

CMDer: Controllable Mode Decomposition-Based Single Motion Synthesis with Diffusion

Junliang Chen¹, Sihang Chen¹, Yoonsang Lee², Xiaojuan Gu³,
Kevin Romond⁴, and Fang-Lue Zhang^{1,5*}

¹ School of Engineering and Computer Science, Victoria University of Wellington,
Wellington, New Zealand junliang.chen@vuw.ac.nz

² Department Of Computer Science, Hanyang University, Seoul, South Korea

³ School of Intelligent Imaging Engineering, Beijing Film Academy, Beijing, China.

⁴ School of Design Innovation, Victoria University of Wellington,
Wellington, New Zealand

⁵ School of Computer Science and Engineering, The University of New South Wales,
Sydney, Australia fanglue.zhang@unsw.edu.au

Abstract. Realistic motion generation from a single example remains challenging due to the scarcity of large-scale, high-quality motion datasets, particularly for animals and custom characters. Existing approaches mainly operate in temporal or spatial domains, where controllability is imposed through joint-level or time-based constraints. However, motion signals exhibit a natural hierarchical structure across frequency scales: low-frequency components encode the fundamental rhythm and movement pattern, while high-frequency components capture individual details and variations. This observation suggests an alternative control dimension based on frequency-selective motion synthesis. We present CMDer (pronounced Commander), a Controllable Mode Decomposition-based Single Motion Synthesis framework that leverages diffusion models to generate diverse motions from a single motion example. CMDer employs Multivariate Variational Mode Decomposition (MVMD) to represent motion sequences as intrinsic frequency modes, providing an interpretable and structured representation for motion synthesis. A lightweight U-Net diffusion model then learns to generate these modes, while a Mode Enhancement Block (MEB) improves local motion modeling. By performing diffusion directly in the decomposed motion modes, CMDer enables frequency-selective control that allows users to preserve global rhythmic patterns while varying fine-grained motion details. Experiments demonstrate that CMDer achieves generation quality comparable to state-of-the-art temporal-domain approaches while introducing a new controllable dimension for motion synthesis. We further showcase diverse applications including mode-based synthesis, cross-modality motion generation, and controllable crowd animation.

* Corresponding author.

1 Introduction

Generating diverse and controllable motions from a single example remains challenging under data-scarce conditions. In such settings, generative models must balance two competing objectives: preserving the structural identity of the exemplar while producing diverse yet controllable variations. Existing approaches [16, 18, 21, 30] improve diversity but typically operate in joint or temporal space, where controllability is imposed along spatial or temporal dimensions rather than across frequency scales. Consequently, they struggle to disentangle global rhythmic structure from fine-grained motion details, leading to an inherent trade-off between coherence and diversity.

Motion exhibits structure across multiple frequency scales: low frequencies capture global rhythmic patterns (the motion’s fundamental signature), whereas high frequencies encode fine, variable details. This hierarchy naturally suggests a control strategy based on decomposing motion into frequency modes and selectively preserving or varying specific bands. Such frequency-domain control complements temporal and spatial conditioning by offering an extra editing axis. In crowd animation, for instance, preserving low-frequency modes across agents ensures coherence while varying high-frequency modes yields diversity, something temporal masking alone struggles to express.

To operationalise this idea, we introduce *CMDer*, a controllable mode-decomposition-based diffusion framework for single-sample motion synthesis. Unlike fixed-basis approaches such as Fourier transforms or wavelet decomposition, which assume stationarity or predefined basis functions, we employ Multivariate Variational Mode Decomposition (MVMD) [24] to adaptively decompose motion into intrinsic modes. MVMD explicitly handles non-stationary multivariate signals and preserves inter-joint correlations through shared centre frequencies across channels. This adaptive decomposition produces meaningful motion modes that serve as a structured latent representation for diffusion.

Building upon this representation, *CMDer* performs diffusion directly in the decomposed mode space. By learning to generate and recombine modes rather than raw joint trajectories, the model achieves two advantages: (1) controllability in the frequency domain, allowing selective preservation or variation of specific frequency components and (2) efficient learning in low-data regions, as smoother decomposed modes reduce overfitting and stabilize generation. A lightweight U-Net architecture combined with a Mode Enhancement Block further captures cross-mode and temporal dependencies while maintaining training efficiency.

Extensive experiments demonstrate that adaptive multivariate decomposition significantly improves motion coverage, diversity balance, and harmonic consistency compared to fixed-basis and direct diffusion baselines. These results confirm that incorporating an adaptive frequency prior provides a principled and effective inductive bias for single-sample generative modeling.

Our key contributions are:

- We propose *CMDer*, a mode-decomposition-based diffusion framework that generates diverse motions from a single sample by operating in decomposed frequency modes, enabling controllable motion generation and editing.

- We introduce ***MEB***, a lightweight motion processing block that improves local motion modeling in the decomposed mode representation while capturing relationships across joints, frequencies, and temporal dynamics.
- We demonstrate ***frequency-selective control*** for motion synthesis and editing, enabling applications such as cross-modality motion generation and controllable crowd animation.

2 Related Work

Spectral-Domain Motion Analysis. Motion signals exhibit rich structure in the frequency domain, which has been exploited for analysis and synthesis. Early work applied Discrete Fourier Transform (DFT) for spectral style transfer [33] and FFT-based control [29]. Recent approaches integrate frequency analysis with learning: spectral graph convolutions [14] and DCT-based transformers [6, 15, 20] capture joint correlations in frequency space. DeepPhase [17, 26] uses FFT-based autoencoders to learn periodic motion manifolds, showing the effectiveness of phase representations for locomotion on large datasets. FFT, however, assumes stationarity: motion statistics remain constant over time, so it is ill-suited to non-periodic or transitional motions typical in single-sample settings. Wavelet transforms offer time-frequency localization [19] and can handle non-stationarity, but they rely on predefined basis functions and have difficulty maintaining multivariate inter-joint coordination.

Adaptive Decomposition Methods. To handle non-stationary, multivariate motion without fixed bases, we adopt variational mode decomposition. Empirical Mode Decomposition (EMD) [12] and its multivariate extension (MEMD) [23] adaptively extract intrinsic modes but suffer from mode mixing and noise sensitivity. Variational Mode Decomposition (VMD) [9] mitigates these issues via constrained optimization, and Multivariate VMD (MVMD) [24] extends VMD to preserve inter-channel correlations. Despite its effectiveness in signal analysis, MVMD remains unexplored for motion synthesis. We are the first to apply MVMD to single-motion generation, leveraging its adaptive, interpretable frequency decomposition for controllable synthesis.

