DexKnot: Generalizable Visuomotor Policy Learning for Dexterous Bag-Knotting Manipulation

Anonymous Authors

Abstract— **Knotting plastic bags is a common task in daily life, yet it is challenging for robots due to the bags' infinite degree of freedom, 3D spatial structure, and complex physical dynamics. Existing methods often struggle in generalization to unseen bag instances or deformations. To address this, we present DexKnot, a framework that combines keypoint affordance with diffusion policy to learn a generalizable bag-knotting policy. Our approach proposes an instance- and deformation-invariant representation of bags through real-world manual deformation. For an unseen bag configuration, the keypoints can be identified by matching the representation to a reference. These keypoints are then provided to a diffusion transformer, which generates robot action based on a small number of human demonstrations. DexKnot enables effective policy generalization by reducing the dimensionality of observation space into a sparse set of keypoints. Experiments show that DexKnot achieves reliable and consistent knotting performance across a variety of previously unseen instances and deformations.**

I. INTRODUCTION

Knotting plastic bags is a common and useful task in daily life, yet it is not easy for robots to handle such highly deformable objects [1]–[5]. In robot manipulation, while significant progress has been made in handling rigid and articulated objects [6], operating deformable objects remains a formidable challenge for two primary reasons. First, their infinite degrees of freedom (DoF) lead to a very high-dimensional observation space, causing difficulties for a policy, either imitation learning or reinforcement learning, to learn and generalize. Second, deformable objects have complex and highly variable mechanical properties and physical dynamics, which are difficult to learn for neural surrogate models or to simulate in commonly used physical simulators.

Extensive research has explored the manipulation of deformable objects, including 1-dimensional (1D) lines like ropes [7]–[9], 2-dimensional (2D) surfaces like clothes [10]–[13], and 3-dimensional (3D) volumetric bodies like plasticine [14]. In comparison, plastic bags [5] present even greater challenges. Geometrically, bags exhibit hollow 3D structures [15] with openings and often contain internal items, requiring more precise and fine-grained manipulation. Dynamically, their softer, highly compliant materials lead to less structural stability. For instance, achieving even simple goal configurations—such as an upright pose—can be difficult, as bags tend to gradually collapse under their own weight. Existing studies on bag manipulation predominantly operate cloth bags without handles focusing on simple tasks, such as bag opening and object insertion. In these settings, bags can be treated as a folded cloth, simplifying both robot manipulation and physical simulation. However, the problem

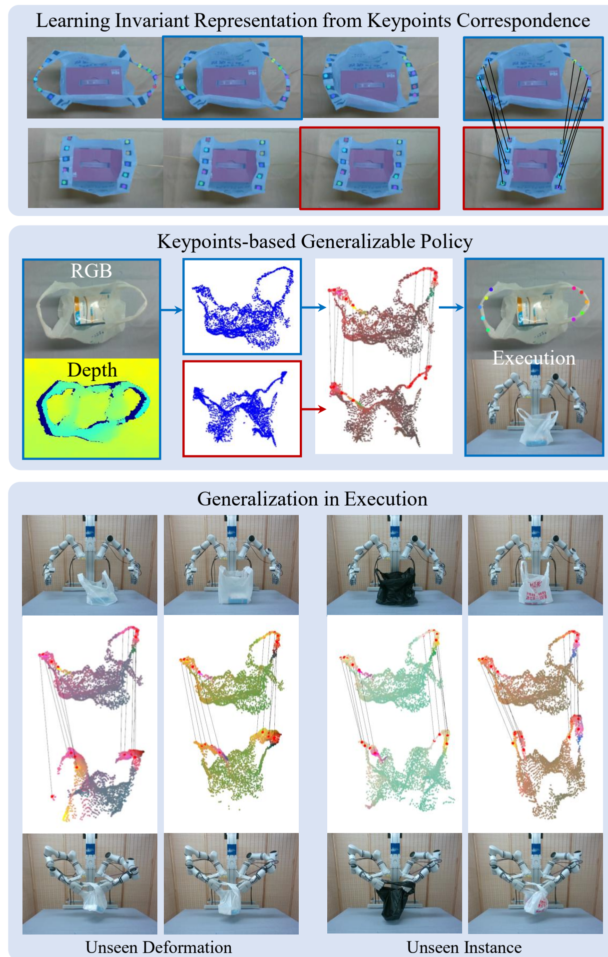

Fig. 1. **Overview of DexKnot. Top row**: Our framework collects keypoints correspondence data through real-world manual deformation, which are used to learn shape-agnostic representation of bags. **Middle row**: For a novel bag configuration, the keypoints are identified by correspondence matching, which guides the policy to execute the knotting task. **Bottom row**: Our framework generalize effectively to unseen deformations and bag instances.

of knotting plastic bags—particularly generalizing across diverse bag instances and initial deformations—remains largely unexplored.

Despite their diverse deformations and sizes, most plastic bags share consistent topological structures (i.e., handles and openings) that enable us to learn invariant representations. This structural consistency also enables us to capture key features that are essential for manipulation while ignoring irrelevant details, motivating our design of a low-dimensional representation scheme. Additionally, the significant sim-to-real gap necessitates a real-world data collection pipeline rather than relying on physical simulation.

With these insights, we present DexKnot, a real-world policy learning framework for generalizable bag knotting, which leverages shape-agnostic contrastive representation learning, keypoints identification through correspondence matching, and keypoints-based generalizable diffusion policy (Figure 1). We choose keypoints as representation because they reduce the dimensionality of the observation space, thus enhancing generalization especially when there are only a few demonstrations. The pipeline of our approach is as follows. First, we perform real-world manual deformation to collect keypoints correspondence data across various bag instances and deformations. Next, we train a PointNet++ [16] encoder to learn shape-agnostic representation of the bags' point clouds, allowing us to identify keypoints for an unseen bag configuration. The keypoints are taken as input by a diffusion transformer [17] (DiT), which generates robot joint angle sequences trained with a few human demonstrations.

