# Pickachu: Vision-Language Model for Generalized 3D Robotic Manipulation

Pranay Junare (junar002@umn.edu)
Aditya Bidwai (bidwa001@umn.edu)
Samra Huseynova (husey012@umn.edu)

## Motivation:

Recent major advancements in large-scale foundational models for vision, language, and action have proven to provide generalization by reasoning about perception and action in a unified manner. Thus, we investigate the usage vision-language models for task generalization in robotic manipulation.

## Problem Definition:

**Input**: Natural language instruction $L$, set of RGB-D images $\mathbb{R}^{H\times W\times 4}$ from 4 static cameras
**Output**: 6-DOF end effector pose in $SE(3)$

## Proposed Idea:

- **Simulation Environment:** CopelliaSim with RLBench tasks
- **Visual Observation:** Four RGB-D cameras
- **Loss function:** Cross-entropy loss
- **Transformer:** Multi-view Transformer with 8 attention layers, SigLIP2[4] for language emb.
- **Training:** Using Imitation-learning
- **Dataset:** Pre-generated dataset provided by seminal work PerAct[2]

## Limitations and Discussions:

- Model evaluated on 25 task variations in simulation using CoppeliaSim.
- Demonstrates strong generalization to unseen instructions, achieving a success rate of **84%** (21/25 tasks completed).
- Consistently succeeds on short, unambiguous language goals (e.g. “leave the coffee out”)
- Struggles with more complex or spatial tasks (e.g: “pick up and place”, “move to bottom shelf”), which accounts for all 4 failures.

## Future Plan:

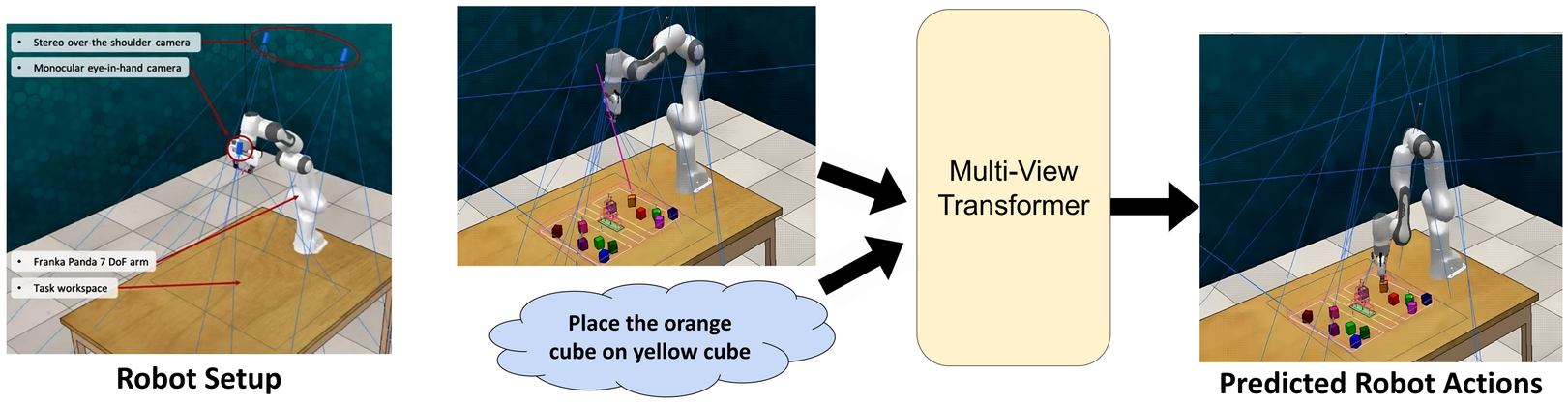
- Scale to different task categories, such as open a jar, turn the tap, open, etc.
- Test our trained model on more complex failure cases in simulation.

