STYLEMOTIF: Multi-Modal Motion Stylization using Style-Content Cross Fusion

Ziyu Guo[†]
The Chinese University of Hong Kong
ziyuguo@link.cuhk.edu.hk

Young Yoon Lee
Roblox
ylee@roblox.com

Joseph Liu
Roblox
josephliu@roblox.com

Yizhak Ben-Shabat
Roblox
ibenshabat@roblox.com

Victor Zordan
Roblox
vbzordan@roblox.com

Mubbasir Kapadia
Roblox
mkapadia@roblox.com

Abstract

We present **STYLEMOTIF**, *a novel Stylized Motion Latent Diffusion model, generating motion conditioned on both content and style from multiple modalities. Unlike existing approaches that either focus on generating diverse motion content or transferring style from sequences,* STYLEMOTIF *seamlessly synthesizes motion across a wide range of content while incorporating stylistic cues from* ***multi-modal*** *inputs, including motion, text, image, video, and audio. To achieve this, we introduce a style-content cross fusion mechanism and align a style encoder with a pre-trained multi-modal model, ensuring that the generated motion accurately captures the reference style while preserving realism. Extensive experiments demonstrate that our framework surpasses existing methods in stylized motion generation and exhibits emergent capabilities for multi-modal motion stylization, enabling more nuanced motion synthesis. Project Page:* https://stylemotif.github.io.

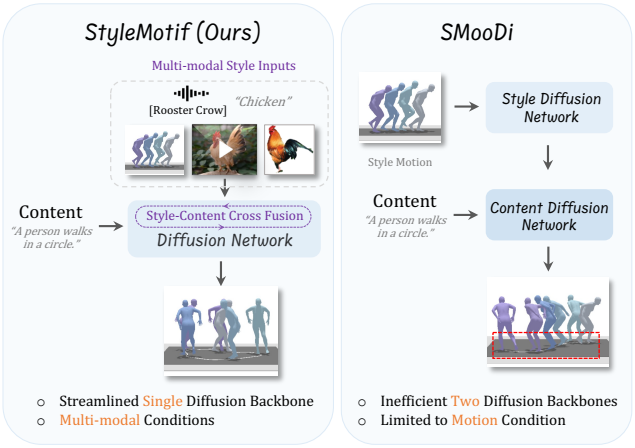

Figure 1. **Comparison of Our Proposed STYLEMOTIF with SMooDi.** SMooDi duplicates and separates two diffusion backbones for content and style, increasing complexity. In contrast, STYLEMOTIF adopts a single diffusion backbone with direct style injection, enabling efficient and realistic motion stylization.

1. Introduction

Human motion generation is a fundamental task in computer graphics and animation, enabling the synthesis of realistic and expressive human movements. Broadly, human motion can be characterized by two complementary aspects: *content*, which defines the underlying action (*e.g.*, walking, jumping), and *style*, which encodes variations such as personal flair, emotional expression, or cultural influences (*e.g.*, jubilant, aggressive). This separation allows for greater control and flexibility in generating motion, making it particularly valuable in creative industries like game development, film production, and virtual reality. However, traditional approaches to stylized motion generation often depend on manual processes such as motion capture or keyframe animation, which are costly, time-consuming, and labor-intensive.

Recent progress in text-to-motion (T2M) diffusion frameworks [6, 27, 54] has greatly advanced the ability to translate natural-language prompts into realistic human motions. By leveraging powerful denoising diffusion models, these approaches capture intricate spatiotemporal dependencies in the data, enabling coherent sequences of human movements to be generated directly from brief textual descriptions. Despite their success in content fidelity and diversity, most current T2M diffusion methods concentrate primarily on *what* action is performed, while overlooking *how* it is performed, namely, its stylistic details. Simply appending a separate style-transfer module to text-driven motion diffusion pipelines can introduce additional complexity and risk of compounding errors in the final output.

In parallel, motion style transfer has been actively studied to infuse stylistic cues from a reference motion (or style

[†] Work done as an intern at Roblox.

data) into another sequence [1, 24, 37, 53]. Although many of these methods effectively disentangle content and style for small-scale tasks, the pipeline becomes cumbersome when a large variety of content motions need to be stylized. Moreover, they commonly assume that the input or target sequences are high-quality motion data. In scenarios where content motions are synthetically generated, or partially noisy, the transfer process can deteriorate, leading to undesirable motion artifacts or compromised style fidelity.

To address the need for simultaneously controlling both content and style, some recent works have merged style encoding with diffusion-based motion generation. Among these, the most recent and representative approach [73] augments a pre-trained latent diffusion model [6] with a style adaptor and classifier-based style guidance, achieving stylized motion from textual prompts and motion-style references. While effective, this method relies on additional training diffusion backbones, which shares structural similarities with ControlNet [66] as shown in Figure 1, which increases model complexity and training overhead. It is also constrained to motion as the primary style input. Concurrent work [33] proposes a bidirectional control flow mechanism to mitigate conflicts between style and content, extending style control to multiple modalities, but this approach also duplicates two separate diffusion backbones with substantial training overhead and limited applicability across modalities.

To this end, we propose **StyleMotif**, a new framework for *multi-modal motion stylization* that unifies text-to-motion diffusion with style conditioning in a *single*-diffusion structure. Specifically, we leverage a pre-trained motion latent diffusion model (MLD) [6] to preserve strong content generation capabilities, and seamlessly integrate *style features* extracted from a dedicated encoder, which is aligned with a multi-modal foundation model [12]. In contrast to previous works, StyleMotif avoids duplicating large portions of the network or relying on specialized style diffusion network. Instead, we introduce a *style-content cross fusion* mechanism, which injects stylistic cues into the diffusion process while maintaining motion realism. As a result, our StyleMotif not only yields more robust stylized outputs but also supports diverse style signals, such as motion, text, images, audio, or video clips, via the alignment in multi-modal feature space.

We summarize our main contributions as follows:

- We present **StyleMotif**, a stylized latent motion diffusion framework that unifies diverse motion content and *multi-modal* styles within a compact *single*-diffusion design.
- We propose a *style-content cross fusion module* that injects stylistic cues into the diffusion denoising process, achieving faithful stylization without compromising motion realism and ensuring efficiency.
- We achieve a unified multi-modal style feature space and unveil new *emergent capabilities* through *multi-modal alignment*, which accommodate various sources including motion, text, images, audio, and video, for flexible and versatile multi-modal style control.
- Extensive experiments demonstrate that StyleMotif consistently outperforms existing methods regarding style expressiveness, content preservation, and efficiency.

