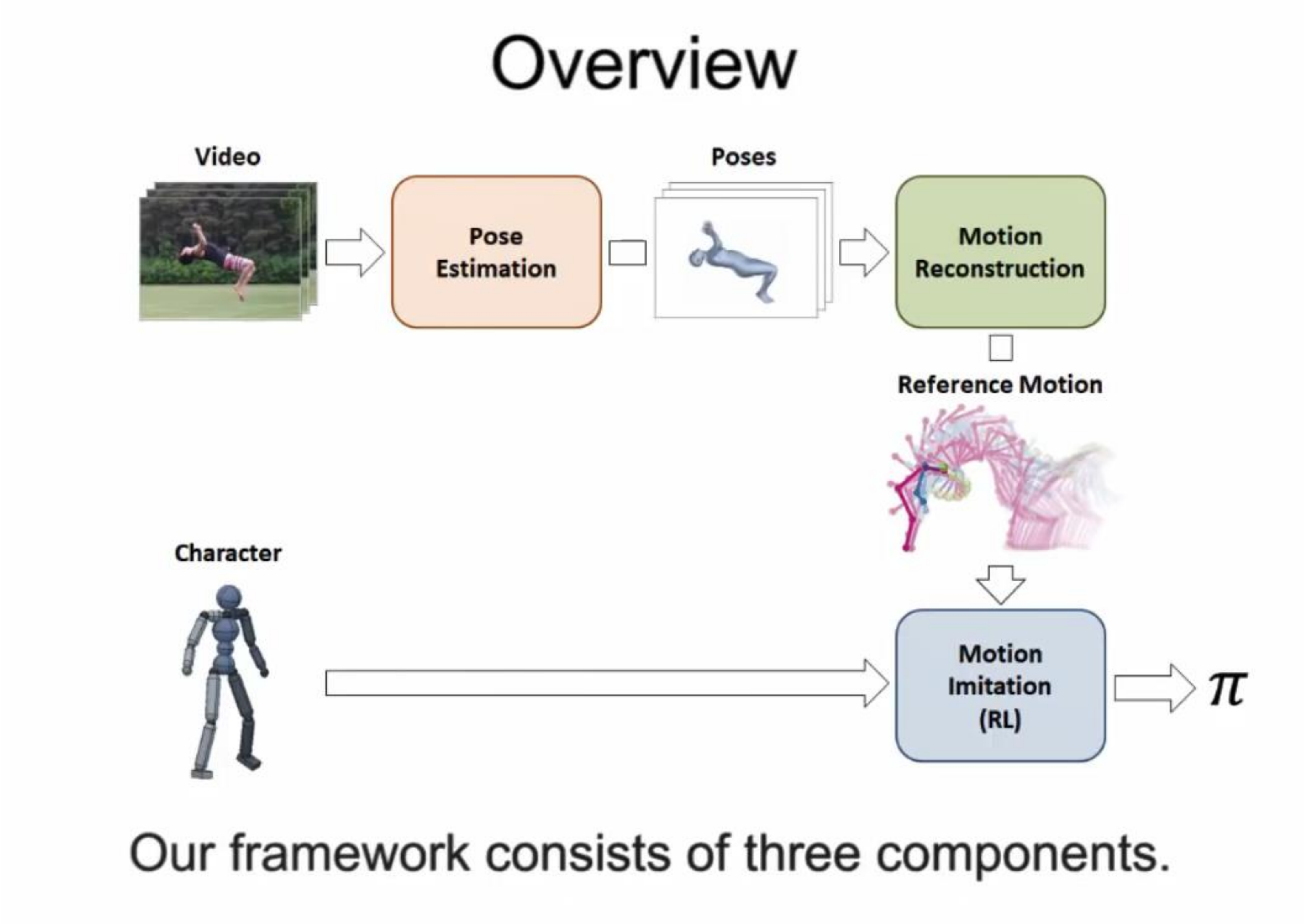
- Large diversity of environments and possible actions

- Control robots by natural language

Example: Learning skills from videos

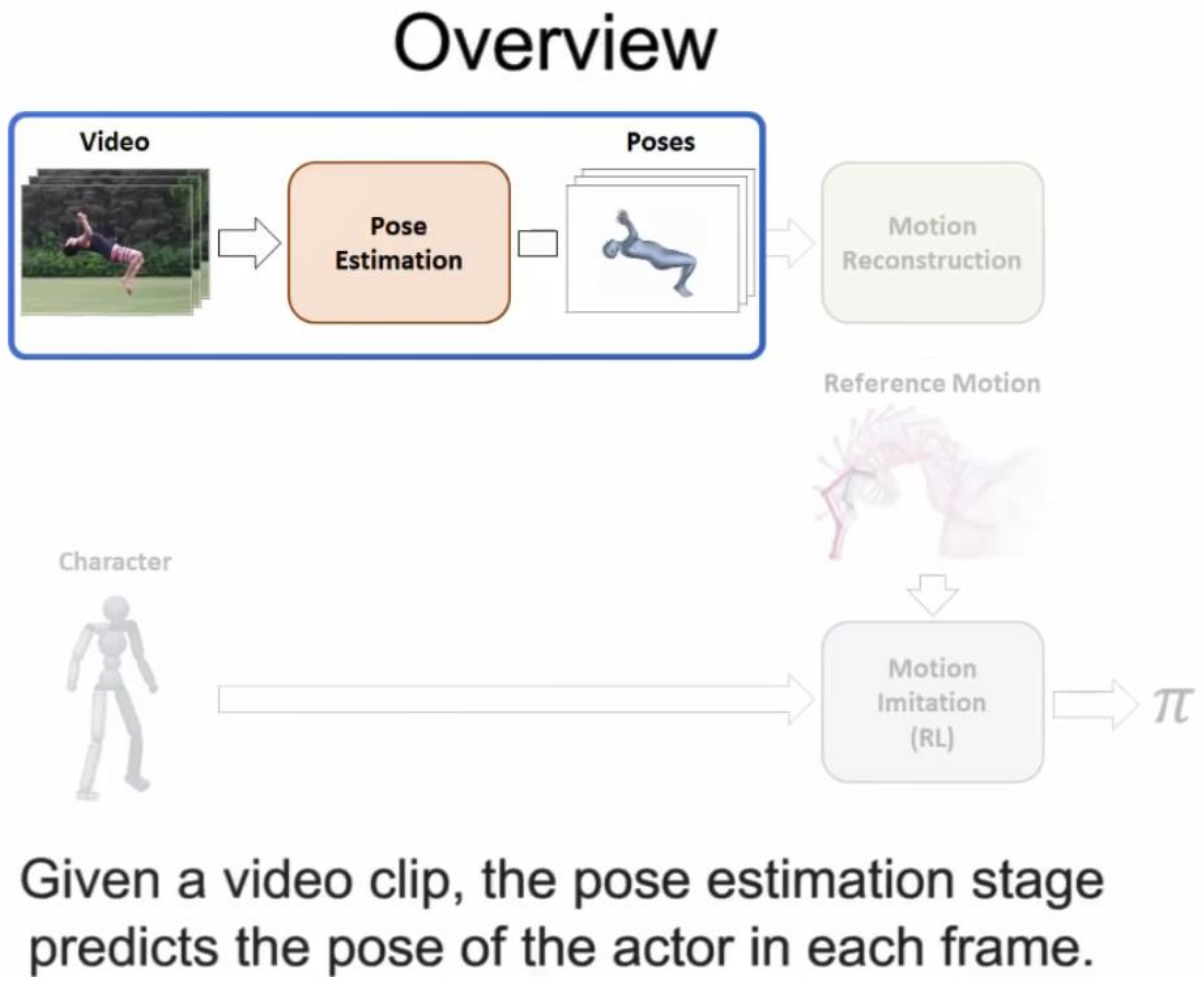
Peng XB, Kanazawa A, Malik J, Abbeel P, Levine S. SFV: Reinforcement learning of physical skills from videos. ACM TOG 2019

Example: Learning skills from videos



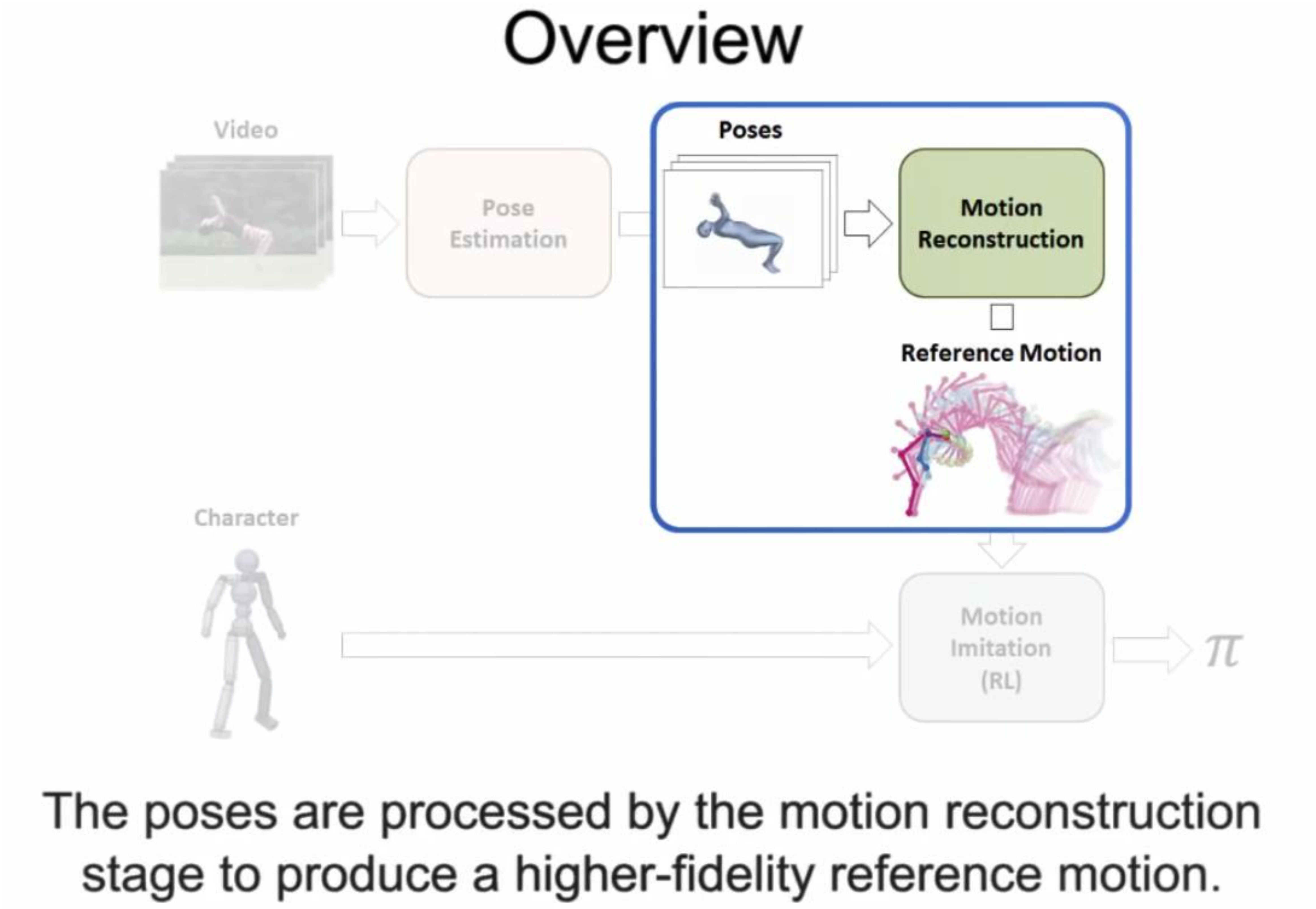
Peng XB, Kanazawa A, Malik J, Abbeel P, Levine S. SFV: Reinforcement learning of physical skills from videos. ACM TOG 2019

Example: Learning skills from videos



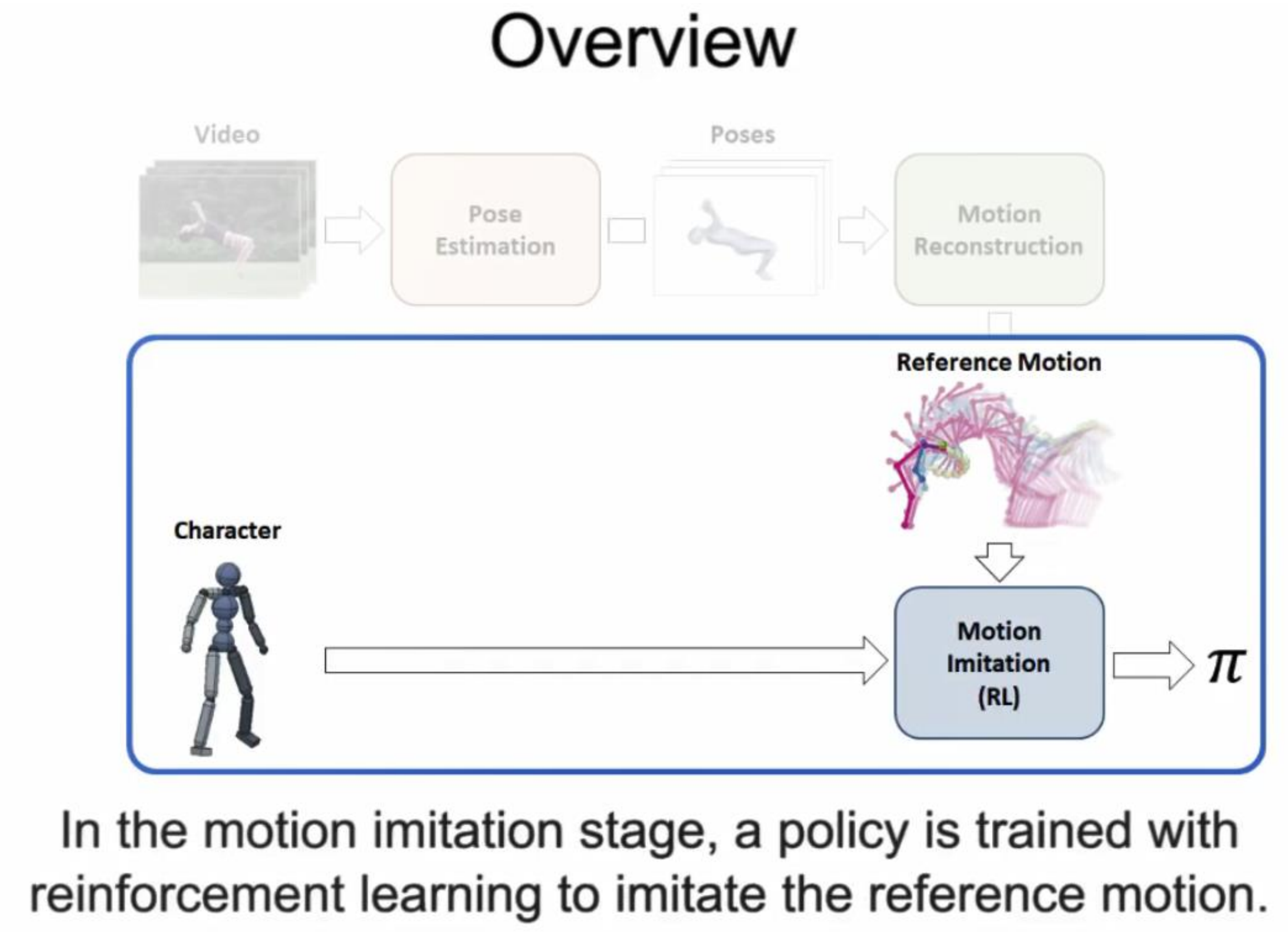
Peng XB, Kanazawa A, Malik J, Abbeel P, Levine S. SFV: Reinforcement learning of physical skills from videos. ACM TOG 2019

Example: Learning skills from videos



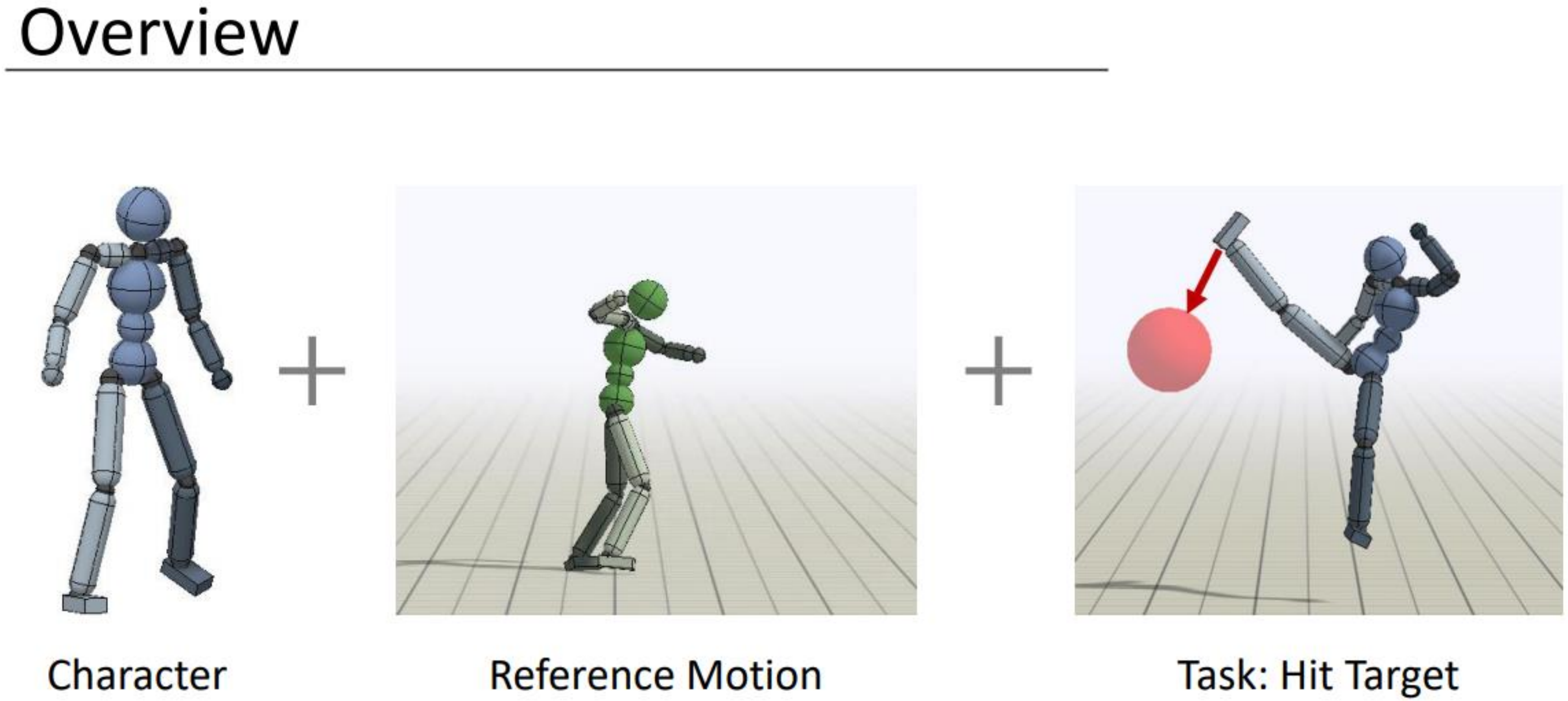
Peng XB, Kanazawa A, Malik J, Abbeel P, Levine S. SFV: Reinforcement learning of physical skills from videos. ACM TOG 2019

Example: Learning skills from videos



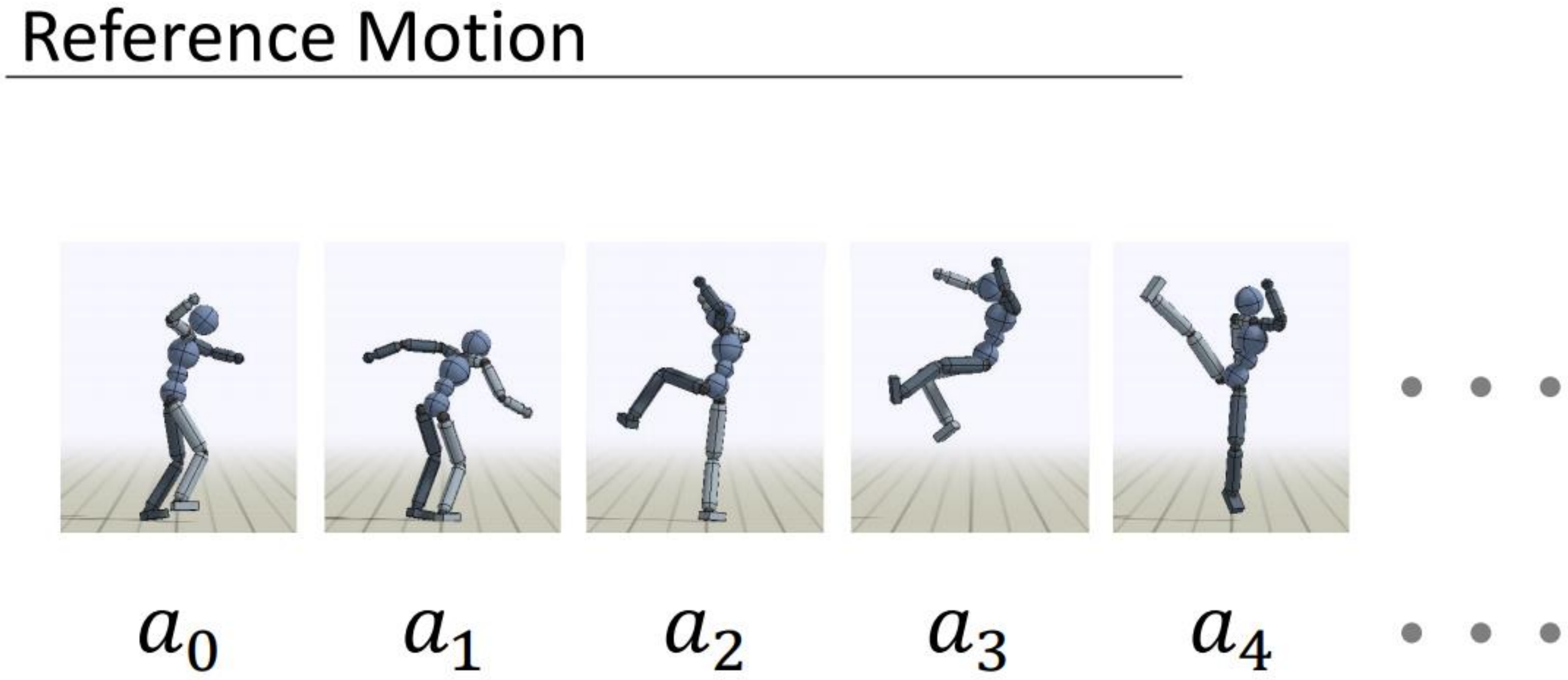
Peng XB, Kanazawa A, Malik J, Abbeel P, Levine S. SFV: Reinforcement learning of physical skills from videos. ACM TOG 2019

Example: Learning skills from videos



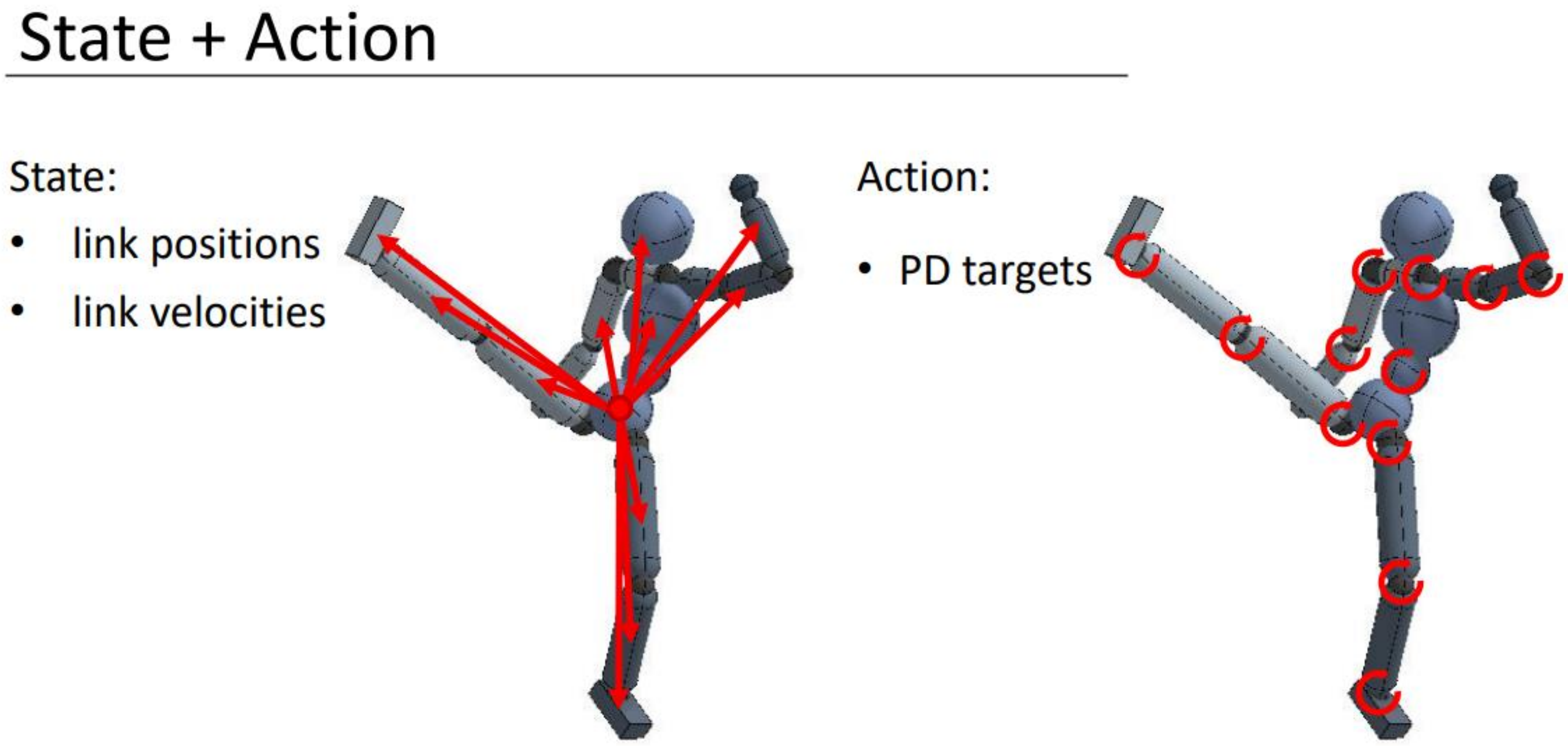
Peng XB, Kanazawa A, Malik J, Abbeel P, Levine S. SFV: Reinforcement learning of physical skills from videos. ACM TOG 2019

Example: Learning skills from videos



Peng XB, Kanazawa A, Malik J, Abbeel P, Levine S. SFV: Reinforcement learning of physical skills from videos. ACM TOG 2019

Example: Learning skills from videos



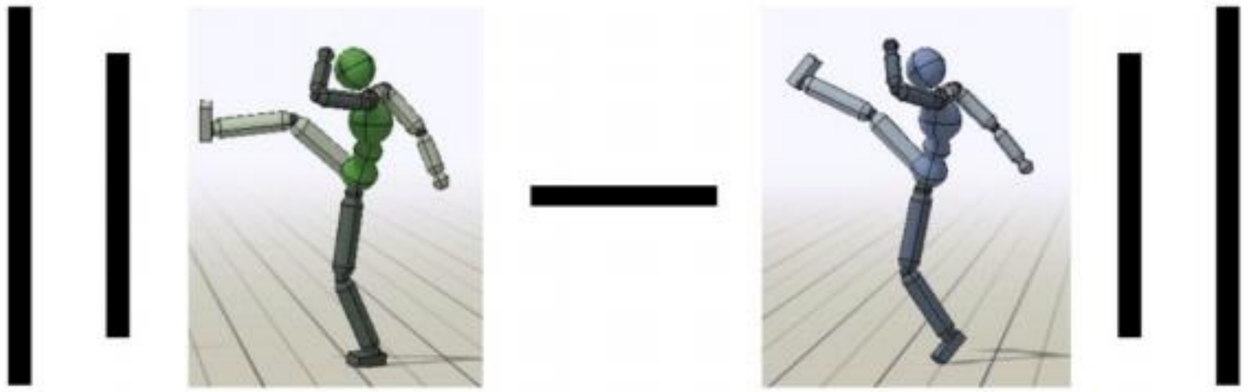
Peng XB, Kanazawa A, Malik J, Abbeel P, Levine S. SFV: Reinforcement learning of physical skills from videos. ACM TOG 2019

Example: Learning skills from videos

Reward

$$r_t = \omega^I r_t^I + \omega^G r_t^G$$

Imitation Objective

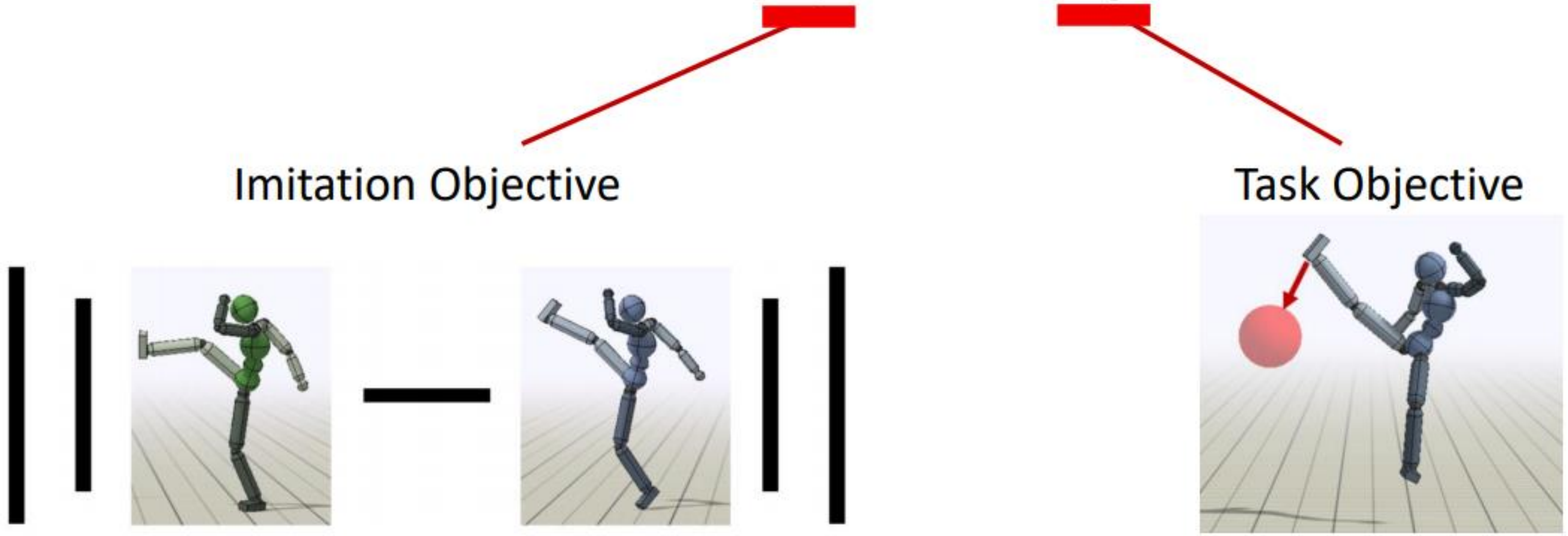


Peng XB, Kanazawa A, Malik J, Abbeel P, Levine S. SFV: Reinforcement learning of physical skills from videos. ACM TOG 2019

Example: Learning skills from videos

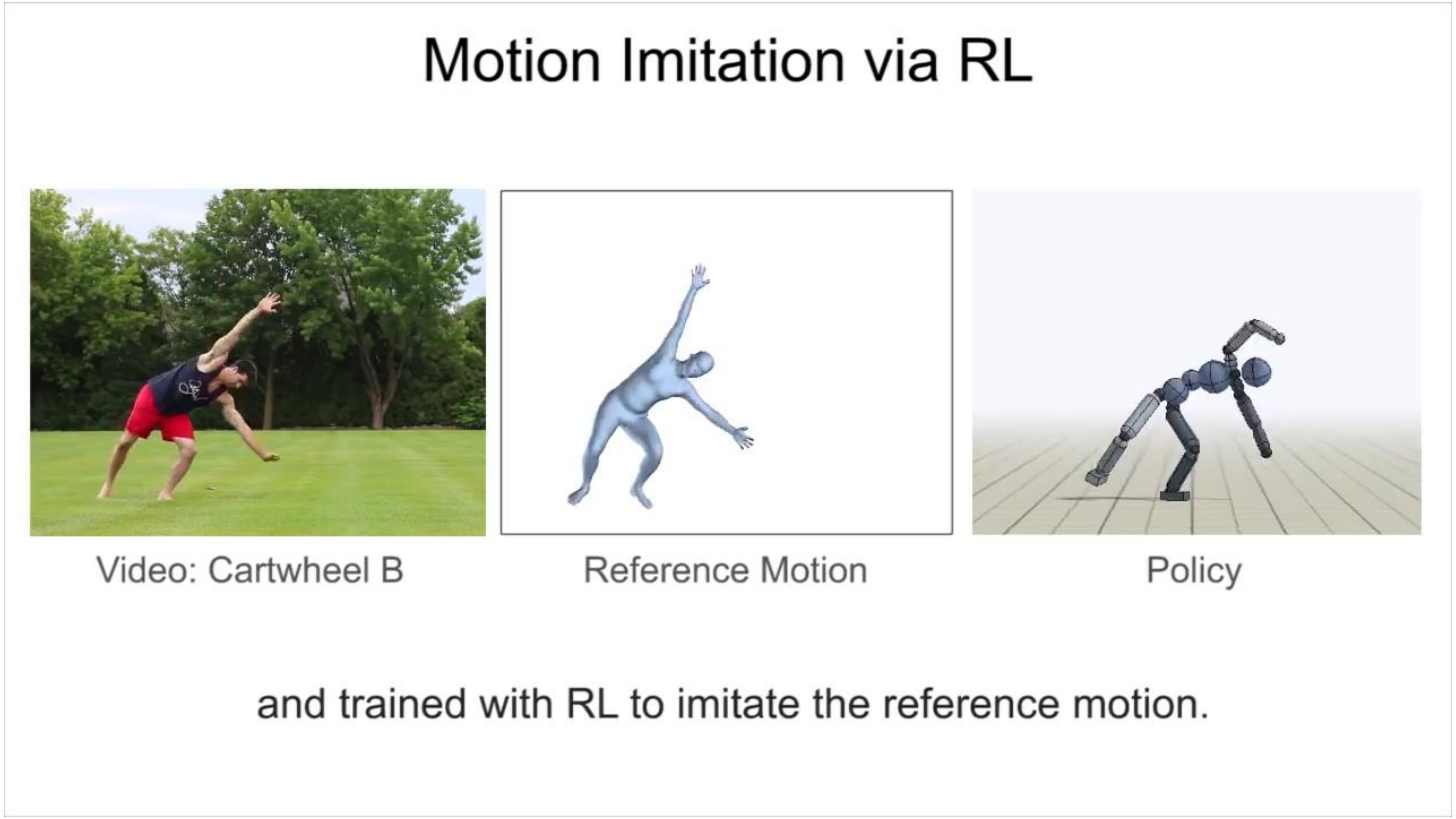
Reward

$$r_t = \omega^I r_t^I + \omega^G r_t^G$$



Peng XB, Kanazawa A, Malik J, Abbeel P, Levine S. SFV: Reinforcement learning of physical skills from videos. ACM TOG 2019

Example: Learning skills from videos



Peng XB, Kanazawa A, Malik J, Abbeel P, Levine S. SFV: Reinforcement learning of physical skills from videos. ACM TOG 2019

ViViDex: Learning Vision-based Dexterous Manipulation from Human Videos

Zerui Chen, Shizhe Chen, Cordelia Schmid, Ivan Laptev

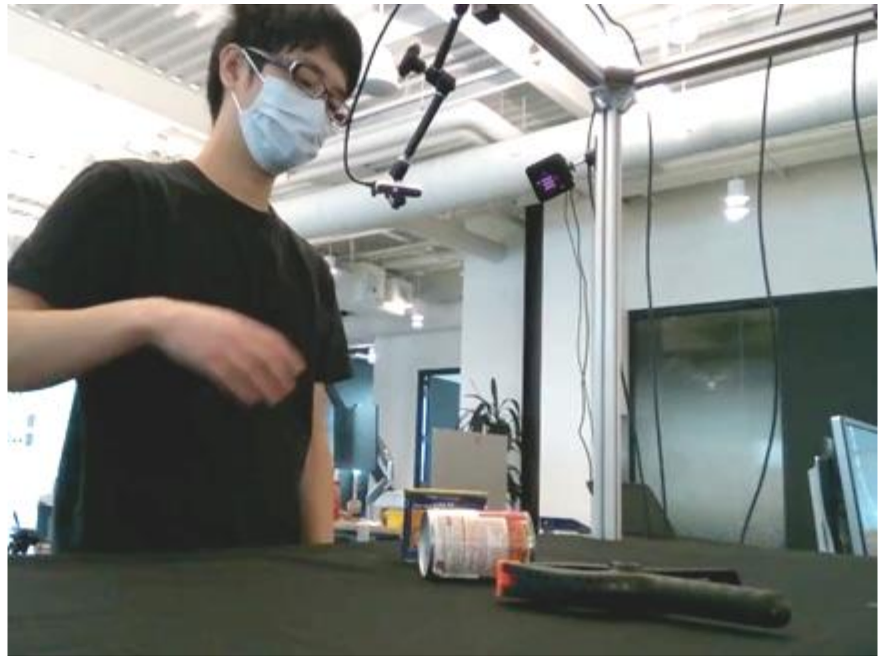
MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Inria

ENS | PSL

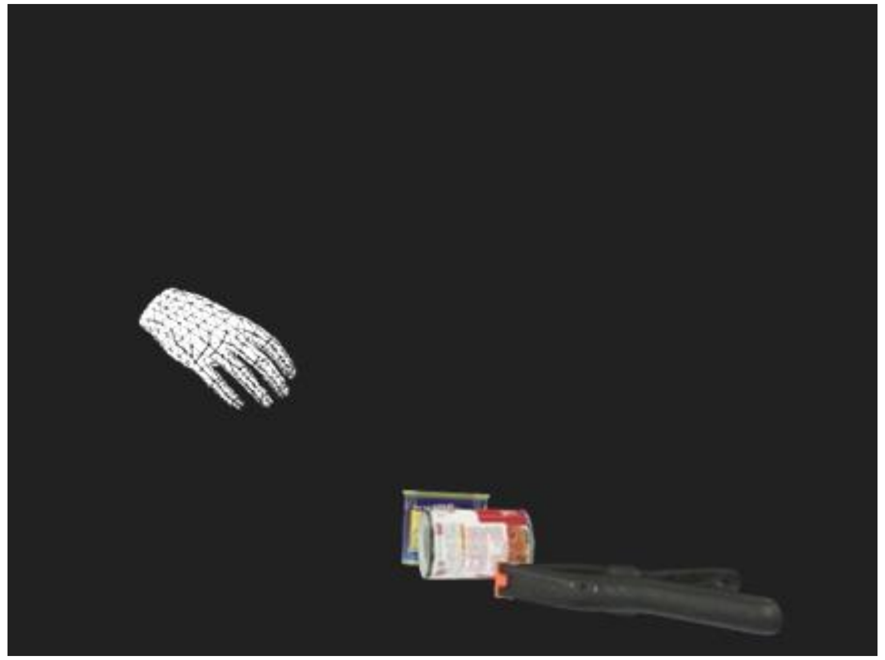
https://zerchen.github.io/projects/vividex.html

Overview of ViViDex



Video Demonstration (DexYCB)

