# SPARD: Single-step Inference with Adaptive Sampling in Residual Diffusion for Human Motion Prediction

**Yiming Zhang[1], Baojia Han[1], Ximing Li[1, 2], Wei Pang[3], Fausto Giunchiglia[4], Xiaoyue Feng[1*], Renchu Guan[1*]**

[1]Key Laboratory of Symbolic Computation and Knowledge Engineering of the Ministry of Education, College of Computer Science and Technology, Jilin University
[2]RIKEN Center for Advanced Intelligence Project, Japan
[3]School of Mathematical and Computer Sciences, Heriot-Watt University
[4]Department of Information Engineering and Computer Science, University of Trento
{fengxy, guanrenchu}@jlu.edu.cn

## Abstract

The task of stochastic human motion prediction has attracted significant attention in recent years due to its wide-ranging applications in robotics, animation, and human-computer interaction. While diffusion models have demonstrated promising progress in this domain, they remain hindered by two critical limitations: (1) slow inference speeds due to their reliance on iterative sampling, and (2) performance degradation resulting from suboptimal sample allocation during generation. To overcome these challenges, we propose SPARD (Single-step Inference with Adaptive Sampling in Residual Diffusion for Human Motion Prediction), a novel framework that achieves efficient single-step inference while maintaining high predictive accuracy. Furthermore, we introduce a novel adaptive noise predictor module that dynamically samples latent representations based on observed motion sequences, ensuring both accuracy and plausibility in generated motions. Extensive experiments on benchmark datasets demonstrate that SPARD significantly outperforms state-of-the-art methods in both inference efficiency and motion quality, achieving a 15× to 18× speedup in sampling time compared to conventional diffusion-based baselines while preserving generation quality.

**Code** — https://github.com/ZYM-JLU/SPARD

## Introduction

The task of human motion prediction (HMP) aims to forecast future human motion sequences based on previously observed ones. With the rapid advancement in deep learning technologies, HMP has gained much attention and achieved significant progress in various domains (Barsoum, Kender, and Liu 2018; Ju et al. 2023; Xu et al. 2023; Zhang et al. 2019; Cen et al. 2024; Zhang, Black, and Tang 2021), including autonomous driving (Paden et al. 2016), animation production (Guo et al. 2020), and human-computer interaction (Tian, Liang, and Zheng 2023; Zhang et al. 2023). However, due to the highly complex and diverse nature of human movements, achieving accurate human motion prediction remains a significant challenge.

*Corresponding Author.


Over the last decade, the field of HMP has undergone a paradigm shift from deterministic models (Fragkiadaki et al. 2015; Cai et al. 2020) to stochastic models (Kundu, Gor, and Babu 2019; Yuan and Kitani 2020b). To effectively model the uncertainty and diversity of human motion, researchers have introduced deep generative models. For example, Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) have been widely adopted to represent the predictive distribution of human motion, enabling the generation of multiple plausible motion sequences. With the advances in generative modeling techniques, diffusion models have demonstrated remarkable success in tasks such as image generation (Ho, Jain, and Abbeel 2020), text-to-image generation (Rombach et al. 2022), and video generation (Ho et al. 2022). Recently, conditional denoising diffusion models have been applied to HMP tasks (Wei et al. 2023; Chen et al. 2023; Barquero, Escalera, and Palmero 2023; Sun and Chowdhary 2024), showing promising performance in capturing multimodal motion dynamics.

However, denoising diffusion models still face two major limitations in HMP. First, the inference process relies on multistep iterative sampling, typically requiring 10 to 100 steps. This requires tens to hundreds of milliseconds for inference time, making these models significantly less efficient compared to VAE-based and GAN-based models. Second, in conditional diffusion models, human motion diversity is primarily modeled through random sampling from a Gaussian distribution during inference. As observed in trajectory prediction (Mao et al. 2023), a limited number of independent and identically distributed samples may fail to adequately capture the full range of human motion patterns. The lack of appropriate sample allocation during independent sampling may cause the model to miss important motion modes, leading to significant degradation in prediction performance.

To address the two key limitations in conditional denoising diffusion models for trajectory prediction, Mao et al. (Mao et al. 2023) proposed the LEapfrog Diffusion model (LED), which replaces Gaussian noise sampling during reverse inference with an initialization trajectory network. However, LED still suffers from several limitations. First, the initialization trajectory is generated based on historical

observations. Consequently, when reducing the number of inference steps, the initialization trajectory must closely approximate the ground truth. This requirement forces the initialization trajectory network to function as a direct predictor, increasing its training difficulty and constraining the expressive capability of the diffusion network. Second, while LED reduces the inference steps compared to standard diffusion models, it still requires at least five iterative diffusion steps, resulting in suboptimal computational efficiency for real-time applications.

To address these challenges, we propose the SPARD (Single-step Inference with Adaptive Sampling in Residual Diffusion for Human Motion Prediction) model, a novel diffusion-based framework for human motion prediction. SPARD offers several key advantages: First, it achieves accurate predictions in a single-step inference process, making its inference speed comparable to VAE-based models and significantly faster than traditional diffusion-based models. Second, we introduce a novel noise predictor module that enables adaptive sampling from the Gaussian noise space based on observed motion. The core innovation of SPARD lies in modeling the residual between observed and predicted human motions. SPARD uses observed human motion as the starting point of the diffusion process and incorporates a noise mechanism to regulate the diversity of predictions. By focusing on the residual, SPARD significantly reduces the number of diffusion steps required. To further improve inference speed, a student model is trained via knowledge distillation, enabling single-step inference. Combined with the noise predictor module, SPARD allows adaptive sampling based on observed motion, overcoming the limitations of traditional independent and identically distributed Gaussian sampling. Notably, with the introduction of the trainable noise predictor module, our approach only requires training this module to adapt to different numbers of predictions, without the need to retrain the entire model to achieve good results, as in other approaches.

