DefWAM: 3D Point Flow Prediction Learns Effective Goal-Conditioned Deformable Object Manipulation

Gaurav Singh, Sergio Orozco, Ryan Duong, Mete Tuluhan Akbulut, George Konidaris, Srinath Sridhar
Brown University, Providence, RI, USA

***Abstract*—Manipulating deformable objects toward desired 3D configurations is a fundamental capability for robots operating in unstructured physical environments. However, goal-conditioned deformable object manipulation in 3D requires reasoning over high-dimensional shape variation, complex dynamics, and complex gripper-object interactions. We propose DefWAM, a 3D World Action Modeling approach that learns goal-conditioned 3D point-flow prediction as a unified objective for modeling both robot behavior and deformable scene evolution. By representing the robot and scene as point clouds in a unified 3D space, and actions as 3D point flows, DefWAM predicts goal-conditioned joint robot and scene flow and models deformable dynamics directly in particle space. This joint 3D point-flow objective enables our approach to learn when and how to use different contact modes, such as grasping, releasing, and re-grasping, to drive deformable objects toward diverse target shapes. Our experiments show that DefWAM outperforms action-only 3D goal-conditioned imitation learning baselines, particularly under large variation in initial and goal configurations. Furthermore, DefWAM's predicted scene evolution enables inference-time steering to avoid collisions by implicitly simulating possible robot-scene futures.**

I. Introduction

Robotic manipulation of deformable objects, such as ropes, cloth, or granular media remains a challenging task due to their high-dimensional configuration spaces. Unlike rigid objects, whose configurations can be described by a 6-DoF pose, deformable objects can undergo a wide range of 3D shape changes governed by complex dynamics. These challenges are further compounded by contact with robot grippers and rigid objects, where even simple actions, such as twisting a rope with a two-finger gripper, can deform the object in ways that are difficult to predict. In this work, we focus on **3D goal-conditioned manipulation of deformable objects** with complex shape configurations, where the objective is to drive the object from its current configuration toward a desired target configuration. This setting motivates models that can learn goal-conditioned robot behaviors while capturing the underlying dynamics of the coupled robot-object system. In particular, the robot must learn not only where to move, but also which behaviors to use and how to execute them to induce the desired shape change.

Prior works on deformable object manipulation often use 3D point or particle-based representations [1, 13, 15], including in goal-conditioned imitation learning [2, 3], demonstrating their value for specifying current and target configurations. However, these approaches remain largely limited to isolated behavior primitives such as pinching or swishing [2, 4] and to target configurations seen during training [2], and because they supervise only on robot actions, the effect of an action on the deformable scene is never an explicit learning target, leaving it unclear whether the learned policy captures scene evolution in a generalizable way. World action models offer an alternative by jointly learning robot actions and scene dynamics [11, 7], providing dense supervision about how actions change the world and enabling learning from heterogeneous, non-repetitive data. Recent work explores goal-conditioned WAMs in 3D [8] but remains limited to rigid objects modeled through 6-DoF poses, whereas deformable manipulation requires capturing high-dimensional 3D shape dynamics.

We propose **DefWAM, a 3D world action model for goal-conditioned deformable object manipulation**. Given the current robot-scene point cloud and a goal point cloud, DefWAM predicts joint robot-scene point flows in a unified 3D particle space, where robot point flows define how the gripper should move and scene point flows predict how the object and surrounding scene evolves jointly. This formulation allows DefWAM to learn from non-repetitive demonstrations covering a wide range of contact modes and shape deformations, while associating different robot behaviors with their induced changes in object shape. By learning actions and dynamics together in 3D, DefWAM generalizes across diverse start and goal configurations and learns to chain multiple behaviors such as grasping, translating, dropping and re-grasping to reach a goal deformable object shape. We evaluate DefWAM across multiple deformable-object manipulation settings that require reasoning over diverse 3D goal configurations and changing contact modes. Across these settings, DefWAM outperforms goal-conditioned 3D imitation learning baselines, particularly when initial and goal configurations vary substantially across demonstrations. We further show that DefWAM's predicted scene evolution provides an additional inference-time capability, allowing the model to autoregressively imagine multiple *"robot-scene futures"* implicitly as point flows and steer around unseen obstacles to avoid collisions of the deformable object.

