Validating Objective Evaluation Metric: Is Fréchet Motion Distance able to Capture Foot Skating Artifacts ?

Antoine Maiorca
antoine.maiorca@umons.ac.be
University of Mons, ISIA Lab
Mons, Belgium

Youngwoo Yoon
youngwoo@etri.re.kr
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea

Thierry Dutoit
thierry.dutoit@umons.ac.be
University of Mons, ISIA Lab
Mons, Belgium

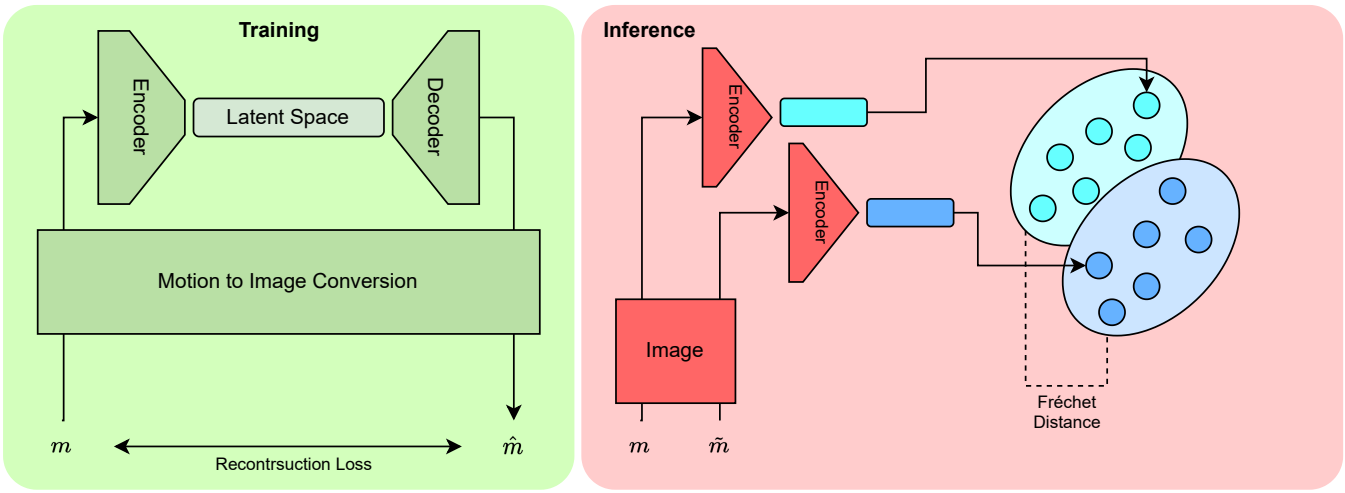

Figure 1: AutoEncoder-based Fréchet Motion Distance. First, a latent representation of motion is learnt to gather relevant information about motion data. Then, Fréchet distance is computed between the whole set of encoded ground truth m and synthesized motion samples $\tilde{m}$.

ABSTRACT

Automatically generating character motion is one of the technologies required for virtual reality, graphics, and robotics. Motion synthesis with deep learning is an emerging research topic. A key component of the development of such an algorithm involves the design of a proper objective metric to evaluate the quality and diversity of the synthesized motion dataset, two key factors of the performance of generative models. The Fréchet distance is nowadays a common method to assess this performance. In the motion generation field, the validation of such evaluation methods relies on the computation of the Fréchet distance between embeddings of the ground truth dataset and motion samples polluted by synthetic noise to mimic the artifacts produced by generative algorithms. However, the synthetic noise degradation does not fully represent motion perturbations that are commonly perceived. One of these artifacts is foot skating: the unnatural foot slides on the ground during locomotion. In this work-in-progress paper, we tested how well the Fréchet Motion Distance (FMD), which was proposed in previous works, is able to measure foot skating artifacts, and we found that FMD is not able to measure efficiently the intensity of the skating degradation.

