# A Generalized Two-stage Approach to Motion Style Transfer

Tong Guo [1], Zhiqian Xia [1], Feng Yu [2], Haifeng Xia [1], Siyu Xia [1]
[1] School of Automation, Southeast University, China
[2] Nanjing Xinduan Intelligent Technology Co., Ltd.
{guot, zhi.xia, hfxia, xsy}@seu.edu.cn

## Abstract

*This paper addresses the generalized motion style transfer problem, where the main challenge lies in the inability to apply discrete semantic labels for effective constraints. A cross-dataset style transfer framework is introduced to address key limitations in existing motion synthesis methods. Unlike prior approaches that rely on paired data or discrete semantic labels, our model transfers style through AdaIN-based adaptive instance normalization to encode content motion and style motion separately. A dual-path training strategy that combines style transfer generation and self-reconstruction ensures content preservation. Then, the Dual-Stream Denoising Network (DSDN) integrates LSTM processing for dynamic correction and spatio-temporal autoencoders for static motion refinement, with an Adaptive Gating Fusion Unit (AGFU) that dynamically balances stride amplitude and joint angles. Innovations include cross-dataset transfer without predefined labels, raw-sequence LSTM processing for motion correction, and adaptive fusion architecture outperforming single-method approaches. Validated on the HumanML3D and Mocap datasets, our framework demonstrates significantly enhanced cross-dataset style transfer efficacy.*

## 1. Introduction

Stylistic expression constitutes a fundamental dimension of human movement, serving as a conduit for personality, affective states, and intent. Consequently, stylistic expressiveness is critical in character and avatar animation. Motion style transfer (MST) addresses this need by combining the content semantics of one motion sequence with stylistic characteristics of another. Although recent deep learning methodologies have shown progress in MST within constrained, narrowly defined motion categories (e.g., benchmarks comprising limited action classes), substantial challenges persist, particularly with respect to large-scale, content-rich datasets such as HumanML3D [13]. Such datasets typically lack explicit stylistic annotations and encompass an extensive diversity of motion types. Prevailing MST approaches frequently depend on pre-defined discrete style labels, impose rigid motion constraints, or operate within supervised learning frameworks. These dependencies lead to limitations in model generalization capacity and often yield synthesized motions that exhibit suboptimal kinematic fidelity or undesirable artifacts.

Current approaches to motion style transfer face several interrelated limitations. First, the absence of systematic cross-dataset registration protocols constitutes a fundamental barrier. Existing work lacks methodologies to reconcile heterogeneous skeletal definitions, coordinate systems, and normalization procedures across diverse motion capture datasets, critically impeding the development of generalizable models capable of robust style transfer between disparate repositories. Compounding this, there exists a pronounced dependency on either meticulously paired training examples or discrete semantic style labels, which severely constrains applicability to the vast majority of unstructured, unpaired motion data lacking explicit annotations. This reliance limits flexibility and hinders the capture of nuanced, continuous stylistic variations. Finally, a critical shortfall lies in the insufficient incorporation of biomechanical priors and constraints during the style fusion process. The prevalent emphasis on perceptual or statistical coherence often occurs at the expense of physical plausibility, leading to kinematically distorted movements characterized by violations of anatomical joint limits, dynamic instability, or implausible energy expenditure, thereby undermining the practical utility of synthesized motion for domains requiring biomechanical fidelity.

In this paper, a novel framework for the transfer of styles is proposed. Our framework comprises two core modules: (1) AdaIN-based Style Transfer Module: Leverages adaptive instance normalization to align content motion with style motion, enabling style fusion while preserving motion semantics. A dual path training strategy, combining style transfer generation and self-reconstruction, ensures content consistency and style fidelity. (2) Dual stream denoising network: Deo et al. [7] proposed MI2A, integrating a denoising convolutional autoencoder (DCAE) with an LSTM network to model spatiotemporal wave dynamics, achieving state-of-the-art accuracy in fluid simulations. In computer vision, convolutional LSTM [31] extended traditional LSTMs with convolutional structures, achieving superior performance in spatiotemporal prediction tasks such as precipitation nowcasting. Inspired by [7, 31], the adaptive fusion of dynamic (LSTM-based temporal modeling) and static (spatio-temporal autoencoder) streams is used to address artifacts. The Adaptive Gating Fusion Unit (AGFU) dynamically allocates weights to correct stride amplitude irregularities and joint angle deviations, guided by physics-informed feature extraction.

Our key contributions are as follows.

(1) Unified motion representation: A skeletal alignment pipeline adapts heterogeneous data sets to the HumanML3D standard, ensuring consistent normalization of motion parameters and format compatibility.

(2) Cross-Motion Transfer: Adaptive Instance Normalization (AdaIN) propagates style features while preserving motion content.

(3) Dual stream denoising network (DSDN): A Dynamic Stream (LSTM) and a Static Stream (spatio-temporal autoencoder) collaboratively remove artifacts, while AGFU dynamically allocates correction weights.

Our framework is validated on a cross-dataset benchmark that integrates HumanML3D (14,616 motions) and Mocap [38] (1,500 stylized samples). Experimental results demonstrate that the proposed Cross-motion Transfer method achieves superior performance in style transfer tasks, confirming effective cross-dataset style alignment. Furthermore, the Dual-Stream Denoising Network (DSDN) module, through synergistic optimization by its Dynamic Stream (LSTM [10, 31, 39]) and Static Stream (spatiotemporal autoencoder), reduces the loss of self-reconstruction by 2.27% and 14.95%, respectively, compared to dynamic-only or static-only configurations, achieving precise joint angle restoration and gait phase coherence in post-processing.

