2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

DiffDef: A Diffusion Model for Generating Multimodal Goal Shapes From Demonstrations for Deformable Object Manipulation

Bao Thach[1,§], Tanner Watts[3,§], Siyeon Kim[1], Britton Jordan[1], Mohanraj Devendran Shanthi[1], Shing-Hei Ho[1], James M. Ferguson[1], Tucker Hermans[1,2], and Alan Kuntz[3]

***Abstract*— Deformable object manipulation is a key capability in many robotic applications. A promising paradigm for this problem is shape servoing, which aims to control deformable objects toward desired goal shapes. However, existing approaches typically rely on impractical goal-shape acquisition methods, such as domain-knowledge engineering or manual manipulation. Moreover, prior methods generally assume a single deterministic goal and fail to handle multimodal goal settings, a common scenario in many real-world tasks where multiple distinct goal shapes can all lead to successful task completion. In this paper, we introduce *DiffDef*, a novel neural network that uses a diffusion model to learn a distribution of feasible goal shapes rather than predicting a single deterministic outcome. This allows *DiffDef* to generate diverse goal configurations while avoiding the mode-averaging artifacts common in deterministic predictors. We evaluate our method on several deformable manipulation tasks inspired by manufacturing and surgical applications, both in simulation and on two physical robotic platforms: the da Vinci Research Kit (dVRK) and a bimanual KUKA-based robotic system. The results demonstrate that *DiffDef* effectively captures multimodal goal distributions and significantly improves task performance in practical robotic settings. Website: sites.google.com/view/diffdef.**

I. Introduction

Deformable object manipulation is a central problem in robotics, with applications in surgery, manufacturing, and everyday household environments. Unlike rigid bodies, deformable objects exhibit effectively infinite degrees of freedom and complex dynamics [1–3], making it difficult to explicitly define task goals or learn policies that generalize across contexts. A key limitation of prior approaches [4–9] is their reliance on manually specified geometric goal shapes [10]. The recent state-of-the-art method, *DefGoalNet* [10], addresses this limitation by learning to generate goal shapes directly from human demonstrations.

However, *DefGoalNet* predicts only a single deterministic goal shape and struggles to represent multimodal goal distributions. Such multimodal settings are common and critical in real-world robotic tasks. Multimodality occurs when multiple distinct shapes can equally accomplish a task under the same conditions. For example, a surgical robot may deform tissue either to the left or to the right, both resulting in successful outcomes in a tissue retraction task (Figure 1). In such cases, deterministic models like *DefGoalNet* "average" these possibilities (Figure 2), producing an ambiguous and physically infeasible goal shape that often leads to task failure.

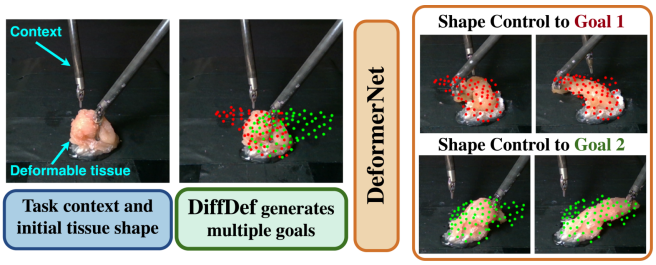

Fig. 1: Multimodal Contextual Shape Servoing: Given the task context and initial shape of the deformable object, *DiffDef* predicts multiple valid goal point clouds (red and green points) by learning a diffusion-based probabilistic model from human demonstrations. Conditioned on one of these predicted goals, *DeformerNet* then computes actions to manipulate the object into the desired shape and successfully accomplishes the task. We experiment on a surgical retraction task, where the robot is tasked with pulling the tissue aside to create tension without colliding with the surgical tool.

To address this limitation, we introduce *DiffDef*, a novel neural network that models multimodal goal distributions by generating goal shapes conditioned on geometric context via a conditional diffusion process [11]. As part of our learning-from-demonstration shape servoing pipeline, *DiffDef* is trained on demonstration trajectories collected from diverse policies (e.g., different human demonstrators performing the same task in distinct ways). This enables the model to capture the full distribution of goal shapes that successfully accomplish a task given the current context. At runtime, *DiffDef* takes as inputs the current deformable object point cloud and a contextual point cloud, and generates a goal shape sampled from the learned distribution. The predicted goal is then passed to *DeformerNet* [12, 13], which computes robot actions to deform the object toward the goal shape, thereby successfully completing the task.

We evaluate our approach in simulation and in real-world experiments on both surgical (tissue retraction) and manufacturing (object packaging) tasks. Across all tasks, the multimodal goal shapes generated by *DiffDef* consistently achieve high success rates, outperforming the state-of-the-art goal generation method *DefGoalNet*. By introducing a generative, multimodal formulation for deformable object goals, our approach brings robotic manipulation of soft

§ These authors contributed equally; bao.thach@utah.edu. [1]Kahlert School of Computing, University of Utah, Salt Lake City, UT, USA; [2]NVIDIA Corporation, Seattle, WA, USA; [3]Department of Electrical and Computer Engineering and Department of Computer Science, Vanderbilt University, Nashville, TN, USA. Research reported in this publication was supported by the Advanced Research Projects Agency for Health (ARPA-H) under the ALISS project, Award Number D24AC00415-00. The ARPA-H award of up to $11,935,038 provided 100% of the financial support for this work. The opinions and findings in this paper is solely the responsibility of the authors and do not necessarily represent the official views of ARPA-H.

materials closer to practical real-world applications.

II. Related Work

Recent advances in machine learning have empowered robots to manipulate rigid objects by effectively leveraging rich, high-dimensional sensory inputs such as 3D point clouds [14–18]. Neural networks, in particular, have become central to tackling complex robotic perception and control problems, including shape reconstruction [18], object pose estimation [17], and grasp synthesis [14, 17, 18]. Furthermore, sophisticated learning frameworks have enabled robots to execute long-horizon, multistep tasks by integrating a diverse set of learned behaviors [19]. Building on these successes, we explore a learning-based framework aimed at controlling the 3D shape of deformable objects.

Traditionally, shape servoing has been addressed through predominantly model-based or learning-free techniques [5–7, 9, 20]. These methods typically rely on manually selected feature points to represent the deformable object, making them susceptible to sensor noise and limiting their ability to generalize to novel object geometries. Shetab-Bushehri *et al.* [9] propose a 3D lattice-based representation that enables accurate 3D control of deformable shapes. However, their method relies on the assumption of consistent feature correspondences over time, which is difficult to uphold in real-world robotic deployments.

