


ReMoDiffuse: Retrieval-Augmented Motion Diffusion Model

Mingyuan Zhang¹, Xinying Guo¹, Liang Pan¹, Zhongang Cai^{1,2}, Fangzhou Hong¹, Huirong Li¹,
Lei Yang², Ziwei Liu¹ 

¹S-Lab, Nanyang Technological University, Singapore

²Sensetime, China

{mingyuan001,XGU0012}@e.ntu.edu.sg, yanglei@sensetime.com, ziwei.liu@ntu.edu.sg

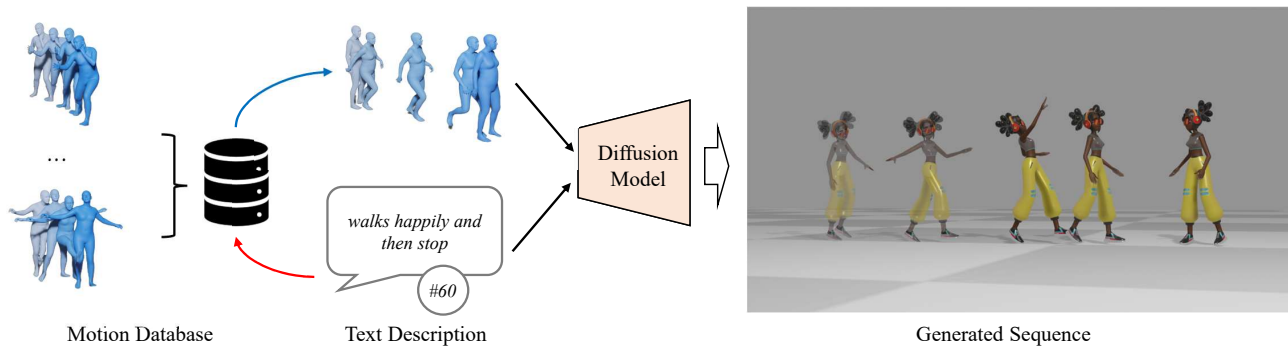


Figure 1: ReMoDiffuse is a retrieval-augmented 3D human motion diffusion model. Benefiting from the extra knowledge from the retrieved samples, ReMoDiffuse is able to achieve high-fidelity on the given prompts.

Abstract


3D human motion generation is crucial for creative industry. Recent advances rely on generative models with domain knowledge for text-driven motion generation, leading to substantial progress in capturing common motions. However, the performance on more diverse motions remains unsatisfactory. In this work, we propose **ReMoDiffuse**, a diffusion-model-based motion generation framework that integrates a retrieval mechanism to refine the denoising process. ReMoDiffuse enhances the generalizability and diversity of text-driven motion generation with three key designs: **1) Hybrid Retrieval** finds appropriate references from the database in terms of both semantic and kinematic similarities. **2) Semantic-Modulated Transformer** selectively absorbs retrieval knowledge, adapting to the difference between retrieved samples and the target motion sequence. **3) Condition Mixture** better utilizes the retrieval database during inference, overcoming the scale sensitivity in classifier-free guidance. Extensive experiments demonstrate that ReMoDiffuse outperforms state-of-the-art methods by balancing both text-motion consistency and motion quality, especially for more diverse motion generation. Project page: <https://mingyuan-zhang.github.io/projects/ReMoDiffuse.html>

github.io/projects/ReMoDiffuse.html

1. Introduction

Human motion generation has numerous practical applications in fields such as game production, film, and virtual reality. This has led to a growing interest in generating manipulable, plausible, diverse, and realistic human motion sequences. Traditional modeling processes are time-consuming and require specialized equipment and a significant amount of domain knowledge. To address these challenges, generic human motion generation models have been developed to enable the description, generation, and modification of motion sequences. Among all forms of human-computer interaction, natural language, in the form of text, provides rich semantic details and is a commonly used conditional signal in human motion generation.

Previous research has explored various generative models for text-driven motion generation. TEMOS uses a Variational-Auto-Encoder (VAE) to synthesize detailed motions, utilizing the KIT Motion-Language dataset [17]. Guo *et al.* [7] propose a two-stage auto-regressive approach for generating motion sequences. More recently, diffusion models have been applied to human motion generation due to their strength and flexibility. MotionDiffuse [27] gener-

 Corresponding author.

ates realistic and diverse actions while allowing for multi-level motion manipulation in both spatial and temporal dimensions. MDM [24] uses geometric losses as training constraints to make predictions of the sample itself. While these methods have achieved impressive results, they are not versatile enough for uncommon condition signals.

Some recent works on text-to-image generation utilize retrieval methods to complement the model framework, providing an retrieval-augmented pipeline to tackle the above issue [22, 4, 3]. However, simply transferring these methods into text-driven motion generation fields is impractical due to three new challenges. *Firstly*, the similarity between the target motion sequence and the elements in database is complicated. We need to evaluate both semantic and kinematic similarities to find out related knowledge. *Secondly*, a single motion sequence usually contains several atomic actions. It is necessary to learn from the retrieved samples selectively. In this procedure, the model should be aware of the semantic difference between the given prompt and retrieved samples. *Lastly*, motion diffusion models are sensitive to the scale in classifier-free guidance, especially when we supply another condition, retrieved samples.

In this paper, we propose a new text-driven motion generation pipeline, ReMoDiffuse, which addresses the above-mentioned challenges and thoroughly benefits from the retrieval techniques to generate diverse and high-quality motion sequences. ReMoDiffuse includes two stages: retrieval stage and refinement stage. In the retrieval stage, we expect to acquire the most informative samples to provide useful guidance for the denoising process. Here we consider both semantic and kinematic similarities and suggest a **Hybrid Retrieval** technique to achieve this objective. In the refinement stage, we design a **Semantics-Modulated Transformer** to leverage knowledge retrieved from an extra multi-modal database and generate semantic-consistent motion sequences. During inference, **Condition Mixture** technique enables our model to generate high-fidelity and description-consistent motion sequences. We evaluate our proposed ReMoDiffuse on two standard text-to-motion generation benchmarks, HumanML3D [7] and KIT-ML [17]. Extensive quantitative results demonstrate that ReMoDiffuse outperforms other existing motion generation pipelines by a significant margin. Additionally, we propose several new metrics for quantitative comparisons on uncommon samples. We find that ReMoDiffuse significantly improves the generation quality on rare samples, demonstrating its superior generalizability.

