GarmentPile: Point-Level Visual Affordance Guided Retrieval and Adaptation for Cluttered Garments Manipulation

Ruihai Wu*[1] Ziyu Zhu*[2,1] Yuran Wang*[1] Yue Chen [1] Jiarui Wang [1] Hao Dong†[1]
[1]CFCS, School of Computer Science, PKU [2]School of EECS, PKU

Abstract

Cluttered garments manipulation poses significant challenges due to the complex, deformable nature of garments and intricate garment relations. Unlike single-garment manipulation, cluttered scenarios require managing complex garment entanglements and interactions, while maintaining garment cleanliness and manipulation stability. To address these demands, we propose to learn point-level affordance, the dense representation modeling the complex space and multi-modal manipulation candidates, while being aware of garment geometry, structure, and inter-object relations. Additionally, as it is difficult to directly retrieve a garment in some extremely entangled clutters, we introduce an adaptation module, guided by learned affordance, to reorganize highly-entangled garments into states plausible for manipulation. Our framework demonstrates effectiveness over environments featuring diverse garment types and pile configurations in both simulation and the real world. Project page: https://garmentpile.github.io/.

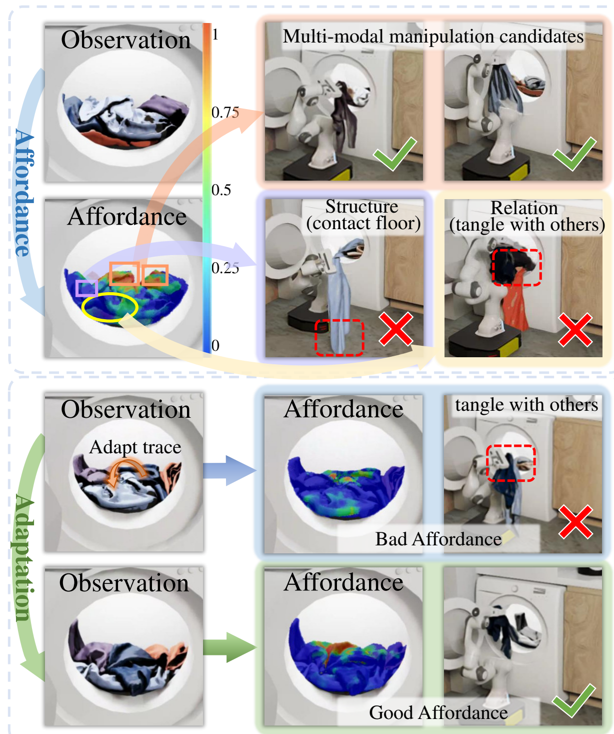

Figure 1. **Point-Level Affordance for Cluttered Garments**. A higher score denotes the higher actionability for downstream retrieval. ***Row 1***: per-point affordance simultaneously reveals 2 garments suitable for retrieval. ***Row 2***: it is aware of garment structures (grasping edges leads other parts contacting floor) and relations (retrieving one garment while dragging nearby entangled garments out), and thus avoids manipulating on points leading to such failures. ***Row 3* and *4***: highly tangled garments may not have plausible manipulation points, affordance can guide reorganizing the scene, and thus garments plausible for manipulation will exist.

1. Introduction

Garments, such as shirts, dresses, and socks, are essential in daily life and pose significant challenges for human-assistive robots. Interacting with various types of garments to perform everyday tasks is crucial yet complex. The complexity of garments comes from their high-dimensional state spaces, intricate kinematics, dynamics and diverse categories. Most studies focus on single-garment manipulation, such as unfolding [14], folding [1, 49], hanging [46], and dressing up [42]. However, many real-life scenarios involve multiple cluttered garments, such as arranging clothes on a bed or retrieving items from a washing machine. In these cases, it is crucial to maintain cleanliness and avoid disturbing adjacent garments (failure cases in Figure 1).

Manipulating cluttered garments presents greater challenges than single-garment or cluttered rigid object manipulation. The more complex states in the clutters and the complicated interrelations between garments make it difficult to distinguish between different garments and infer their relations. Moreover, garment piles often involve multiple plausible retrieval garments (Figure 1, *row 1*), further increasing the demands on the multi-modal representation capability of the learned manipulation policy.

Point-level affordance, derived from 3D point cloud input and representing the **per-point actionability** on the object for downstream tasks, is a suitable representation for cluttered garments manipulation. First, the per-point space supports representing complex states of cluttered scenes. Also, the per-point score can easily represent the multi-modal policy outputs (Figure 1, *row 1*). Most importantly, the feature of each point is extracted from local to global, capable of representing the garment local geometry information for grasping and retrieval, the structural information of each garment, and garment relation indicating whether the action would disturb other garments (Figure 1, *row 2*). For unseen garment clutters, which easily exist as garments have nearly infinite deformations and combinations, the above extracted information (garment geometry, structure and relations) is consistent across scenes, making the representation easily generalize to novel scenarios.

However, affordance alone is not a universal solution. In extreme cases, such as highly tangled garments, the overall affordance is significantly reduced, meaning there may not exist manipulatable positions (Figure 1, *row 3*), and thus need a person or robot to first reorganize the garments to a new state plausible for manipulation (Figure 1, from *row 3* to *row 4*). Therefore, we further introduce a novel design, adaptation module, to mimic what people often do. By iteratively executing the pick-and-place actions, the adaptation module can use learned point-level affordance as the signal to efficiently reorganize cluttered garments.

The absence of suitable simulation environments also partly obstructs the research on cluttered garment manipulation. Previous works have primarily focused on the simulation and manipulation of single garments or simpler deformable objects [4, 36, 37, 45, 47, 53, 55], rather than tackling the challenges posed by cluttered scenarios. To address this, we propose a new evaluation environment based on GarmentLab [29], including 9 garment categories with various deformations and 3 representative scenarios: sofa, washing machine, and basket. Both qualitative and quantitative results from simulations and real-world experiments demonstrate the effectiveness of our framework.

In conclusion, our contributions mainly include:

- We propose to study the novel task of cluttered garments manipulation, and build the pioneering environment with diverse scenarios covering different garment categories.
- We introduce point-level affordance learning for cluttered garments manipulation, with multiple novel designs to efficiently represent highly complex state and action spaces, and multi-modal policy outputs.
- We further develop the adaptation module guided by learned affordance, to efficiently adapt the cluttered garments to states easy to successfully manipulate.
- Extensive experiments in both simulation and real world demonstrate the effectiveness of our framework.

2. Related Work

2.1. Visual Affordance for Robotic Manipulation

Visual affordance [12] indicates actionable possibilities on objects for various manipulation tasks. This approach has been widely used in grasping [7, 19, 51], articulated manipulation [10, 21, 50], and scene interaction [15, 25, 32, 33]. Point-level affordance, in particular, assigns an actionability score to each point, and thus enabled fine-grained geometry understanding and improved cross-shape generalization in diverse tasks, such as articulated [31, 41], bimanual [54] and deformable [45, 46] manipulation. For cluttered garment scenarios, where garments often overlap and entangle, we empower point-level affordance with the awareness of garment structure and inter-relation, and further propose affordance-guided efficient scene adaptation.

