Zhao Yang Gaoling School of Artificial Intelligence, Renmin University of China Beijing Key Laboratory of Big Data Management and Analysis Methods Beijing, China yangyz1230@gmail.com Bing Su* Gaoling School of Artificial Intelligence, Renmin University of China Beijing Key Laboratory of Big Data Management and Analysis Methods Beijing, China subingats@gmail.com Ji-Rong Wen Gaoling School of Artificial Intelligence, Renmin University of China Beijing Key Laboratory of Big Data Management and Analysis Methods Beijing, China jrwen@ruc.edu.cn

ABSTRACT

Text-to-motion generation has gained increasing attention, but most existing methods are limited to generating short-term motions that correspond to a single sentence describing a single action. However, when a text stream describes a sequence of continuous motions, the generated motions corresponding to each sentence may not be coherently linked. Existing long-term motion generation methods face two main issues. Firstly, they cannot directly generate coherent motions and require additional operations such as interpolation to process the generated actions. Secondly, they generate subsequent actions in an autoregressive manner without considering the influence of future actions on previous ones. To address these issues, we propose a novel approach that utilizes a past-conditioned diffusion model with two optional coherent sampling methods: Past Inpainting Sampling and Compositional Transition Sampling. Past Inpainting Sampling completes subsequent motions by treating previous motions as conditions, while Compositional Transition Sampling models the distribution of the transition as the composition of two adjacent motions guided by different text prompts. Our experimental results demonstrate that our proposed method is capable of generating compositional and coherent long-term 3D human motions controlled by a user-instructed long text stream. The code is available at https://github.com/yangzhao1230/PCMDM.

CCS CONCEPTS

 \bullet Computing methodologies \rightarrow Activity recognition and understanding.

KEYWORDS

generative models, human motion generation, diffusion models, product of distributions

MM'23, October 29-November 3, 2023, Ottawa, ON, Canada

© 2023 Association for Computing Machinery.

ACM ISBN 979-8-4007-0108-5/23/10...\$15.00

https://doi.org/3581783.3611887

ACM Reference Format:

Zhao Yang, Bing Su, and Ji-Rong Wen. 2023. Synthesizing Long-Term Human Motions with Diffusion Models via Coherent Sampling. In *Proceedings of ACM Conference (MM'23)*. ACM, Ottawa, ON, Canada, 12 pages. https://doi.org/3581783.3611887



Figure 1: Generating Coherent Long-Term 3D Human Motion from Text Streams. Existing autoregressive methods often lack natural transitions, leading to the use of alignment and interpolations between generated motions. In contrast, our proposed method employs diffusion models with coherent sampling methods, enabling the generation of smooth actions without the need for additional post-processing.

1 INTRODUCTION

Synthesizing complex and realistic 3D human motions is of great significance in virtual reality, game production, film industry, humanmachine interaction, etc. Motion capture techniques [42] have been

^{*}Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

widely applied to obtain motion data, but they are expensive and time-consuming because sophisticated equipment and professional actors are required to perform actions. It is desirable to develop motion generation models that can automatically generate demanded human motions as specified.

Early motion generation methods [36, 11, 3] can only generate motions from a fixed set of pre-defined action classes, which cannot meet highly flexible requirements in open real-world applications. A more convenient way to specify the target motion is by describing it in natural language. Text-to-motion generation [1, 10, 37] aims to generate corresponding motions following any given language instructions. With the development of multi-modal pretraining [46, 17] and generative models [45, 14, 22, 9, 23], text-to-motion generation has made significant progress and attracted a lot of attention. However, most existing methods can only take a single sentence describing a single motion as the input prompt and generate an individual short-term motion.

In practice, a motion is usually not executed separately, but a series of motions are executed successively to perform an event or a complex activity. For example, as shown in Fig. 1, the actor follows a stream of language instructions to complete one shot of a movie and the character in a game performs a series of actions to complete a task. When a stream of text prompts is fed into existing text-tomotion generation models, they can only be separately processed to generate a series of individual motions that are not coherent. Simple linear interpolations among motions cannot lead to natural and smooth transitions.

It is not trivial to extend existing text-to-motion generation models to tackle such long-term compositional motion generation for the following reasons. 1. Due to the semantic changes between different sentences, text streams differ significantly from the pretraining text prompts of these models. 2. It is not plausible to collect sufficient compositional motion-text data for pre-training these models since the number of motion compositions grows exponentially. 3. Non-autoregressive generative models can only generate motions with a fixed number of frames, but the number of sentences contained in streams is variable; autoregressive generative models rely on large amounts of compositional training data to model the transitions.

Relatively much fewer works have been devoted to long-term compositional motion generation. In [25], although compositional motions can be generated, the input can only be multiple action labels rather than text prompts, and thus the motions for composition are limited to a pre-defined action set. In [2], TEACH firstly tackles the problem of text-conditioned compositional motion generation, which employs the conditional variational autoencoder (VAE) [46] to successively generate motions by taking the last few frames of the previous motion as conditions. Although TEACH utilizes previous motion as a condition for generation, it still cannot directly generate coherent motions and requires alignment and interpolation between different motions.

Diffusion models have shown strong empirical performances and controllability in image and motion generation because the generation process is divided into multiple steps and each step only needs to fit a simple Gaussian distribution to reverse the forward diffusion. In this paper, we propose a novel approach for generating long-term compositional human motions from text streams using diffusion models. Specifically, we propose a past-conditioned diffusion model and two coherent sampling methods, namely Past Inpainting Sampling and Compositional Transition Sampling, that enable the direct generation of coherent motions without post-processing such as alignment and interpolation.

The contributions of this paper are summarized as follows.

- To the best of our knowledge, we propose the first diffusion model-based method to tackle the problem of generating long-term compositional human motions from text streams. Our human motion diffusion model can be trained from general text-motion data, and no specific aligned text streams and long-term compositional human motions are required for training. To further improve the generation quality, we can also add a past-conditioned module to train on matched long-term text-action pairs.
- We propose two sampling methods based on diffusion models that allow the model to directly generate coherent motions without the need for post-processing such as alignment and interpolation. These methods are Past Inpainting Sampling and Compositional Transition Sampling. Among them, Compositional Transition Sampling abandons the autoregressive generation paradigm, allowing adjacent motions to influence each other during generation and better aligning with the semantics of real long-term motions.
- We conduct both quantitative and qualitative experimental evaluations to demonstrate the superiority of the proposed method over state-of-the-art text-instructed action composition methods.

2 RELATED WORK

We review relevant works in human motion generation and diffusion generative models.

