—Supplementary Material—
Motion-X: A Large-scale 3D Expressive Whole-body Human Motion Dataset

Jing Lin[1,2*‡], **Ailing Zeng**[1*†], **Shunlin Lu**[1,3*‡],
Yuanhao Cai[2], **Ruimao Zhang**[3], **Haoqian Wang**[2], **Lei Zhang**[1]
[1]International Digital Economy Academy (IDEA)
[2]Tsinghua University, [3]The Chinese University of Hong Kong, Shenzhen
`https://motion-x-dataset.github.io`

Contents

---

[*] Equal Contribution, [‡] Work done during an internship at IDEA
[†] Corresponding author

37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks.

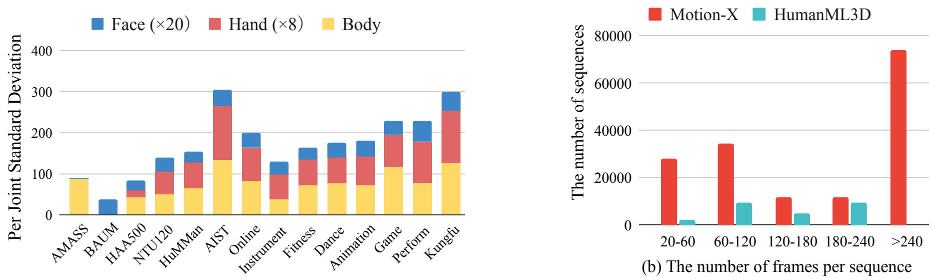
Appendix

A Motion-X: Additional Details

In this section, we provide more details about *Motion-X* that are not included in the main paper due to space limitations, including statistic analyses, data preprocessing, and motion augmentation mechanism.

A.1 Statistic Analyses

Fig. 1(a) shows each sub-dataset standard deviation of body, hand, and face joints. Our dataset has a large diversity of the hand and face joints, filling the gap of the previous body-only dataset in terms of expressiveness. Besides, as shown in Fig. 1(b), *Motion-X* provides a large volume of long motion (> 240 frames), which will be beneficial for long-term motion generation.



(b) The number of frames per sequence

Figure 1: Statistics of motion diversity and length.

A.2 Processing of Each Sub-dataset

We gather 81.1K motion sequences from seven existing datasets and a large volume of online videos with the proposed annotation pipeline. As shown in Tab. 1, due to the lack of comprehensive annotations from their original datasets, we provide well-annotated whole-body motion, comprehensive semantic labels, and whole-body pose descriptions for all datasets. Here we introduce more details about each sub-dataset's data preprocessing.

Dataset	Clip Number	Frame Number	Motion Annotation	Text Annotation	RGB
AMASS [15]	26.3K	5.4M	B,H	S	✗
AIST [35]	1.4K	0.3M	B	S	✓
HAA500 [9]	5.2K	0.3M	-	S	✓
HuMMan [6]	0.7K	0.1M	B	S	✓
GRAB [33]	1.3K	0.4M	B,H	S	✗
EgoBody [44]	1.0K	0.4M	B,H	-	✓
BAUM [40]	1.4K	0.2M	F	S	✓
IDEA400*	12.5K	2.6M	-	-	✓
Online Videos*	32.5K	6.0M	-	-	✓
Motion-X	81.1K	15.6M	B,H,F	S,P	✓

Table 1: Statistics of sub-datasets. B, H, and F are body, hand, and face. S and P are semantic and pose texts. Please note that the semantic text annotations are different from different datasets. Some action datasets [22, 25, 9, 40] only provide action-level texts instead of sentences. Most semantic labels are annotated manually, making specific domain descriptions (e.g., specific instruments) hard to label precisely. * denotes videos are collected in this work.

AMASS [25] is the existing largest-scale motion capture dataset, which provides body motions and almost static hand motions. To fill in the face parameters, we extract facial expressions from the facial dataset BAUM [40] with the SOTA facial reconstruction method EMOCA [10] and perform a data augmentation (in Sec. A.3). For the text labels, we utilize the semantic labels from HumanML3D [15] and annotate the pose description with our whole-body pose captioning module.

IDEA400 is a high-quality and expressive motion dataset recorded by ourselves, providing 13K motion sequences and 2.6M frames. Inheriting the NTU120 [22] categories, we expand them to 400 actions with additional human self-contact motions, human-object contact motions, and expressive whole-body motions (e.g., rich facial expressions and fine-grained hand gestures). There are 36 actors with diverse appearance and clothing. For each action, we have the actors perform three times standing, three times walking, and four times sitting, a total of ten times. We annotate the SMPL-X format pseudo labels with the proposed motion annotation pipeline, which can generate high-quality whole-body motions. To obtain the semantic labels, we use the designed action labels and expand them with the large language model (LLM) [18]. The pose descriptions are annotated with our whole-body pose captioning method. In this subset, we provide monocular videos, the body keypoints, SMPL-X parameters, and action labels. Please see the video in our website for more details and visualization.

AIST [35] is a large-scale dance dataset with multi-view videos and dance genres labels. Instead of using the body-only SMPL annotations from AIST++ [20], we annotate the whole-body motion via our motion annotation pipeline. We obtain the semantic labels by expanding the dance genres label and providing frame-level pose descriptions.

HAA500 [9] is a large-scale human-centric atomic action dataset with manually annotated labels and videos for action recognition. It contains 500 classes with fine-grained atomic action labels, covering sports, playing musical instruments, and daily actions. However, it does not have the motion labels. We annotate the 3D whole-body motion with our pipeline. We use the provided atomic action label as semantic labels. Besides, we input the video into video-LLaMA [42] and filter the human action descriptions as supplemental texts. Pose description is generated by our automatic pose annotation method.

HuMMan [6] is a human dataset with multi-modality data, including multi-view videos, keypoints, SMPL parameters, action labels, etc. It does not provide whole-body pose labels. We estimate the SMPL-X parameters with our annotation pipeline. Besides, we expand the action label with LLM into semantic labels and use the proposed captioning pipeline to obtain pose descriptions.

GRAB [33] is human grasping dataset with body and hand motion. Meanwhile, it provides text descriptions of each grasping motion without corresponding videos. Therefore, similar to AMASS, we extract facial expressions from BAUM to fill in the facial expression. We use the provided text description as semantic labels and annotate pose descriptions based on the SMPL-X parameters.