2.2. Garment Manipulation

Among diverse categories of objects, such as rigid and articulated objects, deformable objects pose particular challenges in representation and manipulation, for the large state and action space, as well as complex dynamics. Manipulating 1D cables and ropes [37, 47, 53] and square-shaped fabrics [11, 23, 43, 45] have been relatively well-studied with vision, RL, or imitation based methods, Garments featuring more diverse categories, shapes and deformations demonstrate more difficulty. Current studies mainly focus on manipulating a single piece of garment, proposing methods for folding [1, 49], unfolding [3, 14, 26], hanging [5, 46] garments, and dressing up [20, 38, 42]. With the development of garment simulation environment [29], we move a step towards building diverse scenarios with cluttered garments in various categories (*e.g.*, shirts, dresses, gloves and socks), and study the challenging cluttered garment manipulation.

2.3. Cluttered Objects Manipulation

Manipulating objects in cluttered environments is essential for tasks like grasping [39, 40, 52], retrieving [16, 22, 48], and rearranging [6, 13]. Cluttered scenes pose unique challenges, such as occlusions and complex spatial interactions, requiring advanced perception and control. Approaches using visual grounding [30, 48], object detection [24], and relational detection [27] help locate and contextualize target objects, while end-to-end frameworks streamline pose prediction [2, 9]. Learned retrieval affordances [27, 40, 44] further enhance adaptability in dense scenes. However, cluttered garments present unique challenges due to diverse shapes and deformations, complicating detection and relational understanding. We address this by proposing a point-level affordance approach with novel designs that capture the complex states and dynamics of garment piles, enabling more precise manipulation.

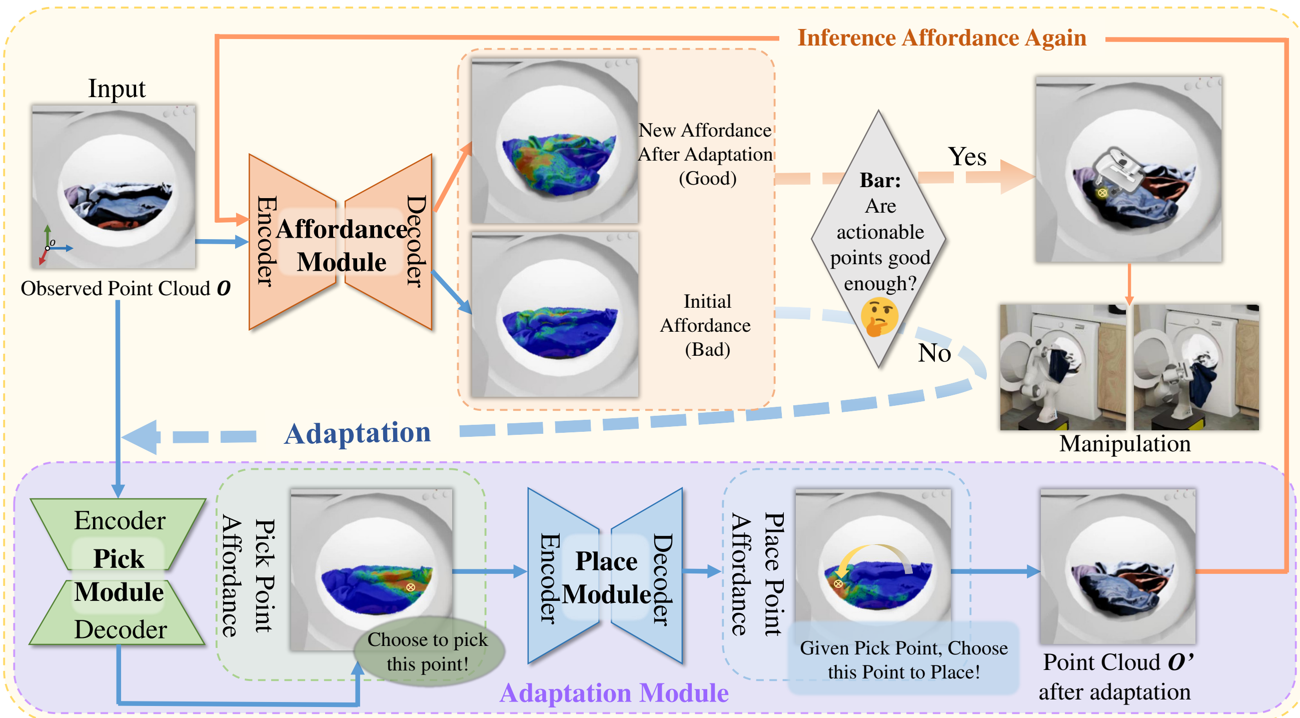

Figure 2. **Framework Overview.** Given the observed point cloud, the Affordance Module predicts the initial point-level manipulation (retrieval) affordance score. When actionability is not good enough, the framework proposes the adaptation pick-place action. It first predicts per-point pick affordance, and selects the pick point with the highest score, conditioned on which it predicts place affordance and selects the place point. After executing adaptation action, it receives a new point cloud and generates new affordance. When actionability is good enough, the robot retrieves on the point with the highest affordance score. This loop is executed until all garments are retrieved.

3. Problem Formulation

Given a clutter of k garments and its 3D point cloud observation $O \in \mathbb{R}^{N\times 3}$, we study garment retrieval, aiming to retrieve k garments one-by-one while avoiding 2 common issues that may lead to uncleanliness or unsafety:

- The target garment contacts the floor during the retrieval (Figure 1, *row 2*, *column 2*).
- When the retrieving one garment, others are dragged out (Figure 1, *row 2*, *column 3*).

We follow previous deformable manipulation studies [28, 37, 45] and use pick-and-place as action primitive. Since placing positions are usually predefined (e.g., placing garments into a basket or washing machine), we use the grasp point $p_{retrieve} \in \mathbb{R}^3$ with heuristic retrieval orientation as **retrieval action**. In case plausible retrieval garment is not available, we use pick-and-place action ($p_{pick} \in \mathbb{R}^3$ and $p_{place} \in \mathbb{R}^3$) as **adaptation action** to reorganize the scene.

We define point-level retrieval / pick / place affordance maps $A^{retrieve}$, A^{pick}, $A^{place}_{p_{pick}} \in \mathbb{R}^N$, each digit normalized to [0,1], indicating per-point actionability for retrieval / pick / place. The point with highest score will be selected.

4. Method

4.1. Overview

We first describe directly learning point-level affordance for cluttered garment manipulation (Section 4.2). Then, we describe learning the adaptation module (Section 4.3), by first learning point-level place affordance supervised by retrieval affordance (Section 4.3.1), and point-level pick affordance supervised by place affordance (Section 4.3.2). Finally, we introduce training strategy (Section 4.4).

