Object Manipulation: Interactive Scene Exploration and Foundation Models

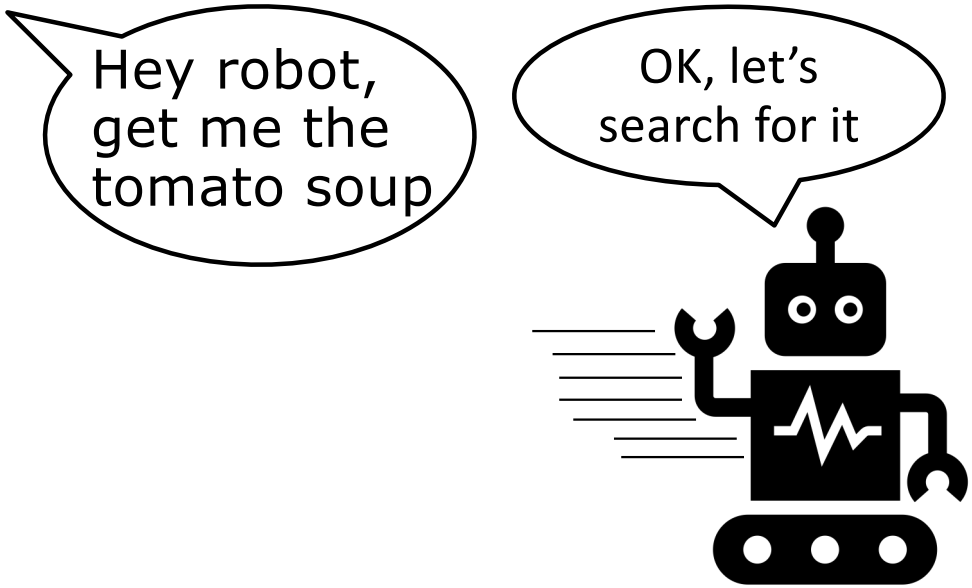
Maren Bennewitz, Rohit Menon, Nils Dengler

Humanoid Robots Lab, University of Bonn

Goal of This Chapter

- Learn about the concepts of **mechanical search** and **manipulation-enhanced mapping** in cluttered spaces
- Grasp the ideas of **foundation models for manipulation**

Motivation


Hey robot, get me the tomato soup
OK, let's search for it


?
FRISCHE VOLLMILCH
KIDNEY-BOHNEN
Kölln
Schoko
CHEEZ-IT
ORIGINAL
40% MORE
French's

Motivation

How to see behind the boxes?

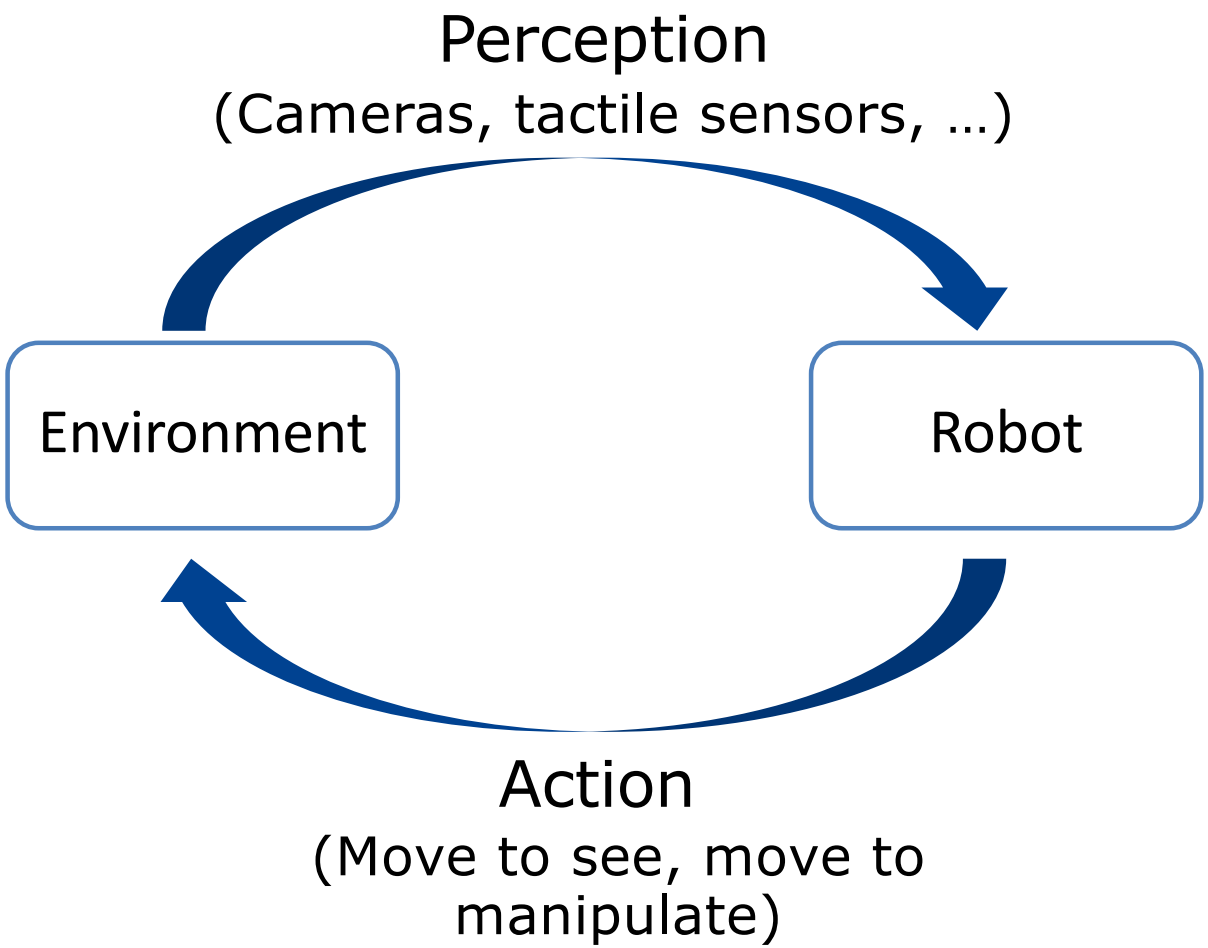
Only possible by re-arranging = **interactive perception**

Why Interactive Scene Exploration?

