REALM: A Real-to-Sim Aligned Benchmark for Generalization in Robotic Manipulation

Martin Sedlacek[1], Pavlo Yefanov[1], Georgy Ponimatkin[1], Simon Pilc[1], Mederic Fourmy[1],
Evangelos Kazakos[1], Cees Snoek[2], Josef Sivic[1], Vladimir Petrik[1]

[1]CIIRC, Czech Technical University [2]VIS Lab, University of Amsterdam

CZECH INSTITUTE OF INFORMATICS ROBOTICS AND CYBERNETICS CTU IN PRAGUE

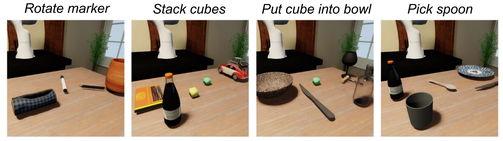
UNIVERSITY OF AMSTERDAM

https://martin-sedlacek.com/realm/

OVERVIEW

Goal

- **Reliable and scalable evaluation** of generalist policies for robot manipulation.
- Understanding failure modes to guide model development, data collection, and evaluations.



Motivation

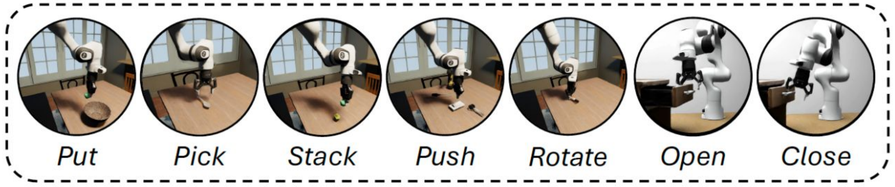
- **Real world performance is the only true benchmark for robotics.**
- Many simulated environments are disconnected from reality and fail to serve as a good proxy.
- Answering *"how does my policy perform on task **X** if **Y** is changed"* without spending hours on a real robot.

Challenges

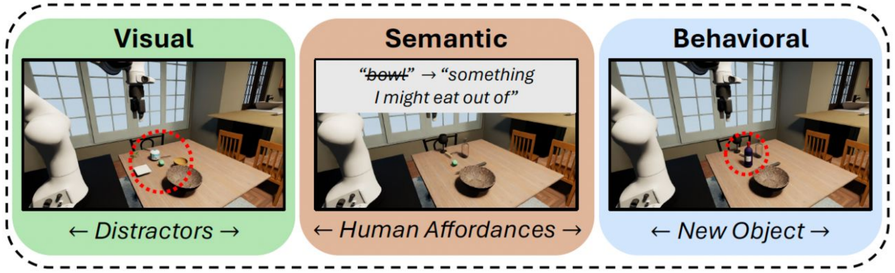
- Realistic robot control and protocol for large-scale that reflects real-world VLA performance.

Contributions

- A real-to-sim aligned simulation for DROID.
- **REALM:** A generalization benchmark supporting **15 perturbations** from **3 categories.**
- Results for 3 state-of-the-art VLA models.



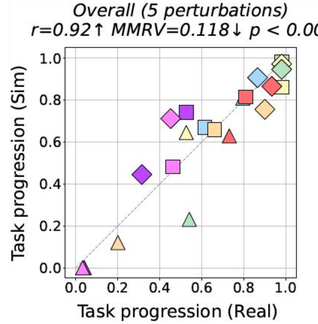
7 manipulation skills



15 perturbations | 3 categories

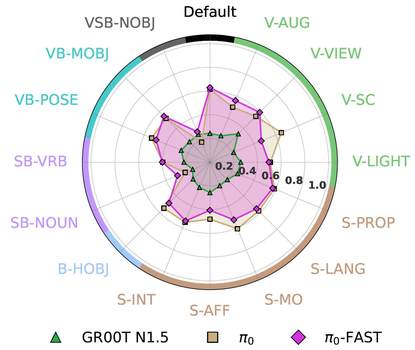
real-to-sim aligned

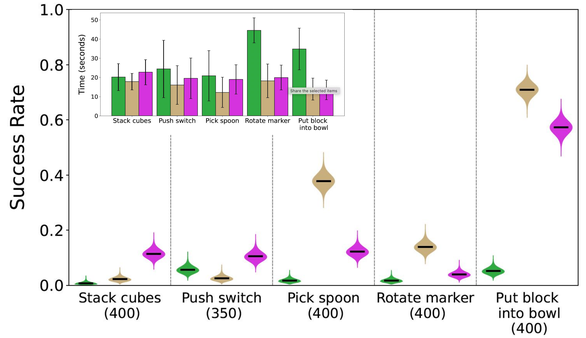
REAL-TO-SIM



- High correlation between sim and real under diverse perturbations / tasks.
- Realistic robot control and high-fidelity simulated visuals.

RESULTS





TAKEAWAYS

- High-fidelity simulation with aligned robot control provides a valuable proxy for real-world performance, helping mitigate the issue of saturated and disconnected simulation benchmarks.
- Despite VLM backbones pre-trained on Internet-scale data, there is a noticeable drop in performance under most perturbations.
- But, performance for most models does not go to zero, indicating they do possess some generalization capabilities.
- Behavioral generalization across skills and objects is still the most challenging, even for cases that are over-represented in the data.