ClusterStyle: Modeling Intra-Style Diversity with Prototypical Clustering for Stylized Motion Generation

Kerui Chen[1*], Jianrong Zhang[1*], Ming Li[2], Zhonglong Zheng[3], and Hehe Fan[1†]

[1] CCAI, Zhejiang University, China
[2] School of Artificial Intelligence, The Chinese University of Hong Kong, Shenzhen, China
[3] Zhejiang Normal University, China

Abstract. Existing stylized motion generation models have shown their remarkable ability to understand specific style information from the style motion, and insert it into the content motion. However, capturing intra-style diversity, where a single style should correspond to diverse motion variations, remains a significant challenge. In this paper, we propose a clustering-based framework, **ClusterStyle**, to address this limitation. Instead of learning an unstructured embedding from each style motion, we leverage a set of prototypes to effectively model diverse style patterns across motions belonging to the same style category. We consider two types of style diversity: global-level diversity among style motions of the same category, and local-level diversity within the temporal dynamics of motion sequences. These components jointly shape two structured style embedding spaces, *i.e.*, global and local, optimized via alignment with non-learnable prototype anchors. Furthermore, we augment the pretrained text-to-motion generation model with the Stylistic Modulation Adapter (SMA) to integrate the style features. Extensive experiments demonstrate that our approach outperforms existing state-of-the-art models in stylized motion generation and motion style transfer. Project page: `https://1233chen.github.io/ClusterStyle/`.

Keywords: Motion generation · Style transfer · Diffusion model

1 Introduction

Stylization is an essential technique for generative modeling [20, 75], enabling the rendering of source content in a target style. In the stylized motion generation [46, 62, 63], the primary aim is to transfer the style from a reference motion sequence to a source motion sequence while preserving the original content, which makes it useful in various real-world applications, including augmented reality [63], animation [1], and virtual avatars [3].

† Corresponding author.
* Equal Contribution.

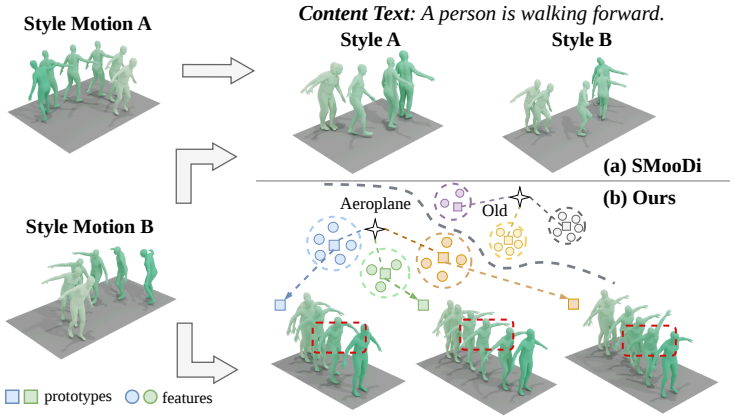

Fig. 1: Comparison between SMooDi [86] and ClusterStyle. Given two style motions with the same style but differing in expression, along with a content text, (a) SMooDi generates similar stylized motions. Meanwhile, the generated motions still reflect elements from the style motions' content, leading to inconsistency with the intended content text. (b) In contrast, our method uses different **prototypes** of the areoplane style, generating diverse motion with varying motion magnitudes (highlighted in red), while maintaining stronger alignment with the content semantics.

Motion style is inherently expressive [4,87], where the same style can exhibit diverse motion variations influenced by the human's mood, intention, or context. For example, given a textual description "a person is walking forward", the "aeroplane" style should encompass a wide range of postures as well as variations in action amplitude, rather than a single rigid walk with both arms raised.

Recently, many works aim to achieve stylized motion generation and motion style transfer using diffusion models [27, 31, 40, 55, 57, 60, 61, 86]. Song et al. [61] leveraged trajectory as an auxiliary condition during the denoising process to improve the preservation of the content. SMooDi [86] introduces a large-scale dataset for stylized motion generation from both the textual description and style motion. Building on the pretrained text-to-motion model, *i.e.*, MLD [11], it uses a ControlNet-like architecture to serve as the style encoder. BiFlow [40] further replaces the style-to-content flow with a bidirectional control flow to reduce conflicts between content and style.

However, these methods embed the style motion into an *unstructured* feature, which often cannot model the intra-style diversity, resulting in relatively uniform motion output. Meanwhile, we find that such unstructured features include both content and style information of the style motion, and the content part may lead to deviations from the target content intended by the text. For example, in Fig. 1(a), given two style motions with the same style but to different extents, and a text prompt "a person is walking forward", SMooDi generates relatively uniform motions, and does not reflect "walking forward".

In this paper, we propose ClusterStyle to address these limitations. It is inspired by recent clustering-based representation learning methods [42, 88] in computer vision, where cluster centers are progressively updated through inter-

actions with pixel features. Specifically, each style category is represented by clustering its style motions into K non-learnable prototypes (*i.e.*, distinct centroids). Then, we build a transformer-based architecture as the style motion encoder, and the extracted features are used to update the prototypes iteratively during training. To facilitate the learning of a high-quality and disentangled style feature space, we encourage the contrastive property across prototypes belonging to different categories (Prototype-Based Inter-Style Learning) as well as among distinct prototypes within the same category (Prototype-Based Intra-Style Learning). Furthermore, we observe that a single motion sequence may also exhibit stylistic differences over time. Thus, we propose hierarchical clustering, which performs clustering on the entire motion sequences (*global-level*) as well as their temporal segments (*local-level*). Finally, we present a Stylistic Modulation Adapter (SMA) to effectively integrate the style information into the text-to-motion model.

Our approach characterizes two appealing advantages. First, we explicitly model the intra-style diversity through learning multiple clustering centroids on both global and local levels of motion sequences. As shown in Fig. 1(b), using different cluster centroids (prototypes), our method can generate expressions of the same style to varying extents, making ClusterStyle a flexible and *interpretable* framework. Second, prototypes focus more on style attributes, enabling a better consistency between the motion and the target content described by the text. We evaluate the proposed approach on the 100STYLE [86] and HumanML3D [25] datasets across two tasks, *i.e.* stylized motion generation and motion style transfer. Extensive experiments demonstrate that ClusterStyle outperforms current state-of-the-art methods in fidelity and content preservation.

Our contributions can be summarized as follows: **1.** We propose ClusterStyle, a clustering-based framework that facilitates diverse stylized motion generation. **2.** We propose to use multiple clustering-based prototypes to model the intra-style diversity and consider both global-level and local-level discrepancies. **3.** We introduce Stylistic Modulation Adapter (SMA) for style attribute injection.

2 Related Work

Text-to-Motion Generation. The task of generating human motion from textual descriptions has seen rapid advancement in recent years. Early studies primarily focus on constructing a shared embedding space for text and motion [2, 21, 50–52, 64]. Further advancements introduce the idea of discretizing the motion representation through vector quantization [26]. For instance, T2M-GPT [80] presents VQ-VAE and GPT along with some training recipes, *e.g.*, corruption strategy, which significantly improve the performance on the large-scale dataset, *i.e.*, HumanML3D [25]. Based on this line, several works improve this paradigm by enhancing the VQ-VAE [23, 85] and leveraging masked token modeling [23, 53, 54].