## 2. Related Works

### 2.1. Image Style Transfer

In the field of style transfer, numerous deep learning-based methods have been developed and have demonstrated some effectiveness [4, 23, 26, 29, 32, 36]. [32] disentangles the visual style of a source image from its content and transfers the style to the intended image. [36] makes an assumption that global inconsistency is dominated by local inconsistencies and devises a generic contrastive coherence-preserving loss applied to local patches. Within the specific domain of image style transfer, Ulyanov et al. [34] introduce instance normalization (IN), significantly improving the quality of stylized output. Subsequently, Huang [19] advances the approach by incorporating an AdaIN layer. AdaIN enables the injection of distinct style statistics into the same network backbone, thereby facilitating the application of multiple target styles. AdaIN has proven highly effective and has been widely adopted in numerous recent image style transfer works [3, 6, 20, 22, 24, 27].

Recently, [28] proposed regularization by neural style transfer, which integrates a neural style transfer engine with a denoiser to allow magnetic field transfer reconstruction. [25] introduces SwapGAN, a multistage deep generative approach that uses three generators and one discriminator within a unified framework to complete the project from start to finish. Inspired by [25, 28], this paper adopts a two-stage framework to perform the style transfer task: the first stage performs style transfer, followed by a second stage for motion correction.

### 2.2. Motion Style Transfer

Recent advances in motion style transfer have benefitted from machine learning and deep learning techniques [1, 2, 9, 12, 14, 17, 18, 21, 35, 37, 40], yet significant challenges persist in the use of datasets and technical implementation. The early methods [1, 21] focused on decoupling and recombining motion content and style. For instance, Aberman et al.c̃ite1 employed generative adversarial networks to unravel style and content without paired data, while MotionPuzzle [21] enabled localized style control over individual body parts. Gram matrix-based alignment [9, 12, 17, 18] was explored to synchronize stylistic characteristics between movements. To further disentangle content and style, AdaIN-based GAN frameworks [1, 19] were proposed, using adversarial training for style transfer. Subsequent works, such as the [38] local mixture autoregressive, captured complex stylistic relationships, while Holden et al. [17, 18] integrated pre-trained motion manifolds for content supervision and Gram matrices for style representation. MotionPuzzle [21] and Guo et al. [14] extended these ideas through bodypart-specific styling and pre-trained integration of motion models.

However, existing methods predominantly focus on style transfer within narrow motion categories (e.g., five motions in Mocap datasets), leaving content-rich datasets like HumanML3D largely unexplored. This dataset, despite its rich diversity of actions, lacks stylistic annotations, hindering its application in style transfer tasks. To address these gaps, our work introduces a novel framework that operates without paired data, transcends fixed motion category constraints, and rectifies stylized motions to achieve cross-motion style transfer. Then, mocap styles are allowed in HumanML3D, overcoming limitations of previous approaches relying on predefined style labels or limited motion repertoires.