Human motion generation. Unconditional human motion generation [50, 53] models human motion distributions without labels or guidance. Conditional human motion generation is gaining more interest. Some conditions are based on the motion itself, such as motion prediction [7, 33, 13] or motion in-betweening [19, 12] from prefix and suffix poses. Other conditions directly describe desired motion, such as action label conditioned [11, 3, 36, 31], free-form text prompt conditioned [10, 37, 1], and music conditioned [26] motion generation. However, current methods struggle with generating coherent long-term motions composed of multiple actions. Recent work, such as [25] and [2], introduce conditional VAE models to generate coherent compositional motions conditioned on multiple action labels or free-form language guidance. In contrast, we propose a diffusion model-based method that generates long-term compositional human motions from text streams, which does not require specific aligned text streams or long-term compositional human motions for training. We also introduce two sampling methods, Past Inpainting Sampling, and Compositional Transition Sampling, that enable direct generation of coherent motions without post-processing.

Diffusion generative models. Diffusion models [45] add gradually increasing Gaussian noises to data and learn to reverse the perturbation via estimating the added noises. Recently, diffusion models have achieved success in various generative tasks, such as image generation [14, 47], audio generation [24], text generation [27, 8], story generation [35], and molecule generation [49, 18, 16]. Large-scale pretrained diffusion models, such as DALL-E2 [40], Stable Diffusion [41], and Imagen [44], have dominated the field of text-to-image generation. Recently, some works applied diffusion models to 3D human motion generation, such as [48], [20], [52], and [51]. However, these methods cannot generate coherent longterm motions with multiple text prompts, which is our focus. [28, 4] clarified the relationship between diffusion models and energy based models [6, 5], and propose methods of multi-conditioned compositional visual generation. We also consider the transition between adjacent actions as a composition of motions guided by different texts.

3 PRELIMINARY

Inspired by non-equilibrium thermodynamics, diffusion models [45] define a Markov chain of the forward diffusion process q to gradually add Gaussian noises to the real data distribution $\mathbf{x}_0 \sim q(\mathbf{x})$. The step sizes are controlled by a predefined variance schedule $\{\alpha_t \in (0, 1)\}_{t=1}^T$. The forward process $q(\mathbf{x}_t | \mathbf{x}_{t-1})$ at each time step t is:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{\alpha_t} x_{t-1}, (1 - \alpha_t) \mathbf{I}\right). \tag{1}$$

After adding enough noise, the data distribution finally becomes equivalent to an isotropic Gaussian distribution.

DDPM [14] learns a prediction network denoted as ϵ_{θ} to reverse the forward diffusion process by predicting and removing the noise added in the corresponding forward step. Let $\beta_t = 1 - \alpha_t$ and $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$. The denoising process $p(\mathbf{x}_{t-1}|\mathbf{x}_t)$ can also be reparameterized as a Gaussian distribution, which can be estimated by ϵ_{θ} and has a form of the following:

$$p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_{t}) = \mathcal{N}(\mathbf{x}_{t-1};\boldsymbol{\mu}_{\theta}(\mathbf{x}_{t},t),\beta_{t})$$

with
$$\boldsymbol{\mu}_{\theta}(\mathbf{x}_{t},t) = \frac{1}{\sqrt{\alpha_{t}}}(\mathbf{x}_{t} - \frac{\beta_{t}}{\sqrt{1 - \bar{\alpha_{t}}}}\boldsymbol{\epsilon}_{\theta}(\mathbf{x}_{t},t)),$$
(2)

where $\tilde{\beta}_t$ is a function of $\{\beta_t\}_{t=1}^T$.

The learning objective of diffusion models is to approximate the mean $\mu_{\theta}(\mathbf{x}_t, t)$ of the Gaussian distribution with respect to the reverse diffusion process. The mean for sampling \mathbf{x}_{t-1} is estimated as the noisy data \mathbf{x}_t minus the predicted noise $\epsilon_{\theta}(\mathbf{x}_t, t)$ at step t. The variational lower bound (ELBO) [22] is adopted to minimize the negative log-likelihood of $p_{\theta}(\mathbf{x}_0)$ [14], and the simplified objective can be written as a denoising objective:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon} \sim \mathcal{N}(0, 1), t} \left[\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) \|^2 \right].$$
(3)

Once $\epsilon_{\theta}(x, t)$ is trained, we can use Eq. (2) to recover $\mu_{\theta}(x, t)$ and conduct ancestral sampling. Applying the properties of Markov chains, we can derive:

$$\kappa_0 = \frac{x_t - \sqrt{1 - \bar{\alpha}_t}\epsilon}{\sqrt{\bar{\alpha}_t}}.$$
(4)

We can also use Eq. (4) to recover clean data $\hat{\mathbf{x}}_{\theta}$ (\mathbf{x}_{t} , t).

For a conditional generation with diffusion models, we use classifierfree guidance [15]. Specifically, given condition *c*, we train a conditional model ϵ_{θ} (xt, *c*, *t*) and an unconditional model ϵ_{θ} (xt, *t*) simultaneously. Then we can sample with:

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{s} \cdot \boldsymbol{\epsilon} \boldsymbol{\theta} \left(\boldsymbol{x}t, \boldsymbol{c}, t \right) - \left(\boldsymbol{s} - 1 \right) \cdot \boldsymbol{\epsilon} \boldsymbol{\theta} \left(\boldsymbol{x}_t, t \right), \tag{5}$$

where *s* is the guidance scale, and *c* denotes the condition.

4 METHOD

4.1 **Problem Formulation**

We tackle the problem of text-conditioned long-term 3D human motion generation. Given an input stream of text prompts S = $\{P^1, P^2, \cdots, P^N\}$, where each prompt P^i describes a desired action and N is the number of prompts, the goal is to generate a long-term realistic and coherent 3D human motion sequence $H = \{X^1, X^2, \cdots, X^N\}$ that sequentially contains the corresponding actions. Each sentence prompt $P^i = \{w_1^i, w_2^i, \cdots, w_{I_i}^i\}$ contains U_i words, where w_u^i is the *u*-th word, and the corresponding generated motion $X^i = \{x_1^i, x_2^i, \cdots, x_{L^i}^i\}$ has a predefined number L^i of frames, where x_l^i is the *l*-th frame. Each frame $x_l^i \in \mathbb{R}^d$, where *d* is the dimensionality of the human motion representation. Following [2], motions are parametrized by SMPL body models [29] and are converted to a 6D rotation representation together with root translation. Therefore d = 135 in our setting. We make the assumption in the following description of our method that there are only two segments of actions $H = \{X^1, X^2\}$ (which can be extended to an arbitrary number of segments). To avoid confusion with the timestep t in the diffusion model, we use $X = \{x^1, x^2, \dots, x^{L_X}\}$ to represent X^1 and $Y = \{y^1, y^2, \dots, y^{L_Y}\}$ to represent X^2 . The corresponding stream of text Prompts is $S = \{P^X, P^Y\}$.

4.2 Overview

Our proposed method consists of two main components: the first is the Past-Conditioned Human Motion Diffusion Model which can simultaneously incorporate text-guided conditions and the conditions of previous motions. The second consists of two coherent sampling methods that enable multiple motions to be seamlessly connected without the need for post-processing. The overview is shown in Fig. 2, the left part (a) illustrates the architecture of the Past-Conditioned Human Motion Diffusion Model, while the right part (b) demonstrates one of our proposed sampling methods, the Compositional Transition Sampling strategy.