Recently, diffusion models have become an increasingly popular choice for motion generation. The foundation of the diffusion model in this domain is laid

by MDM [65] and MotionDiffuse [82]. They both use the Transformer-based denoising network to progressively refine random noises to synthesize human motions guided by textual descriptions. Subsequent works [13, 22, 34, 68, 69, 72] introduce diverse additional information to improve the quality or controllability of motion generation, such as spatial constraints [36, 66, 71], physical constraints [76] and multimodal information [8, 9, 28, 83, 89]. MLD [11] proposes to shift the diffusion process to the latent space. The following efforts have been dedicated to improving performance [19,77] and sampling efficiency [84] through architectural refinement. Motion latent diffusion model is more relevant to our approach, as the current stylized motion generation methods are built upon MLD [11]. We augment this architecture with our designed Stylistic Modulation Adapter (SMA) to effectively fuse text and style.

Stylized Motion Generation. Stylized motion generation aims to synthesize human motions that reflect a desired content and style, which can be specified through text [37, 60] or reference motion [29, 30, 39, 46, 49, 55, 62, 63, 70, 73]. Earlier approaches [1, 33] primarily focused on motion style transfer and employed autoencoder-based architectures with AdaIN [32] to disentangle and recombine content and style features. Following this, Guo [24] explored style transfer in a probabilistic latent space, while MOST [38] improves the recombination technique between the content and style. Recently, diffusion-based methods [31, 57, 61] have increasingly been adopted in this task, primarily leveraging the prior knowledge from pre-trained text-to-motion models. For instance, Hu et al. [31] treat the denoising process as a style transfer stage and utilize CLIP [58] for transfer guidance, and MoMo [57] achieves zero-shot style transfer by exploring the attention mechanism.

To facilitate research in this area, SMooDi [86] proposes a new dataset named 100STYLE, which is paired with content text and style motion for a style-specific task. In addition to introducing a new dataset, SMooDi proposed a ControlNet-like architecture to incorporate style features into the generation process. To mitigate conflicts between content and style, BiFlow [40] refines the style-to-content flow in the SMooDi into a bidirectional control flow. StyleMotif [27] leverages multi-modal style inputs (e.g., text, audio, motion) and proposes a style-content cross-fusion mechanism to effectively integrate this information.

However, these methods all neglect the importance of intra-style diversity, leading to the lack of diversity and distinctiveness in the generated results when guided by different style motions. In this paper, we represent each style category with multiple prototypes, obtained by clustering style embeddings from the full style motion dataset, which are used to model the intra-diversity of style.

Clustering in Vision. Clustering enables models to automatically mine underlying patterns and learn meaningful representations from data. Early clustering methods [18,35,59] primarily relied on raw data representations and hand-crafted priors, which were not effective when faced with complex data. Some methods [5, 6, 14, 74] propose to combine deep learning techniques with prototype-based clustering. For example, Wang et al. [67] considered the representation learning from the clustering perspective, constructing multiple cluster centers

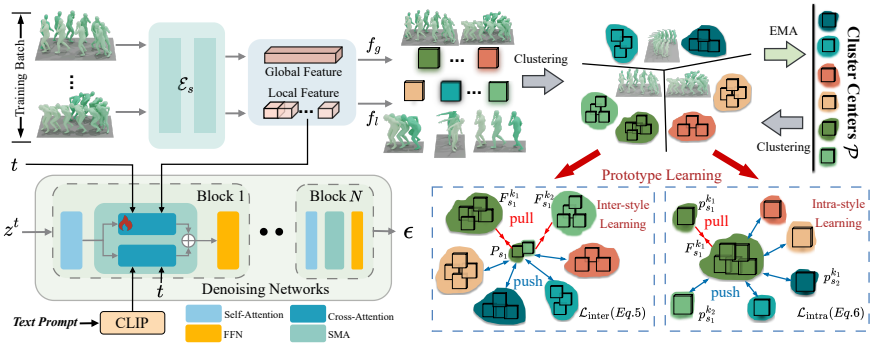

Fig. 2: The overview of the ClusterStyle. Our method consists of a style encoder and a motion latent diffusion model. In the style encoder, we present a cluster-based prototype learning paradigm that represents each style category using a set of non-learnable prototypes (cluster centers) to model the intra-style diversity explicitly. Then, two contrastive losses are proposed for prototype-based intra-style learning and inter-style learning, respectively. To incorporate the style embedding into the diffusion process, we introduce a Style Modulation Adapter (SMA), enabling effective guidance of stylized motion generation.

for each class to improve the recognition ability of the model. This clustering paradigm has inspired a broad spectrum of research across domains, such as image segmentation [7, 15, 41, 42, 48, 79, 88], point cloud analysis [16, 17, 43, 44], and zero-shot learning [45, 56].

Inspired by these methods, we extend the clustering technique to stylized motion generation. Different from previous works that focus on global information of each data instance (*i.e.*, image), ClusterStyle considers both long-term and short-term relationships within style motion sequences, facilitating a more fine-grained understanding of underlying style attributes and expression patterns.

3 Method

In this paper, we aim to model intra-style diversity for stylized 3D motion generation. As illustrated in Fig. 2, we introduce ClusterStyle, a novel framework that comprises a motion latent diffusion model and a style encoder with a cluster-based prototype learning strategy. Sec. 3.1 outlines the problem formulation and provides an overview of our approach. We introduce cluster-based style prototype learning, hierarchical style modeling, and style modulation adapter in Sec. 3.2, Sec. 3.3, and Sec. 3.4, respectively. Finally, implementation details are provided in Sec. 3.5.

3.1 Problem Setting and Overview

Given a style motion sequence $\boldsymbol{X}_s = [\boldsymbol{x}_s^1; \boldsymbol{x}_s^2; \cdots; \boldsymbol{x}_s^{L_s}]$ and a content text c, where $\boldsymbol{x}_s^i \in \mathbb{R}^d$, L_s and d are the length and dimension of style motion. Our goal

is to generate a motion sequence $\boldsymbol{X}_c = [\boldsymbol{x}_c^1; \boldsymbol{x}_c^2; \cdots; \boldsymbol{x}_c^{L_c}]$ with length L_c that is consistent with the text while adhering to the style demonstrated by the style motion $\boldsymbol{X}_s$.