- **Active perception** reduces uncertainty but **insufficient** for completely covered spaces
- **Physical interactions** help to
  - Uncover hidden objects
  - Refine object pose estimates or physical properties
- However, those actions **may amplify uncertainty**
- Informed action sampling to select **high-impact manipulation actions**
- Uncertainty-guided planning to balance risk and discovery

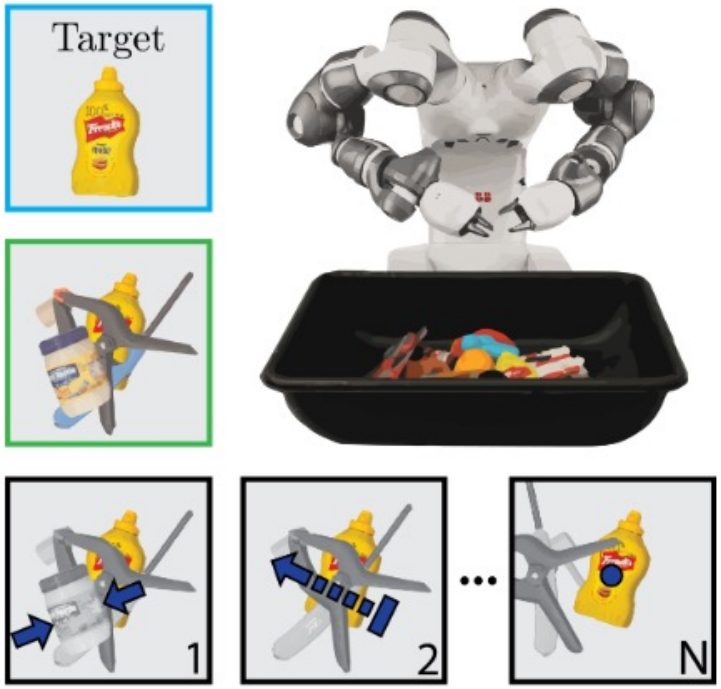
Interactive Perception

- Perception-action loop: sense, plan, act
- Robot **perceives**, **plans** next **action**, **executes** action, **perceives** again



Mechanical Search

- Robots are frequently required to interactively search for and retrieve objects from cluttered **bins**, **shelves**, or **tables**
- Mechanical search: **locate and extract** a given target object



[Danielczuk et al., ICRA 2019]

Key Ideas of Mechanical Search

- Instead of relying solely on vision, use actions like **pushing** or **grasping** objects to **reveal** or **access** the target object
- Typically, only **partial observability**: incomplete or uncertain knowledge of the scene
- Necessary **gather more information** through interaction with the objects

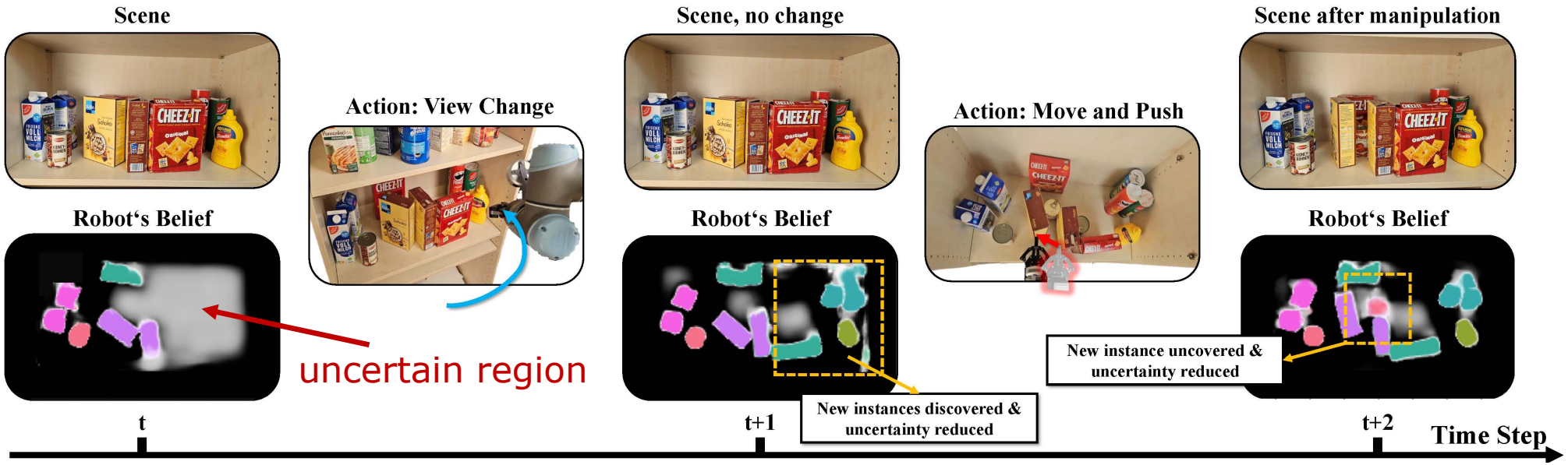
Bejjani et al., IROS 2021

Summary: Mechanical Search

- **Goal**: Retrieval of a specific object, even if occluded by or "buried" in clutter
- **Actions**: Push, pull, grasp, remove distractors
- **Benefits**: Enables task completion despite limited visibility, robust to clutter and occlusions
- **Drawbacks**: Focused on object-level tasks, **not** scene understanding

