Learning Robust and Scalable Motion Matching with Lipschitz Continuity and Sparse Mixture of Experts

Tobias Kleanthous
tobiask@global.tencent.com
Production R&D
Tencent
Edinburgh, UK

Antonio Martini
antonio@global.tencent.com
Production R&D
Tencent
London, UK

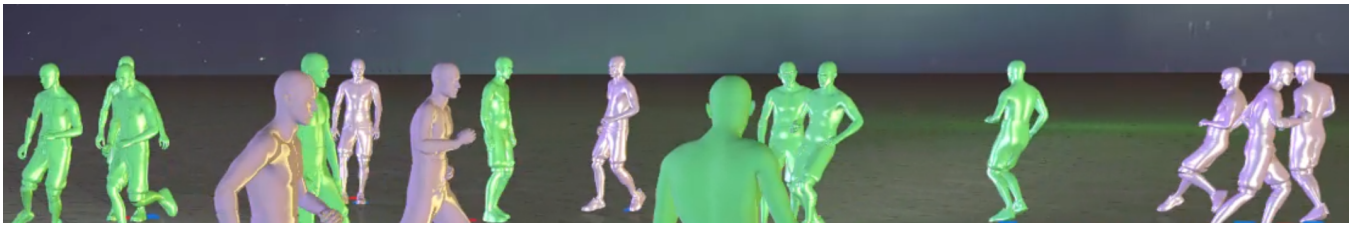


Figure 1: Our method, enabling scalable motion matching across a variety of styles and larger character counts concurrently.

ABSTRACT

Motion matching(Büttner and Clavet [2015]; Clavet [2016]) has become a widely adopted technique for generating high-quality interactive animation systems in video games. However, its current implementations suffer from significant computational and memory resource overheads, limiting its scalability in the context of modern video game performance profiles.

"Learned Motion Matching"[Holden et al. 2020] mitigated some of these challenges, however, whilst reducing memory requirements, it resulted in increases in performance costs. In this paper, we propose a novel method for learning motion matching that combines a Sparse Mixture of Experts model architecture and a Lipschitz-continuous latent space for representation of poses.

This approach significantly reduces the computational complexity of the models, while simultaneously improving the compactness of the data that can be stored and the robustness of pose output. As a result, our method enables the efficient execution of motion matching that significantly outperforms other implementations for large character counts, by 8.5x times in CPU execution cost and at 80% of the memory requirements of "Learned Motion Matching", on contemporary video game hardware, thereby enhancing its practical applicability and scalability in the gaming industry.

CCS CONCEPTS

• **Computing methodologies** → **Regularization**; **Procedural animation**; **Motion capture**.

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

ACM ISBN 979-8-4007-0393-5/23/11…$15.00
https://doi.org/10.1145/3623264.3624442

KEYWORDS

character animation, animation, neural networks, motion matching, regularization

ACM Reference Format:
Tobias Kleanthous and Antonio Martini. 2023. Learning Robust and Scalable Motion Matching with Lipschitz Continuity and Sparse Mixture of Experts. In *ACM SIGGRAPH Conference on Motion, Interaction and Games (MIG '23), November 15–17, 2023, Rennes, France.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3623264.3624442

1 INTRODUCTION

Large-scale *"AAA"* games regularly push limits in seeking to create ever richer, more characterful and naturalistic performances to entertain and immerse their players. The ongoing increases in complexity of character animation systems in modern video games has been well documented in an escalating battle for hardware resources, such as CPU and memory. In recent years, video games developers have embraced a wide-range of techniques to deliver hundreds of styles across thousands of animations when striving to capture the imaginations of players at hitherto unprecedented scales, but not without substantial costs in terms of authoring complexity and production[Kleanthous 2021].

Research in the area of interactive character animation has sought to manage the complexity of implementations, whilst extending the expressiveness available to games developers, from parametric animation graphs, to *motion fields*[Lee et al. 2010] and the increasingly popular approach of *motion matching*(Büttner and Clavet [2015]; Clavet [2016]), before the now growing body of research in applying machine learning techniques to animated character control(Holden et al. [2017]; Starke et al. [2020]).

In this paper, we build upon the successful foundations of both motion matching and machine learning approaches to motion synthesis, to improve scalability and robustness of these techniques. We demonstrate meaningful improvements in runtime performance

metrics to broaden their application across the breadth of devices which modern video games are released on.

2 RELATED WORK

In this section, we highlight related works covering animation control, limitations concerning compute complexity of machine learning models and their execution, and approaches to smooth representations of data in *neural networks (NNs)*.

2.1 Motion Matching and Learned Motion Matching

Motion matching has gained popularity for its expressive control, whilst enabling users to move away from the laborious manual creation and maintenance of stately graphs to control the flow and blend of many animations simultaneously(Büttner and Clavet [2015]; Clavet [2016]). The core principles of a system require analysing some mixture of user driven control parameters and internally derived parameters stored in a database that can uniquely identify desirable transitions at runtime to assist in striking a balance between continuity and satisfying goals for a given action. Commonly, features such as future and historic trajectories, alongside pose driven features, such as model space velocities of joints, are used.

Whilst the qualitative output and control achievable is desirable, a substantial requirement for additional data storage and CPU cost is incurred when deploying motion matching. *Learned Motion Matching (LMM)*[Holden et al. 2020] addressed issues concerning memory requirements by reducing the database and animations instead to a collection of smaller *feedfoward networks (FFNs)*, however, applicability was limited by the needs to balance a reduction of memory with an increase in CPU costs, an already heavily contested resource in optimising large-scale video games. Other methods to improve runtime performance such as GPU based execution[Raparthi et al. 2020] have proven effective, but make trade offs with latency of control, that prevent responsive player control of characters, and again, are in contention with a resource heavily used in accelerating high-quality rendering.

In this work, we show that the promise of a robust, and comparatively cost-effective execution of such techniques via *NNs* is possible to achieve in a manner that excels beyond prior methods.

2.2 Compute Complexity of Neural Network Inference on Local Hardware

The main performance limitation of using feedforward networks *(FFN)* for *LMM*[Holden et al. 2020], stems from the fact that, as the motion database increases in size, it is necessary to increase *FFN* depth and width in order to preserve motion reproduction accuracy, resulting in a drastic increase in the number of model weights which leads to run-time evaluation times beyond production budget (See also Section 8 in [Holden et al. 2020]). This problem was also recognised in [Maiorca et al. 2022], which attempted to generate a sparsified Mixture of Experts*(MoE)* by post-processing the model weights with a pruning scheme. This weights sparsification approach however suffers from severe limitations as increases in model performance are directly correlated to decreased motion reproduction accuracy, resulting either in visual artifacts or in limited performance gains. An additional problem noted in the same work, is that as sparse matrix-vector operations are not supported in hardware, the theoretical gains do not necessarily translate to faster execution times.

Inspired by recent advances in the field of large language models and conditional computation(Fedus et al. [2022]; Lepikhin et al. [2021]; Zoph et al. [2022]), in this work we show for the first time that, with some small modifications and additional training considerations, *SMoE* type layers applied to motion synthesis result in highly accurate and performant architectures, without requiring any post training simplifications.

2.3 Smooth Latent Manifolds via Lipschitz Continuous Neural Networks

We note, that components of animations retain temporospatial relationships that express smoothness in their raw form, which other methods seek to exploit[Starke et al. 2022]. By maximising the preservation of this relationship in a latent manifold, we seek to induce an intuitive mapping that can be further exploited to improve utilisation and quality. Lipschitz continuity in neural networks provides an interesting avenue of research when seeking to enforce this relationship.

