Shape My Moves: Text-Driven Shape-Aware Synthesis of Human Motions –Supplementary Material–

Ting-Hsuan Liao[1,2] Yi Zhou[2] Yu Shen[2] Chun-Hao Paul Huang[2] Saayan Mitra[2]
Jia-Bin Huang[1] Uttaran Bhattacharya[2]

[1] University of Maryland, College Park [2] Adobe Research
https://shape-move.github.io/

Overview

Our supplementary material covers the following content:
- Additional Implementation Details.
- Discussion on Evaluations.

A. Additional Implementation Details

We provide all implementation details of our data processing in Section A.1, shape data statistics of the dataset we use for our experiments in Section A.1, and the human perceptual study setup in Section A.3. Further, we discuss the principles behind the selection of evaluation data for assessing the Penetrate, Float, and Skate metrics in Section A.4, and the complete text inputs for qualitative results listed in the main manuscript in Section A.5.

A.1. Data Processing

Generating Additional Training Data. We elaborate on our process to generate additional shape data that we use to train the SA-VAE. We leverage the Attributes to Shape (A2S) model from SHAPY [7], which provides a method to link linguistic shape attributes with SMPL-X [18] shape coefficients. SHAPY [7] identifies 30 linguistic attributes, of which 12 are gender-specific attributes that apply either to females or males. By using a discrete 5-level Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree), the model can predict the beta shape parameters of SMPL-X [18].

To diversify the shape data for training, we generate 1,000 additional shapes for each gender. Initially, we randomly sample discrete values within the range [1,5] for each attribute. We then obtain the predicted SMPL-X [18] shape beta parameters from the A2S model. Given that the primary source in HumanML3D [9] is derived from the AMASS [1, 3, 5, 8, 11, 12, 14–17, 20, 23–25] dataset, which is extended from the SMPL-H [19] model, it is necessary to align the body types in our dataset accordingly. To achieve this, we employ the conversion tools provided by SMPL-X [18] to transform the SMPL-X [18] shape beta parameters into SMPL-H [19] beta parameters.

Unifying Gender-Specific Parameters. Within the entire training set, we include 2,000 additional shape beta parameters alongside the original 449 shapes from the HumanML3D dataset [9]. However, these shape beta parameters are gender-specific, implying that female and male models possess distinct body model topologies, which can hinder generalization in model training. Simply put, the same shape β parameter values can produce vastly different bodies for females and males. To address this issue, we propose converting all gender-specific shape β parameters to a neutral gender format. This approach allows all the shape β values to share the same shape space, thereby enhancing the generalization capability of our model.

Labeling Shape β Parameter Values. We use the largest current text-to-motion dataset, HumanML3D [9], for our experiments, which is composed of two large-scale datasets: HumanAct12 [4] and AMASS [1, 3, 5, 8, 11, 12, 14–17, 20, 23–25]. However, HumanAct12 [4] does not provide beta values, only the 3D joint locations. To obtain the complete training data, we use SMPLify [2] to fit the shape β parameters from the joint locations, obtaining approximate shape β values.

Constructing Shape Description Input. We list the shape description template in Table A.1, where $< \cdot >$ denotes a placeholder to be replaced with body measurement attributes. These attributes include height, arm and leg lengths, and the circumferences of the chest, waist, and hips. For each training data point, we randomly select one template to form the shape description, which we then com-

Table A.1. **Template for Shape Description Data Input.** This table provides a structured format for users to input detailed measurements of human body attributes. It includes fields for height, weight, body type, and specific measurements for chest, waist, hips, arms, and legs, ensuring comprehensive data collection for shape-aware motion synthesis.

