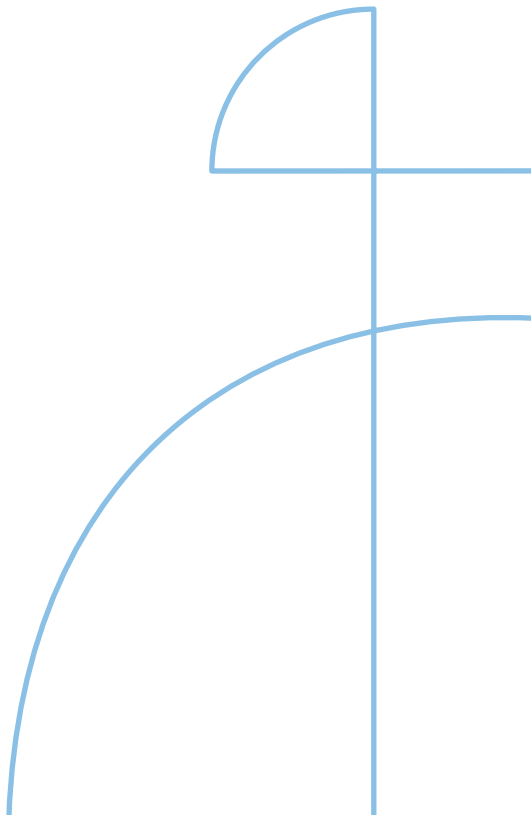
Doctoral Thesis in Computer Science

Robotic Manipulation of Deformable Objects across Object Variations and Human Feedback

MARCO MOLETTA

KTH ROYAL INSTITUTE OF TECHNOLOGY



Robotic Manipulation of Deformable Objects across Object Variations and Human Feedback

MARCO MOLETTA

Academic Dissertation which, with due permission of the KTH Royal Institute of Technology, is submitted for public defence for the Degree of Doctor of Philosophy on Thursday, the 26th November 2026, at 09:00 in Kollegiesalen, Brinellvägen 8, Stockholm, Sweden.

Doctoral Thesis in Computer Science
KTH Royal Institute of Technology
Stockholm, Sweden 2026

AI-Acknowledgment: AI-Acknowledgment: This thesis was produced in part with the aid of the generative AI tools [Claude Opus 5.5]. The tools were to improve language and readability. This included identifying repetitions, suggesting rephrasings, and checking consistency in the text. Generative AI was also used for minor editing of some figures. The author has critically reviewed and revised all generated material and bears full responsibility for the content of the thesis.

© Marco Moletta

ISBN 978-91-8106-749-1
TRITA-EECS-AVL-2026:95

Printed by: Universitetsservice US-AB, Sweden 2026

A Roberta e Carlo.

Abstract

Robots operating in human environments must handle objects whose appearance, geometry, physical behavior, and desired mode of manipulation vary across tasks and users. This is especially challenging for deformable objects, such as garments and textiles, whose behavior depends on contact, material response, deformation, and task intent. This thesis investigates methods for improving the generalization, robustness, and personalization of robotic manipulation systems for deformable objects. The thesis addresses this problem through three complementary directions. First, it studies object-centric representations, including skeleton-graphs and keypoints, as compact descriptions of task-relevant structure. These representations support the transfer of human-specified folding plans across garment instances and improve the generalization of imitation learning policies. Second, it considers mechanical-property variation, studying interaction-derived representations that capture object-specific mechanical context and support prediction across different deformation behaviors. Third, it investigates human feedback as a mechanism for refining policies after offline training, improving robustness through corrective interventions and personalization through feedback on preferred manipulation behavior. Overall, the results show that robust deformable object manipulation requires different mechanisms for different sources of variation. Structured representations help address visible object changes, interaction reveals mechanical properties that affect action outcomes, and human feedback provides information about failures and preferences that task success alone does not capture. The thesis contributes toward manipulation systems that can operate more reliably across changing objects, dynamics, and user-specific requirements.

Sammanfattning

Robotar som verkar i mänskliga miljöer måste kunna hantera objekt vars utseende, geometri, fysiska beteende och önskade manipulationssätt varierar mellan uppgifter och användare. Detta är särskilt utmanande för deformerbara objekt, såsom plagg och textilier, vars beteende beror på kontakt, materialrespons, deformation och uppgiftsintention. Denna avhandling undersöker metoder för att förbättra generalisering, robusthet och personalisering i robotsystem för manipulation av deformerbara objekt. Avhandlingen behandlar detta problem genom tre kompletterande riktningar. För det första studeras objektcentrerade representationer, inklusive skelettgrafer och nyckelpunkter, som kompakta beskrivningar av uppgiftsrelevant struktur. Dessa representationer stödjer överföring av människospecificerade vikningsplaner mellan olika plagginstanser och förbättrar generaliseringen hos imitationsinlärda policyer. För det andra behandlas variation i mekaniska egenskaper genom interaktionsbaserade representationer som fångar objektspecifik mekanisk kontext och stödjer prediktion över olika deformationsbeteenden. För det tredje undersöks mänsklig återkoppling som en mekanism för att förfina policyer efter offline-träning, genom att förbättra robusthet med korrigerande interventioner och personalisering genom återkoppling på föredraget manipulationsbeteende. Sammantaget visar resultaten att robust manipulation av deformerbara objekt kräver olika mekanismer för olika variationskällor. Strukturerade representationer bidrar till att hantera synliga objektförändringar, interaktion avslöjar mekaniska egenskaper som påverkar resultatet av handlingar, och mänsklig återkoppling tillför information om misslyckanden och preferenser som enbart uppgiftens framgång inte fångar. Avhandlingen bidrar därmed till manipulationssystem som kan fungera mer tillförlitligt över föränderliga objekt, dynamik och användarspecifika krav.

Acknowledgements

This PhD thesis marks the end of a long academic journey and of my formal education. These years have deepened my interest in science and understanding of it, but taught me even more about myself through the people I met along the way. I have always tried to learn something from everyone I meet, believing that each person has something worth teaching, and those below are the ones I learned the most from. Thank you for shaping this journey and making it so special.

First and foremost, thank you, Dani, for giving me the opportunity to pursue this PhD and for always letting me explore the research directions I believed in. I am very grateful for the trust you placed in me and for your invaluable advice, which has guided me both in research and beyond the academic world. A close second is Michael. Thank you for supporting me from day one, for always being available to help and to discuss ideas and projects, and for all the whisky tastings. I wish you all the best in your endeavors. Alberta, thank you for being the best collaborator I could have asked for, and even more for being one of my best friends. I hope California brings you a lot of joy. Alfredo, Gustaf, and Giovanni, thank you for the intellectually stimulating discussions and, just as much, for all the jokes in the office. To my other collaborators, Maciej, Thomas, Alex, and Chiara, thank you for the opportunity to work with you; I learned a lot from each of you. Farzaneh, Nona, Aniss, Ciwan, and Zehang, thank you for your immense kindness, for the great conversations, and for always being open to sharing different points of view. Haofei, thank you for being a great office mate and travel companion in Austin. Lakshadeep and Yuchong, thank you for keeping me updated on football and for the fun talks over lunch. To the rest of Dani's group, old and new (Jens, Chao, Zheyu, Katharina, Mohammed, Yufei, Arda, Ziqi, Thoranna, Helena, Yunfan, Jiacheng, Diluka, and everyone else), thank you for making the group such a welcoming place. Thank you also to Marcel for being such a kind (former) office mate, to Loizos, Piotr, and Jonne for the fun and very competitive user studies in the robot lab, and to Ermanno and Donato for the great time spent together at IROS in Pittsburgh. Matteo, thank you for sharing my devotion to SRV, for always being up for nerding out about guitars and music, and for the fantastic days roaming around New England together with Kim and Matilda. More broadly, I would like to thank the whole RPL department, its faculty, Grace, and all the lovely people I met there, for the lunch conversations and the fun and stimulating events.

Adi, Yonk, Seb, and CJ, I could not have asked for a better crew of fellow pirates. Thank you for every sailing adventure so far; the seven seas still await. To the rest of my Stockholm fam, Carin, Thomas, Linn, Josie, Raj, Ann-Marie, Emma, Kristina, Anastasiia, and Sofiya, thank you for all the dinners, BBQs, parties, game nights, and amazing weekends out in the Swedish countryside. To Miguel, Björn, and Marghe, thank you for being the best bandmates on the planet and for all the jam sessions on our road to the Rock & Roll Hall of Fame. Special thanks also go to the musicians and artists whose work inspires everything creative I do, this thesis included. Your music and lyrics, and the unforgettable live performances I have been lucky enough to witness, enrich my soul and will stay with me forever.

Voglio ringraziare anche i famigerati membri de La Cumpa (Cadò, Campa, Fede, Grego, Berti, Minaccia, Sam, Bajo, Bori, Tex, José, Berna e Tinor), la mia seconda famiglia, per essere al mio fianco fin dalla scuola materna, per tutte le prime esperienze vissute insieme e per la voglia di festeggiare come se fossimo ancora diciottenni. A Giulia, grazie per essere come una sorella per me e per sopportare tutti i video di animali scemi che ti mando; ricordati dell'anguria che devi ancora a me e a Berti. A tutti gli amici degli anni di Trento: Luca e Jack, del leggendario Interno 11, i Bagafioi (Pietro, Matteo, Ponta, Ste, Anti e Signo) e il resto del gruppo San Pio X (Fella, Bea, Veronica, Marta, Allegra, Oscar, Robert e Samuele): grazie per aver reso indimenticabili i miei anni di università e per continuare a ritrovarci da Nardini ogni volta che possiamo. Al gruppo del liceo, in particolare Davidino, Tobia, Mancio, Pego e Ferra: grazie per essere sopravvissuti insieme a me alle ore della Cadore e per i weekend in giro per l'Europa (alcuni memorabili e altri meno) a sturare lavandini. A Olivier, Martin, Dima, Cesare, Massimo, Omar, Davide, Simone, Anna, Giacomo e Camilla: grazie per aver riempito di calore i miei primi anni a Stoccolma, facendomi sentire un po' più vicino a casa. A Matteo, grazie per aver condiviso alcuni degli anni più assurdi delle nostre vite, fondando un'azienda insieme, tra alti e bassi, e passando ventiquattro ore al giorno sotto lo stesso tetto.

Ai nonni Basilio e Teresa, grazie per le vostre storie, che mi accompagnano da quando ero bambino, e per gli anni meravigliosi vissuti insieme sulla neve. A tutti gli zii e i cugini, grazie per essermi sempre stati vicini, anche da lontano.

A Marghe, grazie per il tuo amore e il tuo affetto. Averti al mio fianco rende ogni sfida più semplice ed ogni gioia più grande.

A mamma Roberta, papà Carlo e Mattia: grazie per il vostro amore e il vostro sostegno, per essere fin dal mio primo giorno di vita i miei esempi e il mio porto sicuro; questo lavoro è dedicato a voi.

Infine, a chi ha visto iniziare questo mio percorso ma non ha potuto vederlo concludersi, ai nonni Nino e Maria, agli amici Bigolo e Sime: ovunque io vada, vi porto sempre con me.

With gratitude,
Con affetto,
Marco Moletta
Stockholm, October 2026

Contents

Acknowledgements **v**

Contents **vii**

Acronyms **xi**

I Overview **1**

1 Introduction **3**
1.1 Motivation, Challenges, and Thesis Scope 3
1.2 Aim and Research Questions 6
1.3 Thesis Contributions 7
1.4 Thesis Outline 10

2 Variations in Deformable Objects **13**
2.1 Physical and Mechanical Properties of Deformable Objects 13
2.2 Manipulation Preferences 16
2.3 Summary and Discussion 19

3 Representations for Deformable Object Manipulation **21**
3.1 From Observations to State Representations 21
3.2 Direct Sensory Observations 24
3.3 Object-Centric Representations for Visual and Geometric Structure . 24
3.4 Object-Centric Representations for Dynamics and Deformation Context 27
3.5 Summary and Discussion 30

4 Human Feedback for Personalization and Robustness **33**
4.1 Imitation Learning and Visuomotor Diffusion Policies 34
4.2 Human Feedback in Robot Learning 35
4.3 Summary and Discussion 41

5 Conclusions and Future Work **43**
5.1 Relation to Foundation Models in Robotics 44
5.2 Limitations 46
5.3 Future Work 47

6 Summary of Included Papers **49**

References **55**

II Included Publications **61**

A A Virtual Reality Framework for Human-Robot Collaboration in Cloth Folding **A1**
A.1 Introduction A1
A.2 Related Work A3
A.3 Proposed framework A5
A.4 Experiments A7
A.5 Limitations and Future work A12
A.6 Conclusion A12
References A12

B EDO-Net: Learning Elastic Properties of Deformable Objects from Graph Dynamics **B1**
B.1 Introduction B1
B.2 Related Work B3
B.3 Problem Formulation B4
B.4 Method B5
B.5 Environment and Implementation Details B7
B.6 Experiments B9
B.7 Conclusions B13
B.8 Acknowledgements B13
References B13

C Real-Time Operator Takeover for Visuomotor Diffusion Policy Training **C1**
C.1 Introduction C1
C.2 Background & Related Work C3
C.3 Methodology C5
C.4 Experiments C6
C.5 To Takeover or Not to Takeover C10
C.6 Conclusion C12
References C14

D Preference Aligned Visuomotor Diffusion Policies for Deformable Object Manipulation D1
D.1 Introduction D1
D.2 Related Work D3
D.3 Method D5
D.4 Experiments D10
D.5 Conclusion D15
References D15

E On the Generalization Capabilities, Design Choices and Limitations of Keypoint Imitation Learning E1
E.1 Introduction E2
E.2 Related Work E3
E.3 Method E5
E.4 Experiments E6
E.5 Discussion and Limitations E14
E.6 Conclusion E15
References E16

Acronyms

BC: Behavior Cloning
DOM: Deformable Object Manipulation
GNN: Graph Neural Network
IL: Imitation Learning
IIL: Interactive Imitation Learning
KIL: Keypoint Imitation Learning
OOD: Out-of-Distribution
VLA: Vision-Language-Action
VR: Virtual Reality

Part I

Overview

Chapter 1

Introduction

1.1 Motivation, Challenges, and Thesis Scope

A central ambition in robotics is to build systems that can perceive the world, interact with it physically, and adjust their behavior to the situations they encounter [1]. Recent progress in artificial intelligence has greatly expanded what robots can learn from data, but autonomous physical interaction remains difficult to scale beyond structured settings. Everyday environments contain objects whose appearance, geometry, mechanical response, and desired mode of manipulation vary across tasks, users, and contexts. Robots are increasingly expected to operate outside controlled industrial settings and to assist humans in dynamic, unstructured environments, from private homes to hospitals and warehouses. In such settings, relevant tasks include folding laundry, sorting garments for reuse or recycling, assisting a person during dressing, handling food, manipulating soft tissue during surgery, harvesting fruit, and processing flexible materials in industrial production. Spanning domestic service, healthcare, agriculture, logistics, and manufacturing, these tasks expose a central limitation of current robotic manipulation systems: many objects that humans handle routinely are deformable, and their behavior depends on properties that are difficult to observe, predict, and control [2].