Specifically, following previous works [27, 40, 86], we build our approach on a motion latent diffusion model [11], which consists of a VAE and latent diffusion model, pretrained on the HumanML3D dataset [25]. During training, with the style motion-text pair $\{\boldsymbol{X}_s, c_s\}$ available, we first feed the style motion into the VAE encoder to obtain the latent feature z_s. We gradually add Gaussian noise ϵ to z_s over T times via $z_s^t = \sqrt{\alpha^t} z_s^0 + \sqrt{1-\alpha^t}\epsilon$, where t is the timestep and $\{\alpha^t\}_{t=0}^T$ is the noise variance schedule. Then, we propose a clustering-based style encoder which uses a set of prototypes to represent a single style class for intra-style diversity (Sec. 3.2). It also captures style information hierarchically by modeling both global-level motion and local-level temporal segments (Sec. 3.3), obtaining the style embedding f_s. Then, we train a denoising autoencoder to predict the noise conditioned on f_s and c_s, the loss function can be formulated as:

$$\mathcal{L}_{\text{diff}}^s = \mathbb{E}_{z_s, \epsilon, t, c_s, f_s}\Big[\|\epsilon - \epsilon_\theta(z_s^t, t, c_s, f_s))\|_2^2\Big]. \tag{1}$$

Furthermore, to prevent forgetting content-related prior knowledge, we also train the diffusion model on the content motion-text pair $\{\boldsymbol{X}_c, c\}$ from the HumanML3D dataset,

$$\mathcal{L}_{\text{diff}}^c = \mathbb{E}_{z_c, \epsilon, t, c, f_c}\Big[\|\epsilon - \epsilon_\theta(z_c^t, t, c, f_c))\|_2^2\Big], \tag{2}$$

where z_c and f_c denote the motion latent features from the VAE encoder and the style encoder, respectively. Different from existing methods [86] that employ a ControlNet-like [81] architecture for style feature fusion, we utilize a cross-attention to understand the textual description [77, 78] and further propose a Style Modulation Adapter (Sec. 3.4) for style feature integration.

Combining these components, we optimize ClusterStyle using the following overall training objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diff}}^s + \mathcal{L}_{\text{diff}}^c + \lambda_{\text{style}}\mathcal{L}_{\text{style}}, \tag{3}$$

where λ is a hyper-parameter to balance the weight of $\mathcal{L}_{\text{style}}$, which supervises the training of the style encoder (see Sec. 3.3 for more details).

3.2 Clustering-based Style Prototype Learning

The design of ClusterStyle is motivated by a key insight: motion style is inherently diverse, a single style should not be confined to a fixed form, but can manifest through multiple dynamic patterns. To achieve this, we raise two questions: ❶ *How to model the intra-style diversity?* ❷ *How can intra-style diversity be automatically mined without relying on manual annotation?* To answer these questions, we propose a clustering-based framework for style prototype learning, which enables the discovery of underlying sub-style patterns within each style category.

Cluster Center Initalization. As a response to question ❶, we propose to represent each style category using multiple prototypes. Each prototype captures a distinct sub-style pattern and acts as a reference point that encourages variation within the same style. This helps organize the style features into a well-structured and diverse style space. Specifically, for a style category $\boldsymbol{s}$, we assume it contains K_g sub-style patterns and define a cluster using a set of prototypes (*a.k.a.* cluster centers) $P_s = \{p_s^i\}_{i=1}^{K_g}$, where $p_s^i \in \mathbb{R}^{1\times d'}$ and d' is the dimension of the prototype. We encode the style motion feature using a transformer-based style encoder $\mathcal{E}_s$, which can be computed as $f_g = \mathcal{E}_s(\boldsymbol{X}_s)$. Our goal is to map the style feature f_g to the closest prototype within P_s, which represents the corresponding style cluster. To extend this idea, the prototype clusters for all style categories are collectively defined as $\mathcal{P} = \{P_s\}_{s=1}^{S}$, where S is the total number of style categories.

Prototype Assignment. It is challenging to directly supervise the assignment of features to prototypes due to the lack of manual annotations that specify or differentiate these prototypes. Driven by question ❷, we formulate this process as an unsupervised optimal transport problem. Formally, given a series of style motion features $F_s = [f_{g,s}^1, f_{g,s}^2, \cdots, f_{g,s}^{N_s}] \in \mathbb{R}^{d'\times N_s}$, where N_s denotes the total motion numbers of the style category s. Correspondingly, we define a prototype matrix $P_s' \in \mathbb{R}^{d'\times K_g}$ representing K_g style-specific prototypes. Note that each column of F_s and P_s' is L2-normalized. Based on a binary assignment matrix $L_s \in \{0,1\}^{K_g\times N_s}$, we compute the prototype assignment map as $A_s = P_s' L_s \in \mathbb{R}^{d'\times N_s}$, where each column of A_s represents the aggregated prototype assigned to the corresponding style motion feature. We measure the inner product similarity $\langle\cdot,\cdot\rangle_I$ and maximize $\langle A_s, F_s\rangle_I$ to determine L_s. To prevent all style motion features from being assigned to a single prototype, we introduce a balancing constraint that encourages each prototype to be selected approximately $\frac{N_s}{K_g}$ times on average. We adopt the Sinkhorn algorithm [12] to solve this problem, which introduces entropy regularization to enable fast and stable computation of the transport plan:

$$\max_{L_s} \langle A_s, F_s\rangle_I + \mu \ \ \mathrm{KL}\ (L_s || \frac{\mathbf{1}}{K_g N_s} \mathbf{1}_{K_g} \mathbf{1}_{N_s}^\top), \tag{4}$$

where KL is used to smooth the distribution via a hyper-parameter μ. There are two additional constraints imposed on relaxed L_s: (1) $\boldsymbol{L}_s \in \mathbb{R}_+^{K_g\times N_s}$ ensuring that each feature is assigned once; and (2) $\boldsymbol{L}_s \mathbf{1}_{N_s} = \frac{N_s}{K_g}\mathbf{1}_{K_g}$, encouraging balanced usage of prototypes.

Prototype-based Contrastive Learning. After tackling questions ❶ and ❷, the next step aims to learn the high-quality representation of $\mathcal{P}$ and style motions. We propose two prototype-based contrastive losses $\mathcal{L}_{\text{inter}}^g$ and $\mathcal{L}_{\text{intra}}^g$, corresponding to Prototype-based Inter-Style Learning and Prototype-based Intra-style Learning, respectively. Specifically, with the style motion feature f_g, the prototype-based inter-style learning is designed to enhance inter-style discrimination by pulling each motion feature closer to the prototype of its style category and pushing it away from prototypes of all other styles. The inter-style

contrastive loss $\mathcal{L}^g_{\text{inter}}$ is formulated as follows:

$$\mathcal{L}^g_{\text{inter}} = -\log \frac{\exp(-\text{sim}[f_g, s])}{\Sigma^S_{s'=1}\exp(-\text{sim}[f_g, s'])}, \tag{5}$$

where $\text{sim}[f_g, s] = \min\{\cos(f_g, p^k_s)\}^{K_g}_{k=1}$ is a function to measure the similarity between the style feature f_g and P_s, $\cos(\cdot)$ denotes the cosine similarity. We provide an ablation study in the Supplement to evaluate the impact of different similarity metrics on model performance.

