HandDiffuse: Generative Controllers for Two-Hand Interactions via Diffusion Models

Pei Lin

ShanghaiTech University, School of Information Science and Technology, Shanghai, China
linpei2024@shanghaitech.edu.cn

Abstract

Existing hands datasets are largely short-range and the interaction is weak due to the self-occlusion and self-similarity of hands, which can not yet fit the need for interacting hands motion generation. To rescue the data scarcity, we propose HandDiffuse12.5M, a novel and real dataset that consists of temporal sequences with strong two-hand interactions. HandDiffuse12.5M has the largest scale and richest interactions among the existing two-hand datasets. We further present a strong baseline method HandDiffuse for the controllable motion generation of interacting hands using various controllers. Specifically, we apply the diffusion model as the backbone and design two motion representations for different controllers. To reduce artifacts, we also propose Interaction Loss which explicitly quantifies the dynamic interaction process. Our HandDiffuse enables various applications, i.e., motion in-betweening and trajectory controled generation. Experiments show that our method outperforms the state-of-the-art techniques in motion generation. The vivid two-hand motions generated by our method can also construct synthetic datasets and enhances the accuracy of existing hand motion capture algorithms.

Datasets — https://handdiffuse.github.io/

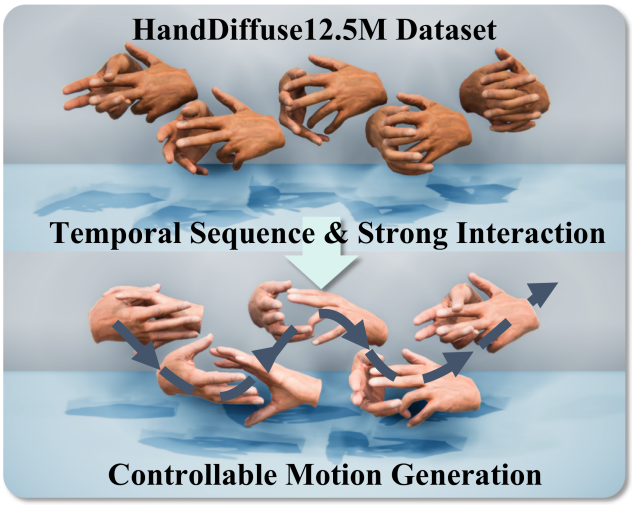

Figure 1: **Overview:** The proposed HandDiffuse12.5M benchmark dataset consists of temporal sequences with strong interaction. Based on it, we propose HandDiffuse, a strong baseline for the motion generation of interacting hands.

Introduction

The tightly interacting hands play a crucial role in human emotional expression and communication, serving as a vital component for physical interaction with oneself. Controllable hand motion generation not only enhances the experience of augmented reality/virtual reality (AR/VR) but also help robotics and avatars to express their emotion in another dimension.

So far, lots of works have emerged in the field of human body motion generation (Tevet et al. 2022b; Zhang et al. 2022a; Chen et al. 2023; Jiang et al. 2023), but the generation of temporal hands motions in strong interaction remains blank as the lack of strong interacting hands datasets. However, the dataset acquisition is time-consuming and expensive due to the interacting hands present self-occlusion, self-similarity, and complex articulations.

There are only a few datasets (Tzionas et al. 2016; Moon et al. 2020; Zuo et al. 2023; Li et al. 2023a; Moon et al. 2023) focusing on hands with strong interactions. Most of them, i,e, InterHand2.6M (Moon et al. 2020), Two-hand 500K (Zuo et al. 2023) and RenderIH (Li et al. 2023a), provides large-scale discrete frames but the available temporal sequences remain sparse. The severe lack of sufficient temporal motions makes them unsuitable for the motion generation task of two-hand interactions. Only recently, the concurrent work Re:InterHand (Moon et al. 2023) provides large-scale interaction data, yet it focuses on the motion capture of interacted hands under diverse lighting conditions.

To break the data scarcity, in this paper, we first contribute a real dataset, *HandDiffuse12.5M*, which focuses on capturing diverse temporal sequences with strong two-hand interactions. Our dataset consists of various modalities, including synchronous 50-view videos, 3D key points, and hand poses. It covers interacting hand motions in random and complex poses, daily communications, finger dances, and Chinese daily sign language with 250K temporal frames($\approx$ 2 hours), resulting in roughly 12.5M pictures. To address the issue

of occlusion, which is unavoidable in hand interactions, we manually annotate the key points in 240K images. Note that our HandDiffuse12.5M dataset owns the largest scale not only in the duration but also in the number of pictures among the existing available interacting hands datasets. Thus, it not only opens up the research for temporally consistent two-hand motion generation but also brings huge potential for future exploration in hand modeling.

Based on our novel dataset, we further provide a strong baseline algorithm named *HandDiffuse*, which tailors the diffusion model (Ho, Jain, and Abbeel 2020) to generate vivid interacting hands motions under diverse explicit control. Note that hand motion is an abstract form of emotional expression and ten fingers always gather together during interactions. Therefore, it is difficult to describe the motions with concrete text prompts and the control for motion generation needs to be explicit. To this end, we adopt several explicit controllers as the conditions for our diffusion models, including the discrete keyframes, the trajectories of wrists and 2D keypoints. These conditionings enable various downstream applications, ranging from hand motion in-betweening to trajectory-aided hand motion completion which connects hand motions with body motions (see Figure1), and can also assist in hand motion reconstruction. In HandDiffuse, we propose Interacting Hands DDIM which is inspired by modular design and come up with two specific motion representations for different controllers, as well as a novel Interaction Loss for the dynamic process of hand interaction, which helps to reduce the artifacts. Finally, we provide a thorough evaluation of our approach as well as a comparison with state-of-the-art motion generation methods (Tevet et al. 2022b; Liang et al. 2023; Lee et al. 2024; Müller et al. 2024). We also provide results to show that HandDiffuse can contribute to data augmentation for other datasets. Our preliminary outcomes indicate that controllable hand motion generation remains a challenging problem, and our dataset will consistently benefit further investigations of this new research direction. To summarize, our main contributions include:

- We contribute a large-scale real dataset named HandDiffuse12.5M, which provides accurate and temporally consistent tracking of human hands under diverse and strong two-hand interactions.
- We are the first to specifically target on hand-only interaction motion generation and we propose a strong baseline for generating interacting hand motions from various explicit controllers, enabling vivid motion in-betweening, trajectory conditioning and hands reconstruction.
- We introduce Interacting Hands DDIM, two motion representations as well as an Interaction Loss to significantly improve the generation results of two-hand interactions. The generated motions can contribute to data augmentation for other datasets.
- Our dataset, related codes will be made publicly available to stimulate further research.

