HumanMAC: Masked Motion Completion for Human Motion Prediction

https://lhchen.top/Human-MAC

Ling-Hao Chen[1*†] Jiawei Zhang[2†] Yewen Li[3] Yiren Pang[2] Xiaobo Xia[4*‡] Tongliang Liu[4]
[1]Tsinghua University [2]Xidian University
[3]Nanyang Technological University [4]The University of Sydney
{thu.lhchen}@gmail.com {zjw}@stu.xidian.edu.cn {yewen001}@e.ntu.edu.sg
{yrpang}@outlook.com {xiaoboxia.uni}@gmail.com {tongliang.liu}@sydney.edu.au

Abstract

Human motion prediction is a classical problem in computer vision and computer graphics, which has a wide range of practical applications. Previous effects achieve great empirical performance based on an encoding-decoding style. The methods of this style work by first encoding previous motions to latent representations and then decoding the latent representations into predicted motions. However, in practice, they are still unsatisfactory due to several issues, including complicated loss constraints, cumbersome training processes, and scarce switch of different categories of motions in prediction. In this paper, to address the above issues, we jump out of the foregoing style and propose a novel framework from a new perspective. Specifically, our framework works in a masked completion fashion. In the training stage, we learn a motion diffusion model that generates motions from random noise. In the inference stage, with a denoising procedure, we make motion prediction conditioning on observed motions to output more continuous and controllable predictions. The proposed framework enjoys promising algorithmic properties, which only needs one loss in optimization and is trained in an end-to-end manner. Additionally, it accomplishes the switch of different categories of motions effectively, which is significant in realistic tasks, e.g., the animation task. Comprehensive experiments on benchmarks confirm the superiority of the proposed framework. The project page is available at https://lhchen.top/Human-MAC.

1. Introduction

The problem of Human Motion Prediction (HMP) focuses on predicting possible future pose sequences from a sequence of observed motions [6, 16, 7, 20, 92, 84], which has a wide range of applications [83, 29, 90, 61, 45, 89, 87, 30, 99], *e.g.*, autonomous driving [50] and healthcare [71]. This problem is complicated and challenging since the predicted motions are always required to be *continuous*, *diverse*, and *realistic* simultaneously [14].

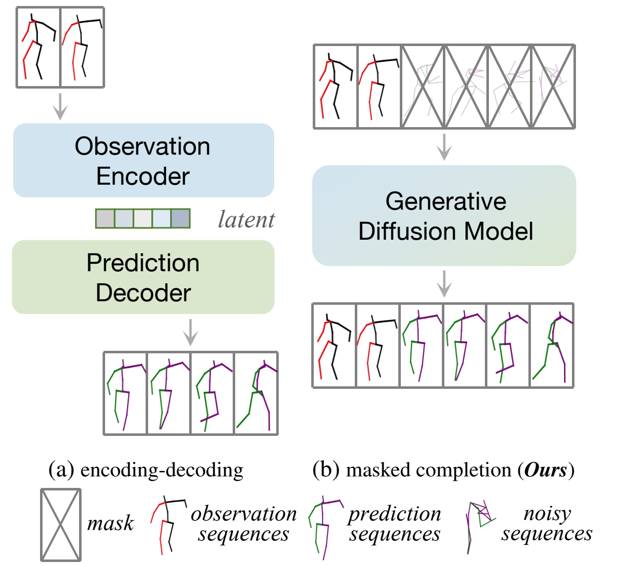


Figure 1: Comparison between the encoding-decoding fashion and masked motion completion. (a) The methods in the encoding-decoding fashion encode the observation into a latent explicitly and then decodes the latent into prediction results. (b) The proposed motion diffusion model generates motions from noise in the training stage. In the inference stage, it treats HMP as a masked completion task.

Prior state-of-the-art methods work in an encoding-decoding fashion to tackle the HMP problem, which conditions on previous motion frames and predicts unobserved motions [61, 41, 8, 14, 82]. Technically, these methods first encode the previous motion frames to latent representations explicitly and then decode the latent representations into prediction results (see Figure 1a) [54, 75, 5, 22, 2].

Although these methods enjoy the good performance in some scenarios, they are still unsatisfactory in practice. We detail the issues from three aspects. (1) Most state-of-the-art methods rely on multiple loss constraints for high-quality prediction results, *e.g.*, the average pairwise distance [4], final displacement error [19, 33], and adversarial loss [6, 20]. Consequently, they need carefully designed hyper-parameters to balance different loss constraints, which makes it laborious for method applica-

*This project is led by Ling-Hao Chen and Xiaobo Xia jointly.
†Equal contribution.
‡Corresponding author.

tions. (2) Previous state-of-the-art methods need multi-stage training [89, 5]. That is to say, the learning of the encoder/decoder and sampling in the latent space is performed in different stages. To make matters worse, complex pipelines always require an additional stage of engineering tuning. (3) For the methods, it is hard to realize the switch of different categories of motions, *e.g.*, switch from `WalkDog` to `Sitting`, which is pivotal for result diversity. The reason is that these methods are largely limited to observed motion sequences for prediction, which include few such switches.

In this paper, we get out of the encoding-decoding fashion and propose the **MA**sked motion **C**ompletion framework to handle the HMP problem (*aka* **HumanMAC**). To the best of our knowledge, this is the first time to tackle HMP from the perspective of masked completion. Specifically, in the training stage, we integrate observed and predicted motions, and learn a motion diffusion model to generate motions from random noise. In the inference stage, we predict possible future motions given observed motions. Note that traditional diffusion models [24, 65, 39, 48, 40, 85] have to make the prediction from random noise with a denoising procedure, which cannot make use of observed motions, resulting in uncontrollable results. We thereby propose a new method named DCT-Completion in the inference stage, which can employ observed motions for future motion prediction. Particularly, we add noise to observations to obtain a noisy spectrum of observed motions. In each denoising step of our method, we involve this noisy spectrum with a masking mechanism, where we combine the denoised and noisy spectrum with a mask. As the noisy spectrum is transformed from observations, the mechanism makes prediction motions conditioned on observed motions. Controllable predictions are produced accordingly (see Figure 1b).

Compared with the methods in the encoding-decoding fashion, HumanMAC enjoys the following great properties. (1) During training, there is only one loss function in the objective of HumanMAC. This avoids previous several hyper-parameters that balance different losses, and facilitates method applications. (2) HumanMAC is trained in an end-to-end manner, which is remarkably simpler to implement than multi-stage training. (3) HumanMAC can achieve more diverse prediction results that contain the switch of different categories of motions, even though the switch is rare or not presented in observed motion sequences. This is benefited from that we holistically model the whole sequence, *i.e.*, the integrated observed and predicted motions. The generated motion is therefore controlled to be continuous between the observed frames and predicted frames. Due to the continuity of motions, *e.g.*, from `Squatting` to `Standing Up` and from `Standing Still` to `Walking` for we humans, the trained model can naturally complete the switch of different categories of motions.

Before delving into details, we clearly summarize our contributions as follows:

- We carefully discuss the three issues of previous paradigms in human motion prediction, which can inspire follow-up research for algorithm improvement.
- To address the issues, we propose a novel framework from the new perspective of masked completion. The proposed method enjoys promising algorithmic properties that include only one loss function in the objective, an end-to-end training manner, and great motion switch ability.
- We conduct a series of experiments on benchmark datasets to justify our claims. In both qualitative and visualization comparisons with the state-of-the-art methods, our method achieves superior performance, which creates a simple and strong baseline for future research. Comprehensive ablation studies and discussions are also provided.