In prototype-based intra-style learning, each motion feature is contrasted against all prototypes by treating the assigned prototype as positive, and all others, even within the same style category, as negatives. Given the style motion feature f_g, the corresponding prototype p^k_s, we define the negative prototypes as $\hat{P}$, which consists of all prototypes in $\mathcal{P}$ except the assigned p^k_s. Then, the intra-style contrastive loss $\mathcal{L}^g_{\text{intra}}$ can be written as:

$$\mathcal{L}^g_{\text{intra}} = -\log \frac{\exp(\cos(f_g, p^k_s)/\tau)}{\exp(\cos(f_g, p^k_s)/\tau) + \Sigma_{\hat{p}\in\hat{P}}\beta\, \exp(\cos(f_g, \hat{p})/\tau)}, \tag{6}$$

where $\beta = 1$ if $\hat{p} \not\sim f_g$, else 5, $\not\sim$ indicates that $\hat{p}$ and f_g are from different style categories, τ is a temperature hyper-parameter which is set to 0.05 following [10]. Then, the final training objective is defined as follows:

$$\mathcal{L}^g_{\text{style}} = \mathcal{L}^g_{\text{inter}} + \mathcal{L}^g_{\text{intra}}. \tag{7}$$

By doing so, the proposed two loss functions encourage the model to focus more on style-relevant patterns and to learn style representations that are disentangled from content information, thereby preventing the target content from being influenced by content information embedded in the style motion.

Prototype Update. Unlike traditional classifiers optimized by gradient descent, our prototypes are non-parametric and non-learnable statistics computed as the centroids of their assigned style-feature sets. Concretely, for category s and prototype k, let $\hat{f}^k_{g,s}$ denote the mean of features assigned to that prototype. At each training iteration, we update the prototype p^k_s by EMA via below formular:

$$p^k_s \leftarrow \lambda_p p^k_s + (1-\lambda_p)\hat{f}^k_{g,s}, \tag{8}$$

where $\lambda_p \in [0, 1]$ is a momentum coefficient.

Prototype-based Guidance. The prototypes can serve as guiding signals to generate diverse stylized motion. Specifically, we define the prototype-based guidance function

$$\epsilon_\theta(z^t, t, c, s) = \epsilon_\theta(z^t, t, c, s) + \gamma_g \nabla_{z^t} G_g(z^t, t, p^k_s), \tag{9}$$

where $G_p(z^t, t, p^k_s) = 1 - \cos(f^{x^0}_g, p^k_s)$, γ_g is the guidance weight. During the inference, z^t is iteratively optimized to approach the sub-style pattern associated with the target prototype p^k_s. Here, z^t is the noisy latent at the timestep t, and we transform it to z^0 using the predict noise, which is then decoded into x^0 through $\mathcal{D}(z^0)$. $f^{x^0}_g$ is the style feature of x^0, and $\cos(\cdot)$ denotes the cosine similarity. Please find more details in Supplement.

3.3 Hierarchical Style Modeling

It is worth noting that motion inherently exhibits temporal dynamics where a single motion sequence may involve stylistic variations over time. To capture a fine-grained understanding of style patterns within a motion sequence, we propose a hierarchical style modeling framework that independently clusters features at both the global level (entire motion sequences) and the local level (temporal segments). Specifically, as we mentioned in Sec. 3.2, we encode the style motion sequence X_s into $f_m \in \mathbb{R}^{L_s \times d'}$, the global style embedding f_g is processed by an average pooling operation over the temporal dimension. For the local style embedding, we first divide f_m into L_w non-overlapping temporal segments with a window size of w, where $L_s = L_w \times w$. Then we apply the same pooling operation within each segment, obtaining $f_l \in \mathbb{R}^{L_w \times d'}$. We perform an identical clustering-based learning that clusters the local-level features into K_l local prototypes. A prototype-based contrastive loss, $\mathcal{L}^l_{style}$, is then computed to guide the learning of fine-grained style representations. The final style loss $\mathcal{L}_{style}$ is formulated as follows:

$$\mathcal{L}_{\text{style}} = \mathcal{L}^g_{\text{style}} + \mathcal{L}^l_{\text{style}}. \tag{10}$$

Finally, the overall style representation $f_s \in \mathbb{R}^{(L_w+1)\times d'}$, which encapsulates both hierarchical information and intra-style diversity, is constructed by concatenating the global-level feature f_g and local-level features f_l, and is subsequently used to control the style in the generation process. Note that the guidance mechanism (Sec. 3.2, Eq. (9)) used in the global prototype approach can also be applied to local prototypes.

3.4 Style Modulation Adapter

With the style representation f_s available, we design a Style Modulation Adapter (SMA) to effectively integrate f_s into the pre-trained text-to-motion model, while preserving its prior knowledge. As shown in Fig. 2, the proposed SMA takes both text embedding c and style representation f_s as conditions. It employs two cross-attention modules for the content branch and style branch, respectively, enabling separate control over semantic content and stylistic rendering. Specifically, in the content branch, the text embedding is first computed with $O_{\text{content}} = \text{Softmax}(\frac{QK^\top}{\sqrt{d}})V$ to obtain the content feature, where $Q = XW^q_c$, $K = \boldsymbol{c}W^k_c$, $V = \boldsymbol{c}W^v_c$ with $W^{q,k,v}_c \in \mathbb{R}^{d\times d}$. As for the style branch, we use another cross-attention to model the interaction between f_s and the denoised latent motion z, which can be calculated as $O_{\text{style}} = \text{Softmax}(\frac{Q(K')^\top}{\sqrt{d}})V'$, where $K' = f_sW^k_s$, $V' = f_sW^v_s$ with $W^{k,v}_s \in \mathbb{R}^{d'\times d}$. We then combine O_{content} and O_{style} together, leading to the output O_{SMA}

$$O_{\text{SMA}} = O_{\text{content}} + \lambda O_{\text{style}}, \tag{11}$$

where λ is a trainable parameter that modulates the balance between content and style.

Table 1: Quantitative comparisons of ClusterStyle with existing state-of-the-art methods on the stylized motion generation task. The best and second best results are **bold** and underlined.

Methods	Venue	FID ↓	FSR ↓	MM Dist ↓	R-Precision (Top-3) ↑	Diversity →	SRA ↑
Motion Puzzle [33]	TOG 2022	6.127	0.185	6.467	0.290	6.576	63.769
Aberman [1]	TOG 2020	3.309	0.347	5.983	0.406	8.816	54.367
ChatGPT+MLD	-	0.614	0.131	4.313	0.605	8.836	4.819
SMooDi [86]	ECCV 2024	1.609	0.124	4.477	0.571	9.235	72.418
BiFlow [40]	ARXIV 2025	1.527	0.118	4.292	0.613	9.303	77.042
StyleMotif [27]	ARXIV 2025	1.551	**0.097**	4.354	0.586	7.567	77.650
ClusterStyle (Ours)	-	**1.137**	0.113	**3.610**	**0.708**	8.719	**78.101**

3.5 Implementation Details

ClusterStyle adopts a frozen CLIP ViT-L/14 model as the text encoder. For the style encoder, we use 6 transformer layers with a dimension of 512. We set the global prototype number K_g to 3, local prototype number K_l to 30, and momentum coefficient λ_p to 0.95. Following [77], we use a pretrained VAE with 5 latent vectors. The denoising autoencoder consists of $N = 9$ layers of transformer blocks with a dimension of $d = 256$. During training, ClusterStyle is optimized using the AdamW optimizer with a batch size of 128, and the learning rate is set to 2e-5 with a linear warm-up period of 500 iterations, followed by 3000 iterations of training with a constant learning rate. We first train the style encoder for 1800 iterations, after which prototype updates are frozen during the remaining training process. Training our model takes about 1 hour on a single NVIDIA A6000 GPU. Note that only $W_s^{k,v}$ in the Style Modulation Adapter (SMA) and parameters in the style encoder are updated. During inference, we generate stylized motions over 50 steps using the DDIM strategy. Following SMooDi [86], we use classifier-free guidance for content texts and classifier guidance for style motions. Please find more information corresponding to the inference in the supplementary material.