Related Works

Hand Dataset. Although hand datasets have significantly accelerated the development of VR/AR, embodied AI and human-object interaction. Obtaining ground truth for specific dual-hand interaction motion from real-world captured data is challenging (Moon et al. 2020; Li et al. 2023a). Moon et al. (Moon et al. 2020) have attempted to combine manual and automated annotations, but manual annotations are labor-intensive. Recently, Moon et al. (Moon et al. 2023) have improved the detection accuracy of 2D keypoints and reconstructed hands to facilitate fully automated annotation schemes, but these methods are still time-consuming.

As a result, some works have established synthetic hand datasets to enhance annotation precision (Zuo et al. 2023; Li et al. 2023a; Zimmermann and Brox 2017; Lin, Wilhelm, and Martinez 2021; Gao et al. 2022). Among them, RenderIH (Li et al. 2023a) increases the diversity of data through re-lighting, while DART (Gao et al. 2022) provides a wide range of hand accessories to bridge the gap between synthetic and real data. However, synthetic datasets fail to balance the diversity of motion and the continuity of sequences. Some works (Li et al. 2023a; Zuo et al. 2023) sample single-frame actions for the left and right hands separately from real datasets and solve the occlusion problem between the hands through optimization. However, such optimization methods are not effective for temporal sequences. Therefore, very few papers (Tzionas et al. 2016; Moon et al. 2020, 2023) have collected temporal interacting hand datasets (Li et al. 2023a).

Hands Capture & Reconstruction. The reconstruction of both hands is a significant challenge in the area of human motion capture (Zuo et al. 2023). The release of the InterHand2.6M (Moon et al. 2020) has motivated many regression-based methods (Rong et al. 2021; Baowen et al. 2021; Li et al. 2022; Hampali et al. 2022; Di and Yu 2022; Fan et al. 2021; Kim, In Kim, and Baek 2021; Hao et al. 2022; Moon 2023). These regression-based methods significantly boost the development of VTuber and VR/AR devices due to their balance between real-time performance and accuracy. Some works (Li et al. 2022; Park et al. 2023) proposed a Transformer-based network with the cross-attention between right and left hands. Recently, Zuo et al. (Zuo et al. 2023) constructed a learning-based prior to capture dual-hands which achieved great results. Building on the impressive performance of diffusion models in image generation (Ronneberger, Fischer, and Brox 2015; Song, Meng, and Ermon 2020; Ho, Jain, and Abbeel 2020), DiffHand (Li et al. 2023b) combines hand reconstruction with diffusion models. The existing hands capture algorithms can also get enhanced by synthetic data. RenderIH(Li et al. 2023a) and DART(Gao et al. 2022) have shown that the synthetic hand datasets can improve the capture accuracy.

Human Motion generation. In recent years, there has been a growing interest in the field of human motion generation, with many studies exploring the generation of human motions based on conditioning signals. These conditioning signals are used to guide the generation of specific types of motions, such as motion in-betweening (Kaufmann

Dataset	Temporal Sequence	Source	# Views	Manually Annotation	Appearance
RenderIH (Li et al. 2023a)	No	Synthesis	-	No	Natural
Two-hand 500K (Zuo et al. 2023)	Partial	IMU+Synthesis	-	No	-
InterHand2.6M (Moon et al. 2020)	Partial	RGB	80	Yes	Lab
Re:InterHand (Moon et al. 2023)	Yes	RGB	26	No	Relighting
HandDiffuse12.5M(Ours)	Yes	RGB	50	Yes	Lab

Table 1: **Dataset Comparisons.** We compare our HandDiffuse12.5M dataset with other existing publicly available interacting-hand datasets. Our HandDiffuse12.5M dataset contains both strong interaction and temporal sequences, and has the largest scale.

et al. 2020; Harvey et al. 2020; Qin, Zheng, and Zhou 2022), trajectory control (Zhang et al. 2021; Rempe et al. 2023; Karunratanakul et al. 2023; Xie et al. 2023) and unconstrained motion synthesis (Pavlakos et al. 2019; Yan et al. 2019; Zhao, Su, and Ji 2020). Leading by the pioneering work of MDM (Tevet et al. 2023), human motion generation models based on diffusion architecture (Liang et al. 2023; Zhang et al. 2022b; Du et al. 2023) have been shown to gain better motion diversity and expressiveness compared to prior works based on GAN (Lin and Amer 2018) or VAE (Cai et al. 2021; Petrovich, Black, and Varol 2021). However, most existing works in this area have primarily focused on generating whole-body motions, only few works(Lee et al. 2024; Zhang et al. 2024) try to generate hands' motion in recent. Utilizing the capabilities of the diffusion models, we propose a novel method HandDiffuse to enable interacting hands motion generation tasks with control signals such as key frames and trajectory.

HandDiffuse12.5M Dataset

Capture system Existing hand motion capture solutions include IMU gloves, elastic sensor gloves, marker-based systems, and multi-view camera capture. However, due to the severe occlusion in hand interactions, marker-based systems are not suitable; IMU gloves and elastic sensor-based capture solutions theoretically address the issue of visual occlusion, but the IMU gloves are strict to calibration and it is easy to accidentally touch the sensor during the hands interaction; the elastic sensor-based capture solutions lacks the restoration of all degrees of freedom for finger joints. Therefore, we have opted for a multi-view RGB approach combined with manual annotation, similar to InterHand2.6M. Our capture system consists of 50 mounted video cameras recording interaction sequences at a frame rate of 30-60 frames per second (fps). We adjust the resolution to 3840 × 2160 to allocate more pixels for capturing hand movements. We design four categories of hand interaction motions for recording: random but complex hand interactions, daily communication, finger tutting dance, and sign language. Each motion sequence is controlled to last approximately one minute limited by the hardware. More detailed description about the capture system has been shown in Appendix.