Pose Estimation

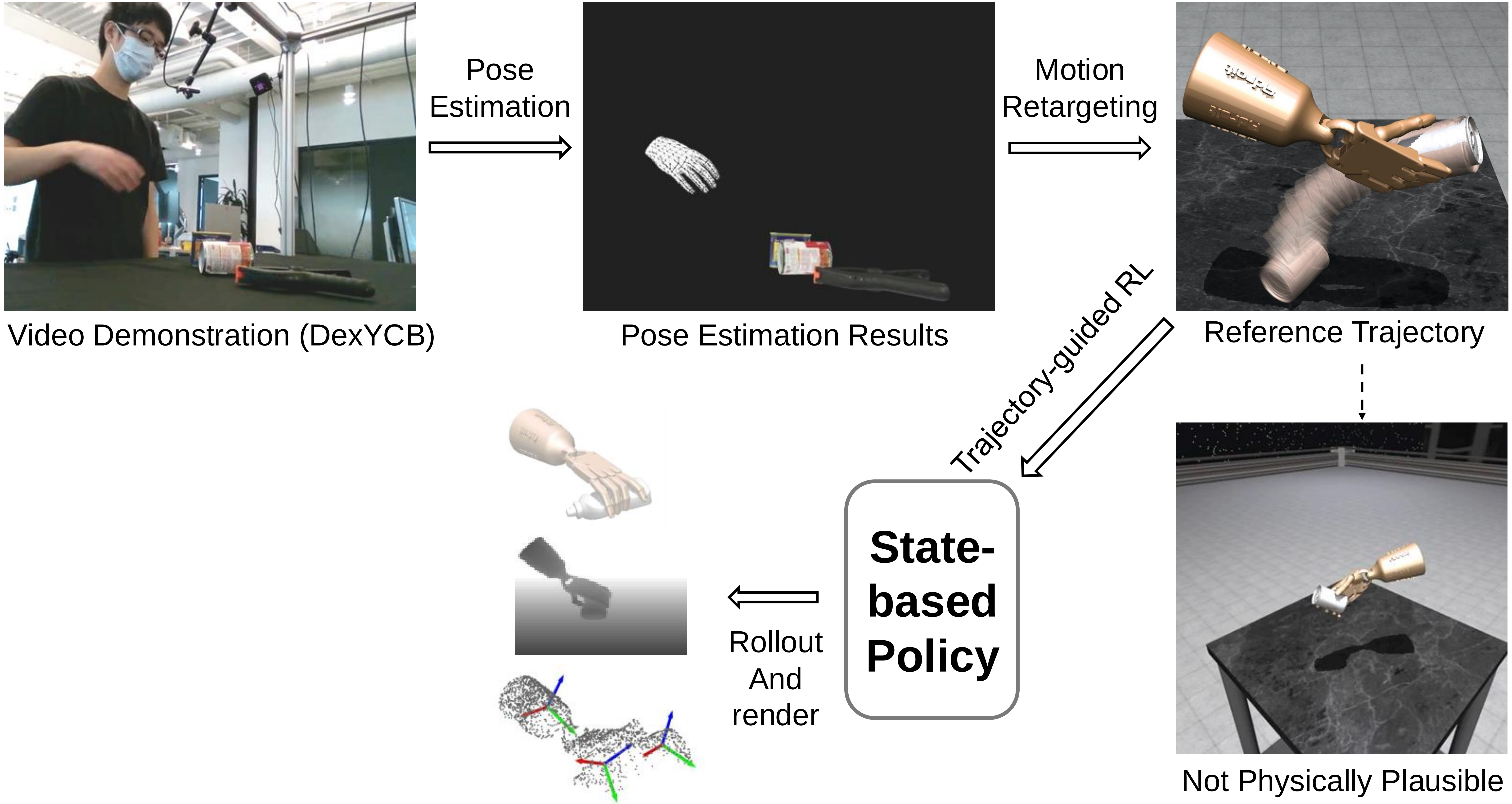


Pose Estimation Results

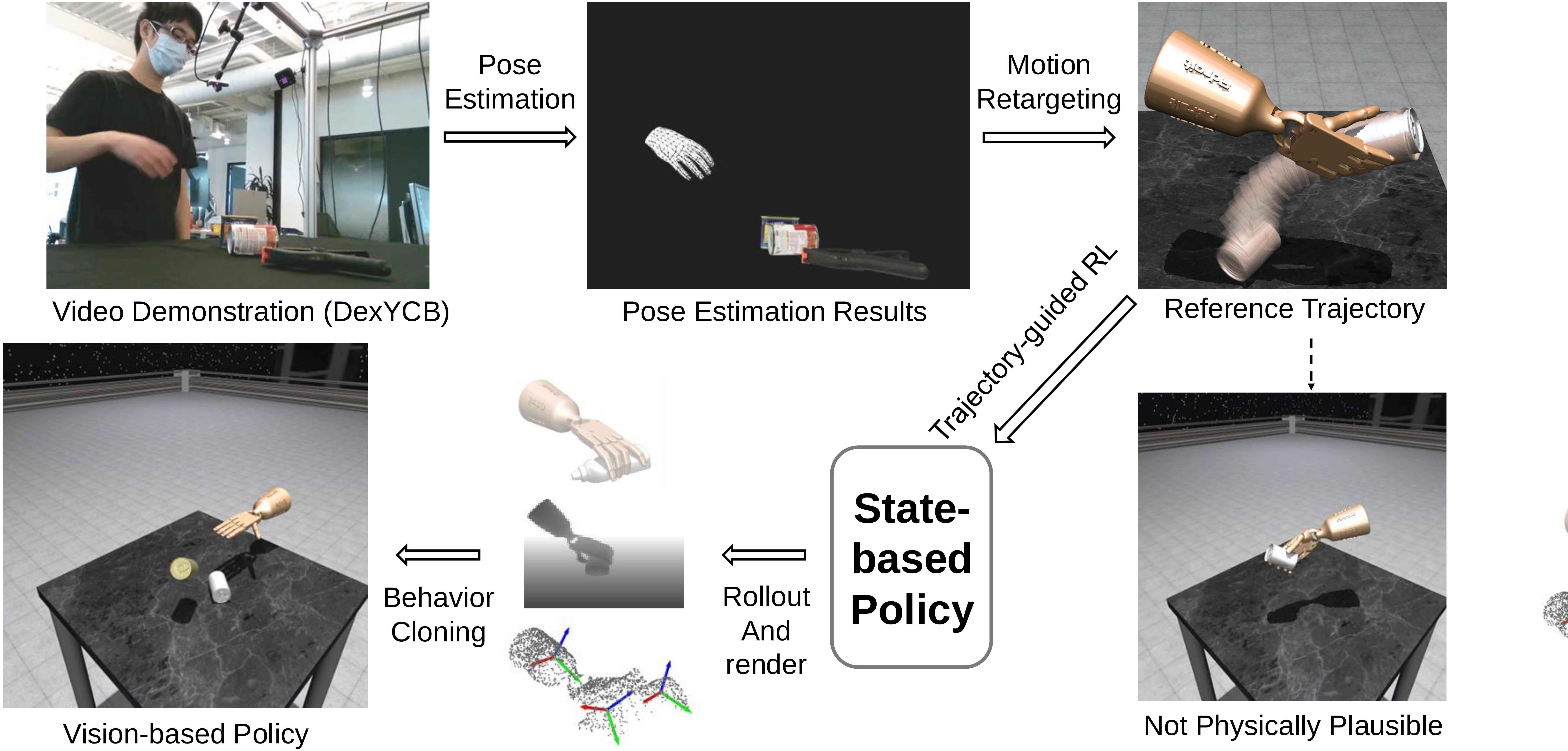
Motion Retargeting

Reference Trajectory

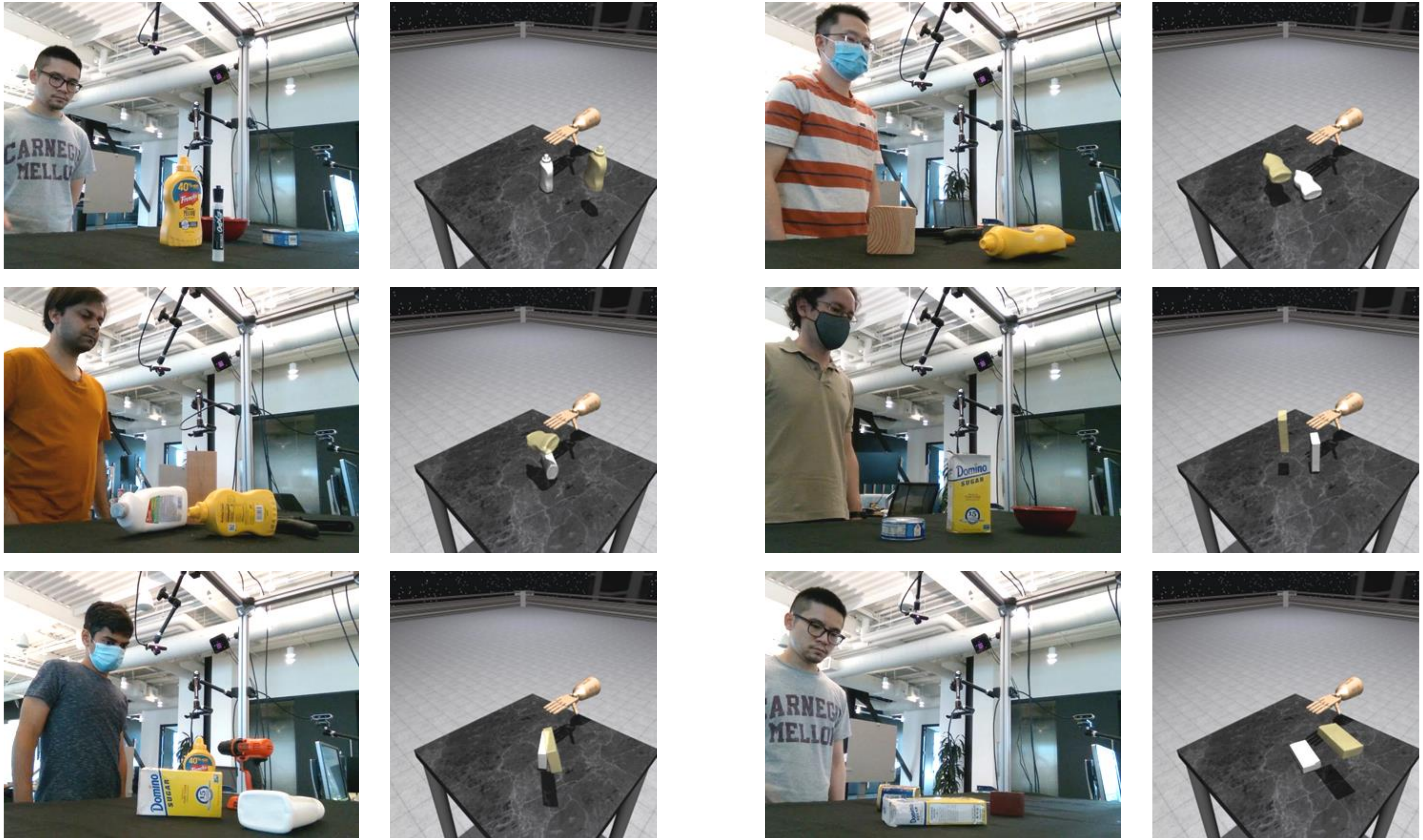
Overview of ViViDex



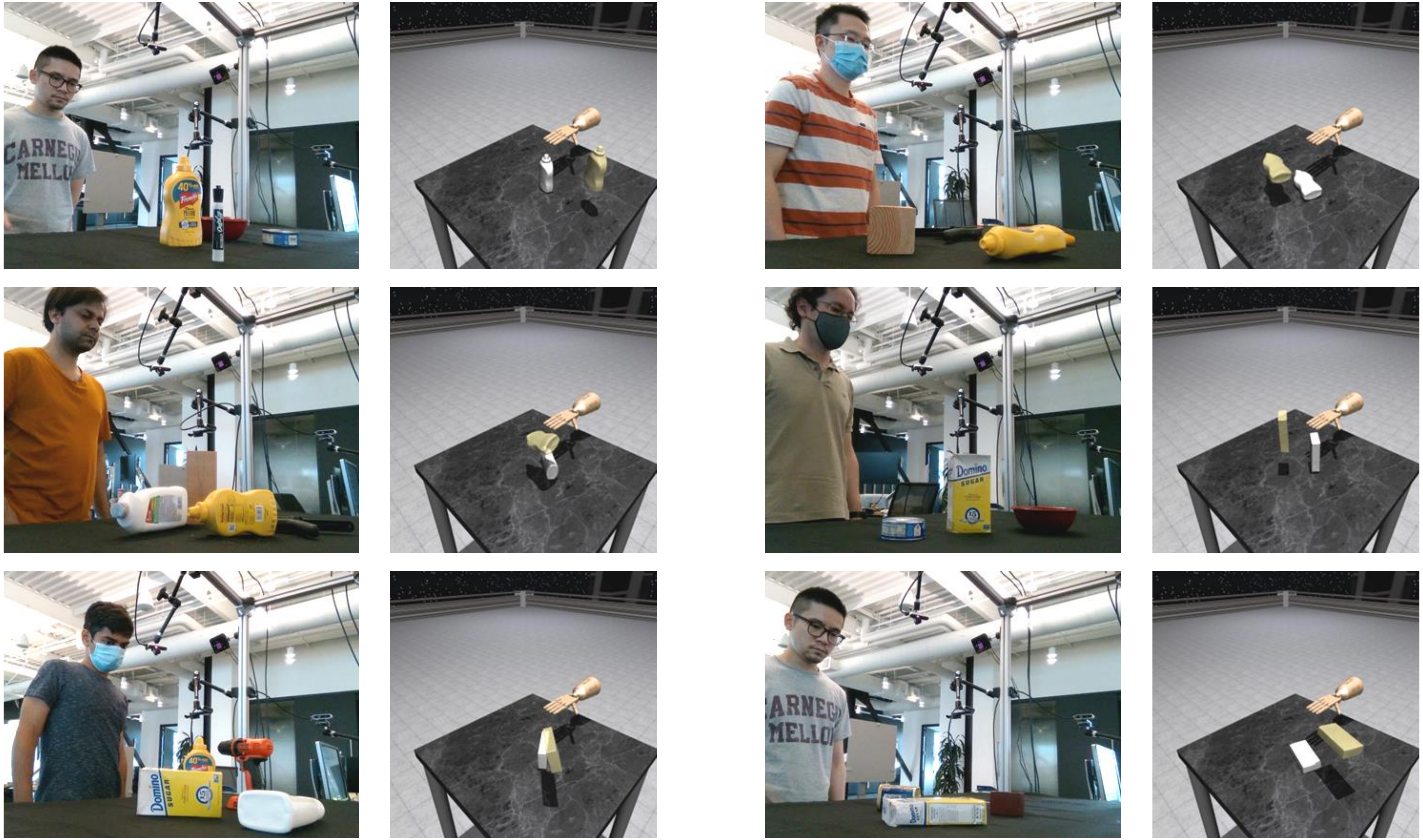
Overview of ViViDex



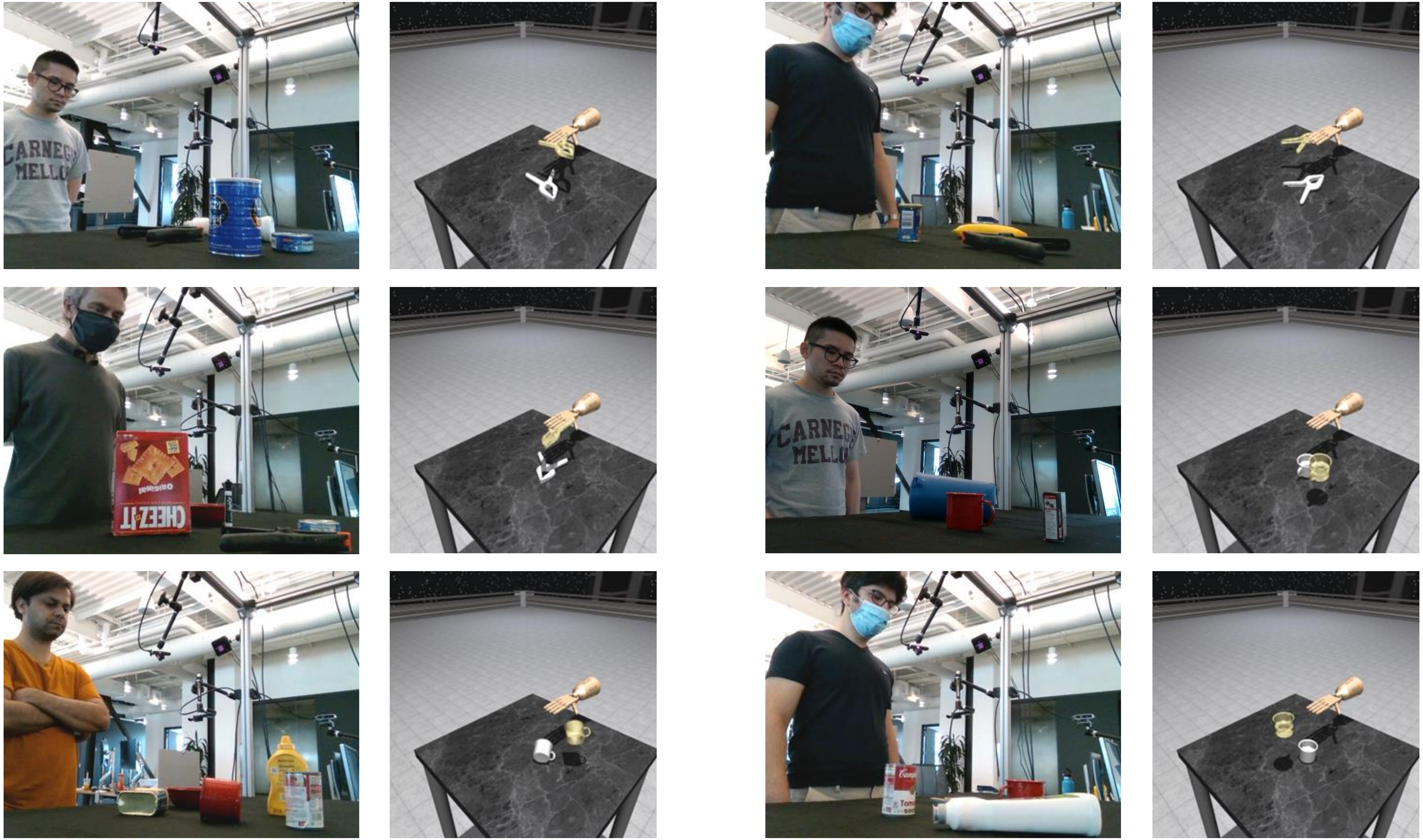
Results for relocation policies



Results for relocation policies

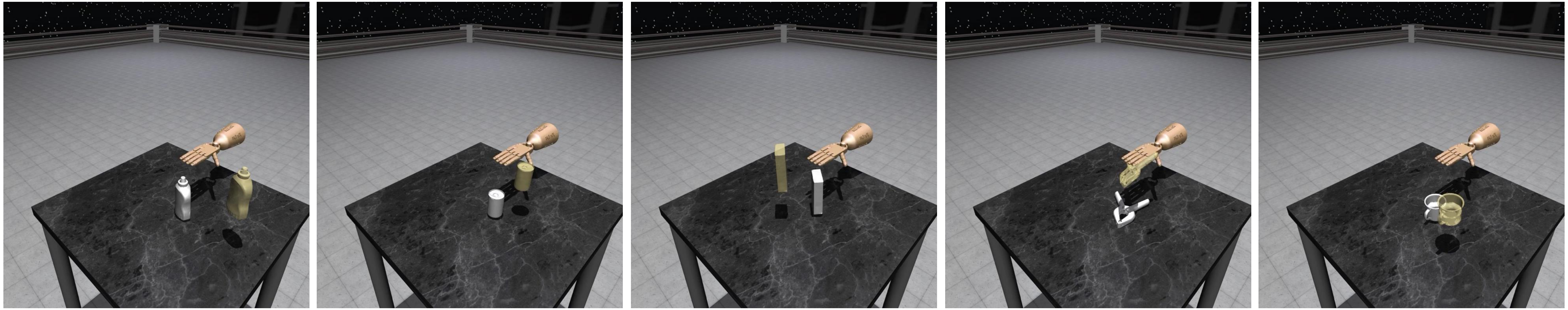


Results for relocation policies

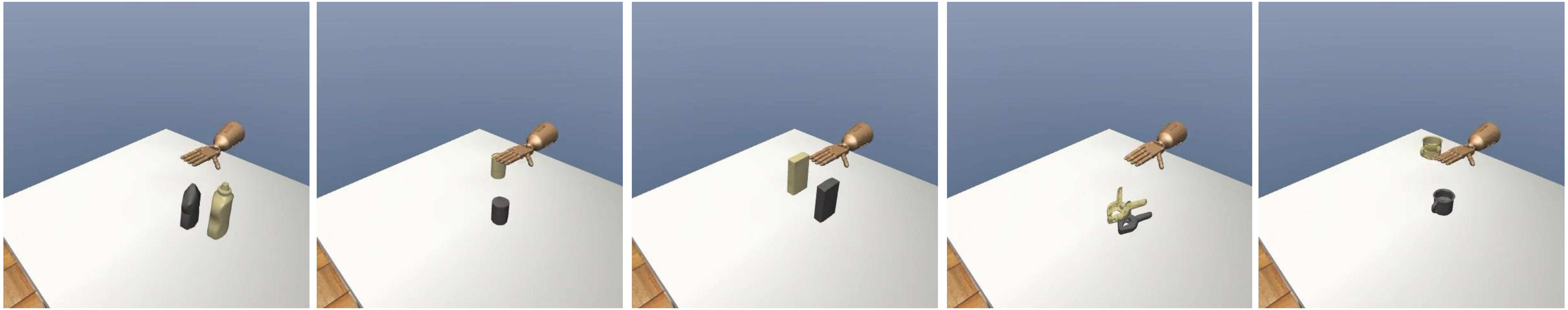


Comparison with the state of the art on the relocation task

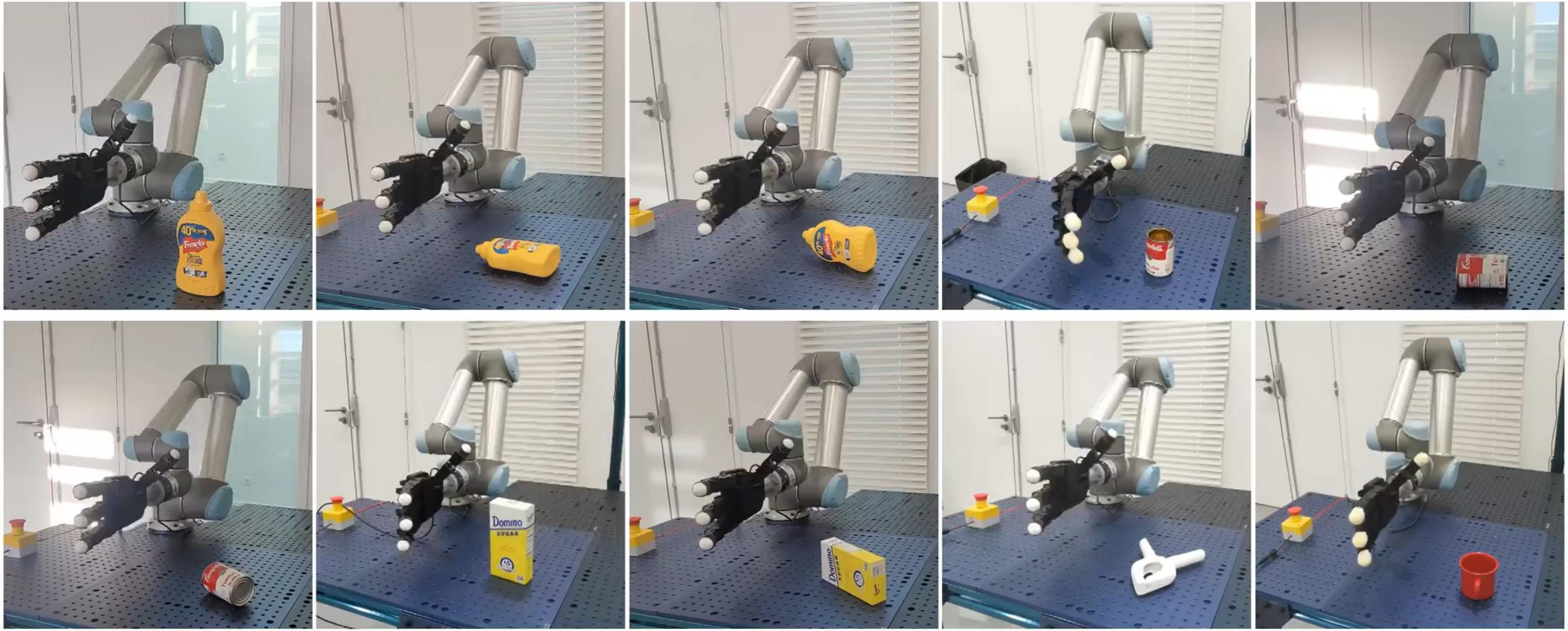
ViViDex visual policies



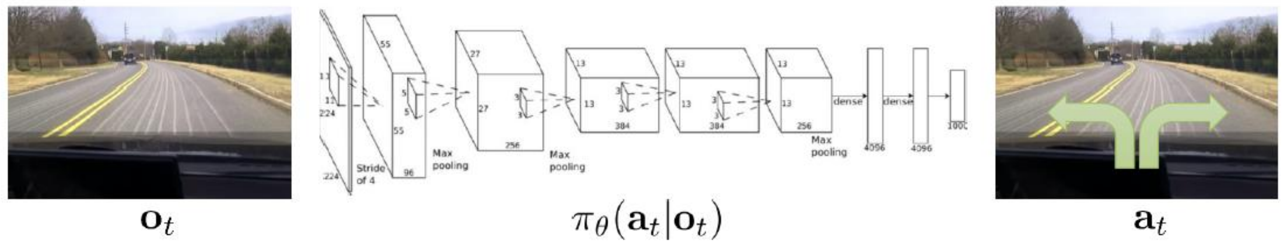
DexMV state-based policies

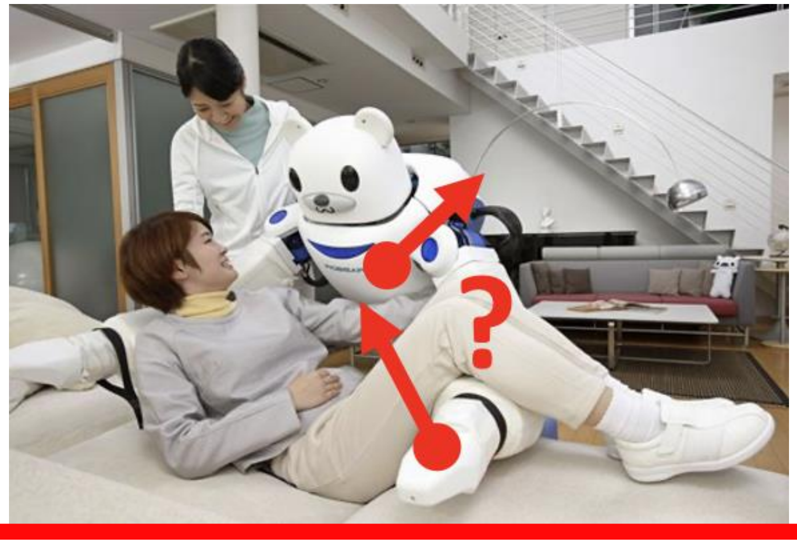


Real experiments with Allegro robot: seen objects

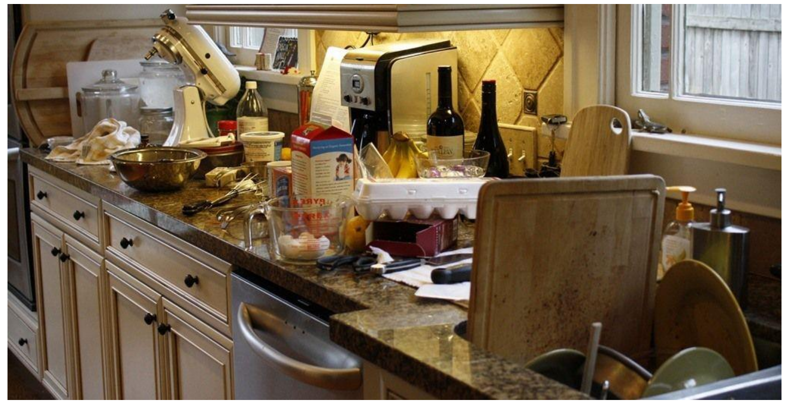


Challenges





- Supervision is costly or not unvaiable



- Large diversity of environments and possible actions

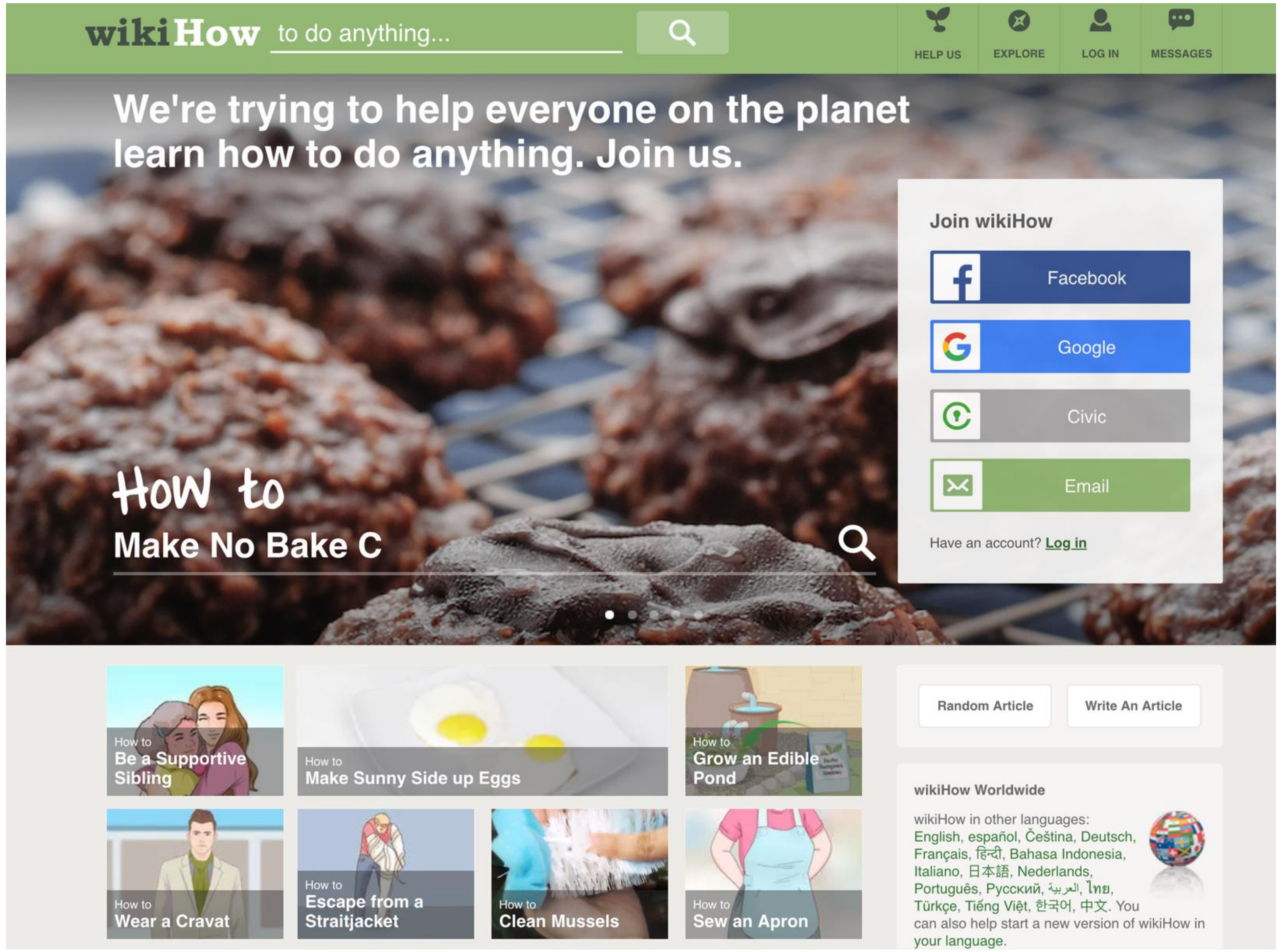
Learn from how-to videos

- Control robots by natural language

Learning from procedural videos

[Alyarac et al., CVPR 2016]

Going WikiHow scale



Step 1: Scrap ~130K tasks from WikiHow

Examples of scrapped tasks

- ~~How to Be Healthy~~
- How to Cook Quinoa in a Rice Cooker
- How to Sew an Apron
- How to Break a Chain
- ~~How to April Fool your Girlfriend~~
- ....

Step 2: Filter out non-visual tasks

HowTo100M dataset

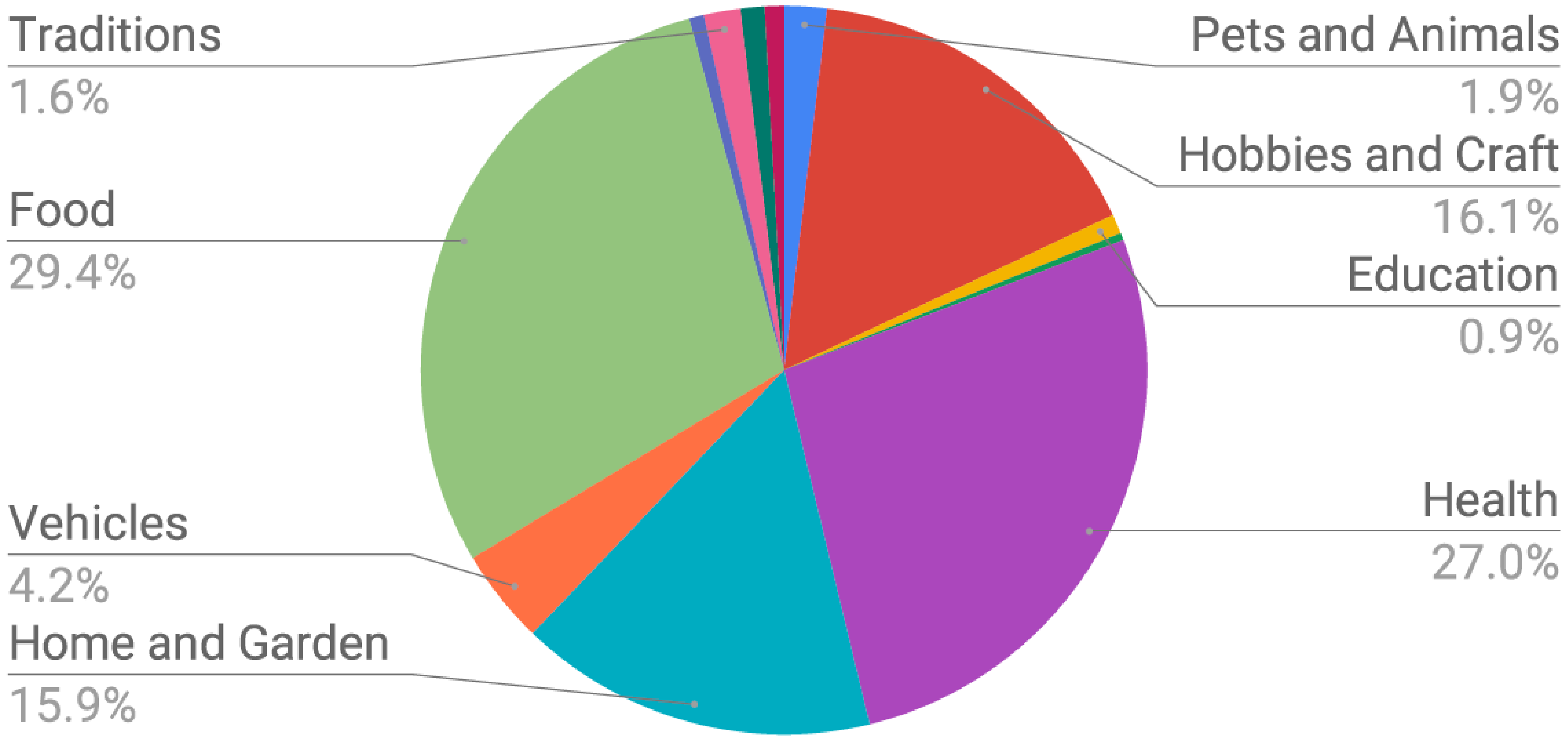
23 k
wikiHow
visual tasks

1.2 M
YouTube
videos

136 M
Short
clips

ASR
Supervision
*ASR: Automated Speech Recognition

0
Manual
annotation



[Miech, Zhukov, Alayrac, Tapaswi, Laptev and Sivic, ICCV 2019]

HowTo100M dataset: Examples

[Miech, Zhukov, Alayrac, Tapaswi, Laptev and Sivic, ICCV 2019]

Video description datasets

Dataset	Clips	Captions	Videos	Duration	Source	Year
Charades [42]	10k	16k	10,000	82h	Home	2016
MSR-VTT [52]	10k	200k	7,180	40h	Youtube	2016
YouCook2 [61]	14k	14k	2,000	176h	Youtube	2018
EPIC-KITCHENS [5]	40k	40k	432	55h	Home	2018
DiDeMo [11]	27k	41k	10,464	87h	Flickr	2017
M-VAD [46]	49k	56k	92	84h	Movies	2015
MPII-MD [37]	69k	68k	94	41h	Movies	2015
ANet Captions [22]	100k	100k	20,000	849h	Youtube	2017
TGIF [23]	102k	126k	102,068	103h	Tumblr	2016
LSMDC [38]	128k	128k	200	150h	Movies	2017
How2 [39]	185k	185k	13,168	298h	Youtube	2018
HowTo100M	**136M**	**136M**	**1.221M**	**134,472h**	Youtube	2019

[Miech, Zhukov, Alayrac, Tapaswi, Laptev and Sivic, ICCV 2019]

Some of our work in this domain

- **Learning from Narrated Instruction Videos**,
  J.-B. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev and S. Lacoste-Julien; *In CVPR'16, PAMI 2017*
- **Joint Discovery of Object States and Manipulation Actions**,
  J.-B. Alayrac, J. Sivic, I. Laptev and S. Lacoste-Julien.; *In Proc. ICCV'17*
- **Cross-task weakly supervised learning from instructional video**,
  D. Zhukov, J.-B. Alayrac, R.G. Cinbis, D. Fouhey, I. Laptev and J. Sivic; *in Proc. CVPR'19*
- **HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips**,
  A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev and J. Sivic; *In Proc. ICCV'19*
- **End-to-End Learning of Visual Representations from Uncurated Instructional, Videos**, A. Miech*, J.-B. Alayrac*, L. Smaira, I. Laptev, J. Sivic and A. Zisserman; *In Proc. CVPR'20*
- **Look for the Change: Learning Object States and State-Modifying Actions from Untrimmed Web Videos**,
  T. Souček, J.-B. Alayrac, A. Miech, I. Laptev and J. Sivic; *In Proc CVPR'22, PAMI'24*
- **GenHowTo: Learning to Generate Actions and State Transformations from Instructional Videos**,
  Tomáš Souček, Dima Damen, Michael Wray, Ivan Laptev, Josef Sivic, *In proc CVPR'24*

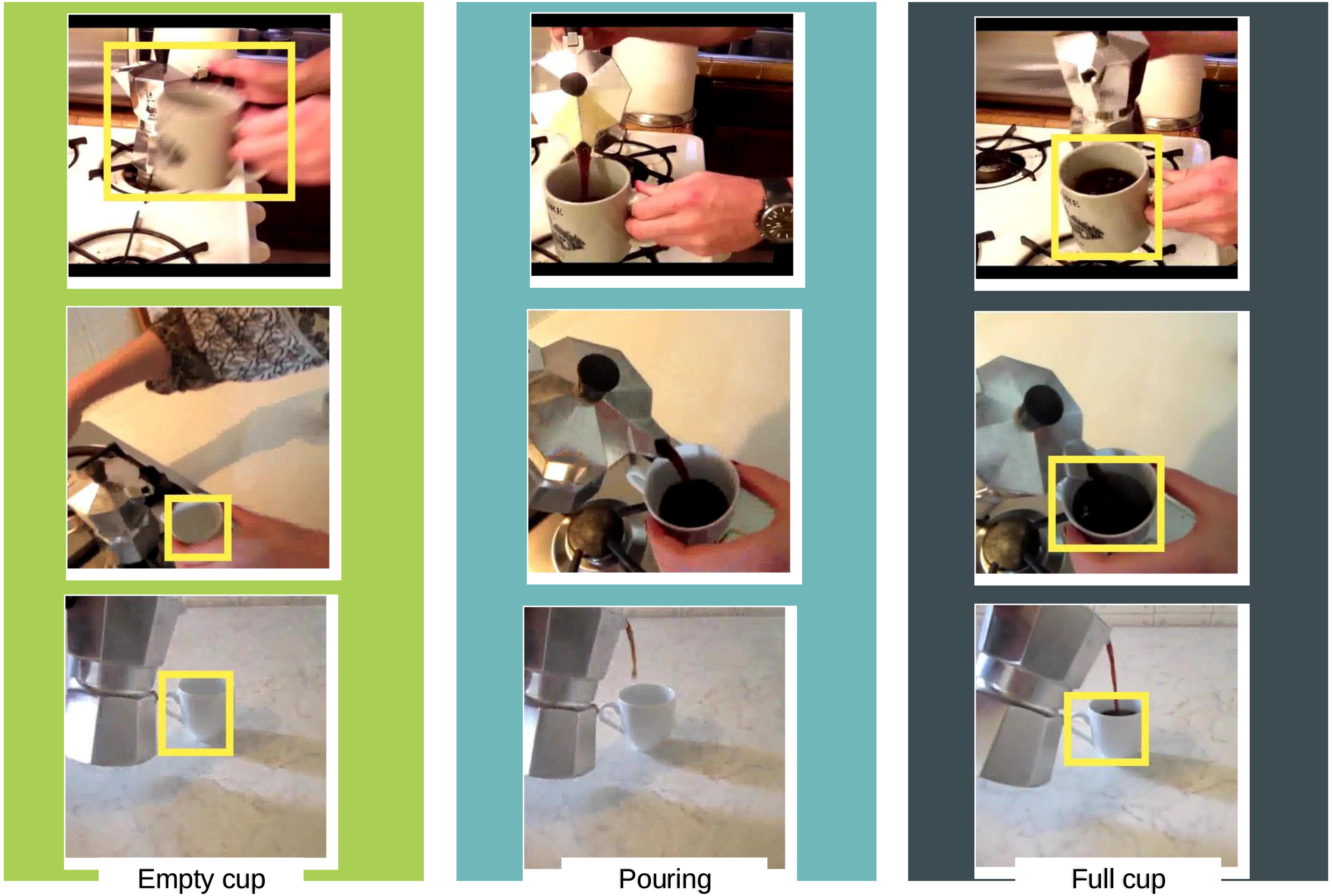
Recognition

Actions and state changes

Generation

Learn how actions change states of objects

Pour coffee

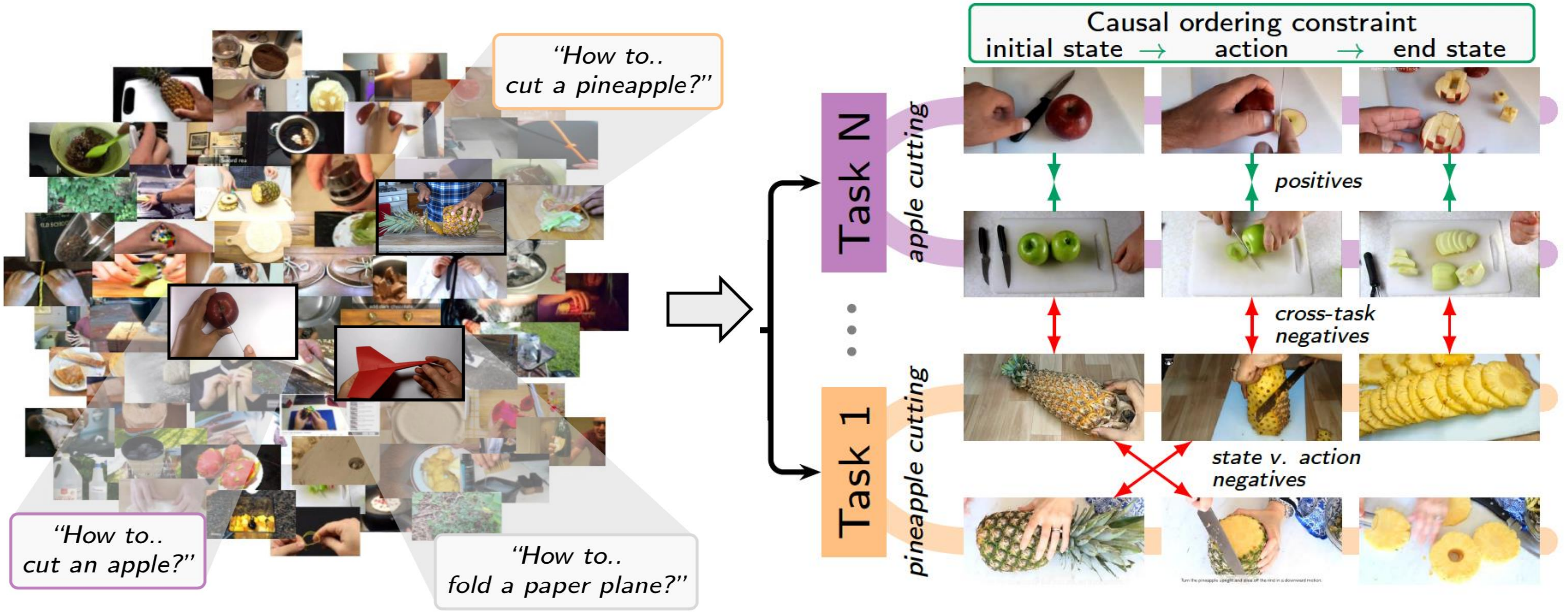


[Alayrac et al., ICCV 2017]

Goal #1: Learn to localize object state changes

Input: videos with noisy video-level labels

Output: temporal localization of object states and state-modifying actions

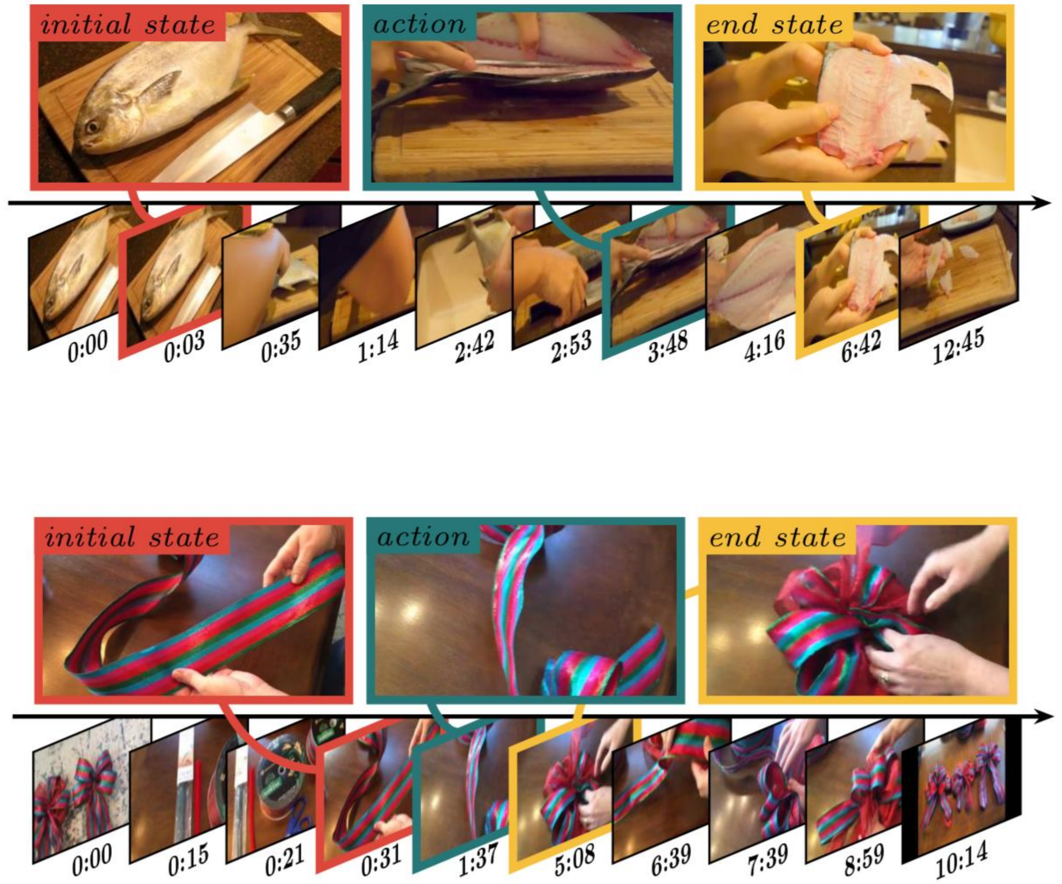


[**Tomas Soucek**, Jean-Baptiste Alayrac, Antoine Miech, Ivan Laptev, and Josef Sivic
Multi-Task Learning of Object States and State-Modifying Actions from Web Videos, CVPR 2022, PAMI 2024]

Challenges

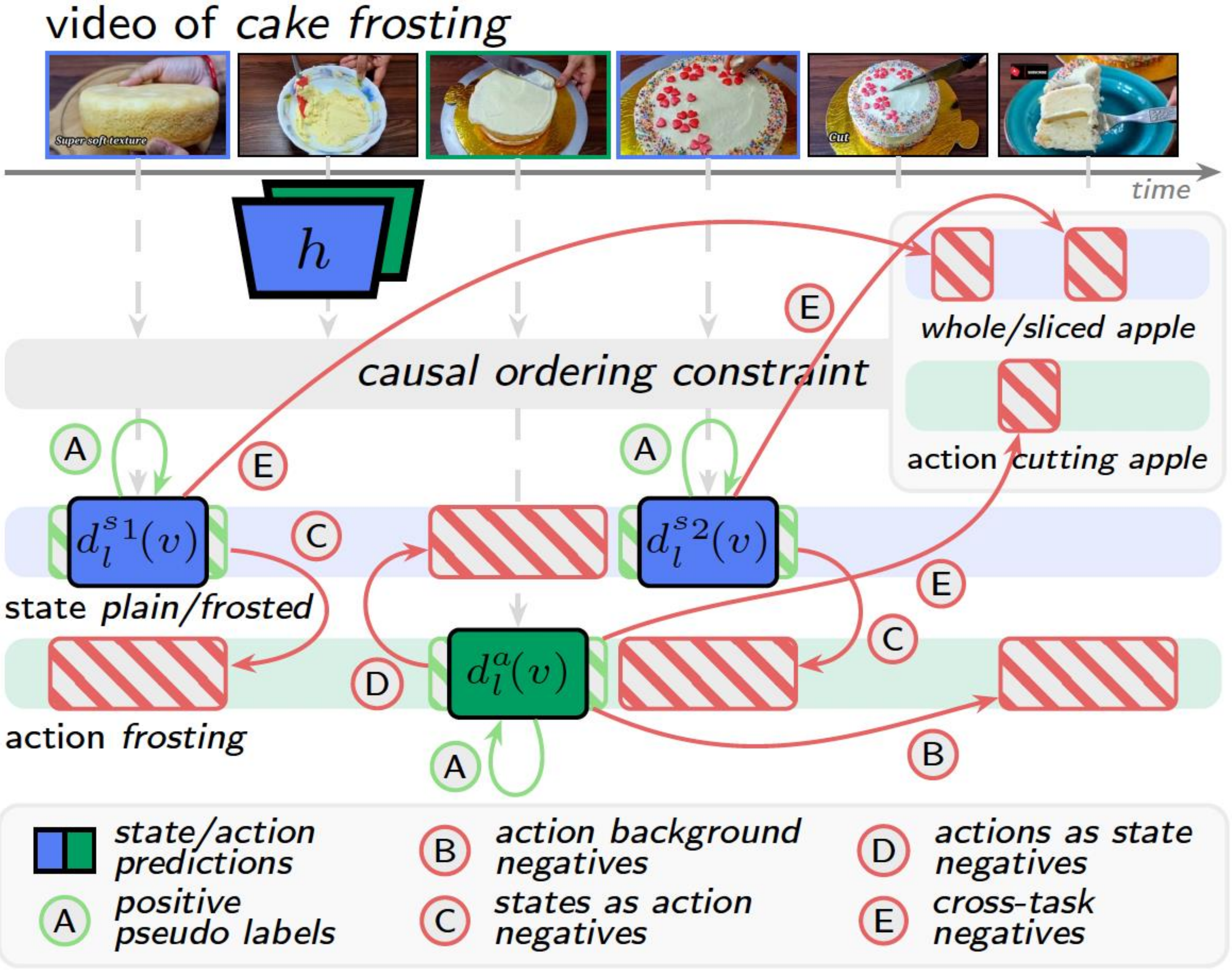
Visual variability

Long videos

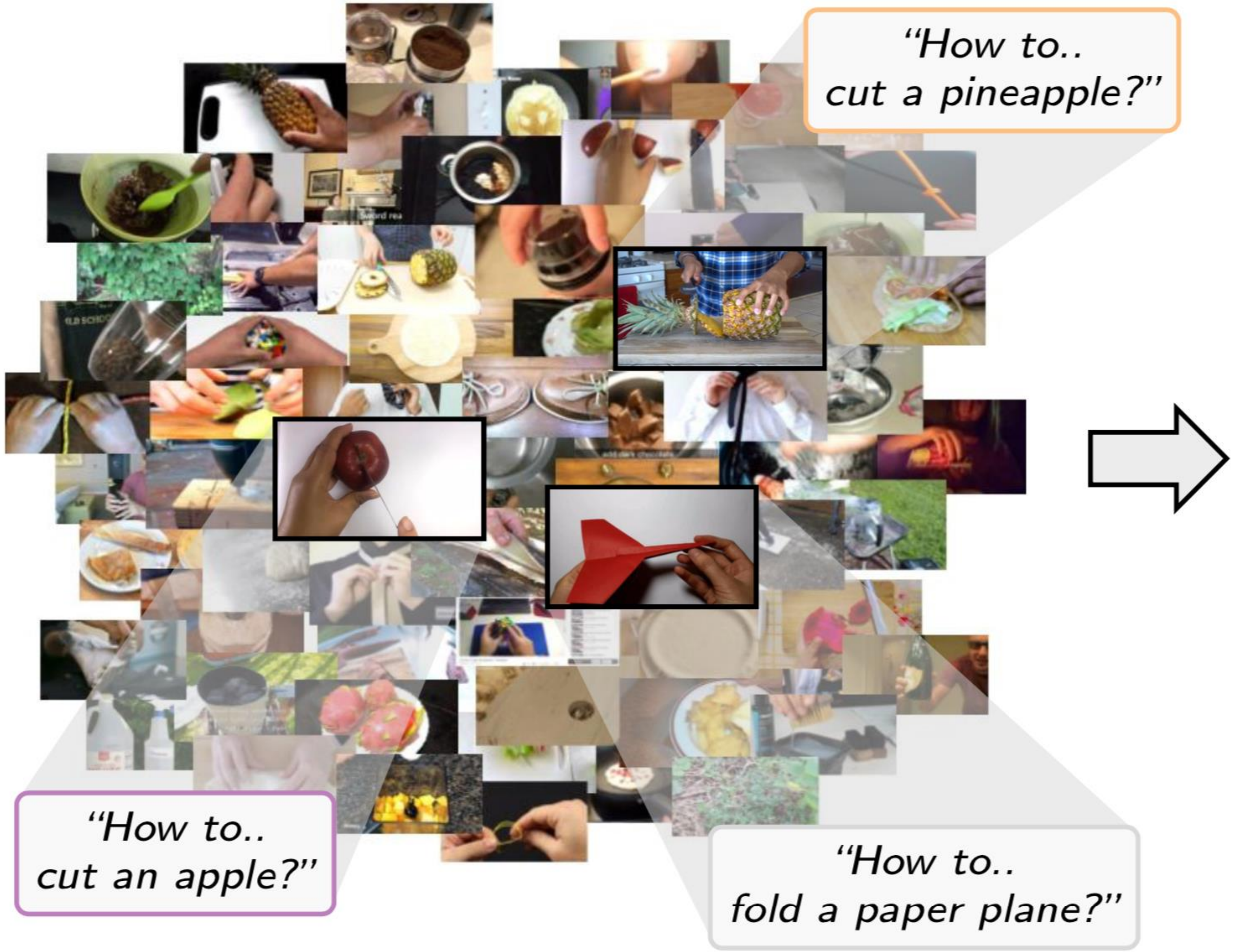


In-the-wild, uncurated, noisy data

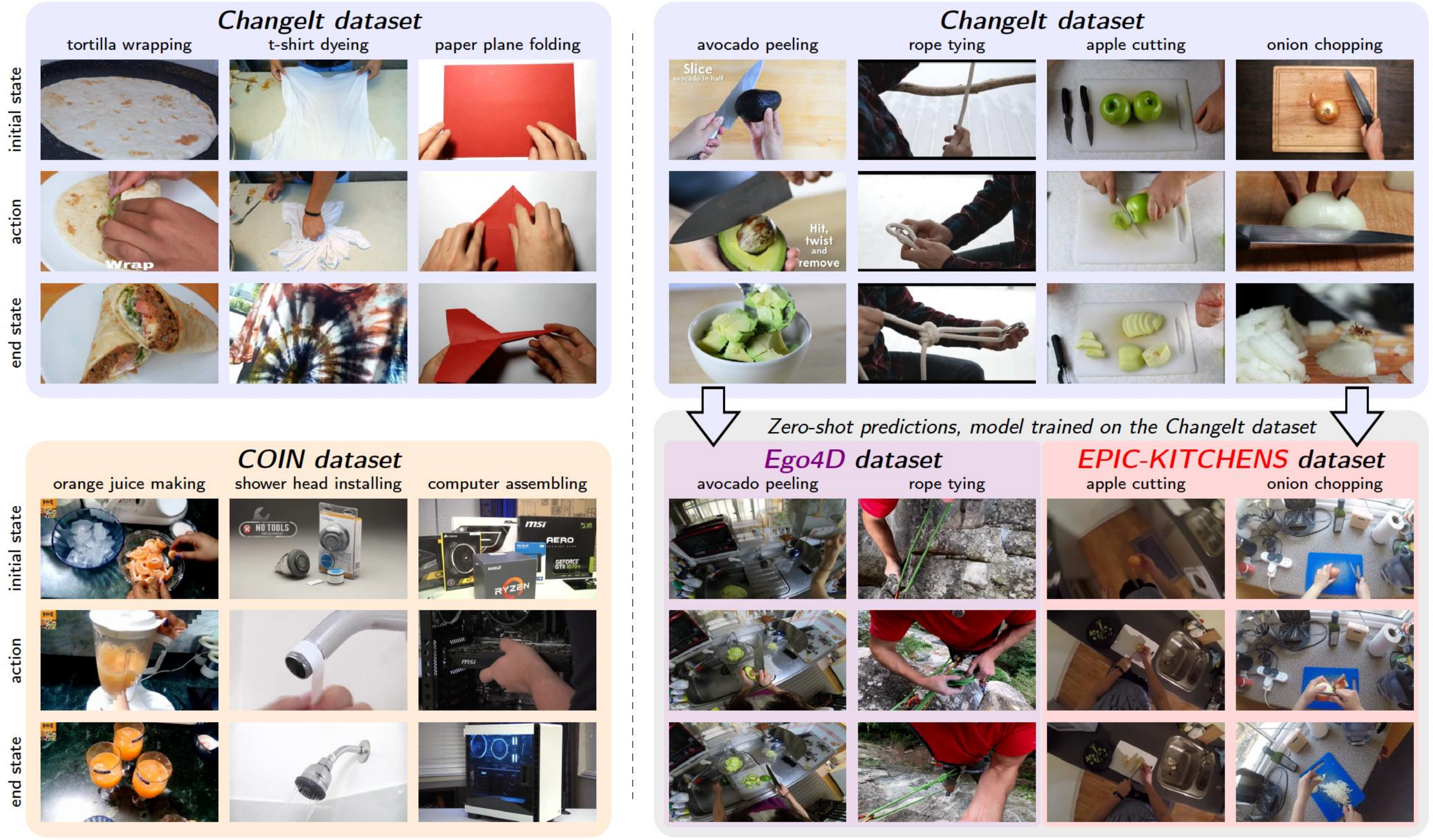
Contribution 1: Constraints for self-supervised learning



ChangeIt dataset



- **44 interactions** such as "How to cut an apple?"
- **34,000+** videos, **2600+** hours
- Up to **15mins** long, **4.6mins** on average
- Auto-annotated with the **noisy video-level** category label
- **667** videos manually annotated with **temporal labels**.


ChangeIt dataset
tortilla wrapping
t-shirt dyeing
paper plane folding
initial state
action
end state
Wrap
ChangeIt dataset
avocado peeling
rope tying
apple cutting
onion chopping
Slice
avocado in half
Hit,
twist
and
remove
COIN dataset
orange juice making
shower head installing
computer assembling
initial state
action
end state
NO TOOLS
msi
AERO
GEFORCE GTX 1070
RYZEN
Zero-shot predictions, model trained on the ChangeIt dataset
Ego4D dataset
avocado peeling
rope tying
EPIC-KITCHENS dataset
apple cutting
onion chopping

Look for the Change: Learning Object States and State-Modifying Actions from Untrimmed Web Videos

T. Souček J.B. Alayrac A. Miech I. Laptev J. Sivic

CVPR 2022

Look for the Change: Learning Object States and State-Modifying Actions from Untrimmed Web Videos

T. Souček J.B. Alayrac A. Miech I. Laptev J. Sivic

CVPR 2022

Goal #2: Generate changes of object states

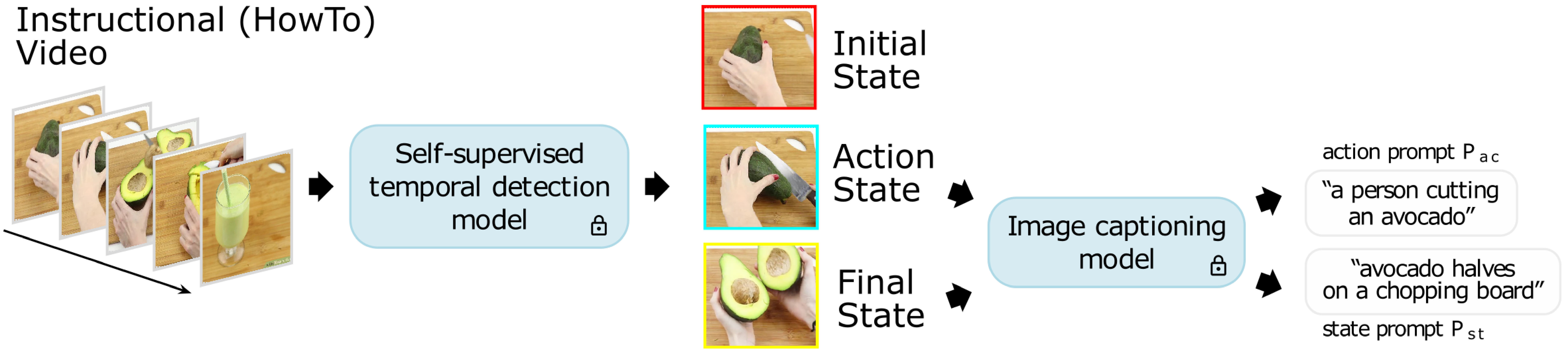
[**Tomas Soucek**, Dima Damen, Michael Wray, Ivan Laptev and Josef Sivic
GenHowTo: Learning to Generate Actions and State Transformations from Instructional Videos, CVPR 2024]

Challenges:

1. Change the object

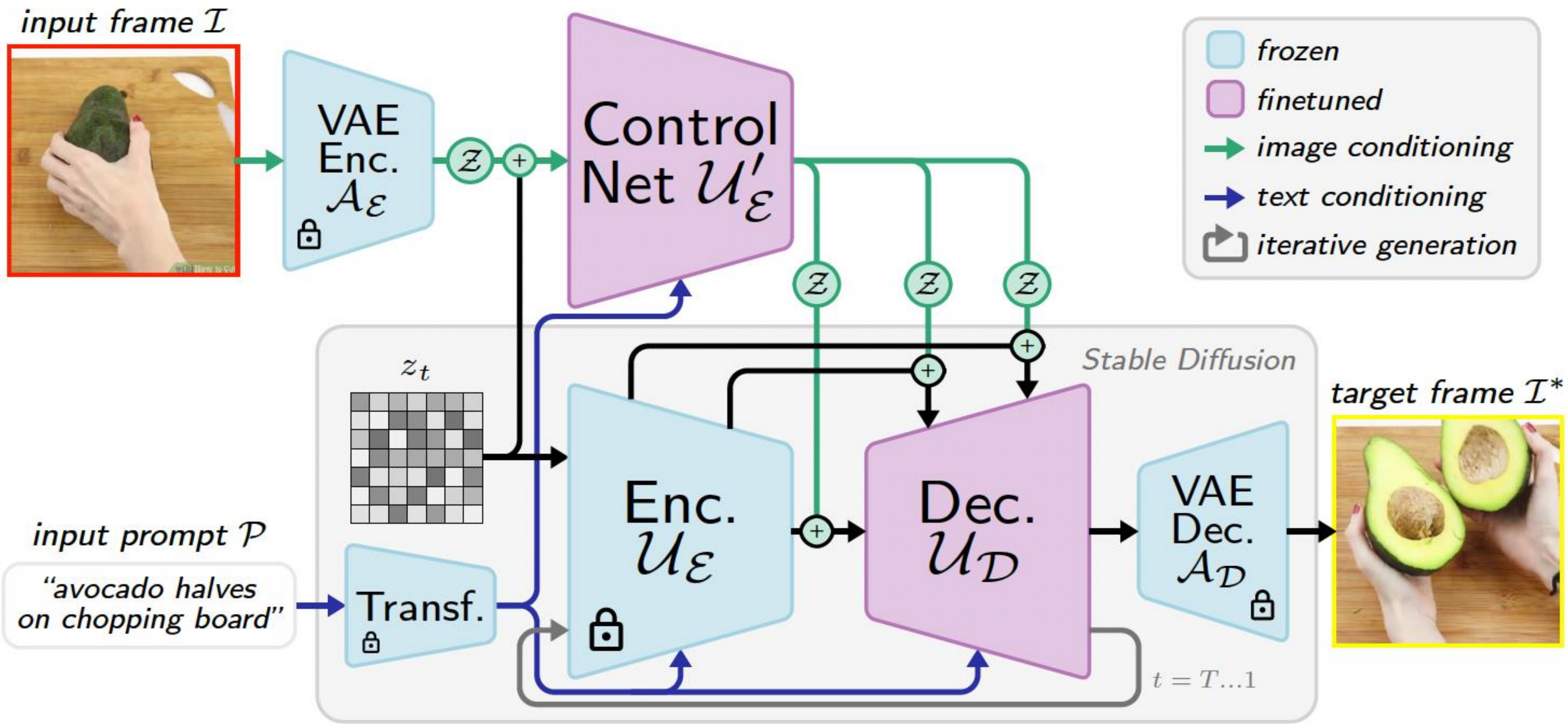
2. Keep the scene context

Contribution 1: Dataset of annotated image triplets



[Tomas Soucek, Jean-Baptiste Alayrac, Antoine Miech, Ivan Laptev, and Josef Sivic. Multi-task learning of object state changes from uncurated videos, PAMI 2024.]

Contribution 2: Method



Contribution 2: Method

Preserves the scene while changing the object state

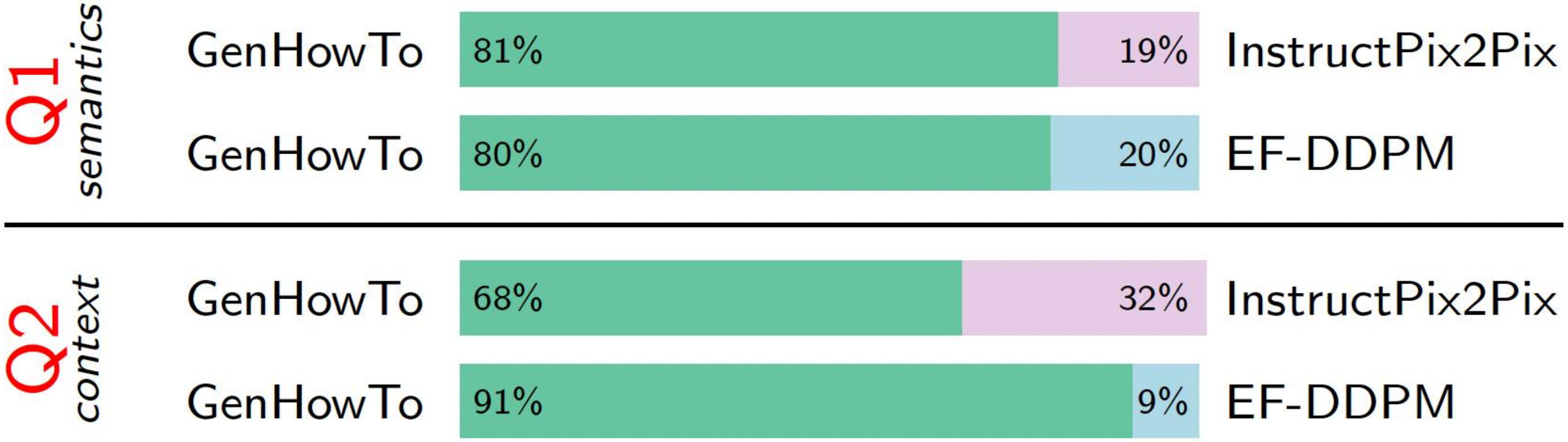
Experiments: quantitative evaluation

	Method	$\text{Acc}_{\text{ac}} \uparrow$	$\text{Acc}_{\text{st}} \uparrow$
	test set categories unseen during training		
(a)	Stable Diffusion	0.51	0.50
(b)	Edit Friendly DDPM	0.60	0.61
(c)	InstructPix2Pix	0.55	0.63
(d)	*CLIP (manual prompts)*	0.52	0.62
(e)	**GenHowTo**	**0.66**	**0.74**
	test set categories seen during training		
(f)	Edit Friendly DDPM†	0.69	0.80
(g)	**GenHowTo**†	**0.77**	**0.88**
(h)	*Real images*	0.96	0.97

† Models trained also on the test set *categories*.

Experiments: user study

Q1: “Which image better represents the final state described as <input prompt> of the same object as in the first image?”.



Q2: “Which image better preserves the consistency of the scene?” to verify how well the methods preserve the background.