2. Related Work

Human Motion Generation. Recent progress in human motion generation [3, 5, 8, 13, 14, 41–43, 46, 55, 57, 58, 62] has been driven by transformer [14, 34] and diffusion models [2, 8, 49, 66], showing great potential in producing realistic and diverse motions. These approaches have shown great potential in producing realistic and diverse motion sequences. For example, Momask [14] improves motion generation using a residual VQ-VAE. Similarly, LaMP [34] introduces a motion-aware text encoder and a motion-to-text language model to enhance motion quality through text conditioning. Diffusion models, in particular, have become a key approach in motion generation [6, 22, 27, 48, 54, 61, 65, 67]. MDM [54] introduces a motion diffusion model that operates directly on raw motion data to capture the relationship between motions and input text conditions. MLD [6] improve efficiency by embedding the diffusion process in latent space, reducing computational cost. They also allow conditioning on specific constraints, such as predefined trajectories [27, 55, 59] or human-object interactions [7, 40, 57], enabling greater control and diversity. Our work leverages pretrained motion latent diffusion model [6] and multi-modal foundation models [12] to achieve stylized human motion generation while maintaining high motion quality.

Motion Stylization. Motion stylization [1, 24, 28, 37, 39, 44, 45, 50, 51, 53, 56, 60] involves transferring stylistic features from a reference motion to a source motion/textual prompt, enabling creative transformations while preserving the original motion content. Early methods, like those in motion style transfer [1, 24], typically separate motion content and style for recombination. For instance, Aberman et al. [1] used a generative adversarial network to decouple style from content without paired data, while Motion Puzzle [24] allows style control for individual body parts, and Guo et al. [15] utilized pretrained motion models for better style integration. Recent approaches, like SMooDi [73], generate stylized motion from text and style sequences using style guidance. However, these models face two main challenges: *(1)* inefficient architectures that duplicate diffusion backbones for content and style , and (*2) limited to style motion for guidance*. Another concurrent approach, MulSMo [33], attempts to address some of these by introducing a bidirectional control flow between style and content networks, reducing conflicts. Still, it remains restricted to a few modalities and duplicates diffusion networks. In this

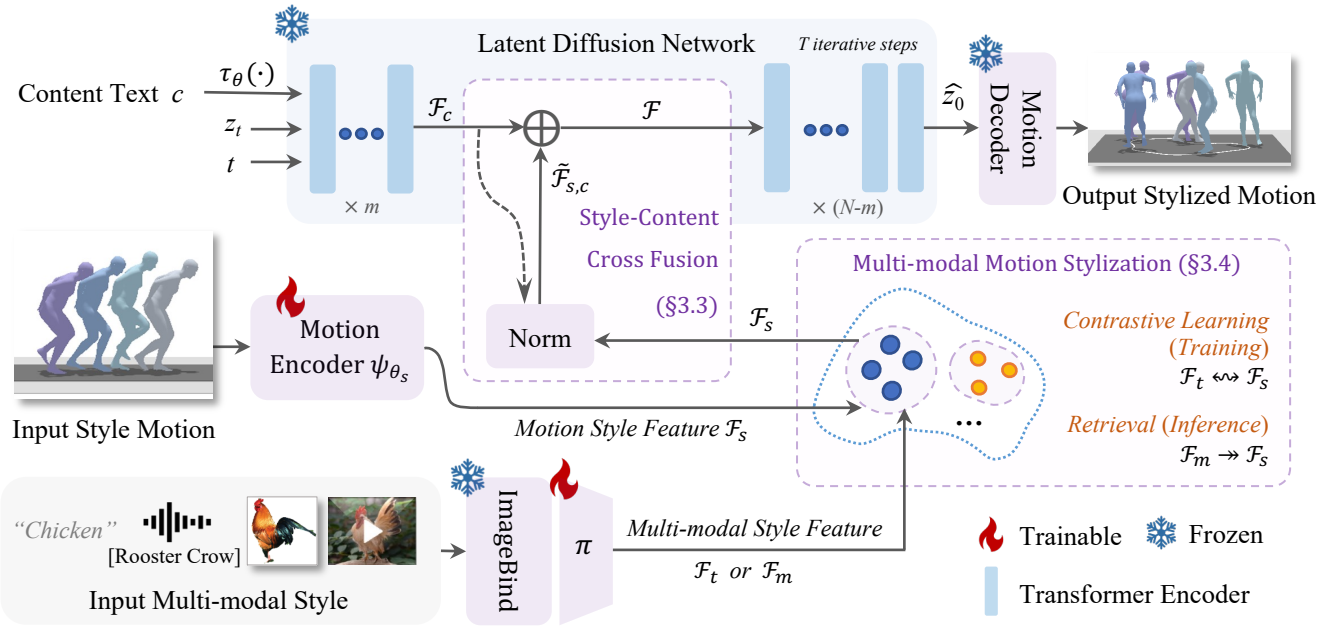

Figure 2. **Overall Pipeline of StyleMotif**, a diffusion framework for multi-modal motion stylization. Given a text prompt and a reference style from various modalities, our model extract style features and fuse them with content by style-content cross fusion. Through multi-modal alignment with contrastive learning, we enable seamless multi-modal conditioning and flexible stylization across motion, text, images, audio, and video.

work we eliminate redundant diffusion backbone, resulting in a simpler, more effective, and efficient approach to stylized motion generation. Our method also enables efficient multi-modal motion stylization, supporting text, image and multi-modal inputs while maintaining both high stylization quality and motion realism.

Multi-modality Learning. Recent advancements in multi-modality learning [9, 10, 38] have revolutionized various domains by enabling joint understanding across diverse data types. Foundational models like CLIP [47] and its extension [16, 68, 69, 75] establish robust vision-language alignment, while ImageBind [12], Point-Bind [17], and Language-Bind [74] expand this paradigm to diverse modalities, demonstrating emergent cross-modal capabilities. The rise of multi-modal large language models [4, 19, 21, 30, 31, 64, 72] further enhances semantic understanding through unified text-visual processing, achieving superior performance across 2D images [11, 29, 70], 3D point clouds [17, 18, 52], and complex reasoning scenarios [20, 26, 71]. For human motion, some approaches [5, 14, 25] incorporate multiple modalities into human motion understanding and generation tasks. Our work introduces the first framework for multi-modal guided motion stylization, achieving seamless style injection from diverse modalities through a single diffusion architecture with style-content cross fusion.

