Generation of Novel Fall Animation with Configurable Attributes

Siyuan Peng
peng2000@umd.edu
University of Maryland, College Park
College Park, MD, USA

Kate Ladenheim
klad@umd.edu
University of Maryland, College Park
College Park, MD, USA

Snehesh Shrestha
snehesh@umd.edu
University of Maryland, College Park
College Park, MD, USA

Cornelia Fermüller
fermulcm@umd.edu
University of Maryland, College Park
College Park, MD, USA

Figure 1: A demonstration of the process: from RGB human images to rendered animations

ABSTRACT

It takes less than half a second for a person to fall [8]. Capturing the essence of a fall from video or motion capture is difficult. More generally, generating realistic 3D human body motions from motion capture (MoCap) data is a significant challenge with potential applications in animation, gaming, and robotics. Current motion datasets contain single-labeled activities, which lack fine-grained control over the motion, particularly for actions as sparse, dynamic, and complex as falling. This work introduces a novel human falling dataset and a learned multi-branch, Attribute-Conditioned Variational Autoencoder model to generate novel falls. Our unique dataset introduces a new ontology of the motion into three phases: Impact, Glitch, and Fall. Each branch of the model learns each phase separately and the fusion layer learns to fuse the latent space together. Furthermore, we present encompassing data augmentation techniques and an inter-phase smoothness loss for natural plausible motion generation. We successfully generated high quality images, validating the efficacy of our model in producing high-fidelity, attribute-conditioned human movements.

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**; **Hierarchical representations**.

MOCO '24, May 30–June 02, 2024, Utrecht, Netherlands

ACM ISBN 979-8-4007-0994-4/24/05
https://doi.org/10.1145/3658852.3659087

KEYWORDS

Motion Synthesis, Human Body, VAE, Dataset, Fall

ACM Reference Format:
Siyuan Peng, Kate Ladenheim, Snehesh Shrestha, and Cornelia Fermüller. 2024. Generation of Novel Fall Animation with Configurable Attributes. In *9th International Conference on Movement and Computing (MOCO '24), May 30–June 02, 2024, Utrecht, Netherlands.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3658852.3659087

1 INTRODUCTION

The task of generating 3D human body motions, particularly from motion capture (MoCap) data remains a critical yet challenging endeavor, with potential applications in computer animation, game development, and robotics. The key challenge lies in generating motion sequences that have realistic and diverse body movements using limited data. In this work, our goal is to leverage a combination of "falling movement sequences" attributes to generate an infinite number of 3D human motions. Many works have been conducted in this field that use a single action category as auxiliary input [7, 22, 27, 30, 35, 40, 43, 44] or a description [23, 33, 42] of a single action category. These methods often generate simple or cyclic movements, which lack granular control over the motion.

Instead, our goals are to use the data of a single performing artist and represent the action as a sequence of movements, each characterized by a set of attributes specific to that artist's vision and analysis of motion.

We address this problem by training a Conditional Variational Autoencoder (CVAE) [32] on a limited set of 3D human MoCap data of a performer. Using a multi-branch encoder-decoder architecture with a fusion layer in the middle, each of the (three) motion phases has its encoder-decoder network. We fuse the latent distributions

with a fully connected layer to allow communication between the different phases and output the motion for each phase.

With no existing dataset, we collected recordings of a single artist. The artist performed dramatic falls that conformed to the movement score for "Animating Death," which was also developed by the artist. This score is part of choreographies of falling and dying [19]. This project includes artistic works "Monumental Death" [17] and "COMMIT!" [18] which use these dramatic falling choreographies as primary material. Falls in these works conform to the aforementioned score.

During data collection, the artist performed falls based on a sequence of attributes: Impact, Glitch, and Fall. The Impact is the embodied site of initiation for the motion; the Glitch is a performed moment of panic, confusion, or shuddering ecstasy that sets the stage for the next segment; and finally, the Fall is the passage of the body from upright to lying down. To maintain parameters across the recordings, the artist performed falls with randomized attributes in the three phases. Utilizing the markerless Captury MoCap System [1], we collected approximately 100 trials of the artist performing dramatic falling actions, labeled with these attributes and granular sub-definitions of expressive motion.

