BOTH2Hands: Inferring 3D Hands from Both Text Prompts and Body Dynamics

Wenqian Zhang, Molin Huang, Yuxuan Zhou, Juze Zhang, Jingyi Yu, Jingya Wang, Lan Xu
ShanghaiTech University
{zhangwq2022,huangml,zhouyx2,zhangjz,yujingyi,wangjingya,xulan1}@shanghaitech.edu.cn

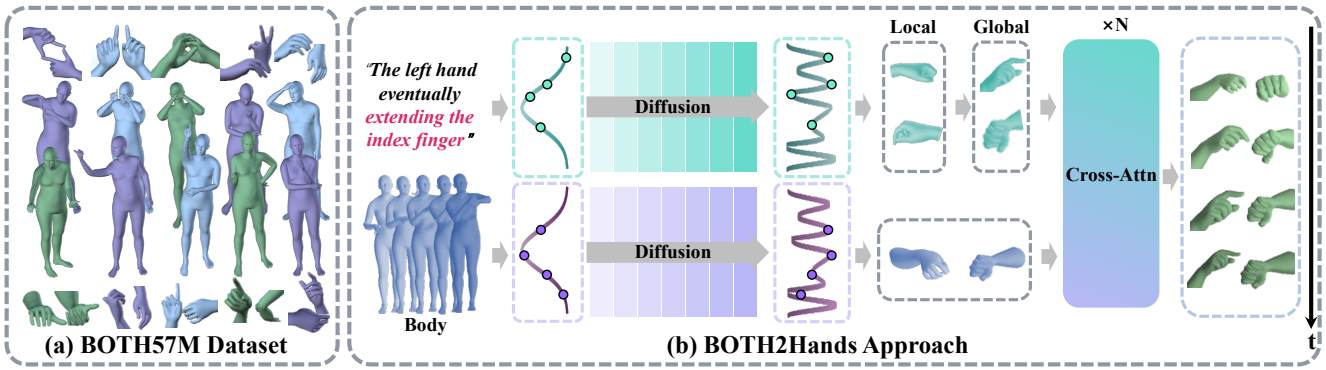

Figure 1. (a) Our BOTH57M dataset contains rich gestures and body movements. (b) BOTH2Hands is the only model that can handle text prompts and body dynamics as input, generating realistic hand motions at present.

Abstract

The recently emerging text-to-motion advances have inspired numerous attempts for convenient and interactive human motion generation. Yet, existing methods are largely limited to generating body motions only without considering the rich two-hand motions, let alone handling various conditions like body dynamics or texts. To break the data bottleneck, we propose BOTH57M, a novel multi-modal dataset for two-hand motion generation. Our dataset includes accurate motion tracking for the human body and hands and provides pair-wised finger-level hand annotations and body descriptions. We further provide a strong baseline method, BOTH2Hands, for the novel task: generating vivid two-hand motions from both implicit body dynamics and explicit text prompts. We first warm up two parallel body-to-hand and text-to-hand diffusion models and then utilize the cross-attention transformer for motion blending. Extensive experiments and cross-validations demonstrate the effectiveness of our approach and dataset for generating convincing two-hand motions from the hybrid body-and-textual conditions. Our dataset and code will be released to the community for future research, which can be found at github.

1. Introduction

The recent years have witnessed the tremendous progress of human motion generation, especially for the recently emerging text-to-motion setting [6, 61, 62, 74]. It enables novices to conveniently generate desired motions in a natural interactive manner. Yet, realistic human motions require the generation of companion motions of hands. Actually, we humans tend to incorporate a wide variety of hand motions with body movements in our daily communications.

The recent text-to-motion advances [8, 29, 62, 74] mostly focus on generating body motions only. In contrast, the convenient generation of two-hand motions from text prompts has significantly fallen behind, mainly due to severe data scarcity. The wider adopted text-motion datasets [14, 38] embrace limited two-hand motions and corresponding textual annotations. Only recently, the concurrent work Motion-X [30] provides a large-scale dataset with expressive human motions and paired text prompts. Yet, it still lacks detailed annotations for the hand motions, making the fine-grained generation challenging, let alone enabling explicit finger-level controls. On the other hand, various methods [43, 50, 51] synthesize two-hand motions with body motion as extra conditions. Such a body-to-hand setting implicitly reasons the inherent correlations of human motions between the body and hands, and hence effectively

handles specific scenarios like speeches or daily conversations [71, 79]. However, only body-level reasoning falls short of providing explicit and direct controls of hand motions, especially in a human-interpretable manner like text prompts.

To tackle the above challenges, in this paper, we present *BOTH2Hands* – a novel scheme to generate two-hand motions under a novel and hybrid setting: from both text prompts and body dynamics, as illustrated in Fig. 1. By organically combining the explicit and implicit conditions, our approach enables vivid and fine-grained hand motion generation. Nevertheless, generating hand motions in such a novel setting is challenging. First, it requires fusing and balancing the conditions from two very different modalities [23], which may point to diverse generation results. Second, the fundamental data scarcity for two-hand motion generation remains, while such a novel multi-modal further constitutes barriers to data annotations.

Specifically, we first introduced a large-scale multi-modal dataset, named *BOTH57M*, for two-hand motion modeling. Our dataset includes accurate hands and body motions with paired finger-level hand annotations and body descriptions, under diverse activities, covering 57.4 million frames of 8.31 hours with 23,477 textual annotations. To handle the occlusion, we adopt a camera dome with 32 RGB input views and utilize the off-the-shelf motion capture approach [17] to faithfully recover the skeletal motions of both the hands and body. We then provide two types of textual annotations for the captured motions: one describes the full body motions in general, while another focuses on fine-grained hand motions with finger-level and highly precise annotations. Note that our BOTH57M dataset is the first of its kind to open up future research for two-hand motion generation under hybrid conditions of both body dynamics and text prompts. Our accurate motions and expressive annotations also bring substantial potential for future direction in multi-modal control or human behavior analysis.

Based on our novel dataset, we further propose BOTH2Hands, a strong baseline approach to generate vivid two-hand motions from diverse conditions like body motions and text prompts. We tailor the recent diffusion models [20] into a two-stage mechanism for this novel task. Our core idea is to optimize the potential of the diffusion model using each modality separately and subsequently utilize a cross-attention transformer to blend them into a two-hand motion generation with multi-conditioning. Specifically, we warm up two parallel body-to-hand and text-to-hand diffusion models in the first stage. Then, we leverage a cross-attention transformer for motion blending, where two conditioned results are alternately inserted into the attention layers to generate convincing and vivid two-hand motions. Finally, we present a thorough evaluation of our approach against various state-of-the-art motion generation methods using our dataset. We also perform cross-validation on both our dataset and the concurrent Motion-X [30] dataset, demonstrating the enhancement of our dataset for the two-hand generation task. To summarize, our main contributions include:

- We propose a novel scheme to generate fine-grained two-hand motions under a novel setting: from both implicit body dynamics and explicit text prompts.
- We contribute a large-scale multi-modal dataset for two-hand generation, with accurate body and hand motions as well as rich finger-level textural annotations.
- We combine parallel diffusion structures with a subsequent cross-attention transformer, to effectively generate hand motions from various conditions.
- To tackle the data scarcity, we will release our dataset, codes, and pre-trained models for future exploration.

