

AStF: Motion Style Transfer via Adaptive Statistics Fusor

Hanmo Chen^{*}
hmc@stu.xidian.edu.cn
Hangzhou Institute of
Technology, Xidian
University
Hangzhou, Zhejiang
China

Chenghao Xu^{*}
chx@stu.xidian.edu.cn
Xidian University
Xi'an, Shaanxi, China

Jiexi Yan[†]
yanjiexi@xidian.edu.cn
Xidian University
Xi'an, Shaanxi, China

Cheng Deng[†]
chdeng@mail.xidian.edu.cn
Xidian University
Xi'an, Shaanxi, China

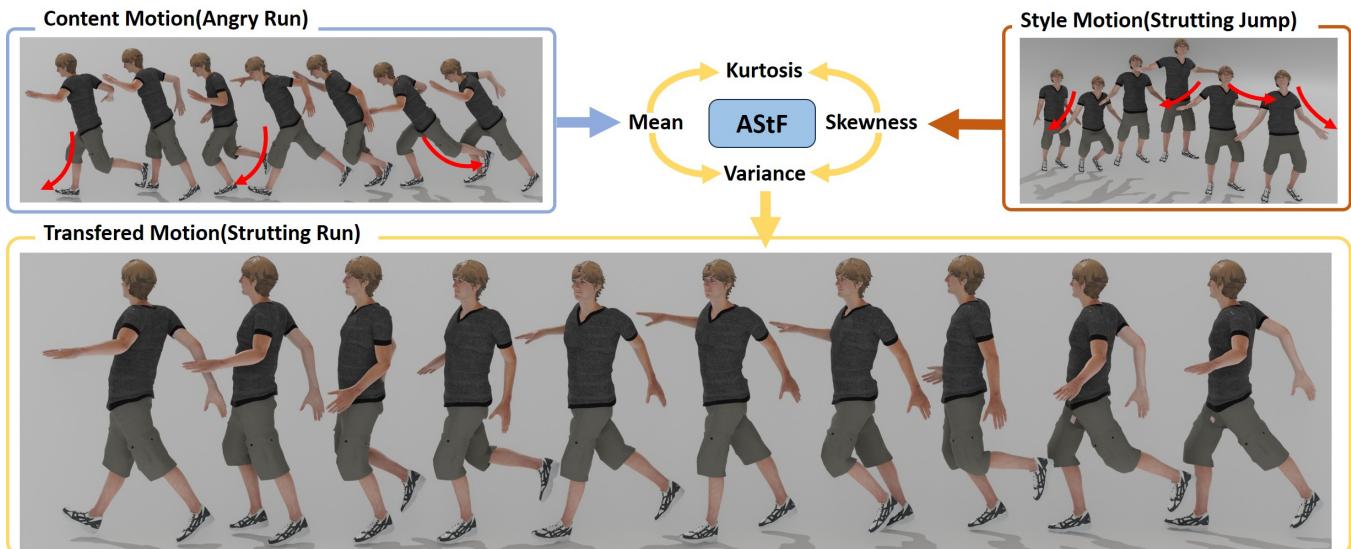


Figure 1: Motion style transfer via Adaptive Statistics Fusor(AStF). The left red solid curved represents the content we desire to preserve, while the right red solid curved arrow indicates the action with style characteristics, which is the style we aim to transfer.

Abstract

Human motion style transfer allows characters to appear less rigidity and more realism with specific style. Traditional arbitrary image style transfer typically process mean and variance which is proved effective. Meanwhile, similar methods have been adapted for motion style transfer. However, due to the fundamental differences between images and motion, relying on mean and variance is insufficient to fully capture the complex dynamic patterns and spatiotemporal coherence properties of motion data. Building upon this, our key insight is to bring two more coefficient, skewness and kurtosis, into the analysis of motion style. Specifically, we propose a novel

Adaptive Statistics Fusor (AStF) which consists of Style Disentanglement Module (SDM) and High-Order Multi-Statistics Attention (HOS-Attn). We trained our AStF in conjunction with a Motion Consistency Regularization (MCR) discriminator. Experimental results show that, by providing a more comprehensive model of the spatiotemporal statistical patterns inherent in dynamic styles, our proposed AStF shows proficiency superiority in motion style transfers over state-of-the-arts. Our code and model are available at <https://github.com/CHMimilanlan/AStF>.

CCS Concepts

- Computing methodologies → Motion processing; Motion capture; Unsupervised learning; Neural networks.

Keywords

Style Transfer, Contrastive Learning, Motion Generation

ACM Reference Format:

Hanmo Chen, Chenghao Xu, Jiexi Yan, and Cheng Deng. 2025. AStF: Motion Style Transfer via Adaptive Statistics Fusor. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3746027.3754938>

^{*}Both authors contributed equally to this research.

[†]Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '25, Dublin, Ireland

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 979-8-4007-2035-2/2025/10
<https://doi.org/10.1145/3746027.3754938>

1 Introduction

Recent years have witnessed growing attention on computer-generated human animation [4, 6, 32, 36], owing to its critical role in bridging the gap between virtual character behaviors and human perceptual realism [45]. This technology finds extensive applications in virtual reality system, video game, and anime. Within this domain, human motion constitutes a fundamental component whose quality directly determines animation verisimilitude [5, 7, 30, 47]. In the real world, human motions express multiple characteristics including emotion, physiological conditions, age et al., which collectively form motion style descriptors. This understanding has propelled motion style transfer into a forefront research challenge [29, 44, 52]. The primary objective of motion style transfer is to seamlessly transfer stylistic traits from a stylized motion sequence to another content motion sequence.

The primary challenge in motion style transfer lies in effectively disentangling and integrating motion style while preserving the physical plausibility of the content [8, 21, 37, 43]. Early approaches [16, 17] employed feedforward neural networks to disentangle motion styles, enabling direct style transfer. However, these methods often lacked diversity in the generated styles. More recently, inspired by the success of image style transfer techniques such as AdaIN [19], AdaAttn [28], and StyA2K [51], contemporary motion style transfer methods [14, 20, 23] have widely adopted the AdaIN paradigm. This approach adaptively modulates the normalized content features using the mean and standard deviation of the style features, aligning the statistical distribution of the content features with that of the style, thereby achieving effective style transfer.