We adopt the SMPL human model [24] to represent the human body and actions, providing access to human joints' location, rotation, and surface meshes. Following the methodology of Petrovich et al. [30] and Lee et al. [20], we employ a mixed loss function combining joint and mesh information.

Different from previous works, the falling movement is complex and has multi-phase labels. Thus from a computational point of view, it is more challenging to accurately represent and generate the falling movements. We tackle this problem by using a multi-branch design with each branch specialized for one phase. However, a challenge arises from the multi-branch design: abrupt motion changes between different phases. To mitigate this issue, we designed an inter-phase smoothness loss. By controlling the variance of the motion sequence's first derivative and modeling the transition data using spline interpolation, we achieve smooth transitions between phases. With a limited number of recorded motion sequences, there are only a few samples for each falling attribute. We introduced whole-body movement data augmentation using Fourier transformation. Transforming the data into frequency space and manipulating the phase and magnitude, we effectively increased data variation, leading to a more robust model.

Our contributions to the field of human pose representation and animation are as follows.

(1) We collected a unique falling pose dataset with multi-attribute labeling (2) We developed an attribute-conditioned multi-branch 3D human body motion synthesis model (3) We implemented human body pose data augmentation using Fourier transformation, and (4) We designed an Inter-Phase Smoothness Loss to smooth the transition between phases.

Further, our work presents a unique collaboration between artists and computer scientists; one in which the point of view of a particular artist drives the creative output of a machine-learning model. Instead of animations built off of aggregate data, ultimately erasing the identities and particularities of the contributing performers, our resulting tool celebrates the particular creative vision and embodied attributes of a single artist. The resulting animation tool offers new creative possibilities for falling animations, which could be extended to a variety of other choreographed motions.

2 RELATED WORK

2.0.1 Machine Learning: Prior to deep learning, researchers applied optimization methods to 3D human motion prediction and synthesis tasks [9, 21]. Methods like inverse kinematics [15, 38] and motion graphs [2], however, need manual tuning and cannot generate complex and diverse human movements. With the recent development of generative models like GANs [6] and Diffusion [12], 3D human body motion tasks have received significant attention. Yang, Ceyuan, et al [40] utilize GANs on pose sequences and semantic consistency to control the dynamics of human motion. With the help of large motion datasets, Lin and Amer [23] treat class labels as text conditioning and feed them to an RNN-based GAN network. [3] build a probabilistic function conditioned on previous frame actions. The limitations of GAN-based networks include accumulated error in long sequences; it is difficult to train them and hard to model spatial information. Denoising Diffusion models have shown remarkable performance in generating diverse and realistic images and videos. Recent works adopted this methodology for motion modeling with promising results. The MDM model [33] is a transformer-based diffusion model designed for various tasks, including text-to-motion and action-to-motion. A significant contribution is that it predicts on samples rather than noise. MotionDiffuse[42], a diffusion-based human motion synthesis model, is capable of responding to fine-grained manipulation of body parts. PhysDiff[41] is designed to integrate physical constraints into the diffusion process, enhancing the physical plausibility of existing models. The downside of the diffusion-based model, however, is the need for a vast amount of data to generate high-quality and diverse motion sequences. Variational Autoencoder (VAE) has been a popular method for solving human body motion synthesis tasks. Habibie et al. (2017) [11] designed a VAE with a recurrent design, showing the potential of VAEs in capturing the temporal dependencies. Yan et al. (2018) [39] utilize the concept of motion modes to design their MT-VAE model capable of generating multiple diverse facial and full-body motions. Generating motion frame by frame, He et al. (2018) combine VAE and RNN design to generate consistent and diverse video sequences. Our work builds upon Petrovich et.al's ACTOR[30] network design to extract sequence-level embeddings and generate holistic body movement.

2.0.2 Artistic Animation: Our work also builds on embodied data collection, translation, and generative animation in artistic contexts. Shaw [31] describes projects *Synchronous Objects* and *Motion Bank;* the former presenting alternative visualizations of embodied data and the latter providing a platform for annotating choreography specific to an artist's vision. Choreographic motion capture tools have been explored by Whitley [36] and collaborators, though this project creates sequences from prerecorded motions. Wayne McGregor's *Living Archive* [25] uses machine learning processes to generate new choreography from the embodied data of McGregor's dancers. Ellsworth [5] used GANs in the artwork *Cellular Automaton* to extend spatial configurations of pre-recorded motion.