2. Related Works

Motion Generation. Currently, numerous works focus on motion generation under various conditions such as text and label [1, 3, 9, 14, 44, 46, 47, 62, 74–76], speech and music [15, 34, 68–70, 78, 79] and objects [4, 7, 16, 59]. Other interesting works use brand-new algorithms [18] or focus on new scenes [31]. Among these, text-to-motion generation is a challenging task due to the difficulty of aligning natural language with time and space [24, 45]. MotionClip [61] aligns text with other modalities, enhancing model mapping text to motion. As diffusion model [10, 20, 56, 57] was introduced in various tasks [8, 52, 52, 54, 55], it also performs well in motion generation. For instance, Human Motion Diffusion Model [62] introduced text conditions and showed good results. Other works like T2M-GPT [74] applied the transformer architecture in this task and proved its effectiveness. MLD (motion-latent-diffusion) [6] has attempted to generate motions in the latent space. Some other works like InterGen [29] focus on human interaction scenes achieving good results. Full-body motion generation holds considerable significance in some specific domains like human object interaction (HOI) [13, 27, 59, 60, 64] and speech [2, 35, 67, 70], due to its extensive applicability.

Hands are equally important as bodies in motion [42], presenting unique challenges due to their high density in small spatial occupancy. Previous hand generation works concentrate on physics-based issues [32, 48, 77] such as surface contact [72] and reconstructing hands [28, 53]. Other data-driven [22, 39, 41, 58] works often focus on specific scenarios [73], modeling the individual as a whole rather than considering parts separately [12, 65]. Some works like Body2Hands [43] take body as condition and achieve impressive results [50, 51, 71, 79]. Previous studies often focus on full-body or body motions only, rather than generating hands aligned with both body motions and text controls.

Motion Dataset. These days various motion datasets have

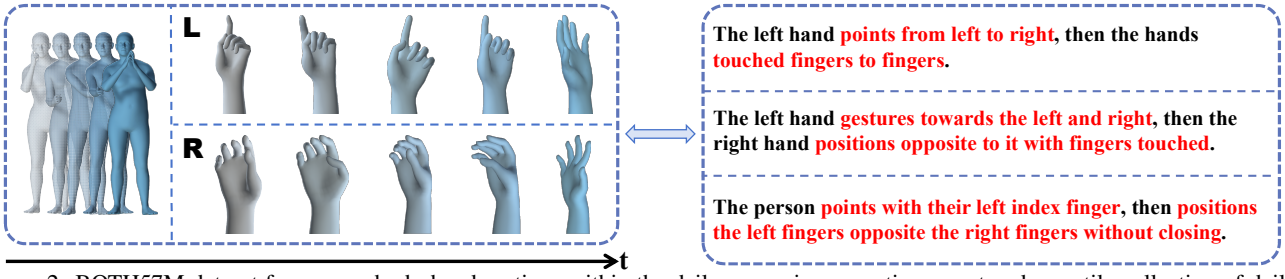

Figure 2. BOTH57M dataset focuses on body-hand motions within the daily scene, incorporating a vast and versatile collection of daily gestures. Each segment of hands within the dataset has been manually annotated three different times. For more details, please refer to the supplementary materials.

Dataset	#Frame	#Cam-view	Vocab	RGB	Annotations	Hand Joints std ↑	#Detailed annotation	
							Body	Hand
GRAB	1.6M	54	×	×	×	0.675	–	–
EgoBody	220K	6	×	✓	×	0.316	–	–
BEAT	32M	16	–	✓	–	0.076	–	–
MotionX	13.7M	–	2898	✓	✓	0.186	1	1
BOTH57M(Ours)	57.4M	32	4140	✓	✓	0.422	3	3

Table 1. **Dataset comparisons.** We conduct a comparison of datasets that encompass body and hand motions. **Vocab.** denotes the distinct vocabulary numbers used for annotation. **Annotation** refers to text annotations. **Hand Joint Standard Deviation** reflects the standard deviation of hand joint positions, indicating the diversity of hand motions. **Detailed annotation** refers to the number of text annotations for specific skeleton parts in each motion clip.

been presented. Action-labeled datasets like BABEL [49] offer verb-object phrases as conditions, which is unnatural for human communication. Datasets such as KIT [46] or HumanML3D [14] provide detailed natural annotations, while they ignore hands. Other datasets focus on hand scenarios like Hand-Object Manipulation [11], 3D Interacting Hand [26, 40]. Yet, such scenes mostly focus on hands, they hardly contain both hand and body data. Full-body datasets like GRAB [59] contain rich hand gestures but are narrowed down to HOI scenes. BEAT [33] uses speech text as conditions, lacking standard motion description. Currently, the largest full-motion dataset Motion-X [30] contains descriptions aligned with motions but lacks annotations focusing on hands, and the diversity of hand movement is less rich than their body motions. More data including rich daily hand gestures with detailed annotation is needed for body hand motion synthesis.

3. BOTH57M Dataset

Overview. We introduce the BOTH57M, a unique body-hand motion dataset comprising 1,384 motion clips and 57.4M frames, with 23,477 manually annotated motions and a rich vocabulary of 4,140 words. The dataset focuses on hands and body motion in daily various activities, referencing the book "Dictionary of Gestures " [5] and supplementing with custom-designed movements. To the best of our knowledge, this is the only dataset that provides hybrid and detailed annotations of both body and hands at present, providing huge potential for future research. Tab. 1 shows a detailed comparison of various body hand datasets with ours. The rich vocabulary and hand diversity underscores our advantage in tackling the text/body-to-hand task.

Data Collecting. We utilize 32 RGB cameras to build a dense-view system for body-hand motion capturing. During data collection, participants are instructed to perform movements listed in the "Dictionary of Gestures" excluding unfriendly gestures. Subsequently, manual annotations are implemented. Three annotators are required to annotate full-body motions. For hand motions, three other annotators individually annotate finger-level actions for the left and right hand, focusing on the changing process of finger movements and gaining detailed records for prominent finger gestures. Fig. 2 offers a comprehensive exemplification of our dataset. For a comprehensive understanding of data collection and processing, as well as an in-depth explanation of our collected motions and text annotations, please refer to the supplementary material.

