Extending Text-Driven 3D Human Motion Generation with Diffusion Models using Data-Augmented Training

Félix Fourreau **Mathis Wauquiez**
École des Ponts ParisTech (ENPC)
ENS Paris-Saclay
{felix.fourreau, mathis.wauquiez}@eleves.enpc.fr

January 2025

Abstract

The paper we worked on explores text-driven 3D human motion generation using diffusion-based models that operate directly on Skinned Multi-Person Linear Model (SMPL) representations of the human body. However, unlike *the timeline-based approach proposed in the STMC paper, we worked on a* ***single-prompt*** *setup for motion generation while keeping the MDM-SMPL architecture from the paper. We also integrate the* ***data augmentation*** *strategies proposed in the TMR++ study.*

We used three popular datasets (HumanML3D, KIT-ML, and BABEL), and evaluated both pretrained diffusion models on HumanML3D from the STMC paper and custom-trained *ones under various augmentation setups. Although the pretrained STMC model does not use data augmentation we did not manage to outperform it using our data augmentation. We also trained a* ***ConvUNet1D with attention*** *denoiser instead of MDM-SMPL's transformer encoder based denoiser, in hope of faster generation and better denoising, but unfortunately our architecture only yielded poor denoising performance, as compared to the denoiser of MDM-SMPL. Furthemore, our findings conclude that text augmentations from TMR++ could benefit motion generation, evaluated on the original datasets.*

1. Introduction

Generating 3D human motion from textual prompts can have significant impacts on making 3D animation more accessible and less time consuming, with applications such as content creation for games, VR experiences, and film.

Our Framework. We used the *diffusion-based* MDM-SMPL approach proposed in STMC [14], but we used a *single text prompt* input instead of the original timeline test input from the original paper. To evaluate the models, we run experiments on three major text-motion datasets: **HumanML3D** [4], **KIT-ML** [4], and **BABEL** [16], comparing the original pretrained model (on 10 000 epochs) against newly trained models (with less epochs) under various data augmentation setups. We also tested our own *ConvUNet1D with attention* architecture but observed no improvement over MDM-SMPL.

2. Related Work

In the past decades, numerous motion generation techniques have been developped. Given the nature of the problem, most approaches are generative. To our knowledge, most popular generative methods have been adapted for human motion generation. For instance, HP-GAN [1] is an example of Generative Adverserial Networks (GANs) (more precisely, improved Wasserstein GANs) for pose generation, conditioned by past poses. Normalizing Flows [18] have been explored as well, with MoGlow [5] for example, which introduces probabilistic, generative, controlable models for motion data. Due to their superior results and training stability, the research community quickly turned to Variational AutoEncoders (VAEs). A lot of work have been done in this direction, including ACTOR [11], that achieves variable-length motion generation conditioned on categorical actions through a class conditioned VAE. TEMOS [12] is a similar VAE-based method that allows motion generation through the use of a joint (text, motion) latent space. Sampling motions from this joint latent space is done by encoding the text to the joint latent space, sampling from this latent space, and then decoding it to a motion. Another interesting and follow-up work is TMR[13], that leverages the same methodology as TEMOS, plus text-motion embeddings contrastive alignment inspired by CoCa [22], using the InfoNCE loss [21]. More recently, diffusion models showed to excel for human motion generation. Among diffusion-based methods, Motion Diffusion Model (MDM)

[20] was one of the first to demonstrate the performance of diffusion models for motion generation, concurrently to MotionDiffuse [23]. Motion Latent-based Diffusion model (MLD) introduces diffusion in a low-dimensional latent space through the usage of a VAE. StableMoFusion [8] compares different denoiser model architectures, stating the superiority of their CondUNet1D against the Diffusion Transformer (DiT) structure for motion generation, achieves footskate reduction by integrating an adequate loss function and enables efficient inference through the use of the DPM-Solver++ sampler [9].