Experiments: qualitative results

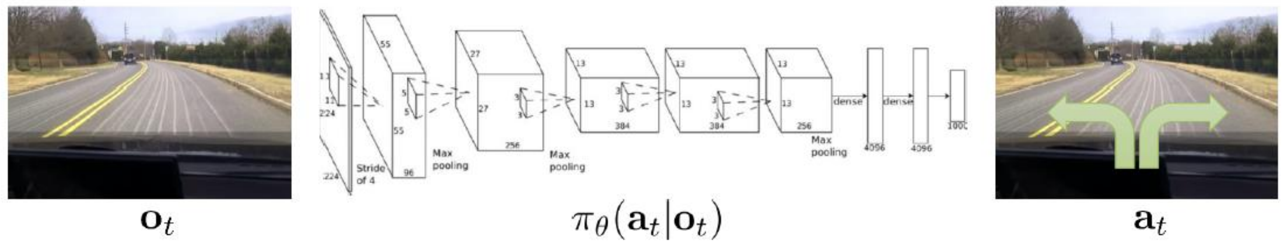
Generated action

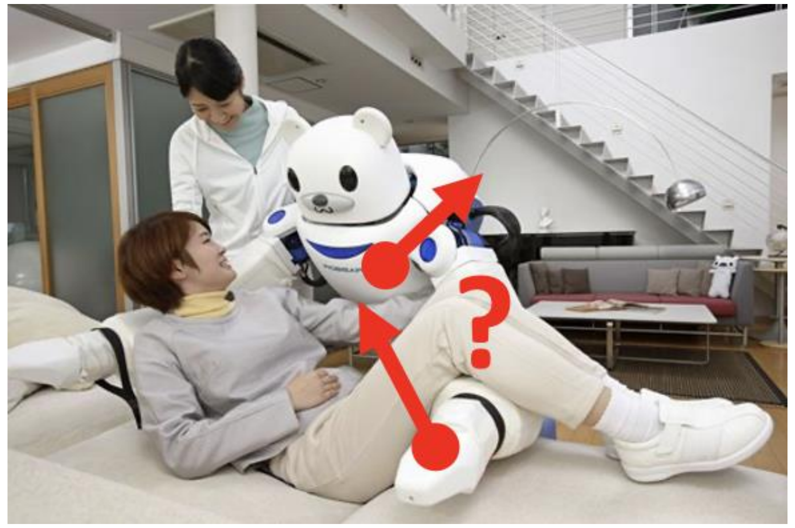
Generated object state

Generated action

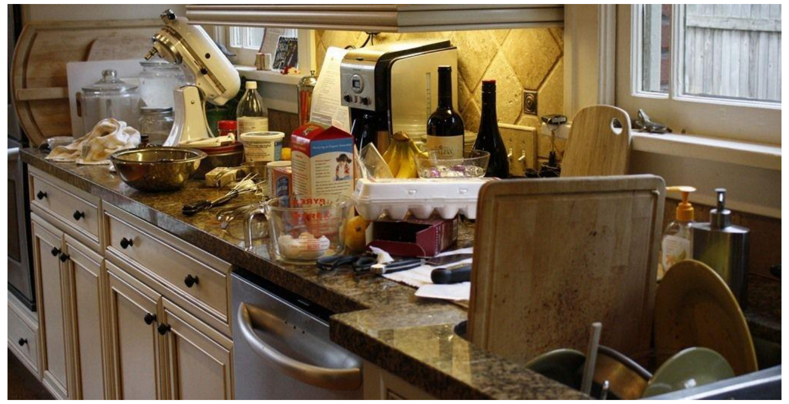
Generated object state

Challenges





- Supervision is costly or not unvaiable



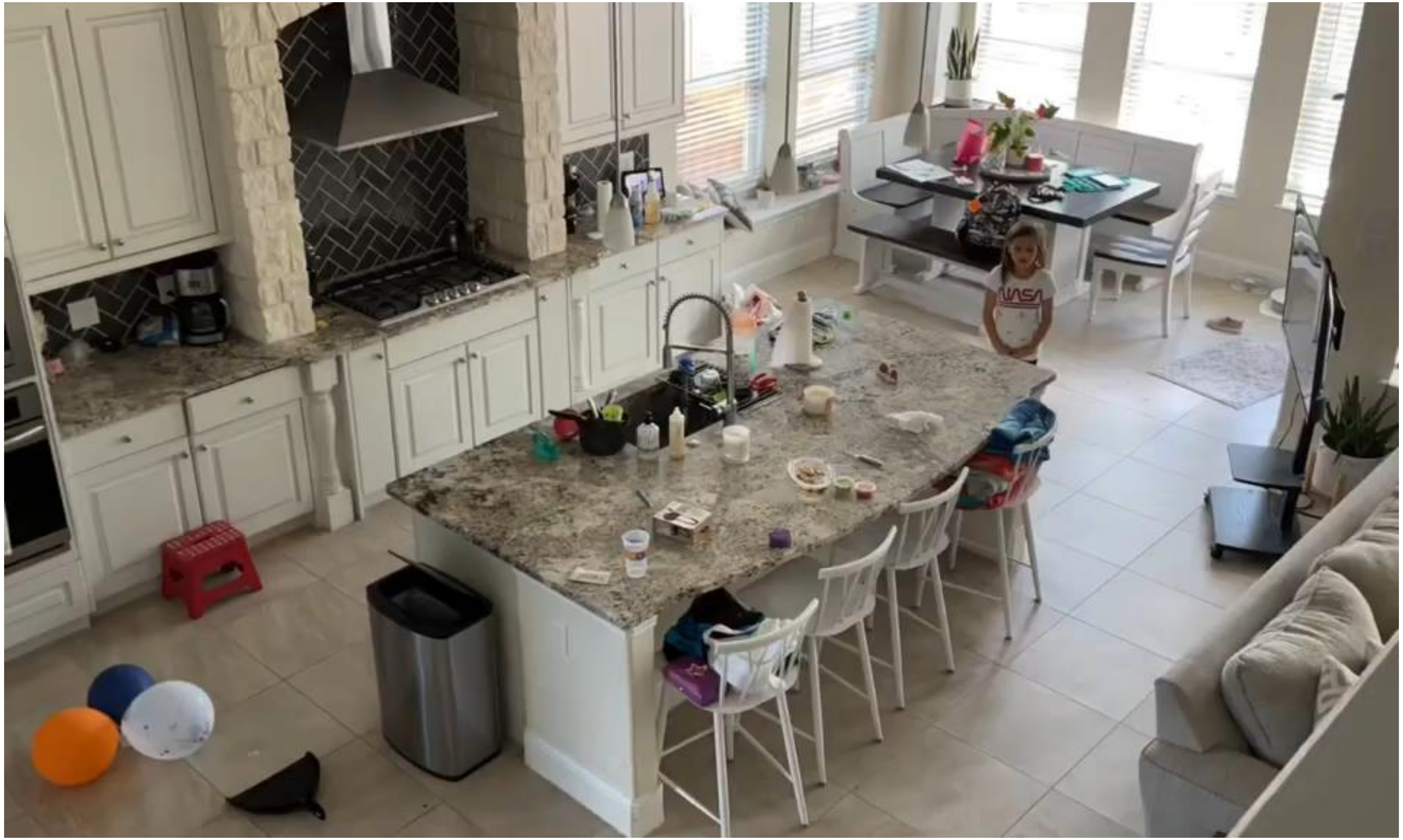
- Large diversity of environments and possible actions

- Control robots by natural language

Use vision-language models

Language-defined goals

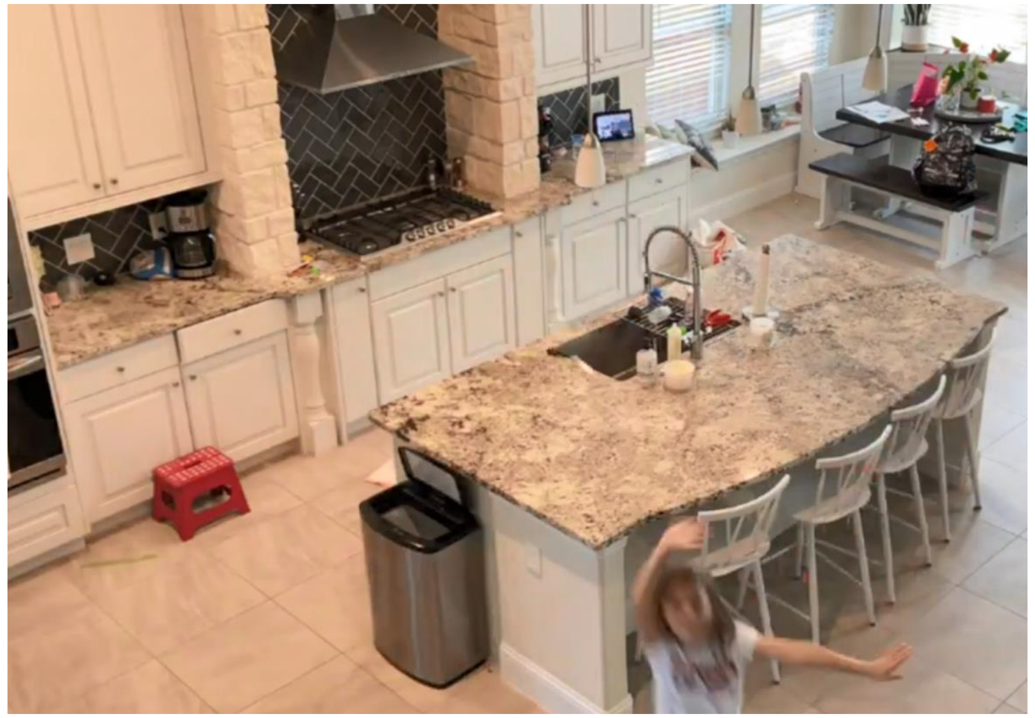
g: Clean the kitchen



Language-defined goals

g : Clean the kitchen

$$\pi_\theta(a_t|o_t, g)$$



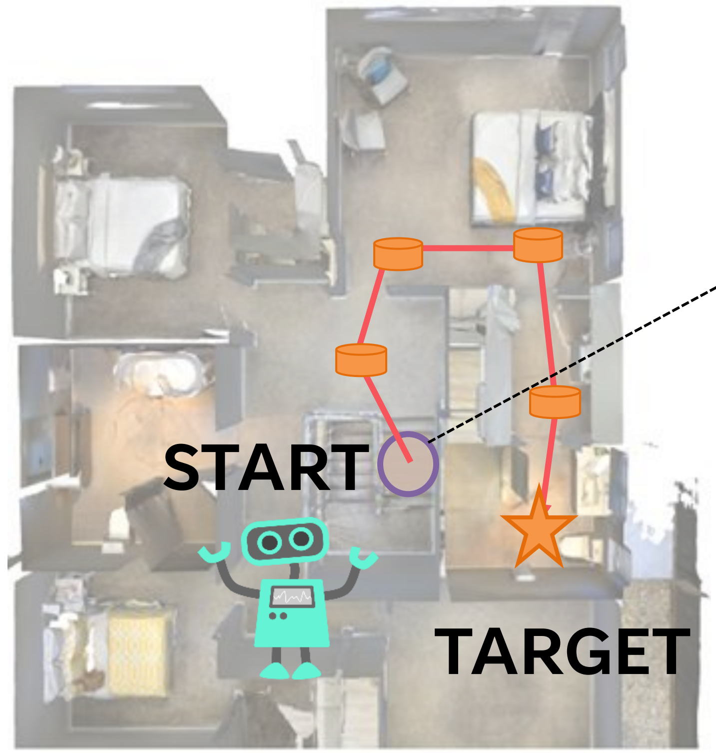
Navigation policy

Manipulation policy

Vision-and-Language Navigation (VLN)

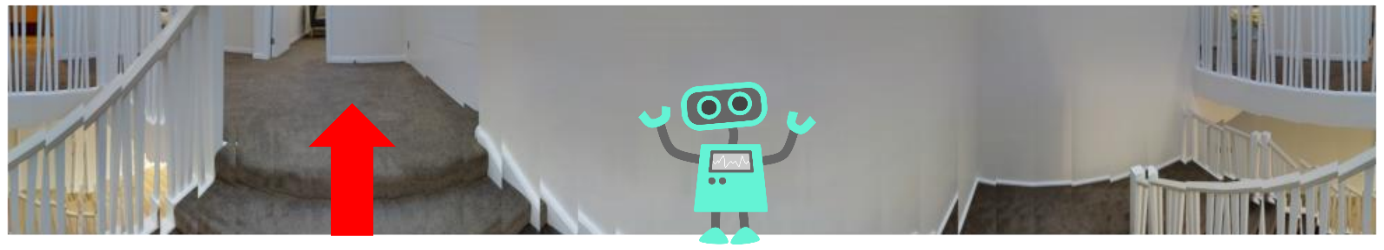
Train autonomous agents that can follow natural language instructions to navigate in realistic environments

Bird's-eye View
(invisible to the agent)



“ Go to the bathroom on the second floor and clean the mirror.

Panoramic Image
(agent's observation)



1) Understand the complicated language

2) Associate the language with visual observation

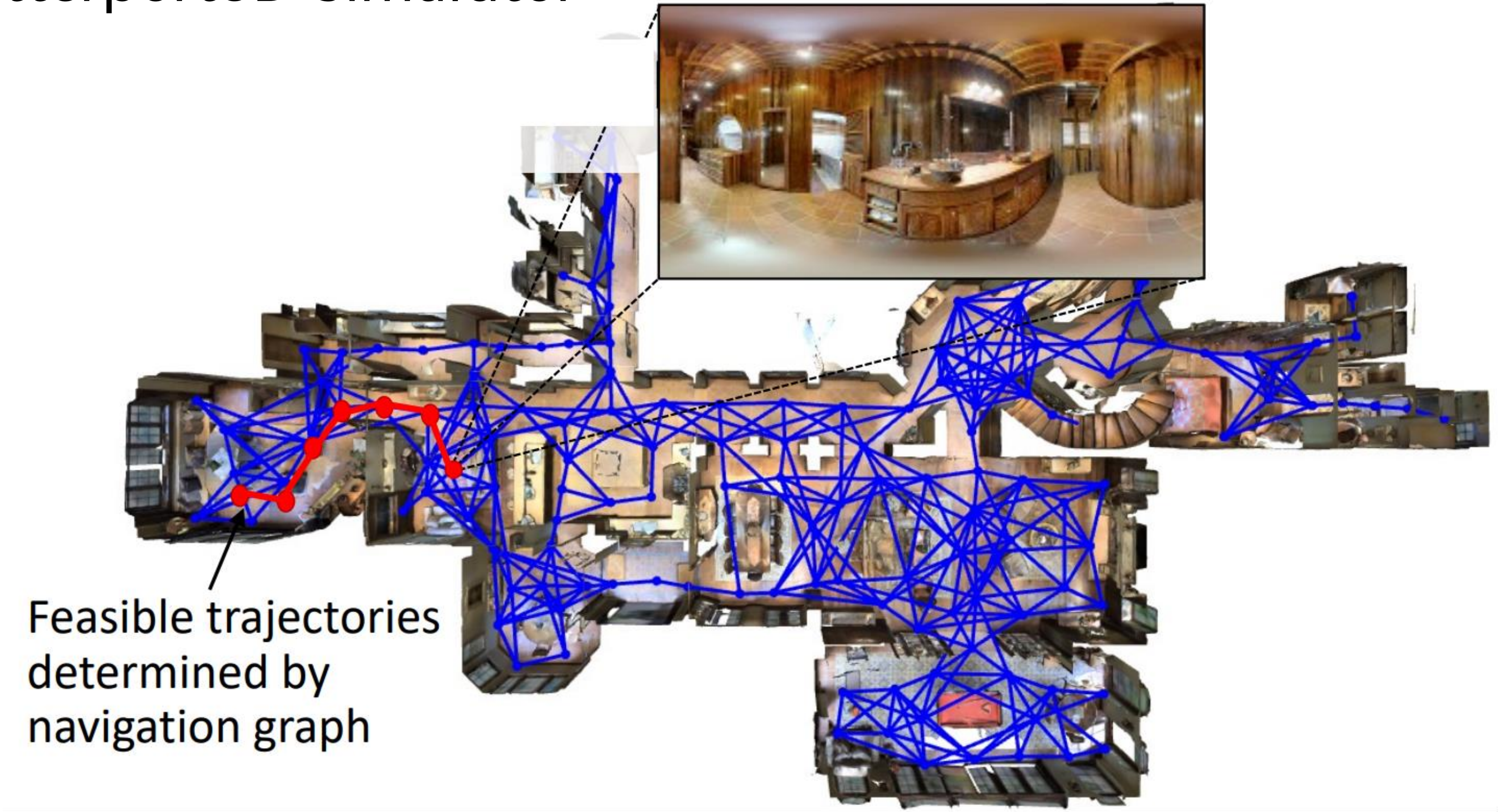
3) Sequentially make actions to explore the unseen house and find the target location

Matterport3D Simulator

- Simulator for embodied visual agents, based on Matterport3D dataset (Chang et. al. 2017)
- Contains 10,800 panoramic images / 90 buildings
- High visual diversity

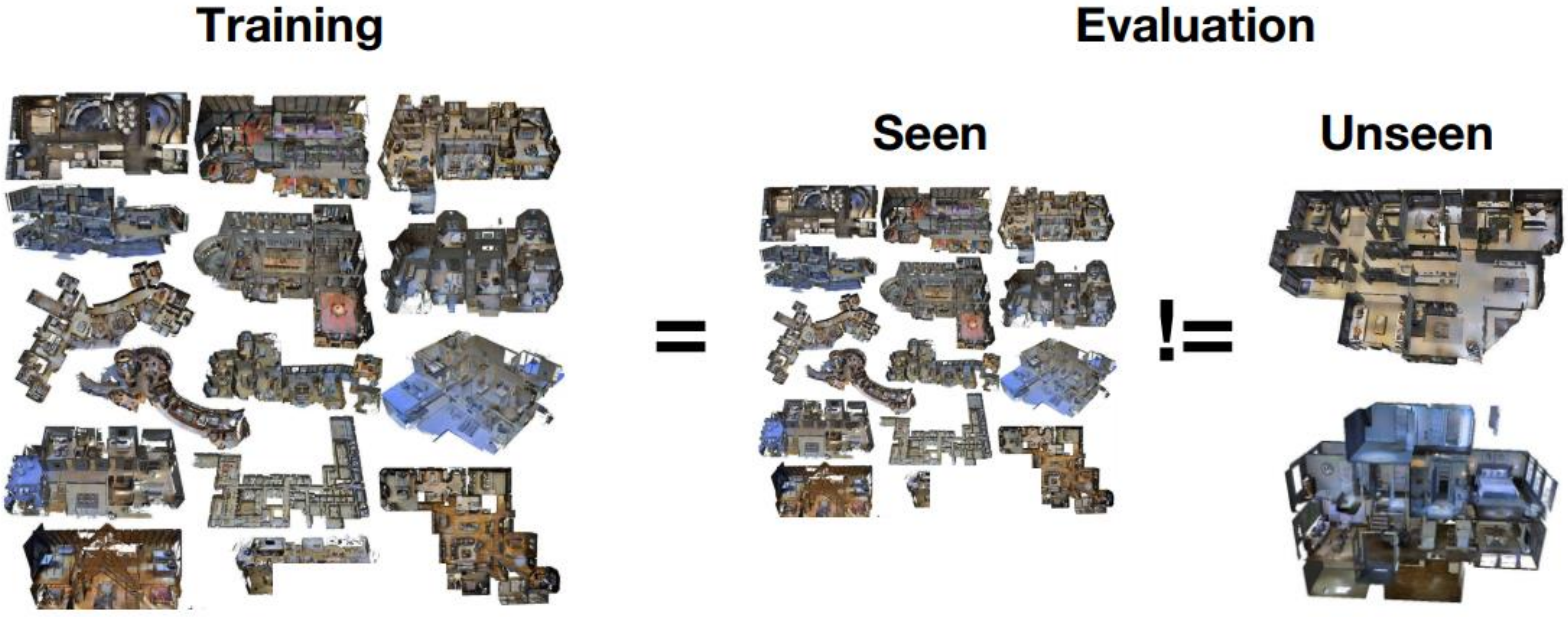
Matterport3D: Learning from RGB-D Data in Indoor Environments, Chang et al., 3DV 2017

Matterport3D Simulator



Matterport3D: Learning from RGB-D Data in Indoor Environments, Chang et al., 3DV 2017

Matterport3D Simulator



Matterport3D: Learning from RGB-D Data in Indoor Environments, Chang et al., 3DV 2017

VLN Challenge 1: Data

Learning good representations for VLN tasks

Existing works extract image features with models pretrained on Internet images

Egocentric images

Internet images

Contain more diverse views of scenes and objects

Require more spatial relation reasoning besides object categories

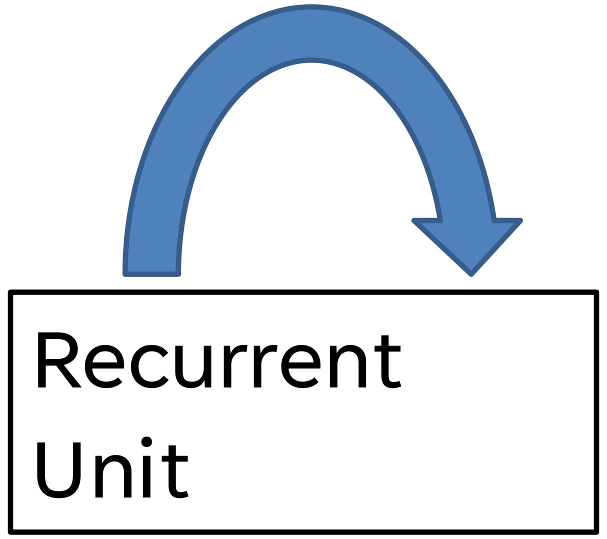
"walk to the back of the couch"

VLN Challenge 2: Modeling history

Keeping track of navigation history in agent's memory

Existing works mainly adopt a fixed-size recurrent unit to encode history

"Go to the bathroom on the second floor and clean the mirror.



Prone to forget previous observations in long navigation trajectories

Helpful to understand the environment

Correct previous navigation decisions and explore new areas

Airbert: In-domain Pretraining for Vision-and-Language Navigation

Pierre-Louis Guhur[1], Makarand Tapaswi[1,2], Shizhe Chen[1], Cordelia Schmid[1], Ivan Laptev[1]

ICCV 2021

Project page: https://airbert-vln.github.io
Code and data: https://github.com/airbert-vln

[1]Inria, École normale supérieure, CNRS, PSL Research University, Paris, France, [2]IIIT Hyderabad, India

Challenges

Training: 61 environments
R2R dataset

Testing: unseen objects
and scenes

“ Walk down the hall toward the Christmas tree. Stop in front of the first Christmas tree.

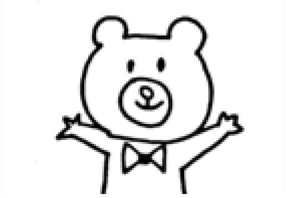
Limited amount and diversity of VLN training data

VLN-BERT: learning from web image-caption pairs

1. Pretraining

Conceptual Captions (image-caption pairs)

Facade of an old shop



a cartoon illustration of a bear waving and smiling

the scenic route through mountain range includes these unbelievably coloured mountains

trees in a winter snowstorm

2. VLN fine-tuning

R2R (path-instruction pairs)

“ Turn around and go straight. Take a left at the wall and go straight.

start

moves from node to node

target

Limitations:

- Out-of-domain pretraining
- Lacks temporal reasoning

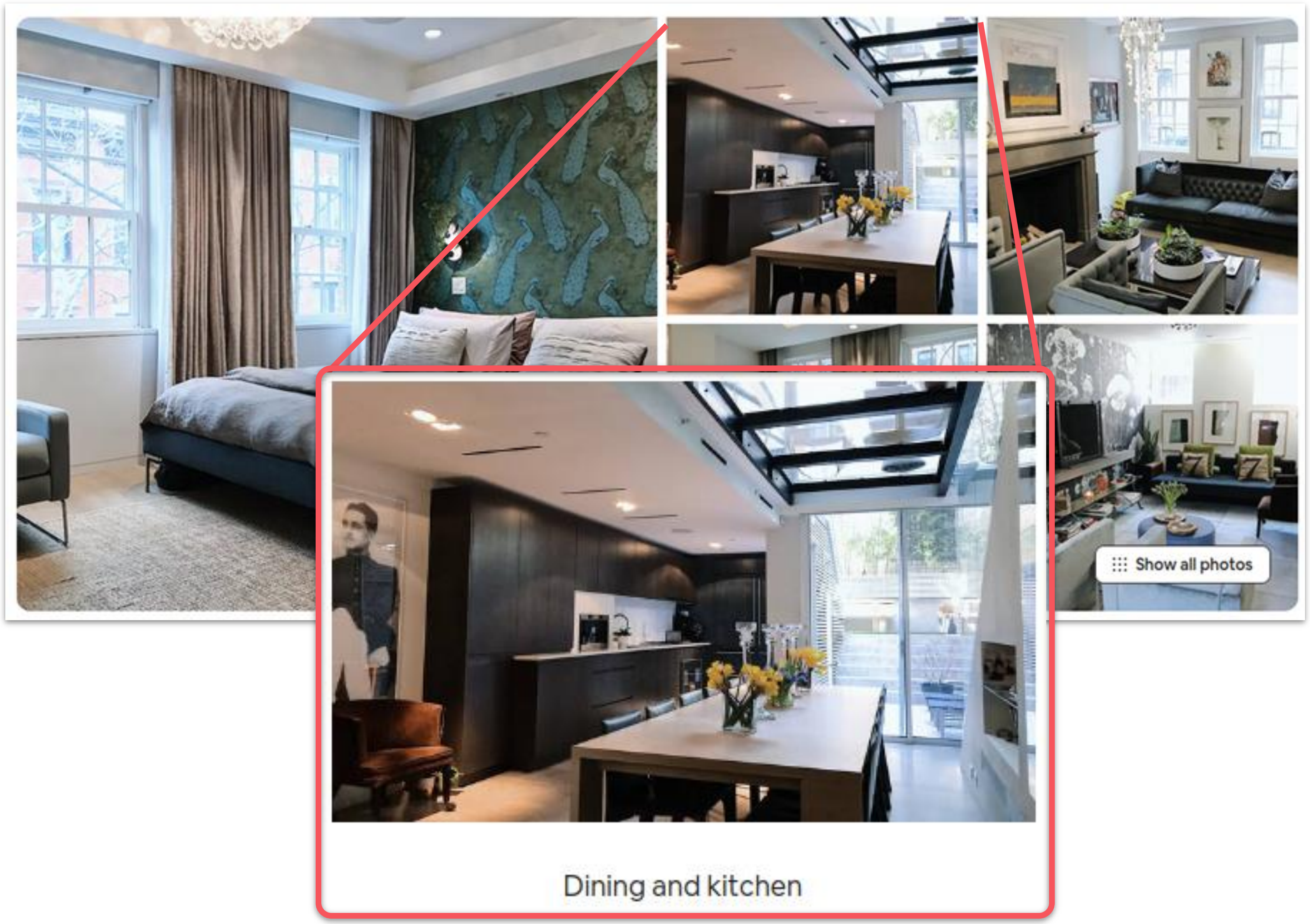
[Majumdar et al., ECCV 2020]

Self-supervised In-domain Pretraining

Collected BnB, a large-scale in-domain dataset

- 150K US listings from AirBnB
- Remove outdoor images
- Remove invalid captions

Dataset	Source	#Envs	#Imgs	#Texts
R2R	Matterport	90	10.8K	21.7K
REVERIE	Matterport	86	10.6K	10.6K
Speaker	Matterport	60	7.8K	0.2M
ConCaps	Web	-	3.3M	3.3M
BnB (ours)	***Airbnb***	***140K***	***1.4M***	***0.7M***



Generating BnB Path Instructions

Input images with caption

Living room opening to the garden

Open kitchen with seating for 4

Bedroom desk

Output image sequence with instructions

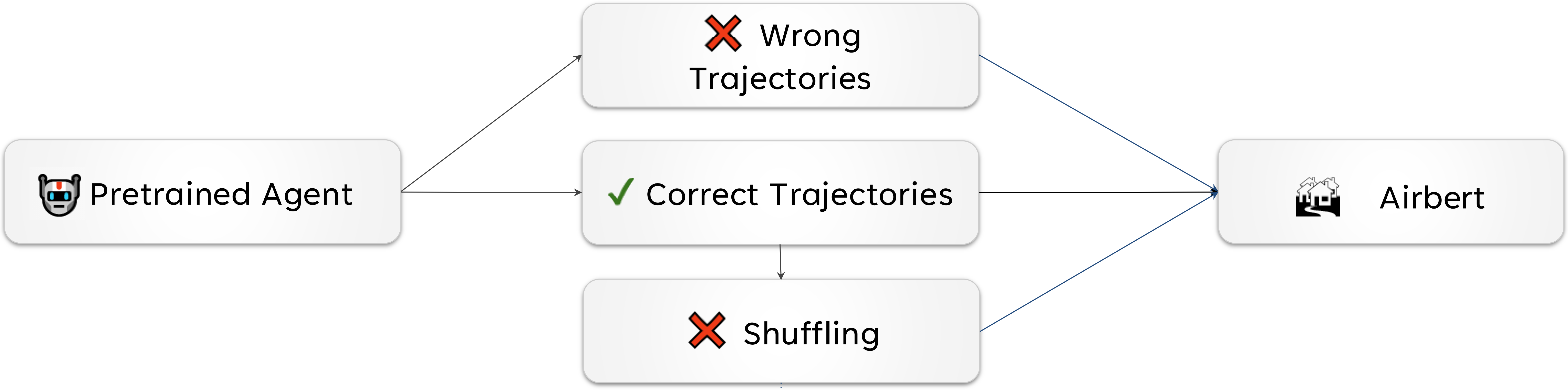
A. Concatenate image captions

“ Living room opening to the garden, open kitchen with seating for 4 and bedroom desk.

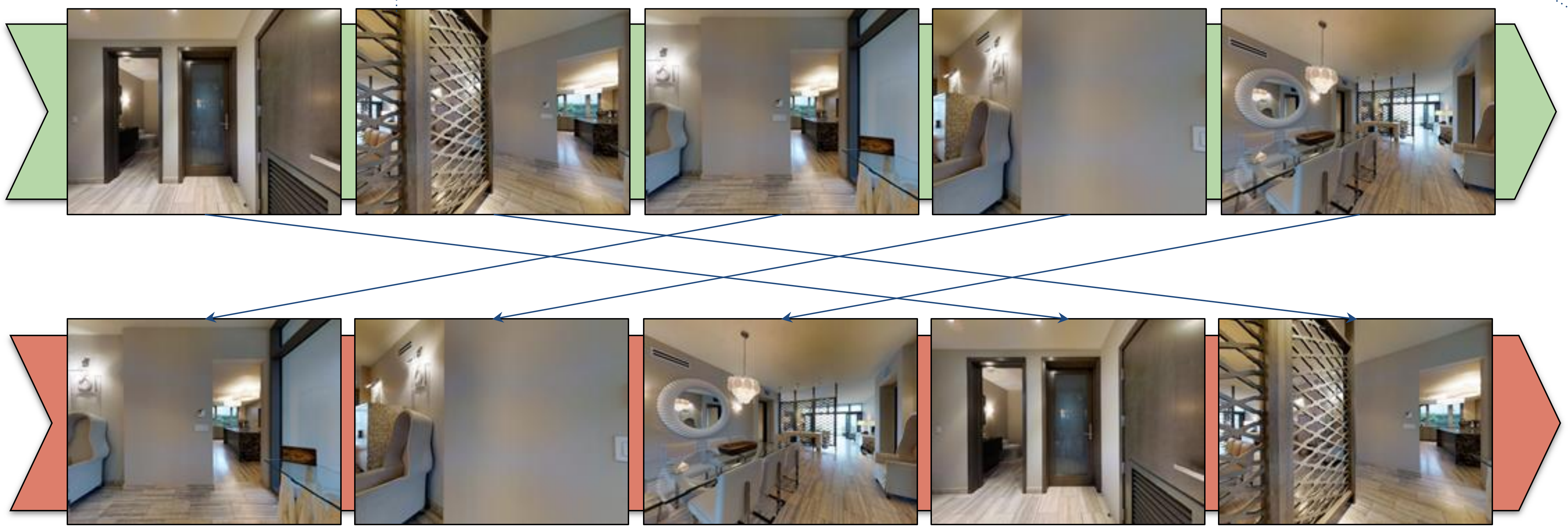
B. Use video ViLBERT captioning model

“ Exit the living room and walk through the bedroom. Stop in front of the two chairs.

VLN Pretraining: Shuffling loss



“ Walk down one flight of stairs and stop on the landing.



Results: building path-instructions out of image-captions

	BnB		Speaker		R2R		Success Rate	
	Mask	Shuffle	Rank	Shuffle	Rank	Shuffle	Seen	Unseen
1	-	-	-	-	✓	-	70.20	59.26
2	✓	-	-	-	✓	-	73.24	64.21
4	✓	-	✓	-	✓	-	70.21	65.52
5	✓	✓	✓	✓	✓	✓	73.83	**68.67**

VLN-BERT baseline

Results: building path-instructions out of image-captions

	BnB		Speaker		R2R		Success Rate	
	Mask	Shuffle	Rank	Shuffle	Rank	Shuffle	Seen	Unseen
1	-	-	-	-	✓	-	70.20	59.26
2	✓	-	-	-	✓	-	73.24	64.21
4	✓	-	✓	-	✓	-	70.21	65.52
5	✓	✓	✓	✓	✓	✓	73.83	**68.67**

VLN-BERT baseline

BnB dataset helps

Results: building path-instructions out of image-captions

	BnB		Speaker		R2R		Success Rate	
	Mask	Shuffle	Rank	Shuffle	Rank	Shuffle	Seen	Unseen
1	-	-	-	-	✓	-	70.20	59.26
2	✓	-	-	-	✓	-	73.24	64.21
4	✓	-	✓	-	✓	-	70.21	65.52
5	✓	✓	✓	✓	✓	✓	73.83	**68.67**

VLN-BERT baseline

BnB dataset helps

Speaker model

Results: building path-instructions out of image-captions

	BnB		Speaker		R2R		Success Rate	
	Mask	Shuffle	Rank	Shuffle	Rank	Shuffle	Seen	Unseen
1	-	-	-	-	✓	-	70.20	59.26
2	✓	-	-	-	✓	-	73.24	64.21
4	✓	-	✓	-	✓	-	70.21	65.52
5	✓	✓	✓	✓	✓	✓	73.83	**68.67**

VLN-BERT baseline

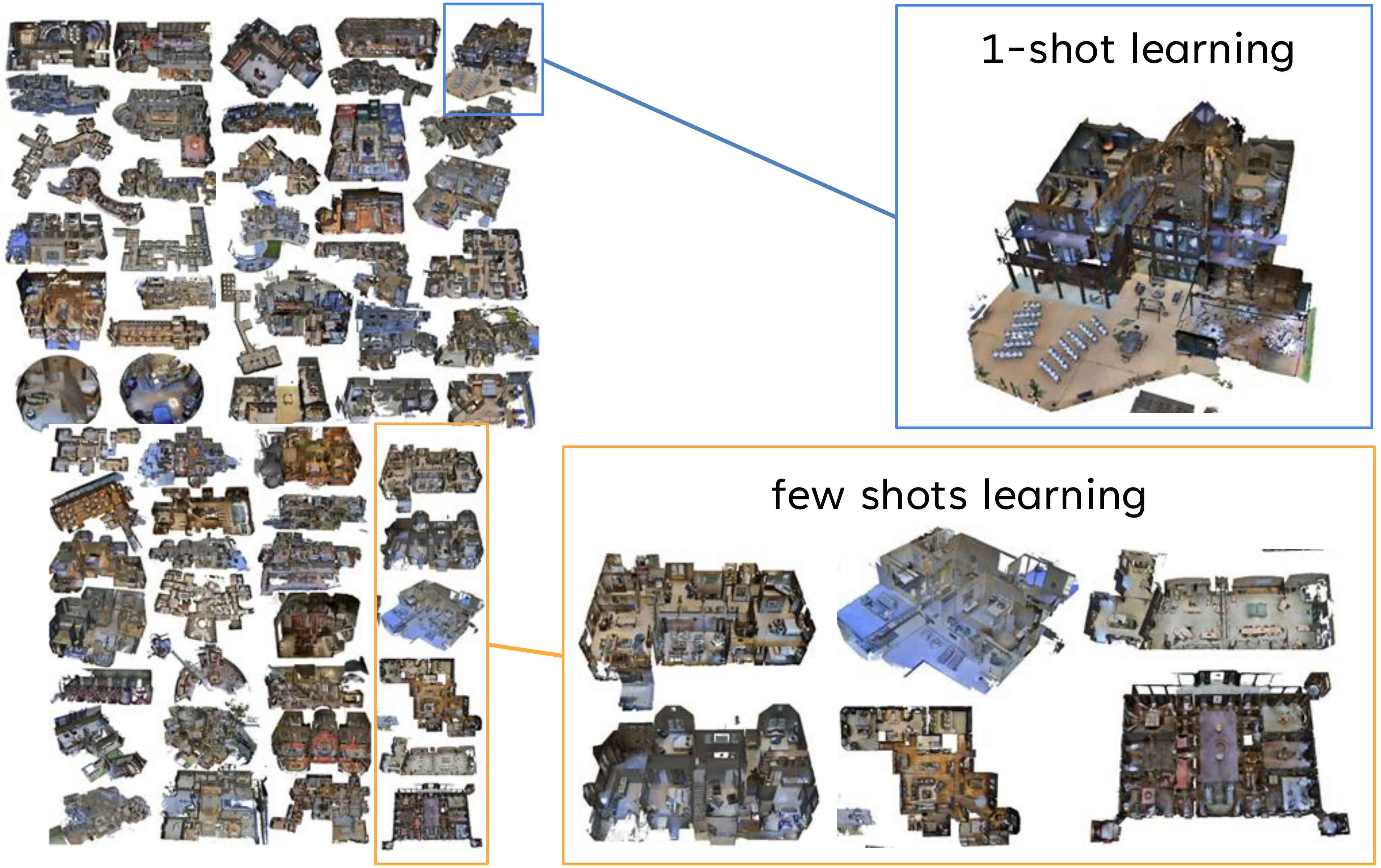
BnB dataset helps

Speaker model

Shuffling loss helps

Results: Few-Shot Learning

Can we learn to navigate given a few environments?

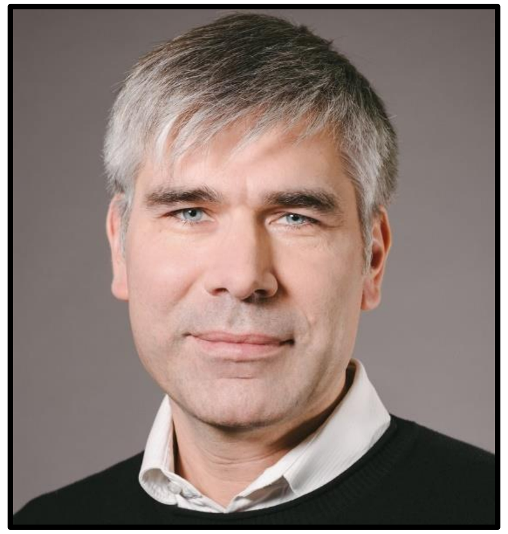


- Constraining the training over a very small number of environments.
- When tested on unseen environments, much of the objects were never observed.
- Airbert achieves much better performance in few-shot setting than VLN-BERT

# Envs	VLN-BERT	Airbert
1	27.06	**49.48**
10% (6)	37.01	**58.04**
Full (61)	57.15	**64.75**

History Aware Multimodal Transformer for Vision-and-Language Navigation



Shizhe Chen   Pierre-Louis Guhur   Cordelia Schmid   Ivan Laptev

NeurIPS 2021

Webpage: https://cshizhe.github.io/projects/vln_hamt.html

VLN Challenges: Modeling history

Keeping track of the navigation state

Environment understanding

Instruction grounding

Turn left
up the stairs.
Go straig
the bedr
the right
past the

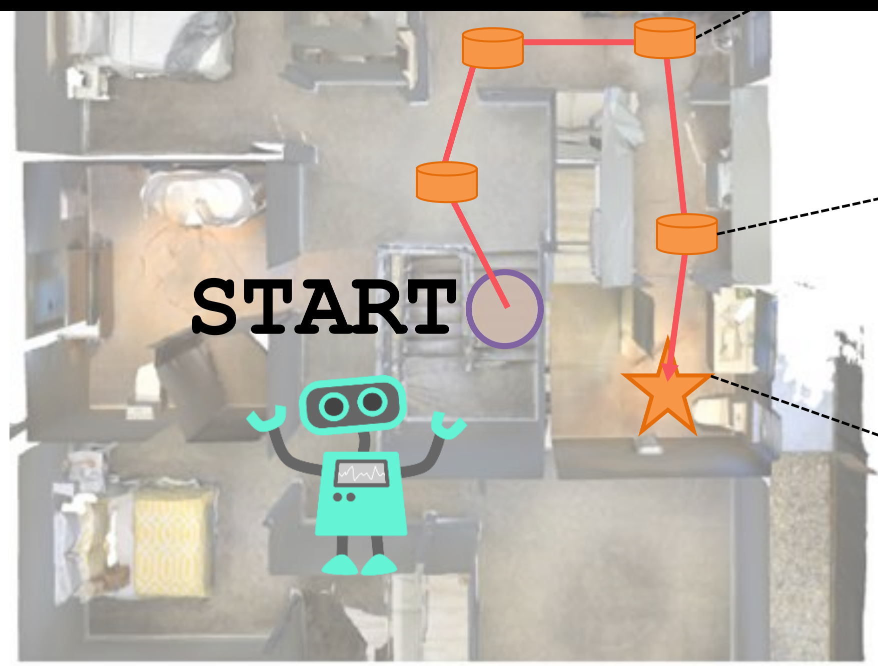
Turn right again and go through the closet.

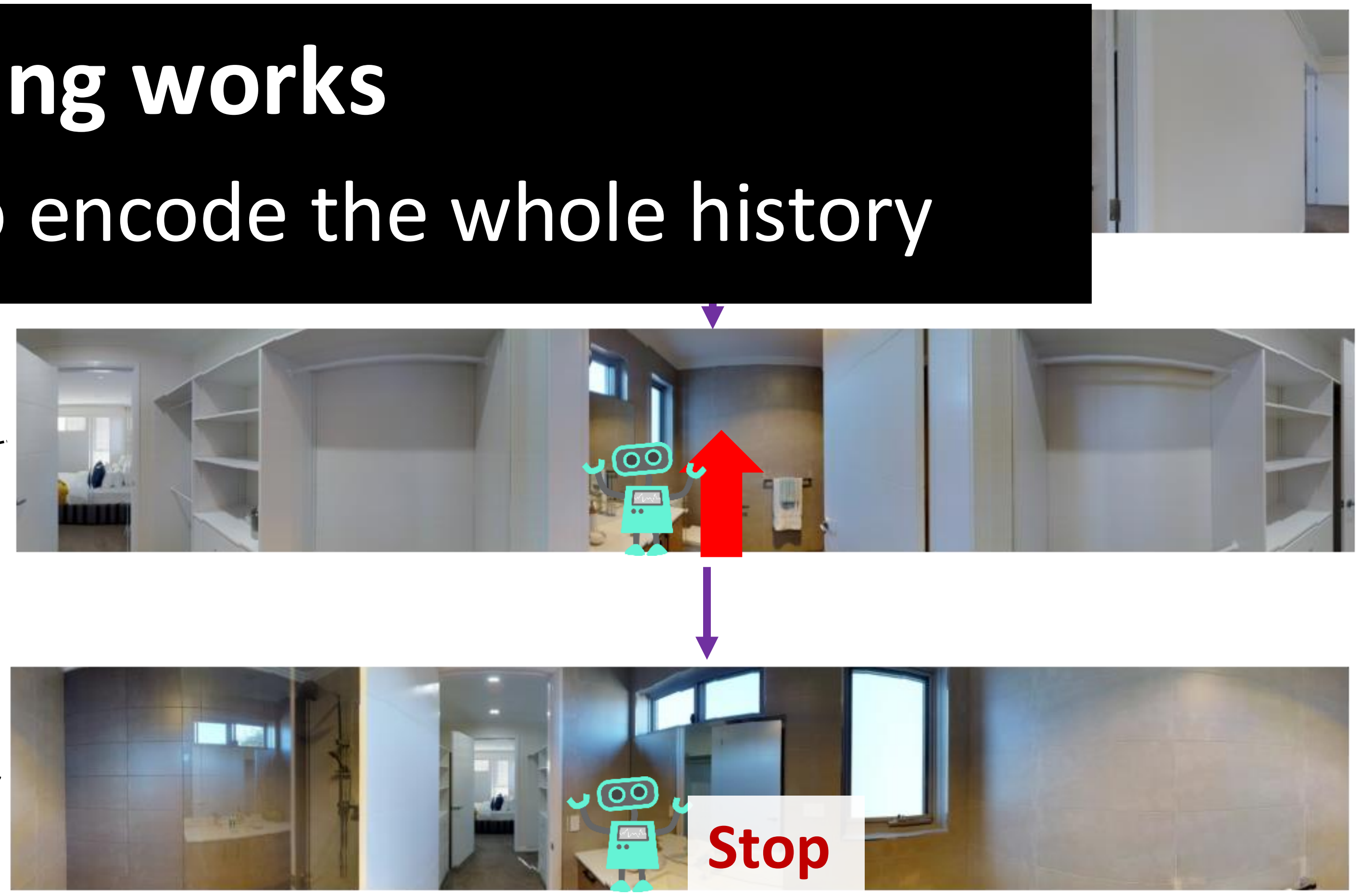
Continue straight, into the bathroom.

Wait right there, in front of the mirror.

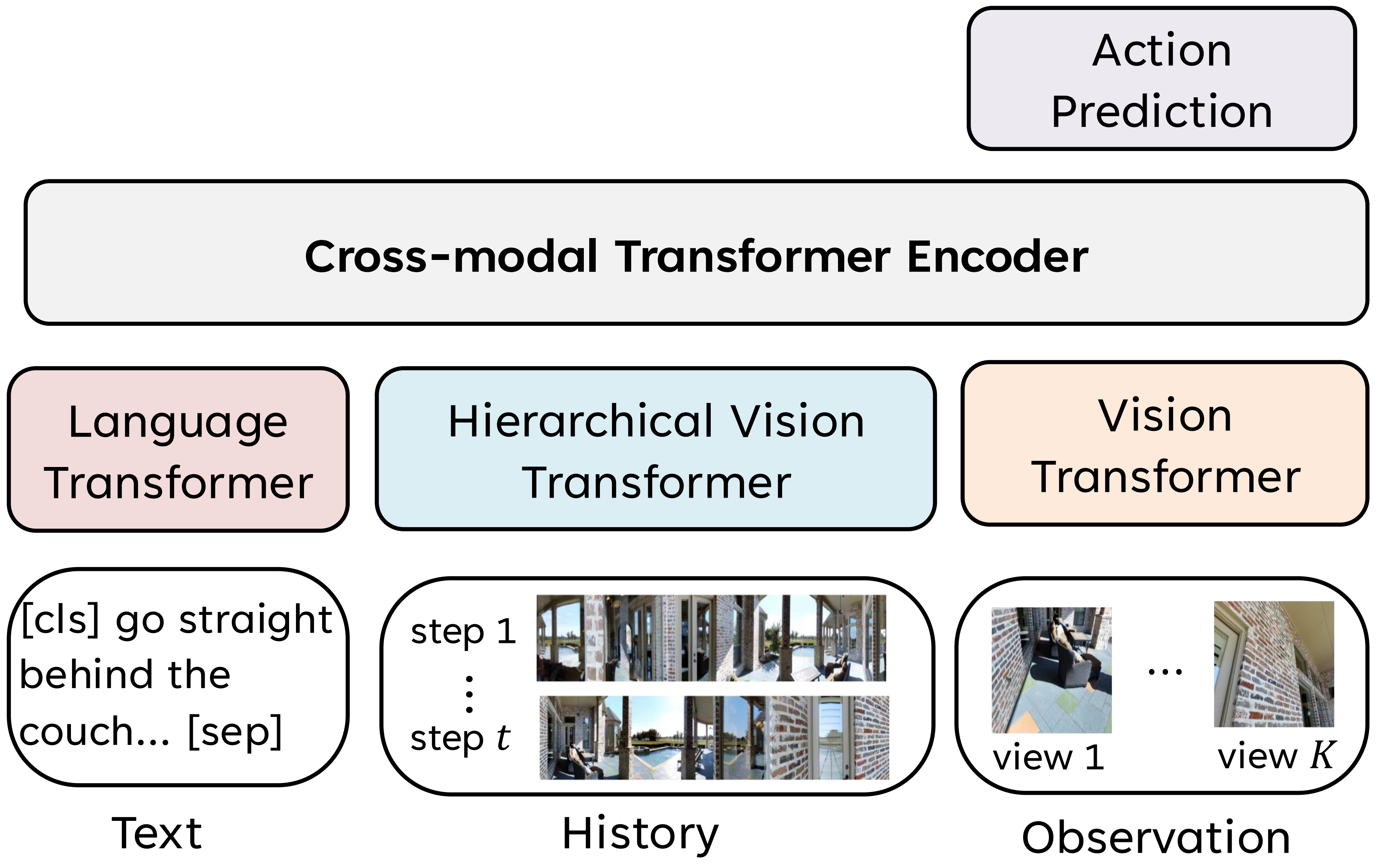
Limitations of existing works

Adopt a fixed-size recurrent unit to encode the whole history





History Aware Multimodal Transformer (HAMT)

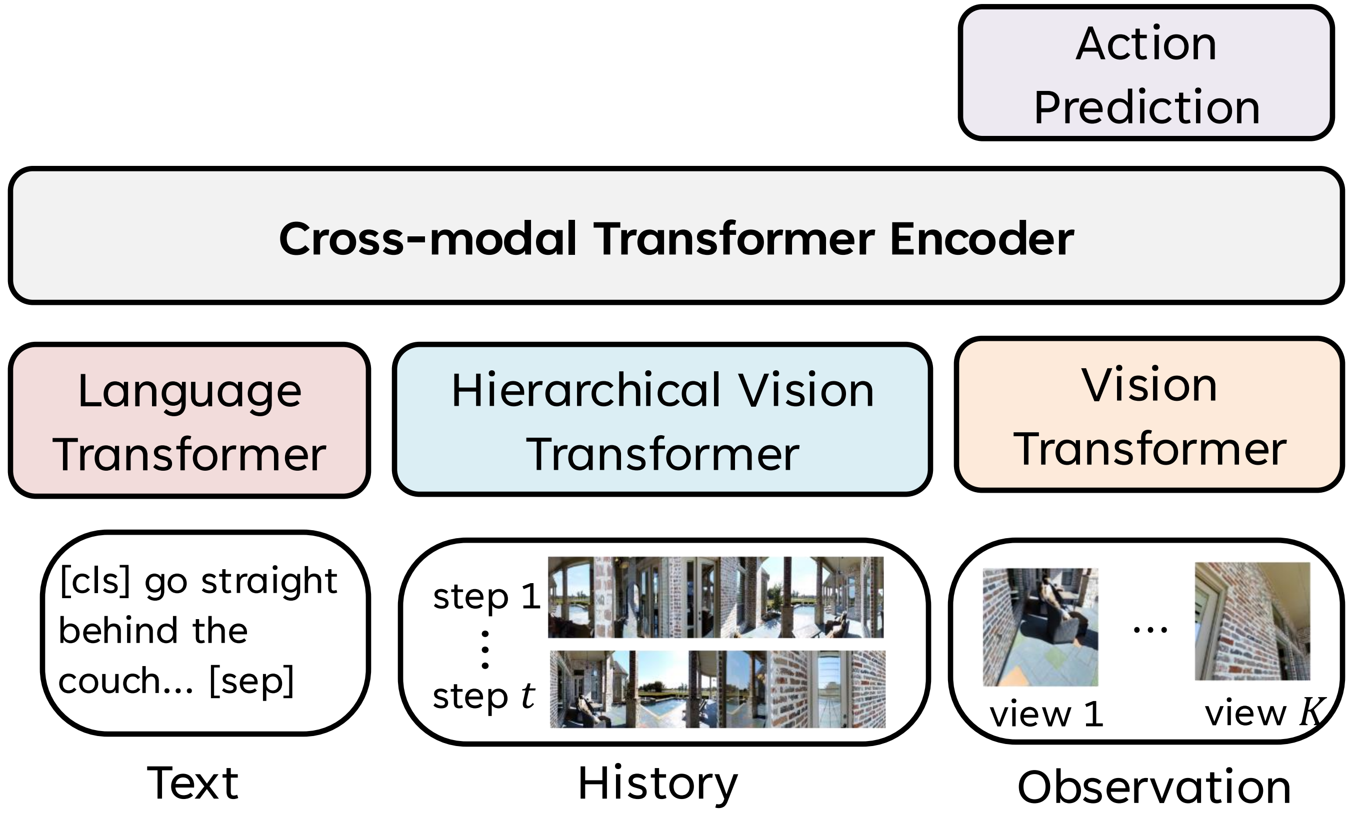


A fully transformer-based architecture for multimodal decision making

History Aware Multimodal Transformer (HAMT)

Long-horizon history modelling
- Learn dependency of all panoramic observations and actions in history sequence

End-to-end optimization for visual representation
- Fully transformer-based architecture allows efficient training

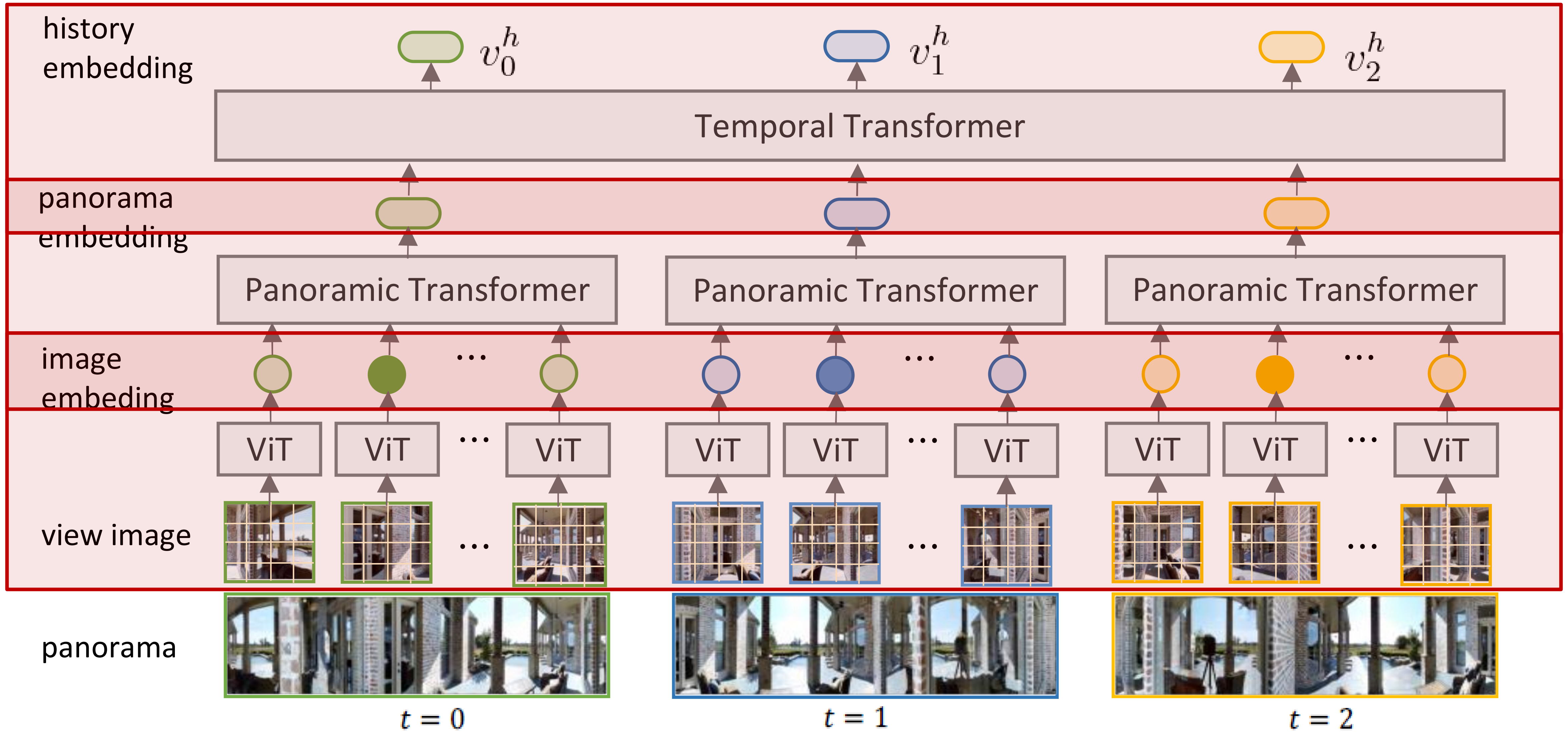


PROBLEMS

- Computational expensive to encode all panoramas
  - K views, T steps → $O(K^2T^2)$
- The action prediction task alone might be insufficient to learn generalizable models

HAMT: Hierarchical History Encoding

- ViT for single view image encoding
- Panoramic Transformer for spatial relation encoding within panorama
- Temporal Transformer for temporal relation encoding across panoramas



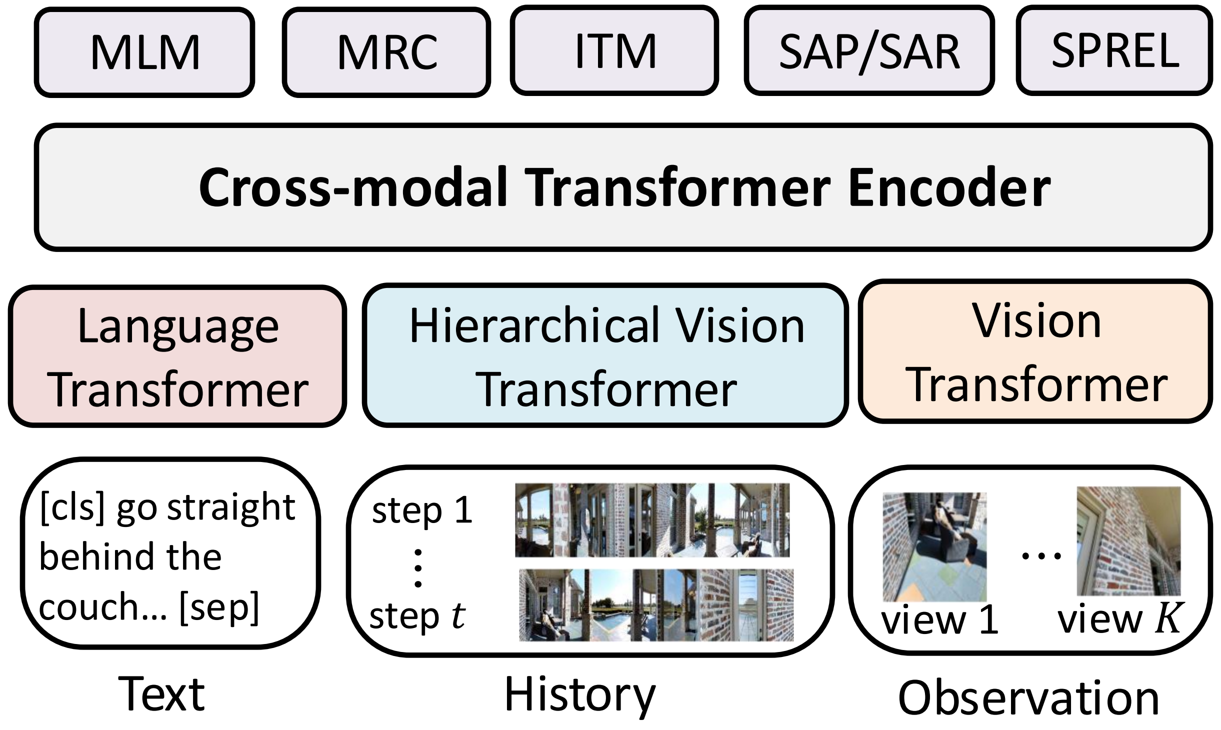
HAMT: End-to-End Training with Proxy Tasks

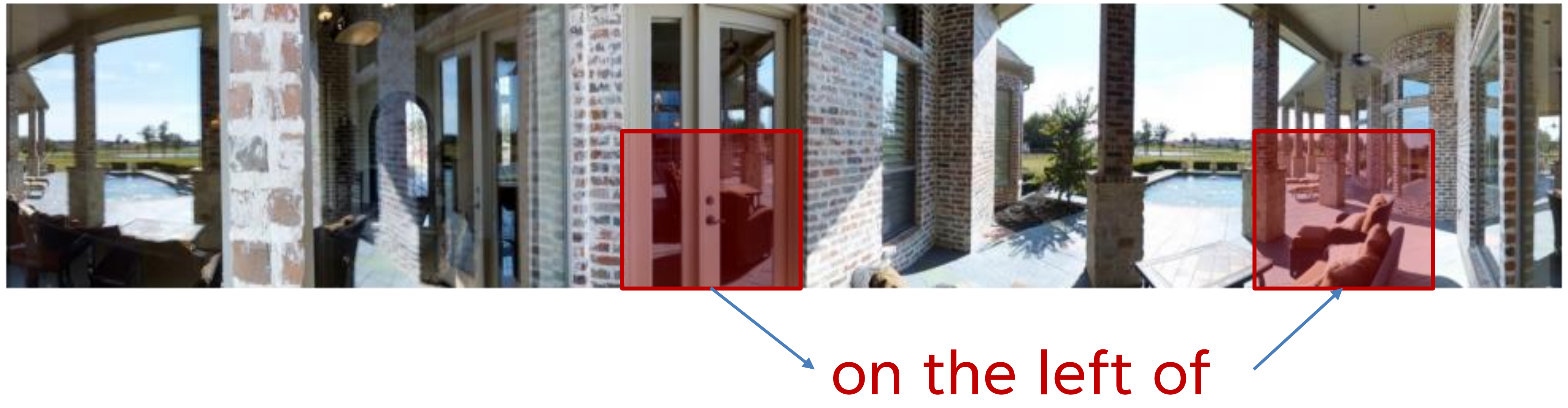
Common vision-and-language proxy tasks

- Masked Language Modelling
- Masked Region Modelling
- Instruction Trajectory Matching

New proxy tasks for VLN

- Single-step Action Prediction/Regression
- Spatial Relationship Prediction





HAMT: Fine-tuning for Sequential Action Prediction

Combine Reinforcement Learning (RL) and Imitation Learning (IL)

$$\Theta \leftarrow \Theta + \underbrace{\mu \frac{1}{T} \sum_{t=1}^{T} \nabla_{\Theta} \log \pi(\hat{a}_t^h; \Theta)(R_t - V_t)}_{\text{Reinforcement Learning (RL)}} + \underbrace{\lambda \mu \frac{1}{T^*} \sum_{t=1}^{T*} \nabla_{\Theta} \log \pi(a_t^*; \Theta)}_{\text{Imitation Learning (IL)}}$$

RL: A3C Algorithm

Rewards: reduced navigation distance, path fildelity etc.

