Realistic Human Motion Generation with Cross-Diffusion Models

Zeping Ren[1*], Shaoli Huang[2†], Xiu Li[1†]

[1] Tsinghua Shenzhen International Graduate School
[2] Tencent AI Lab

Abstract. In this work, we introduce the Cross Human Motion Diffusion Model (CrossDiff[3]), a novel approach for generating high-quality human motion based on textual descriptions. Our method integrates 3D and 2D information using a shared transformer network within the training of the diffusion model, unifying motion noise into a single feature space. This enables cross-decoding of features into both 3D and 2D motion representations, regardless of their original dimension. The primary advantage of CrossDiff is its cross-diffusion mechanism, which allows the model to reverse either 2D or 3D noise into clean motion during training. This capability leverages the complementary information in both motion representations, capturing intricate human movement details often missed by models relying solely on 3D information. Consequently, CrossDiff effectively combines the strengths of both representations to generate more realistic motion sequences. In our experiments, our model demonstrates competitive state-of-the-art performance on text-to-motion benchmarks. Moreover, our method consistently provides enhanced motion generation quality, capturing complex full-body movement intricacies. Additionally, with a pre-trained model, our approach accommodates using in-the-wild 2D motion data without 3D motion ground truth during training to generate 3D motion, highlighting its potential for broader applications and efficient use of available data resources.

1 Introduction

In recent years, the field of human motion synthesis [7, 9, 10, 14, 15, 18, 41, 50] has witnessed significant advancements, primarily driven by the growing demand for high-quality, realistic motion generation in applications such as gaming, virtual reality, and robotics.

A crucial aspect in this research area is generating human motion based on textual descriptions, enabling contextually accurate and natural movements [40]. However, current methods [4, 8, 26, 40] predominantly rely on 3D motion information during training, leading to an inability to capture the full spectrum of intricacies associated with human motion. When using only 3D representation,

* Work done during an internship at Tencent AI Lab.
† Corresponding author.
[3] https://wonderno.github.io/CrossDiff-webpage/

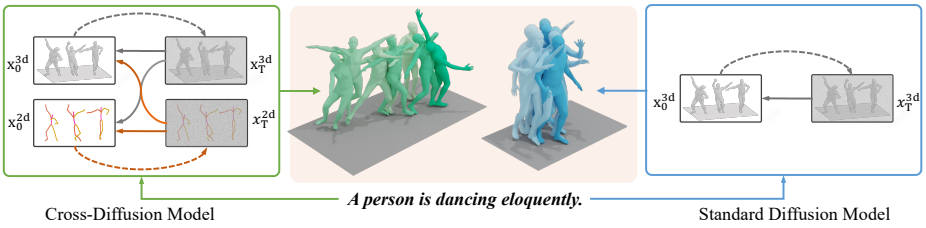

Fig. 1: Our method utilizing the cross-diffusion mechanism (Left) exhibits more full-body details compared to existing methods (Right).

the generation model may struggle to relate text semantics to some body part movements with very small movement variations compared to others, which can lead to overlooking important motion details. This is because the model might focus on more dominant or larger movements within the 3D space, leaving subtle nuances underrepresented. For example, when given a prompt such as "a person is dancing eloquently," as illustrated in Figure 1, the generated motion might lack vitality, display a limited range of movements, and contain minimal local motion details.

To effectively address the limitations and accurately capture the nuances of full-body movement, we introduce the Cross Human Motion Diffusion Model (CrossDiff). This innovative approach seamlessly integrates and leverages both 3D and 2D motion information to generate high-quality human motion sequences. The 2D data representation effectively illustrates the intricacies of human body movements from various viewing angle projections. Due to different view projections in 2D data, small body part movements can be magnified in certain projections, making them more noticeable and easier to capture. This helps the text-to-motion generation models to better associate text descriptions with a wider range of human body motion details, including subtle movements that might have been overlooked when relying solely on 3D representation.

As a result, incorporating 2D information with 3D enables the diffusion model to establish more connections between motion and text prompts, ultimately enhancing the motion synthesis process. The CrossDiff learning framework consists of two main components: unified encoding and cross-decoding. These components work together to achieve more precise and realistic motion synthesis. Furthermore, it is essential to transfer the knowledge acquired in the 2D domain to 3D motion, which leads to an overall improvement in the model's performance.

Unified encoding fuses motion noise from both 3D and 2D sources into a single feature space, facilitating cross-decoding of features into either 3D or 2D motion representations, regardless of their original dimension. The distinctive innovation of our approach stems from the cross-diffusion mechanism, which enables the model to transform 2D or 3D noise into clean motion during the training process. This capability allows the model to harness the complementary information present in both motion representations, effectively capturing

intricate details of human movement that are often missed by models relying exclusively on 3D data.

In experiments, we demonstrate our model achieves competitive state-of-the-art performance on several text-to-motion benchmarks, outperforming existing diffusion-based approaches that rely solely on 3D motion information during training. Furthermore, our method consistently delivers enhanced motion generation quality, capturing complex full-body movement intricacies essential for realistic motion synthesis. A notable advantage of our approach is its ability to utilize 2D motion data without necessitating 3D motion ground truth during training, enabling the generation of 3D motion. This feature underscores the potential of the CrossDiff model for a wide range of applications and efficient use of available data resources.

2 Related Work

2.1 Human Motion Generation

Human motion generation is the process of synthesizing human motion either unconditionally or conditioned by signals such as text, audio, or action labels. Early works [7, 16, 18, 24] treated this as a deterministic mapping problem, generating a single motion from a specific signal using neural networks. However, human motion is inherently stochastic, even under certain conditions, leading to the adoption of deep generative models in more recent research.

