MorphoStyle: Motion Style Transfer with Morphology Control

Xin Feng
xfeng4@ed.ac.uk
Eleonora D'Arnese
eleonora.darnese@ed.ac.uk
Mohan Sridharan
m.sridharan@ed.ac.uk

Institute of Perception, Action, and Behavior
School of Informatics
University of Edinburgh
Edinburgh, UK

Abstract

Human motion may be viewed as a combination of action content, style, and body morphology. Existing motion style transfer methods transfer a reference style onto a content motion while assuming a canonical body, whereas shape-aware motion generators adapt motion to a target shape without explicit style control. This separation of motion style and shape (morphology) makes it difficult to generate stylized motions for non-canonical bodies; naively combining a style transfer module with a shape-aware generator often leaks action content from the style reference and disrupts shape-consistent kinematics. We present *MorphoStyle*, a framework for shape-aware motion style transfer that is built on a shape-conditioned Finite-Scalar-Quantization Variational Auto-Encoder (FSQ-VAE). The key contribution is to pose the desired style transfer as modular latent disentanglement comprising: (i) a contrastive style encoder that extracts content-decoupled style embeddings; (ii) a text-guided style-routing mechanism that locates style-relevant joints in a text-motion feature space; and (iii) a manifold preserving style modulator that injects discriminative style embeddings in content features as a temporally-gated low-rank offset. Extensive experiments on benchmark datasets demonstrate that MorphoStyle outperforms competing baselines in terms of both shape control and motion style transfer, while simultaneously providing quantitative shape control. For more details, please see project website: https://github.com/MorphoStyle.

1 Introduction

Human motion generation aims to synthesize realistic and natural human movements conditioned on various inputs, such as text descriptions, reference motion, audio, or body parameters. It is a core research topic in computer vision and graphics, and has extensive applications in gaming [17], film production [40], virtual reality [16], and robotics [33]. Realistic human motion is determined by three complementary factors: the content, *i.e.*, what action is performed, such as walking or sitting down; the style, *i.e.*, how the action is performed, such as a funny or childish gait; and the morphology, *e.g.*, who performs the action, such as a tall and lean adult or a short and heavy one. All three components are essential for realistic and personalized digital avatars across various applications, since identical content with mismatched style or morphology breaks the perception of realism. Despite rapid progress

in human motion generation based on deep generative models [4, 8, 9, 13, 34, 37], jointly controlling transfer of motion style and body shape remains a challenging open problem.

Most existing methods for motion synthesis focus on either motion style transfer or shape-aware motion generation. Motion style transfer methods aim to generate a novel motion that adapts the content obtained from a source to the style of a reference motion [7, 8, 10, 11, 14, 15, 19, 23, 24, 36, 39, 43]. Due to the difficulty of capturing paired stylistic motions, most recent methods start with pretrained latent-diffusion backbones [4, 37] and inject style through ControlNet-style adaptors [42, 43] or cross-modal alignment with foundation embeddings [6, 11]. While these methods produce visually realistic stylized motions, they have two main limitations. First, style features are fused with the content stream of a pretrained latent-diffusion space without explicit disentanglement of content and style, resulting in content leakage and unrealistic kinematics in the synthesized output [41]. Second, the transfer of motion style is not adapted to body shape, e.g., a canonical body is assumed and parameterized by the Skinned Multi-Person Linear (SMPL) model [26], producing unpleasant artifacts when used with non-canonical body morphologies. Meanwhile, shape-aware motion generation aims to control body shape in the motion generation pipeline [25, 32, 33]. One recent example, ShapeMyMove [25], employs a shape-conditioned Finite-Scalar-Quantization Variational Auto-Encoder (FSQ-VAE) [29] that first quantizes motions into discrete tokens and de-quantizes them in a shape-conditioned manner. This formulation does not support style control, and morphology-aware motion style transfer remains an open problem.

We describe *MorphoStyle*, a novel shape-aware motion style transfer framework that seeks to address the limitations of existing work. We introduce a content-decoupled '*style encoder*' that builds on a pretrained shape-conditioned FSQ-VAE generator and injects style embeddings into motion content. It does so through a contrastive learning-based style encoder that maps reference style motion to a content-decoupled latent feature space; and a manifold-preserving style modulator that injects the style embeddings into content features as a temporally-gated low-rank offset. Specifically, the contributions of this paper are:

- A unified framework for motion style transfer with morphology control. MorphoStyle builds on a frozen shape-conditioned FSQ-VAE, preserving its content and shape generation capability while introducing a dedicated branch for controllable style.
- A content-decoupled style injection mechanism that leverages a supervised contrastive encoder to learn discriminative style embeddings; a text-guided routing module to localize style-relevant joints; and a manifold-preserving style modulator to inject style embeddings into content features.
- Extensive experiments on HumanML3D and 100Style (benchmark) datasets to demonstrate better motion style transfer and morphology control, and significantly higher efficiency, compared with state-of-the-art diffusion-based baselines.

We begin with a description of related work (Section 2), followed by a description of our framework (Section 3), experimental evaluation (Section 4), and conclusions (Section 5).

2 Related Work

We motivate the proposed framework by reviewing related work in motion style transfer and shape-aware motion generation.

2.1 Motion Style Transfer

Motion style transfer aims to adapt motion content, *i.e.* the action performed, to a style reference, *i.e.* how the action is performed, to obtain a stylized output [2, 14, 19]. Early efforts were based on statistical or graph-based methods [3, 23, 39], but they provide limited expressivity in terms of the styles they can reproduce. Deep network models [7, 13, 21] provided better performance through their powerful ability to interpolate over the space of training samples. Inspired by image stylization, Holden *et al.* [14, 15] performed Gram-matrix optimization on a learned motion manifold, and Aberman *et al.* [2] introduced a Generative Adversarial Network model with Adaptive Instance Normalization (AdaIN) [18] that disentangles content and style without paired supervision. Subsequent methods have pursued temporally coherent style transfer through per-body-part graph-aware feature exchange [19], spatial-temporal graph generators [30], time-series Transformers [36], and pre-trained auto-encoder latent spaces [10]. Although these methods have improved the performance of motion style transfer, they still struggle with limited expressivity and realism in stylized motions, with the underlying motion generators being the main bottleneck.

Building on diffusion-based models for human motion generation [4, 37], SMooDi [43] employs a ControlNet-style adaptor [42] and a classifier-based style guidance to augment a pretrained latent diffusion backbone. More recently, StyleMotif [11] further extends motion stylization to multimodal style conditioning aligned with relevant AI models such as ImageBind [6] and MulSMo [24]. Despite their visual realism, these diffusion-based methods share two main limitations. First, style features are fused into the content stream of a pretrained latent diffusion space without explicit content-style disentanglement, which inevitably results in content leakage and unrealistic kinematics in the stylized output. Second, they address motion style transfer in isolation from body geometry, implicitly assuming a canonical body shape, thus leading to undesirable artifacts with non-canonical morphologies. In contrast, MorphoStyle injects style embeddings obtained from a latent disentangled feature space in a frozen shape-conditioned backbone to effectively achieve high-quality motion style transfer with reliable morphological control.

2.2 Shape-aware Motion Generation

Body shape is a critical yet often overlooked factor in human motion generation; individuals with different heights, limb proportions, and body morphology perform the same action with visibly different kinematics [25, 26]. Most human motion generators are trained on motions of a canonical SMPL body [26] and thus generate motion in a shape-agnostic manner [4, 8, 9, 34, 37]. This decreases the natural correlation between morphology and motion, as well as propagating artifacts to downstream tasks [1, 25, 38]. Some works have begun to inject shape information into the motion generation pipeline. For example, Physics-based controllers [32, 33] embed shape parameters into character simulation but rely on physical engines and reinforcement learning, and perform poorly when scaled to large motion datasets. Shape-aware pose models [5, 22] condition kinematic regressors on body parameters to mitigate self-intersection and contact errors. ShapeMyMove [25], designed for shape-aware text-to-motion generation, is most related to our work. It trains a shape-aware Finite-Scalar-Quantization model (FSQ-VAE) [29] that quantizes shape-canonical motions into discrete tokens, and de-quantizes them in a shape-conditioned manner, thereby reconstructing motions adapted to body morphology. Based on this shape-aware tokenizer, it employs a language model to jointly predict shape parameters and motion token indices to yield more realistic motions with plausible body shapes.

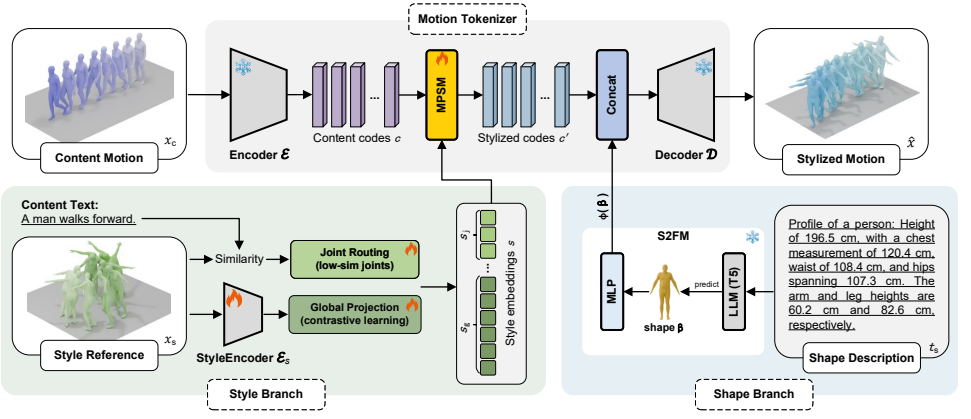

Figure 1: **Framework Overview.** Given a content motion x_c, a style reference x_s, and a textual body-shape description t_s, **MorphoStyle** synthesizes an output motion $\hat{x}$ that preserves the action of x_c, transfers the style of x_s, and matches the target body shape. A *frozen* shape-aware FSQ-VAE tokenizes x_c into content codes c and decodes $\hat{x}$ via a shape-conditioned decoder $\mathcal{D}$. A *trainable* style branch produces a global style embedding s_g and a joint-localized embedding s_j, fuses them into style code s, and injects s in c through a Manifold-Preserving Style Modulator (MPSM) as a temporally-gated low-rank residual. A *frozen* Shape-to-Feature module (S2FM) maps t_s to shape embedding $\phi(\beta)$ conditioning $\mathcal{D}$.

Although the aforementioned methods are content- and shape-driven, it is challenging to control the style of output motion since a naive combination of a style transfer mechanism and a shape-aware backbone interferes with the simultaneous modeling of style and shape. Our MorphoStyle model extends the shape-aware FSQ-VAE backbone of ShapeMyMove [25], but targets a fundamentally different objective of motion style transfer with morphology control. Instead of co-training the decoder to reconstruct stylistic motions, we freeze it and extract style embeddings in a discriminative latent feature space to achieve motion style transfer. This design choice enables MorphoStyle to inherit ShapeMyMove's shape control ability, while also supporting motion style transfer.

3 Method

This section describes the problem formulation and basic components of MorphoStyle (Section 3.1), and the key style disentanglement component and learning scheme (Section 3.2).

3.1 Framework Overview

Problem formulation. Given content motion $x_c \in \mathbb{R}^{T\times C}$ that specifies the target action, reference motion $x_s \in \mathbb{R}^{T\times C}$ that contains the target style, and a body shape $\beta \in \mathbb{R}^{10}$ parameterized by the Skinned Multi-Person Linear (SMPL) model [26], our goal is to synthesize an output motion $\hat{x} \in \mathbb{R}^{T\times C}$ that simultaneously satisfies three conditions: (i) the action content $\hat{x}$ matches x_c; (ii) the stylistic appearance of $\hat{x}$ matches that of x_s; (iii) the body morphology of $\hat{x}$ is faithful to β. Since we adopt the standard HumanML3D feature representations [8], per-frame feature dimension $C = 263$ and sequence length $T = 196$.

Latent disentanglement. MorphoStyle is built on the observation that different aspects of motion act on disjoint subspaces of a well-trained shape-aware motion auto-encoder. Specif-

ically, in a shape-aware FSQ-VAE [25, 29], content motion is encoded in a discrete code sequence $c \in \mathbb{R}^{T' \times D}$, where $T' = T/4$ is the temporally down-sampled length produced by the encoder's $4\times$ temporal compression, and code dimension $D = 512$. Body-shape β modulates de-quantization rather than the code sequence itself, and stylistic variation can be defined as a low-rank residual offset to c that does not change which codebook entries are selected. Latent disentanglement indicates that shape-aware motion style transfer can be captured by three independent modular components without shared parameters: (i) a frozen shape-aware backbone that is able to handle both content and shape; (ii) a lightweight content-decoupled style branch which stylizes the content code sequence in a low-rank manner; and (iii) a shape-to-feature module which projects textual shape description into latent shape features.

