

Lecture 12

Safety and Uncertainty

Q of the Day: Explain intuitively what mutual information is.



<https://forms.gle/AM2ZNkN6ahhxceMq9>



Last Time

☒ active learning

This Time

☐ final presentation logistics

☐ tips on giving a good presentation

☐ safety and uncertainty

Logistics

- This Thursday – Guest lecture from Prof. Sidd Karamcheti on “*HRI in the Era of Foundation Models*”
- Next Tuesday – No Class 😊

Upcoming Deadlines

- Dec. 2 – Final Project Presentations: Part 1
- Dec. 4 -- Final Project Presentations: Part 2
- Dec. 11 – Final Project Report

Before Class:

-
- HRI 25: Final Presentations
- File Edit View Insert Format Tools Extensions Help
- Slide Show
- Menu
- 1 2 3 4 5 6 7
- ## Final Presentation Instructions
- ### Before Class on Dec. 2
- Put your presentations into this slide deck
 - All talks should be **12 min presentation + 2-3 min Q&A**
 - Make sure your slides cover: (1) Motivation, (2) Problem Statement / Research Question, (3) Why it is Hard / Novelty, (4) Key Insight or Hypothesis, (5) Results
 - Andrea will download the slides on her computer on morning of Dec. 2 so that no additional changes can be made and groups presenting on Dec. 4 do not have a time advantage.**
- ### During Class on Dec. 2 and Dec. 4
- We will go start-to-finish through this slide deck.
 - Every team member should present roughly an equal proportion of the presentation.**
 - When you *not* presenting, you must write **feedback to your assigned presenter(s)** via this [Google Doc](#) (the quality of your feedback will be part of your grade). The feedback can include, but is not limited to comments about how well the overall research problem was motivated, the novelty / impact / soundness of the proposed idea, the presented results, and the overall presentation quality. Teams will have a chance to incorporate this feedback into their final project reports.
- 1 2 3 4 5 6 7
- Day 1
Monday, Dec 2
1. Jerry Wang
- Example Project Title
2. David Smith
3. Dean Kahan

- We will go start-to-finish through this slide deck
- Every team member should **present roughly an equal proportion of the presentation.**
- When you *not* presenting, you must **write feedback to your assigned presenter(s) via the Google Doc** (the quality of your feedback will be part of your grade).

Class Feedback: HRI '25 Final Presentations

Instructions

When you are not presenting, you must write up feedback for your assigned presenter(s) down below. The feedback can include, but is not limited to, comments about how well the overall research problem was motivated, the novelty / impact / soundness of the proposed idea, the presented results, and the overall presentation quality. Teams will have a chance to incorporate this feedback into their final project reports.

Note: We assigned you to give feedback to the *same* groups that you gave feedback to during the Midterm Project Pitches so that you are more calibrated to the progress that the team has made since the middle of the semester.

Day 1

Presenter: Jenny Wang
Jack

Angelica

Daehwa

Presenter: Owen Kwon
Sreyas

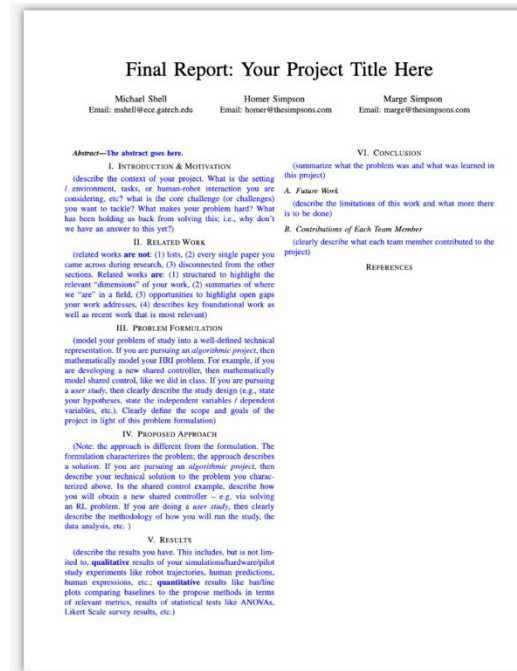
Ding-jun

Dec. 11– Final Project Report

- The final report should present your final findings in a research or survey paper format.
- The length should be maximum 6 pages, double-column.
- Follow the template and answer the questions – you should build on your Mid-term Report, address the feedback you got from me, Yilin, and the class!
- You can also see the grading rubric on Canvas. We are looking to see progress throughout the semester relative to where you started.

The final report should present your final findings in a research or survey paper format. The length should be maximum 6 pages, double-column. The grade will be determined based on the content quality and not on the absolute length (please see the grading rubric below).

Please use the attached Latex template and follow the structure of the subsections.





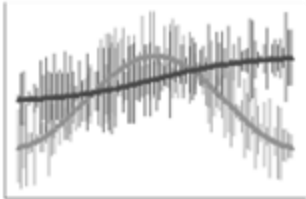
Latex Template (zip file): [final-report-latex.zip](#) ↓

Heuristics for good presentations


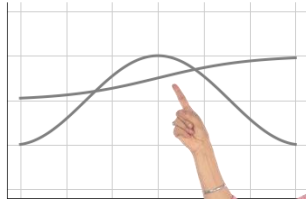
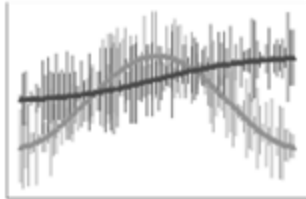




Heuristics for good presentations

Sparse; figures over text!

- Increased information
- Increased viewer effort



focus on slide



focus on speaker

Heuristics for presentations


Distill your ideas and takeaways for the viewer

Key Idea

- Several reactive control strategies have been developed to deal with pHRI
- But the robot returns to its original motion stems from a fundamental limitation of traditional pHRI strategies: they miss the fact that human interventions are often intentional
- Key idea is that because pHRI is intentional, it is also informative
- pHRI provides observations about the correct robot objective function, and the robot can leverage these observations to learn that correct objective.

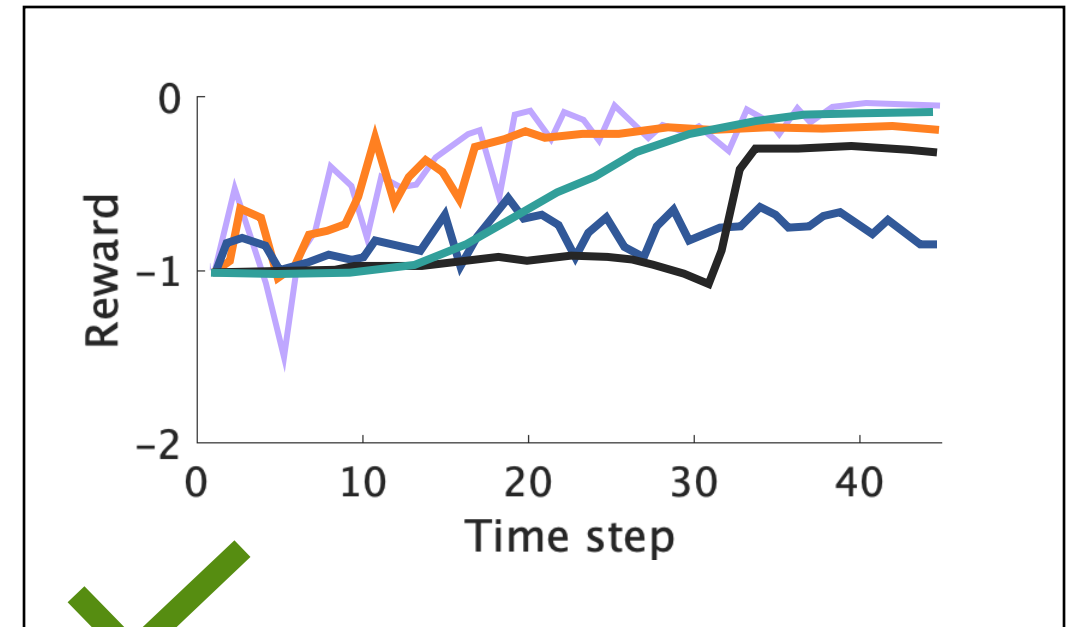
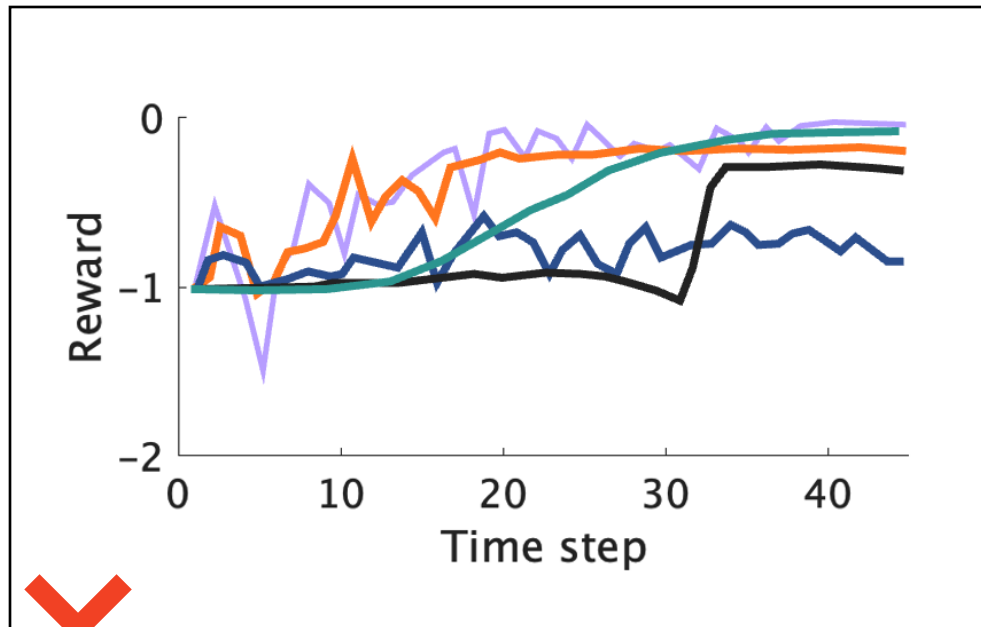
Key Idea

Because pHRI is intentional, it is also informative.

 *Human corrections are observations about the correct robot objective function*

Heuristics for good presentations

Be visual (e.g., make graphs and break them down)



Heuristics for good presentations

If using equations, explain them and build them up

HJ Reachability

$$\max_{\substack{u \\ \text{red}}} \min_{\substack{d \\ \text{blue}}} \nabla_x V(x, t)^\top f(x, \text{red } u, \text{blue } d) + \frac{\partial V}{\partial t} = 0$$



HJ Reachability

$$f(x, \text{red } u, \text{blue } d)$$

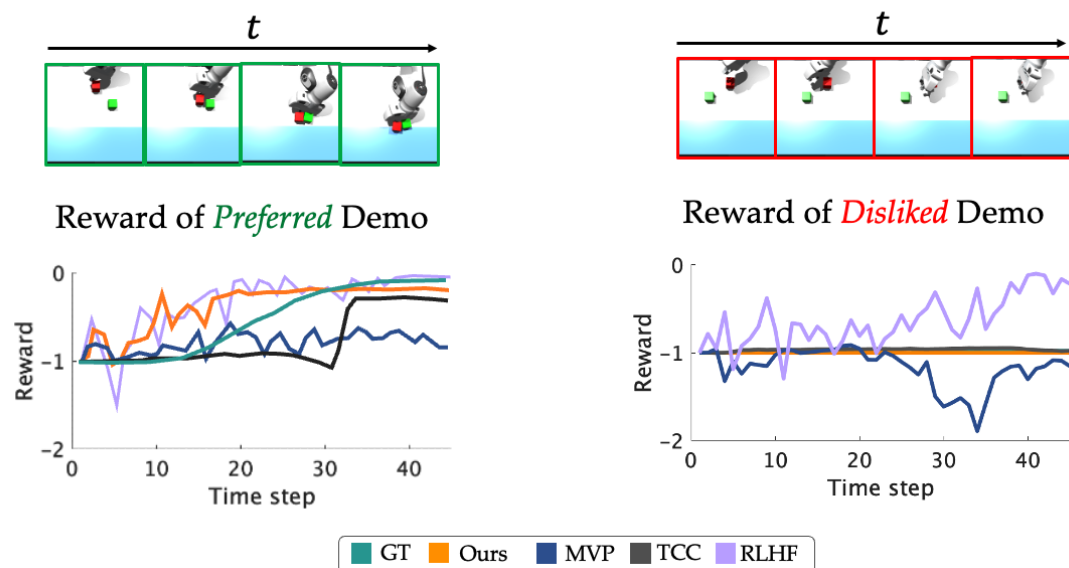
Player 1 *Player 2*



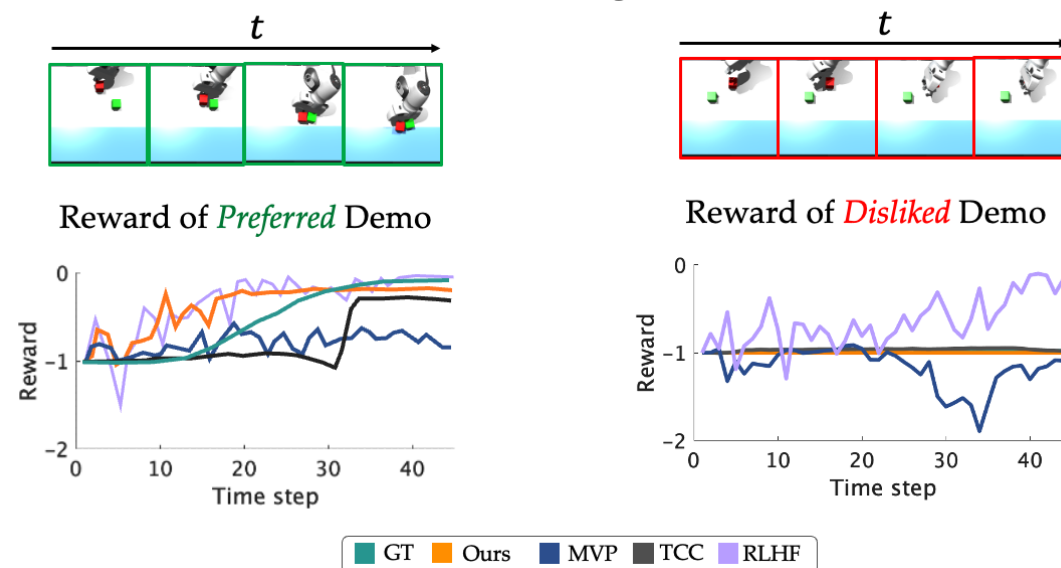
Heuristics for presentations

Use useful titles

Results






Visual rewards with an *aligned representation* are highly correlated* with the ground-truth reward



Go to the Class Website for More Resources

FAQ

TABLE OF CONTENTS

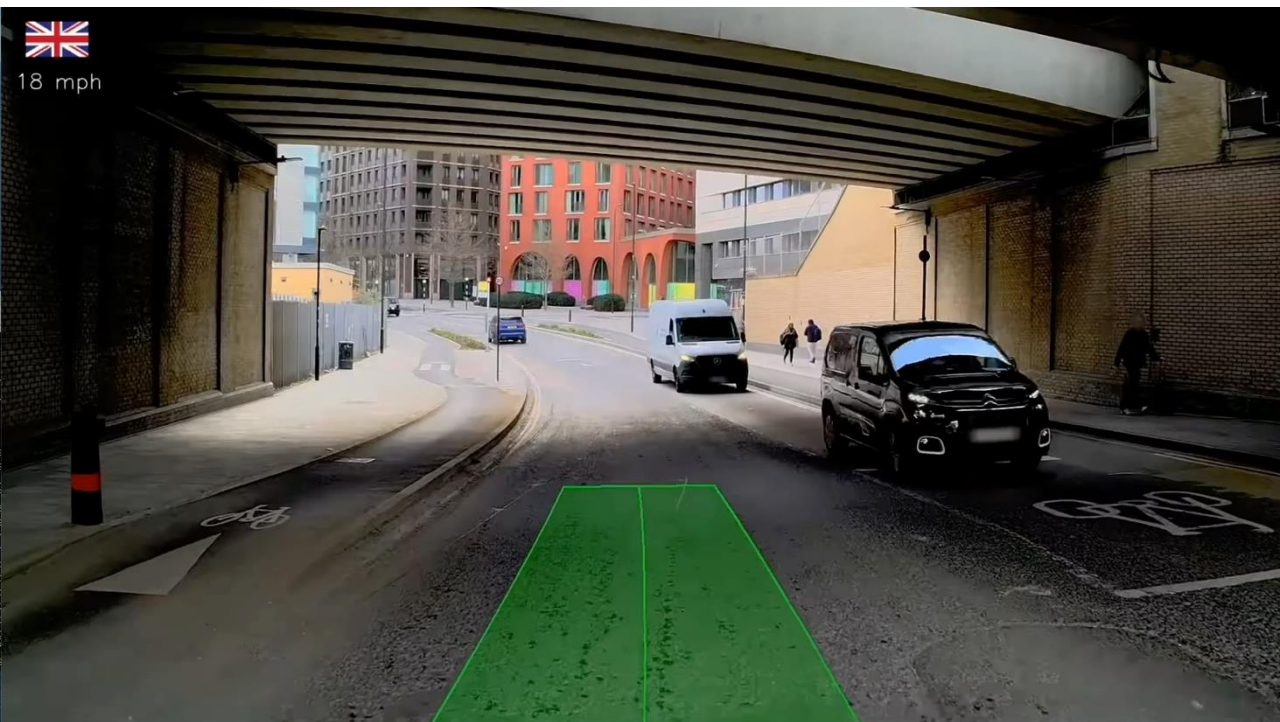
- 1 [I don't have access to robots! What can I do?](#)
- 2 [I don't have access to compute! What can I do?](#)
- 3 [For my class project, I want to test an algorithm with people. Do I need to run a formal user study?](#)
- 4 [How can you effectively read a research paper?](#) 
- 5 [How do you write a good research paper?](#) 
- 6 [How do I make nice figures for a paper or talk?](#) 
- 7 [Is there a textbook for the class?](#)

Now onto....

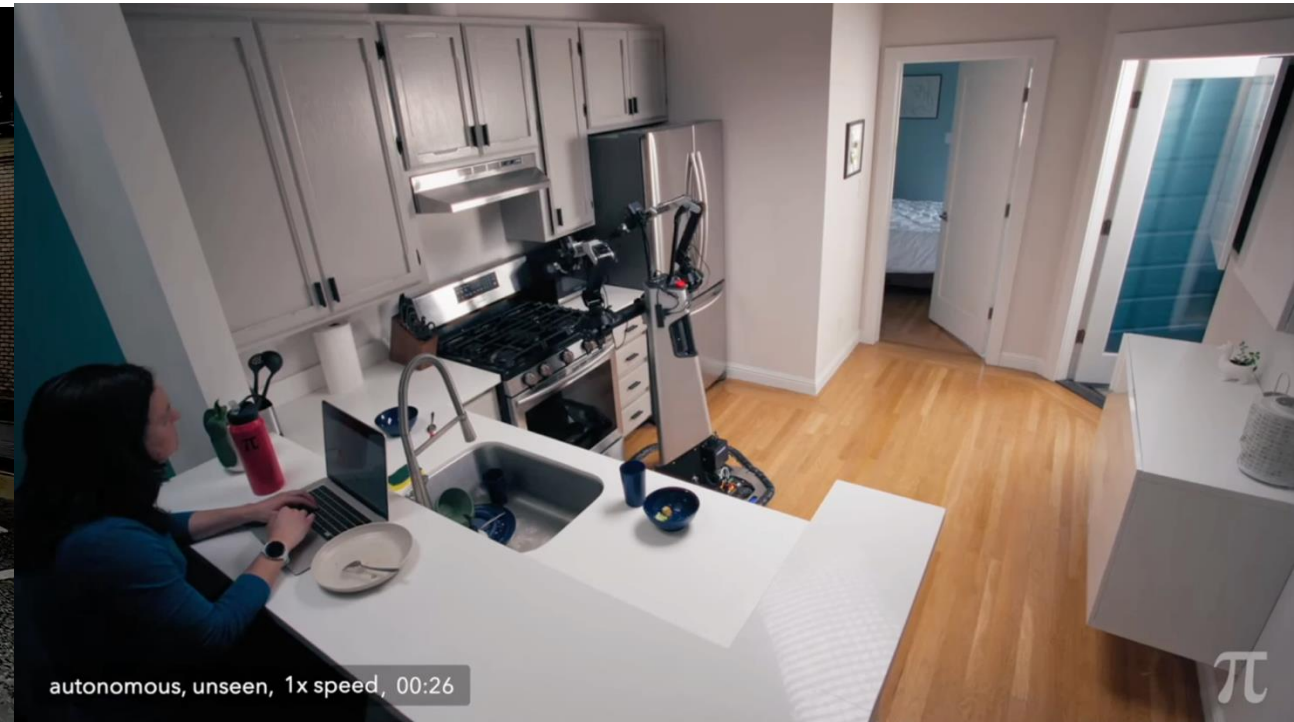
Lecture 12

Safety and Uncertainty

Imagine you are given a robot policy that needs to be deployed.
This policy may have to interact *with* or *in service of* an end-user



Wayve



Physical Intelligence

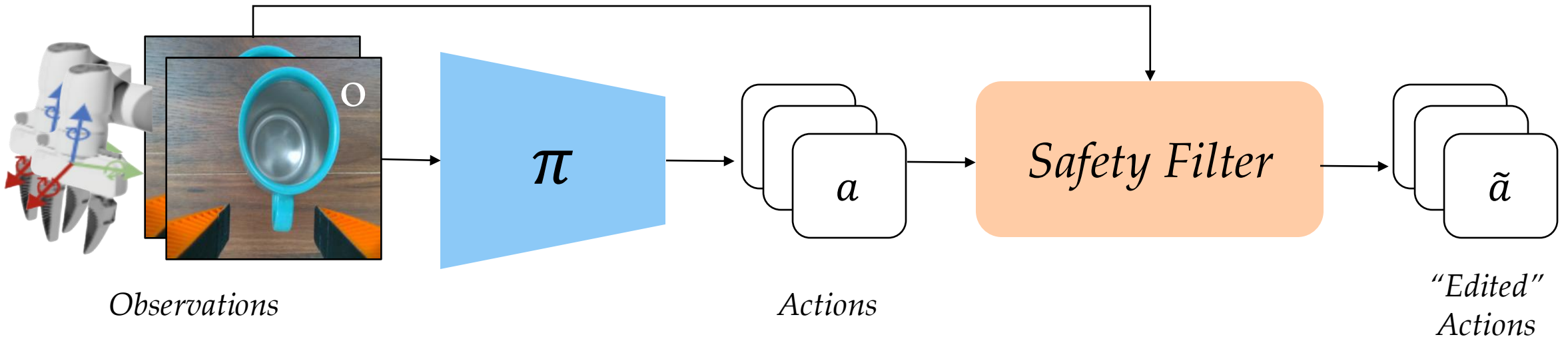
How do we know that the robot will “make safe decisions”?
What does “a safe decision” mean more precisely?

Let's start with some ideas from safe control....

