Decoupling Contact for Fine-Grained Motion Style Transfer

XIANGJUN TANG, State Key Lab of CAD&CG, Zhejiang University, China
LINJUN WU, State Key Lab of CAD&CG, Zhejiang University, China
HE WANG, Department of Computer Science and UCL Centre for Artificial Intelligence, University College London, United Kingdom
YIQIAN WU, State Key Lab of CAD&CG, Zhejiang University, China
BO HU, Tencent Technology Co., Ltd., China
SONGNAN LI, Tencent Technology Co., Ltd., China
XU GONG, Tencent Technology Co., Ltd., China
YUCHEN LIAO, Tencent Technology Co., Ltd., China
QILONG KOU, Tencent Technology Co., Ltd., China
XIAOGANG JIN*, State Key Lab of CAD&CG, Zhejiang University; ZJU-Tencent Game and Intelligent Graphics Innovation Technology Joint Lab, China

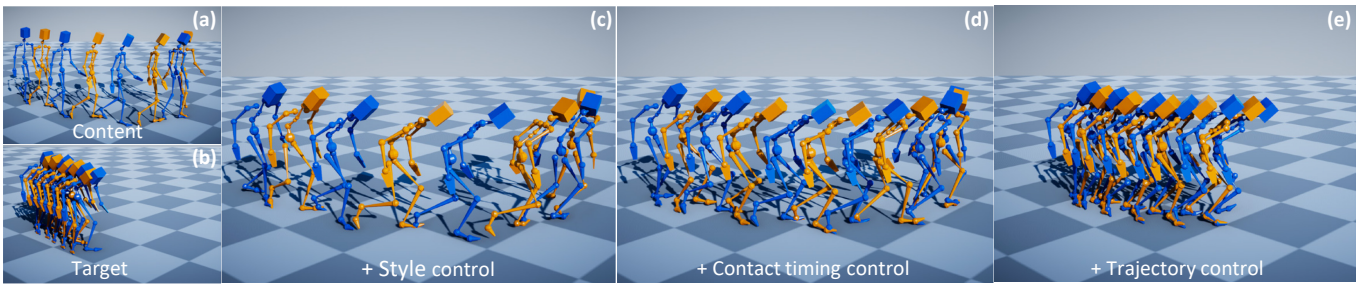

Fig. 1. Our method can independently control style, contact timing, and trajectory, allowing for fine-grained motion style transfer. Given a content motion (a) and an "old man" style (bending, fast pace, and slow speed) target motion (b), our approach allows for the gradual addition of "style" (c), "contact timing" (d), and "trajectory" (e) of the target motion to the content, which previous methods could not achieve. The result in (c) depicts the target motion's bending pose; the result in (d) depicts more frequent contact within the same time duration, indicating a faster pace of the character; and the result in (e) depicts a slower speed, precisely replicating the entire "old man" style. We show every frame in which either foot makes contact with the ground.

Motion style transfer changes the style of a motion while retaining its content and is useful in computer animations and games. Contact is an essential component of motion style transfer that should be controlled explicitly in order to express the style vividly while enhancing motion naturalness and quality. However, it is unknown how to decouple and control contact to achieve fine-grained control in motion style transfer.

In this paper, we present a novel style transfer method for fine-grained control over contacts while achieving both motion naturalness and spatial-temporal variations of style. Based on our empirical evidence, we propose controlling contact indirectly through the hip velocity, which can be further decomposed into the trajectory and contact timing, respectively. To this end, we propose a new model that explicitly models the correlations between motions and trajectory/contact timing/style, allowing us to decouple and control each separately. Our approach is built around a motion manifold, where hip controls can be easily integrated into a Transformer-based decoder. It is versatile in that it can generate motions directly as well as be used as post-processing for existing methods to improve quality and contact controllability. In addition, we propose a new metric that measures a correlation pattern of motions based on our empirical evidence, aligning well with human perception in terms of motion naturalness. Based on extensive evaluation, our method outperforms existing methods in terms of style expressivity and motion quality.

CCS Concepts: • **Computing methodologies** → **Motion capture**; *Motion Transfer*; Neural networks; Motion manifold.

Additional Key Words and Phrases: Style transfer, Motion quality, Editing.

ACM Reference Format:
Xiangjun Tang, Linjun Wu, He Wang, Yiqian Wu, Bo Hu, Songnan Li, Xu Gong, Yuchen Liao, Qilong Kou, and Xiaogang Jin. 2024. Decoupling Contact for Fine-Grained Motion Style Transfer. In *SIGGRAPH Asia 2024 Conference Papers (SA Conference Papers '24), December 3–6, 2024, Tokyo, Japan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3680528.3687609

*Corresponding author

SA Conference Papers '24, December 3–6, 2024, Tokyo, Japan

ACM ISBN 979-8-4007-1131-2/24/12…$15.00
https://doi.org/10.1145/3680528.3687609

1 INTRODUCTION

Motion style transfer has important applications in computer animation and gaming by reducing laborious and costly motion capture and empowering artists in creation. Typically, motion style transfer is accomplished by separating the style from the content so that the style can be transferred to a different content [Aberman et al. 2020; Holden et al. 2016; Jang et al. 2022; Park et al. 2021; Yumer and Mitra 2016]. However, in character motions, it is difficult to distinguish between content and style, which can sometimes lead to ambiguity [Song et al. 2023].

Given such an ambiguity, existing work interprets content and styles differently [Amaya et al. 1996; Jang et al. 2022; Song et al. 2023; Yumer and Mitra 2016], which can be broadly divided into supervised styles and unsupervised styles. Supervised style uses human-labeled actions and treats style as spatial-temporal variations [Ribet et al. 2019] (amplitude, speed, and pose, for example) of an action. Walking motions, for example, may include various stepping strategies (stride, stroll, tramp, march) with varying strides, stepping frequency, knee height during stepping, and so on, all of which are treated as style. Unsupervised style relies on motion similarity rather than human labeling. Clustered motions based on similarity (for example, in the data space or some latent space) can be viewed as variations on the same content [Jang et al. 2022]. Despite the cutting-edge performance of deep learning under both strategies [Aberman et al. 2020; Jang et al. 2022; Song et al. 2023], it is safe to say that a complete separation of motion style and content is difficult to achieve.

If a complete decoupling is difficult, we ask a question: what information cannot be decoupled completely and is it important in motion style transfer? We find contact is such an element and it is crucial. In locomotion, contact is multi-faceted including timing, duration, frequency, pattern, location, etc. It is tightly coupled with both content and style. Although most existing work treats contact as part of the content [Aberman et al. 2020; Jang et al. 2022; Unuma et al. 1995; Yumer and Mitra 2016], it has recently been recognized as being closely related to style as well [Aberman et al. 2020; Dong et al. 2020]. However, it is unknown how to control contact to achieve fine-grained control in motion style transfer. As will be demonstrated later, not explicitly controlling contact frequently results in limited expressiveness of certain styles or even compromising the content itself, e.g. unnatural motions.

To this end, we present a novel style transfer method based on fine-grained contact control that achieves both motion naturalness and spatial-temporal variations of style. Given a source motion containing the action's main content and another motion containing the desired style, our approach allows for the interpolation of the multi-faceted contact information between them, resulting in controllable and smooth transitions. However, because contact is so closely linked to both content and style, naive direct contact control results in unnatural and uncontrollable motions. We propose to control contact indirectly via the hip velocity, which can be further decomposed into high-level and low-level features based on empirical statistical evidence from the data. We discovered that high-level features can largely determine the root trajectory and thus the spatial aspect of contact, such as locations, whereas low-level features can govern the timing. They work together to provide fine-grained control during style transfer.