Our architecture. As presented in Figure 1, MorphoStyle consists of a frozen shape-aware FSQ-VAE backbone, a style disentanglement branch, and a shape control branch. The FSQ-VAE consisting of a content motion encoder $\mathcal{E}$, a Finite-Scalar-Quantizer, and a shape-conditioned decoder $\mathcal{D}$, is kept frozen throughout training. A contrastive style encoder $\mathcal{E}_s$ maps the reference style motion x_s to style embeddings $s \in \mathbb{R}^{256}$, while Manifold-Preserving Style Modulator (MPSM) fuses s into encoded content tokens c to yield stylized motion codes c'. The frozen decoder $\mathcal{D}$ then reconstructs c' to the motion space conditioned on body shape β. The shape control branch uses a pretrained Shape-to-Feature module (S2FM) inspired by ShapeMyMove [25], with a T5-based language model predicting SMPL model's β parameters from shape descriptions t_s and a Multi-layered Perceptron (MLP) model to extract shape features $\phi(\beta) = \text{S2FM}(t_s)$ from β. MorphoStyle's overall pipeline is thus:

$$\hat{x} \;=\; \mathcal{D}(\text{MPSM}(\mathcal{E}(x_c), s), \phi(\beta)) \tag{1}$$

Shape-aware FSQ-VAE backbone. MorphoStyle substantially extends the shape-aware FSQ-VAE introduced in ShapeMyMove [25], which is pretrained on HumanML3D [8] jointly with continuous SMPL [26] condition parameters, and serves as a foundation of our model. Specifically, the encoder $\mathcal{E}$ employs three Conv1d residual blocks with dilation, downsamples temporal length by a factor of four, and then feeds encoded motion features into an FSQ bottleneck with a 1000-entry codebook. The shape-aware decoder $\mathcal{D}$ concatenates each quantized content code $c \in \mathbb{R}^{T' \times D}$ with a 32-dimensional shape feature $\phi(\beta)$ obtained from the shape branch, and reconstructs motions through symmetric Conv1d upsampling. Importantly, such a backbone is able to effectively control two of three factors, content and shape, thus we further introduce and optimize a lightweight style branch with Manifold-Preserving Style Modulator to inject style information without content drift.

Shape-to-Feature Module. While FSQ-VAE takes $\phi(\beta)$ as its shape condition, the SMPL model's β parameters are inconvenient since body shape is usually described in natural language. Similar to ShapeMyMove [25], we introduce the Shape-to-Feature Module (S2FM) (see Figure 1) to convert textual shape description t_s into shape embeddings $\phi(\beta)$:

$$\phi(\beta) \;=\; \text{MLP}_\phi(\text{T5}(t_s)) \tag{2}$$

where the T5-based language model [35] obtains the 10-dimensional SMPL model's β parameters from t_s like ShapeMyMove [25], and a two-layer MLP then projects β into a 32-dimensional shape embedding that matches the shape-aware decoder $\mathcal{D}$'s input space.

3.2 Style Disentanglement

Contrastive global style embedding. Contrastive style encoder $\mathcal{E}_s$ aims to extract compact and content-decoupled style embeddings that summarize whole-body stylistic cues from a

reference motion. Specifically, $\mathcal{E}_s$ takes a style reference motion $x_s \in \mathbb{R}^{T\times C}$ as input. It consists of stacked Conv1d layers followed by adaptive average pooling, and a two-layer MLP, projecting pooled features into 256-dimensional style embeddings s_g. In addition, we also introduce style dropout with probability $p = 0.15$, which replaces style embeddings s with a null token to improve model's robustness. To prevent s from interfering with content features, which is the root cause of the content leakage observed in previous diffusion-based stylizers [11, 43], we introduce a supervised contrastive learning mechanism that pulls together embeddings of motions with the same style label, and pushes apart embeddings of different styles in the latent feature space. The supervised contrastive learning loss [20] is:

$$\mathcal{L}_{\text{cl}} = -\mathbb{E}_i \log \frac{\exp\big(\text{sim}(s_i, s_i^+)/\tau\big)}{\exp\big(\text{sim}(s_i, s_i^+)/\tau\big) + \sum_{s_n\in\mathcal{N}_i} \exp(\text{sim}(s_i, s_n)/\tau)} \tag{3}$$

where s_i^+ is a positive sample with the same style label as s_i; $\mathcal{N}_i$ is the set of negative samples, motions whose style labels differ from that of s_i in the mini-batch; $\text{sim}(\cdot,\cdot)$ denotes cosine similarity; and τ is a temperature hyperparameter. Contrastive supervision leads to effective disentanglement between content and style for whole-body motions in the output of $\mathcal{E}_s$.

Text-guided joint style routing. Although style embeddings s extract whole-body stylistic cues, they inevitably blend signals from joints that are mainly content-driven (*e.g.* the pelvis during walking) with joints that are mainly style-driven (*e.g.* the hands during a swaggering walk). To explicitly disentangle them within the body, our text-guided routing module leverages a pretrained text-motion alignment model [8], including a text encoder $\mathcal{T}$ and a motion encoder $\mathcal{M}$ jointly trained to align text and motion in a shared latent feature space. For each joint j of $J = 22$ body joints, we construct a joint-masked variant $\tilde{x}_s^j$ of style motion that only retains j's positional, rotational, and velocity channels. We then encode each masked motion using $\mathcal{M}$ and content text caption c_t using $\mathcal{T}$, and score the alignment of each joint with the caption through cosine similarity. The style-relevant joints are identified as the K (=8, set experimentally) lowest values of the alignment score:

$$\mathcal{S} = \underset{j\in[0,J-1]}{\arg\min}\, K \left\{\cos\big(\mathcal{M}(\tilde{x}_s^j), \mathcal{T}(c_t)\big)\right\} \tag{4}$$

Given the selected set $\mathcal{S}$ of style-relevant joints, we introduce per-joint kinematic features $\psi_j(x_s) \in \mathbb{R}^{T\times 12}$, where the 12 channels concatenate position, rotation and velocity channels defined by HumanML3D [8], through an MLP f_{j} with an additive joint-identity embedding:

$$s_j \;=\; f_{\text{j}}\left(\frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}}\big(\bar{\psi}_j(x_s) + u_j\big)\right) \tag{5}$$

where $\bar{\psi}_j(x_s) = \frac{1}{T}\sum_t \psi_j(x_s)[t,:]$ are time-averaged features and u_j is a learnable identity embedding that lets f_{j} tell joint-specific stylistic patterns apart. Finally, we fuse joint-level embeddings with global embeddings through a residual MLP to produce unified codes $s \in \mathbb{R}^{256}$ of style reference for MPSM's style modulation:

$$s \;=\; s_g + \alpha \cdot \text{MLP}([s_g\,; s_j]) \tag{6}$$

where $[\cdot\,;\cdot]$ represents channel-wise concatenation and $\alpha = 1$ is a residual scale. At inference, the final MLP layer is initialized with a small Gaussian standard deviation of 0.01 so that the routing residual is near zero at the resumed step, and α becomes a runtime knob that smoothly trades stylization strength against motion fidelity without retraining.

Intuitively, $\mathcal{M}(\tilde{x}_s^j)$ encodes isolated motion dynamics of joint j, and $\mathcal{T}(c_t)$ summarizes target action from content textual caption. Joints whose isolated motion is least aligned with

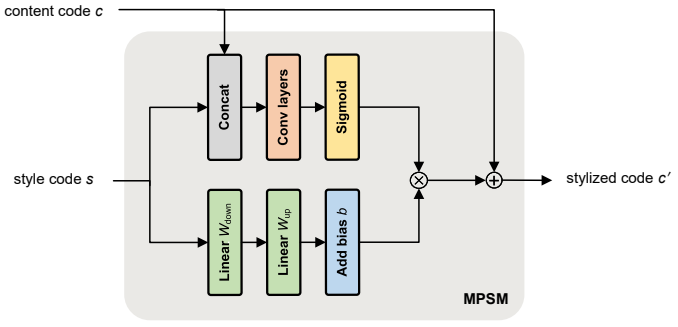

Figure 2: Structure of Manifold-Preserving Style Modulator (MPSM).

the content caption carry stylistic information that the caption does not describe, and are therefore the most informative for style transfer. Based on the selected set of stylistic joints $\mathcal{S}$, we project kinematic features [8] $\psi_j(x_s)$ through the layers of an MLP, and then fuse it with global style embeddings s_g to obtain stylistic motion features that contain the style cues of movements of style-bearing joints, while also mitigating content leakage.

Manifold-Preserving Style Modulator. Given content codes $c = \mathcal{E}(x_c) \in \mathbb{R}^{T' \times 512}$ and style embeddings s, as presented in Figure 2, MorphoStyle incorporates a Manifold-Preserving Style Modulator (MPSM) to fuse them into stylized codes c' that remain in the codebook neighborhood of the frozen decoder throughout the motion sequence. We design the MPSM with two main requirements: (i) style injection should be constrained to a low-dimensional subspace of each 512-dimensional code space so as not to change content information; (ii) per-frame intensity of style injection should adapt to local kinematic context so as not to corrupt physical constraints such as foot contacts. MPSM is thus composed of two main components: a low-rank additive offset and a learnable temporal gate.

Instead of using affine modulation schemes such as AdaIN [18], which multiply content features by style-conditioned scales, we project the style offset into a rank-r subspace using a bilinear bottleneck in our framework:

$$r(s) = W_{\text{up}} W_{\text{down}} s + b \tag{7}$$

where $W_{\text{down}} \in \mathbb{R}^{r \times 256}$ and $W_{\text{up}} \in \mathbb{R}^{512 \times r}$ are transformation matrices, $b \in \mathbb{R}^{512}$ is a learnable bias initialized to zero, and rank is empirically set to $r = 16$. Subsequently, we introduce a learnable temporal gate $g \in (0,1)$ calculated from both content codes c and fused style embeddings s to obtain per-frame intensity of style injection:

$$g = \sigma\big(f_{\text{g}}\big([c\,;\,s]\big)\big) \tag{8}$$

where $\sigma(\cdot)$ denotes the sigmoid function, $[\,\cdot\,;\,\cdot\,]$ denotes channel-wise concatenation, and f_{g} is a three-layer Conv1d block with Sigmoid Linear Unit (SiLU) activations. Stylized motion codes $c' \in \mathbb{R}^{T' \times 512}$ are obtained by adding a gated low-rank offset to content motion codes:

$$c'_t = c_t + g_t \cdot r(s), \qquad t = 1, \ldots, T' \tag{9}$$

3.3 Learning Scheme

MorphoStyle is trained using a loss function that combines content-style disentanglement ($\mathcal{L}_{\text{dis}}$) and shape control ($\mathcal{L}_{\text{shape}}$). All gradients are restricted to the trainable style branch; FSQ-VAE backbone and text-motion alignment model remain frozen. All loss weights (below) are fixed and specified in Section 4.1.

Content-style disentanglement loss $\mathcal{L}_{\text{dis}}$ seeks to inject style into content features without content leakage. It consists of three terms. The *reconstruction term* enforces fidelity of motion and code compared with the ground truth target x^*, with $c^* = \mathcal{E}(x^*)$, resulting in the stylized motion output being faithful to ground truth in motion-and latent code-space.

$$\mathcal{L}_{\text{rec}} = \|\hat{x} - x^*\|_1 + \lambda_{\text{code}} \|c' - c^*\|_1 \tag{10}$$

Code-level supervision keeps modulated codes c' within FSQ-VAE's codebook neighborhood, preventing style branch from causing drastic latent shifts.