II. Related Work

Particle-based representations: Particle and point-cloud representations explicitly encode 3D object structure and scale across object categories and materials [1], and have been widely used to model deformable dynamics via learned

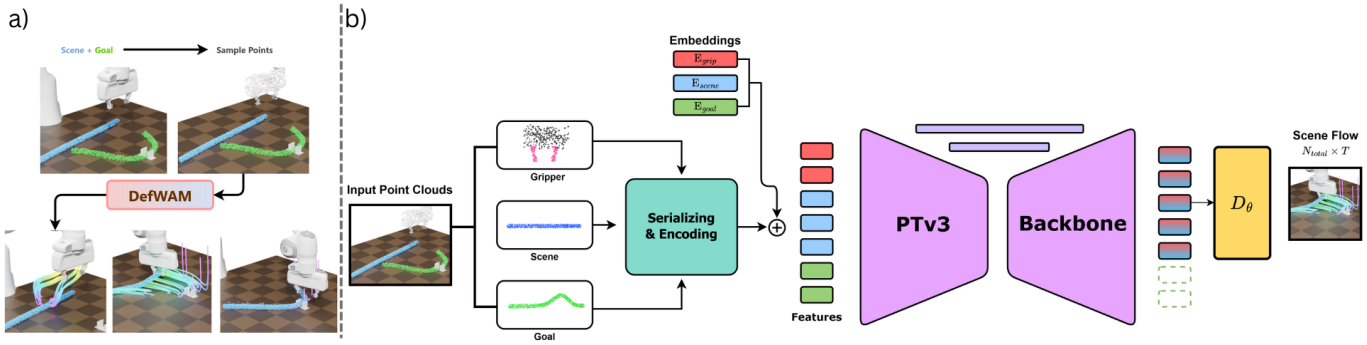

Fig. 1: **(a) Method Overview:** DefWAM learns to manipulate deformable objects by ***predicting joint robot-scene point flows***. This unified 3D objective learns goal-directed manipulation behaviors together with deformable scene dynamics across diverse target shapes.**(b) DefWAM architecture.** Given past robot-scene point clouds and a goal point cloud, DefWAM processes all points in a unified 3D space using a PTv3 backbone with learned type-embeddings. A shared flow head predicts future robot and scene point flows, which are used to recover executable robot actions and model deformable scene evolution.

message passing and Transformer-based architectures, with recent work improving material coverage and partial-view observability [14, 15, 6, 5]. However, planning in low-level action spaces quickly becomes intractable, so these methods typically plan over a fixed set of control particles attached to a predetermined grasp point [6, 14]. **Goal-conditioned imitation learning:** A parallel line of work maps current and goal point clouds directly to actions for deformable tasks [2, 3, 4], but most train on narrow, largely repeated start-goal configurations and demonstrate single-mode behaviors such as pinching, rolling, or swishing. **World action models:** WAMs jointly predict future observations and actions in a single model [11], enabling learning from heterogeneous, non-repetitive data; DyWA [8] extends this to 3D point clouds but represents scene evolution through object-centric 6-DoF poses, which cannot capture high-dimensional deformable shape dynamics. In contrast, DefWAM jointly predicts robot and scene flow in a unified goal-conditioned model, directly modeling robot behavior and deformable dynamics without a separate action-proposal module, and supporting arbitrary contact-mode sequences within a single model.