Despite the significant progress achieved, the effectiveness of existing AdaIN-based methods remains limited when applied to motion style transfer due to fundamental differences between static images and dynamic motions. In image style transfer, style can be mostly regarded as an overall mixture of colors and textures, essentially analogous to a photographic filter. In most image style transfer cases, stylization intensity contributes equally to all region of an image without noticeable asymmetry. It is uncommon for a small region to be heavily stylized while another remains unaffected. Therefore, utilizing mean and variance is sufficient to capture perceptually relevant style information. However, motion styles inherently involve complex spatiotemporal dynamics and asymmetric variations across multiple dimensions [2, 12, 46], including joint position, velocity, and acceleration, as well as trajectory patterns. For example, angry motion styles often manifest as increased limb acceleration, whereas styles associated with elderly individuals exhibit slower movement velocities. Even within the same motion content, such as a punch, different joints may exhibit distinct dynamic behaviors—e.g., rapid arm acceleration while leg movement remains minimal. Moreover, a single joint can display diverse temporal behaviors across different time steps, further illustrating the asymmetric and temporally variant nature of motion [3, 27].

These complex motion dynamics suggest that effective motion style transfer necessitates modeling beyond global first- and second-order statistics, with particular emphasis on higher-order and joint-specific temporal characteristics. While prior approaches relying

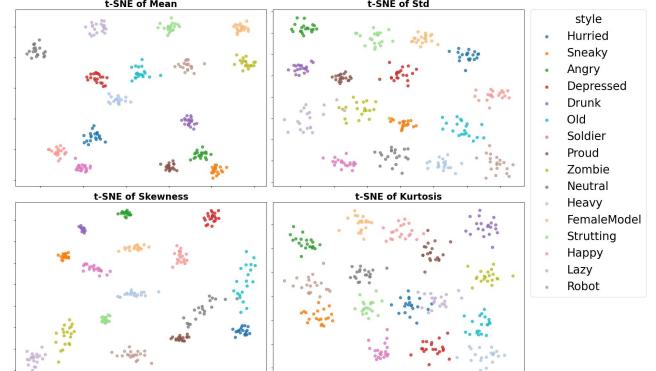


Figure 2: Statistics distribution differences across styles by t-SNE on BFA dataset [1].

solely on mean and variance offer a foundation for style disentanglement, they are inadequate for capturing the nuanced dynamics and asymmetries inherent in motion data. Building on this observation, we explore the incorporation of higher-order statistical descriptors. To validate our hypothesis, we visualize the distribution of various motion styles under different statistical metrics using t-SNE [39], as shown in Figure 2. The resulting scatter plot reveals substantial divergence in the distribution patterns across styles when higher-order statistics are considered. These findings suggest that metrics such as skewness and kurtosis may provide a more expressive and discriminative characterization of motion styles, motivating their integration into the modeling framework.

In this paper, inspired by Zhang et al. [49], we introduce skewness and kurtosis as higher-order dynamic statistics and propose a novel Adaptive Statistics Fusor (ASfF), comprising a Style Disentanglement Module (SDM) and a High-Order Multi-Statistics Attention mechanism (HOS-Attn). Within the ASfF framework, skewness captures the asymmetry in motion sequence distributions, while kurtosis reflects the dynamic intensity and variability of motion styles. Together with mean and variance, these four statistical measures constitute a comprehensive spatiotemporal representation that effectively disentangles the latent characteristics of motion styles. Following the extraction of these statistical features, the HOS-Attn module is designed to adaptively inject the disentangled style statistics into the content motion through spatiotemporal-aware weighting, ensuring both effective style integration and preservation of temporal coherence.

Additionally, contrastive learning is often used in motion style transfer, where discriminator significantly influences the generator. Existing methods [20, 23, 33] have not focused much on the configuration of the discriminator, as they simply input generated motion and style motion into the discriminator separately to compute adversarial loss, leading to a lack of correlation between generated motion and style motion within the discriminator. Therefore, inspired by Ko et al. [26], we propose the Motion Consistency Regularization (MCR) discriminator, which utilizes a correlation loss to guide the discriminator in capturing style consistency between the generated motion and the reference style motion.

To summarize, our contributions are listed as follows:

- We designed AStF by incorporating SDM and HOS-Attn. SDM effectively disentangles four statistics to capture the global, asymmetry, and dynamic features of motion, while HOS-Attn integrates the disentangled style statistics into the content motion, maintaining spatiotemporal coherence.
- We propose MCR discriminator which significantly enhances the style consistency of generated results in motion style transfer by increasing the feature similarity between style motion and generated motion, as well as the internal consistency within the style motion, in the discriminator.
- Our proposed AStF shows superiority performance in motion style transfer compared to existing methods.

2 Related Works

2.1 Image Style Transfer

Since stylized images possess certain artistic significance, style transfer is first applied to the image. Early research on image style transfer mainly relied on hand-craft features until Gatys et al. [13] pioneered transfer style from one image to another using a convolution neural network. To improve efficiency, Johnson et al. [22] combine the benefits of feedforward networks and perceptual loss functions, achieving faster and better transferring. Alternatively, CycleGAN (Zhu et al., 2017), transfer styles between images by utilizing Generative Adversarial Networks (GANs) with local similarity. Ulyanov et al. [38] enhanced stylization quality and better detail preservation by introducing Instance Normalization (IN). Building on IN, Huang et al. [19] introduced Adaptive Instance Normalization (AdaIN), which is a significant advancement that adjusts the mean and variance of content features to match those of an arbitrary style image. However, since AdaIN focus on mean and variance, which are global information and lack of local feature, Liu et al. [28] proposed AdaAttn, utilizing spatial attention score to dynamically calculate style feature statistics. For the purpose of efficiency, Zhu et al. introduced All-to-Key attention to reduce computational complexity and mitigate detail distortion.

2.2 Motion Style Transfer

Contrastive learning [9, 26] plays an essential role in motion style transfer. Holden et al. [17] adopts feedforward control network, achieving the first work in Motion style transfer. Since the effectiveness of AdaIN in image style transfer, Aberman et al. [1] incorporate AdaIN in motion style transfer, along with deterministic autoencoders. Since Generative flow [11] has proven to be effective in the domain of image generation, Wen et al. [41] introduced a generative flow based model, generating motions with less distortion. MoST, introduced by Kim et al. [23], addressed style transfer between motions with differing contents by employing a transformer [40] with part-attentive style modulation and Siamese encoders. Similarly, Guo et al. [14] introduced a generative approach in the latent space, decomposing content and style motion into latent codes to enable flexible stylization. Motion Puzzle, proposed by Jang et al. [20], provides local style editing capabilities via revised AdaIN and attention mechanisms. Notably, the aforementioned works all employed AdaIN. The achievements of latent diffusion model [34] have inspired its application in motion style transfer. Chen et al. [10]

proposed Motion Latent-based Diffusion (MLD), a diffusion process within a Variational AutoEncoder's [25] latent space generates conditional human motions from inputs like text or action classes. MLD provides a foundational model for subsequent diffusion-based motion style transfer works. Zhong et al. [50] proposed SMooDi, which incorporated style guidance and a lightweight adaptor in MLD. Hu et al. [18] fine-tune MLD with semantic-guided loss for few-shot style transfer using minimal style data. Song et al. [35] extended MLD with multi-condition, including disentangling trajectory, content, and style. These works mentioned above are effective in generating stylized motions across varied content and styles.