Learning-based approaches have emerged as a promising direction for 3D shape control by leveraging the generalization strength of modern neural networks. Hu *et al.* [21] utilize Fast Point Feature Histograms [22] to represent the state of deformable objects within a learning framework. However, this architecture struggles to model the complex dynamics of 3D deformable objects [13]. The current state-of-the-art method, *DeformerNet* [12, 13], addresses this issue by employing a PointConv-based [23] neural network that inputs the current and goal point clouds and predicts the robot action to deform the object toward the goal shape. However, all of these methods rely on explicitly defined goal shapes, a key limitation precluding real-world applications.

Although point cloud generative models have achieved impressive results in computer vision and graphics [24–27], their integration into robotic applications has so far been relatively underexplored.

Recent methods in robotic goal generation synthesize images from language instructions [28, 29], but 2D goals cannot capture the complex geometries required for deformable object control. Text-only goals [30, 31] also fall short, as they lack the geometric details needed for 3D shape servoing.

For learning robotic policies from multimodal demonstration datasets, diffusion-based models such as Diffusion Policy [32] and Stein Variational Belief Propagation [33] directly map observations to actions, yet they demand costly demonstrations and large-scale datasets. Our approach instead predicts multimodal goal representations, leaving action execution to *DeformerNet*. This separation improves sample efficiency: action data can be generated at low cost in simulation, while expensive demonstration data is required in much smaller quantities for goal learning [10].

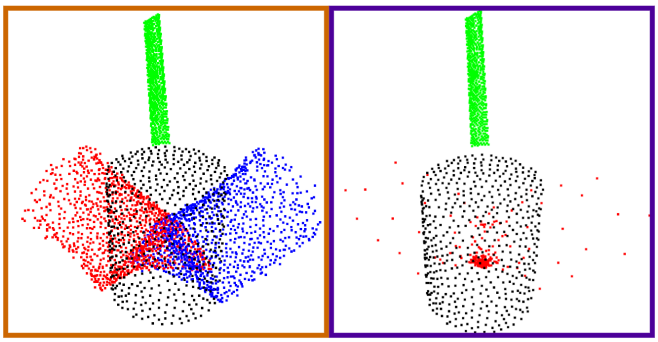

Fig. 2: Simulated retraction results comparing predicted goals from *DiffDef* (left) to *DefGoalNet* (right). *DiffDef* (this paper) effectively captures the underlying bimodal goal distribution, producing two distinct, realistic goal shapes (red and blue). In contrast, *DefGoalNet* (current state-of-the-art) averages the two modes, resulting in a single unrealistic and physically infeasible goal point cloud (red).

III. Problem Formulation

We study robotic manipulation tasks involving deformable objects where success can be achieved through multiple distinct strategies. For example, in tissue retraction, pulling tissue either to the left or right may both generate the necessary exposure of underlying anatomy. In such settings, the robot must reason not about a single goal state, but about a distribution of feasible goal shapes.

Formally, let $\mathcal{P}_c$ denote the current partial-view point cloud of the deformable object, and let $\mathcal{P}_\tau$ represent a contextual point cloud capturing task-relevant environmental features. Together, $(\mathcal{P}_c, \mathcal{P}_\tau)$ define the scene state. Given a task τ, our objective is to model a conditional distribution

$$p(\mathcal{P}_g \mid \mathcal{P}_c, \mathcal{P}_\tau), \tag{1}$$

where $\mathcal{P}_g$ is a feasible goal point cloud that would lead to successful task completion of τ.

We assume access to a multimodal expert demonstration dataset $\mathcal{D}$, where each trajectory contains sequences of object and contextual point clouds illustrating different but equally valid strategies for completing τ. The multimodality of $\mathcal{D}$ ensures that $p(\mathcal{P}_g \mid \mathcal{P}_c, \mathcal{P}_\tau)$ reflects the diversity of successful outcomes rather than collapsing them into a single average goal.

At execution time, the robot samples $\mathcal{P}_g \sim p(\mathcal{P}_g \mid \mathcal{P}_c, \mathcal{P}_\tau)$ and uses a downstream policy (*DeformerNet* [12, 13]) to compute the actions that deform the object toward the sampled goal. We formulate robot action $\mathcal{A}$ as rigid transformation $\mathcal{SE}(3)$, representing the change of robot end-effector's pose.

IV. Methods

Our framework decomposes deformable object manipulation into two subproblems: (i) **goal generation**, where the robot predicts a distribution over feasible goal shapes given the current state and task context, formalized as $p_\theta(\mathcal{P}_g \mid \mathcal{P}_c, \mathcal{P}_\tau)$; and (ii) **goal-conditioned control**, where the robot executes actions to deform the object toward a sampled

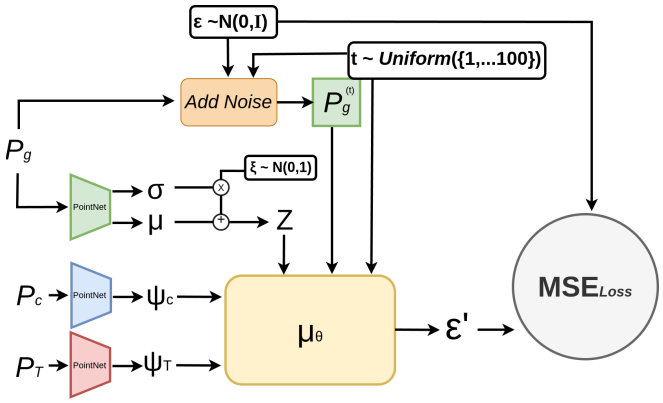

Fig. 3: *DiffDef* architecture for learning multimodal goal distributions. P_g: goal point cloud; P_c: current point cloud; P_T: context. Encoders produce latent representations (z, ψ_c, ψ_T). The noise predictor μ_θ estimates ϵ' to match ground-truth noise ϵ.

goal shape. This separation improves sample efficiency: expensive demonstrations are required only for goal learning, while control policy supervision can be collected at scale in simulation. Since *DeformerNet* [12] already provides a robust solution for goal-conditioned control, our focus in this section is on solving the goal learning problem via *DiffDef*.

