Robot Grocery Shopper

18-848 Autonomous Robotics II
Fall 2025

1. What are we trying to do?

We all use mobile shopping apps, some time, we find it convenient to call an Uber Eat or do Curbside pickup.
However, it still takes time for the associates to pick up the items for us.
Can we make it more convenient for both the customer and the staff working in the shop?

1. What are we trying to do?

We are trying to make a dual-arm robot with mobile base to:

- Self-identify grocery items on the shelf,
- Automatically identify, scan, and calculate the 6D pose of the target items.
- Plan and execute the optimal trajectory to pick the item from the shelf and place into basket
- Maintaining safe during the whole process.

0. Updates from pre-demo

https://drive.google.com/file/d/1bvqm4vzGCk33N7ORQ7mD8Hd8S8-6cLIj/view?usp=share_link

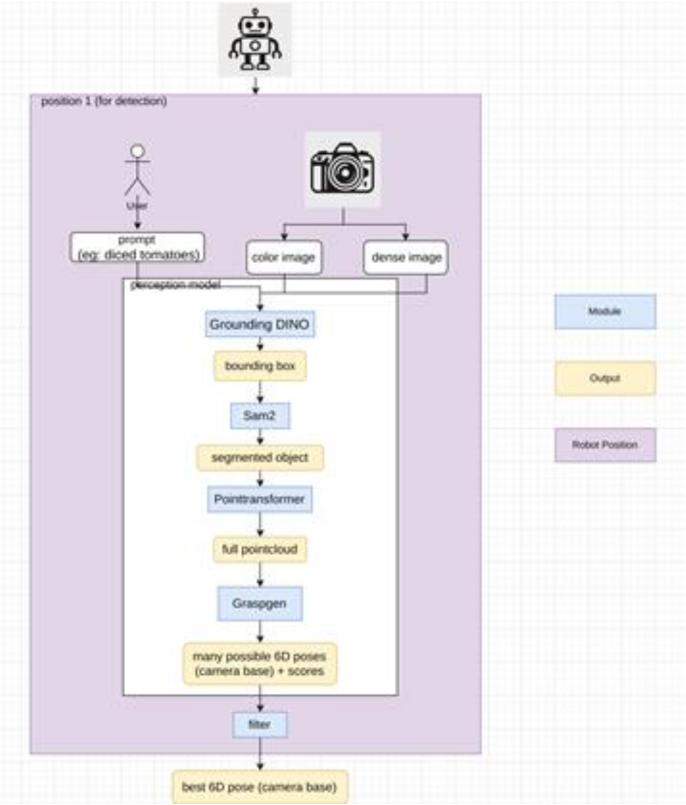
Updates from pre-demo

[redacted]

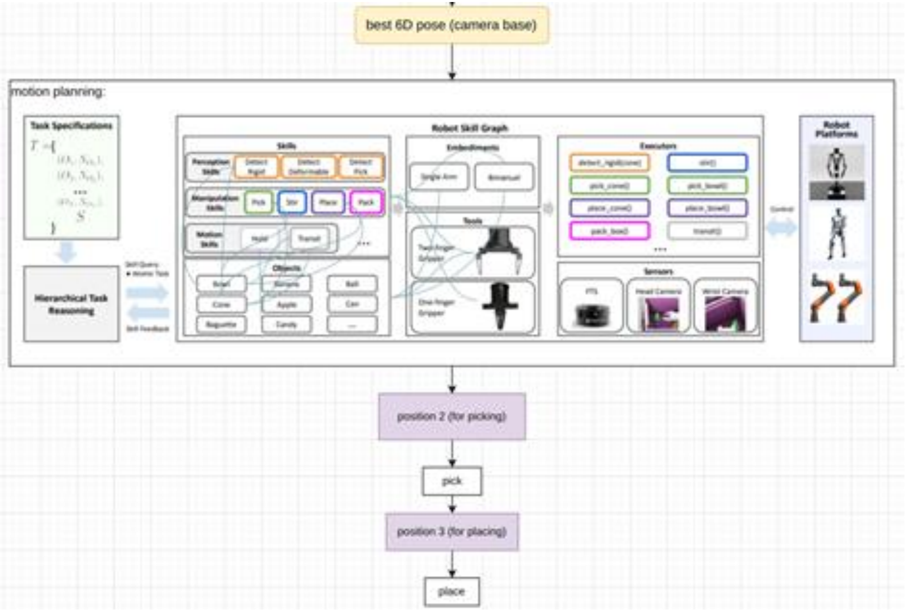
[redacted]

2. Pipeline

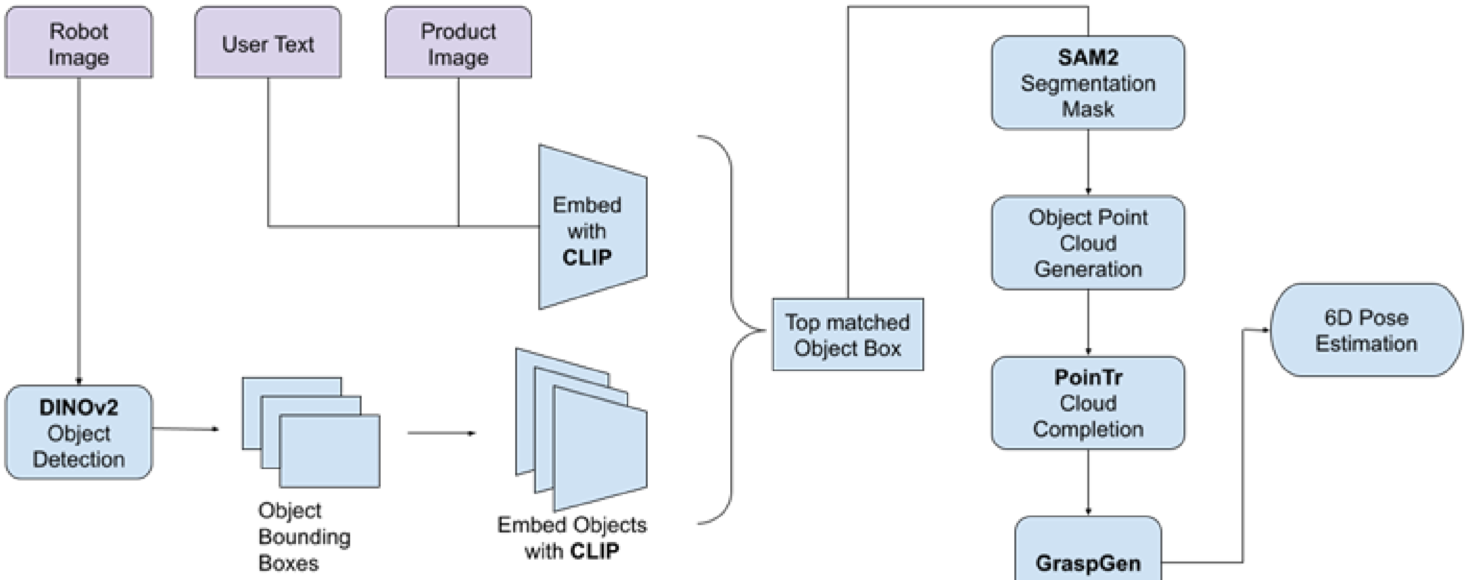
Perception



Manipulation



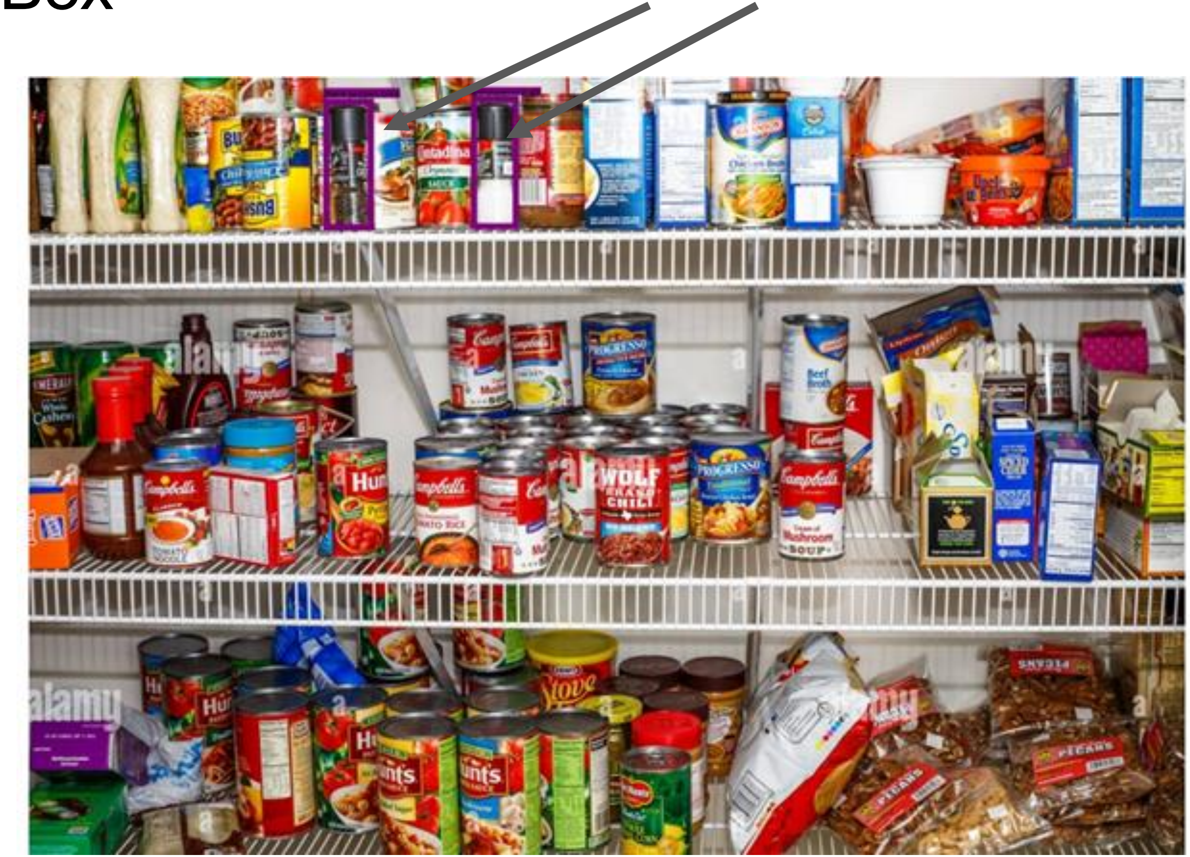
2.1 Perception Pipeline



Get 2D Bounding Box

> Salt and Pepper Shakers

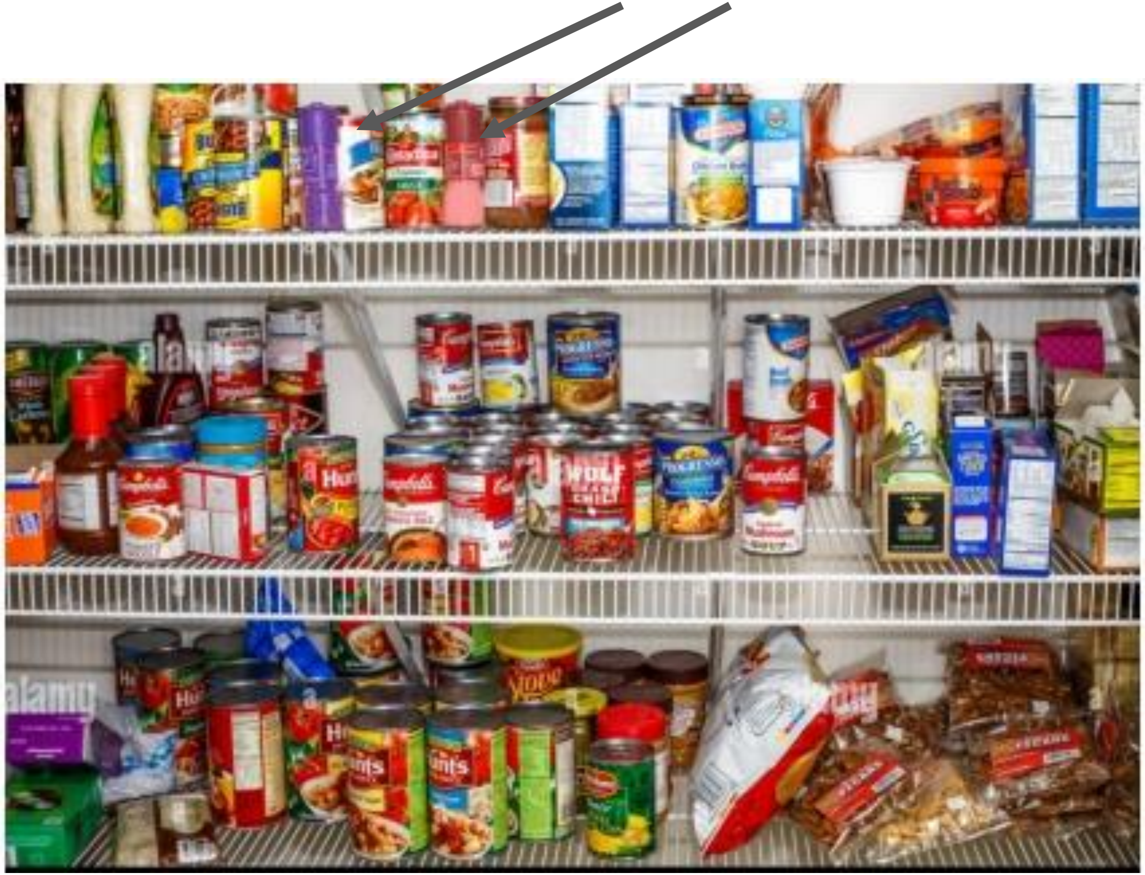
User Input



Get 2D Segmentation Mask

Salt and Pepper Shakers

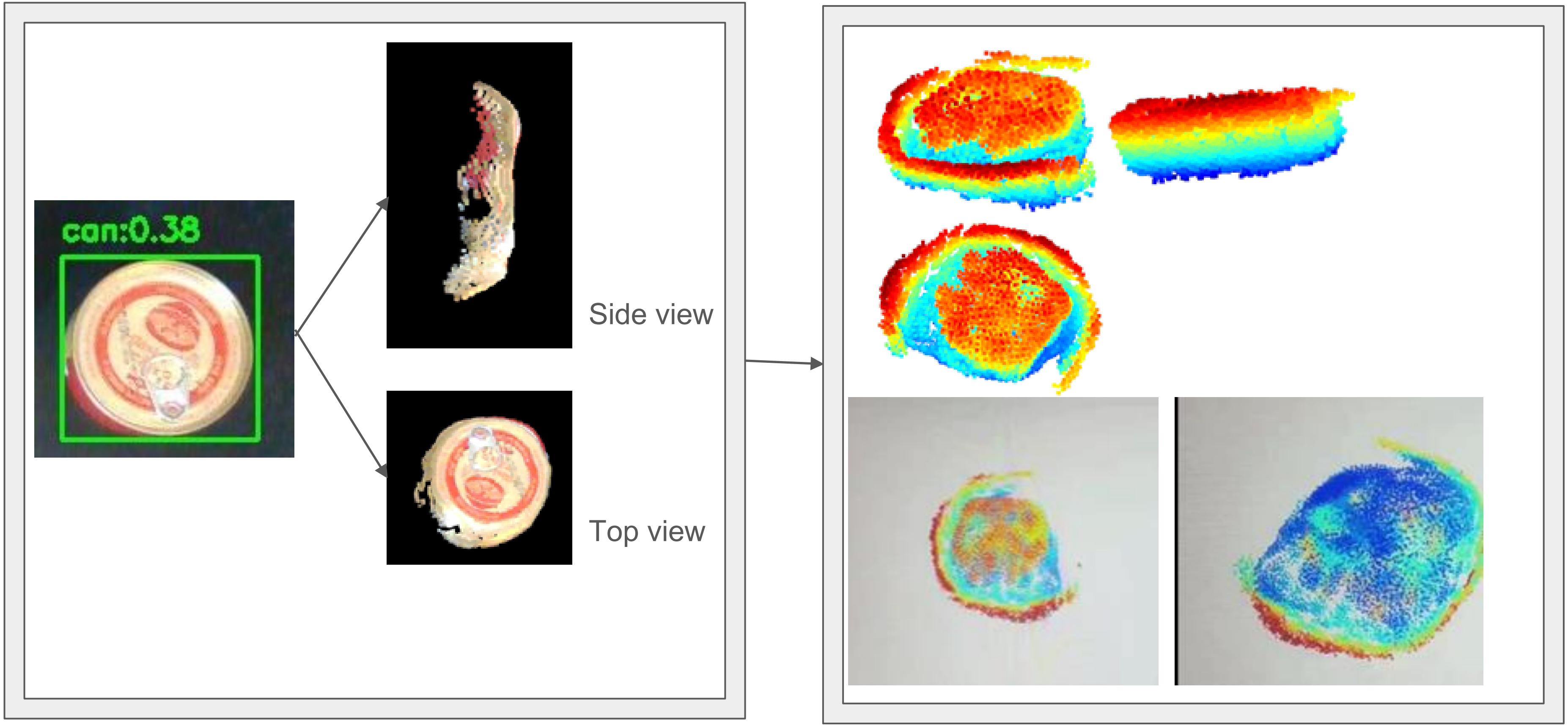
User Input



Original Shelf Image

Isolated Masked Region (Zoomed In)

The performance of PoinTr(point cloud completion model)

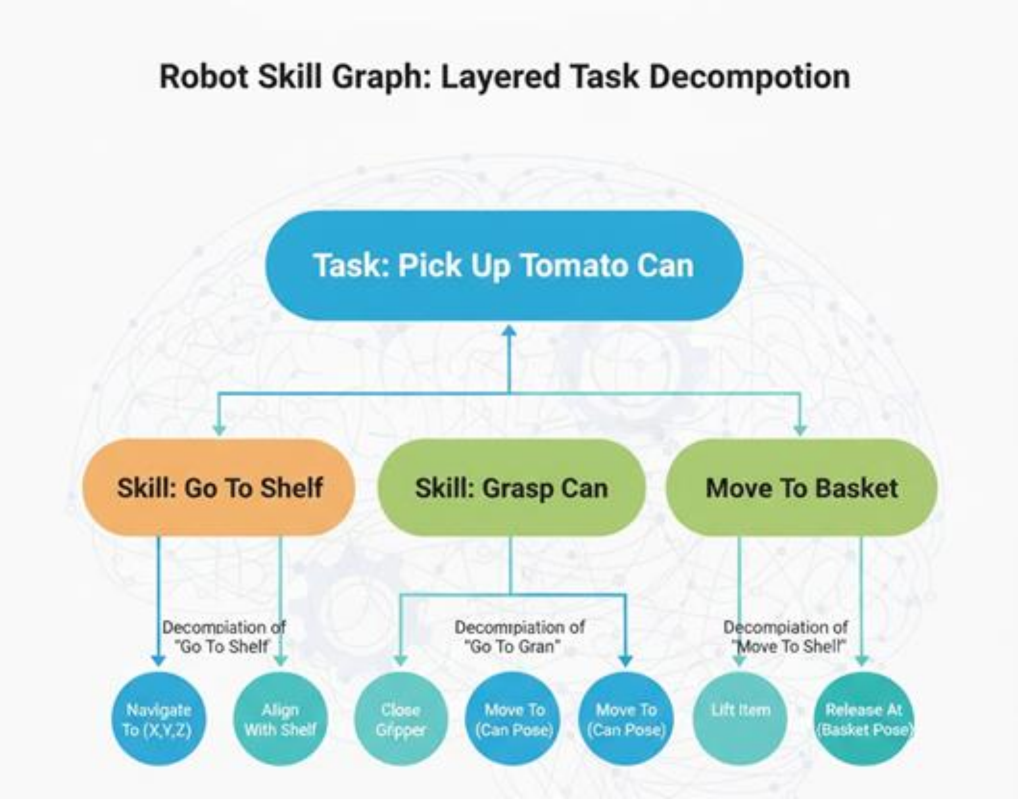


Side view

Top view

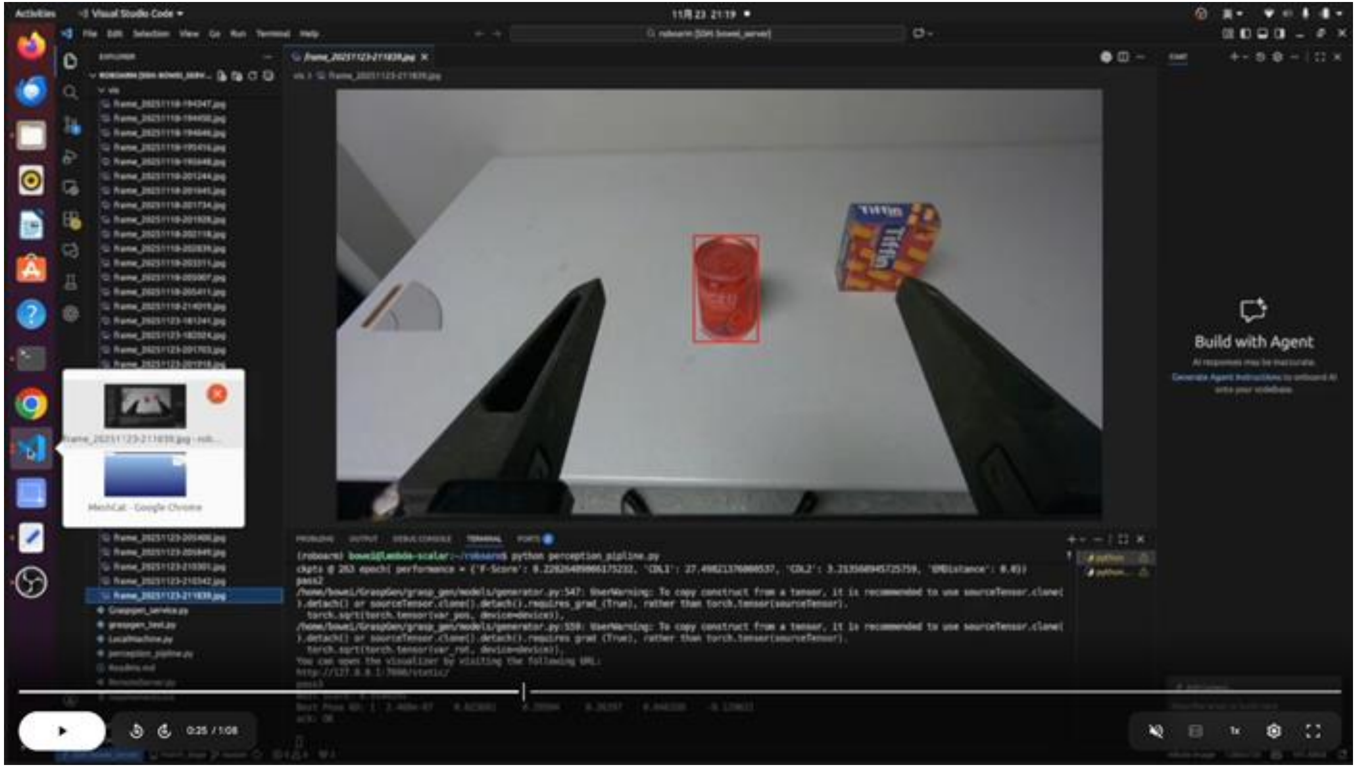
Gasping point showed in the

2.2 Manipulation Pipeline



2.2 Manipulation In Simulation

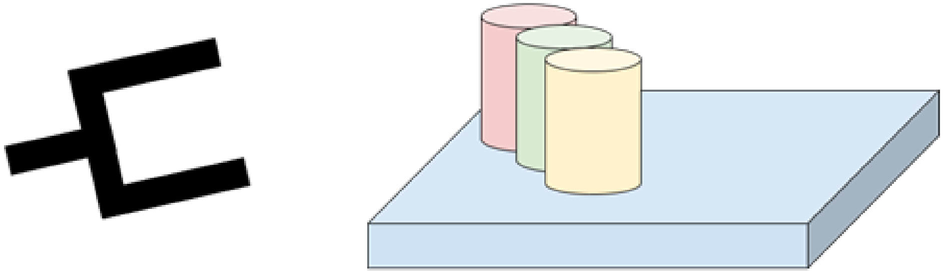
3. System Demo (AKA Manipulation in real life)