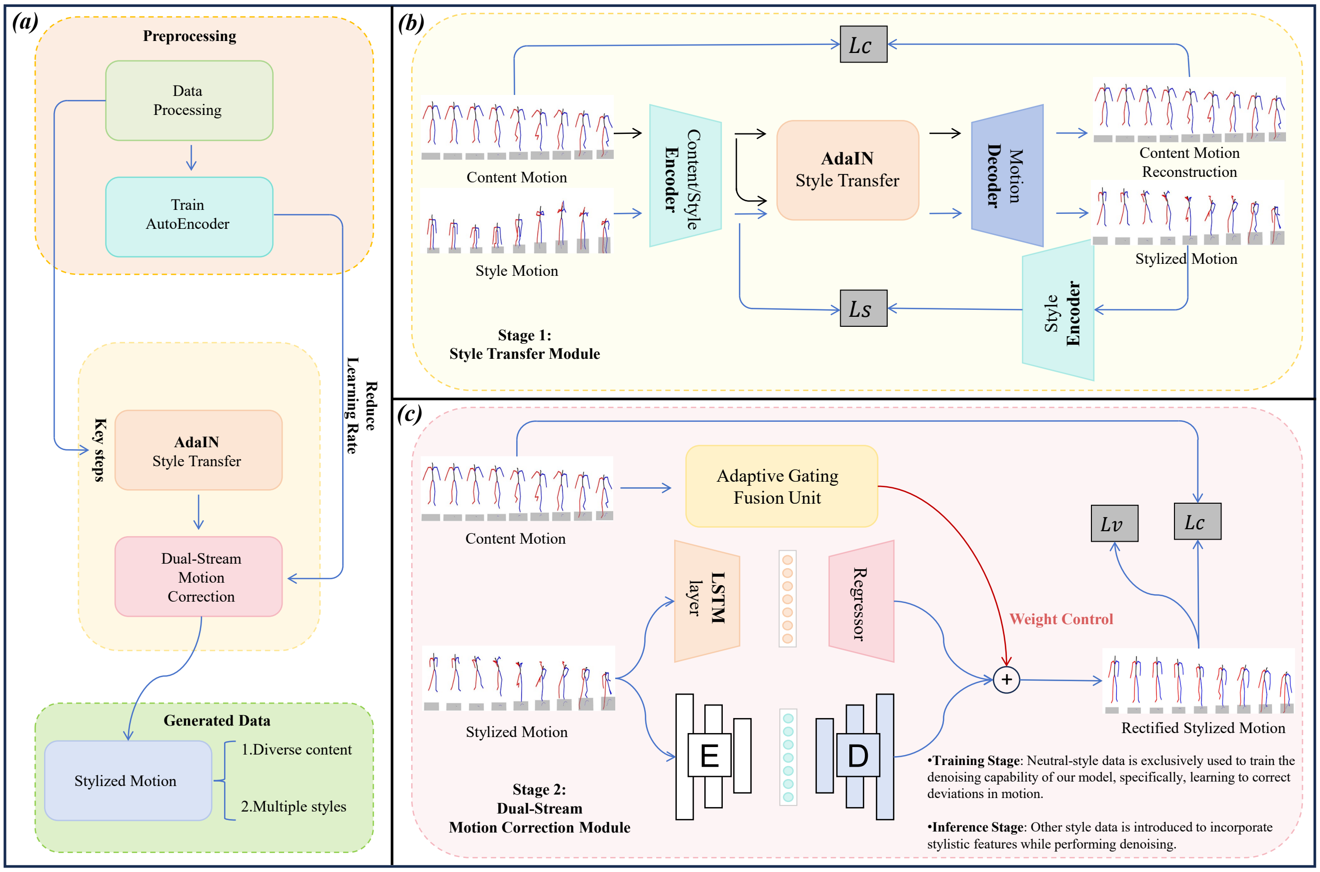



Figure 1. (a) Pipeline: Content and style motions are registered. Stage 1 fuses style features using AdaIN. Stage 2 refines motion via dual-stream denoising. Output preserves content but adopts style. (b) Stage 1: Style Transfer: Feature extractors and AdaIN fuse content and style inputs, producing initial stylized motion. (c) Stage 2: Dual-Stream Correction: The stylized motion is denoised separately in dynamic and static streams.

# 3. Method

## 3.1. Overview

Given a content motion sequence $P_c = (p_1, \ldots, p_T)$ and a style motion sequence $P_s = (p_1, \ldots, p_T)$, our goal is to extract their respective features and generate a novel 3D motion sequence $P_o = (o_1, \ldots, o_T)$ that preserves the semantic content of $P_c$ while adopting the stylistic features of $P_s$.

As shown in Fig.1a, data processing prepares the raw motion sequences. The sanitized data subsequently proceeds through the style transfer module, generating motion imbued with the target stylistic attributes. These style-transferred outputs then undergo refinement via a dedicated motion rectification module. This stage synthesizes the final result, ensuring: (i) Preservation of core kinematic content fidelity, (ii) Accurate transference of stylistic characteristics, and (iii) Correction of positional artifacts inherent in the initial style transfer process.

As represented in Fig.1a, a motion autoencoder trained on the preprocessed dataset – ensures the rectification module's architecture, facilitating motion alignment.

## 3.2. Style Transfer Module

As shown in Fig.1b, our AdaIN Style Transfer Module serves as the core component for transferring motion styles while preserving content semantics. The module operates through a carefully designed dual-path architecture that simultaneously enforces style alignment and content preservation in the latent space.

The content motion $P_c$ is encoded into a content latent representation $C_c$ via a content encoder $E_c$, while the style motion $P_s$ is encoded into a style latent representation $C_s$ using a style encoder $E_s$:

$$C_c = E_c(P_c), \quad C_s = E_c(P_s) \tag{1}$$

To perform style transfer, AdaIN is employed, which aligns the feature statistics of the content motion with those of the style motion. Formally, AdaIN adapts the channel-wise mean $\mu$ and variance $\sigma$ of $C_c$ to match $C_s$:

$$\mathrm{AdaIN}\left(\mathbf{C}_c, \mathbf{C}_s\right) = \sigma\left(\mathbf{C}_s\right)\left(\frac{\mathbf{C}_c - \mu\left(\mathbf{C}_c\right)}{\sigma\left(\mathbf{C}_c\right)}\right) + \mu\left(\mathbf{C}_s\right) \tag{2}$$

This operation ensures that the stylized features inherit the global style characteristics of $C_s$ while retaining the structural content of $C_c$.

#### 3.2.1. Encoder Design.

Inspired by the superior performance of vision transformers [33], our style encoder employs a simplified transformer-based architecture as the style encoder.

Compared to [33], the HumanML3D dataset is adapted as content motion, which lacks discrete semantic labels, rendering the semantic codebook unnecessary. To maintain style transfer effectiveness, the core content encoder structure is kept but the discriminator is omitted due to dataset incompatibility and the absence of discrete labels. Instead, feature-level similarity is enforced using mean and variance statistics losses, inspired by image style transfer techniques [11, 12].

#### 3.2.2. Dual-Path Training Strategy.

The module is optimized through two parallel pathways, each serving a distinct purpose:

Style-Transferred Generation Path: $C_c$ and $C_s$ are processed by AdaIN to create stylized features $\mathbf{C}_{c\to s}$. A motion decoder then transforms $\mathbf{C}_{c\to s}$ into the stylized motion $\mathbf{P}_{c\to s}$. To ensure the generated features match the target style, a style similarity loss ($L_{\text{style}}$) minimizes the difference in mean $\mu$ and standard deviation $\sigma$ between $\mathbf{C}_{c\to s}$ and $C_s$:

$$L_{\text{style}} = \|\mu(\mathbf{C}_{c\to s}) - \mu(\mathbf{C}_s)\|_2^2 + \|\sigma(\mathbf{C}_{c\to s}) - \sigma(\mathbf{C}_s)\|_2^2 \tag{3}$$

Self-Reconstruction Path: This path uses $C_c$ as both the content and style input to AdaIN, producing reconstructed features $\mathbf{C}_{c\to c}$. The motion decoder then generates the reconstructed motion $\mathbf{P}_{c\to c}$. A content similarity loss ($\mathcal{L}_{\text{content}}$) enforces that this reconstructed motion $\mathbf{P}_{c\to c}$ closely matches the original input motion $\mathbf{P}_c$:

$$\mathcal{L}_{\text{content}} = \|\mathbf{P}_{c\to c} - \mathbf{P}_c\|_1 \tag{4}$$

#### 3.2.3. Joint Optimization and Inference.

The total training objective combines both losses to balance style fidelity and content preservation:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{style}}\mathcal{L}_{\text{style}} + \lambda_{\text{content}}\mathcal{L}_{\text{content}} \tag{5}$$

where $\lambda_{\text{style}}$ and $\lambda_{\text{content}}$ are weighting hyperparameters.

