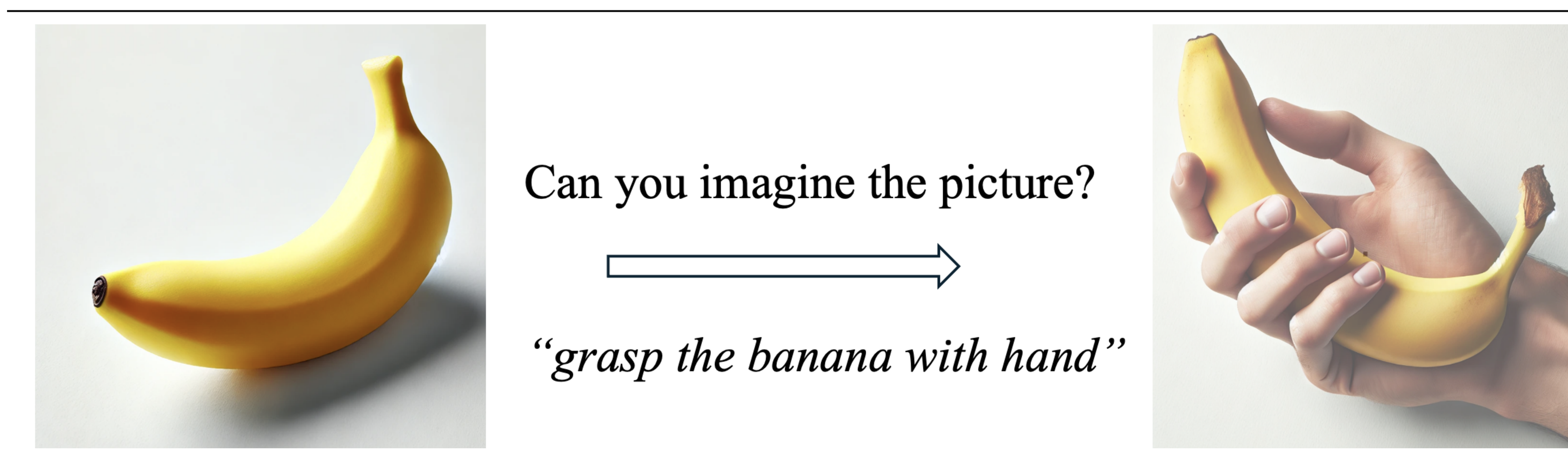
Imagination Policy:

Using Generative Point Cloud Models for Learning Manipulation Policies

Haojie Huang[1] Karl Schmeckpeper[†2] Dian Wang[†1] Ondrej Biza[†12] Yaoyao Qian[‡1] Haotian Liu[‡3] Mingxi Jia[‡4] Robert Platt[12] Robin Walters[1]

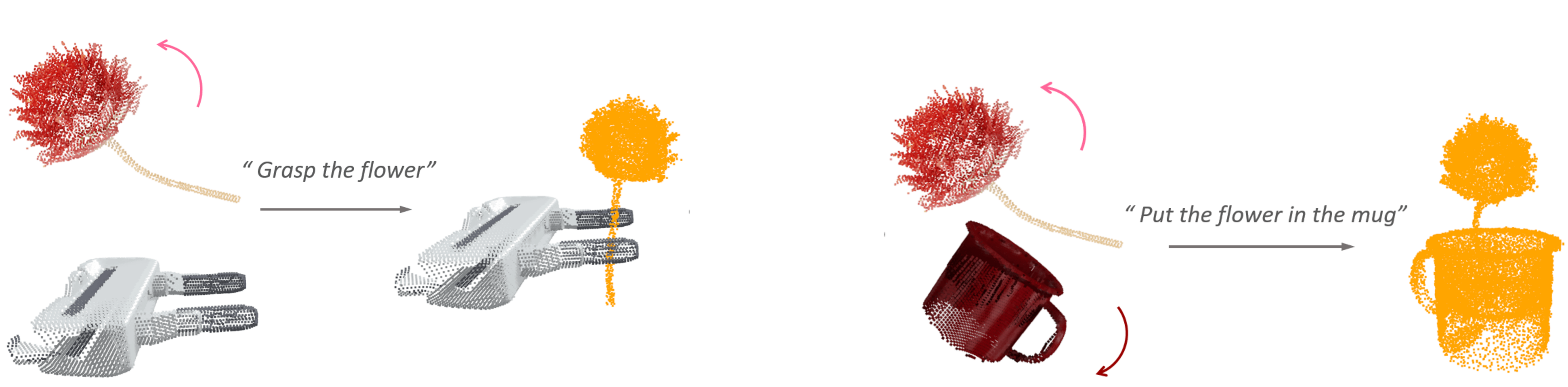
[1] Northeastern University [2] Boston Dynamics AI Institute [3] Worcester Polytechnic Institute [4] Brown University
†‡ Equally Contribution

'Imagine' the goal state



(1). Human can imagine the goal states during planning and perform actions to match those goals.