We evaluated DexKnot's performance and generalization capacities through systematic experiments. The results show that DexKnot has high success rates on both seen and unseen deformations for various seen and unseen bag instances. Compared to 3D Diffusion Policy [18] (DP3), the state-of-the-art imitation learning framework, our approach demonstrates better generalization capacity on out-of-distribution deformations, such as twisted and inclined handle states.

The main contribution of this work is the development of DexKnot, a real-world framework for generalizable bag knotting task with a few demonstrations. Specifically,

- We propose an imitation learning framework leveraging keypoint representation to enable cross-instance and cross-deformation generalization.
- We develop a pipeline for keypoints correspondence data collection, using point tracking to avoid massive annotation and physical simulation.
- We conduct systematic experiments to demonstrate that DexKnot significantly outperforms existing strong baselines on the generalizable bag knotting task.

II. Related Work

A. Deformable Object Manipulation

A traditional line of work in deformable object manipulation is model-based methods [14], [19]–[25], which either build or learn a dynamics model of the object to manipulate. The dynamics models can predict the motion and deformation of objects subject to manipulation inputs, facilitating model predictive control (MPC) or model-based reinforcement learning (MBRL). Recently, significant advances have been made in end-to-end policy learning. Model-free reinforcement learning (MFRL) has demonstrated effectiveness in manipulating rope and cloth [26]–[28]. Imitation learning, especially diffusion policy [29] (DP), has also been applied in many relevant tasks, such as garments [13]. Compared to reinforcement learning (RL), DP is easier to train and more friendly for real-world data collection, which is thus chosen as our policy.

Recent advances in physical simulation have largely facilitated deformable object manipulation tasks, including ropes, cloth [26]–[28], [30], garments [13], tissues [31], and plasticine [32], [33]. Physical simulation is especially crucial for RL, which is expected to learn policies outperforming humans by scalable exploration in virtual environments. However, the sim-to-real gap remains a significant challenge, which is pronounced when objects are highly deformable.

B. Bag Manipulation

Compared to simple deformable objects, bags present more challenges for manipulation [2]–[4]. In terms of policy learning, the primary challenge of bag manipulation is generalization for initial deformations. The state-of-the-art policy learning methods like RL and DP [29] struggle to generalize with high-dimensional inputs but little data. To address this, there are some simple yet effective solutions, such as using airflow [34] or shaking a bag [35]. Another solution is iterative policy, which learns to adjust actions iteratively based on visual feedback to achieve precise goal conditions [5], [36]. Our approach leverages representation learning and diffusion policy, enabling generalization by extracting sparse manipulation-relevant keypoints as representation.

In terms of tasks, many works in bag manipulation focus on opening a bag or inserting objects into a bag [2], while less attention is on knotting a bag. However, knotting a bag is valuable with applications in many scenarios such as supermarket. Our method achieves the knotting task, while not involving any designs specific to knotting. This means our approach is general and can potentially adapt to other tasks.

C. Generalizable Visual Representations

Visual representation aims to encode invariant information across varying situations to facilitate downstream policy. The most straightforward approach is to simply feed RGB-D image into a U-Net, as employed in diffusion policy [29]. However, such dense representations often contain substantial irrelevant information that can distract the policy and impede generalization. Point clouds offer a sparser alternative that better captures spatial structure, and recent work on 3D diffusion policy [18] has demonstrated that using point clouds as visual representations can achieve strong performance and generalization.

Deformable objects exhibit an effectively infinite number of possible states, making it particularly challenging for dense representations to generalize effectively. In contrast, sparse keypoint representations [12], [37] can provide actionable affordance for downstream motion planning or policy learning, facilitating generalization by reducing the dimensionality of observation space. Correspondence matching offers a powerful method for identifying keypoints on novel instances, an approach that has proven effective in garment manipulation [12].

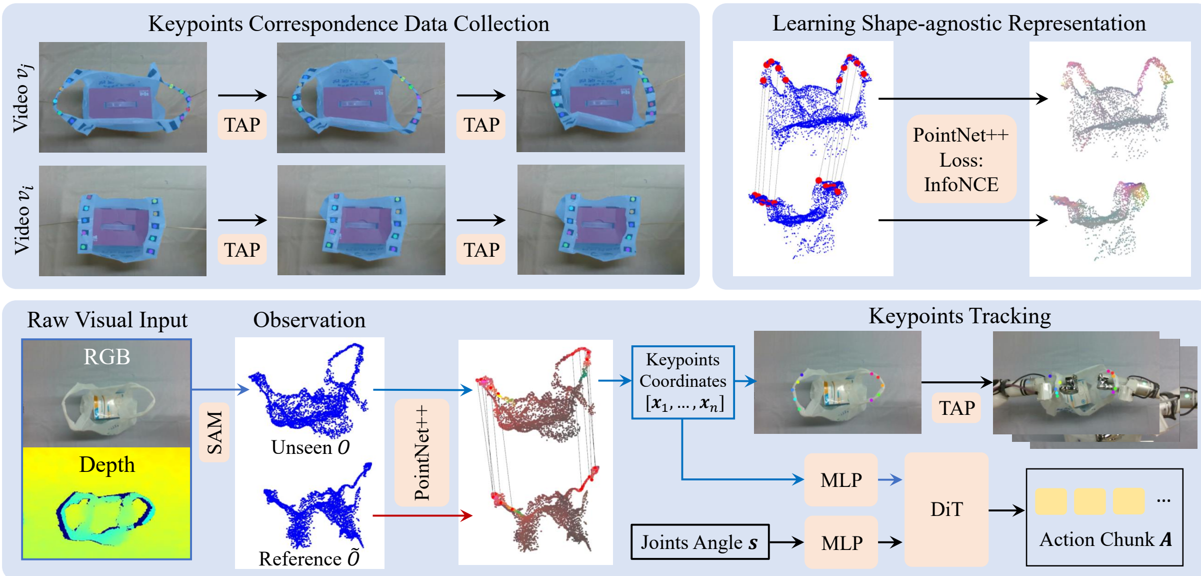

Fig. 2. **Our Proposed Framework. Top left**: For each bag, we perform manual deformation while recording RGB-D videos, and then we track the keypoints for correspondence data construction. **Top right**: The PointNet++ encoder learns to produce similar representations for corresponding keypoints across different deformations using an InfoNCE loss. **Bottom row**: During policy inference, keypoints are identified in the initial frame through representation matching and tracked across subsequent frames using TAP. These keypoint coordinates are combined with robot joint states and fed into a Diffusion Transformer (DiT) to generate an action chunk.

