UMSD:High Realism Motion Style Transfer via Unified Mamba-based Diffusion

Anonymous Author(s)
Submission Id: 923

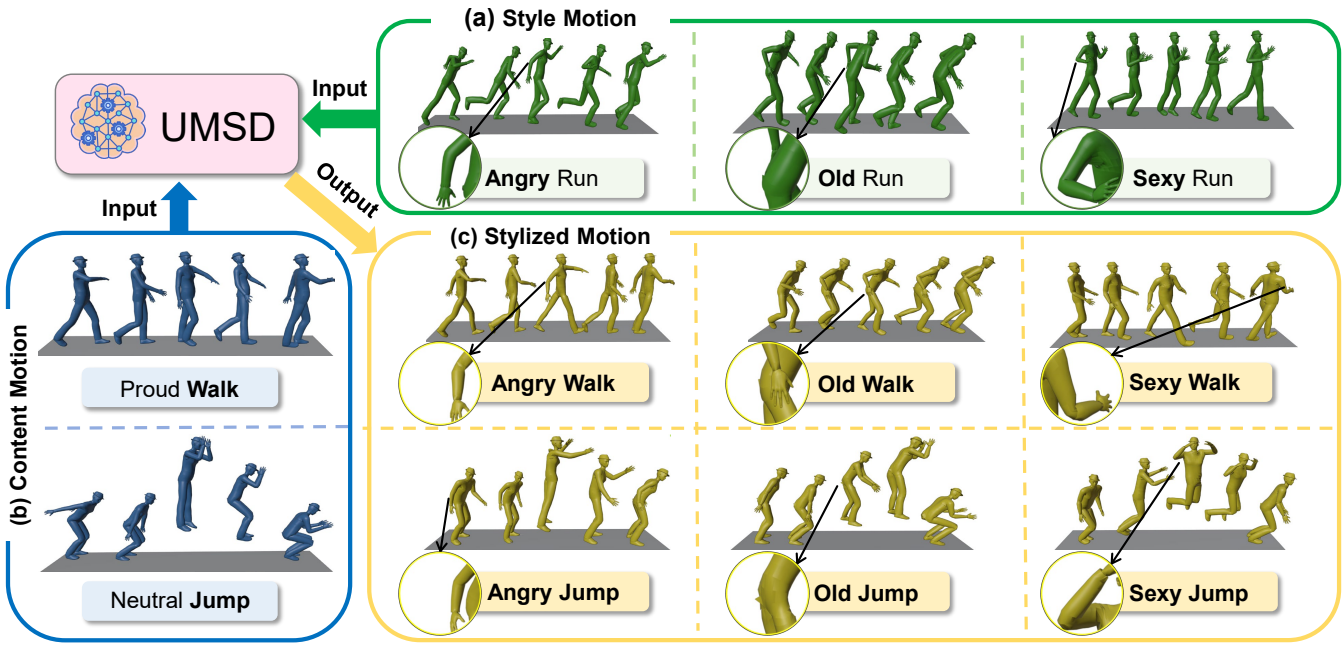

Figure 1: Motion style transfer with UMSD. Our method effectively transfers the style of the style motion to the content of the content motion, ensuring the prominence of the style while preserving the content of the motion to a great extent.

Abstract

Motion style transfer is a significant research direction in the field of computer vision, enabling virtual digital humans to rapidly switch between different styles of the same motion, thereby significantly enhancing the richness and realism of movements. It has been widely applied in multimedia scenarios such as films, games, and the metaverse. However, most existing methods adopt a two-stream structure, which tends to overlook the intrinsic relationship between content and style motions, leading to information loss and poor alignment. Moreover, when handling long-range motion sequences, these methods fail to effectively learn temporal dependencies, ultimately resulting in unnatural generated motions. To address these limitations, we propose a Unified Motion Style Diffusion (UMSD) framework, which simultaneously extracts features from both content and style motions and facilitates sufficient information interaction. Additionally, we introduce the Motion Style Mamba (MSM) denoiser, the first approach in the field of motion style transfer to leverage Mamba's powerful sequence modelling capability. Better capturing temporal relationships generates more coherent stylized motion sequences. Third, we design a Diffusion-based Content Consistency Loss and a Style Consistency Loss to constrain the framework, ensuring that it inherits the content motion while effectively learning the characteristics of the style motion. Finally, extensive experiments demonstrate that our method outperforms state-of-the-art (SOTA) methods qualitatively and quantitatively, achieving more realistic and coherent motion style transfer.

CCS Concepts

• **Computing methodologies** → **Motion processing**; • **Social and professional topics** → **User characteristics**.

Keywords

Motion Style Transfer, Mamba Model, One-stream Structure, Diffusion Generative Models

ACM MM, Dublin, Ireland

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Style is a crucial aspect of human motion, reflecting individual characteristics such as emotions, age, and health status [9], which

are essential for animating human characters and avatars. However, manually extracting these styles is both challenging and time-consuming [3, 36]. Motion style transfer addresses this issue by transferring the style of a style motion onto the content of a content motion, resulting in a stylized motion that retains both characteristics (*e.g.*, an angry walk in Figure 1c). This technique enhances the diversity and realism of digital human motion, making it widely applicable in fields such as film production and game design.

However, achieving style characteristics while preserving motion content poses significant challenges. To address this issue, existing methods [1, 18, 31, 32] typically employ two separate encoders to independently extract features from content and style motions. The extracted information is fused, and a generative model produces stylized motion. Although this approach enables motion style transfer to some extent, the generated motion sequences still fall short in terms of naturalness and continuity for the following reasons: (1) Most current frameworks adopt a two-stream structure, with independent encoders for content and style motion, as shown in Figure 2a. This setup causes the encoders to overlook intrinsic connections between the two motion types, leading to information loss and poor alignment of features in high-dimensional space, ultimately degrading the quality of the generated output. (2) Current frameworks struggle to capture temporal relationships in long-range motion sequences, resulting in generated motion sequences that lack naturalness and coherence.

To address the above-mentioned issues, we propose a UMSD Framework, which employs a one-stream structure to extract features from content and style motions simultaneously. This unified approach enables effective information exchange, generates stylized motion, and overcomes the limitations of using separate encoders, as shown in Figure 2b. Specifically, we introduce a novel UMSD Attention module that integrates cross-attention and self-attention mechanisms. The cross-attention mechanism enables information exchange between content and style motion features, enhancing their complementarity. Meanwhile, the self-attention mechanism independently processes each motion feature, capturing crucial local details and dependencies within the motion sequence.