(2). Imagination Policy generates point clouds to imagine desired key states (pick, preplace, place) which are then translated to actions.



(a) Pick/Single Generation (b) Place/Pair Generation

Figure 1. Illustration of pick generation and place generation. The generated points are colored in orange. A key symmetric property: different rotated observations will not affect the imagined state.

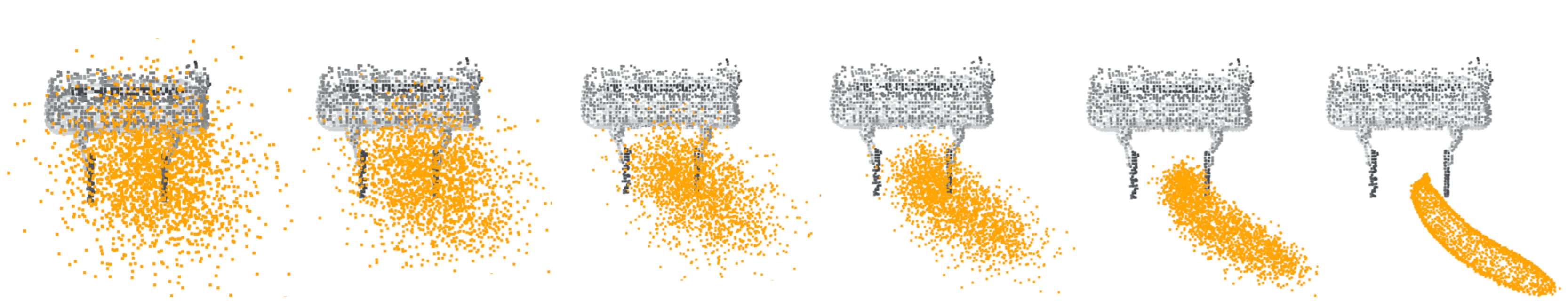


Figure 2. Trajectory of the pick generation process conditioned on the gripper point cloud. ("grasp the banana by the crown").

