Motion Style Slider: Endpoint-Supervised Continuous Style Control for Human Motion Diffusion

Chen-Chieh Liao[1,3⋆], Yichen Peng[1], Yiyi Cai[2], Yûi Ono[3], Hiroki Hanaoka[3], Erwin Wu[1], Hideki Koike[1], and Shuichi Kurabayashi[3]

[1] Institute of Science Tokyo
[2] The University of Tokyo
[3] Cygames, Inc.

Abstract. Existing human motion diffusion methods provide strong motion generation quality [35], and recent style transfer models can inject target style cues [11, 33], but fine-grained *continuous* control of style intensity remains underexplored. In production, style intensity is subjective across artists and directors, so the practical requirement is not a universal absolute unit (e.g., globally correct "2×"), but a reliable monotonic control axis. We propose **Motion Style Slider**, a motion-to-motion style transfer framework for endpoint-supervised continuous control. Given a content motion and a style motion, we construct a style direction in a learned motion-style embedding space and condition diffusion generation with a scalar intensity α. The training objective combines diffusion denoising with latent intensity regularization to encourage smooth and monotonic style scaling without requiring intermediate-intensity ground-truth motions. Our framework is compatible with pretrained motion diffusion backbones and supports heterogeneous style datasets, including the multi-actor style motion dataset [20]. To test out-of-range usability, we additionally introduce a small real-capture over-reaction extension and evaluate large-α behavior against these unseen targets. Experiments measure controllability, interpolation/extrapolation behavior, content preservation, and motion realism, with ablations on direction construction and loss design.

Keywords: Motion Style Transfer · Human Motion Diffusion · Continuous Style Control

1 Introduction

Generating expressive human motion with controllable style is a central goal for animation, virtual characters, and interactive content creation. Recent motion diffusion models have substantially improved realism and diversity [9, 10, 35, 43], while motion style transfer methods have shown strong adaptation from

⋆ The work was partially done during an internship at Cygames, Inc.

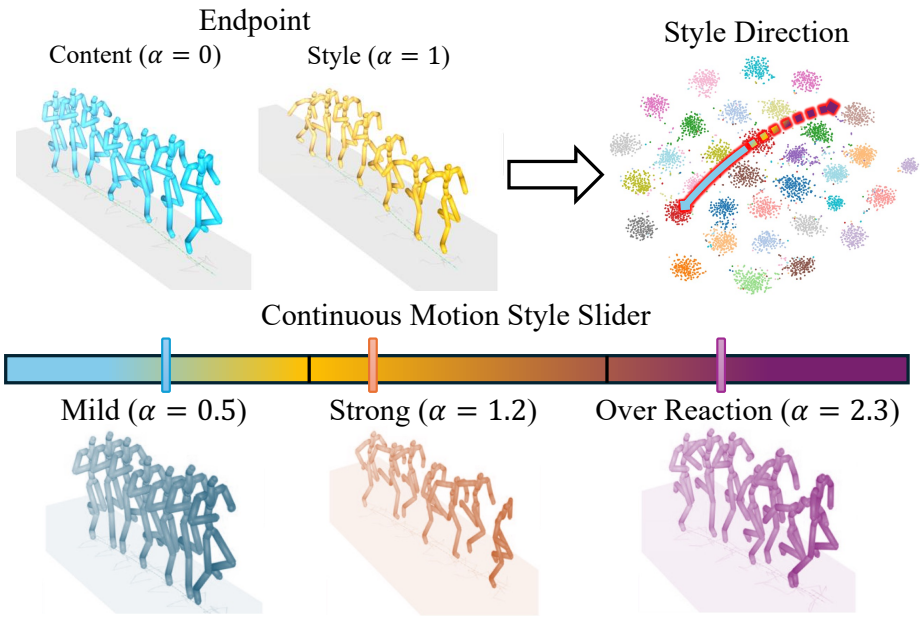

Fig. 1: Motion Style Slider overview. Given a content motion and a target style endpoint, our model produces a continuous family of motions controlled by intensity α, from neutral-style behavior ($\alpha = 0$) to stylized behavior ($\alpha = 1$) and extrapolated over-reaction ($\alpha > 1$).

reference examples [11, 33, 41]. However, most existing pipelines still treat style control as a discrete problem (e.g., *happy*, *sad*) or as weak latent interpolation without explicit intensity supervision. As a result, fine-grained style strength control remains unreliable, and extrapolation beyond seen style strength is rarely validated with real motion targets.

In real game-animation workflows, this appears as iterative direction from directors and animation leads: "more like this", "a bit less", "push it further". Importantly, style intensity is not absolute across users: one person's "2×" may be another's "1.5×". This makes relative controllability the key requirement: as the slider increases, style should change in a predictable order, with smooth transitions and moderate extrapolation headroom beyond observed endpoints.

To address this gap, this paper focuses on **continuous style intensity control** for human motion diffusion. We seek a practical setting where only endpoint supervision is available: a neutral-content motion and a fully stylized motion with the same underlying content. Intermediate-intensity motions are generally unavailable in existing datasets, yet these intermediate states are exactly what users need for controllable generation. This creates a key learning challenge: how to produce smooth, monotonic, and extrapolatable style variation without intermediate ground-truth.

As shown in Figure 1, our central idea is to model style change as a **direction** in a learned style embedding space and to use an explicit scalar intensity variable α for conditioning. Concretely, we compute a style direction from endpoint motions and construct a continuous conditioning code by moving along this direction. The diffusion model receives explicit content-motion conditioning together with style-intensity conditioning, and is trained with denoising loss plus latent intensity regularization. This design makes style intensity control explicit and testable while preserving input content structure.

In this work, we implement this idea on top of an off-the-shelf motion diffusion framework, with style representations derived from motion-language alignment models. Our training and evaluation cover multiple style datasets, enabling analysis across heterogeneous style taxonomies and capture conditions. To directly test out-of-range behavior, we additionally collect a small over-reaction extension with stronger-than-endpoint captures, used as unseen high-strength targets.

