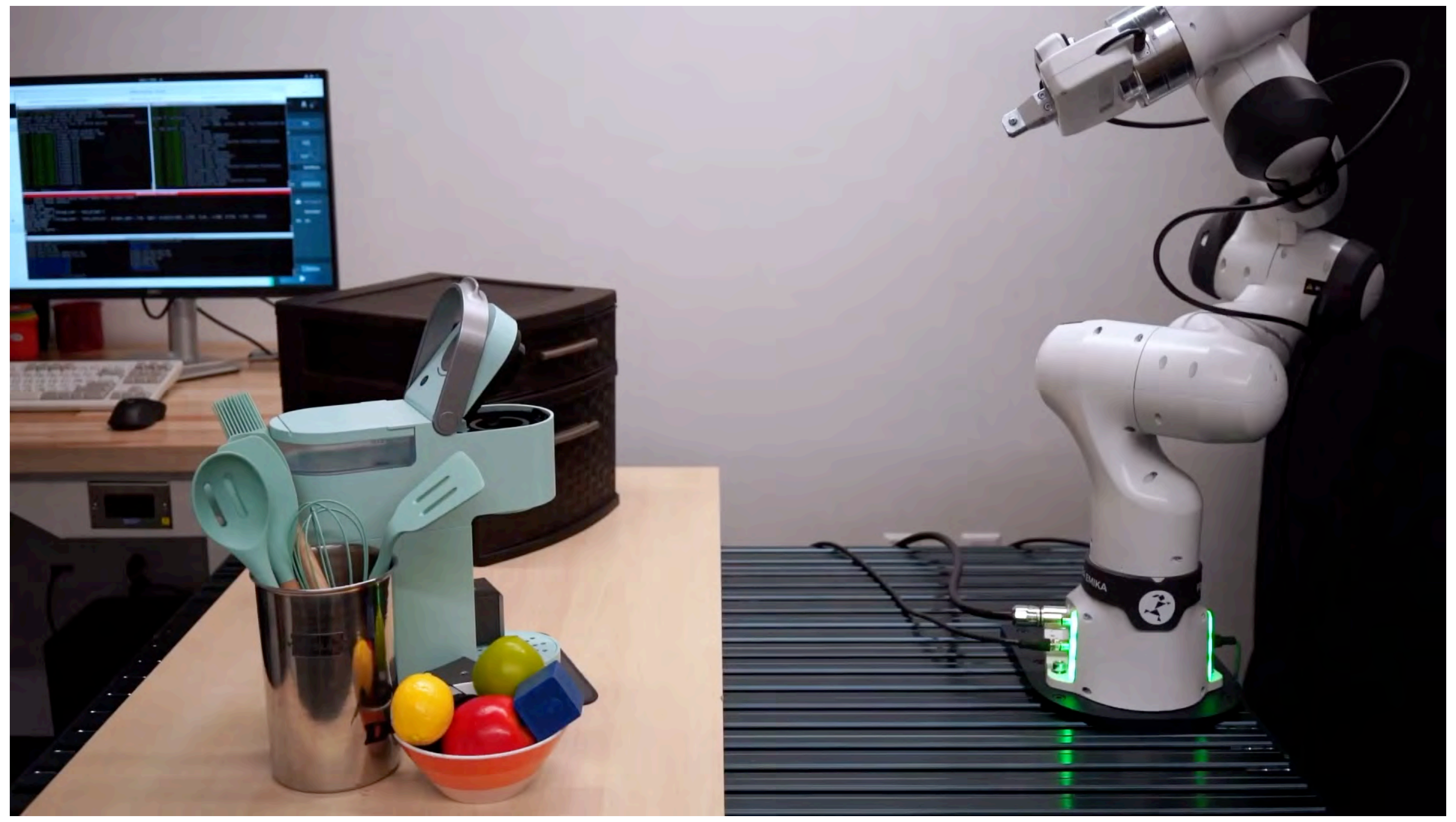


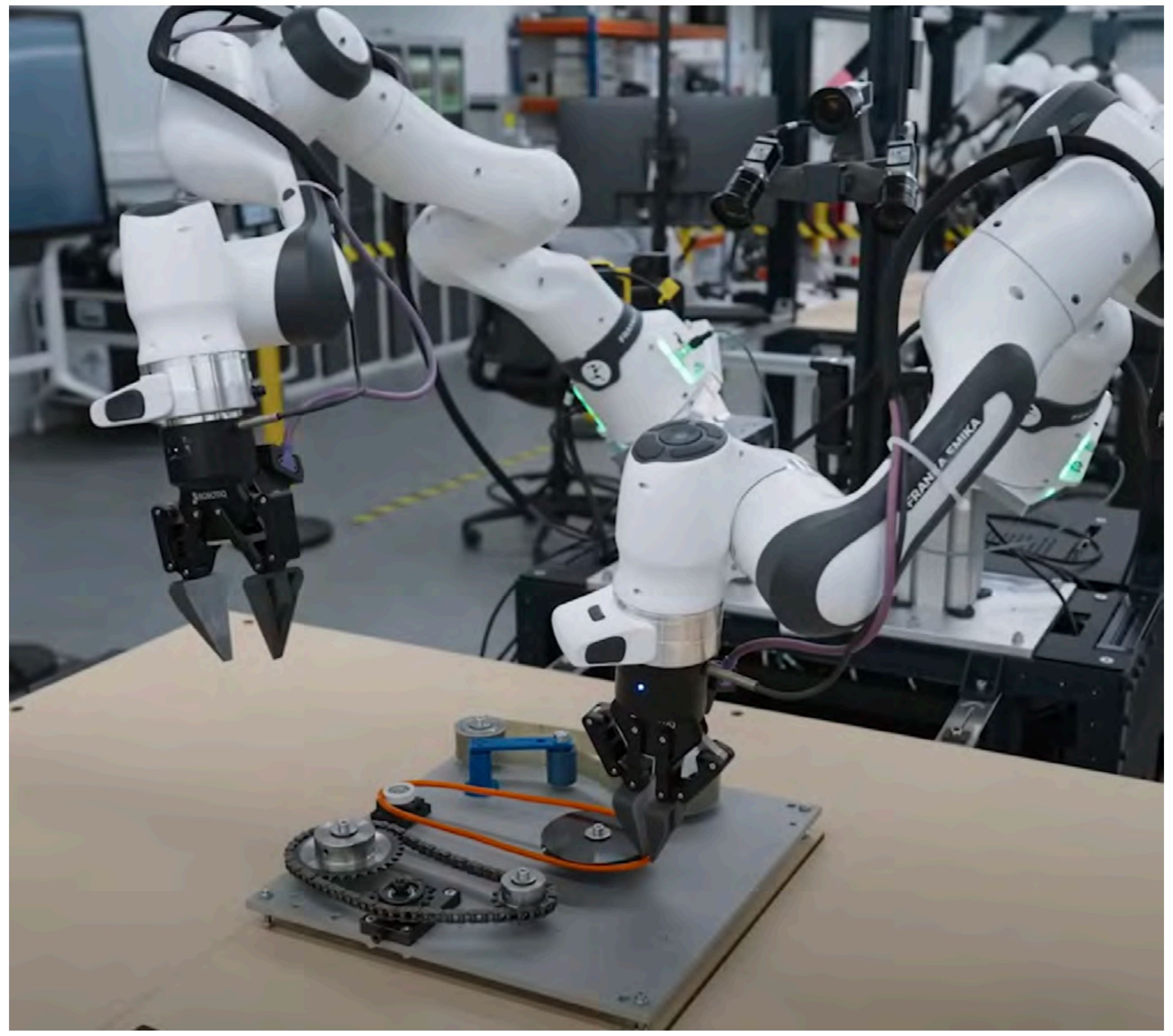
Building General-Purpose Robots through Learning and Composing Generalizable Skills

Xiaolin Fang

Massachusetts Institute of Technology

Towards General-Purpose Embodied Agents

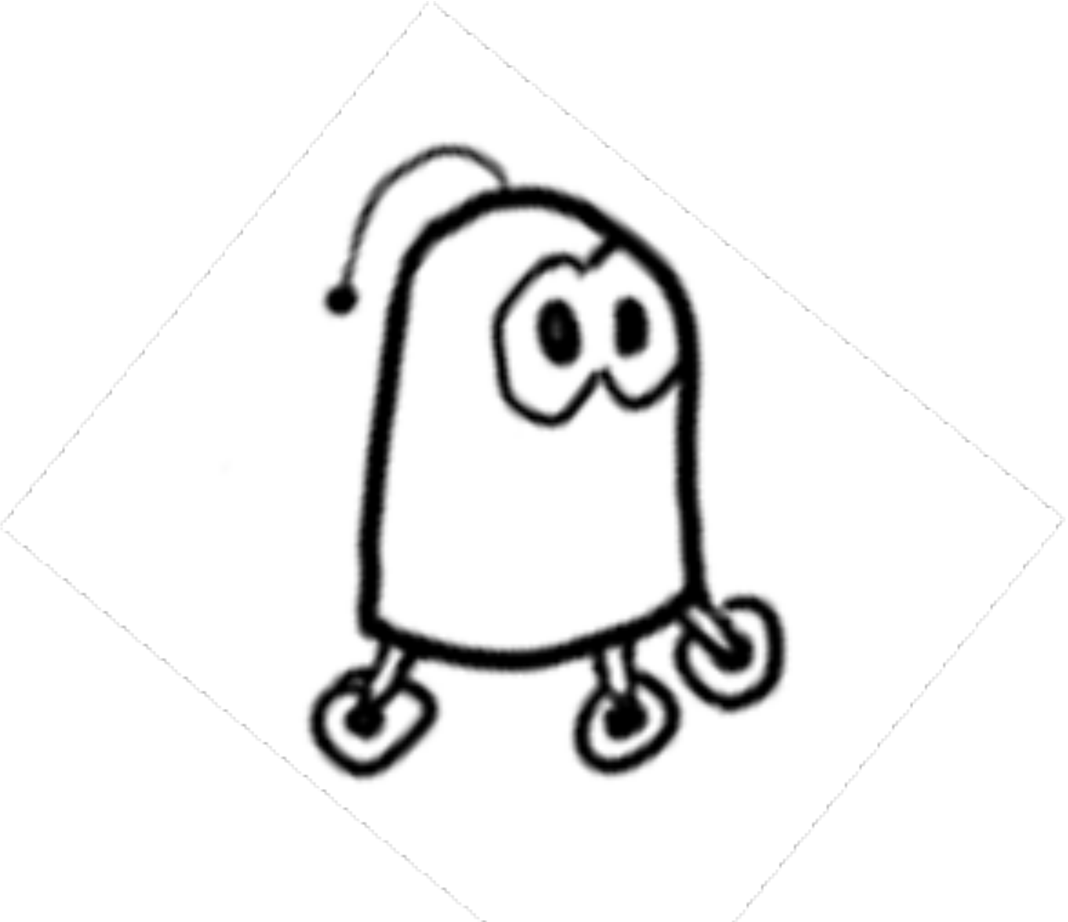
Goal : Having a robot that can do various tasks, with many objects, across different environments

