Single Motion Diffusion

Anonymous Author(s)*

Synthesizing realistic animations of humans, animals, and even imaginary creatures has long been a goal for artists and computer graphics professionals. Compared to the imaging domain, which is rich with large available datasets, the number of data instances for the motion domain is limited, particularly for the animation of animals and exotic creatures (e.g., dragons), which have unique skeletons and motion patterns. In this work, we present a Single Motion Diffusion Model, dubbed SinMDM, a model designed to learn the internal motifs of a single motion sequence with arbitrary topology and synthesize motions of arbitrary length that are faithful to them. We harness the power of diffusion models and present a denoising network designed specifically for the task of learning from a single input motion. Our transformer-based architecture avoids overfitting by using local attention layers that narrow the receptive field, and encourages motion diversity by using relative positional embedding. SinMDM can be applied in a variety of contexts, including spatial and temporal in-betweening, motion expansion, style transfer, and crowd animation. Our results show that SinMDM outperforms existing methods both in quality and time-space efficiency. Moreover, while current approaches require additional training for different applications, our work facilitates these applications at inference time. Our code and trained models are available at https://sinmdm.github.io/SinMDM-page.

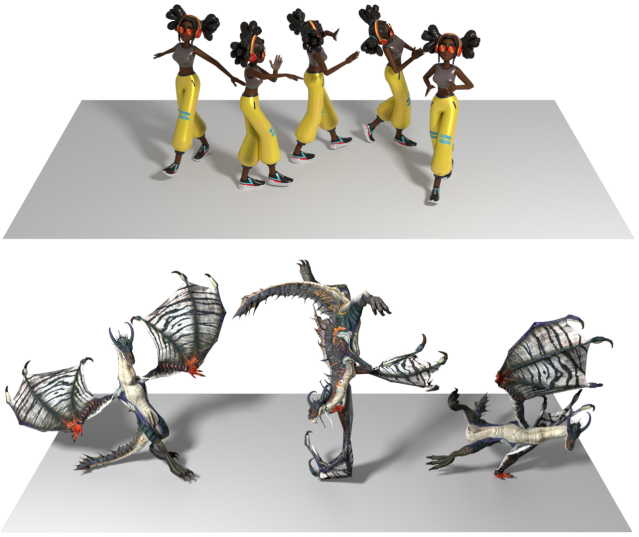


Fig. 1. SinMDM learns the internal motion motifs from a single motion sequence with arbitrary topology and synthesizes motions that are faithful to the core learned motifs of the input sequence. Top: a girl exercising while walking. Bottom: a breakdancing dragon. Left to right: breakdance uprock, breakdance freeze, and breakdance flair.

1 INTRODUCTION

Motion modeling of 3D characters has been a long-pursued task. Motion is challenging as it is a four-dimensional entity, with numerous degrees of freedom, and at the same time, humans are highly susceptible to the natural behavior and correctness of motion. Traditionally, motion sequences have been manually modeled by experts, which is a time-consuming and costly process. In recent years, neural models have offered faster and less expensive tools for modeling motion [Holden et al. 2016; Petrovich et al. 2022; Raab et al. 2022]. In particular, the very recent adaptation of diffusion models into the motion domain provides unprecedented results in both quality and diversity [Kim et al. 2022; Tevet et al. 2022b].

These data-driven methods typically require large amounts of data for training. However, motion data is quite scarce and, moreover, for a non-human skeleton, it is barely existent. The few available datasets contain humanoids only, whose topology and bone ratio are fixed. Animators often customize a skeleton per character (human, animal, or magical creature), for which common data-driven techniques are irrelevant.

In this work, we present a Single Motion Diffusion Model, dubbed SinMDM, that trains on a *single motion* input sequence. Our model enables modeling motions of arbitrary skeletal topology, which often have no more than one animation sequence to learn from. SinMDM can synthesize a variety of variable-length motion sequences that retain the core motion elements of the input and can handle complex and non-trivial skeletons. For example, our model can generate a diverse clan based on one flying dragon or one hopping ostrich.

Learning from a single instance has been explored for the imaging domain [Shaham et al. 2019; Shocher et al. 2019] and for the motion domain [Li et al. 2022], using the GAN architecture [Goodfellow et al. 2014]. Indeed until recently, GANs have been the dominant approach for generative models. We find diffusion models [Ho et al. 2020] more suitable for single input learning, as the descriptive ability attained by gradual denoising yields a light-weight model that, compared to prior art, is simpler in architecture and more efficient in terms of the number of parameters and training time. Furthermore, we demonstrate that diffusion models can be used with limited data, contrary to their reputation for requiring large amounts.

Most motion diffusion models utilize transformers. However, common vanilla transformers are not suitable for learning a single sequence, as their receptive field encompasses the entire motion. To learn local motion sequences, the receptive field must be small enough, analogously to the use of patch-based discriminators [Isola et al. 2017; Li and Wand 2016] in GAN-based techniques. The use of a narrow receptive field promotes diversity and reduces overfitting.

In our work, we design and examine two core networks with a temporally narrow receptive field (Fig. 2a) to denoise the motion signal. The first, p_{qna}, is a QnA-based transformer [Arar et al. 2022] designed with a small receptive field, and the second, p_{unet}, is a degenerated UNet consisting of a sequence of single-level convolutional layers with skip connections. We analyze and compare their performance with Ganimator [2022], a GAN-based single motion method, and show that the configuration of p_{qna} performs generally better than the other two alternatives: First, it is much faster and consumes less memory. Additionally, it facilitates various applications at inference time and does not require special training for

specific tasks. Lastly, and perhaps most importantly, it demonstrates higher-quality results.

We present many use cases of SinMDM, all of which are applied at inference time. One application we present is *motion composition*, where a given motion sequence is composed jointly with a synthesized one, either temporally or spatially. Special cases of motion composition include in-betweening and motion expansion. Another application we present is *harmonization*, along with its special case, style transfer. Here, a reference motion is changed such that it matches the learned motion motifs. Additional applications that we present are *long sequence generation* and *crowd animation*.

Finally, we suggest two comprehensive benchmarks for single-motion evaluation. The first is built upon the artistically crafted MIXAMO [2021] dataset, utilizing metrics that do not require an additional feature-extracting model. The second is based on the HumanML3D [2022] dataset and enables metrics that use latent features, such as single-FID. In Sec. 5 we show that our model outperforms current works on both benchmarks.

2 RELATED WORK

2.1 Single-Instance Learning

The goal of single-instance generation is to learn an unconditional generative model from a single instance and generate diverse samples with similar content by capturing the internal statistics of patches. The type of instance depends on the input domain. Most single-instance learning research has been focused on the domain of imaging. The first works on this topic are SinGAN [Shaham et al. 2019] and InGAN [Shocher et al. 2019]. SinGAN uses a patch-based discriminator [Isola et al. 2017; Li and Wand 2016] and an image pyramid to generate diverse results hierarchically. InGAN [Shocher et al. 2019], uses a conditional GAN to solve the same problem using geometry transformation. More recent approaches include ExSinGAN [Zhang et al. 2021b], which trains multiple modular GANs to model the distribution of structure, semantics, and texture, and ConSinGAN [Hinz et al. 2021], trains several stages sequentially and improves SinGAN. Many works in the imaging domain follow and improve the aforementioned pioneering works [Asano et al. 2020; Chen et al. 2021; Granot et al. 2022; Lin et al. 2020; Sun and Liu 2020; Sushko et al. 2021; Yoo and Chen 2021; Zhang et al. 2022b; Zheng et al. 2021].

Several works have been introduced in other domains, such as for the shapes domain [Son et al. 2022] and for the 3D scenes domain [Son et al. 2022]. In the motion domain, the only work that learns a single motion is Ganimator [Li et al. 2022]. Ganimator follows SinGAN, hence it uses a GAN architecture, with a patch-based discriminator and a temporal pyramid.

The vast majority of single-instance learning works use a GAN architecture [Goodfellow et al. 2014]. Until recently, GANs have been the dominant approach for generative models. However, we are currently seeing a trend towards using diffusion models [Ho et al. 2020; Song et al. 2020a] as an alternative to GANs.