To maintain long-range dependencies within the motion sequence, we first introduce the Mamba model [12] for motion style transfer and propose the Motion Style Mamba (MSM) denoiser. MSM leverages the robust sequence modeling capability of State Space Models (SSM) to generate more coherent and natural motions by learning temporal dependencies. Also, we propose a diffusion-based style consistency loss and a diffusion-based content consistency loss, which constrain the UMSD framework to inherit motion content while learning motion style effectively. Extensive experiments on two benchmark datasets demonstrate that our method outperforms SOTA approaches. Our contributions are as follows:

- We propose a novel UMSD framework employing a one-stream structure. This structure enables simultaneous feature extraction from content and style motion while facilitating extensive interaction between them.
- We apply the Mamba model to the motion style transfer field for the first time and propose the MSM denoiser. This leverages SSM's strong sequential modeling capability to better preserve long-range dependencies in motion sequences.
- We propose a Diffusion-based Content and Style Consistency Loss, which separately constrains the UMSD framework to more comprehensively retain input motion content while effectively learning style features.
- We conduct extensive experiments to evaluate our framework, and the results show that the proposed UMSD framework outperforms SOTA methods in both qualitative and quantitative metrics.

2 Related Work

2.1 Motion Style Transfer

Motion style transfer is an advanced and essential research area in computer vision. Early methods [3, 36] rely on handcrafted feature extraction to design various motion styles, which is inefficient and difficult to apply in practical contexts such as film and gaming. In recent years, new methods [1, 18, 23, 26, 31, 32, 41, 42] based on deep learning techniques [7, 11, 24, 33, 37] have been proposed, effectively addressing the inefficiency of manual feature extraction. For instance, Aberman *et al.* [1] used unsupervised learning to transfer motion style by learning from a collection of style-labeled motions. Building on this, Finestyle [31] and Most [18], respectively, designed a bidirectional interaction flow fusion module and an innovative motion style transformer, enabling effective learning of motion content and style feature transfer. MCM-LDM [32] achieved high-quality motion style transfer by disentangling and finely integrating three key elements: motion trajectory, content, and style, ensuring the core content is preserved. However, most of these approaches adopted a two-stream structure, with two separate encoders extracting content and style motion features. This can cause the encoders to overlook intrinsic connections between the two motion types, leading to information loss and poor alignment in high-dimensional spaces.

2.2 Diffusion Generative Models

Diffusion models are highly regarded for their exceptional performance in various research fields, including image generation [8, 28, 39], video generation [5, 28, 43], reinforcement learning [10, 17], and motion generation [4, 13, 34, 35]. For example, MDM [34] utilized uses a transformer-based diffusion model for condition-guided motion generation. MLD [6] introduced diffusion models in the latent space of a motion VAE, significantly improving high-fidelity motion generation. Alexanderson *et al.* [2] explored diffusion models for audio-driven motion generation, demonstrating how auditory cues can guide motion synthesis. However, in the above diffusion models, the denoiser often struggles to effectively learn the temporal dependencies in long sequences, which limits its performance in motion style transfer tasks.

3 Methodology

Overview. We propose a novel UMSD framework to achieve high naturalness in motion style transfer, as detailed in Section 3.1. First, we apply UMSD Attention to extract features from content and style motions and enable information exchange, as described in Section 3.2. Then, we introduce an MSM denoiser, which generates more coherent stylized motion sequences (Section 3.3). Section 3.4 presents two loss functions to supervise content and style consistency.

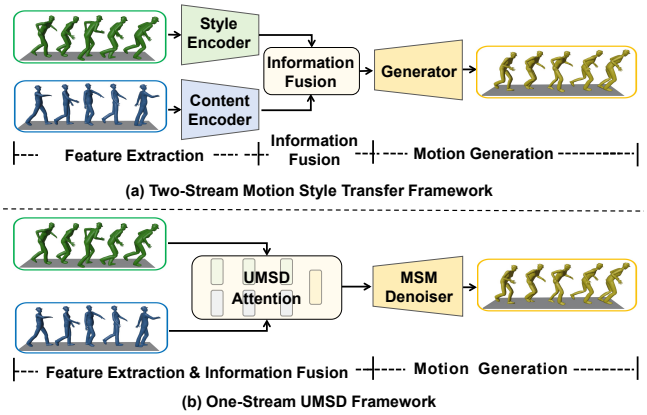

Figure 2: Comparison between two-stream frameworks and our one-stream UMSD. (a) Existing methods first extract content and style motion features separately, then perform feature fusion and motion generation. (b) In contrast, UMSD unifies feature extraction and information fusion, enabling direct generation of stylized motion.

Pose Representation. We categorize the motion sequences input into the UMSD framework into two types: content motion and style motion. Since the stylized motion output is strongly correlated with the input content and each motion is clearly defined by joint rotations (unit quaternions) [21], we represent the content motion sequence as joint rotations $\boldsymbol{m}^{c,1:N} = \{\boldsymbol{m}^{c,i}\}_{i=1}^{N} \in \mathbb{R}^{4J\times N}$. Additionally, as style can be inferred from the relative motion of joint positions, we use joint positions to represent the style motion sequence $\boldsymbol{n}^{s,1:N} = \{\boldsymbol{n}^{s,i}\}_{i=1}^{N} \in \mathbb{R}^{3J\times N}$, where $J = 21$ is the number of joints in the human skeleton [1], and N represents the number of poses in a motion sequence. Here, $\boldsymbol{m}$ and $\boldsymbol{n}$ denote the motion content of content and style motion, respectively, and c and s indicate the motion style for content and style motion. The UMSD framework aims to learn the motion style s while retaining the motion content m, thereby generating a stylized motion $\boldsymbol{m}^{s,1:N} = \{\boldsymbol{m}^{s,i}\}_{i=1}^{N} \in \mathbb{R}^{4J\times N}$ that combines both characteristics.

3.1 Unified Motion Style Diffusion Framework

Existing frameworks [1, 18, 31] for motion style transfer typically use a two-stream structure with separate encoders for content and style motion feature extraction. This structure often leads to information loss and misalignment in high-dimensional space. Moreover, when handling long-range motion sequences, these frameworks struggle to effectively model sequence dependencies and temporal relationships, resulting in stylized motions that lack natural flow and coherence. We propose a UMSD framework to address these issues, as illustrated in Figure 3.

Our UMSD framework is a one-stream structure based on the diffusion model [4]. Taking the content motion $\boldsymbol{m}_t^{c,1}$ at noising step $\boldsymbol{t}$ as an example, diffusion is regarded as a Markov noising process. The motion sequence follows a forward noising process, $q(\boldsymbol{m}_t^{c,1:N} \mid \boldsymbol{m}_{t-1}^{c,1:N})$, where $\boldsymbol{m}_0^{c,1:N}$ is drawn from the data distribution. The forward noising process is defined as:

$$q(\boldsymbol{m}_t^{c,1:N} \mid \boldsymbol{m}_{t-1}^{c,1:N}) = \mathcal{N}(\sqrt{\alpha_t}\boldsymbol{m}_{t-1}^{c,1:N}, (1-\alpha_t)I), \tag{1}$$

where $\alpha_t \in (0,1)$ are constant hyperparameters. As α_t approaches 0, we approximate $\boldsymbol{m}_T^{c,1:N} \sim \mathcal{N}(0,1)$. We use $T = 1000$ timesteps. The forward noising process for the style motion $\boldsymbol{n}_t^{s,1:N}$ follows the same procedure. From this point, we denote the full-length sequences of content motion, style motion, and stylized motion at noising step $\boldsymbol{t}$ as $\boldsymbol{m}_t^c$, $\boldsymbol{n}_t^s$, and $\boldsymbol{m}_t^s$, respectively.

