Group Project Proposal

THE ROBOTICS INSTITUTE

NeuroGrip

16-662 Team 3A

Author: **Yu-Hsin (Thomas) Chan**
Andrew ID: yuhsinch
E-mail: *yuhsinch@andrew.cmu.edu*

Author: **Boxiang (William) Fu**
Andrew ID: boxiangf
E-mail: *boxiangf@andrew.cmu.edu*

Author: **Joshua Pen**
Andrew ID: jpen
E-mail: *jpen@andrew.cmu.edu*

Author: **Jet Situ**
Andrew ID: jets
E-mail: *jets@andrew.cmu.edu*

February 7, 2025

Carnegie Mellon University
Robotics Institute

1 Overview

The overall aim of the project is to use text-based natural language commands to control the Franka arm to pick up and drop off items from the shelf. On a high level, user prompts are given to a pre-trained vision-language model (VLM) along with a continuous video feed of the shelf. The VLM outputs a prompt for the Franka arm to complete a particular grasping task. This could involve picking up an user-specified item from the shelf, and/or placing an item on the specified level of the shelf.

2 Forward Process

The forward process involves first receiving an user command prompt. Along with a video feed of the shelf, the prompt is fed into a vision-language model to output a command prompt for the Franka arm. Using the video feed, the object to be manipulated is located and the surface dimensions of the item is determined using image segmentation. Next, the homogeneous transforms for the end-effector, position on the shelf, camera, and object are calculated so that all items are in the same frame of reference. Inverse kinematics is used to move the end-effector to a Cartesian offset from the object to be manipulated. The object is then grabbed by the end-effector and moved to its desired location using positional way-points.

3 Reset Process

Whenever the task is interrupted or is unable to continue, the reset process is initiated. This involves back-tracking on the position way-points from the forward process. The end-effector returns the object to its original location. If this is not possible, the object will be dropped at a designated drop-zone so the object is not broken during the reset process. Finally, the robot arm returns its pose to its default position and lets the user know that the task was not able to be completed. The robot arm remains in this state as it awaits the user for further input.

4 Motion Generation Problem

For the outer-loop of the motion generation problem, we intend to use Cartesian way-points to gradually move the Franka arm from its starting position to its destination. The way-points should be incremental steps to prevent the robot arm from acting in undesired ways. Since we anticipate multiple planning queries, we will most likely be using probabilistic roadmaps (PRMs) to construct a map in 3D-Cartesian space. This will be mapped to joint-angle space using inverse kinematics.

For the inner-loop, we intend on using positional PID control to move the joints to the desired outer-loop joint-angles. The Franka arm has a built-in library `frankapy` and functions `goto_joints()` and `goto_pose()` that could be helpful to use for implementing controls. Furthermore, some of the team members have experience using the MoveIt ROS package, which could be used to implement the inner-loop control.

5 Variability Explored

The variability explored in the task space is the many combinations of actions, objects, and locations the user prompt can issue to the Franka arm. The action space is $\{Get, Put\}$, the object space can include 5 or more unique objects, and the location space specifies which tier from the shelf to get or put the item. For example, we can have the task {*Get the red book from the 3rd row of the shelf*} or {*Put the teddy bear on the top row of the shelf*}. The total number of possible tasks is $2 \times O \times L$, where O is the number of objects and L is the number of rows on the shelf.

6 What Could be Learned

From this project, we could learn how to integrate large language models to robotic applications. In particular, using natural languages for manipulator control and grasping. Such tasks would be widely applicable to real-world scenarios such as warehousing and distribution centers. The precise getting and putting of designated objects to specified locations would greatly aid in sorting efficiency and accuracy.

7 Hardware Setup

A shelf would be set up at a fixed location from the Franka arm. A camera would be fixed directly facing the shelf and have the whole shelf in frame. A computer would be connected to the Franka arm and be running an operations terminal that allows natural language inputs. The natural language inputs is sent to a vision-language model to parse. A command for the Franka arm is generated as output. This is then fed to the Franka arm for motion generation and control. A sketch of the setup is displayed in Figure 1.

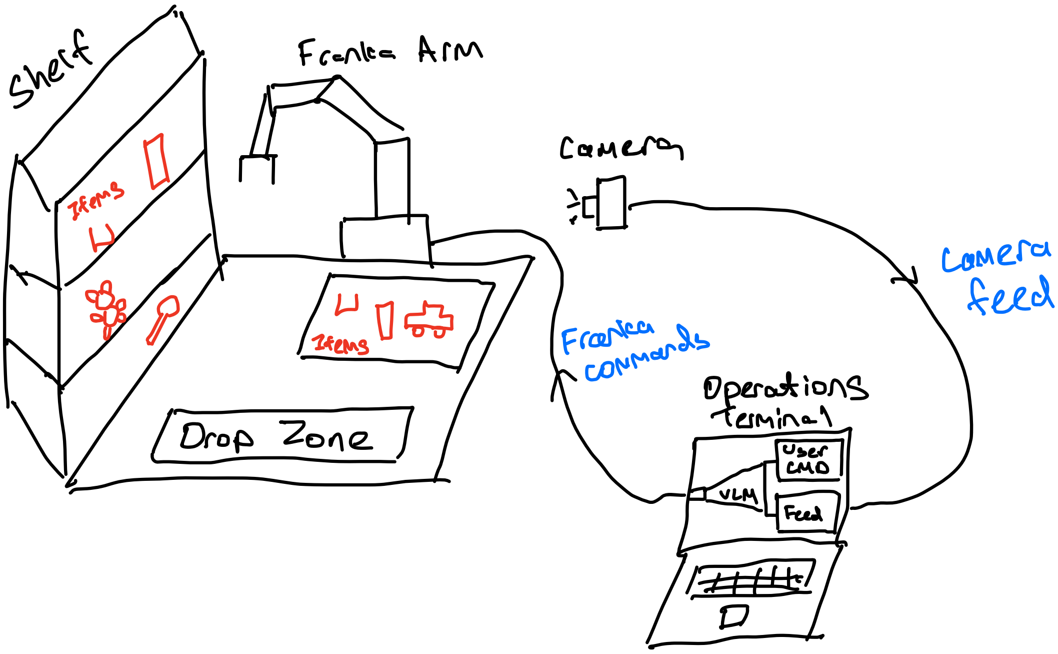

Figure 1: Preliminary Hardware Setup