VLM-Robobench

A VLM powered benchmark platform for robot manipulation

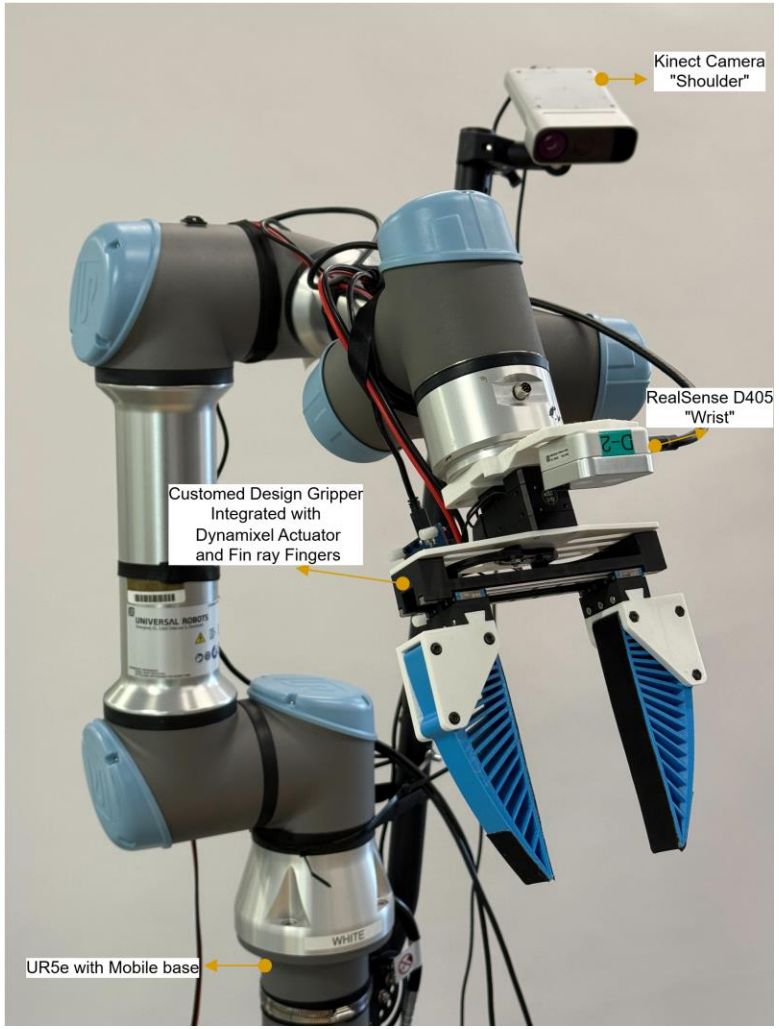
Pengyuan Guo, Zhonghao Mai

10/21/2025

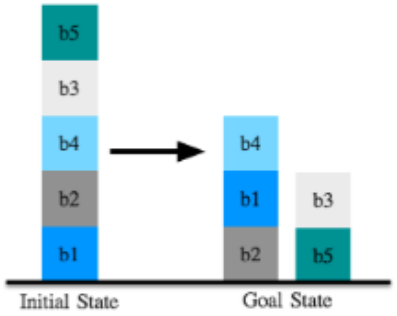
PURDUE UNIVERSITY | College of Engineering

Motivation & Background

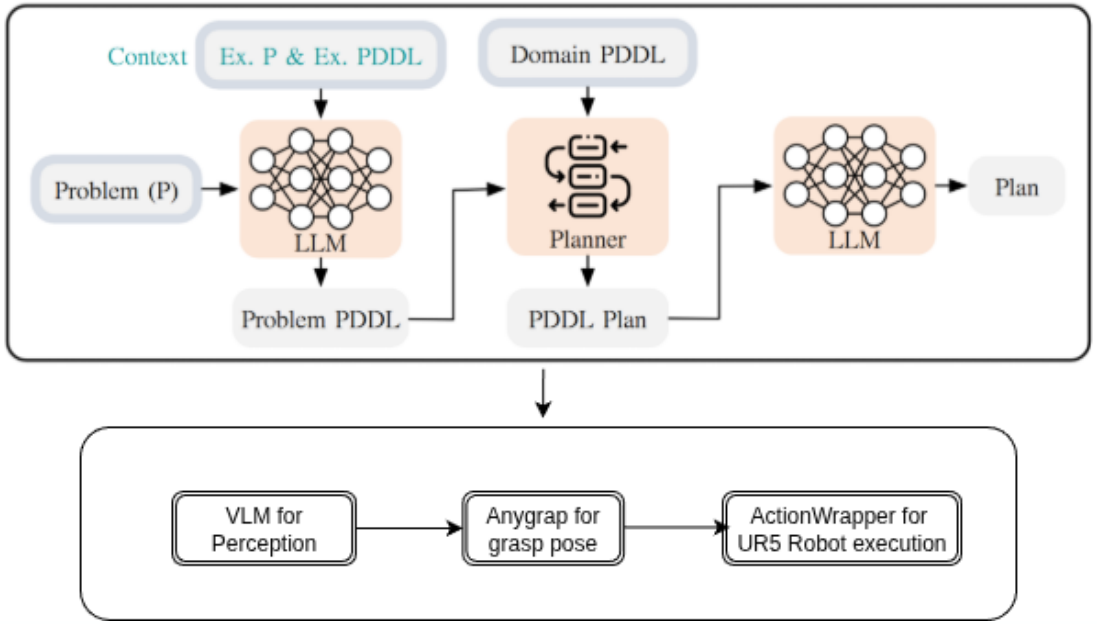
- **Reproducible Real-World Platform**
  - Most robot manipulation benchmarks (e.g., Behavior-1K) focus on simulated environments.
  - We propose an **open-source, real-world testbed** for **VLMs and embodied agents**, built from **accessible hardware and reproducible software**.
- **Leveraging the Power of Vision-Language Models (VLMs)**
  - VLMs are increasingly capable in **spatial reasoning** and **visual grounding**.
  - Our goal: evaluate how well these models can understand, reason, and act in physical manipulation tasks.
  - We aim to design **benchmark tasks and insightful metrics** to measure VLM performance in robotics.
- **Towards Embodied, Agentic Robotics**
  - Inspired by **agentic architectures** (e.g., ReAct: Reason → Act → Observe), which enable long-horizon reasoning.
  - We build an embodied agent combining **LLM + PDDL planner + VLM + GraspNet + Execution module**, enabling **zero-shot manipulation** through reasoning, tool use, and reflection—**without policy training**.



