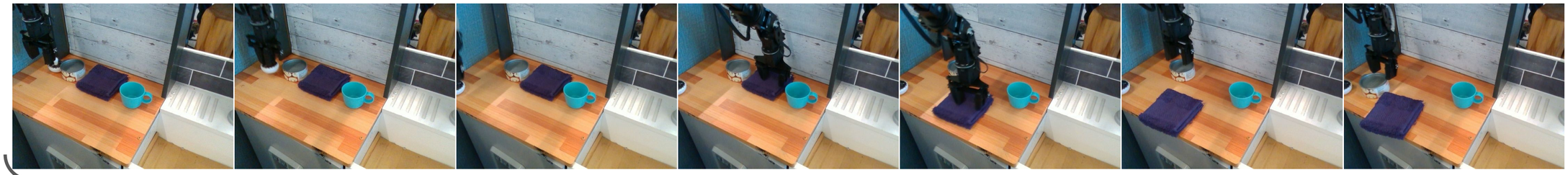
NILS: Natural Language Instruction Labeling for Scalability

Robot Video Demo



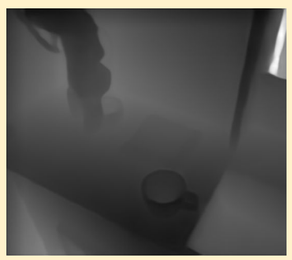
Stage 1: Object Retrieval

VLMs

Detected Objects: Table, Spinning Top, Towel, Cup, Tin Can

Stage 2: Object-Centric Scene Annotation

Bounding Boxes & Segmentation Masks



Predicted Depth

Gripper Positions

Object Movement

spinning top is next to (on the left of) tin can

towel is next to (on the right of) tin can

cup is next to(on the right of) towel

Object Relations

…

Object States

…

Gripper States

Stage 3 : Keystate Detection + Label Generation

State Changes Keystates: -

Gripper Position Keystates: 97,103,109

Object Movement Keystates: 95, 102,109

Object Relations: Keystates: 94, 110

96,102,110

Templated Language of Observations

Object movements: The tinned food moved.tinned food moved 100.0 pixels to the left 54.0 pixels forward
Object relation changes:
Initial relations: tinned food is on the left of towel
Final relations: tinned food is next to spinning top
Gripper proximity: The robot gripper was close to the tinned food
Global position changes: tinned food moved from top left of table to bottom left of table

LLM

Keystate: 102

Generated Language Labels

Move the tin can from the top left of the table to the bottom left of the table

Relocate the tin can to the left of the spinning top

Move the tin can to the left and then forward