Fitness Aware Human Motion Generation with Fine-Tuning

Kiril Bikov[*]
University of Cambridge, UK

Shiye Su[*]
University of Cambridge, UK

Deepro Choudhury[*]
University of Cambridge, UK

Zhilin Guo
University of Cambridge, UK

Weihao Xia
University of Cambridge, UK

Mehmet Salih Celiktenyildiz
Bilkent University, Turkiye

Chenliang Zhou
University of Cambridge, UK

Param Hanji
University of Cambridge, UK

Cengiz Oztireli[†]
University of Cambridge, UK

Abstract

Diffusion models have recently gained considerable attention in 3D human motion generation due to their ability to handle complex human movements. However, existing models fail to incorporate the nuances presented by individual physical fitness levels. Therefore, we address this gap by integrating Functional Movement Screen (FMS) scores into diffusion models through fine-tuning, enabling the generation of fitness-aware motions. This approach transforms FMS data into HumanML3D format, optimises a base diffusion model, and introduces conditioning based on FMS scores. As a result, our fine-tuned model is capable of generating motions tailored to individual fitness levels and shows significant improvements in motion generation fidelity. Producing synthetic human motions conditioned on fitness levels is a novel approach that can be highly beneficial for various fields such as healthcare, sports, and entertainment.

Keywords: Diffusion; Human Motion; Fine-Tuning

1 Introduction

Human motion generation has been focusing on type-based motion, conditioning on pre-defined action classes (e.g. “jumping”, “throwing a ball”) or mechanical description of actions (e.g. “running in a circle with arms raised”). While this level of specificity suffices for some tasks, other applications are crucially dependent on the *quality of the movement* that reflects the fitness level of the subject. Furthermore, induced by the absence of physical constraints during training, human motion diffusion models are highly artefact-prone.

Therefore, we take the first steps in making 3D motion diffusion models *fitness-aware*. With our proposed fine-tuning approach, the models can generate 3D motions consistent with a given level of athleticism or injury severity. To achieve this goal, we incorporate Functional Movement Screen (FMS) score Xing et al. (2022) into existing human motion diffusion models Tevet et al. (2023) to generate motions that are graded by quality. FMS is used to identify movement pattern asymmetries

[*] Equal contribution

[†] Corresponding author

38th Workshop on Fine-Tuning in Machine Learning (NeurIPS 2024).

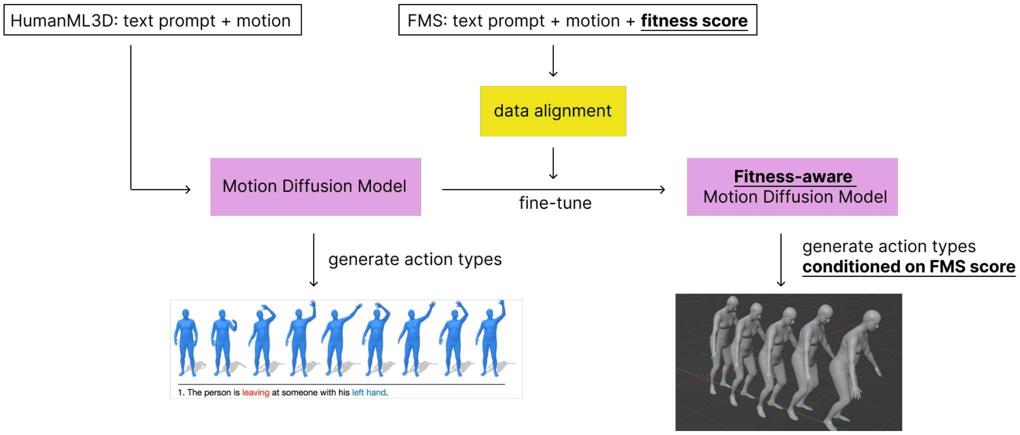

HumanML3D: text prompt + motion
FMS: text prompt + motion + fitness score
data alignment
Motion Diffusion Model
fine-tune
Fitness-aware Motion Diffusion Model
generate action types
generate action types conditioned on FMS score
1. The person is leaving at someone with his left hand.

Figure 1: **Method Overview**. Firstly, an MDM Tevet et al. (2023) is trained on HumanML3D with text-motion pairs Guo et al. (2022b). Then, model is fine-tuned conditioned on fitness level using FMS data, where the text-motion pairs are enriched with an FMS grade label.

or deficiencies, providing a simple measure to rate and rank basic movement patterns essential for daily physical activity. By conditioning on the score, we aim to create a more holistic and accurate human motion representation by taking into account individuals' physical capabilities. To the best of our knowledge, our work is the first attempt to close this research gap. Our research could facilitate the development of personalised and accurate human movement models, essential to applications including healthcare, sports, and entertainment.

In this work, we present Fitness-Aware 3D Human Motion Generation. We incorporate FMS score Xing et al. (2022) as condition for existing human motion generation methods Tevet et al. (2023), enabling the production of 3D human motions conditioned on fitness quality.

To summarize, our main contributions are:

1. A fitness-aware 3D human motion generation method, a previously unseen class of human motion generation methods.
2. A pipeline to transform FMS data into format compatible with HumanML3D and hyper-parameter optimization that improves specificity and multimodality of motion diffusion models.
3. Comprehensive evaluation, including training of a customized FMS classifier to validate that our model indeed generates movements that are distinct according to FMS value.

2 Related Work

2.1 Human Motion Generation

Human motion generation is an essential area of study in computer vision and animation which has various applications such as virtual reality Lee et al. (2023a); Winkler et al. (2022); Ye et al. (2022), gaming Holden et al. (2017); Starke et al. (2022), human behavior analysis Chong et al. (2020); Guo et al. (2023); Liu et al. (2022); Yang et al. (2023); Zhang et al. (2023d), and robotics Christen et al. (2023); Li et al. (2023b); Wan et al. (2022); Yamane et al. (2013). The generated motion can be conditioned on a diverse range of multi-modal inputs, such as action labels Guo et al. (2020); Lee et al. (2023b); Petrovich et al. (2021); Xu et al. (2023), textual descriptions Ahuja and Morency (2019); Dabral et al. (2023); Guo et al. (2022b,a); Kong et al. (2023); Petrovich et al. (2022), incomplete pose sequences Harvey et al. (2020); Wan et al. (2022); Duan et al. (2021), control signals Holden et al. (2017); Liao et al. (2022); Pi et al. (2023); Shi et al. (2023); Zhang et al. (2023b), and auditory inputs like music Lee et al. (2019); Li et al. (2022, 2021); Pang et al. (2023).

Yet to our knowledge, no prior work has attempted the generation of human motion conditioned on fitness level.

2.2 Diffusion Models and Diverse Generative Techniques

Diffusion models are a type of generative model inspired by the thermodynamic stochastic diffusion process, where a neural network learns to gradually recover unseen images from a noisy distribution Song and Ermon (2020); Zhang et al. (2022). Diffusion generative models have shown impressive results across a range of fields Chou et al. (2023); Liu et al. (2023); Müller et al. (2023); Rombach et al. (2022); Yu et al. (2023); Po et al. (2023); Long et al. (2023). They have been successfully applied in various areas such as image synthesis Rombach et al. (2022); Ramesh et al. (2022); Saharia et al. (2022); Sinha et al. (2021); Vahdat et al. (2021), video creation Luo et al. (2023); Yang et al. (2022), adversarial attacks Zhuang et al. (2023); Nie et al. (2022), motion prediction Wei et al. (2023); Chen et al. (2023a), music-to-dance synthesis Li et al. (2023a); Tseng et al. (2023), and text-to-motion conversion Chen et al. (2023b); Ren et al. (2023); Tevet et al. (2022); Yuan et al. (2022); Zhang et al. (2022).

Recently, diffusion models have been leveraged frequently for generating human motion Zhang et al. (2022); Tevet et al. (2023); Chen et al. (2023b); Zhang et al. (2023c); Wang et al. (2023). Among these methods, MotionDiffuse Zhang et al. (2022) marks the first use of diffusion models in text-to-motion applications. MDM Tevet et al. (2023) processes motion diffusion by working with raw motion data to learn how it relates to the specified input conditions. MLD Chen et al. (2023b) employs a VAE model to encode motion and diffuse it within a latent space. ReMoDiffuse Zhang et al. (2023c) retrieves motion related to textual input to aid in motion generation. Fg-T2M Wang et al. (2023) applies a fine-grained approach to extract both local and overall semantic linguistic features. Recently, EMDM Zhou et al. (2023) leverages kinematics-based methods and textual descriptions to synthesize human motion through a diffusion model, capturing complex denoising distributions in a few sampling steps. Moreover, PriorMDM Shafir et al. (2023) introduced a fine-tuning approach to enhance the MDM for generating human interactions from text-motion data.

