Objective Evaluation Metric for Motion Generative Models: Validating Fréchet Motion Distance on Foot Skating and Over-smoothing Artifacts

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

Hugo Bohy
hugo.bohy@umons.ac.be
ISIA Lab, University of Mons
Mons, Hainaut, Belgium

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

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

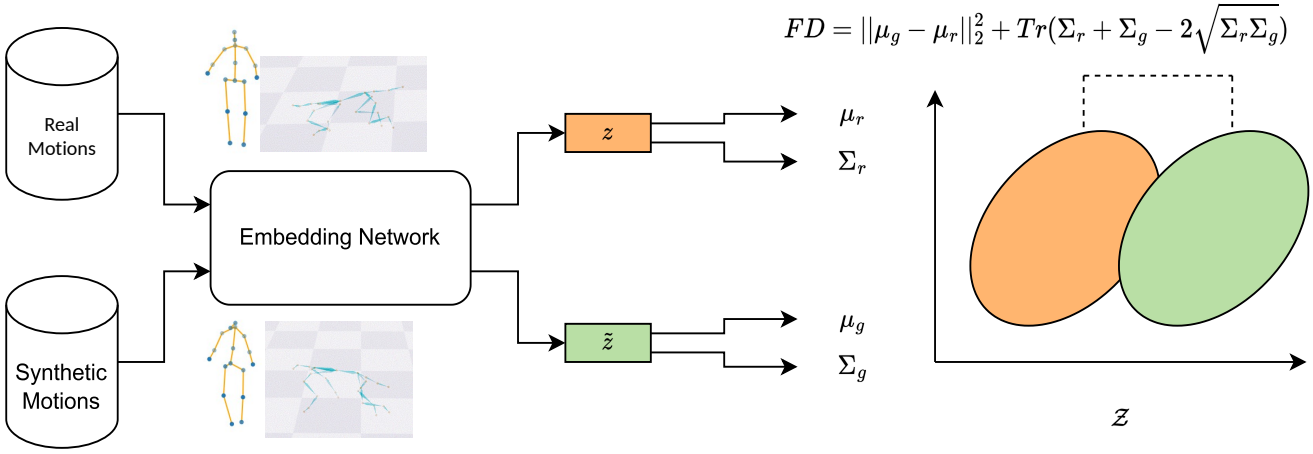

Figure 1: Overview of the objective evaluation metric using Fréchet distance (FD) between the distributions of embeddings of the synthetic and real motions. This gives a score that assesses the quality and diversity of the generated motion samples, allowing the evaluation and comparison of motion-generative models. As an embedding network that maps raw motion data into latent feature spaces, two autoencoder architectures (one with 1D convolutions and another one with Transformer layers) are tested.

ABSTRACT

Nowadays, Deep Learning-powered generative models are able to generate new synthetic samples nearly indistinguishable from natural data. The development of such systems necessarily involves the design of evaluation protocols to assess their performance. Quantitative objective metrics, such as Fréchet distance, in addition to human-centered subjective surveys, have become a standard for evaluating generative algorithms. Although motion generation is a popular research field, only a few works addressed the problem of the design and validation of a robust objective evaluation metric for motion-generative models. These previous works proposed to degrade ground truth motion samples with synthetic noises (*e.g.*, Gaussian, Salt& Pepper) and studied the behavior of the proposed metric. However, this degradation does not mimic common motion artifacts produced by generative models. In this work, we propose (1) to validate Fréchet distance-based objective metrics on motion datasets degraded by two realistic motion artifacts, *foot skating* and *over-smoothing*, often found in motion synthesis results, and (2) a *Fréchet Motion Distance* (FMD), using Transformer-based feature extractor, able to capture the motion artifacts and also robust towards the variation of motion length.

MIG '23, November 15–17, 2023, Rennes, France

ACM ISBN 979-8-4007-0393-5/23/11.
https://doi.org/10.1145/3623264.3624443

CCS CONCEPTS

• **Computing methodologies** → **Motion processing**; • **Computer systems organization** → *Robotics*; • **Networks** → Network reliability.

KEYWORDS

Objective evaluation metric, Motion-generative model evaluation, Fréchet motion distance

ACM Reference Format:

Antoine Maiorca, Hugo Bohy, Youngwoo Yoon, and Thierry Dutoit. 2023. Objective Evaluation Metric for Motion Generative Models: Validating Fréchet Motion Distance on Foot Skating and Over-smoothing Artifacts. In *ACM SIGGRAPH Conference on Motion, Interaction and Games (MIG '23), November 15–17, 2023, Rennes, France.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3623264.3624443

1 INTRODUCTION

Generative models are designed to produce new data samples that are similar to real samples. Recent deep learning (DL)-based models have the ability to generate unseen but plausible modalities, enabling them to create realistic and meaningful samples in diverse domains such as images [Gao et al. 2023; Kim et al. 2022; Ramesh et al. 2022; Rombach et al. 2021], texts [Brown et al. 2020; Guo et al. 2017; Wolf et al. 2020] and audio [Dhariwal et al. 2020; Lee et al. 2021; Leng et al. 2022]. In a motion generation context, the synthesis of realistic motion is a relevant task in a broad range of industrial applications such as in robotics or video games industry. Speech- [Yoon et al. 2020], text- [Petrovich et al. 2021] or music-driven [Starke et al. 2022] motion generation, motion style transfer [Aberman et al. 2020a,b] or trajectory-controlled locomotion [Holden et al. 2017; Ling et al. 2020; Zhang et al. 2018] are examples of popular use cases in this context.

An essential element in the development of generative models is the establishment of an appropriate evaluation metric that measures the performance of the model. This evaluation typically considers factors such as *quality*, *i.e.*, the similarity between the generated and real samples, as well as *diversity* of the generated samples in terms of their variability. For generative tasks, objective evaluation is not straightforward unlike other paradigms that have ground truths such as classification. The nature of one-to-many relationships in generative tasks (for example, many different but plausible dance motions can be synthesized for the same input music) and the human subjectivity regarding the quality of the synthetic modalities are factors that hinder the generative model evaluation and comparison.

Subjective surveys are a common method to assess the performance of generative models [Tilmanne et al. 2012; Yoon et al. 2022]. It involves human judges that rate the plausibility and naturalness of synthetic samples. However, this type of evaluation has significant drawbacks as pointed out in [Borji 2018]. First, relying on human judgment implies, in essence, being sensitive to the subjectivity of judges. Indeed, some individuals might not have the same perception of the quality of the generated modality, which will result in a large variance in the collected results if the number of participants is not sufficient. In addition, gathering a large group of judges requires a non-neglectable amount of time and resources to conduct the evaluation. This method of evaluation also suffers from low reproducibility issues that hinder comparison with other previous methods. For these reasons, objective evaluation metrics are needed that evaluate qualitatively and at low cost. The need for evaluation metrics is even greater in recent learning-based generative models, as they involve numerous training attempts. An evaluation tool to compare the attempts is necessary for faster development.

