

CoMusion: Towards Consistent Stochastic Human Motion Prediction via Motion Diffusion

Jiarui Sun[✉] and Girish Chowdhary[✉]

University of Illinois Urbana-Champaign
{jsun57,girishc}@illinois.edu

Abstract. Stochastic Human Motion Prediction (HMP) aims to predict multiple possible future human pose sequences from observed ones. Most prior works learn motion distributions through encoding-decoding in the latent space, which does not preserve motion’s spatial-temporal structure. While effective, these methods often require complex, multi-stage training and yield predictions that are inconsistent with the provided history and can be physically unrealistic. To address these issues, we propose **CoMusion**, a single-stage, end-to-end diffusion-based stochastic HMP framework. **CoMusion** is inspired from the insight that a smooth future pose initialization improves prediction performance, a strategy not previously utilized in stochastic models but evidenced in deterministic works. To generate such initialization, **CoMusion**’s motion predictor starts with a Transformer-based network for initial reconstruction of corrupted motion. Then, a graph convolutional network (GCN) is employed to refine the prediction considering past observations in the discrete cosine transformation (DCT) space. Our method, facilitated by the Transformer-GCN module design and a proposed variance scheduler, excels in predicting accurate, realistic, and consistent motions, while maintaining appropriate diversity. Experimental results on benchmark datasets demonstrate that **CoMusion** surpasses prior methods across metrics, while demonstrating superior generation quality.

Keywords: Stochastic Human Motion Prediction · Diffusion Models

1 Introduction

Human Motion Prediction (HMP) aims to forecast human movements based on observed motion trajectories. This task has a wide range of applications [6, 30, 41, 65, 74, 76, 77, 79, 83], spanning autonomous driving [55], robotics [19], animation creation [71], and healthcare [63]. A considerable body of research tackles the deterministic HMP problem, aiming to predict a single, most probable future pose sequence [14, 47, 49]. Among these works, the top-performing models [40, 47] demonstrate that graph convolutional networks (GCN) is very suitable for HMP. By coupling GCN with discrete cosine transformation (DCT), these GCN-DCT methods treat human poses as graphs and explicitly model spatial-temporal relations among joints, which benefits motion prediction. However, the deterministic methods fall short in contexts such as autonomous driving in crowded areas, where predicting different possible human motions is crucial.

Recent research has shifted toward the stochastic paradigm of HMP, adopting generative models to learn conditional motion distributions [4, 5, 11, 18, 34, 37, 50, 60, 70, 78]. These stochastic approaches, while effective in certain scenarios, face several issues. First, most top-performing methods involve *complex multi-stage training processes* to enhance prediction performance. Methods such as [50, 78] necessitate multiple training rounds for motion mode coverage and motion validity. Recent diffusion model (DM)-based methods also require extra training stages for output post-processing [70] and motion encoding-decoding [4]. The multiple training stages require laborious engineering efforts in model tuning, making them less appealing for many applications. Second, stochastic HMP works often generate *inconsistent or even unrealistic motion* with respect to the provided history [15, 78]. To regularize predictions and enhance diversity, these methods either incorporate explicit diversity-promoting losses [50] or construct additional sampling spaces [15] to avoid posterior collapse. Such methods frequently result in sub-optimal predictions that deviate from the historical motion context. This deviation can, at times, result in pose sequences that are entirely unrealistic from a physical standpoint. Although recent works [4] have begun to address this issue, ensuring the predicted motion that is both realistic and seamlessly synchronized with the provided motion history remains a significant challenge.

Intuitively, to mitigate these issues, it is reasonable for stochastic models to utilize the GCN-DCT design proven effective in deterministic contexts [47, 51], as their strong performance suggests a potential reduction in the cumbersome training pipelines and prediction inconsistency. Surprisingly, this is not the case. Most stochastic methods learn motion distributions through encoding-decoding in the latent space [4, 58, 70, 78], not preserving motion’s spatial-temporal structure. This raises the question why such *model design gap* exists between deterministic and stochastic HMP works. To study this, we delve into the efficacy of the GCN-DCT design in deterministic models. We find that these models excel primarily due to GCN’s spatial modeling capability and, crucially, DCT’s prowess in temporal modeling. As joint motion is temporally smooth, the DCT space significantly decreases the variability among elements between the last-observation-padded sequence and the groundtruth, as illustrated in Fig. 1, making it easier to learn than in the pose space. This ease of learning is evidenced by the universal use of global residual connections [14] for residual learning across these methods [10, 49, 51, 52]. However, this is not the case for stochastic models. The DM-based models primarily predict noise, and methods predicting

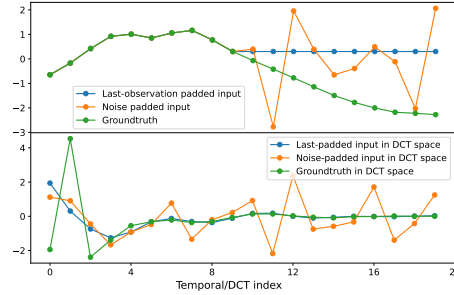


Fig. 1: Top: Three joint motion trajectories (length 20), last 10 features vary among the last-observation-padded, noise-padded and groundtruth sequences. **Bottom:** Their corresponding DCT values.

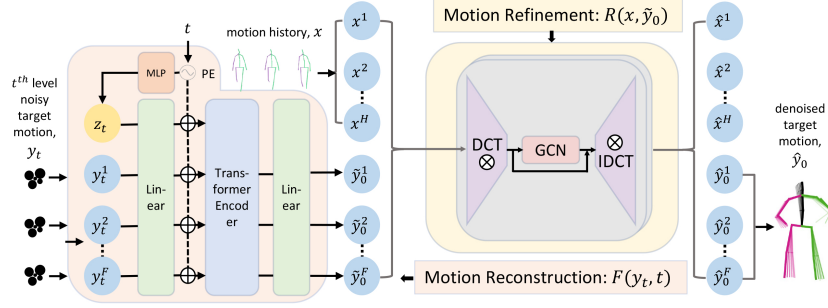


Fig. 2: Architecture of CoMusion’s predictor $G_\theta(\cdot)$. Inputs include the t^{th} level target noisy motion y_t , motion history x , and time step t . The motion predictor operates in two stages: (1) the Transformer-based motion reconstruction module $F(\cdot)$ initially reconstructs $\tilde{y}_0$ from y_t and t , and (2) the GCN-based motion refinement module $R(\cdot)$ then generates the complete motion sequence using the concatenated inputs of x and $\tilde{y}_0$. IDCT stands for Inverse DCT and PE for Positional Encoding.

motion equally find no learning benefits from concatenating motion history with noise, due to a much larger input-groundtruth discrepancy in the DCT space, a challenge also depicted in Fig. 1. This fundamental difference explains why the GCN-DCT design has not been extensively adopted in stochastic HMP methods.