Activities
Visual Studio Code
11月23 21:19
File Edit Selection View Go Run Terminal Help
roboarm [SSH: bowei_server]
frame_20251123-211939.jpg
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
Build with Agent
AI responses may be inaccurate.
Generate Agent Instructions to onboard AI onto your codebase.
frame_20251123-211939.jpg - rob...
MeshCat - Google Chrome
0:25 / 1:08
1x

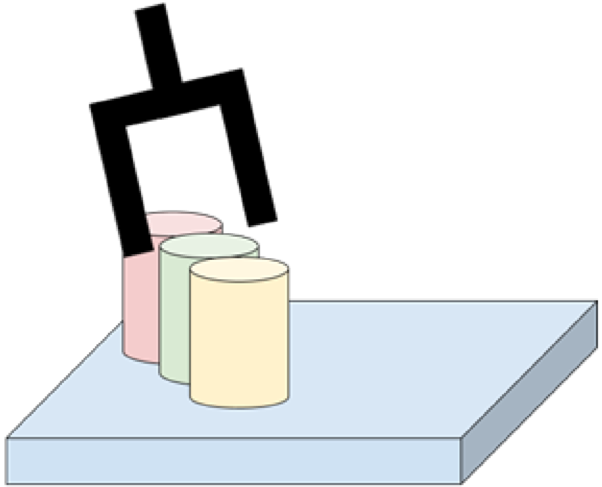
```
(roboarm) bowei@lambda-scalar:~/roboarm$ python perception_pipline.py
ckpts @ 263 epoch( performance = {'F-Score': 0.22826489866175232, 'CDL1': 27.49821376000537, 'CDL2': 3.2135689457257 59, 'EMDistance': 0.0}}
pass2
/home/bowei/GraspGen/grasp_gen/models/generator.py:547: UserWarning: To copy construct from a tensor, it is recommended to use sourceTensor.clone(
).detach() or sourceTensor.clone().detach().requires_grad_(True), rather than torch.tensor(sourceTensor).
  torch.sqrt(torch.tensor(var_pos, device=device)),
/home/bowei/GraspGen/grasp_gen/models/generator.py:559: UserWarning: To copy construct from a tensor, it is recommended to use sourceTensor.clone(
).detach() or sourceTensor.clone().detach().requires_grad (True), rather than torch.tensor(sourceTensor).
  torch.sqrt(torch.tensor(var_rot, device=device)),
You can open the visualizer by visiting the following URL:
http://127.0.0.1:7000/static/
pass3
```

Approaching Crowded Shelf Space

Approaching Crowded Shelf Space



Approaching Crowded Shelf Space



Updates from pre-demo

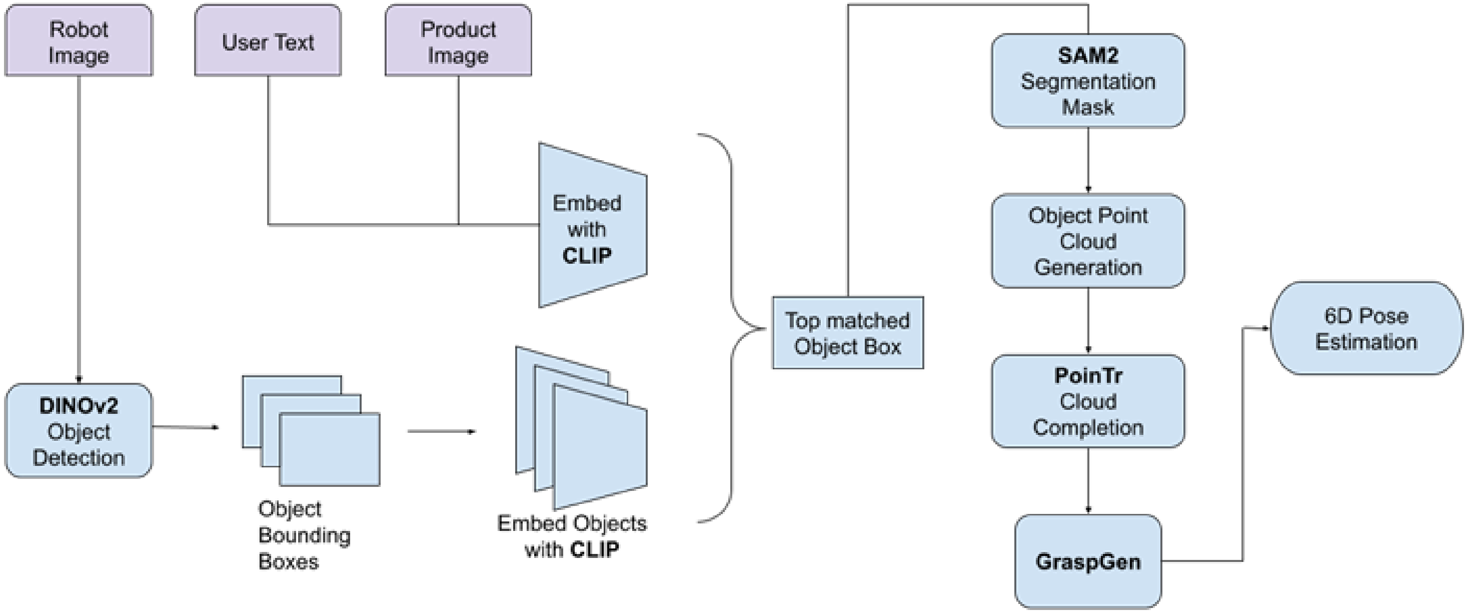
4. Future Work

- Dual-arm collaboration
- More comprehensive collision detection
- More user-friendly UI and LLM understanding

Thanks

Arm Grocery Shopper Team
Ranit, Bowei, Yikuan, Jianlin, Yuanliu, Han, Lihao

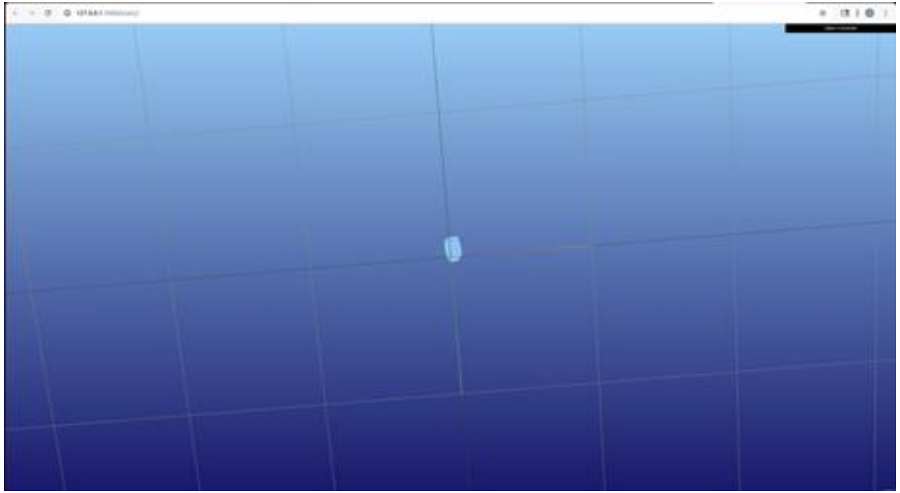
Vision Pipeline

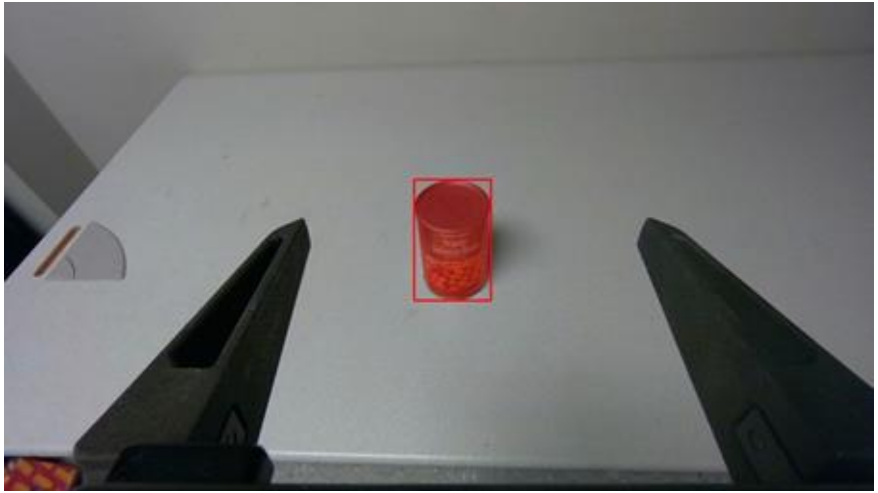

Robot Image
User Text
Product Image
Embed with **CLIP**
DINOv2 Object Detection
Object Bounding Boxes
Embed Objects with **CLIP**
Top matched Object Box
SAM2 Segmentation Mask
Object Point Cloud Generation
PoinTr Cloud Completion
GraspGen
6D Pose Estimation

Fixing Grasp Gen

1. Upright Filter
2. Gripper Offset

Upright Filter

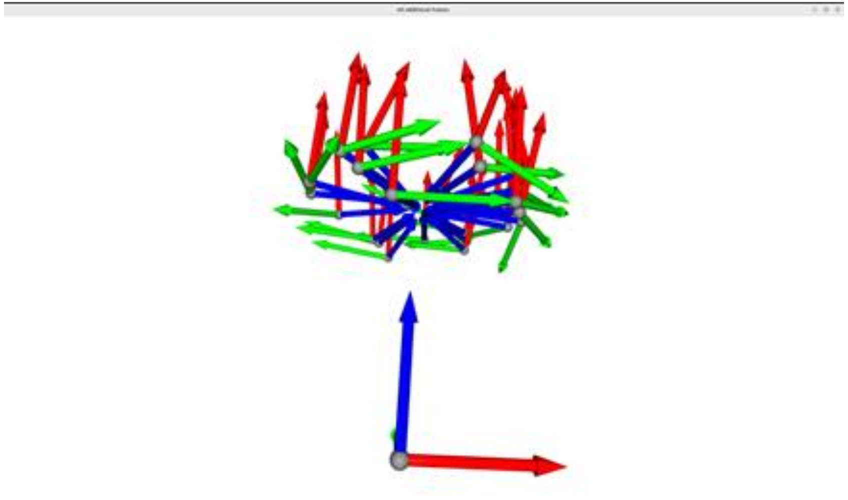




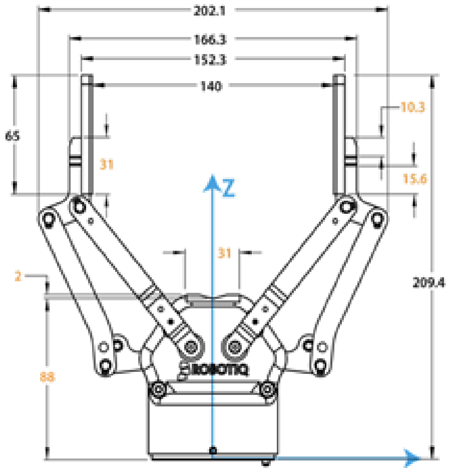
Filtered pose close to the camera(pose with the smallest Z value): in “can” case, pose directly facing towards the object will be the best pose, not the side poses.

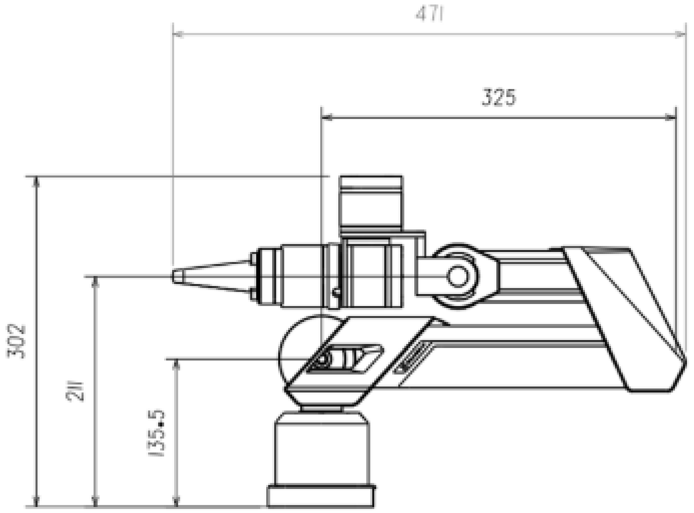
Upright Filter

Issue: Given the upright filter around cam Z axis=True



Gripper Offset





Skill graph (motion planning)

Input: best 6D pose (camera base)

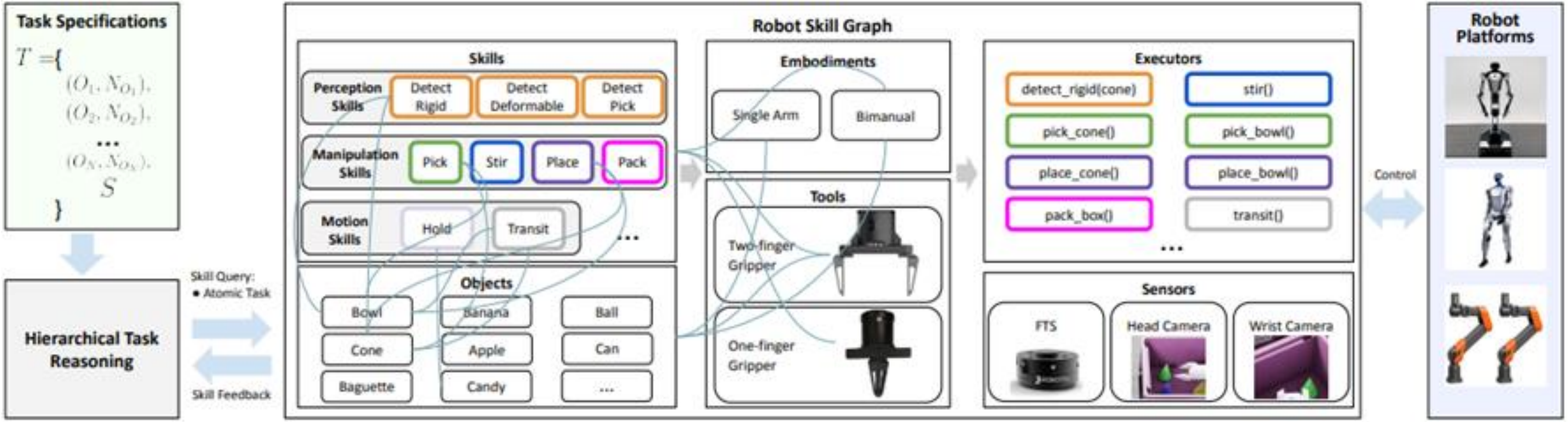
Output: action

Implementation:

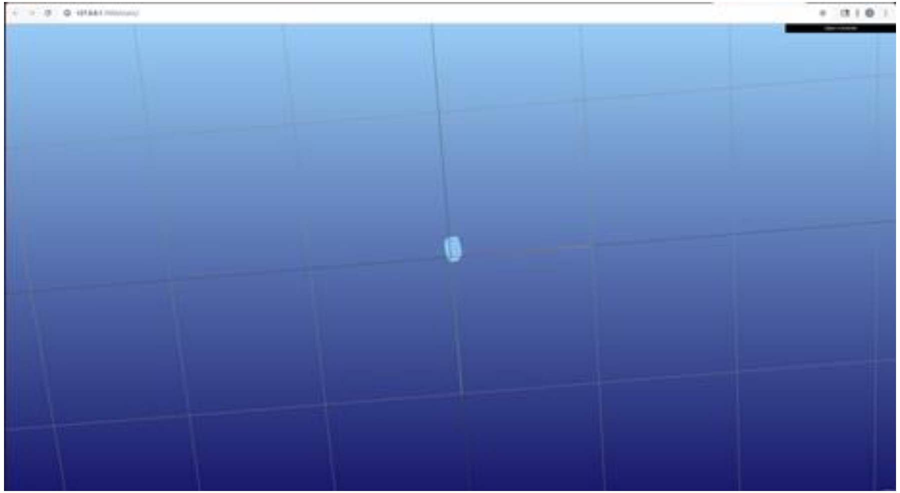
Process 1: transform 6D pose (camera base) to 6D pose (robot base)

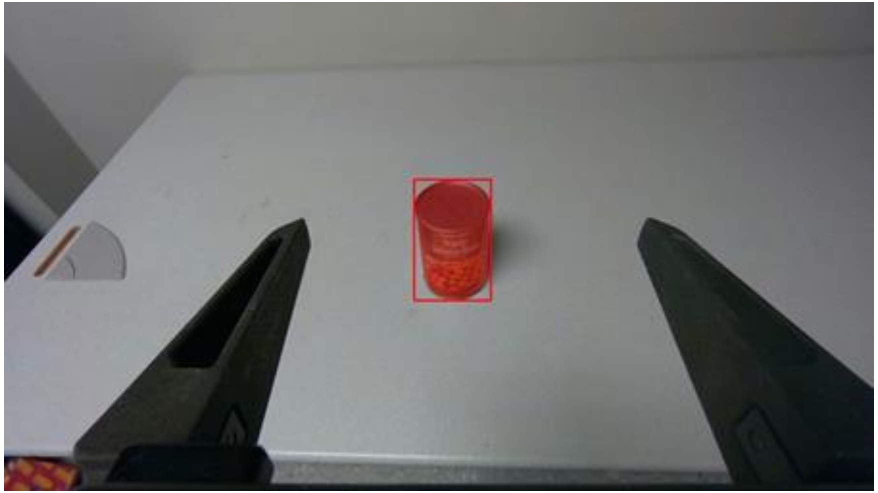
Process 2: planning trajectory according to 6D pose (robot base)

Process 3: reach target position



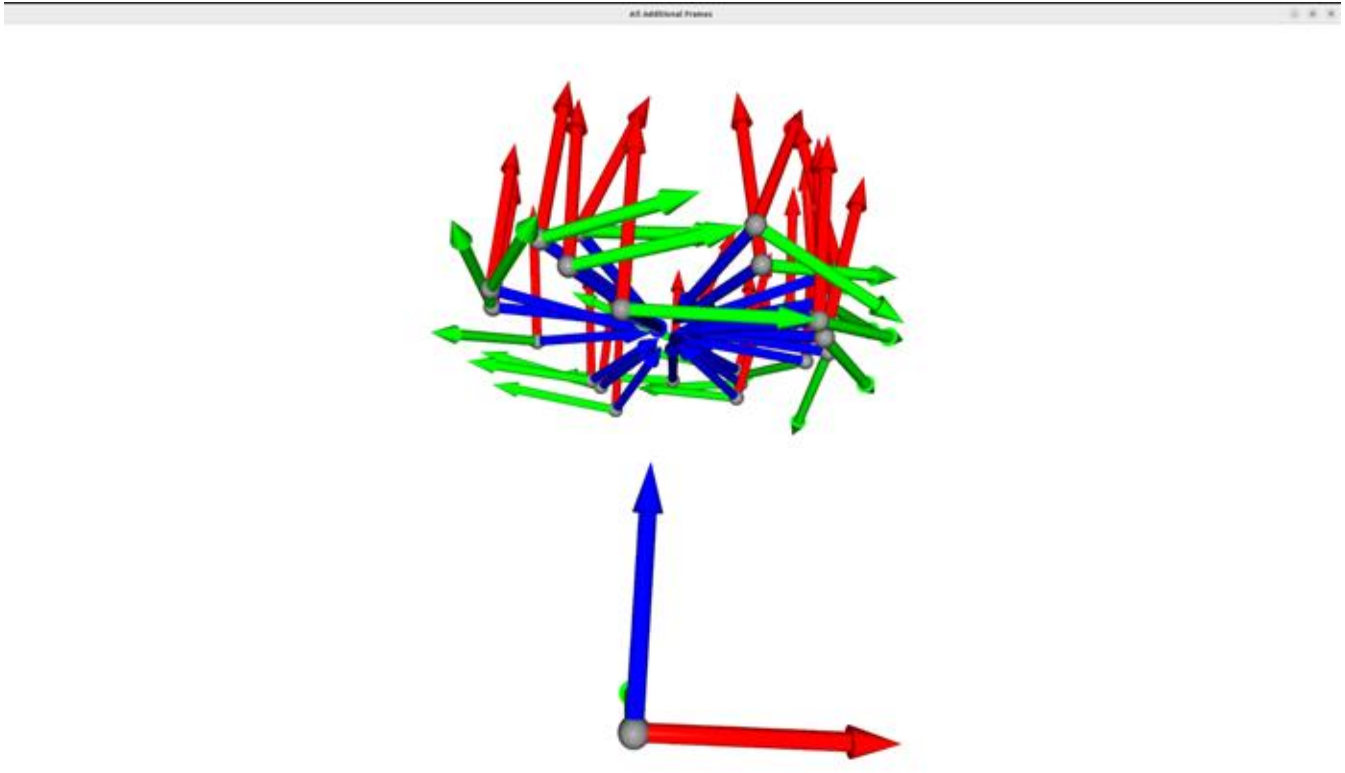
11.19.2025





Filtered pose close to the camera(pose with the smallest Z value): in “can” case, pose directly facing towards the object will be the best pose, not the side poses.