A function f is defined as Lipschitz continuous when there is some constant c that satisfies

$$||f(x) - f(y)|| \leq c \cdot ||x - y|| \tag{1}$$

In broad terms, applied to *NNs*, by enforcing Lipschitz continuity we can ensure there that the similarity of a given pair of inputs can lead to a direct relationship with the clear similarity of the pair's respective *NN* outputs. For example, in terms of maximising expressable continuity of consecutive frames of animation. By enforcing minimized weights of layers within some learned bounds(Gouk et al. [2020]; Liu et al. [2022]) the benefits are two-fold, a more robust and better performing model can be learned with such an approach, that also more readily generalizes across unseen data[Miyato et al. 2018]. This is an intriguing proposition in the context of animation in machine learning, where common approaches have relied on a more compact encoding of the animation data as a working representation. Expressing a close relative relationship between animation data and their latent representation presents a benefit for the predictability of their usage.

Various techniques, such as *variadic autoencoders (VAEs)*[Ling et al. 2020], have been applied to address smoothness and continuity when modelling animation tasks, with mixed results, often suffering from over smoothing. *Periodic phase autoencoders (PAEs)*[Starke et al. 2022] that learn approximations of temporospatial relationships which can be used to align animation representations, have shown great success, but do not fully imbue that ability within a single model, instead solving a more easily learned relationship that can be subsequently used in other models such as *phase functioned neural networks (PFNNs)*[Holden et al. 2017].

Influenced by [Liu et al. 2022] applying Lipschitz continuous *NNs* for geometric interpolation, we show in this work, that a Lipschitz continuous encoder, applied to animation data and techniques seen in *LMM*[Holden et al. 2020], can improve the quality of results,

not only within the decoded latent pose representation, but the predictive components that leverage the encoder's latent output, by better representing the original relationships of animation frames in the latent manifold.

3 CORE CONCEPTS

In this section, we discuss the core details of our method and our application of *SMoE* and Lipschitz continuity to *NNs* for animation control.

3.1 Sparse Mixture of Experts (SMoE)

We employ a SMoE layer architecture with top-2 routing [Zoph et al. 2022]. As we are not dealing with discrete tokens like in NLP applications, and in order to avoid any form of motion jitter, we avoid the probabilistic dropping of the second expert as typically done in existing work(Lepikhin et al. [2021]; Zoph et al. [2022]) As illustrated in Figure 2, a gating layer takes x as input and routes it to the best matched top-2 experts selected out of a set $\{E_i(x)\}_{i=1}^{N}$ of N experts. The gating logits $h(x) = LINEAR(x)$ are calculated by passing the input through a linear layer, the final gate values are then obtained via a softmax operation which results in a probability distribution over the N experts

$$p_i(x) = \frac{e^{h(x)_i}}{\sum_j^N e^{h(x)_j}} \tag{2}$$

The input x is then routed to the experts with the top-2 gate values. The final results is computed as the weighted sum of the top-2 active experts output weighted by their associated normalised gate values:

$$y = \frac{p_{1st}(x)}{p_{1st}(x) + p_{2nd}(x)} E_{1st}(x) + \frac{p_{2nd}(x)}{p_{1st}(x) + p_{2nd}(x)} E_{2nd}(x) \tag{3}$$

Where $1st$ and $2nd$ refer to the indices of respectively the highest and second highest gate values in Eq.2. Experts $E_i(x)$ are feed-forward networks(FFN) sharing the same architecture and employing *GELU*[Hendrycks and Gimpel 2023] activation functions, which we have found to provide superior performance to ELU, ReLU and LeakyReLU activations. A *SMoE* architecture can be implemented as a single layer which can then be integrated as a part of a larger model. Further, it is worth noting that single experts have no predefined interpretable meaning, how data is routed to experts and what each single expert learns is automatically determined during training.

3.2 Lipschitz Layers for Smooth Continuity in Latent Space

We implement a Lipschitz linear layer as in section 4.1.1 of [Liu et al. 2022]. A learnable Lipschitz bound c is used to normalize the weight matrix W with rows R, using scaling vector created by the function $Scale$.

$$Scale(c, W) = Min(\frac{softplus(c)}{Max(\sum_{r=1}^{R} |W_r|, \epsilon)}, 1.0) \tag{4}$$

$$\epsilon = 1x10^{-7} \tag{5}$$

$$WeightNormalization(W_i) = W_i \cdot Scale(c, W) \tag{6}$$

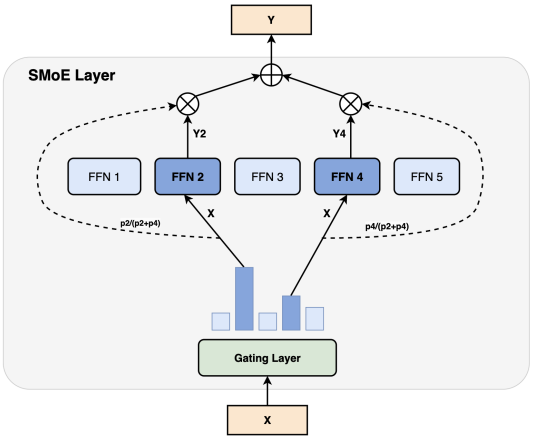


Figure 2: Example of activations in a SMoE layer with 5 experts and top-2 gating

This weight normalization is then applied when treating the Lipschitz layer interchangeably with a linear layer in a model, with bias b.

$$LipschitzLinear(X) = X * WeightNormalization(W) + b \tag{7}$$

For initialisation, for W we use the same Kaiming initialization[He et al. 2015] used in *PyTorch's*[Paszke et al. 2019] linear layer implementation and then initialize c by taking the $L\infty$-norm of W. At runtime, c remains constant and the normalization of the weight matrices within a model can thus be evaluated as constant when preparing the model for inference.

4 ARCHITECTURE

In this section, we highlight the architecture of our method and models.

4.1 Pose Representation

To represent a single pose, we combine a hierarchy of J joints and a pair of linear and angular root motion velocities.

Individual joints are formed of a position and rotation. We use positions and root motion velocities of the form $P \in \mathbb{R}^3$. For rotations, after experimentation across Euler angles and quaternions, as [Pavllo et al. 2018], we decided on 6D representations due to their superior ability to better express continuity in a neural network(Li et al. [2022]; Zhou et al. [2020]), we denote as $R \in \mathbb{R}^6$.

Poses are assessed during training both in local and model (or character) spaces, where model space is the result of a forward kinematics operator $FK(P, R)$ on the pose hierarchy.

4.2 Model Architecture

We structure our method similarly as seen in *LMM*[Holden et al. 2020], in this section, we detail differences according to the unique approaches in our method.

4.2.1 Pose Autoencoder. Our pose autoencoder structure follows similar patterns in related work(Holden et al. [2020]; Ling et al. [2020]), we train the complete encoder and decoder pairing, with

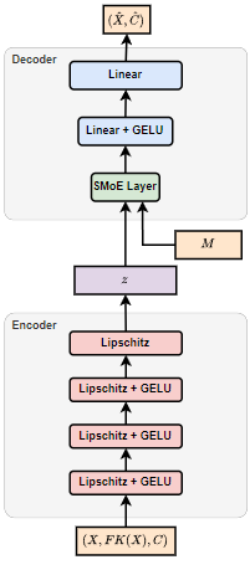


Figure 3: Example flow of our pose autoencoder.

the ability to discard the encoder component at runtime. The inputs to our model is the concatenation of the pose in local space X, the model space evaluation after applying forward kinematics $FK(X)$ and any applicable tagging, such as foot contacts $C \in 0.0, 1.0$. The output is our pose prediction in local space $\hat{X}$, including root motion velocities, and any probabilities for any associated tagging per pose $\hat{C}$.

