PhysHO: Physics-Based Dynamic 3D Gaussian Human and Object from Monocular Video

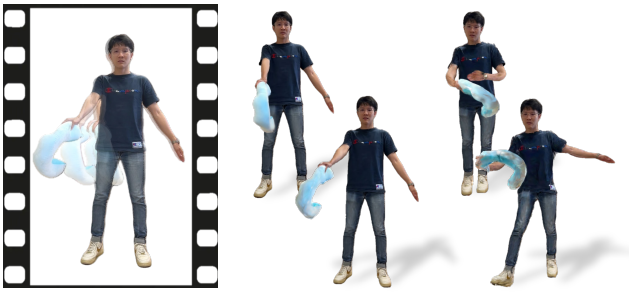
Suyi Jiang Gim Hee Lee
Department of Computer Science, National University of Singapore
jiang_suyi@u.nus.edu, gimhee.lee@nus.edu.sg

Abstract

Physically plausible reconstruction of human–object dynamics from a single video remains under-explored in physics-based methods. Most prior approaches omit human-generated internal actuation by assuming motion driven solely by gravity and simple contacts. They also rely on idealized constitutive laws that underfit heterogeneous and anisotropic materials. We introduce ***PhysHO****, which tightly couples SMPL-driven Linear Blend Skinning (LBS) with a Material Point Method (MPM) simulator to address these gaps. Our key insight is to use LBS as an interpretable actuation prior and MPM to propagate those forces through contact under physical constraints. Concretely, we derive targeted actuation with a PD controller guided by LBS trajectories and gate it per particle via a learnable LBS-impact factor so that only particles inside the SMPL volume are directly actuated. We model real materials with residual neural constitutive laws layered on expert elastic–plastic models and conditioned on per particle to capture heterogeneity and anisotropy. We stabilize monocular learning with structure-preserving 3D flow supervision and a progressive and loss-balanced training schedule. Our PhysHOreconstructs observed dynamics with high fidelity, and predicts future motion and simulates outcomes under novel human actions. Experimental results demonstrate robust human-driven interactions beyond gravity-only scenes. Project:* https://suezjiang.github.io/physho/.

1. Introduction

Reconstructing physically plausible human–object dynamics from video is increasingly important for VR/AR content creation, digital humans, robotics simulation, *etc*. Recent progress in physics-based reconstruction that couples differentiable rendering with differentiable simulation has made notable strides, especially for dynamic object reconstruction from multi-view inputs. However, most prior systems assume simple scenes where motion is driven only by gravity and ground-plane contacts, and they struggle in common real-world settings where ***external actuation arises from human–object interaction***. Physics-based reconstruction of human-driven scenes is challenging for two reasons. First, interaction dynamics are driven not only by gravity but also by internal forces generated by human motion, which must be identified and modeled from observations. Second, real materials rarely match ideal homogeneous and isotropic assumptions. In reality, heterogeneity and anisotropy are ubiquitous and difficult to represent and infer accurately.



Monocular Human-Object Video | Physics-Based Reconstruction | Novel Pose Simulation

Figure 1. We propose PhysHO, a framework that reconstructs physically grounded results from monocular human-object videos. Our method can further simulate predictions under novel motions.

Dynamic 3D Gaussian Splatting has achieved impressive visual quality. Some methods extend Gaussians into 4D volumes [10, 63, 69, 70]. Others model time-varying flows [12, 34], and template/motion-basis approaches encode articulation as linear transforms [6, 17, 42, 45, 62, 68]. GART [30] combines SMPL [37] with latent bones and learned skinning to reconstruct dynamic human scenes. Despite high-fidelity renderings, these methods primarily overfit observed frames and are not governed by physical laws, which limit their ability to do extrapolation to novel actions and prediction. Conversely, physics-based approaches couple differentiable renderers such as Neural Radiance Field (NeRF) and 3D Gaussian Splatting (3DGS) [26, 41] with differentiable simulators such as Material Point Method

(MPM) [21, 23, 55] to infer materials and recover object dynamics from videos [3, 31, 64, 74]. Despite the impressive results, they typically omit internal actuation generated by humans and often rely on idealized constitutive models that underfit heterogeneous and anisotropic behavior. Although PhysRig [71] injects actuation via learned velocity controls, it requires synthetic 3D supervision and therefore limiting transfer to real videos.

We propose **PhysHO**, a monocular framework that reconstructs physically plausible human–object dynamics by tightly coupling SMPL-driven Linear Blend Skinning (LBS) with an MPM simulator. Our central insight is to use LBS as an actuation prior that explains where and how the human injects internal forces, and to use MPM as the physics engine that propagates those forces through contact to objects under conservation laws and material behavior. This yields three key novelties: 1) ***A targeted actuation mechanism*** achieved with a PD controller driven by LBS trajectories and modulated by a learnable per-particle LBS-impact factor that injects the minimum forces necessary only inside the SMPL volume, preventing spurious object actuation; 2) ***Neural residual constitutive laws*** layered on expert elastic and plastic models conditioned by per-particle features, which capture heterogeneous and anisotropic materials while retaining stability; 3) ***A structure-preserving 3D flow*** supervision strategy with RGB, optical flow and as-rigid-as-possible (ARAP) regularization that anchors the simulator when learning from a single view to enable generalization beyond overfitting to observed frames.

Concretely, we first build a mass-preserving fixed-count canonical 3D Gaussian reconstruction using LBS with the original SMPL bones and fixed skinning. We then fine-tune the Gaussian parameters to maintain rendering fidelity under physics-driven deformations using the deformation gradients. We treat LBS trajectories as reference motion, inject PD-controller forces, and gate them per particle via the LBS-impact factor such that object and exterior particles move only through physical interaction. For materials, we optimize spatially varying Young's modulus and Poisson's ratio and learn residual neural corrections conditioned on per-particle latents. To stabilize optimization, we first refine per-frame particle positions to obtain 3D flow supervision, then train progressively from early to later frames with loss-balanced scheduling that allocates more iterations to harder frames. Our design choices deliver temporally consistent, physically grounded reconstructions, robust human-driven interactions, and accurate predictions under novel motions.

Our ***main contributions*** are summarized as follows:

- A human-object reconstruction framework that combines LBS with MPM. We use LBS to localize human internal actuation and MPM to propagate forces to objects under physical constraints. This combination enables simulation and prediction from a single video.
- A PD-controller actuation model with a learnable LBS-impact factor that applies minimal spatially targeted forces only within the SMPL volume to prevent spurious actuation on objects and improving interaction fidelity.
- Neural residual constitutive laws built on expert models with spatially varying Young's modulus and Poisson's ratio and per-particle latent conditioning to represent heterogeneous and anisotropic materials.
- A monocular training pipeline that leverages a structure-preserving 3D flow prior (RGB + optical flow + ARAP) and progressive loss-balanced scheduling for high-quality reconstruction, future-motion prediction, and realistic outcomes under novel human actions.

2. Related Work

Dynamic Reconstruction. NeRF [41] and 3D Gaussian Splatting [26] can be extended to dynamic scenarios by learning a time-conditioned deformation field [1, 2, 15, 25, 63, 69, 70]. Some [6, 45, 46, 58, 60, 67, 72] learn motion bases to decompose and represent dynamic motion. Some [12, 18, 33] directly optimize the scene flow from dynamic observation. These methods typically rely on overfitting a time-conditioned function to deform spatial points without modeling the underlying structure or physical properties. Thus they are not physically grounded and cannot generalize to simulate future dynamics.

Human Reconstruction. Many existing methods leverage the parametric model SMPL [37] to assist dynamic human reconstruction from monocular or multi-view videos. These methods learn pose-dependent warping functions [4, 20, 27, 42, 50, 62], learn 3D Gaussian maps [5, 32], deform SMPL mesh [47], or combine with latent motion bases [30, 73] to represent temporal non-rigid deformations. Although these methods can reconstruct visually plausible human motions, their representations of network-based deformation or kinematic transformation are not physically grounded, resulting in reconstructions that lack physical realism.

Physics-Based Dynamic Reconstruction. Previous physics-based reconstruction works primarily focus on object-centric scenarios. For forward simulation, some methods [14, 24, 64] integrate 3DGS [26] with various simulation methods [21, 23, 40, 43, 55] to allow simulation of reconstruction. For inverse reasoning of physical properties, various differentiable simulation methods [7, 9, 11, 21, 22, 28, 57, 66] are explored. Some [31, 51] integrates NeRF/3DGS with MPM to learn material properties from videos. Phyrecon [44] leverages physical simulation to aid geometry learning. Some methods [35, 36] distill material property using video generation models. Some methods [3, 39] employ neural networks to improve the expressiveness of physical models. Spring-Gaus [74] incorporates a spring-mass system for dynamic reconstruc-

tion. Physrig [71] learns skeleton-driven dynamics from synthetic 3D data. A recent work, MPMAvatar [29], utilizes mesh-based simulation to reconstruct human clothing from multi-view videos. In terms of physics-based human motion reconstruction, some [8, 13, 38, 48, 49, 56, 61] leverage reinforcement learning to imitate human motions. Some [16, 53] optimize the control torques of the human skeleton. Yet these methods focus on simplified skeleton without geometry or appearance. Our method targets human-object dynamic reconstruction, aiming to achieve physically grounded results under the challenging monocular setting.

