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

Supplementary Material

A. More Details of BOTH57M

In this section, we delve deeper into the specifics of our dataset. The following paragraphs introduce the camera dome setting, data reconstruction pipeline with data quality discussion, data text annotation, and data statics.

A.1. Camera Dome Setting

Our RGB cameras are mostly arranged evenly in three layers(low, medium, and high) at 270°. Other cameras are arranged on the highest layer at the remaining 90°. We zoom in on some of the cameras in order to ensure that they are sufficiently fine so that the motion of hands can be captured in detail. Moreover, we placed three 5500W fill lights in the dome to support hand motion capture with high-precision zoom that highly requires brightness. Fig. 8 shows the cameras and fill lights setting in the dome, the red cameras are zoomed in for capturing hands.

A.2. Capture System Synchronization

In order to avoid motion blurring, the RGB cameras are set to work at 59.97 FPS with the resolution adjusted to 3840×2160. Furthermore, cameras are available only if they are all turned on and set to the intended focal length. When being captured, everyone is required to make a T-pose at the beginning and end of each shoot so that the initial orientation alignment of ground truth can be easily performed, i.e. the origin of the coordinate system is at the initial position of the person and the forward direction is towards the z-coordinate with up direction is towards the y-coordinate, using a right-handed system.

A.3. Actor Requirements

During shooting, actors should perform motions in "Dictionary of Gestures" [5] excluding unfriendly gestures. They are required to expand the pose in the book into a complete body-hand motion sequence and perform it. Performers are encouraged to increase the diversity of hand gestures and body motions without changing the original meaning of the book.

A.4. Data Reconstruction and Quality

For data reconstruction, we adopt ViTPose [66] for predicting 2D body keypoints, and MediaPipe [37] for estimating 2D hand keypoints from RGB images. Markerless mocap then derives SMPLH parameters and 3D coordinates from these 2D results. Fig. 9 provides an overlay of our dataset. We are confident that our dataset exhibits sufficient full-body quality in 32-camera views with partial camera zoom-in for hand motion. We quantitatively validated our dataset using a subset with manually labeled 2D keypoints, triangulated for 3D joint coordinates as a gold standard. Our annotations show **31.4 mm** MPJPE and **25.2 mm** PA-MPJPE overall, with hand metrics at **6.51 mm** MPJPE and **3.97 mm** PA-MPJPE, which is comparable to the widely adopted datasets, e.g., 36.1mm for Human3.6M [21] (full-body), 18.4mm for Grab [59] (body/hand marker-based), 2.78mm for InterHand [40] (hand only), 33.5mm for Motion-X [30] (pseudo-GT), etc. Nevertheless, we will release all the raw captured images, and welcome the community to further improve the tracking results.

Body part interact with	# Frame	Time
Head et al.	16.33M	2.33h
UpperBody et al.	5.64M	0.81h
Hand et al.	34.59M	5.01h
LowerBody et al.	1.27M	0.16h
Total	57.83M	8.31h

Table 5. BOTH57M Subjects.

A.5. Data Text Annotation

For data text annotation, we segment each motion into clips of 5 seconds. We also offer the data together with the book "Dictionary of Gestures" [5] to the annotators. For each motion sequence, three annotators are supposed to describe the motion according to the emotion from the corresponding poses in the book in "Physical Description + Meaning of Action" format. "Physical Description" means to describe the action precisely; "Meaning of Action" means the emotion the action wants to express. For example, "Put the five fingers of your right hand together in front of your right ear." is a physical description, and "Such an action is meant to express listening carefully." is the meaning of the action. The other three annotators are required to focus on only hand movements and use detailed descriptions to explain how the fingers move. The annotations contain only physical descriptions such as "Left-hand index finger straight, other fingers are naturally bent, right-hand thumb extended, other fingers in a fist". Fig. 10 shows the motion of our dataset as well as the corresponding annotation in detail.

A.6. Data Statics

BOTH57M is the only dataset that provides hybrid and detailed annotations of both body and hands at present. It consists of 1,384 motion clips and 57.4M frames, with 23,477 manually annotated motions and a rich vocabulary of 4,140 words. The data-capturing system includes 32 synchronized high-resolution RGB cameras with 59.97 FPS and