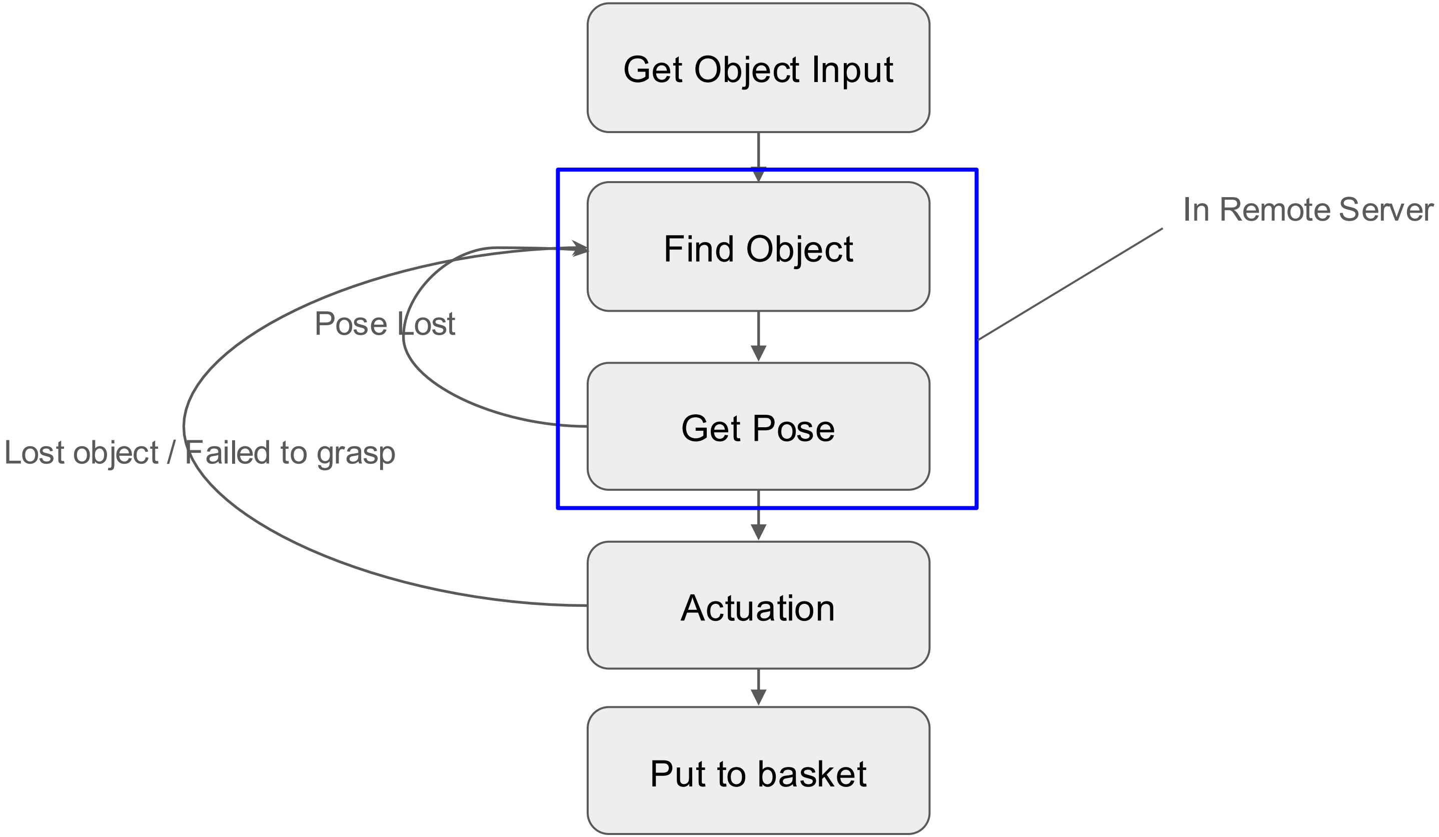
issue: Given the upright filter around cam Z axis=True



Motions in sim environment

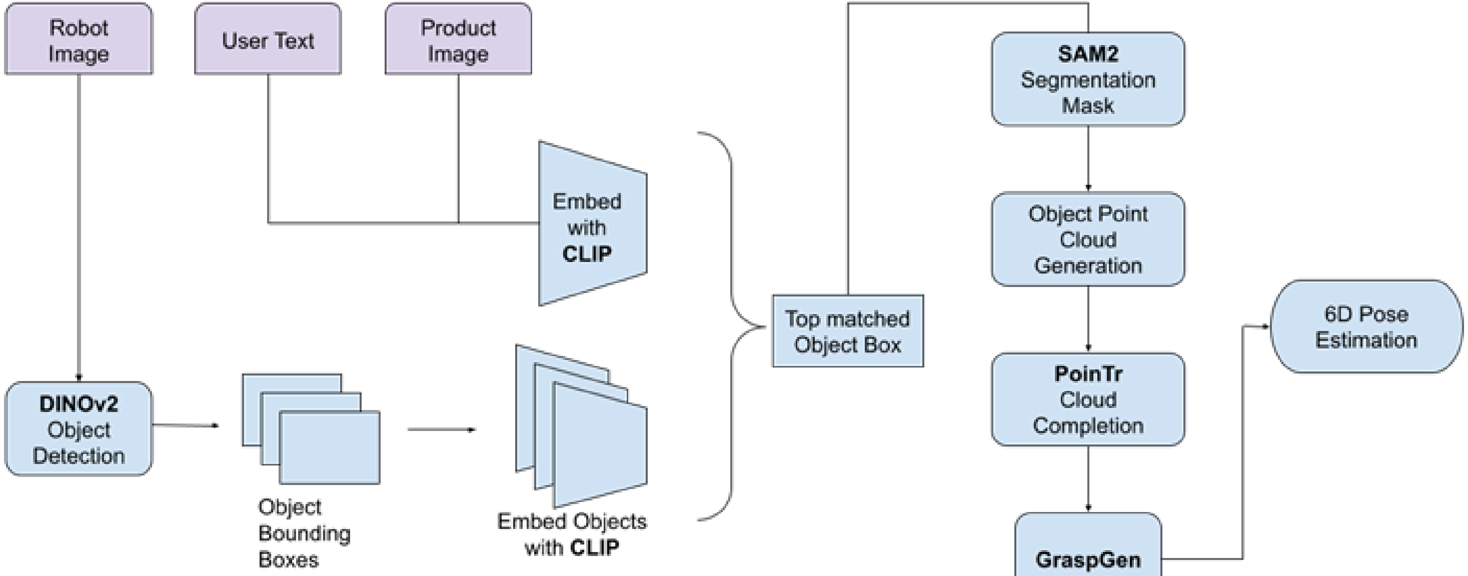
11.7.2025

Demo Day!



Get Object Input
Find Object
Get Pose
Actuation
Put to basket
Pose Lost
Lost object / Failed to grasp
In Remote Server

Vision pipeline


Robot Image
User Text
Product Image
Embed with **CLIP**
DINOv2 Object Detection
Object Bounding Boxes
Embed Objects with **CLIP**
Top matched Object Box
SAM2 Segmentation Mask
Object Point Cloud Generation
PoinTr Cloud Completion
GraspGen
6D Pose Estimation

Pipeline

Initial: Go to the detection position

Perception

Input：color image + dense image + user input (text prompt)

Output：best 6D pose (camera base)

Implementation:

1. Dino: text prompt→bounding box
2. Sam2 : boudingbox—> segmented object
3. Pointtransformer: mask + dense information →full pointcloud
4. Graspgen: full pointcloud—->many possible 6D poses + scores(camera base)
5. Filter: choose the best grasp position

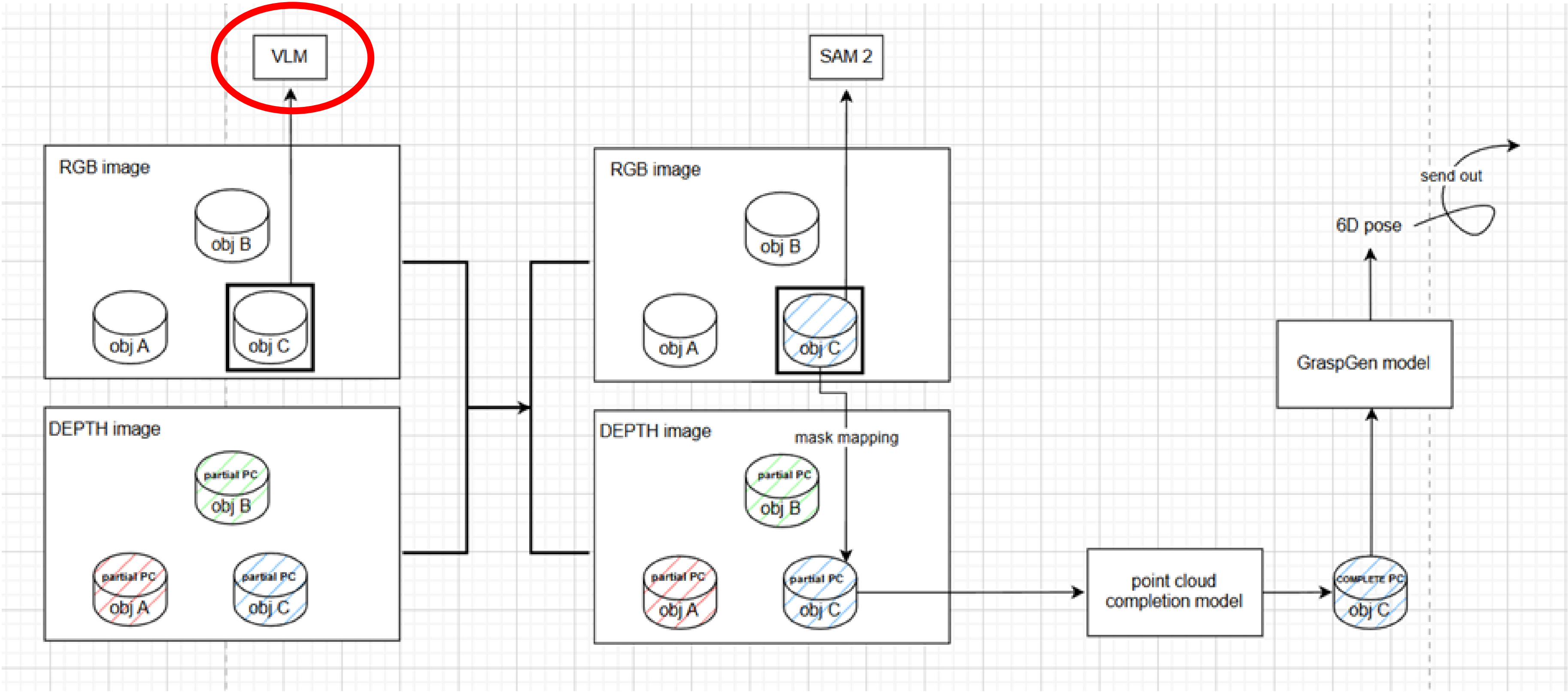
   Implementation: Relative to the camera's most recent position

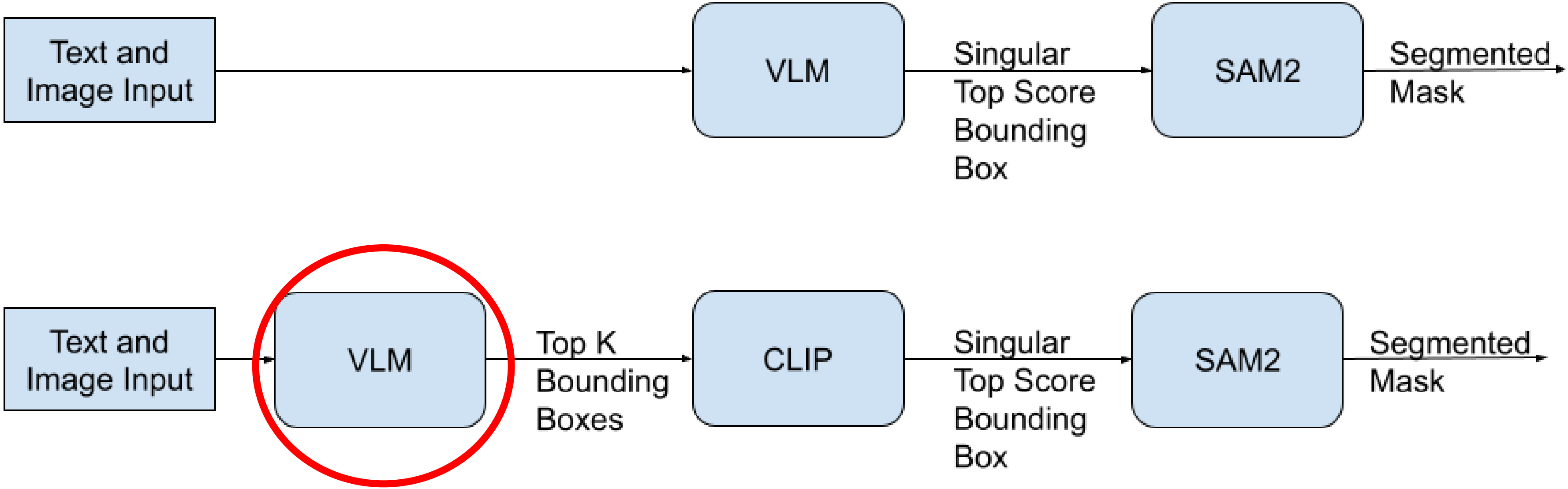
Manipulation – Safe Trajectory Selection

https://drive.google.com/file/d/17BLArBKfSiBW9LJ05nhuEEca3uQim3fs/view?usp=sharing

10.20.2025

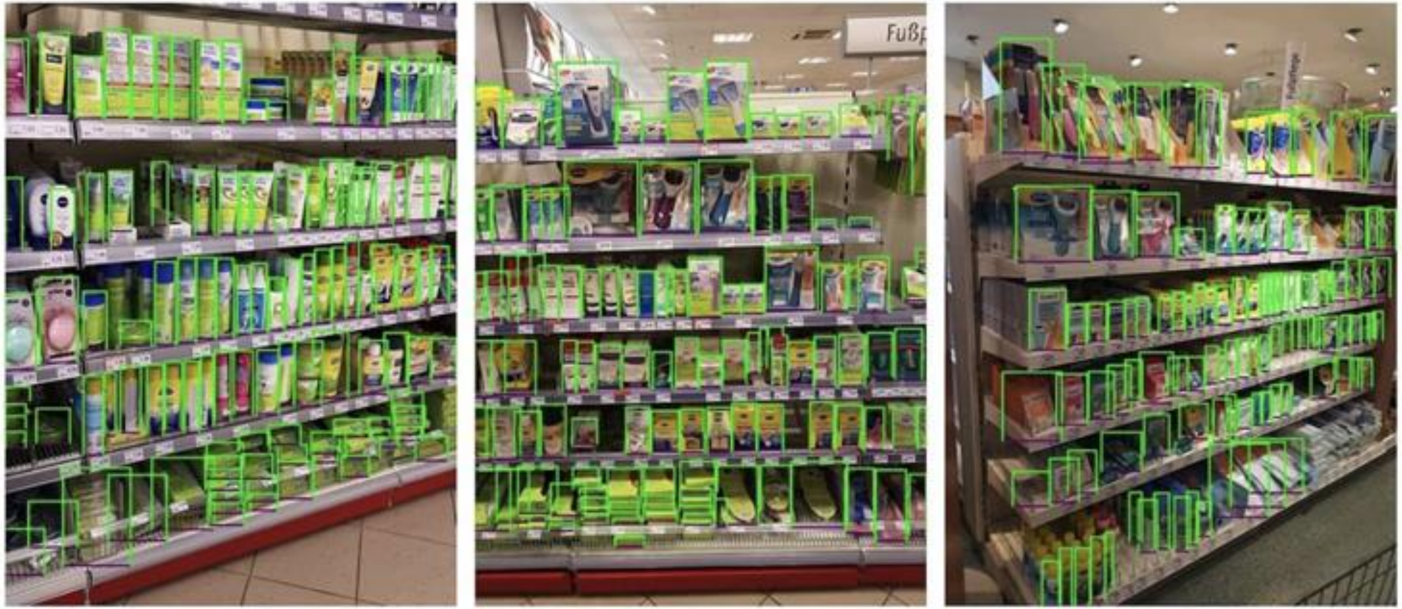
A general view of vision pipeline

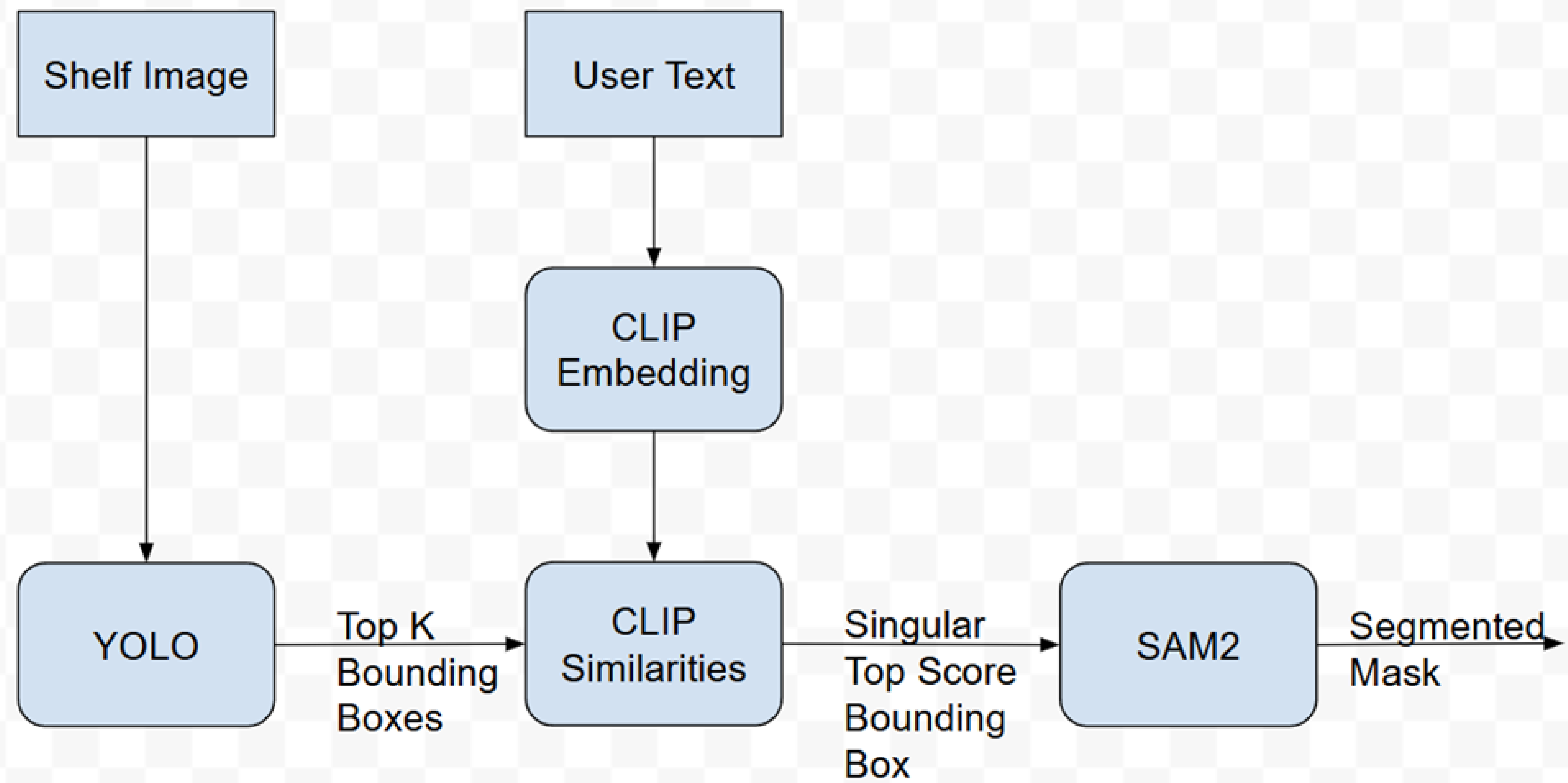



Text and Image Input
VLM
Singular Top Score Bounding Box
SAM2
Segmented Mask
Text and Image Input
VLM
Top K Bounding Boxes
CLIP
Singular Top Score Bounding Box
SAM2
Segmented Mask