III. DefWAM: 3D Point Flow Prediction as a Goal-Conditioned Policy

We formulate goal-conditioned manipulation as a world action modeling problem [11]. Given an observation history $o_{\leq t}$, and a goal state g, we want to predict both the robot action a_t and the corresponding future world state o_{t+1}. To this end, we want to learn the conditional distribution $\pi_\theta(a_t, o_{t+1} \mid o_{\leq t}, g)$. For our setting, we represent all the objects including the robot as well as the goal, as 3D points. Let $P_t^r \in \mathbb{R}^{N_r \times 3}$, $P_t^s \in \mathbb{R}^{N_s \times 3}$, and $P^g \in \mathbb{R}^{N_g \times 3}$ be the robot, scene and goal points respectively. P^s and P^g can be obtained by rendering depth or sampling points on scene geometry in a simulation environment. Similarly P^r can be obtained by sampling points on its geometry from the known URDF, which we assume is available to us. Given this point-based representation, DefWAM instantiates world action modeling as joint point-flow prediction over a future horizon. Given the past n frames of robot and scene points and the goal point cloud, DefWAM predicts point flows over a horizon H as

$$(\Delta P^r_{t+1:t+H}, \Delta P^s_{t+1:t+H}) = f_\theta(P^r_{t-n:t}, P^s_{t-n:t}, P^g), \quad (1)$$

where n is the number of history frames and H is the prediction horizon. Here, $\Delta P^r_{t+1:t+H} \in \mathbb{R}^{H \times N_r \times 3}$ represents the robot motion used to recover actions, and $\Delta P^s_{t+1:t+H} \in \mathbb{R}^{H \times N_s \times 3}$ represents the corresponding scene motion. These can be supervised using ground truth point-flows, which can be obtained using the simulation state.

A. Method

Flow Prediction Backbone. We use Point Transformer V3 (PTv3) [10] as our backbone (Fig. 1(b)) similar to PointWorld [5], with each point represented by its 3D coordinate and optional color. Given $\mathbf{P}^r_{t-n:t}$, $\mathbf{P}^s_{t-n:t}$, and $\mathbf{P}^g$, PTv3 processes all points as a single unified set, adding learned type embeddings $\mathbf{E}_r, \mathbf{E}_s, \mathbf{E}_g \in \mathbb{R}^d$ to distinguish robot, scene, and goal tokens. Its encoder-decoder attention exchanges information across the three sources and outputs a per-point feature for every input point. Following PointWorld [5], a shared per-point head D_θ maps each robot and scene feature to H future displacements, producing the joint flow $(\Delta\mathbf{P}^r_{t\to t+1:t+H}, \Delta\mathbf{P}^s_{t\to t+1:t+H})$. We train end-to-end with a Huber loss against ground-truth displacements $\Delta\mathbf{P}^*_{t\to t+h} = \mathbf{P}_{t+h} - \mathbf{P}_t$ over the union of scene and robot points:

$$\mathcal{L}_{\text{flow}} = \sum_{h=1}^{H} \mathcal{L}_{\text{Huber}}\left(\Delta\mathbf{P}_{t\to t+h}, \Delta\mathbf{P}^*_{t\to t+h}\right), \quad (2)$$

jointly supervising both terms to learn the coupling between robot motion and scene deformation.

Action Extraction via Inverse Kinematics. The predicted robot point flow specifies where each sampled point on the robot geometry should move over the next H steps, so we recover executable joint trajectories by solving damped least-squares inverse kinematics independently at each timestep, matching the URDF forward-kinematics points to the predicted ones, and obtain a binary grasp open/close action by thresholding the predicted gripper width. The resulting joint

