Editing Motion with Words: A Training-Free Framework for Text-Guided Motion Stylization

Anonymous Authors

Anonymous Institute

Abstract. Editing motion style through textual prompts without motion exemplars remains an open challenge. Existing text-guided editing methods struggle to manipulate global motion styles while preserving action content. We present TextStyler, a training-free framework that repurposes pre-trained motion diffusion models for text-driven global stylization—editing the "how" of motion execution while maintaining the "what" of action semantics, without style-specific training or stylized motion references. We implement a KV cache that retains content-critical key-value pairs from the original motion during inversion, ensuring temporal coherence and action fidelity. During stylized denoising, we inject target style information through a text-aligned loss guidance while preserving cached semantic features through masked cross-attention. This framework enables stylization without requiring style-specific training or motion references. Through extensive experiments on the HumanML3D dataset, we show the superiority of our method over the current method.

Keywords: Diffusion · Human motion · Motion stylization.

1 Introduction

Text-driven human motion generation has advanced significantly with diffusion models, yet a critical gap remains: editing the style of motions through text alone—transforming a "walk" into a "drunk walk" or a "punch" into a "monkey punch"—without requiring motion exemplars or task-specific fine-tuning. While existing methods enable basic text-to-motion synthesis, they conflate action semantics ("what" is performed) and motion style ("how" it is performed), limiting their ability to manipulate global stylistic attributes while preserving the original motion's intent.

Existing methods often struggle to guarantee strict content preservation and style independence, resulting in semantic drift or structural inconsistencies in the generated motions. Moreover, many motion style transfer methods depend heavily on stylized motion datasets that are typically limited in both scale and style diversity. These approaches either compromise motion quality through distribution shift, require multi-modal references, or lack zero-shot capability, leaving unmet the need for flexible textual style control in pre-trained models. Further, the inherent scarcity of motion data annotated with nuanced style labels poses fundamental challenges for learning-based approaches, which typically fail

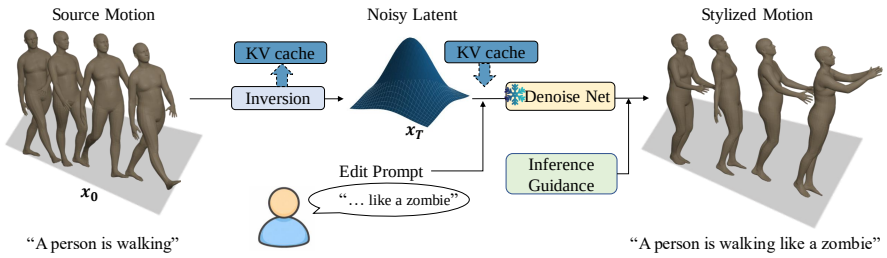

Fig. 1. Editing Motion with Words: A Training-Free Framework for Text-Guided Motion Stylization

to generalize beyond seen style-action combinations. Consequently, their performance degrades substantially when applied to unseen styles or cross-domain scenarios, revealing weak cross-style generalization capability. Therefore, reducing reliance on annotated style datasets and improving zero-shot or few-shot style transfer remain critical challenges in motion stylization and constitute the primary motivation of this work.

Recent image editing breakthroughs demonstrate that hierarchical diffusion features can enable disentangled control through attention manipulation. However, directly adapting these techniques to motion fails due to critical differences: human motions require temporal coherence in root trajectories while enabling stylistic variations in limb movements, unlike the spatial consistency requirements in images. Moreover, text conditioning for motion exhibits stronger semantic-style coupling than in vision-language models - "zombie" modifies both movement style and body posture in ways that text encoders struggle to isolate.

In this work, we introduce TextStyler, a novel training-free framework that achieves global motion stylization solely through text prompts by reusing a pretrained motion diffusion model. As shown in Figure 1, given a human motion and a style editing prompt, TextStyler can successfully extract the style information expressed in the edited prompt without altering the original motion content. The TextStyler method is built upon two key insights: 1) utilizing a KV-cache attention manipulation to preserve the temporal consistency and content semantics of generated human motions during inversion and denoising through hierarchical activation caching and disentanglement of root and local features; 2) text-aligned style guidance, which injects target style information through a pretrained text-motion alignment model, thereby ensuring style fidelity while avoiding semantic drift. Our contributions are summarized as follows:

- We propose TextStyler, the first training-free text-driven motion stylization framework, which requires neither stylized motion examples, model fine-tuning, nor paired datasets.
- We design a KV-cache-based attention mechanism that decomposes motions into root and local joint features, and caches key-value pairs during the

DDIM[17] inversion process, thereby preserving temporal consistency and motion semantics during style editing. In addition, a loss-guided semantic alignment mechanism is introduced, where the loss term modulates style-related features through a pretrained text-motion model during the late denoising stage, thereby ensuring consistency between style and text while avoiding content distortion.