Experiments: Datasets

- R2R
- RxR

VLN with Fine-grained Instructions
Require continuous instruction grounding

- REVERIE
- R2R-Last

VLN with High-level Instructions
Require scene memory

- CVDN

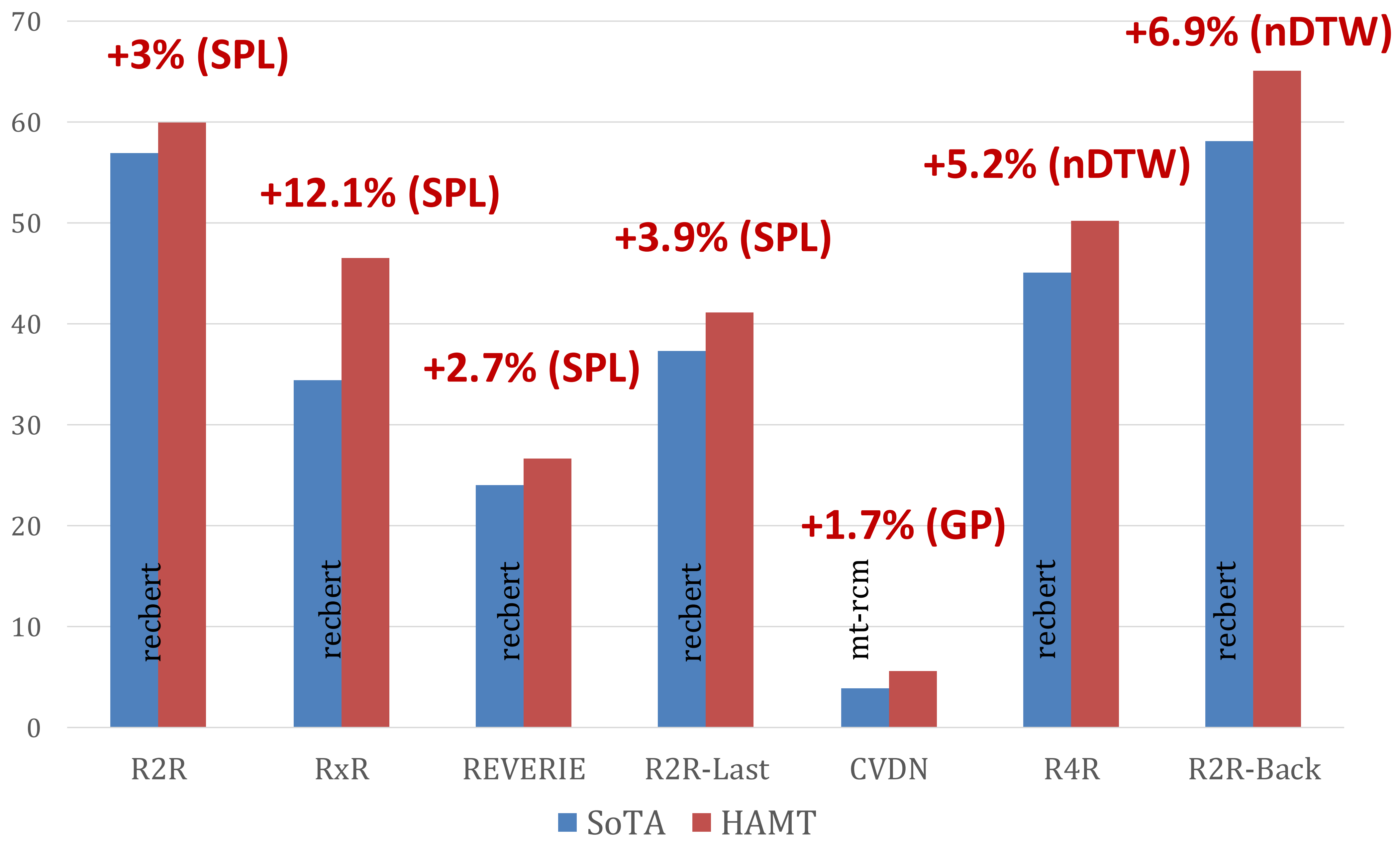
Vision-and-Dialogue Navigation
Emphasize dialogue history understanding

- R4R: concatenate 2 paths in R2R
- R2R-Back: R2R + Return back to the start

Long-horizon VLN
Emphasize long-term scene memory
& instruction grounding

Experiments: Comparison with SoTA

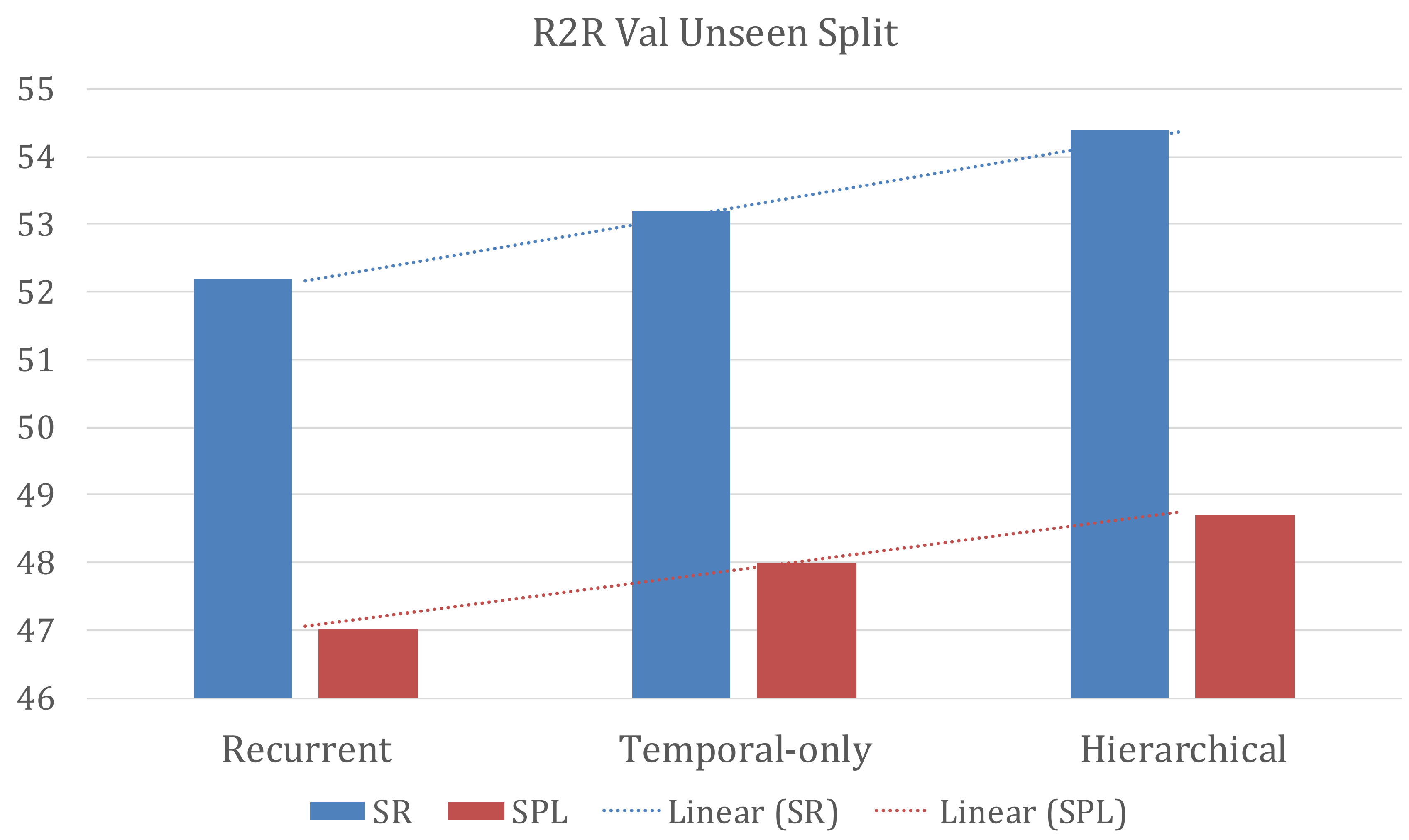
HAMT outperforms state of the art on all datasets



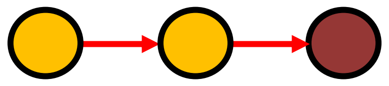
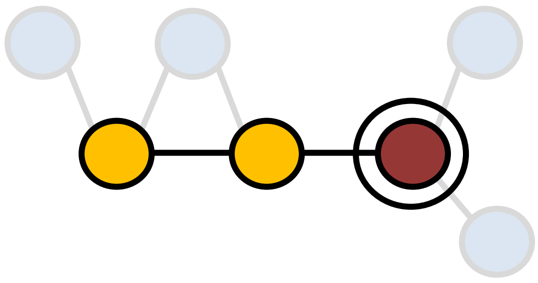
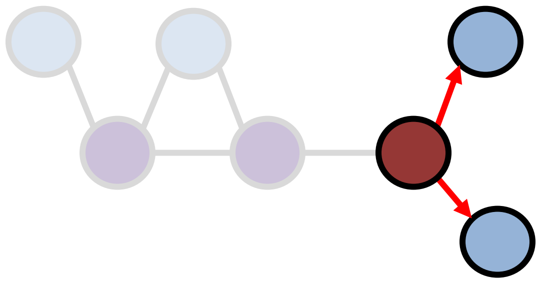
Experiments: Ablation

How important is the history encoding?

- Recurrent: a fixed-size vector to encode the whole history
- Temporal-only: select only one view per panorama to improve efficiency
- **Hierarchical: hierarchically encode all panoramas**



Limitations of HAMT

	Navigation Memory	Visual Memory	Action Space
HAMT			

Sequence
no structure of the house

Local actions
hard to backtrack many steps

Current location

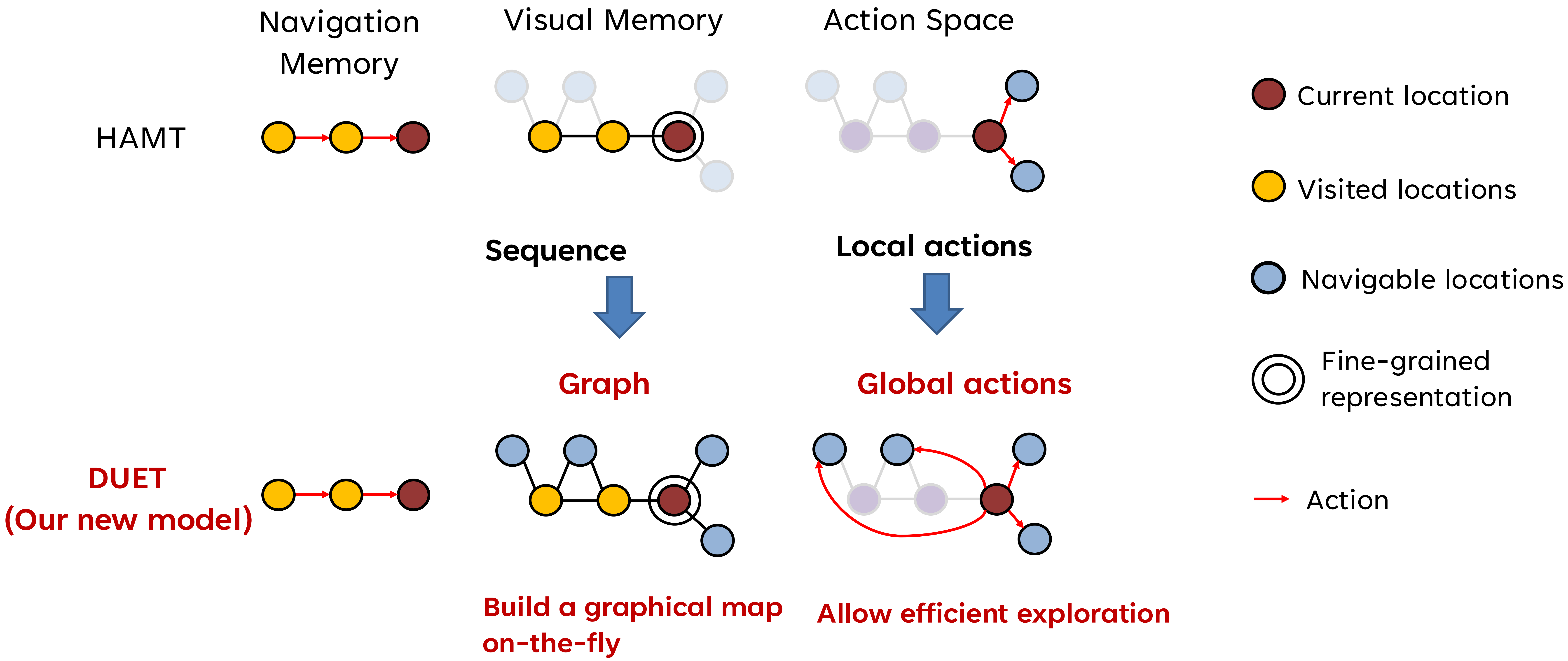
Visited locations

Navigable locations

Fine-grained representation

Action

Limitations of HAMT



DUET: Experimental Results

REVERIE dataset

	SR	SPL	RGS	RGSPL
HAMT	30.40	26.67	14.88	13.08
DUET	**52.51**	**36.06**	**31.88**	**22.06**

SOON dataset

Split	Methods	TL	OSR↑	SR↑	SPL↑	RGSPL↑
Val Unseen	GBE [8]	28.96	28.54	19.52	13.34	1.16
	DUET (Ours)	36.20	**50.91**	**36.28**	**22.58**	**3.75**
Test Unseen	GBE [8]	27.88	21.45	12.90	9.23	0.45
	DUET (Ours)	41.83	**43.00**	**33.44**	**21.42**	**4.17**

Winner of VLN Challenges hosted in Human Interaction for Robotics Navigation Workshop at ICCV 2021


1ST PLACE IN THE
REVERIE CHALLENGE 2021

Shizhe Chen[1], Pierre-Louis Guhur[1], Makarand Tapaswi[2]
Cordelia Schmid[1] and Ivan Laptev[1]
[1] Inria, École normale supérieure, CNRS, PSL Research University
[2] IIIT Hyderabad

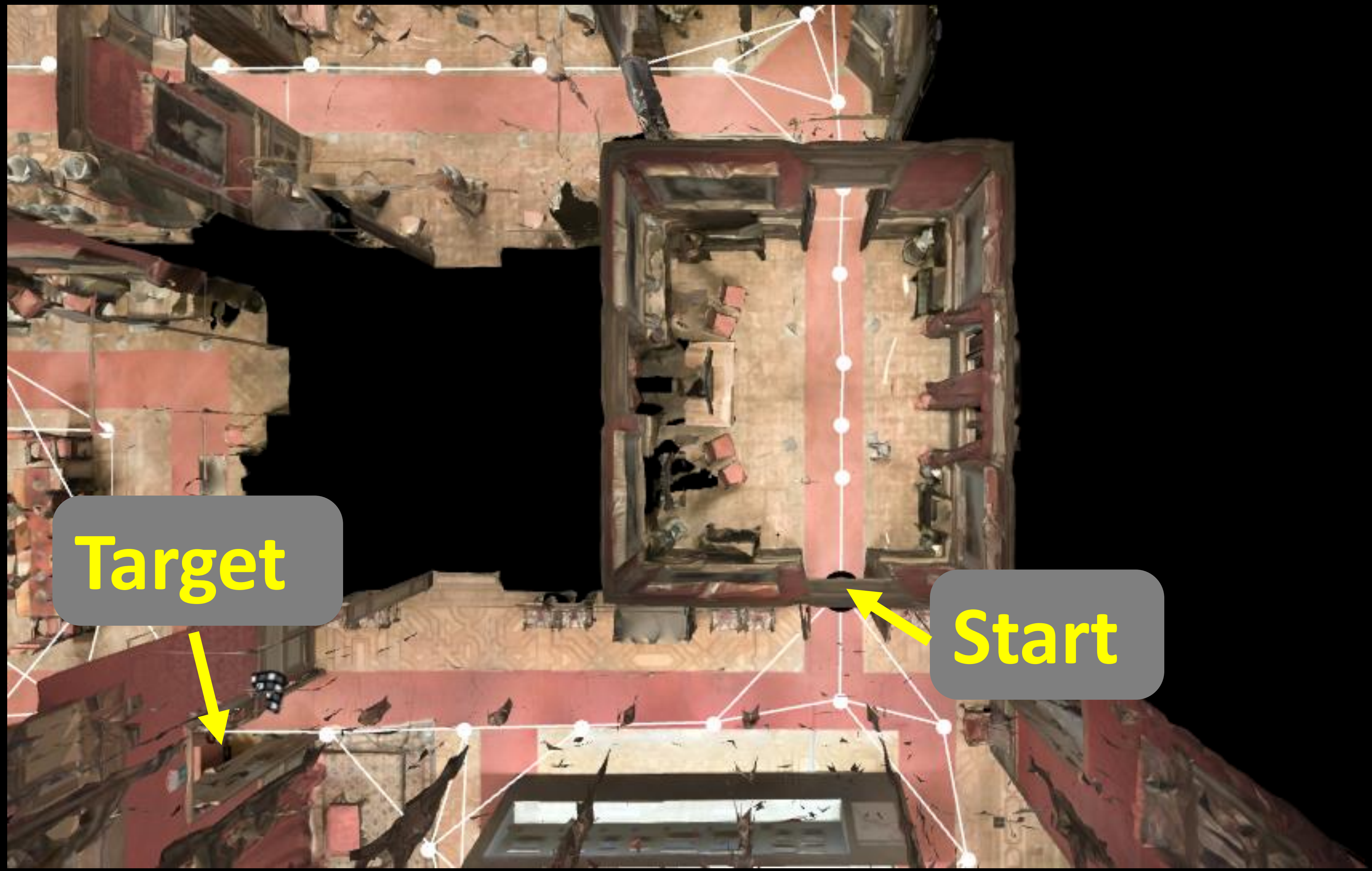
presented at the

Human Interaction for Robotic Navigation Workshop
at the IEEE/CVF International Conference on Computer Vision (ICCV) 2021

Oct. 16 2021

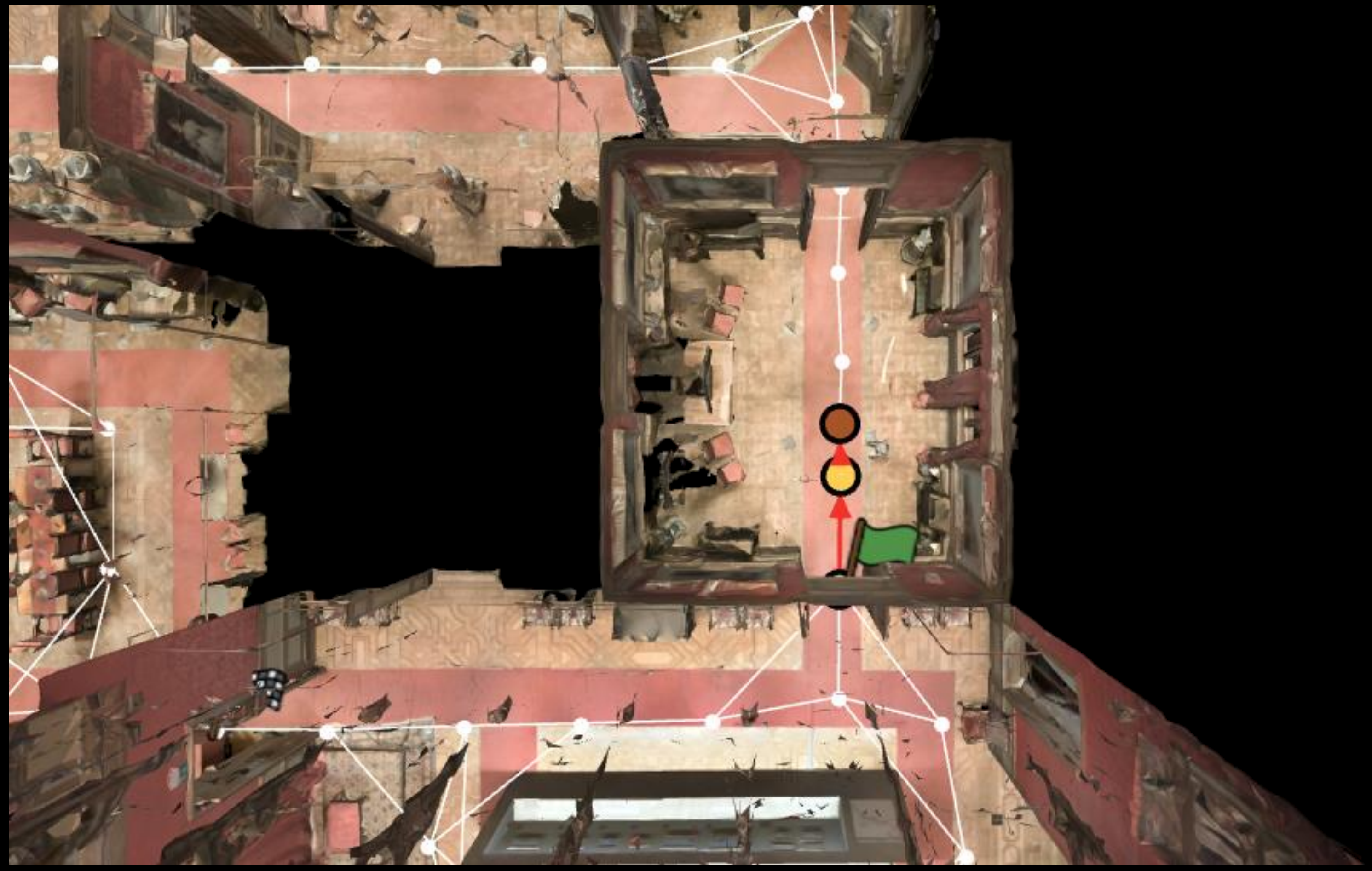
SIGNED, Dr. Qi Wu
On behalf of the 2020 REVERIE Challenge Organizers
Yuankai Qi Fengda Zhu Qi Wu

Instruction: Exit the roped off hall, follow the red carpet, turn right, continue straight down the red carpet, enter room at the end, stop once inside the room.

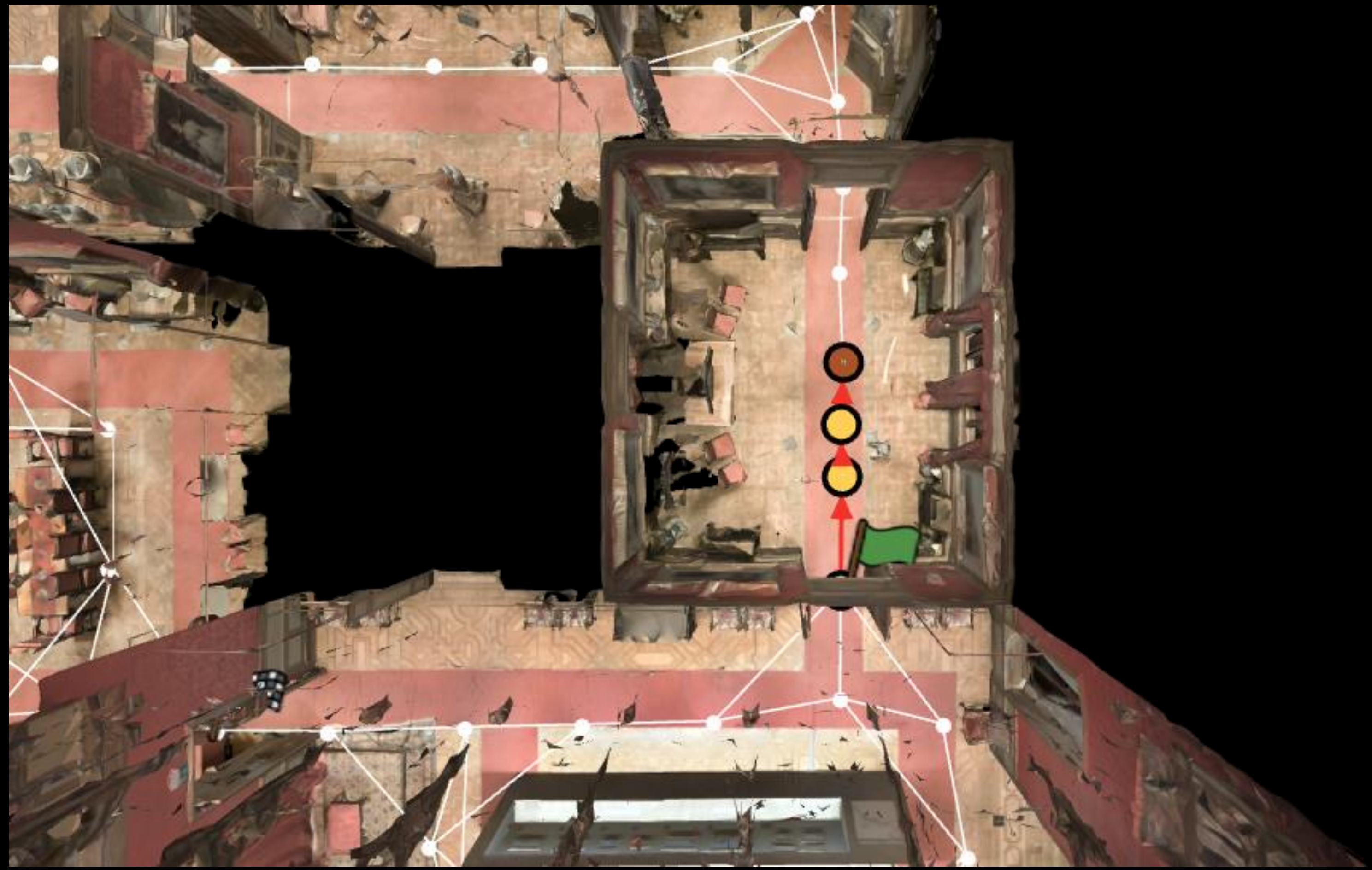


Instruction: **Exit the roped off hall, follow the red carpet**, turn right, continue straight down the red carpet, enter room at the end, stop once inside the room.

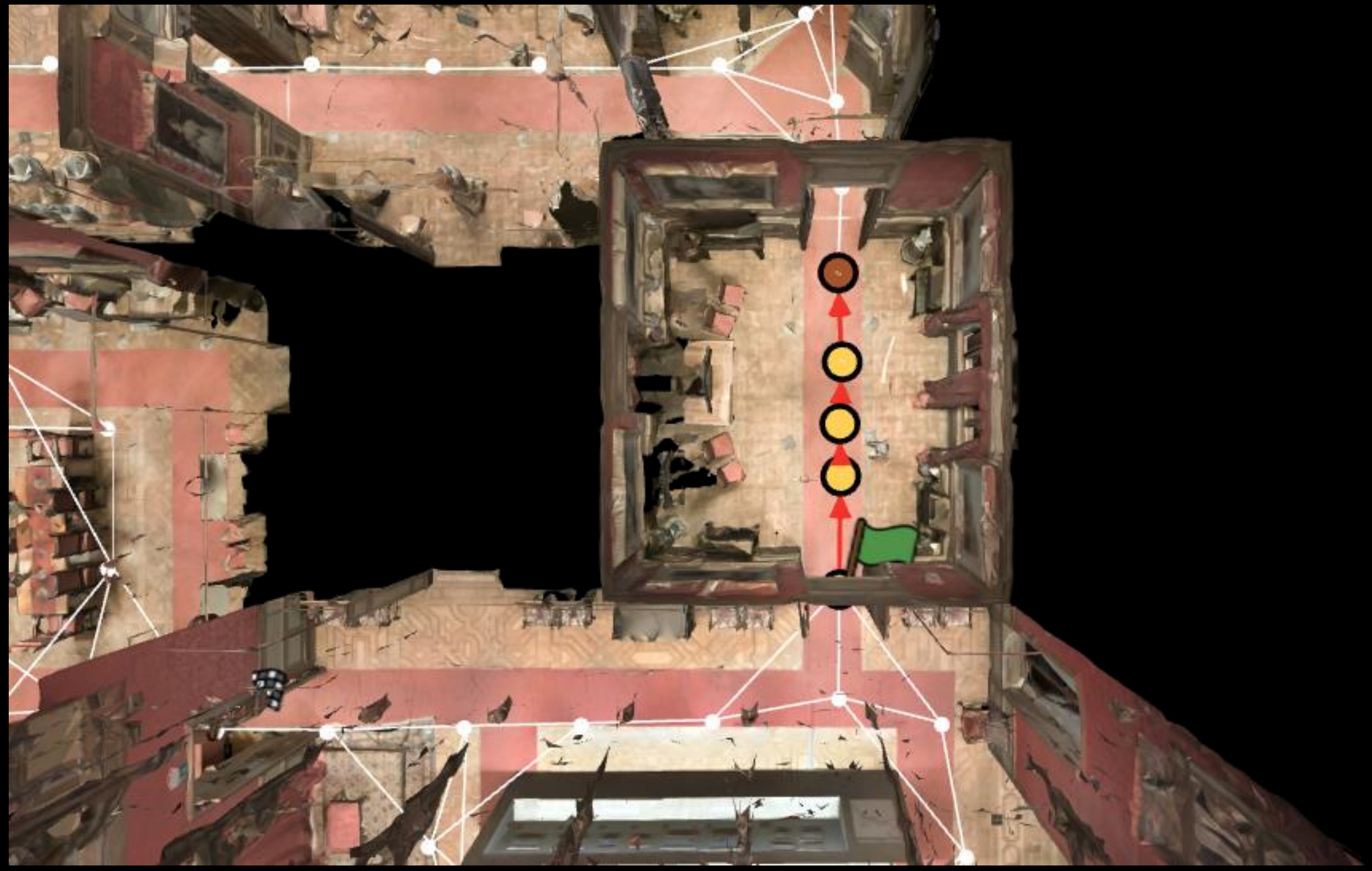
Instruction: **Exit the roped off hall, follow the red carpet**, turn right, continue straight down the red carpet, enter room at the end, stop once inside the room.



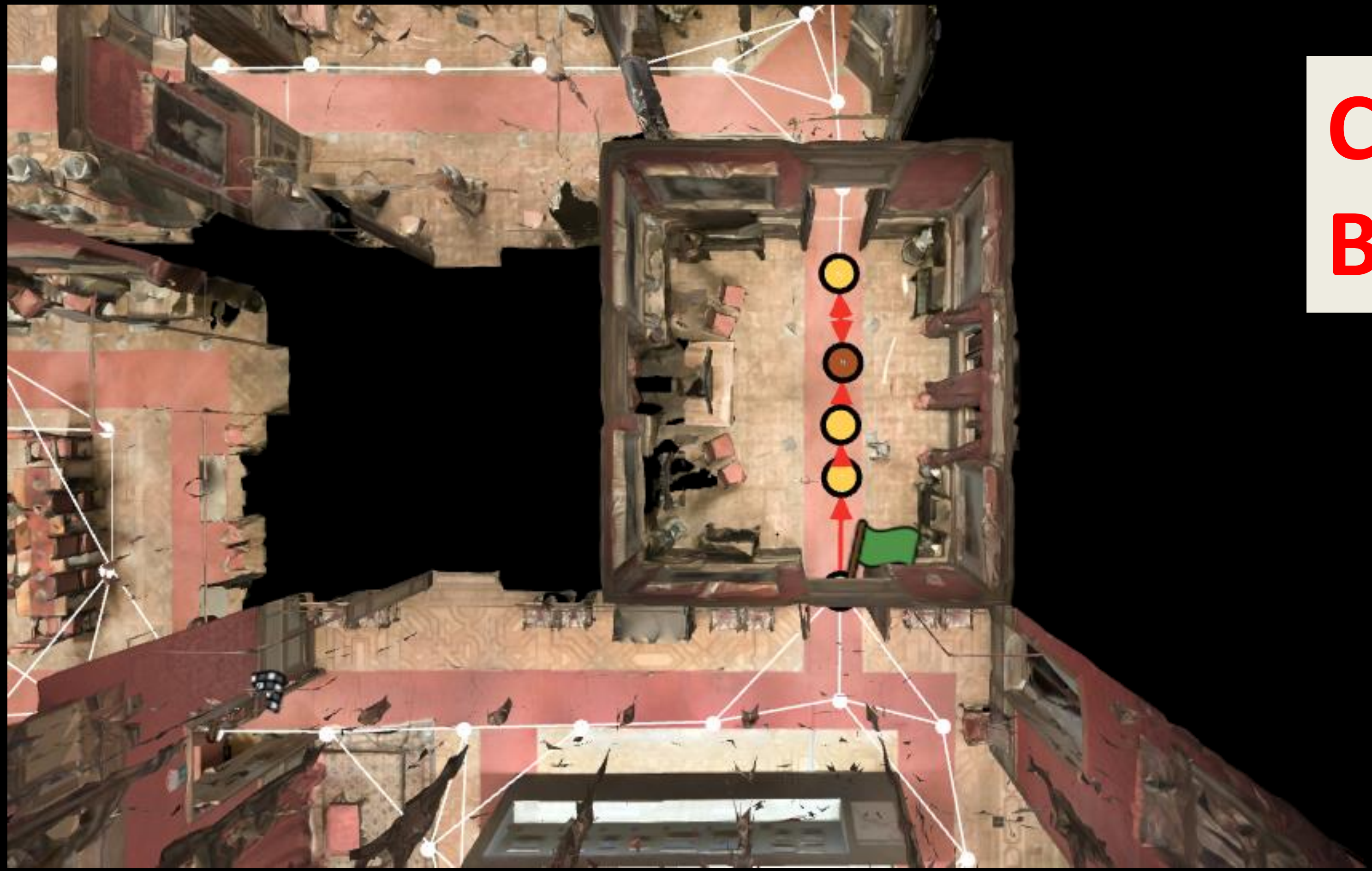
Instruction: **Exit the roped off hall, follow the red carpet**, turn right, continue straight down the red carpet, enter room at the end, stop once inside the room.



Instruction: **Exit the roped off hall, follow the red carpet**, turn right, continue straight down the red carpet, enter room at the end, stop once inside the room.

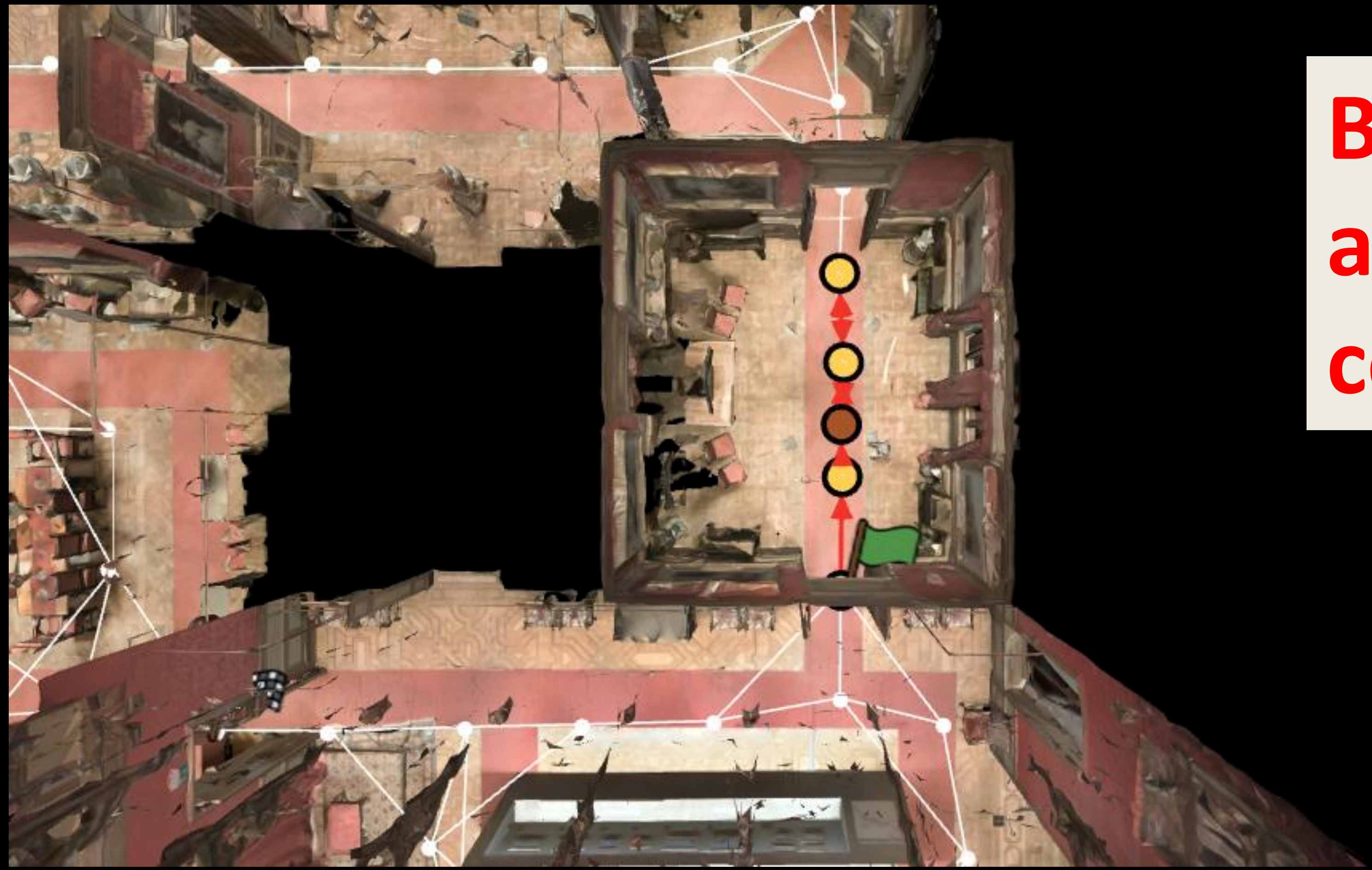


Instruction: Exit the roped off hall, follow the red carpet, **turn right**, continue straight down the red carpet, enter room at the end, stop once inside the room.



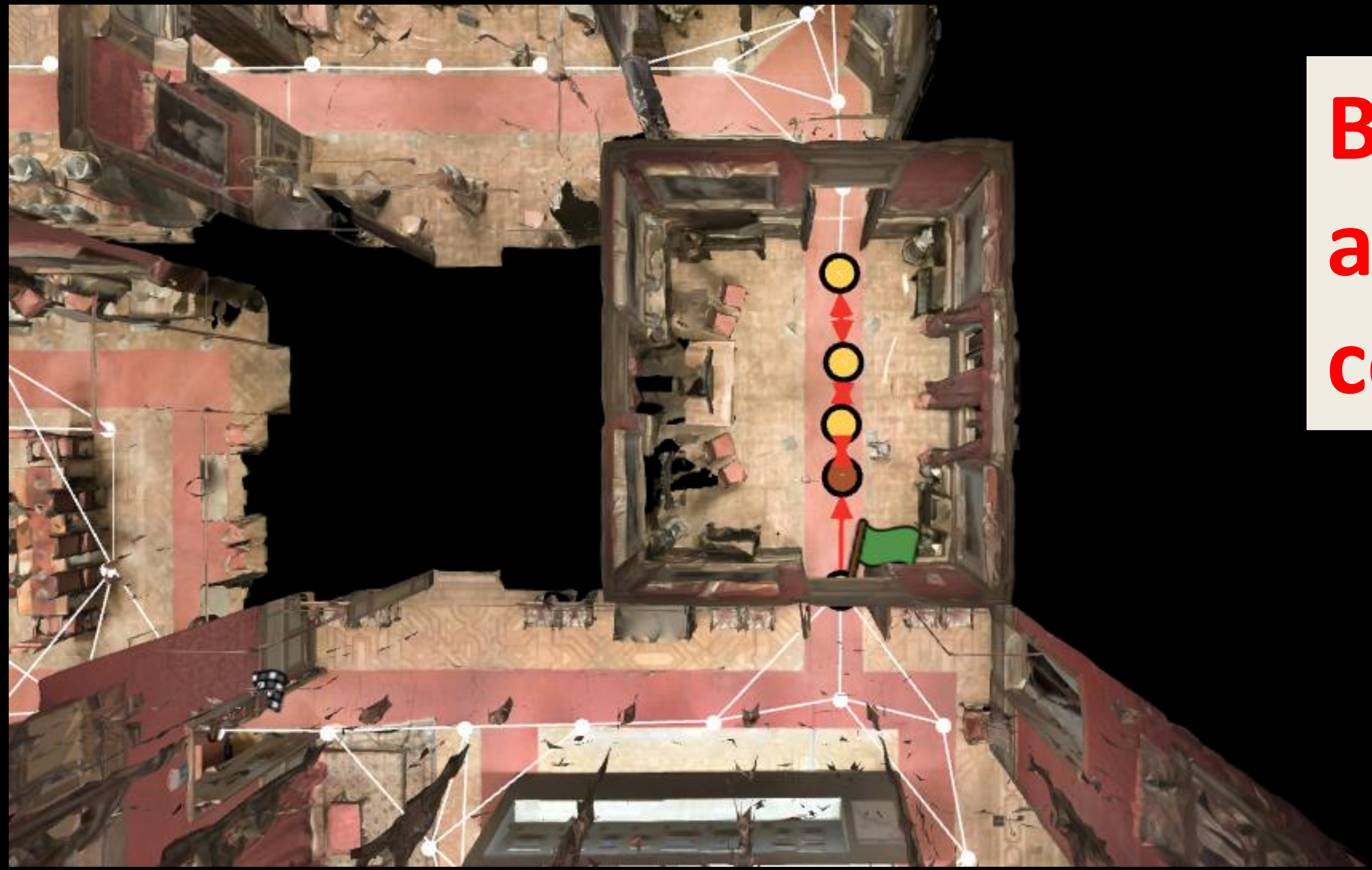
Cannot turn right.
Back Track

Instruction: Exit the roped off hall, follow the red carpet, turn right, continue straight down the red carpet, enter room at the end, stop once inside the room.



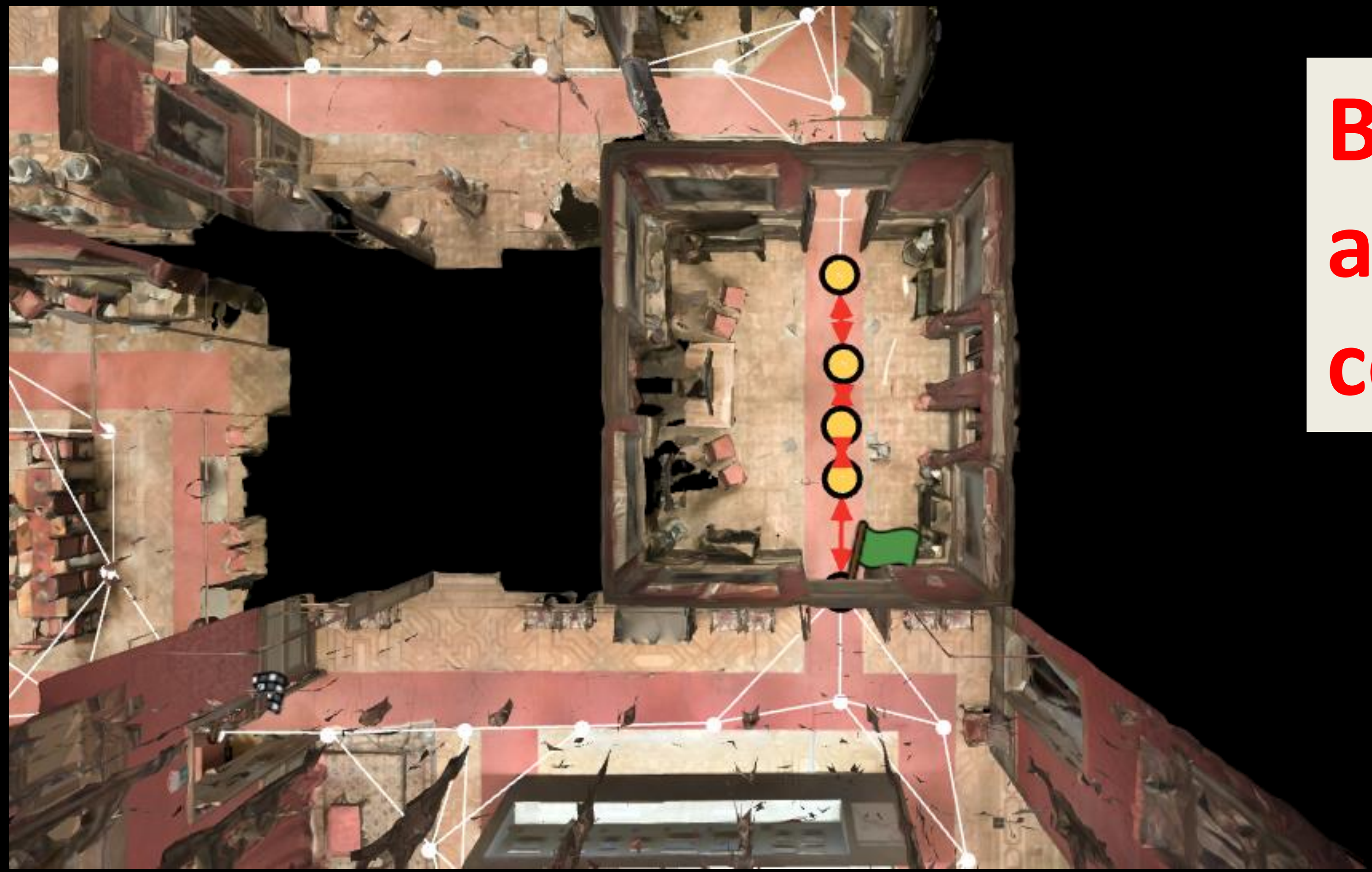
Back tracking according to the constructed map.

Instruction: Exit the roped off hall, follow the red carpet, turn right, continue straight down the red carpet, enter room at the end, stop once inside the room.



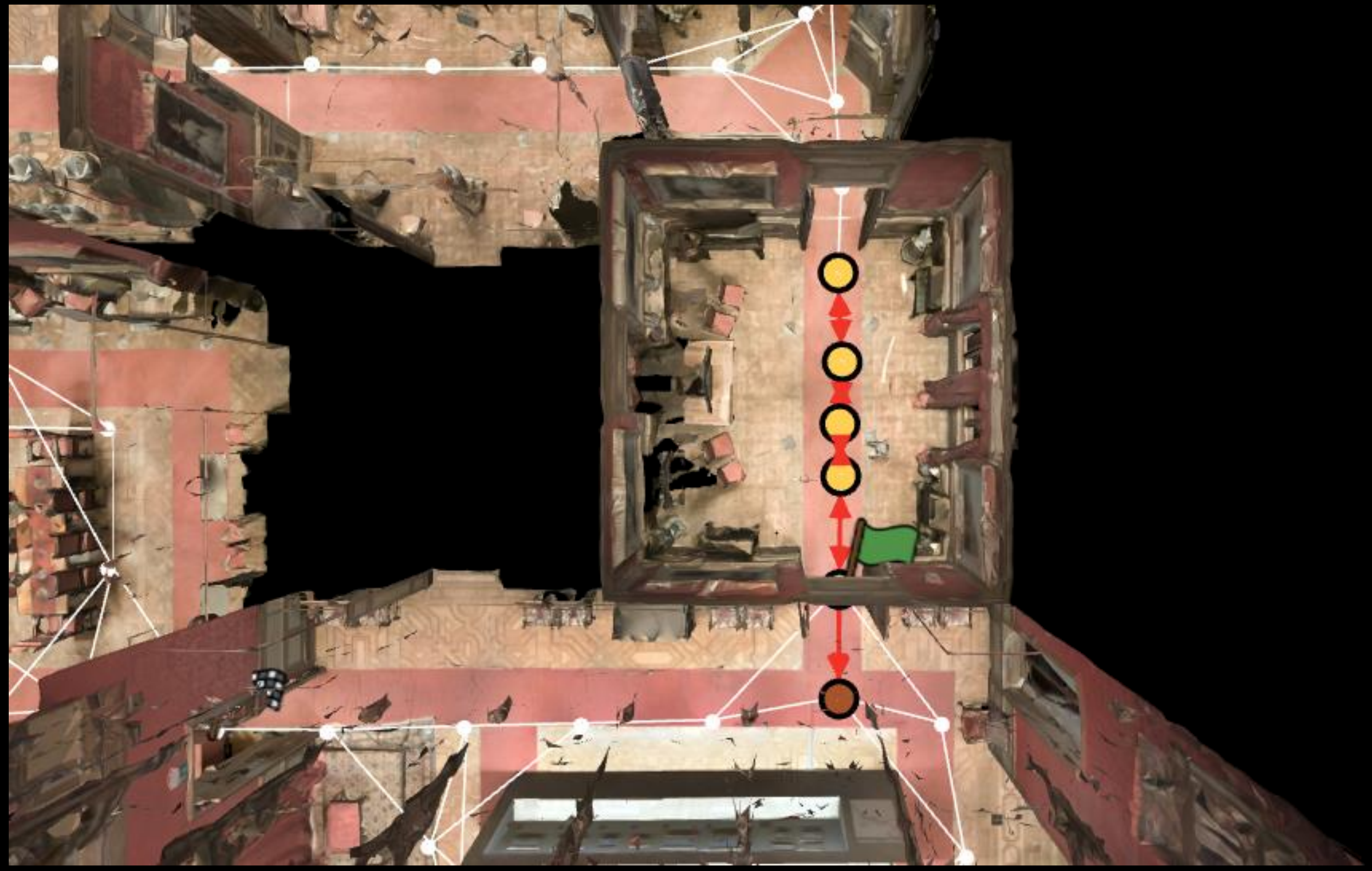
Back tracking according to the constructed map.

Instruction: Exit the roped off hall, follow the red carpet, turn right, continue straight down the red carpet, enter room at the end, stop once inside the room.

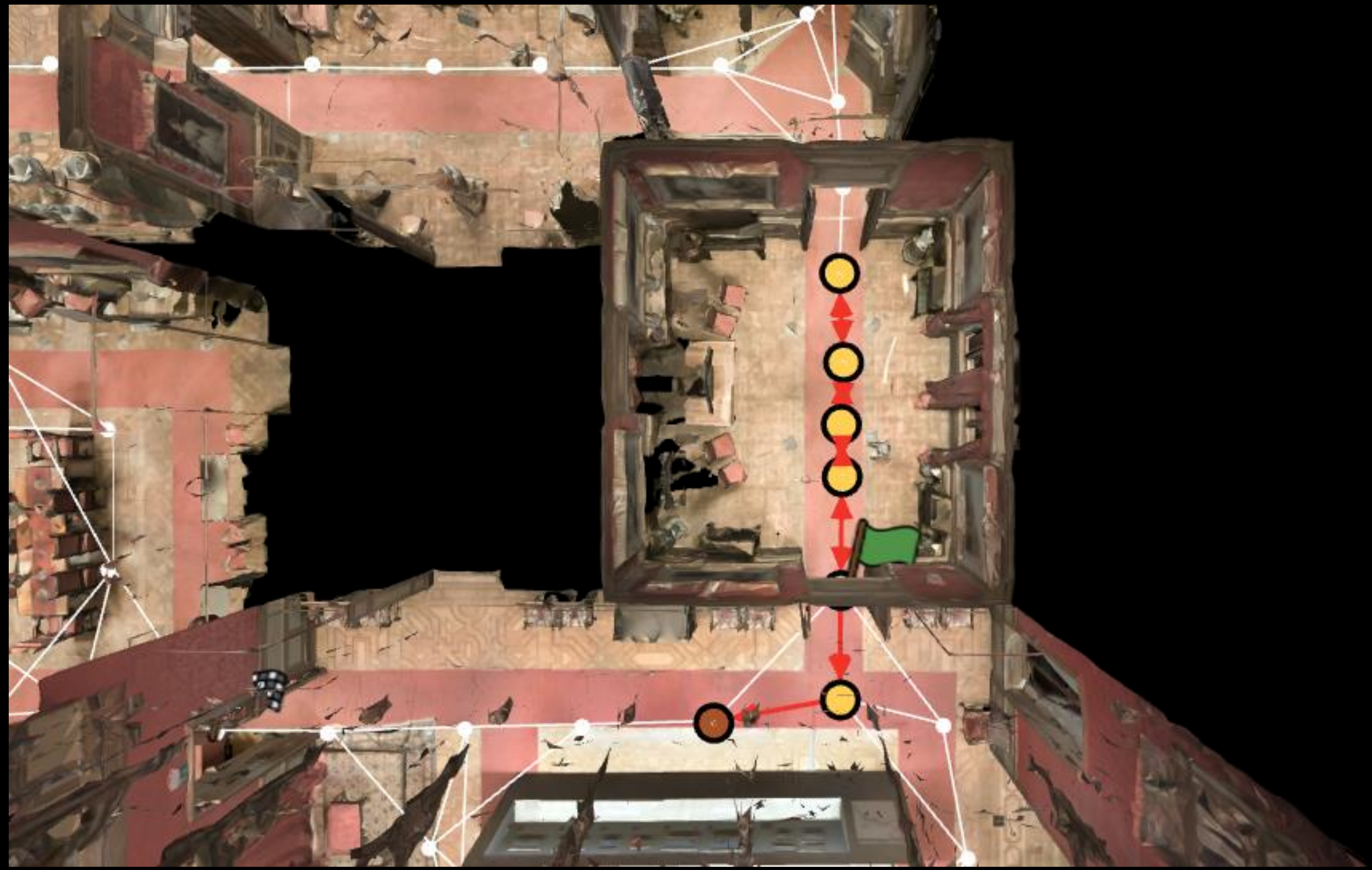


Back tracking according to the constructed map.

