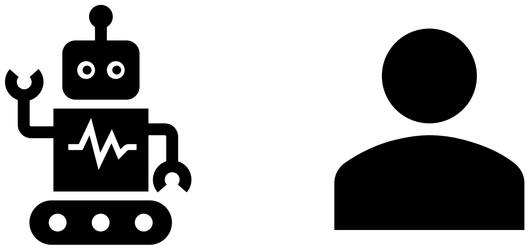
Augmenting Robotic Capabilities with Natural Language

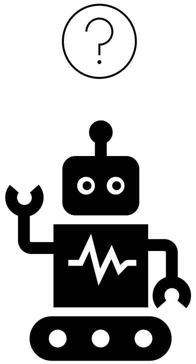
Albert Yu

October 27, 2025

Home Robots



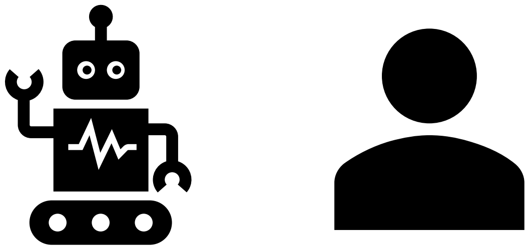
$\sim 1000s$ of demos



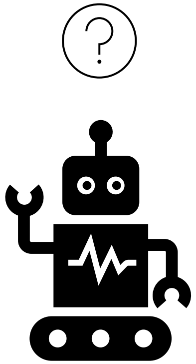
Demonstrations are expensive

Incapable of complex situations/tasks

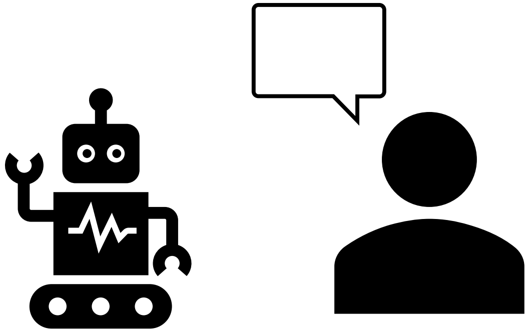
Language for Home Robots



$\sim 1000s$ of demos

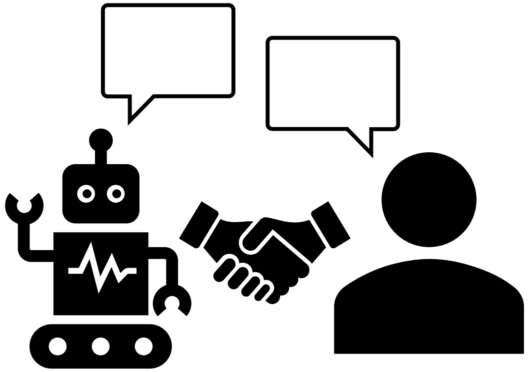


Task Instructions for Sample Efficiency



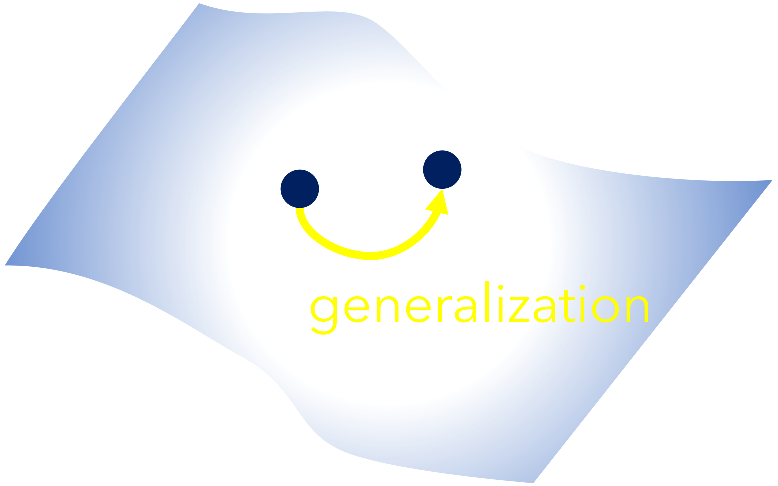
$\sim 10s$ of demos

Collaboration for Challenging Tasks

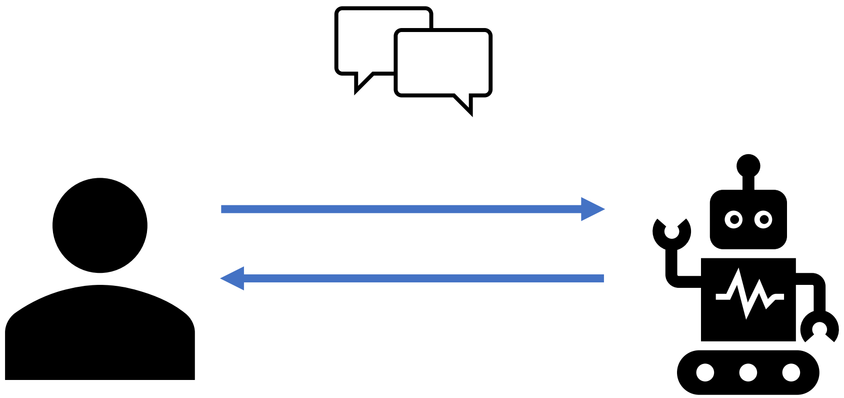


Language to Improve Robotic Capabilities

Semantics for Generalization



Communicative Medium



Outline

1. Language as Semantics for Generalization
   a) DeL-TaCo: Demo-Language Task Conditioning
   b) Lang4Sim2Real: Language to Bridge the Sim2Real Gap
2. Language as a Communicative Medium
   a) MicoBot: Bidirectional Dialog for Human-Robot Collaboration
3. Future Directions

Outline

1. Language as Semantics for Generalization
   a) DeL-TaCo: Demo-Language Task Conditioning
   b) Lang4Sim2Real: Language to Bridge the Sim2Real Gap
2. Language as a Communicative Medium
   a) MicoBot: Bidirectional Dialog for Human-Robot Collaboration
3. Future Directions

Using Both Demonstrations and Language Instructions to Efficiently Learn Robotic Tasks

Collaborators: Ray Mooney

ICLR 2023

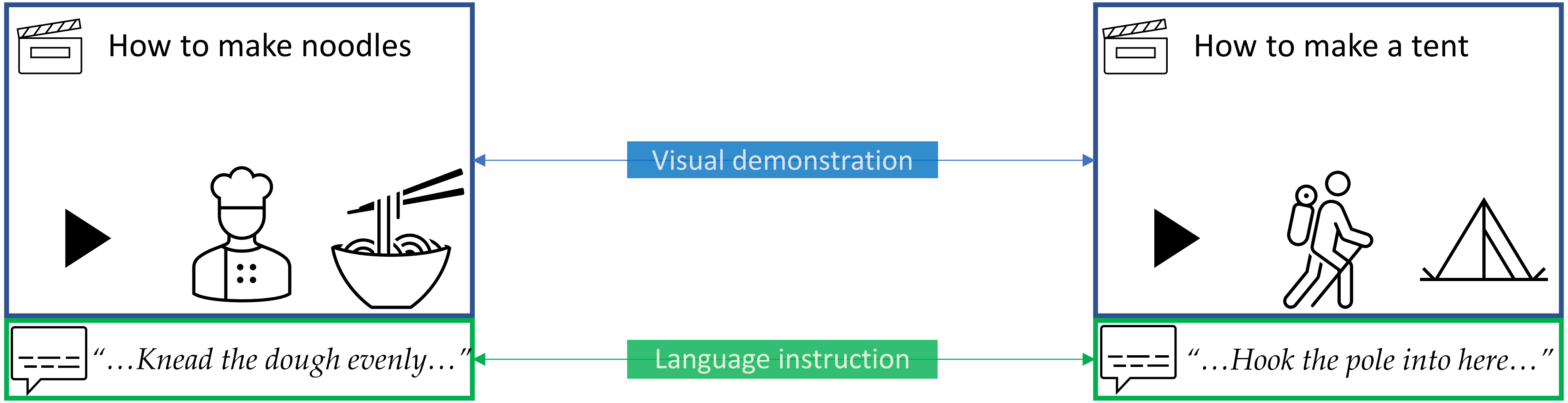
Learning new skills

Language instructions only

Demonstrations only

Imagine learning from only one modality…

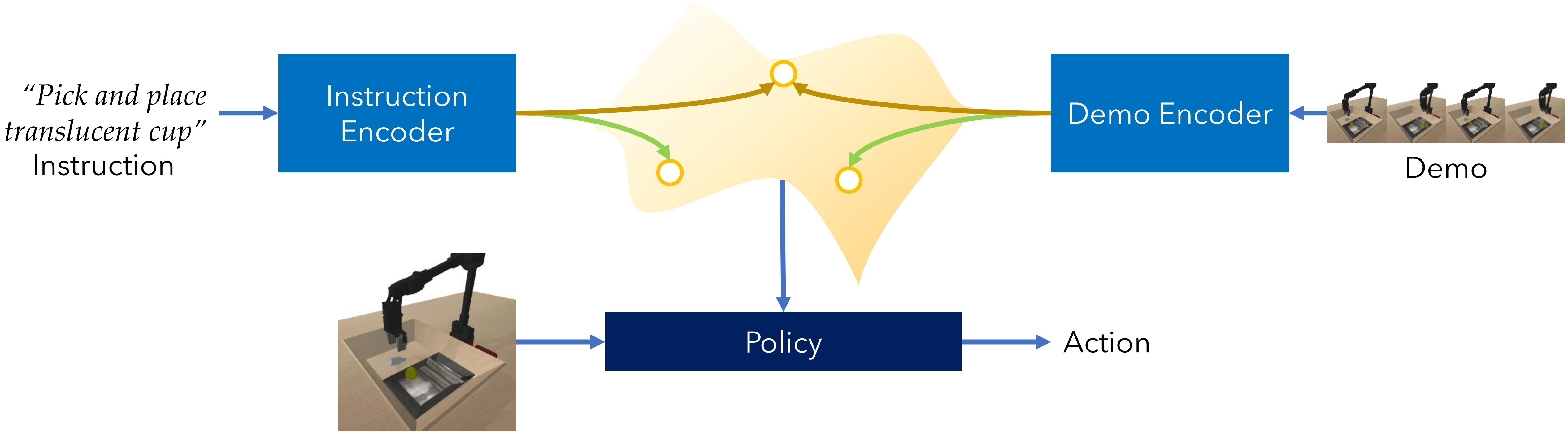
Hypothesis: Multimodality helps Robots learn new tasks



Prior Work

MCIL (RSS 2021)
Lynch & Sermanet

BC-Z (CoRL 2022)
Jang, et. al.

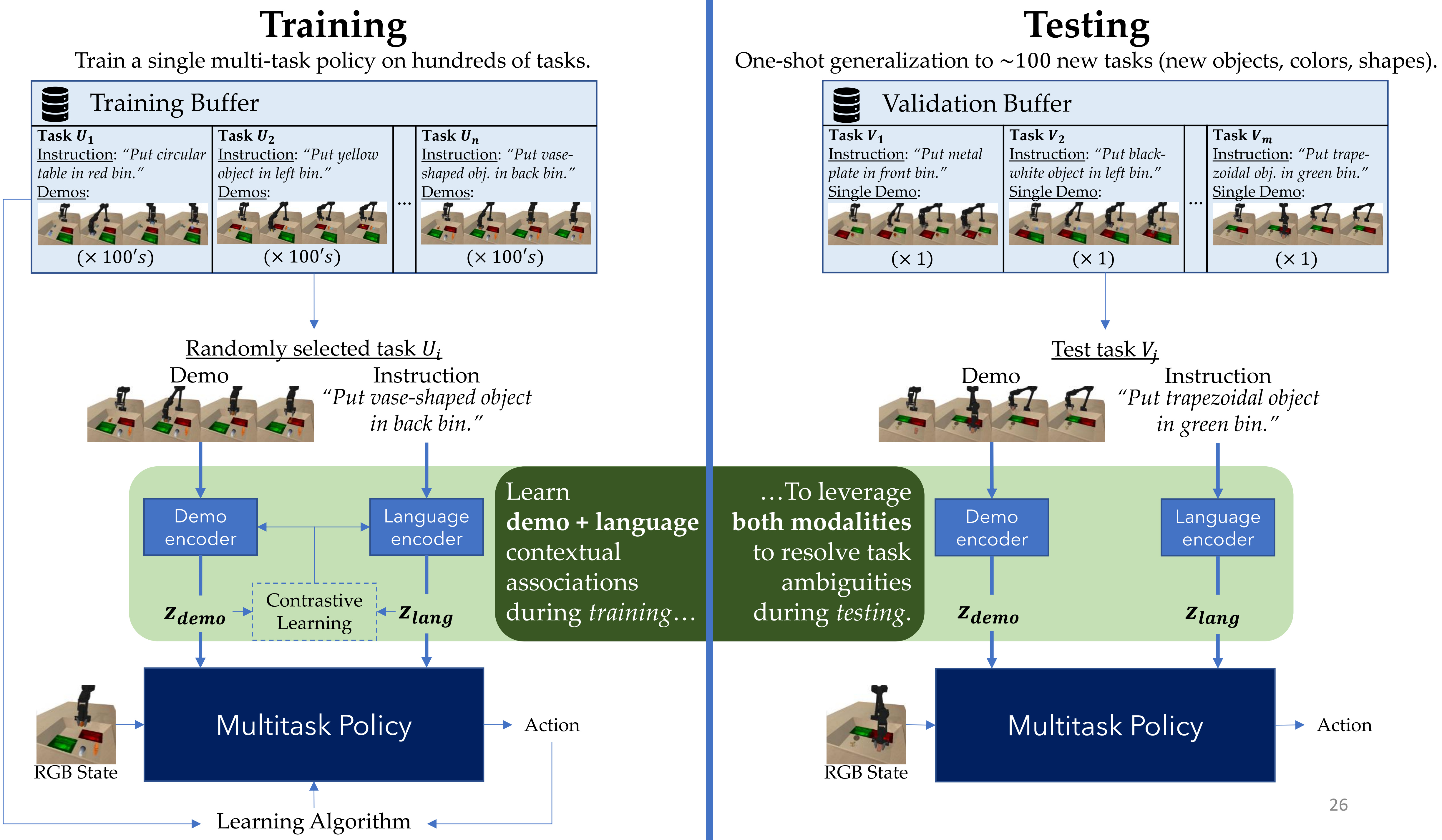


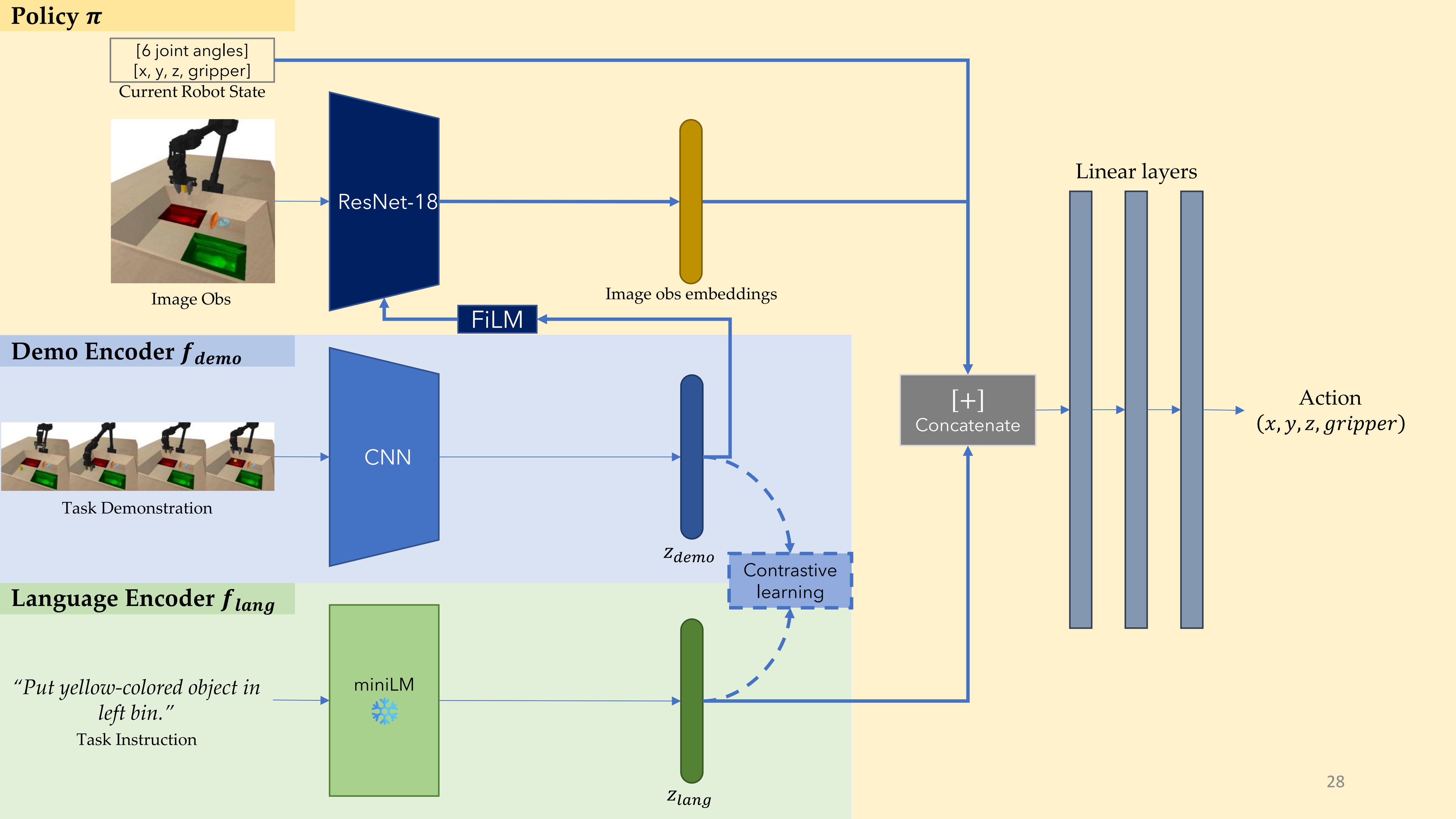
Existing methods do not simultaneously use both demonstrations and language to teach robots new tasks.

