Deform360: A Massive Multi-view Visuotactile Dataset for Deformable World Models

Wanjia Fu, George Konidaris

Brown University

Abstract. Predicting object dynamics (i.e., world modeling) is a fundamental challenge for robotic manipulation, and modeling deformable objects presents a particularly difficult case due to their high-dimensional state spaces and complex material properties. While current world models approach this through two distinct paradigms: learning the dynamics over the 2D pixel space or more explicit 3D geometric space. A systematic understanding of their relative strengths and limitations remains elusive due to the lack of diverse, large-scale real-world data. To address this, we present Deform360, a large-scale visuotactile dataset featuring 198 daily-life objects, 1,980 interaction sequences, and over 215 hours of observations from 41 surround-view cameras and bimanual tactile grippers to capture both global motion and contact-induced local deformations. Leveraging a novel markerless visuotactile 3D tracking pipeline to extract dense geometry and motion, we systematically evaluate current state-of-the-art world models, comparing 2D video models against 3D particle models. Finally, we provide a preliminary demonstration indicating the real-world applicability of our dataset by performing robot planning tasks using Model Predictive Control (MPC) on deformable objects. Our analysis reveals key insights into the trade-offs between structural priors and scalability, providing a solid benchmark for future research in generalizable deformable object-centric world modeling.

Keywords: World Models · Deformable Object Manipulation · Visuotactile Perception

1 Introduction

World modeling (*i.e.*, action-conditioned future prediction) has emerged as a powerful paradigm for robots to simulate environments and plan actions [3, 4, 32, 38]. Despite recent progress, accurately predicting the behavior of deformable objects remains a significant challenge [14, 19, 58]. Deformable bodies possess theoretically infinite degrees of freedom, and contact-induced local deformations are frequently occluded by end-effectors or the object itself [13, 31, 54]. Consequently, accurately capturing deformable dynamics has attracted significant interests, particularly through tactile sensing to observe occluded physical interactions and provide critical ground-truth signals.

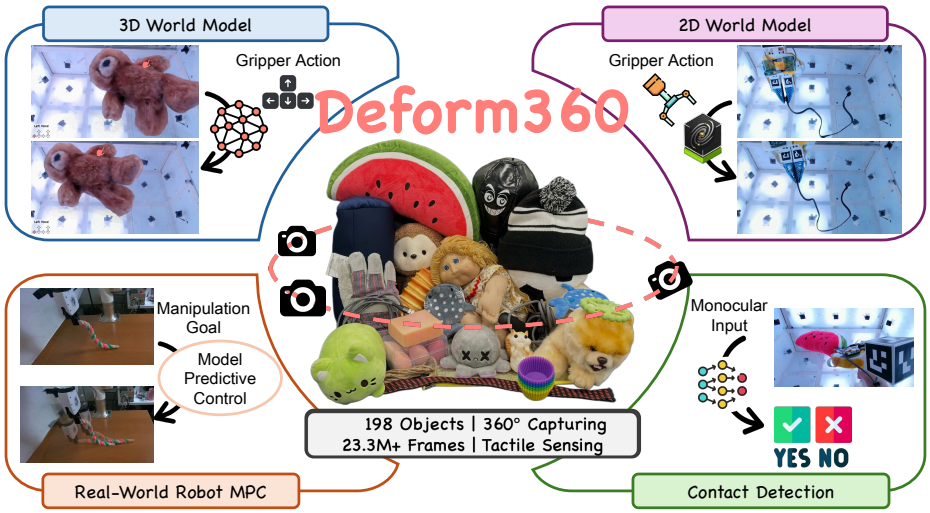

Fig. 1: Overview of Deform360. We collect a massive multi-view visuotactile dataset with 198 deformable objects (a subset is shown), supporting 2D and 3D world models, contact detection, and real-world robot planning tasks.

Current models predicting deformable dynamics follow two paradigms: predicting dynamics over 2D pixel space (*e.g.*, video generation) [35, 50, 66, 68] or 3D geometric space [9, 30, 45, 81, 82]. Evaluating their relative strengths requires diverse, large-scale, multi-view real-world datasets, which are essential for verifying generalizability and extracting ground-truth 3D states. However, existing datasets often lack object diversity [6, 9, 30, 51, 80], rely on synthetic data [5, 32, 45, 46, 64, 84], or miss high-fidelity annotations and contact-induced deformation modeling [12, 38, 43, 47].

To address this, we present Deform360, a massive and diverse visuotactile dataset for large-scale deformable dynamics modeling. The dataset encompasses **198** daily-use objects across **1,980** interaction sequences, spanning a wide spectrum of materials. Our capture system features 41 surround-view cameras and bimanual tactile-equipped UMI grippers [11, 86], ensuring 360° observability of global dynamics and fine-grained contact-induced deformations. Delivering over **215.7 hours** and **23.3 million frames** of visuotactile recordings, Deform360 provides a foundational benchmark for 3D dynamics evaluation of world models. Leveraging this data, we develop a novel markerless 3D tracking pipeline to scalably generate high-fidelity particle annotations across varying materials.

Using these annotations, we formulate three core tasks to evaluate deformable world modeling: 1) contact prediction, demonstrating the utility of our synchronized tactile data by modeling the coupling between visual observations and physical contact events. 2) world model benchmarking, providing a systematic evaluation of current state-of-the-art 2D video models [50] against 3D particle models [27, 30, 82] across multiple generalization settings. 3) robot planning,

Table 1: Comparison of real-world datasets containing deformable objects. We benchmark existing datasets based on key modalities for learning dynamics, including availability of 3D mesh data (Mesh), camera calibration (Cam Calib), *markerless* (whether tracking does not rely on physical markers), inclusion of tactile data (Tactile), multi-view coverage (360° View), and number of camera views (Views). We also report the scale of each dataset in terms of deformable objects (# obj.), total frames (# frames), number of manipulation actions (# actions), and number of object categories (# cat.). '✓' indicates the presence of a specific modality, and '✗' indicates its absence.

Dataset	Mesh	Calib.	Markerless	Tactile	360°	Views	# obj.	# frames	# actions	# cat.
ClothSim2Real [6]	✓	✗	✓	✗	✗	1	3	∼4.9k	2	1
ClothTrack [12]	✗	✗	✗	✗	✗	5	8	∼173k	4	1
DDER [9]	✗	✗	✗	✗	✗	11	5	∼1.8M	5	2
Robo360 [47]	✗	✓	✓	✗	✓	86	17	∼2M	7	7
PokeFlex [51]	✓	✓	✓	✓	✓	6	18	21.3k	2	6
DOT [43]	✓	✓	✗	✗	✓	42	22	∼689k	9	3
HMDO [76]	✓	✓	✓	✗	✓	10	12	21.6k	2	1
HCOS [21]	✓	✗	✓	✗	✗	1	27	64	3	1
PhysTwin [30]	✗	✓	✓	✗	✗	3	11	7.2k	4	4
PGND [82]	✓	✓	✓	✗	✓	4	8	∼1M	6	6
Deform360 (Ours)	✓	✓	✓	✓	✓	41	198	∼23.3M	13	17

evaluating the real-world applicability of our dataset by deploying learned world models in a model predictive control (MPC) framework for deformable object manipulation. Our evaluation reveals that while 3D particle models excel in low-data regimes due to structured physical priors, 2D video models leverage their scalability to achieve better generalization when provided with extensive data.

In summary, this paper makes four key contributions: 1) A large-scale, multi-view, visuotactile real-world dataset of deformable object interactions, including 198 diverse daily-use objects with 1,980 interactions from two tactile-equipped UMI grippers, captured using 41 surround-view cameras. 2) High-quality 3D reconstructions and tracking results for each episode, derived from a markerless multi-view visuotactile perception pipeline. 3) A systematic evaluation comparing state-of-the-art 2D video world models and 3D particle dynamics models on dynamics reconstruction and future prediction, revealing critical insights into their respective trade-offs in structural priors and scalability. 4) A preliminary demonstration of real-world applicability through robot planning using MPC on deformable objects collected in Deform360.

2 Related Work

2.1 Deformable Dataset

Existing datasets for deformable objects often focus on specific categories or lack essential modalities for comprehensive world modeling (Tab. 1). Early benchmarks primarily targeted thin-shell objects such as cloth [6, 12, 21] or linear objects like ropes and cables [9,31,67]. While these datasets provide valuable insights into specific manipulation tasks, they are often limited in object diversity and camera viewpoints. More recent work has attempted to scale object diversity

and sensory richness. Robo360 [47] provides a large-scale omnispective dataset with 86 camera views, covering various categories like bags, food, and ropes, though it lacks high-fidelity annotations and tactile sensing. PokeFlex [51] introduces a multi-modal dataset of volumetric deformable objects and interaction wrenches, focusing on poking and dropping actions. HMDO [76] and DOT [43] focus on hand manipulation and textureless object tracking, respectively, providing detailed surface reconstructions but with a restricted range of interaction types. Simulation-based repositories like ClothesNet [84] offer massive object diversity but lack real-world physical feedback and interaction complexity.