Idea from safe control:

Safety filters modify *any* policy's output so that present actions will not result in future failure

Aside: In LLM land this is called an “output guardrail” ☺





The Four Ingredients for Safety

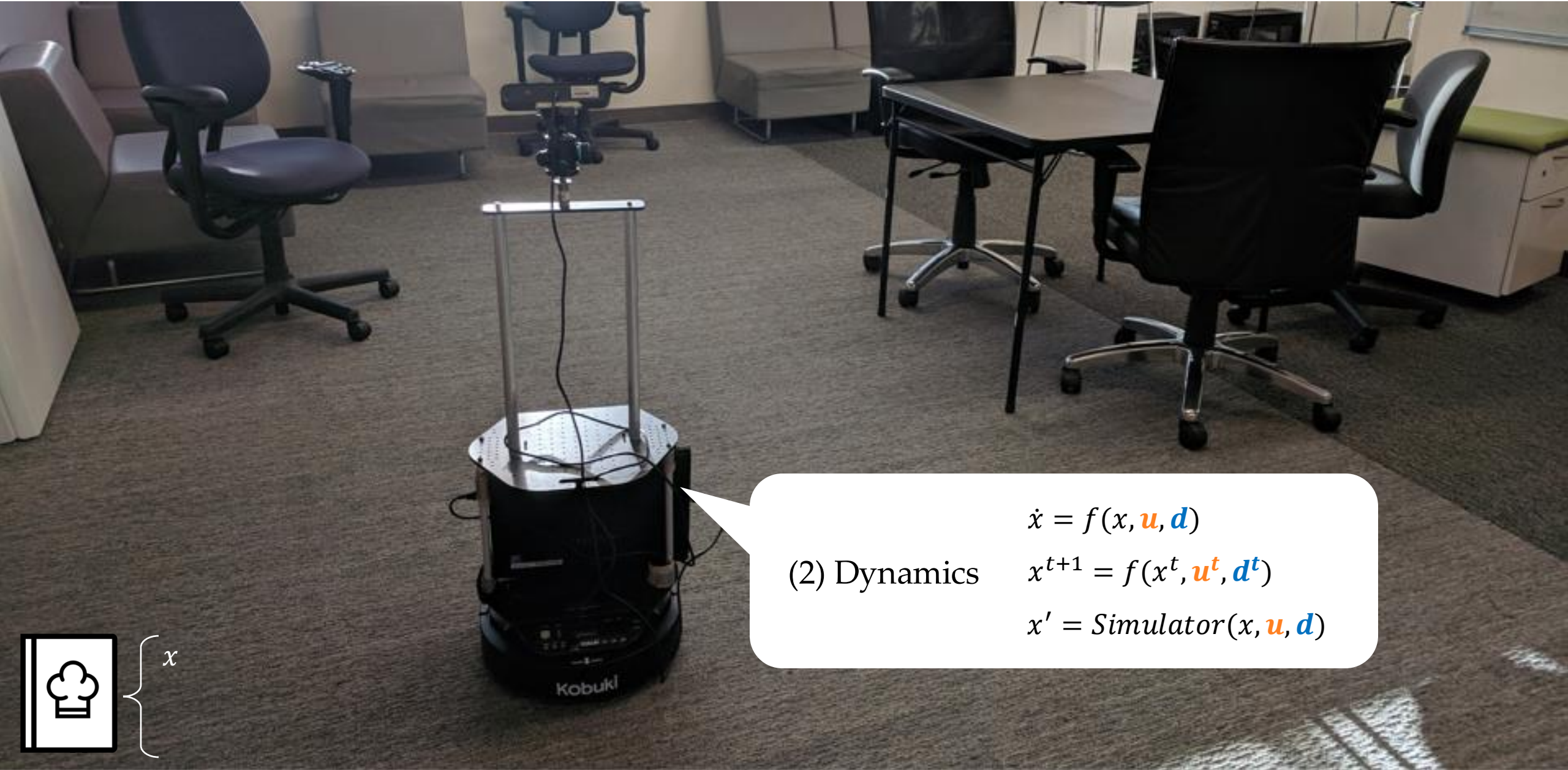


(1) State Space $x \in \mathbb{R}^4$

xy-position, velocity, heading



The Four Ingredients for Safety



(2) Dynamics

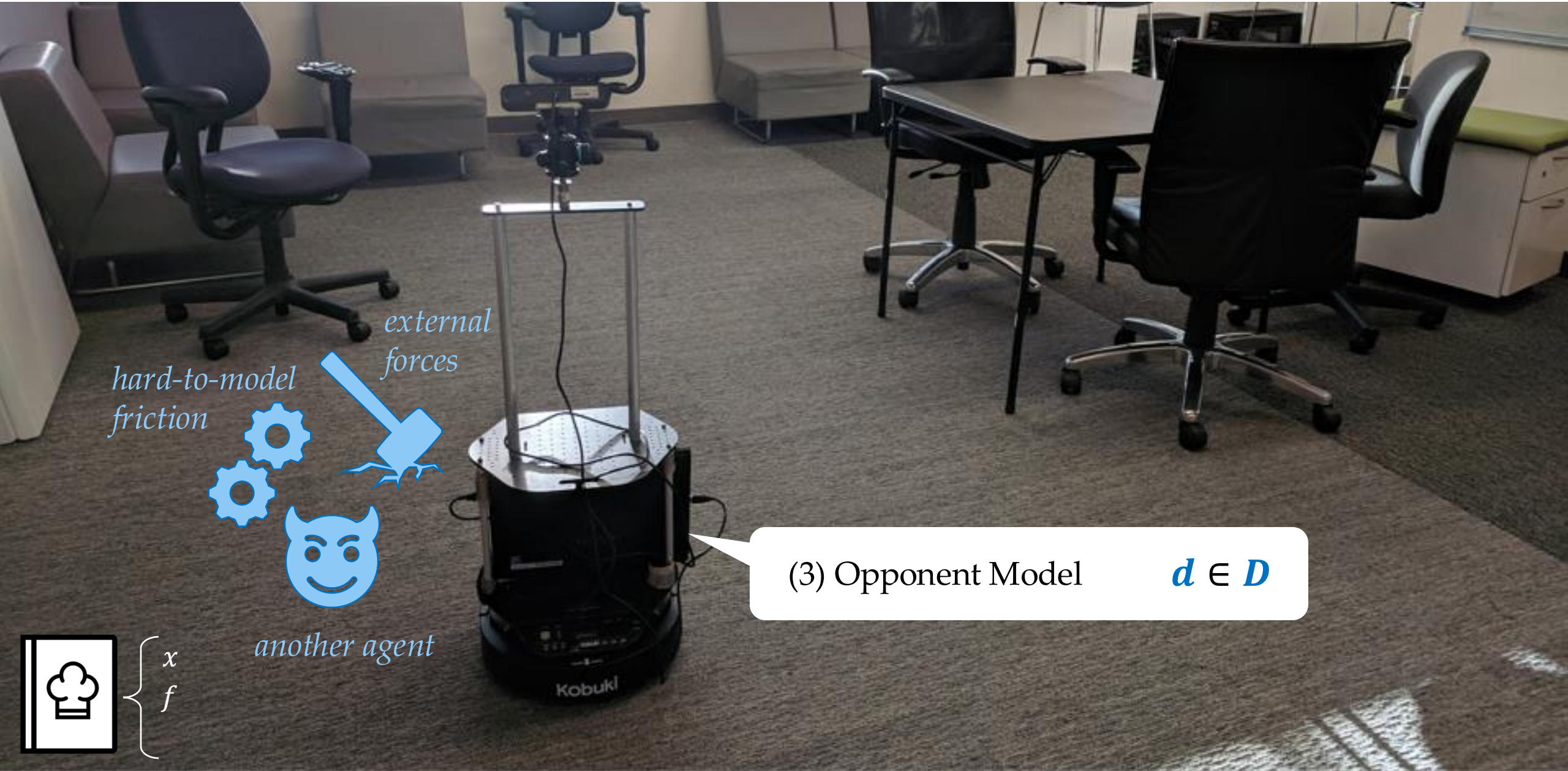
$$\dot{x} = f(x, \mathbf{u}, \mathbf{d})$$

$$x^{t+1} = f(x^t, \mathbf{u}^t, \mathbf{d}^t)$$

$$x' = \text{Simulator}(x, \mathbf{u}, \mathbf{d})$$



The Four Ingredients for Safety



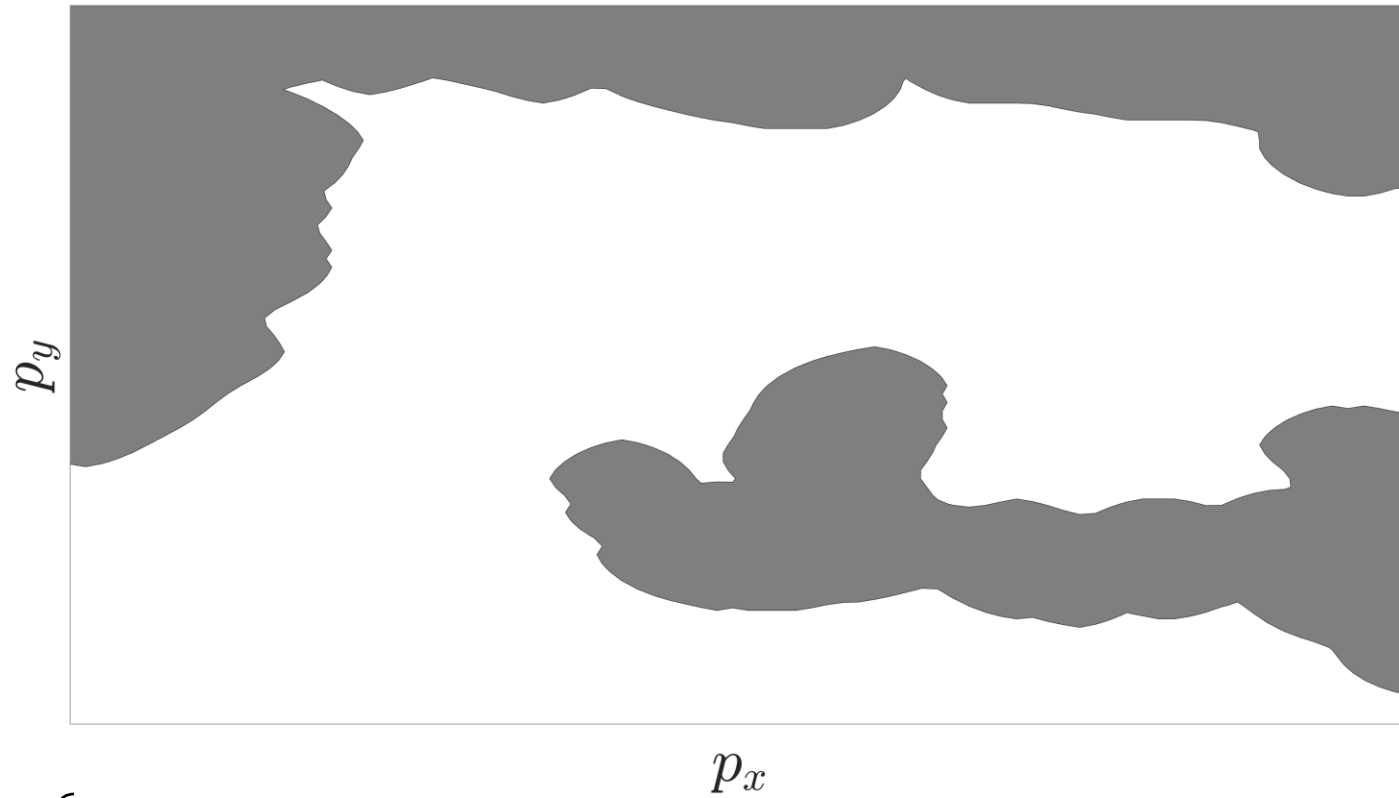
The Four Ingredients for Safety

(4) Failure Set $\mathcal{F} \subset X$



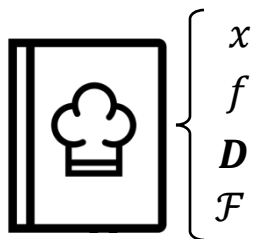
Note: now we are dealing with *constraints* rather than *rewards*

Let's cook up a safety filter!

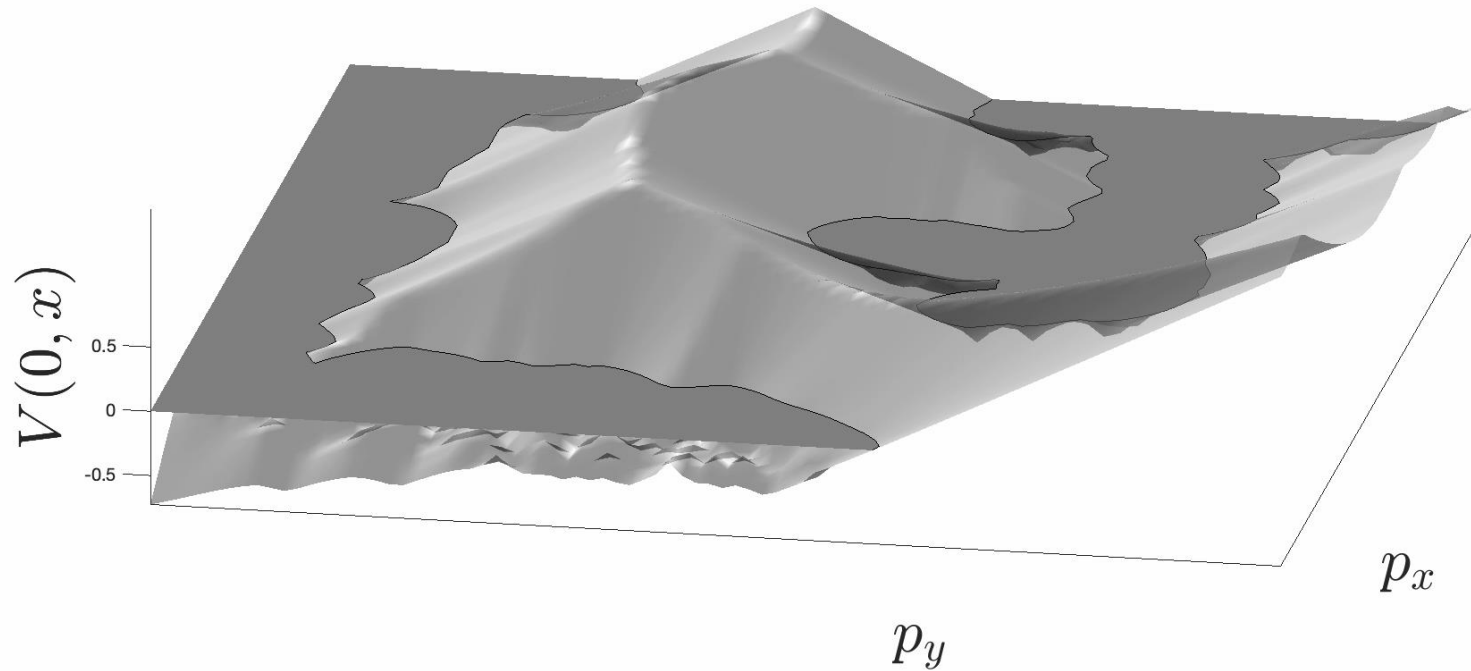


Failure Set
(i.e., our safety specification)

$$\mathcal{F} \subset \mathcal{X}$$

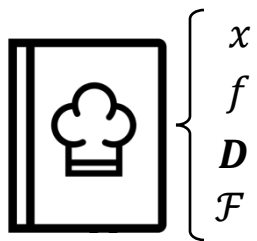


Let's cook up a safety filter!

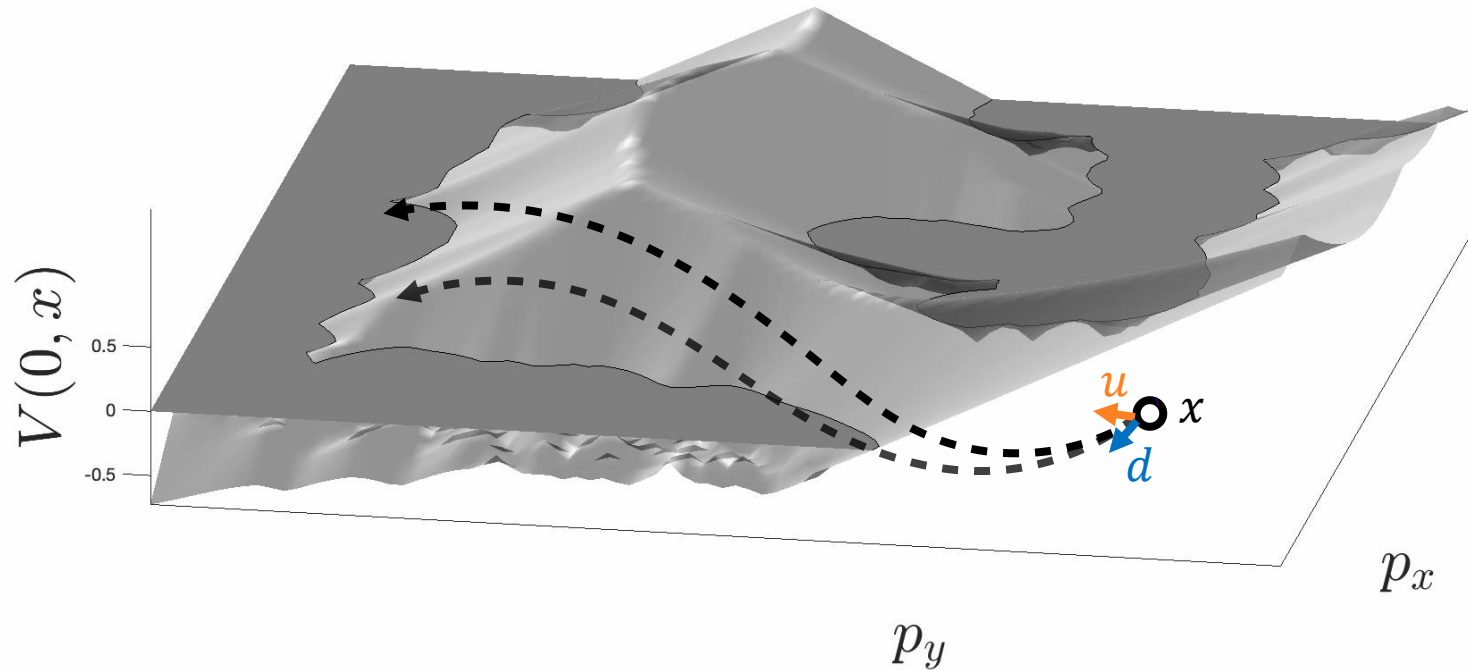


Encode Failure Set

$$\mathcal{F} = \{x: \ell(x) \leq 0\}$$



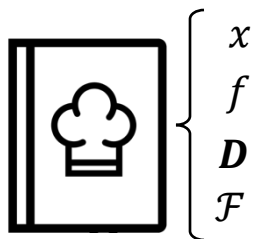
Let's cook up a safety filter!



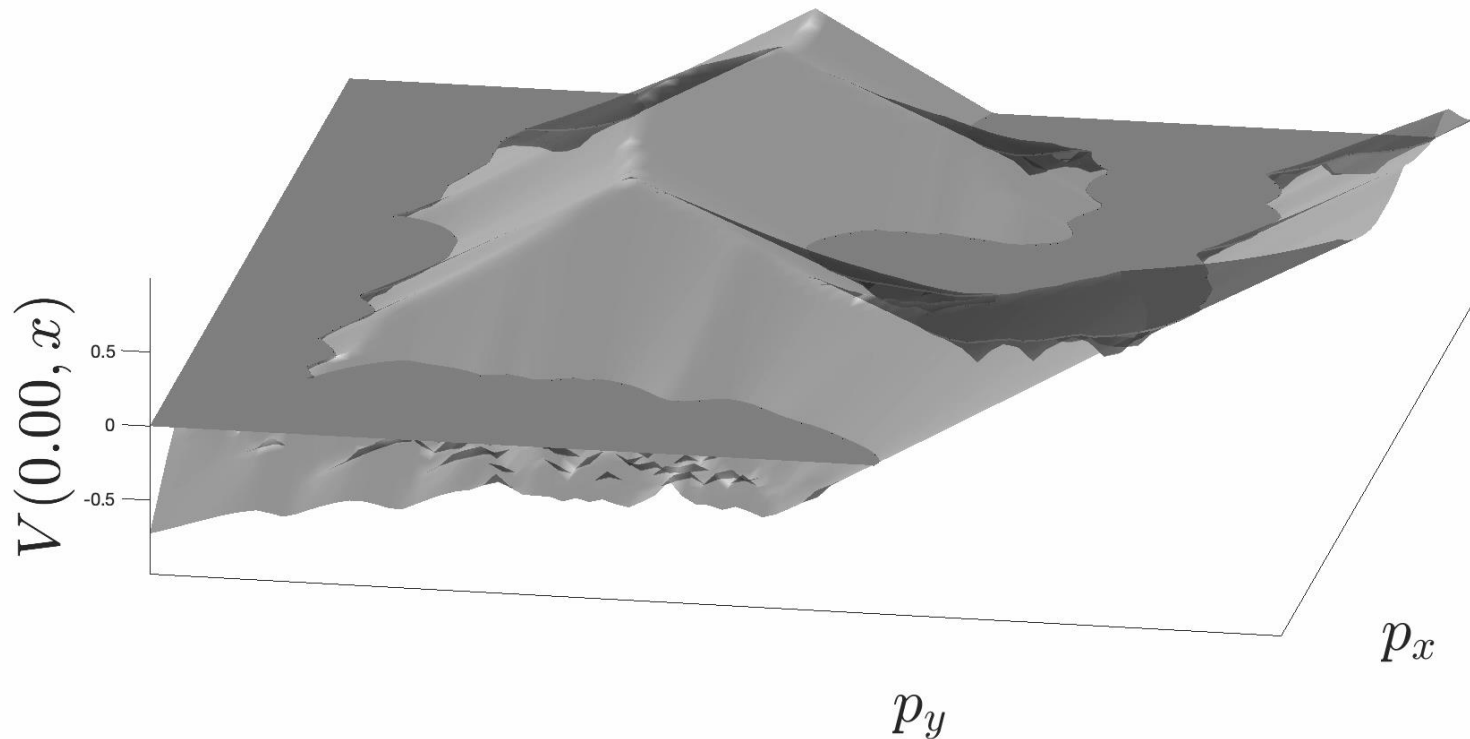
Pose Safety Critical Game

$$V(x) := \max_{\pi_u} \min_{\pi_d} \left(\min_{t \geq 0} \ell(\zeta_x^{u,d}(t)) \right)$$

V “remembers” the closest system got to failure under best robot strategy π_u and worst opponent strategy π_d



Let's cook up a safety filter!



Solve Safety Game

$$V(x) := \max_{\pi_u} \min_{\pi_d} \left(\min_{t \geq 0} \ell(\zeta_x^{u,d}(t)) \right)$$

*Many solvers: exact grid-based PDE solvers [1],
adversarial RL [2,3], self-supervised learning [4]*

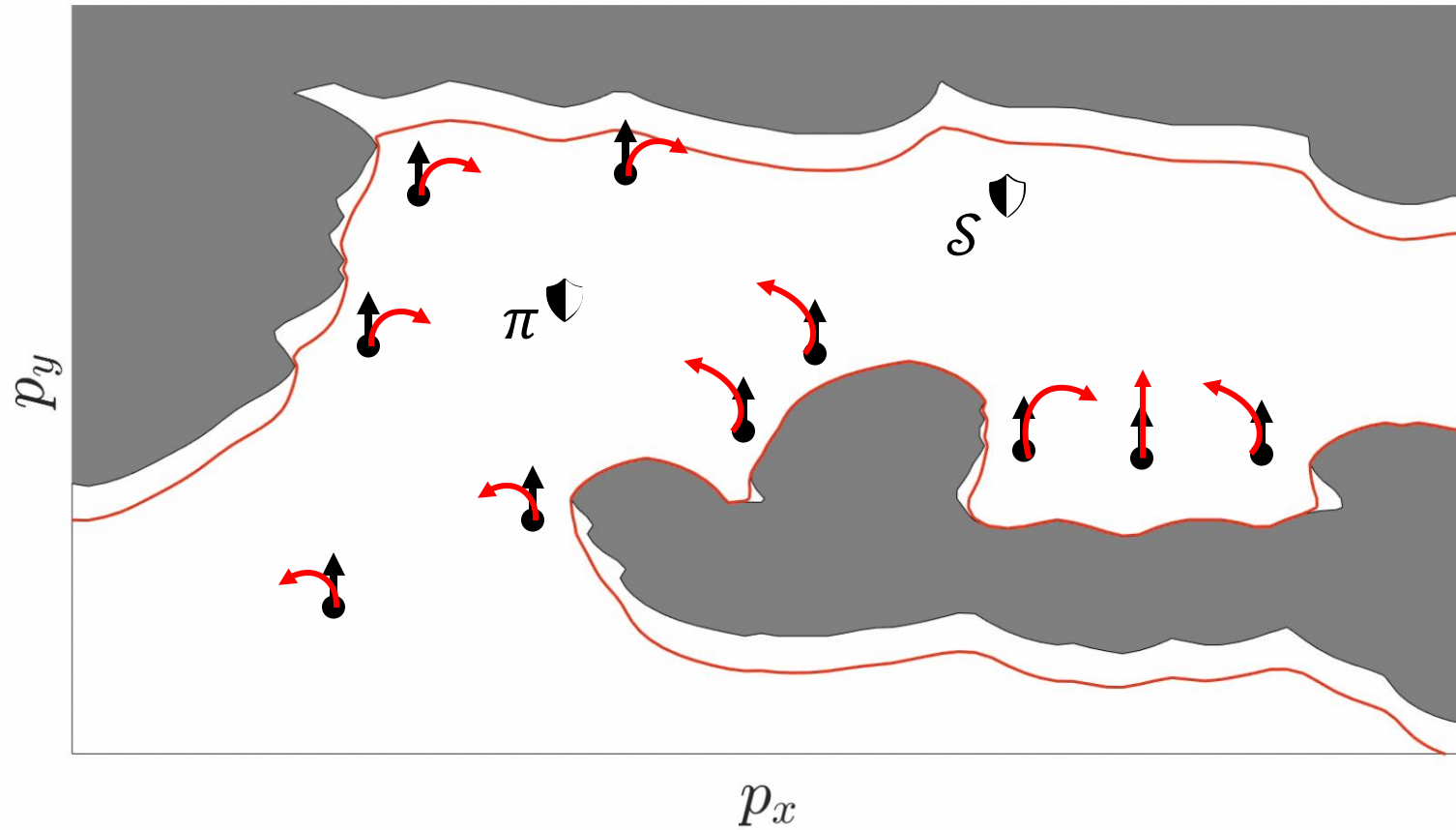
[1] Mitchell, Journal of Scientific Computing 2008

[2] Pinto, et al. ICML 2017

[3] Hsu, et al. L4DC 2023

[4] Bansal & Tomlin, ICRA 2021

Let's cook up a safety filter!