The main contributions of this research are summarized as follows:

- We propose SPARD, an efficient residual diffusion framework for human motion prediction that combines knowledge distillation with single-step inference, achieving inference speed comparable to VAE-based models and significantly outperforming existing diffusion-based models.
- We propose a novel noise predictor module that performs adaptive sampling of observed human motions, achieving significant improvements in prediction accuracy.
- We conduct extensive experiments on benchmark datasets, showing that the proposed method outperforms existing baselines. Furthermore, ablation studies validate the effectiveness of the proposed model.

# Related Work

## Human Motion Prediction

Early research on human motion prediction primarily focused on deterministic models, employing regression models such as RNNs (Fragkiadaki et al. 2015; Gui et al. 2018; Martinez, Black, and Romero 2017), Transformers (Mao, Liu, and Salzmann 2020; Aksan et al. 2021), and Graph Convolutional Networks (GCNs) (Dang et al. 2021; Li et al. 2020; Mao et al. 2019) to predict the most likely pose sequences. However, as researchers recognized the inherent stochasticity of human motion, the field shifted toward probabilistic approaches, with GANs and VAEs emerging as mainstream solutions that generated diverse predictions through latent space sampling. The recent success of diffusion models led to their adoption in this domain, beginning with MotionDiff (Wei et al. 2023) which pioneered a two-stage training strategy: the first stage utilized a transformer-based noise prediction module to predict noise, while the second stage fine-tuned the diffusion model's output. To address motion consistency issues, Belfusion (Barquero, Escalera, and Palmero 2023) first trained a VAE-based latent space for motion representation and then employed a UNet-based network to predict noise in this latent space. HumanMAC (Chen et al. 2023) introduced an end-to-end model that formulated motion prediction as a masked completion problem in the Discrete Cosine Transform (DCT) space, which improved training efficiency and simplified the training process. Comusion (Sun and Chowdhary 2024) modified the loss function to directly predict future human motions instead of noise. By incorporating GCN-DCT methods for refinement, Comusion transformed MotionDiff's two-stage training strategy into an end-to-end training process, which simplified the workflow and improved prediction accuracy. Despite the notable advancements of conditional diffusion models in human motion prediction, two major limitations remain:

- Long inference time. Existing diffusion-based models typically require 10 to 100 sampling steps, limiting their applicability in scenarios with strict time constraints.
- Independent and identically distributed (i.i.d.) sampling. Diffusion-based models sample from Gaussian noise distributions in an i.i.d. manner, leading to inadequate sample allocation that may cause the model to miss potentially important motion modes, resulting in a degradation of prediction performance.

## Residual Diffusion Model

The Residual Diffusion Model (Yue, Wang, and Loy 2023) was initially proposed for image super-resolution tasks. Traditional conditional diffusion models establish a mapping between high-resolution images and Gaussian noise by progressively adding noise to the high-resolution image, while the low-resolution image serves merely as a conditional signal to implicitly guide the reconstruction during the reverse generation process. However, the significant distributional discrepancy between Gaussian noise and the high-resolution image necessitates a large number of diffusion steps (typically 1,000 iterations), leading to prolonged inference time. Yue et al. (Yue, Wang, and Loy 2023) introduced Reshift, a novel approach that directly models the residual diffusion process between low-resolution and high-resolution images. Unlike diffusion models that use Gaussian noise as the diffusion endpoint, Reshift treats the noisy low-resolution (LR)

image itself as the final state, substantially simplifying the learning process. This innovation enables high-quality reconstruction with only a few inference steps (e.g., 15 steps), significantly reducing computational overhead and accelerating inference. The method employs a direct image reconstruction loss between the predicted and ground-truth images. RDDM (Liu et al. 2024) further refines the framework by decomposing the loss into two components: residual diffusion and noise diffusion. The residual diffusion component governs the deterministic aspects of the model, ensuring close alignment between the generated results and the low-resolution inputs; while the noise diffusion component regulates diversity, enabling high-quality output generation with preserved variability. ReshiftL (Yue, Wang, and Loy 2024) incorporates perceptual regularization into the Reshift framework, reducing the required diffusion steps to just four while achieving state-of-the-art performance in tasks such as image inpainting and deblurring. Due to the residual diffusion model's efficient diffusion process and flexible generation capabilities, it demonstrates promising potential across a wide range of conditional generation tasks.

## Methods

To address the critical challenges of computational efficiency in diffusion modeling and controlled sampling in motion prediction, we propose a novel three-stage training framework that incorporates three core modules: the residual diffusion module, model distillation, and the noise predictor module. The following sections provide a detailed introduction to each component.

### Problem Definition

HMP tasks aim to forecast future human motion sequences based on past observations. In a stochastic HMP task, given a historical human pose sequence of $N$ frames, denoted as $\mathbf{X}_{1:N} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \in \mathbb{R}^{N\times 3J}$, where $J$ represents the number of human joints, the objective is to predict $K$ possible future sequences, each consisting of $F$ frames, with one prediction represented as $\hat{\mathbf{X}}_{N+1:N+F} = \{\mathbf{x}_{N+1}, \mathbf{x}_{N+2}, \cdots, \mathbf{x}_{N+F}\} \in \mathbb{R}^{F\times 3J}$.

