

# Diffusion-based Human Motion Style Transfer with Semantic Guidance

Lei Hu<sup>1,2</sup> , Zihao Zhang<sup>1</sup> , Yongjing Ye<sup>1</sup> , Yiwen Xu<sup>1,2</sup> , and Shihong Xia<sup>†1,2</sup>

<sup>1</sup>Institute of Computing Technology, Chinese Academy of Sciences

<sup>2</sup>University of Chinese Academy of Sciences

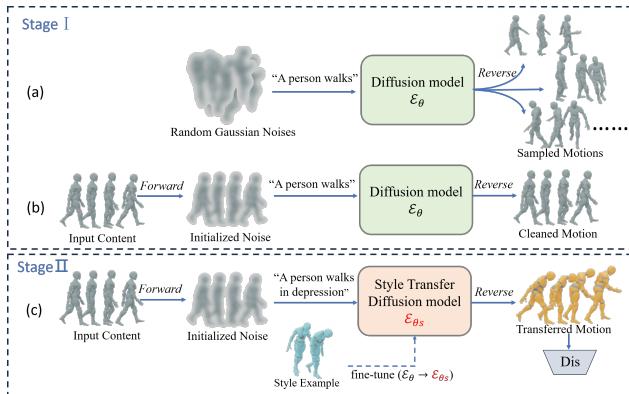
## Abstract

3D Human motion style transfer is a fundamental problem in computer graphic and animation processing. Existing AdaIN-based methods necessitate datasets with balanced style distribution and content/style labels to train the clustered latent space. However, we may encounter a single unseen style example in practical scenarios, but not in sufficient quantity to constitute a style cluster for AdaIN-based methods. Therefore, in this paper, we propose a novel two-stage framework for few-shot style transfer learning based on the diffusion model. Specifically, in the first stage, we pre-train a diffusion-based text-to-motion model as a generative prior so that it can cope with various content motion inputs. In the second stage, based on the single style example, we fine-tune the pre-trained diffusion model in a few-shot manner to make it capable of style transfer. The key idea is regarding the reverse process of diffusion as a motion-style translation process since the motion styles can be viewed as special motion variations. During the fine-tuning for style transfer, a simple yet effective semantic-guided style transfer loss coordinated with style example reconstruction loss is introduced to supervise the style transfer in CLIP semantic space. The qualitative and quantitative evaluations demonstrate that our method can achieve state-of-the-art performance and has practical applications. The source code is available at <https://github.com/hlcddy/diffusion-based-motion-style-transfer>.

## CCS Concepts

- Computing methodologies → Motion processing; Artificial intelligence;

## 1. Introduction



**Figure 1:** The concept figure of our framework. (a) The pre-trained text-to-motion generative model; (b) The motion-to-motion translation model based on diffusion; (c) The fine-tuning for human motion style transfer.

† Corresponding author: xsh@ict.ac.cn

The stylized human motion can greatly enhance the artistic tension of the character, making it more lively and expressive in animation, gaming, and virtual reality. Currently, the primary way to acquire 3D human motion is motion capture. Nevertheless, it is infeasible to capture all of the infinite stylistic variations, and the capture relies on skilled actors, thereby restricting the scale of current style motion datasets. Consequently, human motion style transfer, which aims to transfer the style from a reference style example to other content motions, emerges as a compelling technique to enrich the human motion variations.

It is challenging to provide a precise mathematical definition of motion style since variations in both motion content and style are manifested in low-level motion features, such as joint rotations and global velocity. Therefore, style transfer, as a non-trivial problem, has been widely studied for decades. Prior approaches [HPP05, MXH\*10, XWCH15] typically model style by examining differences between motion examples. However, it is not easy to obtain paired motion segments manually, and the laborious correspondence and preprocessing (e.g., aligning motion pairs through dynamic temporal warping [RJ93]) hinders the model's ability to scale incrementally. Recently, AdaIN-based methods [AWL\*20, JPL22] have emerged as the predominant approach for style transfer, offering the capability to disentangle con-

tent and style for unpaired motions and to cluster style features in the latent space. Nevertheless, AdaIN-based techniques necessitate datasets with balanced style distributions to train the clustered latent space. In practical scenarios, however, we may only have access to a limited number of stylistic examples (or even just a single clip) compared to a vast amount of neutral motion. Additionally, the AdaIN-based methods require re-training the model when confronted with unseen action types or styles that fall outside the dataset's distribution. Thus, the key issue to be addressed in human motion style transfer is: Given only one unseen style example, how to obtain a style transfer model capable of transferring a variety of content motions by few-shot learning.

Facing the challenge above, in this paper, we introduce a two-stage framework to split the human motion modeling and style transfer, as illustrated in Figure 1. In stage I, **(a)** we pre-train a diffusion-based text-to-motion generative model  $\epsilon_\theta$  on a text-motion dataset. This model, functioning as a generative prior, can sample from Gaussian noise and generate diverse natural human motions from scratch under text guidance. We chose the text-motion generative model as a prior because the text prompts can provide more motion information relative to the content/style labels in the subsequent style transfer learning. **(b)** If we add noise to the input content motion by the forward process and denoise it by the reverse process of the diffusion model, the pre-trained text-to-motion generative model  $\epsilon_\theta$  can be seamlessly converted into a motion-to-motion translation model. And the noising and denoising processes inevitably introduce some variations to the input motion.

Given that motion styles can be viewed as special variations of human motions, we are inspired to regard the reverse process of the diffusion model as the style translation process. Specifically, in stage II, **(c)** given one style example, we fine-tune the diffusion network ( $\epsilon_\theta \rightarrow \epsilon_{\theta_s}$ ) to enable the reverse process of  $\epsilon_{\theta_s}$  to convert neutral noised motions into stylized motions. To guide the few-shot learning, we also employ a pre-trained motion-semantic discriminator to distinguish whether the transferred motion is semantically consistent with the content of the source motion and the target style.

The proposed two-stage learning framework enables us to use a large text-motion dataset with unbalanced style distribution to pre-train a robust prior model. This ensures that the model is suitable for a wide range of content motion inputs during the second stage of style transfer fine-tuning. When presented with a style example, we only need to follow the stage II and fine-tune the  $\epsilon_\theta$  to obtain a corresponding style transfer model, without re-training on the whole dataset.

During the fine-tuning, to perform the style transfer and preserve the original content of the input motion, we introduce two loss functions: the style example reconstruction loss  $\mathcal{L}_{sr}$  and the semantic-guided transfer learning loss  $\mathcal{L}_s$ . Specifically, for  $\mathcal{L}_{sr}$ , we first use the generative prior  $\epsilon_\theta$  and motion in-painting technique to generate an aligned neutral motion corresponding to the style example. Subsequently, we enforce the style transfer diffusion model  $\epsilon_{\theta_s}$  to translate the corresponding neutral motion into a motion resembling the style example. For  $\mathcal{L}_s$ , we use a pre-trained motion-semantic discriminator to assess the alignment of the output motions with the semantics of the original content and the style of the example motion in the CLIP [RKH<sup>21</sup>] space.

The contributions of our work can be summarized as follows:

- We propose a novel style transfer framework, which regards the reverse process of the diffusion model as a style translation process.
- We propose a novel semantic-guided style transfer learning strategy, which can supervise human motion style transfer in the CLIP semantic space.
- Extensive experiments demonstrate that the proposed framework achieves unseen style transfer on only one style example through fine-tuning.

## 2. Related Works

In this section, we will first review work related to human motion style transfer (2.1) and then review techniques related to motion generation (2.2).

### 2.1. Human Motion Style Transfer

**Example-based style transfer.** Human motion style transfer has been a long-standing problem in the field of computer graphics for decades. The early example-based methods mostly regard the style transfer as a translation problem and employ the examples to learn the translation model [BH00, HPP05, WFH07, IAF09, MXH<sup>10</sup>, XWCH15, YM16]. [HPP05] learn the mapping between motions with identical content but different styles by a linear time-invariant system, which can then be used to transfer a new motion to the learned style in a real-time manner. [MXH<sup>10</sup>] employed a Bayesian network to effectively model the relationship between user-specified parameters and latent variation parameters, which capture the style variation. [XWCH15] employs a series of local mixtures of auto-regressive models to learn style differences and achieve notable results on unlabeled, heterogeneous motion data. [YM16] proposes a spectral style transfer to circumvent the costly and time-consuming requirement of style databases. Typically, the matching motion pairs necessary for these approaches are either manually provided or searched in motion databases.

**Human motion style transfer in latent space.** In recent years, with the development of deep learning, various advanced methods have been employed on human motion style transfer tasks. [HSK16, HHKK17], [DHS<sup>19</sup>] and [WDL<sup>21</sup>] construct the motion manifold by the convolutional network, CVAE, and CycleGAN [ZPIE17], respectively. They explicitly define the Gram matrix [GEB16] of the deep features as the human motion style, while the feature as content. The concept proposed by [DHS<sup>19</sup>] bears a resemblance to our approach, which uses a few style examples and a neutral motion dataset to achieve style transfer. Nonetheless, their transfer is limited to locomotion.

