# UniMoGen: Universal Motion Generation

Aliasghar Khani
Autodesk Research
Canada
aliasghar.khani@autodesk.com

Arianna Rampini
Autodesk Research
Canada
arianna.rampini@autodesk.com

Evan Atherton
Autodesk Research
Canada

Bruno Roy
Autodesk Research
Canada

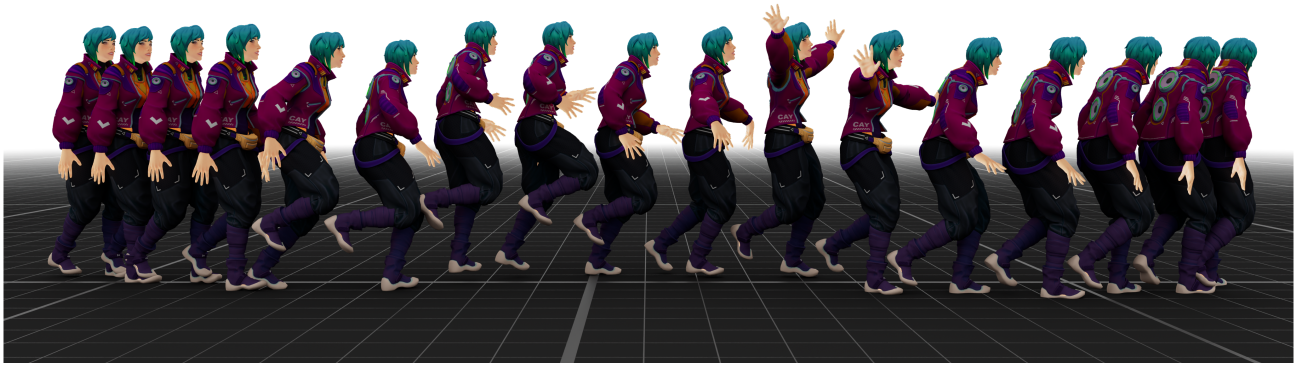

**Figure 1: UniMoGen generates realistic and diverse character motions in real time, controllable via action type, trajectory, and past motion context. It supports arbitrary skeleton topologies by operating in a skeleton-agnostic manner, and can produce long, smooth motion sequences that transition seamlessly across different styles. The figure shows a sample motion sequence generated by UniMoGen.**

## Abstract

Motion generation is a cornerstone of computer graphics, animation, gaming, and robotics, enabling the creation of realistic and varied character movements. A significant limitation of existing methods is their reliance on specific skeletal structures, which restricts their versatility across different characters. To overcome this, we introduce UniMoGen, a novel UNet-based diffusion model designed for skeleton-agnostic motion generation. UniMoGen can be trained on motion data from diverse characters, such as humans and animals, without the need for a predefined maximum number of joints. By dynamically processing only the necessary joints for each character, our model achieves both skeleton agnosticism and computational efficiency. Key features of UniMoGen include controllability via style and trajectory inputs, and the ability to continue motions from past frames. We demonstrate UniMoGen 's effectiveness on the 100STYLE dataset, where it outperforms state-of-the-art methods in diverse character motion generation. Furthermore, when trained on both the 100STYLE and LAFAN1 datasets, which use different skeletons, UniMoGen achieves high performance and improved efficiency across both skeletons. These results highlight UniMoGen's potential to advance motion generation by providing a flexible, efficient, and controllable solution for a wide range of character animations.

## CCS Concepts

• **Computing methodologies → Motion processing**.

## Keywords

Motion Generation, Diffusion models, Unet, skeleton-agnostic

## 1 Introduction

The generation of realistic and diverse character motions is essential in various domains, including computer graphics, animation, gaming, and robotics. Motion generation enables the creation of lifelike animations that enhance user experiences in films, video games, virtual reality, and robotic simulations [Holden et al. 2016]. Previous research has demonstrated significant progress in data-driven approaches to motion generation [Chen et al. 2024; Guo et al. 2025; Li et al. 2024; Ling et al. 2024; Tevet et al. 2025, 2023; Zhao et al. 2024; Zhu et al. 2023]. However, these techniques are often tailored to specific skeletal structures, limiting their applicability to characters with different topologies. This presents a major challenge in developing a universal model capable of generating motion for a wide range of characters, such as humans, animals, and fantastical creatures, each with distinct skeletal configurations.

Recent advancements in motion generation have aimed to address the challenge of producing animations using diffusion and auto-regressive models, but limitations remain. For instance, MDM [Tevet et al. 2023] introduced the first motion diffusion model conditioned on text input. While pioneering, it does not incorporate trajectory information for controllability or utilize past motion frames for auto-regressive generation. Building on this, CAMDM [Chen et al. 2024] employs an auto-regressive diffusion framework with a transformer-based architecture to generate high-quality motions based on user control signals and prior motion, achieving real-time performance. Although this method improves controllability by leveraging trajectory and past motion, it passes all inputs and conditions into the transformer at once. This results in unnecessarily long input sequences, leading to increased memory usage and slower generation times. MotionLLaMA [Ling et al. 2024], on the other hand, proposes a transformer-based auto-regressive model that tokenizes motion sequences and, given text or audio, generates motion sequences through next-token prediction. While this approach is compatible with language modeling frameworks, applying tokenization to continuous motion data, which is highly sensitive to small value changes, can lead to subtle but important information loss [Li et al. 2025]. This degradation in precision negatively affects the quality of the generated motions. In addition to these limitations, like many other methods, these methods are designed for a single skeletal structure and require separate training for each distinct skeleton, restricting their generalizability. In contrast, AnyTop [Gat et al. 2025] introduces a diffusion model capable of generating motions for arbitrary skeletons by integrating topology information into a transformer-based denoising network. However, it requires specifying a maximum number of joints in the skeletons; if a skeleton has fewer joints, the model pads them, resulting in unnecessary computational and time overhead.