A number of concurrent works in the imaging domain use diffusion models to learn from single images. Like our p_{unet}, Wang et al. [2022] and Nikankin et al. [2022] drop the image pyramid structure and use a UNet with a limited depth. A different work [Kulikov et al. 2022] constructs a multi-scale diffusion process from down-sampled versions of the training image, as well as their blurry versions.

Ganimator [Li et al. 2022] is our immediate comparison reference, as it is the only single-motion learning work. Section 5 and our supplementary video show that SinMDM outperforms it quantitatively and qualitatively. In addition, Ganimator uses a complex architecture that combines a temporal hierarchy of motions with a skeletal hierarchy of joints. Our model uses neither hierarchies, which makes it simple to implement while achieving better results.

2.2 Diffusion Models

Diffusion models use a stochastic diffusion process, as modeled in thermodynamics [Sohl-Dickstein et al. 2015; Song and Ermon 2020], to generate samples from a data distribution. These models are adapted for image generative applications. Dhariwal and Nichol [2021] introduce the concept of classifier-guided diffusion for conditional generation, which is later adapted in the GLIDE [Nichol et al. 2022] model. Ho and Salimans [2022] propose the Classifier-Free Guidance approach, which can trade-off between fidelity and diversity in the generated samples. This approach has been demonstrated to achieve better results compared to other methods, as shown by Nichol et al. [2022].

Local editing of images may be viewed as an inpainting problem, in which a portion of the image is held constant while the model denoises the remaining part [Saharia et al. 2022; Song et al. 2020b]. In our work, we adapt this technique for motion composition of specific body parts or temporal intervals.

In the motion domain, several very recent works [Kim et al. 2022; Tevet et al. 2022b; Zhang et al. 2022a] introduce diffusion-based synthesis, where the most prominent one is MDM [Tevet et al. 2022b]. MDM utilizes a lightweight network that can be trained on a single mid-range GPU and uses a transformer rather than the common UNet, and predicts motion rather than noise. Like MDM, SinMDM presents a lightweight architecture and predicts motion rather than noise. Unlike MDM, our work uses a QnA-based transformer, as the receptive field of a vanilla transformer is the full motion, inducing over-fitting.

2.3 Motion Synthesis Models

In recent years, we witness prosperity in the domain of motion synthesis using neural networks [Holden et al. 2016, 2015]. Most of these models focus on specific motion-related tasks, conditioned on some limiting factors, which can be high-level guidance such as action [Cervantes et al. 2022; Guo et al. 2020; Petrovich et al. 2021; Tevet et al. 2022b] or text [Ahuja and Morency 2019; Bhattacharya et al. 2021; Guo et al. 2022; Petrovich et al. 2022; Tevet et al. 2022a,b; Zhang et al. 2021c], can be parts of a motion such as motion prefix [Aksan et al. 2019; Barsoum et al. 2018; Habibie et al. 2017; Hernandez et al. 2019; Yuan and Kitani 2020; Zhang et al. 2021a] or in-betweening [Duan et al. 2021; Harvey and Pal 2018; Harvey et al. 2020; Kaufmann et al. 2020], motion retargeting or style transfer [Aberman et al. 2020a,b, 2019; Holden et al. 2017; Villegas et al. 2018], and even music [Aristidou et al. 2022; Lee et al. 2018; Li et al. 2021; Sun et al. 2020]. Fewer models are fully unconditioned [Holden

et al. 2016; Raab et al. 2022; Starke et al. 2022] and they learn the motion manifold from the input data in an unsupervised manner.

The architecture of motion synthesis models can be roughly divided into autoregressive [Fragkiadaki et al. 2015; Ghorbani et al. 2020; Guo et al. 2020; Habibie et al. 2017; Jang and Lee 2020; Maheshwari et al. 2022; Petrovich et al. 2021; Zhou et al. 2018], GAN-based [Degardin et al. 2022; Wang et al. 2020; Yan et al. 2019; Yu et al. 2020], and more recently, diffusion-based [Kim et al. 2022; Tevet et al. 2022b; Zhang et al. 2022a]. Our work belongs to the latter category.

3 PRELIMINARY

In this work, we present SinMDM, a novel framework that learns the internal motion motifs of a *single motion* of arbitrary topology, and generates a variety of synthesized motions that retain the core motion elements of the input sequence.

At the crux of our approach lays a denoising diffusion probabilistic model (DDPM) [Ho et al. 2020]. We consider diffusion models to be more appropriate for single input learning compared to previous methods, and suggest a light-weight model, efficient in time and space and simple in architecture. This is achieved through the gradual denoising process, which enhances the model's descriptive capability. Our generative network is a transformer whose attention layers are replaced by the recently introduced QnA layers [Arar et al. 2022]. In addition, we examine an alternate architecture based on the commonly used UNet, adjusted and degenerated such that it can train on a single input.

In this section, we briefly recap DDPM. Then, in the following section, we describe our method and focus on the unique architecture that serves our goals. Next, we detail the experiments conducted to validate our approach (Sec. 5), describe various applications enabled by SinMDM (Sec. 6), and summarise with conclusions (Sec. 7). The readers are encouraged to watch the supplementary video in order to get a full impression of our results.

3.1 Denoising Diffusion probabilistic Models (DDPM)

DDPMs [2020] have become the de-facto leading generative networks technique. They are mainly dominant in the imaging domain [Dhariwal and Nichol 2021], and just recently pioneering works have successfully applied this approach in the motion domain [Tevet et al. 2022b; Zhang et al. 2022a]. Denoising networks learn to convert unstructured noise to samples from a given distribution, by performing an iterative process of removing small amounts of Gaussian noise at each step.

Given an input motion sequence x_0, we apply a Markov noising process of T steps, $\{x_t\}_{t=0}^{T}$, such that

$$q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I), \tag{1}$$

where $\alpha_t \in (0,1)$ are constant hyper-parameters. When α_t is small enough, we can approximate $x_T \sim \mathcal{N}(0, I)$.

We apply unconditional motion synthesis that models x_0 as the reversed diffusion process of gradually cleaning x_T, using a generative network p_θ. Following Tevet et al. [2022b] we choose to predict the input motion, denoted $\hat{x}_0$ [Ramesh et al. 2022] rather than predicting ϵ_t, hence

$$\hat{x}_0 = p_\theta(x_t, t). \tag{2}$$

We apply the widespread diffusion loss, via

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t\sim[1,T]}\|x_0 - p_\theta(x_t, t)\|_2^2. \tag{3}$$

Synthesis at inference time is applied through a series of iterations, starting with pure noise x_T. In each iteration, a clean version of the current sample x_t is predicted using a generator p_θ. This predicted clean sample $\hat{x}_0$ is then "re-noised" to create the next sample x_{t-1}, with the process being repeated until $t = 0$ is reached.

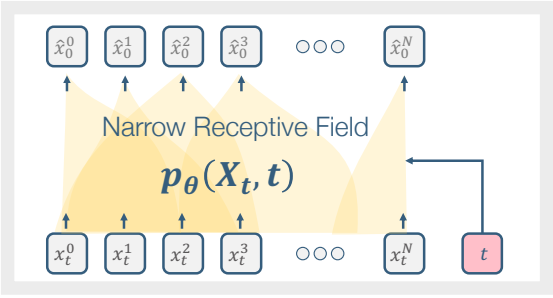
4 GENERATIVE NETWORK

Our goal is to construct a model that can generate a variety of synthesized motions that retain the core motion motifs of a single learned input sequence. More formally, we would like to construct the generative network p_θ in Eq. 2, that synthesizes a motion $\hat{x}_0$ from a noised motion x_t. We study two alternative architectures of p_θ, a QnA-based transformer dubbed p_{qna}, and a degenerated UNet dubbed p_{unet}. The former is our preferred architecture, which outperforms the latter. The alternative architecture is studied mainly for ablation, and despite its simplicity, it outperforms current state-of-the-art in many aspects. In the rest of this work, unless otherwise specified, descriptions apply to both of these architectures. Our architecture is depicted in Fig. 2.

