InteractFusion: Synthesizing Interactive Motions with Noise-Decoupled Diffusion Models

Junyi Zhang
Shanghai Jiao Tong University
junyizhang@sjtu.edu.cn

Hongchi Xia
Shanghai Jiao Tong University
xiahongchi@sjtu.edu.cn

Abstract

Human motion modeling is very important for many modern graphic applications, especially in the virtual reality scene. Creating the real human motion effect plays a crucial role in the construction of the whole virtual scene. However, the traditional human motion modeling method has a high threshold and is not easy for more people to master and use. In order to eliminate such expertise barriers and allow virtual reality technology to be developed more widely and deeply, recent research has proposed that motion generation can be carried out through natural language. However, it remains challenging to achieve diverse and fine-grained motion generation with various text inputs. In this paper, we propose a method to solve this problem. We take advantage of the diffusion model, a powerful tool proposed by recent studies which has a powerful ability to understand natural language and generate visual effects, to achieve interactive human motion modeling through natural language as a medium.

Keywords: Motion modeling, Diffusion model

1 Introduction

Human motion modeling is a critical component of animating virtual characters to imitate vivid and rich human movements, which has been a vital topic for many applications, such as film-making, game development, and virtual YouTuber animation. To mimic human motions, virtual characters should be capable of moving around naturally, reacting to environmental stimuli, and meanwhile expressing sophisticated emotions. Despite decades of exciting technological breakthroughs, it requires sophisticated equipment (e.g., expensive motion capture systems) and domain experts to produce lively and authentic body movements. In order to remove skill prerequisites for layman users and potentially scale to the mass audience, it is vital to create a versatile human motion generation model that could produce diverse, easily manipulable motion sequences.

Overcoming these proficiency barriers and facilitating the widespread and extensive development of virtual reality technology has become a focal point of recent research endeavors. One such proposal suggests that motion generation can be accomplished through natural language, thereby eliminating the need for specialized expertise. However, effectively achieving diverse and nuanced motion generation based on a range of textual inputs remains a significant challenge that researchers continue to tackle. The quest for generating motion that is both varied and intricately detailed in response to different text descriptions persists as an active area of exploration.

To tackle this problem, we note that Diffusion Models, a type of generative models, which have been gaining significant popularity in the past several years. We first warp the interaction of humans and objects with specific representations and feed into diffusion pipeline, which applies Gaussian noises to it and then denoise to recover the interactions. We train the diffusion by predicting the noise by a designed neural network which takes the motions, text and objects information as inputs. What's more, we propose an extra temporal and interaction shared noise in diffusion models, which improves the previous pipeline that only blindly adds noise and overlook the consistency constraints.

Overall, we propose our method InteractDiffuse which can successfully generate interactions among humans and objects.

2 Related Works

2.1 Diffusion Models

This paper proposes a new motion generation pipeline based on the Denoising Diffusion Probabilistic Model (DDPM) [1]. One of the principal advantages of DDPM is that the formation of the original motion sequence can be retained. It means that we can easily apply more constraints during the denoising process. In the later sections, we will explore more potential of DDPM in different types of conditions. Besides, benefiting from this nature, DDPM can generate more diverse samples.

2.2 Conditional Intent Generation

The increasing maturity of various generative models stimulates researchers' enthusiasm to study conditional motion generation. Text-driven intent generation can be regarded as learning a joint embedding of text feature space and intent feature space. [2] proposes an auto-regressive pipeline. It first encodes language descriptions into features and then auto-regressively generates motion frames conditioned on the text features. However, this method is hard to capture the global relation due to the auto-regressive scheme. Moreover, the generation quality is inferior. Instead, [3] softly fuses text features into generation and can yield the whole sequence simultaneously.

3 Method

3.1 Representation of interactiveness motion

3.1.1 Representation for Human-Object Interaction

Human-object interaction refers to the dynamic relationship between humans and objects in a given context or environment. It encompasses the various ways in which humans interact, manipulate, and engage with objects to achieve specific goals or perform tasks. To generate reasonable interaction motion between human and object, we need better represenions for them. Here, following [4], we represent the human mesh using the SMPL-X parametric body model. SMPL-X parametrizes the full human body along with the hands and the face as a differentiable function SMPLX(β, r, ϕ, t), consisting of body shape parameters $\beta \in \mathbb{R}^{10}$, the root translation $t \in \mathbb{R}^3$, the axis-angle rotations for the body joints $r \in \mathbb{R}^{J\times 3}$ (J = 55), and the face expression parameters $\phi \in \mathbb{R}^{10}$. It maps the parameters to a body mesh with 10,475 vertices. To improve the stability and the convergence characteristics of our model, we use the 6D continuous representations $\theta \in \mathbb{R}^{J\times 6}$to represent body joint rotations. We downsample all the objects in the dataset to 300 vertices for faster optimization. For object pose representation, we follow the traditional 6d pose representation. The object's 6-DOF pose is represented using a rotation matrix $\mathbf{R} \in \mathbb{R}^9$ and a translation vector $\mathbf{T} \in \mathbb{R}^3$.