Building on this insight, we suggest that incorporating a pre-processing step to reconstruct smooth future motion from noise could simplify the learning process. By using sequences padded with the reconstructions as inputs for a GCN-DCT design, we can mirror the reduced learning difficulty observed in deterministic HMP works [47, 49, 51]. To this end, we introduce **CoMusion**, a *single-stage* DM-based framework tailored to *consistent* HMP with its predictor architecture shown in Fig. 2. To generate a smooth future pose sequence, **CoMusion**’s motion predictor starts with a Transformer-based network for initial reconstruction of corrupted motion. Then, a GCN is employed to refine the generated motion in the DCT space, using the concatenation of provided history and reconstructed sequence. As such, **CoMusion** explicitly captures the spatio-temporal dependencies of human motion as a graph, which most stochastic HMP works have overlooked. Importantly, **CoMusion** adopts a direct motion prediction strategy [64], diverging from the common noise prediction scheme [4, 11, 25]. This approach allows **CoMusion** to integrate a structure-aware loss that accounts for skeletal structure, further easing its learning process. Moreover, with a simple yet effective adjustment to the standard cosine variance scheduler, we additionally elevate the accuracy and diversity of **CoMusion**’s generated motion samples.

Our contributions are summarized as follows. (1) We propose a single-stage, end-to-end diffusion framework for stochastic HMP, generating significantly more coherent and realistic motion than previous methods. (2) We design a motion generator which combines Transformer and GCN to capture spatial-temporal dynamics of human motion in DCT space. To the best of our knowledge, **CoMusion** represents the first exploration of integrating GCN-DCT design with DMs for

stochastic HMP. (3) We conduct comprehensive analyses to validate the efficacy of **CoMusion**. Benchmark results show that **CoMusion** outperforms previous approaches, achieving an improvement of at least 35% in fidelity metrics, establishing it as a robust new baseline in the field.

2 Related Work

Human Motion Prediction. Early efforts in HMP focused on deterministic settings [1, 8, 9, 17, 21, 28, 39, 53], aiming to predict one most likely pose sequence. A key development in this domain was initiated by Mao *et al.* [51], which popularized modeling motion in DCT space using GCNs [10, 14, 49, 51, 52]. However, deterministic methods cannot model motion distributions and are thus not suitable for stochastic HMP. To this end, generative methods [3, 5, 20, 22, 37, 38, 75, 78] are proposed. However, it is surprising that the top-performing stochastic works [4, 7, 15, 45, 58] rarely utilize the GCN-DCT design which has been proven effective in deterministic settings. Instead, they opt for learning via encoding-decoding in the latent space, which does not preserve motion’s spatial-temporal structure. While few works [11] attempt to explicitly exploit spatial-temporal patterns from a motion completion perspective, they, along with encoding-decoding methods, face issues such as complex, multi-stage training pipelines [46, 70] and sub-optimal predictions that are often inconsistent with the provided history.

Recently, few DM-based approaches [4, 11, 70] are proposed due to their ability to produce more diverse, higher-quality samples compared to generative adversarial network (GAN) and variational autoencoder (VAE)-based methods. Wei *et al.* proposed MotionDiff [70], a two-stage framework that consists of a Transformer-based noise predictor for motion generation, and a pretrained network to enhance sample diversity as a post-processing step. To address history-future inconsistency, Barquero *et al.* [4] also took a two-stage approach to model the diffusion process in the behavioral space rather than the coordinate space. Chen *et al.* proposed HumanMAC [11], which uses a Transformer-based module with mask modeling to achieve single-stage learning. Our **CoMusion** takes this one step further by exploring the potential of the GCN-DCT design with a pre-processing Transformer unit to model the motion denoising process. This synthesis allows **CoMusion** to capture the intricate spatial-temporal dynamics of human motion, and it ensures the efficiency of single-stage training coupled with unmatched performance in generating consistent, realistic prediction samples.

Denoising Diffusion Models. Denoising diffusion probabilistic models (DDPMs) [18, 25, 54, 61, 69, 73] have recently received significant attention and have been applied in many fields [12, 16, 26, 36, 56, 72] due to their superior sample quality and diversity. DDPMs define a diffusion process in which random noise is gradually added to the data using a Markov chain, and then learn how to reverse the process to generate desired data samples. In the field of text-driven human motion synthesis, prior works such as MDM [64], MotionDiffuse [81], PhysDiff [80] and MotionGPT [29] leverage DMs with Transformers to generate human

motion via natural languages. The advancements in DMs have also enabled works such as Diffusion-Conductor [84], EDGE [66], and MoFusion [13] in generating human motions synchronized with audio and music.

One bottleneck of DDPMs is their efficiency, as they typically require a large number of denoising steps to generate one sample. Numerous efforts have aimed to tackle this issue, devising methods for fast sampling [43, 44, 82] and reducing the resolution of data [57] through auto-encoding. Though not in the thousands, HMP works such as [11, 70] still require a large number of denoising steps even when using advanced samplers [61]. Benefiting from learning motion directly instead of noise with a motion predictor that reduces learning difficulty through the GCN-DCT design, CoMusion achieves state-of-the-art performance in both prediction accuracy and fidelity with only a few diffusion steps, without the need for fast sampling techniques, while maintaining an appropriate level of diversity.

3 Methodology

3.1 Problem Definition

Given a motion history $x^{1:H} = \{x^i\}_{i=1}^H$ of length H , our objective is to predict the subsequent F poses $x^{H+1:H+F} = \{x^i\}_{i=H+1}^{H+F}$. In stochastic HMP, we forecast multiple pose sequences from a single motion history, denoting one predicted sequence as $y^{1:F} := x^{H+1:H+F}$. Each pose at time step i is represented as $x^i \in \mathbb{R}^{J \times 3}$ where J being the number of body joints. Superscripts in $x^{1:H}$ and $y^{1:F}$ may be omitted when contextually clear.

3.2 Conditional Motion Diffusion

Let $\{y_t\}_{t=0}^T$ denote a general Markov noising process where y_0 represents the true data samples. The forward unconditional diffusion transitions are denoted as:

$$q(y_t|y_{t-1}) = \mathcal{N}(y_t; \sqrt{\alpha_t}y_{t-1}, (1 - \alpha_t)\mathbf{I}), \quad (1)$$

where $\{\alpha_t\}_{t=0}^T \in [0, 1]$ control the noise level, and can either be fixed [25] or learned [33]. To estimate the true data distribution, the reverse diffusion process is constructed to progressively denoise the corrupted data samples y_t from $t = T$ to $t = 1$ as:

$$p_\theta(y_{t-1}|y_t) = \mathcal{N}(y_{t-1}; \mu_\theta(y_t, t), \sigma_\theta^2(y_t, t)\mathbf{I}). \quad (2)$$

In our HMP context, we need to extend the above formulation to the conditional case. Specifically, the reverse diffusion transition in Eq. (2) becomes:

$$p_\theta(y_{t-1}|y_t, x) = \mathcal{N}(y_{t-1}; \mu_\theta(y_t, x, t), \sigma_\theta^2(y_t, x, t)\mathbf{I}), \quad (3)$$

where x represents the motion history and y_t represents the t^{th} level target noisy motion. In the seminal work [25], Ho *et al.* proposed to (1) fix $\sigma_\theta^2(\cdot)$, (2) predict noise ϵ using a noise predictor $\epsilon_\theta(\cdot)$ instead of μ by reparameterization, and (3) optimize the model with objective $L(\theta) = \mathbb{E}_t[\|\epsilon - \epsilon_\theta(\cdot)\|^2]$. These techniques are quickly adopted by later DM-based works [11, 70] due to the simplicity and great performance they offer.