To achieve the above, we propose a new model to learn a motion manifold where the correlations between motions and other factors (trajectory, contact timing and style) are explicitly modeled. Specifically, we first employ different neural networks to encode the hip trajectory, contact timing of the content motion, and style features of the style motion separately. We then fuse them to control the trajectory and contact timing, as well as a latent variable that captures the style features. Following that, we obtain a controllable motion manifold into which hip controls can be easily added for final motion synthesis using a Transformer-based decoder. The core of our approach is the motion manifold, which, as demonstrated later, can enhance motion quality and contact controllability. It is versatile in that it can directly generate motions, and can also be used as post-processing for existing methods [Aberman et al. 2020; Jang et al. 2022]. In addition, we propose *Contact Precision-Recall*, which measures the match between the synthesized contact and hip velocity based on our empirical observation of their high correlation. This metric aligns better with human perception in measuring motion quality than Fréchet Motion Distance (FMD) and foot skating metrics. We demonstrate that our method outperforms existing methods in terms of style expressivity and motion quality through extensive validation.

The contributions of our work can be summarized as follows:

- A novel method for fine-grained control of motion style transfer, which improves the expressiveness and naturalness of motion style transfer.
- A new transformer-based model for motion manifold based on empirical evidence that allows us to control contacts by hip velocity, resulting in more natural and refined motions.
- Our manifold can be combined with existing motion transfer methods to improve motion quality and controllability.

2 RELATED WORK

2.1 Controllable Motion Generation

Our method allows for contact control via hip velocity, creating a continuous space for contact editing. A related field is controllable motion generation, which can be formulated as motion planning problems [Beaudoin et al. 2008; Levine et al. 2012; Safonova and Hodgins 2007; Wang et al. 2015, 2013]. However, motion planning problems require complex optimization [Chai and Hodgins 2007] and frequently lead to slow computation. By searching in structured data, data-driven methods [Arikan and Forsyth 2002; Kim et al. 2023; Kovar et al. 2002; Min and Chai 2012; Shen et al. 2017] can avoid slow optimizations but require unaffordable memory space to cover diverse control situations. Deep neural networks can leverage compressed data representations [Holden et al. 2020]. One method for incorporating controllability is to include constraints as regularization in the loss function [Chiu et al. 2019; Martinez et al. 2017; Wang et al. 2019]. When different contacts are required, however, simply adding constraints will not yield high-quality results. Learning the conditional probability via a generative model, such as VAE [Chen et al. 2020; Ling et al. 2020; Petrovich et al. 2021, 2022; Tang et al.

2022; Zhang et al. 2023a], GAN [Ahn et al. 2018], flows [Alexanderson et al. 2020], and diffusion models [Alexanderson et al. 2023; Ao et al. 2023; Chen et al. 2023; Ghorbani et al. 2023; Tevet et al. 2022], can produce controlled natural motions. However, controlling contact or interpolating contact between two motions without affecting the style hasn't been thoroughly studied.

2.2 Motion Style Transfer

To achieve motion style transfer, early work aligns two motions to characterize their differences [Hsu et al. 2005], or by modeling the style in frequency domains [Bruderlin and Williams 1995; Pullen and Bregler 2002; Unuma et al. 1995; Yumer and Mitra 2016], but they largely handle relatively small amounts of data. Recent methods relying on a large labeled dataset learn the mapping between two different domains [Almahairi et al. 2018; Dong et al. 2020] or model the style directly as the common features across all motions with the same style label [Mason et al. 2022; Xia et al. 2015]. The style can be represented as one-hot embedding [Chang et al. 2022; Park et al. 2021; Smith et al. 2019] or style variable [Brand and Hertzmann 2000]. However, these representations lack the details of the given style sequence. Another strategy models the style as the variance of the latent vectors, using Gram Matrix [Holden et al. 2017a] or AdaIN [Aberman et al. 2020; Park et al. 2021]. To extract the fine-grained style variance, the following methods [Jang et al. 2022, 2023; Kim et al. 2024; Song et al. 2023] incorporate the body-part level attention mechanism. However, these methods do not handle temporal difference [Dong et al. 2020] and hardly handle hip velocity correctly since hip velocity is coupled with both style and content. Stylization is a related area of style transfer in which specific style motions are generated without involving the original motions. Autoregressively stylized motion generators [Mason et al. 2022, 2018; Tang et al. 2023; Tao et al. 2022; Xia et al. 2015] generate high-quality diverse motion while adhering to predefined constraints such as trajectory or keyframes. These methods, however, cannot incorporate the content of another motion.

3 METHODOLOGY

3.1 Hip Velocity-Contact Timing Relationship

Explicitly controlling contact would enable fine-grained style transfer, which enhances both the style expressiveness and the content quality. A naive approach would be to constrain the contact velocity and position. However, since the contact is tightly coupled with both style and content, this would result in unnatural and uncontrollable motions, as demonstrated in the Appendix. To this end, we aim to find a proxy, which can be leveraged to control the contact and can also be easily decoupled from the style and the content. Notably, while phase [Starke et al. 2022] is a popular implicit motion representation related to contact, it is not an ideal choice for two reasons. First, it is relevant to both style and content [Tang et al. 2023] and cannot be easily decoupled. Second, editing an implicit representation for flexible motion control is not straightforward. Instead, we find that *hip velocity* might be a good proxy because it is loosely related to the characteristics that are essential for expressing the style or content, such as body movement and poses, but also easily controllable and correlated to the contact. As illustrated in Fig. 2, although the hip speed magnitude is not explicitly related to the contact, the timing change of the hip speed trend such as the increasing trend to decreasing trend corresponds to the switch timing in contact.

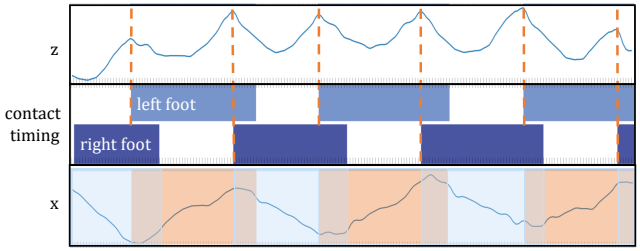

Fig. 2. The diagram illustrates the correlation between the contact pattern and the hip speed of a walking sequence from the STYLE100 Dataset. The "z" axis of our frame points to the character's forward-facing direction and the "x" axis points to the character's left, both in the local coordinate system at the current frame. In the middle (contact timing), there are two rows of bars (light blue and blue). The top bars represent the left foot contact duration and the bottom ones represent the right foot. The first row shows the "z" component of the hip velocity, in which there is a peak value when the foot makes contact with the ground. The curve in the third row depicts the "x" component of the hip velocity, with orange and light blue rectangles representing left and right foot contact duration, respectively. The "x" component of the hip velocity decreases (increase in negative "x" axis) during the right leg contact, and increases during the left leg contact.

To achieve the goal of controlling the contact by hip velocity, the remaining questions are whether the hip-contact relationship exists for other diverse motions and whether the hip velocity sequence of a motion is adequate for inferring the contact timing pattern. We conduct an experiment to explore these questions. Specifically, we model the potential hip-contact relationship with a convolution neural network, denoted by $f_\delta(\cdot)$, trained on two different motion datasets. The network takes the hip velocity $\mathbf{h} \in \mathcal{R}^{T\times 3}$ as input and predicts the contact states $c_t \in \mathcal{R}^{T\times 2}$ for two legs. The experimental results on two datasets both demonstrate high prediction precision, validating that the hip velocity sequence is sufficient to predict contact patterns for diverse motions. Notably, scaling the hip velocity sequence by a single factor before feeding it into the $f_\delta(\cdot)$ still preserves the high prediction precision, indicating that the contact pattern is more closely associated with the change of hip speed trend than with speed magnitude, which aligns well with the observation from Fig. 2. The details are shown in the Appendix.

To control the contact by the hip, our method aims to generate motions adhering to the learned hip-contact relationship by $f_\delta(\cdot)$ to control the contact, synthesizing high-quality results without foot-skating artifacts and can control contact flexibility.