Compared with prior motion style transfer pipelines that emphasize one-shot transfer quality, our goal is reliable relative control. We therefore center evaluation on: (i) style monotonicity with respect to α, (ii) interpolation and modest out-of-range behavior, (iii) content preservation, and (iv) motion realism.

In summary, our main contributions are:

- We formulate diffusion-based motion-to-motion style transfer as an **endpoint-supervised continuous control** problem, where style intensity is explicitly parameterized by a scalar slider α and interpreted as a relative control coordinate.
- We propose a style-direction conditioning and latent regularization scheme that enables smooth and monotonic style scaling without requiring intermediate-intensity ground-truth motions.
- We provide a multi-dataset evaluation protocol, including a new real-capture over-reaction split for out-of-range analysis, and report ablations that isolate the effects of direction design and loss terms.

2 Related Work

2.1 Text-to-Motion Generation

Text-to-motion (T2M) generation has progressed from sequence modeling and motion tokenization to diffusion-based synthesis. Early works introduce motion tokenization and reciprocal text–motion generation frameworks such as TM2T and T2M [9, 10]. Building on discrete motion tokens, subsequent works explore language–motion large models including MotionGPT, T2M-GPT, MotionGPT-2, and unified large motion models [17, 37, 42, 45], which treat human motion as a language-like sequence for autoregressive generation. Other extensions further explore masked motion modeling, retrieval augmentation, efficient sequence modeling, keyframe control, or expressive whole-body generation [2, 7, 8, 25, 44, 46].

More recently, diffusion-based approaches significantly improved motion realism and diversity. Representative methods include MotionDiffuse [43], MDM [35],

retrieval-augmented and prior-based variants [32,44], and latent diffusion frameworks such as MLD [33]. These models learn strong motion priors from large text–motion datasets and enable high-quality motion synthesis conditioned on textual descriptions; recent controllable variants further incorporate spatial constraints or real-time control [4, 19, 40].

Despite these advances, most T2M systems primarily control action semantics (e.g., "walk" or "jump") rather than the intensity or strength of style. Consequently, they provide limited mechanisms for calibrated stylistic variation.

2.2 Motion Style Transfer

Motion style transfer (MST) aims to modify the stylistic characteristics of a motion sequence while preserving its underlying content. Styles may correspond to emotional expressions, performer-specific traits, or character-driven motion patterns, enabling expressive virtual characters and animation systems.

Style transfer was first popularized in the image domain [3, 6, 15, 18, 22], where neural networks disentangle content and style representations and recombine them to synthesize stylized outputs. These ideas later inspired motion-based approaches [1,5,12,13,27,38]. Early methods relied on optimization-based frameworks that iteratively adjusted motion features to match target style statistics [12], while subsequent learning-based models improved efficiency and scalability [5, 12].

Later work explored more flexible stylization strategies, including label-free or weakly supervised style learning [16, 38], cross-content style transfer [21, 24], spatial–temporal modeling and stochastic style generation [27], kinematic constraints for physically plausible motion [36], and real-time stylization frameworks for interactive applications [26].

More recently, diffusion-based generative models have significantly improved stylization fidelity and motion realism. Diffusion-based MST methods [14,30,47] leverage strong generative priors to synthesize high-quality stylized motions, while domain-specific approaches such as dance stylization further extend these ideas to specialized motion domains [31]. Latent diffusion frameworks further enable flexible motion–style combinations using multi-condition representations, including MCM-LDM [33] and StyleMotif [11].

Despite these advances, most existing methods treat style as a categorical condition or perform transfer toward a single target style instance. As a result, stylization strength is typically evaluated at a fixed operating point, and continuous or calibrated control of style intensity remains largely unexplored.

2.3 Motion Representation Learning and Style Conditioning

Learning structured motion representations is crucial for controllable motion generation and stylization. Recent work has explored cross-modal representation learning between motion and language to obtain semantically meaningful embedding spaces.

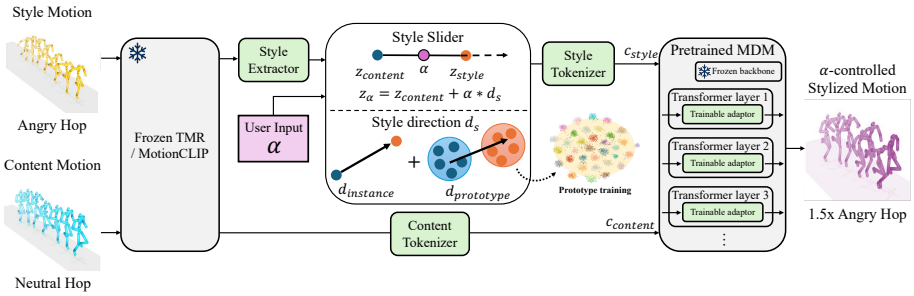

Fig. 2: Method pipeline. Given endpoint motions, we compute a style direction in latent space, sample an intensity value, build a style condition token, and train a diffusion denoiser with endpoint-aware supervision and latent intensity regularization.

MotionCLIP [34] aligns motion representations with CLIP features to enable text-driven motion editing and retrieval. Similarly, TMR [29], building on text–motion representation learning such as TEMOS [28], learns retrieval-oriented text–motion embeddings that capture semantic relationships between motions and language descriptions. Such embedding spaces provide a natural basis for defining semantic directions or conditioning signals for controllable motion generation. We use a TMR-based motion encoder because it provides a frozen, motion-level semantic space with stable retrieval behavior across different actions and styles. Our contribution is not to introduce a new representation model; instead, we use this representation to define endpoint style directions, train a controllable diffusion adaptor, and evaluate whether generated motions move predictably along the intended direction.