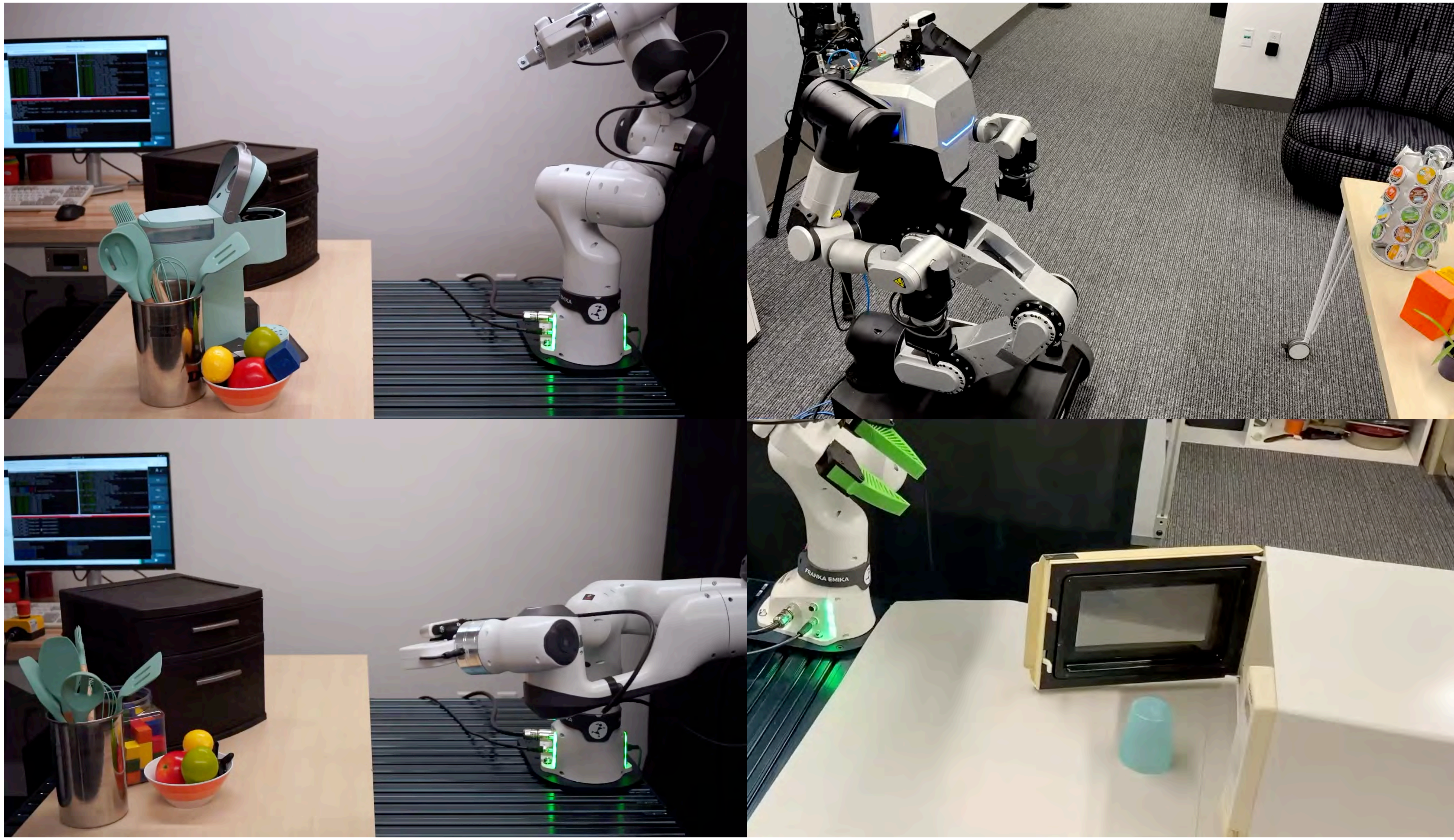



Towards General-Purpose Embodied Agents

Generalization

- Task Description
- Visual Changes
- Embodiment
- Spatial Constraints

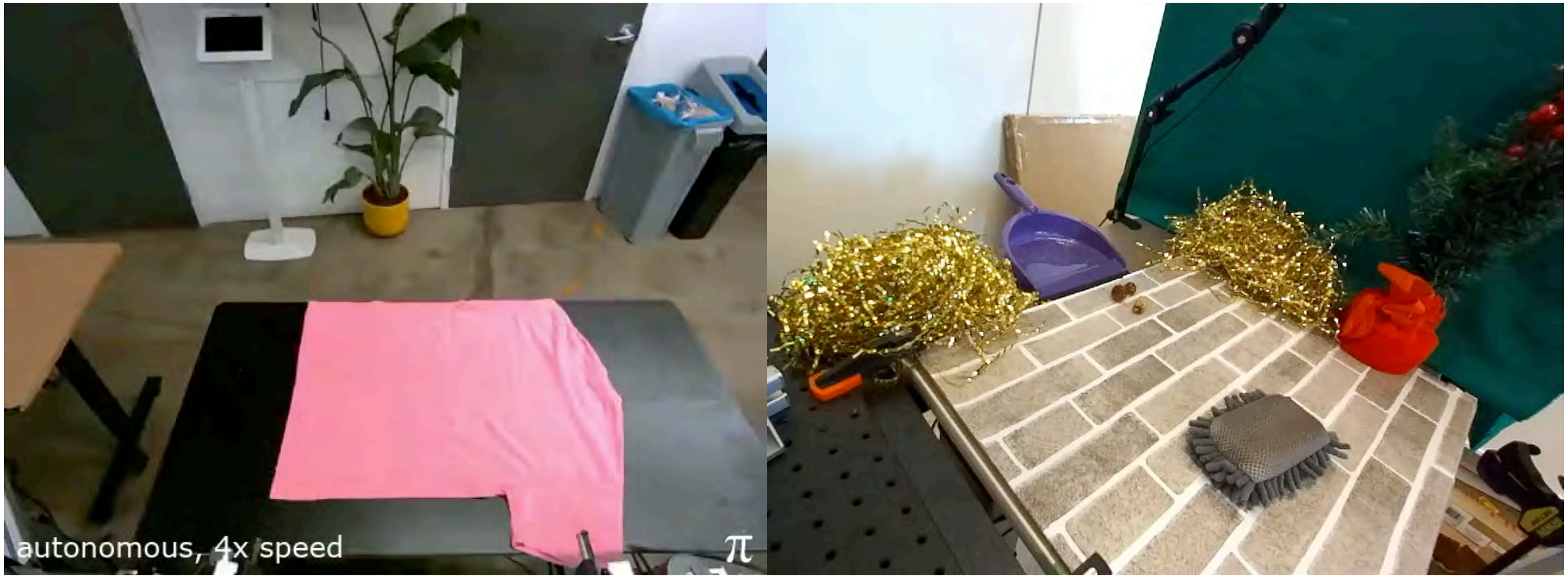




Learning diverse robot behaviors from Data

$$\pi_\theta(s_{:t}\ , a_{:t}\ ,\ g) \rightarrow a_t$$

Dataset $D = \{(s_i, a_i, g)\}$



[FAST: Efficient Robot Action Tokenization for Vision-Language-Action Models, Physical Intelligence, Pertsch et.al, 2025]

[OpenVLA: An Open-Source Vision-Language-Action Model, Kim et.al, 2024]

How do we learn diverse robot behaviors that can generalize to different environments?

With Low Data Collection Effort per Task

Why should we care about data efficiency?

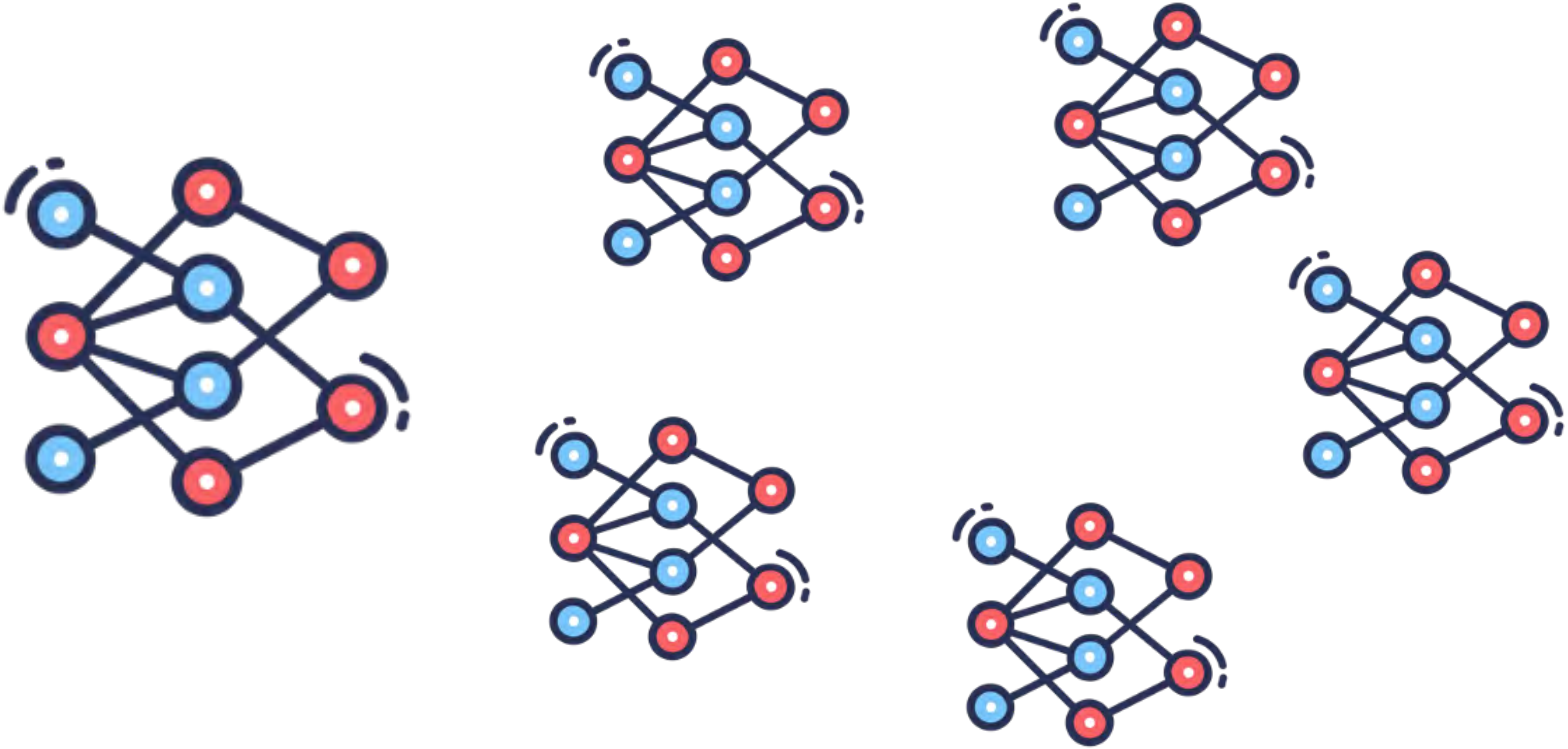
How do we learn diverse robot behaviors that can generalize to different environments?