Our Current Pipeline & Structure

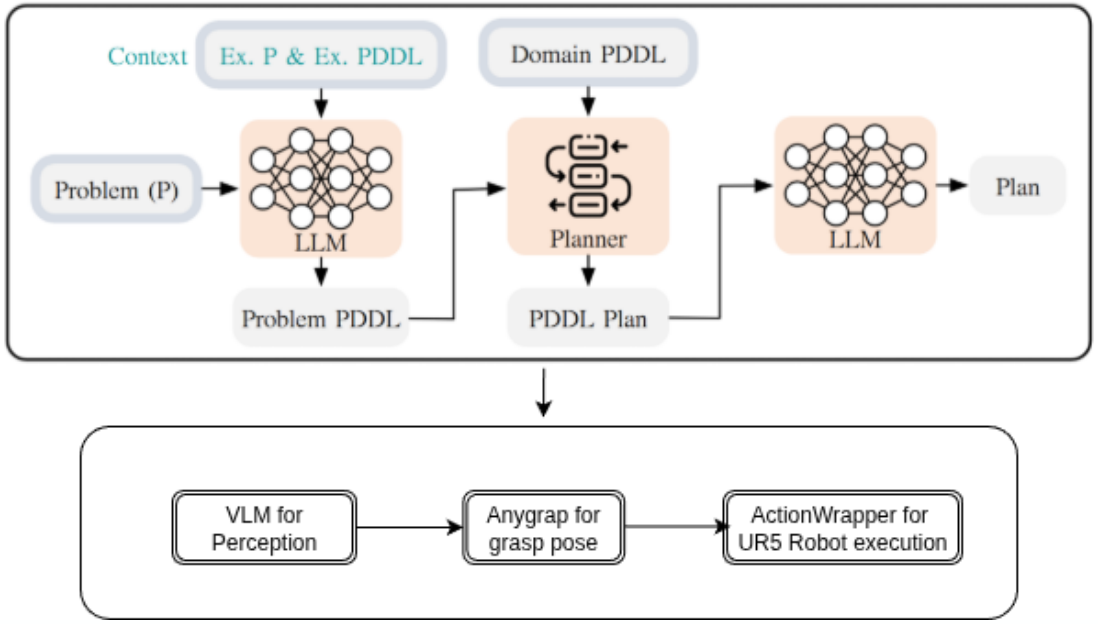


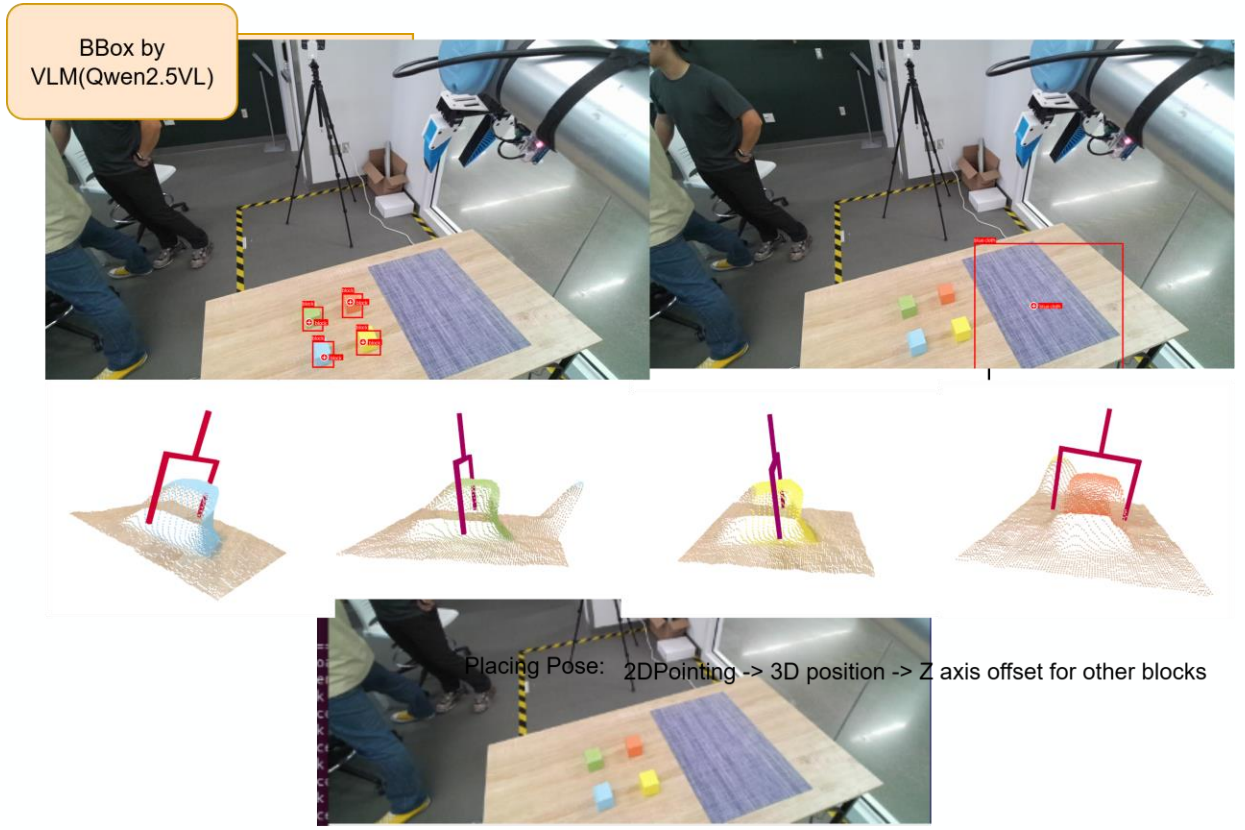
"You are a robot with a fixed base and one gripper that can pick up and stack objects on a table.
Your goal is to create a stack on the table with five blocks.
The final sequence from bottom to top should be: "



To develop a reproducible, VLM-powered benchmark platform that integrates LLMs and VLMs for reasoning and task planning—making robotic manipulation more accessible, flexible, and scalable for research.