BC-Z authors: Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, Chelsea Finn

DeL-TaCo
Joint Demo-Language Task Conditioning

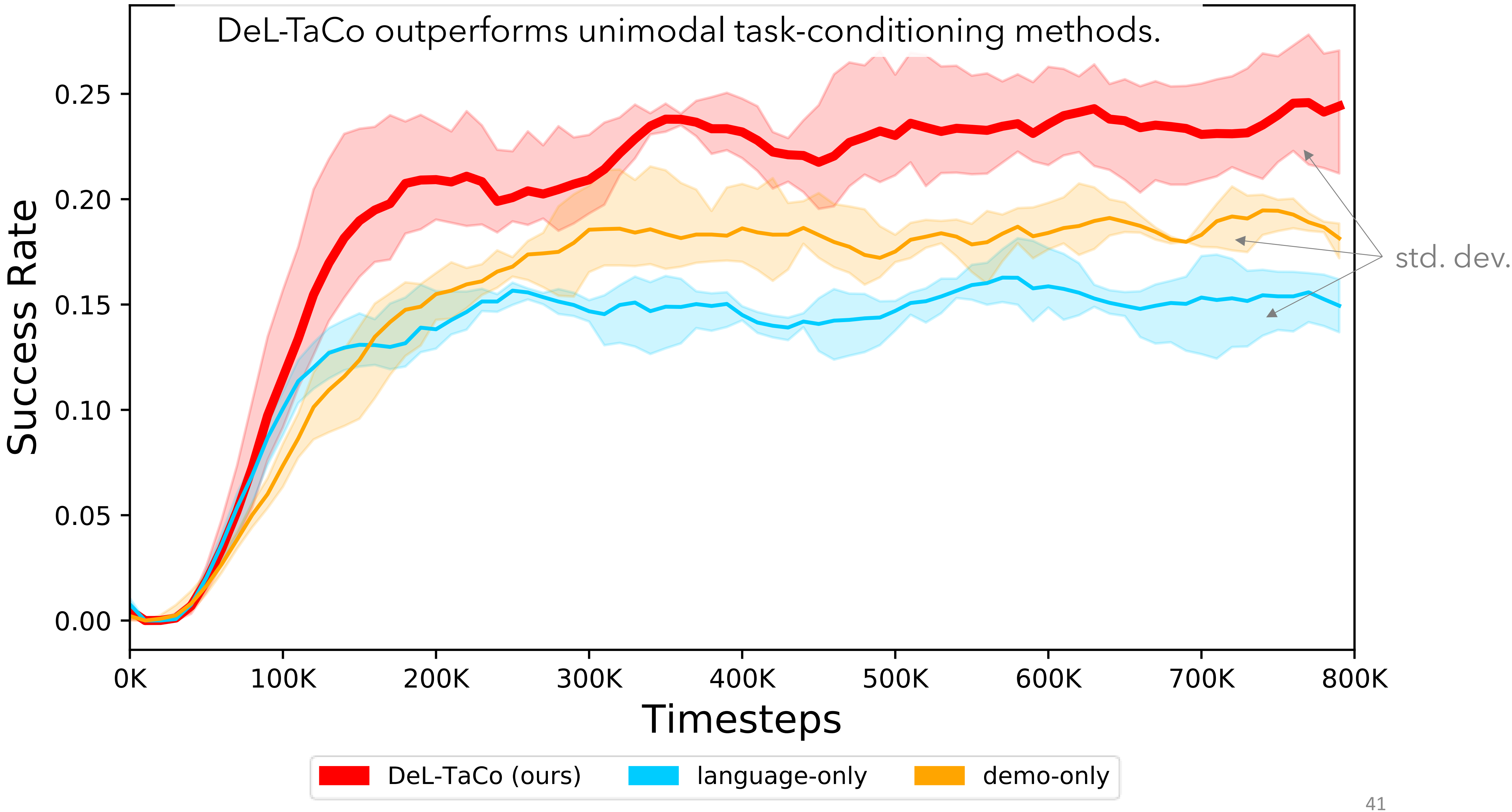
Leverage benefits of learning with **both demonstrations and language**


Training
Train a single multi-task policy on hundreds of tasks.
Training Buffer
Task U_1
Instruction: "Put circular table in red bin."
Demos:
(× 100's)
Task U_2
Instruction: "Put yellow object in left bin."
Demos:
(× 100's)
...
Task U_n
Instruction: "Put vase-shaped obj. in back bin."
Demos:
(× 100's)
Randomly selected task U_i
Demo
Instruction
"Put vase-shaped object in back bin."
Demo encoder
Language encoder
z_{demo}
Contrastive Learning
z_{lang}
Learn demo + language contextual associations during training…
RGB State
Multitask Policy
Action
Learning Algorithm
Testing
One-shot generalization to ~100 new tasks (new objects, colors, shapes).
Validation Buffer
Task V_1
Instruction: "Put metal plate in front bin."
Single Demo:
(× 1)
Task V_2
Instruction: "Put black-white object in left bin."
Single Demo:
(× 1)
...
Task V_m
Instruction: "Put trape-zoidal obj. in green bin."
Single Demo:
(× 1)
Test task V_j
Demo
Instruction
"Put trapezoidal object in green bin."
…To leverage both modalities to resolve task ambiguities during testing.
Demo encoder
Language encoder
z_{demo}
z_{lang}
RGB State
Multitask Policy
Action



Policy π
[6 joint angles]
[x, y, z, gripper]
Current Robot State
Image Obs
ResNet-18
Image obs embeddings
FiLM
Linear layers
Demo Encoder f_{demo}
CNN
Task Demonstration
z_{demo}
[+]
Concatenate
Action
$(x, y, z, gripper)$
Contrastive learning
Language Encoder f_{lang}
miniLM
"Put yellow-colored object in left bin."
Task Instruction
z_{lang}

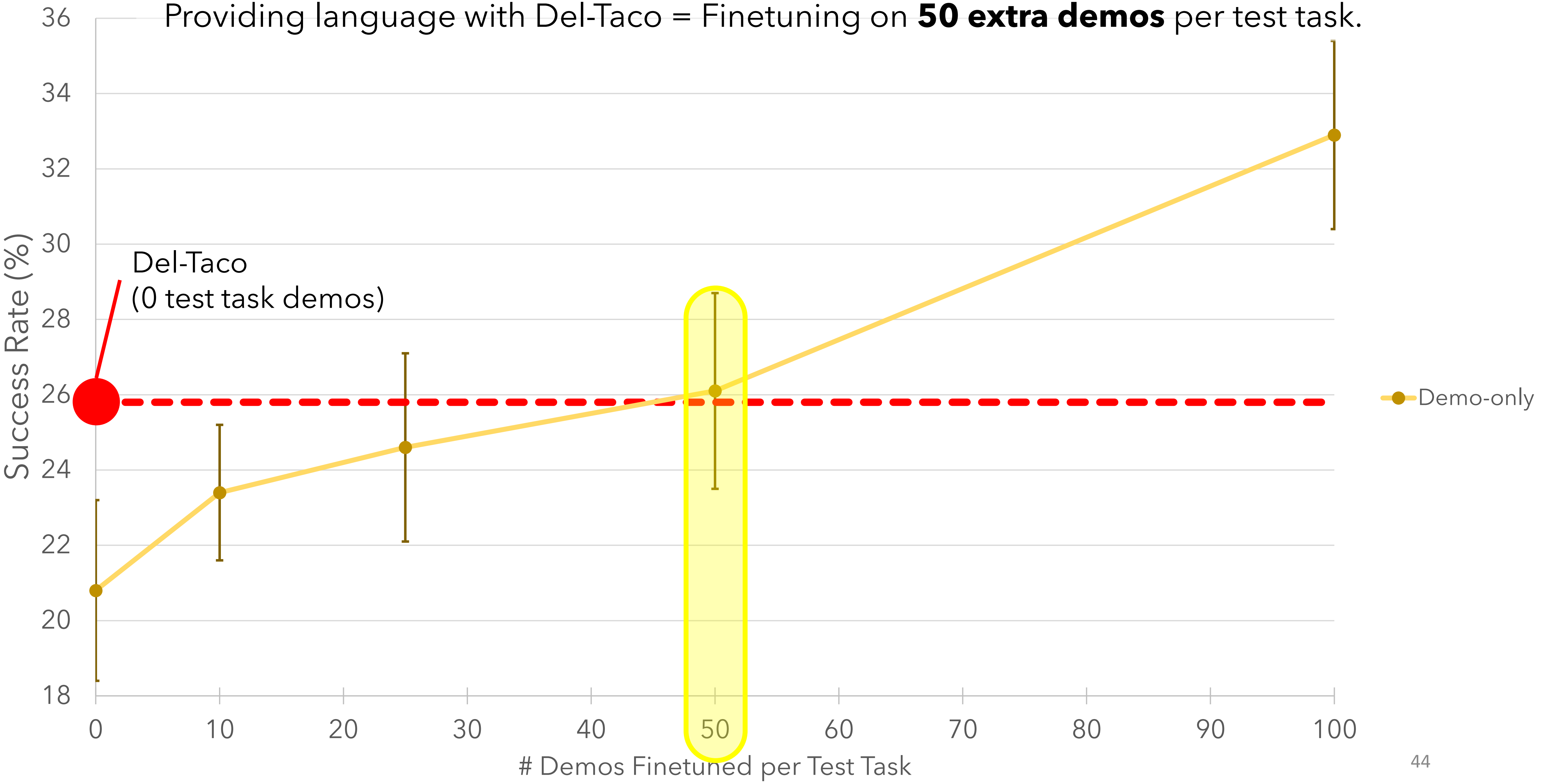
Generalization Performance on Novel Objects, Colors, Shapes



3 seeds per line

How Many Demonstrations is Language Worth?

Providing language with Del-Taco = Finetuning on **50 extra demos** per test task.



Key Takeaways

Generalization: Learning from both demonstrations and language with DeL-TaCo leads to better generalization than learning from one modality.

Sample-Efficiency: DeL-TaCo reduces the need for ~50 demonstrations over policies conditioned only on demonstrations.

Outline

1. Language as Semantics for Generalization
    a) DeL-TaCo: Demo-Language Task Conditioning
    b) Lang4Sim2Real: Language to Bridge the Sim2Real Gap
2. Language as a Communicative Medium
    a) MicoBot: Bidirectional Dialog for Human-Robot Collaboration
3. Future Directions

Natural Language Can Help Bridge the Sim2Real Gap

https://robin-lab.cs.utexas.edu/lang4sim2real/ | https://arxiv.org/pdf/2405.10020

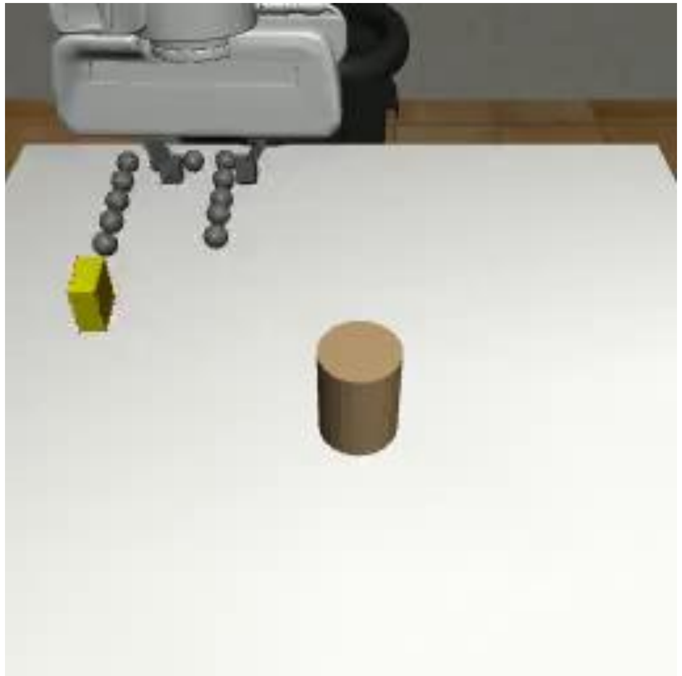
Collaborators: Ray Mooney, Roberto Martín-Martín, Addie Foote

RSS 2024

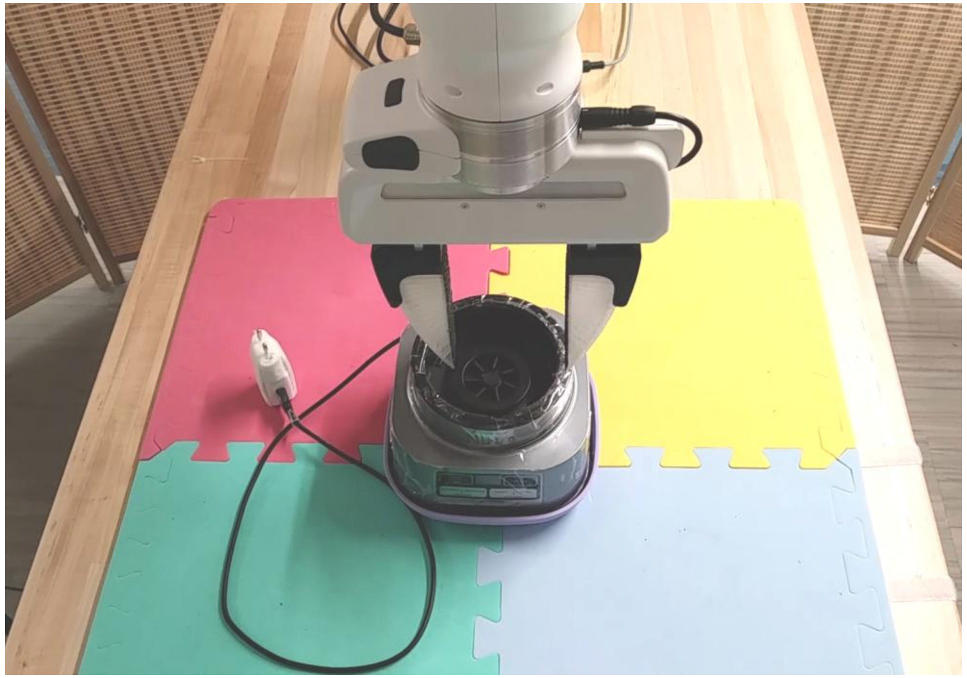
Sim2Real

Sim Data $

Real-world Data $$$$

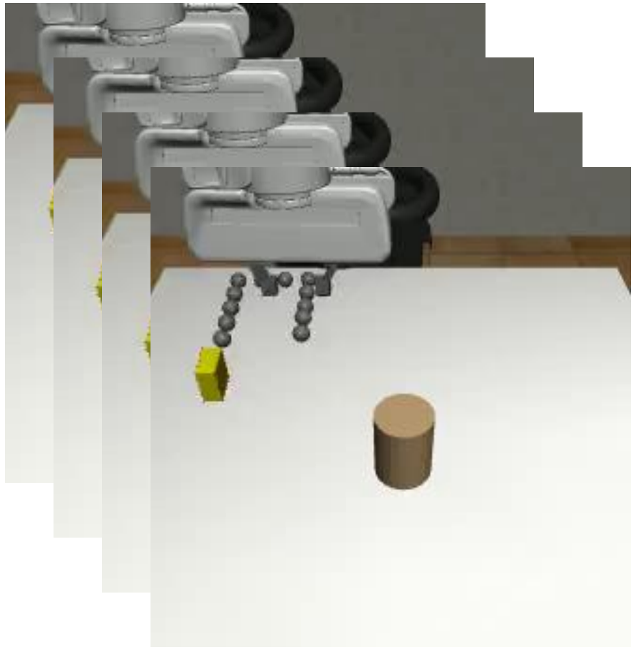


Domain Gap

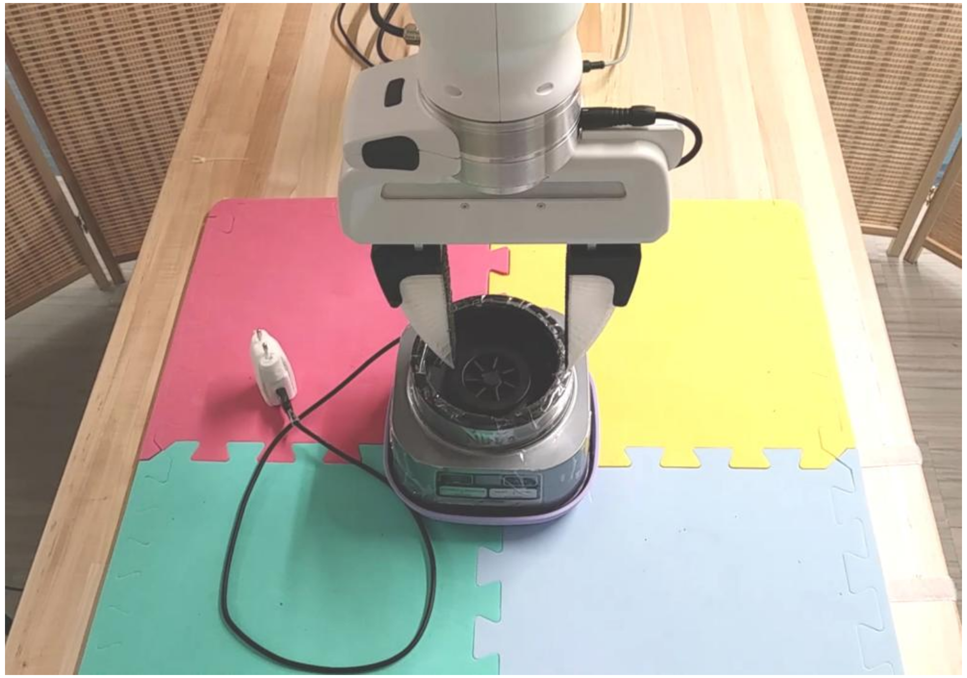


Hypothesis: Semantic information in language enables robots to bridge wide visual sim2real gaps.

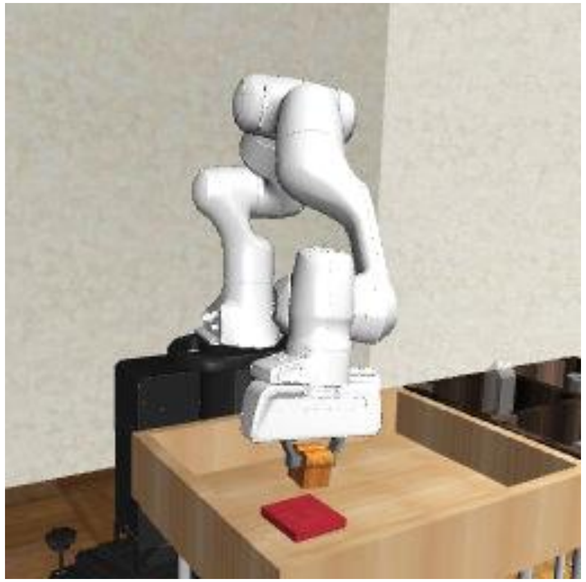
Sim



Real



Semantically Similar States → Similar Actions

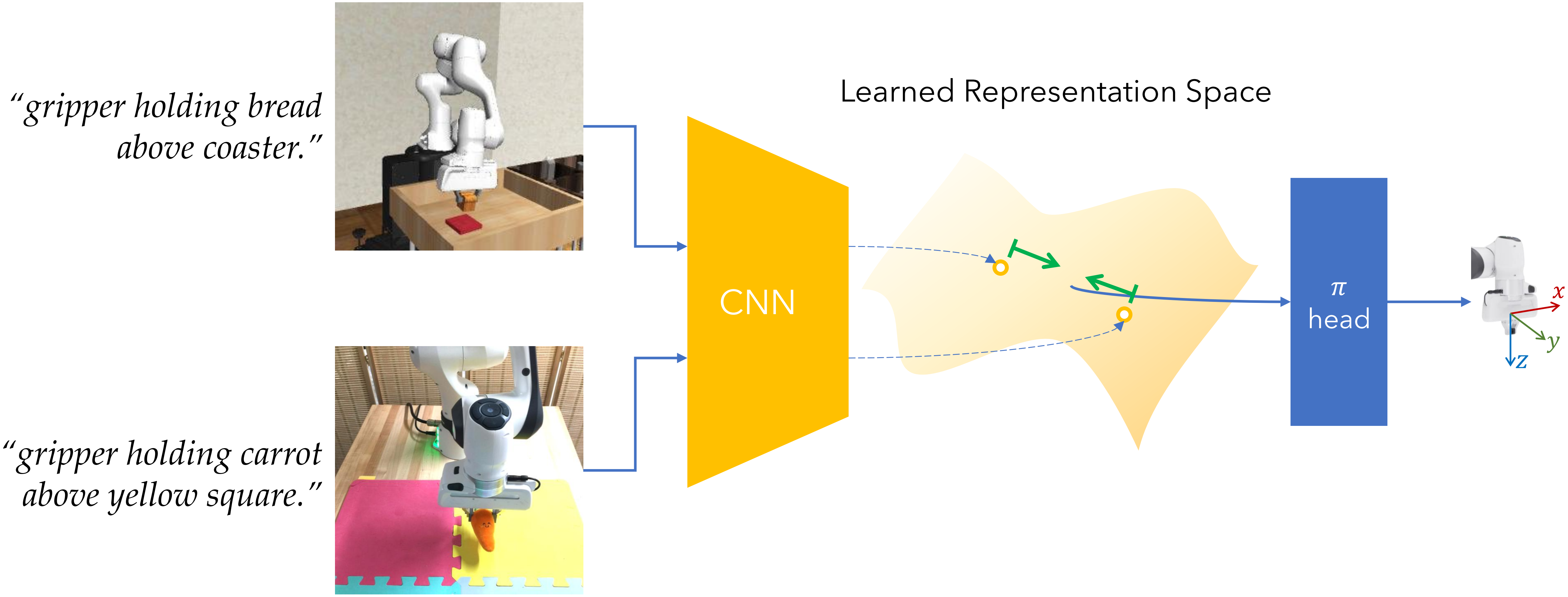


“gripper holding bread above coaster.”