1. A <gender> standing <height> cm tall, with a chest circumference of <chest> cm, waist circumference of <waist> cm, and hip circumference of <hip> cm. The length of the arm is <arm> cm and leg height is <leg> cm.
2. This individual, a <gender>, has a height of <height> cm. Their body measurements include a chest of <chest> cm, a waist of <waist> cm, and hips measuring <hip> cm. The arm extends <arm> cm and the legs are <leg> cm long.
3. The <gender> stands <height> cm tall. They have a chest circumference of <chest> cm, a waist of <waist> cm, and hips that measure <hip> cm. The arm and leg measurements are <arm> cm and <leg> cm, respectively.
4. Describing the <gender>: They are <height> cm in height, with body measurements that include a chest of <chest> cm, a waist of <waist> cm, and hips of <hip> cm. Additionally, their arm measures <arm> cm, and their leg height is <leg> cm.
5. Here is a <gender> with a height of <height> cm. Their measurements are as follows: chest <chest> cm, waist <waist> cm, and hips <hip> cm. The lengths of the arm and legs are <arm> cm and <leg> cm respectively.
6. You see a <gender> whose physical stature includes a height of <height> cm. Notable measurements are a chest circumference of <chest> cm, waist circumference of <waist> cm, and hip circumference of <hip> cm, with an arm length of <arm> cm and leg height of <leg> cm.
7. A <gender> with these dimensions: <height> cm tall, chest <chest> cm, waist <waist> cm, hips <hip> cm, arm <arm> cm, and leg height <leg> cm.
8. Consider a <gender> who is <height> cm tall. They have a chest of <chest> cm, waist of <waist> cm, and hips of <hip> cm. Their arm is <arm> cm long, and their legs measure <leg> cm in height.
9. Profile of a <gender>: Height of <height> cm, with a chest measurement of <chest> cm, waist of <waist> cm, and hips spanning <hip> cm. The arm and leg heights are <arm> cm and <leg> cm, respectively.
10. A detailed look at a <gender>: They stand <height> cm tall, and feature a chest of <chest> cm, a waist of <waist> cm, and hips measuring <hip> cm. Their arm length is <arm> cm, and the leg height is <leg> cm.

bine with the original motion description as the final text input for our method.

A.2. Shape Data Statistics

We report the shape data statistics for the HumanML3D dataset [9], which comprises 14,616 motion sequences totaling 28.59 hours of motion. The dataset includes 449 unique subjects sourced from AMASS [1, 3, 5, 8, 11, 12, 14–17, 20, 23–25], with a demographic distribution of 263 male and 186 female subjects. Additionally, the dataset contains 1,191 subjects from HumanAct12 [4]. Following previous works, we utilize the first 10 principal components of the shape parameters to construct the shape β. We present the mean and standard deviation of each principal component in the shape β in the dataset in Figure B.1, and for the dataset including the additional 2,000 shape data in Figure B.2, to illustrate the diversity of body shapes used for training.

A.3. Human Perceptual Study

We provide the layout shown to participants on Amazon Mechanical Turk during our perceptual study in Figure B.3. For the baseline methods that do not synthesize shape β parameters simultaneously, we employ SMPLify [2] to fit the shape β parameters and render those animations for the user study.

A.4. Evaluation Data

We describe our method for filtering evaluation data to compute the Penetrate, Float, and Skate metrics. Following the approach of HUMOS [22], we exclude motion sequences where the lowest joint is higher than 0.25 meters from the ground in at least 5 frames. This procedure ensures that all motions sequences we evaluate on are motions on the ground plane. We perform this filtering because motion sequences off the ground plane (which are typically supported by other objects and scene components, *e.g.*, stairs) create ambiguities in defining the above metrics.

A.5. Full Text Inputs for Qualitative Results

Here, we provide the full text input for the qualitative results we show in the main manuscript (Figure 4). We encourage reviewers to view additional comparative results on the webpage.

- *Top row.* Consider a man who is 182 cm tall. They have a chest of 101 cm, waist of 90 cm, and hips of 96 cm.

Table A.2. **Physical Plausibility Comparison** and highlight the best and second-best results.

