Arbitrary Motion Style Transfer with Multi-condition Motion Latent Diffusion Model

Wenfeng Song[1], Xingliang Jin[1], Shuai Li[2,3*], Chenglizhao Chen[4†], Aimin Hao[3,5], Xia Hou[1], Ning Li[1], Hong Qin[6]

[1]Beijing Information Science and Technology University, China [2]Zhongguancun Laboratory, China

[3]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University

[4]College of Computer Science and Technology, China University of Petroleum (East China)

[5]Research Unit of Virtual Human and Virtual Surgery (2019RU004), Chinese Academy of Medical Sciences

[6]Department of Computer Science, Stony Brook University (SUNY at Stony Brook), Stony Brook, New York 11794-2424, USA

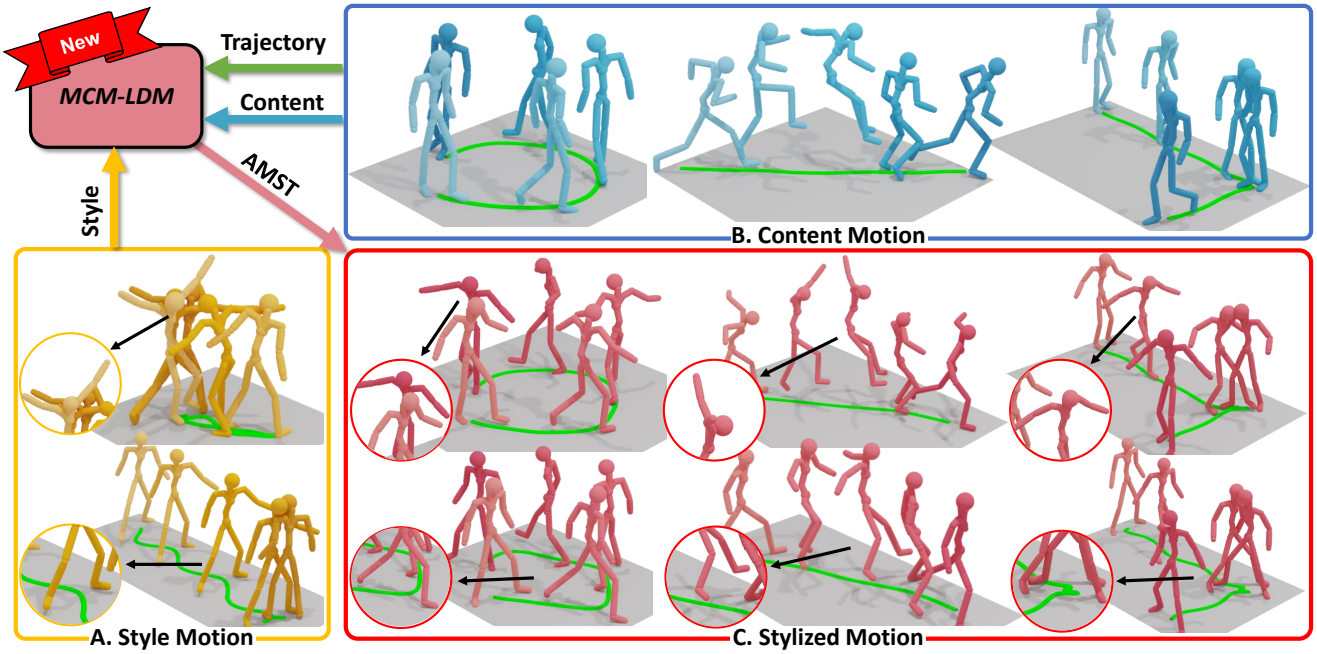

Figure 1. **Arbitrary motion style transfer with our MCM-LDM.** The black arrows point to the highlighted style features. These results illustrate our method's capability to maintain the essence of original content while seamlessly infusing it with new stylistic characteristics and trajectory considerations.

Abstract

Computer animation's quest to bridge content and style has historically been a challenging venture, with previous efforts often leaning toward one at the expense of the other. This paper tackles the inherent challenge of content-style duality, ensuring a harmonious fusion where the core narrative of the content is both preserved and elevated through stylistic enhancements. We propose a novel Multi-condition Motion Latent Diffusion Model (MCM-LDM) for Arbitrary Motion Style Transfer (AMST). Our MCM-LDM significantly emphasizes preserving trajectories, recognizing their fundamental role in defining the essence and fluidity of motion content. Our MCM-LDM's cornerstone lies in its ability first to disentangle and then intricately weave together motion's tripartite components: motion trajectory, motion content, and motion style. The critical insight of MCM-LDM is to embed multiple conditions with distinct priorities. The content channel serves as the primary flow, guiding the overall structure and movement, while the trajectory and style channels act as auxiliary components and synchronize with the primary one dynamically. This mechanism ensures that multi-conditions can seamlessly integrate into the main flow, enhancing the overall animation without overshadowing the core content. Empirical evaluations underscore the model's proficiency in achieving fluid

*,† Corresponding authors

and authentic motion style transfers, setting a new benchmark in the realm of computer animation. The source code and model are available at https://github.com/XingliangJin/MCM-LDM.git.

1. Introduction and Motivation

Computer animation, an intricate melding of computational prowess and artistic flair, has continually pushed the boundaries of what is conceivable in digital realms. Among its myriad ventures, Arbitrary Motion Style Transfer (AMST) stands out as an area of heightened intrigue and profound challenge. The vision it encapsulates is tantalizing: melding distinct motion styles onto varied content, much like casting the intense fervor of martial arts onto the delicate pirouettes of a ballet dancer or infusing the serenity of a meandering stream with the tumultuous dynamism of a waterfall. However, the road to actualizing this vision is fraught with complexities that have stymied even advanced methodologies.

Previous methods in motion style transfer, including Motion Puzzle [26] and others [1, 21, 22, 36, 42, 43], have made significant strides in AMST. However, two main challenge still exists. **Content-Style Duality:** The critical challenge in AMST is the dual imperative of maintaining content integrity while seamlessly integrating a distinct, often contrasting, style. This intricate process involves not just superimposing stylistic elements but intricately weaving them into the fabric of the original content. As exemplified in Fig. 1-C, the goal is to capture the essence of the style from style motions (Fig. 1-A) while preserving the core attributes and dynamics of the content motion (Fig. 1-B). Achieving this preservation is difficult due to the complexities of disentangling the intertwined latent spaces representing content and style.