3. Preliminary

3.1. Template-Based Human Reconstruction

GART [30] reconstructs a canonical space of human 3D Gaussians denoted as $\mathcal{G} = \{(\mu_c^i, R_c^i, S_c^i, \eta_c^i, h_c^i)\}_{i=1}^{\|\mathcal{G}\|}$ from a monocular video. Each Gaussian comprises the 3D mean $\mu_c^i \in \mathbb{R}^3$, rotation $R_c^i \in \mathrm{SO}(3)$, diagonal scaling matrix $S_c^i \in \mathbb{R}^{3\times3}$, opacity factor $\eta_c^i \in (0, 1]$, and spherical-harmonics coefficients $h_c^i \in \mathbb{R}^{\text{sph}}$. Given the SMPL [37] pose θ, each Gaussian is given by linear blend skinning:

$$\begin{aligned} \mu_{lbs}^i &= A_{rot}^i\, \mu_c^i + A_t^i, \qquad R_{lbs}^i = A_{rot}^i\, R_c^i, \\ A^i(\theta) &= \sum_k W_k^i\, B_k(\theta), \end{aligned} \tag{1}$$

where $B_k(\theta) \in \mathrm{SE}(3)$ is the k-th bone transform and W_k^i is the k-th skinning weight for the i-th Gaussian kernel. In GART, the bone set includes both SMPL skeleton bones and learnable latent bones. The skinning weights to all bones are optimized jointly.

3.2. Material Point Method

The Material Point Method (MPM) [21, 23, 55] is a differentiable solver for continuum materials that couples Lagrangian particles with an Eulerian background grid. It advances the state by solving the momentum equation:

$$\rho \frac{D\mathbf{v}}{Dt} = \nabla \cdot \boldsymbol{\sigma} + \rho \mathbf{g} + \boldsymbol{f}_{ex}, \tag{2}$$

where ρ is density, $\mathbf{v}$ is velocity, and $\nabla \cdot \boldsymbol{\sigma}$ is the internal force induced by the divergence of the Cauchy stress. The terms $\rho\mathbf{g}$ and $\boldsymbol{f}_{ex}$ denote gravity and other external forces, respectively.

As shown in Algorithm 1, each particle carries the state $\boldsymbol{s}_n = \{\boldsymbol{x}_n, \boldsymbol{v}_n, \boldsymbol{C}_n, \boldsymbol{F}_n\}$ at frame n, where $\boldsymbol{x}_n$ is position, $\boldsymbol{v}_n$ is velocity, $\boldsymbol{C}_n$ is an affine velocity matrix, and $\boldsymbol{F}_n$ is deformation gradient. Over T substeps, an MPM integrator $I(\cdot)$ updates this state using stresses from an elastic constitutive law $\mathcal{E}$ and a plastic return-mapping law $\mathcal{P}(\cdot)$. Concretely, the elastic Cauchy stress is $\boldsymbol{\sigma} = \mathcal{E}(\boldsymbol{F}, E, \nu)$ given Young's modulus E and Poisson's ratio ν, and plasticity projects the trial deformation via $\boldsymbol{F} = \mathcal{P}(\boldsymbol{F}^{trial})$.

Algorithm 1: MPM at frame n with T substeps

Input: $\boldsymbol{s}_n = \{\boldsymbol{x}_n, \boldsymbol{v}_n, \boldsymbol{C}_n, \boldsymbol{F}_n\}$
Output: $\boldsymbol{s}_{n+1} = \{\boldsymbol{x}_{n+1}, \boldsymbol{v}_{n+1}, \boldsymbol{C}_{n+1}, \boldsymbol{F}_{n+1}\}$
/* Unpack particle state at frame n */
$\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{C}, \boldsymbol{F} \leftarrow \boldsymbol{s}_n$;
/* Substepping for stability & accuracy */
for $i \leftarrow 1 \dots T$ **do**
  /* Elastic Cauchy stress */
  $\boldsymbol{\sigma} = \mathcal{E}(\boldsymbol{F}, E, \nu)$;
  /* MPM Integrator */
  $\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{C}, \boldsymbol{F}^{trial} = I(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{C}, \boldsymbol{F}, \boldsymbol{\sigma})$;
  /* Plastic return-mapping */
  $\boldsymbol{F} = \mathcal{P}(\boldsymbol{F}^{trial})$;
/* Write back updated particle state */
$\boldsymbol{s}_{n+1} \leftarrow \boldsymbol{x}, \boldsymbol{v}, \boldsymbol{C}, \boldsymbol{F}$;

4. Method

Fig. 2 shows of our proposed **PhysHO**, which reconstructs physically plausible human-object dynamics from a monocular video by tightly coupling SMPL-driven Linear Blend Skinning (LBS) with an MPM simulator. The key insight is to use LBS as an ***actuation prior*** to explain where and how the human injects internal forces, and to use MPM as the ***physics engine*** that transmits those forces to objects through contact while obeying conservation laws and realistic material behavior. The method comprises four components: 1) A mass-preserving canonical 3D Gaussian representation with LBS and physics-aware fine-tuning; 2) LBS-integrated dynamics via a PD controller gated by an LBS-impact factor; 3) Residual neural constitutive laws on top of expert elastic and plastic models; 4) A training pipeline that builds structure-preserving 3D flow supervision and optimizes progressively with loss-balanced scheduling.

4.1. Representation

We represent the scene with a mass-preserving fixed-count set of 3D Gaussians that double as MPM particles. To this end, we learn a canonical LBS-driven space using SMPL bones with fixed skinning and perform physics-aware fine-tuning to adapt Gaussian parameters via deformation gradient–based covariance updates.

Canonical Gaussians and Mass Preservation. Similar to prior physics-based reconstructions, we require pre-reconstructed 3D Gaussians for the human and the object to preserve conservation of mass. The set size is fixed, and these Gaussians serve as simulation particles. Given predicted SMPL poses and N input frames, we follow GART [30] to learn a canonical set:

$$\mathcal{G} = \{(\mu_c^i, R_c^i, S_c^i, \eta_c^i, h_c^i)\}$$

parameterized by the LBS transform in Eqn. 1. However, unlike GART that introduces latent bones and learned skin-

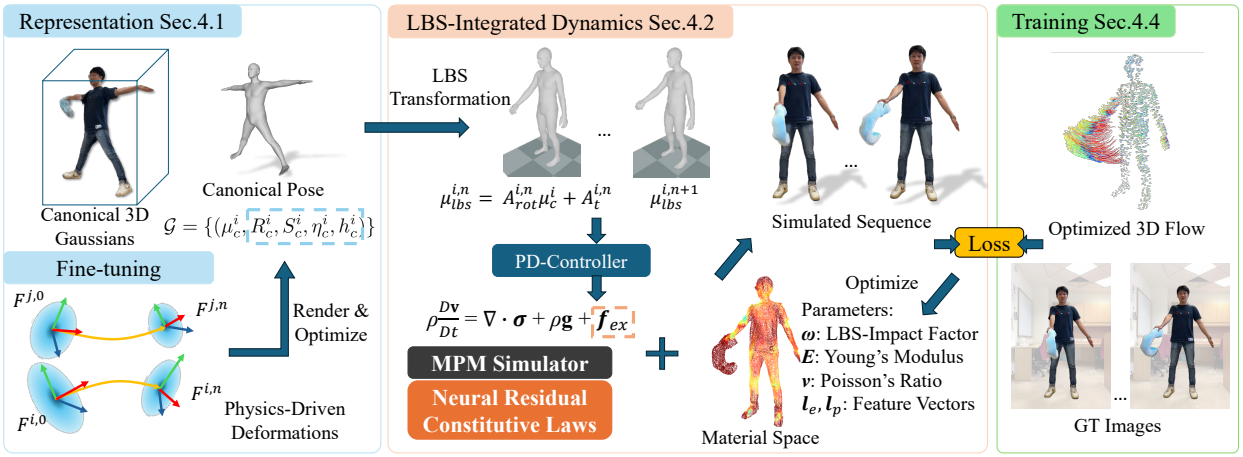


Figure 2. **PhysHO framework.** We couple SMPL-driven LBS with an MPM simulator, where LBS provides a localized actuation prior inside the human and MPM propagates forces through contact to objects under physical constraints. Residual neural constitutive laws model heterogeneous and anisotropic materials. Training uses a structure-preserving 3D flow prior and progressive loss-balanced optimization.

ning weights, we use the original SMPL bones with fixed skinning in Eqn. 1. This matches our setting in which the learning clip for the human canonical space exhibits no significant non-rigid deformation beyond body spin (*cf*. Sec. 5.1). After training, the LBS-transformed Gaussians for pose θ^n at frame n are:

$$\mathcal{G}^n = \{(\mu_{lbs}^{i,n}, R_{lbs}^{i,n}, S_{lbs}^{i,n}, \eta_{lbs}^{i,n}, h_{lbs}^{i,n})\}.$$

Initialization and Physics-Aware Covariance. We initialize the simulation from the first LBS frame by using the Gaussian means as particle positions:

$$\boldsymbol{x}_0 \leftarrow \{\mu_{lbs}^{i,0}\} \quad \text{with } \mathcal{G}^0.$$

Following [65], the Gaussian i at frame n is deformed by the deformation gradient $F^{i,n}$ to give the covariance:

$$\Sigma^{i,n} = F^{i,n} R_{lbs}^{i,0} S_{lbs}^{i,0} (S_{lbs}^{i,0})^\top (R_{lbs}^{i,0})^\top (F^{i,n})^\top. \tag{3}$$

Fine-Tuning under Physics-Driven Deformations. The Gaussian orientation, scaling, opacity, and spherical-harmonic radiance coefficients are initially learned in the LBS-driven canonical space. Directly applying Eqn. 3 across time can degrade rendering because these parameters are not yet adapted to physics-induced deformations. We therefore perform a physics-aware fine-tuning step. Specifically, we set the LBS means as desired particle positions at each frame n of the body spin stage:

$$\boldsymbol{x}_n \leftarrow \{\mu_{lbs}^{i,n}\},$$

and the particle velocities are estimated by central differences:

$$\boldsymbol{v}_n = \frac{\boldsymbol{x}_{n+1} - \boldsymbol{x}_{n-1}}{2 \cdot \Delta t}. \tag{4}$$

Given $(\boldsymbol{x}_n, \boldsymbol{v}_n)$, we compute $F^{i,n}$ using Alg. 1 with the stress fixed to $\boldsymbol{\sigma} = 0$ and plasticity $\mathcal{P}$ set to $\boldsymbol{F} = \boldsymbol{F}^{trial}$.

