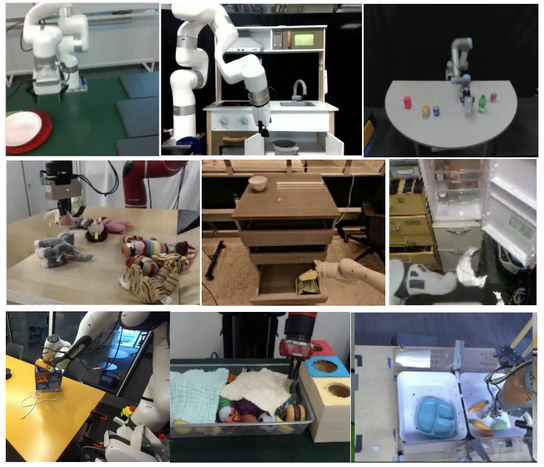
Multi-modality Real Data

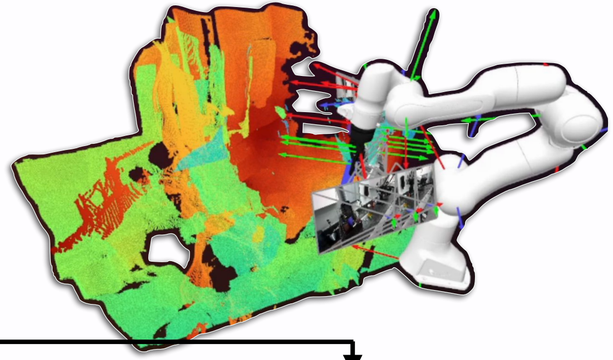


176k Manipulator Episode
from 463 Real Scenes

Robo2VLM

Scene-interaction Understanding

- Semantic Segmentation
- Manipulation Phase Classification



Object info, current phase

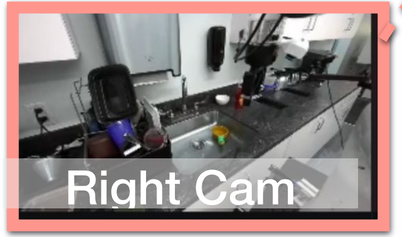
Keyframe selection

Embodied Question Template

Query

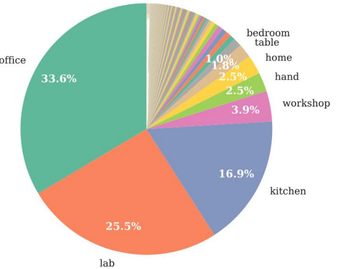
Visual Language Grounding

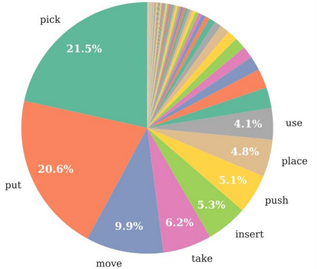
- Embodied Question Conversion
- Spatial Query Projection



Selected camera view

VQA categories





Sample Question

Question

Which language instruction best describes the robot's trajectory shown in the image?

ID: Mon_Oct_23_13_15_02_2023_q1 Episode: Open the tap_190

trajectory_understanding

Choices

A. Drop the book into the platform

B. Move the pen to the drawer

C. Open the tap (Correct)

D. Grab the cup with the gripper