InterAct: Advancing Large-Scale Versatile 3D Human-Object Interaction Generation

Supplementary Material

We will release the complete dataset including the consolidated MoCap data and our synthetic data, as well as models for benchmarking tasks. In this supplementary material, we introduce: (**1**) an overview of the structure of our InterAct dataset in Sec. A, along with a **website featuring demo videos** of benchmark tasks, demonstrating how interaction correction and augmentation improve data quality and scale; (**2**) additional illustrations and details of our data curation, correction, and augmentation processes in Sec. B; (**3**) further implementation details and experimental results in Sec. C, which were omitted from the main paper due to space limitations; (**4**) licensing information in Sec. D; and (**5**) a discussion of limitations and potential negative societal impacts of our work in Sec. E.

A. Dataset

A.1. Structure

Each sequence folder contains the following files:

Motion Data. The human motion data (`human.npz`) contains SMPL parameters [62, 73]. We have standardized all data to a common global coordinate system and aligned ground heights for consistency. The object data (`object.npz`) includes the object names (indicating the corresponding point cloud file), as well as information on object angles and translations.

Interaction Representation. We convert the motion data into interaction representations, resulting in two files: `markers.npy`, containing the human representation, and `motion.npy`, which includes both human motion (represented by markers) and object motion (represented by BPS [68]).

Annotations. The annotations include multiple text descriptions (`text.txt`) and action labels (`action.txt`).

Visualizations. We also provide corresponding videos for the interaction data within each sequence folder for visual reference.

A.2. Webpage

Our website is at `https://sirui-xu.github.io/InterAct`. It includes demo videos comparing the raw data with our corrected and augmented versions, as well as showcasing generated results across six interaction generation tasks. These demos complement the quantitative results in the main paper, demonstrating that our unified multi-task model, trained on our extensive dataset, achieves state-of-the-art performance. We also compare our marker-based representation to joint position and rotation-based representations, demonstrating that markers as representative surface vertices are better suited for interaction modeling, since contact occurs on surfaces rather than joints.

B. Data Collection, Annotation, and Unification

B.1. Text and Action Annotation

In Sec. 3.1 of the main paper, we describe our process for manually annotating interaction sequences with detailed text instructions. Here, we elaborate on how we leverage large language models for automatic annotation augmentation, and generate additional action labels that facilitate further tasks.

We use GPT-4 [60] to rephrase, simplify, and generate action labels for the annotations we collected. Specifically, for rephrasing and simplification, we send two messages to the API: a system message containing the requirements and the annotation that needs to be processed. For example, the system message we use for the simplification task is:

> "I will provide a few sentences describing a human interacting with an object. Your task is to shorten the description while retaining its meaning, ensuring the object's name remains unchanged.

To generate action labels, we prompt GPT-4 to identify from 15 predefined action labels and categorize the annotations accordingly. Specifically, the prompt we use is:

> "I will provide a few sentences describing a human interacting with an object, and you need to select the single most fitting word to describe interactions from the following set: [Carry, Sit, Swing, Exercise, Rotate, Move, Hold, Drink, Eat, Play, Adjust, Lift, Kick, Pass, Manipulate]. Your response should be one word only."

We include some of our manually annotated action labels in the prompt to enhance in-context learning and response accuracy of the language model.

B.2. Processing of Each Sub-dataset

GRAB [77] provides full 3D body shape, pose, and 3D object pose data captured using MoCap markers. We utilize their SMPL-X [62] human annotation, downsample interaction sequences from 120 to 30 fps, and align the data to our unified global coordinate system and ground heights.

BEHAVE [4] contains HOI video frames captured from multi-view RGBD sequences. We use their SMPL-H [73] human annotation. We align their data into our unified global coordinate system and ground heights.

InterCap [28] contains HOI video frames captured from multi-view RGBD sequences. We use their SMPL-X [62] human annotation. We align their data into our unified global coordinate system and ground heights.

Chairs [29] contains HOI video frames captured from multi-view RGBD sequences. We only select sequences that contain rigid objects. We only select sequences that contain rigid objects. To correct the tilted human and object, we calculate the ground normal vector using the lowest point set of the human and object in each frame. We use their SMPL-X [62] human annotation, interpolating interaction sequences from 10 to 30 fps. We align their data into our unified global coordinate system and ground heights.