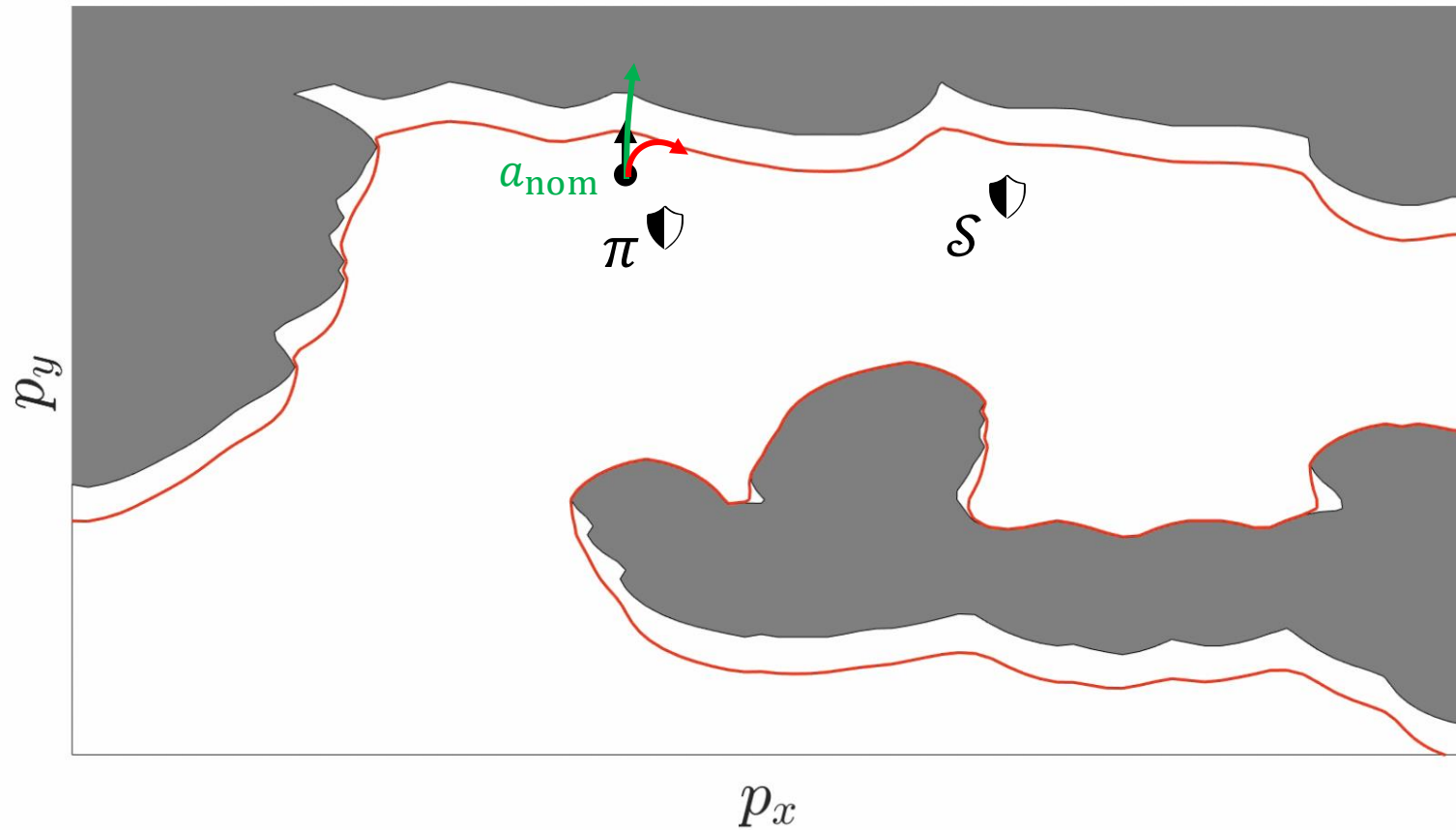


Safety Policy

$$\pi^{\text{shield}}, \quad \mathcal{S}^{\text{shield}} = \{x : V(x) > 0\}$$

Safe Set (i.e., "Monitor")

Let's cook up a safety filter!



Safety Policy

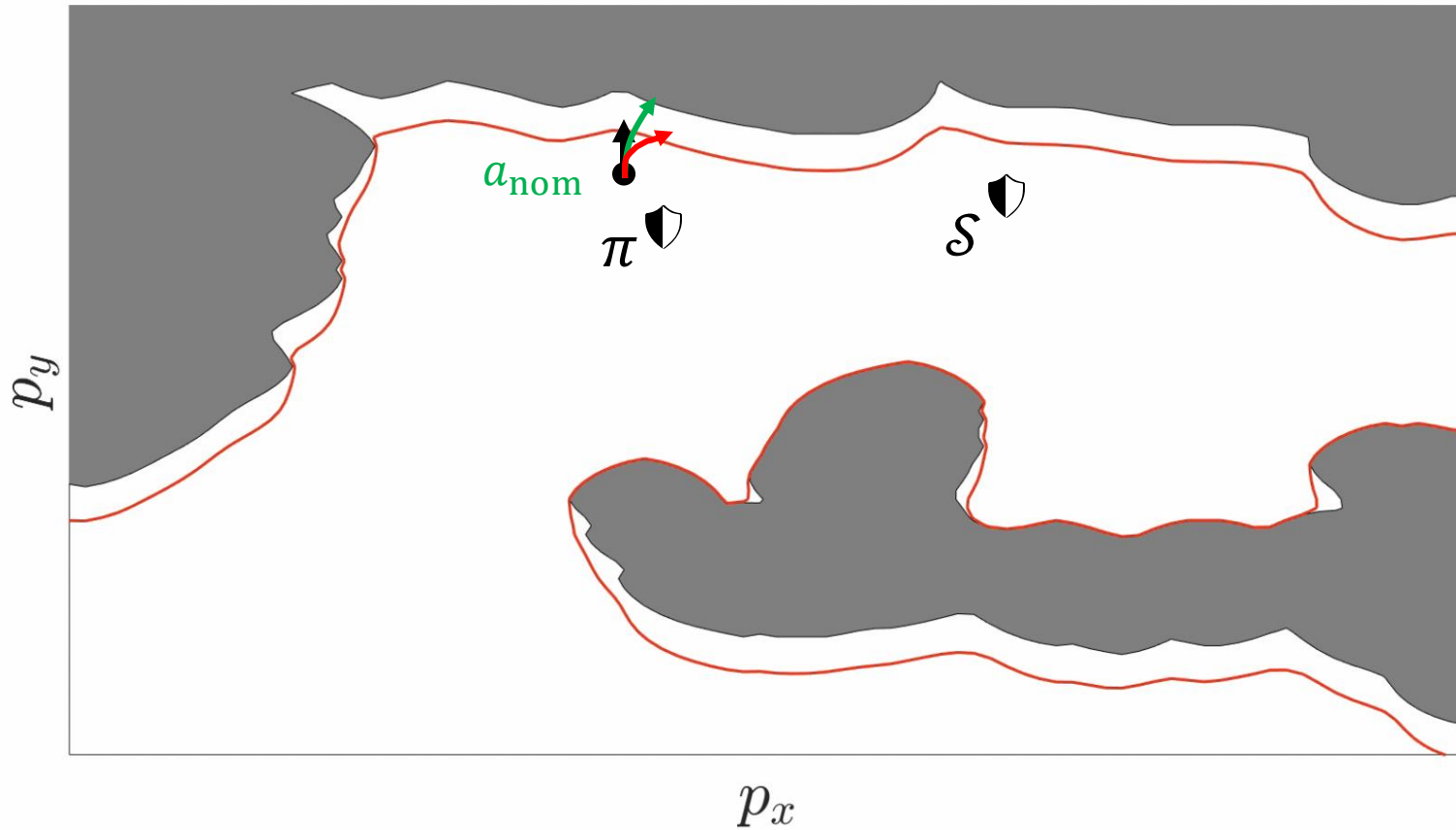
$$\pi^{\text{shield}}, \quad \mathcal{S}^{\text{shield}} = \{x : V(x) > 0\}$$

Safe Set (i.e., "Monitor")

Safety Filter

$$a^* = \begin{cases} \pi^{\text{shield}}, & x \text{ near bdry } \mathcal{S}^{\text{shield}} \\ \text{[any policy here]}, & x \in \mathcal{S}^{\text{shield}} \end{cases}$$

Let's cook up a safety filter!



Safety Policy

$$\pi^{\bullet}, \quad S^{\bullet} = \{x : V(x) > 0\}$$

Safe Set (i.e., "Monitor")

Safety Filter

$$a^* = \arg \min_a ||a - a_{\text{nom}}||_2^2$$

$$\text{s.t. } V(x(t + \delta)) \geq 0$$

**Note: there are many filtering variants!*

[Wabersich, et al. "Data-driven safety filters." Control Systems Magazine, 2023]

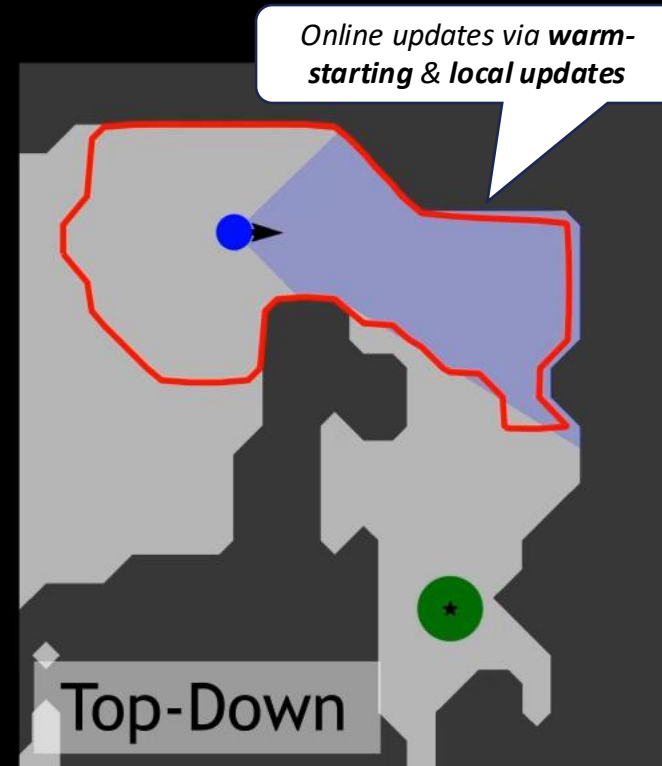
[Hsu, et al. "The Safety Filter." Annual Review of Control, Robotics, and Autonomous Systems, 2023]

"Find an action that is similar to the base policy as long as the next state is still safe"

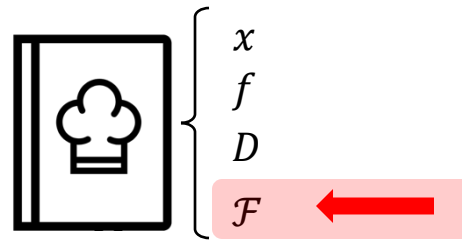
Vision-Based Robot *Without* Safety Strategy



Robot *With* Safety Filter Updated Online



What “matters” about safe robot behavior
can be hard to specify ...



So far, the safety representations we have seen are....



$$\mathcal{F} = \{x : \| x_R - x_H \|_2 \leq \epsilon\}$$



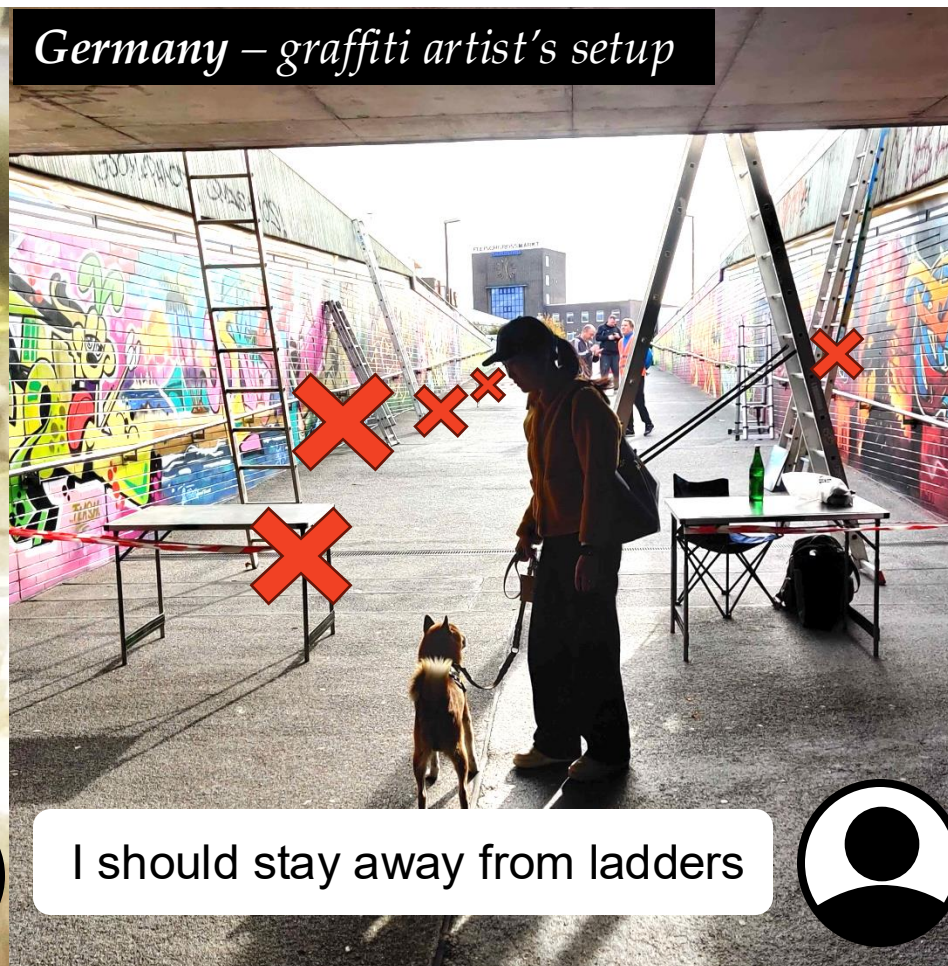
$$\mathcal{F} = \{x : \| x - \text{SLAM}(x) \|_2 \leq \epsilon\}$$

But in the open world, there are many more constraints....

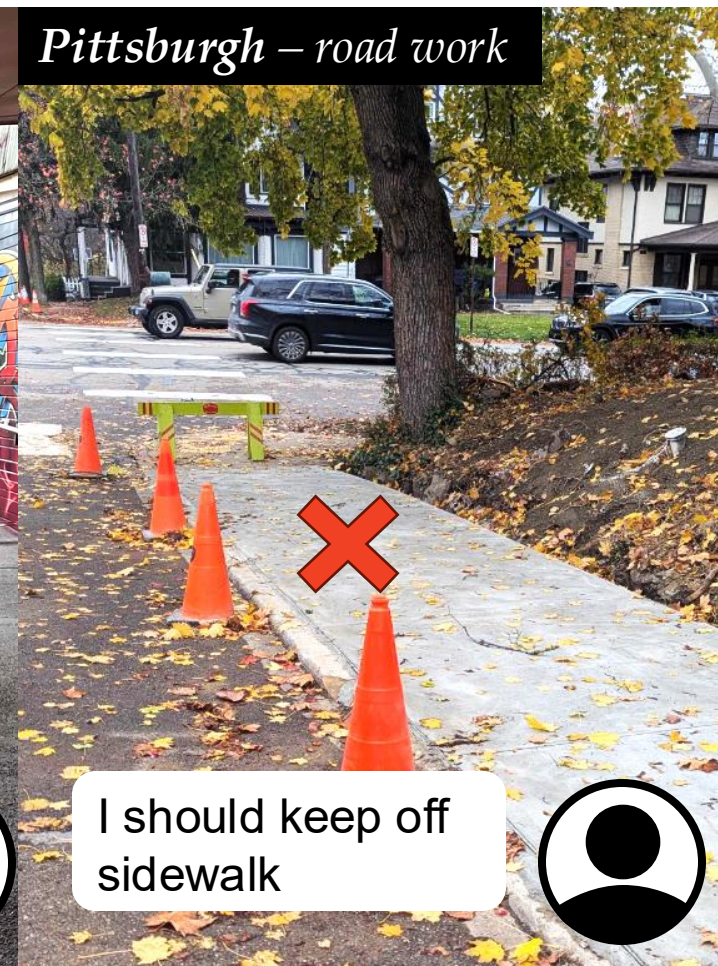
Brazil – caution tape



Germany – graffiti artist's setup



Pittsburgh – road work



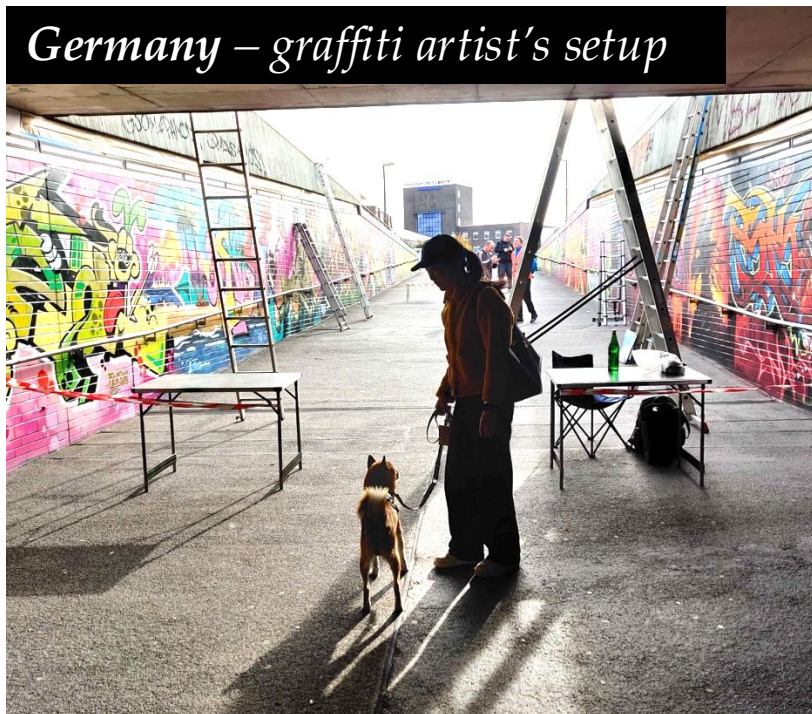
Real images taken by my students!

But in the open world, there are many more constraints....

Brazil – caution tape



Germany – graffiti artist's setup



Pittsburgh – road work



Spills



Accidents



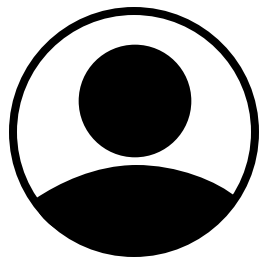
Fragile objects



Sensitive personal areas

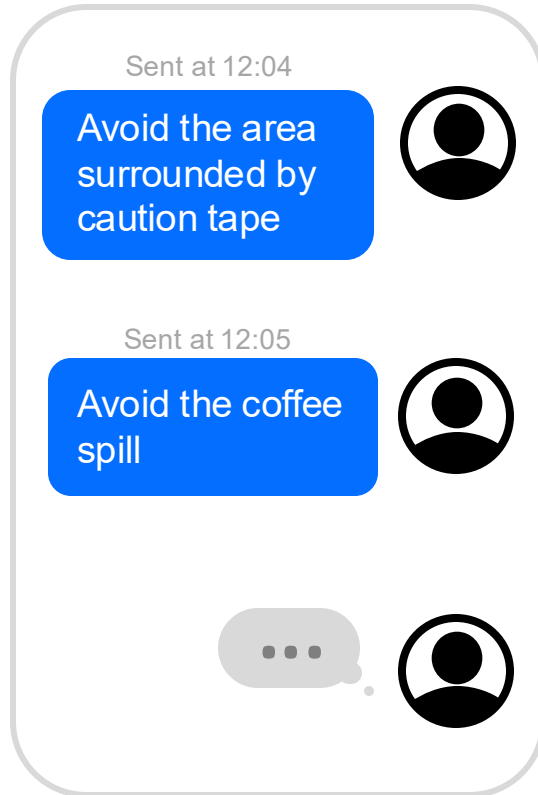


MPPI



How can robots refine their safety representations to include **semantically-meaningful safety constraints**?

Language Feedback



Idea:

Vision-language models enable a flexible way to communicate safety constraints to the robot

Offline



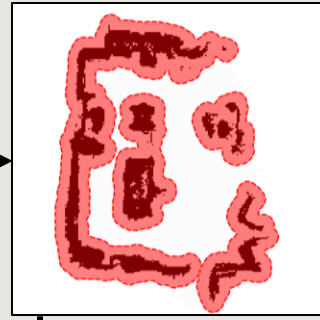
Robot

Failure Set



$$\hat{\mathcal{F}}_E^0$$

Safe Set & Policy



$$\mathcal{S}^{\blacktriangleleft,0}, \pi^{\blacktriangleleft,0}$$

Online



Human



Robot

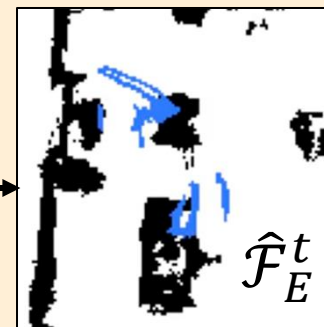
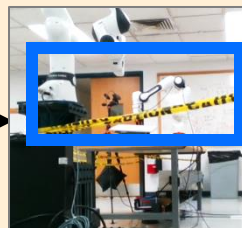
Avoid the area surrounded by caution tape.



RGB

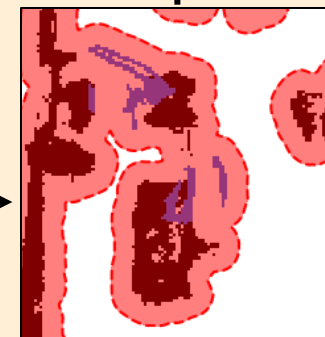
OWL-ViT

Detection



Semantic Failure Set $\hat{\mathcal{F}}_E^t$

Warm-Start
HJ Reachability



Updated Safe
Set & Policy

π_R^*

Safety
Filter

$\mathcal{S}^{h,t}, \pi^{h,t}$

From the human's POV...

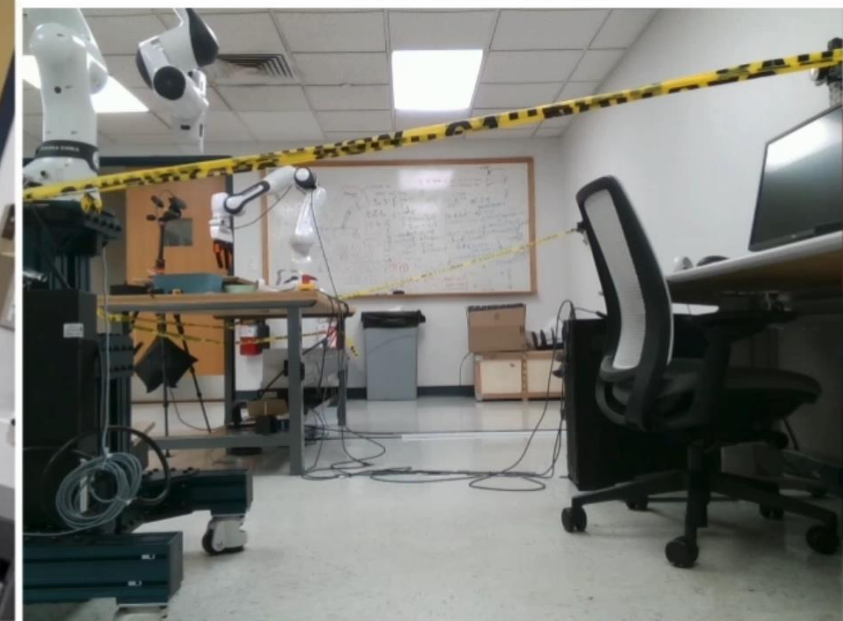
Language Feedback



From the robot's POV...