Some methods integrate style modeling with motion generation. [MSZ<sup>18</sup>, MSK22, WYF<sup>21</sup>] regard the motion style as a condition in real-time motion synthesis. [MSZ<sup>18</sup>] introduce residual adapters to adjust the weights of PFNN [HKS17] to transfer the locomotion style during the real-time controlling. [WYF<sup>21</sup>] proposes a generative flow model and uses the style feature as a condition to synthesize the stylized human motion in an autoregressive manner. [MSK22] extend the work [MSZ<sup>18</sup>] and proposes a style

modulation network to extract style information from motion clips and use it to modulate the hidden layers of the motion synthesis network. However, most of these methods are also restricted to style transfer for locomotion.

**AdaIN-based human motion style transfer.** Due to the remarkable performance in image style transfer [HB17], AdaIN has been extensively employed in the human motion style modeling [AWL\*20, PJL21, JPL22, SJL\*23]. It implicitly decouples the motion into content and style codes and enables style transfer through code recombination. This type of method can be regarded as a special case of human motion style transfer methods in latent space. [AWL\*20] is the first work to use AdaIN in the human motion style transfer task. [JPL22] achieves style control over different body parts using part-based AdaIN. FineStyle [SJL\*23] proposes a fine-grained style transfer framework by introducing the content label and fusing it with style code in the latent space for more accurate modeling. [AZL23] uses AdaIN to modulate the features in the latent diffusion module to achieve stylized gesture generation with different style prompts. The AdaIN-based approach has shown promising results in addressing style transfer between non-locomotion. However, the methods require a dataset with balanced style motion to learn the clustered latent space. Therefore, it is difficult to learn style information when given only one unseen style example.

Our framework does not require the dataset to be stylistically balanced when pre-training the generative prior, and in the fine-tuning stage, we need only one style example to generate a style transfer diffusion model that can be applied to a variety of content motion inputs.

## 2.2. Human Motion Generation

Generating natural-looking human motion with user-specific constraints has always been a research focus in the field of human motion synthesis. As for traditional constraints, auto-regressive networks [HKS17, SZKZ20, WCX19, LCVDP20] have demonstrated flexible motion control results given joystick signals. As for sparse constraints such as VR trackers, physics-based motion tracking and generation methods [YLHX22, WWY22] have achieved remarkable performance. Recently, there has been a substantial surge of interest in cross-modal constraints, especially those with semantic prompts [PBV21, ZHZX22, APBV22, GZZ\*22, JCL\*23, ZZC\*23, ZHZX23]. The primary approach of these works focuses on utilizing appropriate generative models in latent space. For example, ACTOR [PBV21] and TEACH [APBV22] leverage transformer-VAE architecture to construct a mapping between label/text prompts and motions. Similarly, Vector-Quantized VAE(VQ-VAE) [VDOV\*17], which quantize continuous latent into discrete representations, demonstrate its ability on text-to-motion tasks [JCL\*23, ZZC\*23, ZHZX23] combined with Generative Pre-trained Transformer [BMR\*20].

Unlike encoder-decoder frameworks, diffusion-based methods directly sample from the original motion space, providing a more stable training approach for motion generation. Denoising Diffusion Probabilistic Models [HJA20] combined with CLIP textual features is wildly used for text-driven motion synthesis

tasks [TRG\*23, ZGP\*23, ZCP\*24]. Among them, the conditioned-synthesis pipeline introduced by MDM [TRG\*23] showcases its flexibility in adapting to various new tasks, including physics constraints [YSI\*23] and music-dance synthesis [TCL23]. Moreover, PriorMDM [STKB24] uses MDM as a generative prior and extends its ability on motion inpainting. SinMDM [RLT\*24] employs the U-Net architecture in the diffusion model and uses only single-segment motion as training data. Particularly, they implement style transfer as a special case of harmonization. [YYY\*24] connects the noise spaces of different style motion based on Dual Diffusion Implicit Bridges [SSME23] and ensures the preservation of motion content by keyframe manifold constraint gradients. In our work, inspired by the StyleDiffusion in image [WZX23], we consider the reverse process of the diffusion model as a style translation process, which models the style difference between style-neutral examples.

## 3. Method

In this section, we will first introduce the motion representation in Sec 3.1. Then, we review the preliminary knowledge of the diffusion models (Section 3.2). In Sec 3.3, we will describe stage I: pre-training of generative model and motion-semantic discriminators. In Sec 3.4, we will introduce stage II: the fine-tuning for style transfer. Finally, in Sec 3.5, we describe the style transfer at inference time and the post-process.

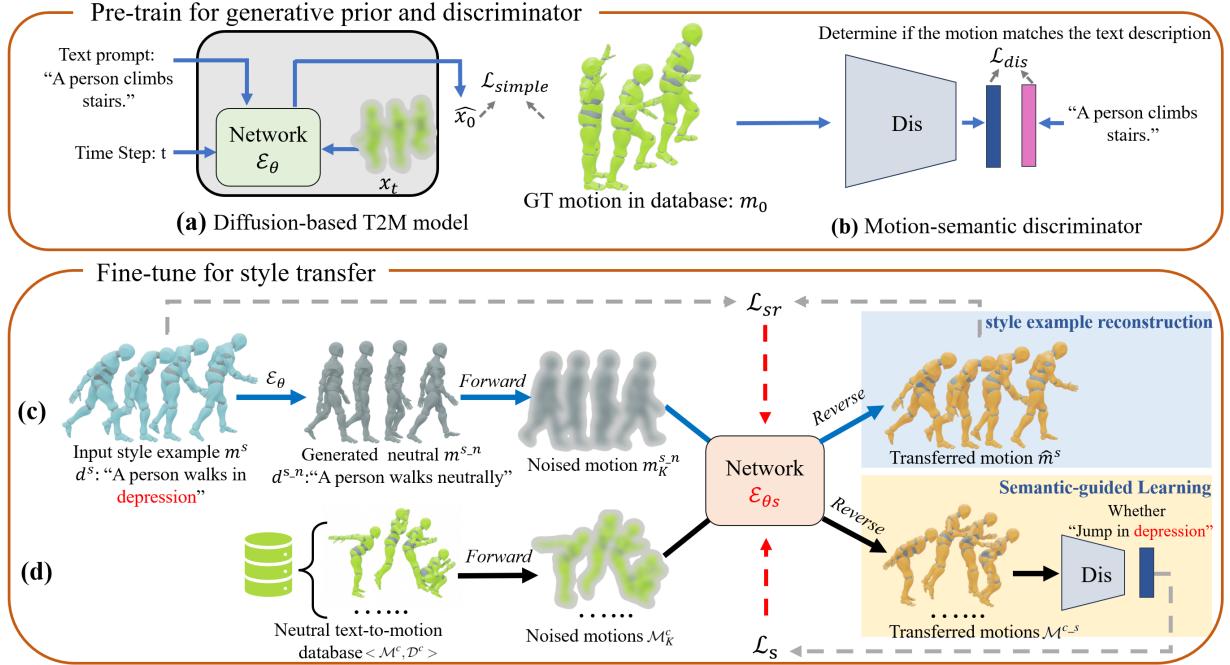
### 3.1. Motion Representation

A motion sequence  $m \in \mathcal{M}$  is represented by a temporal set of  $f$  poses. Each pose feature comprises the following three components, namely, the root joint vector  $\mathbf{o} \in \mathbb{R}^4$ , local joint positions relative to the root  $\mathbf{p} \in \mathbb{R}^{(J-1) \times 3}$  and local joint rotations  $\mathbf{r} \in \mathbb{R}^{J \times 6}$  relative to their parent joint coordinate frame, where  $J$  denotes the number of joints. Additionally, the root joint vector  $\mathbf{o}$  includes the y-axis angular velocity  $\mathbf{o}^r$ , xz-plane linear velocity  $\mathbf{o}^x, \mathbf{o}^z$ , and absolute root height on the y-axis  $\mathbf{o}^h$ . The rotation representation in  $\mathbf{r}$  is the 6D vector defined in the work [ZBL\*19], which is actually the first two column vectors of the rotation matrix. The y-axis angular velocity  $\mathbf{o}^r$  of the root joint is obtained from the change in the human-facing direction. The facing direction is calculated similarly to PFNN [HKS17], which averages the vector between the hip joints and the vector between the shoulder joints and takes the cross product with the upward direction.

To obtain facing-independent motion features, we further localize the  $\mathbf{o}^x, \mathbf{o}^y, \mathbf{p}$ , and the root joint rotation  $\mathbf{r}_0$  based on the facing direction of the current pose. These processed features are concatenated to formulate  $\mathbf{v} = [\mathbf{o}, \mathbf{p}, \mathbf{r}]$  as pose representation, and the motion  $m$  of length  $f$  can be expressed as  $m = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_f]$ . We denote the style motion and content motion as  $m^s$  and  $m^c$ , respectively. (The motion lengths of  $m^s$  and  $m^c$  can be different.)