We form an encoder with several linear layers, each with a learnable Lipschitz constant (see section 3.2). As per [Holden et al. 2020], the resultant latent encoding z is then concatenated with standardized motion matching features M, extracted from the training data. Finally, we feed the components into our decoder, where we use the *SMoE* architecture to distribute across 128 significantly smaller experts prior to reforming the original pose dimensions in 2 following linear layers (see Figure 3 and Encoder, Decoder and *SMoE* structures in Tables 5, 6 and 4 respectively). Lastly, we perform sigmoid activation on the final 2 components of the pose output to recover probabilities of respective foot contact.

In experimentation, we found applying the Lipschitz layers in our decoder to produce a degradation in the quality of our final results, with little impact on the regularization observed in the pose latent. We believe this is, in part, due to the lack of enforcement of smoothness on all components of the concatenated motion matching features, and as a result, applying a Lipschitz bound on the decoder negatively impacts the ability for the weights to learn usefully from this additional input.

Whilst we experimented with a larger latent, eschewing the provision of the motion matching features as input, we did not sufficiently solve the need for recovery of those features for runtime evaluation elsewhere in the system. We propose further investigation to improve on this in the future to ensure full Lipschitz continuity of all components of the model.

4.2.2 Stepper. As in [Holden et al. 2020], we produce a stepper model that is autoregressively trained to predict output deltas of the latent z and motion matching features M, δz and δM respectively. This model effectively replaces the point in a traditional animation system or motion matching implementation, where we are able to sample an animation on a future frame, instead predicting what the values will be.

Each step is equivalent to a timestep of

$$\delta t = \frac{1}{fps} \tag{8}$$

where fps, or frames per second, is $60Hz$ in our data. Thus, training starts at sampled frame of the concatenated inputs (z_0, M_0) and is then predicted autoregressively on some number of N predictions up to (z_n, M_n), in our case, for $N = 30$ further updates.

$$\delta\hat{z}_n, \delta\hat{M}_n = Stepper(z_n, M_n) \tag{9}$$

$$\hat{z}_{n+1} = \hat{z}_n + \delta\hat{z}_n \tag{10}$$

$$\hat{M}_{n+1} = \hat{M}_n + \delta\hat{M}_n \tag{11}$$

Consequently, we are motivated as in [Liu et al. 2022], to ensure our latent representation of a pose is smooth such that we may sub-sample on intervals where

$$\delta t < \frac{1}{fps} \tag{12}$$

thereby, being able to smoothly sample in-between frames, as we do when sampling the original source animation. We cumulatively gather losses for the predictions and velocity similarities over N and then take the mean. In all cases, we use a mean squared error *MSE*, where the predicted latent and predicted latent delta are weighted twice that of the equivalent for motion matching features.

Our implementation is entirely based on a *SMoE* layer, detailed in table 7 in Appendix E. Furthermore, we induce a robustness to noisy input at training time, as detailed in section 5.8.

4.2.3 Projector. We implement a *projector* model as in [Holden et al. 2020], with the same cost functions, whose purpose is to provide a replacement for the KD-tree (or other spatial partitioning system) as seen in *motion matching*[Clavet 2016]. As with our *stepper* implementation, we form our projector entirely as a *SMoE*, further detailed in Appendices E and F.

5 TRAINING

We provide further implementations of our training in this section. Our models and training are implemented using *PyTorch*[Paszke et al. 2019]. All training was performed across internal compute servers utilising *Nvidia V100* GPUs.

Owing to the nature of the implementation using multiple models, we are able to train the stepper and projector models concurrently after training has completed on the pose autoencoder.

For all models, we train using the Adan optimizer[Xie et al. 2023], which we found to have a significant accuracy and performance advantage over commonly used alternatives. We set a learning rate $lr = 0.02$ with no weight-decay, and used a simple learning rate scheduler that decreased by factor $\gamma = 0.99$ for every 3 virtual epochs where the weighted moving average of overall reported losses do not reduce.

In practice, our models completed training to a desired level on assessed metrics before fully converging, thus, given an open-ended number of steps and compute time, we would expect to see further improvements in the quality of results.

At the time of writing, a full set of models can be trained with sub-centimetre accuracy and be used in-game in 9 hours.

5.1 SMoE Expert Utilization

In order to ensure efficient experts utilisation, we add an auxiliary load balancing loss term during training. We follow the approach of [Fedus et al. 2022] which is summarised below for convenience.

Given N experts indexed by $i = 1$ to N and a batch $\mathcal{B}$ with T batch entries, the auxiliary loss is computed as the scaled dot-product between vectors f and P,

$$\text{LoadBalancingLoss} = \alpha \cdot N \cdot \sum_{i=1}^{N} f_i \cdot P_i \tag{13}$$

where α is the loss relative weight within the overall model cost function and it is set to 0.01 for all models, f_i is the fraction of batch entries dispatched to expert i,

$$f_i = \frac{1}{T} \sum_{x \in \mathcal{B}} \mathbb{1}\{\text{argmax } p(x) = i\} \tag{14}$$

and P_i is the fraction of the gating layer probability allocated for expert i,

$$P_i = \frac{1}{T} \sum_{x \in \mathcal{B}} p_i(x). \tag{15}$$

The auxiliary loss of Eq.13 encourages uniform routing of the batch entries since it is minimized under a uniform distribution, this is when both vectors f_i and P_i have values of $1/N$. See also [Fedus et al. 2022] for additional details.

Similarly to what suggested in [Zoph et al. 2022], we have found top-2 routing to be a good compromise between model accuracy and run-time performance. However, contrarily to what observed in past work on large language models[Zoph et al. 2022], where larger batch sizes result in model deterioration, we have instead found that for motion data, larger batch sizes consistently results in more efficient and accurate models.

5.2 Pose Losses in Autoencoder Training

When evaluating poses during training of the autoencoder, we consider the weighting per joint and evaluate the results both in local and model spaces. As noted in section 5.3, the impact of error in local space produces a larger proportional error on child joints further up the pose hierarchy, so local space accuracy is of some importance to guiding our models towards producing more accurate results. We create a pose representation for a number of joints J output from our autoencoder $\hat{P}$ and compare to ground truth poses in our training set P.

For positional and root motion velocity losses, we evaluate across batch $\mathcal{B}$, with per-joint weights W using an $L1$ loss per joint. We found this discouraged an averaging effect over reconstruction of positions compared to $L2$

$$PositionalLoss(P, \hat{P}, W) = \frac{1}{\mathcal{J}} \sum_{j=1}^{\mathcal{J}} W_j \cdot ||P_j - \hat{P}_j||_1 \tag{16}$$

For rotational losses, we take the two normalized orthogonal vector components of the 6D rotational representation, Rx and Ry, and minimize the predicted rotation from ground truth as formed in [Hempel et al. 2022], see Appendix G for details on $AngleLoss(R, \hat{R})$. We found this to produce better continuity of smooth rotations over time when assessing our results compared to using an $L1$ or $L2$ distance metric on the components that make up our rotation representation. We see further improvements by integrating with a normalization loss on the components.

$$NormLoss(\hat{R}) = |1.0 - ||\hat{R}x||_2| + |1.0 - ||\hat{R}y||_2| \tag{17}$$

$$6DLoss(R, \hat{R}) = AngleLoss(R, \hat{R}) + NormLoss(\hat{R}) \tag{18}$$

$$RotationalLoss(R, \hat{R}, W) = \frac{1}{\mathcal{J}} \sum_{j=1}^{\mathcal{J}} W_j \cdot 6DLoss(R_j, \hat{R}_j) \tag{19}$$

When assessing results in training, we split our weight evaluation equally between rotational and positional losses.