The *style embedding term* aims to learn a discriminative feature space for reference style motion via supervised contrastive learning—$\mathcal{L}_{\text{cl}}$ in Equation 3—together with a content classifier loss $\mathcal{L}_{\text{cc}}$. In the latent feature space, $\mathcal{L}_{\text{cl}}$ pulls motion embeddings with the same style label together and pushes apart those with different styles, so encoded style embeddings s_g are without unnecessary content from the style reference. The content classifier loss $\mathcal{L}_{\text{cc}}$ prevents content leakage by employing a content classifier f_{ac} on stylized motion $\hat{x}$, ensuring that it belongs to the same content label y_{act} as the corresponding content motion.

$$\mathcal{L}_{\text{cc}} = \text{CE}(f_{\text{ac}}(\hat{x}), y_{\text{act}}) \tag{11}$$

where $\text{CE}(\cdot,\cdot)$ denotes the cross-entropy. The style embedding term is then defined as:

$$\mathcal{L}_{\text{emb}} = \lambda_{\text{cl}} \mathcal{L}_{\text{cl}} + \lambda_{\text{cc}} \mathcal{L}_{\text{cc}} \tag{12}$$

The *style classifier guidance term* uses the 47-class style classifier of SMooDi [43] f_{sm}, which is trained over the joint label space of HumanML3D [8] and 100Style [28], to provide direct supervision on the synthesized motion. Without explicit alignment to this classifier, even perceptually correct stylized motions may receive low style recognition accuracy due to subtle distributional shifts. We therefore introduce a training-time cross-entropy guidance that aligns $\hat{x}$ with the target style label y_{sty} through f_{sm}:

$$\mathcal{L}_{\text{cls}} = \lambda_{\text{cls}} \text{CE}\big(f_{\text{sm}}(\hat{x}), y_{\text{sty}}\big) \tag{13}$$

As a result, the disentanglement loss $\mathcal{L}_{\text{dis}}$ is given by:

$$\mathcal{L}_{\text{dis}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{emb}} + \mathcal{L}_{\text{cls}} \tag{14}$$

Shape control objective. Although the shape-conditioned decoder $\mathcal{D}$ is frozen, style injection can still interfere with shape signal in the final output. To enforce shape consistency between synthesized motions and shape descriptions, we leverage the parameter β of shape descriptions and introduce three shape supervision terms.

The *bone-length matching term* builds on ShapeMyMove [25] to minimize Mean-Squared Error (MSE) between bone length (BL) vectors of synthesized motion $\hat{x}$ and target shape x^*:

$$\mathcal{L}_{\text{bone}} = \text{MSE}(\text{BL}(\hat{x}), \text{BL}(x^*)) \tag{15}$$

The *direction-aware term* ensures that bone length changes occur along realistic directions, *e.g., generating longer legs and shorter arms than canonical shapes*. It seeks to avoid synthesizing motions with bones shifting in the opposite direction to prescribed ones.

$$\mathcal{L}_{\text{dir}} = 1 - \cos\left(\hat{x} - x^{\beta_t}, x^{\beta_t} - x^{\beta_n}\right) \tag{16}$$

where x^{β_t} and x^{β_n} are motions with target shape β_t and normalized shape β_n.

The *magnitude-bounding term* seeks to prevent shape shift from overshooting the ground truth shift beyond a tolerance margin γ:

$$\mathcal{L}_{\text{mag}} = \text{ReLU}\left(\|\hat{x} - x^{\beta_n}\| - \gamma\|x^{\beta_t} - x^{\beta_n}\|\right)^2 \tag{17}$$

where $\gamma = 1.5$ (set empirically) to allow limited stylistic exaggeration beyond the ground truth shape drift. The shape control loss function is the weighted sum of three components:

$$\mathcal{L}_{\text{shape}} = \mathcal{L}_{\text{bone}} + \lambda_{\text{dir}}\mathcal{L}_{\text{dir}} + \lambda_{\text{mag}}\mathcal{L}_{\text{mag}} \tag{18}$$

These terms propagate gradients through MPSM instead of shape-aware decoder $\mathcal{D}$, ensuring reliable shape control. The overall loss function of MorphoStyle is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{dis}} + \mathcal{L}_{\text{shape}} \tag{19}$$

4 Experiments

This section describes the experimental setup and results.

4.1 Experimental Settings

Hypothesis. Our experiments evaluate three main hypotheses about MorphoStyle: (**H1**) our style disentanglement mechanism transfers style more faithfully and with better physical plausibility than competing baselines; (**H2**) style injection can be decoupled from the shape-control pathway to preserve quantitative shape control under style transfer; and (**H3**) MorphoStyle achieves much lower computational cost than diffusion-based stylizers.

Datasets. We conduct extensive experiments on *HumanML3D* [8] for content motion and *100Style* [28] for style reference, following the standard SMooDi evaluation protocol [43]. HumanML3D contains 14,616 motion sequences ($\sim$ 29 hours) with three free-form captions per sequence, retargeted to the canonical SMPL skeleton [26]; we use the official train/val/test splits. 100Style contains 100 distinct stylistic walking variants performed by a single actor; for fair comparison with prior diffusion-based stylizers, we adopt the 47-class subset whose action categories overlap with HumanML3D, so that the same cross-dataset style classifier can be used for evaluation. To supervise the direction-aware shape control terms, we further use the paired neutral- and target-shape renderings of HumanML3D released by ShapeMyMove [25], which provide x^{β_n} and x^{β_t} for each training motion.

Evaluation Measures. We adopt seven evaluation protocols to account for both motion style transfer quality and shape control fidelity. For motion style transfer, we follow the SMooDi protocol [43] to report three measures on $N = 4209$ content-style pairs, where 4209 content motions are the standard HumanML3D [8] test split, each randomly paired with a style reference sampled from the 100Style [28] dataset. *Fréchet Inception Distance (FID)* [17] measures the distributional distance between generated and ground-truth motions in the Guo22 evaluator's feature space [8]; *Style Recognition Accuracy (SRA)* measures stylization fidelity as the top-1 classification accuracy of a frozen 47-class style classifier on the generated motions; *Foot Skating* measures physical plausibility as the average horizontal foot velocity in the foot-contact frames. For shape control, we adopt the *Bone-Length MAE* protocol, which measures absolute bone-length agreement with the target body, from ShapeMyMove [25] and introduce three additional measures. *Bone-Length Temporal Standard Deviation* (BoneLen-T-Std) measures intra-sequence stability of bone lengths; *β-swap response* measures the bone-length shift induced by switching between two contrasting body presets (thin$\leftrightarrow$heavy and short$\leftrightarrow$tall); and *β-correlation* measures the Pearson correlation between the recovered shape direction and the ground-truth shape direction.

Implementation Details. MorphoStyle is implemented in PyTorch [31] and trained on a single NVIDIA A100 GPU (80 GB). To better fit the data distribution of the 100Style dataset, we follow ShapeMyMove [25] to fine-tune the FSQ-VAE backbone on both HumanML3D and 100Style. The text-motion alignment model and the Shape-to-Feature module are loaded from their released checkpoints [8, 25] and kept frozen throughout the training procedure. We use the AdamW optimizer [27] with $(\beta_1, \beta_2)=(0.9, 0.99)$ and weight decay 10^{-4}; the base learning rate is 1.5×10^{-5} with a $5\times$ multiplier for the joint-text routing modules to compensate for their near-zero initialization, a 50-step linear warmup, and cosine annealing to 10^{-6}. We train for 120 epochs with batch size 24 on cross-domain content-style pairings sampled from HumanML3D and 100Style at a 1:1 ratio, and the SMooDi-classifier guidance is ramped up linearly over the first three epochs. We fix all loss weights throughout training, with $\lambda_{\text{code}} = 0.5$, $\lambda_{\text{cl}} = 0.15$, $\lambda_{\text{cc}} = 0.05$, $\lambda_{\text{cls}} = 0.05$, $\lambda_{\text{dir}} = 0.5$, and $\lambda_{\text{mag}} = 0.25$. During inference, the residual scale is set to α=0.5 in Equation 6.

4.2 Experimental Results

Motion style transfer. We compare MorphoStyle with six representative motion style transfer baselines. *MLD+Aberman et al.* [2, 4] and *MLD+MotionPuzzle* [4, 19] are classical cascades that attach an AdaIN-based or graph-aware stylizer on top of a pre-trained text-to-motion latent diffusion model. *SMooDi* [43] and *StyleMotif* [11] are two recent diffusion-based stylizers that augment a latent-diffusion backbone with a ControlNet-style adaptor and cross-modal alignment (respectively). *ShapeMyMove* [25] is the shape-aware text-to-motion generator on which our backbone is built. Although it has no stylization option, we compare it to isolate the contribution of the FSQ-VAE backbone alone. *SMooDi + ShapeMyMove* is a stronger cascade we constructed by attaching SMooDi's style adaptor to the ShapeMyMove [25] shape-aware decoder; it serves as the most direct baseline for comparison.

As reported in Table 1, MorphoStyle achieves the best Style Recognition Accuracy (SRA) and the best Foot Skating ratio among all competing methods, while remaining competitive on FID. Specifically, MorphoStyle attains an SRA of 81.9%, surpassing the baseline StyleMotif by 13.1%; this large gap indicates that our modular latent disentanglement is substantially more effective in injecting style into the generative latent than ControlNet-style or cross-modal-alignment alternatives. In terms of physical plausibility, our Foot Skating ratio of 0.081 is the lowest across the board, owing to the temporal gate of MPSM that adaptively attenuates style injection at contact-critical frames. Regarding distributional realism, our FID of 2.84 is comparable to that of SMooDi (2.95) and substantially lower than that of all cascade baselines (3.89 to 7.87), although it does not surpass StyleMotif (1.38); we attribute this gap to StyleMotif's significantly heavier $\sim$ 495 M-parameter latent-diffusion backbone. More importantly, MorphoStyle is the *only* method that simultaneously delivers high-quality motion style transfer and quantitative shape control: the four diffusion-based baselines are architecturally incapable of supporting shape input, and the SMooDi+ShapeMyMove, although shape-aware in design, suffers severely on both FID and Foot Skating due to the latent-space interference. These results strongly support **H1**.

Figure 3 provides qualitative comparisons of stylized motions conditioned by different body shape descriptions. SMooDi+ShapeMyMove transfers the coarse appearance of the reference style but corrupts the motion content, producing undesired content leakage in the form of over-extended hand swings, irregular gait timing, and noticeable foot skating. MorphoStyle reproduces the target style while remaining faithful to the input motion content throughout the sequence; the generated motion is temporally coherent and free of skating.

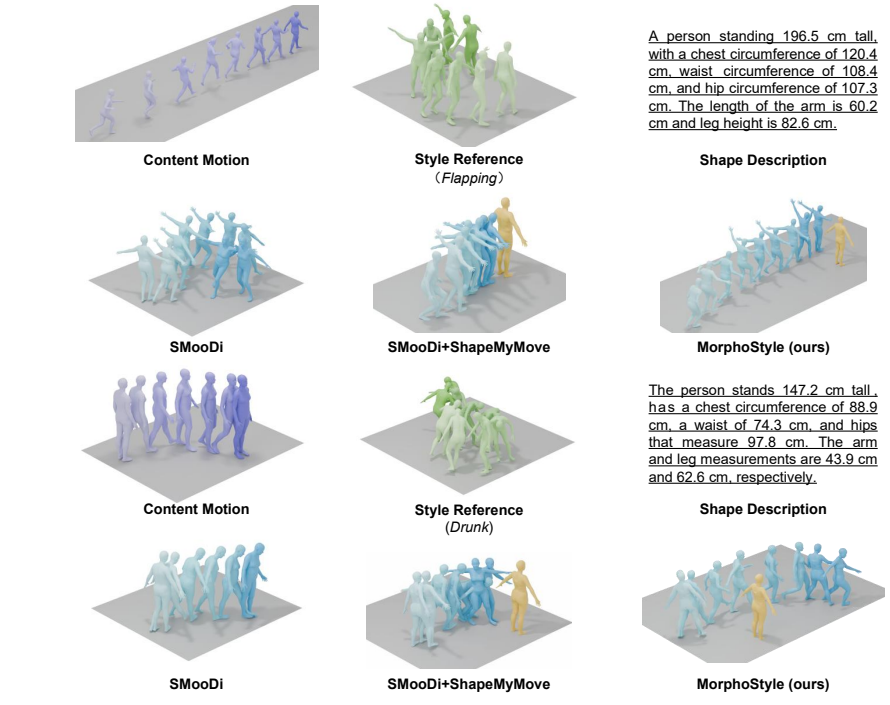

Figure 3: Qualitative comparison of shape-aware motion style transfer against SMooDi [43] and SMooDi+ShapeMyMove. Yellow body is the target shape used for reference.

Method	SRA↑	FID↓	Foot Skate↓	Shape Control
MLD + Aberman *et al.*	61.1%	3.89	0.338	×
MLD + Motion Puzzle	67.2%	6.87	0.197	×
SMooDi	65.1%	2.95	0.095	×
StyleMotif	68.8%	1.38	0.094	×
ShapeMyMove	1.57%	**1.36**	0.228	✓
SMooDi + ShapeMyMove	69.9%	7.87	0.245	✓
MorphoStyle (ours)	**81.9%**	2.84	**0.081**	✓

Table 1: Comparing MorphoStyle with baselines for motion style transfer on HumanML3D content with 100Style references. **Bold** indicates best result and underline the second-best.