To summarize, our contributions are threefold: **1)** We carefully design a retrieval-augmented motion diffusion model which efficiently and effectively explores the knowledge from retrieved samples; **2)** We suggest new metrics to evaluate the model’s generalizability under different scenarios comprehensively; **3)** Extensive qualitative and quantitative

experiments show that our generated motion sequences achieve higher generalizability on both common and uncommon prompts.

2. Related Work

2.1. Diffusion Models

Diffusion models [10, 14] is a new class of generative models that have achieved impressive progress on text-to-image generation tasks. Prafulla Dhariwal and Alex Nichol [5] propose a diffusion model-based generative model, which first outperforms Generative Adversarial Networks(GAN) and establishes a new state-of-the-art text-driven image generation task. Their success with this advanced generative model quickly attract attention from worldwide researchers. GLIDE [13] designs classifier-free guidance and proves its superiority compared to the CLIP guidance used in previous works. DALL-E2 [19] attempts to bridge the text embedding and image embedding from the CLIP [18]. It includes another diffusion model which tries to synthesize an image embedding from the text embedding.

Recently, some works have focused on employing retrieval methods as complements to the model framework, providing an idea to enhance the generalizability. KNN-Diffusion [22] uses k-Nearest-Neighbors (kNN) to train an efficient text-to-image model without any text, enabling the model to adapt to novel samples. RDM [3] replaces the retrieval examples with the user-assigned images. Then it can effectively transfer artistic style from these images into the generated one. Re-Imagen [4] leverages knowledge from the external database to free the model from memorizing rare features, striking a good balance between fidelity and diversity.

2.2. Text-Driven Motion Generation

Text-driven motion generation has witnessed significant progress recently. Earlier works focus on learning a joint embedding space between motion sequences and language descriptions deterministically. JL2P [1] attempts to create a joint embedding space by applying the same reconstruction task on both text and motion embedding. Specifically, JL2P encodes the input text and motion data separately by two different encoders for each modality. A motion decoder is then applied on both embeddings to reconstruct the original motion sequences, which are expected to be the same as the initial input. Ghosh *et al.* [6] further develop this idea by manually dividing each pose sequence into an upper one and a lower one to represent two different body parts. In addition, the proposed method integrates a pose discriminator to improve the generation quality further. MotionCLIP [23] attempts to enhance the generalizability of text-to-motion generation. It enforces the motion embedding to be similar

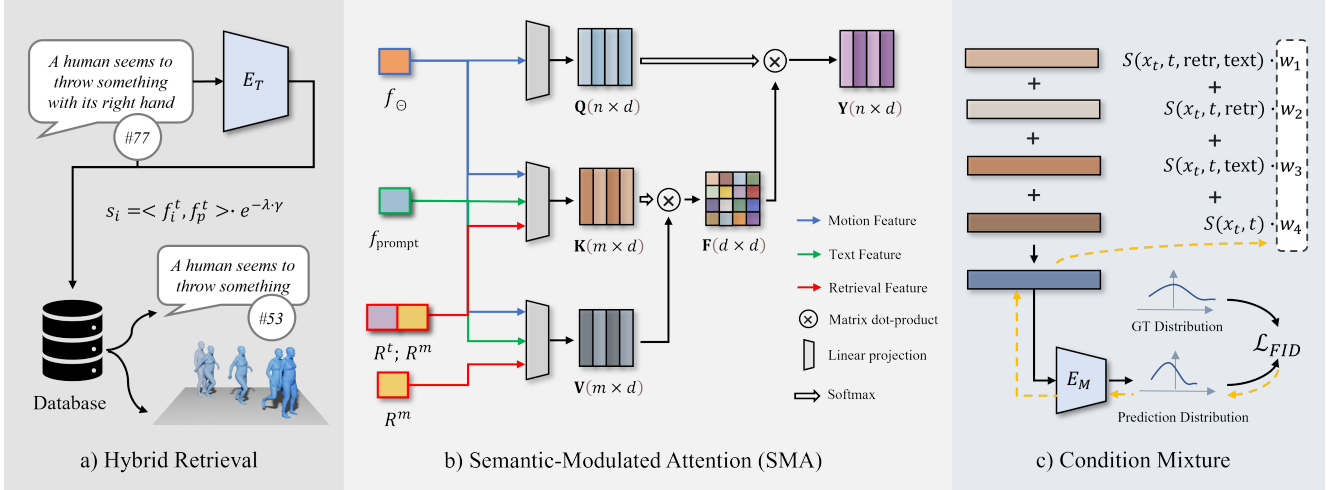


Figure 2: **Overview** of the proposed ReMoDiffuse. a) Hybrid retrieval database stores various features of each training data. The pre-processed text feature and relative difference of motion length are sent to calculate the similarity with the given language description. The most similar ones are fed into the semantics-modulated transformer (SMT), serving as additional clues for motion generation. b) Semantics-modulated transformer incorporates N identical decoder layers, including a semantics-modulated attention (SMA) layer and an FFN layer. The figure shows the detailed architecture of SMA module. CLIP’s extracted text features f_{prompt} from the given prompt, features R^t and R^m from the retrieved samples, and current motion features f_{Θ} will further refine the noised motion sequence. c) To synthesize diverse and realistic motion sequences, starting from the pure noised sample, the motion transformer repeatedly eliminates the noise. To better mix outputs under different combinations of conditions, we suggest a training strategy to find the optimal hyper-parameters w_1, w_2, w_3 and w_4 .

to the text and image embedding from critical poses. These two embeddings are acquired from CLIP [18], which excels at encoding texts and images into a joint space. Consequently, MotionCLIP can generate motion sequences with unseen descriptions.