CCS CONCEPTS

• **Computing methodologies** → **Motion capture**.

KEYWORDS

Deep Neural Network, Motion Generation, Generative Model Evaluation

ACM Reference Format:
Antoine Maiorca, Youngwoo Yoon, and Thierry Dutoit. 2023. Validating Objective Evaluation Metric: Is Fréchet Motion Distance able to Capture Foot Skating Artifacts ?. In *ACM International Conference on Interactive Media Experiences (IMX '23), June 12–15, 2023, Nantes, France.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3573381.3596460

IMX '23, June 12–15, 2023, Nantes, France

ACM ISBN 979-8-4007-0028-6/23/06.
https://doi.org/10.1145/3573381.3596460

1 INTRODUCTION

Motion synthesis is a popular field in graphics and robotics. That research area also impacts the development of realistic immersive space. In many interactive media experiences, virtual characters are animated to interact with human users [25, 33, 39]. Hence,

animating a virtual character with realistic and plausible motions is a task that matters in computer vision. Recent generative methods are able to produce diverse motions with outstanding quality in various context [10, 14, 22, 30, 37, 44, 46].

One crucial component in developing generative models is an evaluation metric that measures how well the models generate realistic and natural-looking movements. Since the ability to create virtual characters that appear genuine and realistic is paramount to delivering an immersive and convincing user experience [20], being able to evaluate the generative models' performance and point out the motion artifacts in the avatar animation is an important task in the immersive experience design.

Unlike the tasks where ground truth or goal exist (*e.g.*, classification, game playing), there is usually no single answer for generative tasks, and this makes assessing the performance of generative models difficult. The quality and diversity of synthesized samples from a generative model are key indexes in the evaluation. Assessing precisely these two is then fundamental to being able to compare the performance of generative algorithms. One instinctive manner to achieve this goal is to involve humans in the loop of the generative method evaluation. Humans observe and rate the quality of the generated samples through user studies. These subjective measurements rely on human perception so that the metric is highly correlated with the ultimate goal of the generative model that generates indistinguishable and pleasing results for humans. The human-based evaluation was usually performed to clearly and effectively identify the quality of the samples of motion synthesized from their proposed generative models [1, 38]. However, this kind of evaluation has significant drawbacks as pointed out in [4]: the subjectivity of judges, low reproducibility issues and the non-neglectable amount of time and resources to conduct the evaluation hinder fast comparison at low cost with other previous methods. For these reasons, there is a need for objective evaluation metrics that evaluate qualitatively and at low cost. The need for evaluation metrics is even greater in recent learning-based generative models because those involve numerous training attempts and an evaluation tool to compare attempts is necessary for faster development.

One popular metric that aims to assess the quality and diversity of a set of generated modalities is by measuring the *Fréchet Distance* (FD) [8] between the distribution of ground truth and synthesized sample datasets and is computed as

$$FD = ||\mu_r - \mu_g||_2^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{\frac{1}{2}}) \quad (1)$$

where (μ_r, Σ_r) and (μ_g, Σ_g) are the mean and covariance matrix pairs of respectively ground truth and generated sample distributions. This equation assumes that the distributions are multivariate Gaussian. In the image synthesis field, the standard method employed to estimate the real and synthesized distributions is by extracting the last activation maps from the *Inception* network [36]. The distribution parameters are estimated from these sets of feature maps and FD is computed with these parameters. This score is called *Fréchet Inception Distance* (FID) and is now the de-facto standard in evaluating image synthesis models [7, 35]. The audio synthesis field adopted a modified version of FID called *Fréchet Audio Distance* [18]. The motion synthesis field also used FD to evaluate the quality and diversity of the generated samples [11, 40, 42]. However, these works proposed different methods to compute the score and very few analyses have been done on these methods to evaluate the validity of the proposed metric [27, 44]. These analyzes showed that the proposed metric is efficient to capture spatial noise on motion data but lacks of reliability regarding temporal disturbance. However, these artifacts do not represent any real artifacts produced by generative models or during a standard motion capture recording. In this paper, we propose to validate the FMD on foot skating artifacts. Foot skating is a motion artifact where the character's foot slides when on contact with the ground. In deep neural animation, it is often caused by a regressive model converging to the mean pose. This work aims to analyze the FMD behavior on motion data polluted by foot skating artifacts with various intensity. We hope this work will be helpful in the development of a robust metric for the evaluation of motion generative models.