In contrast, Deform360 provides massive-scale real-world visuotactile sequences, including high-fidelity 2D/3D annotations, 41-view calibrated and synchronized video and tactile data for 198 diverse objects. With 23.3 million frames and 215.7 hours of data, it significantly exceeds the scale and sensory richness of existing benchmarks, providing a robust foundation for dynamics evaluation.

2.2 World Models

Action-conditioned 2D Video Models. World models enable robots to learn environmental dynamics and plan actions through imagination [22–24,37,57,72]. Recently, video diffusion models have emerged as a powerful paradigm for world simulation [2, 7, 15–18, 20, 35, 42, 50, 53, 61, 66, 70]. Action-conditioned variants like [68], Vid2World [26], PAN [62], and Cosmos [50] add action conditioning to the video models. These methods predict the 3D dynamics *implicitly* in the latent space [24,25,37,85]. While they exhibit remarkable scalability and capture physics from internet-scale data through massive pre-training, they often suffer from 3D and temporal inconsistencies during long-horizon predictions [38].

3D World Models. To incorporate stronger structural priors, 3D dynamics models represent objects using explicit geometry, such as meshes or particles [28, 45, 46, 56, 82]. These models derive transition functions through either physics-based optimization or data-driven learning. Physics-based approaches utilize differentiable simulators [1, 10, 49, 73] – such as DiffCloth [44], PhysTwin [30], and PhysGaussian [75] – to embed physical priors like mass conservation, achieving long-term stability but remaining susceptible to state estimation noise. Conversely, learning-based approaches approximate dynamics using MLPs [82], Graph Neural Networks [45, 67, 83] or Transformers [27]. While free from specific simulator inaccuracies [3], they typically struggle in low-data scenarios compared to physics-driven methods.

3 The Deform360 Dataset

The Deform360 dataset is constructed to address the scarcity of high-fidelity, large-scale real-world data for deformable object modeling. Unlike existing benchmarks that often rely on synthetic environments or restricted object categories, Deform360 provides a dense, synchronized visuotactile record of complex physical interactions across a vast distribution of materials. 198 daily-life objects and 1,980 interaction sequences are included.

3.1 Visuotactile Capture System

Each interaction sequence is recorded using a synchronized visuotactile rig designed for 360° observability. The visual modality consists of RGB videos from $N = 41$ cameras, denoted as $\{\mathbf{V}_n\}_{n=1}^{N} \in \mathbb{R}^{T\times 3\times H\times W}$, where T represents the number of frames. The cameras operate at a resolution of $H, W = 720 \times 1280$ at 30 frames per second. Each camera n is characterized by its intrinsic matrix $\mathbf{K}_n \in \mathbb{R}^{3\times 3}$ and extrinsic matrix $\mathbf{E}_n \in \mathbb{R}^{4\times 4}$, representing its fixed pose in the world coordinate system. This dense spatial coverage is critical for resolving the contact-induced local deformations that are often occluded by end-effectors or the object's own geometry. Simultaneously, we record tactile signals $\mathbf{T} \in \mathbb{R}^{T\times D}$ from sensors mounted on the bimanual UMI grippers [11, 86]. These signals are software-synchronized with the visual stream at 30 Hz to provide a continuous record of interaction forces and local material response.

3.2 Object Taxonomy and Diversity

The dataset features 198 diverse daily-use objects, spanning a wide range of materials and geometric complexities. This scale – totaling 1,980 interaction sequences – represents an order of magnitude increase over existing real-world deformable benchmarks. To facilitate research into generalizable dynamics, we categorize the object set based on their physical material responses:

- **1D Deformables:** Ropes, cables, and wires with varying stiffness and thicknesses.
- **2D Deformables:** Diverse fabrics, cloths, and thin-shell materials like airbags and paper.
- **3D Volumetric Deformables:** Plush toys, stuffed animals, and foam-based objects that exhibit significant shape change.

3.3 Interaction Protocol and Scale

Data collection is performed using tactile-equipped UMI grippers. The protocol encompasses a wide range of manipulation primitives, including unimanual poking and squeezing, as well as complex bimanual maneuvers such as stretching, folding, and twisting. For each episode, we record the robot's proprioception (6D pose and openness) alongside $\{\mathbf{V}_n\}$ and $\mathbf{T}$. This enables precise action reasoning, ensuring the dataset is suitable for training world models that must understand both global object motion and fine-grained contact physics.

3.4 Dataset Statistics

The Deform360 dataset comprises 198 unique daily-life objects with a total of 1,980 interaction sequences (5 unimanual and 5 bimanual episodes per object). Our capture setup utilizes 41 camera viewpoints recording at 720p resolution and 30 FPS. This results in 74,850 raw videos and a total of **23.3 million**

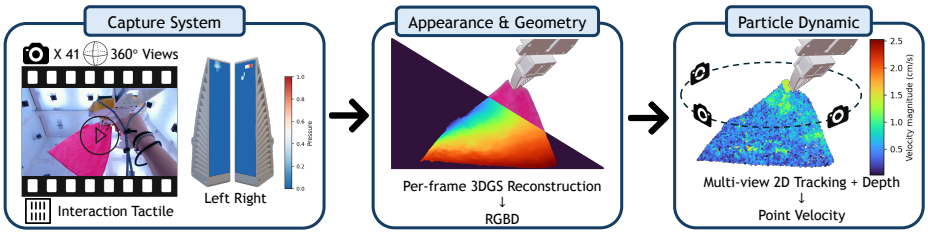

Fig. 2: Overview of our annotation pipeline. We convert multi-view video and tactile streams into dense annotations for deformable object dynamics. We first reconstruct high-fidelity dynamic geometry using per-frame 3D Gaussian Splatting. Then, we perform markerless 2D tracking, lift the tracks into 3D, and optimize for temporal and multi-view consistency and physical plausibility.

frames across all views (614,490 frames per individual viewpoint). The single-view video duration totals 20,483 seconds, which corresponds to **215.7 hours** of cumulative multi-view footage. The average duration of each interaction episode is 10.34 seconds. Furthermore, synchronized tactile streams are recorded for each interaction episode, providing a rich source of physical contact data.

4 Annotation Pipeline

The goal of our annotation pipeline is to convert visuotactile observations (multi-view video and tactile streams) into dense annotations for deformable object dynamics. While Gaussian Splatting-based (GS) approaches [48, 71] achieve high-fidelity dynamic reconstruction, they are optimized for rendering rather than physical tracking; independent Gaussian motions may result in trajectories that lack temporal consistency [19, 75]. To address this, we decouple per-frame geometry recovery from temporal tracking: GS provides high-quality per-frame geometry, while 2D tracking results from each view are lifted into 3D using this geometry to ensure multi-view consistency. We further refine the tracking results with tactile signals, which provide ground-truth contact information. An overview of our pipeline is shown in Figure 2.

4.1 Visuotactile Preprocessing

The pipeline begins with precise multi-view calibration using ArUco grids, ensuring metric-scale 3D reconstruction. Since the raw RGB videos $\{\mathbf{V}_n\}$ are captured with lens distortion, we first undistort all camera streams and refer to the undistorted videos as $\{\mathbf{V}_n\}$ for simplicity throughout the paper. We track the gripper's 6D pose and openness from multiple views using ArUco markers and aggregate them using the RANSAC algorithm.

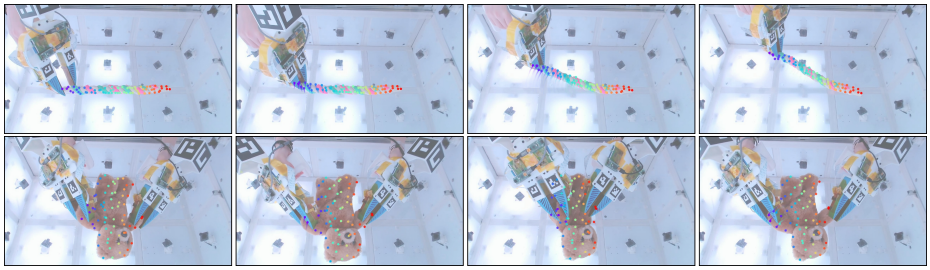

Fig. 3: Visualization of our particle tracking system across different object categories.

4.2 Dynamic Geometry Reconstruction

We reconstruct the 3D dynamic geometry of the interacting objects on a per-frame basis using 3D Gaussian Splatting (3DGS) [34], which represents a 3D scene as a set of K anisotropic Gaussians. Each Gaussian k is characterized by its mean position $\boldsymbol{\mu}_k \in \mathbb{R}^3$, a covariance matrix $\boldsymbol{\Sigma}_k = \mathbf{R}_k \mathbf{S}_k \mathbf{S}_k^T \mathbf{R}_k^T$ (decomposed into a rotation matrix $\mathbf{R}_k$ and a scaling matrix $\mathbf{S}_k$ to ensure positive semi-definiteness), an opacity $\alpha_k \in [0, 1]$, and spherical harmonic coefficients encoding view-dependent color c_k. The color and depth maps are obtained through rasterization of the 3D Gaussians [78].

