ICLR
International Conference On Learning Representations

# RoboMD: Uncovering Robot Vulnerabilities through Semantic Potential Fields

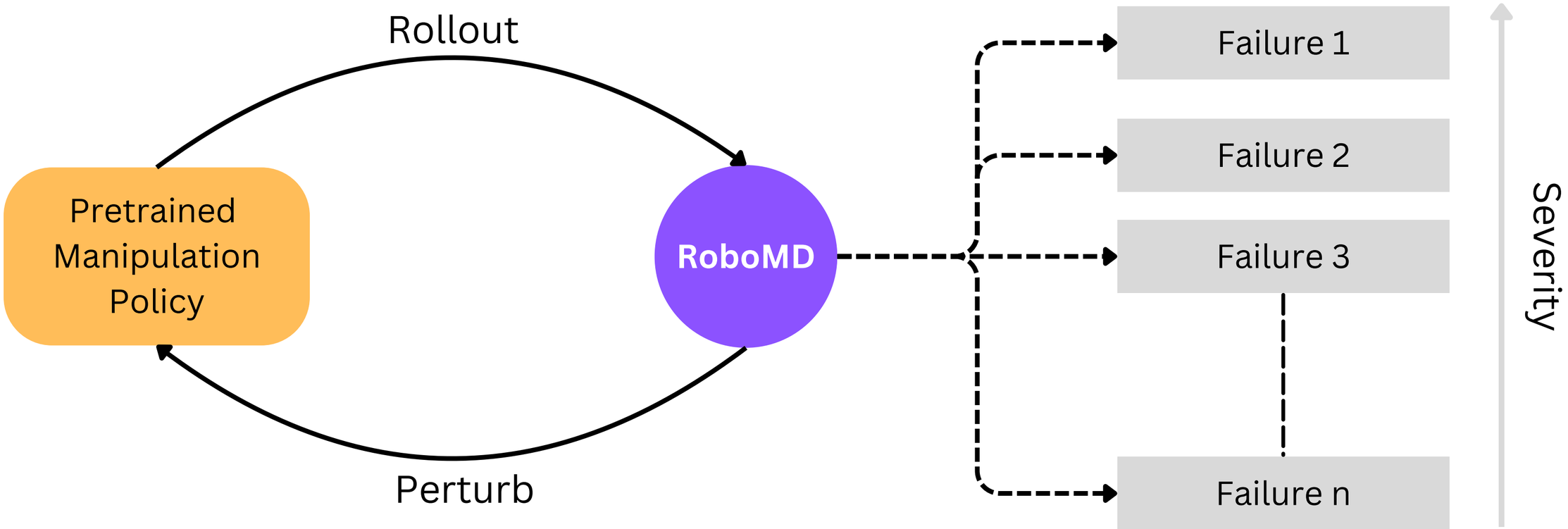
Som Sagar, Jiafei Duan, Sreevishakh Vasudevan, Yifan Zhou,
Heni Ben Amor, Dieter Fox, and Ransalu Senanayake