During inference, the module takes any content-style pair ($P_c$, $P_s$) and generates a stylized motion $\mathbf{P}_{c\to s}$ blending $P_s$'s style with $P_c$'s content. Dual-path design prevents over-stylization and content degradation. This high-quality style transfer generalizes to varied motion combinations for flexible synthesis.

### 3.3. Motion Correction Module

#### 3.3.1. Problem Analysis and Motivation.

AdaIN-based motion style transfer effectively transfers style but inevitably introduces kinematic distortions due to inherent incompatibilities between some motion content and styles. These distortions include dynamic noise and static noise.

This arises primarily because AdaIN aligns global feature distributions, which can disrupt critical local motion details. Furthermore, the unconstrained style deformation process inherently allows for physically implausible results when combining highly disparate motions and styles.

#### 3.3.2. Architecture.

As shown in Fig.1c, the DSDN corrects motion artifacts by processing data through two specialized branches. The Dynamic Stream uses an LSTM-based [15, 16, 39] gated network to capture temporal variations such as inconsistent stride lengths and changes in movement style. The Static Stream employs a spatio-temporal autoencoder [5, 8, 13] with multi-scale convolutions to compress the temporal dimension and then accurately reconstruct spatial joint poses, focusing on repairing static spatial distortions like inaccurate arm or leg angles. An Adaptive Gating Fusion Unit dynamically combines the outputs of both streams, generating content-aware weights.

#### 3.3.3. Dynamic Stream Processing Module.

To address dynamic trajectory misalignment in motion style transfer, Dynamic Stream module is introduced. This LSTM-based solution directly models temporal dependencies in raw motion sequences using gated recurrent units (GRUs). Unlike static autoencoders that downsample inputs, it preserves full temporal resolution to dynamically correct trajectory offsets throughout the motion timeline.

The module processes raw joint-coordinate sequences ($x \in \mathbb{R}^{B\times T\times 263}$) through a single-layer LSTM with 128 hidden dimension to capture contextual features across time. These features are then mapped to dynamic corrections $\mathbf{x}_{\text{dyn}}$ via a lightweight two-layer regressor. By operating directly on raw sequences, the LSTM adapts to gait cycles in real-time—such as detecting sudden phase shifts during running motions—and adjusts trajectories dynamically. This contrasts with CNN-based autoencoders, whose fixed receptive fields fail to model transient temporal logic like stride-length variations. Regressor:

$$\mathbf{x}_{\text{dyn}} = W_2 \cdot \tanh(W_1 \cdot h_t + b_1) + b_2 \tag{6}$$

where $h_t$ denote the output of LSTM.

#### 3.3.4. Static Stream Processing Module.

The static stream employs a spatio-temporal autoencoder to address spatial pose distortions through a symmetric encoder-decoder architecture. The encoder $\mathcal{E}$ compresses the input pose sequence $\mathcal{P} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_{T'})$ into a motion snippet code sequence $\mathcal{C}_s = (\boldsymbol{c}_s^1, \ldots, \boldsymbol{c}_s^T)$ via two 1D convolutional layers with kernel size 4 and stride 2, reducing the temporal dimension($T' = T/4$) while expanding channels to 512. The decoder $\mathcal{D}$ symmetrically reconstructs the original sequence $\hat{\mathcal{P}}$ using transposed convolutions.

$$C_s = E(P), \quad \hat{P} = D(C_s) \tag{7}$$

#### 3.3.5. Adaptive Gating Fusion Unit.

An Adaptive Gating Fusion Unit (AGFU) is proposed to dynamically adjust the fusion ratio between dynamic stream denoising and static stream denoising modules. The proposed module adaptively allocates different gating weights for each time step based on the semantic analysis of input content actions, enabling the optimal combination of the two denoising strategies.

As shown in that combines style transfer2.3, the adaptive gating unit consists of two main components: a feature extractor and a gating decision layer. The feature extractor, composed of two linear transformations and a Tanh activation function, extracts key feature representations from the input content motion sequence. The gating decision layer, consisting of a linear transformation and a Softplus activation function, outputs the gating weights $\alpha$.

Specifically, given an input sequence $\mathbf{x} \in \mathbb{R}^{B \times T \times D}$ (where $B$ is batch size, $T$ is time steps, and $D$ is feature dimension), the input is first flattened into a 2D tensor $\mathbf{x}_{flat} \in \mathbb{R}^{B \times T \times D}$, then processed through the feature extractor:

$$h = \tanh(W_1 x_{flat} + b_1) \tag{8}$$

$$\text{gate\_logits} = W_2 h + b_2 \tag{9}$$

where $W_1 \in \mathbb{R}^{64 \times D}$ and $W_2 \in \mathbb{R}^{1 \times 32}$ are weight matrices of the two linear layers, and $b_1 \in \mathbb{R}^{64}$, $b_2 \in \mathbb{R}^1$ are corresponding bias terms. The Softplus activation function is applied to the gate logits to obtain the final gating weights:

$$\alpha = \sigma(\text{gate\_logits}) \tag{10}$$

where $\sigma$ is the Softplus function defined as $\sigma(x) = \ln(1 + e^x)$, ensuring positive and smooth output values to avoid gradient vanishing issues.

The resulting gating weights $\alpha \in \mathbb{R}^{B \times T \times 1}$ represent the proportion of contribution from the dynamic stream denoising module for each time step, while the remaining portion $(1 - \alpha)$ is allocated to the static stream denoising module. This adaptive mechanism enables the model to intelligently adjust the contribution of both denoising strategies based on the characteristics of input content, improving overall denoising performance.

