Humanoid Robotics

Manipulation 5: Visual Servoing, Interactive Scene Exploration, and Foundation Models for Manipulation

Maren Bennewitz

UNIVERSITÄT BONN

Goal of This Chapter

- Understand principles of **visual servoing**
- Learn about the concepts of **mechanical search** and **manipulation-enhanced mapping** for object search in cluttered spaces
- Understand opportunities and limitations of **foundation models for robotic manipulation**

Visual Servoing

Motivation

- Typical manipulation workflow:
  - Detect once, execute blindly
  - One-shot perception is fragile
- Calibration errors accumulate
- Objects may move after detection
- Contact changes object state
- Vision can correct online

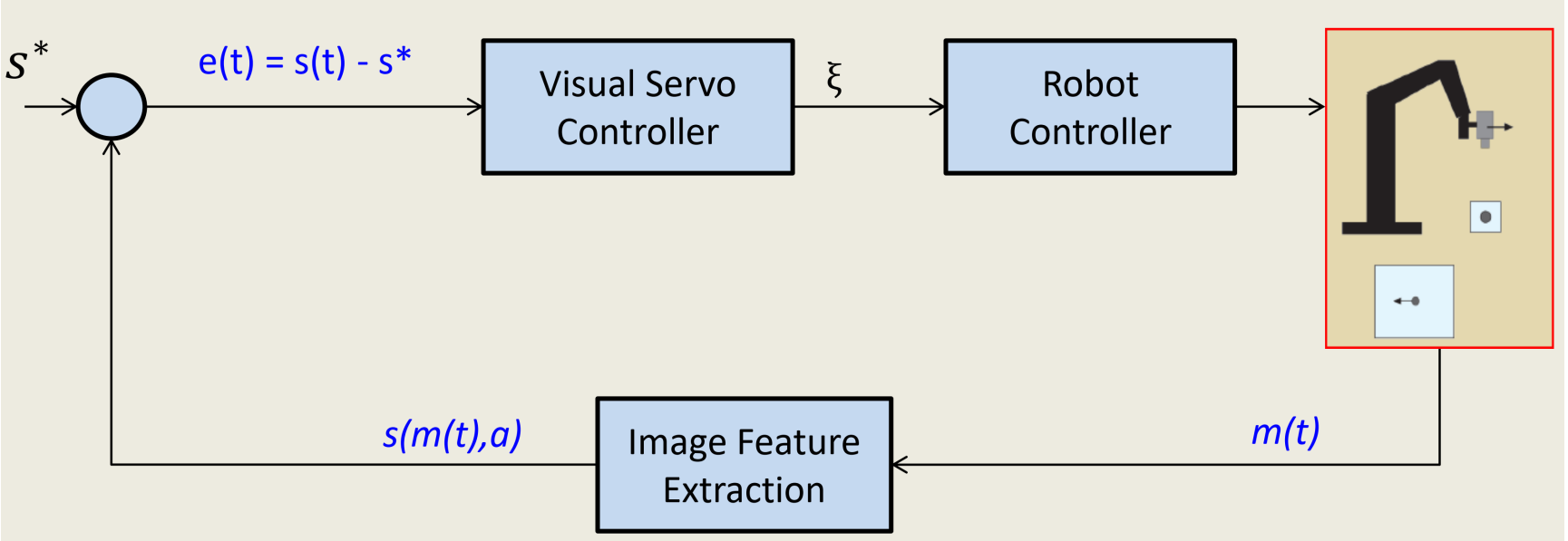
HumanoidRobots, KIT

Coupling Perception and Manipulation

- Visual servoing **couples seeing and acting**
- Action **changes** future **visual evidence**
- Perception **continuously corrects action**
- Manipulation becomes **visually adaptive control**

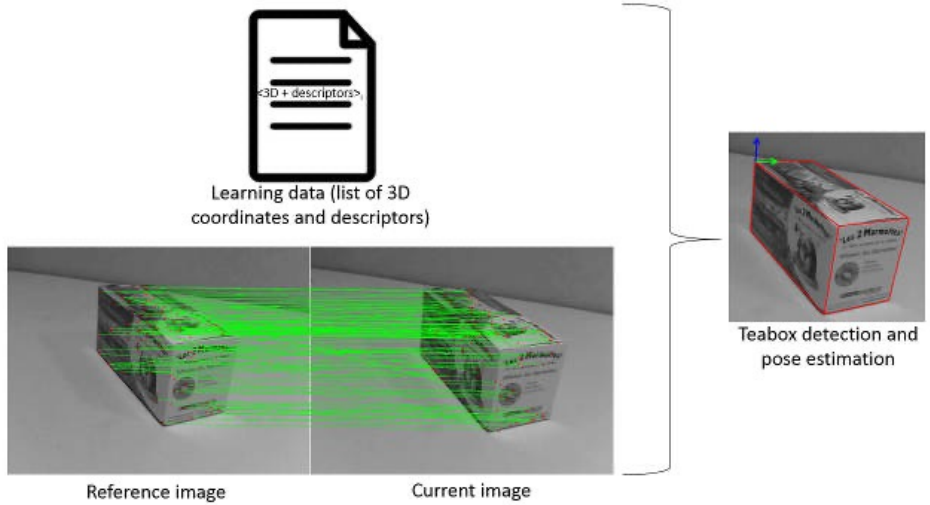
What is Visual Servoing?

- Compare current and desired visual features
- Continuously extract visual features $s(m(t), a)$
- Define error: $e(t) = s(t) - s^*$
- Repeat until error below desired threshold

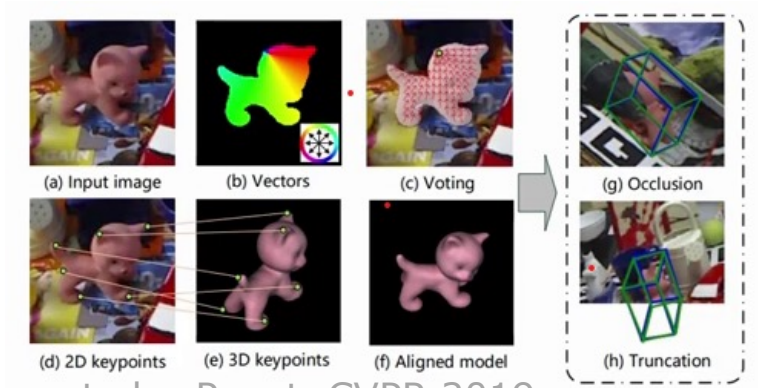


Visual Features

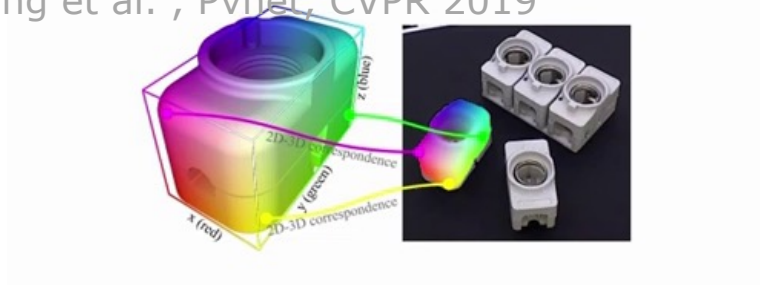
- Measurements $m(t)$ from the image (e.g., the image coordinates of interest points)
- Visual features $s(m(t), a)$
- Desired feature vector s^*
- Examples of visual features:
  - Object centroid or bounding box
  - Line or edge parameters
  - Key points on gripper or object