To improve the diversity of generated motion sequences, previous works introduce variational mechanisms. TEMOS [16] employs a Variational Autoencoder (VAE) [11] to replace the deterministic auto-encoder structures. Besides, different from the recurrent neural networks in the previous works, both motion encoder and motion decoder in TEMOS is based on transformer architectures [25]. Guo *et al.* [7] propose an auto-regressive conditional VAE, which is conditioned on both the text feature and the previously generated frames. Given these conditions, the proposed pipeline will generate four successive frames as a unit. TEACH [2] also exploits auto-regressive models but in a larger length range. It can synthesize a long motion sequence with the given description and the previous sequence. Consequently, it can generate motion sequences with different actions continuously. TM2T [8] regards the text-driven motion generation task as a translation task between natural languages and motion sequences. Most recently, T2M-GPT [26] quantizes motion clips into discrete tokens and use a transformer to automatically generate later tokens.

Inspired by the success of diffusion models in text-to-

image generation tasks, some recent works have adapted this advanced generative model to motion generation tasks. MotionDiffuse [27] is an efficient DDPM-based architecture for plausible and controllable text-driven motion generation. It generates realistic and diverse actions and allows for multi-level motion manipulation in both spatial and temporal dimensions. MDM [24] is a lightweight diffusion model featuring a transformer-encoder backbone. It makes predictions of the sample rather than the noise so that geometric losses are supported as training constraints. Although these methods have outstanding performances on text-driven motion generation tasks, they are not versatile enough for uncommon condition signals. In this paper, we equip the diffusion model-based architecture with retrieval capability, enhancing the generalizability.

3. Our Approach

In this paper, we present a **Retrieval-augmented Motion Diffusion model (ReMoDiffuse)**. We first describe the overall architecture of the proposed method in Section 3. The background knowledge about the motion diffusion model will be discussed in Section 3.2. Then we will introduce our proposed novel retrieval techniques and the corresponding model structure in Section 3.3. Finally, we will introduce the training objective and sampling strategy in Section 3.5.

3.1. Framework Overview

Figure 2 shows the overall architecture of ReMoDiffuse. We establish the whole pipeline based on MotionDiffuse [27], which incorporates diffusion models and a series of transformer decoder layers. To strengthen its generalizability, we extract features from two different modalities to establish the retrieval database. During denoising steps, ReMoDiffuse first retrieves motion sequences based on the extracted text features and relative motion length. These retrieved samples are then fed into the motion transformer layers. As for each decoder layer, the noised sequence is refined by Semantics-Modulated Attention (SMA) layers and then absorbs information from the given description and the retrieved samples. In the classifier-free generation process, we have distinct outputs under different condition combinations. To better fuse these outputs, we finetune our model on the training split to find the optimal combination of hyperparameters w_1, w_2, w_3 and w_4 . We will introduce these components in the following subsections.

3.2. Diffusion Model for Motion Generation

Recently, diffusion models have been introduced into motion generation [27, 24]. Compared to VAE-based pipelines, the most popular motion-generative models in previous works, diffusion models strengthen the generation capacity through a stochastic diffusion process, as evidenced by the diverse and high-fidelity generated results. Therefore, in this work, we build our motion generation framework in a corporation with diffusion models.

Diffusion Models can be parameterized as a Markov chain $p_\theta(\mathbf{x}_0) := \int p_\theta(\mathbf{x}_{0:T}) d\mathbf{x}_{1:T}$, where $\mathbf{x}_1, \dots, \mathbf{x}_T$ are the noised sequences distorted from the real data $\mathbf{x}_0 \sim q(\mathbf{x}_0)$. All \mathbf{x}_t , where $t = 0, 1, 2, \dots, T$, are of the same dimensionality. In the motion generation tasks, each \mathbf{x}_t can be represented by a series of pose $\theta_i \in \mathbb{R}^D, i = 1, 2, \dots, F$, where D is the dimensionality of the pose representation and F is the number of the frames.

In the forward process of diffusion models, the computation of the posterior distribution $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$ is implemented as a Markov chain that gradually adds Gaussian noises to the data according to a variance schedule β_1, \dots, β_T :

$$\begin{aligned} q(\mathbf{x}_{1:T}|\mathbf{x}_0) &:= \prod_{t=1}^T q(\mathbf{x}_t|\mathbf{x}_{t-1}), \\ q(\mathbf{x}_t|\mathbf{x}_{t-1}) &:= \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}). \end{aligned} \quad (1)$$

To efficiently acquire \mathbf{x}_t from x_0 , Ho *et al.* [10] approximate $q(\mathbf{x}_t)$ as $\mathbf{x} := \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$, where $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^t \alpha_s$.

In diffusion models, the aforementioned forwarding Markov chain is reversed to learn the original motion distributions. Expressly, diffusion models are trained to denoise

noisy data \mathbf{x}_t into clean data \mathbf{x}_0 . Following MDM [24], we predict the clean state \mathbf{x}_0 . The training target can be written as:

$$\mathbb{E}_{x_0, \epsilon, t} [\mathbf{x}_0 - S(\mathbf{x}_t, t, \text{retr}, \text{text})], \quad (2)$$

where *retr* and *text* denote the conditions of retrieved samples and the given prompts respectively. Here $t \in \mathcal{U}(0, T)$ denotes the timestamp, which is uniformly sampled from 0 to the maximum diffusion steps T . $S(\mathbf{x}_t, t, \text{retr}, \text{text})$ indicates the estimated clean motion sequence, given the four inputs.