With Low Data Collection Effort per Task

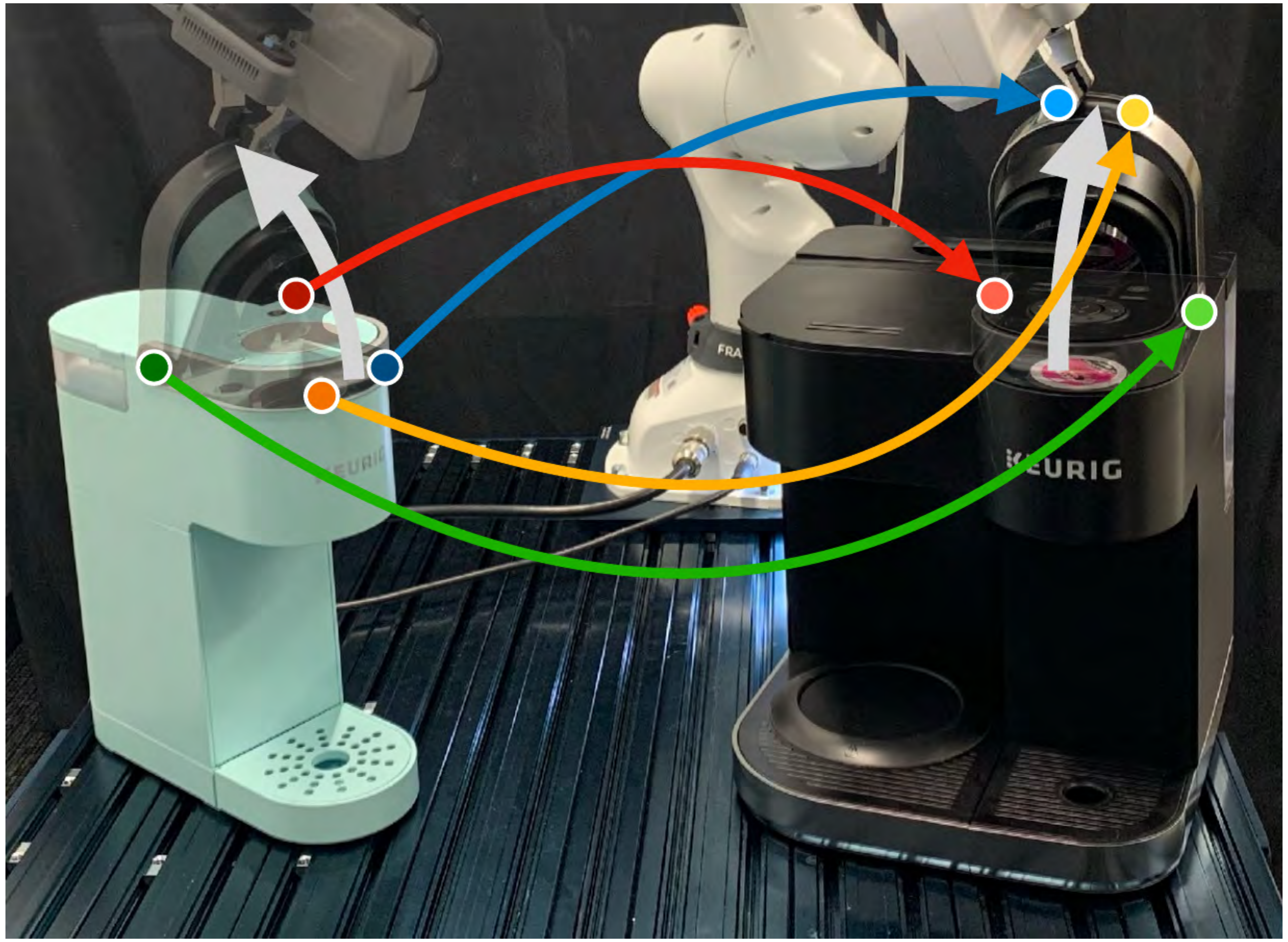
Less data requirement per task → More diverse behaviors

[Open X-Embodiment: Robotic Learning Datasets and RT-X Models, 2024]

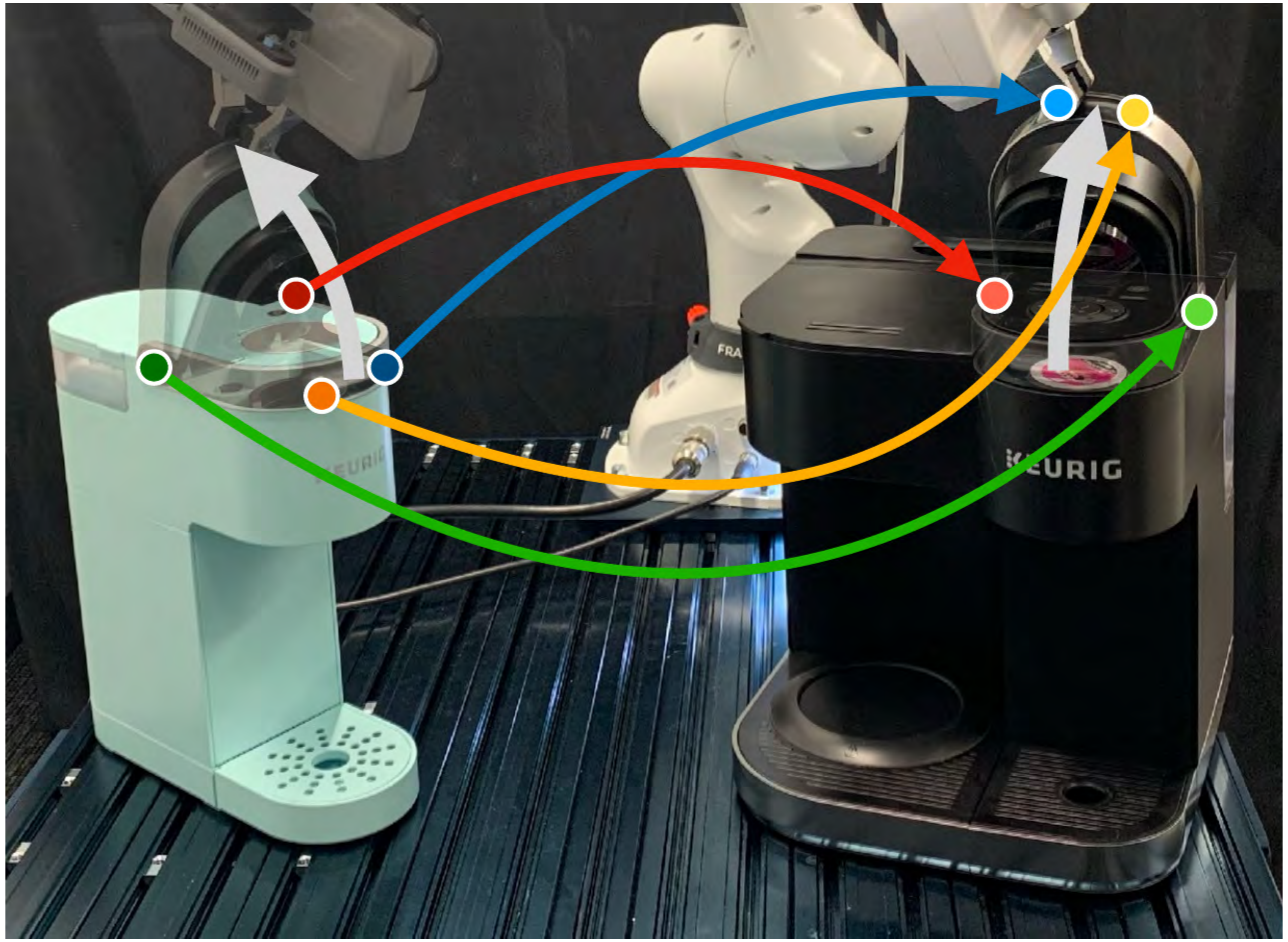
Task-specific Contextual Abstraction from Foundation Models



Data-Efficient Learning for a Single Task



KALM



Keypoint Abstraction using Large Models for Object-Relative Imitation Learning

Xiaolin Fang*, Bo-Ruei Huang*, Jiayuan Mao*, Jasmine Shone,
Joshua B. Tenenbaum, Tomás Lozano-Pérez, Leslie Pack Kaelbling


