Computers & Graphics 129 (2025) 104244

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Special Section on CGVC2024

NeFT-Net: N-window extended frequency transformer for rhythmic motion prediction

Adeyemi Ademola [a], David Sinclair [a], Babis Koniaris [b], Samantha Hannah [a], Kenny Mitchell [a,*]

[a] Edinburgh Napier University, School of Computing, Engineering & Built environment, Merchiston Campus, Edinburgh EH10 5DT, United Kingdom
[b] Heriot-Watt University, School of Mathematical & Computer Science, Edinburgh Campus, Edinburgh EH14 4AP, United Kingdom

ARTICLE INFO

Dataset link: https://github.com/CarouselDancing/NeFT-net

Keywords:
Machine learning
Motion processing
Rendering
Virtual reality

ABSTRACT

Advancements in prediction of human motion sequences are critical for enabling online virtual reality (VR) users to dance and move in ways that accurately mirror real-world actions, delivering a more immersive and connected experience. However, latency in networked motion tracking remains a significant challenge, disrupting engagement and necessitating predictive solutions to achieve real-time synchronization of remote motions. To address this issue, we propose a novel approach leveraging a synthetically generated dataset based on supervised foot anchor placement timings for rhythmic motions, ensuring periodicity and reducing prediction errors. Our model integrates a discrete cosine transform (DCT) to encode motion, refine high-frequency components, and smooth motion sequences, mitigating jittery artifacts. Additionally, we introduce a feed-forward attention mechanism designed to learn from N-window pairs of 3D key-point pose histories for precise future motion prediction. Quantitative and qualitative evaluations on the Human3.6M dataset highlight significant improvements in mean per joint position error (MPJPE) metrics, demonstrating the superiority of our technique over state-of-the-art approaches. We further introduce novel result pose visualizations through the use of generative AI methods.

1. Introduction

In the fields of virtual reality (VR) and computer vision, real-time tracking is crucial for recovering accurate 3D pose data. Human joint pose data is commonly captured using multi-camera or single-camera setups integrated with AI algorithms to obtain depth information and directly recover pose key points and joint orientations. Nevertheless, challenges such as limited sensor range, occlusion, and latency persist in tracking 3D pose data. In order to improve immersion and engagement in patterned motion scenarios, there is a high demand for techniques that minimize latency [1,2] during motion tracking through motion prediction.

Deep learning techniques have significantly advanced the domain of human motion prediction [3,4]. Among these, recurrent neural networks (RNNs) have become particularly popular for predicting sequential human pose data [5,6]. However, when it comes to long-term horizons and periodic motions, RNNs often struggle due to their inability to effectively capture long-term history, which is essential for forecasting periodic motion actions. To address this limitation, recent approaches have incorporated encoders [7] to better represent historical information.

Our work introduces a multi-window extended frequency attention-based human motion prediction technique that utilizes synthetically generated periodic data based on re-timed foot anchor placements, as illustrated in Fig. 2. Our method is motivated by the observation that humans tend to repeat their motions in actions such as dancing to music beats. To validate this, we focus on the context of rhythmic motion prediction, where we demonstrate the effectiveness of our approach by re-timing *Human3.6m* [8] to match these rhythmic patterns. We present results based on analyzing relevant information from significant bones, such as the feet, over a fixed-length period.

Inspired by previous works [9], we represent each sub-sequence of foot anchors in the trajectory space using a Discrete Cosine Transform (DCT).

We then introduce our dual-windowed extended frequency motion attention as weights for DCT-encoded motion aggregation into a future motion estimate. To encode spatial dependencies between joints, we combine the motion estimate with the last observed matching period, using the result as input to a graph convolutional network (GCN) [10]. Our experiment, as shown in Fig. 5, demonstrates that our approach

* Corresponding author.
E-mail addresses: adeyemi.ademola@napier.ac.uk (A. Ademola), D.sinclair@napier.ac.uk (D. Sinclair), B.koniaris@hw.ac.uk (B. Koniaris), S.hannah@napier.ac.uk (S. Hannah), K.Mitchell2@napier.ac.uk (K. Mitchell).