During the sampling process, we can sample \mathbf{x}_{t-1} from a Gaussian Distribution $\mathcal{N}(\mu_\theta(\mathbf{x}_t, t, c), \beta_t)$, where c denotes the condition of *retr* and *text* for simplicity. The mean of this distribution can be acquired from \mathbf{x}_t and $S(\mathbf{x}_t, t, c)$ by the following equation:

$$\begin{aligned} \mu_\theta(\mathbf{x}_t, t, c) &= \sqrt{\bar{\alpha}_t}S(\mathbf{x}_t, t, c) + \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(\mathbf{x}_t, t, c) \\ \epsilon_\theta(\mathbf{x}_t, t, c) &= \left(\frac{\mathbf{x}_t}{\sqrt{\bar{\alpha}_t}} - S(\mathbf{x}_t, t, c) \right) \sqrt{\frac{1}{\bar{\alpha}_t} - 1} \end{aligned} \quad (3)$$

Hence, on the basis of diffusion models, the text-driven motion generation pipeline should be able to predict the start sequence \mathbf{x}_0 , with the given conditions. In this paper, we propose a retrieval technique to enhance this denoising process. We will introduce how we retrieve motion sequences and how to fuse this information.

3.3. Retrieval-Augmented Motion Generation

Basically, there are two stages in retrieval-based pipelines. The first stage is to retrieve appropriate samples from the database. The second stage is acquiring knowledge from these retrieved samples to refine the denoising process of diffusion models. We will thoroughly introduce these two steps.

Hybrid Retrieval. To support this process, we need to extract features for calculating the similarities between the given text description and the entities in the database. Considering that the retrieval procedure is not differentiable, we have to utilize pre-trained models instead of using learnable architectures. An intuitive method is to generate text features on both query text and the data points. Thanks to the pre-trained CLIP [18], we can easily evaluate the semantic similarities from language descriptions. Formally, for each data point $(\text{text}_i, \Theta_i)$, we first calculate $f_i^t = E_T(\text{text}_i)$ as the text-query feature, where E_T is the text encoder in the CLIP model.

Text features usually encourage the retrieval process to select samples with high semantic similarities. These features play a significant role in retrieving suitable samples. However, there is another kind of feature that is vital but easily overlooked, the relative magnitude between the expected motion length and that of each entity in the database.

Hence, the similarity score s_i between i -th data point and the given description prompt and expected motion length L is defined as below:

$$s_i = \langle f_i^t, f_p^t \rangle \cdot e^{-\lambda \cdot \gamma},$$

$$f_p^t = E_t(\text{prompt}), \gamma = \frac{\|l_i - L\|}{\max\{l_i, L\}}, \quad (4)$$

where $\langle \cdot, \cdot \rangle$ denotes cosine similarity between the two given feature vectors, l_i is the length of the motion sequence Θ_i . The similarity score s_i becomes larger when text-query is closer to the prompt feature. When the expected motion length is close to the length of one entity, the corresponding s_i will also increase. This property is significant because the motion sequence with a similar length can provide more informative features for the generation. λ is a hyper-parameter to balance the magnitude of these two different similarities.

To establish the retrieval database, we simply select all the training data as entities. Given the number of retrieved samples k , prompt, and motion length L , we sort all elements by the score s_i in Equation 4. Then the most k similar ones are selected as the retrieved samples $(\text{text}_i, \Theta_i)$ and fed into the semantics-modulated attention components in the motion transformer. We will illustrate the detailed architecture in the next paragraph.

Network Architecture. Similar to MotionDiffuse [27] and MDM [24], we build up our pipeline on the basis of transformer layers as shown in Figure 2. In both semantics-modulated attention modules and FFN modules, following MotionDiffuse [27], we add a stylization block to fuse timestamp t into the motion generation process. First, an embedding vector e_t is obtained from the timestamp t . It should be mentioned that the original design in MotionDiffuse also uses an embedding vector from the given prompt, which is not suitable for classifier-free guidance. Then for each block, a residual shortcut is applied between the input $\mathbf{X} \in \mathbb{R}^{n \times d}$ and the output $\mathbf{Y} \in \mathbb{R}^{n \times d}$, where n is the number of elements and d is the dimensionality.

Two major difficulties should be resolved to better explore knowledge from the retrieved samples. First, in the literature of motion diffusion models [27, 24], the resolution of motion sequences is not reduced through the denoising process. The maximum length of one motion sequence is around 200 frames in the HumanML3D [7] dataset, leading to a dramatic computational cost, especially when we expect to retrieve more samples. Hence, efficiency is highly prioritized for the information fusion component. Second, the semantic relation between the retrieved samples and given prompts is complicated. For example, ‘a person is walking forward’ and ‘a person is walking forward slowly’ are highly similar. However, these two prompts will lead to two distinct motion sequences regarding pace and intensity.

Therefore, the model should know which motion features can be borrowed, guided by the difference between the language descriptions.

Based on these observations, we design two encoders to extract text features and motion features from the retrieved data, respectively. As for motion features, we expect them to be capable of providing low-level information while retaining the computational cost to an acceptable degree. Therefore, we build up a series of encoder layers, which include alternating Semantics-Modulated Attention(SMA) modules and FFN modules. This motion encoder processes raw motion sequences into usable ones. To reduce the computational cost, we down-sample the sequence into 1/4 original FPS, which is denoted as $R^m \in \mathbb{R}^{F' \cdot k \times D}$, where F' is the number of frames after down-sampling and k is the number of retrieved samples. This simple strategy greatly decreases the computation with little information lost. As for the text encoder, the feature $R^t \in \mathbb{R}^{k \times D}$ from the last token is supposed to represent the global semantic information. R^m and R^t constitute the features we needed for the purpose of retrieval-based augmentation.

Semantics-Modulated Attention. These extracted features will be passed to the cross attention component, as shown in Figure 2. The noised motion sequence forms the query vector $Q \in \mathbb{R}^{F \times D}$. As for the key vector K and the value vector V , we consider three sources of data: 1) The motion sequence $f_\Theta \in \mathbb{R}^{F \times D}$ itself. As shown in Figure 2, our proposed transformer does not contain a self-attention module. Instead, we combine the function self-attention into the SMA; 2) The text condition f_{prompt} , which semantically describes the expected motion sequence and is extracted as in MotionDiffuse [27]. Specifically, the prompt is first fed into the pre-trained CLIP model to get a feature sequence, which is further processed by two learnable transformer encoder layers; 3) Features R^m, R^t from the retrieved samples. We simply concatenate $f_\Theta, f_{\text{prompt}}, R^m$ for value vector V and $f_\Theta, f_{\text{prompt}}, [R^m; R^t]$ for key vector K . Here $[\cdot; \cdot]$ denotes the concatenation of both terms. This design allows our proposed method to fuse low-level motion information from the retrieved samples and also to fully consider the semantic similarities. The acquired vectors Q, K, V are sent to perform Linear Attention [21] for efficient computation.