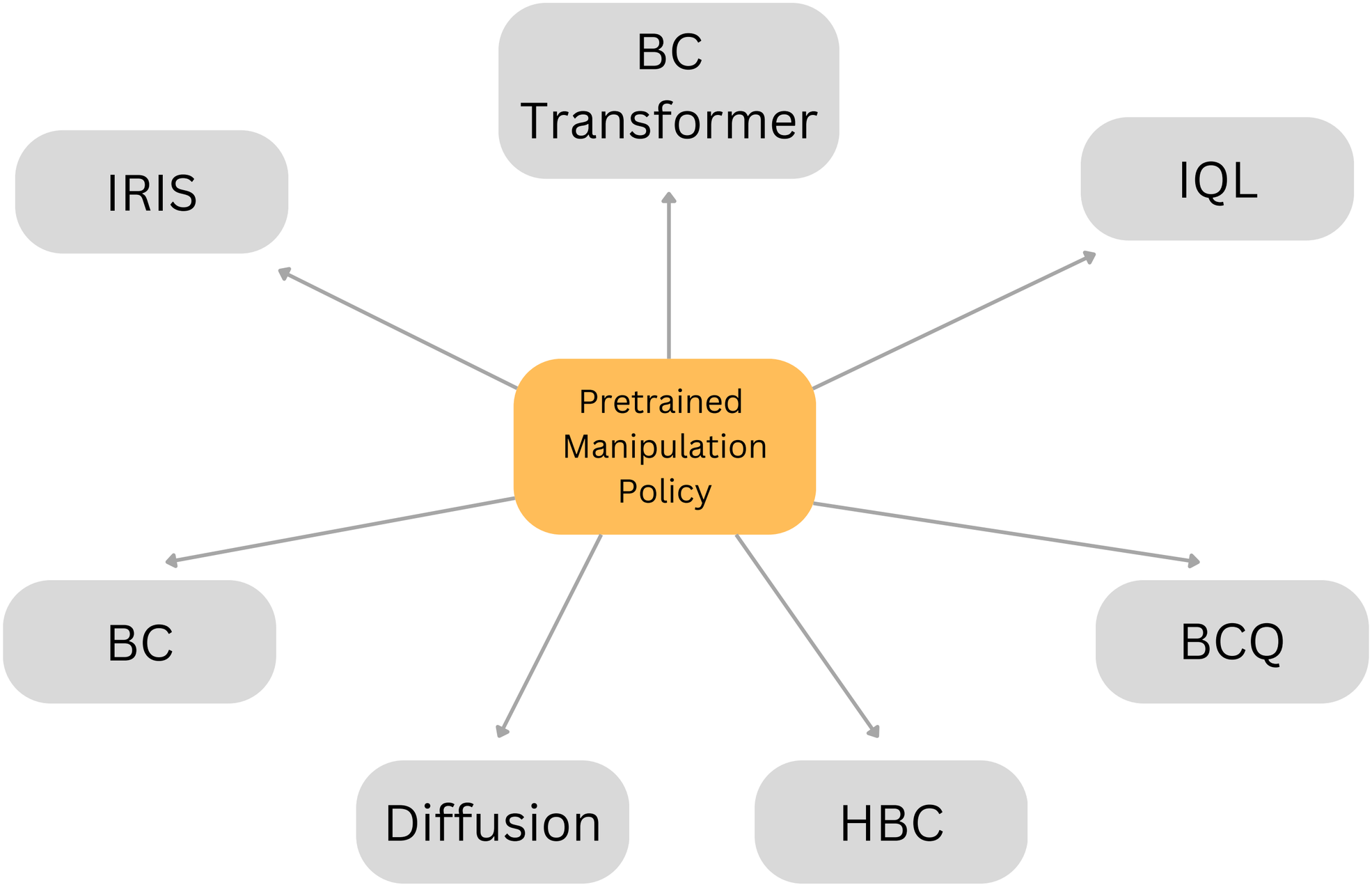
LENS Lab

ASU Arizona State University

UNIVERSITY of WASHINGTON

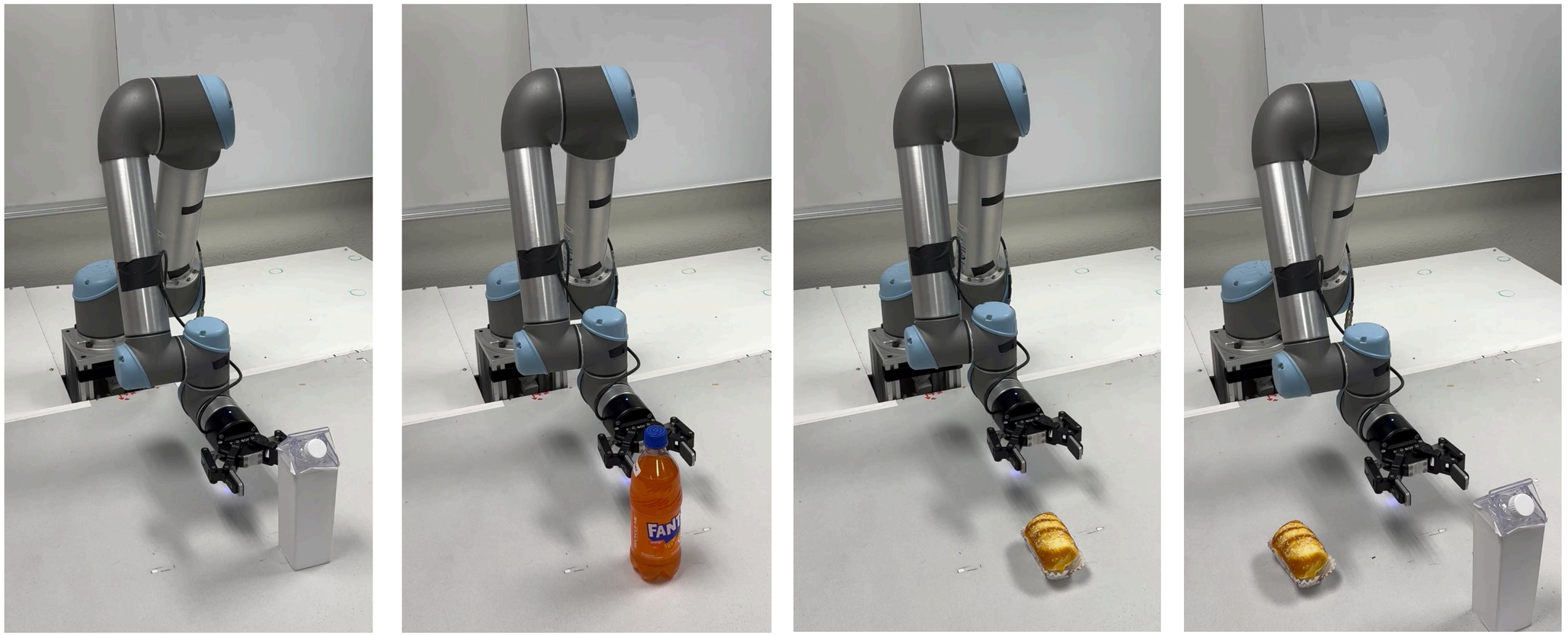
Our framework, **RoboMD**, provides a systematic approach for **diagnosing and quantifying failure modes** in **robot manipulation policies**, enabling robust performance improvement across **diverse and unseen environments**




