RD-Diff: RLTransformer-based Diffusion Model with Diversity-Inducing Modulator for Human Motion Prediction

Haosong Zhang[1,2], Mei Chee Leong[1], Liyuan Li[1], and Weisi Lin[2]

[1] Institute for Infocomm Research (I^2R), A*STAR, Singapore
[2] Nanyang Technological University, Singapore
haosong001@e.ntu.edu.sg

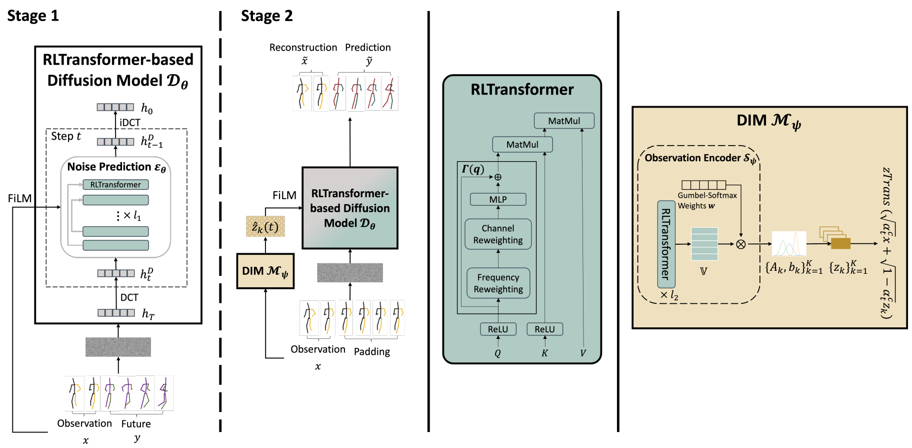
Abstract. Human Motion Prediction (HMP) is crucial for human-robot collaboration, surveillance, and autonomous driving applications. Recently, diffusion models have shown promising progress due to their ease of training and realistic generation capabilities. To enhance both accuracy and diversity of the diffusion model in HMP, we present **RD-Diff: RLTransformer-based Diffusion model with Diversity-inducing modulator**. First, to improve transformer's effectiveness on the frequency representation of human motion transformed by Discrete Cosine Transform (DCT), we introduce a novel Regulated Linear Transformer (RLTransformer) with a specially designed linear-attention mechanism. Next, to further enhance the performance, we propose a Diversity-Inducing Modulator (DIM) to generate noise-modulated observation conditions for a pretrained diffusion model. Experimental results show that our RD-Diff establishes a new state-of-the-art performance on both accuracy and diversity compared to existing methods.

Keywords: Diffusion Model · Human Motion Prediction · Transformer

1 Introduction

Human Motion Prediction (HMP) predicts future human motion based on an observation sequence, finding critical applications in areas such as human-robot collaboration (HRC), surveillance and autonomous driving [30, 40, 46, 47, 54, 63, 75, 81, 90, 103]. Due to the inherent uncertainty and stochastic nature of human movement, predicting accurate and diverse human poses and motions is particularly challenging, especially in safety-critical applications.

Prior research on deterministic HMP aims to regress a single future sequence of human skeletal poses based on observations [3,4,17,23,24,27,48,50,51,56,61, 65,67,70,87]. This approach often leads to the most likely result without considering the uncertainty and multi-modal nature of human motion. To address this, deep generative models have been introduced for better performance. Various approaches have employed generative adversarial networks (GANs) [9, 32, 35] and



(a) Two-stage Training Pipeline (b) RLTransformer (c) Diversity-Inducing Modulator

Fig. 1: RD-Diff Architecture. Our RD-Diff consists of novel Regulated Linear Transformer (RLTransformer), RLTransformer-based diffusion model $\mathcal{D}_{\boldsymbol{\theta}}$ and Diversity-Inducing Modulator (DIM) $\mathcal{M}_{\boldsymbol{\psi}}$. (a) RD-Diff employs a two-stage learning pipeline. (b) RLTransformer is designed with a specialized attention mechanism for frequency-transformed motion sequences to replace the conventional softmax transformer used in both $\mathcal{D}_{\boldsymbol{\theta}}$ and $\mathcal{M}_{\boldsymbol{\psi}}$. (c) $\mathcal{M}_{\boldsymbol{\psi}}$ conducts post-hoc modulation to observation x, and generates noise-modulated conditions $\{\hat{\mathbf{z}}_{\mathbf{k}}(\mathbf{t})\}_{k=1}^{K}$ as diffusion progresses. Through a tailored combination of RLTransformer, $\mathcal{D}_{\boldsymbol{\theta}}$ and $\mathcal{M}_{\boldsymbol{\psi}}$, RD-Diff fosters diverse and reasonable predictions in HMP.

variational autoencoders (VAEs) [11,26,44,60,93,98,100,104] to produce multiple future motion samples from an observed sequence. These approaches establish a conditional distribution for future poses, incorporating multiple loss constraints to ensure sample quality and diversity. However, these encoding-decoding approaches often face difficulties in fidelity and avoiding mode collapse [19,22]. Recently, diffusion models [38] have been introduced in HMP [1,8,19,80,96], showing improved distribution matching and avoidance of mode collapse [28,58,78]. Despite these advancements, there has been limited exploration of achieving strong performance in accuracy and diversity using the diffusion model.

In this paper, we propose RD-Diff, a novel diffusion-based model for HMP. Tailored for frequency-transformed motion sequences, we propose a Regulated Linear Transformer (RLTransformer), which is adopted in RLTransformer-based diffusion model $\mathcal{D}_{\boldsymbol{\theta}}$ to replace the softmax transformer. To further enhance the capability of the diffusion model for both accuracy and diversity, we introduce a novel condition-modulating strategy, Diversity-Inducing Modulator (DIM) $\mathcal{M}_{\boldsymbol{\psi}}$, for the pretrained $\mathcal{D}_{\boldsymbol{\theta}}$. DIM $\mathcal{M}_{\boldsymbol{\psi}}$ facilitates the generation of diverse results by sampling latent noise variables from an imbalanced multimodal distribution using Gumbel-Softmax sampling. Subsequently, it modulates the observed pose sequence as different noise-modulated conditions by introducing scheduled, monotonically decreasing diversity-inducing latent noise variables. The design of RD-Diff is grounded in a nuanced understanding of future human motions, acknowledging their deterministic properties influenced by physical laws and historical movements, as well as the imperative need for diversity in real-world applications. The decomposition of establishing a motion prior and diversity-induced sampling process enhances control and flexibility. The modular design of DIM $\mathcal{M}_{\psi}$ has the potential to be employed alongside any given pretrained model and sampling algorithm. Experiment results on two typical benchmark datasets demonstrate that RD-Diff substantially improves the diversity and accuracy of HMP over existing models, and produces new state-of-the-art on all metrics. In summary, our contributions are three-fold:

1. We propose RD-Diff, a novel generative diffusion model for HMP with a two-stage training pipeline to enhance the diffusion model's capabilities for both accuracy and diversity.
2. We introduce a specifically designed transformer architecture, RLTransformer, for improved representation of frequency domain human motion features.
3. We introduce a novel condition-modulation method, DIM, which enables diverse and accurate sampling from complex imbalanced distributions, generates diversity-inducing Gaussian noise, and modulates condition embedding in the diffusion model to achieve balanced capabilities for both accuracy and diversity in HMP.

2 Related Work

Human Motion Prediction. Traditional methods for HMP typically employ recurrent neural networks (RNNs) [5, 20, 21, 31, 53, 57, 68, 69, 73, 82, 86], transformers [2, 16, 71, 95], graph convolutional networks (GCNs) [25, 49, 62, 64] or multi-layer perceptrons (MLPs) [13, 34] to learn spatial-temporal dependencies. Recently, diverse HMP approaches have gained attention [6,7,10,18,33,45,55,59, 66, 88, 89, 99, 102]. HP-GAN [10] introduces a generative adversarial framework, using a random vector **z** to control diverse future pose generations. MT-VAE [99] uses a Variational Autoencoder (VAE) for conditional data distribution modeling. Many GAN and VAE-based methods often result in similar predictions due to latent vector similarities. Efforts have been made to improve the diversity. Yuan et al. propose DLow [102], explicitly inferring diverse random vectors. [26] disentangles the process of Dlow, while Mao et al. [66] introduce a method for mapping random vectors and poses to a future sequence. STARS [97] employs multi-level spatial-temporal anchors to generate diverse results, improving both diversity and accuracy among the VAE-based methods. However, they may have limited capacity to predict reasonable future motion sequences and have inherent limitations of posterior collapse [96].