3. StyleMotif: Multi-Modal Motion Stylization with Style-Content Cross Fusion

We propose StyleMotif, a novel framework for stylized motion synthesis that combines style-content cross fusion and multi-modal motion stylization, as illustrated in Figure 2. Our approach integrates a style-content cross fusion mechanism that allows for coherent feature blending, ensuring the generated motion accurately reflects the reference style while maintaining the content's realism (§3.2). This mechanism is complemented by a multi-modal motion stylization strategy, which leverages inputs from diverse modalities such as image, video, audio, and text to provide control over the stylization process (§3.3). This framework enables highly flexible and realistic stylized motion synthesis, which offers greater control and diversity.

3.1. Overview

Preliminaries. Motion Latent Diffusion (MLD) [6] formulates a conditional latent diffusion model by training $\epsilon_\theta(z_t, t, c)$ to denoise a sequence of latents $\{z_t\}_{t=0}^{T}$, where $z_t \in \mathbb{R}^{n \times d}$ represents the motion latent at timestep t, conditioned on the distribution $p(z_t|c)$. A conditional domain encoder $\tau_\theta(c)$, such as CLIP [47], enables text-to-motion tasks, and the model is trained as $\epsilon_\theta(z_t, t, \tau_\theta(c))$. To incorporate additional style conditioning, a style condition s can be introduced with its own encoder $\psi_\theta(s)$, which may take the form of a motion or text encoder. This extends the model to $\epsilon_\theta(z_t, t, \tau_\theta(c), \psi_{\theta_s}(s))$. SMooDi [73] adopts a ControlNet [66]-style approach for style conditioning by

creating a trainable copy of the neural network weights θ_s from the original MLD model θ_c. This copy includes zero-initialized linear layers, denoted as $\mathcal{Z}(\cdot;\cdot)$. The output $\mathcal{F}^i(z_t, t, \tau_\theta(c), \psi_\theta(s); \theta_c)$ of the i-th MLD block, now conditioned on style, is computed as:

$$\mathcal{F}^i(z_t, t, \tau_\theta(c), \psi_{\theta_s}(s); \theta_c) = \mathcal{F}^i(z_t, t, \tau_\theta(c); \theta_c) + \mathcal{Z}(\mathcal{F}^i(z_t, t, \tau_\theta(c), \psi_{\theta_s}(s); \theta_s); \theta_{z_i}) \quad (1)$$

A key property of this formulation is that since θ_{z_i} is initialized to zero, thus $\mathcal{F}^i(z_t, t, \tau_\theta(c), \psi_\theta(s)) = \mathcal{F}^i(z_t, t, \tau_\theta(c))$ at the beginning. However, the tradeoff of that method is that it needs to maintain the additional parameters θ_s and θ_{z_i} in order for style transfer.

StyleMotif. We propose **StyleMotif**, eschewing zero linear layers in favor of injecting it directly via statistical manipulation, with the pipeline shown in Figure 2 describing our approach. Our StyleMotif utilizes latent space diffusion within a single generative diffusion backbone, building upon a pretrained MLD model [6]. Instead of perturbing the outputs of each block i via zero initialized layer $\mathcal{Z}(\mathcal{F}^i(z_t, t, \tau_\theta(c), \psi_\theta(s); \theta_s); \theta_{z_i})$, StyleMotif replaces this with a statistically transformed style embedding that is injected into the original MLD branch. This simplifies the model while ensuring high-quality stylization results, whose details are presented in § 3.2. Also, we follow SMooDi's generation guidance and training scheme to ensure high-quality stylized motion synthesis. We use a hybrid guidance strategy that balances content fidelity and style adherence by combining classifier-free and classifier-based techniques during diffusion sampling.

3.2. Style-Content Cross Fusion

Style Encoder Pre-training. To establish a robust foundation for style-content fusion, we combine the content knowledge from the pre-trained MLD's [6] VAE with the style knowledge from the 100STYLE [37] dataset. The MLD's VAE is pre-trained on the HumanML3D [13] dataset, which provides extensive understanding of content motion. Building on this, we further fine-tune the model on the 100STYLE dataset in a variational autoencoding manner to initially align the content and style data distributions in the latent space. After training, we discard the decoder and retain only the encoder as the motion style encoder. This reconstruction task enables the encoder to learn robust motion feature representations, which are essential for supporting the subsequent stylization process. By integrating content and style knowledge in this way, we ensure a strong foundation for seamless style-content fusion.

Style-Content Cross Normalization. To train a stylized diffusion model $\epsilon_\theta(z_t, t, \tau_\theta(c), \psi_\theta(s))$, we effectively fuse content and style features directly within the latent space diffusion process *instead of* duplicating two pre-trained diffusion backbones for content and style or separate control mechanisms [66]. Prior work [23, 32] has proven that feature statistics (mean, variance) effectively encode style. Given the output features $\mathcal{F}^i$ of the i-th block, we derive the content features $\mathcal{F}_c^i = \mathcal{F}^i(z_t, t, \tau_\theta(c); \theta_c)$ from the input text c, and the style features $\mathcal{F}_s = \psi_{\theta_s}(s)$ are extracted from the reference motion sequence s using the pre-trained style encoder. To perform the fusion, we first compute the mean μ_c and variance σ_c^2 of the content features across the feature dimension:

$$\mu_c = \frac{1}{D}\sum_{j=1}^{D} \mathcal{F}_c^{i,j}, \quad (2)$$

$$\sigma_c^2 = \frac{1}{D}\sum_{j=1}^{D} \left(\mathcal{F}_c^{i,j} - \mu_c\right)^2, \quad (3)$$

where $\mathcal{F}_c^{i,j}$ denotes the content features of the i-th block and the j-th feature element. Next, we normalize the *style* features $\mathcal{F}_s$ using these *content* statistics, ensuring that the style features are adapted to the content's statistical properties:

$$\widetilde{\mathcal{F}}_{s,c} = \frac{\mathcal{F}_s - \mu_c}{\sqrt{{\sigma_c}^2 + \eta}}, \quad (4)$$

where η is a constant added for numerical stability. This style-content cross normalization ensures that the style features are smoothly integrated with the content features while maintaining the content's original structure. After that, we add the normalized style features to the content features, formulating our final cross-normalization as

$$\mathcal{F}^i(z_t, t, \tau_\theta(c), \psi_{\theta_s}(s); \theta_c) = \mathcal{F}_c^i + \gamma \cdot \widetilde{\mathcal{F}}_{s,c}, \quad (5)$$

where γ is a parameter used for scaling the normalized value. $\widetilde{\mathcal{F}}_{s,c}$ can be thought of as a perturbation of $\mathcal{F}_c^i$ as it is scaled within its range. Notably, the fusion process is performed only once after the m-th block during the denoising process to avoid distorting the content while effectively introducing the style. This efficient fusion method eliminates the need for additional learnable parameters, as it is based solely on statistical transformations, achieving high-quality results with minimal computational overhead.