4 Experiment

4.1 Experimental Settings

We evaluate our approach on two tasks: stylized motion generation (Sec. 4.2) and motion style transfer (Sec. 4.3). We compare our model with five representative state-of-the-art methods: Aberman [1], Motion Puzzle [33], SMooDi [86], BiFlow [40], and StyleMotif [27]. Meanwhile, we also offer detailed analysis and discussion in Sec. 4.4. Please note that more information about datasets is provided in the supplementary material.

Datasets. We conduct experiments using a single model on two datasets, *i.e.*, HumanML3D [25] and 100STYLE [86]. HumanML3D [25] is used to preserve content-related prior knowledge. 100STYLE [86] is used for providing a wide range of motion styles to guide style-specific learning. During evaluation, content

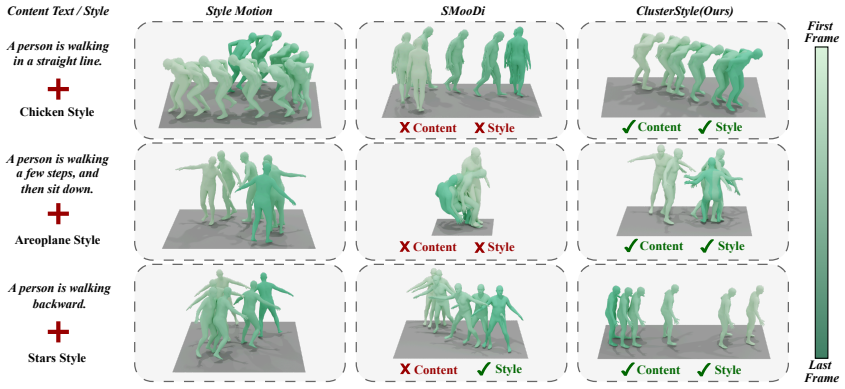

Fig. 3: Qualitative results of stylized motion generation. We compare our method with SMooDi under various text prompts and style motion inputs. Our approach demonstrates better content alignment and achieves more accurate and expressive style rendering. For example, the results of SMooDi are inconsistent with the motion trajectories (*e.g.*, "backward", "straight") and action (e.g., "walk") described in the content. More visual comparisons can be found in the Supplement.

is derived from HumanML3D, while style references are taken from 100STYLE. Please see Supplement for more information.

Evaluation Metrics. Following SMooDi [86], we adopt standard evaluation metrics to comprehensively assess the quality of stylized motion. These metrics include: (1) R-Precision and Multi-modal Distance (MM-Dist) for content preservation; (2) Style Recognition Accuracy (SRA) for style fidelity; (3) Fréchet Inception Distance (FID) for motion quality; (4) *Diversity* for motion diversity; (5) Foot Skating Ratio (FSR) for physical plausibility.

4.2 Stylized Motion Generation

Quantitative Results. Tab. 1 shows quantitative results on the HumanML3D and 100STYLE datasets. We compare our method with current state-of-the-art methods, *e.g.*, SMooDi [86], BiFlow [40], and StyleMotif [27]. Our method achieves R-Precision of 0.708 and SRA of 78.101, surpassing the BiFlow by 15.5% and 1.1%. In terms of FID, our method achieves a score of 1.137, representing a 27% improvement over StyleMotif and demonstrating significantly improved generation fidelity. These results show that the motions generated by our approach are semantically aligned with the content text and stylistically consistent with the reference style motion.

Qualitative Results. In Fig. 3, we present visual comparisons between our model and baseline approach SMooDi [86]. It can be seen that SMooDi struggles to retain essential motion details (*e.g.*, “circular” or “walking” trajectories) and lacks fidelity in expressing target styles such as “aeroplane” or “chicken”. Our approach is able to generate motions that better reflect the semantics of the text prompt, while also transferring correct stylistic characteristics from the

Table 2: Quantitative comparison with the state-of-the-art methods.

Method	FSR ↓	FID ↓	SRA ↑
Motion Puzzle [33]	0.197	6.871	67.233
Aberman [1]	0.338	3.892	61.006
SMooDi [86]	0.095	1.582	65.147
BiFlow [40]	<u>0.087</u>	1.566	<u>70.238</u>
StyleMotif [27]	0.094	<u>1.375</u>	68.810
ClusterStyle (Ours)	**0.078**	**0.768**	**73.849**

Table 3: Ablation studies of different loss functions for the style encoder.

$\mathcal{L}_{\text{style}}$	FID ↓	R-Precision ↑ (Top-3)	Diversity →	SRA ↑
$\mathcal{L}_{\text{inter}}$	1.112	0.712	8.479	75.182
$\mathcal{L}_{\text{intra}}$	1.349	0.673	8.509	67.746
$\mathcal{L}_{\text{entropy}}$	1.331	0.677	8.886	65.465
$\mathcal{L}_{\text{inter}} + \mathcal{L}_{\text{intra}}$	1.137	0.708	8.719	78.101

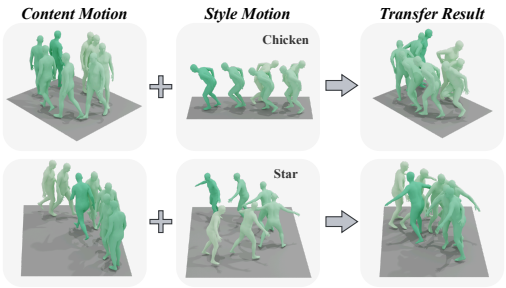

Fig. 4: Qualitative results of motion style transfer. Our approach effectively transfers the target motion style, such as 'Chicken' or 'Star', onto the original motion, preserving its structure while adapting its stylistic characteristics.

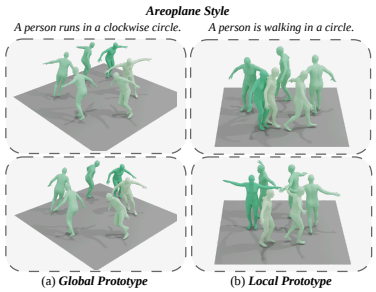

Fig. 5: Visualization of prototype guiding. We visualize how global and local prototypes guide the stylization process for diverse generation results under the 'Aeroplane' style.

reference motion. These observations demonstrate the superior performance of our model, as well as its effectiveness in disentangling motion content and style. Please find more visual results in the supplementary material.

4.3 Motion Style Transfer

Our method utilizes SD-Edit [47] to achieve motion style transfer, relying solely on the pre-trained ClusterStyle model without additional fine-tuning. More implementation details and visualization results are available in the Supplement.