KEURIG
KEURIG
KEURIG

Task-specific Contextual Abstraction from Foundation Models

Task-specific Contextual Abstraction from Foundation Models

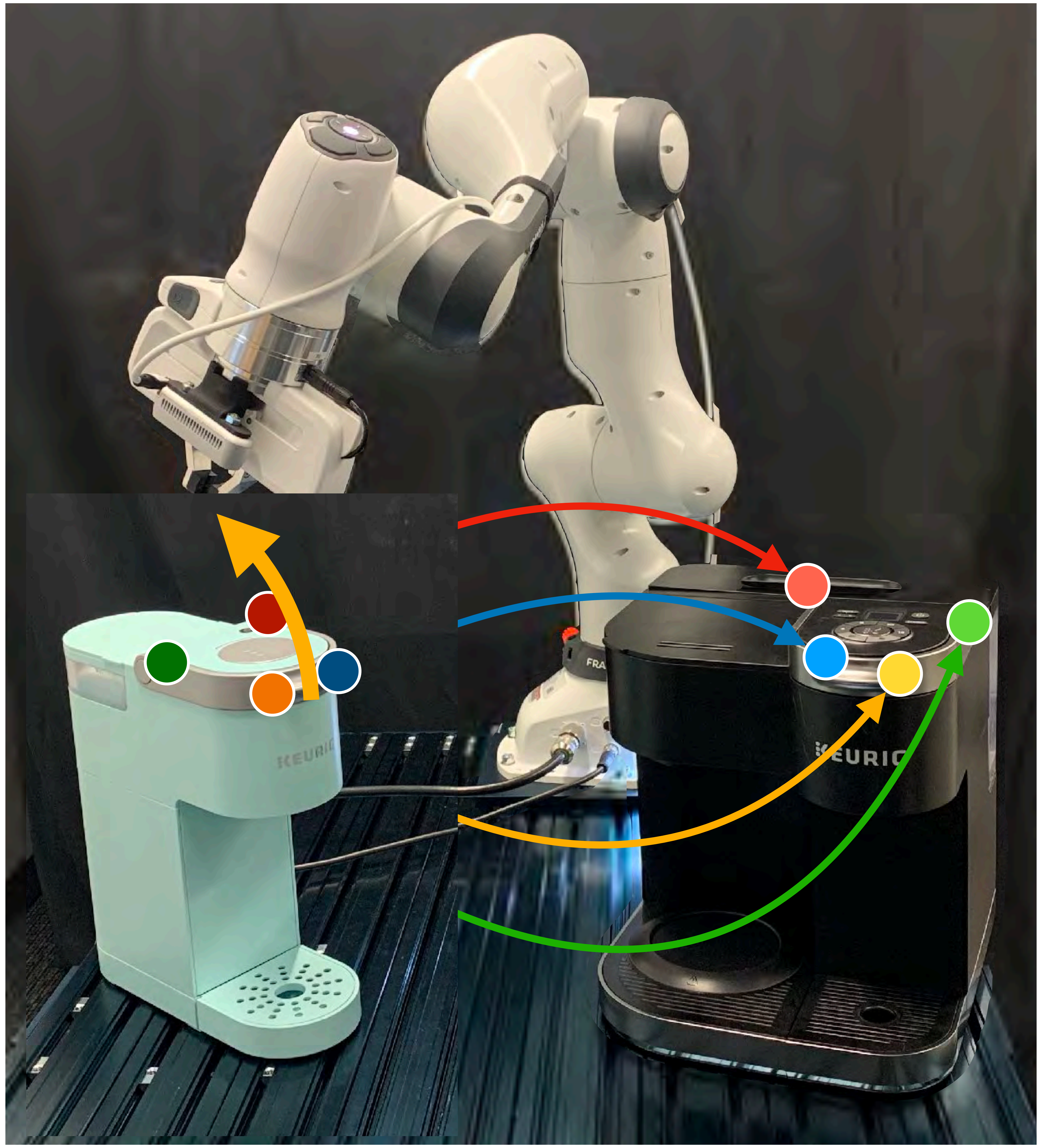
Semantics

Object Parts, Relations, Keypoints

Task-specific Contextual Abstraction from Foundation Models

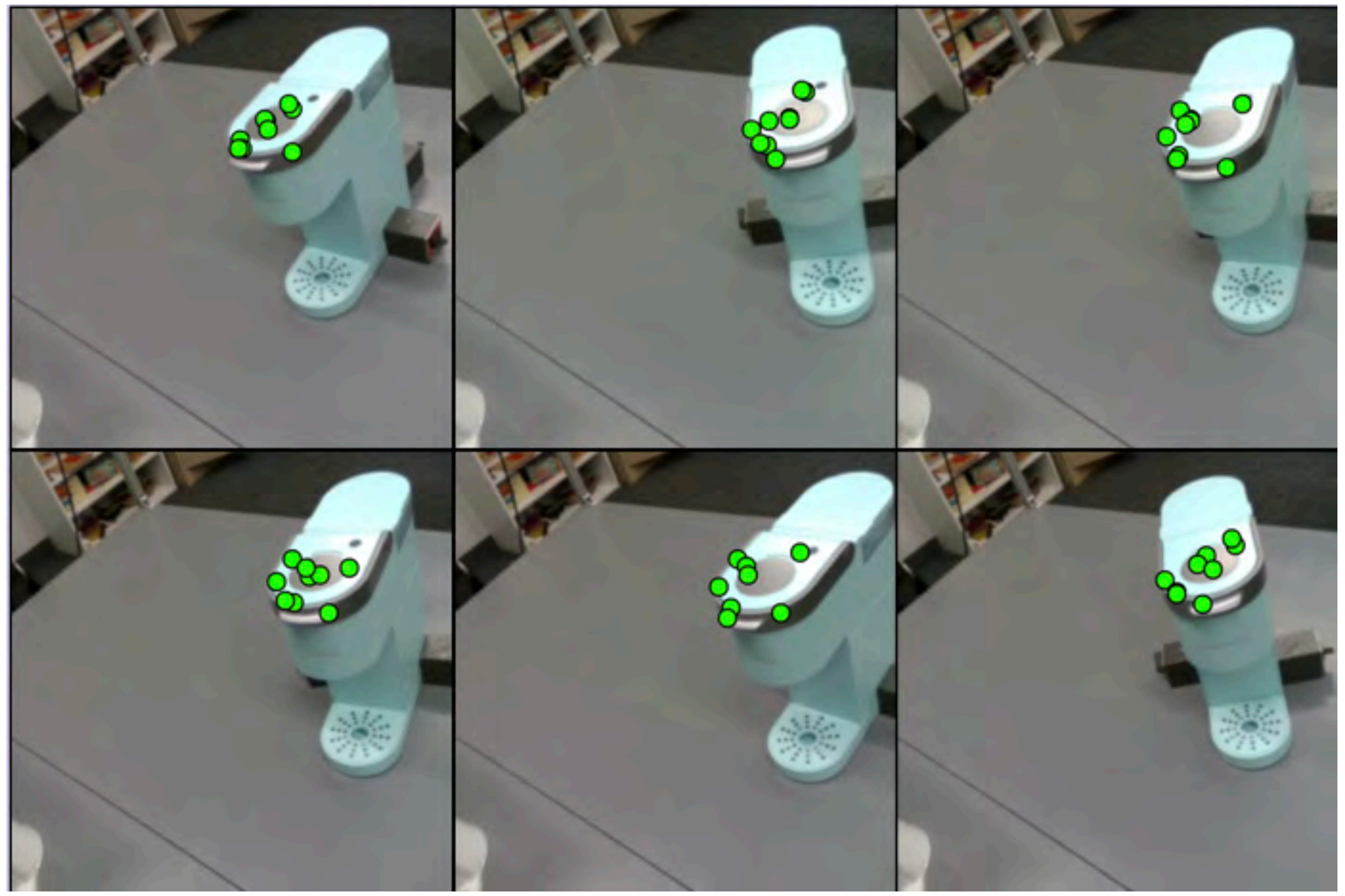
KALM: Keypoints as the contextual abstraction

- State Representation
  - Sparse and Local
- Frame for Action Representation
  - Object-Relative Action Frame



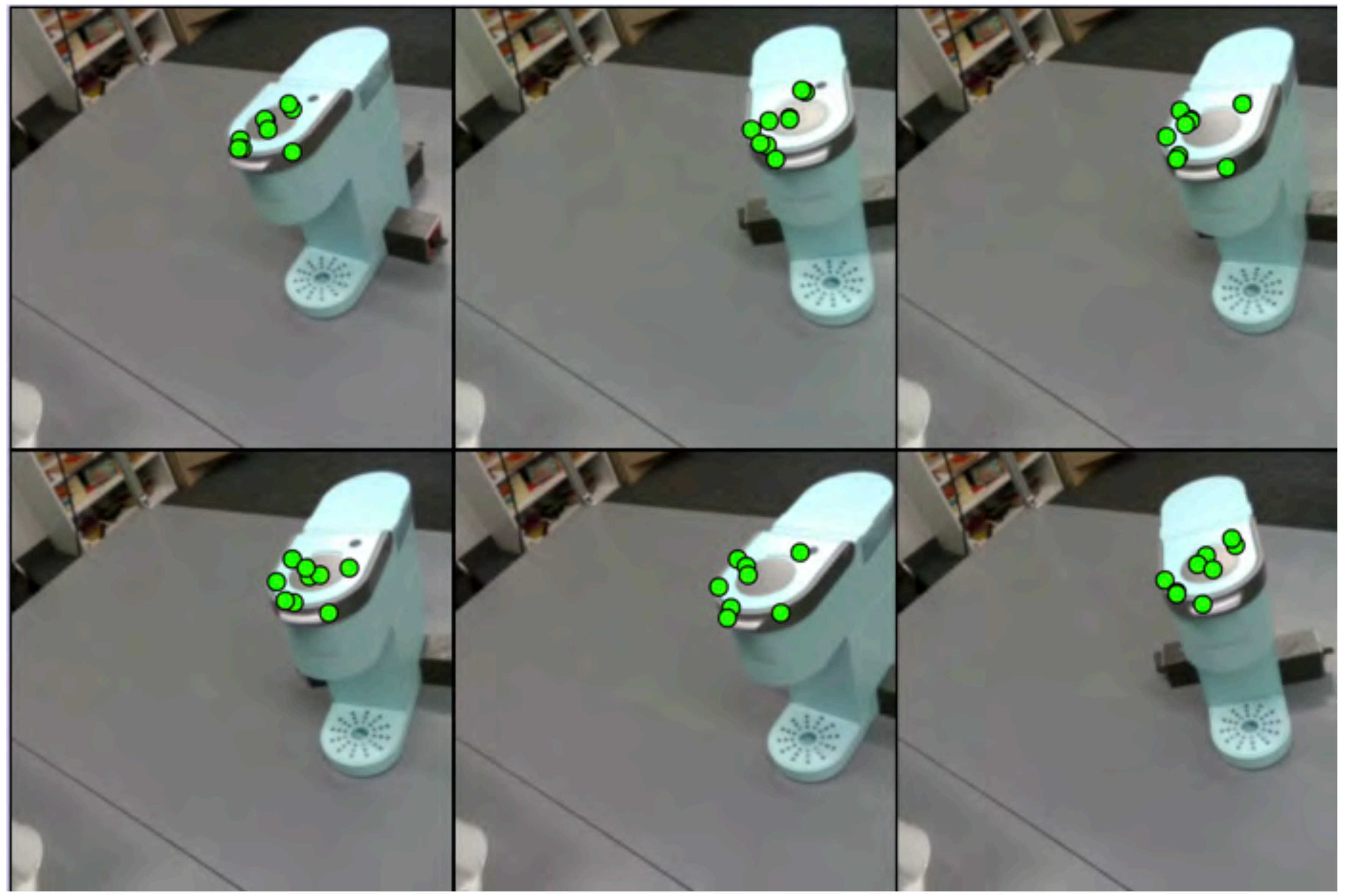
Problem Formulation

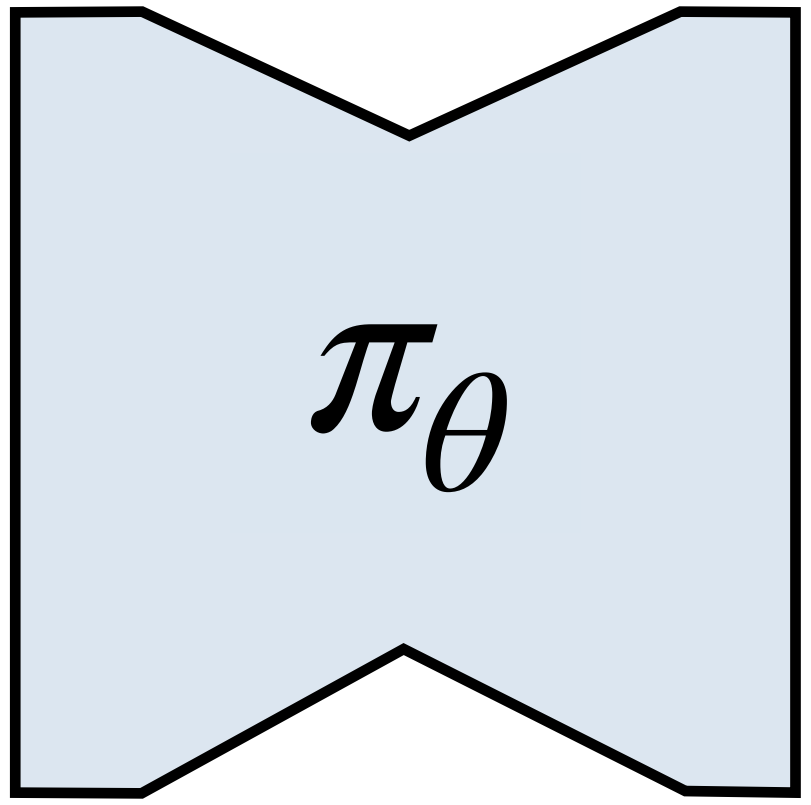
- Goal
  - **Keypoint Extraction** Identify a set of keypoints K_i in each image I_i in D



Problem Formulation

- Goal
  - **Keypoint Extraction** Identify a set of keypoints K_i in each image I_i in D
  - **Trajectory Prediction Model** Learn a keypoint-conditioned trajectory prediction model $\pi_\theta\ (\tau_i \mid K_i, F_i)$





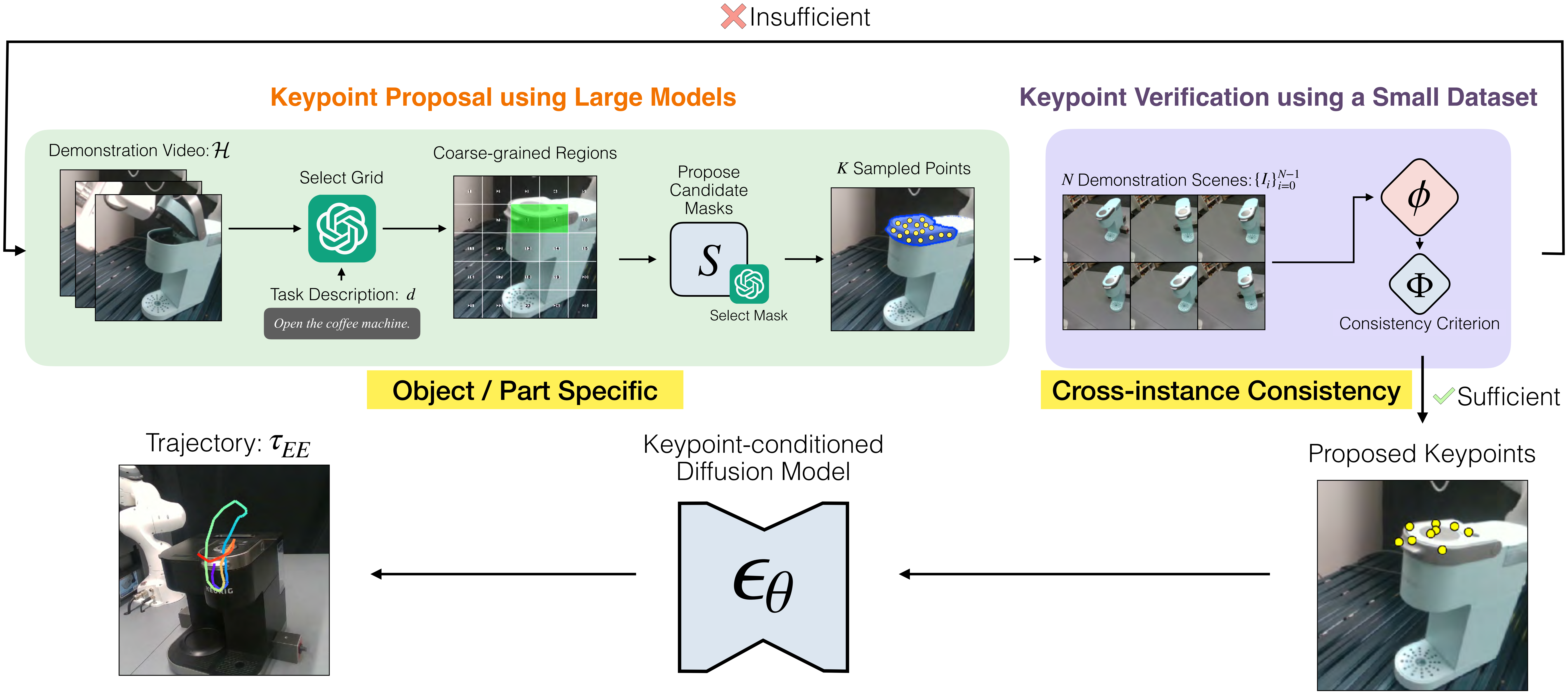
Keypoint-Conditioned
Diffusion Model

Key Contributions

- Keypoint Extraction using Large Models
- Keypoint-Conditioned Action Model

Key Contributions

- Keypoint Extraction using Large Models
- Keypoint-Conditioned Action Model


Insufficient
Keypoint Proposal using Large Models
Keypoint Verification using a Small Dataset
Demonstration Video: $\mathcal{H}$
Select Grid
Coarse-grained Regions
Propose Candidate Masks
K Sampled Points
Task Description: d
Open the coffee machine.
S
Select Mask
N Demonstration Scenes: $\{I_i\}_{i=0}^{N-1}$
ϕ
Φ
Consistency Criterion
Object / Part Specific
Cross-instance Consistency
Sufficient
Trajectory: τ_{EE}
Keypoint-conditioned Diffusion Model
ϵ_θ
Proposed Keypoints