3.2 UMSD Attention

During motion style transfer, the accurate encoding of both style and content is essential. Existing methods [1, 18, 31, 32] typically encode them separately, which causes the encoders to overlook the intrinsic connections between the two motion types. We propose UMSD Attention, which integrates cross-attention and self-attention mechanisms to extract features from both content and style motions, facilitating comprehensive information exchange, as shown in Figure 3.

To begin with, we concatenate the position-encoded content motion sequence $\boldsymbol{m}_t^c$ and style motion sequence $\boldsymbol{n}_t^s$ at noising step $\boldsymbol{t}$, resulting in the unified sequence Z_t^{u1}, defined as $Z_t^{u1} = [Z_t^{c1}; Z_t^{s1}]$. Here, each pose in the motion sequence represents a token. We perform feature extraction and information fusion in three stages, using Z_t^{si} with $i \in \{1,2,3,4\}$ as query embeddings in the following illustration. In the first stage, we employ a cross-attention mechanism to facilitate information exchange between content and style motion features, enhancing their complementarity. This process is expressed as follows:

$$Z_t^{s2} = \text{softmax}\left(\frac{Q_{s1}K_{c1}^T}{\sqrt{d}}\right)V_{c1}. \tag{2}$$

The query Q_{si} is generated by applying a linear projection to Z_t^{si}. At the same time, K_{ci}^T and V_{ci} are produced by linearly projecting Z_t^{ci} to obtain keys and values, respectively, where $i \in \{1,2,3,4\}$ and d represents the dimension of the key. We then apply a self-attention mechanism to independently process the content and style motion features, capturing critical local details and dependencies across motion sequences, yielding Z_t^{s3}. In the third stage, we use a cross-attention mechanism to establish deeper connections between the content and style motions, which helps the final stylized motion better integrate both feature types. The following formula represents this process:

$$Z_t^{s3} = \text{softmax}\left(\frac{Q_{s2}K_{s2}^T}{\sqrt{d}}\right)V_{s2}, \tag{3}$$

$$Z_t^{s4} = \text{softmax}\left(\frac{Q_{s3}K_{c3}^T}{\sqrt{d}}\right)V_{c3}. \tag{4}$$

K_{si}^T and V_{si} represent the keys and values generated from Z_t^{si}, with $i \in \{1,2,3,4\}$, through linear projection. The process for obtaining query embeddings Z_t^{ci} follows the same steps, resulting in Z_t^{c4}. By concatenating it with Z_t^{s4}, we obtain Z_t^{u2} as $Z_t^{u2} = [Z_t^{c4}; Z_t^{s4}]$. Subsequently, the following operation is applied to yield the output Z_t^{out} for the UMSD attention:

$$Z_t^{out} = Z_t^{u1} \oplus \text{LN}(Z_t^{u2}), \tag{5}$$

where $\oplus$ denotes the matrix addition, and $\text{LN}(\cdot)$ is a linear layer.

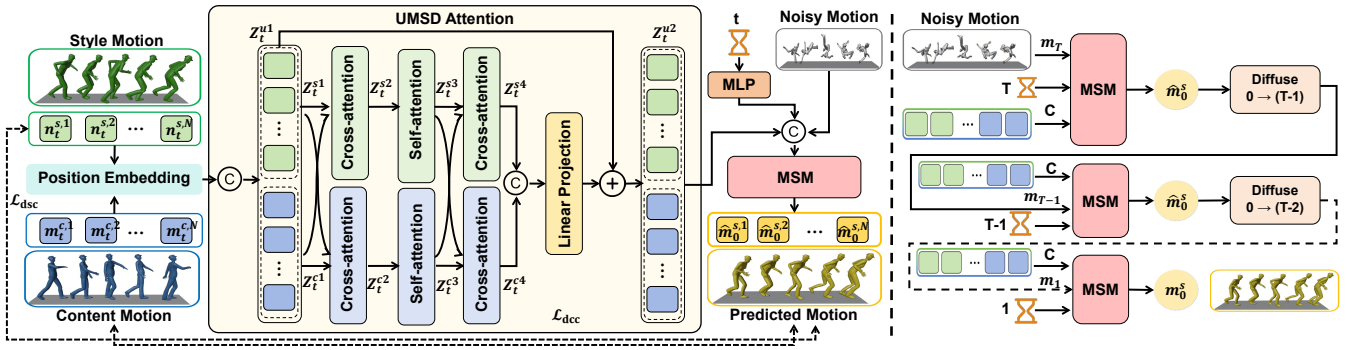

Figure 3: Framework overview. (Left) Overview of the Unified Motion Style Diffusion (UMSD) framework. The framework input is a noisy motion sequence $m_t^{z,1:N}$ of length N at noising step t, with $z \in \{c, s\}$, conditional information C, and t itself. Here, C represents the result of information interaction between content motion $m_t^{c,1:N}$ and style motion $n_t^{s,1:N}$ through UMSD attention. A Motion Style Mamba (MSM) denoiser is then applied to predict the stylized motion $\hat{m}_0^{s,1:N}$. (Right) Sampling MSM. Given the condition C, noisy motion is sampled across the desired motion dimensions, iterating from t down to 1. At each step t, the MSM predicts the clean stylized motion $\hat{m}_0^{s,1:N}$, which is then diffused back to $\hat{m}_{t-1}^{s,1:N}$.

Through this interactive attention mechanism, UMSD attention effectively encodes both style and content, allowing them to mutually reinforce each other. This enables the style transfer results to not only faithfully express the style but also retain the content to a significant extent.

3.3 Motion Style Mamba Denoiser

After employing diffusion-based models [15], we observe that existing denoisers, such as U-Net [29] and transformers [19, 37], struggle to capture temporal relationships in long-range motion sequences effectively [34]. This limitation leads to generated motion sequences that lack natural continuity. To address this issue, we draw inspiration from the Mamba model [12] and apply it for the first time in the motion style transfer field, proposing the Motion Style Mamba (MSM) denoiser. MSM leverages the powerful sequence modelling capability of the State Space Model (SSM) to capture temporal information in motion sequences better, preserving long-term dependencies within them. The structural diagram is shown in Figure 4. Before entering the MSM denoiser, the motion sequence $\boldsymbol{m}_z^t$ at noising step $\boldsymbol{t}$, with $\boldsymbol{z} \in \{c, s\}$, undergoes the following processing:

$$\mathcal{D}_{in} = \text{concat}(\boldsymbol{m}_t^z, \text{MLP}(T), Z_t^{out}), \tag{6}$$

where concat$(\cdot)$ represents concatenation, D_{in} represents the input to the MSM denoiser. The MLP consists of two linear layers and an activation layer, projecting the timestep $\boldsymbol{t}$ into a continuous vector space to form a latent vector optimized for MSM processing.