Stylization Block. Similar to MotionDiffuse [27] and MDM [24], we build up our pipeline on the basis of transformer layers. In both semantics-modulated attention modules and FFN modules, following MotionDiffuse [27], we add a stylization block to fuse timestamp t into the motion generation process. First, an embedding vector e_t is obtained from the timestamp t . Then for each block, a residual shortcut is applied between the input $\mathbf{X} \in \mathbb{R}^{n \times d}$ and the

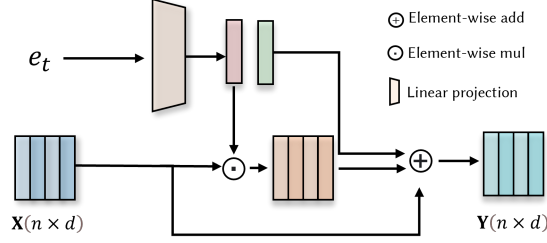


Figure 3: **Architecture** of the stylization block. This module is adapted from MotionDiffuse [27]. We remove the prompt embedding from the original design to better support classifier-free guidance. This module attempts to inject the information of the current timestamp into the feature representation, which is necessary for denoising steps. Specifically, the timestamp embedding e_t is fed into a series of transformation layers. Two embeddings are generated afterward and serve as an additive offset and a multiplicative offset to the original feature map, respectively.

output $\mathbf{Y} \in \mathbb{R}^{n \times d}$, where n is the number of elements and d is the dimensionality. The detailed structure is shown in Figure 3.

3.4. Condition Mixture

Classifier-free guidance enables us to generate motion sequences with both high fidelity and consistency with the given text description. A typical formulation is described as below:

$$\epsilon = w \cdot \epsilon_{\theta}(\mathbf{x}_t, t, \text{text}) - (w - 1) \cdot \epsilon_{\theta}(\mathbf{x}_t, t), \quad (5)$$

where w is a hyper-parameter to balance the text-consistency and motion quality. In our proposed retrieval-augmented diffusion pipeline, the given retrieved samples can be regarded as an additional condition. Therefore, we get four estimations: $S(\mathbf{x}_t, t, \text{retr}, \text{text})$, $S(\mathbf{x}_t, t, \text{retr})$, $S(\mathbf{x}_t, t, \text{text})$, $S(\mathbf{x}_t, t)$. We need four parameters to balance these items. To achieve a better performance, here we suggest a **Condition Mixture** technique to achieve this objective. Specifically, given the pre-trained Semantics-Modulated Transformer (SMT), we optimize the value of w_1, w_2, w_3, w_4 and get the final output \hat{S} as:

$$\hat{S} = w_1 \cdot S(\mathbf{x}_t, t, \text{retr}, \text{text}) + w_2 \cdot S(\mathbf{x}_t, t, \text{text}) + w_3 \cdot S(\mathbf{x}_t, t, \text{retr}) + w_4 \cdot S(\mathbf{x}_t, t). \quad (6)$$

Empirically, we find that the tendency of Frchet Inception Distance (FID) is similar to that of Precision when the hyper-parameters are nearly optimal. Hence, we only attempt to minimize the FID in this procedure.

Contrastive Model. To imitate the evaluator used in the standard evaluation process, we train our contrastive model,

which aims at encoding the paired text descriptions and motion sequences into a joint embedding space. As for the motion encoder, we use a 4-layer ACTOR [15] Encoder. The text encoder is identical to the one we used in ReMoDiffuse. The only difference is that we require a sentence feature instead of a sequence of word features. We train this contrastive learning model with the same loss in Guo *et al.* [7]. 20K and 40K optimization steps are applied for the KIT-ML and HumanML3D datasets, respectively.

Parameter Finetuning. As mentioned before, we only use 50 denoising steps to generate motion sequences in the inference stage. However, it is impractical to calculate the gradient through such several forward times. To simplify the problem, we divide all denoising steps into the first 40 steps and the last ten steps. In the first part, we use grid search to find a better parameter combination. Specifically, for Equation 6, we search w_1 and w_2 from $[-5, 5]$ with step 0.5 to find the best parameter for each model. Here we use inspiration from Re-Imagen [4] that set $w_4 = 0$. Besides, to retain the output’s statistics, we $w_1 + w_2 + w_3 + w_4 = 1$. These two properties enable us to find the optimal combination by only searching the value of w_1 and w_2 . The evaluation metric is the calculated FID between our generated sequences and the natural motion sequences in the training split performed by our trained contrastive model. This search aims to find an optimal combination of w_1 and w_2 to achieve the lowest FID.

In the second stage, we use an end-to-end training scheme to optimize w_1, w_2 , and w_3 . w_4 is acquired by $1 - w_1 - w_2 - w_3$. We use the Adam optimizer to train our model on the training split for 1K steps to find the best parameter combination.

We use the searched parameters during training to perform the first 40 denoising steps. After that, we auto-regressively denoise the motion sequence with learnable w_1, w_2 and w_3 . The training objective here is also reducing FID.

We use the Adam optimizer and train 1K steps for both HumanML3D and KIT-ML datasets to find the best parameter combination.

3.5. Training and Inference

Model Training. Inspired by the classifier-free technique, 10% of the text conditions and 10% of the retrieval conditions are independently randomly masked to approximate $p(\mathbf{x}_0)$. The training object is to minimize the mean square error between the predicted initial sequence and the ground truth, as shown in Equation 2. In the training stage, we typically use a 1000-steps diffusion process.