4.3 Past-conditioned Human Motion Diffusion Model

We model the forward diffusion process of the second action Y as

 $q(Y_t | Y_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}Y_{t-1}, (1 - \alpha_t)I),$ (6) where *t* is the timestep of the diffusion model (*t* = 0 means the clean motion data) and α_t are hyper-parameters to control the noise level of the forward process. We model the reverse distribution as a conditional distribution $p(Y_{t-1} | Y_t, P^Y, X^h)$, where $X^h = X^{L_X - h + 1:L_X}$ and *h* is a predefined hyperparameter that controls the last *h* frames of P_X as the conditional input.

Our model architecture, PCMDM, is inspired by MDM [48] and TEACH [2]. We employ a Transformer Encoder as the backbone for text-conditioned motion generation while incorporating a Past Encoder (PC) to provide contextual information about the previous motion. Note that when the current motion is the first segment, we have $X^h = \emptyset$. As shown in Fig. 2 (b), the text prompt P^Y is encoded through a pre-trained frozen text encoder. We obtain the token



Figure 2: An overview of the proposed method. (Left): The architecture of our diffusion model, which is a Transformer Encoder used for predicting the denoised motion during the reverse process. The current text prompt is encoded using a frozen text encoder to obtain text features, which are added to the timestep embedding to obtain the token z_t . The last h frames of the previous motion are encoded using the Past Encoder and are also used as token inputs. After being concatenated with the noisy second motion $y_t^{1:L_Y}$ at the *t*-th timestep, these tokens are fed into the Transformer Encoder to predict the denoised motion. (Right): Our proposed Compositional Transition Sampling. The compositional motion $m_t^{1:L_X+L_Y}$ is split into the two noisy motions $x_t^{1:L_X}$ and $y_t^{1:L_Y}$ at the *t*-th timestep. We add the first $L_{Tr/2}$ frames of $y_t^{1:L_Y}$ to $x_t^{1:L_X}$ to obtain $x'_t^{1:L'_X}$. We obtain $y'_t^{1:L'_Y}$ in a similar way. We then separately feed them into the diffusion model to predict the denoised motions $\hat{x'}_0^{1:L'_Y}$ and $\hat{y'}_0^{1:L'_Y}$, respectively. They have an overlap of L_{Tr} frames, which is the transition. By composing the overlapping part, the two motions are concatenated to form the denoised compositional motion $\hat{m}_0^{1:L_X+L_Y}$, which is further used to sample $m_{t-1}^{1:L_X+L_Y}$ at the (t-1)-th timestep.



Figure 3: The conventional sampling methods cannot force adjacent motions to align. With our proposed Coherent Sampling methods, we can ensure that the generated motions are coherent and aligned.

 z_t by adding embeddings of text and timestep t. X^h is encoded through the Past Encoder. We concatenate these condition tokens and the current motion Y_t as the input to the Transformer Encoder. Note that our network $Y_\theta(Y_t, t, P^Y, X^h)$ does not predict the noise ϵ as in DDPM [14], but directly predicts the clean human motion sequence Y_0 , which has proved to be effective in human motion generation in [48]. The training objective is:

$$\mathcal{L} = E_{Y_0 \sim q\left(Y_0 | \boldsymbol{P}^{\boldsymbol{Y}}, \boldsymbol{X}^h\right), t \sim [1, T]} \left[\left\| Y_0 - Y_\theta \left(\boldsymbol{Y}_t, t, \boldsymbol{P}^{\boldsymbol{Y}}, \boldsymbol{X}^h \right) \right\|_2^2 \right].$$
(7)

Once $Y_{\theta}\left(Y_t, t, P^Y, X^h\right)$ is trained, we can recover $\mu_{\theta}\left(Y_t, t, P^Y, X^h\right)$ as proven in Eq. (2) and Eq. (4). In the following text, for the sake of convenience, we may use either Y_{θ} or μ_{θ} to represent the diffusion model backbone and we will ignore the conditions P^Y and X^h .

4.4 Sampling Strategies for Ensuring Motion Continuity

With PCMDM, we can generate long-term motions in an autoregressive manner using a general diffusion model sampling method, similar to TEACH [2]. However, we have observed a problem of discontinuity between adjacent motions, as shown in Fig. 3. To address this issue, we have designed two improved sampling methods: Inpainting Sampling and Compositional Transition Sampling. These methods naturally align adjacent motions without the need for alignment and interpolation between motions.

Inpainting Sampling. We cast the long-term 3D human motion generation as an image inpainting problem [43, 32, 34]. In inpainting, the objective is to predict missing pixels of an image using a know region as a condition, while in our case, we aim to predict future motion using the ending frames of past generated motion as a condition.

For a pair of adjacent text streams $S = \{P^X, P^Y\}$, assuming that we have generated X_0 . The last h frames of X_0 are denoted by $x_0^{L_X-h+1:L_X}$. We define $X^{ref} = [x_0^{L_X-h+1:L_X}; 0^{h+1:h+L_Y}]$. We first sample Y_T of length $h + L_Y$ from the standard Gaussian distribution. We then generate a binary mask $m = [1^{1:h}; 0^{h+1:h+L_Y}]$ where positions with respect to history frames $x_0^{L_X-h+1:L_X}$ are set to 1 and other positions are set to 0.

The difference with the original sampling step is that we overwrite the masked history motion (where m = 1) with the reference motion $x_0^{L_X-h+1:L_X}$ at each step before sampling. At the *t*-th step, the reverse sampling process to generate Y_{t-1} given Y_t is as follows:

$$\hat{Y}_0 = Y_\theta \left(Y_t, t \right), \tag{8a}$$

$$\hat{Y}_0 = \boldsymbol{m} \odot \boldsymbol{X}^{ref} + (1 - \boldsymbol{m}) \odot \hat{Y}_0.$$
(8b)

To sample Y_{t-1} , we can use the modified predicted \hat{Y}_0 . This means that Y_{t-1} is sampled while conditioning on the known human motion generated for the previous sentence. This approach called Past Inpainting sampling, connects the generated motion Y_0 with the previous motion X_0 .

Compositional Transition Sampling. TEACH and Past Inpainting Sampling assume that only previously generated actions affect subsequent ones. However, in real continuous motion, previous and future actions may have mutual influences. For example, the transitions from "running, jumping" to "running, sitting down" have significant semantic differences, indicating that autoregressive generation is not sufficient. Therefore, we propose a one-shot sampling method that models the distribution of transitions as the product of two adjacent different conditional action distributions. This method allows for the simultaneous consideration of the mutual interaction between previous and future actions in generating long-term human motion.