2. Related Work

2.1. Human Motion Prediction

Early in the research, traditional works [9, 47, 35, 68, 12, 23, 46, 64, 96, 64] try to predict motions in a deterministic way. In consideration that the prediction of motions is subjective, a series of works propose to produce diverse motions in prediction. At the present stages, the state-of-the-art methods of human motion prediction predict motions in an encoding-decoding way [61, 41, 8, 14, 82], and rely on carefully designed multiple loss constraints [4, 19, 33] to achieve the diversity and authenticity of human motions. Besides, most of the previous state-of-the-art methods need multi-stage training and cannot be trained in an end-to-end way. This property results in its dependency on the pre-training quality of the encoder and decoder. Therefore, we propose a human motion diffusion model with only one loss, which can be trained in an end-to-end way. Furthermore, thanks to the advantages of the completion method in the inference stage, we can switch between different motions for generating richer human motions, which is not achieved by previous works.

2.2. Denoising Diffusion Models

Denoising diffusion probabilistic models [24, 48, 65, 40, 11, 74, 67] are motivated by the second law of thermodynamics. The models try to learn a reverse diffusion process from random noise to the data distribution. Due to its dynamics, the generated results are both diverse and high-quality [52, 78, 76]. Therefore, diffusion models have also produced great impacts on image/video generation [15, 59, 88, 25], drug discovery [81], 3D reconstruction [80], and other research problems [57, 86, 49, 32, 21, 42, 91]. To take advantage of its good properties, we introduce the denoising diffusion model to predict human motions.

Before our work, MotionDiffuse [93] and MDM [70] are prior works that introduce diffusion models into the text-driven motion synthesis area. Tseng *et. al.* propose EDGE [72] to generate motions from music. Alexanderson *et. al.* propose a diffusion-based model [3] to synthesize motion driven by the audio. UDE [98] and MoFusion [13] propose to unify the audio-/text-driven motion generation task [36, 55, 26, 69, 63, 79, 54, 56, 34, 28] in one system. BeLFusion [5] tries to predict motions via the diffusion model in the latent space. However, it disentangles the model training into several stages, which is restricted by the pre-training quality of the encoder and decoder. Our end-to-end training framework can avoid this problem.

3. Preliminaries

3.1. Problem Formulation

We note the observed sequence of the H-frame motion as $\mathbf{x}^{(1:H)} = \left[\mathbf{x}^{(1)}; \mathbf{x}^{(2)}; \ldots; \mathbf{x}^{(H)}\right] \in \mathbb{R}^{H\times 3J}$, where $\mathbf{x}^{(h)} \in \mathbb{R}^{3J}$ is the coordinates of each joints at the frame h and J is the number of joints. Given the observed motion $\mathbf{x}^{(1:H)}$, the objective of the Human Motion Prediction (HMP) problem is to predict the following F motions $\mathbf{x}^{(H+1:H+F)} = \left[\mathbf{x}^{(H+1)}; \mathbf{x}^{(H+2)}; \ldots; \mathbf{x}^{(H+F)}\right] \in \mathbb{R}^{F\times 3J}$.

3.2. Discrete Cosine Transform

The work [46] proposed the Discrete Cosine Transform (DCT) to predict and generate human motions. The DCT operation extracts both current and periodic temporal properties from the motion sequence, which is beneficial for obtaining continuous motions. Therefore, staying precedent [46], we train our model with DCT.

Technically, given a $(H+F)$-frame motion sequence $\mathbf{x} \in \mathbb{R}^{(H+F)\times 3J}$, we project the sequence into the DCT domain via the $\texttt{DCT}(\cdot)$ operation:

$$\mathbf{y} = \texttt{DCT}(\mathbf{x}) = \boldsymbol{D}\mathbf{x}, \tag{1}$$

where $\boldsymbol{D} \in \mathbb{R}^{(H+F)\times(H+F)}$ is the predefined DCT basis [46], and $\mathbf{y} \in \mathbb{R}^{(H+F)\times 3J}$ is the DCT coefficients. As the DCT operation is an orthogonal transform, we can recover the motion sequence from the DCT domain via an inversed Discrete Cosine Transform ($\texttt{iDCT}(\cdot)$) operation:

$$\mathbf{x} = \texttt{iDCT}(\mathbf{y}) = \boldsymbol{D}^\top\mathbf{y}. \tag{2}$$

Due to the smoothness property of human motions [46], we simplify to perform the DCT and iDCT operation by selecting the first L rows of $\boldsymbol{D}$ and $\boldsymbol{D}^\top$ (noted as $\boldsymbol{D}_L, \boldsymbol{D}_L^\top \in \mathbb{R}^{L\times(H+F)}$), following $\mathbf{c} \approx \boldsymbol{D}_L\mathbf{x}$ and $\mathbf{x} \approx \boldsymbol{D}_L^\top\mathbf{c}$. The simplified form of the DCT/iDCT operation can reduce the computational cost by discarding the high-frequency components [46]. In the sequel, if there is no confusion, we will replace $\boldsymbol{D}$ with $\boldsymbol{D}_L$.

4. Methodology

4.1. Model Training

In the training stage, we conduct the DCT operation discussed above on the full motion $\mathbf{x} \in \mathbb{R}^{(H+F)\times 3J}$ for the first L frequency components to obtain the spectrum $\mathbf{y}_0 \in \mathbb{R}^{L\times 3J}$, where $\mathbf{y}_0 = \mathbf{y}$ in Eq. (1). The noisy DCT spectrum $\mathbf{y}_t$ at the timestep t can be calculated by the reparameterization trick:

$$\mathbf{y}_t = \sqrt{\bar{\alpha}_t}\mathbf{y}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \tag{3}$$

where $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$, $\alpha_i \in [0,1]$ are pre-defined variance parameters, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. There are T timesteps in total. In this paper, for noise prediction, we employ a simple network parameterized by $\boldsymbol{\theta}$, whose architecture is shown in Appendix A. We name the noise prediction network as $\texttt{TransLinear}$. At the timestep t, the network outputs predicted noise as $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{y}_t, t)$. Afterward, we optimize the parameters $\boldsymbol{\theta}$ with the noise prediction loss:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{\epsilon},t}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{y}_t, t)\right\|^2\right]. \tag{4}$$

It is worth noting that the loss in Eq. (4) is the *only loss* in our HumanMAC during training. Additionally, HumanMAC is trained in an end-to-end manner by minimizing the loss $\mathcal{L}$ directly. We provide an algorithm flow of model training in Algorithm 1.

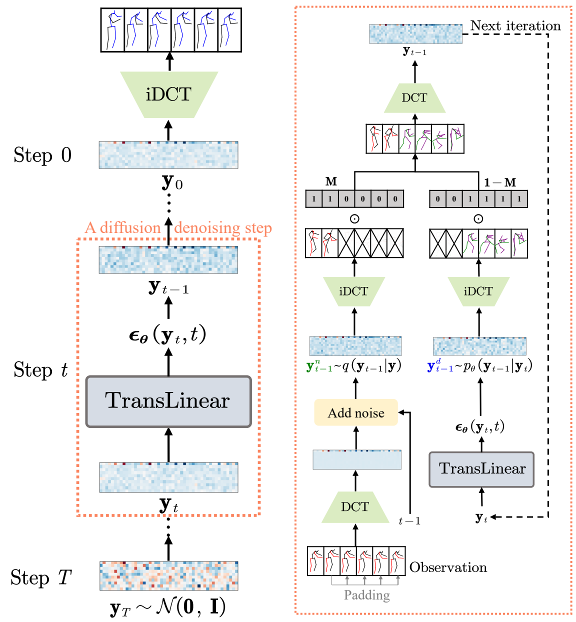
4.2. DCT-Completion in Inference

In the inference stage, given observed motions, we need to use the trained model to make predicted motions. In prior diffusion-based works, with a trained diffusion model, motions can be generated via T denoising steps. We provide the illustration of this generation in Figure 2a. However, the prior procedure cannot make use of observed motions, which makes predicted motions uncontrollable. To address the issue, we propose a new completion algorithm without model retraining, named DCT-Completion, which can complete prediction motions with observation motions. We discuss DCT-Completion as follows.