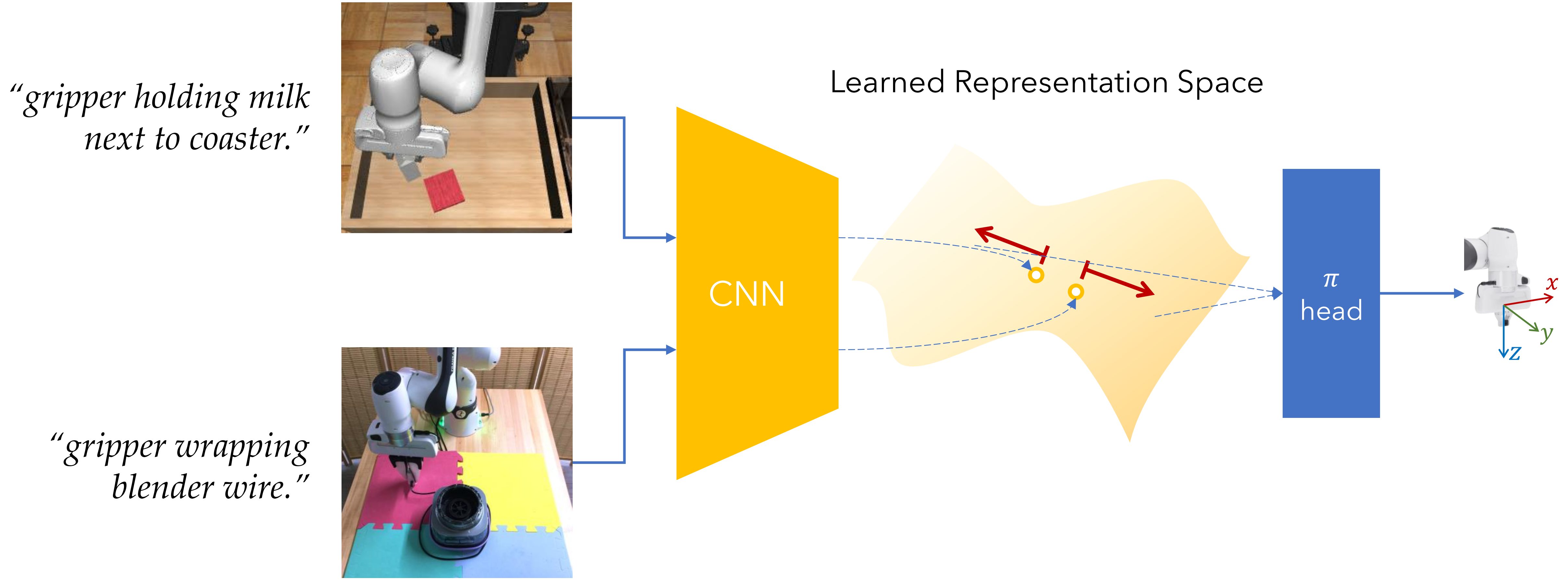
“gripper holding carrot above yellow square.”

From both of these states, the robot should open the gripper to place the object.

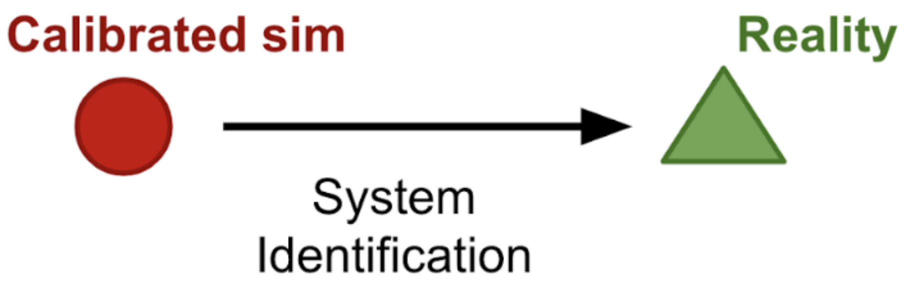
Domain Invariant Representations



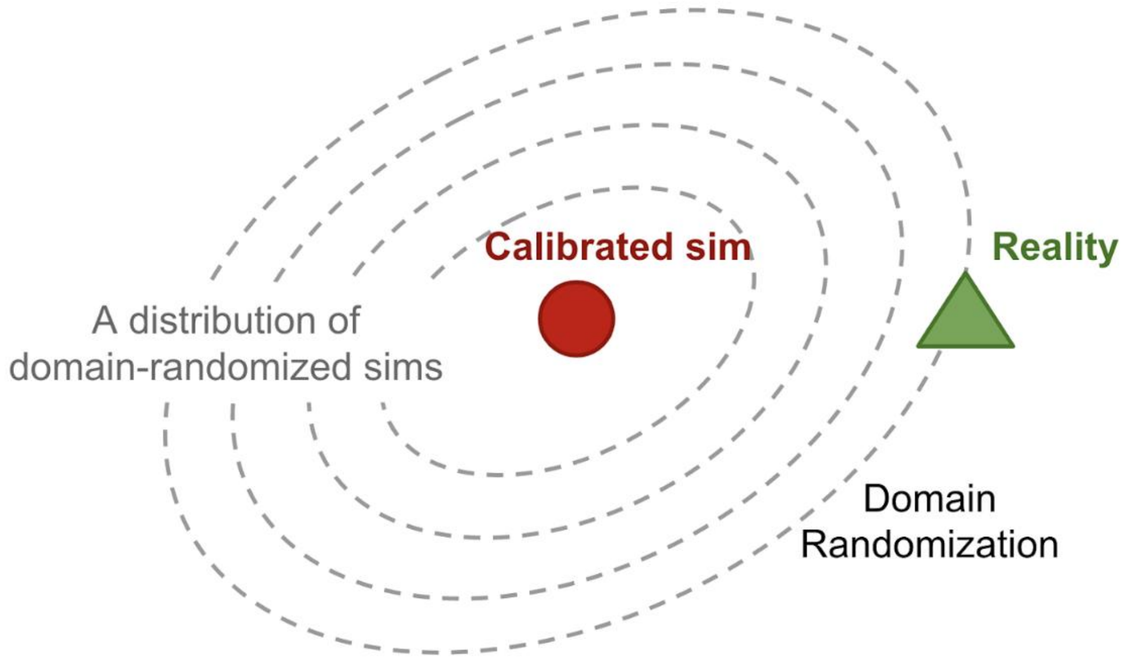
Domain Invariant Representations



Prior Work: Sim2Real



(Yu+ 2017)



(Tobin+ 2017)

Both approaches are very engineering-intensive.
Domain Randomization leads to overly conservative policies.

https://lilianweng.github.io/posts/2019-05-05-domain-randomization/
Yu+2017: Wenhao Yu, Jie Tan, C Karen Liu, and Greg Turk. "Preparing for the unknown: Learning a universal policy with online system identification." RSS 2017.
Tobin+2017: Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. "Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World." IROS 2017.

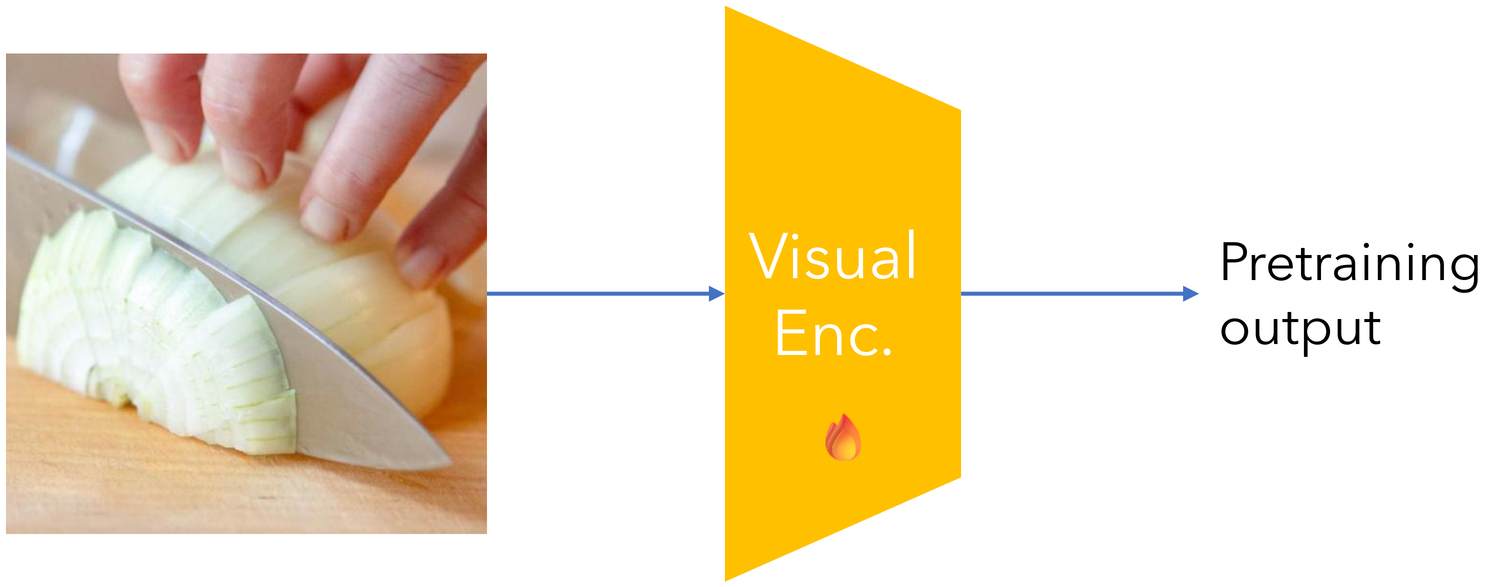
Prior Work: Visual Pretraining

Vision-only Pretraining
Image Classification Masked Image Modeling/Reconstruction Contrastive Learning (via img aug) Frame Temporal Ordering Future Frame Prediction (Radosavovic et al., 2023; Zhao et al., 2022; Gupta et al., 2022b; Seo et al., 2023; Laskin et al., 2020; He et al., 2020; Jing et al., 2023; Zhao et al., 2022; Yuan et al., 2022; Wang et al., 2022)

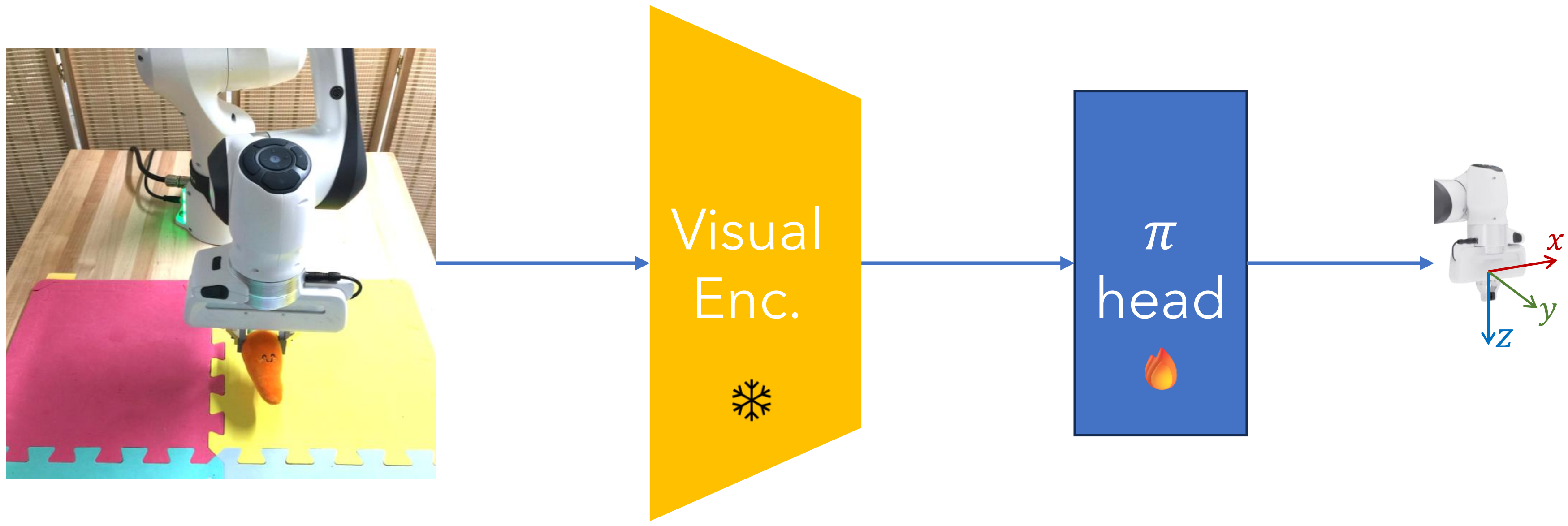
These visual representations are not robust to large domain shifts.