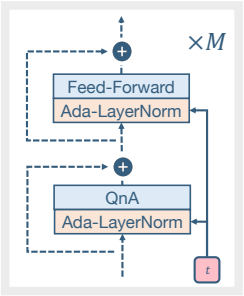
While traditional single-instance techniques use a pyramid of down-sampled instances (images or motions) and learn in a coarse-to-fine fashion, our model introduces a simplified architecture, where the size of the motion (number of joints and number of frames) is fixed throughout the entire training process.

p_{qna} – QnA-based Transformer. We are interested in using a transformer architecture, as it outperforms the common UNet in motion diffusion models [Tevet et al. 2022b]. However, we have found that standard transformers are not as effective for learning from a single sequence because they employ global attention layers, resulting in a receptive field that encompasses the entire sequence, which can cause overfitting. A possible solution would be using local attention in non-overlapping windows, like in ViT [Dosovitskiy et al. 2021]. Nonetheless, non-interleaving windows tend to limits the cross-window interaction, hurting the model's performance. Our solution is to use QnA [Arar et al. 2022], a state-of-the-art shift-invariant local attention layer, that aggregates the input locally in an overlapping manner, much like convolutions, but with the expressive power of attention. The key idea behind QnA is to introduce learned queries, shared by all windows, allowing fast and efficient implementation. In particular, QnA enables local attention with a temporally narrow receptive field. Our QnA-based transformer is the first to be used in the motion domain, and is depicted in Fig. 2b, where global attention layers of a vanilla transformer are replaced by QnA layers.

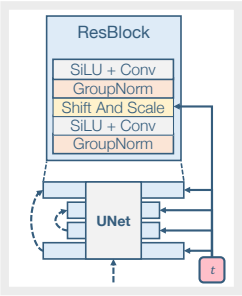
When learning a single motion, using a narrow receptive field can help prevent overfitting, but induces small networks with limited expressiveness. We address this challenge with stochastic depth regularization [Huang et al. 2016], a training procedure where for each sample, a subset of layers is randomly dropped and bypassed by the



(a) Our denoiser predicts the input sequence from a noisy one.



(b) p_{qna} - Transformer-based.



(c) p_{unet} - UNet-based.

Fig. 2. To allow training on a single motion, our denoising network is designed such that its overall receptive field covers only a portion of the input sequence. This effectively allows the network to simultaneously learn from multiple local temporal motion segments, as shown in (a). We consider two alternative architectures to learn from single motion: (b) p_{qna}, a transformer-based approach, and (c) p_{unet}, a convolutional-based approach, utilizing a simplified version of the UNet architecture.

identity function. Effectively this regularization induces stochasticity in the receptive field size.

An important aspect of the technical design involves the use of positional embedding (PE). The traditional use of global PEs can restrict the variety of the output by enforcing the preservation of sub-motion order. To promote more diversity and permit rearrangement of motion components, we employ relative PEs instead [Press et al. 2021; Su et al. 2021]. This way, sub-motions attain a natural internal order while also allowing greater flexibility in their sequencing.

p_{unet} – Degenerated UNet. We further examine the UNet [Nichol et al. 2022] architecture as an alternative to QnA-based transformers, as it is frequently used by diffusion models in the imaging domain. When training it on a single input, a strong overfit occurs, induced by the large receptive field. We mitigate this issue by (a) decreasing the depth of the UNet, and (b) removing the attention layers within it, as global attention is aware of the full motion sequence.

In Sec. 5, we confirm this architectural change to be correct, by showing that the quantitative scores get better when reducing the depth of the UNet. In our experiments, we examine different levels of depth and conclude that the UNet with only one level of depth performs the best. Networks deeper than that tend to synthesize motions too similar to the input one. As a result, the preferred network is merely a sequence of convolutions with residuals and time embedding, hence we call it a degenerated UNet. See Fig. 2c. Despite its simple architecture, motions synthesized by the degenerated UNet are of higher quality compared to the current work Ganimator [2022] (see our video), and of better quantitative results.

Although UNet's attained metric scores are slightly better than the ones obtained by a QnA-based transformer, it requires significantly more training iterations to reach those scores, and more importantly, the qualitative results of QnA-based architecture look better, as shown in Fig. 3 and in our video.

Another issue with using a UNet is that margin padding (*i.e.*, temporal beginning and end) of the motion causes the beginning and the end of the synthesized motion to be too similar to the input motion as well as to each other, aligning with the findings of Xu et al. [2021]. Using non-zero padding techniques (*e.g.*, replications, reflections) does not solve the problem as the network can identify replicated/reflected features as well. The QnA-based transformer is free of this issue, thus it synthesizes diverse beginnings and ends.

4.1 Motion representation

A motion sequence is represented by its dynamic and static features, **D** and **S**, respectively. The former differ at each temporal frame (*e.g.*, joint rotation angles), while the latter are temporally fixed (*e.g.* bone lengths, skeletal topology). **D** and **S** can be combined into global 3D pose sequences using *forward kinematics* (FK).

In our research, we focus on synthesizing the *dynamic features*, leaving the static features intact. That is, we predict dynamics (rotation angles over time) for a fixed skeleton topology with fixed bone lengths. For simplicity, we use the term *motion synthesis* for the generation of dynamic features only.

Let N denote the number of frames in a motion sequence, and F denote the length of the features of all joints together in a single frame. Note that while T denotes the number of diffusion steps, N denotes the number of temporal frames in a motion. We represent the dynamic features of a motion by a tensor $\mathbf{D} \in \mathbb{R}^{N \times F}$. Naturally, the convolution for this representation is 1D, convolving on the temporal dimension (of size N) and holding F features.

Let J denote the number of skeletal joints, and let Q denote the number of rotational features, where rotational features may be Euler angles, quaternions, 6D rotations, etc.

When using the HumanML3D [Guo et al. 2022] dataset, we adhere to its representation, which is a concatenation of root joint velocity and height, and joints velocity, location and rotation, and foot contact labels. When using data from other datasets, we adhere to the representation used by Ganimator [2022], so we can conduct a fair comparison with their results. Their representation consists of a 3D root location, a rotation angle for each joint, and foot contact labels. Altogether, for a general representation $D \in \mathbb{R}^{N \times F}$ we got in this case $F = 3 + JQ + C$, where C is the number of joints that are prone to contact the ground, as elaborated below.

The rotations in both representations are defined in the coordinate frame of their parent in the kinematic chain, and are represented by the 6D rotation features ($Q = 6$) Zhou et al. [2019], which yields the best result in many works [Petrovich et al. 2021; Qin et al. 2022].

A growing number of works use foot contact labels [Gordon et al. 2022; Raab et al. 2022] to mitigate common foot sliding artifacts. Let **C** denote the set of joints that contact the ground in the subject

whose motion is being learned, and let $C = |\mathbf{C}|$. Foot contact labels are usually represented by $\mathbf{L} \in \{0,1\}^{N \times C}$. Clearly, a human, a spider, and a snake possess different C values.

When a dataset provides foot contact label information [Guo et al. 2022], we use it as is. When a dataset does not provide them, we calculate it as done by Li et al. [2022], via

$$\forall j \in \mathbf{C}, n \in [1, N]: \qquad \mathbf{L}^{nj} = \mathbf{1}[\|\Delta_n \mathrm{FK}([\mathbf{D}, \mathbf{S}])^{nj}\|_2 < \epsilon], \quad (4)$$

where $\Delta_n \mathrm{FK}([\mathbf{D}, \mathbf{S}])^{nj}$ denotes the velocity of joint j in frame n retrieved by a forward kinematics operator, and $\mathbf{1}[\cdot]$ is an indicator function.

5 EXPERIMENTS

Our experiments are held on motion capture data from the Mixamo [2021], HumanML3D [2022], and Truebones Zoo [2022] datasets, and on an artist-created animation, using an NVIDIA GeForce RTX 2080 Ti GPU.

5.1 Benchmarks.

We test our framework on two benchmarks. One consists of data from the HumanML3D dataset, and the other from the Mixamo dataset. These two datasets are different in many aspects. The data in HumanML3D fits the SMPL [Loper et al. 2015] topology, and its users normally use SMPL's mean body definition. In contrast, Mixamo provides 70 characters, each possessing their unique bone lengths and some possessing unique topologies. In addition, the motions in the Mixamo dataset are more diverse and more dynamic.