https://doi.org/10.1016/j.cag.2025.104244
Received 4 January 2025; Received in revised form 14 April 2025; Accepted 2 May 2025
Available online 17 May 2025

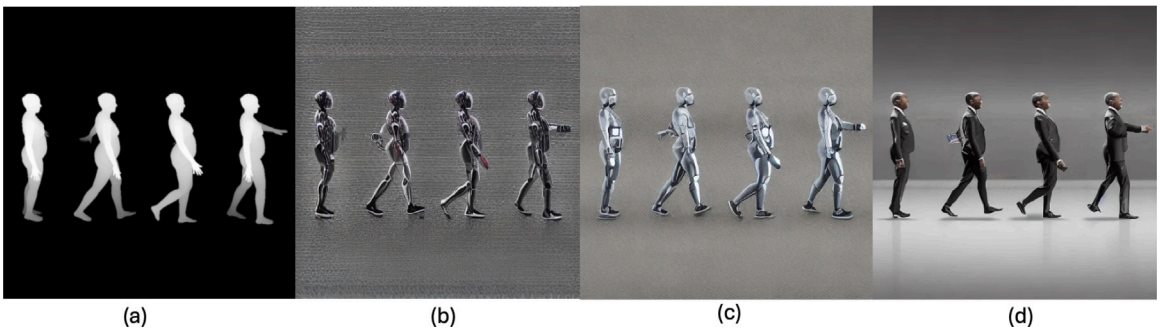


Fig. 1. NeFT-Net stable diffusion walking sequence visualization. Note: We apply the positive prompt "A walking human" to the depth map image on the left (a) with different seeds to achieve (b), (c), and (d).

outperforms state-of-the-art methods in long-term and short-term periodic motion prediction on the Human3.6M walking and walking together datasets. Our work extends *DeFT-Net* [11] developed upon insights from Mao et al. [12], specifically improving 3D pose motion prediction for known periodicity based on foot anchor placements.

To summarize, the main contributions of this paper are:

- a re-timed motion with supervised foot anchor information of periodic cycles, such as walking, for the defined use case of rhythmic motion prediction.
- an improved overall mean per joint position error (MPJPE) results compared to state-of-the-art methods in experiments on the Human3.6M dataset for forecasting short and long-term motions by introducing *MultiWindowDCT* attention aligned on a best fit period of each motion sequence.
- a strategy for photorealistic visualization of human body motion sequences by employing the use of stable diffusion on depth maps as shown in Fig. 1.
- an open source version of our code implementation https://github.com/CarouselDancing/NeFT-net

2. Related work

Human motion prediction relates to a variety of research areas like Computer Vision and Machine learning (ML), where predicting future movements is essential for applications in computer graphics and virtual reality. Section 2.1 details the various traditional techniques employed in the task of motion style synthesis and prediction. Section 2.2 describes how recurrent neural networks (RNN) has been adopted over the years for sequence-to-sequence 3D human motion prediction. Section 2.3 highlights the uniqueness of the attention-based approach compared to other approaches for motion prediction.

2.1. Traditional approaches

Motivated by the inherent probabilistic nature of periodic human motion, early methods such as Boltzmann machines and Hidden Markov Models (HMMs) [13,14] have been widely used to predict motion sequences. Style interpolation techniques are also frequently applied to synthesize motion, often driven by scripts, 2D video inputs, or to generate new choreography for virtual motion capture. While these methods offer robust solutions, they lack the adaptability and precision needed for capturing both short-term and long-term dependencies, particularly in dynamic contexts like dance and rhythmic walking sequences. Other advancements have introduced probabilistic models that leverage large motion databases and low-dimensional representations [15]. These methods utilize implicit empirical distributions and efficient binary tree-based search to approximate the true distribution of human motion. By structuring motion data efficiently, they allow for realistic motion synthesis and robust tracking within Bayesian frameworks, addressing both adaptability and precision challenges.