- We construct the TMB (Text-edited Motion Benchmark), which contains triplets of motion text prompts, original motions, and style-edited motions from HumanML3D[6]. Experimental results on TMB demonstrate that our TextStyler outperforms existing methods in training-free stylized generation capability, which is further validated through qualitative results, quantitative results, and ablation studies.

2 Related works

2.1 Text-guided Motion Editing

Human motion editing has been extensively studied, yet it remains a challenging problem. A fundamental challenge in human motion editing arises from the strong structural dependencies of the human skeleton, where joints are tightly coupled and mutually constrained.

To address this limitation, Goel et al.[4] extended text-guided editing to full motion sequences through heuristic keyframe selection. Methods such as FineMoGen[21], Iterative Motion Editing[4], and COMO[8] rely on foundation models for motion generation and editing, yet they are unable to process arbitrary unlabeled motion inputs and depend heavily on the reasoning capabilities of LLMs. MotionFix[2] introduced a dedicated dataset together with a diffusion-based motion editing framework. However, because the model is trained on a limited set of (source motion, edited motion, instruction) triplets, it struggles to generalize to more complex motion compositions and temporal dependencies. Building upon this line of work, SimMotionEdit[12] further improves editing quality by incorporating auxiliary training tasks.

Existing approaches attempt to bridge this gap through various approaches. Existing methods generally fail to avoid the coupling effect in which local modifications induce unintended global changes, often resulting in artifacts such as unnatural poses or unstable motion dynamics. We therefore investigate how to better balance two competing objectives: preserving the semantic content and dynamics of the original motion while accurately applying text-guided modifications. The goal is to generate edited motions that satisfy user instructions while remaining temporally coherent, physically plausible, and natural.

2.2 Motion Stylization

Achieving effective motion stylization remains a persistent challenge. The key challenge of motion stylization lies in effectively disentangling motion “content”

(what action is performed) from motion "style" (how the action is performed). Recent advancements in motion style transfer[5, 9] have been facilitated by the integration of sophisticated neural architectures and generative models. Existing motion style transfer approaches typically rely on dedicated stylized motion datasets[1, 13]. However, the motion diversity in these datasets is often limited, which constrains the scalability and generalization of such methods, particularly in the context of diffusion-based motion generation models.

Guo et al.[5] leveraged the latent space of pretrained motion models to improve the extraction and fusion of motion content and style representations. Although it alleviates the requirement for paired training data, they remain fundamentally constrained by the predefined style distributions observed during training, thereby limiting their generalization to unseen styles. In contrast, training-free frameworks offer a more promising direction for improving style generalization and adaptability.

SMooDi[22] introduces style adapters and classifier guidance to condition generation on reference motions, but requires paired style-content datasets and struggles with out-of-distribution actions. MulSMo[11] proposes bidirectional control flow between style and content networks, yet still depends on explicit style motion sequences for conditioning. While LoRA-MDM[16] demonstrates style preservation through low-rank adaptation, it requires per-style fine-tuning that limits practical application.

3 Method

This section presents TextStyler, our training-free motion stylization framework. Built upon a pretrained motion diffusion model, TextStyler exploits the KV-cache mechanism to preserve content-related key-value pairs from the original motion during the inversion process, thereby maintaining temporal consistency and motion fidelity. During the stylized denoising stage, semantic information is retained through the cached features, while an additional alignment loss is introduced to enforce semantic consistency between the edited text prompt and the generated motion. Our method achieves style-content disentanglement without requiring model retraining or paired training data.

3.1 Preliminaries

Human Motion Models. We adopt the popular 263-dimensional HumanML3D format to represent human motions $x^{1:N}$, where each pose in the motion sequence encodes root kinematics (height, angular/linear velocities), and joint-level information (positions, velocities, and rotation). We adhere to common practices in motion diffusion generation by employing the Denoising Diffusion Probabilistic Models (DDPM) framework.

Inference Guidance. Inference guidance[3, 20, 19] is a training-free method that can deal seamlessly with the infinite diversity of possible conditioning, modulating the denoising process via gradient-based optimization of target-aligned