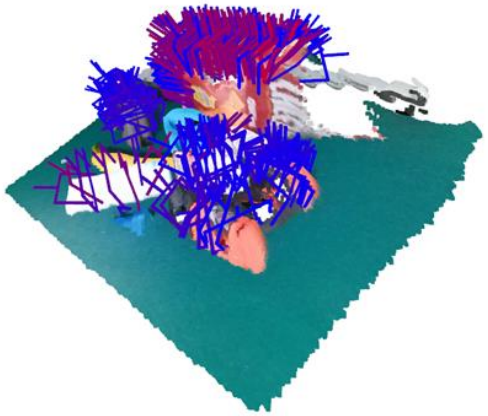
Our Current Pipeline & Structure

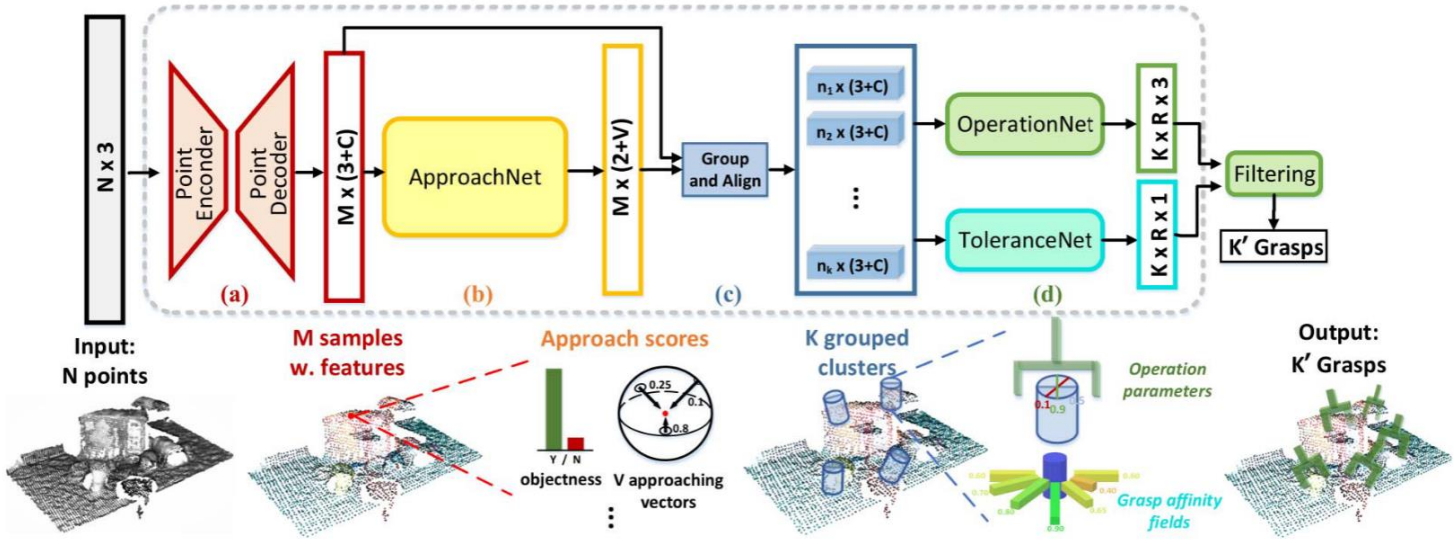




GraspNet for General Object Grasping

General Pipeline for GraspNet (open-sourced, old version of Anygrasp published at CVPR 2020)





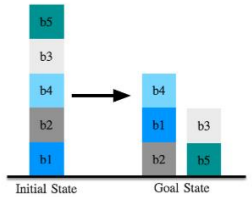
PURDUE UNIVERSITY | College of Engineering

Leveraging LLM+PDDL for Planning

To achieve more complex task, directly using LLM may not be suitable

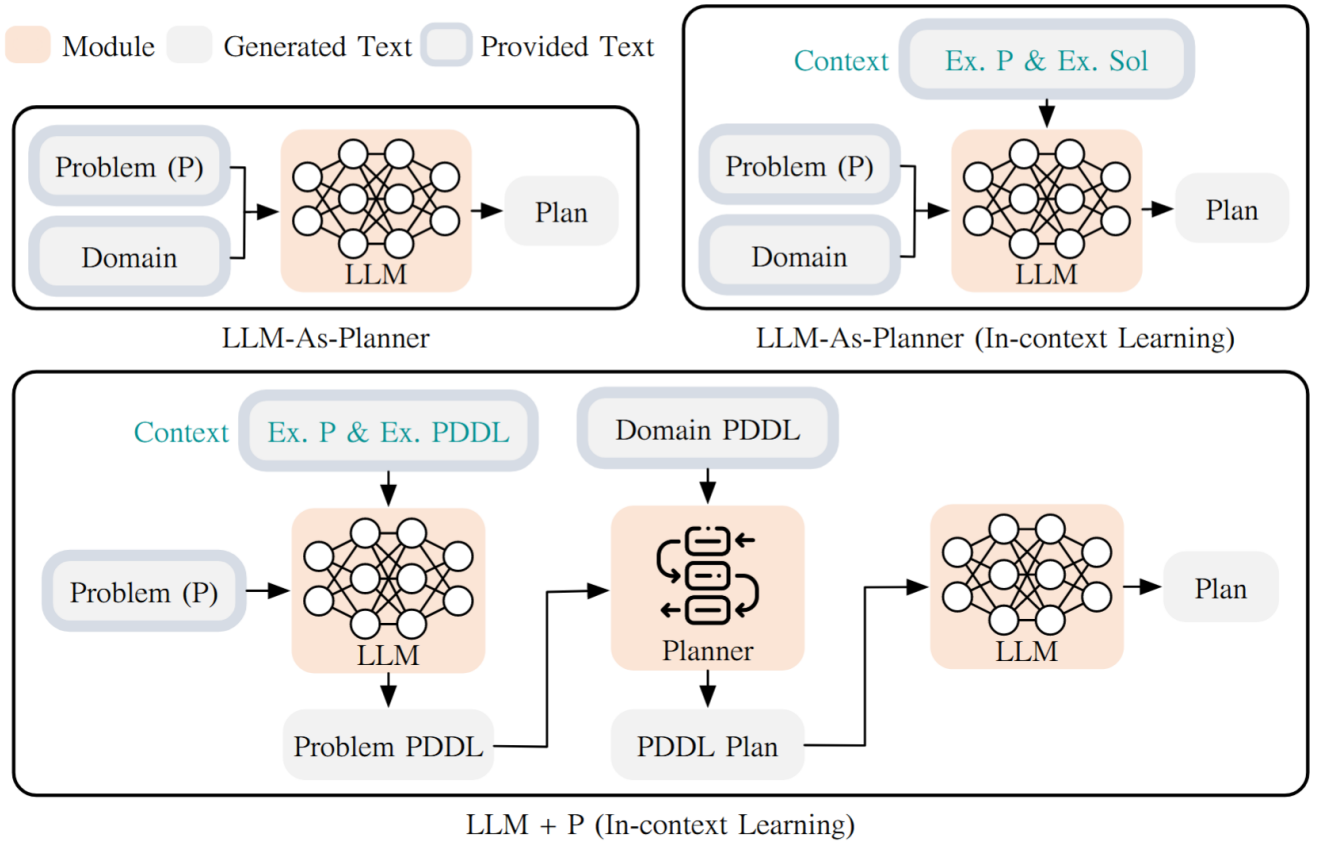
A Failure Example of GPT-4 in Planning

Problem (P1): You have 5 blocks. One cannot place more than one block on another block. b5 is on top of b3. b4 is on top of b2. b2 is on top of b1. b3 is on top of b4. b1 is on the table. b5 is clear. Your arm is empty.
Your goal is to move the blocks.
b1 should be on top of b2.
b3 should be on top of b5.
b4 should be on top of b1.