A. Conditional Goal Distribution Learning

We represent the set of all valid outcomes for a task as a conditional distribution over goal point clouds as in Equation 1, where $\mathcal{P}_\mathrm{c}$ is the current partial-view object point cloud and $\mathcal{P}_\tau$ encodes task-relevant context. Unlike prior deterministic goal networks [10], this formulation captures the inherent multimodality of deformable manipulation tasks.

To parameterize $p_\theta(\mathcal{P}_\mathrm{g} \mid \mathcal{P}_\mathrm{c}, \mathcal{P}_\tau)$, we adopt a *conditional diffusion model* [11]. Given a dataset $\mathcal{D}$ of demonstrations with diverse, successful strategies, we add noise to each goal point cloud $\mathcal{P}_\mathrm{g}$ through the forward diffusion process, and train a noise predictor μ_θ to recover the injected noise conditioned on $(\mathcal{P}_\mathrm{c}, \mathcal{P}_\tau)$. At inference, we generate diverse samples $\mathcal{P}_\mathrm{g} \sim p_\theta$ by reversing this diffusion process.

B. DiffDef Architecture

Figure 3 summarizes our model architecture. A PointNet [34] encoder maps the goal point cloud $\mathcal{P}_\mathrm{g}$ into Gaussian parameters (μ, σ), from which we sample a latent vector z. Two additional PointNet encoders map the current and contextual point clouds $(\mathcal{P}_\mathrm{c}, \mathcal{P}_\tau)$ into feature vectors (ψ_c, ψ_τ). A noisy version of the goal $\mathcal{P}_\mathrm{g}^{(t)}$ is obtained via the forward diffusion process [11, 35] for training supervision, where t denotes the diffusion timestep. Finally, the *noise predictor* μ_θ takes as input $(\mathcal{P}_\mathrm{g}^{(t)}, t, z, \psi_c, \psi_\tau)$ and outputs the predicted noise ϵ'. The model is trained to minimize the reconstruction loss between ϵ' and the ground-truth noise ϵ, with an additional Kullback–Leibler (KL) regularization term encouraging a smooth latent space.

The forward diffusion process [11, 35] injects noise into a goal point cloud $\mathcal{P}_\mathrm{g}$ over multiple time steps $t \sim \mathcal{U}\{1, \ldots, T\}$ to produce a noisy version $\mathcal{P}_\mathrm{g}^{(t)}$:

$$\mathcal{P}_\mathrm{g}^{(t)} = \sqrt{\bar{\alpha}_t}\mathcal{P}_\mathrm{g} + \sqrt{1-\bar{\alpha}_t}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{2}$$

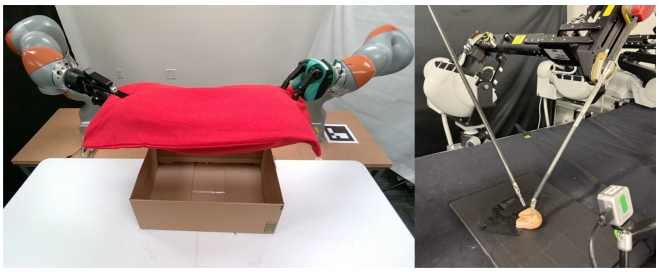

Fig. 4: Physical robot experiment setups. **(Left)** For the object packaging task, we use two KUKA robotic arms to manipulate a soft pillow. **(Right)** For the surgical retraction task, we use a dVRK surgical robot to manipulate *ex vivo* chicken muscle tissue.

where $\bar{\alpha}_t$ is the cumulative noise schedule. In our framework, we use a PointNet-based goal encoder to produce a latent sample

$$z = \mu + \sigma \odot \xi, \quad \xi \sim \mathcal{N}(0, I). \tag{3}$$

Given the point clouds, the current and contextual encoders yield

$$\psi_c = \mathrm{Enc}(\mathcal{P}_\mathrm{c}), \quad \psi_\tau = \mathrm{Enc}(\mathcal{P}_\tau). \tag{4}$$

The noise predictor model then estimates the added noise from all of the encoded information:

$$\epsilon' = \mu_\theta(\mathcal{P}_\mathrm{g}^{(t)}, t, z, \psi_c, \psi_\tau). \tag{5}$$

The training objective is the mean squared error between predicted and ground-truth noise, regularized by a KL penalty on the latent space:

$$\mathcal{L} = \mathbb{E}\left[\|\epsilon - \epsilon'\|^2\right] + \lambda D_\mathrm{KL}\left[q_\varphi(z \mid \mathcal{P}_\mathrm{g}) \,\|\, \mathcal{N}(0, I)\right]. \tag{6}$$

At runtime, we sample $z \sim \mathcal{N}(0, I)$ and initialize $\mathcal{P}_\mathrm{g}^{(T)} \sim \mathcal{N}(0, I)$. The reverse diffusion process [11, 35] iteratively removes noise using μ_θ, ultimately yielding a clean goal $\hat{\mathcal{P}}_\mathrm{g} \sim p_\theta(\mathcal{P}_\mathrm{g} \mid \mathcal{P}_\mathrm{c}, \mathcal{P}_\tau)$. Sampling multiple goals from this distribution produces diverse valid strategies to accomplish the task τ.

C. Integration With Goal-Conditioned Control

Once a goal point cloud is successfully generated by *DiffDef*, the robot deforms the object to reach the goal shape using *DeformerNet* [13], a task-agnostic deformable shape control policy that maps $(\mathcal{P}_\mathrm{c}, \mathcal{P}_\mathrm{g}) \mapsto \mathcal{A}$. Executed in a closed-loop manner [10, 12, 13], this policy drives the object toward the sampled goal, thereby successfully completing the manipulation task τ. Together, *DiffDef* and *DeformerNet* form a unified pipeline: *DiffDef* proposes diverse, task-feasible goal shapes, while *DeformerNet* efficiently executes corresponding deformation actions.