2 RELATED WORK

2.1 Metrics in Motion Generation Evaluation

Several techniques have been employed to assess the quality of synthetic motions. In short-term human motion prediction *i.e.*, predicting future frames given a motion prefix, Euclidean distance between the generated motion and the ground truth evaluates the model accuracy [9, 17, 28, 29]. A popular use case in the interactive motion generation field is the development of methods that animate and control in real time the motion of a biped or quadruped character given its path [14, 46]. In this context, foot skating artifacts (the unnatural foot sliding when it is in contact on the ground) are commonly perceived and measured as an evaluation metric. Subjective evaluation protocols have been conducted by several works in the motion generation context such as in walk cycle generation to evaluate the expression naturalness [38]. More recently, in music-driven motion generation model, virtual characters learn to perform a dance that matches with the input music and user studies have been performed to measure the motion samples rhythmic and aesthetic consistency [1, 15, 21, 23, 48]. *Diversity* and *multimodality* metrics are used to evaluate if the method is able to generate a wide range of motion from a text prompt [37, 47]. However, these metrics are task-oriented and only apply in specific contexts of motion generation. For example, they cannot be used in generative models that take no context in input to synthesize motion.

2.2 Fréchet Inception Distance

The Fréchet Distance [8], known as *Wasserstein 2* distance is introduced in [13] to evaluate GAN on image modalities. This measures the discrepancy between two multivariate Gaussian distributions by Equation 1. Considering the two distributions respectively as the real and synthesized data distribution, it shows that the FID score effectively assesses the quality and diversity of a generated dataset of images. The statistics needed to compute the FID (see Equation 1) are estimated from the real and generated feature maps computed from *InceptionV3* [36] when being fed by the data. The advantages of the FID are the following: (1) it works a single-value metric that evaluate the performance of generative models and in contrast to IS and (2) it is sensitive to intra-class mode collapse *i.e.*, generating very similar samples. However, the FID exhibits some

limitations: It is inconsistent with human judgment when evaluating on other image dataset than *Imagenet* [31], sensitive to the number of samples to estimate FID and even the particular model being evaluated [6]. Many works proposed an improved version of this metric [3, 6, 24, 26]. These limitations make the use of FID as a single and unified metric to assess the performance of a generative model difficult. This is why other metrics such as *Precision* and *Recall* have been proposed in [19, 34] that respectively measure the average sample quality and the coverage of sample manifold.

2.3 FID in Motion Generation Evaluation

Nowadays, most of the recent works mentioned in Section 2.1 employed an adapted version of the FID as a metric to benchmark the proposed model in motion generation where the feature manifold is related to motion data instead of image modalities. In action motion generation, the methods in [11, 42] train an action classifier model as proposed in [43] to embed the motion data. The FD is then computed on the output of the final layer of the feature extractor network. Chang *et al.* [5] used the action classifier part of the proposed backbone generative models as the feature network. However, this method explicitly requires labels of motion data to train the action classifier network which is not suitable for unlabeled and unstructured data. Instead, Wang *et al.* [40] propose to train the *LSTM*-based motion prediction network in [9] to extract a feature space that captures relevant information on motion. The generation of co-speech gesture is also evaluated using FID by [2, 44] from an unsupervised training of an AutoEncoder and the FD is computed between the latent spaces of real and synthesized gesture. [27, 44] establish a analysis of this method and conclude that the latent space is efficient to identify spatial motion degradation but fails to capture temporal discontinuities. Finally, [41] evaluate their action motion generative model using *InceptionV3* [36] pretrained on the *Imagenet* dataset but no analysis on the validity of this metric has been performed.