For instance, Dancing2music [14] employed GANs to generate motion under corresponding conditions. ACTOR [25] introduced a framework based on transformers [44] and VAEs, which, although designed for action-to-motion tasks, can be easily adapted for text-to-motion tasks as demonstrated in TEMOS [26]. Since text and audio are time-series data, natural language processing approaches are commonly used. Works by [6], [1], and [8] utilized GRU-based language models to process motion data along the time axis.

MotionCLIP [39] uses the shared text-image space learned by CLIP [28] to align the feature space of human motion with that of CLIP. MotionGPT [35, 36] directly treats motion as language and addresses the motion generation task as a translation problem. However, conditions like language and human motion differ significantly in terms of distribution and expression, making accurate alignment challenging.

To overcome this issue, T2M-GPT [52] and TM2T [9] encode motion using VQ-VAE [43] and generate motion embeddings with generative pretrained transformers. MotionDiffuse [53] is the first application of diffusion models in text-to-motion tasks. MDM [40] employs a simple diffusion framework to diffuse raw motion data, while MLD [4] encodes motion using a VAE model and diffuses it in the latent space. ReMoDiffuse [54] retrieves the motion related to the text to assist in motion generation. Meanwhile, Fg-T2M [45] utilizes a fine-grained method to extract neighborhood and overall semantic linguistic features. Although these methods attain success, they depend exclusively on 3D

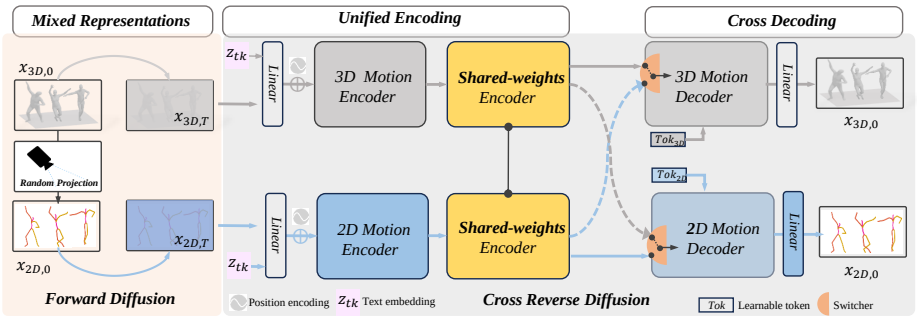

Fig. 2: Overview of our CrossDiff framework for generating human motion from textual descriptions. The framework incorporates both 3D and 2D motion data, using unified encoding and cross-decoding components to process mixed representations obtained from random projection.

motion data during training, which results in a failure to capture sufficient complexities associated with human motion. In contrast, our approach utilizes a cross-diffusion mechanism to leverage the complementary information found in both 2D and 3D motion representations.

2.2 Diffusion Models

Diffusion generative models [11, 35, 36], based on stochastic diffusion processes in Thermodynamics, involve a forward process where samples from the data distribution are progressively noised towards a Gaussian distribution and a reverse process where the model learns to denoise Gaussian samples. These models have achieved success in various domains, including image synthesis [29, 31, 32, 34, 42], video generation [11, 21, 49], adversarial attacks [23, 55], motion prediction [3, 46], music-to-dance synthesis [17, 41], and text-to-motion generation [4, 30, 40, 51, 53].

Classifier-Free Guidance [12] enables conditioned generation without additional model training, balancing fidelity and diversity. In the image domain, inpainting methods [5, 20, 37] iteratively incorporate known information into the diffusion process, maintaining constant image parts while learning to inpaint others. Similar approaches have been applied to motion editing [13, 33, 40]. Differing from these works, our work focuses on leveraging multiple data representations to enhance diffusion model learning.

3 Method

3.1 Overview

Given a textual description c, our objective is to generate multiple human motion sequences $x^{1:N} = \{x^i\}_{i=1}^{N}$, each with a length of N. As illustrated in Figure 2, our

method is carefully designed to efficiently incorporate both 3D and 2D motion data within the learning process of the diffusion model.

During the training phase, we first obtain mixed representations of the data from the provided 3D input using a random projection technique. Afterward, the 2D and 3D data representations are independently diffused and processed through our learning framework, CrossDiff, which primarily consists of unified encoding and cross-decoding components.

The unified encoding module maps both the 2D and 3D data into a shared feature space. These features are then passed through the cross-decoding component, resulting in the generation of two motion representations. These representations are subsequently employed for loss calculation and model learning. In the inference phase, our approach supports not only standard sampling but also mixture sampling.

Preliminary. Denoising diffusion probabilistic models (DDPM) [11] can iteratively eliminate noise from a gaussian distribution to approximate a true data distribution. This technique has had a significant impact on the field of generative research, including text-to-motion applications. In this study, we have adapted DDPM and trained a transformer-based model to gradually reduce noise and generate motion sequences.