EgoBody [44] is a large-scale dataset capturing ground-truth 3D human motions in social interactions scenes. It provides high-quality body and hand motion annotations, lacking facial expression. Thus, we perform a motion augmentation to obtain expressive whole-body motions. Since EgoBody does not provide text information, we manually label the semantic description using the VGG Image Annotation (VIA) [11] and annotate the pose description with the automatic pose captioning pipeline.

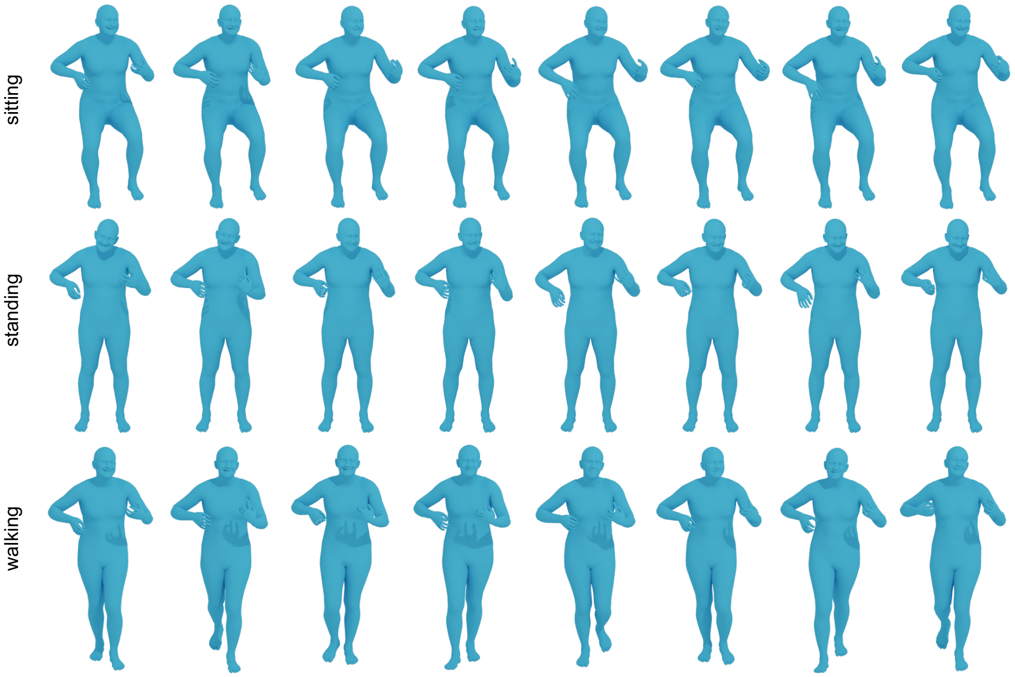
BAUM [40] is a facial dataset with 1.4K audio-visual clips and 13 emotions. We annotate facial expressions from BAUM with the SOTA face reconstruction method EMOCA.

Online Videos. To improve the richness of appearance and motion diversity, especially on professional motions, we collect 33K monocular videos from online sources, covering various real-life scenes. We design action categories as motion prompts and input them into LLM. Then, we collect videos from online sources based on the answer of LLM, after which we filter the candidate videos by transition detection and annotate the whole-body motion, semantic label and pose description for the selected videos.

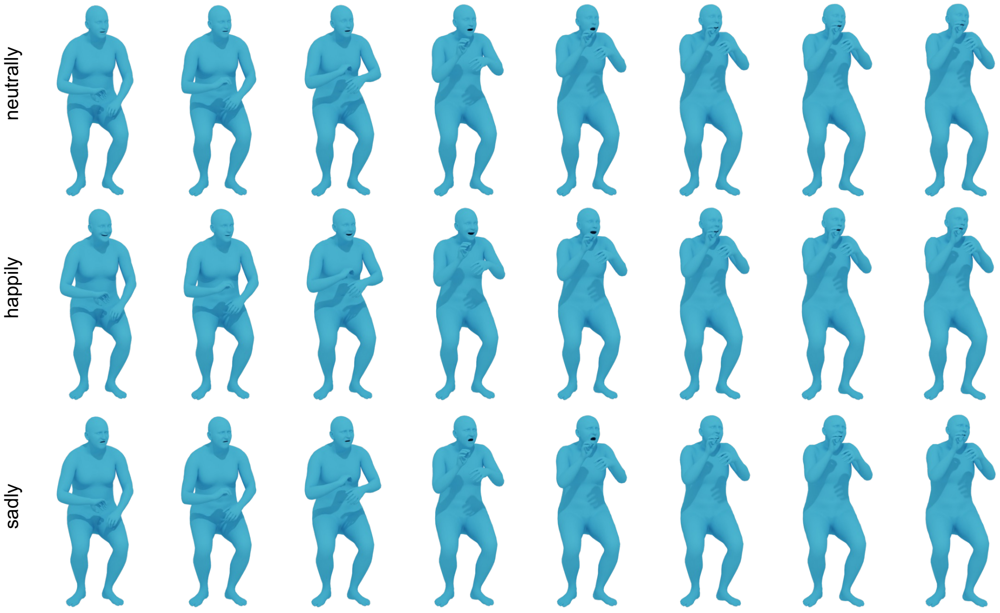
A.3 Motion Augmentation Mechanism

Lower-body Motion Augmentation. *Motion-X* contains some upper-body videos collected from online videos, like the videos in UBody [21], where the lower-body part is invisible. Estimating accurate lower-body motions and global trajectories for these videos is challenging. Thanks to the precise low-body motions provided in AMASS, we can simply perform a lower-body motion augmentation for these sequences, i.e., selecting the closest motion from AMASS based on the SMPL-X parameters and replacing the lower-body motion with it. Meanwhile, we incorporate relevant keywords (e.g., sitting, standing, walking) in the text descriptions. Fig. 2(a) depicts three plausible lower-body augmentations for the motion sequence with the semantic label "a person is playing the guitar happily."

Facial Expression Augmentation. As shown in Tab. 1, the motion capture datasets AMASS, GRAB, and EgoBody do not provide facial expressions. Thus, we perform a facial expression augmentation for these motions by randomly selecting a facial expression sequence from the BAUM [40] dataset to fill the void and incorporating emotion labels (e.g., happy, sad, and surprise) in the semantic description. We perform interpolation for the selected sequence to ensure the same length as the original motion. An example of face expression augmentation is illustrated in Fig. 2(b).