Add noise into observation. For the left branch in Figure 2b, we pad the observation into a full sequence with the final observation frame and project it to the DCT domain, noted as $\mathbf{y}$. Then, we add noise on $\mathbf{y}$ to obtain noisy spectrum of the observation at the timestep $t-1$:

$$\mathbf{y}_{t-1}^n = \sqrt{\bar{\alpha}_{t-1}}\mathbf{y} + \sqrt{1-\bar{\alpha}_{t-1}}\mathbf{z}, \tag{5}$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Note that $\mathbf{y}_{t-1}^n \sim q(\mathbf{y}_{t-1} \mid \mathbf{y})$.



(a) Prior diffusion generation with T steps.

(b) A diffusion DCT-Completion step.

Figure 2: Comparison between the vanilla denoising procedure and the proposed DCT-Completion.

Denoised prediction. For the right denoising branch shown in Figure 2b, it aims to denoise the noisy prediction spectrum $\mathbf{y}_{t-1}$ from $\mathbf{y}_t$:

$$\mathbf{y}_{t-1}^d = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{y}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}\left(\mathbf{y}_t, t\right)\right) + \sigma_t \mathbf{z}, \tag{6}$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t = 1$; else $\mathbf{z} = \mathbf{0}$. Note that $\mathbf{y}_{t-1}^d \sim p_{\boldsymbol{\theta}}(\mathbf{y}_{t-1} \mid \mathbf{y}_t)$.

Masked completion. In a trained diffusion model, the noisy observation spectrum $\mathbf{y}_{t-1}^n$ and denoised prediction spectrum $\mathbf{y}_{t-1}^d$ are approximately the same distribution [24], *i.e.*, $q(\mathbf{y}_{t-1} \mid \mathbf{y}) \approx p_{\boldsymbol{\theta}}(\mathbf{y}_{t-1} \mid \mathbf{y}_t)$. Therefore, we project the denoised and noisy spectrum to the temporal domain via the iDCT operation and then combine them together with a masking mechanism:

$$\mathbf{y}_{t-1} = \mathtt{DCT}[\mathbf{M} \odot \mathtt{iDCT}(\mathbf{y}_{t-1}^n) + (\mathbf{1} - \mathbf{M}) \odot \mathtt{iDCT}(\mathbf{y}_{t-1}^d)], \tag{7}$$

where $\mathbf{M} = [\underbrace{1, 1, \ldots, 1}_{H-\text{dim}}, \underbrace{0, 0, \ldots 0}_{F-\text{dim}}]^\top$ denotes the mask of the prediction, and $\odot$ denotes the Hadamard product. With the masked noisy observation guidance, we can complete the prediction motion in each reverse diffusion step.

For a convenient understanding of readers, we provide the algorithm flow of inference in Algorithm 2.

Motion switch. For the motion switch ability, different from the vanilla HMP task, we provide the final additional M-frame target motions in the inference stage to

Algorithm 1: Training procedure of HumanMAC

Input: motion $\mathbf{x} \in \mathbb{R}^{(H+F)\times 3J}$, noising steps T, the initialized noise prediction network $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$, maximum iterations I_{max}.
Output: the noise prediction network $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$.
for $I = 0, 1, \ldots, I_{\text{max}}$ **do**
  $\mathbf{y}_0 = \mathtt{DCT}(\mathbf{x}) \sim p(\mathbf{y}_0)$;
  $t = \mathtt{Uniform}(\{1, 2, \cdots, T\})$;
  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$;
  $\boldsymbol{\theta} = \boldsymbol{\theta} - \nabla_{\boldsymbol{\theta}} \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}\left(\sqrt{\bar{\alpha}_t}\mathbf{y}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t\right)\right\|^2$;
return the noise prediction network $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$.

Algorithm 2: Inference procedure of HumanMAC

Input: observed motion $\mathbf{x}^{(1:H)} \in \mathbb{R}^{H\times 3J}$, the mask of the observation $\mathbf{M}$, noising steps T, the trained noise prediction network $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$.
Output: competed motion $\mathbf{x} \in \mathbb{R}^{(H+F)\times 3J}$.
$\mathbf{y}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$;
$\mathbf{x} := \mathtt{Pad}(\mathbf{x}) \in \mathbb{R}^{(H+F)\times 3J}$ // observation padding;
$\mathbf{y} = \mathtt{DCT}(\mathbf{x}) \sim p(\mathbf{y})$;
for $t \in T, T-1, \ldots, 1$ **do**
  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$;
  $\mathbf{y}_{t-1}^n = \sqrt{\bar{\alpha}_{t-1}}\mathbf{y} + \sqrt{1-\bar{\alpha}_{t-1}}\boldsymbol{z}$;
  $\mathbf{y}_{t-1}^d = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{y}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}\left(\mathbf{y}_t, t\right)\right) + \sigma_t\mathbf{z}$;
  $\mathbf{y}_{t-1} = \mathtt{DCT}[\mathbf{M} \odot \mathtt{iDCT}(\mathbf{y}_{t-1}^n) + \mathtt{iDCT}((\mathbf{1} - \mathbf{M}) \odot \mathbf{y}_{t-1}^d])$;
return $\mathtt{iDCT}(\mathbf{y}_0)$.

obtain the controllable prediction. Besides, we will mask the motions except for observation and target motions, *i.e.*, $\mathbf{M} = [\underbrace{1, 1, \ldots, 1}_{H-\text{dim}}, \underbrace{0, 0, \ldots 0}_{(F-M)-\text{dim}}, \underbrace{1, 1, \ldots 1}_{M-\text{dim}}]^\top$. As we model the whole sequence directly, the generated motion is controlled to be continuous among the observed, predicted, and target frames. Besides, benefiting from the continuity of motions, the trained model can naturally complete the switch of different categories of motions, *e.g.*, from `Sitting` to `Standing`.

5. Experiments

5.1. Experimental Setup

Datasets. We evaluate our model on two human motion datasets, *i.e.*, Human3.6M [27] and HumanEva-I [62]. For fair comparisons, we follow the same dataset partition settings with previous work. Additionally, we evaluate our method on the CMU-MoCap and AMASS dataset, shown in Section 5.7.