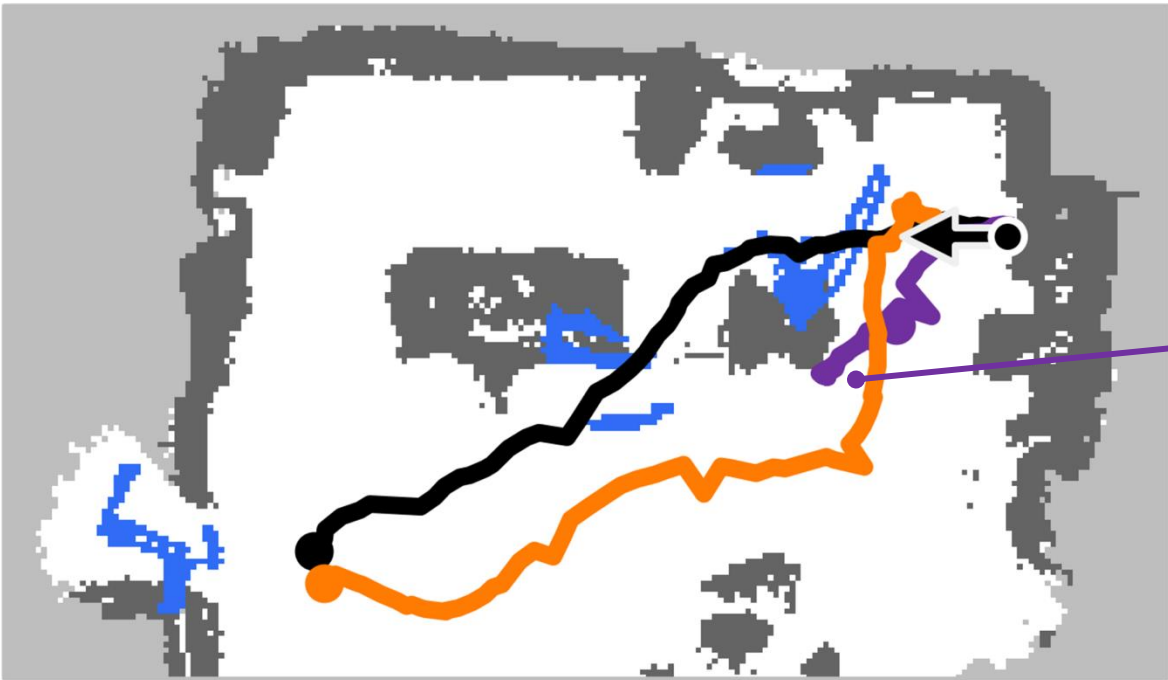
VLM Detections



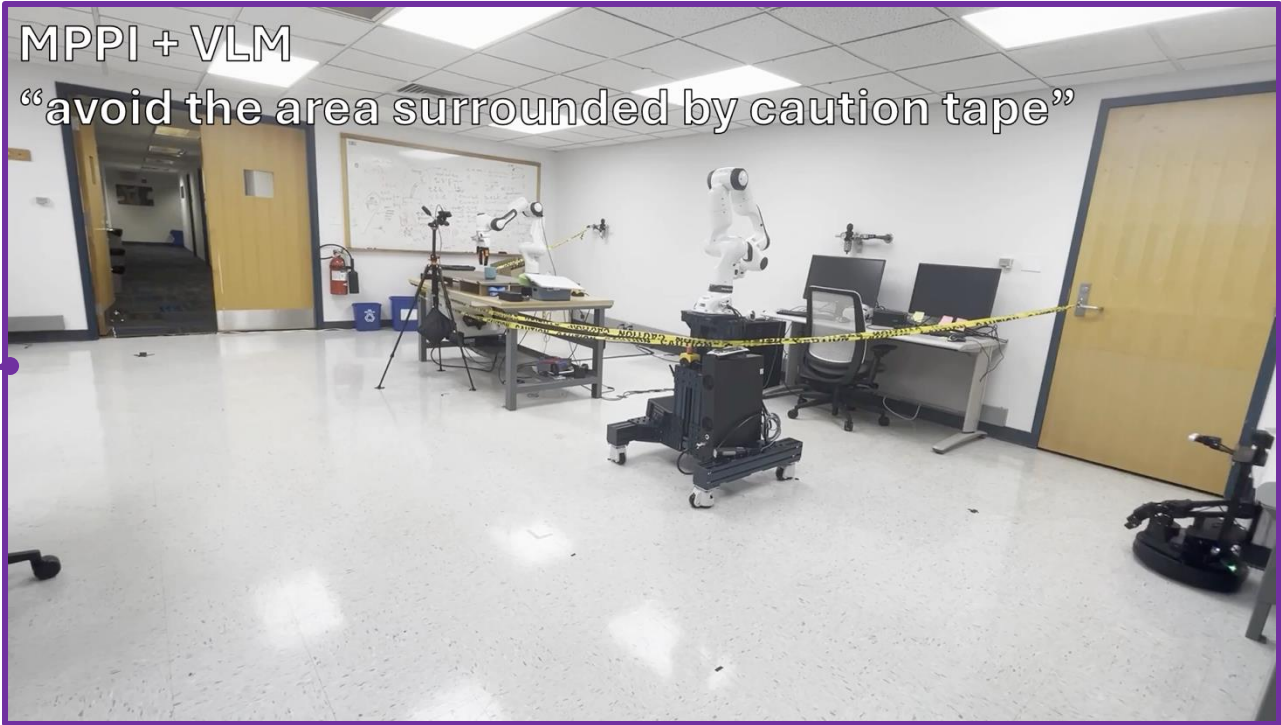


“avoid the dog toys and the laundry”

Quantitative & Qualitative Hardware Results



 Semantic Fail  Plan-SLAM  Plan-Lang  Safe-Lang



Method	Caution Tape Scenario			
	Plan Time (ms)	t -to-Goal	Abides \mathcal{F}_E^*	$\pi_{\mathcal{R}}^{\heartsuit}$ On
Plan-SLAM	7 (± 1)	17.647	X	N/A
Plan-Lang	40 (± 30)	∞	X	N/A
Safe-Lang	31 (± 29)	26.176	✓	29.37%

All modules run asynchronously

So far, we have been safeguarding the robot by *switching*.....

Safety Filter

$$a^* = \begin{cases} \pi^{\text{safe}}, & x \text{ near bdry } \mathcal{S}^{\text{safe}} \\ \text{[any policy here]}, & x \in \mathcal{S}^{\text{safe}} \end{cases}$$

Can we do better?

Safety with Agency: Human-Centered Safety Filter with Application to AI-Assisted Motorsports

Donggeon David Oh*, Justin Lidard*, Haimin Hu, Himani Sinhmar, Elle Lazarski, Deepak Gopinath,
Emily Sumner, Jonathan DeCastro, Guy Rosman, Naomi Leonard, Jaime Fernández Fisac

Safety Filter (before)

$$a^* = \begin{cases} \pi^\downarrow, & V^\downarrow(x) \approx 0 \\ \text{[any policy here]}, & V^\downarrow(x) > 0 \end{cases}$$

Safety Filter (now)

$$\begin{aligned} \mathbf{a}^* &= \arg \min_{a \in A} ||a - \mathbf{a}_{\text{nom}}||_2^2 \\ \text{s. t. } &Q^\downarrow(x, a) \geq \alpha V^\downarrow(x) \end{aligned}$$

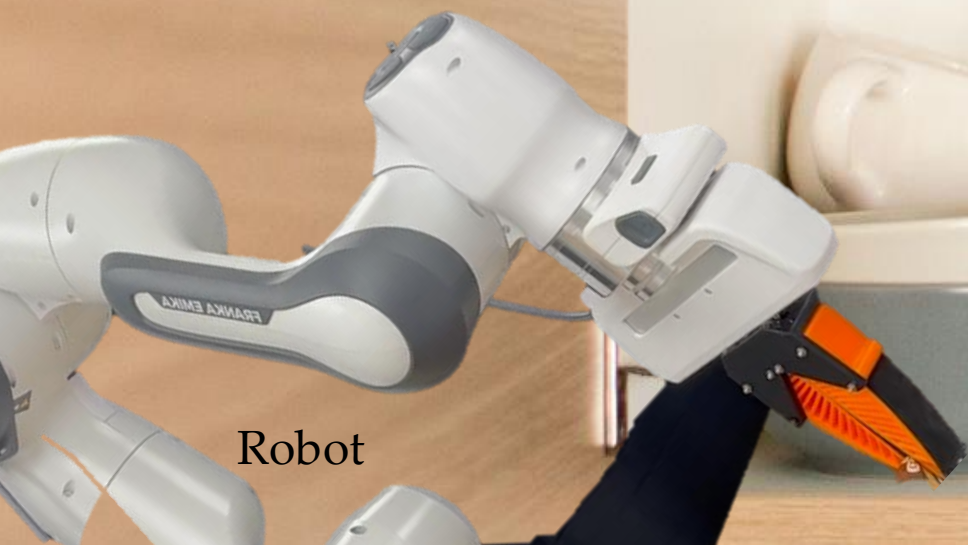
"Find an action that is similar to the base policy as long as the next state is still safe"

$\alpha \in [0,1)$ design parameter controls how quickly the safety value function is allowed to decrease over a single timestep

Can safety filters generalize *beyond physical actions*?



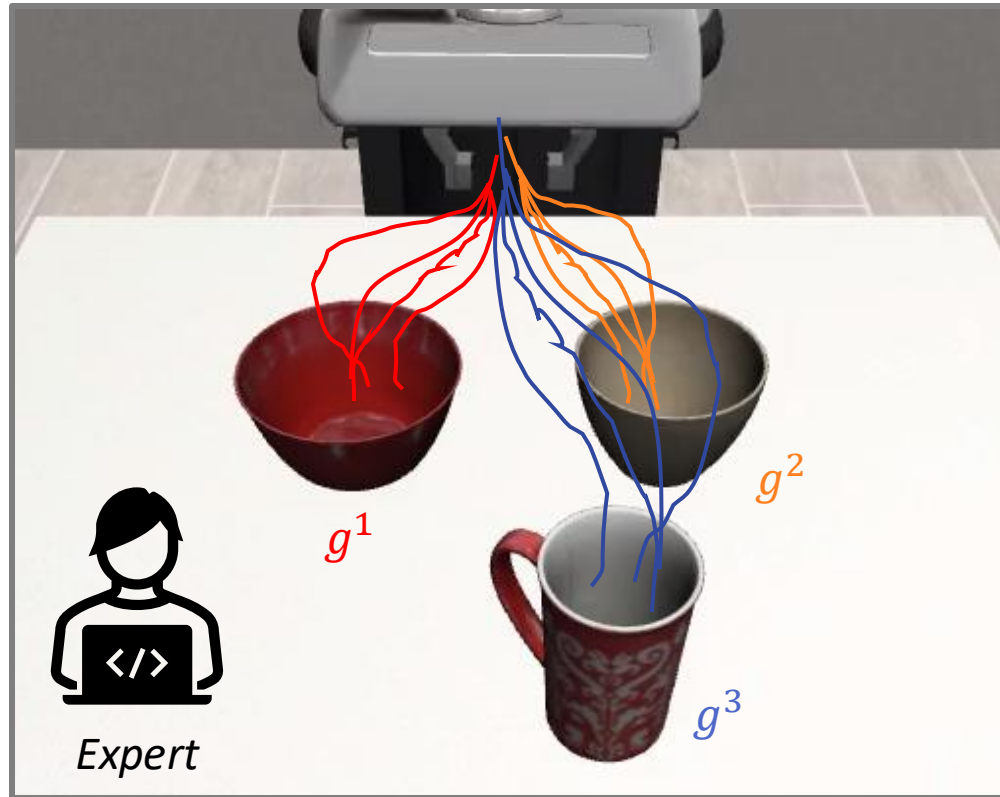
Human User



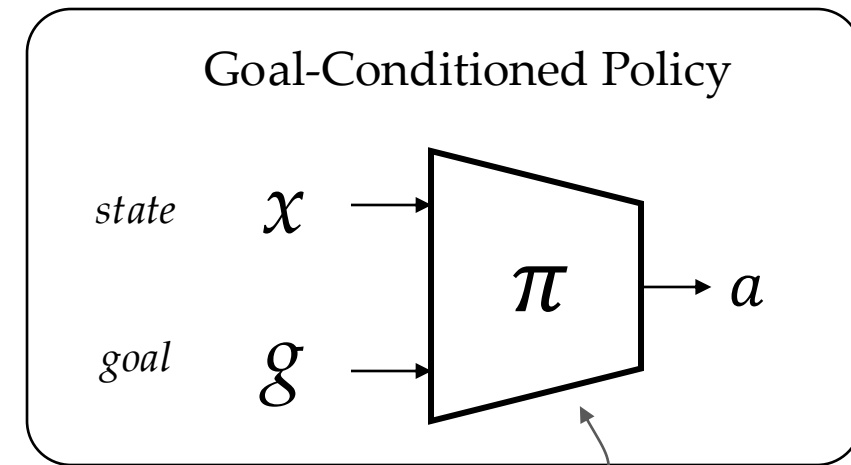
Robot



Goal-conditioned imitation learned policies are useful



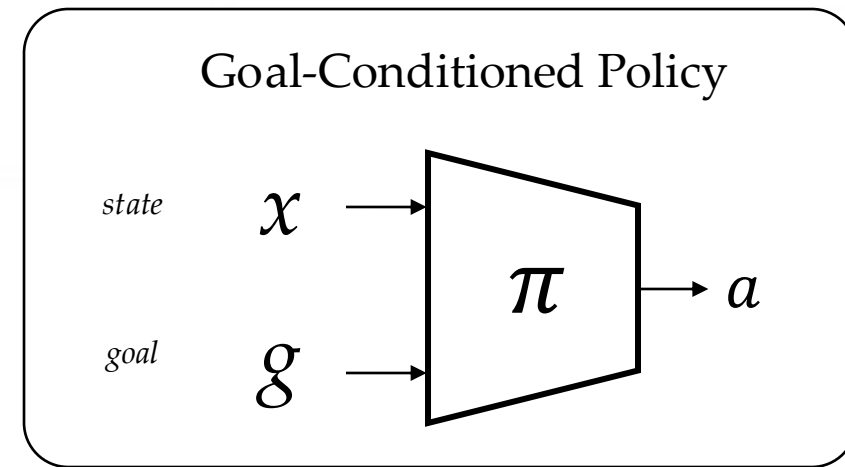
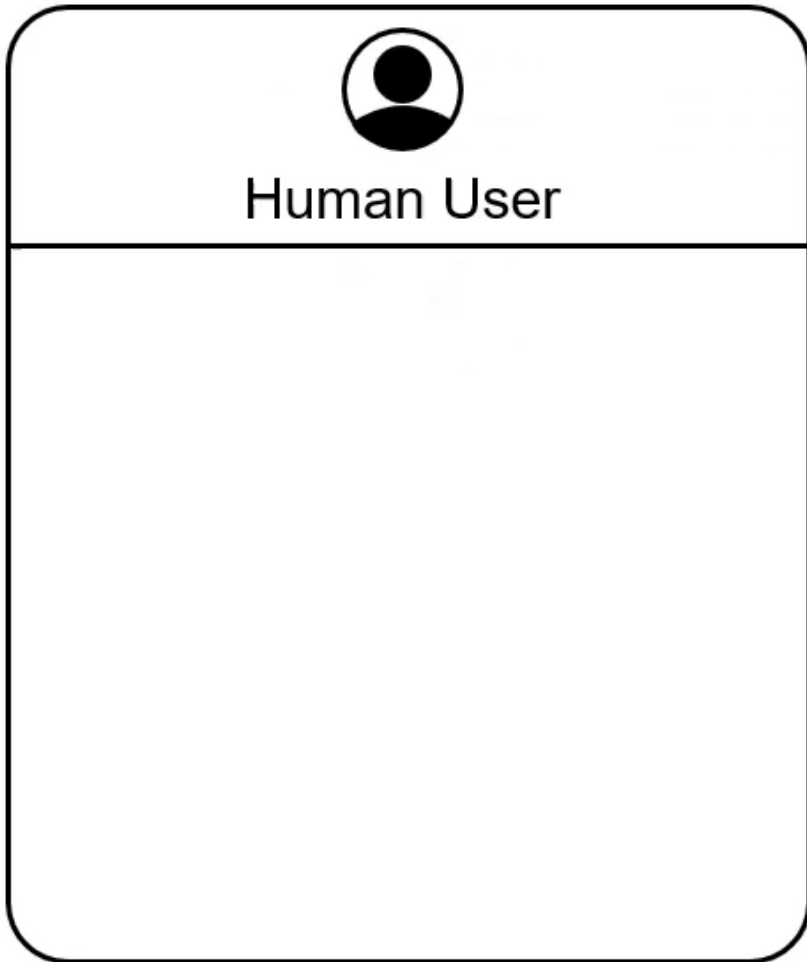
100 successful demos per target object



Training via Behavior Cloning

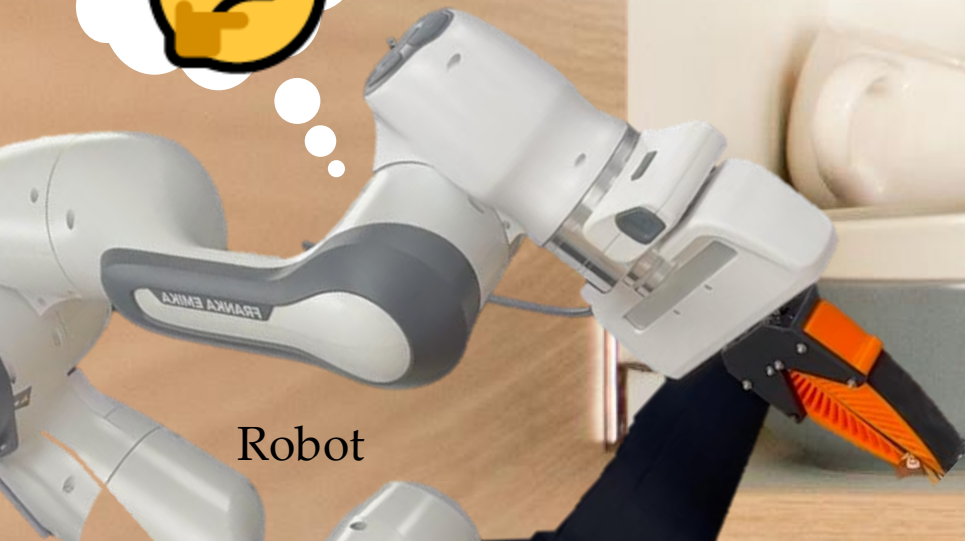
$$\mathcal{L}_{BC}(\mathcal{D}) = \mathbb{E}_{(x^i, a^i, g^i) \sim \mathcal{D}} \left\| \pi(x^i; g^i) - a^i \right\|_2^2$$

Goal-conditioned imitation learned policies are useful
but they aren't guaranteed to safely succeed





Human User

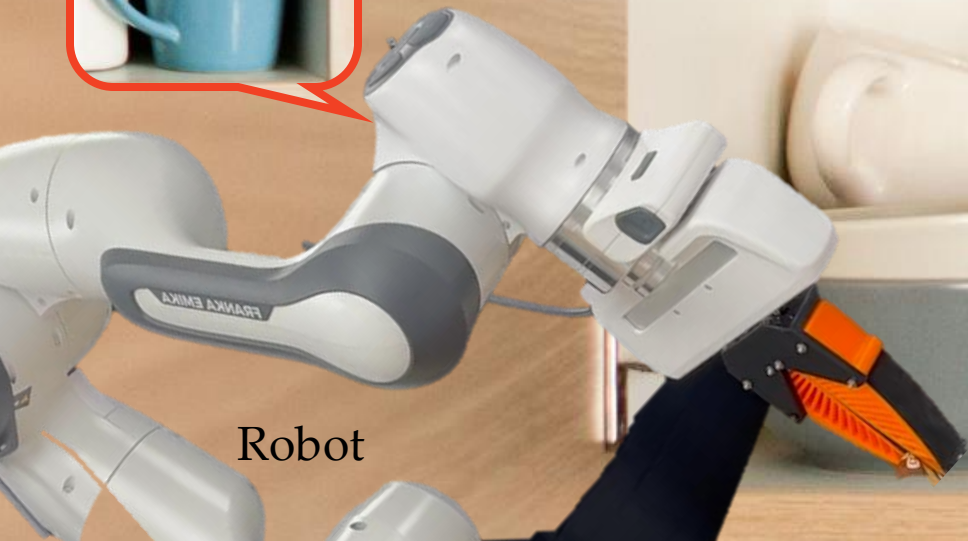


Robot





Human User



Robot



Idea:

alternative suggestion can be modeled as **safety filtering in *goal space***, rather than action space

alternative suggestion can be modeled as **safety filtering in goal space**, rather than action space

Before

$$a^* = \arg \min_{a \in A} ||a - a_{\text{nom}}||_2^2$$

$$\text{s.t. } V(x(t + \delta)) \geq 0$$

"Find an action that is similar to the base policy as long as the next state is still safe"

Ours

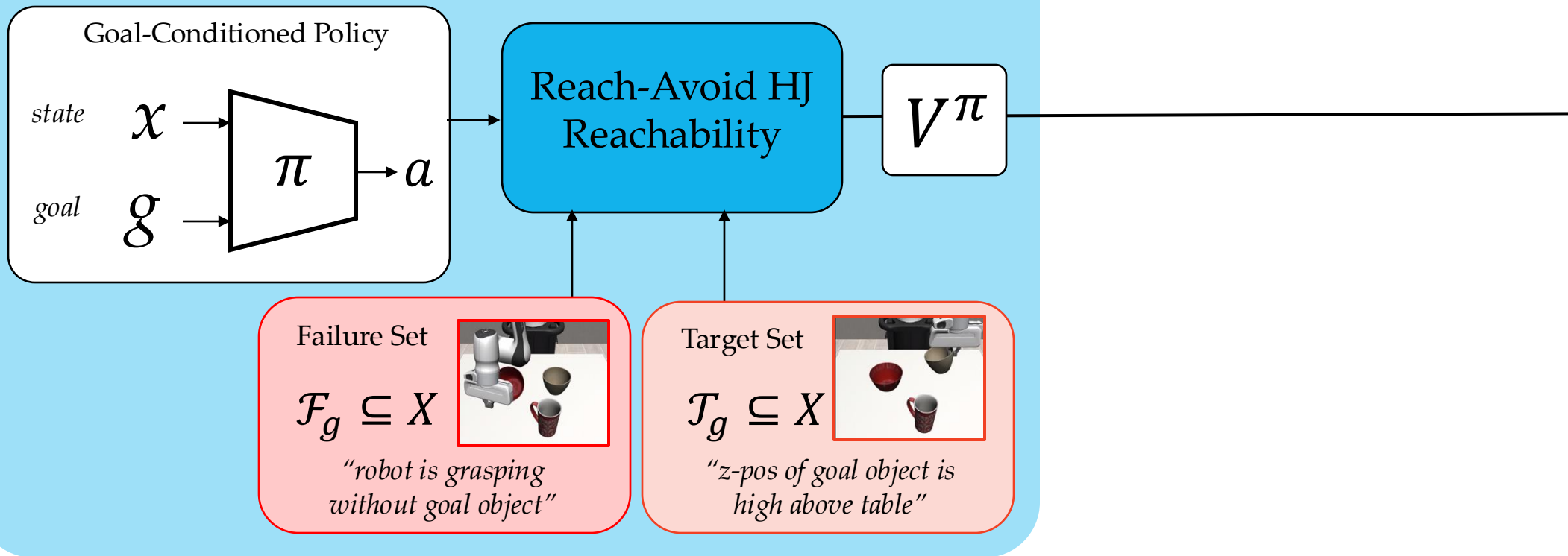
$$g_R = \arg \min_{g \in G} ||g - g_H||_2^2$$

$$\text{s.t. } V^\pi(x; g) \geq 0$$

"Find a goal that is similar to the human's original goal as long as the pre-trained policy is safe"

