

Supplementary Files For UniFolding: Towards Sample-efficient, Scalable, and Generalizable Robotic Garment Folding

Anonymous Author(s)

Affiliation

Address

email

1 Evidence of Human Preferences in *fling* Action

2 Fig. 1 and Fig. 2 show grasping point distribution (showed in NOCS[1] space) of human demonstra-
3 tion data collected by VR. We can see that humans frequently grasp shoulders, collars, and waists
4 in the earlier stage of the unfolding process when the garment is usually more crumpled. Humans
5 will probably grasp shoulders at the later stage of the unfolding process when the garment is more
6 flattened and recognizable.



Figure 1: The grasping point distribution (showed in NOCS[1] space) for *fling* action in human demonstration data through VR. These points are from **earlier** steps of the unfolding process.



Figure 2: The grasping point distribution (showed in NOCS[1] space) for *fling* action in human demonstration data through VR. These points are from **later** steps of the unfolding process.

7 2 How R_C and R_A are Calculated

8 Intuitively speaking, R_C encourages actions that make the garment more flattened and more similar
9 to the canonical pose, and R_A encourages actions that make the garment more aligned with the

10 target pose in planar position and rotation. Please refer to ClothFunnels[2] for the detailed definition
11 of R_C and R_A .

12 3 Details of Human Demonstration Data Collection in VR

13 **VR Recording System** We build a real-time data recording system for collecting human demon-
14 stration data for garment manipulation in Virtual Reality. This system is based on the VR-Garment
15 system implemented in GarmentTracking[3]. It is driven by Unity, and the physics engine for cloth
16 simulation is based on Obi. In practice, this system can effectively collect large amounts of human
17 demonstration data for thousands of garments with different shapes and sizes.

18 **Data Recording Pipeline** The data recording pipeline is similar to that in GarmentTracking[3].
19 Firstly, the volunteer will put on an HTC Vive Pro VR Headset and VRTRIX VR gloves. Secondly, a
20 virtual garment from CLOTH3D[4] dataset will randomly drop on the table in virtual space. Thirdly,
21 the volunteer will use his hands to perform the action primitives defined in the main paper for
22 multiple steps to fully smooth and fold the garment. On average, the whole multi-step manipulation
23 process for one garment only takes about 20s in VR.

24 **Data Post-processing** The raw data generated by the data recording pipeline are videos that con-
25 tain the garment mesh vertices and hand poses of each frame. We use a simple method to automati-
26 cally convert hand poses into robot gripper poses, please see the supplementary files for more details.
27 After data recording, We will perform the following data post-processing steps to generate data that
28 are available for network training: Firstly, we automatically divide the whole video of the garment
29 manipulation process into multiple valid action intervals. The start and ending of each action inter-
30 val are decided by the grasping and releasing states of both human hands. Secondly, we use simple
31 rules to automatically generate labels of action primitive type for all valid action intervals based on
32 patterns of human actions. Thirdly, we re-render the garment mesh in Unity and generate RGB-D
33 image, mask, NOCS[1] map, and gripper poses for the starting frame of each action interval.

34 4 Details of Human Preference Annotation and Learning

35 In practice, we generate 8 comparisons from selected keypoint candidates with a fixed threshold
36 $R_{CA} > c$ for each data sample, which can filter out most of the bad keypoint candidates. We invite
37 two volunteers to annotate the same data samples and drop those annotations where two volunteers
38 do not agree, which can increase data annotation quality.

39 5 Keypoint Prediction for *fling* action

40 The dense features generated by the Transformer model will be used for the pose prediction branch
41 for fling action. This branch will predict two grasp points for *fling* action. The grasp point indicates
42 the location on the garment where the robot should grip and perform the flinging action.