Diffusion Models. Diffusion models, as a promising addition to deep generative models, generate high-quality samples through a forward process inspired by the second law of thermodynamics. Denoising Diffusion Probabilistic Models (DDPM) [58] achieves impressive results by predicting injected noise, but its inference process is time-consuming. Denoising Diffusion Implicit Model (DDIM) [85] addresses this by introducing a non-Markovian diffusion process, accelerating sampling while maintaining quality. Unlike previous stochastic approaches using GANs and VAEs to sample multiple plausible predictions, diffusion models offer a novel direction for considering uncertainty in HMP, generating diverse and contextually appropriate predictions [8, 91]. HumanMAC [19], TransFusion [91], MotionDiff [96], TCD [80] and BeLFusion [8] explore diffusion models in HMP. HumanMAC [19] introduces a mask-completion formulation rather than separated representations for observation and prediction. Without diversity requirement, TCD [80] has produced the highest accuracy. However, most diffusion models could not generate diverse motions while maintaining high fidelity. Our work builds on recent diffusion model developments, prioritizing a balance between accuracy and diversity.

3 Our Method

3.1 Preliminaries

Given the observed human motion $\boldsymbol{x}$, HMP aims to forecast the future motion sequence as $\tilde{\mathbf{y}}$. We adopt HumanMAC [19] as the diffusion baseline, capitalizing on its ability to model entire motion sequences using mask operations. We let $\boldsymbol{h} \in \boldsymbol{R}^{(O+F)\times 3J}$ denote the complete sequence of human motion, where O represents the number of frames in the observation sequence $\boldsymbol{x}$, F represents the number of frames in the future sequence $\tilde{\mathbf{y}}$ to be predicted, 3 denotes the Cartesian coordinates of a body joint, and J is the number of body joints. In line with earlier research [16, 19, 34, 62–64, 95, 104], we apply the Discrete Cosine Transform (DCT) [42] to transform the temporal domain into the frequency domain, facilitating compact yet informative feature extraction. DCT leverages the advantages of frequency-domain representations to capture periodic patterns and temporal dependencies thereby mitigating extreme jittering in human motion.

3.2 Method Overview

The architecture of RD-Diff is illustrated in Fig. 1. During training, RLTransformer-based diffusion model $\mathcal{D}_{\boldsymbol{\theta}}$ is trained in stage 1, building a motion prior; while Diversity-Inducing Modulator (DIM) $\mathcal{M}_{\boldsymbol{\psi}}$ is trained in stage 2, conducting post-hoc modulation to observation x. During inference, RD-Diff samples a result $\tilde{\mathbf{y}}$ from the distribution of $p(\tilde{\mathbf{y}}|\mathbf{x}) = \int p_{\boldsymbol{\theta}}(\tilde{\mathbf{y}}|\mathbf{x}, \hat{\mathbf{z}}) r_{\boldsymbol{\psi}}(\hat{\mathbf{z}}|\mathbf{x}) d\hat{\mathbf{z}}$. $\hat{\mathbf{z}}$ denotes a noise-modulated condition sampled from a distribution generated by $\mathcal{M}_{\boldsymbol{\psi}}$, $p_{\boldsymbol{\theta}}(\tilde{\mathbf{y}}|\mathbf{x}, \hat{\mathbf{z}})$ denotes the conditional distribution modeled by $\mathcal{D}_{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta}$, and $r_{\boldsymbol{\psi}}(\hat{\mathbf{z}}|\mathbf{x})$ denotes the latent distribution modeled by the $\mathcal{M}_{\boldsymbol{\psi}}$ parameterized by $\boldsymbol{\psi}$. Formally, the generation of $\tilde{\mathbf{y}}$ involves two sequential steps:

$$\hat{\mathbf{z}} = \mathcal{M}_{\boldsymbol{\psi}}(\mathbf{x}), \tag{1}$$

$$\tilde{\mathbf{x}}, \tilde{\mathbf{y}} = \mathcal{D}_{\boldsymbol{\theta}}(\mathbf{x}, \hat{\mathbf{z}}). \tag{2}$$

$\mathcal{D}_{\boldsymbol{\theta}}$ outputs not only the predicted future pose sequence $\tilde{\mathbf{y}}$ but also "predicted" observed pose sequence $\tilde{\mathbf{x}}$ for the ease of reconstruction loss in Sec. 3.6.

3.3 Regulated Linear Transformer

Human motion may contain both low-frequency features (e.g., global body movement) and high-frequency features (e.g., fine-grained hand or leg movement). Meanwhile, high-frequency features might be related to noise and errors. Softmax transformer [92], which normalizes weights based on exponential terms, often leads to sharp weight distribution [14, 77, 105]. In contrast, linear transformer [52] tends to be excessively smooth [15,41]. Inspired by [37], a model with focus ability between a linear transformer and a softmax transformer achieves high efficiency and expressiveness in image tasks. However, the fixed mapping function in [37] by adjusting element-wise power may not be ideal for capturing the dynamic nature of human motion sequences. A more adaptive attention mechanism is needed to accommodate the variations in human motion patterns.

We present **Regulated Linear Transformer (RLTransformer)**, featuring a novel attention mechanism built on top of a linear transformer to capture both essential low-frequency movements and subtle high-frequency details, achieving a better dominant frequency focusing. To regulate the linear transformer, we train an adaptive frequency re-weighting function and a channel-wise re-weighting function to adjust the query features. The architecture of the RLTransformer is illustrated in Fig. 1 (b). The similarity function is defined as:

$$\mathrm{Sim}\left(Q_i, K_j\right) = \phi_q\left(Q_i\right)\phi_k\left(K_j\right)^T, \tag{3}$$

where $\phi_q(x) = \mathit{\Gamma}\left(\mathrm{ReLU}(x)\right)$, $\phi_k(x) = \mathrm{ReLU}(x)$, and $\mathit{\Gamma}$ denotes a novel regulated function. To guide the initialization of query vectors and ensure the network understands "what to query", we introduce feature re-weighting on both frequency and channel dimensions of the query vector $q \in \boldsymbol{R}^{N\times D}$, where N is the number of frequency tokens and D is the hidden dimension. Specifically, we initialize a trainable normalized vector $S \in \boldsymbol{R}^{D}$, which is then multiplied with normalized query vector $\hat{q}$ to obtain the frequency re-weighting vector $F \in \boldsymbol{R}^{N}$. Subsequently, frequency re-weighting is applied to q to obtain the frequency-wise regulated query q_f. Channel-wise re-weighting is then conducted on q_f using a learnable vector $C \in \boldsymbol{R}^{D}$, resulting in the channel-wise regulated query q_{fc}. Finally, q_{fc} undergoes an MLP layer and is added to the original q to obtain the regulated query q_r. Overall, the regulation function $\mathit{\Gamma}$ performs the following calculation: $F = \mathrm{norm}(\hat{q} \times S), q_f = q \odot F, q_{fc} = q_f \odot C, q_r = q + \mathrm{MLP}(q_{fc})$, where norm denotes normalization and $\odot$ denotes element-wise multiplication.

Conventional linear transformers compute $K^T V$ first, which leads to the complexity of $\mathcal{O}(ND^2)$. However, we compute QK^T first to save computation cost, considering that the hidden dimension ($D = 512$) is larger than the number of tokens ($N = 20$) in our task, in contrast to image tasks where N is usually larger than D. Moreover, $\mathit{\Gamma}$ can be applied to query and key features, which will be further discussed in Sec. 4.2.

