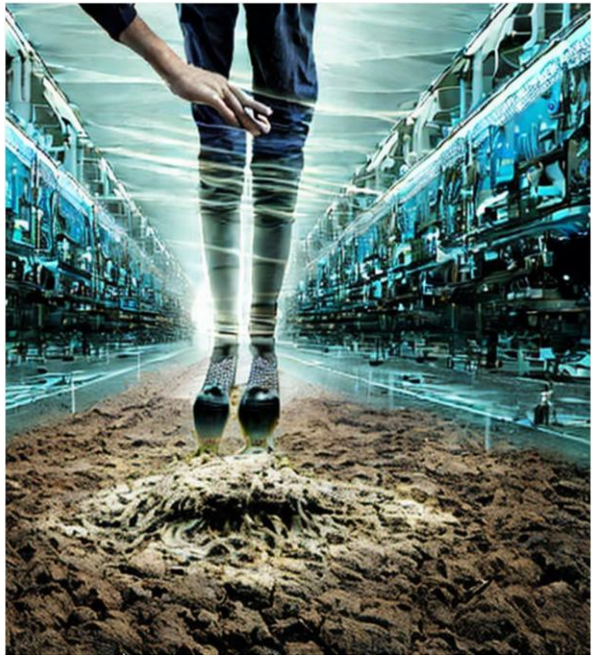
Multi-Modal Vision-Language-Action Foundation Models for Generalizable Robotics

Zsolt Kira
Associate Professor
School of Interactive Computing
Georgia Tech

The great shift



- Modality-specific pipelines
  ▶ DL
    ▶ Transformers

- Scale + Self/semi-supervised learning FTW!
  - Web ▶ Language Models ▶ **Knowledge**
  - DINO/MAE/CLIP/SAM ▶ **Scene Understanding**

Where does robotics go from here?

[Image by Stable Diffusion]

The Reality

- Perception is *still* tied to *known* categories or poor open-vocabulary methods during training
- Brittle to out-of-distribution data
- Limited Open-World abilities
- Even large-scale datasets (RT-X) limited in generalization

% success rates

Method	Seen	Unseen		
		Layouts	Objects	Receptacles
MonolithicRL	91.7 ±1.1	86.3 ±1.4	74.7 ±1.8	52.7 ±2.0
SPA	70.2 ±1.9	72.7 ±1.8	72.7 ±1.8	60.3 ±2.0
SPA-Priv	77.0 ±1.7	80.0 ±1.6	79.2 ±1.7	60.7 ±2.0

Degradation over novelty…

Habitat 2.0
Work by Andrew Szot, Dhruv Batra, and Meta

Open-World Learning (w/ FMs and VLAs)

Reproducible Robotics -> Simulation

Generalization to an Open World

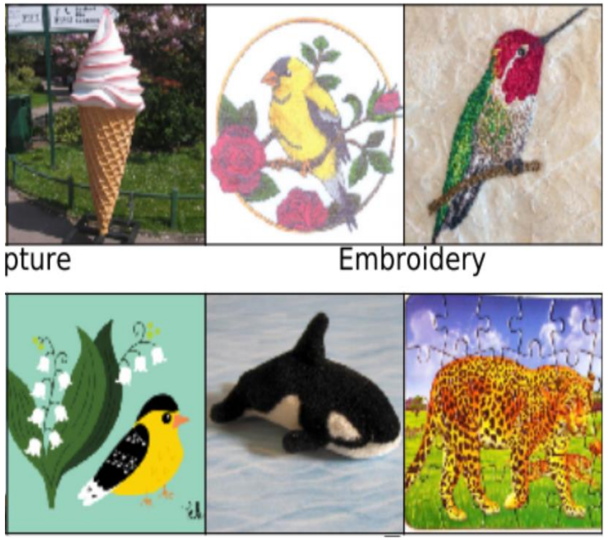
Long-Horizon/Long-Context

Robust Fine-tuning

[NeurIPS 2023 OVMM Challenge, ICML 2023, Neurips 2021]
(w/ Dhruv Batra)

[ICLR 2018/2019, arXiv:2305.10420, ECCV 2022]

[ECCV 2024]

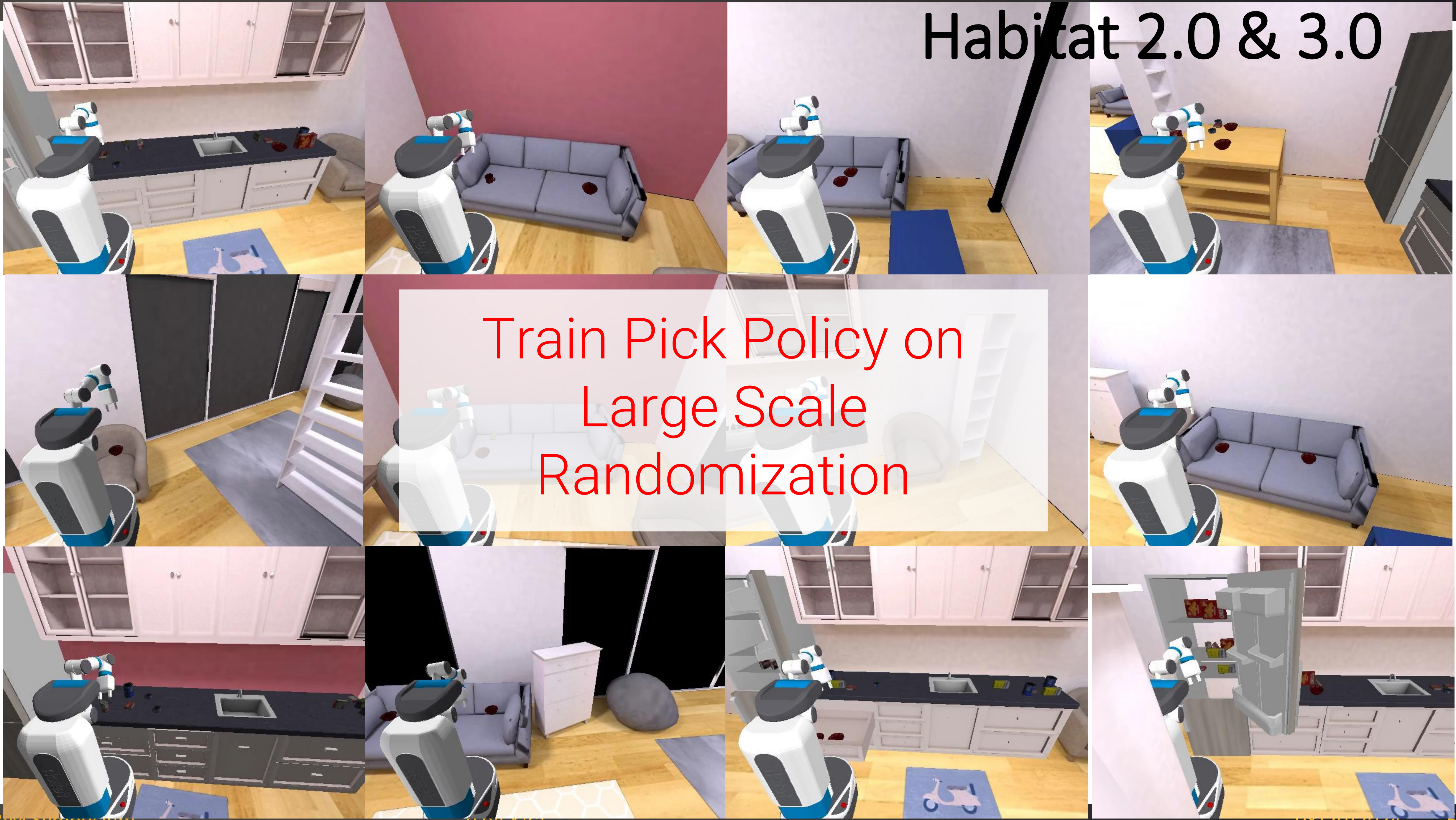


[ImageNet-R]

[CVPR 2023, NeurIPS 2023/2024]

[Middle two images by Stable Diffusion]