5.3 Pose Weighting

In developing our method, we experimented with many different means of weighting individual joints that comprise the pose when assessing losses. There are many potential factors that influence the decision, but we sought to find the means to enable our method to be more rapidly applied across different skeleton topologies, with minimal tuning by the end user.

Our weighting method works on the basis of two assumptions, in local space error on joints that are further down the hierarchy has an oversized impact on accuracy of those descendants further along; in model space, perceptual error becomes more noticeable the further from the root of the hierarchy a joint is. We thus produce two sets of weightings, per joint, one in local space, the other the weights to apply on the model space poses. Our mechanism begins with a simple voting scheme, each time we find a joint with a parent we recursively apply a vote to its parent and so on, until we reach the bottom of the hierarchy that forms the skeleton. At this point, we now have our local space joint weights, we then simply invert the relationship to produce their model space equivalents, those joints with the fewest children, becoming the highest weighted.

In practice, this very simple approach reduced the need for us to continue manually tuning joint weights, we see increased overall accuracy in our a results and we more easily enable support for other skeletons without explicit changes to the method.

5.4 Consecutive Poses in Autoencoder Training

In training our autoencoder, we train in pairs of temporally consecutive poses, as in [Holden et al. 2020]. We also similarly implement an additional loss for minimizing the difference between local and model space velocities in the training data and predictions.

$$VelocityLoss(P, \hat{P}) = \frac{1}{J} \sum_{j=1}^{\mathcal{J}} \left\| \frac{(P_1 - P_0)}{\delta t} - \frac{(\hat{P}_1 - \hat{P}_0)}{\delta t} \right\|_1 \tag{20}$$

5.5 Contact Label Prediction and Velocity Losses in Autoencoder Training

We note that our method supports a user defined number of arbitrary labels, or *tags*, in particular, commonly used tagging for foot or hand contacts. As we mention in section 4.2.1, these tags T are provided as input during training, and a sigmoid activation is applied to produce predicted probabilities $\hat{T}$, we then assess using a binary cross entropy loss $BCE(T, \hat{T})$.

As we highlight in section 5.4, we train our autoencoder in pairs of consecutive poses to evaluate an overall velocity consistency of our predictions. We further conditionally apply a loss on the predicted velocity $\hat{V}$, after forward kinematics, of joints deemed "in-contact" during phases where manual or automated tagging $C \in 0, 1$ indicates a given set of J joints should be in contact and thus, the difference minimized in relative global velocity.

$$ContactLoss(V, \hat{V}, C) = \sum_{j=1}^{\mathcal{J}} C_j \cdot ||V - \hat{V}||_2 \tag{21}$$

5.6 Smoother Latent Regularization

Where N is the number of Lipschitz layers in the encoder, we calculate the cumulative product of each layer's Lipschitz bounds $softplus(c_i)$, as per [Liu et al. 2022]. We then combine with $L1$ and $L2$ metrics, and velocity minimization losses (weighted λ_1, λ_2 and λ_3, respectively), as per [Holden et al. 2020] (see also section 5.4), on the latent values z produced per batch $\mathcal{B}$ to produce a final regularization loss.

$$L1Loss(z) = \lambda_1 \cdot ||z||_1 \tag{22}$$

$$L2Loss(z) = \lambda_2 \cdot |z||_2^2 \tag{23}$$

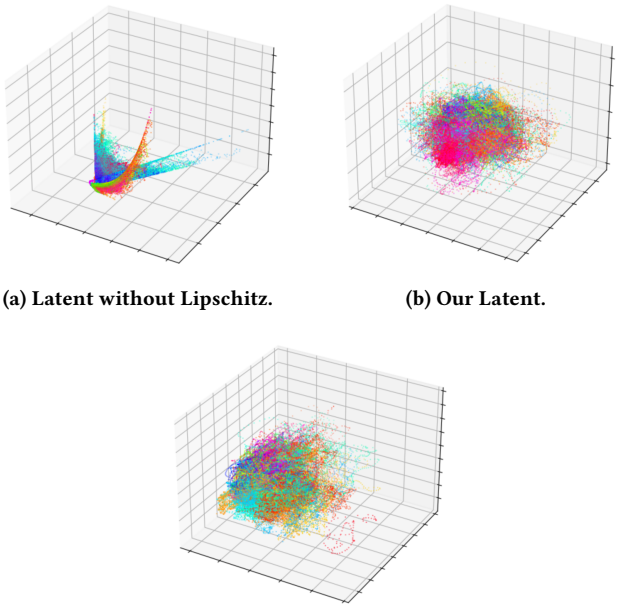
$$LatentVelocityLoss(z, \delta t) = \lambda_3 \cdot \left\| \frac{z_1 - z_0}{\delta t} \right\|_1 \tag{24}$$

$$LipschitzLoss = \alpha \cdot \prod_{i=1}^{N} softplus(c_i) \tag{25}$$

We found beyond a certain value, where minimizing the weights of the *NN* is a primary goal, there was little perceptible impact on our results by further optimization of the weighting α and settled on a value of $\alpha = 0.001$ across all experiments. For our $L1$, $L2$ and latent velocity losses, we used weights $\lambda_1 = 0.1$, $\lambda_2 = 0.1$ and $\lambda_3 = 0.01$. We observe a better overall distribution of samples in the latent for a given training set (see figure 4), which is conducive to improved results in our *stepper* implementation, when complemented with our usage of the *SMoE* architecture. Without the addition of the Lipschitz term, we see substantial banding and regular unpredictable jumps in the progression of a poses relative to another in temporal terms in the latent manifold. Intuitively, we see a relationship between samples of the motion matching features correlate well. By making this addition, we find a substantial benefit in the ability for our method to predict integration along the latent, and further, to correct drift in predictions over time.

5.7 SMoE and Large Batch Sizes

As noted in section 5.1, across all our models, we observed behavior contrary to that seen in applying *SMoE* to large language models[Zoph et al. 2022], whereby, liberally increasing batch sizes substantially improved the quality of our results. Consequently, for all models, we trained batch sizes of 12288, with 100 batches for each of 300 virtual epochs. We decided on these numbers as we observed the limiting factor in returns being a trade off of time available to train and memory available on GPU for such large batch sizes. We also observed a small increase in training time from our usage of *SMoE* compared to less accurate implementations comprised of *FFNs*.



(a) Latent without Lipschitz. **(b) Our Latent.**

(c) Motion matching features.

Figure 4: Latent and motion matching feature PCA projections in 3D.

5.8 Robustness to Noise

As noted in section 7.1 of [Holden et al. 2020], one of the primary functions of the *projector* model, is to correct *drift* of predicted latents from the *stepper* back on to the valid latent manifold. Whilst the *stepper* provides great utility over a shorter window of time, its overall tendency to gradually move out of validity limits the ability to further stagger *projector* calls or to enable concurrent playback of decoded predictions.

In principle, this addition is made to ensure that we compensate for the out of sample results from the *projector*, and never train the *stepper* on ground truth that it has no guarantee of encountering at inference. To account for noise in input, we propose a simple method of applying noise during training. This induces an ability, combined with our Lipschitz latent, for our method to better align the *stepper* output for a greater period of time. We follow the same autoregressive process of training with a simple addition. To simulate initially noisy data for the first step with samples z_0, M_0 of each batch $\mathcal{B}$, we apply gaussian noise using the standard deviation of the latent and motion matching features, $\sigma(z)$ and $\sigma(M)$. We use the standard deviation to give some meaningful reference

unit without introducing a dependency on calculating the variance of error returned post-*projector* training, which would necessitate removing the ability to train the *stepper* and *projector* concurrently.

$$z^n = \mathcal{N}(0,1) \cdot \sigma(z) \cdot \lambda_{scale} \tag{26}$$

$$M^n = \mathcal{N}(0,1) \cdot \sigma(M) \cdot \lambda_{scale} \tag{27}$$

$$z_{initial} = z_0 + z^n \tag{28}$$

$$M_{initial} = M_0 + M^n \tag{29}$$

$$\delta\hat{z}_1, \delta\hat{M}_1 = Stepper(z_{initial}, M_{inital}) \tag{30}$$

We then continue to train our *stepper* without further alterations, after making the first predictions using $z_{initial}$ and $M_{initial}$. After some experimentation, we found a larger value of $\lambda_{scale} = 0.15$ provided us with robust results, see section 6.2.1.