3.4 RLTransformer-based Diffusion Model

In **RLTransformer-based diffusion model** $\mathcal{D}_{\boldsymbol{\theta}}$, the forward process begins with projecting the motion sequence $\boldsymbol{h}$ to the frequency domain using the DCT

```
Algorithm 1: Diversity-Inducing Modulator(DIM)
  Input: Observation sequence $\mathbf{x}$, number of output samples $K$, observation
         encoder $\mathcal{S}_{\boldsymbol{\psi}}$
  Output: $K$ noise-modulated condition embeddings $\{\hat{\mathbf{x}}_{\mathbf{k}}(\mathbf{t})\}_{k=1}^{K}$ for $\mathcal{D}_{\boldsymbol{\theta}}$ at
          diffusion step $t$
  $\{\mathbf{A}_k, \mathbf{b}_k\}_{k=1}^{K} = \mathcal{S}_{\boldsymbol{\psi}}(\mathbf{x})$;
  for $k = 0, 1, \ldots, K$ do
      $\mathbf{z}_k \sim \mathcal{N}(\mathbf{b}_k, \mathbf{A}_k)$;
      $\alpha_t^c = 1 - \beta_t^c$, where $\beta_t^c$ is cosine scheduler;
      $\hat{\mathbf{z}}_{\mathbf{k}}(\mathbf{t}) = \mathrm{zTrans}(\sqrt{\alpha_t^c}\boldsymbol{x} + \sqrt{1-\alpha_t^c}\mathbf{z}_{\boldsymbol{k}})$.
```

operation: $\boldsymbol{h}^{\boldsymbol{D}} = \mathrm{DCT}(\boldsymbol{h})$, where $\boldsymbol{h}^{\boldsymbol{D}} \in \boldsymbol{R}^{(O+F)\times 3J}$. Given that crucial information about human motion resides in lower frequency coefficients and higher frequency terms are likely noise-related, we retain only the first N frequency dimensions of data: $\boldsymbol{h}^{\boldsymbol{D}} \in \boldsymbol{R}^{N\times 3J}$. The noisy DCT pose sequence $\boldsymbol{h}_t^{\boldsymbol{D}}$ at diffusion step t is sampled using the reparameterization trick:

$$\boldsymbol{h}_t^{\boldsymbol{D}} = \sqrt{\bar{\alpha}_t}\boldsymbol{h}_0^{\boldsymbol{D}} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \tag{4}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, $\alpha_i \in [0,1]$ are pre-defined variance parameters, $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, and $\boldsymbol{h}_0^{\boldsymbol{D}}$ corresponds to $\boldsymbol{h}^{\boldsymbol{D}}$ [19].

Regarding noise prediction in the backward process, we adopt RLTransformer for the noise prediction network $\boldsymbol{\epsilon}_\theta$ instead of the U-net used in the original DDPM [38], as shown in Fig. 1 (a). To enable predictions conditioned on the observation, we leverage the FiLM conditioning [76], which carries out feature-wise affine transformation, to guide the generation. The observation sequence $\boldsymbol{x}$ is first padded with the last observed frame to match the length of the complete motion sequence. Subsequently, the padded sequence is processed through the DCT operator to obtain compact historical information. The noise prediction network $\boldsymbol{\epsilon}_\theta$ comprises l_1 layers of RLTransformer and long skip connections between the shallow and deep layers, following U-Net [79]. The introduction of the RLTransformer allows the network to focus on relevant features and adaptively modify the input query before further processing. In addition, we adopt "FreeU" [83] to re-weight feature maps from the backbone and skip connections to enhance generation quality. In the output stage, the motion sequence can be recovered from frequency coefficients using the inverse discrete cosine transform (IDCT): $\boldsymbol{h} = \mathrm{IDCT}(\boldsymbol{h}^{\boldsymbol{D}})$.

3.5 Diversity-Inducing Modulator

Existing diversity-promoting methods, consisting of diversity loss and direct decoding from latent space, might cause premature deviations from the ground truth, leading to implausible predictions [19,91]. As a remedy, we propose a novel **Diversity-Inducing Modulator (DIM)** $\mathcal{M}_{\boldsymbol{\psi}}$, where diversity is promoted by injecting learned decreasing noises to the condition of the diffusion model in all diffusion steps. Fig. 1 (c) and Algorithm 1 illustrate the modulating process. We first introduce an observation encoder $\mathcal{S}_{\boldsymbol{\psi}}$ with Gumbel-Softmax sampling to generate different Gaussian distributions from which noises are sampled. Then, we introduce a modulation strategy with a cosine scheduler to modulate the condition of the diffusion model using the generated noise. Unlike DLow [102], which focuses on diversifying latent flows, our DIM present a more flexible way by adjusting the noise intensity injected into the observation conditions.

Formally, given the observation sequence $\mathbf{x}$, observation encoder $\mathcal{S}_{\boldsymbol{\psi}}$, parameterized by $\boldsymbol{\psi}$, generates K different sets of mean $\mathbf{b}_k \in \boldsymbol{R}^D$ and variance $\mathbf{A}_k \in \boldsymbol{R}^{D\times D}$. Specifically, padded observation sequence is fed into l_2 layers of RL-Transformer to optimize a latent subspace $\mathbb{V} = \text{span}(\mathbf{v_1}, \ldots, \mathbf{v_N})$, with N learned base embeddings $\mathbf{v} \in \boldsymbol{R}^D$. One can conceptualize that the acquired embeddings collectively constitute a "latent subspace". Fundamentally, the latent subspace is optimized, from which random sampling effectively mirrors diverse samples from the target distribution. With the latent subspace, we use Gumbel-Softmax sampling to sample a weight vector $\mathbf{w} \in \boldsymbol{R}^N$ that contains N weights to combine the base embeddings as a point in the latent subspace: $\sum_{i=1}^{N} w_i^k \mathbf{v}_i$, where $\sum_{i=1}^{N} w_i^k = 1$. We repeat the process K times to obtain K points from the latent subspace, which are then passed into two MLPs to get $\{\mathbf{A}_k\}_{k=1}^{K}$ and $\{\mathbf{b}_k\}_{k=1}^{K}$, respectively. Afterwards, $\mathcal{M}_{\boldsymbol{\psi}}$ employs the reparameterization trick to sample latent noise variables from these distributions: $\mathbf{z}_k \sim \mathcal{N}(\mathbf{b}_k, \mathbf{A}_k) = \mathbf{A}_k \boldsymbol{\epsilon} + \mathbf{b}_k$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 1)$. To maintain the essential correspondence between the condition and the output, a scheduling strategy is employed that lessens the noise corruption in the condition as the diffusion progresses from $t = T$ to $t = 0$, causing it to diminish during the final steps, where T is the number of diffusion steps. More specifically, inspired by the forward process of diffusion models, we transform a given $\boldsymbol{x}$ to a noise-modulated condition:

$$\hat{\mathbf{z}}_{\mathbf{k}}(\mathbf{t}) = \text{zTrans}(\sqrt{\alpha_t^c}\boldsymbol{x} + \sqrt{1 - \alpha_t^c}\mathbf{z}_{\boldsymbol{k}}), \tag{5}$$

where $\alpha_t^c = 1 - \beta_t^c$, β_t^c is the cosine scheduler [72], and zTrans denotes z-Transform to rescale the conditioning vector back toward its prior mean and standard deviation. $\mathcal{M}_{\boldsymbol{\psi}}$ processes a different $\hat{\mathbf{z}}_{\mathbf{k}}(\mathbf{t})$ at each step with scheduled fading noise, leading to more diverse outputs. Finally, we map $\mathbf{x}$ and $\hat{\mathbf{z}}_{\mathbf{k}}(\mathbf{t})$ to a future pose sequence using the pre-trained $\mathcal{D}_{\boldsymbol{\theta}}$ introduced in Sec. 3.4.

3.6 Learning Objective

In stage 1, we train $\boldsymbol{\epsilon}_\theta$ in $\mathcal{D}_{\boldsymbol{\theta}}$ by minimizing the mean squared error (MSE): $\mathcal{L}_1 = \mathbb{E}_{t,\boldsymbol{h}_0^{\boldsymbol{D}},\epsilon} \left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\boldsymbol{h}_t^{\boldsymbol{D}}, t\right)\right\|_2^2$. It aims to learn the accurate representation of the observed data and establish a robust prior. In stage 2, we adopt the pre-trained $\mathcal{D}_{\boldsymbol{\theta}}$ to train $\mathcal{M}_{\boldsymbol{\psi}}$, where the condition is encoded in a latent space that is then corrupted with decreasing noise to maximize diversity while keeping the latent and the predicted motion accurate. Stage 2 aims to introduce variability and diversity into predictions. We minimise three loss functions: diversity loss, accuracy loss [102] and a newly proposed reconstruction loss to train the $\mathcal{S}_{\boldsymbol{\psi}}$ in $\mathcal{M}_{\boldsymbol{\psi}}$.

