Supplementary Material
InterDiff: Generating 3D Human-Object Interactions with Physics-Informed Diffusion

Sirui Xu Zhengyuan Li Yu-Xiong Wang* Liang-Yan Gui*
University of Illinois at Urbana-Champaign
{siruixu2, zli138, yxw, lgui}@illinois.edu
https://sirui-xu.github.io/InterDiff/

In this supplementary material, we include additional method details and experimental results: (1) We provide a demo video, which is explained in detail in Sec. A. (2) We present additional information on our approach including the network architecture and learning objectives in Sec. B. (3) We provide additional implementation details in Sec. C. (4) We show additional ablation studies in Sec. D.

A. Visualization Video

In addition to the qualitative results in the main paper, we provide demos on the project website that showcase more comprehensive visualizations of the task, 3D human-object interaction (HOI) forecasting, and further demonstrate the effectiveness of our method. In demos, we visualize that without our proposed physics-informed correction step, pure diffusion produces implausible interactions, which is consistent with the results presented in Sec. 4 of the main paper. In addition, we demonstrate that our method InterDiff can forecast *diverse and extremely long-term* HOIs, while also maintaining their physical validity. Intriguingly, we observe that our method InterDiff consistently produces smooth and vivid HOIs, *even in cases where the ground truth data exhibit jitter patterns* from the motion capture process. Finally, we emphasize the impact and effectiveness of our contact-based coordinate system.

B. Additional Details of Methodology

B.1. Interaction Diffusion

In Sec. 3.1 of the main paper, we have highlighted our proposed InterDiff pipeline. Here, we explain the architecture and the learning objectives in detail.

Architecture. In the reverse diffusion process, the encoder and decoder consist of several transformer layers, respectively. We set the first and last layers as the original transformer layer [8], while the self-attention module in the middle layers is equipped with QnA [2], a local self-attention layer with learnable queries similar to [7]. The encoder contains an additional PointNet [6] that extracts the feature of the object in the canonical pose. This shape encoding is directly added to the encoding of the past interaction, which is further processed by the transformer encoder.

Learning Objectives. As mentioned in the main paper, we disentangle the learning objective into rotation and translation losses for the human state $\boldsymbol{h}$ and the object state $\boldsymbol{o}$, respectively. The original learning objective is denoted as

$$\begin{aligned} \boldsymbol{x}_0(t) &= \mathcal{G}(\boldsymbol{x}_t, t, \boldsymbol{c}), \\ \mathcal{L}_r &= \mathbb{E}_{t\sim[1,T]}\|\boldsymbol{x}_0(t) - \boldsymbol{x}\|_2^2, \end{aligned} \tag{1}$$

where $\boldsymbol{x}_0(t)$ is the result given by the reverse process at step t, and $\boldsymbol{x}$ is the ground truth data, as defined in Sec. 3.1 of the main paper.

The disentangled objectives are denoted as

$$\begin{aligned} \mathcal{L}_h &= \mathbb{E}_{t\sim[1,T]}\|\boldsymbol{h}_0(t) - \boldsymbol{h}\|_2^2, \\ \mathcal{L}_o &= \mathbb{E}_{t\sim[1,T]}\|\boldsymbol{o}_0(t) - \boldsymbol{o}\|_2^2, \end{aligned} \tag{2}$$

where $\boldsymbol{h}_0(t), \boldsymbol{h}$ are the human motion generated by the diffusion model and the ground truth data, respectively. And $\boldsymbol{o}_0(t), \boldsymbol{o}$ are the denoised object motion and the ground truth, respectively.

To promote a smooth interaction over time, we introduce velocity regularizations as:

$$\begin{aligned} \mathcal{L}_{vh} &= \mathbb{E}_{t\sim[1,T]}\|\boldsymbol{h}_0^{H+1:H+F}(t) - \boldsymbol{h}_0^{H:H+F-1}(t)\|_2^2, \\ \mathcal{L}_{vo} &= \mathbb{E}_{t\sim[1,T]}\|\boldsymbol{o}_0^{H+1:H+F}(t) - \boldsymbol{o}_0^{H:H+F-1}(t)\|_2^2. \end{aligned} \tag{3}$$

B.2. Interaction Correction

Architecture. Here, we use SMPL [4]-represented human interactions as example, while we extract markers [10] over

*Equal contribution.