ViSP Robotics Toolbox



Peng et al. , Pvnet, CVPR 2019



Park et al. , ICCV, 2019

Interaction Matrix

- Relates camera motion to feature motion
- Camera velocity $\xi = [v, \omega]$
- Feature dynamics $\dot{s} = L\xi$
- $L \in \mathbb{R}^{6\times k}$ is the interaction matrix or image Jacobian
- L depends on chosen features
- Assuming s^* is constant, $\dot{e}(t) = \dot{s}(t) = L\xi$
- Enables velocity-based visual servoing
- Determine ξ in each control step

Control Law: Reduce Visual Error

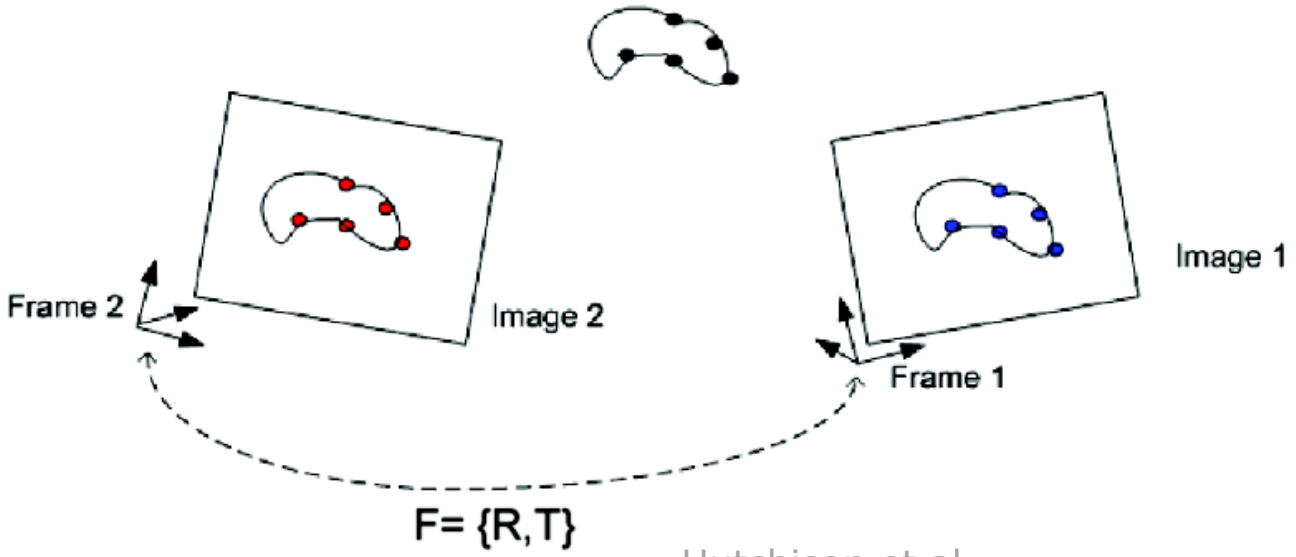
- Desirable to achieve exponential decay of error
- $e(t) = e(t_0)\exp(-\lambda t)$ → $\dot{e}(t) = -\lambda e$
- $L\xi = -\lambda e$
- $u(t) = \xi$, solve for ξ
- Obtain $\xi = -\lambda L^{+} e$, where L^{+} is the Moore-Penrose pseudo-inverse
- In practice, it is impossible to know exactly the value of L or of L^{+}, since these depend on measured data

Image-Based Visual Servoing (IBVS)

- Features $s(t)$ extracted from image data
- Error defined in the image feature space: $e(t) = s(t) - s^*$
- Control signal $\xi = [v, \omega]$ is camera body velocity, i.e., control in Cartesian space
- Computed directly from $s(t)$
- If the feature is a single image point with image plane coordinates x and y, we have $s(t) = (x(t), y(t))$
- Often robust to calibration errors
- Motion can be less intuitive

Position-Based Visual Servoing (PBVS)

- Estimate 6D object pose relative to camera
- Error $e(t)$ defined in 3D pose space, i.e., in SE(3)
- Control signal $\xi = [v, \omega]$ is camera body velocity, i.e., control in Cartesian space
- Sensitive to camera calibration and pose estimation errors



Hutchison et al.

Visual Servoing Examples

PVNet + IBVS

PVNet + PBVS

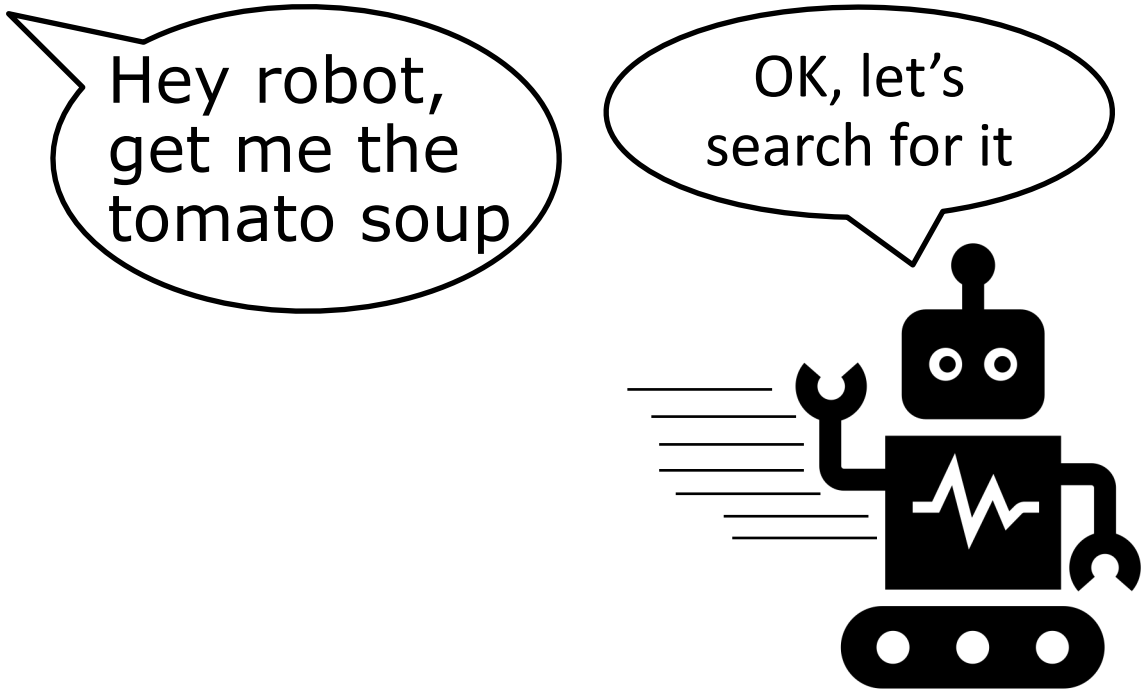
Tao et al. , ICRA, 2026

Visual Servoing: Summary

- Enables closed-loop control: observe, correct, act
- Servoing links vision to control
- IBVS acts through image features
- PBVS acts through 3D pose