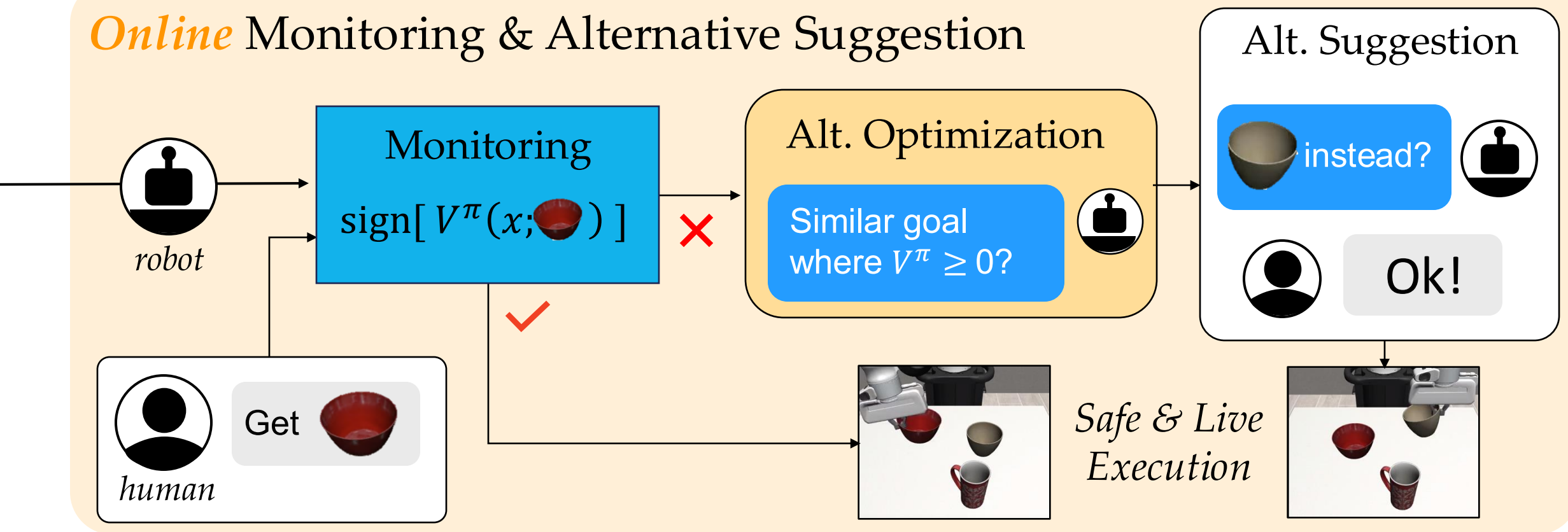
SALT: safety filter goals to suggest Safe ALTERNatives

Offline Safety Analysis



SALT: safety filter goals to suggest Safe ALternatives

Online Monitoring & Alternative Suggestion



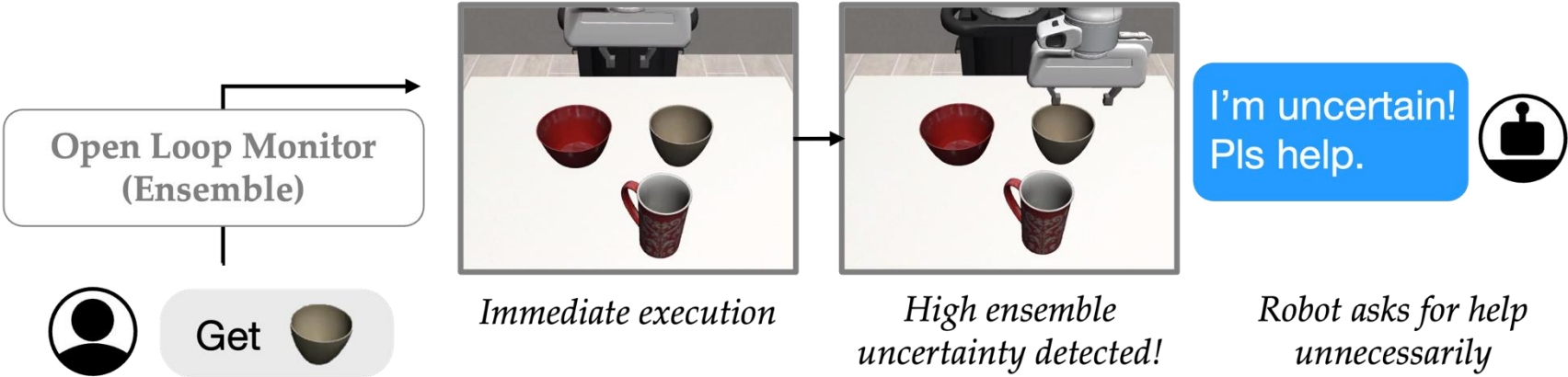
SALT: safety filter goals to suggest Safe ALTERNatives



Human User

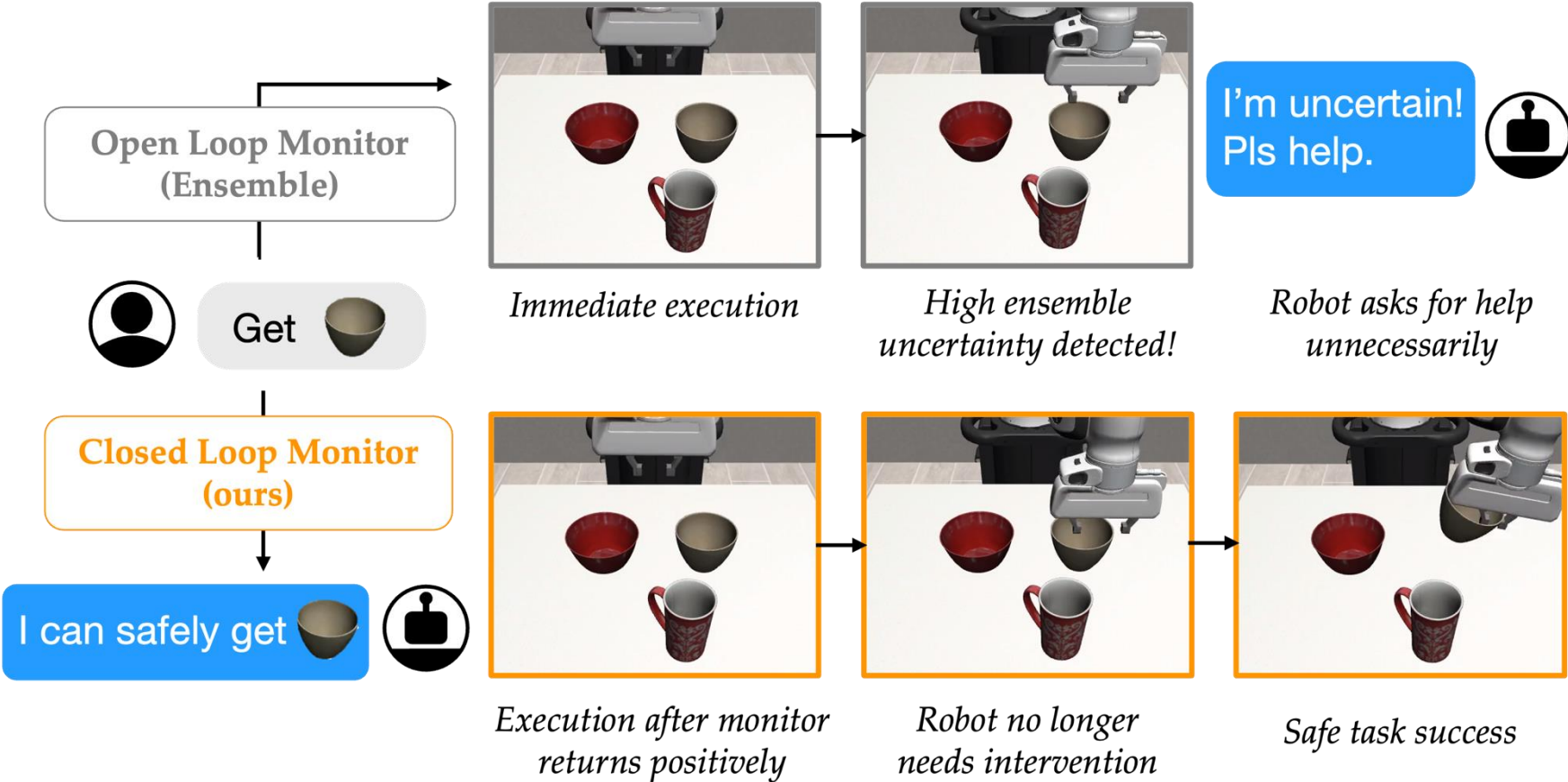
What is the Benefit of SALT as a Runtime Monitor?

	Method	Manipulation			
		TNR % (↑)	TPR % (↑)	FPR % (↓)	FNR % (↓)
Open-Loop →	Ensemble	34.61 (±1.68)	27.45 (±1.96)	32.65 (±1.27)	5.26 (±0.89)
Closed-Loop {	RewardSum	64.47 (±1.61)	4.67 (±0.68)	8.85 (±1.01)	22.01 (±1.21)
	SALT (ours)	61.21 (±1.91)	13.23 (±1.56)	13.23 (±1.23)	12.33 (±0.85)

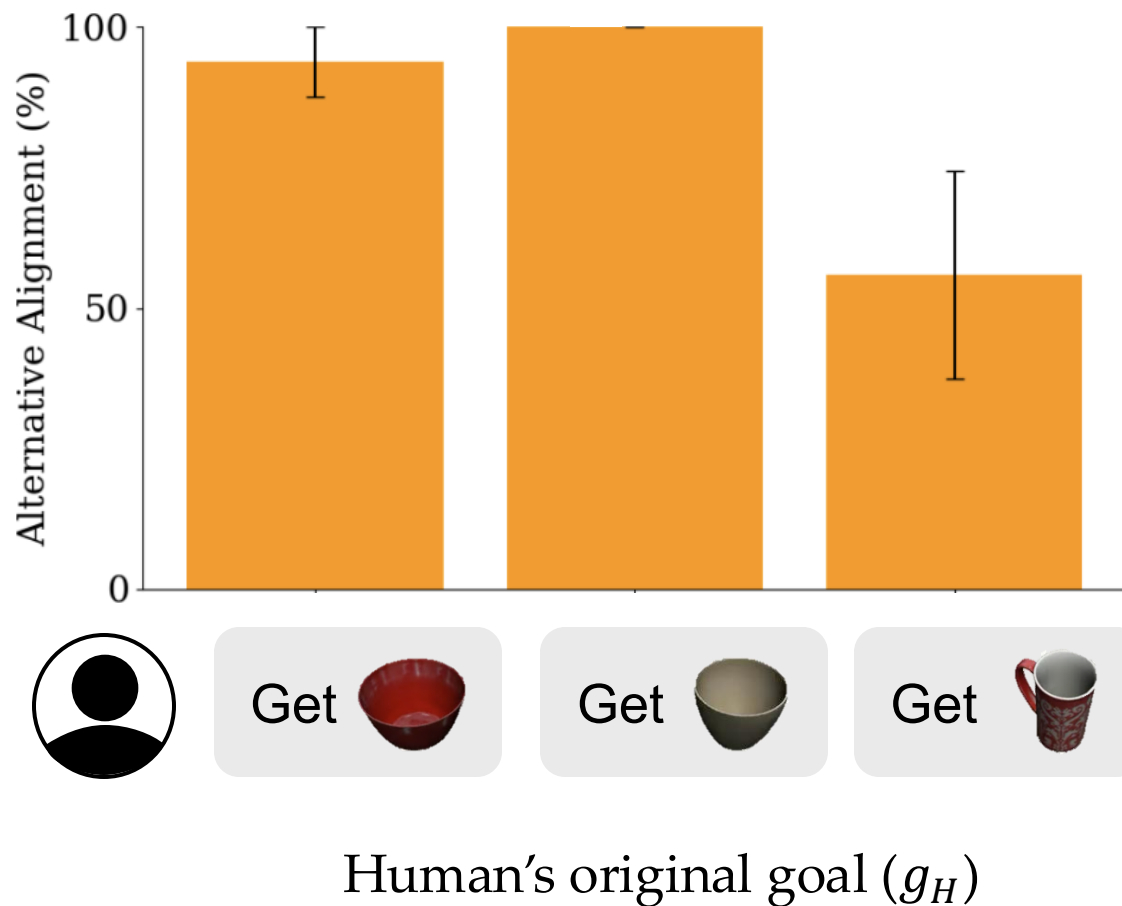


What is the Benefit of SALT as a Runtime Monitor?

Method	Manipulation			
	TNR % (↑)	TPR % (↑)	FPR % (↓)	FNR % (↓)
Open-Loop → Ensemble	34.61 (±1.68)	27.45 (±1.96)	32.65 (±1.27)	5.26 (±0.89)
Closed-Loop {	RewardSum	64.47 (±1.61)	4.67 (±0.68)	22.01 (±1.21)
	SALT (ours)	61.21 (±1.91)	13.23 (±1.56)	12.33 (±0.85)



How Acceptable Are The Proposed Alternatives?



10 expert users from labs at
CMU and UC San Diego.

1,000 random initial conditions x^0 for
each g_H and g_R suggested by SALT

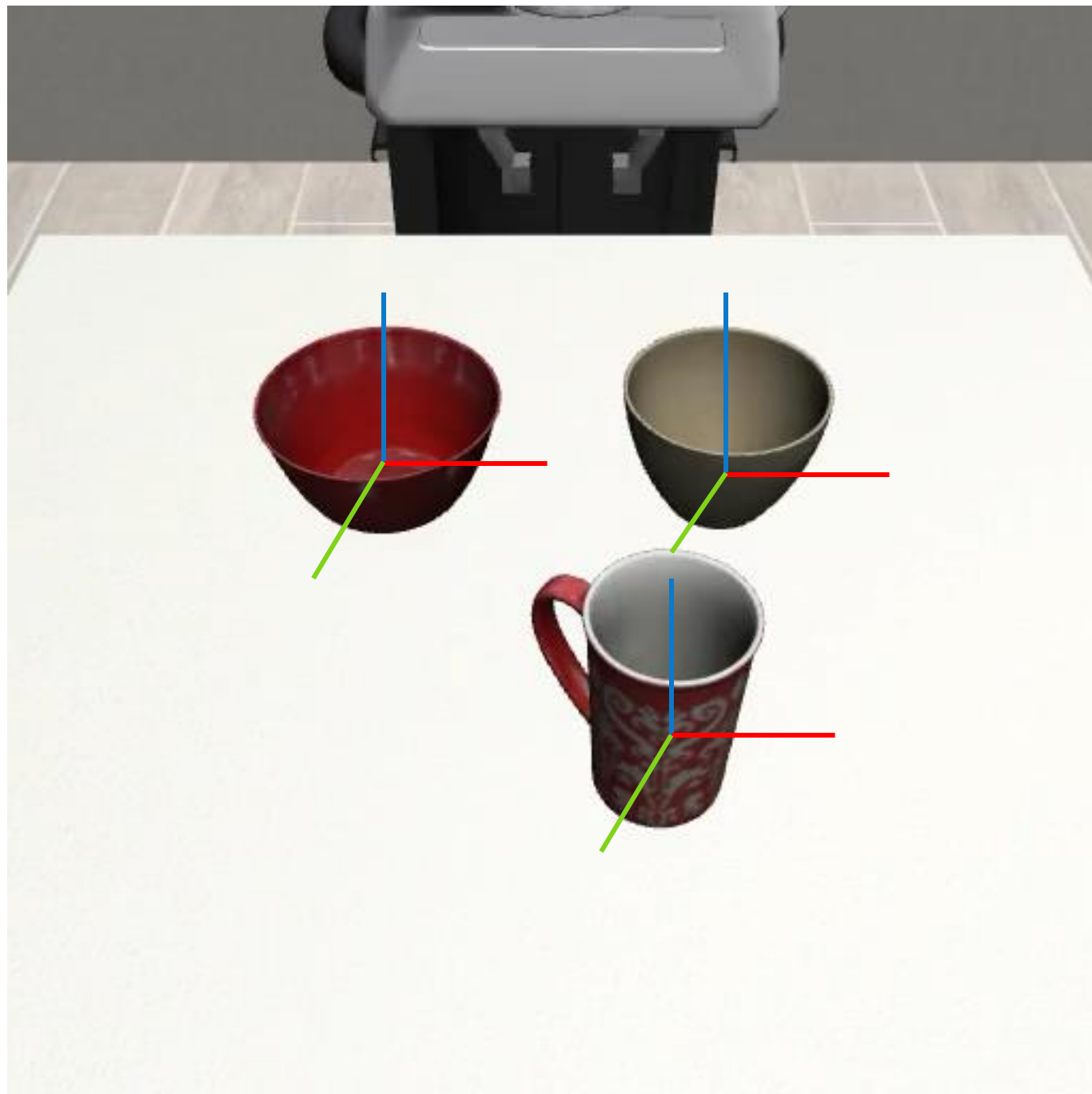


Get



$$g_R = \arg \min_g ||g - \textcolor{teal}{g}_H ||_2^2$$

$$\text{s.t. } V^\pi(x; g) \geq 0$$





Get

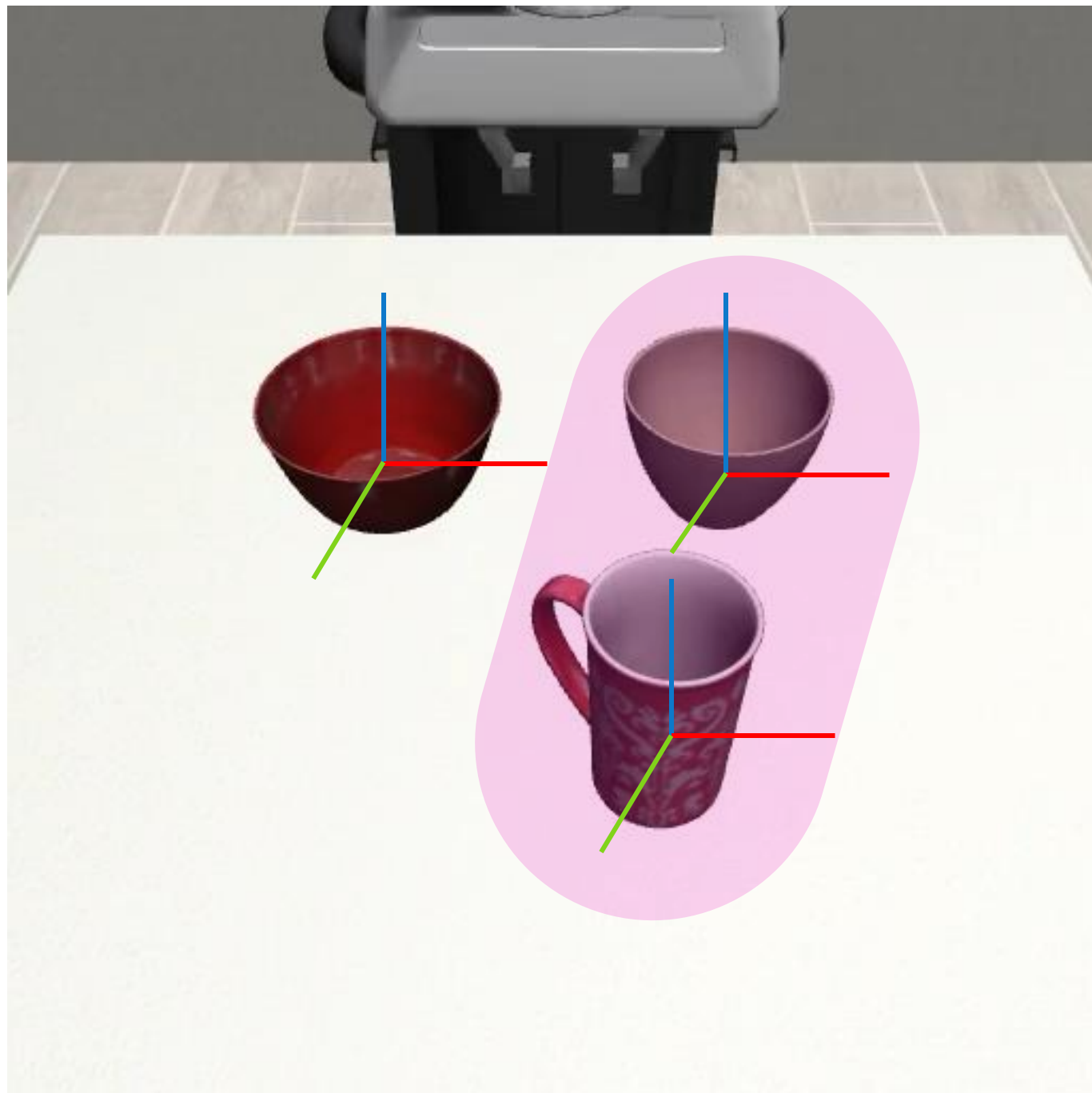
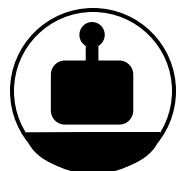


$$g_R = \arg \min_g ||g - g_H||_2^2$$

$$\text{s.t. } V^\pi(x; g) \geq 0$$



instead?





Get



$$g_R = \arg \min_g d(\mathcal{E}(g), \mathcal{E}(g_H))$$

$$\text{s.t. } V^\pi(x; g) \geq 0$$



$$g_R = \arg \min_g d(\mathcal{E}(g), \mathcal{E}(g_H))$$



=

ChatGPT ▾



We: You are a robot in a kitchen. You have a set of items in front of you.
 We: The items are: Red Mug, Red Bowl, Brown Bowl
 We: Given Red Mug, which item is the most similar? Please answer with only one of the letters. The user is organizing their kitchen items by color.

You:

- A) Brown Bowl
- B) Red Bowl



$$g_R = \arg \min_g d(\mathcal{E}(g), \mathcal{E}(g_H))$$

$$g_H = \arg \min_g d(\mathcal{E}(g), \mathcal{E}(g_H))$$



ChatGPT ▾



We: You are a robot in a kitchen. You have a set of items in front of you.
 We: The items are: Red Mug, Red Bowl, Brown Bowl
 We: Given Red Mug, which item is the most similar? Please answer with only one of the letters. The user is organizing their kitchen items by color.

