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, is the name of the policies emerging from this paradigm shift. This nomenclature reflects their modality: vision and language for inputs while action as outputs (Figure 1). Leveraging techniques such as Imitation Learning and Reinforcement Learning, these models 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), and subsequently exploit a vast and variegated robotics dataset to train the Vision Language Action Model.

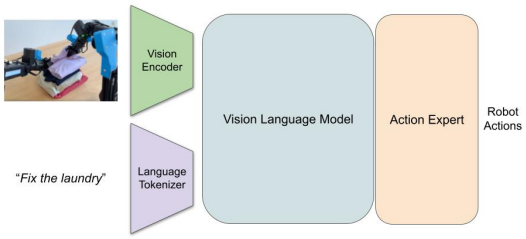

Figure 1: VLA abstraction

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 makes rigorous comparison between proposed models (e.g., π_0, $Gr00t_{N1}$, $OpenVLA$) challenging. While some models like π_0 have demonstrated specific DOM tasks (Shirt/Towel/Laundry Folding), a standardized suite is needed for equitable evaluation and deeper understanding.

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.

2.1 Overall Goal

This project aims to deliver:

- **DOM-Bench Suite:** A well-defined set of deformable object manipulation tasks, environments, and metrics designed to evaluate robotics foundation models.

- **Evaluation Protocol:** A clear procedure for applying the benchmark suite to different models, including specifications for data efficiency testing.
- **Comparative Analysis & Paper:** Implementation of the protocol on selected models, summarization of findings, and dissemination through a research paper.

2.2 Specific Objectives

To achieve the overall goal, the following specific objectives are defined:

2.2.1 Literature Review

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)

Propose a standardized benchmark suite (DOM-Bench) and evaluation protocol. The design will focus on providing multidimensional insights into model capabilities beyond simple task success rates.

- **Multi-Tier Task Structure:** Define tasks across tiers of increasing complexity, each with their own measurable metrics, assessing different capabilities:
  - *Tier 1 (Fundamentals):* Focus on evaluating core shape control and manipulation primitives using standardized objects and refined metrics.
  - *Tier 2 (Complex Interactions):* Focus on evaluating performance on tasks requiring more sophisticated interaction, multi-step execution, or tool use.
  - *Tier 3 (Foundation Model Evaluation):* Focus specifically on evaluating language understanding, reasoning, planning, and generalization capabilities inherent to foundation models when applied to DOM.
- **Generalization Assessment Jaquier et al. (2023):** Systematically evaluate model generalization across key axes:
  - *Environments:* Define evaluation in both simulation and real-world settings (RPL lab).
  - *Tasks:* Utilize the multi-tier structure described above. (See Appendix A for task details).
  - *Embodiments:* Define target robot embodiments for evaluation (Minimum: SO-100; Stretch: Others available at RPL). The protocol should allow for adaptation to different embodiments where feasible.
- **Data Efficiency Evaluation:** Quantify model performance relative to the amount of task-specific fine-tuning data.
  - *Zero-Shot Performance:* Evaluate capabilities directly using the pre-trained foundation model (0 examples).
  - *Few-Shot Performance:* Evaluate after fine-tuning on a small, standardized number of demonstrations (e.g., 10 examples, 30 examples).

Stretch Goals:

- **Context Horizon Quantification:** Design specific long-horizon tasks (potentially extensions of Tier 3 tasks or dedicated scenarios) to quantitatively assess the effective context length a model can utilize for successful task completion. This involves evaluating performance degradation as task complexity and required memory increase.
- **Explainable AI Integration:** Incorporate methods (e.g., representational probing Lu et al. (2025), 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.
- **Further Tasks Expansion:** Include highly complex or nuanced tasks within the tiers (marked in Appendix A) that push the limits of current model capabilities (e.g., intricate knotting, delicate material handling, creative construction).

Protocol Definition: Define a precise experimental protocol specifying setup, initial state randomization, number of trials, permissible interventions (if any), and data logging requirements. Reference Appendix A for detailed task specifications.

2.2.3 Evaluation & Analysis

Implement the defined evaluation protocol using the DOM-Bench suite. Therefore, the first step is to apply the protocol to a core set of publicly available foundation models:

- π_0 Black et al. (2025)
- OpenVLA Kim et al. (2024)
- Gr00t N1 Bjorck et al. (2025)

And a posteriori, analyze the results based on the defined metrics, focusing on comparative performance across models, tiers, generalization axes, and data efficiency levels.

Stretch Goals: Evaluation expansion can be achieved by adding additional models:

- OpenVLA-OFT Moo Jin Kim (2025)
- Octo Ghosh et al. (2024)

Or by conducting a deeper analysis following the extension of the benchmark.

References

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, et al. Gr00t n1: An open foundation model for generalist humanoid robots. https://arxiv.org/abs/2503.14734, 2025.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, et al. $\pi 0$: A vision-language-action flow model for general robot control. https://physicalintelligence.company/blog/pi0, 2025.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, et al. Rt-1: Robotics transformer for real-world control at scale. https://robotics-transformer1.github.io, 2022.

Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, et al. Octo: An open-source generalist robot policy. https://arxiv.org/abs/2405.12213, 2024. Project website: https://octo-models.github.io.