Another line of work studies actor-specific or persona-oriented motion generation. Persona-based models aim to capture performer-specific motion characteristics and enable personalized motion synthesis [20]. Datasets such as PerMo highlight the importance of modeling stylistic variation across actors and performing styles. Our work is complementary to these approaches. Rather than focusing solely on personalized motion generation, we study the problem of **continuous style intensity scaling**. To evaluate this property, we conduct experiments across multiple style datasets [20, 23, 39] and further introduce a real-capture "over-reaction" split to evaluate extrapolation beyond observed style endpoints. This setup emphasizes that controllable stylization should be assessed as a continuous and extrapolatable property rather than a single-point style transfer result.

3 Method

3.1 Problem Formulation

As illustrated in Figure 2, we study motion style control under an *endpoint-supervised* setting. Let $m_c \in \mathbb{R}^{T\times D}$ denote a neutral or content motion, and $m_s \in$

$\mathbb{R}^{T\times D}$ denote its stylized endpoint corresponding to the same content instance. Training data provides only such endpoint pairs (m_c, m_s), while intermediate-intensity motions are typically unavailable.

Our goal is to learn a model that generates motion $\hat{m}_\alpha$ for a continuous intensity parameter $\alpha \geq 0$ such that

$$\begin{aligned} \hat{m}_{\alpha=0} &\approx m_c, \\ \hat{m}_{\alpha=1} &\approx m_s, \end{aligned} \tag{1}$$

where $\hat{m}_\alpha$ changes style smoothly and monotonically with α. Importantly, α represents a *relative control coordinate* rather than an absolute style scale. This design reflects practical animation workflows, where style strength is adjusted incrementally rather than specified by a fixed universal unit. To implement this control mechanism, we build on a pretrained MDM diffusion model [35] and introduce conditioning and training strategies that enable continuous style modulation while preserving the underlying motion content.

3.2 Style Direction in Latent Space

To represent style variation in a structured manner, we operate in a learned motion embedding space. Let $E_s(\cdot)$ denote a frozen style encoder (a TMR-based head [29] in our implementation) that maps a motion sequence to a normalized embedding:

$$z = E_s(m), \qquad z \in \mathbb{R}^d, \qquad \|z\|_2 = 1. \tag{2}$$

Working in this latent space allows style manipulation to occur along semantically meaningful directions rather than directly modifying raw joint coordinates. The normalization stabilizes direction scale across different motions, so the scalar α controls movement along the style direction rather than being dominated by embedding magnitude.

For each training pair, we compute the embeddings of the content and stylized endpoints:

$$z_c = E_s(m_c), \qquad z_s = E_s(m_s). \tag{3}$$

These two points define a local style axis associated with the pair. We estimate a style direction d_s using one of three strategies:

$$d_{\text{instance}} = z_s - z_c, \qquad d_{\text{prototype}} = \mu_s - \mu_c, \tag{4}$$

$$d_s = \begin{cases} d_{\text{instance}}, & \text{instance mode}, \\ d_{\text{prototype}}, & \text{prototype mode}, \\ (1-\beta)d_{\text{prototype}} + \beta d_{\text{instance}}, & \text{hybrid mode}. \end{cases} \tag{5}$$

Here μ_s and μ_c denote style and content/neutral prototypes computed from the training data, and $\beta \in [0, 1]$ balances global and instance-specific style directions.

3.3 Condition Construction

To preserve motion content during stylization, we explicitly provide a content-motion condition to the diffusion model. Let $E_c(\cdot)$ denote a frozen motion encoder that extracts a content representation:

$$h_c = E_c(m_c), \qquad h_c \in \mathbb{R}^{d_c}. \tag{6}$$

In our implementation, E_c corresponds to the TMR motion encoder with masked mean pooling over motion tokens (excluding the CLS token). We project this representation into a diffusion conditioning token:

$$c_{\text{content}} = P_c(h_c) \in \mathbb{R}^{1\times d}, \tag{7}$$

where P_c is a learnable projection module. The encoder E_c remains frozen to maintain stable semantics, while P_c and lightweight MDM adaptor layers are trained to inject content and style conditions into the pretrained denoiser.

Given an intensity value α, we construct the style condition by moving along the style direction:

$$z_\alpha = z_c + \alpha d_s. \tag{8}$$

This equation implements the *style slider*: increasing α moves the embedding along the style axis, continuously strengthening the stylization relative to the content anchor z_c. In implementation variants, we optionally normalize the condition vector or direction for numerical stability. The resulting vector is projected into a style conditioning token c_{style}. Both content and style tokens are then provided to the MDM denoiser:

$$\hat{x}_\theta^0(x_t, t \mid c_{\text{content}}, c_{\text{style}}, c_{\text{text}}). \tag{9}$$

Text conditioning c_{text} is optional and disabled in our default configuration.

During training, we sample the intensity value α using an endpoint-aware scheme:

$$\alpha = \begin{cases} 0, & \text{with probability } p_0, \\ 1, & \text{with probability } p_1, \\ u,\ u \sim \mathcal{U}(0, \alpha_{\max}), & \text{otherwise.} \end{cases} \tag{10}$$

This sampling strategy increases the frequency of endpoint supervision while still exposing the model to intermediate intensity values, improving the reliability of the slider behavior.

3.4 Training Objective

For each sampled α, we define the diffusion target motion as

$$m_\alpha^\star = \begin{cases} m_c, & \alpha = 0, \\ m_s, & \alpha > 0. \end{cases}$$

The standard diffusion denoising objective is then

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{m^\star_\alpha, t, \epsilon}\left[\left\|x^0 - \hat{x}^0_\theta(x_t, t, c_{\text{content}}, c_{\text{style}}, c_{\text{text}})\right\|_2^2\right]. \tag{11}$$

To maintain strong endpoint fidelity, we apply different weights to endpoint and mid-range samples: endpoints use weight 1, while intermediate samples are down-weighted by w_{mid}. Because no real midpoint motions are available, intermediate-α samples are not treated as hard motion-space midpoint targets; instead, the diffusion loss preserves endpoint realism while the latent losses below shape the generated style trajectory.