Manipulation-Enhanced Mapping

- Focus shifts from locating and extracting a specific object to **mapping all objects** in the scene
- Useful for object retrieval in the **long run**



J. Marcos Correia et al., "Map space belief prediction for manipulation-enhanced mapping, RSS 2025

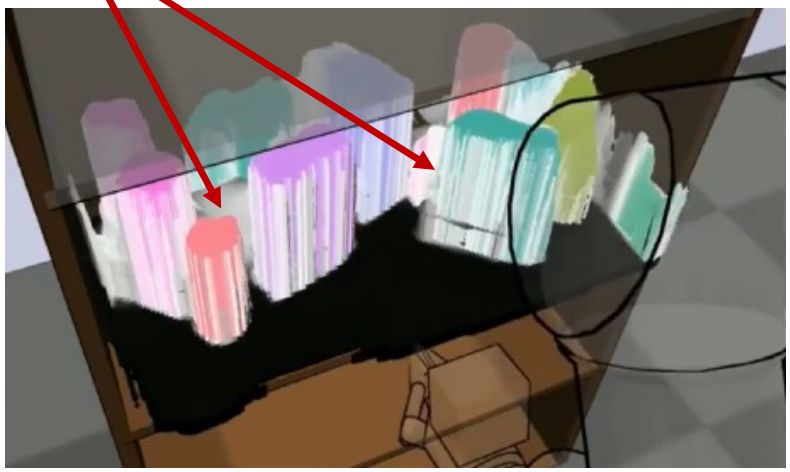
Environment Representation

- Environment represented as a **„belief"** of the scene
- Belief is realized as 3D voxel map, split in **occupancy** and **semantic** representation
- Modeling of the robot's **uncertainty** in the belief