Model Inference. During each denoising step, we use the learned coefficients w_1, w_2, w_3 and w_4 to get \hat{S} as Equa-

Table 1: **Quantitative results on the HumanML3D test set.** For a fair comparison, all methods use the real motion length from the ground truth as the extra given information. ‘↑’(‘↓’) indicates that the values are better if the metric is larger (smaller). We run all the evaluations 20 times. $x^{\pm y}$ indicates that the average metric is x and the the 95% confidence interval is y . The best result and the second best result are in red cells and blue cells, respectively.

Methods	R Precision↑			FID↓	MM Dist↓	Diversity↑	MultiModality↑
	Top 1	Top 2	Top 3				
Real motions	0.511 $^{\pm.003}$	0.703 $^{\pm.003}$	0.797 $^{\pm.002}$	0.002 $^{\pm.000}$	2.974 $^{\pm.008}$	9.503 $^{\pm.065}$	-
Guo <i>et al.</i> [7]	0.457 $^{\pm.002}$	0.639 $^{\pm.003}$	0.740 $^{\pm.003}$	1.067 $^{\pm.002}$	3.340 $^{\pm.008}$	9.188 $^{\pm.002}$	2.090 $^{\pm.083}$
MDM [24]	-	-	0.611 $^{\pm.007}$	0.544 $^{\pm.044}$	5.566 $^{\pm.027}$	9.559 $^{\pm.086}$	2.799 $^{\pm.072}$
MotionDiffuse [27]	0.491 $^{\pm.001}$	0.681 $^{\pm.001}$	0.782 $^{\pm.001}$	0.630 $^{\pm.001}$	3.113 $^{\pm.001}$	9.410 $^{\pm.049}$	1.553 $^{\pm.042}$
T2M-GPT [26]	0.491 $^{\pm.003}$	0.680 $^{\pm.003}$	0.775 $^{\pm.002}$	0.116 $^{\pm.004}$	3.118 $^{\pm.011}$	9.761 $^{\pm.081}$	1.856 $^{\pm.011}$
Ours	0.510 $^{\pm.005}$	0.698 $^{\pm.006}$	0.795 $^{\pm.004}$	0.103 $^{\pm.004}$	2.974 $^{\pm.016}$	9.018 $^{\pm.075}$	1.795 $^{\pm.043}$

Table 2: **Quantitative results on the KIT-ML test set.**

Methods	R Precision↑			FID↓	MM Dist↓	Diversity↑	MultiModality↑
	Top 1	Top 2	Top 3				
Real motions	0.424 $^{\pm.005}$	0.649 $^{\pm.006}$	0.779 $^{\pm.006}$	0.031 $^{\pm.004}$	2.788 $^{\pm.012}$	11.08 $^{\pm.097}$	-
Guo <i>et al.</i> [7]	0.370 $^{\pm.005}$	0.569 $^{\pm.007}$	0.693 $^{\pm.007}$	2.770 $^{\pm.109}$	3.401 $^{\pm.008}$	10.91 $^{\pm.119}$	1.482 $^{\pm.065}$
MDM [24]	-	-	0.396 $^{\pm.004}$	0.497 $^{\pm.021}$	9.191 $^{\pm.022}$	10.847 $^{\pm.109}$	1.907 $^{\pm.214}$
MotionDiffuse [27]	0.417 $^{\pm.004}$	0.621 $^{\pm.004}$	0.739 $^{\pm.004}$	1.954 $^{\pm.062}$	2.958 $^{\pm.005}$	11.10 $^{\pm.143}$	0.730 $^{\pm.013}$
T2M-GPT [26]	0.416 $^{\pm.006}$	0.627 $^{\pm.006}$	0.745 $^{\pm.006}$	0.514 $^{\pm.029}$	3.007 $^{\pm.023}$	10.921 $^{\pm.108}$	1.570 $^{\pm.039}$
Ours	0.427 $^{\pm.014}$	0.641 $^{\pm.004}$	0.765 $^{\pm.055}$	0.155 $^{\pm.006}$	2.814 $^{\pm.012}$	10.80 $^{\pm.105}$	1.239 $^{\pm.028}$

tion 6. To reduce the computation cost introduced by the retrieved samples, we pre-process all f_i^v, f_i^t, R^t, R^m to ensure no repeated computation for different syntheses.

Different from the training stage, we carefully reduce the whole denoising process into 50 steps during inference, which enables our model to generate high-quality motion sequences efficiently.

4. Experiments

4.1. Datasets and Metrics

Datasets. We evaluate our proposed framework using the KIT dataset [17] and the HumanML3D dataset [7], two leading benchmarks in text-driven motion generation tasks. KIT Motion Language Dataset is an open dataset combining human motion and natural language, which contains 3,911 motions and 6,363 natural language annotations. HumanML3D is a scripted 3D human motion dataset that originates from and textually reannotates the HumanAct12 [9] and AMASS datasets [12]. Overall, HumanML3D consists of 14,616 motions and 44,970 descriptions.

Evaluation Metrics. We follow the performance measures employed in MotionDiffuse for quantitative evaluations, namely Frechet Inception Distance (FID), R Precision, Diversity, Multimodality, and Multi-Modal Distance. (1) FID is an objective metric calculating the distance between features extracted from real and generated motion sequences, which highly reflects the generation quality. (2) R-

precision measures the similarity between the text description and the generated motion sequence and indicates the probability that the real text appears in the top k after sorting, and in this work, k is taken to be 1, 2, and 3. (3) Diversity measures the variability and richness of the generated action sequences. (4) Multimodality measures the average variance of generated motion sequences given a single text description. (5) Multi-modal distance (MM Dist for short) represents the average Euclidean distance between the motion feature and its corresponding text description feature.

4.2. Implementation Details

We use similar settings on HumanML3D and KIT-ML datasets. As for the motion encoder, a 4-layer transformer is used, and the latent dimension is 512. As for the text encoder, a frozen text encoder used in the CLIP ViT-B/32, together with 2 additional transformer encoder layers, is built and applied. As for the diffusion model, the variances β_t are pre-defined to spread linearly from 0.0001 to 0.02, and the total number of noising steps is set to be $T = 1000$. Adam is adapted as the optimizer to train the model with a learning rate equal to 0.0002. 1 Tesla V100 is used for training, and the batch size on a single GPU is 128. Pieces of training on KIT-ML and HumanML3D are carried out for 40k and 200k steps respectively.