Methods	Shape Input Capability	Arbitrary Length	Penetrate (cm) ↓	Float (cm) ↓	Skate (%) ↓	Bone Length Variances ↓
Real	–	–	$0.0^{\pm 0.000}$	$0.0352^{\pm 0.000}$	$8.110^{\pm 0.001}$	$0.0^{\pm 0.000}$
SA-VAE (Recon.)	–	–	$0.0289^{\pm 0.000}$	$0.2090^{\pm 0.000}$	$6.443^{\pm 0.001}$	$0.623^{\pm 0.000}$
TM2T [10]	✗	✓	$0.1485^{\pm 0.001}$	$0.2456^{\pm 0.001}$	$8.554^{\pm 0.001}$	$5.339^{\pm 0.032}$
T2M [9]	✗	✓	$0.0939^{\pm 0.001}$	$0.6805^{\pm 0.001}$	$4.250^{\pm 0.060}$	$1.352^{\pm 0.096}$
MLD [6]	✓	✗	$0.3091^{\pm 0.001}$	$0.6558^{\pm 0.011}$	$9.313^{\pm 0.153}$	$2.695^{\pm 0.053}$
MotionDiffuse [27]	✓	✗	$0.2401^{\pm 0.001}$	$0.2703^{\pm 0.001}$	$7.710^{\pm 0.010}$	$0.138^{\pm 0.002}$
MDM [21]	✓	✗	$0.1011^{\pm 0.001}$	$1.7101^{\pm 0.032}$	$8.523^{\pm 0.150}$	$0.666^{\pm 0.010}$
MotionGPT [13]	✓	✓	$0.6986^{\pm 0.017}$	$0.2245^{\pm 0.007}$	$7.889^{\pm 0.078}$	$2.271^{\pm 0.018}$
T2M-GPT [26]	✓	✓	$0.1789^{\pm 0.004}$	$0.5241^{\pm 0.001}$	$6.162^{\pm 0.044}$	$1.176^{\pm 0.007}$
Ours	✓	✓	$0.0268^{\pm 0.001}$	$0.2658^{\pm 0.008}$	$6.143^{\pm 0.123}$	$0.625^{\pm 0.002}$

Table A.3. **Guo *et al.* [9] Benchmark Comparison.** and highlight the best and second-best results.

Methods	Shape Input Capability	Arbitrary Length	RPrecision ↑			FID ↓	MMDist ↓	Diversity ↓
			Top1	Top2	Top3			
Real	-	-	$0.469^{\pm 0.002}$	$0.665^{\pm 0.002}$	$0.769^{\pm 0.003}$	$0.001^{\pm 0.000}$	$3.217^{\pm 0.007}$	$0.000^{\pm 0.000}$
SA-VAE (Recon.)	-	-	$0.454^{\pm 0.003}$	$0.645^{\pm 0.002}$	$0.749^{\pm 0.002}$	$0.125^{\pm 0.001}$	$3.308^{\pm 0.006}$	$0.101^{\pm 0.097}$
TM2T [10]	✗	✓	$0.374^{\pm 0.003}$	$0.559^{\pm 0.003}$	$0.673^{\pm 0.002}$	$1.671^{\pm 0.018}$	$3.843^{\pm 0.009}$	$0.937^{\pm 0.091}$
T2M [9]	✗	✓	$0.408^{\pm 0.003}$	$0.592^{\pm 0.003}$	$0.697^{\pm 0.002}$	$1.230^{\pm 0.023}$	$3.597^{\pm 0.007}$	$0.430^{\pm 0.058}$
MLD [6]	✓	✗	$0.383^{\pm 0.002}$	$0.571^{\pm 0.003}$	$0.680^{\pm 0.003}$	$0.882^{\pm 0.024}$	$3.736^{\pm 0.008}$	$0.020^{\pm 0.070}$
MotionDiffuse [27]	✓	✗	$0.426^{\pm 0.002}$	$0.616^{\pm 0.002}$	$0.723^{\pm 0.003}$	$0.563^{\pm 0.010}$	$3.392^{\pm 0.006}$	$0.320^{\pm 0.070}$
MDM [21]	✓	✗	$0.317^{\pm 0.006}$	$0.490^{\pm 0.006}$	$0.599^{\pm 0.007}$	$0.461^{\pm 0.045}$	$4.180^{\pm 0.035}$	$0.320^{\pm 0.083}$
MotionGPT [13]	✓	✓	$0.128^{\pm 0.002}$	$0.208^{\pm 0.002}$	$0.271^{\pm 0.002}$	$1.020^{\pm 0.034}$	$7.055^{\pm 0.002}$	$0.389^{\pm 0.095}$
T2M-GPT [26]	✓	✓	$0.394^{\pm 0.003}$	$0.576^{\pm 0.003}$	$0.683^{\pm 0.002}$	$0.269^{\pm 0.010}$	$3.710^{\pm 0.008}$	$0.190^{\pm 0.061}$
ShapeMove (Ours)	✓	✓	$0.413^{\pm 0.008}$	$0.601^{\pm 0.005}$	$0.705^{\pm 0.005}$	$0.198^{\pm 0.015}$	$3.533^{\pm 0.016}$	$0.117^{\pm 0.131}$