The idea of sampling from distributions that are the product of different experts originally comes from energy-based models(EBM) [6, 5]. [28] builds on this idea by exploring the connection between diffusion models and energy-based models, and proposes a method of directly adding the weighted noise predicted by diffusion models trained under different conditions to generate images with multiple instructions through the product of these distributions:

$$\hat{\mu}\left(\boldsymbol{x}_{t},t\right) = \sum_{i=1}^{n} w_{i}\left(\mu_{\theta}\left(\boldsymbol{x}_{t},t\mid\boldsymbol{c}_{i}\right)\right),\tag{9}$$

where c_i represents different conditions, and w_i represents the weight of each condition's impact. In theory, the distribution sampled in this way is only equivalent to the product of different distributions when t = 0 and t = T, theoretically. However, the experimental results in generating composite conditional images have shown good performance.

Our goal is to generate multiple actions that can be smoothly connected. It is non-trivial to directly apply this method to long-term human motion generation, because image generation applies all conditions to the same image canvas, while motion only exhibits such combined condition effects in relatively short transitions. Therefore, we define the transition between two adjacent actions X and Y:

$$T\boldsymbol{r} = \{\boldsymbol{x}^{L_X - L_T/2 + 1}, \cdots, \boldsymbol{x}^{L_X}, \boldsymbol{y}^1, \cdots, \boldsymbol{y}^{L_T/2}\},$$
(10)

Algorithn	1 1 Cor	npositional	Transition	Samp	ling

1: **Require** Human Motion Diffusion Model: $\mu_{\theta} (X_t, t \mid P^X)$ and $\mu_{\theta} (Y_t, t \mid P^Y)$; length of $X: L_X$; length of $Y: L_Y$; number of transition frames L_{Tr} ; covariances $\tilde{\beta}_t$; weight function w; Operator **SPLIT** 2: Initialize $M_T \sim \mathcal{N}(0, I)$ 3: **for** t = T, ..., 1 **do** 4: $X'_t, Y'_t \leftarrow$ **SPLIT** (M_t) 5: $\mu^i_M \leftarrow w(i)\mu^i_{\theta}(X'_t, t \mid P^X) + (1 - w(i))\mu_{\theta}(Y'_t, t \mid P^Y)$ 6: $M_{t-1} \sim \mathcal{N}(\mu_M, \tilde{\beta}_t)$ 7: **end for** 8: Return M_0

where L_X , L_Y and L_T are the lengths of X, Y and Tr, respectively. To facilitate modeling, we add the length of $L_T/2$ to the end of Xand the beginning of Y, resulting in X' and Y':

$$X' = \{ \boldsymbol{x}^1, \cdots, \boldsymbol{x}^{L_X}, \boldsymbol{y}^1, \cdots, \boldsymbol{y}^{L_T/2} \},$$

$$Y' = \{ \boldsymbol{x}^{L_X - L_T/2 + 1}, \cdots, \boldsymbol{x}^{L_X}, \boldsymbol{y}^1, \cdots, \boldsymbol{y}^{L_Y} \}.$$
(11)

We represent the whole motion obtained by concatenating *X* and *Y* as M = [X;Y]. We define an operator **SPLIT** to directly obtain *X'* and *Y'* from *M*:

$$\mathbf{SPLIT}(\mathbf{M}) = (\mathbf{X'}, \mathbf{Y'}). \tag{12}$$

Note that we have $X' \cap Y' = Tr$. The Transition between adjacent actions should contain semantic information that simultaneously captures both of their characteristics. Thus we assume that the distribution of Tr is the product of the distributions of $q^{X'}(x)$ and $q^{Y'}(x)$. Therefore we have:

$$q^{Tr}(\mathbf{x}) = \frac{1}{Z} q^{X'}(\mathbf{x}) q^{Y'}(\mathbf{x}), \quad Z = \int q^{X'}(\mathbf{x}) d\mathbf{x} q^{Y'}(\mathbf{x}) d\mathbf{x}, \quad (13)$$

here Z is a normalization constant used to ensure that the integral

where Z is a normalization constant used to ensure that the integral of the probability distribution equals 1. Intuitively, the meaning of this distribution is that the high probability region of $q^{Tr}(\mathbf{x})$ is also a high probability region under the probability distributions of both $q^{X'}(\mathbf{x})$ and $q^{Y'}(\mathbf{x})$. In this way, we have defined the product of distributions at the transition, and therefore, we can use the method described in Eq. (9) to sample the transition:

$$\hat{\mu}_{Tr} (Tr_t, t) = (\mu_{\theta}(X'_t, t \mid P^X) + \mu_{\theta}(Y'_t, t \mid P^Y))/2.$$
(14)

We assume that the influence of the preceding and subsequent actions on the transition is equally important. Therefore, we set the weight coefficients *w* to $\frac{1}{2}$.

In the previous discussion, we consider the transition positions of two actions, and we can obtain the composed transition by simply adding the weighted sum of μ_{θ} for each motion. For other non-overlapping areas, we can directly use the backbone PCMDM to predict μ_{θ} for single-text-guided sampling. The sampling formula for **M** as a whole can be written as:

$$\mu_M^i = w(i)\mu_{\theta}^i(X_t^i, t \mid P^X) + (1 - w(i))\mu_{\theta}(Y_t^i, t \mid P^I),$$
 (15)
where μ_M^i represents the sampling mean at the *i*-th frame, and $w(i)$ is a time-dependent weight function used to determine the position of the *i*-th frame:

$$w(i) = \begin{cases} 1 & \text{if } i \le L_X - L_T/2 \text{ or } i > L_Y + L_T/2 \\ 1/2 & \text{otherwise.} \end{cases}$$
(16)

The complete sampling process is in Alg. 1. With Compositional Transition Sampling, we can directly obtain coherent actions without the need for alignment and interpolation.

5 EXPERIMENTS

5.1 Dataset

In traditional text-to-motion datasets, each motion sequence is associated with a single text. However, the BABEL dataset [38] provides more fine-grained annotations for long-term motion sequences, with each subsequence corresponding to a specific text annotation. In total, there are 10881 motion sequences, with 65926 subsequences and the corresponding textual labels. Following TEACH [2], we create subsequence pairs by taking adjacent pairs of subsequences from the long sequences. For example, if the original data contains a motion sequence ['walk', 'sit down', 'wave right hand'], we can construct two subsequence pairs from it: ['walk', 'sit down'] and ['sit down', 'wave right hand']. The BABEL dataset itself comes with pre-defined train and test splits, and we adopt the default data partitioning of BABEL. After dividing the long sequences into subsequence pairs, there are approximately 15.7k and 5.7k pairs in the training and testing sets respectively.

