Benchmarking Foundation Models for Robotic Deformable Object Manipulation

Author	Andrea Ritossa - ritossa@kth.se
Supervisor	Alberta Longhini, Marco Moletta, ?
Timeline	Summer 2025
Location	Robotics, Perception and Learning Lab, KTH

1 Introduction

Since the publication of RT-1 Brohan et al. (2022), the robotics research is undergoing a prolific period focused around foundation models. The central promise of this paradigm is the development of generalist policies that aim to achieve general-purpose embodied intelligence, capable of complex interactions within unstructured environments. This shift mirrors advancements in Machine Learning, particularly in Natural Language Processing and Computer Vision, where large pretrained models have demonstrated remarkable success.

Robotics foundational models, often referred to as Vision-Language-Action (VLA) models, are the policies emerging from this paradigm shift. This nomenclature reflects their modality: vision and language serve as inputs while actions serve as outputs (Figure 1). Leveraging techniques such as Imitation Learning and Reinforcement Learning, these models have demonstrated remarkable capabilities in dexterous manipulation across multiple tasks, offering a degree of generalization over robot embodiments and environments. The key enabler of this progress is the usage of a pretrained vision language model for initialization of the model (or part of the model), which then exploits a vast and varied robotics dataset to train the Vision Language Action Model.

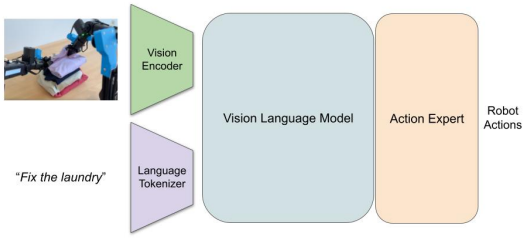

Figure 1: VLA abstraction

2 Goals and Objectives

Due to the rapid pace of development in robotics foundation models, the field currently lacks established, comprehensive comparison benchmarks, particularly for complex interaction scenarios like deformable object manipulation (DOM). This absence makes rigorous comparison between foundational models challenging, while simultaneously presenting an opportunity for standardization. Although some models like π_0 Driess et al. (2024) have demonstrated capabilities in specific DOM tasks (Shirt/Towel/Laundry Folding), a standardized evaluation suite is still missing. Such a suite is essential for equitable evaluation and deeper understanding of model capabilities across different manipulation scenarios.

2.1 Overall Goal

The primary objective of this project is to develop DOM-Bench: a standardized benchmarking suite tailored for evaluating robotics foundation models on deformable object manipulation tasks. This will provide a coherent and equitable basis for comparing model capabilities and limitations.

This project aims to deliver:

- DOM-Bench Suite: A set of deformable object manipulation tasks, environments, and metrics designed to evaluate robotics foundation models.
- Evaluation Protocol and Implementation: Specify the procedure for applying the benchmark suite to different models, e.g: specifications for data efficiency testing and implement the protocol on selected models.
- Paper: Summarization of findings, and dissemination through a research paper.

2.2 Specific Objectives

2.2.1 Literature Review

The first task is to conduct a comprehensive review covering: existing robotics benchmarks, evaluation methodologies, and taxonomies, with a focus on deformable object manipulation and metrics used; Robotics foundation models (VLAs), detailing architectures, training data/recipes, and claimed capabilities relevant to manipulation.

2.2.2 Benchmark Design (DOM-Bench)

Following our review, we will propose DOM-Bench, a benchmark suite and evaluation protocol designed for evaluating robotics foundational models over Deformable Object Manipulation (DOM) tasks. The primary goal of DOM-Bench is to offer deeper insights into a model's capabilities, moving beyond simple task success rates. To achieve this, DOM-Bench will employ a multi-tier task structure. Tasks are organiyzed into tiers of increasing complexity, with each tier designed to evaluate distinct capabilities using specific, measurable metrics. (A detailed brainstorming of potential tasks can be found in Appendix A).

Primitive DOM manipulation *(Minimum Requirement)* The first result is to gather performance over a series of simple tasks (Appendix A.1) whose aim is to evaluate the core deformable object manipulation abilities of the models.

> **Example:**
> Picking a task for each type of deformable object (rope 1D, fabric 2D and liquid 3D) from the DaXBench simulator - {WhipRope, FoldTShirt, PourSoup}. This test will cover a single environment, and a single embodiment (since the DaXBench provides end effector control), therefore we will need to generate forward and inverse kinematics for coherence with the output of the models.

Evaluating Data Efficiency *(Minimum Requirement)* Beyond task performance, a key aspect of DOM-Bench is the measurement of data efficiency. We aim to quantify how model performance changes based on the amount of fine-tuning data provided. Specifically, we plan to record performance under different conditions: zero-shot (no fine-tuning examples), fine-tuned with 10 examples, fine-tuned with 30 examples.