Single Motion Learning. Generating diverse motions from a single exemplar is critical for animating characters with limited training data, particularly animals and stylized rigs. This task requires balancing diversity with controllability. GANimator [16] introduced a progressive synthesis architecture for hierarchical motion control, but necessitates extensive training time. GenMM [18] employed a training-free generative motion matching strategy, yet limited its ability to produce highly diverse motions. SinMDM [21] focused on capturing fine-grained spatio-temporal patterns by controlling narrow receptive fields. Nevertheless, the quality of the synthesized motion in time was unstable. MotionDreamer [30] employs a Transformer to perform codebook switching in the VQ-VAE latent space. However, under single-sample learning, reliably filling and utilizing the codebook requires a sufficiently long training sequence, which is difficult to satisfy with only one short motion clip. These approaches primarily

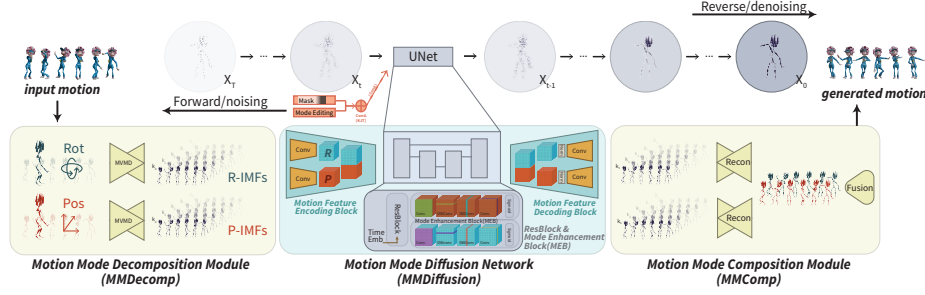


Fig. 1: CMDer pipeline for frequency-selective motion synthesis. Given a single motion, MMDecomp decomposes it into K interpretable frequency modes via MVMD (Sec. 3.1), separating low-frequency rhythm from high-frequency details. MMDiffusion learns to denoise these modes through a lightweight U-Net, enabling users to specify which modes to preserve or vary via conditional masks. MMComp reconstructs the final motion by summing the generated modes. The framework processes position (P) and rotation (R) in separate branches to preserve their distinct characteristics. This provides frequency-selective control that is difficult to express through purely temporal-domain masking. In the conditional setup, K denotes the number of modes, J the number of joints, and T the time length.

operate in the temporal or spatial domain, controlling when (e.g., keyframes, temporal masks) or where (e.g., joint-level masks) motion varies. In contrast, our method introduces a frequency-structured representation that enables selective preservation or variation of motion components across frequency scales. This formulation provides an additional, interpretable control perspective in addition to existing temporal and spatial strategies.

Diffusion Models for Motion Generation. Denoising Diffusion Probabilistic Models (DDPM) [10] have become dominant in generative modeling, with extensions for conditional generation [8], classifier-free guidance [11], and region-based editing [25]. For motion synthesis, MDM [27] introduced transformer-based diffusion for text-to-motion, establishing the DDPM paradigm for motion. Follow-up work explored physical constraints [32], real-time control [7], and multi-view generation [13]. Our work builds on the DDPM framework but differs: (1) we operate in frequency space, enabling frequency-selective control over rhythmic structure and fine-grained motion components, and (2) we focus on *single-sample learning*, requiring lightweight architectures to prevent overfitting.

3 Method

Our goal is to enable diverse, controllable motion synthesis from a single exemplar while preserving its characteristic style. We observe that motion style manifests hierarchically in the frequency domain: low frequencies capture the fundamental rhythm (the motion’s signature), while high frequencies encode in-

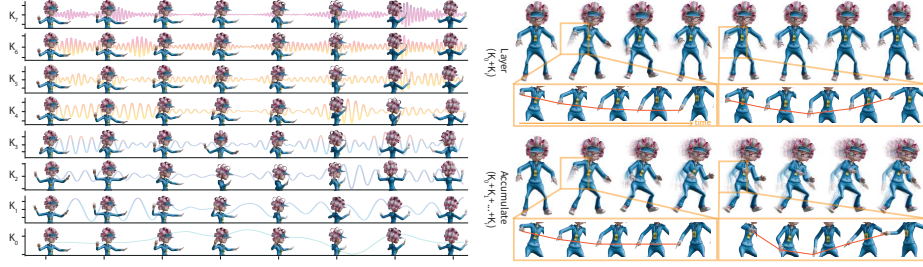


Fig. 2: Decomposition of the dancing motion of the *Sporty Granny* character using MVMD with $k = 8$. Left: Joint trajectories across layers K_i (frames 480–520, sampled every 5 frames). Right: Synthesized motions from $(K_0 + K_i)$ and cumulative sums $(K_0 + \dots + K_i)$ using two-frame ghost images. Higher layers capture increasingly fine motion details. See supplementary video for details.

dividual details. This observation motivates decomposing motion into frequency modes, learning to generate in this mode space, and selectively preserving or varying modes during synthesis.

This frequency-domain representation offers two advantages for single-motion learning: (1) *frequency-level controllability*: allowing users to selectively preserve or vary rhythmic structure and fine-grained motion components; and (2) *stable learning*: improved learning stability in low-data regimes, as decomposed modes are smoother than raw motion signals and may ease optimization.

As illustrated in Fig. 1, *CMDer* consists of three key steps: Motion Mode Decomposition Module (MMDecomp) decomposes motions into intrinsic components (modes); Motion Mode Diffusion Network (MMDiffusion) performs mode-based motion editing and synthesis through feature encoding, ResNet, diffusion, and Mode Enhancement Block; and Motion Mode Composition Module (MMComp) composes generated output back into motion sequences.

Motion Representation. A motion sequence can be viewed as a temporal set of T poses comprising the joints’ global positions $P \in \mathbb{R}^{J \times 3 \times T}$ and local rotations $R \in \mathbb{R}^{J \times Q \times T}$, where J denotes the number of joints and Q is the rotation feature dimension. We use 6D rotation features ($Q = 6$) [34] to avoid gimbal lock. The final representation is $\mathcal{M}_T = [R, P] \in \mathbb{R}^{J \times (Q+3) \times T}$.