3.3 Motion Diffusion Pipeline

Prediction Target. However, the ϵ -prediction target in the diffusion formulation presented above hinders HMP models from enjoying the aforementioned GCN-DCT design’s learning benefits, and also prevents them from leveraging various motion losses that have been extensively studied in previous works [24, 35]. To address these, we choose to predict the future motion directly [64]. Specifically, instead of predicting ϵ , we choose another reparameterization such that $\hat{y}_0 \leftarrow G_\theta(y_t, x, t)$, where $\hat{y}_0$ is the learned approximation of target future motion y_0 . This predicted motion $\hat{y}_0$ is then diffused back through $t - 1$ steps and, together with the provided motion history x , are used to generate the subsequent predictions in the sampling chain. With $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$, the forward diffusion process in Eq. (1) can be simplified as:

$$q(y_t|y_0) = \mathcal{N}(y_t; \sqrt{\bar{\alpha}_t}y_0, (1 - \bar{\alpha}_t)\mathbf{I}). \quad (4)$$

Generator Architecture. With the y_0 -prediction objective, our motion generator can benefit from the effective GCN-DCT design used in deterministic HMP works. In particular, $G_\theta(\cdot)$ leverages the spatio-temporal graph structure of motion data. Our designed network, shown in Fig. 2, consists of (1) a Transformer-based reconstruction module $F(\cdot)$ and (2) a GCN-based refinement module $R(\cdot)$.

First, we utilize a Transformer-based module, denoted as $F(y_t, t)$, to generate an “initial reconstruction” $\tilde{y}_0$ from the target noisy motion y_t , without considering the motion history. $F(y_t, t)$ explicitly models temporal correlations across noisy frames using a Transformer encoder. The time step t is first mapped to an embedding and then projected to the same latent dimension as the noisy motion frames $\{y_t^i\}_{i=1}^F$ via a feedforward network, producing a time token z_t . The noisy motion frames $\{y_t^i\}_{i=1}^F$ are first projected and then summed with positional encodings [67] to obtain positional information. These transformed motion frames are then prepended with the time token z_t and fed into the Transformer encoder. To derive the initial reconstruction $\tilde{y}_0$ from $F(\cdot)$, we discard the first output token which corresponds to t , and subsequently project the remaining learned representations back into the pose dimension of $J \times 3$. Using $F(\cdot)$ as a pre-processing step is crucial as it yields a smoother motion representation in the coordinate space compared to y_t , particularly in the early phases of denoising. The learned representation $\tilde{y}_0$ can be seamlessly integrated with the given motion history x , simplifying the learning process for the refinement module. The effectiveness of $F(\cdot)$ is further demonstrated in Sec. 4.4.

For the refinement module $R(x, \tilde{y}_0)$, we adopt the GCN-based architecture [62] from deterministic HMP research due to its effectiveness. The module $R(\cdot)$ begins by concatenating the inputs x and $\tilde{y}_0$. It then progressively refines the entire motion trajectory by alternating between the pose space and frequency space using DCT and its inverse (IDCT). The process starts by converting the motion trajectory into DCT coefficients, which are then processed by multiple GCN layers to capture the spatial-temporal relationships among joints. Next, the refined motion representation is converted back to the pose space and is

used as the input for the next GCN block. The final predicted future motion $\hat{y}_0$ is obtained by removing the segment that corresponds to the known motion history. That is:

$$\hat{y}_0 \leftarrow [\hat{x}; \hat{y}_0] = R(x, F(y_t, t)) := G_\theta(y_t, x, t). \quad (5)$$

Variance Scheduler. The variance scheduler $\{1 - \alpha_t\}_{t=0}^T$ is essential for the performance of DMs. The linear scheduler [25] and the cosine scheduler [54] are commonly used due to their simplicity and performance they provide. However, the suitability of these schedulers for specific cases, such as ours, warrants further investigation. As mentioned, our generator $G_\theta(y_t, x, t)$ aims to (1) directly predict future motion and (2) incorporate the historical motion sequence $x^{1:H}$ at each step of denoising. This approach differs markedly from the multimodal conditional settings found in other DM applications [2, 80], where the temporal behavioral guidance is not as explicitly defined.

To this end, we study how standard schedulers are designed. Equation (4) is equivalent to:

$$y_t = \sqrt{\bar{\alpha}_t} y_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad (6)$$

where $\epsilon \sim \mathcal{N}(0, I)$. In the theoretical setup, $\bar{\alpha}_0 = 1$ and $\bar{\alpha}_T = 0$, effectively transforming clean data into standard Gaussian noise. Following this premise, both linear and cosine schedulers are empirically designed to have $\bar{\alpha}_0 \approx 1$.

However, our empirical observations (Tab. 4) suggest these standard schedulers, while effective in many scenarios, do not facilitate the diversity and accuracy of samples necessary for our model to excel. We hypothesize that having $\bar{\alpha}_0$ close to 1 prevents $G_\theta(y_t, x, t)$ from producing accurate and diverse samples. Specifically, since we feed the clean motion history directly into $G_\theta(\cdot)$ and aim to predict y_0 by modeling $x^{1:H}$ explicitly, the strong guidance from $x^{1:H}$ and the strong spatial-temporal modeling of $G_\theta(\cdot)$ make the prediction task overly simple as the reverse process progresses. Due to the poor mode coverage of the current HMP datasets with their limited size, predictions may not be multimodal, resulting in sub-optimal sample diversity and accuracy.

To address the issue, rather than developing complex methods to promote sample diversity, we simply modify the original cosine scheduler to better suit our HMP framework. Specifically, we have:

$$\bar{\alpha}_t = \cos\left(\frac{t/T + 1}{2} \cdot \frac{\pi}{2}\right)^2. \quad (7)$$

By setting the offset to 1 as opposed to the commonly used 0.008 and relax the $\bar{\alpha}_0 \approx 1$ restriction, we establish an initial value of $\bar{\alpha}_0 = \cos(\pi/4)^2 = 0.5$. This change ensures that the task of predicting y_0 remains non-trivial even at the final denoising stages when t is close to 0, and consequently, obliges the model to consistently tackle the prediction task throughout the entire diffusion process, without an over-reliance on the clarity of the motion history guidance.