To overcome these challenges, we present UniMoGen, a novel approach to motion generation that is inherently skeleton-agnostic. UniMoGen is built upon a UNet-based diffusion model with attention modules. The UNet architecture enhances efficiency by first downsampling the motion sequence in the temporal dimension and applying attention modules to the shorter sequences. Additionally, the attention modules enable UniMoGen to handle motion data from characters with varying numbers of joints without requiring padding or fixed skeletal templates. By temporally downsampling the motion sequence and processing only the relevant joints for each character, UniMoGen achieves both skeleton agnosticism and computational efficiency, making it suitable for large-scale applications.

This work introduces several key contributions that advance the field of motion generation:

- **Skeleton-Agnostic Architecture:** UniMoGen is the first model to seamlessly handle arbitrary skeletal structures without padding or fixed joint counts, enabling simultaneous training on diverse characters, such as humans and animals, and setting a new standard for universal motion generation.
- **Efficient and Controllable Motion Synthesis:** By leveraging a UNet-based diffusion model with temporal downsampling and attention mechanisms, UniMoGen achieves high computational efficiency while offering fine-grained control through style and trajectory inputs, as well as the ability to continue motion sequences from past frames.
- **Real-Time Generation:** UniMoGen supports real-time motion synthesis, generating motions in just 0.09 seconds on a GPU.

We evaluate UniMoGen on the 100STYLE [Mason et al. 2022] dataset, a comprehensive collection of stylized human locomotion data encompassing 100 different styles, such as walking, running, and sidestepping. In this benchmark, our method outperforms MDM and CAMDM, demonstrating its superior ability to generate diverse and high-quality motions. For example, compared to CAMDM, UniMoGen reduces the percentage of frames with foot penetration from 4.73% to 0.3% on average for both left and right feet, and decreases the average foot sliding distance from 0.98 to 0.56.

To further test its scalability, we train UniMoGen on a combination of the 100STYLE and LAFAN1 [Harvey et al. 2020] datasets, which provide a broad spectrum of human actions, including daily activities like walking, running, and sitting down. In this more comprehensive setting, UniMoGen not only achieves better performance than AnyTop but also does so with improved efficiency, highlighting its robustness across different datasets and its potential for real-world applications. For instance, our method achieves an average foot penetration percentage of 11.05% (across both feet), significantly lower than AnyTop's 26.41%.

The remainder of this paper is structured as follows: Section 2 provides an overview of related work in motion generation. Section 3 delves into the architectural details of UniMoGen, explaining how it achieves skeleton agnosticism and efficiency. Section 4 describes the datasets used, the experimental methodology, and results comparing UniMoGen with baseline methods. Finally, Section 5 concludes the paper and outlines potential avenues for future research.

## 2 Related work

### 2.1 Motion Generation with Diffusion Models

Diffusion models have emerged as a powerful framework for motion generation, leveraging their ability to produce high-quality, diverse samples. For instance, MDM [Tevet et al. 2023] was a pioneering work that adapted diffusion models for motion synthesis, generating sequences from text or style inputs without auto-regression. However, it lacks the temporal continuity typically provided by auto-regressive methods. Building on this direction, FlowMDM [Barquero et al. 2024] introduces a transformer-based bidirectional diffusion model that generates long, smooth, and realistic human motion sequences conditioned on multiple textual descriptions. By combining a bidirectional Transformer, blended positional encodings, and pose-centric cross-attention, it effectively captures both past and future motion dependencies, enabling seamless transitions and eliminating the need for post-processing. Complementary to these text-driven approaches, CAMDM [Chen et al. 2024] focuses on real-time motion generation using motion diffusion probabilistic models. It enables high-quality and diverse character animations in response to dynamic user-supplied control signals. A notable contribution of CAMDM is its support for real-time interactive control. Further enhancing controllability and realism, DART [Zhao et al. 2024] introduces a diffusion-based auto-regressive model that generates long human motion sequences in real time, conditioned on both text and motion history. Operating in a learned latent motion primitive space, DART supports continuous text-driven generation and allows fine-grained spatial control—such as reaching target poses or navigating to specific locations via latent noise optimization and reinforcement learning.

In contrast to our method, all these methods are designed to work with a single skeleton each time and cannot be used to train on a dataset with different skeletons. In addition to that, our method uses 1D convolutions to downsample and upsample the data in the temporal dimension, which helps in reducing attention costs. However, as these methods use transformer architecture, they keep the number of frames untouched and all attention operations are conducted at the original frame length.

## 2.2 Auto-Regressive Motion Generation

Numerous works have explored auto-regressive models that follow the next-token prediction paradigm for motion generation. For example, T2M-GPT [Zhang et al. 2023] employs a Vector Quantized-Variational AutoEncoder (VQ-VAE) to discretize motion sequences into code indices, and a GPT-like model to perform auto-regressive next-index prediction conditioned on previous indices and a text description. Similarly, LaMP [Li et al. 2024] also uses a VQ-VAE to encode motion but adopts a masked prediction strategy instead of standard auto-regression. During inference, LaMP performs iterative masked prediction: it begins with a completely masked sequence, estimates distributions for the masked tokens, samples tokens, and re-masks low-confidence tokens over multiple steps. The generation process is conditioned by motion-informative text features extracted using LaMP's pre-trained text transformer, replacing the commonly used CLIP embeddings. Extending this direction, Motion-LLaMA [Ling et al. 2024] leverages a Large Language Model (LLM) fine-tuned with LoRA to handle various motion-related tasks. It introduces the Holistic Motion (HoMi) tokenizer to convert continuous motion into discrete tokens and performs motion generation in a unified auto-regressive framework using the causal language model (LLaMA3.2-Instruct), predicting the next motion token based on past tokens and conditioning signals such as text or audio.

Like motion diffusion models and unlike our method, all these methods are designed to work with only one skeleton. Moreover, training a tokenizer for the motion data, which is a continuous one and very sensitive to small variations in values, is very challenging and will degrade the quality [Li et al. 2025].

## 2.3 Skeleton-Agnostic Motion Generation

Addressing the long-standing challenge of generating motion for arbitrary skeletons, Gat et al. introduce AnyTop [Gat et al. 2025], a diffusion model designed to generate motions for diverse characters with distinct motion dynamics using only their skeletal structure as input. This work specifically tackles the problem of handling a wide variety of skeletal topologies, including skeletons which vary significantly in structure. AnyTop utilizes a transformer-based denoising network tailored for arbitrary skeleton learning, incorporating topology information and textual joint descriptions to learn semantic correspondences across diverse skeletons. A key design choice is embedding each joint independently at each frame, enabling greater flexibility compared to methods that embed the entire pose. The model demonstrates generalization to unseen skeletons and can produce natural motions for a range of character types like bipeds, quadrupeds, and multi-legged creatures. AnyTop stands out as a skeletal-based approach capable of generating smooth motions on a diverse range of characters using a single unified model without topology-specific adjustments.