### Latent Space

Diffusion-based models like MotionDiff (Wei et al. 2023) and Comusion (Sun and Chowdhary 2024) perform diffusion process in the temporal domain of human motion. However, due to the long motion sequences, the diffusion space may become large, leading to inefficiencies in both training and inference. Belfusion (Barquero, Escalera, and Palmero 2023) introduces a separately trained encoder-decoder to embed the original feature into a latent space, reducing the diffusion dimension. However, this approach requires an additional training process. Inspired by previous work (Mao et al. 2019), we apply the discrete cosine transform (DCT) to convert the time domain of human motion into the frequency domain, reducing dimensionality and obtaining a more compact representation of motion. Since the most important features of human motion are primarily concentrated in low-frequency signals, we model with only the first $l$ rows of the DCT basis. Formally, for an observed human motion sequence $\mathbf{X}$, we apply the DCT operation to obtain:

$$\mathbf{Z} = \mathrm{DCT}(\mathbf{X}). \tag{1}$$

Since DCT operation is an orthogonal transformation, the original motion sequence can be reconstructed from the frequency domain using the inversed DCT (iDCT) operation:

$$\mathbf{X} = \mathrm{iDCT}(\mathbf{Z}). \tag{2}$$

### Residual Diffusion Module

We define a complete sequence of human motion as $\mathbf{X}_0 = \{\mathbf{x}_1, \cdots, \mathbf{x}_N, \mathbf{x}_{N+1}, \cdots, \mathbf{x}_{N+F}\} \in \mathbb{R}^{(N+F)\times 3J}$. The observed human motion sequence $\mathbf{x}_{1:N}$ is treated as the conditioning input. To match the length of $\mathbf{X}_0$, we extend the last observed frame $\mathbf{x}_N$ by repeating it F times, denoted as $\mathbf{Y} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N, \mathbf{x}_N, \cdots, \mathbf{x}_N\} \in \mathbb{R}^{(N+F)\times 3J}$. We apply the DCT operation to the complete motion sequence $\mathbf{X}_0$ and the conditioning sequence $\mathbf{Y}$ to obtain the corresponding DCT coefficients $\mathbf{Z}_0$ and $\mathbf{Z}_y$. We define the residual between observed and complete human motion in DCT space as: $\mathbf{E}_0 = \mathbf{Z}_y - \mathbf{Z}_0$. The transition distribution at step $t$ is modeled as follows:

$$q(\mathbf{Z}_t|\mathbf{Z}_0, \mathbf{Z}_y) = \mathcal{N}\left(\mathbf{Z}_t; \mathbf{Z}_0 + \eta_t\mathbf{E}_0, \kappa^2\eta_t\mathbf{I}\right), t = 1, \cdots, T, \tag{3}$$

where the shifting sequence $\{\eta_t\}_{t=1}^T$ controls the noise schedule in residual diffusion process, satisfying $\eta_1 \to 0$ and $\eta_T \to 1$. The parameter $\kappa$ is a hyperparameter that adjusts the noise variance, and $\mathbf{I}$ is the identity matrix.

Our residual diffusion module $\mathcal{G}_\theta$, which is parameterized by $\theta$, consists of two key components: a temporal network $d$ and a spatial network $r$. The temporal network $d$ consists of multiple stacked TransLinear blocks (Chen et al. 2023) to extract temporal features from the latent representation $\mathbf{Z}_t$. To better capture spatial coherence and joint dependencies, the spatial network $r$ adopts a GCN-DCT (Mao et al. 2019) architecture, which embeds temporal features into DCT space and employs GCN to capture spatial dependencies among human joints. The noisy input $\mathbf{Z}_t$ first passes through the temporal network $d$ to capture temporal dependencies. We then concatenate the historical observed motions with the temporal network $d$'s predictions along the time dimension and feed them into the spatial network $r$, yielding the final prediction of the residual diffusion module $\mathcal{G}_\theta(\mathbf{Z}_t, t, \mathbf{Z}_y)$. At timestep $t$, the output of the residual diffusion module $\mathcal{G}_\theta(\mathbf{Z}_t, t, \mathbf{Z}_y)$ is directly optimized to match $\mathbf{Z}_0$. To avoid additional loss during domain transformation, we optimize $\theta$ in the original motion space by applying the iDCT operation. The training objective is defined as follows:

$$L_\theta = \mathbb{E}_{\substack{\mathbf{X}_0\sim q(\mathbf{X}_0|\mathbf{Y}) \\ t\sim[1,T]}}[\|\mathbf{X}_0 - \mathrm{iDCT}(\mathcal{G}_\theta(\mathbf{Z}_t, t, \mathbf{Z}_y))\|_1]. \tag{4}$$

Additionally, we adopt a classifier-free guidance approach (Ho and Salimans 2021) to train the residual diffusion module $\mathcal{G}_\theta$, enabling flexible control over conditioning observation $\mathbf{Z}_y$ and the diversity of the generated predictions.

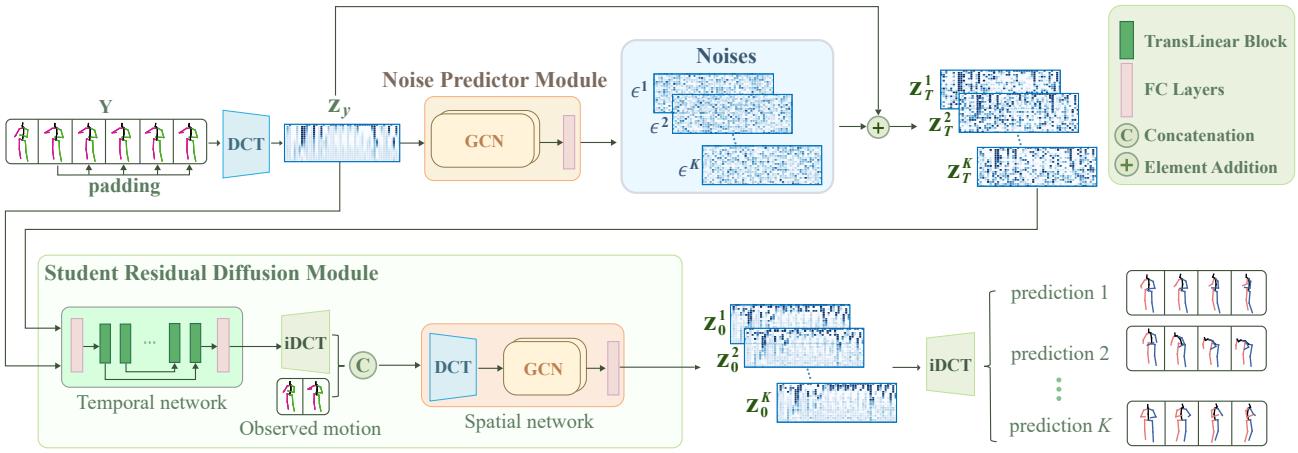


Figure 1: The inference procedure of our SPARD model.