3.1 Motion Mode Decomposition Module (MMDecomp)

Body motion involves interactions among multiple joints. MMDecomp’s objective is to extract frequency-based modes that separate style (low-freq) from variations (high-freq). Treating joint motion signals as multivariate and non-stationary signals enables better disentanglement of underlying motion patterns at different scales. To preserve inter-dependencies among motion signal channels while extracting relevant frequency modes, we employ Multivariate Variational Mode Decomposition (MVMD) [24] to decompose the multivariate signals. Some visualization results of our MVMD decomposition are shown in Fig. 2.

MVMD is an extension of the Variational Mode Decomposition (VMD) [9] algorithm, designed for multichannel or multivariate signals. The core objective of MVMD is to decompose a multichannel signal into a set of multivariate Intrinsic Mode Functions (IMFs), each characterized by a common oscillatory pattern across all channels.

In MVMD, each multivariate mode is modeled as a band-limited, amplitude- and frequency-modulated signal sharing a joint center frequency across channels. To extract these modes, MVMD formulates a variational optimization problem that minimizes the sum of bandwidths of the modes, subject to the constraint that the sum of the modes reconstructs the original multichannel signal (Eq. 1).

$$\min_{\{u_{k,c}\}, \{\omega_k\}} \sum_{k=1}^K \sum_{c=1}^C \left\| \partial_t \left[u_{k,c}^+(t) e^{-j\omega_k t} \right] \right\|_2^2 \quad \text{subject to} \quad \sum_{k=1}^K u_{k,c}(t) = x_c(t) \quad (1)$$

Here, $x_c(t)$ is the input signal in the c -th channel, $u_{k,c}^+(t)$ denotes the analytic signal of the k th mode in channel c , and ω_k is the shared center frequency for mode k across all channels. Lower-indexed modes ($\mathbf{u}_1, \mathbf{u}_2$) capture slower, global patterns; higher-indexed modes ($\mathbf{u}_{K-1}, \mathbf{u}_K$) capture faster, local details.

The solution is obtained using the Alternating Direction Method of Multipliers (ADMM), iteratively updating both the modes and their corresponding center frequencies in the Fourier domain. A full derivation of the variational formulation, analytic signal model, and the ADMM-based optimization strategy is provided in the supplementary materials.

In our framework, each entry $\mathcal{M}_T(i, j, :)$ of the motion representation $\mathcal{M}_T$ corresponds to a multivariate signal of dimension T . We perform MVMD on these multivariate signals to extract their IMFs. These mode-based motion representations make it more accessible for lightweight models to learn.

3.2 Motion Mode Diffusion Network (MMDiffusion)

Existing motion diffusion models [27, 32] typically operate in joint or temporal space, where control is applied through joint-level manipulation or noise conditioning at different diffusion steps. We design a diffusion model that learns to reconstruct mode-based motion representations, capturing structured spatio-temporal motion characteristics and enabling more diverse and controllable motion generation.

Motion Diffusion. We employ a denoising network to progressively reconstruct the distribution of target motion data, the rotation and position IMFs (R -IMFs and P -IMFs), from Gaussian noise. The diffusion process is denoted as $\{x_\tau^{1:N}\}_{\tau=0}^T$, where $x_\tau^{1:N}$ is the concatenation of P -IMFs and R -IMFs with noise at time step τ , N is the sequence length, and T is the total steps. The initial state $x_0^{1:N}$ represents the target distribution. The diffusion process is as Eq. 2.

$$q(x_\tau \mid x_{\tau-1}) = \mathcal{N}(\sqrt{\alpha_\tau} x_{\tau-1}, (1 - \alpha_\tau) I), \quad (2)$$

Conditioned motion synthesis models the distribution $p(x_0|\mathbf{c})$ as a reverse diffusion process, where $\mathbf{c}$ indicates conditional control. This process iteratively de-noises x_T to reconstruct the target data. Instead of predicting ϵ_τ [10], we directly estimate the signal as $\hat{x}_0 = G(x_\tau, \tau, \mathbf{c})$ [21, 22, 27], where G represents the generative network. We employ a simplified objective reconstruction function $\mathcal{L}_{\text{simple}}$ [10], as Eq. 3 shows.

$$\mathcal{L}_{\text{simple}} = E_{x_0 \sim q(x_0|\mathbf{c}), \tau \sim [1, T]} [\|x_0 - G(x_\tau, \tau, \mathbf{c})\|_2^2] \quad (3)$$

The forward generation network G predicts the denoised signal x_τ for the current time step. This predicted signal is subsequently used as input for the next diffusion step $\tau - 1$. The process is repeated iteratively until the time step reaches zero, resulting in the final motion prediction.

Motion Feature Encoding Block. After mode decomposition, the joint rotations and spatial positions IMFs are processed through convolutional blocks that separately embed them into high-dimensional features, which are then concatenated. To support varying skeletal structures, the hidden channel dimensions of rotations and positions are adaptively set as $H_R = 32 \lceil \frac{C_R}{32} \rceil$, $H_P = 32 \lceil \frac{C_P}{32} \rceil$.

U-Net. For each denoising step, our model utilizes a U-Net architecture integrated with a localized MEB. We employ a shallow U-Net consisting of a single layer with a ResBlock and an MEB, aiming to reduce overfitting on individual motions.

Mode Enhancement Block (MEB). In single-motion learning, controlling the receptive field is crucial to prevent overfitting and ensure motion synthesis performance while minimizing the model’s overhead. Our core insight comes from the conflict between single-motion learning and the information loss caused by traditional attention, where motion details are necessary.

To address this, we proposed a convolution-based mode-enhanced mechanism to learn spatio-temporally localised relationships within mode-decomposed motion data. To further enhance locality and efficiency, we employ depth-wise convolutions with small kernels on temporal and spatial dimensions. MEB leverages an anisotropic strip convolutions (implemented as depth-wise $(1, t)$ and $(j, 1)$ convs) to independently model temporal rhythms and joint correlations within each frequency mode k . Unlike conventional attention, we directly propagate the output of the attention module weight as features to preserve motion details. The sigmoid-based gating mechanism adaptively refines features by emphasising key spatio-temporal anchors. This approach improves motion learning efficiency, facilitates better generalization under limited motion data, while reducing the computational overhead of other attention mechanisms. The convolution-based structures also support long-sequence motion generation [16, 31]. Also, rotation and position features are processed through separate branches, enabling each to learn mode-specific attention maps without interference. Concretely, let $Z_i \in \mathbb{R}^{K \times C_i \times T}$ denote the input features for $i \in \{R, P\}$, our MEB can be denoted as Eq. 4.

$$\text{MEB}(Z_R, Z_P) = [Z'_R, Z'_P] \in \mathbb{R}^{(C_R + C_P) \times K \times J}, \quad (4)$$

where $Z'_i = \text{sigmoid}(\sigma \circ \text{BN} \circ \text{Conv}_4 \circ \text{Conv}_3^{dw} \circ \text{Conv}_2^{dw} \circ \sigma \circ \text{BN} \circ \text{Conv}_1(Z_i))$, and $\circ$ denotes function composition (from right to left). Each branch first computes hidden through $\sigma \circ \text{BN} \circ \text{Conv}_1(\cdot)$. Subsequently, we apply the depth-wise convolutional strategy along both the motion and mode axes: $\text{Conv}_2^{dw}(\cdot)$ and $\text{Conv}_3^{dw}(\cdot)$. The branch's mode enhancement results can then be calculated through a convolutional layer $\sigma \circ \text{BN} \circ \text{Conv}_4(\cdot)$ and an active layer $\text{sigmoid}(\cdot)$. Finally, the two branches' results are concatenated as output.