BC Transformer
IRIS
IQL
Pretrained Manipulation Policy
BC
BCQ
Diffusion
HBC

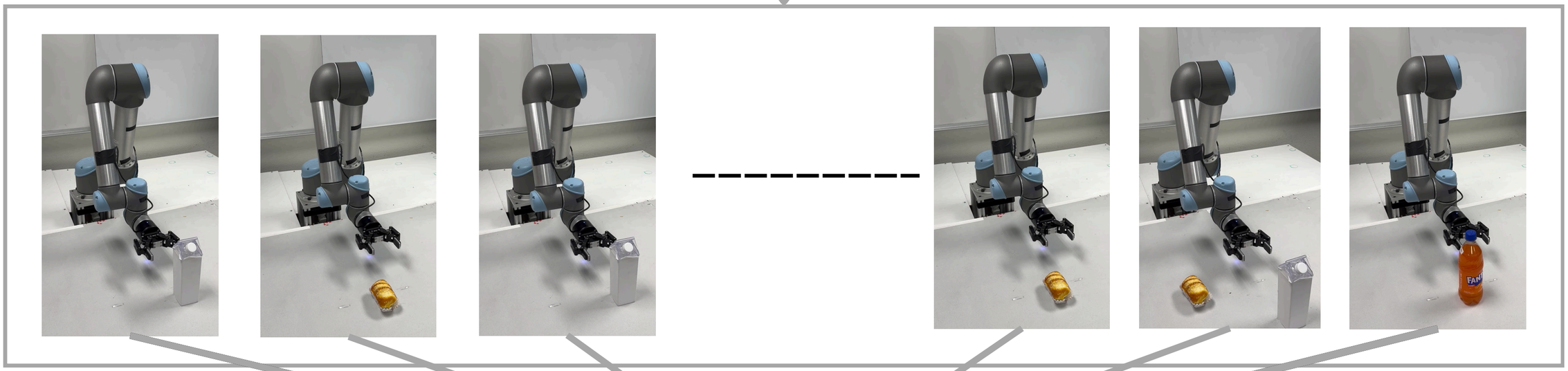
Pretrained Manipulation Policy

Rollout



Pretrained Manipulation Policy

Choose rollout of Milk Carton



Action = “Milk Carton”