2/3D joint coordinates & MANO fitting Upon securing multi-view video sequences, we utilize DWPose(Yang et al. 2023) which is the SOTA 2D keypoints detector to extract the hands 2D keypoints. The subsequent triangulation of these 2D keypoints yields 3D keypoints in world coordinate, which are then integrated into 3D pose by optimizing MANO (Romero, Tzionas, and Black 2017) which is differentiable. The frames with failed fitting will be selected out and manually re-annotated with 2D key points on the images. Figure2 has shown some MANO reprojection results which contains strong and various hands interaction. Due to space limitation, a more detailed procedure for annotating the dataset and the approach we used to fit MANO is illustrated in the Appendix.

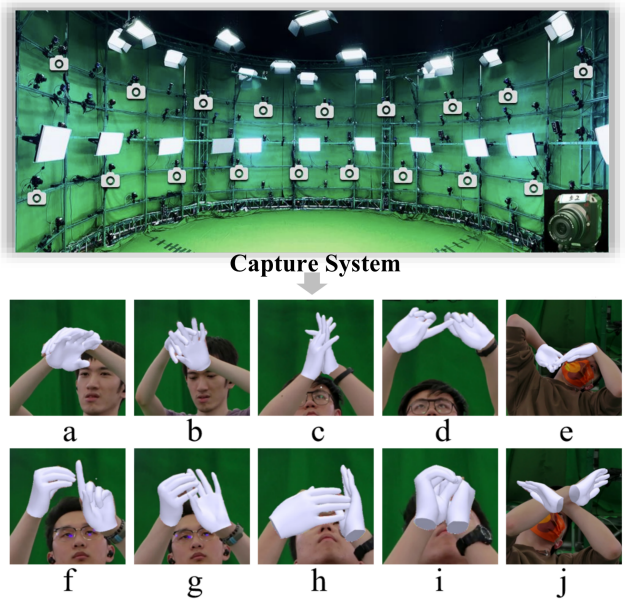

Figure 2: The capture system and reprojectoin of MANO. The proposed HandDiffuse12.5M benchmark dataset consists of strong and various interaction with accurate annotation.

Quantitative Evaluation for our Dataset. Our HandDiffuse12.5M dataset focuses on temporal sequences which has more than 250K temporal frames. This makes it the largest dataset among all of the interacting hands datasets as presented in Table 1. As shown in Figure 3, each sequence has around 1000 to 4000 temporal frames and the average is 2955 temporal frames per motion sequence, with the median of 3630 frames.

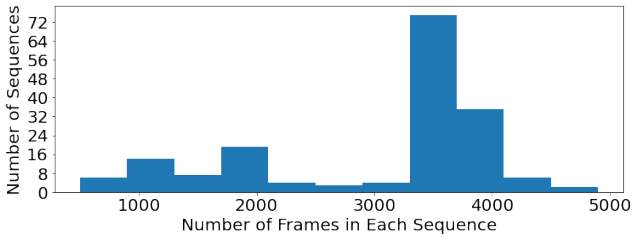

Figure 3: The distribution of HandDiffuse12.5M's temporal frames.

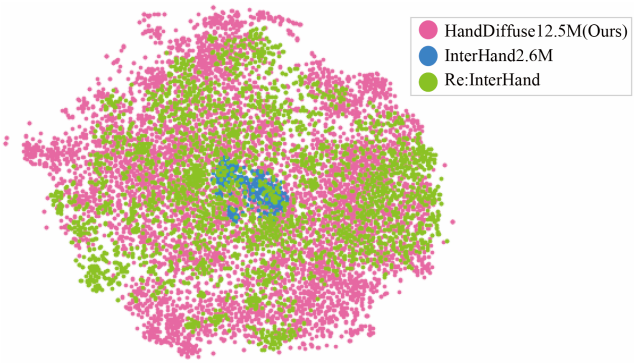

Figure 4: t-SNE visualization of HandDiffuse12.5M(Ours), InterHand2.6M and Re:InterHand. For InterHand2.6M, we only choose its temporal frames. The result indicates the diversity and richness of HandDiffuse12.5M.

Our dataset's average shortest distance between two hands' joints is 3.61cm, superior to InterHand2.6M's 4.04 cm and Re:InterHand's 3.84 cm. We classify a sample as "contacting" if the shortest distance between two hands' vertices is under 3 mm. With a contacting hands ratio of 57.3%, our dataset exceeds InterHand2.6M's 44.74% and Re:InterHand's 52.7%. We randomly selected 5000 images from the dataset to manually annotate the 2D keypoints. The error between the annotation results and the reprojection of 3D keypoints is 11.3 pixels per keypoint in a 3840×2160 image space. The MANO fitting error is 4.3mm per joint.

We also compare our dataset with InterHand2.6M (Moon et al. 2020) and Re:InterHand (Moon et al. 2023), both of which also contain hand interaction motion sequences, by visulizing the t-SNE results shown in Figure 4. Figure 4 illustrates that our dataset exhibits a greater variety of interaction poses compared to InterHand2.6M and Re:InterHand. This is attributed to our dataset's explicit focus on hand interaction sequences for generation tasks.

In summary, HandDiffuse12.5M comprises 12.5 million images and 250K temporal frames, making it the largest two-hand dataset to date. It provides a variety of strongly interactive motion and accurate annotations.

Method

An overview of HandDiffuse is shown in Figure 5. With HandDiffuse12.5M dataset, the goal is to synthesize two hands motion $\mathbf{x^{1:N}}$ without or with designed controllers c. These selected controllers are common in body motion generation and are also able to dictate the synthesis of interacting hands motions, such as key frames (motion inbetweening), trajectory of both wrists (trajectory control) and 2D keypoints(3D hands reconstruction). In contrast to single-person human motion generation, the task of generating interacting hands motion requires the network to learn not only the local motion of each hand but also the dynamic process of global interaction. Therefore, we adopts a modular design to learn the local pose of single hand and dynamic interacting process separately. We also propose two specific hand motion representations for different controllers and Interaction Loss ($\mathbf{Loss}_{interaction}$) to model this dynamic interacting process better.