However, a key limitation of AnyTop is its reliance on a predefined maximum number of joints. For skeletons with fewer joints, it pads the joint dimension with zeros, leading to unnecessary computational and memory overhead. Conversely, if a skeleton exceeds this joint limit, the model must discard the extra joints, resulting in a loss of valuable information.

In contrast, our method leverages a U-Net architecture with attention modules that process joints independently, eliminating padding and enabling efficient training on datasets with different skeletons, such as 100STYLE and LAFAN1. Our skeleton-agnostic design, combined with auto-regressive diffusion and trajectory conditioning, allows for flexible and high-quality motion generation across varied skeletal structures, addressing the computational and data loss issues inherent in methods like AnyTop.

# 3 Method

In UniMoGen, our goal is to train a skeleton-agnostic motion model that combines auto-regressive generation with diffusion-based training. This design enables the model to produce arbitrarily long motion sequences while maintaining high motion quality. For a high-level overview of our diffusion architecture, please refer to Figure 2.

## 3.1 Diffusion Models

Diffusion models are a class of generative models that learn to reverse a gradual noising process to generate data samples [Ho et al. 2020]. They operate by modeling a Markov chain that incrementally adds noise to data over a series of time steps, defined by a forward process $q(\mathbf{x}_t|\mathbf{x}_{t-1})$. This process transforms the original data distribution $\mathbf{x}_0 \sim p_{\text{data}}$ into a noise distribution, typically Gaussian, at the final time step $T$. The reverse process, parameterized by $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$, is learned to denoise the data step-by-step, starting from pure noise to reconstruct samples resembling the training data. One of the possible training objectives minimizes the difference between the clean and predicted data, often using a simplified mean-squared error loss:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0,\epsilon,t}\left[\|x_0 - \hat{x}_0(\mathbf{x}_t, t)\|^2\right],$$

where $x$ is the true noise, and $\hat{x}_0$ is the model's prediction. This framework has shown remarkable success in generating high-quality samples across various domains, including images [Ho et al. 2020; Rombach et al. 2022] and time-series data [Chen et al. 2024; Tevet et al. 2023; Yang et al. 2024], due to its stable training dynamics and ability to capture complex data distributions.

## 3.2 Universal Motion Generation

We propose a novel auto-regressive diffusion model for motion generation, designed to be agnostic to skeleton structures, enabling simultaneous training across diverse skeleton types. Our model, referred to as UniMoGen, takes as input a style index, a diffusion time step, a trajectory, and optionally, past frames. The style index is selected from a predefined set of styles. The trajectory consists of $F$ motion frames, including position and root rotation trajectories. The past frames, when provided, contain $F'$ motion frames comprising root positions and joint rotations. The model then outputs joint rotations and root positions for $F$ motion frames.

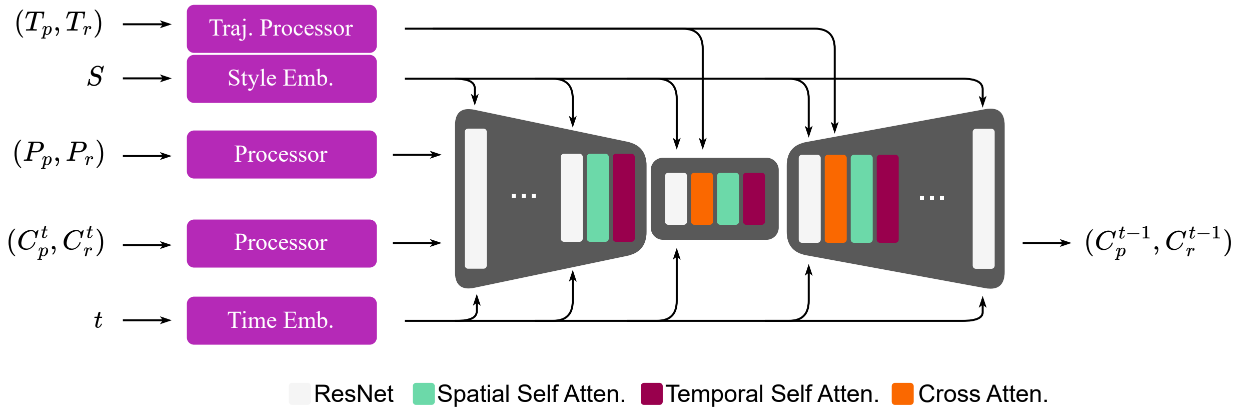

**Figure 2: Overview of the UniMoGen denoising architecture. During training, the model receives style index $S$, past motion inputs as root positions $P_p$ and joint rotations $P_r$, trajectory $(T_p, T_r)$, and diffusion time step $t$. Dedicated modules process each input, and their representations are fused in a UNet-based diffusion network. The network leverages temporal and joint-level self-attention, cross-attention to inject trajectory information, and Feature-wise Linear Modulation (FiLM) [Perez et al. 2018] to condition on time and style. The model outputs future motion $(C_p, C_r)$, enabling controllable, skeleton-agnostic generation across diverse characters. As illustrated in the figure, we omit attention modules in the first and last layers of the UNet and apply them only to the downsampled layers to reduce memory consumption.**

The architecture of UniMoGen is based on a U-Net with 1D convolutions along the temporal dimension. In addition, it incorporates attention modules across both temporal and joint dimensions, cross-attention modules to inject trajectory information, and Feature-wise Linear Modulation (FiLM) [Perez et al. 2018] to condition on time and style. FiLM works by passing the time step and style index through linear layers to get scale and shift parameters. These are then used to modulate the normalized features following Group Normalization [Wu and He 2018], allowing the model to dynamically adapt its behavior based on the temporal and stylistic context.