Granularity of Details: Beyond the broader motion patterns, the devil lies in the details. The style patterns mostly ignore a critical factor: trajectory. A significant challenge arises due to the inherent discrepancies between the trajectories characteristic of the original content and the desired style. As illustrated in Fig. 2-A, conventional methods [1, 26, 36, 42] often directly transpose the content motion's trajectory onto the stylized motion. The copy-based methods, while straightforward, frequently result in unnatural artifacts, such as the common issue of 'foot sliding'.

In addressing the content-style duality, we introduce the Multi-condition Motion Latent Diffusion Model (MCM-LDM), benefiting from the generative capabilities of diffusion models, known for their effectiveness in capturing complex data distributions. MCM-LDM systematically segments motion into tripartite components — content, style, and trajectory — and employs a multi-condition guidance mechanism in the denoising process. This allows the model to generate new styles that are coherent and seamlessly integrated with the content, overcoming the common pitfall of disjointed or unnatural style transfers.

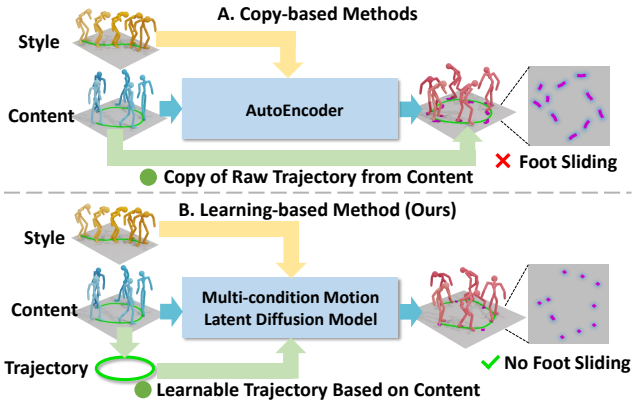

Figure 2. **Comparisons of trajectory.** Our method (B) learns to preserve motion trajectory during style transfer, while other methods [1, 26, 36] (A) copy content trajectory directly onto stylized motions, resulting in foot sliding issue.

To tackle the challenge of Granularity of Details, we propose a custom-designed Multi-condition Denoiser, to skillfully balance these conditions, ensuring the natural dynamics of the original motion are preserved while integrating new stylistic elements. Unlike previous works, we aim for the learning-based manner as shown in Fig. 2-B. The denoiser embeds multiple conditions with distinct priorities to preserve primary content while dynamically integrating style and trajectory as secondary conditions, enabling a sophisticated balance in guiding the diffusion process. This mechanism leads to more authentic and cohesive AMST outcomes (as despite in Fig. 1-C), setting a new standard in the realm of computer animation.

To summarize, our contributions are listed as follows.

- We present the first diffusion-based approach in AMST that integrates trajectory awareness, providing a nuanced solution that addresses previously unexplored aspects of motion style transfer.
- Our innovative MCM-LDM systematically extracts and guides motion through content, style, and trajectory conditions during the diffusion process, effectively addressing the complex challenges of content-style duality and the granularity of motion details.
- We propose a novel Multi-condition Denoiser, which primarily serves the content while adapting style and trajectory as secondary conditions, enabling a sophisticated balance in guiding the diffusion process. This mechanism leads to authentic and cohesive AMST outcomes, setting a new standard in the realm of computer animation.

2. Related Work

2.1. Motion Style Transfer

Initial approaches [3, 5, 7, 23, 25, 31, 33, 46, 48, 51] to motion style transfer predominantly utilized machine learning techniques. However, these methods often result in subop-

timal performance and limited transfer scope. Recent deep learning-based methods [1, 9, 13, 21, 22, 26, 29, 35, 36, 43, 47] have significantly enhanced the quality and scope of motion style transfer. For instance, Aberman et al. [1] introduced a 1D temporal convolution-based network with AdaIN for motion style control in the latent space. Park et al. [36] expanded this to spatio-temporal graph convolution (STGCN) for effective motion content preservation. These methods, however, are constrained by the need for annotated style data and specific styles. Addressing this, Jang et al. [26] introduced AMST for individual body parts using STGCN and a novel body-part style fusion network. Our approach similarly tackles AMST with our MCM-LDM. We aim to extend the network to learn to preserve the content motion trajectory without artifacts.

2.2. Motion Generation via Diffusion Model

Diffusion model has rapidly evolved in the field of motion generation [8, 12, 14, 30, 39]. Initially, methods [2, 27, 38, 40, 45, 52–54] focused on generating motions with diffusion model in the original motion space. For instance, Zhang et al. [52] was the first to use the diffusion model for generating motions from text. Later, Tevet et al. [45] utilized the transformer network structure for the diffusion model to learn condition-guided motion generation. However, applying the diffusion model directly to the original motion space incurs high computational costs and slow inference times. More recent approaches [4, 6, 11, 28, 32, 50] have shifted towards applying diffusion models in the motion latent space. MLD [11] first introduced the use of diffusion models in the continuous latent space of a motion Variational AutoEncoder (VAE). Kong et al. [28] proposed a discrete diffusion model in the Vector Quantised-Variational AutoEncoder (VQ-VAE) latent space for text-to-motion generation. Inspired by these methods, We present a multi-condition motion latent diffusion model, designed for AMST with enhanced efficiency and versatility.

3. New Method

3.1. Method Overview

Our approach achieves AMST by utilizing motion content, style, and trajectory as guiding conditions in the denoising process of our MCM-LDM. As illustrated in Fig. 3, our method begins with extraction and encoding these conditions using our Multi-condition Extraction module, as detailed in Sec. 3.2. To generate stylized motion guided by content, trajectory, and style conditions, we introduce our MCM-LDM, a motion latent diffusion model optimized for multi-condition guidance, described in Sec. 3.3. In Sec. 3.4, we provide a detailed description of our Multi-condition Denoiser (Fig. 4), designed specifically for the multi-condition guided denoising process.

3.2. Multi-condition Extraction

In contrast to conventional motion style transfer methods [1, 26, 36] that use separate inputs for content and style during the training stage, our approach takes the same motion as both the style and content input. Thus, our training task shifts motion style transfer to self-reconstruction. To effectively disentangle and encode condition features from a single motion ($x_{1:L}$, where L is the motion length), we have designed the Multi-condition Extraction module.