5.2 Metrics

For each benchmark, we use a different set of metrics. For the Mixamo benchmark, we use the metrics introduced in Ganimator [Li et al. 2022]. Our motivation for using these metrics is that Ganimator is the only current work that synthesizes motion from a single input, hence we are interested in comparing with it. However, these metrics are based on the values of motion features (*e.g.*, rotation values) while the usage of deep features is the current best practice [Zhang et al. 2018]. HumanML3D offers capabilities for calculating deep features, so we use it to evaluate metrics that depend on them.

Note that attaining a good score on some metrics, but a bad score on others, is not enough: High inter diversity with low coverage (or high SiFID [Shaham et al. 2019]) indicates that the synthesized motions diverge from the input one, and high coverage (low SiFID) with low inter diversity indicates overfit. The ideal outcome would be to achieve a combination of good values for all metrics.

Metrics on the Mixamo benchmark. We use the Mixamo dataset to compare SinMDM and Ganimator. We use the metrics suggested by them, and add several of our own. This group of metrics is applied on the motion itself, and not on deep features.

The metrics in Ganimator consist of (a) *coverage*, which is the rate of temporal windows in the input motion x_0 that are reproduced by the synthesized one, (b) *global diversity*, measuring the distance between $tess(\hat{x}_0)$ and x_0, where $tess(\cdot)$ is a tessellation that minimizes the L2 distance to the input sequence, and (c) *local diversity*, which is the average distance between windows in the synthesized motion $\hat{x}_0$ and their nearest neighbors in the input one.

Table 1. Results on the Mixamo benchmark, comparing current work Ganimator with p_{qna} and p_{unet}. Our models lead in all metrics but one. The best results are in **bold**. The intra-diversity GT value is in the header.

	Coverage ↑	Global Div. ↑	Local Div ↑	Inter Div. ↑	Intra Div. → 0.77	#Param. (M) ↓	#Iter. (K) ↓	Iter. Time (s) ↓	Tot. Time (h) ↓
Ganimator	94.3	1.24	**1.17**	0.09	0.64	21.7	60 (15×4)	0.36	6.00
Ours (p_{qna}) Deep + Sto. Depth	94.2	0.96	0.66	**0.14**	**0.77**	5.14	**20**	0.110	**0.61**
Ours (p_{unet}) 1 Level	**98.4**	**1.38**	0.99	0.11	0.67	**4.67**	100	**0.063**	1.75

The aforementioned metrics are measured relative to the input motion sequence. We add two metrics that are not related to the input motion, (d) *inter diversity*, the diversity between synthesized motions, and (e) *intra diversity*, which is the diversity between sub-windows internal to a motion.

In addition, we measure time-space efficiency values: (f) the number of network parameters, (g) the number of required iterations, (h) the time required for each iteration, and (d) the total running time, which is a multiplication of the last two.

For metrics (a)-(d), a higher score is better. In metric (e), inter diversity, the goal is to be as close as possible to the internal diversity of the input motion. For metrics (f)-(h), a lower score is better.

Metrics on the HumanML3D benchmark. We use this benchmark to measure metrics that are applied on deep features, obtained with a motion encoder by Guo et al. [2022]. The computed metrics are (a) *SiFID* [Shaham et al. 2019], which measures the distance between the distribution of sub-windows in the learned motion and a synthesized one, (b) *inter diversity*, which is the LPIPS distance [Zhang et al. 2018] between various motions synthesized out of one input, and (c) *intra diversity*, which is the LPIPS distance between sub-windows in one synthesized motion.

For metric (a) lower is better, for metric (b) higher is better, and for the latter one, the goal is to be as close as possible to the intra diversity of the learned motion.

5.3 Quantitative Results

In table 1 we compare our QnA-based transformer model p_{qna}, as well as our UNet-based architecture p_{unet}, with Ganimator. The table shows that SinMDM outperforms Ganimator in all metrics except one. Note that although p_{qna} and p_{unet} are quantitatively comparable, p_{qna} is qualitatively better, as shown in our supplementary video.

All the metrics are computed separately on each benchmark motion and then averaged. The metrics that measure time were computed on benchmark motion number 9 only.

The authors of Ganimator published a quantitative comparison on one motion, namely the Gangnam-style dancing sequence. We align with their study and measure our results on this motion as well, as shown in Tab. 2. In this table we use Ganimator's metrics to

Table 2. Results on the Gangnam-style motion, comparing existing techniques with ours. We mark the table leaders for the bottom half only, as MotionTexture and acRNN often achieve good scores on some metrics and poor scores on others, indicating either overfit or divergence. On the lower half, where all models attain high scores in all metrics, our models lead the table. The best results are in **bold**, and second best are underlined. The results for other works were originally reported by Ganimator.

	Coverage ↑	Global Diversity ↑	Local Diversity ↑
MotionTexture [2002]	84.6	1.03	1.04
MotionTexture (Single)	100	0.21	0.33
acRNN [2018]	11.6	5.63	6.69
Ganimator [2022]	97.2	1.29	1.19
Ours (p_{qna})	97.7	1.38	0.94
Ours (p_{unet})	**98.0**	**1.55**	**1.20**

Table 3. Results on the HumanML3D benchmark, comparing current work MDM with our best p_{qna} and p_{unet} versions. MDM attains high SiFID and intra diversity scores, but a low inter diversity score, which indicate of overfit. Our models attain high (but not highest) scores in all metrics, demonstrating that a balance of high scores across all metrics is more crucial than excelling in only a select few.

	SiFID ↓	Inter Div. ↑	Intra Div. → 4.52
MDM	**0.01**	0.0	**4.50**
Ours (p_{qna}) Deep + Sto. Depth	6.68	**1.84**	3.80
Ours (p_{unet}) 1 level	1.88	0.68	4.35

compare with two other, non-single motion works, MotionTexture [Li et al. 2002] and acRNN [Zhou et al. 2018]. Note that also for this specific motion, our results lead the table.

To evaluate our model's performance against another motion diffusion model, we compare it with MDM [Tevet et al. 2022b]. The comparison is conducted on the HumanML3D dataset with metrics based on deep features. The results are shown in Tab. 3. As mentioned above, attaining a high score in one metric only, indicates either overfit or divergence from the input motion. Indeed, MDM yields complete overfit, thus its SiFID and intra diversity scores are perfect (indicating similarity to the input motion), but its inter diversity scores are low. The overfit of MDM is caused by the global attention it uses.

5.4 Qualitative results

Our supplementary video reflects the quality of our results in the best way. It presents multiple synthesized motions, as well a comparison between Ganimator's results to ours, and between p_{qna} and p_{unet}. In addition, Fig. 3 depicts the differences between our two architectures and Ganimator. The figure shows that p_{qna} is faithful to the motifs of the input motion, while p_{unet} tends to be less dynamic, and Ganimator becomes motionless at a certain point.

5.5 Ablation

We examine several architectural variations for each of our studied architectures, p_{qna} and p_{unet}. Each variation is tested on both benchmarks (Mixamo and HumanML3D). For p_{qna}, our proposed

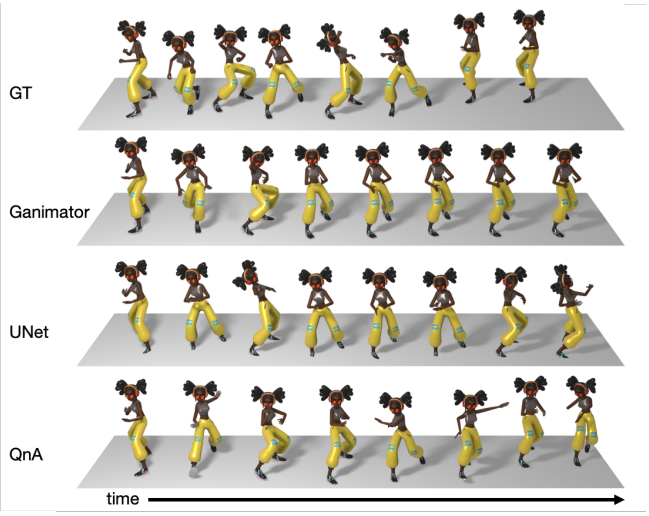


Fig. 3. Qualitative comparison, based on the motion *punch to elbow* from the Mixamo dataset.