Diffusion is modeled as a Markov noising process $\{x_t^{1:N}\}_{t=0}^{T}$ of T steps. For simplicity, we use x_t to denote $x_t^{1:N}$ in the following discussion. Starting with a motion sequence x_0 in original data distribution, the noising process can be described as

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

where $\alpha_t \in (0,1)$ is constant hyper-parameters. When α_T is small enough, we can approximate $x_T \in \mathcal{N}(0,\mathbf{I})$. The reverse process is to progressively denoise x_T from a gaussian distribution to obtain the clean motion x_0. Following [29,40], we predict the clean motion x_0 itself on textual condition c as $\hat{x}_0 = G(x_t, t, c)$. We apply the simple objective [11]

$$\mathcal{L}_{simple} = \mathbb{E}_{t\sim[1,T]}||x_0 - G(x_t,t,c)||_2^2. \tag{2}$$

3.2 Mixed Representations

As the naive diffusion model is trained only on one data distribution (3D poses), we have trained it on a mixture representation of 3D and 2D poses. To obtain 2D data that is closer to the real distribution, we randomly projected the 3D poses into 2D planes in four directions (front, left, right, and back). The 3D poses $x_{3D}^i \in \mathbb{R}^{d_{3D}}$ and 2D poses $x_{2D}^i \in \mathbb{R}^{d_{2D}}$ are represented respectively by d_{3D}-dimensional and d_{2D}-dimensional redundant features, respectively, as suggested by [8]. The pose x^i is defined by a tuple of (r, j^p, j^v, j^r, c^f), where $(r_{3D}, j_{3D}^p, j_{3D}^v, j_{3D}^r, c_{3D}^f)$ is identical to [8]. In addition, $r_{2D} \in \mathbb{R}^2$ represents 2D root velocity. $j_{2D}^p \in \mathbb{R}^{2j}, j_{2D}^v \in \mathbb{R}^{2j}$ and $j_{2D}^r \in \mathbb{R}^{2j}$ represent the local joints positions, velocities and rotations, respectively, with j denoting the number of joints besides the root. $c_{2D}^f \in \mathbb{R}^4$ is a set of binary features obtained by thresholding the heel and toe

joint velocities. Notably, the rotation representation is made up of the sine and cosine values of the angle.

3.3 Cross Motion Diffusion Model

Framework. Our pipeline is illustrated in Figure 2. CLIP [28] is a widely recognized text encoder, and we use it to encode the text prompt c. The encoded text feature and time-step t are projected into transformer dimension and summed together as the condition token z_{tk}. The 2D and 3D motion sequences are projected into the same dimension, concatenated with condition token z_{tk} in time axis and summed with a standard positional embedding. We aim to unify the two domain features in one space but it is too difficult for one linear layer. A straightforward idea is to encode via two separate encoders:

$$z^0_{3D} = \mathcal{E}_{3D}(x_{3D,t}, t, c), z^0_{2D} = \mathcal{E}_{2D}(x_{2D,t}, t, c), \tag{3}$$

where $\mathcal{E}_{3D/2D}(\cdot)$ are 3D/2D L_1-layer transformer encoders [44]. However, We find it more efficient to add another shared-weight encoder to extract shared feature:

$$\{z^i_{3D/2D}\}^{L_2}_{i=1} = \mathcal{E}_{share}(z^0_{3D/2D}), \tag{4}$$

where $\mathcal{E}_{share}(\cdot)$ is a shared-weight L_2-layer transformer encoder, and $\{z^i_{3D/2D}\}^{L_2}_{i=1}$ are the outputs of each shared-weight layer. The whole process is defined as unified encoding.

To output motion in two modality, we use independent L_2-layer transformer decoders [44] for 2D and 3D data. Starting from 2D/3D learnable token embeddings $Tok_{2D/3D}$, each decoder layer takes the output of the previous layer as queries and the output of same-level shared layer z^i as keys and values instead of the last layer. The starting point is to make decoder layers follow the extracting patterns of shared-weight layers rather than gaining deeper embeddings. Finally, a linear layer is added to map the features to the dimensions of the motions. This cross-decoding can be integrated as:

$$\hat{x}_{3D,0} = \mathcal{D}_{3D}(\{z^i_{3D/2D}\}^{L_2}_{i=1}), \hat{x}_{2D,0} = \mathcal{D}_{2D}(\{z^i_{3D/2D}\}^{L_2}_{i=1}), \tag{5}$$

where $\mathcal{D}_{3D/2D}$ are 3D/2D decoders including learnable tokens; and $\hat{x}_{3D/2D,0}$ are predicted clean 3D/2D motion sequences. In summary, with a pair of 3D motion and 2D projected motion, CrossDiff outputs four results via

$$\hat{x}_{iD\to jD,0} = G_{iD\to jD}(x_{iD,t}, t, c) = \mathcal{D}_{jD}(\mathcal{E}_{share}(\mathcal{E}_{iD}(x_{iD,t}, t, c))), \tag{6}$$

where $\hat{x}_{iD\to jD,0}$ are predicted j-dimension clean motion $\hat{x}_{jD,0}$ from i-dimension motion noise $x_{iD,t}$ with $i, j \in \{2, 3\}$.

Training. As mentioned in Section 3.1, we apply a simple objective (Eq. 2) for all outputs:

$$\mathcal{L}_{iD\to jD} = \mathbb{E}_{t\sim[1,T]}||x_{jD,0} - G_{iD\to jD}(x_{iD,t}, t, c)||^2_2. \tag{7}$$

We train our model in two stage. In stage I, CrossDiff is forced to learn the reverse process, motion characteristic of texts in both domains and the motion connection of two distribution via the loss

$$\mathcal{L}_{stageI} = \mathcal{L}_{3D\to 3D} + w_{23}\mathcal{L}_{2D\to 3D} + w_{32}\mathcal{L}_{3D\to 2D} + w_{22}\mathcal{L}_{2D\to 2D}, \tag{8}$$

where w_{23}, w_{32}, w_{22} are relative weights. In stage II, there is only a 3D generation loss:

$$\mathcal{L}_{stageII} = \mathcal{L}_{3D\to 3D}. \tag{9}$$

This helps the model focus on the 3D denoising process and eliminate the uncertainty of the 2D and 3D mapping relationship while retaining the knowledge of diverse motion features.