Inspired by recent work, we adopt the diffusion model as the backbone which has shown remarkable generative ability in image generation and human motion generation. Diffuse is modeled as a Markov noising process and the ground truth $\mathbf{x_0^{1:N}}$ is diffused to $\mathbf{x_t^{1:N}}$ by adding t-step independent Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$. This process can be formulated as

$$q(\mathbf{x_t}|\mathbf{x_0}) = \sqrt{\overline{\alpha}_t}\mathbf{x_0} + (1 - \overline{\alpha}_t),$$

where $\overline{\alpha}_t = \prod_{s=0}^{t} \alpha_s$ and $\alpha_t = 1 - \beta_t$. β_t is the cosine noise variance schedule and $q(\mathbf{x_t})$ is near $\mathcal{N}(0, I)$ when t is big enough. Then, $\mathbf{x_t}$ is sent to a denoiser $\mathcal{G}$ conditioned on timestep t and different conditions c to predict $\hat{\mathbf{x}}_\mathbf{0}$:

$$\hat{\mathbf{x}}_\mathbf{0} = \mathcal{G}(\mathbf{x_t}, t, c).$$

Two Denoisers & Interacting Hands DDIM HandDiffuse adopts a modular design and applies two denoisers: Single Hand Denoiser(**SHDe**) and Interacting Hands Denoiser(**IHDe**) to focus on the different parts of the whole interacting process as shown in Figure 5.

SHDe is used to learn the characteristics of local hand motion, so we set the global translation and global rotation of the hand to 0, and extract the local joints' positions $(J_{local}, \dot{J}_{local})$ and pose(θ) as input. By mirroring the right hand to the left hand, we only trained one SHDe. By minimizing the

$$\mathbf{Loss}_{reconSH} = \|L_{GT} - \hat{L}_0\|_2, \quad (1)$$

where $\hat{L}_0$ is the denoised single hand parameter and L_{GT} is the ground truth, we model the local motion of single hand.

After training the SHDe and freezing its parameters, we concatenate the generated local motions of two hands with the global information which is t-step noise and transform the whole vector to designed motion representation. IHDe focuses on the global information and will also finetune the local motion. During the training stage, after completing a round of denoising, we minimize

$$\mathbf{Loss}_{IHDe} = \mathbf{Loss}_{reconDH} + \mathbf{Loss}_{interaction}, \quad (2)$$

$$\mathbf{Loss}_{reconDH} = \|x_{GT} - \hat{x}_0\|_2, \quad (3)$$

where $\hat{x}_0$ is the denoised two hands parameter and x_{GT} is the ground truth, and $\mathbf{Loss}_{interaction}$ will be illustrated later. In the inference stage, after denoising, we re-add the

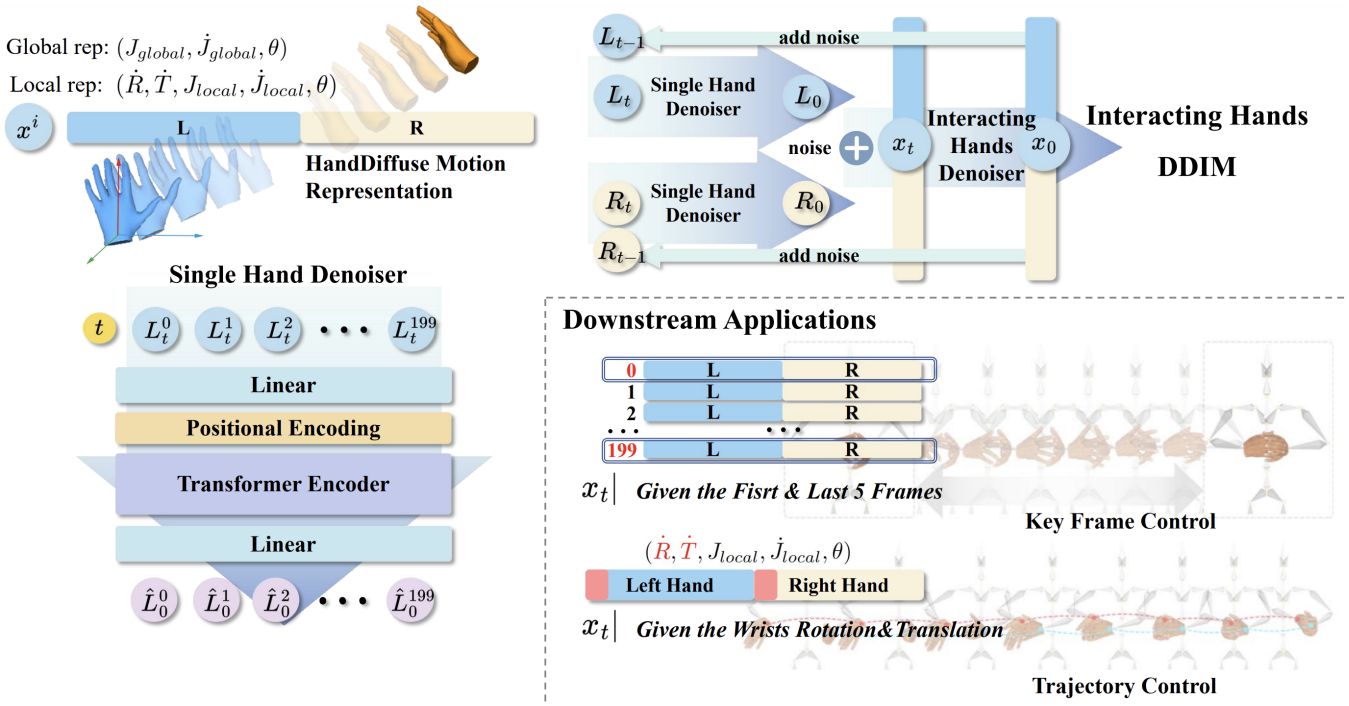

Figure 5: **(Left) Overview of HandDiffuse** We first generate the motions for each hand separately by training Single Hand Denoiser which only focuses on local poses. The generated two single hands' local pose are concatenated with global information(noise) and transformed into designed motion representation. The interacting hands denoiser further optimizes the interaction process. **(Right) Overview of Downstream Applications.** When the control is key frames, we generate the final motions $\hat{\mathbf{x}}_\mathbf{0}^{\mathbf{1:N}}$ by giving the first 5 frames $\mathbf{x}_\mathbf{GT}^{\mathbf{:5}}$ and the last 5 frames $\mathbf{x}_\mathbf{GT}^{\mathbf{-5:}}$ before denoising. When the control os wrists' trajectories, we generate the final motion $\hat{\mathbf{x}}_\mathbf{0}$ by giving the root angular velocity and linear velocity in each frame before denoising. Other controls like 2D keypoints has been illustrated in Appendix due to the space limitation.

noise of $t-1$ steps. We refer to this process as Interacting Hands DDIM. Subsequent evaluations have proven the excellence of the modular design.