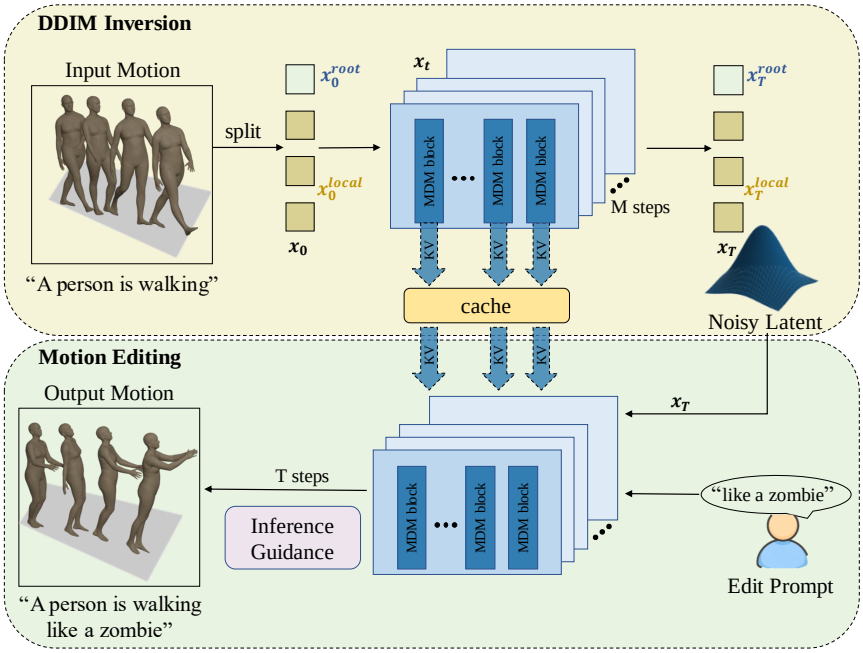

Fig. 2. The editing pipeline of TextStyler. (a) Top: Illustration of the **Key-Value (KV) caching scheme** in the DDIM inversion process. (b) Bottom: The style editing stage utilizes cached content features from (a) to reweight self-attention and cross-attention maps, enabling **style infusion** while preserving *content integrity* via root-related activations.

loss functions. DSG[19] is an inference guidance method that can be a plugin module to existing training-free conditional diffusion models with almost no extra computational overhead. Therefore, we designed a loss function based on motion style information and used DSG to guide the denoising process, generating motions with new styles.

3.2 Pipeline

The full pipeline of our method is illustrated in Fig. 2. Our framework builds on text-conditioned diffusion models that use self-attention to refine motion sequences and cross-attention to inject textual control.The source motion x^{source} is generated by the source text prompt c^{source}. We employ a pre-trained and fixed motion diffusion model[18] to synthesize the output motion x^{edit} by applying an edit prompt c^{edit} onto the source motion x^{source}.

Given source motion $x^{\text{source}} \in \mathbb{R}^{T \times d}$ with T frames, we first invert it into the noisy latent using DDIM inversion. The source prompt c^{source} assists the inversion by conditioning the reverse diffusion process, ensuring the inverted latent retains temporally coherent features aligned with the original motion seman-

tics. This inversion step reconstructs a trajectory in the latent space that, when denoised, faithfully recovers x^{source} under the guidance of c^{source} with minimal distortion.

While we adopt the cross-attention variant of MDM as our backbone, naïve replacement of c^{source} with c^{edit} during denoising often degrades motion integrity or introduces semantic inconsistencies. To address this, we propose two key enhancements, which are detailed subsequently: **Decomposed attention control with KV cache** and **Loss-guided semantic alignment**.

Decomposed attention control with KV cache. To enable preservation of global motion content and localized motion style editing, we partition the motion representation into root and local components. Each frame in a motion sequence is decomposed into root and local components $x = (x^{\text{root}}, x^{\text{local}})$.

Caching mechanism During DDIM inversion, we isolate the key-value activations corresponding to root and local motion parameters through their positional indices in the motion representation. Then, for each attention layer, we store the root-related KV activations $K_t^{\text{root}}, V_t^{\text{root}}$ and local-related activations $K_t^{\text{local}}, V_t^{\text{local}}$ separately on each timestep t, preserving their temporal dependencies. The pseudocode for KV caching is shown in Algorithm 1. The cached results are used in denoising time to ensure a smooth transition between original content and edited style.

Algorithm 1 DDIM inversion with KV cache

Input: t, human motion x_t, M-layer block $\{l_j\}_{j=1}^{M}$
Output: Prediction vector $V_\theta(x_t)$, KV cache C_{root} and C_{local}
for $j = 0$ **to** M **do**
    $Q, K, V \leftarrow W_Q(x_t), W_K(x_t), W_V(x_t)$
    $K_{\text{root}_j}, V_{\text{root}_j} \leftarrow K_{[:,\mathcal{I}_{\text{root}}]}, V_{[:,\mathcal{I}_{\text{root}}]}$
    $K_{\text{local}_j}, V_{\text{local}_j} \leftarrow K_{[:,\mathcal{I}_{\text{local}}]}, V_{[:,\mathcal{I}_{\text{local}}]}$
    $C_{\text{root}} \leftarrow \text{Append}(K_{\text{root}_j}, V_{\text{root}_j})$
    $C_{\text{local}} \leftarrow \text{Append}(K_{\text{local}_j}, V_{\text{local}_j})$
    $x_t \leftarrow x_t + \text{Attn}(Q, K, V)$ ▷ Details of Transformer blocks are omitted
end for
$V_\theta(x_t) \leftarrow \text{MLP}(x_t)$
return $V_\theta(x_t)$, C_{root}, C_{local}