## References:

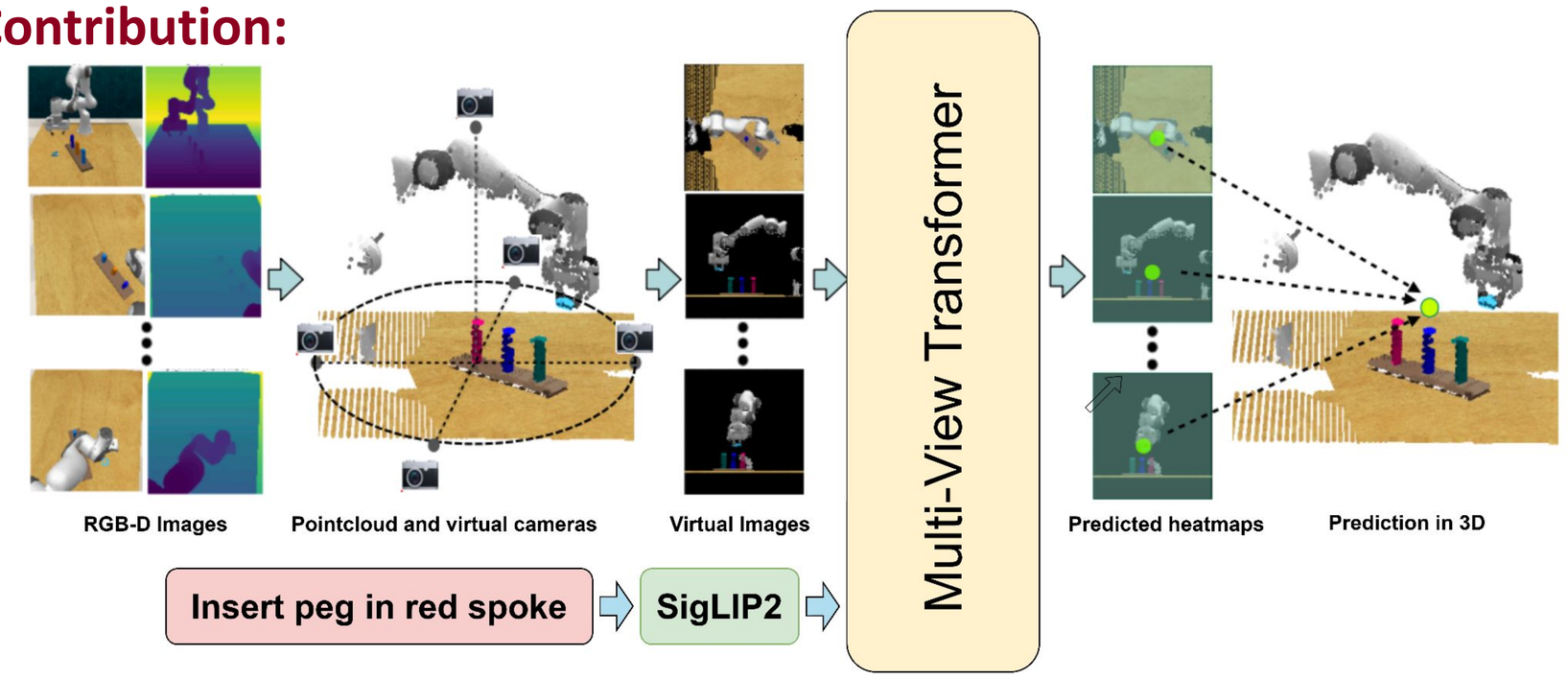
1. "Rvt-2: Learning precise manipulation from few demonstrations.", Goyal, Ankit, Valts Blukis, Jie Xu, Yijie Guo, Yu-Wei Chao, and Dieter Fox, arXiv preprint arXiv:2406.08545 (2024).
2. "Perceiver-actor: A multi-task transformer for robotic manipulation.", Shridhar, Mohit, Lucas Manuelli, and Dieter Fox, Conference on Robot Learning. PMLR, 2023.
3. "Rvt: Robotic view transformer for 3d object manipulation.", Goyal, Ankit, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox, In Conference on Robot Learning, pp. 694-710. PMLR, 2023.
4. "Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features.", Tschannen, Michael, et al., arXiv preprint arXiv:2502.14786 (2025).