III. Method

A. Overview

Our framework addresses the challenge of generalizable bag knotting by combining representation learning with imitation learning. Despite the infinite degrees of freedom inherent to deformable objects, we leverage the topological consistency of plastic bags to learn a shape-agnostic representation. This approach enables identification of sparse keypoints through representation matching for novel bag configurations, significantly reducing observation space dimensionality and thus improving policy generalization.

As shown in Figure 2, our framework operates through three stages:

- **Correspondence Data Collection** (Top left): We perform manual deformation while recording RGB-D videos to capture diverse bag configurations. The keypoints are annotated and tracked to construct correspondence dataset.
- **Shape-Agnostic Representation Learning** (Top right): A PointNet++ encoder learns to produce similar representations for corresponding keypoints across different deformations using contrastive learning with InfoNCE loss, enabling keypoint identification for novel bag configurations.
- **Keypoints-Guided Generalizable Policy** (Bottom row): During inference, keypoints are identified through representation matching in the initial frame and tracked across subsequent frames. These coordinates combined with robot joint states and fed into separate MLPs followed by a Diffusion Transformer (DiT) to generate action chunks for manipulation.

This integrated approach enables effective generalization across diverse bag instances and deformation states by leveraging topological consistency while minimizing the observation space through sparse keypoint representation.

B. Correspondence Data Collection

We develop a pipeline for collecting keypoint correspondence data through real-world manual deformation that avoids both the inaccuracies of physical simulation and the burden of extensive manual annotation. Each bag is marked with n points on the handles, representing keypoints $p_{key}^{(1)}, ..., p_{key}^{(n)}$ that capture the essential topological structure. For each configuration (bag instance and initial deformation), we manually deform the bag while recording an RGB-D video v_i using our robot's head-mounted camera, where i denotes the serial number of the video. This process generates a diverse dataset covering various deformation patterns that would be difficult to achieve through data automation or simulation. To extract the pixel coordinates of the keypoints from v_i, we manually annotate keypoints only in the first frame of each video, then employ Track Any Point [38] (TAP) to propagate these annotations through subsequent frames. To segment the bag from the background, we use Segment Anything [39] (SAM) on the first frame and employ Cutie [40] for mask tracking across frames. The resulting data provides rich 3D information: we obtain precise 3D keypoint coordinates $\boldsymbol{x} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]$ by combining pixel coordinates with depth information, along with complete point cloud observations $\mathcal{O}$ containing n_{pc} points. Finally, we construct our correspondence dataset by randomly matching keypoints across all frames and videos with probability

p_m, creating positive pairs for contrastive learning while maintaining computational efficiency. Key hyperparameters are listed in Table I.

TABLE I
HYPERPARAMETERS IN THE ENCODER AND POLICY

Parameter	Value	Description
n	10	Number of Keypoints
n_{pc}	4096	Number of Points in Point Cloud
p_m	0.001	Probability of Matching
d	512	Encoder Feature Dimension
D	256	DiT Input Dimension
m	150	Number of Negative Point Samples
H	16	Action Chunk Horizon

C. Shape-Agnostic Representation Learning

We formulate the problem of learning deformation-invariant representations as a contrastive learning task that enforces consistency between corresponding keypoints across different bag configurations. This approach enables our system to recognize the same structural features regardless of how the bag is deformed or which specific instance is being manipulated. The core objective is to train a feature extractor F that produces identical representations for equivalent keypoints across different point cloud observations. Formally, given two point cloud observations $\mathcal{O}^{(1)}$ and $\mathcal{O}^{(2)}$ with corresponding keypoints $p^{(1)}_{key,i}$ and $p^{(2)}_{key,i}$, the representations $F(p^{(1)}_{key,i})$ and $F(p^{(2)}_{key,i}) \in \mathbb{R}^d$ extracted by the backbone network F should be identical. Here we implement F with a PointNet++ [16] network due to its ability to capture hierarchical spatial features from point clouds while maintaining permutation invariance. We normalize all extracted representations to unit vectors, so that we can measure the similarity by dot product: $F(p^{(1)}_{key,i}) \cdot F(p^{(2)}_{key,i})$. The learning framework follows a contrastive paradigm where for each anchor keypoint $p^{(1)}_{key,i}$ from $\mathcal{O}^{(1)}$, we consider the corresponding point $p^{(2)}_{key,i}$ from $\mathcal{O}^{(2)}$ as the positive sample, while randomly selecting m points $p^{(2)}_1, p^{(2)}_2, \dots, p^{(2)}_m$ from different locations in $\mathcal{O}^{(2)}$ as negative samples. This construction teaches the network to distinguish between equivalent keypoints other points. We use InfoNCE [41] as the loss function, which has proven effective in contrastive learning scenarios:

$$\mathcal{L} = -\log\left(\frac{\exp(F(p^{(1)}_{key,i}) \cdot F(p^{(2)}_{key,i})/\tau)}{\sum_{j=1}^{m} \exp(F(p^{(1)}_{key,i}) \cdot F(p^{(2)}_{j})/\tau)}\right), \quad (1)$$

The temperature τ modulates the sharpness of the similarity distribution, allowing control over how strongly the model distinguishes between similar and dissimilar pairs.