To ensure our 3DGS models exclusively capture the object, we deploy segmentation models [8, 59] to generate object masks $\mathbf{M}_{\text{obj},n} \in \{0, 1\}^{T \times H \times W}$ for each camera view n. These masks are then used to filter the RGB videos into segmented, object-only streams $\mathbf{S}_n = \mathbf{V}_n \odot \mathbf{M}_{\text{obj},n}$, where $\odot$ denotes element-wise multiplication broadcasted over the color channel. The 3DGS model for each frame is then optimized following the standard objective proposed in [34], where the loss function $\mathcal{L}_{\text{gs}}$ is defined as a weighted combination of the $\mathcal{L}_1$ loss and the structural similarity index (SSIM) loss: $\mathcal{L}_{\text{gs}} = (1 - \lambda_{\text{gs}})\mathcal{L}_1 + \lambda_{\text{gs}}\mathcal{L}_{\text{SSIM}}$, where $\mathcal{L}_1$ is the mean absolute error between the rendered image and the segmented object frame $\mathbf{S}_n$ at time t. Following the original implementation, we set $\lambda_{\text{gs}} = 0.2$. This per-frame reconstruction provides high-fidelity geometry without requiring complex deformation fields.

4.3 Markerless Particle Tracking

Once the per-frame geometry is recovered, we perform dense motion estimation using a multi-view 2D-to-3D lifting approach. We utilize CoTracker3 [33] to track M points sampled from the segmented object mask $\mathbf{M}_{\text{obj},n,t}$ across all camera views, providing persistent 2D trajectories $\mathbf{u}_{n,t} \in \mathbb{R}^{M \times 2}$ for each view n at time t. Following PGND [82], our tracking is performed over a sliding window to effectively address invisible points caused by self-occlusions and gripper interactions. While CoTracker3 offers robust 2D correspondence, 3D lifting is critical to ensure geometric consistency. For each camera view, we back-project the 2D tracks into 3D space using the rendered depth map $\mathbf{D}_{n,t}$ from our optimized

3DGS. The 3D position $\mathbf{P}_{n,t}$ lifted from a 2D track $\mathbf{u}_{n,t}$ is given by:

$$\mathbf{P}_{n,t} = \mathbf{E}_n^{-1}\mathbf{D}_{n,t}(\mathbf{u}_{n,t})\mathbf{K}_n^{-1}\tilde{\mathbf{u}}_{n,t}, \tag{1}$$

where $\tilde{\mathbf{u}}_{n,t}$ is the homogeneous coordinate of $\mathbf{u}_{n,t}$. These single-view 3D estimates are often noisy or inconsistent due to occlusions and depth errors. To obtain a stable particle trajectory in the global frame, we fuse the multi-view estimates into a unified velocity field $\mathbf{v}_t$ using RANSAC.

However, simple algebraic fusion does not guarantee physical plausibility. Crucially, we utilize tactile signals to inspect and enforce the physical plausibility of the particle motions during object-gripper interactions. We minimize a physics-informed objective function $\mathcal{L}_{\text{track}}$ that enforces temporal coherence, local rigidity, and tactile consistency. First, the shape loss $\mathcal{L}_{\text{shape}}$ is defined as the bidirectional Chamfer distance between the predicted next-step positions and the target 3DGS point cloud $\mathbf{P}_{t+1}^{\text{target}}$, which ensures geometric alignment with the recovered surface. To prevent conflicts with contact constraints, particles influenced by tactile sensors $\mathcal{S}_{\text{tactile}}$ are excluded from this loss. Second, the local rigidity loss $\mathcal{L}_{\text{local}}$ enforces As-Rigid-As-Possible (ARAP) constraints to preserve the local metric structure of the deformable object:

$$\mathcal{L}_{\text{local}} = \frac{1}{M}\sum_i \sum_{j\in\mathcal{N}(i)} (\|\mathbf{P}_{i,t+1} - \mathbf{P}_{j,t+1}\| - l_{ij}^{\text{rest}})^2, \tag{2}$$

where l_{ij}^{rest} is the pre-computed rest length between connected particles i and j. Third, we apply a Laplacian regularization to encourage spatial smoothness in the velocity field:

$$\mathcal{L}_{\text{lap}} = \frac{1}{M}\sum_i \left\| \mathbf{v}_{i,t} - \frac{1}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)} \mathbf{v}_{j,t} \right\|^2. \tag{3}$$

Finally, the tactile loss $\mathcal{L}_{\text{tactile}}$ incorporates reduced-order physics from tactile sensors:

$$\mathcal{L}_{\text{tactile}} = \frac{1}{|\mathcal{S}_{\text{tactile}}|}\sum_{i\in\mathcal{S}_{\text{tactile}}} \|\mathbf{v}_{i,t} - \mathbf{v}_{\text{sensor}}\|^2, \tag{4}$$

where the contact set $\mathcal{S}_{\text{tactile}}$ is determined by identifying all object particles within a radius r of the activated taxels of the tactile sensors. Note that $\mathcal{L}_{\text{tactile}}$ assumes localized no-slip. Since our tactile sensors only measure normal-axis pressure, tangential slip is unobservable. Slip events are intentionally avoided during data collections. The total tracking objective is minimized through gradient descent:

$$\mathcal{L}_{\text{track}} = \mathcal{L}_{\text{shape}} + \lambda_{\text{local}}\mathcal{L}_{\text{local}} + \lambda_{\text{lap}}\mathcal{L}_{\text{lap}} + \lambda_{\text{tactile}}\mathcal{L}_{\text{tactile}}. \tag{5}$$

This optimization decouples geometry from tracking, ensuring that particles maintain their physical identity even under large deformations and self-occlusions. We refer to the resulting sequence of configurations, generated by integrating these optimized per-frame velocities starting from the initial particle positions, as the warped point cloud.

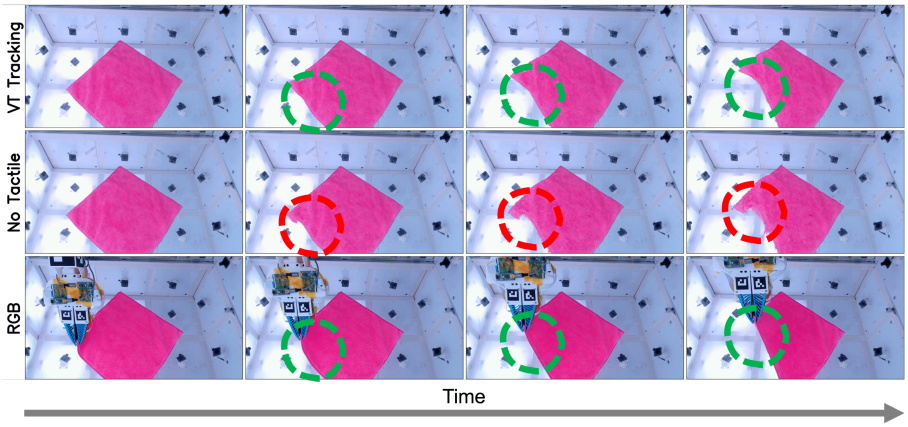

Fig. 4: Benefit of Visuotactile Perception. The three rows compare the warped point clouds – temporally integrated from per-frame tracking – of our full visuotactile approach, tracking without tactile optimization, and the ground truth. Our method yields superior tracking accuracy under occlusions caused by the gripper, confirming the advantage of integrating tactile feedback into the tracking pipeline.

5 Experiments

Our experimental evaluation is designed to demonstrate the utility of Deform360 for learning and benchmarking deformable dynamics across multiple scales and modalities. We first assess the fidelity of our markerless perception pipeline (Sec. 5.1). Next, we investigate the unique visuotactile coupling in our dataset by predicting contacts from visual observations (Sec. 5.2). Then, we conduct a systematic comparison of state-of-the-art world models [27,30,50,82], comparing 3D particle dynamics and action-conditioned video models across three levels of generalization: frame, episode, and object (Sec. 5.3). Finally, we provide a preliminary demonstration of the potential real-world applicability of our dataset by performing robot planning tasks using Model Predictive Control (MPC) on deformable objects (Sec. 5.4). Through these experiments, we aim to provide a comprehensive baseline and reveal critical insights into the future of generalizable deformable world models.

5.1 Evaluation of Reconstruction and Tracking Quality

We evaluate the rendering performance across the 198 objects in the Deform360 dataset (Tab. 2), broken down by object categories (1D, 2D, 3D, and plastic deformables). We utilize standard image quality metrics, including Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS), measured on held-out test views.