The MSM block, the core of the MSM denoiser, leverages the long-range sequential modeling strengths of the SSM to map the timestep $\boldsymbol{t}$ into content and style motion sequences, thereby extracting temporal information while preserving long-range dependencies in the motion sequence. Its structure is as follows:

$$\begin{aligned} \mathcal{D}^0 &= \text{LN}\left(\mathcal{D}_{\text{in}}\right), \\ \mathcal{D}_r^i &= \text{IN}\left(\Phi^+\left(\mu\left(\text{LN}\left(\mathcal{D}^{i-1}\right)\right)\right) + \Phi^-\left(\mu\left(\text{LN}\left(\mathcal{D}^{i-1}\right)\right)\right)\right), \\ \mathcal{D}^i &= \text{IN}\left(\text{LN}\left(\mathcal{D}^{i-1}\right)\right) + \mathcal{D}_r^i, \\ \mathcal{D}^{\text{res}} &= \text{LN}\left(\mathcal{D}_{\text{in}}\right) + \mathcal{D}^3, \end{aligned} \tag{7}$$

where LN$(\cdot)$ denotes a linear layer, IN$(\cdot)$ represents an InstanceNorm layer, $\Phi^+(\cdot)$ refers to forward SSM, and $\Phi^-(\cdot)$ to backward SSM. The original Mamba model, designed for 1-D sequences, is unsuitable for motion style transfer tasks requiring spatial awareness, so we adopt bidirectional sequence modeling here. $\mu(\cdot)$ denotes a Causal Conv1D layer used for feature extraction, $\mathcal{D}_r^i$ and $\mathcal{D}^i$ represent the intermediate result and final output of the right branch in the $i-th$ iteration of the MSM Block respectively, where $i \in \{1, 2, 3\}$. Notably, $\mathcal{D}^0$ represents the input to the MSM Block, while $\mathcal{D}^{res}$ is the output of the residual network containing the MSM Block.

Following the MSM Block structure, we integrate the sequential modelling strengths of SSM with the contextual awareness of the attention mechanism, enhancing our ability to capture fine-grained temporal changes at each timestep of the motion sequence, thereby generating more naturally stylized motion:

$$\sigma = \text{IN}(\text{LN}(\mathcal{D}^{res})) + \text{MHA}(\text{LN}(\mathcal{D}^{res})), \tag{8}$$

where MHA$(\cdot)$ represents Multi-Head Attention and σ denotes the output of the residual network containing MHA. The output of the final MSM denoiser $\mathcal{D}_{out}$, is expressed by the following equation:

$$\mathcal{D}_{out} = \text{FFN}(\sigma) + \text{IN}(\sigma), \tag{9}$$

where FFN$(\cdot)$ represents Feed-Forward Network. With the integration of the Mamba model, our MSM denoiser is better equipped to capture the temporal information of long motions, enabling the generation of more coherent and natural stylized motion sequences.

3.4 Training Objectives

Our objective is to use the output of UMSD Attention as the condition c for the diffusion model, allowing the framework to learn the motion style s while retaining the motion content $\boldsymbol{m}$, thus generating stylized motion $\boldsymbol{m}_t^s$. However, existing diffusion-based motion style transfer methods [32] rely solely on a simple reconstruction loss, which results in insufficient preservation of motion content and style coherence. To address this, we propose the Diffusion-based content consistency loss (*i.e.*, Eq. 10) and style consistency

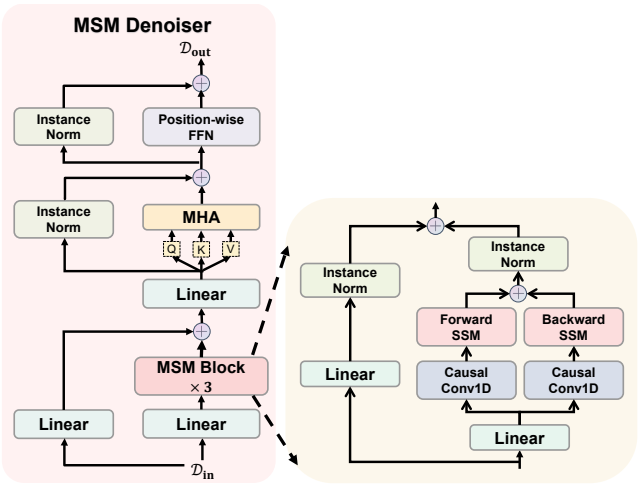

Figure 4: Architecture of MSM denoiser.

loss (Eq. 11), which ensure both accurate retention of content and faithful transfer of style throughout the motion generation process.

Diffusion-based Content Consistency Loss. When the input content motion sequence $\boldsymbol{m}_t^c$ and style motion sequence $\boldsymbol{n}_t^s$ share the same style, i.e., $c = s$, the generated stylized motion $\boldsymbol{m}_t^s$ should closely resemble the content motion $\boldsymbol{m}_t^c$, regardless of the content of style motion $\boldsymbol{n}_t^s$. Based on this observation, we model the distribution $p(\boldsymbol{m}_0^c|\mathrm{C})$ as a reverse diffusion process that iteratively denoises $\boldsymbol{m}_T^c$. Instead of predicting noise ϵ_t following the formula in DDPM [15], we adopt an equivalent approach from Ramesh *et al.* [27], directly predicting the motion itself. Specifically, $\hat{\boldsymbol{m}}_0^s = \hat{\boldsymbol{m}}_0^c = \mathrm{MSM}(\boldsymbol{m}_t^c, \boldsymbol{t}, \boldsymbol{C}) = \mathrm{MSM}(\boldsymbol{m}_t^c, \boldsymbol{t}, \mathrm{U}(\boldsymbol{m}_0^c, \boldsymbol{n}_0^s))$ (see Figure 3, right). The Diffusion-based Content Consistency Loss is expressed as:

$$\mathcal{L}_{\text{dcc}} = \mathbb{E}_{\boldsymbol{m}_0^c, \boldsymbol{n}_0^s \sim \mathcal{M}} \|\mathrm{MSM}(\boldsymbol{m}_t^c, \boldsymbol{t}, \mathrm{U}(\boldsymbol{m}_0^c, \boldsymbol{n}_0^s)) - \boldsymbol{m}_0^c\|_1, \quad (10)$$

where $\mathcal{M}$ denotes the dataset, $\mathrm{MSM}(\cdot)$ represents the Motion Style Mamba denoiser, and $\mathrm{U}(\cdot)$ denotes the UMSD Attention module. Our loss function differs fundamentally from the content consistency loss used in other methods [21, 24] in two main aspects: (1) it is based on a diffusion model, enabling control over the noise addition process to motion via the timestep $\boldsymbol{t}$; (2) the condition in our loss function is the result of fusing content and style motion features, allowing the framework to learn both features better.