Table A.4. **Multimodality.**

Method	MModality
T2M-GPT	$1.428^{\pm 0.055}$
MotionGPT	$8.534^{\pm 0.232}$
Ours	$1.894^{\pm 0.114}$

Table A.5. **Additional VAE Comparisons**

Method	FID ↓	Bone Length Diff. (mm) ↓	Jitter Diff. (m/s^2) ↓
MLD [6]	0.239	85.26	**22.58**
SA-VAE (Ours)	**0.125**	**45.88**	31.49

Their arm is 56 cm long, and their legs measure 77 cm in height. The person demonstrates a motion like a person walks forward while twisting their torso side to side.

- *Bottom row.* This individual, a female, has a height of 168 cm. Their body measurements include a chest of 103 cm, a waist of 91 cm, and hips measuring 111 cm. The arm extends 53 cm and the legs are 71 cm long. The person demonstrates a motion like a person walks backwards and stops.

B. Discussion on Evaluations

We elaborate further on the evaluation metrics we use and the corresponding results. For text-motion alignment and quality, we adhere to the evaluation protocol proposed in Guo *et al.* [9]. We compute all results with a 95% confidence interval, obtained from 20 repeated runs, and report the results in Table A.2 and Table A.3.

We train the text and motion encoder to be shape-aware, following the guidelines in Guo *et al.* [9]. The text encoder utilizes a self-defined vocabulary. Therefore, we retain its original settings and input only motion descriptions. Consequently, in the Guo *et al.* [9] benchmark, the R-Precision

and Multimodal Distance metrics are designed to assess the correlation between motion descriptions and the generated motions.

Note that for the diversity metric, closer values to real motions indicate better performance. Thus, the term listed in the paper represents the absolute difference from the real motion. Another metric proposed in Guo *et al*. [9] for measuring diversity is MultiModality, which computes the distance between motion features. We present the results in Table A.4. Higher values indicate better performance on this metric. However, we find that it might not accurately reflect diversity, as indicated by MotionGPT [13] scoring significantly higher than others. Since MultiModality is computed by generating multiple motion sequences for a single text description and then calculating the average Euclidean distances between paired motion features, we suspect that the high scores may be due to poor alignment with the text input, as suggested by the low R-precision scores. Therefore, we have omitted this metric in the main manuscript but included it in the supplementary material for completeness.

We also extend our evaluation to include non-VQ VAEs for a more comprehensive assessment. In comparisons with the MLD VAE [6] (Table A.5), we observe that while non-quantizing VAEs might capture finer motion details (jitter difference), they underperform in encoding shape information (bone length difference) and overall motion quality (FID).