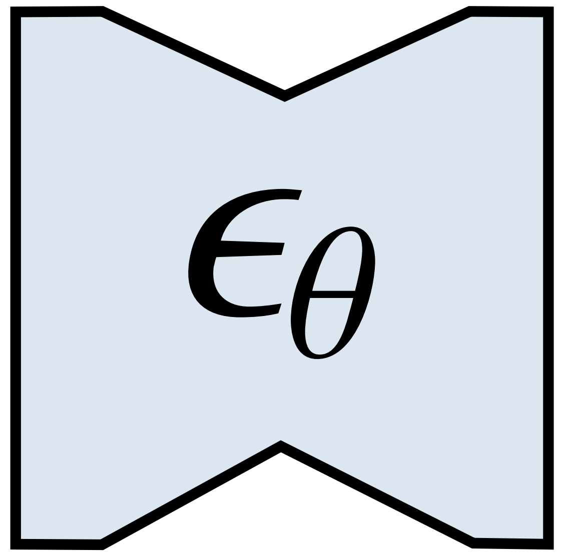
Key Contributions

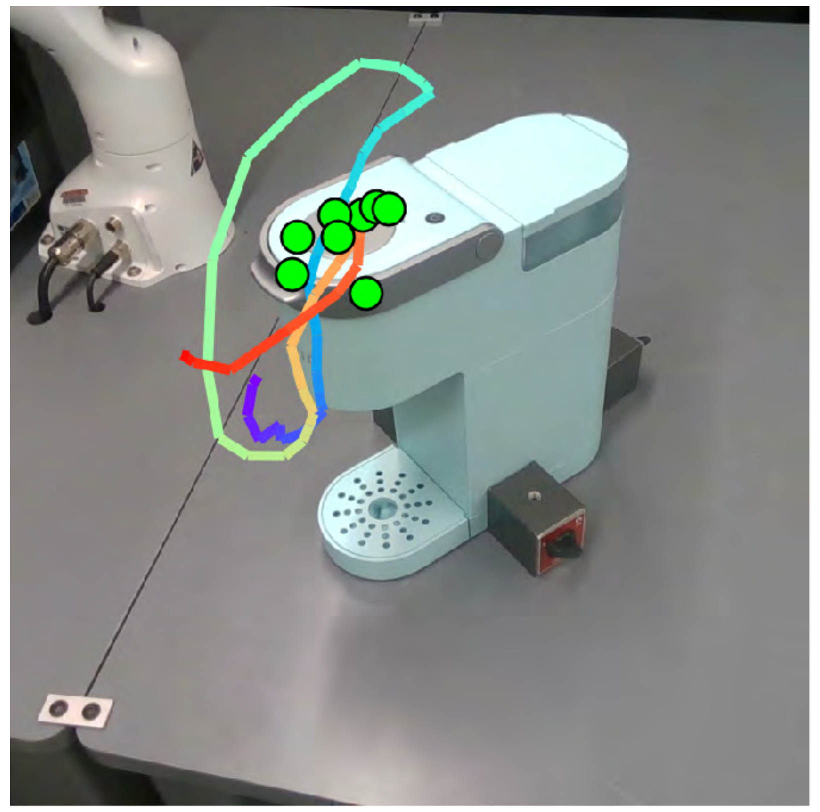
- Keypoint Distillation using Large Models
- Keypoint-Conditioned Action Model

Keypoint-Conditioned Action Model

Proposed Keypoints

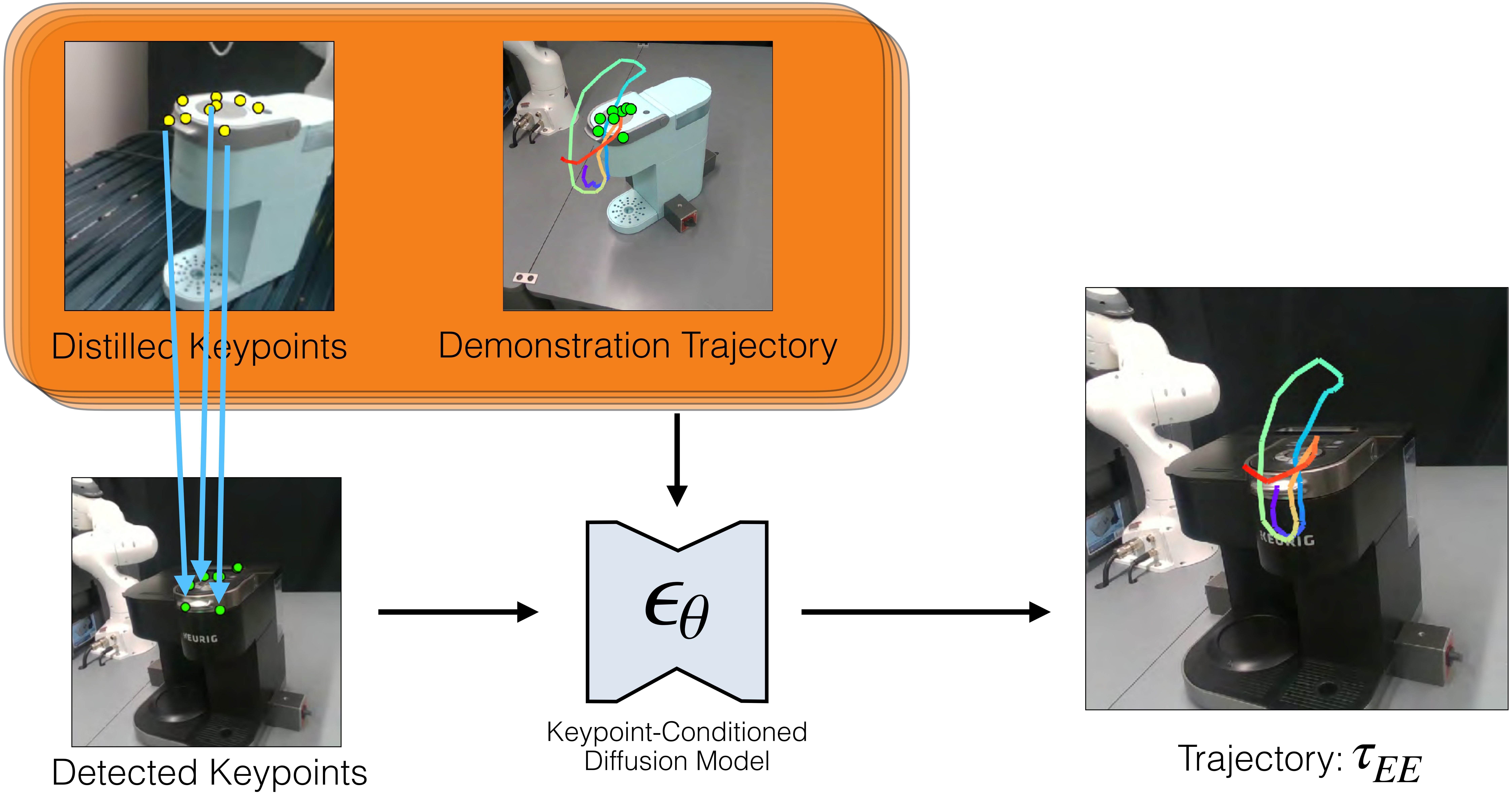
Keypoint-conditioned
Diffusion Model

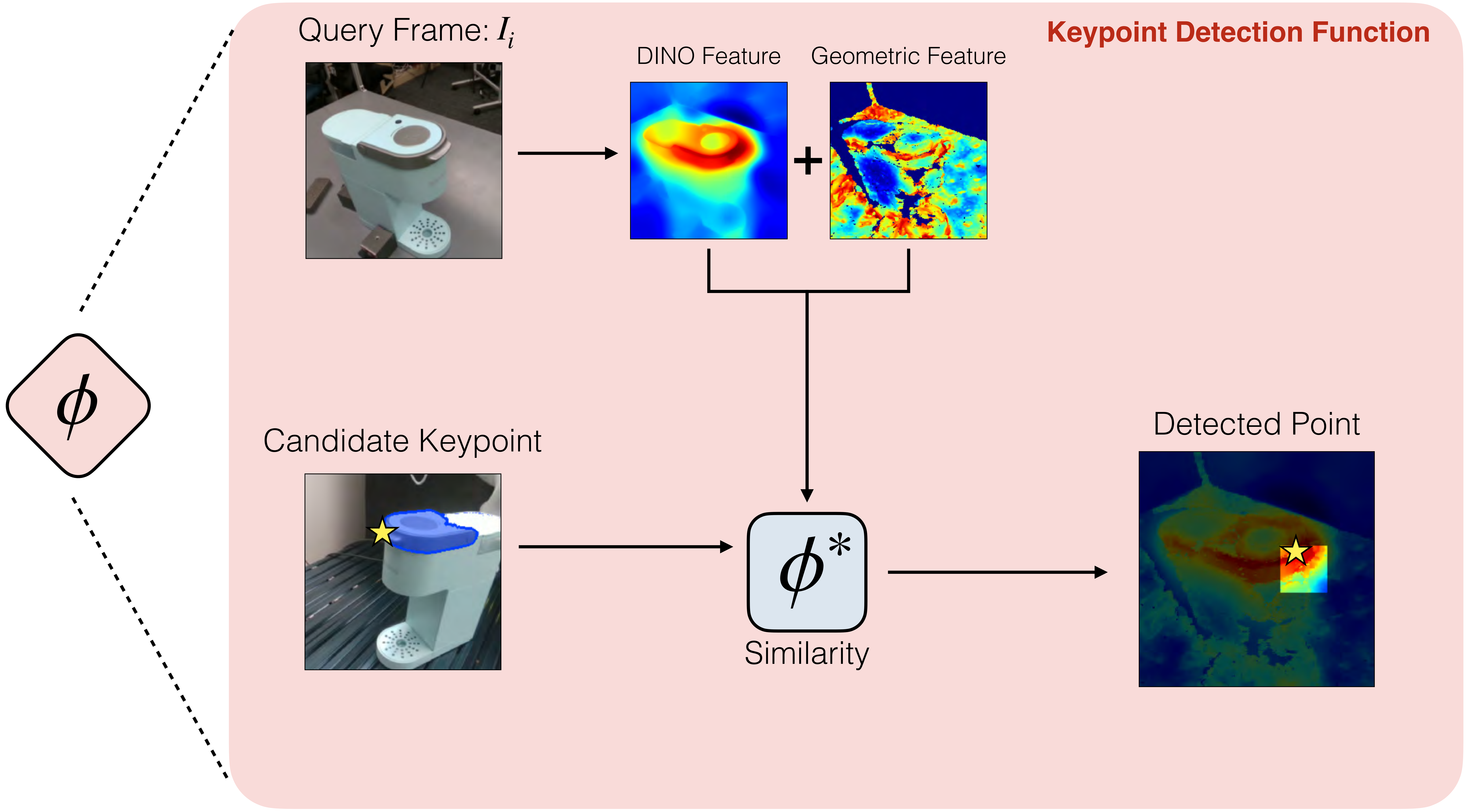




- Input
  - Keypoint locations $\{P_j\}^{|K|}$
  - Keypoint features $\{F_j\}^{|K|}$
- Output
  - A sequence of 6-DoF Poses of the robot's end-effector, relative to the center of key points