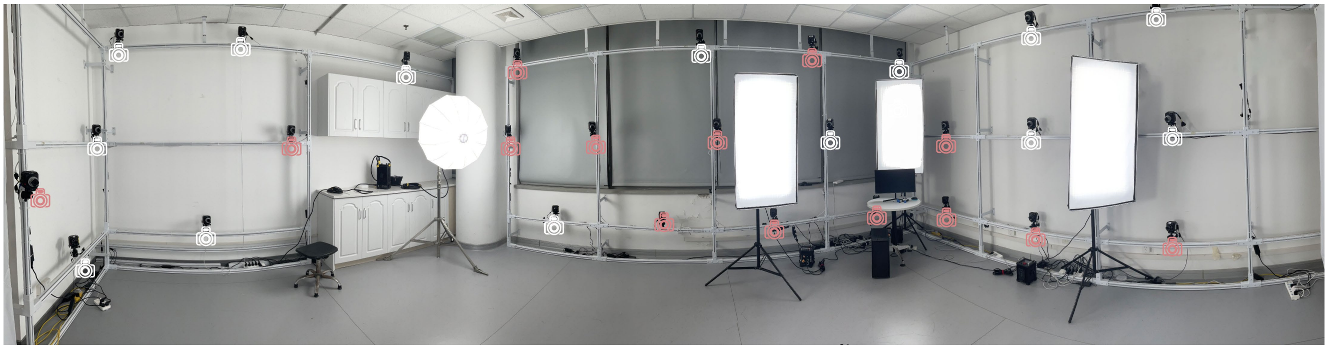


Figure 8. Dome setting, the red cameras are zoomed cameras for capturing hands, and other white cameras are wide-angle cameras for capturing the full body.

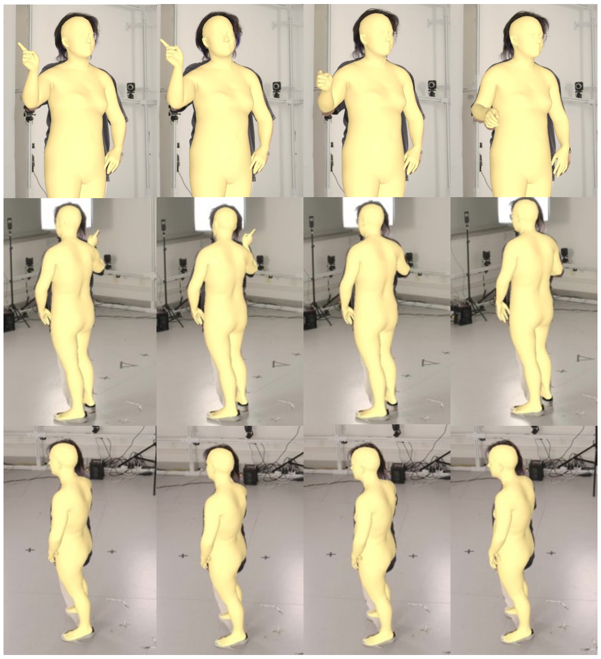


Figure 9. Overlays of BOTH57M in SMPLH. Three different views are presented, arranged from left to right in order of temporal variation.

3840×2160 resolution, which are capable of full body capture tasks. As for subjects of the dataset, we follow "Dictionary of Gestures" to scientifically split the dataset into 36 subjects according to the body part that hands interact with, a statistic is shown in Tab.5.

B. More Experiment Results

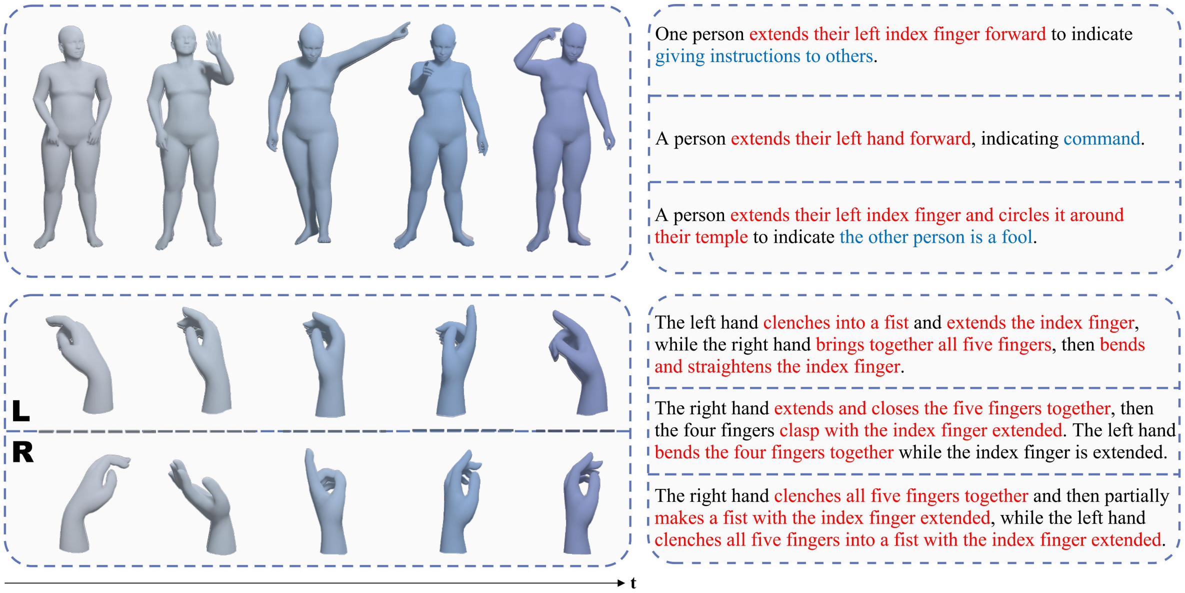
In this section, we provide more experiment visualization results and an extra hyperparameter experiment finding suitable w_B and w_T settings.