3.4 Mixture Sampling

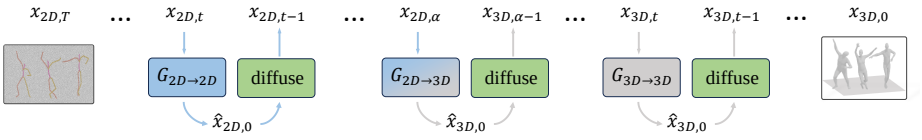

Fig. 3: Overview of Mixture Sampling. The original noise is sampled from a 2D gaussian distribution. From time-step T to α, CrossDiff predicts the clean 2D motion $\hat{x}_{2D,0}$ and diffuses it back to $x_{2D,t-1}$. In the remaining α steps, CrossDiff denoises in the 3D domain and finally obtains the clean 3D motion.

After training, one can sample a motion sequence conditioned on a text prompt in an iterative manner. The standard method [40] gradually anneals the 3D noise from a gaussian distribution, which we still use. We predict the clean sample $\hat{x}_{3D,0}$ and noise it back to $x_{3D,t-1}$ for T steps until $x_{3D,0}$.

Furthermore, utilizing the CrossDiff architecture, we propose a novel two-domain sampling approach. As shown in Figure 3, We first sample 2D gaussian noise which is then denoised with $G_{2D\to 2D}(x_{2D,t}, t, c)$ until time-step α. Next, we project the denoised 2D noise onto the 3D domain using $G_{2D\to 3D}(x_{2D,t}, t, c)$ and continue the denoising process with $G_{3D\to 3D}(x_{3D,t}, t, c)$ for the remaining α steps. Our experiments in supplementary demonstrate the difference between mixture sampling and the vanilla method.

3.5 Learning 3D Motion Generation from 2D Data

Given the complexity and cost associated with collecting high-quality 3D motion data, generating 3D motion from 2D motion data is an attractive alternative. Moreover, generating 3D motion from textual descriptions in an out-of-domain

scenario is approximately a zero-shot task. To achieve this, we first estimated 2D motion from text-related videos using an off-the-shelf model. We then utilized the pretrained model in stage I $G_{2D\to 3D}(x_{2D,t}, t, c)$ to generate corresponding 3D clean motion, with t set to 0 and c set to null condition $\emptyset$. A motion filter is applied to smooth the generated motion. We assume that 2D pose estimation is relatively precise, allowing the processed 2D/3D motion to serve as pseudo-labels for training. The model is fine-tuned with the same objective as stage I, but with different weight hyper-parameters. After training, our model can generate diverse motion according to out-of-domain textual descriptions using mixture sampling (Sec. 3.4).

During training with 2D motion estimated from videos, we encounter significant errors in root estimation due to the uncertainty of camera movement. To address this issue, we decouple the root information $r_{3D/2D}$ and generate it based on other pose features. Please refer to supplementary for more details.

4 Experiments

Our focus is on the text-to-motion generation task, and we conduct experiments on two standard datasets: HumanML3D [8] and KIT Motion-Language (KIT-ML) [27]. In Section 4.1, we introduce these standard datasets and the evaluation metrics used in our experiments. In Section 4.2, we present comparable quantitative results with state-of-the-art methods and outperform diffused methods on the HumanML3D dataset. We also use upper and lower body indices, combined with visualization effects, to illustrate our superior performance in capturing whole-body details. Moreover, we demonstrate that CrossDiff supports training with additional 2D motion sequences in Section 4.3. Finally, we discuss ablation studies in Section 4.4 to analyze the contributions of each component of our proposed method.

4.1 Datasets and Evaluation Metrics

Datasets. The HumanML3D [8] and KIT-ML datasets [27] are widely used in research. KIT-ML provides 6,353 textual descriptions for 3,911 motion sequences, which are all down-sampled to 12.5 FPS. HumanML3D is currently the largest 3D human motion dataset with textual descriptions. The motions are originally from two motion capture datasets, AMASS [22] and HumanAct12 [10], and are rescaled to 20 frames per second. It contains 14,616 motions annotated with 44,970 sequence-level textual descriptions. To enable fair comparisons with previous works [4,8,40], we use the redundant motion representation proposed by [8]. This representation involves re-targeting the joint positions to a default human skeleton template, setting the initial pose at the same position (X=0,Z=0) facing the Z+ direction, and including root global positions, local positions, joint rotations, joint velocities, and foot contact labels. We use the same motion representation for KIT-ML.

To obtain corresponding 2D motion, we utilize orthogonal projection to project 3D motion into 2D planes for four directions (front, left, right, and back). Then, we process the redundant 2D motion representation as described in Section 3.2.

UFC101 [38] is an influential action recognition dataset that contains paired videos and action labels. We obtain 2D joint positions using the ViTPose [48] model and process them into redundant 2D motion representation. We filter out fuzzy data and ultimately select 24 action labels, with each label related to 10-50 motion sequences. We then annotate around five textual descriptions for each label.

Evaluation Metrics. We compare our results with previous works using the same metrics as [8, 40]. These metrics involve evaluating quality with Frechet Inception Distance (**FID**), precision with **R-Precision** and Multi-modal Distance (**MM Dist**), and diversity with Diversity (**DIV**) and Multimodality (**MModality**). These measurements assess generated motion distribution, retrieval ranking accuracy, Euclidean distances between text and motion features, and variance across features and within a single text.