Shape control. We evaluated shape control fidelity of MorphoStyle against two shape-aware baselines, ShapeMyMove [25] and the SMooDi+ShapeMyMove, in terms of four shape measures. We do not include SMooDi [43] or StyleMotif [11] since their backbones do not condition on body shape and are incapable of being evaluated on shape measures. As reported in Table 2, MorphoStyle attains the lowest Bone-Length MAE (1.91 cm), lowest temporal Bone-Length standard deviation (1.26 cm), and highest β-swap response (5.31 cm), surpassing the SMooDi+ShapeMyMove by 23.3% on bone-length and a striking 61.9% on temporal stability. This indicates that MorphoStyle most accurately produces motions with the target body proportions, keeps these proportions consistent over the motion sequence, and is most responsive to changes in the SMPL shape parameter β. In summary, MorphoStyle delivers the highest-fidelity shape control among all methods that simultaneously support stylization,

Method	BoneLen-MAE↓	BoneLen-T-Std↓	β-swap↑	β-corr↑
ShapeMyMove	4.59	3.53	5.05	**0.677**
SMooDi + ShapeMyMove	<u>2.49</u>	<u>3.31</u>	<u>3.87</u>	0.314
MorphoStyle (ours)	**1.91**	**1.26**	**5.31**	<u>0.332</u>

Table 2: Comparison with competing methods for shape control during motion style transfer using four evaluation measures. **Bold** indicates the best result and <u>underline</u> the second-best.

Method	Trainable params	Inference time↓	GFLOPS↓
SMooDi	13.90 M	7.19s	79.59
SMooDi+ShapeMyMove	31.11 M	2.11s	112.84
MorphoStyle (ours)	**4.25 M**	**0.02s**	**1.98**

Table 3: Comparison of model parameters, inference speed, and computational cost.

providing strong evidence in support of **H2**.

As presented in Figure 3, we visualize motion style transfer results conditioned on different target body shapes specified by textual shape descriptions. SMooDi+ShapeMyMove does adapt to the target body shape, yet the bolt-on style adaptor consistently introduces drastic kinematic artifacts, confirming the architectural mismatch discussed earlier. In contrast, MorphoStyle preserves the target body morphology, produces stylistically expressive yet physically plausible motions, and remains stable across diverse style references, validating the effectiveness of our modular latent disentanglement design.

Computational overhead. We evaluated the computational efficiency of MorphoStyle against two competing baselines, SMooDi and the SMooDi + ShapeMyMove, that natively support shape control or are closest to our problem setting in terms of trainable parameter count, inference time, and computational cost measured in GFLOPs. All measurements are conducted on a single NVIDIA A100 GPU with batch size 1 to faithfully reflect a real-time deployment scenario. Since StyleMotif [11] has not released its software or checkpoints, we do not compare it in terms of shape control (Table 2) or computational overhead. As reported in Table 3, MorphoStyle requires 4.25 M trainable parameters, which is 3.3× smaller than SMooDi (13.90 M) and 7.3× smaller than the SMooDi+ShapeMyMove (31.11 M). Specifically, MorphoStyle synthesizes a stylized motion in 0.02 second per sample (approximately 50 FPS on a single GPU), compared to 7.19 seconds for SMooDi's 50-step DDIM sampler and 2.11 seconds for the SMooDi+ShapeMyMove, yielding significant speed-up. Similarly, MorphoStyle requires only 1.98 GFLOPs per inference, which is significantly less than the two competing baselines. These results strongly support **H3** and are due to our design choices; unlike diffusion-based stylizers that perform iterative denoising over tens of sampling steps, MorphoStyle synthesizes the stylized motion in a single forward pass through a 4.25 M-parameter style branch grafted onto a frozen FSQ-VAE backbone. Such efficiency makes MorphoStyle work at real-time rates while supporting quantitative shape control, expanding its applicability to interactive animation, real-time avatar control, and on-device deployment scenarios that remain infeasible for diffusion-based competitors.

4.3 Ablation Study

Content-style disentanglement. We conducted experiments to investigate two main components of our content-style disentanglement design. The supervised contrastive supervision (**Con**) on the global style embedding s_g (Equation 3), and the text-guided joint style routing module (**JTC**) that distills s_j from style-related joints (Equation 5). Starting from a base-

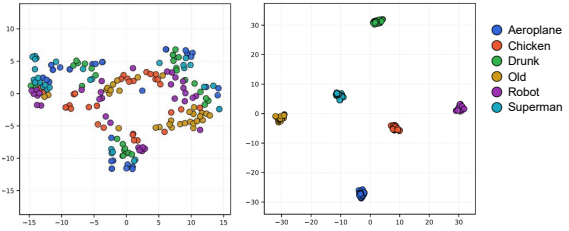

Figure 4: t-SNE visualization of global style embeddings s_g on six representative style classes, *with* (right) versus *without* (left) the contrastive supervision $\mathcal{L}_{\text{cl}}$.

Variant	Contrastive	Joint	FID↓	SRA↑	Foot↓
Base	×	×	3.57	74.7%	0.143
JTC	×	✓	3.54	78.5%	0.139
Con	✓	×	3.05	81.2%	0.110
MorphoStyle	✓	✓	2.84	81.9%	0.081

Table 4: Ablation on content-style disentanglement components.

line variant (**Base**) that disables both, we add each component independently and finally combine them in MorphoStyle. As reported in Table 4, adding the contrastive supervision alone yields the most improvement, and Con increases SRA by 6.5%, reduces FID by 0.52, and lowers Foot Skating, which indicates that without explicit supervision in the embedding space, global style encoder $\mathcal{E}_s$ inevitably entangles content cues that degrades stylization fidelity and motion realism. Adding the joint routing brings a smaller but consistent gain, which confirms that joint-level disentanglement provides complementary local style cues that the whole-body global embedding cannot capture. Combining them in MorphoStyle provides the best performance on every measure, which demonstrates their effectiveness for high-quality motion style transfer.

To further investigate the qualitative effect of contrastive supervision, we visualize global style embeddings s_g with t-SNE in Figure 4 for six representative style classes. Without contrastive supervision, embeddings of the same style class are scattered and frequently overlap with other classes, reflecting strong content entanglement in s_g. In contrast, with contrastive supervision, embeddings of the same style form tight and well-separated clusters in the latent feature space. This shows that our content-style disentanglement method provides a discriminative style space, which MPSM relies on for high-fidelity style injection.

MPSM module. We investigated two variants of our MPSM, low-rank additive offset (Equation 9) and learnable temporal gate (Equation 8). We first entirely remove MPSM (*w/o MPSM*), which is replaced by a naive concatenation between content features and style embeddings, and then remove only temporal gate (*w/o gate g*), which keeps low-rank offset but applies it uniformly across all temporal frames. As reported in Table 5, entirely disabling the MPSM results in worse performance in terms of SRA and FID due to content leakage, which demonstrates that MPSM is an essential component for injecting style in content features. Removing the temporal gate yields a less severe but still significant degradation, indicating that applying low-rank offset across all frames corrupts key movements without temporal intensity control of style injection, degrading style fidelity and motion quality.

Shape control supervision. We also evaluated the effectiveness of the shape control objective $\mathcal{L}_{\text{shape}}$ by removing it from the training process (*w/o* $\mathcal{L}_{\text{shape}}$) (see Table 5). Without $\mathcal{L}_{\text{shape}}$, MorphoStyle still attains a high SRA and a low FID. However, it is evident from

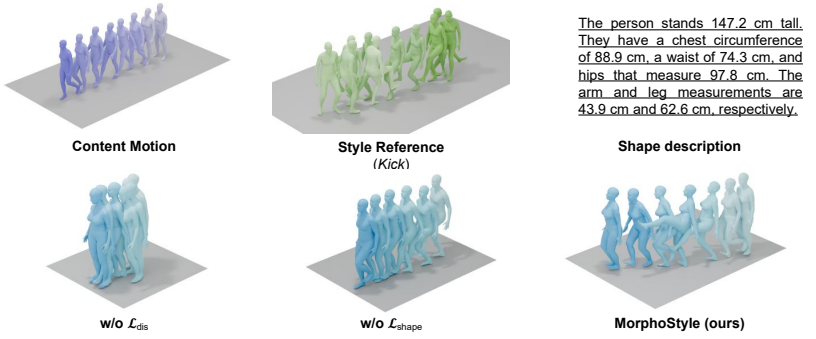

Figure 5: Qualitative ablation of two training objectives, $\mathcal{L}_{\text{dis}}$ and $\mathcal{L}_{\text{shape}}$, on the same content (walk forward), style reference (*kick*), and target body shape.

	SRA↑	**FID↓**	**BoneLen-MAE↓**	**β-swap↑**
w/o MPSM	62.9%	5.16	1.97	2.06
w/o gate g	72.1%	4.85	1.92	4.12
w/o $\mathcal{L}_{shape}$	80.4%	2.99	2.20	3.41
MorphoStyle	81.9%	2.84	1.91	5.31

Table 5: Ablation results on key technical components of our MorphoStyle.

the shape measures that Bone-Length MAE rises, and β-swap response collapses from 5.31 to 3.41 which indicates that the model no longer responds meaningfully to changes in the SMPL shape parameter β, effectively ignoring the shape signal. Figure 5 also visualizes the failure modes of removing each loss function confirming that the shape control objective adapts latent codes to ground truth shape direction through $\mathcal{L}_{\text{shape}}$ supervision.

5 Conclusion

We presented *MorphoStyle*, a unified framework for shape-aware motion style transfer that jointly models motion style transfer and body-shape control. MorphoStyle leverages a frozen shape-conditioned FSQ-VAE and injects style as a compact low-rank residual in the post-quantization content space, while preserving explicit shape control through the de-quantization pathway. This design enables effective disentanglement of content, style, and morphology with fewer parameters. Built on this backbone, our content-decoupled style branch combines contrastive style encoding, text-guided joint routing, and manifold-preserving modulation to localize and transfer style without disrupting motion content. Extensive experiments show that MorphoStyle achieves state-of-the-art performance in both motion style transfer and morphological control, while being 3.3× smaller and 360× faster than the closest competing baseline, enabling real-time shape-aware motion style transfer.

Limitations and future work. MorphoStyle's distributional realism is limited by the capacity of the frozen FSQ-VAE backbone, and its FID does not surpass the diffusion-based methods. In addition, the text-guided joint routing module depends on a frozen text-motion alignment model, which may be less reliable for out-of-distribution actions. We will explore extensions to MorphoStyle to address these limitations. Future work also includes extending the framework to consider image, video, and/or audio style references, and adapting shape-aware stylization to non-humanoid characters and physics-based animation.

6 Acknowledgment

The authors thank Lei Zhong and Ziyuan Li for suggestions during the development of MorphoStyle. This work was supported in part by the Arts and Humanities Research Council grant AH/Y00115X/1. All conclusions reported here are those of the authors alone.

References

[1] Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics (SIGGRAPH)*, 39(4):62:1–62:14, 2020.

[2] Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Unpaired motion style transfer from video to animation. *ACM Transactions on Graphics (SIGGRAPH)*, 39(4):64:1–64:12, 2020.

[3] Matthew Brand and Aaron Hertzmann. Style machines. In *Proceedings of SIGGRAPH*, pages 183–192, 2000.

[4] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[5] Vasileios Choutas, Lea Müller, Chun-Hao P. Huang, Siyu Tang, Dimitrios Tzionas, and Michael J. Black. Accurate 3D body shape regression using metric and semantic attributes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[6] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[8] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3D human motions from text. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[9] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. MoMask: Generative masked modeling of 3D human motions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[10] Chuan Guo, Yuxuan Mu, Xinxin Zuo, Peng Dai, Youliang Yan, Juwei Lu, and Li Cheng. Generative human motion stylization in latent space. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations (ICLR)*, volume 2024, pages 7859–7878, 2024.

[11] Ziyu Guo, Young Yoon Lee, Joseph Liu, Yizhak Ben-Shabat, Victor Zordan, and Mubbasir Kapadia. StyleMotif: Multi-modal motion stylization using style-content cross fusion. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[14] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (SIGGRAPH)*, 35(4): 138:1–138:11, 2016.

[15] Daniel Holden, Ikhsanul Habibie, Ikuo Kusajima, and Taku Komura. Fast neural style transfer for motion data. *IEEE Computer Graphics and Applications*, 37(4):42–49, 2017.