6 EVALUATION

In this section, we present evaluation of the performance of our method in multiple areas. All model inference is handled via our internal technology that enables deploying pre-trained models within video games on consumer hardware and executes experts correctly.

In the development of our method, we compared exporting models from *PyTorch* for execution via *ONNX Runtime*[developers 2021], however, the current implementation across platforms is unable to successfully conditionally reduce the work of evaluating zero weighted experts, instead forming several add-multiply operations for each unused expert. As a consequence, our method requires some degree of specialized implementation for on-device inference to see the theoretical gains in compute complexity reduction.

For examples referenced in this section, we compared performance on a common locomotion task, with animation data sampled from motion captured performed by our colleagues, covering a range of gaits, speeds and orientations. We also further augmented the data by mirroring suitable assets to increase coverage to produce in the region of 42 minutes of sample data. We also sample a motion matching database of multiple features, see Appendix D for details.

6.1 Performance

We evaluated single threaded performance with multiple runs on the same hardware, we present the numbers below for an Intel i9-11900K 3.5Ghz. We trained comparable *FFNs*, as in Appendix E, with the same training data, for all components. Below, we compare both execution cost per inference step, in microseconds (μs), and memory required for weights, in megabytes (Mb). The lowest execution or memory cost is highlighted in **bold**.

Component	FFN(μs)	Ours(μs)	FFN(Mb)	Ours(Mb)
Decoder	60.94	**19.94**	3.86	**3.63**
Stepper	54.02	**1.85**	3.30	**1.45**
Projector	98.57	**3.58**	5.24	**4.87**
Total	213.53	**25.37**	12.40	**9.95**

In expressing our models using *SMoE*, as stated, we are able to substantially reduce the theoretical complexity count of operations for inference of each model compared to equivalent *FFNs*. Further reductions in execution and memory costs could be introduced by quantizing from single-precision floating-point to half.

As seen in *LMM*[Holden et al. 2020], even an *FFN* implementation brings about a substantial reduction in memory usage compared to even compressed animation. Without taking into account additional components used by a motion matching implementation, we see a significantly higher compression ratio using our models versus in-engine compression formats, 9.95Mb versus 36.64Mb. When we factor in the omission of spatial representations of the motion matching features and their database that advantage only grows further.

We implement an optimized reference motion matching system, using *nanoflann*(Blanco and Rai [2014]; Suju and Jose [2017]) to accelerate our spatial searches for samples, with our compressed animation data and storing our motion matching features database as half-precision floats. This allows us to generate new databases and test prior to training models for final usage, but also gives us an important point of comparison. Several components are included and require additional storage, as specified below.

Component	Size(Mb)
Compressed Animation Data	36.64
KD-Tree	3.68
Motion Matching Features DB	12.05
Total	52.37

We compare all three systems below and, again, highlight the lowest execution or memory cost in **bold**.

System	Total CPU Time(μs)	Total Memory (Mb)
FFN	213.53	12.40
Our Method	**25.37**	**9.95**
Motion Matching	78.20	52.37

As we demonstrate, our method provides a solution that does not require the user compromise on performance nor storage, and even performs favourably compared to the time decompressing a single frame in high-quality compression libraries for animation of around 20-30μs[Frechette and contributors 2017] alone.

6.1.1 Further Performance Notes. We give total system costs, but we highlight a number of considerations when evaluating. By default, for all methods, we only trigger the projector or sample search every 100ms of game updates, so this cost can be further amortized over a number of frames. Though we note, unlike both motion matching and *LMM* our projector does not represent the highest cost of our system at runtime.

Furthermore, depending on the strategy employed for decoding poses, one can further optimize characters using our method as per section 6.3 and stagger the frequency of individual pose updates according to an in-game level-of-detail or *"LODing"* strategy.

6.2 Pose Accuracy

Comparing *FFNs* trained with our losses using arrangements as in *LMM* (see Appendix E), our method sees notable improvements in

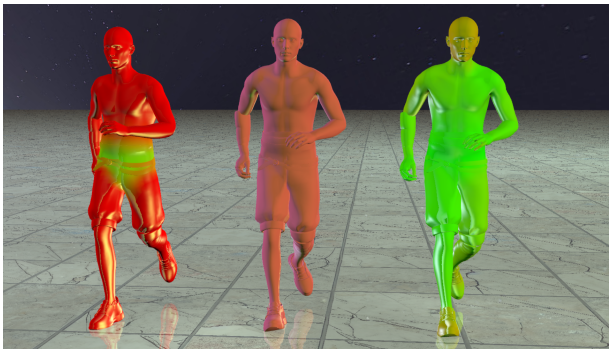


Figure 5: Comparing pose outputs from left to right: *FFN*, ground truth, our method.

pose reconstruction, see figure 5. We color vertices in predictions based on the error of the bones they are skinned to, with a linear green-to-red heatmap, where green sees error of 0.004m or below and red of 0.02m or above.

By increasing the relative size of the *FFNs*, we can see an improvement in accuracy that matches ours, at the expense of additional compute complexity and memory size for model weights, furthermore, when tuning batch sizes and steps required to achieve this accuracy, the training time for the pose autoencoder *FFN* increased to 5 days. However, notably, increasing the size of the models does not leverage the benefit of the Lipschitz regularizing effect on the latent and subsequent predictions output from other models in the system, it also only further increases the already highlighted disadvantage in performance.

6.2.1 Retrieval and Predictive Accuracy. Our implementation compares favourably in error when retrieving samples of the latents produced from training pose autoencoders, however, we see a significant reduction in the error when retrieving the source motion matching features. Our method provides the means to safely predict latents and motion matching features over a longer time period, this allows us to implement decision making similar to traditional motion matching, whereby, we can avoid switching samples that appear to be less desirable than those already playing in the system at runtime, thereby giving us the opportunity to more smoothly play back our resultant choices.

When we introduce noise into our queries and predict future updates via the stepper we see further improvements in the robustness of our results in comparison, with a more pronounced impact on the relative error in the latent representation.

We find the smoothing impact on the latent from the Lipschitz constrained encoder produces long and consistent predictions over an extended period of time. We also tend to see a linear increase in error over time of these predictions. Some examples are presented in figure 8, which entails a search of the motion matching features for the frame at time 0 for the projector and then integration via the stepper a number of samples ahead, prior to decoding. We compare the results of predicting from 40000 randomly drawn samples. We randomly apply Gaussian sampled noise on the first query, then sample autoregressively, as in training. Note, our smoother latent is more predictable for a longer period of time than the results

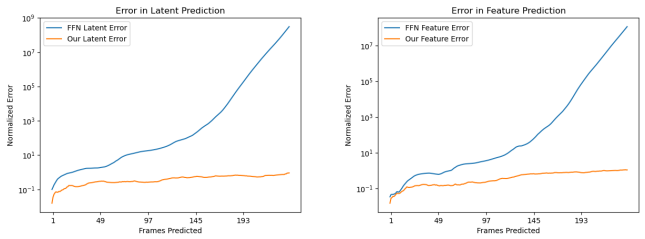


Figure 6: Mean error across autoregressive predictions.