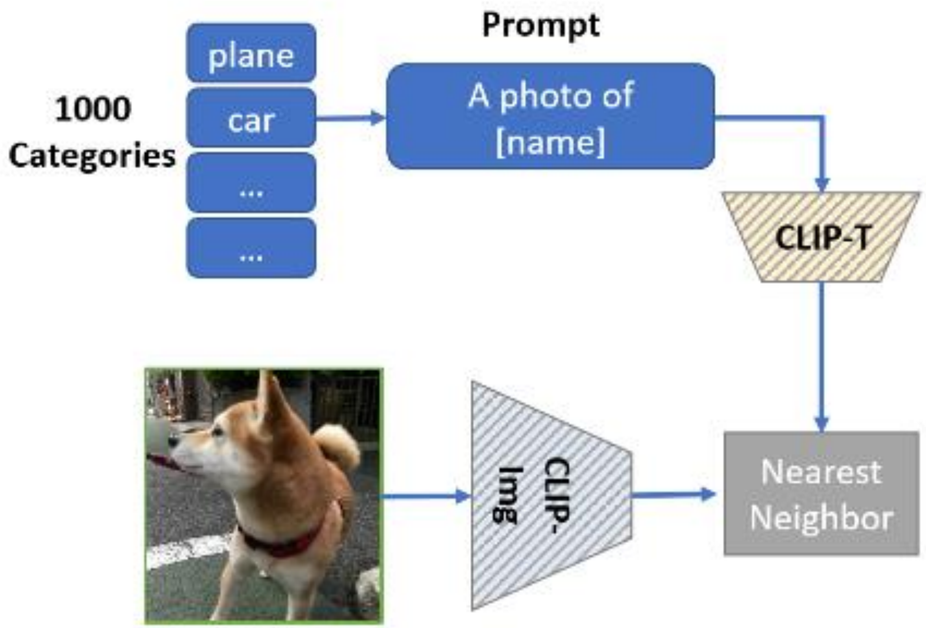
Habitat 2.0 & 3.0



Train Pick Policy on Large Scale Randomization

The Role of Language

- Tremendous progress in language and multi-modal (vision+language) models
- We can leverage these to improve capabilities to learn and name new things



Zero-Shot Learning

Multimodal Large Language Models

Bing's A.I. Chat: 'I Want to Be Alive. 😈'

In a two-hour conversation with our columnist, Microsoft's new chatbot said it would like to be human, had a desire to be destructive and was in love with the person it was chatting with. Here's the transcript.

https://www.nytimes.com/article/ai-artificial-intelligence-chatbot.html

ARTIFICIAL INTELLIGENCE

ChatGPT is about to revolutionize the economy. We need to decide what that looks like.

https://www.technologyreview.com/2023/03/25/1070275/chatgpt-revolutionize-economy-decide-what-looks-like/

Multimodal Large Language Model

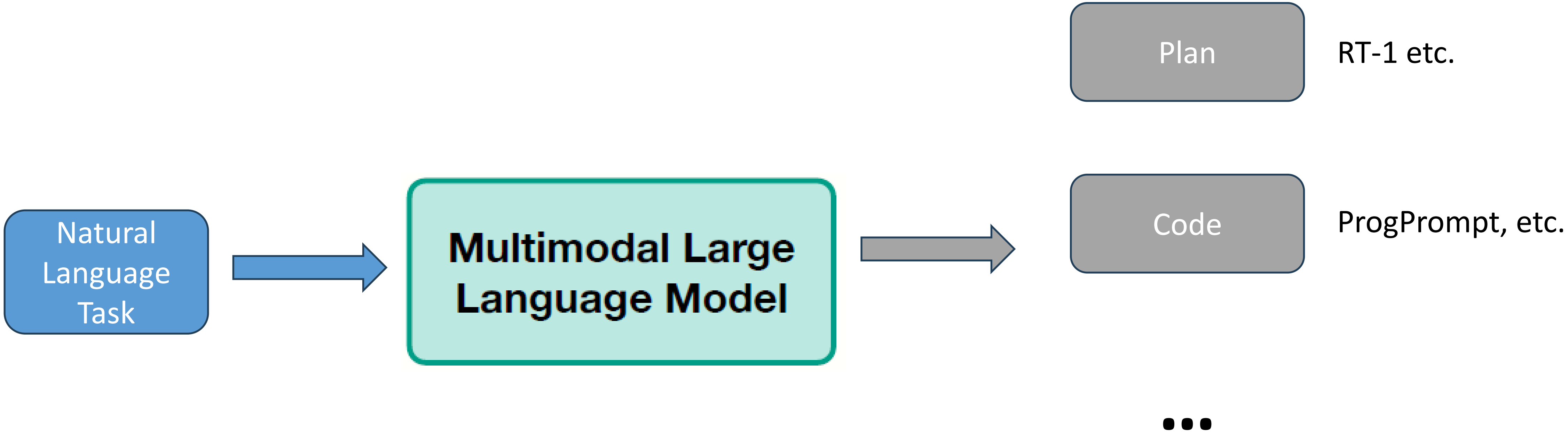
GPT-4o

OPENAI'S LATEST MODEL

Gemini 1.5

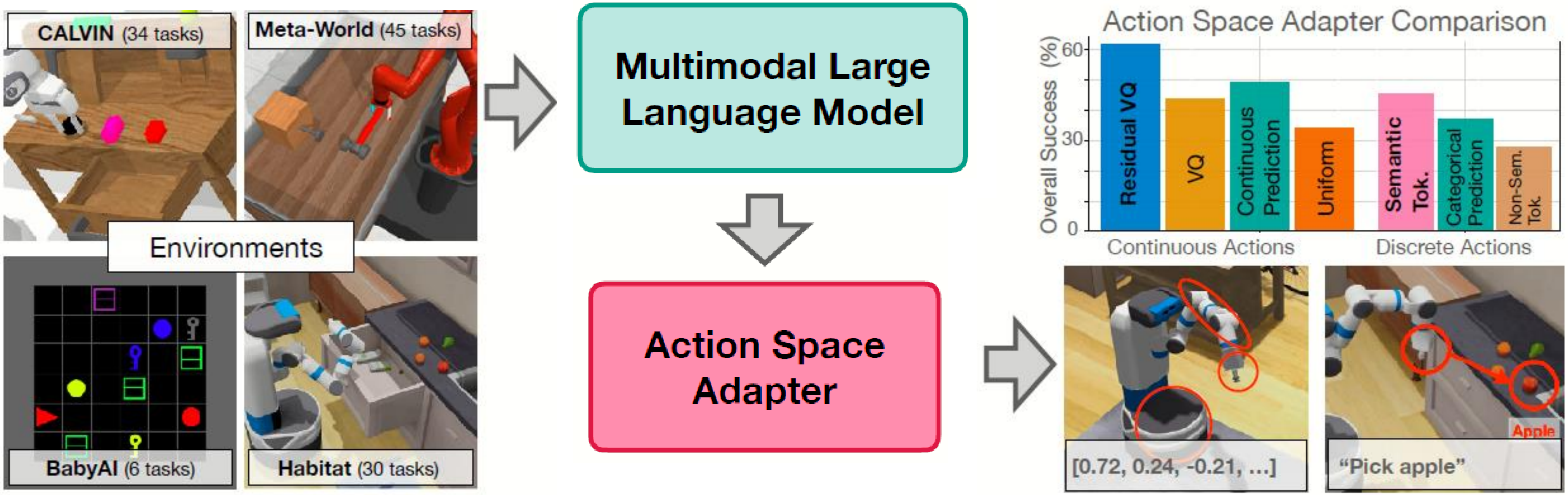
LLAMA 2

Multimodal Large Language Models



What about VLMS for direct task to action?