In addition to diffusion models, which utilize continuous motion representation, discrete token-based approaches using Vector Quantized Variational Autoencoders (VQ-VAEs) have also achieved promising outcomes. Notable examples are TM2T Guo et al. (2022a), T2M-GPT Zhang et al. (2023a), MotionGPT Jiang et al. (2023). Specifically, MoMask Guo et al. (2024a) introduces a masked modeling framework that incorporates a hierarchical quantization scheme and bidirectional transformers to enhance the quality and detail of generated motions. Recently, both CoMo Huang et al. (2024) and FineMoGen Zhang et al. (2024) have utilized language models to generate edit texts for part-based body editing, showing promising outcomes, while CoMo's autoregressive method allows for the generation of motion sequences of arbitrary length. TLControl Wan et al. (2023) uses a disentangled latent space to produce a wide range of human motions, enabling high-fidelity synthesis that aligns with both language descriptions and specific trajectories. Guo et al. Guo et al. (2024b) use the latent space of pre-trained motion models to improve how they extract and integrate motion content and style.

However, these models do not account for the fitness level of the generated 3D human motion, thus coming short of meaningfully differentiating the physique of generated humans.

2.3 Fitness in Human Movement.

Several studies were conducted on classifying movement fitness from human movement Spilz and Munz (2023); Remedios et al. (2020); Chen et al. (2022), and some of these approaches aim to derive Functional Movement Screen score from exercises. A previous study Spilz and Munz (2023) uses a combination of visual neural networks to assign FMS scores to exercise repetitions. Another research paper Remedios et al. (2020) uses an unsupervised approach to identify clustered "phenotypes" from movement to distinguish between two FMS groups. Lin et al. (2023) integrates an I3D network with an attention mechanism and a multilayer perceptron to assess Functional Movement Screening (FMS).

However, these studies only tackles the the classification of movement fitness, and fail to address the generation of 3D human motion conditioned on this fitness level.

3 Method

3.1 Human Motion Diffusion Model.

We formulate the problem of conditional 3D human motion generation as generating a sequence of N frames, $x^{1:N} \equiv \{x^{(1)}, ..., x^{(N)}\}$ given some condition c Tevet et al. (2023). Each frame $x^i \in \mathbb{R}^{J \times D}$, where J is the number of joints and D is the dimension of the joint representation, describes a human pose by the positions of each joint in the model skeleton. For our purposes, we set $D = 3$, representing the x, y, and z Cartesian coordinates, while c serves as a natural language prompt describing the action to be generated. We are interested in the distribution $p(x^{1:N}|c)$, which we model through a reverse diffusion process.

The diffusion forward process, denoted as q, sequentially introduces noise to the input via a Gaussian distribution following the Markov process. The hyperparameter α_t governs the denoising schedule.

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

Motion synthesis is achieved by the denoising (reverse) process, which successively "removes" noise to recover the human motion $x_0^{1:N}$ conditioned on c. The process is modelled by a stack of encoder-only transformers Vaswani et al. (2017). Each frame is projected onto the transformers' hidden dimensions and added to a standard positional embedding. At the first timestep, this is concatenated with the vector z_{tk}, which is the sum of the representations of the condition c and the denoising timestep t.

For text-to-motion, the representation of c is obtained through the linear projection of pre-trained CLIP embeddings Radford et al. (2021) of the text prompt. Meanwhile, t is encoded using a distinct linear projection. The final result of this iterative process yields a prediction $\hat{x}_0^{1:N} = G(x_t, t, c)$, which is trained to draw from the distribution $p(x^{1:N}|c)$ by optimising with the simple L2 objective Ho et al. (2020):

$$\mathcal{L}_c = \mathbb{E}_{x_0 \sim q(x_0|c), t \in [1,T]} \left(\left| x_0 - G(x_t, t, c) \right|_2^2 \right), \tag{2}$$

where T is the number of denoising timesteps.

Sampling is performed iteratively. Starting with a random Gaussian noise at $x_t^{1:N}$ at $t = T$, we apply the G to produce an estimate $\hat{x}_0^{1:N}$, which is then noised by the diffusion forward process back to x_{t-1}^{N-1}. In Tevet et al. (2023), G is trained to learn both the conditioned and unconditioned distributions by randomly setting the condition $x = \varnothing$ for 10% of the samples. Consequently, the sampling process can flexibly tune between relevance to the c and diversity of unconditioned sampling, via a linear combination controlled by s:

$$G_s(x_t, T, c) := G(x_t, t, \varnothing) + s \cdot (G(x_t, t, c) - G(x_t, t, \varnothing)). \tag{3}$$

The overview of our proposed approach is outlined in Figure 1.

3.2 Functional Movement Screen.

To enable fitness-aware 3D movement generation, a tangible and consistent measure of fitness is required. To this end, the Functional Movement Screen (FMS) is a widely used diagnostic tool for identifying physical asymmetries and deficiencies. FMS offers a standardized framework for evaluating movements, making it ideal for our objectives. FMS encompasses seven core movements (Deep Squat, Hurdle Step, In-Line Lunge, Shoulder Mobility Distance, Active Straight Leg Raise, Truck Stability Push Up, and Rotary Stability Quadruped), which are assessed on a scale from 0 to 3. We leverage a recent motion-capture dataset of FMS movements, annotated by type and score, as assembled by Xing et al. (2022). The FMS dataset Xing et al. (2022) captures motion data, including color images, depth images, quaternions, 2D pixel trajectories of joints, and 3D skeleton joint positions, utilizing two Azure Kinect depth sensors. The recordings feature 45 human subjects, aged 18–59, executing 15 distinct movements that illustrate the 7 movement patterns assessed by a standard Functional Movement Screen. These movements are annotated by three FMS experts using the standard 0 to 3 scale, with criteria detailed in the supplementary material's table interpreting each FMS grade. Notably, the dataset does not include movements labeled with grade 0.

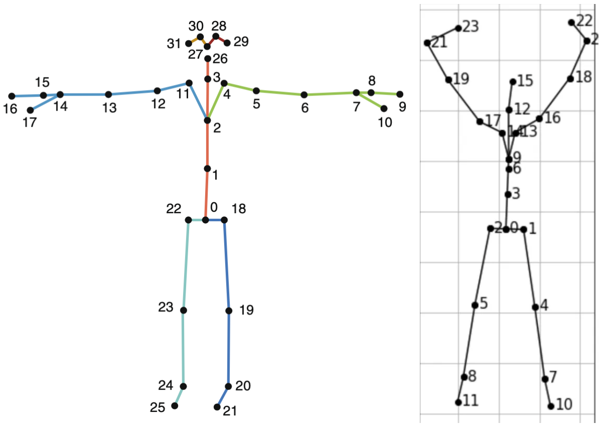


Figure 2: Skeleton models of FMS Xing et al. (2022) and HumanAct12 Guo et al. (2020). (Left) FMS skeleton with 32 joints Xing et al. (2022). (Right) HumanAct12 skeleton with 24 joints Guo et al. (2020).

4 Experiments

4.1 Overview

Following Tevet et al. (2023), we first train a base motion diffusion model (MDM) on HumanML3D dataset Guo et al. (2022b) of text-motion pairs, We then fine-tune the model on the aforementioned FMS dataset Xing et al. (2022) by enriching text-motion pairs with FMS grade labels. We experiment with two approaches to conditioning: (a) specifying the FMS grade as an integer, or (b) as a descriptive adjectival modifier. Thus, our final model demonstrates the capability to generate motions based on both FMS scores and action type descriptions, whereas the base model Tevet et al. (2023) is limited to generating motions solely based on the action type.

We integrated FMS into 3D human motion generation (HMG) to render the process fitness-aware. This consists of: (a) aligning FMS data with existing HMG data for training (Section 4.2), and (b) incorporating FMS into established methods (Section 4.3 and 4.4).

4.2 Data Preprocessing

We use HumanML3D Guo et al. (2022b) to train the motion diffusion model. This dataset, based on HumanAct12 Guo et al. (2020) and AMASS Mahmood et al. (2019), is utilised in recent text-driven human motion generation methods Tevet et al. (2023); Guo et al. (2022b), addressing the challenge of creating diverse and natural motion sequences accurately depicting the content in text descriptions. However, texts alone are not sufficient to accurately and comprehensively describe the corresponding motion clips. To address this, we incorporate FMS movements and annotations from Xing et al. (2022). Below we detail how we map transformed FMS data to minimise distributional shift relative to the HumanML3D.

We leverage scripts provided by Guo et al. (2022b) to transform the HumanAct12 and AMASS to a consistent specification for ingestion by the diffusion model, which we subsequently apply to the transformed FMS dataset.

As shown in Figure 2, the FMS skeleton comprises 32 joints whereas the HumanAct12 comprises 24; moreover, the latter is not a subset of the former. We discard the hands, thumbs, neck, eyes, and ears joints from the FMS skeleton. The FMS 1-2 chain corresponds to the navel-chest axis, whereas HumanAct12 represents this body part with the 3-6-9 chain. Aligning the navel joints, we assume that the corresponding HumanAct12 3-6-9 chain forms a straight line determined by FMS 1-2, and interpolate the positions of HumanAct12 6 and 9 according to the empirically observed mean interjoint distances observed in both datasets. The remaining FMS joints have relatively unambiguous counterparts among the HumanAct12 joints, so we use them directly.

Then, we rotate all FMS positions into a coordinate system that is perpendicularly aligned with the ground, to adjust for the 6° downwards tilt of the Kinect cameras as described in Xing et al. (2022). Lastly, we apply a uniform translation to all FMS positions to align the mean position of the navel joint with that of the HumanAct12 dataset.