3.3. Multi-Modal Motion Stylization

Our StyleMotif extends to support multi-modal motion stylization. To achieve this, we integrate a pre-trained multi-modal foundation model [12], which provides a unified, multi-modal aligned feature space, enabling effective cross-modal alignment between motion and other modalities. With this alignment, our model can flexibly combine multiple

input modalities, such as text, image, audio, and video, to guide the stylization process in a comprehensive manner. This capability results in ***emergent abilities*** for multi-modal motion stylization, allowing for more nuanced outputs.

Motion-Text Pair Curation. To align motion with other modalities, we process a set of motion-text pairs using the curated 100STYLE subset [37], which is carefully selected to avoid conflicts between content and style motions, so that the model can effectively learn the relationships between motion and text. Each motion sequence is paired with a corresponding single textual label, which serves as the text prompt. The curated motion-text pairs will be used for the following cross-model alignment.

Multi-Modal Alignment. To maintain the alignment within the multi-modal space, we freeze the text encoder of ImageBind [12] and introduce a lightweight projection layer after the encoder. This projection layer aligns the feature dimensions of the text and motion encoders. Since the multi-modal model represents each modality with a global feature, our alignment process focuses on these global features to achieve robust alignment between text and motion representations, formalized as:

$$\mathcal{F}_t = \pi\big(\mathcal{E}_{\text{text}}(l)\big), \tag{6}$$

$$\mathcal{F}_s = \psi_{\theta_s}(s), \tag{7}$$

where l and s are the paired input text label and style motion, $\mathcal{F}_t$ and $\mathcal{F}_s$ denote the feature representations from the text and style motion encoders, $\mathcal{E}_{\text{text}}$ and ψ_{θ_s}, respectively, and π represents the projection operation that maps text features into the motion feature space. We employ a contrastive learning loss [63] to align the feature spaces of motion and text, bringing them closer together in the shared multi-modal space. Thus, we obtain a unified space between motion and all the modalities. We formulated it as

$$\mathcal{L}_{\text{align}} = -\frac{1}{2}\sum_{(i,j)} \log \frac{\exp \frac{\mathcal{F}_t^i \cdot \mathcal{F}_s^j}{\tau_0}}{\sum_k \exp \frac{\mathcal{F}_t^i \cdot \mathcal{F}_s^k}{\tau_0}} + \log \frac{\exp \frac{\mathcal{F}_t^i \cdot \mathcal{F}_s^j}{\tau_0}}{\sum_k \exp \frac{\mathcal{F}_t^k \cdot \mathcal{F}_s^j}{\tau_0}}, \tag{8}$$

where t and s represent two modalities (text and style motion) and (i, j) indicates a positive pair in each training batch, k indexes all samples in the batch, including both positive and negative ones, and τ_0 is a temperature parameter. During inference, we obtain the multi-modal (text, image, video, or audio) features from the multi-modal input m,

$$\mathcal{F}_m = \mathcal{E}_{\text{ImageBind}}(m), \tag{9}$$

where $\mathcal{E}_{\text{ImageBind}}$ denotes ImageBind [12]. We use the multi-modal features $\mathcal{F}_m$ to retrieve the most semantically similar motion features from the unified multi-modal space. The retrieved motion features are then used to guide the stylization process, ensuring that fine-grained style details are preserved and accurately reflected in the final output. This approach enhances the model's ability to generate stylized motions that are both contextually relevant and visually coherent.

4. Experiment

We provide extensive quantitative and qualitative analyses across multiple tasks in this section.

4.1. Experimental Settings

Implementation Details. We adopt the pre-trained MLD [6] as the foundation for motion generation. The style encoder of our model derives from the encoder of MLD's VAE, while the projection layer after text encoder is a single Linear layer. During diffusion training, we only enable the style encoder to be trainable while freezing other parameters. The model is optimized using the AdamW optimizer [35] with a constant learning rate of 10^{-5}.

Dataset Settings. We use the HumanML3D [13] dataset as our primary motion content dataset, which collects 14,616 motion sequences from AMASS [36] and annotates 44,970 sequence-level textual description. To train the style network, we utilize the 100STYLE [37] dataset, containing 45,303 style motions. We also adopt the text annotations for 100STYLE from previous work [73], which are pseudo text descriptions generated from MotionGPT [25]. We utilize the consistent root-velocity motion representations from HumanML3D for both the content and style data.

Baselines. We mainly evaluate our method against SMooDi [73] on motion-guided stylization and motion style transfer tasks. We also compare our text-guided stylication with 'ChatGPT+MLD' approach, which utilize ChatGPT to combine style text and content text, and functions as a straightforward text-to-motion model without control capabilities [73]. Additionally, we qualitatively compare our model with SMooDi on motion-guided stylization.

Evaluation Metrics. To evaluate our model, we use several metrics following previous works [73]. We measure R-precision, Multi-modal Distance (MM Dist), Diversity, and Frechet Inception Distance (FID) to assess how well the content from text is preserved and how realistic the generated motions are. We use Style Recognition Accuracy (SRA) to evaluate how accurately the style of the reference motion is reflected in the generated motion. Also, we adopt the Foot Skate Ratio, which helps assess the realism of the generated motion, reducing artifacts like foot sliding. During evaluation, following previous works [73], we randomly select

Method	SRA ↑	FID ↓	MM Dist ↓	R-Precision ↑ (Top-3)	Diversity →	Foot Skate Ratio ↓
***Motion*-Guided Stylization**						
MLD + Aberman [1]	54.37	3.309	5.983	0.406	8.816	0.347
MLD + Motion Puzzle [24]	63.77	6.127	6.467	0.290	6.476	0.185
SMooDi [73]	72.42	1.609	4.477	0.571	9.235	0.124
StyleMotif (Ours)	**77.65**	**1.551**	**4.354**	**0.586**	7.567	**0.097**
***Text*-Guided Stylization**						
MLD + ChatGPT [73]	4.82	0.614	4.313	0.605	8.836	0.131
StyleMotif (Ours)	**56.71**	**0.603**	**3.684**	**0.690**	9.101	**0.101**

Table 1. **Quantitative Results for Motion-Guided and Text-Guided Stylization. Bold** values denote the best performance. As there is no ground-truth reference for Diversity, no value is highlighted in bold; but the metric is provided for reference.