Vision-Language Action Models



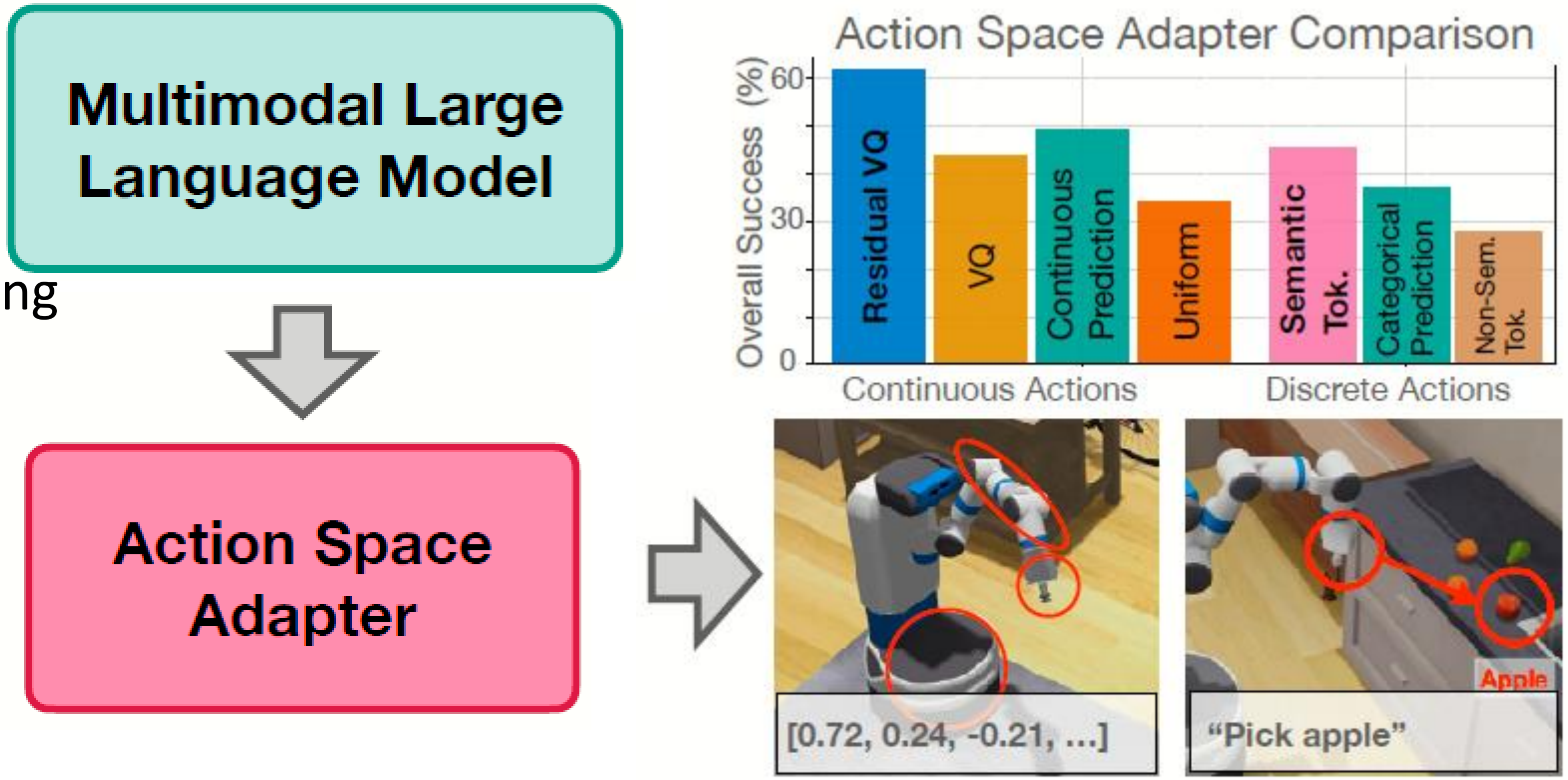
Lots of great concurrent work! OpenVLA, LLARVA, etc.

Szot et al., Grounding Multimodal Large Language Models in Actions, NeurIPS 2024

Vision-Language Action Models

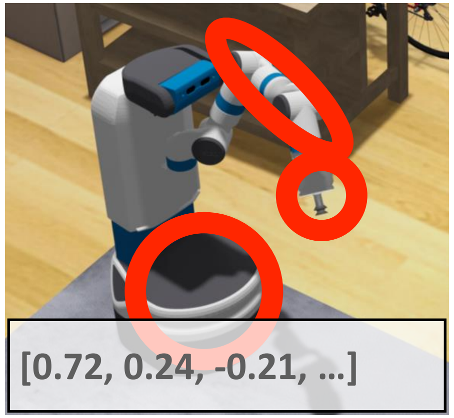
- **Advantages:**
  - Policies driven by textual description of the task!
  - Leverage common sense reasoning inside model
  - Can learn with RL and IL

- **Questions:**
  - How should we represent (tokenize) output actions?
    - Concurrent work tends to just pick one and go with it



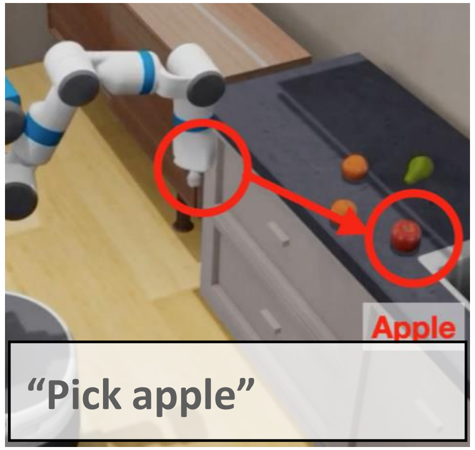
Szot et al., Grounding Multimodal Large Language Models in Actions. NeurIPS 2024.

Action Tokenization



Action is a continuous vector

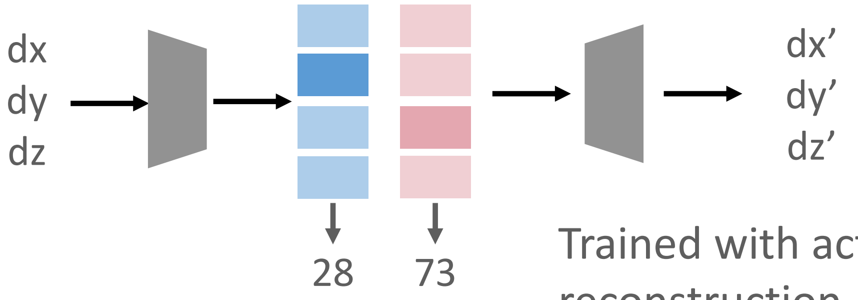
Example: end effector control
[dx, dy, dz]



Action is a selection from a set of discrete choices

Learned Tokenization

Residual VQ-VAE for discrete action tokenization



Trained with action reconstruction

Semantic Tokenization

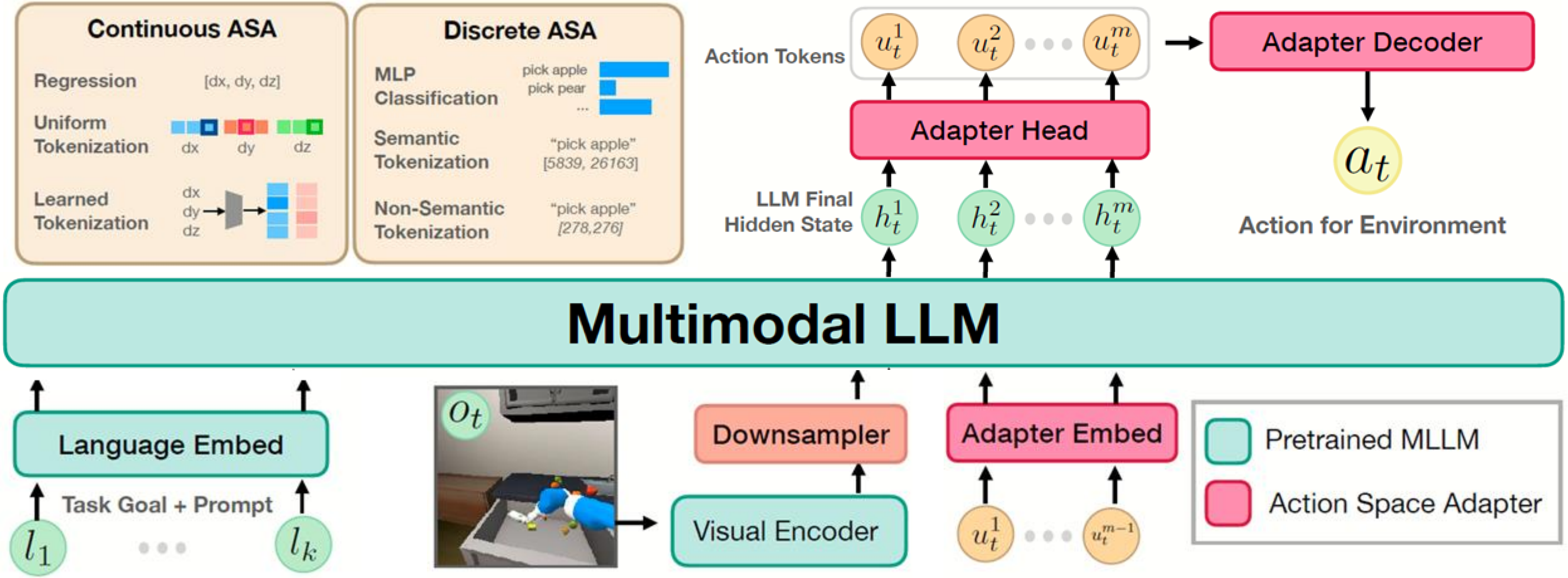
"pick apple"

↓

[278,276]

Describe action with language and tokenize with LLM vocabulary

Szot et al., Grounding Multimodal Large Language Models in Actions. NeurIPS 2024.