3.1.2 Representation for Hand-Object Interaction

Hand-object interaction refers to the dynamic relationship between a human hand and an object during manipulation or engagement. It involves the intricate coordination of the hand's movements, grasp, and manipulation actions to interact with and control objects in the surrounding environment. To generate reasonable interaction motion between hand and object, we need to design great representations. Here we adopt the solution from [5], and follow the MANO pose parameters which is composed of 48 dimensions. That is, initial hand pose $H_0 \in R^{51}$ represented by the 48-dimensional MANO pose parameters. For object representation, in addition to the 6D poses we described before, we simplify the mesh structure in representation by sampling 1024 vertices and symbolize them by vanilla X-Y-Z 3D coordinates.

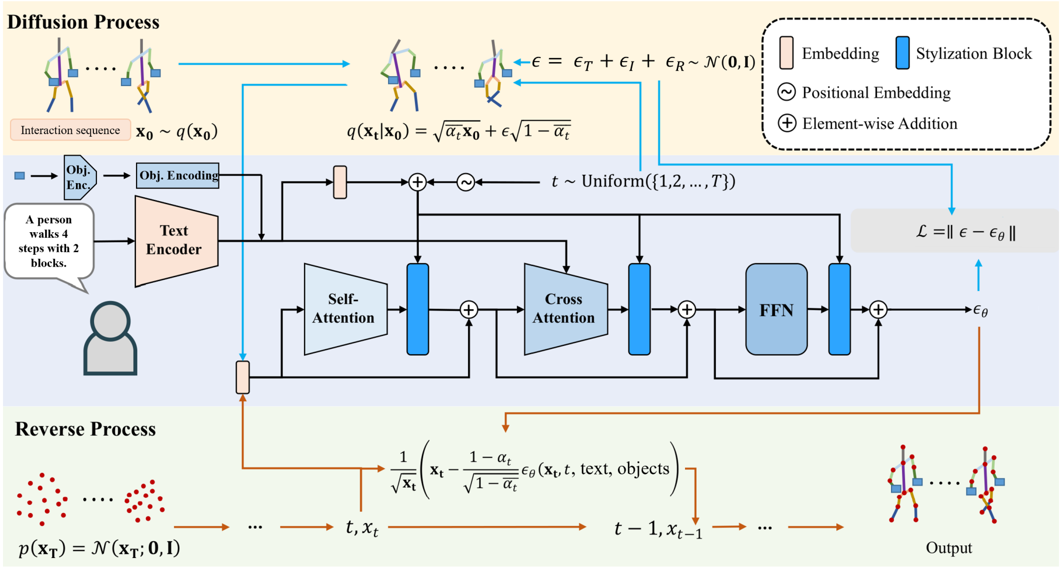

Diffusion Process
Interaction sequence
$\mathbf{x_0} \sim q(\mathbf{x_0})$
$q(\mathbf{x_t}|\mathbf{x_0}) = \sqrt{\overline{\alpha_t}\mathbf{x_0}} + \epsilon\sqrt{1-\overline{\alpha_t}}$
$\epsilon = \epsilon_T + \epsilon_I + \epsilon_R \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
Embedding
Stylization Block
Positional Embedding
Element-wise Addition
Obj. Enc.
Obj. Encoding
$t \sim \text{Uniform}(\{1,2,\dots,T\})$
A person walks 4 steps with 2 blocks.
Text Encoder
$\mathcal{L} = \| \epsilon - \epsilon_\theta \|$
Self-Attention
Cross Attention
FFN
ϵ_θ
Reverse Process
$\frac{1}{\sqrt{\mathbf{x_t}}}\left(\mathbf{x_t} - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha_t}}}\epsilon_\theta(\mathbf{x_t}, t, \text{text, objects}\right)$
$p(\mathbf{x_T}) = \mathcal{N}(\mathbf{x_T}; \mathbf{0}, \mathbf{I})$
...
t, x_t
$t-1, x_{t-1}$
...
Output

