Learning Physics-Guided Residual Dynamics for Deformable Object Simulation

Shivansh Patel[1,2] Kaifeng Zhang[2,3*] Sanjay Pokkali[1*] Svetlana Lazebnik[1] Yunzhu Li[2,3]

[1]UIUC [2]SceniX Inc. [3]Columbia University

***Abstract*—Simulating deformable objects is essential for a wide range of robotic manipulation applications, yet accurately predicting their dynamics remains challenging. We propose Physics-Guided Residual Dynamics (PGRD), a hybrid simulation framework that combines the advantages of physics-based and learning-based approaches. Specifically, PGRD combines an optimizable spring–mass simulator as a backbone with a learned neural network that predicts residual corrections to the physics-based predictions. We adopt a velocity-based formulation to ensure stable simulation and a sliding-window transformer architecture to capture temporal dependencies. We show that PGRD produces more accurate results than both purely physics-based and learning-based methods on a set of diverse real-world deformable objects. We further demonstrate the utility of PGRD in two applications: manipulation planning via Model Predictive Control, including a language-conditioned setting with a generated goal image; and interactive simulation via action-conditioned video prediction by 3D Gaussian Splatting. Project page: https://pgrd-robot.github.io/**

I. Introduction

Accurate simulation of deformable objects is a persistent challenge in robotics and computer vision. Such objects can change shape significantly under external forces, undergoing stretching, bending, crumpling, and twisting. Simulation difficulty stems from material properties such as heterogeneous elasticity and damping, but is also complicated by factors such as self-collisions and friction that come into play during contact-rich interactions.

The literature contains two dominant paradigms for deformable object simulation: physics-based methods [4, 66, 54] and learning-based methods [80, 10, 36] (see Section II for a more detailed survey of related works). Physics-based methods use equations to describe how materials deform and respond to forces, yielding physically plausible and interpretable simulations. However, these methods require precise material parameters that are difficult to obtain in practice [44, 82, 55]. The complex mathematics of these simulators typically restricts practitioners to gradient-free optimization [44], which can tune only a handful of parameters before becoming intractable [82, 91]. Even with optimal parameters, physics-based models are inherently limited by discretization (coarse meshes miss fine-scale deformations [67, 50]) and modeling assumptions (linear elasticity fails under large strains [73, 51]). Learning-based methods sidestep these limitations by fitting observations that are challenging to model analytically. However, purely data-driven models are data-hungry [84] and suffer from poor generalization to unseen scenarios [12, 69]. They often lack the physical consistency required for real-world deployment [12] and, without the inductive bias provided by physical priors, they may learn spurious correlations [72].

To combine the interpretability and generalizability of physics-based models with the flexibility of learning-based models, we propose a hybrid simulation framework of Physics-Guided Residual Dynamics (PGRD). As shown in Fig. 1 and explained in Section III, PGRD uses a spring-mass simulator as its physics backbone and a neural network to predict residual corrections to that model. PGRD follows a two-stage training procedure: first, we optimize the simulator's parameters to match real-world observations; and second, we train the neural network to compensate for the discrepancies between the physics model and the data. Note that making this two-stage approach work is nontrivial, as directly injecting learned corrections into simulated states can destabilize the dynamics and cause error accumulation over time. Instead, we predict residual velocities and integrate them forward in time, yielding position corrections that respect the underlying dynamical structure. This velocity-based formulation enables smooth, stable training and simulation. To model temporal dependencies, we incorporate a sliding-window transformer that refines velocity corrections across multiple timesteps. This temporal history enables the network to capture dynamic phenomena such as momentum, while a gating mechanism ensures stable training.

Compared to purely physics-based or learning-based sim-

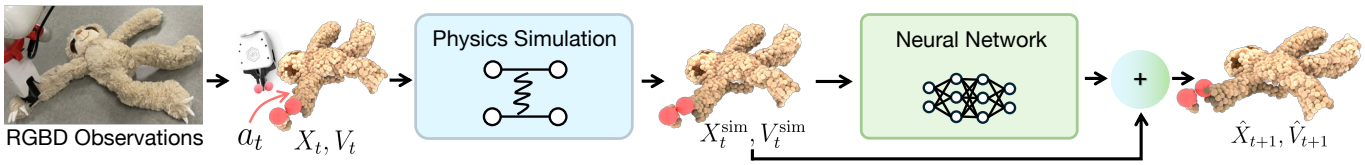

Fig. 1: **Physics-Guided Residual Dynamics Overview.** Given RGBD observations, we extract surface points to instantiate the simulation. Red dots indicate the point where gripper held the object. The framework rolls out an optimized spring-mass physics backbone from the current object state X_t, V_t and robot actions a_t to produce simulation prediction $X_t^{\text{sim}}, V_t^{\text{sim}}$. Subsequently, our residual dynamics network predicts per particle residual velocities, which are added to the simulator velocities and time integrated to obtain the final positions $\hat{X}_{t+1}, \hat{V}_{t+1}$.

ulation methods, PGRD offers several advantages. First, the physics backbone provides priors that improve generalization to new scenarios. Second, it requires significantly less training data than purely learning-based approaches, as the network only needs to predict corrections rather than the entire dynamics from scratch. Third, our approach maintains computational efficiency suitable for real-time applications while achieving higher accuracy than physics-based methods alone.

In Section IV, we validate our approach on a diverse set of real-world deformable objects, including rope, paper, plush toys, flag, and duster. Our experiments demonstrate that PGRD significantly outperforms both purely physics-based and purely learning-based approaches in tracking accuracy across prehensile and non-prehensile manipulation. Notably, only PGRD performs well on the duster, which is a heterogeneous object with a rigid stem and soft feathers. Here, uniform material assumptions in most physics-based deformable simulators fail and learning-based methods struggle with the complexity. Besides a standard tracking evaluation, we also show the advantage of PGRD for action-conditioned video prediction, which we perform by attaching 3D Gaussians to the simulated particles.

Beyond evaluation, we demonstrate two applications of PGRD in Section V. First, PGRD can serve as a forward model for *manipulation planning* via Model Predictive Path Integral (MPPI) control, where candidate action sequences are rolled out and ranked by Chamfer Distance to a target configuration. We demonstrate this on challenging tasks such as cable rerouting through a narrow slot, where PGRD succeeds in 8 out of 10 trials compared to 2 out of 10 for the tuned spring-mass baseline. We further extend planning to a language-conditioned setting, where a goal image generated from a language command replaces the need for a pre-collected target point cloud. Second, PGRD enables *interactive photorealistic simulation* in which users issue manipulation commands and the predicted particle states drive 3D Gaussians to render photorealistic views of the resulting deformation.

In summary, our contributions are as follows. (1) We propose Physics-Guided Residual Dynamics, a hybrid simulation framework that combines an optimizable spring-mass model with learned residual corrections to accurately model deformable object dynamics. (2) We introduce a two-stage training procedure that first optimizes physics parameters using black-box optimization and then trains a neural network to predict residual corrections. (3) We conduct extensive experiments on diverse real-world deformable objects, demonstrating that our approach achieves superior performance compared to existing physics-based and learning-based methods.

II. Related Works

Physics-Based Simulation for Deformable Objects. Traditional physics-based simulation methods rely on analytical models to discretize deformable objects and numerical solvers for the equations of motion of the models. For deformable objects, spring-mass models [4, 43, 39] represent one of the most intuitive and efficient approaches, where objects are modeled as networks of point masses interconnected by springs. Consequently, this representation has been extensively employed to model the dynamics of diverse deformable objects in both computer graphics and robotics [4, 43, 39, 89, 44, 88, 32]. Our work leverages the spring-mass model as a backbone precisely because of its efficiency and interpretability. While more complex approaches like Finite Element Methods (FEM) [65, 53], Position-Based Dynamics (PBD) [54, 49], and Material Point Methods (MPM) [66, 31] offer higher physical fidelity, they incur high computational costs and complexity. We include MPM as a baseline in our experiments to validate this tradeoff.

Despite their physical foundations, the above-mentioned methods require precise material parameters that are difficult to obtain in practice. A recent line of work addresses this challenge by reformulating physics simulators to be compatible with automatic differentiation [39, 68, 26, 9, 14, 20, 25, 30, 59, 60, 46, 42, 18, 32]. These methods identify optimal physics parameters by backpropagating through the simulation process, solving inverse problems to find material properties that best match observed data. While such parameter identification can improve simulation accuracy, it imposes a strong requirement the simulator must be fully differentiable. This is often impractical in contact-rich robotics tasks where discontinuities from collisions [8, 40, 89], friction mode transitions [40, 34, 23], and non-smooth contact geometry [74, 83, 40] lead to exploding or vanishing gradients. In practice, even state-of-the-art differentiable spring-mass approaches such as PhysTwin [32] can fail on highly heterogeneous objects like our duster, as we demonstrate in App. C. Our method does not require the simulator to be differentiable, enabling broader applicability to realistic manipulation scenarios.

Learning-Based Simulation for Deformable Objects. Learning has emerged as a powerful alternative to traditional physics-based simulation, particularly excelling in scenarios requiring real-time performance with complex nonlinear dynamics [80, 10, 36, 90, 79, 6, 1, 27, 11, 16, 47, 77, 78, 45, 28]. These approaches leverage data-driven models, typically neural networks, to learn deformation patterns directly from observation data, bypassing the need for explicit material parameters or complex numerical solvers [37, 52].

Graph Neural Networks (GNNs) have become prominent in this domain due to their natural ability to represent meshes as graphs. GNNs excel at capturing long-range interactions and complex dependencies through message passing [85, 87, 61, 58, 37, 70, 38], allowing them to effectively simulate internal and external forces causing deformation [5, 35]. Notably, GBND [87] establishes a topology over sparse particles, enabling the GNN to efficiently learn and propagate the dynamics across the object. We compare against GBND in our experiments. However, GNN-based methods face significant challenges: the effectiveness of message passing is sensitive to graph structure and is vulnerable to partial observations, while insufficient steps fail to capture global information and excessive steps cause oversmoothing [71, 3].