The diversity loss $\mathcal{L}_{div}$ is optimized to increase the distances between pairs of predictions from the same input $\mathbf{x}$: $\mathcal{L}_{div} = \frac{1}{K(K-1)} \sum_{i=1}^{K} \sum_{j \neq i}^{K} \exp\left(-\frac{\|\tilde{\mathbf{y}}_i, \tilde{\mathbf{y}}_j\|_2}{\sigma}\right)$, where $\tilde{\mathbf{y}}_\mathbf{i}$ and $\tilde{\mathbf{y}}_\mathbf{j}$ denote predicted pose sequence, and σ denotes the scaling factor. Besides, the accuracy loss $\mathcal{L}_{acc}$ is employed: $\mathcal{L}_{acc} = \min_k \|\mathbf{y}, \tilde{\mathbf{y}}_k\|_2, k \in [1, K]$, where $\mathbf{y}$ is the ground truth future pose sequence. Furthermore, we aim to generate one good sample close to the ground truth and align all predictions with the observed pose sequence while maintaining a certain degree of diversity. Therefore, we propose an additional reconstruction loss $\mathcal{L}_{rec}$ for alignment with the observed pose sequence:

$$\mathcal{L}_{rec} = \frac{1}{K} \sum\nolimits_{i=1}^{K} \left(\|\mathbf{x}, \tilde{\mathbf{x}}_i\|_2\right). \tag{6}$$

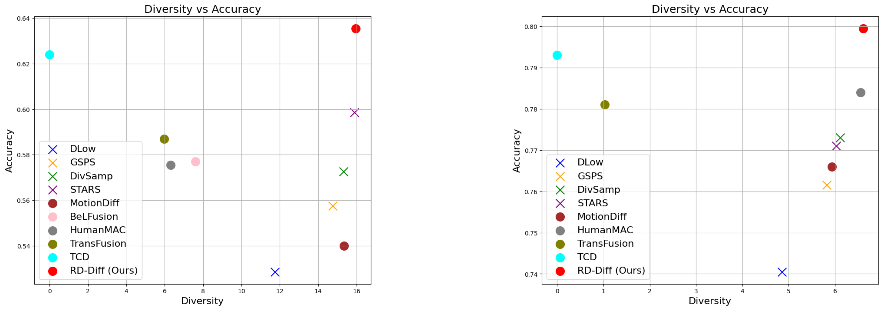
Altogether, the training loss in stage 2 is: $\mathcal{L}_2 = \lambda_{div}\mathcal{L}_{div} + \lambda_{acc}\mathcal{L}_{acc} + \lambda_{rec}\mathcal{L}_{rec}$, where hyper-parameters λ_{div}, λ_{acc}, and λ_{rec}, denoting weights of diversity loss $\mathcal{L}_{div}$, accuracy loss $\mathcal{L}_{acc}$ and reconstruction loss $\mathcal{L}_{rec}$ respectively, are employed to achieve a balance among the three losses. The two-stage training pipeline aims to untangle the direct association between network parameter learning tasks and diverse predictions. This enables the network to generate extreme predictions present in minor target distributions [26].

4 Experiments

4.1 Settings

Datasets. The performance of RD-Diff is evaluated on two widely used HMP datasets: Human3.6M [39] and HumanEva-I [84], following the evaluation protocols established in [66, 102]. (1) **Human3.6M**: It comprises seven subjects, each performing 15 action categories. Training utilizes data from five subjects (S1, S5, S6, S7, S8), with the remaining two subjects (S9, S11) used for testing. There are 17 joints in total for each pose after removing unnecessary joints. The input consists of 25 frames (0.5s at 50fps), predicting 100 frames (2s) into the future. (2) **HumanEva-I**: This dataset involves three subjects, each participating in five action categories. Each pose is described by 15 joints, and we predict 60 future poses (1s at 60fps) based on 15 frames (0.25s). As highlighted in [102], Human3.6M offers a larger dataset with greater motion variation, while HumanEva-I is smaller with less variation.

Evaluation Metrics. Five metrics are employed in the evaluation: (1) **APD** measures the Average Pairwise Distance of predicted results from an input, indicating the diversity of results [7]. (2) **ADE** computes the Average Displacement Error between the ground truth and the most similar result, while (3) **FDE** calculates the Final Displacement Error, focusing only on the last pose. (4) **MMADE** and (5) **MMFDE**, the multimodal versions of ADE and FDE introduced in [102], evaluate accuracy by considering multiple similar past motions $\mathbf{x}_p$ and their corresponding future motions $\{\mathbf{y}_p\}_{p=1}^{P}$. These metrics collectively provide insights into the diversity and accuracy of the predicted results.



(a) Human3.6M **(b)** HumanEva-I

Fig. 2: Performance on both diversity and accuracy (1-error). The dot markers indicate diffusion models, the cross markers indicate other models, and the red dot represents our RD-Diff model. Markers towards the top right corner indicate better performance.

Implementation Details. RD-Diff is trained on both datasets using a 1000-step diffusion model, and sampling is performed with a 100-step DDIM [85]. $\mathcal{D}_{\boldsymbol{\theta}}$ is trained for 1000 epochs, and $\mathcal{M}_{\boldsymbol{\psi}}$ is trained for 200 epochs. Scaling factors for backbone (b_{FreeU}) and skip connection (s_{FreeU}) in "FreeU" are 1.2 and 1.0. For both datasets, the learning rate is $3e^{-4}$ for train stage 1 and $1e^{-3}$ for train stage 2. Variance scheduling in $\mathcal{D}_{\boldsymbol{\theta}}$ is achieved using the cosine scheduler [72]. We set the number of samples $K = 50$ for comparison with previous approaches, and the values for hyperparameters λ_{div}, λ_{acc}, and λ_{rec} are 1, 2, and 0.2 for Human3.6M, and 4, 1, and 0.1 for HumanEva-I, respectively. All experiments are conducted using PyTorch [74] on an NVIDIA Tesla V100 GPU, and Adam [43] is employed as the optimizer for consistency across all experiments. RD-Diff (B) and RD-Diff (L) denote base and large versions of RD-Diff, where regulation function $\mathit{\Gamma}$ in RLTransformer is applied to Q in RD-Diff (B), and applied to both Q and K in RD-Diff (L). Unless otherwise stated, all experiments are conducted on the Human3.6M dataset using RD-Diff (B).

4.2 Comparison with the State-of-the-Arts

Quantitative Results. To show a global view of performance comparison, we create 2D charts as shown in Fig. 2, where APD indicates diversity (horizontal axis) and [1-(ADE+FDE)/2] indicates accuracy (vertical axis), for consistency between two datasets. Our model is nearest to the upper-rightmost corner in both plots and significantly leads over other models, highlighting substantial progress. The quantitative comparisons are shown in Tab. 1, where the previous methods are classified into two categories: deterministic methods (first four rows) and stochastic methods (remaining rows). Deterministic methods produce a single output, lacking APD scores for comparison. All stochastic methods predict 50 future sequences for each input historical pose sequence. By predicting multiple results, stochastic methods can capture diverse possibilities. They are further divided into diffusion-based methods (the 5 methods above RD-Diff) and the

Table 1: Quantitative comparisons where the bold number indicates the top performance and the underlined number for the second position. All the results are calculated using 50 prediction samples for each input historical pose sequence.