While the diffusion loss ensures motion realism and endpoint reconstruction, it does not explicitly enforce consistent behavior along the style axis. We therefore introduce two additional regularization terms.

First, we encourage linear scaling of style strength with respect to α. Using the endpoint direction $\Delta = z_s - z_c$, we define a style projection function

$$g(m) = \left\langle E_s(m) - z_c, \frac{\Delta}{\|\Delta\|_2} \right\rangle. \tag{12}$$

For generated motion $\hat{m}_\alpha$, the linearity constraint becomes

$$\mathcal{L}_{\text{lin}} = \left(g(\hat{m}_\alpha) - \alpha g(m_s)\right)^2. \tag{13}$$

Second, we enforce monotonicity along the style axis. Given two sampled intensities $\alpha_2 > \alpha$, we penalize violations of increasing style strength:

$$\mathcal{L}_{\text{mono}} = \max(0, \gamma - (g(\hat{m}_{\alpha_2}) - g(\hat{m}_\alpha))). \tag{14}$$

Here $\gamma \geq 0$ is a margin encouraging separation between different intensity levels. This loss discourages cases where larger α unexpectedly produces weaker stylization.

The final objective combines these components:

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_{\text{lin}}\mathcal{L}_{\text{lin}} + \lambda_{\text{mono}}\mathcal{L}_{\text{mono}}. \tag{15}$$

In practice, $\mathcal{L}_{\text{lin}}$ and $\mathcal{L}_{\text{mono}}$ are applied to non-endpoint sampled α values, including extrapolation samples when $\alpha > 1$. Together, $\mathcal{L}_{\text{diff}}$ maintains motion realism, while the additional regularizers shape the controllability of the style slider.

3.5 Training and Inference

Training uses endpoint pairs with sampled α values and chosen direction modes (instance, prototype, or hybrid). Model parameters are optimized with AdamW starting from pretrained MDM weights. In our main configuration, we use adaptor-only finetuning with motion-conditioned content input and the text branch disabled.

Table 1: Dataset summary. Statistics from current canonical splits and our over-reaction extrapolation pairs used in this paper.

Dataset	#Styles	#Contents	#Clips	Protocol Use
PerMo	33	10	13,220	main + ablation
Bandai-Namco	14	1	175	main
Xia	7	5	572	main
Ours (over-reaction)	2	5	188	extrapolation-only

At inference time, given a pair (m_c, m_s) and a user-selected intensity α, we compute the corresponding style direction, construct the conditioning token, and sample motions from the diffusion model. The output duration follows the content motion mask; sequences are padded or truncated only for batching, and evaluation is performed on the valid frames. By sweeping α from 0 to values above 1, the system produces both interpolation and extrapolation trajectories along the learned style axis.

4 Experiments

4.1 Datasets

We evaluate on three benchmark style-motion datasets: PerMo [20], Bandai-Namco [23], and Xia [39] (Table 1). All data are converted to the same feature representation as proposed in [9], and endpoint pairs (m_c, m_s) are built where m_c is neutral/content and m_s is stylized.

For extrapolation, we build a game-industry-oriented motion set captured with professional actors, following a protocol and style-definition process similar to Bandai-Namco. During capture, actors are explicitly instructed to perform over-reactions to simulate intensified style motions. We use this set as matched triplets (m_c, m_s, m_{s_2}), where m_{s_2} is a stronger real capture of the same content-style identity. These clips are test-only and excluded from training. We retarget raw captures to the same skeleton/feature space and keep only endpoint-consistent valid triplets after quality-control filtering.

Throughout the experiment, we use canonical train/val/test splits and evaluate on test pairs only.

4.2 Implementation Details

All methods are evaluated with the same canonical pair protocol and the same intensity set $\alpha \in \{0, 0.5, 1.0, 1.5, 2.0\}$ (or ablation-specific subsets). Unified metrics are computed with a shared evaluator using a frozen TMR style encoder head as the feature space.

For our method, default main settings are: instance direction ($d_s = d_{\text{instance}}$), adaptor-only finetuning, pooled motion-content conditioning, and text branch

disabled. Prototype construction details are provided in the supplementary material.

Baselines are evaluated on the same pair splits and alphas, using their own trained checkpoints and the same post-hoc evaluator. This keeps the metric extractor and prototype construction identical across methods.

4.3 Baselines and Variants

We compare against DeepMotionEditing (Aberman et al.) [1] and MCM-LDM [33]. Because these methods expose different controls, we harmonize inputs to endpoint pairs and evaluate with the same pair list, alphas, and feature-space metrics. In particular, MCM-LDM is a strong latent-diffusion style-transfer baseline, but it does not explicitly optimize an endpoint-anchored style direction or a scalar intensity-control objective. This distinction is central to our evaluation: we compare not only endpoint transfer quality, but also whether a method responds predictably as α changes. Additional evaluation on the PerMo dataset with MoST [21], SMooDi [47], MotionCLIP [34], and the HumanML3D T2M evaluator [9] are provided in the supplementary material.

For ablation analysis, we evaluate:

- Direction mode: prototype/instance/hybrid,
- Objective components: with/without $\mathcal{L}_{\text{lin}}$ and $\mathcal{L}_{\text{mono}}$.

4.4 Metrics

We evaluate:

- **Content preservation**: Content Recognition Accuracy (CRA), using nearest prototype in content embedding space.
- **Style fidelity**: Style Recognition Accuracy (SRA), using nearest prototype in style embedding space.
- **Motion realism**: Fréchet Motion Distance (FMD) between generated and target embedding distributions.
- **Control quality**: style linearity error and monotonicity violation from $g(\hat{m}_\alpha)$ trajectories.
- **Extrapolation**: ExtraErr against held-out over-reaction motion m_{s_2} at high alpha.