V. Experiments and Results

We evaluate the performance of *DiffDef* on two representative robotic tasks that highlight its versatility across both surgical and manufacturing domains: (i) **tissue retraction**, which captures the safety-critical requirements of soft-tissue surgery, and (ii) **object packaging**, which reflects deformable object manipulation in industrial automation. For each task, we study both simulated environments and

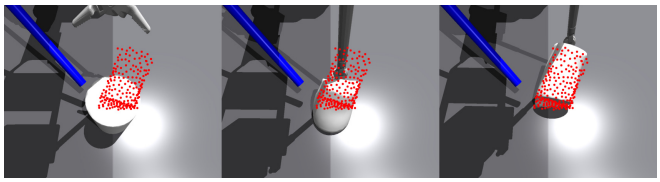

Fig. 5: Representative manipulation sequence of retraction task in simulation. Red point cloud is the goal shape generated by *DiffDef*.

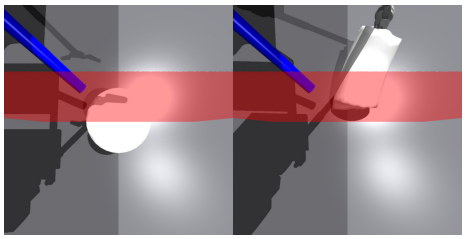

Fig. 6: Tool-conditioned surgical retraction task setup. The objective is to pull the tissue (white) to one side to create tension for subsequent cutting, while avoiding collision with the other surgical tool (blue). **Left:** Initial task state, where a target plane (red) bisects the tissue into two halves. **Right:** The task is deemed successful if the tissue is fully retracted past the target plane without any collision.

physical robot setups. This structure enables us to compare controlled, large-scale simulation results with real-world outcomes under practical sensing and actuation conditions.

In simulation, we utilize the patient-side manipulator of the da Vinci Research Kit (dVRK) surgical robot [36], using the Isaac Gym platform [37]. For physical-robot experiments, we use the physical dVRK system for the surgical task, and two KUKA iiwa 7-DOF robotic arms equipped with Robotiq and Reflex Takktile 2 grippers for the manufacturing task. An Intel® RealSense™ D455 camera is used to capture point clouds for perception. The physical experiment setups are visualized in Figure 4.

A. Tissue Retraction

Retraction is a fundamental task in soft-tissue resection: tissue must be placed under sufficient tension to expose the cut target and enable safe resection. A key challenge is not just whether to retract, but how to select a retraction direction that provides effective exposure while avoiding collisions with other surgical instruments or sensitive anatomy. We design retraction experiments, in both simulation and on physical hardware, to capture this multimodal decision-making process and evaluate the extent to which *DiffDef* can generate diverse feasible retraction goals.

1) Simulated Retraction: We first evaluate *DiffDef* on a simulated tissue retraction task built in Isaac Gym using

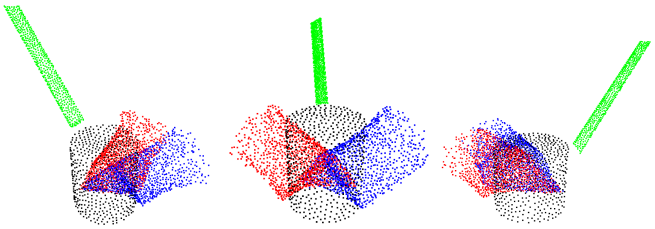

Fig. 7: Example demonstrations for the surgical retraction task. Black: initial tissue state; green: contextual surgical tool; red/blue: two distinct, equally valid goal shapes.

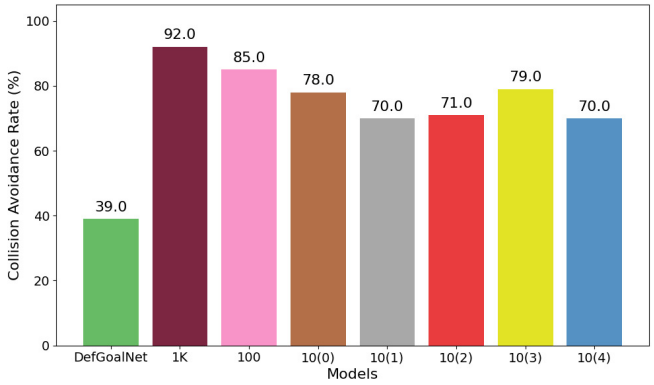

Fig. 8: Simulated retraction—Collision-avoidance rate (higher is better) across multiple dataset sizes. Left to right: *DefGoalNet* trained with 1000 demos vs *DiffDef* with 1000, 100, and 10 demos.

the dVRK patient-side manipulator (PSM) robot [36]. The simulated tissue is modeled as a cylindrical deformable patch, and the cutting instrument (to avoid collisions with) is represented by a cylindrical rigid tool inserted into the scene (see Figure 5). The contextual point cloud $\mathcal{P}_\tau$ corresponds to the visible portion of this tool, constraining feasible retraction directions. The robot must pull the tissue so that it crosses a predefined target plane (representing "safe exposure") without colliding with the tool (see Figure 6).

a) Dataset: In simulation, we collect multimodal demonstrations using a scripted multimodal robot policy (Figures 6 and 7), which produces multiple demonstrations for each context. For each tool pose, two distinct retraction angles are executed (e.g., 30° and 60° relative to the tool axis), creating multiple valid goals per context (see Figure 7). We use tissues with diameters ranging between 1–7 cm, consistent with the typical distribution of tumor sizes observed in surgery [38]. We randomize tissues poses and tool poses to define restricted regions where tissue retraction is not allowed. Note that we use simulated data only for extensive ablation studies and to understand the impact of demonstration dataset size on *DiffDef* performance. All subsequent real robot experiments are conducted with models trained on real human demonstrations.

b) Training and Evaluation: We train *DiffDef* with demonstration datasets of size 10, 100, and 1000, and evaluate on 100 unseen contexts. Baselines include the deterministic *DefGoalNet* [10]. Metrics include: (i) *Collision-avoidance rate*: percentage of executed retraction sequences that successfully avoid tool collision. (ii) *Success percentage*: percentage of points in the final tissue point cloud that successfully pass through the target plane. Note that we only calculate this metric for retraction sequences that are *collision-free*. (iii) *Chamfer distance*: geometric similarity between predicted and ground-truth goal point clouds.

c) Results: Qualitatively, as illustrated in Figure 2, *DiffDef* captures much more effectively the underlying two-mode goal distribution, producing two distinct and equally valid goals. Both predicted goals are realistic and semantically coherent. In contrast, *DefGoalNet* average the two modes, resulting in a single unrealistic and physically impossible goal point cloud. Quantitatively, even with as few