Specifically, the model output $\mathcal{G}_\theta^s(\mathbf{Z}_t, t, \mathbf{Z}_y)$ is formulated as a weighted combination of two components:

$$\mathcal{G}_\theta^s(\mathbf{Z}_t, t, \mathbf{Z}_y) = s\mathcal{G}_\theta(\mathbf{Z}_t, t, \mathbf{Z}_y) + (1-s)\mathcal{G}_\theta(\mathbf{Z}_t, t, \emptyset), \quad (5)$$

where hyperparameter $s$ adjusts the influence of $\mathbf{Z}_y$, allowing flexible control over the generated motion sequences, with $\emptyset$ denoting the absence of input. To further enhance prediction diversity, we incorporate a relaxation technique during training, where the model generates $h$ predictions, and the closest prediction to the ground truth is selected for backpropagation. The training process of Residual Diffusion Module is shown in Algorithm 1.

## Model Distillation

Given an observed human motion sequence $\mathbf{X}_{1:N}$, the residual diffusion module iteratively generates $K$ predicted human motion sequences during the inference process. We refrain from employing the reverse inference scheme used in ResShift (Yue, Wang, and Loy 2023), as it introduces random noise during sampling, which negatively impacts model distillation. Inspired by previous work (Wang et al. 2024), we adopt a deterministic sampling strategy, formulated as follows:

$$\mathbf{Z}_{t-1} = k_t\hat{\mathbf{Z}}_0 + m_t\mathbf{Z}_t + j_t\mathbf{Z}_T, \quad (6)$$

$$\begin{cases} m_t = \sqrt{\frac{\eta_{t-1}}{\eta_t}}, \\ j_t = \eta_{t-1} - \sqrt{\eta_{t-1} - \eta_t}, \\ k_t = 1 - \eta_{t-1} + \sqrt{\eta_{t-1}\eta_t} - \sqrt{\frac{\eta_{t-1}}{\eta_t}}. \end{cases}$$

This deterministic sampling establishes a deterministic mapping between the predictions of the residual diffusion module $\hat{\mathbf{Z}}_0$ and $\mathbf{Z}_T$, generated from Gaussian noise and conditional $\mathbf{Z}_y$. To further accelerate the sampling process, we propose a student residual diffusion module $\mathcal{F}_{\hat{\theta}}$ to learn the deterministic mapping. The loss function for model distillation is defined as follows:

$$L_{\text{distill}} = ||\hat{\mathbf{Z}}_0 - \mathcal{F}_{\hat{\theta}}(\mathbf{Z}_T, T, \mathbf{Z}_y)||_1, \quad (7)$$

where $\mathcal{F}_{\hat{\theta}}(\mathbf{Z}_T, T, \mathbf{Z}_y)$ represents the one-step prediction output of the student module. Since the teacher residual diffusion module only requires a few inference steps (5 steps) to obtain $\hat{\mathbf{Z}}_0$, the computational overhead for teacher module's inference during training remains acceptable. Although distillation may result in a slight decrease in the student module's prediction performance due to the limitations of the

**Algorithm 1:** The training process of Residual Diffusion Module

**Input:** A complete motion sequence $\mathbf{X}_0$, residual diffusion module $\mathcal{G}_\theta$ (consisting of the temporal network $d$ and the spatial network $r$), the number of diffusion steps $T$, hyperparameter $\kappa$, shift squence $\{\eta_t\}_{t=1}^T$ and the number of relaxation predictions $h$.

**Output:** residual diffusion module $\mathcal{G}_\theta$

**repeat**

$\mathbf{X}_{1:N+F} \sim q(\mathbf{X}_0)$
$\mathbf{Z}_0 = \text{DCT}(\mathbf{X}_0)$
$\mathbf{Z}_y = \text{DCT}(\text{Padding}(\mathbf{X}_{1:N}))$
$\mathbf{E}_0 = \mathbf{Z}_y - \mathbf{Z}_0$
$t \sim \text{Uniform}(\{1, \dots, T\})$
$\epsilon^{1:h} \sim N(0, I)$
$\mathbf{Z}_t^{1:h} = \mathbf{Z}_0 + \eta_t\mathbf{E}_0 + \kappa\sqrt{\eta_t}\epsilon^{1:h}$
$\mathbf{Z}_y \leftarrow \emptyset$ with probability $50\%$
$\hat{\mathbf{Z}}_0^{1:h} = d(\mathbf{Z}_t^{1:h}, t, \mathbf{Z}_y)$
$\hat{\mathbf{Z}}_0^{1:h} =$
$r(\text{Concat}(\mathbf{X}_{1:N}, \text{iDCT}(\hat{\mathbf{Z}}_0^{1:h})_{N+1:N+F}))$
$\theta = \theta - \nabla_\theta \min_{i\in\{1,2,\dots,h\}} \left\|\mathbf{X}_0 - \text{iDCT}\left(\hat{\mathbf{Z}}_0^i\right)\right\|_1$

**until** *convergence*;

teacher module, this trade-off is justified by the significant improvement in inference speed and the efficient training of the noise predictor module.