4. BOTH2Hands Algorithm

Based on our novel dataset, our objective is to generate hand motions that align with both textual prompts and body movements. To accomplish this, we propose a novel pipeline called BOTH2Hands to deal with rich conditions to generate lively two-hand motions, as shown in Fig. 3. Our framework consists of a two-stage mechanism: a pair of diffusion-based hand motion denoisers and a cross-attention structured transformer. In the first stage, we feed the body and text controls into two parallel diffusion models. In the second stage, we blend the hand motions generated by two-modality controls. Specifically, we follow EgoEgo [25] and

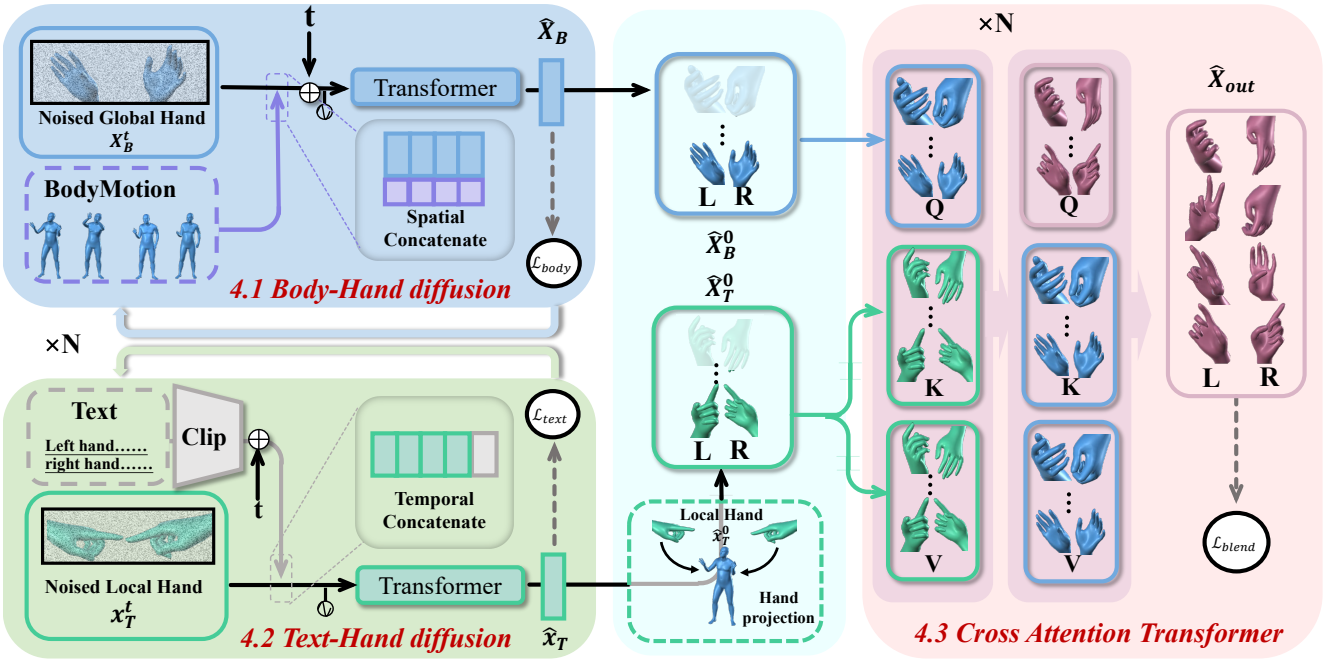

Figure 3. **Overview of BOTH2Hands pipeline.** Our pipeline initially feeds text prompts and body movements into two separate diffusion models. Subsequently, the text-conditioned outcomes are projected into the body-conditioned hand coordinate system using forward kinematics. Finally, we utilize the two sequences of hand motions as inputs into a cross-attention transformer for motion blending.

adopt forward kinematics (FK) to get joint positions and rotations 6D in body motion space. For body-conditioned hand diffusion, our goal is to generate hand motions that are coordinated with the body motions (Sec. 4.1). As for text-conditioned hand diffusion, we first use inverse kinematics (IK) to get local hands to make the denoise process more focused on gesture. However, these generated hands are not in the same coordinate system as the body-conditioned hands. To address this issue, we then employ FK to project the local hands back into the body motion space, thereby eliminating gesture rotation errors while blending (Sec. 4.2). After that, we perform cross-attention motion blending between the text-conditioned hands and the body-conditioned hands (Sec. 4.3). This process ensures the generated hand motions effectively combine the dynamics of body motion with the explicit textual conditions.

4.1. Body-hand motion diffusion

Motion Diffusion Model. In our approach, we adopt the formulation suggested in the denoising diffusion probabilistic model (DDPM) [20], which effectively handles the hand synthesis task. The diffusion model processes the input data ($\mathbf{x}_0$) for t iterations and obtains the noised data at level t. In each iteration, sampled Gaussian noise is added to the data from the previous level. This iterative process is commonly referred to as the forward process and can be represented as a Markov chain with t steps. The transition probability is shown in Eq. 1:

$$q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I}), \quad (1)$$

where β_t is a variance schedule parameter and $\alpha_t = \prod_{i=1}^{t}(1-\beta_i)$. The reverse diffusion process can be modeled as $p_\theta(x_{t-1}|x_t, c_{0:N})$, where θ represents the learned parameters and $c_{0:N}$ represents a set of given conditions (0 indicates no condition). Notable, we can always train a diffusion denoiser with any condition to learn a Gaussian posterior distribution $q(x_{t-1}|x_t, x_0)$. The denoising sampling process can be formulated as:

$$p_\theta(x_{t-1}|x_t, c_{0:N}) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t, c_{0:N}), \sigma_n^2\mathbf{I}). \quad (2)$$

The term $\mu_\theta(x_t, t, c_{0:N})$ is the mean to learn, which can be impacted by the conditions $c_{0:N}$. As mentioned in previous works [10, 23], to be more exact, the updating rule of the mean is:

$$\mu_t^{c_{0:N}} = \mu_t^{c_0} + \sum_{i=1}^{N} s_i(\nabla \log p(c_i|x_t)), \quad (3)$$

where $\mu_t^{c_0}$ is the mean without condition and the gradients of joint condition are noted as $\sum_{i=1}^{N}(\nabla \log p(c_i|x_t))$. The weights s_i controls the strength of conditioning. However, since different conditions pertain to distinct modalities, it is hard to manually configure the strength parameter s_i. Therefore, the preference leans towards the utilization of separate diffusion models to avoid imbalanced control by

different conditions. This approach prevents being affected by other conditions when learning μ_t since the gradient of a single condition is far easier to learn than joint conditions:

$$\mu_t^c = \mu_t + \nabla \log p(c|x_t), \tag{4}$$

where c indicates one single condition.