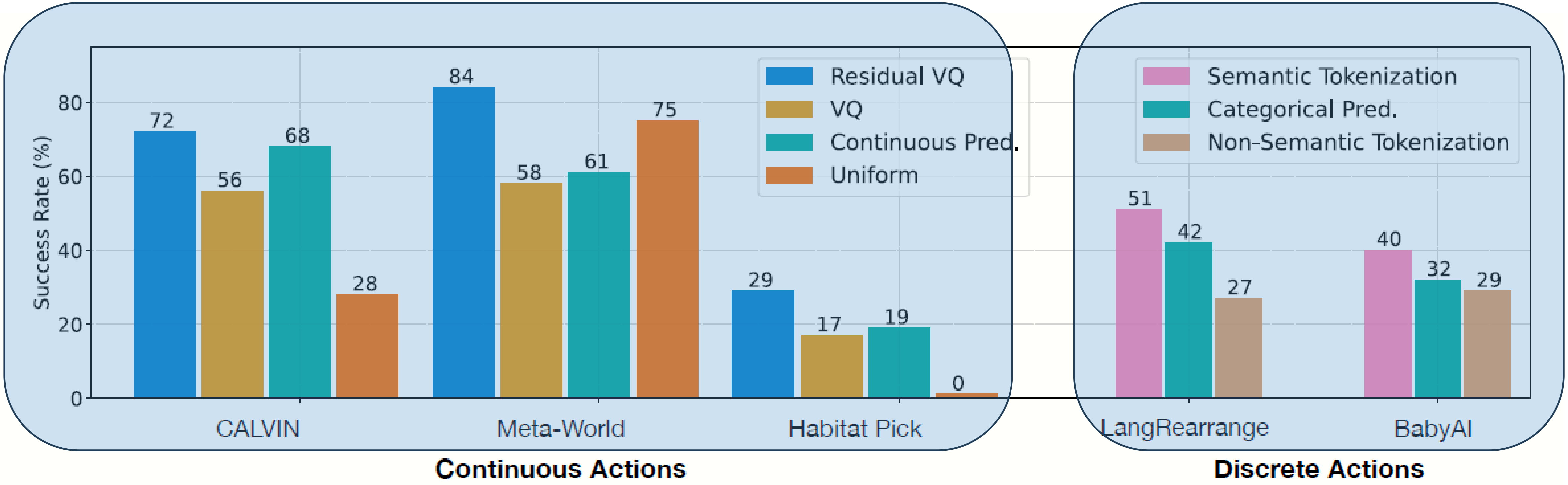
We finetune the ASAs, downsampler, and MLLM

Andrew Szot

ML Ph.D. (co-advised with Dhruv Batra)

Szot et al., Grounding Multimodal Large Language Models in Actions

VLA Results & Findings



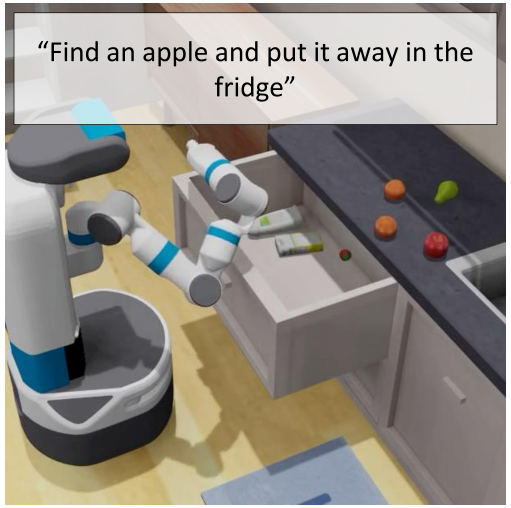
VLA: Results across Spectrum of Generalization

CALVIN | Meta-World | HabPick | BabyAI | LangR

Rotate the pink block towards the right | *Push a button on the coffee machine* | *Pick a lemon* | *Open the blue door* | *Bring something to pour hot coffee in to the TV stand*

		Aggregated		Per Dataset Breakdown										
	Total	Behavior Generalization	Paraphrastic Robustness	Train	Scene	Instruct Rephrasing	Novel Objects	Multiple Rearrange	Referring Expressions	Context	Irrelevant Text	Multiple Objects	Spatial	Conditional Instructs
SemLang	51 ± 1	56 ± 2	47 ± 1	94 ± 3	94 ± 6	92 ± 1	97 ± 0	80 ± 6	31 ± 3	46 ± 14	66 ± 6	2 ± 2	0 ± 0	46 ± 4
Lang	27 ± 12	31 ± 14	24 ± 10	72 ± 13	58 ± 11	74 ± 12	76 ± 29	21 ± 10	10 ± 12	12 ± 11	20 ± 13	0 ± 0	2 ± 3	26 ± 16
Pred	42 ± 2	45 ± 3	38 ± 1	99 ± 1	96 ± 4	92 ± 2	95 ± 4	47 ± 5	26 ± 2	34 ± 2	32 ± 2	0 ± 1	8 ± 1	39 ± 3

Many tasks we want an agent to take actions to autonomously complete



Robotic Manipulation

Navigation

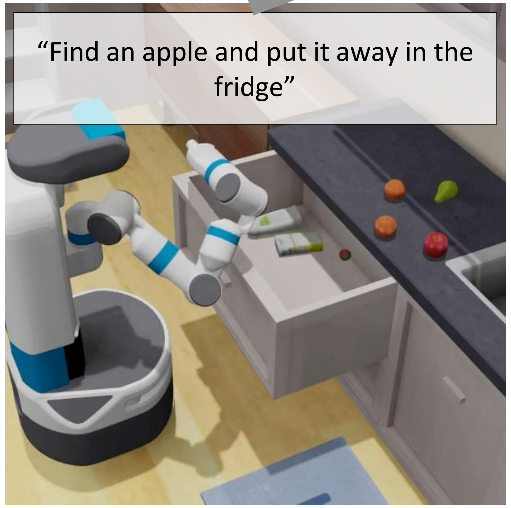
Games

UI Control

Can we have ***one*** policy that does all of these?

How can we create a generalist agent capable of excelling in diverse interactive tasks?