Real-World

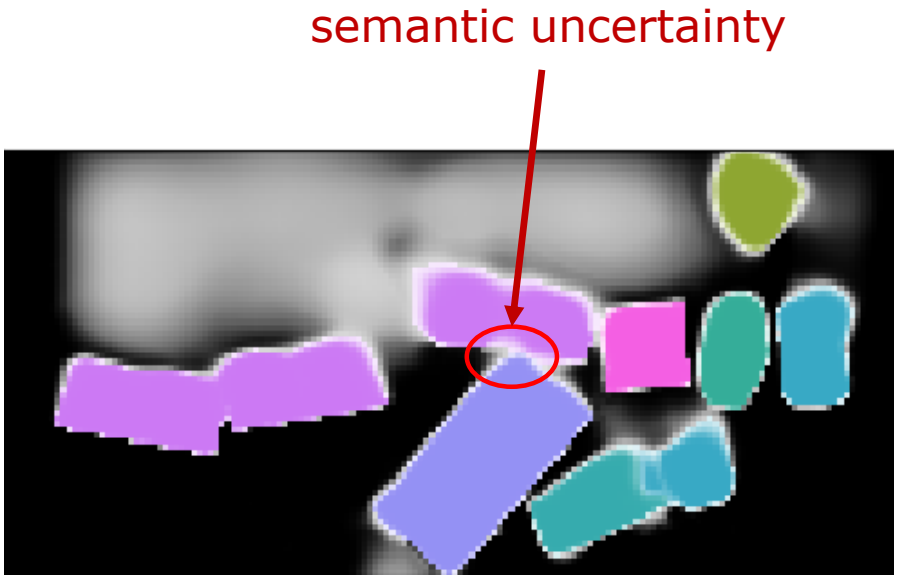
uncertain regions

Belief



Semantic Uncertainty

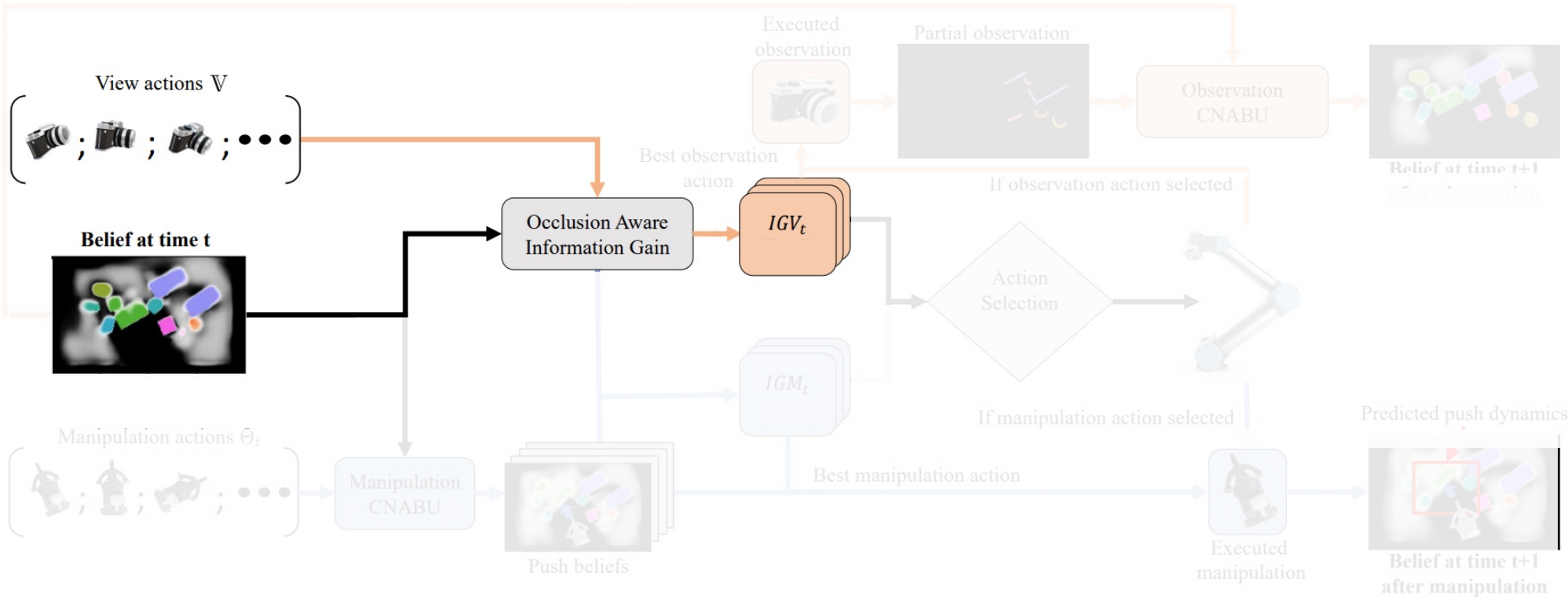
- Hard class labels can be overconfident and incorrect
- **Semantic uncertainty** quantifies **per-pixel label confidence**
- For example, p(box) = 0.9 vs. p(box) = 0.6
- Ambiguity arises from occlusion, noise, and similar textures
- **Uncertainty guides targeted sensing and action planning**



Effect of Actions

- **Active sensing reduces** occupancy and semantic uncertainty
- **Manipulation** actions **increase** uncertainty and disturbance risks, e.g., collisions or object drops
- Post-action sensing updates beliefs and reduces uncertainty
- Manipulation actions only chosen if **expected net information gain high**
- Action–perception loop **balances information gain** against **manipulation risk**

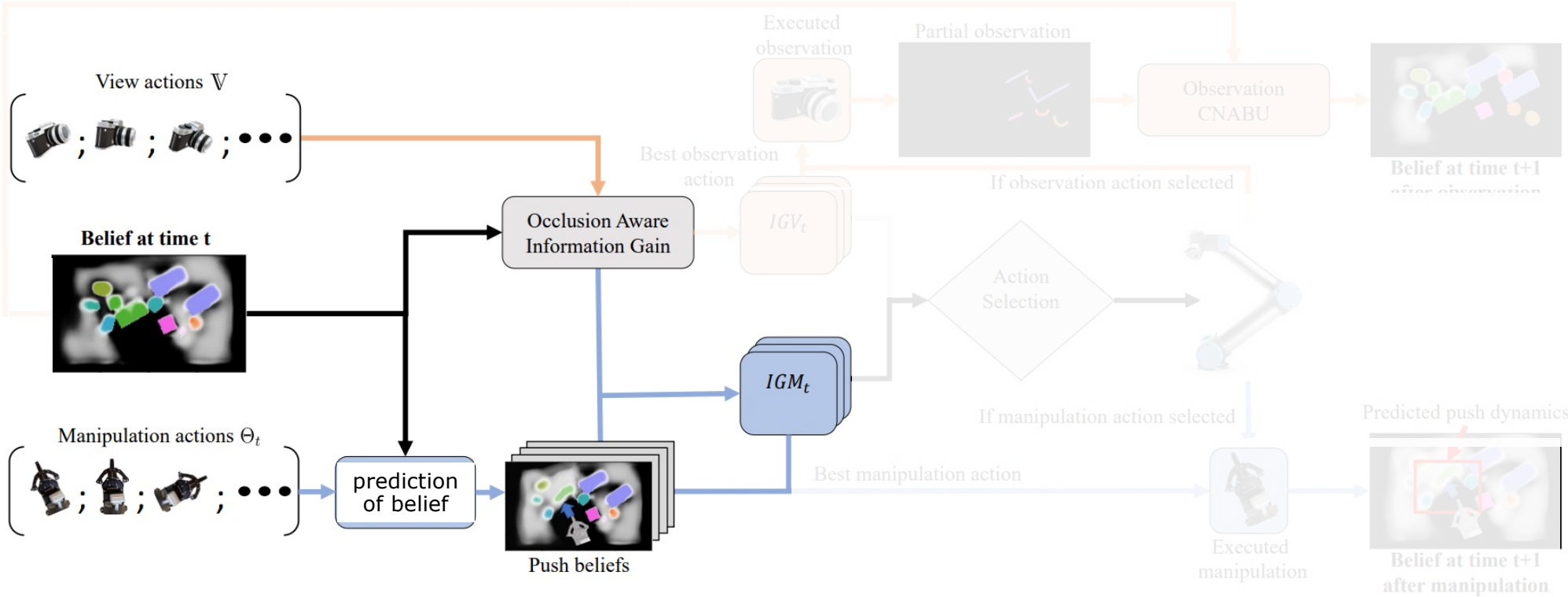
Action Selection



IGV = information gain for viewing action
IGM = information gain for manipulation action