> **Example:**
> Consider evaluating a Vision-Language-Action (VLA) model on a set of tasks: {HangClothes, StoreInCloset, PourRice}. Where HangClothes and StoreInCloset are defined in GarmentLab (a simulated environment, with a specific embodiment), while PourRice will be executed in the RPL lab (a real environment, where we can test different embodiment, e.g., SO-100 and +1 from RPL). To measure data efficiency comprehensively, we need to aggregate performance across these variations. For a single VLA model, the total number of tests required would be calculated based on the variations per task: 1 (HangClothes) + 1 (StoreInCloset) + 2 (PourRice, one for embodiment) = 4 variations. And finally each variation will be tested under 3 different data conditions { 0-shot, 10 demos, 30 demos } resulting in 12 tests.

This approach allows for measuring generalization across tasks, environments (simulation and real), and robot embodiments, similar to the methodology in Bommasani et al. (2023).

Further Tasks Expansion *(Stretch Goal)* Include more tasks for previous evaluation metrics.

Long-Horizon Planning Tasks *(Stretch Goal)* Design specific long-horizon tasks (picking from Appendix A.3) to quantitatively assess the effective context length a model can utilize for successful task completion with DOM. This involves evaluating performance degradation as task complexity and required memory increase.

Instruction Following Evaluation *(Stretch Goal)* Develop tasks that require precise instruction following to evaluate the model's ability to understand and execute complex language-conditioned tasks. This includes tasks with specific spatial, relational, and attribute-based instructions.

Vocabulary Generalization *(Stretch Goal)* Create tasks that test the model's ability to generalize to new vocabulary, including novel objects and actions not seen during training. This involves evaluating the model's robustness and flexibility in handling diverse and complex instructions.

Explainable AI Integration *(Stretch Goal)* Incorporate methods (e.g., representational probing Zhang et al. (2023), attention analysis) to gain insights into the models' internal representations and decision-making processes during DOM tasks. This aims to move beyond black-box evaluation towards understanding *why* models succeed or fail.

2.2.3 Evaluation & Analysis

Finally we will implement the DOM-Bench evaluation protocol across selected foundation models. We will initially focus on a core set of publicly available models, π_0 Driess et al. (2024), OpenVLA Zhou et al. (2023), and Gr00t Liu et al. (2023), and if time allows, we will expand the evaluation to include additional models, OpenVLA-OFT Schmidt et al. (2023) and Octo Brohan et al. (2023). This comprehensive evaluation will establish baseline performance metrics for future research and highlight specific areas where current foundation models excel or require improvement.

Stretch Goals: Evaluation expansion can be achieved by adding additional models such as: OpenVLA-OFT Schmidt et al. (2023) and Octo Brohan et al. (2023).

References

Rishi Bommasani, Percy Liang, and Tony Zhao. Foundation models for decision making: Problems, methods, and opportunities. *arXiv preprint arXiv:2303.04129*, 2023.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

Anthony Brohan, Jeff Abramson, Yevgen Chebotar, Jackie Gu, Daniel Hsu, Julian Kuehne, Wenhe Li, Ishan Misra, Edward Muldrew, Matthias Plappert, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2310.08864*, 2023.

Danny Driess, Zhiao Huang, Yunzhu Liu, Yi Chen, Peter Stone, Yuke Li, and Chiyu Mike Liu. Robotic learning from pixels with parametric implicit functions. *arXiv preprint arXiv:2401.05072*, 2024.

Fei Liu, Zhutian Yoo, Feng Tian, Yuandong Wang, Weich Wang, Bowen Dong, Jeffrey Mahler, Ken Goldberg, et al. Gr00t: A foundation model for robotic manipulation. *arXiv preprint arXiv:2311.01378*, 2023.

Yuvraj Schmidt, Siddhant Jandial, Kaiwen Fischer, Wei-Chiu Ma, Kiana Ehsani, Antonio Torralba, and Shimon Liu. Open-vocabulary 3d visual grounding. *arXiv preprint arXiv:2312.10008*, 2023.

Yuhui Zhang, Fangyun Luo, Hongyang Lee, Tong He, Yan Huang, Anita Shu, Bryan Lu, Jiangfan Han, Jiatao Gu, et al. Probing vision-language models for object and visual feature understanding. *arXiv preprint arXiv:2305.16926*, 2023.

Junqing Zhou, Lechen Li, Xuweiyi Zhang, Yi Cai, Tara Boroushaki, Brian Cheung, Antonio Torralba, Nima Fazeli, and Alberto Rodriguez. Openvla: Open-vocabulary language grounding in 3d scene. *arXiv preprint arXiv:2312.08886*, 2023.

A DOM-Bench Task Suite Details

This appendix provides a brainstorming of tasks within the DOM-Bench suite. TODO: in depth decision of Sim Environments and Real Environments. - GarmentLab: pros comprehensive for garments cons: almost only garments. - DaXBench: pros: covers 1D, 2D and 3D objects cons: sim-to-real gap, no embodiments.