In particular, for the trajectory condition, we employ a transformer-based Trajectory Encoder $\mathcal{E}_{tra}$ to extract and encode the trajectory $t_{1:L}$ of $x_{1:L}$, resulting in trajectory features f_t. However, separating content and style poses a unique challenge due to their inherent overlap within a single motion sequence. We introduce Style Extractor $\mathcal{E}_{sty}$ and Content Encoder $\mathcal{E}_{con}$ specifically designed to isolate the style and content conditions.

For $\mathcal{E}_{sty}$, inspired by image style transfer methods [10, 15, 24] using pre-trained VGG [41] to extract style features, we utilize a pre-trained MotionCLIP [44] as our motion Style Extractor to extract the style features of motion. After training MotionCLIP with our specific data format, the style features f_s are derived from the output of MotionCLIP's encoder. Due to the alignment between the latent space of MotionCLIP and text/image, our $\mathcal{E}_{sty}$ can better capture the style features of the motion.

For $\mathcal{E}_{con}$, we initially feed $x_{1:L}$ through the motion encoder $\mathcal{E}$ of a pre-trained motion VAE, yielding the raw content features z_c in the latent space. Drawing inspiration from ArtFusion[10], our aim is to prevent the model from overly relying on the content information. To achieve this goal, we employ a StyleRemover, which eliminates the style from the content. Specifically, we apply the Instance Normalization layer to z_c before subjecting it to transformer encoding. Finally, we employ linear dimensionality reduction to obtain the final content feature f_c. Given that $\mathcal{E}_{sty}$ extracts the style features, our StyleRemover can naturally learn how to remove the style from the content. Consequently, the resulting content features effectively preserve the content information while eliminating the style. The above extraction of condition features can be expressed as:

$$\begin{aligned} f_t &= \mathcal{E}_{tra}(t_{1:L}), \\ f_s &= \mathcal{E}_{sty}(x_{1:L}), \\ f_c &= \text{StyleRemover}(\mathcal{E}(x_{1:L})). \end{aligned} \tag{1}$$

By encoding each condition separately, we obtain condition features (f_c, f_t, f_s), ensuring that they remain unaffected by other conditions. These independent conditions (f_c, f_t, f_s) facilitate individual guidance for the denoising process in subsequent steps of our MCM-LDM.

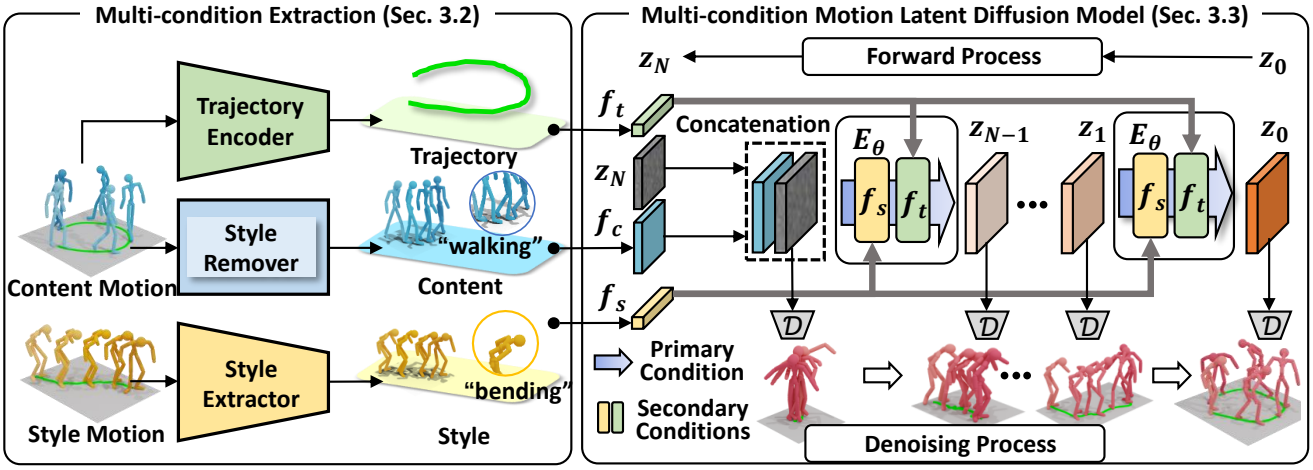

Figure 3. **Method overview.** We have two components: (1) The Multi-condition Extraction obtains content features f_c and trajectory features f_t from the content motion, while the style features f_s are obtained from the style motion. (2) MCM-LDM contains forward process and denosing process. The condition features guide the denoising process through Multi-condition Denoiser.

3.3. MCM-LDM

As demonstrated in Fig. 3-B, extended from the motion latent diffusion model [11], we employ MCM-LDM for motion style transfer. Our MCM-LDM leverages both forward and reverse diffusion processes within a motion latent space, which is defined by the same pre-trained motion VAE used in $\mathcal{E}_{con}$, under the guidance of multiple conditions: content f_c, trajectory f_t, and style f_s. By encoding the original motion $x_{1:L}$ using the VAE encoder $\mathcal{E}$, we obtain the motion latent feature $z_0 = \mathcal{E}(x_{1:L})$.

The overall diffusion process is modeled as a Markov noising process. Starting from a latent motion feature z_0, the forward process progressively adds Gaussian noise to z_0 until its distribution approximates a Gaussian distribution $\mathcal{N}(0, \mathbf{I})$ with a mean of 0 and a covariance matrix of the identity matrix $\mathbf{I}$, indicating uncorrelated variables with equal variances. The forward diffusion process is governed by the conditional probability distribution:

$$q(z_n|z_{n-1}) = \mathcal{N}\left(\sqrt{\alpha_n} z_{n-1}, (1-\alpha_n)\mathbf{I}\right), \quad (2)$$

where z_n is the noisy latent feature sampled at diffusion step n, $n \in \{1, ..., N\}$; $q(z_n|z_{n-1})$ denotes the distribution of z_n given z_{n-1}; The parameter α_n controls the level of noise added to z_{n-1}, gradually transforming it until its distribution approaches $\mathcal{N}(0, \mathbf{I})$. For the reverse process, or the denoising process, starting from a random latent feature z_N, The denoising process then progressively predicts and eliminates the noise at each diffusion step, ultimately reconstructing the original motion latent feature z_0.