Generalist Agent



Robotic Manipulation

Navigation

Games

UI Control

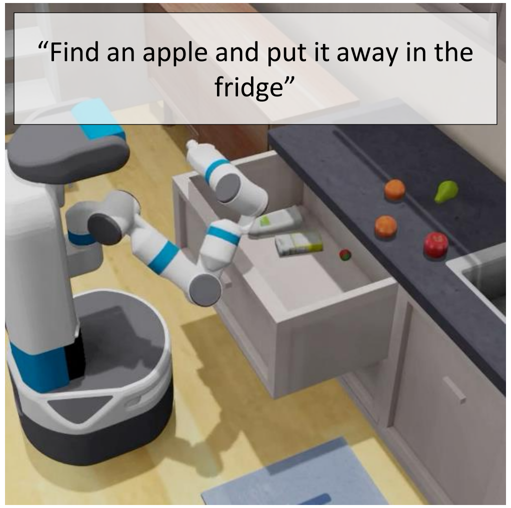
Adapt a pre-trained Multimodal LLM

Multimodal LLM

Interactive Data

Adapt

Generalist Agent



Robotic Manipulation

Navigation

Games

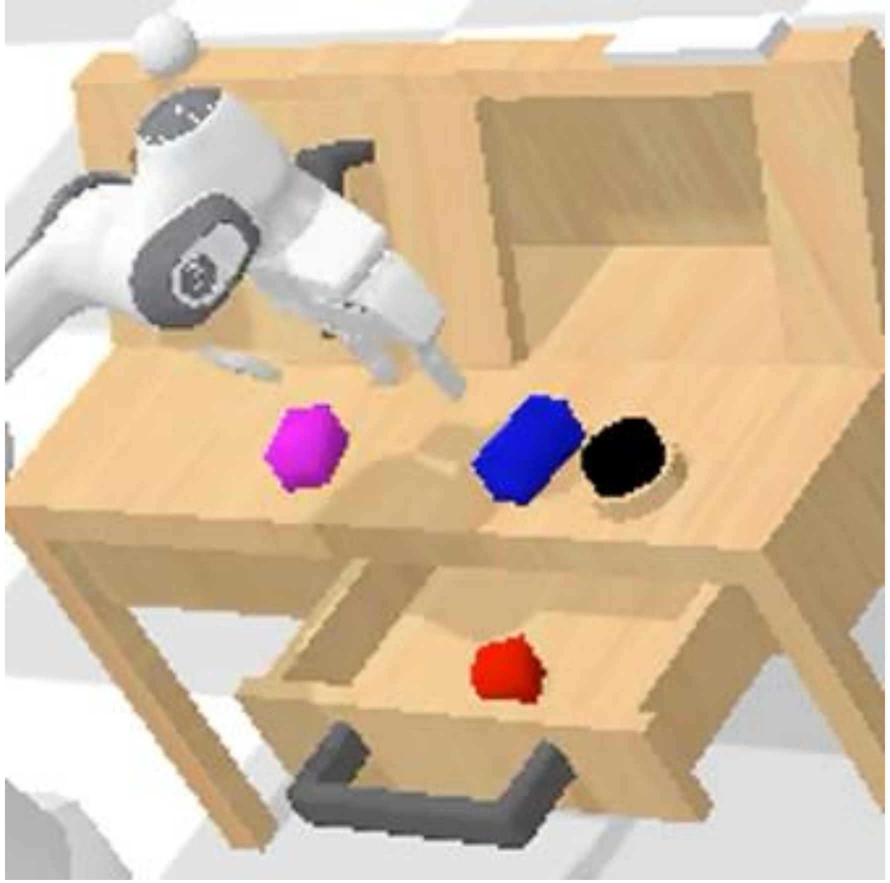
UI Control

Step 1: Collect expert demonstrations in diverse domains for training

From diverse sources, like scripted policies, humans, or RL policies

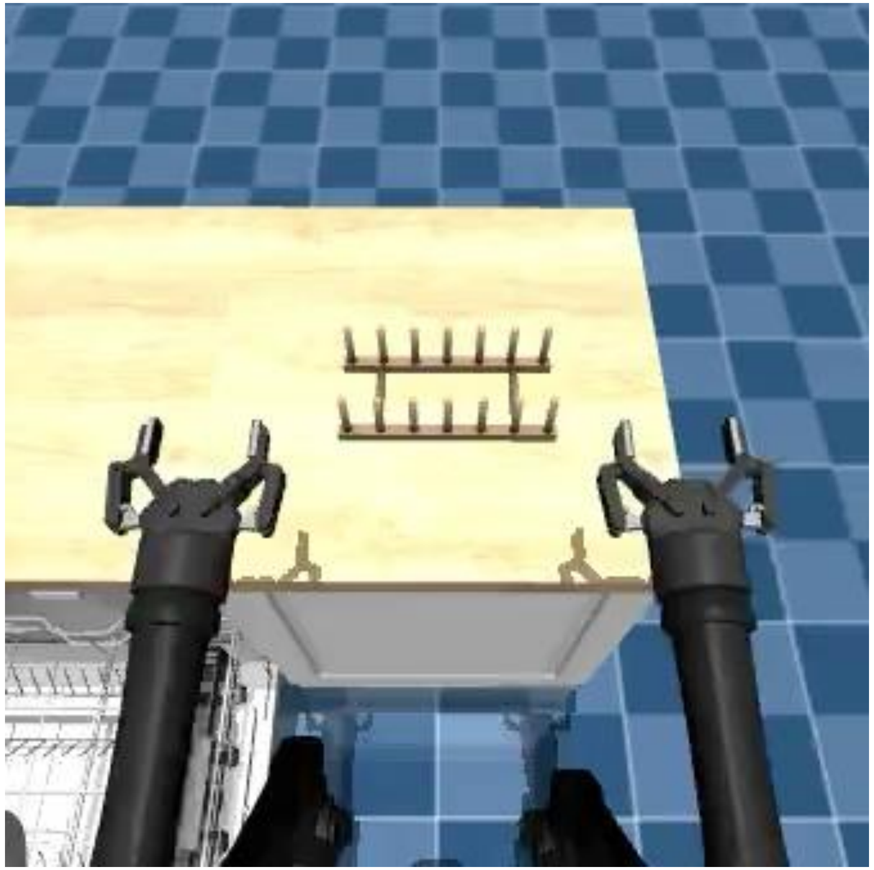
Data - Static Manipulation

"Use the block to pull the handle sideways"



"Move the purple block next to the blue block"

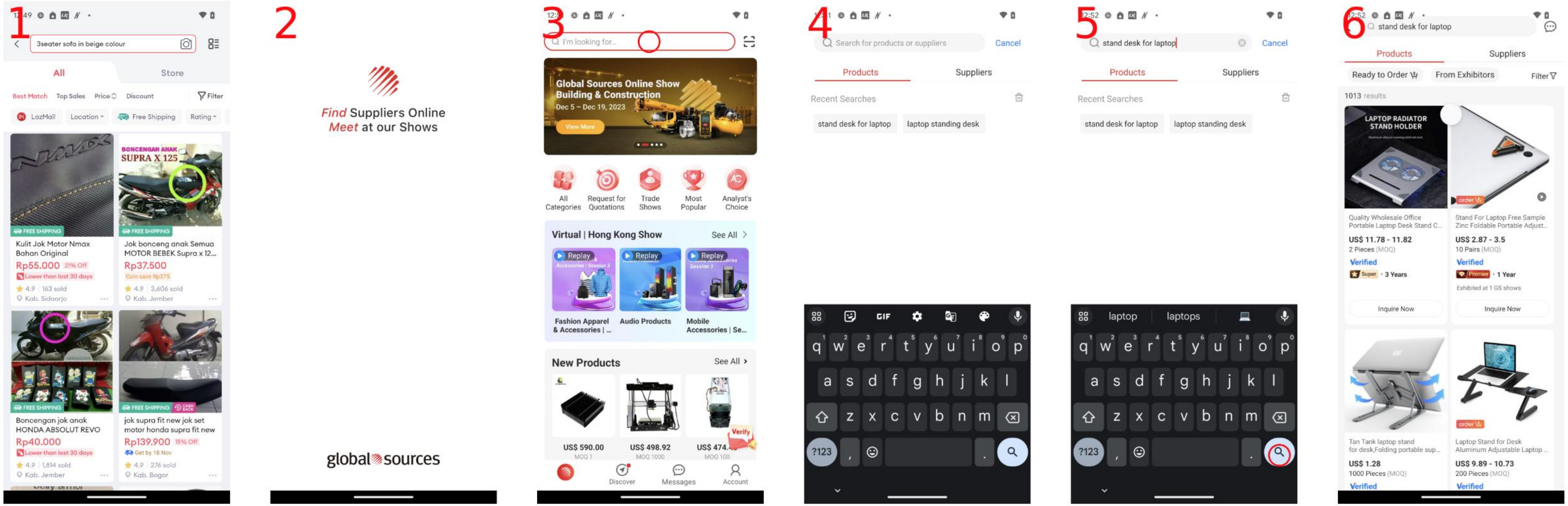
Data - Mobile Manipulation



“Unload the plates from the dishwasher and place them on the rack”

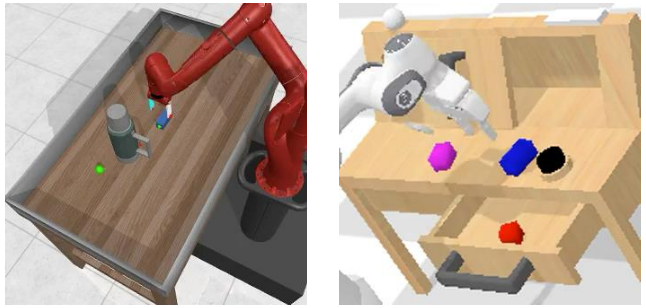
“Pick up the banana”