#### 3.3.6. Loss Function.

Our training process incorporates two key loss terms: content reconstruction loss and temporal smoothness loss. The content reconstruction loss $\mathcal{L}_{\text{rec}}$ enforces consistency between the generated motion and the original input, defined as:

$$\mathcal{L}_{\text{rec}} = \|\mathbf{P}_o - \mathbf{P}_c\|_1 \tag{11}$$

where $\mathbf{P}_o$ denotes the motions after neutral style transfer, and $\mathbf{P}_c$ represents the original input content motion. This loss ensures the preservation of motion structure during style transfer.

Furthermore, temporal jitter is observed in generated motions. Inspired by [30], a temporal smoothness loss $\mathcal{L}_{\text{vel}}$ is introduced to suppress abrupt frame transitions:

$$\mathcal{L}_{\text{vel}} = \|\mathbf{P}_{\text{fused}}[:, 1:] - \mathbf{P}_{\text{fused}}[:, :-1]\|_1 \tag{12}$$

where $\mathbf{P}_{\text{fused}}$ denotes the fused motion sequence, and $[:, 1:], [:, :-1]$ denote operations for shifting the sequence forward or backward. This loss minimizes inter-frame discrepancies, enhancing temporal coherence.

## 4. Experiment

### 4.1. Dataset

A cross-dataset style transfer benchmark is established by integrating HumanML3D as the content dataset and Mocap as the style dataset.

HumanML3D contains 14,616 motion clips (28.59 hours total). Preprocessing included resampling at 20 FPS, truncating motions exceeding 10 seconds, skeleton retargeting to a default template, Z-axis alignment, and mirror augmentation. Text annotations were crowd-sourced. The Mocap dataset provides 1,500 stylized motion samples (30 FPS) spanning 5 actions and 8 distinct styles. Processing involved segmenting motions into 32-frame clips (with 8-frame overlap), root translation normalization, Z-axis rotation alignment, and standardizing all clips to 10 seconds.

To harmonize heterogeneous data formats, Mocap joints are mapped to the HumanML3D skeleton template, normalized motions through root translation and Z-axis rotation, temporally standardized clips to 10 seconds via truncation, and stored bone positions and rotations in unified files.

The harmonized dataset will be released on a public platform. It represents the largest cross-motion style transfer benchmark to date, supporting evaluation across 43 motion classes and 8 fine-grained styles.

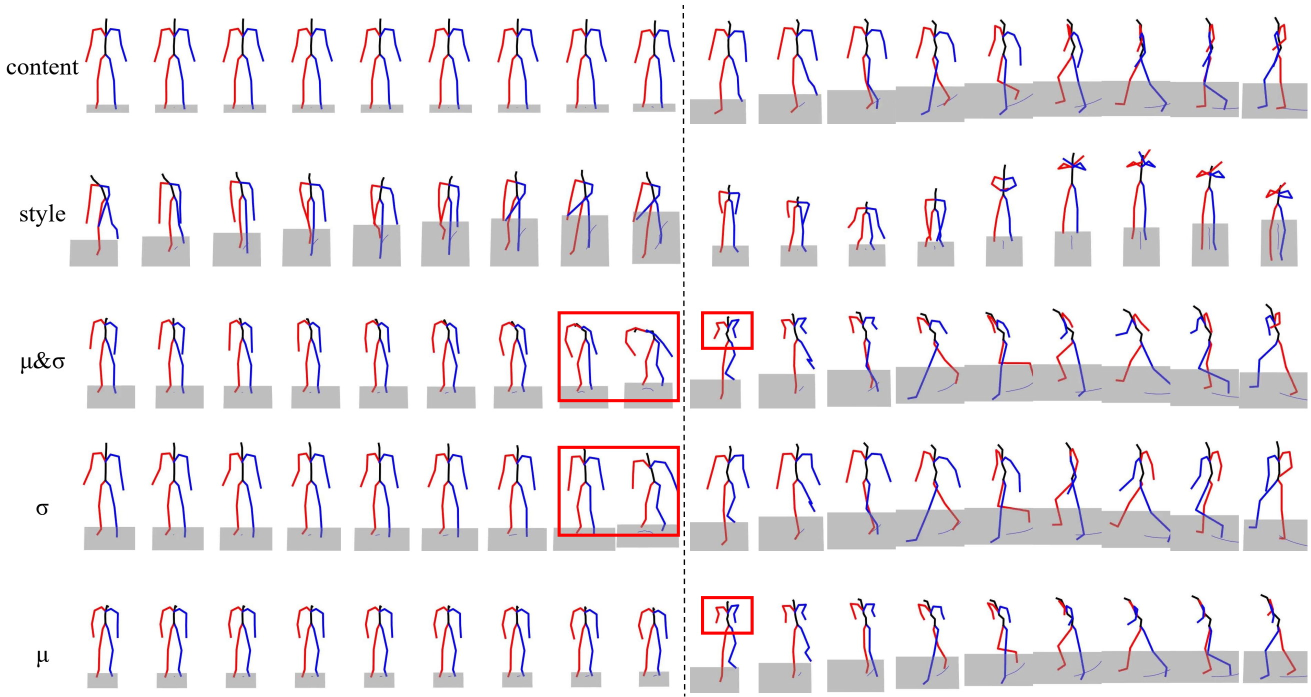



Figure 2. Dynamic Noise and Static Noise Visualization. Here is a 5×2 comparative visualization of motion transfer outcomes: Row 1: Input content motion Row 2: Input style motion Row 3: Complete transfer results Row 4: Variance-only transfer Row 5: Mean-only transfer. Left part, the final two frames of the transfer result in Row 3 exhibit shaking. This artifact mirrors the posture seen in the corresponding frames of Row 4 and identifies Dynamic Noise. Right part, the transfer result in Row 3 displays distortion in the arm movement, deviating from the original content motion in Row 1. This distortion closely resembles the noise pattern present in Row 5, signifying Static Noise.