To incorporate multi-condition guidance into the denoising process, we designed our Multi-condition Denoiser E_θ to predict noise based on the noisy latent feature z_n, the diffusion step n, and the guided conditions (f_c, f_t, f_s). The process of predicting noise can be represented as:

$$E_n^* = E_\theta(z_n, n, f_c, f_t, f_s), \quad (3)$$

where E_n^* denotes the predicted noise at step n. For simplicity, we use $E_\theta(z_n, f_c, f_s, f_t)$ to represent the time-dependent version $E_\theta(z_n, n, f_c, f_s, f_t)$. Further details regarding the specific network design and guided strategy can be found in Sec. 3.4. The objective [20] of our MCM-LDM can be defined as:

$$\mathcal{J} = \mathbb{E}_{E,n,(f_c,f_t,f_s)}\left[\| E - E_\theta(z_n, f_c, f_t, f_s) \|_2^2\right], \quad (4)$$

where E represents the Gaussian noise. In addition, we employ classifier-free guidance [19] during the training of E_θ. Specifically, we use shared weights for training both the full condition model $E_\theta(z_n, f_c, f_t, f_s)$ and the dual condition model $E_\theta(z_n, f_c, f_t, \varnothing)$ without style condition. $\varnothing$ is a zero null style feature. During training, we randomly set $f_s = \varnothing$ by 25% chance to train these two models.

During the inference phase, the style condition f_s is derived from the style motion, while the content condition f_c and trajectory condition f_t are provided by the content motion. The final stylized motion latent feature $\hat{z}_0$ is obtained by progressively predicting the noise in the initial random noise and denoising it. Using the motion decoder $\mathcal{D}$ in the pre-trained VAE, the final stylized motion can be obtained as $\hat{x}_{1:L} = \mathcal{D}(\hat{z}_0)$. By utilizing the classifier-free guidance, the predicted noise in the n diffusion step is computed using

$$E_n^* = \lambda E_\theta(z_n, f_c, f_t, f_s) + (1-\lambda) E_\theta(z_n, f_c, f_t, \varnothing) \quad (5)$$

instead of Equ. 3, where λ is the guidance scale. By adjusting the size of λ, we can adjust the degree of style during style transfer. Despite being trained for self-reconstruction, our MCM-LDM effectively incorporates content, trajectory, and style features for significant style transfer during inference, even with varying content and style motions. Moreover, our MCM-LDM ensures that the stylized motion maintains alignment with the original content motion, owing to the trajectory condition.

3.4. Multi-condition Denoiser

Existing methods [11, 45, 52] for diffusion-based motion generation primarily focus on single-condition guidance. However, in our work, we address the challenge of multi-condition diffusion guidance by introducing our Multi-condition Denoiser E_θ (Fig. 4). E_θ effectively achieves an adaptive balance in guiding the denoising process for our condition features (f_c, f_t, f_s) by distinguishing them into primary and secondary components. Specifically, through both experiments and everyday observations, we have recognized the significance of motion content compared to trajectory and style. Therefore, we assign the primary condition designation to the content f_c, while the trajectory f_t and style f_s serve as secondary conditions for guiding the denoising process in our E_θ. Our E_θ utilizes a transfer-based structure consisting of K layers. We then introduce the guiding strategies of these conditions.

Primary Condition Guidance. The primary condition f_c is integrated as follows:

$$z'_n = \text{Concat}(z_n, f_c), \tag{6}$$

where z'_n is the concatenated feature vector; Concat$(\cdot)$ represents concatenation. By doing so, we combine the primary condition with the noisy latent feature z_n before inputting it into E_θ, ensuring that the primary condition is involved throughout the entire denoising process, exerting its influence on the network.

Secondary Conditions and Their Optimization. For secondary conditions (trajectory f_t and style f_s), we first get the corresponding parameters through:

$$\begin{aligned} \gamma_s, \beta_s, \alpha_s &= \text{MLP}_s(f_s), \\ \gamma_t, \beta_t, \alpha_t &= \text{MLP}_t(f_t), \end{aligned} \tag{7}$$

where γ_s, β_s, α_s, γ_t, β_t, and α_t represent the parameters for the corresponding f_s and f_t conditions; MLP$_s(\cdot)$, MLP$_t(\cdot)$ denotes the different multi-layer perceptron. These parameters are then integrated into E_θ using AdaLN-Zero [37]:

$$\begin{aligned} \hat{z}_{n,k'} &= \hat{z}_{n,k-1} + \alpha_s \text{MSA}(\text{LN}(\hat{z}_{n,k-1})\gamma_s + \beta_s), \\ \hat{z}_{n,k} &= \hat{z}_{n,k'} + \alpha_t \text{MLP}(\text{LN}(\hat{z}_{n,k'})\gamma_t + \beta_t), \end{aligned} \tag{8}$$

where MSA$(\cdot)$ denotes multi-head self-attention, LN$(\cdot)$ denotes layer normalization; $\hat{z}_{n,k-1}$ and $\hat{z}_{n,k}$ represent the output of $(k-1)$-th and k-th layer of E_θ; $\hat{z}_{n,k'}$ represents the intermediate variable in the k-th layer of E_θ; MLP$(\cdot)$ denotes the multi-layer perceptron. By applying the secondary conditions at each layer of our E_θ through AdaLN-Zero, they guide the denoising process in a secondary manner.

By prioritizing the primary condition and incorporating the secondary condition in intermediate layers, our E_θ effectively learns the importance of different conditions, guiding the denoising process. The understanding enables the network to preserve the significance of motion content, resulting in desirable style transfer results. Consequently, our approach achieves desirable results in style transfer, as it retains the majority of the motion content, exhibits the specified style accurately, and preserves the trajectory intact.

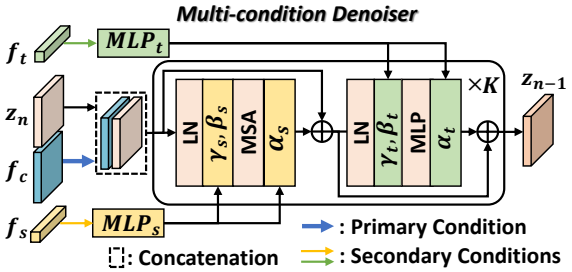

Figure 4. **Architecture of Multi-conditon Denoiser.** We incorporate the content features f_c as a primary condition by concatenating it with the noisy latent feature z_n, achieving a leading role. In contrast, the trajectory features f_t and style features f_s serve as secondary conditions, embedded into content flow dynamically.