	One-Stage	# Loss	Human3.6M					HumanEva-I				
			APD↑	ADE↓	FDE↓	MMADE↓	MMFDE↓	APD↑	ADE↓	FDE↓	MMADE↓	MMFDE↓
acLSTM [97]	✓	1	0.000	0.789	1.126	0.849	1.139	0.000	0.429	0.541	0.530	0.608
DeLi GAN [20]	✓	1	6.509	0.483	0.534	0.520	0.545	2.177	0.306	0.322	0.385	0.371
MT-VAE [83]	✓	3	0.403	0.457	0.595	0.716	0.883	0.021	0.345	0.403	0.518	0.577
BoM [7]	✓	1	6.265	0.448	0.533	0.514	0.544	2.846	0.271	0.279	0.373	0.351
DSF [90]	✗	2	9.330	0.493	0.592	0.550	0.599	4.538	0.273	0.290	0.364	0.340
DLow[89]	✗	3	11.741	0.425	0.518	0.495	0.531	4.855	0.233	0.244	0.343	0.331
GSPS [45]	✗	5	14.757	0.389	0.496	0.476	0.525	5.825	0.233	0.244	0.343	0.331
MOJO [95]	✗	3	12.579	0.412	0.514	0.497	0.538	4.181	0.234	0.244	0.369	0.347
BeLFusion [5]	✗	4	7.602	0.372	**0.474**	**0.473**	**0.507**	-	-	-	-	-
DivSamp [14]	✗	3	15.310	0.370	0.485	0.475	0.516	6.109	0.220	0.234	**0.342**	**0.316**
MotionDiff [75]	✗	4	**15.353**	0.411	0.509	0.508	0.536	5.931	0.232	0.236	0.352	0.320
HumanMAC	✓	1	6.301	**0.369**	0.480	0.509	0.545	**6.554**	**0.209**	**0.223**	**0.342**	0.335

Table 1: Experimental results of quantitative results. Bolded numbers denote the state-of-the-art results. The lower is the better for all metrics except for APD. The symbol '-' indicates that the results are not reported in the baseline work.

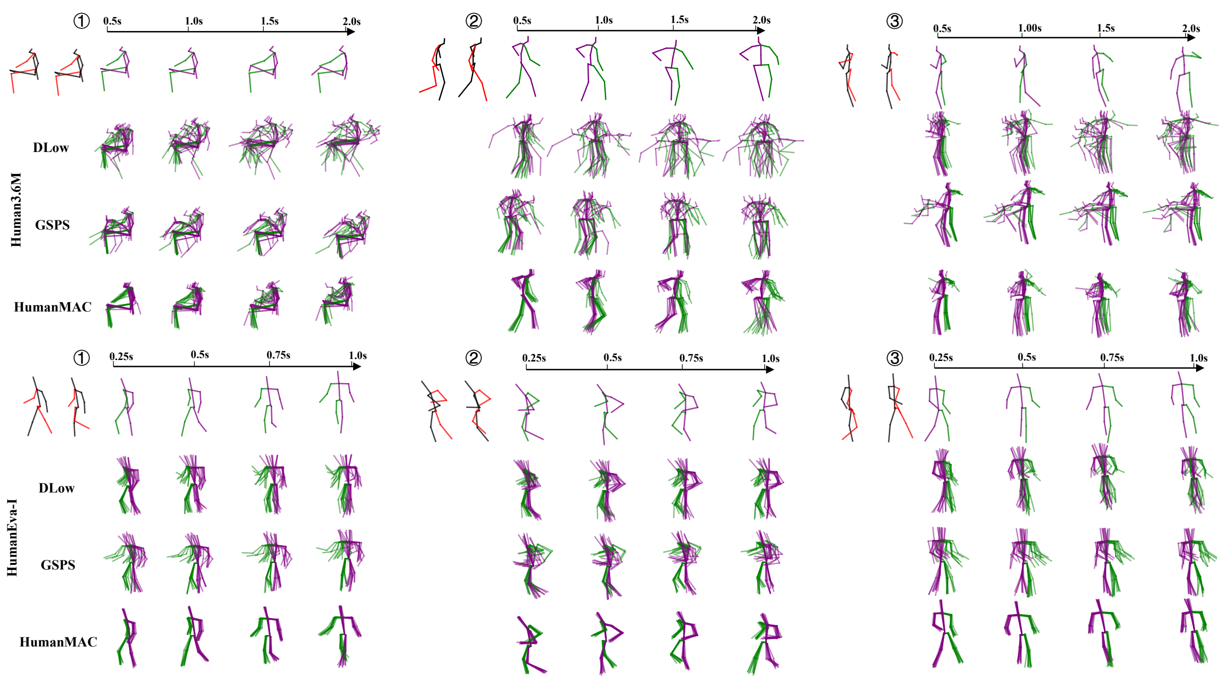


Figure 3: Visualization results. The first and the second row denote the predicted results of Human3.6M [27] and HumanEva-I [62] datasets. The *red-black* skeletons and *green-purple* skeletons denote the observed and predicted motions respectively.

Metrics. Following the previous work [89], we use five metrics to evaluate our model. (1) Average Pairwise Distance (*APD*) is the L2 distance between all motion examples. It is used for measuring result diversity. (2) Average Displacement Error (*ADE*) is defined as the smallest average L2 distance between the ground truth and predicted motion. It measures the accuracy of the whole sequence. (3) Final Displacement Error (*FDE*) is the L2 distance between the prediction results and ground truth in the last prediction frame. (4) Multi-Modal-ADE (*MMADE*) is the multi-modal version of ADE, whose ground truth future motions are grouped by similar observations. (5) Multi-Modal-FDE (*MMFDE*), similarly, is the multi-modal version of FDE.

Baselines. In quantitative comparison, we compare our method with several state-of-the-art works, including acLSTM [97], DeLi GAN [20], MT-VAE [83], BoM [7], DSF [90], DLow [89], GSPS [45], MOJO [95], BeLFusion [5], DivSamp [14], and MotionDiff [75]. In visualization comparison, the competitors are DLow and GSPS.

Implementation details. We train HumanMAC on both datasets with a 1000-step diffusion model and sample with a 100-step DDIM [65]. The Cosine scheduler [48] is exploited for variance scheduling in our model. The noise prediction network contains 8-layer TransLinear blocks for Human3.6M and 4-layer TransLinear blocks for HumanEva-I. Experiments are conducted with PyTorch [51] and one NVIDIA Tesla A100-80GB GPU. Please refer to Appendix B for more details.

5.2. Comparison with the State-of-the-Arts

Quantitative results. Experimental results are provided in Table 1. As can be seen, for Human3.6M, our method achieves state-of-the-art results on both ADE and FDE metrics, which shows the plausibility of our results. Although we cannot achieve as good results as the state-of-the-art baselines in terms of diversity, as shown in Figure 3, our results are more reasonable (*fewer failure examples compared with baselines*). Therefore, we did not achieve the superiority of APD as an artificial indicator by introducing failure examples. For HumanEva-I, we present the competitive results on the MMFDE metric and the state-of-the-art results on other metrics. Our prediction results on HumanEva-I achieve a great trade-off between diversity and authenticity.

Visualization results. We present our qualitative analysis results via visualizing motion sequences in Figure 3. We compare our method with DLow [89] and GSPS [45]. The first and the second row denotes three predicted results of Human3.6M [27] and HumanEva-I [62] datasets respectively. In each frame, we show 10 results of the predicted motion stochastically.

Although the DLow and GSPS methods seem more diverse, both of them generate some failure cases with large distances to ground truths, which are not reasonable motions. For instance, in the second example from Human3.6M, some generated result of DLow is floating in the air. Besides, the third generated result in Human3.6M of GSPS starts with a sudden bend and holds for a few seconds, which does not satisfy the physical constraints of the human center of gravity. We claim that the diversity exhibited by artificial metrics and the comparison of multiple predicted results are unreasonable, resulting in a large number of failure cases. By contrast, the diversity of predicted motions by our method is more reasonable than baseline methods. We provide more empirical evidence in Appendix C.