Our results demonstrate consistently high reconstruction fidelity across all categories, with 3D volumetric objects achieving the highest PSNR of 30.00 dB. This provides a reliable geometric foundation for the subsequent particle tracking and dynamics modeling tasks. As shown in Fig. 4, the integration of tactile feedback significantly enhances the tracking accuracy, especially for objects with fully occluded parts during prehensile manipulations. Quantitatively, the warped point cloud using visuotactile tracking achieves a Chamfer distance error of $2.71 \times 10^{-5} m^2$, five times lower than the $1.41 \times 10^{-4} m^2$ error of the warped point cloud using only visual tracking. This improvement highlights how the integration of tactile cues provides critical ground-truth signals to maintain tracking accuracy through occlusions.

Table 2: Reconstruction Quality. Performance averaged per category.

Category	PSNR ↑	SSIM ↑	LPIPS ↓
1D Deformables	28.85	0.98	0.0451
2D Deformables	25.77	0.95	0.0919
3D Volumetric Deformables	30.00	0.98	0.0481
Global Average	27.66	0.96	0.0708

5.2 Visuotactile Contact Prediction

A unique feature of the Deform360 dataset is the synchronized tactile modality. In this experiment, we investigate the feasibility of predicting local tactile contact events solely from visual observations. We discretize the tactile data into a binary contact/no-contact signal [39–41,55,79], and train a transformer-based encoder [52, 65] to map the visual stream and robot actions to the expected contact signal. Our model achieves a mean accuracy of 88.67% across 36 views (random guessing 50.31%) and an F1-score of 0.8909 (Fig. 5). The results show that the model reliably captures contact-induced deformations from visual cues, highlighting the dataset's utility in learning the coupling between visual surface changes and internal contact physics.

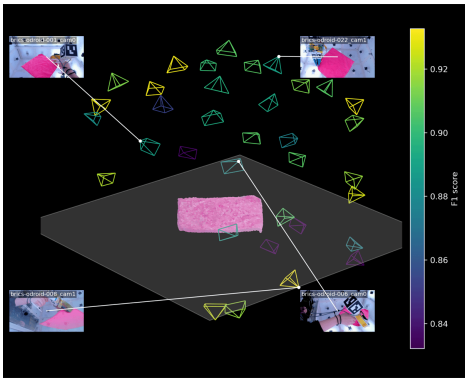

Fig. 5: A visualization of the prediction performance from different camera views.

5.3 World Model Benchmarking

We benchmark current state-of-the-art world models on our dataset to evaluate their ability to capture complex deformable dynamics. Our evaluation spans two primary paradigms: predicting 3D dynamics via particle-based models and predicting 2D dynamics via action-conditioned video models.

3D World Models. We train learning-based 3D world models, such as ParticleFormer [27] and PGND [82], and optimize the differentiable simulation-based PhysTwin [30] using the dense particle trajectories obtained from our

Table 3: Quantitative Results on Per-episode Reconstruction & Resimulation and Future Prediction. We compare the performance of different methods on every episode, divided into reconstruction and resimulation, and future prediction.

Task	Reconstruction & Resimulation					Future Prediction				
Method	CD ↓	Track Error ↓	PSNR ↑	SSIM ↑	LPIPS ↓	CD ↓	Track Error ↓	PSNR ↑	SSIM ↑	LPIPS ↓
PGND	0.032	0.033	26.872	**0.965**	**0.046**	0.073	0.073	25.296	0.963	0.050
ParticleFormer	0.039	0.034	**26.875**	0.964	0.046	0.044	0.041	26.288	**0.964**	**0.047**
PhysTwin	**0.014**	**0.021**	26.783	0.963	0.048	**0.014**	**0.025**	**26.574**	**0.964**	**0.047**

annotation pipeline. We measure the multi-step prediction error using Chamfer distance (CD) and mean squared error (track error) between the ground-truth and the predicted particle positions. We further render the predicted particle trajectories with first-frame 3DGS via linear blend skinning and calculate the image quality metrics against the ground-truth videos [30].

Action-conditioned Video Models. We assess the performance of the *pretrained* generative video model Cosmos-Predict 2.5 2B [50], which has 2 billion parameters. As the pretrained Cosmos model does not natively take robot actions as input, we post-train it on the Deform360 dataset (according to the evaluation setting) to add robot action conditioning (represented by the 6D wrist pose and gripper openness), enabling it to function as an action-conditioned world model. Models are evaluated using the same rendering metrics as 3D dynamics models.

While 3D particle models utilize explicit 3D geometry unlike 2D video models, we evaluate them jointly to highlight the trade-offs between structured priors and implicit generation. This comparison remains practical, as 3D reconstructions are obtainable in real-world robotic deployments [29, 36, 63, 69, 74, 77].

Evaluation settings To rigorously verify the generalizability of the evaluated world models, we adopt three distinct evaluation settings: per-episode, multi-episode, and multi-object.

Per-episode (Frame generalization): Within the same interaction episode, we train the model on the initial T_{train} frames and evaluate its forecasting capability on the subsequent $T - T_{\text{train}}$ frames to verify its frame generalization. In Table 3, the predicted frames seen during training are referred to as "reconstruction" frames, while the predicted frames beyond the training horizon are referred to as "prediction" frames. This setting tests the model's ability to generalize to novel actions given limited dynamic observations.

We observe that the physics-based approach (PhysTwin) outperforms learning-based approaches (PGND and ParticleFormer) in both reconstruction and prediction tasks. This indicates that explicit physical and structural priors remain critical for accurate dynamics modeling in low-data regimes. Note that we do not include Cosmos [50] in this setting as the extremely limited per-episode data is insufficient for the post-training to produce stable predictions.

Multi-episode (Episode generalization): For a given object, we train the model on a subset of E_{train} distinct interaction episodes and test its temporal dynamics prediction on the remaining unseen episodes to verify episode gener-

Table 4: Quantitative Results on Multi-episode Generalization. We compare the performance of different methods on reconstruction and future prediction.

Task	Reconstruction & Resimulation					Future Prediction				
Method	CD ↓	Track Error ↓	PSNR ↑	SSIM ↑	LPIPS ↓	CD ↓	Track Error ↓	PSNR ↑	SSIM ↑	LPIPS ↓
PGND	0.060	0.071	25.513	0.970	0.039	0.130	0.144	23.788	0.971	0.040
ParticleFormer	**0.042**	**0.050**	26.263	**0.971**	0.038	**0.051**	**0.079**	**25.203**	**0.972**	**0.038**
Cosmos	-	-	**27.748**	0.964	**0.034**	-	-	24.950	0.962	0.043

alization. This evaluates whether the model has learned the intrinsic physical properties of the object and can generalize to unseen configurations and interaction sequences. PhysTwin is excluded from this setting as it requires additional per-episode registration to align with different initial configurations, which limits its ability to generalize across episodes without manual intervention.

As seen in Table 4, we observe a distinct trade-off between 2D video models and 3D particle models. Cosmos achieves superior reconstruction performance, as training directly in the 2D space allows it to better preserve high-fidelity visual textures and surface details compared to the particle-to-image rendering pipeline. However, ParticleFormer outperforms in future prediction tasks, which suggests that while the video model excels at visual synthesis, a few interaction episodes are still insufficient for it to fully capture the underlying 3D physical dynamics. In contrast, the explicit 3D structural priors enable more robust generalization of the object's dynamics to novel interaction sequences.

Multi-object (Object generalization):

As visualized in Figure 6, we train the model on a set of O_{train} objects and explicitly test its zero-shot generalization capabilities on the remaining unobserved objects to verify object generalization. This represents the most challenging setting, as it requires the model to infer dynamics and material behaviors for entirely novel object instances based on limited prior knowledge. Since PhysTwin relies on optimization-based fitting of physical parameters to specific object geometries, it cannot natively generalize to novel object categories.

Table 5: Quantitative Results on Multi-Object Generalization. We average train/test scores and report them as future prediction.