Denoising with cached context During editing-phase denoising, we strategically fuse cached and generated activations. For motion self-attention layers, we manipulate the KV activations by blending cached and generated results, as:

$$K_t' = \text{concat}\left(K_t^{\text{root}}, \alpha \cdot K_t^{\text{cached local}} + (1-\alpha) K_t^{\text{local}}\right) \tag{1}$$

$$V_t' = \text{concat}\left(V_t^{\text{root}}, \alpha \cdot V_t^{\text{cached local}} + (1-\alpha) V_t^{\text{local}}\right) \tag{2}$$

Rather than directly replacing the local-related KV activations with the generated ones, we propose to blend the cached local-related KV activations with generated local-related KV activations by a blending hyperparameter α. This blending scheme enables smooth interpolation between the original motion context and newly generated features. This ensures gradual content adaptation while preserving fundamental motion characteristics through the fixed root components. Then the blended result is concatenated with cached root-related KV activations to form the final KV matrices.

For text-conditioned cross-attention, we implement a dual-prompt fusion strategy by concatenating source and edit textual embeddings:

$$K'_t = [K^{\text{source}}, K^{\text{edit}}], V'_t = [V^{\text{source}}, V^{\text{edit}}] \tag{3}$$

To suppress the output motion overly focus on the original context of the prompt, we anneal the similarity score that corresponds to the source textual embedding regions before using softmax to compute the final similarity score, as:

$$\text{Cross-attn}(Q, K, V) = \text{softmax}\left(\frac{QW_Q([\beta_t K^{\text{source}}, K^{\text{edit}}]W_K)^T}{\sqrt{d}}\right) VW_V \tag{4}$$

We found that, rather than choosing β_t as a fixed constant across all denoising timesteps t, drawing β_t from some monotonically decreasing function can benefit the editing process. We hypothesize that the occurrence of this phenomenon again is due to the hierarchical denoising nature of diffusion models: early denoising steps (where global structure is determined) maintain stronger source context, while later steps (focusing on detail refinement) progressively emphasize the style conveyed by the edit prompt. We empirically chose the linear schedule of β_t from 0.5 down to 0.1. The pseudocode for attention manipulation is shown in Algorithm 2.

Algorithm 2 Attention manipulation with KV cache

Input: t, human motion x_t, M-layer block $\{l_j\}_{j=1}^{M}$, KV cache C_{root} and C_{local}, KV blending weight α
Output: denoised motion $\hat{x}_0$
for $j = 0$ **to** M **do**
    $Q, K', V' \leftarrow W_Q(x_t), W_K(x_t), W_V(x_t)$
    $K \leftarrow \text{concat}\left(K_{\text{root}}^{\text{cached}}, \alpha \cdot K_{\text{local}}^{\text{cached}} + (1-\alpha) \cdot K'_{\text{local}}\right)$
    $V \leftarrow \text{concat}\left(V_{\text{root}}^{\text{cached}}, \alpha \cdot V_{\text{local}}^{\text{cached}} + (1-\alpha) \cdot V'_{\text{local}}\right)$
    $x_t \leftarrow x_t + \text{Attn}(Q, K, V)$ ▷ Details of Transformer blocks are omitted
end for
return $\hat{x}_0$

Loss-guided Semantic Alignment. Traditional style transfer methods rely on iterative optimization training between content loss and style loss. Due to the limited availability of motion data with style annotations, models trained on such data usually struggle to generalize well to unseen styles. Therefore, we design a loss function based on motion-style information, which does not require fine-tuning the pretrained model; Instead, it guides the denoising process solely through differentiable loss terms, thereby generating motions with novel styles.

TMR[14] is a benchmark for text-to-motion retrieval, evaluating how related the generated motion is to the given prompt. It has been trained with both a motion and a text encoder to embed cross-modal inputs. Inspired by TMR, let $\mathcal{L}_{\text{alignment}}(x, c)$ denote the guidance loss defining the discrepancy between the source motion and the textual embedding.

$$\mathcal{L}_{\text{alignment}}(x, c) = \lambda_G \cdot \cos\left(E_{\mathcal{M}}(x), E_{\mathcal{T}}(c)\right), \tag{5}$$

where λ_G controls the intensity of the inference guidance, $E_{\mathcal{M}}$ is the motion encoder branch of TMR, and $E_{\mathcal{T}}$ is the text encoder branch of TMR.

