Self-supervised Learning of Action Affordances as Interaction Modes

actaim.github.io

Liquan Wang, Nikita Dvornik, Rafael Dubeau, Mayank Mittal, Animesh Garg

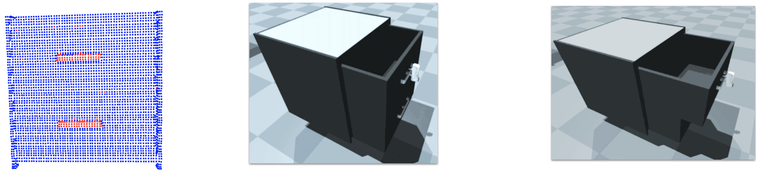
UNIVERSITY OF TORONTO · VECTOR INSTITUTE · ETH zürich

ICRA LONDON 2023

Problem Definition

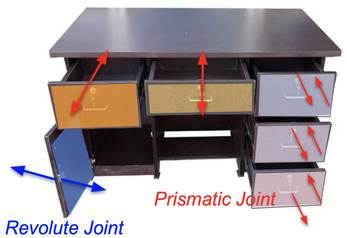
Given an articulated object

- ***Where and how*** we can interact with it
- Discover different ***ways of interactions***



Challenges

① ***Joints***



Different interaction mode varies from ***different types*** and ***numbers*** of joints

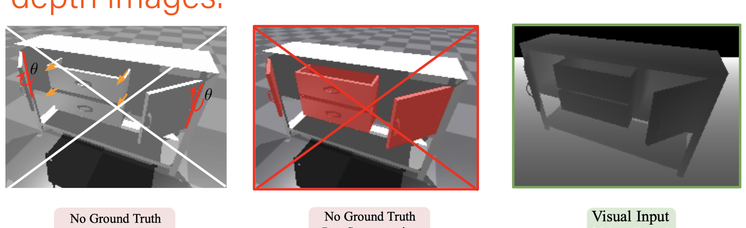
- Left table with 6 different drawers
- Different Types of Motions (Revolute, Prismatic, ⋯)

② ***Generalization***

Generalization over different objects/categories

③ ***privileged information***

Does not require any privileged information including ***object's state information, rewards function computation, or part segmentation***. We only use depth images.



Insight: Interaction Mode

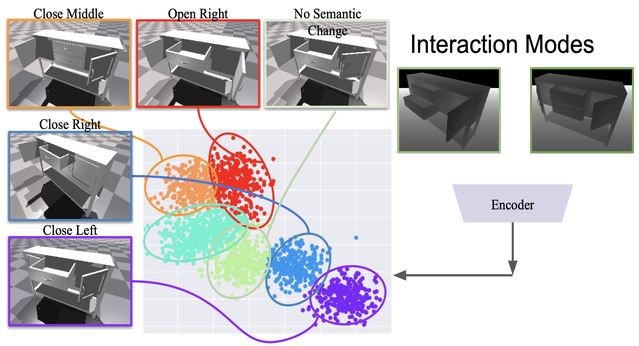
Our Ideas

- ***Interaction modes*** can be defined by pure vision
- Decompose policy into ***action predictor*** and ***mode selection***

$$\mathbb{P}(a|o) = \underbrace{\mathbb{P}(a|o,z)}_{\text{action predictor}} \quad \underbrace{\mathbb{P}(z|o)}_{\text{mode selector}}$$

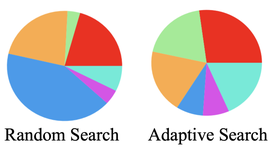
- ***Mode Selector*** Which Interaction mode an object offers implicitly
- ***Action Predictor*** Where and How to interact with the articulated object

Adaptive Data Collection



Interaction modes can be ***distinguished and clustered*** using the differences between initial and final depth image encoding

Using GMM ***adaptive data collection*** to balance the data across different interction modes to avoid generating only fewer rare interaction modes.