Motion Feature Decoding Block. The output rotation and position features from the U-Net are separately passed through normalization layers to standardize the data for a more uniform distribution that stabilizes the model and accelerates convergence. After normalization, the features are fed into convolutional layers, where they are decoded into mode functions for rotation and position, as Eq. 5.

$$\tilde{Z}_i = \text{Conv1D}(\text{SiLU}(\text{GN}_{32}(Z_i))) \quad (5)$$

Loss. The overall training objective is defined as a weighted sum of three mean-squared-error terms and a variational bound (*vb*) term. Loss $\mathcal{L}$ follows:

$$\mathcal{L} = \lambda_R \mathcal{L}_R + \lambda_P \mathcal{L}_P + \mathcal{L}_{VB},$$

where $\mathcal{L}_R$ and $\mathcal{L}_P$ are the Mean-Squared-Error losses for the reconstructed outputs r and p , respectively, each modulated by its corresponding weight λ_R , λ_P , and $\mathcal{L}_{VB}$ is the variational lower-bound loss that regularises the diffusion process.

3.3 Motion Mode Composition Module (MMComp)

The new motion sequences for each modes, generated through conditional diffusion, can be reassembled into a complete motion signal, recovering the original unified representation of joint variations over time. The reconstructed motion signal is $M_{\text{recon}} = \sum_{k=1}^K \mathbf{u}'_k$.

4 Applications

CMDer demonstrates significant potential across various applications related to motion generation and editing. The following subsections demonstrate the applications achieved with *CMDer*, as shown in Fig. 3: General Synthesis Tasks include motion completion, keyframe-guided generation, style transfer, skeletal conditional generation, long motion sequences, and crowd animation, and novel motion synthesis methods and explorations based on the mode concept, including mode editing, harmonious mixing, and modality-driven motion generation. These demonstrations highlight the applicability and controllability of our *CMDer* framework across diverse scenarios.

The diffusion-based motion synthesis tasks can be summarized as synthesizing a new motion $\hat{x}$ from a given reference motion y and a given Region of Interest (ROI) mask m , where $\hat{x}, y, m \in \mathbb{R}^{(J \times (Q+3)) \times T}$. The result is generated through diffusion that is trained with input motion x . The output motion $\hat{x}$ is

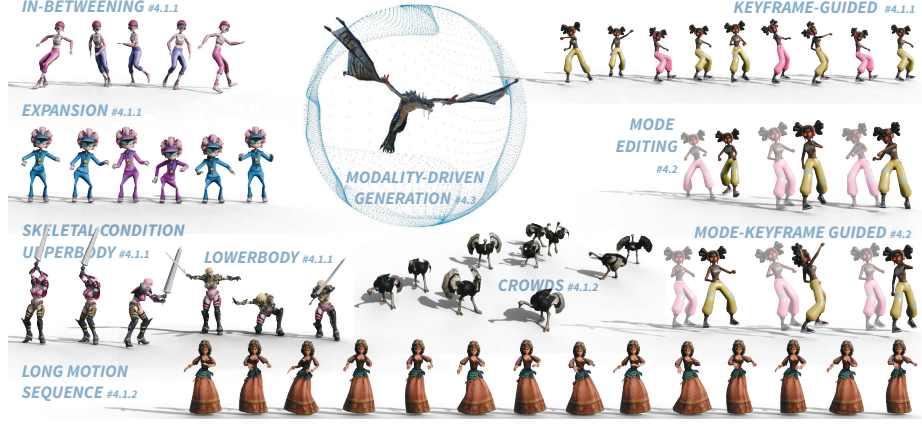


Fig. 3: Various application scenarios of our CMDer. (#4. x. x refers to Section 4. x. x)

sampled from random noise but follows x in ROI ($\hat{x} \odot m$, $\odot$ is element-wise multiplication), while the rest ($\hat{x} \odot (1 - m)$) matches the reference motion y .

Moreover, in our method, with the benefit of MVMD, the motion $\hat{x}$ and y can be decomposed as $\hat{x}'$ and y' , and the ROI mask can be extended to the same shape, that $\hat{x}', y', m' \in \mathbb{R}^{K \times (J \times (Q+3)) \times T}$, which provides another dimension, mode, that facilitates various editing across K different modes.

4.1 Temporal-Spatial Conditional Synthesis

Our *CMDer* is capable of handling conventional motion synthesis and editing tasks, particularly those involving temporal control and joint manipulation. To ensure consistency across all modes and in the resulting motion, the mask m' should be identically configured for each mode k .

Motion Completion. *CMDer* can be used to complete missing or masked parts of the motion, generating, interpolating, or expanding motions that follow learned motion patterns or reference motions. One common application is completing **in-betweening** and **expansion**, where in-betweening synthesizes plausible motion between sparse keyframes, while expansion extends a short motion clip into a longer, temporally coherent sequence.

Keyframe-guided Generation. Keyframe-guided implies that the generated motion should achieve specific designed poses at certain timestamps. To achieve this, the mask m is assigned a value of 1 at the keyframe timestamps, while the target poses should be set using the reference motion.

Skeletal Conditional Generation. Based on the method used in style transfer, by assigning a mask m to specific joints, the user can control the style influence on body parts—such as the **upper** or **lower** body—while preserving the overall cohesion and consistency of the motion.

4.2 Straight-forward Synthesis

For unconditional generation tasks such as unconditioned crowd animation or long-sequence synthesis, *CMDer* simply samples random Gaussian noise and decodes it through the diffusion model.