Recent learning-based methods have explored architectural

improvements, including transformers for handling dense particle systems [62, 75], recurrent neural networks for temporal consistency [48], and attention mechanisms for improved long-range dependency modeling [61]. These methods have demonstrated success across diverse materials, including cloth [39, 41, 58], fluids [61, 37], plasticine [29, 64, 63], and granular materials [37]. The current best method is Particle-Grid Neural Dynamics [86] (PGND), which combines particle representations with spatial grids to learn dynamics while maintaining spatial continuity, and serves as a strong baseline in our experiments. On the down side, approaches like PGND require substantial training data and may struggle with generalization to unseen scenarios or material properties that are significantly different from the training distributions. Our work addresses these limitations by grounding the neural network in physical priors, allowing it to focus on learning only residual corrections rather than the full dynamics from scratch.

Residual Dynamics. The idea of residual learning has been previously employed to model the dynamics of robots [21, 22, 7, 13]. In these works, the robot's model is known, and the residuals primarily reflect simple modeling errors, such as PID gains or backlash. These discrepancies are relatively easy to capture because they often manifest as consistent deviations from the nominal model. In contrast, we model the dynamics of deformable objects, where the residual must compensate for complex, high-dimensional phenomena, including nonlinear stiffness, spatially varying material properties, and contact-rich interactions. The most closely related work in spirit is [2], which models the environment's residual dynamics rather than the robot's. However, it is limited to low-dimensional toy problems such as planar pushing, and its simple architecture does not scale to the high-dimensional particle cloud representations required for deformable object manipulation, making a direct comparison infeasible. PGRD is the first work to extend the residual paradigm to modeling high-dimensional, complex states of deformable objects.

III. Method

This section presents the details of our Physics-Guided Residual Dynamics method, which is illustrated in Fig. 1. We first explain our physics backbone, followed by the neural network architecture.

A. Physics Backbone

Given RGBD observations, we extract surface points to instantiate the simulation. We represent deformable objects using a spring-mass model, wherein the object is discretized into a graph structure consisting of point masses (nodes) interconnected by springs (edges). Each node i has a position $\mathbf{x}_i \in \mathbb{R}^3$ and velocity $\mathbf{v}_i \in \mathbb{R}^3$ that evolve over time according to Newtonian mechanics with a semi-implicit Euler solver. This representation provides computational efficiency and straightforward implementation while capturing essential elastic deformation behaviors. To model manipulation, we apply constraints to particles near the gripper: for prehensile manipulation, they move rigidly with the gripper, whereas for non-prehensile manipulation, the gripper acts as a spherical collision shape that pushes penetrating particles to its surface.

The fidelity of the spring-mass simulator depends critically on five physical parameters collectively denoted as θ: 1) *Stiffness*, which defines the elastic resistance of springs; 2) *Damping*, which controls energy dissipation during motion; 3) *Threshold*, which determines the maximum distance for creating springs between nodes; 4) *Max springs per node*, which limits the connectivity of each mass point for simulation stability; and 5) *Ground friction*, which governs contact interactions with the environment. These parameters collectively encode the material properties and environmental conditions that govern the dynamics of the simulated object.

Given a batch of trajectories sampled from the dataset, each with initial state $(\mathbf{X}_0, \mathbf{V}_0)$ and action sequences $\mathcal{A} = \{a_t\}_{t=1}^{T}$, we roll out the simulator for T time steps with actions $\{a\}_{t=1}^{T}$ under candidate parameters θ to obtain predicted point cloud positions $\{\hat{\mathbf{X}}_t\}_{t=1}^{T}$. At each time step t, we compute a pointwise mean squared error between predicted and ground truth configurations. We average this error over the trajectory horizon to obtain the optimization objective. We minimize this objective using Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [24], which is particularly effective for our problem as it optimizes parameters without requiring gradients. We use CMA-ES because contacts and collisions cause sudden changes in forces, preventing reliable gradient computation in the simulator.

For volumetric objects such as plush toys, modeling only the surface shell often leads to unrealistic collapse, as the lack of internal support points causes the object to flatten under gravity or compression. To address this, we augment the spring-mass model with internal points that provide structural volume. Starting from our multi-view observations, we first reconstruct the complete surface point cloud using RaySt3R [15] and extract a watertight mesh via marching cubes. We then uniformly sample particles within this mesh to populate the interior, ensuring the spring-mass model preserves the object's volume and structural integrity during simulation.

B. Residual Dynamics Framework

While our optimized spring-mass backbone provides a physically grounded baseline, it inherently struggles to capture complex behaviors such as nonlinear stiffness, non-uniform contact friction, and heterogeneous material properties. Instead of attempting to model these intricacies analytically, we introduce a learned residual module to predict the discrepancy between the physics model and observations.

Formally, at each time step t and step size dt, we use the physics backbone to generate an immediate prediction $(X_t^{\text{sim}}, V_t^{\text{sim}})$. We then employ a neural network to predict a per-particle residual velocity $\Delta V_t^{\text{residual}}$ based on the current simulator state and history. We choose a velocity-based residual formulation because it ensures smooth integration and avoids the instability often associated with direct position corrections. The final velocities are obtained by adding the

learned residuals to the simulator's predictions:

$$\hat{V}_{t+1} = V_t^{\text{sim}} + \Delta V_t^{\text{residual}}.$$

These corrected velocities are then time-integrated to obtain the final position updates:

$$\hat{X}_{t+1} = X_t^{\text{sim}} + \Delta V_t^{\text{residual}} \cdot dt.$$

For particles that are rigidly held by grippers, we enforce a boundary condition by zeroing out their residuals. This formulation effectively bridges the gap between the optimized spring-mass model and the observed real-world dynamics.

We train the residual network using a supervised learning objective, minimizing the mean squared error between each particle's predicted position and its corresponding ground truth position across all rollout steps. To ensure the model is robust to error accumulation over long horizons, we employ a multi-step rollout training scheme. During training, we do not reset the simulator to the ground truth state at every step. Instead, we feed the hybrid simulator's *predicted* state from time t back as the input for time $t+1$. This exposes the network to its own past predictions, allowing it to learn how to recover from drift.

C. Network Architecture

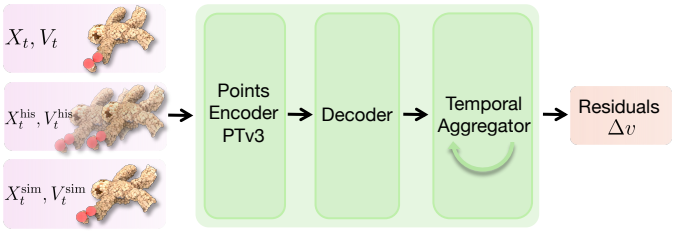

Fig. 2: **Network Architecture.** The points encoder extracts spatiotemporal features, which the decoder projects to estimate initial residual velocity. Finally, the temporal aggregator refines these estimates using a recurrent history to output the final residuals Δv.

To parameterize the residual velocity, we design a network consisting of three primary components: encoder, decoder, and temporal aggregator. The network architecture is shown in Fig. 2. The complete hyperparameters for all the network components are provided in App. D.

Encoder. We employ a Point Transformer V3 (PTv3) [76] architecture to extract per-particle features. At time step t, we construct input features by concatenating the current state (X_t, V_t), history from previous time steps $(X_t^{\text{his}}, V_t^{\text{his}})$, and the simulator's immediate predictions $(X_t^{\text{sim}}, V_t^{\text{sim}})$. The encoder processes these inputs through multiple layers of patch-based self-attention, where each particle attends to others within its local neighborhood. We demonstrate experimentally that the attention mechanism computes position-dependent features for each particle based on its local geometric configuration.

Decoder. To decode per-particle residuals, we utilize a Neural Radiance Field (NeRF) style architecture. We apply Fourier positional encoding to the particle positions and concatenate the resulting embeddings with the features extracted by the encoder. These concatenated features are passed through an MLP decoder to produce an initial velocity correction estimate $\Delta V_t^{\text{initial}}$. Empirically, we found that including the decoder improves performance compared with directly using PTv3 outputs as velocity estimates, as discussed in the decoder ablation in Sec. E.

Temporal Aggregator. To incorporate dynamics from the previous predictions, we refine the base predictions using a sliding-window transformer. At each step t, we update a temporal buffer with the current initial correction $\Delta V_t^{\text{initial}}$. If the buffer contains fewer than W frames, we pad it by replicating the most recent frame. The sequence is projected to a latent embedding, augmented with sinusoidal positional encodings, and processed by the transformer. We extract the output feature vector $\mathbf{h}_t$ corresponding to the final time step and apply two parallel learned projections:

$$\boldsymbol{\delta}_t = \tanh(\text{Proj}_\delta(\mathbf{h}_t)), \quad g_t = \sigma(\text{Proj}_g(\mathbf{h}_t)),$$

where Proj denotes a linear layer, $\tanh$ produces a bounded temporal offset $\boldsymbol{\delta}_t$, and the sigmoid function σ generates gating weights g_t. The final residual velocity is computed as:

$$\Delta V_t^{\text{residual}} = 0.1 \cdot [(1 - g_t) \odot \Delta V_t^{\text{initial}} + g_t \odot \boldsymbol{\delta}_t],$$

where $\odot$ denotes element-wise multiplication. This gating mechanism allows the network to adaptively blend the local base prediction with the temporally refined correction based on the transformer's context.

IV. Experiments

A. Data Collection

We evaluate PGRD on six objects shown in Fig. 3 spanning a range of shapes, material properties, and manipulation types.