In that sense, objective evaluation metrics play a crucial role in assessing the quality and diversity of the generated samples in various domains. In the language field, *perplexity* is a widely used metric that measures confidence in the prediction of the model and has proven to be effective [Liu et al. 2019; Shoeybi et al. 2019; Xu et al. 2020]. Task-specific metrics like *BLEU* [Papineni et al. 2002] and *ROGUE* [Lin 2004] are popular in machine translation and text summarization, respectively [Aghajanyan et al. 2020; Vaswani et al. 2017]. In the image synthesis domain, the *Inception Score* (IS) [Salimans et al. 2016] is used to assess the quality of generated images [Berthelot et al. 2017; Ma et al. 2017]. IS uses the *Inception* image classifier to compute the Kullback-Leibler (KL) divergence between the conditional and marginal distributions estimated by the classifier. However, IS may overlook mode collapses, where generated samples appear natural but lack diversity within specific categories. To address this, the *Fréchet Distance* (FD) [Dowson and Landau 1982] was introduced as an objective evaluation metric. The FD measures the distance between the distribution of the embeddings of real and synthesized samples. The calculation of FD involves the pairs of mean and covariance matrix, denoted (μ_r, Σ_r) and (μ_g, Σ_g), respectively, for the distributions of the embeddings of real and generated samples (see Equation 1). This equation assumes that these distributions are Gaussian.

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

In practice, the real and synthesized embeddings are obtained by extracting the last activation maps from the *Inception* network in the image synthesis field. Distribution parameters are estimated using these feature maps, and the FD, called *Fréchet Inception Distance* (FID), is computed using these parameters. FID has become the standard metric for evaluating image synthesis models [Dhariwal and Nichol 2021; Shah and Bharaj 2022]. A similar approach, known as *Fréchet Audio Distance* (FAD) [Kilgour et al. 2019], has been adopted in the field of audio synthesis.

In motion synthesis, FD-based metrics, as shown in Figure 1, are commonly used to evaluate the quality and diversity of generated motion samples [Guo et al. 2020; Wang et al. 2021; Yan et al. 2019]. However, only a limited number of studies have thoroughly validated the proposed metrics [Maiorca et al. 2022b; Yoon et al. 2020]. These validation protocols typically involve the degradation of motion samples using synthetic noise such as Gaussian or Salt&Pepper. The expected behavior of the score is that the FD increases (higher is worse) according to the intensity of the noises. However, this synthetic motion degradation method does not necessarily capture the common artifacts produced by DL-based motion-generative models. To address this limitation, it is crucial to develop evaluation metrics that are sensitive to real motion artifacts, making them more applicable to a large-scale evaluation of motion-generative models. Such a metric should effectively capture the specific types of artifacts introduced by DL-based motion-generative models, enabling more accurate and relevant assessments of their performance.

Lastly, an objective evaluation metric should be aligned with human perception of the synthesized modalities, since the ultimate

goal of generative models is to generate synthetic samples that are plausible to humans. Therefore, in this work we measure the correlations between the designed objective metric and subjective human ratings.

In this work, we:

- Introduce a validation protocol that involves common motion artifacts: (1) *foot skating* and (2) *over-smoothing*. A motion dataset is polluted by different intensities of both artifacts, and we analyze objective metrics' responses to the degraded motions. To the best of our knowledge, this is the first attempt to validate objective metrics using common motion artifacts.
- Propose a Transformer-based autoencoder (AE) as an unsupervised feature extractor for FD-based metrics. We show that the proposed setting is more robust than the previous one [Yoon et al. 2020] to the variation of motion length, which is a desired property of an objective metric.

2 RELATED WORK

2.1 Evaluating Motion Generative Model

One common method to assess the motion-generative model is by conducting subjective evaluation protocols. These rely on human-based surveys that rate the quality of the generated motion samples. These methods can be applied in various contexts, such as in stylized walk cycle generation to assess expression naturalness [Tilmanne et al. 2012] or in music-driven motion generation to measure rhythmic and aesthetic consistency of motion samples [Au et al. 2022; Huang et al. 2020; Li et al. 2020, 2021b; Zhuang et al. 2022].

In parallel, various methods have been utilized to evaluate the plausibility of synthetic motions based on objective measures. Regarding human motion prediction whose task is to forecast future frames given a motion prefix, the *Euclidean distance* is computed between the few generated frames and the ground truth [Fragkiadaki et al. 2015; Jain et al. 2016; Mao et al. 2019; Martinez et al. 2017]. It measures the ability of the network to predict precisely the future poses. The *Diversity* and *Modality* metrics are common in the field of text-driven motion generation to evaluate if the algorithm is able to generate a wide and diverse range of motion samples from a text prompt [Tevet et al. 2022; Zhang et al. 2022]. In trajectory-controlled animation of virtual characters, *Foot Skating* is an artifact that is perceived as the character's foot sliding when on contact with the ground. This can be measured considering the foot velocity below a height threshold [Zhang et al. 2018]. However, these metrics are only relevant in specific tasks and cannot be applied in every motion generation context. Moreover, they do not encapsulate solely the naturalness and likeliness of the generated samples perceived by human judges.

Then, an evaluation metric based on Fréchet Distance [Dowson and Landau 1982] has first been introduced in [Heusel et al. 2017] to assess the performance of Generative Adversarial Networks (GANs) on image modalities and is called the Fréchet Inception Distance (FID). It relies on *InceptionV3* [Szegedy et al. 2015] to extract the activation maps from the last convolution layer. The mean and covariance matrices are estimated on these features and the FID score is further computed by Equation 1. Its strength comes from its ability to penalize degradation of synthesized samples, as well as the lack of diversity in the generated samples. Since its biased have been exposed *e.g.*, the mismatch between human perception and objective score when it comes to evaluate an unseen dataset [Morozov et al. 2020], the FID sensitivity to the number of samples composing the dataset or even the particular generative model being evaluated [Chong and Forsyth 2020], many works have proposed an improvement of this metric reducing the impact of these biases on the score [Bińkowski et al. 2018; Chong and Forsyth 2020; Liu et al. 2018; Luzi et al. 2023].

In recent studies aforementioned in Section 1, scores built on the Fréchet Distance have been widely used to assess the performance of motion-generative models. In this case, the feature manifold embodies relevant information on the motion modality. For action motion generation, in the same philosophy as FID, an action classifier [Yan et al. 2018] is trained and the mean and covariance matrix pairs are estimated on activation maps [Chang et al. 2022; Guo et al. 2020; Yan et al. 2019]. A fine-tuned version of *InceptionV3* is also used to evaluate the action generative model using the Cartesian position to RGB domain mapping ($\mathbb{R}^3 \rightarrow \mathbb{R}^3$) [Xi et al. 2020].