3.4 Learning Algorithm

As outlined in Sec. 1, aiming to predict motion directly enables us to utilize geometric losses to supervise the model $G_\theta(\cdot)$. To this end, we utilize a structure-aware loss [62] to take into account the details of the structure of human motion. The structure-aware loss weights all joints differently to reflect their relative importance in the motion context. For a single motion trajectory, the loss function for **CoMusion** is expressed as:

$$\mathcal{L}_\theta(G, y_0, x) = \mathbb{E}_{\substack{y_0 \sim q(\cdot|x) \\ t \sim [1, T]}} \mathcal{L}_{\text{rec}}(G_\theta(y_t, x, t), y_0, x). \quad (8)$$

The reconstruction loss is defined by:

$$\mathcal{L}_{\text{rec}} = \frac{1}{J} \sum_{j=1}^J (\gamma \cdot \|(x^j - \hat{x}^j) \cdot \lambda^j\|_1 + \|(y_0^j - \hat{y}_0^j) \cdot \lambda^j\|_1), \quad (9)$$

where the superscript j indicates the joint index, λ^j is the weight assigned to each joint, and γ is a hyperparameter balancing the importances of the reconstruction of motion history and the prediction of future. Importantly, **CoMusion** not only predicts y_0 but also reconstructs x , ensuring a global awareness of the entire motion trajectory and thus enhancing prediction accuracy. The weights for each joint are determined based on the kinematic structure of the human body, prioritizing joints prone to more dynamic movements, and do not require learning. For weight λ^j derivation details, please refer to the supplementary material.

Additionally, we use the relaxation technique [4, 50] to further promote the prediction diversity. For each motion history, we generate k target motion trajectories and only optimize $\mathcal{L}_\theta$ towards the most accurate prediction. That is:

$$\mathcal{L}_{\text{final}} = \min_k \mathcal{L}_\theta(G^k, y_0, x), \quad (10)$$

where G^k is the k^{th} generated motion trajectory based on the original pose sequence $[x; y_0]$.

4 Experiments

4.1 Datasets

Human3.6M [27], the most widely used dataset for stochastic HMP, contains motion clips of 7 subjects performing 15 different actions recorded at 50 Hz. For a fair comparison with previous works, we adopt the evaluation protocol of [4], where a 16-joint skeleton is used for human structure modeling. We predict 2s (100 frames) based on 0.5s (25 frames) of observation.

AMASS [48] unifies multiple Mocap datasets, such as HumanEva-I [59] using a shared SMPL [42] parameterization for human skeleton modeling. As a multi-dataset collection, AMASS can be used to perform cross-dataset evaluation to

Table 1: Quantitative results for Human3.6M dataset [27]. The best results are highlighted in **bold**. The symbol ‘-’ indicates that the results are not reported in the baseline work. For all metrics except for APD, lower is better.

Type	Method	One-Stage	APD $\uparrow$	APDE $\downarrow$	ADE $\downarrow$	FDE $\downarrow$	MMADE $\downarrow$	MMFDE $\downarrow$	CMD $\downarrow$	FID $\downarrow$
GAN-based	HP-GAN [5]	✓	7.214	-	0.858	0.867	0.847	0.858	-	-
	DeLiGAN [23]	✓	6.509	-	0.483	0.534	0.520	0.545	-	-
VAE-based	TPK [68]	✓	6.723	1.906	0.461	0.560	0.522	0.569	6.326	0.538
	Motron [58]	✓	7.168	2.583	0.375	0.488	0.509	0.539	40.796	13.743
	DSF [79]	✗	9.330	-	0.493	0.592	0.550	0.599	-	-
	DLow [78]	✗	11.741	3.781	0.425	0.518	0.495	0.531	4.927	1.255
	GSPS [50]	✗	14.757	6.749	0.389	0.496	0.476	0.525	10.758	2.103
	DivSamp [15]	✗	15.310	7.479	0.370	0.485	0.475	0.516	11.692	2.083
	MotionDiff [70]	✗	15.353	-	0.411	0.509	0.508	0.536	-	-
DM-based	HumanMAC [11]	✓	6.301	-	0.369	0.480	0.509	0.545	-	-
	BeLFusion [4]	✗	7.602	1.662	0.372	0.474	0.473	0.507	5.988	0.209
	Ours	✓	7.632	1.609	0.350	0.458	0.494	0.506	3.202	0.102

examine a model’s generalization ability. Evaluation settings such as frame rate and dataset partition are all set to be the same as in previous works [4] for fair comparisons. We predict 2s (120 frames) into the future based on 0.5s (30 frames) of observation after downsampling.

4.2 Experimental Setup

Evaluation Metrics. Following previous work [4], we use a comprehensive set of metrics to evaluate CoMusion quantitatively. (1) Average Pairwise Distance (APD) computes the averaged ℓ_2 distance between all generated sample pairs to measure sample diversity. (2) The Average and (3) the Final Displacement Errors (ADE and FDE) calculate the averaged all-time and the last-frame ℓ_2 distances respectively between the groundtruth and the closest prediction, measuring sample accuracy. The multimodal versions of ADE and FDE, (4) MMADE, and (5) MMFDE assess a method’s ability to produce multimodal predictions, whose groundtruth is obtained by grouping similar observations. To quantify the realism of motion, (6) the Fréchet Inception Distance (FID) is used to assess the similarity between the distributions of generated and real motions.

Recently, Barquero *et al.* [4] proposed two metrics to better quantify a model’s ability to produce behaviorally consistent motion. (6) The area of the Cumulative Motion Distribution (CMD) measures the difference between the areas under the cumulative true motion and predicted motion distributions, capturing the plausibility of predicted motion at a global level. To analyze to what extent the diversity is properly modeled, (7) the Average Pairwise Distance Error (APDE) is defined as the absolute error between the APD of the multimodal groundtruth and the APD of the predictions. For metric calculation details, please refer to the supplementary material.

Baselines. For Human3.6M quantitative evaluation, CoMusion is compared with GAN-based approaches HP-GAN [5], DeLiGAN [23], VAE-based methods TPK [68], Motron [58], DSF [79], DLow [78], GSPS [50], DivSamp [15],

Table 2: Quantitative results for AMASS dataset [48]. The best results are highlighted in **bold**. The symbol ‘-’ indicates that the results are not reported in the baseline work. As AMASS does not contain class labels, the FID metric is not used for evaluation.

Type	Method	One-Stage	APD $\uparrow$	APDE $\downarrow$	ADE $\downarrow$	FDE $\downarrow$	MMADE $\downarrow$	MMFDE $\downarrow$	CMD $\downarrow$
VAE-based	TPK [68]	✓	9.283	2.265	0.656	0.675	0.658	0.674	17.127
	DLow [78]	✗	13.170	4.243	0.590	0.612	0.618	0.617	15.185
	GSPS [50]	✗	12.465	4.678	0.563	0.613	0.609	0.633	18.404
	DivSamp [15]	✗	24.724	15.837	0.564	0.647	0.623	0.667	50.239
DM-based	HumanMAC [11]	✓	9.321	-	0.511	0.554	0.593	0.591	-
	BeLFusion [4]	✗	9.376	1.977	0.513	0.560	0.569	0.585	16.995
	Ours	✓	10.848	2.328	0.494	0.547	0.469	0.466	9.636

and DM-based methods MotionDiff [70], HumanMAC [11], BeLFusion [4]. A selection of these, representing the most competitive methods, is further evaluated quantitatively on the AMASS dataset. For qualitative analysis, we compare CoMusion with DLow, GSPS, DivSamp, and BeLFusion.