Setup + Approach

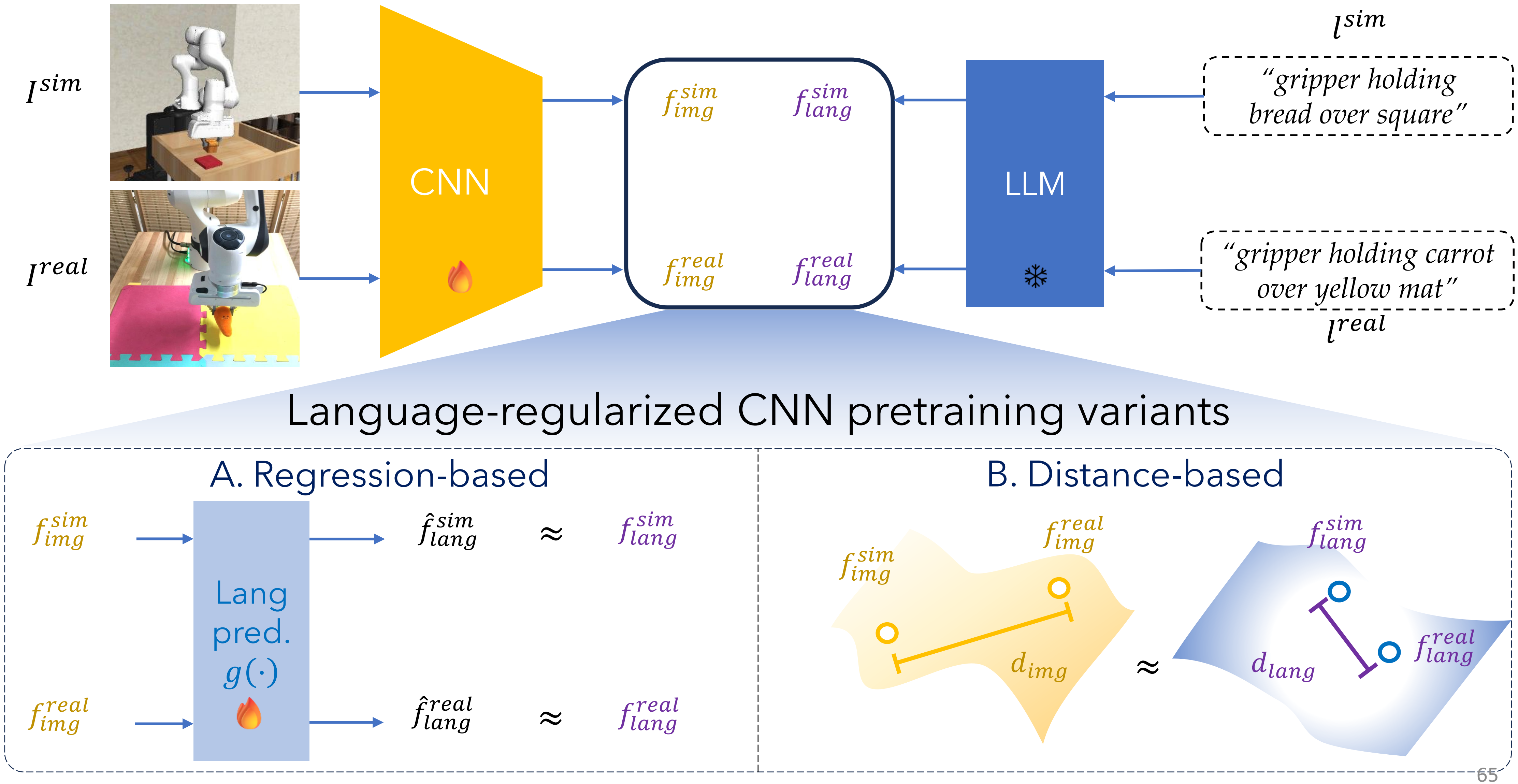
① Pretrain Visual Backbone



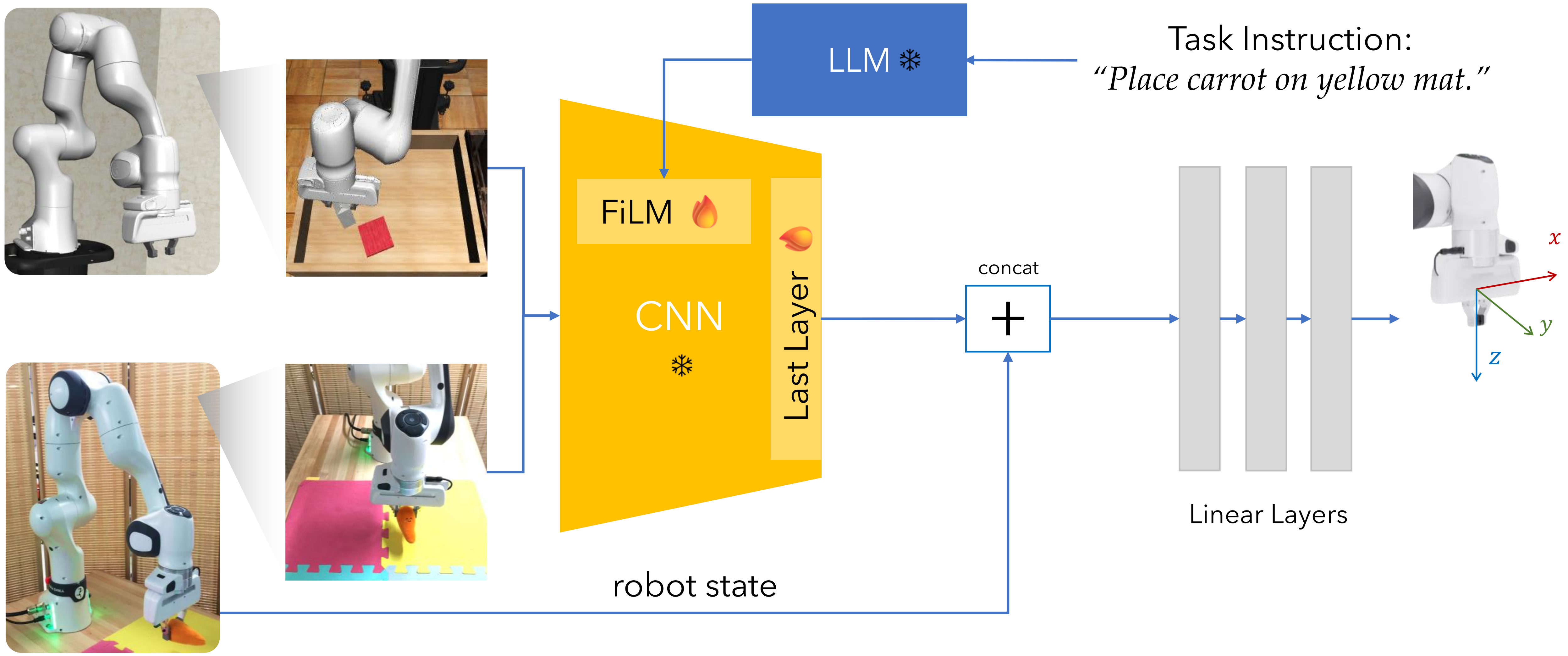
② Finetune Policy Head



① Image-Language Pretraining

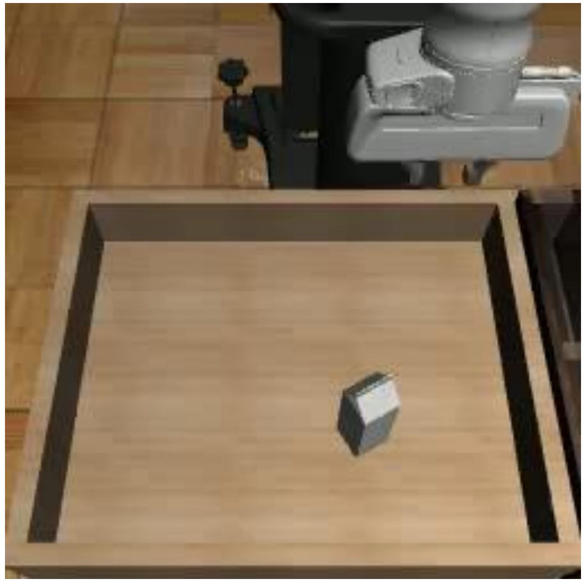


② Multitask, Multidomain Imitation Learning

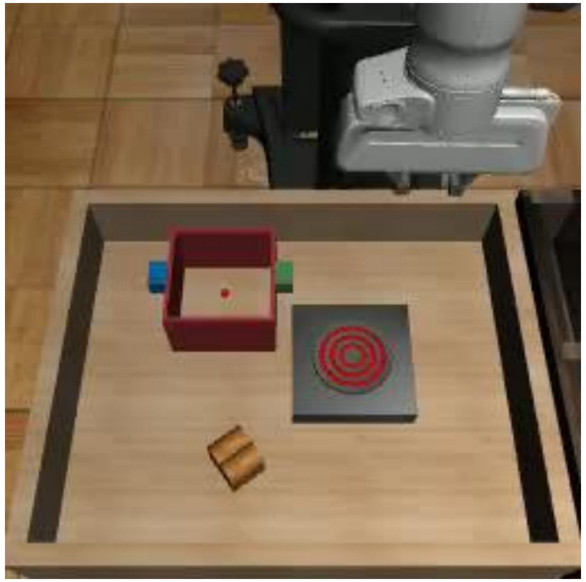


Experiments

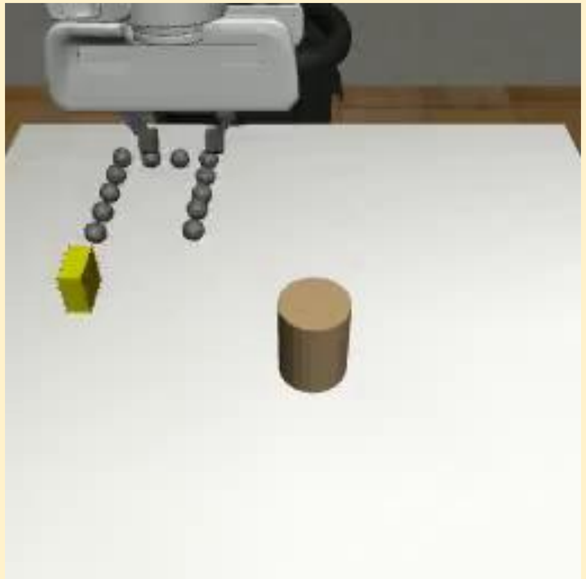
Stack Object

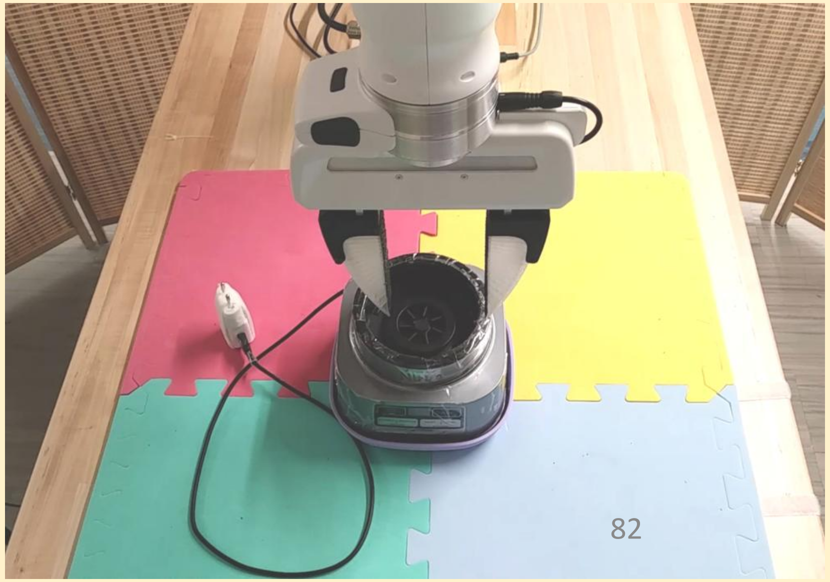


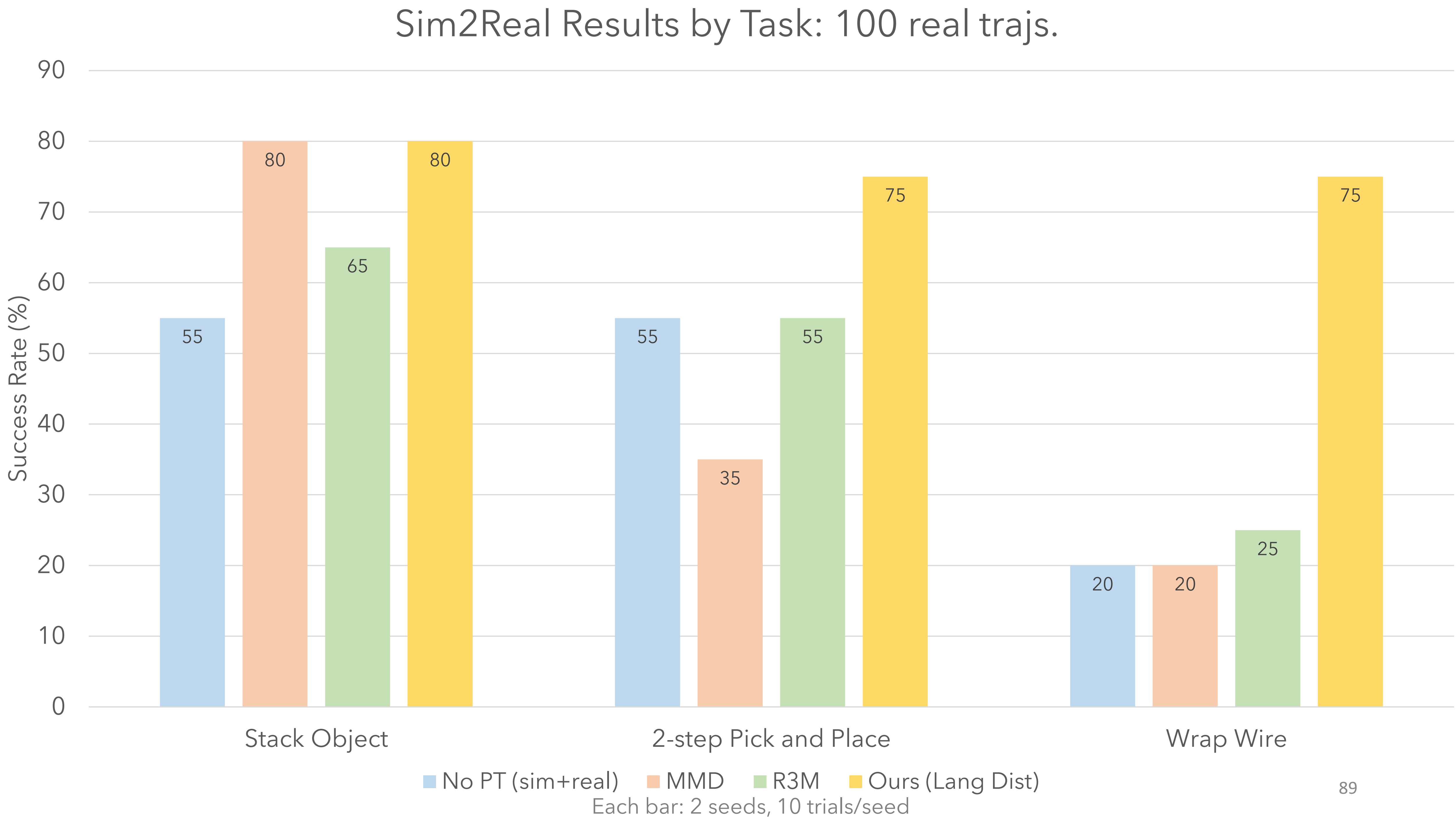
2-step Pick and Place



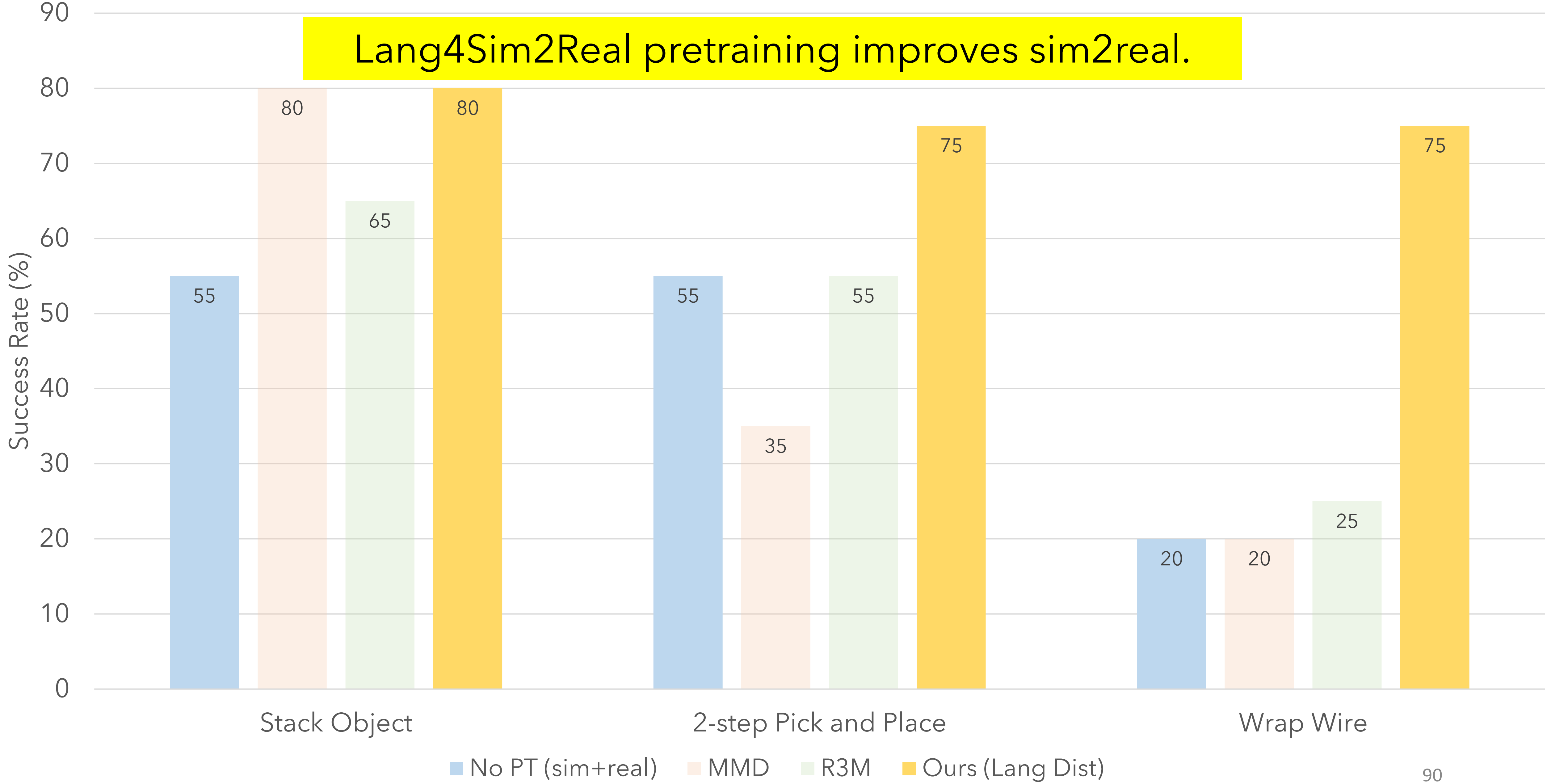
Wrap Wire



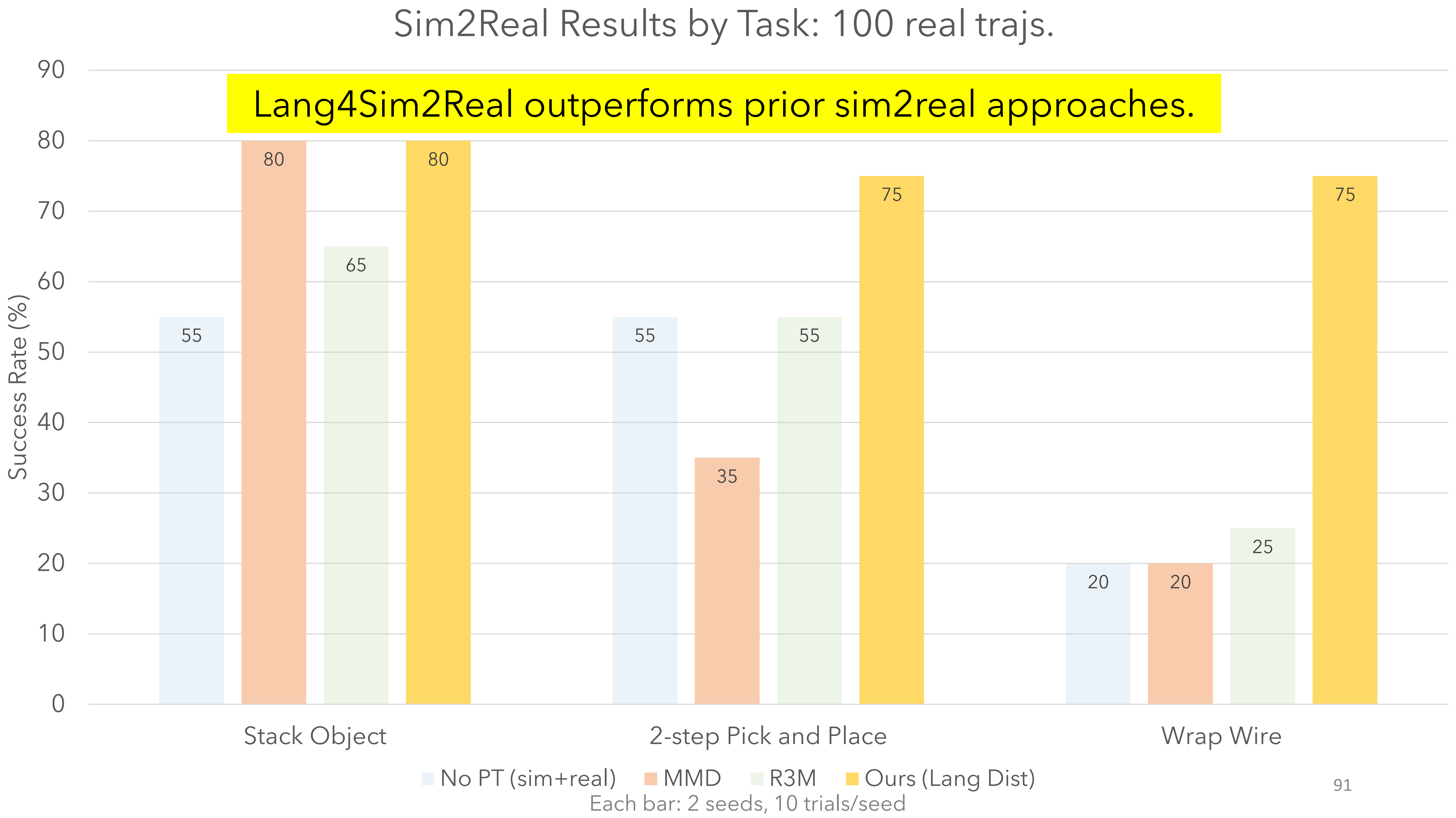



Sim2Real Results by Task: 100 real trajs.
Success Rate (%)
90
80
70
60
50
40
30
20
10
0
55
80
65
80
55
35
55
75
20
20
25
75
Stack Object
2-step Pick and Place
Wrap Wire
No PT (sim+real)
MMD
R3M
Ours (Lang Dist)
Each bar: 2 seeds, 10 trials/seed

Sim2Real Results by Task: 100 real trajs.

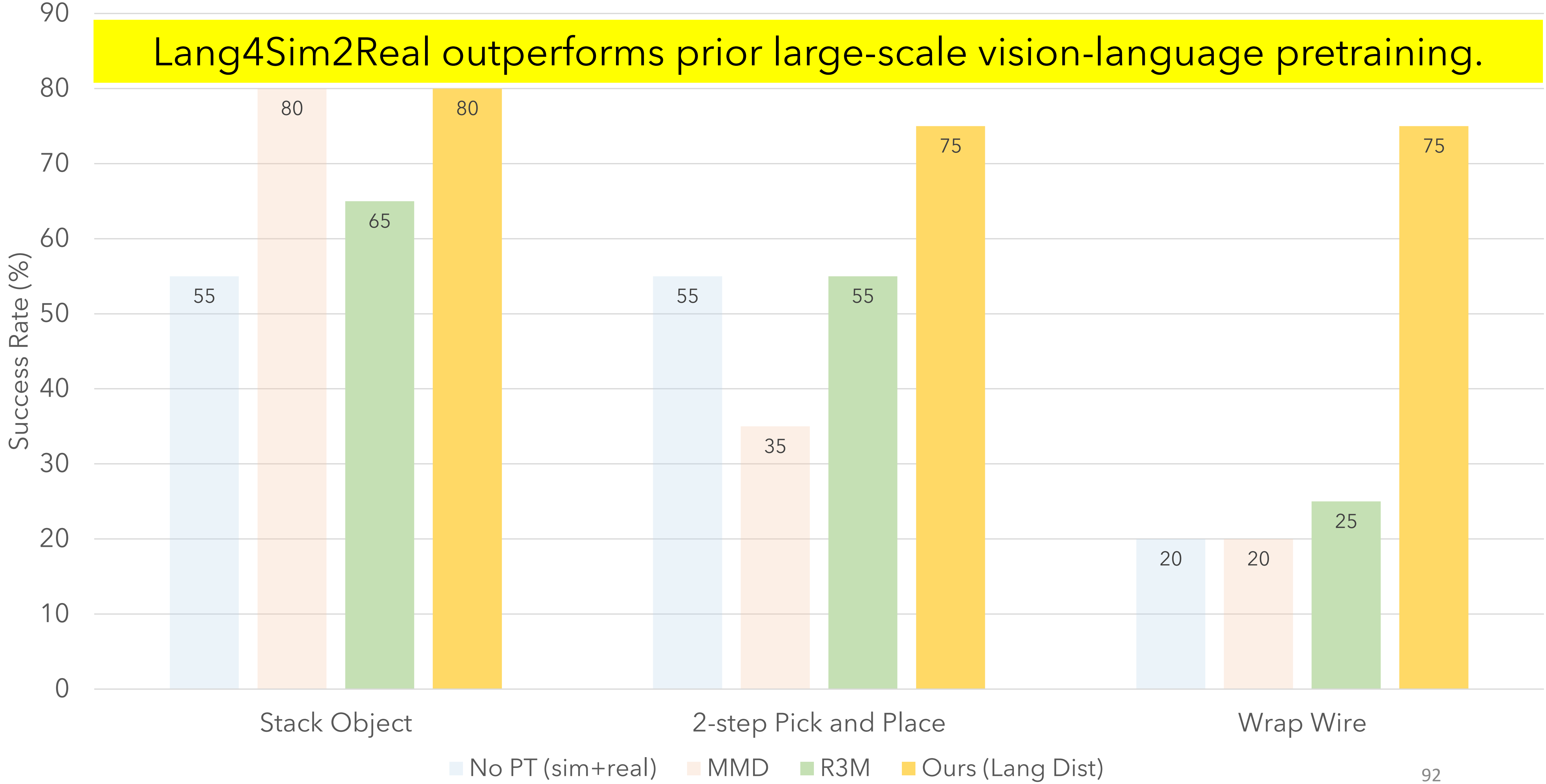


Each bar: 2 seeds, 10 trials/seed

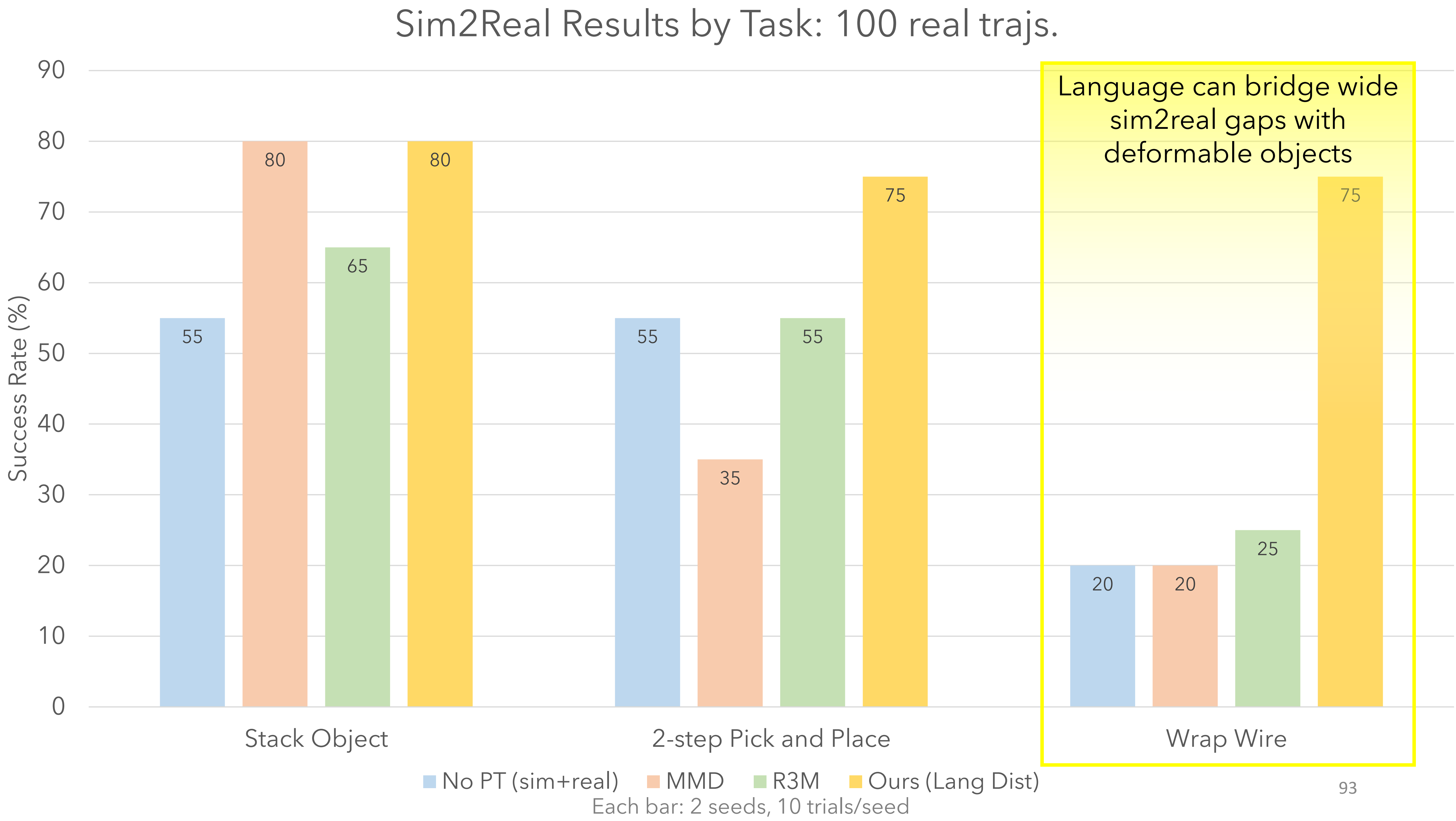

Sim2Real Results by Task: 100 real trajs.
Lang4Sim2Real outperforms prior sim2real approaches.
Success Rate (%)
90
80
70
60
50
40
30
20
10
0
80
80
65
55
55
35
55
75
20
20
25
75
Stack Object
2-step Pick and Place
Wrap Wire
No PT (sim+real)
MMD
R3M
Ours (Lang Dist)
Each bar: 2 seeds, 10 trials/seed

Sim2Real Results by Task: 100 real trajs.