To explicitly align the denoised motion with c^{edit}, we iteratively steer x_t toward latent regions that maximize cross-modal consistency without fine-tuning the base diffusion model by backpropagating $\nabla_{x_t}\mathcal{L}_{\text{alignment}}$ through the denoising steps. Following DSG[19], at each denoising step t, we use the current x_t to estimate the denoised x_0. Subsequently, the guidance loss $L(x_0, c)$ is computed based on the conditioning text signal c. The gradient of this loss is then employed to steer the update direction of x_t:

$$x_t = \mu_\theta(x_t) - \sqrt{n}\sigma_t \frac{\nabla_{x_t} L(x_0(x_t), c)}{||\nabla_{x_t} L(x_0(x_t), c)||_2}, \tag{6}$$

$$\mu_\theta(x_t) = \sqrt{\alpha_{t-1}} x_0(x_t) + \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \cdot \epsilon_\theta(x_t, t), \tag{7}$$

$$\sigma_t = \eta\sqrt{(1 - \alpha_{t-1})/(1 - \alpha_t)}\sqrt{1 - \alpha_t/\alpha_{t-1}}, \tag{8}$$

where $\mu_\theta(x_t)$ denotes the mean of the Gaussian distribution, n denotes the dimension of motion x, α_t is a predefined noise scheduling parameter, and $\epsilon_\theta(x_t, t)$ denotes the noise predicted by the pretrained diffusion model θ.

This iterative scheme can reduce the manifold deviation occurring in the guidance steps, while also promoting the generated motion to conform to the edited text content. Trust Sampling[7] allows for multiple gradient steps on a proxy constraint function at each diffusion step, while scheduling the termination of the optimization when the proxy cannot be trusted anymore. Inspired by it, we estimate the state manifold of the diffusion model to allow for early termination if the predicted noise magnitude of the sample $||\epsilon_\theta(\hat{x_t}, t)||$ exceeds the expected one ϵ_{max} in each diffusion step. The pseudocode for the whole pipeline is shown in Algorithm 3.

Algorithm 3 TextStyler motion styliation pipeline

Input: source human motion x_0^{source}, source text prompt c^{source}, editing prompt c^{edit}, human motion denoising model $G(\theta)$
Output: edited stylized motion x_0^{edit}
$\bar{x}_T, C_{\text{root}}, C_{\text{local}} \leftarrow \text{DDIM-inv}\left(G(\theta), x_0^{\text{source}}, c^{\text{source}}\right)$ ▷ DDIM inversion with KV cache in algorithm 1
for $t = T$ **down to** 0 **do**
    $\hat{x}_t \leftarrow \text{attention-manipulation}\left(G(\theta), \bar{x}_t, C_{\text{root}}, C_{\text{local}}\right)$ ▷ Attention manipulation in algorithm 2
    while $||\epsilon_\theta(\hat{x}_t, t)|| < \epsilon_{max}$ **do**
        $x_t \leftarrow \mu_\theta(\hat{x}_t) - \sqrt{n}\sigma_t \frac{\nabla \mathcal{L}_{\text{alignment}}(\hat{x}_t, c^{\text{edit}})}{\left\|\nabla \mathcal{L}_{\text{alignment}}(\hat{x}_t, c^{\text{edit}})\right\|}$ ▷ Semantic alignment loss guidance
    end while
end for
$x_0^{\text{edit}} \leftarrow x_0$ ▷ The final output
return x_0^{edit}

4 Experiments

4.1 Experiment setting and benchmark

We adopt the transformer decoder variant of MDM as our backbone architecture due to its ability to integrate textual prompts via cross-attention mechanisms. This design enables explicit control over motion generation by conditioning on text inputs, making it well-suited for style transfer tasks.

We introduce the **T**ext-edited **M**otion **B**enchmark (TMB), a text-driven motion editing benchmark comprising (source motion, source prompt, edit prompt) triplets. TMB extends the Motion Transfer Benchmark (MTB)[15] by repurposing its leader motions (HumanML3D[6] sequences filtered for explicit locomotion verbs, e.g., "a person walks forward", "someone walks slowly") as source content. Source prompts retain MTB's original textual descriptions, while edit prompts are distilled from MTB's style-focused follower motions by truncating phrases to isolate style modifiers (e.g., converting "a person moves like a chicken" ⇒ "... like a chicken"). We generate triplets via the Cartesian product pairing of sources and edits. Unlike MTB, which evaluates motion-to-motion transfer, TMB directly tests textual stylization by decoupling edit prompts from predefined target motions.

4.2 Metrics

We aim to assess the motion transfer results based on three criteria: (a) the quality of the output motions, (b) how closely the edited motion aligns with the content of the source motion, and (c) how well the edited motion aligns with the semantics imposed by the editing prompt. To evaluate these aspects, we have chosen specific metrics. Criterion (a) is evaluated using FID, (b) is assessed by

foot contact similarity and trajectory similarity, and criterion (c) is evaluated by the cosine similarity between the edited motion and the editing prompt, computed using the TMR model.

