Skill Chaining for Mobile Manipulation Task-Learning

Bahaa Aldeeb, Karthik Desingh
baldeeb@umn.edu

Introduction:

Advances in robotics are bringing robot home assistance closer to reality. To that end, plenty of recent research has focused on teaching robots to perform tasks such as picking and moving objects. Most of these methods focus on table top applications and light objects. Plenty of training examples is also required for training such models.

We tackle the problem of manipulating heavy objects with limited demonstration data. To do so we build a model that utilizing algorithmic skills (grasp, drag, etc.) capable of dealing with object interactions. We focus our efforts on the task of furniture rearrangement.

Input-output:

RGB-D images from multiple cameras are first processed then voxelized. A voxel-map along with a language instruction is given as input to the model. The model then predicts the skill and the location at which to perform it.

Network:

We use CLIP[1] to encode images and pass them as voxels along with language instructions to a PerAct[2] like model, a transformer based visuo-language 3D model.

Dataset information:

Training data is collected using live robot demonstration. We focus on the task of chair re-organization. Using a fiduciary marker the robot recognizes the chair and follows programmed instructions to grasp then drag it. External cameras record the task execution.

Experiments:

Currently we are focus the task "Go to the right of chair"

References:

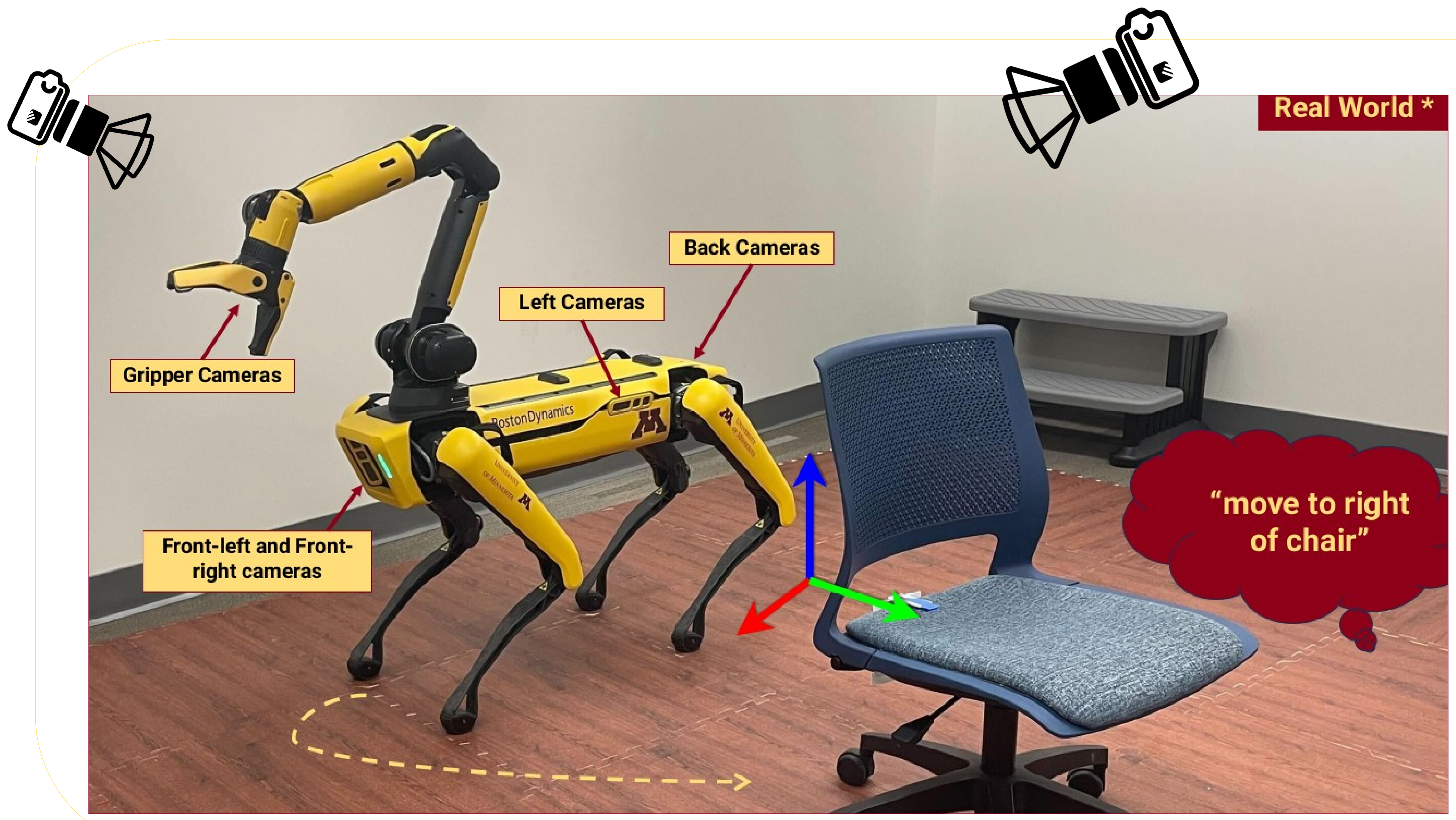
[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