However, this FD-based metric requires a training procedure with labeled data, which makes this method unreliable to evaluate motion where no defined action is performed such as in co-speech gesticulation. To tackle this problem, an unsupervised training paradigm is used: an LSTM based motion prediction network [Fragkiadaki et al. 2015] is trained to predict the next frames of motion. The evaluation is performed using hidden states of the LSTM [Wang et al. 2021]. Automatic generation of co-speech gesture evaluation uses AE to learn a motion latent space [Yoon et al. 2020, 2022]. The statistics are estimated on this manifold, and the FD measures the distance between the set of ground truth and synthesized latent vectors. However, these metrics suffer from the lack of a validation protocol that aims to analyze the behavior of the score to any perturbation perceived on generated samples. Yoon *et al.* [Yoon et al. 2020] validated the score based on the degradation of ground truth motion samples by synthetic noise such as Gaussian or Salt&Pepper.

More importantly, Kucherenko *et al.* [Kucherenko et al. 2023] analyzes the rank correlation between several objective metrics and user-based subjective evaluation in the context of co-speech upper-body gesture evaluation. They claim that, among the objective metrics tested, only the FD score proposed in [Yoon et al. 2020] achieves a statistically significant effect.

2.2 Transformer-based AutoEncoders

With the Transformer came a new area of AE. Transformer architecture has first been introduced in the *Natural Language Processing* (NLP) context [Vaswani et al. 2017] and takes advantage of self-attention mechanisms. It allows the model to assign different weights or importance to different positions in the input sequence. The Transformer's encoder focuses on relevant parts of the sequence while processing it. The decoder attends to the encoded representations to extract relevant information to complete the defined task, *e.g.*, generating the next word. The input sequence of words $s = [s_0, ..., s_T]$ is first converted into tokens $k = [k_0, ..., k_T]$ for computational and representation purposes. This family of neural networks achieves state-of-the-art performance in NLP [Brown et al. 2020; Devlin et al. 2018], computer vision [Bao et al. 2021; Gao

et al. 2023] or audio processing [Baade et al. 2022; Baevski et al. 2020; Gong et al. 2022; Huang et al. 2022a,b].

BERT [Devlin et al. 2018] is a Transformer-based architecture that introduced the use of a *cls* (classification) token in addition to k. This token supposedly holds enough information of the entire input to discriminate between several classes. In this case, the *cls* is used to classify whether two sentences are paired or not. The *cls* token has shown its efficiency in various domains such as NLP [Beltagy et al. 2020] or in computer vision [Dosovitskiy et al. 2020; Yu et al. 2022].

In the field of motion synthesis, frameworks employ Transformers to leverage the human motion synthesis problem in various context [Aksan et al. 2020; Hou et al. 2023; Wang et al. 2022a]. action-conditioned human motion synthesis is performed using a combination of Transformer-based architecture and variational encoding [Petrovich et al. 2021; Wang et al. 2022b]. The action tags are first embedded and then processed by the encoder. Two independent fully-connected layers are used to regress the distribution parameters (μ, σ) based on the output of the encoder corresponding to the action token. Our method is inspired by the one proposed in [Petrovich et al. 2021]. Instead of encoding action tags, we add a *cls* token to the input sequence in order to combine the information contained in it and the capability of the decoder to reconstruct the input motion sequence. This method is explained in Section 3.

3 PROPOSED OBJECTIVE EVALUATION METRIC

We propose an FD-based objective score using a Transformer architecture to extract a latent representation of motion data. More precisely, we employed a modified version of the human action conditioned Transformer architecture [Petrovich et al. 2021], which consists of two components: encoder and decoder. The encoder takes an input sequence and processes it by applying multiple layers of self-attention and feed-forward neural networks. The decoder, on the other hand, generates an output sequence based on the encoded representation, attending to the relevant parts of the input during the decoding process.

In the original work [Petrovich et al. 2021], the Transformer-based model works as a VAE where the action label is first embedded into learnable mean and variance tokens, respectively μ^a_{token} and Σ^a_{token}. These are further aligned with the input motion through the Transformer encoder, and the KL loss is computed on the encoded tokens (μ^a and Σ^a) to fit the statistics of the desired normal distribution $\mathcal{N}(0, 1)$. However, in our context, no action labels are provided with the motion samples. The original framework is hence modified to fit this requirement. Instead of encoding action tokens, a latent vector *cls* is learned during the training process. Then, the encoded *cls*, denoted as z, guides the reconstruction of the input motion samples in the decoder. Hence, the vector z acts as a latent representation of the input motion as in AE architectures. The related architecture is presented in Figure 2.

The input motion is denoted as m and is a sequence of poses $m_0, ..., m_T$. m is first projected linearly. Then, the poses embeddings and the *cls* vector are concatenated and feed the positional encoding (PE) layer. In Transformer-based architectures, PE is a technique that is used to provide information about the positions of elements

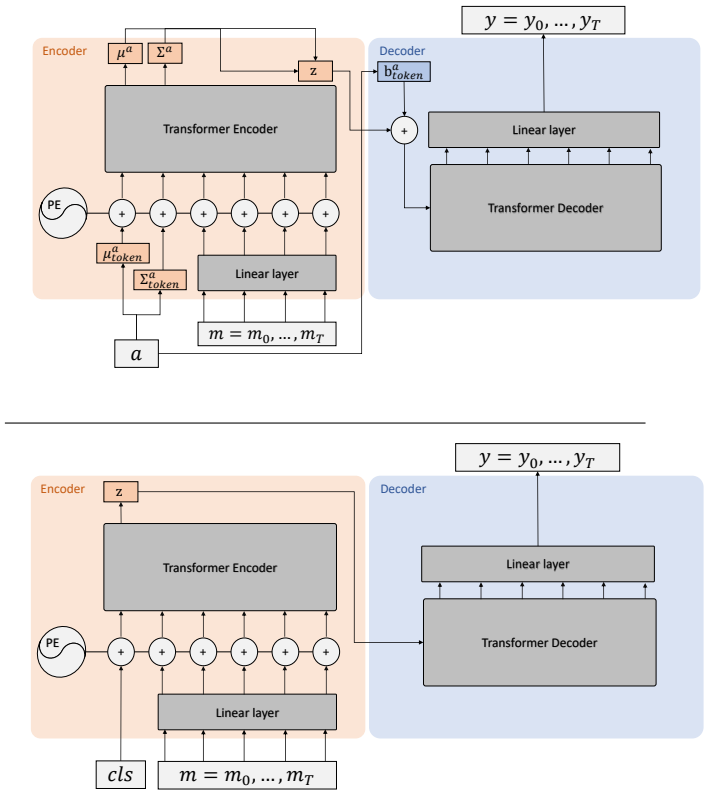

Figure 2: Top: action-conditioned Transformer-based VAE [Petrovich et al. 2021]. - Bottom: our method. The action token is replaced by the *cls* and the variational behavior of the initial network is removed. Our method aims to learn a latent representation of the motion high-dimensional space so that the latent space can be used to extract the statistics needed to compute the FD.

in an input sequence. The purpose of positional encoding is to inject position-related information into the input embeddings before feeding them into the Transformer network. This allows the model to consider the order and position of the poses in the sequence during self-attention computations. Next, the decoder, with the information provided by the latent vector z, aims to reconstruct the input motion m. The mean squared error loss is computed between the decoder output y and the input motion m during the training process.