4.3 Motion Diffusion Model Pretraining

Prior to conditioning the motion diffusion model on the FMS score, we first train the model from scratch for better accuracy and performance beyond that of the official pretrained model provided Tevet et al. (2023). Following Tevet et al. (2023), we adopt a similar transformer-based diffusion model architecture, wherein each denoising step predicts the signal rather than the noise. The model is optimised with a simple L2 reconstruction loss on the joints' cartesian coordinates, as shown in Equation 2. We select different hyperparameters to facilitate a gradual enhancement in the model's ability to denoise by employing a lower learning rate over extended training iterations. Simultaneously, we incorporate weight decay Loshchilov and Hutter (2019) to mitigate overfitting.

4.4 FMS-Conditioned Motion Fine-tuning

After the initial training steps outlined in Section 4.3, we fine-tune the trained MDM on FMS-conditioned movements. We experiment with two representations for the FMS labels:

- **Numerical.** The FMS score is appended to the end of the action text as a number in the format: `FMS score x`. For example: "`a person performs deep squat with their heels on the floor, FMS score 1`".
- **Textual.** The FMS score is appended to the beginning of the action text as a verbatim description with the following mapping:
  - FMS Score 1: `an injured`
  - FMS Score 2: `moderately fit`
  - FMS Score 3: `a very athletic`

  For example: "`an injured person performs deep squat with their heels on the floor`".

During the FMS-conditional finetuning steps, we finetune the pretrained MDM on the FMS dataset with FMS labels. This process results in two fitness-aware methods, which we refer to as the numerical model and textual model respectively.

	R@3 ↑	FID ↓	M.D. ↓	DIV. →	M. ↑
Ground Truth	$0.798_{\pm 0.002}$	$0.002_{\pm 0.000}$	$2.976_{\pm 0.008}$	$9.507_{\pm 0.065}$	–
PT-HP1	$0.623_{\pm 0.006}$	$0.295_{\pm 0.027}$	$5.465_{\pm 0.039}$	$9.657_{\pm 0.108}$	$2.982_{\pm 0.133}$
PT-HP2	$0.632_{\pm 0.007}$	$0.399_{\pm 0.047}$	$5.372_{\pm 0.039}$	$9.647_{\pm 0.083}$	$2.798_{\pm 0.147}$
FT-HP1	$0.586_{\pm 0.004}$	$0.571_{\pm 0.053}$	$5.623_{\pm 0.023}$	$9.366_{\pm 0.079}$	$2.744_{\pm 0.203}$
FT-HP2	$0.614_{\pm 0.006}$	$0.949_{\pm 0.090}$	$5.448_{\pm 0.026}$	$9.537_{\pm 0.084}$	$3.403_{\pm 0.142}$
Base MDM	$0.611_{\pm 0.007}$	$\mathbf{0.544}_{\pm 0.044}$	$5.566_{\pm 0.027}$	$9.560_{\pm 0.086}$	$2.799_{\pm 0.072}$
Fine-tuned MDM	$\mathbf{0.614}_{\pm 0.006}$	$0.949_{\pm 0.090}$	$\mathbf{5.448}_{\pm 0.026}$	$\mathbf{9.537}_{\pm 0.084}$	$\mathbf{3.403}_{\pm 0.142}$

Table 1: **Quantitative Results on HumanML3D**. "↑": higher is better. "↓": lower is better. "→": closer to Ground Truth is better. ± is 95% confidence interval. The fine-tuned MDM model uses numerical FMS scores. Bold indicates best result.

Table 1 shows the quantitative comparison with the base fitness-unaware MDM Tevet et al. (2023). We report five standard evaluation metrics Guo et al. (2020): R-Precision (top$k$, R@K), FID, Multimodal Distance (M.D.), Diversity (DIV.), and Multimodality (M.). MDM is the base fitness-unaware model Tevet et al. (2023). The last four columns are ablation studies on training strategies and hyperparameter settings, where "PT" and "FT" means pretraining (Section 4.3) and finetuning (Section4.4), respectively. Our method achieves significantly better results regarding the five standard metrics. We also conduct ablation studies on the two-stage training strategies (pretraining and finetuning, abbreviated as PT and FT respectively) and two different hyperparameter choices (HP1 and HP2). HP1 is set so that the learning rate matches Tevet et al. (2023)'s value of 1e-4, reduce the batch size to 32, and increase the number of steps to 600,000. HP2 is set so that the learning rate is decreased to 3e-5, the batch size is decreased further to 8, and the number of steps further increased to 2,400,000. The weight decay is set as 3e-06. Results are in the last four columns in the table.

4.5 Classifier Results on Numerical and Textual FMS

To further evaluate our FMS-conditioned models, we build a classifier to recognise FMS scores based on motions. We choose to use a Long-short term memory architecture Yu et al. (2019), implement the following model:

```
LSTM_1(
  (lstm): LSTM(input_size=360, hidden_size=120)
  (relu): ReLU()
  (fc): Linear(in_features=120, out_features=3, bias=True)
  (softmax): Softmax(dim=1)
)
```

The model is used to do 3-label classification given a motion. Classification is performed on the following datasets:

- **Original FMS dataset Xing et al. (2022)** - 1812 movements based on real videos capture from human experiments in an isolated environment.
- **Base Model FMS dataset** - we use the base pre-trained motion diffusion model to generate an evenly distributed dataset. For each combination of action class(15) and FMS score(3), we generate 40 motions to obtain a total of 1800 synthetic motions with FMS labels. This dataset is used as baseline for comparison to the datasets generated by our fine-tuned models.
- **Numerical Model FMS dataset** - similarly to 4.5, we generate 1800 motions with FMS labels.
- **Textual Model FMS dataset** - 1800 motions with FMS labels 4.5.

Since the original FMS dataset Xing et al. (2022) is heavily distributed towards FMS score 2, we use a Focal Loss function to addresses class imbalance Lin et al. (2018). Focal Loss applies a modulating term to the Cross-Entropy Loss in order to focus learning on hard misclassified examples Lin et al. (2018):

$$\mathrm{FL}\left(p_t\right)=-\left(1-p_t\right)^\gamma \log \left(p_t\right) \tag{4}$$

We use an Adam optimizer with a 5×10^{-4} learning rate and a 1×10^{-5} weight decay. The model is trained on the original dataset for 64 epochs and on the remaining 3 datasets for 512 epochs.

Metric	Original	Base	Numerical	Textual
Accuracy	84%	39%	61%	75%
Precision F1	81%	41%	62%	76%
Recall F1	74%	39%	61%	75%
Macro F1	77%	36%	61%	75%
Micro F1	84%	39%	62%	75%

Table 2: Classification Results

Table 2 shows the results for the four datasets. On the original FMS dataset, the classifier achieves 84% accuracy, 81% precision and 74% recall. With the dataset produced by the base diffusion model without fine-tuning, the classifier gets results close to random guessing with 39% (random is 33.3%). On the numerical dataset, where FMS scores are represented as numbers produced by our fine-tuned diffusion model, the classifier achieves significantly higher results than the baseline. The accuracy and the recall are improved by 22% to 61%, precision is improved by 19% to 62%. On the dataset produced by our textual fine-tuned diffusion model, there is an even more significant increase. The accuracy and precision are further increased to 75% and the precision is improved to 76%. Comparing the original dataset to the synthetic textual one produced by our model, there is only 9% difference in the classifier's accuracy. Even more impressive is the fact that the classifier achieves higher recall on the synthetic textual dataset compared to the original real dataset.

Figure 5 in the supplementary materials shows the confusion matrices of the classifier for the four datasets. The original FMS is very imbalanced towards FMS class 2, as the classifier achieves 94% for class 2 but only around 60% for class 1 and 2. With the base dataset, the classifier's predictions deviate towards random classification. On the numerical dataset from our fine-tuned model, the classifier achieves much more balanced predictions in the range of 53% to 69% accuracy. Since the model is fine-tuned on the original FMS dataset, there is still slight bias towards class 2, although is significantly lessened. On the textual dataset from our fine-tuned model, the classifier achieves better results on class 1(80%) and 3(76%) compared to the original dataset. In the textual FMS dataset, all three action class achieve at least 70%.

In summary, the results from the LSTM classifier show that our fine-tuned models produce high-quality movement data that is well-balanced across all FMS scores. On our best synthetic dataset, the classifier achieved higher recall (76%) than the on the original dataset (75%), while also maintaining similar accuracy and precision (75%). Furthermore, with our best synthetic dataset, the classifier achieved higher prediction accuracy on two of the three FMS classes (class 1 - 80% compared to 65%, class 3 - 78% compared to 63%).