We then render each frame using Eqn. 3 and optimize the canonical parameters $\{(R_c^i, S_c^i, \eta_c^i, h_c^i)\}$ by the RGB loss between the rendered image I and ground truth I^*:

$$\mathcal{L}_{RGB} = \|I - I^*\|_1. \tag{5}$$

Insight. This fine-tuning step bridges kinematic LBS and physics-driven deformations by adapting the rendering parameters to deformation gradients while keeping the particle set mass-preserving and fixed.

4.2. LBS-Integrated Dynamics

In the dynamic stage of the input video (*cf*. Sec. 5.1), we learn material properties such that simulated dynamics match observations. To model human motion and its internal actuation, we use the LBS position trajectories as a reference and compute auxiliary driving forces with a PD controller. However, not all particles should receive these forces. Internal actuation originates within the human, and object particles should move only under forces transmitted through contact. Moreover, the reference trajectories in the second stage are imperfect, especially for non-rigid regions. We therefore modulate the PD forces with the LBS-impact factor, which is a learnable per-particle coefficient that controls the force magnitude received by each particle.

Proportional-Derivative (PD) Controller. Given reference positions $\{\mu_{lbs}^{i,n}\}$ and reference velocities $\{v_{lbs}^{i,n}\}$ from Eqn. 4, the additional force on particle i is:

$$f_{PD}^{i,n} = k_p(\mu_{lbs}^{i,n} - x^{i,n}) + k_d(v_{lbs}^{i,n} - v^{i,n}), \tag{6}$$

where k_p and k_d are the proportional and derivative gains, and $(x^{i,n}, v^{i,n})$ is the particle state.

LBS-Impact Factor and Constraints. We gate the actuation with a per-particle coefficient ω^i:

$$f_{ex}^{i,n} = \omega^i f_{PD}^{i,n}, \tag{7}$$

which enters the momentum update in Eqn. 2 and the integrator I in Alg. 1. Following MPM, $f_{ex}^{i,n}$ is transferred to grid nodes by particle-to-grid weights and influences accelerations. In the canonical space, particles outside the SMPL template surface, *i.e.* object particles and exterior human surface are always assigned $\omega^i = 0$. This means that these particles do not receive additional forces from the PD controller. Particles strictly inside the SMPL volume are assigned learnable ω^i to control the strength of $f_{ex}^{i,n}$.

Insight. The LBS-impact factor enforces *targeted actuation*, where only the human interior is directly actuated to prevent spurious forces on objects and to improve interaction fidelity.

4.3. Neural Residual Constitutive Laws

Classical MPM relies on expert constitutive laws and often assumes idealized homogeneous and isotropic materials. Although learning spatially varying Young's modulus $\boldsymbol{E}$ and Poisson's ratio $\boldsymbol{\nu}$ improves heterogeneity, it remains insufficient to capture anisotropy and complex spatial variation. We thus introduce residual neural constitutive laws layered on top of expert elastic-plastic models and conditioned by per-particle latents to provide the expressivity needed for heterogeneous and anisotropic behavior while preserving a stable and physically grounded backbone.

Residual learning for heterogeneous and anisotropic behavior. NCLaw [39] demonstrates that neural constitutive models can capture rich anisotropic dynamics when substituted for expert elastic–plastic laws. However, its formulation is effectively homogeneous in space and struggles with scenes where properties vary across materials and locations. Building on this observation, we introduce per-particle conditioning to encode heterogeneity and learn neural terms as *residuals* on top of expert models to retain physical structure and stabilize monocular training. We use physics-aware modules for elasticity $\mathcal{E}_\theta$ and plasticity $\mathcal{P}_\theta$, extended with per-particle features l_e^i and l_p^i:

$$\boldsymbol{\sigma} = \mathcal{E}_\theta(\boldsymbol{F}, \boldsymbol{l}_e), \qquad \boldsymbol{F} = \mathcal{P}_\theta(\boldsymbol{F}^{trial}, \boldsymbol{l}_p).$$

However, it is ill-posed to learn heterogeneous neural constitutive laws solely from per-frame rendering losses, where the unconstrained predictions readily destabilize the simulator and lead to collapse. To address this issue, we cast these neural terms as residuals over expert models $\mathcal{E}(\cdot)$ and $\mathcal{P}(\cdot)$:

$$\begin{aligned} \boldsymbol{\sigma} &= \mathcal{E}(\boldsymbol{F}, \boldsymbol{E}, \boldsymbol{\nu}) + \mathcal{E}_\theta(\boldsymbol{F}, \boldsymbol{l}_e), \\ \boldsymbol{F} &= \mathcal{P}(\boldsymbol{F}^{trial}) + \mathcal{P}_\theta(\boldsymbol{F}^{trial}, \boldsymbol{l}_p). \end{aligned} \tag{8}$$

The expert terms provide a robust elastic–plastic backbone, and the per particle conditioned residuals model spatial heterogeneity and directional anisotropy.

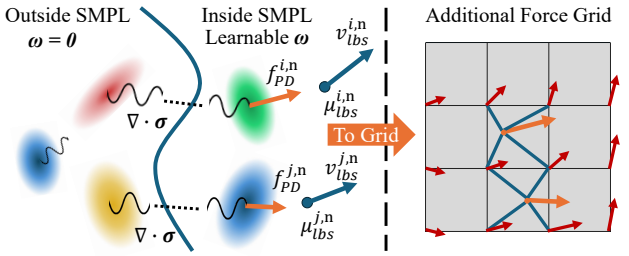


Figure 3. **Targeted actuation.** LBS reference motion drives a PD controller whose forces are modulated per particle by the LBS-impact factor ω^i. Only particles inside the SMPL volume receive direct actuation, while objects move via contact.

Insight. Anchoring to expert elastic–plastic laws yields a well-posed and physically grounded update. The residual neural terms is inspired by NeuMA [3], which optimizes additional LoRA [19] layers $\Delta\mathcal{M}_\theta$ as a residual term to the pretrained NCLaw networks $\mathcal{M}_0$: $\mathcal{M}_\theta := \mathcal{M}_0 + \Delta\mathcal{M}_\theta$. Conditioning these residuals on each particle gives the expressivity needed for heterogeneous and anisotropic responses while maintaining stability and data efficiency.

4.4. Training

Training from a single view is under-constrained, especially with coupled actuation and elastoplastic dynamics. We anchor supervision with a structure-preserving 3D flow estimated per frame from RGB, optical flow, and ARAP. We then optimize the simulator end-to-end to match images and 3D flow, and regularize the LBS-gated actuation and the neural residual constitutive laws. Finally, we adopt a progressive and loss-balanced schedule that fits early frames first and allocates more iterations to harder frames. This design renders monocular optimization well-posed and efficient to yield stable learning of targeted actuation, and heterogeneous and anisotropic material behavior.

Structure-Preserving 3D Flow Supervision. Monocular RGB alone is under-constrained and may drive Gaussians to implausible shapes. We therefore first optimize per-frame particle positions $\boldsymbol{x}_n'$ to obtain a structure-preserving 3D flow. For each frame, we compute deformation gradients from $\boldsymbol{x}_n'$ and render the image. $\boldsymbol{x}_n'$ is then optimized with RGB loss, optical-flow loss, and an as-rigid-as-possible (ARAP) term [54]:

$$\mathcal{L}_{SP-Flow} = \lambda_{rgb}\mathcal{L}_{rgb} + \lambda_{flow}\mathcal{L}_{flow} + \lambda_{arap}\mathcal{L}_{arap}. \tag{9}$$

The optimized 3D flow preserves intrinsic structure and provides 3D supervision for the simulator.

End-to-End Losses. Alg. 2 describes our frame-step simulation. Given state $\boldsymbol{s}_n = \{\boldsymbol{x}_n, \boldsymbol{v}_n, \boldsymbol{C}_n, \boldsymbol{F}_n\}$, we advance to $\boldsymbol{s}_{n+1}$, render for RGB loss, and align positions to the optimized flow:

$$\mathcal{L}_{E2E} = \lambda_{rgb}\mathcal{L}_{rgb} + \lambda_{3Dflow}\|\boldsymbol{x}_{n+1} - \boldsymbol{x}_{n+1}'\|_1. \tag{10}$$

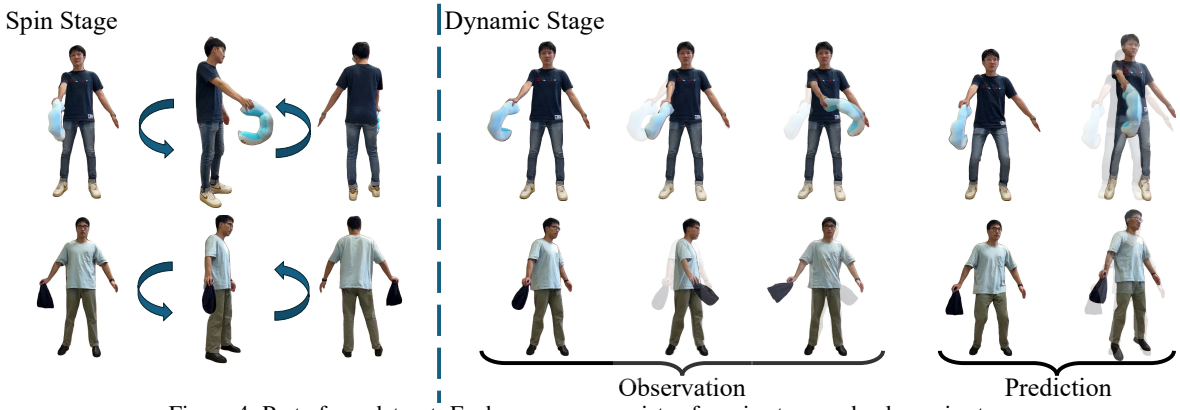


Figure 4. Part of our dataset. Each sequence consists of a spin stage and a dynamic stage.