Noémie Jaquier, Michael C. Welle, Andrej Gams, Kunpeng Yao, Bernardo Fichera, Aude Billard, Danica Kragic, Aleš Ude, and Tamim Asfour. Transfer learning in robotics: An upcoming breakthrough? a review of promises and challenges. https://arxiv.org/html/2311.18044v2/, 2023.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, et al. OpenVLA: An Open-Source Vision-Language-Action Model. https://openvla.github.io/, 2024.

Hong Lu, Hengxu Li, Prithviraj Singh Shahani, Stephanie Herbers, and Matthias Scheutz. Probing a vision-language-action model for symbolic states and integration into a cognitive architecture. https://arxiv.org/abs/2502.04558, 2025.

Percy Liang Moo Jin Kim, Chelsea Finn. Fine-tuning vision-language-action models: Optimizing speed and success. https://arxiv.org/abs/2502.19645, 2025.

A DOM-Bench Task Suite Details

This appendix provides specific examples of tasks within the DOM-Bench suite, categorized by tier. Each task contributes to evaluating the capabilities outlined in Section 2.2 (Benchmark Design).

A.1 Fundamentals

Tier 1 focuses on evaluating core shape control and manipulation primitives, serving as the foundation for more complex interactions. Tasks are primarily conducted in simulation environments such as SoftGym and DaXBench, allowing for consistent and controlled testing. Evaluation centers on basic manipulation success metrics, shape accuracy measurements, and efficiency of execution.

Task Name	**Type**	**Description**
Cloth Fold (Single)	Minimum	Fold a square cloth along a central axis. *Targets: Shape control. Metrics: Fold alignment error, final shape IoU.*
Cloth Spread	Minimum	Unfold a crumpled cloth to maximize coverage. *Targets: Shape control. Metrics: Coverage area (%), wrinkle score.*
Rope Straighten	Minimum	Make a curved rope straight. *Targets: Shape control. Metrics: Final end-to-end distance / rope length ratio, curvature.*
Liquid Pour (Simulated)	Minimum	Pour a target volume between containers. *Targets: State control. Metrics: Volume accuracy, spillage amount.*
Elastic Stretching (Simulated)	Minimum	Stretch a band to a target length. *Targets: Shape control under tension. Metrics: Target length accuracy.*
Sponge Compression (Simulated)	Stretch	Compress a block to target dimensions. *Targets: Shape control under force. Metrics: Dimension accuracy.*

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. These tasks are evaluated in both simulation (Garment Lab) and real-world settings (RPL Lab) to assess transferability. Evaluation emphasizes overall task completion, planning capabilities, interaction quality, and adherence to physical constraints.

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 test true embodied intelligence. Evaluation measures instruction following fidelity, planning success across extended sequences, generalization capabilities to novel scenarios, and resolution of ambiguous instructions.

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

Table 2: Tier 2 Complex Interaction Tasks

Task Name	Type	Description
Language-Conditioned Tasks		
Complex Placement	Minimum	"Place the folded blue napkin **directly on top of** the placemat, **aligned with the bottom edge**." *Targets: Language (spatial, relational, attributes). Metrics: Constraint satisfaction accuracy.*
Precise Action	Minimum	"Pour **exactly 50ml** of the liquid (use rice/beans for real-world safety) **slowly** into the **tallest** container." *Targets: Language (quantity, parameters, attributes). Metrics: Parameter accuracy (volume, qualitative rate), target selection correctness.*
Ordered Arrangement	Minimum	"Arrange the **three different colored** ropes **parallel** to each other, ordered **red, green, blue** from left to right." *Targets: Language (multi-object, spatial, ordering). Metrics: Arrangement accuracy (parallelism, order).*
Goal-Oriented Constraint	Minimum	"Fold the t-shirt **so the logo on the front is hidden**." *Targets: Language (goal constraint), reasoning. Metrics: Goal satisfaction success rate.*
Long-Horizon Planning Tasks		
Laundry Prep	Minimum	"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. Metrics: Sub-goal completion, overall success rate.*
Packing	Minimum	"**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. Metrics: Success rate, packing density (qualitative), item integrity.*
Bandage Prep	Minimum	"**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. Metrics: Sub-goal completion, final state accuracy.*
Generalization Tasks		
Open-Vocabulary Object	Minimum	"Fold the **power cable**" (vs. trained ropes/cloth). *Targets: Object generalization. Metrics: Success rate on unseen object categories.*
Open-Vocabulary Action/Shape	Minimum	"Arrange the rope to form the **letter 'Q'**." or "Use the sponge to **wipe** the surface clean." *Targets: Action/shape generalization. Metrics: Novel task success rate, qualitative shape/action match.*
Instruction Generalization	Minimum	"**Tidy up** the pile of clothes." (vs. explicit fold/stack commands). *Targets: Instruction robustness. Metrics: Success rate with paraphrased/abstract instructions.*
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. Metrics: Qualitative assessment, knot/shape correctness.*
Ambiguity Handling	Stretch	Give ambiguous command like "Prepare the rope." *Targets: Reasoning, potential for clarification query. Metrics: Action plausibility, query generation rate (if applicable).*
Error Recovery (Prompted)	Stretch	Induce failure (e.g., drop object) and instruct: "You dropped it, pick it up and continue." *Targets: Adaptation, planning repair. Metrics: Recovery success rate.*

Table 3: Tier 3 Foundation Model Evaluation Tasks