GPT-4:

```
1. Move b5 from b3 to the table.
2. Move b3 from b4 to b5.
3. Move b4 from b2 to the table.
4. Move b2 from b1 to b5.
5. Move b1 from the table to b2.
6. Move b4 from the table to b1.
```



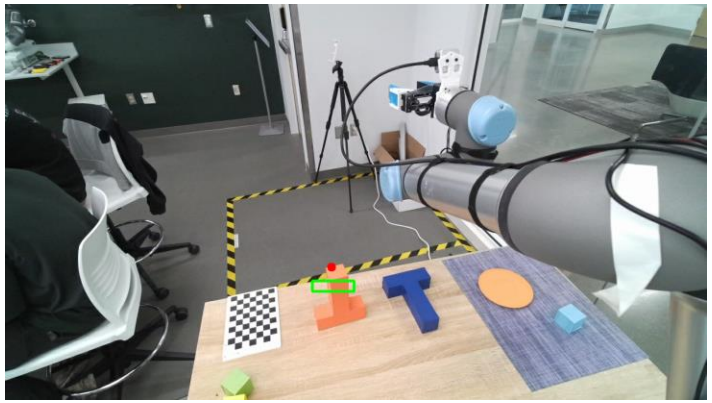
LLM+P, Bo Liu, et al. 2022

Leveraging VLM for Pointing

Qwen2.5VL is very strong in visual grounding, comparing to Chatgpt 4o, 5, Molmo.etc

Qwen2.5VL-3B detection

Molmo-8B: Screwdriver



Molmo-8B: Tip of orange block

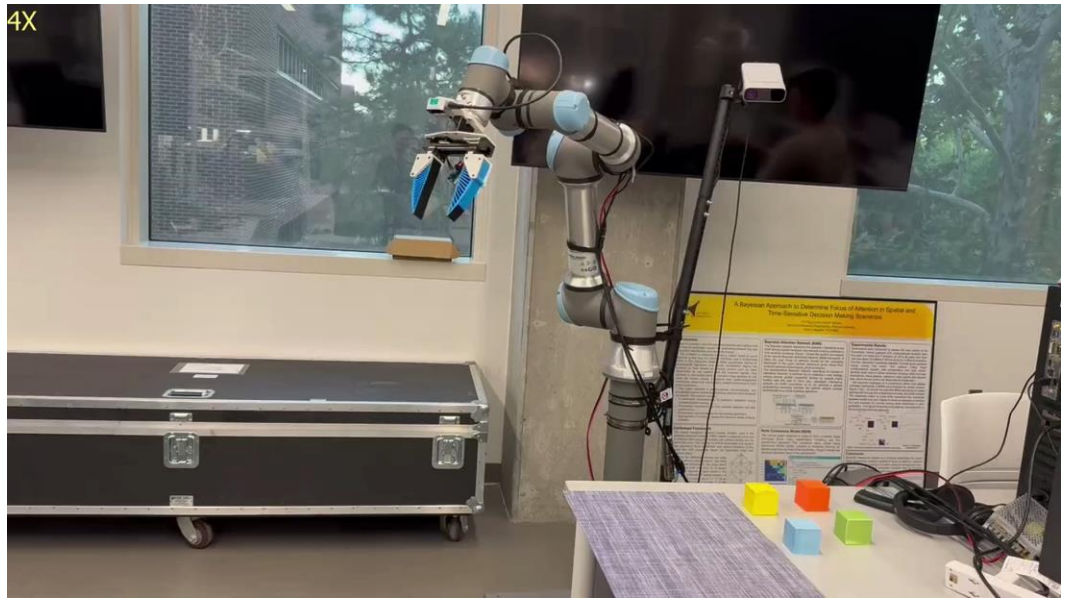
Gpt5 detection

Qwen2.5VL-7B detection

PURDUE UNIVERSITY | College of Engineering

Current Pipeline Showcase

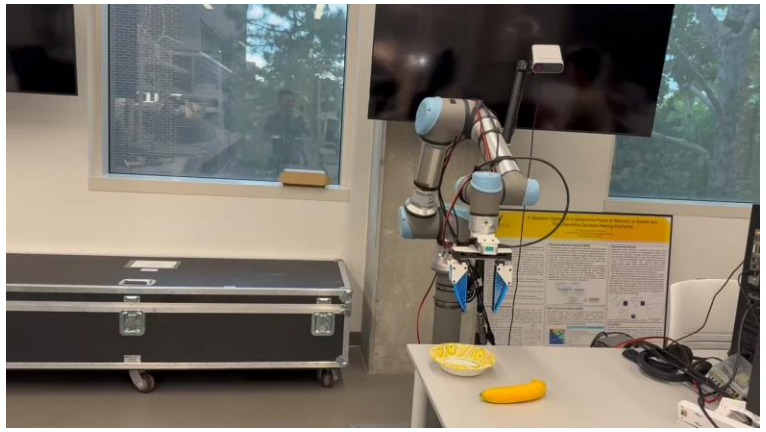
Zero-shot picknplace sequence on different tables:

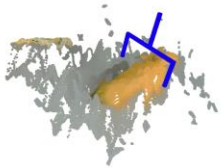


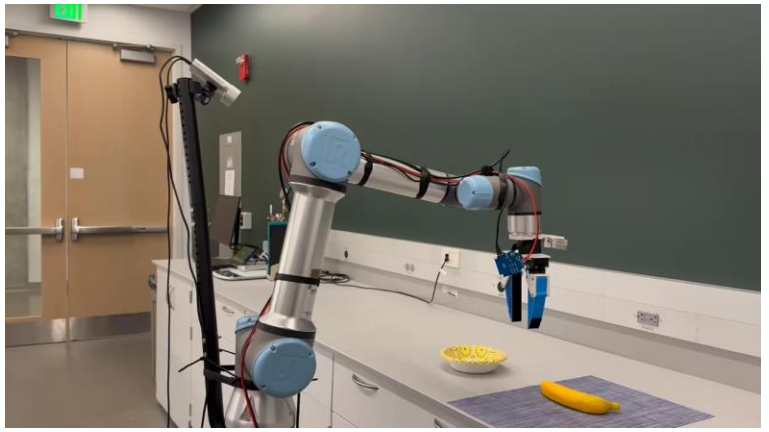
PURDUE UNIVERSITY | College of Engineering