Task	Future Pred				
Method	CD ↓	Err ↓	PSNR ↑	SSIM ↑	LPIPS ↓
PGND	0.429	0.320	22.049	0.969	0.041
ParticleFormer	**0.038**	**0.048**	23.312	**0.969**	0.038
Cosmos	-	-	**25.042**	0.958	**0.037**

In this zero-shot setting, Cosmos shows signs of better generalization and outperforms its 3D counterparts in image quality metrics (Table 5). This suggests that by training on a diverse set of objects, the 2D foundation model's large-scale pre-training allows it to effectively leverage the visual diversity in Deform360 to infer the dynamics of novel object categories. However, we observe that a common failure mode for Cosmos is its failure to strictly follow the provided robot commands during long-horizon predictions, rather than producing incorrect object dynamics. Surprisingly, the implicitly generated object dynamics remain physically reasonable most of the time despite this action misalignment. This

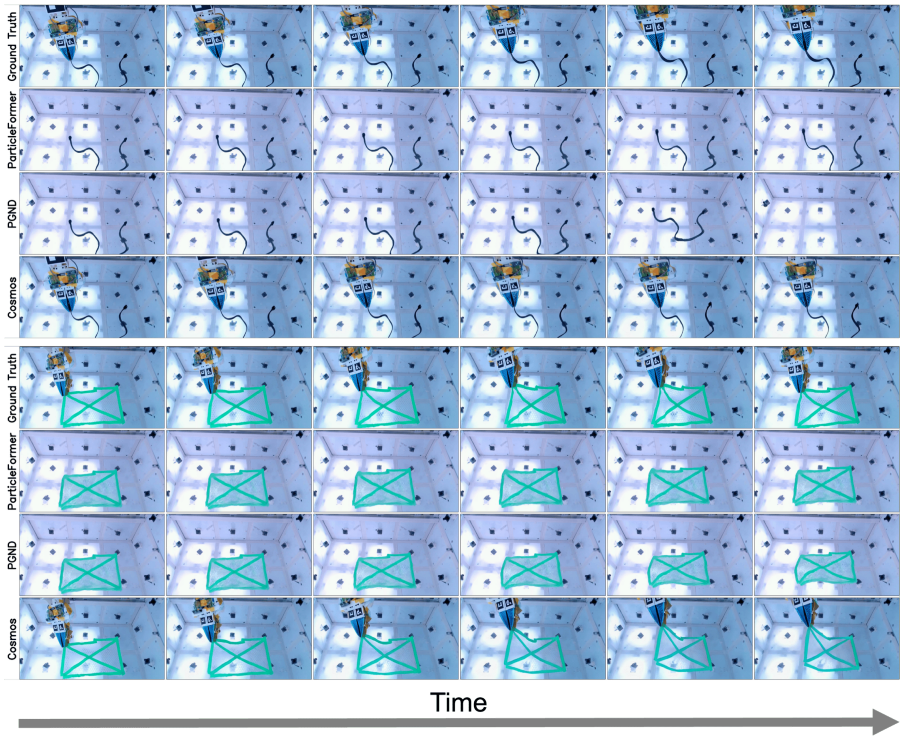

Fig. 6: Visualization of Multi-Object Generalization. We show the predicted future frames for the cable (top) and bubble-wrap (bottom) objects, comparing different world models under the zero-shot object generalization setting.

issue could potentially be resolved with more extensive fine-tuning data or more sophisticated action representations to better align the generative process with high-precision robotic control [60]. Unlike 2D foundation models, current 3D world models lack massive pre-training, which limits their zero-shot generalization capabilities in this setting. While recently PointWorld [28] scale up 3D world models, it had not been open-sourced by the time we finished this work.

5.4 Real-world Robot Planning

To evaluate the real-world applicability of our dataset, we demonstrate that the world models of the objects learned from Deform360 are useful for robotic planning tasks. Crucially, these experiments are deployed using a completely different robot setup (xArm) located in a different lab environment (Fig. 7) in zero-shot. We utilize representations learned from Deform360 via PhysTwin [30] within a MPC framework to plan robot actions for achieving a goal state. We provide preliminary demonstrations indicating the potential real-world applicability, highlighting the utility of Deform360 for building generalizable world models with practical robotic applications.

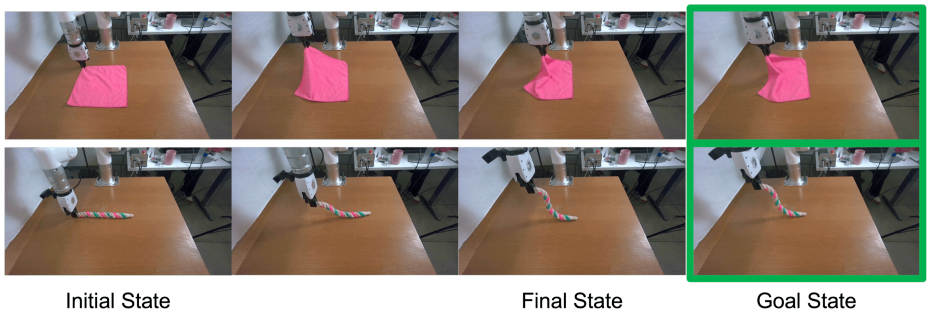

Fig. 7: Real-world Robot Planning with Model Predictive Control. We deploy learned models from Deform360 in an MPC framework to manipulate various deformable objects toward goal state.

We do not deploy Cosmos for real-world planning due to two primary challenges. First, video models are more sensitive to appearance differences across environments, and our current post-training scale is insufficient to support such out-of-distribution visual generalization. Second, designing an effective reward function directly on generated videos remains non-trivial, whereas 3D models can straightforwardly leverage geometric metrics like Chamfer distance [30, 82].

6 Conclusion

We presented Deform360, a large-scale multi-view visuotactile dataset for modeling real-world deformable object dynamics. By developing a novel markerless 3D tracking pipeline, we successfully extracted high-fidelity particle trajectories and dynamic geometry representations across a diverse set of 198 daily-life objects. Leveraging this extensive data, we systematically benchmarked state-of-the-art 3D particle dynamics models against 2D action-conditioned video generation models. Our evaluations revealed a critical trade-off: while 3D particle models exhibited robust physical priors in low-data regimes, video models leveraged their scalability to benefit from massive pre-training, which significantly contributed to their zero-shot visual generalizability. Importantly, we also observed that designing effective reward functions for video models proved non-trivial due to their lack of explicit 3D geometry, creating significant difficulties for their direct application in model-based planning tasks. Despite these challenges, we provide a preliminary demonstration of the potential utility of our dataset by deploying 3D particle models in a real-world Model Predictive Control setup for deformable object manipulation. Ultimately, Deform360 provides a foundational benchmark that we hope will inspire future research toward more generalizable, physically grounded, and scalable world models for robotics.

References

1. Abou-Chakra, J., Sun, L., Rana, K., May, B., Schmeckpeper, K., Minniti, M.V., Herlant, L.: Real-is-Sim: Bridging the Sim-to-Real Gap with a Dynamic Digital Twin for Real-World Robot Policy Evaluation (Apr 2025). https://doi.org/10.48550/arXiv.2504.03597
2. AgiBot-World-Contributors, Bu, Q., Cai, J., Chen, L., Cui, X., Ding, Y., Feng, S., Gao, S., He, X., Hu, X., Huang, X., Jiang, S., Jiang, Y., Jing, C., Li, H., Li, J., Liu, C., Liu, Y., Lu, Y., Luo, J., Luo, P., Mu, Y., Niu, Y., Pan, Y., Pang, J., Qiao, Y., Ren, G., Ruan, C., Shan, J., Shen, Y., Shi, C., Shi, M., Shi, M., Sima, C., Song, J., Wang, H., Wang, W., Wei, D., Xie, C., Xu, G., Yan, J., Yang, C., Yang, L., Yang, S., Yao, M., Zeng, J., Zhang, C., Zhang, Q., Zhao, B., Zhao, C., Zhao, J., Zhu, J.: AgiBot World Colosseo: A Large-scale Manipulation Platform for Scalable and Intelligent Embodied Systems (Apr 2025). https://doi.org/10.48550/arXiv.2503.06669
3. Ai, B., Tian, S., Shi, H., Wang, Y., Pfaff, T., Tan, C., Christensen, H.I., Su, H., Wu, J., Li, Y.: A review of learning-based dynamics models for robotic manipulation. Science Robotics **10**(106), eadt1497 (Sep 2025). https://doi.org/10.1126/scirobotics.adt1497
4. Bansal, H., Lin, Z., Xie, T., Zong, Z., Yarom, M., Bitton, Y., Jiang, C., Sun, Y., Chang, K.W., Grover, A.: VideoPhy: Evaluating Physical Commonsense for Video Generation. In: The Thirteenth International Conference on Learning Representations (Oct 2024)
5. Bear, D., Wang, E., Mrowca, D., Binder, F.J., Tung, H.Y., Pramod, R.T., Holdaway, C., Tao, S., Smith, K.A., Sun, F.Y., Fei-Fei, L., Kanwisher, N., Tenenbaum, J.B., Yamins, D.L., Fan, J.E.: Physion: Evaluating Physical Prediction from Vision in Humans and Machines. In: Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1) (Jun 2021)
6. Blanco-Mulero, D., Barbany, O., Alcan, G., Colomé, A., Torras, C., Kyrki, V.: Benchmarking the Sim-to-Real Gap in Cloth Manipulation. IEEE Robotics and Automation Letters **9**(3), 2981–2988 (Mar 2024). https://doi.org/10.1109/LRA.2024.3360814
7. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., Rombach, R.: Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets (Nov 2023). https://doi.org/10.48550/arXiv.2311.15127
8. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: SAM 3: Segment Anything with Concepts (Nov 2025). https://doi.org/10.48550/arXiv.2511.16719
9. Chen, Y., Zhang, Y., Brei, Z., Zhang, T., Chen, Y., Wu, J., Vasudevan, R.: Differentiable Discrete Elastic Rods for Real-Time Modeling of Deformable Linear Objects. In: 8th Annual Conference on Robot Learning (Sep 2024)
10. Chen, Y., Hu, Y., Sun, L., Kusnur, T., Herlant, L., Jiang, C.: EMPM: Embodied MPM for Modeling and Simulation of Deformable Objects (Jan 2026). https://doi.org/10.48550/arXiv.2601.17251