Fig. 3: **Experimental Objects** (see text).

a) Rope. A mostly one-dimensional object grasped at one end and manipulated by dragging across the table or lifting and lowering motions.

b) Paper. A thin two-dimensional sheet grasped at the top and manipulated by suspending it in the air while performing waving motions, exhibiting bending and twisting.

c) Plush Toy. A three-dimensional volumetric object with long, highly flexible limbs that are challenging to model.

d) Duster. A heterogeneous object with a rigid stem and highly deformable feathers. This is especially challenging because it requires different dynamics predictions for the rigid stem compared to the flexible feathers.

e) Flag. A cloth flag, demonstrating large area deformation dynamics that are challenging to model due to the waving

Method	Metric	Rope	Paper	Plush Toy	Duster	Flag	Teddy Toy
Spring-Mass [32]		$4.4_{\pm 2.0}$	$2.3_{\pm 2.0}$	$3.2_{\pm 0.8}$	$3.8_{\pm 2.0}$	$5.7_{\pm 2.5}$	$5.8_{\pm 3.4}$
MPM [66]		$7.3_{\pm 2.5}$	$15.5_{\pm 4.2}$	$7.4_{\pm 1.8}$	$6.0_{\pm 2.6}$	$23.1_{\pm 7.0}$	–
GBND [87]	MDE ↓	$5.5_{\pm 1.7}$	$3.0_{\pm 1.4}$	$7.7_{\pm 2.6}$	$5.1_{\pm 2.2}$	$30.9_{\pm 6.2}$	$1.5_{\pm 0.4}$
PGND [86]		$3.3_{\pm 1.8}$	$2.1_{\pm 0.5}$	$4.0_{\pm 1.3}$	$3.8_{\pm 0.1}$	$3.2_{\pm 1.4}$	$1.6_{\pm 1.3}$
Ours		$\mathbf{2.6_{\pm 1.31}}$	$\mathbf{1.7_{\pm 0.7}}$	$\mathbf{2.7_{\pm 0.4}}$	$\mathbf{2.9_{\pm 1.8}}$	$\mathbf{2.8_{\pm 1.2}}$	$\mathbf{1.3_{\pm 0.3}}$
Spring-Mass [32]		$4.5_{\pm 2.4}$	$1.7_{\pm 1.2}$	$2.9_{\pm 0.7}$	$4.7_{\pm 2.5}$	$9.1_{\pm 4.7}$	$5.1_{\pm 3.0}$
MPM [66]		$6.1_{\pm 2.4}$	$14.8_{\pm 6.5}$	$7.1_{\pm 1.7}$	$5.2_{\pm 1.7}$	$23.6_{\pm 10.4}$	–
GBND [87]	CD ↓	$6.6_{\pm 2.4}$	$5.0_{\pm 1.7}$	$6.5_{\pm 1.8}$	$6.3_{\pm 2.8}$	$49.1_{\pm 13.0}$	$2.7_{\pm 0.3}$
PGND [86]		$2.7_{\pm 1.4}$	$2.2_{\pm 0.9}$	$3.2_{\pm 0.8}$	$4.0_{\pm 0.1}$	$4.3_{\pm 2.4}$	$1.6_{\pm 1.0}$
Ours		$\mathbf{2.3_{\pm 1.3}}$	$\mathbf{1.3_{\pm 0.6}}$	$\mathbf{2.6_{\pm 0.3}}$	$\mathbf{3.8_{\pm 2.5}}$	$\mathbf{4.3_{\pm 2.2}}$	$\mathbf{1.3_{\pm 0.2}}$
Spring-Mass [32]		$2.4_{\pm 1.3}$	$1.8_{\pm 1.2}$	$1.5_{\pm 0.4}$	$1.9_{\pm 1.2}$	$4.9_{\pm 2.6}$	$2.9_{\pm 1.8}$
MPM [66]		$3.9_{\pm 1.6}$	$11.4_{\pm 5.5}$	$3.9_{\pm 0.9}$	$2.8_{\pm 1.2}$	$19.4_{\pm 7.8}$	–
GBND [87]	EMD ↓	$3.0_{\pm 1.3}$	$2.1_{\pm 0.9}$	$3.1_{\pm 1.1}$	$2.8_{\pm 1.4}$	$24.5_{\pm 6.3}$	$0.8_{\pm 0.2}$
PGND [86]		$1.4_{\pm 0.6}$	$1.5_{\pm 0.5}$	$1.6_{\pm 0.4}$	$1.6_{\pm 0.4}$	$2.3_{\pm 1.3}$	$0.8_{\pm 0.6}$
Ours		$\mathbf{1.2_{\pm 0.6}}$	$\mathbf{1.4_{\pm 0.7}}$	$\mathbf{1.4_{\pm 0.2}}$	$\mathbf{1.4_{\pm 1.1}}$	$\mathbf{2.2_{\pm 1.2}}$	$\mathbf{0.6_{\pm 0.1}}$

TABLE I. **Tracking accuracy across diverse objects.** Physics-Guided Residual Dynamics achieves the lowest tracking error across all objects and metrics, outperforming both physics-based methods and learning-based approaches.

motion. We chose it because its top stick prevents the cloth's surface from collapsing. We model the stick as a rigid constraint, assuming that all points along the stick follow the gripper trajectory.

f) **Teddy Toy.** A volumetric object with higher rigidity than the plush toy. We deform it by poking rather than holding it, representing non-prehensile manipulation.

We collect data using UFACTORY xArm 7 following the data collection procedure from [86]. Four Intel RealSense D455 cameras capture synchronized multi-view observations while the object is manipulated. We apply Grounded SAM 2 for object segmentation and CoTracker [33] for 2D trajectory prediction. These 2D trajectories are lifted to 3D using depth information from all camera views, and persistent point tracks are extracted through an iterative rollout approach. Additional details are provided in App. A.

B. Evaluation Tasks and Metrics

We evaluate our method in two ways: (1) by measuring dynamics prediction accuracy in 3D space using tracking metrics; and (2) by performing action-conditioned video prediction with the help of 3D Gaussian Splatting and assessing rendering quality using visual metrics.

Tracking Metrics. These measure dynamics prediction accuracy in 3D space. We employ three metrics computed between predicted and ground truth point clouds:

- **Mean Distance Error (MDE)** measures the average distance (in cm) between corresponding points in predicted and ground truth configurations, providing a direct assessment of positional accuracy.
- **Chamfer Distance (CD)** computes the bidirectional nearest neighbor distance (in cm) between point clouds, capturing coverage and precision without requiring point correspondences.
- **Earth Mover's Distance (EMD)** quantifies the minimum cost (in cm) of transforming one point cloud into another, providing a holistic measure of distributional similarity.

Visual Metrics. To evaluate rendering quality for action-conditioned video prediction, we assess the rendered images obtained by updating Gaussian positions with our predicted point clouds. We report three metrics:

- **J-Score (IoU)** measures intersection over union of predicted and ground truth masks, quantifying spatial overlap.
- **F Score** evaluates contour accuracy, determining how well the boundary of the predicted object mask matches the boundary of the ground truth.
- **Learned Perceptual Image Patch Similarity (LPIPS)** assesses perceptual similarity using features extracted from predicted and ground truth images, capturing visual differences closer to human perception.

C. Baselines

We compare PGRD against four baselines: two analytical physics-based simulations and two learning-based approaches.
Spring-Mass Model [32]: This is the same as the physics backbone we use for our model. Improvements over this baseline demonstrate that the learned residual model effectively captures the dynamics missed by the physics simulator alone.
Material Point Method (MPM) [66] : MPM combines Eulerian and Lagrangian representations to handle material behaviors and topological changes. This hybrid approach discretizes materials into particles while using a background grid for computing spatial derivatives and enforcing conservation laws. MPM is particularly effective for materials exhibiting both solid and fluid-like behaviors but requires careful parameter tuning and can be computationally expensive.
Graph Based Neural Dynamics (GBND) [87] : This is a learned approach that employs Graph Neural Networks to learn object dynamics directly from data without explicit material parameters. The object is represented as a graph where message passing mechanisms propagate information between nodes to capture long range interactions.

Fig. 4: **Qualitative comparison of PGRD on dynamics prediction.** Columns shaded in green show the input state with gripper positions (red spheres) and the ground truth object configuration after taking the action. The remaining columns show predictions from each method. PGRD accurately captures deformations across all objects. For the duster, only PGRD maintains stem rigidity while allowing deformation for the feathers. MPM does not support non-prehensile manipulation, so its results are unavailable for the Teddy Toy.

Particle-Grid Neural Dynamics (PGND) [86] : This represents the state-of-the-art in learning-based methods, employing a hybrid representation of particles and spatial grids inspired by MPM to model deformable object dynamics. In this method, particles capture the object geometry while the spatial grid discretizes the 3D domain to ensure spatial continuity and improve learning efficiency.

D. Results

As explained in Section IV-B, we evaluate PGRD in two ways: 3D tracking accuracy and action-conditioned video prediction.

Dynamics and Tracking Accuracy. We first analyze our model's ability to track and simulate object states over time, with results summarized in Tab. I. PGRD consistently achieves the lowest error across all objects and metrics, demonstrating that learned residual corrections on physics backbones improve accuracy beyond either pure physics or pure learning approaches. Among physics methods, the optimized spring-mass model generally outperforms MPM, particularly for objects with coherent structure like the rope and plush toy. Among learning methods, PGND substantially outperforms GBND across all objects, indicating that particle representations with physics priors are more effective than graph neural approaches for deformable object simulation.

Qualitatively, Fig. 4 reveals consistent patterns in how different approaches handle deformation. The optimized spring-mass model captures overall motion but lacks the flexibility

to represent local variations in material properties, often appearing too stiff. MPM often struggles to maintain structural cohesion, with object particles separating rather than behaving as connected solids. Learning-based methods show complementary strengths: PGND generates smooth predictions but accumulates drift over time, while GBND struggles with graph topology, producing unphysical artifacts. PGRD leverages its physics backbone for stability while using learned residuals to correct local dynamics, maintaining both plausibility and accuracy throughout the rollout.

The qualitative results in Fig. 4 can give further insights into the limitations of different approaches. MPM fails catastrophically for thin objects like paper and flag, causing them to tear down and lose structure. This happens because it relies on grid-based discretization, which can cause material points to lose cohesion when particle distribution is non-uniform. The duster, with its rigid handle and soft feathers, reveals distinct failure modes across all baselines: the optimized spring-mass model cannot assign spatially varying stiffness and causes the rigid stem to bend unnaturally; MPM fails to maintain volume with material points collapsing toward the floor; GBND predicts erroneous upward motion for all feather particles; and PGND incorrectly compresses the rigid stem. Only PGRD simulates the duster correctly. For the teddy, under non-prehensile manipulation, the object slips out of the optimized spring-mass model. MPM cannot realistically simulate the constraints for non-prehensile manipulation, thus we exclude it from this comparison. GBND predicts minimal particle displacement, producing an overly stiff response. PGND shows the gripper's entire surface in our visualization, but the gripper fingers themselves are not visible because they have pushed into the teddy's interior. Only our PGRD successfully captures the deformation, properly maintaining contact while allowing realistic compression.