Given a novel bag configuration, we use SAM to obtain point cloud observation $\mathcal{O}$, and then identify keypoints using the trained encoder. For keypoint $p_{key,i}$, we compare all point representations $\{F(p_j)\}$ in $\mathcal{O}$ to $F(p^{(ref)}_{key,i})$ in reference observation $\mathcal{O}^{(ref)}$, selecting the most similar point in $\mathcal{O}$:

$$p_{key,i} = \mathrm{argmax}(F(p_j) \cdot F(p^{(ref)}_{key,i})) \quad (2)$$

D. Keypoints-Guided Generalizable Policy

Our policy is designed to leverage the keypoint for generalizable bag manipulation. The key insight is that by reducing the observation space to a compact set of geometrically meaningful keypoints, we can learn effective policies that generalize across diverse bag configurations while requiring minimal demonstration data. The policy operates on the identified keypoint coordinates $\boldsymbol{x}$ obtained through correspondence matching. To maintain temporal consistency and avoid reprocessing the entire point cloud at each step, we employ TAP for continuous keypoint tracking, producing updated coordinates $\boldsymbol{x}_t$ at each time step t. Combined with robot joint angle state $\boldsymbol{s}_t$ as input, the problem can be formulated as learning a policy π that effectively models the action distribution $\pi(\cdot|\boldsymbol{s}_t, \boldsymbol{x}_t)$.

We adopt an action-chunking approach with horizon H to improve temporal coherence and enable long-horizon reasoning. We map keypoint coordinates $\boldsymbol{x}_t$ and robot joint angle $\boldsymbol{s}_t$ into a common embedding space using separate MLPs, yielding $\boldsymbol{z}^x_t$ and $\boldsymbol{z}^s_t$. These embeddings are then concatenated to form full observation $\boldsymbol{z}^{\text{obs}}_t \in \mathbb{R}^{2\times D}$.

For action generation, we use a Diffusion Transformer [17] to generate multi-step actions following diffusion policy paradigm [29], [42], [43]. At each time step t, we bundle next H actions into a chunk $\boldsymbol{A}_t = \boldsymbol{a}_{t:t+H} = [\boldsymbol{a}_t, \boldsymbol{a}_{t+1}, \dots, \boldsymbol{a}_{t+H-1}]$. During training, we sample random diffusion step $t_d = k$, and then add Gaussian noise $\boldsymbol{\epsilon}$ to $\boldsymbol{A}_t$ to get the noised action tokens $\tilde{\boldsymbol{A}}_k = \alpha_k \boldsymbol{A}_t + \sigma_k \boldsymbol{\epsilon}$, where α_k and σ_k are the standard DDPM coefficients. Next, we feed $\tilde{\boldsymbol{A}}_k$ into DiT with observation feature $\boldsymbol{z}^{\text{obs}}_t$. Each DiT layer performs bidirectional self-attention over action tokens, cross-attention to $\boldsymbol{z}^{\text{obs}}_t$, and MLP transformations, predicting original noise $\boldsymbol{\epsilon}$. By minimizing the discrepancy between the predicted and true noise, the model learns to reconstruct the ground-truth action chunk $\boldsymbol{A}_t$. At inference time, iterative denoising steps recover the intended multi-step action sequence from the learned distribution.

IV. EXPERIMENTS

A. Experiment Setup

Robot Platform. All data collection and evaluation experiments are conducted using RealMan RM75-6F dual-arm robot equipped with PsiBot G0-R 6-DoF dexterous hands and a head-mounted Intel RealSense D435 RGB-D camera (Figure 3). Actions are recorded and executed at approximately 10 Hz.

Deformation State Definitions. We define five distinct deformation states to standardize data collection and evaluation (Figure 4):

- Vertical-Compressed (VC): The handles are oriented vertically and in a compressed rope-like state.
- Horizontal-Compressed (HC): The handles are oriented horizontally and in a compressed rope-like state.
- Diagonal-Compressed (DC): The handles are oriented diagonally and compressed into a rope-like state, which

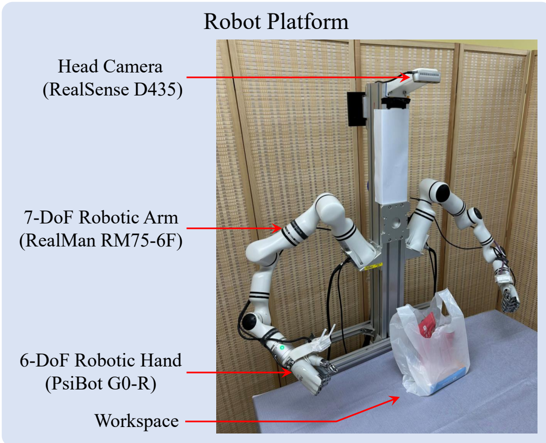

Fig. 3. **Robot setup**. Our robot platform includes a RealMan RM75-6F dual-arm with PsiBot G0-R 6-DoF dexterous hands and a head-mounted Intel RealSense D435 RGB-D camera.

can be considered as an interpolated state of VC and HC.

- Twisted-Flat (TF): The handles are twisted inward and splayed flat.
- Inclined-Flat (IF): The handles lean to one side and splayed flat.

These deformation states are consistently used across all data collection and evaluation procedures.

Data. For keypoints correspondence, we use six plastic bags of varying sizes and shapes, each marked with $n = 10$ keypoints on handles (Figure 5, top row). Note that there are some additional markers on the opening of some bags, which are not used as keypoints. Two experimenters manually deform each bag while the head-mounted camera records the manual deformation process.