HODome [115] contains a total of HOI video frames captured from 76 viewpoints. The human body data are captured from these multi-view images using EasyMocap [2]. We further process these representations to standard SMPL-H [73] annotation, downsample interaction sequences from 60 to 30 fps, and align them into the same global coordinate system and ground heights.

OMOMO [38] contains object and human motion captured using a Vicon system comprised of 12 cameras controlled by Vicon Shogun. We use their SMPL-X [62] human annotation. We align their data into our unified global coordinate system and ground heights.

IMHD [123] contains video frames captured from both RGB cameras and the object-mounted Inertial Measurement Unit (IMU). We process their human representations to standard SMPL-H [73] annotation, downsample interaction sequences from 60 to 30 fps, and align them into the same global coordinate system and ground heights.

B.3. Interaction Correction

In this section, we provide the formulation or explanation of learning objectives for interaction correction and augmentation, which are omitted from Sec. 3.2 of the main paper due to space constraints.

Hand Correction. We define the hand poses as $\{\mathbf{hand}_i\}_{i=1}^{L}$ of arbitrary length L. The learning objectives are,

(**1**) *Penetration Loss*. Given the signed distance field of the human $\mathbf{sdf}_i$, we employ a penetration loss to penalize the body-object interpenetration,

$$E_{\text{pene}} = -\sum_{i=1}^{L}\sum_{d_o} \min(\mathbf{sdf}_i(\boldsymbol{v}_{\boldsymbol{o}_i}[k]), 0), \tag{1}$$

where $\boldsymbol{v}_{\boldsymbol{o}_i}[k]$ refers to the object vertex of index k at frame i.

(**2**) *Smooth Loss.* We employ a smoothing loss to avoid excessive speed and acceleration changes.

$$\begin{aligned} E_{\text{smooth}} = &\sum_{i=1}^{L-1} \|\mathbf{hand}_{i+1} - \mathbf{hand}_i\|_2^2 \\ &+ \sum_{i=1}^{L-2} \|(\mathbf{hand}_{i+2} - \mathbf{hand}_{i+1}) - (\mathbf{hand}_{i+1} - \mathbf{hand}_i)\|_2^2 \end{aligned} \tag{2}$$

where $\mathbf{hand}_i$ represents the hand pose at time step i

(**3**) *Prior Loss.* We apply a prior loss to maintain natural hand poses and prevent the hand joints from exceeding their range of motion (RoM) due to excessive guidance from the Contact Loss. We set the RoM constraints as $(\mathbf{hand}_{\max}, \mathbf{hand}_{\min})$ for all joint, derived from the statistical analysis of the GRAB [77] dataset. The loss is defined as,

$$\begin{aligned} E_{\text{prior}} = &\sum_{i=1}^{L} \| \min(\mathbf{hand}_i - \mathbf{hand}_{\min}, 0)\|_2^2 \\ &+ \| \max(\mathbf{hand}_i - \mathbf{hand}_{\max}, 0)\|_2^2 \end{aligned} \tag{3}$$

Hyperparameters of Contact Promotion. We include the formulation of the contact indicator c_i for contact promotion loss defined in Sec. 3.2 of the main paper.

$$c_i = \begin{cases} 1 & \min_j d_j[i] \le \epsilon \\ 0 & \min_j d_j[i] > \epsilon_2 \\ -12.5 \min_j d_j[i] + 1.25 & \text{otherwise,} \end{cases} \tag{4}$$

where $\min_j d_j[i]$ refers to hand-object chamfer distance, $\epsilon = 0.02$ indicates the contact threshold following [4], and $\epsilon_2 = 0.10$ indicates the non-contact threshold. The expression $-12.5 \min_j d_j[i] + 1.25$ provides the linear interpolation between these two phases.

Full-Body Correction. In full-body correction, we jointly optimize the full-body human pose $\{\boldsymbol{h}_i\}_{i=1}^{L}$ and the object pose $\{\boldsymbol{o}_i\}_{i=1}^{L}$, given the ground truth counterparts $\{\boldsymbol{h}_i^*\}_{i=1}^{L}$ and $\{\boldsymbol{o}_i^*\}_{i=1}^{L}$. The overall objective is defined as,

$$E = \lambda_{\text{pene}} E_{\text{pene}} + \lambda_{\text{smooth}} E_{\text{smooth}} + \lambda_{\text{rec}} E_{\text{rec}}, \tag{5}$$

where each component of the loss is defined as follows:

(**1**) *Penetration Loss.* Same as defined for hand correction.