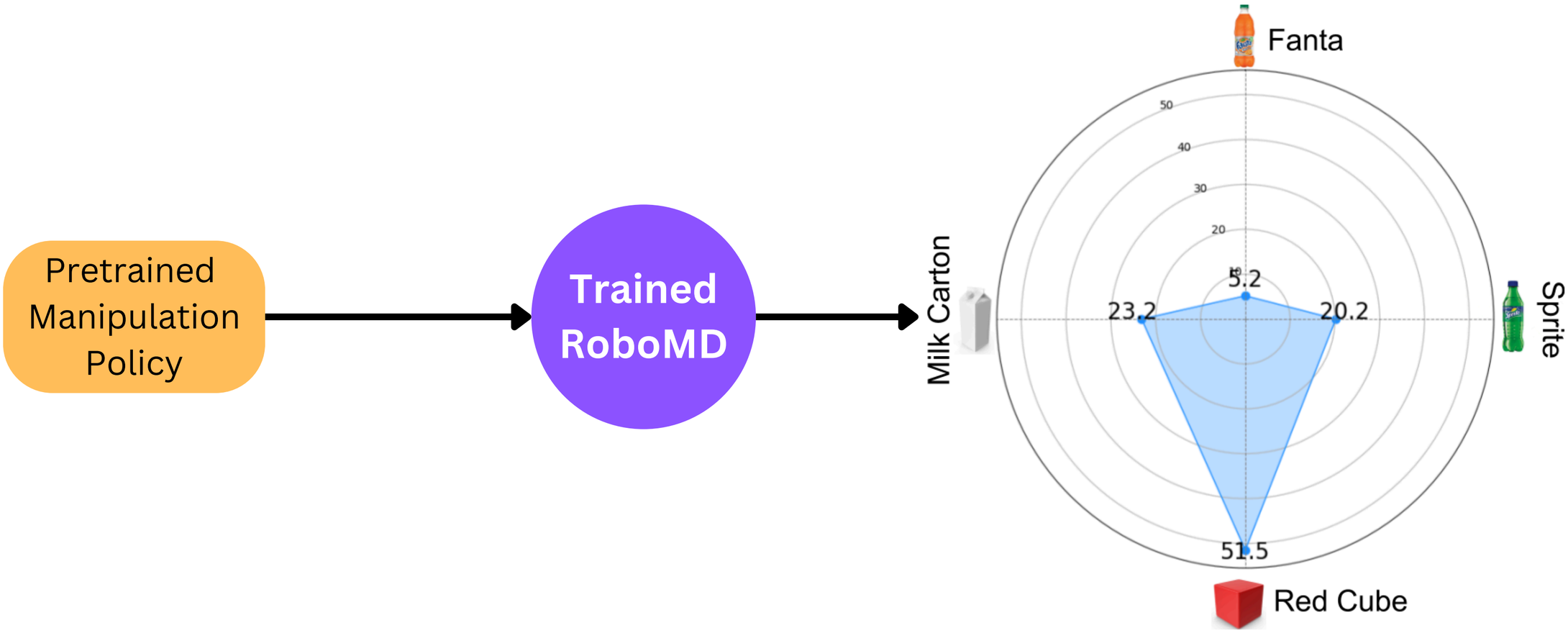
Training RoboMD

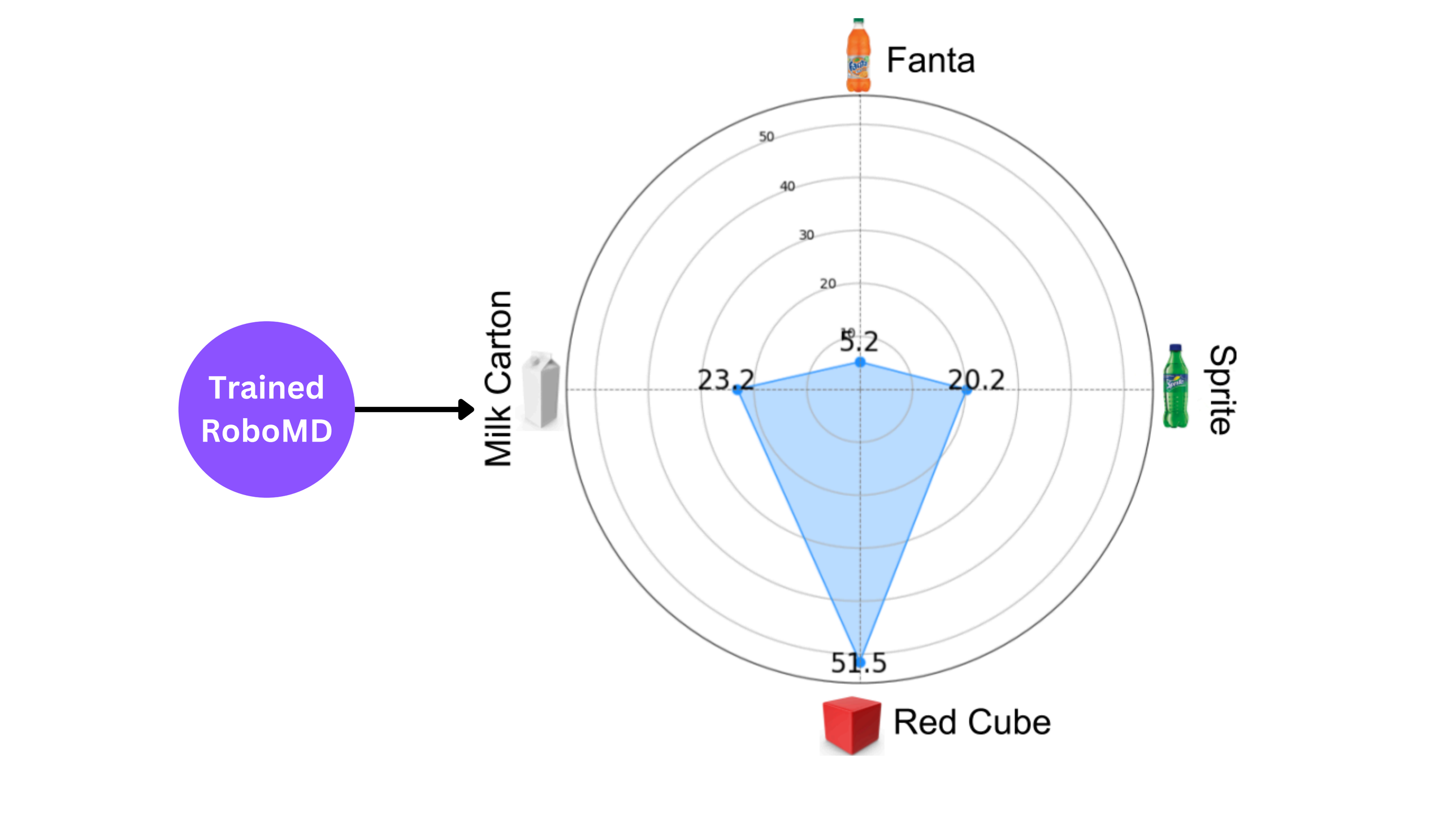
Reward based on failure during rollout

Pretrained Manipulation Policy

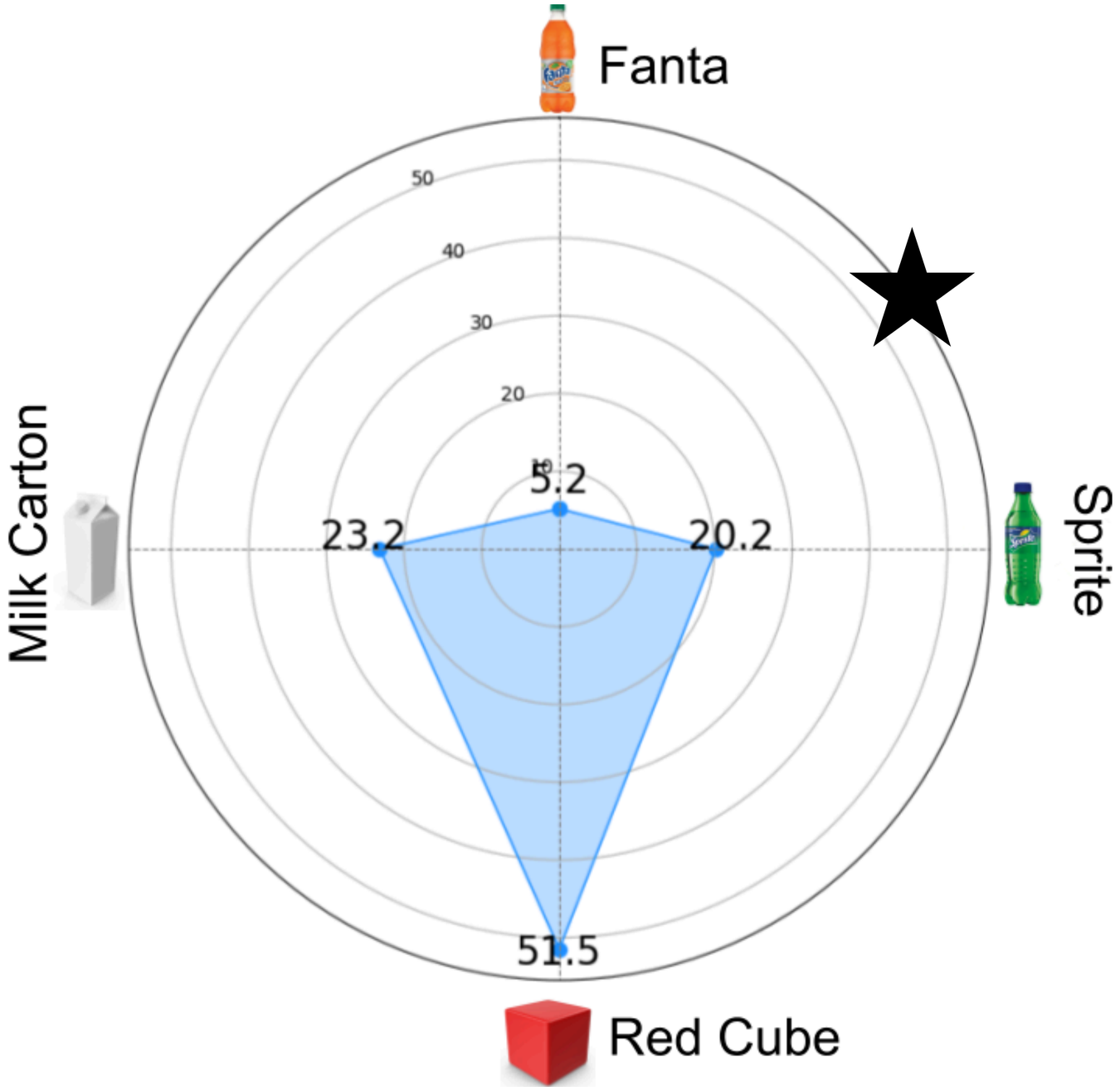
Through **RL RoboMD** observes multiple observations and learns a distribution of **failure over multiple actions (Bread, Fanta, Milk Carton)**