4.3 Baselines

TextStyler performs text-guided motion stylization without requiring additional training. Given an arbitrary source motion, users provide an editing text specifying the target style, and the model generates a stylized motion while preserving the original motion content as much as possible. As summarized in Table 1, no existing method addresses exactly the same task setting as TextStyler. With the exception of MoMo[15], all comparable approaches require additional model training. Among them, MulSMo[10] directly synthesizes stylized motions from text prompts without performing editing on existing motions, while MotionFix[2] has not demonstrated effectiveness on stylized human motion generation. MOST[9], SMooDi[22], LoRA-MDM[16], and MoMo[15] cannot directly control motion style through text descriptions and instead require stylized motion examples as inputs.

Table 1. Motion stylization method comparison

Method	Text-guided	Training-free	Stylization	Content Preserve
MulSMo[10]	✓	×	✓	×
MotionFix[2]	✓	×	×	✓
MOST[9]	×	×	✓	✓
SMooDi[22]	×	×	✓	✓
LoRA-MDM[16]	×	×	✓	✓
MoMo[15]	×	✓	✓	✓
TextStyler(Ours)	✓	✓	✓	✓

We first compare TextStyler with MoMo, since both methods operate in a training-free manner and are capable of transferring stylistic characteristics between motions. MoMo is originally designed as a motion-to-motion transfer framework, where the generated motion preserves the style of the follower motion while following the trajectory of the leader motion. To adapt MoMo for our task, the motion x^{source} generated from the original text prompt c^{source} is treated as the leader motion, while the stylized motion generated from the edited text prompt c^{edit} is treated as the follower motion. The original MoMo pipeline is then applied to evaluate its ability to preserve the source motion trajectory while incorporating stylistic characteristics from the stylized motion.

MulSMo, MOST, SMooDi, and LoRA-MDM are all style transfer methods that require additional training. Our experiments focus on the publicly available implementations of SMooDi and LoRA-MDM. For comparison with SMooDi, the stylized motion generated by TextStyler using the edited text prompt c^{edit} is

provided as the input motion to SMooDi. For LoRA-MDM, the pretrained model parameters are kept unchanged, and its control text is used as the original text prompt c^{source} in TextStyler to generate the corresponding source motion x^{source}, enabling evaluation under identical metrics and settings.

In addition, we compare against a simple baseline that directly regenerates a complete stylized motion sequence using the text-to-motion diffusion model MDM conditioned on the edited text prompt.

4.4 Quantitative and Qualitative Results

Table 2. Motion stylization performance comparison on TMB benchmark. Bold and underlined values indicate the best and second-best results, respectively.

Method	**Metrics**			
	FID ↓	Contact Sim. ↑	Trajectory Sim. ↑	Semantic Sim ↑
MDM	-	0.312	0.384	-
MoMo	2.50	<u>0.793</u>	<u>0.830</u>	0.690
SMooDi	2.66	0.658	0.771	0.600
LoRA-MDM	<u>2.44</u>	0.780	0.816	**0.730**
TextStyler	**2.41**	**0.818**	**0.871**	<u>0.728</u>

Quantitative Results Table 2 presents quantitative comparisons between TextStyler and the aforementioned methods on the TMB benchmark. TextStyler achieves superior performance in both contact similarity and trajectory similarity, demonstrating its ability to effectively preserve the original motion content during the stylization process. In summary, TextStyler attains SOTA performance in contact similarity and trajectory similarity while maintaining strong motion content consistency.

We further conduct ablation studies to analyze the effectiveness of each component in Table 3. In TextStyler, both the KV cache mechanism and DSG module play important roles in motion stylization. Experimental results show that the KV cache achieves better performance in terms of FID and semantic similarity, whereas DSG is more effective at preserving contact similarity and trajectory similarity. The guidance steps in the DSG component iteratively optimize the generated motion to better align with the structure of the original motion, but this comes at the cost of partially sacrificing motion quality. To balance motion quality and similarity to the original motion, we introduced a trust sampling mechanism, which performs early stopping estimation during the guidance process. This strategy significantly improves the FID metric. Since DSG is a training-free method and introduces almost no additional computational overhead, the KV cache and DSG components are further combined to achieve better overall performance.

Table 3. Ablation study of TextStyler components. "DSG." denotes the variant using only the DSG module, while "DSGTS." represents the variant using DSG with trust sampling. "KVDSG." denotes the variant that incorporates both the KV cache and DSG modules without trust sampling.

Model	Metrics			
	FID ↓	Contact Sim. ↑	Trajectory Sim. ↑	Semantic Sim ↑
TextStyler DSG. only	3.11	0.815	**0.881**	0.710
TextStyler DSGTS.	<u>2.54</u>	0.815	0.869	0.715
TextStyler KVDSG.	2.87	**0.818**	0.870	<u>0.724</u>
TextStyler	**2.41**	**0.818**	<u>0.871</u>	**0.728**

Qualitative Results. Figures 3 and 4 present several qualitative results of TextStyler. The human motion skeleton results in Figure 3 demonstrate that TextStyler can successfully achieve text-guided motion stylization generation while preserving the content of the original motion. Figure 4 presents visualized qualitative comparisons of the human motions generated by TextStyler, MOMO, LoRA-MDM, and SmooDi.