4.2. Point-Level Affordance for Retrieval

As described in Introduction and Problem Formulation, the Retrieval Affordance Module (also denoted as Affordance Module for simplicity) $\mathcal{M}_{retrieve}$ predicts the per-point score map $A^{retrieve}$ for each point. Taking as input the point cloud observation O of the garment clutter, we extract the per-point feature using PointNet++ [35] backbone feature extractor $\mathbf{F}_{retrieve}$. As demonstrated in Figure 3 (upper-right), the per-point feature of PointNet++ aggregates the information of local geometry, global structure and garment relations, each of which is essential for predicting whether the manipulation on the target point will succeed (affordance visualizations in 4 with analysis demonstrate such information in the learned affordance map). For the point p, we get the feature $f_p^{retrieve} \in \mathbb{R}^{128}$, and parse it into Multi-Layer Perceptrons (MLPs) with sigmoid [8] activation function for normalization, we can get 1-dimension affordance prediction $\hat{g}_p^{retrieve}$ on p. We define the ground truth retrieval affordance score $g_p^{retrieve}$ on p as 1 (success) or 0 (failure), by directly executing the retrieval action on p and acquiring the manipulation result. We use Binary Cross Entropy (BCELoss) $L_{retrieve}$ to calculate the loss:

$$L_{retrieve} = -\Big(g_p^{retrieve} \cdot \log(\hat{g}_p^{retrieve}) + (1 - g_p^{retrieve}) \cdot \log(1 - \hat{g}_p^{retrieve})\Big) \tag{1}$$

With trained $\mathcal{M}_{retrieve}$, given the 3D point cloud observation $O \in \mathbb{R}^{N\times3}$, we can first infer the point-level retrieval affordance map $A^{retrieval} \in \mathbb{R}^N$ and select the point $p_{retrieval}$ with the highest score for the retrieval action.

4.3. Retrieval Affordance Guided Adaptation

Garments in clutters might be highly entangled, making it difficult to retrieve one garment without disturbing others in some situations, where all points would have low affordance scores, indicating that no point could be manipulated. To deal with this situation, people often reorganize the garments (by picking-placing or stirring), until finding a plausible scene where the subsequent manipulation could be successful. Therefore, we mimic what people often do and propose the adaptation module by iteratively executing the pick-and-place actions to reorganize the scene. The construction of the adaptation module depends on the learned affordance, as it indicates whether the scene is plausible for manipulation. To figure out the specific relationship between affordance and the success rate, we conducted extensive empirical tests and observed that the success rate is significantly high when **the portion of points with high retrieval affordance** (> 0.9) (denoted as P_{high}) exceeds 0.1. Thus, we adapt multiple times until P_{high} exceeds 0.1.

As pick-and-place composites a large action space,

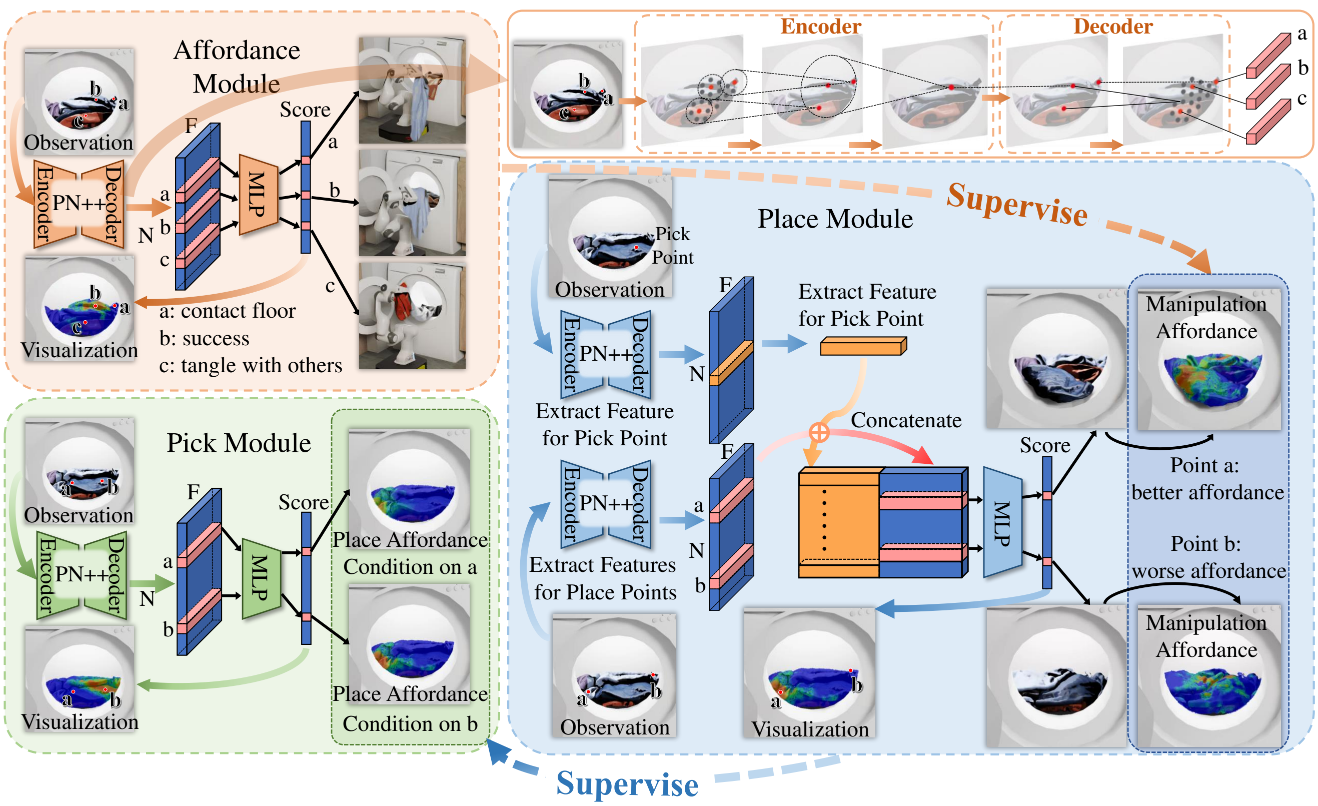

Figure 3. **Learning Framework of Retrieval, Pick and Place Affordance.** Upper-left: the Affordance Module predicts the point-level (retrieval) affordance score for the downstream task. Upper-right: PointNet++ backbone aggregates both local and global features that facilitate incorporating garment geometry, structure and relation information for each point. Lower-right: the Place Module, which predicts the point-level place score conditioned on a pick point for adaptation, is supervised by the trained Affordance Module. Lower-left: the Pick Module, which predicts the point-level place score for adaptation, is supervised by the Place Module.

which is difficult for learning, we disentangle each adaptation action into first predicting the pick point p_{pick} and then the place point p_{place} conditioned on p_{pick}. As the state after placing can be estimated by the learned manipulation (retrieval) affordance, we first learn the actionability for placing on each point given a specific pick point, supervised by the retrieval affordance after the execution of (p_{pick}, p_{place}) action (Section 4.3.1). Then, with the learned place affordance for adaptation, we learn the affordance for p_{pick} (Section 4.3.2).