Visual Scene Exploration

Motivation


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


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

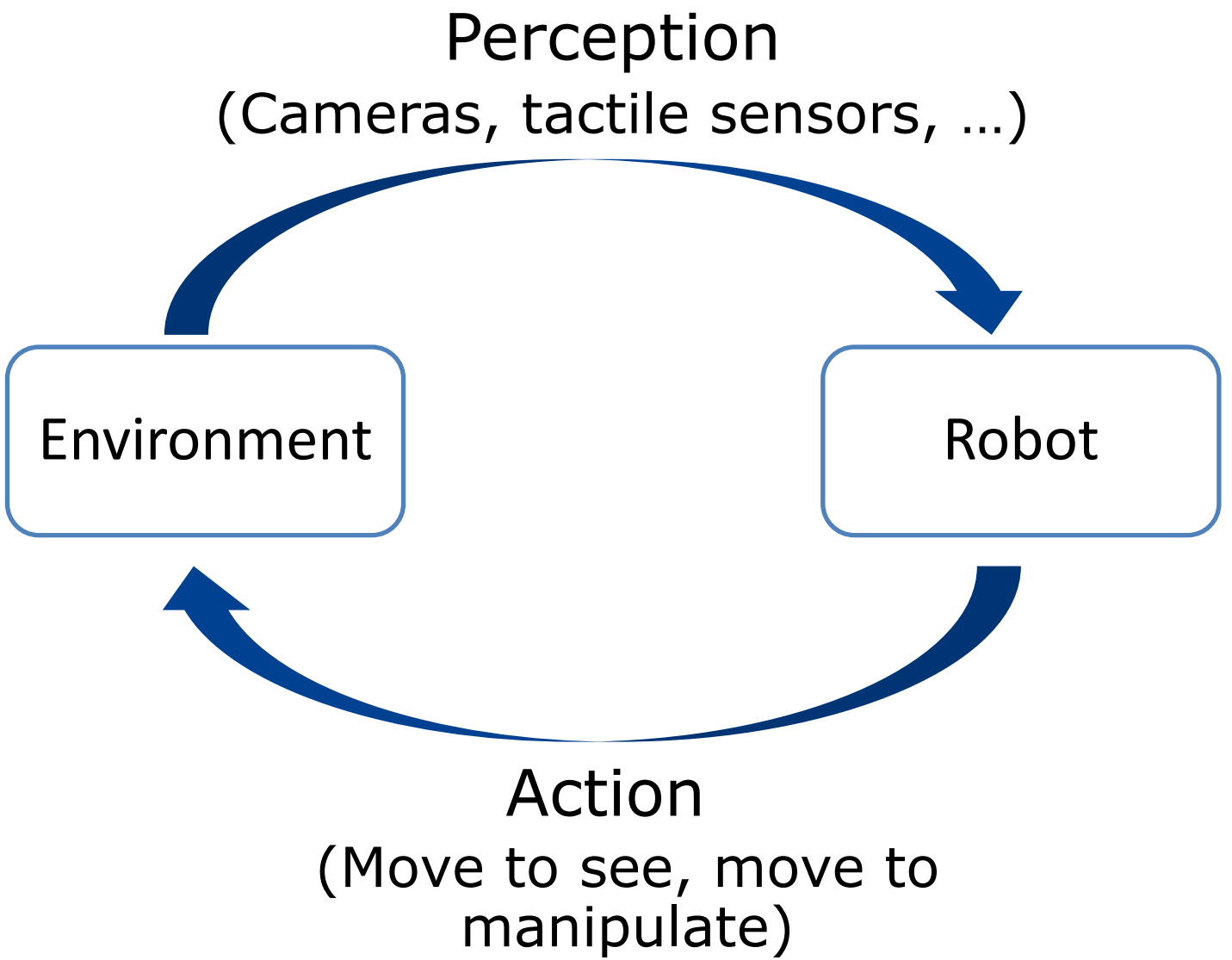
Motivation

How to see behind the boxes?

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

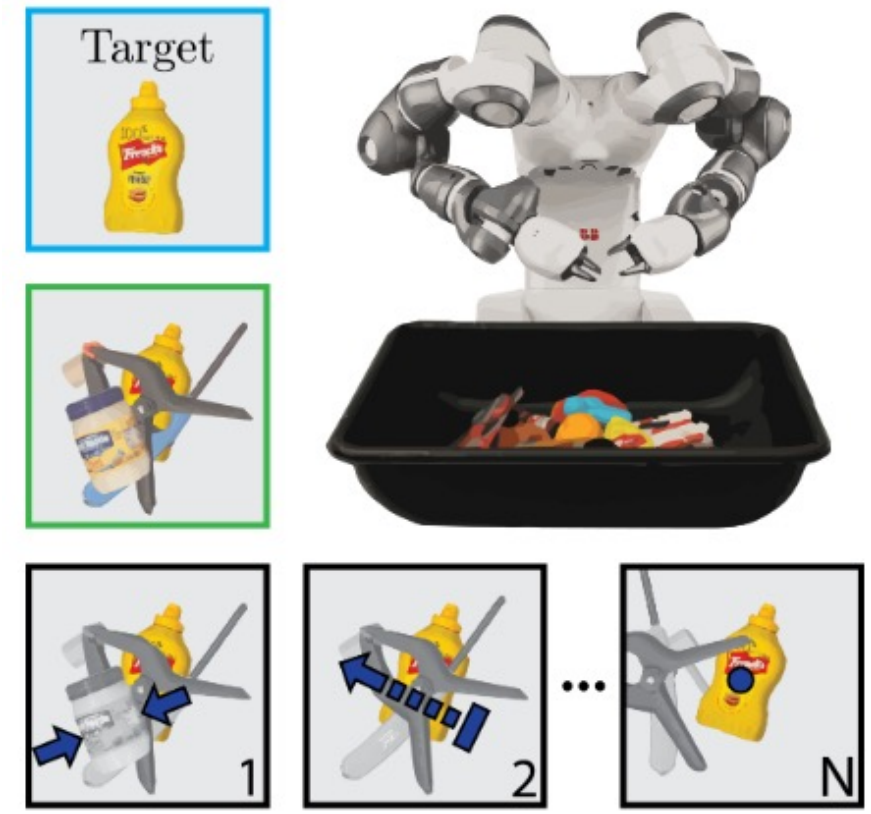
Interactive Perception

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



Mechanical Search for Object Retrieval

- **Search** for and **retrieve** objects from cluttered **bins**, **shelves**, or **tables**
- Involves **locating** (interactive perception) and **extracting** a given target object



[Danielczuk et al., ICRA 2019]

Mechanical Search for Object Retrieval

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

object found

push objects to the left

[Bejjani et al., IROS 2021]

Summary: Mechanical Search

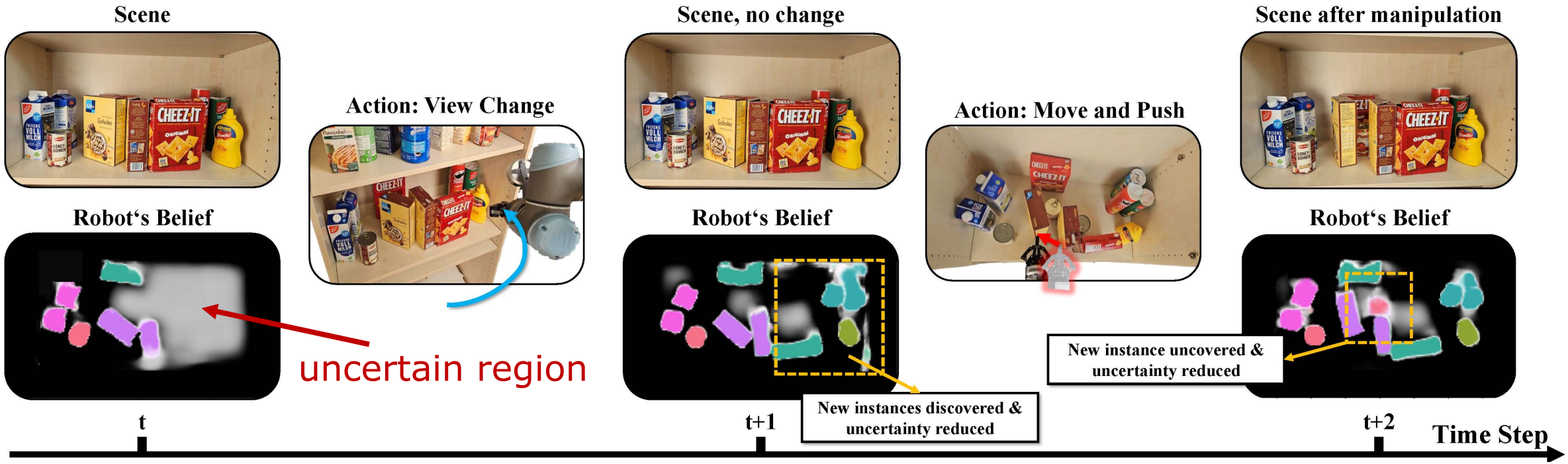
- **Goal**: Retrieval of a specific object, even if occluded by or "buried" in clutter
- **Actions**: Push, pull, or grasp to remove occluders
- Enables task completion despite limited visibility, robust to clutter and occlusions
- **Drawbacks**: Focused on object-level tasks, **not** scene understanding