(a) Lower-body Augmentation of “A person is playing the guitar happily”.



(b) Facial Expression Augmentation of “A man is drinking from a straw”.

Figure 2: Illustration of two motion augmentation methods: (a) lower-body augmentation. (b) facial expression augmentation.

B More Annotation Visual Results

In this section, we present some visual results of the 2D keypoints, SMPL-X parameters, and motion sequences to show the effectiveness of our proposed motion annotation pipeline.

B.1 2D Keypoints

As the main paper claims, we propose a hierarchical Transformer-based model for 2D keypoints estimation. To demonstrate the superiority of our method, we compare it with two widely used methods, Openpose [7] and MediaPipe [41]. We use the PyTorch implementation of Openpose and only estimate the body and hand keypoints as it does not provide the face estimator. As shown in Fig. 3, Openpose and MediaPipe can not achieve accurate results in some challenging poses. Besides, there exists severe missing detection of hands for Openpose and MediaPipe. In contrast, our method performs significantly better, especially the hand keypoint localization.

B.2 SMPL-X Parameters

To register accurate SMPL-X parameters, we elaborately design a learning-based fitting method with several training loss functions. We compare our method with two SOTA learning-based methods, Hand4Whole [29] and OSX [21]. As shown in Fig. 4, our method achieves a much better alignment result than the other models, especially on some difficult poses, which benefits from the iterative fitting process. Notably, Hand4Whole [29] and OSX [21] can only estimate the local positions without optimized global positions, which will suffer from unstable and jittery global estimation. Furthermore, we compare with the widely used fitting method SMPLify-X [30], using their officially released codes, in Fig. 5. Our method is more robust than SMPLify-X and can obtain better results about physically plausible poses, especially in challenging scenes (e.g., hard poses, low-resolution inputs, heavy occlusions). The results from the side view demonstrate that our method can properly deal with depth ambiguity and avoid the lean issue.

B.3 Motion Sequences

To highlight the expressiveness and diversity of our proposed motions, we illustrate examples of the same semantic label, like *dance ballet*, with six motion styles in Fig. 6. This one-to-many (text-to-motion) information can benefit the diversity of motion generation. Then, we demonstrate more motion visualization in Fig. 7 and 8 for different motion scenes. These motions show different facial expressions, hand poses, and body motions.

C Experiment

C.1 Experiment Setup

Motion Representation. To capture the 3D expressive whole-body motion, we use SMPL-X [30] as our motion representation. A pose state is formulated as:

$$\mathbf{x} = \{\theta_b, \theta_h, \theta_f, \psi, \mathbf{r}\}. \tag{1}$$

Here, $\theta_b \in \mathbb{R}^{22\times 3}$ and $\theta_h \in \mathbb{R}^{30\times 3}$ denote the 3D body rotations and hand rotations. $\theta_f \in \mathbb{R}^3$ and $\psi \in \mathbb{R}^{50}$ are the jaw pose and facial expression. $\mathbf{r} \in \mathbb{R}^3$ is the global translation.

Evaluation Metrics. We adopt the same evaluation metrics as [15, 17], including Frechet Inception Distance (FID), multimodality, diversity, R-precision, and multimodal distance. We pretrain a motion feature extractor and a text feature extractor for the new motion presentation with contrastive loss to map the text and motion into feature space and then evaluate the distance between the text-motion pairs. For each generated motion, its ground-truth text description and 31 mismatched text description randomly selected from the test set compose a description pool. We rank the Euclidean distances between the generated motion and each text in the pool and then calculate the average accuracy at the top-k positions to derive R-precision. Multimodal distance is computed as the Euclidean distance between the feature vectors of generated motion and its corresponding text description in the test set. Additionally, We include the average temporal standard deviation as a supplementary metric to evaluate the diversity and temporal variation of whole-body motion.

