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

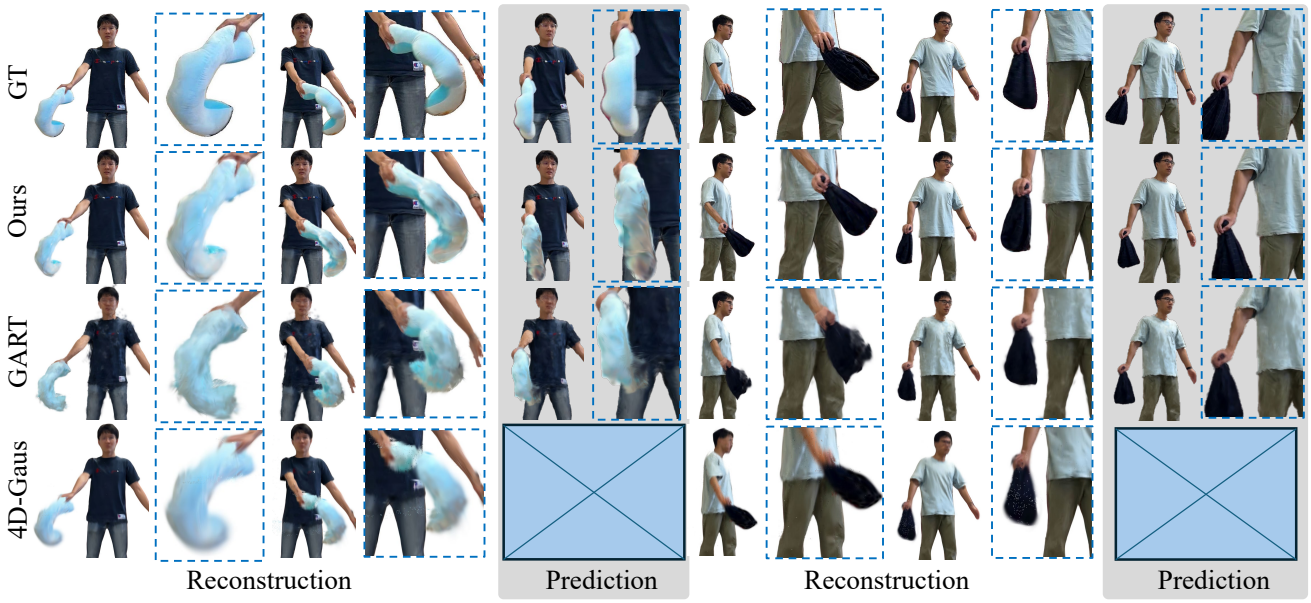

Figure 9. More qualitative comparison of dynamic reconstruction and future prediction with GART [30] and 4D-Gaus [63].