Data - UI Control

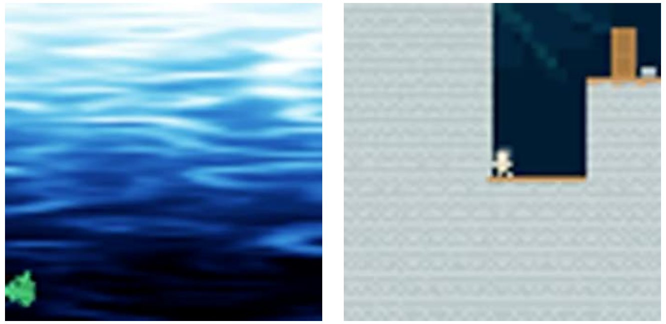


"Find me a standing desk for my laptop from the GlobalSources app"

Static Manipulation

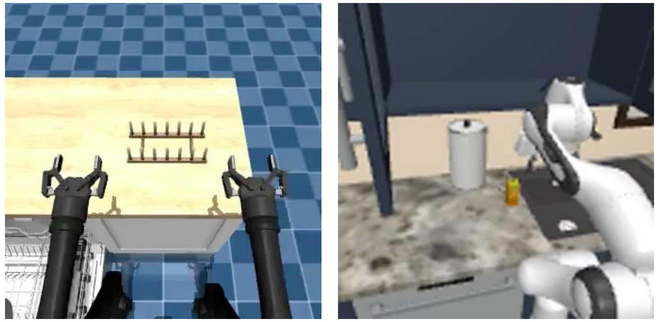


Games



Navigation

Mobile Manipulation



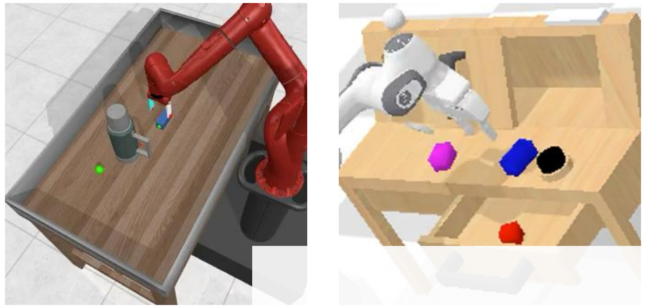
UI Control

Real Robots

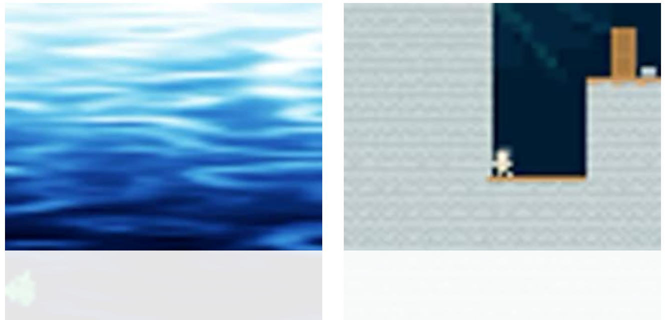
Character Control

Planning

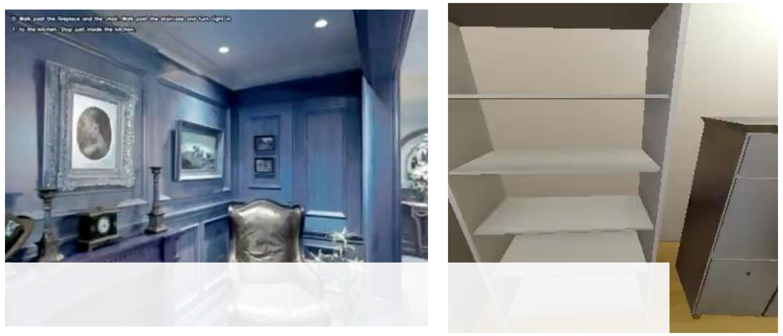
Static Manipulation



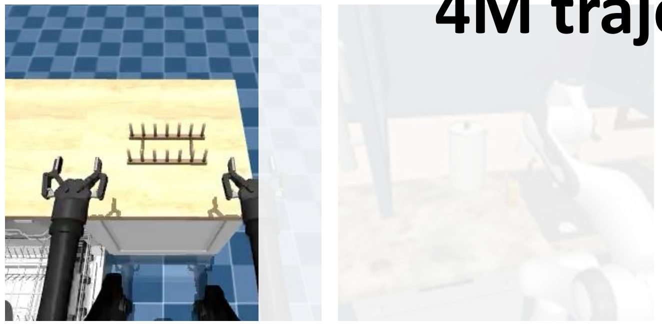
Games



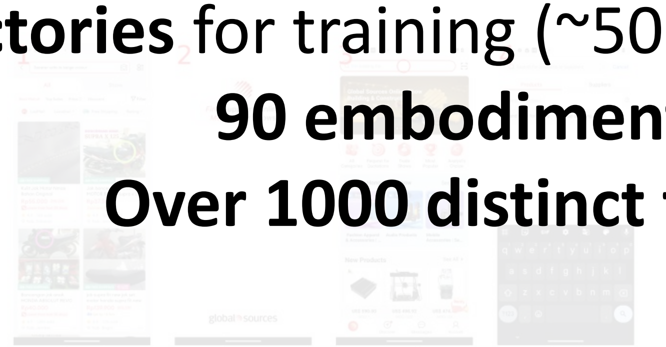
Navigation



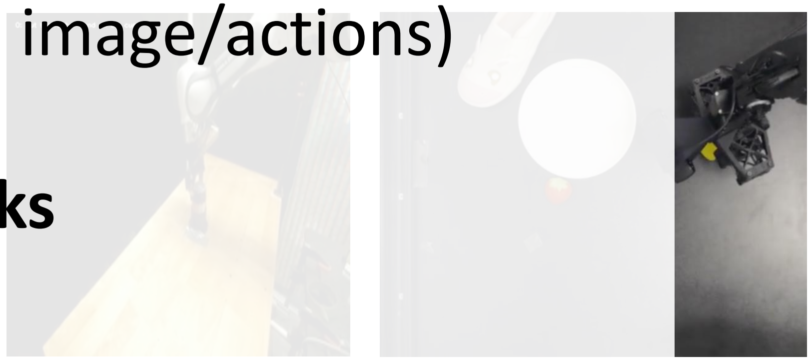
Mobile Manipulation



UI Control



Real Robots

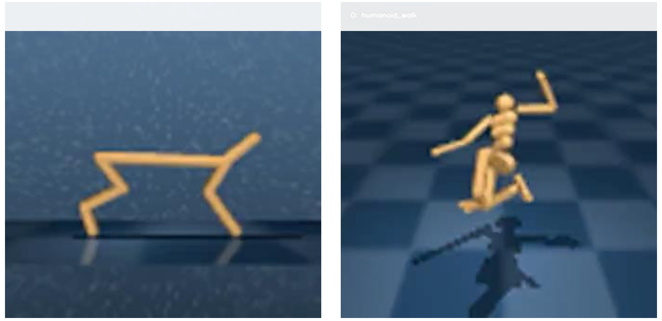


4M trajectories for training (~500M image/actions)

90 embodiments

Over 1000 distinct tasks

Character Control



Planning

Evaluation

New Tasks

Find an apple and put it away in the fridge.

Novel Objects
Find a pear and put it away in the fridge.

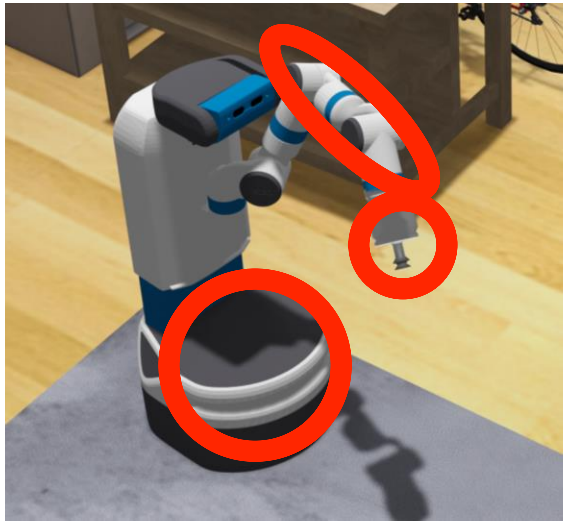
Context
I am hungry for something sweet and healthy. Put a snack for me on the table.