Algorithm 2: PhysHO Update of Frame $n \rightarrow n+1$

Input:
Particle state at frame n: $\boldsymbol{s}_n = \{\boldsymbol{x}_n, \boldsymbol{v}_n, \boldsymbol{C}_n, \boldsymbol{F}_n\}$;
LBS refs: $\{\mu_{\text{lbs}}^{i,n}, v_{\text{lbs}}^{i,n}\}$;
Params: $\{\boldsymbol{E}, \boldsymbol{\nu}, \boldsymbol{\omega}, \boldsymbol{l}_e, \boldsymbol{l}_p, \theta\}$
Output:
Particle state at frame $n+1$:
$\boldsymbol{s}_{n+1} = \{\boldsymbol{x}_{n+1}, \boldsymbol{v}_{n+1}, \boldsymbol{C}_{n+1}, \boldsymbol{F}_{n+1}\}$

$\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{C}, \boldsymbol{F} \leftarrow \boldsymbol{s}_n$;
/* Targeted actuation from LBS prior */
foreach *particle* i **do**
  /* PD controller */
  $f_{PD}^{i,n} \leftarrow k_p(\mu_{\text{lbs}}^{i,n} - x^{i,n}) + k_d(v_{\text{lbs}}^{i,n} - v^{i,n})$;
  /* LBS-impact gating (Eqn. 7) */
  $f_{ex}^{i,n} \leftarrow \omega^i f_{PD}^{i,n}$; // ω^i=0 outside SMPL
/* Substeps for stability */
for $t \leftarrow 1 \dots T$ **do**
  /* Elasticity stress via expert + residual (Eqn. 8) */
  $\boldsymbol{\sigma} \leftarrow \mathcal{E}(\boldsymbol{F}, \boldsymbol{E}, \boldsymbol{\nu}) + \mathcal{E}_\theta(\boldsymbol{F}, \boldsymbol{l}_e)$;
  /* One MPM substep with external actuation */
  $\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{C}, \boldsymbol{F}^{trial} \leftarrow I(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{C}, \boldsymbol{F}, \boldsymbol{\sigma}, \boldsymbol{f}_{ex})$;
  /* Plasticity return-mapping via expert + residual (Eqn. 8) */
  $\boldsymbol{F} \leftarrow \mathcal{P}(\boldsymbol{F}^{trial}) + \mathcal{P}_\theta(\boldsymbol{F}^{trial}, \boldsymbol{l}_p)$;
$\boldsymbol{s}_{n+1} \leftarrow \{\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{C}, \boldsymbol{F}\}$;

We manually set the density ρ, and jointly optimize per-particle $\boldsymbol{E}$, $\boldsymbol{\nu}$, the LBS-impact factors $\boldsymbol{\omega}$ from Eqn. 7, the feature vectors $(\boldsymbol{l}_e, \boldsymbol{l}_p)$, and the parameters of $\mathcal{E}_\theta(\cdot)$ and $\mathcal{P}_\theta(\cdot)$ from Eqn. 8. We further introduce a regularizer $\mathcal{R}$ to limit residual magnitude and injected actuation:

$$\mathcal{R} = \lambda_{law}\Big(\|\mathcal{E}_\theta(\boldsymbol{F}, \boldsymbol{l}_e)\|_2 + \|\mathcal{P}_\theta(\boldsymbol{F}^{trial}, \boldsymbol{l}_p)\|_2\Big) + \lambda_\omega \|\boldsymbol{\omega}\|_2. \tag{11}$$

Progressive Loss-Balanced Optimization. Material parameters govern global dynamics, and thus aggressively fitting later frames when early frames are inaccurate destabilizes training [71]. We therefore train progressively where we begin with a short prefix of frames and expand the window once early dynamics stabilize. After each cycle, we compute per-frame losses and allocate more iterations to higher-loss frames, which accelerates convergence and avoids wasting updates on already well-fit frames.

5. Experimental Results

In this section, we conduct both qualitative and quantitative evaluations to demonstrate the effectiveness of our approach. We collect a dataset to conduct these evaluations (Sec. 5.1). We compare our method with state-of-the-art dynamic reconstruction baselines and evaluate rendering accuracy on both the observed sequences and the future prediction sequences (Sec. 5.2). We further perform ablation studies to analyze each component in our framework (Sec. 5.3).

5.1. Dataset

To address the lack of monocular datasets capturing dynamic scenes of human and elastodynamic objects, we collect a dataset at 1080p and 30 FPS for training and evaluation. The camera is static and its vertical axis is aligned with the gravity direction. The dataset contains 8 sequences with 6 objects. As shown in Fig. 4, each sequence consists of two stages: (1) a human spin clip for reconstructing high-quality 3D Gaussians, and (2) a dynamic clip of human and object. We further divide the dynamic clip into the observation part and the prediction part. For all sequences, we detect SMPL pose sequences using off-the-shelf estimators [52, 59].

5.2. Comparison

We compare our method against state-of-the-art monocular human reconstruction [30] and monocular dynamic reconstruction [63] approaches. For the reconstruction part, we train GART following a similar strategy to ours by first reconstructing 3D Gaussians on the spin stage, and then optimizing on the dynamic stage to allow the motion bases to learn temporal deformation. For 4D-Gaus, to ensure it produces meaningful results, we convert the transformation

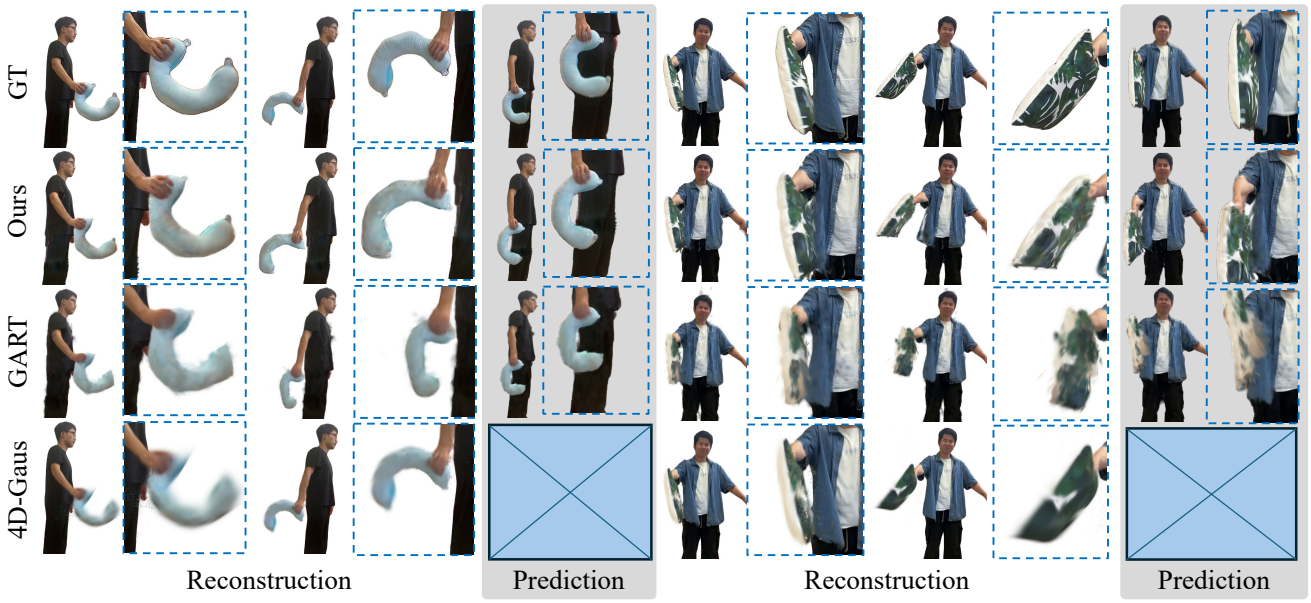


Figure 5. Qualitative comparison of dynamic reconstruction and future prediction with GART [30] and 4D-Gaus [63].