(**2**) *Smooth Loss.* The smooth loss incorporates additional terms for object motion.

$$\begin{aligned} E_{\text{s}} = &\sum_{i=1}^{L-1} \|\boldsymbol{h}_{i+1} - \boldsymbol{h}_i\|_2^2 + + \sum_{i=1}^{L-1} \|\boldsymbol{o}_{i+1} - \boldsymbol{o}_i\|_2^2 \\ &+ \sum_{i=1}^{L-2} \|(\boldsymbol{h}_{i+2} - \boldsymbol{h}_{i+1}) - (\boldsymbol{h}_{i+1} - \boldsymbol{h}_i)\|_2^2 \\ &+ \sum_{i=1}^{L-2} \|(\boldsymbol{o}_{i+2} - \boldsymbol{o}_{i+1}) - (\boldsymbol{o}_{i+1} - \boldsymbol{o}_i)\|_2^2 \end{aligned} \tag{6}$$

(**3**) *Reconstruction Loss.* The reconstruction loss promote the optimized human and object pose to be close to the ground truth $\boldsymbol{h}_i^*$ and $\boldsymbol{o}_i^*$,

$$E_{\text{rec}} = \sum_{i=1}^{L} \|\boldsymbol{h}_i - \boldsymbol{h}_i^*\| + \|\boldsymbol{o}_i - \boldsymbol{o}_i^*\| \tag{7}$$

B.4. Interaction Augmentation

We align the human body to maintain interaction with the object to achieve contact invariance as we highlight in Sec. 3.2 of the main paper. In addition to the learning objective (**1**) E_{align}, introduced in the main paper to ensure contact consistency, we incorporate two additional objectives: (**2**) E_{reg}, a regularization term that penalizes excessive deviations from the original human body pose, with a particular focus on key body parts not in contact with the object; and (**3**) E_{smooth}, as defined in Sec. B.3, which promotes temporal smoothness between frames by restricting velocity and acceleration. The regularization term is derived from the reconstruction loss described in Sec. B.3, and is reformulated as follows:

$$E_{\text{reg}} = \beta \sum_{i=1}^{L} m_m \|\boldsymbol{h}_i - \boldsymbol{h}_i^*\| + \frac{1}{\beta} \sum_{i=1}^{L} \|\boldsymbol{h}_i - \boldsymbol{h}_i^*\|,$$

where m_m is a mask applied to joints that are not involved in the interaction, and $\beta = 5$ is selected to emphasize these vertices. The overall objective for augmentation is then defined as:

$$E_{\text{aug}} = E_{\text{align}} + E_{\text{reg}} + E_{\text{s}}, \tag{8}$$

where the optimization runs for 300 iterations to update the human body pose $\boldsymbol{h}$.

C. Additional Implementation Details and Experimental Analysis

C.1. Marker-Based Representation with Shape Variance

Our method uses the marker-based representation that couples pose and shape. Despite the coupled representation, our model can generate diverse human shapes during interaction generation. Although our task does not focus on specific human shape control and our text descriptions do not specify detailed characteristics like height or body type, the model inherently captures variability from the training data. This results in diverse human shapes, with heights varying from 1.71m to 1.81m across 50 batches of generation.

C.2. Text-Conditioned Interaction Generation

Additional Implementation Details. We split each subdataset into training and testing sets using a 9:1 ratio. Unlike HOI-Diff [63], which computes the Mean Squared Error (MSE) for all motion representations simultaneously, we calculate the loss for the human marker representation and the object motion representation separately. To balance these components, we assign the object motion loss and the contact label loss weights of 0.9 relative to the human motion loss. During training, the model generates 300 frames per sequence—longer sequences are cropped, shorter ones are zero-padded, and padded regions are masked during loss calculation. All experiments were conducted on a single NVIDIA A40 GPU over eight days.