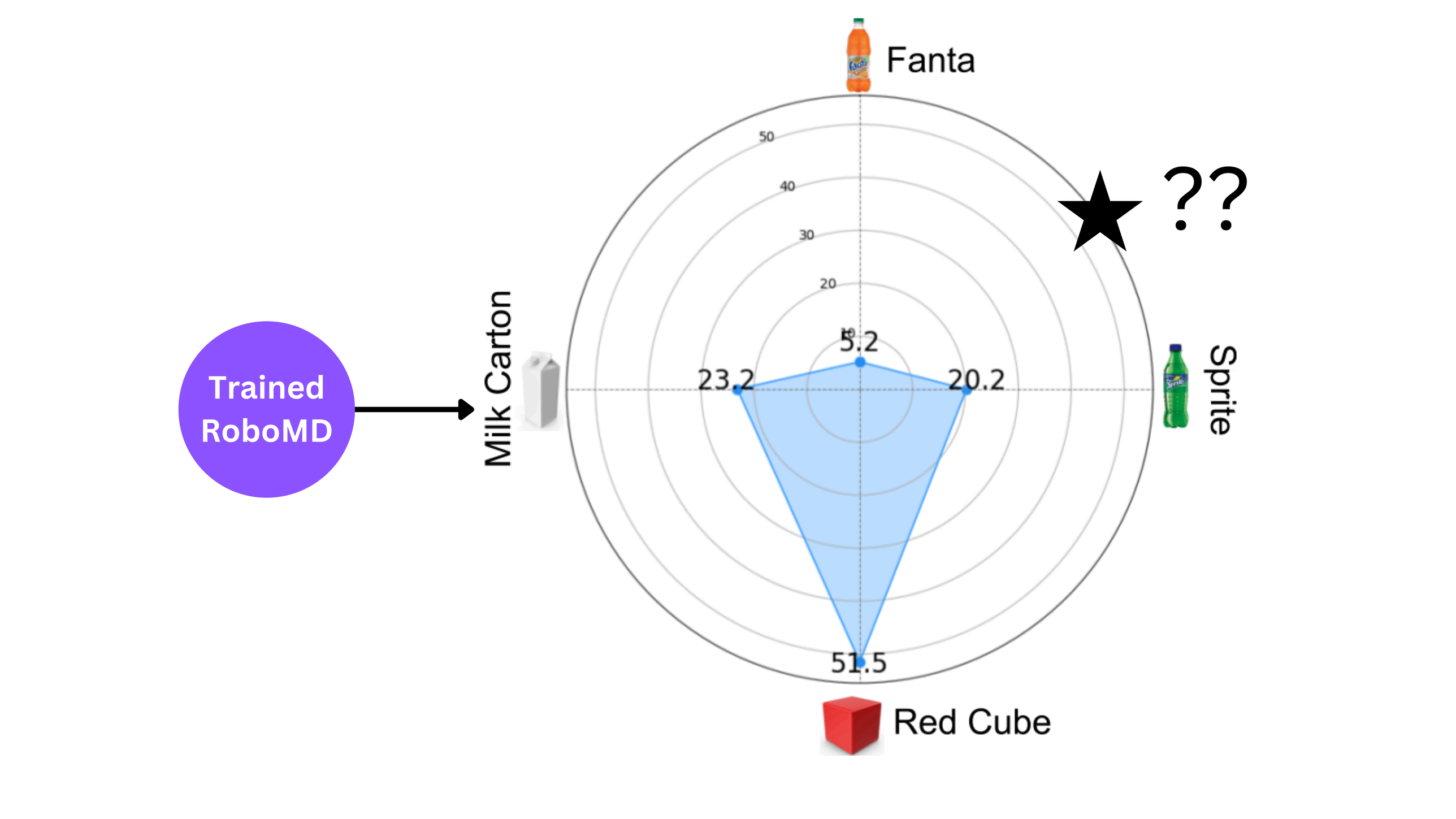
**Training RoboMD**


Pretrained Manipulation Policy
Trained RoboMD
Fanta
Sprite
Milk Carton
Red Cube
50
40
30
20
10
5.2
23.2
20.2
51.5


Trained RoboMD
Fanta
Milk Carton
Sprite
Red Cube
50
40
30
20
10
5.2
23.2
20.2
51.5