5 Limitations

While our work to condition human motion diffusion models on fitness grades is novel and promising, some limitations remain. One main challenge in human motion generation is the low availability of diverse and comprehensive datasets. The FMS dataset has some such limitations:

- The dataset focuses on seven specific FMS movements collected from 45 participants, which does not capture the full diversity of human movements necessary for broader applications in accurate motion generation. The complexity and variety of human motion might require larger datasets for the model to generate fully accurate motions.
- The dataset lacks participant diversity, with data collected from 45 individuals aged 18-59 recruited on a university campus. This limits the model's ability to generalise across different demographic groups.
- Each motion in the FMS dataset is annotated from 0 to 3 by FMS experts. While the use of experts promotes quality, this labelling is unavoidably subjective.
- HumanML3D already allows sampling with textual descriptions of physique (e.g., "the person is walking like they are tired"). FMS scores may be less intuitive for users compared to descriptive text about physique.
- The AMASS dataset includes shape parameters that describe variations in physique, such are not included in our FMS fine-tuning method.

Finally, interpretability is another limitation. Diffusion models often operate as "black boxes“, making it difficult to understand how and why they generate specific motions. This lack of interpretability can be a limitation, especially in fields like healthcare where understanding the rationale behind a movement pattern is crucial.

6 Conclusion

In this paper we present the first fitness-aware 3D human motion generation method. Our method conditions human motion diffusion models to generate 3D movements based on FMS scores, enabling the generation of motions sensitive to the subject's fitness level. Our contribution extends diffusion models to incorporate fitness awareness, allowing the model to focus on the physical condition of the individual instead of relying on type-based motions.

References

Ahuja, C. and Morency, L. (2019). Language2pose: Natural language grounded pose forecasting. In *International Conference on 3D Vision (3DV)*, pages 719–728. IEEE.

Chen, K.-Y., Shin, J., Hasan, M. A. M., Liaw, J.-J., Yuichi, O., and Tomioka, Y. (2022). Fitness movement types and completeness detection using a transfer-learning-based deep neural network. *Sensors*, 22(15):5700.

Chen, L. et al. (2023a). Humanmac: Masked motion completion for human motion prediction. *arXiv preprint arXiv:2302.03665*.

Chen, X. et al. (2023b). Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010.

Chong, E. et al. (2020). Detection of eye contact with deep neural networks is as accurate as human experts. *Nature Communications*, 11(1):6386.

Chou, G. et al. (2023). Diffusion-sdf: Conditional generative modeling of signed distance functions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2262–2272.

Christen, S., Yang, W., Pérez-D'Arpino, C., Hilliges, O., Fox, D., and Chao, Y. (2023). Learning human-to-robot handovers from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9654–9664.

Dabral, R. et al. (2023). Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9760–9770.

Duan, Y. et al. (2021). Single-shot motion completion with transformer. *arXiv preprint arXiv:2103.00776*.

Guo, C. et al. (2020). Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029.

Guo, C. et al. (2022a). Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision (ECCV)*.

Guo, C. et al. (2024a). Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Guo, C., Mu, Y., Zuo, X., Dai, P., Yan, Y., Lu, J., and Cheng, L. (2024b). Generative human motion stylization in latent space. *arXiv preprint arXiv:2401.13505*.

Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., and Cheng, L. (2022b). Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161.

Guo, Y. et al. (2023). Student close contact behavior and covid-19 transmission in china's classrooms. *PNAS Nexus*, 2(5):pgad142.

Harvey, F. et al. (2020). Robust motion inbetweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–61.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

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

Huang, Y., Wan, W., Yang, Y., Callison-Burch, C., Yatskar, M., and Liu, L. (2024). Como: Controllable motion generation through language guided pose code editing. *arXiv preprint arXiv:2403.13900*.

Jiang, B. et al. (2023). Motiongpt: Human motion as a foreign language. In *Advances in Neural Information Processing Systems*, volume 36, pages 20067–20079.

Kong, H. et al. (2023). Priority-centric human motion generation in discrete latent space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14806–14816.

Lee, H. et al. (2019). Dancing to music. In *Advances in Neural Information Processing Systems*, volume 32.

Lee, S., Starke, S., Ye, Y., Won, J., and Winkler, A. (2023a). Questenvsim: Environment-aware simulated motion tracking from sparse sensors. *arXiv preprint arXiv:2306.05666*.

Lee, T., Moon, G., and Lee, K. (2023b). Multiact: Long-term 3d human motion generation from multiple action labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1231–1239.

Li, B. et al. (2022). Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1272–1279.

Li, R. et al. (2021). Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412.

Li, R. et al. (2023a). Finedance: A fine-grained choreography dataset for 3d full body dance generation. *arXiv preprint arXiv:2212.03741*.

Li, Z., Peng, X., Abbeel, P., Levine, S., Berseth, G., and Sreenath, K. (2023b). Robust and versatile bipedal jumping control through reinforcement learning. In *Robotics: Science and Systems XIX*, Daegu, Republic of Korea.

Liao, Z. et al. (2022). Skeleton-free pose transfer for stylized 3d characters. In *European Conference on Computer Vision*, pages 640–656. Springer.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2018). Focal Loss for Dense Object Detection.

Lin, X. et al. (2023). Automatic evaluation of functional movement screening based on attention mechanism and score distribution prediction. *Mathematics*, 11(24):4936.

Liu, X. et al. (2022). Close contact behavior-based covid-19 transmission and interventions in a subway system. *Journal of Hazardous Materials*, 436:129233.

Liu, Y. et al. (2023). Syncdreamer: Generating multiview-consistent images from a single view image. *arXiv preprint arXiv:2309.03453*.

Long, X. et al. (2023). Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint arXiv:2310.15008*.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Luo, Z. et al. (2023). Videofusion: Decomposed diffusion models for high-quality video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10209–10218.

Mahmood, N., Ghorbani, N., Troje, N. F., Pons-Moll, G., and Black, M. J. (2019). Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451.

Müller, N. et al. (2023). Diffrf: Rendering-guided 3d radiance field diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4328–4338.

Neverhood, J. and Xiaosong, C. (2021). Stop motion obj. `https://github.com/neverhood311/Stop-motion-OBJ`.

Nie, W. et al. (2022). Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*.

Pang, K. et al. (2023). Bodyformer: Semantics-guided 3d body gesture synthesis with transformer. *ACM Transactions on Graphics (TOG)*, 42(4):1–12.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A., Tzionas, D., and Black, M. J. (2019). Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985.

Petrovich, M., Black, M., and Varol, G. (2021). Action-conditioned 3d human motion synthesis with transformer vae. In *International Conference on Computer Vision (ICCV)*.

Petrovich, M., Black, M., and Varol, G. (2022). Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*.

Pi, H. et al. (2023). Hierarchical generation of human-object interactions with diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15061–15073.

Plappert, M., Mandery, C., and Asfour, T. (2016). The KIT motion-language dataset. *Big Data*, 4(4):236–252.

Po, R. et al. (2023). State of the art on diffusion models for visual computing. *arXiv preprint arXiv:2310.07204*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Remedios, S. M., Armstrong, D. P., Graham, R. B., and Fischer, S. L. (2020). Exploring the application of pattern recognition and machine learning for identifying movement phenotypes during deep squat and hurdle step movements. *Frontiers in Bioengineering and Biotechnology*, 8:364.

Ren, Z. et al. (2023). Diffusion motion: Generate text guided 3d human motion by diffusion model. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Rombach, R. et al. (2022). High resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://github.com/CompVis/latent-diffusion https://arxiv.org/abs/2112.10752.

Saharia, C. et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pages 36479–36494.

Shafir, Y., Tevet, G., Kapon, R., and Bermano, A. H. (2023). Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*.

Shi, M. et al. (2023). Phasemp: Robust 3d pose estimation via phase-conditioned human motion prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14725–14737.

Sinha, A. et al. (2021). D2c: Diffusion-decoding models for few-shot conditional generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 12533–12548.

Song, Y. and Ermon, S. (2020). Improved techniques for training score-based generative models. In *Advances in Neural Information Processing Systems*, volume 33, pages 12438–12448.

Spilz, A. and Munz, M. (2023). Automatic assessment of functional movement screening exercises with deep learning architectures. *Sensors*, 23(1):5.

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

Tevet, G. et al. (2022). Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer.

Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., and Bermano, A. H. (2023). Human motion diffusion model. In *International Conference on Learning Representations*.

Tseng, J. et al. (2023). Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 448–458.

Vahdat, A. et al. (2021). Score-based generative modeling in latent space. In *Advances in Neural Information Processing Systems*, volume 34, pages 11287–11302.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wan, W., Dou, Z., Komura, T., Wang, W., Jayaraman, D., and Liu, L. (2023). Tlcontrol: Trajectory and language control for human motion synthesis. *arXiv preprint arXiv:2311.17135*.

Wan, W. et al. (2022). Learn to predict how humans manipulate large-sized objects from interactive motions. *IEEE Robotics and Automation Letters*, 7(2):4702–4709.

Wang, Y. et al. (2023). Fg-t2m: Fine-grained text driven human motion generation via diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22035–22044.

Wei, D. et al. (2023). Human joint kinematics diffusion-refinement for stochastic motion prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6110–6118.

Winkler, A., Won, J., and Ye, Y. (2022). Questsim: Human motion tracking from sparse sensors with simulated avatars. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–8.

Xing, Q.-J., Shen, Y.-Y., Cao, R., Zong, S.-X., Zhao, S.-X., and Shen, Y.-F. (2022). Functional movement screen dataset collected with two azure kinect depth sensors. *Scientific Data*, 9(1):104.