Table 4. Ablation results on both benchmarks, comparing architectural variations. The best results are in blue and green for the p_{qna} and p_{unet} networks, respectively. The intra-diversity GT value is in the header.

(a) Mixamo benchmark

	Coverage ↑	Global Div. ↑	Local Div ↑	Inter Div. ↑	Intra Div. → 0.77	#Param. (M) ↓	#Iter. (K) ↓	Iter. Time (s) ↓	Tot. Time (h) ↓
p_{qna}									
Shallow	93.0	0.95	0.65	0.13	0.74	2.83	20	0.065	0.36
Deep	92.3	0.71	0.52	0.13	0.76	5.14	20	0.110	0.61
+ Sto. Depth	94.2	0.96	0.66	0.14	0.77	5.14	20	0.110	0.61
p_{unet}									
1 level	98.4	1.38	0.99	0.11	0.67	4.67	100	0.063	1.75
2 levels	99.7	1.05	0.75	0.11	0.67	14.64	100	0.159	4.42
3 levels	100.0	1.01	0.70	0.09	0.70	36.36	100	0.280	7.78

(b) HumanML3D benchmark

	SiFID ↓	Inter Div. ↑	Intra Div. → 4.52
p_{qna}			
Shallow	7.74	1.68	4.12
Deep + Sto. Depth	6.68	1.84	3.80
p_{unet}			
1 level	1.88	0.68	4.35
2 levels	0.77	0.29	4.50
3 levels	0.41	0.06	4.53

QnA-based transformer, we study a shallow architecture (4 layers), a deep architecture (8 layers), and a deep architecture with stochastic depth. For p_{unet}, the degenerated UNet, we study three down-sampling variations, of 1, 2, and 3 stages. In Tab. 4 we show ablation results for the benchmarks on both datasets. We observe that the best p_{qna} model is the deep one with stochastic depth for

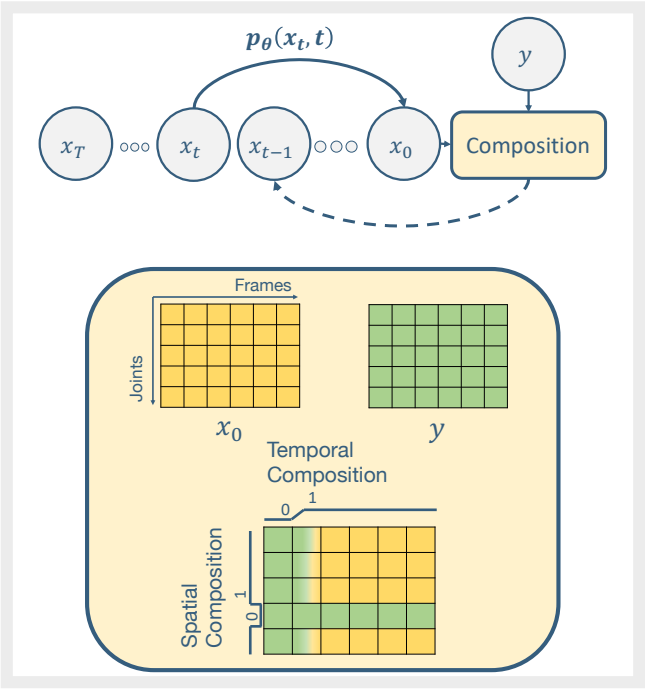


Fig. 4. Motion composition. Parts from a reference motion y, are composed with the synthesized motion $\hat{x}_0$, according to a composition map.

both benchmarks, and the best p_{unet} model is the shallowest for the Mixamo benchmark. On HumanML3D, some of the best results are attained by the 3 levels model, but its low inter diversity indicates that it overfits, hence the shallowest model wins again.

6 APPLICATIONS

Single-motion learning using diffusion models enables various applications. In this section, we present several applications, all of which are applied at inference time, without the need to re-train the network. Application at inference time is enabled by both p_{qna} and p_{unet}. Note that the only current single motion synthesis work, Ganimator [2022], requires a unique training procedure for most of the applications that it facilitates, and uses a dedicated technique for each application. Our diffusion model, however, facilitates numerous applications with just a few techniques.

In the following, we choose to show *motion composition*, where a given motion sequence is composed jointly with a synthesized one, either temporally or spatially. Special cases of motion composition include: in-betweening, motion expansion, and trajectory control. Another application we present, along with its special case, style transfer, is *harmonization*, where a reference motion is changed such that it matches the learned motion motifs. Additional applications we present are *one shot long motion generation* and *crowd animation*. The applications presented here are also demonstrated in our supplementary video.

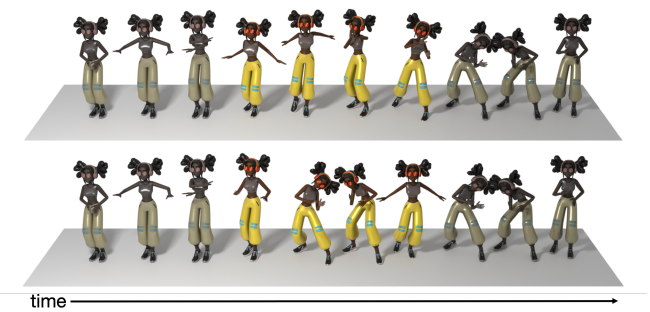


Fig. 5. Temporal composition – In-betweening. Both top and bottom show results of the same input, introducing diverse outputs. The beginning and the end of the motion are given by the reference sequence and can be distinguished according to their faded tone. Observe that the beginning and the end are identical in both sequences. The center of each motion is synthesized.

6.1 Motion Composition

Given a reference motion sequence y, and a region of interest (ROI) mask m, our goal is to synthesize a new motion $\hat{x}_0$, such that the regions of interest $\hat{x}_0 \odot m$ are synthesized from random noise, while the complementary area remains as close as possible to the given motion y, *i.e.*, $y \odot (1 - m) \approx \hat{x}_0 \odot (1 - m)$, where $\odot$ is element-wise multiplication. The network should output a coherent motion sequence, where the transition between given and synthesized parts is seamless. Moreover, the reference motion can be an arbitrary one, on which our network has *not* been trained.

When using a binary mask [Avrahami et al. 2022], as the reference motion y deviates from the motion the network was trained on, the blending between the given and synthesized parts becomes less smooth. To mitigate this issue, we change the ROI mask such that the borders between the given and the synthesized motion segments are linearly interpolated, as depicted in Fig. 4.

We fix the motion segments that need to remain unchanged and sample the parts that need to be filled in. Each step of the iterative inference process (described in Sec. 3.1) is slightly changed, such that parts of y are assigned into $\hat{x}_0$ according to the indices of the mask. That is, $\hat{x}_0 \odot (1 - m) \leftarrow y \odot (1 - m)$.

Temporal composition – use cases: in-betweening, motion expansion. Temporal composition is the action of filling in selected frame sequences. *In-betweening* [Harvey et al. 2020], depicted in Fig. 5, is a special case of temporal composition, where the filled-in part is at the temporal interior of the sequence, and the reference y is from the same distribution as the learned motion. Another special case of temporal composition is *motion expansion*, the motion domain's equivalent of image outpainting [Lin et al. 2021; Teterwak et al. 2019; Yu et al. 2019], where the network generates content that resides beyond the edges of a reference motion sequence. In the case of motion expansion, the ROI mask is zeroed in the center frames, and assigned ones in the outer regions. See Fig. 7 and 6.

Spatial composition – use cases: trajectory control, joints control. Motion composition can be applied spatially, by assigning selected

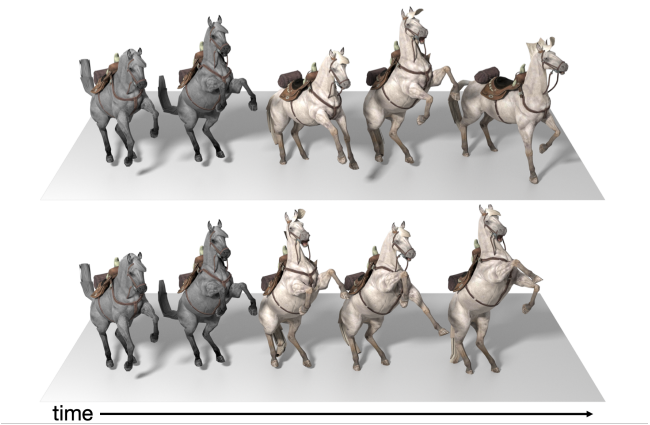


Fig. 6. Temporal composition – motion expansion. Both top and bottom show results of the same input, introducing diverse outputs. The prefix of the motion is provided by the reference sequence and can be identified by its faded color. Note that the prefixes are identical in both sequences. The rest of the motion is synthesized.