3 Method

3.1 Preliminary

Given a style motion M_S and a content motion M_C as inputs, motion style transfer aims to disentangle style features from M_S and apply them to M_C , thereby generating the stylized motion M_G . Existing motion stylization frameworks [1, 14, 23, 48] predominantly adopt Adaptive Instance Normalization (AdaIN), which is inherited image style transfer, as their core component. AdaIN achieves image style transfer by aligning the mean and variance of a content feature map with those of a style feature map, which can be formulated as:

$$\text{AdaIN}(x, y) = \sigma(y) \left(\frac{x - \mu(x)}{\sigma(x)} \right) + \mu(y). \quad (1)$$

Though AdaIN aligns feature between style and content through mean and variance matching, this approach suffers inefficiency when applied to motion sequences. Unlike static images, motion sequence exhibits two intrinsic temporal properties which are asymmetry and dynamic. Asymmetry is reflected in irregular motion trajectories over time, while dynamics is manifested through varying joint speeds and accelerations. Using only the mean and variance is insufficient for capturing high-order and fine-grained style features of motion. Additionally, as presented in Figure 2, the result of t-SNE indicates significant differences in the distribution of statistics across various styles. Consequently, we incorporate skewness and kurtosis to enhance motion style transfer effectiveness, proposing out AStF, which will be detailed in following subsections.

3.2 Method Overview

An overview of AStF pipeline is illustrated in Figure 3. Content motion ' M_C ' and style motion ' M_S ' are first processed through Content Statistics Encoder and Style Statistics Encoder respectively, generating latent motion embeddings e_S and e_C . In our Statistics Encoder, input motion is processed through projection layers to obtain motion embeddings. As previous works [14, 23] utilize a learnable style token to aggregate the style features across an entire motion sequence, which could lead to a lack of style feature extraction capability, our Statistics Encoder replace it with the statistics features of motion embeddings extracted by Simple SDM, which will be detailed in section 3.3. We concatenate the statistics features with the corresponding embedding and pass the result to transformer [40] encoder, generating latent code e_S and e_C . Subsequently, these two embeddings are input into SDM, projecting e_S and e_C to query Q , key K and value V , then calculate four statistics consist of mean, variance, skewness and kurtosis, represented by μ ,

σ^2 , γ and β respectively. This process is formulated as:

$$\text{SDM} : (e_s, e_c; \theta_{\text{SDM}}) \mapsto (\mu, \sigma^2, \gamma, \beta), \quad (2)$$

where θ_{SDM} denote the learnable parameter of SDM. Subsequently, four types of statistics are input into HOS-Attn along with Q to perform Statistics-wise Cross Attention and Gate Self Attention, which ultimately results in transferred motion embedding e_T . The HOS-Attn process is formulated as:

$$\text{HOS} : (\mu, \sigma^2, \gamma, \beta; \theta_{\text{HOS}}) \mapsto (e_T), \quad (3)$$

where θ_{HOS} denote the learnable parameter of HOS-Attn. After acquiring e_T , the motion decoder is utilized to decode the motion embedding e_T into the desired motion sequence M_G .

3.3 Style Disentanglement Module

Unlike conventional approaches that rely solely on mean and variance to align motion style distributions, we propose a Style Disentanglement Module by introducing skewness and kurtosis as higher-order statistics, which we named as SDM. In Statistical Analysis, skewness and kurtosis are higher-order statistics that describe the shape of a probability distribution. Skewness measures the asymmetry of a probability distribution, indicating whether data are concentrated on one side. Kurtosis quantifies the tailedness and peakedness of a distribution, revealing whether extreme values are more frequent or data are more dispersed. Specifically, as presented in Figure 3 (b), we defined Simple SDM and SDM. Simple SDM is implemented in the Statistic Encoder to replaced the learnable style token utilized in previous works [14, 23], achieving adaptive style features extraction. SDM take e_c and e_s as input, following instance normalization and linear to obtain the key triplet (Q, K, V) , representing query, key and value. We define mean as μ , variance as σ^2 , skewness as γ , kurtosis as β , we utilize the following definitions and formulas to calculate each statistics of (Q, K, V) :

$$\left\{ \begin{array}{l} \mu = \frac{1}{F} \sum_{f=1}^F x_{d,f,j}, \\ \sigma^2 = \frac{1}{F} \sum_{f=1}^F (x_{d,f,j} - \mu)^2, \\ \gamma = \frac{1}{F} \sum_{f=1}^F \left(\frac{x_{d,f,j} - \mu}{\sigma} \right)^3, \\ \beta = \frac{1}{F} \sum_{f=1}^F \left(\frac{x_{d,f,j} - \mu}{\sigma} \right)^4, \end{array} \right. \quad (4)$$

where d , f , and j indicate feature dimensions, frames, and joints, respectively. This calculation for each (Q, K, V) tensors yield 12 independent statistics (e.g., $\mu_Q, \sigma_Q^2, \gamma_Q, \beta_Q$ for Q). These statistics are then classified into four domain-specific statistics groups: $[\mu_Q, \mu_K, \mu_V]$, $[\sigma_Q^2, \sigma_K^2, \sigma_V^2]$, $[\gamma_Q, \gamma_K, \gamma_V]$, $[\beta_Q, \beta_K, \beta_V]$. This structured aggregation enables explicit modeling of both hierarchical dynamics and cross-tensor correlations, providing a compact and expressive representation for motion style transfer.