Implementation Details. We train CoMusion as a 10-step DM with standard DDPM [25] sampling. We implement $F(y_t, t)$ using 8 Transformer encoder layers with a latent dimension of 512. We use a 2-layer MLP to project the time step embedding to the transformer dimension. The GCN-based refinement module $R(x, \tilde{y}_0)$ consists of 3 blocks, each of which contains 2 GCN layers with a latent dimension of 256. The latent dimension and the dropout ratio of the refinement module are 256 and 0.5. Adam [32] is used for all experiments with 0.0001 as the initial learning rate. For both datasets, CoMusion is trained for 500 epochs, and the learning rate starts to decay after the 200th epoch. More implementation details can be found in the supplementary material.

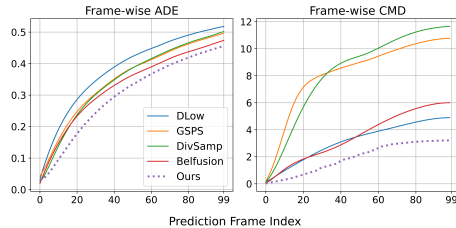


Fig. 3: Left: ADE computed at each prediction frame of state-of-the-art methods. Right: CMD computed up to each prediction frame. Both experiments are conducted on Human3.6M dataset.

4.3 Results Compared with State of the Art

Quantitative Results. The main quantitative comparison results are shown in Tabs. 1 and 2. For Human3.6M, we observe in Tab. 1 that CoMusion outperforms previous methods on the accuracy metrics (ADE and FDE) by large margins, which underlines the plausibility of our predicted motions. More notably, our method excels in generating behaviorally consistent and realistic future motions, evidenced by substantial improvements of **35%** in CMD and **51%** in FID over previous state of the art. While our model does not achieve the highest scores in the diversity metric (APD), it shows the best performance in APDE. This

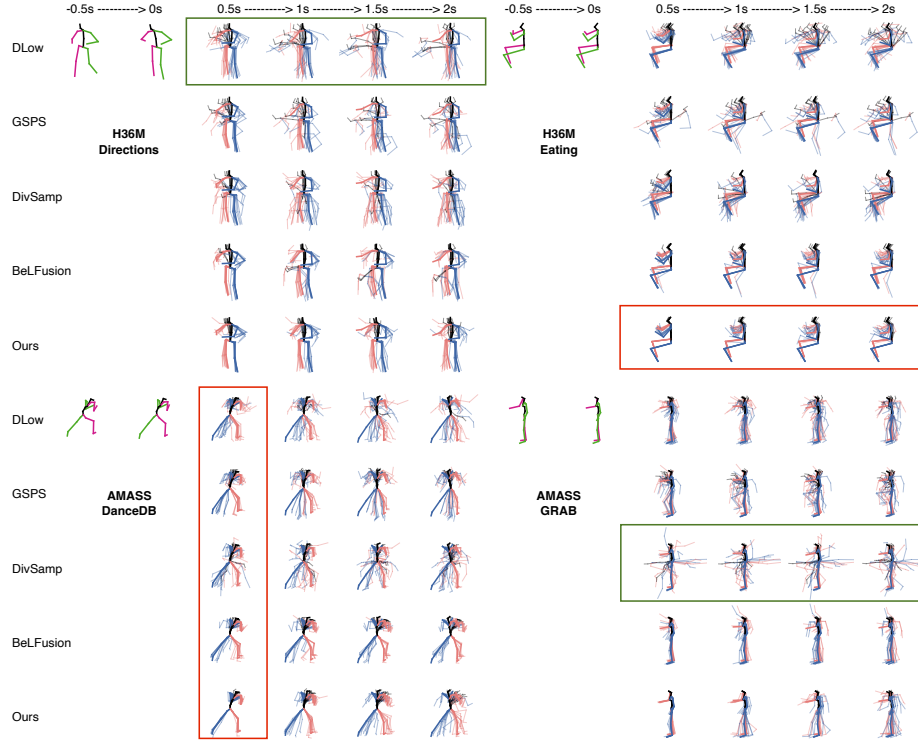


Fig. 4: Qualitative results of **CoMusion** compared with baseline methods. The upper block of rows corresponds to results obtained from the Human3.6M dataset, while the lower block of rows represents results from the AMASS dataset. The *green-purple* and the *blue-orange* skeletons denote the observed history and the predictions respectively.

demonstrates **CoMusion**’s capability to properly model the stochasticity of future motion based on the past. Furthermore, the frame-wise ADE and CMD results shown in Fig. 3 indicate that **CoMusion** can consistently outperform previous methods at each prediction frame. For AMASS results in Tab. 2, we achieve a competitive performance in APDE, and obtain state-of-the-art results in all other metrics (a **43%** improvement on CMD) except APD. This indicates again that **CoMusion** can generate consistent, realistic motion with proper diversity.

Qualitative Results. In Fig. 4, we compare **CoMusion** against multiple state-of-the-art methods qualitatively on both datasets, superimposing 10 predictions beneath the groundtruth motion at each prediction frame. Two actions, **Directions** and **Eating**, from Human3.6M, and two sub-datasets, **DanceDB** and **GRAB** from AMASS, are showcased for this comparison.

The visualizations first confirm that **CoMusion** is capable of generating natural and coherent stochastic predictions that are well-aligned with the motion history, as our predicted motions are qualitatively more reasonable and contain

Table 3: Ablation on CoMusion’s general architecture. In the *Sched.* column, ✓ denotes use of our proposed scheduler.

$F(y_t, t)$	$R(x, \tilde{y}_0)$	<i>Sched.</i>	APD ↑	APDE ↓	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓	CMD ↓	FID ↓
✗	✗	✗	12.880	4.854	0.959	1.000	0.987	1.004	966.716	1.047
✓	✗	✗	3.727	4.441	0.502	0.669	0.632	0.731	3.176	0.167
✗	✓	✗	6.858	1.835	0.539	0.678	0.625	0.694	197.105	0.474
✓	✓	✗	7.602	1.446	0.382	0.489	0.521	0.537	3.323	0.282
✓	✓	✓	7.632	1.609	0.350	0.458	0.494	0.506	3.202	0.102

fewer anomalies. For instance, take the predictions at 0.5s for **DanceDB** from AMASS, highlighted in the lower left red box, where the initial poses predicted by CoMusion are closely aligned with the groundtruth and then gradually depict variations over time. This is in stark contrast to other baselines, which often exhibit sudden motion discontinuities. Second, CoMusion demonstrates its ability to produce diverse predictions that are adapted to the context. For example, in the **Eating** action of Human3.6M (upper right red box), the predicted motions by CoMusion display various arm movements while maintaining the legs in a stationary position in most cases, signifying a realistic portrayal of the **Eating** action. More importantly, CoMusion tends to generate much fewer unreasonable poses when compared to other methods. For instance, in the **Directions** action for DLow (upper left green box), we observe an unnatural sudden bend from many predictions, and in the **GRAB** action for DivSamp (lower right green box), many predicted poses start to float in the air, violating real-world physical rules. These comparative insights highlight the advantage of CoMusion in producing more plausible and contextually appropriate predictions than previously established methods. more examples are provided in the supplementary material.

4.4 Ablation Study

In this section, we conduct an ablation analysis on the Human3.6M dataset to investigate how different design choices of CoMusion affect its motion modeling ability. Additional studies are included in the supplementary material.