Method	Human3.6M					HumanEva-I				
	APD ↑	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓	APD ↑	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓
ERD [31]	-	0.722	0.969	0.776	0.995	-	0.382	0.461	0.521	0.595
acLSTM [53]	-	0.789	1.126	0.849	1.139	-	0.429	0.541	0.530	0.608
LTD [67]	-	0.516	0.756	0.627	0.795	-	0.415	0.555	0.509	0.613
MSR [27]	-	0.508	0.742	0.621	0.791	-	0.371	0.493	0.472	0.548
Pose-Knows [94]	6.723	0.461	0.560	0.522	0.569	2.308	0.269	0.296	0.384	0.375
MT-VAE [99]	0.403	0.457	0.595	0.716	0.883	0.021	0.345	0.403	0.518	0.577
HP-GAN [10]	7.214	0.858	0.867	0.847	0.858	1.139	0.772	0.749	0.776	0.769
BoM [12]	6.265	0.448	0.533	0.514	0.544	2.846	0.271	0.279	0.373	0.351
GMVAE [29]	6.769	0.461	0.555	0.524	0.566	2.443	0.305	0.345	0.408	0.410
DeLiGAN [36]	6.509	0.483	0.534	0.520	0.545	2.177	0.306	0.322	0.385	0.371
DSF [101]	9.330	0.493	0.592	0.550	0.599	4.538	0.273	0.290	0.364	0.340
DLow [102]	11.741	0.425	0.518	0.495	0.531	4.855	0.251	0.268	0.362	0.339
GSPS [66]	14.757	0.389	0.496	0.476	0.525	5.825	0.233	0.244	0.343	0.331
DivSamp [26]	15.310	0.370	0.485	0.475	0.516	6.109	0.220	0.234	0.342	0.316
STARS [97]	15.884	0.358	0.445	0.442	0.471	6.031	0.217	0.241	0.328	0.321
MotionDiff [96]	15.353	0.411	0.509	0.508	0.536	5.931	0.232	0.236	0.352	0.320
BeLFusion [8]	7.602	0.372	0.474	0.473	0.507	-	-	-	-	-
HumanMAC [19]	6.301	0.369	0.480	0.509	0.545	6.554	0.209	0.223	0.342	0.335
TransFusion [91]	5.975	0.358	0.468	0.506	0.539	1.031	0.204	0.234	0.408	0.427
TCD [80]	-	0.356	0.396	0.463	0.445	-	0.199	0.215	-	-
RD-Diff (B)	15.714	0.347	0.401	0.445	0.444	6.561	0.199	0.218	0.321	0.308
RD-Diff (L)	**15.950**	**0.342**	**0.387**	**0.441**	**0.431**	**6.607**	**0.191**	**0.210**	**0.317**	**0.302**

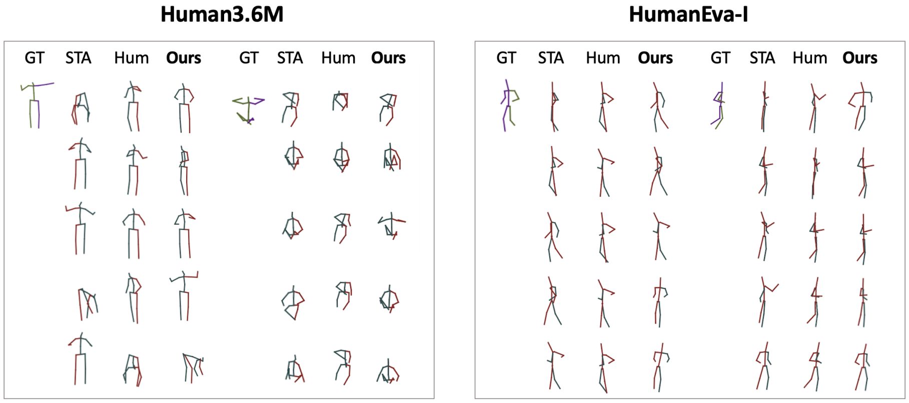


Fig. 3: End poses in 5 samples. STA and Hum denote STARS [97] and HumanMAC [19], respectively.

rest. These results indicate that our RD-Diff achieves superior capability among other models. This suggests that the RD-Diff effectively captures the variability in training data and covers a broad range of variability that previous methods cannot encompass.

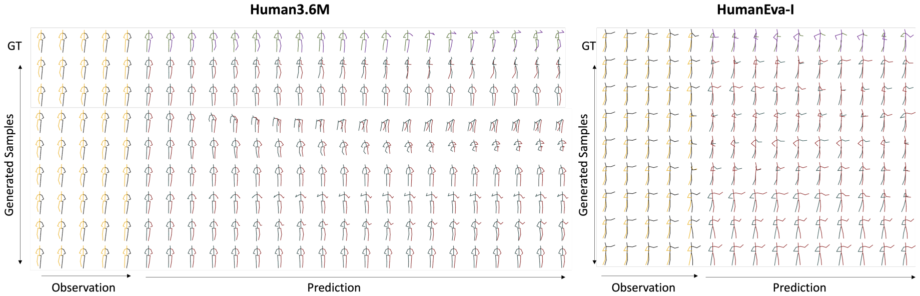


Fig. 4: Given observation frames, 8 samples are generated by RD-Diff on Human3.6M and HumanEva-I datasets, respectively. GT denotes the ground truth motion sequence.

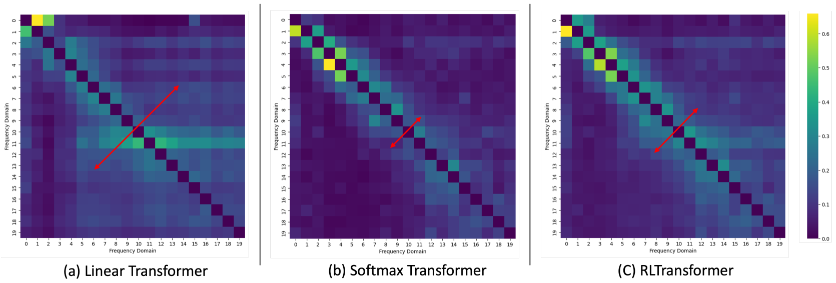
Table 2: Comparison with baseline models.

Method	Human3.6M					HumanEva-I				
	APD ↑	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓	APD ↑	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓
(1) Baseline	6.325	0.371	0.484	0.513	0.549	6.300	0.217	0.229	0.346	0.349
(2) Baseline + RLTransformer	6.452	0.365	0.417	0.479	0.468	6.263	0.211	0.225	0.342	0.325
(3) Baseline + DIM	15.657	0.351	0.408	0.449	0.457	6.515	0.203	0.222	0.323	0.326
(4) **RD-Diff**	15.714	0.347	0.401	0.445	0.444	6.561	0.199	0.218	0.321	0.308

Qualitative Results. To illustrate the diversity of the predicted poses, we visualize five end poses from pose sequences predicted by STARS [97], HumanMAC [19], and our proposed method, in Fig. 3. STARS realizes high diversity but with some failure cases regarding physical constraints. HumanMAC [19] generates realistic poses that typically lack diversity. By contrast, while maintaining high fidelity across different datasets, our method produces more diverse and reasonable results than STARS and HumanMAC. Furthermore, Fig. 4 shows 8 samples generated for each dataset, which shows our RD-Diff can produce diverse and reasonable motion sequences.

4.3 Ablation Study

Comparison with Baseline. Tab. 2 presents ablation studies confirming the efficacy of the design components in our RD-Diff. In Method (1), the baseline method adopts a vanilla softmax transformer and diffusion model, following HumanMAC [19], where the observed pose sequence x directly serves as a condition embedding. In Method (2), we replace the softmax transformer with RLTransformer, resulting in diversity boosts but slight accuracy drops. In Method (3), adding DIM to the baseline significantly boosts performance. In Method (4), we further replace the softmax transformer in Baseline and DIM (i.e., Method (3)) with RLTransformer, which further boosts the performance. It is worth noting that the DIM is a strong contribution, where the idea of inducing stochasticity in the condition applies to any diffusion-based system. The results show that DIM can promote diversity and accuracy with the only cost of a second training stage.



(a) Linear Transformer (b) Softmax Transformer (C) RLTransformer

Fig. 5: The attention weights heatmaps of linear transformer, softmax transformer, and our RLTransformer on DCT frequency domain from a motion clip in Human3.6M. For better visualization, we replace the diagonal attention scores with zero, then normalise the rest of the scores from 0 to 1. Red arrows indicate the range of attention focus.

Table 3: Re-weighting methods in RLTransformer. FR and CR denote frequency re-weighting and channel re-weighting, respectively. The "-" sign means removing a module from RD-Diff.

Method	APD↑	ADE↓	FDE↓	MMADE↓	MMFDE↓
RD-Diff w/o FR	15.701	0.350	0.406	0.447	0.455
RD-Diff w/o CR	15.707	0.348	0.403	0.447	0.450

Table 4: Number of RLTransformer Layers. l_1 denotes the numbers of layers in RLTransformer-based diffusion model.

	Human3.6M					HumanEva-I				
l_1	APD↑	ADE ↓	FDE↓	MMADE ↓	MMFDE ↓	APD↑	ADE ↓	FDE↓	MMADE ↓	MMFDE ↓
2	**17.976**	0.424	0.472	0.508	0.496	**6.904**	0.208	0.227	0.335	0.319
4	16.294	0.363	0.433	0.494	0.492	6.561	**0.199**	0.218	**0.321**	**0.308**
6	15.644	0.355	0.423	0.475	0.498	6.568	0.201	0.219	0.343	0.309
8	15.714	0.347	**0.401**	**0.445**	**0.444**	6.493	0.202	**0.217**	0.349	0.314
10	14.771	**0.346**	0.410	0.446	0.467	6.435	0.202	0.221	0.344	0.312

RD-Diff suggests a way to build a motion prior (i.e., $\mathcal{D}_{\boldsymbol{\theta}}$) and induce stochasticity in the condition through DIM. Ablation studies on one-stage training will be detailed in the supplementary material.