the body meshes as reference. The skeleton-based interaction will follow the same process but use joints as reference. We represent the object motion under *every* reference system as a spatial-temporal graph $\mathbf{G}^{1:H} \in \mathbb{R}^{H\times(1+|\mathcal{M}|)\times D_o}$, where D_o is the number of features for object poses, $1+|\mathcal{M}|$ correspond to 1 ground reference system and $|\mathcal{M}|$ marker-based reference systems, as mentioned in Sec. 3.2.2 of the main paper. Following [5], we first replicate the last frame F times and get $\widehat{\mathbf{G}}^{1:H+F} \in \mathbb{R}^{(H+F)\times(1+|\mathcal{M}|)\times D_o}$, then transform it into the frequency domain. Specifically, given the defined M discrete cosine transform (DCT) [1] bases $\mathbf{C} \in \mathbb{R}^{M\times(H+F)}$, the graph is processed as

$$\tilde{\mathbf{G}}^{1:H+F} = \mathbf{C}\widehat{\mathbf{G}}^{1:H+F}. \tag{4}$$

After applying multiple spatial-temporal graph convolutions to obtain the result $\tilde{\mathbf{G}'}^{1:H+F}$, we convert it back to the temporal domain, denoted as

$$\widehat{\mathbf{G}'}^{1:H+F} = \mathbf{C}^{\mathsf{T}}\tilde{\mathbf{G}'}^{1:H+F}, \tag{5}$$

where we extract the future frames $\widehat{\mathbf{G}'}^{H:H+F}$. As described in Sec. 3.2.2 of the main paper, from this graph, we index the specific future object motion with the informed reference system s and then convert the motion back to the ground reference.

Learning Objectives. Similar to the loss functions introduced for interaction diffusion, we denote two objectives as

$$\begin{aligned}\mathcal{L}_o &= \|\widehat{\boldsymbol{o}}^{1:H+F} - \boldsymbol{o}^{1:H+F}\|_2^2, \\ \mathcal{L}_{vo} &= \|\widehat{\boldsymbol{o}}^{2:H+F} - \widehat{\boldsymbol{o}}^{1:H+F-1}\|_2^2,\end{aligned} \tag{6}$$

where we denote the obtained object motion including the recovered past motion as $\widehat{\boldsymbol{o}}^{1:H+F}$, while the ground truth object motion is $\boldsymbol{o}^{1:H+F}$. We adopt the contact loss $\mathcal{L}_c$ to encourage body vertices and object vertices close to the object surface and body surface, respectively. And the penetration loss $\mathcal{L}_p$ employs the signed distances of human meshes to penalize mutual penetration between the object and human. For more details, please refer to [9]. Note that for skeletal representation, we do not apply $\mathcal{L}_c$ and $\mathcal{L}_p$.

C. Additional Details of Experimental Setup

Additional Implementation Details. For interaction diffusion, the weight of each loss term $(\lambda_h, \lambda_o, \lambda_{vh}, \lambda_{vo}) = (1, 0.1, 0.2, 0.02)$. For interaction prediction, the weight of each loss term $(\lambda_o, \lambda_{vo}, \lambda_c, \lambda_p) = (1, 0.1, 1, 0.1)$.

D. Additional Ablation Studies

Effect of the number of DCT bases. In Figure A, we compare the performance when different numbers of DCT bases are used for the interaction predictor. The results show that as the number of DCT bases increases, the translation error increases, while the rotation error decreases. The reason might be that rotation is more difficult to learn and requires more parameters. However, translation relative to the reference system is very easy to model. To balance the two errors, we choose the number as 10.

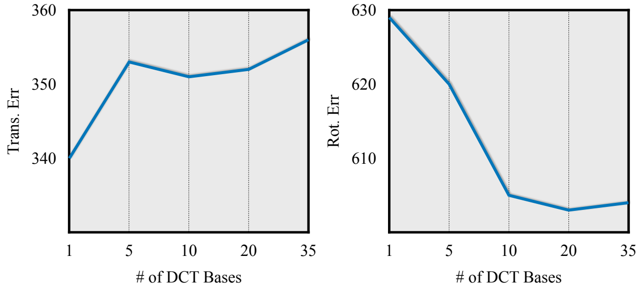

Figure A. **Ablation study** on the BEHAVE dataset [3]. We evaluate the long-term forecasting where we autoregressively generate 100 frames of future interactions. To balance the performance in predicting rotations and translations, we set the number of DCT bases to 10.

References

[1] Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 1974. 2

[2] Moab Arar, Ariel Shamir, and Amit H. Bermano. Learned queries for efficient local attention. In *CVPR*, 2022. 1

[3] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *CVPR*, 2022. 2

[4] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM transactions on graphics*, 2015. 1

[5] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *ICCV*, 2019. 2

[6] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017. 1

[7] Sigal Raab, Inbal Leibovitch, Guy Tevet, Moab Arar, Amit H Bermano, and Daniel Cohen-Or. Single motion diffusion. *arXiv preprint arXiv:2302.05905*, 2023. 1

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1

[9] Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. SAGA: Stochastic whole-body grasping with contact. In *ECCV*, 2022. 2

[10] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *CVPR*, 2021. 1