A.1 Fundamentals

Tier 1 focuses on evaluating core shape control and manipulation primitives, serving as the foundation for more complex interactions. Tasks are easily accessible in simulation environments such as SoftGym and DaXBench, allowing for controlled testing.

Task Name	**Goal**	**Sim / Real**	**Description**
Single Cloth Fold	Minimum	Sim	Fold a square cloth along a central axis. *Targets: Shape control of a 2D Object.*
Rope Straighten	Minimum	Sim	Make a curved rope straight. *Targets: Shape control of a 1D object.*
Liquid Pour	Minimum	Sim	Pour a target volume between containers. *Targets: Control of a viscous object.*
Sponge Compression	Stretch	Real	Compress a block to target dimensions. *Targets: Shape control of a 3D object*

Table 1: Tier 1 Fundamental Tasks

A.2 Complex Interactions

Tier 2 advances to multi-step sequences, tool use, constrained motion, and basic object interactions, representing intermediate complexity in deformable object manipulation. The tasks available in simulation are present in the Garment Lab suite while real-world settings will be assessed in the RPL Laboratory.

Task Name	**Goal**	**Sim / Real**	**Description**
Hang Clothes	Minimum	Sim	Place a cloth item onto a hanger. *Targets: Multi-step manipulation, object interaction.*
Wash in Sink	Minimum	Sim	Simulate wiping/scrubbing an object in a sink area. *Targets: Tool use (implicit sponge), coverage.*
Knot Tying	Minimum	Real	Tie an overhand knot in a rope. *Targets: Topological manipulation.*
Object Wrapping	Stretch	Real	Wrap a simple rigid object (e.g., box) with paper/cloth. *Targets: Multi-step manipulation, surface following.*
WashWithSponge	Stretch	Real	Use a sponge to wipe a designated area. *Targets: Tool use, coverage. Metrics: Success rate, area coverage percentage.*
Spreading Viscous Material	Stretch	Real	Spread simulated paste/dough on a surface with a tool. *Targets: Tool use, distribution control.*
Knot Untying	Stretch	Real	Untie a specific simple knot. *Targets: Inverse manipulation, state recognition.*
Surgical Suturing	Stretch	Real	Pass a needle/thread through marked points on phantom skin. *Targets: Fine manipulation, precision, tool use.*
Store In Closet	Stretch	Sim	Fold a cloth item and place it in a designated shelf area. *Targets: Long-horizon (implicit), combined skills.*

Table 2: Tier 2 Complex Interaction Tasks

A.3 Foundation Model Evaluation

Tier 3 specifically evaluates capabilities unique to foundation models, including language understanding, reasoning, planning, generalization, and adaptation. These tasks are primarily conducted in real-world settings at the RPL Lab to fill the missing literature of tasks within DOM simulators ad hoc for foundational models.

Task Name	**Type**	**Description**
Language-Conditioned Tasks		
Complex Placement	Stretch	"Place the folded blue napkin **directly on top of** the placemat, **aligned with the bottom edge**." *Targets: Language (spatial, relational, attributes).*
Rice Pouring	Stretch	"Pour **exactly 50 grams** of the rice **slowly** into the **tallest** container." *Targets: Language (quantity, parameters, attributes).*
Ordered Arrangement	Stretch	"Arrange the **three different colored** ropes **parallel** to each other, ordered **red, green, blue** from left to right." *Targets: Language (multi-object, spatial, ordering).*
Goal-Oriented Constraint	Stretch	"Fold the t-shirt **so the logo on the front is hidden**." *Targets: Language (goal constraint), reasoning.*
Long-Horizon Planning Tasks		
Laundry Prep	Stretch	"Prepare the laundry: **separate the whites from the colors** (use distinct cloth types), then **fold the towels** and **place the shirts** in the basket." *Targets: Planning, sequencing, multi-skill execution.*
Packing	Stretch	"**Fold two shirts, one pair of pants, and three pairs of socks**, then **arrange them neatly** in the small suitcase." *Targets: Planning, multi-object folding, spatial arrangement.*
Bandage Prep	Stretch	"**Unroll the gauze**, **cut a 15cm strip** (robot indicates cut point, human assists/simulates cut if needed), and **fold it into a square pad**." *Targets: Planning, tool interaction (simulated/assisted), sequencing.*
Vocabulary Generalization Tasks		
Open-Vocabulary Object	Stretch	"Fold the **power cable**" (vs. trained ropes/cloth). *Targets: Object generalization.*
Open-Vocabulary Action/Shape	Stretch	"Arrange the rope to form the **letter 'Q'**." or "Use the sponge to **wipe** the surface clean." *Targets: Action/shape generalization.*
Instruction Generalization	Stretch	"**Tidy up** the pile of clothes." (vs. explicit fold/stack commands). *Targets: Instruction robustness.*
Creative/Complex Shaping	Stretch	"Tie the two ropes together using a **bow knot**." or "Shape the play-doh into a **recognizable animal**." *Targets: Advanced generalization, complex manipulation.*

Table 3: Tier 3 Foundation Model Evaluation Tasks