Action Conditioned Video Prediction. Given an initial observation and a sequence of actions, we evaluate whether accurate dynamics predictions translate to realistic future frame generation. We leverage 3D Gaussian Splatting (3DGS) as the rendering module: Gaussian primitives are fit to the observed point cloud at the initial timestep, where each Gaussian is parameterized by its 3D position, covariance, opacity, and color. As PGRD predicts future point cloud configurations, we update only the Gaussian positions and covariances while keeping appearance attributes fixed. This allows the rendering to maintain visual consistency with the initial observation while reflecting the predicted deformations. To render a frame at timestep t, we project the updated Gaussians onto the camera view and alpha-blend them in depth order to produce the output image. As shown in Table II, PGRD consistently achieves the best $\mathcal{J}$-Score and $\mathcal{F}$-Score across all objects, demonstrating better spatial overlap and foreground detection compared to both physics-based and learning-based baselines. PGRD also maintains the lowest LPIPS scores, indicating that the physics backbone enables the residual model to preserve object appearance during rendering.

V. Applications

A. Planning using PGRD

In our planning experiments, we use PGRD as a forward model for model-based control: given a state s_t and an action a_t, it predicts the next state s_{t+1}, where actions specify gripper positions that interact with the object. To plan over this forward model, we integrate PGRD with Model Predictive Path Integral (MPPI) control [19]. At each timestep, MPPI samples a batch of candidate action sequences $\{\tau_i\}$ spanning a planning horizon H, rolls out PGRD to predict future states, and updates the action distribution toward lower-cost regions. Following standard MPC practice, only the first action is executed before replanning with the newly observed state. We use the Chamfer Distance (CD) between the predicted and a pre-collected target point cloud as the planning objective.

Standard Manipulation Tasks. We evaluate planning on the rope with four start and goal configurations shown in Fig. 5 (top): (1) lifting the rope up, (2) lowering it, (3) rotating it such

Method	Metric	Rope	Paper	Plush Toy	Duster	Flag	Teddy Toy
Spring-Mass [32]		$3.17_{\pm 1.51}$	$6.82_{\pm 1.47}$	$5.30_{\pm 0.87}$	$6.66_{\pm 0.82}$	$4.60_{\pm 1.95}$	$5.43_{\pm 1.99}$
MPM [66]		$2.44_{\pm 1.45}$	$4.51_{\pm 0.94}$	$5.22_{\pm 0.88}$	$5.92_{\pm 0.87}$	$3.78_{\pm 1.12}$	–
GBND [87]	$\mathcal{J}$-Score / IoU ($\times 10$) ↑	$0.53_{\pm 0.55}$	$5.57_{\pm 1.40}$	$3.70_{\pm 1.49}$	$5.13_{\pm 1.04}$	$5.08_{\pm 1.42}$	$7.52_{\pm 0.66}$
PGND [86]		$3.18_{\pm 2.13}$	$7.08_{\pm 0.90}$	$5.86_{\pm 0.93}$	$6.69_{\pm 0.46}$	$5.47_{\pm 1.66}$	$7.13_{\pm 1.02}$
Ours		$\mathbf{3.95_{\pm 1.69}}$	$\mathbf{7.53_{\pm 0.82}}$	$\mathbf{6.21_{\pm 0.78}}$	$\mathbf{6.77_{\pm 0.83}}$	$\mathbf{5.54_{\pm 1.89}}$	$\mathbf{7.82_{\pm 1.14}}$
Spring-Mass [32]		$6.59_{\pm 1.60}$	$4.71_{\pm 2.61}$	$5.76_{\pm 1.12}$	$5.32_{\pm 0.97}$	$2.08_{\pm 1.81}$	$4.73_{\pm 2.72}$
MPM [66]		$4.92_{\pm 2.45}$	$4.73_{\pm 0.82}$	$5.13_{\pm 0.86}$	$5.00_{\pm 1.08}$	$2.49_{\pm 0.97}$	–
GBND [87]	$\mathcal{F}$-Score ($\times 10$) ↑	$1.95_{\pm 1.91}$	$4.95_{\pm 1.35}$	$3.59_{\pm 1.15}$	$4.11_{\pm 1.09}$	$2.54_{\pm 2.04}$	$7.71_{\pm 0.96}$
PGND [86]		$5.87_{\pm 2.96}$	$5.39_{\pm 1.65}$	$5.53_{\pm 1.11}$	$5.29_{\pm 0.97}$	$2.65_{\pm 1.98}$	$7.27_{\pm 1.54}$
Ours		$\mathbf{7.38_{\pm 1.85}}$	$\mathbf{5.91_{\pm 1.76}}$	$\mathbf{6.30_{\pm 1.07}}$	$\mathbf{5.51_{\pm 1.02}}$	$\mathbf{2.68_{\pm 2.14}}$	$\mathbf{8.03_{\pm 1.75}}$
Spring-Mass [32]		$2.6_{\pm 0.9}$	$5.6_{\pm 2.2}$	$5.9_{\pm 2.0}$	$7.0_{\pm 1.9}$	$9.9_{\pm 3.0}$	$3.6_{\pm 1.7}$
MPM [66]		$3.8_{\pm 1.4}$	$6.5_{\pm 1.4}$	$7.7_{\pm 1.7}$	$6.8_{\pm 1.7}$	$9.1_{\pm 2.5}$	–
GBND [87]	LPIPS ($\times 100$) ↓	$5.2_{\pm 1.7}$	$5.7_{\pm 1.3}$	$10.5_{\pm 2.0}$	$6.4_{\pm 1.3}$	$8.8_{\pm 1.3}$	$2.2_{\pm 0.5}$
PGND [86]		$2.6_{\pm 1.3}$	$5.2_{\pm 1.1}$	$6.5_{\pm 1.8}$	$6.6_{\pm 0.9}$	$8.9_{\pm 1.9}$	$2.0_{\pm 0.8}$
Ours		$\mathbf{2.3_{\pm 1.1}}$	$\mathbf{4.9_{\pm 1.4}}$	$\mathbf{5.4_{\pm 2.1}}$	$\mathbf{5.9_{\pm 1.9}}$	$\mathbf{8.5_{\pm 2.9}}$	$\mathbf{1.7_{\pm 0.9}}$

TABLE II. **Action-conditioned video prediction with dynamics prediction and 3DGS.** PGRD produces the most accurate visual predictions, demonstrating superior rendering quality.

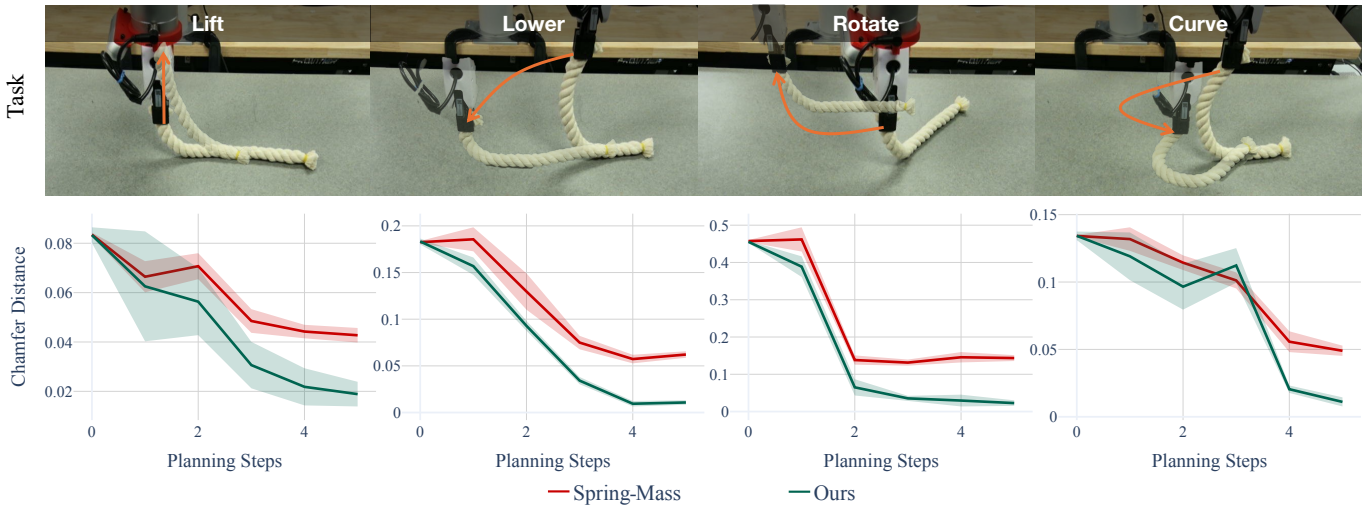

Fig. 5: **Planning results on rope manipulation tasks.** Top row shows start and end configurations for four tasks. The bottom row shows the Chamfer Distance to the target during MPPI planning. PGRD (green) converges faster and achieves lower final error than the optimized spring-mass baseline (red) across all tasks.

that one end remains roughly stationary while the other moves, and (4) deforming it into a curved shape. We compare PGRD against the spring-mass backbone, providing both methods with the same MPPI planning budget. We run 10 trials for each task using the same start and end configurations. As shown in Fig. 5, PGRD's CD decreases smoothly and saturates at a value close to the target, whereas the spring-mass model converges more slowly and somtimes fails to reach a comparable final error. Additional planning results on the plush toy are provided in App. B.

Cable Rerouting. We further test PGRD on a more challenging task: cable rerouting through a narrow slot. This requires the robot to move the rope and thread it through the slot. We run 10 trials with different start and end configurations. The target configuration is provided to the planner.