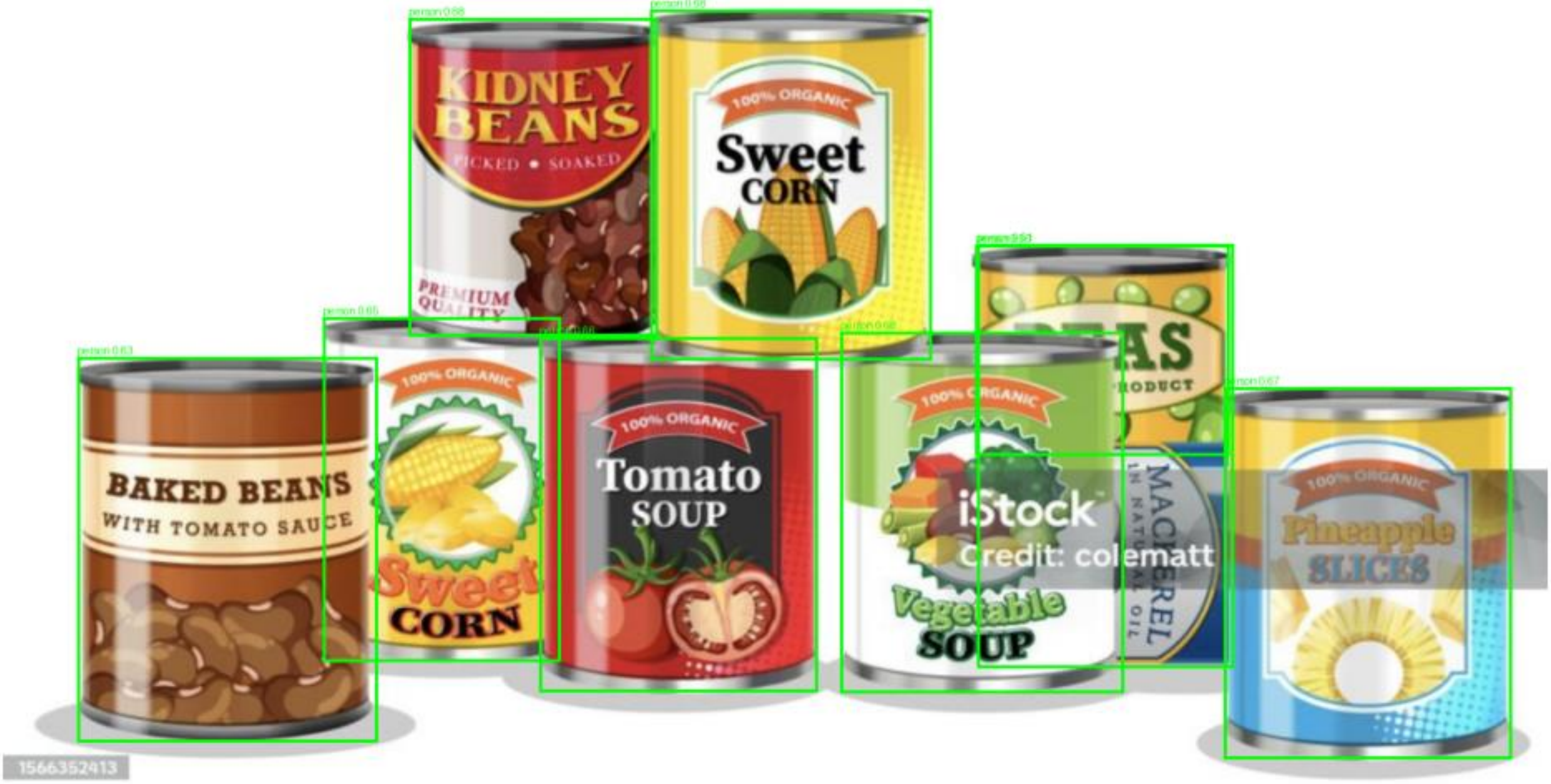
Using Yolo v5 trained on SKU Dataset

- 110K images of densely packed shelves of products
- Generic “product” label that detects generic packaging items like cans, boxes, bottles, etc.



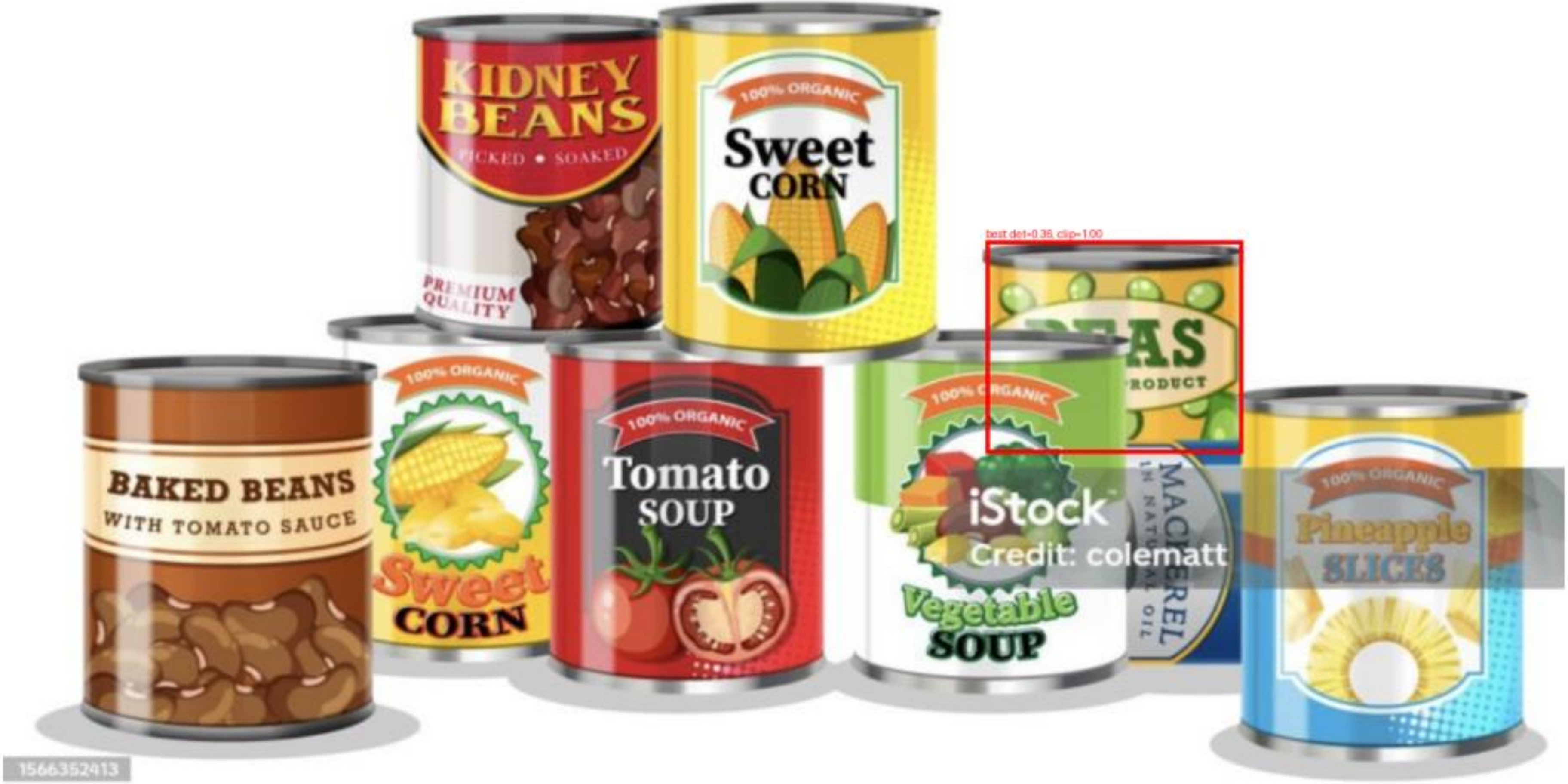

Shelf Image
User Text
CLIP Embedding
YOLO
Top K Bounding Boxes
CLIP Similarities
Singular Top Score Bounding Box
SAM2
Segmented Mask

All Box Detections (pre-CLIP)



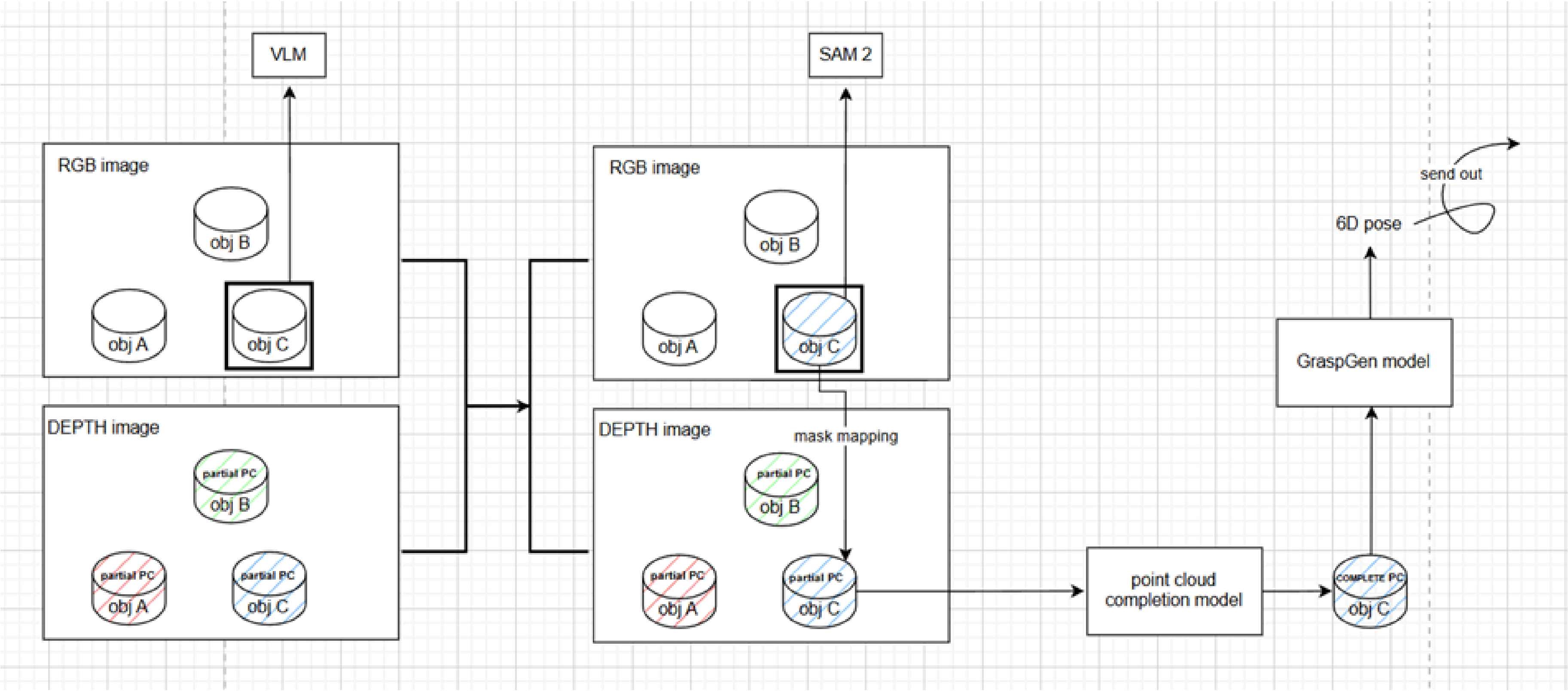
"Can of Peas"

CLIP Re-ranked Detection

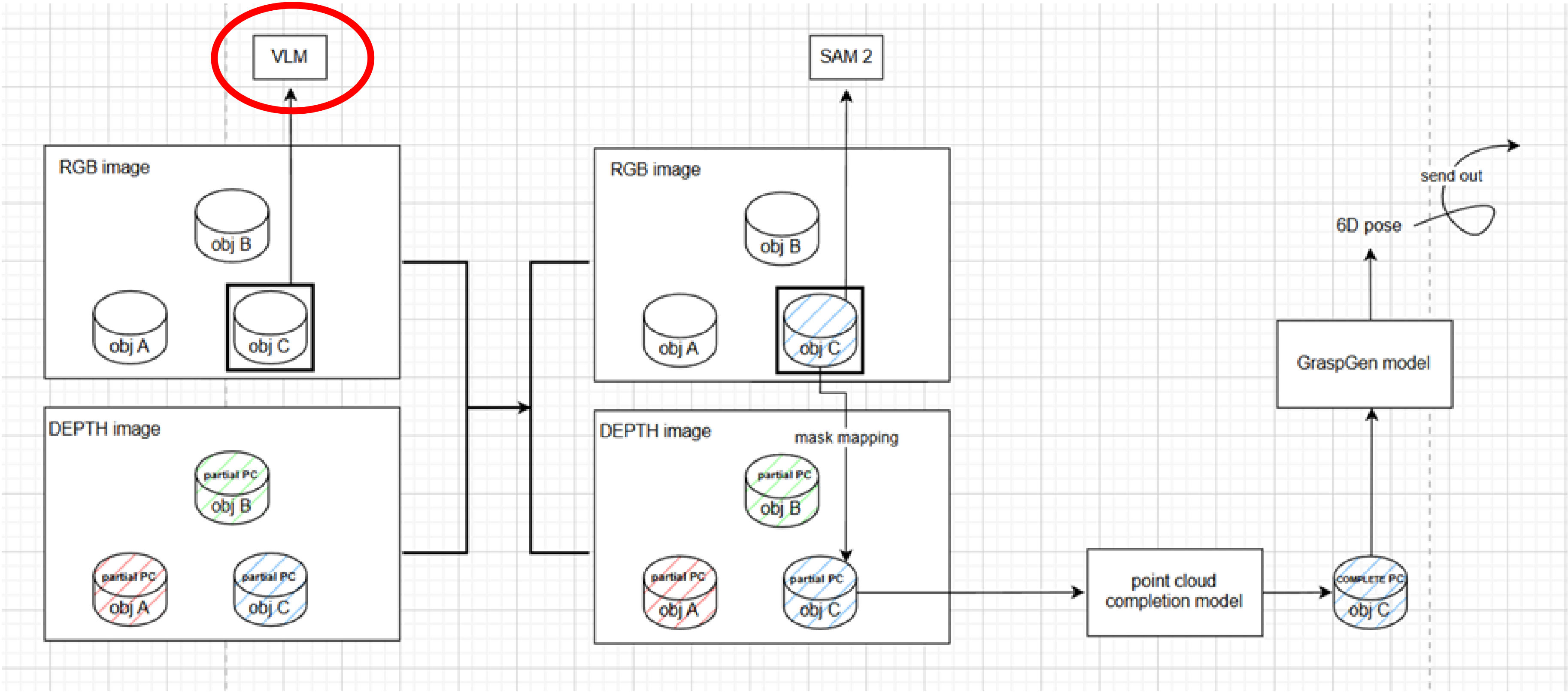


10.06.2025

A general view of vision pipeline



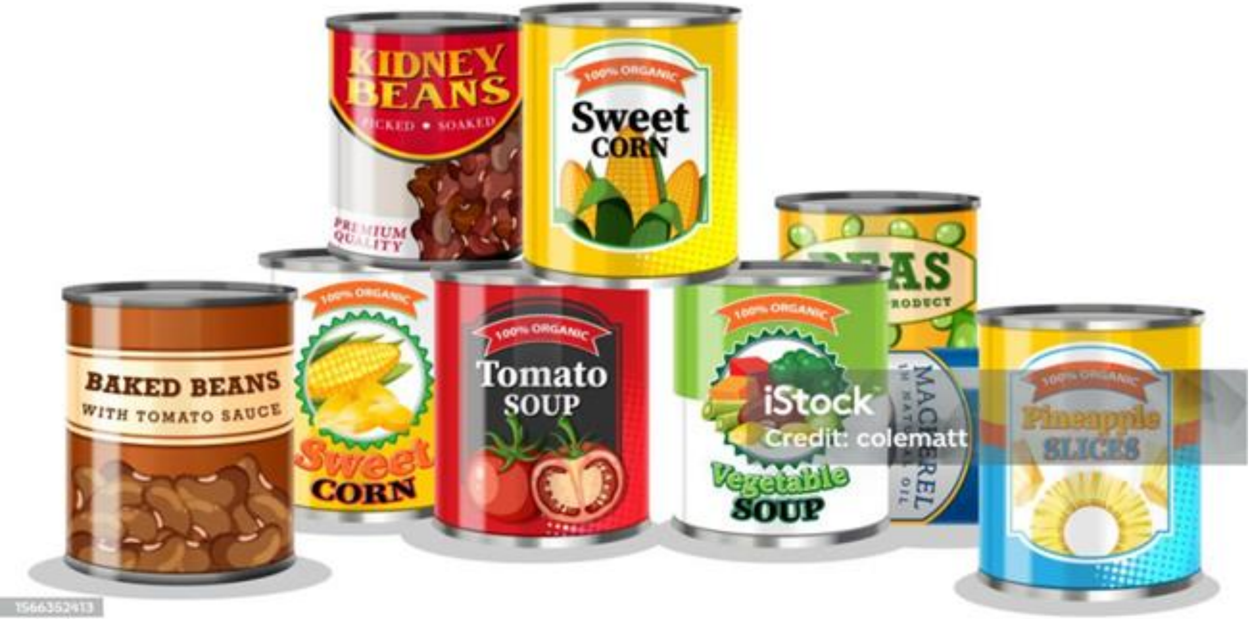
Issue: One Shot Object Detection using VLMs