Pose representation in this work follows the schema used by Guo *et al.* [7]. The pose is defined as a tuple of length seven: $(r^{va}, r^{vx}, r^{vz}, r^h, \mathbf{j}^p, \mathbf{j}^v, \mathbf{j}^r)$, where $r^{va} \in \mathbb{R}$ is the root angular velocity along Y-axis, and $r^{vx}, r^{vz} \in \mathbb{R}$ are the root linear velocities along X-axis and Z-axis respectively.

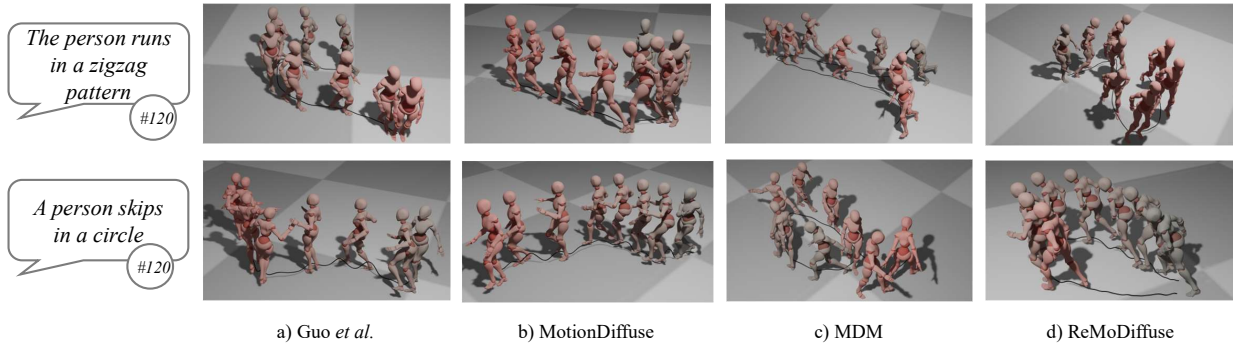


Figure 4: Visual Comparison between previous works and ReMoDiffuse. We draw black lines to show the translation path. As for both given conditions, only ReMoDiffuse conveys accurate action and path condition.

$r^h \in \mathbb{R}$ is the root height. $\mathbf{j}^p, \mathbf{j}^v \in \mathbb{R}^{J \times 3}$ are the local joints positions and velocities. $\mathbf{j}^r \in \mathbb{R}^{J \times 6}$ is the 6D local continuous joints rotations. J denotes the number of joints, and in HumanML3D and KIT-ML, J is 22 and 21 separately.

4.3. Main Results

Table 1 and Table 2 show the comparison between our proposed ReMoDiffuse and four other existing works, including recent diffusion models-based algorithms [24, 27], one VAE-based generative model [7], and one GPT-style generative model [26].

Compared to other diffusion model-based pipelines, our proposed ReMoDiffuse achieves a better balance between the condition-consistency and fidelity. It should be noted that, ReMoDiffuse is the first work to achieve state-of-the-art on both metrics, which demonstrates the superiority of the proposed pipeline.

4.4. Ablation Study

Retrieval Techniques. First, we investigate the influence of different retrieval techniques. To directly evaluate the similarity between the target samples and the given samples, we use retrieved samples as generated results and calculate the FID metric for them. We try different λ to balance the terms of semantic similarity and kinematic similarity. The results are shown in Figure 5. $\lambda = 0$ means that the kinematic similarity will not influence the retrieval process, whose retrieval quality is unacceptable. This result supports our claim that kinematic similarity is significant to the retrieval quality. The optimal value of λ is 0.1 for both KIT-ML and HumanML3D datasets.

Motion Refinement. We further evaluate the proposed cross attention component of our retrieval-augmented motion generation. In Table 3, when using the text feature, FID is enhanced remarkably. It strongly supports our claims that text features are highly significant in hybrid retrieval,

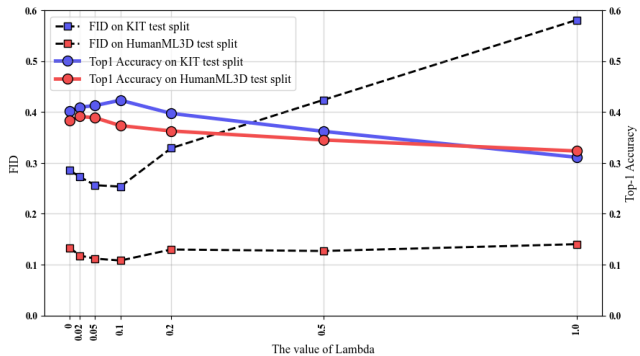


Figure 5: The retrieval performance of different λ . λ is used to balance semantic and kinematic similarity in the retrieval stage. A larger λ indicates the retrieval process focuses more on the kinematic similarity.

Table 3: **Ablation of the proposed architecture.** All results are reported on the KIT testset. ‘T’ and ‘M’ denote the usage of semantic similarity and kinematic similarity respectively. These two factors are considered in both retrieval and refinement stages.

	Retrieval	Attention	#Samples	Stride	FID \downarrow
a)	-	-	-	-	0.245 \pm .008
b)	T	M	2	4	0.314 \pm .012
c)	T& M	M	2	4	0.192 \pm .008
d)	T	T & M	2	4	0.307 \pm .010
e)	T& M	T& M	2	4	0.155 \pm .006
f)	T& M	T& M	1	4	0.186 \pm .008
g)	T& M	T& M	3	4	0.217 \pm .009

which is not discussed in the text-to-image generation tasks. Besides, the proposed retrieval techniques outperform the baseline by a remarkable margin.