B.1. Hyperparameter experiments

We establish 11 sets of hyperparameter experiments with an interval of 0.1, starting at 0, under the premise that $w_B + w_T = 1$. We limit $w_B + w_T = 1$ to ensure the cross-attention transformer layer returns the same order of magnitude of loss as the previous diffusion process. Tab. 6 shows detailed results under different hyperparameter settings. The evaluation metrics are composed of the following sections: Motion-retrieval precision (R Precision), which we adhere to the top 3 results, is used to measure the alignment degree between conditions and motions. Fréchet Inception Distance (FID) [19] is employed to evaluate the feature distribution between the generated actions and the ground truth motion. Multi-modal Distance (MM-Dist) calculates the distance between hand motions and body text conditions. Diversity assesses the richness of generation motion by calculating the variance of data. Multimodality (MModality) measures the variety of generated motions under the same input conditions. The experimental results demonstrate that utilizing solely the body or text as the condition, corresponding to setting $w_T = 0$ or $w_B = 0$ in the experiment, can not get the best results. As the weight of the text condition w_T increases, the alignment performance (R Precision) between motion and condition improves, but the authenticity of the motion (FID) deteriorates. As the weight of the body condition w_B increases, the authenticity of the motion (FID) improves, but the Multi-modal Distance (MM-Dist) also increases, indicating that the distance between the generated hand motion result and the condition has grown farther in latent space. Under the experiment results, we decide to use a group with a good match degree between motion and condition, and relatively realistic motion, specifically setting the hyperparameters of the pipeline as $w_B = 0.8$ and $w_T = 0.2$.

B.2. More evaluation results

We present more visualization results of comparing methods to further validate the effectiveness of the BOTH2Hands algorithm. As shown in Fig. 11, under



One person extends their left index finger forward to indicate giving instructions to others.

A person extends their left hand forward, indicating command.

A person extends their left index finger and circles it around their temple to indicate the other person is a fool.

The left hand clenches into a fist and extends the index finger, while the right hand brings together all five fingers, then bends and straightens the index finger.

The right hand extends and closes the five fingers together, then the four fingers clasp with the index finger extended. The left hand bends the four fingers together while the index finger is extended.

The right hand clenches all five fingers together and then partially makes a fist with the index finger extended, while the left hand clenches all five fingers into a fist with the index finger extended.

Figure 10. Detailed data structure of BOTH57M. We provide body annotations with general motion meaning, marked blue in the figure, and detailed finger-level annotations, the physical descriptions are marked red in the figure.

Table 6. Quantitative evaluation of our condition hyperparameter setting, red and blue results 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.