3. Methodology

MDM-SMPL. The MDM-SMPL model uses the Denoising Diffusion Probabilistic Model (DDPM) [7] framework. The denoiser used is a transformer, operating on human poses in SMPL format, $\boldsymbol{\theta} \in \mathbb{R}^{72}$. Each motion sequence can be written as $x = (x^1, \ldots, x^N)$, with N the number of frames. Diffusion models involves three processes: the forward process, were we gradually noise an input sequence x_0 by progressively adding Gaussian noise to it, the reverse process, were we learn to denoise the T noising steps and therefore train a denoiser adapted to a wide range of noise values, and the inference process, used for sampling from our diffusion model, in which we sample Gaussian noise x_T, and gradually refine it to $x_{T-1}, ..., \hat{x}_0$.

Mathematically, the forward process can be represented by these equations:

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 \; + \; \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{1}$$

The values $\bar{\alpha}_t$ are defined by the noise schedule, and such that $\bar{\alpha}_T = \prod_{i=1}^{T} \alpha_t \approx 0$. We used a cosine noise schedule, as proposed by [3].

Then, according to the DDPM framework, the reverse process objective is to train our denoiser $\hat{\boldsymbol{G}}_{\boldsymbol{\theta}}$ minimizing the simplified loss function:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{\epsilon}, t, \boldsymbol{x}_0, C} \left\| \hat{\boldsymbol{G}}_{\boldsymbol{\theta}} \left(\boldsymbol{x}_t, t, C \right) - \boldsymbol{x}_0 \right\|_2^2 \tag{2}$$

Unlike STMC's *multi-track* approach, we use a **single text prompt** C at test time. More precisely, C is the embeddings gotten from a frozen ViT-B/32 CLIP sentence transformer [17] applied to the textual motion description. In order to perform classifier-free guidance [6], C is set to 0 10% of the time, so that $\hat{\boldsymbol{G}}_{\boldsymbol{\theta}}$ learns both conditioned and unconditioned denoising.

As said, during the inference process, for every time step t, we predict the clean sample $\hat{x}_0 = \hat{\boldsymbol{G}}_{\boldsymbol{\theta}} \left(\boldsymbol{x}_t, t, C \right)$ from x_t, and then noise it back to x_{t-1}, until we reach $t = 0$.

More exactly, we define $\hat{x}_0$ through $\hat{\boldsymbol{G}}_{\boldsymbol{\theta}}^{s} \left(\boldsymbol{x}_t, t, C \right)_s$, with:

$$\hat{\boldsymbol{G}}_{\boldsymbol{\theta}}^{s} \left(\boldsymbol{x}_t, t, C \right) = \hat{\boldsymbol{G}}_{\boldsymbol{\theta}} \left(\boldsymbol{x}_t, t, 0 \right) + s . (\hat{\boldsymbol{G}}_{\boldsymbol{\theta}} \left(\boldsymbol{x}_t, t, C \right) - \hat{\boldsymbol{G}}_{\boldsymbol{\theta}} \left(\boldsymbol{x}_t, t, 0 \right))$$

This allows diversity and fidelity trade-off by manipulating the value of s. In the following experiments, s was set to 2.5. We also either *adopt a pretrained* G_θ (from STMC's code release) or *train our own* on selected datasets.

TMR++ Data Augmentations. We used the text augmentation pipeline proposed in TMR++ [2], which:

- Takes original textual descriptions and paraphrases them using large language models (e.g., Llama-2).
- Converts the original textual descriptions into short action-style keywords (e.g., "run" vs. "A person is running forward").
- Samples textual embeddings from either the real annotations, action-style descriptions, paraphrased texts, or from average embeddings of some paraphrases during training time, to improve textual diversity. The respective probabilities for the textual embeddings calculation are: $(p_{\text{gt}}, p_{\text{act}}, p_{\text{par}}, p_{\text{avg}}) = (0.4, 0.1, 0.2, 0.3)$

Though TMR++ primarily studies cross-dataset retrieval, we only exploit the resulting *augmented text annotations* to strengthen our single-prompt motion generation model.

ConvUNet1D with Attention. We also implement a *ConvUNet1D* architecture inspired by 2D image-diffusion [19] and incorporate temporal self-attention blocks. The hope was to improve the performance of the transformer-based MDM-SMPL. However, as shown in Section 4, this attempt did not yield improvement in metrics such as TMR-Score or recall.