You:

- A) Brown Bowl
- B) Red Bowl

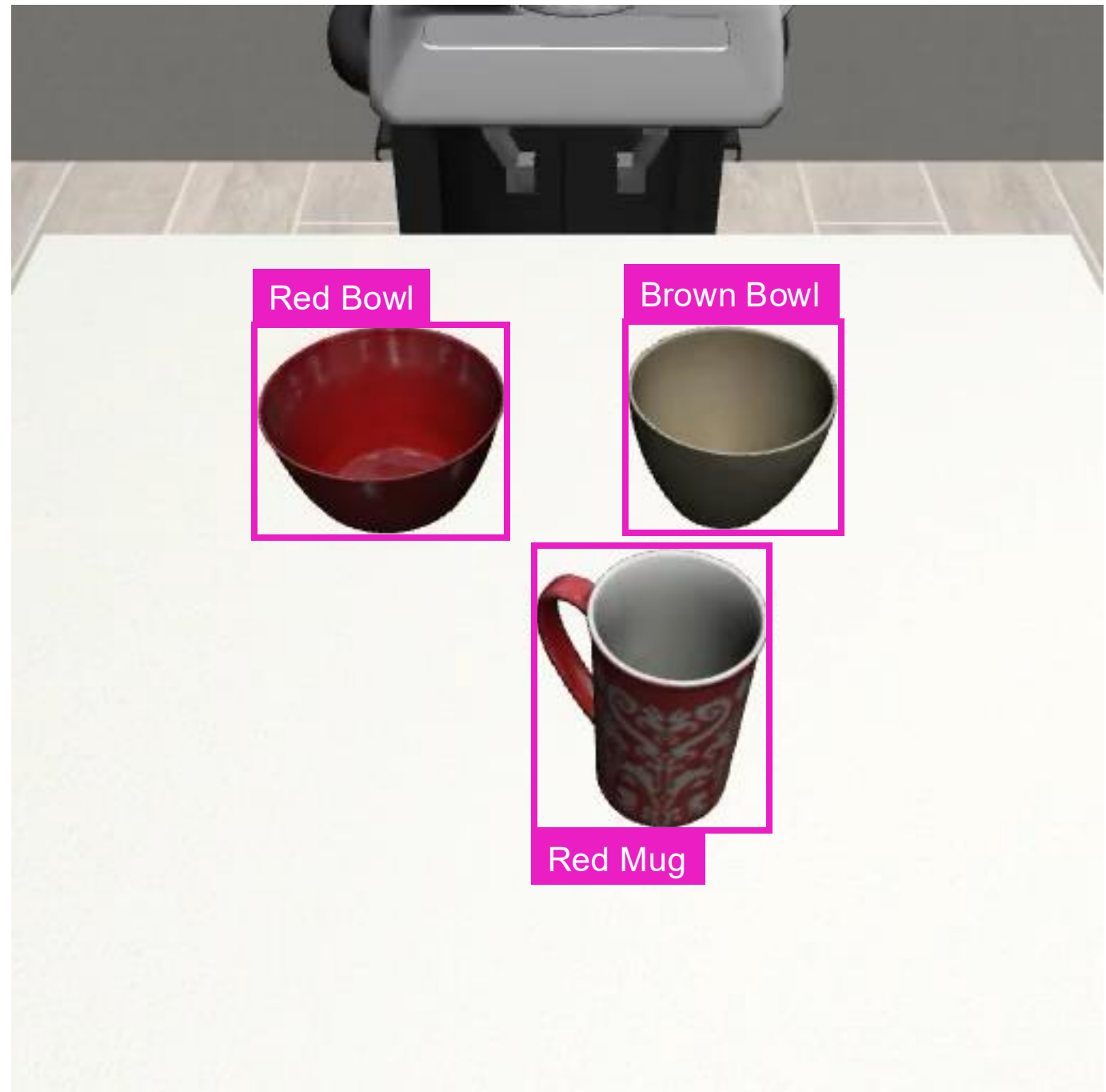
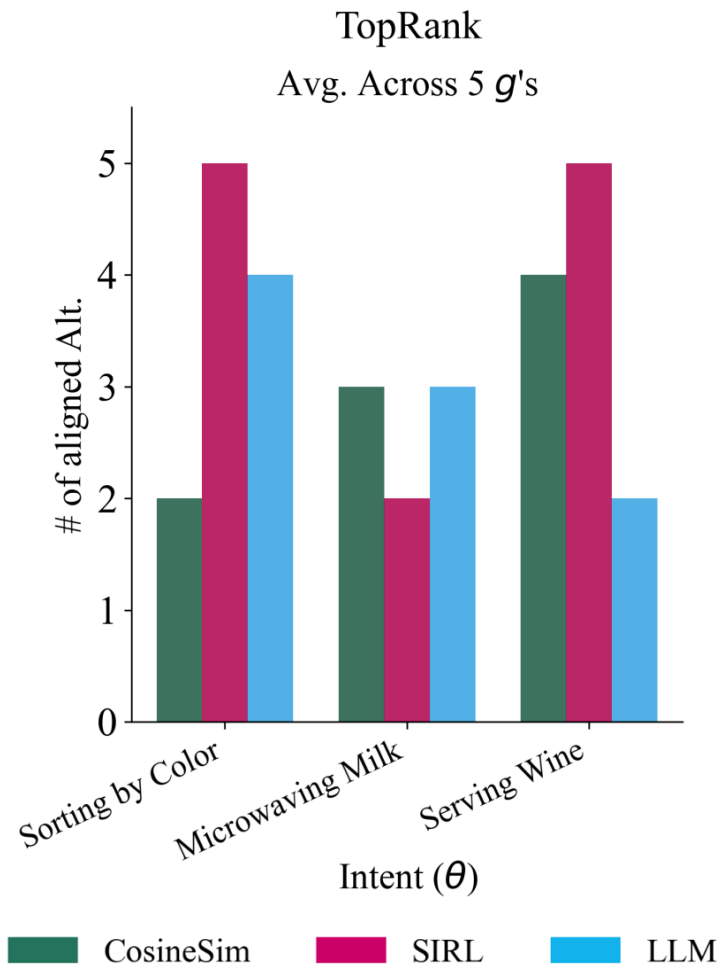


B) Red Bowl



$$g_R = \arg \min_g d(\mathcal{E}(g), \mathcal{E}(g_H))$$

$$g_H = \arg \min_g d(\mathcal{E}(g), \mathcal{E}(g_H))$$





Get

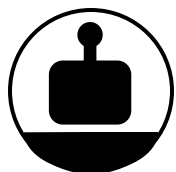


$$g_R = \arg \min_g d(\mathcal{E}(g), \mathcal{E}(g_H))$$

$$\text{s.t. } V^\pi(x; g) \geq 0$$



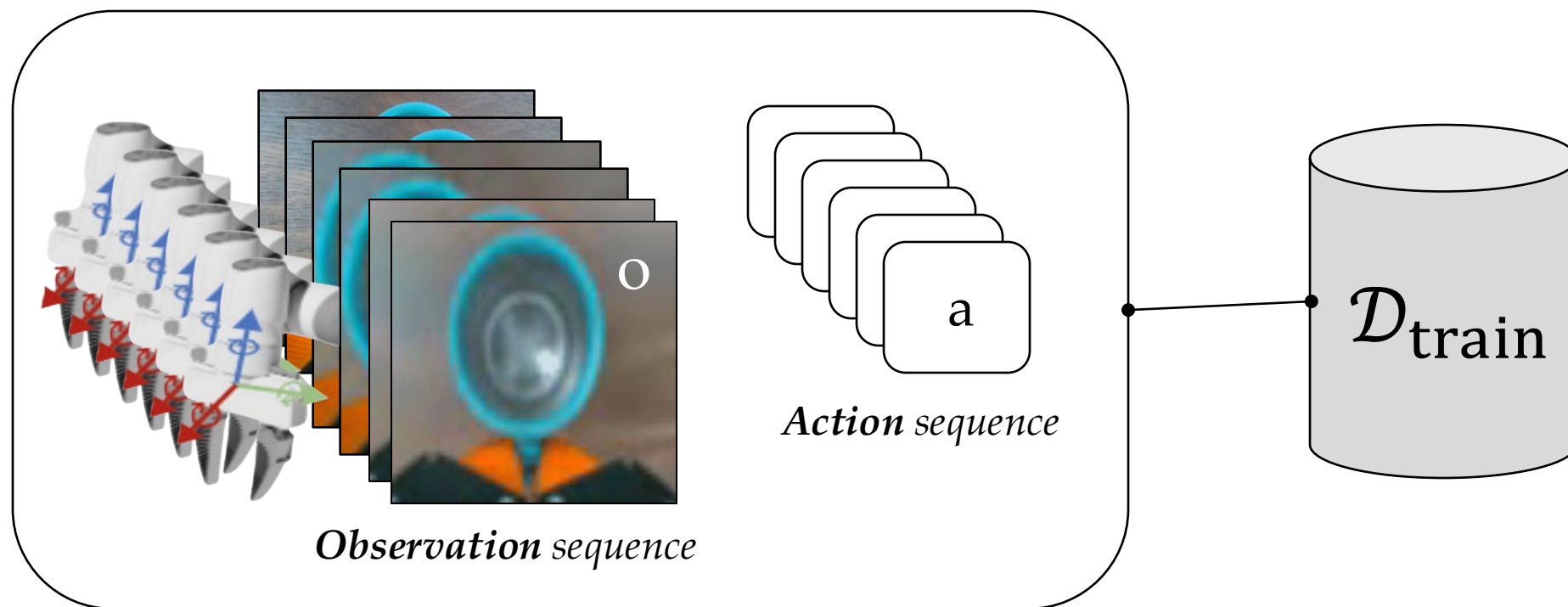
instead?



Hmm... so maybe safety is more than
just avoiding collisions.

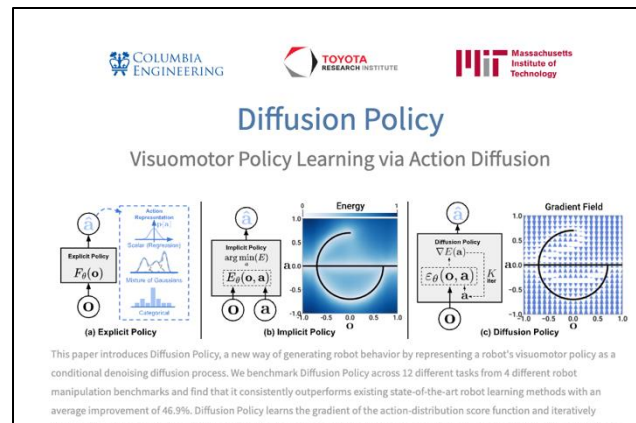
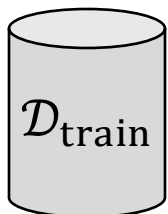
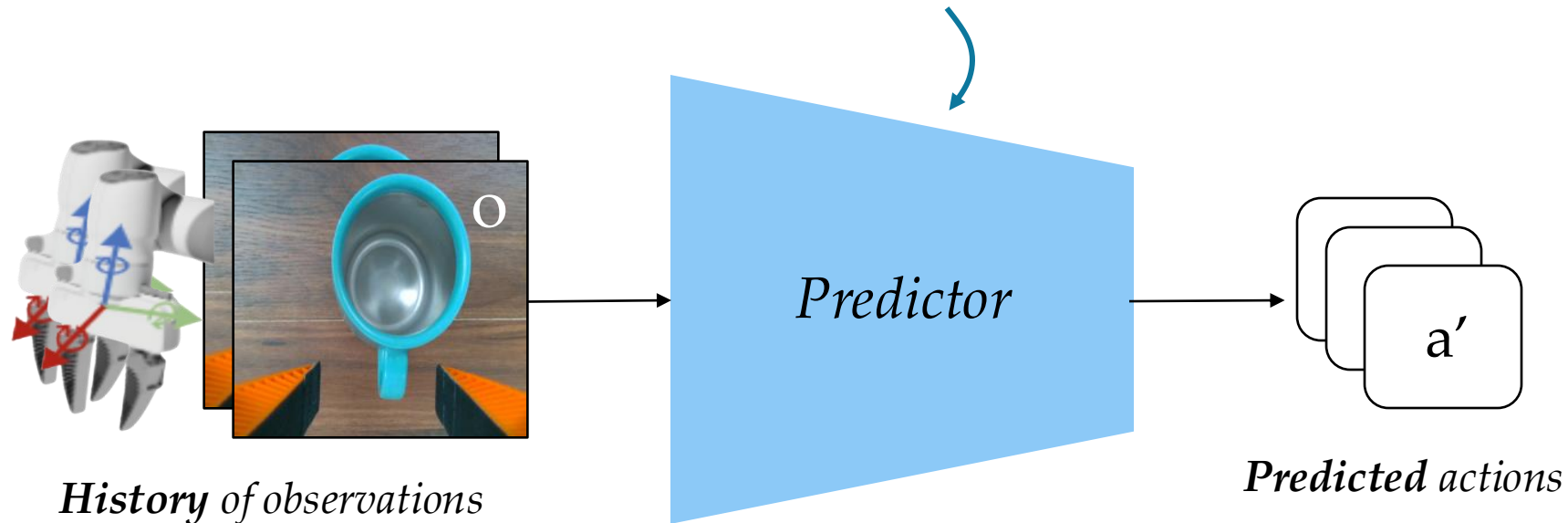
What are other safety problems for
robots, and *how can humans help*?

Let's talk about robot models that are trained on data like...



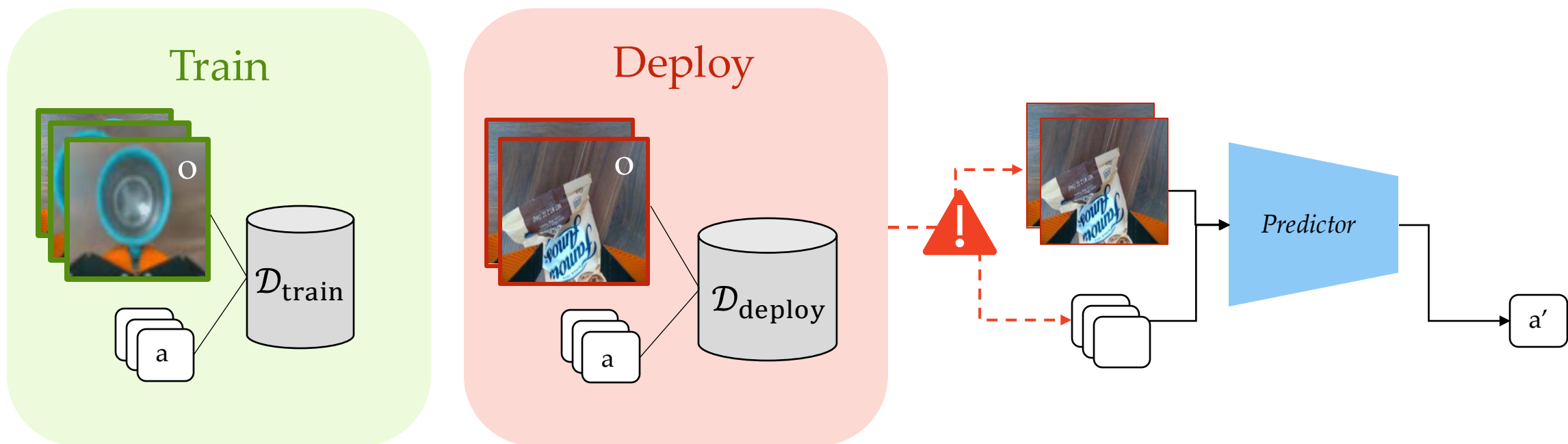
.... and these **models** learn to predict actions ...

e.g., Behavior Cloned Policies



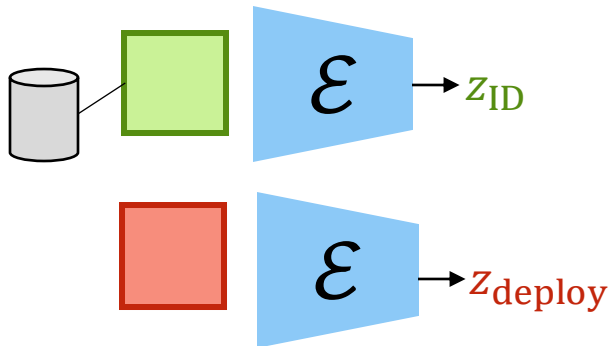
...(& many more)...

How do we detect *out-of-distribution* (OOD) scenarios in these models like these?



How do we detect *out-of-distribution* (OOD) scenarios in these models like these?

Embedding Distances
(e.g., Cosine Sim)

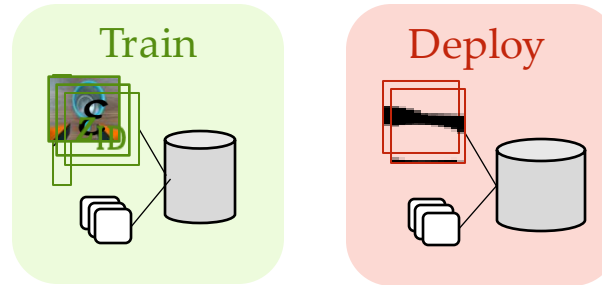


$$\frac{z_{ID} \cdot z_{deploy}}{\|z_{ID}\| \|z_{deploy}\|} < \epsilon \Rightarrow \text{!}$$

[Majumdar et al., arXiv 2025]

[Sinha et al., RSS 2024]

[Luo et al., ICRA 2024]



Foundation Models

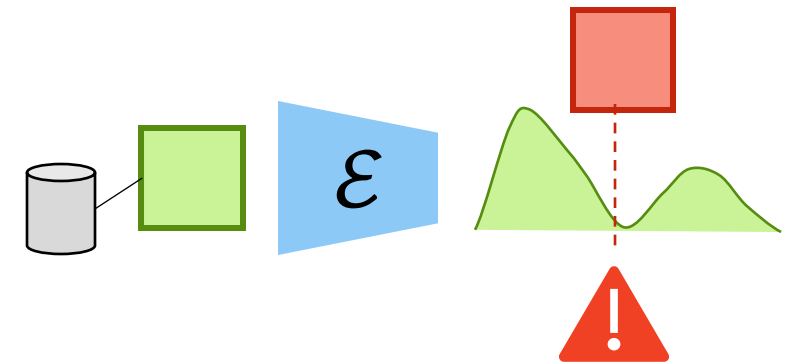


[Ganai et al., arXiv 2025]

[Sinha et al., RSS 2024]

[Elhafsi et al., Autonomous Robots 2023]

Density Estimators



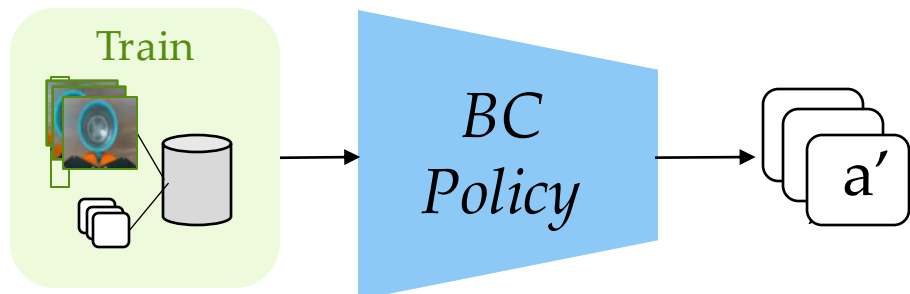
[Xu et al., RSS 2025]

[Liu et al., Neurips 2021]

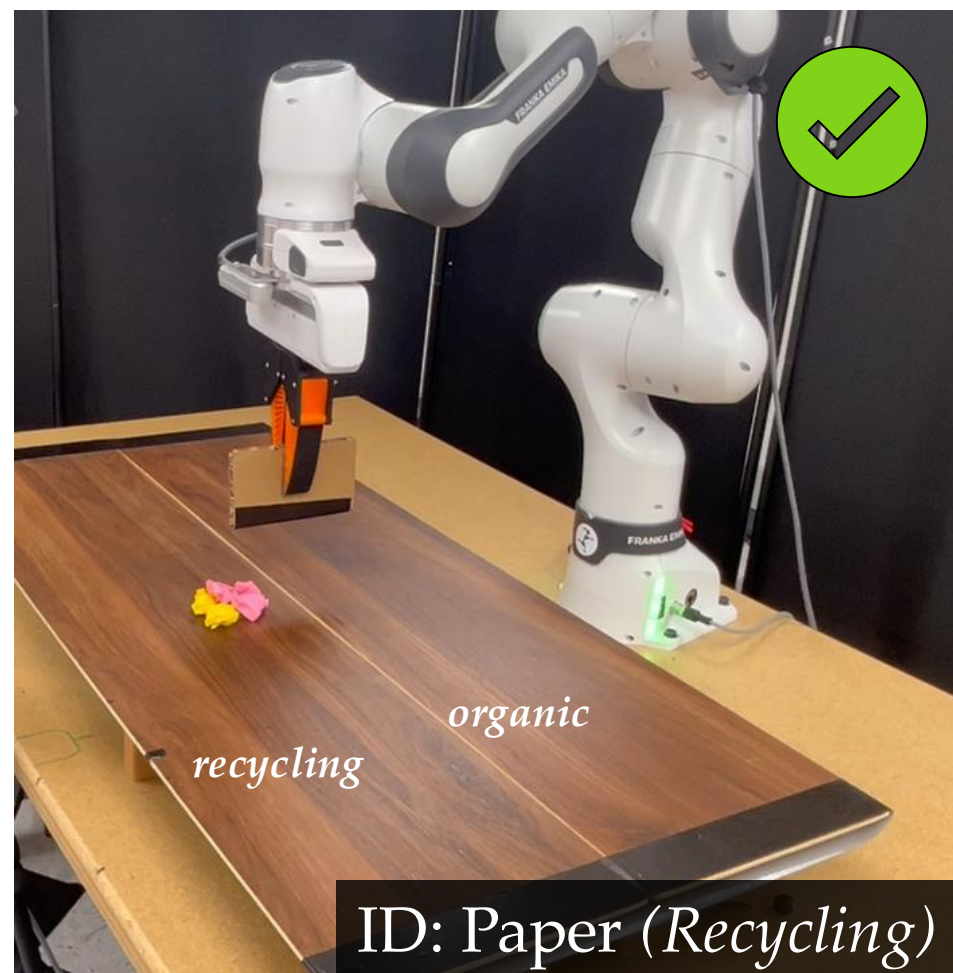
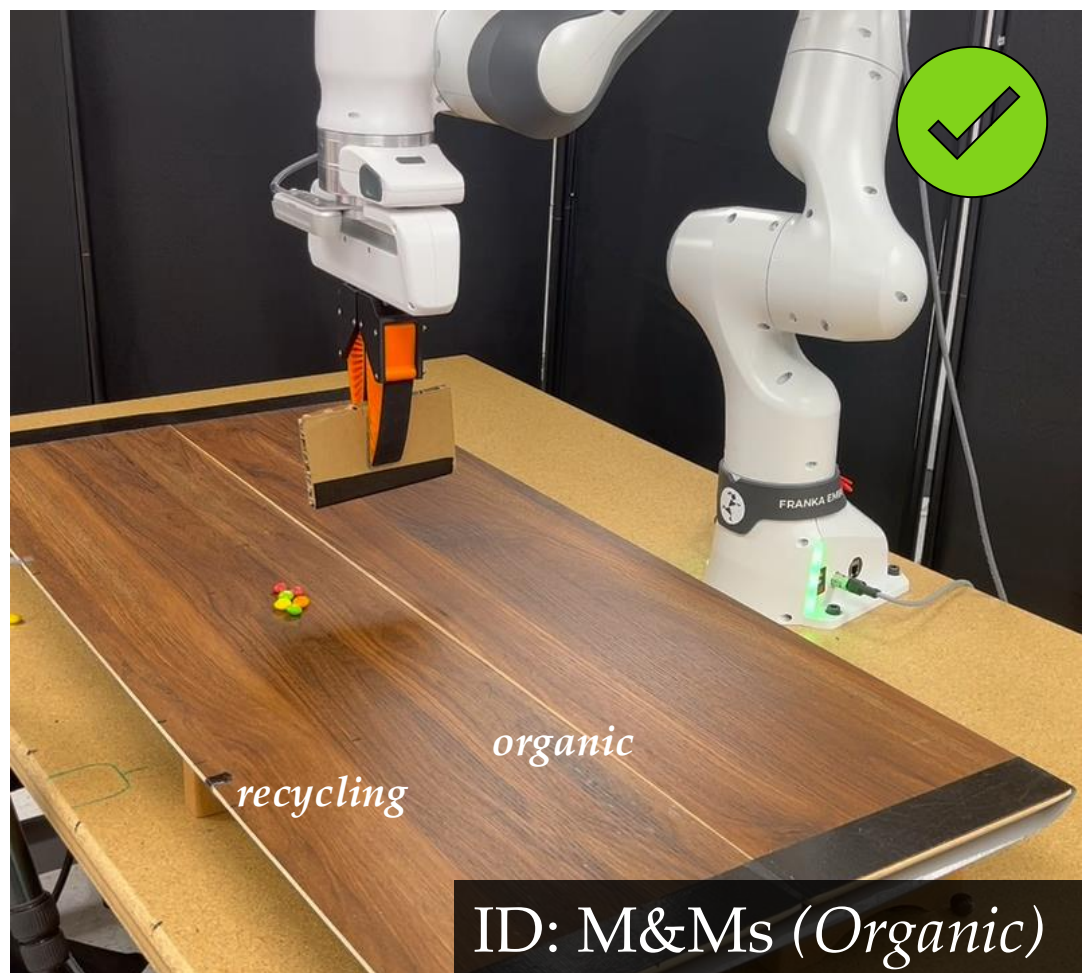
...(& many more)...

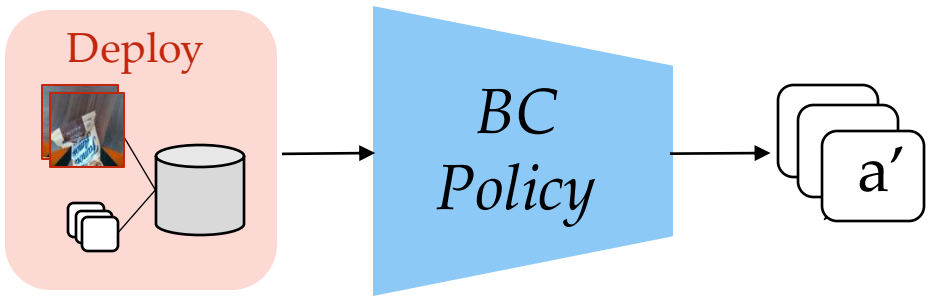
But what should robots *do* once they detect an OOD condition?

How do we go from *detection* to
mitigation of OOD conditions?