Deformable Object Manipulation (DOM) differs fundamentally from rigid object manipulation. While a rigid object can often be represented compactly by its 6-DoF pose together with fixed geometry and a limited set of physical parameters, this description is insufficient for deformable objects. Their state also depends on the object's shape and deformation, which evolve under robot actions, gravity, environmental contacts, and material-dependent properties such as stiffness, elasticity, damping, and friction [3]. This couples perception, modeling, planning, and control: the robot must estimate a partially observable state, predict how actions will change both geometry and visibility, and act despite sensing ambiguities caused by occlusions, self-contact, and visually similar observations that may correspond to different underlying configurations [4].

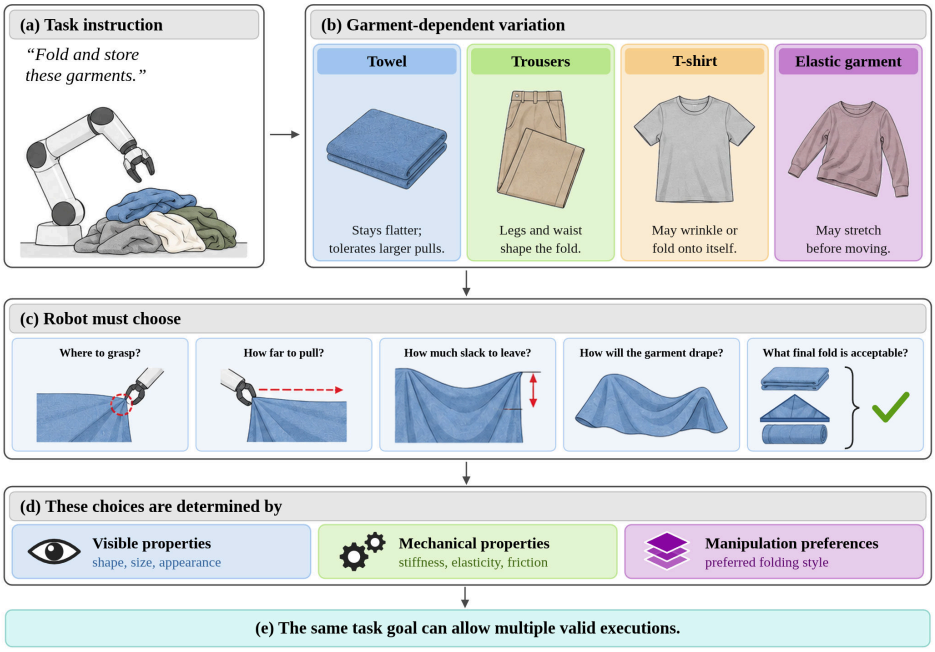


Figure 1.1: Under-specified garment manipulation task. The same instruction can require different manipulation choices depending on the garment type, visible properties, mechanical properties, and user-dependent manipulation preferences.

Beyond the challenges of representing and predicting deformation, the robot's manipulation strategy also influences how effectively it can control a deformable object. Many tasks, including garment folding and rope manipulation, can benefit from dual-arm coordination, allowing the robot to control multiple contact points and guide the object's deformation. However, coordinating the two arms introduces additional constraints on motion planning, grasping, and force distribution, particularly when the object deforms during interaction. These issues are part of the broader problem of cooperative robotic manipulation, which encompasses modeling, control, planning, and learning for coordinated manipulation strategies [5].

Another challenge is variation across objects [6]. Some properties are visible from static observations, such as shape, size, color, texture, and object category. Others are mechanical, such as elasticity, stiffness, friction, damping, and bending resistance, and may only become apparent through interaction. These variations directly affect manipulation outcomes: the same grasp, pull, or place action can produce different deformations depending on the object and its contact with the environment. Practical methods must therefore represent task-relevant object structure while also accounting for properties that are not fully observable before acting.

A further difficulty is that many deformable-object manipulation tasks are under-specified by the task goal alone. A garment is not simply correctly or incorrectly folded: the same object can be folded in multiple valid ways depending on the user's goal, the garment type, the desired storage format, or the preferred manipulation strategy. We refer to these user-dependent choices as manipulation preferences [7]. Consider a robot asked to "fold and store these garments", as depicted in Figure 1.1. The same instruction may apply to a towel, trousers, a thin synthetic shirt, and an elastic garment, but each object requires different manipulation choices. A towel may remain flat under larger pulls, a thin shirt may wrinkle or fold onto itself, trousers impose structure through legs and waist regions, and an elastic garment may stretch before moving as intended. The robot must therefore decide where to grasp, how far to pull, how much slack to leave, how the material will drape, and what final fold is acceptable. These choices depend on visible properties, mechanical properties, and the user's preferred folding style. A robot that only optimizes for task success may therefore produce behavior that is technically valid but misaligned with the user's expectation. This issue is familiar in assistive robotics, where a robot should adapt its behavior to the person it assists rather than forcing the person to adapt to the robot's behavior [8], [9]. The same principle applies to domestic tasks such as garment manipulation: the difficulty is not only to make the robot complete the task, but to make it complete the task in a way that remains reliable across different objects and acceptable to the user.

Recent progress in robot learning has made these challenges more approachable by reducing the need to manually specify every model, controller, and task objective. In Imitation Learning (IL), and in Behavior Cloning (BC) in particular, a robot learns a policy from examples of expert behavior rather than deriving each action from an explicit analytical model [10]. Visuomotor policies extend this idea by mapping sensory observations directly to actions, with recent generative architectures, such as diffusion, enabling expressive action modeling for complex manipulation tasks [11]. This is attractive for DOM because high-fidelity analytical models involve a trade-off between physical accuracy and computational cost, while faster approximations may lose realism and transfer poorly to real objects [12]. More recently, this direction has been scaled further by Vision-Language-Action (VLA) models, which build visuomotor policies on top of pretrained vision-language models and train them on demonstrations pooled across many robots and tasks, aiming for a single generalist policy that can be instructed in natural language and adapted to new tasks with little additional data [13], [14], [15]. Another example is learned world models, i.e. action-conditioned predictive models of how a scene evolves, which are trained on large-scale video datasets and used for planning, data generation, and policy evaluation [16]. Both developments move the expected source of generalization from task-specific structure towards scale. As discussed in Section 5.1, they do not remove all the sources of variation studied in this thesis, but they change the setting in which the proposed mechanisms are most useful.

In parallel, progress in learning deformable-object dynamics has shown that data-driven models can approximate how object state evolves during manipulation

without requiring fully specified analytical models [6], [17]. For example, graph-based representations and Graph Neural Networks (GNNs) can model interactions between local parts of a deformable object and allow to predict how deformations propagate under external actions [18]. Such models make it possible to account, at least implicitly, for physical and mechanical properties that affect manipulation outcomes. However, this progress does not by itself solve the broader generalization problem: methods that work for one object, scene, or task setting may still fail when these properties, the environment, or the desired user-specific strategy change at deployment.

Taken together, these observations motivate the central view of this thesis: DOM is not only a problem of task learning, but of responding to variation. A robot's manipulation policy must remain effective when objects differ in visible properties such as shape, size, and appearance; when hidden mechanical properties such as elasticity or damping change the outcome of its actions; and when the desired execution depends on user-specific manipulation preferences.

1.2 Aim and Research Questions

Building on this framing, this thesis aims to improve the adaptability of robotic manipulation systems for deformable objects, so that a policy remains effective when object appearance, object mechanics, or user preferences differ from those seen during training. It treats DOM as an adaptation problem involving three sources of variation, each addressed through a different mechanism: visual and geometric variation through object-centric representations such as keypoints and graphs; mechanical variation through interactive perception of properties that affect deformation dynamics; and preference variation through human feedback, using corrective interventions and preference-based supervision to improve robustness and personalize manipulation policies. The following research questions structure the thesis around these three sources of variation.

Research Question 1. *How can we improve generalization across visual and geometric variations in deformable object manipulation?*

This question addresses the need for representations that preserve task-relevant object structure while reducing sensitivity to irrelevant visual and geometric variations across different instances, scenes, and object appearances. This is studied in Paper A and Paper E, through the use of sparse, geometric representations and keypoint imitation learning.

Research Question 2. *How can we account for properties that influence the deformation dynamics of deformable objects?*

This question addresses the fact that some mechanical properties relevant for manipulation cannot be reliably inferred from static visual observations alone. Paper B studies how a context of interactions, such as pulling motions, can be encoded

into latent representations that condition a dynamics model and improve prediction across objects with different mechanical properties.

Research Question 3. *How can we improve robustness and personalize deformable object manipulation policies?*

This question addresses how human feedback can improve policies trained only from offline demonstrations, both by correcting failures during deployment and by selecting user-preferred behavior when multiple successful executions are possible. The thesis studies two forms of human feedback: Paper C uses human intervention to correct policy failures, while Paper D accounts for human feedback to align policies toward user-preferred manipulation strategies.

1.3 Thesis Contributions

The contributions are organized around five included papers and three additional publications that provide supporting context:

Paper A. Paper A [19] introduces a Virtual Reality (VR) framework for specifying and executing cloth-folding plans using object-centric keypoint representations, referred to in the paper as skeletons. The framework allows a user to define folding actions through an intuitive virtual interface and transfer them to novel garments of the same category by grounding the actions in the skeleton representation of the object rather than in raw visual observations. This contributes to RQ1 by showing that skeleton representations can support generalization across different instances of the same category. It also partially contributes to RQ3 by providing an example of DOM with human feedback, where the user specifies task intent and supervises automatic execution.

Paper B. Paper B [20] proposes EDO-Net, a graph-based dynamics model conditioned on latent representations of mechanical properties inferred from interaction data. The method uses exploratory interactions, such as pulling, to encode information about the mechanical properties of deformable objects. This latent representation is then used to condition a forward dynamics model, allowing the model to adapt its predictions to objects with different deformation behavior. This contributes to RQ2 by showing that interactive perception can encode properties such as elasticity and stiffness without requiring explicit property labels during training or deployment.

Paper C. Paper C [21] introduces a real-time operator takeover framework for visuomotor diffusion policies. The method allows a human operator to intervene during policy execution when the robot reaches unreliable or Out-of-Distribution (OOD) states. Corrective segments are then extracted from the intervention and used to retrain and improve the BC policy. This contributes to RQ3 by studying

human feedback as a mechanism for robustness, showing how targeted corrections can address policy failures more efficiently than collecting additional full demonstrations of the same type.

Paper D. Paper D [22] proposes a preference alignment method for visuomotor diffusion policies, in a DOM scenario. The method adapts a pretrained policy using novel preferred and non-preferred demonstrations, allowing the robot to shift its behavior towards preferred manipulation strategies. This contributes to RQ3 by addressing personalization in DOM, where multiple task executions may be valid but not equally desirable to the user. The paper treats preferences as a learning signal that complements task demonstrations and helps distinguish preferred and non-preferred behaviors.

Paper E. Paper E [23] studies Keypoint Imitation Learning (KIL) with visual foundation models. The paper evaluates whether keypoint-based object-centric representations can improve generalization compared with raw visual inputs in visuomotor policy learning. It analyzes different keypoint extraction strategies, policy design choices, and limitations of current visual foundation models for DOM. This contributes to RQ1 by studying how compact object-centric representations can support generalization across objects, scenes, and visual variation.

Additional publications. In addition to the included papers, three related publications provide background and context for the thesis. Paper X1 [6] reviews robotic cloth manipulation, with emphasis on generalization, task variation, object property variation, benchmarks, and open challenges. This review provides the broader literature context for the thesis and motivates the focus on adaptation. Paper X2 [24] introduces Elastic Context, showing how elasticity can be encoded from force and displacement measurements and used to condition data-driven dynamics models to improve predictions. Paper X3 [25] studies coarsened graph representations for deformable object dynamics, providing additional support for the graph-based modeling direction used in the thesis.

Overview of the thesis contributions. Figure 1.2 summarizes how the included papers relate to each other and to the three research questions. The papers are organized into two interconnected tracks: *object properties adaptation*, which deals with variation in the object itself, and *human feedback and personalization*, which deals with variation in what the user expects from the robot. Paper A bridges the two: it represents garments as skeleton-graphs and defines folding plans on their nodes, allowing plans to be replicated across garments of the same category (RQ1), while a user specifies task intent and supervises execution through a VR interface (partially RQ3). The object properties track then develops the two components of this representation. Paper B builds on the graph structure, whose connectivity lets GNNs propagate the effect of an action through the object, and augments

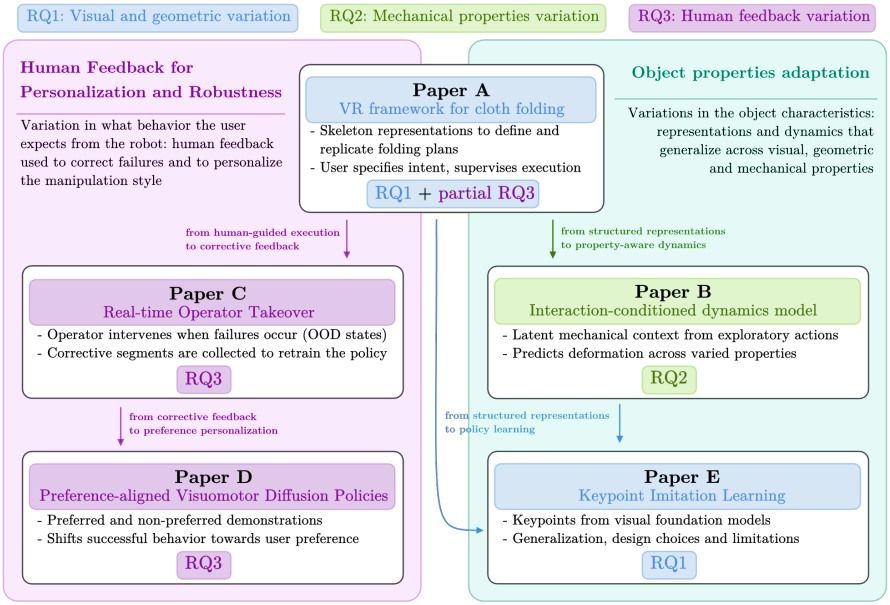


Figure 1.2: Overview of the thesis contributions and their relation to the three research questions.

it with a representation of mechanical properties (RQ2). Paper E instead retains only the nodes as keypoints and replaces fixed folding plans with learned policies, evaluating their generalization to unseen objects and scenes (RQ1). In the human feedback track (RQ3), Paper C moves human intervention from plan specification to real-time correction of a running visuomotor policy, and Paper D extends this from correcting failures to personalizing successful behavior.

List of papers

List of the papers included in the thesis:

- **A: Moletta, M.**, Wozniak, M. K., Welle, M. C., Kragic, D. *A virtual reality framework for human-robot collaboration in cloth folding.* In 2023 IEEE-RAS International Conference on Humanoid Robots (Humanoids). IEEE.
- **B:** Longhini, A.*, **Moletta, M.***, Reichlin, A., Welle, M. C., Held, D., Erickson, Z., Kragic, D. *Edo-net: Learning elastic properties of deformable objects from graph dynamics.* In 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE.
- **C: Moletta, M.**, Welle, M. C., Ingelhag, N., Munkeby, J., Kragic, D. *Real-time operator takeover for visuomotor diffusion policy training.* In 2026 IEEE

International Conference on Automation Science and Engineering (CASE). IEEE.

- **D: Moletta, M.**, Welle, M. C., Kragic, D. *Preference aligned visuomotor diffusion policies for deformable object manipulation.* In 2026 IEEE Robotics and Automation Letters (RAL).