3.2 Architecture

As shown in Fig. 3, our architecture consists of two stages. In the first stage, we apply three different networks to extract the features of style z_s, contact timing z_{ct} and trajectory z_{tj} from three input motions (M_s, M_{ct} and M_h), respectively. In the second stage, these three variables are composed to generate motions. Specifically, the

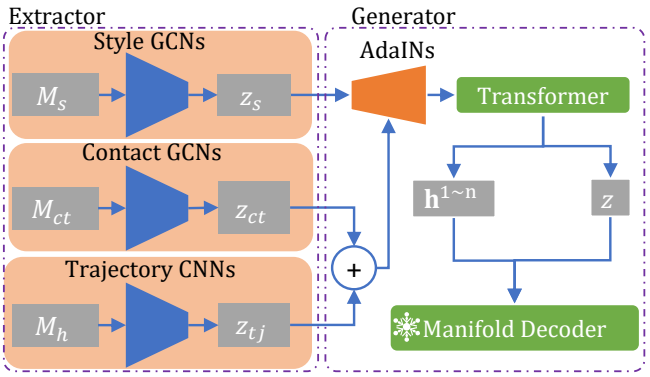

Fig. 3. Overview of our pipeline. The grey blocks represent the data and others indicate the network. Trapezium shapes indicate the presence of downsampling and upsampling in the network. The "❄" denotes the manifold decoder is frozen. Note that only the hip velocity of M_h is used as an input to Trajectory CNNs (see Appendix for details).

contact timing and trajectory features are composed into a content feature, which is modulated with the style feature by AdaIN blocks. Then a transformer predicts the hip velocity that satisfies the contact and trajectory conditions and a latent variable z for encoding the remaining style variations. Lastly, a manifold synthesizes the intended motion. We employ a conditional variational auto-encoder (CVAE) as the manifold, using the hip as the condition, illustrated in Fig. 4. Since the relationship is a kind of temporal pattern, our CVAE encodes the randomness of the whole sequence instead of encoding frame-level characteristics as in [Ling et al. 2020; Tang et al. 2022]. Besides, we apply an attention mechanism that employs hip velocity as the query to emphasize the relationship between hip velocity and motion dynamics. See Appendix for details.

Our manifold has two advantages. First, the manifold synthesizes motion sequences with leg movements that are compatible with hip velocity, reducing foot skating artifacts. Second, it decouples the hip velocity from the motion, giving us great control over the trajectory and contact timing, as shown in Sec. 4.

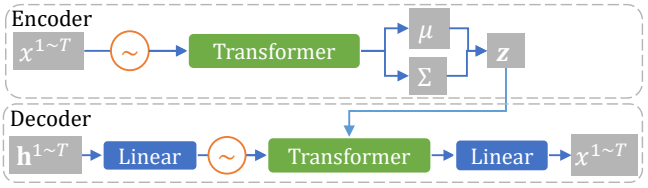

Fig. 4. Overview of our CVAE. The ~ in a circle represents the global positional embedding.

3.3 Losses

Instead of sampling three sequences from the dataset for every training step, we sample a style sequence M_s and a content sequence M_c, similar to previous motion style transfer methods. During training, we extract z_s from M_s, z_{tj} from M_c, and extract z_{ct} alternatively from M_s and M_c. We utilize the reconstruction loss L_{rec} and the cycle consistency loss L_{cyc} as described in [Aberman et al. 2020; Jang et al. 2022]:

$$\begin{aligned} L_{\text{cyc}} &= ||G(M_c, G_{scc}, G_{scc}) - M_c||_1 + ||G(G_{scc}, M_s, M_s) - M_s||_1, \\ L_{\text{rec}} &= ||G_{ccc} - M_c||_1 + ||G_{sss} - M_s||_1, \end{aligned} \tag{1}$$

where $||\cdot||_1$ represents L1 norm and $G_{sch} = G(M_s, M_c, M_h)$ denotes the synthesized motion with style extracted from M_s, contact timing from M_c, and trajectory from M_h. For instance, G_{scc} represents a motion with the style derived from M_s, while maintaining the contact timing and trajectory of M_c. Besides, to separate the trajectory, we integrate a trajectory loss, as indicated by:

$$L_{\text{tj}} = ||\mathbf{H}_{scc} - \mathbf{h}_c||_2^2 + \alpha_{tj}||proj(\mathbf{H}_{ssc}) - proj(\mathbf{h}_c)||_2^2, \tag{2}$$

where $\mathbf{H}_{scc}$ denotes the hip velocity of G_{scc}, and $proj(\cdot)$ extracts the trajectory by projecting the hip position onto the ground. Since both the trajectory and contact timing of $\mathbf{H}_{scc}$ should converge to $\mathbf{h}_c$, the first term does not extract the trajectory explicitly. In our experiment, we empirically set $\alpha_{tj} = 0.2$ to allow for a small deviation in trajectory. In addition, we propose a contact loss L_{ctt} to separate the contact timing, as denoted by:

$$L_{\text{ctt}} = ||f_\delta(\mathbf{H}_{scc}) - f_\delta(\mathbf{h}_c)||_2^2 + ||f_\delta(\mathbf{H}_{ssc}) - f_\delta(\mathbf{h}_s)||_2^2, \tag{3}$$

where $f_\delta(\cdot)$ is the learned function that captures the relationship between the hip velocity and the contact timing. We further propose a style loss L_{style} to enhance the style, leveraging the encoder of CVAE:

$$L_{\text{style}} = ||g(E(G_{scc})) - g(E(M_s))||_2^2, \tag{4}$$

where $E(\cdot)$ represents the latent variable from the final layer of the CVAE encoder, and g denotes the Gram matrix. As a result, the loss of our architecture is defined as:

$$L = L_{\text{rec}} + \alpha_0 L_{\text{cyc}} + \alpha_1 L_{\text{style}} + \alpha_2 L_{\text{tj}} + \alpha_3 L_{\text{ctt}}, \tag{5}$$

where all α_i are empirically set to 0.5 in our experiments.

To train the motion manifold, we employ a β-VAE training procedure, which aims to minimize both the reconstruction and KL-divergence losses. Note that our manifold learns the hip-contact relationship by representation learning without explicitly incorporating any related loss. The implementation details and data formatting are shown in the Appendix.

4 STYLE TRANSFER CONTROLLING

Trajectory Controlling. Transferring style without adjusting velocity may introduce unnaturalness if the character moves at a completely different speed in the style sequence than it does in the content sequence. By scaling the average magnitude of the hip velocity before feeding it into the transformer decoder, our method can modify speed magnitude without affecting contact or style. The experiments are shown in Sec.5.4. The results are shown in the Fig. 6. Besides, animators can conveniently set the trajectory from another motion without impacting the style by replacing the hip sequence input to the manifold decoder with the target hip sequence. This operation also changes the contact timing because it conforms to the hip velocity. The result is depicted in the final image of Fig. 6. To maintain contact timing, our method allows for the interpolation of trajectory features z_{tj} to establish a gradual transition from one trajectory to another. The results are shown in the third row of Fig. 5 and Fig. 9.

Contact Timing Controlling. We can automatically edit the contact timing by linearly interpolating z_{ct} between the content motion and the style motion. As demonstrated in the second row of Fig. 5, when interpolating from 0 to 1, the contact timing of the resulting motion approaches that of the faster-paced target sequence, resulting in a lighter and more agile style, while the spatial-temporal variations of "high-knees" and trajectory remain unchanged.

Our manifold also provides additional flexibility for controlling the contact timing, eliminating the need for reference motions, by solving the following optimization:

$$\arg\min_{\hat{\mathbf{h}}} \lambda||proj(\hat{\mathbf{h}}) - proj(\mathbf{h})||_2^2 + ||f_\delta(\hat{\mathbf{h}}) - c_t||_2^2 + \sigma||\hat{\mathbf{h}} - \mathbf{h}||_2^2, \quad (6)$$

where $c_t \in \mathcal{R}^{T\times 2}$ stands for the desired contact timing, λ is the weight for maintaining the trajectory, and σ is the weight for the regularization term. We begin the optimization by setting the initial value to the original hip velocity. The optimization aims to find the hip velocity that maintains approximately the existing trajectory while achieving the desired contact timing. After finding the intended hip velocity, our manifold modifies the leg movements accordingly. In Fig. 7, we demonstrate the use of this method to add footsteps and switch legs.

Style Controlling. Like previous methods [Aberman et al. 2020; Jang et al. 2022], interpolating between two style features z_s can affect the spatial-temporal variations of the resulting motion, as shown in the first row of Fig. 5.