Figure 1: **InteractDiffuse method overview.**

3.1.3 Representation for Human-Human Interaction

As for Human-Human Interaction, which basically describe the dynamic exchange, communication, and collaboration between individuals. It encompasses the diverse ways in which humans interact with one another. Here we mainly focus on the human motion when interactions happen. Following [6], we treat human as skeleton model in our representation, which is composed of k skeletons and the action goes through totally T frames. We represent the action as a sequence $P = \{P_1, P_2, ..., P_T\}$, where $P_t \in R^{3\times d}$ and $d = 3 \times k$. To be more specific, $P_t = [J_1(t), ..., J_k(t)]$ and $J_i(t) = [x_i(t), y_i(t), z_i(t)]$. The goal is to generate a reaction $Y = \{Y_1, Y_2, ..., Y_T\}$ a sequence of skeleton poses from $X = \{X_1, X_2, ..., X_T\}$ a sequence representing the action motion.

3.1.4 Overall Representations

Overall, we represents the interactions by the combination of three components: Human-Object, Hand-Object and Human-Human Interaction. To be more specific, an interaction happens among humans H and objects O, which can be represented as $I = \{I_0, I_1, ..., I_T\}$, where $I_i = (I_{i,H}, I_{i,O})$ with $I_{i,H} = \{I_{i,h}|h \in H\}$ and $I_{i,O} = \{I_{i,o}|o \in O\}$. $I_{i,h} = (P_i, \text{SMPLX}(\beta_i, r_i, \phi_i, t_i), H_i)$ represents the human action at t frame and $I_{i,o} = (\mathbf{R}, \mathbf{T})$ represents the states of objects at t frame.

3.2 Interaction Diffusion Model

Following the literature on the diffusion model in the image synthesis field, we first build up a text-conditioned motion generation pipeline using a denoising diffusion probabilistic model (DDPM). This model is the basis of our proposed method. For the denoising process, we follow the Cross-Modality Linear Transformer [3] to process input sequences conditioned on the given text prompts and objects information. Beyond the direct application of text-driven motion generation, we take one step further to explore methods that are conditioned on interaction representation during denoising. Specifically, we experiment with Temporal Shared Noise in our proposed InteractDiffuse. We decouple the noise into three parts: one shared among temporal dimension, one shared among interacting instances, and one independent noise, to enhance the performance. The overall pipeline is shown in 1. We introduce each part of this architecture in the following subsections.

3.2.1 Preliminary of DDPM

We build our text-driven motion generation pipeline based on denoising diffusion probabilistic model (DDPM), or diffusion models. A probabilistic model is learned to gradually denoises a Gaussian noise to generate a target output, such as a 2D image or 3D point cloud. Formally, diffusion models are formulated as $p_\theta(x_0) := \int p_\theta(x_{0:T})dx_{1:T}$, where $x_0 \sim q(x_0)$ is the real data, and $x_1, x_2, ..., x_T$ are the latent data. They generally have a diffusion process and a reverse process. To approximate posterior $q(x_{1:T}|x_0)$, the diffusion process follows a Markov chain to gradually add Gaussian noise to the data until its distribution is close to the latent distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, according to variance schedules given by β_t. Instead of repeatedly adding noises on x0, we follow other works to formulate the diffusion process as $q(x_t|x_0) = \sqrt{\bar{\alpha}_t}x_0 + \epsilon\sqrt{1-\bar{\alpha}_t}$ with $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Hence, we can simply sample a noise ϵ and then directly generate x_t by this formulation. Instead of predicting x_{t-1}, here we predict the noise term ϵ.

3.2.2 Intent-Driven Interaction

We model the intent-driven motion generation as a conditional diffusion model, with intent as the condition. The pipeline overview can be found in Fig. 1. To be more specific, we adopt a neural network $\epsilon_\theta(x_t, t, \text{text}, \text{objects})$, which is essential for denoising steps. Previous works mainly utilize UNet-like structure as the denoising model. However, the target motion sequences are variable-length in the motion generation task, making convolution-based networks unsuitable. Therefore, we follow the Cross-Modality Linear Transformer, as shown in Fig. 1. Similar to the machine translation task, our proposed model includes a text encoder and a motion decoder. To meet the requirement of the diffusion models, we further customize each layer of the motion decoder. We introduce our design for $\epsilon_\theta(x_t, t, \text{text}, \text{objects})$ part by part.