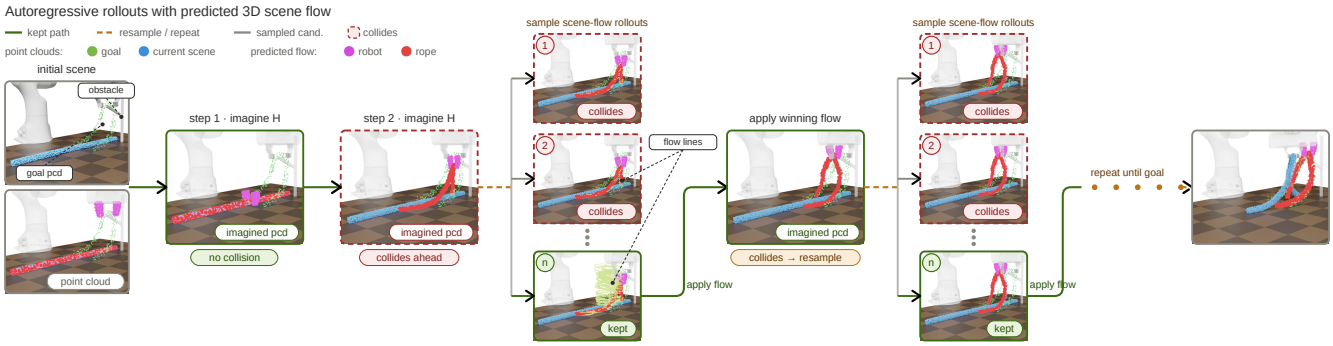

Fig. 2: **Inference-time steering.** DefWAM generates multiple predicted scene-flow rollouts (shown in magenta/red), which can be checked for collisions constraints, greedily keeping the first collision-free one (green = goal, blue = current scene). Zoom in for a better view.

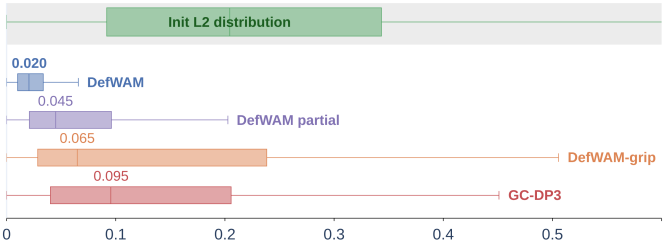

Fig. 3: Per-particle L2 distance between the final rope state and goal configuration in meters. *Init L2* refers to the starting state distance distribution.

Method	CD ↓	EMD ↓	HD ↓	IoU ↑
GC-DP3 [13]	$1.05_{\pm 0.18}$	$2.41_{\pm 0.54}$	$4.25_{\pm 0.90}$	$55.66_{\pm 7.05}$
SculptDiff [2]	$0.92_{\pm 0.17}$	$2.07_{\pm 0.57}$	$\mathbf{3.85}_{\pm 0.93}$	$58.78_{\pm 7.76}$
Ours	$\mathbf{0.82}_{\pm 0.18}$	$\mathbf{1.76}_{\pm 0.64}$	$3.91_{\pm 0.95}$	$\mathbf{64.14}_{\pm 8.27}$

TABLE I: Quantitative comparison on granular media manipulation. CD, EMD, and Hausdorff (HD) are scaled by 10^2. 3D IoU is shown as %.

sequence and grasp action can be sent to existing controllers, so DefWAM produces executable actions without a separate learned action head.

B. Inference-Time Steering with Predicted Scene Flow

Since DefWAM predicts temporally stable robot motion and scene flow, its predictions can be rolled forward autoregressively in a *chain-of-thought style*, producing imagined robot-scene futures before executing any action. Since the predicted scene and obstacles all live in the same 3D point-cloud space, potential collisions can be checked directly on the predicted points, enabling the ability to search for valid actions by sampling multiple possible paths. This yields different coupled robot-scene rollouts, which we greedily collision-check in point-cloud space to steer toward a feasible one, as shown in Fig 2. To implement this, we train with two observation frames similar to [13] and the previous observation $\{P^r_{t-1}, P^s_{t-1}\}$ mainly disambiguates the robot's direction of motion. Thus multiple samples are generated by perturbing the previous robot points P^r_{t-1} with a small global 3D offset, $\tilde{P}^r_{t-1} = P^r_{t-1} + \delta$, $\delta \sim \mathcal{N}(0, \sigma^2 I_3)$, which is enough to alter the inferred motion direction while preserving robot geometry.