### Noise Predictor Module

To regulate the randomness of the human motion predicted by the student residual diffusion module, we first revisit the model's inference process. Given the observed human motion sequence $\mathbf{Y}$, we project it into the DCT domain to obtain $\mathbf{Z}_y$. Next, Gaussian noise is randomly added to $\mathbf{Z}_y$ to produce the latent variable $\mathbf{Z}_T$. The student model $\mathcal{F}_{\hat{\theta}}$ then performs deterministic single-step predictions, where different values of $\mathbf{Z}_T$ lead to corresponding human motion predictions $\hat{\mathbf{Z}}_0$. In the residual diffusion-based architecture, the deterministic component of the prediction is driven by the observed human motion embedding $\mathbf{Z}_y$, while the added Gaussian noise introduces diversity and randomness into the generated motions. We parameterize the Gaussian noise sampling process into a learnable process to regulate the randomness of human motion predictions. We propose a novel noise predictor module $\mathcal{M}_\phi$ to model this process. For each observed human action, the noise predictor module $\mathcal{M}_\phi$ generates K noise samples, which are then used to generate K predicted actions through the student residual diffusion model. Specifically, the noise predictor $\mathcal{M}_\phi$ takes the embedded human motion sequence $\mathbf{Z}_y$ as input and outputs a set of noise samples $[\epsilon^1, \epsilon^2, \ldots, \epsilon^K]$. The noise predictor adopts a simple network architecture: it first employs a GCN network to model spatial dependencies between human joints, followed by a fully connected layer to adjust the output to match the target distribution. Formally, this can be expressed as follows:

$$\epsilon^{1:K} = \alpha \cdot \tanh(\text{GCN}(\mathbf{Z}_y)). \tag{8}$$

Since we simulate sampling from a standard normal distribution Gaussian noise, we apply the $3\sigma$ principle for Gaussian noise, setting the hyperparameter $\alpha = 3$ and using the tanh activation function to constrain the output within the range of $[-3, 3]$. The noise predictor loss is defined as follows:

$$L_{\text{noise\_predictor}} = \min_{i \in \{1,2,\ldots,K\}} ||\mathbf{X}_0 - \hat{\mathbf{X}}_0^i||_1. \tag{9}$$

In the above, $\hat{\mathbf{X}}_0^i$ represents the $i^{th}$ generated motion prediction. This loss ensures that at least one of the generated motion sequences is sufficiently close to the ground-truth human motion. During training, we freeze the parameters of the student model $\mathcal{F}_{\hat{\theta}}$ and only train the noise predictor module $\mathcal{M}_\phi$. The visualization of the inference process is shown in Figure 1.

## Experimental evaluation

### Experimental Setup

**Datasets.** We evaluated our method on two datasets: Human3.6M (Ionescu et al. 2013) and AMASS (Mahmood et al. 2019). To ensure a fair comparison, all experimental settings followed the configurations established in previous work (Barquero, Escalera, and Palmero 2023).

**Evaluation Metrics.** Following previous work (Chen et al. 2023), we use a comprehensive set of metrics to quantitatively evaluate the SPARD model: (1) Average Pairwise Distance (APD) computes the average distance between all pairs of motion samples, measuring the diversity of the predictions. (2) Average Displacement Error (ADE) calculates the average distance between the ground truth and the closest generated motion, assessing the accuracy of the entire sequence. (3) Final Displacement Error (FDE) computes the distance between the predicted result and the ground truth in the final prediction frame, evaluating the precision of the last step. (4) Multi-Modal Average Displacement Error (MMADE) is the multi-modal version of ADE, where the ground truth motions are grouped based on similar observations, assessing the model's ability to generate multi-modal predictions. (5) Multi-Modal Final Displacement Error (MMFDE) is the multi-modal version of FDE, evaluating the accuracy of the last frame in a multi-modal context.

**Baselines** We quantitatively evaluate our proposed approach against several baseline methods, including HP-GAN (Barsoum, Kender, and Liu 2018), DeLiGAN (Gurumurthy, Kiran Sarvadevabhatla, and Venkatesh Babu 2017), DSF (Yuan and Kitani 2020a), GMVAE (Dilokthanakul et al. 2016), MT-VAE (Yan et al. 2018), BoM (Bhattacharyya, Schiele, and Fritz 2018), TPK (Walker et al. 2017), DLow (Yuan and Kitani 2020b), GSPS (Mao, Liu, and Salzmann 2021), DivSamp (Dang et al. 2022), MotionDiff (Wei et al. 2023), HumanMAC (Chen et al. 2023), BeLFusion (Barquero, Escalera, and Palmero 2023), and CoMusion(Sun and Chowdhary 2024). For qualitative analysis, we compare our method with DLow, GSPS, DivSample, and Belfusion.

### Results Compared with State of the Art

**Quantitative Results.** As shown in Table 1, our model achieves state-of-the-art performance on all accuracy metrics in both datasets. Specifically, our ADE improves by 9.14% on Human3.6M and 7.89% on AMASS compared to CoMusion, mainly due to the effectiveness of our residual diffusion module and the noise predictor module. Although the diversity metrics of our model are slightly lower compared to other models, the superior results on MMADE and MMFDE indicate that the quality of the generated motion diversity surpasses that of other baseline models. The reduction in the diversity metric (APD value) may be due to the model distillation and the residual diffusion module, where the deterministic observation embeddings reduce the randomness in the model's predictions. However, we believe this trade-off is necessary, as it strengthens the correlation between the predicted motions and the observed motions, thereby improving the quality of the diverse predictions and avoiding the generation of unreasonable predictions solely for the sake of diversity.

**Qualitative Results.** Figure 2 presents 10 overlaid predictions over time for three actions from Human3.6M (Sitting, Discussion, and Greeting), where the solid lines represent the true results and the faded lines represent the predicted results. As shown in Figure 2, our SPARD model exhibits

