Language-Conditioned Semantic Search-Based Policy (SBP) for Robotic Manipulations Tasks

Jannik Sheikh*, Andrew Melnik*, Gora Chand Nandi**, Robert Haschke*
*Bielefeld University, **IIIT-Allahabad

UNIVERSITÄT BIELEFELD

Indian Institute of Information Technology (IIIT), Allahabad

Results

Method	Input	Success Rate First Setting	Success Rate Second Setting
Baseline	Static RGB & Gripper RGB	38%	30.4%
HULC	Static RGB & Gripper RGB		41.8%
Ours	Static RGB & Gripper RGB	**61.4%**	**57.2%**

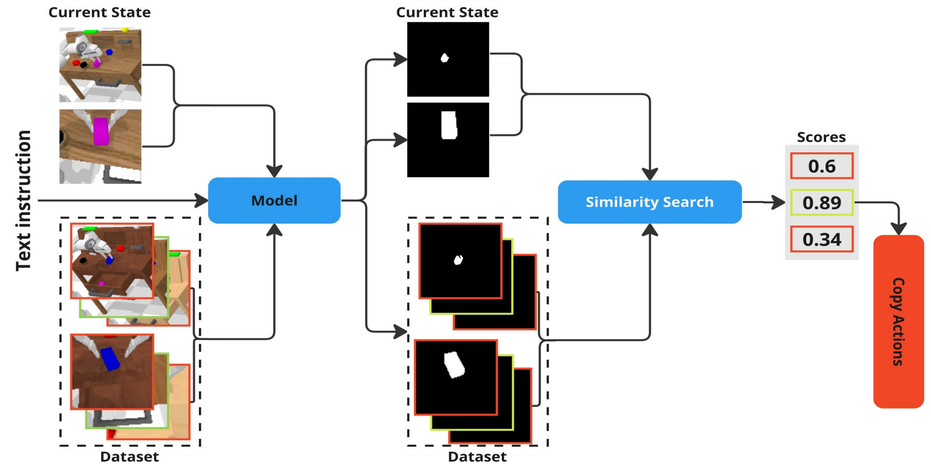
Table 1: Combined results for the Zero-Shot Multi Environment in different evaluation settings.

Task	Success Rate	Task	Success Rate
push pink block left	100%	rotate pink block left	80%
push red block left	100%	rotate red block left	90%
push blue block left	70%	rotate blue block left	30%
push pink block right	90%	rotate pink block right	40%
push red block right	20%	rotate red block right	70%
push blue block right	80%	rotate blue block right	40%
push into drawer	0%	unstack block	70%
lift pink block drawer	90%	stack block	0%
lift red block drawer	70%	turn on led	90%
lift blue block drawer	90%	turn off led	50%
lift pink block slider	50%	turn on lightbulb	70%
lift red block slider	20%	turn off lightbulb	80%
lift blue block slider	10%	place in drawer	100%
lift pink block table	40%	place in slider	30%
lift red block table	30%	move slider right	80%
lift blue block table	50%	move slider left	70%
open drawer	100%	close drawer	90%

Table 2: Our results over all tasks in the first evaluation setting.

Overview

- Instead of training a complex policy, we search in a demonstration dataset for the most similar state and copy the corresponding actions
- Each state is transformed into a latent space capturing the object of interest
- We apply a weighted similarity score between the static and gripper camera, heavily focusing on the gripper camera
- SBP shows promising results and zero-shot capabilities. Foundation Models seemlessly align with our proposed framework



Overview of our framework. We obtain a binary mask of the object of interest in the static and gripper camera views and then find the most similar state in the dataset and start cloning the corresponding actions.

Visualization of the clustered natural language instructions with PCA in a 2D space. For clustering, we fit K-Means to the train embeddings of size 768 generated by the GTE model and set the number of clusters to k, where k represents the total number of tasks, 34. Each data point represents a unique natural language instruction corresponding to a task, and the cluster labels denote the respective tasks.