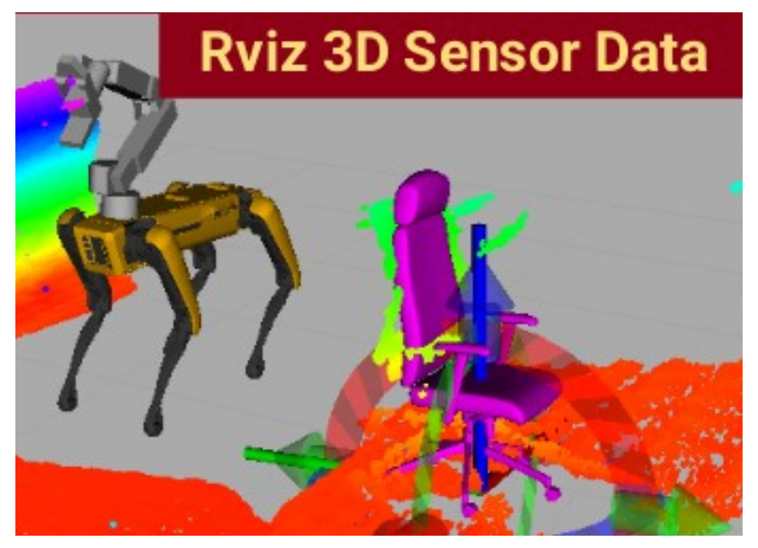
[2] Shridhar, Mohit, Lucas Manuelli, and Dieter Fox. "Perceiver-actor: A multi-task transformer for robotic manipulation." Conference on Robot Learning. PMLR, 2023.

computer science & engineering

Can robots learn to manipulate heavy objects through skill chaining?

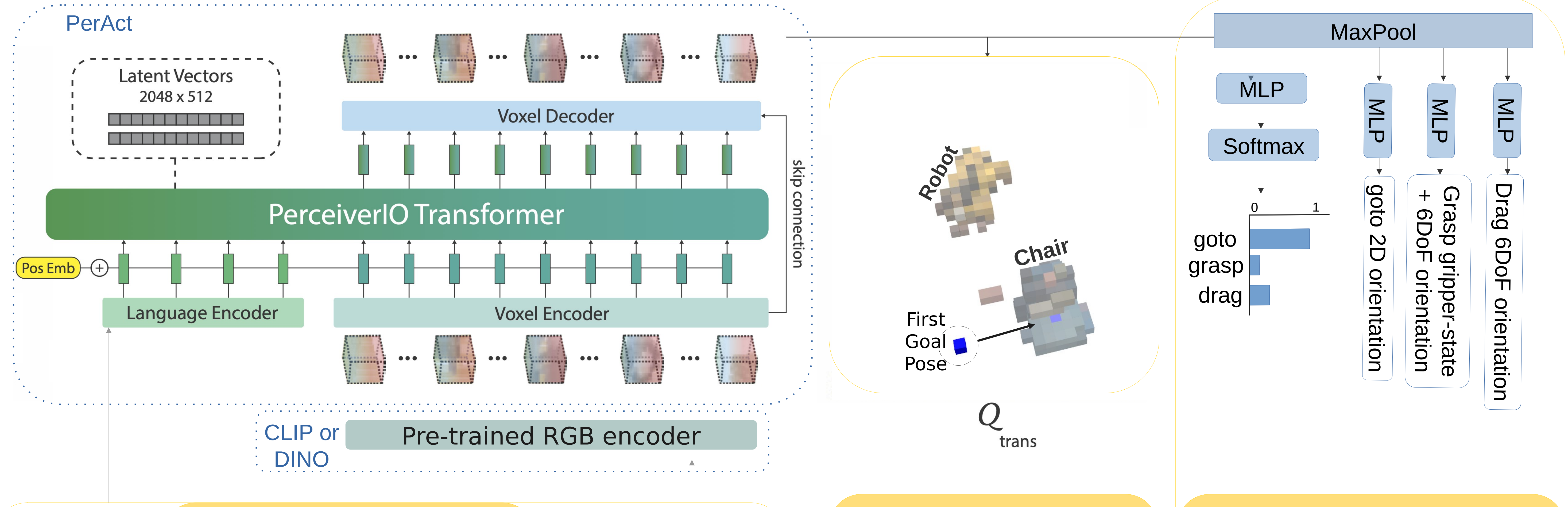
Data Collection





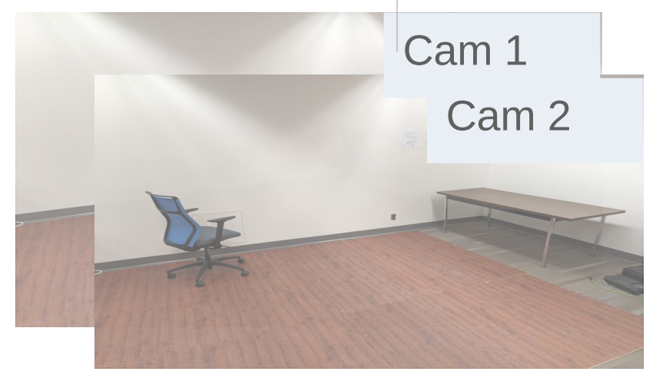
While robot cameras track the object of interest, data is collected using external cameras.
Rviz is used to trigger the robot to perform scripted demonstrations.

Model



"Drag the chair next to table"

The model input consists of Text Prompt as well as rgb-d views from multiple camera.



The model predicts a 3D Q-map indicating where a skill should be predicted.

Separately, the model predicts what skill to apply and the required parameters.

MINNESOTA ROBOTICS INSTITUTE
UNIVERSITY OF MINNESOTA

Robotics: Perception & Manipulation Lab