### 3.2. Preliminary of the Diffusion Models

The diffusion models [HJA20, SME21] consist of two processes: Forward diffusion process and Reverse denoising process. The forward process is modeled as a Markov chain, where the Gaussian noise is gradually added to real data  $m_0 \in \mathcal{M}$  until  $m_T$  is equivalent to an isotropic Gaussian noise. For a single forward step  $t$ , the



**Figure 2:** Overview of our learning process. In the pre-training stage, we train the diffusion-based T2M model  $\epsilon_\theta$  (a) as generative prior and a motion-semantic discriminator (b) for measuring the similarity between motion and textual prompts. In the fine-tuning stage, we employ the style example reconstruction loss (c) to utilize the style information provided by example  $m^s$ . Concurrently, we use the pre-trained motion-semantic discriminator to guide the transferred motion to match the semantics of the original content and the style of the example motion (d). Notably, input style example, generated neutral motion, content motion, and output transferred motion are colored light blue, grey, green, and yellow, respectively.

noising process can be formulated as follows (note the time step  $t$  is not the same concept as the motion frame  $f$ ):

$$q(m_t | m_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}m_{t-1}, (1 - \alpha_t)I) \quad (1)$$

where  $q$  is the conditional probability distribution. The  $\alpha_t \in (0, 1)$  are constant hyper-parameters and  $\alpha_1 > \alpha_2 > \dots > \alpha_T$ ,  $I$  is an identity matrix. When  $\alpha_t$  is small enough,  $m_T$  can be approximated as a standard normal distribution  $\mathcal{N} \sim (0, 1)$ . With the reparameterization trick, we can sample  $m_t$  at any  $t$  step directly based on the original data  $m_0$ :

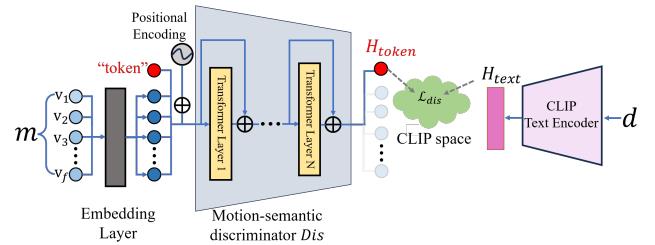
$$m_t = \sqrt{\bar{\alpha}_t}m_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon \quad \epsilon \sim \mathcal{N}(0, 1) \quad (2)$$

where  $\bar{\alpha}_t$  is the cumulative product from  $\alpha_1$  to  $\alpha_t$ . From the Equation 2, it can be seen that the noised feature  $m_t$  consists of an  $m_0$  term and a random Gaussian noise  $\epsilon$  term, and the effect of  $m_0$  on  $m_t$  gradually diminishes as  $t$  increases.

The reverse process is the inverse of the above process. The true transition probability  $q(m_{t-1} | m_t)$  is approximated using a neural network  $\epsilon_\theta$ . Instead of predicting the noise  $\epsilon_t$  as originally proposed by [HJA20], we follow [RDN\*22, TRG\*23] and directly predict the initial motion:

$$\hat{m}_{0(t)} = \epsilon_\theta(m_t, t, c) \quad (3)$$

Where the  $c$  is the conditioning signal in conditioned motion synthesis.  $\hat{m}_{0(t)}$  is the initial motion predicted from  $m_t$  at  $t$  step.



**Figure 3:** The architecture of our motion-semantic discriminator. We introduce a trainable "token" (short for "semantic token") for motion pooling and align the motion-semantic feature  $H_{token}$  with the text feature  $H_{text}$  in the CLIP space.

### 3.3. Pre-train for Generative Prior and Discriminator

As illustrated in Figure 2, we will first pre-train a diffusion-based text-to-motion (T2M) model  $\epsilon_\theta$  as generative prior and a motion-semantic discriminator  $Dis$  in stage I.

Previously, autoencoders [HSK16, DHS\*19] were commonly used to build motion manifolds for downstream tasks. While reconstruction of input in autoencoders can learn a good motion prior from the dataset, the statistical model only contains modal information about the 3D motion and no other control signals. The T2M

generative model  $\epsilon_\theta$  allows us to use semantics as a condition to generate motion by sampling in the noise space, which is beneficial for our subsequent fine-tuning of style transfer.

**Pretraining of T2M generative model  $\epsilon_\theta$ .** Given the motion-text pair  $\langle m, d \rangle$  in dataset, the training process of  $\epsilon_\theta$  is similar to [TRG<sup>\*</sup>23]. Text features produced by the CLIP text encoder are first added with time-step embeddings, and then concatenated with noised motion feature  $m_t$  sampled by the forward process (see Equation 2). Finally, the concatenated features are fed into the network  $\epsilon_\theta$  to clean the motion:

$$\hat{m}_{0(t)} = \epsilon_\theta(m_t, t, d) \quad (4)$$

The  $\epsilon_\theta$  is trained by encouraging the prediction  $\hat{m}_{0(t)}$  at each time step  $t$  to be close to the initial signal  $m_0$ , i.e.,  $m$ , the loss function can be written as:

$$\mathcal{L}_{simple} = \mathbb{E}_{m, t \sim [1, T]} [\|m_0 - \epsilon_\theta(m_t, t, d)\|_2^2] \quad (5)$$

**Pretraining of motion-semantic discriminator  $Dis$ .** Different from the original discriminator network [CWD<sup>\*</sup>18] which only judges fake or real, our proposed motion-semantic discriminator aims to determine whether the motion matches the semantics of a given text. This judgment is achieved by comparing the cosine similarity between the motion features and the text features (see Figure 2 (b)).

The architecture of the motion-semantic discriminator is shown in Figure 3. In order to process motion of different lengths, we follow [PBV21, HZZ<sup>\*</sup>23] and introduce a trainable parameter called "semantic token". The "semantic token" has the same dimension as the pose embeddings, and it will be concatenated with them and fed into the discriminator  $Dis$  after position encoding. Among the output features, we only keep the feature  $H_{token}$  at the position corresponding to the "semantic token" for pooling purposes. During pre-training, the pooled motion feature is encouraged to be aligned with its paired text feature in the CLIP space using contrast learning:

$$\mathcal{L}_{dis} = \mathbb{E}_{m, d} [1 - \frac{Dis(m) \cdot E_{text}(d)}{\max(\|Dis(m)\| \times \|E_{text}(d)\|, \delta)}] \quad (6)$$

Where  $\langle m, d \rangle$  is from the motion-text dataset and  $E_{text}$  is the off-the-shelf CLIP text encoder. The loss  $\mathcal{L}_{dis}$  measures the cosine similarity between vectors  $Dis(m)$  and  $E_{text}(d)$ , where  $\delta$  is a small value to avoid division by zero, which is set to  $10^{-6}$  in our setting.

### 3.4. Fine-tune for Style Transfer

The overview of fine-tuning for style transfer is shown in the bottom half of Figure 2. Given one style example  $m^s$  and a prompt  $d^s$  for its content and style, we aim to obtain a style transfer diffusion model  $\epsilon_{\theta_s}$  related to  $m^s$  through fine-tuning ( $\theta \rightarrow \theta_s$ , different style examples will correspond to different fine-tuned model). After fine-tuning,  $\epsilon_{\theta_s}$  can transfer the style of  $m^s$  to diverse neutral motions while preserving the original motion content. To achieve that, we introduce a style example reconstruction loss  $\mathcal{L}_{sr}$  (see Figure 2(c)) coordinated with a CLIP-based semantic-guided learning loss  $\mathcal{L}_s$  (see Figure 2(d)) to jointly fine-tune the network.

**Style example reconstruction.** As shown in Figure 2(c), in order to fully utilize the style information provided by the input example  $m^s$ , we design the style example reconstruction loss  $\mathcal{L}_{sr}$ . Specifically, we first obtain the neutral motion  $m^{s-n}$  corresponding to  $m^s$ , then we add noise to the neutral motion through the forward process and finally reconstruct the style example by the reverse process to learn the difference between  $m^{s-n}$  and  $m^s$ .

To generate style-neutral motion pair  $\langle m^s, m^{s-n} \rangle$ , previous translation-based style transfer methods obtain matching motions by manually providing [HPP05] or by KNN search over a mocap dataset [XWCH15] and use DTW [RJ93] or GTW [ZDIT12] to apply spatial and temporal warps to align the two motions. However, these procedures are tedious and time-consuming. In this work, we propose to use the pre-trained  $\epsilon_\theta$  as a statistical model to get the well-aligned motion pair  $\langle m^s, m^{s-n} \rangle$ . Specifically, we first add G-step noise to  $m^s$  to obtain the noised motion  $m_G^s$  to remove motion details by the forward process (see Equation 2). During the reverse process, we use  $m_G^s$  as the initial noised motion feature coordinated with a neutral text prompt  $d^{s-n}$  to generate paired natural motion  $m_{0(t)}^{s-n}$  at step  $t$ :

$$m_{0(t)}^{s-n} = \epsilon_\theta(m_G^s, t, d^{s-n}) \quad \text{where } t : \text{from } G \text{ to } 0 \quad (7)$$

Where  $d^{s-n}$  is obtained by modifying  $d^s$ , i.e., we replace style-related vocabulary in  $d^s$  while preserving the content-relevant descriptions. For instance, in Figure 2(c), we substitute "neutrally" for "in depression" in  $d^s$ . This substitution can be achieved by AMR [LFT<sup>\*</sup>15] or a large language model like ChatGPT 3.5 [Ope24]. Upon obtaining  $m_{0(t)}^{s-n}$ , we can infer the denoised motion at step  $t-1$  by the forward process, thus progressively generating final neutral motion  $m^{s-n}$ , i.e.,  $m_{0(0)}^{s-n}$ .