Then we will elaborate on the specific training details for conducting training on such a dataset. Assuming we have a training data example with a text stream of ['walk', 'sit down'], and the corresponding motion sequence consist of two subsequences aligned with the prompts. Since MDM [48] does not have a Past Encoder, its training objective is to independently generate the corresponding subsequences for each prompt. On the other hand, TEACH [2] and PCMDM have a Past Encoder, so the training objective for the "walk" subsequence is to generate the motion based on the prompt alone, while for the "sit down" subsequence, the training objective is to generate the second motion by taking both the previous motion subsequence and the prompt "sit down" as conditions to feed into the model.

5.2 Evaluation Metrics

TEACH uses Average Positional Error (APE) and Average Variational Error (AVE) for evaluation, both of which calculate the distance between generated motions and ground truth motions. Such metrics are not suitable for evaluating generative models. Following recent work [10, 48] on text-guided human motion generation, we use other alternative metrics to evaluate our method.

To this end, following [20], we adapt CLIP [39] and train a MotionCLIP with a motion encoder and a text encoder on the training set through contrastive learning. We concatenate adjacent motion pairs as long-term motion and separate adjacent text descriptions with a comma as long-term text. These are then used as inputs to the MotionCLIP, as shown in Fig. 4. These two encoders can be used to evaluate the following metrics: **Frechet Inception Distance (FID)** represents the distribution divergence from generated samples to real data. A lower value implies better FID results. **R Precision** is calculated by putting the ground-truth text and a set of randomly selected mismatched descriptions from the test set into a pool for each generated motion. We calculate the Euclidean distance between the motion feature and the text feature of each description in the pool and count the average top-3 accuracy. **Diversity** measures the variance of the generated motions across all generated human motions. **Matching Scores** calculate the distance between the motion feature and the corresponding text feature extracted by the encoders.

In addition, a key metric for evaluating the quality of generated long-term human motion is the coherence between adjacent actions. Therefore, we use the Transition Distance proposed by [2]. **Transition Distance** quantifies the degree of discontinuity by calculating the average transition distance, which is defined as the Euclidean distance between the body poses of the last frame of the preceding action and the first frame of the subsequent action. Our method does not require alignment and outperforms previous methods even without it.



Figure 4: The training architecture of MotionClip.

5.3 Implementation details

We trained our human motion diffusion models (both MDM and PCMDM) for 50,000 steps using the AdamW optimizer [21, 30] with a fixed learning rate of 10^{-4} . The minibatch size was set to 32 for both training and evaluation. The MDM architecture is the same as in [48], and our PCMDM is an MDM with a past-conditioned encoder, which is a linear layer. During testing, we randomly selected 1000 text prompts and their corresponding lengths from the test set, generated continuous motion pairs, and calculated the metrics for these 1000 generated data points for the first four metrics. We repeated this experiment 10 times. For the Transition Distance, we conducted a single experiment on the entire test set following [2]. We conducted experiments on various types of GPUs, mainly NVIDIA Tesla V100-32GB GPUs and NVIDIA Tesla A800-80GB GPUs. We used the code and model weights provided for the Teach baselines experiments.

5.4 Comparison with State-of-the-art

TEACH is the only existing work that focuses on long-term 3D human motions from text streams, so we mainly compared our proposed method with all the baseline models proposed in TEACH. Among them, TEACH-Independent is a VAE model trained directly on single text-action pairs, TEACH-Joint is a VAE trained on longer sequences by connecting adjacent text-action pairs, and TEACH incorporates information from previous actions into the training of subsequent actions using a Past-condition module similar to our proposed method. MDM is a diffusion model trained directly on single text-action pairs, and PCMDM is a diffusion model with previous action information injected.

Method	$\mathrm{FID}\downarrow$	R-Precision(TOP 3)↑	$MultimodalDist \downarrow$	diversity \rightarrow
Real	$0.009^{\pm 0.001}$	$0.773^{\pm 0.002}$	$21.860^{\pm 0.006}$	$15.034^{\pm 0.078}$
TEACH-Independent [2] TEACH-Joint [2] TEACH [2]	$12.256^{\pm.0.284} \\ 13.084^{\pm.0.284} \\ 7.312^{\pm.0.0190}$	$\frac{0.816}{0.783^{\pm.0.004}}$ $0.864^{\pm.0.005}$	$21.874^{\pm.0.115}$ $22.218^{\pm.0.103}$ $21.017^{\pm.0.078}$	$13.905^{\pm.0.066} \\ 13.624^{\pm.0.094} \\ 14.2483^{\pm.0.070}$
MDM [48] MDM w/ Inpainting Sampling MDM w/ Compositional Sampling	$7.476^{\pm 0.232} \\ 7.411^{\pm .0.242} \\ 7.057^{\pm .0.232}$	$\begin{array}{c} 0.770^{\pm 0.0.008} \\ 0.772^{\pm .0.005} \\ 0.782^{\pm .0.006} \end{array}$	$\begin{array}{c} 22.020^{\pm0.098} \\ 22.038^{\pm.0.109} \\ 21.868^{\pm.0.110} \end{array}$	$\begin{array}{c} 13.876^{\pm.0.071} \\ 13.788^{\pm.0.083} \\ 14.112^{\pm.0.0871} \end{array}$
PCMDM PCMDM w/ Inpainting Sampling PCMDM w/ Compositional Sampling	$\frac{5.396^{\pm.0.187}}{5.431^{\pm.0.176}}$ 5.242 ^{$\pm.0.131$}	$\begin{array}{c} 0.775^{\pm.0.010} \\ 0.778^{\pm.0.008} \\ 0.799^{\pm.0.007} \end{array}$	$21.653^{\pm.0.120}$ $21.646^{\pm.0.126}$ $21.412^{\pm.0.087}$	$\frac{\underline{14.5338}^{\pm.0.063}}{\underline{14.501}^{\pm.0.084}}$ $\underline{14.652}^{\pm.0.068}$

Table 1: Comparison with Previous State of the Art. We report the performance of TEACH, MDM, and our PCMDM with our proposed sampling methods. *w/ Inpainting Sampling*" and *w/ Compositional Sampling*" refer to using our proposed novel sampling methods on diffusion models. We run all the evaluations 10 times. Bold indicates the best result, <u>underline</u> indicates the second-best result, \pm indicates a 95 confidence interval, and \rightarrow indicates that closer to real is better.

Methods	Transition Dist↓		
	w/ align.	w/out align.	
TEACH-Independent [2]	0.151	0.177	
TEACH-Joint [2]	0.107	0.122	
TEACH [2]	0.107	0.122	
MDM [48]	n/a	0.248	
MDM w/ Inpainting Sampling	n/a	0.116	
MDM w/ Compositonal Sampling	n/a	0.009	
PCMDM	n/a	0.119	
PCMDM w/ Inpainting Sampling	n/a	0.090	
PCMDM w/ Compositonal Sampling	n/a	0.014	

Table 2: We measure the transition distance for generated samples. "w/ align." indicates that the second motion is translated and rotated to match the first motion. Our method achieves significantly better results than other methods without the need for alignment.