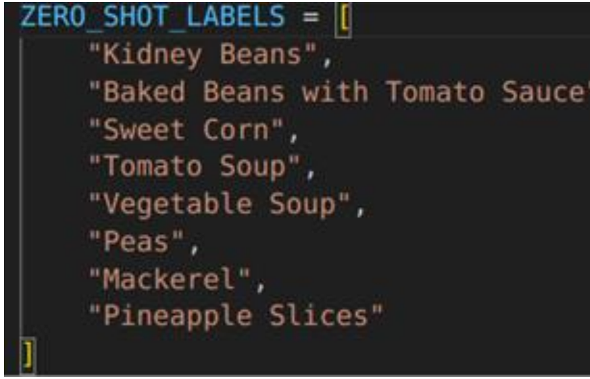


Test 1: Testing Multiple VLMs

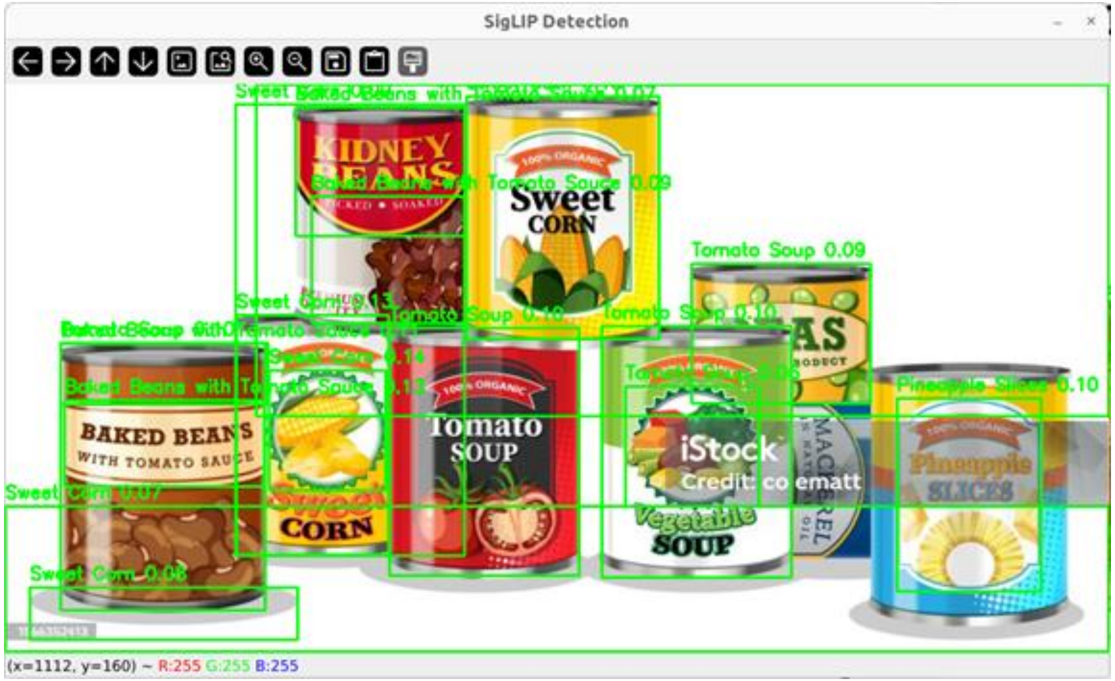
Performance Comparison

1. SIgLIP By Google DeepMind (Vision-language embedding model)
2. OWLV2 By Google Research (Open-vocabulary object detector)
3. QWen2.5-VL By Alibaba (General-purpose multimodal LLM (VLM + LLM))



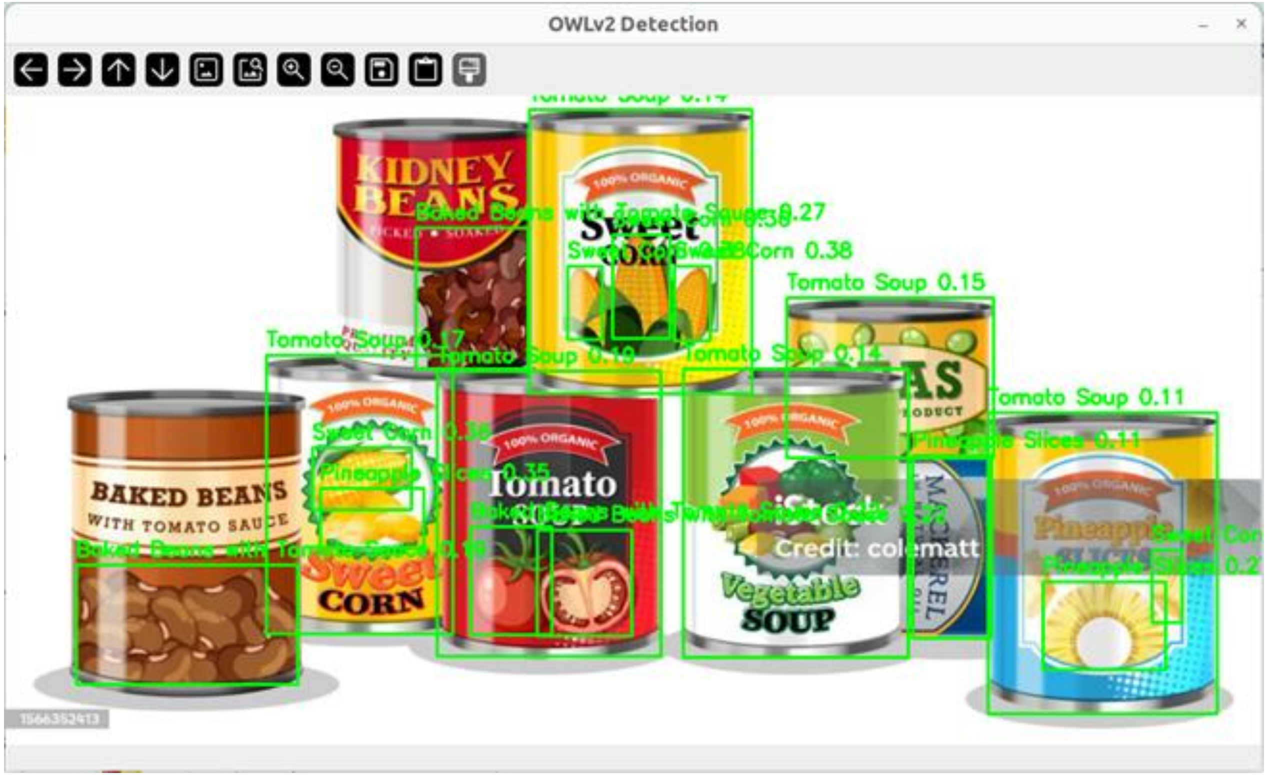

```
ZERO_SHOT_LABELS = [
        "Kidney Beans",
        "Baked Beans with Tomato Sauce
        "Sweet Corn",
        "Tomato Soup",
        "Vegetable Soup",
        "Peas",
        "Mackerel",
        "Pineapple Slices"
]
```

SigLIP(Sigmoid CLIP)

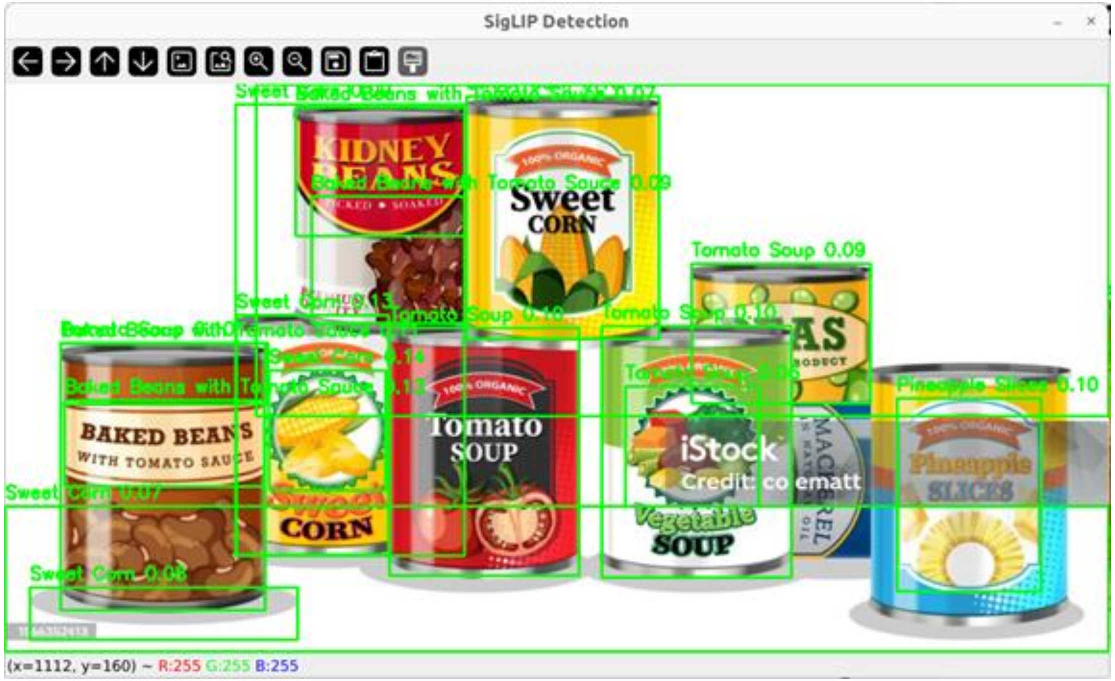


(x=1112, y=160) ~ R:255 G:255 B:255

OWLV2(Open-World Localization v2)



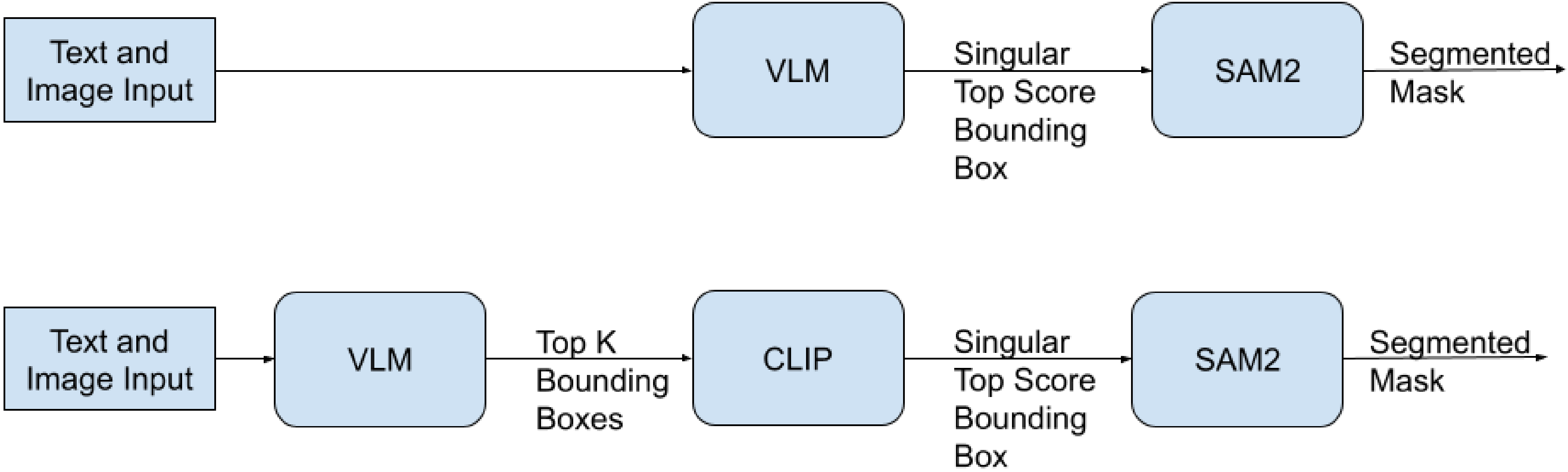
QWen2.5-VL



(x=1112, y=160) ~ R:255 G:255 B:255

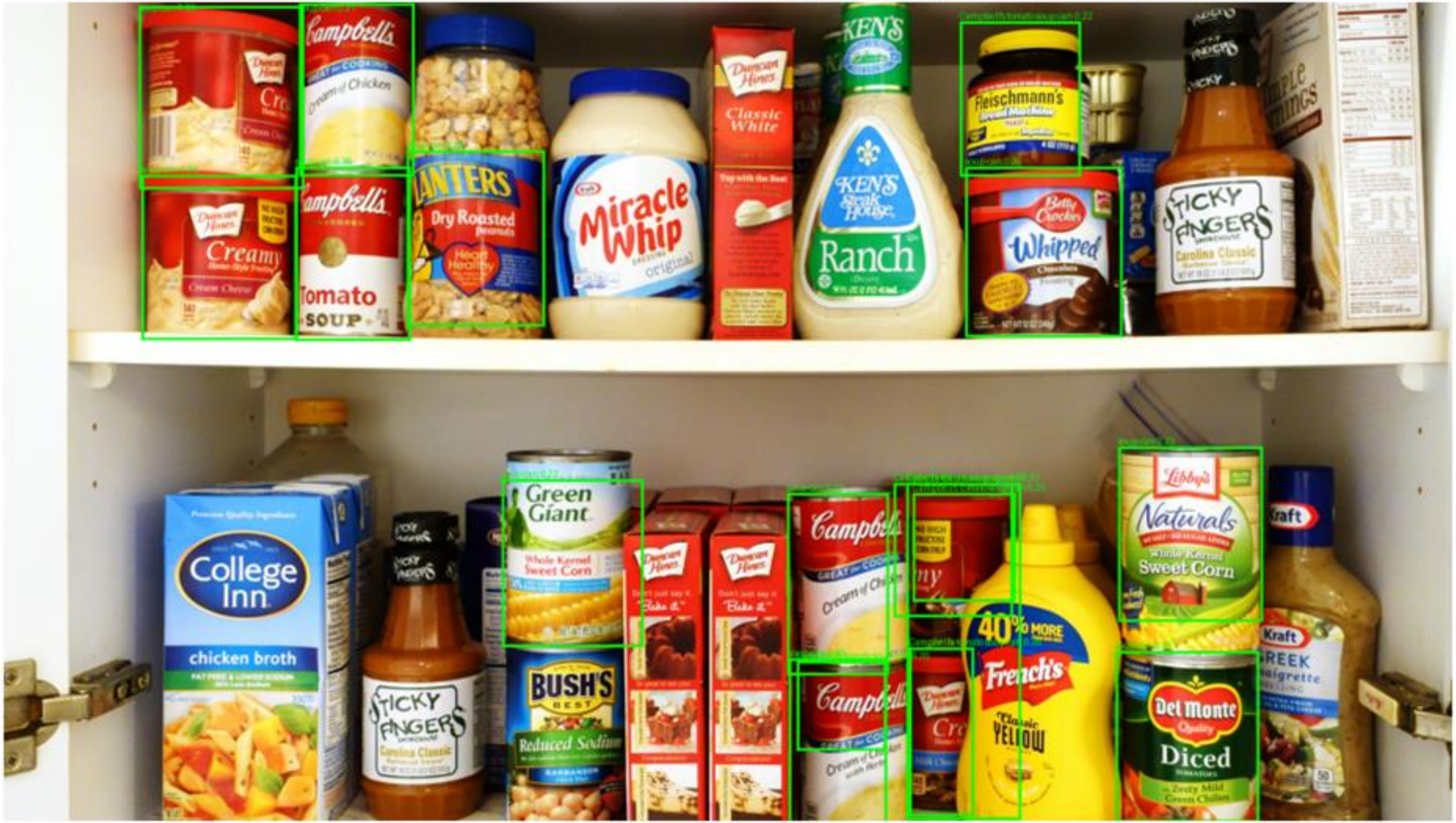
Model	**Type**	**Strength**	**Weakness**	**Best Use**
OWLv2	Detector + Text	Accurate boxes, multi-class	Limited open-world generalization	Zero-shot detection, SAM pre-masking
SigLIP	Embedding	Strong alignment, fast	No spatial grounding	Retrieval, embedding similarity
Qwen 2.5-VL	Generative VLM	Rich semantics, multilingual	No detection, heavy	Reasoning, VQA, semantic grounding

Test 2: Adding CLIP Reranking

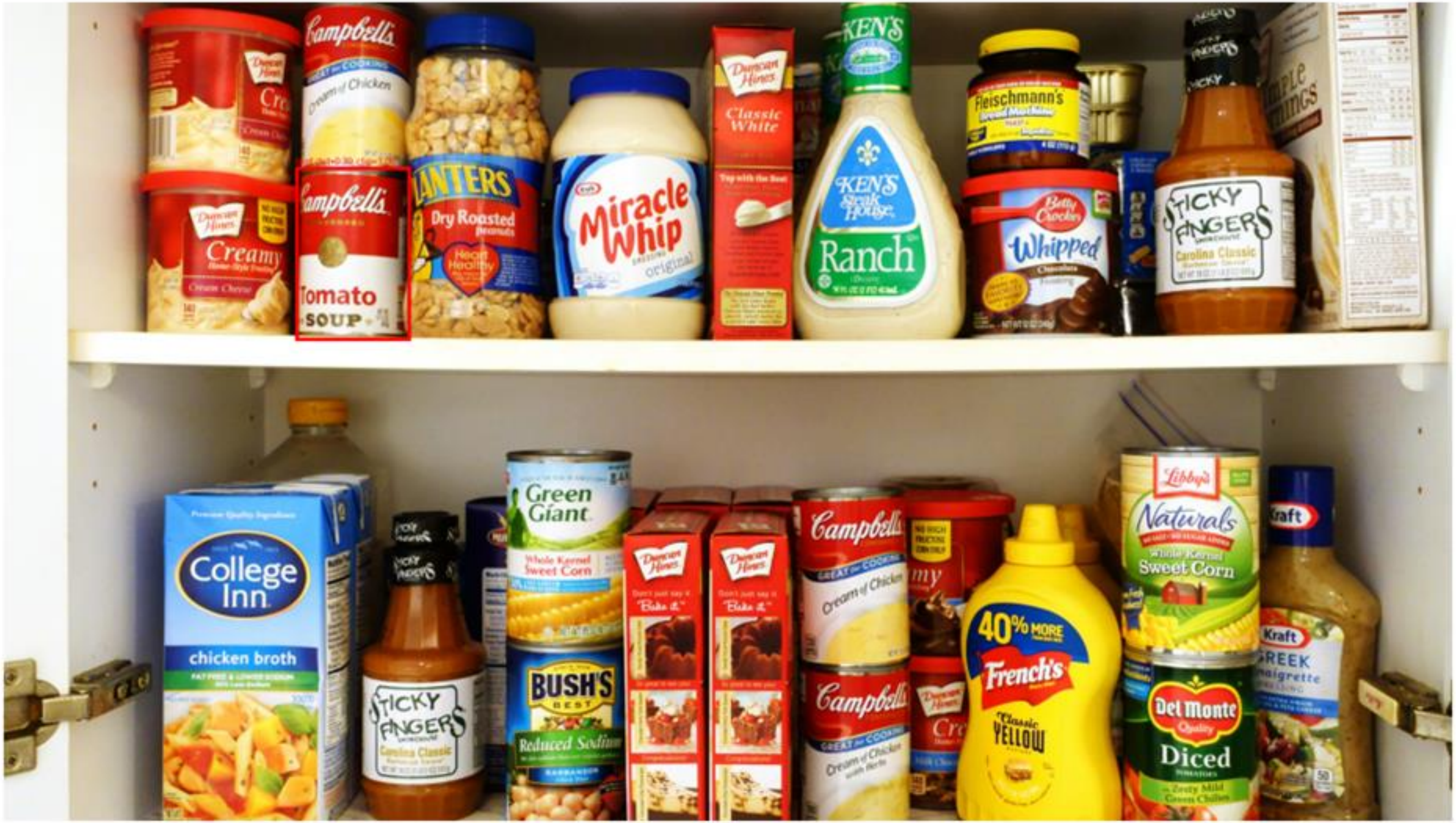

Text and Image Input
VLM
Singular Top Score Bounding Box
SAM2
Segmented Mask
Text and Image Input
VLM
Top K Bounding Boxes
CLIP
Singular Top Score Bounding Box
SAM2
Segmented Mask

"Campbell's red and white tomato soup can"

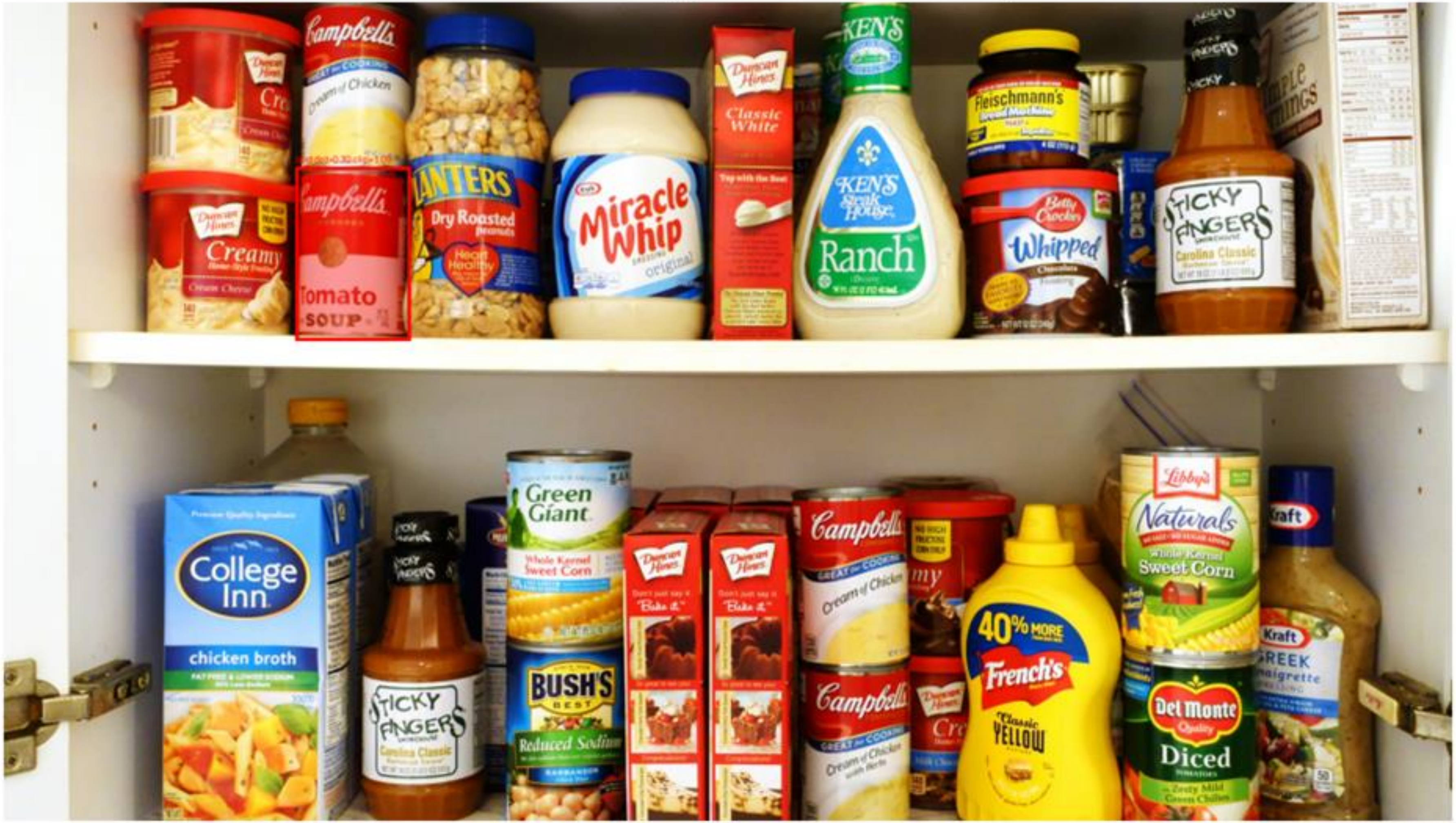
All OWLv2 Detections (pre-CLIP)