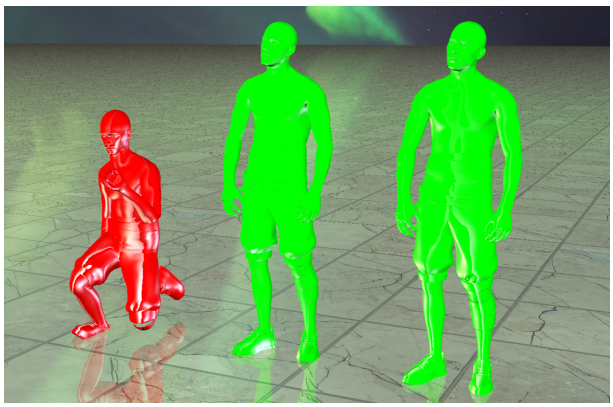
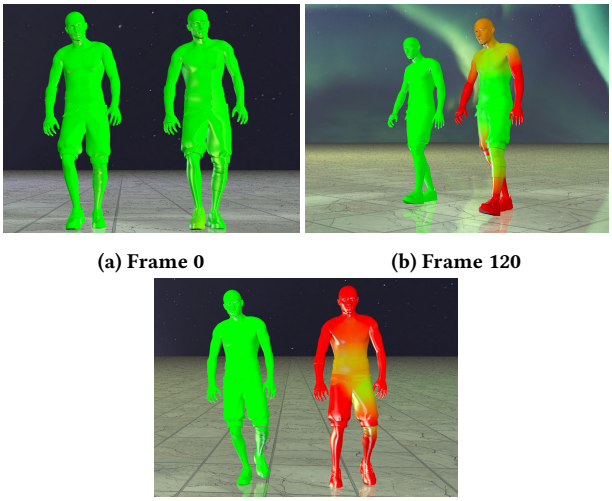


Figure 7: Comparing one second of forward prediction, left, *FFN*, middle, ground truth, right, ours. We see that without the Lipschitz regularization, pose predictions rapidly drift away from the valid manifold.

from an *FFN* with no Lipschitz bounds. There is some contrast in the duration at which our predictions remain coherent, when we compare to *FFN* output without Lipschitz regularization, the decoded pose does not correlate with any plausible mapping from the latent manifold to a real pose, as in our method, see figure 7 Figure 8 shows some examples of the typical quality of predictions we see in our method. At frame 0, we see a close relationship with ground truth, typical mean accuracy across the pose in the region of 0.003m. At frame 120, we still see a relationship with ground truth, but it is possible to discern differences when comparing directly, typical mean accuracy across the pose in the region of 0.02m or below. At frame 240, we see a coherent pose, but there are now significant visible differences, typical mean accuracy in the region of 0.03m. Whilst we are at some points able to predict some 240 frames or more from an initial sample, we find sufficient number of outlier cases where we produce an unusable pose or sample that does not align with valid posing on the latent manifold. As a consequence, we choose to limit all predictions to a maximum of 120 frames, or 2 seconds.

6.3 Smooth Pose Blending in Latent Space

In replicating the behavior of motion matching, additional expenses are added to a runtime implementation due to the necessity of providing the means to blend poses as a character transitions between



(a) Frame 0 (b) Frame 120

(c) Frame 240

Figure 8: Comparing accuracy of pose retrieval and predictions over time, character on the left, ground truth, on the right, our method.

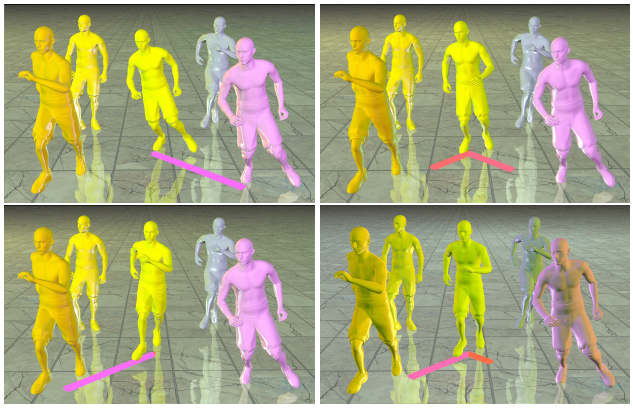


Figure 9: Our method displays plausible character posing when blending between disconnected samples in the latent pose manifold. Clockwise from top-left, we see incremental blending from the ground-truth sample on the front-right over to the front-left. The character in the middle, sampling the blend latents, produces smooth, plausible, posing throughout the transition.

constituent animations that make the complete scene. In order to address the potential performance penalty of cross-blending or decoding multiple poses[Holden et al. 2020], *inertialization*[Bollo 2018] has been leveraged in other systems to reduce costs to a fixed transition.

Whilst effective, *inertialization* implementations vary from complex polynomial solutions to user-tuned spring or velocity based solutions. In pratice, these can be challenging to predictably tune. In seeking to reduce the complexity and increase the robustness of our method, we experimented with the linearizing properties of our latent regularization (as in [Liu et al. 2022]) to support an intuitive alternative to traditional means of linear blending between poses, as in Figure 9.

Our method for interpolation of pose P prior to decoding becomes a simple matter of weighted linear interpolation, we take N active samples X and their respective weights W to generate a single sample Y that can then be decoded and presented in-game.

$$WeightSum = \sum_{i=1}^{N} W_i \tag{31}$$

$$Y = \frac{1}{WeightSum} \sum_{i=1}^{N} X_i W_i \tag{32}$$

$$P = DECODE(Y) \tag{33}$$

Whilst we see a great deal of success in using this method to sample a final pose for presentation in-game, such usage is not a prerequisite for a useful implementation of our method. We note in Appendix C, that without Lipschitz regularization, this method is not possible.

As in [Liu et al. 2022], we can apply extrapolation to generate new data, however, such usage rapidly produces implausible poses, we believe there is interesting potential for investigating further in the future, see Appendix B for examples.

7 CONCLUSION

We demonstrate several advantageous developments on prior state-of-the-art, to enable the desirable behaviour of motion matching like animation control at a scale previously unachievable without further compromises. We note, that the superior improvements in memory reduction as were seen in *LMM*[Holden et al. 2020] no longer need to be evaluated as a trade off for execution costs when implementing a motion matching system by using our method. Indeed, we propose that our method instead unlocks both greater potential for larger animated character counts and execution on more limited contemporary video game platforms, without necessitating reductions in quality or difficult to implement performance strategies.

Whilst we chose to try to strike a balance between performance and reconstruction quality, we note that the complexity and memory usage reductions from our method leave a greater capacity for scaling up or down the size and number of experts within the models to achieve different desired quality goals.

We discuss potential avenues of further investigation in Appendix A.

ACKNOWLEDGMENTS

We thank our colleagues on the Production R&D team for their valuable feedback during development and participation in hours of motion capture in the depths of summer. We also thank our colleagues at Sharkmob for the use of their time and motion capture facilities in producing data for our research.

REFERENCES

Jose Luis Blanco and Pranjal Kumar Rai. 2014. nanoflann: a C++ header-only fork of FLANN, a library for Nearest Neighbor (NN) with KD-trees. https://github.com/jlblancoc/nanoflann.

David Bollo. 2018. Inertialization: High-Performance Animation Transitions in Gears of War. *GDC '18* (March 2018). https://www.youtube.com/watch?v=BYyv4KTegJI

Michael Büttner and Simon Clavet. 2015. Motion matching - the road to next gen animation. *Nucl.ai '2015* (July 2015).