Spatial Relationships
Find an apple and put it in the receptacle to the right of the kitchen counter.

⋮

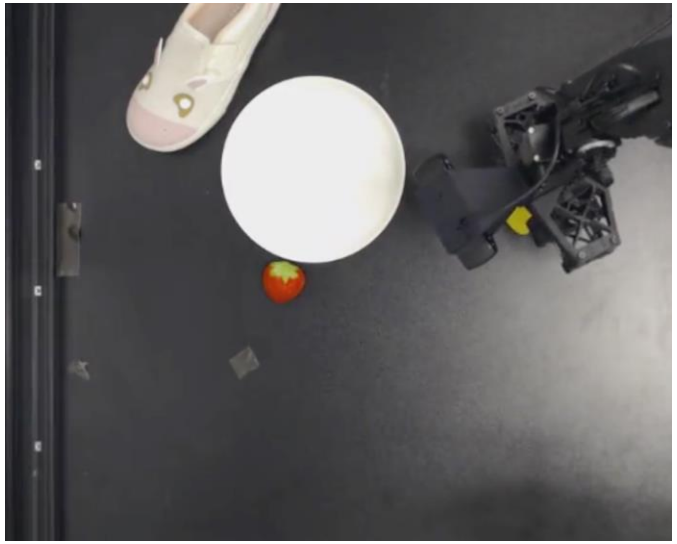
New Embodiments

New control spaces and robot types



New Environments

New platform with limited data



Future Work

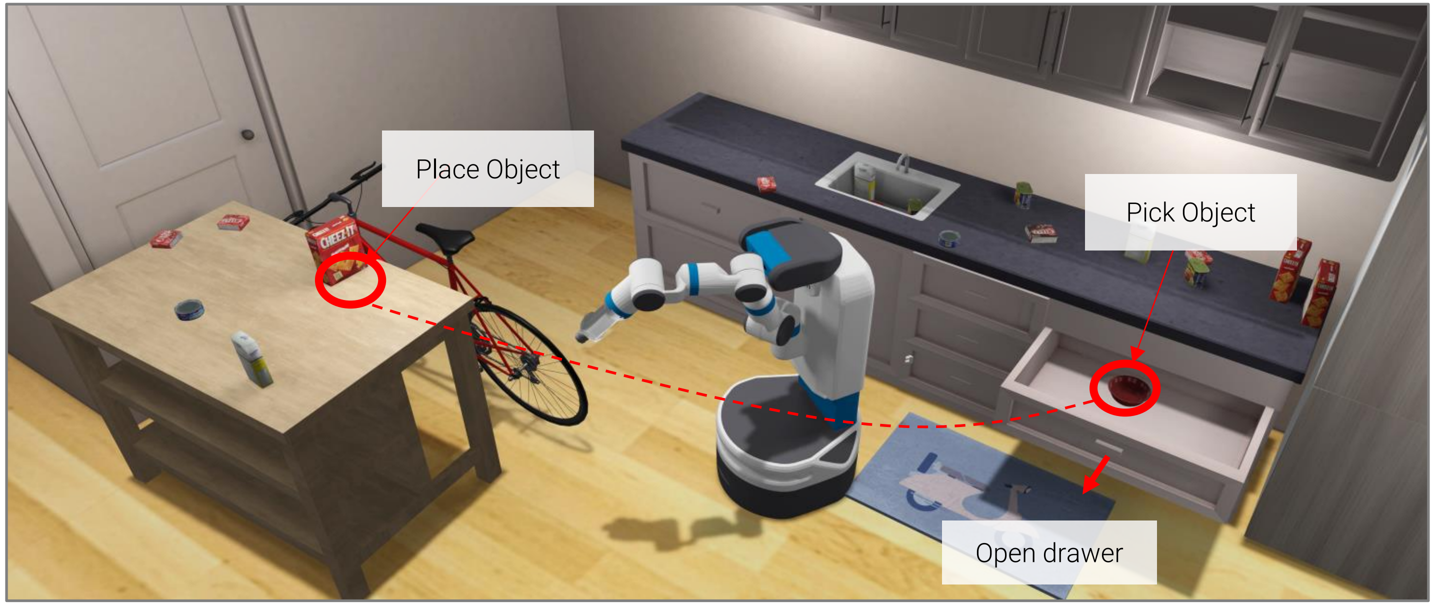
- Adaption to new environments by investigating:
  - # of new demonstrations vs. success rate with supervised fine-tuning
  - # of experiences vs. success rate with reinforcement learning
- Investigating how online data collection can boost performance
- Insights from model training

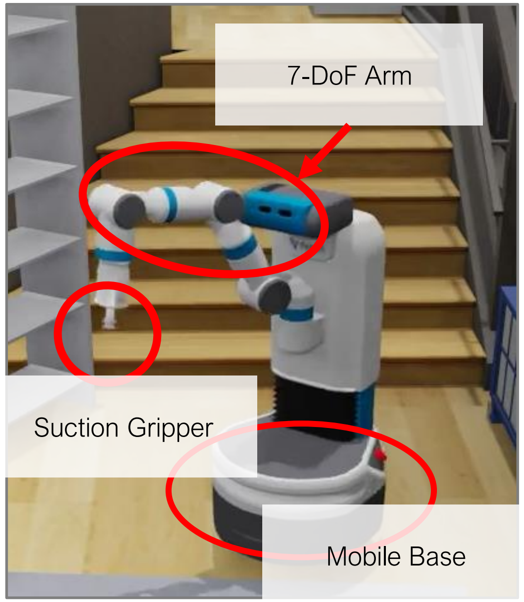
Reinforcement Learning via Auxiliary Task Distillation

Consider the task of using a robot to rearrange an object in the house

- Fetch-Robot with 10-DOF and a suction gripper
- Requires diverse skills like Navigating, Opening a cabinet, Picking up, and Placing

Can long-horizon robot control be learnt end-to-end without using demonstrations or a curriculum?