w_B/w_T	R Precision↑			FID↓	MM-Dist↓	Diversity→	MModality↑
	Top1	Top2	Top3				
1.0/0.0	$0.036^{\pm0.026}$	$0.064^{\pm0.034}$	$0.094^{\pm0.038}$	$0.201^{\pm0.018}$	$1.402^{\pm0.016}$	$3.983^{\pm0.062}$	$1.262^{\pm0.138}$
0.9/0.1	$0.032^{\pm0.022}$	$0.063^{\pm0.028}$	$0.093^{\pm0.034}$	$\mathbf{0.198}^{\pm0.022}$	$1.402^{\pm0.012}$	$3.968^{\pm0.074}$	$1.298^{\pm0.124}$
0.8/0.2	$0.037^{\pm0.014}$	$\mathbf{0.075}^{\pm0.020}$	$\mathbf{0.115}^{\pm0.028}$	$0.201^{\pm0.020}$	$\mathbf{1.392}^{\pm0.008}$	$3.969^{\pm0.082}$	$\mathbf{1.312}^{\pm0.034}$
0.7/0.3	$0.031^{\pm0.020}$	$0.060^{\pm0.024}$	$0.088^{\pm0.015}$	$\mathbf{0.199}^{\pm0.020}$	$1.404^{\pm0.008}$	$3.968^{\pm0.060}$	$\mathbf{1.303}^{\pm0.104}$
0.6/0.4	$0.030^{\pm0.024}$	$0.059^{\pm0.028}$	$0.089^{\pm0.032}$	$0.210^{\pm0.018}$	$1.405^{\pm0.012}$	$3.980^{\pm0.060}$	$1.291^{\pm0.126}$
0.5/0.5	$0.032^{\pm0.012}$	$0.069^{\pm0.024}$	$0.096^{\pm0.032}$	$0.223^{\pm0.026}$	$1.401^{\pm0.012}$	$3.939^{\pm0.048}$	$1.267^{\pm0.246}$
0.4/0.6	$\mathbf{0.038}^{\pm0.010}$	$0.074^{\pm0.016}$	$0.104^{\pm0.024}$	$0.237^{\pm0.038}$	$\mathbf{1.391}^{\pm0.010}$	$3.939^{\pm0.070}$	$1.250^{\pm0.136}$
0.3/0.7	$0.038^{\pm0.016}$	$0.074^{\pm0.026}$	$0.110^{\pm0.026}$	$0.231^{\pm0.028}$	$1.392^{\pm0.014}$	$3.927^{\pm0.056}$	$1.287^{\pm0.112}$
0.2/0.8	$0.035^{\pm0.012}$	$0.069^{\pm0.014}$	$0.108^{\pm0.016}$	$0.235^{\pm0.026}$	$1.392^{\pm0.010}$	$3.922^{\pm0.066}$	$1.231^{\pm0.130}$
0.1/0.9	$\mathbf{0.041}^{\pm0.020}$	$\mathbf{0.079}^{\pm0.020}$	$\mathbf{0.113}^{\pm0.032}$	$0.233^{\pm0.026}$	$1.392^{\pm0.010}$	$3.934^{\pm0.052}$	$1.223^{\pm0.268}$
0.0/1.0	$0.033^{\pm0.010}$	$0.067^{\pm0.022}$	$0.104^{\pm0.030}$	$0.237^{\pm0.028}$	$1.393^{\pm0.008}$	$3.972^{\pm0.058}$	$1.266^{\pm0.130}$

the same text condition but different body conditions, our method can still generate hand results that correspond to the text and body. We also illustrate more cross-validation visualization results. As depicted in Fig. 12, under the challenge of more vague text descriptions and professional dance movements, our dataset can still provide lively hand gestures that align with text controls, demonstrating the broad generalization capabilities of the BOTH57M dataset. Finally, we present a gallery of our inference results for the text/body-to-hand task in Fig. 13.

C. Applications

Our BOTH2Hands algorithm, trained on BOTH57M dataset, has demonstrated excellent generalization capabilities even on other human body datasets that only offer body-level descriptions. We conducted experiments by inputting body and text data from the HumanML3D [14] dataset into BOTH2Hands, successfully generating hand movements that are remarkably vivid. In a bid to challenge the generation of hand movements in more professional actions, we proceed to perform inference on InterHuman [29], a dataset encompassing professional movements such as ballet, boxing, and more. Fig. 14 illustrates how our method can effectively augment data for existing body and text datasets.

D. Limitation and Broader Impact

In this section, we discuss the border impact and limitations.