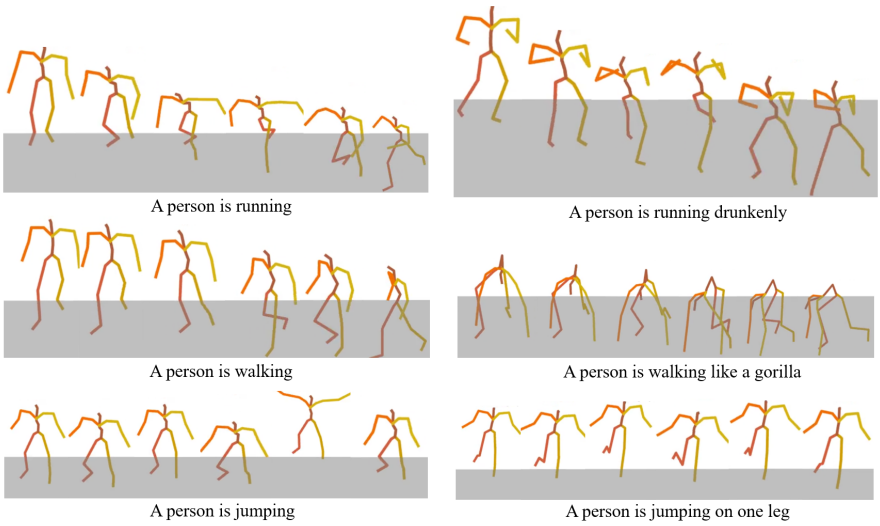

Fig. 3. The qualitative results. In the left column, we display three different motions generated by the original text prompts; In the right column, we show the corresponding stylized motions generated by edited text prompts.