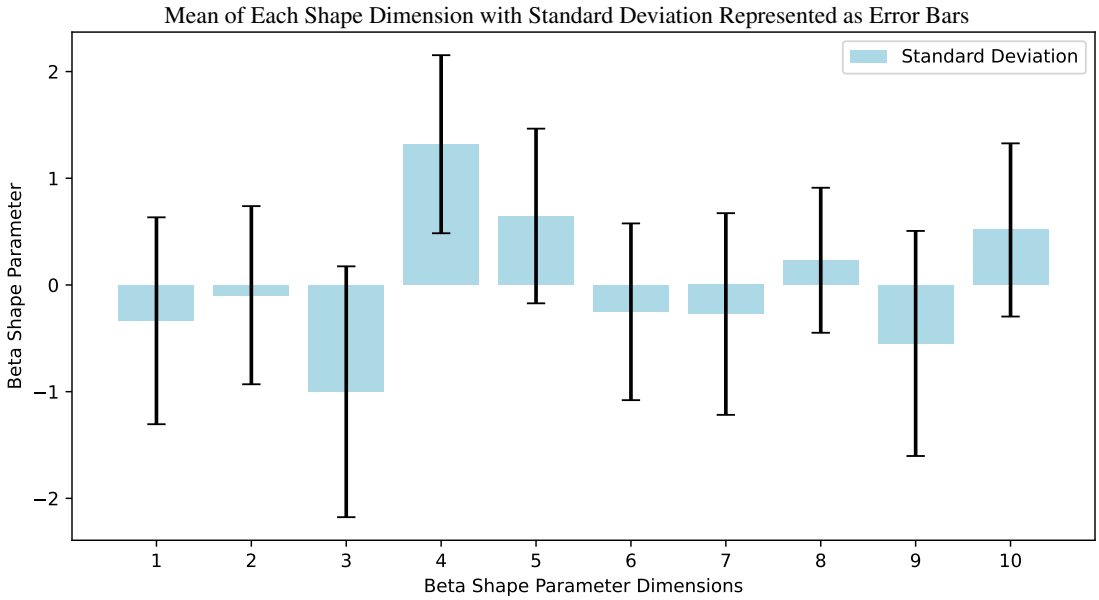

Figure B.1. **Shape Data Statistics in HumanML3D.** We show the mean and standard deviations of the shape β parameters to indicate the diversity of body shapes available in the base dataset.