As shown in Tab. 1, our proposed PCMDM with Compositional Transition Generation achieves state-of-the-art results on FID and diversity, indicating that our method can generate high-quality long-term human motion. Our method performs slightly worse than Teach on R-precision and MultimodalDist, which may be due to the fact that we used a smaller text encoder, Clip base, compared to Teach's DistilBert. It is also possible that our MotionCLIP was not fed with sufficient training data, resulting in encoder features that are not good enough, as our metrics are closer to real-world test data.

In addition, as shown in Tab. 2, we can see that our method's generated actions are much more consistent than the baselines. It is worth noting that our method outperformed the baseline by a factor of 10 on Transition Distance without any post-processing. This indicates that our proposed method can generate more coherent and smooth long-term human motion sequences, which is crucial for real-world applications such as virtual reality and robotics. **5 5**

5.5 Ablation Study

We provide an extensive analysis of the proposed PCMDM with a special sampling method, focusing on the influence of adjustable but



Figure 5: Abalation study on text condition scale *s*. We observe s = 2 obtains the best performance.

important parameters: (1) inpainting frames and transition frames and (2) text condition scale.

Effect of inpainting frames h and transition frames L_{Tr} . As shown in Fig. 3, when the values are set to smaller values, the performance is relatively better. This may be because using too many frames to guide coherent generation can harm the semantic information of individual motions.

Effect of text condition weight *s*. The conditional weight coefficient *s* in Eq.(5) is a crucial parameter in the classifier-free guidance, where a larger value leads to higher fidelity to the condition but may decrease the faithfulness to the true distribution of the original data. Through experiments, we found that s = 2 works best for our task.

5.6 Qualitative Results of Long-Term Motion Generation

In Section 5.4, we simplify the challenging task of generating coherent long-term 3D human motion by following the setting of [2], where the text stream contains two text prompts. However, our method is easily extendable to generate longer coherent 3D human motions for streams with more than two text prompts. In Fig. 6, we present several examples of our method generating motions with two or more challenging prompts. Our results demonstrate that our approach is capable of generating highly realistic 3D human motions, with smooth and natural transitions between successive

Zhao Yang, Bing Su and Ji-Rong Wen

Method	$\mathrm{FID}{\downarrow}$	R-Precision(Top 3)↑	$MultimodalDist \downarrow$	diversity \rightarrow
REAL	$0.009^{\pm 0.001}$	$0.773^{\pm 0.002}$	$21.860^{\pm 0.006}$	$15.034^{\pm 0.078}$
inpainting frames $h = 2$ inpainting frames $h = 4$ inpainting frames $h = 8$ inpainting frames $h = 12$	$5.431^{\pm 0.176}$ $5.416^{\pm 0.179}$ $5.491^{\pm 0.164}$ $5.613^{\pm .0.171}$	$\begin{array}{c} 0.778^{\pm 0.008} \\ 0.778^{\pm 0.009} \\ \textbf{0.779}^{\pm 0.006} \\ 0.775^{\pm 0.007} \end{array}$	$\begin{array}{c} \textbf{21.646}^{\pm 0.126} \\ \textbf{21.675}^{\pm 0.115} \\ \textbf{21.692}^{\pm 0.093} \\ \textbf{21.740}^{\pm 0.089} \end{array}$	$\begin{array}{c} \textbf{14.501}^{\pm .0.084} \\ \textbf{14.483}^{\pm 0.091} \\ \textbf{14.470}^{\pm 0.085} \\ \textbf{14.436}^{\pm 0.121} \end{array}$
transition frames $L_{Tr} = 2$ transition frames $L_{Tr} = 4$ transition frames $L_{Tr} = 8$ transition frames $L_{Tr} = 12$	$5.242^{\pm.0.131}$ $5.269^{\pm0.122}$ $5.285^{\pm0.151}$ $5.627^{\pm0.174}$	$\begin{array}{c} 0.799^{\pm.0.007} \\ \textbf{0.804}^{\pm0.0.006} \\ 0.803^{\pm0.0.007} \\ 0.787^{\pm0.0.007} \end{array}$	$21.412^{\pm .0.087}$ $21.399^{\pm 0.084}$ $21.457^{\pm 0.089}$ $22.628^{\pm 0.092}$	$\begin{array}{c} \textbf{14.652}^{\pm .0.068} \\ \textbf{14.636}^{\pm .0.082} \\ \textbf{14.639}^{\pm .0.102} \\ \textbf{14.600}^{\pm .0.067} \end{array}$

Table 3: Ablation study on the number of inpainting frames in inpainting sampling and transition frames in compositional



Figure 6: Qualitative results of our proposed method on long-term human motion generation. Our method generates coherent and diverse 3D human motions for streams with more than two prompts. Different colored motions represent different text prompts. On the left, we show the overall motion in a single image to observe positional changes. On the right, we display the motion frame by frame, making it easier to see details of the body pose changes.

actions. Fig. 6 presents our generated motions in two different formats. On the left, we render the motion onto a single image, making it easier to observe motions with positional changes in space. On the right, we display the long-term human motion frame by frame, making it easier to see details of the body pose changes.

6 CONCLUSION

In this paper, we propose a novel approach for generating longterm 3D human motions from text streams, by utilizing a pastconditioned human motion diffusion model and two coherent sampling methods called Past Inpainting and Transition Composition Sampling. Our approach generates coherent long-term motions corresponding to each sentence in the input stream without the need for post-processing. Our experimental results demonstrate that our proposed method is able to generate high-quality 3D human motions controlled by a user-instructed long text stream.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China No. 61976206 and No. 61832017, Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098, Beijing Academy of Artificial Intelligence (BAAI), the Fundamental Research Funds for the Central Universities, the Research Funds of Renmin University of China 21XNLG05, and Public Computing Cloud, Renmin University of China.