Simon Clavet. 2016. Motion Matching and The Road to Next-Gen Animation. *GDC '16* (March 2016). https://www.gdcvault.com/play/1023280/Motion-Matching-and-The-Road

ONNX Runtime developers. 2021. ONNX Runtime. https://onnxruntime.ai/. Version: x.y.z.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *J. Mach. Learn. Res.* 23 (2022), 120:1–120:39. http://jmlr.org/papers/v23/21-0998.html

Nicholas Frechette and Animation Compression Library contributors. 2017. Animation Compression Library. https://github.com/nfrechette/acl.

Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J. Cree. 2020. Regularisation of neural networks by enforcing Lipschitz continuity. *Machine Learning* 110, 2 (Dec. 2020), 393–416. https://doi.org/10.1007/s10994-020-05929-w

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. arXiv:1502.01852 [cs.CV]

Thorsten Hempel, Ahmed A. Abdelrahman, and Ayoub Al-Hamadi. 2022. 6d Rotation Representation For Unconstrained Head Pose Estimation. In *2022 IEEE International Conference on Image Processing (ICIP)*. 2496–2500. https://doi.org/10.1109/ICIP46576.2022.9897219

Dan Hendrycks and Kevin Gimpel. 2023. Gaussian Error Linear Units (GELUs). arXiv:1606.08415 [cs.LG]

Daniel Holden, Oussama Kanoun, Maksym Perepichka, and Tiberiu Popa. 2020. Learned Motion Matching. *ACM Trans. Graph.* 39, 4, Article 1 (July 2020). https://doi.org/10.1145/3386569.3392440

Daniel Holden, Taku Komura, and Jun Saito. 2017. Phase-functioned neural networks for character control. *ACM Transactions on Graphics* 36, 4 (July 2017), 1–13. https://doi.org/10.1145/3072959.3073663

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.

Tobias Kleanthous. 2021. Making The Believable Horses of Red Dead Redemption II. *GDC '21* (July 2021). https://www.youtube.com/watch?v=8vtCqfFAjKQ

Yongjoon Lee, Kevin Wampler, Gilbert Bernstein, Jovan Popović, and Zoran Popović. 2010. Motion fields for interactive character locomotion. In *ACM SIGGRAPH Asia 2010 papers on - SIGGRAPH ASIA '10*. ACM Press. https://doi.org/10.1145/1882262.1866160

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2021. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=qrwe7XHTmYb

Peizhuo Li, Kfir Aberman, Zihan Zhang, Rana Hanocka, and Olga SorkineHornung. 2022. GANimator: Neural Motion Synthesis from a Single Sequence. *ACM Trans. Graph.* 41, 4, Article 138 (July 2022). https://doi.org/10.1145/3528223.3530157

Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel van de Panne. 2020. Character Controllers Using Motion VAEs. *ACM Trans. Graph.* 39, 4, Article 40 (July 2020). https://doi.org/10.1145/3386569.3392422

Hsueh-Ti Derek Liu, Francis Williams, Alec Jacobson, Sanja Fidler, and Or Litany. 2022. Learning Smooth Neural Functions via Lipschitz Regularization. *In Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings (SIGGRAPH '22 Conference Proceedings)*. https://doi.org/10.1145/3528233.3530713

Antoine Maiorca, Nathan Hubens, Sohaib Laraba, and Thierry Dutoit. 2022. Towards Lightweight Neural Animation: Exploration of Neural Network Pruning in Mixture of Experts-based Animation Models. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2022, Volume 1: GRAPP, Online Streaming, February 6-8, 2022*, A. Augusto de Sousa, Kurt Debattista, and Kadi Bouatouch (Eds.). SCITEPRESS, 286–293. https://doi.org/10.5220/0010908700003124

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. 2018. Spectral Normalization for Generative Adversarial Networks. arXiv:1802.05957 [cs.LG]

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Dario Pavllo, David Grangier, and Michael Auli. 2018. QuaterNet: A Quaternion-based Recurrent Model for Human Motion. arXiv:1805.06485 [cs.CV]

Nagaraj Raparthi, Eric Acosta, Alan Liu, and Tim McLaughlin. 2020. GPU-based Motion Matching for Crowds in the Unreal Engine. In *SIGGRAPH Asia 2020 Posters*. ACM. https://doi.org/10.1145/3415264.3425474

Sebastian Starke, Ian Mason, and Taku Komura. 2022. DeepPhase: periodic autoencoders for learning motion phase manifolds. *ACM Trans. Graph.* 41, 4, Article 136 (July 2022). https://doi.org/10.1145/3528223.3530178

Sebastian Starke, Yiwei Zhao, Taku Komura, and Kazi Zaman. 2020. Local motion phases for learning multi-contact character movements. *ACM Transactions on Graphics* 39, 4 (Aug. 2020). https://doi.org/10.1145/3386569.3392450

D Arul Suju and Hancy Jose. 2017. FLANN: Fast approximate nearest neighbour search algorithm for elucidating human-wildlife conflicts in forest areas. In *2017 Fourth International Conference on Signal Processing, Communication and Networking (ICSCN)*. 1–6. https://doi.org/10.1109/ICSCN.2017.8085676

Xingyu Xie, Pan Zhou, Huan Li, Zhouchen Lin, and Shuicheng Yan. 2023. Adan: Adaptive Nesterov Momentum Algorithm for Faster Optimizing Deep Models. arXiv:2208.06677 [cs.LG]

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2020. On the Continuity of Rotation Representations in Neural Networks. arXiv:1812.07035 [cs.LG]

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. 2022. ST-MoE: Designing Stable and Transferable Sparse Expert Models. arXiv:2202.08906 [cs.CL]

A FUTURE WORK

In this paper we show that the *SMoE* architecture unlocks compute complexity reducing behavior outwith the domain of large language models, and likely presents the opportunity to significantly benefit the capacity or complexity of *FFNs* in other domains beyond animation synthesis.

We also see that introducing Lipschitz bounds in motion synthesis allows for predictive benefits and a smoother interpretable latent manifold for animation representation. In this paper, we do not rely on fully developing generality in our method, however, we see interesting avenues in how this method may apply to generate new data, particularly by means of extrapolation or interpolation of encodings of samples that are not seen in the training set.

Furthermore, we also remain curious as to how one may leave behind the approach here and in [Holden et al. 2020] of using multiple models, as the overall performance profile of our method potentially unlocks a singular end-to-end trained variant in the future.

Lastly, whilst motion matching provides flexible means to define the features used to define a match, we find there is still an undesirable requirement for engineering feature selection and weights, beyond the specific control values that define the task, such as event times or trajectory goals for root motion. This remains true, even with the improvements introduced by utilising ideas in [Starke et al. 2022] in this paper. We are keen to see how the onerous feedback loop of this task could be improved to allow end users of such a system more time to focus on the quality of the input and output.

B LATENT POSE EXTRAPOLATION

Whilst we see success in blending around valid samples on the latent manifold. A small amount of isolated extrapolation is possible, but blends poorly. Some examples are highlighted here.

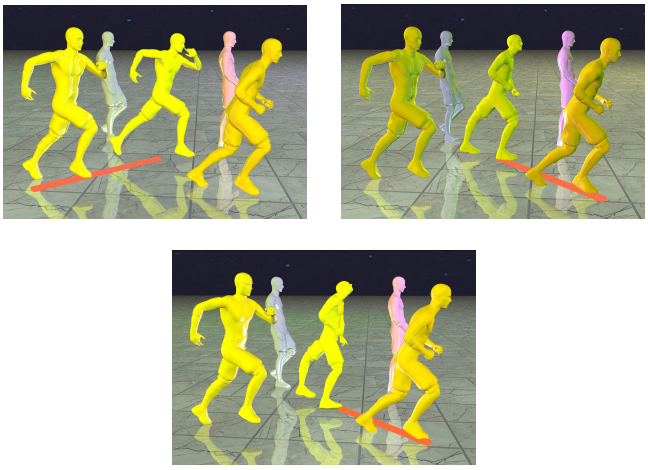


Figure 10: Extrapolation succeeds in a small margin, but rapidly produces implausible and undesirable posing.