5.3. Motion Switch Ability

The proposed HumanMAC enables the switch ability between different categories of motions by applying free-training DCT-Completion. The ability helps produce some motion combinations that are rare or unseen during training. We visualize the results of the switched motion in Figure 4. It shows that various observed motions switch to target motions smoothly. In the case where the directions of the human body facing to differ significantly, HumanMAC accomplishes the switch from one direction to the other smoothly and naturally. For transferring between two motions with *a large distribution gap* (like from `Sitting` to `Standing`), the motion of the upper and lower bodies changes in a natural way. The motion switch ability shows that our model can generate sparsely distributed motion data (out-of-distribution [37, 77], unseen, or long-tailed data). We provide more examples and analysis in Appendix D.

L	Human3.6M			HumanEva-I		
	APD↑	ADE↓	FDE↓	APD↑	ADE↓	FDE↓
5	6.048	0.388	0.501	6.444	0.318	0.340
10	**6.508**	0.389	0.500	**6.554**	**0.209**	0.223
20	6.301	**0.369**	**0.480**	5.985	**0.209**	**0.220**

Table 2: Experimental results of the ablation study on L.

# Layer	Human3.6M			HumanEva-I		
	APD↑	ADE↓	FDE↓	APD↑	ADE↓	FDE↓
2	6.654	0.425	0.542	5.827	0.216	0.230
4	**7.020**	0.496	0.525	**6.554**	0.209	0.223
6	6.944	0.388	0.499	6.287	**0.202**	**0.216**
8	6.301	**0.369**	**0.480**	5.587	0.209	0.221
10	6.512	0.373	0.483	3.350	0.208	0.221

Table 3: Experimental results of the ablation study on the number of layers in the noise prediction network.

5.4. Part-body Controllable Prediction

Due to the flexibility of the mask mechanism in the DCT-Completion algorithm (Eq. (7)), we can complete the F-frame unknown part-body motion by providing the H-frame observation and final F-frame specified part-body motions. As shown in Figure 5, we fix the lower body motions and predict reasonable upper body motions. We present 10 examples of the end poses for all motions. It shows that our method can predict diverse upper body motions when keeping the limited error of the lower body. Besides, it presents some hard cases, *e.g.*, large-angle turning and standing-bending transitions. The baselines GSPS and DLow only can achieve upper-/lower-body control since they need to disentangle the upper and lower body motions. Besides, with disentangled motions, these methods need to retrain the model to achieve the part-body controllable ability. However, our method does not rely on model re-training and can achieve controllable prediction of any part of the body. For more controllable part-body prediction results, please refer to Appendix E.

5.5. Ablation Study

We conduct comprehensive ablation studies on our methods, including (1) the value of L in DCT/iDCT; (2) the design of the noise prediction network; (3) the settings of the diffusion model. We detail them as follows. The gray shadows in tables that are highlighted in all tables indicate our choice for our method.

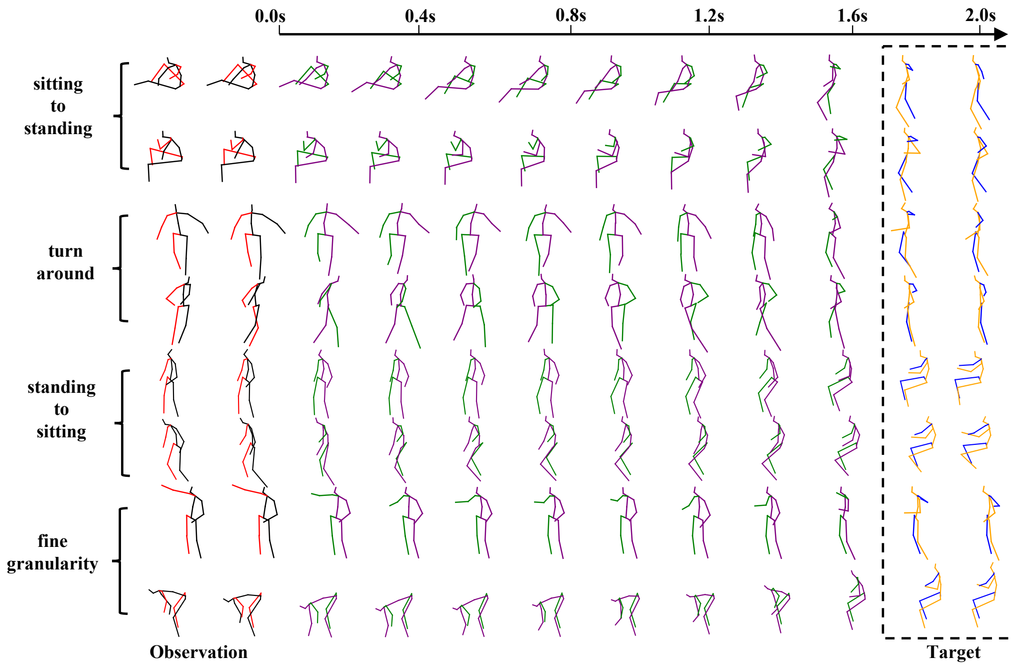


Figure 4: Motion switch results. Visualization of motion switch using DCT-Completion from the Human3.6M dataset. The left side shows randomly sampled motions switched to a standing pose. The right side shows various motions switched to sitting. The whole switching process is fairly smooth. The *red-black* skeletons and *green-purple* skeletons denote the observed and predicted motions respectively.

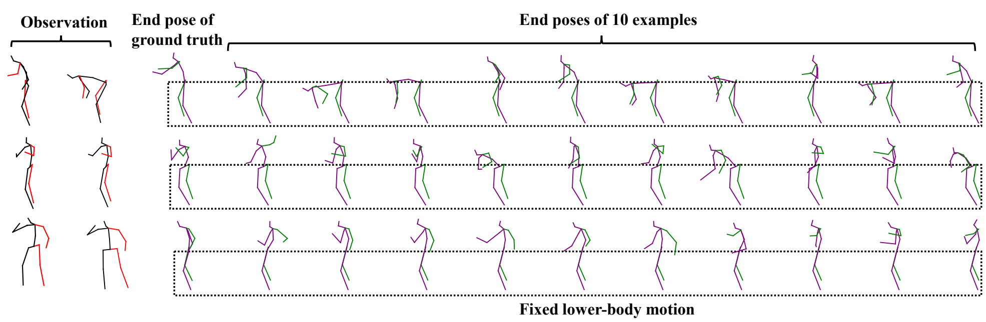


Figure 5: Results about controllable motion prediction, where we fix the lower body to predict the motions of the upper body. The *red-black* skeletons and *green-purple* skeletons denote the observed and predicted motions respectively.