Framework Components. We first evaluate the individual contributions of CoMusion’s components to its performance. We test five variants of our framework by selectively disabling (1) the reconstruction module $F(\cdot)$, (2) the refinement module $R(\cdot)$, and (3) the proposed scheduler. If both modules are disabled, a plain GCN [51] is used. The results in Tab. 3 show that omitting the reconstruction module $F(\cdot)$ consistently leads to inferior performance, evidenced by higher ADE

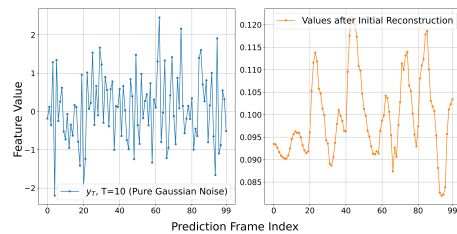


Fig. 5: Left: y_T , a Gaussian trajectory with $F = 100$ frames. Right: $F(y_T, T)$, the reconstructed trajectory. Compared with y_T , $F(y_T, T)$ depicts a much smoother temporal pattern with lower variance.

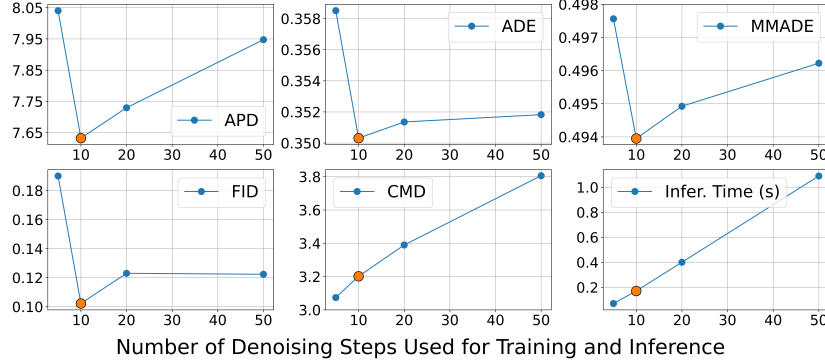


Fig. 6: Ablation results on the number of diffusion steps. The bottom rightmost sub-figure shows the per-sample time spent in seconds on Human3.6M inference.

Table 4: **Left (a):** Ablation on prediction target. **Right (b):** Ablation on variance scheduler. Linear scheduler’s results are not included as it causes CoMusion to diverge.

Target	APD $\uparrow$	ADE $\downarrow$	MMADE $\downarrow$	CMD $\downarrow$	FID $\downarrow$
ϵ	8.266	0.431	0.515	25.968	0.290
y_0 (ours)	7.632	0.350	0.494	3.202	0.102

Scheduler	APD $\uparrow$	ADE $\downarrow$	MMADE $\downarrow$	CMD $\downarrow$	FID $\downarrow$
Cosine	7.602	0.382	0.521	3.323	0.282
Sqrt.	6.988	0.359	0.503	2.832	0.128
ours	7.632	0.350	0.494	3.202	0.102

and FDE for accuracy and increased FID and CMD, indicating less realistic predictions compared to other variants. This is concurrently supported by Fig. 5, where $\tilde{y}_0$ produced by $F(\cdot)$ exhibits a much smoother temporal pattern than the target noisy motion y_t , thereby easing the subsequent task for the refinement module. Furthermore, the refinement module $R(\cdot)$ considerably improves both prediction accuracy and fidelity, benefits that are further enhanced by the proposed variance scheduler. These results demonstrate CoMusion’s effective use of the GCN-DCT design from deterministic works.

Diffusion Model Setups. For DM setup, we study the impacts of (1) the choice of prediction target, (2) the choice of variance scheduler and (3) the number of denoising steps T . The results are summarized in Tab. 4 (a),

(b), Tab. 5 and Fig. 6. From Tab. 4 (a), we first validate the advantage of performing y_0 -prediction over predicting noise. From Tab. 4 (b), we confirm that our proposed scheduler is the best choice for scheduling with the best overall results, as it ensures the y_0 -prediction task remains non-trivial throughout the entire denoising chain. Table 5 further supports our hypothesis regarding the effects of $\bar{\alpha}_0$, demonstrating that setting $\bar{\alpha}_0 = 0.5$ enhances both prediction accuracy and fidelity. From Fig. 6, we validate that $T = 10$ is a reasonable choice for number of diffusion steps with its best overall performance. Crucially, as a single-stage

Table 5: Effect of $\bar{\alpha}_0$.

$\bar{\alpha}_0$	≈ 1	0.9	0.8	0.7	0.6	0.5 (ours)
ADE $\downarrow$	0.407	0.368	0.361	0.356	0.354	0.350
FID $\downarrow$	0.323	0.138	0.123	0.109	0.110	0.102

Table 6: Left (a): Ablation on loss configurations. $\gamma = 0$ means that the model does not try to reconstruct the motion history x . $\lambda^j = 1$ means that all joints are weighted equally. **Right (b):** Ablation on implicit diversity relaxation.

Loss	APD $\uparrow$	ADE $\downarrow$	MMADE $\downarrow$	CMD $\downarrow$	FID $\downarrow$
$\gamma = 0$	7.661	0.351	0.496	2.742	0.123
$\lambda^j = 1$	7.609	0.352	0.494	2.970	0.115
ℓ_2	9.054	0.378	0.509	8.215	0.204
ours	7.632	0.350	0.494	3.202	0.102

k	APD $\uparrow$	ADE $\downarrow$	MMADE $\downarrow$	CMD $\downarrow$	FID $\downarrow$
1	4.061	0.346	0.503	6.094	0.163
2 (ours)	7.632	0.350	0.494	3.202	0.102
3	9.233	0.372	0.506	5.036	0.191

learning framework, **CoMusion** achieves state-of-the-art performance using only a minimal number of diffusion steps, in contrast to previous methods [11, 70]. As indicated in the bottom rightmost sub-figure of Fig. 6 with its low inference time, **CoMusion** is highly efficient and easy to optimize.

Loss Configuration. The designed loss $\mathcal{L}_{\text{final}}$ is essential to **CoMusion**’s performance. As such, we investigate (1) the components of Eq. (9) and (2) the implicit diversity relaxation technique. First, from Tab. 6 (a), we find that both reconstructing x and using structural weights contribute to **CoMusion**’s performance. Moreover, while the ℓ_1 loss may offer sub-optimal diversity, it excels in terms of accuracy and fidelity when compared with the ℓ_2 loss. Second, by varying the relaxation hyperparameter k as described in Eq. (10) and analyzing the results in Tab. 6 (b), we observe that setting k to 2 emerges as the optimal choice for most metrics, which helps **CoMusion** maintaining a good balance between sample diversity, accuracy and fidelity.

5 Conclusion

In this work we present **CoMusion**, a novel end-to-end DM-based framework for stochastic HMP. Benefiting from the GCN-DCT design used in deterministic works, **CoMusion** addresses issues of previous methods as it produces realistic, behaviorally consistent, and properly diverse human motions through single-stage learning. The motion predictor of **CoMusion** features a Transformer-based reconstruction module and a GCN-based refinement module, collaboratively learning future motion from its corrupted form and the provided motion history. By predicting motion directly using this dual design instead of noise and a simple yet effective variance scheduler, **CoMusion** establishes a new paradigm in stochastic HMP. The results obtained from extensive experiments and analyses confirm that **CoMusion** achieves significant performance gains over state-of-the-art baselines on benchmark datasets, demonstrating the efficacy of our method.