2.0.3 Dataset: Previous researches have collected various human body movement datasets, including the popular Human3.6M dataset [13]. It contains 3.6 millions of human poses in 17 scenes. UESTC dataset [14] collects 2.5 thousand movement sequences in 40 simple action categories. With less recorded data, HumanAct12 dataset [10] provides joints coordinates for 12 action categories.

3 METHOD

3.1 Problem setup

When describing human actions, it's desirable to separate body shape from pose [28, 29, 37]. Ignoring body shape, we aim to generate a sequence of pose parameters: specifically, the relative joint rotation in the kinematic tree of the human body. In formal notation, given a combination of falling attribute labels $L_\phi 1, L_\phi 2, L_\phi 3$ (phase ϕ), and time intervals $1, ...N$, we synthesize a sequence of body joint parameters $\hat{P_1}...\hat{P_N}$, that contain the root joint's rotation and translation, as well as the body joint rotation parameters as shown in Fig 2.

3.2 Dataset

3.2.1 Data Collection. We used the Captury [1] motion capture system to record human motion sequences. Eight MoCap cameras in circular formation with four at high altitudes and four at low altitudes were used to ensure good recording of the actor's poses when on the ground. For each trial, the actor's body center location, rotation, and the relative rotations of 24 bones in a kinematic tree were recorded in line with SMPL [24].

We organize all falls into a new ontology consisting of a five-part (impact, glitch, fall, end, and resurrection) movement score for "animating death" [17, 18]. Our collection process focused on the first three parts of the score: the Impact, the embodied site of initiation for the motion; the Glitch, a performed moment of panic, confusion, or shuddering ecstasy that sets the stage for the next segment; and the Fall, the passage of the body from upright to lying down. Our dataset specified body areas of initiation (head, torso, arms, and legs), which were impacted described by the following qualities: (1) **Push**: the impacted body part appears to be shoved in any direction; (2) **Prick**: a localized, sharp action like a needle piercing skin; (3) **Shot**: a forceful, localized action akin to a gunshot; (4) **Contraction**: the hollowing out or concaving of an area of the body; and (5) **Explosion**: a violent, bursting action starting at the point of impact and radiating quickly outwards and away from the body.

The next phase of the choreography is Glitch, which has its own set of aesthetic parameters: (1) **Shake**: a quake or tremor; (2) **Flail**: wild, out-of-control motions that extend outwards in a flinging motion; (3) **Flash**: a single brief, spreading motion akin to a flash of light; (4) **Stutter**: repetitive stop and start motions; (5) **Contort**: twisting or warping the body; (6) **Stumble**: an off-balance, tripping quality; and (7) **Spin**: turning around an axis, which can be full–bodied (the performer turns around their center line) or localized (for example, the spinning of the hand in a circle).

Finally, we defined qualities for the Fall phase: (1) **Release**: where the muscles of the body relax and collapse; (2) **Let Go**: the feeling of a cord being cut, or whatever is holding the body up disappears; (3) **Hinge**: a reference to Horton and Graham modern dance techniques, where the body slowly descends to the ground in a flat shape, akin to a door hinge; and (4) **surrender**: a performance of pleading or giving up, often accompanied by kneeling or other diminutive postures.

3.2.2 Data Augmentation. Due to the limited amount of data and our aim of building a more robust model, we performed data augmentations on existing data. We began by applying a Fast Fourier Transform (FFT) to convert segments of motion time-series data into the frequency domain. This transformation allows us to manipulate the motion data in ways that are not easily achievable in the time domain. We then tweaked the magnitude and phase information and converted it back to the time domain, ensuring the augmented motion remains continuous and retains the natural flow of human movement.

3.3 CVAE Model

We build on the CVAE-based ACTOR model [30], but use action attributes as conditions and extend the model with a multi-branch design. Specifically, for each of the three action phases, an encoder and a decoder are used for extracting information and generating poses from latent spaces.