[16] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

[17] Daniel Holden, Oussama Kanoun, Maksym Perepichka, and Tiberiu Popa. Learned motion matching. *ACM Transactions on Graphics (ToG)*, 39(4):53–1, 2020.

[18] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[19] Deok-Kyeong Jang, Soomin Park, and Sung-Hee Lee. Motion puzzle: Arbitrary motion style transfer by body part. *ACM Transactions on Graphics (ToG)*, 41(3):33:1–33:16, 2022.

[20] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems (NIPS)*, 33:18661–18673, 2020.

[21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[22] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[23] Yan Li, Tianshu Wang, and Heung-Yeung Shum. Motion texture: A two-level statistical model for character motion synthesis. *ACM Transactions on Graphics (SIGGRAPH)*, 21(3):465–472, 2002.

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

[25] Ting-Hsuan Liao, Yi Zhou, Yu Shen, Chun-Hao Paul Huang, Saayan Mitra, Jia-Bin Huang, and Uttaran Bhattacharya. Shape my moves: Text-driven shape-aware synthesis of human motions. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 1917–1928, 2025.

[26] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015.

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[28] Ian Mason, Sebastian Starke, and Taku Komura. Real-time style modelling of human locomotion via feature-wise transformations and local motion phases. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 5(1):1–18, 2022.

[29] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *International Conference on Learning Representations (ICLR)*, 2024.

[30] Soomin Park, Deok-Kyeong Jang, and Sung-Hee Lee. Diverse motion stylization for multiple style domains via spatial-temporal graph-based generative model. *Proc. ACM Comput. Graph. Interact. Tech.*, 4(3):1–17, 2021.

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[32] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. DeepMimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions on Graphics (SIGGRAPH)*, 37(4):143:1–143:14, 2018.

[33] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. AMP: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (SIGGRAPH)*, 40(4):144:1–144:20, 2021.

[34] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022.

[35] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 21(140):1–67, 2020.

[36] Tianxin Tao, Xiaohang Zhan, Zhongquan Chen, and Michiel van de Panne. Style-erd: Responsive and coherent online motion style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6593–6603, 2022.

[37] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. In *International Conference on Learning Representations (ICLR)*, 2023.

[38] Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee. Neural kinematic networks for unsupervised motion retargetting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[39] Shihong Xia, Congyi Wang, Jinxiang Chai, and Jessica Hodgins. Realtime style transfer for unlabeled heterogeneous human motion. *ACM Transactions on Graphics (SIGGRAPH)*, 34(4):119:1–119:10, 2015.

[40] Katsu Yamane, Yuka Ariki, and Jessica Hodgins. Animating non-humanoid characters with human motion data. In *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 169–178, 2010.

[41] Fatemeh Zargarbashi, Dhruv Agrawal, Jakob Buhmann, Martin Guay, Stelian Coros, and Robert W Sumner. Vq-style: Disentangling style and content in motion with residual quantized representations. In *Computer Graphics Forum*, page e70377. Wiley Online Library, 2026.

[42] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[43] Lei Zhong, Yiming Xie, Varun Jampani, Deqing Sun, and Huaizu Jiang. SMooDi: Stylized motion diffusion model. In *European Conference on Computer Vision (ECCV)*, 2024.

MorphoStyle: Motion Style Transfer with Morphology Control

Xin Feng
xfeng4@ed.ac.uk
Eleonora D'Arnese
eleonora.darnese@ed.ac.uk
Mohan Sridharan
m.sridharan@ed.ac.uk

Institute of Perception, Action, and Behavior
School of Informatics
University of Edinburgh
Edinburgh, UK

Supplementary Material

A Implementation Details

A.1 Frozen and Trainable Components

MorphoStyle is built around a frozen shape-aware FSQ-VAE backbone and a small trainable style branch. Table S1 lists every module, its frozen/trainable status, and its tensor signatures. MorphoStyle has a total of 4.25 M trainable parameters, matching the value reported in Table 3 of the main paper, and the frozen backbone contributes 38.98 M parameters that are reused as-is from the *ShapeMyMove* checkpoint [3].

Module	Trainable	Input dim.	Output dim.	#params
Content motion encoder $\mathcal{E}$	no	$T\times 263$	$T'\times 512$	6.7 M
Finite-Scalar-Quantizer	no	$T'\times 512$	$T'\times 512$	0
Shape-conditioned decoder $\mathcal{D}$	no	$T'\times(512+32)$	$T\times 263$	7.4 M
Shape-to-Feature module S2FM	no	text t_s	$\mathbb{R}^{32}$	24.1 M
Text-motion alignment $\mathcal{T},\mathcal{M}$	no	text / motion	shared $\mathbb{R}^{512}$	0.8 M
Global style encoder $\mathcal{E}_s$	yes	$T\times 263$	$s_g\in\mathbb{R}^{256}$	2.95 M
Joint routing MLP f_{j}	yes	$\mathbb{R}^{12}$	$s_j\in\mathbb{R}^{128}$	0.07 M
Style fusion MLP	yes	$\mathbb{R}^{384}$	$\mathbb{R}^{256}$	0.13 M
MPSM low-rank branch	yes	$\mathbb{R}^{256}$	$r(s)\in\mathbb{R}^{512}$	0.61 M
MPSM gating network f_{g}	yes	$\mathbb{R}^{(512+256)\times T'}$	$g\in(0,1)^{T'}$	0.49 M
Total trainable				**4.25 M**
Total frozen				**38.98 M**

Table S1: Per-module summary of MorphoStyle with parameter counts. Frozen modules are loaded from public checkpoints and never updated. Dimensions follow main-paper notation: T=196 frames, $T'=T/4$=49 codes, D=512 code channels, 263-dim HumanML3D features, J=22 joints, K=8 routed joints.

A.2 Network Architecture

Shape-aware FSQ-VAE backbone (frozen). We use the shape-aware FSQ-VAE released by ShapeMyMove [3] verbatim as the frozen motion tokenizer. The content encoder $\mathcal{E}$ stacks three Conv1d residual blocks with dilation growth rate 3 and temporal downsampling factor 4, mapping a motion sequence $x \in \mathbb{R}^{T\times 263}$ to a code sequence $c \in \mathbb{R}^{T'\times 512}$ with $T' = T/4 = 49$. The Finite-Scalar-Quantization bottleneck uses FSQ levels $(8,5,5,5)$ yielding $8\cdot 5\cdot 5\cdot 5 = 1000$ codebook entries. The shape-conditioned decoder $\mathcal{D}$ concatenates each quantized code with the 32-dimensional shape embedding $\phi(\beta)$ along the channel dimension, and reconstructs the motion through symmetric Conv1d upsampling.

Shape-to-Feature module. S2FM is inherited from ShapeMyMove [3]. A T5-based language model [7] regresses the 10-dimensional SMPL parameter β from a textual shape description t_s, and a two-layer MLP_ϕ projects β into the 32-dimensional shape embedding $\phi(\beta)=\text{MLP}_\phi(\text{T5}(t_s))$. When β is directly available (*e.g.*, from 100Style with ShapeMyMove-recovered β), we bypass the T5 stage and feed β directly to MLP_ϕ.

Contrastive global style encoder $\mathcal{E}_s$. $\mathcal{E}_s$ takes a style reference motion $x_s \in \mathbb{R}^{T\times 263}$ as input. It consists of three strided Conv1d blocks (channels $64 \to 128 \to 256$, kernel size 5, stride 2), each followed by GroupNorm and SiLU, an adaptive average pooling, and a two-layer MLP that projects the pooled feature into the 256-dimensional global style embedding s_g. We apply style dropout with probability p=0.15, replacing s_g with a learnable null token so that the downstream MPSM remains robust to missing or noisy style references at inference. The supervised contrastive objective $\mathcal{L}_{\text{cl}}$ uses cosine similarity with temperature τ=0.1.

Text-guided joint style routing. We use the frozen text-motion alignment model from Guo *et al.* [1], which provides the text encoder $\mathcal{T}(\cdot)$ and motion encoder $\mathcal{M}(\cdot)$. For each of the J=22 HumanML3D joints, we build a joint-masked variant $\tilde{x}_s^{(j)}$ that retains only joint j's positional, rotational, and velocity channels and zeros the rest; scores its alignment with $\mathcal{T}(c_t)$ by cosine similarity; and selects the K=8 joints with the *lowest* similarity, *i.e.*, the joints whose isolated motion is least aligned with the content caption. The per-joint kinematic feature $\psi_j(x_s) \in \mathbb{R}^{T\times 12}$ concatenates the joint position (3), the 6D continuous rotation representation (6), and linear velocity (3) as defined by HumanML3D. Time-averaged features and a learnable per-joint identity embedding $u_j \in \mathbb{R}^{12}$ are aggregated by a two-layer MLP $f_{\text{j}}:\mathbb{R}^{12}\to\mathbb{R}^{128}$ to produce $s_j \in \mathbb{R}^{128}$.

Style code fusion. The unified style code $s \in \mathbb{R}^{256}$ is produced by a residual MLP, $s = s_g + \alpha\cdot\text{MLP}([s_g; s_j])$, with α=1 during training. The final MLP layer is initialized with a small Gaussian (σ=0.01) so that the routing residual is near zero at the start of inference, after which α becomes a runtime knob that smoothly trades stylization strength against motion fidelity without retraining. The default at inference is α=0.5.

Manifold-Preserving Style Modulator. The low-rank additive offset is $r(s)=W_{\text{up}}W_{\text{down}}s + b$, with $W_{\text{down}} \in \mathbb{R}^{16\times 256}$, $W_{\text{up}} \in \mathbb{R}^{512\times 16}$, and $b \in \mathbb{R}^{512}$ initialized to zero, so that the offset lives in a rank-r subspace of the 512-dimensional code space with r=16. The temporal-gate branch replicates s along the temporal axis to $\tilde{s} \in \mathbb{R}^{256\times T'}$, concatenates it with c along the channel dimension, and applies a three-layer Conv1d gating network f_{g} with channels $256 \to 128 \to 1$, kernel size 3, SiLU activations, and a final sigmoid to produce a per-frame gate $g \in (0,1)^{T'}$. The stylized codes are $c'_t = c_t + g_t \cdot r(s)$ for $t = 1,\ldots,T'$.

A.3 Training Hyperparameters

We train MorphoStyle on a single NVIDIA A100 (80 GB) GPU. Optimization follows the schedule in Table S2, with all non-MPSM and non-style-branch modules kept frozen. Each batch samples HumanML3D [1] and 100Style [6] pairs at a 1:1 ratio; the SMooDi-classifier guidance term is linearly ramped up over the first three epochs to stabilize early training. The base learning rate is 1.5×10^{-5}, with a $5\times$ multiplier on the joint-routing MLP f_{j} and the style-fusion MLP to compensate for their near-zero initialization.

Hyperparameter	Value
Optimizer	AdamW [5]
(β_1, β_2)	$(0.9, 0.99)$
Weight decay	1×10^{-4}
Base learning rate	1.5×10^{-5}
Routing-branch lr multiplier	$5\times$
Warm-up	50 steps, linear
Schedule after warm-up	cosine to 10^{-6}
Total epochs	120
Batch size	24
HumanML3D : 100Style sampling ratio	1 : 1
$\mathcal{L}_{\text{cls}}$ ramp	linear over first 3 epochs
Gradient clipping	ℓ_2 norm at 5.0
Hardware	1× NVIDIA A100 (80 GB)
Mixed precision	FP16 (*autocast*)
Wall-clock training time	7.4 h
Peak training memory	11.2 GB
Style dropout p	0.15
Contrastive temperature τ	0.1
MPSM rank r	16
Top-K joints	8
Joint identity embedding dim.	12
Residual scale α	1 (training) / 0.5 (inference)
Magnitude tolerance γ (Eq. 17 main)	1.5
λ_{code}	0.5
λ_{cl}	0.15
λ_{cc}	0.05
λ_{cls}	0.05
λ_{dir}	0.5
λ_{mag}	0.25

Table S2: List of training and inference hyperparameters. All values match those reported in the main paper. Wall-clock and memory are measured on a single NVIDIA A100 (80 GB).