3 ANALYZES

3.1 Foot Skating Artifacts

To identify if the FMD is sensitive to foot skating artifacts, we first need to build datasets with different skating intensities. To achieve this goal, we employ the method in [46] that animates a virtual quadruped and controls its trajectory in real-time. All the materials used in this work are provided here [1]. The overview of the model is shown in Figure 2. This model architecture is composed of a pose regression network Ψ and a gating network Φ based on fully connected layers. The pose regression network takes as input motion features at frame f and aims to generate motion features at frame $f+1$. To avoid mean pose regression inducing foot skating artifacts, Mixture-of-Experts technique [16] is used to compute the parameters θ of the regression network: θ is obtained by blending n expert parameters by the coefficients ω computed by the gating network Φ. The input of this model is a subset $\tilde{x}$ of the input motion features x. This subset gathers the information of leg features such as feet position, orientation and velocity. It helps to learn multiple gait cycles of dog locomotion [46].

[1] https://github.com/pauzii/AnimationAuthoring

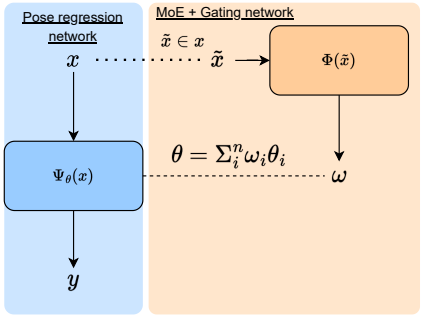

Figure 2: Model proposed in [46] to control in real-time the trajectory of a virtual quadruped. At each frame, the motion of the dog $y = x(f+1)$ is computed by the pose regression network Ψ_θ. Ψ is dynamic model where its parameters $\theta = \Sigma_i^n \theta_i \omega_i$. The n coefficients $\omega_1, ...\omega_n$ blended the n expert parameters $\theta_1, ..., \theta_n$ learned during the training process. The gating network Φ computes the set of coefficients ω from a subset of input motion features $\tilde{x}$.

$$x(f+1) = \Psi_\theta(x(f)) \quad where \quad \theta = \Sigma_i^n \theta_i \omega_i \quad and \quad \omega = \Phi(\tilde{x}(f)) \tag{2}$$

The architecture of the pose regression network Ψ consists of 3 fully connected layers. The number of units in each layer is h_{size} (which was 512 in [46]). Reducing the number of parameters, h_{size}, leads to underfitting: the reduced model to learn the complex behavior of different dog gait cycles tends to converge to a mean pose that minimizes the regression error. This effect induces foot skating artifacts because the resulting motion is stiffer, especially regarding the legs. Hence, reducing h_{size} deteriorates the motion quality and the motion is polluted with more foot skating. Examples of this phenomenon are shown in Figure 3 and we recommend to watch the videos of the degraded motion here.[2]

We generated 5582 frames for each motion using the models with $h_{size} = 512, 256$ and 64. The input control signal (*i.e.*, trajectory of dog locomotion) was the same for the three motions. The skating intensity increases as the number of parameters decreases. And we expect that an objective metric is able to penalize motion with foot skating. For the FMD, which measures the discrepancy between the ground truth and synthesized motion distributions, an effective FMD is expected to increase with the skating intensity.

3.2 Fréchet Motion Distance

While the evaluation in [11, 42] relies on a trained action classifier of human motion to compute the Fréchet distance between the ground truth and generated motion dataset, the method in [27, 44] trains an autoencoder-based feature extractor in an unsupervised manner. Using this autoencoder-based method, there is no need for labeled motion data to build a feature extractor which is essential in FMD. We performed our analyzes on this method with the architecture proposed in [27] since we focus on locomotion data where no extra label (*e.g.*, action class) exist.