Finally, the latent vectors z compose a low-dimensional latent space learned to efficiently represent the high-dimensional motion samples information. FD statistics *i.e.*, μ and Σ for the real and generated motion datasets are estimated in this latent space. The FMD is then computed, giving a score on the quality and diversity of the synthesized motion datasets.

4 ANALYZES

4.1 Datasets

The experiments were performed on two different motion capture (MoCap) datasets of different nature. These datasets have been curated with the specific goal of addressing two key challenges: firstly, the animation of quadrupeds based on their trajectories, and secondly, the development of speech-driven motion generation models. By leveraging these datasets, we can effectively demonstrate the robustness of our proposed FMD in two critical aspects: (1) its ability to handle variations in skeleton structure and (2) its adaptability to different types of motion samples.

4.1.1 Dog Locomotion Dataset. The dog locomotion dataset [Zhang et al. 2018] is a MoCap dataset composed of various quadruped locomotion gaits *e.g.*, walking or running, as well as a few actions such as lying or drinking. 27 bones structure the dog skeleton. Moreover, the pose features for each joint j are the Cartesian positions, the orientations, expressed in exponential maps, towards the forward and upward direction and velocities.

4.1.2 Human Gesture Motion Dataset. The GENEA (Generation and Evaluation of Non-verbal Behaviour for Embodied Agents) Challenge [Yoon et al. 2022] was hosted to compare speech-driven gesture generation systems. Ten participating teams designed their own co-speech gesture-generation system built on the same speech motion dataset modified from the Talking With Hands 16.2M dataset [Lee et al. 2019]. The challenge dataset has 18 hours of full-body motion capture, including fingers, of different people engaging in dyadic conversation. The challenge had two tiers: full- and upper-body gesticulation. For each tier, both the human-likeness of gesture motion and its appropriateness for the specific speech have been evaluated through large-scale subjective human surveys [Kucherenko et al. 2023]. Therefore, these sets of human and synthesized motion samples and their human ratings are useful resources for analyzing the correlation between the proposed objective metric, FMD, and human preferences. The evaluation of this dataset was based on the 3D Cartesian coordinates of the motion. The skeleton for the upper- and full-body tiers is composed of, respectively, 54 and 58 joints.

The evaluation in [Kucherenko et al. 2023] employs a specific syntax to identify each set of motion. For upper-body gestures, **UNA** tier gathers real human gestures, **UBA** and **UBT** are set of motion generated by two models considered as baselines for the challenge, and **US(J—Q)** are tiers submitted by each participating team. Similarly, for full-body gestures, **FNA** is the set of real human motion, **FBT** the motion samples of the baseline and the synthetic motion produced by each proposed method are referred to as **FS(B—I)**. **FSE** is absent because the **E** team decided to withdraw their submission.

4.2 Motion Artifacts

4.2.1 Foot Skating. When animating virtual characters using DL-powered methods, several motion artifacts can arise due to the complexity and challenges of modeling human or quadruped motion. Among others, *foot skating* has been shown to be a common artifact in deep neural animation, especially in trajectory-controlled locomotion [Henter et al. 2020; Ling et al. 2020; Zhang et al. 2018]. Foot skating is defined as the sliding of the character's foot when the joint is in contact with the ground. This is perceived as an artifact that degrades the motion quality as it is considered as an unnatural behavior. An example of foot skating is shown in Figure 3 and in videos here [1]. To create the motion dataset polluted with foot skating, we rely on a state-of-the-art deep neural motion-generative model that animates in real time a quadruped with different locomotion gaits given its trajectory controlled by an external user [Zhang et al. 2018]. This model architecture is made up 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 [Jacobs et al. 1991] is used to calculate the parameters θ of the regression network: θ is obtained by blending n expert parameters with the coefficients ω calculated 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. 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 [Zhang et al. 2018]).

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

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 as 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 as shown in [Maiorca et al. 2022a]. All materials used to animate the quadruped character come from here [2].

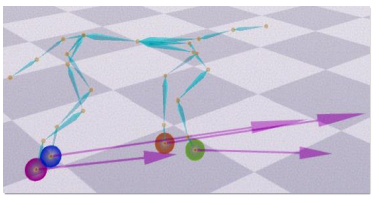

Figure 3: Visualization of foot skating artifacts. The purple arrow represents the foot velocity when in contact with the ground. Here, the skating is intense while the virtual quadruped is walking because the feet's velocity are high.

4.2.2 Motion Over-smoothing. In the context of motion generation, over-smoothing refers to an undesirable characteristic where the generated samples appear excessively smooth, lacking in sharpness and fine details. Over-smoothing can be seen as a result of excessive

[1] https://figshare.com/s/f4f2e64fac44f9b1bece
[2] https://github.com/pauzii/AnimationAuthoring

low-pass filtering. When a motion feature *e.g.*, Cartesian position is excessively low-pass filtered, the high-frequency details, which contribute to sharpness and fine variations, are attenuated. This leads to a smoothed-out appearance on the resulting motion. It is a common artifact encountered in motion-generative models [Chen et al. 2020; Li et al. 2021a]. To mimic the over-smoothing effect, the joint positions are processed by a 1D-Gaussian filter with different intensities ζ. The impact of this process on the resulting motion is shown in Figure 4.

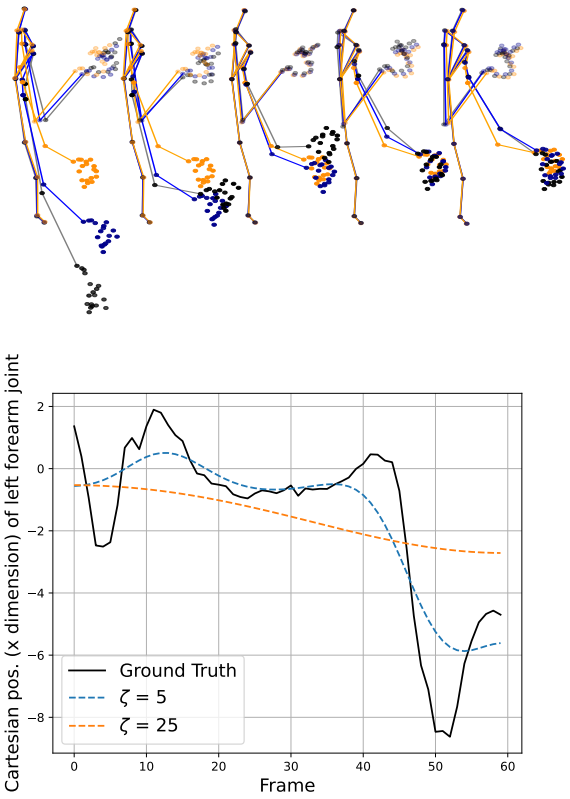

Figure 4: Top: Visualization of ground truth (in black) and filtered motion samples with $\zeta = 5$ and $\zeta = 25$ (in blue and orange respectively). Bottom: Impact on ζ parameter in a Gaussian filter smoothing on 1D signals. The motion sample looks unnatural when the smoothing process is too intense due to the loss of the high-frequency information. With $\zeta = 25$, the filtered motion looks almost static.