Interactive Scene Exploration

- **Active perception** reduces uncertainty but **insufficient** for completely covered spaces
- **Physical interaction** (moving object aside) helps to
  - Uncover hidden objects
  - Refine object pose estimates or properties
- However, those actions change the scene and **may increase uncertainty** in the global map
- Idea: Informed action selection for **manipulation actions**
- Uncertainty-guided planning to balance risk and discovery

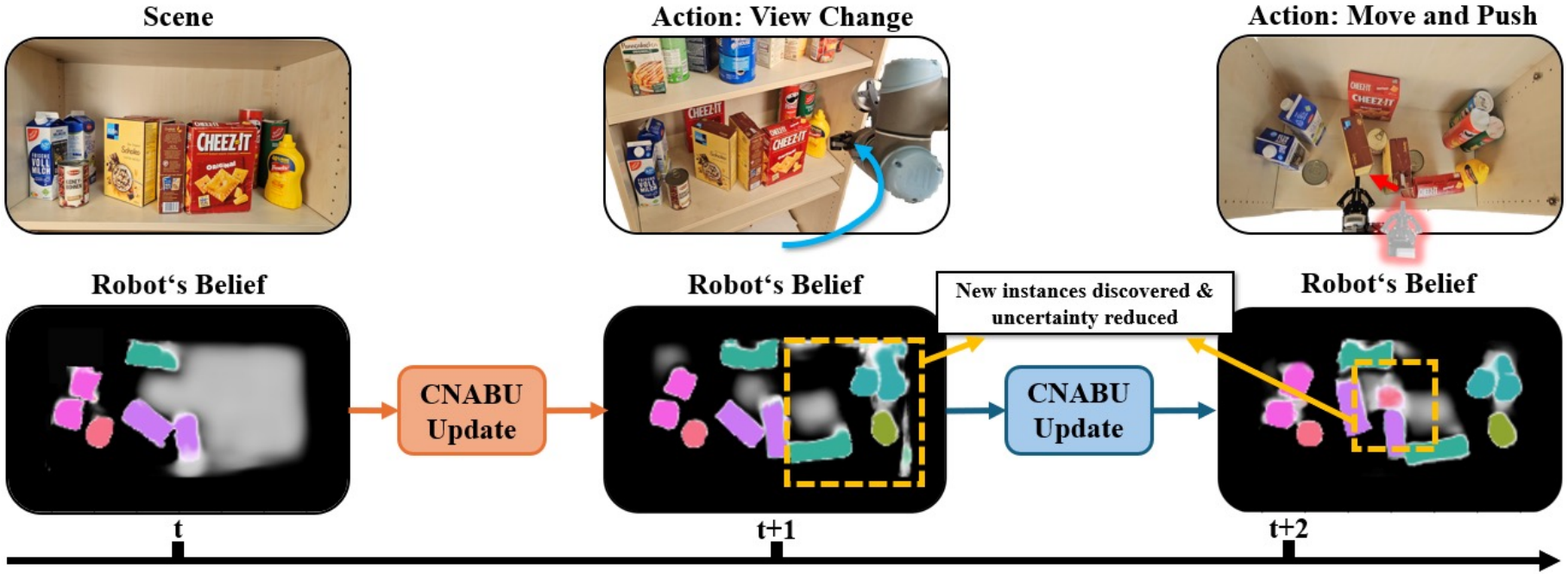
Manipulation-Enhanced Mapping

- 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

Manipulation-Enhanced Mapping



- **Goal:** Map a scene **as completely as possible**
- Select actions by expected **information gain**
- **Predict belief updates** with learned map-space models
- **Calibrated Neural Accelerated Belief Updates (CNABU)**

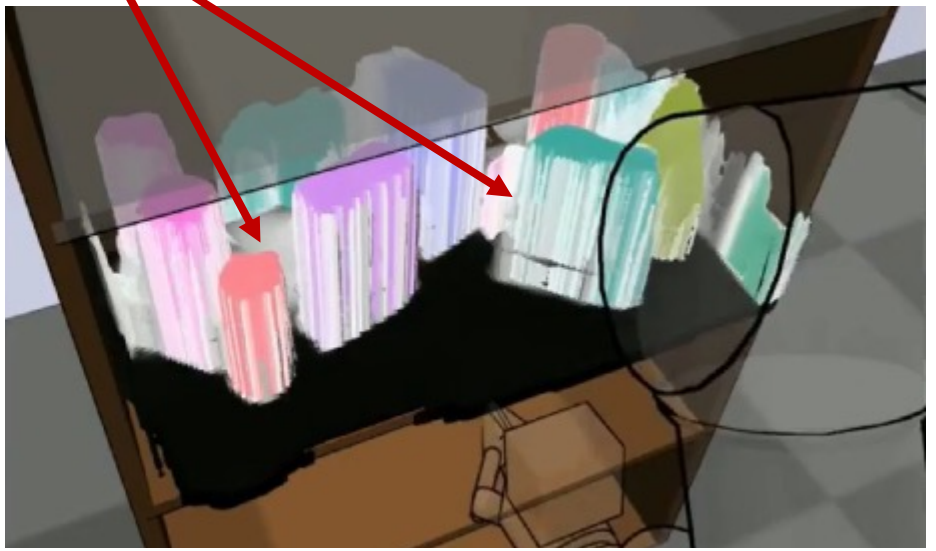
Environment Representation (Belief)

- **3D voxel map** with **occupancy** and **semantic** estimates
- Modeling of the robot's **uncertainty**

