

Enhanced Fine-Grained Motion Diffusion for Text-Driven Human Motion Synthesis

Dong Wei¹, Xiaoning Sun¹, Huaijiang Sun^{1*}, Shengxiang Hu¹, Bin Li², Weiqing Li¹, Jianfeng Lu¹

¹School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China

²Tianjin AiForward Science and Technology Co., Ltd., Tianjin, China

{csdwei, sunxiaoning, sunhuaijiang, hushengxiang, li_weiqing, lujf}@njust.edu.cn, libin@aiforward.com

Abstract

The emergence of text-driven motion synthesis technique provides animators with great potential to create efficiently. However, in most cases, textual expressions only contain general and qualitative motion descriptions, while lack fine depiction and sufficient intensity, leading to the synthesized motions that either (a) semantically compliant but uncontrollable over specific pose details, or (b) even deviates from the provided descriptions, bringing animators with undesired cases. In this paper, we propose *DiffKFC*, a conditional diffusion model for text-driven motion synthesis with KeyFrames Collaborated, enabling realistic generation with collaborative and efficient dual-level control: coarse guidance at semantic level, with only few keyframes for direct and fine-grained depiction down to body posture level. Unlike existing inference-editing diffusion models that incorporate conditions without training, our conditional diffusion model is explicitly trained and can fully exploit correlations among texts, keyframes and the diffused target frames. To preserve the control capability of discrete and sparse keyframes, we customize dilated mask attention modules where only partial valid tokens participate in local-to-global attention, indicated by the dilated keyframe mask. Additionally, we develop a simple yet effective smoothness prior, which steers the generated frames towards seamless keyframe transitions at inference. Extensive experiments show that our model not only achieves state-of-the-art performance in terms of semantic fidelity, but more importantly, is able to satisfy animator requirements through fine-grained guidance without tedious labor.

Introduction

Applying human motion synthesis to film or game industries (Van Welbergen et al. 2010) could greatly reduce the dependence on costly motion capture system. How to *efficiently* generate *high-quality* motions *towards animator needs* becomes a key issue. In terms of efficiency, the technique of text-driven motion synthesis has achieved initial success, where textual descriptions are used to provide an overall, coarse guidance at semantic level, under frameworks of GANs (Ahn et al. 2018) or VAEs (Guo et al. 2022; Petrovich, Black, and Varol 2022), but the synthesized motions often suffer from unrealistic behaviors, such as foot sliding.

*Corresponding author.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

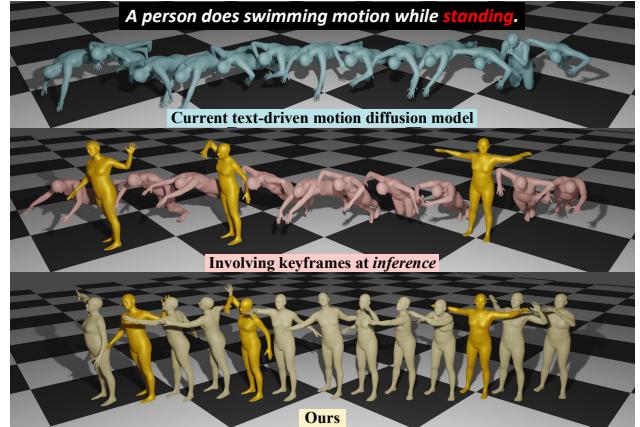


Figure 1: Top: Current text-driven motion diffusion models, such as MDM (Tevet et al. 2023), may miss the short *standing* word and generate a total *swimming* motion. Middle: Directly imposing keyframe (golden) conditions at *inference* fails to solve this problem, and results in heavy discontinuities. Bottom: Our collaborative dual-level control paradigm produces the realistic motion towards animator expectations.

Recent surge of denoising diffusion probabilistic models (DDPMs) has significantly improved the generative quality of text-to-motion synthesis (Zhang et al. 2022; Tevet et al. 2023; Dabral et al. 2023). However, the specific posture details are still uncontrollable for animators. Reaching animator expectations requires more than just semantic fidelity; it involves realizing various intricate details that are difficult to be explicitly expressed using free-form texts, which, unfortunately, are totally ignored in current models. Our research shows that this problem can be effectively solved by incorporating only a few keyframes as direct and fine-grained control, along with coarse textual guidance to establish a collaborative dual-level control paradigm. Additionally, this control paradigm allows for two incidental abilities that are also lacking in current text-to-motion models: correction of semantic misunderstanding and toleration on spelling errors.

Enabling fine-grained control over text-driven motion diffusion models is not trivial. Seemingly, an alternative way is to perform local editing at *inference stage*, which can be roughly divided into two strategies: inpainting (Lug-

mayr et al. 2022; Avrahami, Lischinski, and Fried 2022), which uses keyframes to in-fill the rest frames under the given texts; gradient-based guidance sampling (Dhariwal and Nichol 2021; Nichol et al. 2022), which perturbs the reverse-process mean with the gradient of the keyframe guidance. Nevertheless, such naive approaches cannot yield satisfactory results due to the following reasons: In practical scenarios, such as character animation and gaming, the keyframes provided by animators are often distributed in a *scattered* or *discrete* manner, and therefore tend to be perceived as noise by current motion diffusion models during inference. In this case, directly imposing keyframes at inference could not bring effective or beneficial impact, but result in huge gap between keyframes and the generated content, to which human eyes are very sensitive. Moreover, this defect may become even more pronounced when semantic misunderstanding happens, as shown in Figure 1.

In this paper, we propose *DiffKFC*, a conditional diffusion model for text-driven motion synthesis with KeyFrames Collaborated. To overcome the limitations caused by naive usage of keyframes, we proactively learn the dual-level collaboration between keyframes and textual guidance from early *training stage*. A carefully re-designed transformer structure is employed to fully exploits the internal correlations among multi-modal tokens of texts, keyframes and the diffused target frames. Considering the reduction of animator labor to ensure production efficiency, our DiffKFC should guarantee the effectiveness of keyframe control even when the provided number is small. Such sparsity of keyframes poses *another* challenge to their usage, as the useful keyframe information is prone to be overwhelmed when interacting with other massive tokens. To this end, we customize efficient Dilated Mask Attention (DMA) modules, which dilate the sparse valid tokens in keyframe mask step by step to “complete” the entire mask to be valid, where the scope of attention operation expands from local to global range, so that the keyframe information can gradually expand into the entire temporal dimension.