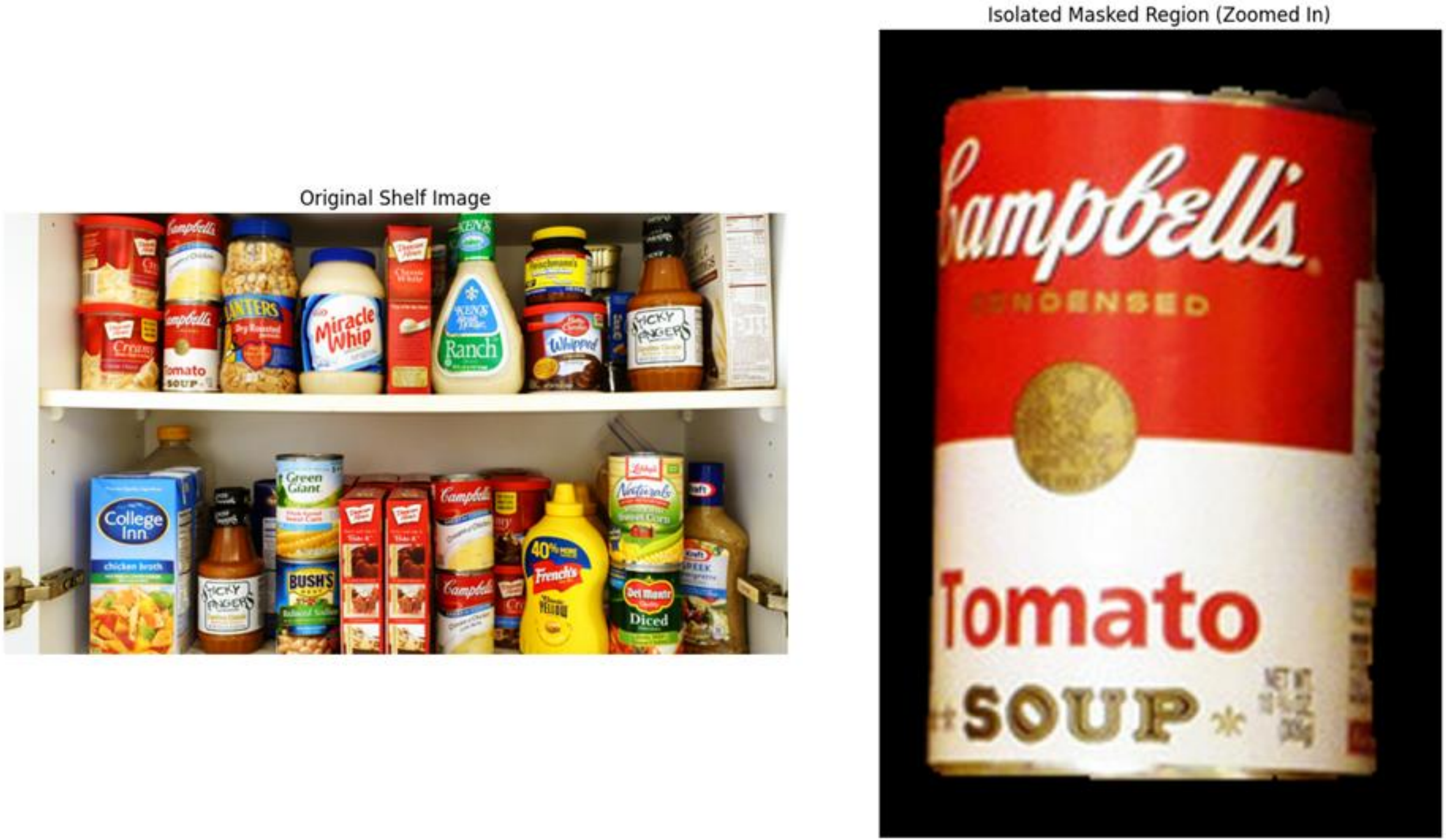
OWLv2 + CLIP Re-ranked Detection



Final Mask Overlay (OWLv2 + CLIP + SAM2)



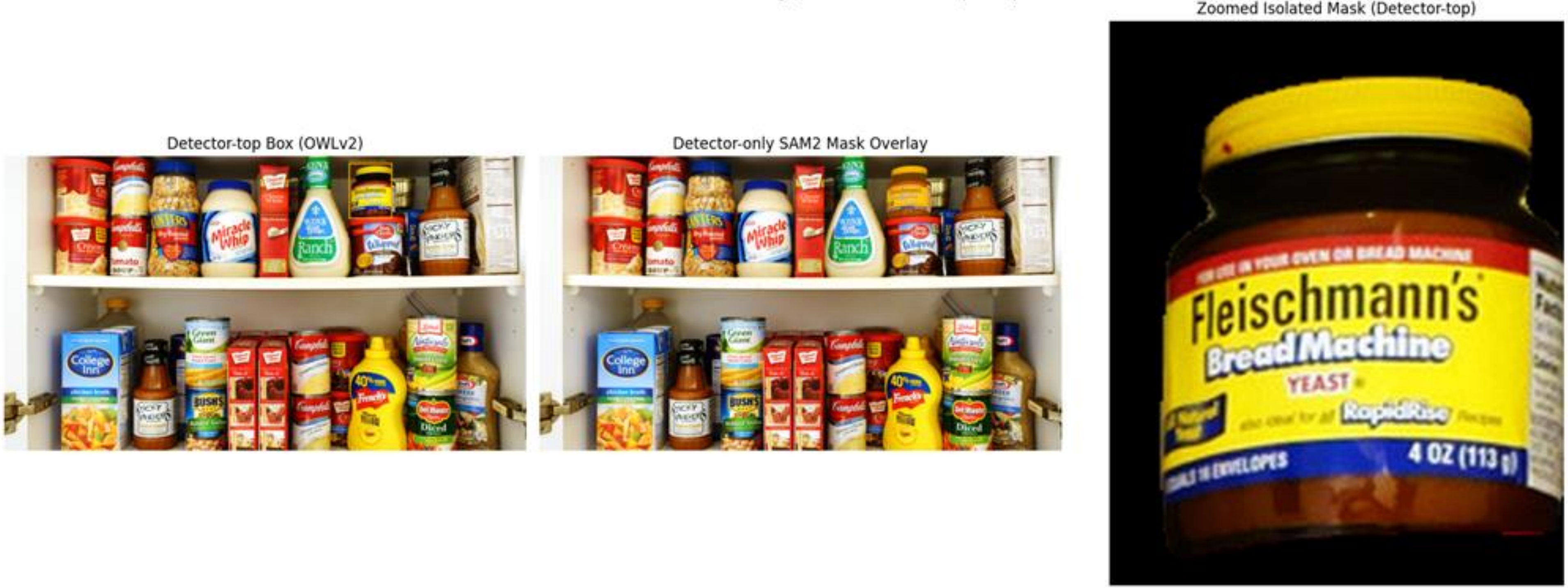
OWLv2 + CLIP + SAM2 — Detected Product Isolation



Original Shelf Image

Isolated Masked Region (Zoomed In)

Baseline: Detector-only (OWLv2 → SAM2) Output



Detector-top Box (OWLv2)

Detector-only SAM2 Mask Overlay

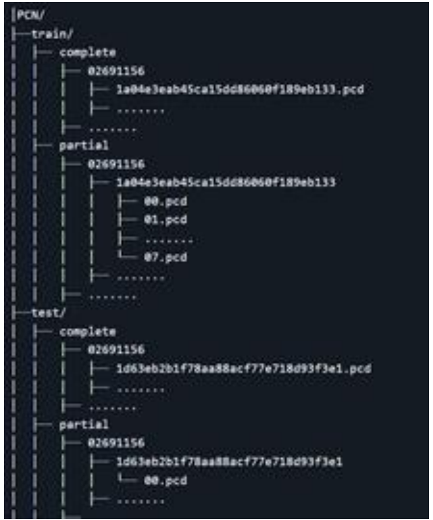
Zoomed Isolated Mask (Detector-top)

Synthetic dataset generation for fine tuning PC completion model

Mesh model from the open resources:
The YCB Object and Model Set: Towards Common Benchmarks for Manipulation Research

Convert mesh model into point cloud

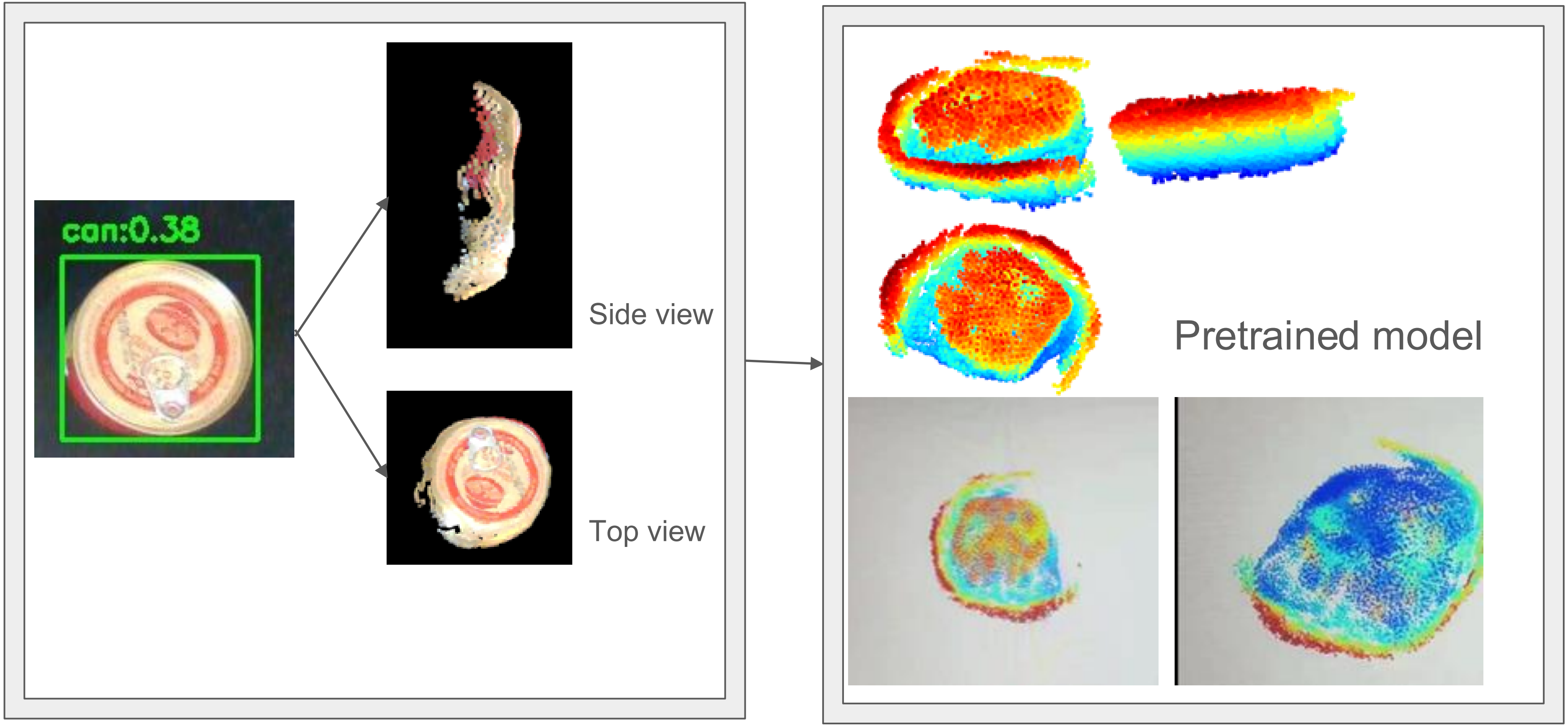
Random sampling complete point cloud from mu
cam-views(N) to generate (1 complete PC, N
partial PC) data pairs.



```
|PCN/
├──train/
|   ├── complete
|   |   ├── 02691156
|   |   |   ├── 1a04e3eab45ca15dd86060f189eb133.pcd
|   |   |   ├── .......
|   |   ├── .......
|   ├── partial
|   |   ├── 02691156
|   |   |   ├── 1a04e3eab45ca15dd86060f189eb133
|   |   |   |   ├── 00.pcd
|   |   |   |   ├── 01.pcd
|   |   |   |   ├── .......
|   |   |   |   └── 07.pcd
|   |   |   ├── .......
|   |   ├── .......
├──test/
|   ├── complete
|   |   ├── 02691156
|   |   |   ├── 1d63eb2b1f78aa88acf77e718d93f3e1.pcd
|   |   |   ├── .......
|   |   ├── .......
|   ├── partial
|   |   ├── 02691156
|   |   |   ├── 1d63eb2b1f78aa88acf77e718d93f3e1
|   |   |   |   └── 00.pcd
|   |   |   ├── .......
```

Re-arrange the PC data pairs into PCN dataset format, used for FT poinTr mo

The performance of PoinTr(point cloud completion model)



Side view

Top view

Pretrained model

PC_completion model(PoinTr) Fine-tuned results

Completed PCD files:
https://drive.google.com/drive/folders/1qgKzcFvokhn9vDwrDxb5LIuXDfEUvHOj?usp=sharing

Online PCD visualizer:
online pcd viewer

9.29.2025

0. Galaxea R1 Lite

- Dual 6-DOF Arms with Grippers
  - 600mm reach (~2ft)
  - Typical load 3kg, Maximum load 5kg.
- 3DOF Body
  - Vertical elevation of 1.7m
- Omnidirectional Chassis
- Intel Core i9-12900HK 32GB RAM 1TB SSD
- Binocular camera x1 (Head)
- Depth Camera x2 (Each side wrist)

0. Galaxea R1 Lite

Keyboard teleop demo:

https://drive.google.com/file/d/1HRkEWN0kjXuUHJKT46ypccUfxe_bq-nG/view?usp=sharing

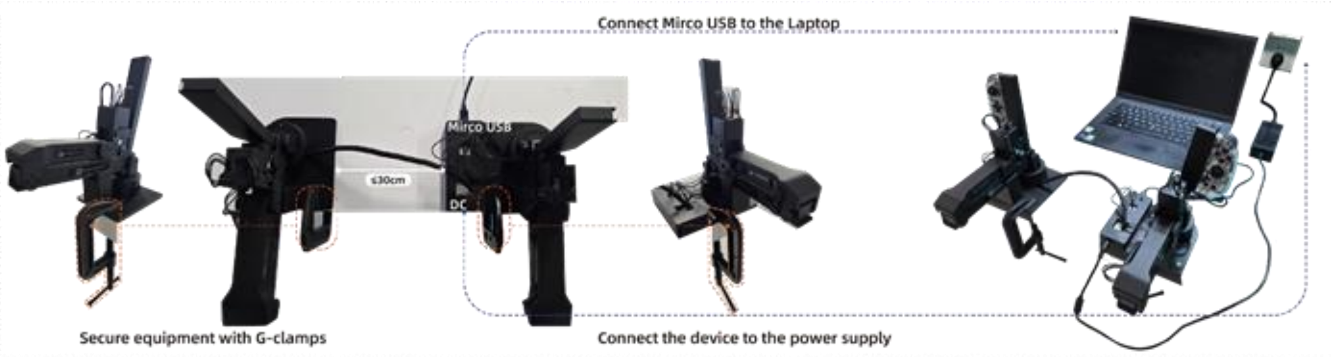
Keyboard teleop real demo:

https://drive.google.com/file/d/1W4uI6dXL5DDIPdimmUsWVhI75n7Ssbsm/view?usp=sharing

Skill graph demo:

https://drive.google.com/file/d/11TkfmO8zgZW7tU07rEvQxNKdpF1jgDS9/view?usp=sharing

0. Galaxea R1 Lite – Isomorphic Teleoperation System

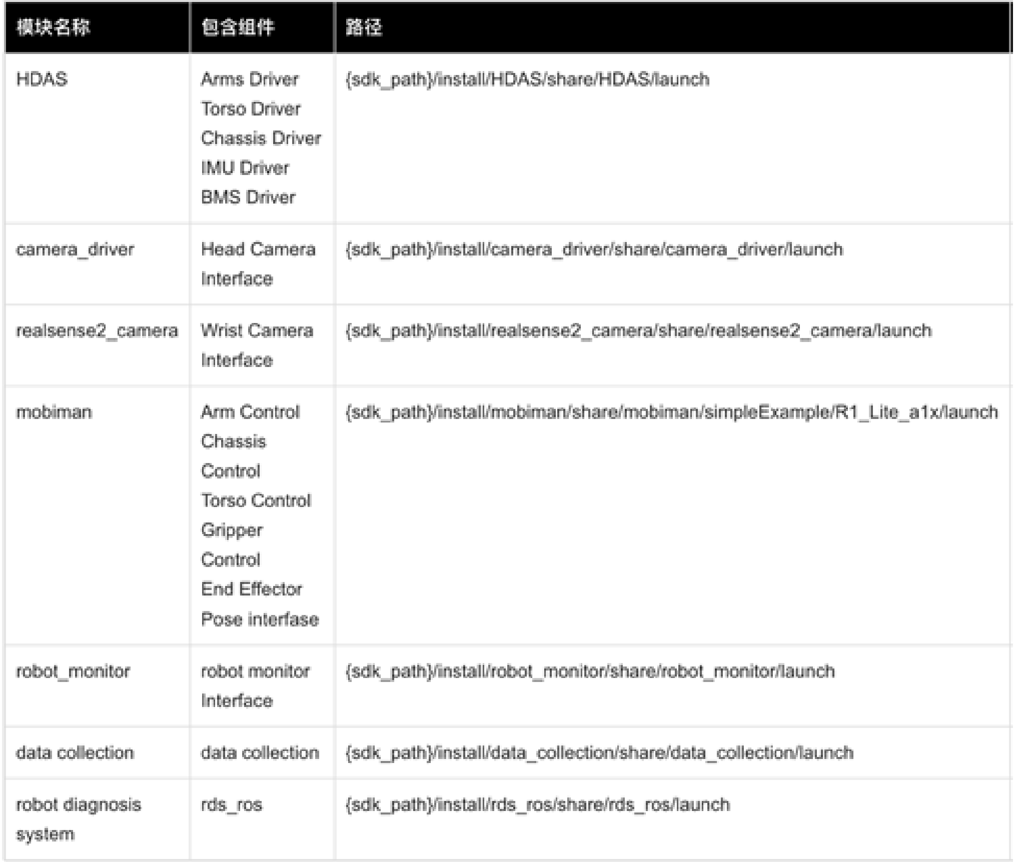


Connect Mirco USB to the Laptop

Secure equipment with G-clamps

Connect the device to the power supply

0. Galaxea R1 Lite – ROS2 Control Framework



模块名称	包含组件	路径
HDAS	Arms Driver Torso Driver Chassis Driver IMU Driver BMS Driver	{sdk_path}/install/HDAS/share/HDAS/launch
camera_driver	Head Camera Interface	{sdk_path}/install/camera_driver/share/camera_driver/launch
realsense2_camera	Wrist Camera Interface	{sdk_path}/install/realsense2_camera/share/realsense2_camera/launch
mobiman	Arm Control Chassis Control Torso Control Gripper Control End Effector Pose interfase	{sdk_path}/install/mobiman/share/mobiman/simpleExample/R1_Lite_a1x/launch
robot_monitor	robot monitor Interface	{sdk_path}/install/robot_monitor/share/robot_monitor/launch
data collection	data collection	{sdk_path}/install/data_collection/share/data_collection/launch
robot diagnosis system	rds_ros	{sdk_path}/install/rds_ros/share/rds_ros/launch

1. Project scope – Whole Picture

- Semantic understanding based path finding
  - The robot will be able to navigate to the corresponding location described by natural language
- Dual Arm collaborative pick and place
  - Target item will be identified and localized by VLM
  - Trajectories and gripping points will be generated by the dual arm task planner
  - Dual arm collaboration schedule will also be generated