Diffusion Model We adopt the mainstream diffusion model design and predict the noise term ϵ to denoise. We optimize the model parameters to decrease a mean squared error as

$$\mathcal{L} = E_{t\in[1,T],x_0\sim q(x_0),\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[||\epsilon - \epsilon_\theta(x_t, t, \text{text}, \text{objects})||] \tag{1}$$

Then we can denoise the motion sequence by estimating $\Sigma(x_t, t, \text{text}, \text{objects})$ step by step and finally get a clean motion sequence, which is conditioned on the given text and objects information.

Text Encoder Here we directly use classical transformer to extract text features. To enhance the generalization ability, we use parameter weights in CLIP to initialize the first several layers. This part of the parameters is frozen and will not be optimized in the later training.

Objects Encoder We represent objects in our interaction as meshes and sample vertices from them. Then we integrate the objects meshes vertices and faces with 6D object pose into our object encoder and generate corresponding encoding, prepared for being fed into the nerual network ϵ_θ.

Linear Self-attention and Cross-attention To simplify calculation of attention blocks in classical Transformer blocks, here we adopt the improvement to the traditional self attention and cross attention block, which boosts the time complextity to linear time. To be more specific, instead of calculating pair-wise attention weights, efficient attention generates global feature map $F_g \in R^{d_k \times d_k}$, where d_k is the dimension of feature after multi-head split. Then Cross-attention replaces **X** in **K** and **V** calculation by the text feature.

With these components, we build up the basic InteractDiffuse system and it has already been with the ability to generate good interactions between humans and objects. However, with the improvements we propose in the next section, we can generate better interactions.

3.2.3 Temporal and Interaction Shared Noise

We propose the temporal shared noise to take into account the 1) temporal consistency and 2) interaction consistency, two nature of the interactiveness motion, in designing the interaction diffusion model. That is to say, unlike the traditional diffusion models that blindly apply noise into the input data, we decouple the noise into three parts: one shared among temporal dimension, one shared among interacting instances, and one independent noise, to enhance the performance.

To be more specific, we decouple the noise ϵ into three parts: ϵ_T, ϵ_I and ϵ_R. To satisfy the requirements of diffusion model, we adopt extra techniques to maintain the distribution of the total noise following the standard normal distribution. ϵ_T is the shared noise throughout the noise addition procedure to force the temporal consistency. ϵ_I is shared noise of every interaction type, which is beneficial to keep interaction consistency. Finally a random ϵ_R which follows normal distribution is applied. This process can be found in the pipeline Fig. 1.

3.3 Rendering the Results

For rendering results, we split it to three sub tasks: human-objects interactions rendering, hand-objects interactions rendering and human-human interactions rendering.

We consider every interaction as a sequence of motions of meshes, represented as SMPL-X parameters and skeletons for human, MANO pose parameters for hands and 6D pose for objects with corresponding meshes.

To render results, we first fix the position of cameras. Then after loading the generated interactions from InteractMotion, we transform the representations of SMPL-X parameters, skeletons, MANO pose parameters and 6D pose to the positions of vertices in the corresponding meshes. For more diverse and continuous interactions, we interpolates among key frames to extend the interactions. Finally, we render every frame separately and incorporate them together to videos.

4 Experiment

We split the interactions into three categories: Human-Objects Interactions, Hand-Objects Interactions and Human-Human Interactions. We use different datasets and render interactions results for datasets respectively.

4.1 Human-Objects Interactions

For Human-Objects Interactions, we choose Grab dataset [7], a dataset of Whole-Body human grasping of objects, consisting of 1.3K sequences of human-object interactions exhibiting multiple intents. We can take texts and objects information as input and generates interactions from our model. We report three types of interactions: Eating Banana, Flying Planes, and Drinking from Bowl. Specifically, we report different generated results of Drinking from Bowl from the same inputs. The results can be found in Fig. 2. We choose four key frames from the generated video. From the key frames photos, we can find that our InteractMotion can generate reasonable motions and interactions of human and objects. What's more, it can support various types of interactions and generate interactions with high variance.