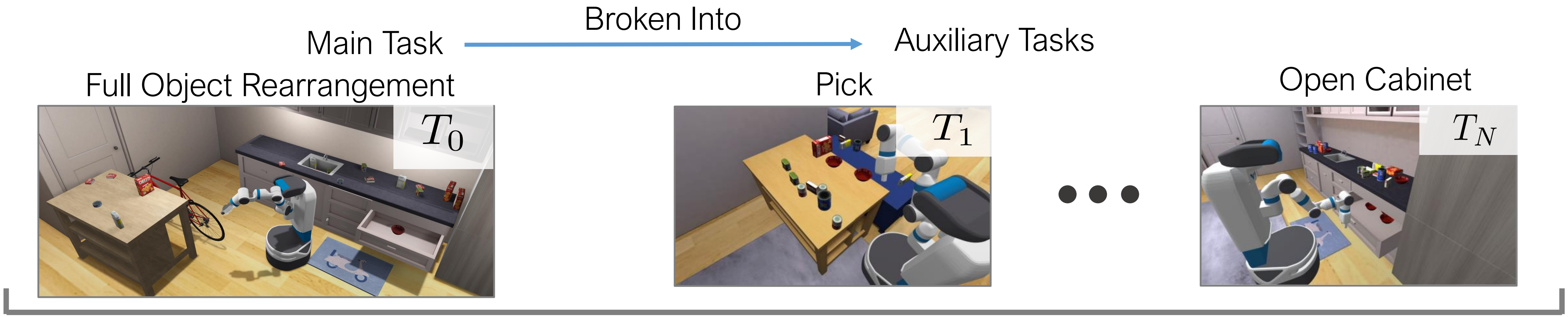


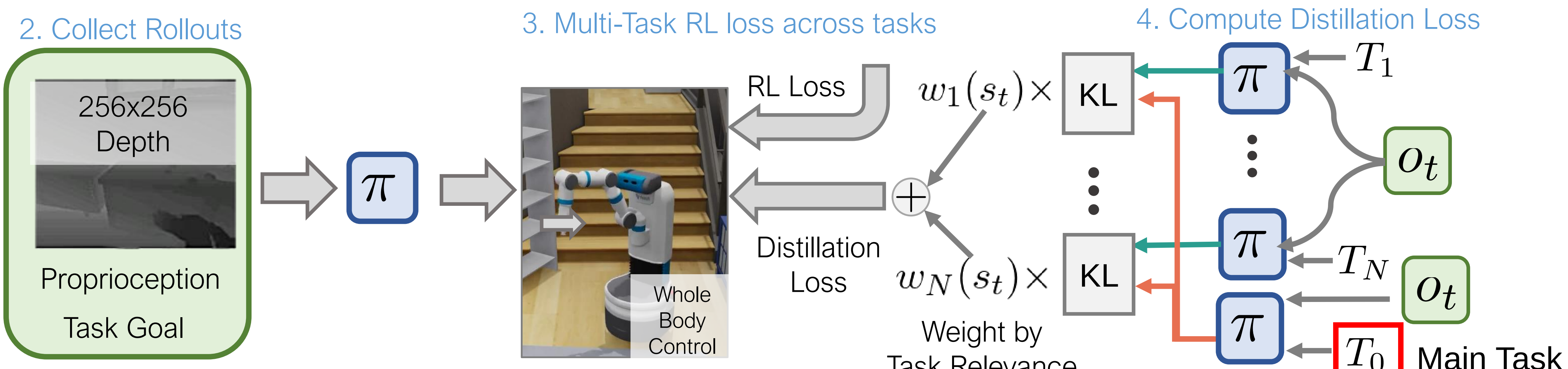


Harish et al., Reinforcement Learning via Auxiliary Task Distillation, ECCV 2024

Yes, by using Auxiliary Tasks!

- Auxiliary tasks carry relevant behaviors which are easier to learn and transferred to the main task
- They are learnt simultaneously along with the main task





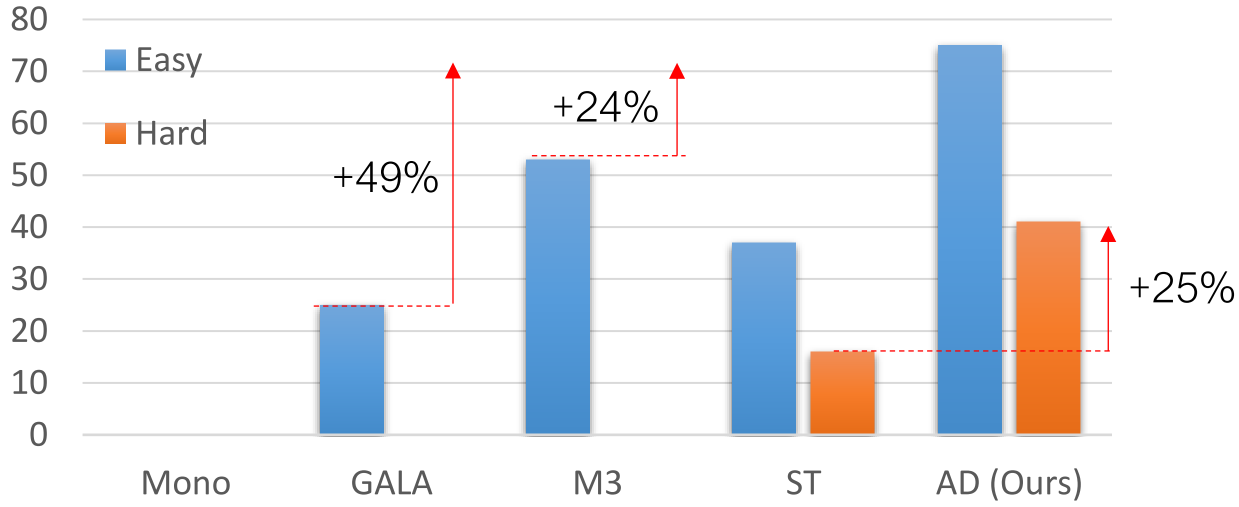
Harish et al., Reinforcement Learning via Auxiliary Task Distillation, ECCV 2024

Results

Outperforms a variety of end-to-end and hierarchical baselines by 2.3x

Easy: Episodes in which the object is placed in an open receptacle
Hard: Object is placed inside a closed receptacle

- M3 (+24%) ➔ Hierarchical RL with STRIPS planner with Navigate, Pick and Place skills
- Mono (+73%) ➔ end to end RL which directly maps observations to actions
- GALA (+24%): Scaling end to end RL with kinematic simulation (2B samples: x4 more than Aux-Distill)
- ST (+25%)➔ Transformer architecture for rearrangement using demonstrations



Harish et al., Reinforcement Learning via Auxiliary Task Distillation, ECCV 2024

AuxDistill

Reinforcement Learning via Auxiliary Task Distillation

ECCV 2024
Wed. Oct 2nd
10:30am

Abhinav Harish

M.S. Thesis Student

Harish et al., Reinforcement Learning via Auxiliary Task Distillation, ECCV 2024

Conclusions

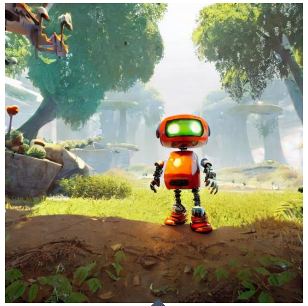
- Already getting benefits of language!
  - Natural task specification
  - Semantic actions
  - Embodiment prompt
- Future goals include combining these ideas into unified architectures
- Focus on:
  - Generalization
  - Long-Horizon / Long Context
  - Robustness

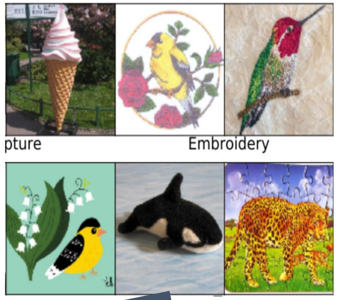
Reproducible Robotics -> Simulation

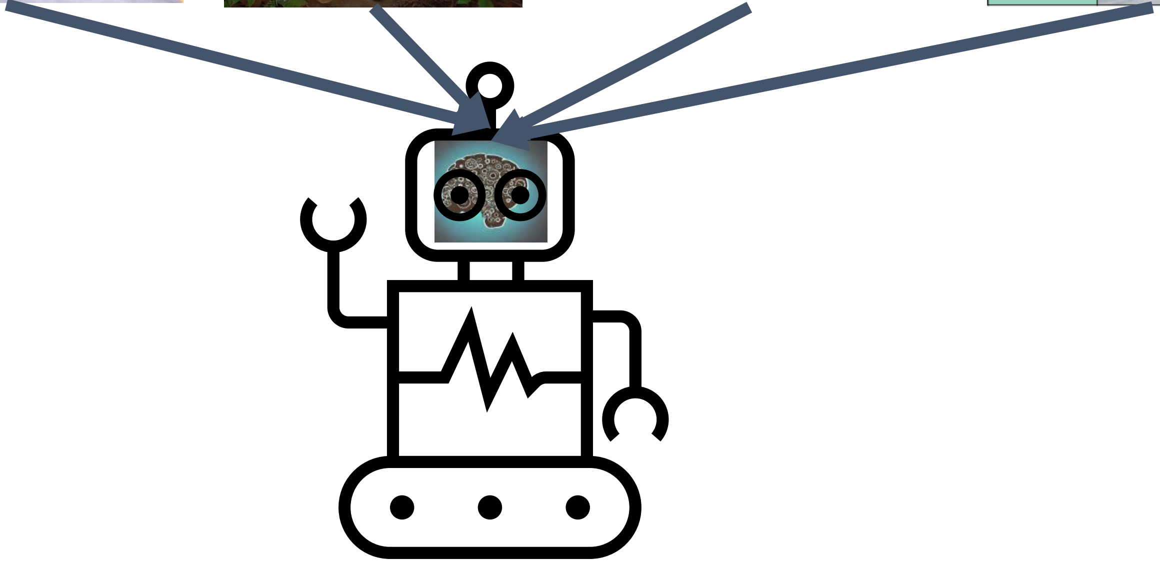
Discovering an Open World

Long-Horizon / Long-Context

Robust Fine-tuning







Acknowledgement and Questions

Andrew Szot

ML Ph.D. (co-advised with Dhruv Batra)

Abhinav Harish

M.S. Thesis Student

Yusuf Ali

CS Ph.D.

Jeremiah Coholich

Robotics Ph.D.

Karmesh Yadav

CS Ph.D. (co-advised with Dhruv Batra)

SAMSUNG RESEARCH AMERICA

MIT-IBM Watson AI Lab

Meta

DARPA

NSF

Office of Naval Research Science & Technology