RLTransformer. As shown in Fig. 5, softmax attention produces a sharp distribution along the diagonal and the top left corner (shortest red arrowed line), and linear attention's distribution is relatively smooth (longest red arrowed line), while RLTransformer shows a sharpness between them. This shows that the softmax transformer has a small "attention window" in frequency tokens (nearest 5 frequency tokens). It might tend to emphasize dominant frequency components, potentially causing a global attention bias and neglecting other crucial information, such as high-frequency components containing fine-grained body movements. Our RLTransformer slightly smooths the sharp distribution in softmax

Table 5: Random noise added to observation. "Baseline 2" refers to "Method (2)" in Tab. 2. b and A denote the mean and variance in a normal distribution $\mathcal{N}(b, A)$.

	Human3.6M								HumanEva-I							
	RD-Diff	Baseline 2	b=0			b=1			RD-Diff	Baseline 2	b=0			b=1		
			$A = 1$	$A = 1.5$	$A = 2$	$A = 1$	$A = 1.5$	$A = 2$			$A = 1$	$A = 4$	$A = 6$	$A = 1$	$A = 4$	$A = 6$
APD ↑	15.714	6.452	10.311	14.360	18.643	10.285	14.464	18.632	6.561	6.263	6.252	7.206	7.641	6.307	7.219	7.745
ADE ↓	0.347	0.365	0.512	0.624	0.794	0.509	0.625	0.793	0.199	0.211	0.316	0.486	0.613	0.302	0.451	0.642
FDE ↓	0.401	0.417	0.679	0.868	0.907	0.643	0.869	0.892	0.218	0.225	0.291	0.415	0.619	0.312	0.409	0.607
MMADE ↓	0.445	0.479	0.671	0.741	0.838	0.686	0.716	0.822	0.321	0.342	0.391	0.546	0.649	0.374	0.547	0.661
MMFDE ↓	0.444	0.468	0.591	0.670	0.809	0.606	0.701	0.790	0.308	0.325	0.364	0.615	0.631	0.384	0.637	0.645

Table 6: Noise schedulers in DIM.

Scheduler	APD↑	ADE↓	FDE↓	MMADE↓	MMFDE↓
None	16.592	0.468	0.596	0.460	0.498
Linear	11.549	0.382	0.425	0.450	0.448
Sqrt	15.236	0.463	0.594	0.453	0.488
Cosine	15.714	0.347	0.401	0.445	0.444

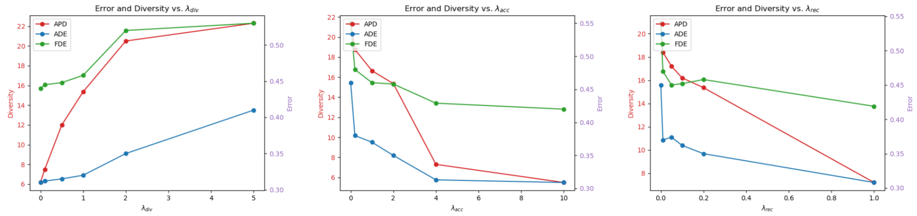
attention, contributing to the performance boosts, as shown in Tab. 2. Additionally, we conduct a quantitative analysis of attention sharpness by calculating the entropy of the attention matrix. A sharper attention map will yield smaller entropy, indicating a greater concentration on specific positions. The sharpness of the linear transformer, softmax transformer, and RLTransformer is 8.33, 8.05, and 8.25, respectively, consistent with the observation from Fig. 5.

We compare GFLOPs and the number of parameters on one layer of softmax transformer and RLTransformer with one video clip in Human3.6M. GFLOPs are 0.089 and 0.091, respectively, whereas the number of parameters is the same for both designs (0.099m). This shows that our RLTransformer can achieve better expressiveness than softmax attention while maintaining the relatively same amount of computation.

To further investigate the individual effects of re-weighting methods in RLTransformer, we remove frequency re-weighting (FR) and channel re-weighting (CR) from RD-Diff, as shown in Tab. 3. The removal of FR has a more pronounced effect than CR as it allows capturing of crucial temporal patterns.

Tab. 4 presents the results of experiments with different numbers of layers in RLTransformer-based diffusion model (l_1) on the Human3.6M dataset. We adopt 8 layers for Human3.6M and 4 for HumanEva-I, yielding the best performance in most error metrics. Afterwards, we fix l_1 and study the number of layers in DIM (l_2), and find l_2=4 is the optimal option for both datasets.

Diversity-Inducing Modulator. We conduct experiments to determine what will happen if random noises, instead of latent noise variables, are added to the observed pose sequence x. We remove observation encoder $\mathcal{S}_\psi$ and draw noise z_k from a random normal distribution $\mathcal{N}(b, A)$, rather than sampling from the Gaussian distribution generated by observation encoder $\mathcal{S}_\psi$. We conduct experiments on different means and variances. As shown in Tab. 5, incorporating noise in observation x enhances result diversity, with noise in larger A leading to more



(a) Diversity Weight λ_{div} **(b)** Accuracy Weight λ_{acc} **(c)** Reconstruction Weight λ_{rec}

Fig. 6: Ablation on weights of different losses. Each subplot uses the left axis (diversity) for the APD metric and the right axis (error) for both ADE and FDE metrics. The weights of the other losses that are not being evaluated in each subplot are fixed at 1.

diverse outcomes. Nevertheless, this comes at the cost of reduced accuracy, resulting in unrealistically generated poses. Consequently, our observation encoder $\mathcal{S}_{\psi}$ in DIM proves superior over directly introducing noise to observations.

We assess the effects of various noise schedulers in DIM, including none (no noise scheduler applied), linear, sqrt, and cosine, as shown in Tab. 6. The method "None" yields the worst performance, where noise intensity does not change as diffusion progresses. In contrast, the cosine scheduler yields the most favourable outcomes for noise modulation. This shows that the noise needs to decrease to a low value early before diffusion ends to ensure the observation information is passed effectively.

Weights of different losses. In Fig. 6, we perform an ablation study on the weight of the different losses. We set λ_{div} to 0, 0.1, 0.5, 1, 2, and 5. As shown in Fig. 6a, as λ_{div} increases, diversity increases, but error increases. We set λ_{acc} to 0, 0.2, 1, 2, 4, and 10. As shown in Fig. 6b, as λ_{acc} increases, the error decreases, but diversity decreases. We set λ_{rec} to 0, 0.01, 0.05, 0.1, 0.2, and 1. As shown in Fig. 6c, as λ_{rec} increases, the error decreases, but diversity decreases. We finally choose λ_{div}, λ_{acc} and λ_{rec} are 1, 2 and 0.2 for Human3.6M, achieving a desirable balance between diversity and accuracy.

5 Conclusion

In conclusion, we present **RD-Diff**, a novel HMP model that improves fidelity and diversity. Our Regulated Linear Transformer (RLTransformer) dynamically regulates linear attention via adaptive modeling of inter-dependencies among different frequencies, optimizing the network's learning process. In addition, our Diversity-Inducing Modulator (DIM) effectively generates noise-modulated conditions for a pretrained diffusion model, utilizing a novel modulation strategy with a cosine scheduler, resulting in more varied and accurate predictions. RD-Diff outperforms existing state-of-the-art methods in quantitative and qualitative comparisons.