Controllability metrics. For each sample, we evaluate $\alpha \in \{0, 0.5, 1.0, 1.5, 2.0\}$ and compute style-strength trajectory $g(\hat{m}_\alpha)$. We report:

$$\text{Mono} = \frac{1}{K-1} \sum_{k=1}^{K-1} \mathbb{I}\big[g(\hat{m}_{\alpha_{k+1}}) \geq g(\hat{m}_{\alpha_k})\big] , \tag{16}$$

$$\text{MonoViol} = \frac{1}{K-1} \sum_{k=1}^{K-1} \max\big(0, g(\hat{m}_{\alpha_k}) - g(\hat{m}_{\alpha_{k+1}})\big) , \tag{17}$$

Table 2: Main quantitative results. FMD/CRA/SRA/MonoViol/LinErr are averaged over PerMo/Bandai-Namco/Xia; ExtraErr is evaluated on the over-reaction set.

Method	FMD ↓	CRA ↑	SRA ↑	MonoViol ↓	LinErr ↓	ExtraErr ↓
DeepMotionEditing [1]	0.588	0.290	0.060	0.25	0.326	1.196
MCM-LDM [33]	**0.381**	**0.808**	0.265	0.11	0.358	1.174
Ours	0.457	0.606	**0.295**	**0.05**	**0.280**	**1.109**

$$\text{LinErr} = \frac{1}{K}\sum_{k=1}^{K} \left| g(\hat{m}_{\alpha_k}) - \alpha_k g(m_s) \right|. \tag{18}$$

In tables, we report monotonicity violation (`MonoViol`), i.e., averaged positive decreases in g across consecutive alphas.

Extrapolation metric. At high intensity, we compare $\hat{m}_{2.0}$ with held-out m_{s_2} in the same embedding space:

$$\text{ExtraErr} = \left\| \phi(\hat{m}_{2.0}) - \phi(m_{s_2}) \right\|_2, \tag{19}$$

where $\phi(\cdot)$ is the frozen TMR style feature extractor.

Standard transfer metrics. Following prior style-transfer evaluations [33], we report FMD, CRA, and SRA for compatibility with existing baselines.

4.5 Main Quantitative Results

Table 2 shows the global comparison. Our method gives the best control-related behavior: highest SRA (0.295), lowest MonoViol (0.05), lowest LinErr (0.280), and lowest ExtraErr (1.109). MCM-LDM remains strong on FMD and CRA, indicating a realism/content advantage under the current setting, while our method better preserves intensity ordering and style-strength controllability.

Table 3 shows the per-dataset behavior on the three main style datasets. On PerMo, our SRA and monotonicity are stronger than MCM-LDM (SRA: 0.140 vs. 0.080; MonoViol: 0.03 vs. 0.22). On Bandai-Namco and Xia, our method keeps lower LinErr; MonoViol is lower on Bandai-Namco and slightly higher on Xia, indicating that controllability gains vary with dataset style geometry.

4.6 Extrapolation to 2× Style Targets

We evaluate out-of-range behavior on our over-reaction captures by comparing generated $\hat{m}_{2.0}$ against real m_{s_2} motions. This isolates whether the model genuinely learns a style-intensity axis beyond trained endpoints. On the over-reaction dataset, ExtraErr is 1.109 (ours), 1.196 (DeepMotionEditing), and 1.174 (MCM-LDM), indicating that ours is closest to the real stronger-style targets.

Table 3: Per-dataset results. Each cell is *Ours/MCM-LDM*.

Dataset	FMD ↓	CRA ↑	SRA ↑	MonoViol ↓	LinErr ↓
PerMo	0.531 / 0.474	0.700 / 0.880	0.140 / 0.080	0.03 / 0.22	0.421 / 0.464
Bandai-Namco	0.143 / 0.237	0.714 / 1.000	0.714 / 0.429	0.01 / 0.06	0.259 / 0.368
Xia	0.279 / 0.346	0.108 / 0.351	0.405 / 0.216	0.07 / 0.05	0.261 / 0.407

Table 4: Ablation study (PerMo). Effect of direction mode and loss terms.

Variant	$\mathcal{L}_{\text{lin}}$	$\mathcal{L}_{\text{mono}}$	Direction	FMD ↓	CRA ↑	SRA ↑	MonoViol ↓	LinErr ↓
Ours	Yes	Yes	instance	**0.531**	**0.70**	**0.14**	0.03	**0.421**
Dir=proto	Yes	Yes	proto	0.584	0.60	0.06	0.04	0.513
Dir=hybrid ($\beta = 0.3$)	Yes	Yes	hybrid	0.588	**0.70**	0.12	0.03	0.509
w/o $\mathcal{L}_{\text{lin}}$	No	Yes	instance	0.605	0.54	0.06	0.03	0.422
w/o $\mathcal{L}_{\text{mono}}$	Yes	No	instance	0.545	0.56	0.04	0.03	0.434

4.7 Ablation Studies

Table 4 shows the effect of direction mode and loss terms on PerMo. The instance direction gives the best FMD, SRA, and LinErr in this setting, while the hybrid direction remains competitive on CRA and MonoViol. Removing either regularizer weakens controllability: without $\mathcal{L}_{\text{lin}}$, FMD/CRA/SRA degrade; without $\mathcal{L}_{\text{mono}}$, CRA/SRA also drop.

4.8 Qualitative Results

Figure 3 provides side-by-side visualizations at matched camera/view settings. We observe that our method produces clearer intensity separation while preserving motion identity, especially in difficult cases with different motion amplitudes (e.g., kick vs. hop). This visual trend is consistent with the higher SRA and lower LinErr/MonoViol in Tables 2 and 3.

Figure 4 illustrates a broader use case beyond the strict paired evaluation setting. We do not claim fully general text- or arbitrary-reference style transfer as the main benchmark objective; rather, this example shows that the reference-based slider can produce meaningful cross-content behavior, supporting its practical use as an animation editing control.