Crowd Animation. Crowd animation requires balancing coherence (agents appear coordinated) with diversity (avoid identical clones). Temporal-domain methods struggle with this: sampling random noise yields no coherence, while copying produces identical clones. Frequency-selective control addresses the trade-off: preserve low-frequency modes across agents for a shared rhythm and coherent appearance, while varying high-frequency modes per agent to introduce individual details and diversity.

Long Motion Sequence. With the ability of full-conv diffusion, *CMDer* can effectively extrapolate and continue motion patterns, allowing the generation of sequences as long as needed.

4.3 Mode Editing

By providing a new dimension on motion, *CMDer* offers a novel motion editing method leveraging mode-related advantages. Modes, which can be viewed as frequency-based sub-motions, are modifiable and remixable through user-controlled adjustments of masks m across the mode dimension.

Mode Layer Editing. *CMDer* allows users to directly scale or mask individual modes, treating each as a frequency-based sub-motion. By amplifying or suppressing selected low- or high-frequency modes, users can adjust global style (e.g., step size, body sway) or fine-grained details without redesigning the whole motion. For example, by setting the 0-1 mode (with $K=8$) mask as 1 and others as 0, we can generate motions with similar base (low-frequency) animations, while having diverse detailed (high-frequency) motions.

Mode Keyframe Editing. Instead of editing in the temporal domain, users can specify keyframe constraints in the mode space, such as enforcing a target pose or trajectory on selected modes at certain times. *CMDer* then diffuses in the remaining unconstrained modes, producing motions that respect the mode-level keyframes while retaining plausible variation in other components.

Mode-Based Crowd Animation. For crowd animation, *CMDer* reuses low-frequency modes across agents to ensure a coherent base locomotion, while sampling distinct high-frequency modes per agent to introduce individual variation. This mode-based control yields crowds that move in synchrony at a global level yet remain diverse in local details, which is difficult to achieve with purely temporal masking or trajectory duplication.

Mode Transfer. Traditional style transfer injects low-frequency noise from reference motion via downsampling/upsampling (e.g., SinMDM’s harmonization [21]). This corresponds to statistical low-frequency components rather than explicitly interpretable motion components. Our mode-based approach instead enables interpretable frequency-band replacement, referred to as Mode Transfer. For mode transfer, we merge the characteristics of both motions x and y in

the synthesized motion. To enable motion fusion and transfer, we first inverse-project the reference motion y into the noise space at an intermediate diffusion step. We then replace the corresponding low-frequency components of motion x with those from the reference motion y , effectively using them as guidance. The model, trained on x , then completes the remaining denoising steps, resulting in a fused motion that combines characteristics of both motions.

4.4 Modality-Driven Motion Generation

Our method exemplifies the potential of motion generation by integrating input signals like music, text, or visual cues to produce expressive, contextually appropriate motion. In our music-driven motion experiment, we map the Root Mean Square (RMS) values of music, which represent average loudness, to the conditional mask m , such that as the RMS increases, more mode layers are activated (set to 1). This heuristic produces visually synchronized results. The results show CMDer enhances context-related expressiveness and facilitates interactions between motions and other modalities across various scenarios.

5 Evaluation and Results

We evaluate our method through comparisons with prior work and ablation studies. Our framework is implemented on an NVIDIA Quadro RTX 6000 GPU. We use Mixamo [1] and Truebone [28] datasets, containing diverse animations of human characters and animals with varying skeletons. To ensure consistency in validation with other methods, we use a Mixamo benchmark, including 20 motion sequences with varying speeds and styles, ranging from 60 to 900 frames.

5.1 Metrics

We employ the following metrics: metrics 1-3 from GANimator [16], metrics 4-6 from SinMDM [21], and metric 7 from the fields of biomechanics and neuro-rehabilitation [3–5]: **1. Coverage**(*Cov.*↑): Coverage rate of generated samples over training samples. **2. Global Diversity**(*GloDiv.*↑): Global diversity by measuring Patched Nearest Neighbours. **3. Local Diversity**(*LocDiv.*↑): Frame diversity within a temporal window in the training sample. **4. Inter Diversity**(*IntDiv.*↑): Differences between sub-windows of synthesized motion. **5. Intra Diversity Difference**(*IntDiff.*↓): Difference in Inter Diversity (*IntDiv.*). **6. Harmonic Mean**(*HM.*↑): A composite evaluation metric that integrates multiple criteria. **7. Spectral Arc Length**(*Smo.*↑): This metric is commonly used to assess motion smoothness [3]. Motions containing noise will be represented by smaller values (details are provided in the supplementary). Metrics with higher values preferred are marked ↑, and lower is better with ↓.

Table 1: Quantitative Results for Mixamo Benchmark (20 motions).

Method	<i>Cov.</i> ↑	<i>GloDiv.</i> ↑	<i>LocDiv.</i> ↑	<i>IntDiv.</i> ↑	<i>IntDiff.</i> ↓	<i>Smo.</i> ↑	<i>HM.</i> ↑
GANimator [16]	96.39	1.12	<u>1.07</u>	0.08	0.26	<u>-7.20</u>	0.72
GenMM [18]	99.26	0.34	0.25	<u>0.12</u>	0.22	<u>-7.32</u>	0.38
SinMDM [21]	97.41	<u>1.30</u>	0.91	0.13	<u>0.19</u>	-7.35	<u>0.76</u>
CMDer(ours)	<u>97.94</u>	1.32	1.10	<u>0.12</u>	0.18	-7.15	0.78

* **Bold** - Best Result; Underline - Second-best Result.

Table 2: Ablation Study on MEB and Mode Decomposition Method on CMDer. Removing MEB significantly reduces coverage while increasing global and local diversity, indicating unstable and distribution-inconsistent generation.

Method	<i>Cov.</i> ↑	<i>GloDiv.</i> ↑	<i>LocDiv.</i> ↑	<i>IntDiv.</i> ↑	<i>IntDiff.</i> ↓	<i>Smo.</i> ↑	<i>HM.</i> ↑
MVMD,w/ MEB	97.94	1.32	1.10	0.12	0.18	-7.15	0.78
MVMD,w/o MEB	56.60	1.77	1.68	0.33	1.02	-13.54	0.00
FFT,w/ MEB	92.61	1.48	1.37	0.13	0.20	-10.93	0.46

5.2 Quantitative Comparisons

Comparison of Single-Motion Generation Metrics. We compare *CMDer* with several state-of-the-art (S.O.T.A) methods (GANimator [16], GenMM [18], and SinMDM [21]) on single instance learning, the evaluation results are shown in Tab. 1. *CMDer* achieves comparable overall performance (*HM*: 0.78 vs SinMDM: 0.76) while excelling in diversity (highest *GloDiv*: 1.32, *LocDiv*: 1.10) and smoothness (highest *Smo*: -7.15). GenMM achieves the highest *coverage*, however, its *global diversity* and *local diversity* are notably lower due to its reliance on motion-matching. SinMDM achieves the highest *intra diversity* but the generated motions lack smoothness. GANimator yields the highest *intra diversity difference*, indicating that its generated motions significantly diverge from the input sample distribution. Overall, *CMDer* surpasses existing methods with consistently robust results among most evaluation metrics, effectively balancing diversity and coverage.