For behavior demonstrations, we use three bags: two are new bags but of the types present in the correspondence data and one completely novel type (Figure 5, bottom left). All bags used in this stage have no markers. A knotting action involves four stages: threading handles; hook the left inner handle with the right index finger and thumb; hook the right outer handle with the left index finger and thumb; tightening the knot. We collected 54 human demonstration trajectories, each comprising 160 action steps, across two initial deformation states (Figure 4, top row):

Evaluation Protocol. We evaluate generalization across initial deformations and bag instances. For cross-deformation generalization, each bag is evaluated in five states: VC and HC, which are present in the demonstrations; DV, TF, and IF, which are not present in the demonstrations (Figure 4, bottom row). For cross-instance generalization, policies are tested on three bags present in behavior demonstrations (Figure 5, bottom left) and three novel bags not present in behavior demonstrations (Figure 5, bottom right).

Metric. We report success rates as the number of successful trials divided by the total attempts across all test conditions.

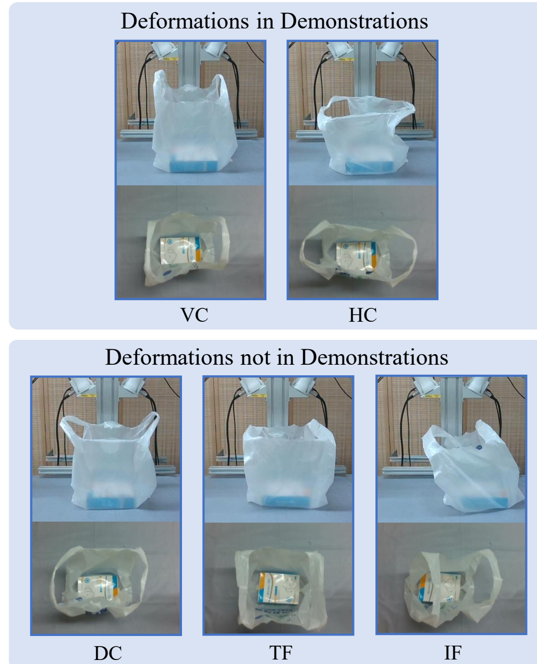

Fig. 4. **Bag deformations in data collection and evaluation**. **Top row**: Deformations included in behavior demonstrations. **Bottom row**: Deformations not included in behavior demonstrations.

Baselines. We compare against state-of-the-art imitation learning approaches and vision-language-action model:

- **DP:** Standard Diffusion Policy [29] trained on our data with raw RGB images as input.
- **DP3:** 3D Diffusion Policy trained on our data with bag point clouds as input.
- π_0**:** Vision-Language-Action model (VLA) π_0 [44] fine-tuned on our data.

B. Generalization Evaluation

We evaluate the generalization capability of our approach against baseline methods across five initial deformations for bag instances seen and unseen by the demonstrations. We note that the π_0 approach performs poorly even on seen bags and often results in hand collisions, so we omit further tests on unseen bags. The standard DP approach demonstrates limited performance either, likely due to the high dimensionality of raw RGB input and the absence of depth information. Consequently, our primary comparative analysis focuses on DP3, which serves as our main baseline due to its strong performance.

Table II shows the success rates of DexKnot and baselines on seen bag instances. For VC and HC (seen deformations) and DC (interpolation deformation), both DP3 and DexKnot achieve high success rates. For TF and IF (out-of-distribution deformations), our approach significantly outperforms DP3, demonstrating better generalization to novel deformations. The results can be explained as follows: The flatten handles are never seen by DP3's encoder, thus leading to wrong behavior of the policy; In contrast, the keypoints on the handles

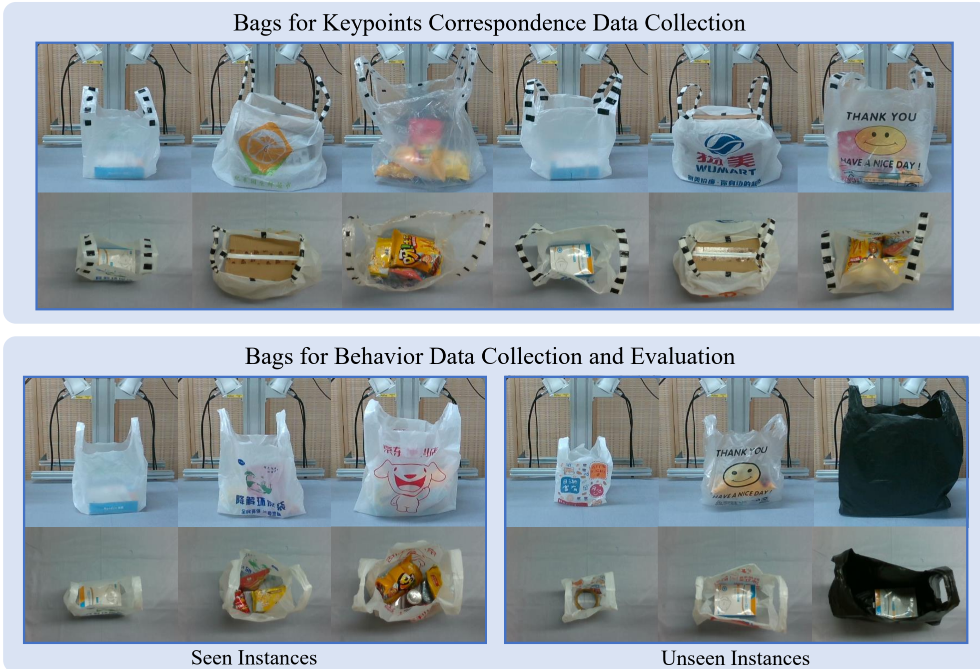

Fig. 5. **Bag instances**. **Top row**: bags used for keypoint correspondence data collection. **Bottom left**: bags used for behavior demonstration data collection. **Bottom right**: novel bags for cross-instance evaluation, not included in keypoints correspondence data or behavior demonstrations.

can still be identified by DexKnot's encoder since it has been pretrained with such states during manual deformation. This performance gap is particularly evident for the IF case: The point cloud deviates a lot from the training data, which cannot be handled by DP3; In contrast, DexKnot can still identify the keypoints, enabling the policy to perform the task.