Method	SRA ↑	FID ↓	Foot Skate Ratio ↓
MLD + Aberman [1]	61.01	3.892	0.338
MLD + Motion Puzzle [24]	67.23	6.871	0.197
SMooDi [73]	65.15	1.582	0.095
StyleMotif (Ours)	**68.81**	**1.375**	**0.094**

Table 2. **Quantitative Results of Motion Style Transfer on HumanML3D [13] dataset.** Our method outperforms previous works in all metrics, which demonstrates effective style-content fusion for high-quality motion style transfer, providing significant advantages for downstream tasks besides motion stylization.

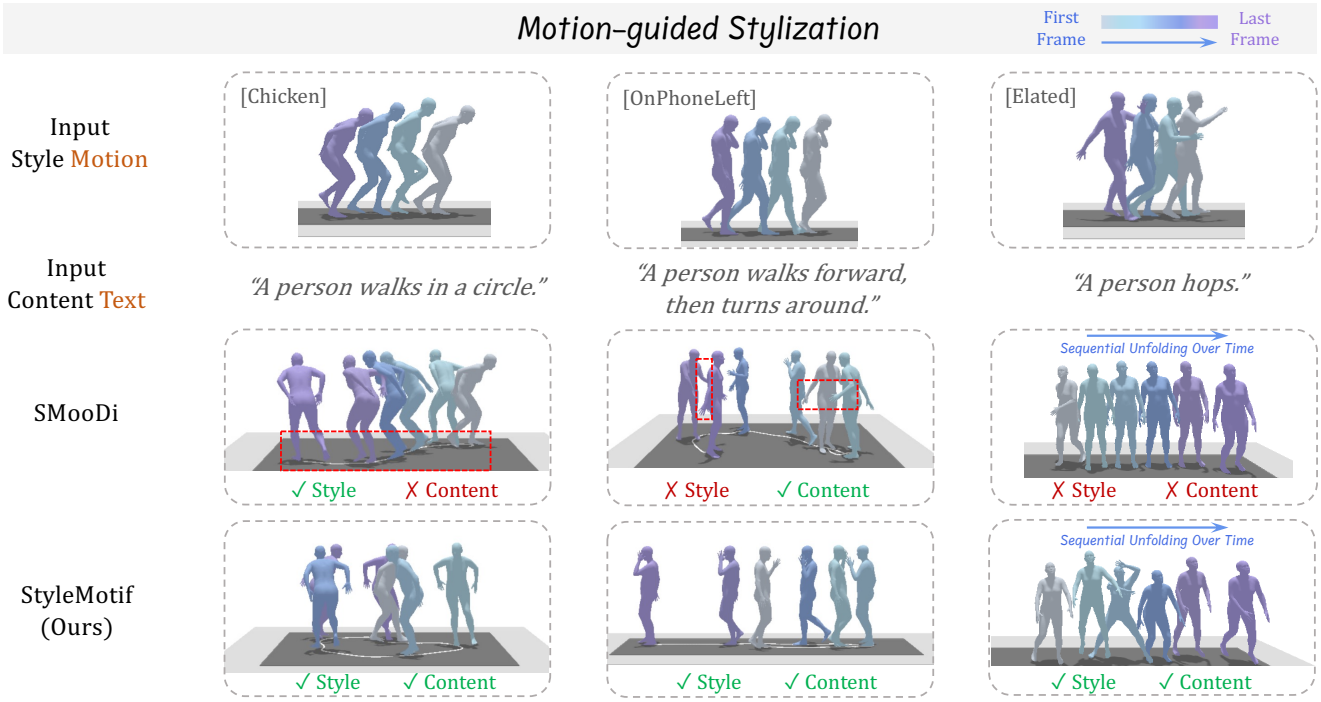

Figure 3. **Qualitative Results of Motion-Guided Stylization.** Our model generates cohesive and realistic motions that effectively align style and content, such as preserving the 'circular' trajectory (first column) and 'hop' content (third column). In contrast, SMooDi [73] struggles to maintain content fidelity and sometimes fails to reflect the specified style (e.g., 'phone on the left' in the second colum).

a content text from HumanML3D [13] and a style motion from 100STYLE [37], and compute SRA for the generated motion using a pre-trained classifier [73].

4.2. Motion-guided Stylization

Quantitative Analysis. In Table 1, we report the quantitative results for motion-guided stylization, where the motion serves as style input while text prompt as content one. Our method outperforms three baseline approaches [1, 24, 73] in all metrics. Specifically, we achieve a ***5.23%*** improvement in SRA while maintaining competitive performance in FID compared to the best baseline [73]. It demonstrates that our method, by effectively utilizing style-content cross fusion, generates motions that better align with the style reference while maintaining content integrity, showcasing the strength of our design in balancing style and content preservation.

Qualitative Analysis. In Figure 3, we show some qualitative results of motion-guided stylization from our model and baseline [73]. Compared with the baseline, our model produces more cohesive stylized motions, with better alignment of style and content. For instance, in the provided examples, SMooDi fails to maintain the *'circular'* trajectory or *'hop'* action specified by the content or cannot accurately reflect the intended style, *'phone on the left'*. This indicates that our approach more effectively integrates style and content, resulting in more realistic and consistent stylized motions.

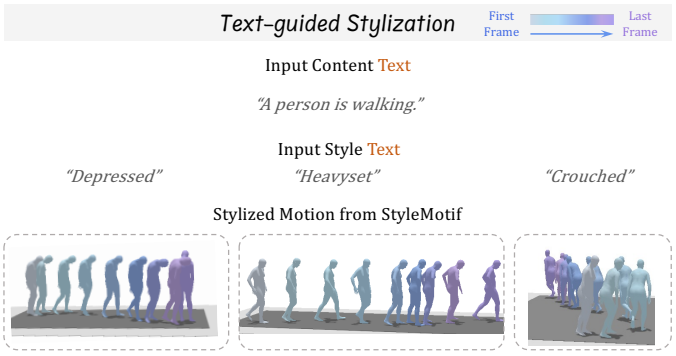

Figure 4. **Qualitative Results of Text-Guided Stylization.** Our model seamlessly integrates textual style descriptions with content, producing visually coherent and stylistically consistent results.

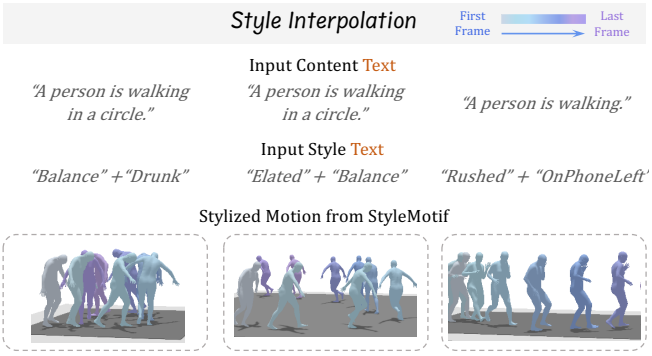

Figure 5. **Qualitative Results of Style Interpolation.** Our model blends multiple style inputs while preserving content integrity, demonstrating effective style-content fusion of our model.