Real-World

uncertain regions

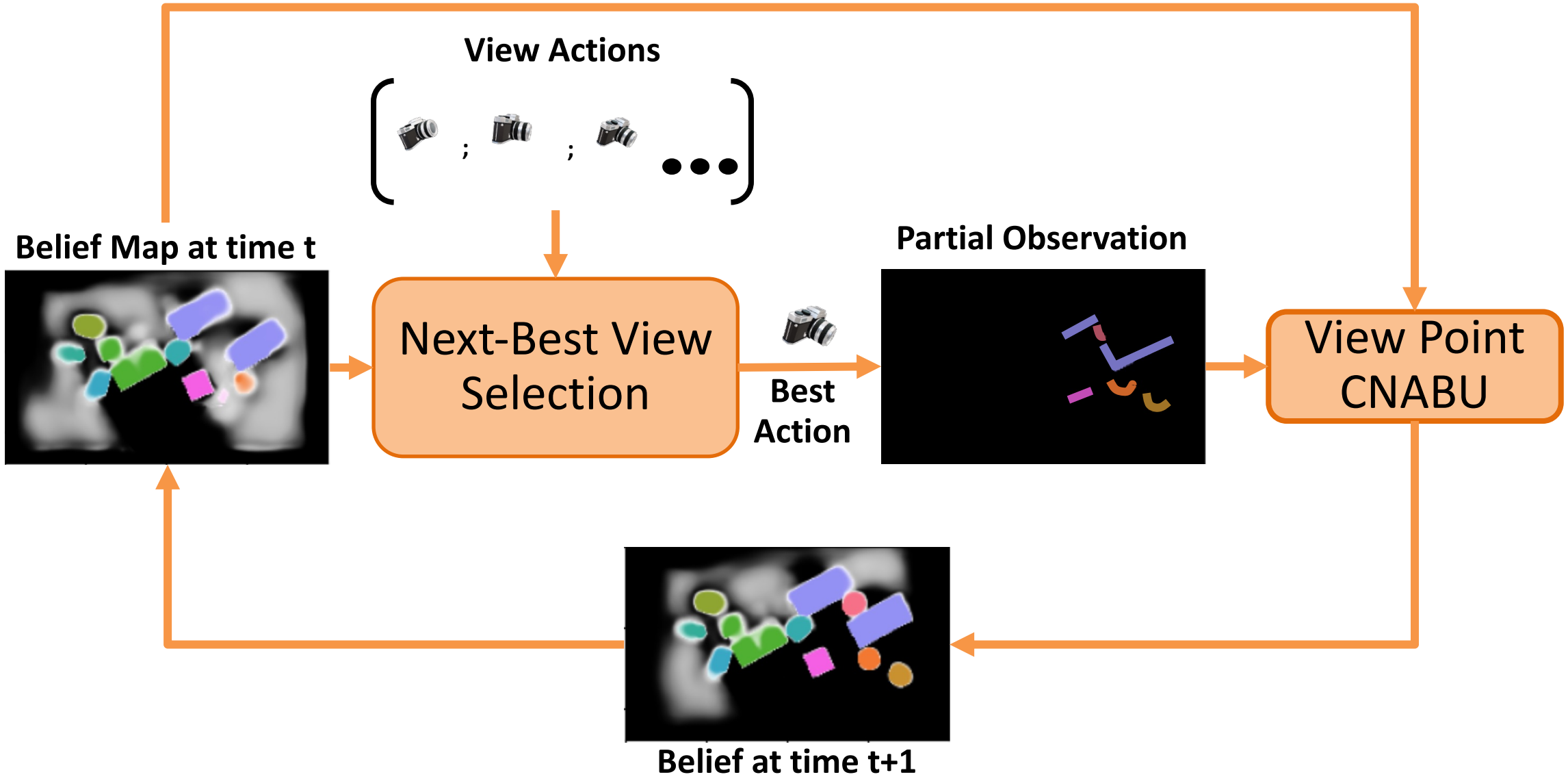
Belief



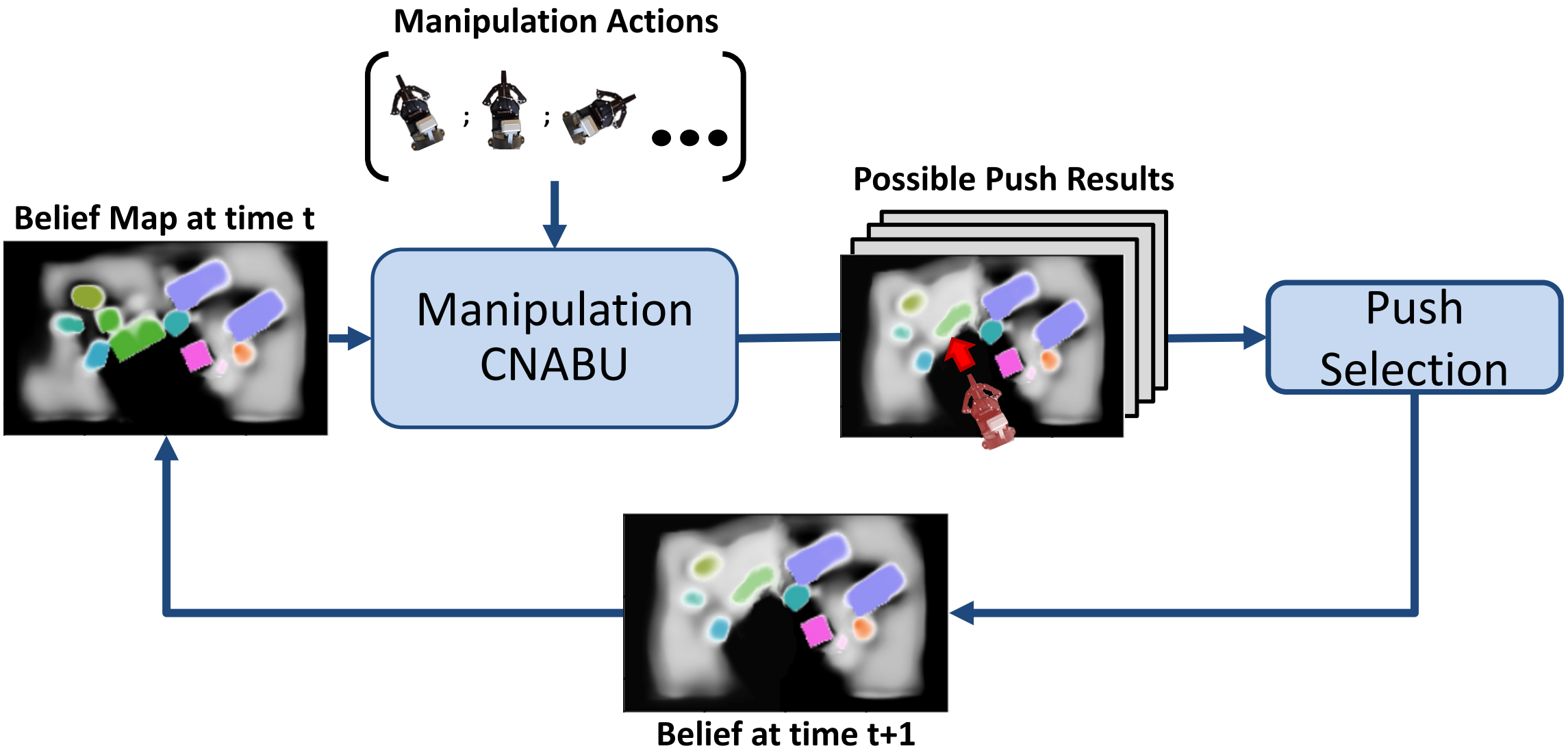
Effect of Actions

- **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 **high expected net information gain**
- Action–perception loop **balances information gain** against **manipulation risk**