2.2. Recurrent Neural Networks (RNN) approaches

RNNs have grown in prominence for 3D human motion prediction tasks [16]. The encoder–decoder model (ERD), first introduced by *Fragkiadaki et al.* [6], incorporates Long Short-Term Memory (LSTM) cells in the latent space for capturing motion dynamics. The work of *Jain et al.* [5] leverages a spatio-temporal graph skeleton, utilizing RNNs as nodes to model kinematic chain joint dependencies. *Aksan et al.* [17] replace dense output layers in the RNN architecture with structural prediction layers to explicitly model joint dependencies that follow a kinematic chain. In the works of *Ghosh et al.* [18], a separate denoising auto-encoder is trained to correct noisy outputs. All these techniques suffer inability to capture long-range motion history trajectories.

However, RNN-based methods have historically struggled with capturing long-term motion history, leading to limitations in predicting prolonged sequences. In response, *Martinez et al.* [19] introduced a sequence-to-sequence (Seq2Seq) architecture incorporating an input-to-output skip connection, which mitigates some of the inherent bias by training the model with its own predictions. Despite improved results over earlier pose-based models [5], the discontinuity between ground truth and predicted frames persisted.

To address this, *Pavllo et al.* [20] adapted the *teacher-forcing technique*, allowing the model to gradually learn from its own outputs, further enhancing prediction accuracy. Additionally, *Chiu et al.* [16] introduced a hierarchical RNN model that operates across multiple time scales to better capture motion variability over different time spans. Furthermore, adversarial training methods proposed by *Gui et al.* [21] enable the generation of smoother motion sequences.

In the work of *Hernandez et al.* [22], human motion forecasting was framed as a tensor imputation problem, with generative adversarial networks (GANs) adapted for long-term prediction. Although these techniques resulted in improved performance, the use of adversarial networks introduces challenges in training, such as instability due to the adversarial nature of the generator-discriminator dynamics, difficulty in achieving convergence, and sensitivity to hyperparameters, particularly when applied to periodic datasets requiring precise foot anchor encoding.

2.3. Beyond recurrent models

Given the drawbacks of RNNs, several works have employed the use of feed-forward networks as an alternative solution [3,9]. The work of *Butepage et al.* [3] introduced a fully connected feed forward to process the recent history poses, investigating techniques to encode temporal historical information via convolution and exploiting the kinematic tree to encode spatial information. *Li et al.* [7] suggest a convolutional sequence-to sequence model (CNN) processing a two-dimensional pose matrix whose column represent the pose at every time step. The model was employed to extract a pose motion prior from long-term motion

history of frames, which, in conjunction with more recent motion history, was used as an input to an auto regressive network for future pose prediction. While more effective than RNN-based frameworks, the manually selected size of the convolutional windows highly influences the temporal encoding of motion sequences. To address this, *Aksan et al.* [23] introduced a spati-temporal transformer encompassing a fully auto-regressive approach to model temporal dependencies given the recursive nature of human motion. *Cai et al.* [24] leverage a transformer architecture on the DCT coefficients extracted from the seed sequence and make joint predictions progressively by following a kinematic tree. Similarly, *Mao et al.* [9] encodes joint sequence via DCT and train a graph convolutional network (GCN) to capture/learn inter-joint dependencies. Since the GCN operates on temporal windows of poses to produce an output, the pose forecast are limited to a predetermined length. To address this they extracted DCT coefficients from shorter sub-sequences in a sliding window fashion aggregated with a 1D attention block. *Guinot et al.* [25] introduced a stacked-attention mechanism utilizing synthetic IMU data to improve long-term dependency handling in dance motion prediction. This method addresses the limitations of traditional RNNs by transforming motion dynamics into the frequency domain using discrete cosine transform (DCT), which better encodes temporal information.

Our work is related to these approaches, but differs in two aspects. First, we introduce windowed inputs of a time-beat signal based on foot anchor pose information to the DCT windowed input so our model can learn periodic motions of short and long term history in the frequency domain. We then introduce an N-window extended frequency model with a focus on motion periodicity.