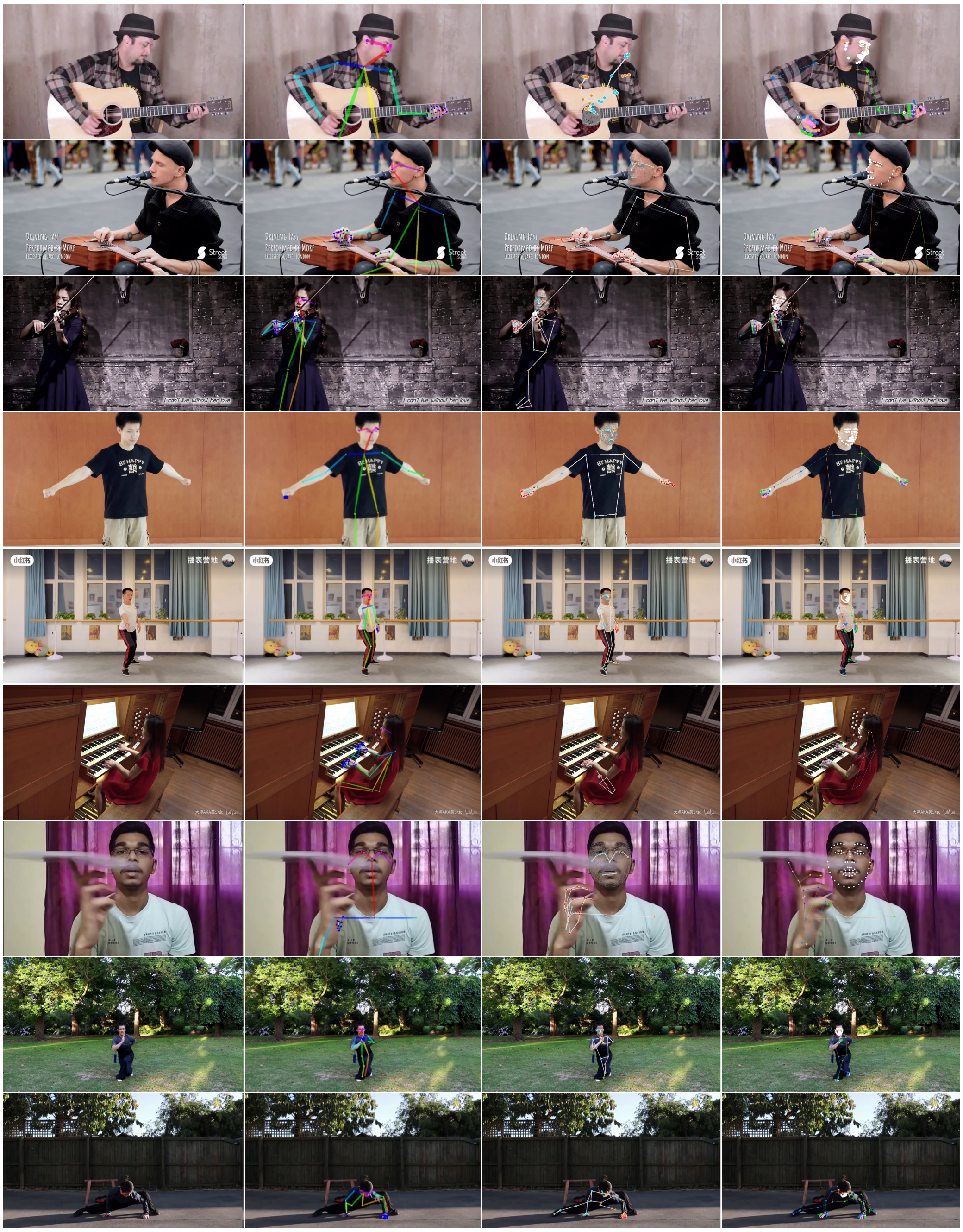

(a) Input Image (b) Openpose (c) MediaPipe (d) Ours

Figure 3: Comparisons of the 2D keypoints annotation quality with widely used methods Openpose [7] and MediaPipe [41].

Computational Costs. We use 8 NVIDIA A100 GPUs for motion annotation and 4 GPUs for motion generation experiments. It takes about 72 hours to annotate 1M frames with our annotation pipeline.

C.2 More Ablation Study

More Comparisons with HumanML3D. Previous motion generation datasets are limited in expressing rich hand and face motions, as they only contain body and minimal hand movements. To

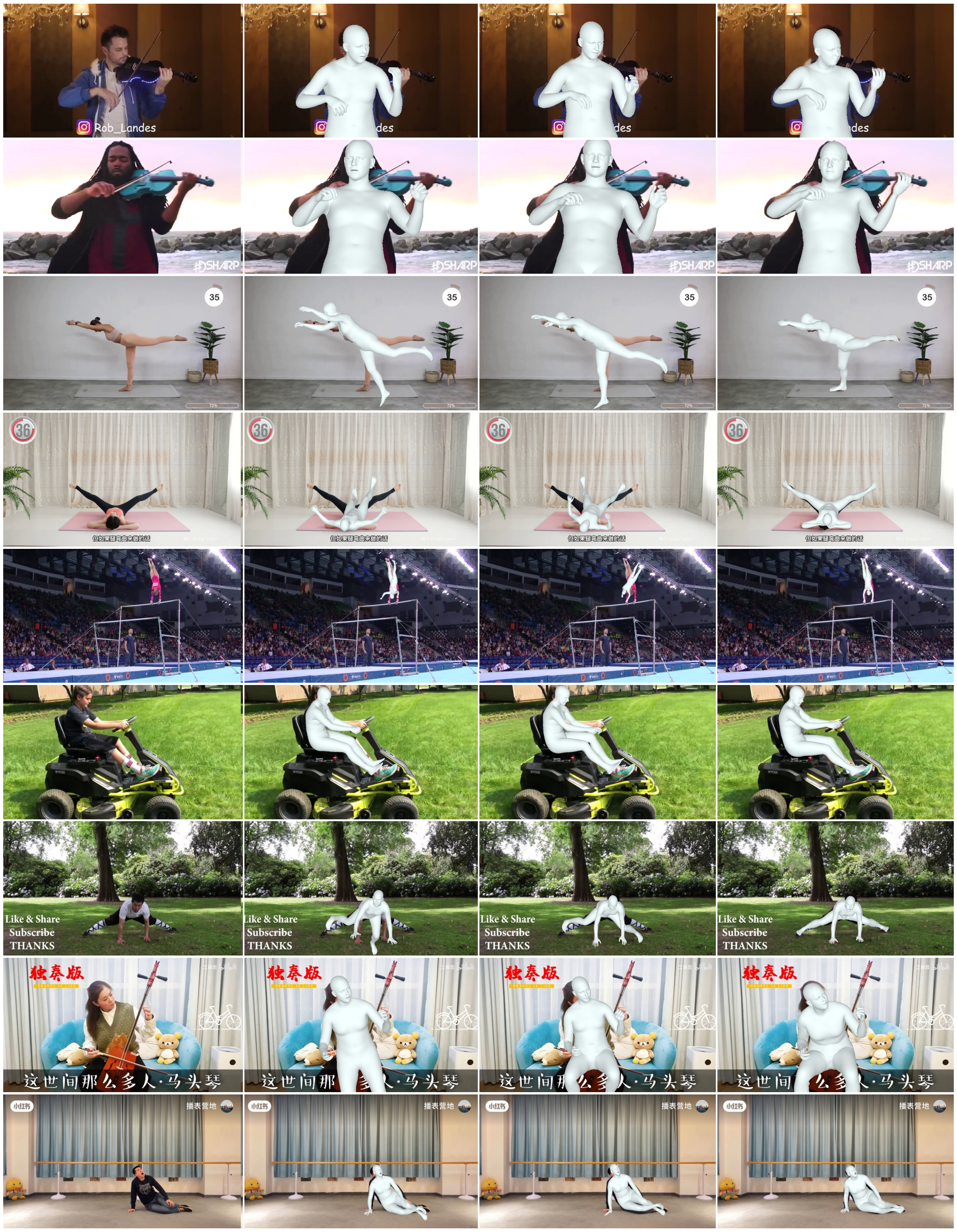

(a) Input Image
(b) Hand4Whole
(c) OSX
(d) Ours

Figure 4: Comparisons of the 3D SMPL-X annotation quality with SOTA methods Hand4Whole [29] and OSX [21].