| Method | Human3.6M | | | | | AMASS | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ↑ APD | ↓ ADE | ↓ FDE | ↓ MMADE | ↓ MMFDE | ↑ APD | ↓ADE | ↓FDE | ↓MMADE | ↓ MMFDE |
| HP-GAN | 7.214 | 0.858 | 0.867 | 0.847 | 0.858 | - | - | - | - | - |
| DeLiGAN | 6.509 | 0.483 | 0.534 | 0.520 | 0.545 | - | - | - | - | - |
| DSF | 9.330 | 0.493 | 0.592 | 0.550 | 0.599 | - | - | - | - | - |
| GMVAE | 6.769 | 0.461 | 0.555 | 0.524 | 0.566 | - | - | - | - | - |
| MT-VAE | 0.403 | 0.457 | 0.595 | 0.716 | 0.883 | - | - | - | - | - |
| BoM | 6.265 | 0.448 | 0.533 | 0.514 | 0.544 | - | - | - | - | - |
| TPK | 6.723 | 0.461 | 0.560 | 0.522 | 0.569 | 9.283 | 0.656 | 0.675 | 0.658 | 0.674 |
| DLow | 11.741 | 0.425 | 0.518 | 0.495 | 0.531 | 13.170 | 0.590 | 0.612 | 0.618 | 0.617 |
| GSPS | 14.757 | 0.389 | 0.496 | 0.476 | 0.525 | 12.465 | 0.563 | 0.613 | 0.609 | 0.633 |
| DivSamp | 15.310 | 0.370 | 0.485 | 0.475 | 0.516 | **24.724** | 0.564 | 0.647 | 0.623 | 0.667 |
| MotionDiff | **15.353** | 0.411 | 0.509 | 0.508 | 0.536 | - | - | - | - | - |
| HumanMAC | 6.301 | 0.369 | 0.480 | 0.509 | 0.545 | 9.321 | 0.511 | 0.554 | 0.593 | 0.591 |
| BeLFusion | 7.602 | 0.372 | 0.474 | 0.473 | 0.507 | 9.376 | 0.513 | 0.560 | 0.569 | 0.585 |
| CoMusion | 7.632 | 0.350 | 0.458 | 0.494 | 0.506 | 10.848 | 0.494 | 0.547 | 0.469 | 0.466 |
| SPARD | 5.571 | **0.318** | **0.428** | **0.466** | **0.475** | 7.634 | **0.455** | **0.528** | **0.439** | **0.441** |

Table 1: Quantitative results of predicting 50 future motions per historical motion. Bold numbers indicate the best results, and underlined numbers represent the second best results. The symbol '-' indicates that the results are not reported in the baselines.

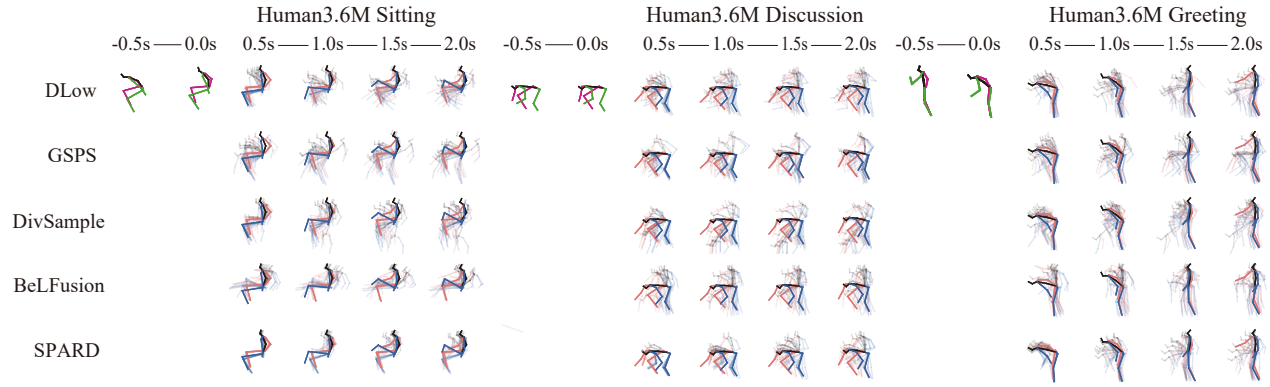


Figure 2: Visualization results of SPARD compared with baseline methods on Human3.6M dataset. Ten randomly generated predictions and ground truth values are stacked per time step. The ground truth is represented by solid lines, while the predictions are shown as faded lines. The green-pink and blue-orange skeletons represent the observed and predicted movements.

close alignment with the ground truth in the early phase (the first 0.5 seconds), demonstrating high accuracy at the start. Over time, the predicted results display an increasing degree of diversity. In contrast, some predictions of other baseline models tend to show noticeable deviations from the ground truth in the early phase. The predictions generated by our model are not only more reasonable, but also better conform to the physical constraints of human motion. This observation to some extent confirms that, while we have reduced the overall diversity of predictions, we have successfully enhanced the quality of the diversity, with the generated actions more closely resembling actual human behavior. This advantage is likely attributable to the deterministic observations in the residual diffusion process and the noise predictor module, which allow the sampling process to adapt to the observed motion patterns rather than relying on random sampling. Interestingly, in the 2-second predictions for the Greeting action, our SPARD model effectively captured the intent behind the actions. The predicted motions spanned the range of possible motion transitions and demonstrated diversity in terms of speed. This characteristic was not observed in the predictions by other models, further emphasizing the superior quality of diversity in our approach.

**Inference time.** We evaluated the generation of 50 predictions for a single sample on a single RTX 4090 GPU with a batch size of 16. Experiments were conducted on the validation set to calculate the average inference time per sample. As shown in Figure 3, HumanMAC exhibited the longest inference time, primarily due to its requirement of 100 inference steps. While both CoMusion and Belfusion required 10 inference steps, Belfusion demonstrated shorter inference time compared to CoMusion. This efficiency arised because Belfusion performs diffusion in the latent space, whereas CoMusion operated in the original action space. In contrast, SPARD, as a single-step prediction approach, achieved significantly faster inference speeds than other diffusion-based methods. Specifically, it achieved an inference speed 15.97 times faster on Human3.6M and 18.21 times faster on AMASS than the fastest diffusion-based Bel-

| Model Distillation | Noise Predictor Module | Human3.6M | | | | | |
|---|---|---|---|---|---|---|---|
| | | ↑ APD | ↓ ADE | ↓ FDE | ↓ MMADE | ↓ MMFDE | ↓ Inference Time |
| × | × | **7.500** | 0.336 | 0.452 | 0.483 | 0.498 | 24.72 |
| ✓ | × | 6.338 | 0.342 | 0.458 | 0.481 | 0.498 | **2.52** |
| ✓ | ✓ | 5.571 | **0.318** | **0.428** | **0.466** | **0.475** | 2.82 |

Table 2: Ablation study on the model distillation and the noise predictor module in Human3.6M. Bold numbers indicate the best results. The unit of inference time is milliseconds.