In terms of model architecture, we adopted an encoder-only transformer similar to MDM(Tevet et al. 2022b). SHDe and IHDe have the same structure but different dimensions.

Motion Representations for Interacting Hands. For different tasks, we have proposed two specific motion representations, which we refer to as local representation (local rep) and global representation (global rep).

The local rep is inspired by HumanML3D(Guo et al. 2022), a widely used approach in human motion generation. In each frame, the pose of left hand is defined by a tuple of $(\dot{R}, \dot{T}, J_{local}, \dot{J}_{local}, \theta)$ and the right hand is $(R_{init}, T_{init}, \dot{R}, \dot{T}, J_{local}, \dot{J}_{local}, \theta)$, where $\dot{R} \in \mathrm{R}^4$ and $\dot{T} \in \mathrm{R}^3$ represent root angular velocity in quaternion and linear velocity respectively. $J_{local} \in \mathrm{R}^{3j}$ and $\dot{J}_{local} \in \mathrm{R}^{3j}$ denote the local joint positions and velocities in root space, with $j = 21$ represents the number of joints. $\theta \in \mathrm{R}^{4i}$ represents the local joint rotations in quaternion representation, following the skeleton structure of MANO with $i = 15$. Each sequence is aligned with the left wrist of the first frame as the coordinate system origin. Both left and right hands are rotated to align the left-hand R vector with the direction (1, 0, 0) which is shown in 5. Therefore, the right hand has $R_{init} \in \mathrm{R}^4$ and $T_{init} \in \mathrm{R}^3$.

The global rep is defined as a tuple of (J_{global}, $\dot{J}_{global}$, θ) for both hands, where J_{global} represents the positions of hand joints in the world coordinate system and two hands are normalized by employing the same approach as local rep.

During the experimental process, we observed that the local rep performed better in the task where the global rotation and translation of the hand are given. The global rep achieved better results in the unconditional generation, the control of key frames(motion in-betweening).

Interaction Loss $\mathbf{Loss}_{interaction}$ The first part of the $\mathbf{Loss}_{interaction}$ is the contact potential loss($\mathbf{Loss}_{potential}$). In each frame, we first calculate a 25×25 distance matrix $\mathbf{D}$ between the 21 left-hand joints and 21 right-hand joints with 4 more joints sampled in each palm. Since the interaction between the hands involves fine-grained movements, we should focus on situations where the finger joints are very close to each other. Inspired by the concept of Contact Potential Field(Yang et al. 2021), we further represent the distances using the elastic potential energy of a spring model and it can be described as

$$\mathbf{P} = 1/2\mathbf{K}\ relu(\tau - \|\mathbf{D}_{pred}\|_2)^2, \tag{4}$$

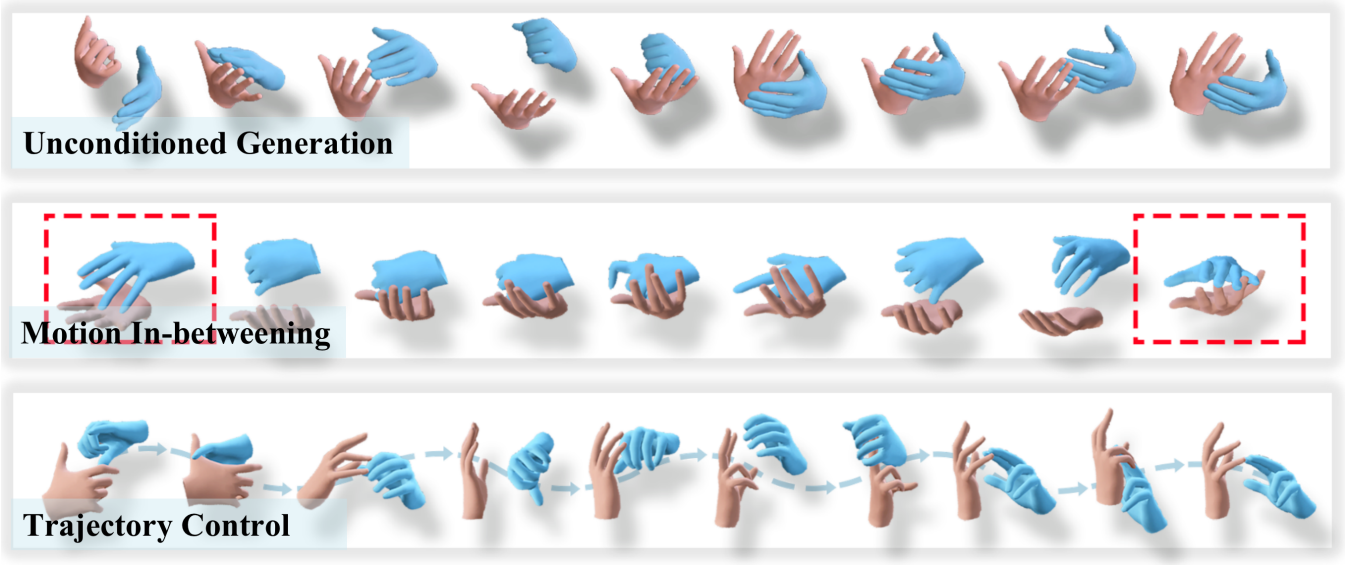

Figure 6: **Two Downstream Applications of HandDiffuse.** The generation results of downstream applications of HandDiffuse.

where $\mathbf{K}$ represents the coefficient of elasticity and τ denotes the distance threshold. The closer the distance between the finger joints, the larger the potential energy obtained, and the distance will be filtered out when it is over the threshold.