5 EXPERIMENTS AND RESULTS

This section proposes multiple metrics (Sec. 5.1) and demonstrates ablation study (Sec. 5.2) and comparisons (Sec. 5.3). All evaluations are based on human animations represented by joint rotations, rather than network output. Human animation reconstruction can be divided into three categories based on the output data used: rotation-based, position-based, and velocity-based, which primarily use rotations, positions, and velocities. The Appendix shows details of these reconstruction methods and a related ablation study. We choose the velocity-based way for our method because it produces smooth motion and reduces foot-skating artifacts. However, because previous methods may not produce reasonable velocity and the velocity-based method will degrade their performance, we choose the most appropriate method for each to allow for fair comparisons.

5.1 Metrics

5.1.1 Manifold Metrics. We compare our manifold to three recent manifolds: MLD [Chen et al. 2023], VQVAE [Zhang et al. 2023b], and MVAE [Ling et al. 2020]. In addition, we propose a variant of our model that encodes frame-level randomness rather than sequence-level randomness, with architecture details provided in the Appendix. We set MLD's hyperparameters to be identical to our method, with the exception of the hip condition. For MVAE, we used a variant [Tang et al. 2022] that treats the hip as a condition. We also included the hip as a condition in VQVAE [Zhang et al. 2023b] to ensure a fair comparison.

We measure the manifolds from two aspects: contact timing controllability and motion quality. To this end, we randomly sample 20 motion sequences from each manifold for each hip trajectory. We employ the precision and recall rate of the predicted contact to evaluate controllability, which we refer to as *Contact Precision-Recall*. Precision is the percentage of the predicted contact change frames (feet touch or lift off the ground) that correctly match the ground-truth contact changes, while recall is the percentage of ground-truth contact changes that correctly match predicted contact changes. When there is no ground truth contact change, we apply $f_\delta(\mathbf{h})$ for replacement, where $\mathbf{h}$ is the predicted hip velocity. These metrics are distinguished by an asterisk (*) in the upper right corner of the results in the tables. To evaluate motion quality, we use three metrics. The first is the foot skating metric [Zhang et al. 2018], which calculates the average foot velocity v_f when the foot height h is within a threshold ($H = 2.5$), as defined by $L_f = v_f \cdot \text{clamp}(2 - 2^{h/H}, 0, 1)$. The second metric is FMD, which calculates the distance between the distribution of the dataset and the distribution of the randomly sampled data. We employ the joints' velocity to calculate the mean and covariance of the distribution. Finally, we conduct a user study in which 20 participants, five of whom are professional animation designers, evaluate the naturalness of motion sequences generated by different manifolds. For each manifold, we randomly sample five sequences with the same hip trajectory, and sample five different hip trajectories in total, resulting in 25 sequences for each manifold. Participants assign scores ranging from 1 to 5 concerning naturalness, with less than 3 denotes unnaturalness and more than 3 indicates naturalness. A score of 3 denotes indistinguishability.

5.1.2 Style Transfer Metrics. We evaluate the controllability of trajectory, contact timing, and style, respectively. Contact timing controllability is measured using contact precision-recall, while trajectory accuracy is the L2 distance between the trajectories of ground-truth motions and those of the generated motions. The style is measured by the FMD and style recognition accuracy (SA), similar to [Jang et al. 2022].

5.2 Ablation study

The experimental results on STYLE100 are shown in Tab. 1 and we leave the results on CMU in the Appendix.

5.2.1 Manifold quality and controllability. Among these manifolds, our manifold demonstrates standout performance, particularly concerning human perception scores on the STYLE100 dataset shown in Tab. 1 (the score of 4 represents the naturalness and we achieve 4.46). Besides, our method consistently achieves high ranks in terms of FMD and foot skating metrics. In terms of contact precision and contact recall rate, indicated by Ct P. and Ct R. in Tab. 1, our method outperforms two other transformer-based networks, MLD and Frame-Level. MLD does not explicitly condition hip velocity while Frame-Level applies the frame-level encoding rather than sequence-level encoding. These experiments validate the significance of the hip condition and long-clip encoding for learning the long-term relationship between contact timing and hip velocity. In addition, the effect of long-clip encoding for learning the relationship is validated by various architectures. VQVAE, which primarily employs convolution layers, achieves comparable contact precision-recall to our method as it also encodes the long sequence characteristic by down-sampling. In contrast, MVAE, which applies

a Mixture-of-Experts MLP and uses frame-level encoding, exhibits lower contact precision-recall.

5.2.2 Contact precision-recall for motion quality evaluation. The contact precision-recall is more meaningful in evaluating motion quality than the common metrics: the foot skating and the FMD. In particular, the foot skating metric only takes into account the foot velocity and can achieve low error by anticipating unintended zero value. Besides, the FMD metric evaluates distribution similarity and may not adequately represent quality, either. In contrast, the contact precision-recall indirectly evaluates the extent to which a sequence satisfies the hip-contact relationship of human motion, reflecting certain types of motion naturalness properly. As demonstrated in Tab. 1, high contact precision-recall metrics consistently correspond to high perception scores while FMD and foot skating do not. Specifically, VQVAE performs poorly in terms of FMD but ranks among the top-performing methods in terms of human perception, which indicates the inaccuracy of the FMD metric in measuring motion quality. Furthermore, as most of the manifolds achieve the foot skating metric that is close to the metric of the dataset (0.25), the foot skating metric is not sufficiently discriminative in this case.

Table 1. Comparisons between our manifold with three previous manifolds and a variation of our manifold. "Percep." represents the human perception score. Fv of the dataset is 0.68.

Methods	Ct P.	Ct R.	FMD_{vel}	Fv	Percep.
STYLE100					
MLD	0.4618*	0.4311*	0.0265	0.92	N/A
MVAE	0.6129	0.4843	**0.0116**	0.84	1.83
Frame-Level	0.5613	0.6480	0.0190	0.95	2.40
VQVAE	0.8633	0.8554	0.0475	**0.67**	3.83
Ours	**0.8782**	**0.8712**	0.0157	0.79	**4.46**

5.2.3 Contact controllability and style. This section evaluates the performance when replacing manifolds in our style transfer framework with other manifolds. We conduct three experiments to evaluate the style effects and controllability of contact timing and trajectory, respectively. We use Motion Puzzle [Jang et al. 2022] as a baseline for style effects. The results are shown in Tab. 2.

The prerequisite for applying L_{style} (Eq. 4) is that the manifold must decouple the style variations from the hip velocity, so that constraining latent variable z does not affect contact timing and trajectory. Otherwise, adding L_{style} significantly degrades performance, as demonstrated in the Appendix. Therefore, we do not employ L_{style} to ensure fair comparisons.

Our manifold achieves the best scores for expressing style, as seen in the FMD and SA metrics for all three experiments in Tab. 2. In the style experiment, other manifold methods tend to prioritize the reconstruction of the content sequence rather than conveying style, resulting in high contact precision-recall but low FMD and SA. We attribute our superior performance to the separation of trajectory and contact timing from spatial-temporal variations of style, so constraining contact timing and trajectory did not significantly affect the style.

Capturing the hip-contact relationship is critical for contact controllability. As evidence, MLD, MVAE, and Frame-Level, which have not learned the relationship, exhibit lower contact precision-recall than our manifold in contact controllability experiments. Notably, the methods that preserve the hip-contact relationship, such as VQ-VAE and our manifold, encounter the challenge of modifying the trajectory or contact timing without alerting each other because both factors are relevant to hip velocity. As demonstrated in the trajectory and contact experiment in Tab. 2, VQVAE struggles in resolving conflicts between meeting the requirements of contact timing and trajectory, which causes both factors to deviate from the intended target, leading to the worst results compared to other methods. Nevertheless, our proposed manifold satisfies contact timing and trajectory requirements effectively, achieving the best contact precision-recall among all the manifolds. In addition, the resulting trajectory differences (8.9 cm) for a two-second sequence are hard to discern visually.

Table 2. Comparisons between different manifolds for our framework. (XZ, Angle) are trajectory metrics, (FMD, SA) are style metrics and (Ct P., Ct R.) are contact timing metrics. Fv represents the foot skating metric. Fv of the dataset is 0.64.