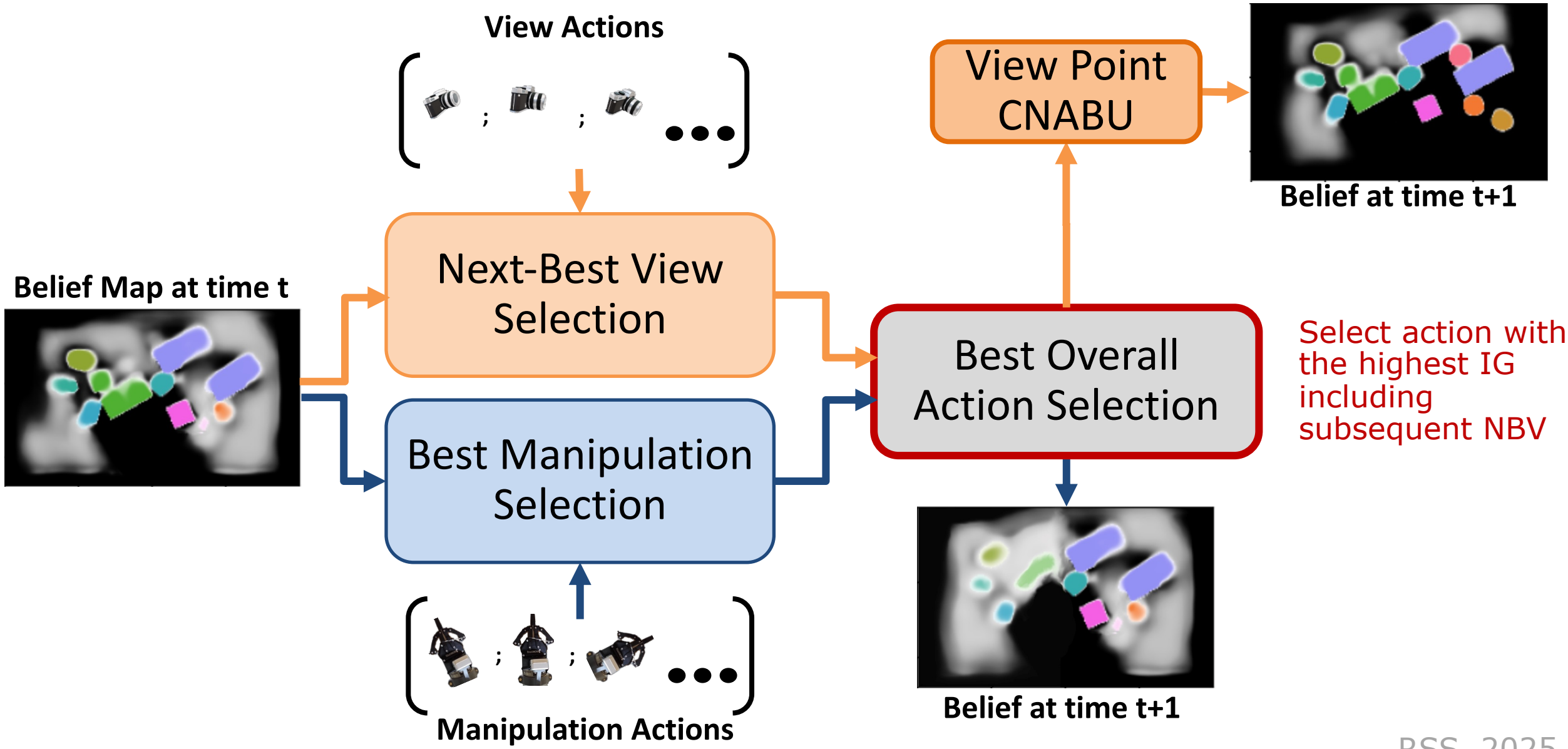
Belief Update After Observation Action



Belief Update After Manipulation Action



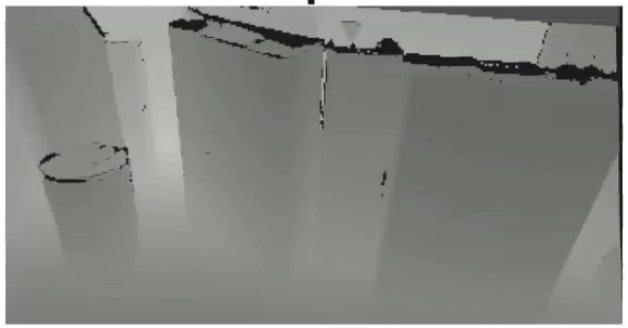
Action Selection



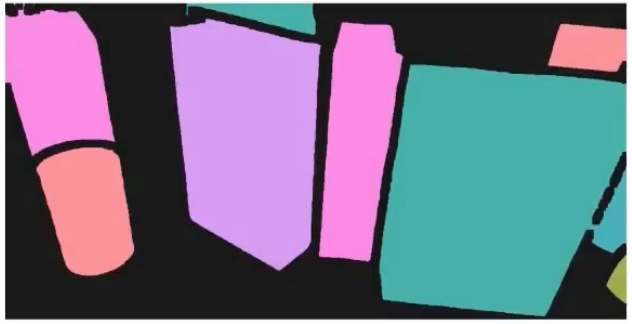
Real-World Experiment

RGB

Depth

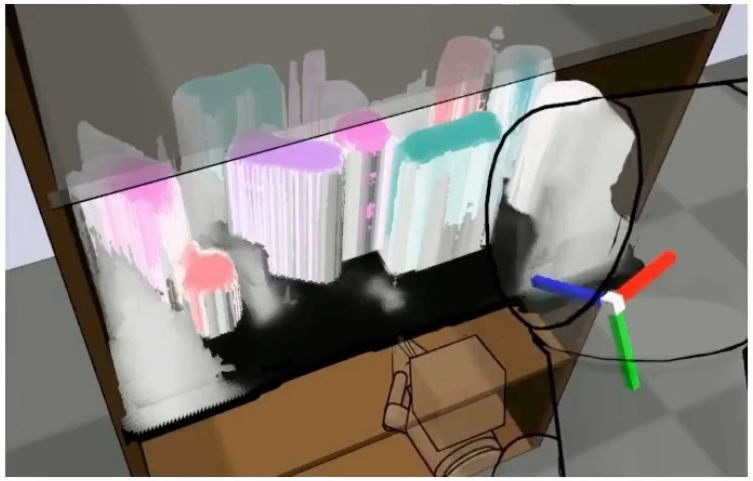


Semantic



External View

Environment Belief



Taken Action: Observation

Summary: Manipulation-Enhanced Mapping

- **Goal**: Map scene as completely as possible
- **Actions**: Push objects, re-orient camera
- **Reduces uncertainty** in semantic occupancy maps
- Yields **more accurate** maps
- Less efficient for single direct object retrieval, but enables efficient object search in the **long run**

Foundation Models for Robot Manipulation

Motivation

- Diverse tasks, heterogeneous robot embodiments
- Unstructured, changing environments
- Leverage large-scale pretraining from vision-language data
- Enable zero- or few-shot generalization to novel tasks
- Bridge semantic understanding with robot actions

Classical Grounding Stack

- However, most foundation models need physical grounding
- Perception yields uncertain scene understanding
- Geometry reasoning gives metric state representation
- Planning yields feasible and goal-directed actions
- Control ensures stable execution
- Monitoring detects runtime failures and triggers recovery

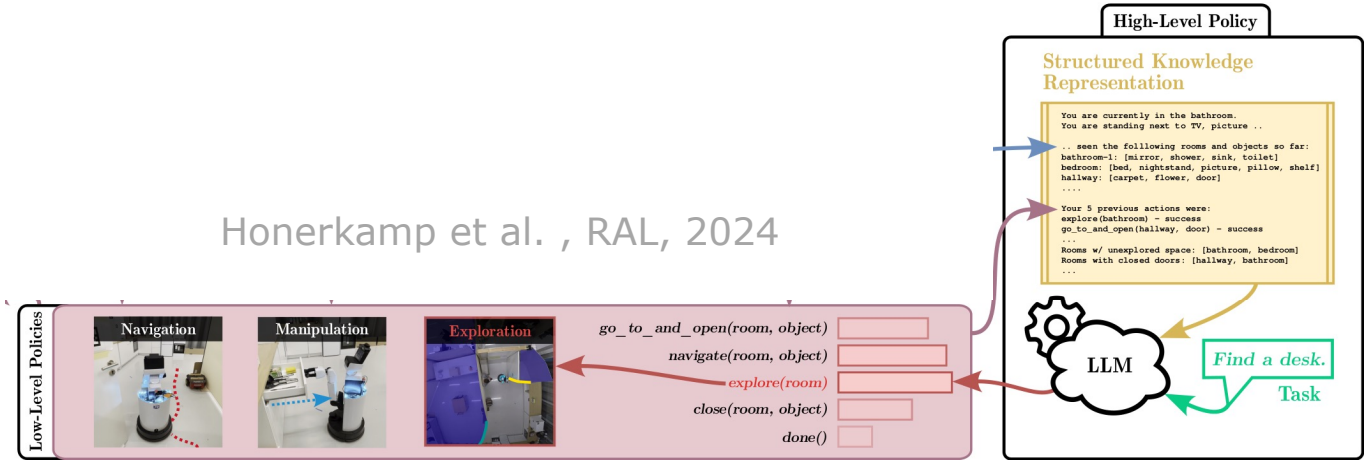
Foundation Model Categories

- **Large Language Models** (LLMs) for reasoning, e.g., ChatGPT
- **Vision Foundation Models** (VFMs) for vision tasks (2D and 3D), e.g., SAM2, SAM3, DINOv3, DepthAnythingV2
- **Vision-Language Models** (VLMs) ground language visually, e.g., CLIP
- **Vision Language Action Models** (VLA) generate robot actions, e.g., $\pi 0$
- **World Models** predict consequences of an input action e.g. PointWorld
- **World Action Models** predict joint evolution of actions and their outcomes, e.g., Unified World Model (UWM)