Keypoint-Conditioned Action Model





Keypoint Detection Function
Query Frame: I_i
DINO Feature
Geometric Feature
+
Candidate Keypoint
ϕ
ϕ^*
Similarity
Detected Point

Generalization

- Novel Objects
- Object Poses
- Camera Views
- Unseen Environments

Keypoints

- State Representation
    - Sparse and Local
- Action Representation
    - Object-Relative Action Frame

Generalization

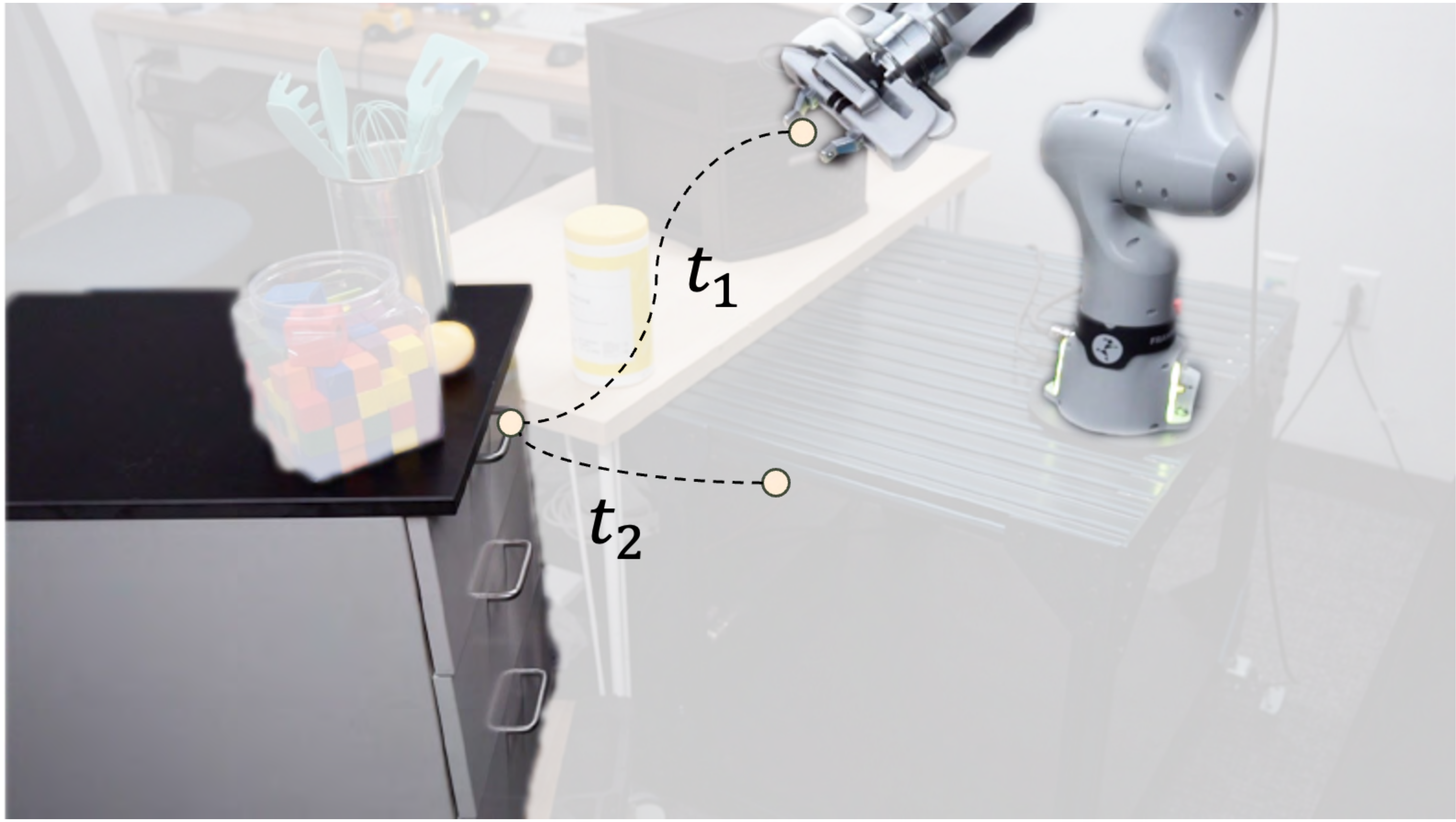
- Novel Objects
- Object Poses
- Camera Views
- Unseen Environments

Keypoints

- State Representation
  - Sparse and Local
- Action Representation
  - Object-Relative Action Frame

Action Model as Trajectory Sampler

Constraints in Novel Environments



Action Model as Trajectory Sampler

Constraints in Novel Environments

Action Model as Trajectory Sampler

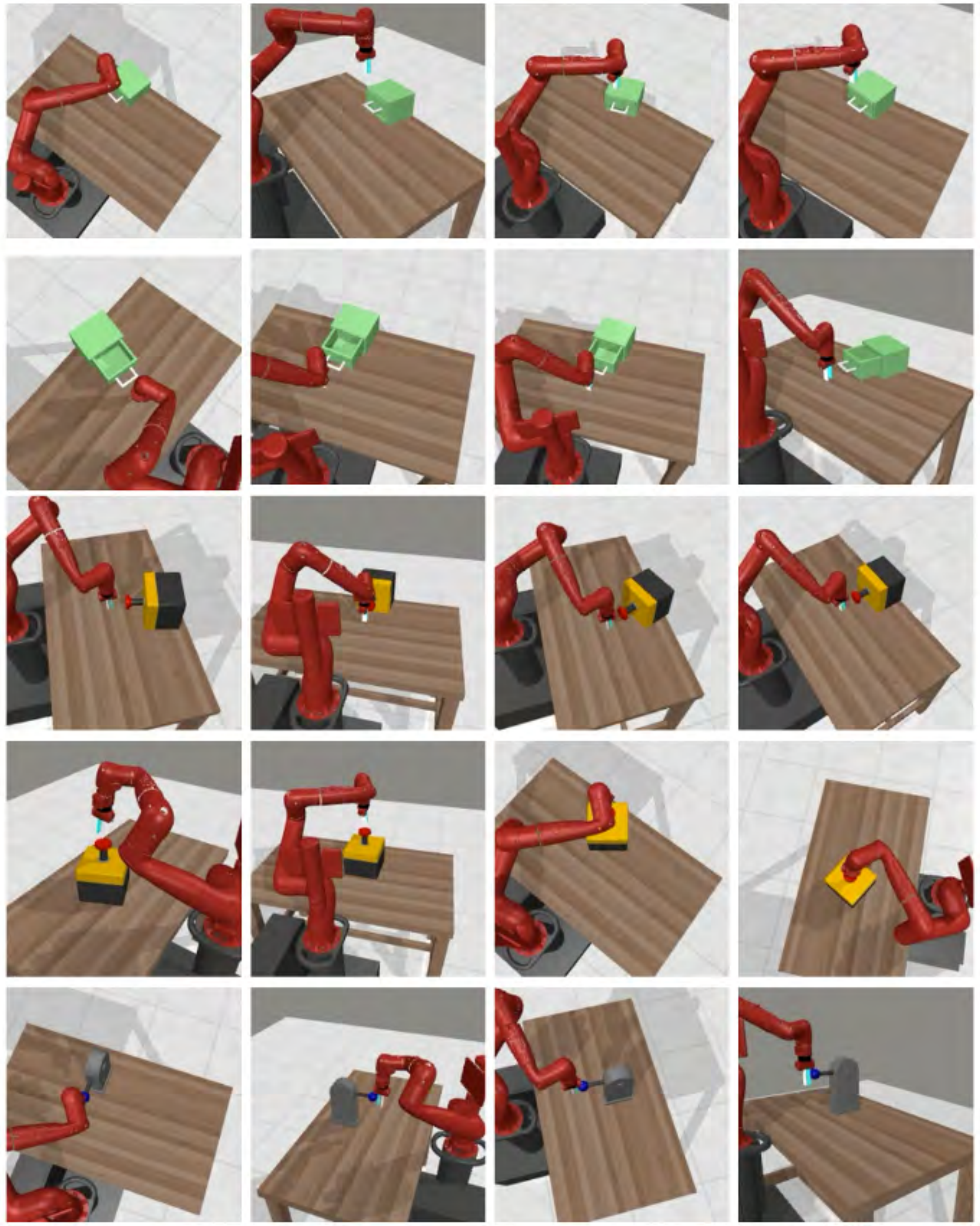
Constraints in Novel Environments

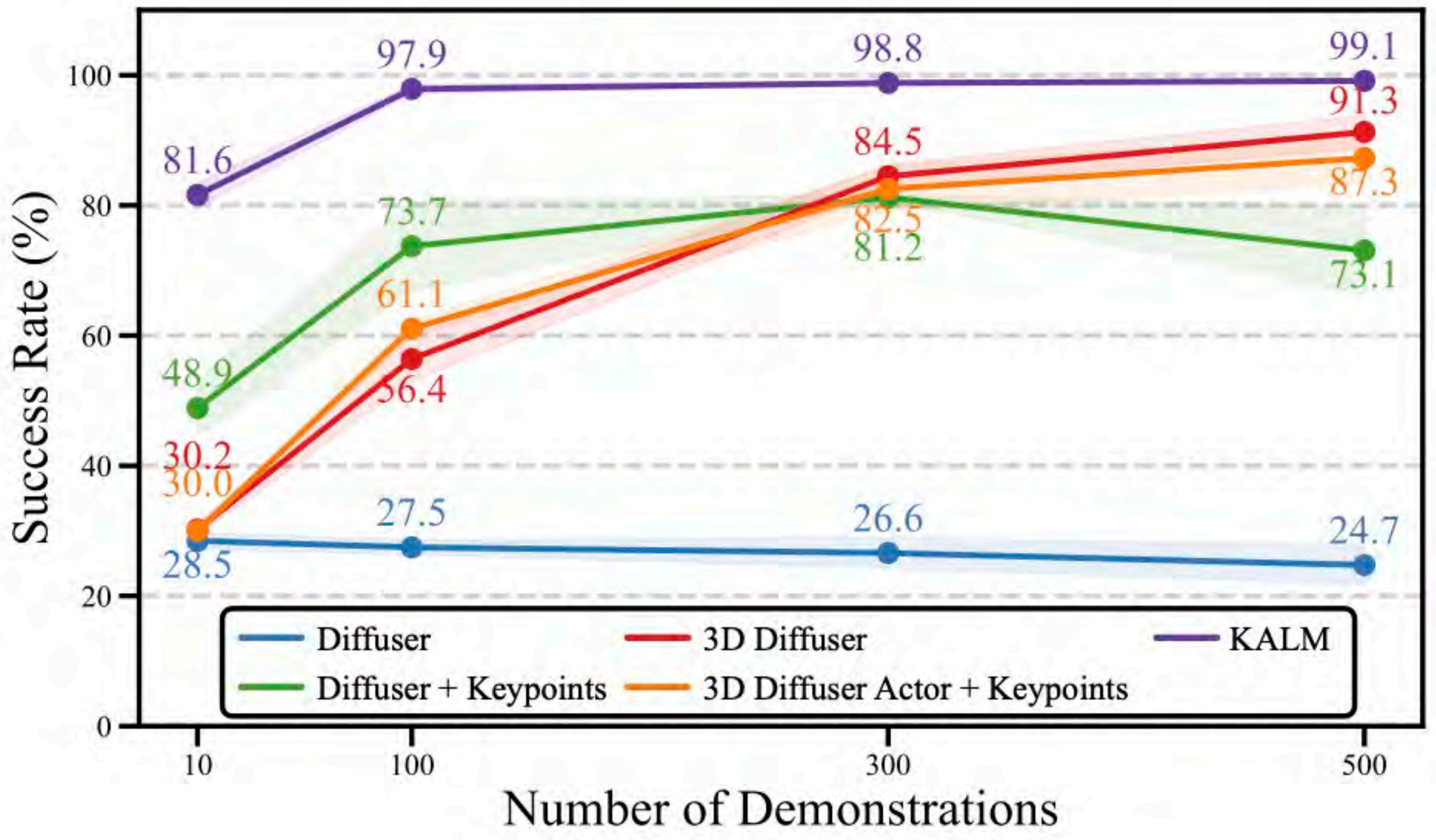
- Sample N trajectories $\{\tau_{EE}^{i}\}_{i=1}^{N}$
- Check whether a trajectory satisfy collision-constraint or robot kinematic constraints in the environment