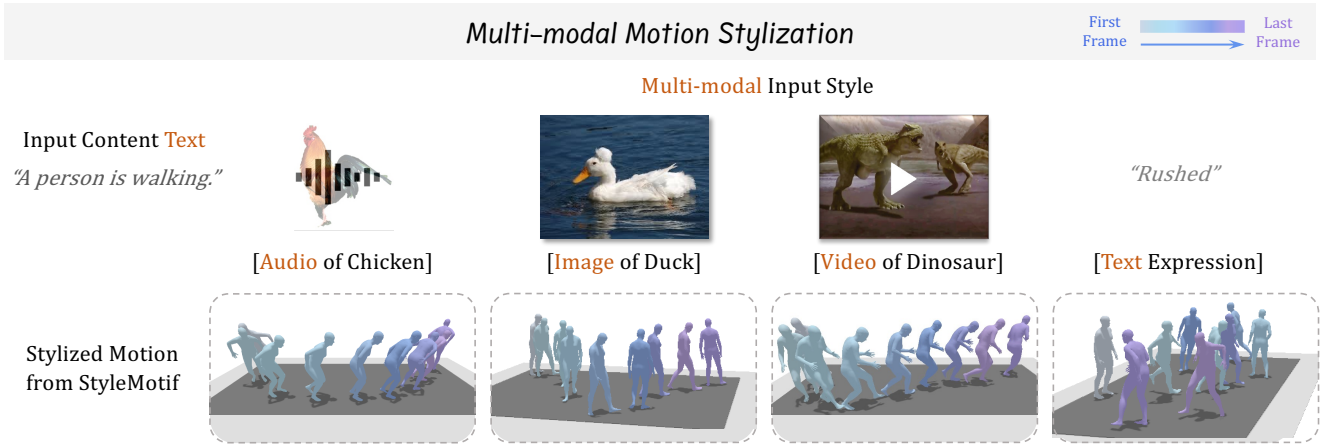

Figure 6. **Qualitative Results of Multi-Modal Motion Stylization.** Our model generates stylized motions guided by diverse modalities (e.g., text, image, video, audio), effectively transferring style while preserving content integrity.

4.3. Text-guided Stylization

Quantitative Analysis. For the text-guided stylization task, we use the HumanML3D [13] dataset for motion content and employ the single text label [37] as the style control. As reported in Table 1 (Bottom), our method significantly outperforms the baseline, achieving ***56.71%*** in SRA, compared to 4.82% for 'ChatGPT + MLD'. Moreover, our method maintains competitive FID, balancing style reflection with content preservation. This indicates that our design, which combines multi-modal alignment, help utilizes the style signal in the motion-text shared spaces and achieve significant effectiveness in text-guided stylization.

Qualitative Analysis. In Figure 4, we showcase qualitative results for text-guided stylization, where our model also demonstrates strong capability in harmonizing style and content, generating high-quality and visually coherent results.

4.4. Motion Style Transfer

For the motion style transfer task, we use the HumanML3D [13] dataset for motion content and the 100STYLE dataset [37] for motion styles. As shown in Table 2, our method outperforms the baseline models across all key metrics. Specifically, we achieve a ***3.66%*** improvement in SRA, indicating better style reflection, and a ***0.207*** reduction in FID reflecting improved realism. These demonstrate that our framework not only improves the alignment between style and content but also provides significant advantages for downstream tasks like motion style transfer with efficient style-content cross fusion. By maintaining content integrity while seamlessly integrating style, StyleMotif ensures more realistic and dynamically consistent stylized motions, which is critical for high-quality motion style transfer.

4.5. Multi-Modal Motion Stylization

With the power of the aligned multi-modal space, our model supports stylization guided by a variety of modalities, including motion, text, image, video, audio. Here, we utilize the input style feature to retrieve the corresponding motion features as the style condition, while the content input is provided as text prompt. The model then generates a stylized motion that incorporates the style from the input modality

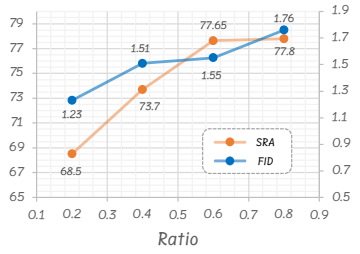

Figure 7. **Ablation Study on Scaling Ratio γ in Eq. 5.** We report both SRA and FID to show the impact of the scaling ratio on both stylization and content preservation.

Style Encoder Pre-training Strategy							
HumanML3D	100STYLE	SRA ↑	FID ↓	MM Dist ↓	R-Precision ↑	Diversity →	Foot Skate Ratio ↓
-	✓	76.73	1.788	4.349	0.571	7.505	0.101
✓	-	76.58	1.635	4.458	0.572	8.534	0.109
✓	✓	**77.65**	**1.551**	**4.354**	**0.586**	7.567	**0.097**

Text in Multi-modal Alignment						
Textual Expression	SRA ↑	FID ↓	MM Dist ↓	R-Precision ↑	Diversity →	Foot Skate Ratio ↓
Brief Description	76.84	1.580	4.378	0.578	7.251	0.099
Detailed Description	75.25	1.622	4.419	0.563	7.764	0.102
Single Text Label	**77.65**	**1.551**	**4.354**	**0.586**	7.567	**0.097**

Table 3. **Ablation Study on Style Encoder Pre-training Strategies and Text Expression for Multi-modal Alignment.** 'w. HumanML3D' and 'w. 100STYLE' denote pre-training with HumanML3D [13] and 100STYLE [37] data respectively.

while maintaining the content integrity as specified by the text prompt. We showcase several examples of multi-modal motion stylization in Figure 6, where different modalities guide the motion generation. For instance, when a text content *"A person is walking."* is provided alongside *an image of a duck* as style input, the model retrieves a relevant motion feature and blends the content with the style of *'Duckfoot'*. As shown, the style from various inputs (text, image, video, audio) is effectively transferred to the generated motion.