During inference, DiffKFC utilizes a simple yet effective temporal smoothness prior as backward propagation to encourage seamless transitions near keyframes, which is different from inference-editing diffusion models that solely rely on keyframes rectification. For user-friendly concerns, we extend the idea of classifier-free guidance (Ho and Salimans 2022), and reserve a variable for users to adjust the relative importance of fine-grained keyframe control under the given semantic space. Remarkably, our model is also compatible with flexible body part editing, such as re-generating the upper body part while remaining consistency and rationality with the fixed, conditional lower body part.

In summary, our contributions are as follows: (a) We are the first to utilize the controllable capability of keyframes for text-driven motion synthesis. It exploits full interaction among multi-modal tokens, enabling a collaborative dual-level control paradigm, i.e., coarse semantic guidance, with keyframes for direct yet fine-grained depiction. (b) We develop dilated mask attention modules to gradually borrow visible useful information from sparse keyframes with local-to-global attention, benefitting its fusion with other tokens.

(c) We show that, with very few keyframes (only 2%), our model achieves 41.6% improvement against state-of-the-art text-to-motion diffusion models in terms of Frechet Inception Distance (FID) on HumanML3D dataset. Meanwhile, it can help produce the exactly expected motions of animators. We also allow two incidental abilities: correcting semantic misunderstanding and tolerating spelling errors.

Related Work

Human Motion Synthesis

Human motion synthesis aims to generate realistic motion sequences based on given conditions. Various guidance that can be easily offered by users, such as action labels (Petrovich, Black, and Varol 2021; Guo et al. 2020), texts (Ahuja and Morency 2019; Ghosh et al. 2021; Guo et al. 2022; Petrovich, Black, and Varol 2022; Shafir et al. 2023; Jiang et al. 2023), music (Li et al. 2021; Aristidou et al. 2022), historical poses (Mao et al. 2019a; Sun et al. 2023) have been employed in motion generation, under frameworks of GANs or VAEs. Particularly, **text-driven motion synthesis** helps animators generate motions with high efficiency, as motion descriptions contained in free-form texts could guide motions towards the target semantics without tedious labor, while suffering from unrealistic behaviors, such as foot sliding and shaky movements. The recent **motion diffusion models** (Zhang et al. 2022; Kim, Kim, and Choi 2023; Tevet et al. 2023; Dabral et al. 2023) have alleviated this problem, where texts are encoded by pre-trained CLIP (Radford et al. 2021), to be fused with motion embeddings to yield higher generative quality. Among them, MDM (Tevet et al. 2023) is the most *representative* and *influential* model which employs a lightweight transformer-encoder backbone. The newly proposed MLD (Chen et al. 2023) further designs a latent diffusion model. GMD (Karunratanakul et al. 2023) additionally involves environment constraints into generations, such as obstacle avoidance. PhysDiff (Yuan et al. 2023) considers the natural laws of physics in the diffusion process to alleviate artifacts such as floating and foot sliding.

However, all the above models are only aimed at generating semantically compliant motions, while ignoring the animator requirements for specific posture details or visual effects that beyond textual descriptions. We propose to incorporate sparse yet useful keyframes for direct and fine-grained depiction, along with coarse textual guidance to form a collaborative dual-level control paradigm.

Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Models (DDPMs) (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020; Song and Ermon 2021) is a kind of generative models inspired by particle diffusion process in thermodynamics, and have gained huge popularity recently for their high generative quality. The basic elements of each sample (e.g., pixels of the image, joints of the human body) can be regarded as heated particles that gradually diffuse towards a complete noisy state, and the models will learn its reverse process (i.e., denoising) to recover the data distribution. Successes have been achieved in many fields, such as image generation (Ho, Jain,

and Abbeel 2020; Dhariwal and Nichol 2021; Lugmayr et al. 2022), point cloud generation (Luo and Hu 2021; Lyu et al. 2022), audio synthesis (Li et al. 2021; Aristidou et al. 2022) and human motion prediction (Wei et al. 2023) and synthesis (Zhang et al. 2022; Kim, Kim, and Choi 2023; Tevet et al. 2023; Dabral et al. 2023).

Concerning the motion synthesis task, **inference-editing with frame-wise conditions** is allowed in diffusion models, with no need for additional training. MDM (Tevet et al. 2023) and FLAME (Kim, Kim, and Choi 2023) can support motion in-betweening, by taking the beginning frames and ending frames as additional conditions, and denoising the rest frames with textual guidance. However, such approaches cannot be analogically applied to keyframe conditions. Unlike them, which condition on *consecutive* frames at the start and end of a sequence, our DiffKFC considers that the keyframes provided by animators are often distributed in a *scattered* or *discrete* manner. Consequently, these keyframes tend to be perceived as noise, and would gradually lose their control capability during inference. Therefore, we propose to proactively learn dual-level collaboration from training stage to avoid the above defect.

Proposed Method

We present an adapted formulation for our task. Existing plain text-driven motion synthesis is aimed at using a free-form textual description, usually a sentence \mathcal{C} , to generate towards the target N -frame motion sequence $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ with each human pose $\mathbf{x}_i \in \mathbb{R}^D$ represented by either joint rotations or positions. We, additionally, incorporate sparse keyframes $\mathcal{X}^{kf} = \mathcal{X} \odot \mathbf{M}$ at training stage, where \mathbf{M} is a binary mask matrix that preserves keyframe values and zeros the target frames (denoted as \mathcal{X}^{ta}) to be generated. Note that mask matrix \mathbf{M} for each sequence could be different.