- **E:** Lips, T.*, **Moletta, M.***, Welle, M. C., Kragic, D., wyffels, F. *On the Generalization Capabilities, Design Choices and Limitations of Keypoint Imitation Learning.* In 2026 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE.

***** : authors contributed equally.

Additionally, a list of papers contributed by the author but not included in the thesis:

- **X1:** Longhini, A., Wang, Y., Garcia-Camacho, I., Blanco-Mulero, D., **Moletta, M.**, Welle, M., ... , Kragic, D. *Unfolding the literature: A review of robotic cloth manipulation.* In 2025 Annual Review of Control, Robotics, and Autonomous Systems.

- **X2:** Longhini, A., **Moletta, M.**, Reichlin, A., Welle, M. C., Kravberg, A., Wang, Y., ... , Kragic, D. *Elastic context: Encoding elasticity for data-driven models of textiles.* In 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE.

- **X3:** Marchetti, G. L.*, **Moletta, M.***, Tegnér, G.*, Shi, P., Varava, A., Kravchenko, A., Kragic, D. *Learning Coarsened Dynamic Graph Representations for Deformable Object Manipulation.* In 2021 International Conference on Advanced Robotics (ICAR). IEEE.

- **X4:** Lippi, M., Welle, M. C., **Moletta, M.**, Marino, A., Gasparri, A., Kragic, D. *Visual action planning with multiple heterogeneous agents.* In 2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN). IEEE.

1.4 Thesis Outline

The remainder of this thesis is organized as follows.

Chapter 2 introduces the sources of variation that make DOM difficult to generalize, and frames them as forms of distribution shift between training and deployment. It first separates *physical properties*, which can be at least partially inferred from static observations (e.g., size, shape, color, texture), from *mechanical properties*, which only become apparent through interaction (e.g., elasticity, damping, bending resistance). It then introduces *manipulation preferences* as a third source

of variation, distinguishes them from hard task constraints, and relates different types of preferences to the papers included in this thesis.

Chapter 3 examines how the choice of representation affects generalization in DOM. It first formalizes the relation between the true object state, sensory observations, and the representation used by the robot. It then discusses direct sensory observations (RGB, RGB-D, depth maps, and pointclouds), object-centric representations that expose visual and geometric structure (masks, object crops, skeleton representations, and keypoints), and representations for dynamics and deformation context (graphs, meshes, particles, and latent property contexts inferred through interaction). Throughout, these representations are connected to the thesis contributions: skeleton representations for plan transfer in Paper A, graph dynamics conditioned on latent mechanical context in Paper B, and keypoints from visual foundation models for imitation learning in Paper E.

Chapter 4 shifts the focus from representing the object to learning and adapting the policy that acts on it. It introduces imitation learning and visuomotor diffusion policies, which form the policy-learning framework used in Papers C, D and E. It then discusses human feedback from the perspective of Interactive Imitation Learning (IIL), covering task specification through the VR interface of Paper A, corrective intervention through real-time operator takeover in Paper C, and preference alignment of visuomotor diffusion policies in Paper D. Corrective and preference feedback are presented as complementary mechanisms: the first improves robustness where the policy is unreliable, and the second reshapes successful behavior toward the user's intent.

Chapter 5 concludes the thesis. It revisits the three research questions and summarizes how the included papers address them. It then discusses the relation of the thesis to foundation models in robotics and explains its main limitations. Finally, it outlines directions for future work.

Chapter 6 provides the abstracts of the included papers together with the author's contributions to each.

Part II contains the included publications, Papers A–E.

Chapter 2

Variations in Deformable Objects

Deformable object manipulation is difficult to generalize because the same nominal task can change across several forms of variation. These variations affect different parts of the manipulation pipeline: perception must handle changes in visual structure, dynamics models must account for different physical responses, and policies must adapt when task success alone does not specify the desired behavior. In this thesis, they are connected by the fact that they all create distribution shift between the conditions observed during training and those encountered at deployment [26].

2.1 Physical and Mechanical Properties of Deformable Objects

A central obstacle in DOM is object variation, since even when the task remains unchanged, the objects to be manipulated may differ in appearance, geometry, structure, and material properties. These differences affect the whole manipulation pipeline: perception may fail under changes in color, texture, or lighting; planning may fail when grasp and place locations no longer correspond to the same object parts across instances; and control may fail when the same action produces a different deformation. Generalization in DOM therefore requires more than learning one successful behavior, as it involves identifying which variations matter for the task and designing representations, models, and policies that can both ignore irrelevant variation and adapt to relevant variation [6].

In this thesis, object properties are divided into two broad categories: *physical properties* and *mechanical properties*, as illustrated in Figure 2.1. Physical properties refer to aspects of the object that can often be observed, or at least partially inferred, from static sensory input, such as size, shape, color, surface texture, object category, and visible geometric structure. Mechanical properties instead describe how the object responds to external forces and actions, including, among others, friction, damping, elasticity, bending resistance, and stretching behavior. This distinction is useful because the two types of properties create different adaptation

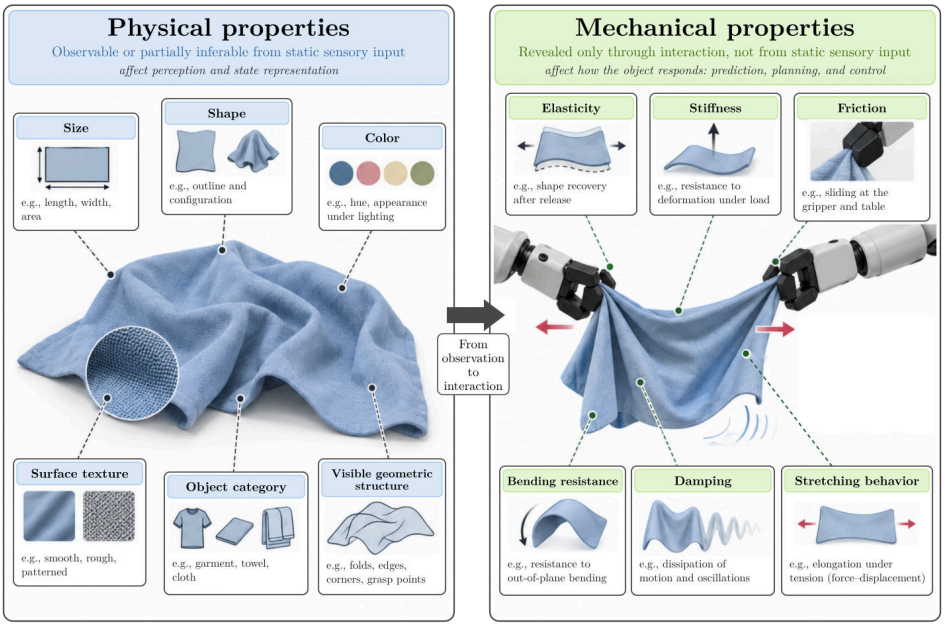


Figure 2.1: Examples of physical and mechanical properties of deformable objects.

problems: physical properties primarily affect perception and state representation, while mechanical properties affect prediction, planning, and control through the object's response to interaction.

Physical properties. Physical properties define the part of object variation that is exposed to perception and often used to specify the task. Shape and size, for instance, help identify the object category and locate task-relevant parts such as grasp points, folding lines, or manipulation targets. Other properties, such as color, surface texture, lighting, and background, may not determine the mechanics of the object, but they still affect perception modules. A policy trained from RGB images can therefore rely on correlations that are not essential to the task, and fail when these visual conditions change even though the physical manipulation problem remains the same [27]. This motivates, for example, domain randomization over camera viewpoint, lighting, object appearance, and texture, with the goal of learning visual policies that are less sensitive to irrelevant visual variation [28], [29].

For deformable objects, visible geometry is especially important because the object state cannot be reduced to a rigid pose. Garments, cloths, ropes, and cables can appear in many configurations, and these configurations constrain which actions are feasible and what information is available to the robot. The representation must therefore preserve the geometric variation needed for manipulation

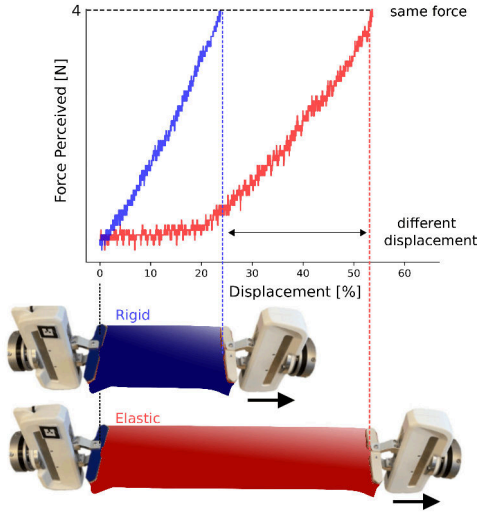


Figure 2.2: Effect of elasticity on the response of textiles to the same interaction. Two textiles are pulled by a dual-arm robot until the same force (4N) is perceived. The stiffer textile (blue) reaches this force at about 24% displacement, while the more elastic one (red) stretches to about 53%. Adapted from [24].

while avoiding unnecessary sensitivity to details that do not support robust action selection. Dense representations, such as RGB images, depth maps, pointclouds, particles, and meshes, preserve rich visual or geometric information, while sparse or object-centric representations, such as graphs and keypoints, reduce dimensionality and expose task-relevant structure. Choosing an appropriate state representation is therefore part of the generalization problem, and involves a trade-off between computational complexity, geometric and physical accuracy, and simplicity [30].

Mechanical properties. Mechanical properties define the part of object variation that is revealed through action: not only what the object looks like, but how it deforms when pushed, pulled, lifted, folded, or released. Two garments may look similar in a camera image but respond differently to the same trajectory. One may stretch, another may resist bending, slide on the table, or retain wrinkles. Figure 2.2 illustrates this for elasticity: when two textiles are pulled until the same force is reached, the more elastic one stretches roughly twice as far as the stiffer one [24]. These differences can change the outcome of an otherwise identical action, so a policy that ignores them may fail to generalize even when perception correctly identifies the object and its visible structure [31].

A single image may show the approximate shape and visible folds of a garment, but it usually does not reveal how the fabric will respond to tension, contact, or motion. The robot may therefore need to perturb the object and observe the resulting

deformation, forces, or motion history. This makes mechanical property estimation an interactive perception problem: information is obtained not only by observing the object, but by interacting with it and interpreting its response [32]. In some of the deformable-manipulation literature, including Paper B, the term *physical properties* is used broadly to cover both visible object properties and *mechanical properties*. In this thesis, the term is reserved for properties observable from static sensory input, while mechanical properties refer to those governing dynamics and deformation. The distinction is not absolute, since material composition affects both appearance and mechanics. It is nonetheless useful: visual and geometric variation motivates object-centric representations, while mechanical variation motivates interactive perception and property-conditioned dynamics. How to combine both in a single policy remains an open challenge, discussed in Chapter 5.

2.2 Manipulation Preferences

Manipulation preferences describe user- or task-dependent choices among multiple valid ways of solving the same manipulation problem. They differ from hard task constraints: a constraint defines what must be satisfied for a behavior to be acceptable, whereas a preference defines which behavior is more desirable when several acceptable behaviors are possible [33]. This follows the planning view in which goals and hard constraints, such as reaching a folded garment state, must hold for any valid plan, while preferences or soft constraints affect the relative quality of alternative feasible plans, such as using a specific folding sequence. In assistive robotics, preferences similarly guide choices between actions that can all achieve the task goal, or modulate aspects such as speed or verbosity without necessarily making non-compliance a task failure [9].

Preferences can appear at several levels of the manipulation process. They may concern the final outcome, intermediate choices, or execution style. This is especially relevant in assistive, domestic, and textile manipulation, where the same nominal task can admit several valid executions depending on the user, context, or intended use. In assisted dressing, the strategy may need to account for a user's mobility, comfort, or safety. In garment folding, one user may prefer folding from left to right, aligning trousers from the waistband, or producing a compact fold for storage, while another may prefer a neater fold, different grasp points, or slower motions that reduce wrinkles and stretching. Preferences should therefore be treated as a source of variation alongside object properties: they change the behavior that the robot should produce, even when the task instruction remains unchanged [7].

Preferences over task execution have received less attention than task success in robotic manipulation. One reason is that many evaluation protocols reduce performance to binary or scalar success measures. These measures are practical for benchmarking, but they hide differences between executions that all satisfy the task goal: a trajectory may succeed while still containing suboptimal actions [34],

or while being inefficient, abrupt, unnecessarily corrective, or misaligned with the user's preferred style. A second reason is that demonstrations encode preferences implicitly. A demonstrator shows not only how to solve the task, but one particular way of solving it. Since several strategies may be valid, this choice can reflect subconscious or user-specific preferences rather than task requirements [35]. Without explicit preference modeling, an IL policy may fail to distinguish between alternatives, collapse them into inconsistent behavior, or reproduce a default strategy that is not preferred by the user.

DOM makes this issue more pronounced. Small differences in grasp point, trajectory, pulling direction, or release timing can lead to different intermediate states and final configurations. In cloth folding, these differences may determine whether the object lies flat, wrinkles, stretches, overlaps correctly, or aligns with the intended fold. At the same time, such qualities are difficult to specify as explicit rules or reward functions. This motivates preference-based learning, where users can provide relative judgments between alternative behaviors or trajectories instead of defining an analytical reward [36]. Such feedback is well suited to properties such as neatness, compactness, symmetry, comfort, or style, which may be easy to recognize in an execution but difficult to formalize.

In this thesis, the goal is not to solve preference learning for DOM in general. Instead, preferences are treated as an important source of variation that becomes especially relevant for deformable objects, where several manipulation strategies can be successful but lead to different outcomes. The corresponding contribution studies preference alignment for visuomotor diffusion policies, using preferred and non-preferred demonstrations to shift an already capable policy toward user-specific manipulation behavior rather than relearning the task from scratch [22].

Types of Manipulation Preferences

Manipulation preferences can be grouped according to how they affect the task. The grouping used here is adapted from the taxonomy of preferences for assistive robotics in [7], which distinguishes between preferences that guide action selection and preferences that configure the execution of a selected action. The same taxonomy also distinguishes primary from modifier preferences, and global from task-specific preferences. Here, these categories are used only to clarify where preferences enter the manipulation problem and how they relate to the thesis contributions, as summarized in Table 2.1.

Decision-making and Configuration preferences. A *decision-making* preference states which valid action, strategy, or sequence of actions the robot should choose (e.g., which fold comes first), while a *configuration* preference states how the selected action should be executed, for example in terms of speed, timing, force, proximity, contact location, or motion style (e.g., pulling the garment gently). In DOM, the two are closely linked: a different grasp, sequence, or execution style can lead to a different deformation and therefore to a different final outcome. In

Table 2.1: Manipulation preference types adapted from [7] and their relation to the thesis papers.

Preference type	**Thesis papers**	**Status in the thesis**
Decision-making	Papers A, D	**Directly addressed**: folding plans, action choices, and preferred strategies.
Configuration	Papers A, D	**Implicitly addressed**: trajectory modification and demonstrated execution style.
Primary / Modifier	Papers A, D	**Primary only**: context dependence is not explicitly modeled.
Global / Task-specific	Papers A, D	**Task-specific only**: preferences are tied to garment-folding tasks.
Safety-related aspects	Papers C, A (partly)	**Related**, but not treated as preferences: robustness and human oversight.

this thesis, decision-making preferences are addressed directly in Papers A and D through folding plans and preferred strategies, while configuration preferences are addressed only implicitly, through trajectory modification and demonstrated execution style (Table 2.1).