Scheduler	Human3.6M			HumanEva-I		
	APD↑	ADE↓	FDE↓	APD↑	ADE↓	FDE↓
Linear	6.231	0.379	0.488	6.320	0.211	0.233
Sqrt	**8.276**	0.582	0.741	**7.636**	0.408	0.490
Cosine	6.301	**0.369**	**0.480**	6.554	**0.209**	**0.223**

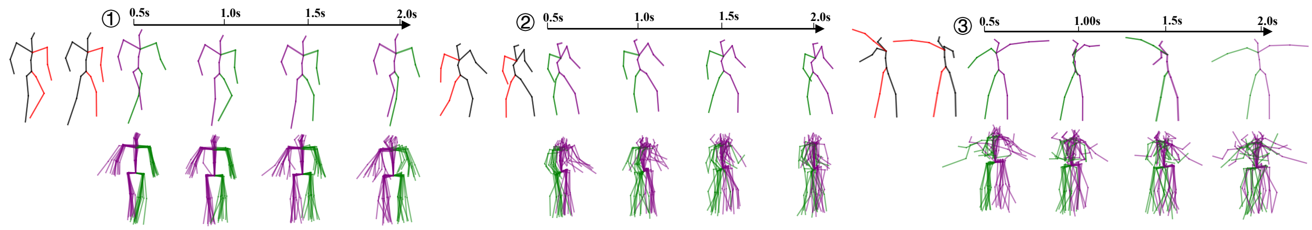
Table 4: Experimental results of the ablation study on different schedulers in the diffusion model.

Value of L. As discussed in Section 3.2, we approximate the DCT and iDCT operation by selecting the first L rows of $\boldsymbol{D}$ as $\boldsymbol{D}_L$ to improve computing efficiency. In Table 2, we show the reasonable choice of different L. The reasonable choices of L are 10 and 20 for Human3.6M [27] and HumanEva-I [62] datasets respectively.

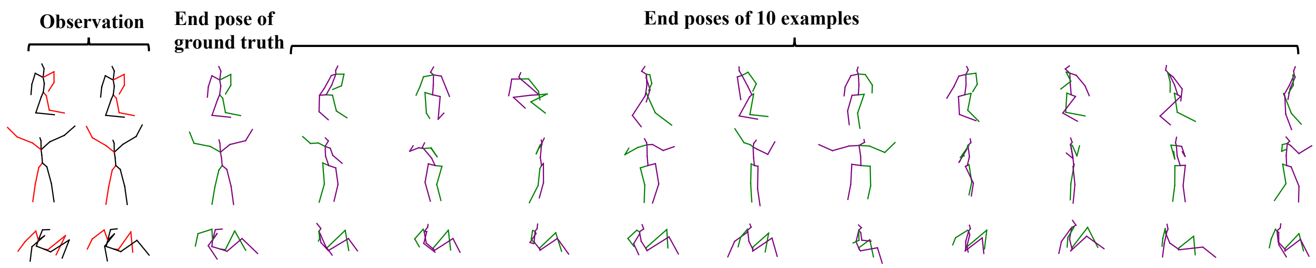
Design of the noise prediction network. In Table 3, we show the reasonable choice of the different numbers of layers in the noise prediction network. The reasonable choices of layers are 8 and 4 for Human3.6M [27] and HumanEva-I [62] respectively. We present other ablation studies on the network design in Appendix F, *e.g.*, the influence of skip connections in the noise prediction network.

# Noising steps	# DDIM steps	Human3.6M		
		APD↑	ADE↓	FDE↓
100	10	**6.492**	0.385	0.493
1000	100	6.301	**0.369**	**0.480**
		HumanEva-I		
100	10	6.199	0.220	0.244
1000	100	**6.554**	**0.209**	**0.223**

Table 5: Experimental results of the ablation study on the number of diffusion steps.



(a) Motion sequence visualization. The first row is the ground truth and the second row is a sample of 10 predictions.



(b) End pose visualization. We visualize the end pose of the prediction from three random examples.

Figure 6: Visualization results of zero-shot adaption ability on the AMASS dataset. The *red-black* skeletons and *green-purple* skeletons denote the observed and predicted motions respectively.

Settings of the diffusion model. We ablate the impacts of both the length of the denoising steps and different diffusion variance schedulers such as “Linear”, “Cosine”, and “Sqrt” on experimental results. As shown in Table 4, the Cosine scheduler is the best choice for scheduling. It is because the noise cannot be increased too much in the beginning steps of the diffusion process to prevent the information from being lost too quickly. The steps of noising and sampling are set to 1000 and 100 respectively in our main results. From the quantitative results in Table 5, reducing both noising steps and sampling steps by a factor of 10 damages the authenticity of generated motions. In addition, this reduction of steps results in jerky motions in visualization results. Moreover, as shown in Tables 2, 3, 4, and 5, the diversity and the authenticity of motions are a pair of tradeoffs. Although some settings are with a higher APD metric, there are many failure cases, which we do not compare in the main results.

5.6. Zero-shot Motion Prediction

We present the generalization ability of the HumanMAC method via zero-shot adaption experiments. Here, we train the HumanMAC model on the Human3.6M dataset and predict motions via the given AMASS dataset [43] observed motions. The experimental setting is shown in Appendix B. The visualization results are shown in Figure 6. As shown in Figure 6, our zero-shot motion prediction results on Human3.6M dataset are reasonable and diverse, which shows the good generalization ability of the HumanMAC framework. For instance, as shown in the first row of Figure 6b, the predicted motion `Kneeling` is vivid, which is unseen in Human3.6M. We provide more results in Appendix H.

5.7. Evaluation on Both CMU-MoCap and AMASS Datasets

We compare our HumanMAC framework with VAE-based and diffusion-based models on both CMU-MoCap [1] and AMASS[43] datasets. All parameters are well-tuned for all baselines. As shown in Table 6, the HumanMAC method presents superior performance on both datasets, which shows the strong prediction ability for the mask-completion mechanism.

5.8. Ablation Study on the DCT Design

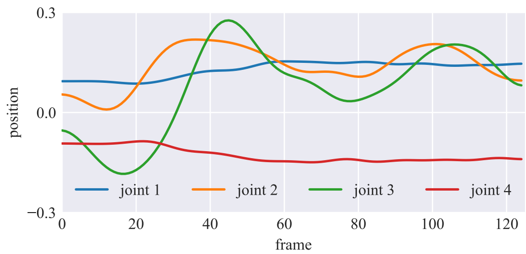
[66] shows the regular periodicity of motions, which inspires us to model human motion with DCT. Furthermore, we show joint motion curves (X-coordinate) of four key points in Figure 7, where the model w/o the DCT suffers from extreme jittering. As shown in Table 7, the DCT makes the prediction more accurate. In conclusion, the DCT design explicitly models the low- and high-frequency signals, making the prediction more accurate and smooth.

5.9. Long Time Series Prediction

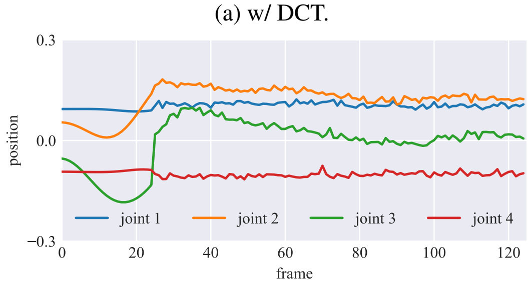
The long-time motion prediction can be implemented in an auto-regressive way. Technically, we can treat the final H-frame predictions as observations and predict further motions recurrently. In Figure 8, we present joint motion curves (X-coordinate). Baselines tend to predict *over-smoothing and almost static* motions. In contrast, ours are *more diverse*, which benefits from our DCT modeling.