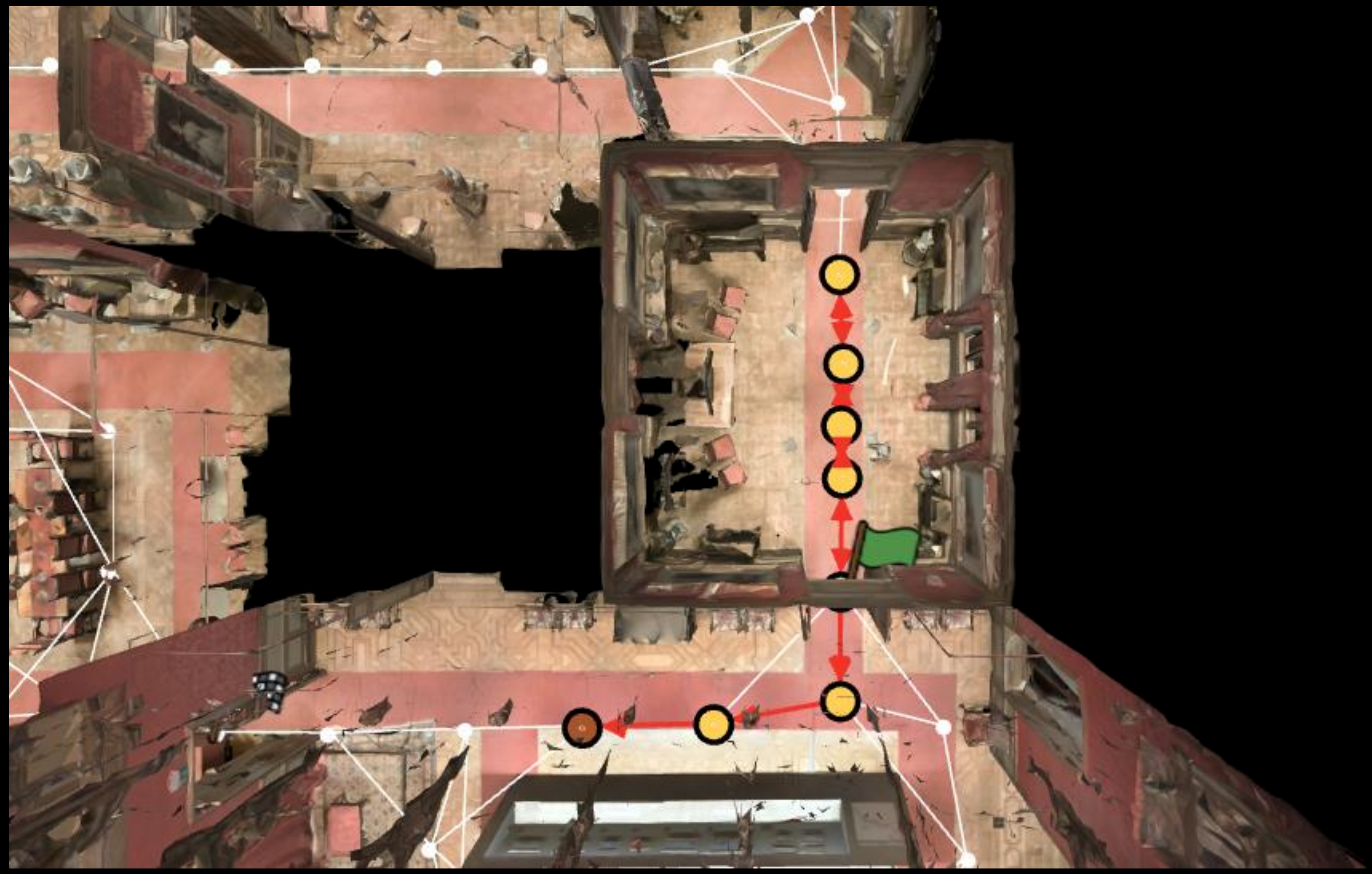
Instruction: **Exit the roped off hall, follow the red carpet**, turn right, continue straight down the red carpet, enter room at the end, stop once inside the room.



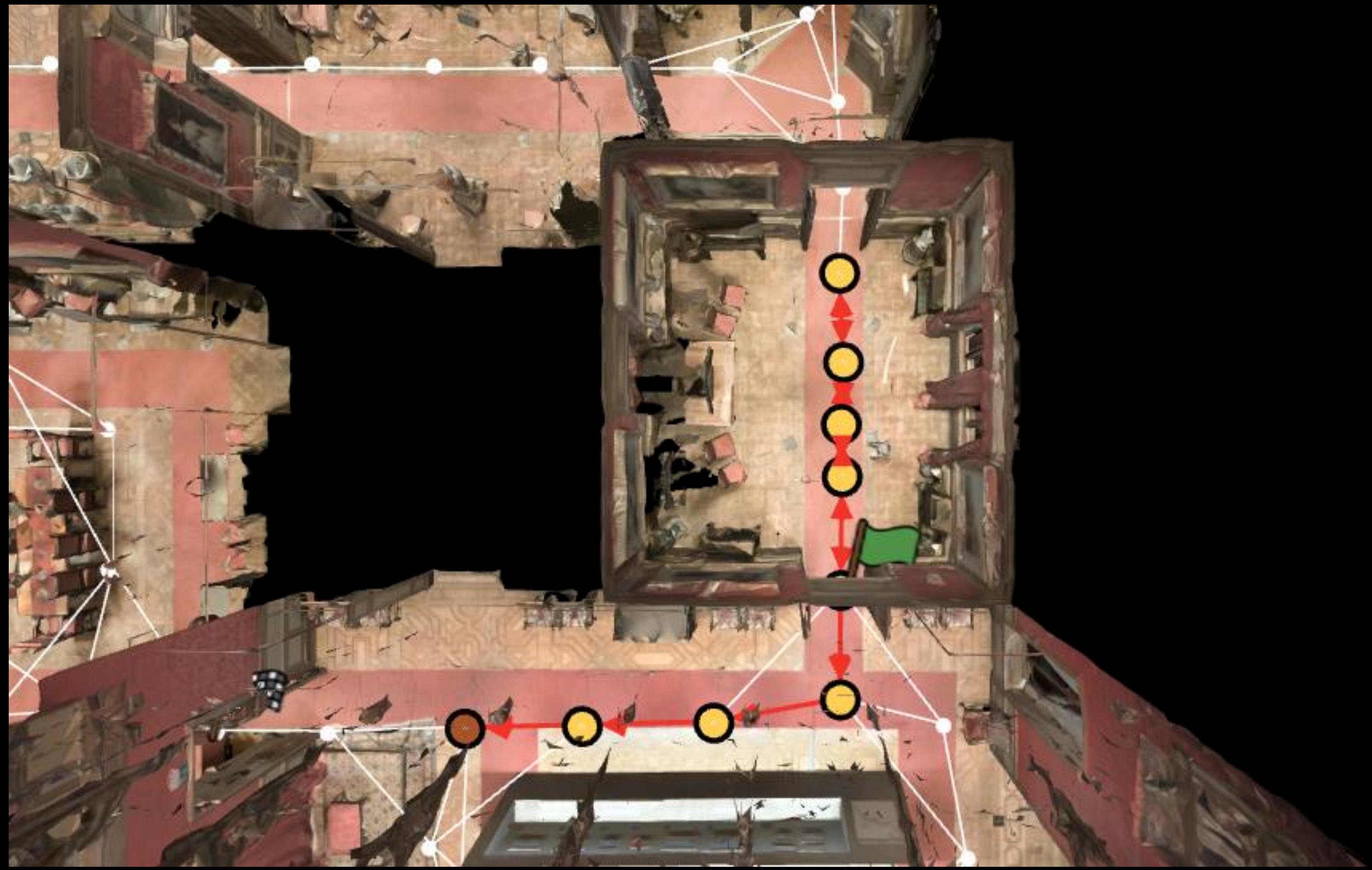
Instruction: Exit the roped off hall, follow the red carpet, **turn right**, continue straight down the red carpet, enter room at the end, stop once inside the room.



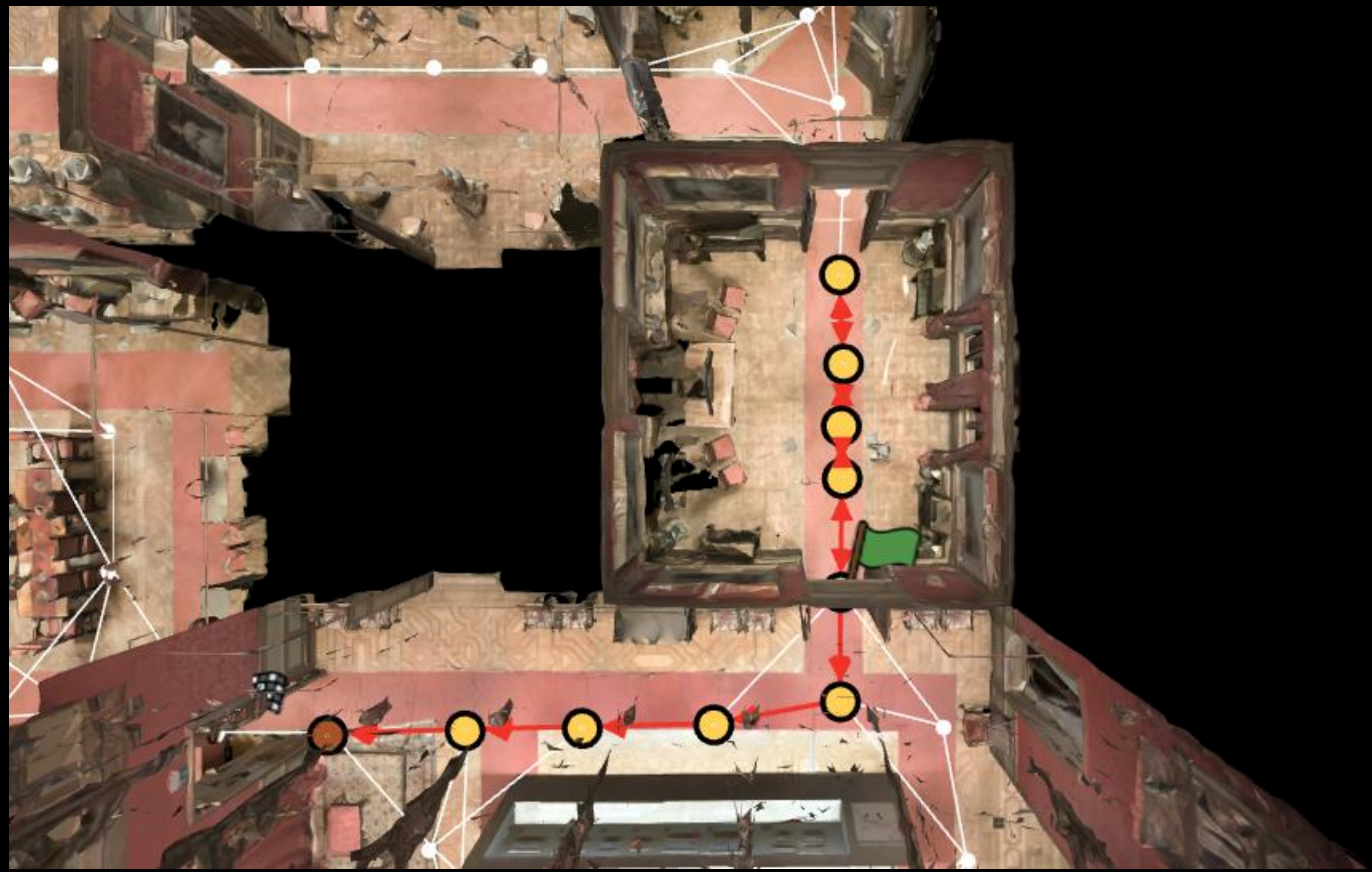
Instruction: Exit the roped off hall, follow the red carpet, turn right, **continue straight down the red carpet,** enter room at the end, stop once inside the room.



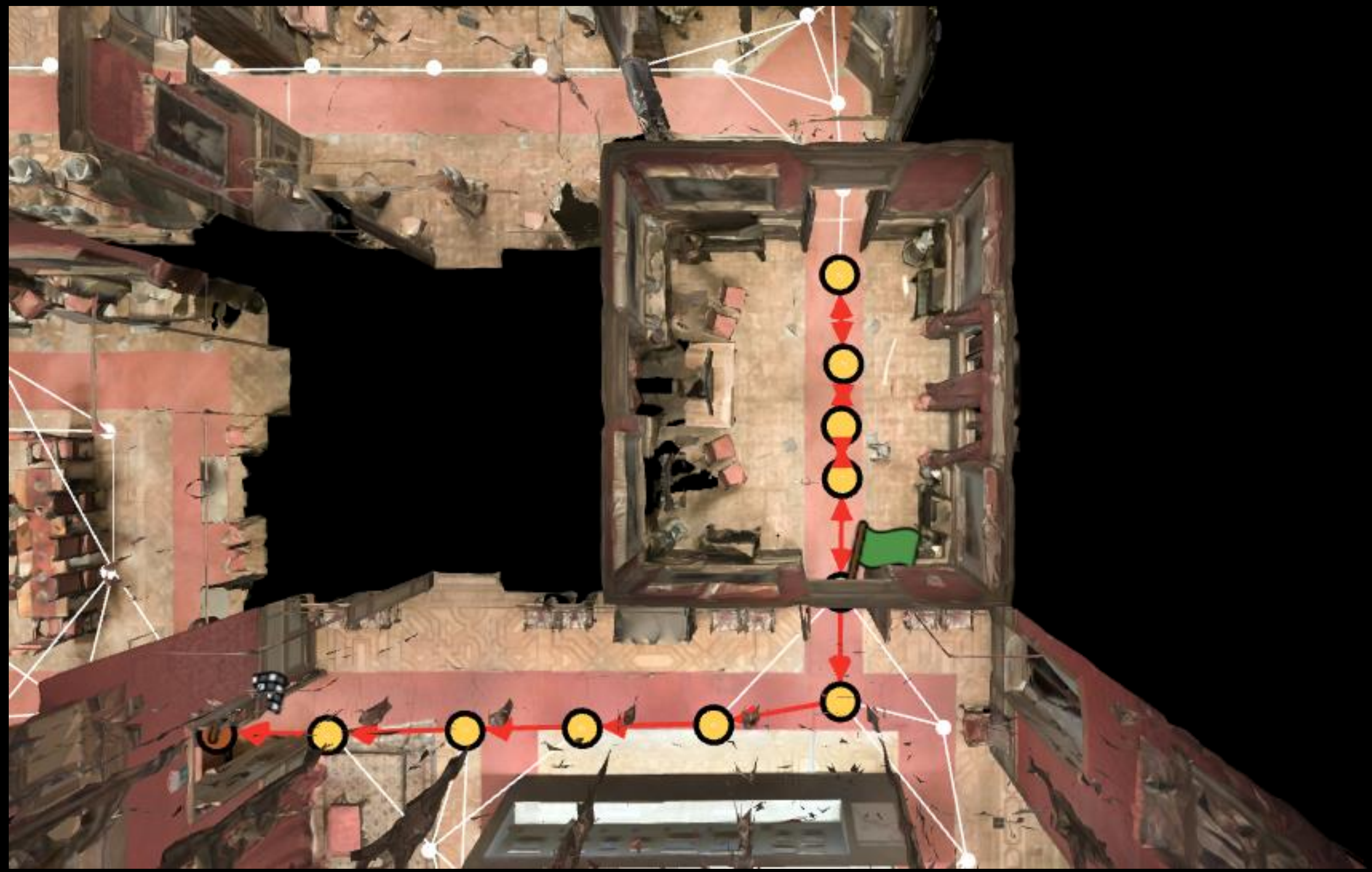
Instruction: Exit the roped off hall, follow the red carpet, turn right, **continue straight down the red carpet,** enter room at the end, stop once inside the room.



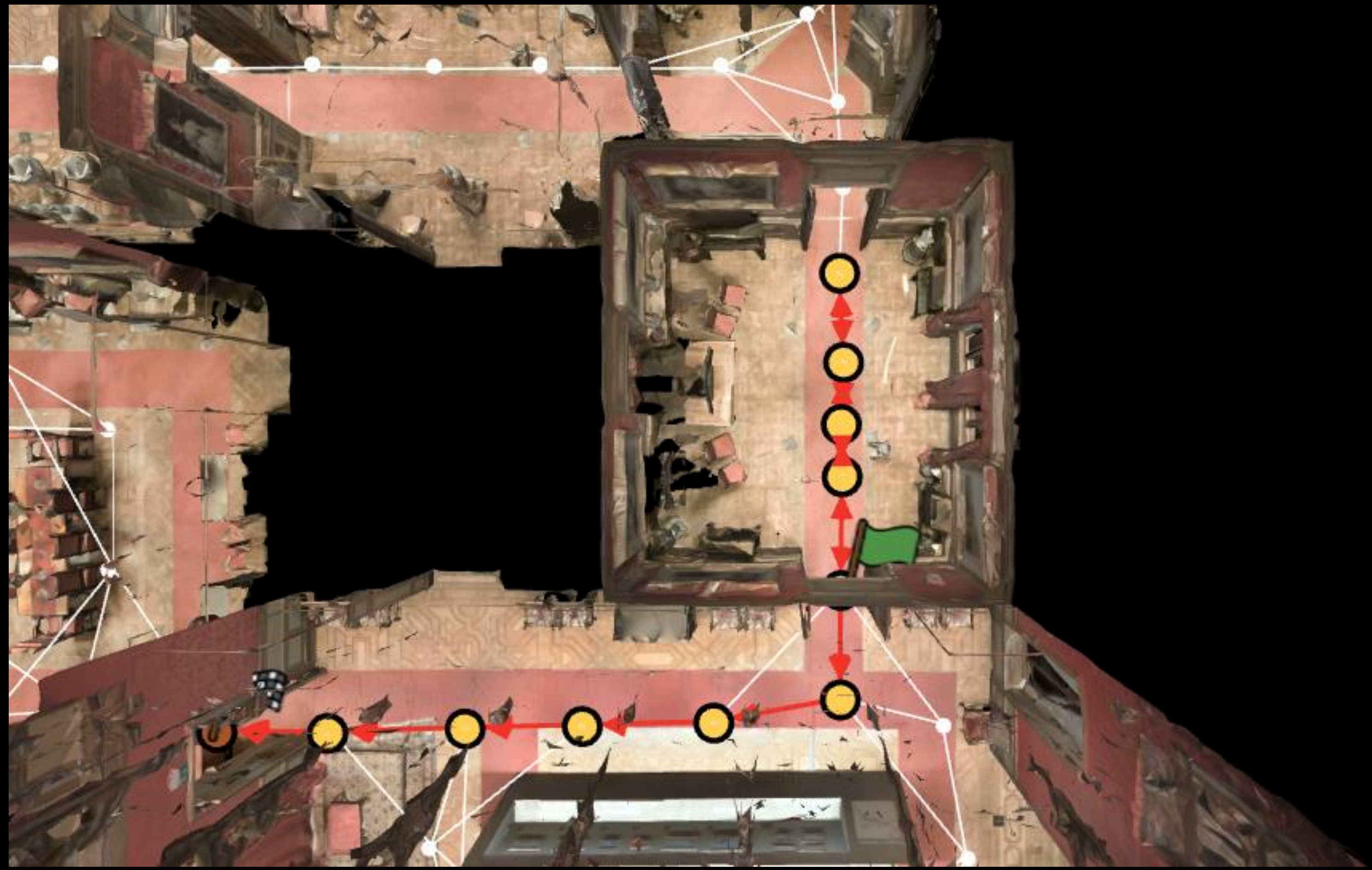
Instruction: Exit the roped off hall, follow the red carpet, turn right, **continue straight down the red carpet,** enter room at the end, stop once inside the room.



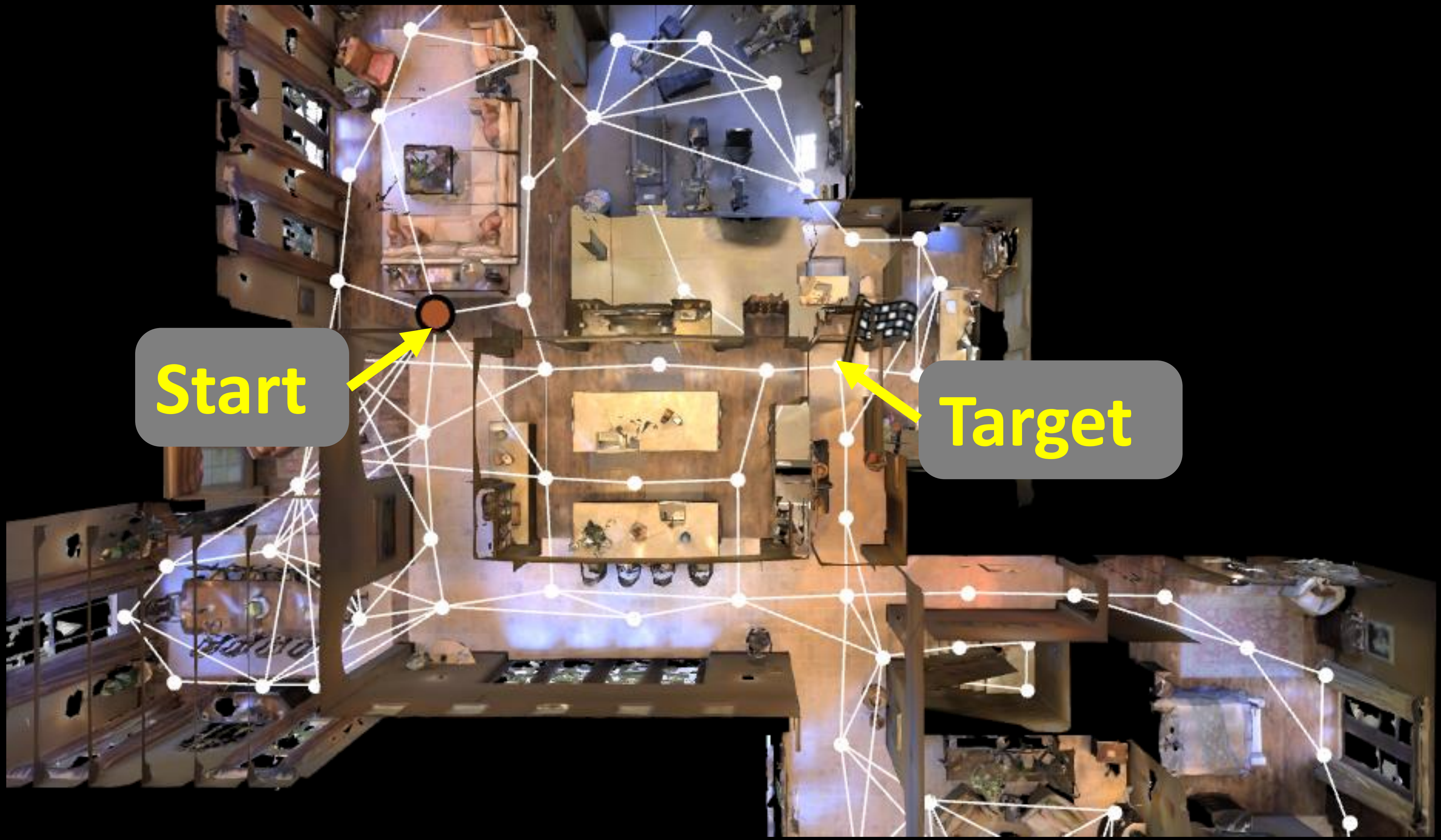
Instruction: Exit the roped off hall, follow the red carpet, turn right, continue straight down the red carpet, **enter room at the end, stop once inside the room.**



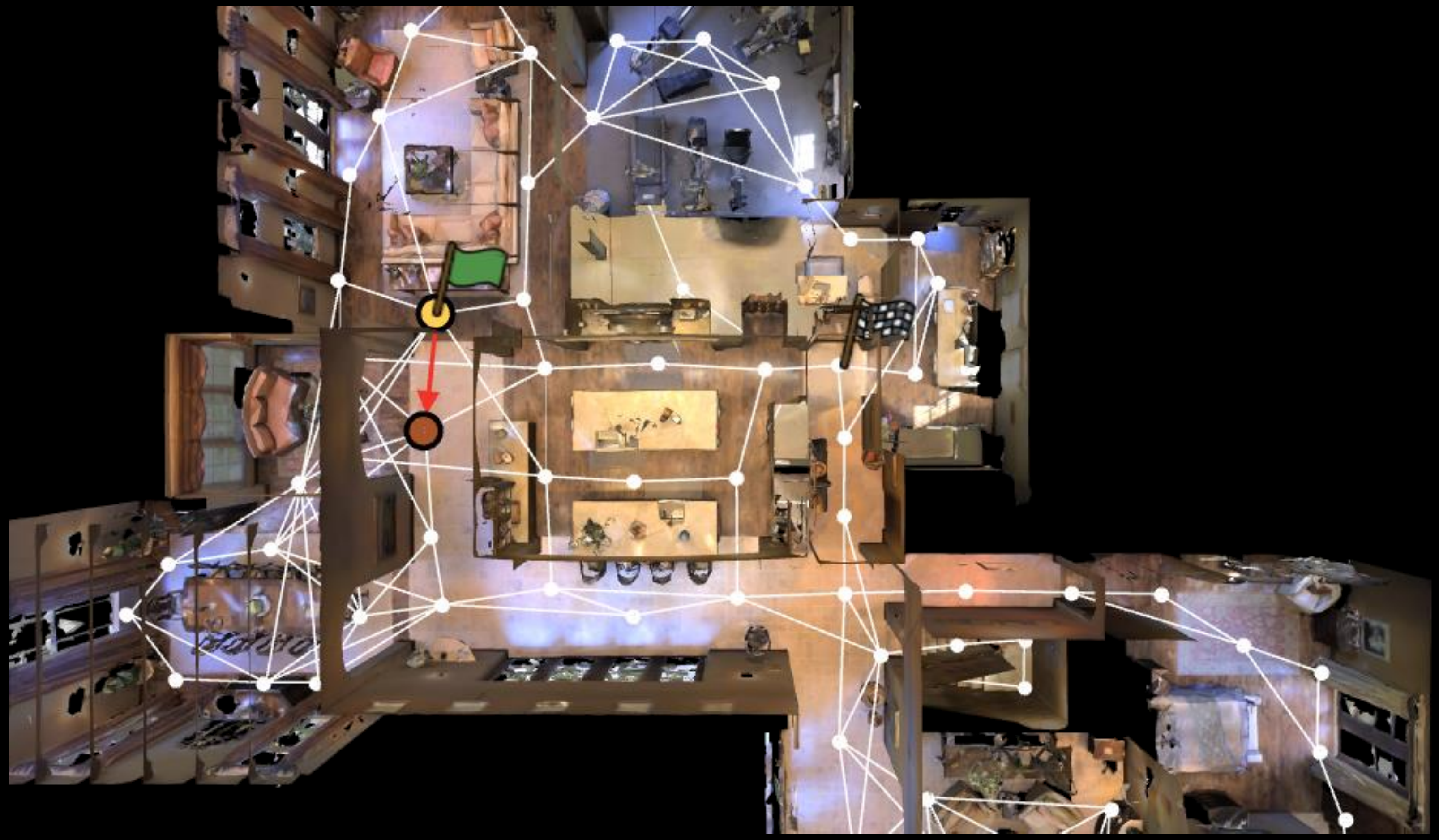
Instruction: Exit the roped off hall, follow the red carpet, turn right, continue straight down the red carpet, **enter room at the end, stop once inside the room.**



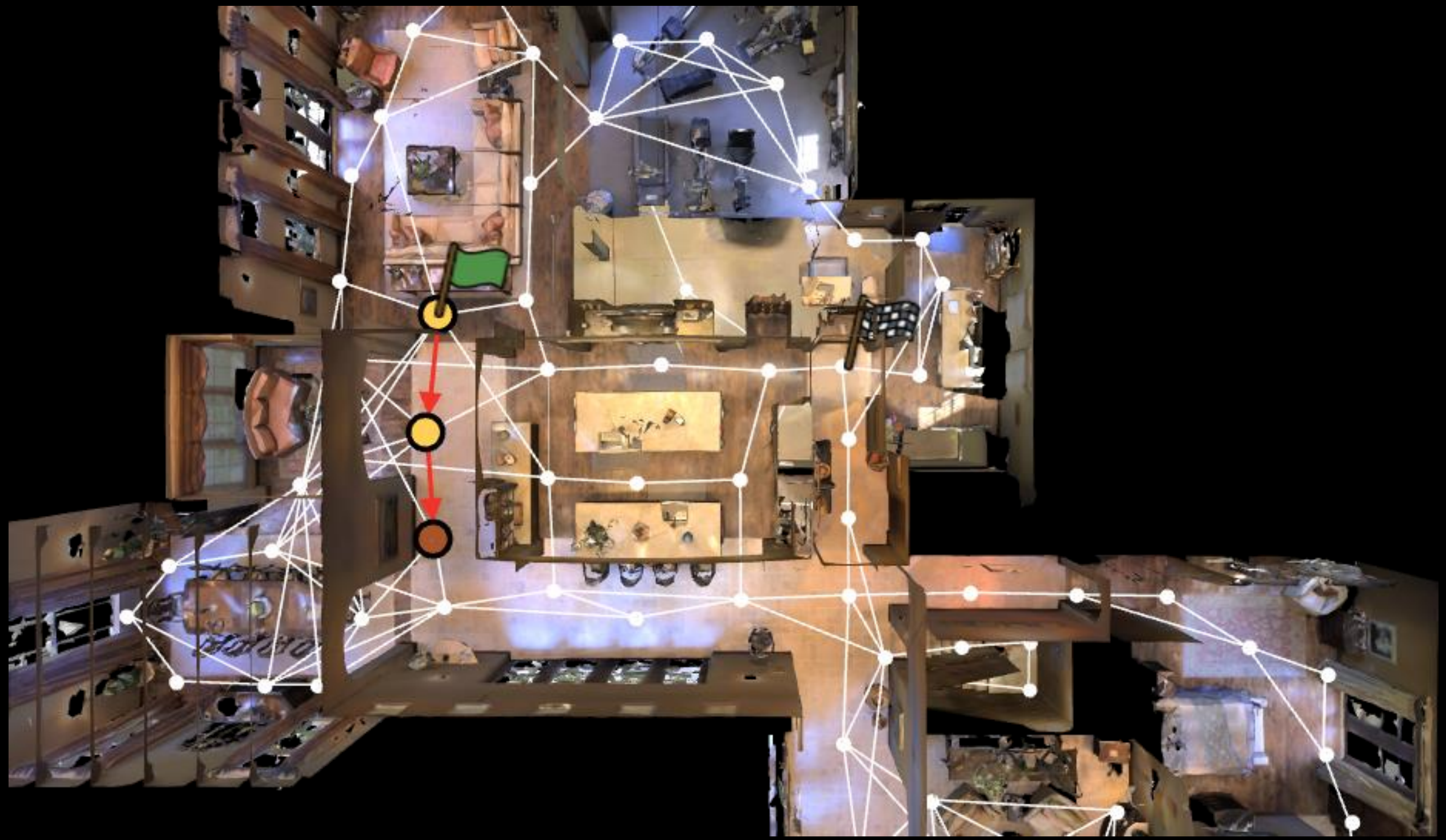
Instruction: Go to second level hallway next to the kitchen and clean the photo above the black bench and that is closest to the kitchen.



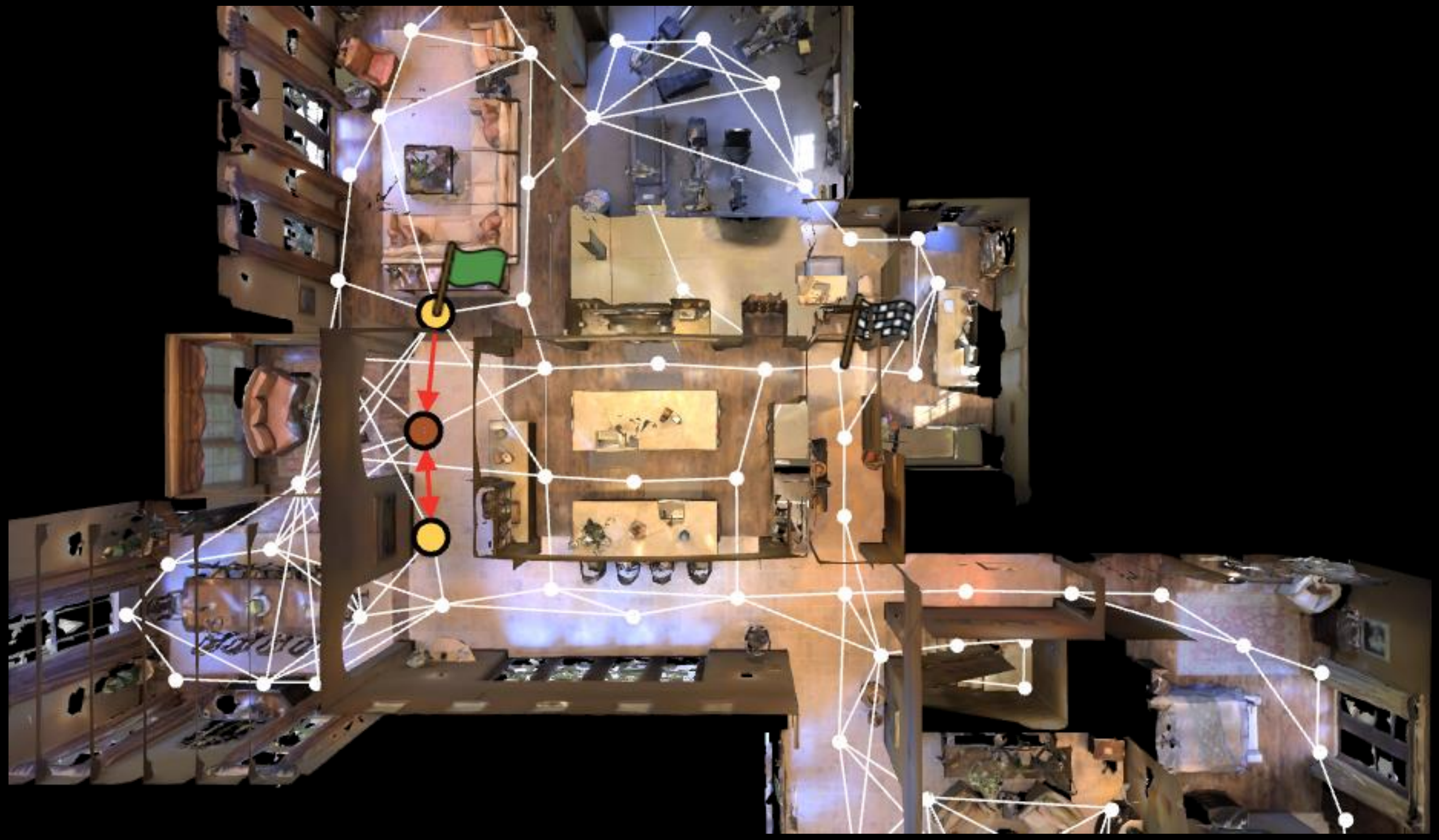
Instruction: Go to second level hallway next to the kitchen and clean the photo above the black bench and that is closest to the kitchen.



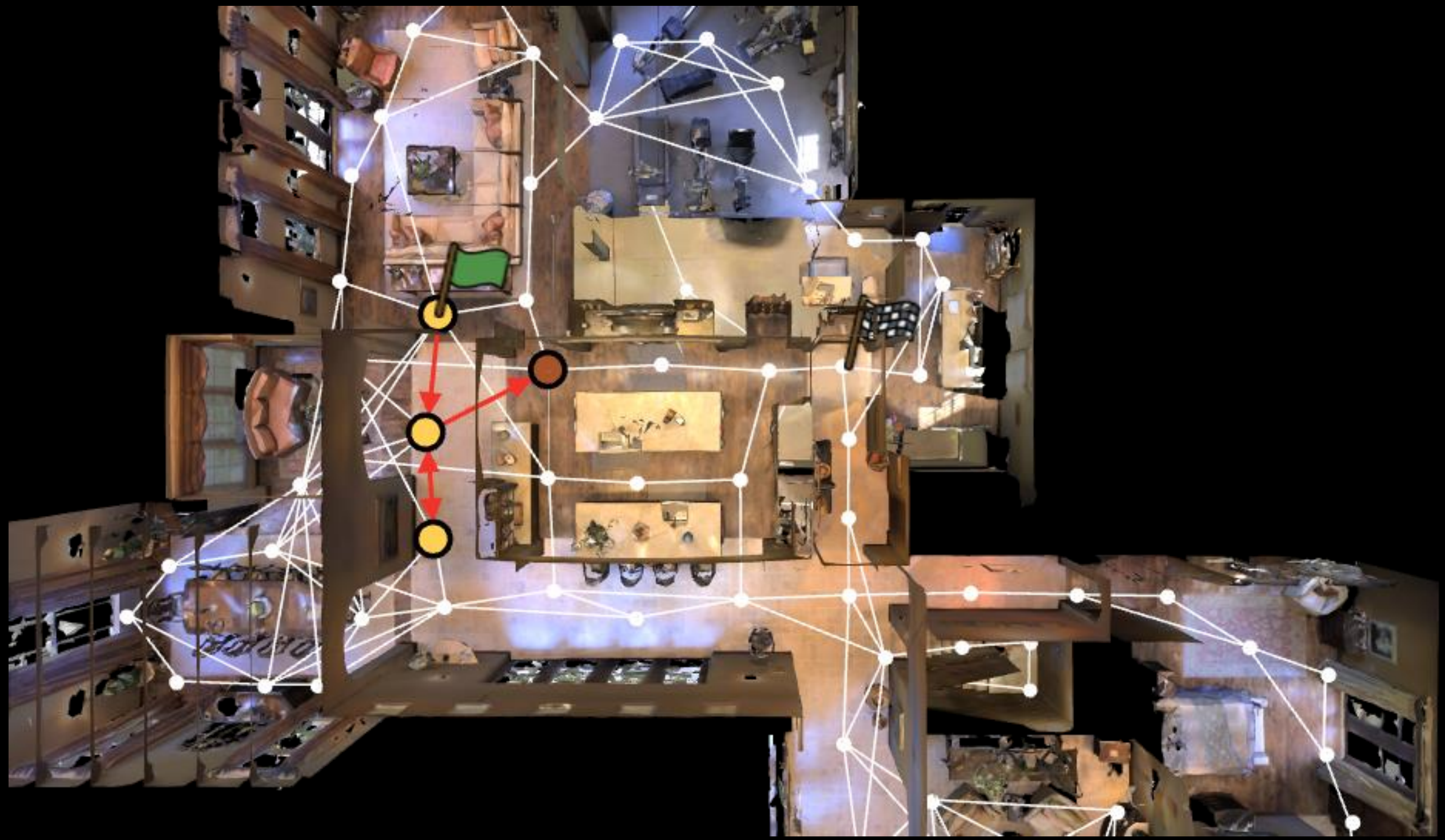
Instruction: Go to second level hallway next to the kitchen and clean the photo above the black bench and that is closest to the kitchen.



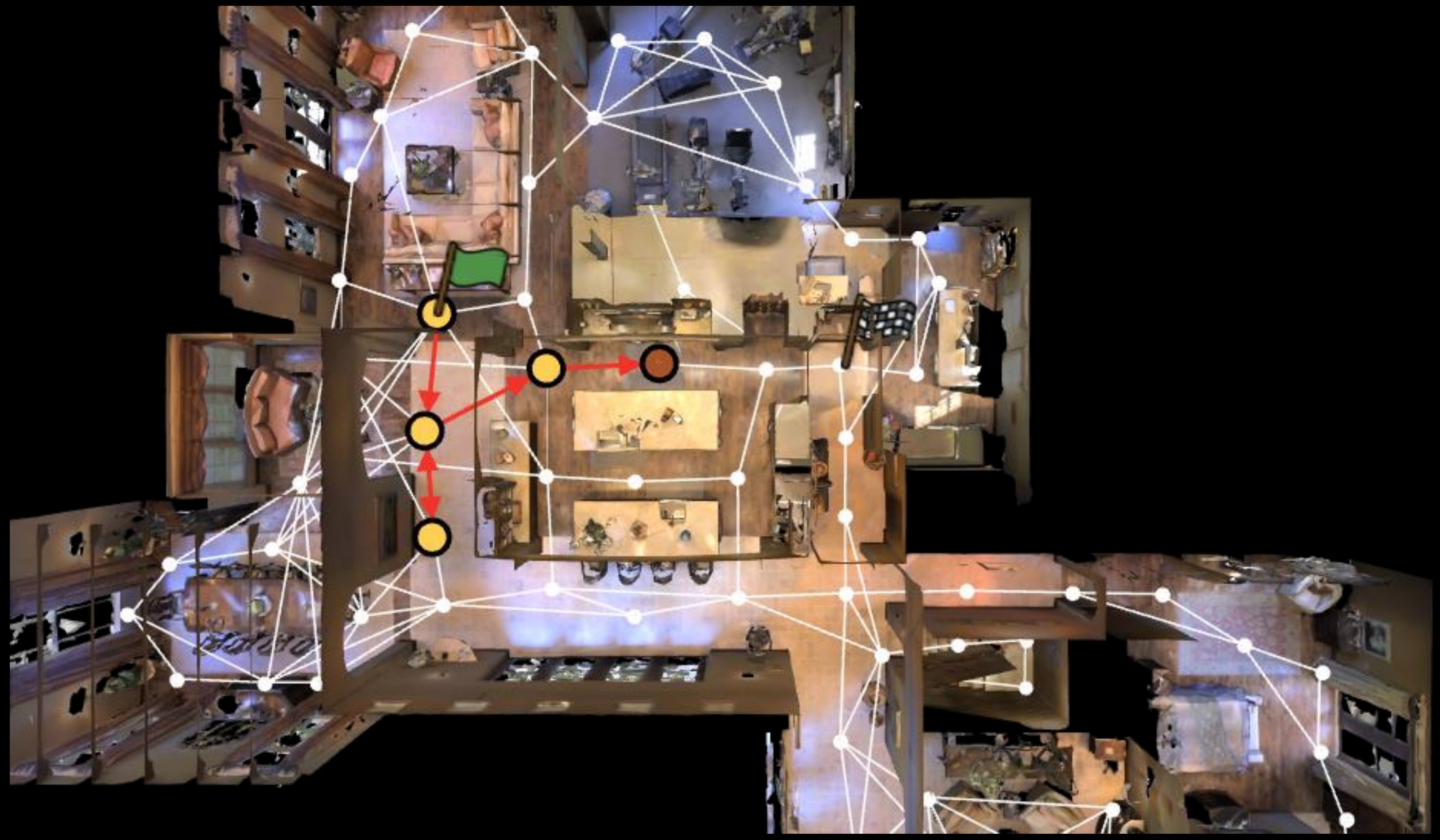
Instruction: Go to second level hallway next to the kitchen and clean the photo above the black bench and that is closest to the kitchen.



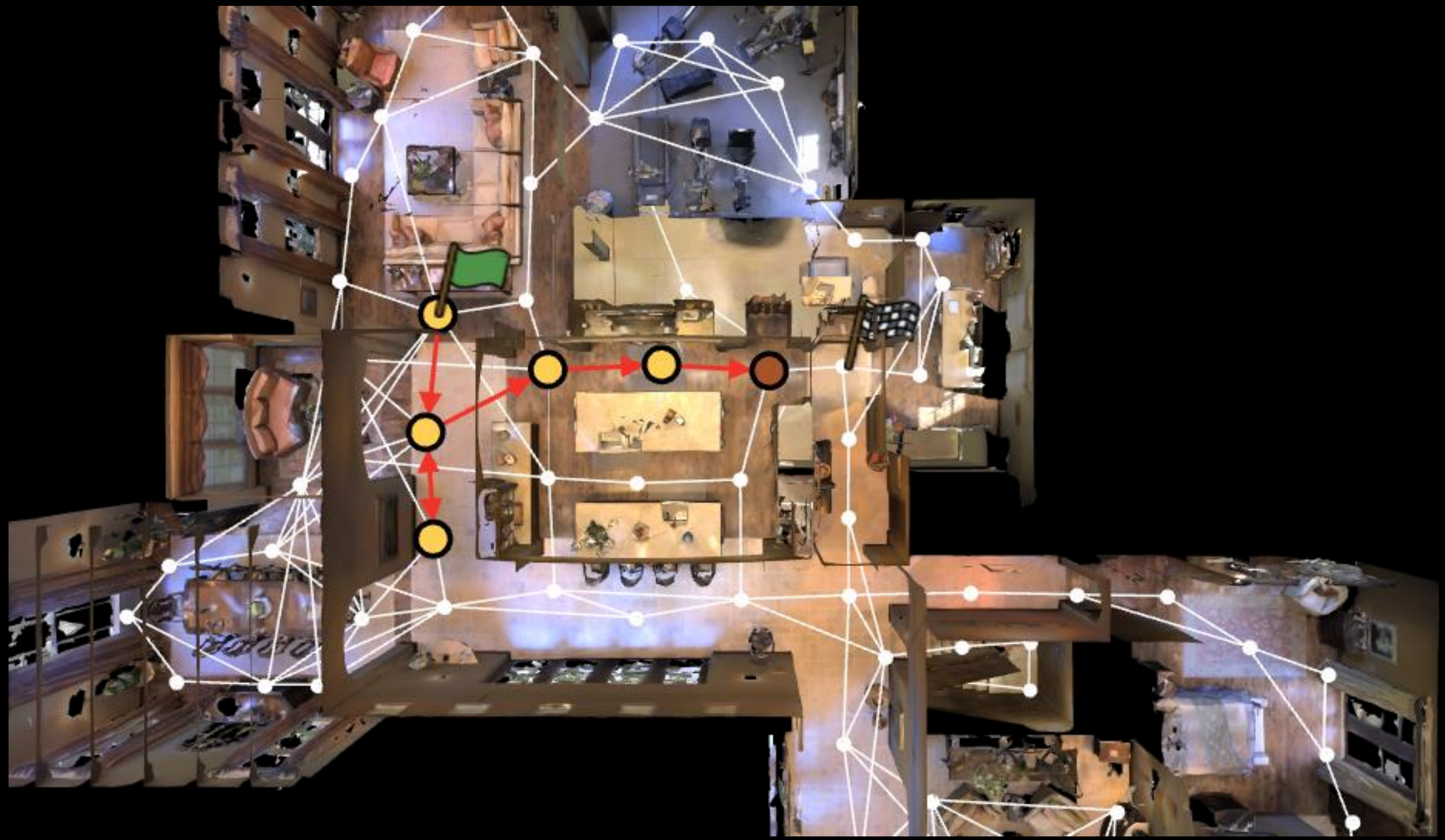
Instruction: Go to second level hallway next to the kitchen and clean the photo above the black bench and that is closest to the kitchen.



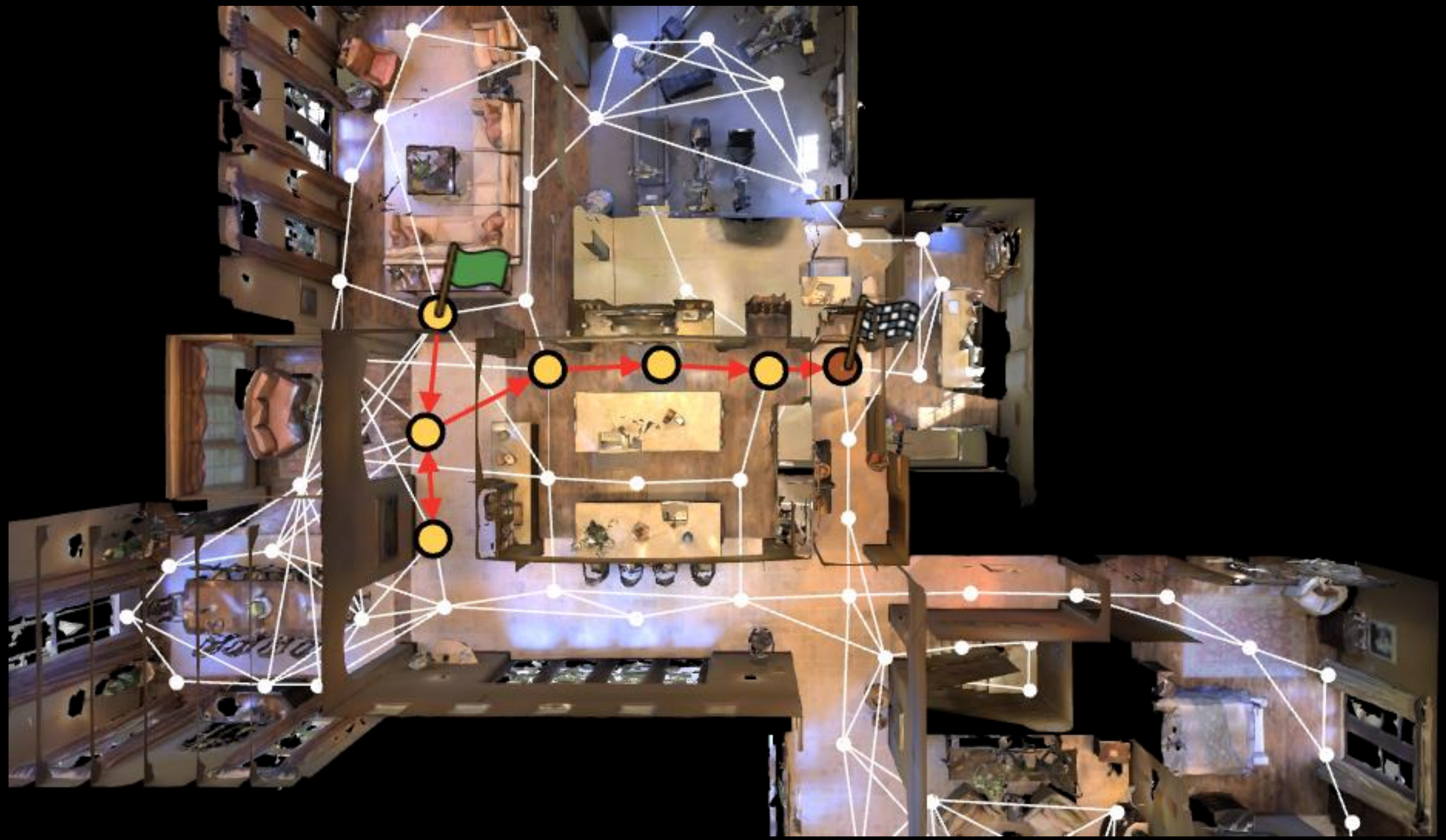
Instruction: Go to second level hallway next to the kitchen and clean the photo above the black bench and that is closest to the kitchen.



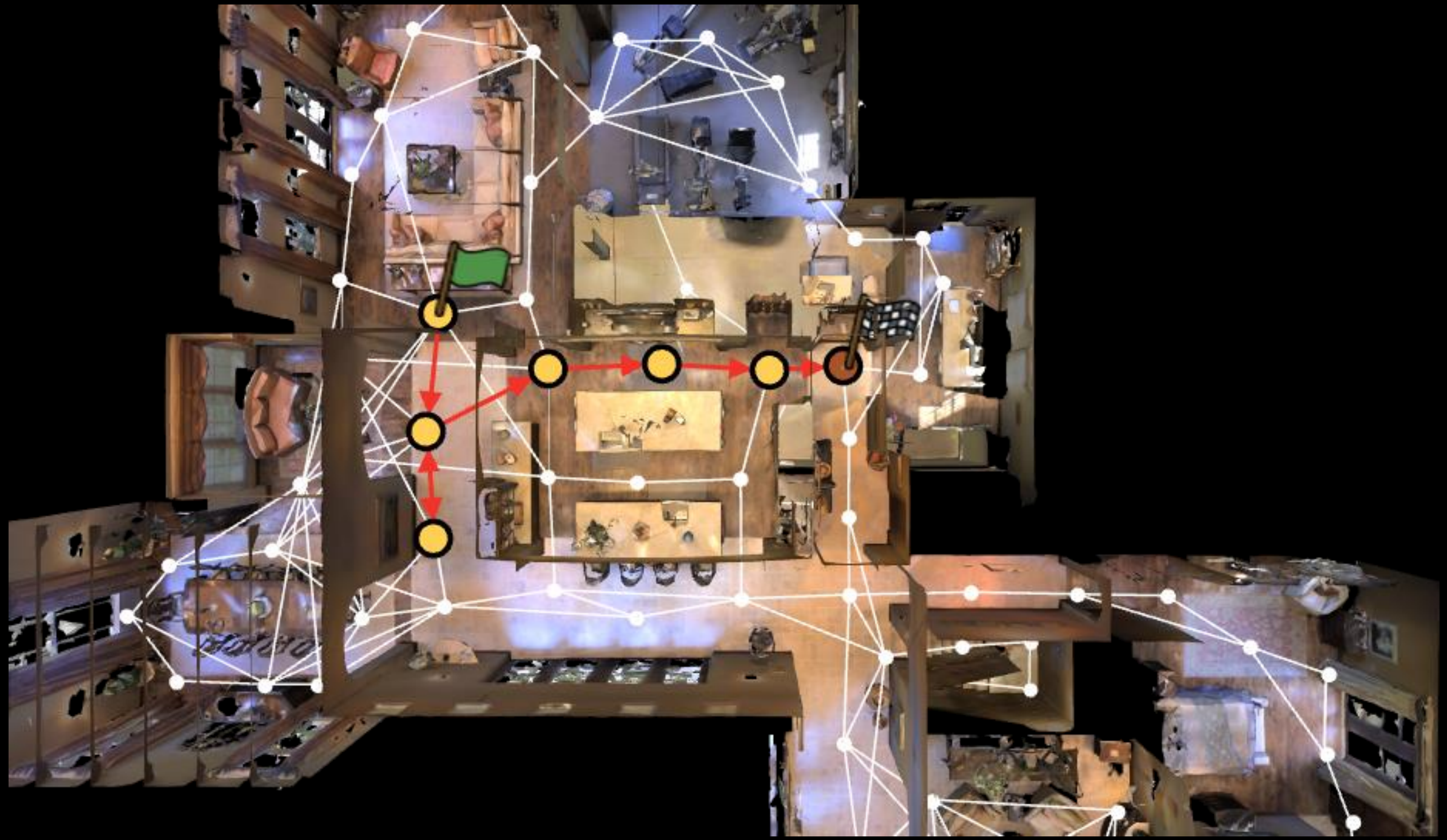
Instruction: Go to second level hallway next to the kitchen and clean the photo above the black bench and that is closest to the kitchen.



Instruction: Go to second level hallway next to the kitchen and clean the photo above the black bench and that is closest to the kitchen.