4.9 User Study

We conducted a Likert-scale user study with 11 university participants and no monetary compensation. The study focused on PerMo and compared our method with MCM-LDM using 12 questionnaire sections (6 motion cases $\times$ 2 methods). Each section showed one triplet generated from the same content–style pair at $\alpha \in \{0.5, 1.0, 1.5\}$. The three videos were anonymized as A/B/C. The A/B/C-to-α assignment was randomized per section, and participants were not told which

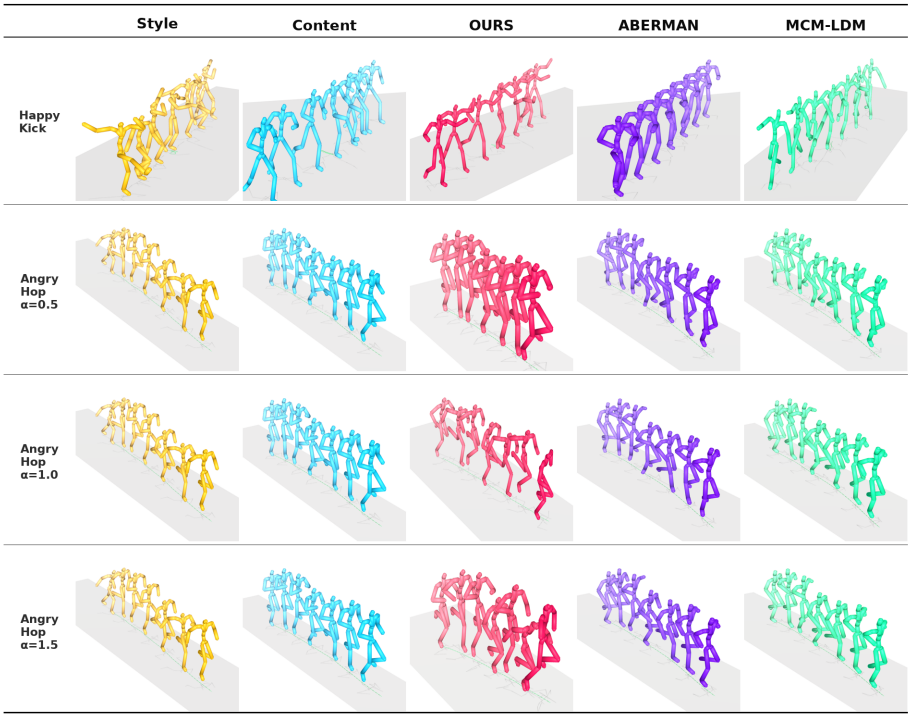

Fig. 3: Qualitative comparison. Columns show style reference, content reference, and generated results from our method and baselines under a unified camera and rendering configuration.

method generated each triplet. Participants answered seven questions: stylization strength for A/B/C, naturalness for A/B/C, and one overall question on whether the style-intensity differences were distinguishable. All ratings used a 7-point Likert scale, where larger values indicate stronger style, higher naturalness, or clearer distinguishability. For analysis, A/B/C were remapped to true low/mid/high α. Detailed user study protocols can be found in the supplementary material.

Table 5 summarizes results on matched sections ($n = 6$). Our method shows consistently higher stylization scores at all intensity levels (low: 3.92 vs. 3.56, mid: 4.61 vs. 3.59, high: 5.18 vs. 3.46) and much higher distinguishability (4.92 vs. 2.36). Naturalness is comparable between methods (ours: 4.64/4.83/4.53, MCM-LDM: 4.74/4.71/4.71 for low/mid/high). For perceived monotonic control, we count a motion case as a pass when the mean participant ratings satisfy style_low $<$ style_mid $<$ style_high after remapping the shuffled A/B/C videos to their true alpha order. As the result, ours achieves a higher pass rate (0.50 vs. 0.00).

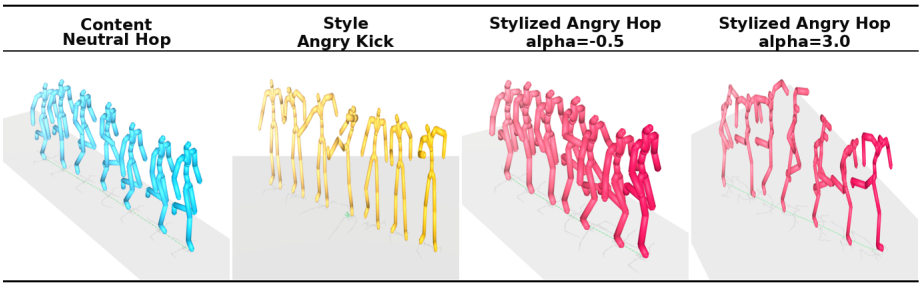

Fig. 4: Cross-content stylization example. Although our quantitative benchmark uses endpoint-paired motions to keep evaluation clean, the learned style direction can also be applied qualitatively across different content motions. Here, a neutral hop is edited using an angry kick reference; negative and large positive α values modulate the output along this reference direction while retaining the hopping content.

Table 5: User study summary (6 matched motion cases, 11 participants). Scores are 1–7 Likert means of low/mid/high (L/M/H) intensity per section.

Method	Distinguish ↑	Stylization (L/M/H) ↑	Naturalness (L/M/H) ↑
Ours	**4.92**	**3.92** / **4.61** / **5.18**	4.64 / **4.83** / 4.53
MCM-LDM	2.36	3.56 / 3.59 / 3.46	**4.74** / 4.71 / **4.71**

Paired sign tests show the same trend but limited statistical significance due to small sample size: Δstyle_high $= +1.73$ ($p = 0.2188$), Δdistinguishability $= +2.56$ ($p = 0.0625$), and monotonic-pass improvement $= +0.50$ ($p = 0.2500$).

5 Discussion

Our results should be interpreted with the task definition in mind: the goal is not only to produce one stylized sample, but to provide a predictable editing axis for style strength. For this setting, SRA, LinErr, MonoViol, and ExtraErr are direct measures of whether the slider follows the requested style direction and intensity ordering. CRA and FMD remain important supporting metrics for content preservation and motion realism, but they do not by themselves measure whether the user can reliably ask for "less", "more", or extrapolated style.