Xu, L. et al. (2023). Actformer: A gan-based transformer towards general actionconditioned 3d human motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2228–2238.

Yamane, K., Revfi, M., and Asfour, T. (2013). Synthesizing object receiving motions of humanoid robots with human motion database. In *IEEE International Conference on Robotics and Automation*, pages 1629–1636. IEEE.

Yang, R. et al. (2022). Diffusion probabilistic modeling for video generation. *arXiv preprint arXiv:2203.09481*.

Yang, X. et al. (2023). Analysis of sars cov-2 transmission in airports based on real human close contact behaviors. *Journal of Building Engineering*, page 108299.

Ye, Y., Liu, L., Hu, L., and Xia, S. (2022). Neural3points: Learning to generate physically realistic full-body motion for virtual reality users. *Computer Graphics Forum*, 41:183–194.

Yu, Y., Si, X., Hu, C., and Zhang, J. (2019). A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270.

Yu, Z. et al. (2023). Surf-d: High-quality surface generation for arbitrary topologies using diffusion models. *arXiv preprint arXiv:2311.17050*.

Yuan, Y. et al. (2022). Physdiff: Physics guided human motion diffusion model. *arXiv preprint arXiv:2212.02500*.

Zhang, J. et al. (2023a). Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhang, J. et al. (2023b). Skinned motion retargeting with residual perception of motion semantics geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13864–13872.

Zhang, M. et al. (2022). Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*.

Zhang, M. et al. (2023c). Remodiffuse: Retrieval-augmented motion diffusion model. *arXiv preprint arXiv:2304.01116*.

Zhang, M., Li, H., Cai, Z., Ren, J., Yang, L., and Liu, Z. (2024). Finemogen: Fine-grained spatio-temporal motion generation and editing. In *Advances in Neural Information Processing Systems*, volume 36.

Zhang, N. et al. (2023d). Close contact behaviors of university and school students in 10 indoor environments. *Journal of Hazardous Materials*, 458:132069.

Zhou, W., Dou, Z., Cao, Z., Liao, Z., Wang, J., Wang, W., Liu, Y., Komura, T., Wang, W., and Liu, L. (2023). Emdm: Efficient motion diffusion model for fast, high-quality motion generation. *arXiv preprint arXiv:2312.02256*, 2.

Zhuang, H., Zhang, Y., and Liu, S. (2023). A pilot study of query-free adversarial attack against stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2384–2391.

Supplementary Material

Training setup

Our method is closely related to human Motion Diffusion Model (MDM) Tevet et al. (2023). Please refer to Figure 3 for schematic. We also provide discussion in the main paper.

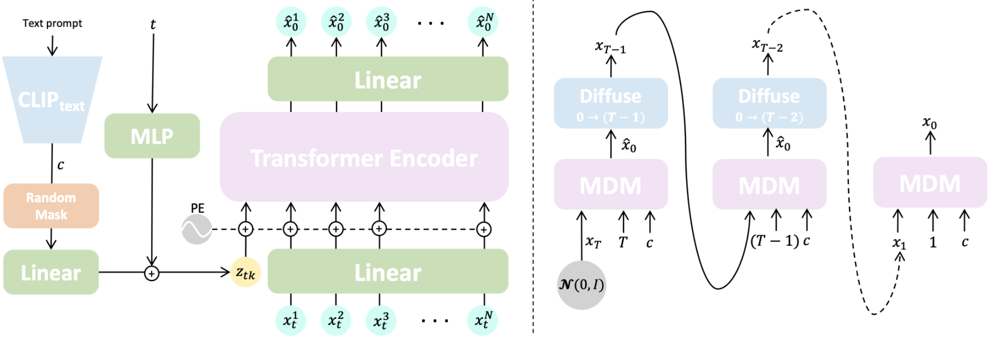


Figure 3: (Left) the Motion Diffusion Model combines a conditioning token with joint positions to iteratively clean a noised $x_t^{1:N}$ back to the clean motion $\hat{x}_0^{1:N}$. (Right) sampling from the motion diffusion model is done by applying iteratively the Motion Diffusion Model to $x_t^{1:N}$ and noising it back to $x_{t-1}^{1:N}$. The figures are taken from Tevet et al. (2023).

All training is performed with the PyTorch framework Paszke et al. (2019). We use AdamW Loshchilov and Hutter (2019) optimiser with default learning rate of 1e-4 and no weight decay. The batch size is 32. Training proceeds by a standard loop: within an epoch, on each batch of motion and condition pairs, we run a forward pass through the model and take a gradient step with the optimiser. We evaluate and save checkpoint every 50,000 steps.

We proceed to train the Motion Diffusion Model Tevet et al. (2023) from scratch and then further fine-tune it for FMS conditioned generation Xing et al. (2022). We employ two hyperparameter choices to train the motion diffusion model. The first hyperparameter choice (HP1) sets learning rate to match Tevet et al. (2023)'s value of 1e-4, reduce the batch size to 32, and increase the number of steps to 600,000. The second hyperparameter choice (HP2) sets learning rate to 3e-5, the batch size is decreased further to 8, and the number of steps further increased to 2,400,000. The weight decay is set as 3e-06.

For FMS-conditional finetuning steps, we start from the 2,400,000-step checkpoint and train a further 600,000 steps only on the FMS dataset with FMS labels Xing et al. (2022). We use the same learning rate of 3e-05, weight decay of 3e-06, and batch size of 8 as in HP2, which shows the superior performance in the MDM pretraining.

Evaluation Metrics

We report five evaluation metrics: R-Precision, Multimodal Distance, Diversity, Multimodality, and FID. For their implementation please see Guo et al. (2020). All metrics require comparing motions and prompt text in latent space. This is achieved by two pretrained bidirectional Gated Recurrent Unit (GRU) encoders, which respectively embed the text and motion into a common representation. The encoders are trained with contrastive loss to minimise the distance between matched text-motion embeddings and separate mismatched text-motion embeddings with some margin Guo et al. (2022b). We describe each metric below:

R-Precision and Multimodal Distance

measure the alignment between the input prompt text and generated output motion Guo et al. (2022b). For top-k R-Precision, we calculate the euclidean distances between a motion embedding and its corresponding text embedding, and to the text embeddings of 31 additional randomly selected mismatched descriptions. The ground truth falling within the top-k rankings is considered "success".

Multimodal Distance is the mean euclidean distance between matched text and motion embeddings in the test set.

FID

quantifies the dissimilarity between the ground truth distribution and the diffusion model's generated distribution. This similarity is again computed as the euclidean distance in the latent space defined by the aforementioned GRU encoders Guo et al. (2020).

Diversity

captures the variance of motions across all text prompts Guo et al. (2020). Two subsets of equal size are sampled from the global set of motions, and their mean euclidean distance in the latent space is reported. Ideally, the diversity of generated motions is similar to the diversity of the ground truth.

Multimodality

calculates the variance of motions generated from a single text prompt Guo et al. (2020). This is computed as the average euclidean distance between a text feature and its corresponding motion features. Conditional on the motion still being *relevant*, as measured by R-Precision and Multimodal Distance, we desire multimodality to be high. However, in practice, the achievable multimodality in a good model is not expected to deviate significantly from the ground truth multimodality.

Datasets

HumanML3D Dataset

Guo et al. (2022b) is a 3D human motion-language dataset that combines the HumanAct12 Guo et al. (2020) and Amass Plappert et al. (2016) datasets. It includes a wide range of human actions, encompassing daily activities like walking and jumping, sports activities such as swimming and playing golf, acrobatics like cartwheels, and artistic movements including dancing.

The dataset comprises 14,616 motion clips and 44,970 text descriptions, with motions downsampled to 20 fps. Each motion clip lasts between 2 to 10 seconds, totalling approximately 28.59 hours of motion data. The average motion length is about 7.1 seconds, and each description is, on average, 12 words long.

Movement	**Variation**
Deep squat	Heels on the floor Heels on a 2-inch board
Hurdle step	Left leg up Right leg up
In-line lunge	Left leg in front Right leg in front
Shoulder mobility	Left arm up Right arm up
Active straight raise	Left leg up Right leg up
Trunk stability push-up	Support on the ground with both hands
Rotary stability	Left limb up Right limb up Left arm and right leg up Right arm and left leg up

Table 3: Description of FMS Movements.

Functional Movement Screen (FMS) Dataset

Xing et al. (2022) is a vision-based autonomous dataset of 1,812 recordings collected using two Azure Kinect depth sensors. The recordings depict 45 human subjects, aged 18–59, performing fifteen distinct movements demonstrating the seven movement patterns evaluated by a standard FMS score. Example descriptions of FMS movements are in Table 3.

Grade	Description
0	Complete with pain in any part of the body.
1	Incomplete.
2	Complete with compensation or deviation from the stand, or both.
3	Complete without compensation.

Table 4: Interpretation of each FMS grade, according to Xing et al. (2022).

The motion data includes colour images, depth images, quaternions, 2D pixel trajectories of joints, and 3D skeleton joint positions. We use only the latter, in line with HumanML3D Guo et al. (2022b). Movements are labelled by three FMS experts with the standard 0 to 3 scale. The criteria for each grade is described in Table 4. The dataset contains no movements labelled with the grade 0.