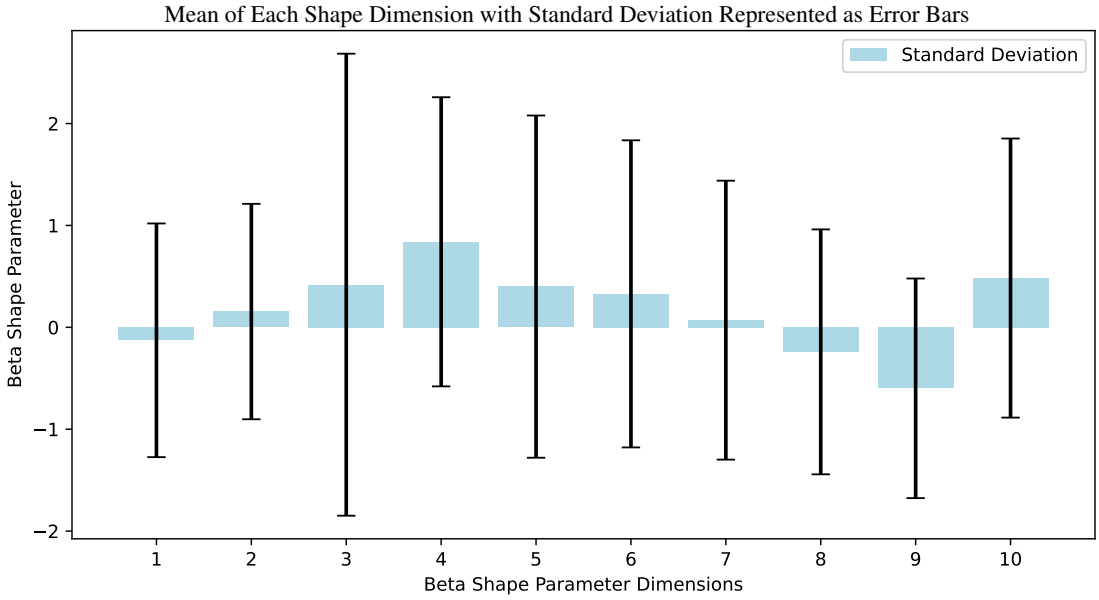

Figure B.2. **Shape Data Statistics in HumanML3D and Additional Synthetic Shape Data.** We show the mean and standard deviations of the shape β parameters to indicate the overall diversity of body shapes in the augmented dataset.