Acknowledgements

This work is supported in part by Navy N00014-19-1-2373, the joint NSF-USDA CPS Frontier project CNS #1954556, USDA-NIFA #2021-67021-34418, and Agriculture and Food Research Initiative (AFRI) grant no. 2020-67021-32799/project accession no.1024178 from the USDA National Institute of Food and Agriculture: NSF/USDA National AI Institute: AIFARMS. Work is also supported in part by NSF MRI grant #1725729 [31].

References

1. Aksan, E., Kaufmann, M., Cao, P., Hilliges, O.: A spatio-temporal transformer for 3d human motion prediction. In: 3DV. pp. 565–574 (2021)
2. Alexanderson, S., Nagy, R., Beskow, J., Henter, G.E.: Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Trans. Graph.* **42**(4), 44:1–44:20 (2023)
3. Aliakbarian, M.S., Saleh, F.S., Salzmänn, M., Petersson, L., Gould, S.: A stochastic conditioning scheme for diverse human motion prediction. In: CVPR. pp. 5222–5231. Computer Vision Foundation / IEEE (2020)
4. Barquero, G., Escalera, S., Palmero, C.: Belfusion: Latent diffusion for behavior-driven human motion prediction. In: ICCV (2023)
5. Barsoum, E., Kender, J., Liu, Z.: HP-GAN: probabilistic 3d human motion prediction via GAN. In: CVPR Workshops. pp. 1418–1427 (2018)
6. Bhattacharyya, A., Schiele, B., Fritz, M.: Accurate and diverse sampling of sequences based on a “best of many” sample objective. In: CVPR. pp. 8485–8493 (2018)
7. Blattmann, A., Milbich, T., Dorkenwald, M., Ommer, B.: Behavior-driven synthesis of human dynamics. In: CVPR. pp. 12236–12246. Computer Vision Foundation / IEEE (2021)
8. Bouazizi, A., Holzbock, A., Kressel, U., Dietmayer, K., Belagiannis, V.: Motion-mixer: Mlp-based 3d human body pose forecasting. In: IJCAI. pp. 791–798 (2022)
9. Bütepage, J., Black, M.J., Kragic, D., Kjellström, H.: Deep representation learning for human motion prediction and classification. In: CVPR. pp. 1591–1599 (2017)
10. Cai, Y., Huang, L., Wang, Y., Cham, T., Cai, J., Yuan, J., Liu, J., Yang, X., Zhu, Y., Shen, X., Liu, D., Liu, J., Magnenat-Thalmann, N.: Learning progressive joint propagation for human motion prediction. In: ECCV. pp. 226–242 (2020)
11. Chen, L.H., Zhang, J., Li, Y., Pang, Y., Xia, X., Liu, T.: Humanmac: Masked motion completion for human motion prediction. In: ICCV (2023)
12. Croitoru, F., Hondru, V., Ionescu, R.T., Shah, M.: Diffusion models in vision: A survey. *CoRR* **abs/2209.04747** (2022)
13. Dabral, R., Mughal, M.H., Golyanik, V., Theobalt, C.: Mofusion: A framework for denoising-diffusion-based motion synthesis. In: CVPR. pp. 9760–9770. IEEE (2023)
14. Dang, L., Nie, Y., Long, C., Zhang, Q., Li, G.: MSR-GCN: multi-scale residual graph convolution networks for human motion prediction. In: ICCV. pp. 11447–11456 (2021)
15. Dang, L., Nie, Y., Long, C., Zhang, Q., Li, G.: Diverse human motion prediction via gumbel-softmax sampling from an auxiliary space. In: MM. pp. 5162–5171 (2022)

16. Dhariwal, P., Nichol, A.Q.: Diffusion models beat gans on image synthesis. In: NeurIPS. pp. 8780–8794 (2021)
17. Fragkiadaki, K., Levine, S., Felsen, P., Malik, J.: Recurrent network models for human dynamics. In: ICCV. pp. 4346–4354 (2015)
18. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: NeurIPS. pp. 2672–2680 (2014)
19. Gui, L., Zhang, K., Wang, Y., Liang, X., Moura, J.M.F., Veloso, M.: Teaching robots to predict human motion. In: IROS. pp. 562–567 (2018)
20. Guo, W., Bie, X., Alameda-Pineda, X., Moreno-Noguer, F.: Multi-person extreme motion prediction. In: CVPR. pp. 13043–13054 (2022)
21. 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: WACV. pp. 4798–4808 (2023)
22. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social GAN: socially acceptable trajectories with generative adversarial networks. In: CVPR. pp. 2255–2264. Computer Vision Foundation / IEEE Computer Society (2018)
23. Gurumurthy, S., Sarvadevabhatla, R.K., Babu, R.V.: Deligan: Generative adversarial networks for diverse and limited data. In: CVPR. pp. 4941–4949 (2017)
24. Harvey, F.G., Yurick, M., Nowrouzezahrai, D., Pal, C.J.: Robust motion in-betweening. *ACM Trans. Graph.* **39**(4), 60 (2020)
25. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
26. Huang, R., Lam, M.W.Y., Wang, J., Su, D., Yu, D., Ren, Y., Zhao, Z.: Fastdiff: A fast conditional diffusion model for high-quality speech synthesis. In: IJCAI. pp. 4157–4163 (2022)
27. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI* **36**(7), 1325–1339 (2014)
28. Jain, A., Zamir, A.R., Savarese, S., Saxena, A.: Structural-rnn: Deep learning on spatio-temporal graphs. In: CVPR. pp. 5308–5317 (2016)
29. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. *CoRR* **abs/2306.14795** (2023)
30. Ju, X., Zeng, A., Jianan, W., Qiang, X., Lei, Z.: Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In: CVPR (2023)
31. Kindratenko, V., Mu, D., Zhan, Y., Maloney, J., Hashemi, S.H., Rabe, B., Xu, K., Campbell, R., Peng, J., Gropp, W.: Hal: Computer system for scalable deep learning. In: Practice and Experience in Advanced Research Computing. p. 41–48. PEARC '20, Association for Computing Machinery, New York, NY, USA (2020)
32. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
33. Kingma, D.P., Salimans, T., Poole, B., Ho, J.: Variational diffusion models. *CoRR* **abs/2107.00630** (2021)
34. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR (2014)
35. Kocabas, M., Athanasiou, N., Black, M.J.: VIBE: video inference for human body pose and shape estimation. In: CVPR. pp. 5252–5262. Computer Vision Foundation / IEEE (2020)
36. Kong, Z., Ping, W., Huang, J., Zhao, K., Catanzaro, B.: Diffwave: A versatile diffusion model for audio synthesis. In: ICLR (2021)
37. Kundu, J.N., Gor, M., Babu, R.V.: Bihmp-gan: Bidirectional 3d human motion prediction GAN. In: AAAI. pp. 8553–8560 (2019)