Moreover, by considering the direction of the distance, we can avoid ambiguity caused by penetration, and the final contact potential loss is described as

$$\mathbf{Loss}_{potential} = |\mathbf{P}_{pred} - \mathbf{P}_{gt}|(1 + \overrightarrow{\mathbf{D}_{pred}} \times \overrightarrow{\mathbf{D}_{gt}}), \quad (5)$$

where $(\times)$ means cross product. More is shown in Appendix.

We observed that MANO has strict requirements for the bone length. Therefore, the second part of the $\mathbf{Loss}_{interaction}$ is the Shape Loss($\mathbf{Loss}_{shape}$). We ensure the rationality of hand shape by calculating the differences in bone lengths between the left and right hands, as well as the differences between the predicted shape and the ground truth (GT). This is described as follows:

$$\mathbf{Loss}_{shape} = |\mathbf{BL}_{left} - \mathbf{BL}_{right}| + |\mathbf{BL}_{pred} - \mathbf{BL}_{GT}|, \quad (6)$$

where BL denotes the bone length.

Downstream Applications For the task of key frames control, we provide the ground truth for the previous five frames $\mathbf{x}^{:5}_{\mathbf{GT}}$ and the last five frames $\mathbf{x}^{-5:}_{\mathbf{GT}}$ for the diffused motion $\mathbf{x}^{\mathbf{1:N}}_{\mathbf{t}}$. For the task of trajectory control, the goal is to generate the reasonable local motion given the global rotation and translation of both hands. In this case, we provide $\dot{R}^{1:N}$ and $\dot{T}^{1:N}$ as control signal for the diffused motion $\mathbf{x}^{\mathbf{1:N}}_{\mathbf{t}}$. The results are presented in Figure 6.

We also explore HandDiffuse by applying the control of 2D keypoints, and the experiments have shown that HandDiffuse can also solve the task of hands reconstruction. Due to the space limitation, the results are presented in Appendix.

Experiment

In this section, we conduct several experiments to evaluate our method HandDiffuse for the task of interacting hands motion generation. We particularly evaluate (1) the comparison of our method against previous state-of-the-art approaches, and (2) the ablation study. The above evaluations are conducted specifically on the two controllers (key frames and trajectory) and other controller(2D keypoints) is evaluated in Appendix due to the space limitation. More experiments on the ability of HandDiffuse to augment existing datasets are presented in Appendix. The models have been trained with $T = 1000$ diffusing steps and a cosine noise schedule. The denoiser has 8 layers, 4 heads and the feed-forward dimension is 1024. We set the length of generated motion N to 200 in all experiments. All of them have been trained on a single *NVIDIA GeForce RTX 3090* GPU for about two days.

Comparisons of different methods We compare our approach with MDM (Tevet et al. 2023)(single human motion generation), InterGen (Liang et al. 2023)(interacting human motion generation), BUDDI(Müller et al. 2024)(latent representation of interacting human) and InterHandGen(Liang et al. 2023)(interacting hands motion generation of single frame). The metrics we applied are common used Fréchet Inception Distance (FID), Diversity and sdf loss(SDF). FID can measure the realistic of the generated motion. Diversity can prove that the motion do not maintain the mean pose for the sake of preserving realistic. SDF would take the positive value while zero otherwise, when the hand joints penetrate with each other. Using the formulation, we can measure the severity of penetration during the whole motion. The strategy we applied to compute these metrics is inspired by MotionCLIP(Tevet et al. 2022a) and is illustated in Appendix. We modify these methods by adapting their motion representation to enable hand motion generation. Specifically,

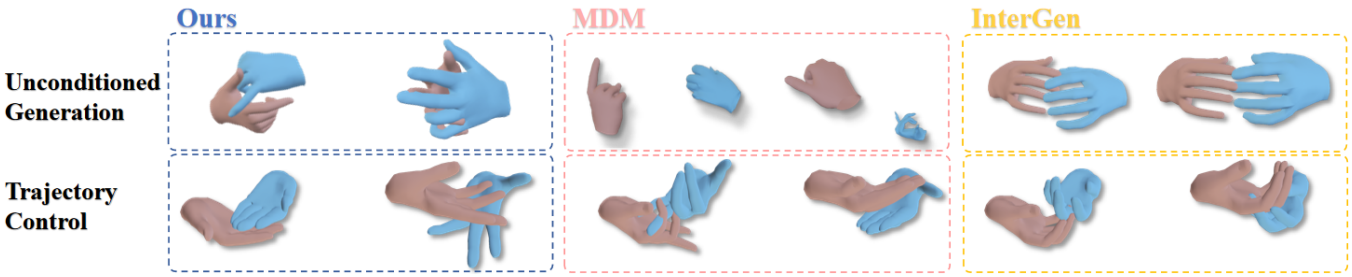

Figure 7: **Qualitative comparison for different methods.** For Unconstrained Generation, MDM tend to generate motion with hands shifting away from each other, while InterGen would generate motions with a mean pose. In terms of Trajectory Control, MDM yields satisfactory results, whereas InterGen generates motion without incorporating interaction.

Method	Unconditioned			In-betweening			Trajectory Control		
	FID↓	Diversity↑	SDF↓	FID↓	Diversity↑	SDF↓	FID↓	Diversity↑	SDF↓
real	0.050	12.034	1.146						
MDM(Tevet et al. 2023)	0.526	10.366	1.74	0.616	10.258	1.621	0.205	11.265	1.572
InterGen(Liang et al. 2023)	0.251	4.928	**0**	**0.269**	5.385	**1.032**	0.402	9.102	3.572
BUDDI(Müller et al. 2024)	0.853	8.282	2.293	1.211	7.341	2.482	0.921	7.497	1.842
InterHandGen(Lee et al. 2024)	0.437	10.238	1.731	0.734	8.620	1.819	0.539	10.124	1.282
HandDiffuse(ours)	**0.161**	**11.829**	1.553	0.273	**11.554**	1.562	**0.173**	**11.758**	**1.141**

Table 2: **Quantitative comparisons** of different methods. We compare the **FID** and **Diversity** for different methods. '↓' indicates that results are better with lower metrics, ↑ is on the contrary. **Bold** means the best result.