LLMs for Reasoning

- Example: “make tea”
- Plan: boil, steep, pour
- Select executable robot skills
- **Risk:** Generates physically infeasible steps
- **Grounding:** Enforce symbolic planning constraints

Honerkamp et al. , RAL, 2024



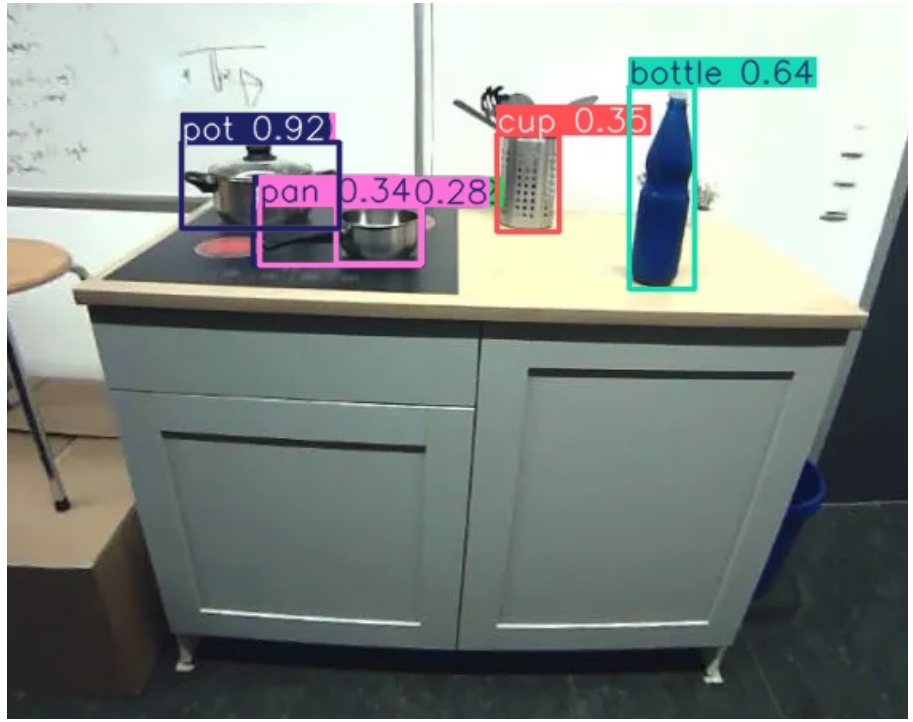
VFMs for Vision Tasks

- Example: monocular depth prediction using DepthAnythingV2
- Extract dense relative depth maps
- Captures class-agnostic geometric structure and depth discontinuities
- **Risk:** Lacks explicit 3D geometric and physical grounding
- **Grounding:** Lift 2D features into 3D spatial representations

Yang et al., NeurIPS, 2024

VLMs for Vision-Language Alignment

- Example: Find “cup”
- CLIP ranks language-image candidates
- Grounded segmentation localizes entities
- **Risks:** May confuse similar objects, language prompts, lacks intrinsic 3D understanding
- **Grounding:** Anchor semantics using VFM masks and 3D maps

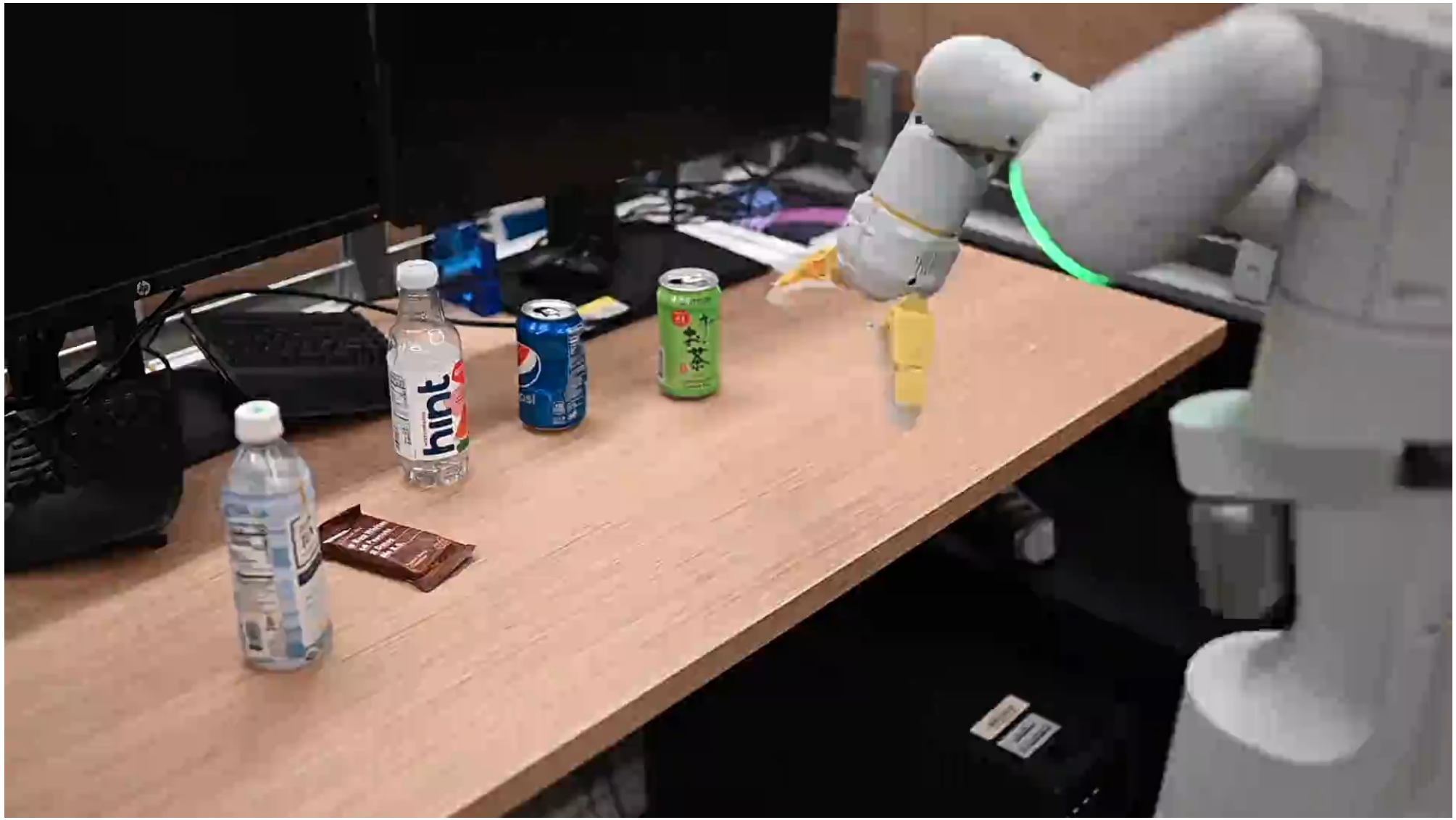


Spatula holder incorrectly detected as cup

VLAs for Robot Actions

- Example: "Grasp the target cup“
- Predict robot actions (e.g., end-effector motions or joint commands)
- Enable language-conditioned control
- **Risk:** Action predictions remain reactive, may collide near clutter, may fail outside training embodiment
- **Grounding:** Integrate inverse kinematics, collision checking, and safety filters