TABLE II
RESULTS ON SEEN BAGS ACROSS DEFORMATIONS

Methods	VC & HC	DC	TF	IF
DP	3/18	2/9	1/9	2/9
DP3	**17/18**	**9/9**	2/9	0/9
π_0	1/18	0/9	1/9	0/9
Ours	16/18	8/9	**8/9**	**4/9**

Table III shows the success rates of DexKnot and baseline methods for bag instances that are never present in the correspondence data or behavior demonstrations. While all methods exhibit reduced performance compared to seen instances, DexKnot significantly outperforms DP3 across all initial deformations, particularly excelling in twisted and inclined cases. These results demonstrate that our approach not only generalizes better to novel deformations but also maintains more consistent performance when presented with unseen instances.

Figure 6 shows qualitative comparisons between DP3 and our approach for a seen bag in three initial deformations. While both methods successfully completed the knotting task under DC configuration, DP3 failed to identify handle locations in TF and IF configurations, leading to the failure of the task. In contrast, DexKnot maintained robust performance across all deformations.

TABLE III
RESULTS ON UNSEEN BAGS ACROSS DEFORMATIONS

Methods	VC & HC	DC	TF	IF
DP	4/18	1/9	0/9	0/9
DP3	14/18	6/9	1/9	0/9
Ours	**15/18**	**8/9**	**6/9**	**4/9**

Our results indicate that while DP3 and DexKnot show comparable performance on seen bag instances, seen deformations, and simple interpolated deformation, our approach demonstrates significantly better generalization when presented with novel deformation patterns, particularly those involving complex spatial transformations such as inclination and twisting.

C. Ablation Studies

To evaluate the contribution of key components in DexKnot, we conducted ablation studies comparing our full framework against two ablated versions on unseen bag instances:

- **Ours w/o TF/IF:** This variant removes exposure to twisted and inclined deformations during the encoder's training phase, testing the importance of diverse manual deformations for learning shape-agnostic representation.
- **Ours w/o TAP:** This variant replaces the TAP-based keypoint tracking with an alternative approach: using Cutie to track the bag's mask and identifying the keypoints by the encoder at each step.

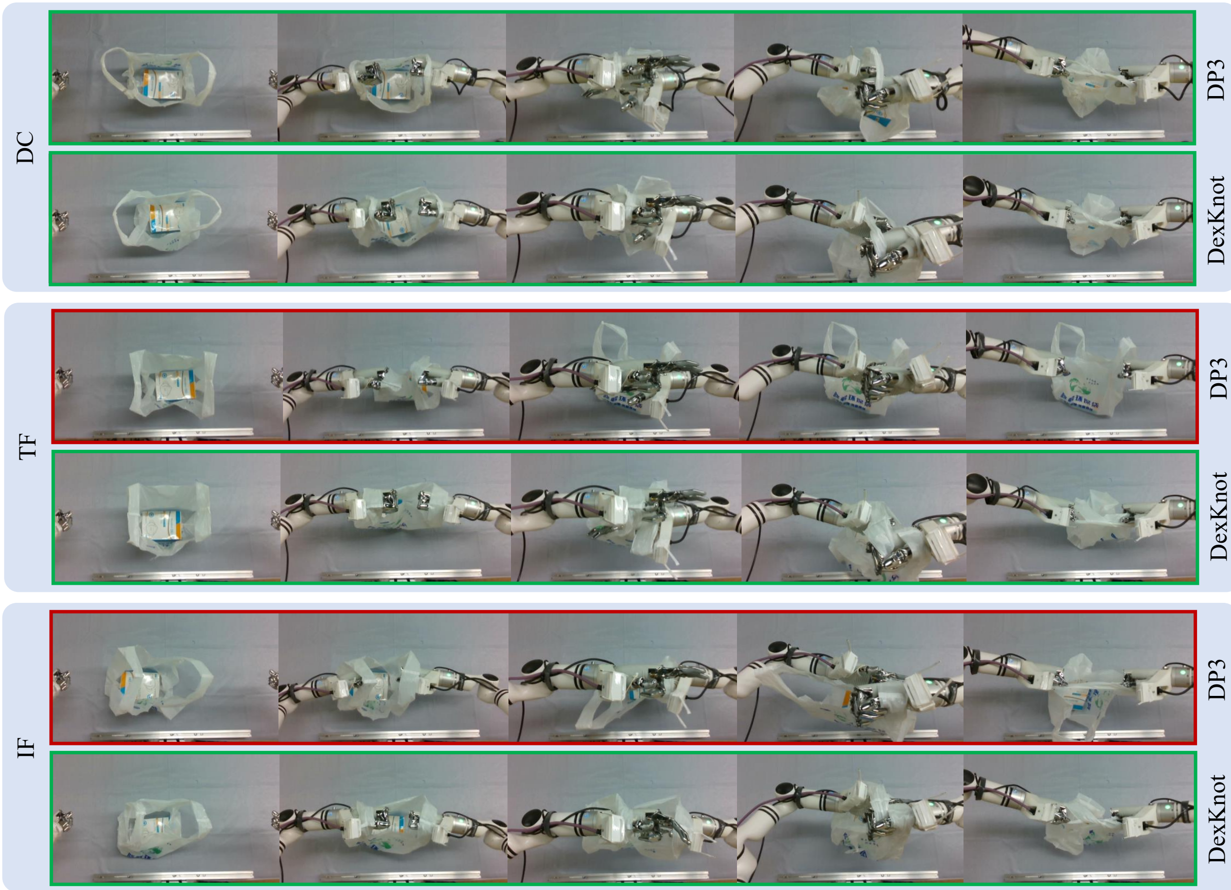

Fig. 6. **Qualitative comparison of policy executions**. Successes and failures are indicated by green and red bounding boxes, respectively. **Top row**: Both DP3 and DexKnot successfully complete the knotting task under Diagonal-Compressed (DC) deformation conditions. **Middle row**: In Twisted-Flat (TF) conditions, DP3 fails to thread the handle while DexKnot successfully accomplishes the task. **Bottom row**: In Inclined-Flat (IF) conditions, DP3 fails to thread the handle while DexKnot successfully accomplishes the task.