38. Lee, N., Choi, W., Vernaza, P., Choy, C.B., Torr, P.H.S., Chandraker, M.: DESIRE: distant future prediction in dynamic scenes with interacting agents. In: CVPR. pp. 2165–2174. IEEE Computer Society (2017)
39. Li, C., Zhang, Z., Lee, W.S., Lee, G.H.: Convolutional sequence to sequence model for human dynamics. In: CVPR. pp. 5226–5234 (2018)
40. Li, M., Chen, S., Zhang, Z., Xie, L., Tian, Q., Zhang, Y.: Skeleton-parted graph scattering networks for 3d human motion prediction. In: European Conference on Computer Vision. pp. 18–36. Springer (2022)
41. Liu, S., Chang, P., Huang, Z., Chakraborty, N., Hong, K., Liang, W., Livingston McPherson, D., Geng, J., Driggs-Campbell, K.: Intention aware robot crowd navigation with attention-based interaction graph. In: ICRA (2023)
42. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: a skinned multi-person linear model. *ACM Trans. Graph.* **34**(6), 248:1–248:16 (2015)
43. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In: NeurIPS (2022)
44. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *CoRR* **abs/2211.01095** (2022)
45. Lucas*, T., Baradel*, F., Weinzaepfel, P., Rogez, G.: Posegpt: Quantization-based 3d human motion generation and forecasting. In: ECCV (2022)
46. Ma, H., Li, J., Hosseini, R., Tomizuka, M., Choi, C.: Multi-objective diverse human motion prediction with knowledge distillation. In: CVPR. pp. 8151–8161 (2022)
47. 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: CVPR. pp. 6427–6436 (2022)
48. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: archive of motion capture as surface shapes. In: ICCV. pp. 5441–5450 (2019)
49. Mao, W., Liu, M., Salzmann, M.: History repeats itself: Human motion prediction via motion attention. In: ECCV. pp. 474–489 (2020)
50. Mao, W., Liu, M., Salzmann, M.: Generating smooth pose sequences for diverse human motion prediction. In: ICCV. pp. 13289–13298 (2021)
51. Mao, W., Liu, M., Salzmann, M., Li, H.: Learning trajectory dependencies for human motion prediction. In: ICCV. pp. 9488–9496 (2019)
52. Mao, W., Liu, M., Salzmann, M., Li, H.: Multi-level motion attention for human motion prediction. *IJCV* **129**(9), 2513–2535 (2021)
53. Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: CVPR. pp. 4674–4683 (2017)
54. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: ICML. pp. 8162–8171 (2021)
55. Paden, B., Cáp, M., Yong, S.Z., Yershov, D.S., Frazzoli, E.: A survey of motion planning and control techniques for self-driving urban vehicles. *IEEE Trans. Intell. Veh.* **1**(1), 33–55 (2016)
56. Pearce, T., Rashid, T., Kanervisto, A., Bignell, D., Sun, M., Georgescu, R., Macua, S.V., Tan, S.Z., Momennejad, I., Hofmann, K., Devlin, S.: Imitating human behaviour with diffusion models. In: ICLR (2023)
57. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10674–10685 (2022)
58. Salzmann, T., Pavone, M., Ryll, M.: Motron: Multimodal probabilistic human motion forecasting. In: CVPR. pp. 6447–6456 (2022)
59. Sigal, L., Balan, A.O., Black, M.J.: Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV* **87**(1-2), 4–27 (2010)

60. Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML. vol. 37, pp. 2256–2265. JMLR.org (2015)
61. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
62. Sun, J., Chowdhary, G.: Towards accurate human motion prediction via iterative refinement. CoRR **abs/2305.04443** (2023)
63. Taylor, W., Shah, S.A., Dashtipour, K., Zahid, A., Abbasi, Q.H., Imran, M.A.: An intelligent non-invasive real-time human activity recognition system for next-generation healthcare. *Sensors* **20**(9), 2653 (2020)
64. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: ICLR (2023)
65. Troje, N.F.: Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of vision* **2**(5), 2–2 (2002)
66. Tseng, J., Castellon, R., Liu, C.K.: EDGE: editable dance generation from music. In: CVPR. pp. 448–458. IEEE (2023)
67. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS. pp. 5998–6008 (2017)
68. Walker, J., Marino, K., Gupta, A., Hebert, M.: The pose knows: Video forecasting by generating pose futures. In: ICCV. pp. 3352–3361 (2017)
69. Wang, Z., Zheng, H., He, P., Chen, W., Zhou, M.: Diffusion-gan: Training gans with diffusion. In: ICLR (2023)
70. Wei, D., Sun, H., Li, B., Lu, J., Li, W., Sun, X., Hu, S.: Human joint kinematics diffusion-refinement for stochastic motion prediction. In: AAAI. pp. 6110–6118 (2023)
71. van Welbergen, H., van Basten, B.J.H., Egges, A., Ruttkay, Z., Overmars, M.H.: Real time animation of virtual humans: A trade-off between naturalness and control. *Comput. Graph. Forum* **29**(8), 2530–2554 (2010)
72. Whang, J., Delbracio, M., Talebi, H., Saharia, C., Dimakis, A.G., Milanfar, P.: Deblurring via stochastic refinement. In: CVPR. pp. 16272–16282. IEEE (2022)
73. Xiao, Z., Kreis, K., Vahdat, A.: Tackling the generative learning trilemma with denoising diffusion gans. In: ICLR (2022)
74. Xu, S., Li, Z., Wang, Y.X., Gui, L.Y.: Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In: ICCV (2023)
75. Xu, S., Wang, Y.X., Gui, L.Y.: Diverse human motion prediction guided by multi-level spatial-temporal anchors. In: ECCV (2022)
76. Xu, S., Wang, Y.X., Gui, L.Y.: Stochastic multi-person 3d motion forecasting. In: ICLR (2023)
77. Yang, J., Zeng, A., Li, F., Liu, S., Zhang, R., Zhang, L.: Neural interactive keypoint detection. In: ICCV. pp. 15122–15132 (2023)
78. Yuan, Y., Kitani, K.: Dlow: Diversifying latent flows for diverse human motion prediction. In: ECCV. pp. 346–364 (2020)
79. Yuan, Y., Kitani, K.M.: Diverse trajectory forecasting with determinantal point processes. In: ICLR (2020)
80. Yuan, Y., Song, J., Iqbal, U., Vahdat, A., Kautz, J.: Physdiff: Physics-guided human motion diffusion model. In: ICCV. pp. 16010–16021 (2023)
81. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. CoRR **abs/2208.15001** (2022)
82. Zhang, Q., Chen, Y.: Fast sampling of diffusion models with exponential integrator. In: ICLR (2023)

83. Zhang, Y., Black, M.J., Tang, S.: We are more than our joints: Predicting how 3d bodies move. In: CVPR. pp. 3372–3382. Computer Vision Foundation / IEEE (2021)
84. Zhao, Z., Bai, J., Chen, D., Wang, D., Pan, Y.: Taming diffusion models for music-driven conducting motion generation. CoRR **abs/2306.10065** (2023)