Primary / Modifier, Global / Task-specific preferences. A *primary* preference states what the robot should do (e.g., handle the garment gently), while a *modifier* preference attaches a condition to it (e.g., only when the garment is wet). A *global* preference applies across tasks (e.g., always move slowly), while a *task-specific* preference applies only to a given activity or object (e.g., fold shirts sleeves-first), and overrides global ones in case of conflict. These distinctions matter because manipulation preferences are often contextual: the same user may prefer different behaviors depending on the object, task, or intended use. In this thesis, both are only partially covered (Table 2.1). In Papers A and D, preferences are expressed for a specific garment-folding task, so they are task-specific by construction, and any conditions on them are contained in the demonstrations or plans rather than represented as separate modifiers. Safety-related aspects are instead kept separate from preferences, and appear in this thesis as human oversight through plan supervision in Paper A and real-time takeover in Paper C. In physically assistive robotics, safety is not treated as a user-modifiable preference, since the robot should be safe by design [7]. However, users may still express preferences for more conservative behavior within the safe set.

2.3 Summary and Discussion

This chapter identified the main sources of variation that a robot manipulating deformable objects must adapt to:

- **Variation in physical properties**, which affects perception and motivates object-centric representations.
- **Variation in mechanical properties**, which affects how objects respond to actions and motivates interactive perception of properties.
- **Variation in user preferences**, which affects which behavior should be selected when several executions are valid.

These variations do not necessarily change the nominal task, but they change what is required for reliable and acceptable execution. They also motivate the main technical directions of the thesis: skeleton representations, graphs, and keypoints for representing deformable objects; interaction-based models for adapting to mechanical properties; and human feedback for correcting or personalizing manipulation policies. The next chapters develop these directions in more detail: Chapter 3 addresses visual, geometric, and mechanical variation through state representations, and Chapter 4 addresses preferences and execution failures through human feedback.

Chapter 3

Representations for Deformable Object Manipulation

This chapter examines how the choice of object representations affects generalization in DOM. The chapter first introduces the role of state representations and the relation between the true object state, sensory observations, and the internal representation used by a planner, controller, dynamics model, or policy. It then reviews the main representation types used throughout the thesis, separating direct sensory observations, such as RGB, RGB-D, depth maps, and pointclouds, from object-centric representations that explicitly expose task-relevant object structure.

3.1 From Observations to State Representations

Following the state-representation-learning view, a state representation can be defined as a compact and task-relevant description, learned or constructed from sensory observations, that provides the information needed for control [37], [38]. In robotic manipulation, this representation defines the interface between sensing and action. This transformation also determines which scene elements are exposed for decision making, which sources of variation are suppressed, and which properties must instead be inferred through interaction. For rigid objects, pose and geometry often provide a sufficiently compact description [4]. For deformable objects, this simplification is no longer sufficient, as the object state is high-dimensional, partially observable, and continuously modified by contact and robot actions [2]. Moreover, deformation is often the intended task outcome rather than a disturbance, as in folding, dressing, wrapping, or shape control. A useful DOM representation must therefore expose task-relevant shape, contact, and material response, while remaining compact and reliable enough to support learning and control [6].

A common way to frame the problem is to distinguish between the underlying object state, the sensory observation, and the representation used by the robot. Let $x_t \in \mathcal{X}$ denote the true deformable object state at time t, let $o_t \in \mathcal{O}$ denote the

available observation, and let $z_t \in \mathcal{Z}$ denote the internal representation used by the downstream model. Since in DOM the full state x_t is generally high-dimensional and only partially observable, the robot must construct a usable representation from sensory data,

$$z_t = \phi(o_t), \tag{3.1}$$

where $\phi : \mathcal{O} \to \mathcal{Z}$ is the representation map from the observation space to the representation space. Depending on the method, ϕ may be a hand-designed perception pipeline, a geometric fitting procedure, a visual foundation model, or a learned neural encoder. This notation is intentionally broad: in this thesis, z_t may describe a set of 3D keypoints extracted from RGB-D observations, a segmentation mask, a graph constructed from a pointcloud, or a latent context vector inferred from an interaction history.

State estimation can be seen as one specific instance of this observation-to-representation problem. In model-based formulations, the robot must infer an internal object state that is consistent with partial and noisy sensory observations [3]. This is especially difficult for deformable objects because the state space is high-dimensional, observations are incomplete, and different configurations may produce similar visual measurements. The representation therefore determines which aspects of deformation, contact, geometry, and material response are exposed to the downstream model, and which ambiguities remain hidden.

Several representation types are used in DOM and throughout this thesis, as summarized in Figure 3.1. They differ in what they expose to the downstream model: raw sensory information, visible object geometry, relational structure, or hidden object-property context. In the following, the term *object-centric* is used for representations that organize information around the manipulated object rather than around the full sensory scene.

Direct sensory observations include RGB and RGB-D images, depth maps, and pointclouds. Strictly speaking, these are not representations, since they are the raw sensor output rather than an abstraction extracted from it. They are nevertheless included here for completeness, because they are often used directly as inputs to encoders, policies, or dynamics models. Direct observations support flexible end-to-end learning, as in RGB-based visuomotor policies, but may also encourage policies to overfit to visual correlations in the training data.

Dense object-centric visual representations include segmentation masks, bounding boxes, object-centered RGB-D crops, and object pointcloud crops. Compared with direct sensory observations, they retain relatively rich information while making the object more explicit, reducing dependence on background, lighting, and scene clutter.

Sparse object-centric geometric representations include keypoints, landmarks, and other low-dimensional descriptions of object structure. They are useful when the task can be expressed through a small number of semantically meaningful points, curves, or geometric relations, as is often the case in manipulation. For example, a cloth-folding plan may depend on corners, edges, or garment-specific

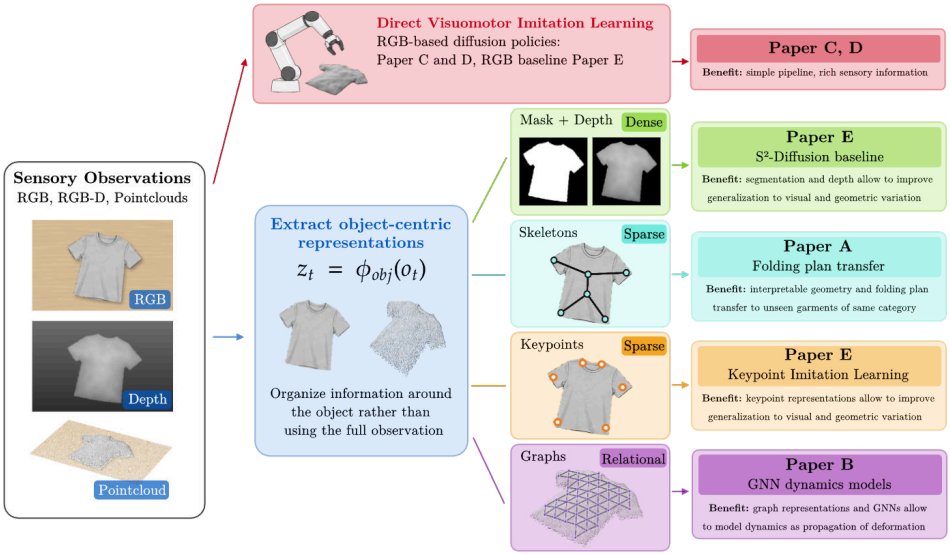


Figure 3.1: Overview of the representations used in this thesis. Sensory observations (RGB, RGB-D, pointclouds) can be used directly as policy input, as in Papers C and D, or mapped to object-centric representations: segmentation and depth (S^2-Diffusion baseline, Paper E), skeletons for folding-plan transfer (Paper A), keypoints for imitation learning (Paper E), and graphs from the object pointcloud for GNN dynamics models (Paper B). Each exposes different information about the object and therefore offers different advantages.

landmarks rather than on every pixel. Sparse representations can improve interpretability and data efficiency, but they depend on reliable extraction and may discard information relevant during execution.

Relational object-centric representations mostly include graphs and meshes. These representations describe the object as a collection of connected or interacting elements, and are therefore especially relevant for dynamics modeling. Their advantage is that they expose interaction structure, but they require estimating and tracking node states, connectivity, and contact relations from noisy and partial observations.

Latent object-property context representations are learned hidden variables that summarize information useful for the downstream model or policy [38]. In this thesis, latent representations are considered object-centric only when they summarize object-level information, such as mechanical response. This is particularly relevant for properties such as elasticity, which may not be visible in a static image but can be inferred from interaction. The resulting latent context can then condition future dynamics prediction or policy execution.

3.2 Direct Sensory Observations

The most direct input to a DOM system is the sensor observation itself. These observations can be passed directly to an encoder without first extracting an explicit intermediate representation. This makes them attractive for end-to-end learning, but also means that the model must learn which parts of the sensory input are relevant for the task and which are incidental [39].

RGB observations are especially common in visuomotor imitation learning because they are simple to collect and preserve rich visual information. Papers C and D use RGB observations as input to diffusion policies, avoiding the need for a handcrafted perception pipeline. The main limitation is that RGB observations entangle task-relevant object structure with nuisance variation. A policy trained directly on RGB may therefore learn visual correlations that hold in the training data but fail under changes in object appearance, scene layout, or camera conditions [28]. This is one of the motivations for the object-centric representations discussed in the following sections.

RGB-D observations and depth maps add metric information about the scene. This is useful because many deformable manipulation actions depend on 3D geometry rather than on image appearance alone. Depth can also be used to lift 2D features, such as keypoints, into 3D robot coordinates. However, depth remains a dense and partial observation. It can be affected by missing measurements, boundary artifacts, reflective or thin materials, self-occlusion, and calibration errors [2].

Pointclouds provide a dense 3D description of the visible object surface. They are often closer to the geometric structure needed for manipulation than RGB images, and can be used directly by policies or converted into more structured forms such as meshes, particles, or graphs [40]. Their limitation is again partial observability: a pointcloud only captures visible surfaces, and it becomes unreliable when the object folds onto itself, is occluded by the robot, or leaves the sensor field of view.

3.3 Object-Centric Representations for Visual and Geometric Structure

Object-centric representations reorganize sensory observations around the manipulated object, or around task-relevant parts of that object, rather than around the full scene [41]. To connect object-centric representations to generalization, it is useful to separate the object instance from the representation used at a given time. An object instance i can be described by a set of properties

$$\omega_i = (p_i, m_i, s_i), \tag{3.2}$$

where p_i denotes visible physical properties such as color, texture, shape, and size, m_i denotes mechanical properties, and s_i denotes semantic or structural information such as object category, garment parts, or task-relevant landmarks. General-

ization in DOM requires maintaining task performance across a distribution of such instances,

$$\omega_i \sim \mathcal{D}_\Omega. \tag{3.3}$$

This is difficult because different components of ω_i affect manipulation in different ways: visible properties influence perception, semantic structure determines where actions should be applied, and mechanical properties determine how the object responds to those actions.

An object-centric representation makes part of this structure explicit by mapping the observation to a representation centered on the manipulated object,

$$z_t = \phi_{\text{obj}}(o_t), \tag{3.4}$$

where $\phi_{\text{obj}} : \mathcal{O} \to \mathcal{Z}_{\text{obj}}$ is an object-centric representation map. The resulting z_t can then be used by a policy,

$$a_{t:t+H} \sim \pi_\theta(\cdot \mid z_t), \tag{3.5}$$

where z_t may be a mask, depth map, object pointcloud, skeleton representation, or keypoint set. Compared with conditioning directly on the full observation, this formulation reduces the burden on the policy: part of the perception problem is handled before policy learning, and the downstream model receives a representation in which the manipulated object is already emphasized [19], [23].

Dense Object-Centric Visual Representations

Dense object-centric representations are useful because they provide a middle ground between raw observations and sparse geometric abstractions, as they preserve more information than sparse representations such as keypoints, but still reduce some of the nuisance variation present in the full scene. Paper E uses this idea in the S^2-Diffusion baseline [27], which conditions a diffusion policy on dense object-centric inputs based on segmentation masks and depth. The comparison with KIL shows that sparse keypoints are not the only useful object-centric representation: dense object-centric observations can provide similar benefits when the object can be segmented and its geometry can be estimated reliably. Their limitation is that they may require larger models and remain sensitive to segmentation errors, missing depth, occlusions, and sensor noise [23].

Sparse Object-Centric Geometric Representations

Compared with dense representations, sparse ones are easier to interpret and can reduce the dimensionality of the policy input. Their limitation is that they compress the object state aggressively and may discard deformation details that matter during execution.

Skeleton Representations The first example of how sparse geometric representations can support DOM is Paper A [19], which uses *skeleton representations* to transfer cloth-folding plans across garment instances with different visible physical properties, such as size, color, and texture. A skeleton representation is a thin, line-like abstraction of the garment's shape that runs along the middle of its silhouette. The skeleton representation is effectively a graph, where nodes are placed at the end points and branching points of the skeleton, and edges connect nodes along the skeleton branches. In Paper A, the representation is used mainly through its nodes, which define pick and place locations for folding. The edges preserve the topological organization of the garment, but are not used for relational reasoning or dynamics modeling here.

The skeleton representation is extracted from an RGB image through a simple perception pipeline. First, the garment is segmented from the background to obtain a binary mask. The mask is then thinned to a one-pixel-wide centerline through *skeletonization* [42], [43], and the resulting skeleton is converted into a graph following [44]. In VR, the user defines a folding plan by selecting pick and place nodes and adjusting the folding trajectory.

Let the skeleton representation of garment instance i be

$$S_i = (V_i, E_i), \tag{3.6}$$

where V_i is the set of nodes and E_i is the set of edges. A single folding action can be represented as

$$u_k = (v_k^{\text{pick}}, v_k^{\text{place}}, \ell_k), \qquad v_k^{\text{pick}}, v_k^{\text{place}} \in V_i, \tag{3.7}$$

where ℓ_k denotes a trajectory parameter, such as the lifting height. A folding plan is then a sequence of such actions,

$$\Pi_i = (u_1, u_2, \ldots, u_M). \tag{3.8}$$

The relevant point is that the plan is defined over the object's geometric structure rather than over raw image coordinates. This allows the same human-specified plan to be reused across garments with different color, texture, scale, and moderate shape variation, as long as the skeleton topology remains compatible. In this sense, the representation provides the generalization mechanism: the human specifies the task in object-centric coordinates, and the robot executes the resulting plan without requiring a previously trained policy. This is particularly useful when demonstration data are scarce, such as at the beginning of data collection, or when an existing policy fails on specific object instances. The limitation is that skeleton representations mainly encode visible geometry. They do not model how a garment stretches, slides, wrinkles, or resists bending during execution, so they address visible physical variation more directly than mechanical variation.

Keypoint Representations A second example of using sparse geometric representations is in Paper E [23], which uses keypoints as sparse object-centric inputs for

IL. In this case, a human specifies three to six semantic keypoints on one reference image for each object category. During training and inference, visual foundation models extract corresponding keypoints in new observations [45]. These 2D keypoints are lifted to 3D using RGB-D data and camera calibration as in [46], encoded by a transformer, and used to condition a diffusion action head.