In addition to prompt revisions, we incorporate the root trajectory of the style example projected onto the horizontal plane as the inpainting features for spatio-temporal alignment. Specifically, we extract the linear velocities of the root joint in  $m^s$  frame-by-frame, denoted as  $O_{hori} = [\mathbf{o}_1^x, \mathbf{o}_1^z, \mathbf{o}_2^x, \mathbf{o}_2^z, \dots, \mathbf{o}_f^x, \mathbf{o}_f^z]$ . After obtaining the denoised motion  $m_{0(t)}^{s-n}$  for each time step  $t$  according to the Equation 7, we substitute the corresponding portion of linear velocity with the velocities from the style example  $m^s$ :

$$m_{0(t)}^{s-n}(O_{hori}) = m^s(O_{hori}) \quad (8)$$

This inpainting acts similarly to DTW [RJ93], allowing the generated motion to be aligned with the style example  $m^s$ . Eventually, we take the final denoised motion  $m_{0(0)}^{s-n}$  as the generated neutral motion  $m^{s-n}$ . In summary, the initial noise  $m_G^s$  sampled by  $m^s$  and the root trajectory inpainting ensure that the generated motion  $m^{s-n}$  is consistent with the content of  $m^s$ , while the neutral text prompt ensures that the generated motion is stylistically neutral.

After obtaining the style-neutral motion pair  $\langle m^s, m^{s-n} \rangle$ , we can encourage the reverse process of diffusion model  $\epsilon_{\theta_s}$  to complete the task of style transfer by simply reconstructing the style example. Specifically, we first add K-step noise to  $m^{s-n}$  through the forward process, resulting in the noised motion  $m_K^{s-n}$ . Then, we perform the reverse process to clean the motion:

$$\hat{m}_{0(t)}^s = \epsilon_{\theta_s}(m_K^{s-n}, t, d^s) \quad \text{where } t : \text{from } K \text{ to } 0 \quad (9)$$

Rather than predicting the motion  $\hat{m}_{0(t)}^s$  to be identity to the  $m^{s,n}$ , we encourage it to be close to  $m^s$  for style transfer:

$$\mathcal{L}_{sr} = \|m^s - \hat{m}_{0(t)}^s\|_2^2 \quad t : from K to 0 \quad (10)$$

**Semantic-guided learning.** As illustrated in Figure 2(d), we resort to diverse neutral motions in the text-to-motion database  $\langle \mathcal{M}^c, \mathcal{D}^c \rangle$  and use semantic-guided style transfer loss  $\mathcal{L}_s$  to simultaneously fine-tune the network  $\epsilon_{\theta_s}$  along with  $\mathcal{L}_{sr}$ . Specifically, we add K-step noise to content motions  $\mathcal{M}^c$  through the forward process, resulting in  $\mathcal{M}_K^c$ . Meanwhile, we modify the set of neutral text prompt  $\mathcal{D}^c$  to  $\mathcal{D}^{c,s}$  with style vocabulary of  $m^s$ . Obtaining the set of stylized text prompts  $\mathcal{D}^{c,s}$  is similar to getting  $d^{s,n}$  in Equation 7. For example, in Figure 2(d), "A person jumps neutrally." is changed to "A person jumps in depression."

We then feed the noised motions  $\mathcal{M}_K^c$  into the network  $\epsilon_{\theta_s}$  for denoising and style transfer. To make the output transferred motions  $\mathcal{M}^{c,s}$  maintain the content of the original motions while having the style corresponding to  $m^s$ , we adopt the pre-trained motion-semantic discriminator  $Dis$  in 3.3 to guide the fine-tuning. Specifically, the transferred motion  $m^{c,s} \in \mathcal{M}^{c,s}$  and the stylized text prompt  $d^{c,s} \in \mathcal{D}^{c,s}$  are fed into  $Dis$  and the CLIP text encoder  $E_{text}$ , respectively. The semantic-guided loss  $\mathcal{L}_s$  is computed by the cosine similarity of the motion and text features:

$$\mathcal{L}_s = \mathbb{E}_{m^{c,s}, d^{c,s}} [1 - \frac{Dis(m^{c,s}) \cdot E_{text}(d^{c,s})}{\max(\|Dis(m^{c,s})\| \times \|E_{text}(d^{c,s})\|, \delta)}] \quad (11)$$

Finally, the style transfer fine-tuning loss is defined as a compound of  $\mathcal{L}_{sr}$  and  $\mathcal{L}_s$ :

$$\mathcal{L}_{total} = \lambda_{sr} \mathcal{L}_{sr} + \lambda_s \mathcal{L}_s \quad (12)$$

### 3.5. Style Transfer at Inference

Given the fine-tuned style transfer diffusion model  $\epsilon_{\theta_s}$ , we can apply the style of  $m^s$  to diverse content motions  $m^c$ . If the source motion  $m^c$  is not stylistically neutral, we can first use the generative prior  $\epsilon_\theta$  to generate the corresponding neutral motion(similar to the style-neutral generation in Sec 3.4) as an intermediate state. Then, the style transfer can be easily achieved by forward process ( $m^c \rightarrow m_K^c$ ) and reverse process ( $m_K^c \rightarrow m^{c,s}$ ).

**Global Velocity.** Global velocity is also a property of motion style, for example, the velocity should decrease when transferring the neutral walking to old walking. However, the translation of global velocity is a challenging task, especially when the root position of the style example is almost static (e.g. given old kicking as the style example). In fact, during the finetuning for the style transfer in Sec 3.4, we learn the translation between temporal-aligned motions. Therefore, to reflect the style differences in the global velocity, we follow the [AWL\*20] and perform a time warping on the global velocity based on the following velocity vector:

$$\mathcal{U} = \frac{1}{f} \sum_{\tau}^f \max_{j \in J} u^j(\tau) \quad (13)$$

Where the  $u^j(\tau)$  is the magnitude of the local velocity of the  $j$ -th joint in time  $\tau$ . The  $\mathcal{U}$  measures the temporal average of the

maximal local joint velocity. We then use the factor  $\mathcal{U}^s / \mathcal{U}^c$  to warp the global velocity of the root joint, where the  $\mathcal{U}^s$  and  $\mathcal{U}^c$  represent the velocity factors for the style example  $m^s$  and content motion  $m^c$ , respectively.

## 4. Implementation Details

We use the Pytorch platform and a single 3090Ti graph card for the pre-training and fine-tuning with a batch size of 64. We employ the AdamW [LH19] optimizer with  $\beta_1 = 0.9, \beta_2 = 0.999$ . The learning rate is set to 0.0001 for all models. When fine-tuning, the  $\lambda_{sr}$  and  $\lambda_s$  are set to 1 and 0.1, respectively. The generative prior  $\epsilon_\theta$  and fine-tuned style transfer model  $\epsilon_{\theta_s}$  have the same Transformer encoder architecture which consists of 8 Transformer layers with 512 latent sizes and 1024 feedforward dimensions. Motion inputs to  $\epsilon_\theta$ ,  $\epsilon_{\theta_s}$ , or the motion-semantic discriminator  $Dis$  need to first pass through a shared embedding layer to encode the raw motion feature  $m$  into 512-dimensional deep features. In contrast, the output layer in  $\epsilon_\theta$  and  $\epsilon_{\theta_s}$  is the inverse of input embedding, which decodes the 512-dimensional features into the raw motion representation. The layer number  $N$  of the motion-semantic discriminator  $Dis$  is set to 8. When generating the style-neutral motion pairs in Sec 3.4, we experimentally find the noisy step  $G = 950, K = 300$  to be an appropriate choice (discuss in Sec 6.1).

At each optimization step in fine-tuning, we need to denoise in a loop to get the stylized motion (see Equation 9). To accelerate the procedure and make the style transfer stable, we use deterministic DDIM like [KKY22] during the forward and reverse processes of  $\epsilon_{\theta_s}$ . Specifically, We perform the total sample steps S=20 in DDIM compared to the T=1000 steps in DDPM. Thus,  $K = 300$  corresponds to 6 steps in the actual accelerated process. However, we still use DDPM when generating the style-neutral motion pair. We fine-tune the model using neutral motion data from the text-motion training dataset as well as style-neutral motion pair  $\langle m^s, m^{s,n} \rangle$  for one epoch. For more information about pre-training and fine-tuning time efficiency, please refer to the Appendix.