Comparison of Periodic Model. We further compare *CMDer* with Periodic AutoEncoder (PAE) from DeepPhase [26] on single-instance performance. The trajectory inputs are randomly generated within the input motion’s projection range on the ground plane. Tested on the Mixamo Benchmark with the same generation duration, single-instance learning on non-periodic motions yields low coverage (*Cov.* = 74.09). Unnatural motion deviations lead to high diversity scores (*GloDiv.* = 1.80, *LocDiv.* = 1.66), making the results unreliable for comparison. A single sample is insufficient to construct a PAE, which also proves the effectiveness of our method in single-sample decomposition.

Ablation Study - Mode Enhancement Block (MEB). We further conduct an ablation study by removing the Mode Enhancement Block (MEB) from

Table 3: Effect of depthwise convolution kernel size on performance metrics

<i>Conv.Size</i>	<i>Cov.</i> ↑	<i>GloDiv.</i> ↑	<i>LocDiv.</i> ↑
3×3	93.03	1.65	1.43
7×7	94.95	1.46	1.27
11×11	97.83	1.37	1.21
21×21	98.10	1.27	1.14

Table 4: MVMD Sensitivity Analysis across different K values.

K	Recon.(%) ↑	Freq. Sep. (Hz) ↑	Stability ↑	Time (s) ↓	Gain Ratio ↑
4	96.75	0.657	<i>0.881</i>	5.28	0.660
8	98.70	0.533	0.895	<i>11.43</i>	<i>0.185</i>
12	<i>99.24</i>	0.499	0.895	16.98	0.043
20	99.53	0.376	0.880	27.47	0.016

CMDer. In this setting, the coverage on the Mixamo benchmark drops dramatically from 97.94 to 56.60, while the global diversity increases from 1.32 to 1.77 and the local diversity from 1.10 to 1.68. This combination of low coverage and abnormally high diversity indicates that the generated motions deviate strongly from the training sample distribution, producing highly uncontrolled and often unnatural behaviours. These results demonstrate that MEB is essential for constraining the model to generate diverse yet distribution-consistent motions, and thus plays a crucial role in the stability and controllability of CMDer.

Ablation Study – Mode Decomposition Method. We ablate the mode decomposition by replacing MVMD with an FFT-based K-band decomposition. Quantitatively, as Tab. 2 shows, the FFT variant achieves lower coverage than MVMD while maintaining comparable diversity, indicating that although it can generate varied samples, it struggles to adequately cover the target motion distribution. Together, these observations support MVMD as a more suitable decomposition for single-motion diffusion in CMDer.

Ablation Study - Depthwise Convolution kernels. To evaluate the effectiveness of the Mode Enhancement Block (MEB), we compare depthwise convolution kernels of varying sizes. As Tab. 3, *Conv.Size* indicates the different sizes of the Depthwise Convolution kernels. Smaller kernels constrain the receptive field of the diffusion model, helping preserve local motion diversity. In contrast, larger kernels increase motion coverage but often lead to overfitting, resulting in less diverse and more homogenized outputs. We select 11×11 as sweet spot balancing diversity (GloDiv: 1.37, LocDiv: 1.21) and coverage (97.83).

Ablation Study – Sensitivity of K and Reconstruction of MVMD. As shown in Tab. 4, $K=4$ fails to capture important signal details, despite exhibiting a relatively high *Benefit Gain Ratio* of 0.66. Increasing K beyond 8 leads to diminishing returns (*Gain* = 0.185). Notably, $K=8$ and $K=12$ achieve the highest *Mode Alignment Stability* (0.895), ensuring robust cross-channel synchronization. However, at $K=20$, the reduced frequency separation (0.376 Hz)

indicates over-decomposition into spurious components. Therefore, $K=8$ was selected as the optimal elbow point, balancing detail extraction and physical interpretability.

User Study. We conducted a perceptual evaluation with 10 participants comparing different methods on 10 motions per metric. Users rated quality, fidelity, and diversity on 5-point Likert scales. As shown in Fig. 4, our results are generally preferred by participants. CMDer achieves significantly higher scores on quality (3.71 vs 3.39) and fidelity (3.64 vs 3.36), with comparable diversity (3.29 vs 3.30). All methods outperform DeepPhase and GANimator.

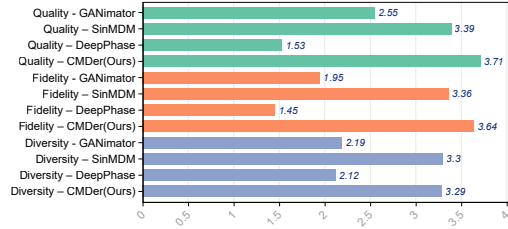


Fig. 4: User Study. The value is the score. Details are provided in the supplementary material.

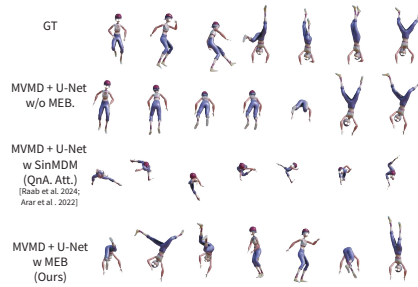


Fig. 5: Ablation. We compared our results with w/o MEB and QnA Attention [2] used in SinMDM [21], both of which failed to generate natural motion. Our Mode Enhancement Block learns joint-frequency and temporal relations, enabling more coherent motion.

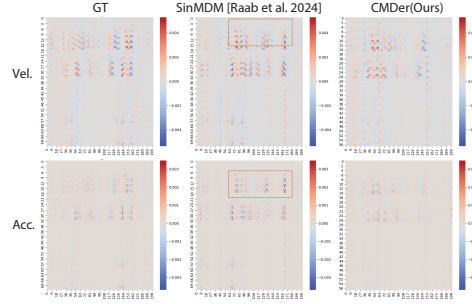


Fig. 6: Comparison of Velocity and Acceleration. We compute and visualize heatmaps of joint spatial velocity and acceleration for motion generated by SinMDM [21], our proposed method, and the Ground Truth. Our results exhibit better continuity with less abrupt changes in spatial.