4.3.1 Point-Level Place Affordance

The adaptation action is composed of a pick point p_{pick} and a place point p_{place}. With a pick point p_{pick}, the Place Affordance Module M_{place} rates the actionability of each point p on whether placing p_{pick} on p will improve the scene. The pick action p_{pick} is difficult to directly get supervision, due to the diversity of the following place actions. On the contrary, the place action p conditioned on p_{pick} can get the direct feedback from the adapted scene by checking the scene actionability (*i.e.*, retrieval affordance) improvement. Therefore, we first train M_{place}.

As demonstrated in Figure 3 (lower-right), for a target place point p, given as input the 3D point cloud O and p_{pick}, two PointNet++, $\mathbf{F}^1_{place}$ and $\mathbf{F}^2_{place}$, respectively extracts the point feature $f^{place_1}_{p_{pick}}$ and $f^{place_2}_{p}$. Then, their feature concatenation is parsed into MLPs to predict the place affordance $\hat{g}^{place}_{p|p_{pick}}$ normalized to [0, 1]. We execute the pick-and-place action from p_{pick} to p and get the new point cloud O', with the new affordance map. If the new affordance map exceeds the initial one by a margin, the ground truth place affordance $g^{place}_{p|p_{pick}}$ is set as 1 otherwise 0. We use BCELoss L_{place} to calculate the loss between $g^{place}_{p|p_{pick}}$ and $\hat{g}^{place}_{p|p_{pick}}$:

$$
\begin{aligned}
L_{place} = - \Big(& g^{place}_{p|p_{pick}} \cdot \log(\hat{g}^{place}_{p|p_{pick}}) \\
& + (1 - g^{place}_{p|p_{pick}}) \cdot \log(1 - \hat{g}^{place}_{p|p_{pick}}) \Big)
\end{aligned} \tag{2}
$$

With the trained $\mathcal{M}_{place}$, given the 3D point cloud observation $O \in \mathbb{R}^{N\times 3}$ and a specific p_{pick}, we can infer the point-level place affordance map $A^{place}_{p_{pick}} \in \mathbb{R}^N$ and select the point p_{place} with the highest score for the place action.

4.3.2 Point-Level Pick Affordance

For adaptation, with a stable Place Module that rates the actionability for each place point conditioned on any pick point, we can further train the Pick Module that rates the actionability g^{pick}_p for each point p, supervised by the best following place action (*i.e.*, the place point select by the learned Place Module) conditioned on pick p.

As demonstrated in Figure 3 (lower-left), taking as input the point cloud observation O of the garment clutter, we extract the per-point feature using PointNet++ backbone feature extractor $\mathbf{F}_{pick}$. For the point p, we get the feature $f^{pick}_p \in \mathbb{R}^{128}$, and parse it into MLPs to predict the pick affordance $\hat{g}^{pick}_p$ normalized to [0, 1]. To get the groud truth score of this point, we use $\mathcal{M}_{place}$ to find the most suitable p_{place} corresponding to p, and then execute the pick-and-place action from p to p_{place} to get the new point cloud O', with the new affordance map. If the new affordance map exceeds the initial one by a margin, the ground truth pick affordance g^{pick}_p is set as 1 otherwise 0.

We use BCELoss L_{pick} to calculate the loss between g^{pick}_p and $\hat{g}^{pick}_p$:

$$
\begin{aligned}
L_{pick} = - \Big(& g^{pick}_p \cdot \log(\hat{g}^{pick}_p) \\
& + (1 - g^{pick}_p) \cdot \log(1 - \hat{g}^{pick}_p) \Big)
\end{aligned} \tag{3}
$$

With the trained $\mathcal{M}_{pick}$, given the 3D point cloud observation $O \in \mathbb{R}^{N\times 3}$, we can infer the point-level pick affordance map $A^{pick} \in \mathbb{R}^N$ and select the point p_{pick} with the highest score for the pick action.

4.4. Inference and Training Details

Figure 2 (caption) describes **inference** pipeline and details.

Figure 3 (caption) describes the **training** pipeline.

For **training data, epoches and computing resources**, we use NVIDIA GeForce 4090 for training. We set batch size to be 128 to train (Retrieval) Affordance and Pick Affordance. While for Place Affordance, we set batch size to be 64 because there are two PointNet++ networks. We collect 20,000 pieces of data and train Retrieval Affordance for 120 epoches, as well as 8,000 pieces of data and train Pick and Place Affordance for 80 epoches. It takes fewer than 24 hours to train each module.

We further use **online data** to boost the robustness during **training**. Since cluttered garments have exceptionally diverse states, the model trained on offline data might not work well in some unseen clutters. Therefore, based on the offline trained models, we collect the scenarios where the models do not well, and further train the models on such data to improve the model robustness. For example, for the (Retrieval) Affordance Module, we randomly sample scenes and use learned affordance model to infer the manipulation point $p_{retrieve}$. If the manipulation failed after the actual execution, we add this data into the buffer, as this point is not actually suitable for manipulation. When the buffer consists on 64 data, we compose a batch of 128 data from the buffer and 64 offline data, to train the model on the scenes in which it makes mistake predictions, while maintaining the knowledge from the previous offline data. We iterate this process as the sampled mistake distributions might have changes, until the model shows consistent performance with low variance. As we can acquire the manipulation or adaptation execution results for each module, the online adaptation proceeded for all modules.