[2] https://www.youtube.com/watch?v=kXnghjpyj_U

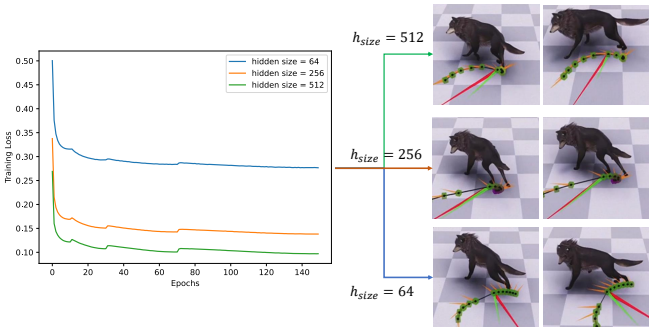

Figure 3: Training loss of the model implemented in [46]. Reducing h_{size} impacts negatively the training curve and leads to more foot skating in the resulting motion. The dog motion becomes stiff when removing network parameters.

The overview of the method to compute FMD is in Figure 1. The motion is m is represented as a sequence of pose $m_0, ..., m_{F-1}$. Each pose is a set of bone position p, orientation r, and position velocity v. The orientations are expressed in exponential map. We followed the strategy of representing feature vectors as images introduced in [27]. But in the previous work, only the Cartesian positions were used to represent the motion features, so the motion was converted into an image following the similarity between the Cartesian and RGB space: $x \equiv R$, $y \equiv G$, $x \equiv B$. In our case, we extended this method for orientations and velocities. The same conversion is used for the velocities since it is defined in Cartesian space. Considering the orientations, the exponential maps representation $\in \mathbb{R}^3$ which is suitable for the motion-to-image conversion, here denoted as I.

A ResNet-34-based autoencoder is first trained to reconstruct the initial image $I(m)$. The motion dataset is split into the training (80% of the whole dataset) and validation set (20%). Then, the set of ground truth m and synthesized motions $\tilde{m}$ are encoded into latent vectors. This latent space embodies an efficient set of motion features that is able to identify motion polluted by spatial noise and temporal discontinuities when computing the Fréchet distance between the latent spaces of the ground truth and artificially degraded motion [27]. In our work, we aim to analyze this metric behavior on motion data polluted by foot skating *i.e.*, determine if the FMD computed between the latent spaces is sensitive to the skating intensity. In this case, the synthesized motion $\tilde{m} = m_{h_{size}}$ where $m_{h_{size}}$ denotes the motion produced by the model with the hidden layer size = h_{size}. Similarly, $FMD_{h_{size}}$ represents the Fréchet distance between the latent vectors of the ground truth validation set and the motion generated by the model whose hidden layer size is h_{size}.

4 RESULTS AND PERSPECTIVES

The results of the evaluation are shown in Table 1. We observed that the FMD did not assess the quality of the motion samples successfully: the best score, *i.e.*, the smallest FMD, is given to the motion dataset with $h_{size} = 256$ which is uncorrelated with the

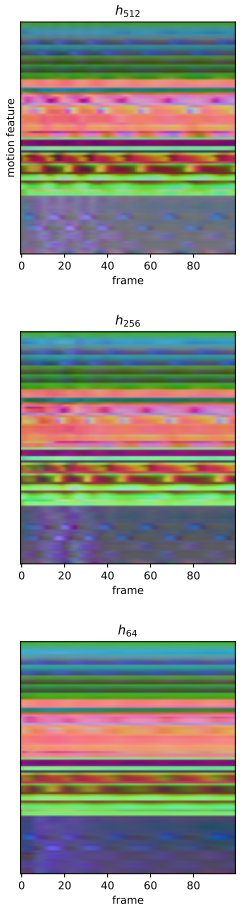

Figure 4: Samples of motion converted into an image representation: $I(m_{512})$, $I(m_{256})$ and $I(m_{64})$. Since each motion feature is expressed in $\mathbb{R}^3$, the RGB conversion is straightforward: the image width and height respectively represent the motion features and the temporal dimension. The set of features (position, orientation and position velocity) are stacked vertically. The stiffness of the motion when decreasing h_{size} is also observed in the image: the horizontal color transition is smoother pointing out the unnatural rigidity of motion.