Multi-modal capability of generation

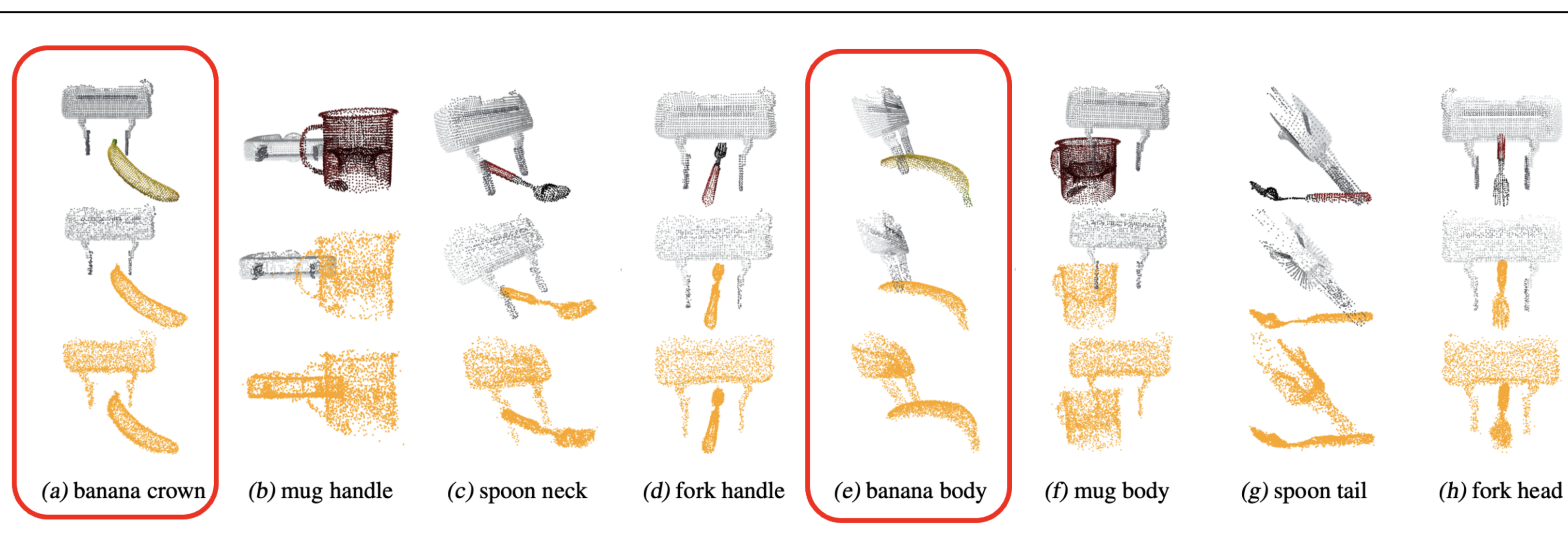
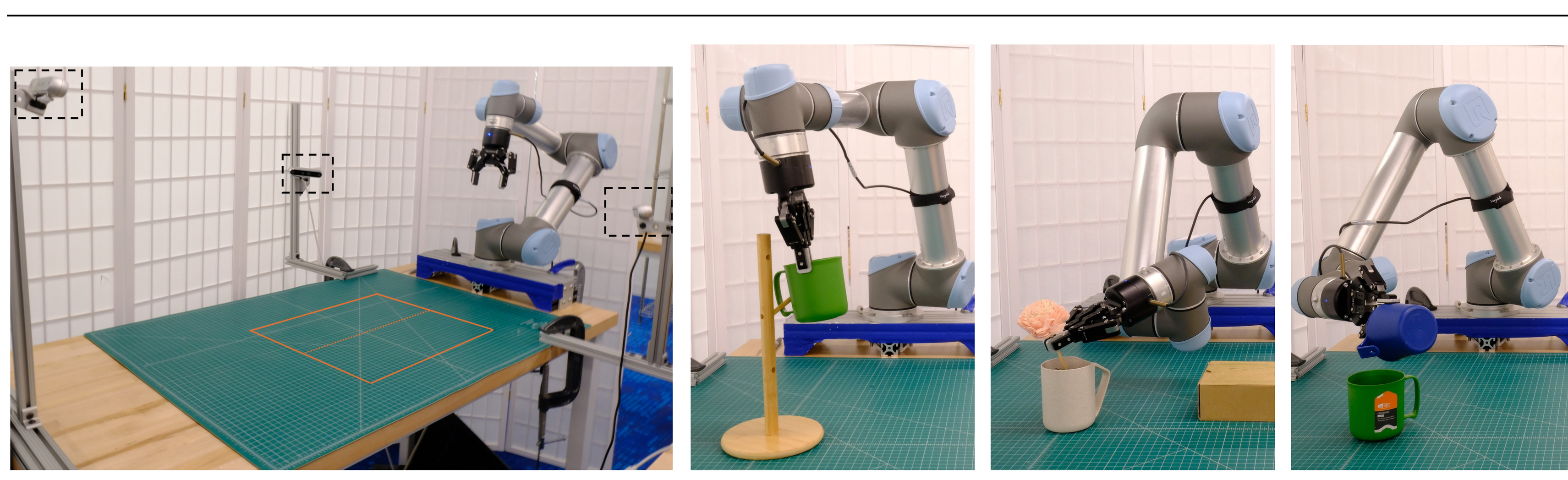


Figure 3. With the same input, the model can generate different configurations.

Real-world Experiments: with only 10 demos



(a) Workspace Settings (b) Mug-Tree (c) Plug-Flower (d) Pour-Ball

Task	# demos	# pick completions	# place completions	# completions	success rate
Mug-Tree	10	15/15 (100%)	12/15 (80.0%)	12 /15	80.0%
Plug-Flower	10	15/15 (100%)	14/15 (93.3%)	14/15	93.3%
Pour-Ball	10	14/15 (93.3%)	14/14 (100%)	14/15	93.3%

Table 1. Performance on real-world experiments.