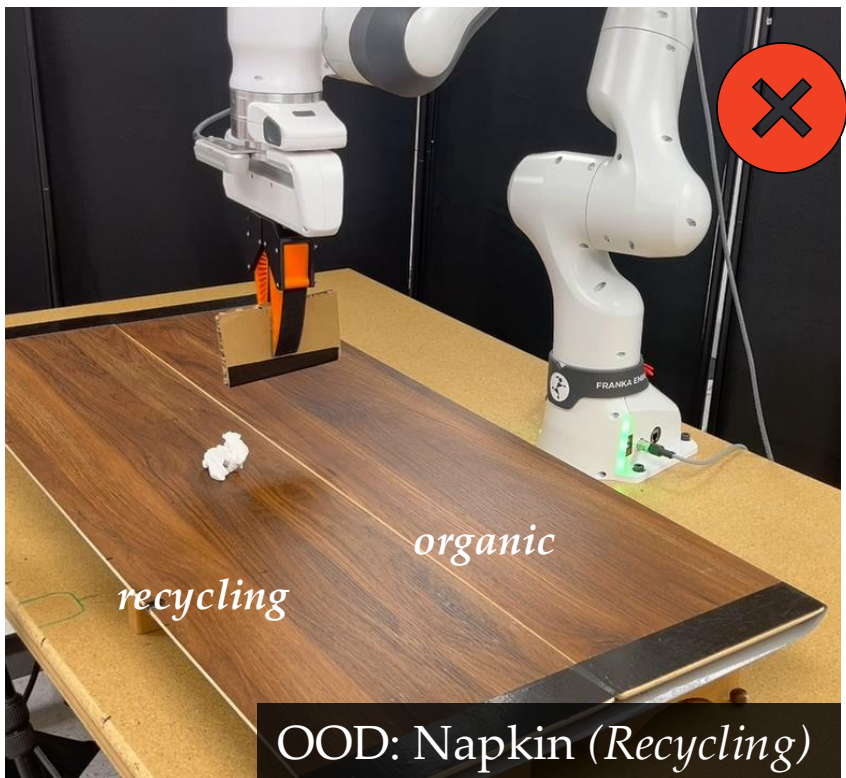
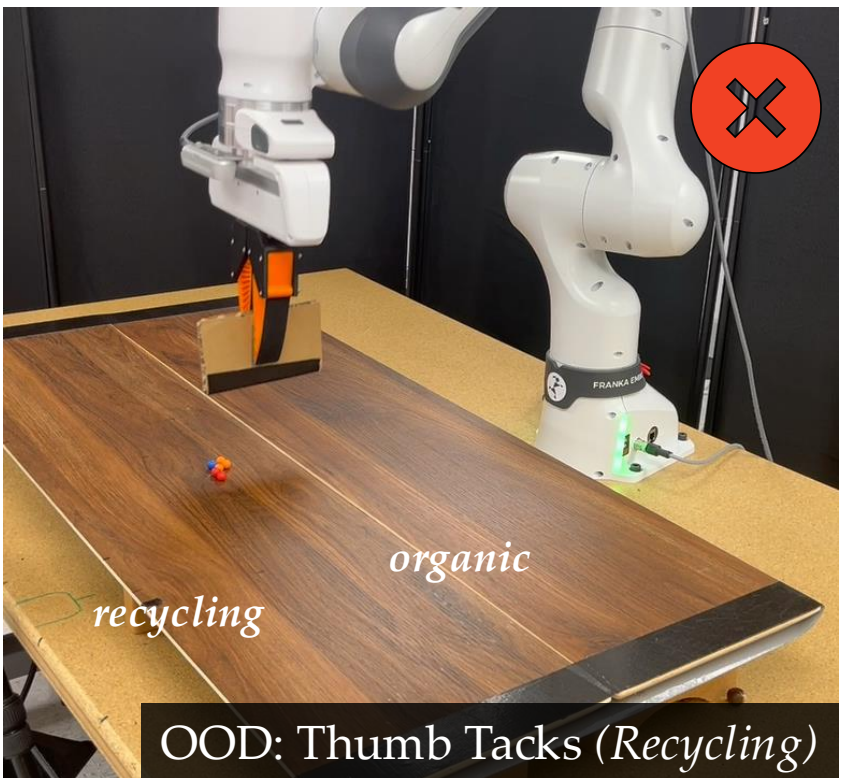
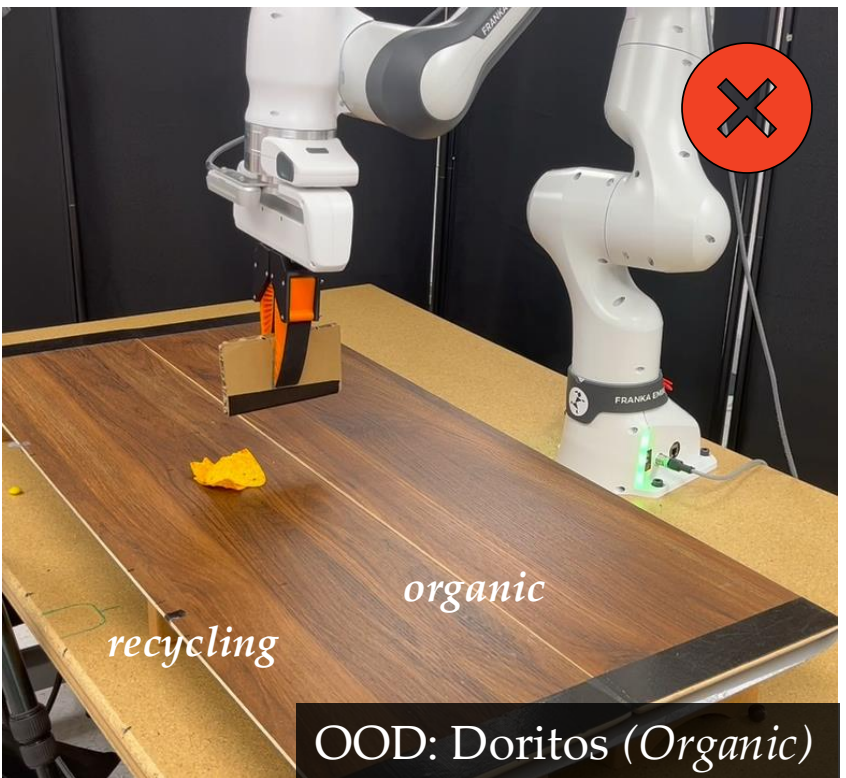


Task: Sort Organic & Recycling Waste

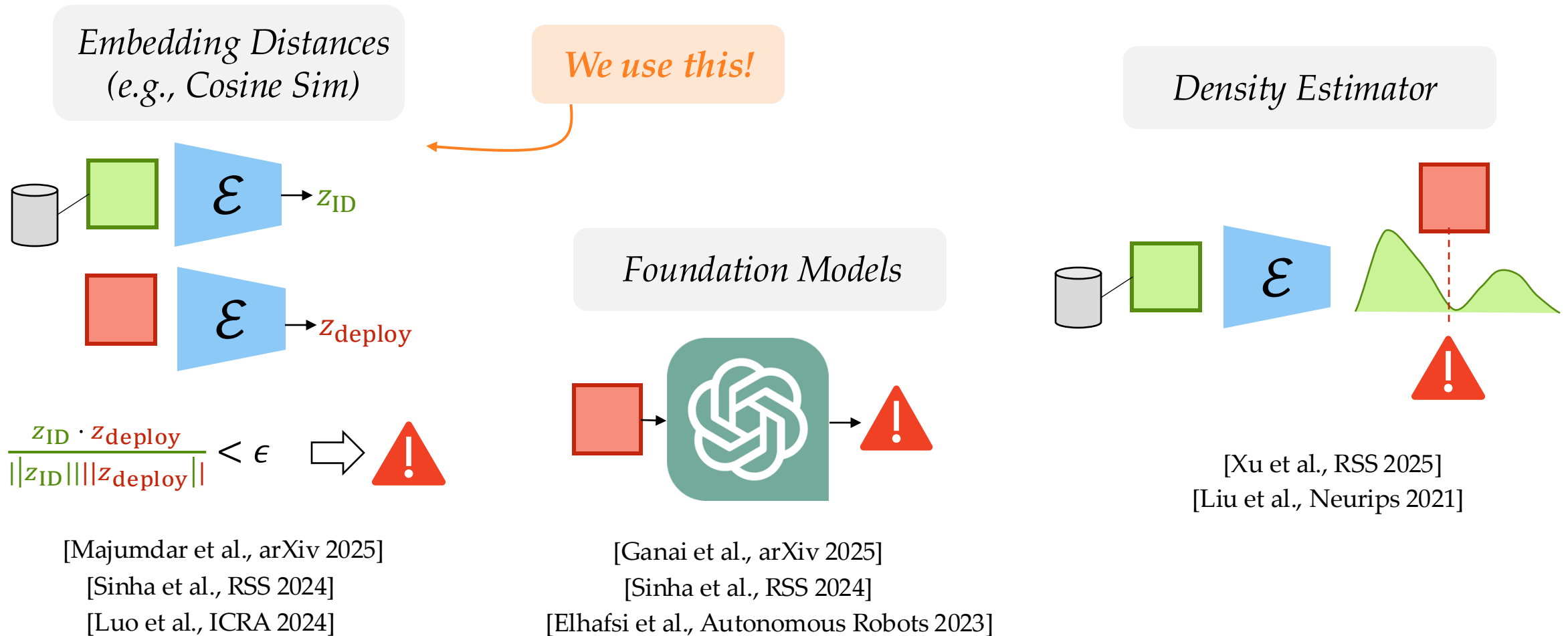




Task: Sort Organic & Recycling Waste



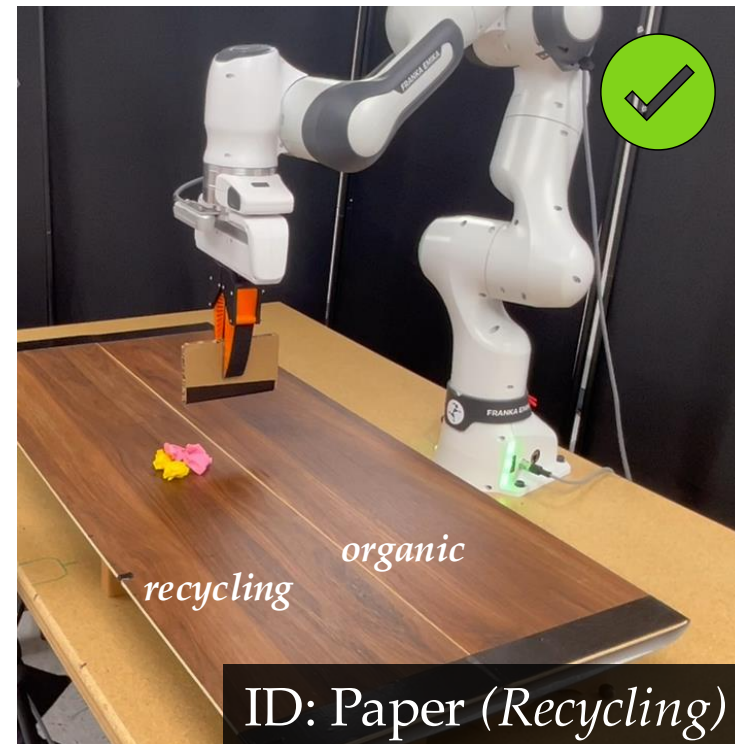
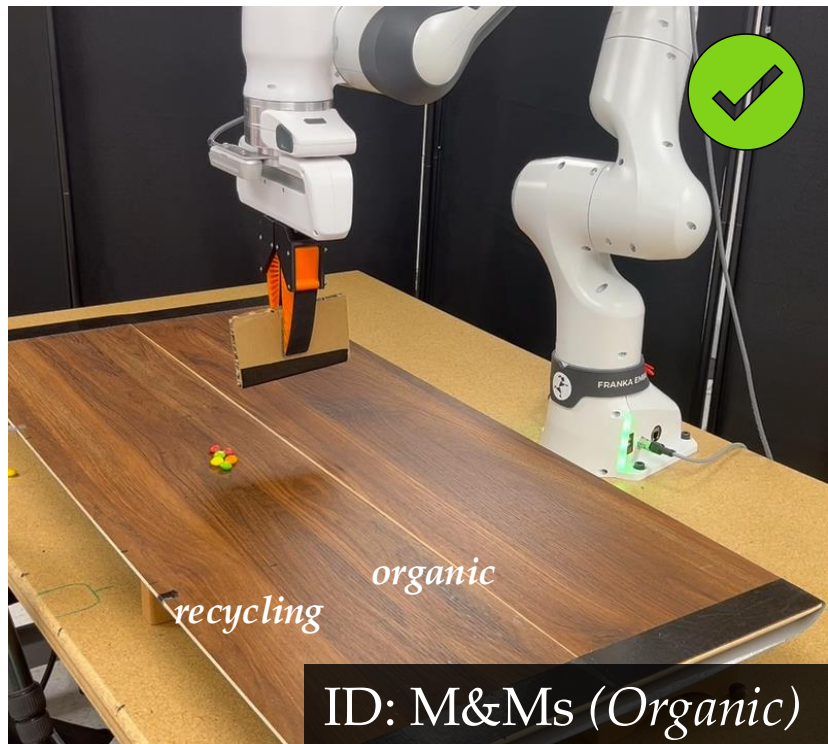
Here is where we can return to these OOD detection methods from before....



So, what should happen next?

It may be tempting to collect more
demonstration data but...

The robot already knows the right behavior in its model, but it doesn't understand correct mapping

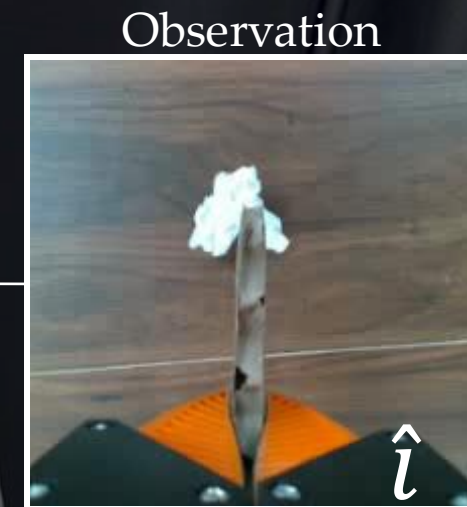
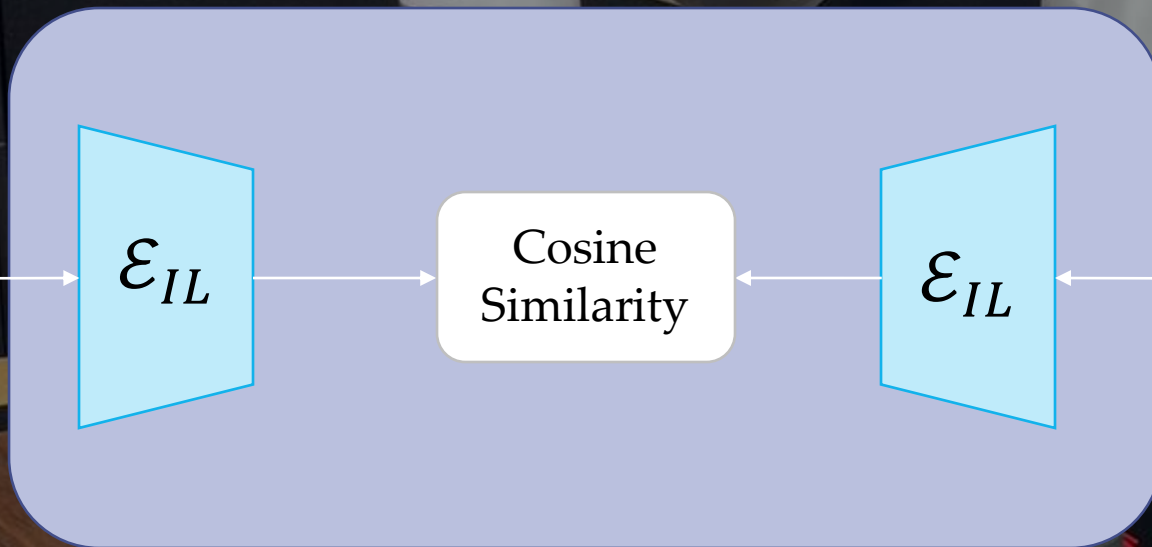
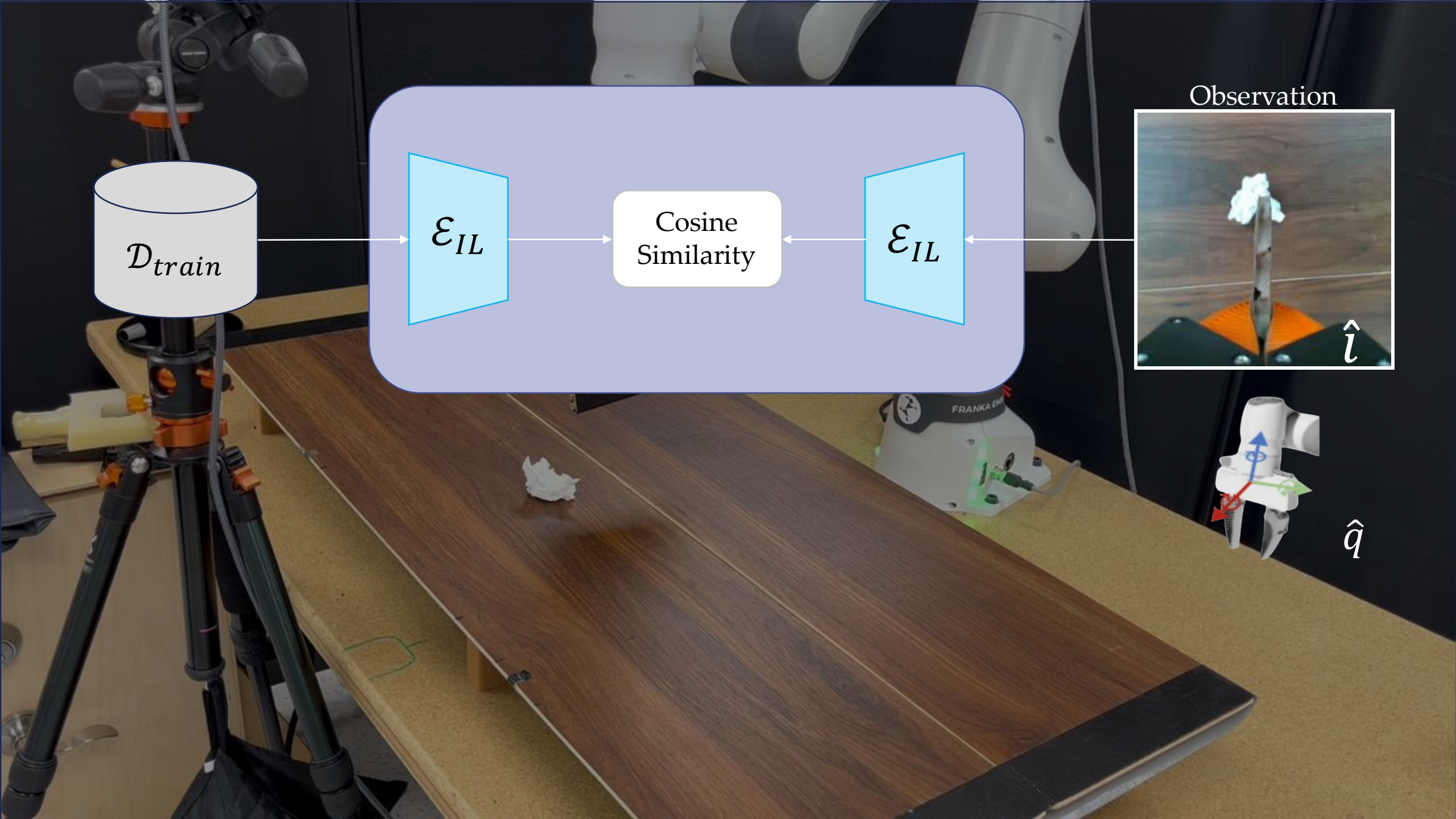


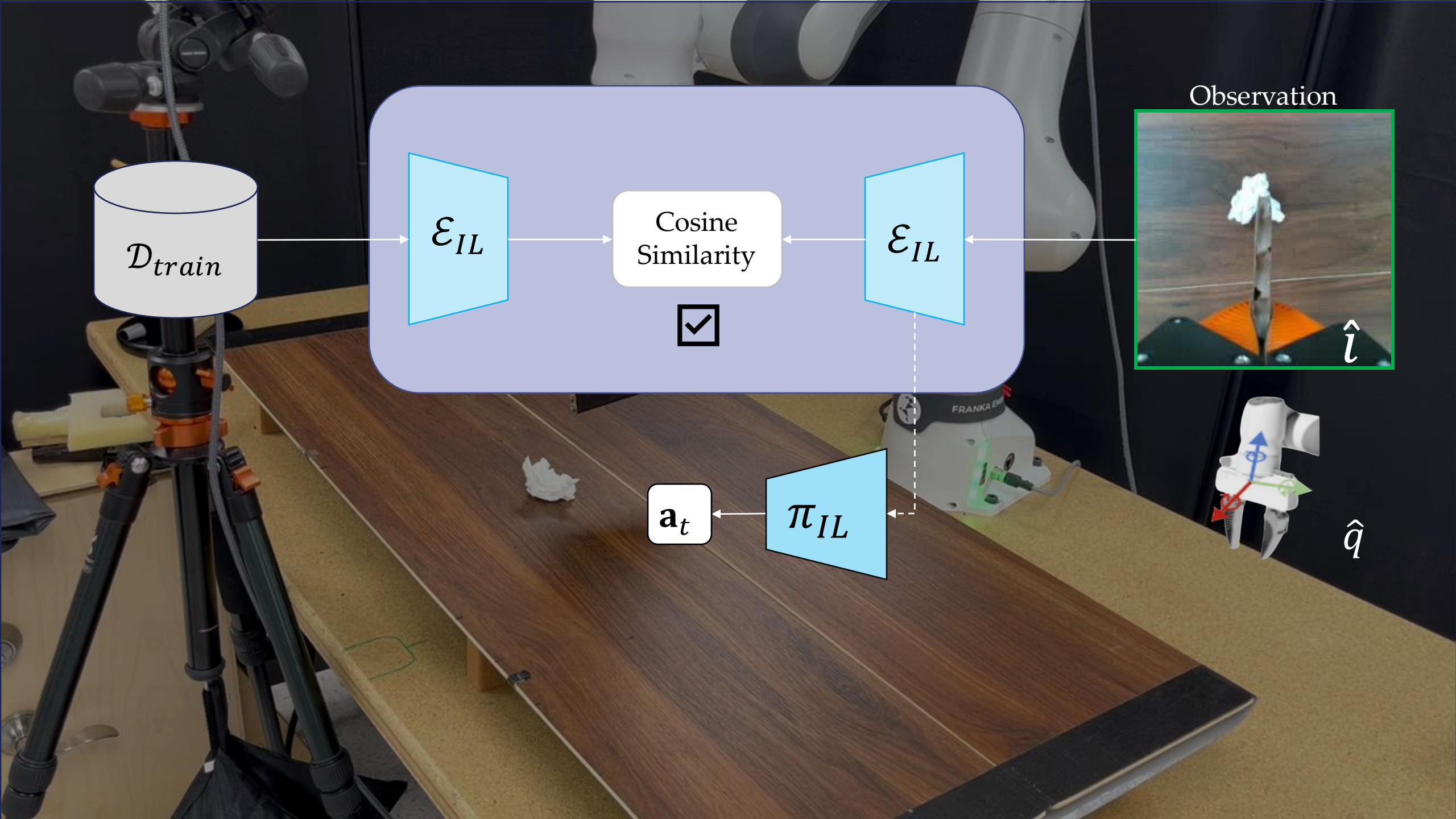
Key idea: In-distribution (ID) behaviors can directly be transferred to OOD conditions that share *functional similarities* with ID conditions.



OOD: Napkin (*Recycling*)

Adapting By Analogy (ABA): A runtime observation intervention approach for generalizing visuomotor policies to OOD obs. via functionally corresponding ID obs.