	CMU-MoCap					AMASS				
	APD↑	ADE↓	FDE↓	MMADE↓	MMFDE↓	APD↑	ADE↓	FDE↓	MMADE↓	MMFDE↓
DLow	15.206	0.516	0.544	0.544	0.547	13.170	0.590	0.612	0.618	0.617
GSPS	12.995	0.456	0.512	0.522	0.534	12.465	0.563	0.613	0.609	0.633
DivSamp	**20.473**	0.441	0.501	**0.508**	0.525	**24.724**	0.564	0.647	0.623	0.667
MoDiff	19.174	0.464	0.523	0.531	0.542	-	-	-	-	-
BeLFusion	-	-	-	-	-	9.376	0.513	0.560	**0.569**	**0.591**
HumanMAC	8.021	**0.433**	**0.467**	0.512	**0.505**	9.321	**0.511**	**0.554**	0.593	**0.591**

Table 6: Evaluation on CMU-MoCap and AMASS datasets. The symbol '-' indicates that the results are not reported in the baseline work.



(a) w/ DCT.



(b) w/o DCT.

Figure 7: Predicted motion curves on w/ or w/o DCT.

	APD↑	ADE↓	FDE↓	MMADE↓	MMFDE↓
w/o DCT	**7.191**	0.444	0.521	0.521	0.550
w/ DCT	6.301	**0.369**	**0.480**	**0.509**	**0.545**

Table 7: Ablation on w/ or w/o DCT process (Human3.6M).

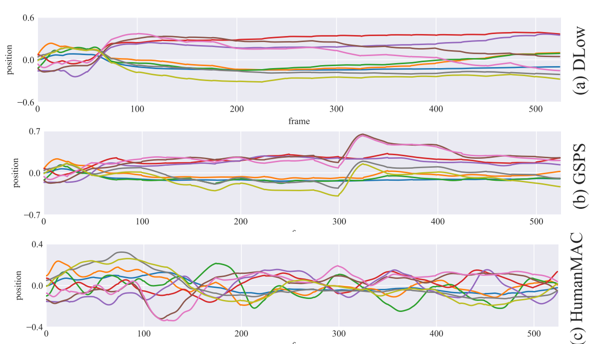


Figure 8: Joints motion curves on the X-coordinate.

6. Limitation and Discussion

HumanMAC is a diffusion-style method for the HMP task. In our diffusion model, the proposed method needs 100 steps DDIM sampling steps, which makes it limited to being a real-time system. Therefore, we will try more methods, *e.g.*, DPM-solver++ [39, 40] to reduce the sampling steps in future work. Besides, although our method is relatively simple at the present stage, we will try to further simplify the network architecture and completion algorithm. Moreover, we will test our method on larger-scale datasets [10, 43, 18, 17, 44, 58, 38] for human motion prediction to build a generalized HumanMAC model and explore its zero-shot transfer ability on the HMP task.

As analyzed in the experiment section, the previous metric for judging diversity is questionable, because of those failure cases. Therefore, in the HMP task, a more reasonable metric for diversity needs to be raised imminently. This is also what we are interested in.

7. Conclusion

In this paper, we target human motion prediction, which is a practical and challenging problem. We precisely reveal the issues of previous methods that work in an encoding-decoding fashion. Different from them, we propose a new learning framework for human motion prediction, which operates in a masked completion fashion. The proposed framework enjoys excellent algorithmic properties and achieves state-of-the-art performance in both quantitative and qualitative comparisons. We believe that this work can offer a new perspective to the research community of human motion prediction and inspire follow-up research.

Acknowledgement

We would like to thank Mr. Yu-Kun Zhou from Xidian University, and Mr. Wenhao Yang from Nanjing University for providing significant suggestions and technical support. Ling-Hao Chen is partially supported by International Digital Economy Academy (IDEA). Xiaobo Xia is supported by Australian Research Council Project DE190101473 and Google PhD Fellowship. Tongliang Liu is partially supported by the following Australian Research Council projects: FT220100318, DP220102121, LP220100527, LP220200949, and IC190100031.

References

[1] Cmu motion capture database. http://mocap.cs.cmu.edu, 2009. 8

[2] Vida Adeli, Mahsa Ehsanpour, Ian Reid, Juan Carlos Niebles, Silvio Savarese, Ehsan Adeli, and Hamid Rezatofighi. Tripod: Human trajectory and pose dynamics forecasting in the wild. In *ICCV*, 2021. 1

[3] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *arXiv preprint arXiv:2211.09707*, 2022. 3

[4] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *CVPR*, pages 5223–5232, 2020. 1, 2

[5] German Barquero, Sergio Escalera, and Cristina Palmero. Belfusion: Latent diffusion for behavior-driven human motion prediction. *arXiv preprint arXiv:2211.14304*, 2022. 1, 2, 3, 5, 15

[6] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *CVPR Workshops*, pages 1418–1427, 2018. 1

[7] Apratim Bhattacharyya, Bernt Schiele, and Mario Fritz. Accurate and diverse sampling of sequences based on a "best of many" sample objective. In *CVPR*, pages 8485–8493, 2018. 1, 5, 15

[8] Andreas Blattmann, Timo Milbich, Michael Dorkenwald, and Bjorn Ommer. Behavior-driven synthesis of human dynamics. In *CVPR*, pages 12236–12246, 2021. 1, 2

[9] Arij Bouazizi, Adrian Holzbock, Ulrich Kressel, Klaus Dietmayer, and Vasileios Belagiannis. Motionmixer: Mlp-based 3d human body pose forecasting. In *IJCAI*, 2022. 2

[10] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *ECCV*, pages 557–577. Springer, 2022. 9

[11] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. *arXiv preprint arXiv:2211.09788*, 2022. 2

[12] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *CVPR*, 2020. 2

[13] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. *arXiv preprint arXiv:2212.04495*, 2022. 3

[14] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Diverse human motion prediction via gumbel-softmax sampling from an auxiliary space. In *ACM MM*, pages 5162–5171, 2022. 1, 2, 5, 15

[15] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, pages 8780–8794, 2021. 3

[16] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *CVPR*, pages 4346–4354, 2015. 1

[17] 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 *CVPR*, pages 5152–5161, 2022. 9

[18] 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 *ACM MM*, pages 2021–2029, 2020. 9

[19] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *CVPR*, pages 2255–2264, 2018. 1, 2

[20] Swaminathan Gurumurthy, Ravi Kiran Sarvadevabhatla, and R Venkatesh Babu. Deligan: Generative adversarial networks for diverse and limited data. In *CVPR*, pages 166–174, 2017. 1, 5, 15

[21] Xizewen Han, Huangjie Zheng, and Mingyuan Zhou. Card: Classification and regression diffusion models. In *NeurIPS*, 2022. 3

[22] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *ICCV*, 2021. 1

[23] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7134–7143, 2019. 2

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 2, 4

[25] Jonathan Ho, Tim Salimans, Alexey A Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *NeurIPS*, 2022. 3

[26] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: zero-shot text-driven generation and animation of 3d avatars. *TOG*, 41(4):1–19, 2022. 3

[27] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE TPAMI*, 36(7):1325–1339, 2013. 4, 5, 6, 7, 13

[28] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *arXiv preprint arXiv:2306.14795*, 2023. 3

[29] Xu Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. Human-art: A versatile human-centric dataset bridging natural and artificial scenes. *arXiv preprint arXiv:2303.02760*, 2023. 1

[30] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. Humansd: A native skeleton-guided diffusion model for human image generation. *arXiv preprint arXiv:2304.04269*, 2023. 1

[31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 13

[32] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *ICLR*, 2020. 3

[33] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *CVPR*, pages 336–345, 2017. 1, 2

[34] Buyu Li, Yongchi Zhao, Shi Zhelun, and Lu Sheng. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *AAAI*, 2022. 3