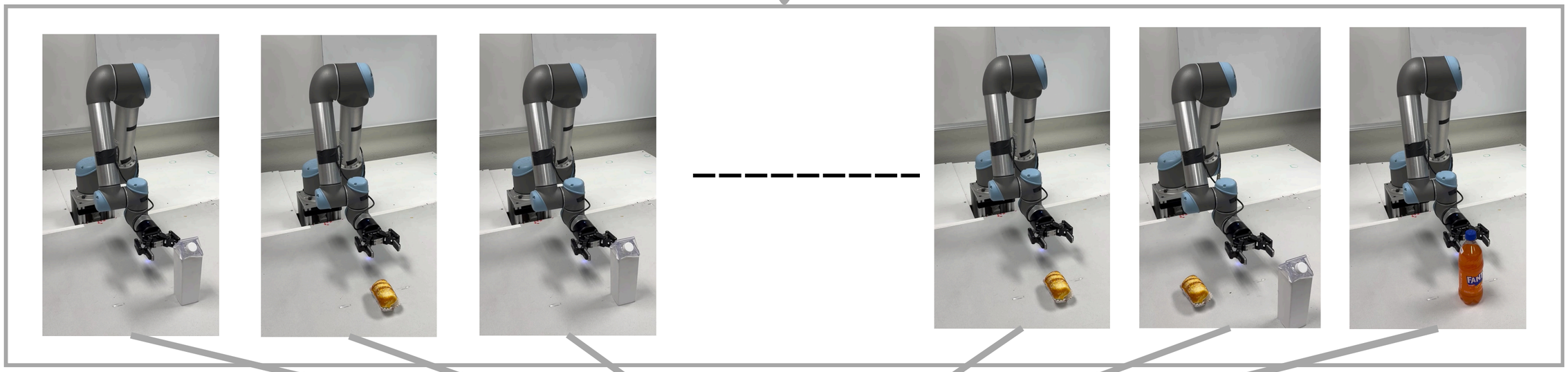
Trained RoboMD




Trained RoboMD
Fanta
Milk Carton
Sprite
Red Cube
??
50
40
30
20
10
5.2
23.2
20.2
51.5

Pretrained Manipulation Policy

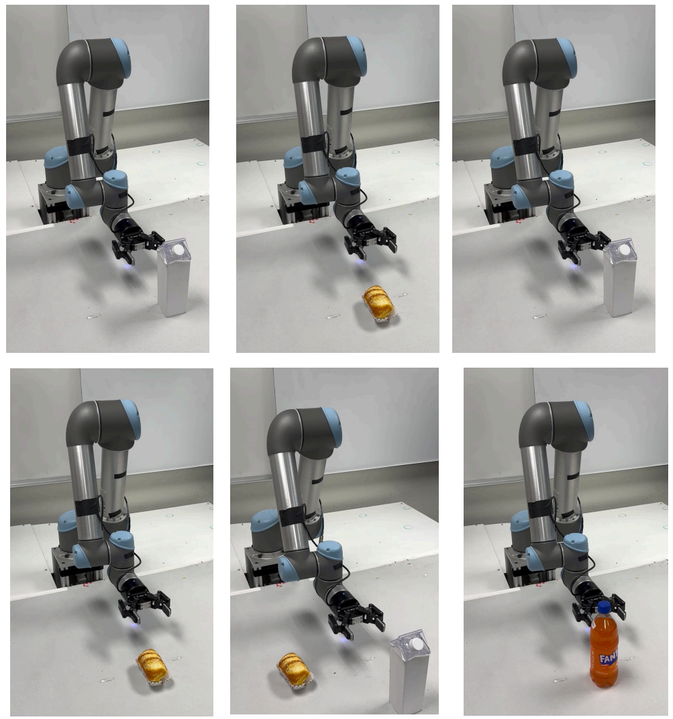
Change RL action space to handle continous action space



Traing RoboMD

“Change object to red cube”,
“Change object to bread”
.....
“Change object to Fanta”

Descriptions



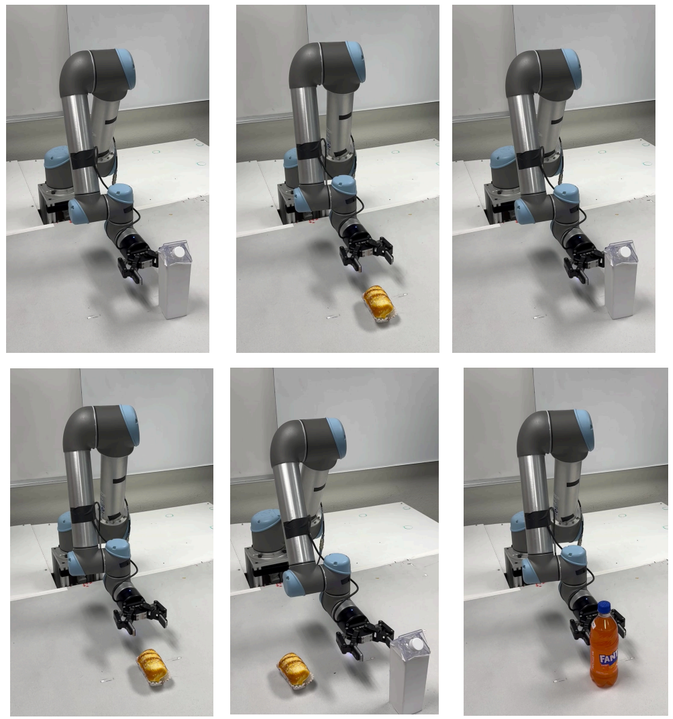
CLIP

Flattened

ViT

“Change object to red cube”,
“Change object to bread”
.....
“Change object to Fanta”