Methods	(XZ	Angle)	(FMD	SA)	(Ct P.	Ct R.)	Fv
Style							
Motion Puzzle	5.5	0.046	87	**0.920**	0.467	0.476	1.68
MLD	3.3	0.027	135	0.751	0.862	0.872	**0.60**
MVAE	3.1	0.017	157	0.763	0.840	0.817	0.99
Frame-Level	2.5	0.027	194	0.571	**0.919**	**0.926**	0.61
VQVAE	4.1	0.032	182	0.603	0.861	0.876	0.53
Ours	**2.4**	**0.013**	**85**	0.879	0.849	0.849	0.61
Contact							
MLD	**5.2**	0.039	105	0.847	0.586	0.593	0.82
MVAE	8.89	0.030	128	0.824	0.649	0.645	0.97
Frame-Level	8.4	0.037	100	0.847	0.610	0.651	0.64
VQVAE	13.4	0.046	115	0.806	0.575	0.586	0.64
Ours	8.9	**0.031**	**70**	**0.931**	**0.741**	**0.784**	**0.58**
Trajectory							
MLD	**6.1**	0.037	129	0.805	0.561	0.566	0.62
MVAE	8.1	**0.027**	260	0.239	0.593	0.571	0.87
Frame-Level	7.7	0.029	159	0.734	0.590	0.651	0.63
VQVAE	13.9	0.047	143	0.719	0.531	0.569	**0.50**
Ours	8.9	0.031	**65**	**0.943**	**0.642**	**0.678**	0.60

5.3 Comparisons

We evaluate our method against previous style transfer methods, including [Aberman et al. 2020], Motion Puzzle [Jang et al. 2022], and a variant of Motion Puzzle that utilizes the similar decoupling formulation (denoted as Motion Puzzle (+ decouple)). Instead of using the velocity-based way as our method, [Aberman et al. 2020] and Motion Puzzle employ the rotation-based way to generate the final animation because [Aberman et al. 2020] uses rotation representation only and it's difficult for Motion Puzzle to accurately learn the joints' velocities. Additionally, [Aberman et al. 2020] sets hip velocity using a heuristic solution and does not preserve the trajectory. Therefore, we omit trajectory metrics (xz and angle) for [Aberman et al. 2020].

We conduct three experiments for evaluating the controllability of style, trajectory and contact timing by setting the corresponding interpolation factor to 0.5 and 1.0, respectively. The style controllability experiment with an interpolation factor of 1 is equivalent to the previous motion style transfer, which can validate that our method not only outperforms previous methods in controlling contact but also maintains superior style effects. We omit the results

of the interpolation of 0.5 in Tab. 3 because it results in a similar conclusion as the interpolation of 1.0. The complete table is shown in the Appendix.

Table 3. Comparisons between our method with previous motion style transfer methods.

Methods	(XZ	Angle)	(FMD	SA)	(Ct P.	Ct R.)	Fv
Style							
Aberman et al.	N/A	N/A	191	0.732	**0.791**	0.773	0.83
Motion Puzzle	5.5	0.046	87	0.920	0.467	0.476	1.68
+ decouple	1.7	**0.010**	**69**	0.940	0.363	0.294	1.91
Ours	**1.6**	0.014	72	**0.943**	0.782	**0.799**	**0.63**
Contact							
Aberman et al.	N/A	N/A	180	0.773	0.232	0.498	2.05
Motion Puzzle	79	0.862	**40**	**0.990**	**0.874**	**0.919**	0.72
+ decouple	3.6	**0.020**	52	0.968	0.458	0.270	1.91
Ours	**2.9**	0.031	70	0.931	0.741	0.784	**0.58**
Trajectory							
+ decouple	**2.7**	**0.015**	**48**	**0.970**	0.400	0.510	1.46
Ours	8.9	0.031	65	0.943	**0.642**	**0.678**	**0.60**

Table 4. The performance of methods using IK post-processing.

Methods	(FMD	SA)	(Ct P.	Ct R.)	Fv
Style					
Aberman et al.	247	0.574	0.839	0.798	0.67
Motion Puzzle	167	0.760	0.800	0.866	0.78
+ decouple	145	0.812	0.733	0.867	0.87

5.3.1 Style effects and motion naturalness. Our method achieves almost the best results for conveying style (FMD, SA) while also preserving the motion's naturalness. Specifically, in terms of motion naturalness, Motion Puzzle fails to generate high-quality results, as indicated by the high foot skating value and low contact precision-recall. Although [Aberman et al. 2020] improves the quality by employing discriminators, it still has worse foot skating artifacts than our method. To reduce foot skating, both [Aberman et al. 2020] and Motion Puzzle employ contact-based IK solvers as post-processing to preserve the contact timing of the content sequence. However, applying this post-processing may undermine the style, as indicated by the increased FMD and decreased SA values presented in Tab. 4. In section 5.4, we show that projecting their results onto our manifold is a better way for post-processing.

5.3.2 Contact controllability. Motion Puzzle's interpolation with a factor of 1.0 is equivalent to reproducing the style sequence, thus easily achieving superior contact precision-recall and style metrics (FMD and SA). However, it devastates the trajectory requirements, demonstrated by the trajectory error of up to 79 cm for factor 1.0. Decoupling the trajectory enables achieving trajectory requirements, but it can lead to leg movement incompatible with velocity, resulting in poor contact precision-recall metrics. Overall, our method modifies the contact timing while barely affecting style or trajectory, exhibiting the best contact controllability among the compared methods.

5.3.3 Trajectory controllability. Neither [Aberman et al. 2020] nor Motion Puzzle decouples the trajectory, so we did not compare them on the trajectory controllability task. Although the (+ decouple) approach achieves a trajectory closer to the desired trajectory, this is accomplished by ignoring the contact timing completely, evidenced by the poor performance in Ct P. and Ct R. Synthesizing a motion with a significantly different trajectory but similar contact timing is a challenging task. Our method performs better in this regard.

5.4 Manifold Capability

The manifold capability is evaluated from three perspectives. First, our manifold is able to preserve contact timing even when scaling the magnitude of the hip speed. Second, our approach allows for the complete replacement of both the trajectory and contact timing without compromising the quality of motion or style. Lastly, our proposed manifold approach can also serve as a post-processing stage for previous style transfer methods, improving the motion quality and providing additional control capability.

All the experimental results are presented in Tab. 5, with the original style transfer results ("Original" in the table) serving as a baseline. Our manifold demonstrates high contact precision-recall in scaling experiments, proving its ability to scale speed magnitude while preserving contact timing. Additionally, after replacing the hip velocity, our manifold achieves the desired trajectory and contact timing while maintaining style.

In addition, both scaling and replacing operations are relevant to the manifold only. Therefore, even without applying our complete approach, the contact timing and trajectory can be controlled by projecting a motion onto our manifold. To evaluate the performance of the projection, we project the results from previous style transfer methods. This projection approach significantly improves the motion quality, with fewer foot skating artifacts when compared to the results in Tab. 3 and better style effects than when using a contact-IK based solver as the post-processing. In conclusion, our manifold can serve as post-processing for other style transfer methods, improving motion quality while providing contact controllability.

Table 5. Manifold capability experiments comprise editing the hip velocity and projecting previous transfer methods' results onto our manifold.

Operation	(XZ	Angle)	(FMD	SA)	(Ct P.	Ct R.)	Fv
Hip editing							
Original	2.6	0.0190	66	0.940	0.773	0.787	0.61
Scaling	0.62	0.0065	59	0.941	0.760	0.763	0.86
Replacing	0.44	0.0044	20	0.992	0.927	0.908	0.57
Projection							
Aberman et al.	4.2	0.074	189	0.670	0.882	0.830	0.41
Motion Puzzle	5.4	0.046	105	0.842	0.832	0.882	0.50
+ decouple	1.2	0.009	84	0.880	0.794	0.839	0.53

6 DISCUSSIONS

Phase-related methods

Previous methods [Holden et al. 2017b; Starke et al. 2020, 2021] can use phase variables to control the contact. They first extract contact-related intermediate variables from motions and then learn the motion distribution based on the intermediate variables. While phase-related methods provide precise control over contact, they have not fully explored the variety of styles. For example, if the phase condition constrains multiple body-part movements, the result may be limited to a specific style [Tang et al. 2023]. Furthermore, these methods model phase using a temporal continuous function

and frequently require auto-regressive manner prediction from previous frames, reducing style diversity even further as the style is influenced by previous frames.