IV. Experiments

We evaluate DefWAM on two deformable settings: **rope manipulation**, simulated using a spring-mass model [6], and **granular media shaping**, simulated using the Material Point Method [9]. The rope environment requires the policy to chain multiple prehensile behaviors, implicitly figuring out where to grasp, how to drag, and when to release purely through goal conditioning, while the granular environment requires the robot to figure out how to move the paddle to form arbitrary goal shapes. Expert demonstrations for specific goal shapes are costly and time consuming for deformable objects because their configurations vary unpredictably. We instead collect unstructured play data in simulation, which covers diverse object configurations and behaviors. We build environment-specific procedural policies that use privileged state information to compose primitives such as grasping, pushing, pick-and-place, and random displacement from randomized starts. Following hindsight relabeling, we use each episode's final scene state as the goal, producing diverse start-goal pairs without goal-specific demonstration collection.

Method	Success Rate (%)
DefWAM	20
DefWAM + Steering	93.3

TABLE II: Success rate on the rope-hanging task with and without inference-time steering.

We compare against the following baselines and ablations. **GC-DP3**: we adapt 3D Diffusion Policy [13] by adding a conditioning latent for the goal point cloud. **SculptDiff** [2]: a point-cloud goal-conditioned diffusion policy built on Point-BERT [12] for clay sculpting. **DefWAM-partial**: evaluates DefWAM using partial point clouds instead of full particle states, testing operation under partial observability. **DefWAM-grip**: an ablation that supervises only the gripper flow and omits the rope flow, isolating the contribution of joint robot-scene flow prediction.

Fig. 3 reports per-particle L2 distance on rope manipulation. DefWAM achieves the lowest L2 distance and the tightest dispersion across trials, while GC-DP3 fails to disambiguate situations effectively, learns spurious correlations, and converges to local minima. Notably, DefWAM remains effective even with partial views, outperforming both GC-DP3 and DefWAM-grip, indicating its applicability to real settings where full point clouds are rarely available. On granular shaping, Table I shows DefWAM achieves the best Chamfer distance, EMD, and 3D IoU and remains competitive

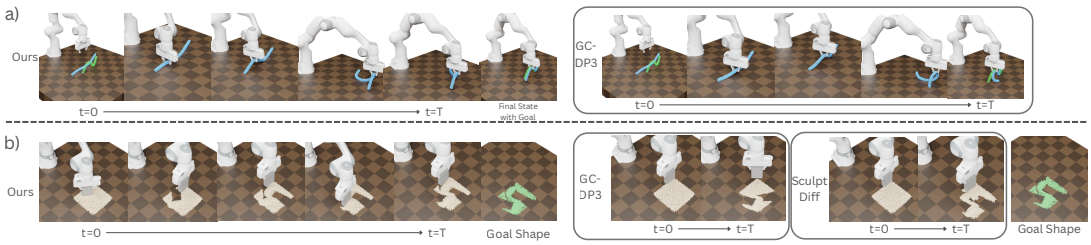

Fig. 4: **Qualitative results.** (a) Rope manipulation rollouts comparing DefWAM and GC-DP3. The leftmost frame shows the initial rope+goal configuration at $t = 0$, intermediate frames show the executed trajectory, and the rightmost frame shows the final rope state overlaid with the goal. (b) Granular shaping rollouts comparing DefWAM, GC-DP3, and SculptDiff. The leftmost frame shows the initial sand pile, intermediate frames show the shaping process, and the rightmost frame shows the final state relative to the target goal shape.

on Hausdorff distance, despite being trained from scratch while SculptDiff uses a pretrained PointBERT. Fig. 4 provides qualitative rollouts for the approaches.