MM'23, October 29-November 3, 2023, Ottawa, ON, Canada

REFERENCES

- Chaitanya Ahuja and Louis-Philippe Morency. 2019. Language2pose: natural language grounded pose forecasting. In 2019 International Conference on 3D Vision (3DV). IEEE, 719–728.
- [2] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. 2022. Teach: temporal action composition for 3d humans. In *International Conference* on 3D Vision 2022.
- [3] Pablo Cervantes, Yusuke Sekikawa, Ikuro Sato, and Koichi Shinoda. 2022. Implicit neural representations for variable length human motion generation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October* 23–27, 2022, Proceedings, Part XVII. Springer, 356–372.
- [4] Yilun Du, Conor Durkan, Robin Strudel, Joshua B Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Grathwohl. 2023. Reduce, reuse, recycle: compositional generation with energy-based diffusion models and mcmc. arXiv preprint arXiv:2302.11552.
- Yilun Du, Shuang Li, and Igor Mordatch. 2020. Compositional visual generation with energy based models. Advances in Neural Information Processing Systems, 33, 6637–6647.
- [6] Yilun Du and Igor Mordatch. 2019. Implicit generation and modeling with energy based models. Advances in Neural Information Processing Systems, 32.
- [7] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. 2015. Recurrent network models for human dynamics. In Proceedings of the IEEE international conference on computer vision, 4346–4354.
- [8] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. 2022. Diffuseq: sequence to sequence text generation with diffusion models. arXiv preprint arXiv:2210.08933.
- [9] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In NIPS.
- [10] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022. Generating diverse and natural 3d human motions from text. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5152–5161.
- [11] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. 2020. Action2motion: conditioned generation of 3d human motions. In Proceedings of the 28th ACM International Conference on Multimedia, 2021–2029.
- [12] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. 2020. Robust motion in-betweening. ACM Transactions on Graphics (TOG), 39, 4, 60–1.
- [13] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. 2019. Human motion prediction via spatio-temporal inpainting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 7134–7143.
- [14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840– 6851.
- [15] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598.
- [16] Emiel Hoogeboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. 2022. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning*. PMLR, 8867–8887.
- [17] Yuqi Huo et al. 2021. Wenlan: bridging vision and language by large-scale multi-modal pre-training. arXiv preprint arXiv:2103.06561.
- [18] Jaehyeong Jo, Seul Lee, and Sung Ju Hwang. 2022. Score-based generative modeling of graphs via the system of stochastic differential equations. In International Conference on Machine Learning. PMLR, 10362–10383.
- [19] Manuel Kaufmann, Emre Aksan, Jie Song, Fabrizio Pece, Remo Ziegler, and Otmar Hilliges. 2020. Convolutional autoencoders for human motion infilling. In 2020 International Conference on 3D Vision (3DV). IEEE, 918–927.
- [20] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. 2022. Flame: free-form languagebased motion synthesis & editing. arXiv preprint arXiv:2209.00349.
- [21] Diederik P Kingma and Jimmy Ba. 2014. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980.
- [22] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.
- [23] Durk P Kingma and Prafulla Dhariwal. 2018. Glow: generative flow with invertible 1x1 convolutions. Advances in neural information processing systems, 31.
- [24] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. [n. d.] Diffwave: a versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*.
- [25] Taeryung Lee, Gyeongsik Moon, and Kyoung Mu Lee. 2022. Multiact: longterm 3d human motion generation from multiple action labels. arXiv preprint arXiv:2212.05897.
- [26] Buyu Li, Yongchi Zhao, Shi Zhelun, and Lu Sheng. 2022. Danceformer: music conditioned 3d dance generation with parametric motion transformer. In

Proceedings of the AAAI Conference on Artificial Intelligence number 2. Vol. 36, 1272–1279.

- [27] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. [n. d.] Diffusion-lm improves controllable text generation. In Advances in Neural Information Processing Systems.
- [28] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. 2022. Compositional visual generation with composable diffusion models. In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII. Springer, 423–439.
- [29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. Smpl: a skinned multi-person linear model. ACM transactions on graphics (TOG), 34, 6, 1–16.
- [30] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.
- [31] Qiujing Lu, Yipeng Zhang, Mingjian Lu, and Vwani Roychowdhury. 2022. Action-conditioned on-demand motion generation. In Proceedings of the 30th ACM International Conference on Multimedia, 2249–2257.
- [32] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: inpainting using denoising diffusion probabilistic models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11461–11471.
- [33] Julieta Martinez, Michael J Black, and Javier Romero. 2017. On human motion prediction using recurrent neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2891–2900.
- [34] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741.
- [35] Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhu Chen. 2022. Synthesizing coherent story with auto-regressive latent diffusion models. arXiv preprint arXiv:2211.10950.
- [36] Mathis Petrovich, Michael J Black, and Gül Varol. 2021. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10985–10995.
- [37] Mathis Petrovich, Michael J Black, and Gül Varol. 2022. Temos: generating diverse human motions from textual descriptions. arXiv preprint arXiv:2204.14109.
- [38] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. 2021. Babel: bodies, action and behavior with english labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 722–731.
- [39] Alec Radford et al. 2021. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning. PMLR, 8748-8763.
- [40] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125.
- [41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10684–10695.
- [42] Javier Romero, Dimitrios Tzionas, and Michael J Black. 2017. Embodied hands: modeling and capturing hands and bodies together. ACM Transactions on Graphics (TOG), 36, 6, 1–17.
- [43] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. 2022. Palette: image-to-image diffusion models. In ACM SIGGRAPH 2022 Conference Proceedings, 1–10.
- [44] Chitwan Saharia et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. arXiv preprint arXiv:2205.11487.
- [45] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In International Conference on Machine Learning. PMLR, 2256–2265.
- [46] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. Advances in neural information processing systems, 28.
- [47] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. [n. d.] Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Repre*sentations.
- [48] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. 2022. Human motion diffusion model. arXiv preprint arXiv:2209.14916.
- [49] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. [n. d.] Geodiff: a geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*.
- [50] Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. 2019. Convolutional sequence generation for skeleton-based action synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 4394– 4402.

MM'23, October 29-November 3, 2023, Ottawa, ON, Canada

- [51] [52]
- Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. 2022. Physdiff: physics-guided human motion diffusion model. *arXiv preprint arXiv:2212.02500*. Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2022. Motiondiffuse: text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*.

Zhao Yang, Bing Su and Ji-Rong Wen

[53] Rui Zhao, Hui Su, and Qiang Ji. 2020. Bayesian adversarial human motion synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6225–6234.

MM'23, October 29-November 3, 2023, Ottawa, ON, Canada



Figure 7: More visualized results.

Hyperparameter	Value
Optimizer	AdamW
Learning Rate	0.0001
Weight Decay	0.0001
Batch Size	32
Transformer Latent Dimension	256
Transformer Heads	4
Transformer Feedforward Dimension	1024
Transformer Num Layers	8
Dropout	0.1
Diffusion Noise Schedule	Cosine
Diffusion Step	1000
Variance	Fixed Small

Table 4: Hyperparameters of PCMDM.

Hyperparameter	Value
Optimizer	AdamW
Learning Rate	0.0001
Weight Decay	0.0001
Batch Size	32
Motion Transformer Latent Dimension	512
Motion Transformer Heads	8
Motion Transformer Feedforward Dimension	768
Motion Transformer Num Layers	6
Transformer Heads	4
Dropout	0.1
Text Encoder	CLIP ViT-B/32
Temperature of Contrastve Loss	0.1

Table 5: Hyperparameters of MotionCLIP.

A APPENDIX

We report more experimental results and more technical details which are not included in the paper due to space limit.

A.1 Hyperparameters.

To facilitate better replication of our experiment, we have provided detailed hyperparameter settings used in the experiment. Hyperparameters used in PCMDM are listed in Tab 4. Hyperparameters used in MotionClip are listed in Tab 5. The former mainly follows the setting of MDM [48], while the latter mainly follows the setting of Flame [20].