Overview of Method. Our objective is to generate motions conformed to the given text, with sparse keyframes controlled towards the expectations of animators. As shown in Figure 2, we propose a conditional diffusion model with our re-designed transformer structure to generate realistic and controllable human motions. We train our model using classifier-free guidance to enable flexible and convenient adjustment on the relative importance of fine-grained control for animators. To generate smooth transitions near keyframes, we develop a simple yet effective smoothness prior at each denoising step. Our design further brings two incidental benefits of correcting semantic misunderstanding and tolerating spelling errors.

Diffusion Model with Keyframe Collaborated

Denoising diffusion probabilistic model can be described as a latent variable model with the *forward* diffusion process and the *reverse* diffusion process. Intuitively, the diffusion process gradually injects small amount of noise to the data until it is totally transformed to isotropic Gaussian noise, while the reverse process learns to gradually eliminate the noise to recover the original meaningful data. Recently, diffusion models have been introduced into human motion synthesis due to their powerful generative capabilities.

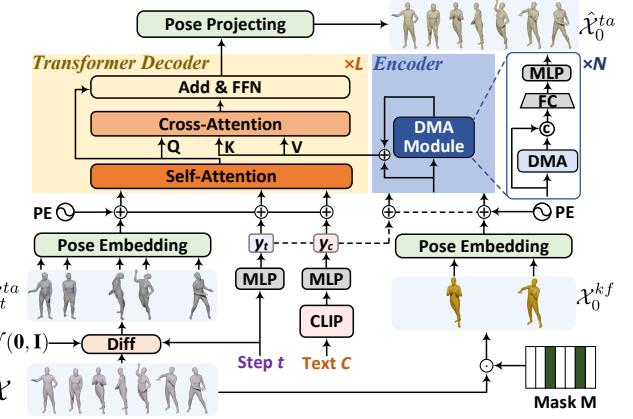


Figure 2: Overview of the proposed DiffKFC method. The encoder of DiffKFC is fed with clean keyframes \mathcal{X}_0^{kf} , while the decoder of DiffKFC is fed with the target frames \mathcal{X}_t^{ta} in a noising step t and the output of the encoder. The diffusion step t and text description \mathcal{C} are projected to the tokens \mathbf{y}_t and \mathbf{y}_c , which are then combined with input tokens. Dilated Mask Attention (DMA) modules are proposed to extract refined tokens from sparse keyframes and texts to conduct better interaction with encoded diffused motion tokens.

Unlike existing motion diffusion models (Zhang et al. 2022; Tevet et al. 2023; Kim, Kim, and Choi 2023) that are generally designed for text-to-motion generation, while bluntly involving the beginning and ending frames only during *inference* for motion in-betweening, our diffusion model, however, regards keyframes as an important element for fine-grained control *since training stage*, taking them as a conditional input to fully learn their internal information and proactively exploit their delicate correlations with texts.

Diffusion Process. Let \mathcal{X}_0^{ta} denotes the ground truth sequence (with N poses) that required to be generated. We define the forward diffusion process as a fixed posterior distribution $q(\mathcal{X}_{1:T}^{ta} | \mathcal{X}_0^{ta})$, which can be modeled as a Markov chain according to a fixed variance schedule β_1, \dots, β_T :

$$q(\mathcal{X}_t^{ta} | \mathcal{X}_{t-1}^{ta}) = \mathcal{N}(\mathcal{X}_t^{ta}; \sqrt{1 - \beta_t} \mathcal{X}_{t-1}^{ta}, \beta_t \mathbf{I}). \quad (1)$$

A special property is that the diffusion process can be represented in a closed form for any diffusion step t .

Reverse Process. Given a set of specified keyframe \mathcal{X}^{kf} and a text prompt \mathcal{C} , we consider the reverse dynamics of the above diffusion process to recover the meaningful sequence \mathcal{X}_0^{ta} from the white noise \mathcal{X}_T^{ta} . We formulate this reverse dynamics with the following parameterization:

$$p_\theta(\mathcal{X}_{t-1}^{ta} | \mathcal{X}_t^{ta}, \mathcal{X}_0^{kf}, \mathcal{C}) = \mathcal{N}(\mathcal{X}_{t-1}^{ta}; \mu_\theta(\mathcal{X}_t^{ta}, \mathcal{X}_0^{kf}, \mathcal{C}, t), \sigma_t^2 \mathbf{I}), \quad (2)$$

where μ_θ is a neural network to estimate the means, $\sigma_t^2 \mathbf{I}$ is a user-defined variance term of Gaussian transition, and \mathcal{X}_0^{kf} denotes the clean (not diffused) keyframe matrix.

Notably, inference-editing diffusion models usually approximate Eq.(2) with $p_\theta(\mathcal{X}_{t-1}^{ta} | \mathcal{X}_t^{ta}, \mathcal{X}_t^{kf}, \mathcal{C})$, which would harm useful information in the observed keyframes and mislead the text prompt, as depicted in (Tashiro et al. 2021).

Loss Function. As revealed in (Benny and Wolf 2022), the prediction of μ_θ can be replaced by either the prediction of ϵ , denoted as ϵ_θ , or the prediction of \mathcal{X}_0 , denoted as \mathcal{X}_θ . Here, we select \mathcal{X}_θ model due to the following reasons: First, \mathcal{X}_θ model could have more influence than ϵ_θ on the sampling mean when t approximates to 0. This is beneficial for fine-tuning using our smoothness prior at the last denoising steps. Second, for the task of motion generation, geometric losses are necessary to constrain the human poses, encouraging coherent and plausible motions. Therefore, the loss function can be derived as follows:

$$\mathcal{L}_{simple} = \mathbb{E}_{\mathcal{X}_0^{ta}, t} \left[\|(\mathcal{X}_0^{ta} - \mathcal{X}_\theta(\mathcal{X}_t^{ta}, \mathcal{X}_0^{kf}, \mathcal{C}, t)) \odot (\mathbf{1} - \mathbf{M})\|_2^2 \right], \quad (3)$$

where $\mathbf{1} - \mathbf{M}$ is a mask that corresponds to \mathcal{X}_0^{ta} . Similar to (Tevet et al. 2023), we adopt auxiliary kinematic losses \mathcal{L}_{phy} including joint positions, foot contact and velocities, to enforce physical properties and prevent artifacts like foot sliding.