Body to Hand Diffusion. As the same body motion may lead to different hand gestures, we need the diffusion probabilistic model to sample the most possible gesture instead of directly matching one gesture on the body. We utilize the widely adopted model SMPLH (SMPL+MANO) [36, 53] as our skeleton, with a total of 52 joints, where the initial 22 joints are body joints and the remaining 30 joints are hand joints. We parameterize the representation of motions as positions and rotations 6D of joint. To fetch parameters, we use FK to calculate the absolute rotations and joint positions (we define absolute parameters as real positions and rotations of joints without the intervention of parent joints), denoted as global motions, where the former 22 joint positions and rotations 6D correspond to global body ($\mathbf{C}_B \in \mathbb{R}^{T\times 22\times 9}$) and the latter 30 joints correspond to global hands ($\mathbf{X}_B \in \mathbb{R}^{T\times 30\times 9}$). We perform the forward process by adding noise to the hand motions step-by-step, fetching the sequences of hand motions $\mathbf{X}_1^t, \mathbf{X}_2^t, ..., \mathbf{X}_T^t$ at noise level t. Followed by the forward process, we conduct a reverse diffusion procedure on the transformer self-attention denoiser to estimate $\mathbf{X}_B^0$. Then we adopt methods in [25] to directly concatenate the noised hand $\mathbf{X}_B^t$ and cleaned body $\mathbf{C}_B^0$ together as denoiser input during the body-hand diffusion process, with the loss shown in Eq. 5:

$$\mathcal{L}_{body} = \mathbb{E}_{\mathbf{x}_0,t}||\hat{\mathbf{X}}_\theta(\mathbf{X}_B^t, t, \mathbf{C}_B^0) - \mathbf{X}_B^0||_1. \tag{5}$$

We directly predict the cleaned motion $\mathbf{X}_B^0$ and use reconstruction loss as diffusion training loss.

4.2. Text-hand motion diffusion

For text-conditioned hand synthesis, we use IK to extract the local positions and rotations 6D from FK motion results. Then we discard body rotations, keeping hand rotations 6D as ground truth, the calculation process is defined below:

$$\mathbf{x}_T^{rot} = Cat(IK(\mathbf{M}_{lhand}^{rot}), IK(\mathbf{M}_{rhand}^{rot})), \tag{6}$$

where $\mathbf{M}$ is the full body motion aligned with text condition, $IK(\cdot)$ is inverse kinematics process and $Cat(\cdot,\cdot)$ is concatenate operation. For joint positions, we first use FK to calculate the 52 absolute joint positions. Then we can obtain hand joints in the origin of the coordinate system by subtracting the positions of their respective wrist for each hand joint:

$$\mathbf{x}_T^{pos} = Cat(\mathbf{M}_{lhand}^{pos} - \mathbf{M}_{lwrist}^{pos}, \mathbf{M}_{rhand}^{pos} - \mathbf{M}_{rwrist}^{pos}). \tag{7}$$

We define these rotation and position groups as local hands parameters ($\mathbf{x}_T \in \mathbb{R}^{T\times 30\times 9}$).

Hand Projection. Relative positions (we define relative parameters as joint positions and rotations relative to parent joints) representation may result in motion drifting due to the need for integrating velocity to obtain absolute positions [63]. Nevertheless, relative rotation representations are advantageous for focusing on gestures and are easy to migrate [36]. Remember that we use motion representation, consisting of positions and rotations 6D. For body-conditioned hand synthesis, absolute positions and rotations of hand joints with body joints are directly applicable. However, predicting the absolute pose of hands without wrist positions is hard and meaningless for text-conditioned synthesis, so for the positions, we prefer absolute representation in the origin of the coordinate system to avoid integration prediction and parent skeleton influence. To focus on the gesture itself, we prefer to use relative rotations. Nonetheless, absolute positions and relative rotations cannot be used for hand blending directly, since the two conditioned hands are on different coordinate systems. In order to mitigate the influence of the spatial reference system, we projected $\mathbf{x}_T$ in local space to $\mathbf{X}_T$ in global space to eliminate their potential space error:

$$\mathbf{X}_T = FK(Cat(IK(\mathbf{C}_B), \mathbf{x}_T)). \tag{8}$$

$FK(\cdot)$ is forward kinematics process, $\mathbf{x}_T$ are text-conditioned hand motions in local space, while $\mathbf{X}_T$ are text-conditioned hand motions in global space.

Text to Hand diffusion. On the text-conditioned diffusion process, we follow the methods proposed in MDM [62] to add the text condition token $\mathbf{c}$ to embed noise t step token. The denoising structure is similar to the body-conditioned motion denoiser, with a slight difference in input dimension. We only feed the noised hands as input since text-conditioned synthesis can not contain body motion. We adopt the reconstruction loss similar to Eq. 5 to predict $\mathbf{x}_T^0$:

$$\mathcal{L}_{text} = \mathbb{E}_{\mathbf{x}_0,t}||\hat{\mathbf{x}}_\theta(\mathbf{x}_T^t, t, \mathbf{c}) - \mathbf{x}_T^0||_1. \tag{9}$$

4.3. Cross-attention hand blending

Inspired by the success of the sharing-weights transformer in InterGen [29], we adopted a cross-attention transformer for gesture blending. The networks are fed with two conditioned hand motions, $\mathbf{X}_T$ and $\mathbf{X}_B$. We sequentially apply hand motions as attention inputs to the transformer, and compute the weighted reconstruction loss between the final output and two types of gestures. Specifically, the $\mathbf{X}_T$ and $\mathbf{X}_B$ are firstly embedded into a common latent space and positionally encoded into the latent states $\mathbf{h}_{text}$ and $\mathbf{h}_{body}$. Then, it is processed by N attention-based blocks to obtain the blending hidden states $\mathbf{h}_{out}^N$. Each block consists of multi-head cross-attention layers ($Attn$) followed by one

feed-forward network (FF). For the first time the hands passing through the cross-attention block, the input hidden layer $\mathbf{h}_{body}$ is embedded into the query matrix ($\mathbf{Q}$); the attention hidden layer $\mathbf{h}_{text}$ is embedded into a key matrix ($\mathbf{K}$) and value matrix ($\mathbf{V}$); finally we embed results into a vector $\mathbf{h}_{out}^{(1)}$. The hand-blending process is detailed below:

$$\mathbf{h}_{out}^{(1)} = Attn(\mathbf{Q}^{(0)}, \mathbf{K}^{(0)}, \mathbf{V}^{(0)}) = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D}})\mathbf{V},$$
$$\mathbf{Q}^{(0)} = \mathbf{h}_{body}W^Q, \mathbf{K}^{(0)} = \mathbf{h}_{text}W^K, \mathbf{V}^{(0)} = \mathbf{h}_{text}W^V, \tag{10}$$

where D is the number of channels in the attention layer; W are trainable weights, and $\mathbf{Q}^{(i)}$, $\mathbf{K}^{(i)}$, $\mathbf{V}^{(i)}$ are transformer matrices under i-th layer. Passing through the attention layer once, we get the output $\mathbf{h}_{out}^{(1)}$. Then we switch the $\mathbf{K}$, $\mathbf{V}$ input to $\mathbf{h}_{body}$, which means we use body-conditioned hands as attention input to emphasize the body movements. The changed attention process is:

$$\mathbf{Q}^{(1)} = \mathbf{h}_{out}^{(1)}W^Q, \mathbf{K}^{(1)} = \mathbf{h}_{body}W^K, \mathbf{V}^{(1)} = \mathbf{h}_{body}W^V. \tag{11}$$