The keypoint representation is

$$K_t = \{k_{t,1}, \ldots, k_{t,N}\}, \qquad k_{t,i} \in \mathbb{R}^3, \tag{3.9}$$

where each $k_{t,i}$ is the 3D position of a semantic keypoint. The policy can then be written as

$$a_{t:t+H} \sim \pi_\theta(\cdot \mid K_t, q_t), \tag{3.10}$$

where q_t is the robot proprioceptive state.

Keypoints are compact, semantic, and easy to connect to manipulation actions. They identify task-relevant object parts while suppressing irrelevant visual details, making them suitable for IL under object and scene variation. Paper E shows that this object-centric structure improves generalization compared with policies trained directly from RGB observations. The paper also compares different extraction strategies, including image matching, instance matching, and tracking, showing that performance depends strongly on the reliability of the perception module: occlusion, large rotations, self-occlusion, depth artifacts, instance detection errors, and tracking drift can all corrupt the representation extracted from the visual foundation model [23].

3.4 Object-Centric Representations for Dynamics and Deformation Context

The previous section considered object-centric representations that mainly expose visible object structure: masks, depth maps, object crops, skeleton representations, and keypoints. However, DOM also requires reasoning about how the object responds to action. This motivates relational and latent representations. They are still object-centric when they describe the manipulated object, but they also allow to address a different aspect of the problem. In this thesis, relational representations are used to describe how object elements are connected and interact with each other, while latent object-property contexts can be used to describe hidden mechanical information inferred through interaction.

Relational Representations for Dynamics

For deformable objects, describing the current geometry is often not enough, since the robot must also predict how it changes in response to an action. In this chapter, the term *dynamics* refers to this state transition: how the object state changes as a consequence of an action. A dynamics model, or forward model, is then a

model that learns to approximate this transition. This should not be confused with *dynamic manipulation*, such as flinging or shaking a cloth [47], where inertial effects dominate. The work in this thesis mainly concerns quasi-static or low-speed manipulation, in which the object remains close to equilibrium at each step, so that its next configuration is mainly determined by geometry, gravity, contact, and material response rather than by velocity [48]. A generic learned forward model can therefore be written as

$$\hat{z}_{t+1} = f_\theta(z_t, a_t), \tag{3.11}$$

where z_t is the current object representation and a_t is the robot action. For deformable objects, this prediction remains difficult because the same action can produce different deformations depending on the current configuration, contact conditions, and hidden mechanical properties. In practice, the forward transition model in Eq. (3.11) is also the defining object of what is now called a world model [16]; its relation to the contributions of this thesis is discussed in Section 5.1.

A key difficulty is that the effect of an action is not confined to where it is applied. In a deformable object, an action such as a grasp and pull directly moves only the material near the contact point, which in turn displaces the surrounding regions of the object. The global deformation therefore emerges from a chain of local interactions spreading through the object. Relational representations mirror this structure by describing the object as a set of connected or interacting elements, where graphs and meshes can represent specific points of interest on a deformable object as nodes and local relations as edges. By passing information along the connections between neighboring elements, a model can propagate the effect of a local action step by step to distant parts of the object [18], [49]. This makes relational representations a natural basis for modeling and learning the dynamics of deformable objects.

Relational representations are used in physics-based models, data-driven dynamics models, and robotic simulators. Physics-based models rely on analytical or numerical descriptions, such as mass-spring systems, finite element methods, position-based dynamics, particles, or continuum mechanics. These models provide physical interpretability and can generate simulated data, but they can be computationally expensive or difficult to tune to match real textiles [3]. Data-driven models instead learn transition functions from state-action-next-state tuples. They can be efficient at inference time and capture empirical regularities from data, but their generalization depends on the coverage of the training distribution [30].

Paper B follows this data-driven direction by using a graph representation together with a GNN [50] dynamics model. In this case, the object representation is a graph state

$$G_t = (V_t, E_t), \tag{3.12}$$

with V_t and E_t, respectively, sets of nodes and edges: nodes encode object elements, such as 3D points on the deformable object, and edges encode spatial or physical relations between them. This representation is well-suited for learning deformable dynamics because GNNs update each node by aggregating information from its

neighbors. A generic message-passing layer can be written as

$$h_v = \gamma\left(\mathbf{x}_v, \sum_{u \in \mathcal{N}(v)} \psi(\mathbf{x}_v, \mathbf{x}_u, e_{vu})\right), \tag{3.13}$$

where $\mathbf{x}_v$ is the feature vector of node v, e_{vu} is the edge feature between nodes v and u, $\mathcal{N}(v)$ is the neighborhood of v, and γ and ψ are learned update and message functions. Eq. (3.13) implements the propagation described above: each message-passing step moves information one edge further, so stacking n steps allows the effect of an action to reach nodes up to n hops away from the contact point, effectively learning how the deformation propagates along the object. This relational inductive bias has been shown to improve dynamics prediction compared with models that do not explicitly encode object structure and pairwise interactions [18].

Relational representations therefore extend object-centric representations from visible structure to interaction structure. Their main limitation is that this structure must be estimated from sensory data. In real manipulation, reliable graph states, node correspondences, and connectivity are difficult to maintain under occlusion, self-contact, missing depth, and large deformations [3]. Moreover, relational structure alone does not capture the hidden mechanical properties that make the same action produce different deformations on different objects. This is addressed in the next subsection through latent object-property contexts.

Latent Representations for Deformation Context

Latent representations are learned hidden variables, typically produced by a neural network encoder, that compress task-relevant information from observations into a vector or feature space [38]. In this thesis, latent representations are used as object-property contexts to encode mechanical properties. Unlike physical properties, these cannot be inferred from a static observation and require interactive perception, in which the robot observes how the object deforms under force and encodes this response into a context variable [32].

This problem is addressed in Paper B. For an object instance i, the robot performs an exploratory interaction, such as pulling the object, and records it as a sequence of graph states G_k^i and forces F_k^i measured at the robot's wrists:

$$O_i = \big((G_1^i, F_1^i), (G_2^i, F_2^i), \ldots, (G_T^i, F_T^i)\big). \tag{3.14}$$

A single observation is generally not sufficient to identify mechanical properties: elasticity is revealed by how the measured force and the resulting deformation evolve together during the interaction. The adaptation module therefore accumulates evidence over the whole interaction. At each step, an encoder Enc_η summarizes the current observation, and a recurrent unit updates the context:

$$e_k = \mathrm{Enc}_\eta(G_k^i, F_k^i), \qquad c_k = \mathrm{RNN}_\eta(e_k, c_{k-1}), \qquad k = 1, \ldots, T, \tag{3.15}$$

starting from an initial context c_0. The final state $c_i = c_T$ is used as the latent context of the object, so that the adaptation module can be written compactly as

$$c_i = f_\eta(O_i, c_0). \tag{3.16}$$

The recurrent formulation lets the estimate be refined as more of the object's response is observed, and does not require interactions of fixed length. Since mechanical properties are assumed constant for a given object, c_i is inferred once and then kept fixed during prediction. Importantly, no property labels are used: f_η is trained jointly with the dynamics model, so the context encodes whatever information about the object's response improves prediction. The current object state is represented separately as a graph G_t^i, and the graph dynamics model predicts deformation conditioned on both the current graph state and the inferred context:

$$\delta\hat{G}_t^i = g_\theta(G_t^i, a_t, c_i), \tag{3.17}$$

with $\hat{G}_{t+1}^i = G_t^i + \delta\hat{G}_t^i$. With $z_t = G_t^i$, this is an instance of the general form introduced above:

$$\hat{z}_{t+1} = f_\theta(z_t, a_t, c_i). \tag{3.18}$$

This makes the role of the latent representation explicit: the same object state and action can lead to different predicted deformations depending on the mechanical-property context of the object [20].

Latent object-property contexts therefore complement geometric and relational representations. Their limitation is that they usually require informative and structured observations from which the relevant mechanical response can be inferred.

3.5 Summary and Discussion

This chapter shows that representation choice is central to generalization in DOM, since each representation exposes some aspects of the object while compressing others [2], [3], [6]:

- **Direct sensory observations** preserve rich information and are easy to use with visuomotor policies, but include nuisance variation. They are used in Papers C and D.
- **Object-centric visual representations**, dense or sparse, identify where to act, but may discard deformation details and depend on reliable extraction. Paper A uses skeletons to transfer folding plans across variation in physical properties, and Paper E evaluates keypoints and dense object-centric inputs for imitation learning under object and scene variation.
- **Relational representations** model how deformation propagates through the object, but require structured state estimation.

- **Latent object-property contexts** capture how the object responds to interaction, but require informative interaction data. Paper B combines them with graph representations to adapt predictions to mechanical variation.

No single representation is therefore sufficient for all sources of variation. A remaining challenge is to integrate structured representations, inferred mechanical context, and human feedback within the same policy, so that it can adapt to what the object looks like, how it responds to interaction, and how the user wants it to be manipulated. This last aspect motivates the next chapter, which shifts from adaptation through representation to adaptation through human feedback.

Chapter 4

Human Feedback for Personalization and Robustness

The previous chapter examined how representations expose different aspects of deformable objects to a manipulation system. A suitable representation, however, does not by itself determine how the robot should act. This chapter therefore shifts the focus from representations to learning and adapting the policy that acts on those representations. In particular, IL provides a practical way to acquire manipulation skills directly from demonstrations when explicit dynamics models, controllers, or task objectives are difficult to design [51]. Nevertheless, the resulting policies are constrained by the states and behaviors represented in their demonstration data [52]. During deployment, they may encounter poorly covered states and fail to recover from errors, or they may complete the task successfully while following a strategy that does not reflect the user's intent. Human input can therefore play a role beyond the initial demonstration phase, allowing an existing manipulation capability to be specified, corrected, or personalized as new requirements become apparent. In this chapter, this perspective is developed mainly through two complementary forms of adaptation: corrective feedback for improving robustness and preference feedback for aligning successful behavior with user intent.

This chapter first introduces IL and visuomotor diffusion policies as the policy-learning framework underlying Papers C, D and E. It then considers human feedback from an interactive-learning perspective, beginning with interfaces for task specification and feedback, illustrated by the explicit folding-plan specification of Paper A. This is followed by corrective intervention through real-time operator takeover in Paper C and preference-based adaptation of visuomotor policies in Paper D [19], [21], [22].

4.1 Imitation Learning and Visuomotor Diffusion Policies

Having addressed how deformable objects and their relevant properties can be represented, the next question is how to generate robot actions to manipulate them. One possible route is model-based manipulation, where a model predicts the outcome of candidate actions and supports planning or control. This was the focus of the previous chapter in Paper B: the object state is represented as a graph, and the learned forward model is conditioned on a latent representation of mechanical properties inferred through interaction. Such approaches are useful because they make the action-outcome relation explicit and can incorporate physical structure or object-specific context into prediction and decision making [20].

However, reliable dynamics modeling for deformable objects remains challenging and often time-consuming. Accurate prediction depends on identifying relevant physical and mechanical properties, which may require extensive interaction, calibration, or simulation, while model errors can directly degrade planning and control. These difficulties are amplified by the high-dimensional, nonlinear, and contact-rich nature of deformation, and by the fact that many of these properties, such as friction, stiffness, bending resistance, and elasticity, are not directly observable [24]. Model-free policy learning avoids this explicit modeling step, although at the cost of greater data requirements and lower interpretability. IL provides one such complementary route by learning a direct mapping from observations (or representations) to actions from human demonstrations [51]. This can be particularly useful when the task does not require precise reasoning about mechanical properties, or when their influence can be captured sufficiently from the demonstrated sensorimotor behavior.

A common formulation to achieve this is BC, where a policy is trained through supervised learning to reproduce actions observed in expert demonstrations [51]. Let

$$\mathcal{D}_{\text{demo}} = \{(o_t, a_t)\}_{t=1}^{N} \tag{4.1}$$

denote a dataset of demonstrated observation-action pairs, where o_t contains the observations available to the policy, potentially together with the robot proprioceptive state, and a_t the corresponding action. For continuous actions, a standard BC objective is

$$\mathcal{L}_{\text{BC}}(\theta) = \mathbb{E}_{(o_t,a_t)\sim\mathcal{D}_{\text{demo}}}\left[\|a_t - \pi_\theta(o_t)\|^2\right]. \tag{4.2}$$

The observation o_t may contain raw RGB or depth images, but it can equally contain object-centric representations such as the keypoints discussed in Chapter 3. The choice of representation and the problem of learning how to act from it can therefore be treated separately.

Recent IL methods increasingly use generative policies rather than deterministic action regression. A relevant example throughout this thesis is the visuomotor diffusion policy, which models a conditional distribution over short action sequences [11]. Let

$$A_t^0 = (a_t, a_{t+1}, \ldots, a_{t+H}) \tag{4.3}$$

denote a demonstrated action sequence and A_t^k its noisy version at diffusion step k. Given sampled noise ϵ, the model is trained through

$$\mathcal{L}_{\text{diff}}(\theta) = \mathbb{E}_{A_t^0, o_t, k, \epsilon}\left[\left\|\epsilon - \epsilon_\theta\left(o_t, A_t^k, k\right)\right\|^2\right], \tag{4.4}$$

where ϵ_θ predicts the noise added to the action sequence. At inference time, the policy iteratively denoises an initially noisy sequence conditioned on o_t, producing an action chunk A_t. Predicting action sequences improves temporal consistency, while the generative formulation can represent multimodal behaviors in which different grasps, trajectories, or manipulation orders may all be valid. VLA models extend this formulation to large pretrained policies, typically pairing a vision-language backbone with a diffusion or flow-matching action head [14], so the limitations discussed next apply to them as well (see Section 5.1).

More expressive policy architectures, however, do not remove the limitations of finite demonstration data. During autonomous execution, compounding errors can drive the policy into states absent from the demonstrations and eventually lead to failure [52], [53]. Moreover, many deformable manipulation tasks admit several successful strategies that differ in trajectory or final configuration, while a generic demonstration dataset does not specify which one a particular user prefers [9]. These two limitations motivate the corrective and preference feedback mechanisms considered in the remainder of this chapter.

4.2 Human Feedback in Robot Learning

Human feedback provides a way to extend IL beyond a fixed offline demonstration dataset. Rather than requiring all relevant states and behaviors to be anticipated during initial data collection, an initial policy can be learned from demonstrations and subsequently refined using information provided by a human during execution or adaptation. This idea is central to Interactive Imitation Learning (IIL), where a human teacher remains in the learning loop as the robot executes its current policy [54], as illustrated in Figure 4.1. By collecting supervision on states induced by the learner itself, IIL can focus human effort on behaviors poorly represented by the original demonstrations and reduce the distribution mismatch that leads to compounding errors in standard BC [53].

Following the taxonomy of Celemin et al. [54], human feedback can be classified along two dimensions, as summarized in Table 4.1. Feedback in the *evaluative space* assesses the quality of a behavior, either absolutely through human reinforcement or relatively through preferences. Feedback in the *transition (state-action) space* instead provides information about how the behavior should be performed, through corrective demonstrations or relative corrections. These modalities differ in how explicitly the human specifies the desired behavior: corrective demonstrations directly provide actions or trajectories, whereas preferences indicate which behavior is more desirable (for example through pairwise comparisons [55]) without prescribing the corresponding actions.