Transition Guidance. The ability to generate natural transitions near keyframes is critical for realism of human motions. However, we find that the motions generated by our keyframes-conditioned model tend to have slight discontinuities near keyframes, which can be easily perceived by human eyes. To mitigate this problem, we leverage the idea of Discrete Cosine Transform (DCT) to define our temporal smoothness prior during inference, motivated by (Mao et al. 2019b). More concretely, we concatenate l generated frames before and after the i -th keyframe \mathbf{x}_i^{kf} to form a new sequence $\mathbf{G}_i = [\hat{\mathbf{x}}_{i-l}, \dots, \hat{\mathbf{x}}_{i-1}, \mathbf{x}_i^{kf}, \hat{\mathbf{x}}_{i+1}, \dots, \hat{\mathbf{x}}_{i+l}]$. We remove the high-frequency DCT basis, and approximate this sequence by $\hat{\mathbf{G}}_i = \mathbf{G}_i \mathbf{D} \mathbf{D}^T$, where $\mathbf{D} \in \mathbb{R}^{(2l+1) \times m}$ encodes the first m DCT bases. Our transition loss is defined as:

$$\mathcal{L}_{tr} = \frac{1}{(2l+1) \cdot K} \sum_{i=1}^K \|\hat{\mathbf{G}}_i - \mathbf{G}_i\|_2^2, \quad (4)$$

where K denotes the number of keyframes.

To apply this idea to diffusion models, we can replace the classifier with our smoothness prior in classifier guidance (Dhariwal and Nichol 2021). With guidance scale r , we perturb the reverse-process mean using the gradient of the transition loss with respect to keyframes:

$$\hat{\mu}_\theta(\mathcal{X}_t^{ta} | \mathcal{X}_0^{kf}) = \mu_\theta(\mathcal{X}_t^{ta} | \mathcal{X}_0^{kf}) + r \cdot \Sigma_\theta(\mathcal{X}_t^{ta} | \mathcal{X}_0^{kf}) \nabla_{\mathcal{X}_t^{ta}} \mathcal{L}_{tr}, \quad (5)$$

Classifier-free Guidance. From the perspective of animators, we wish to adjust the relative importance of fine-grained keyframes \mathcal{X}_0^{kf} on the motions against coarse-grained text description \mathcal{C} . To this end, we extend the core idea of the classifier-free guidance (Ho and Salimans 2022), which suggests interpolating or extrapolating between the conditioned model and the unconditioned model. Then, our classifier-free guidance inference can be implemented as: $\mathcal{X}_\theta(\mathcal{X}_t^{ta}, t, \mathcal{C}, \mathcal{X}_0^{kf}) = \mathcal{X}_\theta(\mathcal{X}_t^{ta}, t, \mathcal{C}, \emptyset) + s \cdot (\mathcal{X}_\theta(\mathcal{X}_t^{ta}, t, \mathcal{C}, \mathcal{X}_0^{kf}) - \mathcal{X}_\theta(\mathcal{X}_t^{ta}, t, \mathcal{C}, \emptyset))$, where s controls the intensity of keyframe control in text-driven motion synthesis. For training, our model learns both the dual-level controlled and coarse text-only controlled distributions by randomly dropping keyframes in 10% of the samples, such that $\mathcal{X}_\theta(\mathcal{X}_t^{ta}, \mathcal{C}, t)$ approximates $p(\mathcal{X}_0^{ta})$.

Network Architecture

To enable variable-length motion generation and capture global temporal information, we implement the denoising network \mathcal{X}_θ with a re-designed Transformer (Vaswani et al. 2017) structure, as depicted in Figure 2. Our model consists of feature extraction, the transformer decoder (orange) and the keyframe encoder (blue). In particular, feature extraction uses MLP to encode the target human poses and the diffused time step. The pre-trained CLIP (Radford et al. 2021) model is employed as the text encoder, similar to MDM. Our transformer decoder contains self-attention layers (SA), cross-attention layers (CA) and feed-forward network (FFN). It first models strong global correlation between any two tokens (including diffused target poses \mathcal{X}_t^{ta} , text \mathcal{C} and diffused step t) via SA mechanism. Then, it accepts the outputs of our keyframe encoder to exchange the keyframe information by CA layers, and is then fed into FFN. After the fusion with keyframe tokens, diffused pose tokens pass the learned information to its own branch, which naturally enriches the representation of each diffused pose token.

However, for our keyframe encoder, we find that directly using ViT encoder fails to borrow visible useful information for training. *This is common as given keyframes are sparse that tend to be overwhelmed when fused into CA layers.* One possible solution is to use motion in-betweening models (like SLERP (Shoemake 1985)) as a pre-processing step, which interpolates keyframes to form a full sequence, to be fed into CA layers. Nevertheless, this lowers the flexibility of keyframe selection, as it requires that at least the first frame and last frame must be given, and may introduce misleading guidance that harms generative realism and blurs motion details. That is where our motivation of DMA lies in.

Dilated Mask Attention Module. Inspired by (Li et al. 2022) for large hole image inpainting, we propose Dilated Mask Attention stack module to handle large number of missing (target) frame tokens (about 95% tokens are invalid). Our key insight is that, since the keyframe matrix contains important yet temporarily unusable information due to sparsity, if we take these visible sparse keyframes as starting points, and find a certain strategy to “complete” the entire sequence, then the “completed sequence” would interact with the diffused target frames and texts more effectively.