#### 4.1.1. Data Choise

Content features are derived from rotational variables following [13], which effectively captures global motion dynamics through root joint velocity and foot-ground contact patterns. This prevents static pose stacking artifacts while maintaining natural foot trajectories. For style representation, skeletal joint positions utilize ($T \times 22 \times 3$ tensor) as in [33], where the spatial configuration of 22 joints provides discriminative style cues across different motion categories.

### 4.2. Style Transfer Module

#### 4.2.1. Loss Function

As shown in Table 1, the loss trends indicate stable optimization with clear content-style trade-offs. The content loss ($\mathcal{L}_{\text{content}}$) decreases from 0.29 (10 epochs) to 0.22 (135 epochs), demonstrating improved content reconstruction accuracy. Notably, the training-validation gap remains small (0.22 v.s. 0.24), suggesting generalization robustness. The style loss ($L_{\text{style}}$) exhibits sharper decay (0.14 $\rightarrow$ 0.008), indicating rapid style constraint satisfaction. The validation style loss slightly exceeds training values (0.009 v.s. 0.008), implying minor distributional divergence.

#### 4.2.2. Result

Style transfer can introduce dynamic and static noise (Fig. 2). Dynamic noise causes visible jitter ("Transfer Variance Only 1"), propagating to the output ("Transfer Mean and Variance 1"). Static noise creates spatial inconsistencies, like abnormal arm articulation ("Transfer Mean Only 2"), preserved in the combined result ("Transfer Mean and Variance 2"). Section 4.2 quantitatively evaluates noise suppression using our Motion Correction Module.

#### 4.2.3. Comparative Analysis with State-of-the-Art

A rigorous comparative evaluation was conducted against the Fine-style benchmark [33], revealing fundamental limitations in cross-modal style transfer methodologies.

**Adversarial Discriminator Failure in Heterogeneous Transfers** Fine-style's domain discriminator achieves high style transfer accuracy (55.63%)[33] when evaluating mocap-to-mocap style consistency. However, it catastrophically fails to recognize stylistic attributes (0.05%) in transfers between HumanML3D (content) and mocap (style) domains. This discrimination collapse indicates that Fine-

Table 1. Training and Validation Loss Comparisons of Style Transfer Modules. $*$ indicates lower than Fine-Style method. **Bold** indicates the best performing metric per column

| Method | Phase | $\mathcal{L}_{\text{content}}$ | $\mathcal{L}_{\text{style}}$ |
|---|---|---|---|
| Proposed | Training | $\mathbf{0.22 \pm 0.02}^{*}$ | $\mathbf{0.008 \pm 0.001}$ |
| | Validation | $\mathbf{0.24 \pm 0.03}^{*}$ | $\mathbf{0.009 \pm 0.001}$ |
| Fine-Style [33] | Training | $0.25 \pm 0.04$ | / |
| | Validation | $0.27 \pm 0.04$ | / |

style embeddings reside outside the target style manifold when handling cross dataset style transfer.

**Content Reconstruction Deficiencies** Quantitative analysis reveals Fine-style exhibits 13.6% higher mean content reconstruction error (0.25 ± 0.025 vs. 0.22 ± 0.020. Visual inspection confirms this manifests as: articulation errors in upper-limb kinematic chains, temporal inconsistencies during periodic motions, etc.

**Non-physical Motion Artifacts** Significant root displacement artifacts were observed in Fine-style outputs (Fig. 3), violating biomechanical constraints. As shown, it can be observed that the content reconstruction preserves the underlying motion effectively. In contrast, the style transfer results exhibit significant positional drift over time. This issue is clearly visible in the final frame of the first row (highlighted by the red box), where the blue trajectory curve indicates an anomalous shift during the motion sequence.

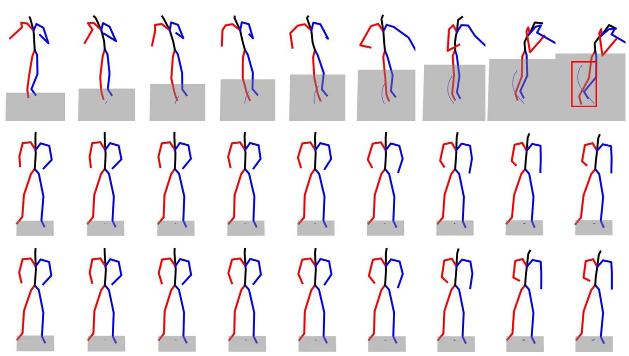


Figure 3. Row 1: Style transfer results. Row 2: Content reconstruction results. Row 3: Original motion sequences. Note the overall positional drift in the style transfer output (Row 1), highlighted by the red box in the last frame, which captures an anomalous trajectory (blue). The content reconstruction maintains motion consistency.

## 4.3. Motion Correction Module

### 4.3.1. Training and Inference Phase Explanation.

As mentioned above, the style transfer framework does not employ a discriminator. Instead, it relies on a loss of content similarity at the data level and a loss of style similarity at the feature level. However, this approach may lead to unintended motion artifacts or noise in the synthesized movements. To address this challenge, a novel DSDN is proposed that synergizes dynamic and static processing with adaptive fusion, effectively eliminating motion shifts induced during style transfer while preserving fidelity.

During training, two key datasets are utilized: original unstyled motion data (content), and a neutral-styled version of this content, interpreted as noisy content data. Data analysis reveals that neutral-style noise exhibits systematic similarities to noise generated by other stylization methods. The core objective is to train the model to decompose and remove this noise component exclusively from the neutral-styled data. Through this focused training using a single, representative noise source, the model learns a generalized noise decomposition strategy applicable to diverse styles.

At inference, the original unstyled content data (for extracting core content features) and the same content stylized with seven other target styles. A denoise process occurs.