43 **Keypoint Candidate Prediction** After analyzing the statistics of human demonstration data in
44 VR, we find that humans will frequently grasp recognizable keypoints on the garment (e.g. cuff,
45 shoulder, waist) for fling action (please see the supplementary file for more details). Motivated
46 by this observation, we choose to directly learn possible keypoint candidates purely from human
47 demonstration data. However, the distribution of these keypoint candidates on the garment is multi-
48 modal, so we firstly predict K possible keypoint candidates $\mathcal{P} = \{\mathbf{P}_1, \dots, \mathbf{P}_K\}$, then supervise
49 them with the variety (Minimum-over-N) loss[5] in Eq. 1:

$$L_{kp}(\mathcal{P}, \mathbf{P}^*) = \min_{\{\mathbf{P}_1, \dots, \mathbf{P}_K\} \in \mathcal{P}} \{d(\mathbf{P}^*, \mathbf{P}_1), d(\mathbf{P}^*, \mathbf{P}_2), \dots, d(\mathbf{P}^*, \mathbf{P}_K)\} \quad (1)$$

50 where \mathbf{P}^* is the ground-truth keypoint, and $d(\cdot, \cdot)$ is the distance metric. Intuitively, L_{kp} only
51 supervises the predicted keypoint closet to the ground-truth keypoint, which encourages the variety

52 of the K predicted keypoints. For fling action, we have two ground-truth keypoints $\{\mathbf{P}_{left}^*, \mathbf{P}_{right}^*\}$
 53 for dual-arm robots, so the final loss is shown in Eq. 2:

$$L_{kp}(\mathcal{P}, \mathbf{P}_{left}^*, \mathbf{P}_{right}^*) = (L_{kp}(\mathcal{P}, \mathbf{P}_{left}^*) + L_{kp}(\mathcal{P}, \mathbf{P}_{right}^*))/2 \quad (2)$$

54 As for the prediction of keypoint candidates \mathcal{P} , an intuitive way is to use attention-based offset
 55 voting[] to directly regress keypoints in 3D task space (the coordinate frame of the input point
 56 cloud) as shown in Eq. 3:

$$\mathbf{P}_j = \frac{1}{N} \sum_{i=1}^N a_{i,j} (\mathbf{x}_i + \mathbf{o}_{i,j}), \quad s.t. \sum_{i=1}^N a_{i,j} = 1 \quad (3)$$

57 where \mathbf{P}_j is the j -th keypoint prediction, $a_{i,j} \in [0, 1]$ is the attention score, $\mathbf{x}_i \in \mathbf{X}$ is the i -th
 58 point in the input point cloud \mathbf{X} , and $\mathbf{o}_{i,j}$ is the 3D offsets of the j -th keypoint \mathbf{P}_j respective to
 59 the i -th point \mathbf{x}_i . The attention score $a_{i,j}$ and offsets $\mathbf{o}_{i,j}$ are predicted by MLP with dense features
 60 generated by Transformer as input.

61 **Prediction in Canonical Space** In practice, we find that regressing keypoint candidates in canon-
 62 ical space (Normalized Object Coordinate Space, NOCS[1]) is much easier than regressing them
 63 directly in task space. So we additionally predict per-point NOCS coordinate $\mathbf{c}_i \in \mathcal{C}$ for the in-
 64 put point cloud with dense features generated by the Transformer. Due to the bilateral symmetry
 65 property of most garments, we use the symmetric Huber loss defined in Eq. 4 to supervise NOCS
 66 prediction \mathcal{C} :

$$L_{nocs}(\mathcal{C}, \mathcal{C}^*) = \min \left\{ \frac{1}{N} \sum_{i=1, \dots, N} Huber(\mathbf{c}_i, \mathbf{c}_i^*), \frac{1}{N} \sum_{i=1, \dots, N} Huber(\mathbf{c}_i, \mathbf{c}_i^{*sym}) \right\} \quad (4)$$

67 where $\mathbf{c}_i^* \in \mathcal{C}^*$ is the original ground-truth NOCS coordinate of i -th point, and \mathbf{c}_i^{*sym} is the sym-
 68 metrical ground-truth NOCS target of i -th point.