3. Method overview

Our technique introduces a unique approach to improving human motion prediction by incorporating periodic patterns and adapting a multi-window of poses Z_i. Each Z_i consists of three concatenated slices S_i, $S_{i+p+\text{offset}}$, and $S_{i+2p+2\text{offset}}$ from the motion history $S_1 = [s_1, s_2, s_3, \ldots, s_N]$. Here, p represents the period, and offset allows flexibility in adjusting the relative positions of these slices. This technique captures long-term temporal dependencies by analyzing different periods within human motion data, thus enhancing our model's ability to forecast future poses with improved performance. As shown in Fig. 2, we synthesize 3D pose data by interpolating frames containing motion foot anchor information from natural walking sequences in the Human3.6M dataset. We apply *spherical interpolation* for pose rotations and *linear interpolation* for pose translations to ensure smooth periodic motions. Since future frame forecasting from past sequences is the main goal, our method parallels approaches that utilize Discrete Cosine Transform (DCT) to encode motion, suppress high frequencies, and smooth jittery motions as seen in prior work [9,12]. To adapt the attention model to periodic motion cycles, we fold pose tensors to learn smooth motion transitions. Our model utilizes window slices of encoded periodic motion. For instance, if the first window captures the current motion, the second window integrates the immediate history, and the third slice looks two steps further back. This three-slice stack model enables more robust short- and long-term motion forecasting.

3.1. Foot anchor frame interpolation

As our goal is to learn from periodic walking sequence motions and forecast future pose motions, similar to *Cao et al.* [26], we rely on frame annotations based on the right foot placement at every nth frame. For periodic actions, such as walking and walking together, *linear interpolation* is applied to the root joint for smooth transitions between frames.

In Eq. (1), we compute a weighted average between the translation vectors of two key frames, p_1 and p_2. The interpolation factor $t \in [0, 1]$ controls the degree of blending between these frames. When $t = 0$, the result is entirely p_1, and when $t = 1$, the result is p_2. For intermediate values of t, the linear interpolation (lerp) computes a gradual transition between the two translation vectors, creating smooth transitions in position between frames.

$$\text{lerp}(p_1, p_2, t) = (1 - t)p_1 + tp_2 \tag{1}$$

In addition to translation interpolation, we also handle rotational changes between frames. Unlike translations, rotations are more complex and require spherical interpolation to compute smooth rotational transitions. Drawing from Kapoulkine's spherical linear interpolation approximation [27], we define a spherical path between the rotations and create key rotations from the rotation vectors of two consecutive frames.

In Eq. (2), we perform spherical linear interpolation (slerp) between two quaternions, q_1 and q_2, which represent rotations at two keyframes. The angle θ is the shortest angle between the two quaternions, and $t \in [0, 1]$ is the interpolation factor. The sine terms ensure that the interpolation follows the shortest path on the spherical surface, smoothly transitioning between the two rotations. When $t = 0$, the result is the first rotation q_1, and when $t = 1$, the result is q_2. This method provides a constant-speed rotational interpolation, crucial for preserving the natural flow of human motion.

$$\text{slerp}(q_1, q_2, t) = \frac{\sin((1-t)\theta)}{\sin(\theta)} q_1 + \frac{\sin(t\theta)}{\sin(\theta)} q_2 \tag{2}$$

We combine both interpolation techniques to achieve periodic dataset-based foot anchor frame placements and pass these sequences in an encoded DCT fashion to our multi-window frequency transformer. This method allows our model to learn and forecast future motion patterns from periodic sequences efficiently with fewer errors.

4. Multi-window frequency attention

Our multi-window attention presents a novel approach to addressing the complexities of human motion forecasting, particularly in periodic actions such as walking. As natural human motion contains short-term and long-term dependencies, which can be difficult to capture using traditional forecasting models, we address these challenges by incorporating multiple temporal windows representing different segments of the motion history, an adaptive weighting mechanism, and frequency-domain transformation. Through the use of the Discrete Cosine Transform (DCT) [28] and Graph Convolutional Networks (GCNs) [29], our model is more robust to temporal and spatial dependencies present in natural human motion (see Fig. 3).