11. Chi, C., Xu, Z., Pan, C., Cousineau, E., Burchfiel, B., Feng, S., Tedrake, R., Song, S.: Universal Manipulation Interface: In-The-Wild Robot Teaching Without In-The-Wild Robots (Feb 2024). https://doi.org/10.48550/arXiv.2402.10329
12. Coltraro, F., Borràs, J., Alberich-Carramiñana, M., Torras, C.: Tracking cloth deformation: A novel dataset for closing the sim-to-real gap for robotic cloth manipulation learning. The International Journal of Robotics Research **44**(9), 1431–1442 (Aug 2025). https://doi.org/10.1177/02783649251317617
13. Cong, X., Xing, A., Pokhariya, C., Fu, R., Sridhar, S.: Dytact: Capturing dynamic contacts in hand-object manipulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7191–7203 (2025)
14. Cong, X., Yang, H., Chen, L., Zhang, K., Yi, L., Bajaj, C., Huang, Q.: 4drecons: 4d neural implicit deformable objects reconstruction from a single rgb-d camera with geometrical and topological regularizations. arXiv preprint arXiv:2406.10167 (2024)
15. Cong, X., Yang, H., Wang, A., Wang, Y., Yang, Y., Zhang, C., Ma, C.: Viva: Vlm-guided instruction-based video editing with reward optimization. arXiv preprint arXiv:2512.16906 (2025)
16. Du, H., Ye, J., Cong, X., Li, R., Ni, J., Agarwal, A., Zhou, Z., Li, Z., Balestriero, R., Wang, Y.: Videogpa: Distilling geometry priors for 3d-consistent video generation. arXiv preprint arXiv:2601.23286 (2026)
17. Du, Y., Yang, S., Dai, B., Dai, H., Nachum, O., Tenenbaum, J., Schuurmans, D., Abbeel, P.: Learning Universal Policies via Text-Guided Video Generation. In: Advances in Neural Information Processing Systems (Dec 2023)
18. Du, Y., Yang, S., Florence, P., Xia, F., Wahid, A., Ichter, B., Sermanet, P., Yu, T., Abbeel, P., Tenenbaum, J.B., Kaelbling, L.P., Zeng, A., Tompson, J.: Video Language Planning. In: The Twelfth International Conference on Learning Representations (Oct 2023)
19. Duisterhof, B.P., Mandi, Z., Yao, Y., Liu, J.W., Seidenschwarz, J., Shou, M.Z., Ramanan, D., Song, S., Birchfield, S., Wen, B., Ichnowski, J.: DeformGS: Scene Flow in Highly Deformable Scenes for Deformable Object Manipulation (Aug 2024). https://doi.org/10.48550/arXiv.2312.00583
20. Fu, J., Nan, J., Sun, L., Li, H., Qian, J., Barry, J.L., Kitani, K., Konidaris, G.: NovaPlan: Zero-Shot Long-Horizon Manipulation via Closed-Loop Video Language Planning (Feb 2026). https://doi.org/10.48550/arXiv.2602.20119
21. Garcia-Camacho, I., Borràs, J., Calli, B., Norton, A., Alenyà, G.: Household Cloth Object Set: Fostering Benchmarking in Deformable Object Manipulation. IEEE Robotics and Automation Letters **7**(3), 5866–5873 (Jul 2022). https://doi.org/10.1109/LRA.2022.3158428
22. Ha, D., Schmidhuber, J.: World Models (Mar 2018). https://doi.org/10.5281/zenodo.1207631
23. Hafner, D., Lillicrap, T., Ba, J., Norouzi, M.: Dream to Control: Learning Behaviors by Latent Imagination (Mar 2020)
24. Hafner, D., Pasukonis, J., Ba, J., Lillicrap, T.: Mastering Diverse Domains through World Models (Jan 2023). https://doi.org/10.48550/arXiv.2301.04104
25. Hansen, N., Su, H., Wang, X.: TD-MPC2: Scalable, Robust World Models for Continuous Control. In: The Twelfth International Conference on Learning Representations (Oct 2023)
26. Huang, S., Wu, J., Zhou, Q., Miao, S., Long, M.: Vid2World: Crafting Video Diffusion Models to Interactive World Models (May 2025). https://doi.org/10.48550/arXiv.2505.14357

27. Huang, S., Chen, Q., Zhang, X., Sun, J., Schwager, M.: ParticleFormer: A 3D Point Cloud World Model for Multi-Object, Multi-Material Robotic Manipulation. In: Conference on Robot Learning (Jul 2025). https://doi.org/10.48550/arXiv.2506.23126
28. Huang, W., Chao, Y.W., Mousavian, A., Liu, M.Y., Fox, D., Mo, K., Fei-Fei, L.: PointWorld: Scaling 3D World Models for In-The-Wild Robotic Manipulation (Jan 2026). https://doi.org/10.48550/arXiv.2601.03782
29. Hunyuan3D, T., Zhang, B., Guo, C., Liu, H., Yan, H., Shi, H., Huang, J., Yu, J., Li, K., Linus, Wang, P., Lin, Q., Liu, S., Yang, X., Tang, Y., Zhao, Y., Lai, Z., Liang, Z., Zhao, Z.: Hunyuan3D-Omni: A Unified Framework for Controllable Generation of 3D Assets (Sep 2025). https://doi.org/10.48550/arXiv.2509.21245
30. Jiang, H., Hsu, H.Y., Zhang, K., Yu, H.N., Wang, S., Li, Y.: PhysTwin: Physics-Informed Reconstruction and Simulation of Deformable Objects from Videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7219–7230 (2025)
31. Jin, S., Wang, C., Tomizuka, M.: Robust Deformation Model Approximation for Robotic Cable Manipulation. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 6586–6593 (Nov 2019). https://doi.org/10.1109/IROS40897.2019.8968157
32. Kang, B., Yue, Y., Lu, R., Lin, Z., Zhao, Y., Wang, K., Huang, G., Feng, J.: How Far is Video Generation from World Model: A Physical Law Perspective (Nov 2024). https://doi.org/10.48550/arXiv.2411.02385
33. Karaev, N., Makarov, I., Wang, J., Neverova, N., Vedaldi, A., Rupprecht, C.: CoTracker3: Simpler and Better Point Tracking by Pseudo-Labelling Real Videos (Oct 2024). https://doi.org/10.48550/arXiv.2410.11831
34. Kerbl, B., Kopanas, G., Leimkuehler, T., Drettakis, G.: 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Trans. Graph. **42**(4), 139:1–139:14 (Jul 2023). https://doi.org/10.1145/3592433
35. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., Wu, K., Lin, Q., Yuan, J., Long, Y., Wang, A., Wang, A., Li, C., Huang, D., Yang, F., Tan, H., Wang, H., Song, J., Bai, J., Wu, J., Xue, J., Wang, J., Wang, K., Liu, M., Li, P., Li, S., Wang, W., Yu, W., Deng, X., Li, Y., Chen, Y., Cui, Y., Peng, Y., Yu, Z., He, Z., Xu, Z., Zhou, Z., Xu, Z., Tao, Y., Lu, Q., Liu, S., Zhou, D., Wang, H., Yang, Y., Wang, D., Liu, Y., Jiang, J., Zhong, C.: HunyuanVideo: A Systematic Framework For Large Video Generative Models (Mar 2025). https://doi.org/10.48550/arXiv.2412.03603
36. Lai, Z., Zhao, Y., Liu, H., Zhao, Z., Lin, Q., Shi, H., Yang, X., Yang, M., Yang, S., Feng, Y., Zhang, S., Huang, X., Luo, D., Yang, F., Yang, F., Wang, L., Liu, S., Tang, Y., Cai, Y., He, Z., Liu, T., Liu, Y., Jiang, J., Linus, Huang, J., Guo, C.: Hunyuan3D 2.5: Towards High-Fidelity 3D Assets Generation with Ultimate Details (Jun 2025). https://doi.org/10.48550/arXiv.2506.16504
37. LeCun, Y.: A Path Towards Autonomous Machine Intelligence. https://openreview.net/forum?id=BZ5a1r-kVsf (Jun 2022)
38. Li, D., Fang, Y., Chen, Y., Yang, S., Cao, S., Wong, J., Luo, M., Wang, X., Yin, H., Gonzalez, J.E., Stoica, I., Han, S., Lu, Y.: WorldModelBench: Judging Video Generation Models As World Models (Feb 2025). https://doi.org/10.48550/arXiv.2502.20694
39. Li, H., Akl, J., Sridhar, S., Brady, T., Padir, T.: ViTa-Zero: Zero-shot Visuotactile Object 6D Pose Estimation. In: 2025 International Conference on Robotics and Automation (ICRA) (Apr 2025)