demonstrate the expressiveness of our dataset, we conduct a comparison between HumanML3D and *Motion-X* on face, hand, and body, separately. Specifically, we train MLD [8] on each dataset and evaluate the diversity of generated motions and ground-truth motions by computing the average temporal standard deviation of the SMPL-X parameters and joint positions. The SMPL-X parameters include body poses, hand poses, and facial expressions. Body, hand, and face joint positions are represented as root-relative, wrist-relative, and neck-relative, respectively. We randomly choose 300 generated samples from the validation set and repeat the experiment 10 times to report the average

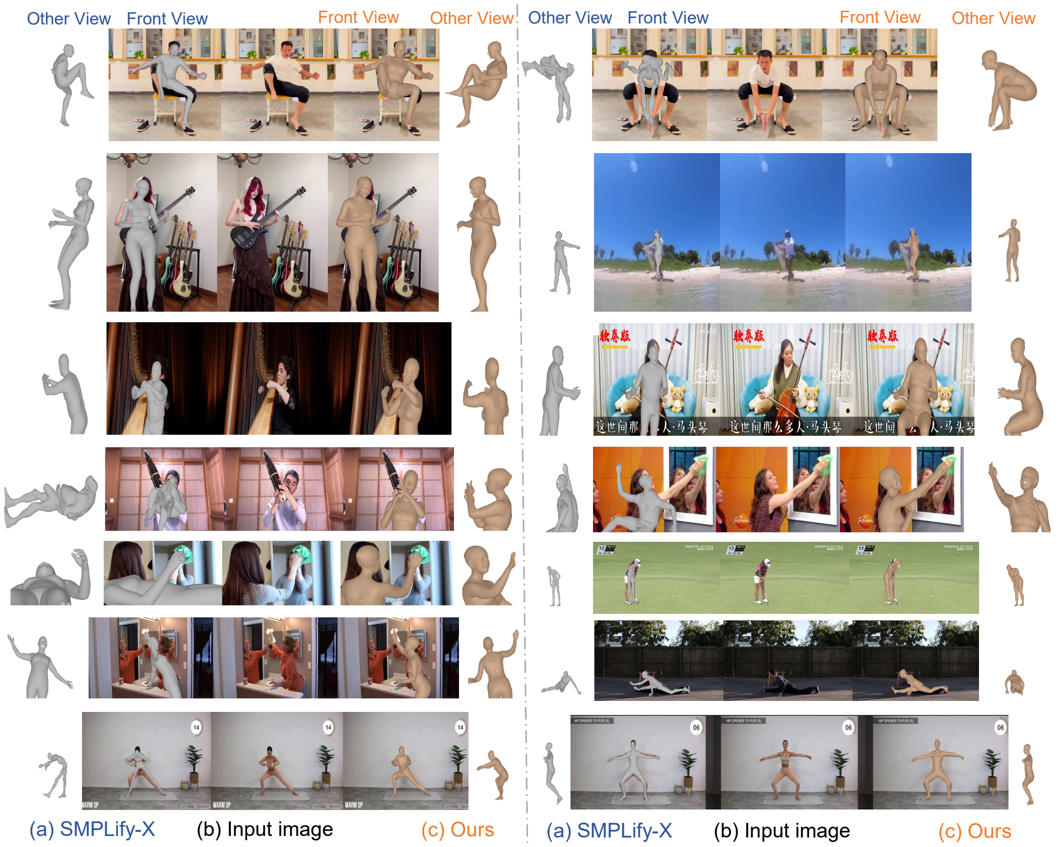


Figure 5: Comparisons of the 3D SMPL-X annotation quality with wide-used fittiing methods SMPLify-X [30].

results. As shown in Tab. 2, the generated and ground-truth motions in *Motion-X* exhibit a higher deviation, especially in hand and face parameters, indicating significant hand and face movements over time. These results demonstrate that the model trained with *Motion-X* can generate more diverse facial expressions and hand motions, demonstrating the ability of our whole-body motion to capture fine-grained hand and face movements and expressive actions.

Method	Joints Position ↑			SMPL-X Param (10^{-2}) ↑		
	Face	Hand	Body	Face	Hand	Body
HumanML3D (GT)	0.00	0.00	88.7	0.00	0.00	9.01
Motion-X (GT)	1.60	8.82	92.2	13.4	5.24	9.95
HumanML3D	0.00	0.00	64.8	0.00	0.00	7.23
Motion-X	1.33	11.4	66.0	7.28	5.30	7.41

Table 2: Temporal standard deviation of the SMPL-X parameters and joint positions on HumanML3D and *Motion-X*. We compare the GT and generated motions with the MLD model trained on HumanML3D and *Motion-X*, respectively.

D Related Work

In this part, we introduce relevant **methods** for human motion generation.

According to different inputs, producing human motions can be divided into two categories: the general motion synthesis from scratch [37, 47, 45, 5] and the controllable motion generation from given text, audio, and music as conditions [1, 32, 46, 43, 8, 15, 2, 13]. Motion synthesis encompasses several tasks, such as motion prediction, completion, and interpolation [5], developed over several decades in computer vision and graphics. These tasks tend to utilize nearby frames with spatio-temporal correlations to infer estimated frames in a deterministic manner [28, 3, 4, 12, 14, 19, 27, 26]. On the other hand, motion generation is a more challenging task that aims to synthesize long-term, diverse, natural human motions.

Many generative models, like GANs, VAEs, and recent diffusion models, have been explored [36, 38, 16, 17, 31]. This work mainly discusses text-conditioned motion generation. This field has evolved from inputting action classes [16, 17] to sentence descriptions [15, 8, 43], and generating motions from 2D to 3D keypoints, to the emerging parametric model (e.g., SMPL [23, 15, 8]). These models have become expressive and comprehensive toward real-world scenarios thanks to the development of related benchmarks. Recently, diffusion model-based methods have rapidly developed and shown advantages in diverse, realistic, and fine-grained motion generation [43, 8, 46, 34, 39]. Some concurrent works [34, 43, 8] introduce novel diffusion model-based motion generation framework to achieve state-of-the-art (SOTA) quality. For example, MLD [8] presents a motion latent-based diffusion model with a representative motion variational autoencoder, showing its efficiency. Based on the proposed Motion-X, HumanTOMATO [24] introduces the first text-aligned whole-body motion generation that can generate high-quality, diverse, and coherent facial expressions, hand gestures, and body motions simultaneously.