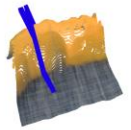
Current Pipeline Showcase

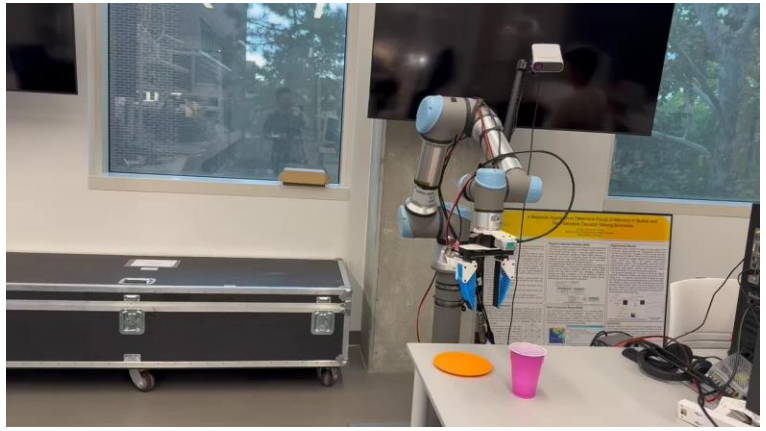
Zero-shot picknplace sequence on different tables:

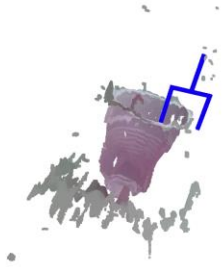


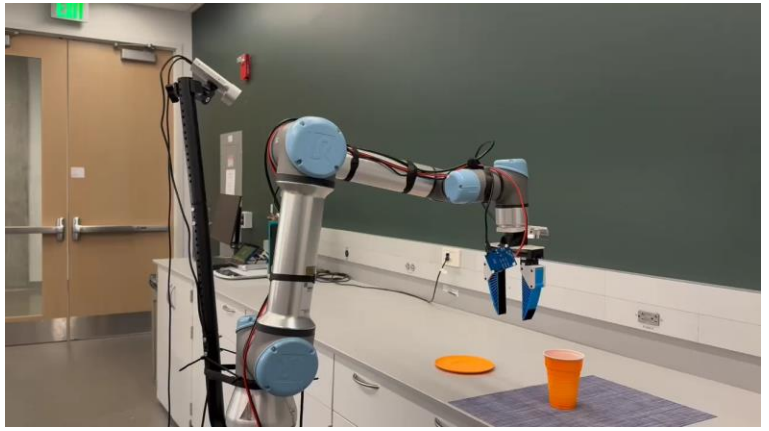


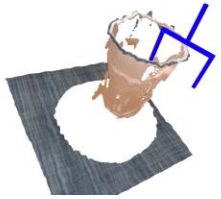






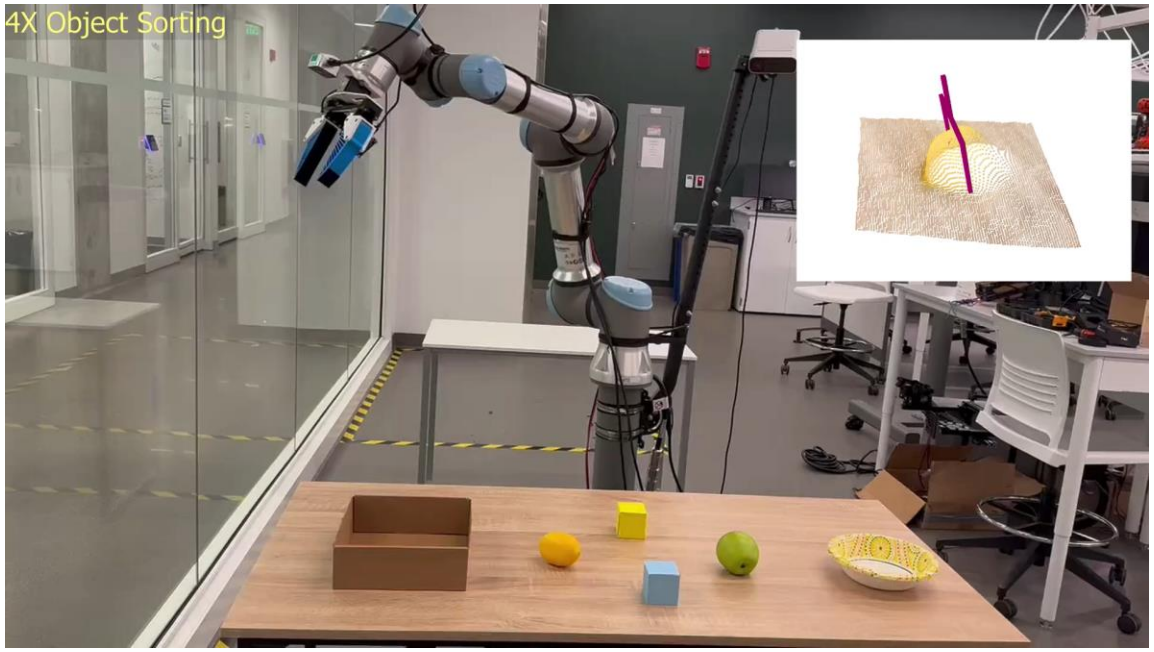


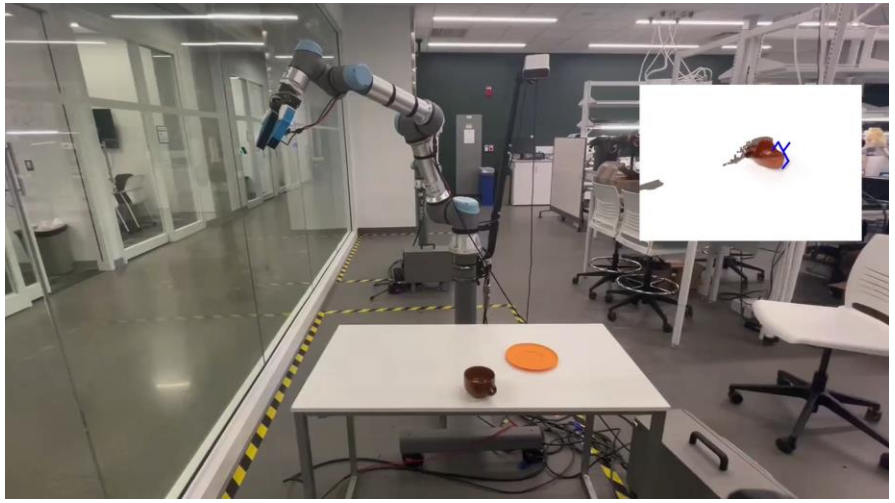




Current Pipeline Showcase

Zero-shot picknplace sequence on different tables:





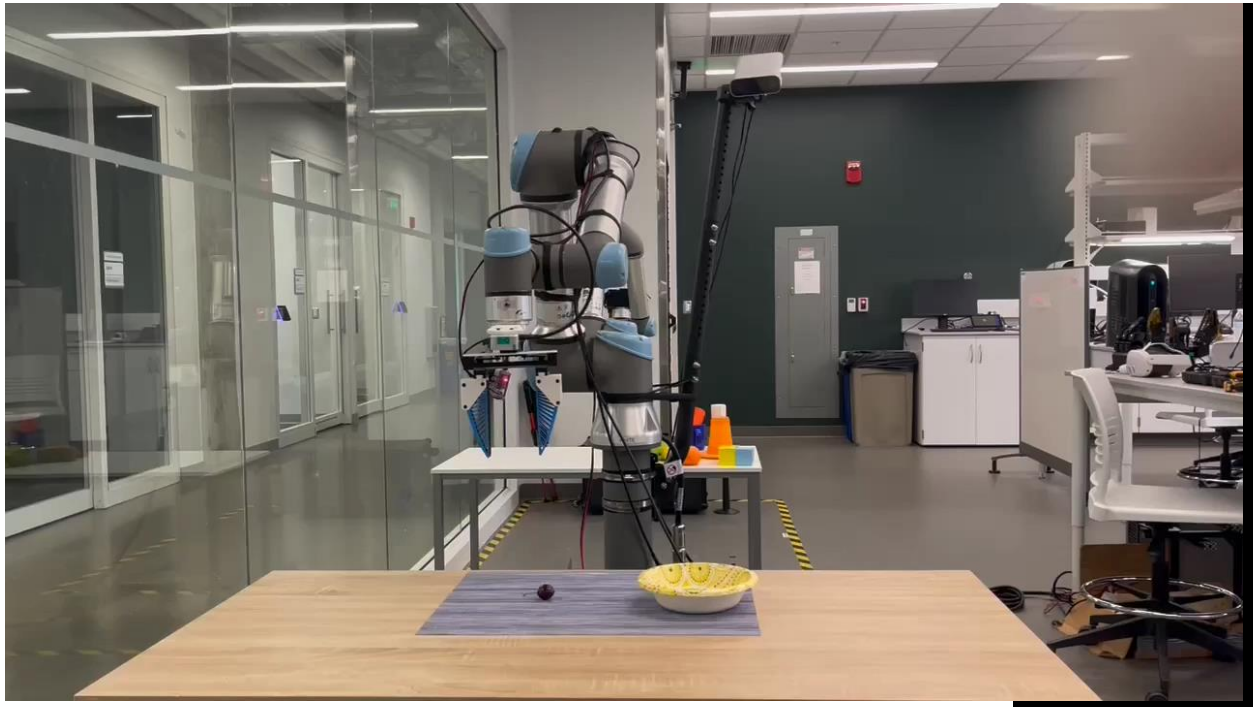
Future Plan

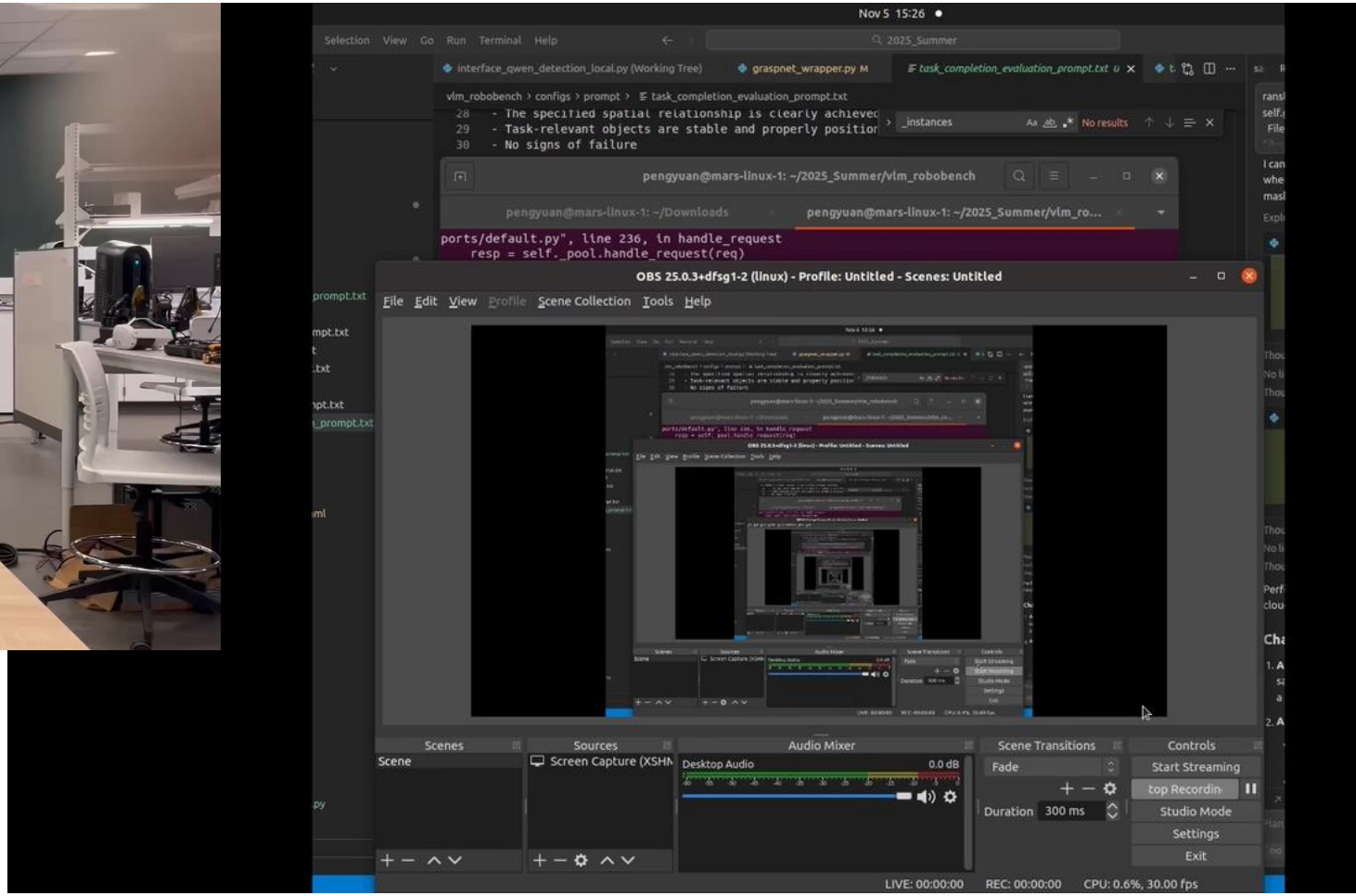
Agent/CoT/PDDL

- **Generalization**
  - We have tools like planning, perception, execution, how can we leverage them to have better generalization, robustness, so the idea of agent come up.
  - Investigate some papers and how can we leverage the ideas (reasoning, agent architecture)
    - AI suite (Andrew Ng, tool use, reflection)
    - MCP (probably, leverage this interface)
    - ReACT
- **Think of how to improve detection capability by getting a closer view of the image.**
  - e.g. enlargement of the image to better analyze the semantic meaning.
  - Refeeding the generated grasp pose to LLM/VLM to verify grasp pose quality