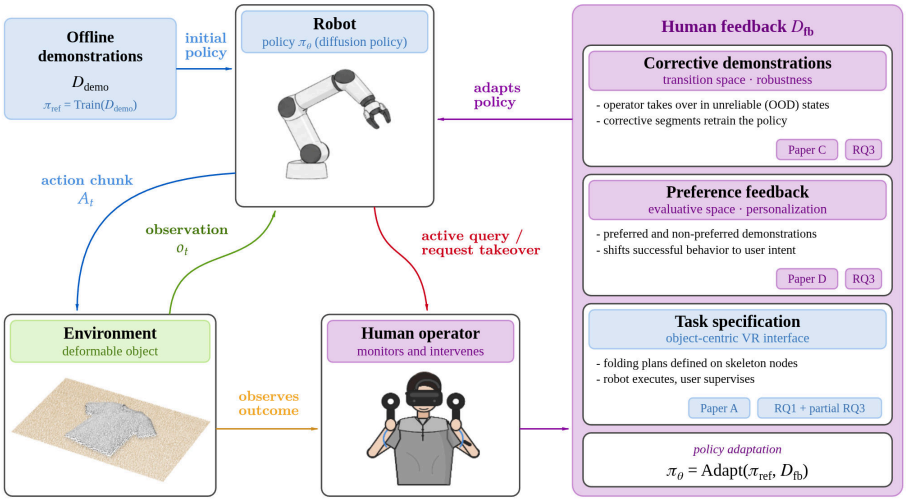


Figure 4.1: IIL loop (adapted from [54]). An initial policy π_{ref} is trained from offline demonstrations $\mathcal{D}_{\text{demo}}$. The robot then acts on the deformable object and observes the resulting states, while a human operator monitors the execution, answers queries, and provides feedback $\mathcal{D}_{\text{fb}}$. This feedback can be corrective demonstrations (Paper C), preferences (Paper D), or task specifications (Paper A), and is used to adapt the policy beyond the initial dataset, $\pi_\theta = \text{Adapt}(\pi_{\text{ref}}, \mathcal{D}_{\text{fb}})$.

Table 4.1: Human feedback modalities in IIL, adapted from the taxonomy of Celemin et al. [54].

	Absolute feedback	**Relative feedback**
Evaluative space	Human reinforcement	Human preferences
Transition space (state–action)	Corrective demonstrations	Relative corrections

The contributions of this thesis span both sides of this taxonomy. Paper C is most closely related to corrective demonstrations in the transition space: when autonomous execution becomes unreliable, the operator provides actions that recover the task and supply supervision in states where the policy lacks suitable behavior. Paper D instead addresses preference feedback in the evaluative space: the objective is not necessarily to correct a failed execution, but to distinguish which of several valid behaviors better reflects the user's intent. The type of feedback is even richer than a conventional pairwise preference, since preferred behavior is communicated through full demonstrations. These trajectories therefore provide both an evaluative signal (by explicitly labeling the behaviors as desired or undesired) and

examples of how that behavior can be executed.

In this thesis, these two forms of feedback are used for two complementary adaptation objectives. *Robustness* concerns missing competence: the policy encounters states or situations for which its original demonstrations provide insufficient guidance. *Personalization* concerns missing intent: the policy may already be capable of completing the task, but does not know which successful strategy the user prefers. Human feedback can therefore either expand the policy's behavior where execution is unreliable or shift an existing capability toward a desired way of performing the task.

A common abstraction for both cases is

$$\begin{aligned} \pi_{\text{ref}} &= \text{Train}\left(\mathcal{D}_{\text{demo}}\right), \\ \mathcal{D}_{\text{fb}} &= \text{CollectFeedback}\left(\pi_{\text{ref}}\right), \\ \pi_{\theta} &= \text{Adapt}\left(\pi_{\text{ref}}, \mathcal{D}_{\text{fb}}\right), \end{aligned} \tag{4.5}$$

where π_{ref} denotes the policy available before feedback-based adaptation and $\mathcal{D}_{\text{fb}}$ the additional human input. In Paper C, π_{ref} is the initial diffusion policy deployed before being refined with corrective takeover demonstrations. In Paper D, it is the pretrained reference policy that anchors adaptation toward preferred behavior.

Interfaces for Task Specification and Feedback

An important practical component of interactive robot learning is the interface through which demonstrations, corrections, and preferences are communicated. Different interfaces expose different levels of control: for example, teleoperation can provide continuous robot trajectories or short corrective segments, while higher-level interfaces can expose task-relevant object structure and allow users to specify actions or preferences without directly controlling every robot motion. Extended-reality (XR) interfaces are particularly relevant for manipulation because they provide a shared three-dimensional interaction space in which robot motion, task information, and virtual object representations can be visualized and manipulated [56], [57]. They have consequently been used extensively for teleoperation and shared control [58], [59], interactive task specification [60], and demonstration collection for cloth manipulation [61], [62].

In this context, Paper A illustrates this at a relatively high level of abstraction [19]. As discussed in Chapter 3, the garment is represented by a skeleton that supports transferring folding actions across garment instances. Here, the same representation also acts as an interface for task specification: users select skeleton nodes as pick-and-place locations, define their sequence, and adjust the trajectory of each folding action, as illustrated in Figure 4.2. The object-centric representation therefore provides a vocabulary for expressing manipulation intent, while geometric correspondence and execution can remain automated. This enables the user to specify and supervise the relevant parts of the task without programming robot trajectories directly. At a lower level of abstraction, in Papers C and D instead

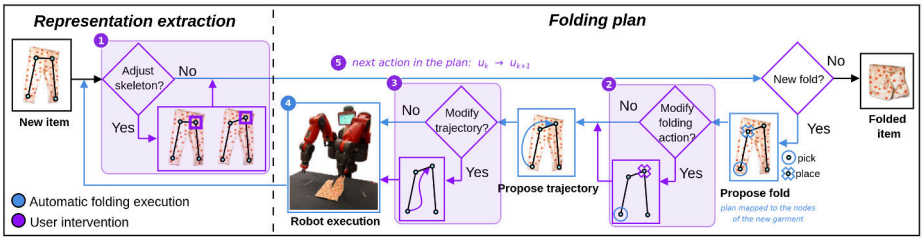


Figure 4.2: Paper A framework for specifying and transferring folding plans through skeleton representations. Blue denotes automatic execution and purple denotes user intervention. When a plan is replicated on a new garment, the user can adjust the skeleton (1), select or correct the pick-and-place nodes of each action (2), and adjust its trajectory (3). The robot then executes the action (4) and proceeds to the next action in the plan (5), repeating until the sequence of folds is complete. The mapping of the plan to the new garment and the robot execution remain automated.

the demonstrations are collected through teleoperation using the Quest2ROS VR interface, which maps Oculus Quest 2 controller velocities to Cartesian end-effector velocity commands and records the resulting 6D actions while the operator holds the trigger [21], [22], [63].

From a personalization perspective, such interaction can be viewed as an explicit form of preference specification. Selecting one folding sequence, set of manipulation points, or trajectory instead of another determines how the garment should be folded, although the preference itself is specified rather than inferred or learned from demonstrations. Importantly, Paper A already shows that even a nominally simple pick-and-place fold can depend on additional continuous parameters: changing the intermediate trajectory height can considerably affect the resulting garment configuration [19]. This illustrates that manipulation intent can be expressed at different levels of articulation, from high-level choices such as which side to fold first to finer execution details. More generally, two folding strategies may use the same pick and place locations while differing in how the cloth is manipulated between them. One strategy may lift and place the garment directly, whereas another may maintain tension during the motion, rotate the grasped part before placement, keep the cloth close to the supporting surface, or smooth local wrinkles before completing the fold. Such differences are difficult to encode with a small set of discrete pick-and-place parameters, yet they may strongly affect the resulting fold and reflect meaningful user preferences. This motivates interfaces capable of capturing preferences at multiple levels of expressiveness, including subtle trajectory- and interaction-level characteristics that cannot be specified through simple task-level actions alone.

Corrective Feedback for Robustness

A central limitation of offline BC is the distribution mismatch between the states represented in the demonstrations and those encountered by the learned policy. Let $d_{\pi^*}(x)$ denote the state-visitation distribution induced by the expert policy (in practice, the human collecting demonstrations) and $d_{\pi_\theta}(x)$ that induced by the learned policy. In general,

$$d_{\pi^*}(x) \neq d_{\pi_\theta}(x), \tag{4.6}$$

because the expert demonstrator generates the training trajectories, whereas the learned policy generates its own states during deployment. Small action errors therefore accumulate and drive the policy into uncovered states (the compounding-error problem, Section 4.1). Hence, the practical objective is to reduce this coverage gap between the two distributions. DAgger addresses this by rolling out the current policy, obtaining expert actions for the encountered states, and aggregating them into the training set, i.e., corrective feedback in the transition space (Table 4.1) [53].

This problem is critical in DOM, where high-dimensional configuration spaces make exhaustive demonstration coverage impractical [2], [3]. It therefore matters where additional supervision is collected, while the policy is relied upon to generalize between the covered regions. Paper C follows this general principle through Real-Time Operator Takeover (RTOT), but its contribution is primarily at the system-integration level rather than a new theoretical formulation of IIL. The framework integrates a live visuomotor diffusion policy with teleoperation so that an operator can seamlessly override the policy when execution approaches an undesirable state and return control immediately after the correction. A ring buffer continuously retains the observations preceding takeover, preserving the context that led to the intervention, while only the subsequent operator-controlled segment is recorded as corrective action data. This allows deployment failures to be converted directly into compact training demonstrations without restarting or repeating the full task.

Using the notation introduced above, if $\mathcal{I}$ denotes the operator-controlled interval and B the number of past observations retained by the buffer, the resulting corrective data can be represented as

$$\mathcal{D}_{\text{corr}} = \{(o_{t-B:t},\, A_t^{\text{corr}})\}_{t\in\mathcal{I}}, \qquad \pi_\theta = \text{Train}(\mathcal{D}_{\text{demo}} \cup \mathcal{D}_{\text{corr}}). \tag{4.7}$$

Rather than recording complete demonstrations that largely repeat already covered behavior, RTOT concentrates supervision near states where the current policy becomes unreliable, reducing operator effort and producing more targeted data. Paper C shows that these takeover demonstrations are substantially shorter while improving performance more than additional initial demonstrations of comparable number. Iterating this process progressively expands the training data toward states encountered by the learned policy.

Corrective takeover therefore addresses missing *competence*: the operator supplies behavior where the policy would otherwise become unreliable. It does not, however, indicate which of several already successful executions is preferable. This

distinction leads to the second form of feedback considered in this chapter, where the objective shifts from robustness under distribution shift to personalization among valid behaviors.

Preference Feedback for Personalization

As noted in Section 4.1, many DOM tasks admit several valid solutions: garments can be folded using different sequences, directions, alignments, or final configurations, while assistive tasks may depend on user-specific strategies. Moreover, as discussed in Chapter 2, success-based evaluations overlook meaningful differences in how a task is executed [64]. Consequently, two behaviors may satisfy the same task objective while only one reflects the user's desired strategy.

A preference can be expressed as a relation between two valid behaviors,

$$\tau^+ \succ \tau^-, \tag{4.8}$$

where τ^+ denotes the preferred and τ^- the less preferred behavior. A policy should then increase the probability of reproducing the preferred strategy without unnecessarily discarding manipulation capabilities already contained in the pretrained policy, thereby keeping the *alignment tax* low [65]. This requirement makes preference adaptation different from simply training a new BC policy on a small set of preferred demonstrations.

Preference learning has traditionally been approached by inferring user preferences from pairwise comparisons between alternative behaviors [66], [67]. In reinforcement-learning-based approaches, these comparisons can be used to learn a reward function, which is subsequently optimized to obtain the desired policy [55], [68]. This provides a flexible mechanism for representing user intent, but introduces separate reward-learning and policy-optimization stages [65]. Direct Preference Optimization (DPO) provides a more direct alternative by avoiding explicit reward-model fitting and directly contrasting preferred and dispreferred behavior while regularizing the adapted policy against a reference model [69]. Conceptually, the objective increases the relative likelihood of τ^+ over τ^- under π_θ, compared with their relative likelihood under π_{ref}. The reference policy therefore acts as an anchor, limiting deviation from previously learned behavior while shifting the policy toward the new preference [65]. Anchoring to the reference policy is especially important when preference data is limited. Let

$$\mathcal{D}_{\text{pref}} = \mathcal{D}_{\text{win}} \cup \mathcal{D}_{\text{lose}}, \tag{4.9}$$

where the *win* dataset contains the demonstrations expressing the preferred behavior, while the *lose* dataset contains the alternative, non-preferred demonstrations. Importantly, a lose demonstration is not necessarily a failed manipulation: it may successfully complete the task while violating the intended preference. Preference optimization can therefore exploit information that ordinary success-based imitation objectives discard.

Paper D [22] applies preference alignment to visuomotor diffusion policies for DOM. Starting from a pretrained reference policy, the method adapts the policy using preferred and non-preferred demonstrations rather than learning a new policy only from the preferred examples. The paper compares several preference-optimization objectives and introduces RKO, which combines binary preference labels with context-dependent weighting so that more informative distinctions between preferred and non-preferred behavior receive greater emphasis. The key aspect is that preference feedback modifies an existing manipulation capability in a targeted and sample-efficient way.

This clarifies the relationship to Paper A. In Paper A, the user explicitly specifies a desired folding strategy through an object-centric representation, and the selected plan is then transferred across garment instances. In Paper D, the desired behavior is instead expressed through demonstrations, and the policy must learn how its action distribution should change to reproduce that preference. The transition is therefore from *explicit task-level preference specification* to *learned policy-level preference alignment*.

4.3 Summary and Discussion

This chapter shows that human feedback can extend IL beyond the initial demonstration dataset, since relevant information often becomes available only after deployment [53], [54]:

- **Task specification** through object-centric interfaces lets the user state the desired strategy explicitly. Paper A uses skeletons as a vocabulary for defining and transferring folding plans.
- **Corrective feedback** addresses missing competence by adding behavior where execution is unreliable. Paper C collects it through real-time operator takeover.
- **Preference feedback** addresses missing intent by reshaping otherwise valid behavior toward the user's strategy. Paper D uses preferred and non-preferred demonstrations to align visuomotor diffusion policies.

Interactive feedback therefore targets the deficiencies that actually emerge instead of requiring exhaustive coverage up front, and its usefulness depends on the interface through which it is provided, which trades off human effort, expressiveness, and the amount of behavior that must be specified explicitly. Remaining challenges are more expressive interfaces for the trajectory- and interaction-level preferences discussed in Section 4.2, and more reliable detection of when intervention is needed. Combining all these mechanisms within a single policy remains open and is discussed in the next chapter.

Chapter 5

Conclusions and Future Work

This thesis investigated how robotic manipulation systems for deformable objects can remain effective when objects, environments, and desired behaviors differ from those present during training. The central argument is that DOM should be considered not only as a problem of learning a task, but also as a problem of *adaptation*. A robot must account for what the object looks like, how it responds to interaction, and how the human wants the task to be performed. These aspects correspond to visual and geometric variation, mechanical variation, and variation in user preferences, and require different but complementary mechanisms. The thesis addressed these challenges through three main directions: object-centric representations for generalization across visual and geometric variations, interactive perception for adapting to mechanical properties, and human feedback for improving policy robustness and personalization. The research questions have been addressed as follows:

Research Question 1. *How can we improve generalization across visual and geometric variations in deformable object manipulation?*

Papers A and E show that object-centric representations can improve generalization by exposing task-relevant object structure while reducing sensitivity to irrelevant visual variation. In Paper A, skeleton representations allow human-defined cloth-folding plans to be transferred across different garment instances [19]. Paper E extends this principle to imitation learning, where keypoints extracted using visual foundation models improve generalization over RGB-based policies, particularly under scene variation [23]. However, the results also show that keypoints do not universally outperform other object-centric representations and that their effectiveness depends on reliable extraction. The main takeaway is therefore not that one representation is universally preferable, but that explicitly exposing relevant object structure can improve generalization when this structure can be robustly perceived.