E Limitation and Broader Impact

E.1 Limitation

There are two main limitations. **(i)** The motion quality of our markless motion annotation pipeline is inevitably inferior to the multi-view mark-based motion capture system. However, as the quantitative and qualitative results demonstrate, our method can perform much better than existing markless methods, thanks to large-scale models pre-trained on massive 2D and 3D keypoints datasets and our elaborately designed fitting pipeline. Besides, a 30 mm PA-MPVPE error would be acceptable for the text-driven motion generation task since the target is to synthesize natural and realistic motions that are semantically consistent with the text input. Furthermore, the experiment on the mesh recovery task has demonstrated that our dataset can also benefit the human reconstruction task, which requires a higher annotation quality. Accordingly, a better motion annotation will be beneficial, and we will leave it as our future work. **(ii)** During our experiment, we found out that existing evaluation metrics are not always consistent with visual results. Besides, SMPL-X parameters may not be the best motion representation for expressive whole-body motion representation. Thus, there is a need for further research on the evaluation metric, motion representation, and model designs for the expressive motion generation task. We leave them as future work.

E.2 Broader Impact

Based on the scalable and automatic annotation way proposed in this work, although there are inevitable errors, large-scale data could be helpful. Meanwhile, this way can boost the direction of “learning from noisy labels" for related tasks, such as text-driven whole-body motion generation. A large-scale 3D human motion dataset would have numerous applications and boost novel research topics in various fields, such as animation, games, virtual reality, and human-computer interaction. Until now, human motion datasets have had no negative social impact yet. Our proposed *Motion-X* will strictly follow the license of previous datasets, and would not present any negative foreseeable societal consequence, either.

F License

All data is distributed under the CC BY-NC-SA (Attribution-NonCommercial-ShareAlike) license. Detailed license and instructions can be found on the page `https://motion-x-dataset.github.io`. Further, we will provide a GitHub repository to solicit possible annotation errors from data users. For the sub-datasets, we would ask the user to read the original license of each original dataset, and we would only provide our annotated result to the user with the approvals from the original Institution. Here, we provide a brief license of the used assets:

- HumanML3D dataset [15] originates from the HumanAct12 [17] and AMASS [25] datasets, which are both released for academic research only and it is free to researchers from educational or research institutes for non-commercial purposes.
- BAUM dataset [40] is CC-BY 4.0 licensed.
- HAA500 dataset [9] is MIT licensed.
- IDEA400 dataset belongs to the International Digital Economy Academy (IDEA) and is licensed under the Attribution-Non Commercial-Share Alike 4.0 International License (CC-BY-NC-SA 4.0).
- HuMMan dataset [6] is under S-Lab License v1.0.
- AIST dataset [20] is CC-BY 4.0 licensed.
- GRAB dataset [33] is released for academic research only and is free to researchers from educational or research institutes for non-commercial purposes.
- EgoBody [44] is under CC-BY-NC-SA 4.0 license.
- Other data is under CC BY-SA 4.0 license.
- SMPLify-X [30] codes are released for academic research only and are free to researchers from educational or research institutes for non-commercial purposes.
- Codes for preprocessing and training are under MIT LICENSE.

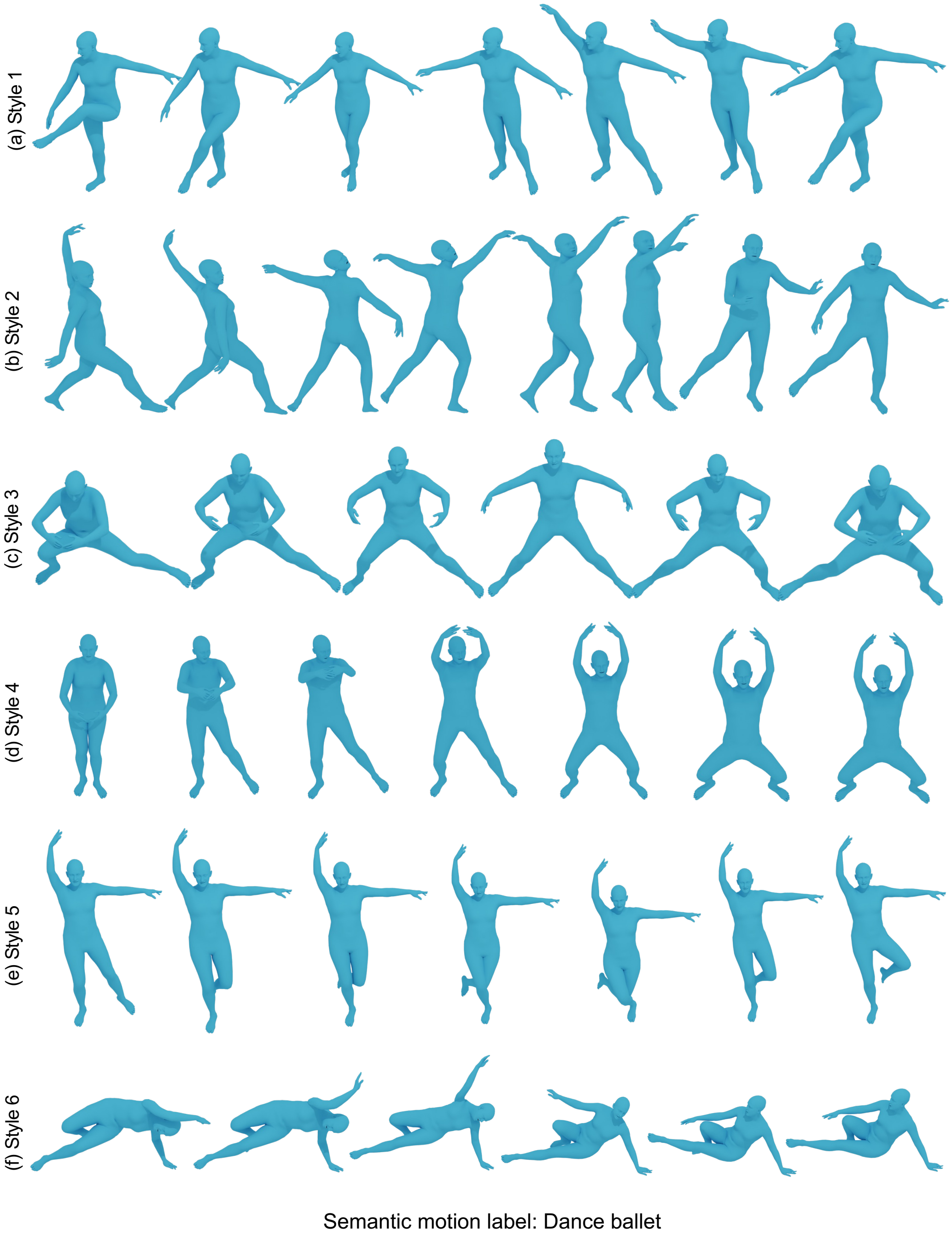


Figure 6: Examples of the same semantic label "dance ballet" with different styles to enhance the motion diversity.