J. Marcos Correia et al., "Map space belief prediction for manipulation-enhanced mapping, RSS 2025

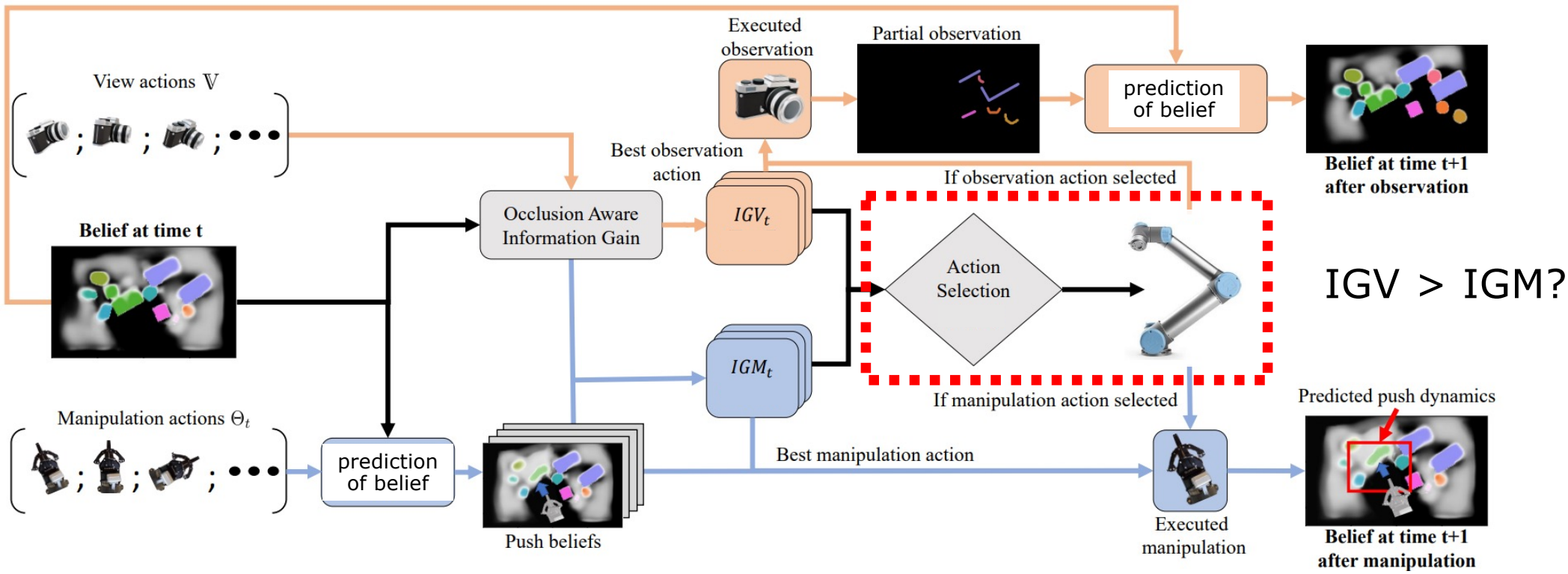
Action Selection



IGV = information gain for viewing action
IGM = information gain for manipulation action