The core idea behind our model is the use of *N=three temporal windows*, each representing a different portion of the motion history: Current Window, Dual Window, Nth-Past temporal Window. This segmentation allows the model to better account for motion patterns over time. The introduction of learnable weights enables the model to dynamically adjust the relative importance of each window. We compute the **deltas**, or differences, between adjacent windows to capture motion changes over time:

$$\Delta_{cp} = \mathbf{X}_c - \mathbf{X}_p \tag{3}$$

$$\Delta_{pd} = \mathbf{X}_p - \mathbf{X}_d \tag{4}$$

Next, the model applies learnable weights α_c, α_p, and α_d to adaptively weight the different temporal windows:

$$\mathbf{X}_{\text{weighted}} = \alpha_c \mathbf{X}_c + \alpha_p \mathbf{X}_p + \alpha_d \mathbf{X}_d \tag{5}$$

This adaptive weighting ensures that the model remains flexible, especially when the nature of the motion changes over time.

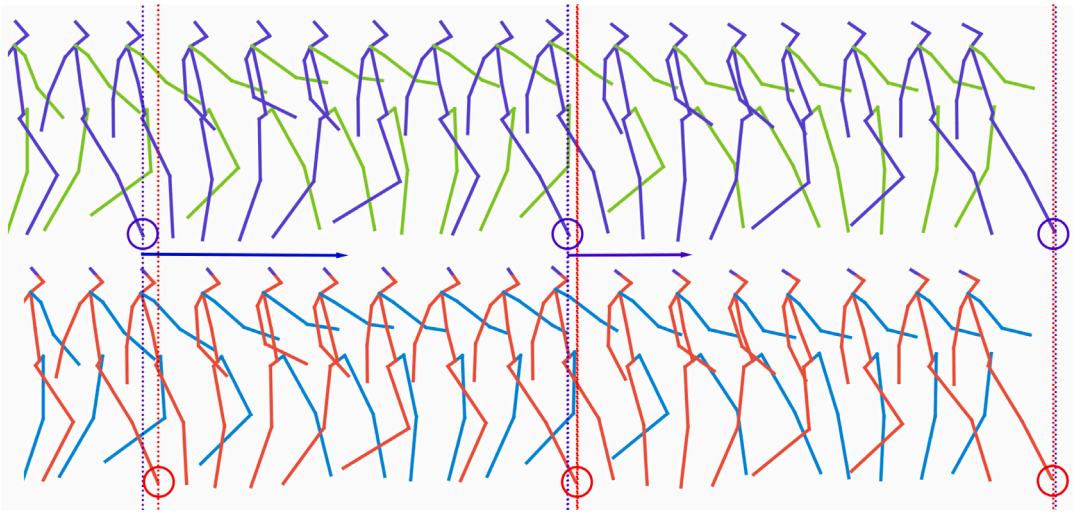


Fig. 2. A skeleton-grid comparison of the fixed DCT motions from the HistRepeatDCT method [12] and our re-timed multi-window extended DCT motions for test subject 5 walking synchronized with right foot anchor placements. The fixed DCT motion sequence is shown as right leg purple/left leg green, and our multi-window extended re-timed DCT motions as right leg red/left leg blue skeleton. Note: The red circles represent foot placement re-timed frames and purple circles define foot placements from start to end of the original sequence.