69 Then we can modify Eq. 3 by replacing \mathbf{x}_i with \mathbf{c}_i to generate K keypoint predictions \mathcal{P}^{nocs} in
 70 canonical space instead of task space, which is shown in Eq. 5:

$$\mathbf{P}_j^{nocs} = \frac{1}{N} \sum_{i=1}^N a_{i,j} (\mathbf{c}_i + \mathbf{o}_{i,j}), \quad s.t. \sum_{i=1}^N a_{i,j} = 1 \quad (5)$$

71 Nextly, we need to find the corresponding 3D location \mathbf{P}_j in task space for j -th keypoint from
 72 NOCS coordinate \mathbf{P}_j^{nocs} in canonical space. Due to the local similarity of the NOCS coordinates,
 73 we can calculate \mathbf{P}_j by weighted sum defined in Eq. 6:

$$\mathbf{P}_j = \frac{\sum_{i=1}^N w_{i,j} \mathbf{x}_i}{\sum_{i=1}^N w_{i,j}}, \quad w_{i,j} = \exp(-\alpha \cdot \|\mathbf{P}_j^{nocs} - \mathbf{c}_i\|_2) \quad (6)$$

74 Intuitively, $w_{i,j}$ is the weight based on the L2-distance between j -th keypoint \mathbf{P}_j^{nocs} and i -th point
 75 \mathbf{c}_i in canonical space. The larger $w_{i,j}$ is, the more likely j -th keypoint \mathbf{P}_j is closer to the i -th point
 76 \mathbf{x}_i in task space. We set $\alpha = 50$ by default.

77 Finally, we can supervise K keypoint candidate predictions both in canonical space and task space
 78 by Eq. 7:

$$L_{kp.all}(\mathcal{P}^{nocs}, \mathcal{P}, \mathbf{P}^{*nocs}, \mathbf{P}^*) = L_{kp}(\mathcal{P}^{nocs}, \mathbf{P}^{*nocs}) + L_{kp}(\mathcal{P}, \mathbf{P}^*) \quad (7)$$

79 6 Additional Garment Details

80 This section presents the parameters of the garments that are used in our experiment. We use a total
 81 of 60 garments, divided into two sets: a test set of 10 long-sleeved and 10 short-sleeved garments,
 82 and a training set of 20 long-sleeved and 20 short-sleeved garments. The garments cover various
 83 materials and textures. Each garment is assigned a unique ID, and its size and material are also listed

84 in the table. The size information indicates the height and width of the garment when fully unfolded.
 85 In addition, we capture an RGB image of each garment from a top-down view.

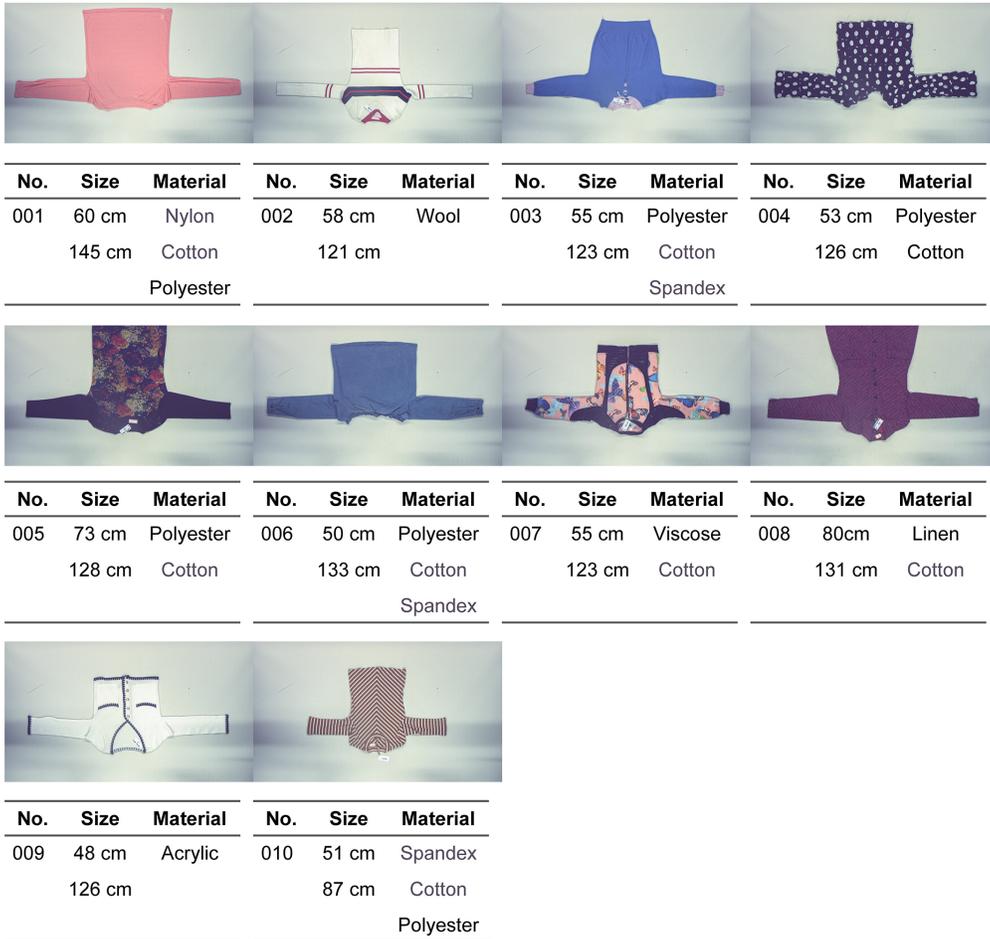
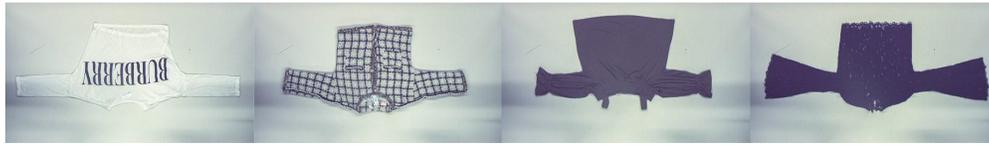