Below, we describe how we apply guidance during the diffusion process, corresponding to the fourth model variant introduced in Sec. 5.2. Our key insight is to perform an additional calculation of object motion in a relative coordinates with respect to markers. By comparing the discrepancy L between the directly regressed object coordinates and those motion derived from human motion markers, we compute gradients to guide the object trajectories. The loss function L is weighted inversely by the distance between each marker and the object to emphasize the influence of closer body parts. We then compute the gradients of this loss with respect to the object's translation $\boldsymbol{o}$ and rotation $\boldsymbol{r}$, and the predicted means are updated during the final 30 iterations of diffusion denoising given 1000 iterations in total:

$$\hat{\boldsymbol{o}} = \hat{\boldsymbol{o}} - \tau_1 \frac{\partial L}{\partial \boldsymbol{o}}, \quad \hat{\boldsymbol{r}} = \hat{\boldsymbol{r}} - \tau_2 \frac{\partial L}{\partial \boldsymbol{r}},$$

where $\tau_1 = 0.1$, $\tau_2 = 0.2$ is selected. This guidance directs the diffusion model to generate object motions that consider the relationship with the human. Similar design choices can be found in [18, 63, 93, 102].

C.3. Action-Conditioned Interaction Generation

Additional Implementation Details. We rephrase the action label as "A person [action] the [object name]" and treat it similarly to text-conditioned generation, utilizing our interaction-aware text encoder to encode the text as introduced in Sec. 5.2 of the main paper. We uses the same data split, motion representation, and loss functions as the text-to-interaction task. Each experiment is conducted on a single NVIDIA A40 GPU over a duration of eight days.

C.4. Object-Conditioned Human Generation

Additional Implementation Details. We follow the dataset splitting strategy proposed in OMOMO [38]. Specifically, we divide HOI interactions into disjoint training and testing set based on object categories. As detailed in Table. A, the training set includes 168 objects, while the testing set consists of 29 unseen objects. Compared to baseline models, our multi-task model separately computes the human marker reconstruction loss and an additional feature reconstruction loss. The weight for the additional feature reconstruction

Dataset	Training Objects	Test Objects
behave	chairwood, keyboard, tablesquare, yogamat, boxmedium, suitcase, basketball, boxsmall, backpack, boxtiny, plasticcontainer, monitor, boxlong, stool, toolbox, chairblack	trashbin, boxlarge, tablesmall, yogaball
chairs	110, 162, 75, 64, 181, 156, 123, 111, 81, 45, 116, 33, 68, 60, 43, 130, 176, 158, 48, 59, 166, 96, 30, 87, 141, 44, 36, 103, 147, 149, 83, 154, 99, 104, 98, 85, 152, 180, 172, 109, 131, 157, 117, 92, 46, 151, 142, 49, 26, 29, 118, 171, 173, 168, 143, 121	15, 17, 24, 25
grab	toothpaste, spheremedium, cubesmall, table, cylindermedium, cubemedium, cubelarge, apple, duck, cubemiddle, wristwatch, waterbottle, flute, pyramidmedium, piggybank, banana, spheresmall, pyramidsmall, eyeglasses, coffeemug, cylindersmall, torussmall, flashlight, knife, stanfordbunny, pyramidlarge, rubberduck, camera, alarmclock, bowl, wineglass, headphones, cylinderlarge, hammer, stamp, torusmedium, toruslarge, hand, toothbrush, watch, doorknob, body, stapler, train, scissors, mug, elephant, lightbulb, fryingpan, gamecontroller, binoculars, airplane, mouse	phone, cup, teapot, spherelarge
imhd	broom, pan, baseball, dumbbell, kettlebell, suitcase	chair, skateboard, golf, tennis
intercap	skateboard, stool, racket, soccerball, fantabottle, suitcase	chair, toolbox, umbrella, cup
neuraldome	pillow, smallsofa, trolleycase, monitor, tennis, baseball, flower, chair, pan, case, table, badminton, pingpong, book, keyboard, trashcan, pink	bigsofa, box, talltable, desk
omomo	floorlamp, vacuum_bottom, mop_top, largetable, largebox, smallbox, vacuum, monitor, mop_bottom, plasticbox, vacuum_top, clothesstand, trashcan, woodchair	smalltable, whitechair, suitcase, tripod, mop

Table A. Data split for the task of object-conditioned human generation.

loss is set to half that of the human marker reconstruction loss. We adopt the same architectural design as the 1-stage and 2-stage baselines in [38]. Our multi-task model consists of four self-attention blocks, each with four attention heads. The dimensions for keys, queries, and values are all set to 256, and each layer produces a 512-dimensional output. The bps feature is configured with a dimension of 256. For training, we use a batch size of 64 over 300k iterations. All experiments are performed on a single NVIDIA A40 GPU, and training the multi-task model, the 1-stage baseline, and both components of the 2-stage baseline takes roughly 23 hours.