3.4 High-Order Multi-Statistics-Attention

To holistically fuse the statistical characteristics of motion features and enhance the correlation between content and style embeddings, we propose HOS-Attn. As presented in Figure 3 (c), HOS-Attn explicitly models four domain-specific statistics groups extracted from the SDM through inter-domain interaction and cross-domain interaction. For each domain-specific statistics group, HOS-Attn computes its cross-domain interaction through Statistics-wise Cross-Attention, which applies dedicated cross-attention to each group of statistics, outputting refined statistics features $\mu_c, \sigma_c^2, \gamma_c, \beta_c$. The refined statistics features are subsequently concatenated with Q to form an augmented feature. This augmented feature is then processed by cross-domain interaction through Gate Self-Attention to integrate global statistical features. Inside the Gate Self-Attention, we first apply self-attention to the augmented feature, then discard the concatenated statistical terms, resulting in output F_c . To preserve the integrity of the original motion semantics while incorporating statistical guidance, a gating residual mechanism is introduced to produce the final output F_o :

$$F_o = \lambda_g \times F_c \oplus (1 - \lambda_g) \times Q, \quad (5)$$

where \oplus indicated to element-wise addition. The similarity coefficient λ_g is calculated based on cosine similarity. The formula of λ_g is as follows:

$$\lambda_g = \frac{\frac{Q \cdot K}{\|Q\| \|K\|} + 1}{2}. \quad (6)$$

3.5 Motion Consistency Regularization

In motion style transfer, prior works equip the discriminator with only a coarse style-classification head, leading to the issue of style fading which is characterized by weak expression and temporal degradation of style in the generated motion. Our analysis suggests the root cause lies in the discriminator's reliance on global features while ignoring local features, allowing the generator to "cheat". Inspired by Ko et al. [26], we introduce Motion Consistency Regularization (MCR) to enhances the discriminator with two complementary roles, style-style and style-generation, to address this problem. As presented in Figure 3(d), in style-style role, MCR splits a style motion into two subsequences to evaluate internal coherence. In style-generation role, MCR computes the similarity between style and generated motion to capture global features. This design effectively mitigates the problem of style fading.

Specifically, MCR discriminator consists of a feature extractor D_0 followed by a classification head D_1 . By taking style motion M_S as input, style-style role first apply random crop to M_S , separating M_S into two subsequences s_1 and s_2 . Then we obtain the intermediate feature z_1 by $z_1 = D_0(s_1)$, and similarly obtain z_2 . Subsequently, we further process z_1 and z_2 by Simple SDM and MCR Module, a lightweight module comprising two linear layers with LeakyReLU activation, to obtain refined features r_1 and r_2 . We compute the MCR loss between r_1 and r_2 within their valid temporal region Ω on the basis of cosine similarity. This process is formulated as:

$$\text{sim}(r_1, z_2; \Omega) \equiv \sum_{(f) \in \Omega} -\frac{r_1[f]}{\|r_1[f]\|_2} \cdot \frac{z_2[f]}{\|z_2[f]\|_2}, \quad (7)$$

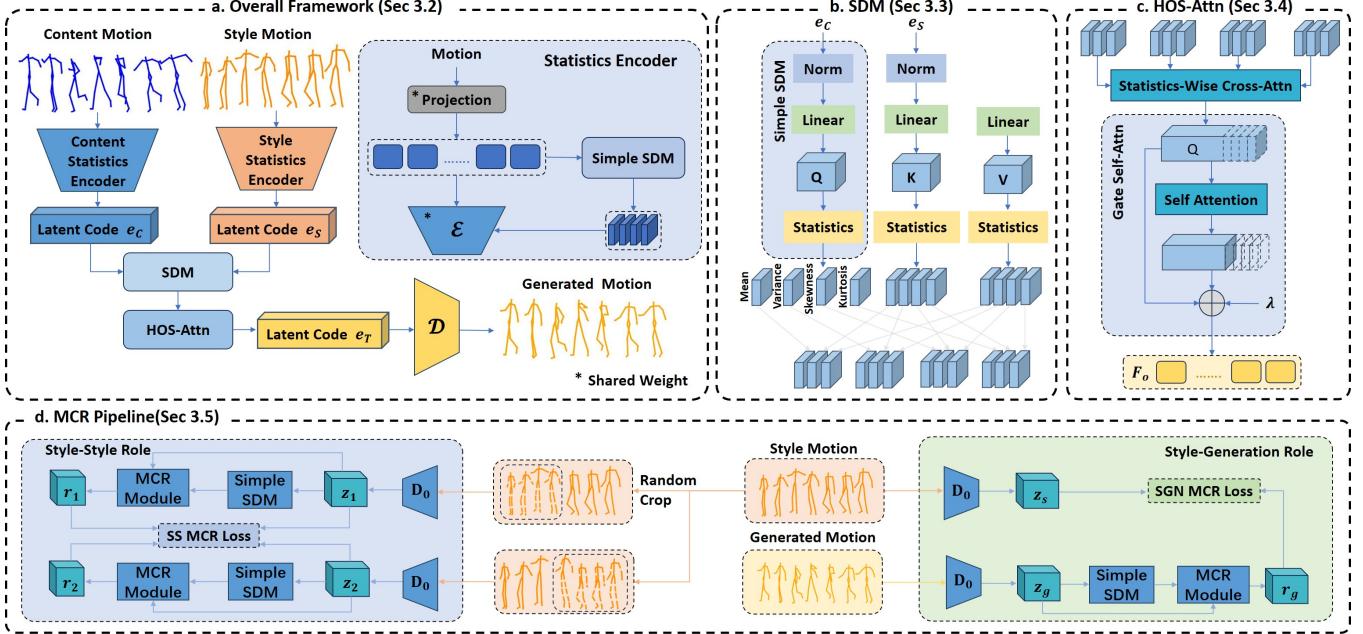


Figure 3: Method Overview. (a) Overall framework of AStF mainly comprising Content and Style Statistics-Encoder, Style Disentanglement Module (SDM), High-Order Multi-Statistics-Attention(HOS-Attn) and decoder. (b) Detailed structure of SDM and Simple SDM. (c) Detailed structure of HOS-Attn. (d) Pipeline of Motion Consistency Regularization (MCR) discriminator.

where $\|\cdot\|_2$ indicates ℓ_2 -norm, f indicates to frame index. The style-style MCR loss is formulated as:

$$\mathcal{L}_{ss} = \frac{1}{3} \text{sim}_{nc}(r_1, F_{sg}(z_2)) + \frac{1}{3} \text{sim}_{nc}(r_2, F_{sg}(z_1)), \quad (8)$$

where $F_{sg}(\cdot)$ indicates to stop-gradient. For the style generation component, the generated motion M_G is passed through D_0 to obtain the corresponding feature representation z_g . Similarly, the style reference motion M_S is processed by the same module to extract the feature z_s . The feature z_g is subsequently passed through the Simple SDM and MCR modules to obtain the refined representation r_g . This facilitates the computation of the style generation MCR loss, defined as follows:

$$\mathcal{L}_{sgn} = \frac{1}{3} \text{sim}_{nc}(r_g, F_{sg}(z_s)). \quad (9)$$

3.6 Loss Function

The training objective of our AStF is defined through loss function for the discriminator \mathcal{L}_D and the generator \mathcal{L}_G . As adversarial loss and R1 loss [31] are proven to be effective in the training of discriminator in previous works [1, 20, 23], we combined our MCR loss, which consists of style-style MCR loss and generation-style MCR loss in equation 8 and 9, with adversarial loss and R1 loss to generate our discriminator loss:

$$\mathcal{L}_D = \mathcal{L}_{adv} + \mathcal{L}_{R1} + \lambda_{MCR} \cdot (\mathcal{L}_{ss} + \mathcal{L}_{sgn}), \quad (10)$$

where λ_{MCR} is the hyperparameter, setting $\lambda_{MCR} = 1$ to balance the contributions. \mathcal{L}_G integrates multiple constraints for content-style disentanglement, which are commonly used in existing methods,

including adversarial (\mathcal{L}_{adv}^G), reconstruction (\mathcal{L}_r), content cycle consistency (\mathcal{L}_{cc}) and style cycle consistency (\mathcal{L}_{cs}) losses. However, during training, we observed that M_G tends to adhere to M_C , indicating insufficient style feature extraction and transfer. To address this issue, we propose a style-align loss \mathcal{L}_a that processes M_G through an encoder to obtain its latent representation e_g , then computes an L2 loss between e_g and the style latent code e_s . This mechanism effectively guides the style motion to properly embed its stylistic characteristics into the M_G . Consequently, our \mathcal{L}_G is formulated as follows:

$$\mathcal{L}_G = \mathcal{L}_{adv}^G + \lambda_r \mathcal{L}_r + \lambda_c (\mathcal{L}_{cc} + \mathcal{L}_{cs}) + \lambda_a \mathcal{L}_a, \quad (11)$$

where λ_* denotes the weight of each loss. We set $\lambda_r = 3$, $\lambda_c = 3$, $\lambda_a = 1$ in our experiment.

4 Experiments

In this section, we evaluate our proposed AStF and analyze its essential characteristics.

4.1 Experimental Settings

Experimental Implementation. For a fair comparison, we follow the experimental protocol of GenMoStyle [14] and take different style motions and content motions as input to analyze the performance of our method. We train our models on an NVIDIA A6000 48 GB. We optimize the generator and discriminator using separate Adam optimizers [24] with the same $\beta_1 = 0.9$ and $\beta_2 = 0.99$, and different learning rates of 1×10^{-5} and 1×10^{-6} , respectively. We set the batch size as 16 and trained for about 400K iterations. To

Table 1: Quantitative results on the test set of BFA [1] and Xia [42] datasets. Style FID ↓, Cnt FID ↓, Sty Acc ↑, Cnt Acc ↑, and Geo Dis ↓ denote style/content FID, style/content accuracy, and geodesic distance respectively. Bold font indicates the best result.

Method	BFA dataset			Xia dataset				
	Style FID ↓	Style Acc ↑	Geo Dis ↓	Style FID ↓	Content FID ↓	Style Acc ↑	Content Acc ↑	Geo Dis ↓
Aberman et al. [1]	5.412	0.525	0.631	0.602	0.00574	0.377	0.575	0.869
MotionPuzzle [20]	3.487	0.745	0.431	0.285	0.00541	0.764	0.573	0.568
Park et al. [33]	3.501	0.731	0.422	0.272	0.00574	0.775	0.481	0.767
MoST [23]	3.107	0.775	0.350	0.188	0.00151	0.699	0.867	0.398
GenMoStyle [14]	1.892	0.845	0.434	0.192	0.00147	0.873	0.817	0.435
AStF (ours)	1.558	0.905	0.414	0.157	0.00101	0.930	0.903	0.440

verify the effectiveness of the proposed model, we compare it with other state-of-art methods in motion style transfer: Aberman et al. [1], MotionPuzzle proposed by Jang et al. [20], Park et al. [33], and MoST proposed by Kim et al. [23], GenMoStyle proposed by Guo et al. [14].

Datasets. Following previous works, we evaluated our AStF on two publicly available motion datasets, Xia dataset [42] and BFA dataset [1]. Xia dataset [42] contains 8 distinct motion styles, including angry, childlike, and strutting, etc., and we categorize the contents into 6 classes, including walking, jumping, kicking, etc.. BFA dataset [1] consists of 16 motion styles, however, not be labeled by content. Both datasets have 31-joint skeleton, and we select 21 key joints for training and evaluation, following the approach proposed by Aberman et al. [1]. Given that each motion sequence in the Xia dataset varies in frame length, we downsample each sequence by a factor of two, then apply zero-padded until each motion sequence reaches a fixed frame length L_m , which we set to 200. For the BFA dataset, we directly split each Biovision Hierarchy (BVH) file into groups of 400 frames, discarding any remaining samples with fewer than 400 frames. We then apply downsampling by a factor of two, resulting in each sample having a fixed frame length of 200.

Evaluation Metrics. For the purpose of effectively demonstrating the robustness of our proposed methodology, we utilize five separate evaluation metrics to quantify both content retention and style fidelity of the generated motion with the input of content motion and style motion. We employ the content and style Fréchet Inception Distance (FID) [15], a widely accepted metric in previous work, to quantify the feature distribution similarity between the input motions and generated motion. We utilize recognition accuracy to measure semantic consistency. Additionally, we compute the geodesic distance [14] between the 3D rotation matrices of content motion and generated motion. For the calculation of content and style FID, we first trained a content classifier and a style classifier on the train set of Xia dataset and BFA dataset. The content classifier and the style classifier are applied to compute content and style accuracy.

4.2 Evaluation Results

Quantitative Result. In order to evaluate the content retention and style fidelity capability of our AStF, the the Fréchet Inception Distance (FID) [15], recognition accuracy and geodesic distance are

utilized as quantitative measures. The detailed numerical quantitative results for BFA dataset [1] and Xia dataset [42] are presented in Table 1. As indicated in Table 1, our AStF achieves the lowest style FID and content FID, as well as the highest style and content accuracy, along with effective geodesic distance. This demonstrates that our AStF outperforms in terms of style fidelity and content retention. In contrast, other methods exhibit certain shortcomings in specific terms. Aberman et al. [1] show the lowest accuracy and the highest FID, indicating failure in most test cases. MotionPuzzle [20] and Park et al. [33] perform well in style accuracy and style FID but has slight deficiencies in content retention, suggesting a weakness in preserving content. MoST [23] excels in content accuracy and FID, and achieves better results in geodesic distance compared to our AStF, however, shows weaknesses in style fidelity, which means MoST tends to align more closely with content motion with insufficient style feature extraction. GenMoStyle [14] yields a satisfactory results on both style fidelity and content retention, however, there is still a noticeable gap compared to our method.

