VoxAct-B: Voxel-Based Acting and Stabilizing Policy for Bimanual Manipulation

I-Chun Arthur Liu, Sicheng He, Daniel Seita*, Gaurav S. Sukhatme*

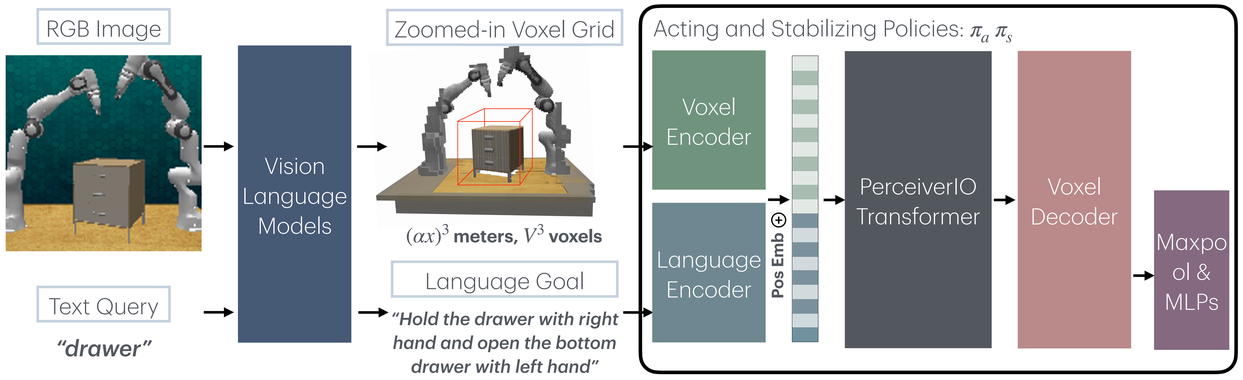
USC University of Southern California

CoRL

Motivation

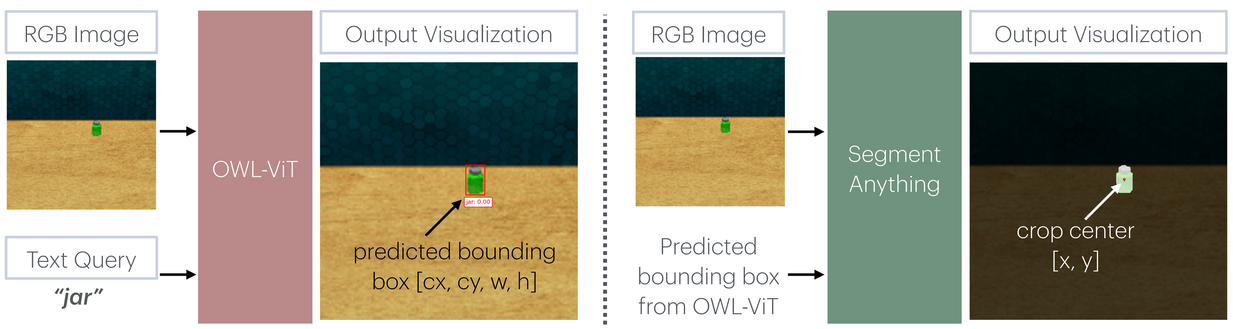
- Bimanual manipulation tasks are challenging due to high dimensional action spaces.
- Prior methods train policies on large datasets or exploit primitive actions.
- How can we design a sample-efficient method without relying on primitive actions?
- A voxel-based method that leverages Vision Language Models (VLMs) to prioritize key regions within the scene and reconstruct a voxel grid.

Method

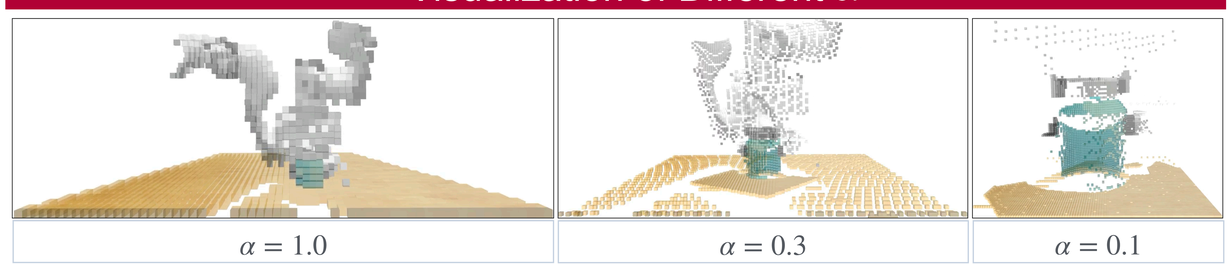


- **Input:** RGB-D images, language goals, and proprioception data.
- VLMs output pose of the object. We use this information to determine the language goal, roles of each arm (i.e., acting or stabilizing), and where to reconstruct a voxel grid.
- **Output:** Discretized pose of the next best voxel, gripper action, collision avoidance flag, and arm ID.

VLMs



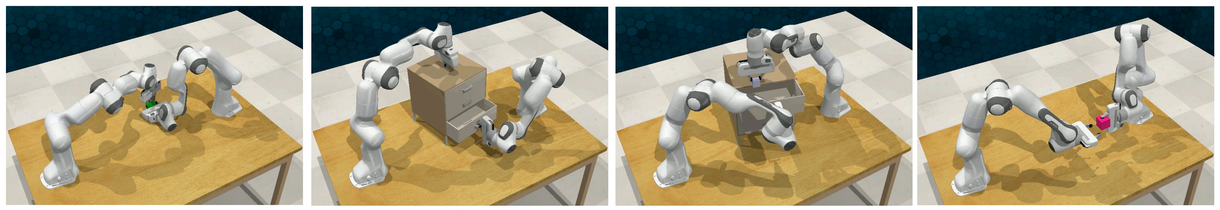
Visualization of Different α



Acting and Stabilizing Policies π_a π_s

- *Acting and stabilizing formulation* exploits the discretized action space that predicts the next best voxel with spatial equivariance properties, enabling more efficient learning from multi-modal demonstrations.
- *Learning to predict arm ID* allows policies to learn to map the appropriate acting or stabilizing actions to a given arm during training.

Environments



Simulation Results

- Half of the training data has the left arm as acting, and the other half has the right arm as acting. Methods are evaluated on 25 episodes of unseen test data.

(1) Comparison with Baselines (2) Multi-Task Results (3) Ablations

(1)

Method	Open Jar		Open Drawer		Put Item in Drawer		Hand Over Item	
	10	100	10	100	10	100	10	100
Bimanual PerActs	8.0	-	36.8	-	5.6	-	0.0	-
Diffusion Policy	4.8	21.6	4.8	5.6	2.4	4.8	0.0	0.0
VoxPoser	8.0	8.0	32.0	32.0	4.0	4.0	0.0	0.0
ACT w/Transformers	4.0	30.4	12.8	28.0	8.8	44.8	1.6	7.2
VoxAct-B (ours)	**40.0**	**59.2**	**73.6**	**72.8**	**39.2**	**49.6**	**19.2**	**14.4**

(2)

	Open Jar	Open Drawer	Put Item in Drawer
ACT w/Transformers	2.7	12.0	14.7
VoxAct-B (ours)	**21.3**	**62.7**	**17.3**

(3)

Method	Open Drawer
VoxAct-B w/o VLMs	19.2
VoxAct-B w/o Segment Anything	67.2
VoxAct-B w/o acting and stabilizing	64.8
VoxAct-B w/o arm ID	68.0
VoxAct-B (ours)	**73.6**

Real-World Results