Quantitative Results. Tab. 2 presents the quantitative results, comparing our method against current state-of-the-art methods. Our method achieves the lowest Foot Skating Ratio (0.078), reflecting a 17% reduction compared to the best-performing method, *i.e.*, StyleMotif, and indicating improved physical realism. For FID and SRA, our methods achieve scores of 0.768 and 73.849, a 44% and 5.03% improvement over StyleMotif, demonstrating superior motion fidelity and transfer accuracy.

Qualitative Results. Fig. 4 shows the visual results of ClusterStyle on motion style transfer. Given styles of "chicken" and "star", we observe that our method effectively preserves the action and trajectory of the content motion, while integrating styles consistent with style motions.

Table 4: Ablation studies of key components.

K_g	K_l	FID ↓	MM Dist ↓	R-Precision ↑ (Top-3)	Diversity →	SRA ↑
1	1	1.443	3.886	0.668	8.136	74.785
1	30	1.143	3.613	0.701	8.483	76.226
3	30	**1.137**	**3.610**	**0.708**	8.719	78.101
5	30	1.170	3.641	0.701	8.366	77.883
10	30	1.198	3.687	0.697	8.183	**78.528**
3	1	1.313	3.769	0.685	8.178	**78.744**
3	5	1.226	3.690	0.695	8.461	78.399
3	10	1.185	3.637	0.701	8.661	77.281
3	30	1.137	3.610	0.708	8.719	78.101
3	50	**1.092**	**3.607**	**0.711**	8.516	77.302

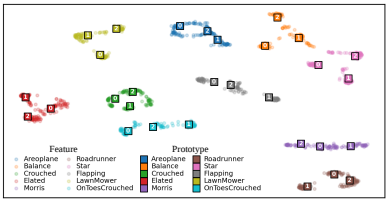

Fig. 6: (Please Zoom in for details.) Visualization of style feature space.

4.4 Discussion

Style Diversity. We explore the capability of our method to generate diverse motions from a single style. In our experiments, we use different global prototypes as global style features to generate stylized motions, as shown in the left column of Fig. 5 and the bottom column of Fig. 1. Furthermore, we randomly combine local prototypes to generate stylized motions, as shown in the right column of Fig. 5. We find that both different global prototypes and local prototypes yield stylistic results that vary in the extent of spread arms. Please note that, at inference time, our method enables diverse motion generation by integrating the style features obtained from the style encoder with different configurations of global and local prototypes.

We also visualize the style feature space in Fig. 6 using t-SNE embeddings. We can observe that prototypes emerge as multiple distinct modes within each style. These diverse prototypes capture the intra-diversity and facilitate the generation of varied stylistic motions. Meanwhile, the visualization also reveals clear inter-style separation, ensuring that our model possesses the discriminative capability to handle similar style motions (*e.g.*, *Star* and *Aeroplane*). To further validate this discriminative power, we provide an additional visual comparison in Fig. 7. The primary distinction between *Star* and *Aeroplane* is the open-leg stance. When conditioned on a *Star* reference motion, SMooDi struggles to differentiate the two and produces Aeroplane-like results. In contrast, our model accurately identifies the subtle discrepancies, consistently generating motions that remain faithful to the target style. More results and analysis are provided in the supplementary material.

Investigating Key Components. We first investigate the effectiveness of the clustering strategy and prototype number (*i.e.*, K_g and K_l) on the task of stylized motion generation, and the results are shown in Tab. 4.

Clustering Strategy: We replace the clustering procedure with a learnable classification head, and the style loss $\mathcal{L}_{\text{style}}$ is replaced by a conventional cross-entropy loss. As indicated in the first row of Tab. 4, this modification leads to a 0.3 reduction in FID, a 5.6% decrease in R-Precision, and a 3.32% drop in SRA. This shows the effectiveness of the proposed cluster-based framework.

Global Prototype Number K_g: The middle four rows of Tab. 4 illustrate the impact of varying the number of global prototypes. K_g =3 yields relatively better

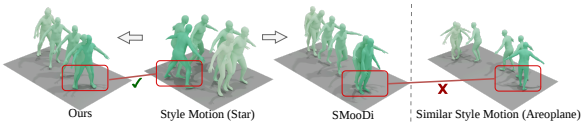

Fig. 7: Qualitative comparison on easily confused Star style motion.

Table 5: Ablation study of different similarity metrics.

sim	FID ↓	FSR ↓	R-Precision ↑ (Top-3)	Diversity →	SRA ↑
L_2	1.191	0.114	0.706	8.553	76.483
L_1	1.322	0.119	0.689	8.936	76.802
cos	1.137	0.113	0.708	8.719	78.101

results compared to other configurations. We observe a trade-off: as the number of global prototypes increases, content consistency metrics (e.g., R-Precision) tend to decline, whereas SRA improves. This suggests that while diversity benefits from a larger set of prototypes, content alignment may be negatively affected.

Local Prototype Number K_l: The last five rows of Tab. 4 show the impact of different number of local prototypes. K_l =30 achieves comparable performance to K_l =50, while being more computationally efficient. It can be seen that increasing local prototypes improves content preservation but degrades SRA, showing an opposite trend to global prototypes.

Ablation Study of Training Objective. Tab. 3 presents the impact of different loss functions used in the style encoder. Training with only inter-style $\mathcal{L}_{\text{inter}}$, $\mathcal{L}_{\text{intra}}$ or $\mathcal{L}_{\text{entropy}}$ leads to performance drops of 2.9%, 10.4%, and 12.6% in SRA, respectively. Here, $\mathcal{L}$entropy denotes a variant where learnable prototypes are trained using a standard cross-entropy loss instead of $\mathcal{L}_{\text{intra}}$. Our prototype-based contrastive learning achieves the best performance by combining $\mathcal{L}_{\text{inter}}$ and $\mathcal{L}_{\text{intra}}$ together.

Impact of different similarity metrics. We adopt cosine similarity in the prototype-based contrastive loss. As shown in Tab. 5, both alternatives lead to a decrease in SRA. In particular, adopting L_1 distance further degrades FID and Top-3 accuracy, indicating reduced motion quality and weaker alignment with the textual content.

5 Conclusion

In this paper, we propose ClusterStyle, a clustering-based framework for capturing intra-style diversity in stylized motion generation. Unlike prior methods that learn a single embedding per style, ClusterStyle represents style features using multiple clustering-based prototypes at both global and local levels. To fuse style features into the pretrained diffusion model, we introduce the Stylistic Modulation Adapter (SMA). Our approach achieves superior performance across stylized motion generation and motion style transfer. We conduct extensive ablation studies and provide visual results to validate the effectiveness of the proposed components. We also show that prototypes can learn diverse sub-style patterns with clear semantic meaning, enhancing both diversity and interpretability.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (92570101), the Major Program of the Natural Science Foundation of Zhejiang Province, China (LD26F020003) and the Earth System Big Data Platform of the School of Earth Sciences, Zhejiang University.