Broader Impact. We are the first to propose generat-

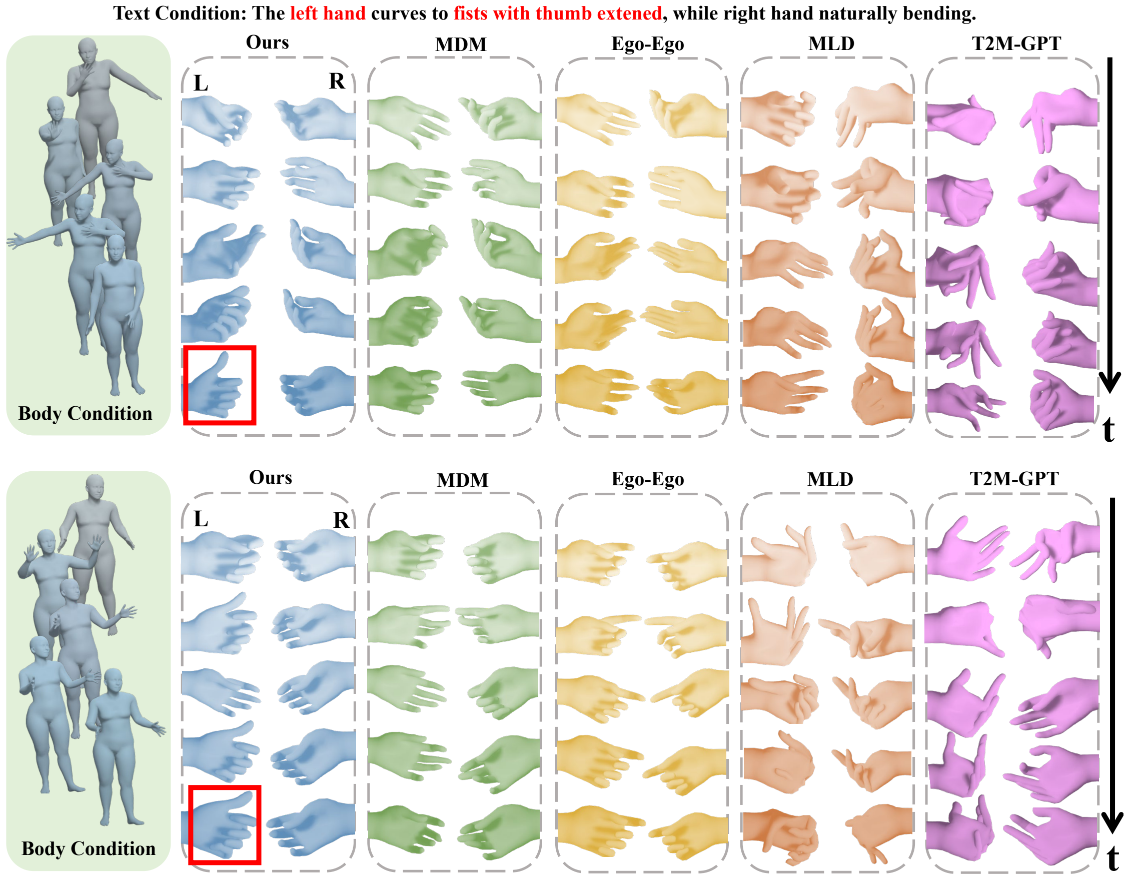

Text Condition: The left hand curves to fists with thumb extened, while right hand naturally bending.
Ours
MDM
Ego-Ego
MLD
T2M-GPT
L
R
Body Condition
t
Ours
MDM
Ego-Ego
MLD
T2M-GPT
L
R
Body Condition
t

Figure 11. More method evaluation results. Our method performs well under different body conditions.

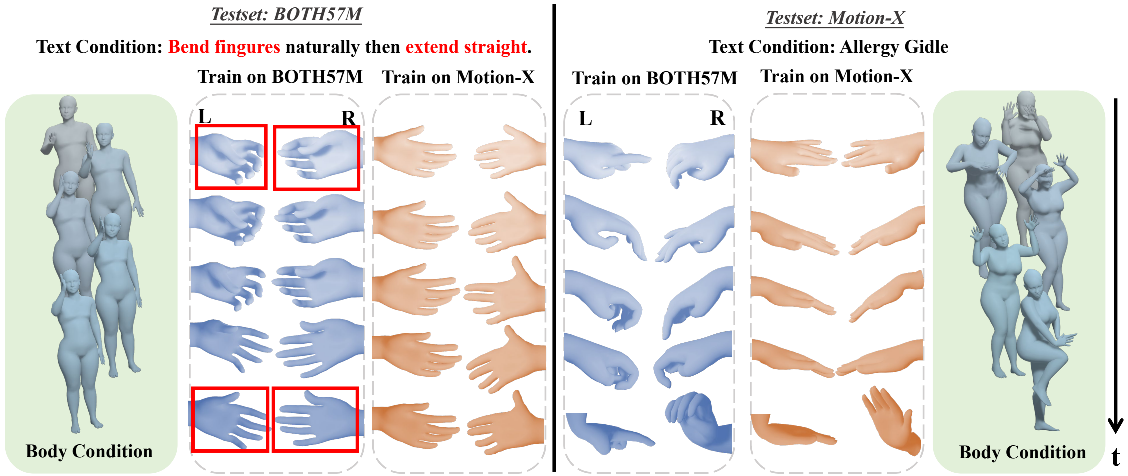


Figure 12. More dataset evaluation results. BOTH57M continues to exhibit robust generalization even under challenging motion and text conditions.

ing fine-grained two-hand motions through implicit control from real-world body input and user-defined explicit text control. This provides a direction that can be studied for user-customized motion generation. Furthermore, the pre-trained BOTH2Hands, based on the hand synthesis through the parallel diffusion structure and large-scale motion data training on the BOTH57M, can enhance data for large body motion datasets with textual descriptions