	Square Pillow						Cloth Bag					
	Full			40-60%			Full			30-50%		
Reconstruction	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Ours	21.69	0.8872	0.1079	20.64	0.8818	0.1150	24.24	0.9473	0.0804	23.58	0.9464	0.0909
GART [30]	18.81	0.9145	0.1282	17.46	0.8728	0.1322	23.45	0.9389	0.0845	21.59	0.9236	0.0984
4D Gaussian [63]	27.64	0.9252	0.1099	25.23	0.9117	0.1180	28.17	0.9495	0.0837	26.35	0.9595	0.0985
	Full			30-50%			Full			20-40%		
Prediction	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Ours	18.94	0.8777	0.1300	18.18	0.8755	0.1398	21.16	0.9247	0.0845	20.20	0.9133	0.0903
GART [30]	18.57	0.8798	0.1564	16.80	0.8573	0.1614	20.63	0.9316	0.0910	18.10	0.9025	0.1036
	C-shape Pillow #1						C-shape Pillow #2					
	Full			30-50%			Full			50-70%		
Reconstruction	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Ours	24.86	0.9489	0.0676	23.96	0.9440	0.0701	24.45	0.9336	0.0651	24.02	0.9305	0.0658
GART [30]	24.80	0.9375	0.0690	22.99	0.9410	0.0738	21.79	0.9294	0.0984	21.12	0.9296	0.0990
4D Gaussian [63]	29.03	0.9689	0.0703	28.79	0.9685	0.0788	26.72	0.9461	0.0839	25.02	0.9372	0.0864
	Full			10-30%			Full			10-30%		
Prediction	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Ours	21.12	0.9453	0.0788	20.41	0.9335	0.0795	21.46	0.9183	0.0945	21.37	0.9176	0.0944
GART [30]	21.69	0.9450	0.0893	19.32	0.9270	0.0914	20.02	0.9138	0.1170	19.80	0.9082	0.1269

Table 1. Quantitative comparison of reconstruction accuracy and future prediction. **FUll** is the full sequence. **#-#%** denotes its subset.

of the human root joint into camera motion during training, which simplifies its training. Without this modification, 4D-Gaus fails to learn the global motion of the human body. As illustrated in Fig. 5, GART struggles to preserve high-quality textures and its latent motion bases cannot model large deformations. 4D-Gaus produces visually blurry reconstructions. As shown in Tab. 1, we evaluate rendering accuracy on the full sequence (**Full**) and specifically on frames with large deformations (**#-#%**). Our method achieves the best performance on LPIPS while remaining competitive or superior on the other metrics. Although the visual quality of GART and 4D-Gaus severely degrades, they still can obtain relatively high PSNR and SSIM scores. This is because in both methods, the 3D Gaussians are continuously optimized to match the ground-truth rendering, allowing them to overfit pixel-aligned metrics. In contrast, in our approach, the appearance of the 3D Gaussians needs to be fixed for learning the physical models. It is thus easier for GART and 4D-Gaus to achieve good PSNR and SSIM scores. However, PSNR and SSIM mainly evaluate pixel-wise alignment, while LPIPS focuses on texture fidelity and perceptual similarity. Therefore, our method consistently outperforms others in LPIPS, indicating superior visual realism and texture preservation.

For the future prediction part, we compare our method against GART, since 4D-Gaus cannot produce results beyond the training frames. As shown in Fig. 5, given future human poses, the motion bases learned by GART fail to model the deformation and could even collapse the object structure. In contrast, our method generates physically rea-

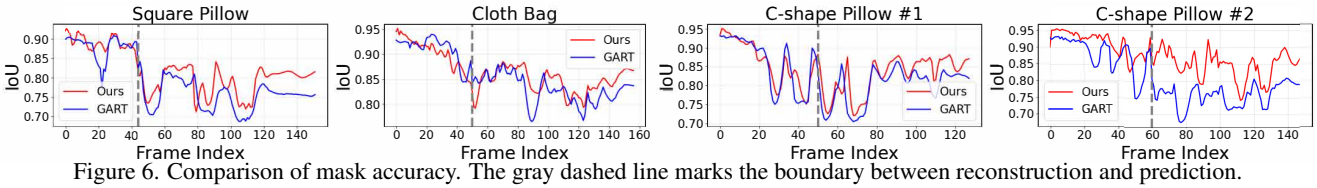


Figure 6. Comparison of mask accuracy. The gray dashed line marks the boundary between reconstruction and prediction.

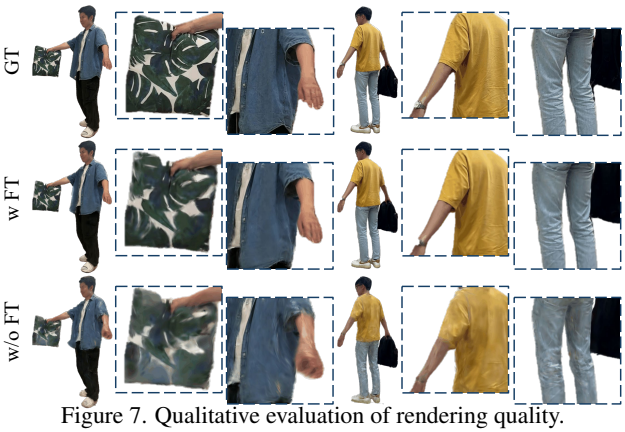


Figure 7. Qualitative evaluation of rendering quality.

	PSNR ↑	SSIM ↑	LPIPS ↓
w Fine-tuning	**27.30**	**0.9464**	**0.0681**
w/o Fine-tuning	25.42	0.9292	0.0854

Table 2. Quantitative evaluation of rendering quality.

	PSNR ↑	SSIM ↑	LPIPS ↓	IoU ↑
Full	**24.03**	**0.9534**	**0.0652**	**0.8845**
w/o $\boldsymbol{l}_e, \boldsymbol{l}_p$	23.54	0.9436	0.0680	0.8636
w/o $\mathcal{E}_\theta, \mathcal{P}_\theta$	22.26	0.9387	0.0664	0.8289

Table 3. Quantitative evaluation of reconstruction accuracy.

sonable dynamics through simulation. As shown in Tab. 1, we achieve better scores on the full sequence (**Full**) and especially on the subset with larger deformations (**#-#%**). This is because GART produces severely degraded results at frames with large deformation, while our physically-based simulation remains robust across the entire sequence.

We further compare the Intersection over Union (IoU) between the rendered alpha masks and the ground truth masks across the full dynamic sequences. As shown in Fig. 6, our method achieves higher IoU than GART in both reconstruction and prediction, particularly on frames with large deformations, demonstrating the advantage of our approach in dynamic representation.

5.3. Ablation Study

We evaluate our method in terms of both rendering quality and physical reconstruction accuracy. For the rendering quality on the spin stage, as shown in Fig. 7, without fine-tuning (**w/o FT**), directly rendering with the physics-aware covariance causes the original canonical 3D Gaussians to no longer fit the observations. The appearance is blurry and the texture is distorted. After fine-tuning (**w FT**), the rendering quality improves, producing textures that more closely match the ground truth. As shown in Tab. 2, fine-tuning

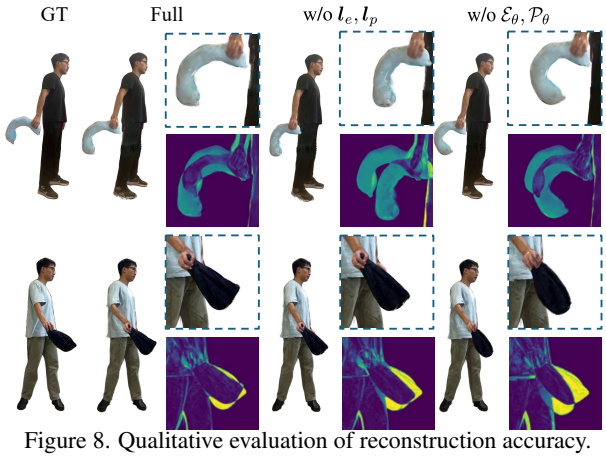


Figure 8. Qualitative evaluation of reconstruction accuracy.

leads to better rendering scores.

For the dynamic physical reconstruction, as shown in Fig. 8, removing the feature vectors (**w/o** $\boldsymbol{l}_e, \boldsymbol{l}_p$) reduces the expressiveness of the physical models, resulting in larger reconstruction errors compared to the ground truth. When residual laws are removed and only the expert constitutive models are used (**w/o** $\mathcal{E}_\theta, \mathcal{P}_\theta$), the simulator struggles to reproduce the observed dynamic behaviors. As shown in Tab. 3, our full model achieves the best rendering accuracy and IoU score. Additionally, without 3D flow supervision, the neural residual models tend to overfit frames with large reconstruction errors, which leads to excessive correction of the expert constitutive models and training failure.

6. Conclusion

We have presented PhysHO, a unified monocular framework that reconstructs physically-plausible dynamic human-object scenarios by integrating LBS-driven actuation with MPM-based simulation. By treating LBS trajectories as controllable internal forces and learning their spatial influence through the LBS-impact factor, our system effectively bridges kinematic motion priors and physics-based dynamics. Furthermore, by extending expert constitutive models with neural residual laws conditioned on per-particle features, our method is capable of representing heterogeneous and anisotropic material behaviors. To ensure robust monocular optimization, we incorporate a structure-preserving 3D flow supervision and a progressive, loss-balanced training schedule, enabling stable convergence. The experimental results demonstrate the superiority of our method in both rendering quality and physical accuracy with the state-of-the-art methods.

Acknowledgment. This research / project is supported by the National Research Foundation (NRF) Singapore, under its NRF-Investigatorship Programme (Award ID. NRF-NRFI09-0008).