References

1. Aberman, K., Weng, Y., Lischinski, D., Cohen-Or, D., Chen, B.: Unpaired motion style transfer from video to animation. ACM Transactions on Graphics (TOG) **39**(4), 64–1 (2020)
2. Ahuja, C., Morency, L.P.: Language2pose: Natural language grounded pose forecasting. In: 2019 International conference on 3D vision (3DV). pp. 719–728. IEEE (2019)
3. Ahuja, K., Ofek, E., Gonzalez-Franco, M., Holz, C., Wilson, A.D.: Coolmoves: User motion accentuation in virtual reality. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **5**(2), 1–23 (2021)
4. Akber, S.M.A., Kazmi, S.N., Mohsin, S.M., Szczęsna, A.: Deep learning-based motion style transfer tools, techniques and future challenges. Sensors **23**(5), 2597 (2023)
5. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: Proceedings of the European conference on computer vision (ECCV). pp. 132–149 (2018)
6. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems **33**, 9912–9924 (2020)
7. Chen, G., Li, X., Yang, Y., Wang, W.: Neural clustering based visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5714–5725 (2024)
8. Chen, K., Wang, J., Zhang, J., Li, M., Lu, Y., Fan, H.: Scaling video understanding via compact latent multi-agent collaboration. arXiv preprint arXiv:2605.00444 (2026)
9. Chen, K., Wu, Z., Hou, W., Li, K., Fan, H., Yang, Y.: Prompt-aware controllable shadow removal. arXiv preprint arXiv:2501.15043 (2025)
10. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PmLR (2020)
11. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18000–18010 (2023)
12. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013)
13. Dabral, R., Mughal, M.H., Golyanik, V., Theobalt, C.: Mofusion: A framework for denoising-diffusion-based motion synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9760–9770 (2023)
14. Ding, C., Zhang, J., Ding, H., Zhao, H., Wang, Z., Xing, T., Hu, R.: Decoupling with entropy-based equalization for semi-supervised semantic segmentation. In: IJCAI (2023)

15. Ding, Y., Li, L., Wang, W., Yang, Y.: Clustering propagation for universal medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3357–3369 (2024)
16. Feng, T., Quan, R., Wang, X., Wang, W., Yang, Y.: Interpretable3d: An ad-hoc interpretable classifier for 3d point clouds. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 1761–1769 (2024)
17. Feng, T., Wang, W., Wang, X., Yang, Y., Zheng, Q.: Clustering based point cloud representation learning for 3d analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8283–8294 (2023)
18. Frigui, H., Krishnapuram, R.: A robust competitive clustering algorithm with applications in computer vision. Ieee transactions on pattern analysis and machine intelligence **21**(5), 450–465 (2002)
19. Gao, X., Yang, Y., Xie, Z., Du, S., Sun, Z., Wu, Y.: Guess: Gradually enriching synthesis for text-driven human motion generation. IEEE Transactions on Visualization and Computer Graphics **30**(12), 7518–7530 (2024)
20. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2414–2423 (2016)
21. Ghosh, A., Cheema, N., Oguz, C., Theobalt, C., Slusallek, P.: Synthesis of compositional animations from textual descriptions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1396–1406 (2021)
22. Goel, P., Wang, K.C., Liu, C.K., Fatahalian, K.: Iterative motion editing with natural language. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–9 (2024)
23. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: Momask: Generative masked modeling of 3d human motions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1900–1910 (2024)
24. Guo, C., Mu, Y., Zuo, X., Dai, P., Yan, Y., Lu, J., Cheng, L.: Generative human motion stylization in latent space. arXiv preprint arXiv:2401.13505 (2024)
25. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5152–5161 (2022)
26. Guo, C., Zuo, X., Wang, S., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: ECCV (2022)
27. Guo, Z., Lee, Y.Y., Liu, J., Ben-Shabat, Y., Zordan, V., Kapadia, M.: Stylemotif: Multi-modal motion stylization using style-content cross fusion. arXiv preprint arXiv:2503.21775 (2025)
28. Han, B., Peng, H., Dong, M., Ren, Y., Shen, Y., Xu, C.: Amd: Autoregressive motion diffusion. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 2022–2030 (2024)
29. He, T., Gao, L., Song, J., Li, Y.F.: Towards open-vocabulary scene graph generation with prompt-based finetuning. In: European conference on computer vision. pp. 56–73. Springer (2022)
30. He, T., Gao, L., Song, J., Li, Y.F.: Toward a unified transformer-based framework for scene graph generation and human-object interaction detection. IEEE Transactions on Image Processing **32**, 6274–6288 (2023)
31. Hu, L., Zhang, Z., Ye, Y., Xu, Y., Xia, S.: Diffusion-based human motion style transfer with semantic guidance. In: Computer Graphics Forum. vol. 43, p. e15169. Wiley Online Library (2024)
32. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017)

33. Jang, D.K., Park, S., Lee, S.H.: Motion puzzle: Arbitrary motion style transfer by body part. ACM Transactions on Graphics (TOG) **41**(3), 1–16 (2022)
34. Jin, P., Wu, Y., Fan, Y., Sun, Z., Yang, W., Yuan, L.: Act as you wish: Fine-grained control of motion diffusion model with hierarchical semantic graphs. Advances in Neural Information Processing Systems **36**, 15497–15518 (2023)
35. Jolion, J.M., Meer, P., Bataouche, S.: Robust clustering with applications in computer vision. IEEE transactions on pattern analysis and machine intelligence **13**(8), 791–802 (1991)
36. Karunratanakul, K., Preechakul, K., Suwajanakorn, S., Tang, S.: Guided motion diffusion for controllable human motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2151–2162 (2023)
37. Kim, B., Jeong, H.I., Sung, J., Cheng, Y., Lee, J., Chang, J.Y., Choi, S.I., Choi, Y., Shin, S., Kim, J., et al.: Personabooth: Personalized text-to-motion generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22756–22765 (2025)
38. Kim, B., Kim, J., Chang, H.J., Choi, J.Y.: Most: Motion style transformer between diverse action contents. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1705–1714 (2024)
39. Li, W., Su, X., Cao, Y., Xu, H., Xia, X., You, S., Chen, Y., Xu, C.: Vla-attc: Adaptive test-time compute for vla models with relative action critic model. arXiv preprint arXiv:2605.01194 (2026)
40. Li, Z., He, Y., Zhong, L., Shen, W., Zuo, Q., Qiu, L., Dong, Z., Yang, L.T., Yuan, W.: Mulsmo: Multimodal stylized motion generation by bidirectional control flow. arXiv preprint arXiv:2412.09901 (2024)
41. Liang, J., Cui, Y., Wang, Q., Geng, T., Wang, W., Liu, D.: Clusterfomer: clustering as a universal visual learner. Advances in neural information processing systems **36**, 64029–64042 (2023)
42. Liang, J., Zhou, T., Liu, D., Wang, W.: Clustseg: Clustering for universal segmentation. arXiv preprint arXiv:2305.02187 (2023)
43. Liu, X., Han, X., Xia, H., Li, K., Zhao, H., Jia, J., Zhen, G., Su, L., Zhao, F., Cao, X.: Pointcluster: Deep clustering of 3-d point clouds with semantic pseudo-labeling. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–14 (2024)
44. Liu, Y., Tian, B., Lv, Y., Li, L., Wang, F.Y.: Point cloud classification using content-based transformer via clustering in feature space. IEEE/CAA Journal of Automatica Sinica **11**(1), 231–239 (2023)
45. Lu, Y., Quan, R., Zhu, L., Yang, Y.: Zero-shot video grounding with pseudo query lookup and verification. IEEE Transactions on Image Processing **33**, 1643–1654 (2024)
46. Mason, I., Starke, S., Komura, T.: 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 (2022)
47. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
48. Niu, C., Shan, H., Wang, G.: Spice: Semantic pseudo-labeling for image clustering. IEEE Transactions on Image Processing **31**, 7264–7278 (2022)
49. Park, S., Jang, D.K., Lee, S.H.: 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 (2021)

50. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3D human motion synthesis with transformer VAE. In: International Conference on Computer Vision (ICCV) (2021)
51. Petrovich, M., Black, M.J., Varol, G.: Temos: Generating diverse human motions from textual descriptions. In: European Conference on Computer Vision. pp. 480–497. Springer (2022)
52. Petrovich, M., Black, M.J., Varol, G.: TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In: International Conference on Computer Vision (ICCV) (2023)
53. Pinyoanuntapong, E., Saleem, M.U., Wang, P., Lee, M., Das, S., Chen, C.: Bamm: Bidirectional autoregressive motion model. In: Computer Vision – ECCV 2024 (2024)
54. Pinyoanuntapong, E., Wang, P., Lee, M., Chen, C.: Mmm: Generative masked motion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
55. Qian, Z., Yang, D., Li, M., Kou, D., Zhang, L.: Mafd: Fine-grained motion style transfer with adaptive signal fusion. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
56. Qu, H., Wei, J., Shu, X., Wang, W.: Learning clustering-based prototypes for compositional zero-shot learning. arXiv preprint arXiv:2502.06501 (2025)
57. Raab, S., Gat, I., Sala, N., Tevet, G., Shalev-Arkushin, R., Fried, O., Bermano, A.H., Cohen-Or, D.: Monkey see, monkey do: Harnessing self-attention in motion diffusion for zero-shot motion transfer. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–13 (2024)
58. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
59. Reynolds, D.: Gaussian mixture models. In: Encyclopedia of biometrics, pp. 827–832. Springer (2015)
60. Sawdayee, H., Guo, C., Tevet, G., Zhou, B., Wang, J., Bermano, A.H.: Dance like a chicken: Low-rank stylization for human motion diffusion. arXiv preprint arXiv:2503.19557 (2025)
61. Song, W., Jin, X., Li, S., Chen, C., Hao, A., Hou, X., Li, N., Qin, H.: Arbitrary motion style transfer with multi-condition motion latent diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 821–830 (2024)
62. Tang, X., Wu, L., Wang, H., Hu, B., Gong, X., Liao, Y., Li, S., Kou, Q., Jin, X.: Rsmt: Real-time stylized motion transition for characters. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–10 (2023)
63. Tao, T., Zhan, X., Chen, Z., van de Panne, M.: Style-erd: Responsive and coherent online motion style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6593–6603 (2022)
64. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: Motionclip: Exposing human motion generation to clip space. In: European Conference on Computer Vision. pp. 358–374. Springer (2022)
65. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=SJ1kSyO2jwu

66. Wan, W., Dou, Z., Komura, T., Wang, W., Jayaraman, D., Liu, L.: Tlcontrol: Trajectory and language control for human motion synthesis. In: European Conference on Computer Vision. pp. 37–54. Springer (2024)
67. Wang, W., Han, C., Zhou, T., Liu, D.: Visual recognition with deep nearest centroids. arXiv preprint arXiv:2209.07383 (2022)
68. Wang, Y., Leng, Z., Li, F.W., Wu, S.C., Liang, X.: Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22035–22044 (2023)
69. Wei, M., Xie, X., Shi, G.: Acmo: Attribute controllable motion generation. arXiv preprint arXiv:2503.11038 (2025)
70. Wen, Y.H., Yang, Z., Fu, H., Gao, L., Sun, Y., Liu, Y.J.: Autoregressive stylized motion synthesis with generative flow. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13612–13621 (2021)
71. Xie, Y., Jampani, V., Zhong, L., Sun, D., Jiang, H.: Omnicontrol: Control any joint at any time for human motion generation. arXiv preprint arXiv:2310.08580 (2023)
72. Xie, Z., Wu, Y., Gao, X., Sun, Z., Yang, W., Liang, X.: Towards detailed text-to-motion synthesis via basic-to-advanced hierarchical diffusion model. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6252–6260 (2024)
73. Xu, J., Xu, H., Ni, B., Yang, X., Wang, X., Darrell, T.: Hierarchical style-based networks for motion synthesis. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. pp. 178–194. Springer (2020)
74. Yan, X., Misra, I., Gupta, A., Ghadiyaram, D., Mahajan, D.: Clusterfit: Improving generalization of visual representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6509–6518 (2020)
75. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
76. Yuan, Y., Song, J., Iqbal, U., Vahdat, A., Kautz, J.: Physdiff: Physics-guided human motion diffusion model. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16010–16021 (2023)
77. Zhang, J., Fan, H., Yang, Y.: Energymogen: Compositional human motion generation with energy-based diffusion model in latent space. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17592–17602 (2025)
78. Zhang, J., Fan, H., Yang, Y.: Towards decompositional human motion generation with energy-based diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 30650–30660 (2026)
79. Zhang, J., Wu, T., Ding, C., Zhao, H., Guo, G.: Region-level contrastive and consistency learning for semi-supervised semantic segmentation. In: IJCAI (2022)
80. Zhang, J., Zhang, Y., Cun, X., Zhang, Y., Zhao, H., Lu, H., Shen, X., Shan, Y.: Generating human motion from textual descriptions with discrete representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14730–14740 (2023)
81. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
82. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. IEEE transactions on pattern analysis and machine intelligence **46**(6), 4115–4128 (2024)

83. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: Remodiffuse: Retrieval-augmented motion diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 364–373 (2023)
84. Zhang, Z., Liu, A., Reid, I., Hartley, R., Zhuang, B., Tang, H.: Motion mamba: Efficient and long sequence motion generation. In: European Conference on Computer Vision. pp. 265–282. Springer (2024)
85. Zhong, C., Hu, L., Zhang, Z., Xia, S.: Attt2m: Text-driven human motion generation with multi-perspective attention mechanism. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 509–519 (October 2023)
86. Zhong, L., Xie, Y., Jampani, V., Sun, D., Jiang, H.: Smoodi: Stylized motion diffusion model. In: European Conference on Computer Vision. pp. 405–421. Springer (2024)
87. Zhong, L., Yang, Y., Li, C.: Smoogpt: Stylized motion generation using large language models. arXiv preprint arXiv:2509.04058 (2025)
88. Zhou, T., Wang, W., Konukoglu, E., Van Gool, L.: Rethinking semantic segmentation: A prototype view. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2582–2593 (2022)
89. Zhou, Z., Wang, B.: Ude: A unified driving engine for human motion generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5632–5641 (2023)