3.3.1 Encoder. The input are a sequence of body poses from P_1 to P_K, where P_1 to P_M are poses for phase 1, P_{M+1} to P_N are inputs for phase 2, and P_{N+1} to P_K are for phase 3. Since the operations are similar for each phase, we will explain the steps in detail for phase 1 only. After converting the phase label into learnable tokens, we prepend them to the pose sequences. These tokens, similar to those used in [30], are used for pooling purposes in the temporal dimension. Due to the self-attention mechanism, these tokens will aggregate (or pool) information from the entire action sequences. A similar implementation can be seen in the BERT [4] model for sentiment prediction. Positional Encoding (PE) has been proven to be a vital part of various works, such as Transformer [34] and NeRF [26] architectures. We also take advantage of it and add it to the input, and the encoder encodes all data into a low-dimensional latent space. To extract the distribution parameters μ and Σ, we simply take the first two outputs of the corresponding encoder for each phase.

3.3.2 Embedded Space. With three separate encoders in our model, merging information from the different phases is essential to produce smooth outputs. After extracting the two distribution parameters μ and Σ from the encoder outputs, we use the reparameterization trick introduced in [16] to allow gradients to pass in the sampling process. Then, we concatenate sampled latent vectors and pass them to the fully connected fusion layer. The model is expected to exchange and learn features from neighboring branches and generate poses accordingly.

3.3.3 Decoder. The goal of the decoder is, given a latent vector z and one of the falling attributes $sL_{a_i}, L_{b_j}, L_{c_k}$, to generate a novel sequence of human body parameters $\hat{P_1}...\hat{P_N}$. Again, we have three very similar attention decoders so we only describe one decoder. As one of the inputs of the decoder network, time information is added by the positional encoding and fed into the decoder as a

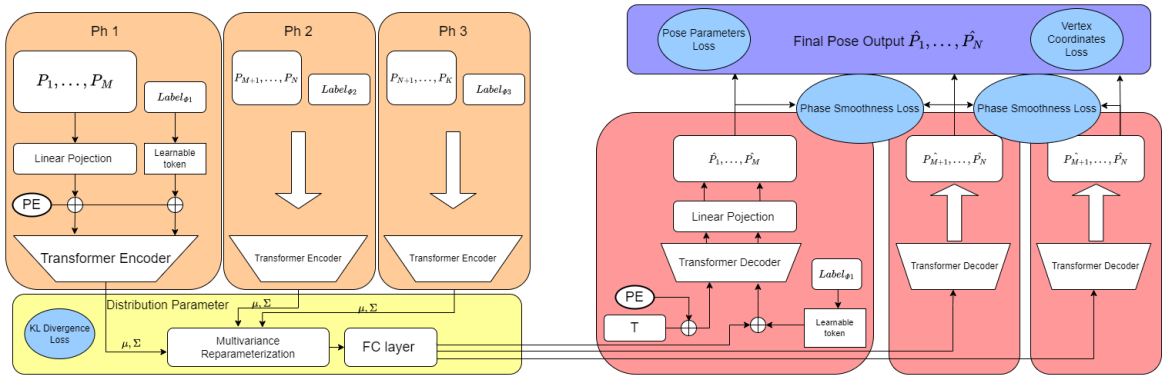

Figure 2: Attribute-Conditioned Variational Autoencoder model architecture: Encoder (orange box), embedded space (yellow box), and decoder (red box). Taking a sequence of body poses, we split them into inputs for different phases. We prepend two learnable variables (derived from action attributes) to the input sequences. Adding positional information to the input through a positional encoder (PE), we feed them into the transformer encoder. However, we only take the first two outputs of the encoder as the distribution parameters μ and Σ into the embedded space. The encoder uses a fully connected layer as a fusion layer between different phases. The input of the decoder has two parts: time information (the length of the generated sequences) and the attribute-biased latent space. The final multi-phase sequences are the concatenation of all the outputs from the three phases.

query. To add falling attributes to the decoder, we sum a learnable token with the latent space to shift it to an attribute-dependent space. The output is then going through a fully connected layer to get the final output motion sequences for the phase: $\hat{P_1}$ to $\hat{P_M}$. To get a complete synthesized motion, we simply concatenate all generated poses.