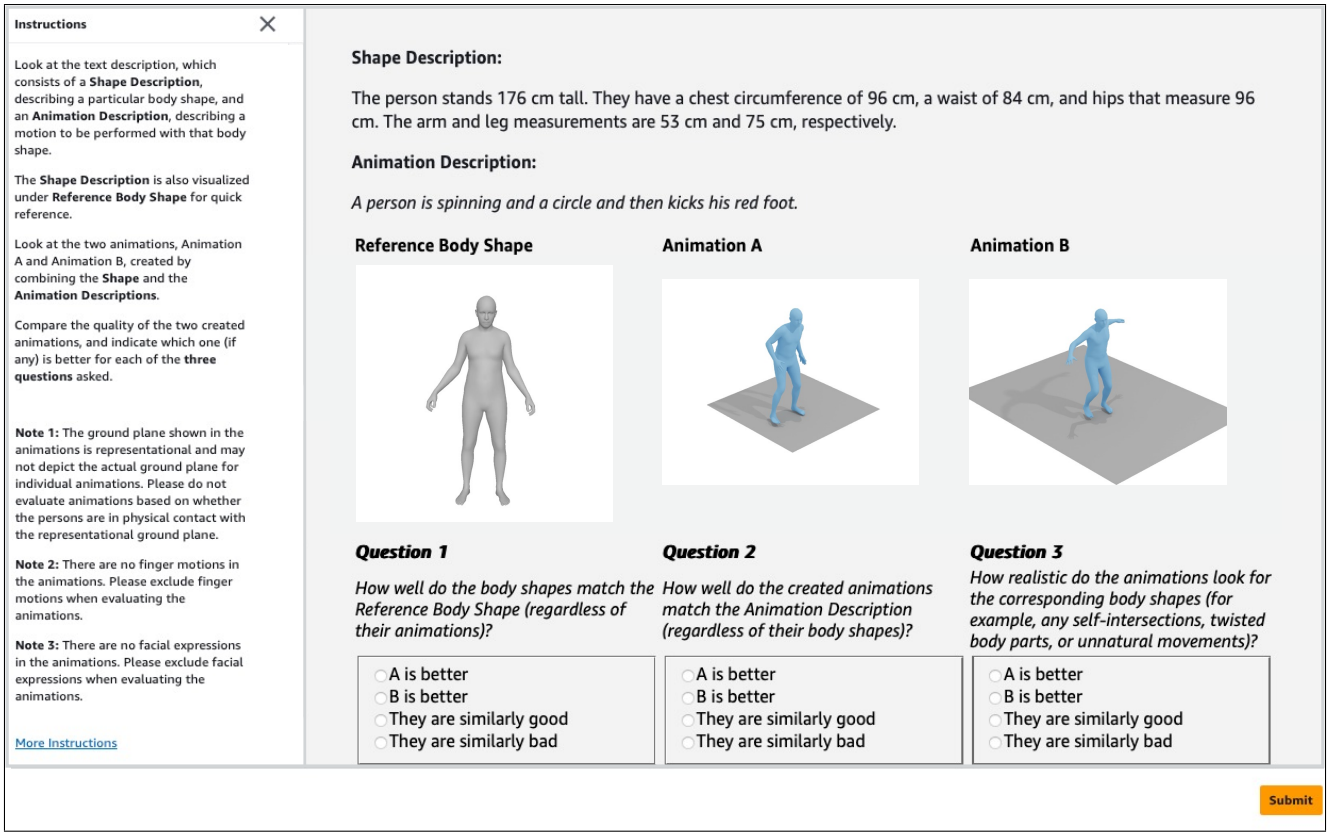

Figure B.3. **Layout of the Perceptual Study.** We show the layout with the collapsible menu item "Instructions" on the left and the body of the study on the right. The body consists of the text descriptions at the top, the reference body shape (ground truth body shape corresponding to the shape description) and the two animations to compare in the middle, and the Q&A block at the bottom.

References

[1] Ijaz Akhter and Michael J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *CVPR*, 2015. 1, 2

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, 2016. 1, 2

[3] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering human bodies in motion. In *CVPR*, 2017. 1, 2

[4] Judith Butepage, Michael J Black, Danica Kragic, and Hedvig Kjellstrom. Deep representation learning for human motion prediction and classification. In *CVPR*, 2017. 1, 2

[5] Carnegie Mellon University. CMU MoCap Dataset, 2003. 1, 2

[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 *CVPR*, 2023. 3, 4

[7] Vasileios Choutas, Lea Müller, Chun-Hao P Huang, Siyu Tang, Dimitrios Tzionas, and Michael J Black. Accurate 3d body shape regression using metric and semantic attributes. In *CVPR*, 2022. 1

[8] Saeed Ghorbani, Kimia Mahdaviani, Anne Thaler, Konrad Kording, Douglas James Cook, Gunnar Blohm, and Nikolaus F. Troje. MoVi: A large multipurpose motion and video dataset. *arXiv preprint arXiv: 2003.01888*, 2020. 1, 2

[9] 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 *CVPR*, 2022. 1, 2, 3, 4

[10] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *ECCV*, 2022. 3

[11] Fabian Helm, Nikolaus Troje, Mathias Reiser, and Jörn Munzert. Bewegungsanalyse getäuschter und nicht-getäuschter 7m-würfe im handball. *47. Jahrestagung der Arbeitsgemeinschaft für Sportpsychologie, Freiburg.*, 2015. 1, 2

[12] Ludovic Hoyet, Kenneth Ryall, Rachel McDonnell, and Carol O'Sullivan. Sleight of hand: Perception of finger motion from reduced marker sets. In *ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, 2012. 1, 2

[13] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *NIPS*, 2024. 3, 4

[14] Matthew M. Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 33(6):220:1–220:13, 2014. 1, 2

[15] Eyes JAPAN Co. Ltd. Eyes Japan MoCap Dataset, 2008.

[16] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 2019.

[17] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation mocap database hdm05. Technical Report CG-2007-2, Universität Bonn, 2007. 1, 2

[18] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019. 1

[19] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *SIGGRAPH Asia*, 36(6), 2017. 1

[20] L. Sigal, A. Balan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1):4–27, 2010. 1, 2

[21] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023. 3

[22] Shashank Tripathi, Omid Taheri, Christoph Lassner, Michael Black, Daniel Holden, and Carsten Stoll. Humos: Human motion model conditioned on body shape. In *ECCV*, 2024. 2

[23] Matt Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total Capture: 3d human pose estimation fusing video and inertial sensors. In *BMVC*, 2017. 1, 2

[24] ACCAD/Ohio State University. Advanced Computing Center for the Arts and Design, 2024.

[25] Simon Fraser University and National University of Singapore. SFU Motion Capture Database, 2011. 1, 2

[26] 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 *CVPR*, 2023. 3

[27] 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. 3