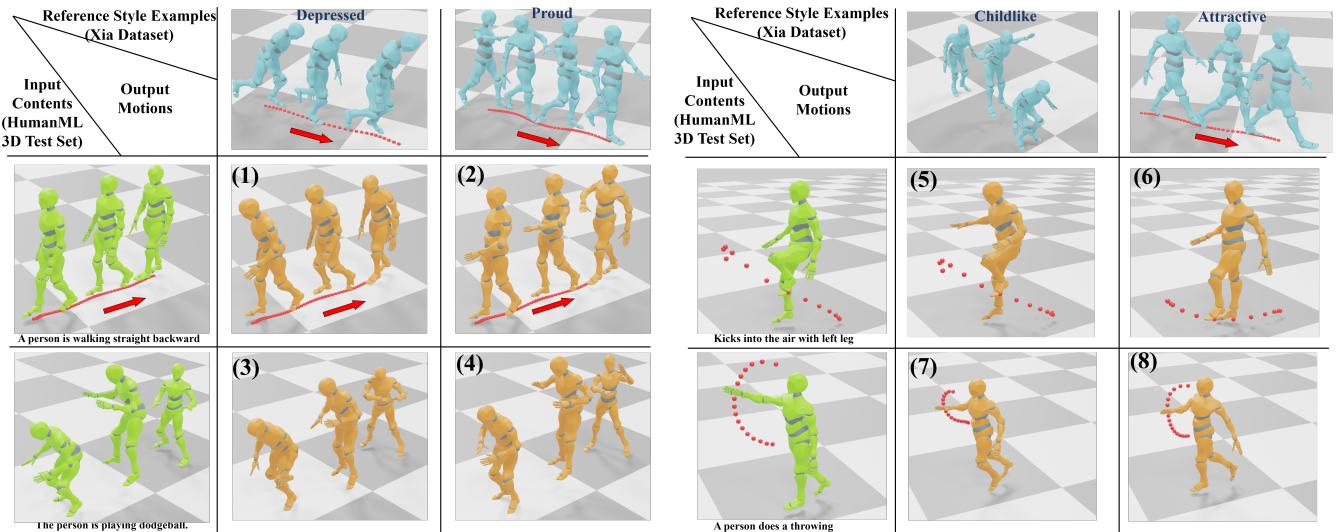
## 5. Experiments and Evaluation

In Sec 5.1, we first qualitatively assess the ability of our method to transfer unseen style to diverse motion contents given only a segment of style motion example. Next, in Sec 5.2, we quantitatively and qualitatively compare our method with state-of-the-art approaches on the dedicated style motion datasets (with content and style annotation).

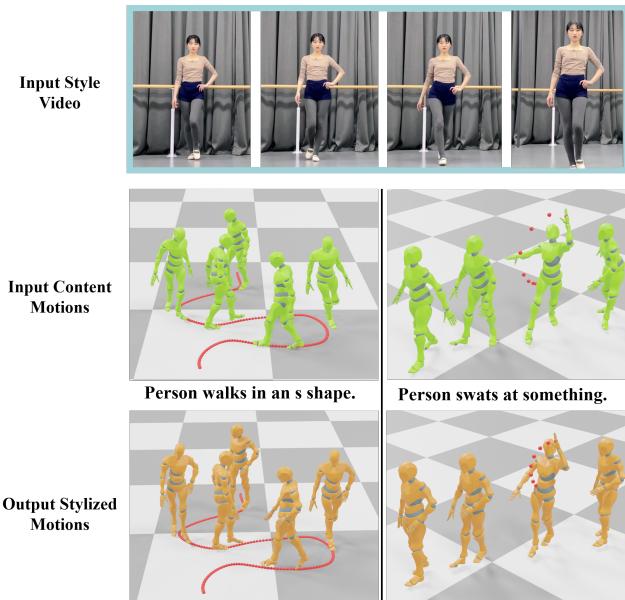
### 5.1. Transfer Unseen Style to Diverse Content Motions

In this part, we design experiments to demonstrate that our two-stage framework can acquire the generative ability from large-scale text-motion datasets and then be transformed into a style transfer model ( $\epsilon_\theta \rightarrow \epsilon_{\theta_s}$ ) by fine-tuning with a single style example. The datasets we used are as follows:

**HumanML3D** [GZZ\*22] is a stylistically unbalanced (predominantly neutral), text-annotated large 3D human motion dataset. It encompasses a wide range of human contents extracted from the AMASS [MGT\*19], including daily activities (e.g. walking and



**Figure 4:** Qualitative results of our few-shot style transfer. We can transfer unseen motion styles (from Xia dataset) to different content motions (from HumanML3D test set).



**Figure 5:** Qualitative results. We can learn the human motion style from video and transfer it to different content motions.

jumping), sports, and artistic pursuits such as dancing. We follow the official train-test split.

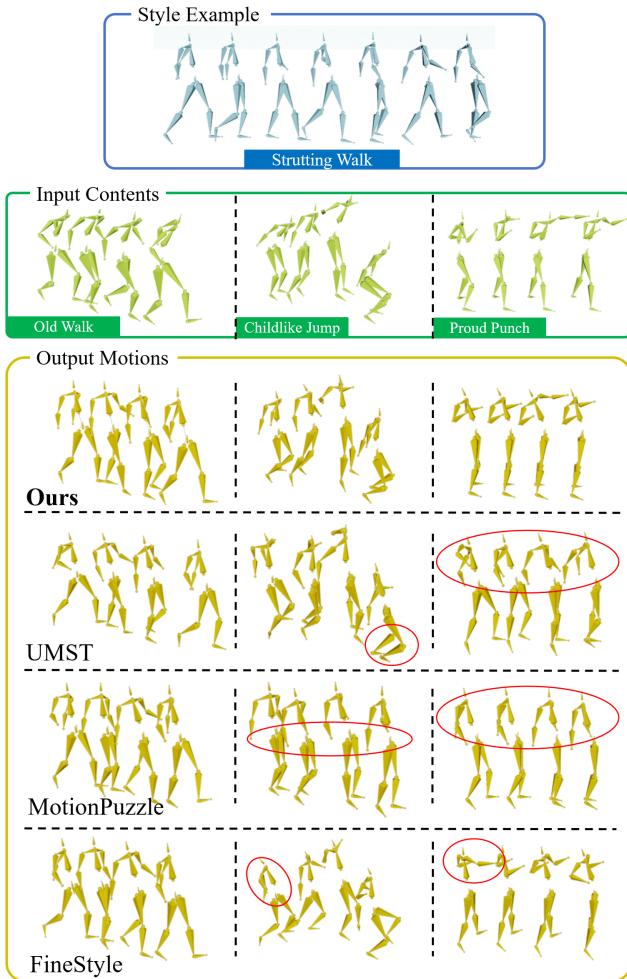
**Xia dataset** [XWCH15] is a dedicated style motion dataset (with content/style labels), which has been widely used in previous motion style transfer work [AWL\*20, PJL21, JPL22, SJL\*23]. It comprises a total of 79,829 frames of motion data and consists of 6 motion contents (walk, run, jump, etc.) and 8 styles (angry, childlike, depressed, etc.). We down-sample the framerate of the Xia

dataset to be consistent with HumanML3D and retarget the motion to the SMPL [LMR\*23] skeleton structure. Since this dataset does not contain textual annotations, we simply annotate the motion with text based on the content and style labels. For example, an annotation might read 'a person is running in depression'.

In stage I, we pre-train the text-to-motion generative prior  $\epsilon_0$  on the HumanML3D train set. In stage II, we sample an unseen style example  $m^s$  from the Xia dataset for style fine-tuning ( $\epsilon_0 \rightarrow \epsilon_{0s}$ ). At inference, we input content motions from the HumanML3D test set to be converted to the style corresponding to  $m^s$ .

It is difficult to quantitatively evaluate the style transfer performance due to the domain differences between the two datasets. Therefore, we qualitatively evaluate our method as shown in Figure 4. The results (1)(2) show that different motion styles (depressed, proud) can be transferred to the same backward motion, despite the style example containing only forward movements. The transferred motions (3)(4) demonstrate that our method can well preserve the action of the content motions while expressing the appropriate stylistic characteristics of the style examples. The results (5)~(8) demonstrate the ability of our method to handle the style transfer between heterogeneous motions, where the given style examples are locomotion and the content motions are non-locomotion. Notably, the stylized output motions exhibit distinct end-effector trajectories compared to the content motions, with the trajectory morphology and body pose collectively conveying the corresponding style.

**Transfer style from video.** Given our framework's capability to fine-tune using just one style example, we are prompted to leverage a diverse array of video resources containing stylized human movements as style exemplars. Specifically, we first use HybrIK [LXC\*21] to extract the human poses in the video, and then apply a Butterworth low-pass filter to remove the jitter from the estimated motion. The cut-off frequency of the filter is set to 3,



**Figure 6:** Qualitative comparison with state-of-the-art methods on Xia test set.

and we exclude the 3D global translation of the estimated motion during this process. Then, we proceed with the fine-tuning stage pipeline to derive a transfer model  $\epsilon_{\theta_s}$  associated with the video. Subsequently, the style transfer diffusion model  $\epsilon_{\theta_s}$  can apply the human style of the video to diverse content motions. As depicted in Figure 5, we observe that the walking style from the video can be applied to other walking trajectories, such as the S-shape. Furthermore, although the akimbo is considered part of the motion style in the video, the semantics of the content motion are effectively preserved when waving the left hand. For more qualitative results, please refer to our supplementary video.