References

1. Ahn, H., Mascaro, E.V., Lee, D.: Can we use diffusion probabilistic models for 3d motion prediction? In: arXiv preprint arXiv:2302.14503 (2023) 2
2. Aksan, E., Kaufmann, M., Cao, P., Hilliges, O.: A spatio-temporal transformer for 3d human motion prediction. In: 2021 International Conference on 3D Vision (3DV). pp. 565–574 (2021) 3
3. Aksan, E., Cao, P., Kaufmann, M., Hilliges, O.: Attention, please: A spatio-temporal transformer for 3d human motion prediction. arXiv preprint arXiv:2004.08692 **2**(3), 5 (2020) 1
4. Aksan, E., Kaufmann, M., Cao, P., Hilliges, O.: A spatio-temporal transformer for 3d human motion prediction. In: 2021 International Conference on 3D Vision (3DV). pp. 565–574. IEEE (2021) 1
5. Aksan, E., Kaufmann, M., Hilliges, O.: Structured prediction helps 3d human motion modelling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7144–7153 (2019) 3
6. Aliakbarian, S., Saleh, F., Petersson, L., Gould, S., Salzmann, M.: Contextually plausible and diverse 3d human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11333–11342 (2021) 3
7. Aliakbarian, S., Saleh, F.S., Salzmann, M., Petersson, L., Gould, S.: A stochastic conditioning scheme for diverse human motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5223–5232 (2020) 3, 8
8. Barquero, G., Escalera, S., Palmero, C.: Belfusion: Latent diffusion for behavior-driven human motion prediction. In: arXiv preprint arXiv:2211.14304 (2022) 2, 4, 10
9. Barsoum, E., Kender, J., Liu, Z.: Hp-gan: Probabilistic 3d human motion prediction via gan. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1418–1427 (2018) 1
10. Barsoum, E., Kender, J., Liu, Z.: Hp-gan: Probabilistic 3d human motion prediction via gan. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 1418–1427 (2018) 3, 10
11. Bhattacharyya, A., Schiele, B., Fritz, M.: Accurate and diverse sampling of sequences based on a "best of many" sample objective. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8485–8493 (2018) 1
12. Bhattacharyya, A., Schiele, B., Fritz, M.: Accurate and diverse sampling of sequences based on a "best of many" sample objective. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8485–8493 (2018) 10
13. Bouazizi, A., Holzbock, A., Kressel, U., Dietmayer, K., Belagiannis, V.: Motionmixer: Mlp-based 3d human body pose forecasting. In: arXiv preprint arXiv:2207.00499 (2022) 3
14. Cai, H., Gan, C., Han, S.: Efficientvit: Enhanced linear attention for high-resolution low-computation visual recognition. arXiv preprint arXiv:2205.14756 (2022) 5
15. Cai, H., Li, J., Hu, M., Gan, C., Han, S.: Efficientvit: Multi-scale linear attention for high-resolution dense prediction (2024) 5

16. Cai, Y., Huang, L., Wang, Y., Cham, T.J., Cai, J., Yuan, J., Thalmann, N.M.: Learning progressive joint propagation for human motion prediction. In: Computer Vision–ECCV 2020: 16th European Conference. vol. 16, pp. 226–242 (2020) 3, 4
17. Cai, Y., Huang, L., Wang, Y., Cham, T.J., Cai, J., Yuan, J., Liu, J., Yang, X., Zhu, Y., Shen, X., et al.: Learning progressive joint propagation for human motion prediction. In: European Conference on Computer Vision. pp. 226–242. Springer (2020) 1
18. Cai, Y., Wang, Y., Zhu, Y., Cham, T.J., Cai, J., Yuan, J., Liu, J., Zheng, C., Yan, S., Ding, H., et al.: A unified 3d human motion synthesis model via conditional variational auto-encoder. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11645–11655 (2021) 3
19. Chen, L.H., Zhang, J., Li, Y., Pang, Y., Xia, X., Liu, T.: Humanmac: Masked motion completion for human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9544–9555 (October 2023) 2, 4, 6, 10, 11
20. Chiu, H.k., Adeli, E., Wang, B., Huang, D.A., Niebles, J.C.: Action-agnostic human pose forecasting. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1423–1432. IEEE (2019) 3
21. Corona, E., Pumarola, A., Alenya, G., Moreno-Noguer, F.: Context-aware human motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6992–7001 (2020) 3
22. Croitoru, F.A., Hondru, V., Ionescu, R.T., Shah, M.: Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(9), 10850–10869 (Sep 2023). https://doi.org/10.1109/tpami.2023.3261988, http://dx.doi.org/10.1109/TPAMI.2023.3261988 2
23. Cui, Q., Sun, H.: Towards accurate 3d human motion prediction from incomplete observations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4801–4810 (2021) 1
24. Cui, Q., Sun, H., Yang, F.: Learning dynamic relationships for 3d human motion prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6519–6527 (2020) 1
25. Dang, L., Nie, Y., Long, C., Zhang, Q., Li, G.: Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11467–11476 (2021) 3
26. Dang, L., Nie, Y., Long, C., Zhang, Q., Li, G.: Diverse human motion prediction via gumbel-softmax sampling from an auxiliary space. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 5162–5171 (2022) 1, 3, 8, 10
27. Dang, L., Nie, Y., Long, C., Zhang, Q., Li, G.: Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11467–11476 (2021) 1, 10
28. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: Advances in Neural Information Processing Systems. vol. 34, pp. 8780–8794 (2021) 2
29. Dilokthanakul, N., Mediano, P.A., Garnelo, M., Lee, M.C., Salimbeni, H., Arulkumaran, K., Shanahan, M.: Deep unsupervised clustering with gaussian mixture variational autoencoders. arXiv preprint arXiv:1611.02648 (2016) 10

30. Fragkiadaki, K., Levine, S., Felsen, P., Malik, J.: Recurrent network models for human dynamics. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4346–4354 (2015) 1
31. Fragkiadaki, K., Levine, S., Felsen, P., Malik, J.: Recurrent network models for human dynamics. In: Proceedings of the IEEE international conference on computer vision. pp. 4346–4354 (2015) 3, 10
32. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014) 1
33. Gu, C., Yu, J., Zhang, C.: Learning disentangled representations for controllable human motion prediction. Pattern Recognition **146**, 109998 (2024) 3
34. Guo, W., Du, Y., Shen, X., Lepetit, V., Alameda-Pineda, X., Moreno-Noguer, F.: Back to mlp: A simple baseline for human motion prediction. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4809–4819 (2023) 3, 4
35. Gurumurthy, S., Kiran Sarvadevabhatla, R., Venkatesh Babu, R.: Deligan: Generative adversarial networks for diverse and limited data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 166–174 (2017) 1
36. Gurumurthy, S., Kiran Sarvadevabhatla, R., Venkatesh Babu, R.: Deligan: Generative adversarial networks for diverse and limited data. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 166–174 (2017) 10
37. Han, D., Pan, X., Han, Y., Song, S., Huang, G.: Flatten transformer: Vision transformer using focused linear attention (2023) 5
38. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851 (2020) 2, 6
39. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE transactions on pattern analysis and machine intelligence **36**(7), 1325–1339 (2013) 8
40. Jain, A., Zamir, A.R., Savarese, S., Saxena, A.: Structural-rnn: Deep learning on spatio-temporal graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5308–5317 (2016) 1
41. Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F.: Transformers are rnns: Fast autoregressive transformers with linear attention. In: International Conference on Machine Learning. pp. 5156–5165. PMLR (2020) 5
42. Khayam, S.A.: The discrete cosine transform (dct): theory and application. Michigan State University **114**(1), 31 (2003) 4
43. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: arXiv preprint arXiv:1412.6980 (2014) 9
44. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 1
45. Kundu, J.N., Gor, M., Babu, R.V.: Bihmp-gan: Bidirectional 3d human motion prediction gan. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8553–8560 (2019) 3
46. Lee, M.L., Behdad, S., Liang, X., Zheng, M.: Task allocation and planning for product disassembly with human–robot collaboration. In: Robotics and Computer-Integrated Manufacturing. vol. 76, p. 102306 (2022) 1