Qualitative Result. To further intuitively demonstrate the effectiveness of our AStF, we generated motions on the test sets of the Xia dataset [42] and BFA dataset [1] with our trained AStF, the qualitative results are presented in Table 3. As Xia dataset [42] is labeled by content, we present more qualitative results on Xia dataset. Since Aberman et al. failed in most cases, we have omitted its visualizations for the sake of clarity. For the purpose of clear demonstration, we applied red indications in each figure in Table 3. Dashed circles and arrows indicate undesired content which may reflect to style characteristic in the content motion, while solid lines represent the desired content which need to be transferred from style motion to content motion. For instance, as in case (1) in Table 3, we take childlike jump as content motion and depressed walk as style motion. The action of arms-outstretched in the content motion is the content that we do not desire to retain, as it reflects a characteristic of childlike style. On the other hand, the head-down, hunched posture, which represents the style of depressed in the style motion, is what we aim to extract and transfer to the content motion. Additionally, the action of jump in content motion is also what we desire to retain. As presented in the experiment results of case (1), the generated motion of our AStF effectively transferred the depressed style while retaining the content jump. MoST [23] failed to discard the childlike characteristics in the content motion. MotionPuzzle [20] appeared not to adequately preserve the

Table 2: Ablation experiment results

Model Variants	Style FID ↓	Content FID ↓	Style Acc ↑	Content Acc ↑	Geo Dis ↓
w/o Entire MCR Loss	0.204	0.00130	0.723	0.856	0.452
w/o Style-Style MCR Loss	0.188	0.00118	0.782	0.881	0.437
w/o Style-Generation MCR Loss	0.191	0.00111	0.773	0.889	0.423
w/o Simple SDM	0.203	0.00222	0.710	0.851	0.343
w/o Style Align Loss	0.353	0.00218	0.637	0.838	0.245
w/o Skew	0.166	0.00113	0.805	0.847	0.430
w/o Kurt	0.160	0.00135	0.799	0.832	0.447
w/o Skew and Kurt	0.179	0.00168	0.718	0.857	0.450
$\lambda_{\text{MCR}} = 0.5$	0.198	0.00151	0.837	0.876	0.471
$\lambda_{\text{MCR}} = 2$	0.186	0.00194	0.875	0.801	0.523
$\lambda_{\text{MCR}} = 3$	0.207	0.00217	0.804	0.735	0.551
$\lambda_a = 0.5$	0.189	0.00128	0.871	0.901	0.416
$\lambda_a = 2$	0.175	0.00187	0.892	0.812	0.462
$\lambda_a = 3$	0.161	0.00198	0.919	0.784	0.501
AStF (Full Model)	0.157	0.00115	0.930	0.903	0.440

jump content. Park et al. [33] and Aberman et al. [1] were insufficient in transferring depressed style to content motion. Overall, the qualitative experimental results demonstrate that our AStF effectively achieves style fidelity and content retention, successfully transferring style while preserving content. While Most [23] and GenMoStyle [14] effectively preserving content, it fails to achieve sufficient style transfer, with generated motions tending to align more closely with content motion. MotionPuzzle [20], Aberman et al. [1] and Park et al. [33] show inadequate content preservation, producing awkward motions with indistinct styles.

4.3 Ablation Study

To validate the necessity of key components in our AStF, we conduct ablation experiments focusing on five critical elements: (1) proposed MCR Loss, (2) Simple SDM conducted in Statistics Encoder, (3) high-order statistics in SDM, (4) style-align loss, (5) hyper-parameters in our loss function. To thoroughly demonstrate the impact of these elements on content retention and style fidelity of our AStF, all experiments are conducted on the Xia dataset [42], since it has motions with labeled styles and contents. Table 2 presents the results of ablation experiments. Overall, through the aforementioned ablation experiments, we observed the significant impact of each module on enhancing the model’s style fidelity and content retention.

The Impact of proposed MCR Loss. The MCR Loss contains style-style MCR Loss and style-generation MCR Loss. We systematically evaluate their contributions through three ablation configurations: without style-style MCR loss, without style-generation MCR loss, and without the entire MCR loss. Our ablation experiments indicated that, both style-style and style-generation MCR Losses independently improved style and content FID and accuracy. The full model, compared to the variant without the entire MCR Loss, achieved a 8.7% improvement in style accuracy and a 4.7% improvement in content accuracy, highlighting the impact of MCR Loss on overall performance.

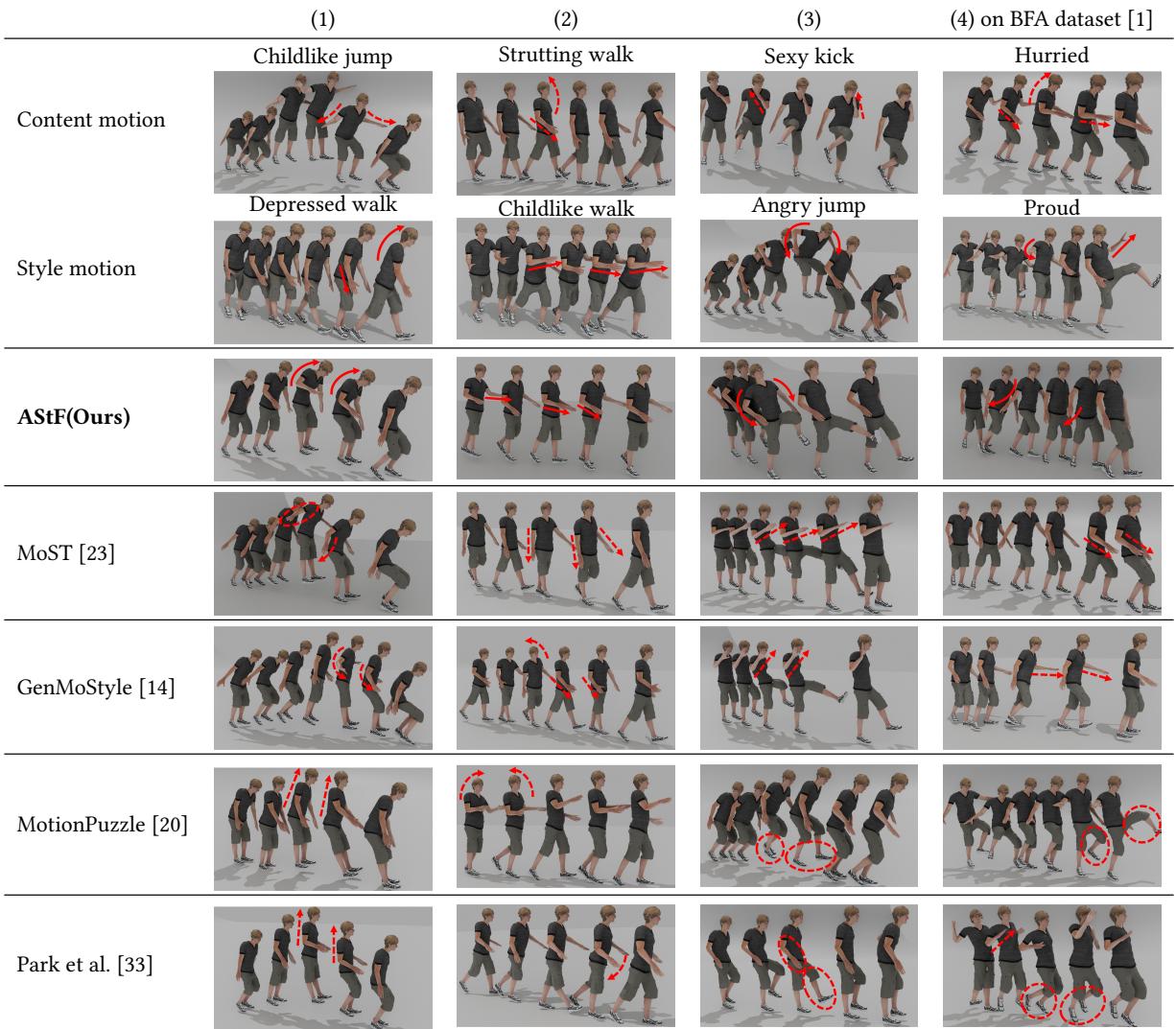
The Impact of Simple SDM. We tested the scenario of using a learnable style token instead of Simple SDM in our Statistics Encoder. The ablation experiments results indicate that using learnable style token significantly reduces style accuracy and results in the highest style FID value. In contrast, our Simple SDM achieves a 22.0% improvement in style accuracy and a 0.046 enhancement in style FID.