Suppose that the input of our keyframe encoder is denoted as $\mathbf{Z} \in \mathbb{R}^{d \times N}$. The attention of DMA can be derived by:

$$\text{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T + \mathbf{M}'}{\sqrt{d}}\right)\mathbf{V}, \quad (6)$$

where \mathbf{Q} , \mathbf{K} , \mathbf{V} are the learnable query, key, value matrices, respectively. The mask $\mathbf{M}' \in \mathbb{R}^{N \times N}$ is defined as:

$$\mathbf{M}'_{ij} = \begin{cases} 0, & \text{if token } j \text{ is valid,} \\ -\infty, & \text{if token } j \text{ is invalid,} \end{cases} \quad (7)$$

where only valid tokens participate in computing weights, while invalid tokens are ignored in current state.

The mask \mathbf{M}' indicates whether a token is valid, and it is initialized by the input keyframe mask \mathbf{M} . We design a valid-token-dilation strategy, which automatically activates invalid neighbors to become valid step by step. As illustrated

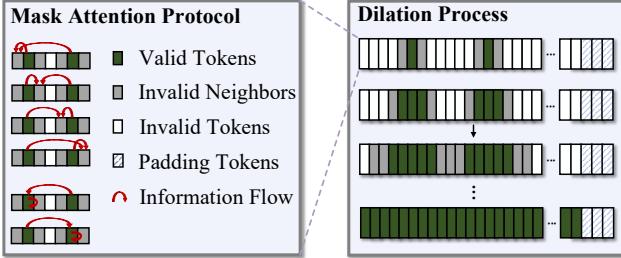


Figure 3: Illustration of dilation strategy. Left: Mask attention protocol. In the range of valid tokens and their invalid neighbors, the output of attention is computed as the weighted sum of valid tokens while invalid tokens are ignored. Right: Dilation process. The invalid neighbors are automatically updated to be valid through each attention. Finally, the whole sequence becomes fully valid except padding tokens (no benefit to interact with diffused tokens).

in Figure 3, at every dilation step, attention operation is conducted between every (**valid token, invalid neighbor**) pair, and therefore dilates the valid region. After several times of such operation, the whole sequence is updated to be fully valid (except padding region). Our dilation on valid regions is a gradual process, which forces keyframe information to gradually expand to the entire temporal dimension, with the following advantages: (a) benefitting full interaction in CA layers with tokens of diffused frames and texts; (b) alleviating the problems caused by keyframe sparsity, which helps to generate more natural and detailed motions.

In practice, we remove layer normalization to reduce the importance of invalid tokens, and take advantage of feature concatenation to replace residual learning similar to popular diffusion ViT backbone (Bao et al. 2023). The input and output of attention are concatenated to be fed into a fully-connected layer:

$$\mathbf{Z}'_k = \text{FC}([\text{DMA}(\mathbf{Z}_{k-1}) \parallel \mathbf{Z}_{k-1}]), \quad \mathbf{Z}_k = \text{MLP}(\mathbf{Z}'_k), \quad (8)$$

where \mathbf{Z}_k is the output of the k -th DMA block. A global skip connection is also adopted to prevent gradient exploding.

Experiments

In this section, to evaluate our proposed model, we introduce datasets, evaluation metrics, implementation details, and comparable baselines. Results, visualized comparisons with discussion are followed. Ablation study is conducted to show the impact of each component.

Datasets & Evaluation Metrics

KIT Motion-Language dataset (Plappert, Mandery, and Asfour 2016) is a text-to-motion dataset that contains 3,911 human motion sequences and 6,353 sentences of textual descriptions. Dataset split procedure is consistent with prior (Ahuja and Morency 2019; Petrovich, Black, and Varol 2022; Kim, Kim, and Choi 2023).

HumanML3D (Guo et al. 2022) is a new dataset that annotates existing 3D motion capture datasets AMASS (Mah-

mood et al. 2019) and HumanAct12 (Guo et al. 2020), containing 14,616 motions with 44,970 textual descriptions. Each frame is represented by the concatenation of root velocities, joint positions, joint velocities, joint rotations, and foot contact labels. We follow (Guo et al. 2022; Tevet et al. 2023) to use this representation as well as on KIT dataset.

Evaluation Metrics. We use the following matrices of text-to-motion task provided by (Guo et al. 2022). (1) *Frechet Inception Distance (FID)*: similarity between the distribution of generations and ground truth. (2) *R-Precision*: the alignment between the generated motion and corresponding text. (3) *Diversity*: the average body joint differences between each pairs of generated motions that are randomly split.

Particularly, to evaluate our distinctive fine-grained control property, we further employ (4) *Average Displacement Error (ADE)*: the average discrepancies over all timesteps (keyframe excluded) between the desired motion (i.e., GT) and the generated motion closest to GT, which is a metric borrowed from (Yuan and Kitani 2020) for diverse human motion prediction. We design (5) *Keyframe Generative Error (K-Err)*: the average discrepancies between the keyframe and the corresponding generated frame. (6) *Keyframe Transition Smoothness (KS)*: the average smoothness between the keyframe and its adjacent generated frames.

Baselines & Implementation Details

Baselines. We compare the performance of our model with text-to-motion baselines: JL2P (Ahuja and Morency 2019), Hier (Ghosh et al. 2021), T2M (Guo et al. 2022), MDM (Tevet et al. 2023), MoFusion (Dabral et al. 2023), MLD (Chen et al. 2023), GMD (Karunratanakul et al. 2023) and PhysDiff (Yuan et al. 2023).

Implementation Details. As for the diffusion model, the number of diffusion steps t is 1,000 with cosine beta scheduling following (Tevet et al. 2023; Kim, Kim, and Choi 2023). We employ a frozen *CLIP-ViT-B/32* (Radford et al. 2021) model to encode the text prompt \mathcal{C} . For each motion sequence, we select 5% frames as keyframes for fine-grained control, and at least 1 frame will be chosen as keyframe. The encoder of our DiffKFC has 8 layers, and their dilated step size (the number of invalid neighbors of each valid token) is set to $\{2, 2, 4, 4, 6, 6, 8, N\}$. As for the decoder, we build up 8 transformer layers with 8 heads, latent dimension as 512 and feed-forward size 1,024. During inference, the transition guidance scale r and classifier-free guidance scale s are set as 100.0 and 2.5, respectively. We use Adam optimizer with learning rate set to 0.0001. Our model is trained under Pytorch framework using NVIDIA RTX 3090, with batch size 64 for 500K steps on HumanML3D and 200K steps on KIT.