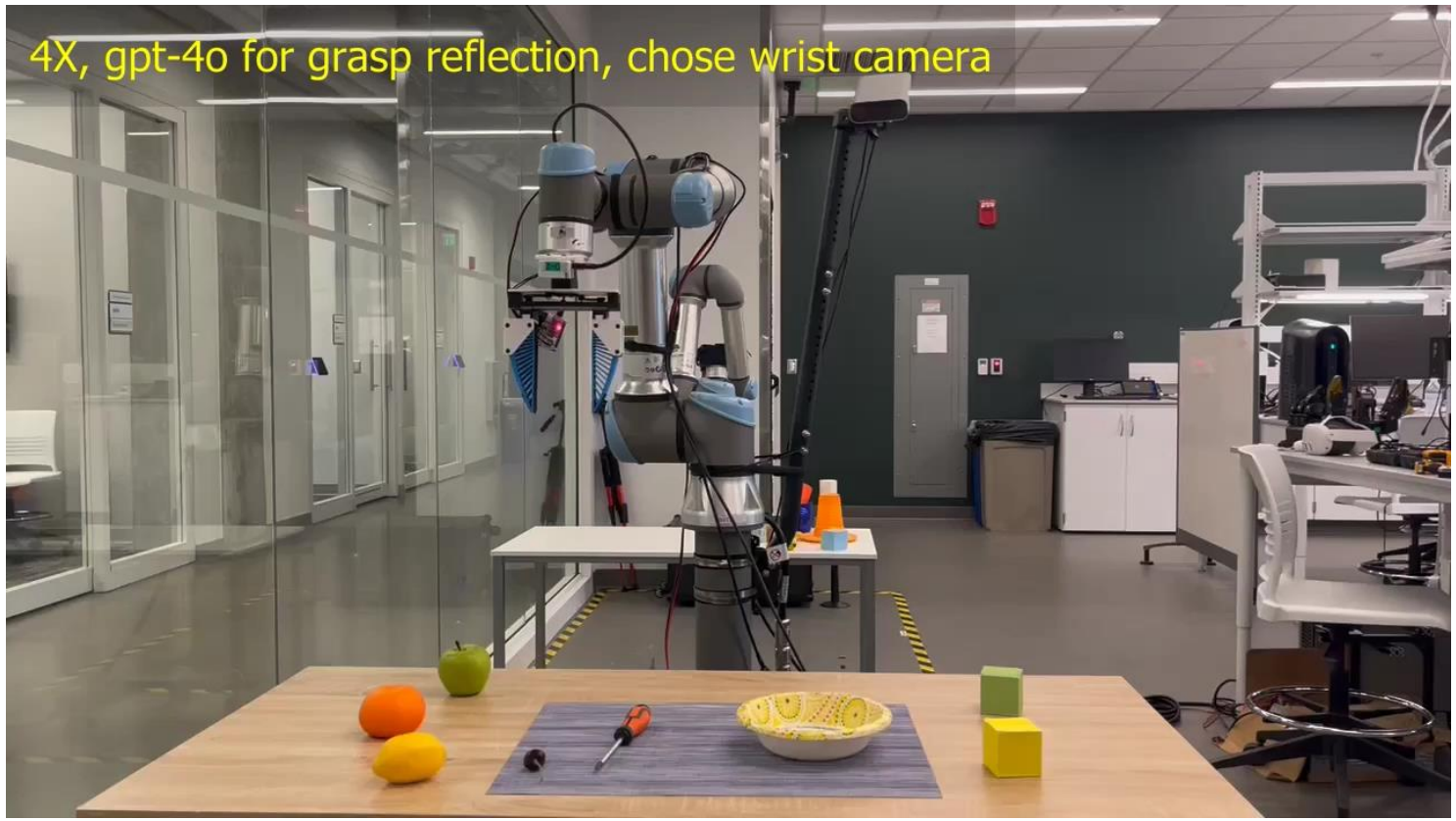
- **Leverage stronger VLM model or other small modules to optimize the architecture.**
  - e.g. Molmo (Excels at pointing, not very accurate in bbox).
  - Adding move actions
- **What kind of novelty or contribution can we present.**
  - Reproducibility
  - Benchmarking on different modules in the pipeline to show insight **(Benchmark tasks/metrics design)**
  - In-the-wild zero-shot task planning and execution
- **Closed-loop/Harder task**
  - Drawer organization
  - Monitoring precondition/effect
- **Question**
  - If we want to add contributions in PDDL+agent/reasoning, which direction is more promising? Any Suggestions?