3.3.4 Inference. During inference, we randomly sample vectors from a normal distribution. Shifted toward the attribute-defined latent space, the sampled vector is combined with a learned attribute token. An arbitrary length of empty sequences is also fed into the decoder as input. This allows us to synthesize any length of falling movements with any combination of attributes.

3.3.5 Loss. We utilize a combined loss comprising the human body model's parameters, KL divergence, and vertex reconstruction loss. To enforce smooth transitions between phases, we also designed a movement smoothness loss, which is added to the final objective.

For each phase, we have L2 Euclidean distance losses between the human body model parameters, denoted as L_B, and L2 loss between the reconstructed human model vertex L_V, along with the standard Kullback-Leibler (KL) divergence loss. The total in-phase loss can be expressed as the weighted sum of the above losses. A key challenge in multi-phase pose generation is to maintain a smooth transition between different phases. We implemented an inter-phase smoothness loss to tackle this problem. The loss comprises two parts: velocity smoothness and displacement smoothness. We take the last 10% of the data from previous phase and the first 10% of the data from current phase as transitional data. By calculating the first derivative of the joint displacement, we minimize the variance of the velocity. The displacement smoothness constraint is implemented using spline interpolation. For transitional data, we model a movement curve using the interpolation and then calculate the L2 Euclidean distance between the interpolation and the predicted joint position. The total loss is a weighted sum between the in-phase loss and the smoothness loss.

4 RESULTS

4.0.1 Experiment Setup. To enhance the robustness of our model, we applied the pre-trained weights from ACTOR [30] on the UESTC dataset [14] on all the encoder-decoder branches while discarding the unfit weights. We trained our model on the falling dataset for 3000 epochs, using a batch size of 20 and a learning rate of 1e-4.

4.0.2 Qualitative Result. In Figures 3, 4, and 5, we present visualizations of a sample generated by our model.

4.0.3 Quantitative Result. We generated two sets of random motions using both our multi-branch model and the original ACTOR model as a point of comparison [30]. Through a randomized blind rating on a Likert scale ranging from 1 to 10, where higher scores indicate better conformity to specified attributes and overall quality, our model consistently outperformed the original by an average margin of 20%.

4.0.4 Discussion and Limitation. We implemented data augmentation by varying the amplitude of the movement. However, due to the randomness of the action and the augmentation process, there were instances where parts of the human body, such as the elbow, ended up inside the human model. Future work could explore more realistic and robust methods of data augmentation. As we add more branches to the model, the training time and resource requirements increased. This is compounded by the fact that we input entire motion sequences into the model. Longer input motions would further escalate the need for computational resources.

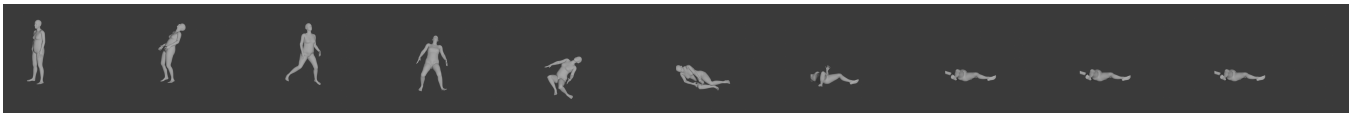

Figure 3: Impact Location: Torso; Impact Attribute: Push; Glitch Attribute: Stumble; Fall Attribute: Surrender

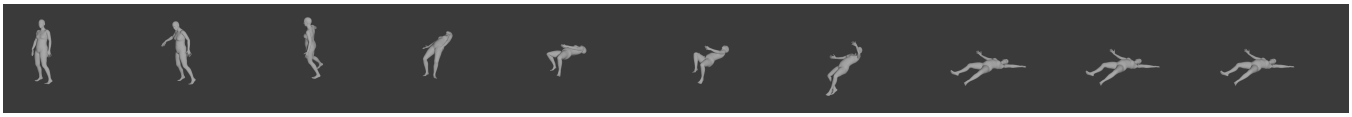

Figure 4: Impact Location: Arms; Impact Attribute: Prick; Glitch Attribute: Spin; Fall Attribute: Let go