Results

Comparisons with plain text-driven baselines. Table 1 shows the results of DiffKFC and baselines, which prove the efficacy of incorporating keyframes. Our generative quality surpasses other models (including the most advanced motion diffusion models) by a significant margin, with *FID* obtaining 74.4% improvement over SOTA PhysDiff (Yuan et al. 2023) on HumanML3D, and 59.4% improvement over SOTA MLD (Chen et al. 2023) on KIT, under keyframe

	HumanML3D			KIT		
	FID ↓	R-Precision (Top-3) ↑	Diversity →	FID ↓	R-Precision (Top-3) ↑	Diversity →
Real	0.002 ^{±.000}	0.797 ^{±.002}	9.503 ^{±.065}	0.031 ^{±.004}	0.779 ^{±.006}	11.08 ^{±.097}
JL2P(3DV'19)	11.02 ^{±.046}	0.486 ^{±.002}	7.676 ^{±.058}	6.545 ^{±.072}	0.483 ^{±.005}	9.073 ^{±.100}
Hier(ICCV'21)	6.532 ^{±.024}	0.552 ^{±.004}	8.332 ^{±.042}	5.203 ^{±.107}	0.531 ^{±.007}	9.563 ^{±.072}
T2M(CVPR'22)	1.067 ^{±.002}	0.740 ^{±.003}	9.188 ^{±.002}	2.770 ^{±.109}	0.693 ^{±.007}	10.91 ^{±.119}
MoFusion(CVPR'23)	-	0.492	8.82	-	-	-
MLD(CVPR'23)	0.473 ^{±.013}	0.772^{±.002}	9.724 ^{±.082}	0.404 ^{±.027}	0.734^{±.007}	10.80 ^{±.117}
PhysDiff(ICCV'23)	0.433	0.631	-	-	-	-
GMD(ICCV'23)	0.212	0.670	9.440	-	-	-
MDM(ICLR'23)	0.544 ^{±.044}	0.611 ^{±.007}	9.559 ^{±.086}	0.497 ^{±.021}	0.396 ^{±.004}	10.85 ^{±.109}
DiffKFC	0.111^{±.031}	0.686 ^{±.006}	9.498^{±.082}	0.164^{±.026}	0.420 ^{±.007}	10.98^{±.108}

Table 1: Results of baselines and DiffKFC (with keyframe rate 5%) on HumanML3D and KIT datasets. → means results are better when closer to that of real motion. We evaluate with 20 times of running for each metric, under 95% confidence interval. Bold indicates best results; “-” means unavailable results.

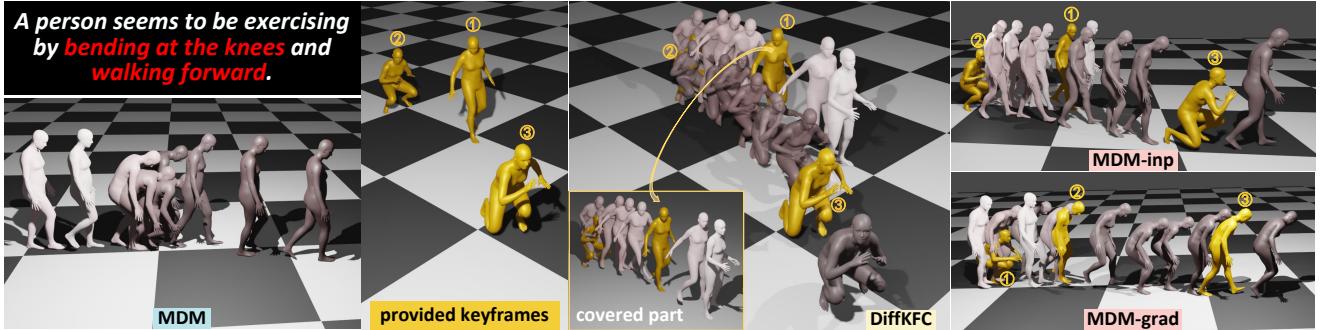


Figure 4: The description is ambiguous on whether “bending *then* walking” or “bending *while* walking”. MDM generates motions that compliant to the former semantics, but obviously do not meet animator needs (reflected in GT keyframes). Our DiffKFC not only removes this ambiguity (identifying the latter semantics), but also generates coherent and keyframe-aligned motions. Visualization of naive inference-editing approaches are also given. MDM-inp yields a continuous walking motion interspersed with single-frame of sudden knee bending, which is a typical unrealistic phenomenon; while for MDM-grad, the generated pose at keyframe positions are different from real keyframes, indicating its failure towards desired visual effects.

rate 5% of fine-grained control. Our *Diversity* is the closest to that of real motions, which means that our generations are more aligned with the natural law of human behaviors. Our method also achieves the comparable performance with baselines in terms of *R-Precision*.