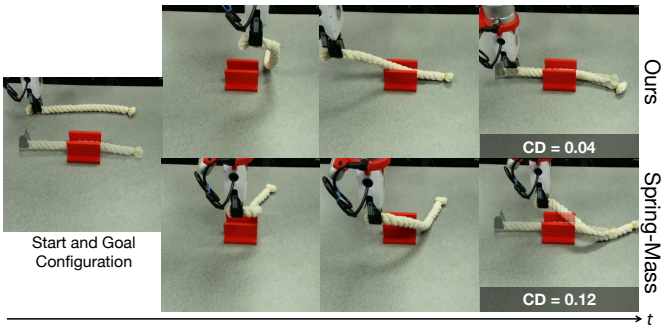

Fig. 6: **Cable Rerouting Execution.** Comparison of PGRD (top) and the Spring-Mass backbone (bottom) for rope rerouting through a slot. PGRD navigates the rope through the opening while the Spring-Mass backbone exhibits repeated collisions with the slot edges.

As demonstrated in Fig. 6, PGRD successfully navigates the rope through the slot. The learned residuals compensate for inaccuracies in the physics backbone, enabling precise trajectory execution that avoids collisions with the edges. In contrast, the spring-mass backbone struggles with these fine dynamics. Small errors cause the rope to collide with the edges, forcing the planner to retry the motion multiple times before completing the task.

Quantitatively, Fig. 7 shows the distributions of start and end configurations overlaid (left) and comparisons of chamfer distances (right). PGRD achieves significantly lower CD than the optimized spring-mass. We also evaluate success rates for this task. A trial is considered successful if the rope passes entirely through the slot and the gripper remains below the slot height in the final configuration. Under this criterion, PGRD succeeds in 8/10 trials, whereas the spring-mass model succeeds in only 2/10 trials. These results demonstrate that the improved model enables reliable execution of tasks that would otherwise be infeasible.

Language-Conditioned Planning. The planning experiments above require a target point cloud, so it must be collected beforehand. To relax this requirement, we integrate PGRD with a language-conditioned goal-generation pipeline, enabling the planner to operate solely on language commands.

Given an initial RGB image and a language command (e.g., "pass the rope through the red slot"), we use Nano Banana Pro to synthesize a goal image. Fig. 9 shows an example of initial and generated goal images. Note that here we use a single camera rather than the four in the standard planning

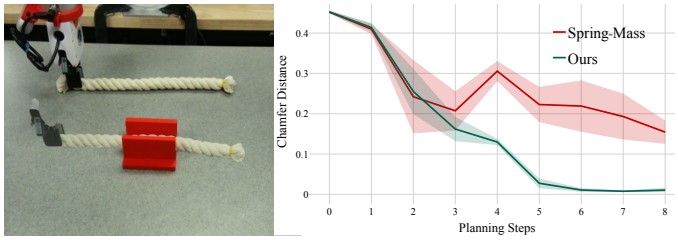

Fig. 7: **Cable rerouting through narrow slot.** Left: distribution of start and end configurations. Right: Chamfer Distance during MPPI planning steps. PGRD achieves lower distance than the spring-mass baseline, successfully threading the rope through the slot in 8 out of 10 trials compared to 2 out of 10 for the baseline.

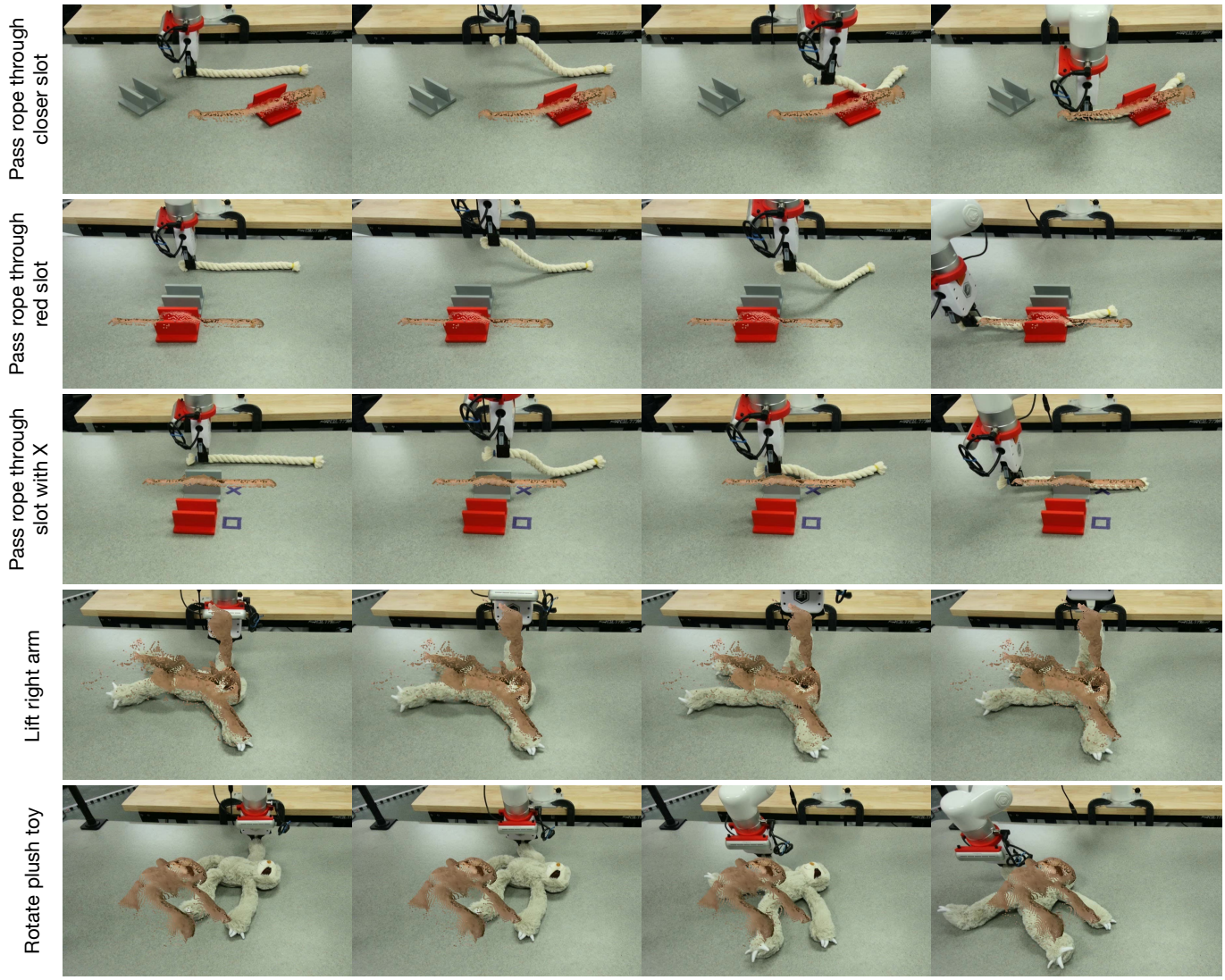

Fig. 8: **Execution of PGRD with generated goals.** Each row shows a task, with the target point cloud overlaid. The target is obtained from a language-conditioned goal image. PGRD successfully executes a range of tasks, including passing rope through a specified slot and moving a plush toy to target configurations.

Fig. 9: **Language-conditioned goal image generation.** Given the initial image (left) and the command "Pass the rope through the slot closer to the camera," Nano Banana Pro synthesizes a goal image (right) that correctly places the rope through the intended slot.

experiments, since the goal image is generated from a single viewpoint. To estimate depth from the generated image, we use Depth Anything V2 [81]. Following Patel *et al.* [57], we resolve the scale and shift ambiguity inherent to monocular depth estimation by fitting an affine transform to align the predicted depth to the initial depth map from the depth camera. We restrict the alignment to background table points, which remain consistent in both images. Since the object appears in a different configuration in the goal image than in the initial image, including object points for this alignment would degrade the quality of the alignment. The resulting depth map is unprojected into 3D with known camera parameters to obtain the target point cloud, which is passed directly to the MPPI planner.

To determine where to grasp the object, we use AnyGrasp [17] to generate grasp candidates on the initial point cloud. For the rope, we select the candidate closest to the right end. For the plush toy, the language command implies a specific region to manipulate, which we identify by extracting DINOv2 [56] features from both the initial and goal images to establish dense correspondences. We locate the object point with the largest displacement between the two images and select the AnyGrasp candidate closest to that point.

Fig. 8 shows real-world execution across tasks, where the object state progressively aligns with the overlaid target point cloud. Beyond removing the need to physically collect the goal

configuration, this pipeline opens up a richer space of goals, including spatial references such as "pass rope through closer slot" and "lift right arm" of the plush toy.

B. Interactive Photo-Realistic Simulation.

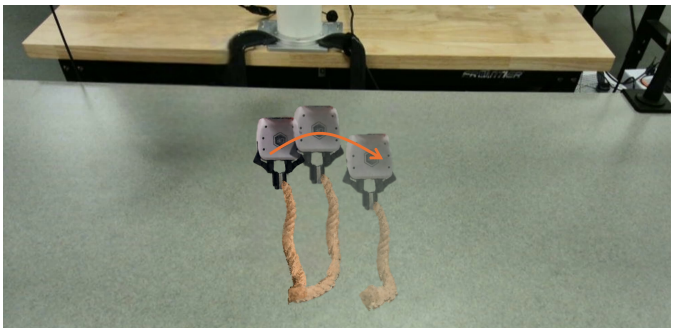

Fig. 10: **Interactive rope manipulation with PGRD rendered using 3DGS.** The visualization overlays three timesteps, with later configurations in lighter opacity. As PGRD predicts the dynamics, the 3D Gaussians attached to the particles are updated to render images that maintain visual consistency.

Finally, as shown in Fig. 10, PGRD supports interactive, photo-realistic simulation by combining its predicted dynamics with 3D Gaussian Splatting. Users issue manipulation commands via keyboard input; at each step, the physics backbone propagates the object state, the learned residual network applies per-particle corrections, and the updated Gaussians are rendered to produce a photo-realistic view of the resulting deformation. Because only Gaussian positions are updated while appearance attributes remain fixed, the rendering maintains visual consistency with the initial observation throughout the interaction.

