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 A (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 inputs, vision and language, and outputs, action. Leveraging techniques such as Imitation Learning, Reinforcement Learning, and Self-Supervised Learning, these models have demonstrated remarkable capabilities in dexterous manipulation across multiple tasks, offering a degree of generalization over robot embodiments and environments. A key enabler of this progress is the rise of large-scale pre-training which utilize vast datasets to construct general-purpose representations, applicable across diverse robotic settings, tasks, and embodiments Noémie Jaquier (2023).

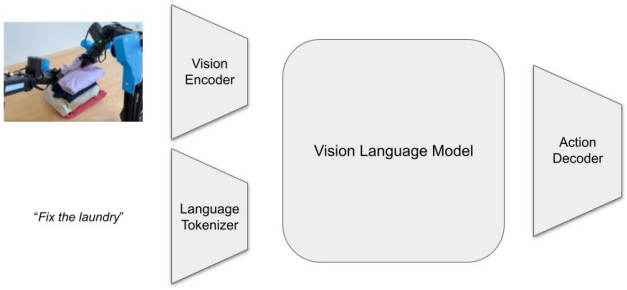

Figure 1: VLA abstraction

2 Goals and Objectives

Due to their rapid success, the field lacks a standardized benchmarking literature, a critical component for ensuring coherent and equitable expansion. While some work, such as the π_0 model Physical Intelligence (2025), showcase tasks involving deformable object manipulation (e.g., laundry folding), a unified approach for comparing models remains absent. Current efforts often focus on testing single models rather than comparing multiple foundational models to elucidate their differences. This proposal aims to address this gap by developing a Deformable Object Manipulation Benchmarking suite for Robotics Foundation Models, fostering a deeper understanding of their capabilities and limitations. Specifically, we will empirically evaluate how the choices made when building a model (architectural, training recipe, tasks covered, etc.) affect performance on the model capabilities. For example, we will compare the Action Decoder: Diffusion Transformer NVIDIA (2025) vs flow matching Physical Intelligence (2025) vs decoding directly from the pretrained vision language model B (2024).

2.1 Overall Goal

This project aims to deliver:

1. Benchmark suite to evaluate robotics foundational models against deformable object manipulation
2. Evaluation protocol, applied to the selected policies
3. Summarization of the work in a paper

2.2 Specific Objectives:

Literature Review Conduct a comprehensive review, including but not limited to:

- Existing robotics benchmarks, evaluation methodologies, and taxonomies focusing on deformable object manipulation.
- Robotics foundational models, emphasizing architectures, training data, and intended capabilities.

Benchmark Design Propose a standardized benchmark suite that integrates and/or extends existing ones and an evaluation protocol. This includes defining clear metrics to provide multidimensional insights of the models' abilities beyond mere performance, such as:

- *Generalization*: Formalize the variations across environments (ex: 1. RPL Laboratory 2. Simulation (GarmentLab, DaXBench, ...), tasks (ex: Cloth Folding, Rope knotting, Sponge ..., Liquid Pouring), and robot embodiments (..., SO-100). (FORMALIZE BETTER?)
- *Data efficiency*: Specify how to measure performance across scaling the number of samples the policy consumes: zero-shot, few-shot performance. (FORMALIZE BETTER?)
- (Stretch Goal) *Context horizon*: Expand the type of tasks taken into account to quantitatively investigate the horizon each model is able to sustain to solve a task: short-term versus long-term task performance. (FORMALIZE BETTER?)
- (Stretch Goal) *Explainable AI*: usage of explainable AI methods such as probing C (2025) to gain measures of understanding from the internal feature representations of the model. (FORMALIZE BETTER?)

Evaluation Implement the defined evaluation protocol and apply it to a selected set of publicly available foundation models, following the benchmark's recipe. A candidate list of the foundational models to test is:

- π_0, Physical Intelligence (2025)
- OpenVLA, B (2024)
- (Stretch Goal) Gr00t NVIDIA (2025)
- (Stretch Goal) Octo D (2024)

Significance: This work addresses the current lack of standardized benchmarking for robotics foundation models, particularly for the challenging deformable object manipulation tasks. By providing a common evaluation framework, this benchmark aims at facilitating comparisons and guide future research towards the development of more capable robotic systems.

References

A. Rt-1: Robotics transformer for real-world control at scale. `robotics-transformer1.github.io`, 2022.

B. OpenVLA: An Open-Source Vision-Language-Action Model. `https://openvla.github.io/`, 2024.

C. Probing a vision-language-action model for symbolic states and integration into a cognitive architecture. `https://arxiv.org/abs/2502.04558`, 2025.

D. Octo: An open-source generalist robot policy. `https://arxiv.org/abs/2405.12213`, 2024.

Michael C. Welte Noémie Jaquier. Transfer learning in robotics: An upcoming breakthrough? a review of promises and challenges. `https://arxiv.org/html/2311.18044v2/`, 2023.

NVIDIA. Gr00t n1: An open foundation model for generalist humanoid robots. https://arxiv.org/abs/2503.14734, 2025.

Physical Intelligence. $\pi 0$: A vision-language-action flow model for general robot control. https://www.physicalintelligence.ai/pi0, 2025.