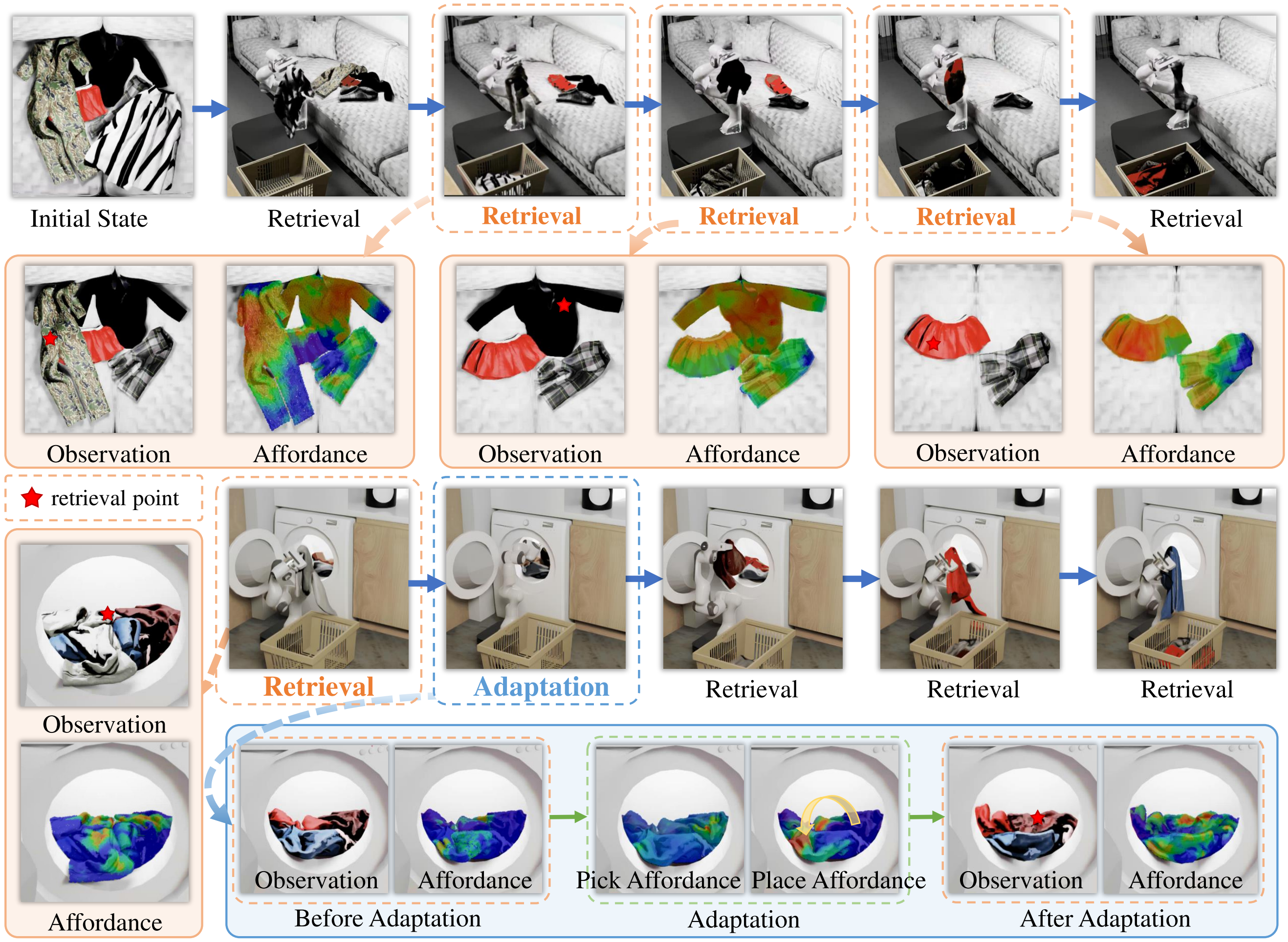

Figure 4. **Example Manipulation Sequences** in WashingMachine and Sofa.

5. Experiments

5.1. Environment, Assets, Data and Evaluation

We use GarmentLab [29] built on Isaac Sim [57] as the environment, which supports simulating multi-material interaction coupling of garments. We load 9 categories (dress, onesie, glove, hat, scarf, trousers, underpants, skirt and top) of 126 different garments from ClothesNet [56]. We construct 3 different types of representative and realistic scenes:

- **WashingMachine** that contains piled garments and represents manipulation with spatial constraints.
- **Basket** that contains piled garments in the basket, representing manipulation with another spatial constraint type.
- **Sofa** that contains piled garments on the sofa, representing manipulation in the open space.

Each scene contains a maximum number of 5 garments in different types with various deformations. To obtain relatively complete point cloud, we place the sensor in front of washing machine for garments inside it, while above (top-down) garments on sofa and in baskets. Same poses are placed in the real world. We use the success rate of manipulation for evaluation.

5.2. Baselines and Ablations

Due to the lack of direct studies in cluttered garments, we compare with methods on affordance learning, cluttered scene understanding and deformable object manipulation:

- **Where2Act** [31] that predicts primitive-level (*i.e.*, grasping or pulling) per-point actionability.
- **Support-M**, the modified version of **Supporting Relation** [18] from cluttered rigid objects to cluttered garments, which first segments each object in the scene, and then retrieves the object not supported by others.
- **GPT-Fabric-M**, the modified **GPT-Fabric** [36] to use GPT-4o [17] to infer spatial relations in garment clutters and retrieve the garment causing least influence to others.

To demonstrate the necessity of the proposed adaptation module, we compare with the following ablated versions:

- **Ours w/o Adaptation** that removes the adaptation part.
- **Ours w/o Pick Afford** that randomly selects the picking point instead of using trained pick affordance.
- **Ours w/o Place Afford** that randomly selects the placing point instead of using trained place affordance.

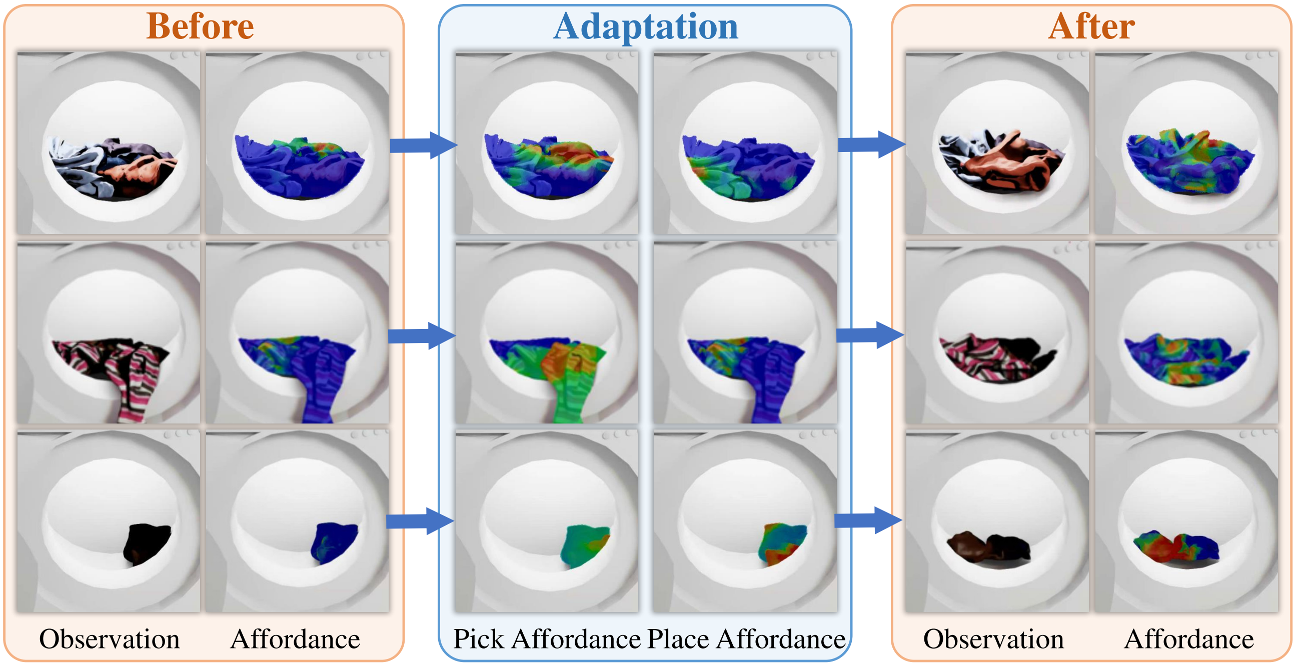

Figure 5. **Retrieval Affordance before and after Adaptation, Adaptation Action indicated by Pick Affordance and Place Affordance**. When predicted retrieval affordance is not good enough(*columns 1, 2*), the adaptation procedure will be triggered. The point with highest score in pick affordance will be chosen as p_{pick} (*column 3*) while the point with the highest score in place affordance will be chosen as p_{place} (*column 4*). After executing pick and place for adaptation, retrieval affordance has improved significantly (*columns 5, 6*).