Task	Unconditioned		In-betweening		Trajectory Control	
	FID↓	SDF↓	FID↓	SDF↓	FID↓	SDF↓
real	0.050	1.146	-	-	-	-
w Global rep.	**0.161**	**1.553**	**0.273**	**1.562**	0.418	1.347
w Local rep.	0.431	1.962	0.567	1.971	**0.173**	**1.141**
w/o SHDe	0.168	1.575	0.282	1.571	0.182	1.538
w/o $Loss_{potential}$	0.171	1.606	0.317	1.628	0.210	1.614
w/o $Loss_{shape}$	0.390	1.583	0.512	1.579	0.380	1.245
Complete model	**0.161**	**1.553**	**0.273**	**1.562**	**0.173**	**1.141**

Table 3: **Quantitative comparisons.** "w/o SHDe" means IHDe is the only denoiser. **SDF** is only calculated on frames with penetraion.

we adapted InterHandGen to generate continuous frames following its original pipeline in which we first generated 200 frames of left-hand motion, then generated the right-hand motion based on the left hand. We adapted BUDDI for continuous frame generation and treat the motion of each frame as a token, and retained the "Person Embed" component to distinguish different hands. As depicted in Figure 7, both MDM and InterGen exhibit limitations when generating hand motion. They tend to produce motion sequences with severe joint penetration, unnatural interaction or static mean pose. The resulst of HandDiffuse has the highest diversity as shown in Table 2. Because the motion generated by InterGen without condition is static, the FID is lower than us and the penetration does not exist.

Ablation Study At first, we demonstrate the effectiveness of the different motion representations in different tasks in Table 3. The qualitative result is shown in Appendix. By incorporating the $\mathbf{Loss}_{interaction}$, we mitigate the issue of joint penetration, resulting in improved motion generation quality. Table 3 demonstrates that our Interacting Hands DDIM with modular deisgn achieves the highest performance when using the proper motion representation and $\mathbf{Loss}_{interaction}$.

Conclusion

We present HandDiffuse12.5M, the largest interacting hands dataset with strong interaction and temporal sequences, to open up the research direction of interacting hands motion generation. We propose a strong baseline method HandDiffuse which is the first interacting hands motion generation approach along with some useful controllers: key frames control, trajectory control and 2D keypoints control. Our evaluation in multi-stage demonstrates the benefit of our approach againist others and HandDiffuse fills the blank in human motion generation.

Acknowledgments

We would like to express our sincere gratitude to all the students and faculty members of the VDI Laboratory at ShanghaiTech University for their invaluable assistance and guidance throughout this experiment. We also appreciate the use of the filming equipment provided by the VDI Laboratory.

Additionally, we would like to acknowledge our collaborators: Sihang Xu, Hongdi Yang, Yiran Liu, Xin Chen, Jingya Wang, Jingyi Yu, and Lan Xu, for their contributions to this work.

References

Baowen, Z.; Yangang, W.; Xiaoming, D.; Yinda, Z.; Ping, T.; Cuixia, M.; and Hongan, W. 2021. Interacting Two-Hand 3D Pose and Shape Reconstruction from Single Color Image. In *International Conference on Computer Vision (ICCV)*.

Cai, Y.; Wang, Y.; Zhu, Y.; Cham, T.-J.; Cai, J.; Yuan, J.; Liu, J.; Zheng, C.; Yan, S.; Ding, H.; et al. 2021. A unified 3d human motion synthesis model via conditional variational auto-encoder. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11645–11655.

Chen, X.; Jiang, B.; Liu, W.; Huang, Z.; Fu, B.; Chen, T.; and Yu, G. 2023. Executing your Commands via Motion Diffusion in Latent Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18000–18010.

Di, X.; and Yu, P. 2022. LWA-HAND: Lightweight Attention Hand for Interacting Hand Reconstruction. arXiv:2208.09815.

Du, Y.; Kips, R.; Pumarola, A.; Starke, S.; Thabet, A.; and Sanakoyeu, A. 2023. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 481–490.

Fan, Z.; Spurr, A.; Kocabas, M.; Tang, S.; Black, M.; and Hilliges, O. 2021. Learning to Disambiguate Strongly Interacting Hands via Probabilistic Per-pixel Part Segmentation. In *International Conference on 3D Vision (3DV)*.

Gao, D.; Xiu, Y.; Li, K.; Yang, L.; Wang, F.; Zhang, P.; Zhang, B.; Lu, C.; and Tan, P. 2022. DART: Articulated Hand Model with Diverse Accessories and Rich Textures. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022. Generating Diverse and Natural 3D Human Motions From Text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5152–5161.

Hampali, S.; Sarkar, S. D.; Rad, M.; and Lepetit, V. 2022. Keypoint Transformer: Solving Joint Identification in Challenging Hands and Object Interactions for Accurate 3D Pose Estimation. In *IEEE Computer Vision and Pattern Recognition Conference*.

Hao, M.; Sheng, J.; Wentao, L.; Chen, Q.; Mengxiang, L.; Wanli, O.; and Ping, L. 2022. 3D Interacting Hand Pose Estimation by Hand De-occlusion and Removal. In *European Conference on Computer Vision (ECCV)*.

Harvey, F. G.; Yurick, M.; Nowrouzezahrai, D.; and Pal, C. 2020. Robust motion in-betweening. *ACM Transactions on Graphics*, 39(4).

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Jiang, B.; Chen, X.; Liu, W.; Yu, J.; Yu, G.; and Chen, T. 2023. MotionGPT: Human Motion as a Foreign Language. *arXiv preprint arXiv:2306.14795*.

Karunratanakul, K.; Preechakul, K.; Suwajanakorn, S.; and Tang, S. 2023. GMD: Controllable Human Motion Synthesis via Guided Diffusion Models. *arXiv preprint arXiv:2305.12577*.

Kaufmann, M.; Aksan, E.; Song, J.; Pece, F.; Ziegler, R.; and Hilliges, O. 2020. Convolutional Autoencoders for Human Motion Infilling. In *2020 International Conference on 3D Vision (3DV)*, 918–927.

Kim, D. U.; In Kim, K.; and Baek, S. 2021. End-to-End Detection and Pose Estimation of Two Interacting Hands. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 11169–11178.

Lee, J.; Saito, S.; Nam, G.; Sung, M.; and Kim, T.-K. 2024. InterHandGen: Two-Hand Interaction Generation via Cascaded Reverse Diffusion. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, L.; Tian, L.; Zhang, X.; Wang, Q.; Zhang, B.; Liu, M.; and Chen, C. 2023a. RenderIH: A Large-scale Synthetic Dataset for 3D Interacting Hand Pose Estimation. *arXiv preprint arXiv:2309.09301*.