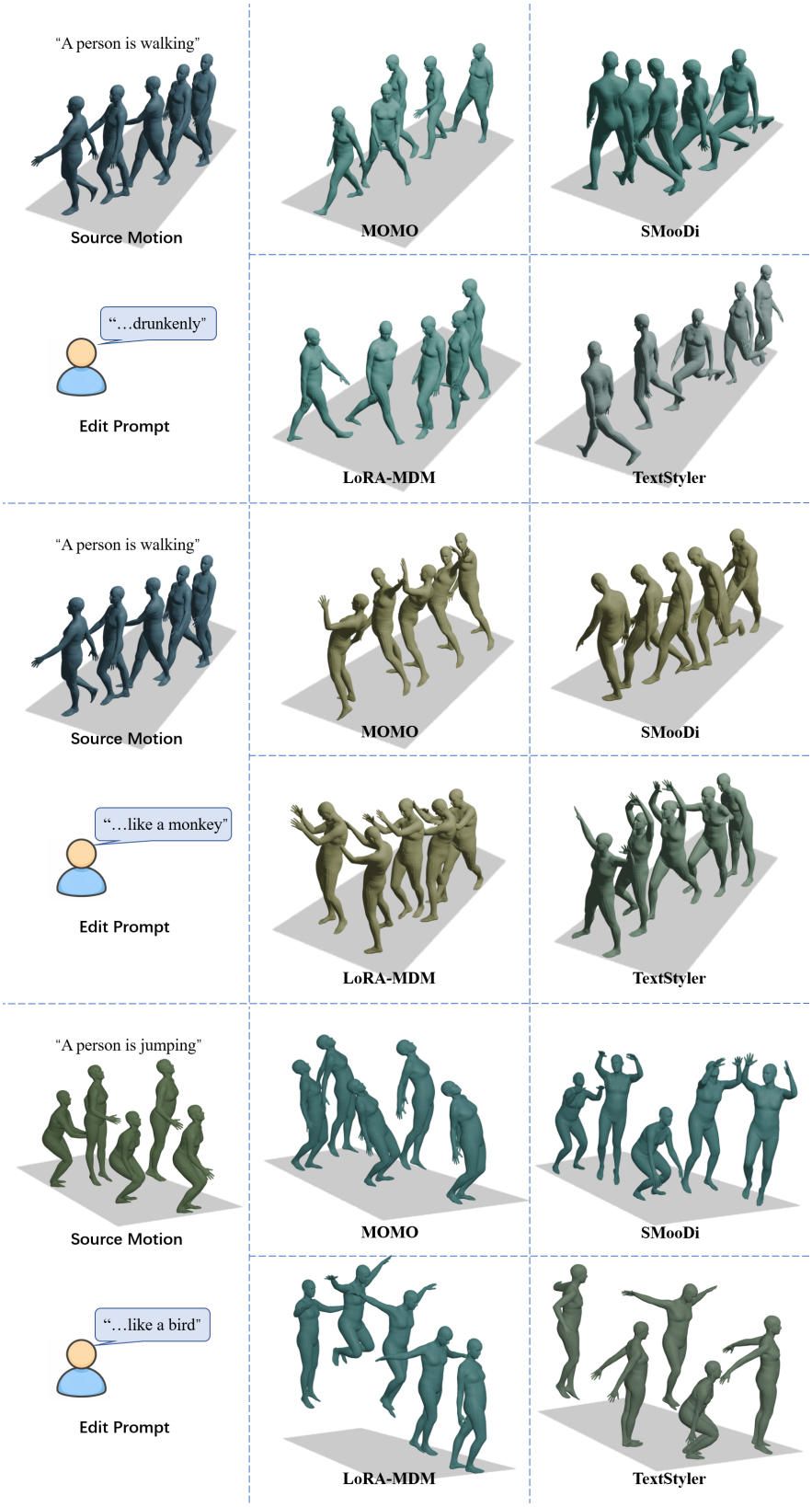

Fig. 4. Comparison of motion stylization qualitative results from different models

The comparison shows that LoRA-MDM can fit the target style well, but causes considerable distortion to the original motion. Although MoMo can generate motions with prominent style characteristics, it is difficult for the style to be effectively adapted to the given text prompt due to the misalignment between style and content attention features in the mixed attention mechanism. The results of SmooDi sometimes overemphasize style and lose plausibility, while in other cases, they ignore the target style in order to conform to the original motion content. This is mainly because SmooDi suffers from overfitting on stylized motion datasets. When handling out-of-distribution text prompts, it often generates motions lacking vividness. Our TextStyler successfully manipulates motion stylization while preserving the original motion's intent.

5 Conclusion

We present TextStyler, a training-free diffusion-based motion stylization framework. TextStyler enables text-guided global motion stylization while preserving motion semantics, without requiring style-specific training or reference motions. We also introduce a new benchmark, TMB, containing selected motion prompts, the paired human motion, and edit prompt triplets from HumanML3D. Extensive quantitative and qualitative experiments on the TMB benchmark demonstrate that TextStyler achieves state-of-the-art performance compared with existing motion stylization approaches.

%subsubsectionAcknowledgments.

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. Athanasiou, N., Cseke, A., Diomataris, M., Black, M.J., Varol, G.: Motionfix: Text-driven 3d human motion editing. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
3. Bansal, A., Chu, H.M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., Goldstein, T.: Universal guidance for diffusion models. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp. 843–852 (2023), https://api.semanticscholar.org/CorpusID:256846836
4. 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)
5. 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)
6. 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)
7. Huang, W., Jiang, Y., Van Wouwe, T., Liu, K.: Constrained diffusion with trust sampling. Advances in Neural Information Processing Systems **37**, 93849–93873 (2024)

8. Huang, Y., Wan, W., Yang, Y., Callison-Burch, C., Yatskar, M., Liu, L.: Como: Controllable motion generation through language guided pose code editing. In: European Conference on Computer Vision. pp. 180–196. Springer (2024)
9. 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)
10. 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)
11. 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 (2025), https://arxiv.org/abs/2412.09901
12. Li, Z., Cheng, K., Ghosh, A., Bhattacharya, U., Gui, L., Bera, A.: Simmotionedit: Text-based human motion editing with motion similarity prediction. arXiv preprint arXiv:2503.18211 (2025)
13. 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)
14. Petrovich, M., Black, M.J., Varol, G.: Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9488–9497 (2023)
15. 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)
16. Sawdayee, H., Guo, C., Tevet, G., Zhou, B., Wang, J., Bermano, A.H.: Dance like a chicken: Low-rank stylization for human motion diffusion (2025), https://arxiv.org/abs/2503.19557
17. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
18. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022)
19. Yang, L., Ding, S., Cai, Y., Yu, J., Wang, J., Shi, Y.: Guidance with spherical gaussian constraint for conditional diffusion. arXiv preprint arXiv:2402.03201 (2024)
20. Yu, J., Wang, Y., Zhao, C., Ghanem, B., Zhang, J.: Freedom: Training-free energy-guided conditional diffusion model. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 23117–23127 (2023), https://api.semanticscholar.org/CorpusID:257622962
21. Zhang, M., Li, H., Cai, Z., Ren, J., Yang, L., Liu, Z.: Finemogen: Fine-grained spatio-temporal motion generation and editing. Advances in Neural Information Processing Systems **36**, 13981–13992 (2023)
22. Zhong, L., Xie, Y., Jampani, V., Sun, D., Jiang, H.: Smoodi: Stylized motion diffusion model. ArXiv **abs/2407.12783** (2024), https://api.semanticscholar.org/CorpusID:271244425