The architecture of PCMDM has been explained clearly in the main text. MotionCLIP consists of a Motion Encoder (a Transformer Encoder) and a pre-trained CLIP Text Encoder. We encode the action-text pairs in the training set separately with the Motion Encoder and Text Encoder and then use contrastive learning to train MotionCLIP by comparing the encoded results.

A.2 More visualized results.

We provide more visual results in Fig. 7. Different colored motions represent different text prompts. We show the overall motion in a single image.

A.3 Compositional Diffusion Models.

In the main text, we mentioned that conditional compositional sampling could be performed using a pre-trained diffusion model [28]. We will provide further explanation for this. We know that diffusion models are consistent with denoising score matching [47]:

$$\nabla_x \log q_\sigma(x) \approx -\frac{\epsilon_\theta(x,t)}{\sigma_t} \tag{17}$$

For the sake of exposition, we will prove this from the perspective of score matching with the score $\nabla_x \log q_\sigma(x)$, and our explanation is mainly based on [4].

For the simple combination distribution of two conditions, i.e. a product model $q^{\text{prod}}(x) \propto q^1(x)q^2(x)$. The true score of this distribution is given by:

$$\nabla_{x} \log \tilde{q}_{t}^{\text{prod}}(x_{t}) = \nabla_{x} \log \left(\int dx_{0} q^{1}(x_{0}) q^{2}(x_{0}) q(x_{t} \mid x_{0}) \right)$$
(18)

However, this score is difficult to obtain. We can approximate this true score through another simple score summation:

$$\nabla_{x} \log q_{t}^{\text{prod}}(x_{t}) = \nabla_{x} \log \left(\int dx_{0}q^{1}(x_{0}) q(x_{t} \mid x_{0}) \right) + \nabla_{x} \log \left(\int dx_{0}q^{2}(x_{0}) q(x_{t} \mid x_{0}) \right)$$
(19)

This is consistent with the formula for conditional combination derived from the perspective of the diffusion model in [28]. Note that when t = 0 we have:

$$\nabla_x \log \tilde{q}_t^{\text{prod}} = \nabla_x \log q_t^{\text{prod}}(x_t)$$
(20)

So we can use this sum-based method to approximate the true product distribution.

A.4 Generalization Evaluation.

While the proposed method demonstrates promising results in generating long-term human motion, the evaluation may be limited to specific datasets or scenarios, which may not fully represent the diversity and complexity of real-world applications. Conducting evaluations on a wider range of datasets and scenarios would strengthen the robustness and generalizability of the proposed method. MM'23, October 29-November 3, 2023, Ottawa, ON, Canada

Table 6: Remove Duplicates.

Method	$\mathrm{FID}\downarrow$	R-Precision (TOP 3) \uparrow	$Diversity \rightarrow$
Real	$0.018^{\pm 0.002}$	$0.821^{\pm 0.002}$	$15.415^{\pm 0.099}$
TEACH	$8.893^{\pm 0.019}$	$0.931^{\pm 0.004}$	$14.631^{\pm 0.051}$
MDM w/ IS	$7.662^{\pm 0.255}$	$0.824^{\pm 0.005}$	$14.202^{\pm 0.054}$
MDM w/ CS	$7.889^{\pm 0.272}$	$0.830^{\pm 0.006}$	$14.303^{\pm 0.045}$
PCMDM w/ IS	$5.483^{\pm 0.111}$	$0.858^{\pm 0.006}$	$14.827^{\pm 0.097}$
PCMDM w/ CS	$5.215^{\pm 0.153}$	$0.877^{\pm 0.004}$	$14.946^{\pm 0.073}$

Table 7: Query GPT-3.

Method	FID \downarrow	R-Precision (TOP 3) ↑	Diversity \rightarrow
Real	$0.009^{\pm 0.001}$	$0.297^{\pm 0.002}$	$14.985^{\pm 0.042}$
TEACH	$20.296^{\pm 0.483}$	$0.353^{\pm 0.005}$	$13.501^{\pm 0.090}$
MDM w/ IS	$18.670^{\pm 0.386}$	$0.358^{\pm 0.006}$	$12.775^{\pm 0.050}$
MDM w/ CS	$18.853^{\pm 0.209}$	$0.364^{\pm 0.009}$	$12.849^{\pm 0.074}$
PCMDM w/ IS	$21.062^{\pm 0.320}$	$0.332^{\pm 0.008}$	$13.043^{\pm 0.064}$
PCMDM w/ CS	$20.863^{\pm 0.140}$	$0.333^{\pm 0.009}$	$13.112^{\pm 0.057}$

Since BABEL is the only dataset suitable for our task, we designed two types of experiments based on BABEL to evaluate the generalization of our method. The first experiment involved exclusively using text streams in the test set that do not overlap with the action phrases present in the training set. The action phrase represents the smallest unit of a single action. For instance, if the action phrase "swing the golf club" appeared in the training set, we would remove all instances of data in the test set that include this particular action phrase. In total, there were 1992 unique phrases in the test set. After eliminating duplicates, the test set contained 1229 unique action phrases that did not appear in the training set. The results of the first experiment are shown in 6.

It can be observed that the diffusion model methods based on Inpainting Sampling (IS) and Compositional Sampling (CS) did not exhibit a significant decrease in FID, while TEACH [1] experienced a decrease from 7.312 to 8.893. Therefore, it can be concluded that our method demonstrates a significantly better generalization. In the original BABEL [3] dataset, the repeated action phrases are mostly short phrases with fewer words, e.g. "walk" and "stand". After removing duplicate action phrases, the average prompt length in the test set increased. The R-Precision, which is used to evaluate the accuracy of action and text retrieval, improves as the text prompt length increases. Longer text prompts provide more information, leading to an improvement in R-Precision for all methods, including real data.

In the second experiment, we obtained aliases for each action phrase by querying GPT-3 and replacing the text prompts in the test set with these aliases. This task presented a challenge because the original words in the dataset were mostly common words, and GPT-3 had the potential to transform the action phrases into combinations of less common words. For instance, the phrase "rotate left leg to the left behind" could be transformed by GPT-3 into "swivel left leg to the rear." The results of the second experiment are shown in 7.

Many of the text prompts given by GPT-3 contain uncommon words that our model has not seen in the training dataset. As a result, in such a setting, the performance of all methods sharply declines. The best-performing approach in this scenario is the MDM method based on our proposed Coherent Sampling. MDM [2] does not require a Past Encoder and is trained solely on individual subsequencetext pairs. Because in the case where the text prompt satisfies OOD, the generation of the previous subsequence already shows relatively poor performance. If we continue to encode the poorly generated subsequence using the Past Encoder and feed it to the following subsequence, it will result in error accumulation. Although Coherent Sampling does not show significant improvements in the test metrics, it is still necessary. Because using MDM's standard sampling method can only generate multiple disjointed subsequences, we need to use Coherent Sampling to make the generated sequences sufficiently coherent. The coherent Sampling technique is one of the most important contributions of our work.