intensity of the foot skating degradation. The latent representation learned during the training procedure of the autoencoder in [27] can be a major factor that makes that method not reliable to capture relevant motion information so that the FMD is sensitive to foot skating artifacts. In further research, it may be necessary to find a latent space or embeddings that are suitable for the evaluation of motion generative models. We need to carefully design this evaluation metric to measure precisely the performance of these models with respect to foot skating artifacts

However, foot skating is not the only perturbation generative models produce on the resulting motion. We may extend our analysis to other common artifacts such as pose freezing, sudden turning, or unbalanced poses. In interactive avatar context, since it seems

Generation model size (h_{size})	Fréchet Motion Distance ↓
64	89.99
256	**83.18**
512	101.60

Table 1: Fréchet Motion Distance evaluation of motion generated by model with $h_{size} = 64, 256$ and 512 (lower is better). The FMD in [27] is not proportional to foot skating intensity

impossible to list exhaustively all the artifacts that a neural network might induce on the synthesized motion, large-scale user studies are often performed to identify which model performs better in term of motion generation as in [45]. A relevant objective metric should be correlated with the human perception on the motion samples quality.

Finally, the inefficiency of the latent representation might come from the encoding process: The autoencoder used in this work is based on the ResNet-34 architecture [12]. Recent works have addressed and improved the encoding motion into a latent compact representation process as in [32]. Taking advantage of improved autoencoder architectures might lead to a better understanding on the latent information.

5 CONCLUSION

The evaluation of motion-generative models requires objective metrics that are capable of effectively penalizing any degradation in the motion samples. We validated the FMD proposed in [27] with more realistic motion artifacts, foot skating, and showed that FMD was not successful in capturing foot skating artifacts, a common perturbation in deep neural animation, which highlights the need for further metric development that can identify and account for such nuances. Developing an objective metric for evaluating the performance of motion-generative models is advantageous for designing interactive and immersive applications that involve human-avatar interaction. This is because the naturalness and realism of avatar animation significantly contribute to enhancing the overall user experience. Any new objective metric developed should consider these types of analyses to ensure that there is no bias or mismatch between human judgment and the designed objective score. The ultimate goal of developing such metrics is to facilitate the creation of motion-generative models that can produce high-quality, realistic motion that is indistinguishable from real-world motion.

REFERENCES

[1] Ho Yin Au, Jie Chen, Junkun Jiang, and Yike Guo. 2022. ChoreoGraph: Music-Conditioned Automatic Dance Choreography over a Style and Tempo Consistent Dynamic Graph. In *Proceedings of the 30th ACM International Conference on Multimedia* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 3917–3925. https://doi.org/10.1145/3503161.3547797

[2] Uttaran Bhattacharya, Elizabeth Childs, Nicholas Rewkowski, and Dinesh Manocha. 2021. Speech2AffectiveGestures: Synthesizing Co-Speech Gestures with Generative Adversarial Affective Expression Learning. In *Proceedings of the 29th ACM International Conference on Multimedia* (Virtual Event, China) *(MM '21)*. Association for Computing Machinery, New York, NY, USA, 2027–2036. https://doi.org/10.1145/3474085.3475223

[3] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. 2018. Demystifying MMD GANs. In *International Conference on Learning Representations*. https://openreview.net/forum?id=r1lUOzWCW

[4] Ali Borji. 2018. Pros and Cons of GAN Evaluation Measures. *CoRR* abs/1802.03446 (2018). arXiv:1802.03446 http://arxiv.org/abs/1802.03446

[5] Ziyi Chang, Edmund J. C. Findlay, Haozheng Zhang, and Hubert P. H. Shum. 2022. Unifying Human Motion Synthesis and Style Transfer with Denoising Diffusion Probabilistic Models. In *Proceedings of the 2023 International Conference on Computer Graphics Theory and Applications*.