References

[1] Aljaz Bozic, Pablo Palafox, Michael Zollhofer, Justus Thies, Angela Dai, and Matthias Nießner. Neural deformation graphs for globally-consistent non-rigid reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1450–1459, 2021. 2

[2] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 130–141, 2023. 2

[3] Junyi Cao, Shanyan Guan, Yanhao Ge, Wei Li, Xiaokang Yang, and Chao Ma. NeuMA: Neural material adaptor for visual grounding of intrinsic dynamics. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024. 2, 5

[4] Jianchuan Chen, Ying Zhang, Di Kang, Xuefei Zhe, Linchao Bao, Xu Jia, and Huchuan Lu. Animatable neural radiance fields from monocular rgb videos. *arXiv preprint arXiv:2106.13629*, 2021. 2

[5] Jianchuan Chen, Jingchuan Hu, Gaige Wang, Zhonghua Jiang, Tiansong Zhou, Zhiwen Chen, and Chengfei Lv. Taoavatar: Real-time lifelike full-body talking avatars for augmented reality via 3d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10723–10734, 2025. 2

[6] Devikalyan Das, Christopher Wewer, Raza Yunus, Eddy Ilg, and Jan Eric Lenssen. Neural parametric gaussians for monocular non-rigid object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10715–10725, 2024. 1, 2

[7] Yitong Deng, Hong-Xing Yu, Jiajun Wu, and Bo Zhu. Learning vortex dynamics for fluid inference and prediction. In *Proceedings of the International Conference on Learning Representations*, 2023. 2

[8] Zhiyang Dou, Xuelin Chen, Qingnan Fan, Taku Komura, and Wenping Wang. C· ase: Learning conditional adversarial skill embeddings for physics-based characters. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 3

[9] Tao Du, Kui Wu, Pingchuan Ma, Sebastien Wah, Andrew Spielberg, Daniela Rus, and Wojciech Matusik. Diffpd: Differentiable projective dynamics. *ACM Trans. Graph.*, 41(2), 2021. 2

[10] Yuanxing Duan, Fangyin Wei, Qiyu Dai, Yuhang He, Wenzheng Chen, and Baoquan Chen. 4d-rotor gaussian splatting: Towards efficient novel-view synthesis for dynamic scenes. In *Proc. SIGGRAPH*, 2024. 1

[11] Mathieu Dubied, Mike Yan Michelis, Andrew Spielberg, and Robert Katzschmann. Sim-to-real for soft robots using differentiable fem: Recipes for meshing, damping, and actuation. *IEEE Robotics and Automation Letters*, 7:1–1, 2022. 2

[12] Bardienus P Duisterhof, Zhao Mandi, Yunchao Yao, Jia-Wei Liu, Jenny Seidenschwarz, Mike Zheng Shou, Ramanan Deva, Shuran Song, Stan Birchfield, Bowen Wen, and Jeffrey Ichnowski. DeformGS: Scene flow in highly deformable scenes for deformable object manipulation. *WAFR*, 2024. 1, 2

[13] Qiao Feng, Yiming Huang, Yufu Wang, Jiatao Gu, and Lingjie Liu. Physhmr: Learning humanoid control policies from vision for physically plausible human motion reconstruction. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–10, 2025. 3

[14] Yutao Feng, Xiang Feng, Yintong Shang, Ying Jiang, Chang Yu, Zeshun Zong, Tianjia Shao, Hongzhi Wu, Kun Zhou, Chenfanfu Jiang, and Yin Yang. Gaussian splashing: Unified particles for versatile motion synthesis and rendering. *arXiv preprint arXiv:2401.15318*, 2024. 2

[15] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023. 2

[16] Erik Gärtner, Mykhaylo Andriluka, Erwin Coumans, and Cristian Sminchisescu. Differentiable dynamics for articulated 3d human motion reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13190–13200, 2022. 3

[17] Chen Geng, Sida Peng, Zhen Xu, Hujun Bao, and Xiaowei Zhou. Learning neural volumetric representations of dynamic humans in minutes. In *CVPR*, 2023. 1

[18] Zhiyang Guo, Wengang Zhou, Li Li, Min Wang, and Houqiang Li. Motion-aware 3d gaussian splatting for efficient dynamic scene reconstruction. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2024. 2

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 5

[20] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[21] Yuanming Hu, Yu Fang, Ziheng Ge, Ziyin Qu, Yixin Zhu, Andre Pradhana, and Chenfanfu Jiang. A moving least squares material point method with displacement discontinuity and two-way rigid body coupling. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 2, 3

[22] Chenfanfu Jiang, Craig Schroeder, Joseph Teran, Alexey Stomakhin, and Andrew Selle. The material point method for simulating continuum materials. In *ACM SIGGRAPH 2016 Courses*, New York, NY, USA, 2016. Association for Computing Machinery. 2

[23] Chenfanfu Jiang, Craig Schroeder, Joseph Teran, Alexey Stomakhin, and Andrew Selle. The material point method for simulating continuum materials. In *Acm siggraph 2016 courses*, pages 1–52. 2016. 2, 3

[24] Ying Jiang, Chang Yu, Tianyi Xie, Xuan Li, Yutao Feng, Huamin Wang, Minchen Li, Henry Lau, Feng Gao, Yin Yang, and Chenfanfu Jiang. Vr-gs: A physical dynamics-aware interactive gaussian splatting system in virtual reality. *arXiv preprint arXiv:2401.16663*, 2024. 2

[25] Erik Johnson, Marc Habermann, Soshi Shimada, Vladislav Golyanik, and Christian Theobalt. Unbiased 4d: Monocular 4d reconstruction with a neural deformation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6598–6607, 2023. 2

[26] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 1, 2

[27] Muhammed Kocabas, Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats, 2023. 2

[28] Simon Le Cleac'h, Hong-Xing Yu, Michelle Guo, Taylor Howell, Ruohan Gao, Jiajun Wu, Zachary Manchester, and Mac Schwager. Differentiable physics simulation of dynamics-augmented neural objects. *IEEE Robotics and Automation Letters*, 8(5):2780–2787, 2023. Publisher: IEEE. 2

[29] Changmin Lee, Jihyun Lee, and Tae-Kyun Kim. Mpma-vatar: Learning 3d gaussian avatars with accurate and robust physics-based dynamics. In *NeurIPS*, 2025. 3

[30] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19876–19887, 2024. 1, 2, 3, 6, 7

[31] Xuan Li, Yi-Ling Qiao, Peter Yichen Chen, Krishna Murthy Jatavallabhula, Ming Lin, Chenfanfu Jiang, and Chuang Gan. Pac-nerf: Physics augmented continuum neural radiance fields for geometry-agnostic system identification. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[32] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[33] Youtian Lin, Zuozhuo Dai, Siyu Zhu, and Yao Yao. Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21136–21145, 2024. 2

[34] Youtian Lin, Zuozhuo Dai, Siyu Zhu, and Yao Yao. Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21136–21145, 2024. 1

[35] Yuchen Lin, Chenguo Lin, Jianjin Xu, and Yadong MU. OmniphysGS: 3d constitutive gaussians for general physics-based dynamics generation. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[36] Fangfu Liu, Hanyang Wang, Shunyu Yao, Shengjun Zhang, Jie Zhou, and Yueqi Duan. Physics3d: Learning physical properties of 3d gaussians via video diffusion. *arXiv preprint arXiv:2406.04338*, 2024. 2

[37] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, 2015. 1, 2, 3

[38] Zhengyi Luo, Jinkun Cao, Kris Kitani, Weipeng Xu, et al. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10895–10904, 2023. 3

[39] 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. 2, 5, 1

[40] Miles Macklin, Matthias Müller, and Nuttapong Chentanez. Xpbd: position-based simulation of compliant constrained dynamics. In *Proceedings of the 9th International Conference on Motion in Games*, page 49–54, New York, NY, USA, 2016. Association for Computing Machinery. 2

[41] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2

[42] Arthur Moreau, Jifei Song, Helisa Dhamo, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Human gaussian splatting: Real-time rendering of animatable avatars. In *CVPR*, 2024. 1, 2

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

[44] Junfeng Ni, Yixin Chen, Bohan Jing, Nan Jiang, Bin Wang, Bo Dai, Puhao Li, Yixin Zhu, Song-Chun Zhu, and Siyuan Huang. Phyrecon: Physically plausible neural scene reconstruction. 2024. 2

[45] David Novotny, Ignacio Rocco, Samarth Sinha, Alexandre Carlier, Gael Kerchenbaum, Roman Shapovalov, Nikita Smetanin, Natalia Neverova, Benjamin Graham, and Andrea Vedaldi. Keytr: Keypoint transporter for 3d reconstruction of deformable objects in videos. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5585–5594, 2022. 1, 2

[46] Jongmin Park, Minh-Quan Viet Bui, Juan Luis Gonzalez Bello, Jaeho Moon, Jihyong Oh, and Munchurl Kim. Splinegs: Robust motion-adaptive spline for real-time dynamic 3d gaussians from monocular video. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 26866–26875, 2025. 2

[47] Pramish Paudel, Anubhav Khanal, Danda Pani Paudel, Jyoti Tandukar, and Ajad Chhatkuli. ihuman: Instant animatable digital humans from monocular videos. In *European Conference on Computer Vision*, pages 304–323. Springer, 2024. 2

[48] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforce-

ment learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4):1–14, 2018. 3

[49] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4):1–20, 2021. 3

[50] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. 2024. 2

[51] Yidi Shao, Mu Huang, Chen Change Loy, and Bo Dai. Gaussim: Registering elastic objects into digital world by gaussian simulator. In *ICCV*, 2025. 2

[52] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia Conference Proceedings*, 2024. 6

[53] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics (ToG)*, 39(6):1–16, 2020. 3

[54] Olga Sorkine and Marc Alexa. As-rigid-as-possible shape manipulation. *ACM Transactions on Graphics (TOG)*, 26 (3):113, 2007. 5

[55] 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. 2, 3

[56] Chen Tessler, Yoni Kasten, Yunrong Guo, Shie Mannor, Gal Chechik, and Xue Bin Peng. Calm: Conditional adversarial latent models for directable virtual characters. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–9, 2023. 3