J. Marcos Correia et al., "Map space belief prediction for manipulation-enhanced mapping, RSS 2025

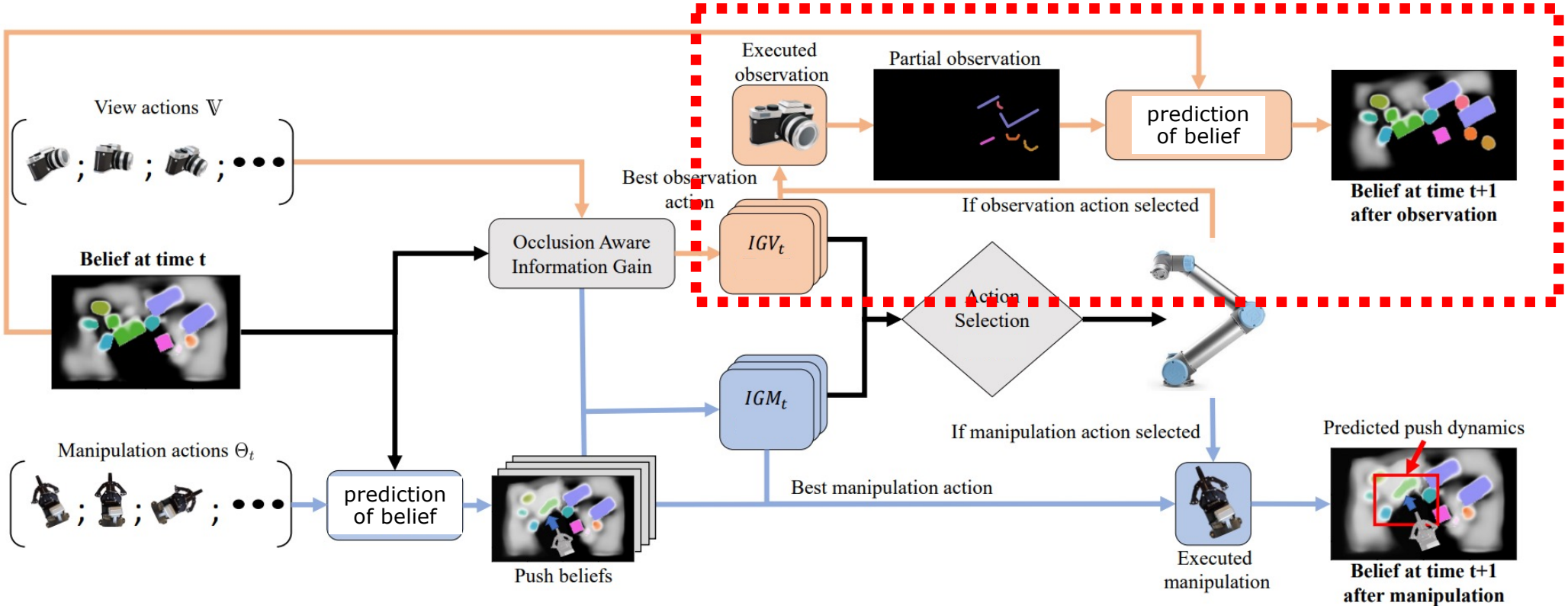
Action Selection



IGV = information gain for viewing action
IGM = information gain for manipulation action

J. Marcos Correia et al., "Map space belief prediction for manipulation-enhanced mapping, RSS 2025

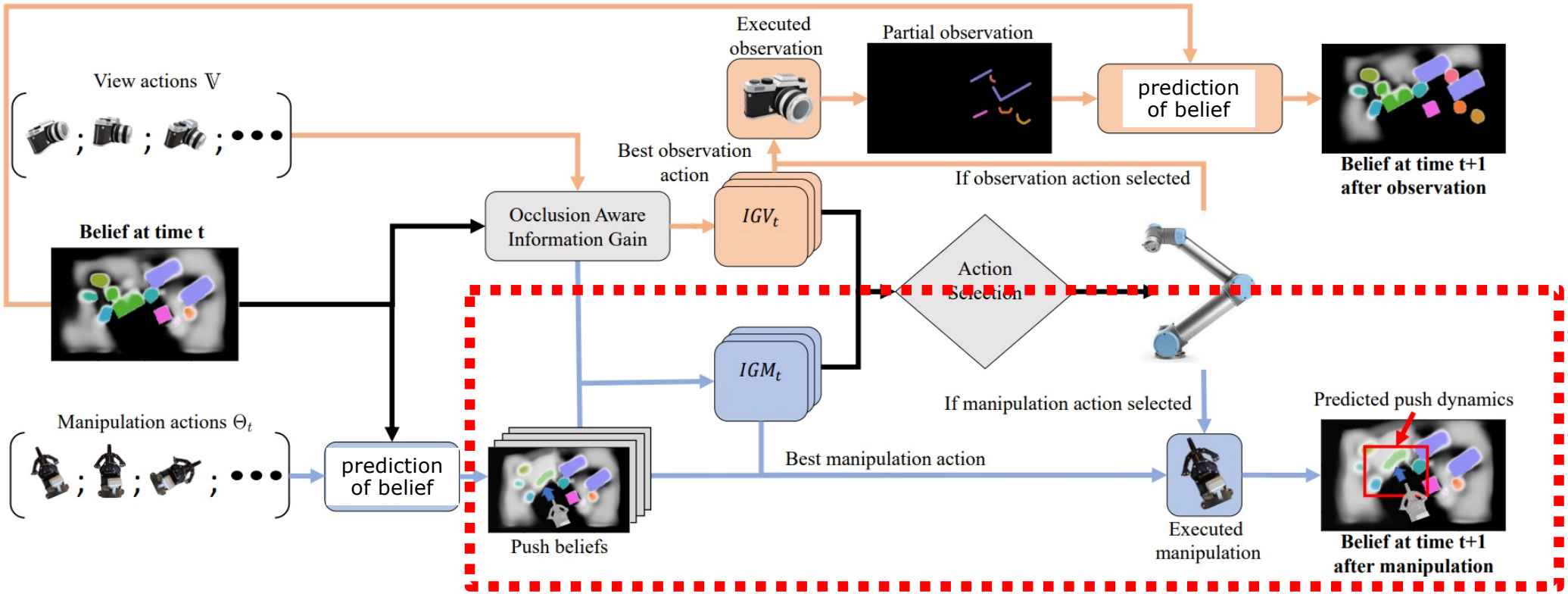
Action Selection



IGV = information gain for viewing action
IGM = information gain for manipulation action

J. Marcos Correia et al., "Map space belief prediction for manipulation-enhanced mapping, RSS 2025

Action Selection



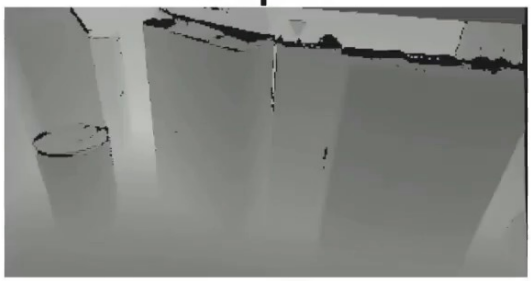
IGV = information gain for viewing action
IGM = information gain for manipulation action