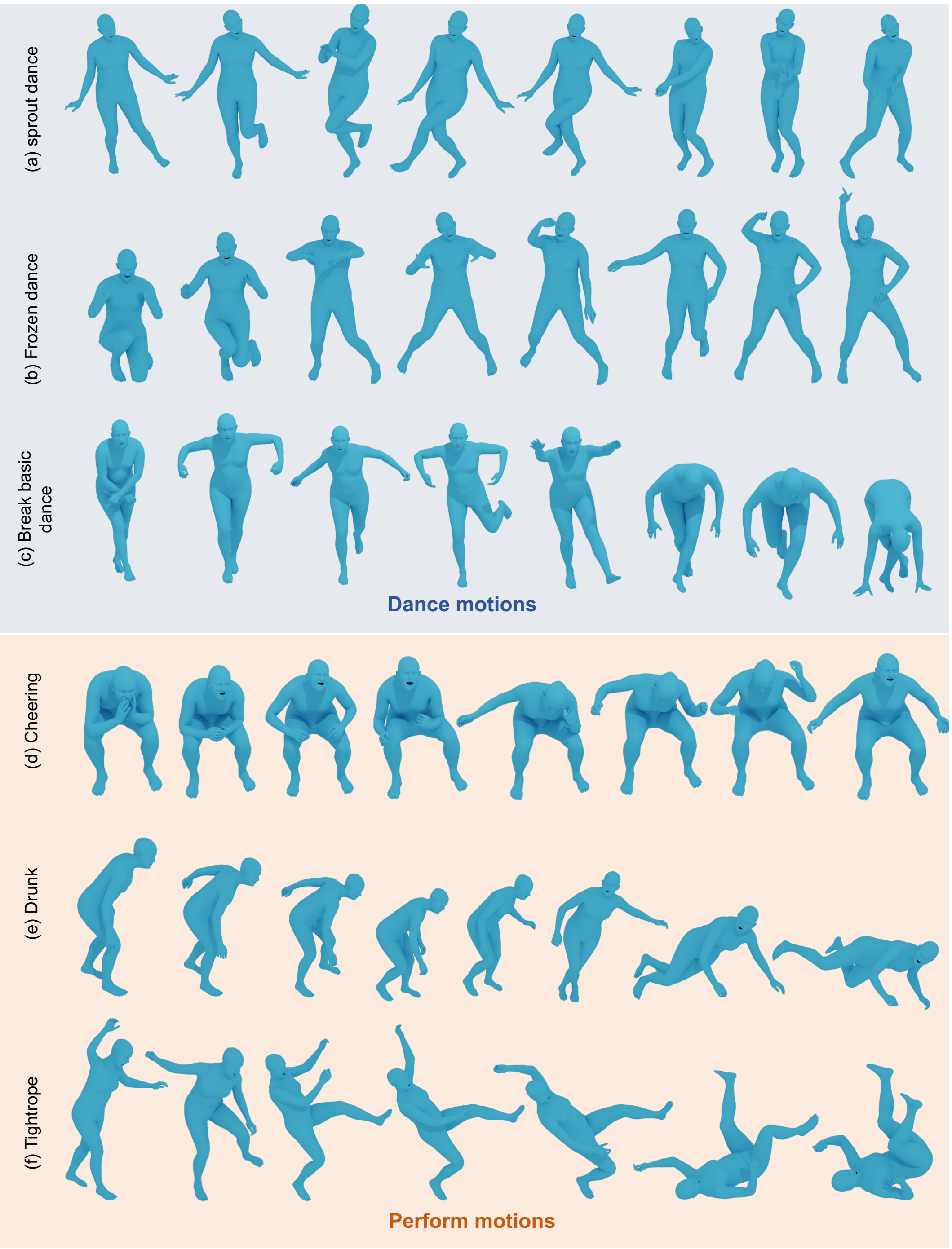


Figure 7: Visualization of the dance and perform motion sequences of *Motion-X*.