5.3 Qualitative Analysis

We conduct a comprehensive qualitative evaluation, including ablation studies and comparisons with S.O.T.A methods.

Ablation Study - Mode Enhancement Block (MEB). We qualitatively evaluate MEB using *Breakdance Freeze* motion from Mixamo [1]. Fig. 5 shows frames sampled with 10-frame intervals from each method. The first row is the Ground Truth. The second, using MVMD without MEB, loses motion detail and appears rigid and unnatural. The third, MVMD with QnA attention [2] as in SinMDM [21], results in distortions by ignoring inter-mode dependencies. The fourth row shows our proposed method, which captures fine details and generates realistic, high-quality motion.



Fig. 7: Comparison of S.O.T.A Methods. Our approach demonstrates superior capability in generating continuous motion and complex motion details.

Comparison of Generation. To evaluate our model’s performance, we conducted a generated sampling comparison with state-of-the-art methods, including GANimator and SinMDM. As illustrated in Fig. 7, our approach demonstrates superior capability in generating continuous motion and complex motion details. GANimator encounters motion locking in the middle of sequences, disrupting the continuity of generated motions. Similarly, SinMDM produces motion sequences with temporal inconsistencies, characterized by abrupt changes. Fig. 6 visualizes the input and generated velocity and acceleration distributions. As observed, SinMDM exhibits a higher frequency of abrupt changes in both velocity and acceleration, indicating inconsistent motion transitions. While our method generates smoother motion sequences, with fewer sudden variations in velocity and acceleration, demonstrating its effectiveness in maintaining temporal consistency and stability.

6 Conclusion

In this paper, we propose *CMDer*, a novel single-sample motion synthesis approach based on controllable mode decomposition within a diffusion framework. Its core innovation lies in mode decomposition and a mode enhancement mechanism, enabling a deeper understanding of motion patterns. Using this framework, we explore a new perspective on motion editing and synthesis, enabling mode editing, harmonious motion mixing, and modality-driven motion generation. *CMDer* also outperforms general motion synthesis tasks, such as motion completion, crowd animation, and long motion sequence generation.

Limitation. Although *CMDer* shows promising results, further improvements are needed, such as for foot-skating artifacts that happen in the diffusion-based model [21, 27]. *CMDer* lays a foundation for motion synthesis in the spectral domain, but further study is still needed to understand spectral characteristics across various motion types, characters, and creatures. Future work will focus on these aspects to enable more practical and interactive motion synthesis.

Acknowledgements

This research was supported by the Wellington Doctoral Scholarship and Tiandi Science and Technology Co. through the Technology Collaboration Project (No. 2025-TD-CXY003).

References

1. Adobe Inc.: Mixamo (2025), <https://www.mixamo.com/>
2. Arar, M., Shamir, A., Bermano, A.H.: Learned Queries for Efficient Local Attention. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10831–10842. IEEE, New Orleans, LA, USA (Jun 2022). <https://doi.org/10.1109/CVPR52688.2022.01057>, <https://ieeexplore.ieee.org/document/9879235/>
3. Balasubramanian, S., Melendez-Calderon, A., Burdet, E.: A robust and sensitive metric for quantifying movement smoothness. *IEEE Transactions on Biomedical Engineering* **59**, 2126–2136 (2012), <https://api.semanticscholar.org/CorpusID:23763595>
4. Bayle, N., Lempereur, M., Hutin, E., Motavasseli, D., Remy-Neris, O., Gracies, J.M., Cornec, G.: Comparison of various smoothness metrics for upper limb movements in middle-aged healthy subjects. *Sensors* **23**(3) (2023). <https://doi.org/10.3390/s23031158>, <https://www.mdpi.com/1424-8220/23/3/1158>
5. Beck, Y., Herman, T., Brozgol, M., Giladi, N., Mirelman, A., Hausdorff, J.: Sparc: A new approach to quantifying gait smoothness in patients with parkinson’s disease. *Journal of NeuroEngineering and Rehabilitation* **15** (06 2018). <https://doi.org/10.1186/s12984-018-0398-3>
6. Cai, Y., Huang, L., Wang, Y., Cham, T.J., Cai, J., Yuan, J., Liu, J., Yang, X., Zhu, Y., Shen, X., Liu, D., Liu, J., Thalmann, N.M.: Learning Progressive Joint Propagation for Human Motion Prediction. In: *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII*. pp. 226–242. Springer-Verlag, Berlin, Heidelberg (2020). https://doi.org/10.1007/978-3-030-58571-6_14, https://doi.org/10.1007/978-3-030-58571-6_14
7. Chen, R., Shi, M., Huang, S., Tan, P., Komura, T., Chen, X.: Taming Diffusion Probabilistic Models for Character Control. In: *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers ’24*. pp. 1–10. ACM, Denver CO USA (Jul 2024). <https://doi.org/10.1145/3641519.3657440>, <https://dl.acm.org/doi/10.1145/3641519.3657440>
8. Dhariwal, P., Nichol, A.: Diffusion Models Beat Gans on Image Synthesis. In: *Proceedings of the 35th International Conference on Neural Information Processing Systems*. pp. 8780–8794. NIPS ’21, Curran Associates Inc., Red Hook, NY, USA (2024)
9. Dragomiretskiy, K., Zosso, D.: Variational Mode Decomposition. *IEEE Transactions on Signal Processing* **62**(3), 531–544 (Feb 2014). <https://doi.org/10.1109/TSP.2013.2288675>, <https://ieeexplore.ieee.org/document/6655981/?arnumber=6655981>
10. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. pp. 6840–6851. NIPS ’20, Curran Associates Inc., Red Hook, NY, USA (2020)