This distinction is important in practical animation workflows. Artists frequently adjust style strength iteratively, and a method that generates plausible individual samples but responds inconsistently to the control value is difficult to use as an editing interface. Our results therefore support the central hypothesis that style intensity should be treated as a relative control coordinate rather than a fixed global scale. The cross-content result in Figure 4 further suggests that the same reference-based control signal can be useful beyond exactly paired motions, while our quantitative evaluation keeps the endpoint-paired setting to ensure clean, reproducible measurement.

The over-reaction benchmark is useful beyond headline scores. It separates interpolation quality from true out-of-range behavior by evaluating against real stronger-style targets m_{s_2}. In this setting, methods can have similar in-range style recognition but different extrapolation error and monotonic behavior, indicating that endpoint supervision alone is insufficient unless the conditioning geometry is well constrained. The lower ExtraErr of our method suggests that learning a consistent style direction not only improves interpolation but also helps maintain a meaningful trajectory when extrapolating beyond the observed endpoints.

Limitations. First, controllability depends on the quality of endpoint pairs and the style embedding calibration used by both training regularizers and evaluation. Noisy or weakly matched pairs reduce direction quality. Second, large-α generation may still produce motion artifacts for rare styles or sparse actors. A possible remedy is to add kinematic or physical constraints (e.g., foot-contact consistency and motion smoothness penalties) and curriculum training that gradually increases the sampled $\alpha_{\max}$. Third, our quantitative benchmark focuses on reference-based endpoint control; extending the slider to unseen-style or text-guided control is an important direction for future work.

6 Conclusion

We presented Motion Style Slider, an endpoint-supervised framework for continuous style-intensity control in motion diffusion. The core design combines latent style-direction conditioning with a scalar control parameter and explicit regularization for linearity and monotonicity, enabling controllable interpolation and extrapolation without intermediate-intensity supervision.

Across the PerMo, Bandai-Namco, and Xia motion dataset, our method improves control-oriented behavior (style fidelity and slider consistency) while remaining competitive on realism/content metrics. We further introduced an over-reaction extrapolation benchmark that evaluates against real stronger-style motions, and used it to measure out-of-range performance directly rather than only in-range transfer quality.

Overall, these results support the view that practical motion stylization needs a reliable relative control axis rather than a single fixed style strength. Future work includes stronger biomechanical constraints for high-α stability, improved cross-dataset style alignment, and larger user studies.

Acknowledgements

We thank Osaka Cygames' Motion Capture Studio for capturing the motion dataset and for their support throughout the project.

References

1. Aberman, K., Weng, Y., Lischinski, D., Cohen-Or, D., Chen, B.: Unpaired motion style transfer from video to animation. ACM Trans. Graph. **39**(4) (Aug 2020). https://doi.org/10.1145/3386569.3392469
2. Cai, Y., Wu, Y., Li, K., Zhou, Y., Zheng, B., Liu, H.: Flooddiffusion: Tailored diffusion forcing for streaming motion generation (2026), https://arxiv.org/abs/2512.03520
3. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8185–8194 (2020). https://doi.org/10.1109/CVPR42600.2020.00821
4. Dai, W., Chen, L.H., Wang, J., Liu, J., Dai, B., Tang, Y.: Motionlcm: Real-time controllable motion generation via latent consistency model. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XVI. p. 390–408. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-72640-8_22, https://doi.org/10.1007/978-3-031-72640-8_22
5. Du, H., Herrmann, E., Sprenger, J., Fischer, K., Slusallek, P.: Stylistic locomotion modeling and synthesis using variational generative models. In: Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games. MIG '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3359566.3360083
6. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2414–2423 (2016). https://doi.org/10.1109/CVPR.2016.265
7. Geng, Z., Han, C., Hayder, Z., Liu, J., Shah, M., Mian, A.: Text-guided 3d human motion generation with keyframe-based parallel skip transformer (2024), https://arxiv.org/abs/2405.15439
8. 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)
9. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from texts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5152–5161 (2022)
10. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: European Conference on Computer Vision (ECCV). pp. 580–597 (2022)
11. Guo, Z., Lee, Y.Y., Liu, J., Ben-Shabat, Y., Zordan, V., Kapadia, M.: Stylemotif: Multi-modal motion stylization using style-content cross fusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13349–13359 (2025)
12. Holden, D., Saito, J., Komura, T.: A deep learning framework for character motion synthesis and editing. ACM Trans. Graph. **35**(4) (Jul 2016). https://doi.org/10.1145/2897824.2925975
13. Holden, D., Saito, J., Komura, T., Joyce, T.: Learning motion manifolds with convolutional autoencoders. In: SIGGRAPH Asia 2015 Technical Briefs. SA '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2820903.2820918