Instruction: Go to second level hallway next to the kitchen and clean the photo above the black bench and that is closest to the kitchen.



Examples in simulation: successful cases

Target: *"cabinet"*

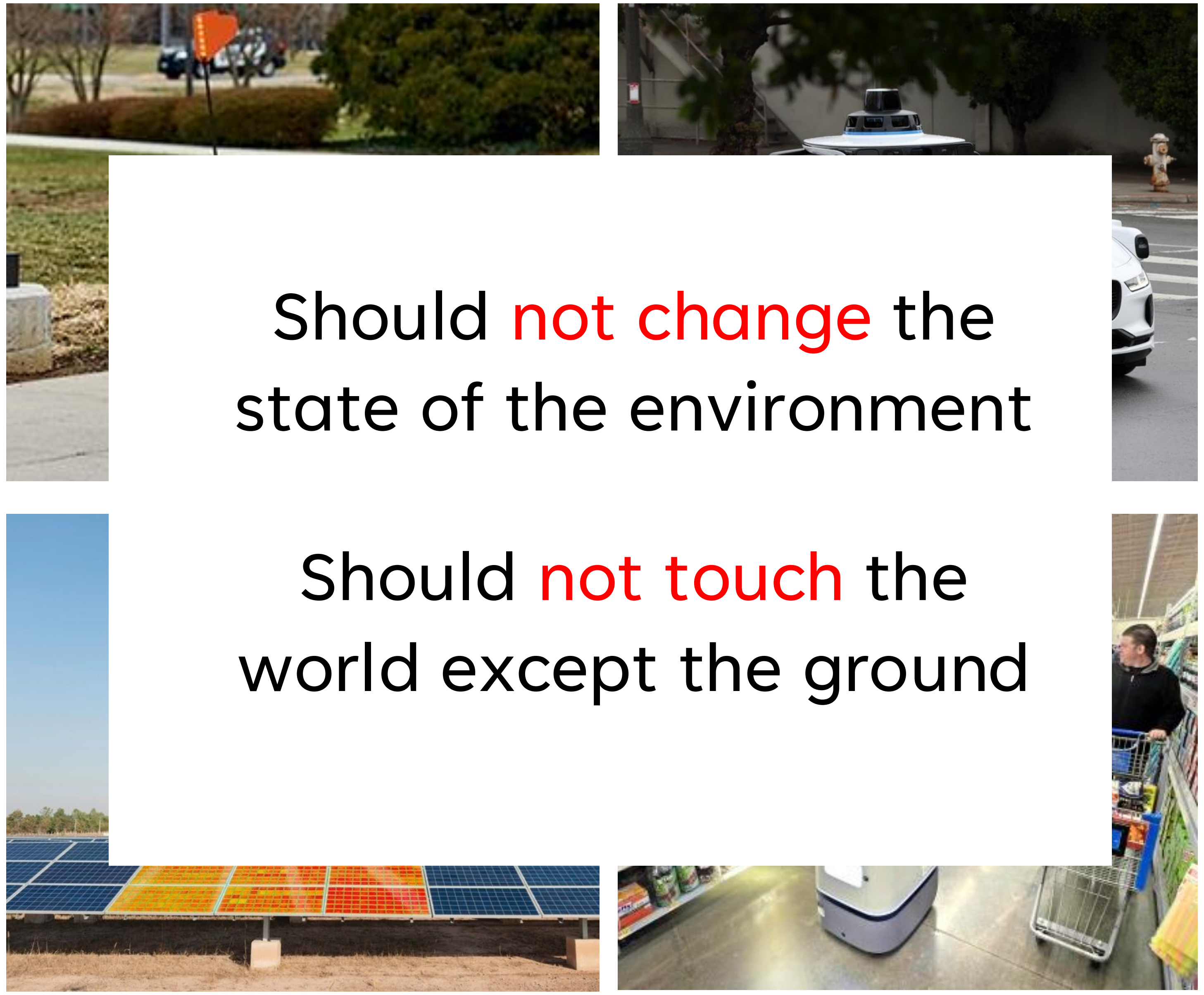
Target: *"chest of drawer"*

Real world examples

Navigation

Manipulation

Navigation vs. Manipulation



Should not change the state of the environment

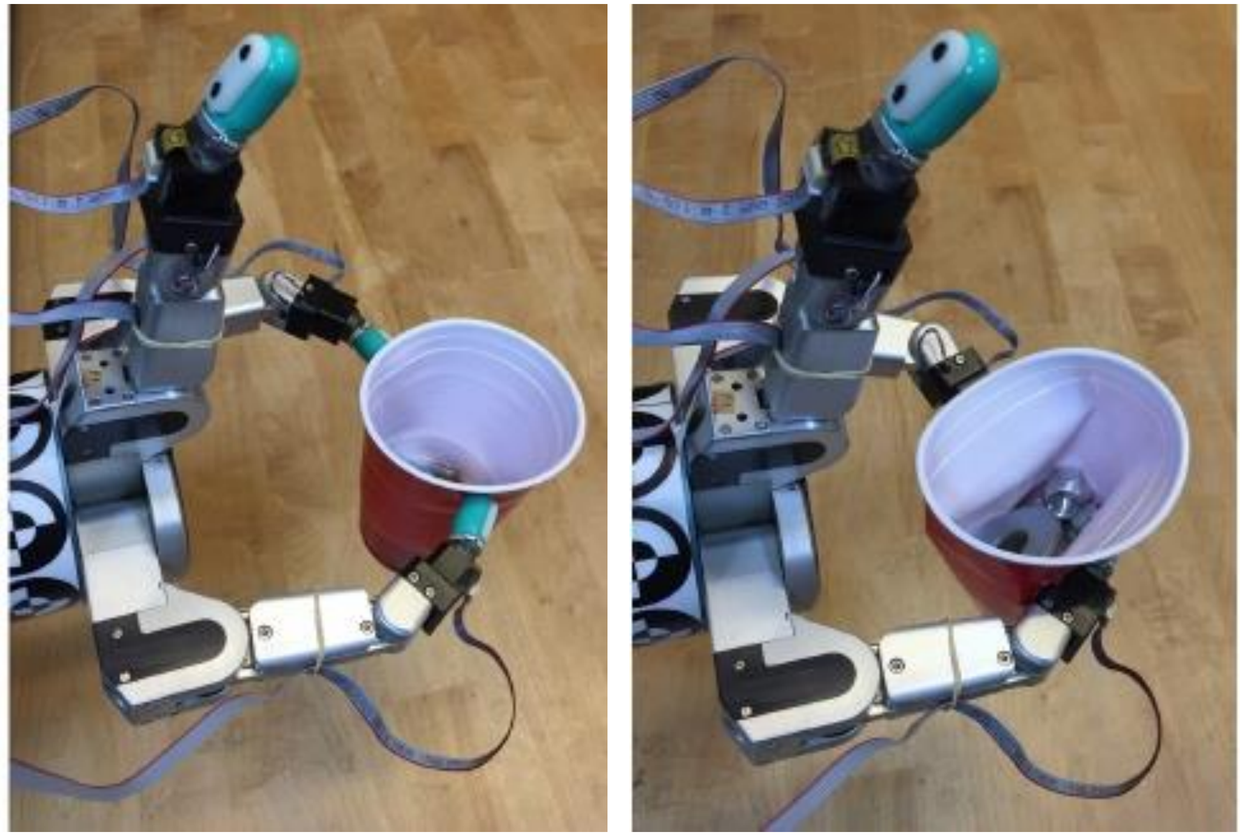
Should not touch the world except the ground

Should change the environment from state A to state B

Physical contacts with the environment are essential

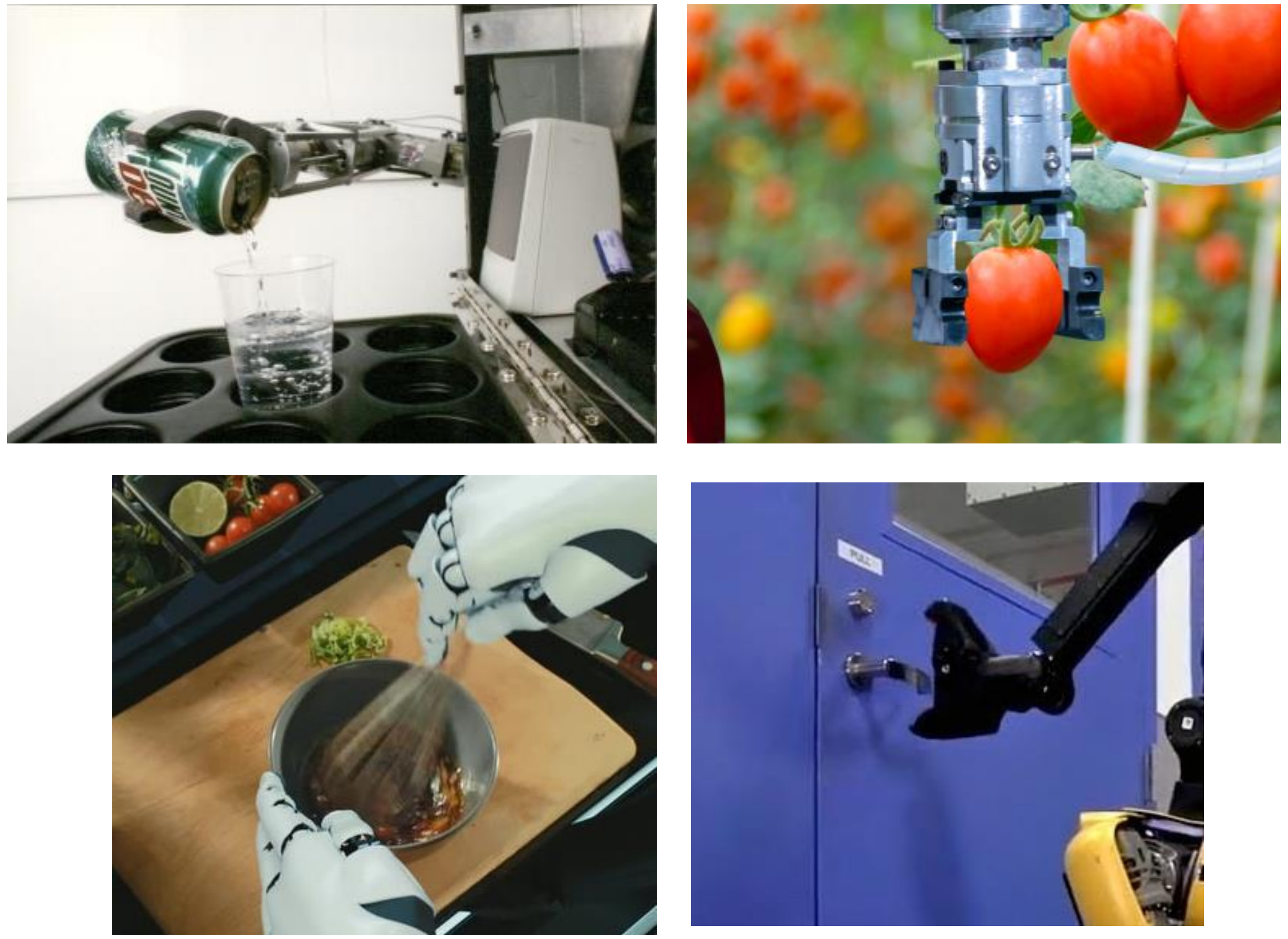
Manipulation Challenges

Results will depend on the gravity, friction, object softness, …



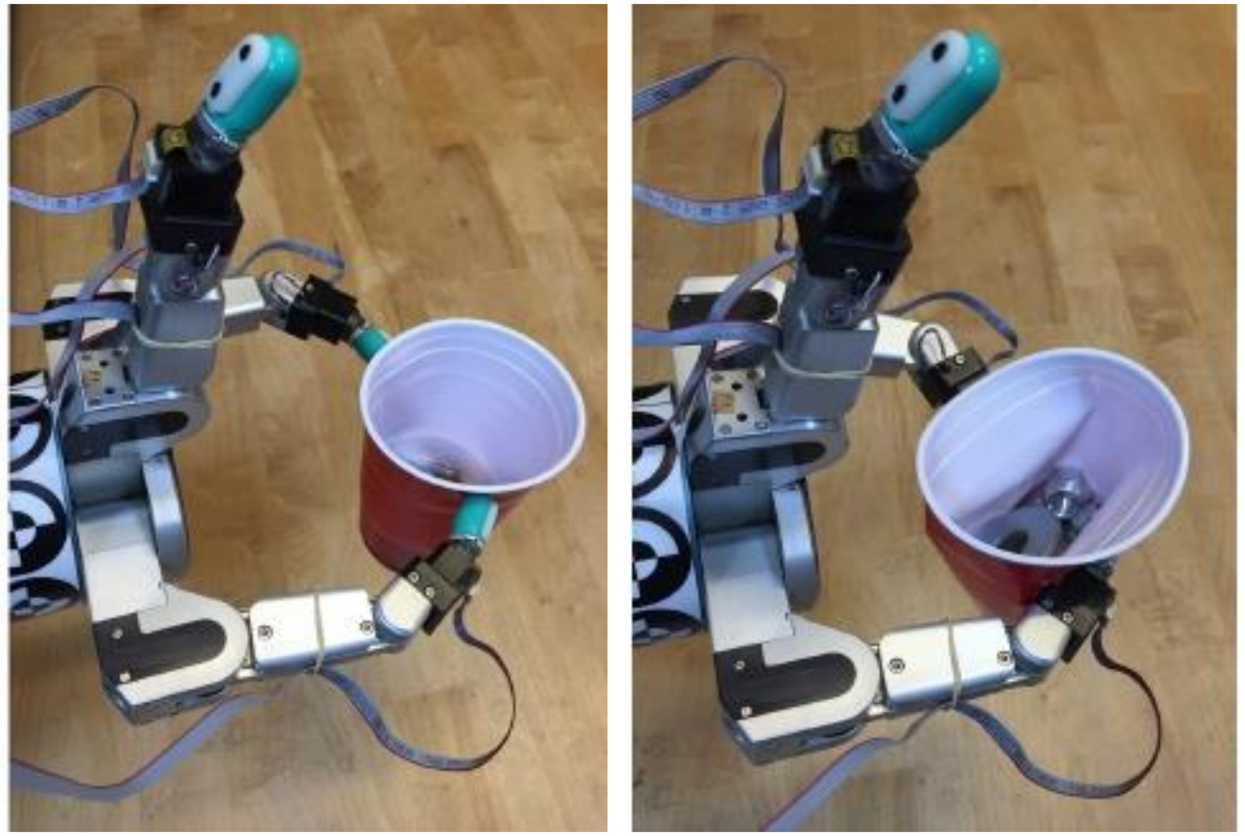
→ **Use physics simulators**

Large action space



Manipulation Challenges

Results will depend on the gravity, friction, object softness, ...

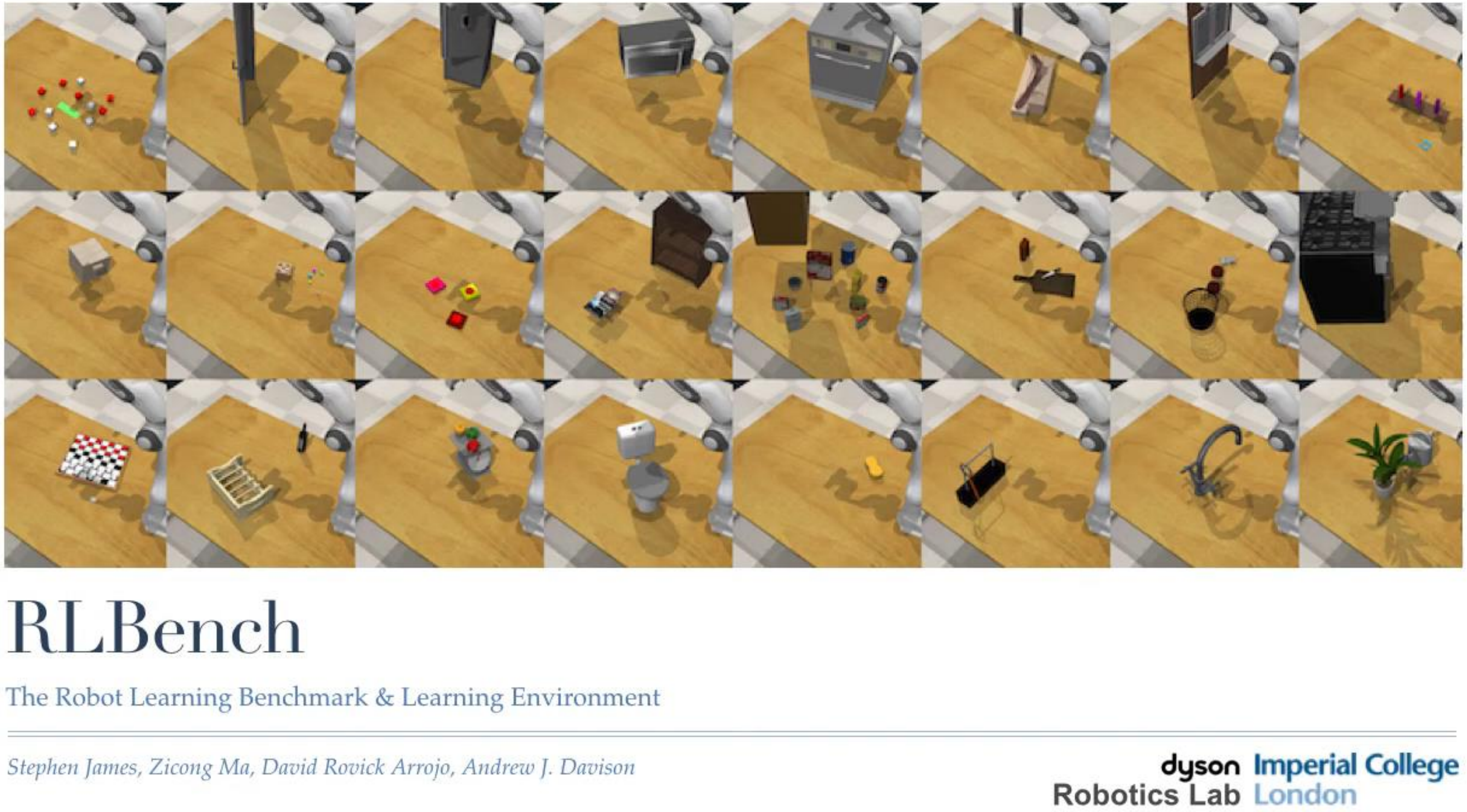


→ **Use physics simulators**

Large action space

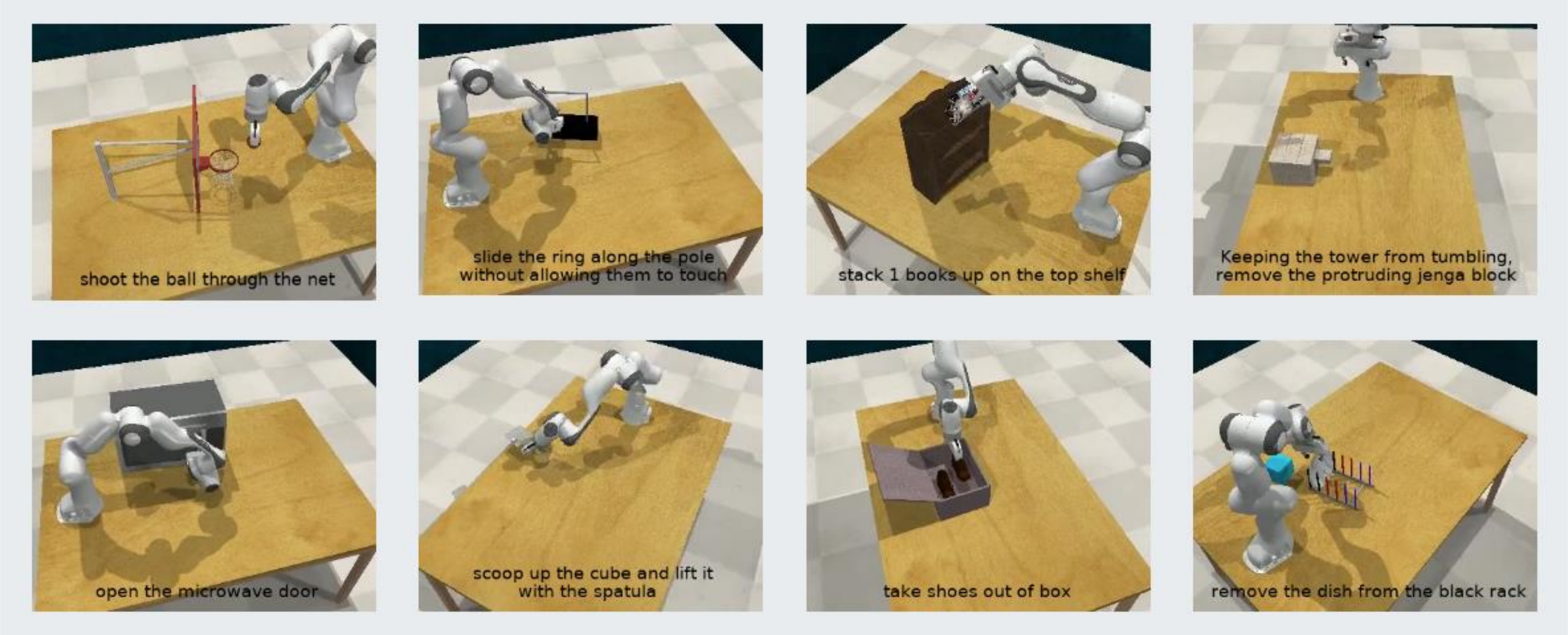
→ **Define tasks by language**

RLBench: Robot Learning Environment James et al., ICRA 2019



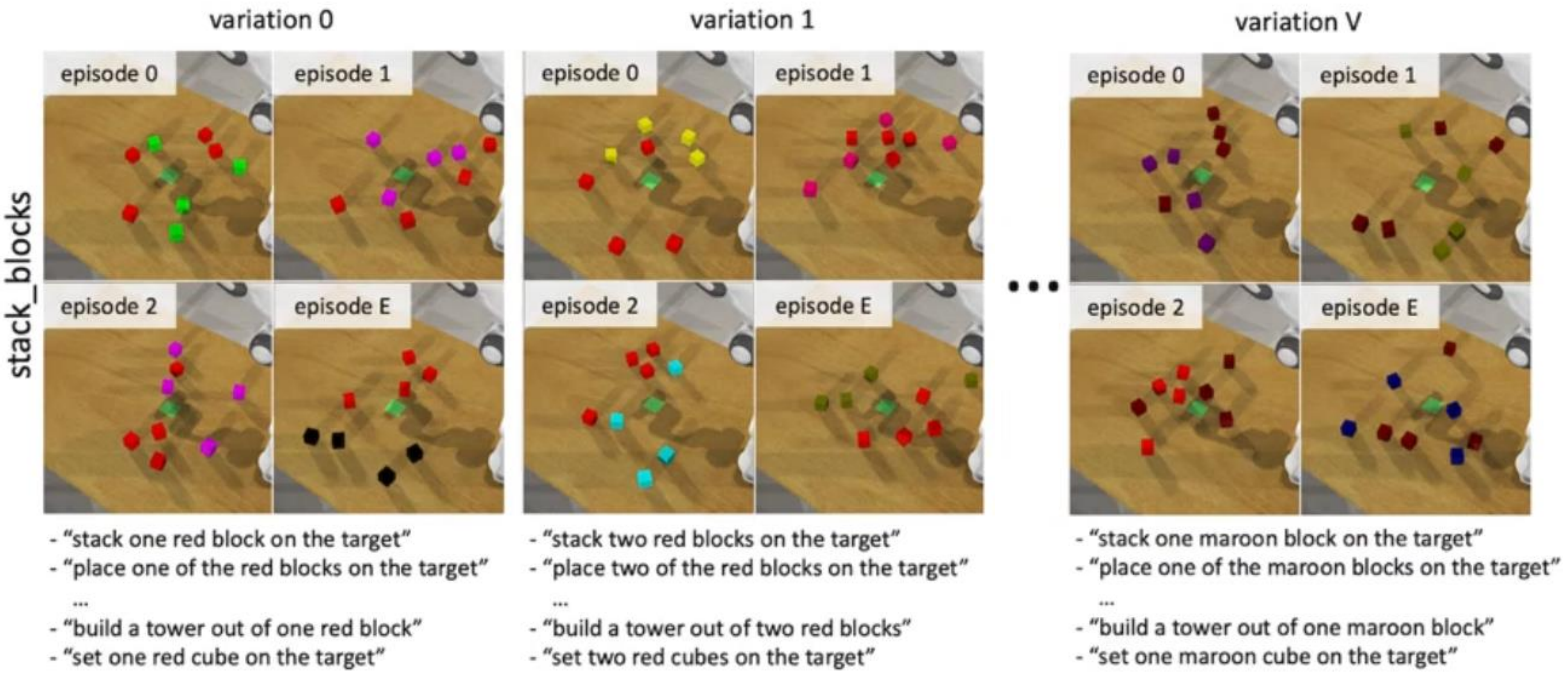
RLBench: Robot Learning Environment James et al., ICRA 2019

Tasks and variations



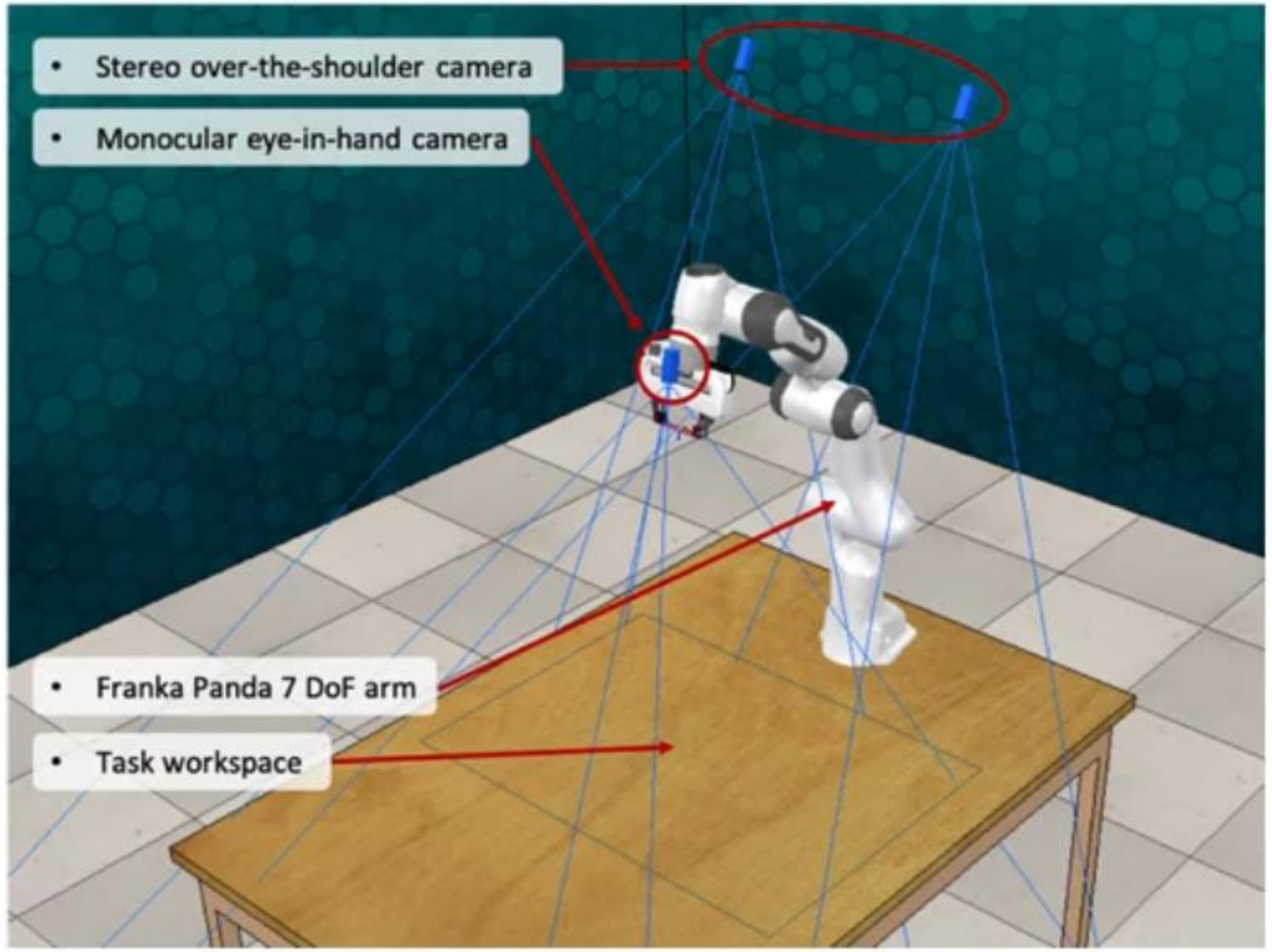
RLBench: Robot Learning Environment James et al., ICRA 2019

Tasks and variations

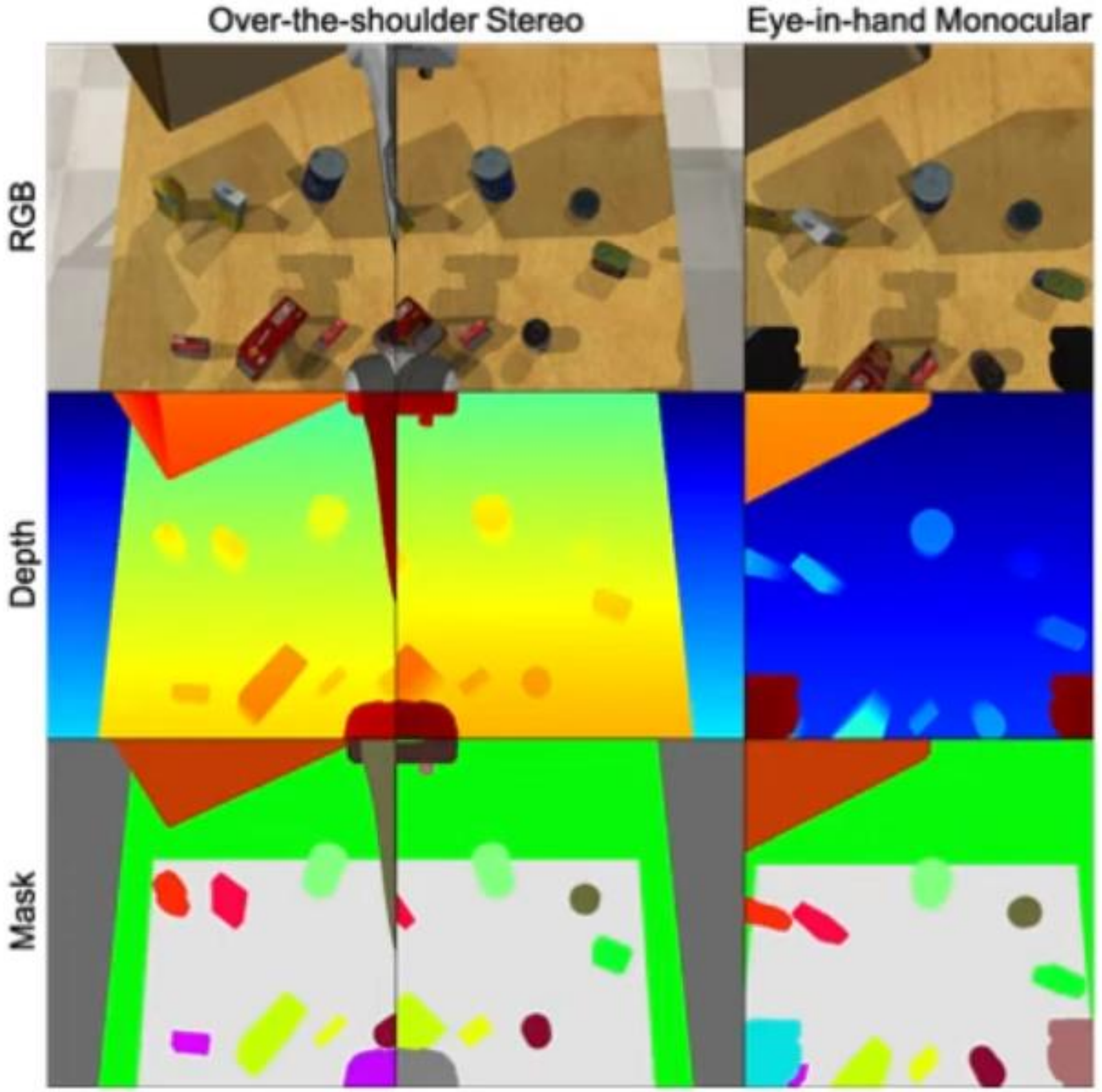


RLBench: Robot Learning Environment James et al., ICRA 2019

Simulation of scenes and observations



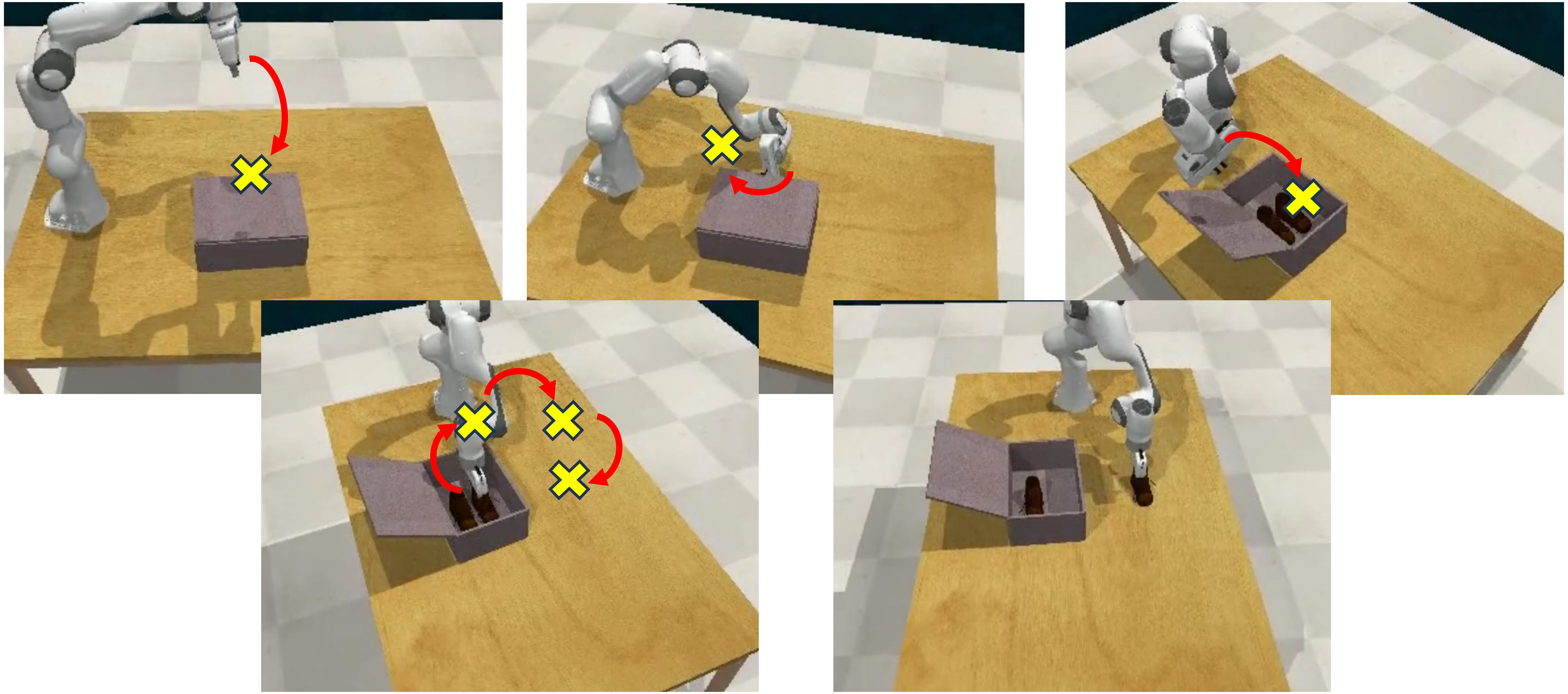
Scene



Observations

RLBench: Robot Learning Environment James et al., ICRA 2019

Demonstrations are defined by **3D waypoints**



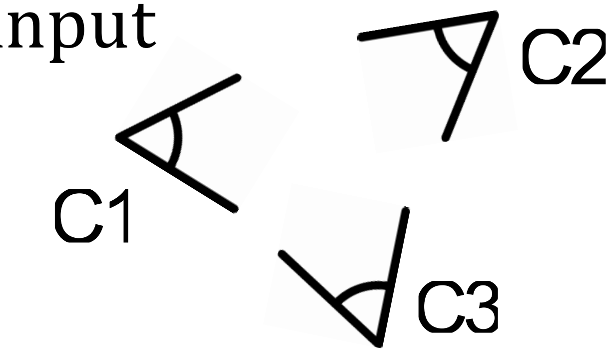
RLBench: Robot Learning Environment James et al., ICRA 2019

Manipulation vs. Navigation

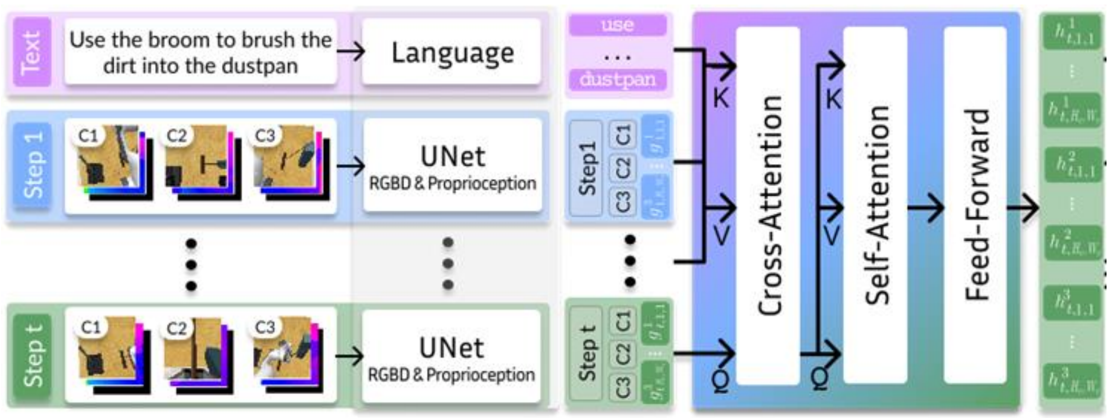
HiveFormer Guhur et al., CoRL 2022

Takes language-defined tasks as input

Uses multi-camera input



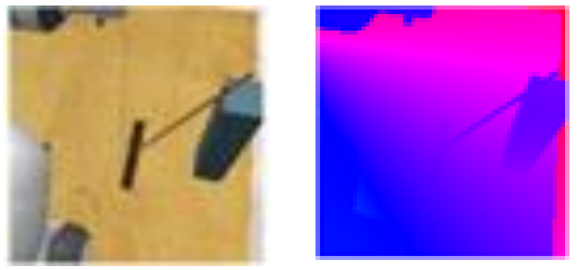
Transformer



Encodes explicit observation history

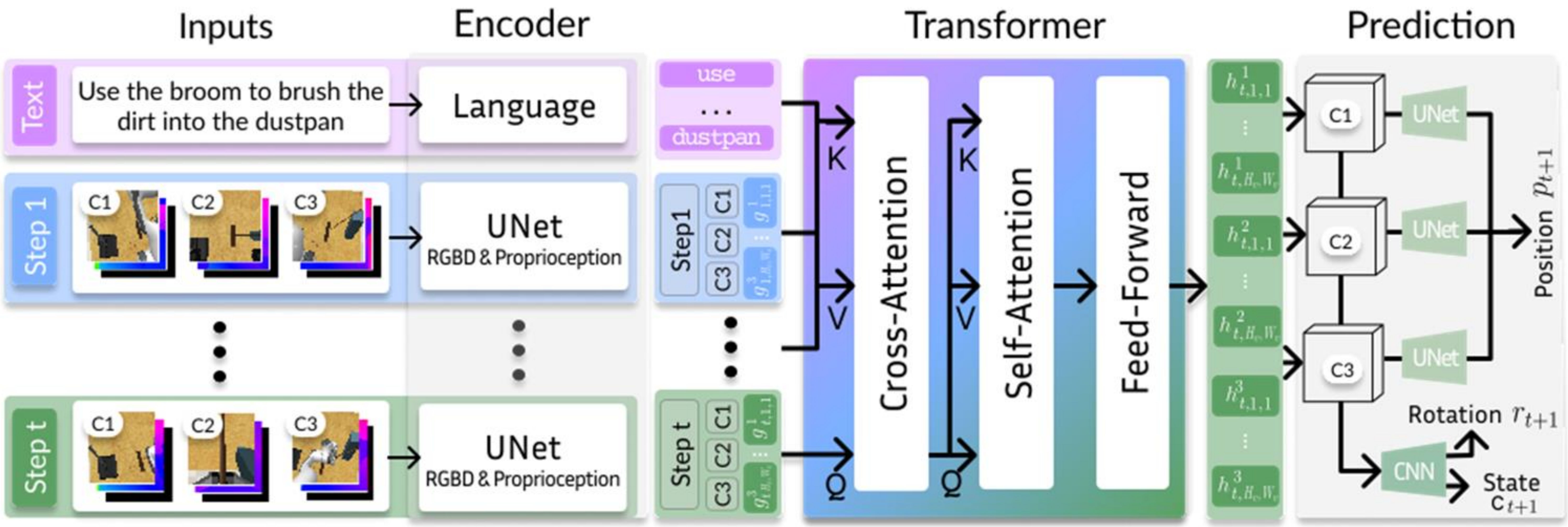
o_1 ... o_t

Uses raw RGB+D for visuomotor policies



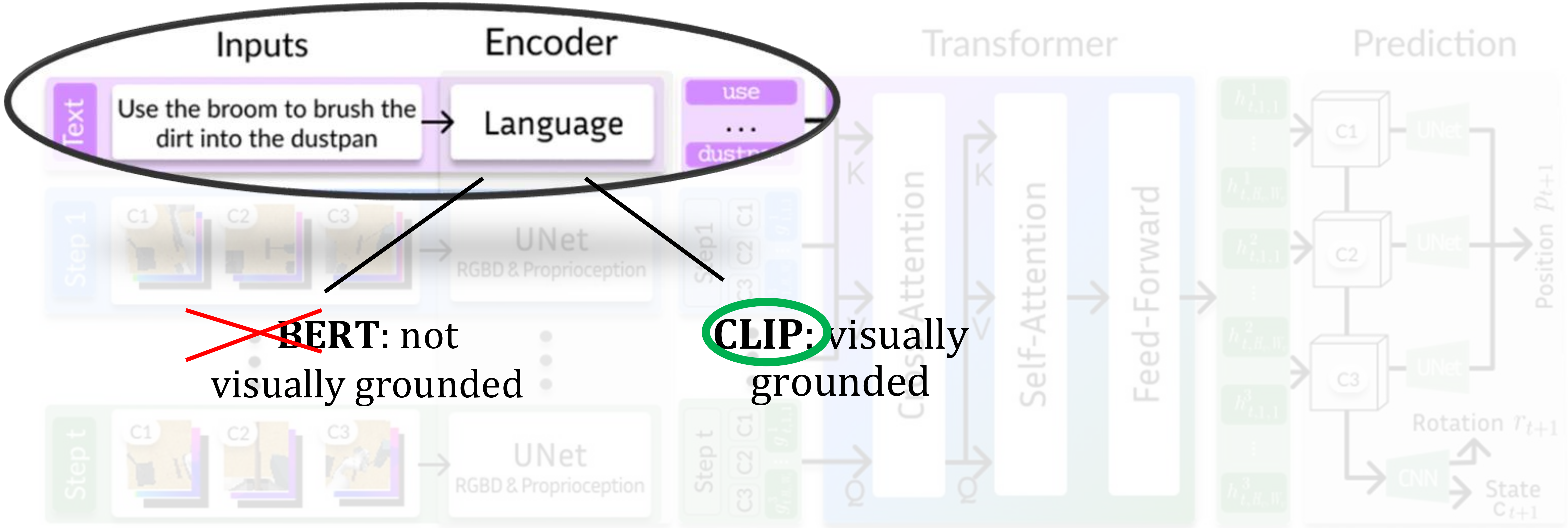
HiveFormer Guhur et al., CoRL 2022

History-aware instruction-conditioned multi-view transformer



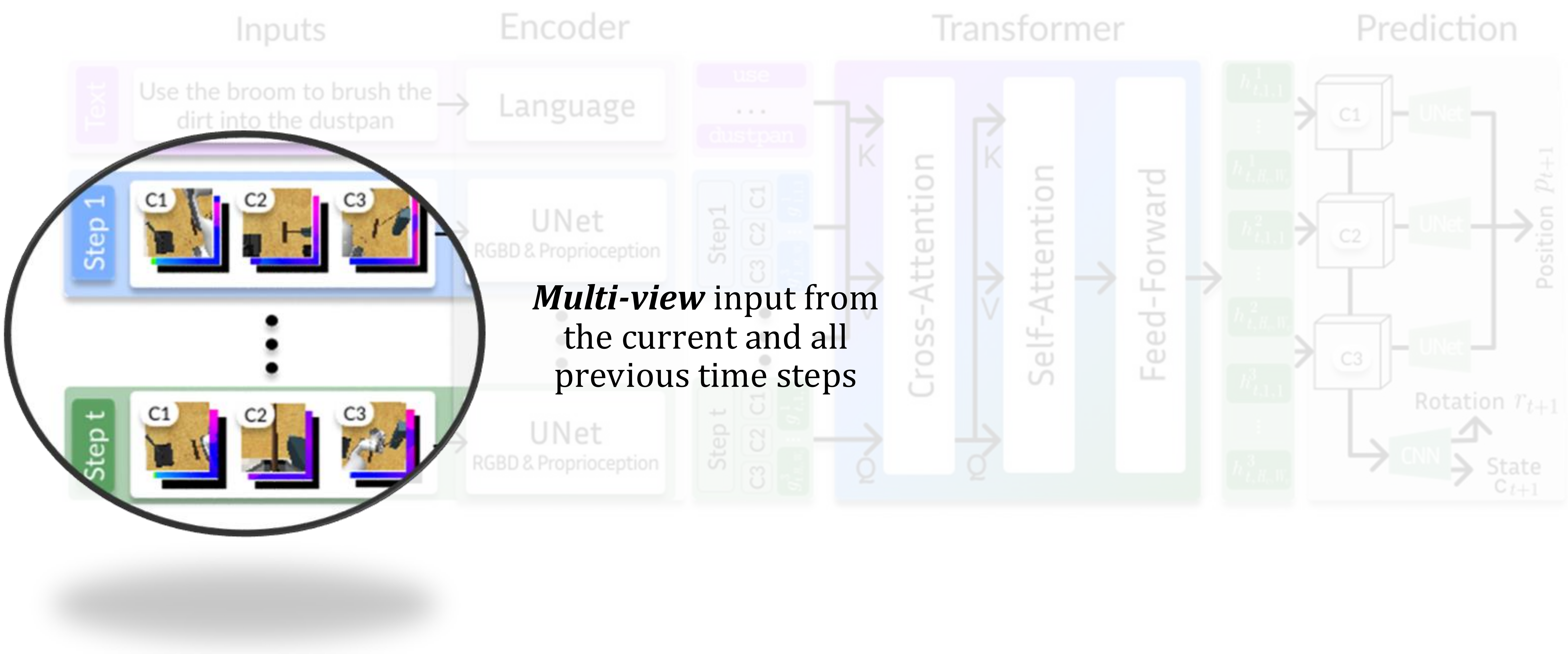
HiveFormer Guhur et al., CoRL 2022

History-aware **i**nstruction-conditioned multi-**vi**ew trans**former**



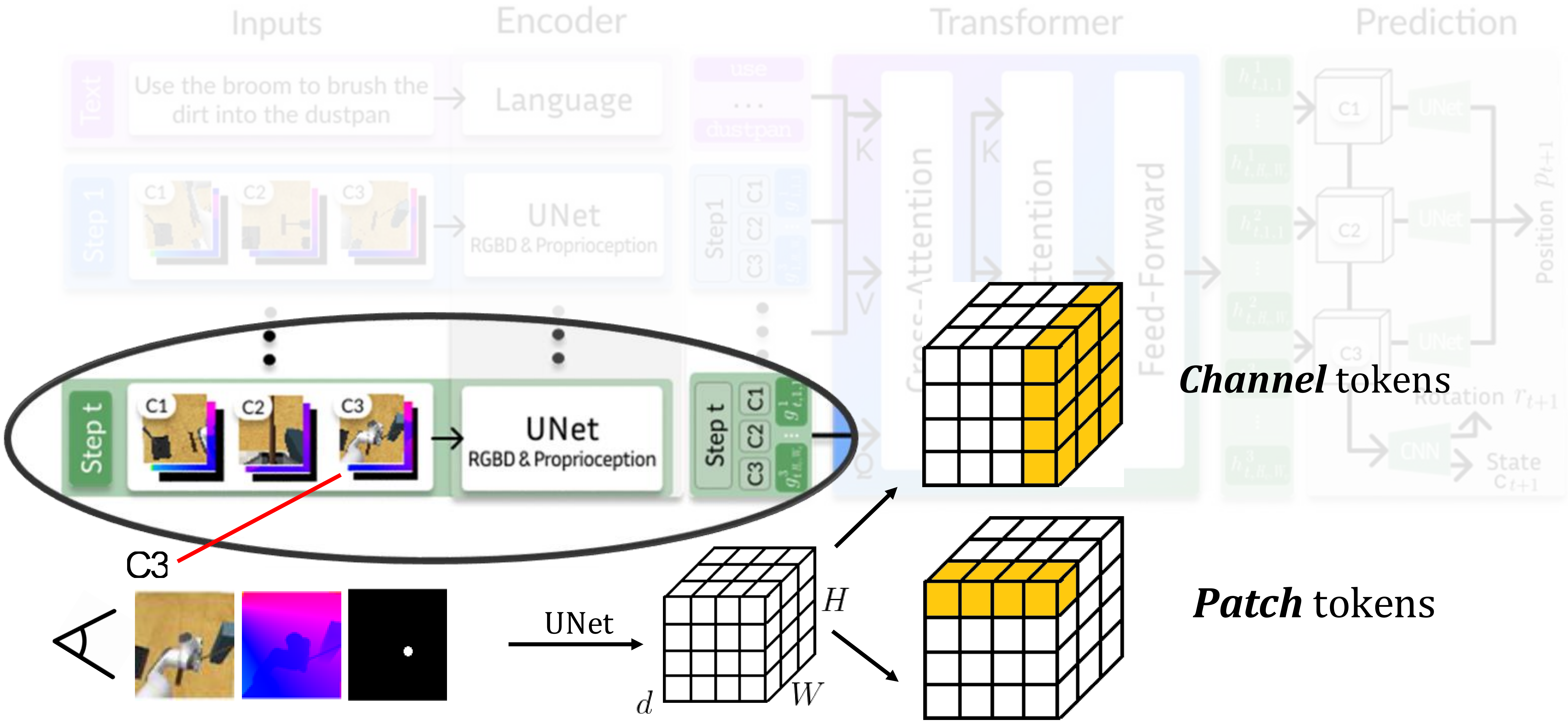
HiveFormer Guhur et al., CoRL 2022

History-aware instruction-conditioned multi-view transformer



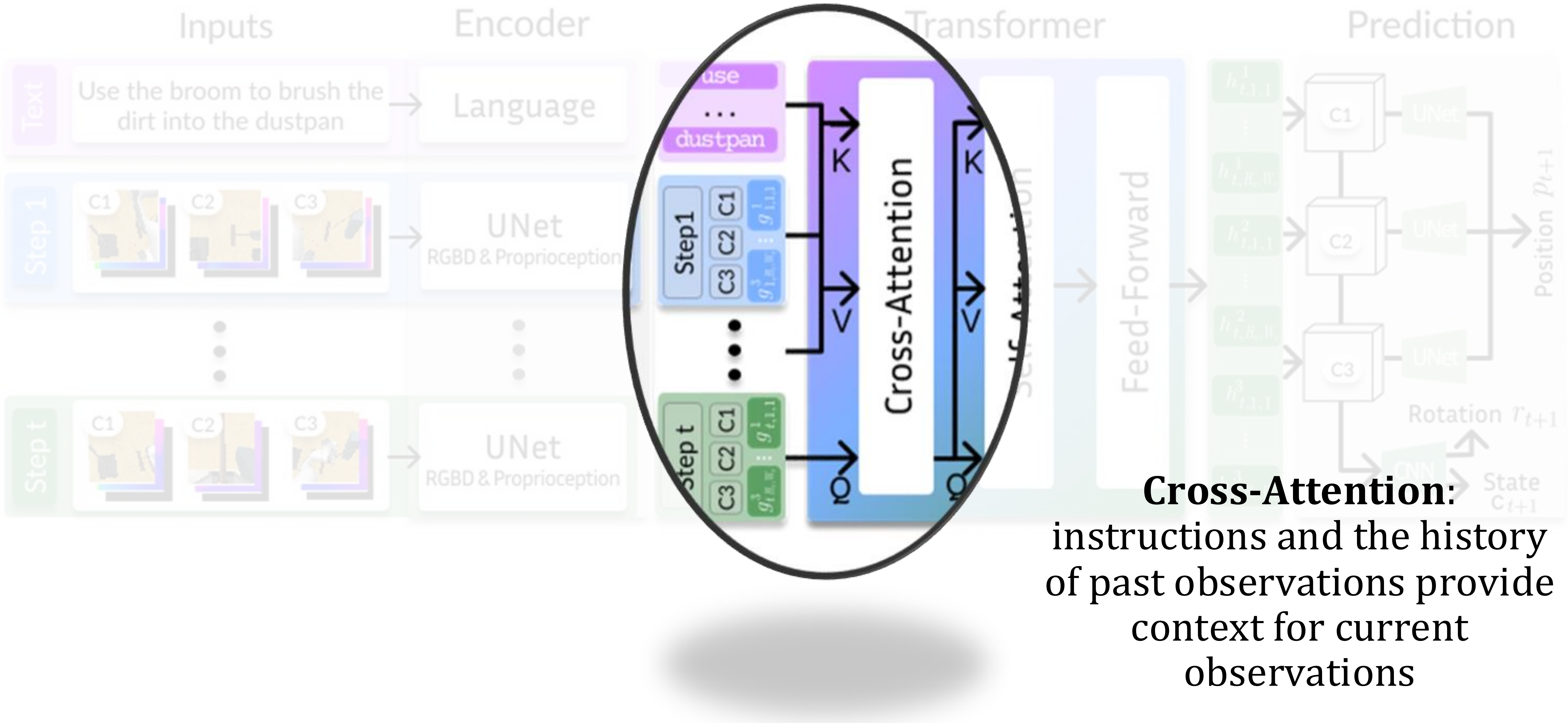
HiveFormer Guhur et al., CoRL 2022

History-aware instruction-conditioned multi-view transformer



HiveFormer Guhur et al., CoRL 2022

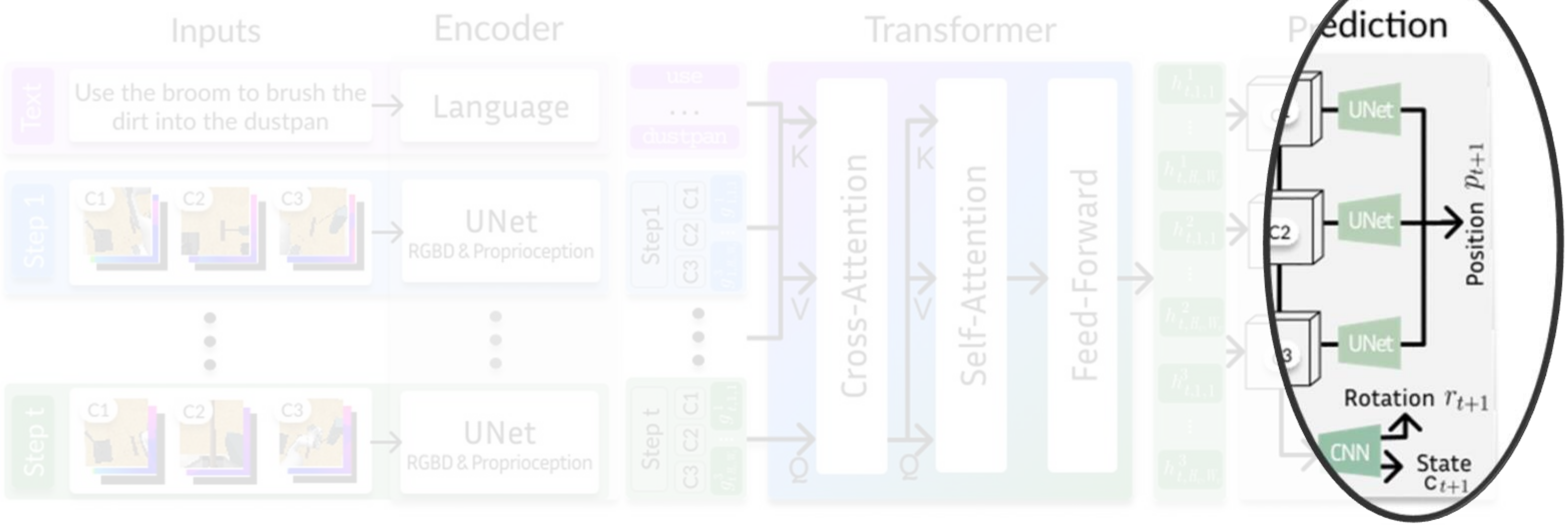
History-aware instruction-conditioned multi-view transformer