Additional Evaluations

Qualitative results

We qualitatively compare samples generated by three models: base fitness-unaware MDM Tevet et al. (2023) (base model), the “numerical” method (numerical model), and the “textual” method (verbatim model). The latter two are obtained through FMS-grade-aware fine-tuning.

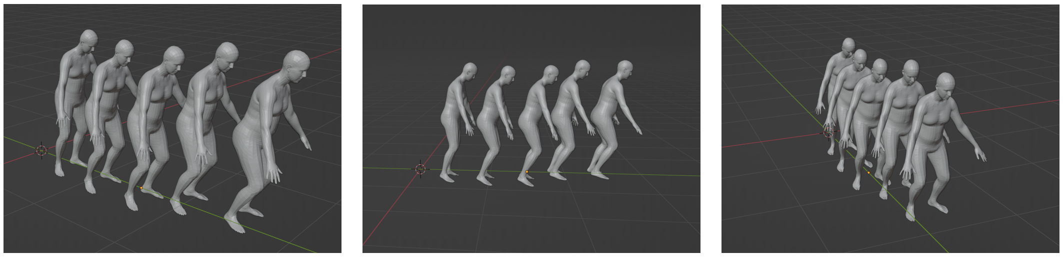
When generating new motion, files with joint coordinates and mp4 videos are produced. We generate motions for 5 action classes using the best numerical and verbatim, as well as for the base model that the authors provide (with numerical FMS scores). After we obtain videos, we further process them to produce SMPL meshes (each object mesh corresponds to one movement frame). We use Blender with a plugin called Stop-motion-obj Neverhood and Xiaosong (2021) to generate videos in SMPL-X Pavlakos et al. (2019) format.

Please refer to our Supplementary Video, which summarises all SMPL-X videos for the five action classes generated by the base, numerical, and textual models. It can be observed the base model hallucinates for action classes “deep squat”, “hurdle step”, and “in line lunge”, producing different and irrelevant movements. In comparison, our numerical model gives considerably better performance for those three classes - the FMS scores are distinctive as well. The textual model hallucinates somewhat as well, but still provides some meaningful movements. For action class “shoulder mobility”, the base model produces satisfactory outputs, the numerical one generates very well-defined movements and the verbatim model hallucinates completely.

Qualitative Evaluation

Upon inspection of the results generated by the base, numerical, and verbatim models, the base model hallucinates for action classes such as “deep squat”, “hurdle step”, and “in-line lunge”, and generates entirely different and irrelevant movements. In comparison, the numerical model performs significantly better for these three classes, with distinct FMS scores as well.While the verbatim model also exhibits hallucinations, it still provides some meaningful movements occasionally.

Figure 4 shows three images of a squat with FMS score 2 generated by the numerical model. The images visualize the movement with five objects depicted at different time frames of the movement. We can observe that the movement looks realistic - it is clear that the person is performing a “squat”. Although there is some slight floating above the ground at frame 4 and 5, the frames represent a consistent motion sequence of a moderately fit person performing a squat. The SMPL-X Pavlakos et al. (2019) videos demonstrate that the numerical model is more accurate and hallucinates far less than the verbatim one. One possible explanation is that the verbatim model adds significantly more text to the initial prompt, which confuses the model.

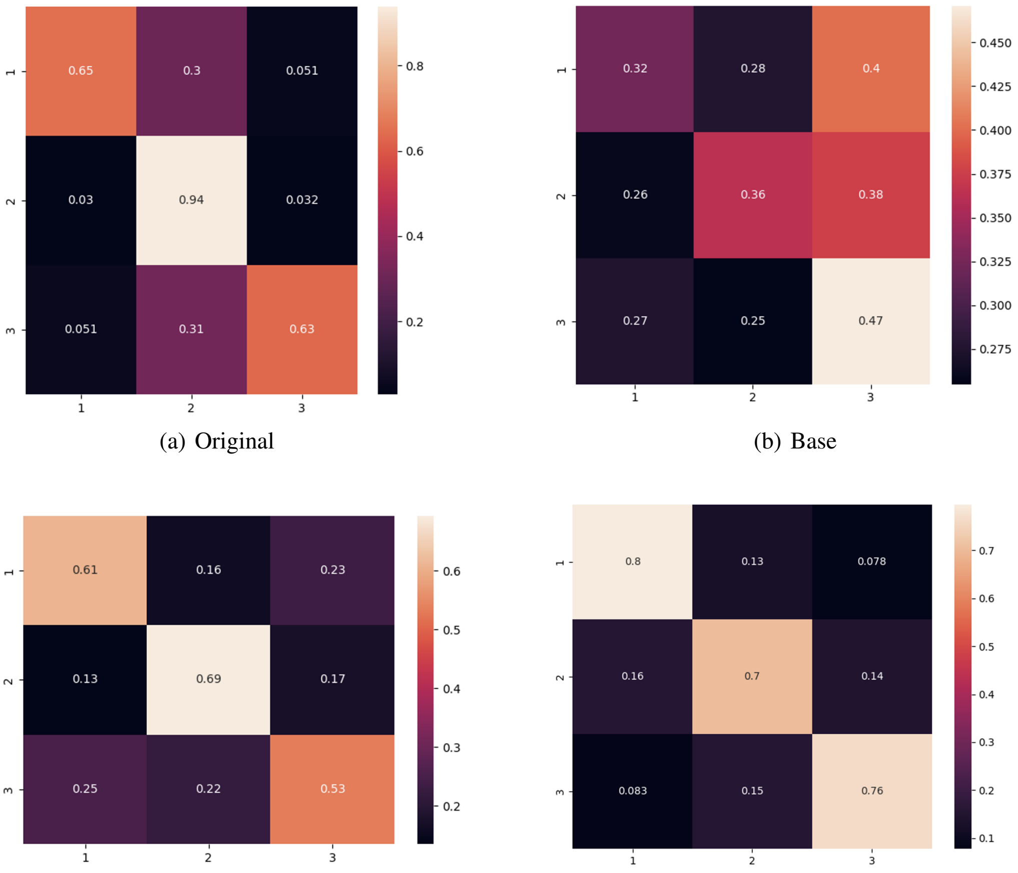


(a) Side View (30) (b) Side View (90) (c) Front View (Above)

Figure 4: **Qualitative Results**. The examples feature the Squat action with an FMS score of 2 from the numerical model at three viewpoints.

Classifier Results

To further evaluate our FMS-conditioned models, we build a classifier to recognise FMS scores based on motions. The architecture of the classifier and the the main results are discussed in section .The model is trained on the original dataset for 64 epochs and on the remaining 3 datasets for 512 epochs, using an Adam optimiser with a 5e-4 learning rate and a 1e-5 weight decay.



(a) Original (b) Base (c) Numerical (d) Textual

Figure 5: Confusion Matrices of the Classifier for Each Dataset.

Supplementary Material

Kiril Bikov[⋆]
University of Cambridge, UK

Shiye Su[⋆]
University of Cambridge, UK

Deepro Choudhury[⋆]
University of Cambridge, UK

Zhilin Guo
University of Cambridge, UK

Weihao Xia
University of Cambridge, UK

Mehmet Salih Celiktenyildiz
Bilkent University, Turkiye

Chenliang Zhou
University of Cambridge, UK

Param Hanji
University of Cambridge, UK

Cengiz Oztireli[⋆⋆]
University of Cambridge, UK

In this supplementary document, we provide further insights, experiments and analyses. Section provides more details on Implementation, Evaluation Metrics, Datasets, and the MDM Tevet et al. (2023). Section presents additional experiments to showcase the advantages of the proposed method. Lastly, Section discuss in more details the limitations, potential applications, and future directions of our proposed method.

Implementation

Training setup

Our method is closely related to human Motion Diffusion Model (MDM) Tevet et al. (2023). Please refer to Figure 1 for schematic. We also provide discussion in the main paper.

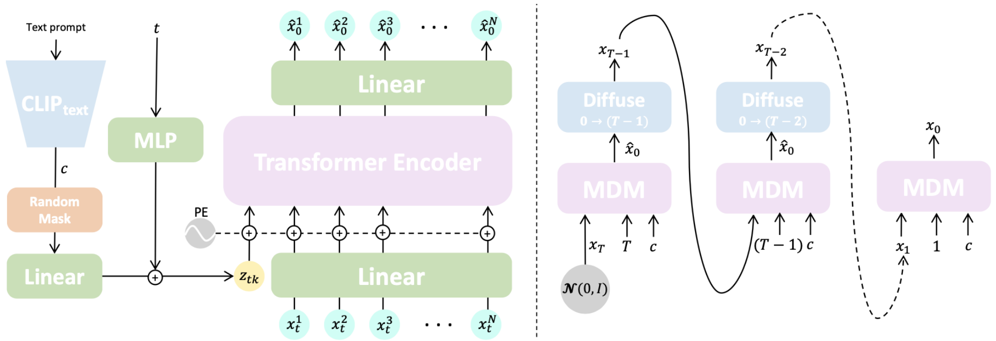


Fig. 1: (Left) the Motion Diffusion Model combines a conditioning token with joint positions to iteratively clean a noised $x_t^{1:N}$ back to the clean motion $\hat{x}_0^{1:N}$. (Right) sampling from the motion diffusion model is done by applying iteratively the Motion Diffusion Model to $x_t^{1:N}$ and noising it back to $x_{t-1}^{1:N}$. The figures are taken from Tevet et al. (2023).