As quantitatively demonstrated in Table IV, both ablated versions show performance degradation across all deformations compared to the full method. The performance drop in **Ours w/o TF/IF** indicates that training the encoder on a diverse set of deformations is crucial for learning shape-agnostic representations that enables generalization in the downstream policy. The inferior results of **Ours w/o TAP** indicates that identifying keypoints initially and then tracking them provides more reliable state estimation than tracking the mask and identifying the keypoints in each frame. These results validate the importance of each component in our complete framework.

TABLE IV

ABLATION STUDY RESULTS ON UNSEEN BAGS

Methods	VC & HC	DC	TF	IF
Ours w/o TAP	13/18	7/9	5/9	**4/9**
Ours w/o TF/IF	**17/18**	7/9	1/9	**4/9**
Ours	15/18	**8/9**	**6/9**	**4/9**

V. CONCLUSION

We have presented DexKnot, a framework that integrates self-supervised keypoint correspondence learning with diffusion policies for generalizable bag knotting. By learning a shape-agnostic representation, our approach encodes crucial manipulation information into a sparse set of keypoints. This strategy dramatically reduces the observation space dimensionality while preserving essential structure for the task, enabling robust generalization to both unseen initial deformations and bag instances. Experimental results demonstrate superior performance over baseline methods, particularly for out-of-distribution deformations.

Despite the advantages, DexKnot also has some limitations. First, although our correspondence data collection pipeline significantly reduces manual effort by requiring only first-frame annotations, the initial annotation requirement remains a notable limitation. Second, the keypoint representation's low dimensionality, while beneficial for generalization, introduces a vulnerability to misidentification errors. This represents an inherent trade-off between representations' sparsity and robustness that warrants further investigation.

REFERENCES

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

[2] L. Y. Chen, B. Shi, D. Seita, R. Cheng, T. Kollar, D. Held, and K. Goldberg, "Autobag: Learning to open plastic bags and insert objects," *arXiv preprint arXiv:2210.17217*, 2022.

[3] L. Y. Chen, B. Shi, R. Lin, D. Seita, A. Ahmad, R. Cheng, T. Kollar, D. Held, and K. Goldberg, "Bagging by learning to singulate layers using interactive perception," in *2023 IROS*. IEEE, 2023, pp. 3176–3183.

[4] A. Bahety, S. Jain, H. Ha, N. Hager, B. Burchfiel, E. Cousineau, S. Feng, and S. Song, "Bag all you need: Learning a generalizable bagging strategy for heterogeneous objects," in *2023 IROS*. IEEE, 2023, pp. 960–967.

[5] C. Gao, Z. Li, H. Gao, and F. Chen, "Iterative interactive modeling for knotting plastic bags," in *Conference on Robot Learning*. PMLR, 2023, pp. 571–582.

[6] H. Geng, Z. Li, Y. Geng, J. Chen, H. Dong, and H. Wang, "Partmanip: Learning cross-category generalizable part manipulation policy from point cloud observations," in *CVPR*, 2023, pp. 2978–2988.

[7] A. Nair, D. Chen, P. Agrawal, P. Isola, P. Abbeel, J. Malik, and S. Levine, "Combining self-supervised learning and imitation for vision-based rope manipulation," in *2017 ICRA*. IEEE, 2017, pp. 2146–2153.

[8] P. Sundaresan, J. Grannen, B. Thananjeyan, A. Balakrishna, M. Laskey, K. Stone, J. E. Gonzalez, and K. Goldberg, "Learning rope manipulation policies using dense object descriptors trained on synthetic depth data," in *2020 ICRA*. IEEE, 2020, pp. 9411–9418.

[9] K. Suzuki, M. Kanamura, Y. Suga, H. Mori, and T. Ogata, "In-air knotting of rope using dual-arm robot based on deep learning," in *2021 IROS*. IEEE, 2021, pp. 6724–6731.

[10] L. Y. Chen, H. Huang, E. Novoseller, D. Seita, J. Ichnowski, M. Laskey, R. Cheng, T. Kollar, and K. Goldberg, "Efficiently learning single-arm fling motions to smooth garments," in *The International Symposium of Robotics Research*. Springer, 2022, pp. 36–51.

[11] T. Weng, S. M. Bajracharya, Y. Wang, K. Agrawal, and D. Held, "Fabricflownet: Bimanual cloth manipulation with a flow-based policy," in *Conference on Robot Learning*. PMLR, 2022, pp. 192–202.

[12] R. Wu, H. Lu, Y. Wang, Y. Wang, and H. Dong, "Unigarmentmanip: A unified framework for category-level garment manipulation via dense visual correspondence," in *CVPR*, 2024, pp. 16 340–16 350.

[13] Y. Wang, R. Wu, Y. Chen, J. Wang, J. Liang, Z. Zhu, H. Geng, J. Malik, P. Abbeel, and H. Dong, "Dexgarmentlab: Dexterous garment manipulation environment with generalizable policy," *arXiv preprint arXiv:2505.11032*, 2025.

[14] H. Shi, H. Xu, Z. Huang, Y. Li, and J. Wu, "Robocraft: Learning to see, simulate, and shape elasto-plastic objects in 3d with graph networks," *The International Journal of Robotics Research*, vol. 43, no. 4, pp. 533–549, 2024.