5.3. Results and Analysis

Figure 4 demonstrates example manipulation trajectories guided by affordance-based adaptation and manipulation. Visualized affordance demonstrates the awareness of following representative features in cluttered garments:

- Spots on garments where the manipulation will not cause garments to fall on the ground. In other words, the policy to some degree understands the overall structure of the garment, and thus will not grasp garment edges as other parts far from the grasping point may fall on the ground.
- Multi-modal retrieval points in clutters. Point-level representation is able and suitable to indicate diverse points in different garments plausible for retrieval. For example, in the Sofa scene, all upside garments have high scores.
- Geometries where the grasp operation will only grasp one garment instead of more. In other words, the junction between 2 garments do not have high affordance scores.
- Spots where the retrieval operation will not drag other garments as the side effect. For example, in the Sofa scene, only upside garments have high affordance scores.

Figure 5 demonstrates the rationality and feasibility of our adaptation policy. What's more, from the procedure of adaptation we can also find out some evidence about how our model is aware of the structure of cluttered garment and take better methods about adapting the pile of garments:

Method	WashingMachine	Sofa	Basket
Ours w/o Adaptation	0.712	0.702	0.693
Ours w/o Pick Afford	0.724	0.704	0.716
Ours w/o Place Afford	0.778	0.743	0.731
Ours	**0.805**	**0.819**	**0.792**

Table 1. **Ablation Study.** Section 5.3 provides detailed analysis.

- *Row 1*: the robot picks the red garment (supporting some upside garments) a little to the front, and thus this garment doe not support others.
- *Row 2*: the sleeve leaks out of the washing machine and thus the garment might easily slip outside the machine when executing any retrieval action. The adaptation model picks the sleeve back into the washing machine.
- *Row 3*: one small garment is hidden on the edge of the washing machine, and thus the robot cannot determine how many garments in the scene. The adaptation module choose to pick right of the front garment to the left, and the hidden garment is revealed.

Table 1 shows results of ablation study. Pick affordance plays a crucial role as only limited points are plausible for picking for adaptation (*i.e.*, picking most points do not improve the scene). Conditioned on a good pick point, place affordance will further contribute to the better adaptation.

Method	WashingMachine	Sofa	Basket
Where2Act [31]	0.585	0.643	0.624
Support-M [18]	0.562	0.784	0.684
GPT-Fabric [36]	0.463	0.408	0.384
Ours	**0.805**	**0.819**	**0.792**

Table 2. **Quantitative Comparisons with Baselines.** Our method outperforms baselines by a margin, with analyses in Section 5.3.

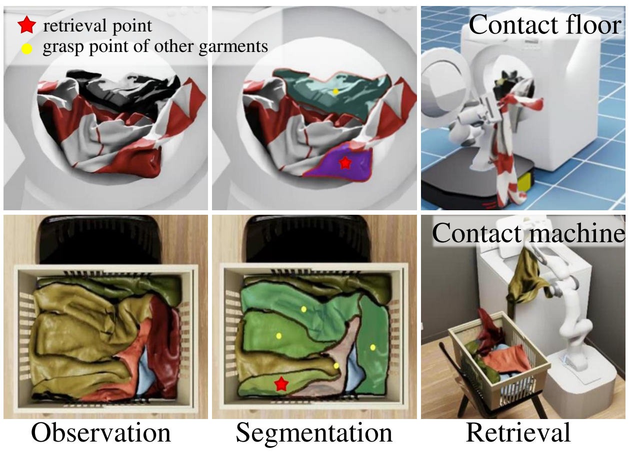

Figure 6. **Manipulation Guided by Segmentation and Support Relation**. Foundation segmentation models cannot precisely segment garments in clutters, misleading the following manipulation.

Table 2 demonstrates the success rate of our method and baselines. For **Where2Act**, as it only denotes affordance for pulling primitive, it cannot precisely guide the downstream retrieval task. For **Support-M**, while it performs well in the Sofa scenario, its success rate drops sharply in WashingMachine and Basket. This is because garments on the sofa are in a relatively open space and thus show relatively intact shapes, making it not so difficult for foundation model to segment garments. However, WashingMachine and Basket scenarios involve severe occlusions and deformations, making garment segmentation significantly more challenging. As demonstrated in Figure 6, foundation segmentation models mistakenly segments part of a garment as a whole garment. More visualizations are demonstrated in the supplementary material. Besides, we observe that segmentation performance is influenced by lighting conditions and garment colors. For example, if the garments in a pile have similar colors or textures, chances are that they might not be segmented to separate individuals. For **GPT-Fabric-M**, as shown in Figure 7, the high complexity of cluttered garments prevents GPT from accurately inferring garment relations. In the supplementary material, we provide the prompts and additional actions guided by GPT-4o. Notably, **Support-M** and **GPT-Fabric-M** require inference times of **3s** and **8s** respectively, which are significantly longer than affordance-based methods, making them less suitable for real-time decision-making and operation.

Figure 7. **Manipulation Guided by GPT-4o.** GPT cannot precisely infer garment relations in complicated clutters.

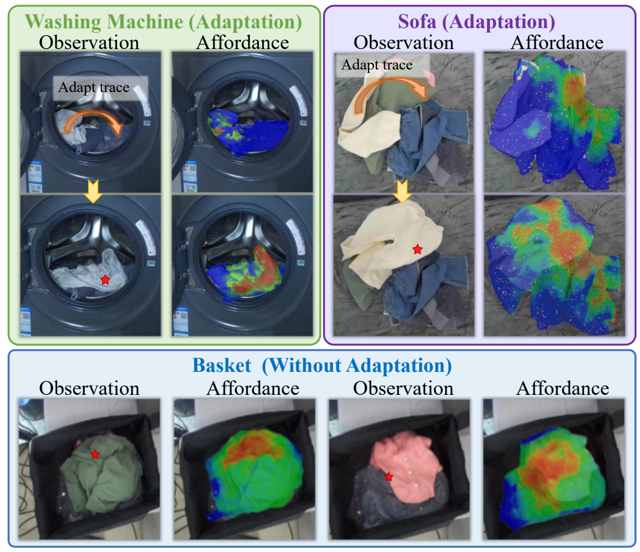

Figure 8. **Real-World Results** of affordance and policy.