4.6. Style Interpolation

Leveraging the aligned multi-modal space, our model enables text-guided style interpolation. Given one content text along with at least two style style texts, our model generates a motion that combines the characteristics of all input styles. When style texts are provided, we retrieve the most similar style motion features from the shared space. These features are then combined by weighted summation. The combined style features are fused with the content features using the proposed style-content cross fusion. In Figure 5, we showcase some qualitative results of style interpolation in, where the generated motions successfully blend the characteristics of both styles while maintaining the content's integrity.

4.7. Ablation Study

Style Encoder Pre-training. In Table 3, we investigate the impact of different pre-training strategies for the style encoder. Our results show that pre-training on both HumanML3D [13] and 100STYLE [37] yields the best performance. Compared to pre-training solely on either one, the combined training on both datasets provides the style encoder with a richer set of prior knowledge. This enables the model to effectively balance content preservation and style reflection, leveraging the diverse characteristics of both datasets to enhance the overall stylization quality.

Cross Normalization Scaling Ratio. In Figure 7, we examine the effect of different scaling ratios γ in Eq. 5 on the stylization performance. As shown, the model achieves the best performance with $\gamma = 0.6$, striking an optimal balance between style reflection and content preservation. The choice of scaling ratio influences both SRA and FID, where the optimal value improves stylization without sacrificing content integrity.

Text Expression for Multi-modal Alignment. In Table 3, we also explored different text representations for contrastive learning alignment. Specifically, we tested single labels (*e.g.*, *"Old"*), brief descriptions (*e.g.*, *"An old person"*), and more detailed descriptions (*e.g.*, *"An old person is moving slow and stiff"*). Our experiments show that using a single label produced the best results, providing a clearer and more concise style signal for the model while avoiding unnecessary complexity in the text representation.

5. Conclusion

In this work, we introduce **StyleMotif**, a novel stylized motion diffusion model capable of generating motion conditioned on both content and style from multiple modalities. Unlike prior approaches that either focus on motion generation across various content types or style transfer between sequences, StyleMotif effectively synthesizes motion while incorporating stylistic cues from *multi-modal* inputs, including text, image, video, and audio. To achieve this, we introduce a *style-content cross fusion* mechanism and align a style encoder with a pre-trained multi-modal model. Through extensive experiments across diverse applications, we demonstrate that StyleMotif outperforms existing methods in stylized motion generation, producing high-quality, realistic results that faithfully adhere to the given style references. Moreover, our model exhibits ***emergent capabilities*** for multi-modal motion stylization, enabling richer and more nuanced motion synthesis. These findings indicate the potential of StyleMotif in advancing stylized motion generation and open new avenues for future research in multi-modal-driven motion synthesis and style-aware generative models.

Acknowledgements

We would like to extend our sincere thanks to the Core AI and Avatar Movement team at Roblox for their guidance and support throughout this project. In particular, we are grateful to Haomiao Jiang, Vincent Petrella, Sean Palmer, and Simone Guggiari for their valuable feedback and insights. We also thank Lei Zhong and Zhe Li for their technical advice.

References

[1] K. Aberman, Y. Weng, D. Lischinski, D. Cohen-Or, and B. Chen. Unpaired motion style transfer from video to animation. *TOG*, 2020. 2, 6

[2] Nefeli Andreou, Xi Wang, Victoria Fernández Abrevaya, Marie-Paule Cani, Yiorgos Chrysanthou, and Vicky Kalogeiton. Lead: Latent realignment for human motion diffusion. *arXiv preprint arXiv:2410.14508*, 2024. 2

[3] Z. Cen, H. Pi, S. Peng, Z. Shen, M. Yang, S. Zhu, H. Bao, and X. Zhou. Generating human motion in 3d scenes from text descriptions. In *CVPR*, 2024. 2

[4] Kexin Chen, Yuyang Du, Tao You, Mobarakol Islam, Ziyu Guo, Yueming Jin, Guangyong Chen, and Pheng-Ann Heng. Llm-assisted multi-teacher continual learning for visual question answering in robotic surgery. *ICRA 2024*, 2024. 3

[5] L.H. Chen, S. Lu, A. Zeng, H. Zhang, B. Wang, R. Zhang, and L. Zhang. Motionllm: Understanding human behaviors from human motions and videos. *ArXiv*, 2024. 2, 3

[6] 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, 3, 4, 5

[7] Sammy Christen, Shreyas Hampali, Fadime Sener, Edoardo Remelli, Tomas Hodan, Eric Sauser, Shugao Ma, and Bugra Tekin. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. In *SIGGRAPH Asia 2024 Conference Papers*, 2024. 2

[8] Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. *arXiv preprint arXiv:2404.19759*, 2024. 2

[9] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11162–11173, 2021. 3

[10] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 3

[11] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 3

[12] 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. 2, 3, 4, 5

[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 *CVPR*, 2022. 2, 4, 5, 6, 7, 8

[14] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *CVPR*, 2024. 2, 3

[15] Chuan Guo, Yuxuan Mu, Xinxin Zuo, Peng Dai, Youliang Yan, Juwei Lu, and Li Cheng. Generative human motion stylization in latent space. *arXiv preprint arXiv:2401.13505*, 2024. 2

[16] Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. Calip: Zero-shot enhancement of clip with parameter-free attention. *AAAI 2023 Oral*, 2022. 3

[17] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023. 3

[18] Zilu Guo, Hongbin Lin, Zhihao Yuan, Chaoda Zheng, Pengshuo Qiu, Dongzhi Jiang, Renrui Zhang, Chun-Mei Feng, and Zhen Li. Pisa: A self-augmented data engine and training strategy for 3d understanding with large models. *arXiv preprint arXiv:2503.10529*, 2025. 3

[19] Ziyu Guo, Ray Zhang, Hao Chen, Jialin Gao, Dongzhi Jiang, Jiaze Wang, and Pheng-Ann Heng. Sciverse: Unveiling the knowledge comprehension and visual reasoning of lmms on multi-modal scientific problems. *arXiv preprint arXiv:2503.10627*, 2025. 3

[20] Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025. 3

[21] Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*, 2023. 3

[22] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *CVPR*, 2023. 2

[23] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 4

[24] Deok-Kyeong Jang, Soomin Park, and Sung-Hee Lee. Motion puzzle: Arbitrary motion style transfer by body part. *ACM TOG*, 2022. 2, 6

[25] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023. 3, 5

[26] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large

multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025. 3

[27] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Gmd: Controllable human motion synthesis via guided diffusion models. In *ICCV*, 2023. 1, 2

[28] Boeun Kim, Jungho Kim, Hyung Jin Chang, and Jin Young Choi. Most: Motion style transformer between diverse action contents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1705–1714, 2024. 2

[29] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024a. *URL https://arxiv. org/abs/2408.03326*. 3