Descriptions



CLIP

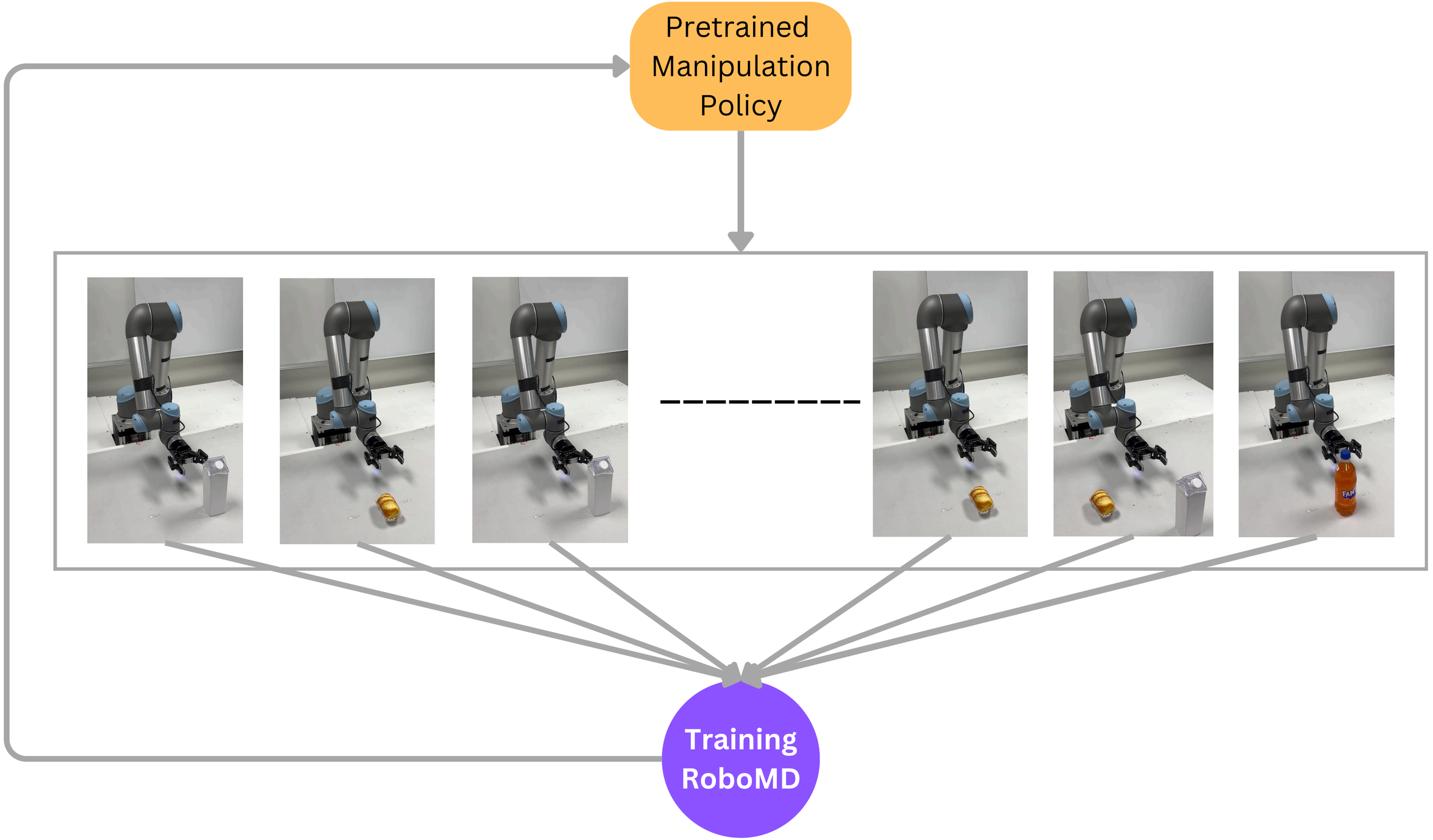
Flattened

ViT

MLP Projector

1/0

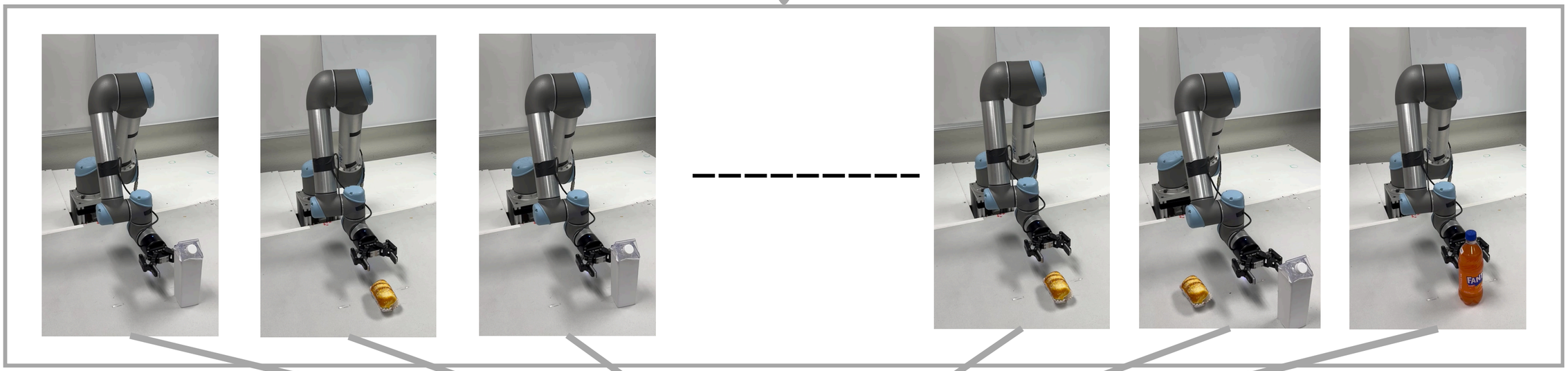
Embedding layer


Pretrained
Manipulation
Policy
Training
RoboMD

Search for closest known embedding

Pretrained Manipulation Policy

Choose rollout of Fanta



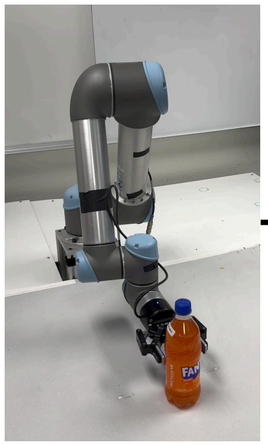
Random action embedding

Training RoboMD

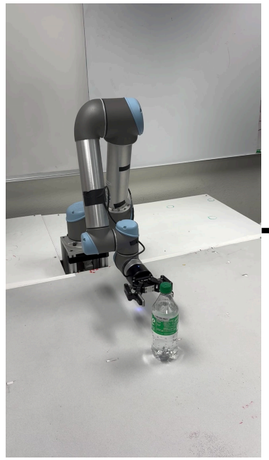
$$Reward \propto \frac{1}{d\left(e_{known}, e_{unknown}\right)}$$

Rank seen and unseen scenarios
based on their likelihood to fail!