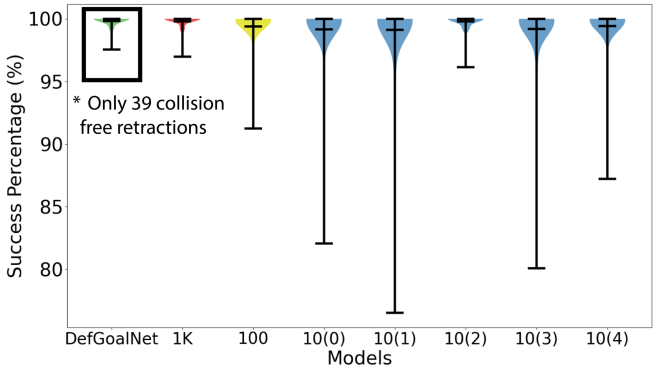

Fig. 9: Simulated retraction results—Success percentage metric. We only calculate this metric on collision-free retraction sequences.

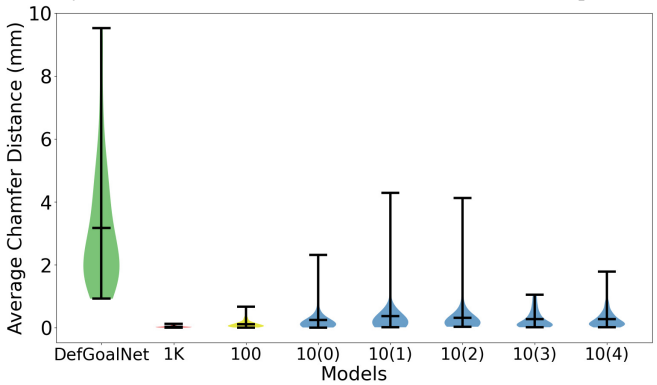

Fig. 10: Simulated retraction results—Violin plots of Chamfer distances (lower is better) between predicted and ground truth goals.

as 10 demonstrations, *DiffDef* achieves a collision avoidance rate of >70% and a median success percentage of >95% (Figures 8 and 9). Notably, *DiffDef* trained on just 10 demonstrations already outperforms *DefGoalNet* trained with the full dataset of 1000 demonstrations. For success percentage, *DiffDef* appears comparable to *DefGoalNet*, but only 39 collision-free retractions from *DefGoalNet* are included in the calculation, compared to the 92 retractions from *DiffDef*. Representative manipulation sequences are shown in Figure 5. The retraction simulation experiments confirm the key advantages of *DiffDef*: robustness to small datasets, the ability to capture multimodal goal distributions, and superior task success compared to *DefGoalNet*.

2) Physical Retraction: Next, we evaluate *DiffDef* on a real dVRK robotic system manipulating *ex vivo* chicken muscle tissue using real human demonstrations. As in surgery, retraction must be coordinated with the cutting tool. We model this by introducing a second dVRK instrument that provides context: the tool's pose dictates which retraction directions are clinically feasible. Conditioning on this context ensures that predictions respect surgical constraints and proves that complicated surgical contexts can be used to inform *DiffDef*, resulting in multiple unique valid goals.

a) Dataset: We collected 42 human expert demonstrations across 21 tool poses (two goal shapes per pose). We train on 32 demonstrations from 16 tool contexts and hold out the remaining 10 demonstrations from 5 unseen contexts for testing.

b) Evaluation: For each test case, we sample one goal point cloud from *DiffDef*. To assess point cloud quality, we

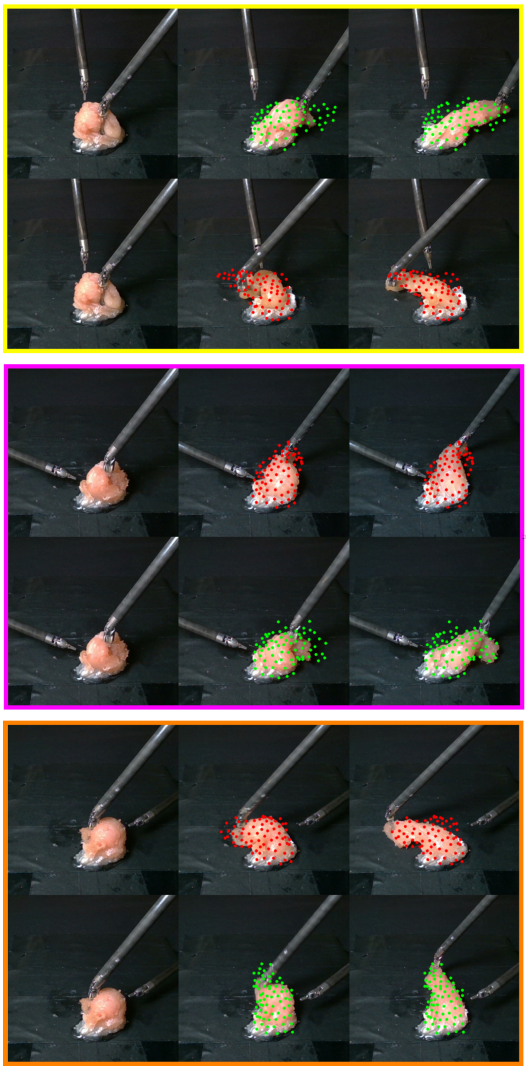

Fig. 11: Example retraction sequences on the physical dVRK surgical robot. Each pair of sequences shares the same task context, but utilize two different goal shapes (green/red points) sampled from *DiffDef*, resulting in two distinct ways the tissue is retracted.

compare the generated goal to the two human-demonstrated goals and report the minimum Chamfer distance, since diffusion does not allow us to explicitly choose between modes. A binary success is recorded if the tissue is retracted past the target plane without tool collision.

c) Results: Across five unseen contexts, chamfer distances were 0.029, 0.015, 0.007, 0.005, and 0.009 millimeters. Figure 12 visualizes the *worst-case* predicted goal point cloud among the test set. Qualitatively, the goals look similar to the ground truth, and the model captures well the goal shape geometry, showcasing *DefGoalNet*'s ability to generalize even when trained on a relatively small dataset.