Unlike previous phase-related methods, we use hip velocity to control contact. Furthermore, we model the intrinsic relationship of human motion without imposing any explicit artificial constraints. In the absence of an explicit constraint, we initially struggled to generate precise out-of-distribution contact timing because we couldn't find a suitable hip velocity as the condition. However, we believe it is appropriate for a style transfer task because rhythm and frequency, which are easier to achieve, can express the style to some extent. Explicitly adhering to out-of-distribution contact patterns [Starke et al. 2020] may reduce motion quality and is undesirable for motion style transfer.

Limitations and Future Work

If the desired contact and trajectory are incompatible, our method may produce an unnatural motion. For example, transferring a walking motion to one that does not rely on the foot to support and move may result in a floating motion (see the failure case in Fig. 11 and our video). However, as shown in our video, the relationship holds for diverse motions including locomotion, dancing, jumping, and so on, which are applicable to the vast majority of human daily motion types. In addition, even for the cases in which the contact control is not applicable, compared to previous methods, our method is still able to provide extra trajectory control and a similar style transfer because of our manifold's diversity. Future research could explore a more general control method.

Additionally, manifold-based motion generation approaches have become popular in recent days. Lots of studies have utilized diffusion networks with either VAE [Chen et al. 2023] or VQVAE [Ao et al. 2023; Jiang et al. 2024; Zhang et al. 2023b] as the manifold. Future research could involve incorporating our manifold and metric to enhance motion quality and editing flexibility in diffusion applications.

7 CONCLUSION

We have presented a novel approach to character motion style transfer that addresses the difficult task of decoupling contact from motions. By extracting style, contact timing, and trajectory features, our approach enables fine-grained control over these aspects independently, resulting in more expressive and natural motion. Modeling the relationship between hip velocity and contact timing using a transformer architecture is the key insight for achieving finer control over contact timing while preserving naturalness. We also propose a new metric for measuring the match between the synthesized contact and hip velocity, which is also closely aligned with human perception in terms of motion naturalness. Furthermore, our proposed manifold is versatile in the sense that it can both directly generate motions and be used as post-processing for existing motion transfer techniques. Experiment results show that our method produces high-quality and expressive results for a wide range of motion styles, outperforming state-of-the-art methods in style expressivity and motion quality.

ACKNOWLEDGMENTS

Xiaogang Jin was supported by Key R&D Program of Zhejiang (No. 2023C01047) and the National Natural Science Foundation of China (Grant No. 62472373). We extend our appreciation to the participants who generously devoted their time to our user study.

REFERENCES

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

Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. 2018. Text2action: Generative adversarial synthesis from language to action. In *2018 IEEE International Conference on Robotics and Automation*. IEEE, 5915–5920.

Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-controllable speech-driven gesture synthesis using normalising flows. *Computer Graphics Forum* 39, 2 (2020), 487–496.

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

Amjad Almahairi, Sai Rajeshwar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. 2018. Augmented cyclegan: Learning many-to-many mappings from unpaired data. In *International Conference on Machine Learning*. PMLR, 195–204.

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

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

Okan Arikan and D. A. Forsyth. 2002. Interactive motion generation from examples. *ACM Transactions on Graphics* 21, 3 (2002), 483–490.

Philippe Beaudoin, Stelian Coros, Michiel van de Panne, and Pierre Poulin. 2008. Motion-motif graphs. In *Proceedings of the 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. 117–126.

Matthew Brand and Aaron Hertzmann. 2000. Style machines. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*. 183–192.

Armin Bruderlin and Lance Williams. 1995. Motion signal processing. In *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*. 97–104.

Jinxiang Chai and Jessica K. Hodgins. 2007. Constraint-based motion optimization using a statistical dynamic model. *ACM Transactions on Graphics* 26, 3 (2007), 8–es.

Ziyi Chang, Edmund JC Findlay, Haozheng Zhang, and Hubert PH Shum. 2022. Unifying human motion synthesis and style transfer with denoising diffusion probabilistic models. In *Proceedings of the 2023 International Conference on Computer Graphics Theory and Applications*.

Wenheng Chen, He Wang, Yi Yuan, Tianjia Shao, and Kun Zhou. 2020. Dynamic future net: Diversified human motion generation. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2131–2139.

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.

Hsu-kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. 2019. Action-agnostic human pose forecasting. In *Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision*. 1423–1432.

Yuzhu Dong, Andreas Aristidou, Ariel Shamir, Moshe Mahler, and Eakta Jain. 2020. Adult2child: Motion style transfer using cyclegans. In *Proceedings of the 11th Annual International Conference on Motion, Interaction, and Games*. 1–11.

Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F. Troje, and Marc-André Carbonneau. 2023. ZeroEGGS: Zero-shot Example-based Gesture Generation from Speech. *Computer Graphics Forum* 42, 1 (2023), 206–216. https://doi.org/10.1111/cgf.14734 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14734

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

Daniel Holden, Oussama Kanoun, Maksym Perepichka, and Tiberiu Popa. 2020. Learned motion matching. *ACM Transactions on Graphics* 39, 4 (2020), 1–12.

Daniel Holden, Taku Komura, and Jun Saito. 2017b. Phase-functioned neural networks for character control. *ACM Transactions on Graphics* 36, 4 (2017), 1–13.

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

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

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

Deok-Kyeong Jang, Yuting Ye, Jungdam Won, and Sung-Hee Lee. 2023. MOCHA: Real-Time Motion Characterization via Context Matching. In *SIGGRAPH Asia 2023 Conference Papers*. 1–11.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. 2024. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems* 36 (2024).

Boeun Kim, Jungho Kim, Hyung Jin Chang, and Jin Young Choi. 2024. MoST: Motion Style Transformer between Diverse Action Contents. *arXiv preprint arXiv:2403.06225* (2024).

Chaelin Kim, Haekwang Eom, Jung Eun Yoo, Soojin Choi, and Junyong Noh. 2023. Interactive locomotion style control for a human character based on gait cycle features. *Computer Graphics Forum* (2023), e14988.

Lucas Kovar, Michael Gleicher, and Frédéric Pighin. 2002. Motion graphs. *ACM Transactions on Graphics* 21, 3 (2002), 473–482.

Sergey Levine, Jack M Wang, Alexis Haraux, Zoran Popović, and Vladlen Koltun. 2012. Continuous character control with low-dimensional embeddings. *ACM Transactions on Graphics* 31, 4 (2012), 1–10.

Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel van de Panne. 2020. Character controllers using motion VAEs. *ACM Transactions on Graphics* 39, 4 (2020), 1–12.

Julieta Martinez, Michael J Black, and Javier Romero. 2017. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2891–2900.

Ian Mason, Sebastian Starke, and Taku Komura. 2022. Real-time style modelling of human locomotion via feature-wise transformations and local motion phases. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 5, 1, 1–18.

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

Jianyuan Min and Jinxiang Chai. 2012. Motion graphs++: A compact generative model for semantic motion analysis and synthesis. *ACM Transactions on Graphics* 31, 6 (2012), 1–12.

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

Mathis Petrovich, Michael J Black, and Gül Varol. 2021. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10985–10995.

Mathis Petrovich, Michael J Black, and Gül Varol. 2022. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*. Springer, 480–497.

Katherine Pullen and Christoph Bregler. 2002. Motion capture assisted animation: Texturing and synthesis. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*. 501–508.

Sarah Ribet, Hazem Wannous, and Jean-Philippe Vandeborre. 2019. Survey on style in 3d human body motion: Taxonomy, data, recognition and its applications. *IEEE Transactions on Affective Computing* 12, 4 (2019), 928–948.