All training is performed with PyTorch framework Paszke et al. (2019). We use AdamW Loshchilov and Hutter (2019) optimiser with default learning rate of 1e-4 and no weight decay. The batch size

[⋆] Equal contribution
[⋆⋆] Corresponding author

is 32. Training proceeds by a standard loop: within an epoch, on each batch of motion and condition pairs, we run a forward pass through the model and take a gradient step with the optimiser. We evaluate and save checkpoint every 50,000 steps.

We proceed to train the Motion Diffusion Model Tevet et al. (2023) from scratch and then further fine-tune it for FMS conditioned generation Xing et al. (2022). We employ two hyperparameter choices to train the motion diffusion model. The first hyperparameter choice (HP1) sets learning rate to match Tevet et al. (2023)'s value of 1e-4, reduce the batch size to 32, and increase the number of steps to 600,000. The second hyperparameter choice (HP2) sets learning rate to 3e-5, the batch size is decreased further to 8, and the number of steps further increased to 2,400,000. The weight decay is set as 3e-06.

For FMS-conditional finetuning steps, we start from the 2,400,000-step checkpoint and train a further 600,000 steps only on the FMS dataset with FMS labels Xing et al. (2022). We use the same learning rate of 3e-05, weight decay of 3e-06, and batch size of 8 as in HP2, which shows the superior performance in the MDM pretraining.

Evaluation Metrics

We report five evaluation metrics: R-Precision, Multimodal Distance, Diversity, Multimodality, and FID Heusel et al. (2017). For their implementation please see Guo et al. (2020). All metrics require comparing motions and prompt text in latent space. This is achieved by two pretrained bidirectional Gated Recurrent Unit (GRU) encoders, which respectively embed the text and motion into a common representation. The encoders are trained with contrastive loss to minimise the distance between matched text-motion embeddings and separate mismatched text-motion embeddings with some margin Guo et al. (2022). We describe each metric below:

R-Precision and Multimodal Distance

measure the alignment between the input prompt text and generated output motion Guo et al. (2022). For top-k R-Precision, we calculate the euclidean distances between a motion embedding and its corresponding text embedding, and to the text embeddings of 31 additional randomly selected mismatched descriptions. The ground truth falling within the top-k rankings is considered "success". Multimodal Distance is the mean euclidean distance between matched text and motion embeddings in the test set.

FID

quantifies the dissimilarity between the ground truth distribution and the diffusion model's generated distribution. This similarity is again computed as the euclidean distance in the latent space defined by the aforementioned GRU encoders Guo et al. (2020).

Diversity

captures the variance of motions across all text prompts Guo et al. (2020). Two subsets of equal size are sampled from the global set of motions, and their mean euclidean distance in the latent space is reported. Ideally, the diversity of generated motions is similar to the diversity of the ground truth.

Multimodality

calculates the variance of motions generated from a single text prompt Guo et al. (2020). This is computed as the average euclidean distance between a text feature and its corresponding motion features. Conditional on the motion still being *relevant*, as measured by R-Precision and Multimodal Distance, we desire multimodality to be high. However, in practice, the achievable multimodality in a good model is not expected to deviate significantly from the ground truth multimodality.

Datasets

HumanML3D Dataset

Guo et al. (2022) is a 3D human motion-language dataset that combines the HumanAct12 Guo et al. (2020) and Amass Plappert et al. (2016) datasets. It includes a wide range of human actions, encompassing daily activities like walking and jumping, sports activities such as swimming and playing golf, acrobatics like cartwheels, and artistic movements including dancing.

The dataset comprises 14,616 motion clips and 44,970 text descriptions, with motions downsampled to 20 fps. Each motion clip lasts between 2 to 10 seconds, totalling approximately 28.59 hours of motion data. The average motion length is about 7.1 seconds, and each description is, on average, 12 words long.

Movement	**Variation**
Deep squat	Heels on the floor Heels on a 2-inch board
Hurdle step	Left leg up Right leg up
In-line lunge	Left leg in front Right leg in front
Shoulder mobility	Left arm up Right arm up
Active straight raise	Left leg up Right leg up
Trunk stability push-up	Support on the ground with both hands
Rotary stability	Left limb up Right limb up Left arm and right leg up Right arm and left leg up

Table 1: Description of FMS Movements.

Functional Movement Screen (FMS) Dataset

Xing et al. (2022) is a vision-based autonomous dataset of 1,812 recordings collected using two Azure Kinect depth sensors. The recordings depict 45 human subjects, aged 18–59, performing fifteen distinct movements demonstrating the seven movement patterns evaluated by a standard FMS score. Example descriptions of FMS movements are in Table 1.

Grade	Description
0	Complete with pain in any part of the body.
1	Incomplete.
2	Complete with compensation or deviation from the stand, or both.
3	Complete without compensation.

Table 2: Interpretation of each FMS grade, according to Xing et al. (2022).

The motion data includes colour images, depth images, quaternions, 2D pixel trajectories of joints, and 3D skeleton joint positions. We use only the latter, in line with HumanML3D Guo et al. (2022). Movements are labelled by three FMS experts with the standard 0 to 3 scale. The criteria for each grade is described in Table 2. The dataset contains no movements labelled with the grade 0.

Additional Evaluations

Qualitative results

We qualitatively compare samples generated by three models: base fitness-unaware MDM Tevet et al. (2023) (base model), the "numerical" method (numerical model), and the "textual" method (verbatim model). The latter two are obtained through FMS-grade-aware fine-tuning.

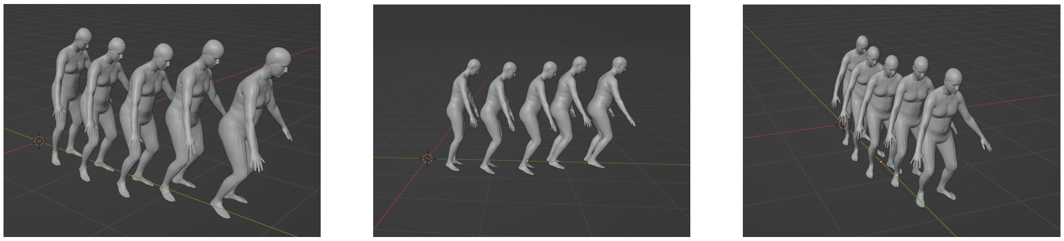
When generating new motion, files with joint coordinates and mp4 videos are produced. We generate motions for 5 action classes using the best numerical and verbatim, as well as for the base model that the authors provide (with numerical FMS scores). After we obtain videos, we further process them to produce SMPL meshes (each object mesh corresponds to one movement frame). We use Blender with a plugin called Stop-motion-obj Neverhood and Xiaosong (2021) to generate videos in SMPL-X Pavlakos et al. (2019) format.

Please refer to our Supplementary Video, which summarises all SMPL-X videos for the five action classes generated by the base, numerical, and textual models. It can be observed the base model hallucinates for action classes "deep squat", "hurdle step", and "in line lunge", producing different and irrelevant movements. In comparison, our numerical model gives considerably better performance for those three classes - the FMS scores are distinctive as well. The textual model hallucinates somewhat as well, but still provides some meaningful movements. For action class "shoulder mobility", the base model produces satisfactory outputs, the numerical one generates very well-defined movements and the verbatim model hallucinates completely.

Qualitative Evaluation.

Below, we present a qualitative comparison with samples generated by three models: the base fitness-unaware MDM Tevet et al. (2023) (base model), the "numerical" method (numerical model), and the "textual" method (verbatim model). The latter two are obtained through FMS-grade-aware fine-tuning.

Upon inspection of the results generated by the base, numerical, and verbatim models, the base model hallucinates for action classes such as "deep squat", "hurdle step", and "in-line lunge", and generates entirely different and irrelevant movements. In comparison, the numerical model performs significantly better for these three classes, with distinct FMS scores as well. While the verbatim model also exhibits hallucinations, it still provides some meaningful movements occasionally. For the action class "shoulder mobility", the base model produces satisfactory outputs, the numerical one generates very well-defined movements, and the verbatim model completely hallucinates.



(a) Side View (30°)

(b) Side View (90°)

(c) Front View (Above)

Fig. 2: **Qualitative Results**. The examples feature the Squat action with an FMS score of 2 from the numerical model at three viewpoints.

Figure 2 shows three images of a squat with FMS score 2 generated by the numerical model. The images visualize the movement with five objects depicted at different time frames of the movement. We can observe that the movement looks realistic - it is clear that the person is performing a "squat". Although there is some slight floating above the ground at frame 4 and 5, the frames represent a consistent motion sequence of a moderately fit person performing a squat. The SMPL-X Pavlakos et al. (2019) videos demonstrate that the numerical model is more accurate and hallucinates far less than the verbatim one. One possible explanation is that the verbatim model adds significantly more text to the initial prompt, which confuses the model. Presumably, with longer added text for the FMS score, it becomes unclear for the model which part of the text describes the action class and which part describes the FMS score. This ambiguity may lead to significantly more hallucinations and unrealistic movements.