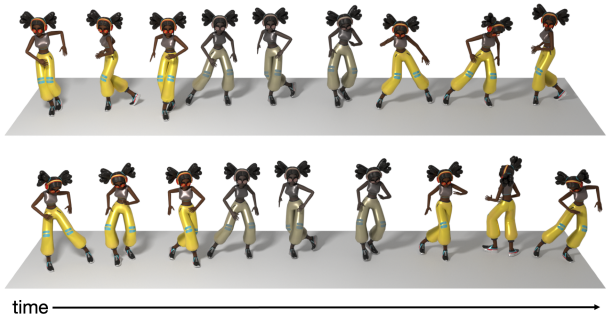


Fig. 7. Temporal composition – motion expansion. Both top and bottom show results of the same input, introducing diverse outputs. The center of the motion is provided by the reference sequence and can be identified by its faded color. Note that the centers are the same in both sequences. The initial and final time segments are synthesized.

joint indices to the ROI mask. In Fig. 8 we illustrate control over the upper body, where the motion of the upper body is determined by a reference motion and assigned to the target motion. The model synthesizes the rest of the joints yielding a motion with the given sequence in the upper body, and with the learned motifs in the lower body. A composition can be both spatial and temporal, and all it takes is an ROI mask where several frame sequences are zeroed, *i.e.*, taken from the reference motion, and in the complementary part, several joints are zeroed (see Fig. 4).

6.2 Harmonization

Given a synthesized motion sequence x_0, we would like to integrate a portion of an unseen motion, y, into it. The portion of y can be temporal, *i.e.*, several frames, or spatial, *i.e.*, several joints, or both. SinMDM overrides a window in x_0 with the desired portion of y and denotes the outcome y_0. Next, y_0 is harmonized such that it matches the core motion elements learned by our network, using a

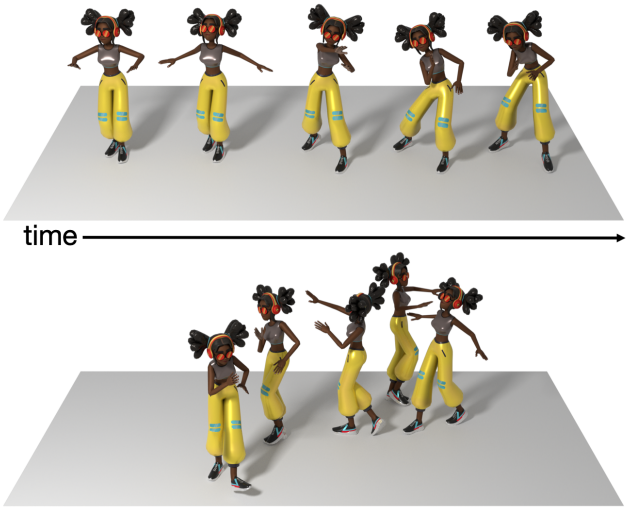


Fig. 8. Spatial composition. Top: reference motion, unseen by the network. Bottom: composed motion. The referenced sequence is *warm-up*, and the learned one is *walk in circle*. In the composed result, the top body part performs a warm-up activity, and the bottom body part walks in a curvy line.

linear low-pass filter ϕ_N as suggested by Choi et al. [2021], via

$$\hat{x}_0 = \phi_N(y_{t-1}) + x_0 - \phi_N(x_0), \tag{5}$$

where y_{t-1} is the noised version of the motion y at step $t-1$.

Note the difference between harmonization and motion composition: Both assign parts of an unseen sequence y into a synthesized motion x_0. However, the former changes the assigned part such that it matches the learned distribution, while the latter aims to keep the assigned part unchanged.

Style transfer. We implement style transfer as a special case of harmonization, where instead of using a portion from y, we use all of it. That is, we fully override x_0. We use a style motion x learned by the model, and a content motion y, unseen by the model. Once applying harmonization, the result possesses the content of y and the style of x, as depicted in Fig. 9.

6.3 Straight-forward Applications

In this section we present applications that may require special techniques in existing works, but require no special technique when conducted using our model.

Long motion sequences. Our network can synthesize variable-length motions, even very long ones, with no additional training. Imputed to its small receptive field, the network can hallucinate a sequence as long as requested. An example of a one-minute animation is introduced in Fig. 10. Note that no single motion current work can generate an arbitrary length without retraining or using an auto-regressive procedure.

Crowd animation. Although trained on a single sequence, during inference SinMDM can generate a crowd performing a variety of

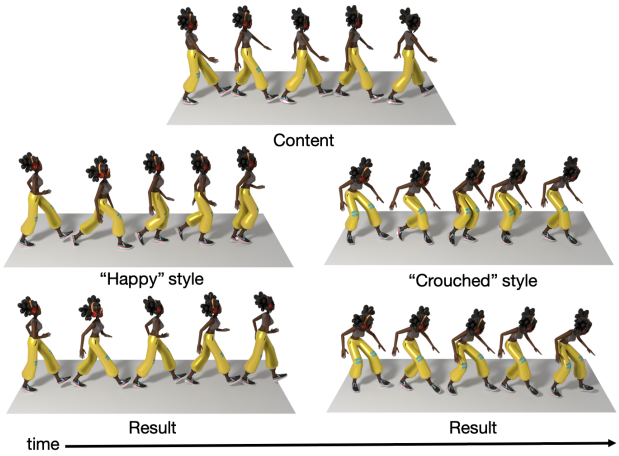


Fig. 9. Style transfer is a special case of the harmonization application, where a reference motion is adjusted such that it matches the learned motion motifs. The style motion is learned by the model, and the content motion is unseen by it. Top: one content, unseen by the network, is applied to both styles. Left: a "happy" style, learned by the network, and below it the harmonized result. Right: a "crouched" style, learned by the network, and below it the harmonized result. Note that the character in both results is using the exact step rhythm and size as the character in the content motion.

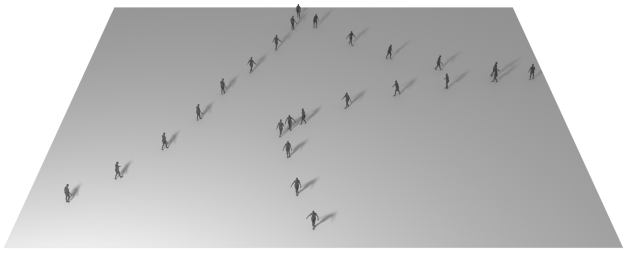


Fig. 10. Long motion. The learned sequence is a 10 seconds motion, depicting a person walking back, then turning and walking back again. The synthesized motion is a 60 seconds sequence, depicting a person walking back, and occasionally turning and walking back again.

similar motions, each sampled from a different Gaussian noise $x_T \sim \mathcal{N}(0, I)$, as illustrated in Fig. 11.