Comparisons of fine-grained control property. To demonstrate the effectiveness of DiffKFC in enabling our distinctive fine-grained control, firstly, we need to prove the necessity of introducing keyframes into *diffusion training stage*, while existing text-driven motion diffusion baseline cannot acquire this control capability by *naive inference-editing*. Secondly, we need to prove that even if existing baseline has keyframes participating in *training*, simple fusion of texts and keyframes still fails to yield satisfactory improvements. *Only when both aspects above are proven can we conclude that DiffKFC is meaningful in enabling keyframe control to meet animator expectations.*

Since we build up our model on the basis of MDM transformer layers, we choose MDM as the baseline to conduct

	FID ↓	R-P Top-3 ↑	ADE ↓	K-Err ↓	KS →
Real	0.002	0.797	-	-	0.125
MDM	0.544	0.611	1.868	1.826	1.131
MDM-inp	0.386	0.642	1.498	-	0.847
MDM-grad	0.449	0.625	1.674	1.629	0.914
MDM-fus	0.308	0.658	0.343	0.219	0.254
w/o TG	0.119	0.686	0.198	0.037	0.161
w/ TG	0.111	0.686	0.205	0.041	0.136

Table 2: Results of MDM with different keyframe using and our DiffKFC on HumanML3D. Keyframe rate 5% is for all.

the following experiments to compare with. (a) **MDM-inp**, MDM inpainting at *inference*, which uses keyframes to infill the missing frames with textual guidance; (b) **MDM-grad**, MDM with gradient-based guidance sampling at *inference*, which perturbs the reverse-process mean with the

	FID ↓	R-P Top-3 ↑	Diversity →	ADE ↓	K-Err ↓
Vanilla Enc.	0.477	0.620	8.957	1.212	0.715
Unified Enc.	0.293	0.655	9.150	0.476	0.181
DiffKFC	0.111	0.686	9.498	0.205	0.041

Table 3: Ablation studies of network architecture designs. ‘Enc.’ is the abbreviation of the keyframe encoder.

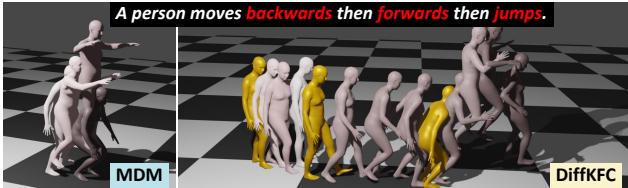


Figure 5: The motion generated by MDM is semantically incorrect (a total *jump* action), but our DiffKFC corrects this misunderstanding and generates the desired motion.

gradient of keyframe guidance, under the given texts; (c) **MDM-fus**, MDM with text-keyframe simple fusion at *training*, which uses state-of-the-art motion interpolation model (Oreshkin et al. 2022) to interpolate keyframes into a full sequence, and operates cross-attention between this sequence and text directly for condition fusion. In table 2, we show these comparisons with ours.

Apart from the best *FID* and *R-Precision*, DiffKFC significantly outperforms different keyframe using on MDM in terms of *ADE*, *K-Err* and *KS*. It means our generations are very close to the exactly expected motions, with natural keyframe transitions, which implies that our design is valid for fine-grained control towards animator needs.

Additionally, we remove transition guidance in our DiffKFC (i.e., w/o TG), and get worse *KS*, which proves that this strategy is helpful to generate smoother transitions.

Visualization. In Figure 4, our DiffKFC produces motions more than just realistic and semantically compliant; it ensures that the generations precisely adhere to the given keyframes, where animator requirements for specific posture details are predominantly reflected. However, this cannot be realized by naive inference-editing approaches. MDM-inp fails to exert effective fine-grained control over large number of missing frames, leading to undesired motion interspersed with unrealistic sudden pose change. MDM-grad can generate coherent motions, but the generated pose at keyframe positions are different from real keyframes, making the whole sequence less likely to meet the expectations.

We also show visualizations of correcting semantic misunderstanding and tolerating spelling errors in Figure 5 and 6, where our DiffKFC constantly outperforms the baseline. Notably, in all our figures, darker colors of human body indicate later occurrence.

Ablation Study

Discussion on the effectiveness of DMA. We discuss its effectiveness by conducting the following two experiments on HumanML3D, shown in Table 3. (a) We substitute our

Keyframe rate	FID ↓	R-P Top-3 ↑	Diversity →	ADE ↓	K-Err ↓
10%	0.079	0.694	9.500	0.124	0.039
5%	0.111	0.686	9.498	0.205	0.041
2%	0.253	0.665	9.441	0.387	0.050
0%	0.597	0.602	9.576	1.831	1.829

Table 4: Comparison with different keyframe rates.



Figure 6: True description: “A person walks forward and waves his left hand.” When spelling waves mistakenly into walks or wavs, MDM fails to extract the information and causes different failures, but our DiffKFC can tolerate both of these errors and produce the desired motion.

DMA-stack structure with a vanilla transformer encoder (i.e., Vanilla Enc.). Results are worse, as the information in sparse keyframes is overwhelmed when fused into CA layers, indicating that our dilation strategy is valid; (b) For conditional diffusion model, a common design for condition encoding is to directly concatenate the condition part and the diffused target part, which are then together fed into a unified encoder (i.e., Unified Enc.). However, such unified encoding would make it hard to distinguish between keyframes and noised targets, which harms the useful information and leads to unsatisfactory results. In our DiffKFC, we separately encodes keyframes and diffused target frames, with the specially designed DMA for keyframe encoding, and therefore avoid the above problem.

Discuss on keyframe rate. Compared to our keyframe number 5% of the total frames, we further set it to 10%, 2% and 0%, respectively, for evaluation. In Table 4, results of 0% keyframe deteriorate significantly (i.e., plain text-driven). However, when keyframes are involved, results of every setting all get improved. Although the more keyframes the better the results, it still proves that only few keyframes are sufficient to help yield satisfactory performance, which allows reduction on the dependence of dense drawing by animators.

Conclusion

In this paper, we propose DiffKFC, a conditional diffusion model for text-driven motion synthesis with keyframes collaborated. We carefully re-design a transformer structure to allow full interaction among multi-modal tokens, enabling realistic yet efficient generation under dual-level control paradigm. We customize dilated mask attention modules that gradually borrow visible useful information with local-to-global attention to overcome the challenge caused by keyframe sparsity. Our model not only achieves state-of-the-art performance in semantic fidelity, but also generates motions that satisfy animator expectations on details.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NO. 62176125, 61772272).