Alla Safonova and Jessica K. Hodgins. 2007. Construction and optimal search of interpolated motion graphs. *ACM Transactions on Graphics* 26, 3 (2007), 106–es.

Yijun Shen, He Wang, Edmond S. L. Ho, Longzhi Yang, and Hubert P. H. Shum. 2017. Posture-based and action-based graphs for boxing skill visualization. *Computers and Graphics* 69, Supplement C (2017), 104–115.

Harrison Jesse Smith, Chen Cao, Michael Neff, and Yingying Wang. 2019. Efficient neural networks for real-time motion style transfer. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 2, 2, 1–17.

Wenfeng Song, Xingliang Jin, Shuai Li, Chenglizhao Chen, Aimin Hao, and Xia Hou. 2023. FineStyle: Semantic-aware fine-grained motion style transfer with dual interactive-flow fusion. *IEEE Transactions on Visualization and Computer Graphics* (2023).

Sebastian Starke, Ian Mason, and Taku Komura. 2022. DeepPhase: periodic autoencoders for learning motion phase manifolds. *ACM Transactions on Graphics* 41, 4 (2022), 1–13.

Sebastian Starke, Yiwei Zhao, Taku Komura, and Kazi Zaman. 2020. Local motion phases for learning multi-contact character movements. *ACM Transactions on Graphics* 39, 4, Article 54 (2020).

Sebastian Starke, Yiwei Zhao, Fabio Zinno, and Taku Komura. 2021. Neural animation layering for synthesizing martial arts movements. *ACM Transactions on Graphics* 40, 4 (2021), 1–16.

Xiangjun Tang, He Wang, Bo Hu, Xu Gong, Ruifan Yi, Qilong Kou, and Xiaogang Jin. 2022. Real-time controllable motion transition for characters. *ACM Transactions on Graphics* 41, 4 (2022), 1–10.

Xiangjun Tang, Linjun Wu, He Wang, Bo Hu, Xu Gong, Yuchen Liao, Songnan Li, Qilong Kou, and Xiaogang Jin. 2023. RSMT: Real-time stylized motion transition for characters. In *SIGGRAPH '23 Conference Proceedings* (August 6-10).

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.

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

Munetoshi Unuma, Ken Anjyo, and Ryozo Takeuchi. 1995. Fourier principles for emotion-based human figure animation. In *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*. 91–96.

He Wang, Edmond SL Ho, and Taku Komura. 2015. An energy-driven motion planning method for two distant postures. *IEEE Transactions on Visualization and Computer Graphics* 21, 1 (2015), 18–30.

He Wang, Edmond SL Ho, Hubert PH Shum, and Zhanxing Zhu. 2019. Spatio-temporal manifold learning for human motions via long-horizon modeling. *IEEE Transactions on Visualization and Computer Graphics* 27, 1 (2019), 216–227.

He Wang, Kirill A Sidorov, Peter Sandilands, and Taku Komura. 2013. Harmonic parameterization by electrostatics. *ACM Transactions on Graphics* 32, 5 (2013), 1–12.

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

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

He Zhang, Sebastian Starke, Taku Komura, and Jun Saito. 2018. Mode-adaptive neural networks for quadruped motion control. *ACM Transactions on Graphics* 37, 4 (2018), 1–11.

Haotian Zhang, Ye Yuan, Viktor Makoviychuk, Yunrong Guo, Sanja Fidler, Xue Bin Peng, and Kayvon Fatahalian. 2023a. Learning physically simulated tennis skills from broadcast videos. *ACM Transactions On Graphics* 42, 4 (2023), 1–14.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. 2023b. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14730–14740.

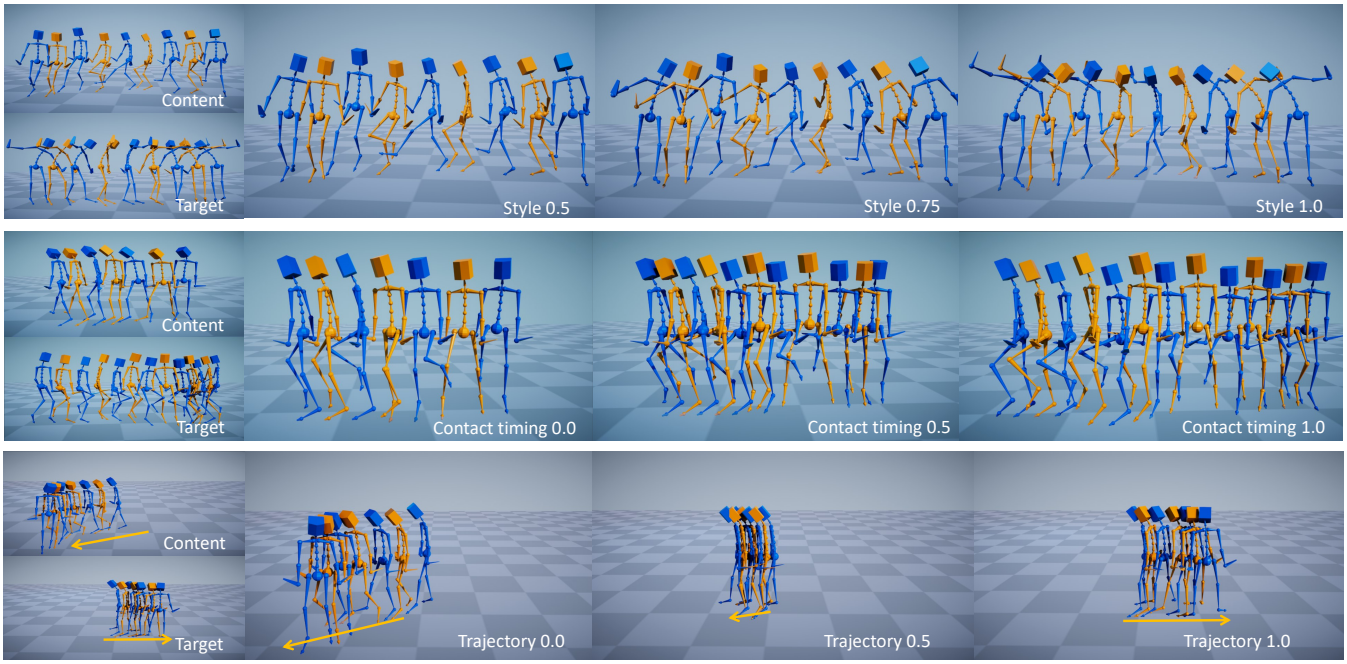

Fig. 5. Our method allows users to control trajectory, contact timing, and style of motions separately or jointly. To improve visibility, we intentionally increase the spatial distance between two adjacent skeletons. Blue skeletons represent the frames when the character makes contact with the ground, while orange skeletons represent the midpoint between two blue frames. Our method allows for separate interpolation of style (first row), contact timing (second row), and trajectory (third row). The latent style/contact/trajectory space interpolation parameter here varies from content sequence (0.0) to target sequence (1.0).

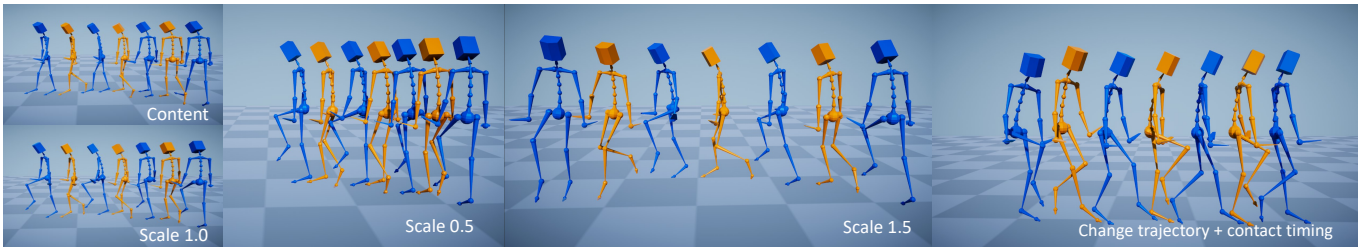

Fig. 6. Our method allows changing the trajectory by scaling the magnitude of the hip velocity (Scale 0.5, Scale 1.0, and Scale 1.5), as well as changing the trajectory and contact timing at the same time by setting its hip velocity from another motion (change trajectory + contact timing). In this setting, all results are transferred to the "high knees" style.