### 4.3.2. Loss Function.

Table2 presents the training and validation losses of the motion correction module. The four loss components quantify distinct corrective capabilities: (1) $L_{rec}^{D}$ evaluates the temporal consistency of dynamic flow rectification, (2) $L_{rec}^{S}$ measures spatial regularization performance for static flow correction, (3) $L_{rec}^{F}$ reflects the cross-modal synergy between static and dynamic flow denoising, while (4) $L_{vel}^{F}$ enforces temporal continuity to suppress motion artifacts. The model is trained for 800 epochs using the Adam optimizer (learning rate 0.001).

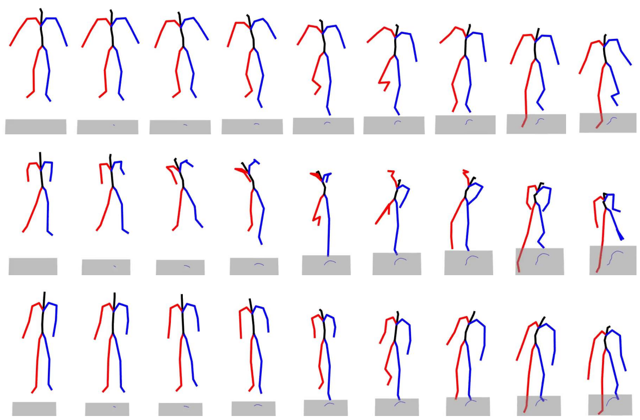
Quantitative analysis: Superiority of dual-stream denoising: $L_{rec}^{F}$ (0.3236 ± 0.01) demonstrates lower error than both single-stream counterparts ($L_{rec}^{D}$: 0.3311, $L_{rec}^{S}$: 0.3805), confirming that joint optimization of static and dynamic flows yields synergistic benefits.

This result substantiates the effectiveness of our dual-

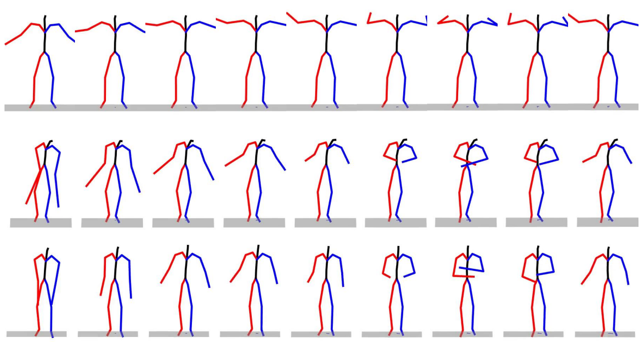
Table 2. Training and Validation Loss of Motion Correction Module

| Phase | $L_{rec}^{D}$ | $L_{rec}^{S}$ | $L_{rec}^{F}$ | $L_{vel}^{F}$ |
|---|---|---|---|---|
| Training Set | $0.3311 \pm 0.01$ | $0.3805 \pm 0.01$ | $0.3236 \pm 0.01$ | $0.0996 \pm 0.00013$ |
| Validation Set | $0.4221 \pm 0.02$ | $0.3895 \pm 0.02$ | $0.3760 \pm 0.02$ | $0.0986 \pm 0.00011$ |

stream architecture in disentangling motion components while maintaining temporal coherence.



(a) Walking Downstairs Sequence



(b) Waving Motion Sequence

Figure 4. Results: original motion sequence (Row 1), style transfer sequence (Row 2), and motion correction sequence (Row 3). (a) The initial style transfer results exhibit noticeable deformations in the human body and typographical elements. After applying the correction algorithm, the posture becomes anatomically plausible while the style characteristics are preserved. (b) Unnatural elongation of the subject's arm is observable across the first few frames. Following correction, the figure is restored to a physically coherent configuration.

#### 4.3.3. Result.

Key results are visualized in Fig.4a and Fig.4b, demonstrating human motion style transformations.

Fig.4a shows walking downstairs sequence. The top row depicts the original walking-downstairs motion. The middle row shows the result after applying style transfer. While the depressed characteristic (e.g., slumped posture) is evident, the figure's arms exhibit significant distortion. The bottom row displays the outcome after motion correction. This step successfully eliminates the distorted arm artifacts while effectively preserving the core features of the depressed style, particularly the bent posture and lethargic movement characteristic in the latter part of the sequence.

Fig.4b shows waving motion sequence. The top row illustrates the original waving motion. The middle row presents the Depressed style transferred version. Although the depressed mood is conveyed, the upper body now shows noticeable tilting. The bottom row features the result following motion correction. The corrected motion exhibits a naturalized body posture. Crucially, key indicators of the depressed style – specifically, the constrained and melancholic nature of the hand gestures – remain distinctly preserved.

## 5. Conclusion

This paper proposes a novel two-stage 3D motion style transfer framework with three key innovations: 1) Unified Motion Representation: A skeletal alignment pipeline standardizes datasets into HumanML3D, enabling label-free cross-dataset style transfer. 2) AdaIN-based Style Transfer: A module using adaptive instance normalization disentangles motion content from style, enabling transfer while preserving anatomical fidelity. Dual-path training ensures effective propagation. 3) Dual-Stream Denoising Network: Combining dynamic LSTM and static autoencoding, guided by AGFU, effectively corrects biomechanical artifacts. Experiments on HumanML3D and Mocap datasets confirm the framework's enhanced cross-dataset transfer capability. The AdaIN approach successfully disentangles style semantics and the DSDN outperforms single-method denoising in artifact correction. Future research will address data distribution differences and real-time editing applications.

## 6. Acknowledgements

This work was supported by Key Laboratory of Intelligent Processing Technology for Digital Music (Zhejiang Conservatory of Music), Ministry of Culture and Tourism under Grant 2023DMKLB003, the Big Data Computing Center of Southeast University, and Research Fund for Advanced Ocean Institute of Southeast University, Nantong (No. GP20010002).