[30] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *ICLR 2025 Spotlight*, 2024. 3

[31] Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18061–18070, 2024. 3

[32] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017. 4

[33] Zhe Li, Yisheng He, Lei Zhong, Weichao Shen, Qi Zuo, Lingteng Qiu, Zilong Dong, Laurence Tianruo Yang, and Weihao Yuan. Mulsmo: Multimodal stylized motion generation by bidirectional control flow. *arXiv preprint arXiv:2412.09901*, 2024. 2

[34] Zhe Li, Weihao Yuan, Yisheng He, Lingteng Qiu, Shenhao Zhu, Xiaodong Gu, Weichao Shen, Yuan Dong, Zilong Dong, and Laurence T Yang. Lamp: Language-motion pretraining for motion generation, retrieval, and captioning. *arXiv preprint arXiv:2410.07093*, 2024. 2

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

[36] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, 2019. 5

[37] 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. 2, 4, 5, 6, 7, 8

[38] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. In *European Conference on Computer Vision*, pages 407–426. Springer, 2022. 3

[39] Soomin Park, Deok-Kyeong Jang, and Sung-Hee Lee. Diverse motion stylization for multiple style domains via spatial-temporal graph-based generative model. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 2021. 2

[40] Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. *arXiv preprint arXiv:2312.06553*, 2023. 2

[41] Mathis Petrovich, Or Litany, Umar Iqbal, Michael J Black, Gul Varol, Xue Bin Peng, and Davis Rempe. Multi-track timeline control for text-driven 3d human motion generation. In *CVPR*, 2024. 2

[42] Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Pu Wang, Minwoo Lee, Srijan Das, and Chen Chen. Bamm: Bidirectional autoregressive motion model. *arXiv preprint arXiv:2403.19435*, 2024.

[43] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *CVPR*, 2024. 2

[44] Ziyun Qian, Zeyu Xiao, Zhenyi Wu, Dingkang Yang, Mingcheng Li, Shunli Wang, Shuaibing Wang, Dongliang Kou, and Lihua Zhang. Smcd: High realism motion style transfer via mamba-based diffusion. *arXiv preprint arXiv:2405.02844*, 2024. 2

[45] Sigal Raab, Inbal Leibovitch, Guy Tevet, Moab Arar, Amit H Bermano, and Daniel Cohen-Or. Single motion diffusion. *arXiv preprint arXiv:2302.05905*, 2023. 2

[46] Sigal Raab, Inbar Gat, Nathan Sala, Guy Tevet, Rotem Shalev-Arkushin, Ohad Fried, Amit H Bermano, and Daniel Cohen-Or. Monkey see, monkey do: Harnessing self-attention in motion diffusion for zero-shot motion transfer. *arXiv preprint arXiv:2406.06508*, 2024. 2

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. 2021. 3

[48] Davis Rempe, Zhengyi Luo, Xue Bin Peng, Ye Yuan, Kris Kitani, Karsten Kreis, Sanja Fidler, and Or Litany. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *CVPR*, 2023. 2

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[50] Wenfeng Song, Xingliang Jin, Shuai Li, Chenglizhao Chen, Aimin Hao, Xia Hou, Ning Li, and Hong Qin. Arbitrary motion style transfer with multi-condition motion latent diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 821–830, 2024. 2

[51] Xiangjun Tang, Linjun Wu, He Wang, Bo Hu, Xu Gong, Yuchen Liao, Songnan Li, Qilong Kou, and Xiaogang Jin. Rsmt: Real-time stylized motion transition for characters. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–10, 2023. 2

[52] Yiwen Tang, Zoey Guo, Zhuhao Wang, Ray Zhang, Qizhi Chen, Junli Liu, Delin Qu, Zhigang Wang, Dong Wang, Xuelong Li, et al. Exploring the potential of encoder-free architectures in 3d lmms. *arXiv preprint arXiv:2502.09620*, 2025. 3

[53] Tianxin Tao, Xiaohang Zhan, Zhongquan Chen, and Michiel van de Panne. Style-erd: Responsive and coherent online motion style transfer. 2022. 2

[54] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023. 1, 2

[55] Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. Tlcontrol: Trajectory and language control for human motion synthesis. *arXiv preprint arXiv:2311.17135*, 2023. 2

[56] Yu-Hui Wen, Zhipeng Yang, Hongbo Fu, Lin Gao, Yanan Sun, and Yong-Jin Liu. Autoregressive stylized motion synthesis with generative flow. In *CVPR*, 2021. 2

[57] Qianyang Wu, Ye Shi, Xiaoshui Huang, Jingyi Yu, Lan Xu, and Jingya Wang. Thor: Text to human-object interaction diffusion via relation intervention. *arXiv preprint arXiv:2403.11208*, 2024. 2

[58] Qi Wu, Yubo Zhao, Yifan Wang, Yu-Wing Tai, and Chi-Keung Tang. Motionllm: Multimodal motion-language learning with large language models. *arXiv preprint arXiv:2405.17013*, 2024. 2

[59] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. In *ICLR*, 2024. 2

[60] Jingwei Xu, Huazhe Xu, Bingbing Ni, Xiaokang Yang, Xiaolong Wang, and Trevor Darrell. Hierarchical style-based networks for motion synthesis. In *ECCV*, 2020. 2

[61] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, 2023. 2

[62] Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. Interdreamer: Zero-shot text to 3d dynamic human-object interaction. *arXiv preprint arXiv:2403.19652*, 2024. 2

[63] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1179–1189, 2023. 5

[64] Senqiao Yang, Jiaming Liu, Ray Zhang, Mingjie Pan, Ziyu Guo, Xiaoqi Li, Zehui Chen, Peng Gao, Yandong Guo, and Shanghang Zhang. Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. *AAAI 2025*, 2023. 3

[65] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *ICCV*, 2023. 2

[66] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2, 3, 4

[67] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *PAMI*, 2024. 2

[68] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *CVPR 2022*, 2022. 3

[69] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *ECCV 2022*. Springer Nature Switzerland, 2022. 3

[70] Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. Llama-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *ICLR 2024*, 2024. 3

[71] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *ECCV 2024*, 2024. 3

[72] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, et al. Mavis: Mathematical visual instruction tuning with an automatic data engine. *arXiv preprint arXiv:2407.08739*, 2024. 3

[73] Lei Zhong, Yiming Xie, Varun Jampani, Deqing Sun, and Huaizu Jiang. Smoodi: Stylized motion diffusion model. In *ECCV*, 2024. 2, 3, 5, 6

[74] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023. 3

[75] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. *ICCV 2023*, 2023. 3