After this, we swap $\mathbf{K}$, $\mathbf{V}$ input again and repeat this process until getting the final output in latent space:

$$\mathbf{h}_{out}^{(N)} = FF(Attn(\mathbf{Q}^{(N-1)}, \mathbf{K}^{(N-1)}, \mathbf{V}^{(N-1)})). \tag{12}$$

We use blending loss to supervise the learning of weights W.

$$\mathcal{L}_{blend} = \mathbb{E}_{X_{GT}, X_B, X_T}||X_{GT} - (w_B X_B + w_T X_T)||_1, \tag{13}$$

where w_B and w_T are hyperparameters controlling weights of different hand motion parts. We set w_B and w_T to positive numbers and $w_B + w_T = 1$. X_{GT} is GT hand motion.

5. Experiment

We design various experiments to evaluate the validity of our method and dataset. For method evaluation, we compare our approach and baseline with existing human motion synthesis methods (Sec. 5.1). To assess the richness and effectiveness of the BOTH57M, we train our method on training sets of BOTH57M and Motion-X separately. And subsequently evaluated trained models through the test sets (Sec. 5.2). Additionally, an ablation study is performed to verify the importance of hand projection and blending loss (Sec. 5.3).

5.1. Methods Evaluation

We compare our approach with several other methods in the task of generating hands based on textual and body conditions. We introduce latent text-to-motion methods T2M-GPT [74] and MLD [6], diffusion-based method MDM [62], and body-conditioned motion synthesis method EgoEgo [25] for comparison. We align the input and output dimensions for unbiased comparison, keeping text and body conditions the same for all methods. In MLD, we employ two encoder-decoder structures for the body and hands. In the latent diffusion process, we merge the cleaned latent body token with the noisy hand token. The combined token is then denoised to predict latent hands, which are subsequently fed into the hand decoder. For T2M-GPT, we train an encoder-decoder structure to derive body features and then add up body and text tokens. For non-latent space methods, we directly concatenate the body conditions onto the noised hands as input and follow [62] to add text conditions into it. All methods use the same implementation details as they presented. For the structures added to other methods, we keep dimensions the same as the original framework. For our method, all transformers consist of N=4 blocks, a latent dimension of 512, and 4 attention heads. We use a frozen CLIP-ViTL-14 model as the text encoder. As for other parameters, the diffusion timesteps are set to 1000 during training and inference; the AdamW optimizer is used with a fixed learning rate of $1e^{-4}$; and hyperparameter w_B is set to 0.8, w_T is set to 0.2; the motion blending process (method in Sec.4.3) is performed 3 times. All the methods are trained with the BOTH57M training set on a single NVIDIA GeForce RTX 2080 Ti GPU for about 2 days. For more inference results, please refer to supplementary materials.

Following [14], our evaluation metrics include Motion-retrieval precision (R Precision), Fréchet Inception Distance (FID) [19], Multi-modal Distance (MM-Dist), Diversity and MultiModality (MModality). And we randomly split BOTH57M into the train (80%), val (5%), and test (15%) set, adopting SMPLH as motion representation. Tab. 2 presents detailed quantitative results from the same test set, showing our methods reconstructed motions closest to the real motion. Fig. 4 shows BOTH2Hands achieves good alignment between hand motions and conditions. Non-latent methods perform well on body conditions, but poorly on text conditions. Latent methods struggle with body conditions. Fig. 6 demonstrates our blending block beats our diffusion baseline. Text-conditioned hands lack body alignment, while body-conditioned hands fail to meet prompt requirements. Nevertheless, we also see marginal improvements in our evaluation results. This phenomenon implies that multi-conditioned generation performance may not be adequately reflected by widely used metrics such as R-precision, which are effective for single-conditioned generation evaluation. Two key reasons are listed below. First, in a full-body setting, hand motion constitutes a small portion, limiting the metric-based improvements. Second, the correlation between hand motions and text is non-linear, the metrics increase brought by enhancing hand motion is limited due to complicated hand-text alignment. We plan to ex-

Table 2. **Quantitative evaluation** of our design with baselines and others. The red one and blue one indicate the best result and the second best result. We use a 95% confidence interval, approximated by the mean value plus or minus twice the standard deviation.

Methods	R Precision↑			FID↓	MM-Dist↓	Diversity→	MModality↑
	Top1	Top2	Top3				
Real	$0.034^{\pm.020}$	$0.067^{\pm.026}$	$0.109^{\pm.030}$	$0.181^{\pm.012}$	$1.391^{\pm.006}$	$3.980^{\pm.090}$	-
T2M-GPT	**0.042**$^{\pm.014}$	$0.073^{\pm.014}$	$0.104^{\pm.020}$	$0.461^{\pm.016}$	$1.398^{\pm.010}$	$3.689^{\pm.094}$	$1.178^{\pm.100}$
MDM	$0.039^{\pm.020}$	**0.077**$^{\pm.016}$	**0.114**$^{\pm.024}$	$0.257^{\pm.024}$	$1.397^{\pm.008}$	$3.887^{\pm.074}$	$1.273^{\pm.086}$
MLD	$0.036^{\pm.012}$	$0.071^{\pm.014}$	$0.106^{\pm.020}$	$0.296^{\pm.026}$	$1.400^{\pm.0014}$	$3.826^{\pm.078}$	$1.191^{\pm.178}$
Ego-Ego	$0.034^{\pm.022}$	$0.070^{\pm.030}$	$0.109^{\pm.032}$	$0.287^{\pm.026}$	$1.398^{\pm.012}$	$3.810^{\pm.090}$	$1.240^{\pm.090}$
BOTH2Hands (Ours)	$0.037^{\pm.014}$	$0.075^{\pm.020}$	**0.115**$^{\pm.028}$	**0.201**$^{\pm.020}$	**1.392**$^{\pm.008}$	$3.969^{\pm.082}$	**1.312**$^{\pm.034}$
BOTH2Hands-Text	$0.035^{\pm.020}$	$0.067^{\pm.026}$	$0.109^{\pm.030}$	**0.198**$^{\pm.012}$	**1.391**$^{\pm.006}$	$3.980^{\pm.090}$	**1.274**$^{\pm.138}$
BOTH2Hands-Body	**0.039**$^{+.012}$	**0.076**$^{\pm.024}$	$0.112^{\pm.026}$	$0.203^{\pm.016}$	$1.392^{\pm.010}$	$3.955^{\pm.098}$	$1.266^{\pm.120}$