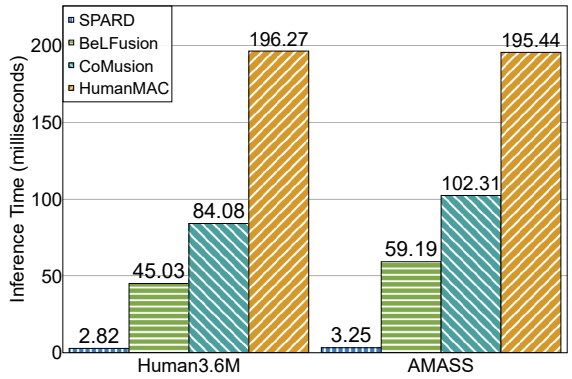


Figure 3: Inference time of diffusion-based methods.

fusion method, and even slightly outperformed the VAE-based DLow method (3.42/4.33ms for H36M/AMASS). This makes our model particularly advantageous in scenarios with strict time constraints.

## Ablation Studies

We conducted systematic ablation studies on model distillation, the noise prediction module in Human3.6M. Table 2 clearly shows that model distillation incurs a slight performance loss in terms of prediction diversity. This may be attributed to the fact that the residual diffusion module of the teacher model adopts a classifier-free sampling strategy, while the residual diffusion module of the student model receives deterministic observation embeddings $\mathbf{Z}_y$, reducing randomness and thereby affecting diversity performance. The student model exhibits some improvement in terms of the MMADE and MMFDE metrics, while showing a slight decline in the accuracy metrics ADE and FDE. Despite the minor loss in some accuracy indicators, the model distillation reduces the inference time by nearly 10 times. We believe this slight performance trade-off is a reasonable compromise in exchange for a significant boost in inference efficiency. Additionally, the noise predictior module greatly enhances the model's prediction accuracy, further demonstrating the necessity of adaptive action sampling and confirming the critical role of the noise predictior module in optimizing model performance.

## Hyperparameter Analysis

We conducted a systematic sensitivity analysis of the classifier-free guidance coefficient $s$ and the relaxation parameter $h$ in the residual diffusion module. As shown in Table 3 and Table 4, the experimental results indicate that both parameters have a significant impact on the diversity and accuracy of the generated results.

**Impact of the Classifier-Free Guidance Coefficient $s$.** When the observed motion embedding $Z_y$ is gradually introduced as a conditional guidance, the diversity of the generated results decreases due to the incorporation of deterministic information, and all accuracy metrics improve. However, as the value of $s$ continues to increase, further reduction in diversity negatively impacts the accuracy metrics. Based on a trade-off analysis between diversity and accuracy, we select $s$ = 0.3 as the optimal parameter.

| $s$ | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 |
|---|---|---|---|---|---|---|---|
| ↑ APD | **8.284** | 7.893 | 7.500 | 7.150 | 6.856 | 6.605 | 6.447 |
| ↓ ADE | 0.344 | 0.339 | 0.336 | 0.335 | **0.334** | 0.335 | 0.335 |
| ↓ FDE | 0.461 | 0.454 | 0.452 | **0.451** | **0.451** | 0.456 | 0.459 |
| ↓ MMADE | 0.484 | **0.483** | **0.483** | 0.486 | 0.489 | 0.493 | 0.498 |
| ↓ MMFDE | 0.499 | **0.498** | 0.499 | 0.502 | 0.508 | 0.517 | 0.526 |

Table 3: Performance metrics for different values of $s$ in Human3.6M. Bold numbers indicate the best results.

**Impact of the Relaxation Parameter $h$.** As shown in Table 4, increasing the value of $h$ results in a noticeable improvement in the diversity of model predictions, while also enhancing the accuracy metrics. However, this comes at the cost of increased training time. Considering the trade-off between training efficiency and generation quality, we choose $h = 10$ as the optimal parameter.

| $h$ | ↑ APD | ↓ ADE | ↓ FDE | ↓ MMADE | ↓ MMFDE | ↓ Training Time |
|---|---|---|---|---|---|---|
| 1 | 3.192 | 0.351 | 0.498 | 0.522 | 0.573 | **39.98** |
| 5 | 6.149 | **0.335** | **0.452** | 0.485 | 0.502 | 66.95 |
| 10 | **7.500** | 0.336 | **0.452** | **0.483** | **0.499** | 113.49 |

Table 4: Performance metrics for different values of $h$ in Human3.6M. Bold numbers indicate the best results. Training time (in seconds) per epoch with a batch size of 64.

## Conclusion

In this paper, we propose a novel SPARD model. Compared to existing diffusion model-based methods in human motion prediction, SPARD achieves efficient single-step inference for the first time, offering an order-of-magnitude improvement in inference speed over traditional iterative sampling methods. Additionally, the model adaptively samples based on observed motions, significantly enhancing prediction accuracy. Both quantitative and qualitative experimental results demonstrate that our method exhibits clear advantages over state-of-the-art approaches.

## Acknowledgments

Our work is supported by the National Science and Technology Major Project under Grant No. 2021ZD0112500, the National Natural Science Foundation of China (No. 62172187 and No. 62372209). Fausto Giunchiglia's work is funded by European Union's Horizon 2020 FET Proactive Project (No. 823783).

## References

Aksan, E.; Kaufmann, M.; Cao, P.; and Hilliges, O. 2021. A spatio-temporal transformer for 3d human motion prediction. In *2021 International Conference on 3D Vision (3DV)*, 565–574. IEEE.

Barquero, G.; Escalera, S.; and Palmero, C. 2023. Belfusion: Latent diffusion for behavior-driven human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2317–2327.

Barsoum, E.; Kender, J.; and Liu, Z. 2018. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 1418–1427.

Bhattacharyya, A.; Schiele, B.; and Fritz, M. 2018. 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*, 8485–8493.