Besides using the evaluation model from [8], we introduce a new metric measuring the FID (Fréchet Inception Distance) of upper and lower body movements, denoted as **FID-U** and **FID-L**. This enables a fine-grained analysis of human motion and better comprehension of upper and lower body dynamics. We split joints into two groups using the root joint as a boundary and train separate evaluators, following a similar approach to [8]. This effectively evaluates generated motion quality for both body segments, offering a deeper understanding of human motion complexities and advancing research on new motion generation models.

Implementation Details. We use the AdamW [19] optimizer, setting the learning rate to 1e-4 and 1e-5 in stages I and II, respectively. Our model is trained for 4k epochs in stage I and 1k epochs in stage II, using 8 Tesla V100 GPUs with a mini-batch size of 32. The number of diffusion steps is set to 1K. The loss weights (w_{23}, w_{32}, w_{33}) is set to $(1, 1, 1)$ in stage I and $(0.1, 0.1, 1)$ in training on UFC101 datasets.

4.2 Comparisons on Text-to-motion

Comparative Analysis of Standard Metrics. In our evaluation, we test our models 20 times and compare their performance with existing state-of-the-art methods. These methods include Language2Pose [1], T2G [2], Hier [6], T2M [8], MotionDiffuse [53], Fg-T2M [45], MDM [40], T2M-GPT [52], MLD [4] and ReMoDiffuse [54]. As illustrated in Table 1, our model exhibits competitive performance when compared to these leading methods. However, it is important to note that the KIT-ML dataset primarily consists of "walk" movements and lacks intricate details. Consequently, this dataset does not present the same challenges

Table 1: Quantitative results on the HumanML3D and KIT-ML test set. The overall results on KIT-ML are shown on the **right**, while the results of both widely-used and newly-proposed metrics on HumanML3D are shown on the **left**. The red and blue colors indicate the best and second-best results, respectively.

Methods	HumanML3D							KIT-ML						
	R Precision (top 3)↑	FID↓	MM Dist↓	DIV→	MModality↑	**FID-U↓**	**FID-L↓**	R Precision (top 3)↑	FID↓	MM Dist↓	DIV→	MModality↑	**FID-U↓**	**FID-L↓**
Real	0.797	0.002	2.974	9.503	-	-	-	0.779	0.031	2.788	11.08	-	-	-
Language2Pose	0.486	11.02	5.296	7.676	-	-	-	0.483	6.545	5.147	9.073	-	-	-
T2G	0.345	7.664	6.030	6.409	-	-	-	0.338	12.12	6.964	9.334	-	-	-
Hier	0.552	6.532	5.012	8.332	-	-	-	0.531	5.203	4.986	9.563	-	-	-
T2M	0.740	1.067	3.340	9.188	2.090	-	-	0.693	2.770	3.401	10.91	1.482	-	-
MotionDiffuse	0.782	0.630	3.113	9.410	1.553	-	-	0.739	1.954	2.958	11.10	0.730	-	-
Fg-T2M	0.783	0.243	3.109	9.278	1.614	-	-	0.745	0.571	3.114	10.93	1.019	-	-
MDM	0.611	0.544	5.566	9.559	2.799	**0.825**	**0.840**	0.396	0.497	9.191	10.847	1.907	**0.925**	**0.973**
T2M-GPT	0.775	0.141	3.121	9.722	1.831	**0.145**	**0.607**	0.745	0.514	3.007	10.921	1.570	**0.602**	**0.715**
MLD	0.772	0.473	3.196	9.724	2.413	**0.541**	**0.553**	0.734	0.404	3.204	10.80	2.192	**0.563**	**0.772**
ReMoDiffuse	0.795	0.103	2.974	9.018	1.795	**0.125**	**0.565**	0.765	0.155	2.814	10.80	1.239	**0.205**	**0.644**
Ours	0.730	0.162	3.358	9.577	2.620	**0.118**	**0.281**	0.704	0.474	3.308	10.77	1.742	**0.434**	**0.625**

that our method is specifically designed to address. It indicates we can steadily generate high-quality and precise motion while pay attention to rich diversity. In other words, our generated results are not only consistent with the textual description, but also more expressive.

Comparative Analysis of Fine-grained Metrics. We compare the fine-grained metrics for our upper and lower body with those from four recent studies [4, 40, 52, 54]. As demonstrated in Table 1, our generated motion is more robust and detailed. Our low FID scores for both the upper and lower body indicate that our synthesized motion effectively captures full-body movement rather than focusing solely on specific semantic parts. In contrast, ReMoDiffuse and T2M-GPT achieves a low FID score for the upper body but a high score for the lower body. This suggests that their generation process exhibits unbalanced attention towards different body parts, primarily translating textual descriptions into upper body characteristics rather than capturing the entire body's motion.

Figure 4 displays qualitative comparisons with existing methods. Our method can "march" with arm swings, "wobble" using hands for balancing and alternate between defense and attack in a "fight". MDM exhibits a certain rigidity, while T2M-GPT and ReMoDiffuse appear to lack dynamism. For direct comparisons, please refer to the supplementary videos provided. We conducted a user study on motion performance, in which participants were asked two questions to assess the vitality and diversity of the motions. The results, presented in Figure 5(a), confirm our analysis. In summary, our method demonstrates a superior ability to interpret semantic information and generate more accurate and expressive motions.