[57] Artur P Toshev, Harish Ramachandran, Jonas A Erbesdobler, Gianluca Galletti, Johannes Brandstetter, and Nikolaus A Adams. Jax-sph: A differentiable smoothed particle hydrodynamics framework. *arXiv preprint arXiv:2403.04750*, 2024. 2

[58] Qianqian Wang, Vickie Ye, Hang Gao, Weijia Zeng, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. 2024. 2

[59] Yufu Wang and Kostas Daniilidis. Refit: Recurrent fitting network for 3d human recovery. In *International Conference on Computer Vision*, 2023. 6

[60] Yikai Wang, Xinzhou Wang, Zilong Chen, Zhengyi Wang, Fuchun Sun, and Jun Zhu. Vidu4d: Single generated video to high-fidelity 4d reconstruction with dynamic gaussian surfels. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2

[61] Yinhuai Wang, Qihan Zhao, Runyi Yu, Hok Wai Tsui, Ailing Zeng, Jing Lin, Zhengyi Luo, Jiwen Yu, Xiu Li, Qifeng Chen, et al. Skillmimic: Learning basketball interaction skills from demonstrations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17540–17549, 2025. 3

[62] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, 2022. 1, 2

[63] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20310–20320, 2024. 1, 2, 6, 7

[64] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. *arXiv preprint arXiv:2311.12198*, 2023. 2

[65] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4389–4398, 2024. 4

[66] Tianju Xue, Shuheng Liao, Zhengtao Gan, Chanwook Park, Xiaoyu Xie, Wing Kam Liu, and Jian Cao. Jax-fem: A differentiable gpu-accelerated 3d finite element solver for automatic inverse design and mechanistic data science. *Computer Physics Communications*, page 108802, 2023. 2

[67] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2853–2863, 2022. 2

[68] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *CVPR*, 2022. 1

[69] Zeyu Yang, Zijie Pan, Xiatian Zhu, Li Zhang, Jianfeng Feng, Yu-Gang Jiang, and Philip HS Torr. 4d gaussian splatting: Modeling dynamic scenes with native 4d primitives. *arXiv preprint arXiv:2412.20720*, 2024. 1, 2

[70] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. In *International Conference on Learning Representations (ICLR)*, 2024. 1, 2

[71] Hao Zhang, Haolan Xu, Chun Feng, Varun Jampani, and Narendra Ahuja. Physrig: Differentiable physics-based skinning and rigging framework for realistic articulated object modeling. 2025. 2, 3, 6

[72] Xinyu Zhang, Haonan Chang, Yuhan Liu, and Abdeslam Boularias. Motion blender gaussian splatting for dynamic scene reconstruction. *arXiv preprint arXiv:2503.09040*, 2025. 2

[73] Yiqun Zhao, Chenming Wu, Binbin Huang, Yihao Zhi, Chen Zhao, Jingdong Wang, and Shenghua Gao. Surfel-based gaussian inverse rendering for fast and relightable dynamic human reconstruction from monocular videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[74] Licheng Zhong, Hong-Xing Yu, Jiajun Wu, and Yunzhu Li. Reconstruction and simulation of elastic objects with spring-mass 3d gaussians. *European Conference on Computer Vision (ECCV)*, 2024. 2

PhysHO: Physics-Based Dynamic 3D Gaussian Human and Object from Monocular Video

Supplementary Material

In the supplementary material, we first provide the implementation details in Sec. A. We then include further discussions of our method in Sec. B, covering the limitations as well as ethical and social impacts. Finally, we present additional experimental results in Sec. C, including more comparisons, visualizations of more results, and application.

A. Implementation Details

A.1. Computational Cost

We perform all training and simulation on a single RTX 4090 GPU. The input is 30 FPS video sequence. For spin stage reconstruction, the number of training frames is around 300, and training takes about 10 minutes. During the finetuning of canonical 3D Gaussians (Sec. 4.1, Eqn. 5), we first pre-compute the deform gradients at each frame, and the finetuning takes about 10 minutes. In the dynamic stage, the number of training frames is around 50-60. Optimization of structure-preserving 3D flow (Sec. 4.4, Eqn. 9) takes about 1 hour. For the learning of the physics model, the training takes around 6 hours with our progressive loss-balanced optimization strategy. After training, the inference for 100 frames takes about 3 minutes.

A.2. Neural Residual Constitutive Laws

For the neural residual constitutive laws, we adopt the same MLP architecture as NCLaw [39] for both $\mathcal{E}_\theta, \mathcal{P}_\theta$. Each MLP consists of layers with dimensions (13, 64, 64, 9). The input of the network is the concatenation of $\boldsymbol{\Sigma}$ (singular values of $\boldsymbol{F}$), $\boldsymbol{F}^T\boldsymbol{F}$, and the determinants $det(\boldsymbol{F})$. As in Sec. 4.3, each particle is assigned optimizable feature vectors $\boldsymbol{l}_e, \boldsymbol{l}_p$ of dimension 64, which are added to the output of the first MLP hidden layers of $\mathcal{E}_\theta, \mathcal{P}_\theta$ respectively. The summation results are then passed into the subsequent layers. During training, in addition to the regularization term in Eqn. 11, we also impose a regularization term on the per-particle features to constrain their influence on the residual constitutive models:

$$\mathcal{R}_{feat} = \lambda_{feat}(\|\boldsymbol{l}_e\|_2 + \|\boldsymbol{l}_p\|_2).$$

During inference, we apply activation truncation to ensure stable simulation of animation:

$$\mathcal{O} = \begin{cases} \mathbf{0}, & if \|\mathcal{O}\|_{Fro} < \epsilon \\ \mathcal{O}, & else. \end{cases} \tag{12}$$

Here, $\mathcal{O}$ is the output matrix of $\mathcal{E}_\theta$ or $\mathcal{P}_\theta$. $\|\cdot\|_{fro}$ is the Frobenius norm and ϵ is the threshold setting to $1e{-}3$.

A.3. Hyperparameter

In Eqn. 6 of Sec. 4.2, we set the gains of the PD controller as $k_p{=}2e2$ and $k_d{=}2e1$. In Eqn. 8, we choose corotated elasticity and identity plasticity as expert constitutive models $\mathcal{E}$ and $\mathcal{P}$. For the optimization of $\boldsymbol{E}$ and $\boldsymbol{\nu}$, we set their boundaries as $3e3{<}log(\boldsymbol{E}){<}1.5e5$ and $0.2{<}\boldsymbol{\nu}{<}0.4$. As for the loss weight terms, we set $\lambda_{rgb}{=}1$, $\lambda_{flow}{=}0.1$, $\lambda_{arap}{=}1e5$, $\lambda_{3Dflow}{=}1e2$, $\lambda_{law}{=}1e1$, $\lambda_{\omega}{=}1e{-}1$, and $\lambda_{feat}{=}1$. In Alg.2, the number T of steps per frame is 80.

B. Discussion

B.1. Limitations

Although our PhysHO is capable of reconstructing the physically plausible human–object dynamics, it still has several limitations. First, our method currently cannot handle topology changes. This is mainly because the human body and objects are jointly reconstructed and physically coupled. To support topology change, it requires decoupled modeling of the human and the object, as well as explicit reasoning about the contact force at the interaction point. Second, our pipeline heavily relies on the accuracy of monocular human pose estimation, especially for joints interacting with objects. Inaccurate and inconsistent estimation leads to incorrect modeling of internal driving forces, which significantly deteriorates the accuracy of the physical simulation. More robust and temporally consistent monocular pose estimation could alleviate this issue. Finally, our method remains computationally expensive. The physics-driven optimization and simulation take several hours of training, while non-physics-based reconstruction methods already operate at a minute-level runtime. Improving the computational efficiency of physically grounded reconstruction remains an important direction for future research.

B.2. Ethical and Social Impact

Human data inherently contains sensitive personal information. All data used in this work are collected with informed consent from participants and are strictly limited to academic research. Any future data release will follow the same purpose restriction to avoid misuse of identifiable human information.

From a broader social perspective, although current reconstructions remain distinguishable from real videos, advancements in photorealistic human modeling may blur the boundary between synthetic and real content. This could

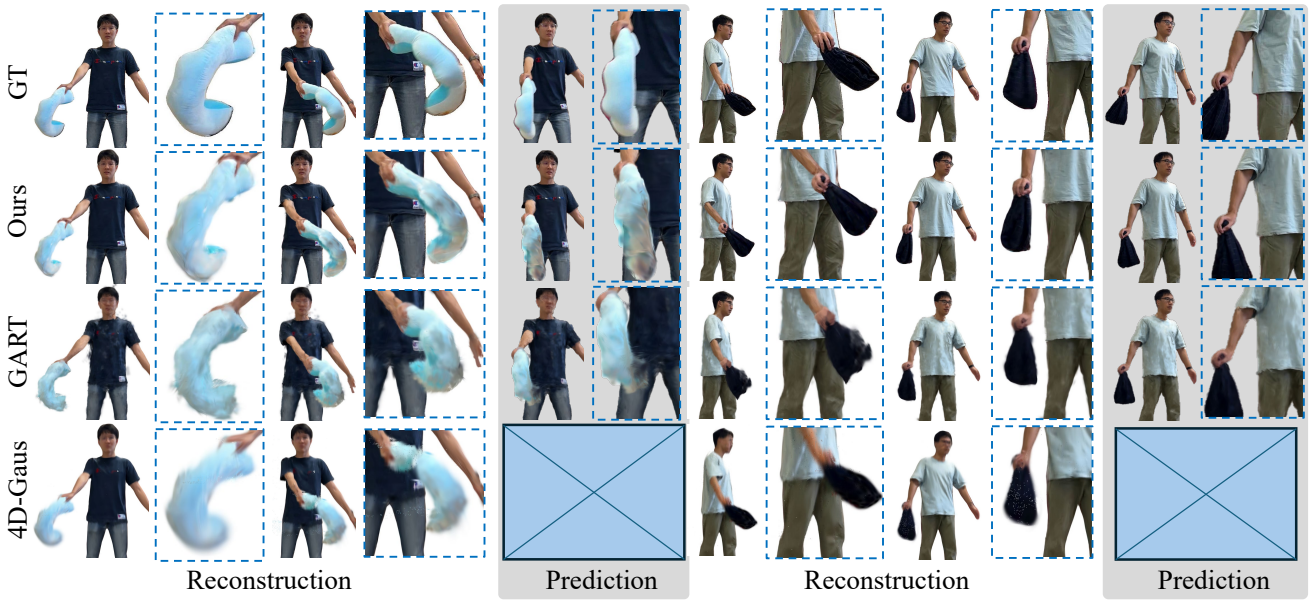


Figure 9. More qualitative comparison of dynamic reconstruction and future prediction with GART [30] and 4D-Gaus [63].