4.5. Analysis on More Diverse Generation

Metrics on Diverse Generation. To fairly compare the generalization ability of our proposed ReMoDiffuse and

other existing works, *e.g.* MotionDiffuse [27], we propose several new metrics. Specifically, inspired by imbalanced regression task [20], here we propose two variants of the original Multimodality Distance. First, we give the definition of sample’s **Rareness**. As for a test prompt p , we calculate its rareness r_p as:

$$r_p = 1 - \max_i \{ \langle E_T(t_i), E_T(\text{prompt}) \rangle \}, \quad (7)$$

where E_T denotes the text encoder in the CLIP [18] model, t_i is the motion description in the training set, and $\langle \cdot, \cdot \rangle$ represents the cosine similarity of the two given vectors. Intuitively, this formulation measures the maximum similarity between the given prompt and training prompts. If this similarity is larger, then the rareness will be lower, and vice versa.

Based on the definition of rareness, we sort all samples in increasing order and define the following metrics: **1) tail 5% MM**, the average Multimodality Distance of the last 5% samples; **2)balanced MM**. we evenly divide the distance space into 100 bins and then calculate the average distance for each bin. Then balanced MM Dist denotes the average distance of all bins. The minimum value is almost 0, meaning some captions in the test split are similar to some of the training split. The maximum value is less than 0.25. We divide the whole distribution into 100 bins as the requirement of our proposed *balanced MM*. Most test data concentrate in interval [0.03, 0.07]. In supplementary material, we will show the data distribution and some examples of different rarenesses.

Results and Analysis. Table 4 shows the generalization ability of three different methods. As for the baseline model, we simply drop out the retrieval technique. From this table we can find that, with our proposed retrieval technique, ReMoDiffuse outperforms both the baseline model and state-of-the-art methods by a remarkable margin.

4.6. Qualitative Results

To illustrate the effectiveness of ReMoDiffuse, we provide a qualitative comparison between previous works and ReMoDiffuse. More examples are available in the project page. As shown in Figure 4, ReMoDiffuse stands out as the only approach that effectively conveys text descriptions that involve both action and path information. In contrast, Guo *et al.*’s method falls short in capturing path descriptions. MotionDiffuse performs well in action categories, but it lacks precision in providing path details. Meanwhile, MDM captures path information, but its generated actions are incorrect. In the examples evaluated, ReMoDiffuse demonstrates its capability to appropriately structure and present the content.

Table 4: **Evaluation of Generalization Ability.** All results are reported on the KIT testset. The best results are in **bold**.

Method	MM ↓	tail 5% MM ↓	balanced MM ↓
MotionDiffuse	2.958	5.928	4.285
Baseline	3.371	6.173	4.661
Ours	2.814	5.439	4.028
Δ	0.557	0.734	0.633

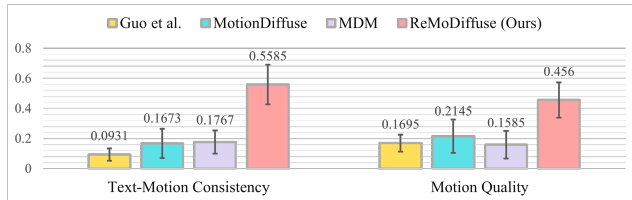


Figure 6: **The result of our user study.** The results indicate that our approach outperforms other methods significantly in both of text-motion consistency and motion quality.

4.7. User Study

We randomly selected 25 samples from the HumanML3D test set to compare Guo *et al.*, MotionDiffuse, MDM, and ReMoDiffuse. We collected 55 responses in total, and the results are shown in Figure 6. It can be seen that in most cases, testers believed that the motions generated by our method were more consistent with the given text descriptions, and the generated animations were more natural, which validates the superior quality of our proposed motion generation framework.

5. Conclusion

In this paper, we present ReMoDiffuse, a retrieval-augmented motion diffusion model for text-driven motion generation. Equipped with a multi-modality retrieval technique, the semantics-modulated attention mechanism, and a learnable condition mixture strategy, ReMoDiffuse efficiently explores and utilizes appropriate knowledge from an auxiliary database to refine the denoising process without expensive computation. Quantitative and qualitative experiments are conducted to demonstrate that ReMoDiffuse has achieved superior performance in text-driven motion generation, particularly for uncommon motions.

Social Impacts. This technique can be used to create fake media when combined with 3D avatar generation. The manipulated media conveys incidents that never truly happened and can serve malicious purposes.

Acknowledgement. This study is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOE-T2EP20221-0012), NTU NAP, and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

- [1] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728. IEEE, 2019. [2](#)
- [2] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Teach: Temporal action composition for 3d humans. *arXiv preprint arXiv:2209.04066*, 2022. [3](#)
- [3] Andreas Blattmann, Robin Rombach, Kaan Oktay, and Björn Ommer. Retrieval-augmented diffusion models. *arXiv preprint arXiv:2204.11824*, 2022. [2](#)
- [4] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022. [2](#), [6](#)
- [5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021. [2](#)
- [6] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1396–1406, 2021. [2](#)
- [7] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. [1](#), [2](#), [3](#), [5](#), [6](#), [7](#), [8](#)
- [8] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. *arXiv preprint arXiv:2207.01696*, 2022. [3](#)
- [9] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020. [7](#)
- [10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. [2](#), [4](#)
- [11] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. [3](#)
- [12] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5442–5451, 2019. [7](#)
- [13] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. [2](#)
- [14] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. [2](#)
- [15] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021. [6](#)
- [16] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. *arXiv preprint arXiv:2204.14109*, 2022. [3](#)
- [17] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. [1](#), [2](#), [7](#)
- [18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. [2](#), [3](#), [4](#), [9](#)
- [19] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. [2](#)
- [20] Jiawei Ren, Mingyuan Zhang, Cunjun Yu, and Ziwei Liu. Balanced mse for imbalanced visual regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7926–7935, 2022. [9](#)
- [21] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3531–3539, 2021. [5](#)
- [22] Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*, 2022. [2](#)
- [23] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. *arXiv preprint arXiv:2203.08063*, 2022. [2](#)
- [24] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. [2](#), [3](#), [4](#), [5](#), [7](#), [8](#)
- [25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. [3](#)
- [26] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. [3](#), [7](#), [8](#)
- [27] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. [1](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#)