4. Experiments

In this section, we conduct a series of experiments to evaluate the effectiveness of our MCM-LDM. Firstly, we provide an overview of the dataset setting and implementation details. Secondly, we present the quantitative metrics to assess the quality of the stylized motions. Next, we compare our MCM-LDM with other state-of-the-art methods. Following that, we conduct ablation studies to analyze the impact of our main components. Additionally, we include a user study to evaluate the performance of our MCM-LDM. More results and details are provided in the supplements.

4.1. Dataset and Implementation Details

Dataset. As our MCM-LDM aims to achieve arbitrary motion style transfer, our training data do not need any further annotated style labels. Therefore, we use a large 3D human motion dataset HumanML3D [17] to train our model, which consists of 14,616 diverse motion sequences. The motions within the dataset are originally taken from AMASS [34] and HumanAct12 [16] datasets with pre-processing.

Implementation Details. We use an off-the-shelf pre-trained VAE model from MLD [11], with a latent space size of 7×256. Following [17], the motion is represented as a combination of 3D joint rotations, positions, velocities, and foot contact, and the trajectory is obtained by the rotation and velocity of the root node. For our Content Encoder, we employ a dimension reduction from 7 to 6. The classifier-free guidance scale λ is set to 2.5. Our Multi-condition Denoiser E_θ utilizes a 9-layer architecture with a dimension of 1,024 and 4 heads. We train our model with a batch size of 128 for 400 epochs, requiring a total training time of 6.67 hours with a single RTX 3090.

4.2. Quantitative Metrics

In this section, we present five metrics that we employ to quantitatively assess the quality of the stylized motions: Fréchet Motion Distance (FMD), Content Recognition Accuracy (CRA), Style Recognition Accuracy (SRA), Trajectory Similarity Index (TSI), and Foot Sliding Factor (FSF). The first three metrics are used to evaluate the overall motion quality, content preservation, and style expression, respectively. They have been widely used in previous motion style transfer methods [26, 36, 42, 43]. We further propose TSI and FSF to evaluate trajectory similarity between the stylized motion and the content motion, and the foot sliding factors of the stylized motions. We then provide a detailed introduction of these metrics.

FMD, CRA, and SRA. To evaluate the quality of stylized motions, we employ FMD as a quantitative metric, which is a variant of Fréchet Inception Distance (FID) [18]. We train a content classifier as a feature extractor using [49] on a subset of the HumanML3D test set with annotated content labels. The FMD is computed based on the feature vectors obtained from the final pooling layer of the classifier, comparing the real and generated motion sequences. A lower FMD value indicates higher motion quality. For the CRA metric, we utilize the same content classifier to evaluate this metric. A higher CRA value implies that the generated motion has a higher potential to preserve the content of the original content motion. Similarly, for the SRA metric, we train a style classifier in another subset of the HumanML3D test set with our annotated style label. The recognition accuracy of style is calculated to obtain the SRA value. A higher SRA value indicates better style performance of the stylized motions.

TSI and FSF. To better evaluate the trajectory preservation and the degree of foot sliding in the stylized motions, we employ the TSI and FSF metrics. Specifically, we calculate the distance between the stylized motion trajectory and the original content motion trajectory using the Euclidean distance to obtain the TSI metric. For FSF metric, we calculate the foot sliding displacement generated by the feet during ground contact for each stylized motion. For more details, please refer to our supplementary material.

Methods	FMD↓	CRA↑ (%)	SRA↑ (%)	TSI↓	FSF↓
Real Motions	–	99.24	100.00	–	–
1DConv+AdaIN [1]	42.68	31.18	57.00	0.22	2.05
STGCN+AdaIN [36]	129.44	**60.43**	17.66	**0.11**	**0.93**
Motion Puzzle [26]	113.31	26.31	46.33	0.22	2.43
Ours	**27.69**	35.75	**58.00**	0.40	1.28

Table 1. **Quantitative evaluation.** '↑' ('↓') indicates that the value is better if the metric is larger (smaller); The **bold fonts** denote best performers. The results demonstrate that our MCM-LDM achieves balanced performance in all metrics.

4.3. Comparison with State-of-the-art Methods

In this section, we qualitatively and quantitatively compare four models, including Motion Puzzle [26], Conv1D+AdaIN from Aberman et al. [1], STGCN+AdaIN from Park et al. [36] and ours. To ensure a fair comparison, we retrain Motion Puzzle using the HumanML3D dataset [17], which is the same dataset used for training our model. Since the original methods of Aberman et al.[1] and Park et al. [36] are designed for style-labeled motion data, we retrain two models, Conv1D+AdaIN and STGCN+AdaIN, using their key components along with the Motion Puzzle [26]'s loss function for arbitrary style transfer. The Conv1D+AdaIN model corresponds to Aberman et al. [1]'s method, which utilizes 1D convolution and AdaIN. On the other hand, the STGCN+AdaIN model represents Park et al. [36]'s method, which incorporates spatio-temporal graph convolution and AdaIN.

Qualitative Evaluation. As shown in Fig. 5, our MCM-LDM demonstrates the ability to generate style transfer results with attractive style features while avoiding the foot sliding issue. Specifically, in the first row of Fig. 5, our stylized motion exhibits more pronounced hand movements that align with the style motion compared to other methods. Notably, Motion Puzzle [26] and 1DConv+AdaIN [1] in the first row suffer from significant foot sliding, as indicated by the purple line, while STGCN+AdaIN fails to transfer the style. In contrast, our results maintain appropriate foot contact, as we successfully incorporate the motion trajectory as an additional diffusion condition in our network. This approach ensures the preservation of the motion trajectory without directly copying it from the content motion, leading to improved foot contact alignment.