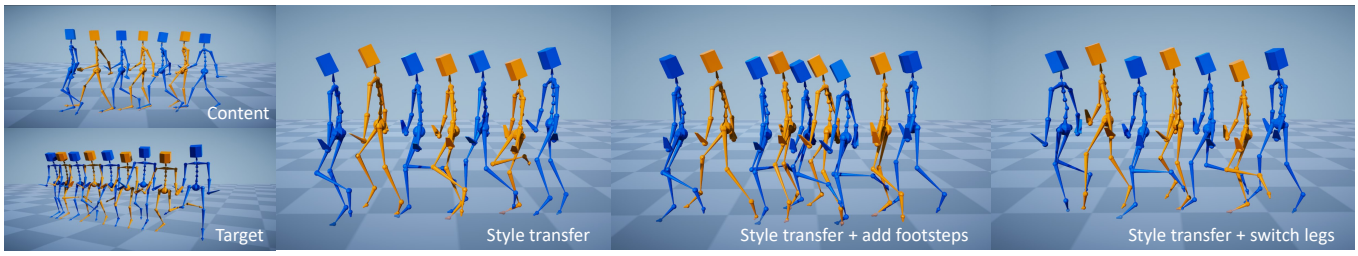

Fig. 7. Our method allows for fine-grained manual control of the contact timing patterns. Following the style transfer process (second column), we show examples of additional control by adding footsteps (second to last column) or switching legs (last column).

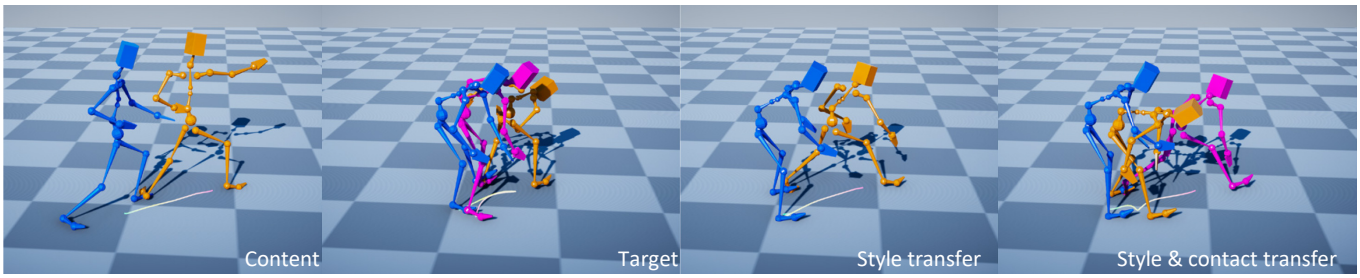

Fig. 8. Our contact control can also be employed in non-locomotion cases. This figure shows a martial art case. The color of the character changes from blue to orange and to purple over time. The character in the content sequence takes a step forward into a bow stance. In the target sequence, the character keeps the left leg static and moves the right leg twice. Our method can transfer the style and contact from the target sequence to the content sequence.

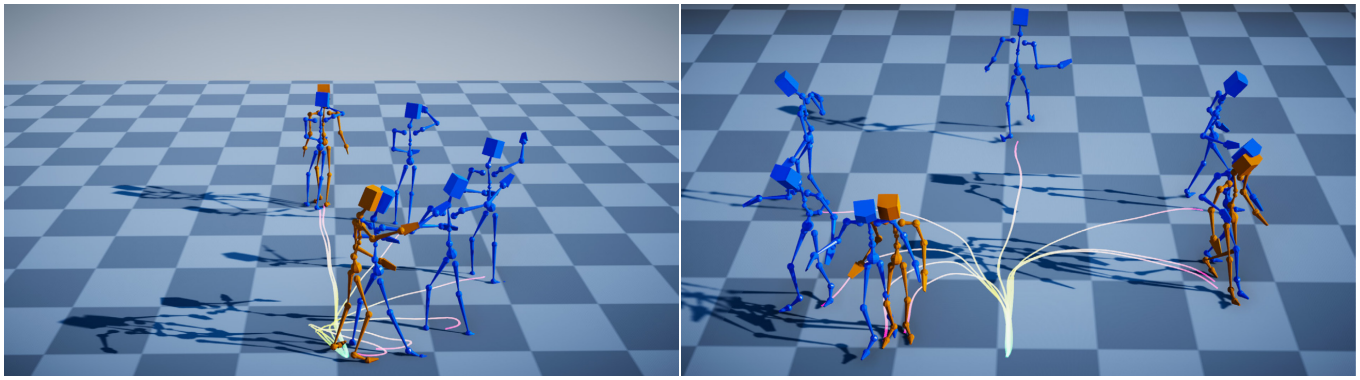

Fig. 9. We show our trajectory interpolation results. The orange characters in each image represent the last frame of the content and target motions, respectively. The blue characters are the last frame of the interpolated motions. The first image demonstrates the interpolation between forward locomotion and a dance motion with intricate trajectory. The second image showcases the interpolation between two motions involving walking in different directions.

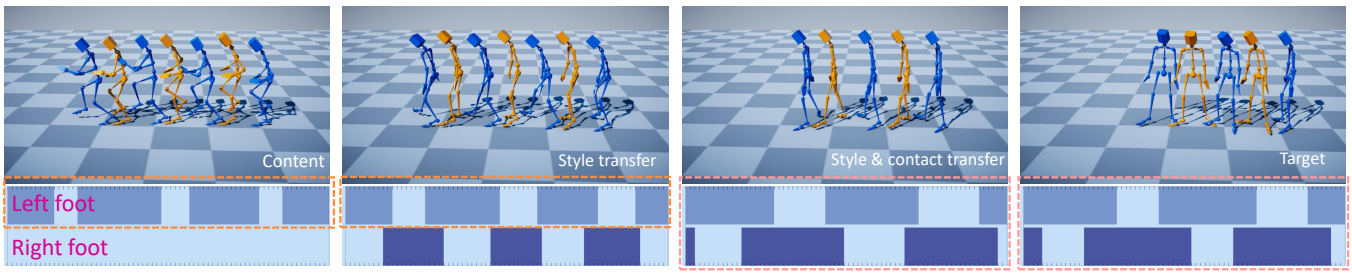

Fig. 10. This case presents our contact control when transferring style from a motion that makes contact using only the left foot to a motion with a "duck foot" style. The style transfer result showcases the "duck foot" style while preserving the contact timing pattern of the left foot, and the style & contact transfer modifies the contact as well. Similar contact timing patterns are highlighted using the same color boxes. Blue skeletons represent the frames when the left foot of the character makes contact with the ground, while orange skeletons represent the midpoint between two blue frames.

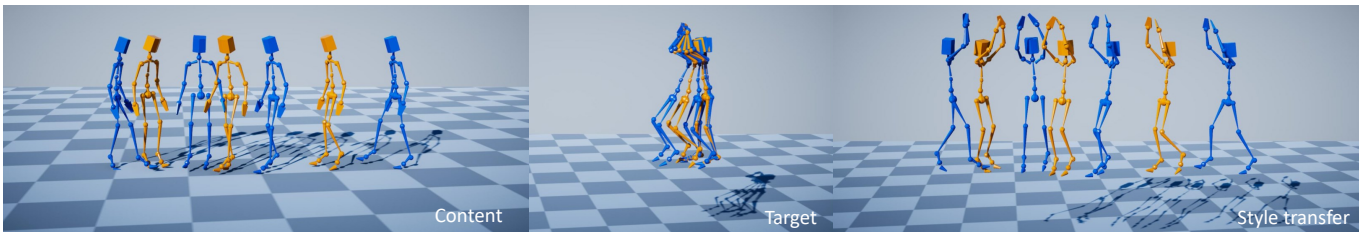

Fig. 11. A failure case. Transferring a walking motion to one that does not rely on the foot for support and movement may result in a floating motion.