Model Training

ActAIM inputs:

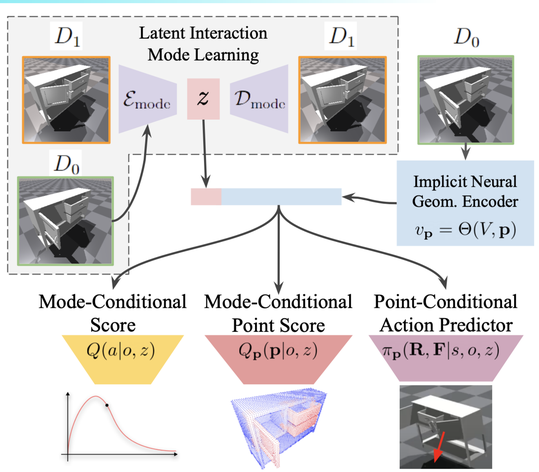
- Initial and final depth images

ActAIM outputs:

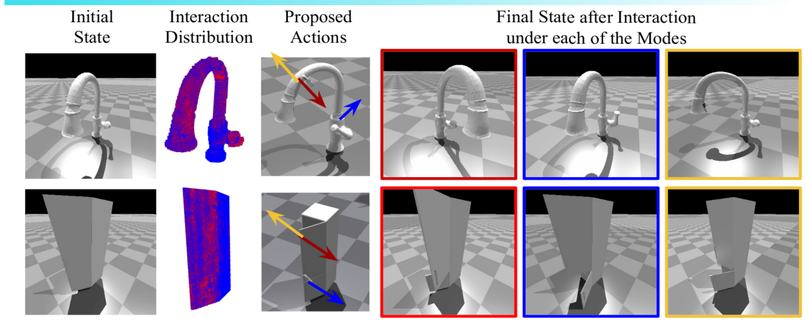
- Mode score
- Mode-conditional interaction point distribution
- Point-conditional action

Improve Scalability:

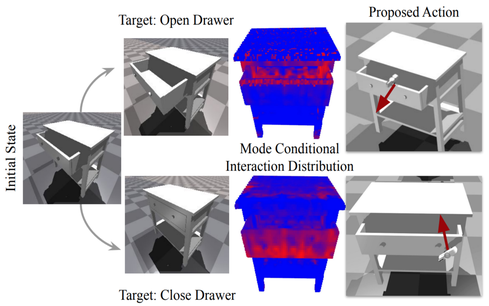
Implicit Neural Geometry (Local Geometry Feature)



Qualitative results



Generative results illustrate valid ***interaction point distribution, interaction modes,*** and ***correpondent low-level action***



Goal-Conditional results illustrate ActAIM takes in ***initial and target depth images***, and predicts ***mode-conditional*** interaction point distribution and correspondent low-level action

