Background

Grasping is a crucial skill for advanced robotic applications and has huge implications for fields like assistive robotics, manufacturing, or logistics. Multi-fingered hands, compared to 2-jaw grippers, offer advantages in terms of their compatibility with a wider range of object categories and also the manipulability after grasping.

Furthermore, leveraging the development of Large Language Model (LLM) and Visual Language Model (VLM), we have the potential to achieve interactive robotic tasks given language inputs. Together with partners from 33 academic labs, Google has released **RT-1-X** [1]**,** a robotics transformer (RT) model derived from **RT-1** [2] and also the whole Open X-Embodiment dataset [1]. The model shows skills transfer across many robot embodiments and is able to generalize to novel tasks. **Voxposer** [3] addresses this challenge in a modularized manner. They use LLM and prompt engineering to generate code which contains function calls for VLM and grasping/manipulation modules. **VL-grasp** [4] has a similar pipeline where they published a new visual-grounding dataset dedicated to their application scenery. **VIMA** [5] uses multimodal prompts to train an attention agent to control the robot arm in various tasks.

Motivations

So far, none of the existing work combines LLMs and VLMs with both multi-fingered hands or even humanoid robots for the purpose of interactive grasping and manipulation. To address this challenge, we need to devise a strategy for solving grasping and manipulation tasks using multi-fingered hands in an interactive manner, capitalizing on the advancements in LLMs and VLMs.

We have built the whole experimental setup in fig 1., including the hand, the Realsense 415 camera, and the robot arm in the gazebo simulator. A dataset of 180k grasp samples has been generated from the simulation. We will use DLR-HIT Hand II [1] as a five-fingered hand to grasp the unknown objects. It is a 15 degrees of freedom (DOF) fully actuated hand.

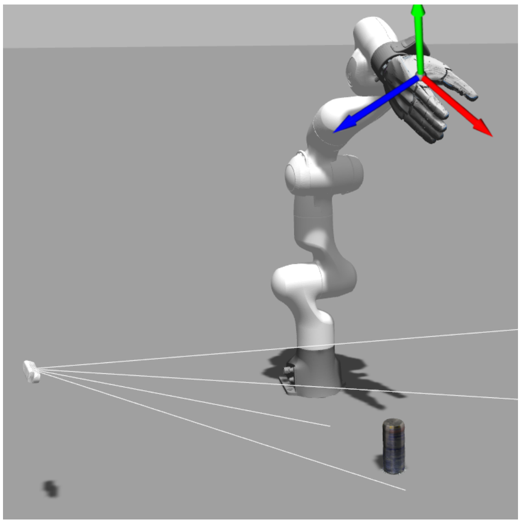

fig 1. Experimental setup in gazebo.

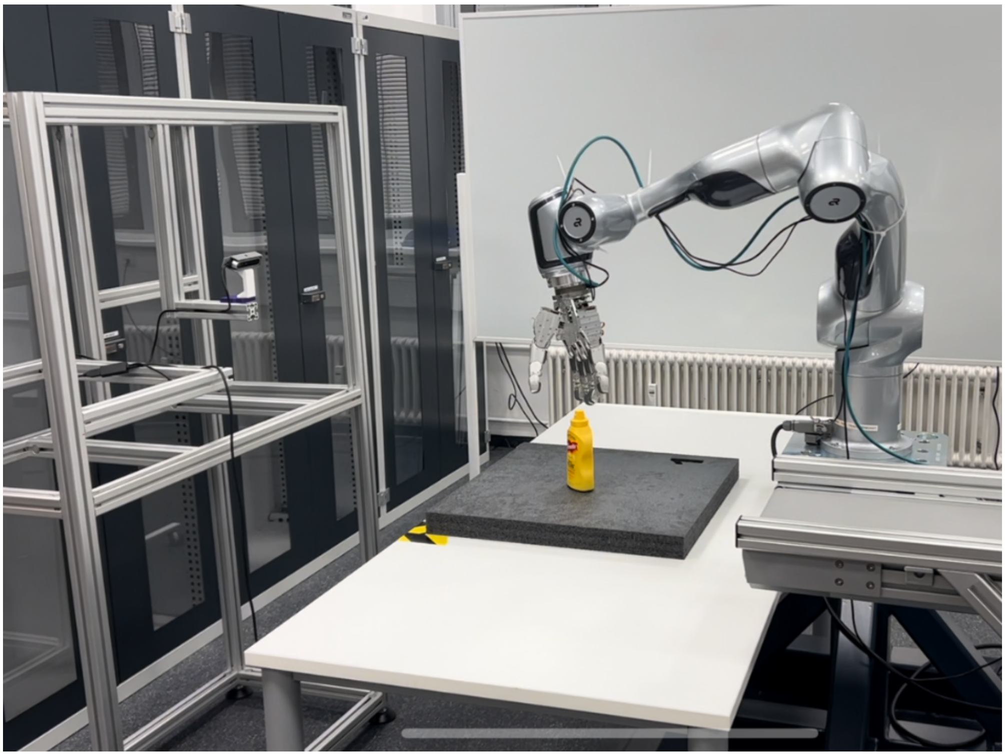

fig 2. Real-world experiment

Topic 1:

- Literature research for LLM-for-robot based methods, specially RT-1,2 from google as end-to-end based approaches.
- Replicate RT-1-X on simulation or preferrably on the real-world set up. (code)
- Finetune RT-1-X to adapt to multi-fingered hands with different tasks.
- Conduct real-world grasping experiments

Topic 2:

- Literature research for LLM-for-robot based methods, especially Voxposer or other similar approaches.
- Replicate one of these methods on simulation or preferrably on the real-world set up.
- We have available LLM, VLM and grasp planner ready to use.
- Conduct real-world grasping experiments

Requirements

- Solid programming in Python
- Familiar with Pytorch, tensorflow.
- Knowledge of robotics, computer vision and basic knowledge of LLMs, VLMs.
- It will be great if you have previous experience with robot arms such as Franka, UR and so on.
- Able to work independently.

Application and Contract

This thesis is an external master thesis in cooperation with Agile Robots AG, located in the south of Munich. If you are interested in the offered thesis topics and would like to gain practical experience with the state-of-the-art LLM robotics methods, feel free to contact me.

Qian Feng

Website: https://www.ce.cit.tum.de/air/people/qian-feng-msc/
E-mail: qian.feng@tum.de
Phone: +49 17643306435

References

[1] Open X-Embodiment : Robotic Learning Datasets and RT-X Models

[2] RT-1: Robotics Transformer for real-world control at scale

[3] VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models

[4] VL-Grasp: a 6-Dof Interactive Grasp Policy for Language-Oriented Objects in Cluttered Indoor Scenes

[5] VIMA: General Robot Manipulation with Multimodal Prompts arXiv:2210.03094v2

[6] Code as Policies: Language Model Programs for Embodied Control

More to go: https://github.com/GT-RIPL/Awesome-LLM-Robotics