40. Li, H., Dikhale, S., Cui, J., Iba, S., Jamali, N.: HyperTaxel: Hyper-Resolution for Taxel-Based Tactile Signals Through Contrastive Learning. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7499–7506 (Oct 2024). https://doi.org/10.1109/IROS58592.2024.10802001
41. Li, H., Jia, M., Akbulut, T., Xiang, Y., Konidaris, G., Sridhar, S.: V-HOP: Visuo-Haptic 6D Object Pose Tracking. In: Robotics: Science and Systems (RSS) XXI (Feb 2025)
42. Li, H., Sun, L., Hu, Y., Ta, D., Barry, J., Konidaris, G., Fu, J.: NovaFlow: Zero-Shot Manipulation via Actionable Flow from Generated Videos (Oct 2025). https://doi.org/10.48550/arXiv.2510.08568
43. Li, X., Guo, Y., Tu, Y., Ji, Y., Liu, Y., Ye, J., Zheng, C.: Textureless Deformable Object Tracking With Invisible Markers. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(9), 7243–7254 (Sep 2025). https://doi.org/10.1109/TPAMI.2024.3463422
44. Li, Y., Du, T., Wu, K., Xu, J., Matusik, W.: DiffCloth: Differentiable Cloth Simulation with Dry Frictional Contact. ACM Trans. Graph. **42**(1), 2:1–2:20 (Oct 2022). https://doi.org/10.1145/3527660
45. Li, Y., Wu, J., Tedrake, R., Tenenbaum, J.B., Torralba, A.: Learning Particle Dynamics for Manipulating Rigid Bodies, Deformable Objects, and Fluids (Apr 2019). https://doi.org/10.48550/arXiv.1810.01566
46. Li, Y., Wu, J., Zhu, J.Y., Tenenbaum, J.B., Torralba, A., Tedrake, R.: Propagation Networks for Model-Based Control Under Partial Observation. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 1205–1211 (May 2019). https://doi.org/10.1109/ICRA.2019.8793509
47. Liang, L., Bian, L., Xiao, C., Zhang, J., Chen, L., Liu, I., Xiang, F., Huang, Z., Su, H.: Robo360: A 3D Omnispective Multi-Material Robotic Manipulation Dataset (Dec 2023). https://doi.org/10.48550/arXiv.2312.06686
48. Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3D Gaussians: Tracking by Persistent Dynamic View Synthesis (Aug 2023). https://doi.org/10.48550/arXiv.2308.09713
49. Newbury, R., Collins, J., He, K., Pan, J., Posner, I., Howard, D., Cosgun, A.: A Review of Differentiable Simulators. IEEE Access **12**, 97581–97604 (2024). https://doi.org/10.1109/ACCESS.2024.3425448
50. NVIDIA, Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., Dworakowski, D., Fan, J., Fenzi, M., Ferroni, F., Fidler, S., Fox, D., Ge, S., Ge, Y., Gu, J., Gururani, S., He, E., Huang, J., Huffman, J., Jannaty, P., Jin, J., Kim, S.W., Klár, G., Lam, G., Lan, S., Leal-Taixe, L., Li, A., Li, Z., Lin, C.H., Lin, T.Y., Ling, H., Liu, M.Y., Liu, X., Luo, A., Ma, Q., Mao, H., Mo, K., Mousavian, A., Nah, S., Niverty, S., Page, D., Paschalidou, D., Patel, Z., Pavao, L., Ramezanali, M., Reda, F., Ren, X., Sabavat, V.R.N., Schmerling, E., Shi, S., Stefaniak, B., Tang, S., Tchapmi, L., Tredak, P., Tseng, W.C., Varghese, J., Wang, H., Wang, H., Wang, H., Wang, T.C., Wei, F., Wei, X., Wu, J.Z., Xu, J., Yang, W., Yen-Chen, L., Zeng, X., Zeng, Y., Zhang, J., Zhang, Q., Zhang, Y., Zhao, Q., Zolkowski, A.: Cosmos World Foundation Model Platform for Physical AI (Jul 2025). https://doi.org/10.48550/arXiv.2501.03575
51. Obrist, J., Zamora, M., Zheng, H., Hinchet, R., Ozdemir, F., Zarate, J., Katzschmann, R.K., Coros, S.: PokeFlex: A Real-World Dataset of Volumetric Deformable Objects for Robotics. IEEE Robotics and Automation Letters **10**(10), 10586–10593 (Oct 2025). https://doi.org/10.1109/LRA.2025.3600148

52. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision. Transactions on Machine Learning Research (Jul 2023)
53. Patel, S., Mohan, S., Mai, H., Jain, U., Lazebnik, S., Li, Y.: Robotic Manipulation by Imitating Generated Videos Without Physical Demonstrations (Jul 2025). https://doi.org/10.48550/arXiv.2507.00990
54. Pokhariya, C., Shah, I.N., Xing, A., Li, Z., Chen, K., Sharma, A., Sridhar, S.: MANUS: Markerless Grasp Capture using Articulated 3D Gaussians. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2197–2208 (2024)
55. Qin, Y., Huang, B., Yin, Z.H., Su, H., Wang, X.: DexPoint: Generalizable Point Cloud Reinforcement Learning for Sim-to-Real Dexterous Manipulation. In: Proceedings of The 6th Conference on Robot Learning. pp. 594–605. PMLR (Mar 2023)
56. Sanchez-Gonzalez, A., Godwin, J., Pfaff, T., Ying, R., Leskovec, J., Battaglia, P.: Learning to Simulate Complex Physics with Graph Networks. In: Proceedings of the 37th International Conference on Machine Learning. pp. 8459–8468. PMLR (Nov 2020)
57. Sutton, R.S.: Dyna, an integrated architecture for learning, planning, and reacting. SIGART Bull. **2**(4), 160–163 (Jul 1991). https://doi.org/10.1145/122344.122377
58. Tang, T., Tomizuka, M.: Track deformable objects from point clouds with structure preserved registration. The International Journal of Robotics Research **41**(6), 599–614 (May 2022). https://doi.org/10.1177/0278364919841431
59. Team, G.R., Abdolmaleki, A., Abeyruwan, S., Ainslie, J., Alayrac, J.B., Arenas, M.G., Balakrishna, A., Batchelor, N., Bewley, A., Bingham, J., Bloesch, M., Bousmalis, K., Brakel, P., Brohan, A., Buschmann, T., Byravan, A., Cabi, S., Caluwaerts, K., Casarini, F., Chan, C., Chang, O., Chappellet-Volpini, L., Chen, J.E., Chen, X., Chiang, H.T.L., Choromanski, K., Collister, A., D'Ambrosio, D.B., Dasari, S., Davchev, T., Dave, M.K., Devin, C., Di Palo, N., Ding, T., Doersch, C., Dostmohamed, A., Du, Y., Dwibedi, D., Egambaram, S.T., Elabd, M., Erez, T., Fang, X., Fantacci, C., Fong, C., Frey, E., Fu, C., Gao, R., Giustina, M., Gopalakrishnan, K., Graesser, L., Groth, O., Gupta, A., Hafner, R., Hansen, S., Hasenclever, L., Haves, S., Heess, N., Hernaez, B., Hofer, A., Hsu, J., Huang, L., Huang, S.H., Iscen, A., Jacob, M.G., Jain, D., Jesmonth, S., Jindal, A., Julian, R., Kalashnikov, D., Karagozler, M.E., Karp, S., Kecman, M., Kew, J.C., Kim, D., Kim, F., Kim, J., Kipf, T., Kirmani, S., Konyushkova, K., Ku, L.Y., Kuang, Y., Lampe, T., Laurens, A., Le, T.A., Leal, I., Lee, A.X., Lee, T.W.E., Lever, G., Liang, J., Lin, L.H., Liu, F., Long, S., Lu, C., Maddineni, S., Majumdar, A., Maninis, K.K., Marmon, A., Martinez, S., Michaely, A.H., Milonopoulos, N., Moore, J., Moreno, R., Neunert, M., Nori, F., Ortiz, J., Oslund, K., Parada, C., Parisotto, E., Paryag, A., Pooley, A., Power, T., Quaglino, A., Qureshi, H., Raju, R.V., Ran, H., Rao, D., Rao, K., Reid, I., Rendleman, D., Reymann, K., Rivas, M., Romano, F., Rubanova, Y., Sampedro, P.P., Sanketi, P.R., Shah, D., Sharma, M., Shea, K., Shridhar, M., Shu, C., Sindhwani, V., Singh, S., Soricut, R., Sterneck, R., Storz, I., Surdulescu, R., Tan, J., Tompson, J., Tunyasuvunakool, S., Varley, J., Vesom, G., Vezzani, G., Villalonga, M.B., Vinyals, O., Wagner, R., Wahid, A., Welker, S., Wohlhart, P.,