VI. Conclusions

This paper presented Physics-Guided Residual Dynamics (PGRD), a hybrid simulation framework that bridges physics-based and learning-based approaches for deformable object simulation. By combining an optimized spring-mass backbone with learned velocity residuals, PGRD achieves superior accuracy while maintaining physical plausibility. Our velocity-based residuals, combined with a sliding-window transformer for temporal aggregation, yield stable predictions, avoiding the instability issues plaguing naive residual learning. Extensive experiments across real-world objects demonstrate that PGRD consistently outperforms a variety of SOTA methods. PGRD handles challenging scenarios involving heterogeneous materials, volumetric deformations, and contact-rich interactions where competing methods fail. Beyond tracking, PGRD also shows promise for applications like action-conditioned video prediction, planning, and interactive simulation.

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] Anurag Ajay, Jiajun Wu, Nima Fazeli, Maria Bauza, Leslie P Kaelbling, Joshua B Tenenbaum, and Alberto Rodriguez. Augmenting physical simulators with stochastic neural networks: Case study of planar pushing and bouncing. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3066–3073. IEEE, 2018.

[3] Hugo Attali, Davide Buscaldi, and Nathalie Pernelle. Rewiring techniques to mitigate oversquashing and oversmoothing in gnns: A survey. *arXiv preprint arXiv:2411.17429*, 2024.

[4] David Baraff and Andrew Witkin. Large steps in cloth simulation. In *Proceedings of 25th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '98)*, pages 43 – 54. SIGGRAPH, July 1998.

[5] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. *Advances in neural information processing systems*, 29, 2016.

[6] Dominik Bauer, Zhenjia Xu, and Shuran Song. Doughnet: A visual predictive model for topological manipulation of deformable objects. In *European Conference on Computer Vision*, pages 92–108. Springer, 2024.

[7] Leonard Bauersfeld, Elia Kaufmann, Philipp Foehn, Sihao Sun, and Davide Scaramuzza. Neurobem: Hybrid aerodynamic quadrotor model. *arXiv preprint arXiv:2106.08015*, 2021.

[8] Onur Beker, Nico Gürtler, Ji Shi, A René Geist, Amirreza Razmjoo, Georg Martius, and Sylvain Calinon. A smooth analytical formulation of collision detection and rigid body dynamics with contact. *arXiv preprint arXiv:2503.11736*, 2025.

[9] Filipe De Avila Belbute-Peres, Thomas Economon, and Zico Kolter. Combining differentiable pde solvers and graph neural networks for fluid flow prediction. In *international conference on machine learning*, pages 2402–2411. PMLR, 2020.

[10] Zhenfang Chen, Kexin Yi, Yunzhu Li, Mingyu Ding, Antonio Torralba, Joshua B Tenenbaum, and Chuang Gan. Comphy: Compositional physical reasoning of objects and events from videos. *arXiv preprint arXiv:2205.01089*, 2022.

[11] Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Advances in neural information processing systems*, 31, 2018.

[12] Loris Di Natale, Bratislav Svetozarevic, Philipp Heer, and Colin N Jones. Physically consistent neural networks for building thermal modeling: theory and analysis. *Applied Energy*, 325:119806, 2022.

[13] Runpei Dong, Ziyan Li, Xialin He, and Saurabh Gupta. Learning humanoid end-effector control for open-

vocabulary visual loco-manipulation. *arXiv preprint arXiv:2602.16705*, 2026.

[14] Tao Du, Kui Wu, Pingchuan Ma, Sebastien Wah, Andrew Spielberg, Daniela Rus, and Wojciech Matusik. Diffpd: Differentiable projective dynamics. *ACM Transactions on Graphics (ToG)*, 41(2):1–21, 2021.

[15] Bardienus P. Duisterhof, Jan Oberst, Bowen Wen, Stan Birchfield, Deva Ramanan, and Jeffrey Ichnowski. Rayst3r: Predicting novel depth maps for zero-shot object completion, 2025. URL https://arxiv.org/abs/2506.05285.

[16] Ben Evans, Abitha Thankaraj, and Lerrel Pinto. Context is everything: Implicit identification for dynamics adaptation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2642–2648. IEEE, 2022.

[17] Hao-Shu Fang, Chenxi Wang, Hongjie Fang, Minghao Gou, Jirong Liu, Hengxu Yan, Wenhai Liu, Yichen Xie, and Cewu Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics*, 39(5):3929–3945, 2023.

[18] Junpeng Gao, Mike Y Michelis, Andrew Spielberg, and Robert K Katzschmann. Sim-to-real of soft robots with learned residual physics. *IEEE Robotics and Automation Letters*, 2024.

[19] Carlos E Garcia, David M Prett, and Manfred Morari. Model predictive control: Theory and practice—a survey. *Automatica*, 25(3):335–348, 1989.

[20] Moritz Geilinger, David Hahn, Jonas Zehnder, Moritz Bächer, Bernhard Thomaszewski, and Stelian Coros. Add: Analytically differentiable dynamics for multi-body systems with frictional contact. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020.

[21] Florian Golemo, Adrien Ali Taiga, Aaron Courville, and Pierre-Yves Oudeyer. Sim-to-real transfer with neural-augmented robot simulation. In *Conference on Robot Learning*, pages 817–828. PMLR, 2018.

[22] Joshua Gruenstein, Tao Chen, Neel Doshi, and Pulkit Agrawal. Residual model learning for microrobot control. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7219–7226. IEEE, 2021.

[23] Mathew Halm and Michael Posa. Set-valued rigid body dynamics for simultaneous frictional impact. *arXiv preprint arXiv:2103.15714*, 2021.

[24] Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary computation*, 9(2):159–195, 2001.

[25] Eric Heiden, Miles Macklin, Yashraj Narang, Dieter Fox, Animesh Garg, and Fabio Ramos. Disect: A differentiable simulation engine for autonomous robotic cutting. *arXiv preprint arXiv:2105.12244*, 2021.

[26] Philipp Holl, Vladlen Koltun, and Nils Thuerey. Learning to control pdes with differentiable physics. *arXiv preprint arXiv:2001.07457*, 2020.

[27] Suning Huang, Qianzhong Chen, Xiaohan Zhang, Jiankai Sun, and Mac Schwager. Particleformer: A 3d point cloud world model for multi-object, multi-material robotic manipulation. *arXiv preprint arXiv:2506.23126*, 2025.

[28] 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.

[29] Zhiao Huang, Yuanming Hu, Tao Du, Siyuan Zhou, Hao Su, Joshua B Tenenbaum, and Chuang Gan. Plasticinelab: A soft-body manipulation benchmark with differentiable physics. *arXiv preprint arXiv:2104.03311*, 2021.

[30] Krishna Murthy Jatavallabhula, Miles Macklin, Florian Golemo, Vikram Voleti, Linda Petrini, Martin Weiss, Breandan Considine, Jérôme Parent-Lévesque, Kevin Xie, Kenny Erleben, et al. gradsim: Differentiable simulation for system identification and visuomotor control. *arXiv preprint arXiv:2104.02646*, 2021.

[31] Chenfanfu Jiang, Craig Schroeder, Andrew Selle, Joseph Teran, and Alexey Stomakhin. The affine particle-in-cell method. *ACM Transactions on Graphics (TOG)*, 34(4): 1–10, 2015.

[32] 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. *arXiv preprint arXiv:2503.17973*, 2025.

[33] Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6013–6022, 2025.

[34] Hajun Kim, Dongyun Kang, Min-Gyu Kim, Gijeong Kim, and Hae-Won Park. Online friction coefficient identification for legged robots on slippery terrain using smoothed contact gradients. *IEEE Robotics and Automation Letters*, 2025.

[35] Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *International conference on machine learning*, pages 2688–2697. Pmlr, 2018.

[36] Thomas Kipf, Elise Van der Pol, and Max Welling. Contrastive learning of structured world models. *arXiv preprint arXiv:1911.12247*, 2019.

[37] Yunzhu Li, Jiajun Wu, Russ Tedrake, Joshua B Tenenbaum, and Antonio Torralba. Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids. *arXiv preprint arXiv:1810.01566*, 2018.

[38] Yunzhu Li, Jiajun Wu, Jun-Yan Zhu, Joshua B Tenenbaum, Antonio Torralba, and Russ Tedrake. Propagation networks for model-based control under partial observation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 1205–1211. IEEE, 2019.

[39] Junbang Liang, Ming Lin, and Vladlen Koltun. Differentiable cloth simulation for inverse problems. *Advances in neural information processing systems*, 32, 2019.

[40] Quentin Le Lidec, Louis Montaut, Yann de Mont-Marin, Fabian Schramm, and Justin Carpentier. End-to-end and highly-efficient differentiable simulation for robotics. *arXiv preprint arXiv:2409.07107*, 2024.

[41] Xingyu Lin, Yufei Wang, Zixuan Huang, and David Held. Learning visible connectivity dynamics for cloth smoothing. In *Conference on Robot Learning*, pages 256–266. PMLR, 2022.

[42] Min Liu, Gang Yang, Siyuan Luo, and Lin Shao. Softmac: Differentiable soft body simulation with forecast-based contact model and two-way coupling with articulated rigid bodies and clothes. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12008–12015. IEEE, 2024.

[43] Tiantian Liu, Adam W Bargteil, James F O'Brien, and Ladislav Kavan. Fast simulation of mass-spring systems. *ACM Transactions on Graphics (TOG)*, 32(6):1–7, 2013.

[44] Bryn Lloyd, Gábor Székely, and Matthias Harders. Identification of spring parameters for deformable object simulation. *IEEE transactions on visualization and computer graphics*, 13(5):1081–1094, 2007.

[45] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *2024 International Conference on 3D Vision (3DV)*, pages 800–809. IEEE, 2024.

[46] Pingchuan Ma, Tao Du, Joshua B Tenenbaum, Wojciech Matusik, and Chuang Gan. Risp: Rendering-invariant state predictor with differentiable simulation and rendering for cross-domain parameter estimation. *arXiv preprint arXiv:2205.05678*, 2022.