Baselines

- Diffuser : RGB Image
- Diffuser + Keypoint : RGB Image and 2D Pixel Locations of the Keypoints
- 3D Diffuser Actor : RGB-D Image
- 3D Diffuser Actor + Keypoint: RGB-D Image and 3D Positions of the Keypoints
- KALM (Ours) : 3D Keypoint Locations and Keypoint Features

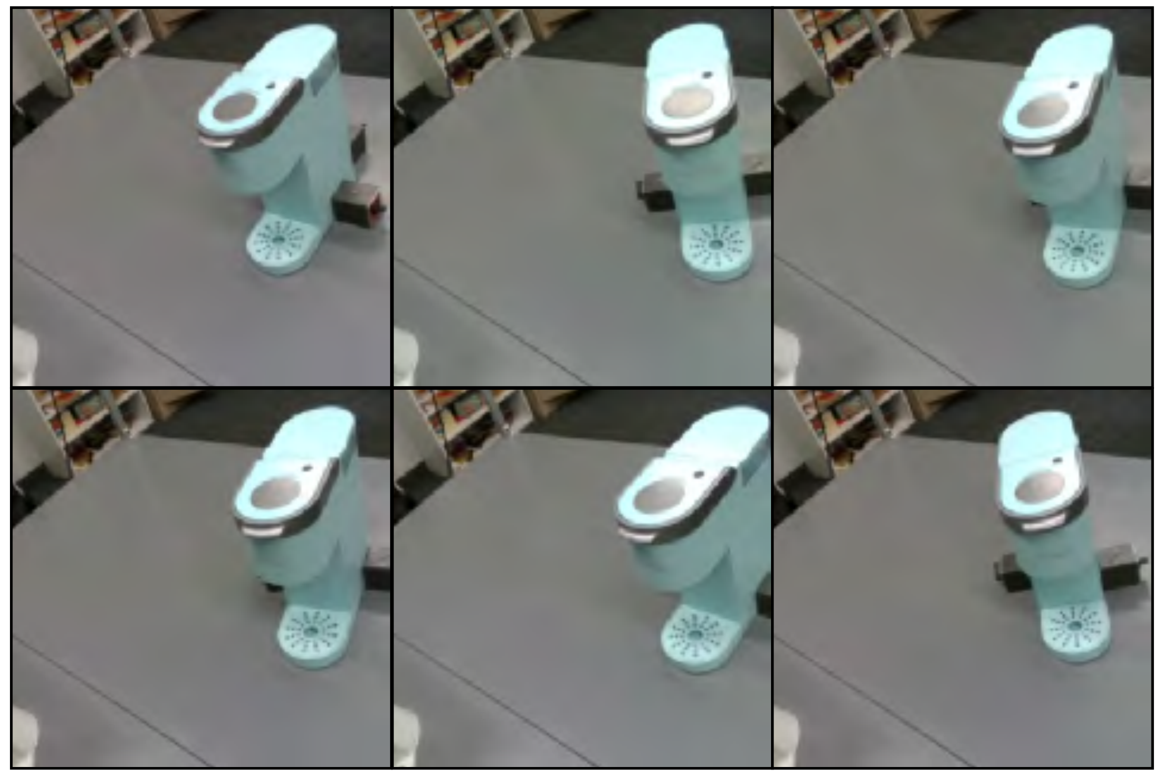
Contextual Abstraction Improves Data Efficiency





