Supplementary Material
StyleMotif: Multi-Modal Motion Stylization using Style-Content Cross Fusion

Overview

- Appendix A: User Study.
- Appendix B: Efficiency Analysis.
- Appendix C: Additional implementation details.
- Appendix D: Additional ablation study.

A. User Study

To further evaluate the perceptual quality of our results, we conducted a user study comparing our method against SMooDi [9]. The study involved expert participants, including professional animators, motion researchers, and PhD students in related fields. Each participant was shown a series of motion pairs (over 200 in total), with each pair consisting of outputs generated by our method and by SMooDi. For each pair, the participant was asked to choose the version they subjectively preferred based on motion plausibility, smoothness, and overall quality. The results show that our method was preferred in **79.8%** of the comparisons, demonstrating a significant perceptual advantage over the baseline. This consistent preference underscores the effectiveness of our method in producing high-quality stylized motions.

B. Efficiency Analysis

We compare the efficiency of our method with SMooDi [9] in terms of learnable parameters and inference speed (seconds per sample), under the same diffusion step setting. As shown in Table 1, our model reduces the number of trainable parameters by ***43.9%***, significantly easing training. While the overall parameter count remains comparable, our single-branch design boosts inference speed by ***22.5%***, outperforming SMooDi's structure. Notably, our style encoder is deeper and thus accounts for most of the computational cost, but *our single-banch design allows for highly parallelizable operations*. In contrast, SMooDi, which duplicates and separates two pre-trained diffusion backbones for content and style, requires output summation after each block, limiting parallel efficiency despite fewer overall parameters. Consequently, our method achieves faster practical inference and more efficient training.

Method	Overall Parameter	Learnable Parameter	Inference Time
SMooDi [9]	468M	13.9 M	4.0 s
StyleMotif	**462 M**	**7.8 M**	**3.1 s**
Improvement	*1.3%*	*43.9%*	*22.5%*

Table 1. **Efficiency Comparison.** For inference time, we report the average time cost (s) per sample on a single NVIDIA A100 GPU.

C. Additional Implementation Details

C.1. Training Details

Our model is trained on a single NVIDIA A100 GPU for 50 epochs with a batch size of 64. We employ the AdamW optimizer [5] with a constant learning rate of 1e-5. The pre-trained generation network used is MLD [1]. Following SMooDi [8], we randomly assign the content text to be ϕ and mask out 10% of the style motion sequence along the temporal dimension. During training, the number of diffusion steps is set to 1000, which is reduced to 50 during inference. Regarding the hyperparameters, we adopt $\gamma = 0.6$ for Eq. 5 and $\tau_0 = 1.0$ for Eq. 8 in the main paper. The proposed style-content cross fusion is adopted only once after the 4-th transformer encoder block of MLD, *i.e.* $m = 4$. For the style encoder pre-training, we train for 200 epochs on the 100STYLE dataset [6], starting from the pre-trained encoder of MLD's VAE. For multi-modal alignment, we freeze the text encoder of ImageBind [2] and train only a single linear projection layer using curated motion-text data pairs for 200 epochs via contrastive learning.

C.2. Implementation Details of Generation Guidance and Learning Scheme

We follow [8] to adopt the style guidance mechanisms and the learning scheme. Here we provide a detailed description.

C.2.1. Style Guidance

We utilize both *classifier-free* and *classifier-based* style guidance to balance content and style in the generated motion:

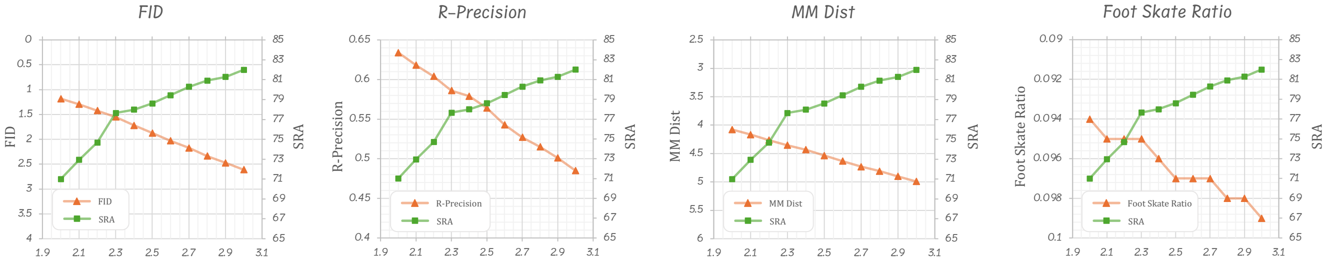

Figure 8. **Impact of Style Guidance Intensity Ratio τ in Eq. 10 for Motion-guided Stylization**. Higher ratios improve style accuracy while degrade content preservation.

Classifier-Free Guidance. We incorporate dual conditioning for content and style. The denoising process is guided by contrasting outputs with and without style input:

$$\begin{aligned}\epsilon_\theta(z_t, t, \tau_\theta(c), \psi_\theta(s)) &= \epsilon_\theta(z_t, t, \emptyset, \emptyset)\\ &+ w_c(\epsilon_\theta(z_t, t, \tau_\theta(c), \emptyset) - \epsilon_\theta(z_t, t, \emptyset, \emptyset))\\ &+ w_s(\epsilon_\theta(z_t, t, \tau_\theta(c), \psi_\theta(s)) - \epsilon_\theta(z_t, t, \tau_\theta(c), \emptyset)) \quad (9)\end{aligned}$$

where w_c and w_s control the strength of content-text alignment and style transfer, respectively. This decomposition enables flexible control over the trade-off between content fidelity and style adherence. Note that during diffusion training, we adopt the same w_c and w_s with [9].

Classifier-Based Guidance. To refine stylization, we augment the diffusion process with gradient updates from a pre-trained style feature extractor f [8]. The guidance is computed as:

$$\epsilon_\theta \leftarrow \epsilon_\theta + \tau \nabla_{z_t} \|\hat{\mathcal{F}}_s - \mathcal{F}_s\|_2 \quad (10)$$

where $\hat{\mathcal{F}}_s$ is the encoded style features of the decoded motion $\hat{x}_0$, from the predicted clean latent $\hat{z}_0$, while $\mathcal{F}_s$ is the input style features. τ modulates the intensity ratio of style guidance. This approach ensures stronger alignment with the reference style while maintaining motion realism. For the style interpolation, we utilize the style motions retrieved from the multi-modal features to calculate the classifier-based gradient. Note that different from [9], we utilize $\mathcal{L}_2$ normalization when calculating the gradient.

The combination of these two guidance mechanisms offers complementary benefits that the classifier-free term provides a balanced trade-off between content and style, while the classifier-based term enforces precise style fidelity.

C.2.2. Learning Scheme

The training process of our framework combines the following objectives:

Standard Denoising Loss. The foundational loss ensures accurate noise prediction during training:

$$\mathcal{L}_{std} = \mathbb{E}_{\epsilon, \psi_\theta(s)}[\|\epsilon_\theta(z_t, t, \tau_\theta(c), \psi_\theta(s)) - \epsilon\|^2] \quad (11)$$

Content Prior Preservation Loss. To mitigate content-forgetting, we follow [8] to include samples from HumanML3D dataset [3] to preserve the model's ability to handle diverse motions:

$$\mathcal{L}_{pr} = \mathbb{E}_{\epsilon', \psi_\theta(s)'}[\|\epsilon_\theta(z'_t, t, \tau_\theta(c'), \psi_\theta(s)') - \epsilon'\|^2] \quad (12)$$

Cycle Prior-preservation Loss. This loss encourages robust alignment of style and content across datasets by swapping styles and contents between pairs of motion sequences:

$$\begin{aligned}\mathcal{L}_{cyc} = \mathbb{E}[\|&\epsilon_\theta(z_t^{sh}, t, \tau_\theta(c), \psi_\theta(s)^{hs})\\ &+ \epsilon_\theta(z_t^{hs}, t, \tau_\theta(c'), \psi_\theta(s)^{sh}) - \epsilon - \epsilon'\|^2]\end{aligned} \quad (13)$$

Here, z^{hs} and z^{sh} represent cross-dataset style-content combinations, ensuring the model does not overfit to a narrow style distribution while maintaining content consistency.