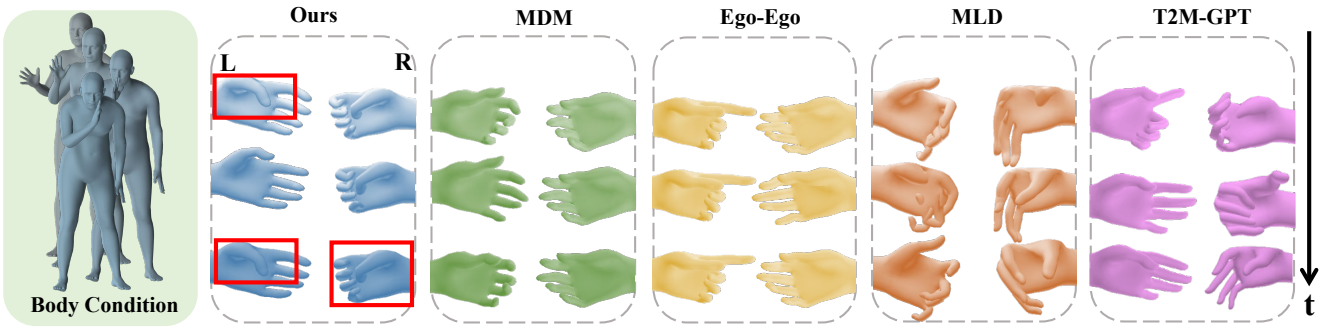

Figure 4. **Qualitative comparisons.** BOTH2Hands algorithm with other methods [6, 25, 62, 74] are given two conditions: text and motion. Text conditions are listed at the top, and body conditions are listed at the left side with no hands, and the temporal order is from top to bottom. The motions that follow the conditions are circled red.

Training Set	Test Set	R precision↑	FID↓	Diversity→	MModality ↑
Real(GT)	Motion-X	$0.044^{\pm.006}$	$0.353^{\pm.068}$	$3.224^{\pm.190}$	–
	BOTH57M	$0.034^{\pm.020}$	$0.181^{\pm.012}$	$3.980^{\pm.090}$	–
Motion-X	Motion-X	**0.048**$^{\pm.018}$	**0.364**$^{\pm.040}$	$3.203^{\pm.116}$	**1.087**$^{\pm.072}$
	BOTH57M	$0.026^{\pm.008}$	$2.399^{\pm.076}$	$1.828^{\pm.200}$	$0.581^{\pm.090}$
BOTH57M	Motion-X	$0.030^{\pm.002}$	$0.858^{\pm.024}$	$4.002^{\pm.054}$	$1.259^{\pm.224}$
	BOTH57M	**0.037**$^{\pm.014}$	**0.201**$^{\pm.020}$	$3.969^{\pm.082}$	**1.312**$^{\pm.034}$

Table 3. **Cross-dataset comparisons** of BOTH57M and Motion-X. We train our pipeline on the training set of them, then evaluate the models on their test sets. Real(GT) means the GT data in the training set is used for evaluation. We use a 95% confidence interval, approximated by the mean value plus or minus twice the standard deviation.

plore more suitable metrics for future studies and hope our released code and dataset can serve as a solid foundation for such exploration.

5.2. Dataset Evaluation

To highlight the richness of the hand motions and the accuracy of text prompts, we compare our BOTH57M with Motion-X [30], the largest full-body motion dataset with text currently. We train BOTH2Hands method on training sets of Motion-X and BOTH57M separately. Then validate methods on respective test sets. The comparison results are presented in Tab. 3. We add the GT data in the training set to the evaluation as the standard. The model trained on Motion-X training set performs well on the test set. However, the model trained on BOTH57M provides better alignment from text to hands, and its hand diversity is also better than the model trained on Motion-X. Fig. 5 shows our qualitative results. Given body and text conditions on test sets, our method trained on BOTH57M always performs better on text and body conditions. It also performs well on general motion prompts due to general motion annotations that contain hand descriptions in BOTH57M.

Method	R Precision↑	FID↓	MM-Dist↓
GT	$0.034^{\pm.020}$	$0.181^{\pm.012}$	$1.391^{\pm.006}$
Ours	**0.037**$^{\pm.014}$	**0.201**$^{\pm.020}$	**1.392**$^{\pm.008}$
w/o hand proj	$0.036^{\pm.014}$	$0.204^{\pm.026}$	$1.393^{\pm.008}$
w/o blending loss	$0.034^{\pm.020}$	$0.210^{\pm.022}$	$1.392^{\pm.010}$

Table 4. **Ablation study of BOTH2Hands algorithm.** Hand projection will fully improve the method results. We use a 95% confidence interval, approximated by the mean value plus or minus twice the standard deviation.

5.3. Ablation Study

To validate the importance of hand projection and blending loss, we perform an ablation study by evaluating the effects of excluding these elements from the BOTH2Hands

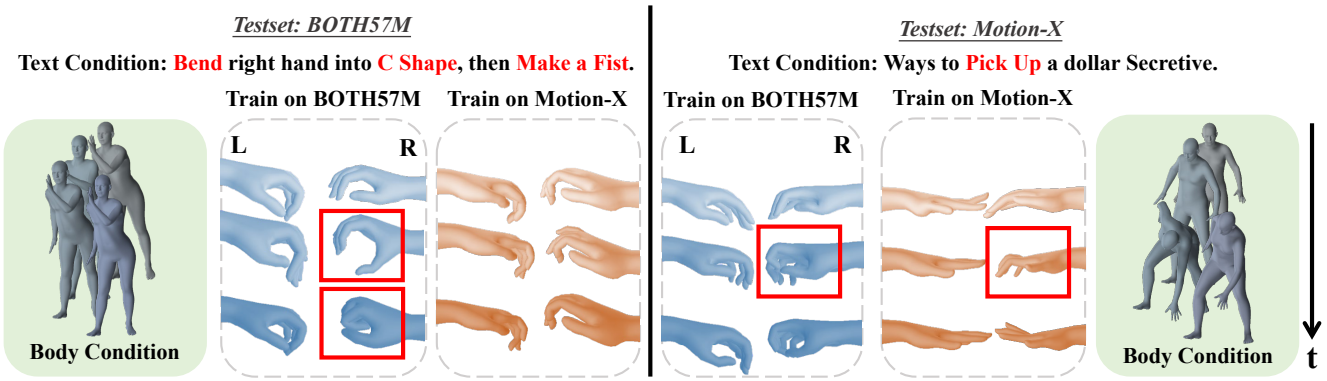

Figure 5. **Dataset Evaluation.** We train the BOTH2Hands algorithm on the training set of BOTH57M and Motion-X. Then sample on the test set of BOTH57M dataset (left) and Motion-X dataset (right). The poses that follow the conditions are circled red.