We perform the retraction procedure on three distinct tool pose configurations. For each configuration, we execute the full seroving pipeline using *DeformerNet* on two different goals generated by *DiffDef*. For each goal predicted by *DiffDef* we run the servoing pipeline three times to assess the robustness of our method, resulting in $3 \times 2 \times 3 = 18$ trials in total. Across these trials, we observe a retraction success rate of 100%. 6 representative sequences are shown

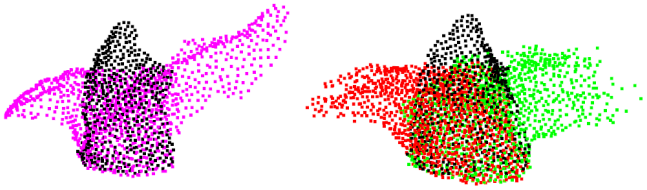

Fig. 12: Worst-case predicted goals from *DiffDef* (right) vs. ground-truth (left) in physical retraction. Black: initial tissue state; Purple: two ground-truth goals; Red/green: two predicted goals. Even in the worst case, predictions closely resemble expert demonstrations.

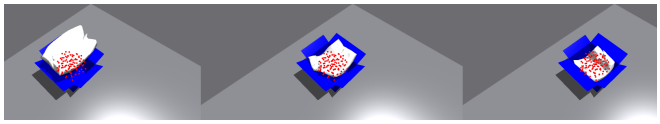

Fig. 13: Sample manipulation sequence on the object packaging task, in simulation. The red point cloud visualizes the goal shape generated by *DiffDef*.

in Figure 11.

B. Object Packaging

In manufacturing, the deformable object packaging task involves safely deforming and compressing an object so that it fits into a container. This scenario is common in warehouse operations, where the deformable item to be packaged may initially be larger than the container's opening. This task highlights continuous multimodal distributions, as an infinite number of feasible packaging arrangements exist.

1) Simulated Packaging: We set up an object packaging task in Isaac Gym where a deformable pillow must be placed inside a box with varying poses. In this setting, the partial point cloud of the box serves as contextual input. Unlike retraction, goal shape non-determinism here arises from continuous rotations of the pillow around the vertical axis. This results in a complex goal distribution characterized by a single high-variance mode, rather than multiple distinct modes.

a) Dataset: Only for large-scale simulation experiments, demonstrations are collected using a scripted multimodal robot policy (Figure 14). For physical robot experiments, we use real human demonstrations (Figure 18). The pillow rotation angle is sampled from a continuous uniform distribution over $(-\pi, \pi]$, yielding a complex continuous goal distribution characterized by a single high-variance mode.

b) Training and Evaluation: We train *DiffDef* with 10, 100, and 1000 demonstrations and compare it against *DefGoalNet*, which is trained with 1000 demonstrations. A representative manipulation sequence is shown in Figure 13. The primary evaluation metric is *coverage percentage*, defined as the fraction of object volume contained within the box after servoing with *DeformerNet*. We additionally report the Chamfer distance between the final object shape and human expert goals. For each test scenario, we randomize both the deformable object size and the container pose. We then execute the *DeformerNet* policy conditioned on the goal predicted by *DiffDef*, record the resulting object point cloud, and compute the percentage of points successfully contained.

c) Results: We present qualitative results in Figure 15, and quantitative results in Figure 16 and Figure 17.

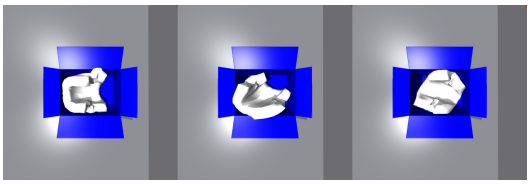

Fig. 14: Example demonstrations for the simulated object packaging task. For a given container, the deformable object can be arranged in multiple valid configurations.

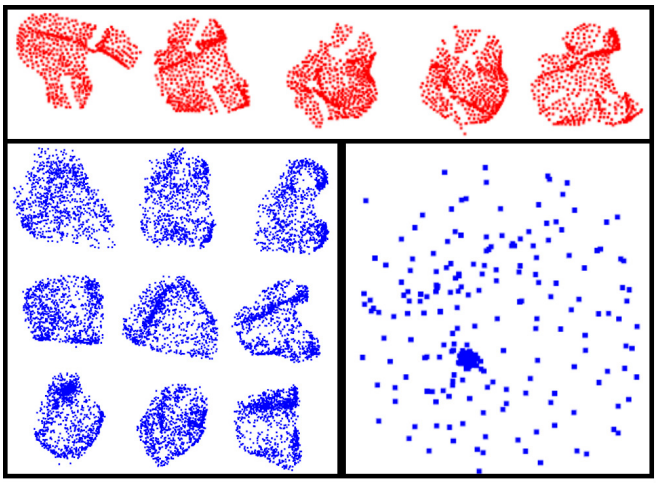

Fig. 15: (Red) Ground-truth goal point clouds collected from expert demonstrator. **(Blue)** Predicted goals from *DiffDef* (bottom left) compared to *DefGoalNet* (bottom right). *DiffDef* generates diverse, multimodal goals, while *DefGoalNet*'s prediction is physically infeasible.

DiffDef achieves nearly 100% median coverage with only 100 demonstrations, outperforming *DefGoalNet* trained with 1000 demonstrations. *DiffDef* again captures effectively the complex continuous goal distribution, while *DefGoalNet* produces physically impractical goal point clouds. Figure 17 shows the Chamfer distance between predicted and ground-truth goal point clouds on a test set of 100 unseen demonstrations. Similar to the surgical retraction task, *DiffDef* significantly outperforms *DefGoalNet*.

2) Physical Object Packaging: Finally, we validate *DiffDef* on physical hardware using the dual-arm KUKA iiwa system (Figure 4-left). Two 7-DOF arms equipped with Robotiq and Reflex Takktile 2 grippers are tasked with placing a soft pillow into a cardboard box.

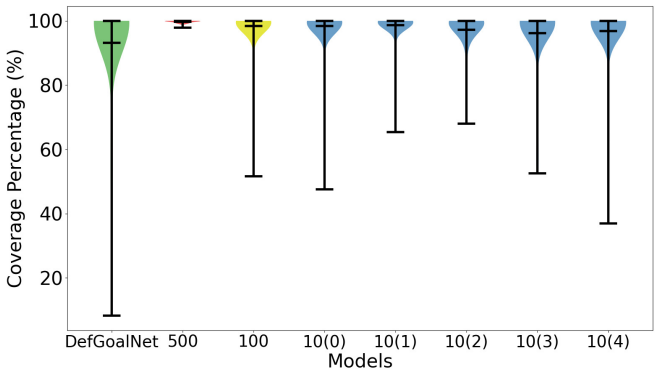

Fig. 16: Simulated object packaging—Violin plots of the coverage percentage metric (higher is better). Left to right: *DefGoalNet*, followed by *DiffDef* trained with 500, 100, and 10 demos.