7 CONCLUSIONS

We have explored the use of diffusion models on single motion sequence synthesis and designed a motion denoising transformer with a narrow receptive field. Training on single motions is particularly useful in motion domains, where the number of data instances is scarce. Particularly, for animals and imaginary creatures, which

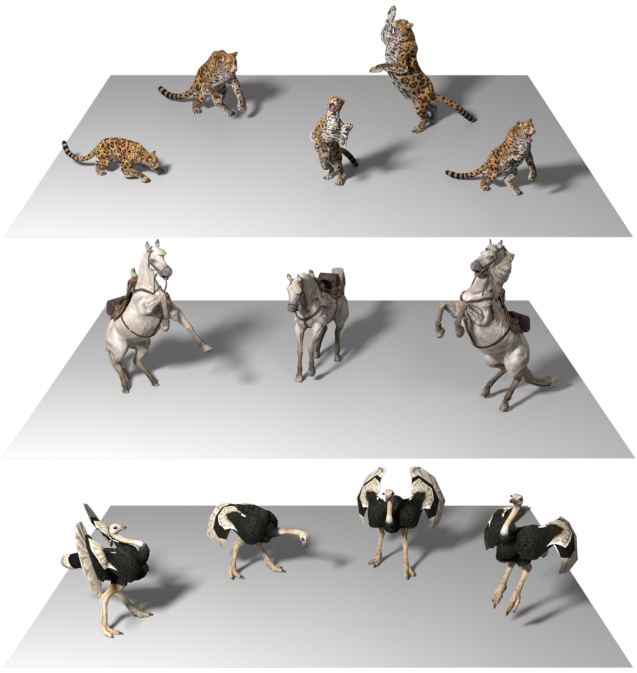


Fig. 11. Crowd animation. Groups of jaguars, horses, and ostriches. In each group, no motion is like the other, and yet they are all learned from a single motion sequence.

have unique skeletons and motion motifs. The motion of such creatures cannot be captured easily nor learned from the human motion data available.

Our experiments on several datasets demonstrate that our lightweight diffusion-based method significantly outperforms current work both in quality and time-space performance. Moreover, our approach allows the synthesis of particularly long motions, and enables a variety of motion manipulation tasks, including spatial and temporal in-betweening, motion expansion, harmonization, style transfer, and crowd animation.

The innate limitation of our method, common to all models (in all domains) that learn a single instance, is the limited ability to synthesize out-of-distribution. However, the main limitation of our diffusion-based approach is the relatively long inference time. This is due to the iterative nature of diffusion models.

Finally, our work shows the competence of diffusion models to learn from limited data, which contradicts their reputation for requiring large amounts of data. Nevertheless, in the future, we would like to address the single input limitations, by possibly learning from available motion data of creatures with rather compatible skeletons.

8 ACKNOWLEDGMENTS

We are grateful to Panayiotis Charalambous, Andreas Aristidou and Brian Gordon for reviewing earlier versions of the manuscript. This research was supported in part by the Israel Science Foundation (grants no. 2492/20 and 3441/21), Len Blavatnik and the Blavatnik family foundation, and the Tel Aviv University Innovation Laboratories (TILabs).

REFERENCES

Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. 2020a. Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 62–1.

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), 64–1.

Kfir Aberman, Rundi Wu, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. 2019. Learning Character-Agnostic Motion for Motion Retargeting in 2D. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 75.

Adobe Systems Inc. 2021. Mixamo. https://www.mixamo.com Accessed: 2021-12-25.

Chaitanya Ahuja and Louis-Philippe Morency. 2019. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*. IEEE, IEEE Computer Society, Washington, DC, USA, 719–728.

Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. 2019. Structured prediction helps 3d human motion modelling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE Computer Society, Washington, DC, USA, 7144–7153.

Moab Arar, Ariel Shamir, and Amit H Bermano. 2022. Learned Queries for Efficient Local Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, Washington, DC, USA, 10841–10852.

A Aristidou, A Yiannakidis, K Aberman, D Cohen-Or, A Shamir, and Y Chrysanthou. 2022. Rhythm is a Dancer: Music-Driven Motion Synthesis with Global Structure. *IEEE Transactions on Visualization and Computer Graphics* 1 (2022).

Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. 2020. A critical analysis of self-supervision, or what we can learn from a single image. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, OpenReview.net. https://openreview.net/forum?id=B1esx6EYvr

Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, Washington, DC, USA, 18208–18218.

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

Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. 2021. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, IEEE Computer Society, Washington, DC, USA, 1–10.

Pablo Cervantes, Yusuke Sekikawa, Ikuro Sato, and Koichi Shinoda. 2022. Implicit Neural Representations for Variable Length Human Motion Generation.

Jinshu Chen, Qihui Xu, Qi Kang, and MengChu Zhou. 2021. Mogan: Morphologic-structure-aware generative learning from a single image.

Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. 2021. Ilvr: Conditioning method for denoising diffusion probabilistic models. In 2021 IEEE. In *CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Washington, DC, USA, 14347–14356.

Bruno Degardin, João Neves, Vasco Lopes, João Brito, Ehsan Yaghoubi, and Hugo Proença. 2022. Generative Adversarial Graph Convolutional Networks for Human Action Synthesis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. IEEE Computer Society, Los Alamitos, CA, USA, 1150–1159.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems* 34 (2021), 8780–8794.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, OpenReview.net. https://openreview.net/forum?id=YicbFdNTTy

Yinglin Duan, Tianyang Shi, Zhengxia Zou, Yenan Lin, Zhehui Qian, Bohan Zhang, and Yi Yuan. 2021. Single-shot motion completion with transformer.

Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. 2015. Recurrent network models for human dynamics. In *Proceedings of the IEEE international conference on computer vision*. IEEE Computer Society, Washington, DC, USA, 4346–4354.

Saeed Ghorbani, Calden Wloka, Ali Etemad, Marcus A. Brubaker, and Nikolaus F. Troje. 2020. Probabilistic Character Motion Synthesis using a Hierarchical Deep Latent Variable Model. *Computer Graphics Forum* 1 (2020). https://doi.org/10.1111/cgf.14116

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

Brian Gordon, Sigal Raab, Guy Azov, Raja Giryes, and Daniel Cohen-Or. 2022. FLEX: Extrinsic Parameters-free Multi-view 3D Human Motion Reconstruction. In *European Conference on Computer Vision*. Springer, Springer International Publishing, Berlin/Heidelberg, Germany, 176–196.

Niv Granot, Ben Feinstein, Assaf Shocher, Shai Bagon, and Michal Irani. 2022. Drop the gan: In defense of patches nearest neighbors as single image generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, Washington, DC, USA, 13460–13469.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022. Generating Diverse and Natural 3D Human Motions From Text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, Washington, DC, USA, 5152–5161.

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*. ACM, New York, NY, USA, 2021–2029.

Ikhsanul Habibie, Daniel Holden, Jonathan Schwarz, Joe Yearsley, and Taku Komura. 2017. A recurrent variational autoencoder for human motion synthesis. In *28th British Machine Vision Conference*. BMVC, UK.

Félix G Harvey and Christopher Pal. 2018. Recurrent transition networks for character locomotion. In *SIGGRAPH Asia 2018 Technical Briefs*. ACM, New York, NY, USA, 1–4.

Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. 2020. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 60–1.

Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. 2019. Human motion prediction via spatio-temporal inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE Computer Society, Washington, DC, USA, 7134–7143.

Tobias Hinz, Matthew Fisher, Oliver Wang, and Stefan Wermter. 2021. Improved techniques for training single-image gans. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. IEEE Computer Society, Washington, DC, USA, 1300–1309.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance.

Daniel Holden, Ikhsanul Habibie, Ikuo Kusajima, and Taku Komura. 2017. Fast neural style transfer for motion data. *IEEE computer graphics and applications* 37, 4 (2017), 42–49.

Daniel Holden, Jun Saito, and Taku Komura. 2016. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–11.

Daniel Holden, Jun Saito, Taku Komura, and Thomas Joyce. 2015. Learning motion manifolds with convolutional autoencoders. In *SIGGRAPH Asia 2015 technical briefs*. ACM, New York, NY, USA, 1–4.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. 2016. Deep networks with stochastic depth. In *European conference on computer vision*. Springer, Springer International Publishing, Berlin/Heidelberg, Germany, 646–661.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE Computer Society, Washington, DC, USA, 1125–1134.

Deok-Kyeong Jang and Sung-Hee Lee. 2020. Constructing human motion manifold with sequential networks. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, The Eurographics Association and John Wiley and Sons Ltd., Hoboken, NJ, USA, 314–324.

Manuel Kaufmann, Emre Aksan, Jie Song, Fabrizio Pece, Remo Ziegler, and Otmar Hilliges. 2020. Convolutional autoencoders for human motion infilling. In *2020 International Conference on 3D Vision (3DV)*. IEEE, IEEE Computer Society, Washington, DC, USA, 918–927.