4.3 Learning from 2D Data

After being trained on a 3D dataset, our model can learn 3D motion generation from 2D data. By fine-tuning the model with the UCF101 dataset [38], we ef-

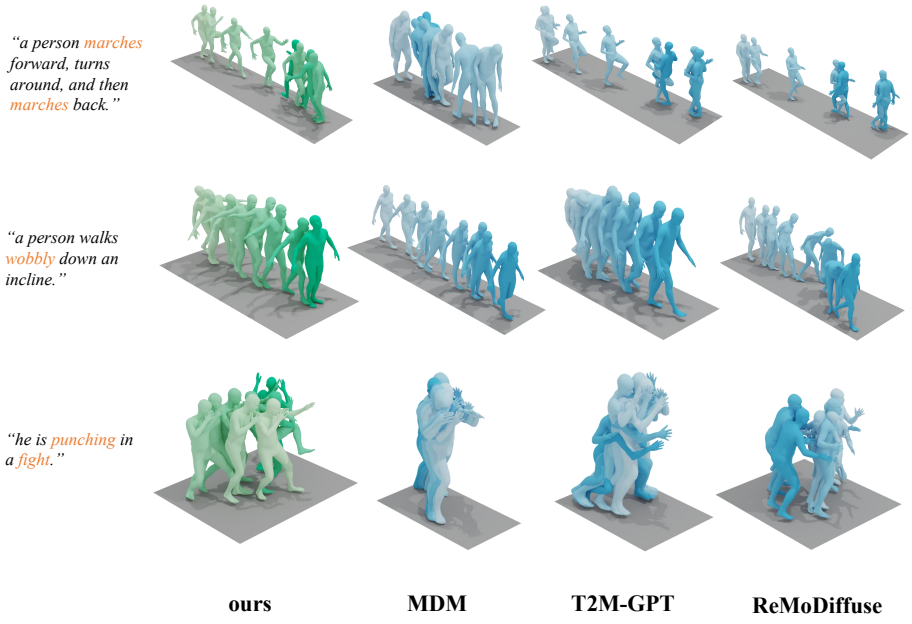

Fig. 4: Qualitative results on HumanML3D dataset. We compare our method with MDM [40], T2M-GPT [52] and MLD [4]. We find that our generated actions better convey the intended semantics.

fectively address a zero-shot problem arising from the absence of ground-truth 3D motion. Our sampling strategy reaches optimal performance when $\alpha = 1$. As depicted in Figure 6, the generated motions for various activities, such as pulling up, biking, table tennis, and baseball, are showcased alongside their textual prompts. Notably, pulling up and biking are entirely outside the HumanML3D motion domain. Although playing table tennis and baseball share similarities with actions like slapping or throwing in HumanML3D, the original model is unable to synthesize motion under those specific descriptions. However, after fine-tuning, our generated motion more closely resembles the given sports rather than merely mimicking the actions. Despite some activities being beyond the scope of the HumanML3D domain, our fine-tuned model successfully synthesizes specific motions by leveraging the weak 2D data. This demonstrates its remarkable adaptability and potential for efficient use of available motion data since 2D motion and related textual descriptions are easier to obtain than 3D motion.

4.4 Ablation Studies

Our model features separate pipelines for 2D and 3D inputs, allowing us to train solely on 3D motion sequences, which is an improvement over MDM [40].

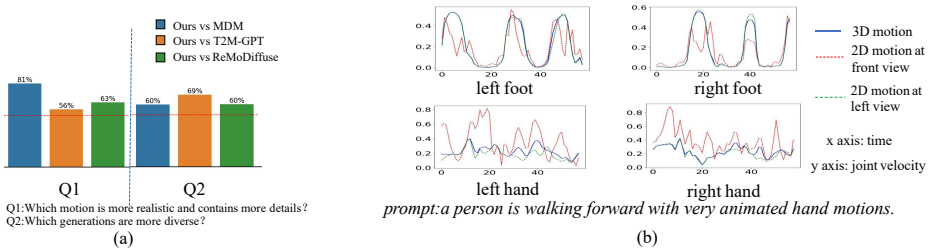

Fig. 5: (a)The result of the user study. (b) Difference between 3D and 2D motion data distribution. The time axis is represented on the x-axis, while the normalized joint velocity is represented on the y-axis. The 3D motion is represented by a blue full line, while the 2D motion is represented by red and green dashed lines, indicating the front and left view, respectively.

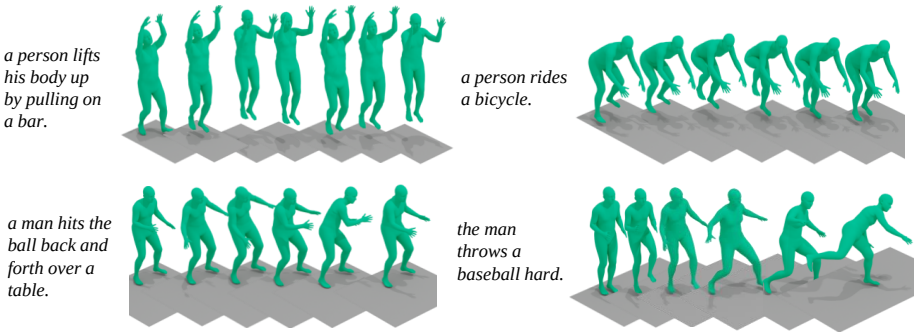

Fig. 6: Generating 3D movements without training on paired 3D motion and textual descriptions.

We investigate the impact of introducing 2D data on the performance of 3D generation and demonstrate the effectiveness of using a shared-weights encoder.