[6] Min Jin Chong and David Forsyth. 2020. Effectively unbiased fid and inception score and where to find them. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6070–6079.

[7] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 8780–8794. https://proceedings.neurips.cc/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf

[8] D.C Dowson and B.V Landau. 1982. The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis* 12, 3 (1982), 450–455. https://doi.org/10.1016/0047-259X(82)90077-X

[9] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. 2015. Recurrent Network Models for Human Dynamics. *2015 IEEE International Conference on Computer Vision (ICCV)* (2015), 4346–4354.

[10] Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F Troje, and Marc-André Carbonneau. 2022. ZeroEGGS: Zero-shot Example-based Gesture Generation from Speech. *arXiv preprint arXiv:2209.07556* (2022).

[11] 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.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

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

[14] Daniel Holden, Taku Komura, and Jun Saito. 2017. Phase-Functioned Neural Networks for Character Control. *ACM Trans. Graph.* 36, 4, Article 42 (jul 2017), 13 pages. https://doi.org/10.1145/3072959.3073663

[15] Ruozi Huang, Huang Hu, Wei Wu, Kei Sawada, Mi Zhang, and Daxin Jiang. 2020. Dance revolution: Long-term dance generation with music via curriculum learning. *arXiv preprint arXiv:2006.06119* (2020).

[16] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive Mixtures of Local Experts. *Neural Computation* 3, 1 (03 1991), 79–87. https://doi.org/10.1162/neco.1991.3.1.79 arXiv:https://direct.mit.edu/neco/article-pdf/3/1/79/812104/neco.1991.3.1.79.pdf

[17] A Jain, AR Zamir, S Savarese, and A Saxena. 2016. Deep learning on spatio-temporal graphs. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA*. 27–30.

[18] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2019. Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms.. In *INTERSPEECH*. 2350–2354.

[19] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2019. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems* 32 (2019).

[20] Marc Erich Latoschik, Daniel Roth, Dominik Gall, Jascha Achenbach, Thomas Waltemate, and Mario Botsch. 2017. The effect of avatar realism in immersive social virtual realities. In *Proceedings of the 23rd ACM symposium on virtual reality software and technology*. 1–10.

[21] Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. 2020. Learning to generate diverse dance motions with transformer. *arXiv preprint arXiv:2008.08171* (2020).

[22] Peizhuo Li, Kfir Aberman, Zihan Zhang, Rana Hanocka, and Olga Sorkine-Hornung. 2022. GANimator: Neural Motion Synthesis from a Single Sequence. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 138.

[23] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. 2021. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13401–13412.

[24] Shaohui Liu, Yi Wei, Jiwen Lu, and Jie Zhou. 2018. An improved evaluation framework for generative adversarial networks. *arXiv preprint arXiv:1803.07474* (2018).

[25] Joan Llobera and Caecilia Charbonnier. 2021. Interactive characters for virtual reality stories. In *ACM International Conference on Interactive Media Experiences*. 322–325.

[26] Lorenzo Luzi, Carlos Ortiz Marrero, Nile Wynar, Richard G Baraniuk, and Michael J Henry. 2023. Evaluating generative networks using Gaussian mixtures of image features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 279–288.

[27] Antoine Maiorca, Youngwoo Yoon, and Thierry Dutoit. 2022. Evaluating the Quality of a Synthesized Motion with the Fréchet Motion Distance. In *ACM SIGGRAPH 2022 Posters* (Vancouver, BC, Canada) *(SIGGRAPH '22)*. Association

for Computing Machinery, New York, NY, USA, Article 9, 2 pages. https://doi.org/10.1145/3532719.3543228

[28] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. 2019. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9489–9497.