RT-2: Vision-Language-Action Model

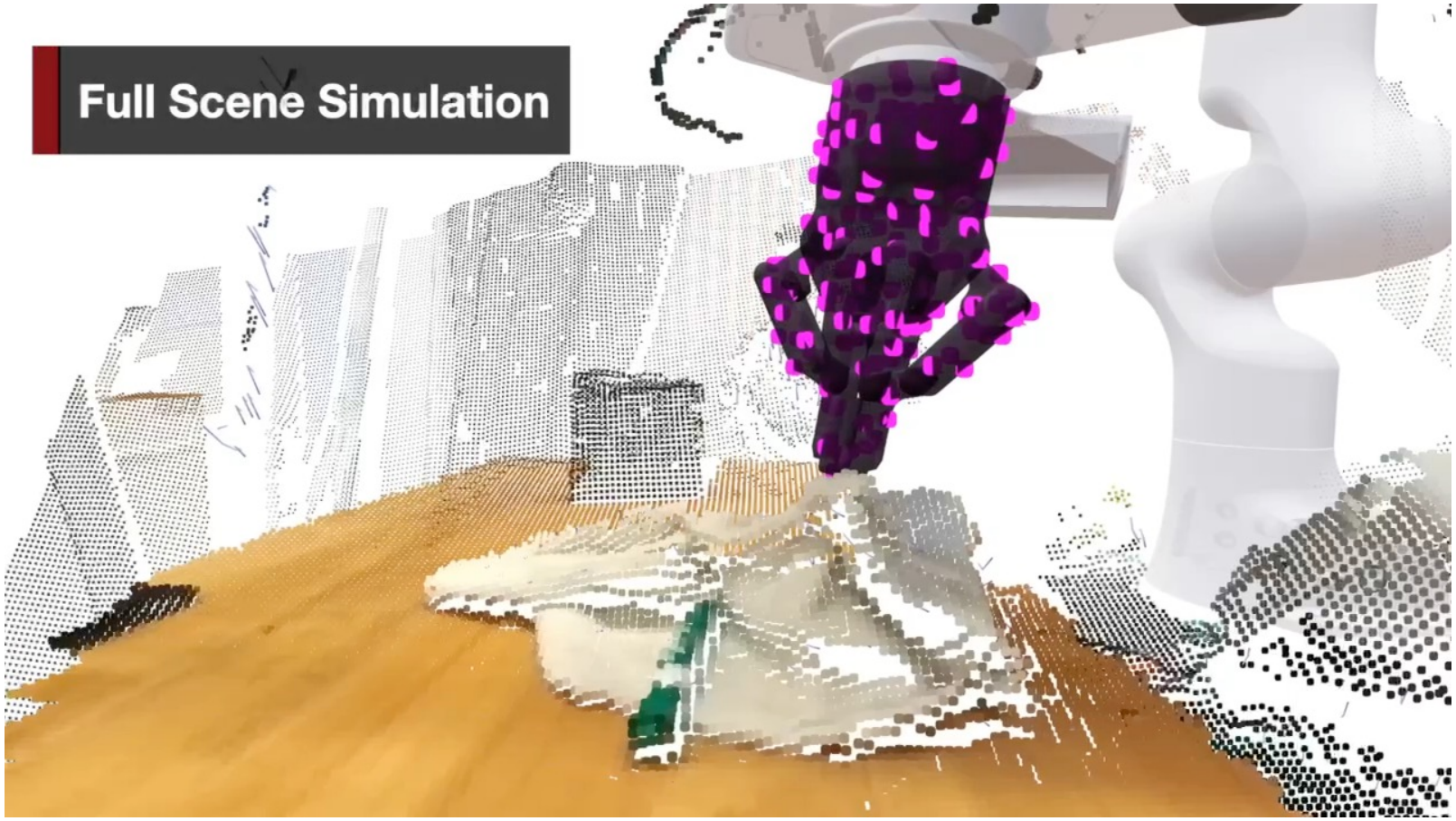


[Zitkovich, et al. "RT-2: Vision-language-action models transfer web knowledge to robotic control", CoRL2023]

World Models

- Example: Candidate push predicts cup motion
- Predict: Cup translates leftward
- Compare possible action outcomes
- **Risk:** Inaccurate contact dynamics modeling, errors accumulate across rollouts
- **Grounding:** MPC constrains predicted rollouts

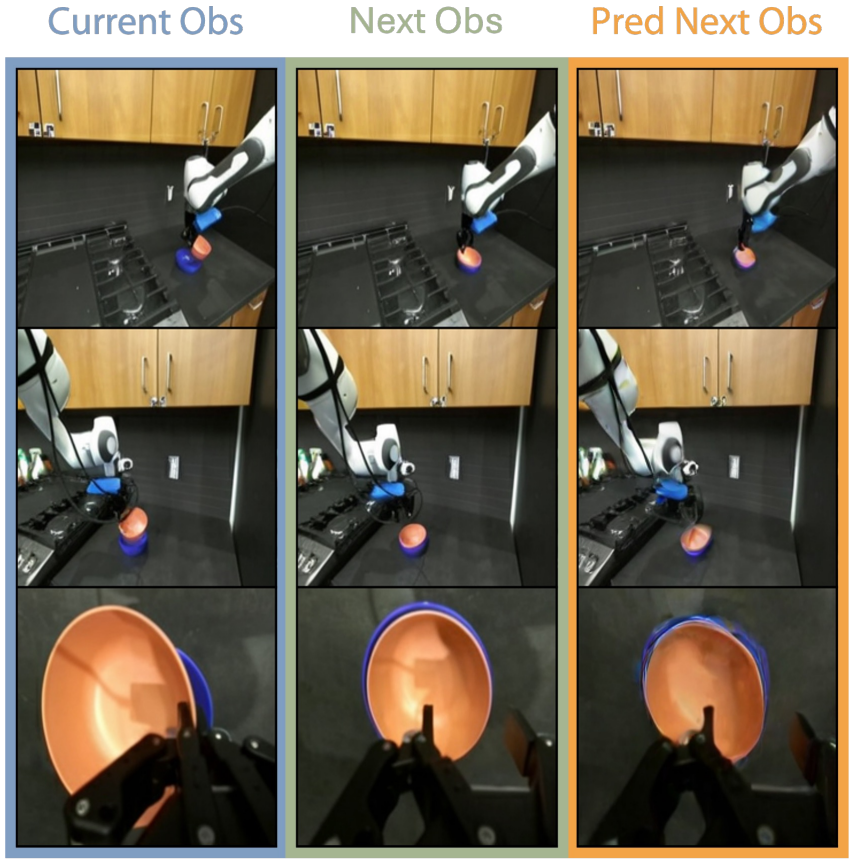
PointWorld World Model



Huang et al. , CVPR, 2026

World Action Models

- Example: Put orange bowl on top of blue one
- Predict next action observation pair
- Couple action and video
- **Risk:** Errors compound across predicted frames, affect subsequent robot actions
- **Grounding:** Correct via real-time sensory feedback



Zhu et al. , RSS, 2025

Challenges & Future Directions

- Ensure safe execution
- Handling data heterogeneity across robots
- Scaling to high-dimensional control spaces
- Integrating real-time feedback loops
- Improving model explainability & debuggability

Summary (1)

- Visual servoing **links vision to control** to achieve **closed-loop control**
- **Mechanical search:** interacting with 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 **leverage large-scale pretraining** from vision-language data and **bridge semantic understanding with robot actions**
- But **physical grounding** is still an open research topic

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
- *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