4.2.3 Subjective Study Correlation. While it seems impossible to list and measure precisely every type of motion artifact that can occur in the various motion-generative systems, the evaluation of these is often performed by user studies as we expect people to consider multiple factors comprehensively, including naturalness, pleasantness, and appropriateness to the input context. As explained in Section 4.1.2, the GENEA dataset allows us to explore and analyze the correlation between the proposed objective score and the human rates. More concretely, we compute the Kendall rank correlation [Kendall 1948] between the median user rates and FMDs in a similar way in [Kucherenko et al. 2023]. The Kendall-τ correlation is a statistical measure used to assess the strength and direction of association between two variables. It is particularly well-suited for data that involve rankings or ordinal scales, where the order or ranking of data points is more important than their actual numerical values.

4.3 Experimental Protocol

To evaluate the quality and diversity of a generated motion dataset, the FMD is computed between the distribution of the generated and ground truth embeddings using Equation 1. The embeddings are the samples z of the latent space computed from the Transformer AE. To validate if the proposed FMD efficiently evaluate motion-generative models, we test if the metric is sensitive to common motion artifacts. Moreover, since it appears difficult to exhaustively point out and reproduce each motion artifact that makes motion samples unpleasant for humans, we also measure the correlation between the results of the subjective evaluation protocol and the objective FMD score. We compare the results of our proposed objective evaluation metric and the method in [Yoon et al. 2020] that relies on Conv1D-based AE to learn a latent representation of motion data.

The motion dataset is split into fixed length samples. We consider them of length 15, 28, 30, 32, 45 and 60 frames. It allows us to challenge the robustness of the method towards the number of frames and the selection of 28 and 32 frames enables us to assess if motion of similar durations leads to similar FMD outcomes. In fact, the metric must not be sensitive to the number of motion frames. A robust metric ensures that generative models are evaluated based on their inherent motion generation capabilities, rather than being influenced by the length of the motions they produce.

80% and 20% of the dog locomotion set are respectively used as the training and validation set. Considering the human motion gesture dataset, the training set is composed of 18 h of gestures and 40 min for the validation set. We trained the proposed model with Pytorch/Fastai frameworks with a batch size of 256, during 100 epochs, estimating the learning rate and using the super-convergence training method [Smith and Topin 2019]. We stacked 8 attention + feed-forward encoder and decoder layers and employed multi-head attention with 8 heads.

5 RESULTS

5.1 Motion Reconstruction

The AEs were trained by reconstructing the input motion. Figure 5 presents the mean squared error (MSE) between the ground truth and reconstructed motion samples from the validation set for all the configurations tested. Videos about the motion reconstruction are shared here [3] [4]. We observed that the Transformer-AE is more powerful to reconstruct the input motion sample regardless of the length of the motion samples. We believe that this is due to self-attention mechanism that allows one to attend to relevant parts of the input sequence and capturing long-range dependencies.

[3] Co-speech gestures motion reconstruction video
[4] Dog locomotion reconstruction video

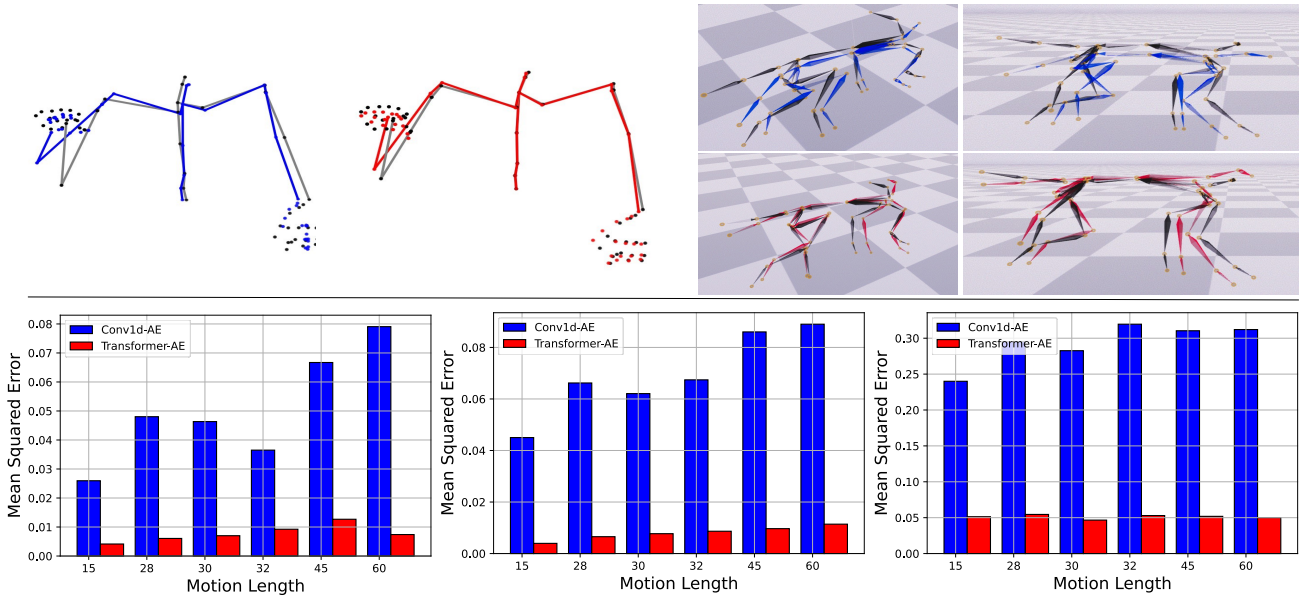

Figure 5: Top: Examples of pose reconstruction in upper body gesticulation (left) and dog walking (right). Bottom: Reconstruction error by Transformer- (red) and Conv1D- (blue) based AE on the validation set of upper- (left), full- (mid) body gestures and dog locomotion (right). The Transformer AE is better at reconstructing the input motion in every tested configurations.

5.2 Fréchet Motion Distance

The FMD is computed between each generated set of motion and the validation set of the MoCap dataset. Both mean-covariance pairs for the validation (real) and submitted (generated) set, respectively (μ_r, Σ_r) and (μ_g, Σ_g), are estimated on the latent dimension. The FMD is further computed by Equation 1.