## References


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

[2] Andreas Aristidou, Qiong Zeng, Efstathios Stavrakis, Kangkang Yin, Daniel Cohen-Or, Yiorgos Chrysanthou, and Baoquan Chen. Emotion control of unstructured dance movements. In *Proceedings of the ACM SIGGRAPH / Eurographics Symposium on Computer Animation*, pages 9:1–9:10, New York, NY, USA, 2017. Association for Computing Machinery. 2

[3] Kyungjune Baek, Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Hyunjung Shim. Rethinking the truly unsupervised image-to-image translation, 2021. 2

[4] Seth Grace Banaga and Carol O'Sullivan. Stroke rehabilitation through proxy agency and embodiment in mixed reality. In *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 964–967, 2022. 2

[5] Manon P. Bart, Sita Dawanse, Nicholas J. Savino, Viet Tran, Tianhong Wang, Sanjaya Lohani, Farris Nefissi, Pascal Bassène, Moussa N'Gom, and Ryan T. Glasser. Classification of single photons in higher-order spatial modes via convolutional neural networks, 2024. 4

[6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains, 2020. 2

[7] Indu Kant Deo and Rajeev K. Jaiman. Predicting wave dynamics using deep learning with multistep integration inspired attention and physics-based loss decomposition, 2025. 1, 2

[8] Saaketh Desai, Sadhvikas Addamane, Jeffrey Y. Tsao, Igal Brener, Laura P. Swiler, Remi Dingreville, and Prasad P. Iyer. Autoscilab: A self-driving laboratory for interpretable scientific discovery, 2024. 4

[9] Han Du, Erik Herrmann, Janis Sprenger, Noshaba Cheema, somayeh hosseini, Klaus Fischer, and Philipp Slusallek. Stylistic Locomotion Modeling with Conditional Variational Autoencoder. In *Eurographics 2019 - Short Papers*. The Eurographics Association, 2019. 2

[10] Samuel J. Edwards and Michael D. Levine. Data-driven extreme response estimation, 2025. 2

[11] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style, 2015. 4

[12] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016. 2, 4

[13] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5142–5151, 2022. 1, 4, 6

[14] Chuan Guo, Yuxuan Mu, Xinxin Zuo, Peng Dai, Youliang Yan, Juwei Lu, and Li Cheng. Generative human motion stylization in latent space, 2024. 2

[15] Rahul Halder, Giovanni Stabile, and Gianluigi Rozza. Physics informed neural network framework for unsteady discretized reduced order system, 2024. 4

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 4

[17] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Trans. Graph.*, 35(4), 2016. 2

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

[19] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization, 2017. 2

[20] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation, 2018. 2

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

[22] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019. 2

[23] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiri Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks, 2018. 2

[24] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation, 2019. 2

[25] Yu Liu, Wei Chen, Li Liu, and Michael S. Lew. Swapgan: A multistage generative approach for person-to-person fashion style transfer. *IEEE Transactions on Multimedia*, 21(9): 2209–2222, 2019. 2

[26] Tanya Marwah, Gaurav Mittal, and Vineeth N. Balasubramanian. Attentive semantic video generation using captions, 2017. 2

[27] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization, 2019. 2

[28] Guoyao Shen, Yancheng Zhu, Mengyu Li, Ryan McNaughton, Hernan Jara, Sean B. Andersson, Chad W. Farris, Stephan Anderson, and Xin Zhang. Regularization by neural style transfer for mri field-transfer reconstruction with limited data, 2025. 2

[29] Junxiao Shen, John Dudley, and Per Ola Kristensson. Simulating realistic human motion trajectories of mid-air gesture typing. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 393–402, 2021. 2

[30] Kang Shen, Haifeng Xia, Guangxing Geng, Guangyue Geng, Siyu Xia, and Zhengming Ding. Deitalk: Speech-driven 3d facial animation with dynamic emotional intensity modeling. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 10506–10514, New York, NY, USA, 2024. Association for Computing Machinery. 5

[31] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai kin Wong, and Wang chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting, 2015. 2

[32] Yang Shi, Pei Liu, Siji Chen, Mengdi Sun, and Nan Cao. Supporting expressive and faithful pictorial visualization design with visual style transfer. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):236–246, 2023. 2

[33] Wenfeng Song, Xingliang Jin, Shuai Li, Chenglizhao Chen, Aimin Hao, and Xia Hou. Finestyle: Semantic-aware fine-grained motion style transfer with dual interactive-flow fusion. *IEEE Transactions on Visualization and Computer Graphics*, 29(11):4361–4371, 2023. 4, 6, 7

[34] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization, 2017. 2

[35] Matthias Wright and Björn Ommer. Artfid: Quantitative evaluation of neural style transfer, 2022. 2

[36] Zijie Wu, Zhen Zhu, Junping Du, and Xiang Bai. Ccpl: Contrastive coherence preserving loss for versatile style transfer, 2022. 2

[37] Shihong Xia. Modeling style and variation in human motion. In *2010 4th International Universal Communication Symposium*, pages 207–207, 2010. 2

[38] Shihong Xia, Congyi Wang, Jinxiang Chai, and Jessica Hodgins. Realtime style transfer for unlabeled heterogeneous human motion. *ACM Trans. Graph.*, 34(4), 2015. 2

[39] Changzhi Yu, Shuangping Han, Kai Song, and Liantuan Xiao. Spatiotemporal deep learning network for photon-level block compressed sensing imaging, 2025. 2, 4

[40] M. Ersin Yumer and Niloy J. Mitra. Spectral style transfer for human motion between independent actions. *ACM Trans. Graph.*, 35(4), 2016. 2