# *Can Multi-Modal Learning help perform complex 3D manipulation tasks?*

## Objective:




## Contribution:




## Experimental Results and Comparisons:

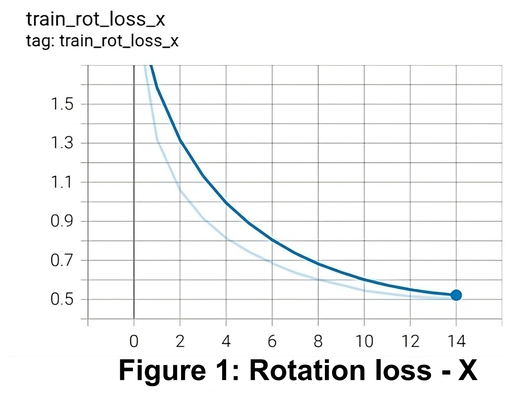



Figure 1: Rotation loss - X

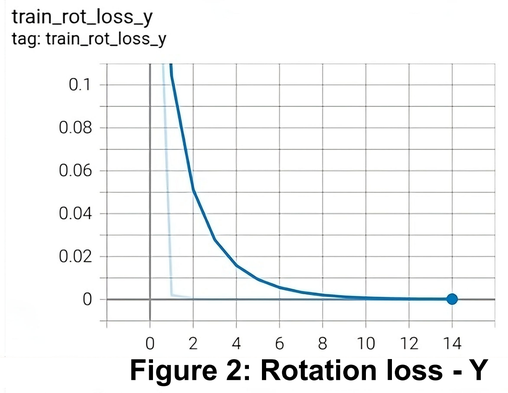



Figure 2: Rotation loss - Y


Figure 3: Rotation loss - Z

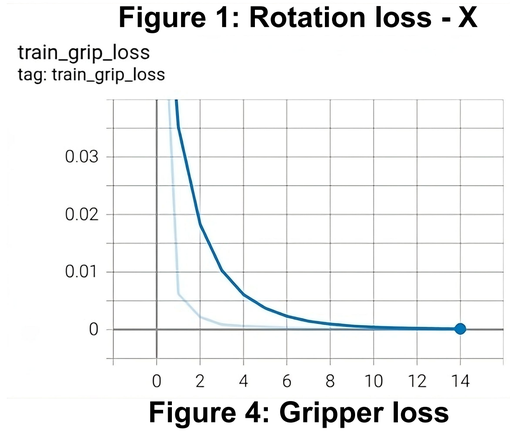



Figure 4: Gripper loss

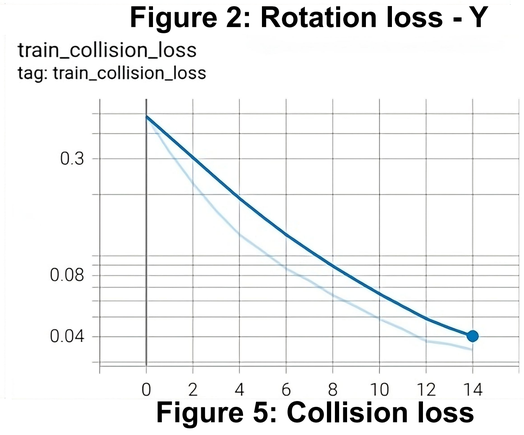



Figure 5: Collision loss

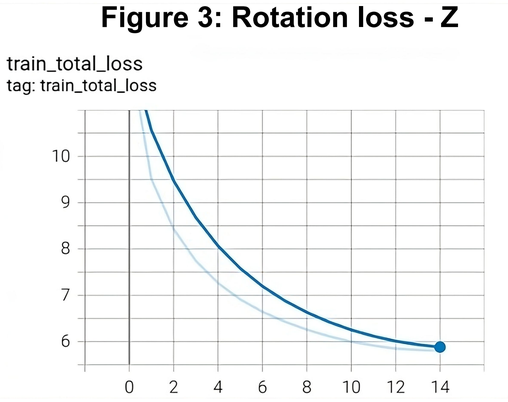
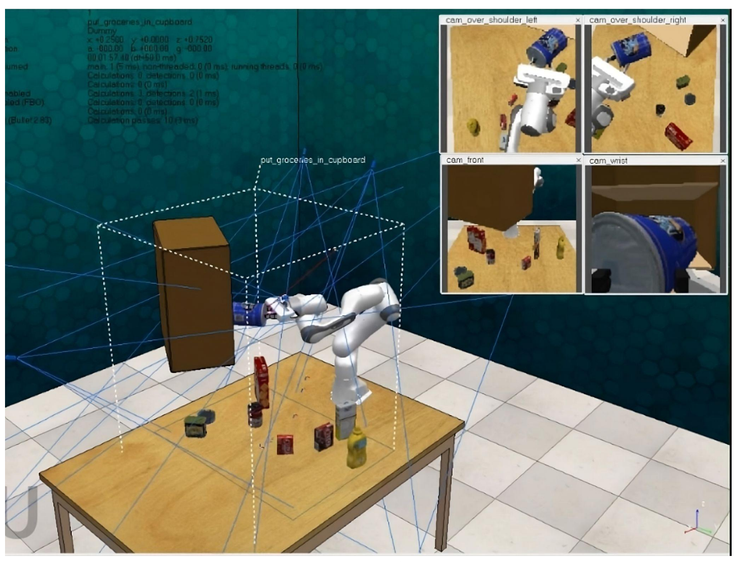
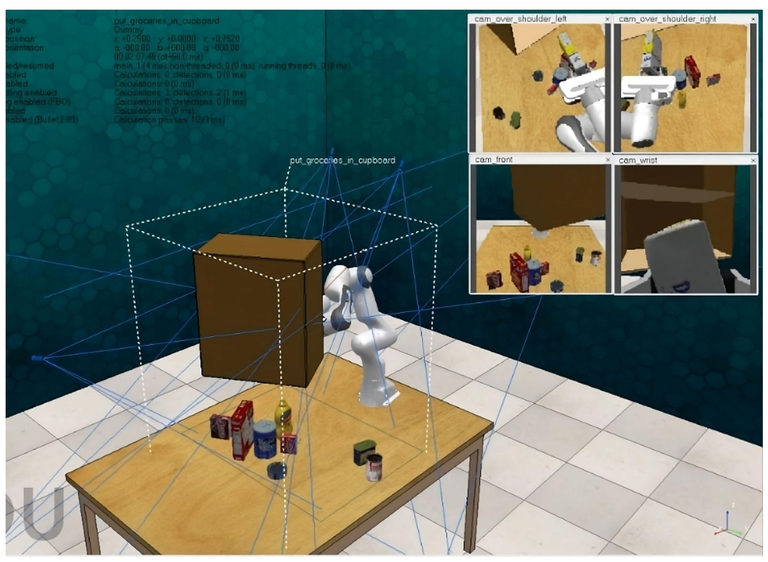



Figure 6: Total loss




**Task:** Place coffee in cupboard




**Task:** Place sugar in cupboard

computer science & engineering

Webpage

MINNESOTA ROBOTICS INSTITUTE
UNIVERSITY OF MINNESOTA