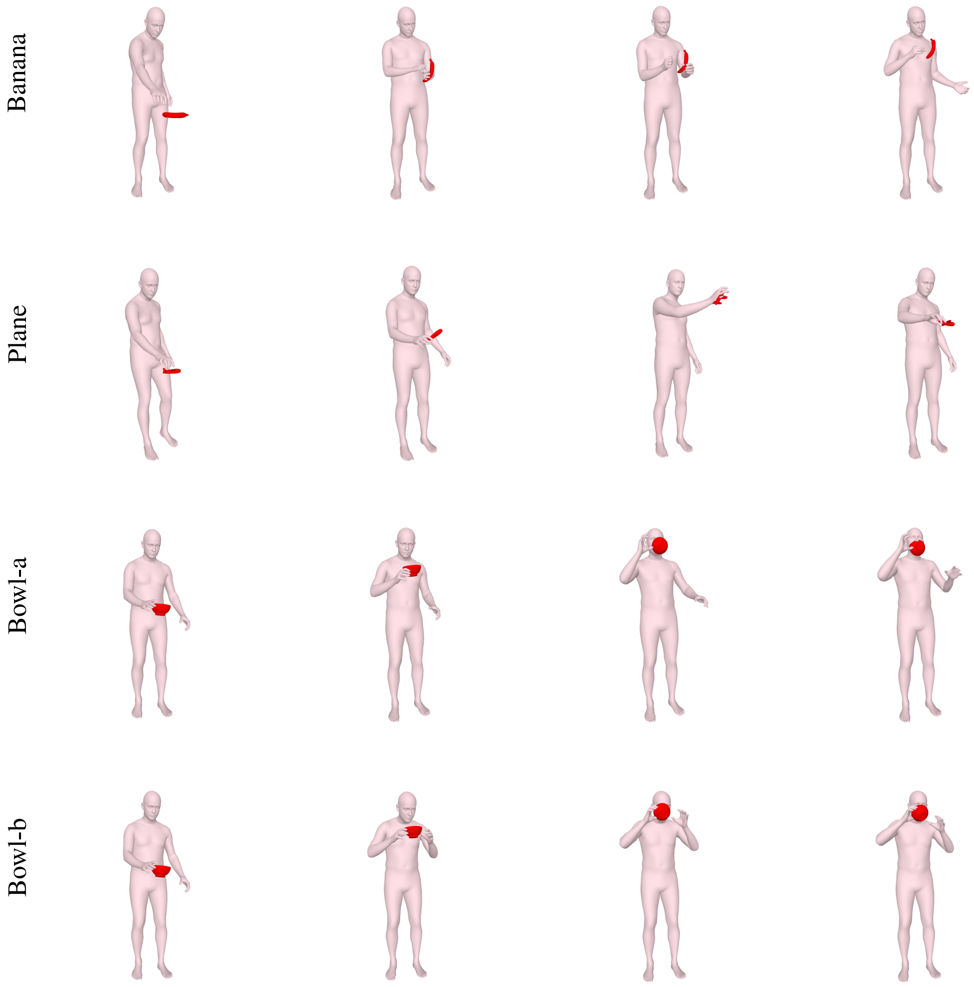
4.2 Hand-Objects Interactions

For Hand-Objects Interactions, we choose HOI4D dataset [8], a dataset concerning human object interactions in 4D spatial-temporal space as a real-world dataset that contains dynamic HOM data spanning various rigid and articulated object categories, but we focus on hands here. We can take texts and objects information as input and generates interactions from our model. We report two types of interactions: open a scissor and close a scissor. The results can be found in Fig. 3. We choose four key frames from the generated video. We can find that in the frames we show in the table, the hand-object interactions are quite reasonable and realistic. It proves that our model can generate hand-objects interactions with high quality.

4.3 Human-Human Interactions

For Human-Human Interactions, we choose K3HI Dataset [9] and SBU dataset [10], two datasets concerning human and human interactions in the representations of skeletons. SBU dataset contains 8 classes of simple interaction motions: walking toward, walking away, kicking, pushing, shaking hands, hugging, exchanging, and punching while K3HI contains the same 8 classes as SBU aside from the "hugging" class which is replaced by "pointing". We can only take texts as input and generates interactions from our model. We report four types of interactions: pointing, pushing, exchanging and kicking. The results can be found in Fig. 4. We choose two key frames from the generated video. We can see that in the representation of skeletons, the interactions among humans are realistic and with high variance. It shows that our InteractMotion support various human-human interactions.

Figure 2: **Running results of InteractMotion with Grab Dataset** We report three types of interactions: Eating Banana, Flying Planes, and Drinking from Bowl.