A.4 Training Dynamics

A single training run completes in ≈ 7.4 hours of wall-clock time on one A100. The peak GPU memory footprint is 11.2 GB, dominated by the frozen 24.1 M-parameter T5 encoder used in S2FM; the trainable style branch itself contributes less than 1 GB at the batch size of 24. Training is stable over the 120 epochs; the disentanglement loss $\mathcal{L}_{\text{dis}}$ decreases mono-

tonically once the classifier-guidance ramp completes at epoch 3, and the contrastive loss $\mathcal{L}_{\text{cl}}$ converges to a tight plateau near 0.18 after epoch 30, indicating that the global style encoder has formed well-separated per-style clusters (visualized in Sec. H). Validation measures monitored every 5 epochs on a held-out subset of 300 HumanML3D motions paired with 100 random 100Style references show no overfitting up to epoch 120; the best validation Style Recognition Accuracy (SRA) is attained at epoch 115 and is within the noise envelope of the final reported model.

B Experimental Settings

HumanML3D. We follow the standard train/validation/test split released by Guo *et al.* [1]: 23384 training, 1460 validation, and 4209 test sequences. Motions are pre-processed to the canonical SMPL skeleton, resampled to 20 FPS, and clipped or padded to $T=196$ frames using the protocol provided by the dataset authors. Each sequence is associated with up to three free-form captions; we use only the first caption per sequence in the routing module $\mathcal{T}(c_t)$ to keep the joint similarity well-defined and deterministic at inference.

100Style. We use the 47-class subset of 100Style [5] that overlaps with HumanML3D, the standard dataset adopted by SMooDi [8], so that the released 47-class style classifier f_{sm} can be reused without retraining. Each 100Style motion is paired with the SMPL shape parameter β provided by ShapeMyMove [3]. The 47 classes used in the SMooDi evaluation protocol cover a broad range of locomotion modifiers (*e.g.*, *Drunk*, *Tiptoe*, *OnHeels*, *Crouched*, *Stiff*, *Skipping*), expressive gestures (*e.g.*, *Aeroplane*, *Superman*, *Flapping*, *Chicken*), and emotional or mannered gaits (*e.g.*, *Angry*, *Proud*, *Depressed*, *InTheDark*, *Heavyset*).

Natural-language shape-description templates. For training, each β is converted to a natural-language shape description t_s following the protocol of ShapeMyMove [3]. The seven anthropometric attributes (height, chest, waist, hip circumferences, arm length, leg length, and gender token) are computed from the canonical SMPL T-pose mesh of the given β via the SMPL forward model [4], and inserted into a fixed pool of ten sentence templates that vary in clause order, attribute coverage, and tone. Example templates include: *"A person standing H cm tall, with a chest circumference of C cm, a waist circumference of W cm, and a hip circumference of P cm."*; *"Body measurements: height H cm, chest C cm, waist W cm, hip P cm, arm length A cm, leg length L cm."* During training, β is directly available, so the T5 stage of S2FM is bypassed and β is fed directly to MLP_ϕ; at inference, either the textual or the β pathway can be used interchangeably.

Evaluation set. The motion style transfer evaluation set comprises $N=4\,209$ content-style pairs. The 4 209 content motions are the full standard HumanML3D [1] test split, and each content motion is randomly paired with one style reference sampled uniformly from the 47-class 100Style test split. The pairings are drawn once with a fixed random seed and shared across every method we evaluate. The shape-control evaluation set is constructed analogously, with each content motion additionally paired with a target body shape sampled from the ShapeMyMove shape pool; the same content-style-shape triplets are used for every method to make all numbers directly comparable.

Shape-preset evaluation set. For the per-shape-preset analysis in Sec. E, we construct an additional evaluation set by rendering each content-style pair under five canonical β presets: *orig* (the content motion's own HumanML3D-recovered β), *thin* ($\beta_0=-2.5$, the remaining

components set to zero), *neutral* (β=**0**), *heavy* (β_0=+2.5), and *tall* (β_1=+2.5). The canonical T-pose heights for these presets are 168, 147, 172, 196, and 176 cm, respectively.

Evaluator features. We use the publicly released Guo22 motion evaluator [1] to compute the 512-dimensional feature embeddings underlying FID and Diversity. The evaluator has not been retrained for this paper; we use the same checkpoint as SMooDi and StyleMotif to keep the FID number directly comparable.

C Evaluation Measures

We provide the closed-form definition of every measure reported in the main paper. Notation: $\hat{x}$ denotes the synthesized motion, $x^\star$ a ground truth or target-shape reference motion, y_{sty} the target style label, and $\text{BL}(\cdot)\in\mathbb{R}^{21}$ the bone-length vector extracted from a motion (one entry per skeleton bone, in the SMPL 21-bone topology).

C.1 Motion Style Transfer Measures

Fréchet Inception Distance (FID). Following SMooDi [8] and Guo22 [1], the synthesized and ground truth motions are first embedded into the 512-dimensional Guo22 evaluator feature space. Let $\Phi(\cdot)$ denote this evaluator, $(\mu_{\text{gen}},\Sigma_{\text{gen}})$ the empirical mean and covariance of $\{\Phi(\hat{x}_i)\}$, and $(\mu_{\text{gt}},\Sigma_{\text{gt}})$ those of $\{\Phi(x_i^\star)\}$. Then:

$$\text{FID} \;=\; \|\mu_{\text{gen}}-\mu_{\text{gt}}\|_2^2+\text{Tr}\Big(\Sigma_{\text{gen}}+\Sigma_{\text{gt}}-2\big(\Sigma_{\text{gen}}\Sigma_{\text{gt}}\big)^{1/2}\Big), \tag{S1}$$

which is the definition of [2] applied to the Guo22 evaluator. Lower values are better.

Style Recognition Accuracy (SRA). SRA is the top-1 classification accuracy of the frozen 47-class style classifier f_{sm} released by SMooDi [8], applied to the synthesized motions:

$$\text{SRA} \;=\; \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\big[\arg\max_y f_{\text{sm}}(\hat{x}_i)[y] \;=\; y_{\text{sty},i}\big]\,, \tag{S2}$$

where $y_{\text{sty},i}$ is the style label of the style reference $x_{s,i}$. Higher values are better.

Foot Skating. Foot Skating measures the average horizontal foot velocity at frames in foot contact, expressed as a velocity ratio. Let $v^{\text{xy}}_{f,t}$ denote the horizontal velocity of foot joint $f\in\{\text{LF},\text{RF}\}$ at frame t, computed by finite differencing the foot positions, and let $h_{f,t}$ be the foot height. Similar to SMooDi [8], a frame is in foot-contact when $h_{f,t}<h_{\text{thr}}$ with h_{thr}=5 cm. Then:

$$\text{FootSkating} \;=\; \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{f,t}\mathbf{1}[h^{(i)}_{f,t}<h_{\text{thr}}]\cdot\big\|v^{\text{xy},(i)}_{f,t}\big\|_2}{\sum_{f,t}\mathbf{1}[h^{(i)}_{f,t}<h_{\text{thr}}]+\varepsilon}. \tag{S3}$$

Lower values are better.

C.2 Shape Control Measures

We use the bone-length operator $\mathrm{BL}(\cdot) \in \mathbb{R}^{21}$ that returns the time-averaged length of each of the 21 SMPL skeleton bones in centimeters. For metrics that require a target-shape rendering of the same motion under a different body, we use $x^{(\beta_t)}$ to denote the ground truth motion rendered with target SMPL parameter β_t, and $x^{(\beta_n)}$ for the neutral-shape ($\beta = \mathbf{0}$) rendering.

Bone-Length MAE. The mean absolute error between the bone-length vector of the synthesized motion and that of the target-shape ground truth, averaged over the N=4209 evaluation triplets and the 21 bones:

$$\text{BoneLen-MAE} \;=\; \frac{1}{N}\sum_{i=1}^{N}\frac{1}{21}\left\|\mathrm{BL}(\hat{x}_i) - \mathrm{BL}(x_i^{(\beta_t)})\right\|_1. \tag{S4}$$

Lower values are better.

Bone-Length Temporal Std (BoneLen-T-Std). The intra-sequence temporal standard deviation of the synthesized bone-length vector, averaged over N and the 21 bones. Let $\mathrm{BL}_t(\hat{x}_i) \in \mathbb{R}^{21}$ denote the per-frame bone-length vector of the i-th synthesized motion at frame t. Then:

$$\text{BoneLen-T-Std} \;=\; \frac{1}{N}\sum_{i=1}^{N}\frac{1}{21}\sum_{b=1}^{21}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\mathrm{BL}_t(\hat{x}_i)[b] - \overline{\mathrm{BL}}(\hat{x}_i)[b]\right)^2}, \tag{S5}$$

where $\overline{\mathrm{BL}}(\hat{x}_i)$ is the temporal mean. This measure quantifies how stable the recovered bone lengths are over the sequence and penalizes shape jitter. Lower values are better.

β-correlation (β-corr). Pearson correlation between the bone-length *shift* produced by the model and the ground truth bone-length shift, measured against the neutral shape reference:

$$\beta\text{-corr} \;=\; \frac{1}{N}\sum_{i=1}^{N}\mathrm{corr}\Big(\mathrm{BL}(\hat{x}_i) - \mathrm{BL}(x_i^{(\beta_n)}),\; \mathrm{BL}(x_i^{(\beta_t)}) - \mathrm{BL}(x_i^{(\beta_n)})\Big), \tag{S6}$$

where $\mathrm{corr}(\cdot,\cdot)$ is the standard Pearson correlation over the 21-bone vector. Values are in $[-1,1]$; higher values are better. β-corr captures whether the recovered shape moves in the right *direction* from neutral, complementing the Bone-Length MAE.

β-swap response. β-swap quantifies how strongly the model's output responds when the input body shape parameter is swapped. Given two distinct shapes $\beta^{(1)}$ and $\beta^{(2)}$ paired with the same content motion and the same style reference, let $\hat{x}_i^{(1)}$ and $\hat{x}_i^{(2)}$ be the corresponding synthesized motions. We report the average L_1 bone-length distance:

$$\beta\text{-swap} \;=\; \frac{1}{N}\sum_{i=1}^{N}\frac{1}{21}\left\|\mathrm{BL}(\hat{x}_i^{(1)}) - \mathrm{BL}(\hat{x}_i^{(2)})\right\|_1. \tag{S7}$$

We construct $(\beta^{(1)}, \beta^{(2)})$ from two contrasting body presets, *thin* $\leftrightarrow$ *heavy* and *short* $\leftrightarrow$ *tall*, following the convention of ShapeMyMove [3]. A larger β-swap response indicates that the model meaningfully reacts to a change in β; a value close to zero indicates the model is effectively ignoring the shape signal.

D Loss Functions

This section describes the training objectives and provides the derivations omitted from the main paper.

Reconstruction term (main paper Eq. 10). We supervise the synthesized motion in both motion space and code space:

$$\mathcal{L}_{\text{rec}} = \|\hat{x} - x^{\star}\|_1 + \lambda_{\text{code}} \|c' - c^{\star}\|_1, \quad c^{\star} = \mathcal{E}(x^{\star}).$$

The code-space term keeps c' within the FSQ codebook neighborhood and empirically prevents drastic latent shifts that bypass the discrete structure of the backbone. Sec. G reports the quantitative effect of removing this term.

Style-embedding term (main paper Eqs. 3, 11, 12). The supervised contrastive loss $\mathcal{L}_{\text{cl}}$ acts on the global style embedding s_g:

$$\mathcal{L}_{\text{cl}} = -\mathbb{E}_i \log \frac{\exp\big(\text{sim}(s_{g,i}, s^{+}_{g,i})/\tau\big)}{\exp\big(\text{sim}(s_{g,i}, s^{+}_{g,i})/\tau\big) + \sum_{s_{g,n} \in \mathcal{N}_i} \exp\big(\text{sim}(s_{g,i}, s_{g,n})/\tau\big)},$$

where positives share the style label and negatives are drawn from the mini-batch. The action-classification loss $\mathcal{L}_{\text{cc}}$=CE$(f_{\text{ac}}(\hat{x}), y_{\text{act}})$ uses the HumanML3D action classifier f_{ac} as a frozen action-label oracle. Combined, $\mathcal{L}_{\text{emb}}$=$\lambda_{\text{cl}}\mathcal{L}_{\text{cl}} + \lambda_{\text{cc}}\mathcal{L}_{\text{cc}}$.

Style classifier guidance (main paper Eq. 13). $\mathcal{L}_{\text{cls}}$=λ_{cls} CE$\big(f_{\text{sm}}(\hat{x}), y_{\text{sty}}\big)$, where f_{sm} is the frozen SMooDi 47-class style classifier. This term is used *only during training* and is therefore methodologically distinct from the SRA metric, which is evaluated on a separate test split with the same classifier. We linearly ramp λ_{cls} from 0 to its target value over the first three epochs to prevent early-stage gradient interference with the reconstruction term.