Diffusion-based Style Consistency Loss. We adopt a similar approach as above. If content motion sequence $\boldsymbol{m}_t^c$ and style motion sequence $\boldsymbol{n}_t^s$ share the same content, i.e., $\boldsymbol{m} = \boldsymbol{n}$, the stylized motion $\boldsymbol{m}_t^s$ should ideally be as close as possible to the style motion $\boldsymbol{n}_t^s$. We use the MSM denoiser to directly predict the motion itself, i.e., $\hat{\boldsymbol{m}}_0^s = \hat{\boldsymbol{n}}_0^s = \mathrm{MSM}(\boldsymbol{n}_t^s, \boldsymbol{t}, \boldsymbol{C}) = \mathrm{MSM}(\boldsymbol{n}_t^s, \boldsymbol{t}, \mathrm{U}(\boldsymbol{m}_0^c, \boldsymbol{n}_0^s))$. The diffusion-based style consistency loss is expressed as follows:

$$\mathcal{L}_{\text{dsc}} = \mathbb{E}_{\boldsymbol{m}_0^c, \boldsymbol{n}_0^s \sim \mathcal{M}} \|\mathrm{MSM}(\boldsymbol{n}_t^s, \boldsymbol{t}, \mathrm{U}(\boldsymbol{m}_0^c, \boldsymbol{n}_0^s)) - \boldsymbol{n}_0^s\|_1. \quad (11)$$

Additionally, we adopt three existing geometric losses, $\mathcal{L}_{\text{pos}}$, $\mathcal{L}_{\text{foot}}$, and $\mathcal{L}_{\text{vel}}$, which control positions, foot contact, and velocities, respectively [30, 34, 35]. Our total training loss function is a combination of the above five losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{dcc}} + \mathcal{L}_{\text{dsc}} + \mathcal{L}_{\text{pos}} + \mathcal{L}_{\text{vel}} + \mathcal{L}_{\text{foot}}. \quad (12)$$

4 Experiments

In this section, we conduct a series of experiments to evaluate the effectiveness of the UMSD framework. First, we provide the details of the implementation of the experiments. Then, we perform both quantitative and qualitative comparisons between UMSD and state-of-the-art (SOTA) methods on two datasets. In the qualitative comparison, we separately assess the model's motion style transfer capability for both short and long motion sequences.

Additionally, to further evaluate UMSD's generalization ability, we test style transfer using previously unseen styles from the training dataset. In the ablation study, we demonstrate the effectiveness of each module in the UMSD framework. A user study is also conducted to compare our method's performance with SOTA methods from a more intuitive perspective. More experimental results can be found in the technical appendix.

4.1 Implementation Details

We train and test our model based on the Xia dataset [38] and BFA dataset [1]. We reduce the original 120fps motion data to 60fps and obtain approximately 1500 motion sequences in total. Our framework is implemented in PyTorch and trains on NVIDIA A800 GPUs, with a learning rate of e^{-6}, using the AdamW optimizer [20]. The training process takes about 10 hours.

4.2 Quantitative Evaluation

We employ the metrics of FMD, KMD, Diversity, CRA, and SRA [16, 25, 31, 32] to evaluate our framework quantitatively. The first two metrics are variants of Fréchet Inception Distance (FID) [14] and Kernel Inception Distance (KID) [22], respectively, measuring the distribution discrepancy between generated and real motion sequences. Diversity quantifies the diversity of generated motions. We train our feature extractor to compute the values of these three metrics. Lower FMD and KMD values indicate that the generated motions are closer to real motions, implying higher generation quality. Conversely, higher Diversity values reflect more extraordinary richness in the generated motions. We compute the content preservation degree and style recognition accuracy of the generated motions using our self-trained content classifier and style classifier for the CRA and SRA metrics. Higher values of these two metrics indicate better quality of the generated stylized motions.

We conduct quantitative comparisons on two mainstream datasets, the Xia [38] and BFA [1] datasets. The results are presented in Tables 1 and 2. It can be observed that our proposed method outperforms SOTA methods [1, 18, 31, 32] on most metrics in the Xia dataset [38]. This superior performance primarily stems from our UMSD framework's ability to facilitate sufficient information interaction between content and style motions. Although our method lags behind Aberman *et al.*'s method [1] regarding Diversity on the Xia dataset, this is mainly because our model architecture prioritizes generating more natural and realistic stylized motions rather than maximizing motion diversity.

Our UMSD framework achieves even more outstanding results on the long-sequence BFA dataset [1], surpassing SOTA methods across all metrics. This significant improvement primarily benefits from our proposed MSM denoiser, which leverages the State Space Model's (SSM) powerful sequence modelling capability to better

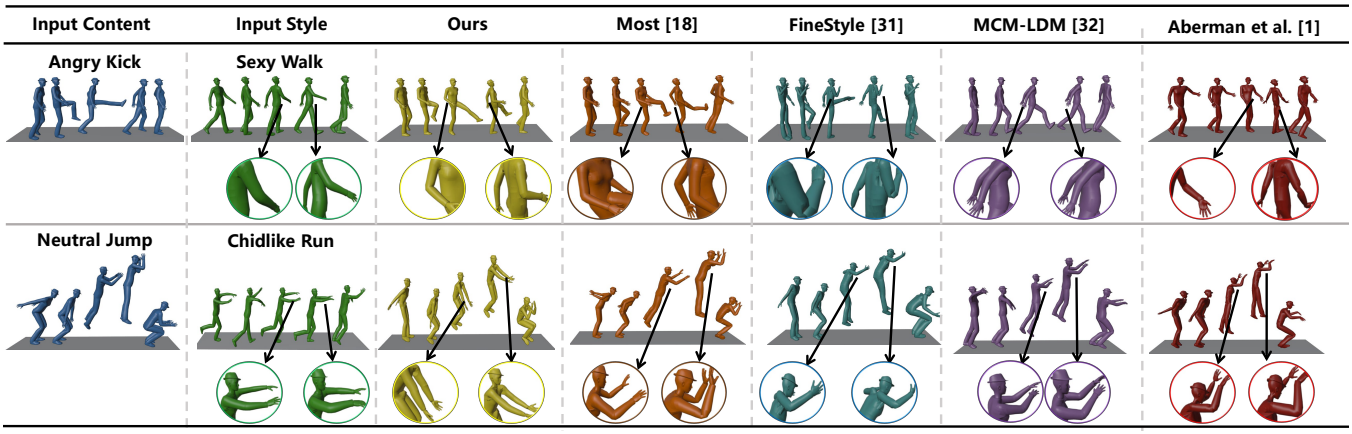

Figure 5: Qualitative comparison of short sequences. We provide two sets of cases comparing the style transfer effects with SOTA methods [1, 18, 31, 32]. We zoom in on critical areas that reflect style characteristics for a more intuitive assessment. The results indicate that our UMSD framework performs better.