Li, L.; Zhuo, L.; Zhang, B.; Bo, L.; and Chen, C. 2023b. DiffHand: End-to-End Hand Mesh Reconstruction via Diffusion Models. arXiv:2305.13705.

Li, M.; An, L.; Zhang, H.; Wu, L.; Chen, F.; Yu, T.; and Liu, Y. 2022. Interacting Attention Graph for Single Image Two-Hand Reconstruction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2751–2760.

Liang, H.; Zhang, W.; Li, W.; Yu, J.; and Xu, L. 2023. InterGen: Diffusion-based Multi-human Motion Generation under Complex Interactions. *arXiv preprint arXiv:2304.05684*.

Lin, F.; Wilhelm, C.; and Martinez, T. 2021. Two-Hand Global 3D Pose Estimation Using Monocular RGB. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2373–2381.

Lin, X.; and Amer, M. R. 2018. Human motion modeling using dvgans. *arXiv preprint arXiv:1804.10652*.

Moon, G. 2023. Bringing Inputs to Shared Domains for 3D Interacting Hands Recovery in the Wild. In *CVPR*.

Moon, G.; Saito, S.; Xu, W.; Joshi, R.; Buffalini, J.; Bellan, H.; Rosen, N.; Richardson, J.; Mallorie, M.; Bree, P.; Simon, T.; Peng, B.; Garg, S.; McPhail, K.; and Shiratori, T. 2023. A Dataset of Relighted 3D Interacting Hands. In *NeurIPS Track on Datasets and Benchmarks*.

Moon, G.; Yu, S.-I.; Wen, H.; Shiratori, T.; and Lee, K. M. 2020. InterHand2.6M: A Dataset and Baseline for 3D Interacting Hand Pose Estimation from a Single RGB Image. In *European Conference on Computer Vision (ECCV)*.

Müller, L.; Ye, V.; Pavlakos, G.; Black, M. J.; and Kanazawa, A. 2024. Generative Proxemics: A Prior for 3D Social Interaction from Images. *Computer Vision and Pattern Recognition (CVPR)*.

Park, J.; Jung, D. S.; Moon, G.; and Lee, K. M. 2023. Extract-and-Adaptation Network for 3D Interacting Hand Mesh Recovery. arXiv:2309.01943.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A. A.; Tzionas, D.; and Black, M. J. 2019. Expressive Body Capture: 3D Hands, Face, and Body From a Single Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Petrovich, M.; Black, M. J.; and Varol, G. 2021. Action-conditioned 3D human motion synthesis with transformer VAE. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10985–10995.

Qin, J.; Zheng, Y.; and Zhou, K. 2022. Motion in-betweening via two-stage transformers. *ACM Transactions on Graphics (TOG)*, 41(6): 1–16.

Rempe, D.; Luo, Z.; Bin Peng, X.; Yuan, Y.; Kitani, K.; Kreis, K.; Fidler, S.; and Litany, O. 2023. Trace and Pace: Controllable Pedestrian Animation via Guided Trajectory Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13756–13766.

Romero, J.; Tzionas, D.; and Black, M. J. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6).

Rong, Y.; Wang, J.; Liu, Z.; and Loy, C. C. 2021. Monocular 3D Reconstruction of Interacting Handsvia Collision-Aware Factorized Refinements. In *International Conference on 3D Vision*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Tevet, G.; Gordon, B.; Hertz, A.; Bermano, A. H.; and Cohen-Or, D. 2022a. MotionCLIP: Exposing Human Motion Generation to CLIP Space. *arXiv preprint arXiv:2203.08063*.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Bermano, A. H.; and Cohen-Or, D. 2022b. Human Motion Diffusion Model. *arXiv preprint arXiv:2209.14916*.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-or, D.; and Bermano, A. H. 2023. Human Motion Diffusion Model. In *The Eleventh International Conference on Learning Representations*.

Tzionas, D.; Ballan, L.; Srikantha, A.; Aponte, P.; Pollefeys, M.; and Gall, J. 2016. Capturing Hands in Action using Discriminative Salient Points and Physics Simulation. *International Journal of Computer Vision (IJCV)*.

Xie, Y.; Jampani, V.; Zhong, L.; Sun, D.; and Jiang, H. 2023. OmniControl: Control Any Joint at Any Time for Human Motion Generation. *arXiv preprint arXiv:2310.08580*.

Yan, S.; Li, Z.; Xiong, Y.; Yan, H.; and Lin, D. 2019. Convolutional sequence generation for skeleton-based action synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4394–4402.

Yang, L.; Zhan, X.; Li, K.; Xu, W.; Li, J.; and Lu, C. 2021. CPF: Learning a Contact Potential Field to Model the Hand-Object Interaction. In *ICCV*.

Yang, Z.; Zeng, A.; Yuan, C.; and Li, Y. 2023. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4210–4220.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2022a. MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model. *arXiv preprint arXiv:2208.15001*.

Zhang, S.; Ma, Q.; Zhang, Y.; Qian, Z.; Kwon, T.; Pollefeys, M.; Bogo, F.; and Tang, S. 2022b. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European Conference on Computer Vision*, 180–200. Springer.

Zhang, S.; Zhang, Y.; Bogo, F.; Pollefeys, M.; and Tang, S. 2021. Learning Motion Priors for 4D Human Body Capture in 3D Scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 11343–11353.

Zhang, W.; Huang, M.; Zhou, Y.; Zhang, J.; Yu, J.; Wang, J.; and Xu, L. 2024. BOTH2Hands: Inferring 3D Hands from Both Text Prompts and Body Dynamics. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhao, R.; Su, H.; and Ji, Q. 2020. Bayesian Adversarial Human Motion Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zimmermann, C.; and Brox, T. 2017. Learning to Estimate 3D Hand Pose from Single RGB Images. Technical report, arXiv:1705.01389. Https://arxiv.org/abs/1705.01389.

Zuo, B.; Zhao, Z.; Sun, W.; Xie, W.; Xue, Z.; and Wang, Y. 2023. Reconstructing Interacting Hands with Interaction Prior from Monocular Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.