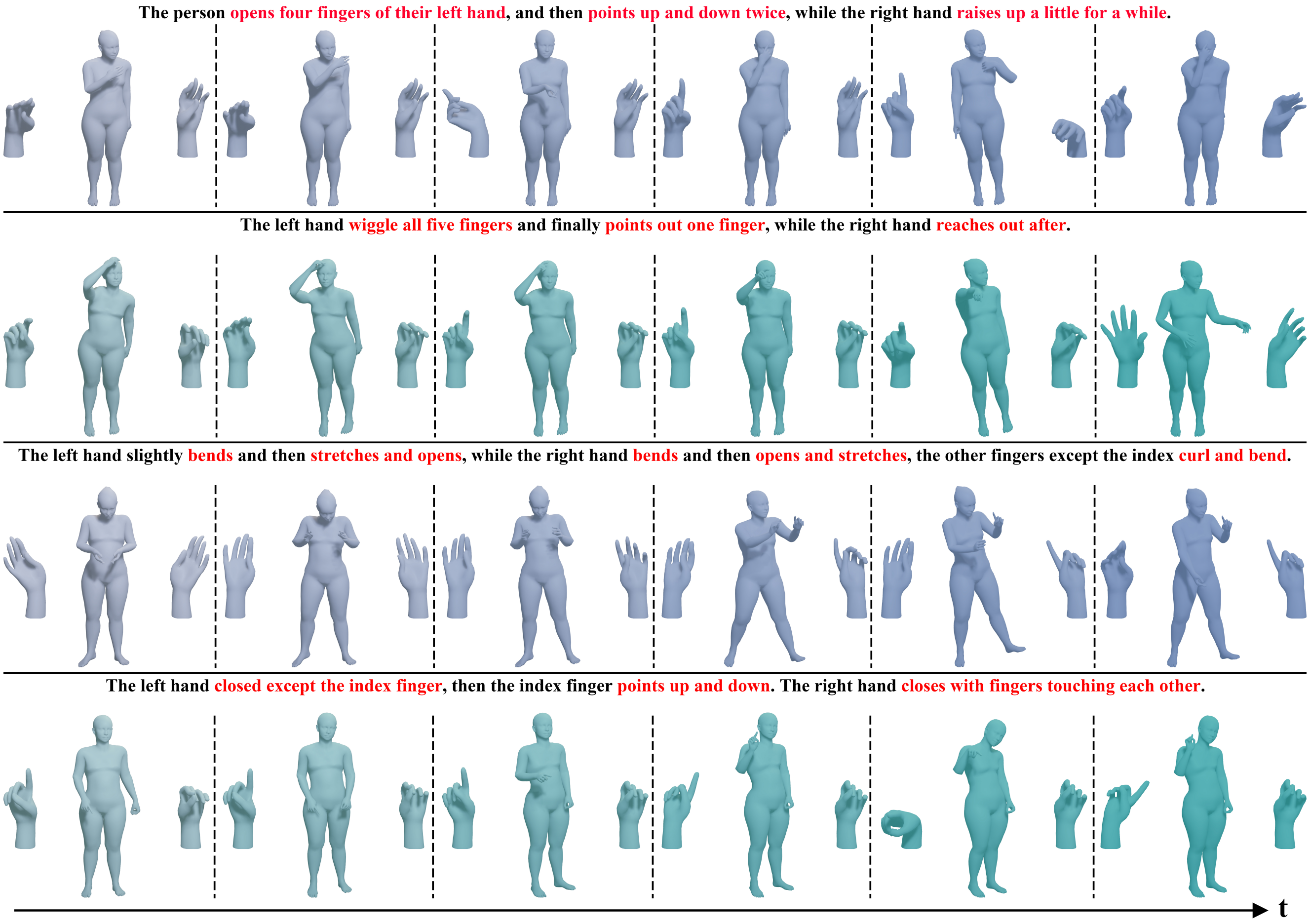
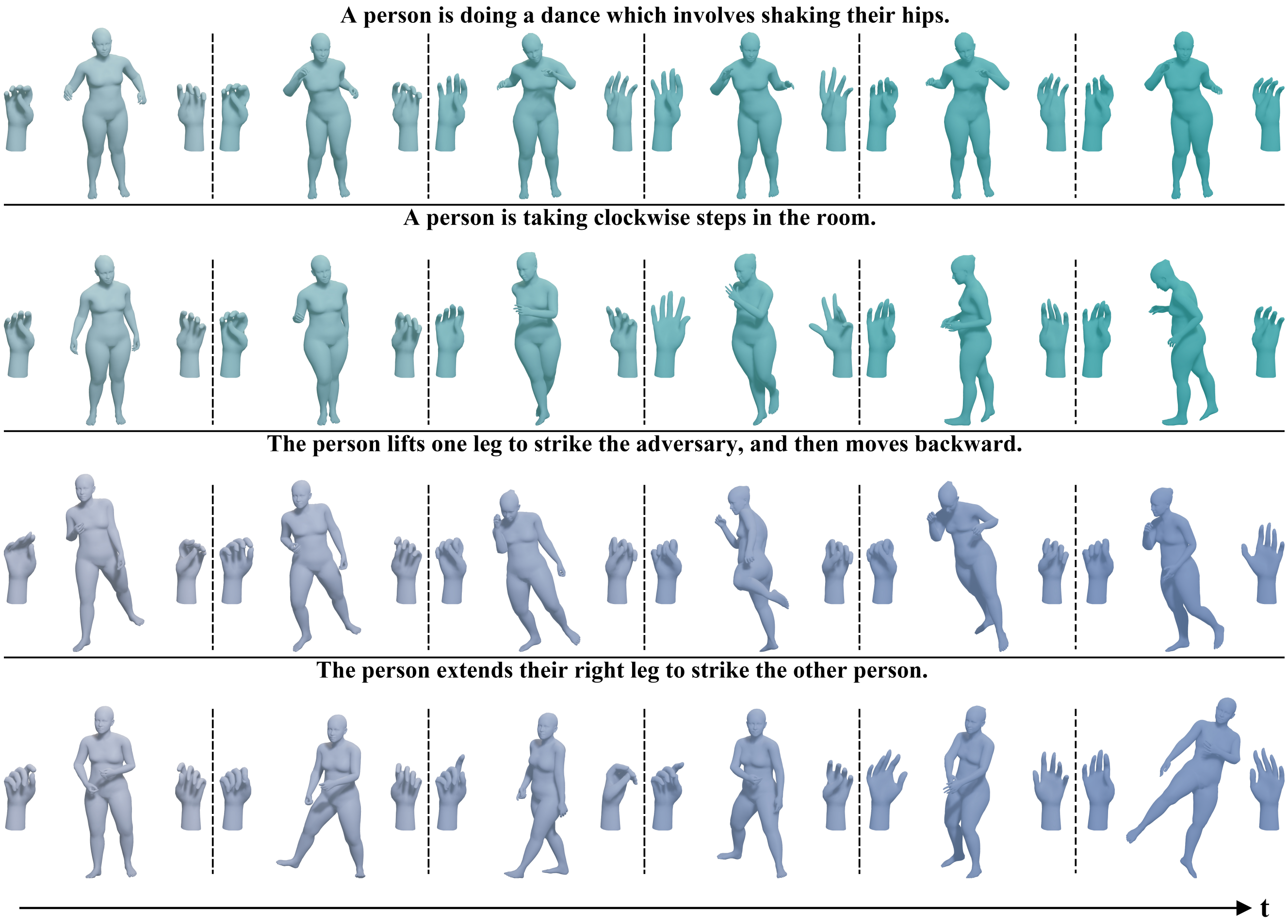