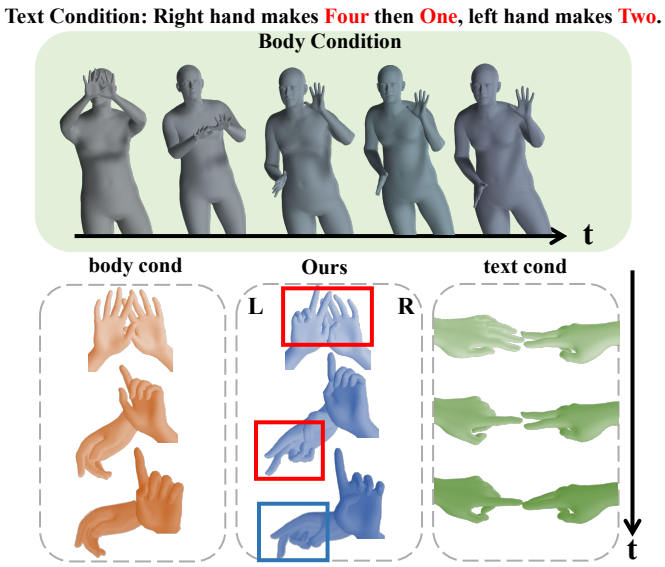

Figure 6. **Qualitative results of baseline comparison.** The motions following texts are circled red, the motions following body are circled blue.

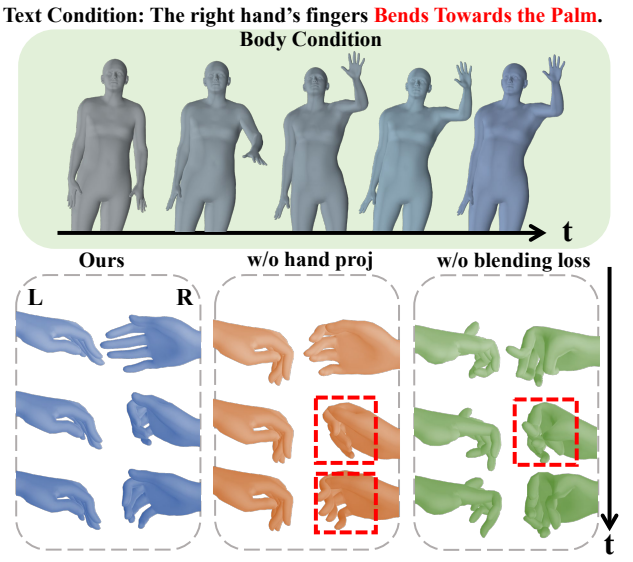

Figure 7. **Qualitative results of ablation study.** The errors of motions are circled red.

algorithm. Hand projection can be removed directly. But for blending loss, we choose linear distance loss as an alternative. Numerical results in Tab. 4 indicate our method performs better with hand projection. This process allows the transformer to focus solely on the motion. Blending loss also highly improves hand motion quality, proving that learning the hand from previous output is effective. As mentioned in Sec. 5.2, marginal improvements found in the ablation study also suffer from the insensitivity of the existing metrics used. But obvious improvements shown in Fig. 7 still prove the effectiveness of our hand projection and blending loss.

6. Conclusion

We introduce BOTH57M, the first comprehensive body-hand dataset that incorporates precise gestures and body movements, paired with meticulous finger-level hand annotations and body descriptions, which spans a variety of activities, consisting of 57.4 million frames in 8.31 hours, supplemented with 23,477 text annotations. Based on this dataset, we introduce BOTH2Hands, a robust algorithm designed to generate hand movements under two conditions: body movements and text prompts. Subsequently, we employ a cross-attention transformer for motion blending. We also conduct a series of detailed evaluations to demonstrate the robustness of our methods and the enhancement of our dataset for the two-hand generation task. We believe the BOTH57M could boost future exploration in multi-modal control and the analysis of human behavior.

7. Acknowledgement

This work was supported by National Key R&D Program of China (2022YFF0902301), Shanghai Local college capacity building program (22010502800). We also acknowledge support from Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI).

References

[1] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728. IEEE, 2019. 2

[2] Tenglong Ao, Zeyi Zhang, and Libin Liu. Gesturediffuclip: Gesture diffusion model with clip latents. *ACM Transactions on Graphics (TOG)*, 42(4):1–18, 2023. 2

[3] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Teach: Temporal action composition for 3d humans. In *2022 International Conference on 3D Vision (3DV)*, pages 414–423. IEEE, 2022. 2

[4] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 361–378. Springer, 2020. 2

[5] François Caradec and Philippe Cousin. *Dictionary of gestures: Expressive comportments and movements in use around the world*. MIT Press Boston, MA, 2018. 3, 1

[6] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 1, 2, 6, 7

[7] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5031–5041, 2020. 2

[8] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9760–9770, 2023. 1, 2

[9] Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. Posescript: 3d human poses from natural language. In *European Conference on Computer Vision*, pages 346–362. Springer, 2022. 2

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2, 4

[11] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[12] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7214–7223, 2020. 2

[13] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Computer Graphics Forum*, pages 1–12. Wiley Online Library, 2023. 2

[14] 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*, pages 5152–5161, 2022. 1, 2, 3, 6, 5

[15] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. Learning speech-driven 3d conversational gestures from video. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pages 101–108, 2021. 2

[16] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11807–11816, 2019. 2

[17] Yannan He, Anqi Pang, Xin Chen, Han Liang, Minye Wu, Yuexin Ma, and Lan Xu. Challencap: Monocular 3d capture of challenging human performances using multi-modal references. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11400–11411, 2021. 2

[18] Yannan He, Garvita Tiwari, Tolga Birdal, Jan Eric Lenssen, and Gerard Pons-Moll. Nrdf: Neural riemannian distance fields for learning articulated pose priors. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6, 2

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 4

[21] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 1

[22] Sophie Jörg, Jessica Hodgins, and Alla Safonova. Data-driven finger motion synthesis for gesturing characters. *ACM Transactions on Graphics (TOG)*, 31(6):1–7, 2012. 2

[23] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2151–2162, 2023. 2, 4

[24] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8255–8263, 2023. 2

[25] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17142–17151, 2023. 3, 5, 6, 7

[26] Lijun Li, Linrui Tian, Xindi Zhang, Qi Wang, Bang Zhang, Liefeng Bo, Mengyuan Liu, and Chen Chen. Renderih: A large-scale synthetic dataset for 3d interacting hand pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20395–20405, 2023. 3

[27] Quanzhou Li, Jingbo Wang, Chen Change Loy, and Bo Dai. Task-oriented human-object interactions generation with implicit neural representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3035–3044, 2024. 2