Lang4Sim2Real outperforms prior large-scale vision-language pretraining.



Each bar: 2 seeds, 10 trials/seed

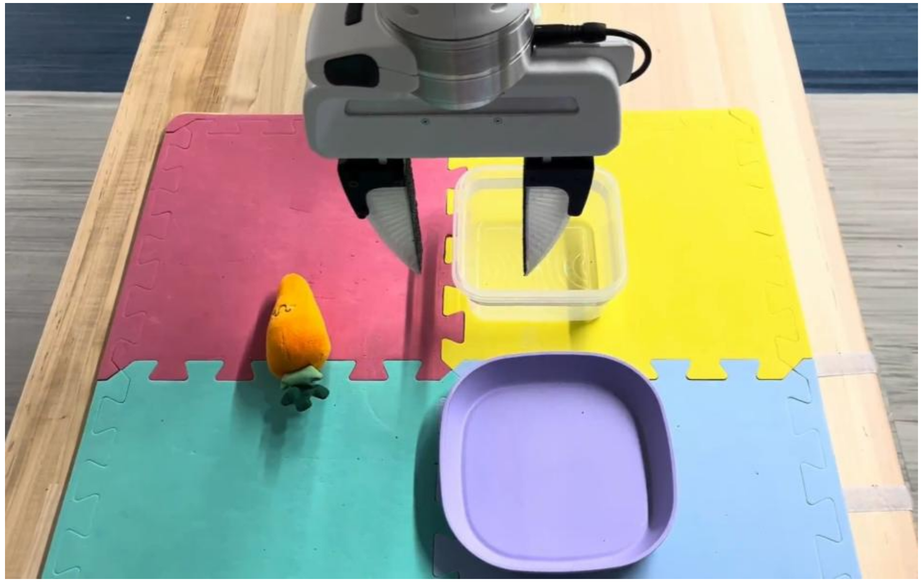


Sim2Real Results by Task: 100 real trajs.
Language can bridge wide sim2real gaps with deformable objects
Success Rate (%)
0
10
20
30
40
50
60
70
80
90
55
80
65
80
55
35
55
75
20
20
25
75
Stack Object
2-step Pick and Place
Wrap Wire
No PT (sim+real)
MMD
R3M
Ours (Lang Dist)
Each bar: 2 seeds, 10 trials/seed

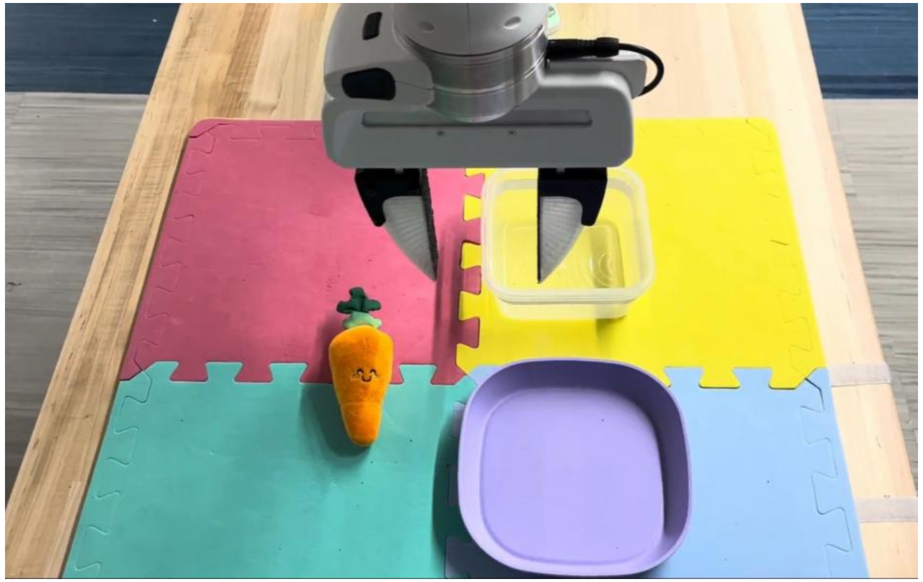
2-step Pick and Place

Our π rollouts (5×)

Successes



Failures

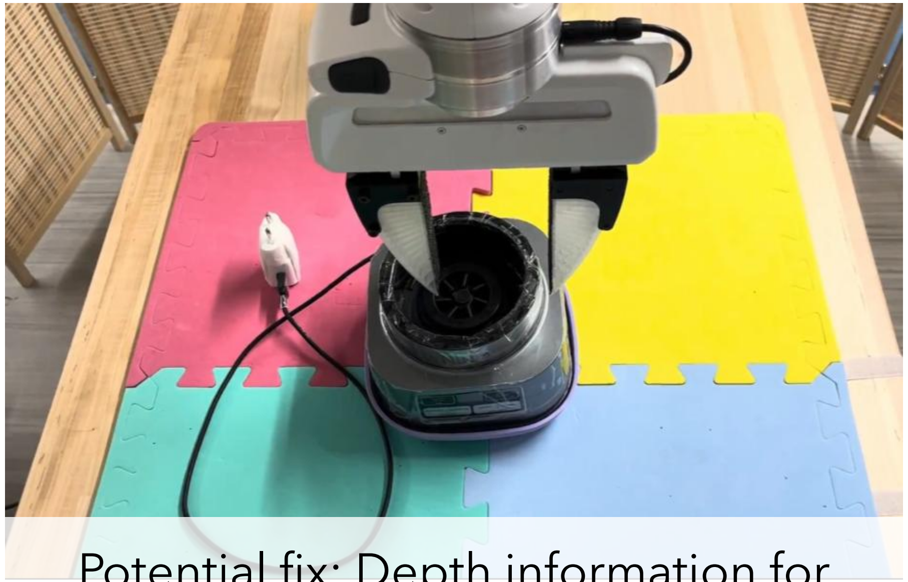


Wrap Wire

Our π rollouts (5×)

Successes

Failures



Potential fix: Depth information for more accurate grasping

Key Takeaways

Language can bridge the sim2real gap with domain-invariant representations.

Our method enables leveraging low-fidelity sim data for sim2real transfer **on deformable objects**.

Outline

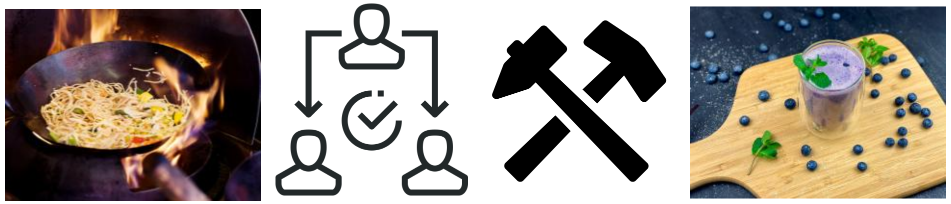
1. Language as Semantics for Generalization
   a) DeL-TaCo: Demo-Language Task Conditioning
   b) Lang4Sim2Real: Language to Bridge the Sim2Real Gap
2. Language as a Communicative Medium
   a) MicoBot: Bidirectional Dialog for Human-Robot Collaboration
3. Future Directions

Mixed-Initiative Dialog for Human-Robot Collaborative Manipulation

Collaborators: Ray Mooney, Roberto Martín-Martín, Chengshu Li, Luca Macesanu, Arnav Balaji, Ruchira Ray

Under Review, ICRA 2026

Imagine preparing for a dinner party with a friend…



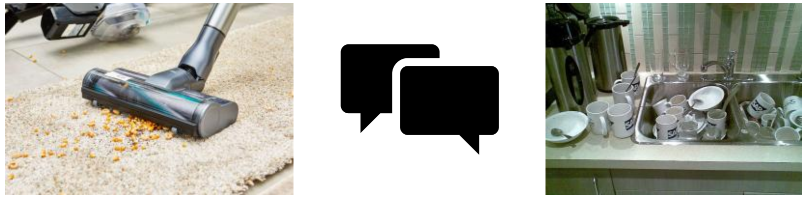
Allocate tasks by skillset

I can cook, but don't know how to make drinks. Can you?

Sure, I make the best smoothies.

Can you help wash dishes?

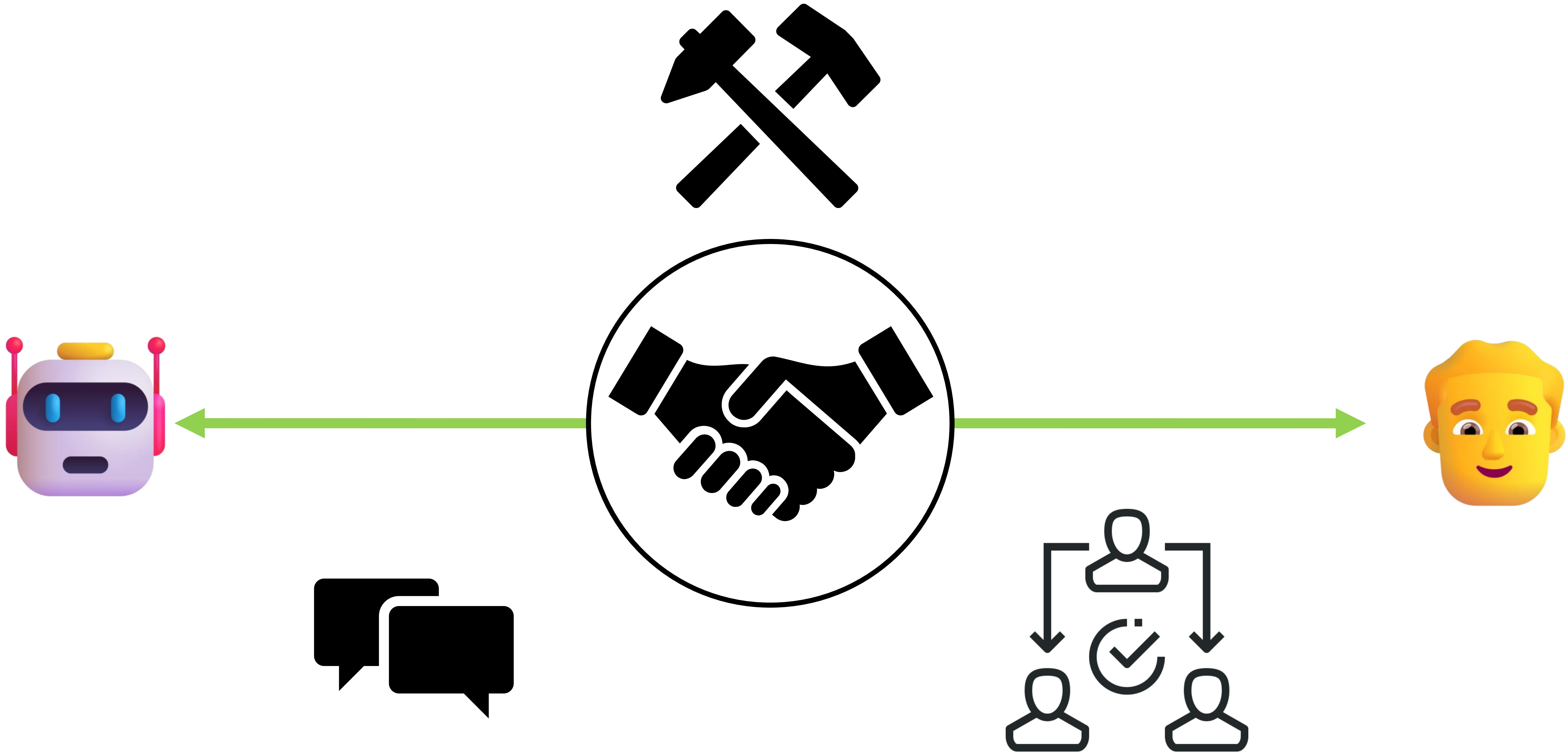
No, I'm busy vacuuming the floor.



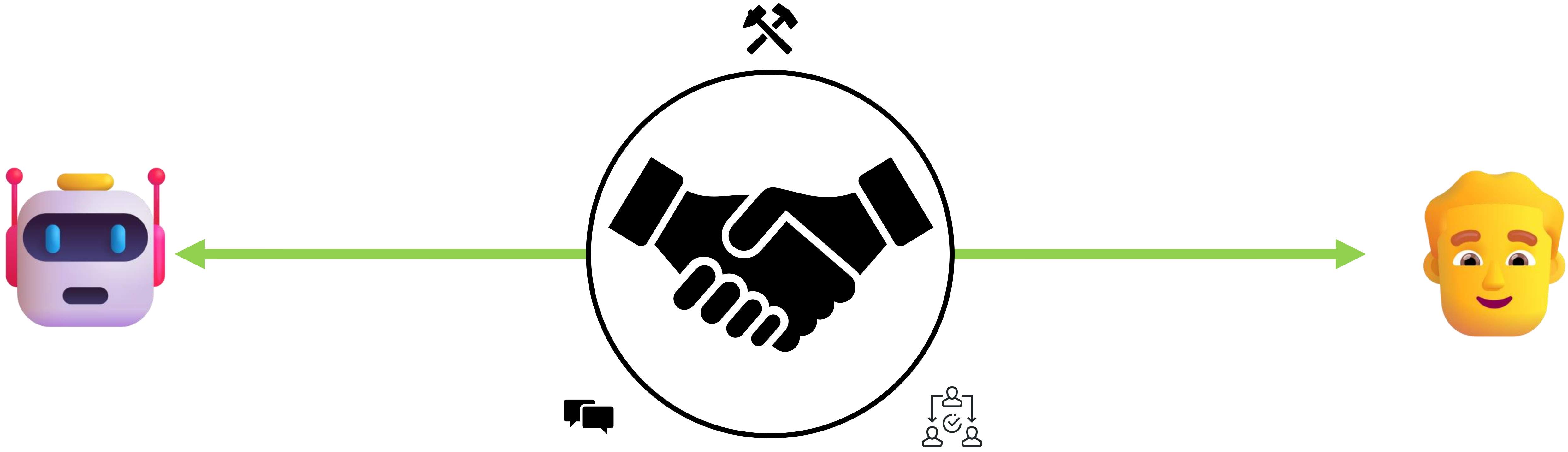
Adapt to availability and willingness

<

Robots struggle on complex tasks…



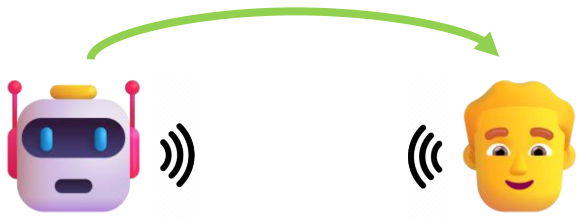
...and need to collaborate and communicate with humans.



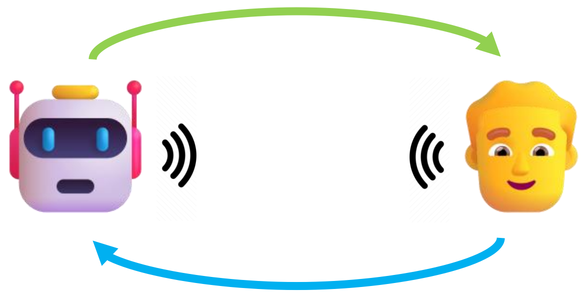
Hypothesis:
A collaborative robot must leverage 🤖 + 👨 physical capabilities and adapt to changing 👨 preferences and helpfulness through **bidirectional communication**.

Prior Work

Human-Initiated Dialog
Most Modern Chatbots



Robot-Initiated Dialog
In Human-Robot Interaction (HRI)
MOSAIC (Wang+, CoRL 2024)
RoCo (Mandi+, ICRA 2024)
KnowNo (Ren+, CoRL 2023)



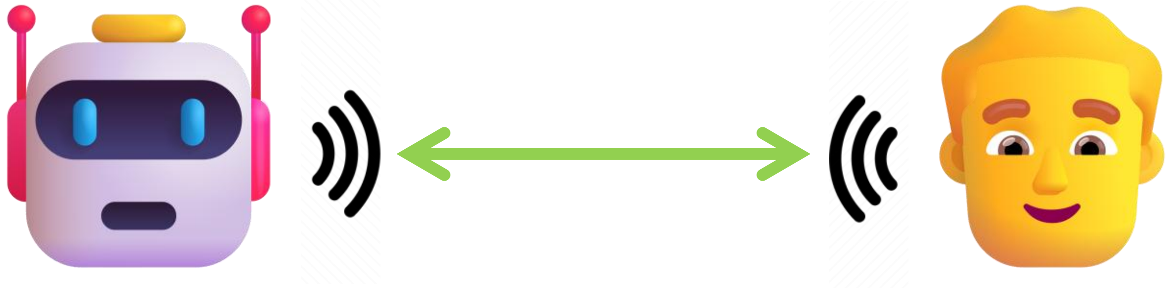
Mixed-Initiative Dialog
Jaime R. Carbonell, 1970
IEEE Man-Machine Systems
(Deng+, 2023)
(Chen+, ACL 2023)

Prior works in Mixed-Initiative Dialog do not study grounded language for physical tasks.
Prior HRI + dialog works are single initiative, limiting the flexibility and adaptability of collaboration.