Action in the loop with reflection on grasp pose & task completion



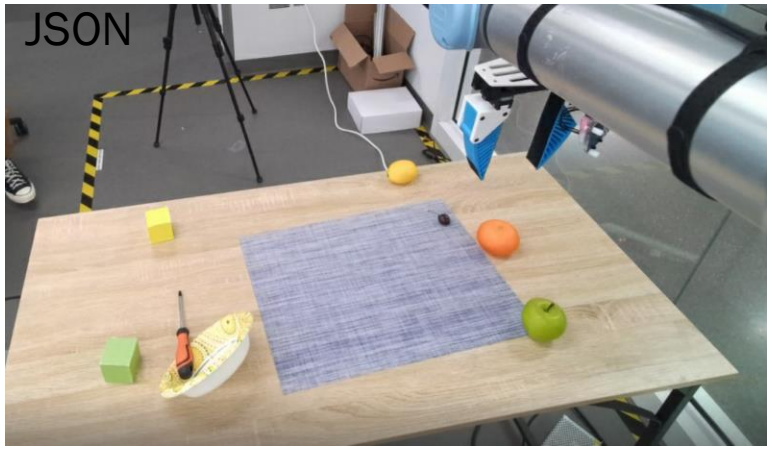


Grasp Reflection Model: gpt-5
Task Completion Model: gpt-5
Bbox Model: Qwen2.5VL-7b

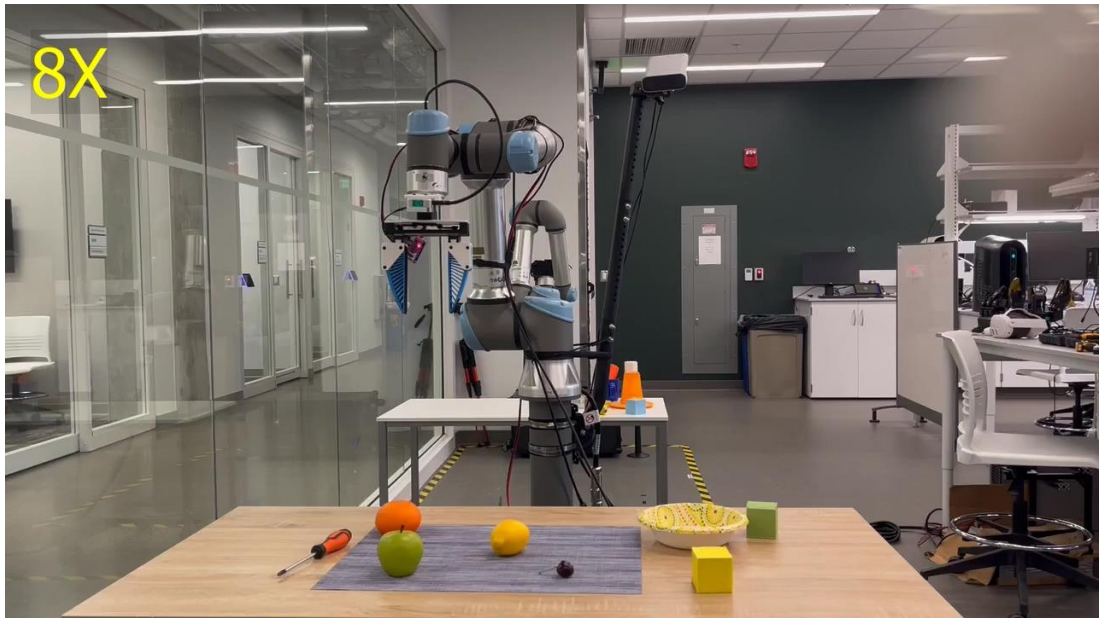


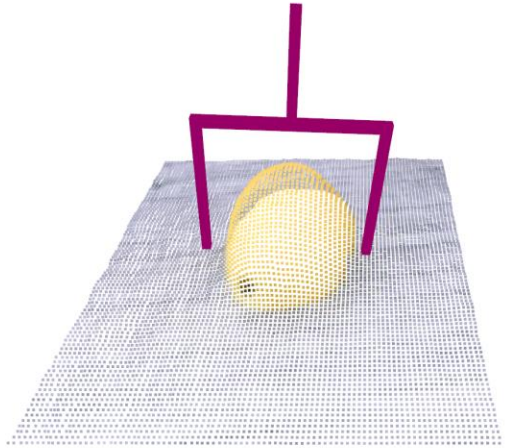
Screwdriver PicknPlace

Task Completion Eval, output in JSON

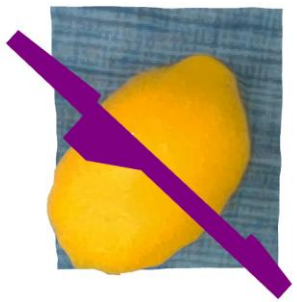



```
{
  "success": true,
  "completion_level": "SUCCESS",
  "confidence": 1.0,
  "task_relevant_objects": {
    "required_objects": [
      "screwdriver",
      "bowl"
    ],
    "observed": [
      "screwdriver",
      "bowl"
    ],
    "missing": []
  },
  "goal_achievement": {
    "target_relationship": "screwdriver in bowl",
    "is_achieved": true,
    "evidence": "The screwdriver is placed inside the bowl on the table."
  },
  "overall_reasoning": "The screwdriver is successfully placed inside the bowl as required by the task.",
  "failure_indicators": [],
  "elapsed_time": 6.123018026351929
}
```






```
{
  "action": "REQUEST_TOP_DOWN_VIEW",
  "object_correctness": {
    "is_correct_object": true,
    "confidence": 0.85,
    "reasoning": "The gripper jaws are centered around and aligned with the yellow point cloud that clearly represents the lemon. The gripper opening encloses the object's lateral boundaries and the approach direction is consistent with a side grasp on the lemon."
  },
  "collision_check": {
    "overlaps_with_surface": true,
    "reasoning": "The lower portions of the gripper legs appear to intersect the visible table point cloud in this viewpoint: both vertical fingers extend down into or very near the populated grid of surface points. Because the table points are visible at the same depth as the finger ends, there is likely overlap or a very small clearance that cannot be confirmed from this angle."
  },
  "overall_reasoning": "While the grasp is correctly positioned on the target lemon, the current side-perspective shows the gripper fingers extending into the area occupied by the table point cloud. That indicates a potential collision or at least uncertain clearance. Per conservative safety rules, a top-down or additional viewpoint is needed to confirm finger clearance and to adjust approach height or lateral placement before execution.",
  "confidence": 0.7,
```


```
  "overall_reasoning": "While the pose is targeted correctly at the lemon, the gripper visualization overlaps the visible table/towel surface in this perspective. Because of that apparent intersection with the surface and limited perspective (depth/vertical clearance unclear), it is unsafe to execute the grasp without additional viewpoints to confirm no collision with the table.",
  "confidence": 0.7,
  "suggestions": "Provide a top-down / higher-elevation view (camera looking along the gripper approach axis) to confirm vertical clearance between gripper fingers/body and the table. If top-down confirms collision, raise the approach height, shift the grasp slightly along the lemon's long axis toward the center, or rotate the gripper so fingers contact only the fruit surface and not the table. Also provide a depth map or side view to verify gripper finger lengths relative to object thickness.",
  "elapsed_time": 17.41938066482544
```

Thank You

PURDUE UNIVERSITY | College of Engineering