Research Question 2. *How can we account for properties that influence the deformation dynamics of deformable objects?*

Paper B shows that information obtained through interaction can be used to infer mechanically relevant properties influencing the deformation dynamics of deformable objects, which otherwise would not be observable from static sensory input. EDO-Net learns a latent representation from interaction history and uses it to condition graph-based dynamics prediction [20]. This enables the same dynamics model to adapt its predictions across objects with different elastic behavior without requiring explicit property labels at deployment. Together with the explicit elasticity representation explored in Paper X2, this demonstrates that interaction can serve as a form of perception when the properties relevant for manipulation are revealed only through deformation. The broader takeaway is that mechanical variation cannot simply be made invariant through representation design: when it changes the outcome of an action, the robot must instead infer and account for it.

Research Question 3. *How can we improve robustness and personalize deformable object manipulation policies?*

Papers C and D address different ways in which human feedback can complement autonomous policy learning, in an IIL setting. Paper C uses real-time operator takeover to collect corrective demonstrations specifically in states where the current policy is unreliable [21]. This improves robustness while avoiding the need to anticipate all relevant failures during initial demonstration collection. Paper D instead addresses cases where the policy may successfully perform the task but not according to the user's preferred strategy. Preference-alignment methods adapt a pretrained visuomotor diffusion policy using preferred and non-preferred behavior, enabling more sample-efficient personalization than conventional fine-tuning [22]. Corrective and preference feedback therefore address complementary deficiencies: the former provides information when the policy lacks *competence*, while the latter provides information when it lacks knowledge of the user's *intent*.

Across the three research questions, the main conclusion is that different variations require different forms of adaptation. Object-centric representations can suppress variations that are irrelevant to the task; interaction can reveal hidden variations that affect object dynamics; and human feedback can provide information that cannot be inferred from task success alone. These mechanisms should therefore be seen as complementary components of adaptive DOM rather than independent solutions.

5.1 Relation to Foundation Models in Robotics

In recent years, robot learning has started to shift towards large pretrained policies that are trained on data from many robots and tasks and then adapted to specific ones. VLA models adapt vision-language models to output robot actions and are trained on demonstrations pooled across robots and tasks [15], [70], while

world models learn action-conditioned predictions of future observations, increasingly from large video corpora [16]. A natural question is whether the problems studied in this thesis are solved by scale. The evidence suggests that scale addresses some sources of variation much better than others.

Visual and geometric variation. This is where VLAs have made the clearest progress. Web-scale pretraining and cross-embodiment data provide robustness to appearance, background, and object instance that task-specific RGB policies lack, and recent VLAs have shown substantial progress in DOM in the real world [71]. It is therefore unlikely that keypoint-only policy inputs, as studied in Paper E, will be the primary route to broad visual generalization; Paper E's own finding that keypoints perform on par with, but not better than, dense object-centric inputs is consistent with this. At the same time, VLA research is reintroducing explicit intermediate structure: models that predict depth and trajectory traces before acting and allow users to edit them [72], and preference-alignment methods that use keypoints proposed by a vision-language model [73]. In this view, the skeleton representations of Paper A and the keypoints of Paper E are less a replacement for a pretrained policy than an interface to it: a way to ground, inspect, and steer its behavior, and to adapt it with few demonstrations. DOM also remains data-intensive at scale; a deformable-specific VLA still relied on tens of hours of folding demonstrations to outperform a general VLA baseline [74]. This cost suggests a possible continued role for intermediate representations, which can reduce the number of demonstrations needed to adapt a policy to new objects.

Mechanical variation. This is the source of variation that current foundation models in robotics address least. They are trained on demonstrations that rarely contain exploratory interactions, condition on short observation histories, and are never asked to infer properties such as stiffness or elasticity, while vision-language backbones show limited physical understanding [75]. The principle of Paper B, summarizing an interaction history into a latent context and conditioning prediction on it, addresses a need that scale alone does not remove: since novel objects will always appear at deployment, a robot must be able to infer their properties on the spot rather than rely only on what it saw during training. This principle does not depend on the graph representation and could equally condition a world model or the action expert of a VLA. The graph-based state estimation used in Paper B, on the other hand, is the part that is least likely to scale to the settings in which VLAs operate.

Robustness and personalization through human feedback. Human feedback remains necessary at scale, both to correct failures and to specify preferred behavior. VLAs are still brittle under heavy distribution shift, and improving them from deployment data has become a central part of their training: recent methods combine demonstrations, autonomous rollouts, and online human corrections pro-

vided during autonomous execution [71]. The design of Paper C, namely seamless takeover of a running policy, retention of the context preceding the intervention, and recording of only the corrective segment, applies unchanged to such models. Personalization has received comparatively less attention. VLAs are trained to reproduce behavior from large, multi-operator datasets, and language conditioning specifies what to do more readily than how to do it. Preference-alignment methods for VLAs have shown that policies can be steered towards objectives such as task success, safety, or efficiency [73], [76], and a natural next step is to align them with user-specific manipulation style. This is particularly relevant for deformable objects, where the desired manipulation behavior is hard to describe in words and several successful executions can lead to different outcomes, which is precisely the direction explored in Paper D.

Overall, foundation models in robotics change the default starting point of a manipulation system, from a policy trained for a single task to a pretrained generalist, but they do not remove the three sources of variation studied in this thesis. However broad the training data, deployment will continue to present objects with unseen material properties, situations the policy has not covered, and users with their own preferences. The contributions of this thesis are therefore best read as mechanisms that operate around such models. Since all policy experiments here use task-specific policies, validating these mechanisms on pretrained VLAs is a direct next step (Section 5.3).

5.2 Limitations

The main limitation of the thesis is that these mechanisms are primarily studied separately. The thesis therefore demonstrates complementary components of adaptation rather than a unified system capable of simultaneously adapting to object appearance, mechanics, execution failures, and user preferences.

The representation-based approaches also depend strongly on perception. Skeletons and keypoints reduce input dimensionality and irrelevant visual variation, but failures in keypoint extraction, tracking, segmentation, depth estimation, or object identification directly affect downstream manipulation. Occlusions, self-occlusions, large object rotations, multiple similar objects, and large deformations remain difficult. Sparse representations additionally discard information about local geometry, wrinkles, contact, or tension that may be important for some tasks, while denser representations retain more information at the cost of increased complexity and sensitivity to nuisance variation.

Mechanical adaptation introduces a different set of limitations. Paper B mainly considers elastic variation and assumes that informative interaction histories and structured graph states are available, but the same real deformable objects may vary simultaneously across a spectrum of different mechanical properties. A single exploratory interaction may therefore be insufficient to infer all properties relevant for a task. Interaction also has a cost and may alter the object state, meaning that

the robot must eventually decide not only how to infer mechanical context, but when exploration is necessary and which action is informative.

The human-feedback approaches remain dependent on the availability and quality of human input. Operator takeover requires detecting when intervention is needed, while automatic uncertainty and out-of-distribution detection remain imperfect. Corrective demonstrations may also contain operator-specific motion that should not necessarily be reproduced. Similarly, preferences may depend on the user, object, task, and context and may evolve over time. Strong alignment toward a new preference can also degrade previously learned capabilities, creating a trade-off between personalization and capability retention.

Finally, the experimental evaluations necessarily study a restricted subset of the variations that occur in real deployment. Most experiments isolate particular changes in object properties, scene conditions, policy behavior, or preferences. In practical applications, these variations occur simultaneously. The presented methods should therefore be interpreted as mechanisms for improving adaptability rather than as a complete solution to open-ended DOM. In addition, all policy-learning experiments use task-specific policies trained from limited demonstrations rather than pretrained generalist models; whether the observed gains carry over to large VLAs remains to be tested.

5.3 Future Work

A central future direction is to **combine geometric, mechanical, and human-derived information within the same manipulation policy**. Structured representations such as keypoints could be augmented with latent mechanical context inferred through interaction, allowing the policy to reason both about where relevant object parts are and how they are expected to respond. Integrating such property information directly into visuomotor diffusion policies and VLAs, for example as an additional context token for the action expert, would connect the dynamics-based and model-free approaches investigated in this thesis. Pretrained foundation models offer a natural basis for this integration: rather than training task-specific policies, the mechanisms developed here could be applied as adaptation layers on top of a generalist VLA, and an important first step is to evaluate whether their benefits carry over at that scale. Similarly, conditioning world models on interaction-derived mechanical context, possibly collected online, could bring object-specific dynamics into models that currently learn them mainly from appearance.

Mechanical adaptation should also be extended beyond predefined pulling interactions and primarily elastic variation. Future systems should **learn when and how to explore**, selecting actions such as pulling, sliding, bending, lifting, or pressing according to which mechanical properties remain uncertain. The inferred context should account for multiple properties simultaneously, potentially combining visual, force, and tactile information.

For object-centric representations, future work should improve **robust perception under occlusion, large deformation, orientation changes, and multiple instances**. Rather than choosing exclusively between sparse and dense representations, hybrid approaches could combine semantic keypoints with local depth, segmentation, or pointcloud information, retaining compact task structure while preserving geometric details where they are needed.

Human-feedback methods should become more autonomous and selective. For corrective learning, this requires **better detection of when intervention is necessary** and methods that separate useful recovery information from operator-specific motion. For personalization, policies should support **continual adaptation across users and contexts** while limiting capability loss. More expressive interfaces are also needed to communicate preferences at different levels, ranging from high-level task outcomes to trajectory, timing, contact, and coordination preferences.

These directions also motivate **closing the loop between human feedback and representation learning**. A repeated intervention may indicate not only a deficient policy but also an incorrect representation or insufficient mechanical context. Future systems could therefore use human feedback to identify which component of the manipulation pipeline should be adapted rather than always modifying the final policy.

Finally, evaluation should increasingly consider **combined sources of variation**. An unseen garment may simultaneously have different appearance, geometry, mechanical properties, and a user-specific desired manipulation strategy. Benchmarks that characterize these dimensions explicitly would make it possible to evaluate whether methods genuinely adapt across interacting distribution shifts rather than solving them independently.

The longer-term objective is therefore a manipulation system that does not assume that every relevant condition can be anticipated in the original training set. Instead, it should generalize through structured representations where possible, actively obtain information when object behavior is uncertain, identify when its current capabilities are insufficient, and incorporate human feedback when the desired behavior cannot be inferred autonomously.

Chapter 6

Summary of Included Papers

This chapter contains abstracts of the included papers and contributions by the author of the thesis. The symbol * denotes equal contribution.

Paper A

A virtual reality framework for human-robot collaboration in cloth folding

Marco Moletta, Maciej K. Wozniak, Michael C. Welle, Danica Kragic
In *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*

Abstract: We present a virtual reality (VR) framework to automate the data collection process in cloth folding tasks. The framework uses skeleton representations to help the user define the folding plans for different classes of garments, allowing for replicating the folding on unseen items of the same class. We evaluate the framework in the context of automating garment folding tasks. A quantitative analysis is performed on three classes of garments, demonstrating that the framework reduces the need for intervention by the user. We also compare skeleton representations with RGB images in a classification task on a large dataset of clothing items, motivating the use of the proposed framework for other classes of garments.
Contributions by the author: Designed the method and its implementation, designed and implemented the experiments, and wrote the paper.

Paper B

EDO-Net: Learning Elastic Properties of Deformable Objects from Graph Dynamics

Alberta Longhini*, **Marco Moletta***, Alfredo Reichlin, Michael C. Welle, David Held, Zackory Erickson, Danica Kragic

In *2023 IEEE International Conference on Robotics and Automation (ICRA)*

Abstract: We study the problem of learning graph dynamics of deformable objects that generalizes to unknown physical properties. Our key insight is to leverage a latent representation of elastic physical properties of cloth-like deformable objects that can be extracted, for example, from a pulling interaction. In this paper we propose EDO-Net (Elastic Deformable Object - Net), a model of graph dynamics trained on a large variety of samples with different elastic properties that does not rely on ground-truth labels of the properties. EDO-Net jointly learns an adaptation module, and a forward-dynamics module. The former is responsible for extracting a latent representation of the physical properties of the object, while the latter leverages the latent representation to predict future states of cloth-like objects represented as graphs. We evaluate EDO-Net both in simulation and real world, assessing its capabilities of: 1) generalizing to unknown physical properties, 2) transferring the learned representation to new downstream tasks.

Contributions by the author: Co-designed the methodology and its implementation, carried out most of the simulated experiments, and wrote half of the paper.

Paper C

Real-time operator takeover for visuomotor diffusion policy training

Marco Moletta, Michael C. Welle, Nils Ingelhag, Jesper Munkeby, Danica Kragic
In *2026 IEEE International Conference on Automation Science and Engineering (CASE)*

Abstract: We present a Real-Time Operator Takeover (RTOT) paradigm that enables operators to seamlessly take control of a live visuomotor diffusion policy, guiding the system back to desirable states or providing targeted corrective demonstrations. Within this framework, the operator can intervene to correct the robot's motion, after which control is smoothly returned to the policy until further intervention is needed. We evaluate the takeover framework on three tasks spanning rigid, deformable, and granular objects, and show that incorporating targeted takeover demonstrations significantly improves policy performance compared with training on an equivalent number of initial demonstrations alone. Additionally, we provide an in-depth analysis of the Mahalanobis distance as a signal for automatically identifying undesirable or out-of-distribution states during execution. Supporting materials, including videos of the initial and takeover demonstrations and all experiments, are available on the project website: https://operator-takeover.github.io/
Contributions by the author: Co-designed the methodology and its implementation, carried out most of the experiments, and wrote most of the paper.

Paper D

Preference aligned visuomotor diffusion policies for deformable object manipulation

Marco Moletta, Michael C. Welle, Danica Kragic
In *2026 IEEE Robotics and Automation Letters (RAL)*

Abstract: Humans naturally develop preferences for how manipulation tasks should be performed, which are often subtle, personal, and difficult to articulate. Although it is important for robots to account for these preferences to increase personalization and user satisfaction, they remain largely underexplored in robotic manipulation, particularly in the context of deformable objects like garments and fabrics. In this work, we study how to adapt pretrained visuomotor diffusion policies to reflect preferred behaviors using limited demonstrations. We introduce RKO, a novel preference-alignment method that combines the benefits of two recent frameworks: RPO and KTO. We evaluate RKO against common preference learning frameworks, including these two, as well as a baseline vanilla diffusion policy, on real-world cloth-folding tasks spanning multiple garments and preference settings. We show that preference-aligned policies (particularly RKO) achieve superior performance and sample efficiency compared to standard diffusion policy fine-tuning. These results highlight the importance and feasibility of structured preference learning for scaling personalized robot behavior in complex deformable object manipulation tasks.
Contributions by the author: Designed the method and its implementation, designed and implemented the experiments, and wrote the paper.