5.2.1 Foot Skating. To validate the proposed metric on foot skating artifacts, we need to verify whether the FMD has an appropriate behavior toward this kind of artifacts *i.e.*, scoring higher motion samples with more intense foot skating. Figure 6 shows the evolution of the score with the hidden size of the animation model h_{size}. Reducing the size of the network increases the intensity of foot-skating of the resulting animation. First, the score from both models is sensitive to skating artifacts, except when decomposing the motion into 32-frame samples and using Conv1D-AE as feature extractor. Then, the scores given by the Transformer model are relatively steady regarding the motion frames compared to the scores measures using the convolutional model. Therefore, considering foot skating, the proposed Transformer-based FMD is more stable to the variation of motion length than the original FMD in [Yoon et al. 2020].

5.2.2 Over-smoothing. The dataset polluted by over-smoothing artifacts are the ground truth full- and upper-body gestures tiers **UNA** and **FNA** from the GENEA challenge. The FD is computed between the validation and the filtered dataset. We study the impact of the ζ parameter *i.e.*, the intensity of the smoothing process on the metric. Figure 6 shows the variation in the score with ζ. First, for both models, a plateau occurs in the score profile when analyzing motion samples of 15 frames for full- and upper-body gestures. This means that, in this configuration, the metric scores similarly dataset polluted by higher degree of over-smoothness. However, considering the analyzes for 28 to 60 frames, both scores are sensitive to the intensity of smoothing degradation. We observe that the motion length has less impact on the score value when the Transformer is employed as a feature extractor, which makes this method more stable to the variation of motion duration to over-smoothing artifacts.

5.2.3 User Rates Correlation. While it is essential for a metric to effectively capture perceptual motion artifacts, it needs to exhibit a significant correlation with human judgments. It should not only detect motion artifacts accurately but also align with how humans perceive and evaluate the plausibility of the motions. Table 1 presents the FMD score for each set of motion considering different length of motion samples. We use the subjective evaluation results presented in [Kucherenko et al. 2023]: the median user rates indicates the human-likeliness of the motion set (higher is better). Table 1 has been arranged in descending order according to user ratings. The motion samples with higher user scores appear at the top, while those with lower scores are listed towards the bottom. There was a system (**USQ**) rated higher than human motion (**UNA**); the authors in [Kucherenko et al. 2023] highlighted the multi-factor constraints that could explain that score *e.g.*, motion artifacts in the ground truth motions from the difficulty to record clean fingers motion in motion capture systems. The same observation is made concerning **FSA** and **FNA**.

We calculated the Kendall-τ correlation between the user and objective scores, reported in Figure 7. Each correlation value is negative since the FMD gives a low score for close synthetic motion distribution, but Figure 7 reports the absolute value of the correlation. We found a statistically significant correlation in the

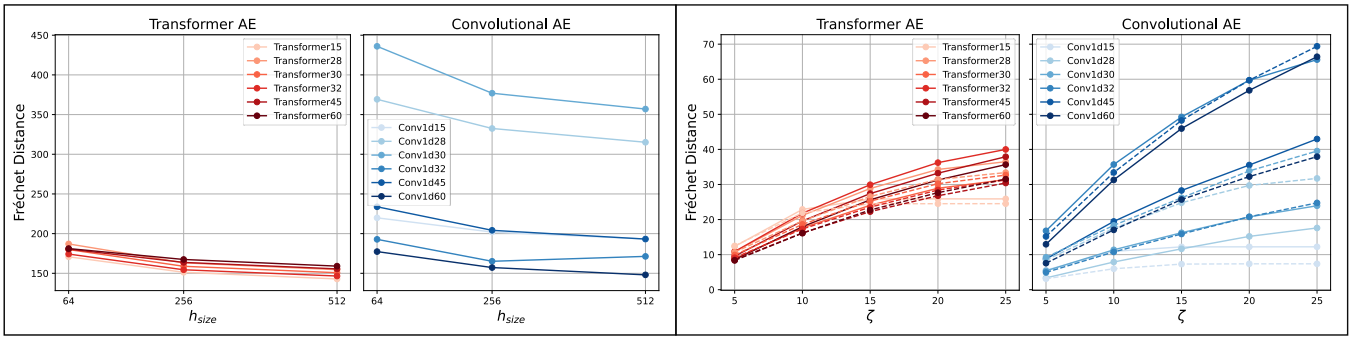

Figure 6: FMD scores on foot skating (left) and over-smoothing (right) motion artifacts. The dashed lines represent the score for full-body gestures. The score computed by both methods effectively penalizes the motion artifact in every tested configuration, except for the Conv1D-AE with a motion length of 32 frames. The Transformer-AE feature extractor allows the score to be more stable with respect to motion samples length than the score from Conv1D-AE.

upper-body tier between the median user rating and the proposed objective metric (Transformer FMD) regardless of the motion length (p-*value* < 0.05). However, using Conv1D-AE as a feature extractor, the significant effect did not appear in every motion length tested. For the full-body tier, even if the correlation profile of the Transformer-based FMD seems more stable toward the variation of motion length than the Conv1D, the significant effect was not observed for a number of motion frames smaller than 45.

Moreover, both Transformer- and Conv1D-based models seem to fail to assess efficiently some of the submitted dataset *e.g.*, **USN** or **FSH**. We observed a large gap between the user ratings and the objective score. Concerning **USN**, we also observed that the mean reconstruction error is higher than the other upper-body systems: 0.63 ± 0.24 for **USN** and 0.075 ± 0.031 in all other systems. This problem of not properly evaluating out-of-distribution samples (but plausible) is due to the fact that the feature extractor is trained on a small motion dataset, and we believe it can be improved by training on a larger and general motion dataset.

6 DISCUSSION AND PERSPECTIVES

This work introduced a motion-generative model evaluation metric, the FMD score, which is designed to address two prevalent motion artifacts–foot skating and oversmoothing–and is robust to the length of motion samples. We believe that this effect is due to the strength of the Transformer architecture. It makes use of the attention mechanism that dynamically weights the importance of different parts of the input sequence. This adaptive attention allows the model to focus on the most relevant spatio-temporal features. 1D convolutions, on the other hand, apply fixed filters across the entire input, which may limit their ability to adapt to varying patterns. Furthermore, Transformers outperform CNN-based methods in the understanding of global context in computer vision [Hatamizadeh et al. 2023] or NLP [Lauriola and Moschitti 2021]. It makes Transformer suitable for capturing long-term dependencies in a sequence rather than 1D convolutions.