C.5. Human-Conditioned Object Generation

Additional Implementation Details. We split the dataset into training and testing sets using a 10:1 ratio. In our multi-task model, we define $\boldsymbol{\eta}$ as the distance between human markers and the object. Following the object-conditioned human generation setup, the additional feature reconstruction loss is weighted at 0.5 times that of the human marker reconstruction loss. Our model leverages the same transformer architecture and diffusion framework as the object-conditioned human generation model. For training, we use a batch size of 64 for 260k iterations, with each experiment running on a single NVIDIA A40 GPU and taking approximately 19 hours.

C.6. Interaction Prediction

Addtional Implementation Details. We use the same transformer architecture and diffusion framework as InterDiff [102]. We substitute the original SMPL representation in InterDiff with marker positions and adjust the corresponding loss function accordingly. As a result, the input and output representations now include marker positions, object angles, object translations, with object geometry as conditions for the diffusion process. Compared to InterDiff, We increase the loss weight for marker MSE by a factor of ten compared to the SMPL MSE loss. For both the InterAct and BEHAVE datasets, we split the data, using 90% for training and the remaining 10% of the BEHAVE dataset for testing. For each experiment, the training process is conducted on a single NVIDIA A40 GPU over a span of two days.

D. License

All data are shared under the CC BY-NC-SA (Attribution-NonCommercial-ShareAlike) license. We will also establish a GitHub repository https://github.com/wzyabcas/InterAct to receive user feedback on any

annotation errors. Users must review and follow the original licenses for each sub-dataset. Please find the licenses of corresponding assets in the code directories, and below is a summary of the licenses for the assets we have used:

1. GRAB [77] uses Software Copyright License for non-commercial scientific research purposes
2. BEHAVE [4] uses Software Copyright License for non-commercial scientific research purposes
3. InterCap [28] uses Software Copyright License for non-commercial scientific research purposes
4. Chairs [29] does not indicate their license but requires to follow the license of each sub-modules they used
5. HODome [115] uses Apache License
6. OMOMO [38] does not indicate their license
7. IMHD [123] uses Dataset Copyright License for Non-commercial Scientific Research Purposes

E. Discussion

Limitations. Despite that our dataset significantly expands the number of objects to 217 – nearly ten times more than existing HOI datasets – we acknowledge that it still possesses scale limitations. While our dataset advances the field by providing enriched annotations and supporting new tasks like text-to-HOI and action-to-HOI, it does not cover the full diversity of in-the-wild object categories encountered in real-world interactions. Although our experiments demonstrate that models trained on our dataset exhibit generalization to out-of-distribution objects within the dataset's scope, as presented in Table 6 and our webpage, achieving robust generalization to a broader range of unseen objects will require further expansion. Therefore, while our work makes a significant step toward larger-scale datasets, future efforts are needed to overcome these scale limitations and enhance model performance.

Another limitation of our method lies in the inherent challenges of denoising and correcting full-body HOI data, especially when dealing with significant noise in the original datasets. Issues like floating objects often stem from errors in the initial data rather than shortcomings of our approach. The severity of these errors can be too great for our correction process to handle effectively with the current hyper-parameter configuration, as large distances between the human and object may be identified as no contact and thus remain uncorrected. While adjusting hyper-parameters, such as employing a more aggressive contact threshold, can resolve specific issues like floating objects, we choose to apply unified hyper-parameters across all data for simplicity. Consequently, our method may not correct all artifacts in every scenario. Despite this, it effectively reduces noticeable noise and penetration issues, significantly improving the overall quality and plausibility of the motion data compared to the original.

Ethics Discussion and Potential Negative Societal Impact. We collect data on real behavioral information, which could potentially raise privacy concerns. And our correction and augmentation could be used to generate fake data and misleading information. However, we ensure that all collected and generated human data are processed into a format using SMPL [45] or markers, which significantly reduces identifying details compared to the raw data or images from their original own data. This processed representation effectively enhances privacy. Additionally, all annotations are gathered by the authors with participant consent. We meticulously review all augmented annotations generated by the language model to ensure they do not contain any harmful information or breach privacy.