C LATENT BLENDING WITHOUT LIPSCHITZ REGULARIZATION

Without enforcing the Lipschitz bounds on an *NN*, blending of the latents does not become a definable task. In our experiments with our *FFN* implementation, we are unable to move smoothly along the latent manifold and instead produce small changes around an apparently averaged pose that do not correlate with the ground truth input poses.

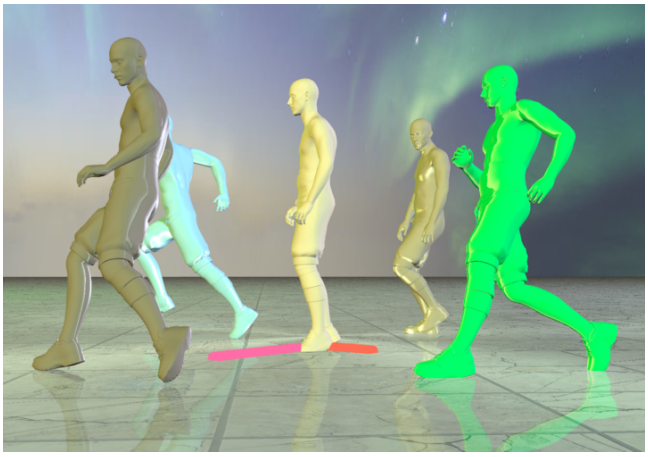


Figure 11: Without enforcing Lipschitz regularization, linear blending is not possible on the latent manifold.

D ANIMATION DATA AND MOTION MATCHING FEATURES IN EVALUATION

We defer to [Clavet 2016] for fuller details on the base functionality and implementation details of motion matching, but state here the features that were extracted from animation data to train the locomotion example in our paper. We note that the method in this paper was also validated on quadrupeds and other styles of bipedal locomotion and actions with varying sizes of motion matching features.

We demonstrate on animation database of human data as below.

Frames	Sample Rate	Joints Per Pose	Uncompressed(Mb)
154130	$60hz$	96	398.63

We provided a motion matching feature vector per-frame of animation, with the following components.

Table 1: Locomotion Motion Matching Features

Feature	Length
Left Foot Linear XYZ Velocities	3
Right Foot Linear XYZ Velocities	3
Pelvis Linear XYZ Velocities	3
Left Foot Position	3
Right Foot Position	3
Root Position XZ at Time -0.1s	2
Root Direction XZ at Time -0.1s	2
Root Position XZ at Time 0.3s	2
Root Direction XZ at Time 0.3s	2
Root Position XZ at Time 0.6s	2
Root Direction XZ at Time 0.6s	2
Root Position XZ at Time 1.0s	2
Root Direction XZ at Time 1.0s	2
Phase Embedding	10
Total	41

When implementing our reference motion matching database, we store these values as half-precision floating point, but when used for training, we take the full single-precision values.

E LAYER DETAILS FOR MODELS IN PAPER EXPERIMENTS

The following tables give further information regarding arrangement of particular models referenced in our experiments.

Table 2: Layers in our Pose Encoder

Layer	Units
Lipschitz + GELU	512
Lipschitz + GELU	512
Lipschitz + GELU	512
Lipschitz	32

Table 3: Layers in our Pose Decoder

Layer	Units/Count
Linear + Softmax	128
Experts	128
Linear + GELU	256
Linear	872

Table 4: Layers in a single Pose Decoder Expert

Layer	Units
Linear + GELU	32
Linear + GELU	32
Linear + GELU	32
Linear	32

Table 5: Layers in FFN Pose Encoder

Layer	Units
Linear + ELU	512
Linear + ELU	512
Linear + ELU	512
Linear + ELU	512
Linear	32

Table 6: Layers in FFN Pose Decoder

Layer	Units
Linear + ReLU	512
Linear + ReLU	512
Linear + ReLU	512
Linear + ReLU	512
Linear	872

Table 7: Layers in SMoE Stepper

Layer	Units/Count
Linear + Softmax	64
Experts	64

Table 8: Layers in Single Stepper Expert

Layer	Units/Count
Linear + GELU	32
Linear + GELU	32
Linear	73

Table 9: Layers in FFN Stepper

Layer	Units/Count
Linear + ReLU	512
Linear + ReLU	512
Linear + ReLU	512
Linear + ReLU	512
Linear	73

Table 10: Layers in SMoE Projector

Layer	Units/Count
Linear + Softmax	64
Experts	64

Table 11: Layers in Single Projector Expert

Layer	Units/Count
Linear + GELU	64
Linear + GELU	64
Linear + GELU	64
Linear + GELU	64
Linear	73

Table 12: Layers in FFN Projector

Layer	Units/Count
Linear + ReLU	512
Linear + ReLU	512
Linear + ReLU	512
Linear + ReLU	512
Linear + ReLU	512
Linear	73

F PROJECTOR IMPLEMENTATION DETAILS

Our projector is implemented as in [Holden et al. 2020], aside from the addition of *SMoE* architecture. The process is as described, we sample a scale n^{σ} from the uniform distribution, a value *noise* from the unit Gaussian to generate noise on the motion matching features x. We then apply feature weights w and search for a result in our KD-tree of ground truth, this is then used create an index to select a motion matching feature vector and latent from our training data X and Z. We then invoke a prediction from the *projector* using the

weighted noisy motion matching features.

$$n^{\sigma} = \mathcal{U}(0,1) \tag{34}$$

$$noise = \mathcal{N}(0,1) \tag{35}$$

$$y = x + n^{\sigma} \cdot noise \tag{36}$$

$$y_{weighted} = y * \sqrt{w} \tag{37}$$

$$index = KDTree(y) \tag{38}$$

$$y_{actual}, z_{actual} = X[index], Z[index] \tag{39}$$

$$\hat{y}, \hat{z} = Projector(y_{weighted}) \tag{40}$$

$$\tag{41}$$

Our losses are also similar, with the weights $\lambda_1 = 1.0$, $\lambda_2 = 10.0$ and $\lambda_3 = 0.3$. We encourage accuracy of the predictions and maintaining the distance of the query observed in the source KD-tree query.

$$YLoss = \lambda_1 \cdot ||y_{actual} - \hat{y}||_1 \tag{42}$$

$$ZLoss = \lambda_2 \cdot ||z_{actual} - \hat{z}||_1 \tag{43}$$

$$QueryDistLoss = \lambda_3 \cdot \left\| ||y_{actual} - y||_2^2 - ||\hat{y} - y||_2^2 \right\|_1 \tag{44}$$

$$\tag{45}$$

For $KDTree(y_{weighted})$, we use FAISS[Johnson et al. 2019] to search a KD-tree of pre-weighted ground truth to accelerate training times.

G GEODESIC LOSS IMPLEMENTATION

We clarify the implementation of the geodesic loss used in our pose losses here. We take the same approach as in [Hempel et al. 2022], forming rotation matrices from our ground truth and predicted 6D rotations, M and $\hat{M}$, respectively. In the interest of making compatible with balancing with our normalization and positional losses, we normalize the result by 2π.

$$GeodesicAngleDelta(M, \hat{M}) = \arccos\left(\frac{tr(M\hat{M}) - 1}{2}\right) \tag{46}$$

$$AngleLoss(R, \hat{R}) = \frac{GeodesicAngleDelta(Matrix(R), Matrix(\hat{R}))}{2\pi} \tag{47}$$