Cross-Attention: instructions and the history of past observations provide context for current observations

HiveFormer Guhur et al., CoRL 2022

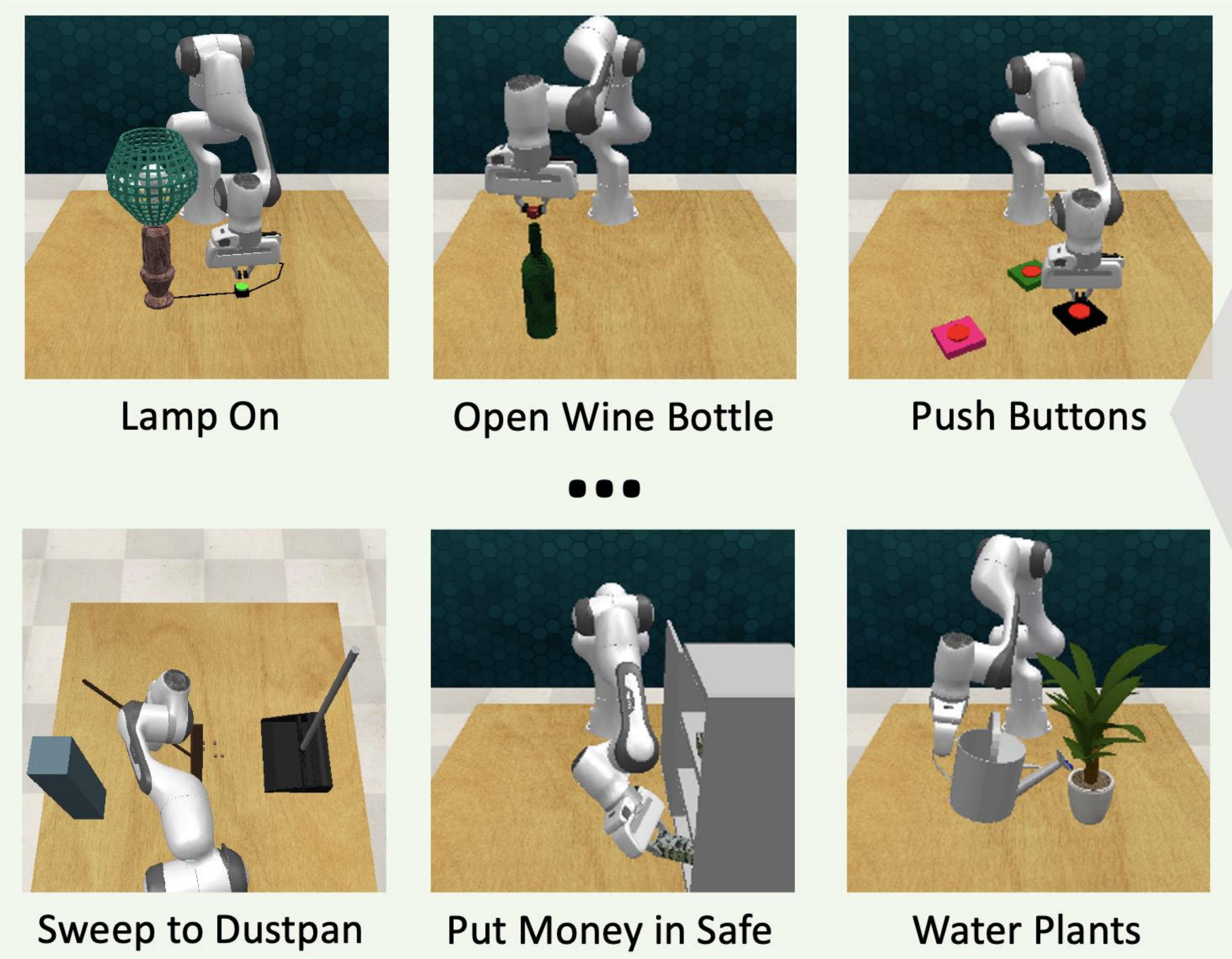
History-aware instruction-conditioned multi-view transformer



Behavior Cloning loss for training; Single and Multi-task training

HiveFormer: Evaluation steup

17 RLBench tasks



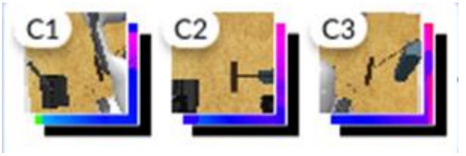
Task text descriptions are not needed

HiveFormer: Results 10 tasks • Single-task setting

	Visual Tokens	Point Clouds	Gripper Position	Multi-View	History	Attn	Mask Obs	SR
R1	×	×	×	×	×	×	×	72.9 ± 4.1
R2	Channel	×	×	✓	×	Self	×	73.1 ± 4.5
R3	Channel	✓	×	✓	×	Self	×	77.1 ± 5.8
R4	Channel	✓	✓	✓	×	Self	×	78.1 ± 5.8
R5	Channel	✓	✓	✓	✓	Self	×	81.8 ± 5.2
R6	Channel	✓	✓	✓	✓	Self	✓	82.3 ± 5.3
R7	Patch	✓	✓	✓	✓	Self	✓	84.4 ± 6.4
R8	Patch	✓	✓	✓	✓	Cross	✓	88.4 ± 4.9

Transformer with multi-view, depth and gripper: +5.2%
w/ vs. w/o history: +3.7%
Patch vs. channel tokens: +2.1%
Cross- vs. Self-Attention: +4%
Overall: +15.5%

HiveFormer: Results 10 tasks • Single-task setting

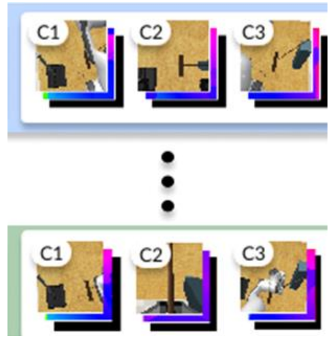


	Visual Tokens	Point Clouds	Gripper Position	Multi-View	History	Attn	Mask Obs	SR
R1	×	×	×	×	×	×	×	72.9 ± 4.1
R2	Channel	×	×	✓	×	Self	×	73.1 ± 4.5
R3	Channel	✓	×	✓	×	Self	×	77.1 ± 5.8
R4	Channel	✓	✓	✓	×	Self	×	78.1 ± 5.8
R5	Channel	✓	✓	✓	✓	Self	×	81.8 ± 5.2
R6	Channel	✓	✓	✓	✓	Self	✓	82.3 ± 5.3
R7	Patch	✓	✓	✓	✓	Self	✓	84.4 ± 6.4
R8	Patch	✓	✓	✓	✓	Cross	✓	88.4 ± 4.9

+5.2
%

Transformer with multi-view, depth and gripper: +5.2%
w/ vs. w/o history: +3.7%
Patch vs. channel tokens: +2.1%
Cross- vs. Self-Attention: +4%
Overall: +15.5%

HiveFormer: Results 10 tasks • Single-task setting

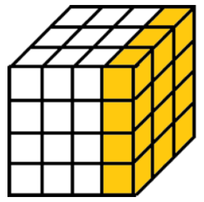


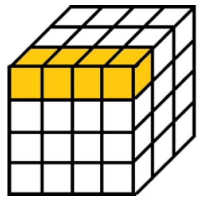
	Visual Tokens	Point Clouds	Gripper Position	Multi-View	History	Attn	Mask Obs	SR
R1	×	×	×	×	×	×	×	72.9 ± 4.1
R2	Channel	×	×	✓	×	Self	×	73.1 ± 4.5
R3	Channel	✓	×	✓	×	Self	×	77.1 ± 5.8
R4	Channel	✓	✓	✓	×	Self	×	78.1 ± 5.8
R5	Channel	✓	✓	✓	✓	Self	×	81.8 ± 5.2
R6	Channel	✓	✓	✓	✓	Self	✓	82.3 ± 5.3
R7	Patch	✓	✓	✓	✓	Self	✓	84.4 ± 6.4
R8	Patch	✓	✓	✓	✓	Cross	✓	88.4 ± 4.9

+3.7%

Transformer with multi-view, depth and gripper: +5.2%
w/ vs. w/o history: +3.7%
Patch vs. channel tokens: +2.1%
Cross- vs. Self-Attention: +4%
Overall: +15.5%

HiveFormer: Results 10 tasks • Single-task setting



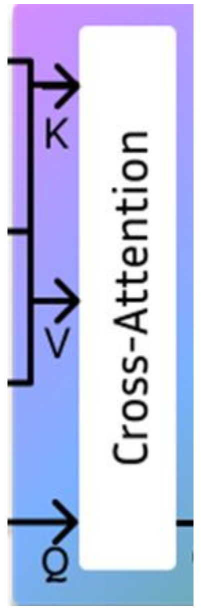


	Visual Tokens	Point Clouds	Gripper Position	Multi-View	History	Attn	Mask Obs	SR
R1	×	×	×	×	×	×	×	72.9 ± 4.1
R2	Channel	×	×	✓	×	Self	×	73.1 ± 4.5
R3	Channel	✓	×	✓	×	Self	×	77.1 ± 5.8
R4	Channel	✓	✓	✓	×	Self	×	78.1 ± 5.8
R5	Channel	✓	✓	✓	✓	Self	×	81.8 ± 5.2
R6	Channel	✓	✓	✓	✓	Self	✓	82.3 ± 5.3
R7	Patch	✓	✓	✓	✓	Self	✓	84.4 ± 6.4
R8	Patch	✓	✓	✓	✓	Cross	✓	88.4 ± 4.9

+2.1%

Transformer with multi-view, depth and gripper: +5.2%
w/ vs. w/o history: +3.7%
Patch vs. channel tokens: +2.1%
Cross- vs. Self-Attention: +4%
Overall: +15.5%

HiveFormer: Results 10 tasks • Single-task setting

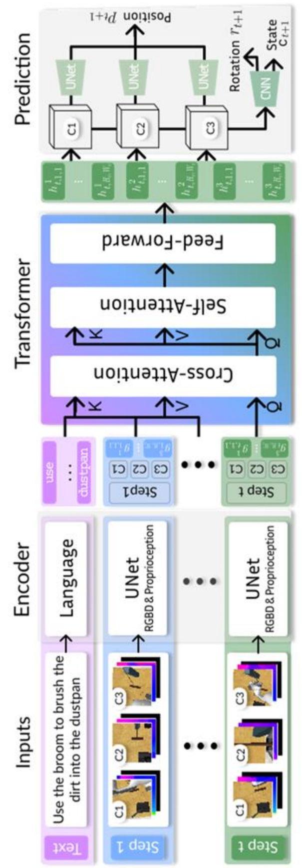


	Visual Tokens	Point Clouds	Gripper Position	Multi-View	History	Attn	Mask Obs	SR
R1	×	×	×	×	×	×	×	72.9 ± 4.1
R2	Channel	×	×	✓	×	Self	×	73.1 ± 4.5
R3	Channel	✓	×	✓	×	Self	×	77.1 ± 5.8
R4	Channel	✓	✓	✓	×	Self	×	78.1 ± 5.8
R5	Channel	✓	✓	✓	✓	Self	×	81.8 ± 5.2
R6	Channel	✓	✓	✓	✓	Self	✓	82.3 ± 5.3
R7	Patch	✓	✓	✓	✓	Self	✓	84.4 ± 6.4
R8	Patch	✓	✓	✓	✓	Cross	✓	88.4 ± 4.9

+4 %

Transformer with multi-view, depth and gripper: +5.2%
w/ vs. w/o history: +3.7%
Patch vs. channel tokens: +2.1%
Cross- vs. Self-Attention: +4%
Overall: +15.5%

HiveFormer: Results 10 tasks • Single-task setting



	Visual Tokens	Point Clouds	Gripper Position	Multi-View	History	Attn	Mask Obs	SR
R1	×	×	×	×	×	×	×	72.9 ± 4.1
R2	Channel	×	×	✓	×	Self	×	73.1 ± 4.5
R3	Channel	✓	×	✓	×	Self	×	77.1 ± 5.8
R4	Channel	✓	✓	✓	×	Self	×	78.1 ± 5.8
R5	Channel	✓	✓	✓	✓	Self	×	81.8 ± 5.2
R6	Channel	✓	✓	✓	✓	Self	✓	82.3 ± 5.3
R7	Patch	✓	✓	✓	✓	Self	✓	84.4 ± 6.4
R8	Patch	✓	✓	✓	✓	Cross	✓	88.4 ± 4.9

+15.5%

Transformer with multi-view, depth and gripper: +5.2%
w/ vs. w/o history: +3.7%
Patch vs. channel tokens: +2.1%
Cross- vs. Self-Attention: +4%
Overall: +15.5%

HiveFormer: Results 74 tasks • Single-task setting

Manually group 74 RLBench tasks into 9 subsets

	Planning	Tools	Long Term	Rot. Invar.	Motion Planning	Screw	Multi Modal	Precision	Visual Occlusion	Avg
Num. of tasks	9	11	4	7	9	4	5	11	14	74
Auto-λ [14]	58.9	20.0	2.3	73.1	66.7	48.2	47.6	34.6	40.6	44.0
Ours (w/o hist)	78.9	46.7	10.0	84.6	73.3	72.6	60.0	63.8	57.9	60.9
Ours (one view)	57.7	23.2	12.3	57.8	63.2	35.6	40.7	33.7	37.1	40.1
Ours	81.6	53.0	16.9	84.2	72.7	80.9	67.1	64.7	60.2	65.4

+21.4 %

HiveFormer generalizes well to many tasks: +21.4% over [14]
History matters especially **Planning**, **Tools** and **Long-Terms** tasks
Multi-view matters especially for **Screw**, **Precision** and **Visual Occlusion** tasks

HiveFormer: Results 74 tasks • Single-task setting

Manually group 74 RLBench tasks into 9 subsets

	Planning	Tools	Long Term	Rot. Invar.	Motion Planning	Screw	Multi Modal	Precision	Visual Occlusion	Avg
Num. of tasks	9	11	4	7	9	4	5	11	14	74
Auto-λ [14]	58.9	20.0	2.3	73.1	66.7	48.2	47.6	34.6	40.6	44.0
Ours (w/o hist)	78.9	46.7	10.0	84.6	73.3	72.6	60.0	63.8	57.9	60.9
Ours (one view)	57.7	23.2	12.3	57.8	63.2	35.6	40.7	33.7	37.1	40.1
Ours	81.6	53.0	16.9	84.2	72.7	80.9	67.1	64.7	60.2	65.4

HiveFormer generalizes well to many tasks: +21.4% over [14]
History matters especially **Planning**, **Tools** and **Long-Terms** tasks
Multi-view matters especially for **Screw**, **Precision** and **Visual Occlusion** tasks

HiveFormer: Results 74 tasks • Single-task setting

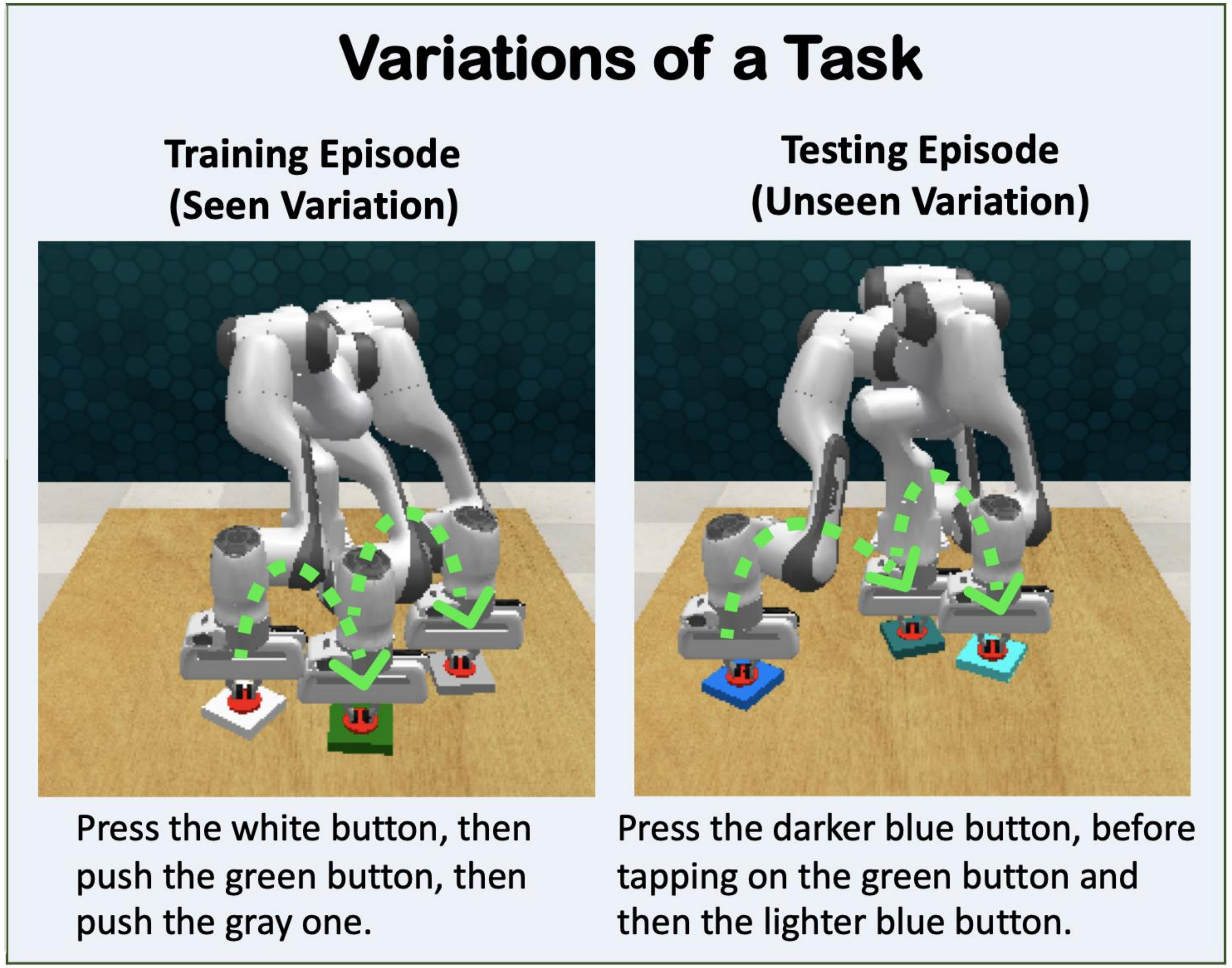
Manually group 74 RLBench tasks into 9 subsets

	Planning	Tools	Long Term	Rot. Invar.	Motion Planning	Screw	Multi Modal	Precision	Visual Occlusion	Avg
Num. of tasks	9	11	4	7	9	4	5	11	14	74
Auto-λ [14]	58.9	20.0	2.3	73.1	66.7	48.2	47.6	34.6	40.6	44.0
Ours (w/o hist)	78.9	46.7	10.0	84.6	73.3	72.6	60.0	63.8	57.9	60.9
Ours (one view)	57.7	23.2	12.3	57.8	63.2	35.6	40.7	33.7	37.1	40.1
Ours	81.6	53.0	16.9	84.2	72.7	80.9	67.1	64.7	60.2	65.4

HiveFormer generalizes well to many tasks: +21.4% over [14]
History matters especially **Planning**, **Tools** and **Long-Terms** tasks
Multi-view matters especially for **Screw**, **Precision** and **Visual Occlusion** tasks

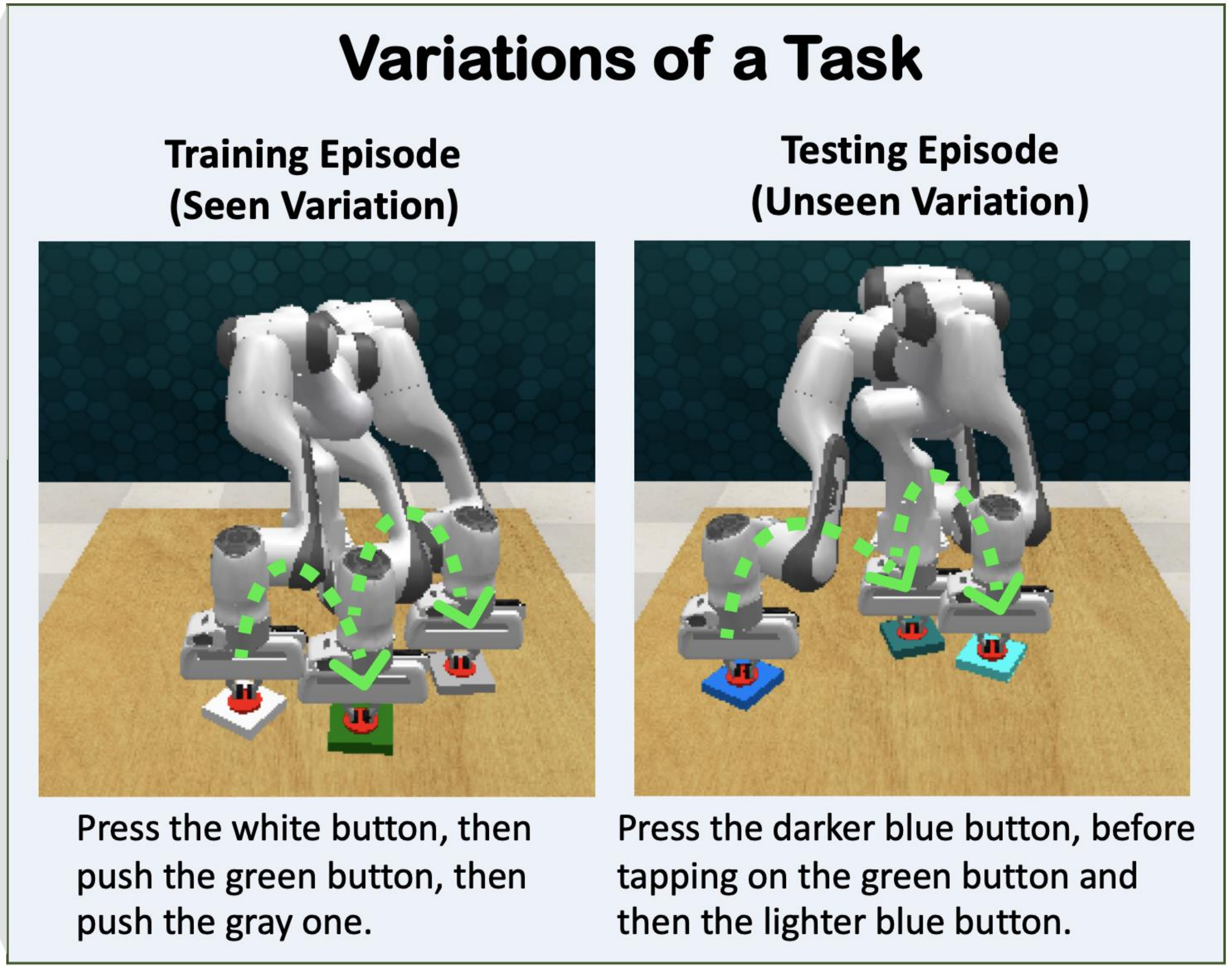
HiveFormer: Evaluation setup

Task variations



Evaluate on unseen task variations
Task text descriptions become crucial

HiveFormer: Results Task variations

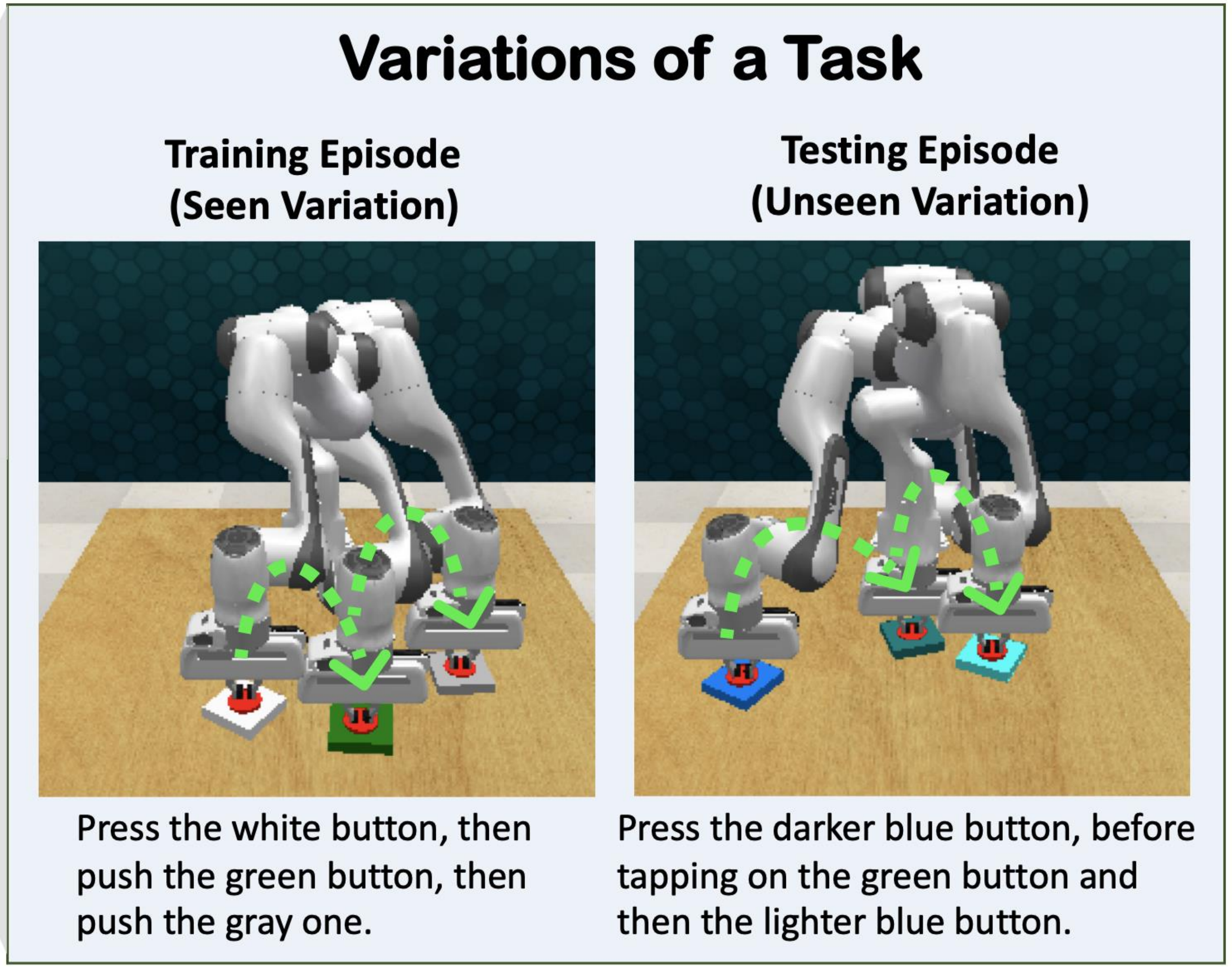


# Demos Per Variation	Instr.	Push Buttons Seen Synt.	Push Buttons Unseen Synt.	Push Buttons Unseen Real	Tower Seen Synt.	Tower Unseen Synt.	Tower Unseen Real
10	Seq.	96.4	71.1	65.7	71.6	49.8	19.4
50	Seq.	99.4	83.1	70.9	74.3	52.1	20.6
100	Seq.	100	86.3	74.2	77.4	56.2	24.1

Generalization to unseen variations

Generalization to natural language extractions

HiveFormer: Results Task variations



# Demos Per Variation	Instr.	Push Buttons Seen Synt.	Push Buttons Unseen Synt.	Push Buttons Unseen Real	Tower Seen Synt.	Tower Unseen Synt.	Tower Unseen Real
10	Seq.	96.4	71.1	65.7	71.6	49.8	19.4
50	Seq.	99.4	83.1	70.9	74.3	52.1	20.6
100	Seq.	100	86.3	74.2	77.4	56.2	24.1

Generalization to unseen variations

Generalization to natural language expressions

Robust visual sim-to-real transfer for robotic manipulation

Ricardo Garcia Robin Strudel Shizhe Chen Etienne Arlaud

Ivan Laptev Cordelia Schmid

OCTOBER 1 - 5, 2023
IEEE/RSJ International Conference on Intelligent Robots and Systems

Inria ENS | PSL PR[AI]RIE

PaRis Artificial Intelligence Research InstitutE

Robust visual sim-to-real transfer for robotic manipulation

Ricardo Garcia Robin Strudel Shizhe Chen Etienne Arlaud

Ivan Laptev Cordelia Schmid

OCTOBER 1 - 5, 2023

IEEE/RSJ International Conference on Intelligent Robots and Systems

Inria

ENS | PSL

PR[AI]RIE

PaRis Artificial Intelligence Research InstitutE

Robust visual sim-to-real transfer for robotic manipulation

Ricardo Garcia Robin Strudel Shizhe Chen Etienne Arlaud
Ivan Laptev Cordelia Schmid

OCTOBER 1 - 5, 2023
IEEE/RSJ International Conference on Intelligent Robots and Systems

ENS | PSL

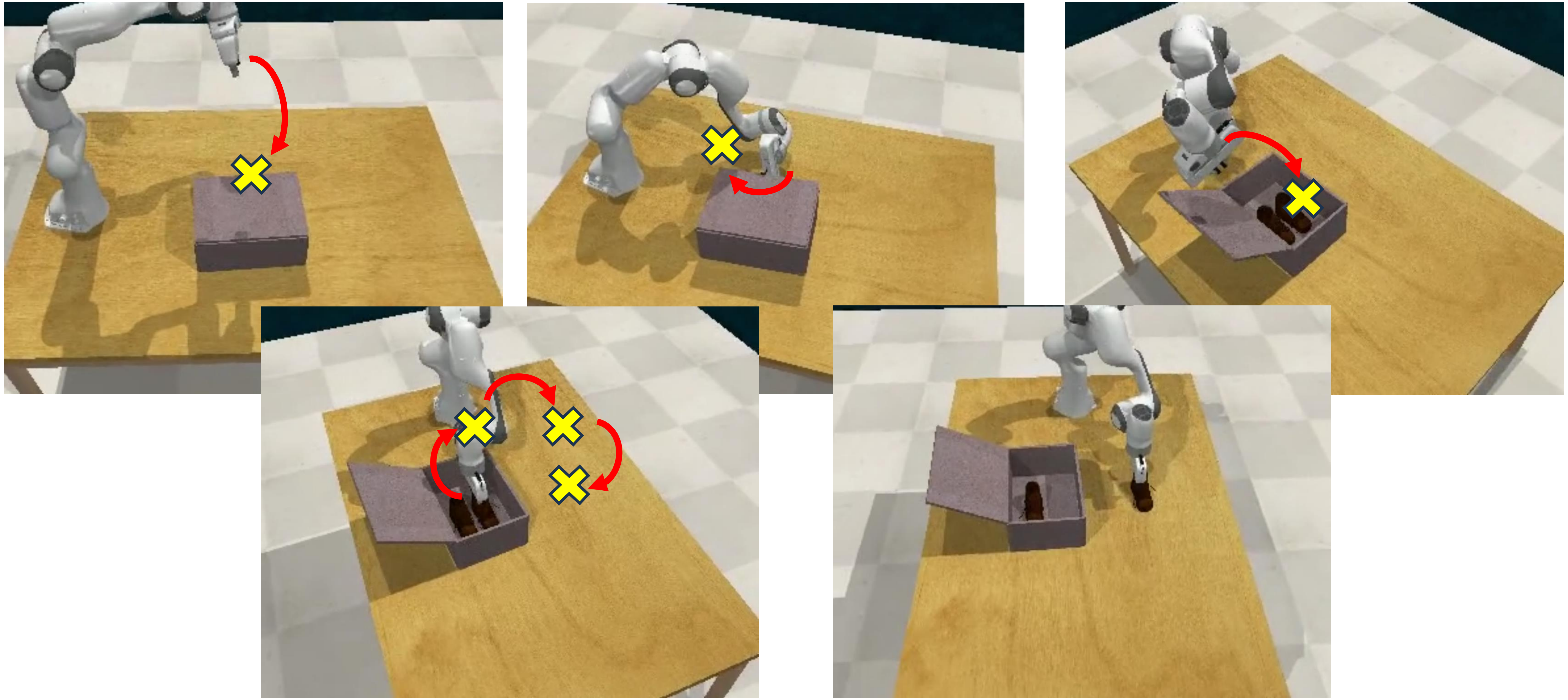
PR[AI]RIE
PaRis Artificial Intelligence Research InstitutE

PolarNet: 3D Point Clouds for Language-Guided Robotic Manipulation

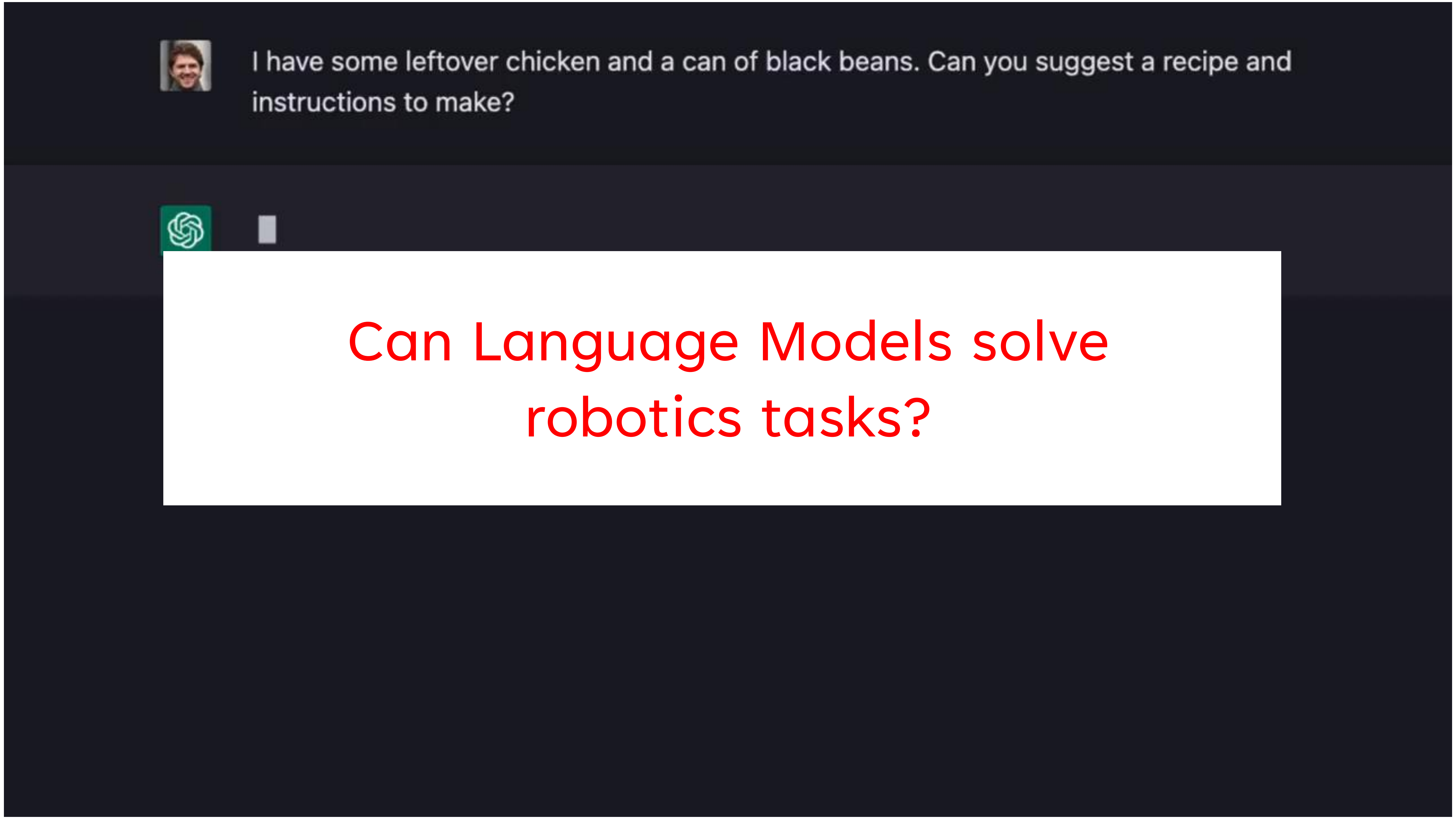
CoRL Submission #247

Supervision

Can we plan tasks without waypoint supervision?



Language Models



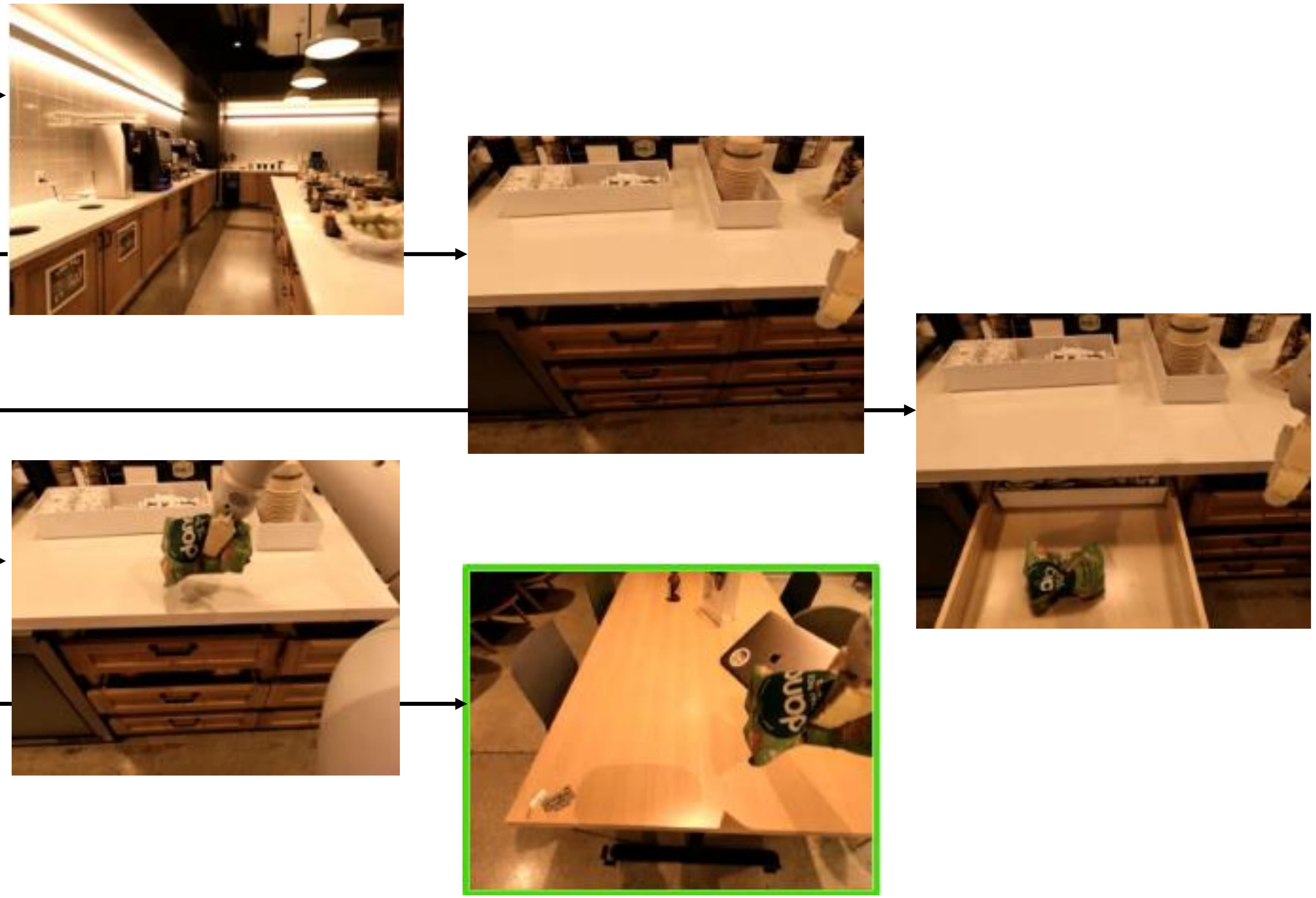
Can Language Models solve robotics tasks?

Language Models for task planning

> Bring me the rice chips from the drawer

LLM:

- Go to the drawers
- Open the top drawer
- Take the rice chips out of the drawer
- Bring it to the user
- Put it down



LLMs for planning: SayCan Ahn et al., CoRL 2022

MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

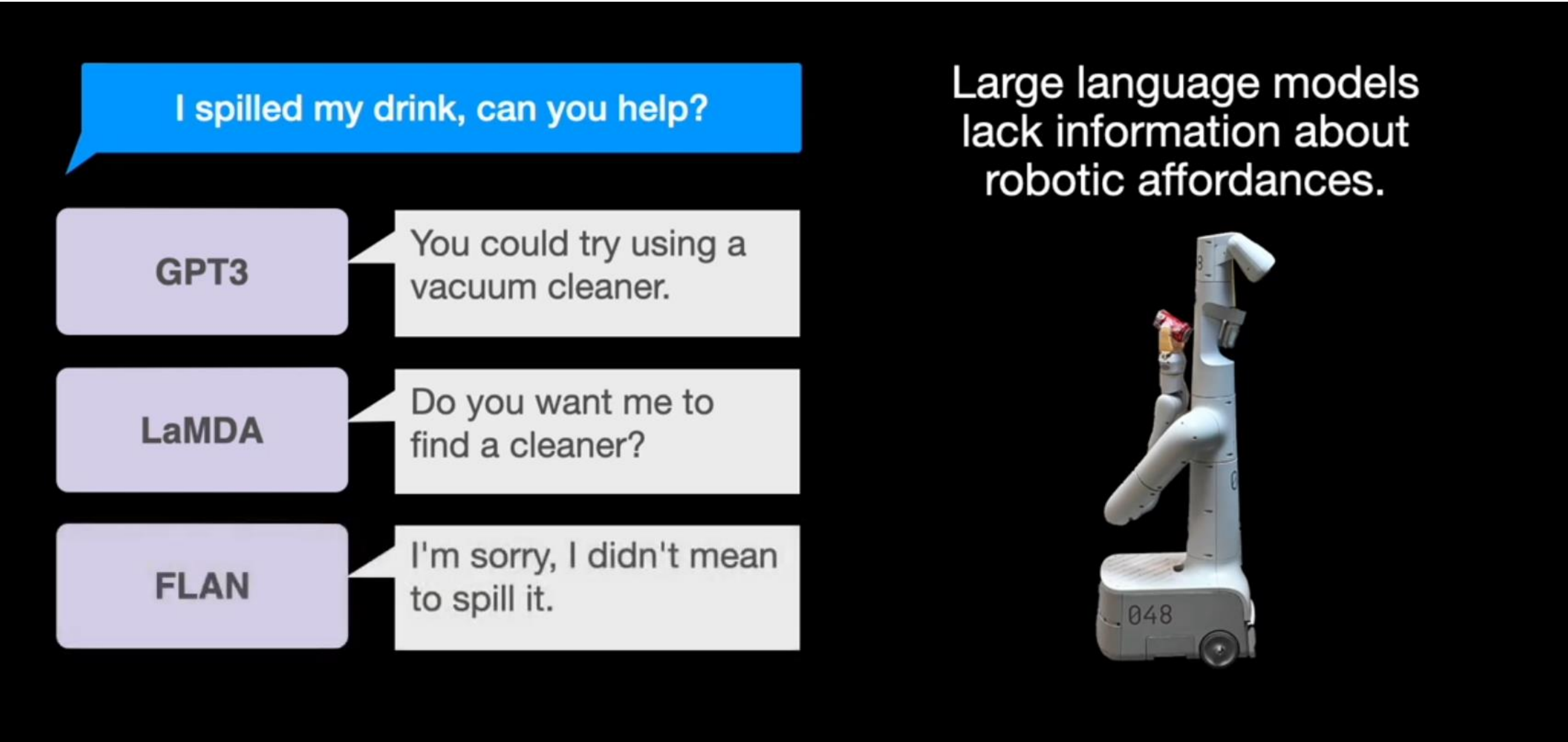
Solve long-horizon tasks from natural language instructions by grounding large language models in the real world.

"I spilled my drink, can you help with that?"

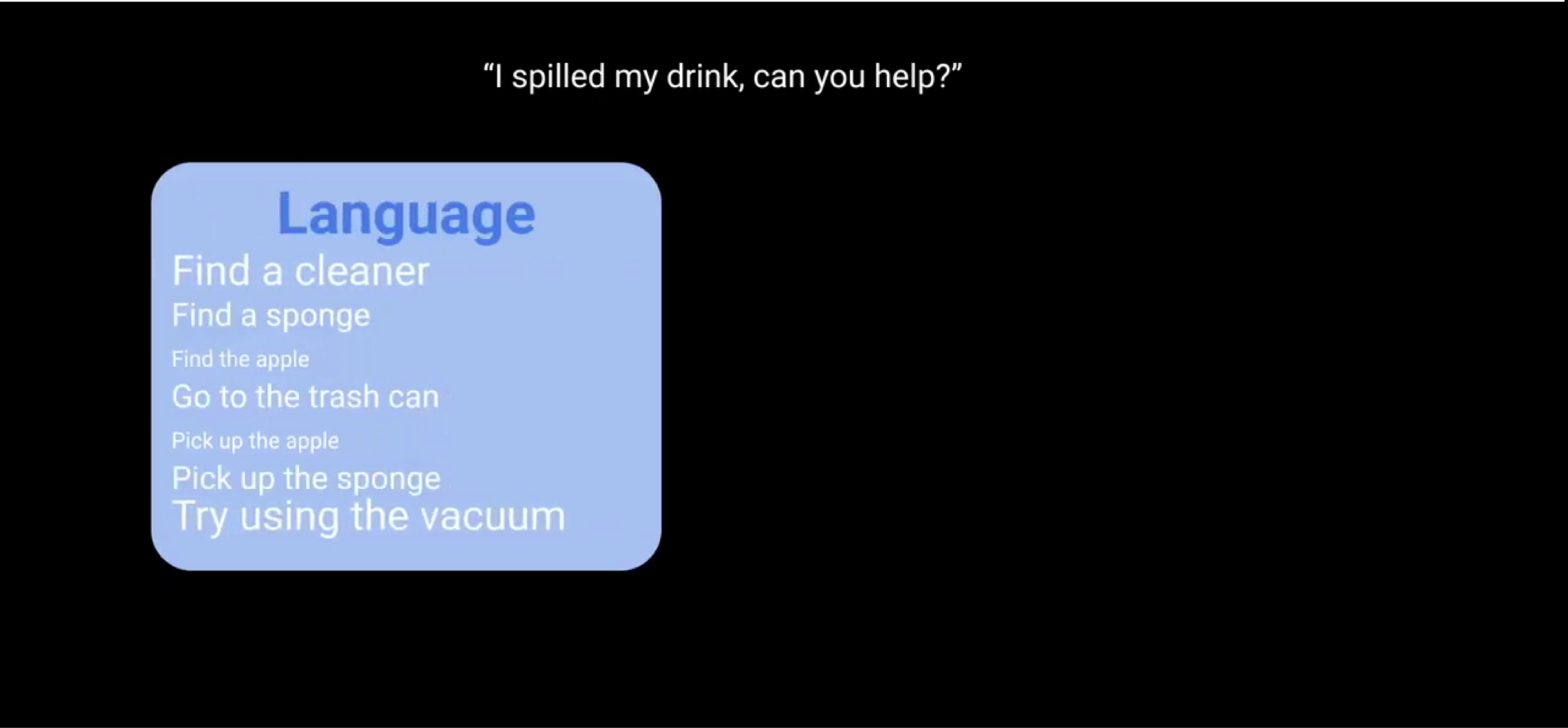
"I just worked out, can you bring me a snack and a drink to recover?"

"I finished a can of coke, can you throw away the can for me?"

LLMs for planning: SayCan Ahn et al., CoRL 2022

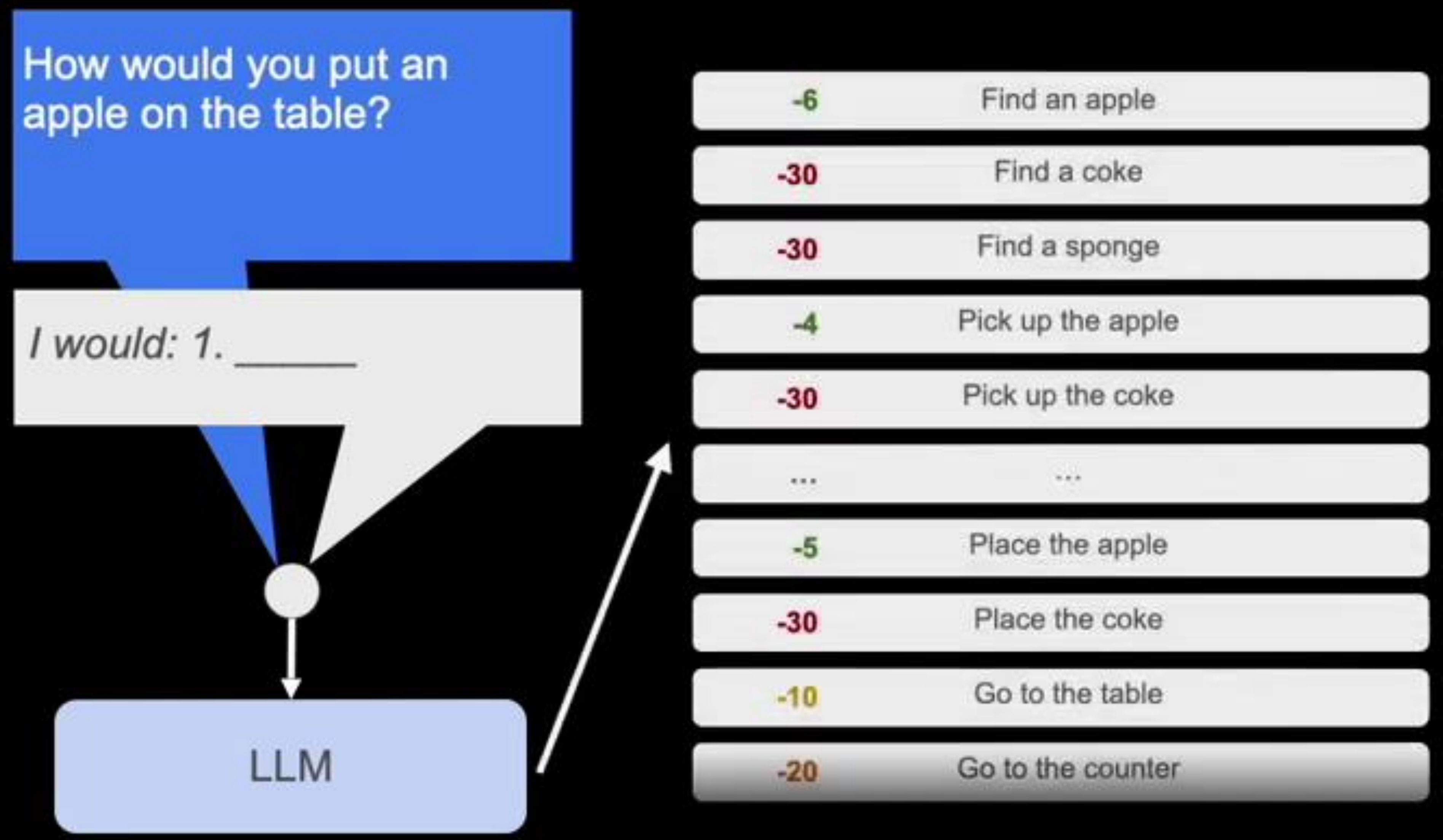


LLMs for planning: SayCan Ahn et al., CoRL 2022



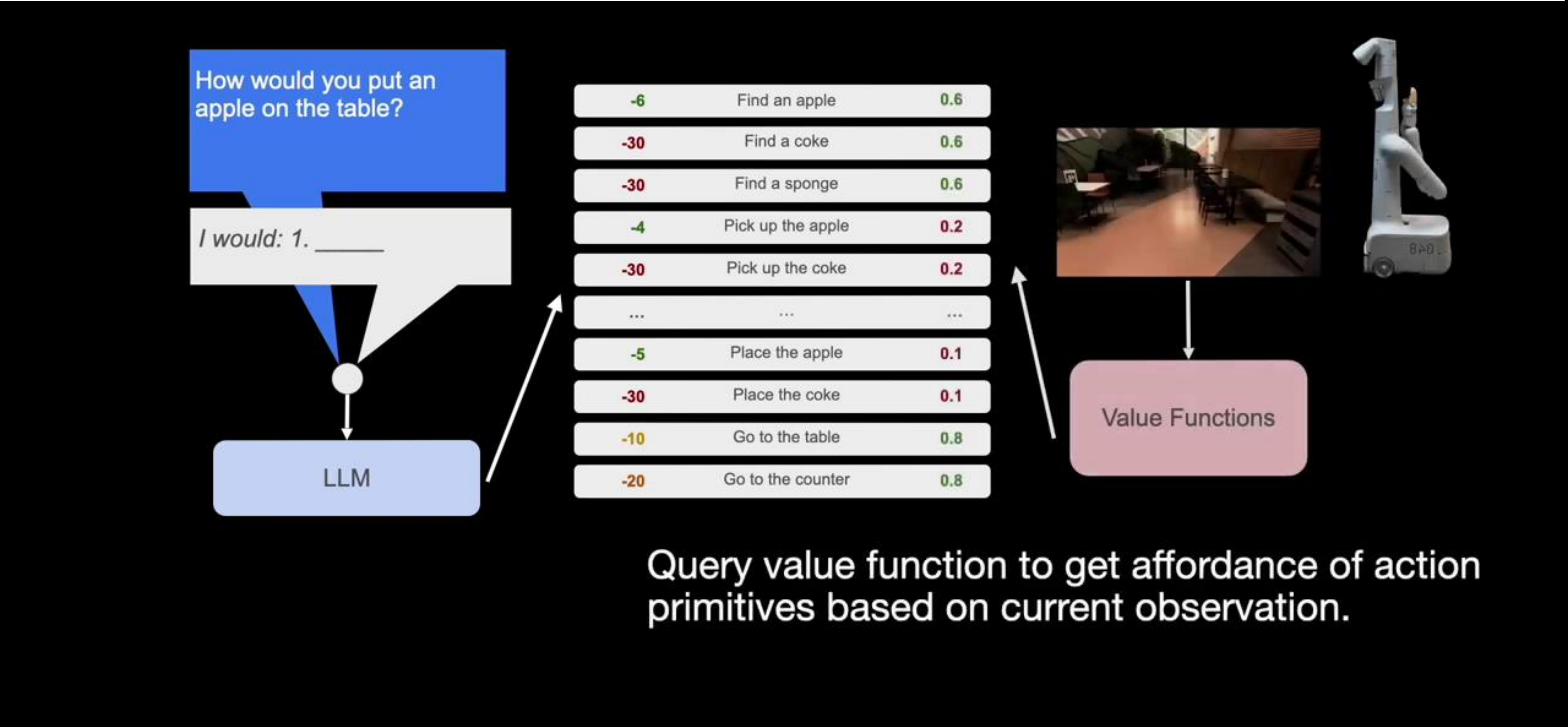
LLMs for planning: SayCan Ahn et al., CoRL 2022

LLMs for planning: SayCan Ahn et al., CoRL 2022

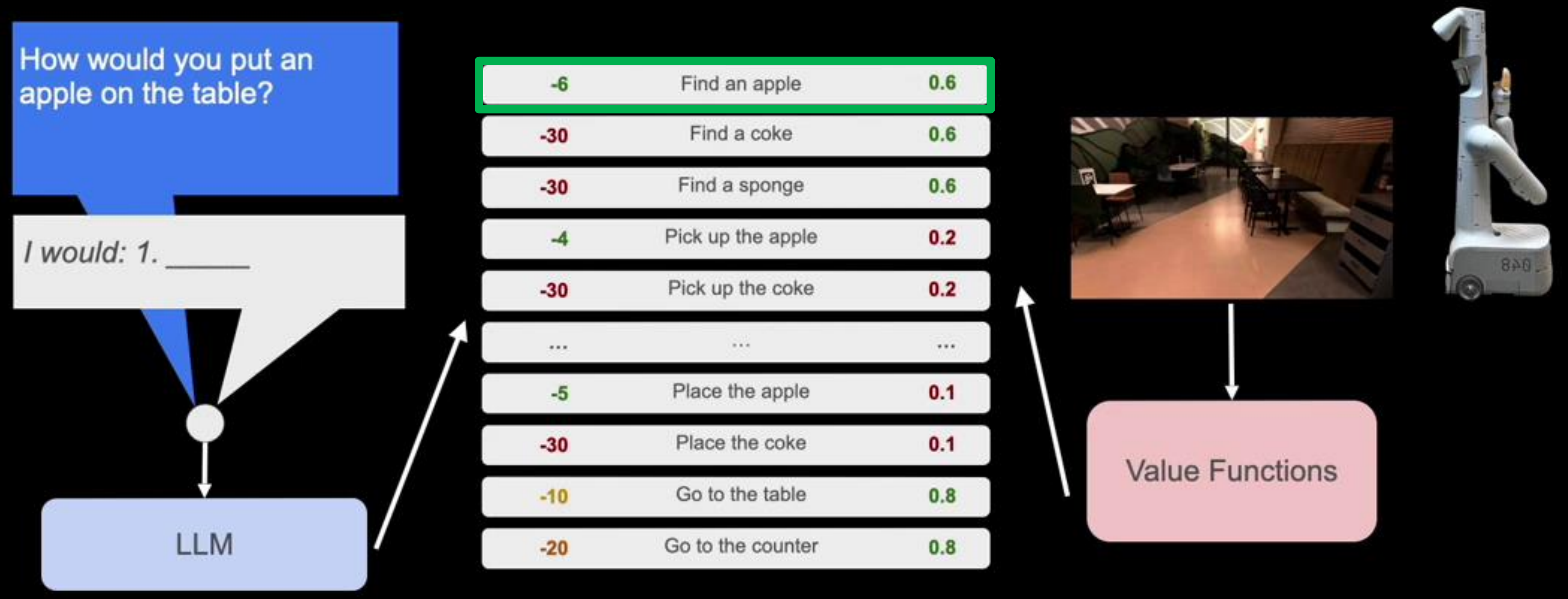


Query language model to rank action primitives based on the instruction.