## 5.2. Comparisons with SOTA

In this section, we quantitatively and qualitatively compare our style transfer method with UMST [AWL\*20], MotionPuzzle [JPL22], and FineStyle [SJL\*23]. The implementation of these methods follows their published code.

**Quantitative evaluation.** To quantitatively compare our style

transfer framework with state-of-the-art methods, we conducted experiments on dedicated style motion datasets: Xia and Bandai-2 [KLI\*23], respectively. The Bandai-2 dataset is similar in composition to Xia but has a larger scale, containing 10 content and 7 styles totaling 384,931 frames of motion. For the Xia dataset, we follow [SJL\*23] to split the train-test set. For the Bandai-2, we uniformly select test samples from each content and style, resulting in a total of 3780 transfer combinations.

Since our approach is two-stage, for a fair comparison, we pre-train and fine-tune the models ( $\epsilon_\theta, \epsilon_\theta \rightarrow \epsilon_{\theta_s}$ ) under the same data conditions (Xia or Bandai-2 training set) as the baselines. Meanwhile, the baselines will also fine-tune the networks with the input style example for the same number of steps before testing. We follow [JPL22] and use three metrics: Content Recognition Accuracy (CRA), Style Recognition Accuracy (SRA), and Fréchet Motion Distance (FMD) to assess the content preservation, style accuracy, and quality of the transferred human motions, respectively.

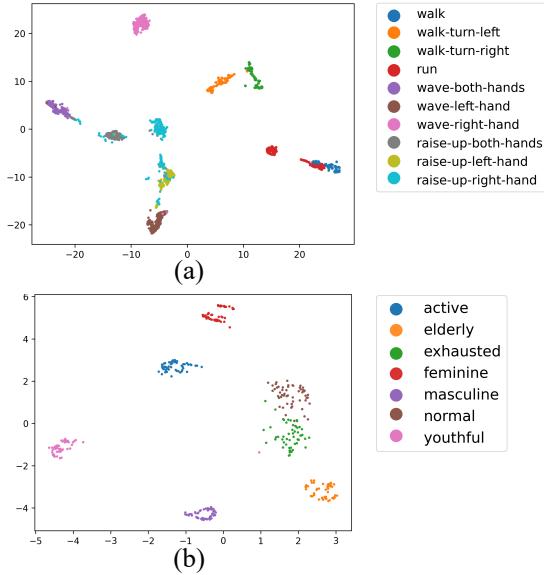
To compute the CRA, we pre-train a content classifier [YXL18] to categorize the generated motions and examine whether they match the label of the source content motions. Similarly, we pre-train another style classifier to categorize the style of the generated motions. The FMD, a derivative of the Fréchet Inception Distance (FID) [HRU\*17], can evaluate the resemblance between generated and real motions by comparing their deep feature distributions. In accordance with [JPL22], we employ the above content classifier to extract the latent features from the classifier’s final pooling layer and calculate the FMD score.

The quantitative results are presented in Table 1. Our method achieves state-of-the-art performance on both the Xia and Bandai-2 test sets. From the table, our CRA significantly outperforms other methods, achieving 83.72% and 93.41% accuracy on the Xia and Bandai-2 datasets, respectively. Regarding motion quality, our method not only yields the lowest FMD scores but also remains unaffected by the dataset scale (3.84 and 2.33 on the Xia and Bandai-2 datasets, respectively). Furthermore, our approach achieves the highest SRA on both datasets, demonstrating a robust balance between content preservation and style transfer. The high performance of our method may be attributed to the explicit definition of style as a particular variation of motion and adding it through the denoising process, which is more interpretive and stable. Also, the style transfer guided by text prompts has more detailed information compared to style labels.

**Qualitative evaluation.** To evaluate the performance of style transfer on heterogeneous motions, we showcase the qualitative results on the Xia test set in Figure 6. The input style example is a clip of “strutting walk”, while the content motions encompass “old walk”, “childlike jump” and “proud punch”. In the first column, where the content motion shares the same action type (i.e., walking) as the style example, all methods successfully execute style transfer from “old walk” to “strutting walk”. In the second column, MotionPuzzle fails to persevere the source content, yielding a drifting motion, while both UMST and FineStyle generate motions with unnatural poses. These samples show that our approach achieves robust style transfer and content preservation even among challenging heterogeneous motions. For more qualitative results, please refer to the Appendix and supplementary video.

**Table 1:** Quantitative comparison with state-of-the-art methods. FMD is short for Fréchet Motion Distance. CRA and SRA stand for Content Recognition Accuracy and Style Recognition Accuracy, respectively. The down arrow indicates that smaller values are better and vice versa.

Methods	Xia [XWCH15]			Bandai-2 [KLI*23]		
	FMD↓	CRA↑ (%)	SRA↑ (%)	FMD ↓	CRA↑(%)	SRA↑(%)
UMST [AWL*20]	5.63	62.88	38.97	23.38	53.78	54.23
MotionPuzzle [JPL22]	22.79	40.19	22.56	38.63	42.67	25.79
FineStyle [SJL*23]	5.34	63.91	55.58	24.84	58.89	56.01
Ours	<b>3.84</b>	<b>83.72</b>	<b>70.77</b>	<b>2.33</b>	<b>93.41</b>	<b>60.11</b>

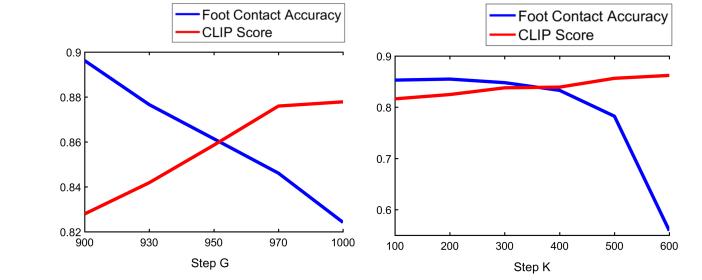


**Figure 7:** Hidden feature visualization via t-SNE for our transferred motions based on (a) content labels and (b) style labels.

### 5.3. Visualization by t-SNE

To intuitively visualize the comprehensive outcomes of our style transfer, we employ t-distributed stochastic neighbor embedding (t-SNE) [VdMH08] to illustrate the clustering efficacy of the transferred motions. Specifically, we project the hidden features, extracted by the final pooling layer of content and style classifiers [YXL18], into a two-dimensional space. This experimental analysis is conducted on the Bandai-2 test set. As depicted in Figure 7(a), the hidden features of our stylized motions exhibit discriminative cluster boundaries based on the content labels. Furthermore, these clusters demonstrate the proximity relationships between motion contents, with similar motions grouped closely together.

Recognizing that identical styles may manifest differently across heterogeneous motion contents, we project the stylized motions with the “raise-up-both-hands” content into a two-dimensional space and color them according to the style labels. As shown in Figure 7(b), the clustering reveals that the output stylized motions exhibit clear style distinctions and can be effectively differentiated by the classifier. In summary, the clustering results indicate that our



**Figure 8:** Ablation study on noise step G and K

stylized motions possess adequate discriminative features, thereby further validating the effectiveness of our approach.

## 6. Ablation Study

We conducted the ablation study, with a primary focus on three key aspects: the selection of noise steps, the reliability of neutral motion generation, and the key components of the framework.

### 6.1. Selection of Noise Steps $G$ and $K$

We seek to find the most appropriate noise step  $G$  during neutral motion generation and the noise step  $K$  during style transfer.

According to Equation 2, a smaller noise step  $G$  during the forward process preserves more detail in the style example  $m^s$ , while a larger  $G$  will make the noised motion  $m_G^s$  closer to the standard normal distribution, thus encouraging the motion generated by Equation 7 to be guided by a neutral textual prompt. Therefore, we wish to find a proper  $G$  such that the generated motion retains the content of  $m^s$  while being stylistically neutral. We use the Foot Contact Accuracy [VCH\*21] to measure the degree of alignment between the  $\langle m^s, m^{s-n} \rangle$ , where  $m^{s-n}$  is the neutral motion corresponding to  $m^s$  (defined in Sec 3.4). On the other hand, we measure neutrality by comparing the similarity of the generated neutral motions  $m^{s-n}$  with the neutral textual descriptions via the motion-semantic discriminator  $Dis$  in the CLIP space. Similarly, we select  $K$  with the same metrics to find the balance between content preservation and style transfer.