In the joint attention module, we compute an attention mask based on the skeleton topology, restricting each joint to attend only to its ancestors, thus preserving kinematic constraints. Following the U-Net paradigm, the encoder downsamples the temporal dimension, and the decoder upsamples it, allowing the temporal attention module to operate on shorter sequences, which reduces computational overhead.

During training, the model receives the following inputs: style index $S \in \mathbb{R}$, time step $t \in \mathbb{R}$, position trajectory $T_p \in \mathbb{R}^{F \times 2}$, root rotation trajectory $T_r \in \mathbb{R}^{F \times 6}$, past root positions $P_p \in \mathbb{R}^{F' \times 3}$, past joint rotations $P_r \in \mathbb{R}^{F' \times J \times 6}$, current root positions $C_p \in \mathbb{R}^{F \times 3}$, and current joint rotations $C_r \in \mathbb{R}^{F \times J \times 6}$, where $J$ is the number of joints of a skeleton. To train the model, Gaussian noise is added to $C_p$ and $C_r$ according to the diffusion schedule. These noisy versions are then concatenated with $P_p$ and $P_r$, respectively, and the model is trained to denoise them. Conditioning is applied in two ways: the model uses cross-attention to incorporate information from the trajectory inputs $T_p$ and $T_r$, while FiLM layers condition the model on the style index $S$ and time step $S$.

As seen in the representations of $P_r$ and $C_r$, the joints are kept as a separate dimension and are processed by the joint-wise attention mechanism that supports variable-length inputs (i.e., varying numbers of joints) and facilitates information sharing across joints. This design enables the model to accommodate skeletons with different joint counts, eliminating the need for padding and avoiding unnecessary computational overhead, thereby ensuring efficient handling of diverse skeletal structures.

The loss functions used during training include the diffusion loss $\mathcal{L}_{\text{d}}$; mean squared error (MSE) between the predicted and ground truth angular velocity of joint rotations, denoted as $\mathcal{L}_{\text{av}}$; MSE between the ground truth and predicted global position of joints, $\mathcal{L}_{\text{gp}}$; and MSE between the ground truth and predicted velocity of global position of joints, $\mathcal{L}_{\text{vgp}}$. In addition, we include a foot contact loss $\mathcal{L}_{\text{foot}}$ , defined as the $L_2$ norm of the predicted global velocity of toe joints on frames where those joints are in contact with the ground. In the ablation study section, we demonstrate the effectiveness of combining these auxiliary losses with the original diffusion loss.

Finally, to support auto-regressive generation, UniMoGen reuses the last $F'$ frames of the generated motion as the past frames for the next generation step. This mechanism enables the model to produce temporally coherent motion sequences of arbitrary length by chaining together successive predictions.

## 3.3 Implementation Details

Our U-Net architecture consists of three layers in both the encoder and decoder. Each layer doubles the number of feature channels and reduces the temporal resolution by half, except for the final encoder layer and the first decoder layer, which preserve the temporal resolution.

During training, we drop $S$ with a probability of 10% to enable Classifier-Free Guidance [Ho and Salimans 2021] during inference. Additionally, we drop $P_p$ and $P_r$ with a probability of 50% to allow the model to learn both to generate motion from scratch, without any past context, and to continue an existing motion sequence when past frames are provided. Furthermore, we apply a Gaussian filter to the trajectory positions at random and occasionally rotate the entire motion path to encourage robustness and invariance to trajectory transformations.

For the diffusion process, we adopt a cosine beta scheduler with 50 steps for the DDPM training phase and 4 steps for DDIM during inference, balancing quality and efficiency. Optimization is performed using the Adam optimizer with a learning rate of $1 \times 10^{-4}$ , along with an exponential learning rate decay where the decay factor (gamma) is set to 0.9999. We trained both experiments on $8 \times$ H100 GPUs: 34K steps for the 100Style dataset and 164K steps for the combined 100Style and LAFAN1 datasets.

## 4 Experiments

In this section, we evaluate the performance of UniMoGen through a series of experiments designed to assess its motion generation quality, physical plausibility, and skeleton-agnostic capabilities. We describe the datasets used, the evaluation metrics, the baseline methods for comparison, quantitative and qualitative results, and an ablation study to analyze key design choices.

### 4.1 Dataset

We evaluate our method using two diverse motion capture datasets, 100style [Mason et al. 2022] and LAFAN1 [Mason et al. 2022], each characterized by distinct skeleton structures (i.e., a single skeleton type per dataset). The 100style dataset comprises 1, 372 clips, totaling 4, 094, 607 frames, and encompasses 100 distinct styles. On the other hand, the LAFAN1 dataset includes 1540 clips with a total of 978, 844 frames and 15 styles. For both training and evaluation, we utilize these datasets in their original forms without retargeting, preserving their original skeleton configurations. For motion representation, we adopt the 6D rotation representation proposed by [Zhou et al. 2019] for joint rotations, along with 3D root positions. Prior to training, we apply min-max normalization to the root position data, scaling it to the range $[-1, 1]$. The rotation data, however, is left unnormalized since it already falls within the same range. For both datasets, we split the data into training, validation, and test sets with 75%, 15%, and 10% of the clips, respectively, ensuring that the style distribution is preserved across all splits. The validation set is used for ablation studies and selecting the best model checkpoint, while the test set is reserved for final comparisons with other methods.

### 4.2 Metrics

To demonstrate the effectiveness of UniMoGen, we employ a comprehensive set of evaluation metrics. First, we use the Fréchet Inception Distance (*FID*) to measure the distributional similarity between ground truth and generated motion sequences. Following prior work [Chen et al. 2024], we train a motion classifier on joint positions from the training set and use its feature activations to embed both real and generated motions. FID is then computed as the Fréchet distance between the resulting feature distributions. In addition to FID, we compute two diversity metrics designed to quantify variation in generated motion. The *Diversity (intra-motion)* measures the variance of each joint's spatial location over time within a single sequence, averaged across all joints and motions. The *Diversity (inter-motion)* measures the variance of joint positions across different motions, averaged across all joints.