Method	WashingMachine	Sofa	Basket
Where2Act [31]	9 / 15	10 / 15	8 / 15
Support-M [18]	8 / 15	12 / 15	9 / 15
GPT-Fabric [36]	6 / 15	7 / 15	6 / 15
Ours	**12 / 15**	**13 / 15**	**12 / 15**

Table 3. **Real-World Success Rate** in different scenarios.

5.4. Real-World Evaluation

We use Franka Panda as the robot, and Kinect to capture depth images, due to its slight noise problem demonstrated by previous studies in articulated object [34] and garment manipulation [46]. Table 3 reports success rates. Figure 8 shows real-world observations, affordance and adaptation. Supplementary video shows more demonstrations.

6. Conclusion

We propose point-level affordance, a dense representation, to capture complex cluttered garments scenarios and guide downstream manipulation. Since cluttered garments may require initial adaptation before manipulation, we further introduce affordance-guided adaptation. Experiments in diverse scenarios demonstrate the superiority of our method.

Acknowledgment. This project was supported by National Natural Science Foundation of China (62376006) and National Youth Talent Support Program (8200800081).

References

[1] Yahav Avigal, Lars Berscheid, Tamim Asfour, Torsten Kröger, and Ken Goldberg. Speedfolding: Learning efficient bimanual folding of garments. In *IROS*, 2022. 1, 2

[2] Michel Breyer, Jen Jen Chung, Lionel Ott, Roland Siegwart, and Juan Nieto. Volumetric grasping network: Real-time 6 dof grasp detection in clutter. In *Conference on Robot Learning*, pages 1602–1611. PMLR, 2021. 2

[3] Alper Canberk, Cheng Chi, Huy Ha, Benjamin Burchfiel, Eric Cousineau, Siyuan Feng, and Shuran Song. Cloth funnels: Canonicalized-alignment for multi-purpose garment manipulation. *ICRA*, 2023. 2

[4] Lawrence Yunliang Chen, Baiyu Shi, Daniel Seita, Richard Cheng, Thomas Kollar, David Held, and Ken Goldberg. Autobag: Learning to open plastic bags and insert objects, 2023. 2

[5] Wei Chen, Dongmyoung Lee, Digby Chappell, and Nicolas Rojas. Learning to grasp clothing structural regions for garment manipulation tasks. In *IROS*, 2023. 2

[6] Sang Hun Cheong, Brian Y Cho, Jinhwi Lee, ChangHwan Kim, and Changjoo Nam. Where to relocate?: Object rearrangement inside cluttered and confined environments for robotic manipulation. In *ICRA*, 2020. 2

[7] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5031–5041, 2020. 2

[8] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *CoRR*, 2017. 4

[9] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *CVPR*, 2020. 2

[10] Samir Yitzhak Gadre, Kiana Ehsani, and Shuran Song. Act the part: Learning interaction strategies for articulated object part discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15752–15761, 2021. 2

[11] Aditya Ganapathi, Priya Sundaresan, Brijen Thananjeyan, Ashwin Balakrishna, Daniel Seita, Jennifer Grannen, Minho Hwang, Ryan Hoque, Joseph E Gonzalez, Nawid Jamali, et al. Learning dense visual correspondences in simulation to smooth and fold real fabrics. In *ICRA*, 2021. 2

[12] James J Gibson. The theory of affordances. *Hilldale, USA*, 1(2):67–82, 1977. 2

[13] Ankit Goyal, Arsalan Mousavian, Chris Paxton, Yu-Wei Chao, Brian Okorn, Jia Deng, and Dieter Fox. Ifor: Iterative flow minimization for robotic object rearrangement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14787–14797, 2022. 2

[14] Huy Ha and Shuran Song. Flingbot: The unreasonable effectiveness of dynamic manipulation for cloth unfolding. In *Conference on Robot Learning*. PMLR, 2022. 1, 2

[15] Mohammed Hassanin, Salman Khan, and Murat Tahtali. Visual affordance and function understanding: A survey. *ACM Computing Surveys (CSUR)*, 54(3):1–35, 2021. 2

[16] Baichuan Huang, Shuai D Han, Jingjin Yu, and Abdeslam Boularias. Visual foresight trees for object retrieval from clutter with nonprehensile rearrangement. *IEEE Robotics and Automation Letters*, 7(1):231–238, 2021. 2

[17] Raisa Islam and Owana Marzia Moushi. Gpt-4o: The cutting-edge advancement in multimodal llm. *Authorea Preprints*, 2024. 6

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 6, 8

[19] Mia Kokic, Danica Kragic, and Jeannette Bohg. Learning task-oriented grasping from human activity datasets. *IEEE Robotics and Automation Letters*, 5(2):3352–3359, 2020. 2

[20] Stelios Kotsovolis and Yiannis Demiris. Model predictive control with graph dynamics for garment opening insertion during robot-assisted dressing. In *ICRA*, 2024. 2

[21] Yuxuan Kuang, Junjie Ye, Haoran Geng, Jiageng Mao, Congyue Deng, Leonidas Guibas, He Wang, and Yue Wang. Ram: Retrieval-based affordance transfer for generalizable zero-shot robotic manipulation. *CoRL*, 2024. 2

[22] Andrey Kurenkov, Joseph Taglic, Rohun Kulkarni, Marcus Dominguez-Kuhne, Animesh Garg, Roberto Martín-Martín, and Silvio Savarese. Visuomotor mechanical search: Learning to retrieve target objects in clutter. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8408–8414. IEEE, 2020. 2

[23] Robert Lee, Daniel Ward, Akansel Cosgun, Vibhavari Dasagi, Peter Corke, and Jurgen Leitner. Learning arbitrary-goal fabric folding with one hour of real robot experience. *arXiv preprint arXiv:2010.03209*, 2020. 2

[24] Dayou Li, Pengkun Wei, Chenkun Zhao, Shuo Yang, Yibin Li, and Wei Zhang. A mobile manipulation system for automated replenishment in the field of unmanned retail. In *2023 IEEE International Conference on Mechatronics and Automation (ICMA)*, pages 644–649. IEEE, 2023. 2

[25] Xueting Li, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Putting humans in a scene: Learning affordance in 3d indoor environments. In *CVPR*, 2019. 2

[26] Yinxiao Li, Danfei Xu, Yonghao Yue, Yan Wang, Shih-Fu Chang, Eitan Grinspun, and Peter K Allen. Regrasping and unfolding of garments using predictive thin shell modeling. In *ICRA*, 2015. 2

[27] Yitong Li, Ruihai Wu, Haoran Lu, Chuanruo Ning, Yan Shen, Guanqi Zhan, and Hao Dong. Broadcasting support relations recursively from local dynamics for object retrieval in clutters. In *Robotics: Science and Systems*, 2024. 2