[29] Julieta Martinez, Michael Black, and Javier Romero. 2017. On Human Motion Prediction Using Recurrent Neural Networks. 4674–4683. https://doi.org/10.1109/CVPR.2017.497

[30] Qianhui Men, Hubert P.H. Shum, Edmond S.L. Ho, and Howard Leung. 2022. GAN-based reactive motion synthesis with class-aware discriminators for human–human interaction. *Computers & Graphics* 102 (2022), 634–645. https://doi.org/10.1016/j.cag.2021.09.014

[31] Stanislav Morozov, Andrey Voynov, and Artem Babenko. 2020. On self-supervised image representations for GAN evaluation. In *International Conference on Learning Representations*.

[32] 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.

[33] Nicolas Robitaille, Philip L Jackson, Luc J Hébert, Catherine Mercier, Laurent J Bouyer, Shirley Fecteau, Carol L Richards, and Bradford J McFadyen. 2017. A Virtual Reality avatar interaction (VRai) platform to assess residual executive dysfunction in active military personnel with previous mild traumatic brain injury: proof of concept. *Disability and Rehabilitation: Assistive Technology* 12, 7 (2017), 758–764.

[34] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. 2018. Assessing generative models via precision and recall. *Advances in neural information processing systems* 31 (2018).

[35] Nisarg A Shah and Gaurav Bharaj. 2022. Towards Device Efficient Conditional Image Generation. (2022).

[36] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the Inception Architecture for Computer Vision. *CoRR* abs/1512.00567 (2015). arXiv:1512.00567 http://arxiv.org/abs/1512.00567

[37] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Amit H Bermano, and Daniel Cohen-Or. 2022. Human Motion Diffusion Model. *arXiv preprint arXiv:2209.14916* (2022).

[38] Joëlle Tilmanne, Alexis Moinet, and Thierry Dutoit. 2012. Stylistic gait synthesis based on hidden Markov models. *EURASIP Journal on Advances in Signal Processing* 2012 (2012), 1–14.

[39] Matthijs van der Boon, Leonor Fermoselle, Frank ter Haar, Sylvie Dijkstra-Soudarissanane, and Omar Niamut. 2022. Deep Learning Augmented Realistic Avatars for Social VR Human Representation. In *ACM International Conference on Interactive Media Experiences* (Aveiro, JB, Portugal) *(IMX '22)*. Association for Computing Machinery, New York, NY, USA, 311–318. https://doi.org/10.1145/3505284.3532976

[40] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. 2021. Scene-aware Generative Network for Human Motion Synthesis. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 12201–12210.

[41] Wang Xi, Guillaume Devineau, Fabien Moutarde, and Jie Yang. 2020. Generative model for skeletal human movements based on conditional DC-GAN applied to pseudo-images. *Algorithms* 13, 12 (2020), 319.

[42] Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. 2019. Convolutional Sequence Generation for Skeleton-Based Action Synthesis. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 4393–4401. https://doi.org/10.1109/ICCV.2019.00449

[43] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (Apr. 2018). https://doi.org/10.1609/aaai.v32i1.12328

[44] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity. *ACM Transactions on Graphics* 39, 6 (2020).

[45] Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2022. The GENEA Challenge 2022: A large evaluation of data-driven co-speech gesture generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 736–747.

[46] He Zhang, Sebastian Starke, Taku Komura, and Jun Saito. 2018. Mode-Adaptive Neural Networks for Quadruped Motion Control. *ACM Trans. Graph.* 37, 4, Article 145 (jul 2018), 11 pages. https://doi.org/10.1145/3197517.3201366

[47] 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).

[48] Wenlin Zhuang, Congyi Wang, Jinxiang Chai, Yangang Wang, Ming Shao, and Siyu Xia. 2022. Music2dance: Dancenet for music-driven dance generation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18, 2 (2022), 1–21.