Wu, C., Wulfmeier, M., Xia, F., Xiao, T., Xie, A., Xie, J., Xu, P., Xu, S., Xu, Y., Xu, Z., Yan, J., Yang, S., Yang, S., Yang, Y., Yu, H.H., Yu, W., Yuan, W., Yuan, Y., Zhang, J., Zhang, T., Zhang, Z., Zhou, A., Zhou, G., Zhou, Y.: Gemini Robotics 1.5: Pushing the Frontier of Generalist Robots with Advanced Embodied Reasoning, Thinking, and Motion Transfer (Oct 2025)

60. Team, G.R., Choromanski, K., Devin, C., Du, Y., Dwibedi, D., Gao, R., Jindal, A., Kipf, T., Kirmani, S., Leal, I., Liu, F., Majumdar, A., Marmon, A., Parada, C., Rubanova, Y., Shah, D., Sindhwani, V., Tan, J., Xia, F., Xiao, T., Yang, S., Yu, W., Zhou, A.: Evaluating Gemini Robotics Policies in a Veo World Simulator (Jan 2026). https://doi.org/10.48550/arXiv.2512.10675
61. Team, M.L., Cai, X., Huang, Q., Kang, Z., Li, H., Liang, S., Ma, L., Ren, S., Wei, X., Xie, R., Zhang, T.: LongCat-Video Technical Report (Oct 2025). https://doi.org/10.48550/arXiv.2510.22200
62. Team, P.A.N., Xiang, J., Gu, Y., Liu, Z., Feng, Z., Gao, Q., Hu, Y., Huang, B., Liu, G., Yang, Y., Zhou, K., Abrahamyan, D., Ahmad, A., Bannur, G., Chen, J., Chen, K., Deng, M., Han, R., Huang, X., Kang, H., Liu, Z., Ma, E., Ren, H., Shinde, Y., Shingre, R., Tanikella, R., Tao, K., Yang, D., Yu, X., Zeng, C., Zhou, B., Liu, Z., Hu, Z., Xing, E.P.: PAN: A World Model for General, Interactable, and Long-Horizon World Simulation (Nov 2025). https://doi.org/10.48550/arXiv.2511.09057
63. Team, S.D., Chen, X., Chu, F.J., Gleize, P., Liang, K.J., Sax, A., Tang, H., Wang, W., Guo, M., Hardin, T., Li, X., Lin, A., Liu, J., Ma, Z., Sagar, A., Song, B., Wang, X., Yang, J., Zhang, B., Dollár, P., Gkioxari, G., Feiszli, M., Malik, J.: SAM 3D: 3Dfy Anything in Images (Nov 2025). https://doi.org/10.48550/arXiv.2511.16624
64. Tung, H.Y., Ding, M., Chen, Z., Bear, D., Gan, C., Tenenbaum, J., Yamins, D., Fan, J., Smith, K.: Physion++: Evaluating Physical Scene Understanding that Requires Online Inference of Different Physical Properties. Advances in Neural Information Processing Systems **36**, 67048–67068 (Dec 2023)
65. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is All you Need. In: Advances in Neural Information Processing Systems. vol. 30 (2017)
66. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and Advanced Large-Scale Video Generative Models (Apr 2025). https://doi.org/10.48550/arXiv.2503.20314
67. Wang, C., Zhang, Y., Zhang, X., Wu, Z., Zhu, X., Jin, S., Tang, T., Tomizuka, M.: Offline-Online Learning of Deformation Model for Cable Manipulation With Graph Neural Networks. IEEE Robotics and Automation Letters **7**(2), 5544–5551 (Apr 2022). https://doi.org/10.1109/LRA.2022.3158376
68. Wang, Y., Syed, R., Wu, F., Zhang, M., Onol, A., Barreiros, J., Nayyeri, H., Dear, T., Zhang, H., Li, Y.: Interactive World Simulator for Robot Policy Training and Evaluation (Mar 2026). https://doi.org/10.48550/arXiv.2603.08546

69. Wen, B., Tremblay, J., Blukis, V., Tyree, S., Muller, T., Evans, A., Fox, D., Kautz, J., Birchfield, S.: BundleSDF: Neural 6-DoF Tracking and 3D Reconstruction of Unknown Objects (Mar 2023). https://doi.org/10.48550/arXiv.2303.14158
70. Wiedemer, T., Li, Y., Vicol, P., Gu, S.S., Matarese, N., Swersky, K., Kim, B., Jaini, P., Geirhos, R.: Video models are zero-shot learners and reasoners (Sep 2025). https://doi.org/10.48550/arXiv.2509.20328
71. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20310–20320 (2024)
72. Wu, P., Escontrela, A., Hafner, D., Goldberg, K., Abbeel, P.: DayDreamer: World Models for Physical Robot Learning (Jun 2022). https://doi.org/10.48550/arXiv.2206.14176
73. Xian, Z., Zhu, B., Xu, Z., Tung, H.Y., Torralba, A., Fragkiadaki, K., Gan, C.: FluidLab: A Differentiable Environment for Benchmarking Complex Fluid Manipulation. In: The Eleventh International Conference on Learning Representations (Sep 2022)
74. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3D Latents for Scalable and Versatile 3D Generation (Dec 2024). https://doi.org/10.48550/arXiv.2412.01506
75. Xie, T., Zong, Z., Qiu, Y., Li, X., Feng, Y., Yang, Y., Jiang, C.: PhysGaussian: Physics-Integrated 3D Gaussians for Generative Dynamics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4389–4398 (2024)
76. Xie, W., Yu, Z., Zhao, Z., Zuo, B., Wang, Y.: HMDO : Markerless multi-view hand manipulation capture with deformable objects. Graphical Models **127**, 101178 (May 2023). https://doi.org/10.1016/j.gmod.2023.101178
77. Xie, X., Wen, B., Chang, Y., Rabeti, H., Li, J., Yuan, Y., Pons-Moll, G., Birchfield, S.: CARI4D: Category Agnostic 4D Reconstruction of Human-Object Interaction (Jan 2026). https://doi.org/10.48550/arXiv.2512.11988
78. Ye, V., Li, R., Kerr, J., Turkulainen, M., Yi, B., Pan, Z., Seiskari, O., Ye, J., Hu, J., Tancik, M., Kanazawa, A.: Gsplat: An Open-Source Library for Gaussian Splatting. Journal of Machine Learning Research **26**(34), 1–17 (2025)
79. Yin, Z.H., Huang, B., Qin, Y., Chen, Q., Wang, X.: Rotating without Seeing: Towards In-hand Dexterity through Touch. In: Robotics: Science and Systems XIX. vol. 19 (Jul 2023)
80. Yu, M., Lv, K., Wang, C., Jiang, Y., Tomizuka, M., Li, X.: Generalizable whole-body global manipulation of deformable linear objects by dual-arm robot in 3-d constrained environments. The International Journal of Robotics Research **44**(4), 607–639 (2025)
81. Zhang, K., Li, B., Hauser, K., Li, Y.: AdaptiGraph: Material-Adaptive Graph-Based Neural Dynamics for Robotic Manipulation. In: Robotics: Science and Systems. vol. 20 (Jul 2024)
82. Zhang, K., Li, B., Hauser, K., Li, Y.: Particle-Grid Neural Dynamics for Learning Deformable Object Models from RGB-D Videos. In: Robotics: Science and Systems (Jun 2025). https://doi.org/10.48550/arXiv.2506.15680
83. Zhang, M., Zhang, K., Li, Y.: Dynamic 3D Gaussian Tracking for Graph-Based Neural Dynamics Modeling. In: 8th Annual Conference on Robot Learning (Sep 2024)

84. Zhou, B., Zhou, H., Liang, T., Yu, Q., Zhao, S., Zeng, Y., Lv, J., Luo, S., Wang, Q., Yu, X., Chen, H., Lu, C., Shao, L.: ClothesNet: An Information-Rich 3D Garment Model Repository with Simulated Clothes Environment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20428–20438 (2023)
85. Zhou, G., Pan, H., LeCun, Y., Pinto, L.: DINO-WM: World Models on Pre-trained Visual Features enable Zero-shot Planning. In: International Conference on Machine Learning (Jun 2025)
86. Zhu, X., Huang, B., Li, Y.: Touch in the Wild: Learning Fine-Grained Manipulation with a Portable Visuo-Tactile Gripper (Nov 2025). https://doi.org/10.48550/arXiv.2507.15062