[35] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *CVPR*, pages 5226–5234, 2018. 2

[36] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, 2021. 3

[37] Yewen Li, Chaojie Wang, Xiaobo Xia, Tongliang Liu, Bo An, et al. Out-of-distribution detection with an adaptive likelihood ratio on informative hierarchical vae. *NeurIPS*, 2022. 6

[38] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *arXiv preprint arXiv:2307.00818*, 2023. 9

[39] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *NeurIPS*, 2022. 2, 9

[40] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 2, 9

[41] Thomas Lucas, Fabien Baradel, Philippe Weinzaepfel, and Grégory Rogez. Posegpt: quantization-based 3d human motion generation and forecasting. In *ECCV*, 2022. 1, 2

[42] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022. 3

[43] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, 2019. 8, 9

[44] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. The kit whole-body human motion database. In *ICAR*. IEEE, 2015. 9

[45] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Generating smooth pose sequences for diverse human motion prediction. In *ICCV*, 2021. 1, 5, 6, 13, 15

[46] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *ICCV*, pages 9489–9497, 2019. 2, 3

[47] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *CVPR*, pages 2891–2900, 2017. 2

[48] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 2, 5, 13

[49] Alexander Quinn 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. In *ICML*, 2022. 3

[50] Brian Paden, Michal Čáp, Sze Zheng Yong, Dmitry Yershov, and Emilio Frazzoli. A survey of motion planning and control techniques for self-driving urban vehicles. *IEEE Transactions on Intelligent Vehicles*, 1(1):33–55, 2016. 1

[51] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 6

[52] Tim Pearce, Tabish Rashid, Anssi Kanervisto, Dave Bignell, Mingfei Sun, Raluca Georgescu, Sergio Valcarcel Macua, Shan Zheng Tan, Ida Momennejad, Katja Hofmann, et al. Imitating human behaviour with diffusion models. In *ICLR*, 2023. 2

[53] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018. 13

[54] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *ICCV*, pages 10985–10995, 2021. 1, 3

[55] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *ECCV*, pages 480–497. Springer, 2022. 3

[56] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 3

[57] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *ICML*, 2021. 3

[58] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: bodies, action and behavior with english labels. In *CVPR*, pages 722–731, 2021. 9

[59] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

[60] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 13, 15

[61] Tim Salzmann, Marco Pavone, and Markus Ryll. Motron: Multimodal probabilistic human motion forecasting. In *CVPR*, pages 6457–6466, 2022. 1, 2

[62] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 2010. 4, 5, 6, 7, 13

[63] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *CVPR*, pages 11050–11059, 2022. 3

[64] Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. Space-time-separable graph convolutional network for pose forecasting. In *ICCV*, 2021. 2

[65] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2020. 2, 5

[66] Sebastian Starke, Ian Mason, and Taku Komura. Deepphase: Periodic autoencoders for learning motion phase manifolds. *TOG*, 41(4):1–13, 2022. 8

[67] Boyuan Sun, Yuqi Yang, Le Zhang, Ming-Ming Cheng, and Qibin Hou. Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. *arXiv preprint arXiv:2306.04300*, 2023. 2

[68] Yongyi Tang, Lin Ma, Wei Liu, and Wei-Shi Zheng. Long-term human motion prediction by modeling motion context and enhancing motion dynamic. In *IJCAI*, 2018. 2

[69] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *ECCV*, 2022. 3

[70] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Amit H Bermano, and Daniel Cohen-Or. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 3

[71] Nikolaus F Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision*, 2(5):2–2, 2002. 1

[72] Jonathan Tseng, Rodrigo Castellon, and C Karen Liu. Edge: Editable dance generation from music. *CVPR*, 2022. 3

[73] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 13

[74] Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-gan: Training gans with diffusion. In *ICLR*, 2023. 2

[75] Dong Wei, Huaijiang Sun, Bin Li, Jianfeng Lu, Weiqing Li, Xiaoning Sun, and Shengxiang Hu. Human joint kinematics diffusion-refinement for stochastic motion prediction. In *AAAI*, 2023. 1, 5, 15

[76] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *CVPR*, 2022. 2

[77] Xiaobo Xia, Jiale Liu, Jun Yu, Xu Shen, Bo Han, and Tongliang Liu. Moderate coreset: A universal method of data selection for real-world data-efficient deep learning. In *ICLR*, 2022. 6

[78] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. In *ICLR*, 2022. 2

[79] Chen Xin, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. Executing your commands via motion diffusion in latent space. *arXiv*, 2022. 3

[80] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. *arXiv preprint arXiv:2212.14704*, 2022. 3

[81] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. In *ICLR*. 3

[82] Sirui Xu, Yu-Xiong Wang, and Liang-Yan Gui. Diverse human motion prediction guided by multi-level spatial-temporal anchors. In *ECCV*, 2022. 1, 2

[83] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *ECCV*, pages 265–281, 2018. 1, 5, 15

[84] Zigeng Yan, Di-Hua Zhai, and Yuanqing Xia. Dms-gcn: dynamic mutiscale spatiotemporal graph convolutional networks for human motion prediction. *arXiv preprint arXiv:2112.10365*, 2021. 1

[85] Jie Yang, Bingliang Li, Fengyu Yang, Ailing Zeng, Lei Zhang, and Ruimao Zhang. Boosting human-object interaction detection with text-to-image diffusion model. *arXiv preprint arXiv:2305.12252*, 2023. 2

[86] Jie Yang, Chaoqun Wang, Zhen Li, Junle Wang, and Ruimao Zhang. Semantic human parsing via scalable semantic transfer over multiple label domains. In *CVPR*, 2023. 3

[87] Jie Yang, Ailing Zeng, Shilong Li, Feng nd Liu, Ruimao Zhang, and Lei Zhang. Neural interactive keypoint detection. In *ICCV*, 2023. 1

[88] Zhentao Yu, Zixin Yin, Deyu Zhou, Duomin Wang, Finn Wong, and Baoyuan Wang. Talking head generation with probabilistic audio-to-visual diffusion priors. *arXiv:2212.04248*, 2022. 3

[89] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *ECCV*, 2020. 1, 2, 5, 6, 13, 15, 19

[90] Ye Yuan and Kris M Kitani. Diverse trajectory forecasting with determinantal point processes. In *ICLR*, 2019. 1, 5, 15

[91] 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. *arXiv preprint arXiv:2301.06052*, 2023. 3

[92] Jason Y Zhang, Panna Felsen, Angjoo Kanazawa, and Jitendra Malik. Predicting 3d human dynamics from video. In *ICCV*, pages 7114–7123, 2019. 1

[93] 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. 3

[94] Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023. 13

[95] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *CVPR*, pages 3372–3382, 2021. 5, 15

[96] Chongyang Zhong, Lei Hu, Zihao Zhang, Yongjing Ye, and Shihong Xia. Spatio-temporal gating-adjacency gcn for human motion prediction. In *CVPR*, 2022. 2

[97] Yi Zhou, Zimo Li, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. In *ICLR*, 2018. 5, 15

[98] Zixiang Zhou and Baoyuan Wang. Ude: A unified driving engine for human motion generation. *CVPR*, 2023. 3

[99] Jieming Zhu, Quanyu Dai, Liangcai Su, Rong Ma, Jinyang Liu, Guohao Cai, Xi Xiao, and Rui Zhang. Bars: Towards open benchmarking for recommender systems. In *ACM SIGIR*, pages 2912–2923, 2022. 1