Why 2D motion help? To explain the benefits of 2D motions, we compared the distribution differences between 3D and 2D motion data. Hand and feet movements, which are primary indicators of motion, were visualized in both 3D and 2D levels, and their velocities were normalized along the joints dimension. In Figure 5(b), we can clearly see that around the 20th frame, the 2D velocity of the left hand in the front view reaches a higher value while the 3D velocity is quite low, indicating that hand movements in 2D are more prominent than in 3D. The results show that 2D motion captures different details from 3D motion, suggesting that the CrossDiff model can lead 3D motion to learn from the knowledge that 2D motion acquired from text prompts. Specifically, for the given sample, 2D motion might learn "animated hand motions" while 3D motion focuses only on "walking". 2D motion is an explicit feature that we artificially extract to

Table 2: Evaluation of our models with different settings on the HumanML3D dataset. **Bold** indicates best result. The symbol % indicates the percentage of data being used. From top to bottom, we present MDM as baselines, the impact of training with 2D representations, with(w/) or without(w/o) shared-weights encoder.

Methods	R Precision (top 3)↑	FID↓	MM Dist↓	DIV→
MDM	0.611	0.544	5.566	9.559
Ours	**0.730**	**0.162**	**3.358**	**9.577**
50% 3D	0.666	0.586	3.894	9.513
100% 3D	0.685	0.224	3.690	9.445
50% 3D + 100% 2D	0.672	0.422	3.708	9.345
100% 3D + 100% 2D	0.730	0.162	3.358	9.577
w/o shared-weights encoder	0.714	0.187	3.496	9.488
w/ shared-weights encoder	0.730	0.162	3.358	9.577
1 view(front)	0.722	0.186	3.467	9.798
1 view(left)	0.715	0.181	3.412	9.834
4 views	0.730	0.162	3.358	9.577
5 views	0.695	0.202	3.613	9.502

aid the model's learning, and this approach can help improve performance when dealing with arbitrary data.

Influence of 2D Representation Table 2 presents the results from four different experiment settings. The control groups of "100% 3D" and "100% 3D + 100% 2D" demonstrate that when training with paired 3D motion and text, projecting the 3D motion to 2D and building a connection between the 2D motion and text can help boost performance. The visualizations in Figure 1 further highlight the enhanced quality of our generated outputs. The control groups of "50% 3D" and "50% 3D + 100% 2D" prove that additional 2D data can also help improve performance. The additional 2D data indicates other 2D motion without ground truth 3D motion. The experiment in Section 4.3 shows learning 2D motion in the wild can also help with out-of-domain 3D motion learning. As we can see, the combined learning of 2D motion has great potential.

Shared-weights Encoder Without the shared-weights encoder, the model is a simple encoder-decoder framework with two modalities. However, we believe that this is not sufficient to fully fuse the 3D and 2D motion features. Inspired by [47], we found that when learning with data from two modalities, extracting separate and fused feature layer-by-layer is more efficient. The shared-weights encoder serves as a fusing module, while the 3D/2D motion decoder acts as a separate decoder module. The goal is to ensure that the decoder layers follow the same extraction patterns as the shared-weight layers, rather than simply gaining

deeper embeddings. The results presented in Table 2 demonstrate the efficiency of using a shared-weights encoder.

2D Representation from different views To evaluate the impact of different views, we conducted tests on four settings: a) only the front view, b) only the left view, c) four views including front, left, back, and right, and d) five views with an additional top view. When multiple views were used, the 3D motion was paired with a random view projection to formulate the loss. As shown in Table 2, it is reasonable to assume that four views contain more 2D information and outperform one view. Furthermore, the front view and left view do not have much distinction in terms of performance. However, the performance actually decreased with the additional top view. This could be due to the fact that the 2D information becomes more difficult to classify without the condition of the camera view. We did not pass along the camera information because it is difficult to estimate the camera view of in-the-wild videos, and injecting camera information could disrupt the structure of the text-to-2D model, making it difficult for the text-to-3D model to follow. In conclusion, the number of views serves as a practical hyperparameter that can be adjusted through enumeration experiments.

5 Conclusion

In conclusion, the Cross Human Motion Diffusion Model (CrossDiff) presents a promising advancement in the field of human motion synthesis by effectively integrating and leveraging both 3D and 2D motion information for high-quality motion generation. The unified encoding and cross-decoding components of the CrossDiff learning framework enable the capture of intricate details of human movement that are often overlooked by models relying solely on 3D data. Our experiments validate the superior performance of the proposed model on various text-to-motion benchmarks.

Despite its promising results, the CrossDiff model is not without potential weaknesses. One limitation may arise from the model's ability to generalize to unseen or rare motion patterns, especially when trained only on 2D motion data. Additionally, the computational complexity of the model might hinder its real-time application in certain scenarios, such as interactive gaming and virtual reality environments.

Future work could explore methods to enhance the model's generalization capabilities, such as incorporating unsupervised or semi-supervised learning techniques. To further advance our understanding, we propose the accumulation of a sizable 2D motion dataset coupled with relevant textual prompts, enabling the training of a unified motion generation model.

Acknowledgments. This research was partly supported by Shenzhen Key Laboratory of next generation interactive media innovative technology(Grant No: ZDSYS20210623092001004).