The evaluations are conducted on the Bandai-2 test set and the results are presented in Figure 8. In the left subplot, we can see that the metrics reach a balance around  $G=950$ . Similarly, from the right subplot, it is observed that as the number of  $K$  increases,

**Table 2:** Ablation study of style-neutral motion pair generation on Bandai-2 test set.

Methods	Ablation on Paired Neutral motion Generation		
	FMD↓	CRA↑ (%)	SRA↑ (%)
Ours	<b>5.87</b>	98.06	88.61
KNN+DTW	21.76	<b>100</b>	<b>97.78</b>

**Table 3:** Ablation study of the key components of the framework on Xia test set.

Methods	Ablation on Style Transfer		
	FMD↓	CRA↑ (%)	SRA↑ (%)
Ours	<b>3.84</b>	83.72	<b>70.77</b>
with KNN+DTW	5.48	76.86	69.49
with AE	5.84	19.68	13.27
without $\mathcal{L}_s$	4.00	<b>88.01</b>	28.01

the CLIP score increases accordingly, indicating that the style of the generated motion is getting closer to the input style examples. However, after 400 steps, the Foot Contact Accuracy of the generated stylized motion starts to degrade significantly. Therefore, we chose  $K = 300$  as the number of noising/denoising steps in the forward/reverse process during the style transfer diffusion.

## 6.2. The Reliability of Style-Neutral Motion Pairs

To demonstrate that the paired neutral motions generated through our prior model  $\epsilon_\theta$  are reliable, we compare our generated results with the neutral motions produced by KNN+DTW. Specifically, given an style example  $m^s$ , we generate  $m^{s,n}$  using the method described in Sec 3.4, whereas KNN+DTW first searches for a segment of neutral motion (has the same content label as  $m^s$ ) from the dataset that is closest to  $m^s$ , and then aligns it to  $m^s$  via DTW. The comparison results are illustrated in Table 2. Since KNN+DTW searches directly from the dataset, the obtained motions can be easily recognized by the classifier, leading to high CRA as well as SRA. However, the use of DTW alignment compresses or stretches the neutral motions along the temporal axis, leading to an increase in FMD. Our method, on the other hand, is able to achieve a relatively balanced and competitive performance.

In addition to directly evaluating the generated neutral pairs, we also test the effect on the final style transfer, the results are shown in Table 3. our method is able to achieve better style transfer performance with a marginal margin compared to KNN+DTW. For more qualitative results, please refer to our Appendix and supplementary video.

## 6.3. Key Components of the Framework

**Effect of diffusion model.** To validate the effectiveness of the diffusion model, we replace it with an autoencoder(AE) architecture in both the pre-training and fine-tuning stages. Specifically, we adopt the AE architecture in UMST, while omitting the Instance Normalization and AdaIN layers. From Table 3, we observe the performance of our model with AE architecture decreases in all three metrics. This issue may arise from the AE model being

trained to over-fit the motion reconstruction during stage I, making it difficult to achieve effective style transfer through a few-shot fine-tuning.

**Effect of  $\mathcal{L}_s$ .** The semantic-guided style transfer learning loss  $\mathcal{L}_s$  enables us to supervise the fine-tuning in CLIP semantic space. To validate the effectiveness, we remove the loss  $\mathcal{L}_s$  during the fine-tuning stage. It is observed the CRA in Table 3 of the model increases from 83.72% to 88.01% and the SRA decreases dramatically from 70.77% to 28.01%. This is because, without  $\mathcal{L}_s$ , the model's ability to transfer style can only be learned by examining differences in style-neutral motion pair  $\langle m^s, m^{s,n} \rangle$ , and thus is difficult to generalize to heterogeneous motions. Indeed, the  $\mathcal{L}_s$  fully exploits the motion style prior in the text-to-motion dataset, which allows the model  $\epsilon_{\theta_s}$  to learn style transfer both from the example  $m^s$  and from the motions with the same style type as  $m^s$ .

## 7. Discussion

In this paper, we present a novel two-stage framework for human motion style transfer. In stage I, we pre-train a diffusion-based T2M model  $\epsilon_\theta$  as generative prior, which allows us to adapt to heterogeneous motions. We realize that noise in the reverse process can bring about variations in the output motion, and human motion styles, as a particular variation, can also be generated in the reverse process. Therefore, in stage II, we fine-tune the diffusion model based on the input style example, making the reverse process of  $\epsilon_{\theta_s}$  has the capability of style transfer. This two-stage framework allows us to leverage large, stylistically unbalanced text-motion datasets to train a robust generator, while in the fine-tuning for style transfer, we only need a segment of the style example. Extensive experiments demonstrate that our method achieves state-of-the-art performance and has practical applications.

Our system has two main limitations. First, our style transfer model takes the whole motion segment as the input rather than the sliding window, which makes it difficult to combine with real-time motion control methods like PFNN [HKS17]. In the future, improving the denoising efficiency and combining the motion phase [SMK22] may be a solution to achieve diffusion-based real-time style transfer.

Second, we fine-tune an independent style transfer model for each different style example, which may result in model storage stacking as well as difficulties in style interpolation. In future work, we can introduce new conditions (e.g. the hidden features of the style example) in the reverse process of style diffusion to integrate knowledge from different style transfer models. We can also consider introducing the DiffusionBlending technique in PriorMDM [STKB24] for style interpolation and extrapolation.

## Acknowledgment

This work was supported by National Key R&D Program of China (NO. 2022YFB3303202).

## References

- [APBV22] ATHANASIOU N., PETROVICH M., BLACK M. J., VAROL G.: Teach: Temporal action composition for 3d humans. In 2022 International Conference on 3D Vision (3DV) (2022), IEEE, pp. 414–423.

- [AWL\*20] ABERMAN K., WENG Y., LISCHINSKI D., COHEN-OR D., CHEN B.: Unpaired motion style transfer from video to animation. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 64–1. 1, 3, 6, 8, 9
- [AZL23] AO T., ZHANG Z., LIU L.: Gesturediffuclip: Gesture diffusion model with clip latents. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–18. 3
- [BH00] BRAND M., HERTZMANN A.: Style machines. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques* (2000), pp. 183–192. 2
- [BMR\*20] BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SAstry G., ASKELL A., ET AL.: Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901. 3
- [CWD\*18] CRESWELL A., WHITE T., DUMOULIN V., ARULKUMARAN K., SENGUPTA B., BHARATH A. A.: Generative adversarial networks: An overview. *IEEE signal processing magazine* 35, 1 (2018), 53–65. 5
- [DHS\*19] DU H., HERRMANN E., SPRENGER J., FISCHER K., SLUSALLEK P.: Stylistic locomotion modeling and synthesis using variational generative models. In *Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games* (2019), pp. 1–10. 2, 4
- [GEB16] GATYS L. A., ECKER A. S., BETHGE M.: Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 2414–2423. 2
- [GZZ\*22] GUO C., ZOU S., ZUO X., WANG S., JI W., LI X., CHENG L.: Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2022), pp. 5152–5161. 3, 6
- [HB17] HUANG X., BELONGIE S.: Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 1501–1510. 3
- [HHKK17] HOLDEN D., HABIBIE I., KUSAJIMA I., KOMURA T.: Fast neural style transfer for motion data. *IEEE computer graphics and applications* 37, 4 (2017), 42–49. 2
- [HJA20] HO J., JAIN A., ABBEEL P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851. 3, 4
- [HKS17] HOLDEN D., KOMURA T., SAITO J.: Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–13. 2, 3, 10
- [HPP05] HSU E., PULLI K., POPOVIĆ J.: Style translation for human motion. In *ACM SIGGRAPH 2005 Papers*. 2005, pp. 1082–1089. 1, 2, 5
- [HRU\*17] HEUSEL M., RAMSAUER H., UNTERTHINER T., NESSLER B., HOCHREITER S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017). 8
- [HSK16] HOLDEN D., SAITO J., KOMURA T.: A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–11. 2, 4
- [HZZ\*23] HU L., ZHANG Z., ZHONG C., JIANG B., XIA S.: Pose-aware attention network for flexible motion retargeting by body part. *IEEE Transactions on Visualization & Computer Graphics*, 01 (2023), 1–17. 5
- [IAF09] IKEMOTO L., ARIKAN O., FORSYTH D.: Generalizing motion edits with gaussian processes. *ACM Transactions on Graphics (TOG)* 28, 1 (2009), 1–12. 2
- [JCL\*23] JIANG B., CHEN X., LIU W., YU J., YU G., CHEN T.: MotionGPT: Human motion as a foreign language. In *Thirty-seventh Conference on Neural Information Processing Systems* (2023). URL: <https://openreview.net/forum?id=WqizJGNkjn>. 3
- [JPL22] JANG D.-K., PARK S., LEE S.-H.: Motion puzzle: Arbitrary motion style transfer by body part. *ACM Transactions on Graphics (TOG)* 41, 3 (2022), 1–16. 1, 3, 6, 8, 9
- [KKY22] KIM G., KWON T., YE J. C.: Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 2426–2435. 6
- [KLI\*23] KOBAYASHI M., LIAO C.-C., INOUE K., YOJIMA S., TAKAHASHI M.: Motion capture dataset for practical use of ai-based motion editing and stylization, 2023. [arXiv:2306.08861](https://arxiv.org/abs/2306.08861). 8, 9
- [LFT\*15] LIU F., FLANIGAN J., THOMSON S., SADEH N., SMITH N. A.: Toward abstractive summarization using semantic representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Denver, Colorado, May–June 2015), Mihalcea R., Chai J., Sarkar A., (Eds.), Association for Computational Linguistics, pp. 1077–1086. doi:[10.3115/v1/N15-1114](https://doi.org/10.3115/v1/N15-1114). 5
- [LH19] LOSHCHILOV I., HUTTER F.: Decoupled weight decay regularization. In *International Conference on Learning Representations* (2019). URL: <https://openreview.net/forum?id=Bkg6RiCqY7>. 6
- [LMR\*23] LOPER M., MAHMOOD N., ROMERO J., PONS-MOLL G., BLACK M. J.: Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 2023, pp. 851–866. 7
- [LXC\*21] LI J., XU C., CHEN Z., BIAN S., YANG L., LU C.: Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 3383–3393. 7
- [LZCVDP20] LING H. Y., ZINNO F., CHENG G., VAN DE PANNE M.: Character controllers using motion vaes. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 40–1. 3
- [MGT\*19] MAHMOOD N., GHORBANI N., TROJE N. F., PONS-MOLL G., BLACK M. J.: AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision* (Oct. 2019), pp. 5442–5451. 6
- [MSK22] MASON I., STARKE S., KOMURA T.: Real-time style modelling of human locomotion via feature-wise transformations and local motion phases. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 5, 1 (2022), 1–18. 2
- [MSZ\*18] MASON I., STARKE S., ZHANG H., BILEN H., KOMURA T.: Few-shot learning of homogeneous human locomotion styles. In *Computer Graphics Forum* (2018), vol. 37, Wiley Online Library, pp. 143–153. 2
- [MXH\*10] MA W., XIA S., HODGINS J. K., YANG X., LI C., WANG Z.: Modeling style and variation in human motion. In *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (2010), pp. 21–30. 1, 2
- [Ope24] OPENAI.: Chatgpt (3.5) [large language model]. <https://https://chat.openai.com>, 2024. Accessed: 2024-04-07. 5
- [PBV21] PETROVICH M., BLACK M. J., VAROL G.: Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 10985–10995. 3, 5
- [PJJL21] PARK S., JANG D.-K., LEE S.-H.: Diverse motion stylization for multiple style domains via spatial-temporal graph-based generative model. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 4, 3 (2021), 1–17. 3, 6
- [RDN\*22] RAMESH A., DHARIWAL P., NICHOL A., CHU C., CHEN M.: Hierarchical text-conditional image generation with clip latents. arxiv 2022. [arXiv preprint arXiv:2204.06125](https://arxiv.org/abs/2204.06125) (2022). 4
- [RJ93] RABINER L., JUANG B.-H.: "Chapter 4: Pattern-Comparison Techniques". *Fundamentals of speech recognition*. Prentice-Hall, Inc., 1993. ISBN 978-0-13-015157-5. 1, 5
- [RKH\*21] RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SAstry G., ASKELL A., MISHKIN P., CLARK J., ET AL.: Learning transferable visual models from natural language