Classifier Results

To further evaluate our FMS-conditioned models, we build a classifier to recognise FMS scores based on motions. We opt for a long-short term memory (LSTM) architecture Hochreiter and Schmidhuber (1997), given its superior performance in recognizing patterns in sequences and handling variable-length frames. Since we want our classifier to work on various data formats without particular bias or change of accuracy, we opt for LSTM. We implement the following model:

```
LSTM_1(
  (lstm): LSTM(input_size=360, hidden_size=120)
  (relu): ReLU()
  (fc): Linear(in_features=120, out_features=3, bias=True)
  (softmax): Softmax(dim=1)
)
```

The model is employed for 3-label classification given a motion. Classification is conducted on the following datasets:

- Original FMS Data. This includes 1,812 movements based on real videos capture from human experiments in an isolated environment Xing et al. (2022).
- Base Model FMS Data. We use the base pretrained motion diffusion model to generate an evenly distributed dataset. For each combination of action class (15) and FMS score (3), we generate 40 motions to obtain a total of 1800 synthetic motions with FMS labels. This dataset is used as baseline for comparison to the datasets generated by our fine-tuned models.
- Numerical Model FMS Data. Similar to the Base Model FMS Data, we generate 1,800 motions with FMS labels.
- Textual Model FMS Data. Similarly, we generate 1,800 motions with FMS labels.

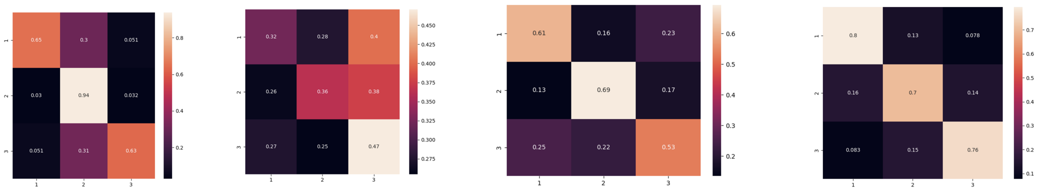
Since the original FMS dataset Xing et al. (2022) is heavily distributed towards FMS score 2, we use focal loss Lin et al. (2017) to addresses class imbalance. The focal loss applies a modulating term to the cross-entropy loss in order to focus learning on hard misclassified examples:

$$\mathrm{FL}\left(p_t\right) = -\left(1-p_t\right)^{\gamma}\log\left(p_t\right). \tag{1}$$

The model is trained on the original dataset for 64 epochs and on the remaining 3 datasets for 512 epochs, using an Adam optimiser with a 5e-4 learning rate and a 1e-5 weight decay.

The results for the four datasets are in the main paper. On the original FMS dataset, the classifier achieves 84% accuracy, 81% precision and 74% recall. With the dataset produced by the base diffusion model without fine-tuning, the classifier gets results close to random guessing with 39% (random is 33.3%). On the numerical dataset, where FMS scores are represented as numbers produced by our fine-tuned diffusion model, the classifier achieves significantly higher results than the baseline. The accuracy and the recall are improved by 22% to 61%, precision is improved by 19% to 62%. On the dataset produced by our textual fine-tuned diffusion model, there is an even more significant increase. The accuracy and precision are further increased to 75% and the precision is improved to 76%. Comparing the original dataset to the synthetic textual one produced by our model, there is only 9% difference in the classifier's accuracy. Even more impressive is the fact that the classifier achieves higher recall on the synthetic textual dataset compared to the original real dataset.

Figure 3 shows the confusion matrices of the classifier for the four datasets. As shown in Figure 3a, the original FMS is very imbalanced towards FMS class 2. The classifier achieves 94% for class 2 but only around 60% for class 1 and 2. With the base dataset, as in Figure 3b, the classifier's predictions deviate towards random classification. On the numerical dataset from our fine-tuned model, the classifier achieves much more balanced predictions in the range of 53% to 69% accuracy,



(a) Original (b) Base (c) Numerical (d) Textual

Fig. 3: Confusion Matrices of the Classifier for Each Dataset.

as shown in Figure 3c. Since the model is fine-tuned on the original FMS dataset, there is still slight bias towards class 2, although is significantly lessened. On the textual dataset from our fine-tuned model, as shown in Figure 3d, all three action class achieve at least 70%. The classifier achieves better results on class 1 (80%) and 3 (76%) compared to the original dataset.

In summary, results from the LSTM classifier show that our fine-tuned models produce high-quality movement data that is well-balanced across all FMS scores. On our best synthetic dataset, the classifier achieved higher recall (76%) than the on the original dataset (75%), while also maintaining similar accuracy and precision (75%). Furthermore, with our best synthetic dataset, the classifier achieved higher prediction accuracy on two of the three FMS classes (class 1 - 80% compared to 65%, class 3 - 78% compared to 63%).

Limitations and Future Directions

Limitations

While our work to condition human motion diffusion models on fitness grades is novel and promising, limitations remain. One challenge in human motion generation is the low availability of diverse and comprehensive datasets. The FMS dataset has such limitations:

- The dataset focuses on seven specific FMS movements collected from 45 participants, which does not capture the full diversity of human movements necessary for broader applications in accurate motion generation. The range of motions, while adequate for FMS assessment, might not be comprehensive for training accurate diffusion models. The complexity and variety of human motion might require larger datasets for the model to generate fully accurate motions and capture subtle nuances
- The dataset lacks participant diversity, with data collected from 45 individuals aged 18-59 recruited on a university campus. Hence, the data may not be representative of the broader population's age, body types, and fitness levels. This limits the model's ability to generalise across different demographic groups.
- Each motion in the FMS dataset is annotated from 0 to 3 by FMS experts. While the use of experts promotes quality, this labelling is unavoidably subjective.
- Non-invasive depth cameras are used to collect this dataset, primarily consists of RGB images, depth images and skeleton data. This also may not capture all the nuances of human movements such as subtle variations in muscle engagement or balance, which are also important for accurate FMS scoring.

Given the complexity of human motions and the subtleties in FMS scoring, the model might "hallucinate" movements that do not align with actual human biomechanics or the intended FMS score. This is evident in a few of the generated motions. We also observe some common artefacts such as foot sliding and ground penetration in our generated samples.

Interpretability is another limitation. Diffusion models, like many deep learning models, often operate as "black boxes", making it difficult to understand how and why they generate specific motions. This lack of interpretability can be a limitation, especially in fields like healthcare where understanding the rationale behind a movement pattern is crucial.

Future Directions

Future work can experiment with more advanced diffusion models, such as PhysDiff Yuan et al. (2023), which explicitly attempt to correct for such pathologies. PhysDiff reduces the prevalence of such artefacts using a physics-guided motion engine to denoise the model's predictions. Their improvements offers better performance and out-of-the box easier fine-tuning, making it a suitable option for further experiments on FMS-conditioned motion generation. Human motion is non-linear and subject to physical and biomechanical constraints. Accounting for these aspects could ensure that generated motions are not only visually plausible but also biomechanically correct.

Future experiments can also explore alternate methods of encoding the FMS scores for conditioning the diffusion model. While we focused on "textual" and "numerical" representations of the FMS labels, another possibility is to directly encode the FMS score in the conditioning token. Future work could involve collecting fitness data using a wider variety of data modalities. Another way to address the scarcity of FMS motion data is to perform mirroring of the current FMS movements. This approach is already effectively applied in the HumanML3D dataset Guo et al. (2022) and can be used to double the size of the FMS dataset.

Potential Applications

We develop a fitness-aware motion generation method capable of generating human movements consistent with a given level of athleticism or injury severity, in addition to action type. Such a method would be beneficial for various domains, including healthcare and sports. We anticipate broad applications of such a fitness-aware motion model. In medicine and physiotherapy, quality of movement, reflecting the extent to which a human is physically able, is central to rehabilitation and diagnosis. Generating fitness-aware motions can enrich the training of medical students and provide useful references for both patient and doctor during the recovery process. Another application in sports education is to generate diverse action samples graded by quality. These samples can similarly provide references and counterexamples to assist in the development of technique and reduce the risk of strain. Finally, character animations in gaming and virtual reality can significantly enhance their vividness and authenticity by controlling details of fitness and physical ability.

Bibliography

Guo, C. et al. (2020). Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029.

Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., and Cheng, L. (2022). Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Neverhood, J. and Xiaosong, C. (2021). Stop motion obj. `https://github.com/neverhood311/Stop-motion-OBJ`.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A., Tzionas, D., and Black, M. J. (2019). Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985.

Plappert, M., Mandery, C., and Asfour, T. (2016). The KIT motion-language dataset. *Big Data*, 4(4):236–252.

Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., and Bermano, A. H. (2023). Human motion diffusion model. In *International Conference on Learning Representations*.

Xing, Q.-J., Shen, Y.-Y., Cao, R., Zong, S.-X., Zhao, S.-X., and Shen, Y.-F. (2022). Functional movement screen dataset collected with two azure kinect depth sensors. *Scientific Data*, 9(1):104.

Yuan, Y., Song, J., Iqbal, U., Vahdat, A., and Kautz, J. (2023). Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16010–16021.