The Impact of high-order statistics. Furthermore, as our AStF involves two additional high-order statistics, skewness and kurtosis, we assessed their importance in our SDM and Simple SDM by conducting three ablation experiments: removing skewness, removing kurtosis, and removing the both. The results indicate that using only mean and variance is insufficient. Adding skewness and kurtosis individually provides slight improvements, while incorporating both statistics results in a 7.2% increase in style accuracy and a 4.6% increase in content accuracy compared to without skewness and kurtosis.

The Impact of proposed style-align loss. To test the critical role of our proposed style-align loss function during generator training, we conducted an ablation experiment without it. The ablation experiment results show a significant decrease in style accuracy compared to full model, while content accuracy remained effective and geodesic distance reached the lowest value among all ablation experiments. This observation indicate that the generated motion tends to align more closely with content motion. However, in most cases, style features were insufficiently extracted from the style motion. In contrast, our style-align loss function resulted in a 29.3% improvement in style accuracy.

The Impact of different weights for losses. As mentioned in Section 3.6, MCR loss and style-align loss are novel losses specifically proposed in this work, while others are directly adopted in previous works, we only conducted ablation experiments on the the weight of MCR loss and style-align loss, which are indicated by λ_{MCR} and λ_a . As indicated in Table 2, while the weight of λ_{MCR}

Table 3: Qualitative Results on Xia [42] and BFA [1] datasets. Please refer to the red indications: dashed circles and arrows represent undesired content, while solid lines indicates to the style that needs to be transferred.



increased, the performance of our AStF decreased in both content retention and style fidelity. As for style-align loss, Increasing λ_a results in a decrease in style FID. However, it also causes an increase in content FID, indicating that an increased weight of style-align loss can lead to deficiencies in content retention.

5 Conclusion

In this work, we introduce AStF, a novel framework that effectively transfers style features from style motion to content motion. Our approach proposes using skewness and kurtosis along with mean and variance, to measure style features. We developed the SDM and HOS-Attn for effective extraction and integration of statistics features. Additionally, the MCR discriminator enhances style consistency through feature alignment. Experiments demonstrate our AStF outperforms existing methods in terms of style fidelity and

content retention. However, our method lacks physical constraints on the joints and capturing subtle style features. In future work, we plan to expand our approach to enhance its capability in subtle styles and physical constraints.

Acknowledgments

Our work is supported in part by the Joint Fund of Ministry of Education of China (8091B022149, 8091B02072404), National Natural Science Foundation of China (62132016, 62302372), and Natural Science Basic Research Program of Shaanxi (2020JC-23).

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 (TOG)* 39, 4 (2020), 64–1.