14. Hu, L., Zhang, Z., Ye, Y., Xu, Y., Xia, S.: Diffusion-based human motion style transfer with semantic guidance. In: Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation. p. 1–12. SCA '24, Eurographics Association, Goslar, DEU (2024). https://doi.org/10.1111/cgf.15169, https://doi.org/10.1111/cgf.15169
15. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 1510–1519 (2017). https://doi.org/10.1109/ICCV.2017.167
16. Jang, D.K., Park, S., Lee, S.H.: Motion puzzle: Arbitrary motion style transfer by body part. ACM Trans. Graph. **41**(3) (Jun 2022). https://doi.org/10.1145/3516429, https://doi.org/10.1145/3516429
17. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: human motion as a foreign language. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23, Curran Associates Inc., Red Hook, NY, USA (2023)
18. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 694–711. Springer International Publishing, Cham (2016)
19. Karunratanakul, K., Preechakul, K., Suwajanakorn, S., Tang, S.: Guided motion diffusion for controllable human motion synthesis. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2151–2162 (2023). https://doi.org/10.1109/ICCV51070.2023.00205
20. Kim, B., Jeong, H.I., Sung, J., Cheng, Y., Lee, J., Chang, J.Y., Choi, S.I., Choi, Y., Shin, S., Kim, J., Chang, H.J.: Personabooth: Personalized text-to-motion generation. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22756–22765 (2025). https://doi.org/10.1109/CVPR52734.2025.02119
21. Kim, B., Kim, J., Chang, H.J., Choi, J.Y.: Most: Motion style transformer between diverse action contents. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1705–1714 (2024). https://doi.org/10.1109/CVPR52733.2024.00168
22. Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada. pp. 10236–10245 (2018), https://proceedings.neurips.cc/paper/2018/hash/d139db6a236200b21cc7f752979132d0-Abstract.html
23. Kobayashi, M., Liao, C.C., Inoue, K., Yojima, S., Takahashi, M.: Motion capture dataset for practical use of ai-based motion editing and stylization (2023), https://arxiv.org/abs/2306.08861
24. Liao, C.C., Kim, J.H., Koike, H., Hwang, D.H.: Content-preserving motion stylization using variational autoencoder. In: ACM SIGGRAPH 2023 Posters. SIGGRAPH '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3588028.3603679, https://doi.org/10.1145/3588028.3603679
25. Liu, M., Liu, Y., Krishnan, G., Bayer, K.S., Zhou, B.: T2M-X: Learning expressive text-to-motion generation from partially annotated data (2024), https://arxiv.org/abs/2409.13251

26. Mason, I., Starke, S., Komura, T.: Real-time style modelling of human locomotion via feature-wise transformations and local motion phases. Proc. ACM Comput. Graph. Interact. Tech. **5**(1) (May 2022). https://doi.org/10.1145/3522618, https://doi.org/10.1145/3522618
27. Park, S., Jang, D.K., Lee, S.H.: Diverse motion stylization for multiple style domains via spatial-temporal graph-based generative model. Proc. ACM Comput. Graph. Interact. Tech. **4**(3) (Sep 2021). https://doi.org/10.1145/3480145, https://doi.org/10.1145/3480145
28. Petrovich, M., Black, M.J., Varol, G.: TEMOS: Generating diverse human motions from textual descriptions. In: Computer Vision – ECCV 2022. pp. 480–497. Springer (2022)
29. 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)
30. Qian, Z., Xiao, Z., Jin, X., Yang, D., Li, M., Wu, Z., Kou, D., Zhai, P., Zhang, L.: Umsd:high realism motion style transfer via unified mamba-based diffusion. In: Proceedings of the 33rd ACM International Conference on Multimedia. p. 7424–7433. MM '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3746027.3754873, https://doi.org/10.1145/3746027.3754873
31. Sawdayee, H., Guo, C., Tevet, G., Zhou, B., Wang, J., Bermano, A.H.: Dance like a chicken: Low-rank stylization for human motion diffusion. In: Computer Graphics Forum. p. e70365. Wiley Online Library (2026)
32. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. In: International Conference on Learning Representations (ICLR) (2024)
33. 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 (CVPR). pp. 821–830 (2024)
34. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: Motionclip: Exposing human motion generation to clip space. In: Computer Vision – ECCV 2022. pp. 358–374. Springer (2022)
35. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. In: International Conference on Learning Representations (ICLR) (2023)
36. Wang, H., Du, D., li, J., Ji, W., Yu, L.: A cyclic consistency motion style transfer method combined with kinematic constraints. Journal of Sensors **2021**, 1–17 (06 2021). https://doi.org/10.1155/2021/5548614
37. Wang, Y., Huang, D., Zhang, Y., Ouyang, W., Jiao, J., Feng, X., Zhou, Y., Wan, P., Tang, S., Xu, D.: MotionGPT-2: A general-purpose motion-language model for motion generation and understanding (2024), https://arxiv.org/abs/2410.21747
38. Wen, Y.H., Yang, Z., Fu, H., Gao, L., Sun, Y., Liu, Y.J.: Autoregressive stylized motion synthesis with generative flow. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13607–13607 (2021). https://doi.org/10.1109/CVPR46437.2021.01340
39. Xia, S., Wang, C., Chai, J., Hodgins, J.: Realtime style transfer for unlabeled heterogeneous human motion. ACM Trans. Graph. **34**(4) (Jul 2015). https://doi.org/10.1145/2766999

40. Xie, Y., Jampani, V., Zhong, L., Sun, D., Jiang, H.: Omnicontrol: Control any joint at any time for human motion generation. In: The Twelfth International Conference on Learning Representations (2024)
41. Yang, Z., Luo, Z., Wang, Z., Liu, Y.: Arbitrary motion style transfer via contrastive learning. Applied Sciences **15**(4), 1817 (2025). https://doi.org/10.3390/app15041817
42. Zhang, J., Zhang, Y., Cun, X., Zhang, Y., Zhao, H., Lu, H., Shen, X., Ying, S.: Generating human motion from textual descriptions with discrete representations. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14730–14740 (2023). https://doi.org/10.1109/CVPR52729.2023.01415
43. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. IEEE Trans. Pattern Anal. Mach. Intell. **46**(6), 4115–4128 (Jun 2024). https://doi.org/10.1109/TPAMI.2024.3355414
44. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: Remodiffuse: Retrieval-augmented motion diffusion model. In: ICCV. pp. 364–373 (2023), https://doi.org/10.1109/ICCV51070.2023.00040
45. Zhang, M., Jin, D., Gu, C., Hong, F., Cai, Z., Huang, J., Zhang, C., Guo, X., Yang, L., He, Y., Liu, Z.: Large motion model for unified multi-modal motion generation. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XIII. p. 397–421. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-72624-8_23
46. Zhang, Z., Liu, A., Reid, I., Hartley, R., Zhuang, B., Tang, H.: Motion mamba: Efficient and long sequence motion generation. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part I. p. 265–282. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-73232-4_15
47. Zhong, L., Xie, Y., Jampani, V., Sun, D., Jiang, H.: Smoodi: Stylized motion diffusion model. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part I. p. 405–421. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-73232-4_23