[47] Pingchuan Ma, Peter Yichen Chen, Bolei Deng, Joshua B Tenenbaum, Tao Du, Chuang Gan, and Wojciech Matusik. Learning neural constitutive laws from motion observations for generalizable pde dynamics. In *International Conference on Machine Learning*, pages 23279–23300. PMLR, 2023.

[48] Xiao Ma, David Hsu, and Wee Sun Lee. Learning latent graph dynamics for visual manipulation of deformable objects. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8266–8273. IEEE, 2022.

[49] Miles Macklin, Matthias Müller, Nuttapong Chentanez, and Tae-Yong Kim. Unified particle physics for real-time applications. *ACM Transactions on Graphics (TOG)*, 33 (4):1–12, 2014.

[50] Dominique Madier. An introduction to the fundamentals of mesh generation in finite element analysis. Technical report, FEA Academy, Montreal, Canada, December 2023. URL https://www.fea-academy.com/pdf/FEA%20Academy%20-%20The%20Fundamentals%20of%20Mesh%20Generation%20in%20Finite%20Element%20Analysis.pdf. Published by FEA Academy.

[51] L Angela Mihai and Alain Goriely. How to characterize a nonlinear elastic material? a review on nonlinear constitutive parameters in isotropic finite elasticity. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2207):20170607, 2017.

[52] Damian Mrowca, Chengxu Zhuang, Elias Wang, Nick Haber, Li F Fei-Fei, Josh Tenenbaum, and Daniel L Yamins. Flexible neural representation for physics prediction. *Advances in neural information processing systems*, 31, 2018.

[53] Matthias Müller and Markus H Gross. Interactive virtual materials. In *Graphics interface*, volume 2004, pages 239–246, 2004.

[54] Matthias Müller, Bruno Heidelberger, Marcus Hennix, and John Ratcliff. Position based dynamics. *Journal of Visual Communication and Image Representation*, 18(2): 109–118, 2007.

[55] J Krishna Murthy, Miles Macklin, Florian Golemo, Vikram Voleti, Linda Petrini, Martin Weiss, Breandan Considine, Jérôme Parent-Lévesque, Kevin Xie, Kenny Erleben, et al. gradsim: Differentiable simulation for system identification and visuomotor control. In *International conference on learning representations*, 2020.

[56] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[57] Shivansh Patel, Shraddhaa Mohan, Hanlin Mai, Unnat Jain, Svetlana Lazebnik, and Yunzhu Li. Robotic manipulation by imitating generated videos without physical demonstrations. *arXiv preprint arXiv:2507.00990*, 2025.

[58] Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter Battaglia. Learning mesh-based simulation with graph networks. In *International conference on learning representations*, 2020.

[59] Yiling Qiao, Junbang Liang, Vladlen Koltun, and Ming Lin. Differentiable simulation of soft multi-body systems. *Advances in Neural Information Processing Systems*, 34:17123–17135, 2021.

[60] Junior Rojas, Eftychios Sifakis, and Ladislav Kavan. Differentiable implicit soft-body physics. *arXiv preprint arXiv:2102.05791*, 2021.

[61] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International conference on machine learning*, pages 8459–8468. PMLR, 2020.

[62] Yidi Shao, Chen Change Loy, and Bo Dai. Transformer with implicit edges for particle-based physics simulation. In *European conference on computer vision*, pages 549–564. Springer, 2022.

[63] Haochen Shi, Huazhe Xu, Samuel Clarke, Yunzhu Li, and Jiajun Wu. Robocook: Long-horizon elasto-plastic object manipulation with diverse tools. *arXiv preprint arXiv:2306.14447*, 2023.

[64] Haochen Shi, Huazhe Xu, Zhiao Huang, Yunzhu Li, and Jiajun Wu. Robocraft: Learning to see, simulate, and shape elasto-plastic objects in 3d with graph networks.

The International Journal of Robotics Research, 43(4): 533–549, 2024.

[65] Eftychios Sifakis and Jernej Barbic. Fem simulation of 3d deformable solids: a practitioner's guide to theory, discretization and model reduction. In *Acm siggraph 2012 courses*, pages 1–50. 2012.

[66] Alexey Stomakhin, Craig Schroeder, Lawrence Chai, Joseph Teran, and Andrew Selle. A material point method for snow simulation. *ACM Transactions on Graphics (TOG)*, 32(4):1–10, 2013.

[67] Shijie Tan, Hongjun Zhou, and Jinjin Zheng. A hybrid model method for accurate surface deformation and incision based on fem and pbd. *Scientific Programming*, 2021(1):8343312, 2021.

[68] Kiwon Um, Robert Brand, Yun Raymond Fei, Philipp Holl, and Nils Thuerey. Solver-in-the-loop: Learning from differentiable physics to interact with iterative pde-solvers. *Advances in neural information processing systems*, 33:6111–6122, 2020.

[69] Vaiva Vasiliauskaite and Nino Antulov-Fantulin. Generalization of neural network models for complex network dynamics. *Communications Physics*, 7(1):348, 2024.

[70] Changhao Wang, Yuyou Zhang, Xiang Zhang, Zheng Wu, Xinghao Zhu, Shiyu Jin, Te Tang, and Masayoshi Tomizuka. Offline-online learning of deformation model for cable manipulation with graph neural networks. *IEEE Robotics and Automation Letters*, 7(2):5544–5551, 2022.

[71] Xinjue Wang, Esa Ollila, and Sergiy A Vorobyov. Graph neural network sensitivity under probabilistic error model. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 2146–2150. IEEE, 2022.

[72] Zizhao Wang, Xuesu Xiao, Zifan Xu, Yuke Zhu, and Peter Stone. Causal dynamics learning for task-independent state abstraction. *arXiv preprint arXiv:2206.13452*, 2022.

[73] Robert J Webster III and Bryan A Jones. Design and kinematic modeling of constant curvature continuum robots: A review. *The International Journal of Robotics Research*, 29(13):1661–1683, 2010.

[74] Keenon Werling, Dalton Omens, Jeongseok Lee, Ioannis Exarchos, and C Karen Liu. Fast and feature-complete differentiable physics for articulated rigid bodies with contact. *arXiv preprint arXiv:2103.16021*, 2021.

[75] William F Whitney, Jacob Varley, Deepali Jain, Krzysztof Choromanski, Sumeet Singh, and Vikas Sindhwani. Modeling the real world with high-density visual particle dynamics. *arXiv preprint arXiv:2406.19800*, 2024.

[76] 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.

[77] Yilin Wu, Wilson Yan, Thanard Kurutach, Lerrel Pinto, and Pieter Abbeel. Learning to manipulate deformable objects without demonstrations. *arXiv preprint arXiv:1910.13439*, 2019.

[78] Zhenjia Xu, Jiajun Wu, Andy Zeng, Joshua B Tenenbaum, and Shuran Song. Densephysnet: Learning dense physical object representations via multi-step dynamic interactions. *arXiv preprint arXiv:1906.03853*, 2019.

[79] Shangjie Xue, Shuo Cheng, Pujith Kachana, and Danfei Xu. Neural field dynamics model for granular object piles manipulation. In *Conference on Robot Learning*, pages 2821–2837. PMLR, 2023.

[80] Mengyuan Yan, Yilin Zhu, Ning Jin, and Jeannette Bohg. Self-supervised learning of state estimation for manipulating deformable linear objects. *IEEE robotics and automation letters*, 5(2):2372–2379, 2020.

[81] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

[82] Xintong Yang, Ze Ji, and Yu-Kun Lai. Differentiable physics-based system identification for robotic manipulation of elastoplastic materials. *The International Journal of Robotics Research*, page 02783649251334661, 2025.

[83] Xiaohan Ye, Kui Wu, Zherong Pan, and Taku Komura. Efficient differentiable contact model with long-range influence. *arXiv preprint arXiv:2509.20917*, 2025.

[84] Rose Yu and Rui Wang. Learning dynamical systems from data: An introduction to physics-guided deep learning. *Proceedings of the National Academy of Sciences*, 121(27):e2311808121, 2024.

[85] 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.

[86] 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.

[87] Mingtong Zhang, Kaifeng Zhang, and Yunzhu Li. Dynamic 3d gaussian tracking for graph-based neural dynamics modeling. *arXiv preprint arXiv:2410.18912*, 2024.

[88] Licheng Zhong, Hong-Xing Yu, Jiajun Wu, and Yunzhu Li. Reconstruction and simulation of elastic objects with spring-mass 3d gaussians. In *European Conference on Computer Vision*, pages 407–423. Springer, 2024.

[89] Yaofeng Desmond Zhong, Jiequn Han, Biswadip Dey, and Georgia Olympia Brikis. Improving gradient computation for differentiable physics simulation with contacts. In *Learning for dynamics and control conference*, pages 128–141. PMLR, 2023.

[90] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.

[91] Yicheng Zhu, David Yang, and Yangming Lee. Deformable and fragile object manipulation: A review and prospects. *Sensors*, 25(17):5430, 2025.

Appendix

We structure the supplement into the following sections:

- A Data collection and processing pipeline for obtaining consistent particle trajectories from multi-view RGBD observations.
- B Planning experiments on plush toy manipulation demonstrating PGRD's effectiveness on volumetric objects with flexible limbs.
- C Comparison to PhysTwin on heterogeneous objects, highlighting limitations of parameter optimization versus learned residual corrections.
- D Network architecture details and hyperparameter specifications for the encoder, decoder, and temporal aggregator.
- E Ablation study on decoder architecture, isolating the contribution of the NeRF-style decoder to tracking accuracy.

A. Data Collection and Processing

We mount four Intel RealSense D455 cameras around the workspace to capture synchronized RGBD observations at 30 Hz. The cameras are calibrated to a shared world coordinate frame using a checkerboard calibration procedure, enabling the fusion of observations across views. For each camera view, we apply Grounded SAM 2 to extract object masks. Using text prompts containing object descriptions, it detects and segments the object in the first frame of a sequence and propagates the mask across subsequent frames. We employ CoTracker [33] as the tracking model due to its stable tracking performance. It predicts a set of 2D trajectories, initialized from uniformly sampled grid locations within the first-frame object mask. For each pixel in the segmented region at time t, we compute its 2D displacement to time $t+1$ from the predicted trajectories, and convert it to a 2D velocity by dividing by the time step.