[28] Yuwei Li, Longwen Zhang, Zesong Qiu, Yingwenqi Jiang, Nianyi Li, Yuexin Ma, Yuyao Zhang, Lan Xu, and Jingyi Yu. Nimble: a non-rigid hand model with bones and muscles. *ACM Transactions on Graphics (TOG)*, 41(4):1–16, 2022. 2

[29] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *arXiv preprint arXiv:2304.05684*, 2023. 1, 2, 5, 3, 6

[30] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 1, 2, 3, 7

[31] Pei Lin, Sihang Xu, Hongdi Yang, Yiran Liu, Xin Chen, Jingya Wang, Jingyi Yu, and Lan Xu. Handdiffuse: Generative controllers for two-hand interactions via diffusion models, 2023. 2

[32] C Karen Liu. Synthesis of interactive hand manipulation. In *Proceedings of the 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 163–171, 2008. 2

[33] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European Conference on Computer Vision*, pages 612–630. Springer, 2022. 3

[34] Xian Liu, Qianyi Wu, Hang Zhou, Yuanqi Du, Wayne Wu, Dahua Lin, and Ziwei Liu. Audio-driven co-speech gesture video generation. *Advances in Neural Information Processing Systems*, 35:21386–21399, 2022. 2

[35] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10462–10472, 2022. 2

[36] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. pages 248:1–248:16. ACM, 2015. 5

[37] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019. 1

[38] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 1

[39] Anna Majkowska, Victor Zordan, and Petros Faloutsos. Automatic splicing for hand and body animations. In *ACM SIGGRAPH 2006 Sketches*, pages 32–es. 2006. 2

[40] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 548–564. Springer, 2020. 3, 1

[41] Christos Mousas, Christos-Nikolaos Anagnostopoulos, and Paul Newbury. Finger motion estimation and synthesis for gesturing characters. In *Proceedings of the 31st Spring Conference on Computer Graphics*, pages 97–104, 2015. 2

[42] Lea Müller, Vickie Ye, Georgios Pavlakos, Michael Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3d social interaction from images. *arXiv preprint arXiv:2306.09337*, 2023. 2

[43] Evonne Ng, Shiry Ginosar, Trevor Darrell, and Hanbyul Joo. Body2hands: Learning to infer 3d hands from conversational gesture body dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11865–11874, 2021. 1, 2

[44] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021. 2

[45] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022. 2

[46] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 2, 3

[47] Matthias Plappert, Christian Mandery, and Tamim Asfour. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *Robotics and Autonomous Systems*, 109:13–26, 2018. 2

[48] Nancy S Pollard and Victor Brian Zordan. Physically based grasping control from example. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 311–318, 2005. 2

[49] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 722–731, 2021. 3

[50] Xingqun Qi, Chen Liu, Lincheng Li, Jie Hou, Haoran Xin, and Xin Yu. Emotiongesture: Audio-driven diverse emo-

tional co-speech 3d gesture generation. *arXiv preprint arXiv:2305.18891*, 2023. 1, 2

[51] Xingqun Qi, Chen Liu, Muyi Sun, Lincheng Li, Changjie Fan, and Xin Yu. Diverse 3d hand gesture prediction from body dynamics by bilateral hand disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4616–4626, 2023. 1, 2

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[53] Javier Romero, Dimitris Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6), 2017. 2, 5

[54] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[55] István Sárándi, Alexander Hermans, and Bastian Leibe. Learning 3d human pose estimation from dozens of datasets using a geometry-aware autoencoder to bridge between skeleton formats. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2956–2966, 2023. 2

[56] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 2

[57] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 2

[58] Matthew Stone, Doug DeCarlo, Insuk Oh, Christian Rodriguez, Adrian Stere, Alyssa Lees, and Chris Bregler. Speaking with hands: Creating animated conversational characters from recordings of human performance. *ACM Transactions on Graphics (TOG)*, 23(3):506–513, 2004. 2

[59] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 581–600. Springer, 2020. 2, 3, 1

[60] Purva Tendulkar, Dídac Surís, and Carl Vondrick. Flex: Full-body grasping without full-body grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21179–21189, 2023. 2

[61] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022. 1, 2

[62] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 5, 6, 7

[63] Timo Von Marcard, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer graphics forum*, pages 349–360. Wiley Online Library, 2017. 5

[64] Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. Saga: Stochastic whole-body grasping with contact. In *European Conference on Computer Vision*, pages 257–274. Springer, 2022. 2

[65] Hao Xu, Tianyu Wang, Xiao Tang, and Chi-Wing Fu. H2onet: Hand-occlusion-and-orientation-aware network for real-time 3d hand mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17048–17058, 2023. 2

[66] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 1

[67] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, and Haolin Zhuang. Qpgesture: Quantization-based and phase-guided motion matching for natural speech-driven gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2321–2330, 2023. 2

[68] Siyue Yao, Mingjie Sun, Bingliang Li, Fengyu Yang, Junle Wang, and Ruimao Zhang. Dance with you: The diversity controllable dancer generation via diffusion models. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8504–8514, 2023. 2

[69] Sheng Ye, Yu-Hui Wen, Yanan Sun, Ying He, Ziyang Zhang, Yaoyuan Wang, Weihua He, and Yong-Jin Liu. Audio-driven stylized gesture generation with flow-based model. In *European Conference on Computer Vision*, pages 712–728. Springer, 2022.

[70] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 469–480, 2023. 2

[71] Lianying Yin, Yijun Wang, Tianyu He, Jinming Liu, Wei Zhao, Bohan Li, Xin Jin, and Jianxin Lin. Emog: Synthesizing emotive co-speech 3d gesture with diffusion model. *arXiv preprint arXiv:2306.11496*, 2023. 2

[72] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. Manipnet: neural manipulation synthesis with a hand-object spatial representation. *ACM Transactions on Graphics (ToG)*, 40(4):1–14, 2021. 2

[73] Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8834–8845, 2023. 2

[74] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 6, 7

[75] 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.

[76] Mengyi Zhao, Mengyuan Liu, Bin Ren, Shuling Dai, and Nicu Sebe. Modiff: Action-conditioned 3d motion generation with denoising diffusion probabilistic models. *arXiv preprint arXiv:2301.03949*, 2023. 2

[77] Wenping Zhao, Jianjie Zhang, Jianyuan Min, and Jinxiang Chai. Robust realtime physics-based motion control for human grasping. *ACM Transactions on Graphics (TOG)*, 32 (6):1–12, 2013. 2

[78] Zhuoran Zhao, Jinbin Bai, Delong Chen, Debang Wang, and Yubo Pan. Taming diffusion models for music-driven conducting motion generation. In *Proceedings of the AAAI Symposium Series*, pages 40–44, 2023. 2

[79] Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. Taming diffusion models for audio-driven co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10544–10553, 2023. 2