- [2] Nikos Athanasiou, Alpár Cseke, Markos Diomataris, Michael J Black, and Gürhan Varol. 2024. MotionFix: Text-driven 3d human motion editing. In *SIGGRAPH Asia 2024 Conference Papers*. 1–11.
- [3] Jean Basset, Pierre Bénard, and Pascal Barla. 2024. Smear: Stylized motion exaggeration with art-direction. In *ACM SIGGRAPH 2024 Conference Papers*. 1–11.
- [4] Jinhe Bi, Yujun Wang, Haokun Chen, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. 2025. LLaVA Steering: Visual Instruction Tuning with 500x Fewer Parameters through Modality Linear Representation-Steering. In *ACL*.
- [5] Jinhe Bi, Yifan Wang, Danqi Yan, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. 2025. PRISM: Self-Pruning Intrinsic Selection Method for Training-Free Multimodal Data Selection. arXiv:2502.12119 [cs.CV] <https://arxiv.org/abs/2502.12119>
- [6] Jinhe Bi, Danqi Yan, Yifan Wang, Wenke Huang, Haokun Chen, Guancheng Wan, Mang Ye, Xun Xiao, Hinrich Schuetze, Volker Tresp, et al. 2025. CoT-Kinetics: A Theoretical Modeling Assessing LRM Reasoning Process. *arXiv preprint arXiv:2505.13408* (2025).
- [7] Yuxuan Bian, Ailing Zeng, Xuan Ju, Xian Liu, Zhaoyang Zhang, Wei Liu, and Qiang Xu. 2024. MotionCraft: Crafting Whole-Body Motion with Plug-and-Play Multimodal Controls. *arXiv preprint arXiv:2407.21136* (2024).
- [8] Ziyi Chang, Edmund JC Findlay, Haozheng Zhang, and Hubert PH Shum. 2022. Unifying human motion synthesis and style transfer with denoising diffusion probabilistic models. *arXiv preprint arXiv:2212.08526* (2022).
- [9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.
- [10] 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.
- [11] Laurent Dinh, David Krueger, and Yoshua Bengio. 2014. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516* (2014).
- [12] Qihang Fang, Chengcheng Tang, Bugra Tekin, and Yanchao Yang. 2024. CigTime: Corrective Instruction Generation Through Inverse Motion Editing. *Advances in Neural Information Processing Systems* 37 (2024), 102011–102035.
- [13] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2414–2423.
- [14] Chuan Guo, Yuxuan Mu, Xinxin Zuo, Peng Dai, Youliang Yan, Juwei Lu, and Li Cheng. 2024. Generative human motion stylization in latent space. *arXiv preprint arXiv:2401.13505* (2024).
- [15] 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, 4 (2017), 42–49.
- [16] Daniel Holden, Ikhisanul Habibie, Ikuo Kusajima, and Taku Komura. 2017. Fast neural style transfer for motion data. *IEEE computer graphics and applications* 37, 4 (2017), 42–49.
- [17] Daniel Holden, Jun Saito, and Taku Komura. 2016. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (ToG)* 35, 4 (2016), 1–11.
- [18] Lei Hu, Zihao Zhang, Yongjing Ye, Yiwen Xu, and Shihong Xia. 2024. Diffusion-based Human Motion Style Transfer with Semantic Guidance. In *Computer Graphics Forum*, Vol. 43. Wiley Online Library, e15169.
- [19] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*. 1501–1510.
- [20] 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.
- [21] Zhongyu Jiang, Wenhao Chai, Zhuoran Zhou, Cheng-Yen Yang, Hsiang-Wei Huang, and Jenq-Neng Hwang. 2025. PackDiT: Joint Human Motion and Text Generation via Mutual Prompting. *arXiv preprint arXiv:2501.16551* (2025).
- [22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II* 14. Springer, 694–711.
- [23] Boeuk 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.
- [24] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [25] Diederik P Kingma, Max Welling, et al. 2013. Auto-encoding variational bayes.
- [26] Minsu Ko, Eunju Cha, Sungjoo Suh, Huijin Lee, Jae-Joon Han, Jinwoo Shin, and Bohyung Han. 2022. Self-supervised dense consistency regularization for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18301–18310.
- [27] Chuqiao Li, Julian Chibane, Yannan He, Naama Pearl, Andreas Geiger, and Gerard Pons-Moll. 2024. Unimotion: Unifying 3d human motion synthesis and understanding. *arXiv preprint arXiv:2409.15904* (2024).
- [28] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. 2021. Adaatttn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6649–6658.
- [29] Iliana Loi, Evangelia I Zacharakis, and Konstantinos Moustakas. 2023. Machine learning approaches for 3D motion synthesis and musculoskeletal dynamics estimation: a Survey. *IEEE transactions on visualization and computer graphics* 30, 8 (2023), 5810–5829.
- [30] Megan J McAllister, Anthony Chen, and Jessica C Selinger. 2025. Behavioural energetics in human locomotion: how energy use influences how we move. *Journal of Experimental Biology* 228, Suppl_1 (2025).
- [31] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. 2018. Which training methods for GANs do actually converge?. In *International conference on machine learning*. PMLR, 3481–3490.
- [32] Lucas Mourot, Ludovic Hoyet, François Le Clerc, François Schnitzler, and Pierre Hellier. 2022. A survey on deep learning for skeleton-based human animation. In *Computer Graphics Forum*, Vol. 41. Wiley Online Library, 122–157.
- [33] Soomin Park, Deok-Kyeong Jang, and Sung-Hee Lee. 2021. Diverse Motion Stylization for Multiple Style Domains via Spatial-Temporal Graph-Based Generative Model. *Proc. ACM Comput. Graph. Interact. Tech.* 4, 3, Article 36 (Sept. 2021), 17 pages. doi:10.1145/3480145
- [34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
- [35] 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.
- [36] Luke Stark. 2024. Animation and Artificial Intelligence. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1663–1671.
- [37] Tianxin Tao, Xiaohang Zhan, Zhongquan Chen, and Michiel van de Panne. 2022. Style-ERD: Responsive and coherent online motion style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6593–6603.
- [38] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2016. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022* (2016).
- [39] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
- [40] 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).
- [41] Yu-Hui Wen, Zhipeng Yang, Hongbo Fu, Lin Gao, Yanan Sun, and Yong-Jin Liu. 2021. Autoregressive stylized motion synthesis with generative flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13612–13621.
- [42] Shihong Xia, Congyi Wang, Jinxiang Chai, and Jessica Hodgins. 2015. Realtime style transfer for unlabeled heterogeneous human motion. *ACM Transactions on Graphics (TOG)* 34, 4 (2015), 1–10.
- [43] Chenghai Xu, Guangtao Lyu, Jiexi Yan, Muli Yang, and Cheng Deng. 2024. LLM Knows Body Language, Too: Translating Speech Voices into Human Gestures. In *ACL*. 5004–5013.
- [44] Chenghai Xu, Jiexi Yan, and Cheng Deng. 2025. Keep and Extent: Unified Knowledge Embedding for Few-shot Image Generation. *IEEE TIP* (2025).
- [45] Chenghai Xu, Jiexi Yan, Muli Yang, and Cheng Deng. 2024. Rethinking Noise Sampling in Class-Imbalanced Diffusion Models. *IEEE TIP* (2024).
- [46] Han Yang, Kun Su, Yutong Zhang, Jiaben Chen, Kaizhi Qian, Gaowen Liu, and Chuang Gan. 2024. UniMuMo: Unified Text, Music and Motion Generation. *arXiv preprint arXiv:2410.04534* (2024).
- [47] Ling-An Zeng, Guohong Huang, Gaojie Wu, and Wei-Shi Zheng. 2024. Light-T2M: A Lightweight and Fast Model for Text-to-motion Generation. *arXiv preprint arXiv:2412.11193* (2024).
- [48] Jiaxu Zhang, Xin Chen, Gang Yu, and Zhigang Tu. 2024. Generative motion stylization of cross-structure characters within canonical motion space. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 7018–7026.
- [49] Shuhao Zhang, Hui Kang, Yang Liu, Fang Mei, and Hongjuan Li. 2025. HSI: A Holistic Style Injector for Arbitrary Style Transfer. *arXiv preprint arXiv:2502.04369* (2025).
- [50] Lei Zhong, Yiming Xie, Varun Jampani, Deqing Sun, and Huaizu Jiang. 2024. Smoodi: Stylized motion diffusion model. In *European Conference on Computer Vision*. Springer, 405–421.
- [51] Mingrui Zhu, Xiao He, Nannan Wang, Xiaoyu Wang, and Xinbo Gao. 2023. All-to-key attention for arbitrary style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 23109–23119.

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