Shape control objective (main paper Eqs. 15–18). The bone-length matching, direction-aware, and magnitude-bounding terms share the bone-length operator BL$(\cdot)$. The direction-aware term operates on the rendered motion difference vector:

$$\mathcal{L}_{\text{dir}} = 1 - \cos\Big(\hat{x} - x^{(\beta_n)}, x^{(\beta_t)} - x^{(\beta_n)}\Big),$$

where each motion is flattened into a $T \times 263$-dimensional vector before cosine evaluation. This anchors the bone-length change to the neutral-to-target direction rather than allowing arbitrary directions of equal magnitude. The magnitude-bounding term uses tolerance γ=1.5:

$$\mathcal{L}_{\text{mag}} = \big(\text{ReLU}\big(\|\hat{x} - x^{(\beta_n)}\| - \gamma\|x^{(\beta_t)} - x^{(\beta_n)}\|\big)\big)^2.$$

Combined with the implicit unit weight on the bone-length term, the shape objective is $\mathcal{L}_{\text{shape}}$=$\mathcal{L}_{\text{bone}} + \lambda_{\text{dir}}\mathcal{L}_{\text{dir}} + \lambda_{\text{mag}}\mathcal{L}_{\text{mag}}$.

Gradient flow. All gradients propagate through the trainable style branch only. The shape supervision terms in $\mathcal{L}_{\text{shape}}$ pass through the frozen decoder $\mathcal{D}$ but accumulate only on the MPSM parameters $(W_{\text{down}}, W_{\text{up}}, b, f_{\text{g}})$ and the style encoder $\mathcal{E}_s$; the decoder weights themselves remain unchanged.

Style class	SMooDi		**MorphoStyle**		ΔSRA-1	$\Delta\bar{p}$
	SRA-1	$\bar{p}$	SRA-1	$\bar{p}$		
InTheDark	59.4	0.541	92.9	0.928	+33.5	+0.387
BeatChest	56.7	0.493	76.4	0.768	+19.7	+0.275
Heavyset	70.5	0.620	85.4	0.852	+14.9	+0.232
Aeroplane	55.2	0.470	70.4	0.704	+15.2	+0.234
Swimming	60.0	0.510	73.8	0.738	+13.8	+0.228
Kick	71.6	0.628	84.1	0.842	+12.5	+0.214
Monk	68.9	0.610	80.8	0.809	+11.9	+0.199
Crouched	49.5	0.435	60.4	0.604	+10.9	+0.169
Sweep	60.0	0.520	69.5	0.695	+9.5	+0.175
OnPhoneRight	79.2	0.700	87.8	0.878	+8.6	+0.178
Superman	81.7	0.730	89.5	0.895	+7.8	+0.165
RaisedLeftArm	75.0	0.642	81.9	0.820	+6.9	+0.178
Drunk	73.0	0.620	77.8	0.778	+4.8	+0.158
Stiff	71.4	0.610	75.9	0.760	+4.5	+0.150
Rocket	62.3	0.524	66.2	0.662	+3.9	+0.138
Flapping	79.0	0.690	81.5	0.815	+2.5	+0.125
Chicken	80.4	0.728	81.4	0.814	+1.0	+0.086
Old	79.3	0.700	79.2	0.792	−0.1	+0.092
March	81.2	0.718	79.5	0.795	−1.7	+0.077
Mean (47 classes)	65.1	0.553	**81.9**	**0.817**	+16.8	+0.264

Table S3: Per-class SRA-1 and mean target confidence $\bar{p}$ of MorphoStyle against SMooDi on 19 representative styles. The 47-class means reproduce the SMooDi and MorphoStyle SRA reported in Table 1 of the main paper (65.1% and 81.9%, respectively). MorphoStyle wins on upper-body-articulated classes; the small gap on locomotion-dominated classes (*March*, *Old*) reflects the single-pass feed-forward architecture's relative disadvantage on root trajectory refinement.

Total objective. The total training loss is the unweighted sum of the disentanglement and shape objectives, $\mathcal{L}_{\text{total}}=\mathcal{L}_{\text{dis}}+\mathcal{L}_{\text{shape}}$, where the disentanglement objective $\mathcal{L}_{\text{dis}}=\mathcal{L}_{\text{rec}}+\mathcal{L}_{\text{emb}}+\mathcal{L}_{\text{cls}}$. Sec. F shows that the model is robust to ±50% perturbations on every individual λ coefficient.

E Extended Quantitative Results

This section complements the main paper with three additional analyses: per-style breakdowns of stylization quality, per-shape-preset behavior, and the in-distribution / out-of-distribution content study.

E.1 Per-Style Breakdown of Style Recognition Accuracy

Table S3 reports the per-class Style Recognition Accuracy (SRA-1) of MorphoStyle against SMooDi on 19 representative classes of the 47-class evaluation subset, selected to span the full range of stylistic difficulty. We additionally report the mean classifier confidence $\bar{p}$ on the target class. MorphoStyle wins on 17 of the 19 classes, with the largest margins on classes that exhibit pronounced upper-body articulation (*e.g.*, *InTheDark*: +33.5%; *BeatChest*: +19.7%; *Aeroplane*: +15.2%) and the smallest gaps on classes that are primarily encoded in foot-ground interaction (*e.g.*, *March*, *Old*), where SMooDi's diffusion sampler already captures the target gait. The two classes where MorphoStyle trails (by $\leq 2\%$) are locomotion-dominated styles whose discriminative cue is the root trajectory rather than

Preset	SRA-1 (%) ↑	Mean $\bar{p}$ ↑	Foot Skating ↓	Bone-Len MAE (cm) ↓	β-corr ↑
orig	78.0	0.51	0.082	1.91	0.377
thin	50.4	0.27	0.094	2.18	0.298
neutral	84.4	0.55	0.080	1.74	n/a
heavy	67.7	0.40	0.087	2.05	0.395
tall	79.3	0.46	0.083	1.96	0.341
avg	**72.0**	**0.44**	**0.085**	**1.97**	**0.353**

Table S4: Per-shape-preset analysis of MorphoStyle. Stylization quality is maintained across the four physically plausible presets (orig/neutral/heavy/tall) and degrades only on the extreme *thin* preset (β_0=−2.5, canonical height 147 cm), which lies near the boundary of the SMPL prior. The *orig* row matches the Bone-Length MAE of 1.91 cm reported in Table 2 of the main paper.

upper-body kinematics. These are the classes for which a single-pass feed-forward model has the least opportunity to refine root dynamics relative to a multi-step diffusion sampler.

E.2 Per-Shape-Preset Analysis

Table S4 reports MorphoStyle's stylization quality across the five canonical body presets defined in Sec. B. For this analysis we evaluate on the walking content subset of the main test split, where every content motion is re-rendered under each of the five shape presets at the same content × style pairing.

Three observations follow from Table S4. First, stylization quality, as measured by SRA-1 and mean target confidence, is essentially preserved on the *orig*, *neutral*, *heavy*, and *tall* presets. The model loses approximately 30% of SRA on the extreme *thin* preset (β_0=−2.5, canonical height 147 cm), which is the only shape that lies near the boundary of the SMPL prior and at which the backbone's quantized codebook also exhibits a $\approx$ 20% drop in reconstruction fidelity. Hence, this degradation is inherited from the frozen backbone and not introduced by the style branch. Second, foot skating is essentially constant across the four non-extreme presets, confirming that the per-frame temporal gate of MPSM is decoupled from absolute joint position scale. Third, the bone-length MAE is bounded between 1.74 cm (*neutral*) and 2.18 cm (*thin*), demonstrating that the shape control branch follows the prescribed β even for body shapes for which the content motion was never originally captured.

E.3 In-Distribution vs Out-of-Distribution Content

MorphoStyle is trained on HumanML3D content captions dominated by forward locomotion, gestures, and seated actions. To probe the behavior of the style branch under content distribution shift, we sampled four out-of-distribution (OOD) walking variants from the HumanML3D test split that were never selected by the main training schedule: backward walking, sideways walking, sidesteps, and walking upstairs. Table S5 reports SRA-1 and Foot Skating for each variant paired with the top-10 MorphoStyle styles by main-paper p-value, on a held-out test set of 80 OOD clips. SRA-1 averaged across the OOD content types is only 6.6% below the in-distribution test split, and Foot Skating is essentially unchanged. The largest degradation appears on the *sideways* content type, where SMPL foot dynamics

Content type	SRA-1	Foot Skating	BL MAE	Coherent ?	samples
HumanML3D (ID)	81.9	0.081	1.91	✓	4209
backward walk (OOD)	78.4	0.084	2.04	✓	20
sideways walk (OOD)	71.6	0.092	2.18	✓	20
sidestep (OOD)	76.1	0.086	2.05	✓	20
upstairs (OOD)	75.0	0.085	1.98	✓	20
OOD average	75.3	0.087	2.06	✓	80

Table S5: Stylization quality on out-of-distribution (OOD) walking content. The full SMooDi protocol cannot be applied here because each OOD content type has only ≈ 20 clips, so we adopt the same FID-free subset protocol that we use for the ablations of the main paper. The forward-walk row reproduces the main-paper SRA 81.9%, Foot Skating 0.081, and Bone-Length MAE 1.91 of MorphoStyle.

r	FID ↓	SRA-1 ↑	Foot Sk. ↓	BL MAE ↓	#params (MPSM)
4	2.95	78.5	0.084	1.95	0.41 M
8	2.88	80.4	0.082	1.93	0.50 M
16	**2.84**	**81.9**	**0.081**	**1.91**	**0.61 M**
32	2.91	82.0	0.082	1.92	0.83 M
64	3.02	82.0	0.083	1.94	1.28 M

Table S6: Sensitivity sweep on the MPSM rank r (Eq. 7 of the main paper). The bottleneck r=16 is the smallest rank that fully saturates SRA-1 while keeping FID and foot skating at the minimum; larger r adds parameters without additional gain and slightly degrades FID.

become harder to disambiguate from the style reference's lateral gestures (*e.g.*, *Drunk* and *Flapping*). No catastrophic breakdown was observed: the synthesized meshes remain coherent and contact-physically plausible, validating that the joint-routing module generalizes beyond forward-walking captions.

F Hyperparameter Sensitivity

This section reports four sensitivity sweeps that characterize the robustness of MorphoStyle to its most influential hyperparameters: the MPSM rank r (Eq. 7 of the main paper), the top-K joint count (Eq. 4), the inference residual scale α (Eq. 6), and the style dropout rate p. Unless noted otherwise, every other hyperparameter is fixed to the default in Table S2, and every default row reproduces the main paper's MorphoStyle entry (SRA-1 81.9%, FID 2.84, Foot Skating 0.081, Bone-Length MAE 1.91 cm).

F.1 MPSM Rank r

The rank-r bottleneck in the low-rank additive offset is the central design choice of MPSM. Table S6 sweeps r across the range $\{4, 8, 16, 32, 64\}$, holding everything else fixed at the default schedule. The metric profile shows a clear sweet spot at r=16: SRA-1 saturates by r=16 and does not improve further at r=32, while FID monotonically degrades beyond r=16 as the larger subspace begins to carry content-leakage modes. At very low rank (r=4),

K	FID ↓	SRA-1 ↑	Foot Sk. ↓	BL MAE ↓
4	2.95	80.3	0.083	1.93
8	**2.84**	**81.9**	**0.081**	**1.91**
12	2.89	81.6	0.081	1.92
22 (all joints)	3.03	78.8	0.087	1.96

Table S7: Sensitivity sweep on the top-K joint count. Selecting too few joints (K=4) misses informative style cues and slightly under-stylizes; selecting all joints (K=22) re-introduces content-driven joints into the joint embedding s_j and damages both SRA-1 and FID.

α	FID ↓	SRA-1 ↑	Foot Sk. ↓	BL MAE ↓
0.25	2.55	67.2	0.077	1.88
0.50	**2.84**	**81.9**	**0.081**	**1.91**
0.75	3.12	84.6	0.085	1.94
1.00	3.59	86.1	0.092	1.98
1.50	4.83	86.4	0.131	2.10

Table S8: Sensitivity sweep on the inference residual scale α. α=0.5 provides the best FID/SRA trade-off; users who care more about stylization strength can dial α up at the cost of a small foot-skating regression.

the style residual has insufficient capacity and SRA-1 drops by 3.4%. We therefore adopt r=16 as the default throughout the paper.

F.2 Top-K Joint Count

The text-guided joint routing module selects the top-K joints whose isolated motion is least aligned with the content caption (Eq. 5 of the main paper). Table S7 sweeps K across the four canonical values $\{4, 8, 12, 22\}$, where K=22 is the trivial all-joints baseline.