To further evaluate realism, we compute *Foot Penetration* and *Foot Sliding*. Foot penetration quantifies the fraction of frames in which any of the toes (left or right) intersects the ground, indicating physical implausibility. The foot sliding distance is a critical metric for evaluating motion realism, which measures the moving distance (in meters) of the character's toes when the joint height is below a threshold (0.01 m), capturing unnatural sliding artifacts. Finally, we evaluate *Trajectory Distance*, which comprises two components: the position difference, measured as the distance between the trajectory and the root joint positions along the $x$ and $z$ axes; and the rotation difference, which compares the root joint's orientation with the trajectory's target rotation.

### 4.3 Baselines

We compare UniMoGen against several state-of-the-art approaches. First, we consider MDM [Tevet et al. 2023], the pioneering diffusion-based motion generation model that generates motion solely from text or style inputs and is not auto-regressive. Second, we include CAMDM [Chen et al. 2024], a transformer-based auto-regressive diffusion model that generates motion sequences conditioned on style, trajectory, and past motion, representing the current state-of-the-art in motion generation using trajectory, style, and past motion. Lastly, as UniMoGen is skeleton-agnostic, we compare it with AnyTop [Gat et al. 2025], a skeleton-agnostic motion generation method designed to handle diverse skeleton structures.

All baseline methods were trained until convergence. On the 100Style dataset, this required 400k steps for CAMDM, 324k for MDM, and 34k for our method. When training on the combination of 100Style [Mason et al. 2022] and LAFAN1 [Mason et al. 2022] datasets, convergence was reached after 176K steps for AnyTop and 164K steps for our method.

### 4.4 Results

To highlight the performance of UniMoGen, we present both quantitative and qualitative results. For a fair comparison with MDM, which only conditions on style, we generate 500 samples per style. Similarly, using our method, we generate 500 samples per style by randomly selecting (past motion, trajectory) pairs corresponding to that style from the test set. In contrast, for CAMDM, we use the entire test set and generate one sample for each (style, past motion, trajectory) pair using both UniMoGen and CAMDM. Table 1 compares UniMoGen with MDM and CAMDM on the 100style dataset across generation metrics (FID and diversity) and physical plausibility metrics (number of foot penetration frames and foot sliding distance).

The two top rows show that, despite operating with only 4 denoising steps (i.e., 250 times fewer steps than those required by MDM), UniMoGen achieves a significantly lower FID than MDM, reflecting a meaningful enhancement in distribution alignment and perceptual quality. Additionally, UniMoGen is conditioned on past frames and trajectory, which increases the complexity of the task:

satisfying multiple, potentially conflicting or difficult-to-model constraints simultaneously, such as style, precise past frames, a specific future trajectory, and maintaining physical plausibility, is more challenging than generating plausible motion with fewer constraints, as MDM does. Finally, the substantial improvement in diversity highlights that, unlike MDM which often produces static or repetitive motion, UniMoGen is able to generate a broader and more expressive set of styles while remaining faithful to the input signals. Meanwhile, UniMoGen outperforms CAMDM in diversity as well as all physical plausibility metrics, achieving lower foot penetration and reduced sliding, shown in the bottom two rows. For visual examples, refer to Figure 4.

Given that CAMDM is positioned as MDM's real-time counterpart, UniMoGen demonstrates a superior trade-off between quality and efficiency, enabling physically plausible motion synthesis in real-time, while additionally supporting skeleton-agnostic generation. Each motion generation by UniMoGen takes 4 seconds on a CPU and only 0.09 seconds on a GPU.

**Table 1: Comparison with MDM and CAMDM on the 100Style dataset. Compared to MDM, our model operates with 250x fewer inference steps and is more controllable by conditioning on past frames and trajectory. Despite these constraints, it outperforms MDM in terms of FID and diversity and achieves comparable overall results for foot penetration and sliding. Compared to CAMDM, our method shows superior performance across all metrics, with CAMDM having a slight advantage only in FID.**

| Method | FID ↓ | Diversity ↑ (intra-motion) | Diversity ↑ (inter-motion) | Left Pen. ↓ (Frames)% | Right Pen. ↓ (Frames)% | Ft. Slid. ↓ (m) |
|---|---|---|---|---|---|---|
| MDM | 2.64 | 0.026 | 0.083 | **0.12** | **0.15** | **0.41** |
| UniMoGen | **2.22** | **0.078** | **0.213** | 0.30 | 0.36 | 0.61 |
| CAMDM | **2.20** | 0.052 | 0.161 | 4.73 | 4.73 | 0.98 |
| UniMoGen | 2.24 | **0.078** | **0.213** | **0.26** | **0.35** | **0.56** |

**Table 2: Comparison of trajectory errors with CAMDM on the 100Style dataset. This table reports trajectory-following accuracy. Our method consistently outperforms CAMDM, demonstrating more precise adherence to the given trajectories.**

| Method | Mean Position Error (m) ↓ | Mean Rotation Error (deg) ↓ |
|---|---|---|
| CAMDM | 0.07 | 8.13 |
| UniMoGen | **0.01** | **6.99** |

A key advantage of UniMoGen is its ability to train on multiple skeleton types simultaneously, enabling the development of a large, universal model capable of handling diverse skeletal structures without modification. To evaluate this capability, we conduct a cross-dataset comparison between UniMoGen and AnyTop using a combined dataset consisting of 100Style and LAFAN1, two datasets with distinct skeleton types. Table 3 reports quantitative results for this comparison. As shown, UniMoGen outperforms AnyTop across all metrics, despite not using a text encoder to encode joint names. Furthermore, our method is more efficient than AnyTop as it avoids joint padding, which introduces unnecessary computational and time overhead. Instead, UniMoGen maintains joints as a separate dimension and leverages attention mechanisms to efficiently process varying skeletal structures. As shown in Figure 5, this approach enables UniMoGen to generate motions across different skeleton types. The left and right panels illustrate motions generated for the skeletons of the LAFAN1 and 100Style datasets, respectively.

For more qualitative results and comparisons, please refer to the supplemental video.