5 Conclusion

In this paper, we focus on human motion modeling, which is crucial for animating virtual characters to mimic realistic and expressive movements. To address this problem, we utilize Diffusion Models, a popular type of generative models. Our approach involves warping the interactions between humans and objects using specific representations, which are then fed into a diffusion pipeline. In this pipeline, Gaussian noises are applied to the representations and then denoised to recover the interactions. We train the diffusion model by predicting the noise using a neural network that takes into account motion, text, and object information. Additionally, we propose incorporating an additional temporal and interaction shared noise into the diffusion models, which improves upon the previous method that blindly adds noise and ignores consistency constraints. In summary, our method, InteractDiffuse, effectively generates interactions between humans and objects.

Figure 3: **Running results of InteractMotion with HOI4D Dataset** We report two types of interactions: Opening the scissor and Closing the scissor.

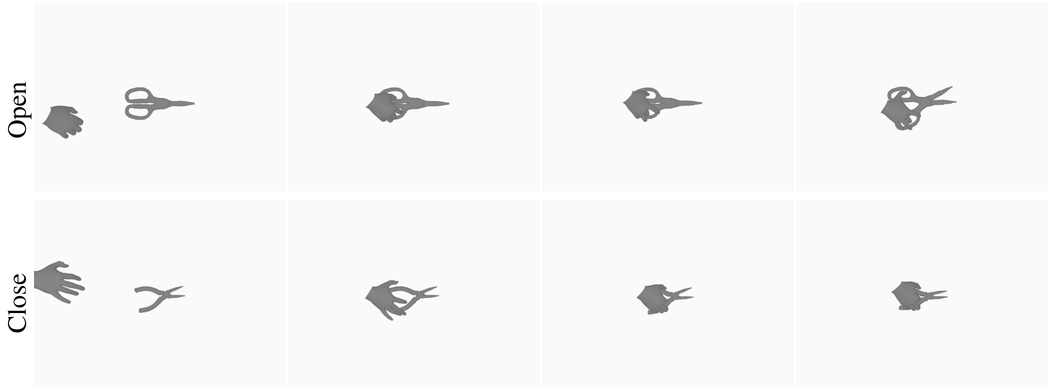
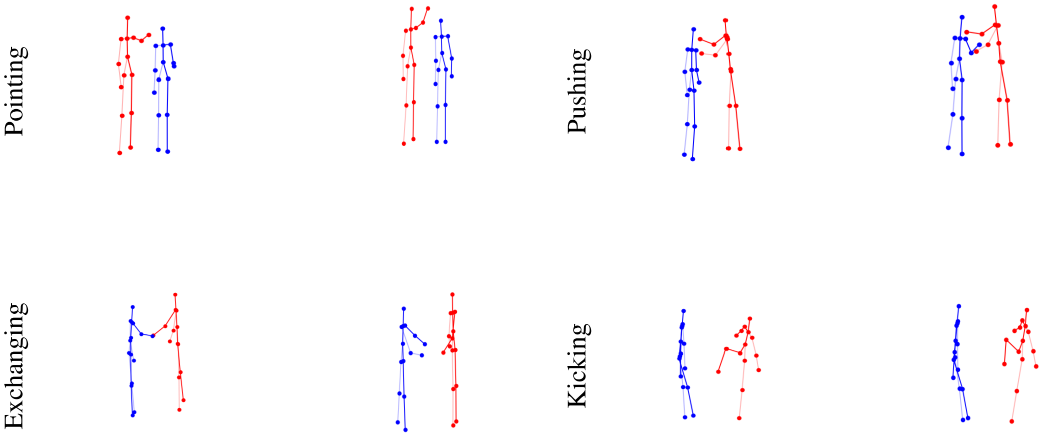


Figure 4: **Running results of InteractMotion with K3HI and SBU Dataset** We report four types of interactions: Pointing, Pushing, Exchanging and Kicking.



References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[2] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, June 2022.

[3] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.

[4] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions, 2023.

[5] Juntian Zheng, Qingyuan Zheng, Lixing Fang, Yun Liu, and Li Yi. Cams: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis, 2023.

[6] Baptiste Chopin, Hao Tang, Naima Otberdout, Mohamed Daoudi, and Nicu Sebe. Interaction transformer for human reaction generation, 2023.

[7] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020.

[8] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21013–21022, June 2022.

[9] Saman Nikzad and Hossein Ebrahimnezhad. Two-person interaction recognition from bilateral silhouette of key poses. *Journal of Ambient Intelligence and Smart Environments*, 9:483–499, 06 2017.

[10] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L. Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–35, 2012.