4. Experiments

We conducted experiments using three annotated datasets to evaluate the performance of the model:

- **HumanML3D** [4]: This dataset contains 14,000 motions, each paired with multiple verbose textual descriptions, covering a wide range of actions and movements.
- **KIT-ML** [15]: A smaller dataset primarily focused on locomotion, with descriptive sentences that provide context for the annotated motions.
- **BABEL** [16]: This dataset often uses concise, imperative-style motion labels, such as "sit" or "wave hand," offering a distinct annotation style.

Each of these annotations is linked to a 3D motion representation from the AMASS dataset [10], a comprehensive database that unifies various motion capture datasets within a common framework and parameterization.

The baseline for comparison was the MDM-STMC pretrained model provided in the STMC GitHub repository. We evaluated its performance on the three datasets, obtaining the following FID scores:

The pretrained model performed exceptionally well on HumanML3D, achieving a remarkably low FID score. However, its performance on KIT-ML and BABEL datasets was significantly worse, with FID scores more than double that

Table 1. FID Scores for Pretrained Model on Different Datasets

Dataset	FID Score
HumanML3D	0.062
KIT-ML	0.105
BABEL	0.236

of HumanML3D. This highlights the model's limited generalization capabilities and underscores the importance of data augmentation to improve cross-dataset performance.

Regarding training efficiency, it is important to note that the pretrained model was trained for 10,000 epochs, which was beyond our computational resources. Based on the loss curves shown in Figure 1, we determined that training for 100 epochs provided a good balance between performance and training time.

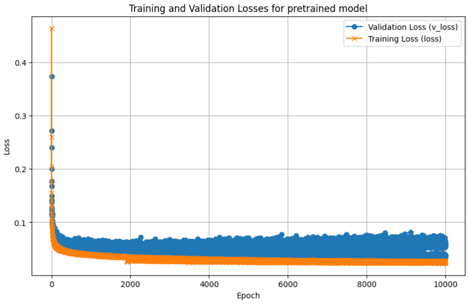

Figure 1. Loss of the pretrained model over 10,000 epochs.

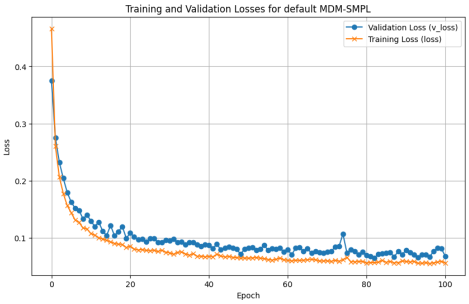

Figure 2. Loss of a model trained on HumanML3D for 100 epochs.

To establish a new baseline, we retrained the MDM-SMPL model on HumanML3D annotations for 100 epochs. The results are shown below:

Further experiments involved training the model on KIT-ML, HumanML3D + KIT-ML, and HumanML3D + KIT-ML with data augmentation, all for 100 epochs. The results across datasets are summarized in Table 3.

The model trained solely on KIT-ML performed poorly across all datasets due to its limited size and diversity.

Table 2. FID Scores for Model Trained on HumanML3D

Dataset	FID Score
HumanML3D	0.322
KIT-ML	0.425
BABEL	0.448

Merging HumanML3D and KIT-ML significantly improved performance, and the addition of data augmentation further boosted results. Initially, augmented models underperformed due to incorrect implementation of probability-based training described in TMR++. After correcting this, we adopted the augmentation strategy with predefined probabilities $(p_{\text{gt}}, p_{\text{act}}, p_{\text{par}}, p_{\text{avg}}) = (0.4, 0.1, 0.2, 0.3)$, which prioritized original data while allowing diversity from augmented datas.

Table 4. FID Scores for HumanML3D + KIT-ML (Augmented) with and Without Probability Training.

Dataset	No Probability Training	With Probability Trainin
HumanML3D	0.340	0.265
KIT-ML	0.358	0.274
BABEL	N/A	0.344

While further optimization of probability parameters could enhance performance, these results already demonstrate the significant impact of data augmentation on improving model generalization.