References

- Ahn, H.; Ha, T.; Choi, Y.; Yoo, H.; and Oh, S. 2018. Text2action: Generative adversarial synthesis from language to action. In *ICRA*, 5915–5920.
- Ahuja, C.; and Morency, L.-P. 2019. Language2pose: Natural language grounded pose forecasting. In *3DV*, 719–728. IEEE.
- Aristidou, A.; Yiannakidis, A.; Aberman, K.; Cohen-Or, D.; Shamir, A.; and Chrysanthou, Y. 2022. Rhythm is a Dancer: Music-Driven Motion Synthesis with Global Structure. *TVCG*.
- Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended diffusion for text-driven editing of natural images. In *CVPR*, 18208–18218.
- Bao, F.; Li, C.; Cao, Y.; and Zhu, J. 2023. All are Worth Words: a ViT Backbone for Score-based Diffusion Models. *CVPR*.
- Benny, Y.; and Wolf, L. 2022. Dynamic Dual-Output Diffusion Models. In *CVPR*, 11482–11491.
- Chen, X.; Jiang, B.; Liu, W.; Huang, Z.; Fu, B.; Chen, T.; Yu, J.; and Yu, G. 2023. Executing your Commands via Motion Diffusion in Latent Space. *CVPR*.
- Dabral, R.; Mughal, M. H.; Golyanik, V.; and Theobalt, C. 2023. MoFusion: A Framework for Denoising-Diffusion-based Motion Synthesis. *CVPR*.
- Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *NeurIPS*, 34: 8780–8794.
- Ghosh, A.; Cheema, N.; Oguz, C.; Theobalt, C.; and Slusallek, P. 2021. Synthesis of compositional animations from textual descriptions. In *ICCV*, 1396–1406.
- Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022. Generating diverse and natural 3d human motions from text. In *CVPR*, 5152–5161.
- Guo, C.; Zuo, X.; Wang, S.; Zou, S.; Sun, Q.; Deng, A.; Gong, M.; and Cheng, L. 2020. Action2motion: Conditioned generation of 3d human motions. In *ACM MM*, 2021–2029.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851.
- Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.
- Jiang, B.; Chen, X.; Liu, W.; Yu, J.; Yu, G.; and Chen, T. 2023. MotionGPT: Human Motion as a Foreign Language. *NeurIPS*.
- Karunratanakul, K.; Preechakul, K.; Suwajanakorn, S.; and Tang, S. 2023. Guided motion diffusion for controllable human motion synthesis. In *ICCV*, 2151–2162.
- Kim, J.; Kim, J.; and Choi, S. 2023. Flame: Free-form language-based motion synthesis & editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 8255–8263.
- Li, R.; Yang, S.; Ross, D. A.; and Kanazawa, A. 2021. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, 13401–13412.
- Li, W.; Lin, Z.; Zhou, K.; Qi, L.; Wang, Y.; and Jia, J. 2022. Mat: Mask-aware transformer for large hole image inpainting. In *CVPR*, 10758–10768.
- Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 11461–11471.
- Luo, S.; and Hu, W. 2021. Diffusion probabilistic models for 3d point cloud generation. In *CVPR*, 2837–2845.
- Lyu, Z.; Kong, Z.; Xudong, X.; Pan, L.; and Lin, D. 2022. A Conditional Point Diffusion-Refinement Paradigm for 3D Point Cloud Completion. In *ICLR*.
- Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 5442–5451.
- Mao, W.; Liu, M.; Salzmann, M.; and Li, H. 2019a. Learning trajectory dependencies for human motion prediction. In *ICCV*, 9489–9497.
- Mao, W.; Liu, M.; Salzmann, M.; and Li, H. 2019b. Learning trajectory dependencies for human motion prediction. In *ICCV*, 9489–9497.
- Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *ICML*.
- Oreshkin, B. N.; Valkanas, A.; Harvey, F. G.; Ménard, L.-S.; Bocquelet, F.; and Coates, M. J. 2022. Motion Inbetweening via Deep Δ -Interpolator. *arXiv preprint arXiv:2201.06701*.
- Petrovich, M.; Black, M. J.; and Varol, G. 2021. Action-conditioned 3D human motion synthesis with transformer VAE. In *ICCV*, 10985–10995.
- Petrovich, M.; Black, M. J.; and Varol, G. 2022. TEMOS: Generating diverse human motions from textual descriptions. In *ECCV*, 480–497. Springer.
- Plappert, M.; Mandery, C.; and Asfour, T. 2016. The KIT motion-language dataset. *Big data*, 4(4): 236–252.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.
- Shafir, Y.; Tevet, G.; Kapon, R.; and Bermano, A. H. 2023. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*.
- Shoemake, K. 1985. Animating rotation with quaternion curves. In *SIGGRAPH*, 245–254.
- Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2256–2265. PMLR.
- Song, Y.; and Ermon, S. 2021. Scorebased generative modeling through stochastic differential equations. *ICLR*.
- Sun, X.; Sun, H.; Li, B.; Wei, D.; Li, W.; and Lu, J. 2023. DeFeeNet: Consecutive 3D Human Motion Prediction with Deviation Feedback. In *CVPR*, 5527–5536.

Tashiro, Y.; Song, J.; Song, Y.; and Ermon, S. 2021. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. *NeurIPS*, 34: 24804–24816.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-Or, D.; and Bermano, A. H. 2023. Human motion diffusion model. *ICLR*.

Van Welbergen, H.; Van Basten, B. J.; Egges, A.; Ruttkay, Z. M.; and Overmars, M. H. 2010. Real time animation of virtual humans: a trade-off between naturalness and control. In *Computer Graphics Forum*, 2530–2554.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NeurIPS*, 30.

Wei, D.; Sun, H.; Li, B.; Lu, J.; Li, W.; Sun, X.; and Hu, S. 2023. Human joint kinematics diffusion-refinement for stochastic motion prediction. In *AAAI*, volume 37, 6110–6118.

Yuan, Y.; and Kitani, K. 2020. Dlow: Diversifying latent flows for diverse human motion prediction. In *ECCV*, 346–364. Springer.

Yuan, Y.; Song, J.; Iqbal, U.; Vahdat, A.; and Kautz, J. 2023. PhysDiff: Physics-Guided Human Motion Diffusion Model. *ICCV*.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2022. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*.