Paper E

On the Generalization Capabilities, Design Choices and Limitations of Keypoint Imitation Learning

Thomas Lips*, **Marco Moletta***, Michael C Welle, Danica Kragic, Francis wyffels
In *2026 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*

Abstract: RGB-based imitation learning requires many demonstrations to generalize to unseen objects or scenes, motivating research into intermediate representations to improve generalization for robotic manipulation. Visual foundation models enable one-shot extraction of keypoints to provide such representation. However, it remains unclear how to integrate them into imitation learning optimally and when they outperform alternative representations. We combine approaches from previous works on keypoint imitation learning (KIL) and investigate several design choices to provide practical guidelines. Using over 2000 real-world rollouts, we also assess the generalization capabilities of KIL to unseen objects and scene variations. KIL achieves a 75% overall success rate across five tasks, significantly outperforming the RGB baseline (47%) and performing on par with S2-diffusion (73%). Finally, we explore the limitations of the foundation models used for keypoint extraction and extend KIL to tasks with multiple object instances. Our results confirm KIL as a data-efficient approach for robot learning, though it does not outperform alternative representations and inherits limitations of the foundation models used for keypoint extraction. All rollout videos, demonstrations, and results are available at https://kil-manipulation.github.io/.
Contributions by the author: Co-designed the methodology and its implementation, carried out half of the experiments, and wrote half of the paper.

References

[1] B. Siciliano and O. Khatib, *Springer handbook of robotics*. Springer, 2008.

[2] J. Zhu et al., "Challenges and outlook in robotic manipulation of deformable objects", *IEEE Robotics & Automation Magazine*, vol. 29, no. 3, pp. 67–77, 2022.

[3] H. Yin, A. Varava, and D. Kragic, "Modeling, learning, perception, and control methods for deformable object manipulation", *Science Robotics*, vol. 6, no. 54, eabd8803, 2021.

[4] R. Laezza, "Robot learning for deformable object manipulation tasks", Ph.D. dissertation, Chalmers University of Technology, 2024.

[5] C. Smith et al., "Dual arm manipulation—a survey", *Robotics and Autonomous systems*, vol. 60, no. 10, pp. 1340–1353, 2012.

[6] A. Longhini et al., "Unfolding the literature: A review of robotic cloth manipulation", *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 8, no. 1, pp. 295–322, 2025.

[7] G. Canal, "Adapting robot behavior to user preferences in assistive scenarios", Ph.D. dissertation, Universidad Politécnica de Cataluña, 2020.

[8] T. L. Chen et al., "Robots for humanity: Using assistive robotics to empower people with disabilities", *IEEE Robotics & Automation Magazine*, vol. 20, no. 1, pp. 30–39, 2013.

[9] G. Canal, G. Alenyà, and C. Torras, "Adapting robot task planning to user preferences: An assistive shoe dressing example", *Autonomous Robots*, vol. 43, no. 6, pp. 1343–1356, 2019.

[10] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration", *Robotics and autonomous systems*, vol. 57, no. 5, pp. 469–483, 2009.

[11] C. Chi et al., "Diffusion policy: Visuomotor policy learning via action diffusion", *The International Journal of Robotics Research*, vol. 44, no. 10-11, pp. 1684–1704, 2025.

[12] Y. Zhu, D. Yang, and Y. Lee, "Deformable and fragile object manipulation: A review and prospects", *Sensors*, vol. 25, no. 17, p. 5430, 2025.

[13] M. J. Kim et al., "Openvla: An open-source vision-language-action model", *arXiv preprint arXiv:2406.09246*, 2024.

[14] K. Black et al., π_0*: A vision-language-action flow model for general robot control*, 2024. arXiv: 2410.24164 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2410.24164

[15] K. Kawaharazuka, J. Oh, J. Yamada, I. Posner, and Y. Zhu, "Vision-language-action models for robotics: A review towards real-world applications", *IEEE Access*, 2025.

[16] B. Hou et al., "World model for robot learning: A comprehensive survey", *The International Journal of Robotics Research*, 2026.

[17] A. Billard and D. Kragic, "Trends and challenges in robot manipulation", *Science*, vol. 364, no. 6446, eaat8414, 2019.

[18] A. Sanchez-Gonzalez, J. Godwin, T. Pfaff, R. Ying, J. Leskovec, and P. Battaglia, "Learning to simulate complex physics with graph networks", in *International conference on machine learning*, PMLR, 2020, pp. 8459–8468.

[19] M. Moletta, M. K. Wozniak, M. C. Welle, and D. Kragic, "A virtual reality framework for human-robot collaboration in cloth folding", in *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*, IEEE, 2023, pp. 1–7.

[20] A. Longhini et al., "Edo-net: Learning elastic properties of deformable objects from graph dynamics", in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2023, pp. 3875–3881.

[21] M. Moletta, M. C. Welle, N. Ingelhag, J. Munkeby, and D. Kragic, "Real-time operator takeover for visuomotor diffusion policy training", in *2026 IEEE 22nd International Conference on Automation Science and Engineering (CASE)*, IEEE, 2026, pp. 2602–2609.

[22] M. Moletta, M. C. Welle, and D. Kragic, "Preference aligned visuomotor diffusion policies for deformable object manipulation", *IEEE Robotics and Automation Letters*, vol. 11, no. 4, pp. 4985–4992, 2026. DOI: 10.1109/LRA.2026.3665075

[23] T. Lips, M. Moletta, M. C. Welle, D. Kragic, and F. wyffels, "On the generalization capabilities, design choices and limitations of keypoint imitation learning", *arXiv preprint arXiv:2605.26649*, 2026.

[24] A. Longhini et al., "Elastic context: Encoding elasticity for data-driven models of textiles", in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 1764–1770. DOI: 10.1109/ICRA48891.2023.10160740

[25] G. L. Marchetti et al., "Learning coarsened dynamic graph representations for deformable object manipulation", in *2021 20th International Conference on Advanced Robotics (ICAR)*, IEEE, 2021, pp. 955–960.

[26] P. W. Koh et al., "Wilds: A benchmark of in-the-wild distribution shifts", in *International conference on machine learning*, PMLR, 2021, pp. 5637–5664.

[27] Q. Yang, M. C. Welle, D. Kragic, and O. Andersson, "S2-diffusion: Generalizing from instance-level to category-level skills in robot manipulation", *IEEE Robotics and Automation Letters*, 2025.

[28] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world", in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, IEEE, 2017, pp. 23–30.

[29] S. James, A. J. Davison, and E. Johns, "Transferring end-to-end visuomotor control from simulation to real world for a multi-stage task", in *Conference on Robot Learning*, PMLR, 2017, pp. 334–343.

[30] V. E. Arriola-Rios, P. Guler, F. Ficuciello, D. Kragic, B. Siciliano, and J. L. Wyatt, "Modeling of deformable objects for robotic manipulation: A tutorial and review", *Frontiers in Robotics and AI*, vol. 7, p. 82, 2020.

[31] A. Longhini, M. C. Welle, Z. Erickson, and D. Kragic, "Adafold: Adapting folding trajectories of cloths via feedback-loop manipulation", *IEEE Robotics and Automation Letters*, vol. 9, no. 11, pp. 9183–9190, 2024.

[32] J. Bohg et al., "Interactive perception: Leveraging action in perception and perception in action", *IEEE Transactions on Robotics*, vol. 33, no. 6, pp. 1273–1291, 2017.

[33] A. Gerevini and D. Long, "Preferences and soft constraints in pddl3", in *ICAPS workshop on planning with preferences and soft constraints*, 2006, pp. 46–53.

[34] Z. Liu et al., "Preference-calibrated human-in-the-loop reinforcement learning for robotic manipulation", *arXiv preprint arXiv:2606.03949*, 2026.

[35] K. Gandhi, S. Karamcheti, M. Liao, and D. Sadigh, "Eliciting compatible demonstrations for multi-human imitation learning", in *Conference on Robot Learning*, PMLR, 2023, pp. 1981–1991.

[36] D. Sadigh, A. Dragan, S. Sastry, and S. Seshia, "Active preference-based learning of reward functions", in *Proceedings of Robotics: Science and Systems*, Cambridge, Massachusetts, Jul. 2017. DOI: 10.15607/RSS.2017.XIII.053

[37] T. Lesort, N. Díaz-Rodríguez, J.-F. Goudou, and D. Filliat, "State representation learning for control: An overview", *Neural Networks*, vol. 108, pp. 379–392, 2018.

[38] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives", *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[39] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies", *Journal of Machine Learning Research*, vol. 17, no. 39, pp. 1–40, 2016.

[40] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations", *arXiv preprint arXiv:2403.03954*, 2024.

[41] Y. Zhu, A. Joshi, P. Stone, and Y. Zhu, "Viola: Imitation learning for vision-based manipulation with object proposal priors", in *Conference on Robot Learning*, PMLR, 2023, pp. 1199–1210.

[42] T.-C. Lee, R. L. Kashyap, and C.-N. Chu, "Building skeleton models via 3-d medial surface axis thinning algorithms", *CVGIP: graphical models and image processing*, vol. 56, no. 6, pp. 462–478, 1994.

[43] H. Sundar, D. Silver, N. Gagvani, and S. Dickinson, "Skeleton based shape matching and retrieval", in *2003 Shape Modeling International.*, IEEE, 2003, pp. 130–139.

[44] F. Reinders, M. E. Jacobson, and F. H. Post, "Skeleton graph generation for feature shape description", in *Data Visualization 2000: Proceedings of the Joint EUROGRAPHICS and IEEE TCVG Symposium on Visualization in Amsterdam, The Netherlands, May 29–30, 2000*, Springer, 2000, pp. 73–82.

[45] R. Bommasani et al., "On the opportunities and risks of foundation models", *arXiv preprint arXiv:2108.07258*, 2021.

[46] M. Levy, S. Haldar, L. Pinto, and A. Shirivastava, "P3-po: Prescriptive point priors for visuo-spatial generalization of robot policies", in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 4167–4174.

[47] H. Ha and S. Song, "Flingbot: The unreasonable effectiveness of dynamic manipulation for cloth unfolding", in *Conference on Robot Learning*, PMLR, 2022, pp. 24–33.

[48] D. Blanco-Mulero, "Towards efficient robotic manipulation of deformable objects by learning dynamics models and adaptive policies", Ph.D. dissertation, Aalto University, 2024.

[49] T. Pfaff, M. Fortunato, A. Sanchez-Gonzalez, and P. W. Battaglia, "Learning mesh-based simulation with graph networks", *arXiv preprint arXiv:2010.03409*, 2020.

[50] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model", *IEEE transactions on neural networks*, vol. 20, no. 1, pp. 61–80, 2009.

[51] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, and J. Peters, "An algorithmic perspective on imitation learning", *Foundations and Trends® in Robotics*, vol. 7, no. 1-2, pp. 1–179, 2018.

[52] S. Ross and D. Bagnell, "Efficient reductions for imitation learning", in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Y. W. Teh and M. Titterington, Eds., ser. Proceedings of Machine Learning Research, vol. 9, Chia Laguna Resort, Sardinia, Italy: PMLR, May 2010, pp. 661–668.

[53] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning", in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, G. Gordon, D. Dunson, and M. Dudík, Eds., ser. Proceedings of Machine Learning Research, vol. 15, Fort Lauderdale, FL, USA: PMLR, Apr. 2011, pp. 627–635.

[54] C. Celemin et al., "Interactive imitation learning in robotics: A survey", *Foundations and Trends® in Robotics*, vol. 10, no. 1-2, pp. 1–197, 2022.

[55] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences", *Advances in neural information processing systems*, vol. 30, 2017.

[56] Z. Makhataeva and H. A. Varol, "Augmented reality for robotics: A review", *Robotics*, vol. 9, no. 2, p. 21, 2020.

[57] M. Wonsick and T. Padir, "A systematic review of virtual reality interfaces for controlling and interacting with robots", *Applied Sciences*, vol. 10, no. 24, p. 9051, 2020.

[58] S. Xu, S. Moore, and A. Cosgun, "Shared-control robotic manipulation in virtual reality", *arXiv preprint arXiv:2205.10564*, 2022.

[59] C. Barentine, A. McNay, R. Pfaffenbichler, A. Smith, E. Rosen, and E. Phillips, "A vr teleoperation suite with manipulation assist", in *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 442–446.

[60] M. Ostanin and A. Klimchik, "Interactive robot programing using mixed reality", *IFAC-PapersOnLine*, vol. 51, no. 22, pp. 50–55, 2018.

[61] B. Waymouth et al., "Demonstrating cloth folding to robots: Design and evaluation of a 2d and a 3d user interface", in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, IEEE, 2021, pp. 155–160.

[62] J. Borras, A. Boix-Granell, S. Foix, and C. Torras, "A virtual reality framework for fast dataset creation applied to cloth manipulation with automatic semantic labelling", in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2023, pp. 11 605–11 611.

[63] M. C. Welle, N. Ingelhag, M. Lippi, M. Wozniak, A. Gasparri, and D. Kragic, "Quest2ros: An app to facilitate teleoperating robots", in *7th International Workshop on Virtual, Augmented, and Mixed-Reality for Human-Robot Interactions*, 2024.

[64] Y. R. Wang et al., "Roboeval: Where robotic manipulation meets structured and scalable evaluation", *arXiv preprint arXiv:2507.00435*, 2025.

[65] L. Ouyang et al., "Training language models to follow instructions with human feedback", *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[66] J. Fürnkranz and E. Hüllermeier, "Preference learning and ranking by pairwise comparison", in *Preference learning*, Springer, 2010, pp. 65–82.

[67] C. Wirth, R. Akrour, G. Neumann, and J. Fürnkranz, "A survey of preference-based reinforcement learning methods", *Journal of Machine Learning Research*, vol. 18, no. 136, pp. 1–46, 2017.

[68] E. Bıyık, D. P. Losey, M. Palan, N. C. Landolfi, G. Shevchuk, and D. Sadigh, "Learning reward functions from diverse sources of human feedback: Optimally integrating demonstrations and preferences", *The International Journal of Robotics Research*, vol. 41, no. 1, pp. 45–67, 2022.

[69] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model", *Advances in neural information processing systems*, vol. 36, pp. 53 728–53 741, 2023.

[70] M. U. Din, W. Akram, L. S. Saoud, J. Rosell, and I. Hussain, "Vision language action models in robotic manipulation: A systematic review", *arXiv preprint arXiv:2507.10672*, 2025.

[71] P. Intelligence et al., "$\pi^*_{0.6}$: A vla that learns from experience", *arXiv preprint arXiv:2511.14759*, 2025.

[72] J. Lee et al., "Molmoact: Action reasoning models that can reason in space", in *2026 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2026, pp. 2585–2592.

[73] Z. Zhang et al., "Grape: Generalizing robot policy via preference alignment", in *2026 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2026, pp. 16 868–16 875.

[74] T. Su et al., "Demavla: A vision-language-action foundation model for generalizable deformable manipulation", *arXiv preprint arXiv:2605.31286*, 2026.

[75] W. Chow, J. Mao, B. Li, D. Seita, V. Campagnolo Guizilini, and Y. Wang, "Physbench: Benchmarking and enhancing vision-language models for physical world understanding", in *International Conference on Learning Representations*, vol. 2025, 2025, pp. 97 959–98 108.

[76] W. Xia, Y. Yang, H. Wu, X. Ma, T. Kong, and D. Hu, "Human-assisted robotic policy refinement via action preference optimization", *Advances in Neural Information Processing Systems*, vol. 38, pp. 36 746–36 768, 2025.

Part II

Included Publications