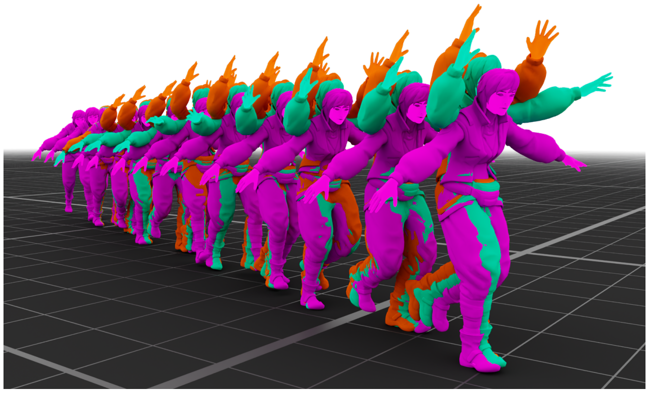

**Figure 3: Style blending with UniMoGen. Visualization of motions generated by blending two styles: Aeroplane and Arms Above Head. Purple shows 100% Aeroplane and 0% Arms Above Head, Green shows a blend of 35% Aeroplane and 65% Arms Above Head, and Orange shows 0% Aeroplane and 100% Arms Above Head. The smooth transition illustrates the expressive and continuous nature of the learned style space.**

**Table 3: Comparison with AnyTop on the combined 100Style and LAFAN1 datasets. The first two rows report results on 100Style, while the following rows correspond to LAFAN1. As the results indicate, UniMoGen consistently outperforms AnyTop across both datasets by a substantial margin.**

| Method | FID ↓ | Diversity ↑ (intra-motion) | Diversity ↑ (inter-motion) | Left Pen. ↓ (Frames)% | Right Pen. ↓ (Frames)% | Ft. Slid. ↓ (m) |
|---|---|---|---|---|---|---|
| AnyTop | 14.69 | 1.31e-5 | 3.96e-5 | 26.66 | 19.69 | 1.49 |
| UniMoGen | **2.191** | **0.08** | **0.23** | **12.64** | **10.90** | **1.21** |
| AnyTop | 4.197 | 7.42e-5 | 2.03e-4 | 19.75 | 33.06 | 1.81 |
| UniMoGen | **1.423** | **0.14** | **0.37** | **9.24** | **12.85** | **1.82** |

*Style Blending.* To further illustrate the flexibility of UniMoGen's style conditioning, we present a style blending experiment, where we interpolate between two style embeddings (e.g., 30% Style A and 70% Style B). As shown in Figure 3, the generated motions smoothly transition between the characteristics of both styles, demonstrating the continuous and expressive nature of the style embedding space learned by our model. Style blending animations examples are also included in the supplemental video.

## 4.5 Ablation Study

In this part, we conduct ablation studies to evaluate the impact of key design choices in UniMoGen, reporting FID, Penetration (Frames), and Sliding (m) on the validation set of the 100STYLE dataset. First, in Table 4, we assess the effect of min-max normalization on root positions, a preprocessing step to stabilize training by scaling data to a fixed range. An analysis of the raw data reveals that while all six components of the joint rotation representation fall within the range $[-1, 1]$, the X, Y, and Z components of the root position vary significantly in scale. Specifically, the ranges for X, Y, and Z are $[-3.52, 3.63]$, $[0.77, 1.21]$, and $[-2.91, 4.01]$, respectively, with the Y-axis exhibiting the smallest variation. This imbalance hinders effective learning of the Y-axis component, contributing to increased foot penetration errors. To mitigate this issue, we apply min-max normalization to the X, Y, and Z components of the root position and compare model performance with and without normalization to evaluate its role in ensuring stable convergence.

**Table 4: Min-Max Normalization of Root Positions. Normalizing root position values using min-max scaling leads to improved performance.**

| Configuration | FID ↓ | Left Pen. (Frames)% ↓ | Right Pen. (Frames)% ↓ | Sliding (m) ↓ |
|---|---|---|---|---|
| W Min-Max Norm | 2.31 | **3.47** | **3.75** | **0.53** |
| W/O Min-Max Norm | **2.26** | 7.57 | 8.75 | 0.60 |

Second, we evaluate the use of a cosine noise scheduler with fewer diffusion steps, to balance generation quality and computational efficiency, testing 50 steps against the standard 1000 steps. As shown in Table 5, we can achieve both better results and faster generations using cosine scheduler.

**Table 5: Cosine Scheduler with 100 Diffusion Steps. Employing a cosine scheduler enables effective training with only 100 diffusion steps, resulting in improved performance despite the reduced step count.**

| Configuration | FID ↓ | Left Pen. (Frames)% ↓ | Right Pen. (Frames)% ↓ | Sliding (m) ↓ |
|---|---|---|---|---|
| Cos. Sched. (100) | **2.24** | **0.99** | **1.32** | **0.49** |
| Lin. Sched. (1000) | 2.31 | 3.47 | 3.75 | 0.53 |

Third, we compare separate attention modules for spatial (joint) and temporal dimensions, which allow specialized feature processing, against a single attention module that merges both dimensions before processing. The results in Table 6 demonstrate the advantage of decoupled attention mechanisms.

**Table 6: Separate vs. Merged Attention Modules. Utilizing separate attention modules enhances both performance and computational efficiency compared to merged attention.**

| Configuration | FID ↓ | Left Pen. (Frames)% ↓ | Right Pen. (Frames)% ↓ | Sliding (m) ↓ |
|---|---|---|---|---|
| Separate Spatial/Temporal | **2.30** | **0.88** | **0.65** | **0.47** |
| Merged Attention | 2.31 | 3.47 | 3.75 | 0.53 |

Fourth, we investigate the inclusion of positional encoding, as used in transformer models [Vaswani et al. 2017], to capture positional relationships in both temporal and spatial (joints) dimensions, testing its impact on motion coherence. Table 7 shows that including positional encodings leads to improved performance.

**Table 7: Positional Encoding. As expected, incorporating positional encodings leads to improved performance in our model.**

| Configuration | FID ↓ | Left Pen. (Frames)% ↓ | Right Pen. (Frames)% ↓ | Sliding (m) ↓ |
|---|---|---|---|---|
| With Positional Encoding | **2.30** | **1.14** | **1.04** | **0.48** |
| Without Positional Encoding | 2.31 | 3.47 | 3.75 | 0.53 |

Fifth, we examine dataset balancing to ensure equitable representation of styles, mitigating bias toward overrepresented categories. This strategy yields improvements across metrics, as reported in Table 8.