[15] D. Seita, P. Florence, J. Tompson, E. Coumans, V. Sindhwani, K. Goldberg, and A. Zeng, "Learning to rearrange deformable cables, fabrics, and bags with goal-conditioned transporter networks," in *2021 ICRA*. IEEE, 2021, pp. 4568–4575.

[16] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[17] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4195–4205.

[18] 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.

[19] Y. Li, J. Wu, R. Tedrake, J. B. Tenenbaum, and A. Torralba, "Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids," *arXiv preprint arXiv:1810.01566*, 2018.

[20] W. Yan, A. Vangipuram, P. Abbeel, and L. Pinto, "Learning predictive representations for deformable objects using contrastive estimation," in *Conference on Robot Learning*. PMLR, 2021, pp. 564–574.

[21] H. Shi, H. Xu, S. Clarke, Y. Li, and J. Wu, "Robocook: Long-horizon elasto-plastic object manipulation with diverse tools," *arXiv preprint arXiv:2306.14447*, 2023.

[22] Y. Wang, Y. Li, K. Driggs-Campbell, L. Fei-Fei, and J. Wu, "Dynamic-resolution model learning for object pile manipulation," *arXiv preprint arXiv:2306.16700*, 2023.

[23] C. Li, Z. Ai, T. Wu, X. Li, W. Ding, and H. Xu, "Deformnet: Latent space modeling and dynamics prediction for deformable object manipulation," in *2024 ICRA*. IEEE, 2024, pp. 14 770–14 776.

[24] D. Bauer, Z. Xu, and S. Song, "Doughnet: A visual predictive model for topological manipulation of deformable objects," in *European Conference on Computer Vision*. Springer, 2024, pp. 92–108.

[25] K. Zhang, B. Li, K. Hauser, and Y. Li, "Adaptigraph: Material-adaptive graph-based neural dynamics for robotic manipulation," *arXiv preprint arXiv:2407.07889*, 2024.

[26] J. Matas, S. James, and A. J. Davison, "Sim-to-real reinforcement learning for deformable object manipulation," in *Conference on Robot Learning*. PMLR, 2018, pp. 734–743.

[27] Y. Wu, W. Yan, T. Kurutach, L. Pinto, and P. Abbeel, "Learning to manipulate deformable objects without demonstrations," *arXiv preprint arXiv:1910.13439*, 2019.

[28] X. Lin, Y. Wang, J. Olkin, and D. Held, "Softgym: Benchmarking deep reinforcement learning for deformable object manipulation," in *Conference on Robot Learning*. PMLR, 2021, pp. 432–448.

[29] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, p. 02783649241273668, 2023.

[30] R. Laezza, R. Gieselmann, F. T. Pokorny, and Y. Karayiannidis, "Reform: A robot learning sandbox for deformable linear object manipulation," in *2021 ICRA*. IEEE, 2021, pp. 4717–4723.

[31] X. Liang, F. Liu, Y. Zhang, Y. Li, S. Lin, and M. Yip, "Real-to-sim deformable object manipulation: Optimizing physics models with residual mappings for robotic surgery," in *2024 ICRA*. IEEE, 2024, pp. 15 471–15 477.

[32] Z. Huang, Y. Hu, T. Du, S. Zhou, H. Su, J. B. Tenenbaum, and C. Gan, "Plasticinelab: A soft-body manipulation benchmark with differentiable physics," *arXiv preprint arXiv:2104.03311*, 2021.

[33] S. Li, Z. Huang, T. Chen, T. Du, H. Su, J. B. Tenenbaum, and C. Gan, "Dexdeform: Dexterous deformable object manipulation with human demonstrations and differentiable physics," *arXiv preprint arXiv:2304.03223*, 2023.

[34] Z. Xu, C. Chi, B. Burchfiel, E. Cousineau, S. Feng, and S. Song, "Dextairity: Deformable manipulation can be a breeze," *arXiv preprint arXiv:2203.01197*, 2022.

[35] N. Gu, Z. Zhang, R. He, and L. Yu, "Shakingbot: dynamic manipulation for bagging," *Robotica*, vol. 42, no. 3, pp. 775–791, 2024.

[36] C. Chi, B. Burchfiel, E. Cousineau, S. Feng, and S. Song, "Iterative residual policy: for goal-conditioned dynamic manipulation of deformable objects," *The International Journal of Robotics Research*, vol. 43, no. 4, pp. 389–404, 2024.

[37] J. Grannen, P. Sundaresan, B. Thananjeyan, J. Ichnowski, A. Balakrishna, M. Hwang, V. Viswanath, M. Laskey, J. E. Gonzalez, and K. Goldberg, "Untangling dense knots by learning task-relevant keypoints," *arXiv preprint arXiv:2011.04999*, 2020.

[38] C. Doersch, Y. Yang, M. Vecerik, D. Gokay, A. Gupta, Y. Aytar, J. Carreira, and A. Zisserman, "TAPIR: Tracking any point with per-frame initialization and temporal refinement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 061–10 072.

[39] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[40] H. K. Cheng, S. W. Oh, B. Price, J.-Y. Lee, and A. Schwing, "Putting the object back into video object segmentation," in *CVPR*, 2024, pp. 3151–3161.

[41] Z. Lai, S. Purushwalkam, and A. Gupta, "The functional correspondence problem," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 772–15 781.

[42] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu, "Rdt-1b: a diffusion foundation model for bimanual manipulation," *arXiv preprint arXiv:2410.07864*, 2024.

[43] Y. Zhong, X. Huang, R. Li, C. Zhang, Z. Chen, T. Guan, F. Zeng, K. N. Lui, Y. Ye, Y. Liang *et al.*, "Dexgraspvla: A vision-language-action framework towards general dexterous grasping," *arXiv preprint arXiv:2502.20900*, 2025.

[44] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, "pi0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.