Overview of Imagination Policy

Architecture Design: : From observation to imagination

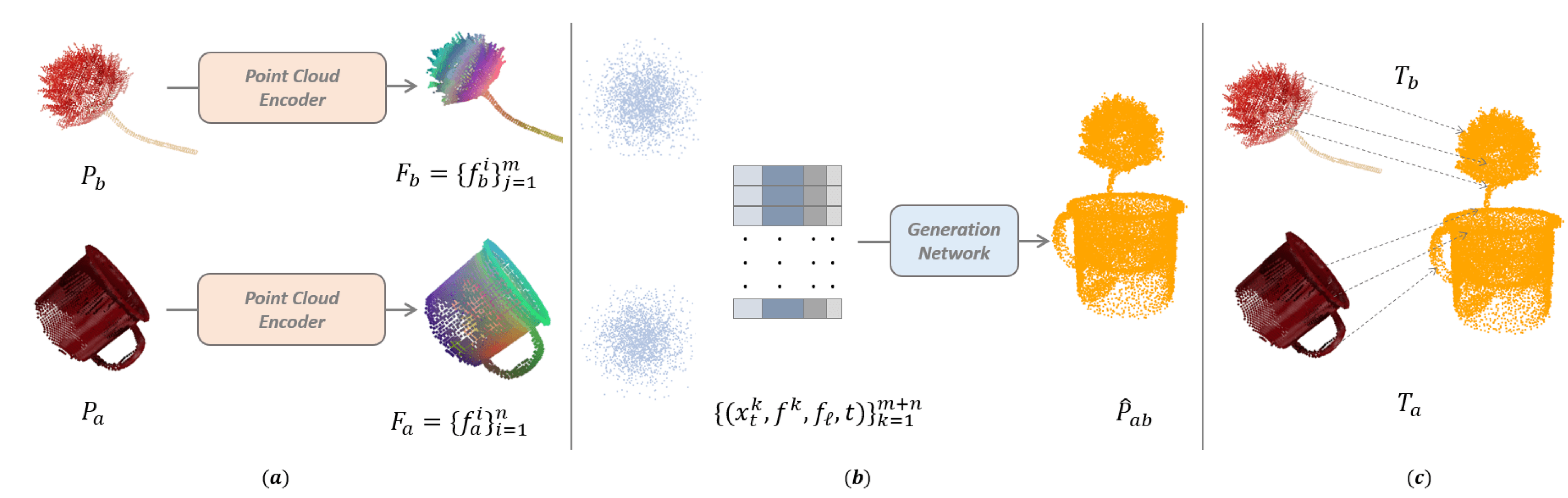


Figure 5. Architecture of Imagination Policy.

We factor action inference into two parts, point cloud generation (Figure 5ab) and transformation inference (Figure 5c).

(a). Encoding the observed point features as F_a and F_b.

(b). Conditional point cloud generation from random Gaussian noise.

(c). Estimating the rigid transformation (T_a and T_b) from the observed point cloud to the generation using correspondence.

Pick/place actions can be calculated with the two rigid transformation matrices. This transforms action inference into a local generative task.