For each camera view, we lift the 2D pixel velocities to 3D using the corresponding depth measurements. Specifically, for a pixel at position (u, v) with depth d and 2D velocity (v_u, v_v), we compute the 3D velocity by back-projecting both the current and next pixel positions to 3D coordinates using camera intrinsics, then computing the difference. This yields per-pixel 3D velocities in the camera frame, which we transform to the world frame using the calibration parameters. We aggregate velocities across all camera views by averaging the 3D velocity estimates for overlapping spatial regions, thereby improving robustness to noise and partial observations from individual views.

Given the 3D point cloud with per-point velocities at each timestep, we extract temporally consistent particle trajectories using an iterative rollout procedure. Starting from frame 0, we initialize particles at the point cloud positions $\{x_i^0\}$. For each subsequent timestep t, we propagate particles forward using their current velocities: $x_i^{t+1} = x_i^t + v_i^t \cdot \Delta t$. To update velocities at the new positions, we perform k-nearest neighbor search (with $k = 5$) between the propagated particle positions and the observed point cloud at time $t + 1$. Each particle's velocity is updated to the mean velocity of its k nearest neighbors in the observed cloud. This iterative process maintains correspondence across frames while being robust to tracking noise, yielding persistent point tracks suitable for training our dynamics model. The entire processing pipeline runs at approximately 2 Hz for sequences with 1000 particles.

B. Planning on Plush Toy

We extend our planning experiments to the plush toy, a volumetric object with long, flexible limbs. The plush toy requires accurate prediction of both volumetric deformation and limb dynamics, testing the framework's ability to handle three-dimensional objects. Similar to rope, we design four manipulation objectives with varying complexity, as visualized in Fig. 11 (top): (1) lifting the object, (2) lowering it, (3) rotating it, and (4) bending it into a curved configuration. Following the same experimental setup, we conduct 10 trials per objective with identical initial and target states, allocating equal computational budget to both PGRD and the spring-mass baseline. The results in Fig. 11 demonstrate that PGRD exhibits consistent convergence behavior, reaching target configurations with reasonable residual error. The spring-mass baseline shows slower error reduction and occasionally plateaus at suboptimal solutions, particularly for tasks involving complex limb deformation, where the learned residuals provide crucial corrections to the physics backbone predictions.

C. Comparison to PhysTwin

PhysTwin [32] represents another approach to modeling deformable objects using spring-mass representations. Similar to PGRD, it employs a spring-mass simulator as its physical backbone. However, it attempts to recover spatially varying material parameters through first-order gradient-based optimization, with a differentiable simulator. While this strategy can identify heterogeneous material properties in principle, we find that it struggles with objects exhibiting extreme variations in properties, such as our duster which has a rigid stem and highly deformable feathers.

We evaluate PhysTwin on the duster object to assess its ability to handle heterogeneous materials. The execution results are shown in Fig. 12. As time progresses from left to right in the figure, the simulation exhibits catastrophic failure. In the leftmost frame, the duster appears correctly initialized. However, by the middle frame, the rigid stem begins to bend unnaturally at the junction between the stem and feathers. By the final frame, the rigid stem completely collapses, with the entire structure folding in on itself. This failure mode reveals limitation of relying solely on first-order optimization to infer spatially varying parameters.

In contrast, our approach as discussed in Sec. IV-D successfully models the duster by combining a tuned spring-mass simulator with learned residual corrections. Rather than

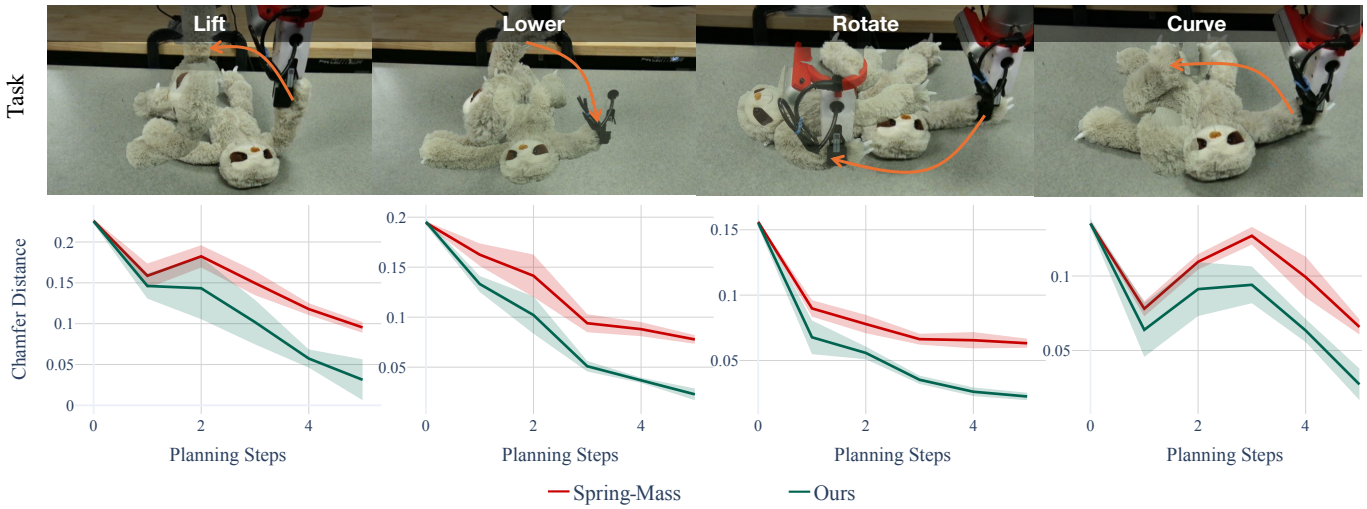

Fig. 11: **Planning results on plush toy manipulation tasks.** Top row displays the initial and target configurations for four manipulation objectives. Bottom row plots the Chamfer Distance evolution throughout MPPI planning. PGRD (green) consistently achieves superior

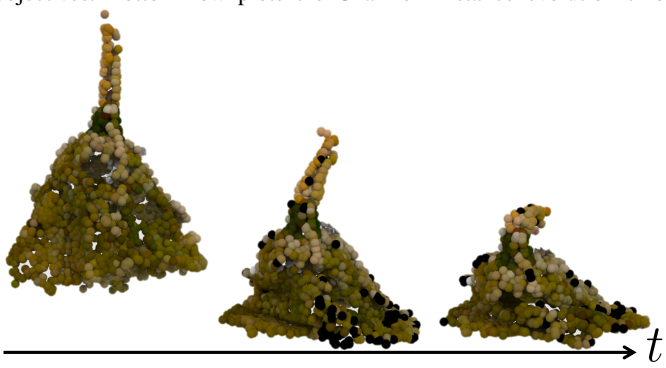

Fig. 12: **PhysTwin execution on duster.** Time progresses from left to right. The duster initially maintains its shape, but the rigid stem begins to bend unnaturally at intermediate time steps and completely collapses by the final frame. This demonstrates that spatially varying stiffness parameters recovered through first-order optimization fail to maintain the structural integrity of heterogeneous objects.

attempting to capture all material heterogeneity through parameter optimization alone, our residual network learns to compensate for discrepancies between the physics model and reality.

D. Network Architecture Details

Our network network consists of three primary components: an encoder, a decoder, and a temporal aggregator. The hyperparameter values for each component are provided in Table III. The encoder uses a Point Transformer V3 (PTv3) [76] architecture to extract per-particle features through patch-based self-attention, where the feature dimension determines the size of the latent representation. The input channels correspond to the concatenated position and velocity information from the current state, history timesteps, and the physics simulator predictions. The decoder employs a conditional NeRF-style MLP to map spatial locations to residual velocities, where hidden layers refers to the number of fully connected layers, hidden dimension specifies the width of each layer, and output channels indicates the dimensionality of the predicted residual velocity (3 for xyz components). The decoder takes as input the per-particle features from the encoder, concatenated with Fourier positional encodings of particle positions. The temporal aggregator employs a transformer encoder to refine predictions using temporal context. Here, the embedding dimension defines the size of the latent space for temporal features, attention heads specifies the number of parallel attention mechanisms, transformer layers indicates the depth of the transformer stack, and feedforward dimension determines the width of the intermediate feedforward network within each transformer layer. Finally, the gating scale controls the magnitude of temporal corrections applied to the base velocity predictions, preventing large adjustments that could destabilize the simulation.

Parameter	**Value**
Encoder (PTv3)	
Feature dimension	64
Input channels	18
Decoder	
Hidden layers	2
Hidden dimension	64
Output channels	3
Temporal Aggregator	
Embedding dimension (d)	64
Attention heads (H)	4
Transformer layers (L)	2
Feedforward dimension (d_{ff})	128
Gating scale	0.1

TABLE III. Network architecture hyperparameters.

E. Ablation Study: Decoder Architecture

In Sec. III-C, we describe the use of a NeRF-style decoder to produce per-particle residual velocities from the PTV3 features. We empirically found that including this decoder

improves performance compared with directly using PTv3 outputs as velocity estimates. To isolate the decoder's contribution, we compare two architectural variants on the rope object: (1) **PTv3 direct**, where the encoder features are directly projected to velocity corrections without the NeRF-style decoder, and (2) **PGRD (full)**, which is our model. Both variants use the same encoder and temporal aggregator.

The results demonstrate that the decoder improves tracking accuracy. Without the decoder, the model achieves 2.8 cm MDE, 0.026 CD, and 0.13 EMD on rope. In contrast, the full architecture with the decoder achieves 2.6 cm MDE, 0.023 CD, and 0.012 EMD. These results confirm that the positional encoding and MLP decoder effectively refine the spatially varying residual corrections, particularly improving the distributional similarity measured by EMD.