\mathcal{D}_{train}

\mathcal{E}_{IL}

Cosine
Similarity



\mathcal{E}_{IL}

Observation

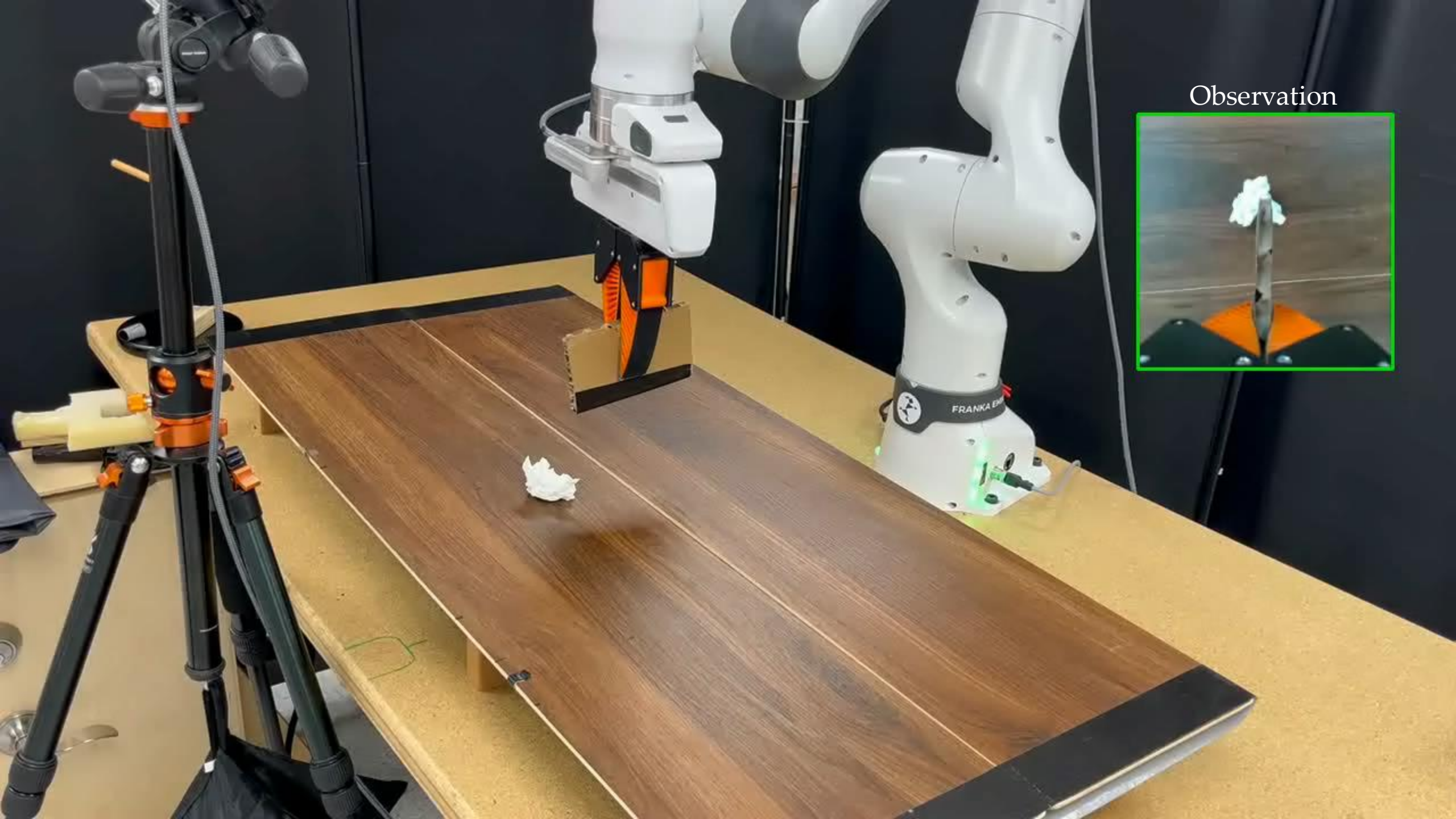


\hat{l}

\mathbf{a}_t

π_{IL}

\hat{q}



Observation



\mathcal{D}_{train}

Observation



\hat{l}

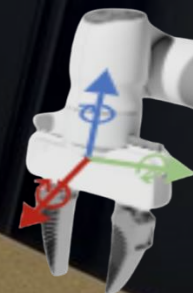
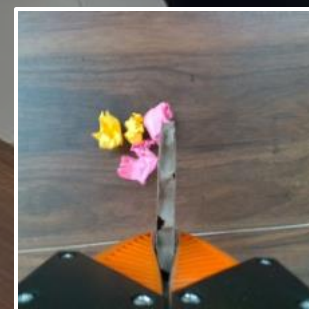


\hat{q}

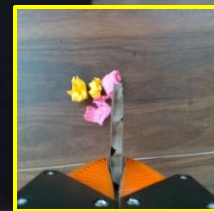
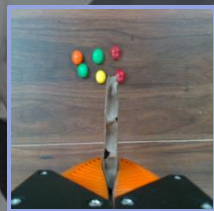


Functional Correspondences (φ)

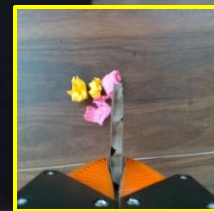
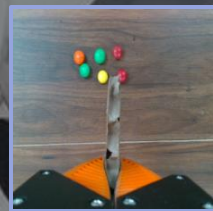
Observation



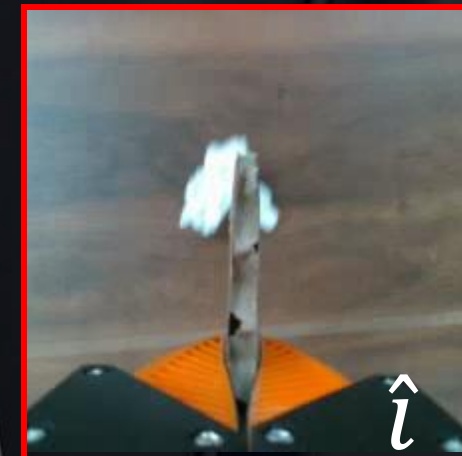
ABA retrieves the top k functionally corresponding training observations by first matching *semantic segments* of the images (e.g., via grounded segment anything)



ABA checks if there are multiple behavior modes in the retrieved samples.



Observation



Update (φ)



Robot

What correspondences
do you see?

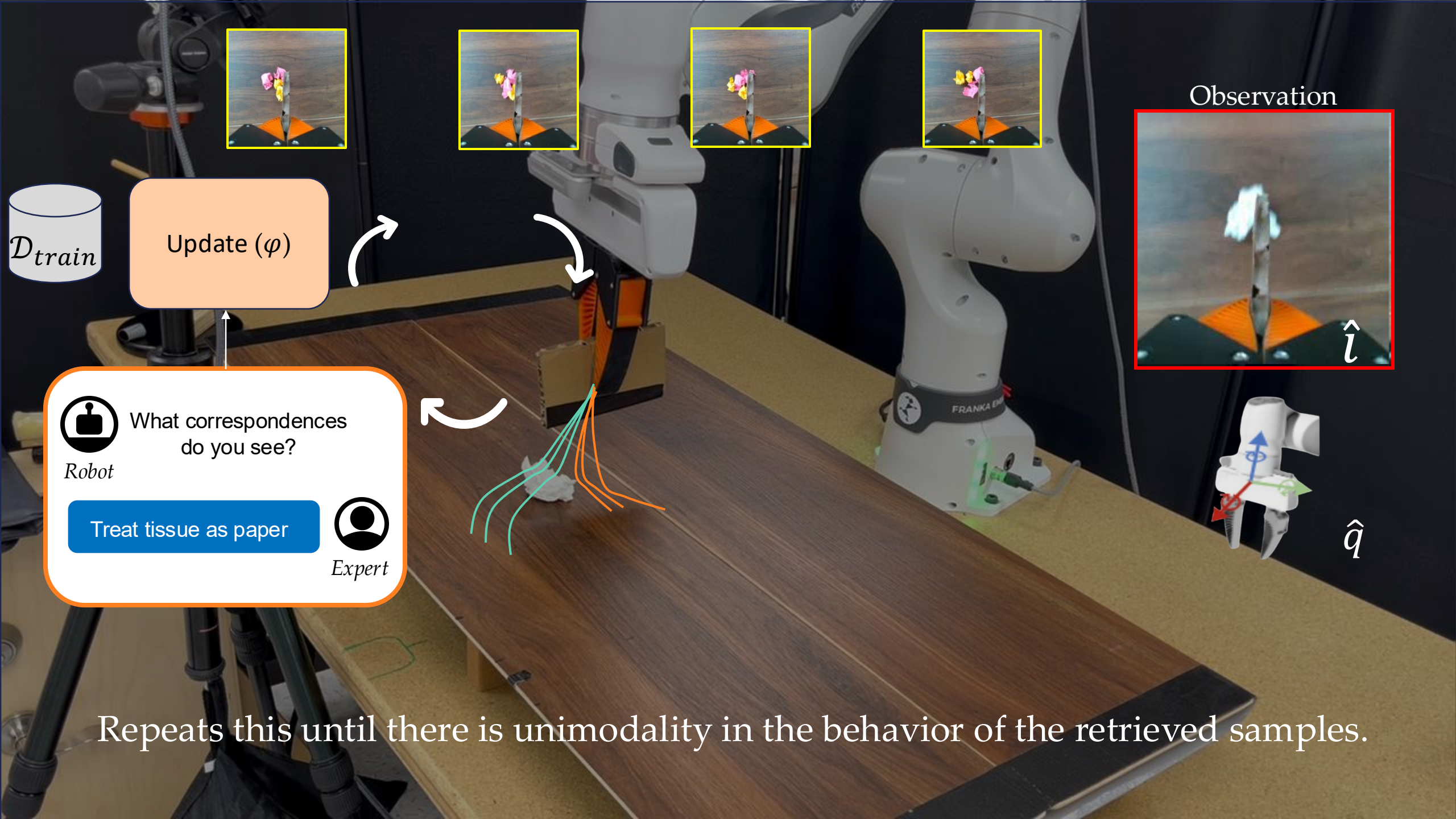
Treat tissue as paper



Expert



If multimodal behaviors are present, **ABA** seeks feedback from the expert to refine the set of functional correspondences (φ)



\mathcal{D}_{train}

Update (φ)



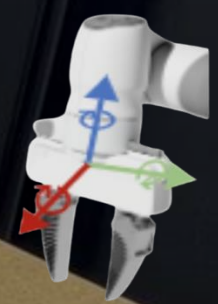
What correspondences do you see?

Robot

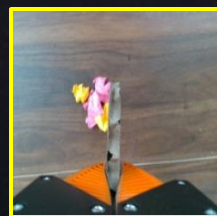
Treat tissue as paper



Expert



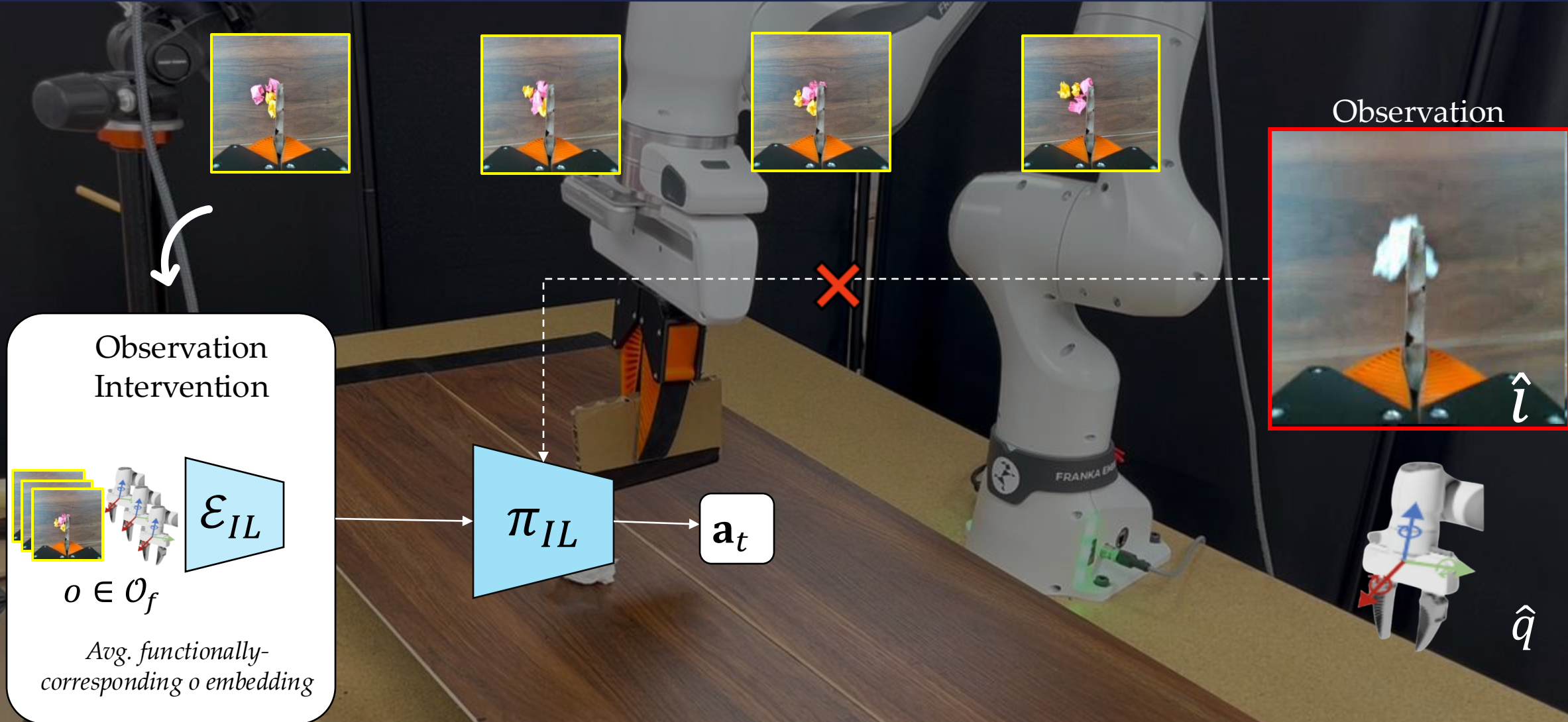
Repeats this until there is unimodality in the behavior of the retrieved samples.



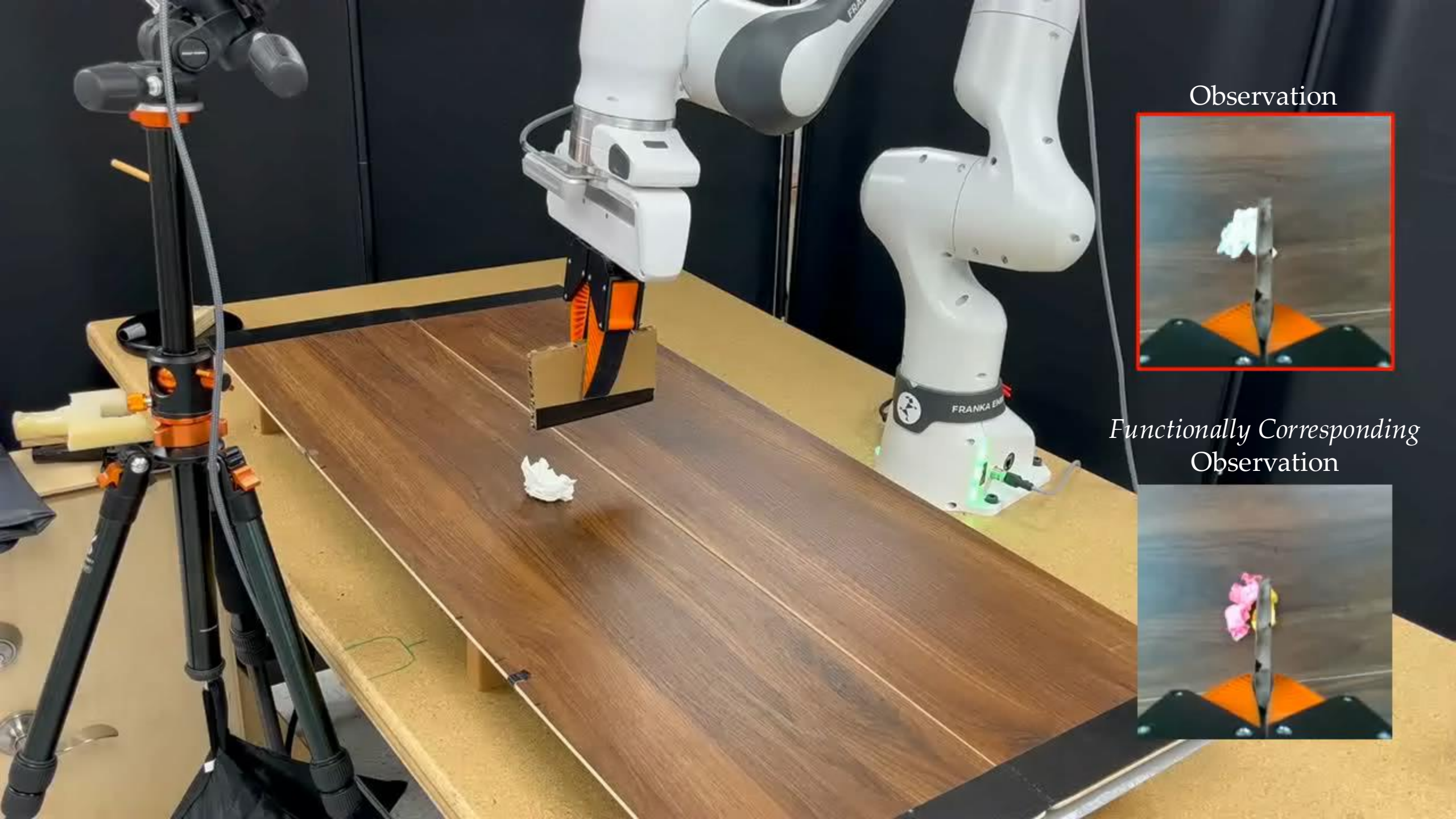
Observation



Great, we identified the correct behavior!



ABA intervenes on the policy using the avg embeddings of the retrieved functionally corresponding observations.



Observation



*Functionally Corresponding
Observation*



What kind of features maximally help observation interventions?

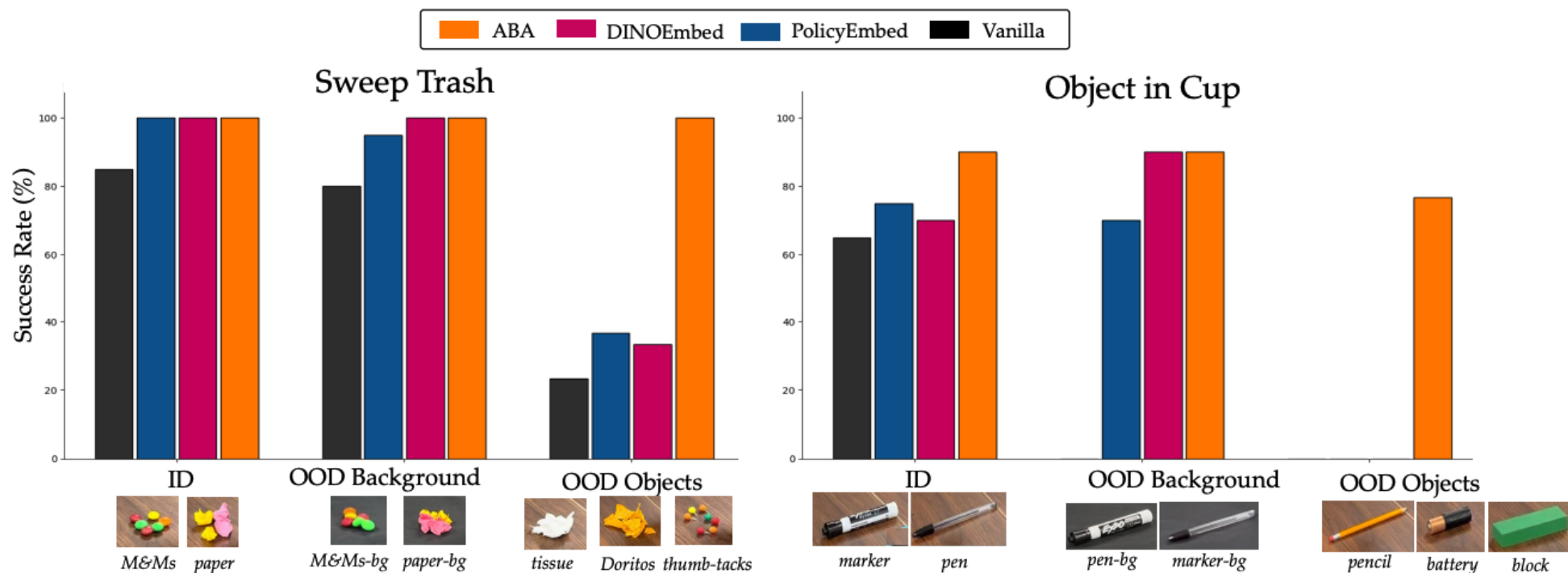


Figure 3: Task Success in ID and OOD Environments. We report the task success rate averaged across 10 rollouts (per each ID and OOD conditions) and averaged across ID, OOD background, or OOD object conditions. For both the sweep-trash and the object-in-cup tasks, we see that **ABA** consistently achieves the highest task success rate compared to baselines.

How efficient is ABA at seeking expert feedback in OOD environments?

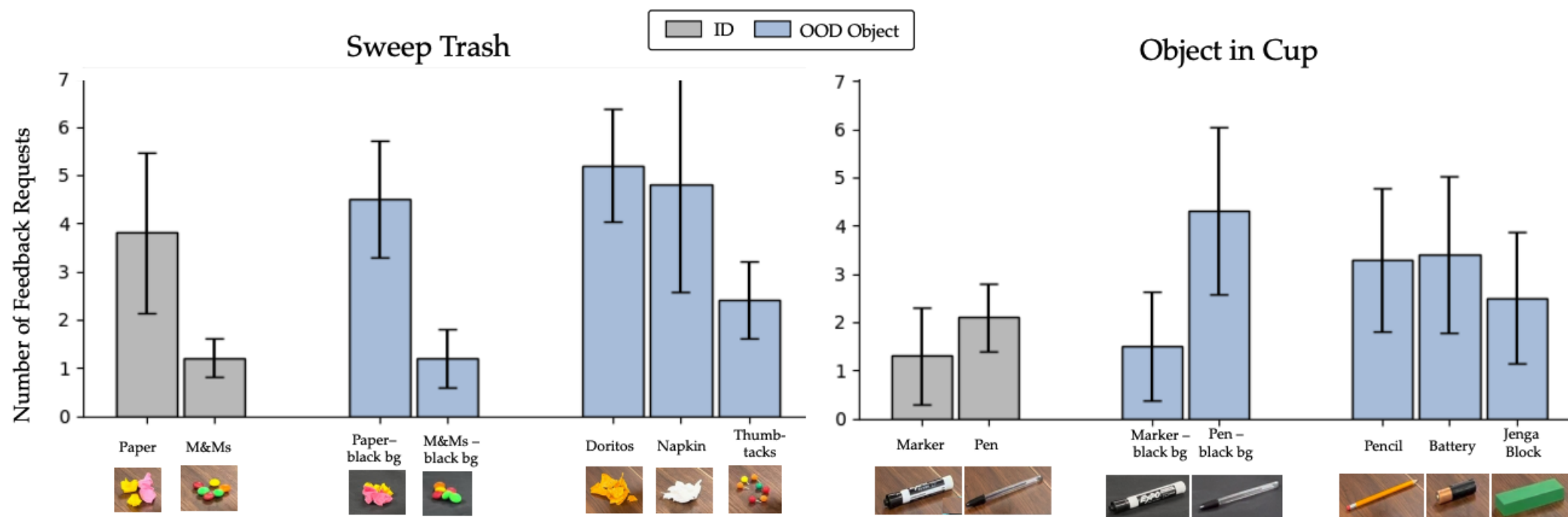


Figure 4: **Expert Feedback Requested by ABA.** We show mean and standard error for the number of feedback requests across 10 rollouts per each environment. We find that ABA infrequently queries the expert for correspondances, given that sweep-trash has 70 timesteps and object-in-cup has 120.

This is just a small *slice* of the problem of safely deploying robots...


EAIS SP25


[Home](#)
[Schedule](#)
[Staff](#)
[Syllabus](#)

Search EAIS SP25

Embodied Artificial Intelligence Safety

Spring 2025. 16-886. Monday / Wednesday 11:00-12:20.



 Make me popcorn.

Announcements

Week 8 Lecture Notes

Mar 24 · 0 min read

System-level anomalies [\[notes\]](#) and [\[slides\]](#) posted online.

Course Overview

Safety is a nuanced concept. For embodied systems, like robots, we commonly equate safety with collision-avoidance. But out in the “open world” it can be much more: for example, a safe mobile manipulator should understand when it is not confident about a requested task and understand that areas roped off by caution tape should never be breached. However, designing systems with such a nuanced understanding is an outstanding challenge, especially in the era of large robot behavior models.

In this graduate seminar class, we study the question of if (and how) the rise of modern artificial intelligence (AI) models (e.g., deep neural trajectory predictors, large vision-language models, and latent world models) can be harnessed to unlock new avenues for generalizing safety to the open world. From a foundations perspective, we study safety methods from two complementary communities: *control theory* (which enables the computation of safe decisions) and *machine learning* (which enables uncertainty quantification and anomaly detection). Throughout the class, there will also be several guest lectures from experts in the field. Students will practice essential research skills including reviewing papers, writing project proposals, and technical communication.

Prerequisites

The course is open to graduate students and advanced undergraduates. While there are no strict prerequisites, familiarity with sequential decision-making, machine learning, optimization, and probability are highly encouraged. Experience with high-level programming languages like Python or MATLAB are also strongly encouraged.

Consider taking 16-886 Embodied AI Safety in Spring 2026 ☺