11. Ho, J., Salimans, T.: Classifier-Free Diffusion Guidance (Jul 2022). <https://doi.org/10.48550/arXiv.2207.12598>, <http://arxiv.org/abs/2207.12598>
12. Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., Yen, N.C., Tung, C.C., Liu, H.H.: The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-Stationary Time Series Analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* **454**(1971), 903–995 (Mar 1998). <https://doi.org/10.1098/rspa.1998.0193>, <https://royalsocietypublishing.org/doi/10.1098/rspa.1998.0193>
13. Kapon, R., Tevet, G., Cohen-Or, D., Bermano, A.H.: MAS: Multi-View Ances-tral Sampling for 3D Motion Generation Using 2D Diffusion. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1965–1974. IEEE, Seattle, WA, USA (Jun 2024). <https://doi.org/10.1109/CVPR52733.2024.00192>, <https://ieeexplore.ieee.org/document/10656690/>
14. Kipf, T.N., Welling, M.: Semi-Supervised Classification with Graph Convolutional Networks (Feb 2017). <https://doi.org/10.48550/arXiv.1609.02907>, <http://arxiv.org/abs/1609.02907>
15. Li, M., Chen, S., Zhao, Y., Zhang, Y., Wang, Y., Tian, Q.: Dynamic Multiscale Graph Neural Networks for 3D Skeleton Based Human Motion Prediction. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 211–220. IEEE, Seattle, WA, USA (Jun 2020). <https://doi.org/10.1109/CVPR42600.2020.00029>, <https://ieeexplore.ieee.org/document/9157584/>
16. Li, P., Aberman, K., Zhang, Z., Hanocka, R., Sorkine-Hornung, O.: GANimator: Neural Motion Synthesis from a Single Sequence. *ACM Transactions on Graphics* **41**(4), 1–12 (Jul 2022). <https://doi.org/10.1145/3528223.3530157>, <https://dl.acm.org/doi/10.1145/3528223.3530157>
17. Li, P., Starke, S., Ye, Y., Sorkine-Hornung, O.: Walkthedog: Cross-morphology motion alignment via phase manifolds. In: SIGGRAPH, Technical Papers (2024). <https://doi.org/10.1145/3641519.3657508>
18. Li, W., Chen, X., Li, P., Sorkine-Hornung, O., Chen, B.: Example-Based Motion Synthesis via Generative Motion Matching. *ACM Transactions on Graphics* **42**(4), 1–12 (Aug 2023). <https://doi.org/10.1145/3592395>, <https://dl.acm.org/doi/10.1145/3592395>
19. Liu, Z., Gortler, S.J., Cohen, M.F.: Hierarchical Spacetime Control. In: *Proceedings of the 21st annual conference on Computer graphics and interactive techniques - SIGGRAPH '94*. pp. 35–42. ACM Press, Not Known (1994). <https://doi.org/10.1145/192161.192169>, <http://portal.acm.org/citation.cfm?doid=192161.192169>
20. Mao, W., Liu, M., Salzmann, M., Li, H.: Learning Trajectory Dependencies for Human Motion Prediction. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9488–9496. IEEE, Seoul, Korea (South) (Oct 2019). <https://doi.org/10.1109/ICCV.2019.00958>, <https://ieeexplore.ieee.org/document/9009559/>
21. Raab, S., Leibovitch, I., Tevet, G., Arar, M., Bermano, A.H., Cohen-Or, D.: Single Motion Diffusion. In: *The Twelfth International Conference on Learning Representations (ICLR)* (2024), <https://openreview.net/pdf?id=DrhZneqz4n>
22. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical Text-Conditional Image Generation with Clip Latents (Apr 2022). <https://doi.org/10.48550/arXiv.2204.06125>, <http://arxiv.org/abs/2204.06125>
23. Rehman, N., Mandic, D.P.: Multivariate Empirical Mode Decomposition. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*

- 466(2117), 1291–1302 (May 2010). <https://doi.org/10.1098/rspa.2009.0502>, <https://royalsocietypublishing.org/doi/10.1098/rspa.2009.0502>
24. Rehman, N.u., Aftab, H.: Multivariate Variational Mode Decomposition. *IEEE Transactions on Signal Processing* **67**(23), 6039–6052 (Dec 2019). <https://doi.org/10.1109/TSP.2019.2951223>, <https://ieeexplore.ieee.org/document/8890883/?arnumber=8890883>
 25. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10674–10685. IEEE, New Orleans, LA, USA (Jun 2022). <https://doi.org/10.1109/CVPR52688.2022.01042>, <https://ieeexplore.ieee.org/document/9878449/>
 26. Starke, S., Mason, I., Komura, T.: Deepphase: periodic autoencoders for learning motion phase manifolds. *ACM Trans. Graph.* **41**(4) (Jul 2022). <https://doi.org/10.1145/3528223.3530178>, <https://doi.org/10.1145/3528223.3530178>
 27. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human Motion Diffusion Model (Oct 2022). <https://doi.org/10.48550/arXiv.2209.14916>, <http://arxiv.org/abs/2209.14916>
 28. Truebones Motions Animation Studios: Truebones Motions Animation Studios (2025), <https://truebones.gumroad.com/>
 29. Unuma, M., Anjyo, K., Takeuchi, R.: Fourier principles for emotion-based human figure animation. In: Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques. p. 91–96. SIGGRAPH '95, Association for Computing Machinery, New York, NY, USA (1995). <https://doi.org/10.1145/218380.218419>, <https://doi.org/10.1145/218380.218419>
 30. Wang, Y., chuan guo, Mu, Y., Javed, M.G., Zuo, X., Lu, J., Jiang, H., cheng, L.: Motiondreamer: One-to-many motion synthesis with localized generative masked transformer. In: The Thirteenth International Conference on Learning Representations (2025), <https://openreview.net/forum?id=d23EVDJRJ6g>
 31. Yan, S., Li, Z., Xiong, Y., Yan, H., Lin, D.: Convolutional Sequence Generation for Skeleton-Based Action Synthesis. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4393–4401. IEEE, Seoul, Korea (South) (Oct 2019). <https://doi.org/10.1109/ICCV.2019.00449>, <https://ieeexplore.ieee.org/document/9008261/>
 32. Yuan, Y., Song, J., Iqbal, U., Vahdat, A., Kautz, J.: PhysDiff: Physics-Guided Human Motion Diffusion Model. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15964–15975. IEEE, Paris, France (Oct 2023). <https://doi.org/10.1109/ICCV51070.2023.01467>, <https://ieeexplore.ieee.org/document/10378047/>
 33. Yumer, M.E., Mitra, N.J.: Spectral Style Transfer for Human Motion Between Independent Actions. *ACM Transactions on Graphics* **35**(4), 1–8 (Jul 2016). <https://doi.org/10.1145/2897824.2925955>, <https://dl.acm.org/doi/10.1145/2897824.2925955>
 34. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the Continuity of Rotation Representations in Neural Networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5738–5746. IEEE, Long Beach, CA, USA (Jun 2019). <https://doi.org/10.1109/CVPR.2019.00589>, <https://ieeexplore.ieee.org/document/8953486/>