Contextual Abstraction Enables Generalization

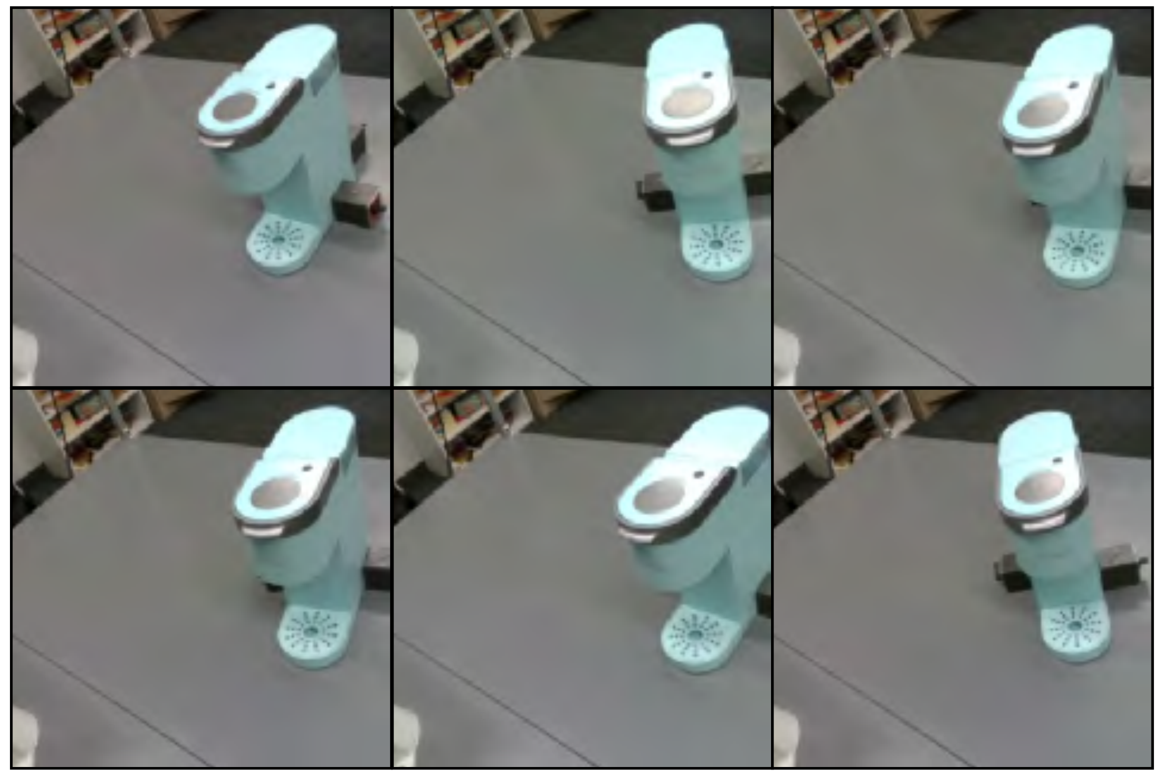
Training Scenes



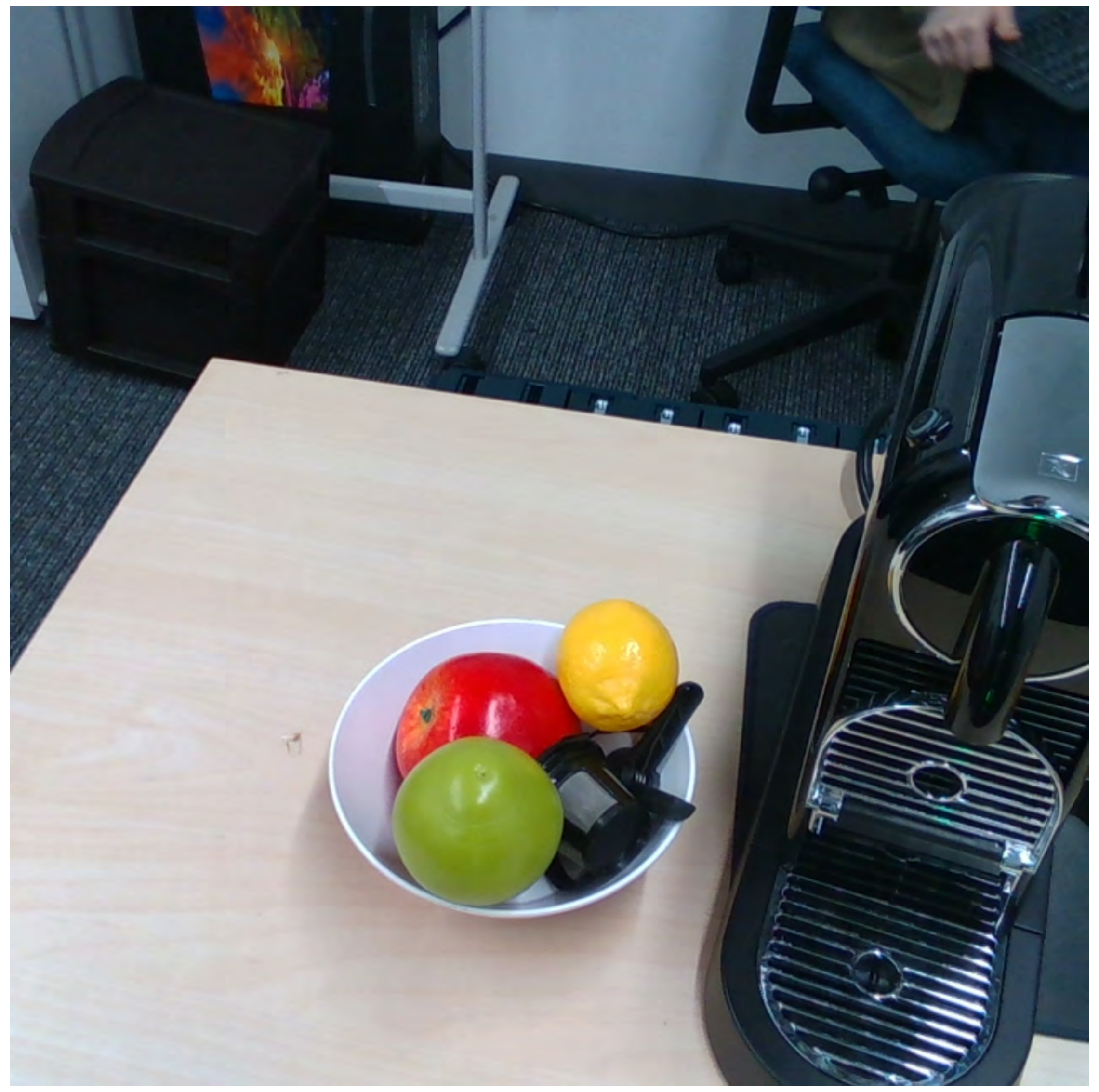
Distilled Keypoints

Contextual Abstraction Enables Generalization

Training Scenes



Distilled Keypoints

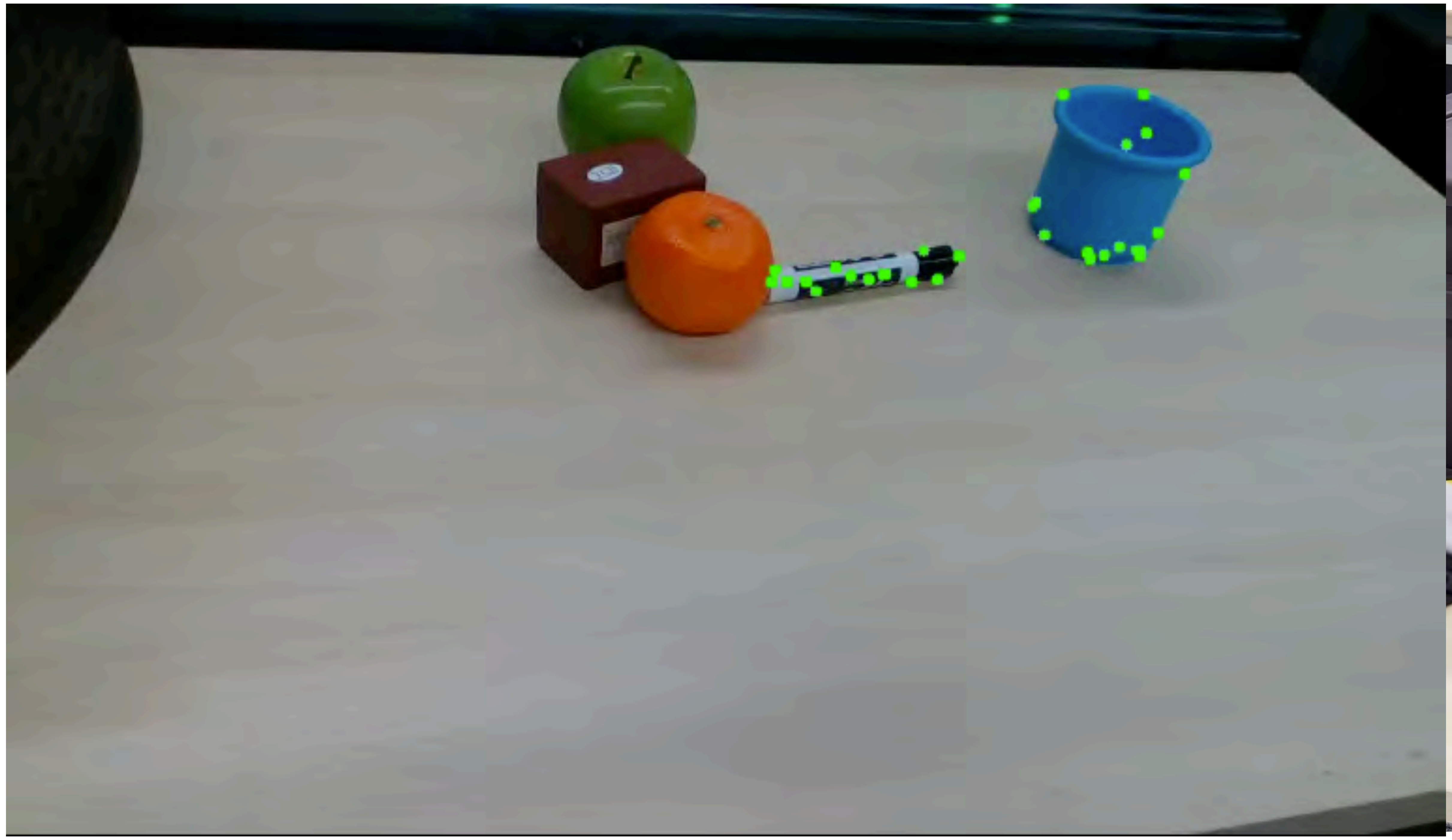


Contextual Abstraction Enables Generalization

Contextual Abstraction Enables Generalization

Tasks	Without Verification		KALM (Ours)	
	View	Cross Object	View	Cross Object
Lifting Handle	1/10	0/10	9/10	6/10
Opening Drawer	2/10	0/10	6/10	7/10
Pouring into Bowl	6/10	2/10	8/10	6/10

Reactive Policy - Real-Time Tracking



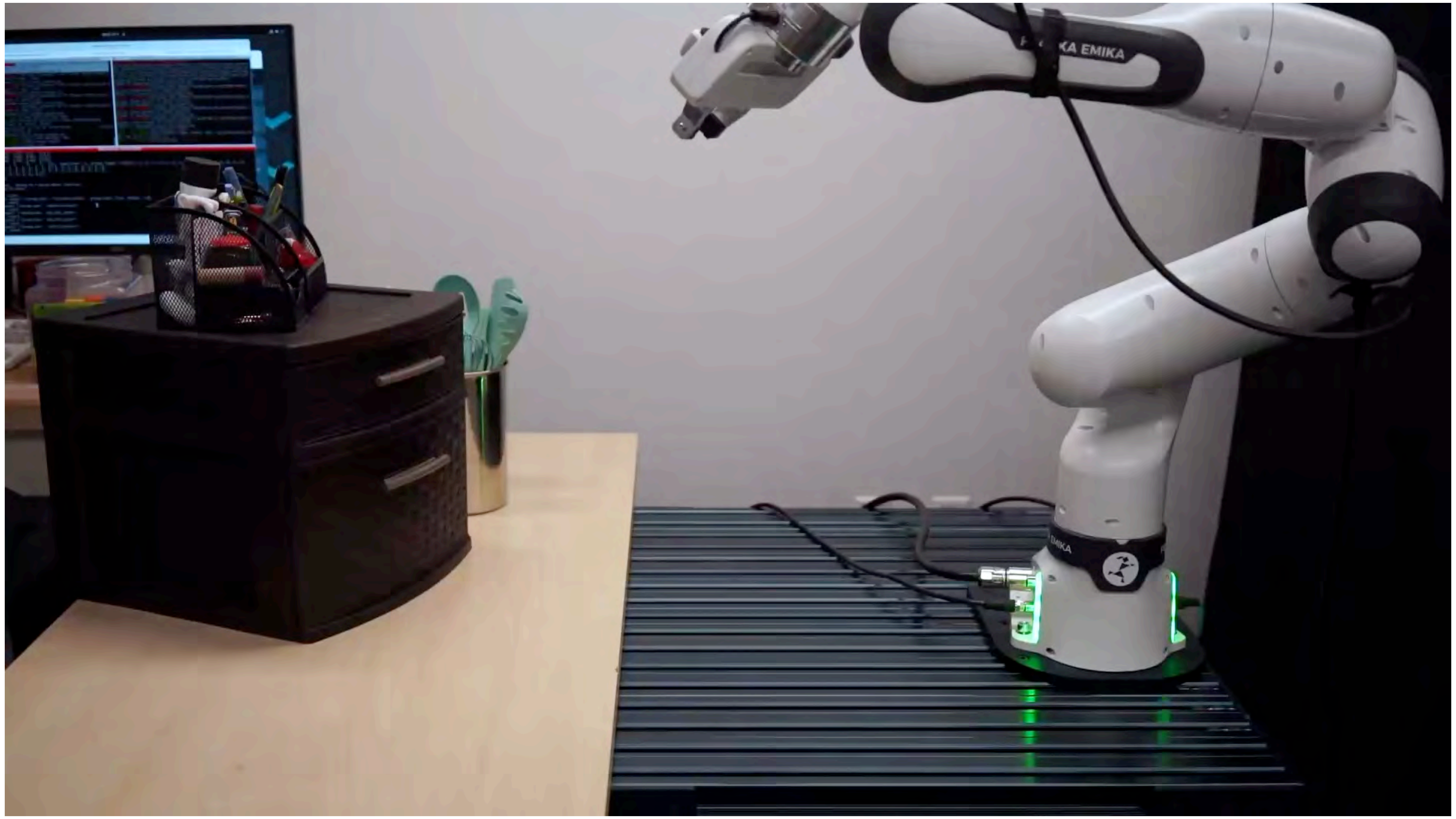
Conclusions

- Leveraging foundation models - let them do the tasks they are good at

- Sparse representation enables data-efficient imitation learning

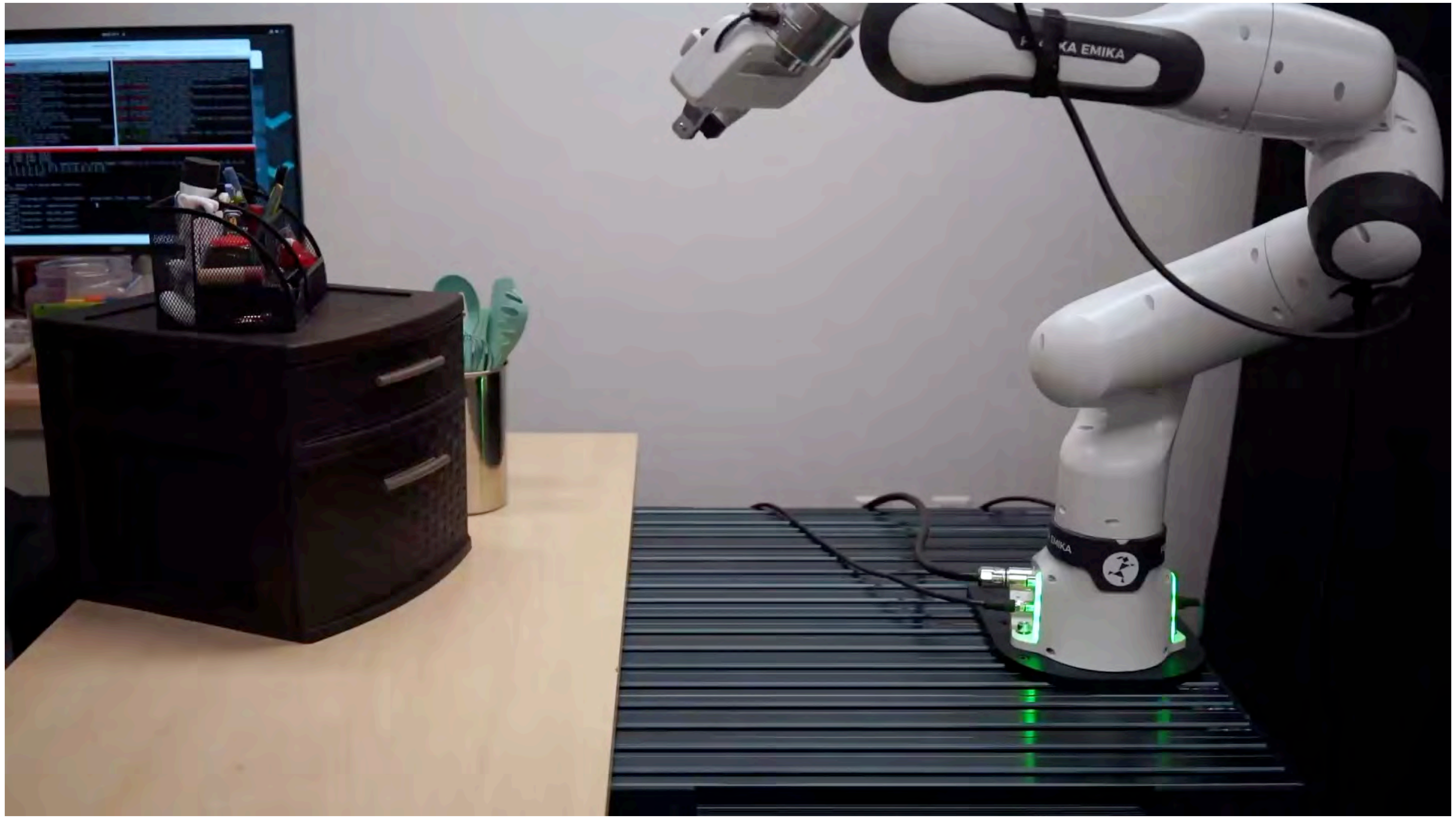
- Pick your space

Open Problems



- Implicit Constraints beyond Demonstration Data: Inferring latent rules
- Abstraction vs. Fidelity: Choose the right level for the task

Open Problems



- Implicit Constraints beyond Demonstration Data: Inferring latent rules
- Abstraction vs. Fidelity: Choose the right level for the task

← Foundation Models?

Conclusions

- Leveraging foundation models - let them do the tasks they are good at
- Sparse representation enables data-efficient imitation learning
- Pick your space