The person opens four fingers of their left hand, and then points up and down twice, while the right hand raises up a little for a while.
The left hand wiggle all five fingers and finally points out one finger, while the right hand reaches out after.
The left hand slightly bends and then stretches and opens, while the right hand bends and then opens and stretches, the other fingers except the index curl and bend.
The left hand closed except the index finger, then the index finger points up and down. The right hand closes with fingers touching each other.
t

Figure 13. Qualitative results generated by our BOTH2Hands algorithm. We showcase text prompts at the top of each motion. From left to right, the temporal order is indicated.

like HumanML3D [14] and InterHuman [29]. The research also contributes to balancing control signals during the multi-signal control diffusion process. Furthermore, text-controllable and body-aligned hand synthesis has potential in many practical application scenarios, such as Virtual Reality (VR), Augmented Reality (AR) games, animations, and human communication studies, etc. Fig. 15 provides the gallery of data examples captured by our dome with 32 synchronized high-resolution RGB cameras.

Limitation. While this work has achieved substantial advancements in the novel task of generating hand motions in alignment with body and text and offering an intuitive control method, it has limitations. First, the fine-grained spatial control provided by text descriptions is sufficient but lacks temporal alignment methods. Second, we focused on the connection between hand gestures and the body as a whole but did not take into account how different parts of the body might separately affect the hands. Third, more suitable metrics that can precisely reflect the alignment between hand, body, text, and other complex conditions should be presented, we welcome the community to focus on it and hope our contribution will push the field forward. Finally, although BOTH2Hands can generate vivid hand poses, it remains challenging to determine whether a two-handed interaction gesture is generated when the wrists are close enough. Hand-interaction poses represent a hugely different meaning from poses that do not interact. Future work may focus on aligning text control temporally and explore how individual body parts relate to hand movements, especially hand-to-hand interaction. Improved metrics are needed to better represent hand and multi-condition alignment.