Quantitative Evaluation. We also provide multidimensional quantitative evaluations, primarily focusing on the model's generated quality, content preservation, style performance, trajectory preservation, and foot sliding degree. The results of the quantitative evaluations are presented in Table 1. From the results, our MCM-LDM significantly outperforms other methods in the FMD metric, indicating better quality in stylized motions. This can be attributed to the powerful generation capability of the diffusion model and the effectiveness of our multi-condition guidance. Additionally, we achieve a remarkable balance between content preservation and style performance. Our MCM-LDM has the best SRA and the second-best CRA. Though our CRA is slightly lower than STGCN+AdaIN model, our method focuses on harmonizing style with content, which may result in minor content modifications for a more integrated style. Notably, STGCN+AdaIN, despite its higher CRA, has the lowest SRA of 17.66, as it tends to reconstruct the original content motion. Our excellent performance in style performance and content preservation is attributed to our successful application of style and content conditional

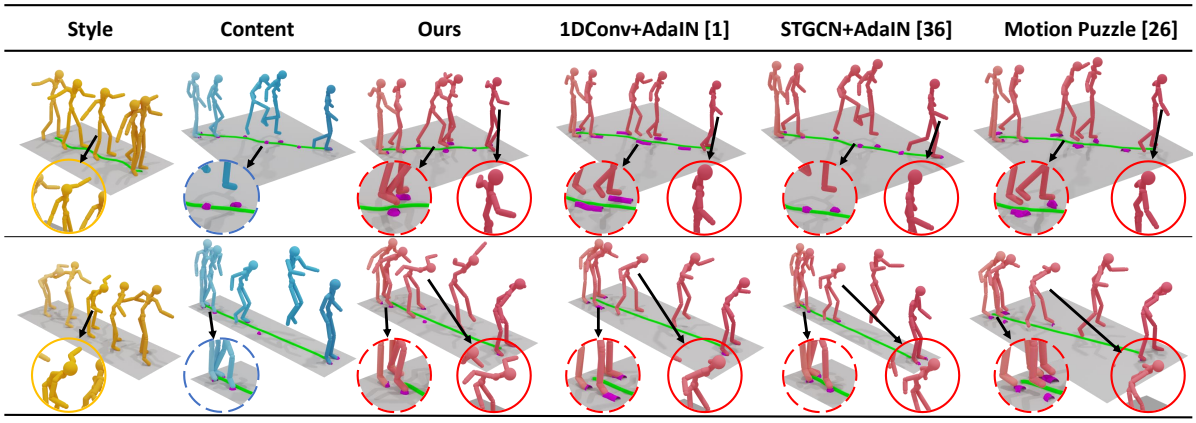

Figure 5. **Qualitative evaluation.** We provide two groups of style transfer cases. The purple line denotes the foot contact with the floor. We zoom in on the details of foot contact as well as stylistic features for a more straightforward evaluation. The results show that our MCM-LDM performs better style performance while avoiding the foot sliding issue.

Methods	FMD↓	CRA↑ (%)	SRA↑ (%)	TSI↓
w/o StyleRemover	34.78	**93.43**	16.88	**0.10**
w/o f_t	29.48	30.18	**65.11**	0.93
w/o MotionCLIP	138.55	28.18	18.00	0.67
Ours	**27.69**	35.75	58.00	0.40

Table 2. **Ablation study.** The results validate the importance of StyleRemover in $\mathcal{E}_{con}$, pre-trained MotionCLIP in $\mathcal{E}_{sty}$, and trajectory condition f_t to our approach.

Methods	FMD↓	CRA↑ (%)	SRA↑ (%)	TSI↓
w Con.	32.56	33.62	58.66	0.46
w AdaIN	33.43	30.25	**59.55**	0.51
w Pri. f_s	30.09	31.75	58.44	0.46
w Pri. f_t	32.81	32.93	55.44	0.49
Ours	**27.69**	**35.75**	58.00	**0.40**

Table 3. **Experiments of four guidance strategies in E_θ.** 'w Con.' and 'w AdaIN' represent the fusion mechanisms of concatenation and AdaIN for incorporating the secondary conditions into E_θ. 'w Pri. f_s' and 'w Pri. f_t' respectively represent treating style or trajectory as a primary condition.

adaptive guidance for diffusion-based motion generation.

We further evaluate the trajectory preservation and foot sliding using TSI and FSF. Disregarding the STGCN+AdaIN [36] model, which tends to reconstruct the original content motion, our MCM-LDM achieves the lowest FSF. As for the TSI metric, our TSI scores are lower compared to other methods. This is because other methods directly replicate the trajectory from the original content motion to preserve the trajectory, naturally resulting in high trajectory similarity but inevitably leading to foot sliding issues. In contrast, our MCM-LDM treats trajectories as an additional condition, allowing the network to learn trajectory preservation. This achieves a trade-off between trajectory accuracy and avoiding foot sliding.

4.4. Ablation Study

In this section, we conduct several ablation experiments on trajectory condition (Table 2), components in the Multi-condition Extraction (Table 2), and Guidance Strategy in E_θ (Table 3).

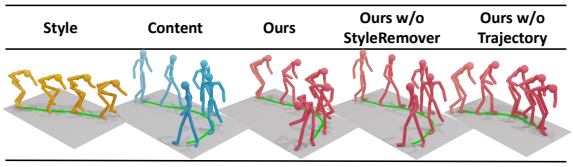

Figure 6. **Visualization of ablation study**. We present the visualization results of two ablation experiments: without our StyleRemove and without the trajectory condition. The results showcase their importance.

Importance of the Trajectory Condition. To assess the impact of the trajectory condition, we design a denoising network $E_\theta(z_n, f_c, f_s)$ that excludes the trajectory condition (Table 2: 'w/o f_t'), relying solely on content f_c and style f_s for guidance. The results show an improvement in the SRA score from 58.00 to 65.11, indicating enhanced style performance. However, the TSI score experienced a significant decrease from 0.40 to 0.93. Fig. 6 visualizes this decline, revealing the network's failure to preserve the original motion trajectories adequately. Such a deficiency is unacceptable for motion-style transfer. Therefore, by incorporating trajectories as an additional condition to E_θ, our MCM-LDM effectively retains the motion trajectories.

Importance of the Components in Multi-condition Extraction. We conduct separate experiments to evaluate the importance of the StyleRemover in $\mathcal{E}_{con}$ and the pre-trained MotionCLIP in $\mathcal{E}_{sty}$. Firstly, we remove the StyleRemover module from $\mathcal{E}_{con}$ (Table 2: 'w/o StyleRemover'). The results show that the SRA score decreases

Methods	Realism	Content Preservation	Style Performance
1DConv+AdaIN [1]	$3.91^{\pm 0.56}$	$3.91^{\pm 0.65}$	$3.86^{\pm 0.63}$
STGCN+AdaIN [36]	$3.65^{\pm 0.78}$	$3.89^{\pm 0.76}$	$3.01^{\pm 0.74}$
Motion Puzzle [26]	$3.79^{\pm 0.79}$	$3.90^{\pm 0.70}$	$3.85^{\pm 0.69}$
Ours	$\mathbf{4.48}^{\pm 0.43}$	$\mathbf{4.45}^{\pm 0.45}$	$\mathbf{4.43}^{\pm 0.51}$

Table 4. **User study.** The results show that our MCM-LDM outperforms other methods in terms of realism, content preservation, and style performance.

from 58.00 to 16.88, while the CRA score increases from 35.75 to 93.43. Fig. 6 further visualizes the style transfer result of this experiment. We find that our MCM-LDM without StyleRemover fails to transfer the style to the content, resulting in a direct reconstruction of the content motion. These results demonstrate that our StyleRemover effectively prevents the model from excessively relying on content, thereby achieving successful motion style transfer. Secondly, we experiment with the exclusion of the pre-trained MotionCLIP and the use of a transformer-based encoder to extract style features (Table 2: 'w/o MotionCLIP'). The results demonstrate a significant decrease in all metrics, indicating that the encoder involved in the training process is ineffective in extracting style features, leading to unsuccessful style transfer. This underscores the importance of our pre-trained MotionCLIP to capture style characteristics.