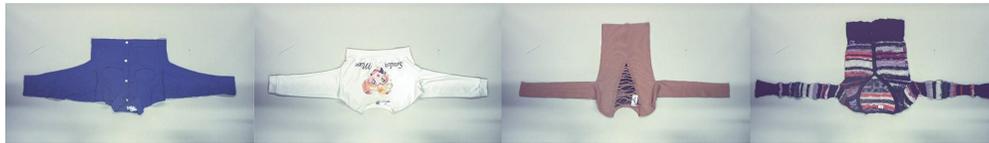
Figure 3: Long-sleeve Shirts (Test Set)

			
No. Size Material	No. Size Material	No. Size Material	No. Size Material
011 49 cm Cotton 63 cm	012 63 cm Spandex 70 cm Viscose	013 55 cm Cotton 83 cm	014 60 cm Polyester 83 cm Cotton
			
No. Size Material	No. Size Material	No. Size Material	No. Size Material
015 46 cm Viscose 133 cm Nylon	016 66 cm Cotton 65 cm Linen	017 56 cm Polyester 95 cm Cotton	018 69 cm Cotton 90 cm
			
No. Size Material	No. Size Material		
019 41 cm Polyester 73 cm Cotton	020 43 cm Polyester 67 cm Cotton		

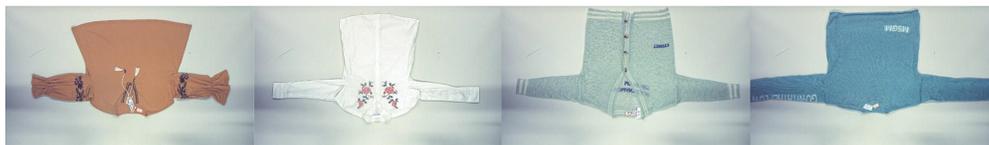
Figure 4: Short-sleeve T-Shirts (Test Set)



No.	Size	Material									
021	50 cm 144 cm	Nylon	022	53 cm 115 cm	Acrylic	023	51 cm 110 cm	Cotton	024	53 cm 143 cm	Polyester Cotton



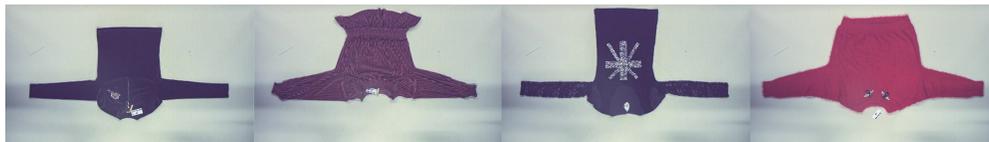
No.	Size	Material	No.	Size	Material	No.	Size	Material	No.	Size	Material
025	46 cm 133 cm	Polyester Cotton	026	38 cm 141 cm	Cotton	027	57 cm 140 cm	Polyester Cotton Spandex	028	63 cm 155 cm	Wool