A person is doing a dance which involves shaking their hips.

A person is taking clockwise steps in the room.

The person lifts one leg to strike the adversary, and then moves backward.

The person extends their right leg to strike the other person.

Figure 14. Inference on HumanML3D [14] and InterHuman [29]. We pretrain our BOTH2Hands algorithm on BOTH57M, and inference two-hand motions using text and body conditions of HumanML3D and InterHuman. Our results faithfully match the daily motion in HumanML3D and challenging professional motions in InterHuman.

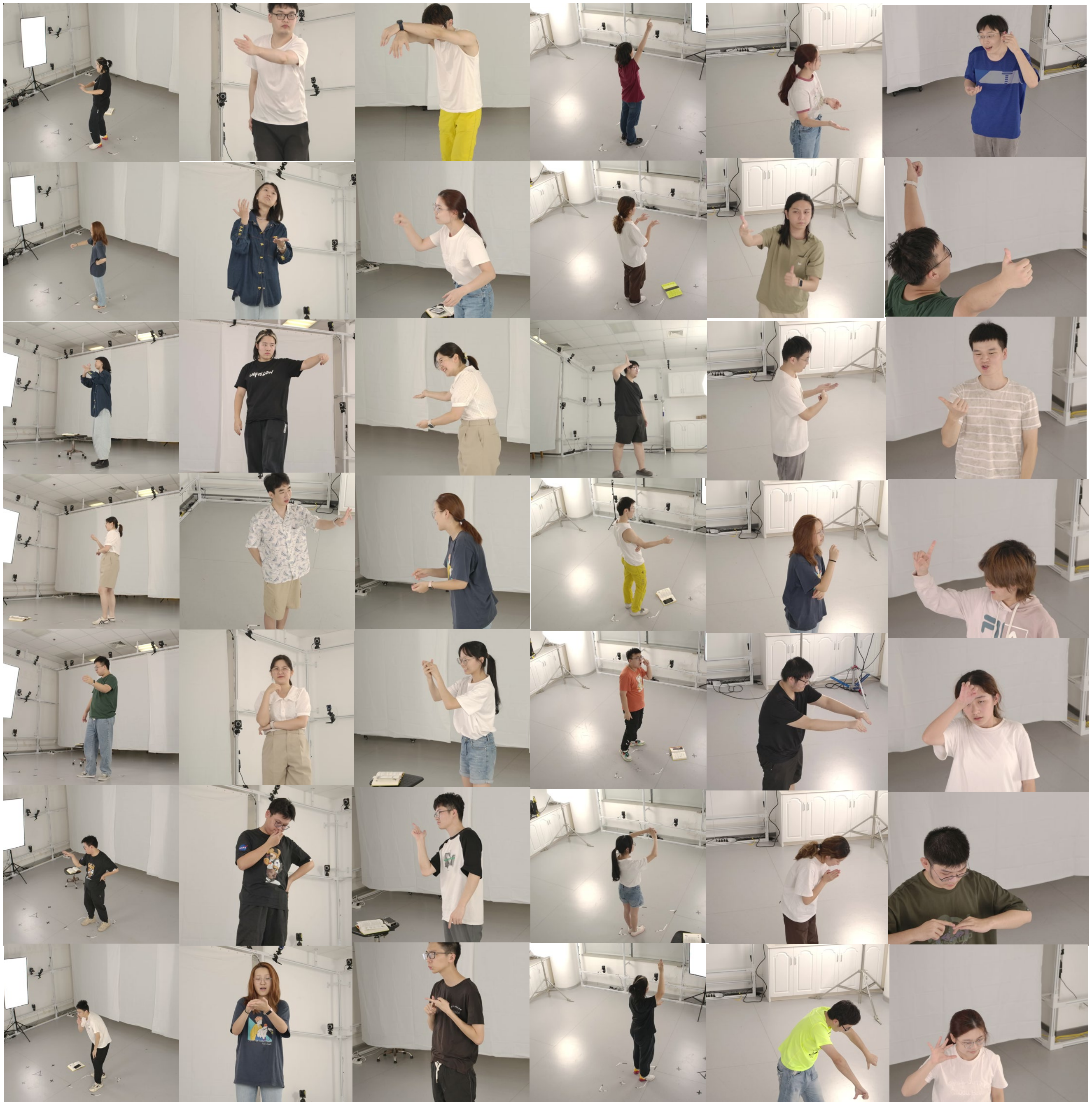


Figure 15. Data examples captured by our dome with 32 synchronized high-resolution RGB cameras. Our dataset includes a variety of body-hand motions under various daily scenes.