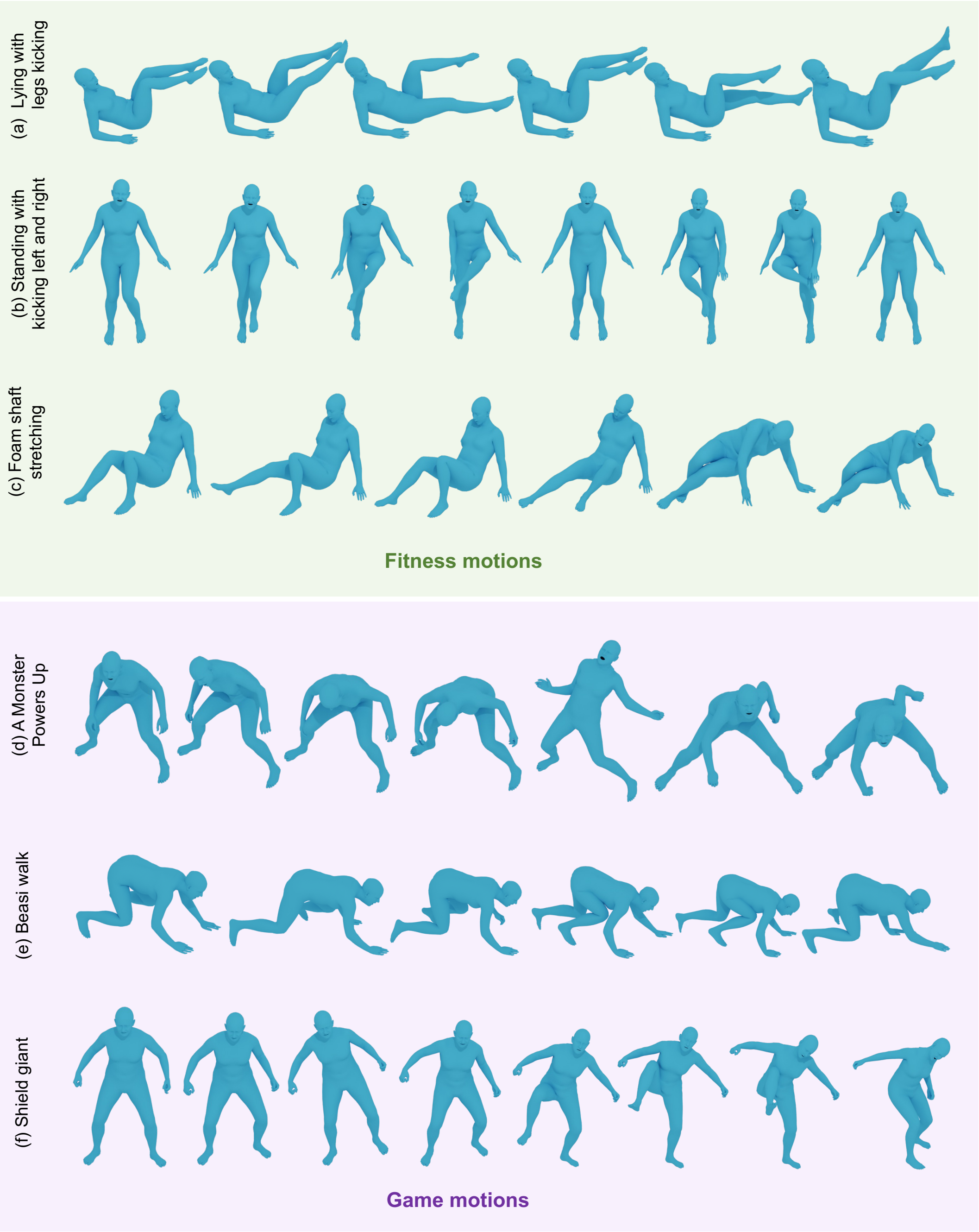

(a) Lying with legs kicking
(b) Standing with kicking left and right
(c) Foam shaft stretching
Fitness motions
(d) A Monster Powers Up
(e) Beasi walk
(f) Shield giant
Game motions

Figure 8: Visualization of the fitness and game motion sequences of *Motion-X*.

References

[1] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action. In *ICRA*, 2018.

[2] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *3DV*, 2019.

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

[4] Yujun Cai, Lin Huang, Yiwei Wang, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Xu Yang, Yiheng Zhu, Xiaohui Shen, et al. Learning progressive joint propagation for human motion prediction. In *ECCV*, 2020.

[5] Yujun Cai, Yiwei Wang, Yiheng Zhu, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Chuanxia Zheng, Sijie Yan, Henghui Ding, et al. A unified 3d human motion synthesis model via conditional variational auto-encoder. In *ICCV*, 2021.

[6] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *ECCV*, 2022.

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[8] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. Executing your commands via motion diffusion in latent space. In *CVPR*, 2023.

[9] Jihoon Chung, Cheng-hsin Wuu, Hsuan-ru Yang, Yu-Wing Tai, and Chi-Keung Tang. Haa500: Human-centric atomic action dataset with curated videos. In *ICCV*, 2021.

[10] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *CVPR*, 2022.

[11] Abhishek Dutta and Andrew Zisserman. The VIA annotation software for images, audio and video. In *ACM MM*, 2019.

[12] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *ICCV*, 2015.

[13] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *ICCV*, 2021.

[14] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. Learning human motion models for long-term predictions. In *3DV*, 2017.

[15] 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.

[16] Chuan Guo, Xinxin Zuo, Sen Wang, Xinshuang Liu, Shihao Zou, Minglun Gong, and Li Cheng. Action2video: Generating videos of human 3d actions. *IJCV*, 2022.

[17] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *ACM MM*, 2020.

[18] Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*, 2023.

[19] Manuel Kaufmann, Emre Aksan, Jie Song, Fabrizio Pece, Remo Ziegler, and Otmar Hilliges. Convolutional autoencoders for human motion infilling. In *3DV*, 2020.

[20] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, 2021.

[21] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *CVPR*, 2023.

[22] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. In *TPAMI*, 2019.

[23] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM TOG*, 2015.

[24] Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. Humantomato: Text-aligned whole-body motion generation. *arxiv:2310.12978*, 2023.

[25] 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.

[26] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *ECCV*, 2020.

[27] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *ICCV*, 2019.

[28] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *CVPR*, 2017.

[29] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *CVPRW*, 2020.

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

[31] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *ICCV*, 2021.

[32] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *ECCV*, 2022.

[33] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *ECCV*, 2020.

[34] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. In *ICLR*, 2023.

[35] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *ISMIR*, 2019.

[36] Zhenyi Wang, Ping Yu, Yang Zhao, Ruiyi Zhang, Yufan Zhou, Junsong Yuan, and Changyou Chen. Learning diverse stochastic human-action generators by learning smooth latent transitions. In *AAAI*, 2020.

[37] Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. Convolutional sequence generation for skeleton-based action synthesis. In *ICCV*, 2019.

[38] Ping Yu, Yang Zhao, Chunyuan Li, Junsong Yuan, and Changyou Chen. Structure-aware human-action generation. In *ECCV*, 2020.

[39] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *ICCV*, 2023.

[40] Sara Zhalehpour, Onur Onder, Zahid Akhtar, and Cigdem Eroglu Erdem. Baum-1: A spontaneous audio-visual face database of affective and mental states. *IEEE Transactions on Affective Computing*, 2016.

[41] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020.

[42] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-finetuned visual language model for video understanding. 2023.

[43] 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.

[44] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *ECCV*, 2022.

[45] Yan Zhang, Michael J Black, and Siyu Tang. Perpetual motion: Generating unbounded human motion. *arXiv preprint arXiv:2007.13886*, 2020.

[46] 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.

[47] Rui Zhao, Hui Su, and Qiang Ji. Bayesian adversarial human motion synthesis. In *CVPR*, 2020.