**Table 8: Dataset Balancing. Balancing the dataset across styles through oversampling yields better motions.**

| Configuration | FID ↓ | Left Pen. (Frames)% ↓ | Right Pen. (Frames)% ↓ | Sliding (m) ↓ |
|---|---|---|---|---|
| Balanced Dataset | 2.33 | **1.98** | **2.21** | **0.52** |
| Unbalanced Dataset | **2.31** | 3.47 | 3.75 | 0.53 |

Finally, we analyze the role of auxiliary losses, which regularize training and enhance output quality. Table 9 confirms that including auxiliary losses improves the overall results.

**Table 9: Auxiliary Losses. Incorporating auxiliary losses significantly enhances the physical plausibility of the generated motions.**

| Configuration | FID ↓ | Left Pen. (Frames)% ↓ | Right Pen. (Frames)% ↓ | Sliding (m) ↓ |
|---|---|---|---|---|
| With Auxiliary Losses | 2.31 | **3.47** | **3.75** | **0.53** |
| Without Auxiliary Losses | **2.27** | 5.77 | 6.23 | 0.81 |

## 5 Conclusion

In this paper, we introduced UniMoGen, a novel skeleton-agnostic auto-regressive diffusion model for motion generation that addresses the limitations of existing methods in handling diverse skeletal structures and computational efficiency. By leveraging a U-Net architecture with 1D convolutions for temporal downsampling and upsampling, attention modules for joint and temporal dimensions, cross attention for conditioning on trajectory, and FiLM for conditioning on style and time, our model generates high-quality motion sequences conditioned on style, trajectory, and optional past frames. The use of attention masks based on skeleton topology ensures kinematic consistency, while processing joints in a separate dimension eliminates the need for padding, a common inefficiency in prior work like AnyTop [Gat et al. 2025]. Our experiments on the 100STYLE and LAFAN1 datasets demonstrate that UniMoGen outperforms state-of-the-art baselines, including CAMDM [Chen et al. 2024] and MDM [Tevet et al. 2023]. For example, it achieves a lower FID score than MDM, as well as fewer foot penetration frames, reduced sliding distances, and improved trajectory adherence compared to CAMDM. The ability to train on multiple skeleton types simultaneously enables a universal model applicable to diverse datasets, as shown in our superior performance against AnyTop on the combined 100STYLE and LAFAN1 datasets.

Our ablation studies further validate the importance of key design choices, such as min-max normalization, separate spatial and temporal attention, dataset balancing, and auxiliary losses, which collectively enhance motion quality and training stability. By addressing the computational and structural limitations of transformer-based models, which process full frame sequences, our method offers a scalable and efficient solution for real-world motion synthesis applications. At the end, it is worth noting that although UniMoGen uses style indices, this can easily be replaced with text input by using a text encoder instead of the style embedding layer.

Looking ahead, future work could explore integrating additional conditioning signals, such as environmental constraints or multi-modal inputs, to further enhance motion realism. Additionally, optimizing our approach for larger datasets with even more diverse skeletons presents promising directions. UniMoGen lays a strong foundation for flexible, high-quality motion generation, paving the way for advancements in animation, gaming, and virtual reality.

## References

German Barquero, Sergio Escalera, and Cristina Palmero. 2024. Seamless human motion composition with blended positional encodings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 457–469.

Rui Chen, Mingyi Shi, Shaoli Huang, Ping Tan, Taku Komura, and Xuelin Chen. 2024. Taming diffusion probabilistic models for character control. In *ACM SIGGRAPH 2024 Conference Papers*. 1–10.

Inbar Gat, Sigal Raab, Guy Tevet, Yuval Reshef, Amit H Bermano, and Daniel Cohen-Or. 2025. AnyTop: Character Animation Diffusion with Any Topology. *arXiv preprint arXiv:2502.17327* (2025).

Ziyan Guo, Zeyu Hu, Na Zhao, and De Wen Soh. 2025. MotionLab: Unified Human Motion Generation and Editing via the Motion-Condition-Motion Paradigm. *arXiv preprint arXiv:2502.02358* (2025).

Felix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher J Pal. 2020. Robust motion in-betweening. *ACM Transactions on Graphics* 39, 4 (2020).

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

Jonathan Ho and Tim Salimans. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

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

Xueqing Li, Zehan Li, Boyu Zhu, Ruihao Jing, Jian Kang, Jie Li, Xiao-Lei Zhang, and Xuelong Li. 2025. Bridging the Gap between Continuous and Informative Discrete Representations by Random Product Quantization. *arXiv preprint arXiv:2504.04721* (2025).

Zhe Li, Weihao Yuan, Yisheng He, Lingteng Qiu, Shenhao Zhu, Xiaodong Gu, Weichao Shen, Yuan Dong, Zilong Dong, and Laurence T Yang. 2024. LaMP: Language-Motion Pretraining for Motion Generation, Retrieval, and Captioning. *arXiv preprint arXiv:2410.07093* (2024).

Zeyu Ling, Bo Han, Shiyang Li, Hongdeng Shen, Jikang Cheng, and Changqing Zou. 2024. MotionLLaMA: A Unified Framework for Motion Synthesis and Comprehension. *arXiv preprint arXiv:2411.17335* (2024).

Ian Mason, Sebastian Starke, and Taku Komura. 2022. Real-Time Style Modelling of Human Locomotion via Feature-Wise Transformations and Local Motion Phases. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 5, 1, Article 6 (may 2022). doi:10.1145/3522618

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

Guy Tevet, Sigal Raab, Setareh Cohan, Daniele Reda, Zhengyi Luo, Xue Bin Peng, Amit Haim Bermano, and Michiel van de Panne. 2025. CLoSD: Closing the Loop between Simulation and Diffusion for multi-task character control. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=pZISppZSTv

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. 2023. Human Motion Diffusion Model. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=SJ1kSyO2jwu

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

Yuxin Wu and Kaiming He. 2018. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*. 3–19.

Yiyuan Yang, Ming Jin, Haomin Wen, Chaoli Zhang, Yuxuan Liang, Lintao Ma, Yi Wang, Chenghao Liu, Bin Yang, Zenglin Xu, et al. 2024. A survey on diffusion models for time series and spatio-temporal data. *arXiv preprint arXiv:2404.18886* (2024).