No.	Size	Material	No.	Size	Material	No.	Size	Material	No.	Size	Material
029	57 cm 128 cm	Viscose Cotton Spandex	030	66 cm 129 cm	Polyester Cotton	031	70 cm 143 cm	Viscose Nylon	032	65 cm 167 cm	Viscose Nylon



No.	Size	Material	No.	Size	Material	No.	Size	Material	No.	Size	Material
033	57 cm 137 cm	Polyester Cotton Spandex	034	63 cm 151 cm	Polyester Cotton Spandex	035	53 cm 91 cm	Polyester Cotton Spandex	036	49 cm 121 cm	Cotton



No.	Size	Material	No.	Size	Material	No.	Size	Material	No.	Size	Material
037	53 cm 125 cm	Polyester Nylon	038	56 cm 133 cm	Polyester Nylon Spandex	039	66 cm 134 cm	Cotton Nylon Spandex	040	60 cm 142 cm	Polyester

Figure 5: Long-sleeve Shirts (Train Set)



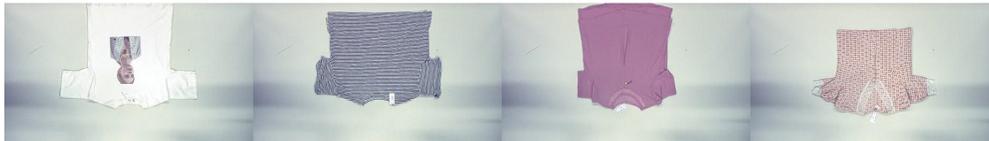
No.	Size	Material									
041	60 cm	Cotton	042	66 cm	Viscose	043	59 cm	Cotton	044	68 cm	Polyester
	69 cm			60 cm	Cotton		75 cm			68 cm	Nylon



No.	Size	Material	No.	Size	Material	No.	Size	Material	No.	Size	Material
045	49 cm	Polyester	046	61 cm	Cotton	047	55 cm	Polyester	048	60 cm	Polyester
	67 cm	Cotton		70 cm			65 cm	Cotton		75 cm	Cotton



No.	Size	Material	No.	Size	Material	No.	Size	Material	No.	Size	Material
049	61 cm	Viscose	050	60 cm	Viscose	051	66 cm	Polyester	052	69 cm	Polyester
	69 cm	Cotton		66 cm	Cotton		78 cm	Cotton		69 cm	Cotton
		Spandex			Spandex						Spandex



No.	Size	Material	No.	Size	Material	No.	Size	Material	No.	Size	Material
053	68 cm	Cotton	054	58 cm	Viscose	055	67 cm	Polyester	056	53 cm	Polyester
	85 cm			78 cm	Cotton		60 cm	Cotton		76 cm	Cotton
					Spandex			Spandex			



No.	Size	Material	No.	Size	Material	No.	Size	Material	No.	Size	Material
057	50 cm	Polyester	058	58 cm	Polyester	059	42 cm	Spandex	060	61 cm	Polyester
	98 cm	Cotton		67 cm	Cotton		70 cm	Cotton		83 cm	Nylon
					Spandex						

Figure 6: Short-sleeve T-Shirts (Train Set)

86 **References**

- 87 [1] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas. Normalized object
88 coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the*
89 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019.
- 90 [2] A. Canberk, C. Chi, H. Ha, B. Burchfiel, E. Cousineau, S. Feng, and S. Song. Cloth
91 funnels: Canonicalized-alignment for multi-purpose garment manipulation. *arXiv preprint*
92 *arXiv:2210.09347*, 2022.
- 93 [3] H. Xue, W. Xu, J. Zhang, T. Tang, Y. Li, W. Du, R. Ye, and C. Lu. Garmenttracking: Category-
94 level garment pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision*
95 *and Pattern Recognition*, pages 21233–21242, 2023.
- 96 [4] H. Bertiche, M. Madadi, and S. Escalera. Cloth3d: clothed 3d humans. In *Computer Vision–*
97 *ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part*
98 *XX 16*, pages 344–359. Springer, 2020.
- 99 [5] L. A. Thiede and P. P. Brahma. Analyzing the variety loss in the context of probabilistic trajec-
100 tory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*,
101 pages 9954–9963, 2019.