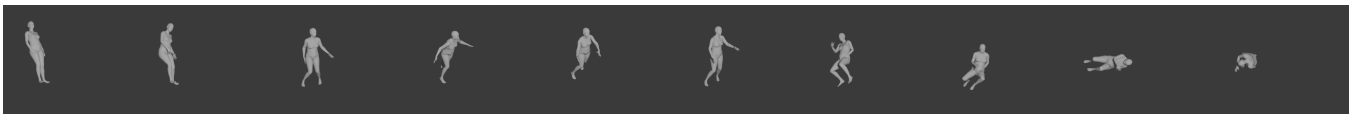

Figure 5: Impact Location: Legs; Impact Attribute: Push; Glitch Attribute: Contort; Fall Attribute: Hinge

5 CONCLUSIONS

We present not only a unique dataset featuring complex labeling but also a novel attribute-conditioned variational autoencoder designed to focus on each phase of the input data. The multibranch design achieved more granular control over the motion sequences. Despite the limited amount of data, we successfully generated diverse and realistic 3D human movement data by applying data augmentation and smoothness losses.

ACKNOWLEDGMENTS

To Robert, for the bagels and explaining CMYK and color spaces. To Ziyan Chen, for explaining SMPL.

REFERENCES

[1] [n. d.]. Captury MOTION CAPTURE REDEFINED: GO MARKERLESS. https://captury.com/. Accessed: 2023-11-12.

[2] Okan Arikan, David A. Forsyth, and James F. O'Brien. 2003. Motion Synthesis from Annotations. In *ACM SIGGRAPH 2003 Papers* (San Diego, California) *(SIGGRAPH '03)*. Association for Computing Machinery, New York, NY, USA, 402–408. https://doi.org/10.1145/1201775.882284

[3] Emad Barsoum, John Kender, and Zicheng Liu. 2018. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 1418–1427.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[5] Michelle Ellsworth. 2021. 2021 Professor of Distinction Michelle Ellsworth on the Post-Verbal Social Network. Youtube Video. https://www.youtube.com/watch?v=8d-t8aIOMuI

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative Adversarial Networks. *Commun. ACM* 63, 11 (oct 2020), 139–144. https://doi.org/10.1145/3422622

[7] Anand Gopalakrishnan, Ankur Mali, Dan Kifer, Lee Giles, and Alexander G Ororbia. 2019. A neural temporal model for human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12116–12125.

[8] Brenda E Groen, Ellen Smulders, D De Kam, Jacques Duysens, and Vivian Weerdesteyn. 2010. Martial arts fall training to prevent hip fractures in the elderly. *Osteoporosis international* 21 (2010), 215–221.

[9] Gutemberg Guerra-Filho, Cornelia Fermüller, and Yiannis Aloimonos. 2005. Discovering a language for human activity. In *Proceedings of the AAAI 2005 Fall Symposium on Anticipatory Cognitive Embodied Systems*.

[10] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. 2020. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2021–2029.

[11] I. Habibie, D. Holden, J. Schwarz, J. Yearsley, and T. Komura. 2017. A recurrent variational autoencoder for human motion synthesis. *Procedings of the British Machine Vision Conference 2017* (2017). https://doi.org/10.5244/c.31.119

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2014. 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 (jul 2014), 1325–1339.

[14] Yanli Ji, Feixiang Xu, Yang Yang, Fumin Shen, Heng Tao Shen, and Wei-Shi Zheng. 2019. A large-scale varying-view rgb-d action dataset for arbitrary-view human action recognition. *arXiv preprint arXiv:1904.10681* (2019).

[15] O. Khatib, E. Demircan, V. De Sapio, L. Sentis, T. Besier, and S. Delp. 2009. Robotics-based synthesis of human motion. *Journal of Physiology-Paris* 103, 3 (2009), 211–219. https://doi.org/10.1016/j.jphysparis.2009.08.004 Neurorobotics.