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. 2023. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14730–14740.

Kaifeng Zhao, Gen Li, and Siyu Tang. 2024. DartControl: A diffusion-based autoregressive motion model for real-time text-driven motion control. In *The Thirteenth International Conference on Learning Representations*.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2019. On the Continuity of Rotation Representations in Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. 2023. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 4 (2023), 2430–2449.

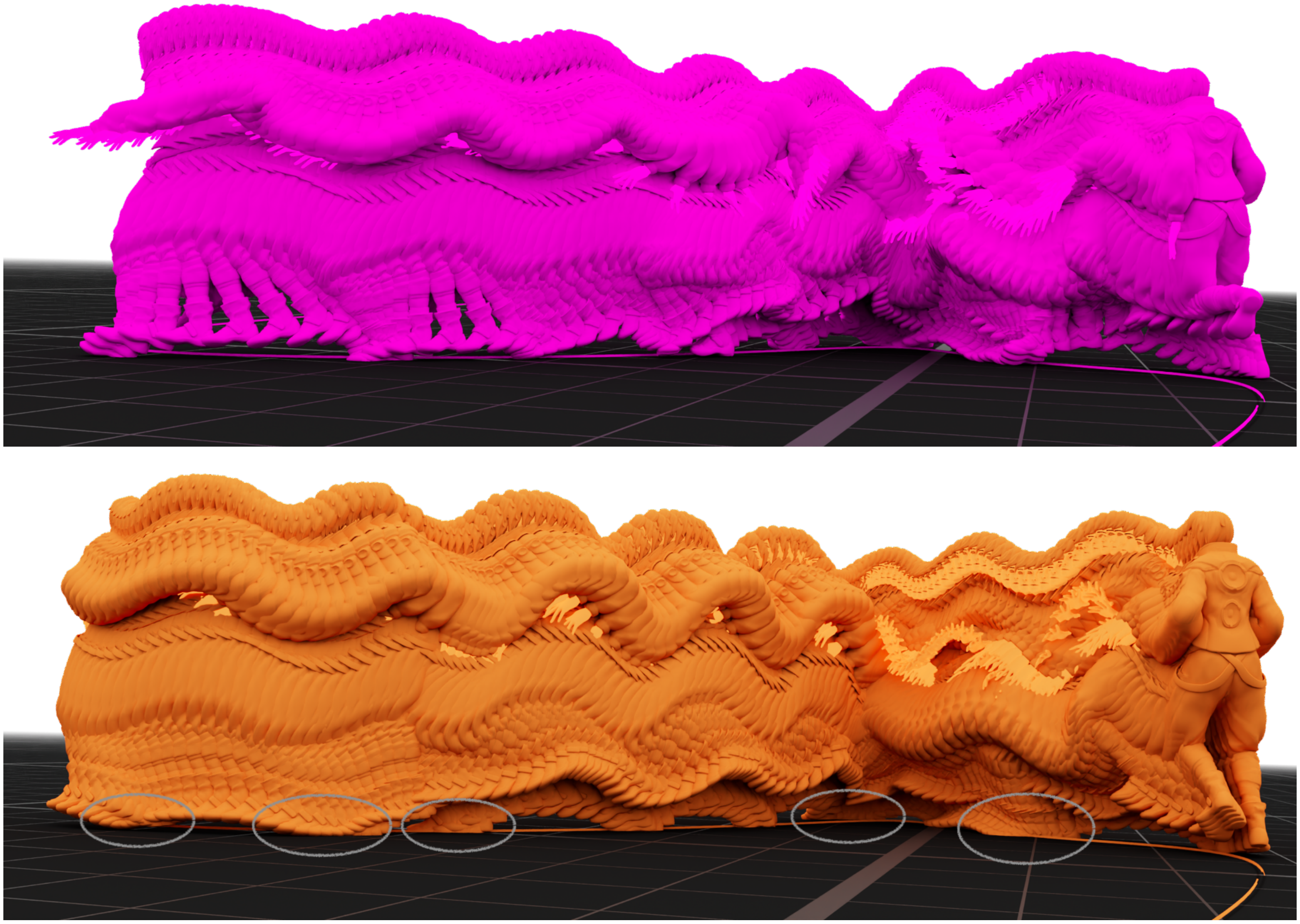

**Figure 4: Onion skinning visualization of UniMoGen and CAMDM results. The top and bottom figures compare motion outputs from UniMoGen and CAMDM, given the same past frames, style, and trajectory. As shown, our model exhibits noticeably less foot sliding and penetration. These issues are highlighted with ellipses in the CAMDM results for clarity.**

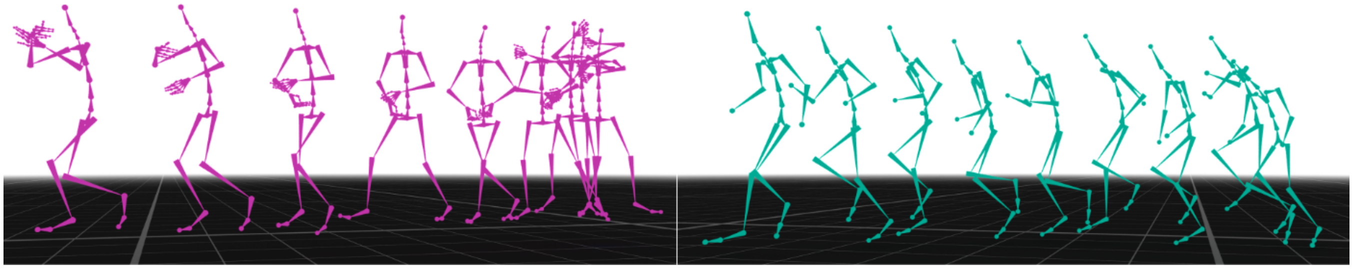

**Figure 5: Multi-Skeleton Generation. Left: a motion generated for the skeleton of LAFAN1. Right: a motion generated for the skeleton of 100Style. Both the skeletons are generated by the same model, which is trained on the combination the two datasets.**