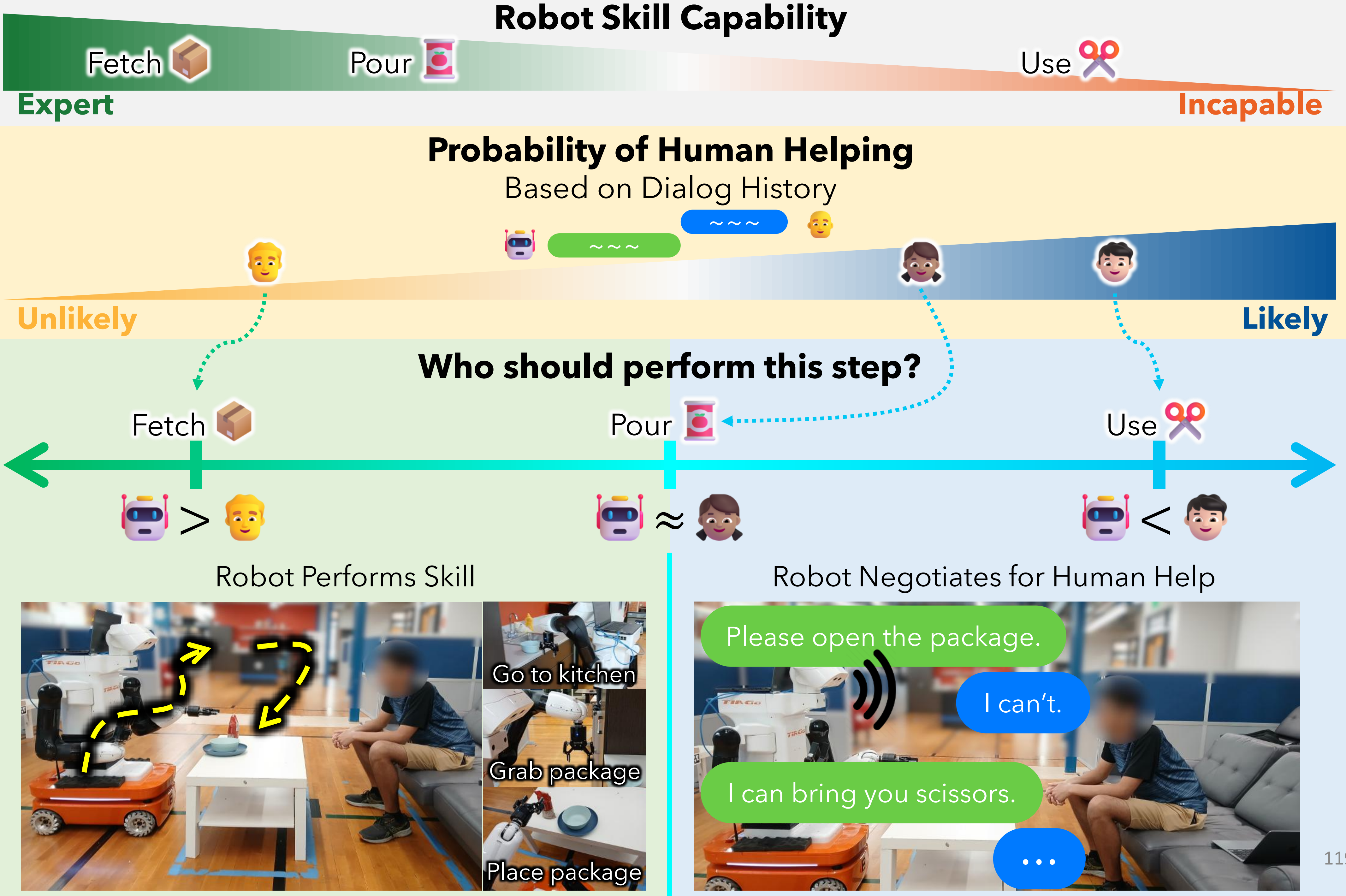
MOSAIC: Huaxiaoyue Wang, Kushal Kedia, Juntao Ren, Rahma Abdullah, Atiksh Bhardwaj, Angela Chao, Kelly Y Chen, Nathaniel Chin, Prithwish Dan, Xinyi Fan, Gonzalo Gonzalez-Pumariega, Aditya Kompella, Maximus Adrian Pace, Yash Sharma, Xiangwan Sun, Neha Sunkara, Sanjiban Choudhury. "MOSAIC: Modular Foundation Models for Assistive and Interactive Cooking." CoRL 2024.
RoCo: Zhao Mandi, Shreeya Jain, Shuran Song. "RoCo: Dialectic Multi-Robot Collaboration with LLMs." ICRA 2024.
KnowNo: Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, Anirudha Majumdar. "Robots That Ask For Help: Uncertainty Alignment for Large Language Model Planners." CoRL 2023.
Deng: Yang Deng, Wenxuan Zhang, Yifei Yuan, Wai Lam. "Knowledge-enhanced Mixed-initiative Dialogue System for Emotional Support Conversations." ACL 2023.
Chen: Maximillian Chen, Xiao Yu, Weiyan Shi, Urvi Awasthi, Zhou Yu. "Controllable Mixed-Initiative Dialogue Generation through Prompting." ACL 2023.

MICoBot

Mixed **I**nitiative **Co**llaborative Ro**bot**



Real-world human + robot manipulation
with mixed-initiative dialog



Robot Skill Capability
Fetch
Pour
Use
Expert
Incapable
Probability of Human Helping
Based on Dialog History
~~~
~~~
Unlikely
Likely
Who should perform this step?
Fetch
Pour
Use
>
≈
<
Robot Performs Skill
Go to kitchen
Grab package
Place package
Robot Negotiates for Human Help
Please open the package.
I can't.
I can bring you scissors.
...

Metric: Who does what?

We want $f(a)$ such that:

$$f_{\text{🤖}}(a) > f_{\text{👱}}(a) \rightarrow \text{🤖 Performs action } a$$

$$f_{\text{🤖}}(a) < f_{\text{👱}}(a) \rightarrow \text{👱 Performs action } a$$

If $r_t = -1, \forall t$:

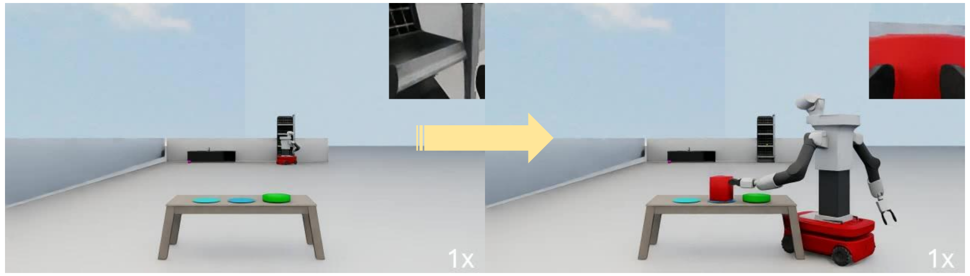
$$Q(s,a) = -\mathbb{E}[\text{\# timesteps to success on } a \text{ from } s]$$

Acquiring Q-functions

$Q(s,a) = -\mathbb{E}[\#\text{ timesteps to success on } a \text{ from } s]$

$Q_R(s,a)$

$Q_H(s,a)$



Train Q_R from OmniGibson data on rough approximation of real env.

Travel time Estimates
+ GPT Time estimates to perform a

Compare: $Q_R(s,a) \gtrless Q_H(s,a)$
Time as Proxy for Effort

Task Allocation

$$Q_R(s,a) \gtreqless Q_H(s,a)$$

Task Allocation

	$Q_R(s,a) \gtrless Q_H(s,a)$
Account for human willingness to help	$Q_R(s,a) \gtrless \frac{1}{p_{help}} Q_H(s,a)$

Task Allocation

	$Q_R(s,a) \gtrless Q_H(s,a)$
Account for human willingness to help	$Q_R(s,a) \gtrless \frac{1}{p_{help}} Q_H(s,a)$
Account for higher value of human effort	$Q_R(s,a) \gtrless \frac{\alpha}{p_{help}} Q_H(s,a)$

where $\alpha > 1$

Adapting to Human Willingness with p_{help}

$$Q_R(s,a) \gtrless \frac{\alpha}{p_{help}} Q_H(s,a)$$

What else does MicoBot need to decide?

High-Level Strategy

Continue doing task?
Initiate Dialog?
Respond to Dialog?

Propose a way to split the task?

Negotiate for human help?

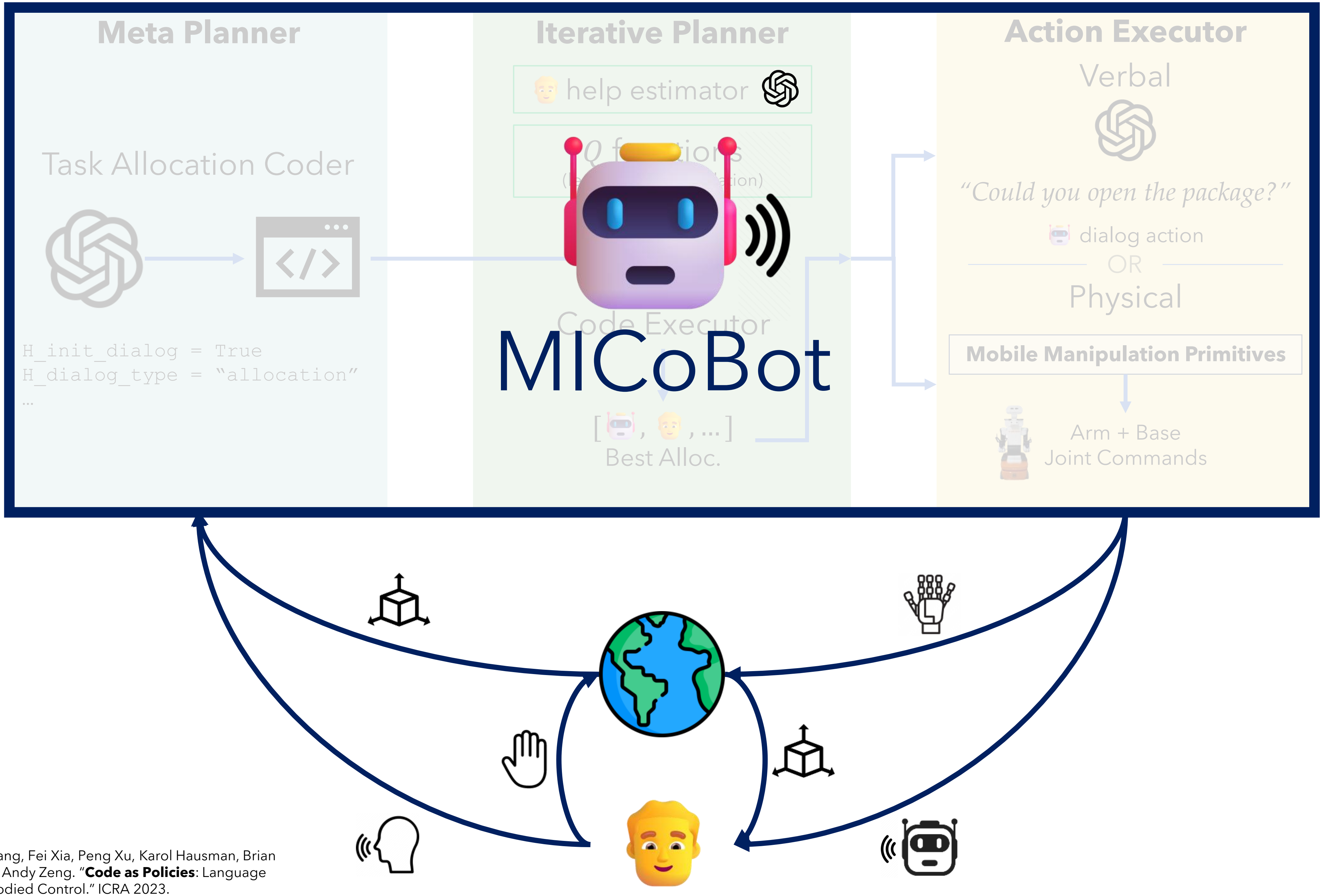
Task Allocation

Who is best suited to perform each step?

Low-level Actions

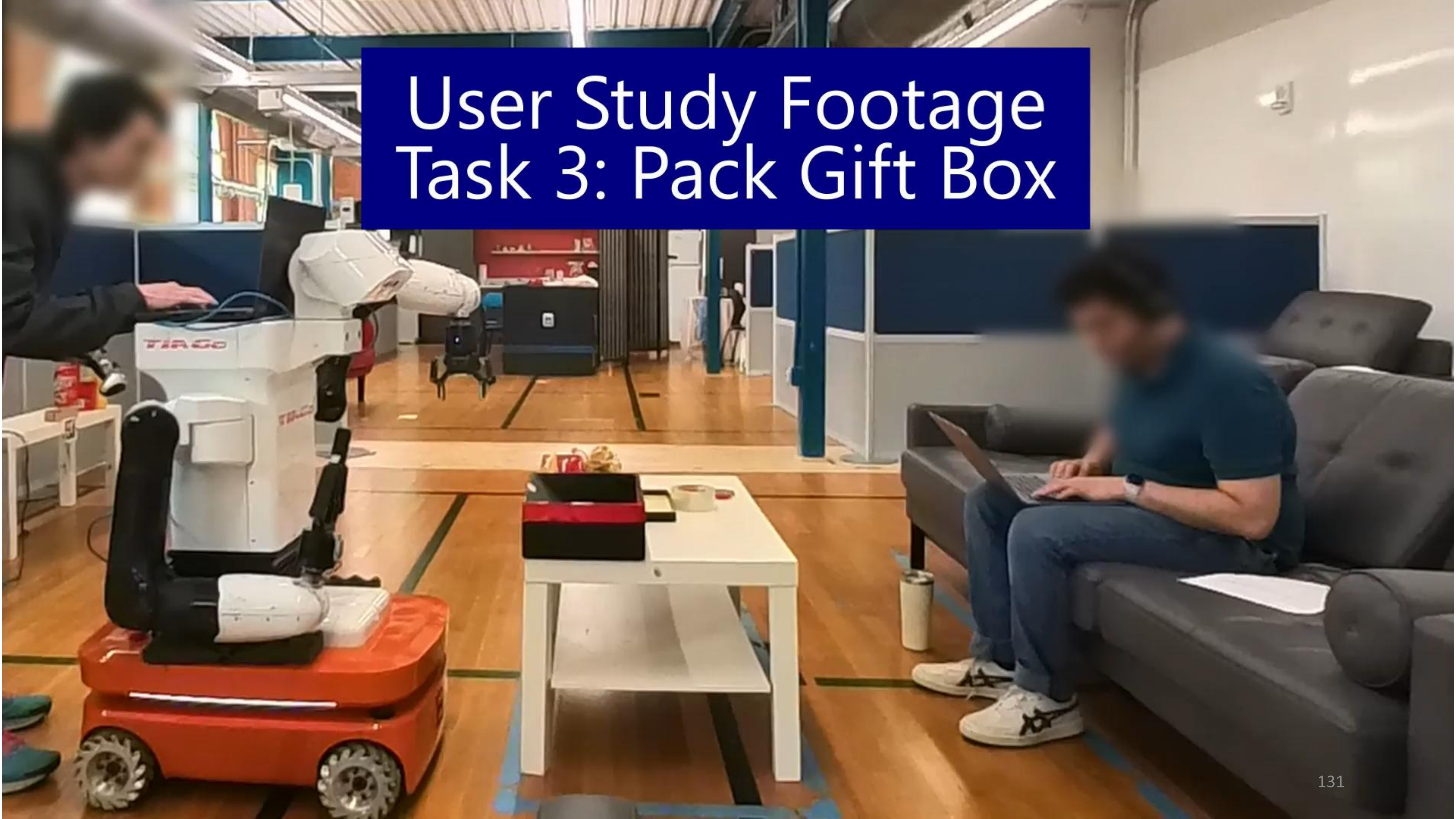
What low-level physical actions to execute?

What words to say to the human?

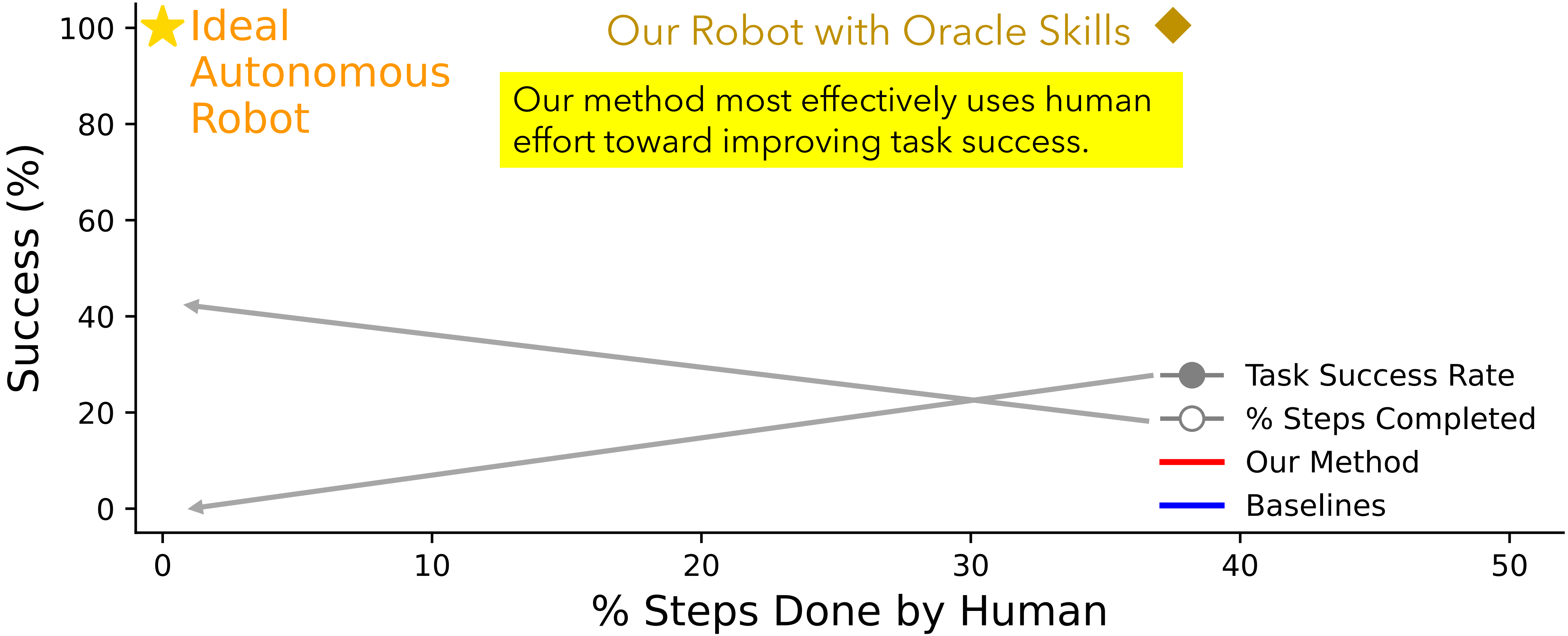


Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. "**Code as Policies**: Language Model Programs for Embodied Control." ICRA 2023.

User Study Footage
Task 3: Pack Gift Box



Real World User Studies ($n = 18$, 3 tasks)



How does MicoBot Adapt to Human Helpfulness?

Human Helpfulness Estimate Over Time

Dialog over Time

🧑 : Please cut the package and pour it into the bowl.

: Please cut the package and pour it into the bowl.

: [*Brings bowl and package to coffee table.*]

: Good job.

: …please open the package with scissors on the coffee table.

: There are no scissors on the coffee table.

: …I can bring scissors so you can open the package…

: …You should bring the scissors AND open the package…

MicoBot can negotiate, adapt to, and convince reluctant human collaborators.