However, it is important to acknowledge the limitations associated with objective evaluation metrics based on the Fréchet distance, despite their widespread usage. First, the FD-based evaluation is designed to jointly estimate the quality and diversity of synthetic

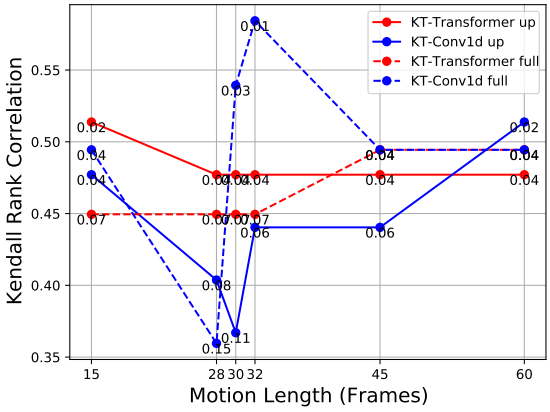

Figure 7: Kendall rank correlation between FMD and user median rates regarding the length of each motion sample. The p-value is indicated at each related point on the graph. The dashed lines refer to the full-body tier correlation.

modalities. Moreover, the Fréchet distance measures the overall similarity between two distributions, but does not explicitly capture semantic alignment or meaningful differences between generated samples and real data. It may not account for specific features or artifacts that should be relevant to capture in the evaluation of generative models.

In addition, we must take into account that Equation 1 which computes FD is built upon the hypothesis that the distributions are multivariate Gaussian. Hence, the Gaussian structure of the latent space is an essential feature in this case. In this work, we did not take into account the latent space distribution, but we observed the expected behavior for the metric on the tested artifacts and found a statistically significant rank correlation between it and the user rates. Nevertheless, the non-Gaussian nature of the tested distributions may induce an inaccurate FD score *e.g.*, that is

Table 1: FMD scores given conditions of the upper and full body tier. The FMD is computed between the motion in the validation set of the GENEA dataset and the motions in each condition. Higher user rate is better, and lower FMD is better. The conditions were presented in descending order of the median user rates.

Cond. in Upper-body Tier	Median User Rate ↑	FMD_{15} ↓		FMD_{28} ↓		FMD_{30} ↓		FMD_{32} ↓		FMD_{45} ↓		FMD_{60} ↓	
		Trans.	Conv.	Trans.	Conv.	Trans.	Conv.	Trans.	Conv.	Trans.	Conv.	Trans.	Conv.
USQ	69	35.15	17.23	31.9	14.57	27.34	18.89	33.47	56.17	32.29	32.06	30.01	56.93
UNA	63	**5.49**	**1.91**	**5.29**	**1.52**	**4.4**	**2.348**	**5.51**	**7.46**	**5.2**	**5.199**	**4.85**	**7.28**
USJ	53	52.64	29.17	50.42	25.58	41.69	23.1	51.75	84.06	49.31	43.3	44.5	81.44
USO	48	53.56	29.95	48.83	15.76	40.17	18.5	49.97	79.80	45.87	69.97	43.1	110.19
USN	44	353.01	354.54	320.86	274.18	279.38	236.29	337.79	787.68	318.5	589.1	329.72	1227.08
USK	41	65.04	29.5	60.75	17.41	51.29	20.04	63.86	87.59	59.51	59.33	56.63	110.41
USM	41	26.77	12.13	25.55	7.38	20.9	11.07	26.28	44.73	25.25	24.22	23.87	37.84
UBT	36	141.93	82.76	136.5	59.05	112.38	63.85	135.87	234.86	127.08	151.34	122.18	196.88
UBA	33	158.8	92.07	154.54	113.11	126.83	90.75	156.45	262.89	145.34	254.33	139.23	374.76
USP	29.5	88.64	70.71	87.88	75.7	74.52	69.61	90.95	194.63	82.58	193.94	83.35	275.86
USL	22	145.77	83.06	138.52	41.6	116.83	52.94	143.87	225.51	132.56	141.04	125.34	199.41
Cond. in Full-body Tier	Median User Rate ↑	FMD_{15} ↓		FMD_{28} ↓		FMD_{30} ↓		FMD_{32} ↓		FMD_{45} ↓		FMD_{60} ↓	
		Trans.	Conv.	Trans.	Conv.	Trans.	Conv.	Trans.	Conv.	Trans.	Conv.	Trans.	Conv.
FSA	71	37.41	13.13	36.51	26.98	33.28	29.08	33.28	25.06	30.28	48.41	32.03	40.26
FNA	70	**6.168**	**1.41**	**5.97**	**4.07**	**5.52**	**5.07**	**5.52**	**2.43**	**5.05**	**8.7**	**5.21**	**5.36**
FSC	53	56.22	13.41	52.96	44.87	49.51	36.07	56.22	20.24	43.04	105.76	43.58	55
FSI	46	41.16	18.07	38.07	43.25	35.59	41.21	38.07	26.92	32.33	83.64	32.77	43.6
FSF	38	29.72	12.92	28.75	23.3	26.14	28.24	28.75	30.27	24.33	50.08	25.04	46.62
FSG	38	52.81	16.87	52.6	40.51	47.3	56.59	52.6	32.19	44.04	106.01	44.8	61.73
FSH	36	27.48	6.14	27.4	16.71	24.86	20.06	27.4	13.45	22.44	34.51	23.27	24.68
FSD	34	110.17	38.99	103.18	133.66	100.19	171.5	103.18	141.11	81.17	289	81.84	244.92
FSB	30	81.73	36.68	78.36	135.5	74.24	169.78	81.73	146.06	68.84	249.51	63.21	242.11
FBT	27.5	144.27	40.78	142.57	122.28	131.4	118.06	131.4	80.86	119.71	271.97	117.78	129.97

sensitive to imperceptible perturbations [Luzi et al. 2023]. Further investigations are needed on this side.

Moreover, we believe that ensuring similarity between the original (high-dimensional) and latent (low-dimensional) spaces *e.g.*, while maintaining proximity between similar samples is beneficial for FMD. However, the training procedure aiming to reduce the mean squared error between the ground truth and the decoded motion does not provide this feature. One solution could be the use of Graph Neural AE regularized with a structure-preserving distance as proposed in [Ahmed et al. 2021].

7 CONCLUSION

In this paper, we tackled the challenge of objectively evaluating motion-generative models and, more specifically, the validation of such a metric. We analyzed the behavior of the FMD to two common artifacts in this context, the foot skating and over-smoothing artifacts. Additionally, we studies the relationship between FMD and human judges' rating of motion likeliness. We found that FMD is sensitive to any artifacts tested and proportional to the intensity of degradation. Moreover, FMD achieved a statistically significant correlation with the user study in every motion length tested with respect to upper body gesticulation, but not for full-body gestures. We also conducted experiments on FMD robustness to motion length variation and proposed a Transformer-based AutoEncoder leading to a more robust FMD than the metric proposed in [Yoon et al. 2020]. Even though these results seem encouraging, we believe that there is room for improvement: considering other motion artifacts, datasets, and analyzing its robustness to framerate variation will be a relevant addition to this work. We hope that this work will pave the way for the design of a consistent objective metric to evaluate the performance of motion-generative models.

ACKNOWLEDGMENTS

This work was partially supported by the Industrial Fundamental Technology Development Program (20023495) funded by MOTIE, Korea.

REFERENCES

Kfir Aberman, Peizhuo Li, Sorkine-Hornung Olga, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. 2020a. Skeleton-Aware Networks for Deep Motion Retargeting. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 62.

Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. 2020b. Unpaired Motion Style Transfer from Video to Animation. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 64.

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. 2020. Better fine-tuning by reducing representational collapse. *arXiv preprint arXiv:2008.03156* (2020).

Imtiaz Ahmed, Travis Galoppo, Xia Hu, and Yu Ding. 2021. Graph regularized autoencoder and its application in unsupervised anomaly detection. *IEEE transactions on pattern analysis and machine intelligence* 44, 8 (2021), 4110–4124.

Emre Aksan, Peng Cao, Manuel Kaufmann, and Otmar Hilliges. 2020. Attention, please: A spatio-temporal transformer for 3d human motion prediction. *arXiv preprint arXiv:2004.08692* 2, 3 (2020), 5.

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

Alan Baade, Puyuan Peng, and David Harwath. 2022. Mae-ast: Masked autoencoding audio spectrogram transformer. *arXiv preprint arXiv:2203.16691* (2022).

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* 33 (2020), 12449–12460.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254* (2021).

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150* (2020).

David Berthelot, Thomas Schumm, and Luke Metz. 2017. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717* (2017).

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

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

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*.

Wenheng Chen, He Wang, Yi Yuan, Tianjia Shao, and Kun Zhou. 2020. Dynamic future net: Diversified human motion generation. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2131–2139.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341* (2020).

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

DC Dowson and BV666017 Landau. 1982. The Fréchet distance between multivariate normal distributions. *Journal of multivariate analysis* 12, 3 (1982), 450–455.

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.

Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. 2023. Masked Diffusion Transformer is a Strong Image Synthesizer. arXiv:2303.14389 [cs.CV]

Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. 2022. Contrastive audio-visual masked autoencoder. *arXiv preprint arXiv:2210.07839* (2022).

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.

Jiaxian Guo, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang. 2017. Long Text Generation via Adversarial Training with Leaked Information. *arXiv preprint arXiv:1709.08624* (2017).

Ali Hatamizadeh, Hongxu Yin, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. 2023. Global context vision transformers. In *International Conference on Machine Learning*. PMLR, 12633–12646.

Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. 2020. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–14.

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).

Daniel Holden, Taku Komura, and Jun Saito. 2017. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–13.

Shuaiying Hou, Hongyu Tao, Hujun Bao, and Weiwei Xu. 2023. A Two-part Transformer Network for Controllable Motion Synthesis. *arXiv preprint arXiv:2304.12571* (2023).

Po-Yao Huang, Vasu Sharma, Hu Xu, Chaitanya Ryali, Haoqi Fan, Yanghao Li, Shang-Wen Li, Gargi Ghosh, Jitendra Malik, and Christoph Feichtenhofer. 2022a. MAViL: Masked Audio-Video Learners. *arXiv preprint arXiv:2212.08071* (2022).

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. 2022b. Masked autoencoders that listen. *Advances in Neural Information Processing Systems* 35 (2022), 28708–28720.

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).

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

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.

Maurice George Kendall. 1948. Rank correlation methods. (1948).

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.

Dongjun Kim, Yeongmin Kim, Wanmo Kang, and Il-Chul Moon. 2022. Refining Generative Process with Discriminator Guidance in Score-based Diffusion Models. *arXiv preprint arXiv:2211.17091* (2022).

Taras Kucherenko, Pieter Wolfert, Youngwoo Yoon, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2023. Evaluating gesture-generation in a large-scale open challenge: The GENEA Challenge 2022. *arXiv preprint arXiv:2303.08737* (2023).

Ivano Lauriola and Alessandro Moschitti. 2021. Answer sentence selection using local and global context in Transformer models. In *ECIR 2021*. https://www.amazon.science/publications/answer-sentence-selection-using-local-and-global-context-in-transformer-models

Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S Srinivasa, and Yaser Sheikh. 2019. Talking with hands 16.2 m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 763–772.

Sang-gil Lee, Heeseung Kim, Chaehun Shin, Xu Tan, Chang Liu, Qi Meng, Tao Qin, Wei Chen, Sungroh Yoon, and Tie-Yan Liu. 2021. Priorgrad: Improving conditional denoising diffusion models with data-driven adaptive prior. *arXiv preprint arXiv:2106.06406* (2021).

Yichong Leng, Zehua Chen, Junliang Guo, Haohe Liu, Jiawei Chen, Xu Tan, Danilo Mandic, Lei He, Xiangyang Li, Tao Qin, et al. 2022. Binauralgrad: A two-stage conditional diffusion probabilistic model for binaural audio synthesis. *Advances in Neural Information Processing Systems* 35 (2022), 23689–23700.

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).

Maosen Li, Siheng Chen, Zihui Liu, Zijing Zhang, Lingxi Xie, Qi Tian, and Ya Zhang. 2021a. Skeleton graph scattering networks for 3d skeleton-based human motion prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*. 854–864.

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

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. 2020. Character controllers using motion vaes. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 40–1.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

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.

Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. 2017. Pose guided person image generation. In *Advances in Neural Information Processing Systems*. 405–415.

Antoine Maiorca, Nathan Hubens, Sohaib Laraba, and Thierry Dutoit. 2022a. Towards Lightweight Neural Animation: Exploration of Neural Network Pruning in Mixture of Experts-based Animation Models. *arXiv preprint arXiv:2201.04042* (2022).

Antoine Maiorca, Youngwoo Yoon, and Thierry Dutoit. 2022b. 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

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.

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

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. https://doi.org/10.3115/1073083.1073135

Mathis Petrovich, Michael J. Black, and Gül Varol. 2021. Action-Conditioned 3D Human Motion Synthesis with Transformer VAE. In *International Conference on Computer Vision (ICCV)*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems* 29 (2016).

Nisarg A Shah and Gaurav Bharaj. 2022. Towards Device Efficient Conditional Image Generation. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press. https://bmvc2022.mpi-inf.mpg.de/0689.pdf

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053* (2019).

Leslie N Smith and Nicholay Topin. 2019. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, Vol. 11006. SPIE, 369–386.

Sebastian Starke, Ian Mason, and Taku Komura. 2022. Deepphase: Periodic autoencoders for learning motion phase manifolds. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–13.

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

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).

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.

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).

Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. 2022a. Towards diverse and natural scene-aware 3d human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20460–20469.

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.

Weiqiang Wang, Xuefei Zhe, Huan Chen, Di Kang, Tingguang Li, Ruizhi Chen, and Linchao Bao. 2022b. NEURAL MARIONETTE: A Transformer-based Multi-action Human Motion Synthesis System. *arXiv preprint arXiv:2209.13204* (2022).

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

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.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. MEGATRON-CNTRL: Controllable Story Generation with External Knowledge Using Large-Scale Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 2831–2845. https://doi.org/10.18653/v1/2020.emnlp-main.226

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

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

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).

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.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917* (2022).

He Zhang, Sebastian Starke, Taku Komura, and Jun Saito. 2018. Mode-adaptive neural networks for quadruped motion control. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–11.

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).

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.