As an exploratory improvement, we experimented with replacing MDM-SMPL's Transformer encoder-based denoiser with a ConvUNet1D architecture enhanced with attention mechanisms. This approach was intended to achieve faster generation times and potentially more effective denoising. However, the training process for this architecture failed to converge as effectively as the original MDM-SMPL model. Consequently, we abandoned the idea after initial attempts:

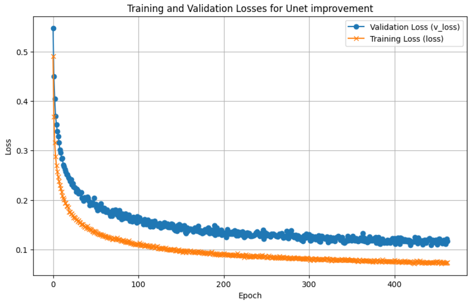

Figure 3. Training performance comparison of ConvUNet1D with attention versus MDM-SMPL.

Table 3. FID Scores Across Datasets for Models Trained with Different Configurations (100 Epochs).

Training Configuration	HumanML3D (FID)	KITML (FID)	BABEL (FID)
Model trained on KIT-ML	1.1599	1.3290	1.2737
Model trained on HumanML3D + KIT-ML	0.3057	0.3459	0.4349
Model trained on HumanML3D + KIT-ML (Augmented)	**0.2649**	**0.2742**	**0.3444**

5. Conclusion

In this work, we extended the MDM-SMPL framework for text-driven 3D human motion generation by integrating data augmentation strategies from TMR++. Our key findings are as follows:

- Model Train on simple datasets: The model trained without data augmentation achieved good scores on their datasets of training but struggled to generalize to other form of annotation.
- Impact of Data Augmentation: Incorporating data augmentation improved the performance of custom-trained models across all evaluated datasets. However, it could still be improved by tuning the probabilities parameters and also by training with more epochs .
- Architectural Exploration: The introduction of a ConvUNet1D architecture with attention mechanisms did not enhance performance compared to the original MDM-SMPL model, suggesting that further architectural innovations are necessary.
- Generalization Challenges: Significant disparities in performance across different datasets emphasize the ongoing challenge of creating models that generalize well beyond their training data.

Overall, while data augmentation contributed to better generalization, it still could benefits from better training setups. It Future work will explore more advanced augmentation techniques, alternative architectures, and optimized training protocols to further enhance the capabilities of text-driven 3D human motion generation.

References

[1] Emad Barsoum, John R. Kender, and Zicheng Liu. HP-GAN: probabilistic 3d human motion prediction via GAN. *CoRR*, abs/1711.09561, 2017. 1

[2] Léore Bensabath, Mathis Petrovich, and Gül Varol. A cross-dataset study for text-based 3d human motion retrieval, 2024. 2

[3] Ting Chen. On the importance of noise scheduling for diffusion models, 2023. 2

[4] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, 2022. 1, 2

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

[6] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 2

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 2

[8] Yiheng Huang, Hui Yang, Chuanchen Luo, Yuxi Wang, Shibiao Xu, Zhaoxiang Zhang, Man Zhang, and Junran Peng. Stablemofusion: Towards robust and efficient diffusion-based motion generation framework, 2024. 2

[9] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models, 2023. 2

[10] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[11] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae, 2021. 1

[12] Mathis Petrovich, Michael J. Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions, 2022. 1

[13] Mathis Petrovich, Michael J. Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis, 2023. 1

[14] Mathis Petrovich, Or Litany, Umar Iqbal, Michael J. Black, Gül Varol, Xue Bin Peng, and Davis Rempe. Multi-track timeline control for text-driven 3d human motion generation, 2024. 1

[15] Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. *Big Data*, 4(4):236–252, 2016. 2

[16] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with english labels. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 722–731, 2021. 1, 2

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2

[18] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows, 2016. 1

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 2

[20] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model, 2022. 2

[21] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 1

[22] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022. 1

[23] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model, 2022. 2