The composite loss function is defined as:

$$\mathcal{L}_{all} = \mathcal{L}_{std} + \lambda_{pr}\mathcal{L}_{pr} + \lambda_{cyc}\mathcal{L}_{cyc} \quad (14)$$

where λ_{pr} and λ_{cyc} control the relative importance of content preservation and cyclic consistency, respectively.

C.3. Implementation Details of Motion Style Transfer

The motion style transfer task involves generating a stylized motion sequence by combining a content motion sequence and a style motion sequence. Although our model is trained with text as the content input rather than motion, it can seamlessly support this task without requiring additional training. Specifically, we utilize the DDIM reverse process [7] to derive the noised latent code $\bar{z}_T$ corresponding to the content motion sequence, formulated as:

$$\begin{aligned}z_{t+1} = &\sqrt{\frac{\alpha_{t+1}}{\alpha_t}} z_t + \left(\sqrt{\frac{1}{\alpha_{t+1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1}\right)\\ &\cdot \epsilon_\theta(z_t, t, \tau_\theta(c), \emptyset),\end{aligned} \quad (15)$$

where α represents the noise scale and t denotes time step. The noised latent code $\bar{z}_T$ can be obtained at the final reverse step T, which substitutes original Gaussian noise and serves as the input for the diffusion network. For this task, the number of denoising steps is set to 30.

C.4. Dataset Details

The 100STYLE dataset [6] contains style labels that may inherently convey content-related meanings, such as *"TwoFootJump,"* which could conflict with content text like *"a person is running."* To mitigate such conflicts between style labels and content text, we apply a filtering process to the 100STYLE dataset during evaluation, following the approach of SMooDi [9]. Specifically, we organize the style labels in the 100STYLE dataset into six categories based on [4]: character (CHAR), personality (PER), emotion (EMO), action (ACT), objective (OBJ), and motivation (MOT). Since the ACT group includes labels that describe specific actions, which overlap with content-related information, we exclude motions from this group when computing the SRA metric for content text derived from the HumanML3D dataset. Additionally, for multi-modal alignment, we refine the style motions by retaining only motions with *"forward"* movement and removing others. This selective filtering minimizes noise and maintains the stability of the text-motion alignment space, facilitating more effective multi-modal motion stylization.

D. Additional Ablation Study

Impact of Style Guidance Ratio. In Figure 8, we analyze the effect of varying the style guidance ratio τ in Eq. 10 by adjusting the style guidance weights. Our findings show that increasing this ratio enhances the SRA metric but simultaneously degrades R-Precision, MM Dist, FID, and the foot skate ratio. This trade-off suggests that higher style guidance ratios improve style accuracy at the cost of content preservation. Notably, when the absolute value of the style guidance ratio exceeds 2.3, the SRA improvement plateaus, while the other metrics continue to deteriorate relatively more rapidly. Based on these observations, we identify $\tau = 2.3$ as the optimal balance point for the style guidance ratio.

References

[1] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *CVPR*, 2023. 1

[2] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023. 1

[3] 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 *CVPR*, 2022. 2

[4] Hye Ji Kim and Sung-Hee Lee. Perceptual characteristics by motion style category. In *Eurographics (Short Papers)*, 2019. 3

[5] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[6] Ian Mason, Sebastian Starke, and Taku Komura. 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*, 2022. 1, 3

[7] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2023. 2

[8] Lei Zhong, Yiming Xie, Varun Jampani, Deqing Sun, and Huaizu Jiang. Smoodi: Stylized motion diffusion model. In *European Conference on Computer Vision*, pages 405–421. Springer, 2024. 1, 2

[9] Lei Zhong, Yiming Xie, Varun Jampani, Deqing Sun, and Huaizu Jiang. Smoodi: Stylized motion diffusion model. In *ECCV*, 2024. 1, 2, 3