J. Marcos Correia et al., "Map space belief prediction for manipulation-enhanced mapping, RSS 2025

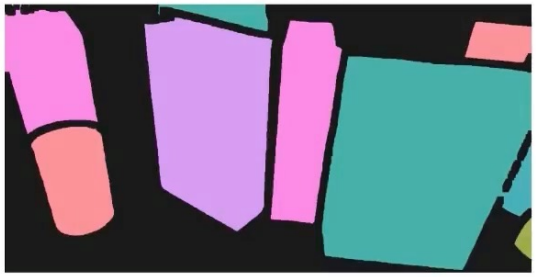
Real-World Experiment

RGB

Depth

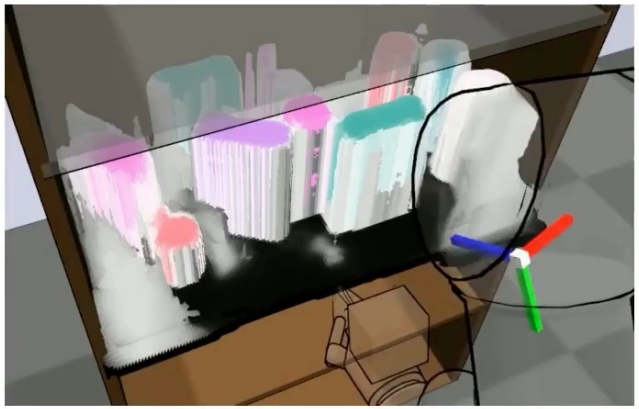


Semantic



External View

Environment Belief



Taken Action: Observation

Summary: Manipulation-Enhanced Mapping

- **Goal**: Increase map accuracy and reduce uncertainty through physical interaction
- **Actions**: Push objects, re-orient camera
- **Benefits:** Reduces uncertainty in semantic/occupancy maps, enables better planning in the long run
- **Drawbacks:** Less efficient for single direct object retrieval

Application in the Horticultural Domain

Foundation Models for Robot Manipulation

Foundation Models in General

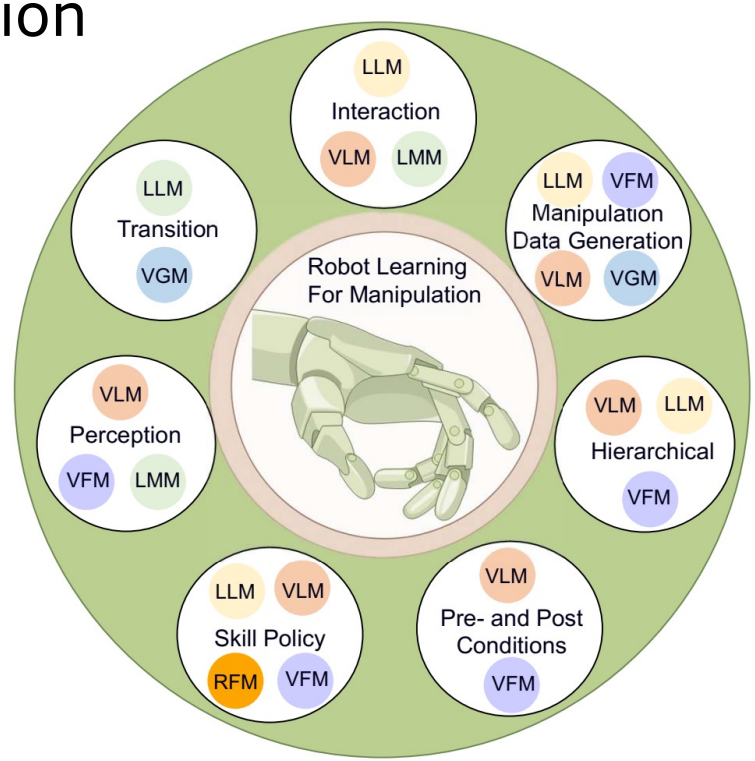
- Pretrained on massive internet-scale data
- Zero- or few-shot task generalization
- Bridge vision, language, and action
- Reduce manual labeling needs
- Enable modular, plug-and-play pipelines

Categories of Foundation Model

- **Large Language Models** (LLMs) for reasoning, e.g., ChatGPT
- **Visual Foundation Models** (VFMs) for segmentation, e.g., SAM
- **Visual Generative Models** (VGMs) for data synthesis, e.g., DALL-E, SORA
- **Vision-Language Models** (VLMs) for multimodal alignment, e.g., CLIP
- **Large Multimodal Models** (LMMs) across modalities, e.g., ULIP
- **Robot Foundation Models** (RFMs) for general policies, can use multimodal input (sensor data), e.g., RT-2
- **Vision-Language-Action Models** (VLA) common variant of RFM, e.g., π_0

Roles in Manipulation Pipeline

- **LLMs** facilitate human–robot communication and action planning
- **VFMs** support open-world object segmentation
- **VGMs** generate simulated manipulation scenarios
- **VLMs** align vision and language representations
- **LMMs** integrate multimodal sensory information
- **RFMs/VLAs** learn end-to-end manipulation policies