Jihoon Kim, Jiseob Kim, and Sungjoon Choi. 2022. FLAME: Free-form Language-based Motion Synthesis & Editing.

Vladimir Kulikov, Shahar Yadin, Matan Kleiner, and Tomer Michaeli. 2022. SinDDM: A Single Image Denoising Diffusion Model.

Juheon Lee, Seohyun Kim, and Kyogu Lee. 2018. Listen to dance: Music-driven choreography generation using autoregressive encoder-decoder network.

Chuan Li and Michael Wand. 2016. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European conference on computer vision*. Springer, Springer International Publishing, Berlin/Heidelberg, Germany, 702–716.

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.

Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. 2021. Learn to dance with aist++: Music conditioned 3d dance generation. , arXiv–2101 pages.

Yan Li, Tianshu Wang, and Heung-Yeung Shum. 2002. Motion texture: a two-level statistical model for character motion synthesis. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*. ACM, New York, NY, USA, 465–472.

Chieh Hubert Lin, Hsin-Ying Lee, Yen-Chi Cheng, Sergey Tulyakov, and Ming-Hsuan Yang. 2021. InfinityGAN: Towards Infinite-Pixel Image Synthesis. In *International Conference on Learning Representations*. OpenReview.net, OpenReview.net.

Jianxin Lin, Yingxue Pang, Yingce Xia, Zhibo Chen, and Jiebo Luo. 2020. TuiGAN: Learning versatile image-to-image translation with two unpaired images. In *European Conference on Computer Vision*. Springer, Springer International Publishing, Berlin/Heidelberg, Germany, 18–35.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)* 34, 6 (2015), 1–16.

Shubh Maheshwari, Debtanu Gupta, and Ravi Kiran Sarvadevabhatla. 2022. MUGL: Large Scale Multi Person Conditional Action Generation with Locomotion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. IEEE Computer Society, Los Alamitos, CA, USA, 257–265.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (Eds.). PMLR, PMLR, 16784–16804. https://proceedings.mlr.press/v162/nichol22a.html

Yaniv Nikankin, Niv Haim, and Michal Irani. 2022. SinFusion: Training Diffusion Models on a Single Image or Video.

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)*. IEEE Computer Society, Washington, DC, USA, 10985–10995.

Mathis Petrovich, Michael J. Black, and Gül Varol. 2022. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*. Springer International Publishing, Berlin/Heidelberg, Germany.

Ofir Press, Noah A Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation.

Jia Qin, Youyi Zheng, and Kun Zhou. 2022. Motion In-Betweening via Two-Stage Transformers. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–16.

Sigal Raab, Inbal Leibovitch, Peizhuo Li, Kfir Aberman, Olga Sorkine-Hornung, and Daniel Cohen-Or. 2022. MoDi: Unconditional Motion Synthesis from Diverse Data.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. 2022. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*. ACM, New York, NY, USA, 1–10.

Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. 2019. SinGAN: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE Computer Society, Washington, DC, USA, 4570–4580.

Assaf Shocher, Shai Bagon, Phillip Isola, and Michal Irani. 2019. Ingan: Capturing and retargeting the" dna" of a natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE Computer Society, Washington, DC, USA, 4492–4501.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*. PMLR, PMLR, PMLR, 2256–2265.

Minjung Son, Jeong Joon Park, Leonidas Guibas, and Gordon Wetzstein. 2022. SinGRAF: Learning a 3D Generative Radiance Field for a Single Scene.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020a. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*. OpenReview.net, OpenReview.net.

Yang Song and Stefano Ermon. 2020. Improved techniques for training score-based generative models. *Advances in neural information processing systems* 33 (2020), 12438–12448.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020b. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*. OpenReview.net, OpenReview.net.

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.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding.

Guofei Sun, Yongkang Wong, Zhiyong Cheng, Mohan S Kankanhalli, Weidong Geng, and Xiangdong Li. 2020. DeepDance: music-to-dance motion choreography with adversarial learning. *IEEE Transactions on Multimedia* 23 (2020), 497–509.

Wenyue Sun and Bao-Di Liu. 2020. ESinGAN: Enhanced single-image GAN using pixel attention mechanism for image super-resolution. In *2020 15th IEEE International Conference on Signal Processing (ICSP)*, Vol. 1. IEEE, IEEE Computer Society, Washington, DC, USA, 181–186.

Vadim Sushko, Dan Zhang, Juergen Gall, and Anna Khoreva. 2021. Generating Novel Scene Compositions from Single Images and Videos.

Piotr Teterwak, Aaron Sarna, Dilip Krishnan, Aaron Maschinot, David Belanger, Ce Liu, and William T Freeman. 2019. Boundless: Generative adversarial networks for image extension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE Computer Society, Washington, DC, USA, 10521–10530.

Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. 2022a. MotionCLIP: Exposing Human Motion Generation to CLIP Space.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. 2022b. Human motion diffusion model.

Truebones Motions Animation Studios. 2022. Truebones. https://truebones.gumroad.com/ Accessed: 2022-1-15.

Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee. 2018. Neural kinematic networks for unsupervised motion retargetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, Washington, DC, USA, 8639–8648.

Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. 2022. SinDiffusion: Learning a Diffusion Model from a Single Natural Image.

Zhenyi Wang, Ping Yu, Yang Zhao, Ruiyi Zhang, Yufan Zhou, Junsong Yuan, and Changyou Chen. 2020. Learning diverse stochastic human-action generators by learning smooth latent transitions. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. AAAI Press, Washington, DC, USA, 12281–12288.

Rui Xu, Xintao Wang, Kai Chen, Bolei Zhou, and Chen Change Loy. 2021. Positional encoding as spatial inductive bias in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, Washington, DC, USA, 13569–13578.

Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. 2019. Convolutional sequence generation for skeleton-based action synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE Computer Society, Washington, DC, USA, 4394–4402.

Jihyeong Yoo and Qifeng Chen. 2021. SinIR: Efficient General Image Manipulation with Single Image Reconstruction. In *International Conference on Machine Learning*. PMLR, PMLR, PMLR, 12040–12050.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. 2019. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*. IEEE Computer Society, Washington, DC, USA, 4471–4480.

Ping Yu, Yang Zhao, Chunyuan Li, Junsong Yuan, and Changyou Chen. 2020. Structure-aware human-action generation. In *European Conference on Computer Vision*. Springer, Springer International Publishing, Berlin/Heidelberg, Germany, 18–34.

Ye Yuan and Kris Kitani. 2020. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision*. Springer, Springer International Publishing, Berlin/Heidelberg, Germany, 346–364.

Jia-Qi Zhang, Xiang Xu, Zhi-Meng Shen, Ze-Huan Huang, Yang Zhao, Yan-Pei Cao, Pengfei Wan, and Miao Wang. 2021c. Write-An-Animation: High-level Text-based Animation Editing with Character-Scene Interaction. In *Computer Graphics Forum*, Vol. 40. Wiley Online Library, The Eurographics Association and John Wiley and Sons Ltd., Hoboken, NJ, USA, 217–228.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2022a. MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE Computer Society, Washington, DC, USA, 586–595.

Yan Zhang, Michael J Black, and Siyu Tang. 2021a. We are more than our joints: Predicting how 3d bodies move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, Washington, DC, USA, 3372–3382.

ZiCheng Zhang, CongYing Han, and TianDe Guo. 2021b. ExSinGAN: Learning an Explainable Generative Model from a Single Image.

Zicheng Zhang, Yinglu Liu, Congying Han, Hailin Shi, Tiande Guo, and Bowen Zhou. 2022b. PetsGAN: Rethinking Priors for Single Image Generation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI*. AAAI Press, Washington, DC, USA, 3408–3416. https://ojs.aaai.org/index.php/AAAI/article/view/20251

Zilong Zheng, Jianwen Xie, and Ping Li. 2021. Patchwise generative convnet: Training energy-based models from a single natural image for internal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, Washington, DC, USA, 2961–2970.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2019. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, Washington, DC, USA, 5745–5753.

Yi Zhou, Zimo Li, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. 2018. Auto-Conditioned Recurrent Networks for Extended Complex Human Motion Synthesis.

In *International Conference on Learning Representations*. OpenReview.net, OpenReview.net.