References

1. Ahuja, C., Morency, L.P.: Language2pose: Natural language grounded pose forecasting. In: 2019 International Conference on 3D Vision (3DV). pp. 719–728. IEEE (2019)
2. Bhattacharya, U., Rewkowski, N., Banerjee, A., Guhan, P., Bera, A., Manocha, D.: Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In: 2021 IEEE virtual reality and 3D user interfaces (VR). pp. 1–10. IEEE (2021)
3. Chen, L.H., Zhang, J., Li, Y., Pang, Y., Xia, X., Liu, T.: Humanmac: Masked motion completion for human motion prediction. arXiv preprint arXiv:2302.03665 (2023)
4. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18000–18010 (2023)
5. Chung, H., Sim, B., Ye, J.C.: Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12413–12422 (2022)
6. Ghosh, A., Cheema, N., Oguz, C., Theobalt, C., Slusallek, P.: Synthesis of compositional animations from textual descriptions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1396–1406 (2021)
7. Ghosh, P., Song, J., Aksan, E., Hilliges, O.: Learning human motion models for long-term predictions. In: 2017 International Conference on 3D Vision (3DV). pp. 458–466. IEEE (2017)
8. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5152–5161 (2022)
9. Guo, C., Zuo, X., Wang, S., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: European Conference on Computer Vision. pp. 580–597. Springer (2022)
10. Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2021–2029 (2020)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
12. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
13. Karunratanakul, K., Preechakul, K., Suwajanakorn, S., Tang, S.: Gmd: Controllable human motion synthesis via guided diffusion models. arXiv preprint arXiv:2305.12577 (2023)
14. Lee, H.Y., Yang, X., Liu, M.Y., Wang, T.C., Lu, Y.D., Yang, M.H., Kautz, J.: Dancing to music. Advances in neural information processing systems **32** (2019)
15. Li, J., Kang, D., Pei, W., Zhe, X., Zhang, Y., He, Z., Bao, L.: Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11293–11302 (2021)
16. Li, M., Chen, S., Zhao, Y., Zhang, Y., Wang, Y., Tian, Q.: Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 214–223 (2020)

17. Li, R., Zhao, J., Zhang, Y., Su, M., Ren, Z., Zhang, H., Tang, Y., Li, X.: Finedance: A fine-grained choreography dataset for 3d full body dance generation. arXiv preprint arXiv:2212.03741 (2023)
18. Li, Z., Zhou, Y., Xiao, S., He, C., Huang, Z., Li, H.: Auto-conditioned recurrent networks for extended complex human motion synthesis. arXiv preprint arXiv:1707.05363 (2017)
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
20. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11461–11471 (2022)
21. Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., Tan, T.: Videofusion: Decomposed diffusion models for high-quality video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10209–10218 (2023)
22. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5442–5451 (2019)
23. Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., Anandkumar, A.: Diffusion models for adversarial purification. arXiv preprint arXiv:2205.07460 (2022)
24. Pavllo, D., Grangier, D., Auli, M.: Quaternet: A quaternion-based recurrent model for human motion. arXiv preprint arXiv:1805.06485 (2018)
25. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3d human motion synthesis with transformer vae. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10985–10995 (2021)
26. Petrovich, M., Black, M.J., Varol, G.: Temos: Generating diverse human motions from textual descriptions. In: European Conference on Computer Vision. pp. 480–497. Springer (2022)
27. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. Big data **4**(4), 236–252 (2016)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
29. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
30. Ren, Z., Pan, Z., Zhou, X., Kang, L.: Diffusion motion: Generate text-guided 3d human motion by diffusion model. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
32. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
33. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. arXiv preprint arXiv:2303.01418 (2023)

34. Sinha, A., Song, J., Meng, C., Ermon, S.: D2c: Diffusion-decoding models for few-shot conditional generation. Advances in Neural Information Processing Systems **34**, 12533–12548 (2021)
35. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015)
36. Song, Y., Ermon, S.: Improved techniques for training score-based generative models. Advances in neural information processing systems **33**, 12438–12448 (2020)
37. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
38. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
39. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: Motionclip: Exposing human motion generation to clip space. In: European Conference on Computer Vision. pp. 358–374. Springer (2022)
40. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022)
41. Tseng, J., Castellon, R., Liu, K.: Edge: Editable dance generation from music. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 448–458 (2023)
42. Vahdat, A., Kreis, K., Kautz, J.: Score-based generative modeling in latent space. Advances in Neural Information Processing Systems **34**, 11287–11302 (2021)
43. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
45. Wang, Y., Leng, Z., Li, F.W., Wu, S.C., Liang, X.: Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22035–22044 (2023)
46. Wei, D., Sun, H., Li, B., Lu, J., Li, W., Sun, X., Hu, S.: Human joint kinematics diffusion-refinement for stochastic motion prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 6110–6118 (2023)
47. Xu, X., Wu, C., Rosenman, S., Lal, V., Che, W., Duan, N.: Bridgetower: Building bridges between encoders in vision-language representation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 10637–10647 (2023)
48. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose: Simple vision transformer baselines for human pose estimation. Advances in Neural Information Processing Systems **35**, 38571–38584 (2022)
49. Yang, R., Srivastava, P., Mandt, S.: Diffusion probabilistic modeling for video generation. arXiv preprint arXiv:2203.09481 (2022)
50. Yoon, Y., Cha, B., Lee, J.H., Jang, M., Lee, J., Kim, J., Lee, G.: Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Transactions on Graphics (TOG) **39**(6), 1–16 (2020)
51. Yuan, Y., Song, J., Iqbal, U., Vahdat, A., Kautz, J.: Physdiff: Physics-guided human motion diffusion model. arXiv preprint arXiv:2212.02500 (2022)
52. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: T2m-gpt: Generating human motion from textual descriptions with discrete representations. arXiv preprint arXiv:2301.06052 (2023)

53. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. arXiv preprint arXiv:2208.15001 (2022)
54. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: Remodiffuse: Retrieval-augmented motion diffusion model. arXiv preprint arXiv:2304.01116 (2023)
55. Zhuang, H., Zhang, Y., Liu, S.: A pilot study of query-free adversarial attack against stable diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2384–2391 (2023)