[16] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[17] Kate Ladenheim. 2022. Monumental Death. Artwork. https://www.kateladenheim.com/work/monumental-death

[18] Kate Ladenheim. 2023. COMMIT! Live Performance. https://www.kateladenheim.com/work/commit

[19] Kate Ladenheim. 2023. Logics of Embodied Control. University Lecture. https://www.wac.ucla.edu/announcements/archive/2023/fall/november/kate-ladenheim-public-talk

[20] Kyungho Lee, Seyoung Lee, and Jehee Lee. 2018. Interactive character animation by learning multi-objective control. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 1–10.

[21] Yi Li, Cornelia Fermüller, Yiannis Aloimonos, and Hui Ji. 2010. Learning shift-invariant sparse representation of actions. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2630–2637.

[22] Zimo Li, Yi Zhou, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. 2017. Auto-conditioned recurrent networks for extended complex human motion synthesis. *arXiv preprint arXiv:1707.05363* (2017).

[23] Xiao Lin and Mohamed R Amer. 2018. Human motion modeling using dvgans. *arXiv preprint arXiv:1804.10652* (2018).

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 851–866.

[25] Wayne McGregor. 2019. Living Archive. Live Performance. https://waynemcgregor.com/productions/living-archive/

[26] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2021. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Commun. ACM* 65, 1 (dec 2021), 99–106. https://doi.org/10.1145/3503250

[27] Dirk Ormoneit, Michael J Black, Trevor Hastie, and Hedvig Kjellström. 2005. Representing cyclic human motion using functional analysis. *Image and Vision Computing* 23, 14 (2005), 1264–1276.

[28] Ahmed AA Osman, Timo Bolkart, and Michael J Black. 2020. Star: Sparse trained articulated human body regressor. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 598–613.

[29] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10975–10985.

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

[31] Norah Zuniga Shaw. 2014. Animate inscriptions, articulate data and algorithmic expressions of choreographic thinking. *Choreographic Practices* 5, 1 (2014), 95–119. https://doi.org/10.1386/chor.5.1.95_1

[32] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems* 28 (2015).

[33] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. 2022. Human motion diffusion model. *arXiv preprint arXiv:2209.14916* (2022).

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[35] Zhiyong Wang, Jinxiang Chai, and Shihong Xia. 2021. Combining Recurrent Neural Networks and Adversarial Training for Human Motion Synthesis and Control. *IEEE Transactions on Visualization and Computer Graphics* 27, 1 (2021), 14–28. https://doi.org/10.1109/TVCG.2019.2938520

[36] Alexander Whitley, Sönke Kirchhof, and Daniel Strutt. 2023. Digital Dance Studio VR (DDS-VR): An Innovative User-Focused Immersive Software Application for Digital Choreographic Composition, Planning, Teaching, Learning, and Rehearsal.. In *ACM SIGGRAPH 2023 Immersive Pavilion* (Los Angeles, CA, USA) *(SIGGRAPH '23)*. Association for Computing Machinery, New York, NY, USA, Article 5, 2 pages. https://doi.org/10.1145/3588027.3595602

[37] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. 2020. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6184–6193.

[38] Katsu Yamane, James J. Kuffner, and Jessica K. Hodgins. 2004. Synthesizing Animations of Human Manipulation Tasks. In *ACM SIGGRAPH 2004 Papers* (Los Angeles, California) *(SIGGRAPH '04)*. Association for Computing Machinery, New York, NY, USA, 532–539. https://doi.org/10.1145/1186562.1015756

[39] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. 2018. MT-VAE: Learning Motion Transformations to Generate Multimodal Human Dynamics. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[40] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. 2018. Pose guided human video generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 201–216.

[41] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. 2023. PhysDiff: Physics-Guided Human Motion Diffusion Model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 16010–16021.

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

[43] Dongsheng Zhou, Xinzhu Feng, Pengfei Yi, Xin Yang, Qiang Zhang, Xiaopeng Wei, and Deyun Yang. 2019. 3D human motion synthesis based on convolutional neural network. *IEEE Access* 7 (2019), 66325–66335.

[44] Dongsheng Zhou, Chongyang Guo, Rui Liu, Chao Che, Deyun Yang, Qiang Zhang, and Xiaopeng Wei. 2021. Hierarchical learning recurrent neural networks for 3D motion synthesis. *International Journal of Machine Learning and Cybernetics* 12 (2021), 2255–2267.