- supervision. In *International conference on machine learning* (2021), PMLR, pp. 8748–8763. 2
- [RLT\*24] RAAB S., LEIBOVITCH I., TEVET G., ARAR M., BERMANO A. H., COHEN-OR D.: Single motion diffusion. In *The Twelfth International Conference on Learning Representations (ICLR)* (2024). 3
- [S JL\*23] SONG W., JIN X., LI S., CHEN C., HAO A., HOU X.: Finestyle: Semantic-aware fine-grained motion style transfer with dual interactive-flow fusion. *IEEE Transactions on Visualization and Computer Graphics* (2023). 3, 6, 8, 9
- [SME21] SONG J., MENG C., ERMON S.: Denoising diffusion implicit models. In *International Conference on Learning Representations* (2021). URL: <https://openreview.net/forum?id=St1giarCHLP>. 3
- [SMK22] STARKE S., MASON I., KOMURA T.: Deepphase: Periodic autoencoders for learning motion phase manifolds. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–13. 10
- [SSME23] SU X., SONG J., MENG C., ERMON S.: Dual diffusion implicit bridges for image-to-image translation. In *International Conference on Learning Representations* (2023). 3
- [STKB24] SHAFIR Y., TEVET G., KAPON R., BERMANO A. H.: Human motion diffusion as a generative prior. In *The Twelfth International Conference on Learning Representations* (2024). URL: <https://openreview.net/forum?id=dTpBEdN9kr>. 3, 10
- [SZKZ20] STARKE S., ZHAO Y., KOMURA T., ZAMAN K.: Local motion phases for learning multi-contact character movements. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 54–1. 3
- [TCL23] TSENG J., CASTELLON R., LIU K.: Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 448–458. 3
- [TRG\*23] TEVET G., RAAB S., GORDON B., SHAFIR Y., COHEN-OR D., BERMANO A. H.: Human motion diffusion model. In *The Eleventh International Conference on Learning Representations* (2023). URL: <https://openreview.net/forum?id=SJ1kSy02jwu>. 3, 4
- [VCH\*21] VILLEGAS R., CEYLAN D., HERTZMANN A., YANG J., SAITO J.: Contact-aware retargeting of skinned motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 9720–9729. 9
- [VdMH08] VAN DER MAATEN L., HINTON G.: Visualizing data using t-sne. *Journal of machine learning research* 9, 11 (2008). 9
- [VDOV\*17] VAN DEN OORD A., VINYALS O., ET AL.: Neural discrete representation learning. *Advances in neural information processing systems* 30 (2017). 3
- [WCX19] WANG Z., CHAI J., XIA S.: Combining recurrent neural networks and adversarial training for human motion synthesis and control. *IEEE transactions on visualization and computer graphics* 27, 1 (2019), 14–28. 3
- [WDL\*21] WANG H., DU D., LI J., JI W., YU L.: A cyclic consistency motion style transfer method combined with kinematic constraints. *Journal of Sensors* 2021 (2021), 1–17. 2
- [WFH07] WANG J. M., FLEET D. J., HERTZMANN A.: Multifactor gaussian process models for style-content separation. In *Proceedings of the 24th international conference on Machine learning* (2007), pp. 975–982. 2
- [WWY22] WINKLER A., WON J., YE Y.: Questsim: Human motion tracking from sparse sensors with simulated avatars. In *SIGGRAPH Asia 2022 Conference Papers* (2022), pp. 1–8. 3
- [WYF\*21] WEN Y.-H., YANG Z., FU H., GAO L., SUN Y., LIU Y.-J.: Autoregressive stylized motion synthesis with generative flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 13612–13621. 2
- [WZX23] WANG Z., ZHAO L., XING W.: Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 7677–7689. 3
- [XWCH15] XIA S., WANG C., CHAI J., HODGINS J.: Realtime style transfer for unlabeled heterogeneous human motion. *ACM Transactions on Graphics (TOG)* 34, 4 (2015), 1–10. 1, 2, 5, 6, 9
- [YLHX22] YE Y., LIU L., HU L., XIA S.: Neural3points: Learning to generate physically realistic full-body motion for virtual reality users. In *Computer Graphics Forum* (2022), vol. 41, Wiley Online Library, pp. 183–194. 3
- [YM16] YUMER M. E., MITRA N. J.: Spectral style transfer for human motion between independent actions. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–8. 2
- [YSI\*23] YUAN Y., SONG J., IQBAL U., VAHDAT A., KAUTZ J.: Phys-diff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 16010–16021. 3
- [YXL18] YAN S., XIONG Y., LIN D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence* (2018), vol. 32. 8, 9
- [YY\*24] YIN W., YU Y., YIN H., KRAGIC D., BJORKMAN M.: Scalable motion style transfer with constrained diffusion generation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2024), vol. 38, pp. 10234–10242. 3
- [ZBL\*19] ZHOU Y., BARNES C., LU J., YANG J., LI H.: On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 5745–5753. 3
- [ZCP\*24] ZHANG M., CAI Z., PAN L., HONG F., GUO X., YANG L., LIU Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024). 3
- [ZDIT12] ZHOU F., DE LA TORRE F.: Generalized time warping for multi-modal alignment of human motion. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (2012), IEEE, pp. 1282–1289. 5
- [ZGP\*23] ZHANG M., GUO X., PAN L., CAI Z., HONG F., LI H., YANG L., LIU Z.: Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 364–373. 3
- [ZHGX22] ZHONG C., HU L., ZHANG Z., XIA S.: Learning uncoupled-modulation cvae for 3d action-conditioned human motion synthesis. In *European Conference on Computer Vision* (2022), Springer, pp. 716–732. 3
- [ZHGX23] ZHONG C., HU L., ZHANG Z., XIA S.: Attt2m: Text-driven human motion generation with multi-perspective attention mechanism. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 509–519. 3
- [ZPIE17] ZHU J.-Y., PARK T., ISOLA P., EFROS A. A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 2223–2232. 2
- [ZZC\*23] ZHANG J., ZHANG Y., CUN X., ZHANG Y., ZHAO H., LU H., SHEN X., SHAN Y.: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 14730–14740. 3