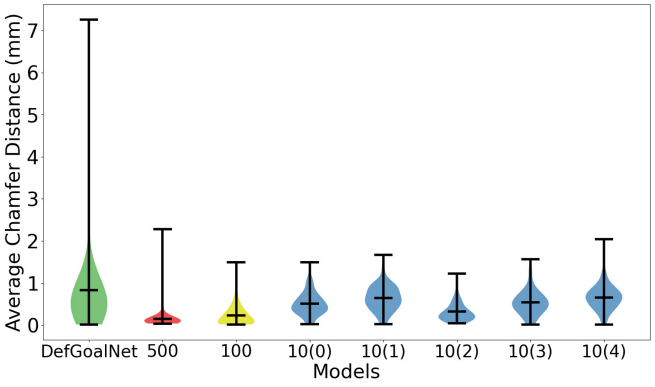

Fig. 17: Simulated object packaging—Violin plots of Chamfer distances (lower is better) between predicted and ground truth goals.

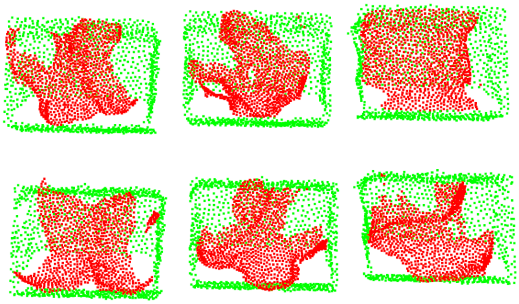

Fig. 18: Example human demonstrations for real-robot object packaging task. Green: container (context); Red: six equally valid goal shapes representing different pillow placement strategies.

a) Dataset: We collected 36 demonstrations across six distinct box poses, with six valid pillow placements for each pose (Figure 18). We use 30 demonstrations from five box poses for training and hold out the remaining six demonstrations from the sixth pose for testing.

b) Results: On the held-out box pose, the predicted goal shapes closely matched the human demonstrations (Figure 19). We evaluated the full servoing pipeline (*DiffDef* + *DeformerNet*) on two previously unseen box poses, sampling three predicted goals per pose and conducting three trials for each goal ($2 \times 3 \times 3 = 18$ trials in total). All 18 trials were successful, with representative sequences shown in Figure 20.

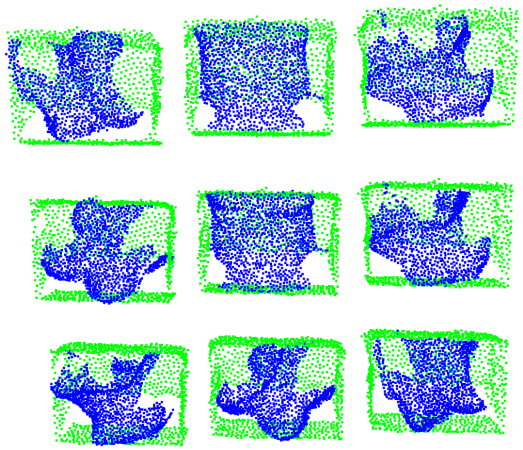

Fig. 19: Predicted goal point clouds (blue) sampled from the goal distribution learned by *DiffDef*, in real-robot packaging experiment.

Fig. 20: Six object packaging sequences with the KUKA system. Each context (container's pose) supports multiple valid packaging strategies corresponding to distinct goals predicted by *DiffDef*.

C. Summary of Results

Across all tasks and domains, *DiffDef* consistently generates diverse and realistic goal shapes that closely resemble human demonstrations, leading to high task success rates. The surgical retraction task highlights context-aware multimodality under safety-critical constraints in medical settings, while the object packaging task illustrates continuous multimodality in industrial manipulation. Together, these results demonstrate that *DiffDef* is a general and versatile framework for deformable object manipulation, well-suited for diverse real-world conditions and application domains.

VI. Conclusion

We have presented a pipeline for learning deformable object manipulation from demonstrations. At the heart of this pipeline is *DiffDef*, a neural network that learns a multimodal distribution over diverse goal shapes capable of successfully completing the given task. We evaluate *DiffDef* across a broad range of robotic tasks spanning both surgery and manufacturing applications. Our experiments show that *DiffDef* consistently outperforms the current state-of-the-art goal prediction method—which models a single deterministic goal rather than a distribution as in our approach—across multiple metrics and task domains, while also requiring less training data. Notably, our approach enables effective multimodal task-specific goal generation from only a limited number of demonstrations, while still benefiting from

a generic control policy trained on a large and diverse dataset that is cheap to acquire and agnostic to the specific downstream task.

REFERENCES

[1] J. Sanchez, J. A. C. Ramon, B.-C. Bouzgarrou, and Y. Mezouar, "Robotic Manipulation and Sensing of Deformable Objects in Domestic and Industrial Applications: A Survey," *Intl. Journal of Robotics Research*, vol. 37, no. 7, pp. 688–716, 2018.

[2] J. Zhu, A. Cherubini, C. Dune, D. Navarro-Alarcon, F. Alambeigi, D. Berenson, F. Ficuciello, K. Harada, J. Kober, X. Li, *et al.*, "Challenges and outlook in robotic manipulation of deformable objects," *IEEE Robotics & Automation Magazine*, vol. 29, no. 3, pp. 67–77, 2022.

[3] Y. Wu, W. Yan, T. Kurutach, L. Pinto, and P. Abbeel, "Learning to manipulate deformable objects without demonstrations," *Robotics: Science and Systems*, 2020.

[4] D. Navarro-Alarcon, H. M. Yip, Z. Wang, Y.-H. Liu, F. Zhong, T. Zhang, and P. Li, "Automatic 3-D Manipulation of Soft Objects by Robotic Arms With an Adaptive Deformation Model," *IEEE Trans. on Robotics*, vol. 32, no. 2, pp. 429–441, 2016.