Keyframe action inference: pick, preplace and place

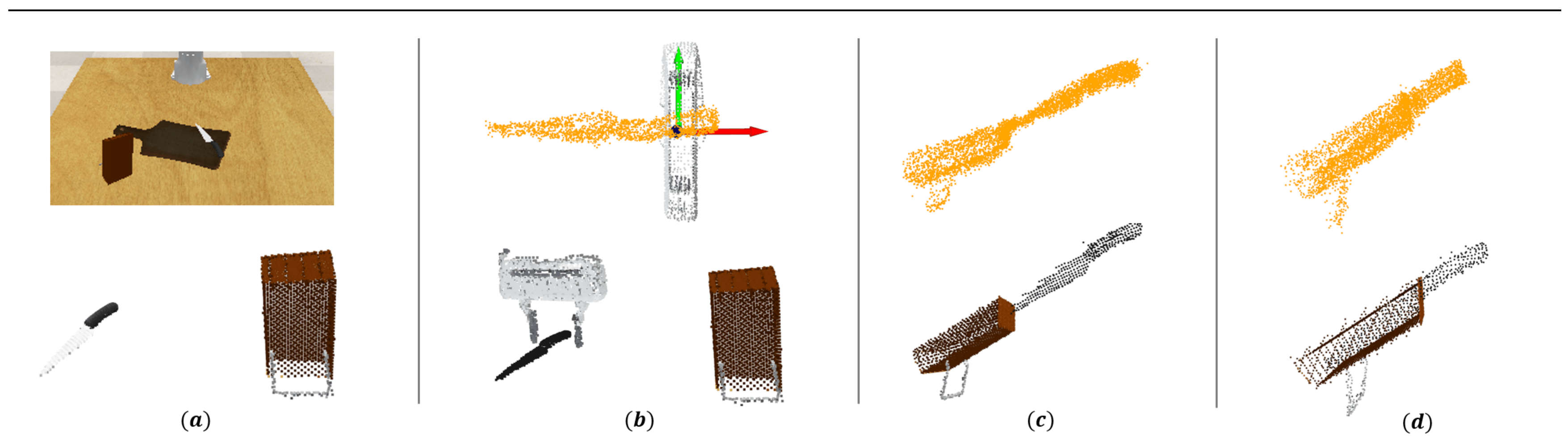


Figure 6. Illustration of the keyframe pipeline on *Insert-Knife*: **(a)** the RGB-D image and the segmented point clouds, **(b)** pick generation, **(c)** preplace generation, and **(d)** place generation.

Simulated Experiments: single model for multitasks

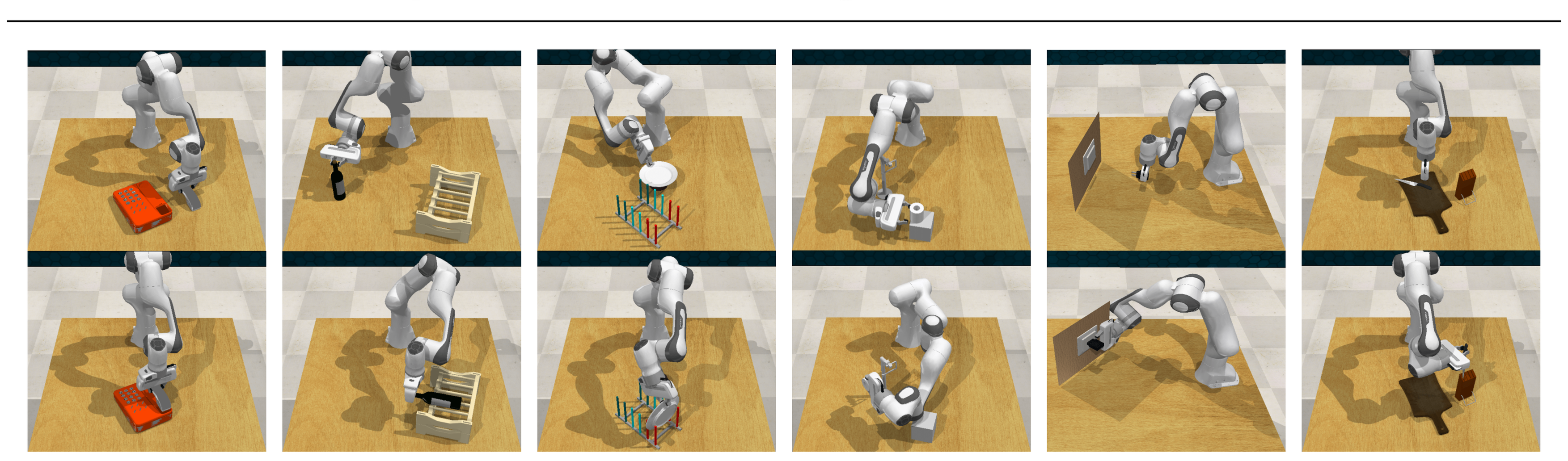


Figure 7. 3D pick-place tasks from RLBench

Model	# demos	phone-on-base	stack-wine	put-plate	put-roll	plug-charger	insert-knife
Imagination Policy	1	4.00	2.67	1.33	2.78	0	0
Imagination Policy	5	78.67	**97.33**	0	1.39	24.00	38.67
Imagination Policy	10	**90.67**	**97.33**	34.67	**23.61**	**26.67**	**42.67**
RVT	10	56.00	18.67	**53.33**	0	0	8.00
PerAct	10	66.67	5.33	12.00	0	0	0
3D Diffusor Actor	10	29.33	26.67	12.00	0	0	0
RPDiff	10	62.67	32.00	5.33	0	0	2.67
Key-Frame Expert		98.67	100	74.6	56	72	90.6

Table 2. Performance comparisons on RL benchmark. Success rate (%) on 25 tests when using 1,5, or 10 demonstration episodes for training. Results are averaged over 3 runs. Even with only 5 demos, our method can outperform existing baselines by a significant margin.