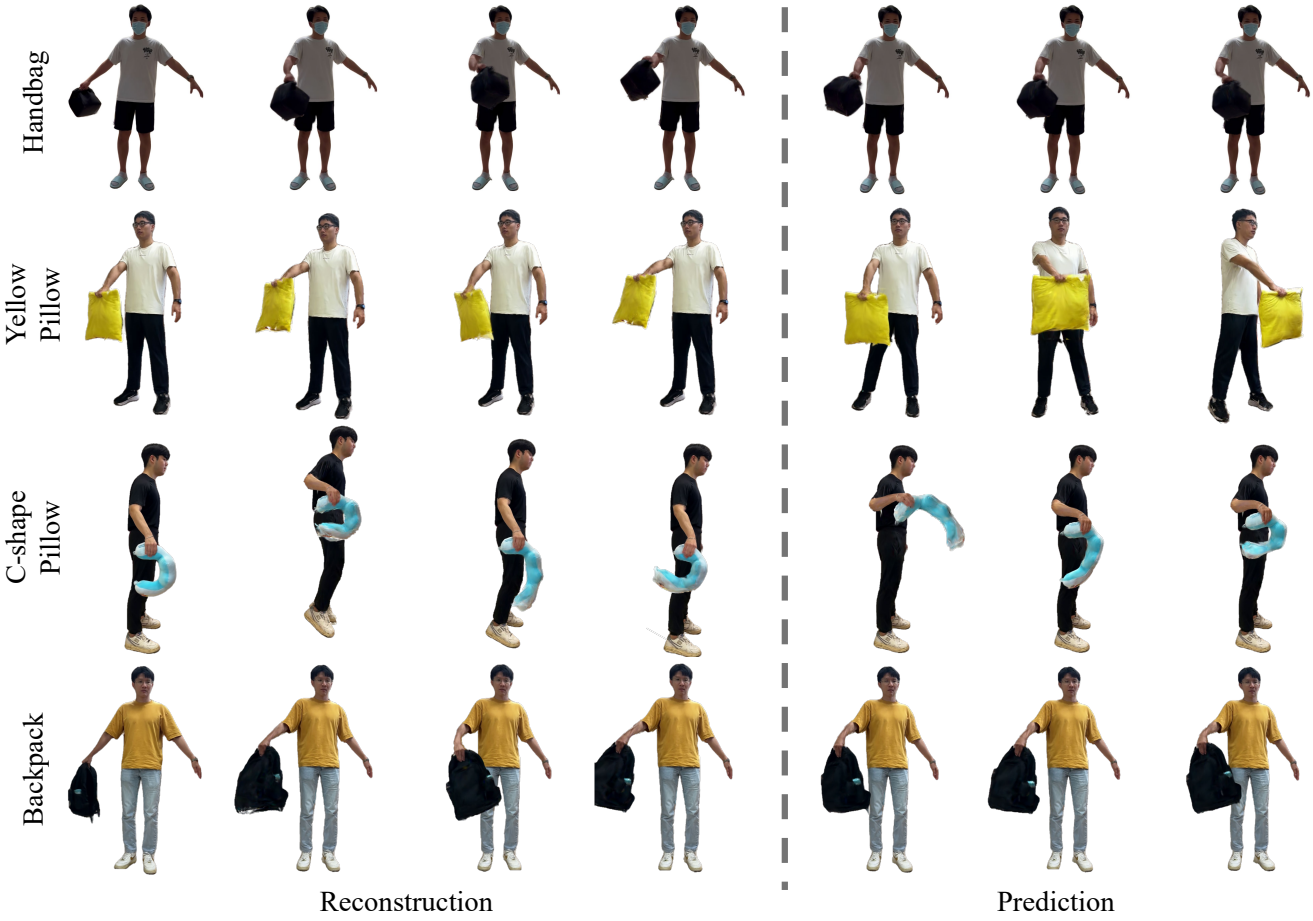


Figure 10. More results of our method, including both reconstruction and prediction.

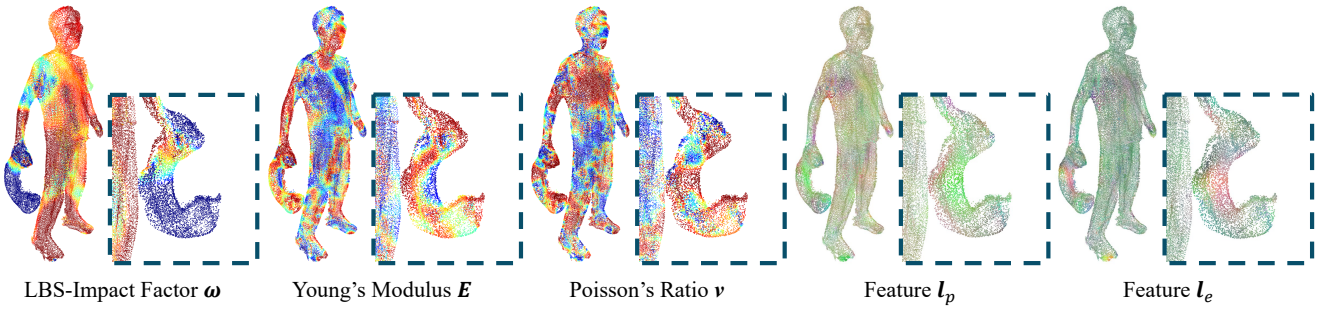


Figure 11. Visualization of the learned material space, including LBS-impact factor, Young's modulus, Poisson's ratio and features $\boldsymbol{l}_e$,$\boldsymbol{l}_p$.

Figure 12. Simulated animations on novel pose sequences.

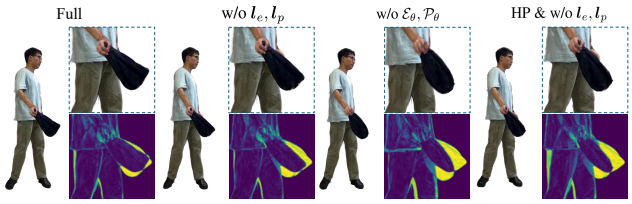


Figure 13. More qualitative evaluation of reconstruction accuracy.

Figure 14. Rotating-view rendering results.

open opportunities for misuse such as identity impersonation or manipulative media fabrication. Therefore, we emphasize that the technology should be applied responsibly, with transparency and clear usage constraints.

	PSNR ↑	SSIM ↑	LPIPS ↓	IoU ↑
HP & w/o $\boldsymbol{l}_e, \boldsymbol{l}_p$	21.55	0.9396	0.0690	0.8184

Table 4. More quantitative evaluation of reconstruction accuracy.

	PhysHO	GART	4D-Gaus
time	$\sim$ 4.5 hours	$\sim$ 8 mins	$\sim$ 15 mins
GPU memory	$\sim$ 8 GB	$\sim$ 4 GB	$\sim$ 2.5 GB

Table 6. Time and GPU usage.

C. More Experimental Results

C.1. More Comparison

As shown in Fig. 9, we provide more qualitative comparisons with GART and 4D-Gaus. Our method achieves the best rendering quality.

C.2. More Ablation Study

We further evaluate the results by enforcing homogeneous properties and removing feature vectors (**HP & w/o** $\boldsymbol{l}_e, \boldsymbol{l}_p$). As shown in Fig. 13, this leads to larger errors for both the human (see right leg) and the object. Tab. 4 complements Tab. 3 in the main paper with the results.

C.3. More Results

As illustrated in Fig. 10, we provide additional reconstruction and prediction results across multiple sequences. As shown in Fig. 14, we render the dynamic results of a rotating view after training. In Tab. 6, we report the training time and GPU usage for the three methods. Our material space, similar to GART, is represented using voxel grids. the parameters of each particle are obtained through spatial interpolation. As shown in Fig. 11, we visualize the material space for an optimized example. The LBS-impact factor on the object is zero across the region that does not contact the human body, indicating that the particles in this region are unaffected by additional actuation. Meanwhile, the spatially varying $\boldsymbol{E}, \boldsymbol{\nu}, \boldsymbol{l}_e, \boldsymbol{l}_p$ reveal that our model captures heterogeneous material properties across the reconstruction.

C.4. Application

As shown in Fig. 12, given novel pose sequences, our method not only recovers the human motion but also realistically simulates the physically driven non-rigid deformations of objects arising from human interactions, demonstrating strong generalization beyond observed training frames. Note that conventional LBS-based methods cannot represent such physically plausible effects.