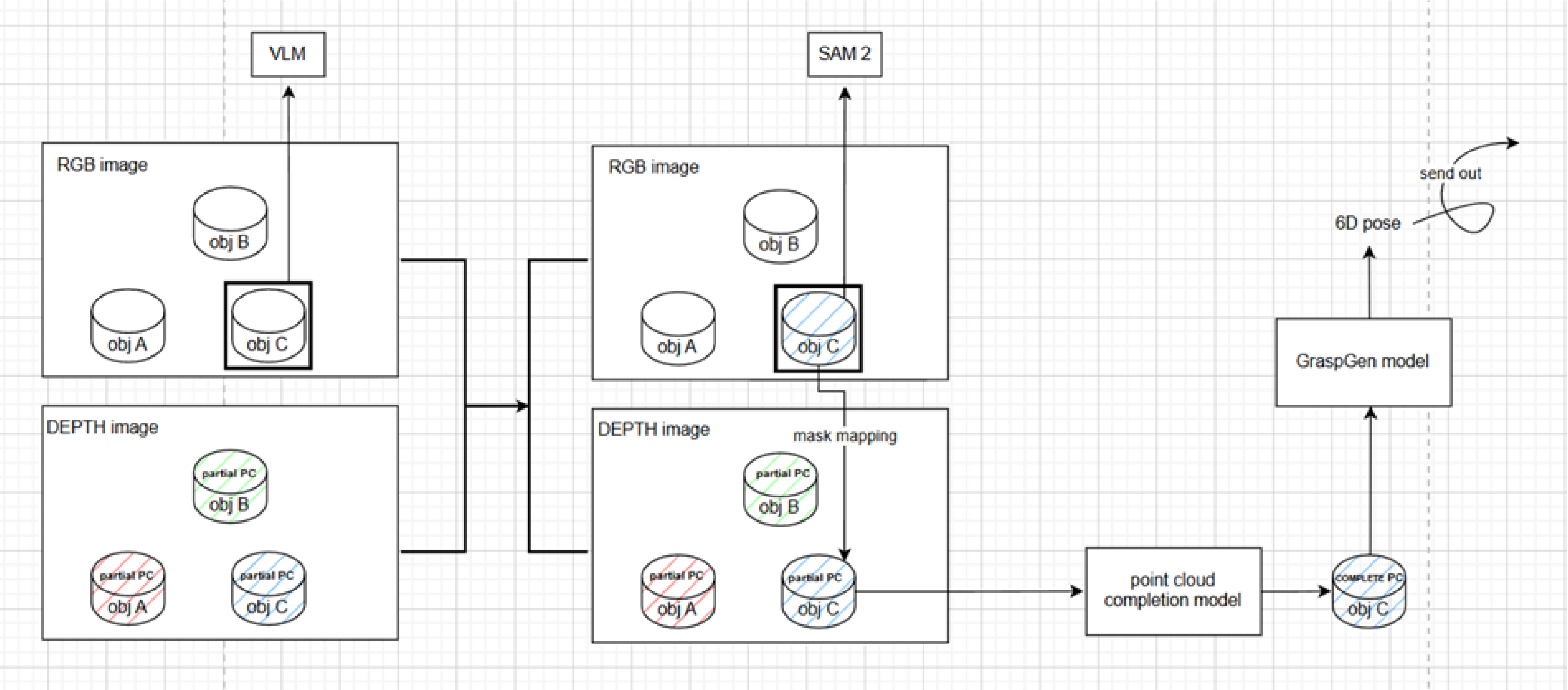
1. Project scope – Whole Picture

- Semantic understanding based path finding
  - The robot will be able to navigate to the corresponding location described by natural language
- **Dual Arm collaborative pick and place**
  - **Target item will be identified and localized by VLM**
  - **Trajectories and gripping points will be generated by the dual arm task planner**
  - **Dual arm collaboration schedule will also be generated**

2. Environment Setup

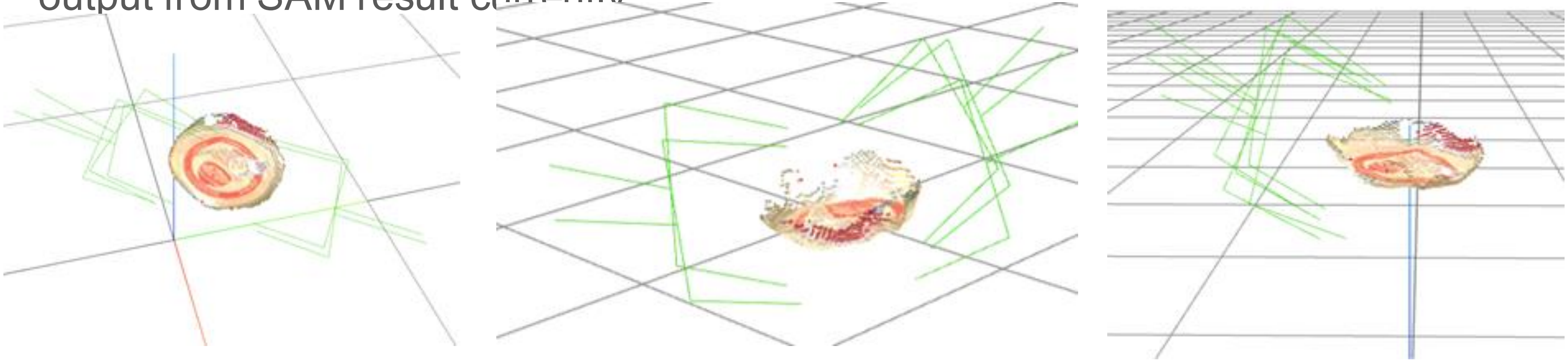
- Fixed base dual arm robot
- Shelf with fixed relative translation
- Experiment will be focused on limited variations of objects
  - Cereal Box (Biscuit Box, Pasta box)
  - Soup Can
  - Candy bag
- We also will take into consideration the collaboration of dual arms
- We will also stream the FPV video to a monitoring device, if the auto grasping fails, the human tele-operator will step in.

3. Perception



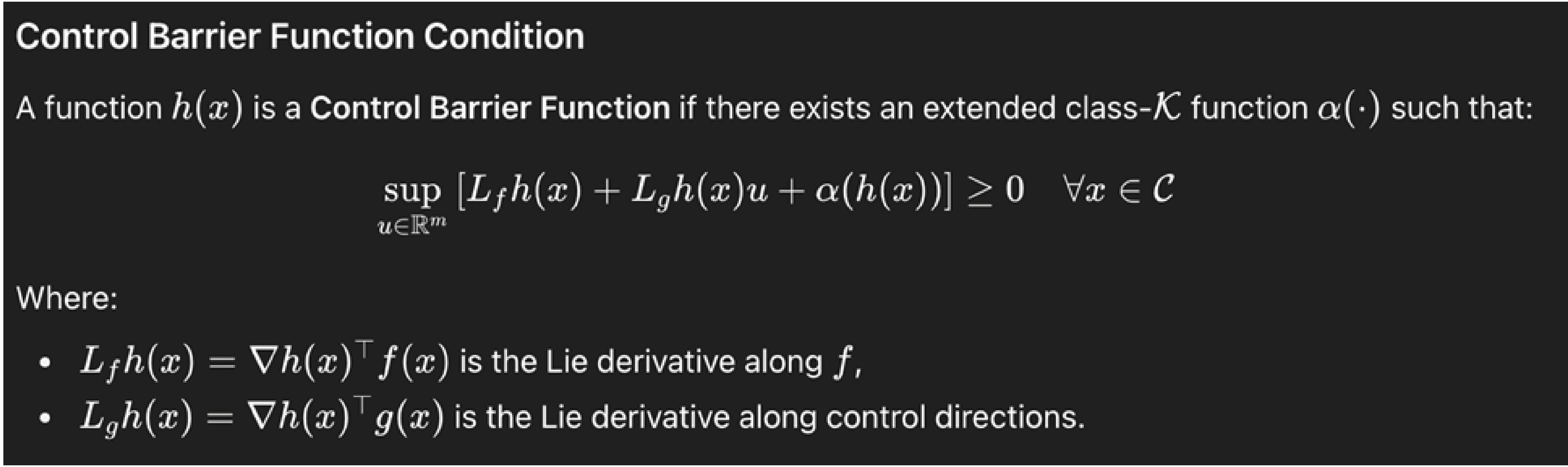
4. Manipulation – End-Goal generation

Based on the segmented mask, we can generate the gripper pose (SE4 set) directly. Since the format of the pointcloud after pc completion model is not aligned with the input requirement of the GraspGen model, we only used the output from SAM result currently



4. Manipulation – Safe Trajectory Selection

Control Barrier Function



Control Barrier Function Condition

A function $h(x)$ is a **Control Barrier Function** if there exists an extended class-$\mathcal{K}$ function $\alpha(\cdot)$ such that:

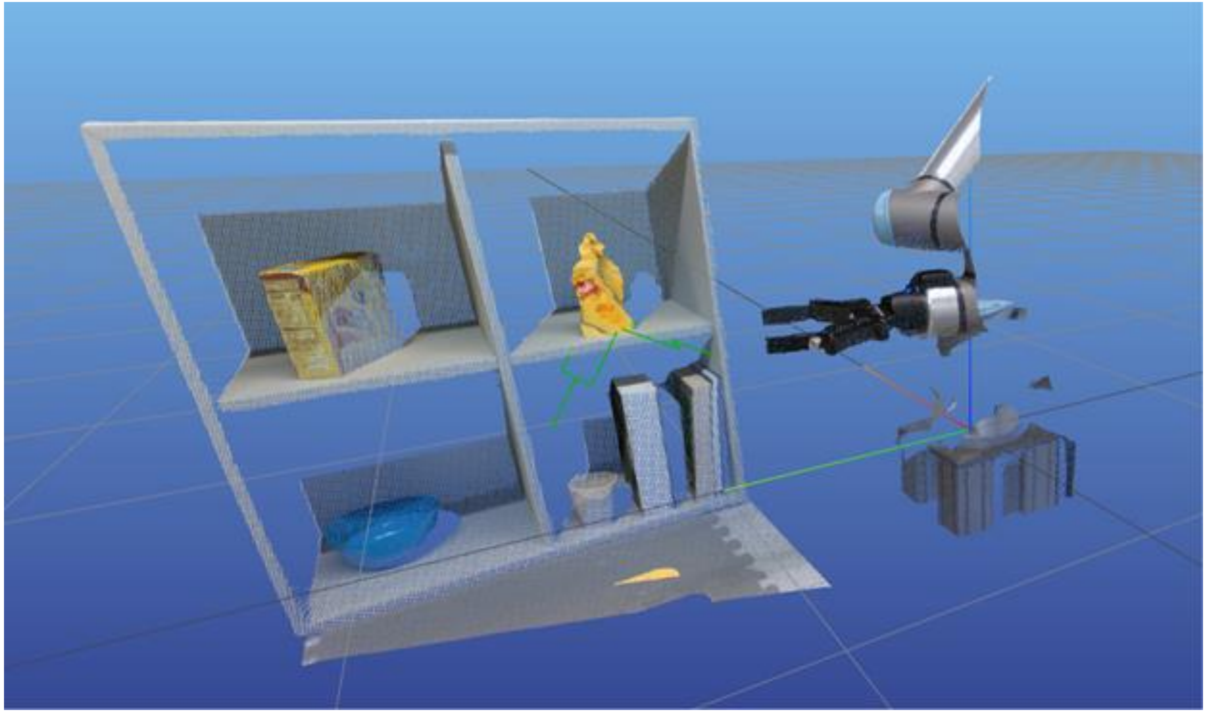
$$\sup_{u \in \mathbb{R}^m} \left[L_f h(x) + L_g h(x) u + \alpha(h(x)) \right] \geq 0 \quad \forall x \in \mathcal{C}$$

Where:

- $L_f h(x) = \nabla h(x)^\top f(x)$ is the Lie derivative along f,
- $L_g h(x) = \nabla h(x)^\top g(x)$ is the Lie derivative along control directions.

4. Manipulation – Safe Trajectory Selection

Control Barrier Function – Obstacle Localization

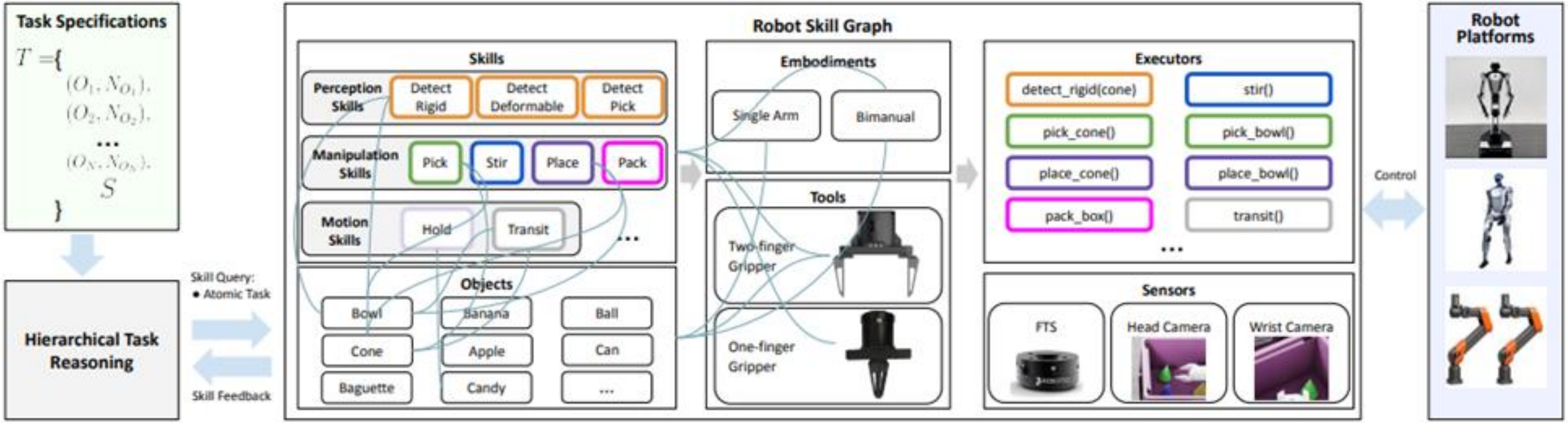


4. Manipulation – Safe Trajectory Selection

https://drive.google.com/file/d/17BLArBKfSiBW9LJ05nhuEEca3uQim3fs/view?usp=sharing

4. Manipulation – Trajectory Execution

HTR and Skill Graph



5. Plan forward

- Will be implementing the middleware of the robot and actuate all components of the new robot in current pipeline
- Upon completion, we will first try a small demo of single arm pick and place

9.22.2025

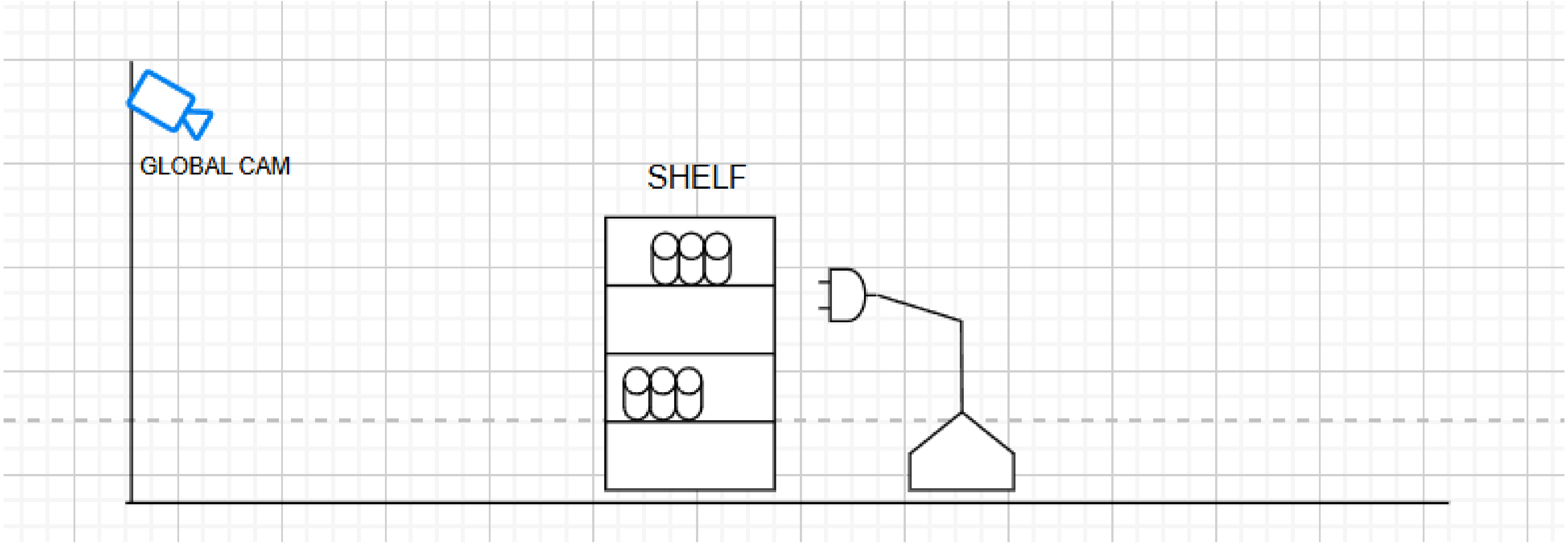
Required softwares

1. Simulation Platform:
    - Gazebo, Mujoco
1. Communication middleware:
    - ROS2
1. Teleoperation set:
    - Apple Vision Pro

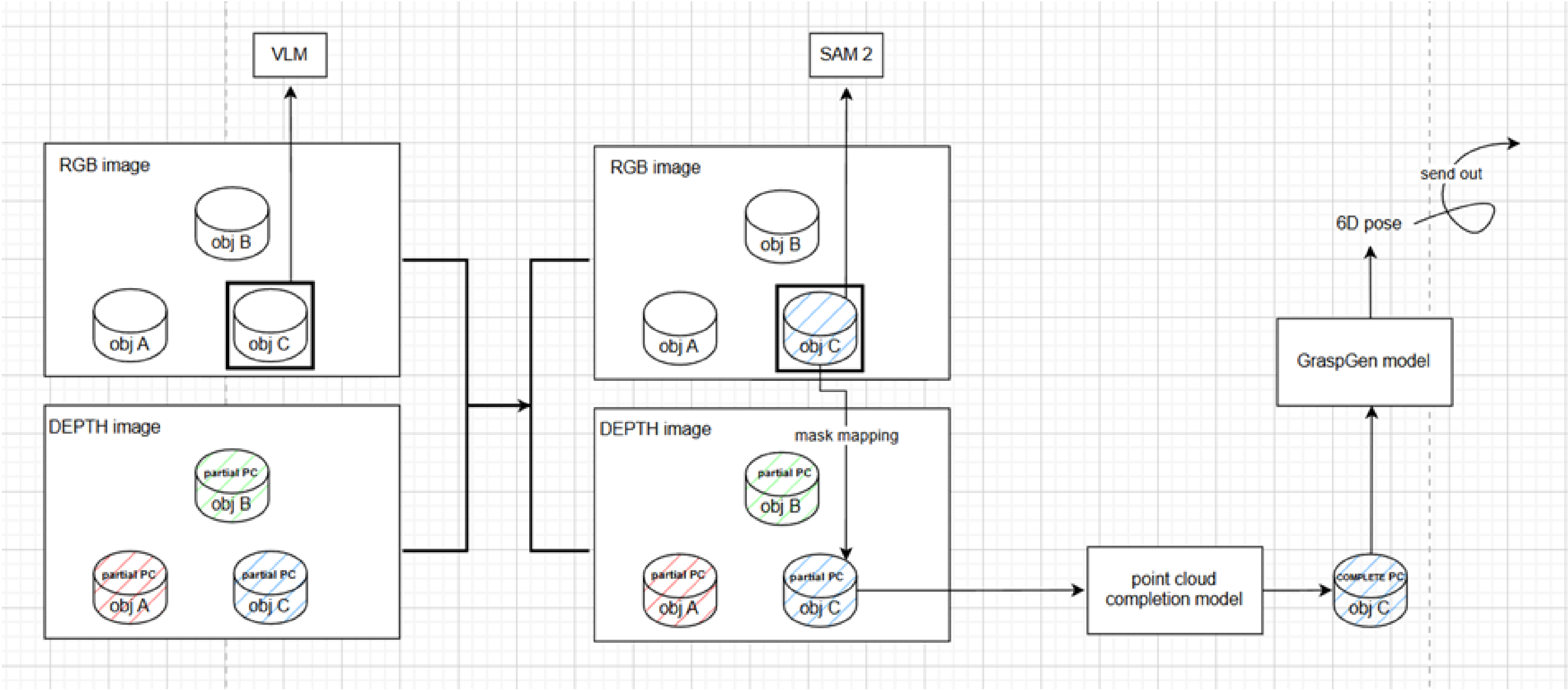
Grasping pose generation + 6D pose estimation pipeline

- Intended Object segmentation based on human language:
    - VLM: GroupViT or DINO
    - Seg model: SAM2
- Point Cloud Completion from Partial Observation:
    - point completion model(PointTr)
- Grasping pose generation + 6D pose estimation:
    - GraspGen model for grasping pose generation (May need some finetune for unseen objects)
    - ICP(point-cloud registration) in WBCD competition help for 6D pose estimation.