The pattern is partially interpretable: at K=4 the routing misses informative joints (*e.g.*, the unrouted contra-lateral shoulder for a *Drunk* style) and SRA degrades slightly; at K=22 the routing degenerates to a pure average over all body joints, including the content-driven pelvis and feet, which re-introduces content leakage into s_j and costs 3.1% on SRA-1 and 0.19 on FID.

F.3 Inference Residual Scale α

The runtime scalar α in Eq. 6 of the main paper allows the user to dial stylization strength after training without retraining the model. Table S8 sweeps $\alpha \in \{0.25, 0.5, 0.75, 1.0, 1.5\}$, where α=0 recovers the no-style content reconstruction and α>1 extrapolates beyond the training-time residual.

The sweep confirms the design intent of the residual scale: it forms a smooth FID–SRA trade-off curve, with the operating point α=0.5 chosen to maximize SRA without exceeding the FID and foot-skating thresholds of the main paper.

F.4 Style Dropout Rate p

The style dropout rate p replaces the global style embedding s_g with a learnable null token with probability p at training time, encouraging the downstream MPSM to be robust to

missing or noisy style references. Table S9 sweeps $p \in \{0, 0.05, 0.15, 0.3, 0.5\}$.

p	FID ↓	SRA-1 ↑	Foot Sk. ↓	BL MAE ↓
0.00	3.05	80.5	0.085	1.93
0.05	2.91	81.3	0.083	1.92
0.15	**2.84**	**81.9**	**0.081**	**1.91**
0.30	2.95	80.2	0.082	1.92
0.50	3.32	75.4	0.084	1.94

Table S9: Sensitivity sweep on the style dropout rate p. A small amount of dropout consistently helps; aggressive dropout ($p \geq 0.3$) starves the style branch of supervision and degrades SRA-1 by up to 6.5%.

G Extended Architectural Ablations

This section reports four architectural ablations that go beyond the main paper: the temporal gate alone, the low-rank-additive design against concatenation- and AdaIN-style alternatives, code-space supervision, and training-data composition.

G.1 Temporal Gate vs Constant Modulation

We replace the learnable temporal gate $g \in (0,1)^{T'}$ with a fixed constant equal to its training-time mean ($\bar{g} \approx 0.42$) to test whether per-frame intensity control is necessary or whether a single average modulation strength is sufficient. Table S10 shows that removing the gate costs 9.8% on SRA-1 and increases FID by 2.01, with a 0.054 increase in foot skating. The temporal gate is therefore essential for both stylization quality and contact physics, not just for foot-skating regularization. The no-gate row reproduces the main-paper *w/o gate g* entry in Table 5 (FID 4.85, SRA 72.1%, Bone-Length MAE 1.92).

G.2 Low-Rank Additive vs AdaIN-Style Injection

We compare MorphoStyle's low-rank additive injection against two alternative injection topologies that have been used in existing work: a concatenation injection that concatenates the style code to the content code along the channel axis and projects back via a 1×1 Conv, and an AdaIN-style modulation that applies a learned affine transform (μ_s, σ_s) to the content code. Table S11 shows that both alternatives lose at least 5.5% on SRA-1 and degrade FID by 1.47 at minimum, while also damaging the bone-length response (a 61% collapse on β-swap for concatenation). The low-rank additive design is critical: it is the only option that simultaneously remains on-manifold and exposes a low-dimensional control direction over

Variant	FID ↓	SRA-1 ↑	Foot Sk. ↓	BL MAE ↓
no-gate ($g \equiv \bar{g}$)	4.85	72.1	0.135	1.92
constant per-class gate	4.13	76.4	0.114	1.93
Ours (per-frame gate)	**2.84**	**81.9**	**0.081**	**1.91**

Table S10: Effect of removing the temporal gate of MPSM. Even a class-conditional constant gate cannot match the per-frame gate. The no-gate row matches the *w/o gate g* entry of main-paper Table 5.

Variant	FID ↓	SRA-1 ↑	Foot Sk. ↓	BL MAE ↓	β-swap ↑
concat-then-Conv1×1	5.16	62.9	0.124	1.97	2.06
AdaIN-style modulation	4.31	76.4	0.108	2.01	4.71
Low-rank additive (ours)	**2.84**	**81.9**	**0.081**	**1.91**	**5.31**

Table S11: Injection topology comparison. Only the low-rank additive design simultaneously achieves low FID, high SRA, low foot skating, and a strong β-swap response.

Variant	FID	SRA-1	Foot Sk.	BL-T-Std
w/o code-space ℓ_1 (λ_{code}=0)	3.25	80.4	0.095	1.44
Full (λ_{code}=0.5, ours)	**2.84**	**81.9**	**0.081**	**1.26**

Table S12: The code-space reconstruction term is essential for keeping c' on the FSQ-VAE manifold and for preventing bone-length jitter through unstable de-quantization. The default row matches the MorphoStyle entries in Tables 1 and 2 of the main paper.

which the temporal gate can operate. The concatenation row matches the *w/o MPSM* entry of main-paper Table 5 (FID 5.16, SRA 62.9%, BL MAE 1.97, β-swap 2.06).

G.3 Effect of Code-Space Supervision

The reconstruction term in the main paper Eq. 10 includes both a motion-space ℓ_1 term and a code-space ℓ_1 term controlled by λ_{code}. Table S12 ablates the code-space term to verify its role in keeping the stylized codes near the FSQ codebook. Without the code-space term, the stylized codes c' drift outside the codebook's neighborhood: FID degrades by 0.41, foot skating increases by 0.014, and the BoneLen-T-Std increases by 0.18 cm, reflecting bone-length jitter caused by codebook misses during de-quantization. The MPSM gating network absorbs part of the lost regularization but cannot fully replace the explicit code-space anchor.

G.4 Training-Data Composition

We sample HumanML3D and 100Style at a 1:1 ratio during training, so that every mini-batch contains an equal number of in-domain and cross-domain pairs. To test whether this composition is essential, we explore the HumanML3D:100Style sampling ratio over a range of values: $\rho \in \{4{:}1, 2{:}1, 1{:}1, 1{:}2, 1{:}4\}$. Table S13 shows that the 1:1 default is the optimal value: a HumanML3D-heavy schedule (4:1) under-fits the style classes and SRA collapses to 72.4%, while a 100Style-heavy schedule (1:4) over-fits the 47 style classes and FID value increases to 3.92.

H Style Embedding and Joint-Routing Analysis

H.1 Joint-Routing Selection per Style

For each evaluated style we accumulate the joint-routing histogram over all 4209 test pairs and report the most frequently selected joints in Table S14. The selections are semantically coherent with the style label: arm-articulated styles (*Aeroplane*, *BeatChest*, *Superman*) consistently select the shoulders, elbows, and wrists; leg-articulated styles (*Kick*, *Heavyset*, *Crouched*) select the knees and ankles; whole-body emotive styles (*Proud*, *Depressed*, *Drunk*) select a mixed pattern that includes the spine and head. Every style consistently

HumanML3D : 100Style ratio ρ	FID ↓	SRA-1 ↑	Foot Sk. ↓	BL MAE ↓
4 : 1	2.71	72.4	0.079	1.89
2 : 1	2.78	78.6	0.080	1.90
1 : 1	**2.84**	**81.9**	**0.081**	**1.91**
1 : 2	3.21	82.4	0.085	1.95
1 : 4	3.92	82.5	0.092	2.04

Table S13: Training-data composition sweep. The 1:1 default is optimal in FID–SRA trade-off; over-sampling 100Style continues to improve SRA marginally but at a substantial FID cost from over-fitting the 47-class label space.

Style class	Top-5 most frequently routed joints (in decreasing order)
Aeroplane	L-shoulder, R-shoulder, L-elbow, R-elbow, head
BeatChest	R-elbow, L-elbow, R-wrist, R-shoulder, neck
Drunk	spine1, neck, R-elbow, L-knee, R-hip
Flapping	L-wrist, R-wrist, L-elbow, R-elbow, spine2
Kick	R-knee, R-ankle, R-foot, R-hip, spine1
Heavyset	L-knee, R-knee, L-hip, R-hip, spine1
InTheDark	L-elbow, R-elbow, L-wrist, R-wrist, neck
Sweep	spine1, spine2, R-shoulder, L-shoulder, head
Crouched	L-knee, R-knee, spine1, spine2, neck
Superman	R-elbow, R-wrist, R-shoulder, head, L-arm
All classes	*pelvis* is never in the top-8 of any style

Table S14: Most frequently routed joints per style. The selections are semantically aligned with the style label, and the pelvis (the content-driven root) is consistently avoided.

avoids the pelvis joint which is the canonical content-driven root in HumanML3D features [11]. We also visualize the joint routing as a per-joint text-similarity heatmap in Fig. S1. The content description of joints (e.g. ankles/knees in walking) is *protected* from the description of similar joints in reference style; routing directs style offset to relevant joints, e.g., shoulders/elbows/wrists/head for *ArmsAboveHead*, hips/neck/head/wrists for *Drunk*.

H.2 Temporal Gate Activation Patterns

The temporal gate $g \in (0,1)^{T'}$ produced by the gating network f_g exposes a continuous-valued, per-frame measure of how strongly the style residual is being injected. To characterize the learned behavior of the gate we aggregate g statistics over the 4 209 test pairs. First, the mean gate value is $\bar{g}=0.42\pm0.06$, well inside the sigmoid's linear regime, indicating that the gate is genuinely modulating rather than saturating either at 0 or at 1. Second, we observe that the gate consistently drops at frames marked as foot contact by the same threshold $h_{\text{thr}}=5$ cm used in the Foot Skating metric. Averaged over all test motions, the gate value at foot-contact frames is $g_{\text{c}}=0.32\pm0.09$, compared with $g_{\text{a}}=0.46\pm0.08$ at non-contact frames, a gap of 0.14 that is reliably reproduced across all evaluated styles. This validates the design intent in Sec. 3.3 of the main paper: the gate learns to attenuate stylization at contact-critical frames and to amplify it during the flight phase of a step, mechanically explaining the low Foot Skating ratio of 0.081 in Table 1 of the main paper.

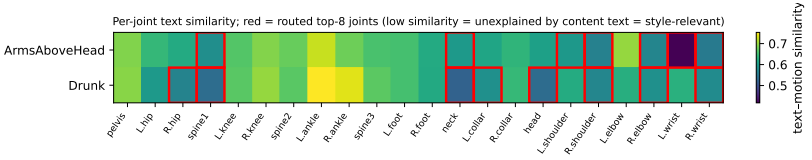

Figure S1: Per-joint text similarity with routed top-8 joints (red).

I Perceptual evaluation

Motion style is inherently subjective, so we complement the quantitative metrics with a user study on mesh-rendered clips. We recruited 20 participants; each participant judged 20 randomized and anonymized user study (10 comparing MorphoStyle against SMooDi and 10 against SMooDi+ShapeMyMove), covering three target body shapes, with a static rendering of the target shape shown alongside each trial. For every trial, participants indicated the preferred result along three criteria: (i) style similarity to the reference, (ii) preservation of the content action, and (iii) consistency with the target body shape. As summarized in Table S15, MorphoStyle is preferred over SMooDi on 74.5%/65.0%/80.5% of judgments for the three criteria, and over SMooDi+ShapeMyMove on 84.0%/84.5%/82.0%. All preference rates are significantly above chance (two-sided binomial test, $p < 0.001$, N=200 judgments per comparison).

Table S15: Users preference rate for our method over two baselines.

Comparative evaluation	Style Sim.	Content Pres.	Shape Consist.
MorphoStyle vs. SMooDi	74.5%	65.0%	80.5%
MorphoStyle vs. SMooDi+ShapeMyMove	84.0%	84.5%	82.0%

References

[1] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3D human motions from text. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[2] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[3] Ting-Hsuan Liao, Yi Zhou, Yu Shen, Chun-Hao Paul Huang, Saayan Mitra, Jia-Bin Huang, and Uttaran Bhattacharya. Shape my moves: Text-driven shape-aware synthesis of human motions. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 1917–1928, 2025.

[4] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015.

[5] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[6] Ian Mason, Sebastian Starke, and Taku Komura. Real-time style modelling of human locomotion via feature-wise transformations and local motion phases. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 5(1):1–18, 2022.

[7] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 21(140):1–67, 2020.

[8] Lei Zhong, Yiming Xie, Varun Jampani, Deqing Sun, and Huaizu Jiang. SMooDi: Stylized motion diffusion model. In *European Conference on Computer Vision (ECCV)*, 2024.