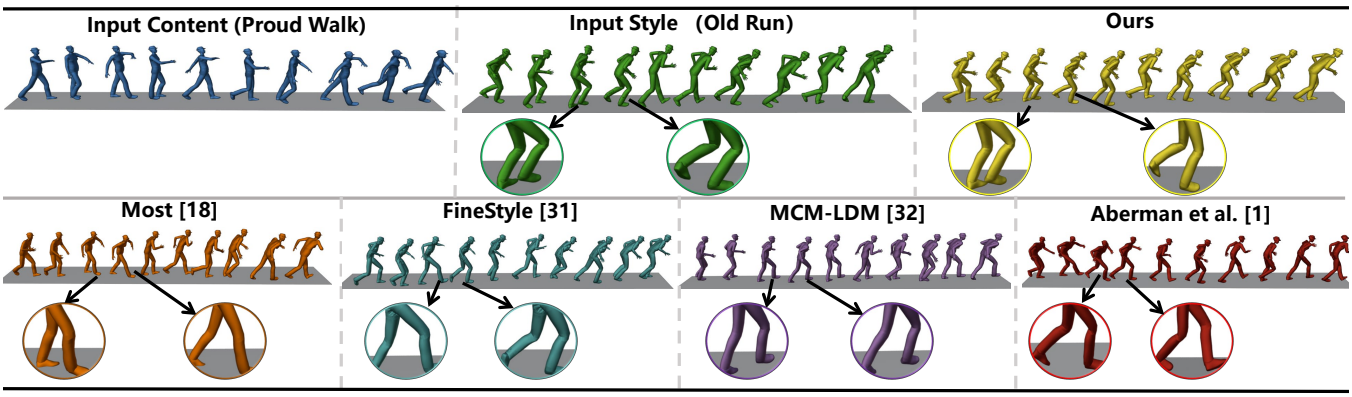

Figure 6: Qualitative comparison of long sequences. We select two long-sequence samples from the Xia dataset [38] for motion style transfer. The results demonstrate that our UMSD framework outperforms SOTA methods [1, 18, 31, 32].

capture temporal dependencies in motion sequences. The MSM denoiser generates more natural and coherent motion sequences by effectively modelling long-range temporal relationships.

4.3 Qualitative Evaluation

We qualitatively compare the visual effects of motion style transfer between our proposed UMSD framework and SOTA methods [1, 18, 31, 32] from three aspects: style expressiveness, content preservation, and motion realism. The content motion and style motion used in the experiments are sourced from the Xia dataset [38]. Ideally, the model can learn the style of the style motion while preserving the content of the content motion, thereby generating stylized motion that combines the characteristics of both.

Short Sequence Evaluation. As shown in Figure 5, we conduct two sets of motion style transfer experiments with motions fewer than 190 frames to fairly compare our UMSD with other methods trained on short motion clips. The results demonstrate that our UMSD framework generates better visual effects. For example, in the first row, we transfer the sexy style to the kick motion. In the input style motion, the right arm is at a right angle, and the left arm hangs straight down. Our generated kick motion maintains these characteristics. Other methods [1, 18, 31, 32], however, fail to capture this feature and produce awkward motions. This is due to our use of a one-stream architecture, where style and content information exchanged during encoding. Thus, our style transfer achieves both effective style representation and content preservation.

Long Sequence Evaluation. To evaluate the ability of our framework to generate long motion sequences, we select two samples from the Xia dataset [38], each with over 190 frames, as long-sequence content and style motions for qualitative comparison. As shown in Figure 6, when transferring the old style to the walk motion, the motion generated by our method shows more pronounced leg curvature, a smaller gap between the thighs, and greater consistency with the input style compared to SOTA methods [1, 18, 31, 32]. It also better captures the characteristics of the old style, demonstrating stronger style expressiveness. This superior performance is primarily attributed to the powerful sequence modelling capability

Table 1: Quantitative evaluation on Xia dataset [38]. '↑' ('↓') indicates that the value is better if the metric is larger (smaller); The bold fonts denote best performers. The results demonstrate that our UMSD outperforms SOTA in terms of overall quality and diversity.

Methods	FMD↓	KMD↓	Diversity↑	CRA↑	SRA↑
Aberman *et al.* [1]	24.15	0.94	**3.06**	36.36	54.55
MCM-LDM [32]	22.94	1.09	2.68	62.50	50.23
FineStyle [31]	23.43	0.97	2.94	73.21	46.15
MoST [18]	22.13	1.01	2.66	75.00	40.42
UMSD (Ours)	**16.45**	**0.65**	3.04	**82.35**	**58.82**

Table 2: Quantitative evaluation on BFA dataset [1]. The results show that our UMSD outperforms SOTA methods, even on long-sequence datasets.

Methods	FMD↓	KMD↓	Diversity↑	CRA↑	SRA↑
Aberman *et al.* [1]	29.76	1.69	2.80	54.49	19.46
MCM-LDM [32]	31.15	1.85	2.81	71.43	35.71
FineStyle [31]	23.54	1.19	2.05	47.62	14.29
MoST [18]	28.29	1.60	2.43	56.25	12.50
UMSD (Ours)	**21.63**	**0.88**	**3.15**	**82.55**	**46.53**

of the SSM module, which more effectively captures the temporal information of the motion sequence, thereby preserving long-term temporal dependencies within the sequence.

4.4 Generalizability Evaluation

Our model can extract styles from arbitrary motion clips. However, in practical applications, motion style transfer models will likely encounter style categories outside the training dataset. Whether the model can still transfer styles from unseen styles determines its generalizability and practical utility.

To compare the generalizability of our proposed UMSD framework with other methods, we select two styles, "happy" and "sneaky", from the BFA dataset [1] (which is not involved in training) as unseen style A and unseen style B (second column), respectively, for testing. We then transfer these two unseen styles to two content motions, "run" and "jump" (first column), and compare the stylized motions generated by different methods.

The results are shown in Figure 7, where we zoom in on key body parts that reflect stylistic characteristics. In the first row, our UMSD framework successfully captures the arm-spreading motion when handling unseen style A. In contrast, other SOTA methods [1, 18, 31, 32] generate unnatural movements and fail to transfer the style effectively. In the second row, we transfer unseen style B to a proud jump motion—our method preserves the original jumping motion while accurately adopting the spinal curvature characteristic of the unseen style, while other approaches fail to achieve this. The generalizability comparison shows that our UMSD exhibits stronger generalization capability and practical utility, making it more effective for real-world applications in multimedia fields such as film production, game design, and virtual reality.

Table 3: Ablation study on Xia dataset [38]. We ablate the key modules of our UMSD, including UMSD Attention, SSM, MHA, loss functions. The results validate their importance.