Quantitative results

Categories: faucet, table, cabinet, door, window, fridge, kitchen pot, kettle,

Test Set	Unseen States of Training Objects							Unseen instances of Training Categories							Unseen Categories			
Sample Success Rate % ↑																		
Random Interaction	5.61	6.05	8.47	14.59	18.15	5.66	9.76	7.99	3.38	6.47	13.86	20.92	4.51	9.52	13.32	10.42	6.04	9.93
Where2Act [6]	33.32	7.05	7.05	17.88	11.64	4.07	13.50	**32.99**	13.78	6.94	18.89	15.00	5.42	15.50	18.43	9.49	4.14	5.34
ActAIM-PN++	41.25	**44.92**	32.46	21.64	46.27	19.52	34.34	21.04	31.21	31.91	19.21	36.14	12.43	25.32	**21.37**	18.34	21.61	20.44
ActAIM **[ours]**	**49.26**	41.36	**36.21**	**28.64**	**58.31**	**19.68**	**38.91**	21.98	**38.10**	**35.54**	**21.03**	**41.61**	**16.19**	**29.07**	21.09	**24.68**	**24.13**	**23.30**
Weighted Modes Ratio % ↑																		
Random Interaction	5.61	5.27	7.62	12.77	15.61	5.26	8.69	4.47	3.12	6.13	9.73	16.72	3.92	7.35	13.32	10.23	5.87	9.81
Where2Act [6]	11.77	6.06	6.25	14.50	8.59	3.51	8.44	10.86	9.71	6.02	11.00	10.63	5.17	8.89	18.43	8.81	3.91	5.19
ActAIM-PN++	29.11	25.52	20.48	12.81	45.54	17.48	25.16	14.28	24.18	26.66	17.29	25.92	10.44	19.79	18.51	15.76	16.09	16.79
ActAIM **[ours]**	**39.20**	**36.49**	**25.56**	**18.76**	**57.42**	**17.29**	**32.45**	**15.12**	**34.58**	**32.55**	**18.90**	**33.21**	**14.56**	**24.82**	**18.92**	**17.68**	**17.31**	**17.97**
Weighted Normalized Entropy % ↑																		
Random Interaction	5.19	4.45	6.80	10.49	10.41	4.18	6.92	7.09	2.82	4.89	5.74	11.30	3.01	5.81	10.02	7.41	5.79	7.74
Where2Act [6]	12.12	5.08	5.41	9.03	5.15	3.23	6.67	**15.62**	8.31	5.93	6.75	5.30	4.23	7.68	**17.84**	7.60	3.91	4.89
ActAIM-PN++	24.60	38.28	28.48	17.85	32.66	8.74	25.10	6.51	13.02	16.43	7.22	14.76	6.00	10.66	15.78	12.34	12.40	13.51
ActAIM **[ours]**	**34.79**	**36.49**	**35.64**	**25.48**	**41.76**	**9.31**	**30.58**	7.14	**19.38**	**24.15**	**7.77**	**19.27**	**8.16**	**14.31**	16.31	**16.17**	**15.58**	**16.02**

ActAIM outperforms baselines on ***sample success rate*** and preditive interction mode ***diversity***

sample-success-rate

$$ssr = \frac{\text{\# successful proposed interaction}}{\text{\# total proposed interaction}}$$

weighted modes ratio

$$\eta = ssr \times \frac{\text{\# successful discovered mode}}{\text{\# total GT modes}}$$

weighted normalized entropy

$$\tilde{\mathcal{H}} = ssr \times \frac{\mathcal{H}(\mathcal{M})}{\mathcal{H}_{max}}$$

Test Set		Unseen States of Training Objects							Unseen instances of Training Categories							Unseen Categories			
Eval Task	**Sample Success Rate % ↑**																		
Dec-DoF	Where2Act-Push	**26.54**	21.12	4.02	23.77	20.27	5.17	16.82	10.12	15.54	15.13	**43.04**	9.42	6.63	16.64	2.52	**23.90**	**43.52**	23.31
(Common)	ActAIM **[ours]**	25.32	**37.45**	**19.31**	**62.91**	**67.32**	**61.23**	**45.59**	**20.31**	**36.31**	**18.24**	41.21	**29.31**	**31.42**	**29.47**	**15.17**	22.50	32.51	**23.39**
Inc-DoF	Where2Act-Pull	12.52	0.27	0.42	0.02	0.98	0.06	2.38	0.00	1.56	0.00	0.04	0.46	0.00	0.34	2.79	0.06	**37.52**	13.46
(Rare Mode)	ActAIM **[ours]**	**24.15**	**16.21**	**11.34**	**28.14**	**49.31**	**17.56**	**24.45**	**12.41**	**14.25**	**9.48**	**10.46**	**15.98**	**13.12**	**12.62**	**7.84**	**13.12**	21.41	**14.12**

ActAIM dominates baseline on sample success rate in ***goal-conditional*** sampling