47. Lee, M.L., Liu, W., Behdad, S., Liang, X., Zheng, M.: Robot-assisted disassembly sequence planning with real-time human motion prediction. In: IEEE Transactions on Systems, Man, and Cybernetics: Systems. vol. 53, pp. 438–450 (2022) 1
48. Li, B., Tian, J., Zhang, Z., Feng, H., Li, X.: Multitask non-autoregressive model for human motion prediction. IEEE Transactions on Image Processing **30**, 2562–2574 (2020) 1
49. 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) 3
50. Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., Tian, Q.: Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 1
51. 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) 1
52. Li, R., Su, J., Duan, C., Zheng, S.: Linear attention mechanism: An efficient attention for semantic segmentation (2020), https://arxiv.org/abs/2007.14902 5
53. 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) 3, 10
54. Liu, W., Liang, X., Zheng, M.: Dynamic model informed human motion prediction based on unscented kalman filter. In: IEEE/ASME Transactions on Mechatronics. vol. 27, pp. 5287–5295 (2022) 1
55. Liu, Z., Lyu, K., Wu, S., Chen, H., Hao, Y., Ji, S.: Aggregated multi-gans for controlled 3d human motion prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 2225–2232 (2021) 3
56. Liu, Z., Su, P., Wu, S., Shen, X., Chen, H., Hao, Y., Wang, M.: Motion prediction using trajectory cues. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13299–13308 (2021) 1
57. Liu, Z., Wu, S., Jin, S., Liu, Q., Ji, S., Lu, S., Cheng, L.: Investigating pose representations and motion contexts modeling for 3d motion prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022) 3
58. 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) 2, 3
59. Lyu, K., Liu, Z., Wu, S., Chen, H., Zhang, X., Yin, Y.: Learning human motion prediction via stochastic differential equations. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 4976–4984 (2021) 3
60. Ma, H., Li, J., Hosseini, R., Tomizuka, M., Choi, C.: Multi-objective diverse human motion prediction with knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8161–8171 (2022) 1
61. Ma, T., Nie, Y., Long, C., Zhang, Q., Li, G.: Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6437–6446 (June 2022) 1

62. Mao, W., Liu, M., Salzmann, M.: History repeats itself: Human motion prediction via motion attention. In: Computer Vision–ECCV 2020: 16th European Conference. vol. 16, pp. 474–489 (2020) 3, 4
63. Mao, W., Liu, M., Salzmann, M.: Generating smooth pose sequences for diverse human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13309–13318 (2021) 1, 4
64. Mao, W., Liu, M., Salzmann, M., Li, H.: Learning trajectory dependencies for human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9489–9497 (2019) 3, 4
65. Mao, W., Liu, M., Salzmann, M.: History repeats itself: Human motion prediction via motion attention. In: European Conference on Computer Vision. pp. 474–489. Springer (2020) 1
66. Mao, W., Liu, M., Salzmann, M.: Generating smooth pose sequences for diverse human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13309–13318 (2021) 3, 8, 10
67. Mao, W., Liu, M., Salzmann, M., Li, H.: Learning trajectory dependencies for human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9489–9497 (2019) 1, 10
68. Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2891–2900 (2017) 3
69. Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2891–2900 (2017) 3
70. Martínez-González, A., Villamizar, M., Odobez, J.M.: Pose transformers (potr): Human motion prediction with non-autoregressive transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2276–2284 (2021) 1
71. Martínez-González, A., Villamizar, M., Odobez, J.M.: Pose transformers (potr): Human motion prediction with non-autoregressive transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2276–2284 (2021) 3
72. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171 (2021) 7, 9
73. Pascanu, R., Mikolov, T., Bengio, Y.: On the difficulty of training recurrent neural networks. In: International Conference on Machine Learning. pp. 1310–1318 (2013) 3
74. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems. vol. 32 (2019) 9
75. Pavllo, D., Grangier, D., Auli, M.: Quaternet: A quaternion-based recurrent model for human motion. In: arXiv preprint arXiv:1805.06485 (2018) 1
76. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer (2017) 6
77. Qin, Z., Sun, W., Deng, H., Li, D., Wei, Y., Lv, B., Yan, J., Kong, L., Zhong, Y.: cosformer: Rethinking softmax in attention. In: International Conference on Learning Representations (2022) 5
78. 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) 2

79. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation (2015) 6
80. Saadatnejad, S., Rasekh, A., Mofayezi, M., Medghalchi, Y., Rajabzadeh, S., Mordan, T., Alahi, A.: A generic diffusion-based approach for 3d human pose prediction in the wild. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 8246–8253 (2023) 2, 4, 10
81. Sajedi, S., Liu, W., Eltouny, K., Behdad, S., Zheng, M., Liang, X.: Uncertainty-assisted image-processing for human-robot close collaboration. In: IEEE Robotics and Automation Letters. vol. 7, pp. 4236–4243 (2022) 1
82. Sang, H.F., Chen, Z.Z., He, D.K.: Human motion prediction based on attention mechanism. Multimedia Tools and Applications **79**(9), 5529–5544 (2020) 3
83. Si, C., Huang, Z., Jiang, Y., Liu, Z.: Freeu: Free lunch in diffusion u-net (2023), https://arxiv.org/abs/2309.11497 6
84. Sigal, L., Balan, A.O., Black, M.J.: Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International journal of computer vision **87**(1), 4–27 (2010) 8
85. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: arXiv preprint arXiv:2010.02502 (2020) 3, 9
86. Song, S., Lan, C., Xing, J., Zeng, W., Liu, J.: An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: Proceedings of the AAAI conference on artificial intelligence. vol. 31 (2017) 3
87. Su, P., Liu, Z., Wu, S., Zhu, L., Yin, Y., Shen, X.: Motion prediction via joint dependency modeling in phase space. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 713–721 (2021) 1
88. Tang, J., Sun, J., Lin, X., Zheng, W.S., Hu, J.F., et al.: Temporal continual learning with prior compensation for human motion prediction. Advances in Neural Information Processing Systems **36** (2024) 3
89. Tanke, J., Zaveri, C., Gall, J.: Intention-based long-term human motion anticipation. In: 2021 International Conference on 3D Vision (3DV). pp. 596–605. IEEE (2021) 3
90. Tian, S., Liang, X., Zheng, M.: An optimization-based human behavior modeling and prediction for human-robot collaborative disassembly. In: 2023 American Control Conference (ACC). pp. 3356–3361 (2023) 1
91. Tian, S., Zheng, M., Liang, X.: Transfusion: A practical and effective transformer-based diffusion model for 3d human motion prediction (2023) 4, 6, 10
92. 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) 5
93. Walker, J., Marino, K., Gupta, A., Hebert, M.: The pose knows: Video forecasting by generating pose futures. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3332–3341 (2017) 1
94. Walker, J., Marino, K., Gupta, A., Hebert, M.: The pose knows: Video forecasting by generating pose futures. In: Proceedings of the IEEE international conference on computer vision. pp. 3332–3341 (2017) 10
95. Wang, J., Xu, H., Narasimhan, M., Wang, X.: Multi-person 3d motion prediction with multi-range transformers. In: Advances in Neural Information Processing Systems. vol. 34, pp. 6036–6049 (2021) 3, 4
96. 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) 2, 3, 4, 10

97. Xu, S., Wang, Y.X., Gui, L.Y.: Diverse Human Motion Prediction Guided by Multi-level Spatial-Temporal Anchors, p. 251–269. Springer Nature Switzerland (2022). https://doi.org/10.1007/978-3-031-20047-2_15, http://dx.doi.org/10.1007/978-3-031-20047-2_15 3, 10, 11
98. Yan, X., Rastogi, A., Villegas, R., Sunkavalli, K., Shechtman, E., Hadap, S., Lee, H.: Mt-vae: Learning motion transformations to generate multimodal human dynamics. In: Proceedings of the European conference on computer vision (ECCV). pp. 265–281 (2018) 1
99. Yan, X., Rastogi, A., Villegas, R., Sunkavalli, K., Shechtman, E., Hadap, S., Yumer, E., Lee, H.: Mt-vae: Learning motion transformations to generate multimodal human dynamics. In: Proceedings of the European conference on computer vision (ECCV). pp. 265–281 (2018) 3, 10
100. Yuan, Y., Kitani, K.: Diverse trajectory forecasting with determinantal point processes. In: arXiv preprint arXiv:1907.04967 (2019) 1
101. Yuan, Y., Kitani, K.: Diverse trajectory forecasting with determinantal point processes. arXiv preprint arXiv:1907.04967 (2019) 10
102. Yuan, Y., Kitani, K.: Dlow: Diversifying latent flows for diverse human motion prediction. In: European Conference on Computer Vision. pp. 346–364. Springer (2020) 3, 7, 8, 10
103. Zhang, X., Yi, D., Behdad, S., Saxena, S.: Unsupervised human activity recognition learning for disassembly tasks. In: IEEE Transactions on Industrial Informatics (2023) 1
104. Zhang, Y., Black, M.J., Tang, S.: We are more than our joints: Predicting how 3d bodies move. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3372–3382 (2021) 1, 4
105. Zhao, G., Lin, J., Zhang, Z., Ren, X., Su, Q., Sun, X.: Explicit sparse transformer: Concentrated attention through explicit selection. arXiv preprint arXiv:1912.11637 (2019) 5