A general view of experiment setting



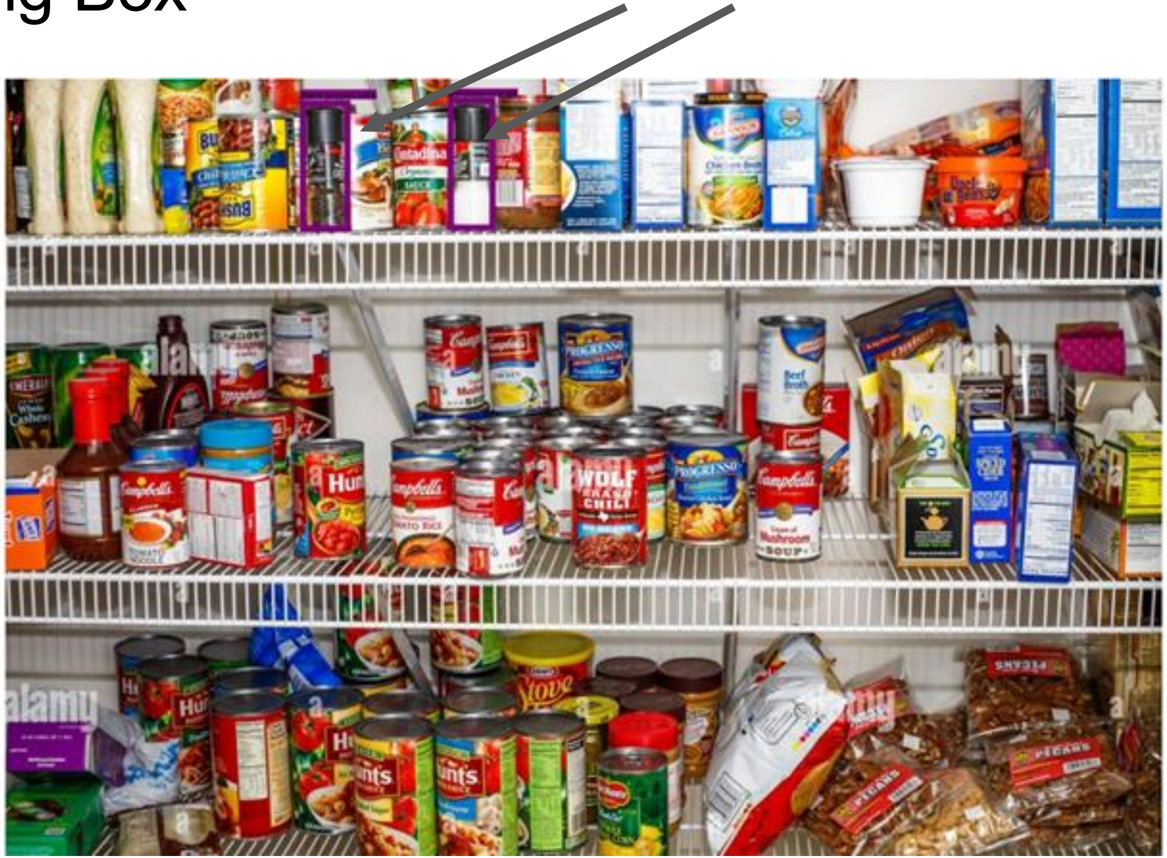
A general view of vision pipeline



1. Get 2D Bounding Box

Salt and Pepper Shakers

User Input

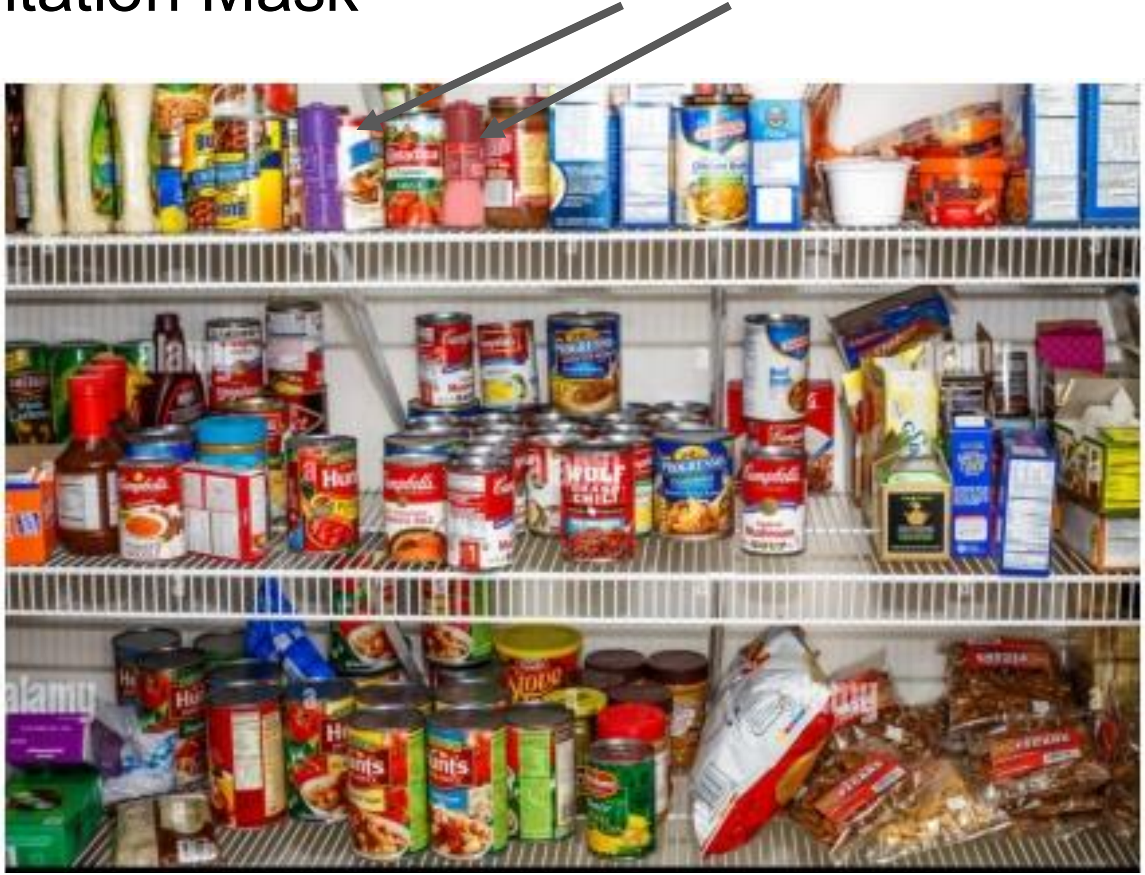


2. Get 2D Segmentation Mask

❝

Salt and Pepper Shakers

User Input

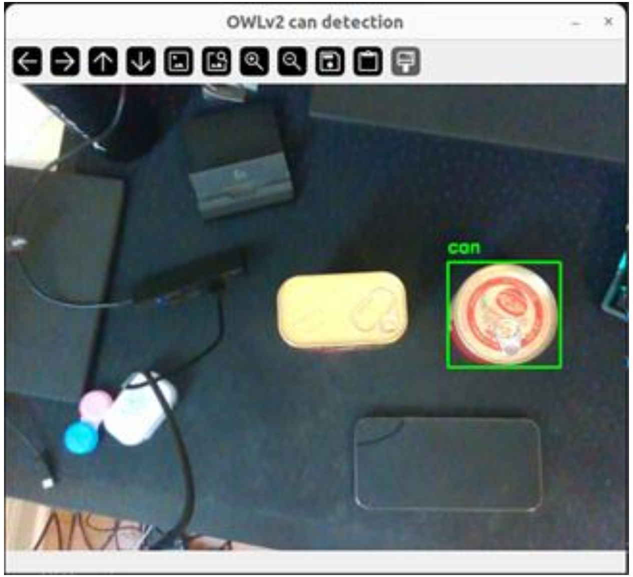


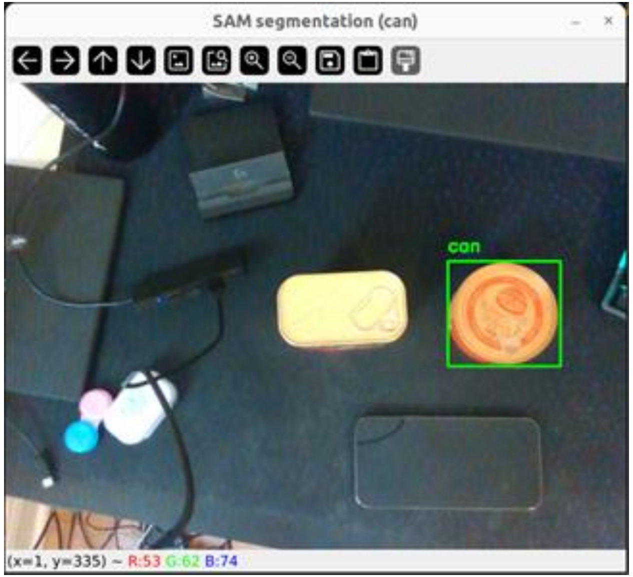
3. Use Mask to filter the Realsense Pointcloud (3D)

4. Run PointTr model for Pointcloud completion
5. Run graspGen to calculate 6D Pose Position

VLM Pipeline

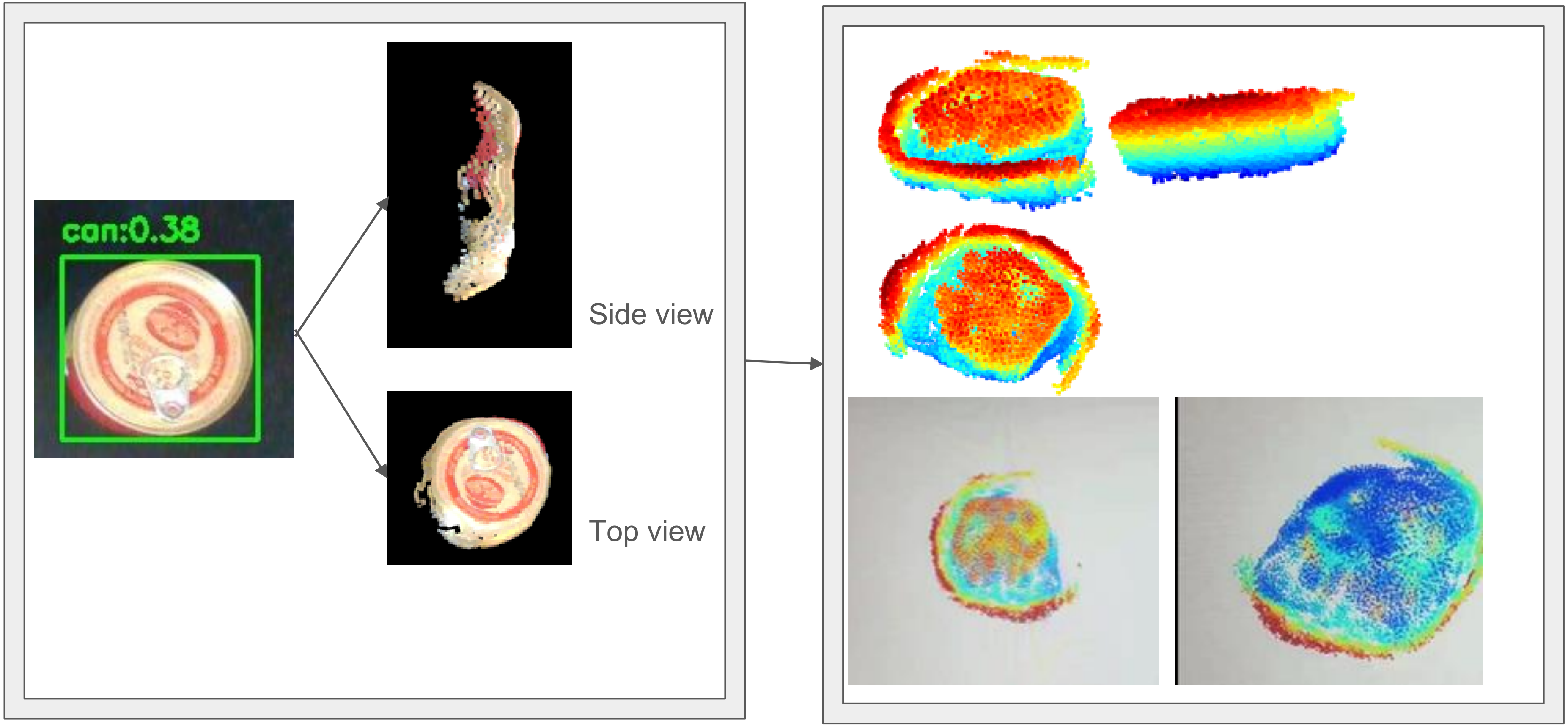
Combined with a VLM model and SAM model, we can get the segmented can mask from a given image input





(x=1, y=335) ~ R:53 G:62 B:74

The performance of PoinTr(point cloud completion model)



The paper’s evaluation metric for both seen and unseen cases

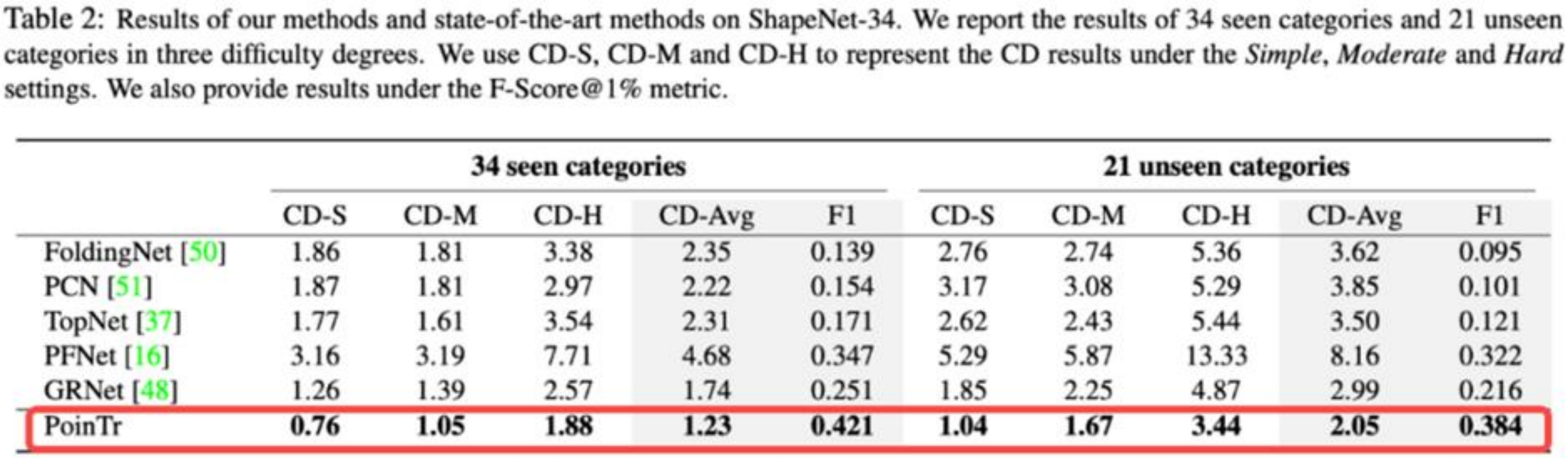


Table 2: Results of our methods and state-of-the-art methods on ShapeNet-34. We report the results of 34 seen categories and 21 unseen categories in three difficulty degrees. We use CD-S, CD-M and CD-H to represent the CD results under the *Simple*, *Moderate* and *Hard* settings. We also provide results under the F-Score@1% metric.

	34 seen categories					21 unseen categories				
	CD-S	CD-M	CD-H	CD-Avg	F1	CD-S	CD-M	CD-H	CD-Avg	F1
FoldingNet [50]	1.86	1.81	3.38	2.35	0.139	2.76	2.74	5.36	3.62	0.095
PCN [51]	1.87	1.81	2.97	2.22	0.154	3.17	3.08	5.29	3.85	0.101
TopNet [37]	1.77	1.61	3.54	2.31	0.171	2.62	2.43	5.44	3.50	0.121
PFNet [16]	3.16	3.19	7.71	4.68	0.347	5.29	5.87	13.33	8.16	0.322
GRNet [48]	1.26	1.39	2.57	1.74	0.251	1.85	2.25	4.87	2.99	0.216
PoinTr	**0.76**	**1.05**	**1.88**	**1.23**	**0.421**	**1.04**	**1.67**	**3.44**	**2.05**	**0.384**

CD means chamfer distance, the smaller the better

A grocery dataset can potentially be used for training PC completion model

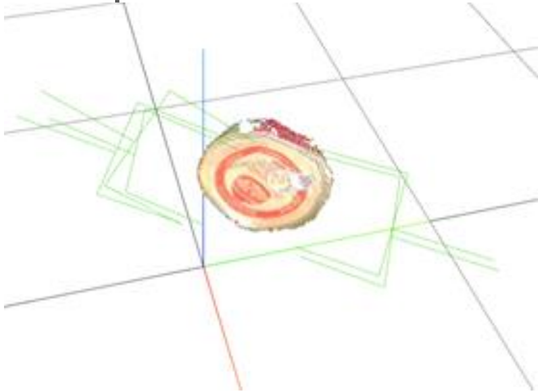
The 3DGrocery100 dataset is a pointcloud dataset with 3D points and colors meant to support and compare pointcloud classification algorithms to improve the state of the art.

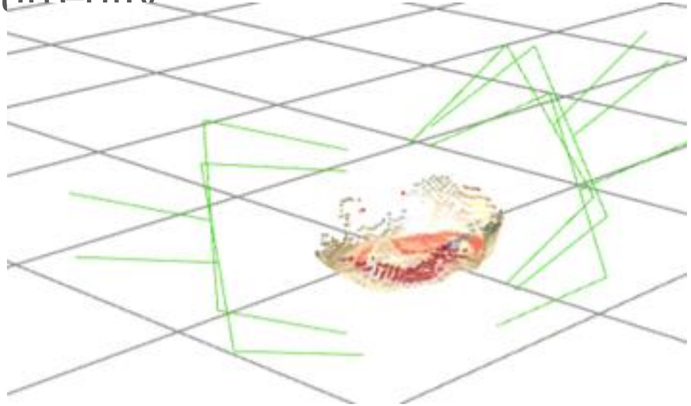
3Dgrocery100 dataset

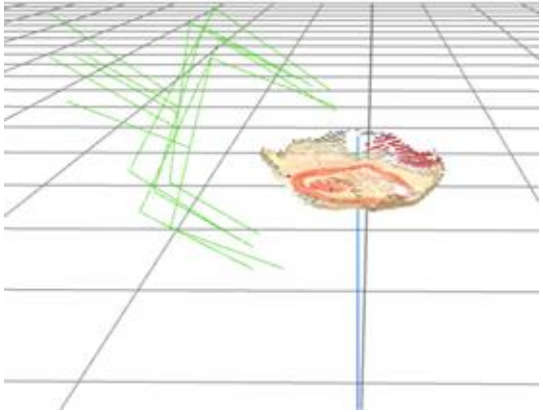
Point cloud data isn't complete, need multi images fusion techniques to collect complete grocery point cloud data to train Point cloud completion model.

The Performance of GraspGen

Based on the segmented mask, we can generate the gripper pose (SE4 set) directly. Since the format of the pointcloud after pc completion model is not aligned with the input requirement of the GraspGen model, we only used the output from SAM result currently







VLM Pipeline Demo

1. https://drive.google.com/file/d/14Cbqs2uV1GmdFiiMEWIyZINmpy_xSHaS/view?usp=drivesdk
2. https://drive.google.com/file/d/1fVgC0HAKgAGY6hCCHRbEG9Y9LISU0FP7/view?usp=drivesdk

R1lite

Real machine:

https://drive.google.com/file/d/11Wh_pQgFOO1oNAed2ib11gnmWZ77HTN9/view?usp=drive_link

https://drive.google.com/file/d/1g-T-tIV-8kcIomoXEz3_yDI346ANmPI1/view?usp=drive_link

Simulation:

https://drive.google.com/file/d/1VxBnEZR92pkkPFPpRkItpPn_WV5tXa6i/view?usp=drive_link

Teleop: Robot tracks hand’s operation

Enable: pinch your right index finger

Usage:

keeping the pinch right index finger:

- Control the robot to move by moving left hand (relative to robot base)
- Control the gripper’s rotation by rotating the left wrist (absolute)
- Control the gripper’s on/off by opening and closing of the hand

releasing pinch: adjust the position of hand for the starting position of the hand for the next teleop

Demo: https://drive.google.com/file/d/1OOLrau5uZt-vgEDRh9e_Wxf9z7mD2AAK/view?usp=share_link https://drive.google.com/drive/folders/1J_BP6FoNayC411fi5XLnelsUyMFDnblK?usp=drive_link

https://drive.google.com/drive/folders/1t72oMd9dy88RH0O50viPiOVc0gjA8fqU?usp=drive_link