Settings	FMD↓	KMD↓	Diversity↑	CRA↑	SRA↑
Ours w/o UMSD Attention	18.83	0.89	2.74	80.03	57.36
Ours w/o SSM	18.08	0.79	2.90	78.31	55.04
Ours w/o MHA	19.19	0.73	2.97	79.96	54.86
Ours w/o $\mathcal{L}_{\text{dcc}}$	17.79	0.80	2.81	72.35	51.70
Ours w/o $\mathcal{L}_{\text{dsc}}$	19.54	0.97	2.78	74.68	48.34
Ours (Full)	**16.45**	**0.65**	**3.04**	**82.35**	**58.82**

4.5 Ablation Study

Here, we conduct ablation experiments on our key components in Table 3, including UMSD Attention, SSM, MHA, loss functions $\mathcal{L}_{\text{dcc}}$ and $\mathcal{L}_{\text{dsc}}$. In Table 4, we perform a module comparison of our MSM with alternative structures, such as STGCN and iTransformer.

Importance of UMSD Attention and Loss Functions. To evaluate the effectiveness of the UMSD attention module, we remove it and instead use two independent encoders to extract content and style motion features separately. These features are then fused to generate stylized motion. As shown in Table 3, all evaluation metrics exhibit degradation, demonstrating that our one-stream structure enables more comprehensive information exchange, thereby reducing misalignment in high-dimensional space. We also conduct ablation experiments on the loss functions, confirming their effectiveness in constraining the model to capture motion content and style features better.

Importance of the MSM Denoiser. We conduct ablation studies on the MSM Denoiser to ensure a fair comparison. Specifically, we replace the MSM module with STGCN [40] and iTransformer [19], respectively, retrain the framework, and perform comparative experiments. As shown in Table 4, the performance deteriorates significantly after substituting the MSM module with either of these alternatives. Our MSM module surpasses both alternatives across all quantitative evaluation metrics.

These results demonstrate that STGCN [40] and iTransformer [19] structures are substantially inferior to our model in capturing the global temporal dynamics of motion sequences, particularly in maintaining long-range temporal dependencies. Furthermore, this comparison highlights our method's superior sequence modelling capability, which more effectively captures temporal information in motion sequences and generates more coherent and natural stylized motion sequences.

Additionally, we conduct ablation studies on the MSM Denoiser's internal SSM and MHA structures in Table 3. Removing either results in degraded performance, demonstrating that the combination of SSM's sequential modelling strengths and the contextual awareness of the attention mechanism effectively captures fine-grained variations in motion sequences, leading to optimal results.

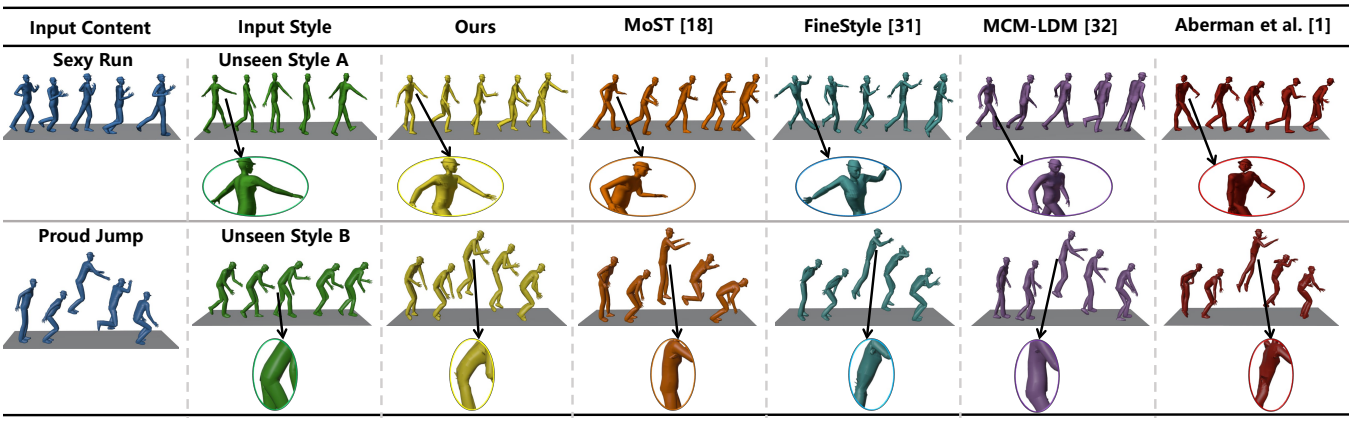

Figure 7: Generalization evaluation. We select a unseen style from the BFA dataset [1], which are not included in the training process, and transfer them to the sexy run and proud jump motions. The stylized motions are presented in the last four columns. Our method achieves better results than SOTA methods.

Table 4: Module comparison. We replace the MSM module with STGCN and iTransformer. The results demonstrate the importance of the MSM module within our framework.

Modules	FMD↓	KMD↓	Diversity↑	CRA↑	SRA↑
STGCN [40]	26.40	1.93	2.45	18.18	11.09
iTransformer [19]	27.24	2.46	2.13	15.52	13.03
MSM (Ours)	**16.45**	**0.65**	**3.04**	**82.35**	**58.82**

4.6 User Study

In addition to qualitative and quantitative comparisons, we conduct a user study to evaluate various methods' motion style transfer results. We convert the generated stylized motions into videos and include them in a questionnaire. 50 volunteers assess the motions based on three criteria: (1) Content Preservation (CP): Does the generated motion retain the content of the content motion? (2) Style Expressiveness (SE): Does the generated motion capture the style characteristics? (3) Motion Realism (MR): Is the generated motion realistic and natural? Volunteers rate each criterion from 1 (not achieved) to 10 (fully achieved).

After collecting all the questionnaires, we set the confidence level of 95% and calculate the average scores from the 50 volunteers. The results, shown in Table 5, indicate that our method achieves the highest CP, SE, and MR scores, with particularly outstanding performance in MR. Additionally, we perform an ANOVA test to examine the significance of these differences. The overall ANOVA establishes considerable distinctions among CP (F=14.847, p<0.01), SE (F=8.299, p<0.01), and MR (F=32.313, p<0.01). The post-hoc analysis suggests that our UMSD framework scores significantly higher than other methods [1, 18, 31, 32] across all three metrics (all p<0.01).

These results further demonstrate the superior performance of our framework. The UMSD framework achieves outstanding results primarily through its one-stream structure, which extracts features from content and style motions while enabling effective information interaction. Additionally, our proposed MSM denoiser utilizes the powerful sequence modelling capability of state space models to generate more coherent and natural motions, further enhancing the realism of stylized motions.

Table 5: User study results. Our UMSD framework outperforms SOTA methods in terms of Content Preservation (CP), Style Expressiveness (SE), and Motion Realism (MR).