: …Ok, if you bring the scissors, I'll open the package…

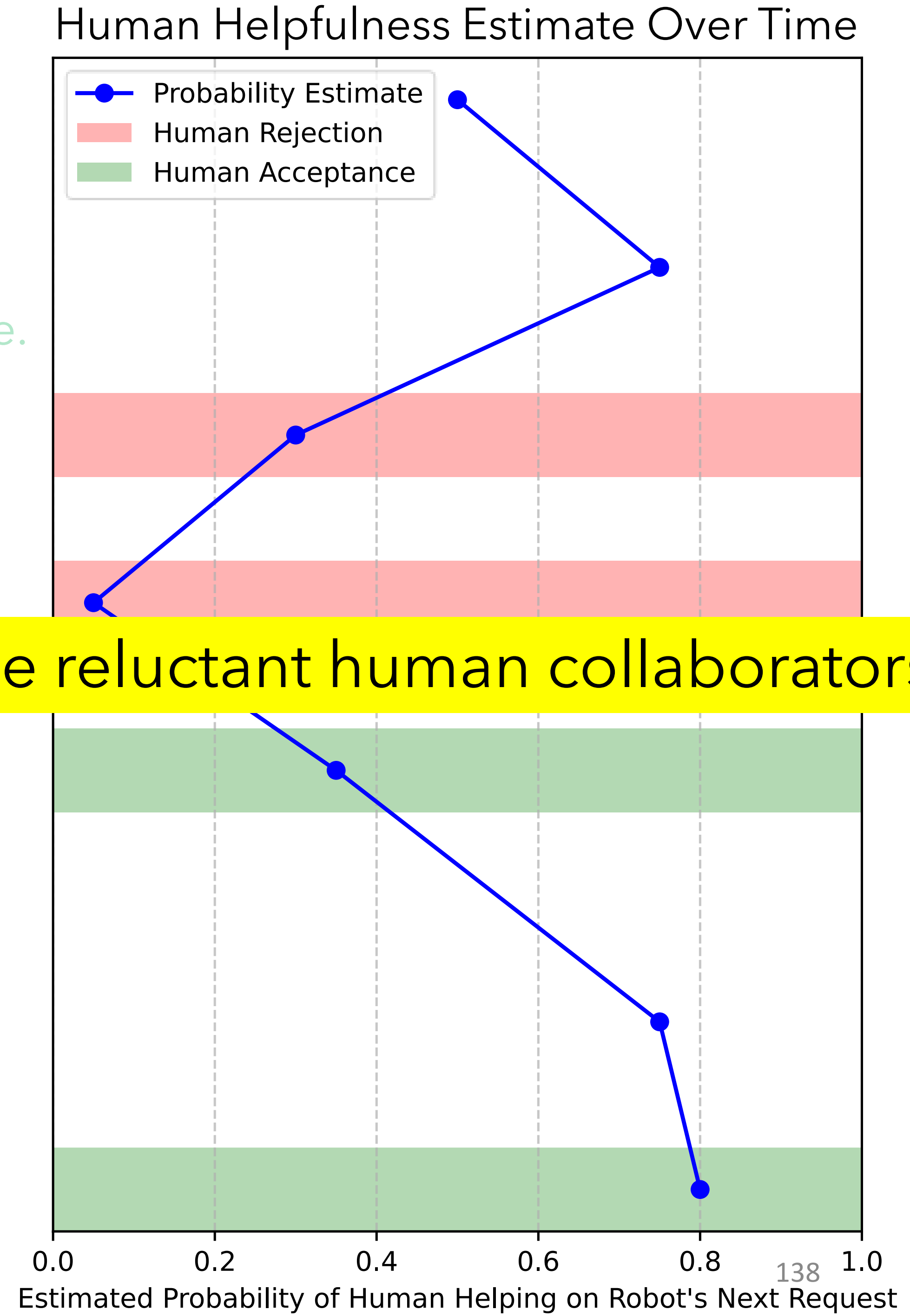
: …I'll bring the scissors then!

: [*Brings the scissors to the coffee table.*]

: Thanks for bringing the scissors.

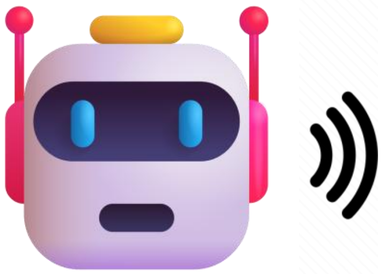
: …please open the package with scissors…

: Ok, I will do that now.



Key Takeaways

1. MicoBot is the first work to enable mixed-initiative dialog for real-world HRI.
2. Mixed-initiative dialog improves collaborative manipulation compared to single-initiative dialog.
3. MicoBot adapts to human collaborators of varying helpfulness and communicative styles.



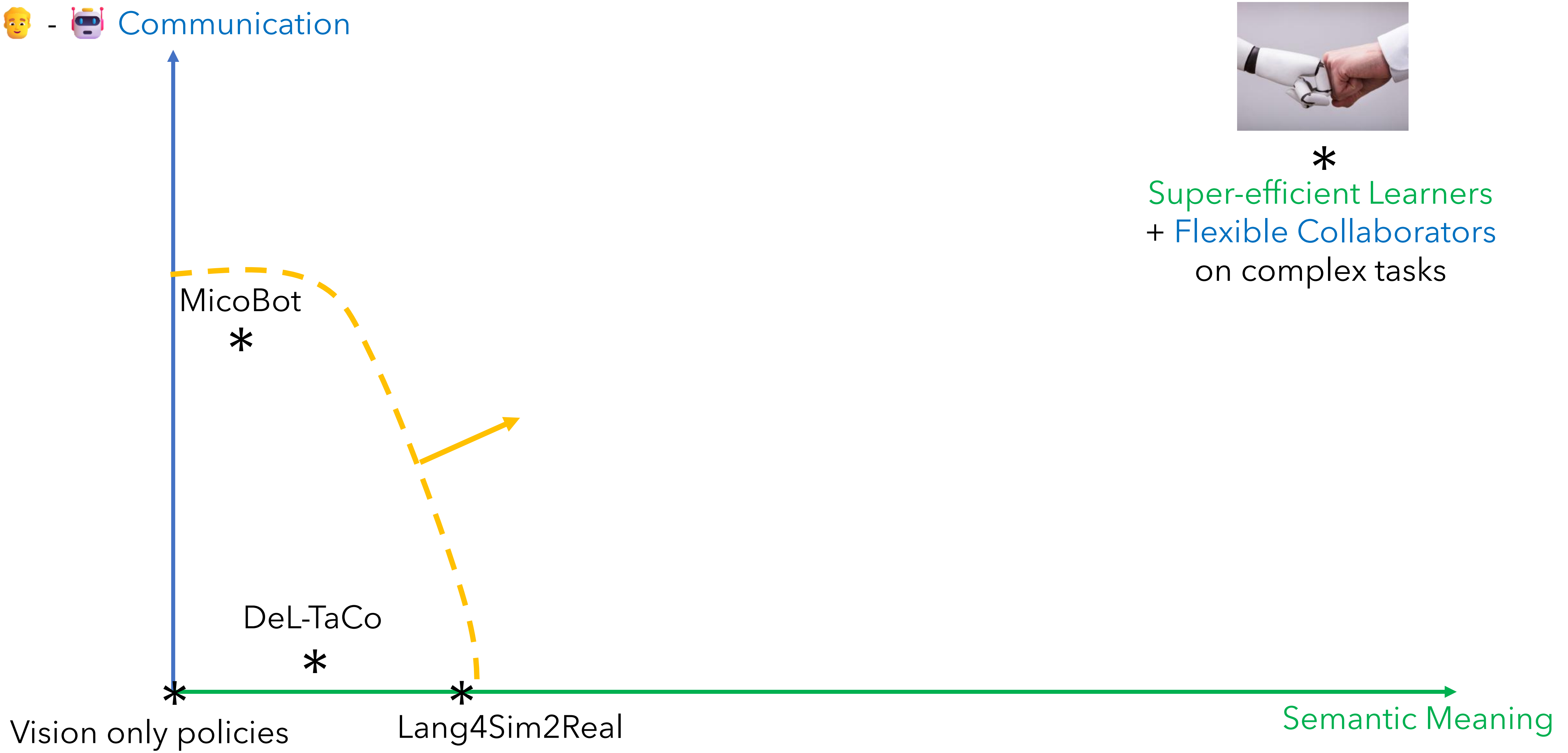
MICoBot

Mixed-initiative dialog
for real-world human+robot manipulation

Outline

1. Language as Semantics for Generalization
    a) DeL-TaCo: Demo-Language Task Conditioning
    b) Lang4Sim2Real: Language to Bridge the Sim2Real Gap
2. Language as a Communicative Medium
    a) MicoBot: Bidirectional Dialog for Human-Robot Collaboration
3. Future Directions

Language for Home-Worthy Real Robots



So far, language has functioned to communicate…

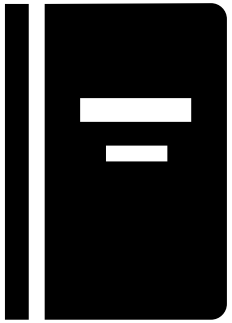
Task Allocations (MicoBot)

Task Commands (Del-Taco)

Scene Descriptions (Lang4Sim2Real)

Untapped:
- How to do a task?
- How to behave in general?
- Fine-grained movements

How do we quickly learn to perform specialized tasks?

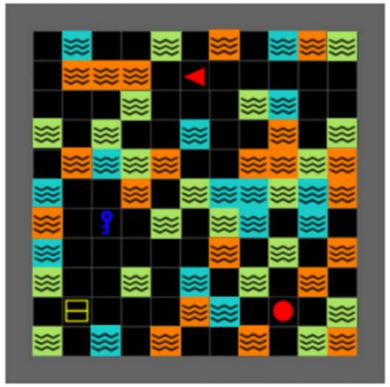


Reading Guidebooks + Experience

Hypothesis: Robots can learn faster and more safely through guidebooks + experience.

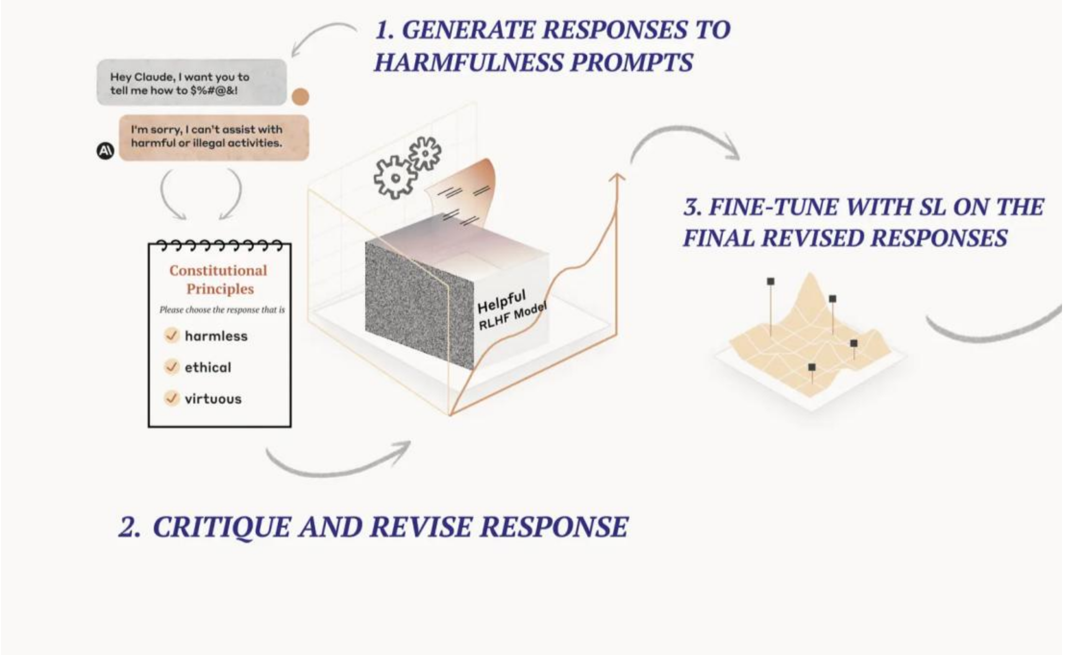
Prior Work

RL with Language Constraints
(Yang+ NeurIPS 2020, Lou+ AAMAS 2024)



"You can walk on red and green but not blue."

Constitutional AI
(Anthropic 2022)

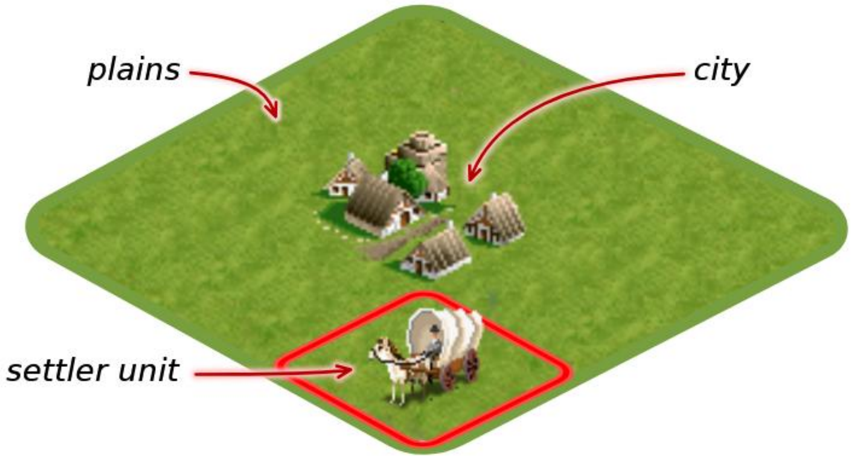


These works focus on safety and do not demonstrate guidebooks can accelerate learning for robotics.

Yang: Tsung-Yen Yang, Michael Y Hu, Yinlam Chow, Peter J. Ramadge, Karthik Narasimhan. "Safe Reinforcement Learning with Natural Language Constraints." NeurIPS 2020.
Lou: Xingzhou Lou, Junge Zhang, Ziyan Wang, Kaiqi Huang, Yali Du. "Safe Reinforcement Learning with Free-form Natural Language Constraints and Pre-Trained Language Models." AAMAS 2024.
Anthropic: Yuntao Bai, …, Jared Kaplan. "Constitutional AI: Harmlessness from AI Feedback." Arxiv 2022.

Prior Work: Learning from Manuals

Learning to Play Civilization II from Manuals
(Branavan+, JAIR 2012)



Relevant text: *"Use settlers to irrigate land near your city"*

Predicted action words: *"irrigate", "settler"*

Predicted state words: *"land", "near", "city"*

Settlers unit, candidate action 1: ***irrigate***

Features:

action = **irrigate** and action-word = "*irrigate*"
action = **irrigate** and state-word = "*land*"
action = **irrigate** and terrain = plains
action = **irrigate** and unit-type = settler
state-word = "*city*" and near-city = true

Settlers unit, candidate action 2: ***build-city***

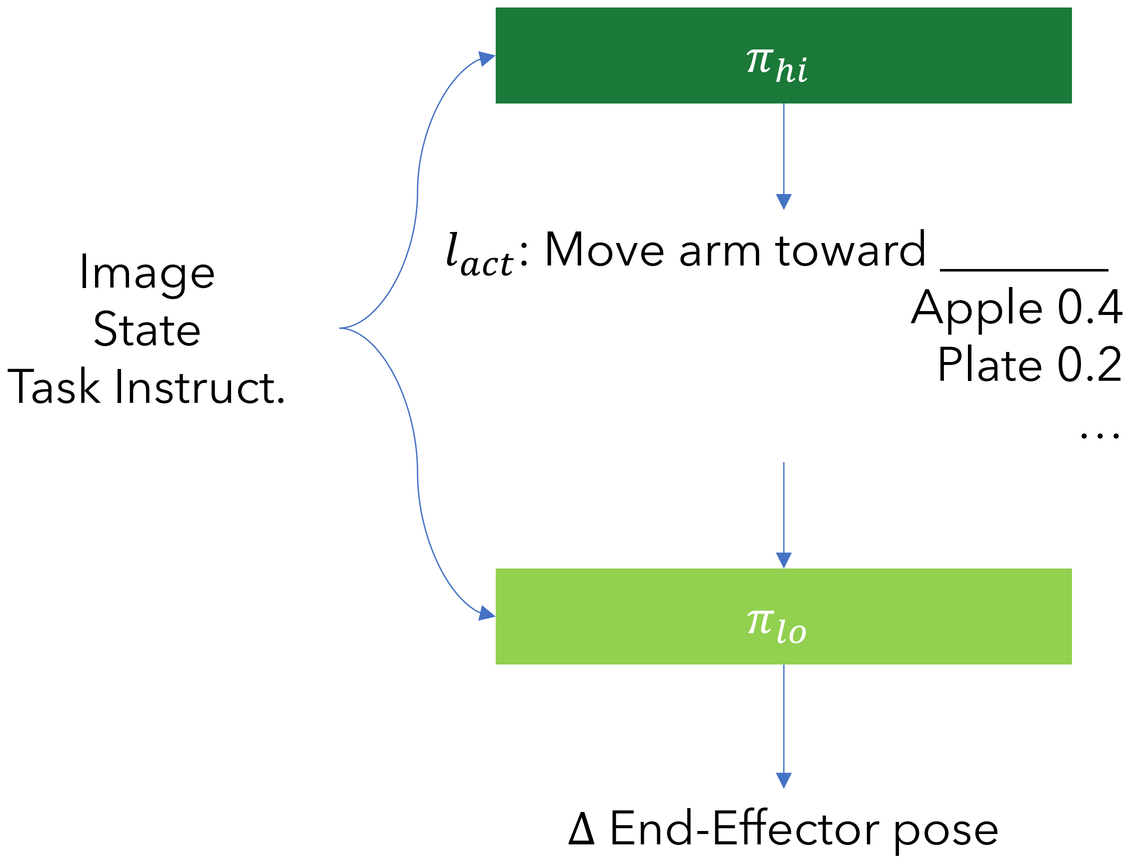
Features:

action = **build-city** and action-word = "*irrigate*"
action = **build-city** and state-word = "*land*"
action = **build-city** and terrain = plains
action = **build-city** and unit-type = settler
state-word = "*city*" and near-city = true

Demonstrates manuals can improve win rate in Monte-Carlo search, but does not apply this to robot learning.

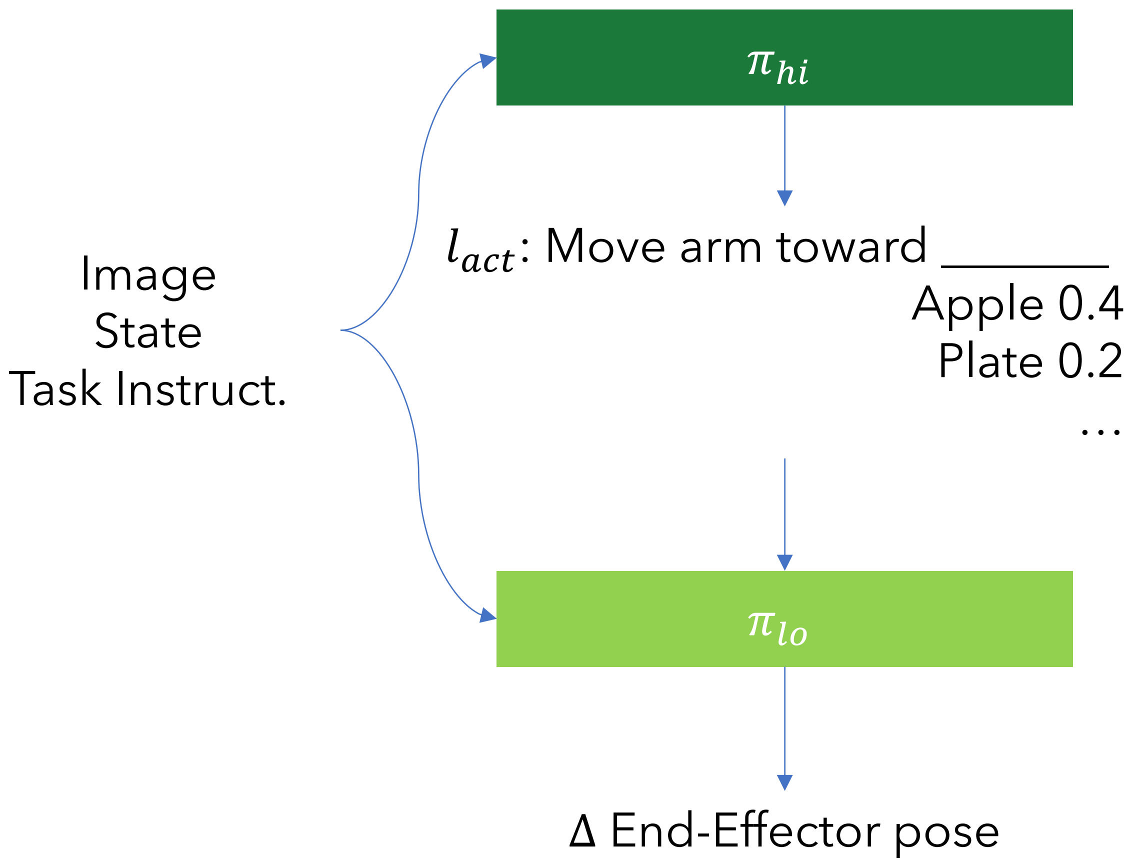
SRK Branavan, David Silver, Regina Barzilay. "Learning to Win by Reading Manuals in a Monte-Carlo Framework." JAIR 2012.