Efficiency of the Condition Mechanisms in Multi-condition Denoiser. We further conduct experiments with multi-condition settings in our Multi-condition Denoiser E_θ. First, we conduct experiments to explore using various conditions as the primary condition. When style f_s is considered the primary condition (Table 3: 'w Pri. f_s'), the SRA metric slightly increases, while other metrics exhibit a significant decrease. Conversely, when trajectory f_t is treated as the primary condition (Table 3: 'w Pri. f_t'), all metrics decrease noticeably. This decrease can be attributed to the lack of trajectory information, which negatively impacts the performance of style transfer. These findings highlight the crucial role of treating content as the primary condition to guide the style transfer process effectively. Secondly, we conduct experiments involving using two other fusion mechanisms where the secondary conditions are incorporated into our Multi-condition Denoiser E_θ. These fusion mechanisms include concatenation and AdaIN (Table 3: 'w Con.' and 'w AdaIN'). The results indicate that both fusion mechanisms lead to a slight increase in the SRA metric but a decrease in other metrics. To achieve a more balanced style transfer effect, we utilize the AdaLN-Zero as our fusion mechanism.

4.5. User Study

In this section, we present a user study evaluating stylized motions of our MCM-LDM in comparison with other methods, including Conv1D+AdaIN [1], STGCN+AdaIN [36], and Motion Puzzle [26]. Participants are asked to rate results generated by these methods on a scale of 1 (significantly inaccurate) to 5 (significantly accurate), based on three metrics: (1) Realism: the naturalness of the stylized motion, (2) Content Preservation: the level of the stylized motion to preserve the content information from content motion, and (3) Style Performance: the level of the stylized motion to perform the style features from style motion.

As shown in Table 4, our method achieves the highest score in three metrics. Moreover, we conduct an ANOVA test to statistically examine the differences. The overall ANOVA establishes considerable distinctions among Realism (F=11.749, p<0.01), Content Preservation (F=6.864, p<0.01), and Style Performance (F=30.619, p<0.01). The post-hoc analysis suggests that our method is significantly higher than other methods across three metrics (all p<0.01). These results further validate the effectiveness of our MCM-LDM in style transfer, making it more favored by users.

4.6. Limitation and Discussion

Although MCM-LDM could transfer arbitrary motion style with multi-conditions, it still has some limitations. First, our MCM-LDM could not generate animations with arbitrary trajectories, which is the same as the content, and the user study may have been biased due to the statistic computation. Second, MCM-LDM tends to be less effective with motions extending beyond the training dataset's temporal scope. Another limitation is the model's capability to handle content actions that involve intricate interactions with the environment. Possible directions include exploring advanced trajectory modification techniques, expanding the training datasets to encompass longer motion sequences, and enhancing the model's ability to understand and replicate environmental interactions.

5. Conclusion and Future Work

We introduced a pioneering approach to AMST through our MCM-LDM. Our model marks a significant advancement in the field of computer animation, particularly in its nuanced handling of motion trajectory, content, and style. We proposed a Multi-condition Denoiser to disentangle and harmoniously integrate the tripartite components of motion—trajectory, content, and style. This ensures a seamless integration of various conditions, thereby maintaining the authenticity of the animation and enhancing its overall appeal. In the future, we aim to extend our model's capabilities to handle more nuanced expressions and subtle human gestures, thereby enhancing its utility in creating emotionally resonant animations. The potential integration of facial and finger movements within our framework could lead to more comprehensive and lifelike character animations.

Acknowledgments

This paper is supported by Beijing Natural Science Foundation (L232102, 4222024), National Natural Science Foundation of China (62102036, 62272021, 62172246), R&D Program of Beijing Municipal Education Commission (KM202211232003), Beijing Science and Technology Plan Project Z231100005923039, National Key R&D Program of China (No. 2023YFF1203803), the Youth Innovation and Technology Support Plan of Colleges and Universities in Shandong Province (2021KJ062), USA NSF IIS-1715985 and USA NSF IIS-1812606 (awarded to Hong QIN).

References

[1] Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Unpaired motion style transfer from video to animation. *ACM Transactions on Graphics*, 39(4): 64:1–64:12, 2020. 2, 3, 6, 7, 8

[2] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics*, 42(4):1–20, 2023. 3

[3] Kenji Amaya, Armin Bruderlin, and Tom Calvert. Emotion from motion. In *Graphics Interface*, pages 222–229, 1996. 2

[4] Tenglong Ao, Zeyi Zhang, and Libin Liu. Gesturediffuclip: Gesture diffusion model with clip latents. *ACM Transactions on Graphics*, 42(4):1–18, 2023. 3

[5] Andreas Aristidou, Qiong Zeng, Efstathios Stavrakis, KangKang Yin, Daniel Cohen-Or, Yiorgos Chrysanthou, and Baoquan Chen. Emotion control of unstructured dance movements. In *Proceedings of the Eurographics Symposium on Computer Animation*, pages 1–10, 2017. 2

[6] German Barquero, Sergio Escalera, and Cristina Palmero. Belfusion: Latent diffusion for behavior-driven human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2317–2327, 2023. 3

[7] Matthew Brand and Aaron Hertzmann. Style machines. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques*, pages 183–192, 2000. 2

[8] Joao Carvalho, An T Le, Mark Baierl, Dorothea Koert, and Jan Peters. Motion planning diffusion: Learning and planning of robot motions with diffusion models. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1916–1923, 2023. 3