Cai, Y.; Huang, L.; Wang, Y.; Cham, T.-J.; Cai, J.; Yuan, J.; Liu, J.; Yang, X.; Zhu, Y.; Shen, X.; et al. 2020. Learning progressive joint propagation for human motion prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, 226–242. Springer.

Cen, Z.; Pi, H.; Peng, S.; Shen, Z.; Yang, M.; Zhu, S.; Bao, H.; and Zhou, X. 2024. Generating human motion in 3D scenes from text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1855–1866.

Chen, L.-H.; Zhang, J.; Li, Y.; Pang, Y.; Xia, X.; and Liu, T. 2023. Humanmac: Masked motion completion for human motion prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9544–9555.

Dang, L.; Nie, Y.; Long, C.; Zhang, Q.; and Li, G. 2021. Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11467–11476.

Dang, L.; Nie, Y.; Long, C.; Zhang, Q.; and Li, G. 2022. Diverse human motion prediction via gumbel-softmax sampling from an auxiliary space. In *Proceedings of the 30th ACM international conference on multimedia*, 5162–5171.

Dilokthanakul, N.; Mediano, P. A.; Garnelo, M.; Lee, M. C.; Salimbeni, H.; Arulkumaran, K.; and Shanahan, M. 2016. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*.

Fragkiadaki, K.; Levine, S.; Felsen, P.; and Malik, J. 2015. Recurrent network models for human dynamics. In *Proceedings of the IEEE international conference on computer vision*, 4346–4354.

Gui, L.-Y.; Wang, Y.-X.; Liang, X.; and Moura, J. M. 2018. Adversarial geometry-aware human motion prediction. In *Proceedings of the european conference on computer vision (ECCV)*, 786–803.

Guo, C.; Zuo, X.; Wang, S.; Zou, S.; Sun, Q.; Deng, A.; Gong, M.; and Cheng, L. 2020. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2021–2029.

Gurumurthy, S.; Kiran Sarvadevabhatla, R.; and Venkatesh Babu, R. 2017. Deligan: Generative adversarial networks for diverse and limited data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 166–174.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022. Video diffusion models. *Advances in Neural Information Processing Systems*, 35: 8633–8646.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2013. 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.

Ju, X.; Zeng, A.; Wang, J.; Xu, Q.; and Zhang, L. 2023. Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 618–629.

Kundu, J. N.; Gor, M.; and Babu, R. V. 2019. Bihmp-gan: Bidirectional 3d human motion prediction gan. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 8553–8560.

Li, M.; Chen, S.; Zhao, Y.; Zhang, Y.; Wang, Y.; and Tian, Q. 2020. 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*, 214–223.

Liu, J.; Wang, Q.; Fan, H.; Wang, Y.; Tang, Y.; and Qu, L. 2024. Residual denoising diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2773–2783.

Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5442–5451.

Mao, W.; Liu, M.; and Salzmann, M. 2020. History repeats itself: Human motion prediction via motion attention. In

*Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, 474–489. Springer.

Mao, W.; Liu, M.; and Salzmann, M. 2021. Generating smooth pose sequences for diverse human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13309–13318.

Mao, W.; Liu, M.; Salzmann, M.; and Li, H. 2019. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9489–9497.

Mao, W.; Xu, C.; Zhu, Q.; Chen, S.; and Wang, Y. 2023. Leapfrog diffusion model for stochastic trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5517–5526.

Martinez, J.; Black, M. J.; and Romero, J. 2017. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2891–2900.

Paden, B.; Čáp, M.; Yong, S. Z.; Yershov, D.; and Frazzoli, E. 2016. A survey of motion planning and control techniques for self-driving urban vehicles. *IEEE Transactions on intelligent vehicles*, 1(1): 33–55.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Sun, J.; and Chowdhary, G. 2024. CoMusion: Towards Consistent Stochastic Human Motion Prediction via Motion Diffusion. In *European Conference on Computer Vision*, 18–36. Springer.

Tian, S.; Liang, X.; and Zheng, M. 2023. An optimization-based human behavior modeling and prediction for human-robot collaborative disassembly. In *2023 American Control Conference (ACC)*, 3356–3361. IEEE.

Walker, J.; Marino, K.; Gupta, A.; and Hebert, M. 2017. The pose knows: Video forecasting by generating pose futures. In *Proceedings of the IEEE international conference on computer vision*, 3332–3341.

Wang, Y.; Yang, W.; Chen, X.; Wang, Y.; Guo, L.; Chau, L.-P.; Liu, Z.; Qiao, Y.; Kot, A. C.; and Wen, B. 2024. Sinsr: diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 25796–25805.

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

Xu, S.; Li, Z.; Wang, Y.-X.; and Gui, L.-Y. 2023. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14928–14940.

Yan, X.; Rastogi, A.; Villegas, R.; Sunkavalli, K.; Shechtman, E.; Hadap, S.; Yumer, E.; and Lee, H. 2018. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *Proceedings of the European conference on computer vision (ECCV)*, 265–281.

Yuan, Y.; and Kitani, K. 2020a. Diverse Trajectory Forecasting with Determinantal Point Processes. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yuan, Y.; and Kitani, K. 2020b. Dlow: Diversifying latent flows for diverse human motion prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, 346–364. Springer.

Yue, Z.; Wang, J.; and Loy, C. C. 2023. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36: 13294–13307.

Yue, Z.; Wang, J.; and Loy, C. C. 2024. Efficient diffusion model for image restoration by residual shifting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, J. Y.; Felsen, P.; Kanazawa, A.; and Malik, J. 2019. Predicting 3d human dynamics from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7114–7123.

Zhang, X.; Yi, D.; Behdad, S.; and Saxena, S. 2023. Unsupervised human activity recognition learning for disassembly tasks. *IEEE Transactions on Industrial Informatics*, 20(1): 785–794.

Zhang, Y.; Black, M. J.; and Tang, S. 2021. We are more than our joints: Predicting how 3d bodies move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3372–3382.