[28] Xingyu Lin, Yufei Wang, Jake Olkin, and David Held. Softgym: Benchmarking deep reinforcement learning for deformable object manipulation. In *CoRL*, 2020. 3

[29] Haoran Lu, Ruihai Wu, Yitong Li, Sijie Li, Ziyu Zhu, Chuanruo Ning, Yan Shen, Longzan Luo, Yuanpei Chen, and Hao Dong. Garmentlab: A unified simulation and benchmark for garment manipulation. In *Advances in Neural Information Processing Systems*, 2024. 2, 6

[30] Yuhao Lu, Yixuan Fan, Beixing Deng, Fangfu Liu, Yali Li, and Shengjin Wang. Vl-grasp: a 6-dof interactive grasp policy for language-oriented objects in cluttered indoor scenes. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 976–983. IEEE, 2023. 2

[31] Kaichun Mo, Leonidas J Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *CVPR*, 2021. 2, 6, 8

[32] Tushar Nagarajan and Kristen Grauman. Learning affordance landscapes for interaction exploration in 3d environments. In *NeurIPS*, 2020. 2

[33] Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, and Kristen Grauman. Ego-topo: Environment affordances from egocentric video. In *CVPR*, 2020. 2

[34] Chuanruo Ning, Ruihai Wu, Haoran Lu, Kaichun Mo, and Hao Dong. Where2explore: Few-shot affordance learning for unseen novel categories of articulated objects. *arXiv preprint arXiv:2309.07473*, 2023. 8

[35] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 2017. 3

[36] Vedant Raval, Enyu Zhao, Hejia Zhang, Stefanos Nikolaidis, and Daniel Seita. Gpt-fabric: Folding and smoothing fabric by leveraging pre-trained foundation models. *arXiv preprint arXiv:2406.09640*, 2024. 2, 6, 8

[37] Daniel Seita, Pete Florence, Jonathan Tompson, Erwin Coumans, Vikas Sindhwani, Ken Goldberg, and Andy Zeng. Learning to rearrange deformable cables, fabrics, and bags with goal-conditioned transporter networks. In *ICRA*, 2021. 2, 3

[38] Zhanyi Sun, Yufei Wang, David Held, and Zackory Erickson. Force-constrained visual policy: Safe robot-assisted dressing via multi-modal sensing. *RA-L*, 2024. 2

[39] Martin Sundermeyer, Arsalan Mousavian, Rudolph Triebel, and Dieter Fox. Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes. In *ICRA*, 2021. 2

[40] Chenxi Wang, Hao-Shu Fang, Minghao Gou, Hongjie Fang, Jin Gao, and Cewu Lu. Graspness discovery in clutters for fast and accurate grasp detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15964–15973, 2021. 2

[41] Yian Wang, Ruihai Wu, Kaichun Mo, Jiaqi Ke, Qingnan Fan, Leonidas Guibas, and Hao Dong. Adaafford: Learning to adapt manipulation affordance for 3d articulated objects via few-shot interactions. *European conference on computer vision (ECCV 2022)*, 2022. 2

[42] Yufei Wang, Zhanyi Sun, Zackory Erickson, and David Held. One policy to dress them all: Learning to dress people with diverse poses and garments. In *RSS*, 2023. 1, 2

[43] Thomas Weng, Sujay Man Bajracharya, Yufei Wang, Khush Agrawal, and David Held. Fabricflownet: Bimanual cloth manipulation with a flow-based policy. In *CoRL*, 2022. 2

[44] Ruihai Wu, Kai Cheng, Yan Zhao, Chuanruo Ning, Guanqi Zhan, and Hao Dong. Learning environment-aware affordance for 3d articulated object manipulation under occlusions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[45] Ruihai Wu, Chuanruo Ning, and Hao Dong. Learning foresightful dense visual affordance for deformable object manipulation. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 2, 3

[46] Ruihai Wu, Haoran Lu, Yiyan Wang, Yubo Wang, and Hao Dong. Unigarmentmanip: A unified framework for category-level garment manipulation via dense visual correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 8

[47] Yilin Wu, Wilson Yan, Thanard Kurutach, Lerrel Pinto, and Pieter Abbeel. Learning to manipulate deformable objects without demonstrations. *RSS*, 2020. 2

[48] Kechun Xu, Shuqi Zhao, Zhongxiang Zhou, Zizhang Li, Huaijin Pi, Yifeng Zhu, Yue Wang, and Rong Xiong. A joint modeling of vision-language-action for target-oriented grasping in clutter. *arXiv preprint arXiv:2302.12610*, 2023. 2

[49] Han Xue, Yutong Li, Wenqiang Xu, Huanyu Li, Dongzhe Zheng, and Cewu Lu. Unifolding: Towards sample-efficient, scalable, and generalizable robotic garment folding. In *CoRL*, 2023. 1, 2

[50] Chengbo Yuan, Chuan Wen, Tong Zhang, and Yang Gao. General flow as foundation affordance for scalable robot learning. *arXiv preprint arXiv:2401.11439*, 2024. 2

[51] Andy Zeng, Shuran Song, Kuan-Ting Yu, Elliott Donlon, Francois R Hogan, Maria Bauza, Daolin Ma, Orion Taylor, Melody Liu, Eudald Romo, et al. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. In *ICRA*, 2018. 2

[52] Andy Zeng, Shuran Song, Kuan-Ting Yu, Elliott Donlon, Francois R Hogan, Maria Bauza, Daolin Ma, Orion Taylor, Melody Liu, Eudald Romo, et al. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. *IJRR*, 2022. 2

[53] Kaifeng Zhang, Baoyu Li, Kris Hauser, and Yunzhu Li. Adaptigraph: Material-adaptive graph-based neural dynamics for robotic manipulation. In *RSS*, 2024. 2

[54] Yan Zhao, Ruihai Wu, Zhehuan Chen, Yourong Zhang, Qingnan Fan, Kaichun Mo, and Hao Dong. Dualafford: Learning collaborative visual affordance for dual-gripper object manipulation. *ICLR*, 2023. 2

[55] Sun Zhaole, Jihong Zhu, and Robert B Fisher. Dexdlo: Learning goal-conditioned dexterous policy for dynamic manipulation of deformable linear objects. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16009–16015. IEEE, 2024. 2

[56] Bingyang Zhou, Haoyu Zhou, Tianhai Liang, Qiaojun Yu, Siheng Zhao, Yuwei Zeng, Jun Lv, Siyuan Luo, Qiancai Wang, Xinyuan Yu, Haonan Chen, Cewu Lu, and Lin Shao. Clothesnet: An information-rich 3d garment model repository with simulated clothes environment. *ICCV*, 2023. 6

[57] Zhehua Zhou, Jiayang Song, Xuan Xie, Zhan Shu, Lei Ma, Dikai Liu, Jianxiong Yin, and Simon See. Towards building ai-cps with nvidia isaac sim: An industrial benchmark and case study for robotics manipulation. In *Proceedings of the 46th International Conference on Software Engineering: Software Engineering in Practice*, page 263–274. ACM, 2024. 6