[9] Ziyi Chang, Edmund J. C. Findlay, Haozheng Zhang, and Hubert P. H. Shum. Unifying human motion synthesis and style transfer with denoising diffusion probabilistic models. In *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 64–74, 2023. 3

[10] Dar-Yen Chen. Artfusion: Arbitrary style transfer using dual conditional latent diffusion models. *arXiv preprint arXiv:2306.09330*, 2023. 3

[11] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 3, 4, 5

[12] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9760–9770, 2023. 3

[13] Han Du, Erik Herrmann, Janis Sprenger, Klaus Fischer, and Philipp Slusallek. Stylistic locomotion modeling and synthesis using variational generative models. In *Proceedings of the ACM SIGGRAPH Conference on Motion, Interaction and Games*, pages 1–10, 2019. 3

[14] Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 481–490, 2023. 3

[15] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016. 3

[16] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the International Conference on Multimedia*, pages 2021–2029, 2020. 5

[17] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 5, 6

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 6629–6640, 2017. 6

[19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 4

[21] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics*, 35(4):1–11, 2016. 2, 3

[22] Daniel Holden, Ikhsanul Habibie, Ikuo Kusajima, and Taku Komura. Fast neural style transfer for motion data. *IEEE Computer Graphics and Applications*, 37(4):42–49, 2017. 2, 3

[23] Eugene Hsu, Kari Pulli, and Jovan Popović. Style translation for human motion. *ACM Transactions on Graphics*, 24(3): 1082–1089, 2005. 2

[24] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1510–1519, 2017. 3

[25] Leslie Ikemoto, Okan Arikan, and David Forsyth. Generalizing motion edits with gaussian processes. *ACM Transactions on Graphics*, 28(1):1–12, 2009. 2

[26] Deok-Kyeong Jang, Soomin Park, and Sung-Hee Lee. Motion puzzle: Arbitrary motion style transfer by body part. *ACM Transactions on Graphics*, 41(3):1–16, 2022. 2, 3, 6, 7, 8

[27] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8255–8263, 2023. 3

[28] Hanyang Kong, Kehong Gong, Dongze Lian, Michael Bi Mi, and Xinchao Wang. Priority-centric human motion generation in discrete latent space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14806–14816, 2023. 3

[29] Shigeru Kuriyama, Tomohiko Mukai, Takafumi Taketomi, and Tomoyuki Mukasa. Context-based style transfer of tokenized gestures. In *Computer Graphics Forum*, pages 305–315, 2022. 3

[30] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *arXiv preprint arXiv:2304.05684*, 2023. 3

[31] C. Karen Liu, Aaron Hertzmann, and Zoran Popović. Learning physics-based motion style with nonlinear inverse optimization. *ACM Transactions on Graphics*, 24(3):1071–1081, 2005. 2

[32] Yunhong Lou, Linchao Zhu, Yaxiong Wang, Xiaohan Wang, and Yi Yang. Diversemotion: Towards diverse human motion generation via discrete diffusion. *arXiv preprint arXiv:2309.01372*, 2023. 3

[33] Wanli Ma, Shihong Xia, Jessica K Hodgins, Xiao Yang, Chunpeng Li, and Zhaoqi Wang. Modeling style and variation in human motion. In *Proceedings of the Eurographics Symposium on Computer Animation*, pages 21–30, 2010. 2

[34] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5442–5451, 2019. 5

[35] Ian Mason, Sebastian Starke, He Zhang, Hakan Bilen, and Taku Komura. Few-shot learning of homogeneous human locomotion styles. *Computer Graphics Forum*, 37(7):143–153, 2018. 3

[36] Soomin Park, Deok-Kyeong Jang, and Sung-Hee Lee. Diverse motion stylization for multiple style domains via spatial-temporal graph-based generative model. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 4(3):1–17, 2021. 2, 3, 6, 7, 8

[37] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 5

[38] Sigal Raab, Inbal Leibovitch, Guy Tevet, Moab Arar, Amit Haim Bermano, and Daniel Cohen-Or. Single motion diffusion. In *International Conference on Learning Representations*, 2024. 3

[39] Zhiyuan Ren, Zhihong Pan, Xin Zhou, and Le Kang. Diffusion motion: Generate text-guided 3d human motion by diffusion model. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2023. 3

[40] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 3

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 3

[42] Wenfeng Song, Xingliang Jin, Shuai Li, Chenglizhao Chen, Aimin Hao, and Xia Hou. Finestyle: Semantic-aware fine-grained motion style transfer with dual interactive-flow fusion. *IEEE Transactions on Visualization and Computer Graphics*, 29(11):4361–4371, 2023. 2, 6

[43] Tianxin Tao, Xiaohang Zhan, Zhongquan Chen, and Michiel van de Panne. Style-erd: Responsive and coherent online motion style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6583–6593, 2022. 2, 3, 6

[44] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Proceedings of the European Conference on Computer Vision*, pages 358–374, 2022. 3

[45] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *International Conference on Learning Representations*, 2023. 3, 5

[46] Munetoshi Unuma, Ken Anjyo, and Ryozo Takeuchi. Fourier principles for emotion-based human figure animation. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques*, pages 91–96, 1995. 2

[47] Yu-Hui Wen, Zhipeng Yang, Hongbo Fu, Lin Gao, Yanan Sun, and Yong-Jin Liu. Autoregressive stylized motion synthesis with generative flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13607–13607, 2021. 3

[48] Shihong Xia, Congyi Wang, Jinxiang Chai, and Jessica Hodgins. Realtime style transfer for unlabeled heterogeneous human motion. *ACM Transactions on Graphics*, 34(4):1–10, 2015. 2

[49] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7444–7452, 2018. 6

[50] Siyue Yao, Mingjie Sun, Bingliang Li, Fengyu Yang, Junle Wang, and Ruimao Zhang. Dance with you: The diversity controllable dancer generation via diffusion models. In *Proceedings of the ACM International Conference on Multimedia*, pages 8504–8514, 2023. 3

[51] M Ersin Yumer and Niloy J Mitra. Spectral style transfer for human motion between independent actions. *ACM Transactions on Graphics*, 35(4):1–8, 2016. 2

[52] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 3, 5

[53] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 364–373, 2023.

[54] Mengyi Zhao, Mengyuan Liu, Bin Ren, Shuling Dai, and Nicu Sebe. Modiff: Action-conditioned 3d motion generation with denoising diffusion probabilistic models. *arXiv preprint arXiv:2301.03949*, 2023. 3