A. Inference-Time Steering

We evaluate inference-time steering on a rope-hanging task, where the robot must place the rope onto a stand peg and the most direct predicted trajectory often deforms the rope into the peg. Since DefWAM predicts full scene flow, predicted rollouts can be collision-checked before execution and steered toward a feasible alternative. Across 10 starting configurations with 3 trials each, steering substantially improves the success rate (Table II).

V. Conclusion and Limitations

We introduced DefWAM, a goal-conditioned 3D world action model for deformable object manipulation that jointly predicts robot and scene point flows in a unified particle space. By supervising both robot motion and deformable scene evolution, DefWAM learns effective strategies from heterogeneous, non-repetitive demonstrations, generalizes across diverse start and goal configurations, and outperforms action-only imitation learning baselines on rope manipulation, and granular shaping. We further show that predicted scene flow enables inference-time steering through imagined robot-scene futures. **Limitations.** Specifying goals as 3D point clouds is impractical for many real-world settings, where interfaces such as language, sketches, or demonstrations would be more natural. Our analytical action extraction also relies on an empirically set grasp threshold that a learned controller could replace to improve flexibility across robots and tasks.

References

[1] Bo Ai, Stephen Tian, Haochen Shi, Yixuan Wang, Tobias Pfaff, Cheston Tan, Henrik I Christensen, Hao Su, Jiajun Wu, and Yunzhu Li. A review of learning-based dynamics models for robotic manipulation. *Science Robotics*, 10(106):eadt1497, 2025.

[2] Alison Bartsch, Arvind Car, Charlotte Avra, and Amir Barati Farimani. Sculptdiff: Learning robotic clay sculpting from humans with goal conditioned diffusion policy. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7307–7314. IEEE, 2024.

[3] Alison Bartsch, Arvind Car, and Amir Barati Farimani. Pinchbot: Long-horizon deformable manipulation with guided diffusion policy. *arXiv preprint arXiv:2507.17846*, 2025.

[4] Cheng Chi, Benjamin Burchfiel, Eric Cousineau, Siyuan Feng, and Shuran Song. Iterative residual policy: for goal-conditioned dynamic manipulation of deformable objects. *The International Journal of Robotics Research*, 43(4):389–404, 2024.

[5] Wenlong Huang, Yu-Wei Chao, Arsalan Mousavian, Ming-Yu Liu, Dieter Fox, Kaichun Mo, and Li Fei-Fei. Pointworld: Scaling 3d world models for in-the-wild robotic manipulation. *arXiv preprint arXiv:2601.03782*, 2026.

[6] Hanxiao Jiang, Hao-Yu Hsu, Kaifeng Zhang, Hsin-Ni Yu, Shenlong Wang, and Yunzhu Li. Phystwin: Physics-informed reconstruction and simulation of deformable objects from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7219–7230, 2025.

[7] Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, and Jinwei Gu. Cosmos policy: Fine-tuning video models for visuomotor control and planning. In *International Conference on Learning Representations (ICLR)*, 2026.

[8] Jiangran Lyu, Ziming Li, Xuesong Shi, Chaoyi Xu, Yizhou Wang, and He Wang. Dywa: Dynamics-adaptive world action model for generalizable non-prehensile manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11058–11068, 2025.

[9] Deborah Sulsky, Zhen Chen, and Howard L Schreyer. A particle method for history-dependent materials. *Computer methods in applied mechanics and engineering*, 118(1-2):179–196, 1994.

[10] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4840–4851, 2024.

[11] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru,

You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

[12] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19313–19322, 2022.

[13] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. *arXiv preprint arXiv:2403.03954*, 2024.

[14] Kaifeng Zhang, Baoyu Li, Kris Hauser, and Yunzhu Li. Adaptigraph: Material-adaptive graph-based neural dynamics for robotic manipulation. *arXiv preprint arXiv:2407.07889*, 2024.

[15] Kaifeng Zhang, Baoyu Li, Kris Hauser, and Yunzhu Li. Particle-grid neural dynamics for learning deformable object models from rgb-d videos. *arXiv preprint arXiv:2506.15680*, 2025.