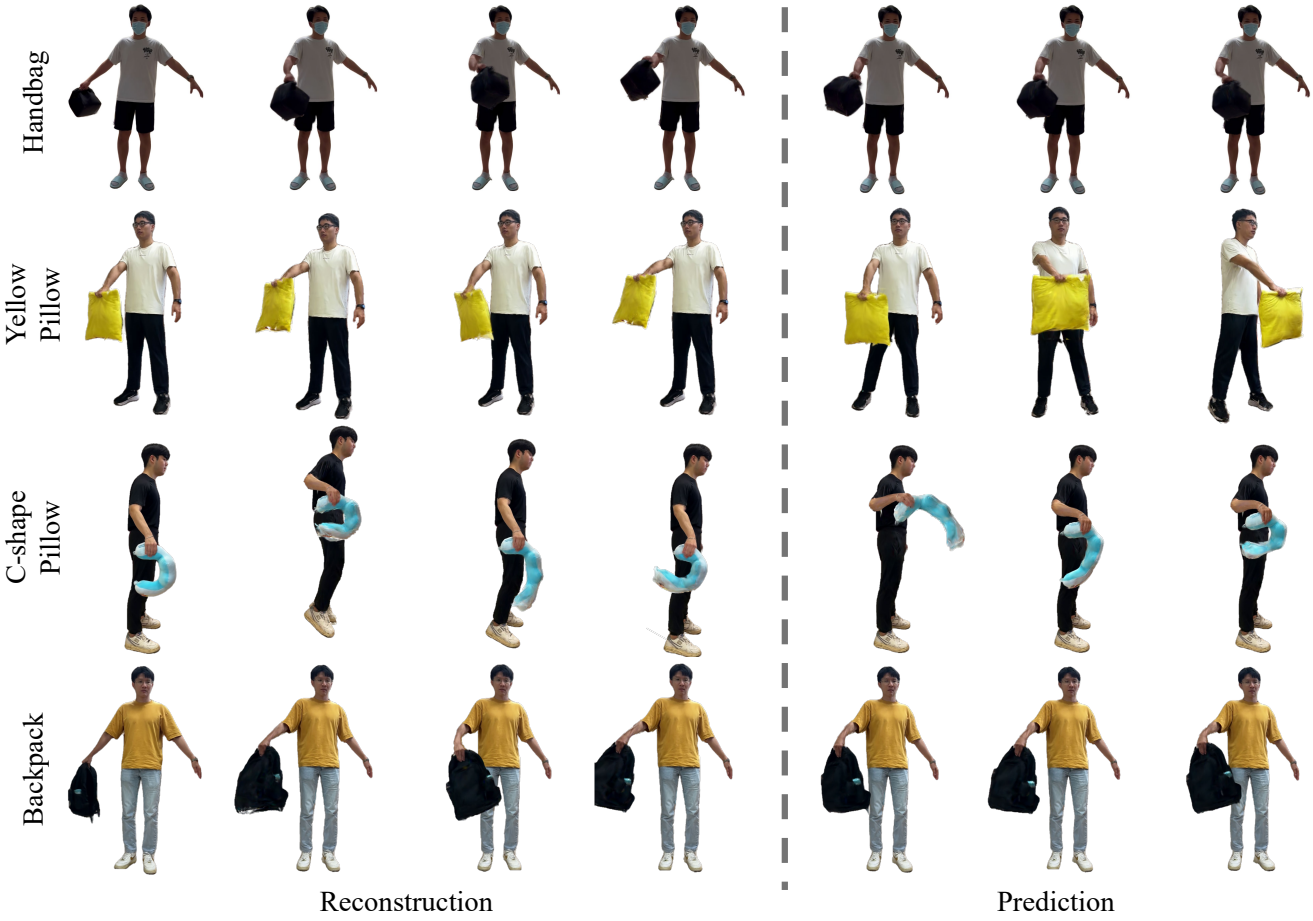

Figure 10. More results of our method, including both reconstruction and prediction.

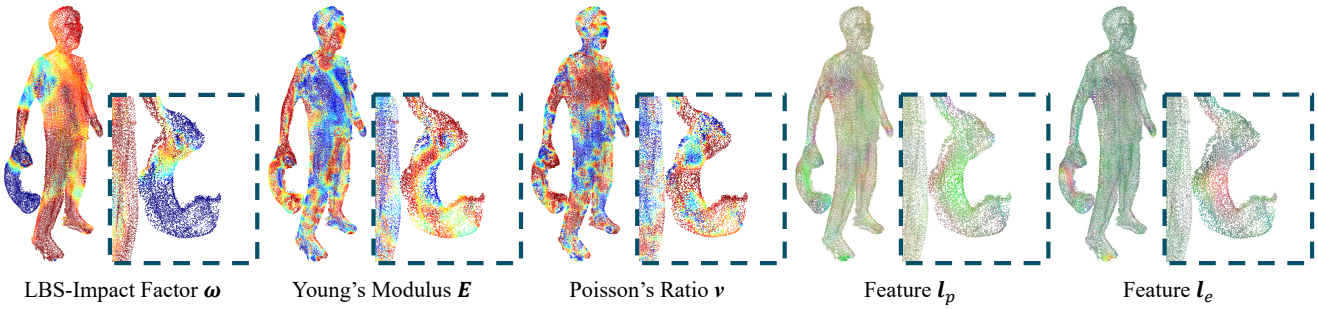

Figure 11. Visualization of the learned material space, including LBS-impact factor, Young's modulus, Poisson's ratio and features $\boldsymbol{l}_e$,$\boldsymbol{l}_p$.

Figure 12. Simulated animations on novel pose sequences.

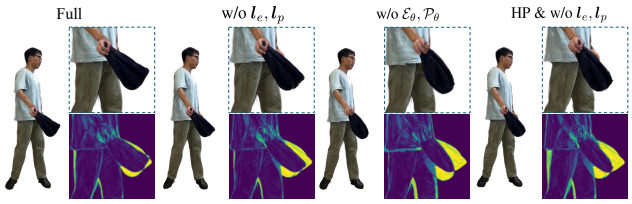

Figure 13. More qualitative evaluation of reconstruction accuracy.

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.