Proposed Architecture



Hierarchy similar to: RT-H (Karamcheti+ RSS 2024), YAY-Robot (Shi+ RSS 2024)

Proposed Architecture



Scarce Trajectory Data

$\boldsymbol{o_t} \rightarrow$						
$a_t \rightarrow$						
$l_{act} \rightarrow$	"Grasp plate"			"Move to counter"		

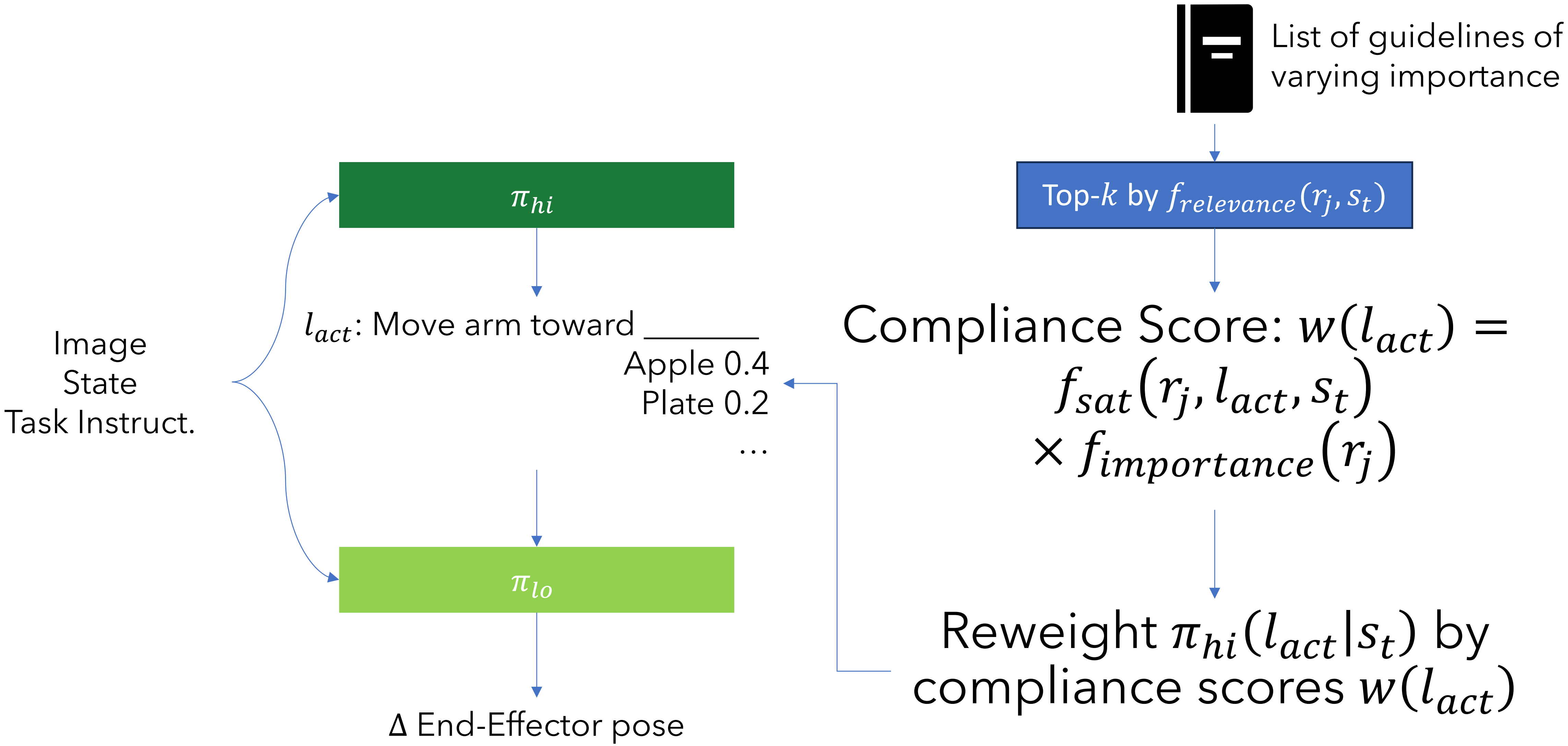
Training

I. π_{lo} BC

II. π_{hi} BC

III. π_{hi} RL w/ rule compliance

Phase III. π_{hi} Guideline Compliance



Experimental Questions

How compliant and performant is π?

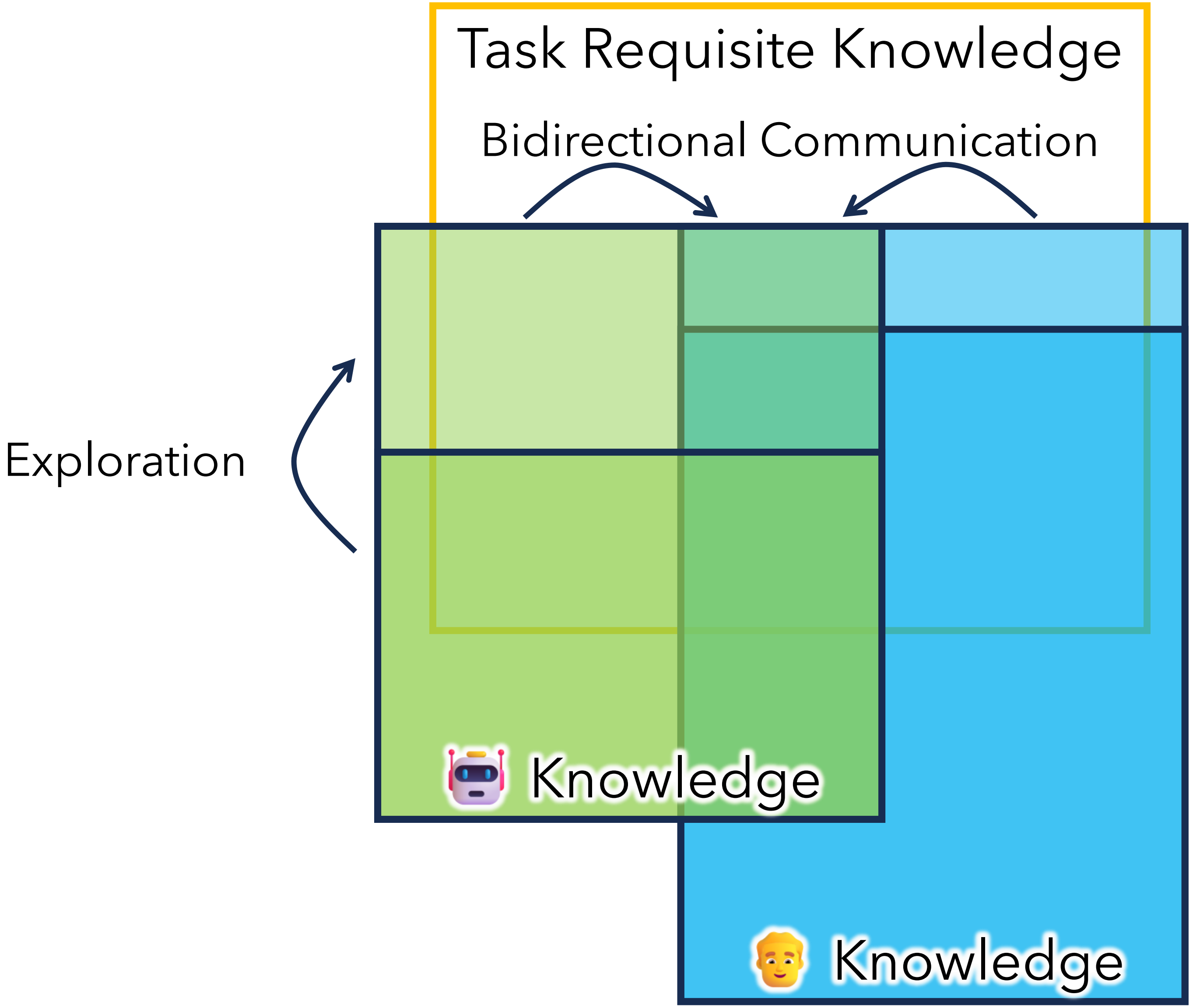
Does guidebook enable faster exploration?

Potential Tasks

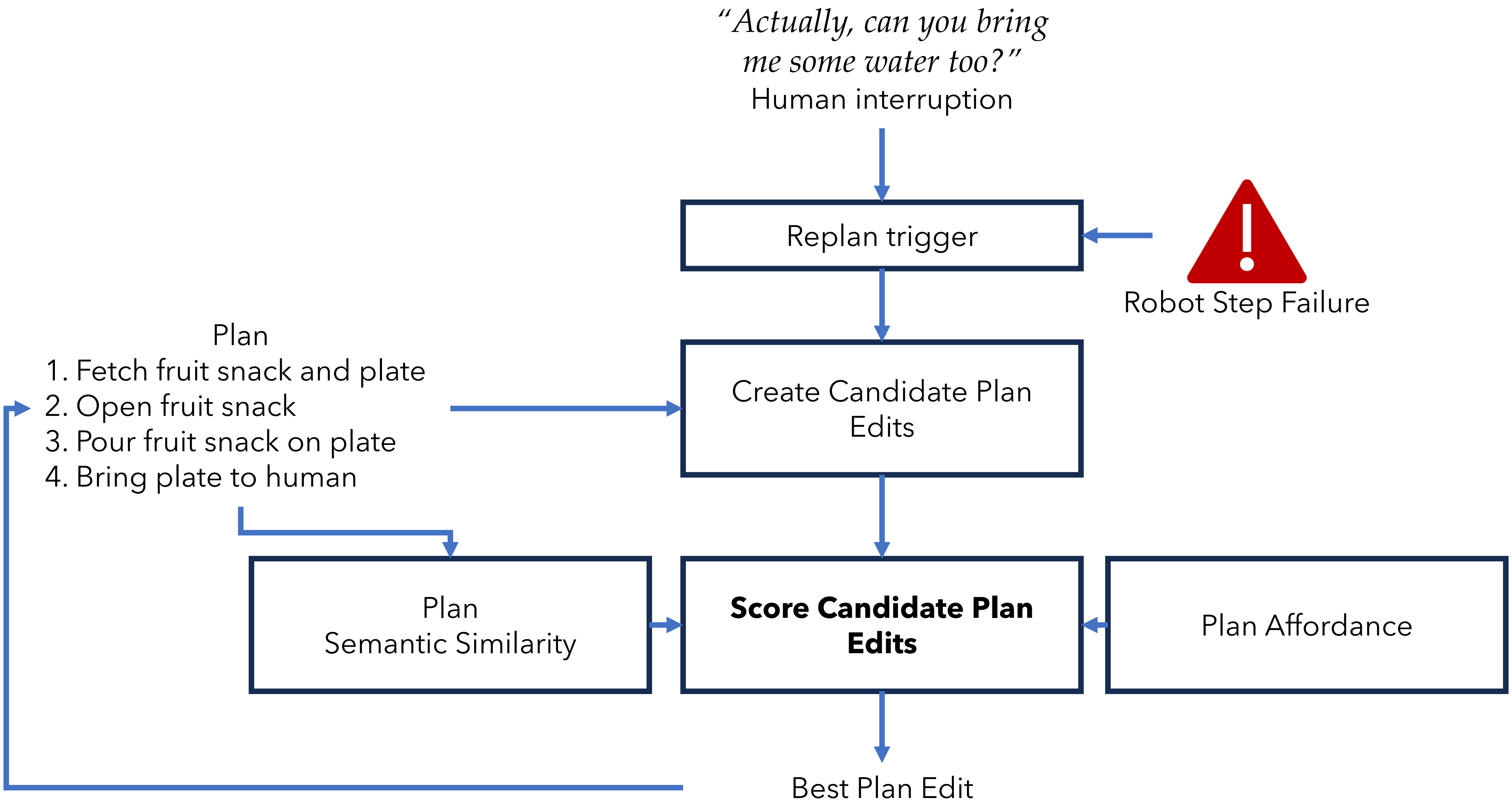
Long-Term Future Directions

Mixed-Initiative Dialog For...	
Task Allocation (MicoBot)	Information Exchange
Replanning	Low-Level Feedback

Information Exchange

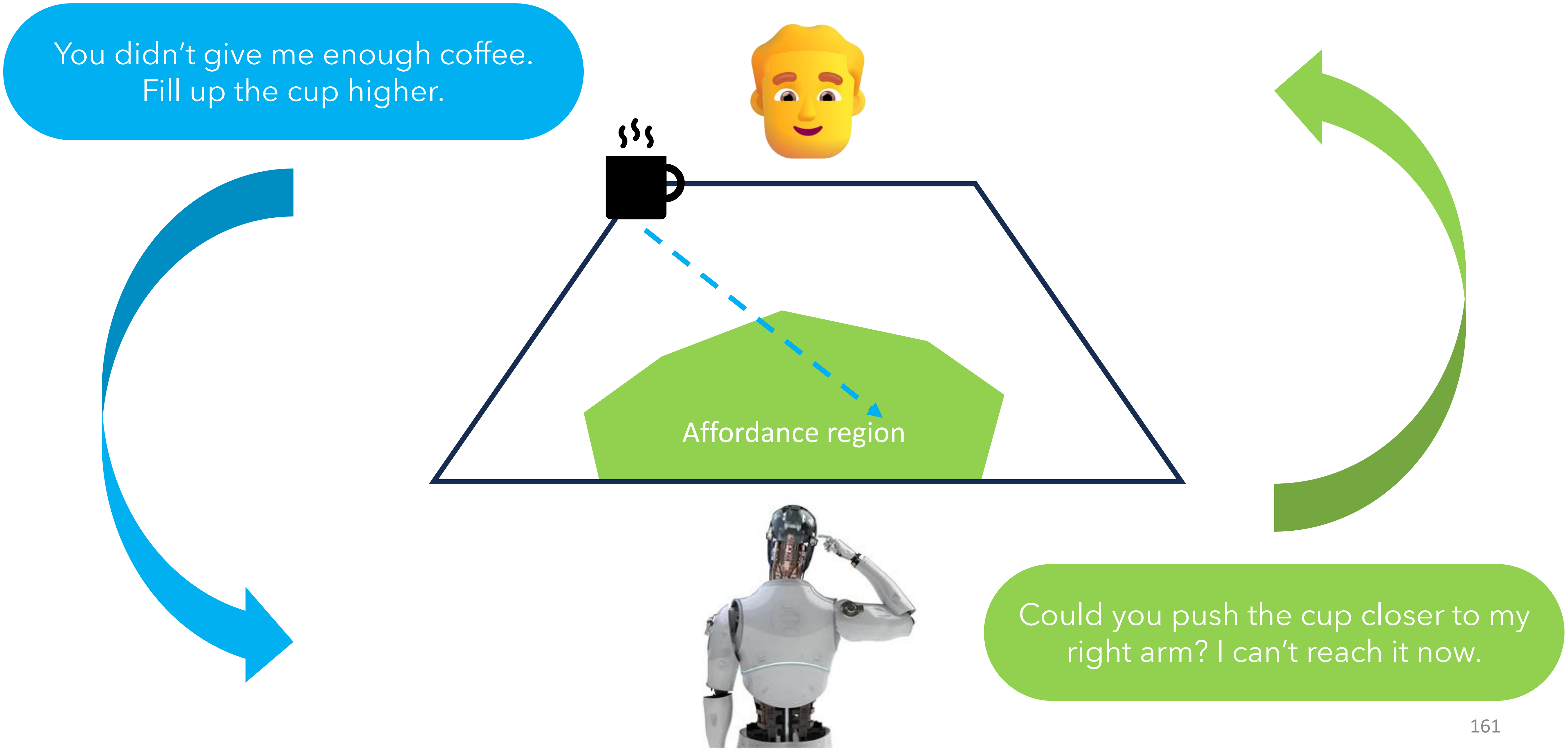


Replanning

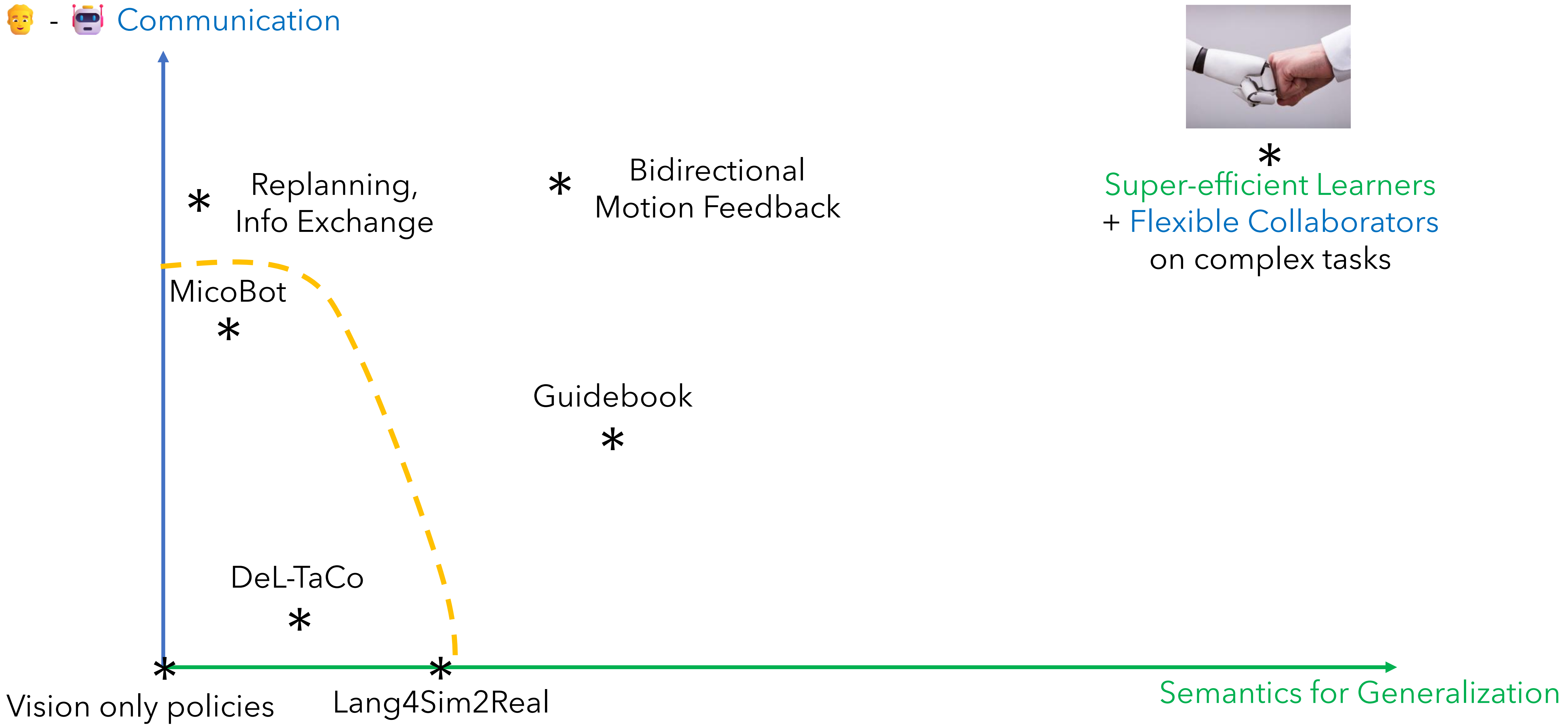


"Actually, can you bring me some water too?"
Human interruption
Replan trigger
Robot Step Failure
Plan
1. Fetch fruit snack and plate
2. Open fruit snack
3. Pour fruit snack on plate
4. Bring plate to human
Create Candidate Plan Edits
Plan Semantic Similarity
Score Candidate Plan Edits
Plan Affordance
Best Plan Edit

Bidirectional Motion Feedback



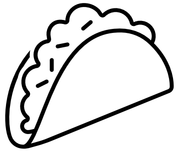
Language for Home-Worthy Real Robots



Conclusion

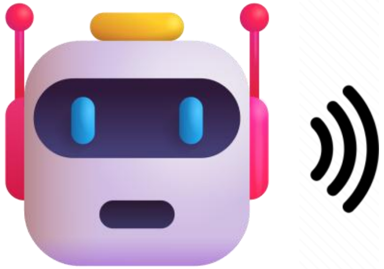
Language augments robotic capabilities.

Generalize to new tasks and domains with tens of demos



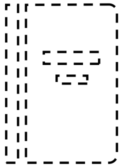
Semantics for Generalization

Perform long-horizon tasks by negotiating for human help



- Communication

Future Work:



Learning Safer Behaviors Faster with guidebooks

Long-Term Future Directions:

Information Exchange
Replanning
Bidirectional Motion Feedback