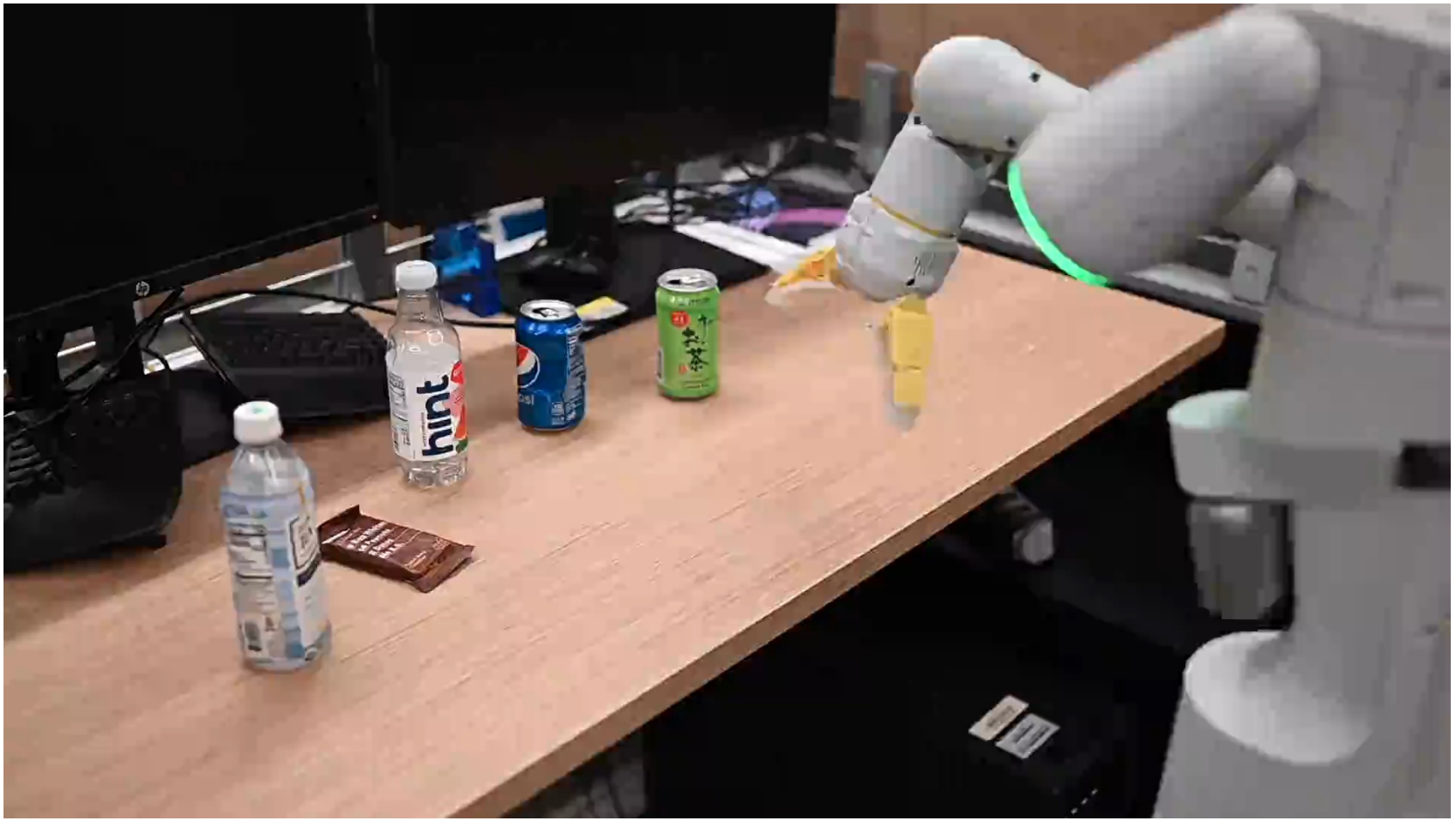


[Li et al., arXiv 2024]

Framework of Robot Learning for General Manipulation

- *Task Instruction: "Please make a cup of tea and hand it to me"*
- **Interaction:** interpret "make tea and hand me"
- **Pre-conditions:** detect kettle & cup affordances
- **Hierarchy:** boil → steep → pour → hand
- **State:** sense water temperature & cup pose
- **Policy:** compute delta pose commands
- **Post-conditions:** verify tea filled & delivered

RT-2: Vision-Language-Action Model



[Zitkovich et al., CoRL 2023]

Long-Term Vision

- Single **generalist policy** that outputs actions for a wide range of **different skills** for a wide range of **different robots**
- Need **only few additional data** samples for each robot and new application
- Instead of programming robots, just **user instructions**

Challenges & Future Directions

- Ensuring safety and execution stability
- Handling data heterogeneity across robots
- Scaling to high-dimensional control spaces
- Integrating real-time tactile feedback loops
- Improving model explainability & debuggability

Summary (1)

- **Mechanical search:** interacting with cluttered objects to retrieve occluded targets
- **Manipulation-enhanced mapping:**
  - Plans manipulation and observation actions to improve spatial and semantic understanding of the scene
  - Enables robots to **reduce uncertainty** and build more **complete and accurate** maps of cluttered spaces

Summary (2)

- **Foundation models** in robotics are large, pre-trained models that can generalize across multiple tasks, environments, and robots by leveraging vast amounts of data
- Enable **scalable, adaptable robotic systems** through transfer learning, multimodal reasoning, and few-shot or zero-shot task execution

Literature

- *Mechanical Search: Multi-step Retrieval of a Target Object Occluded by Clutter,* Danielczuk, Kurenkov, Balakrishna, Matl, Wang, Martín-Martín, Garg, Savarese, and Goldberg, IEEE/RAS Int. Conf. on Robotics and Automation (ICRA), 2019
- *Occlusion-Aware Search for Object Retrieval in Clutter,* Bejjani, Agboh, Dogar, and Leonetti, IEEE/ RSJ Int. Conf. on Int. Robots and Systems (IROS), 2021
- *Safe Leaf Manipulation for Accurate Shape and Pose Estimation of Occluded Fruits* Yao, Pan, Bennewitz, and Hauser, IEEE Int. Conf. on Robotics & Automation (ICRA), 2025
- *What Foundation Models Can Bring for Robot Learning in Manipulation,* Li, Jin, Sun, Yu, Shi, Hao, Hao, Liu, Sun, Zhang, and Fang, arXiv, 2024
- *RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control,* Zitkovich, Yu, Xu, Xu, Xiao, Xia, Wu, Wohlhart, Welker, Wahid , Vuong, Conf. on Robot Learning (CoRL), 2023