LLMs for planning: SayCan Ahn et al., CoRL 2022

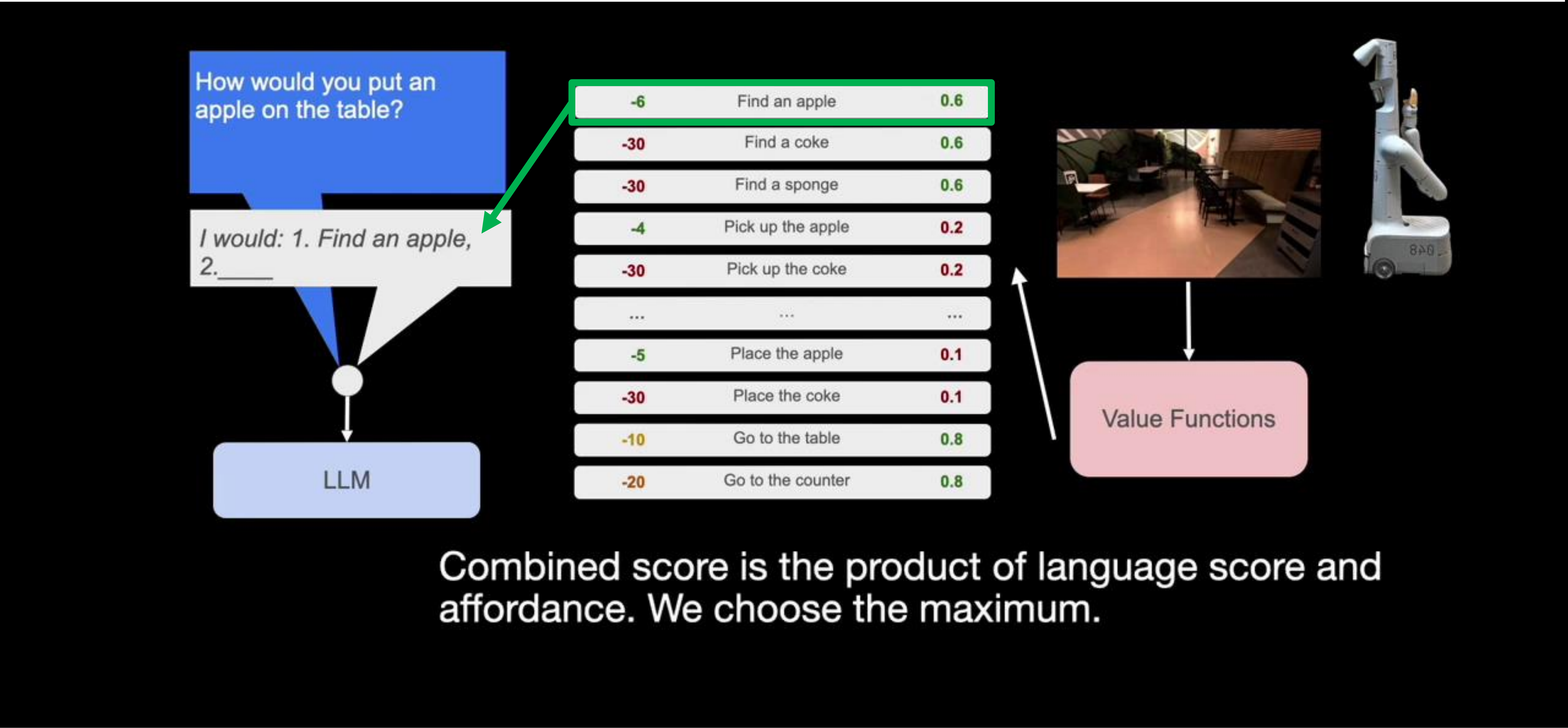


LLMs for planning: SayCan Ahn et al., CoRL 2022



Combined score is the product of language score and affordance. We choose the maximum.

LLMs for planning: SayCan Ahn et al., CoRL 2022



LLMs for planning: SayCan Ahn et al., CoRL 2022

LLMs for planning: SayCan Ahn et al., CoRL 2022

The process is repeated until the task is finished.

Supplementary Video for
"Do As I Can, Not As I Say:
Grounding Language in Robotics Affordances"

Robotics at Google and Everyday Robots

MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

LAIKA: Robot-Dog Explorer Demo

LAIKA: Functionality

- Understands human instructions, e.g. *“go to bicycle and check if it is broken”*
- Finds and navigates to desired objects
- Reports on the state of found objects

LAIKA: Hardware setup

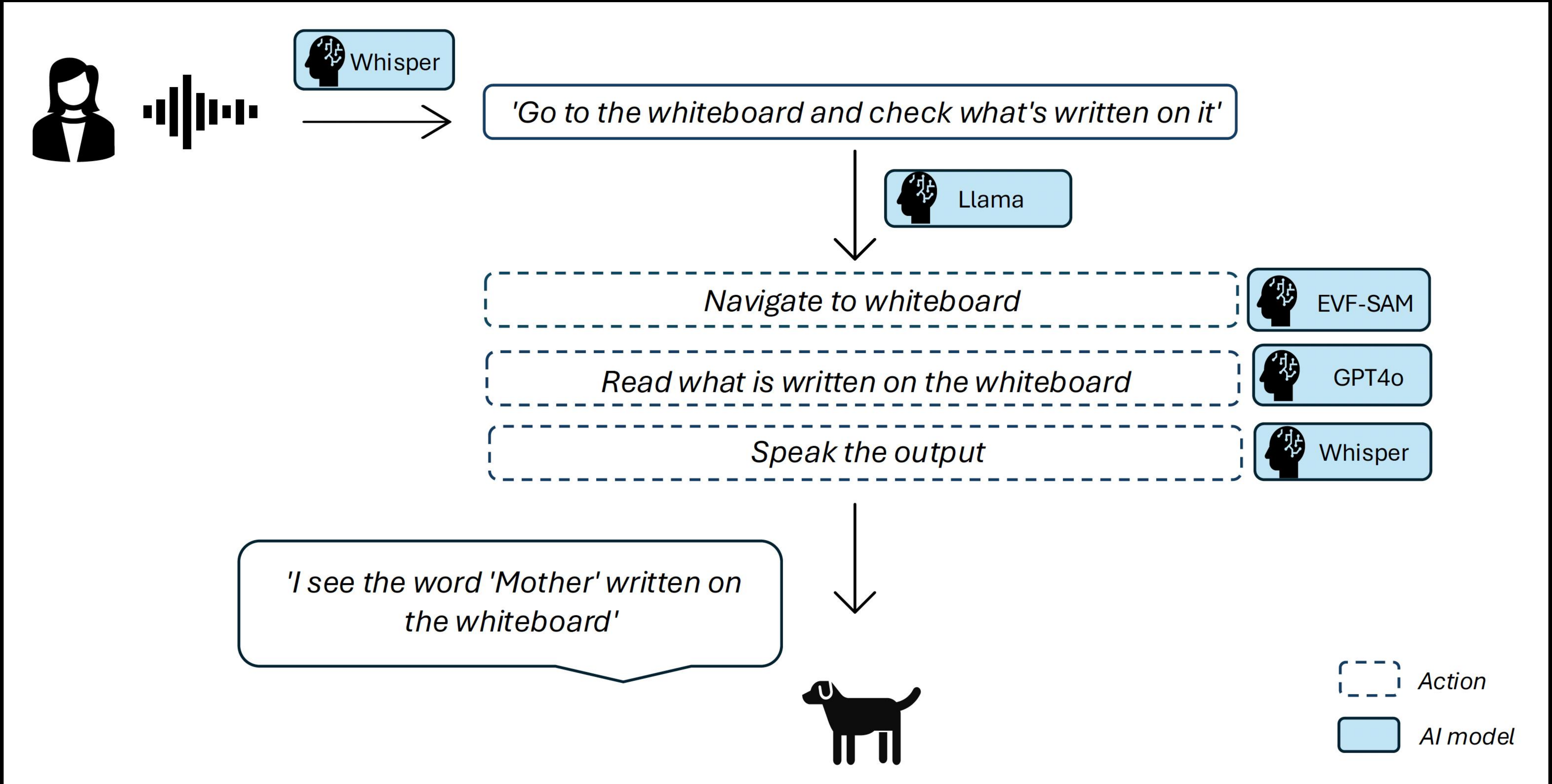
Robot Dog
Unitree Go2

360 camera
Ricoh Theta Z1

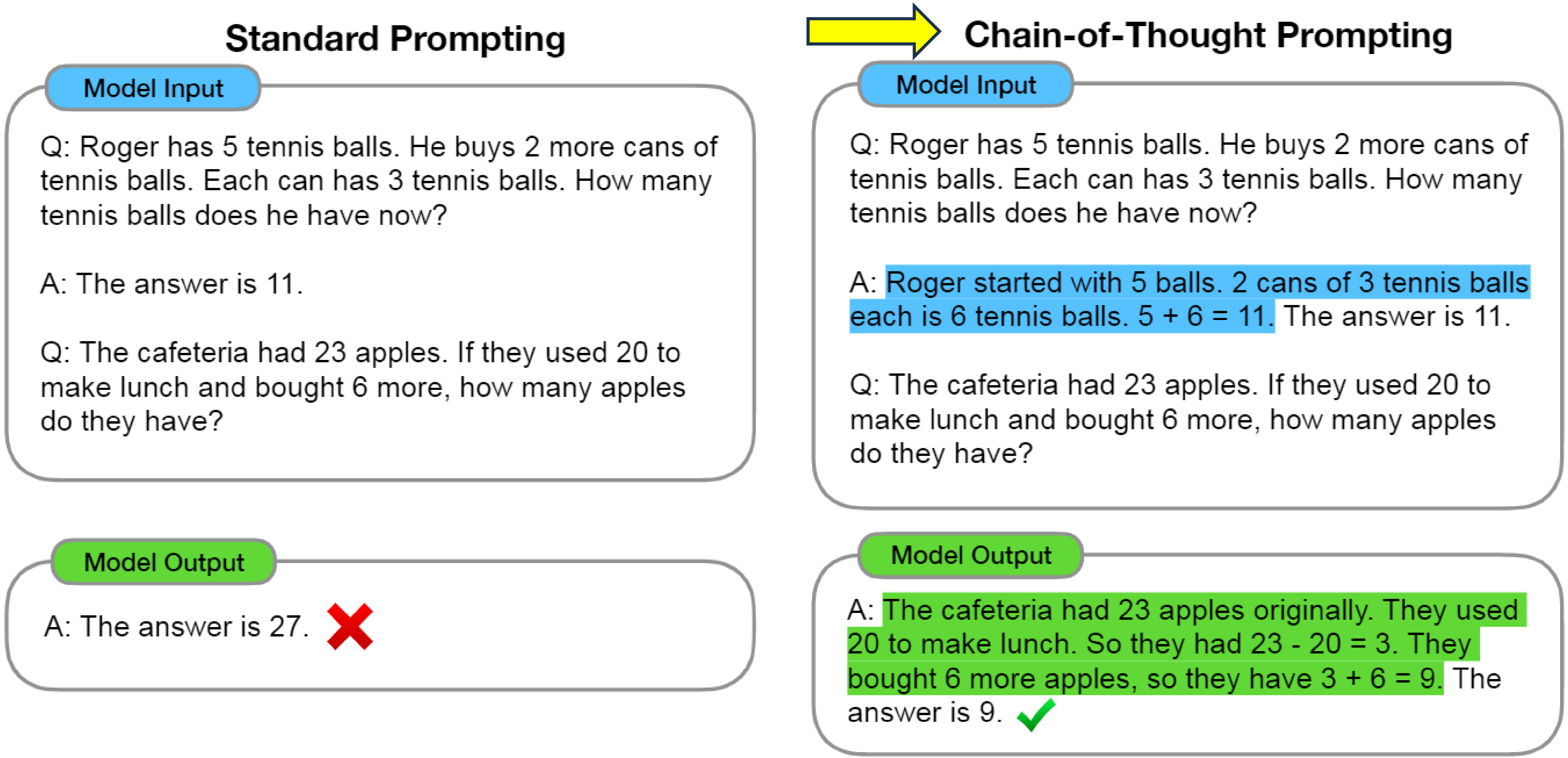
VR headset
Meta Quest Pro

Metawall Screen

LAIKA: Software pipeline



LLMs produce hallucinations



Chain-of-Thought Prompting Elicits Reasoning in Large Language Model, Wei et al., NeurIPS 2022

Chain-of-thought planning

User

Pick up the object that is different from all other objects

Plan: Pick rxbar chocolate

User

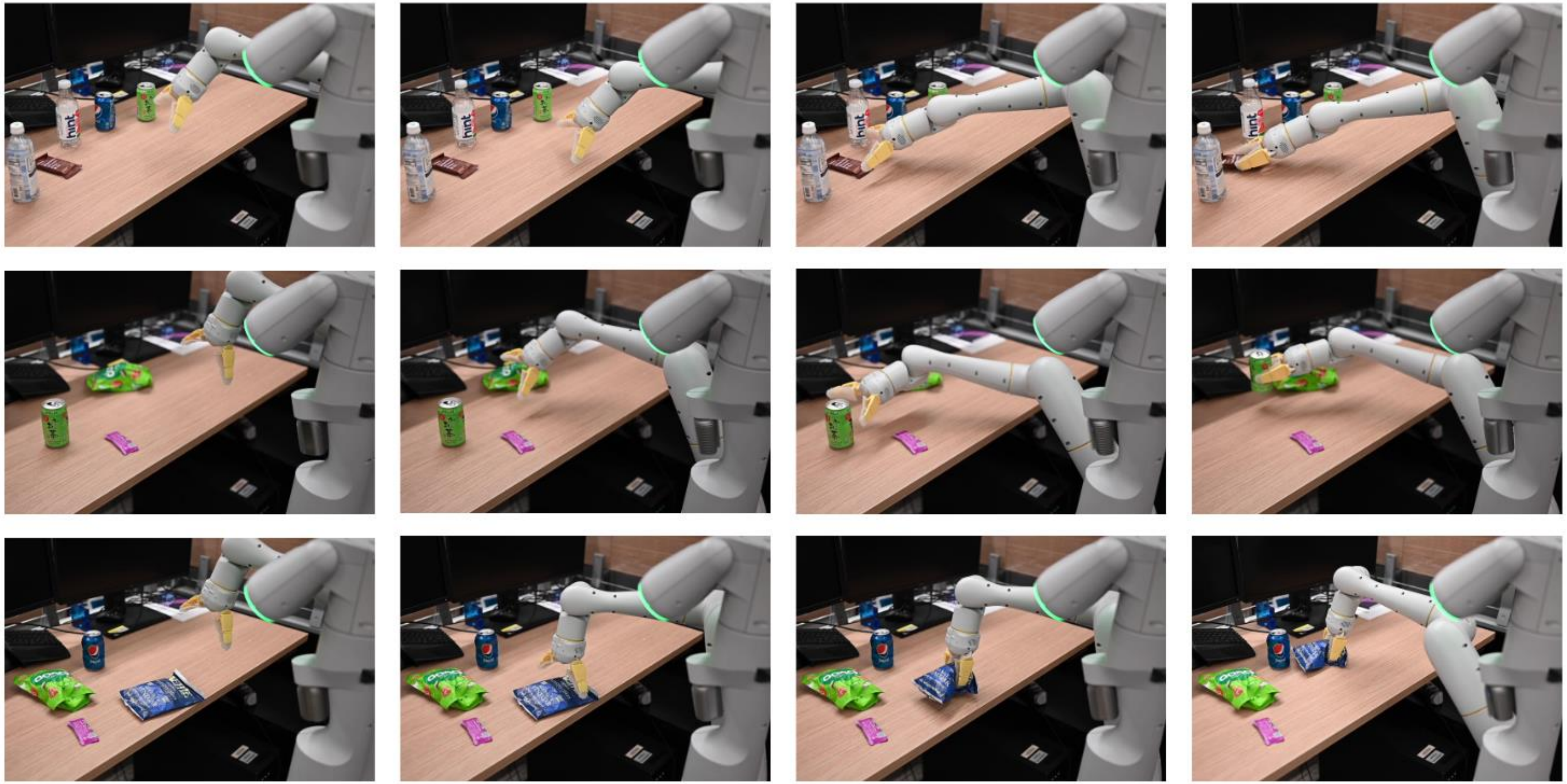
Move the green objects together

Plan: Move green can near green rice chip bag

User

Move the blue objects together

Plan: Move blue chip bag near pepsi



RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control, Brohan et al., CoRL 2023

More recent work on LLM-based planning

- Robots That Ask For Help: Uncertainty Alignment for Large Language Model Planners, Ren et al., CoRL 2023
- Bootstrap Your Own Skills: Learning to Solve New Tasks with Large Language Model Guidance, Zhang et al., CoRL 2023
- VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models, Huang et al., CoRL 2023
- Open-World Object Manipulation using Pre-Trained Vision-Language Models, Stone et al., CoRL 2023
- Language-guided Robot Grasping: CLIP-based Referring Grasp Synthesis in Clutter, Tziafas et al., CoRL 2023
- SLAP: Spatial-Language Attention Policies, Parashar et al., CoRL 2023
- Large Language Models as Commonsense Knowledge for Large-Scale Task Planning, Zhao et al., NeurIPS 2023
- ProgPrompt: Generating Situated Robot Task Plans using Large Language Models, Singh et al., ICRA 2023
- ManipLLM: Embodied Multimodal Large Language Model for Object-Centric Robotic Manipulation, Li et al., CVPR 2024

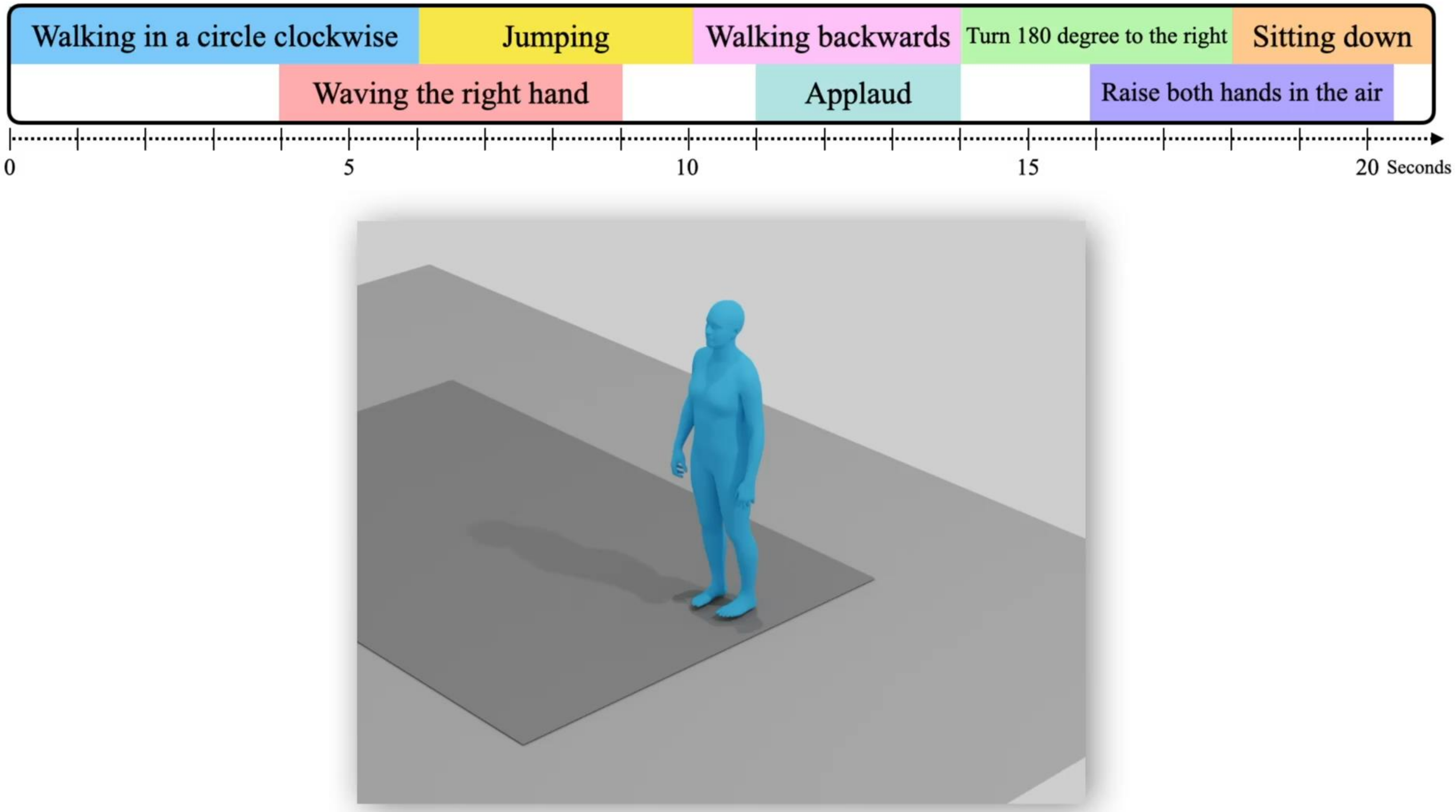
Beyond Robotics: Animation

Recent Advances in Video Generation






Recent Work in Animation

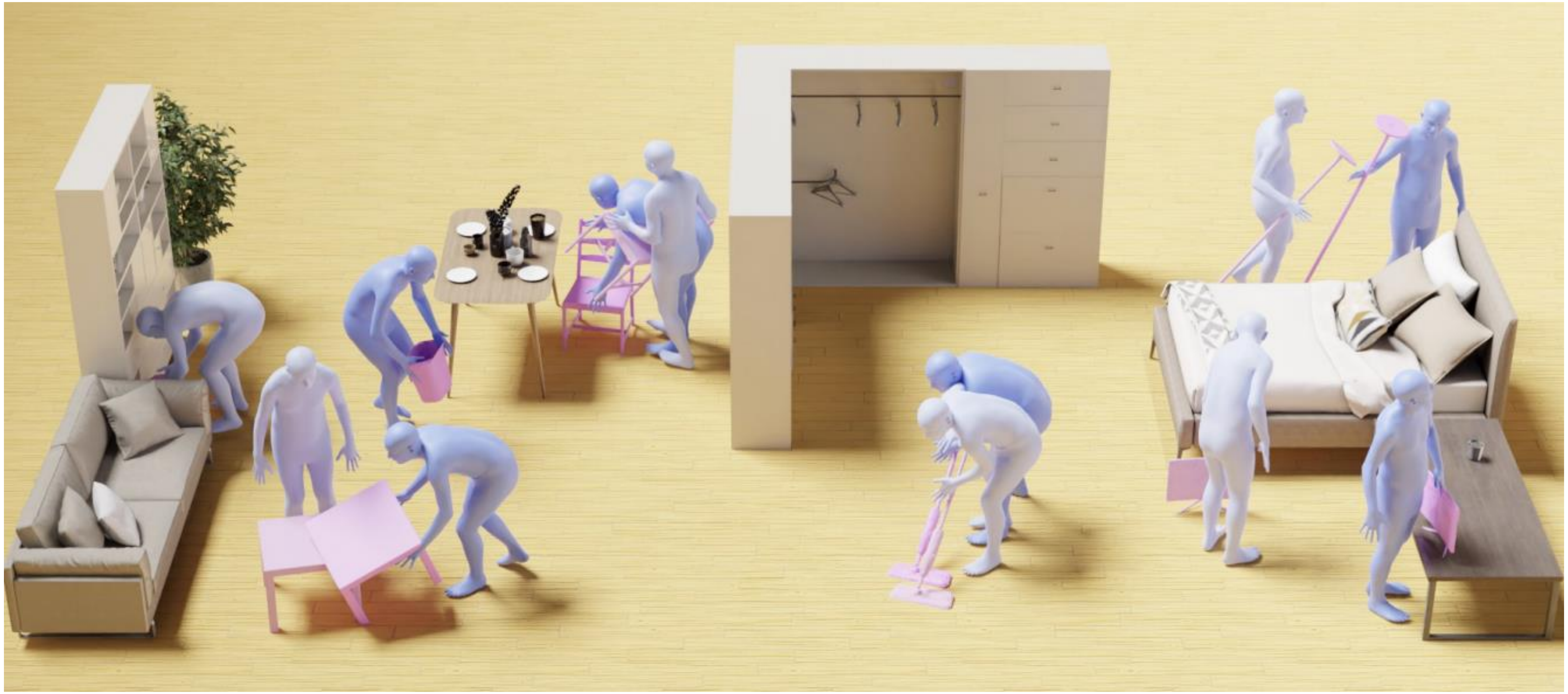


Multi-Track Timeline Control for Text-Driven 3D Human Motion Generation. *Petrovich et al., CVPRW 2024*

Recent Work in Animation

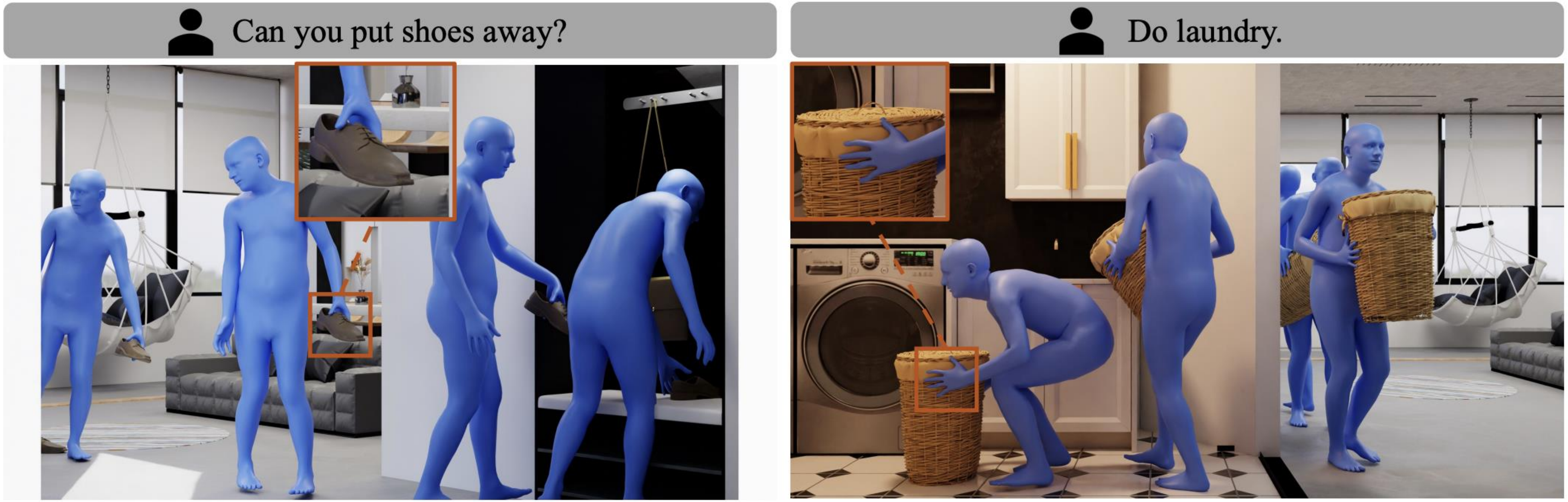
GMD: Guided Motion Diffusion for Controllable Human Motion Synthesis. *Karunratanakul et al., ICCV 2023*

Recent Work in Animation



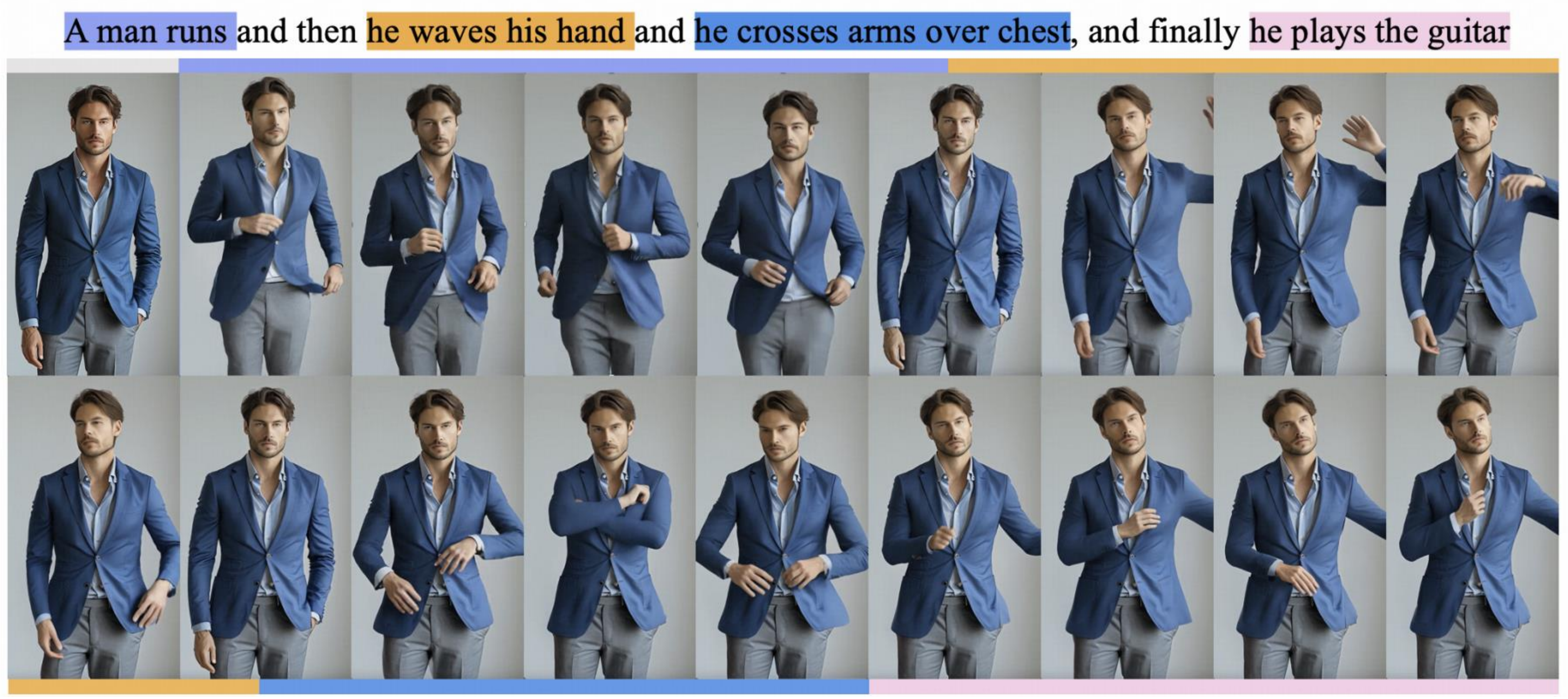
Object Motion Guided Human Motion Synthesis. *Li et al., arXiv 2023*

Recent Work in Animation



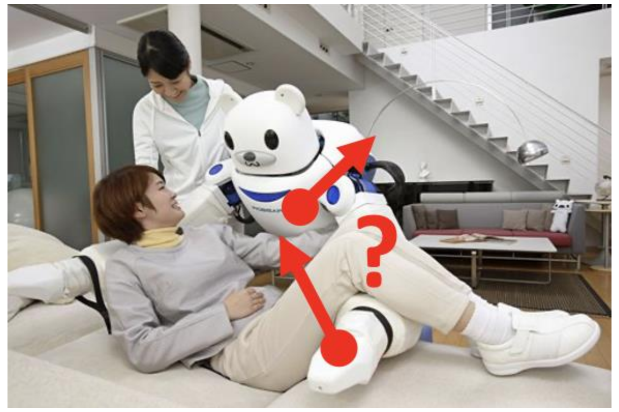
Human-Object Interaction from Human-Level Instructions. *Wu et al., arXiv 2024*

Recent Work in Animation

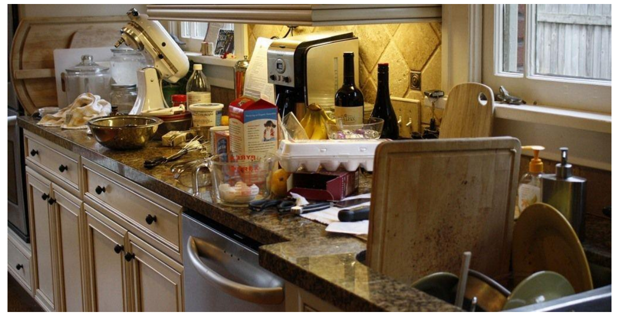


Fleximo: Towards Flexible Text-to-Human Motion Video Generation. *Zhang et al., arXiv Nov 2024*

Summary



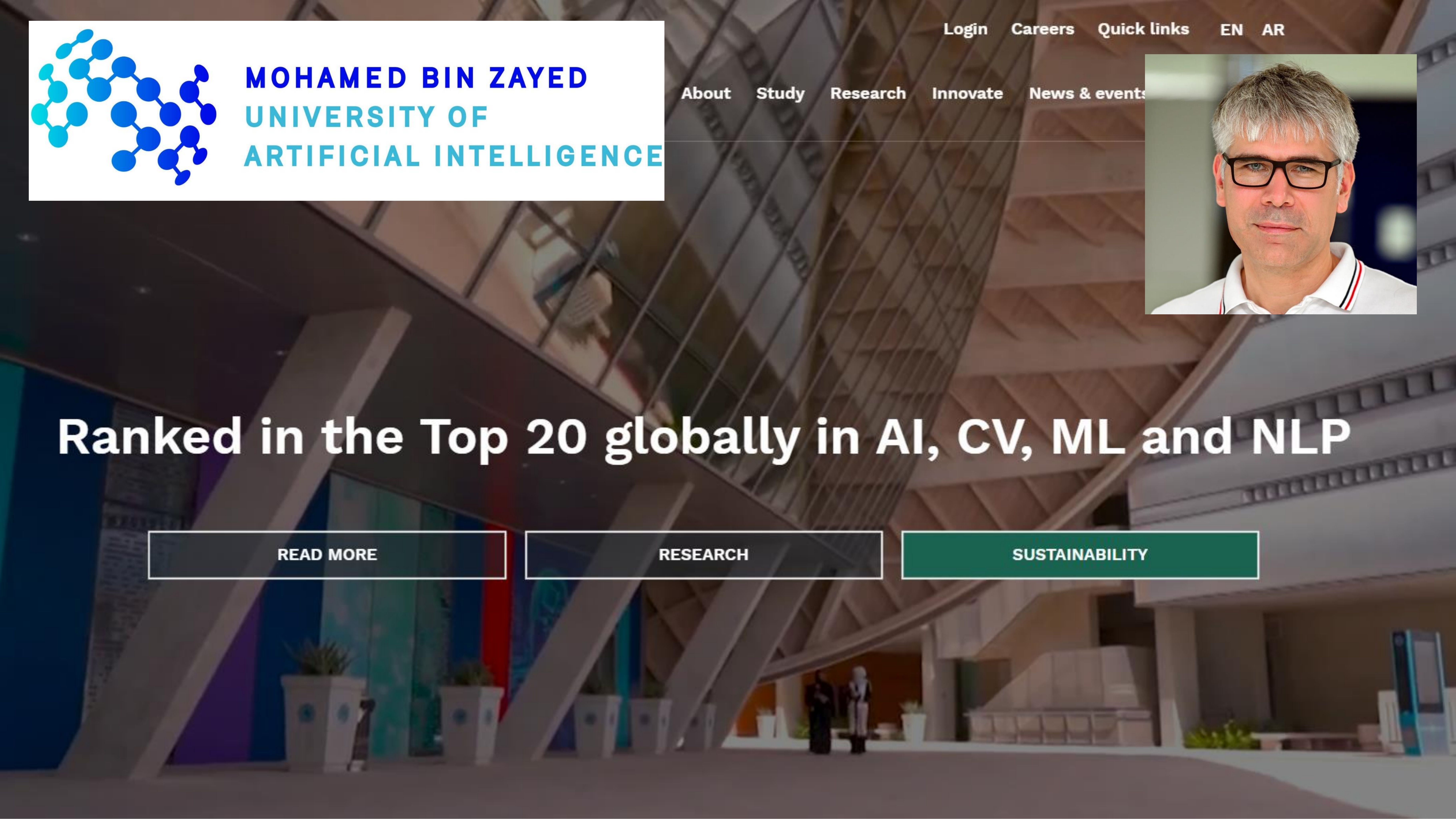
- Learning from human demonstrations



- Learning from video

- Language-driven planning

- Language-driven animation


MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE
Login
Careers
Quick links
EN
AR
About
Study
Research
Innovate
News & events
Ranked in the Top 20 globally in AI, CV, ML and NLP
READ MORE
RESEARCH
SUSTAINABILITY

MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

Embodied Perception

Models and learning methods for embodied computer vision

Ivan Laptev


Fabio Pizzati


Rocktim Jyoti Das


RidouaneGhermi


Amine Boudjoghra


Kamila Zhumakhanova


Rikhat Akizhanov


Harsh Singh


Rusiru Achchige


Mukul Ranjan

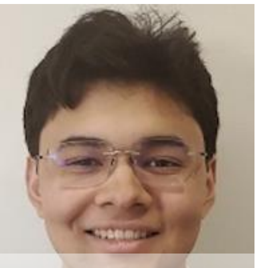

Mingfei Han


Kartik Kuckreja


Junaid Ansari

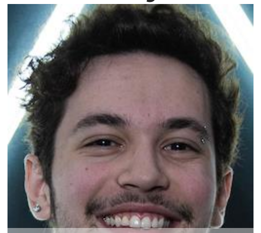

David Romero


Diana Turmakhan


Jason Banks


Abdul Butt


Abdullah Sohail


Gustavo Stahl

Postdocs *RAs* *PhD students* *MSc students*

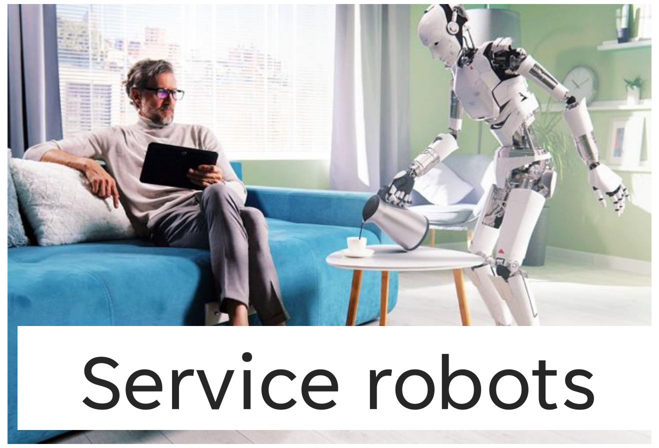
MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

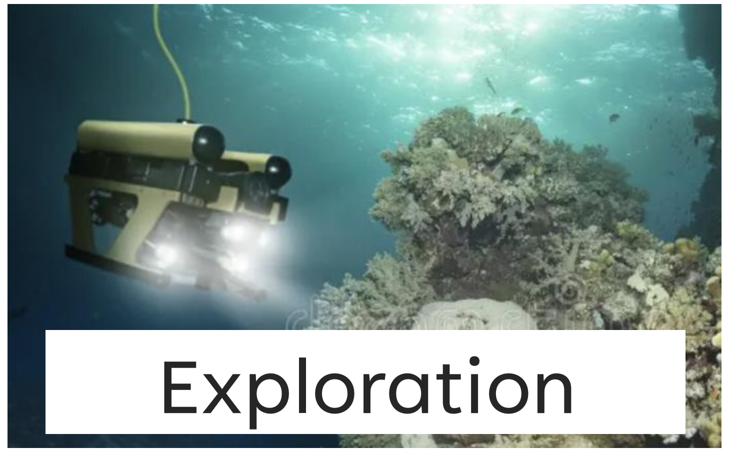
Embodied Perception

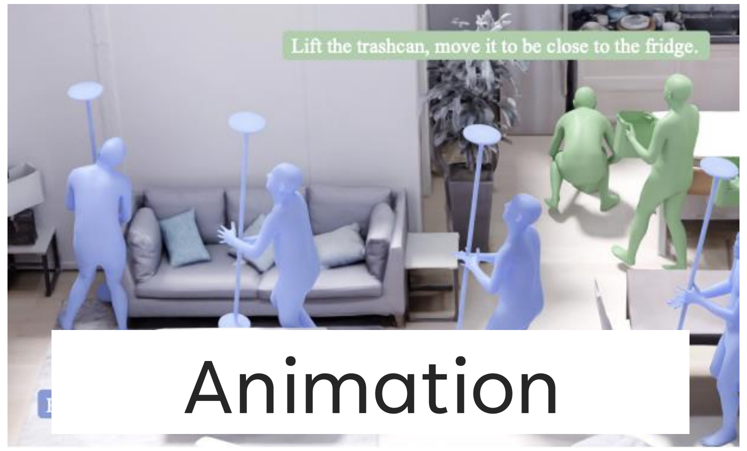
What will happen to the scene after action X? (prediction)

Physics-informed
Language-aware
Sensor-driven
Learnable
World Models

What actions are needed for state transition A → B? (planning)

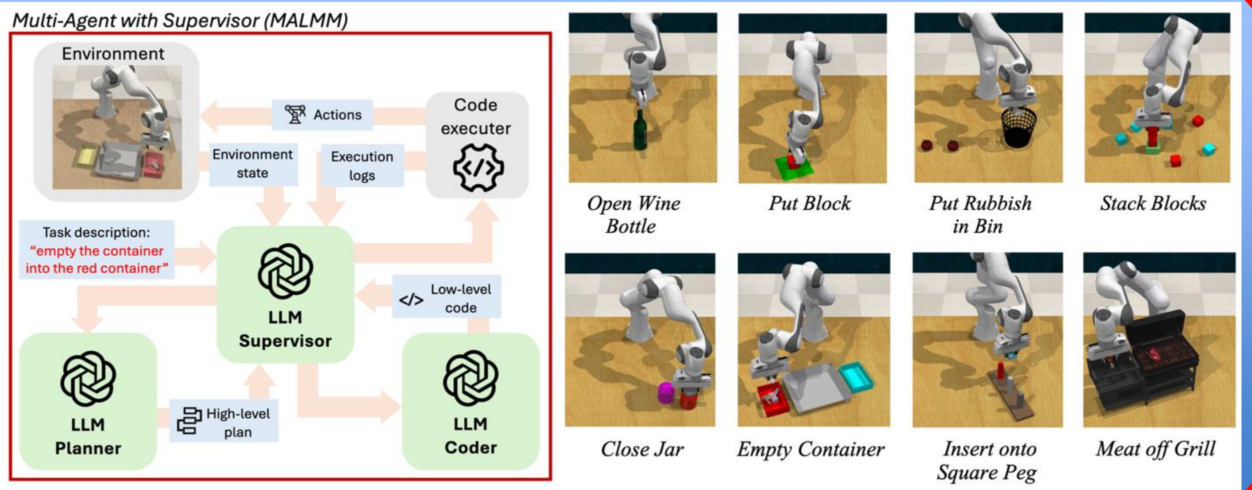


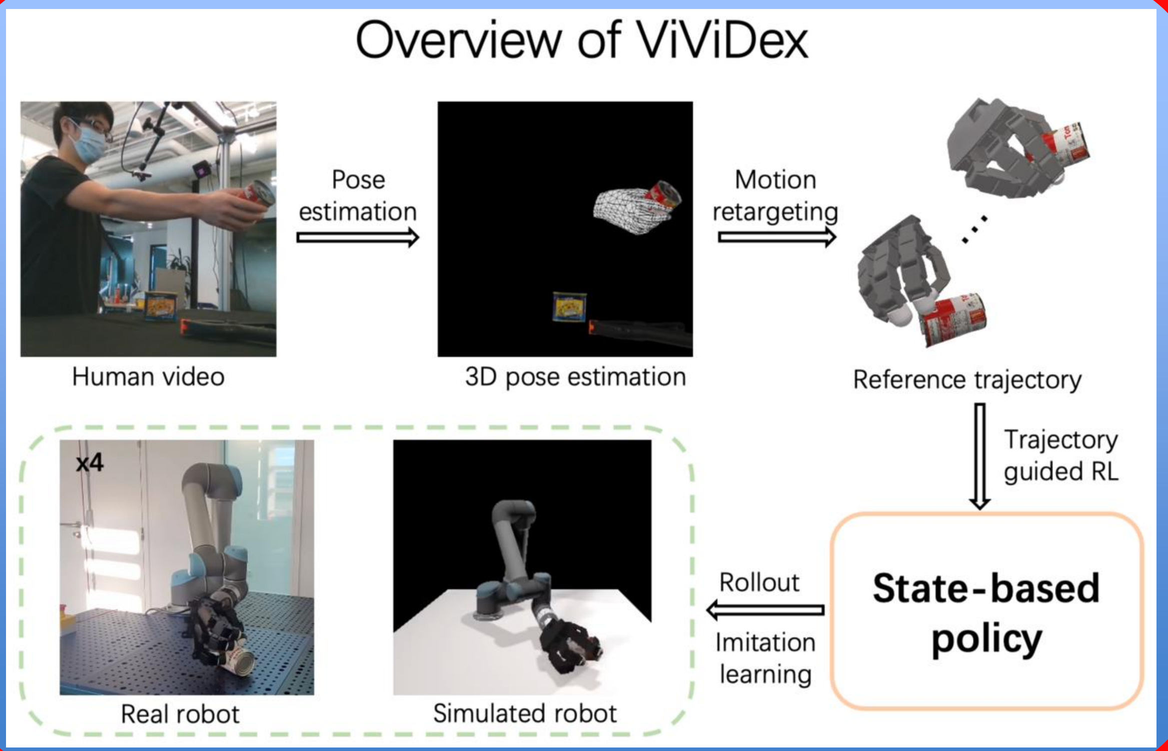


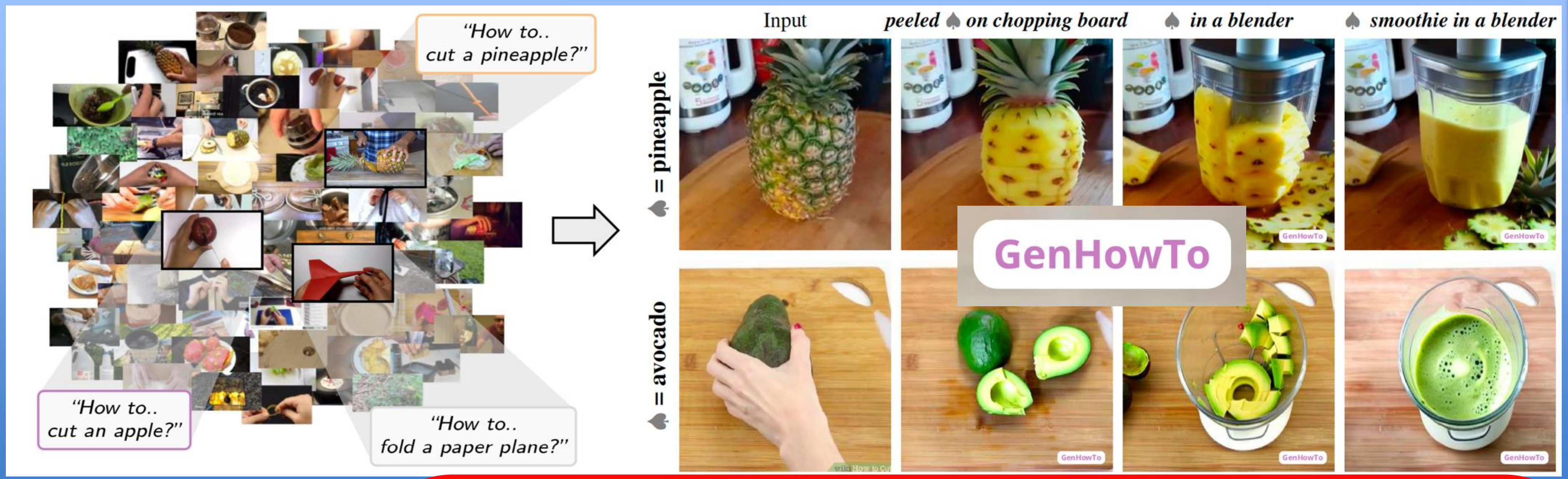


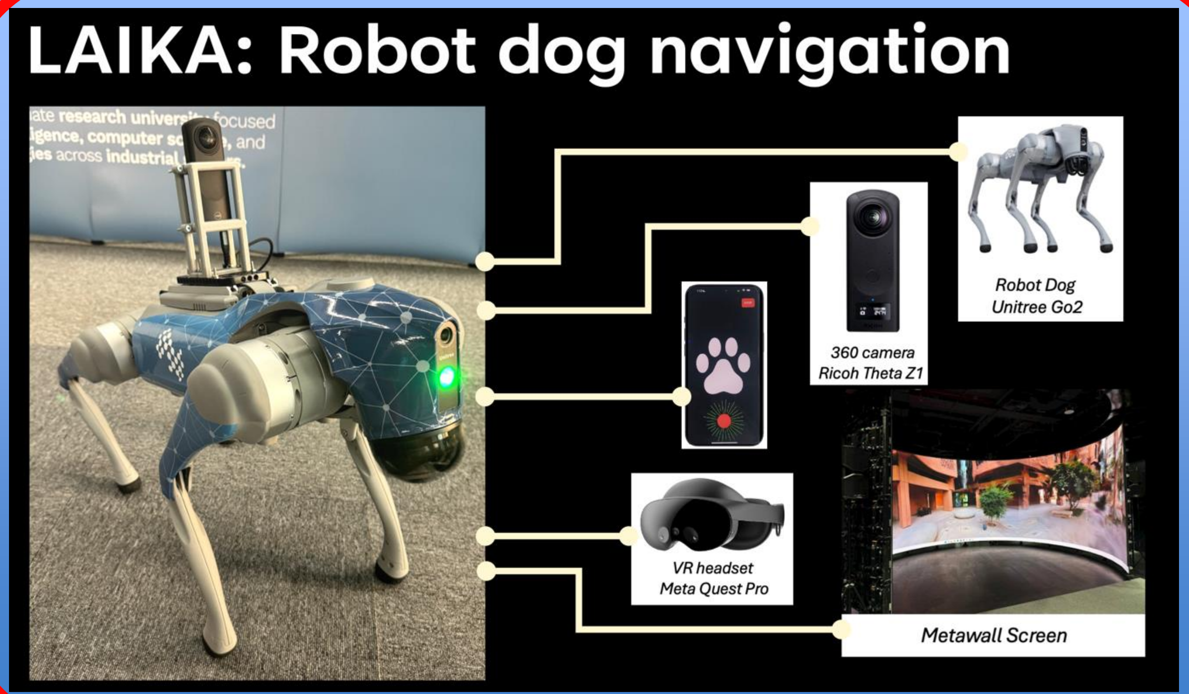
MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

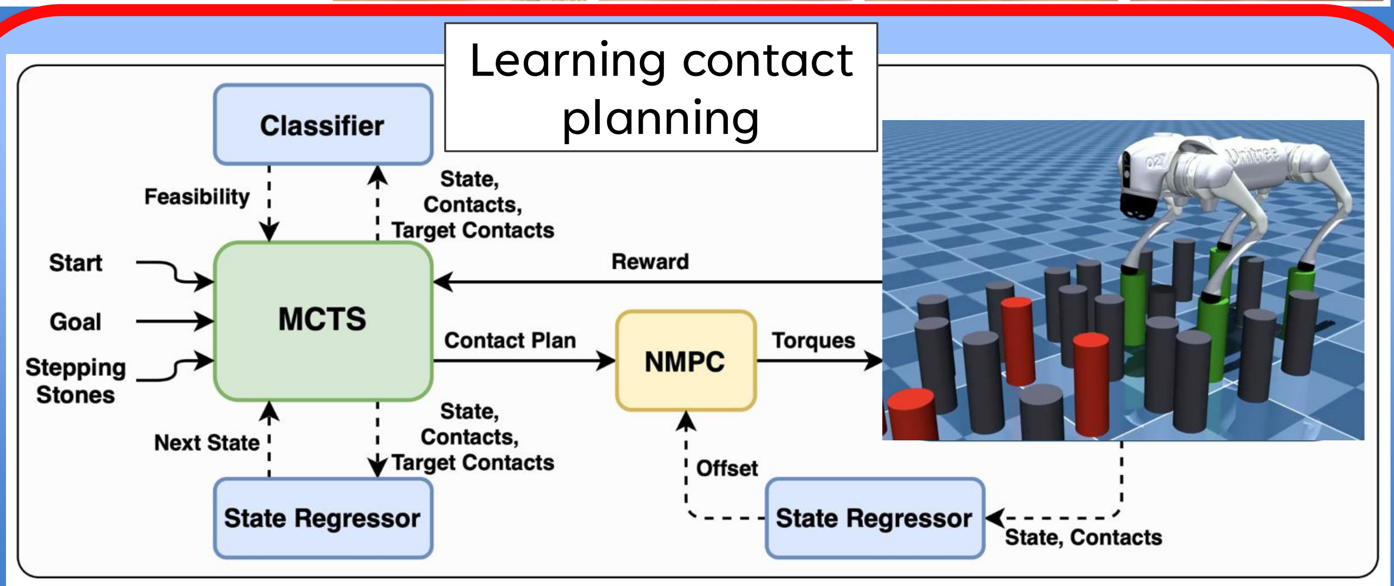
EP Team: Recent projects











MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

Building a new lab for Embodied Perception

- Internships are available
- PhD application is open
- Competitive Internship and PhD salaries
- Departments of CV, NLP, ML, Robotics

Contact: Ivan.Laptev@mbzuai.ac.ae