[5] J. Qi, D. Li, Y. Gao, P. Zhou, and D. Navarro-Alarcon, "Model predictive manipulation of compliant objects with multi-objective optimizer and adversarial network for occlusion compensation," *arXiv preprint arXiv:2205.09987*, 2022.

[6] F. Alambeigi, Z. Wang, R. Hegeman, Y.-H. Liu, and M. Armand, "A robust data-driven approach for online learning and manipulation of unmodeled 3-d heterogeneous compliant objects," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4140–4147, 2018.

[7] ——, "Autonomous data-driven manipulation of unknown anisotropic deformable tissues using unmodelled continuum manipulators," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 254–261, 2018.

[8] J. Zhu, D. Navarro-Alarcon, R. Passama, and A. Cherubini, "Vision-based manipulation of deformable and rigid objects using subspace projections of 2d contours," *Robotics and Autonomous Systems*, vol. 142, p. 103798, 2021.

[9] M. Shetab-Bushehri, M. Aranda, Y. Mezouar, and E. Ozgur, "Lattice-based shape tracking and servoing of elastic objects," *arXiv preprint arXiv:2209.01832*, 2022.

[10] B. Thach, T. Watts, S.-H. Ho, T. Hermans, and A. Kuntz, "Defgoalnet: Contextual goal learning from demonstrations for deformable object manipulation," in *IEEE Intl. Conf. on Robotics and Automation*, 2024, pp. 3145–3152.

[11] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[12] B. Thach, B. Y. Cho, T. Hermans, and A. Kuntz, "Deformernet: Learning bimanual manipulation of 3d deformable objects," *ArXiv*, 2023.

[13] B. Thach, B. Y. Cho, A. Kuntz, and T. Hermans, "Learning visual shape control of novel 3d deformable objects from partial-view point clouds," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2022, pp. 8274–8281.

[14] A. Mousavian, C. Eppner, and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2901–2910.

[15] A. Murali, A. Mousavian, C. Eppner, C. Paxton, and D. Fox, "6-dof grasping for target-driven object manipulation in clutter," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6232–6238.

[16] X. Deng, Y. Xiang, A. Mousavian, C. Eppner, T. Bretl, and D. Fox, "Self-supervised 6D Object Pose Estimation for Robot Manipulation," *IEEE Intl. Conf. on Robotics and Automation*, pp. 3665–3671, 2020.

[17] Q. Lu, M. Van der Merwe, and T. Hermans, "Multi-fingered active grasp learning," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 8415–8422.

[18] M. Van der Merwe, Q. Lu, B. Sundaralingam, M. Matak, and T. Hermans, "Learning continuous 3d reconstructions for geometrically aware grasping," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 11 516–11 522.

[19] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[20] D. Navarro-Alarcon, H. M. Yip, Z. Wang, Y.-H. Liu, F. Zhong, T. Zhang, and P. Li, "Automatic 3-d manipulation of soft objects by robotic arms with an adaptive deformation model," *IEEE Transactions on Robotics*, vol. 32, no. 2, pp. 429–441, 2016.

[21] Z. Hu, T. Han, P. Sun, J. Pan, and D. Manocha, "3-D Deformable Object Manipulation Using Deep Neural Networks," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4255–4261, 2019.

[22] R. B. Rusu, G. Bradski, R. Thibaux, and J. Hsu, "Fast 3D recognition and pose using the Viewpoint Feature Histogram," *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, pp. 2155–2162, 2010.

[23] W. Wu, Z. Qi, and L. Fuxin, "Pointconv: Deep convolutional networks on 3d point clouds," *ArXiv*, 2018.

[24] D. W. Shu, S. W. Park, and J. Kwon, "3d point cloud generative adversarial network based on tree structured graph convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3859–3868.

[25] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. Guibas, "Learning representations and generative models for 3d point clouds," in *International conference on machine learning*. PMLR, 2018, pp. 40–49.

[26] M. S. Arshad and W. J. Beksi, "A progressive conditional generative adversarial network for generating dense and colored 3d point clouds," in *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020, pp. 712–722.

[27] A. Pumarola, S. Popov, F. Moreno-Noguer, and V. Ferrari, "C-flow: Conditional generative flow models for images and 3d point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7949–7958.

[28] S. van Waveren, C. Pek, I. Leite, J. Tumova, and D. Kragic, "Large-scale scenario generation for robotic manipulation via conditioned generative models," *preprint*, 2022.

[29] Y. Du, S. Yang, B. Dai, H. Dai, O. Nachum, J. Tenenbaum, D. Schuurmans, and P. Abbeel, "Learning universal policies via text-guided video generation," *Advances in neural information processing systems*, vol. 36, pp. 9156–9172, 2023.

[30] O. Mees, L. Hermann, and W. Burgard, "What matters in language conditioned robotic imitation learning over unstructured data," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 205–11 212, 2022.

[31] H. Zhou, X. Yao, O. Mees, Y. Meng, T. Xiao, Y. Bisk, J. Oh, E. Johns, M. Shridhar, D. Shah, *et al.*, "Bridging language and action: A survey of language-conditioned robot manipulation," *arXiv e-prints*, pp. arXiv–2312, 2023.

[32] 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.

[33] J. Pavlasek, J. J. Z. Mah, R. Xu, O. C. Jenkins, and F. Ramos, "Stein variational belief propagation for multi-robot coordination," *IEEE Robotics and Automation Letters*, 2024.

[34] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017.

[35] S. Luo and W. Hu, "Diffusion probabilistic models for 3d point cloud generation," in *IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2837–2845.

[36] P. Kazanzides, Z. Chen, A. Deguet, G. S. Fischer, R. H. Taylor, and S. P. DiMaio, "An open-source research kit for the da Vinci® Surgical System," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, May 2014, pp. 6434–6439.

[37] J. Liang, V. Makoviychuk, A. Handa, N. Chentanez, M. Macklin, and D. Fox, "GPU-Accelerated Robotic Simulation for Distributed Reinforcement Learning," *arXiv:1810.05762*, 2018.

[38] A. Ernst, D. Feller-Kopman, H. D. Becker, and A. C. Mehta, "Central airway obstruction," *American journal of respiratory and critical care medicine*, vol. 169, no. 12, pp. 1278–1297, 2004.