Methods	CP↑	SE↑	MR↑
Aberman *et al.* [1]	$6.52^{\pm 0.33}$	$7.21^{\pm 0.48}$	$6.62^{\pm 0.77}$
MCM-LDM [32]	$8.15^{\pm 0.67}$	$8.70^{\pm 0.15}$	$5.81^{\pm 0.13}$
FineStyle [31]	$8.30^{\pm 0.39}$	$8.27^{\pm 0.60}$	$7.01^{\pm 0.47}$
MoST [18]	$8.16^{\pm 0.32}$	$7.42^{\pm 0.53}$	$7.60^{\pm 0.22}$
UMSD (Ours)	$\mathbf{9.59}^{\pm 0.27}$	$\mathbf{9.72}^{\pm 0.61}$	$\mathbf{9.12}^{\pm 0.48}$

5 Conclusion

In this work, we propose the UMSD framework, which employs a one-stream structure to extract features from content and style motion, enabling comprehensive information interaction and avoiding the limitations of using two separate encoders. We also introduce the MSM denoiser, which, for the first time in motion style transfer, leverages the robust sequential modelling capacity of SSM to learn temporal information and enhance motion coherence. Additionally, we propose two loss functions to guide model training. Finally, extensive experiments on two benchmark datasets demonstrate that our method surpasses SOTA approaches. We hope our work inspires further research in this field, leading to more practical applications in real-world scenarios.

Future Work. The field of motion style transfer presents several promising directions for future research, including cross-modal style transfer and style transfer under physical constraints. The supplementary materials provide detailed discussions.

References

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

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

[3] Kenji Amaya, Armin Bruderlin, and Tom Calvert. 1996. Emotion from motion. In *Graphics Interface*, Vol. 96. 222–229.

[4] Shivangi Aneja, Justus Thies, Angela Dai, and Matthias Nießner. 2024. Facetalk: Audio-driven motion diffusion for neural parametric head models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21263–21273.

[5] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. 2024. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7310–7320.

[6] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. 2023. Executing your Commands via Motion Diffusion in Latent Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18000–18010.

[7] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. 2022. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13608–13618.

[8] Chengdong Dong, Vijayakumar Bhagavatula, Zhenyu Zhou, and Ajay Kumar. 2024. Towards More Accurate Fake Detection on Images Generated from Advanced Generative and Neural Rendering Models. *arXiv preprint arXiv:2411.08642* (2024).

[9] S Ali Etemad and Ali Arya. 2014. Classification and translation of style and affect in human motion using RBF neural networks. *Neurocomputing* 129 (2014), 585–595.

[10] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. 2024. Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems* 36 (2024).

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems* 27 (2014).

[12] Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752* (2023).

[13] Bo Han, Hao Peng, Minjing Dong, Yi Ren, Yixuan Shen, and Chang Xu. 2024. AMD: Autoregressive Motion Diffusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 2022–2030.

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems* 30 (2017).

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

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

[17] Bingyi Kang, Xiao Ma, Chao Du, Tianyu Pang, and Shuicheng Yan. 2024. Efficient diffusion policies for offline reinforcement learning. *Advances in Neural Information Processing Systems* 36 (2024).

[18] Boeun Kim, Jungho Kim, Hyung Jin Chang, and Jin Young Choi. 2024. MoST: Motion Style Transformer between Diverse Action Contents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1705–1714.

[19] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. 2023. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625* (2023).

[20] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

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

[22] Bińkowski Mikołaj, J Sutherland Dougal, Arbel Michael, and Gretton Arthur. 2018. Demystifying mmd gans. In *International Conference on Learning Representations*. 1–36.

[23] Yuxuan Mu, Xinxin Zuo, Peng Dai, Youliang Yan, Juwei Lu, Li Cheng, et al. 2023. Generative Human Motion Stylization in Latent Space. In *International Conference on Learning Representations*.

[24] Soomin Park, Deok-Kyeong Jang, Sung-Hee Lee, Deok-Kyeong Jang, and Sung-Hee Lee. 2021. 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 (2021), 1–17.

[25] Sigal Raab, Inbal Leibovitch, Peizhuo Li, Kfir Aberman, Olga Sorkine-Hornung, and Daniel Cohen-Or. 2023. Modi: Unconditional motion synthesis from diverse data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13873–13883.

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

[27] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

[28] Rahul Ravishankar, Zeeshan Patel, Jathushan Rajasegaran, and Jitendra Malik. 2024. Scaling Properties of Diffusion Models for Perceptual Tasks. *arXiv preprint arXiv:2411.08034* (2024).

[29] Olaf Ronneberger, Philipp Fischer, Thomas Brox, Philipp Fischer, and Philipp Fischer. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*. 234–241.

[30] Mingyi Shi, Kfir Aberman, Andreas Aristidou, Taku Komura, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. 2020. Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. *Acm Transactions on Graphics* 40, 1 (2020), 1–15.

[31] Wenfeng Song, Xingliang Jin, Shuai Li, Chenglizhao Chen, Aimin Hao, and Xia Hou. 2023. FineStyle: Semantic-Aware Fine-Grained Motion Style Transfer with Dual Interactive-Flow Fusion. *IEEE Transactions on Visualization and Computer Graphics* (2023).

[32] Wenfeng Song, Xingliang Jin, Shuai Li, Chenglizhao Chen, Aimin Hao, Xia Hou, Ning Li, and Hong Qin. 2024. Arbitrary Motion Style Transfer with Multi-condition Motion Latent Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 821–830.

[33] Yixuan Sun, Zhangyue Yin, Haibo Wang, Yan Wang, Xipeng Qiu, Weifeng Ge, and Wenqiang Zhang. 2024. Pixel-level Semantic Correspondence through Layout-aware Representation Learning and Multi-scale Matching Integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17047–17056.

[34] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. 2022. Human motion diffusion model. *arXiv preprint arXiv:2209.14916* (2022).

[35] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. 2023. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 448–458.

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

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017).

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

[39] Xingqian Xu, Jiayi Guo, Zhangyang Wang, Gao Huang, Irfan Essa, and Humphrey Shi. 2024. Prompt-free diffusion: Taking" text" out of text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8682–8692.

[40] Bing Yu, Haoteng Yin, Zhanxing Zhu, Haoteng Yin, and Zhanxing Zhu. 2018. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 3634–3640. doi:10.24963/ijcai.2018/505

[41] Jiaxu Zhang, Xin Chen, Gang Yu, and Zhigang Tu. 2024. Generative Motion Stylization of Cross-structure Characters within Canonical Motion Space. In *ACM Multimedia*.

[42] Lei Zhong, Yiming Xie, Varun Jampani, Deqing Sun, and Huaizu Jiang. 2024. SMooDi: Stylized Motion Diffusion Model. *arXiv preprint arXiv:2407.12783* (2024).

[43] Qiang Zhou, Shaofeng Zhang, Nianzu Yang, Ye Qian, and Hao Li. 2024. Motion Control for Enhanced Complex Action Video Generation. *arXiv preprint arXiv:2411.08328* (2024).