# Rank seen and unseen scenarios based on their likelihood to fail!



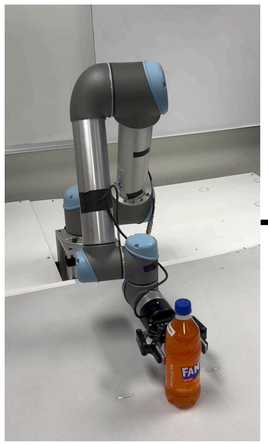
Seen



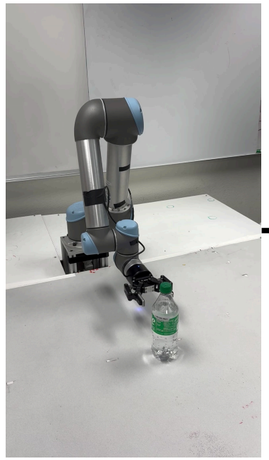
Unseen

**Trained RoboMD**

# Rank seen and unseen scenarios based on their likelihood to fail!



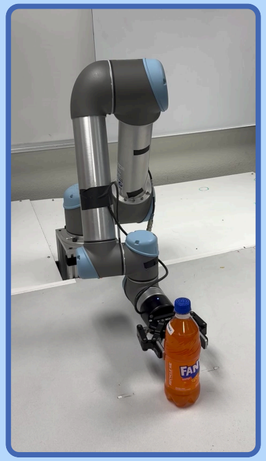
Seen



Unseen

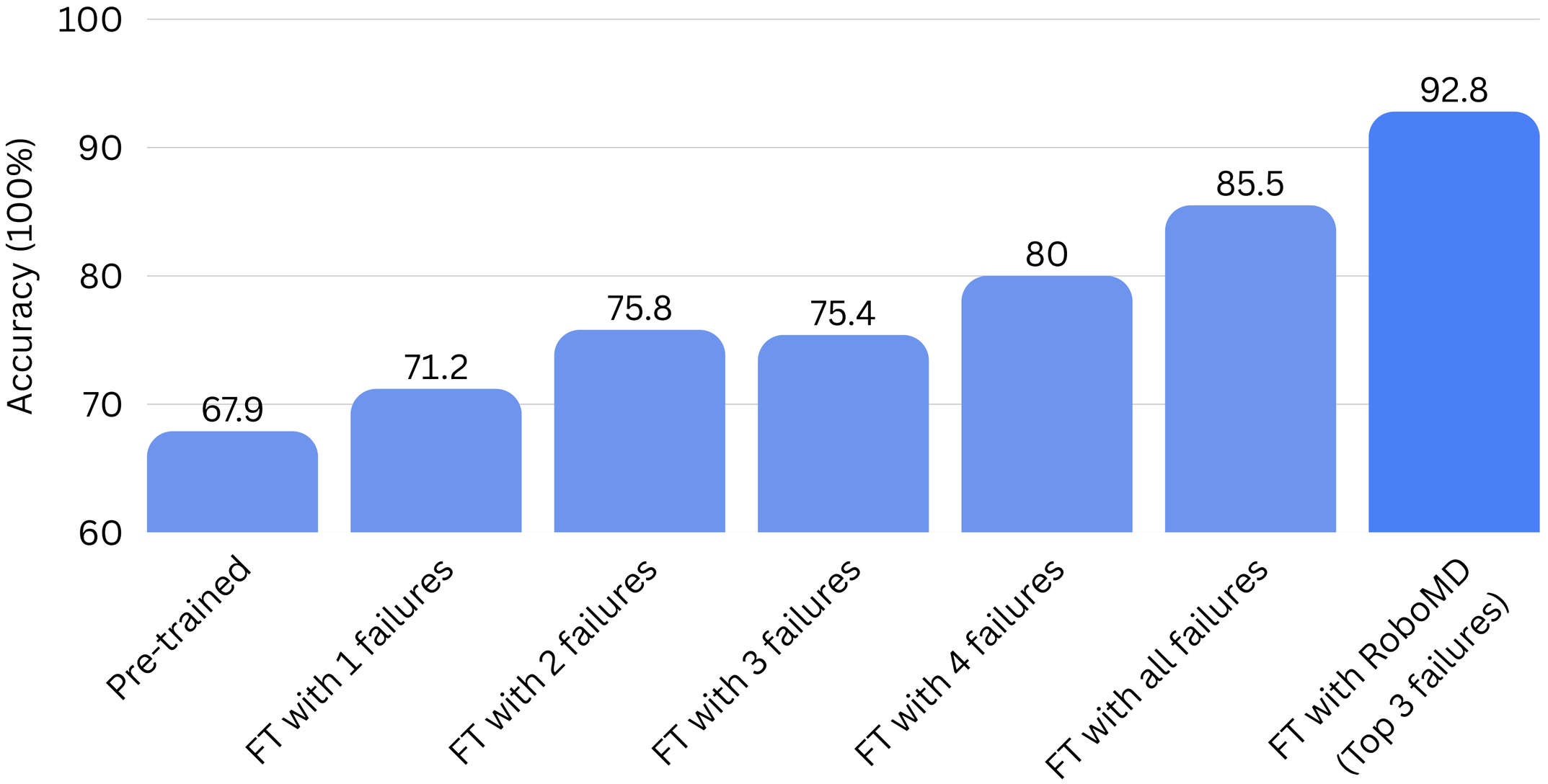
**Trained RoboMD**

**The manipulation policy is more likely to fail with Fanta bottle than water bottle.**



>

# Improve the Manipulation Policy with RoboMD



Paper

Github