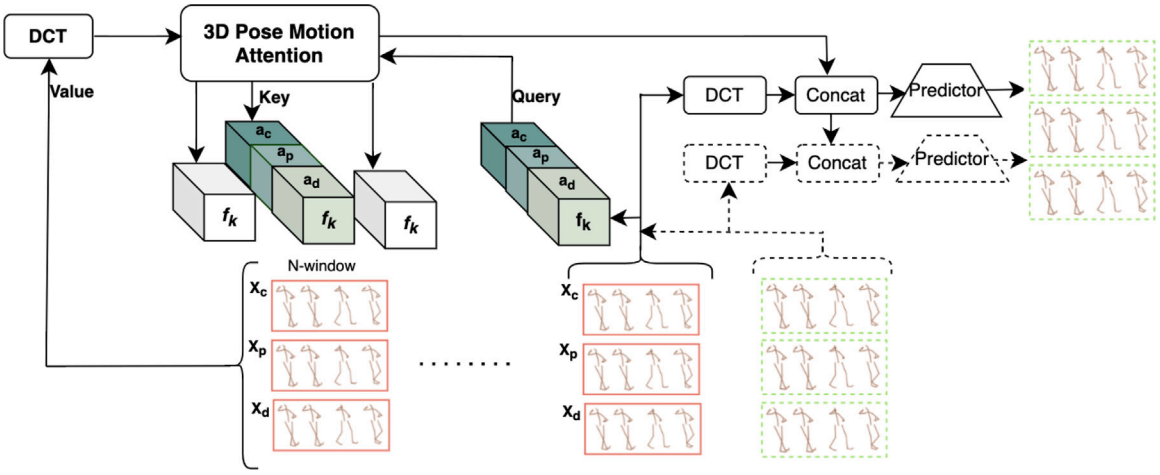


Fig. 3. Overview of NeFT-Net. Our re-timed DCT input poses are shown within the solid red boxes with the multi-window extended history, and the predicted poses are shown within dotted green boxes. The last observed poses are initially used as query. For every consecutive poses in the history (key), we compute an attention score to weigh the multi-window DCT coefficients (values) of the corresponding sub-sequence. The weighted sum of such values is then concatenated with the DCT coefficients of the last observed sub-sequence to predict the future. This comprises the transformer model of *OurMultiWindowDCT*.

Frequency Domain Transformation (DCT)

In addition to our N-window temporal representation, the model leverages the frequency domain through the *discrete cosine transform (DCT)* to handle periodic motion patterns. DCT transforms the motion data from the time domain to the frequency domain, which is particularly useful for periodic actions like walking, where repeating patterns occur. The DCT is defined as:

$$\mathbf{X}_{\text{DCT}}(k) = \sum_{n=0}^{N-1} \mathbf{X}(n) \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right] \tag{6}$$

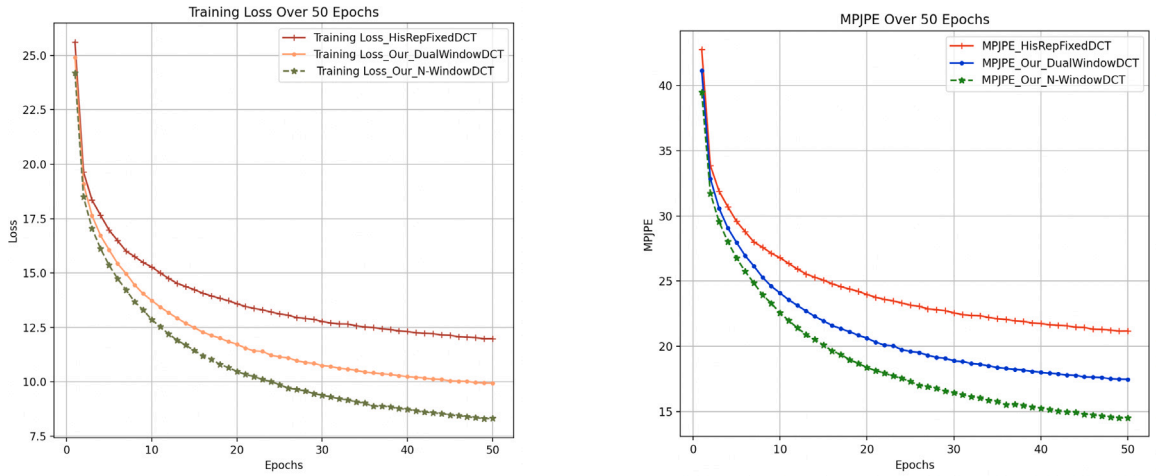
where k represents the frequency index and N is the length of the sequence. Applying DCT to the weighted windows yields:

$$\mathbf{X}_{\text{DCT-weighted}} = \text{DCT}(\mathbf{X}_{\text{weighted}}) \tag{7}$$

We apply the same principle in our Multi-windowDCT approach, where the DCT is applied to sequences from the current, dual, and n-past temporal windows. This transformation emphasizes the dominant frequencies in the motion while suppressing high-frequency noise, leading to smoother and more accurate predictions (see Table 1).

Table 1
Following baseline setting MPJPE Batch evaluation results for test Subject 5 comparison on our re-timed interpolated vs original History Repeats Itself DCT [12] method with Human3.6M datasets for predicting human motion at various frames for activities *walking* and *walking together*.

	Walking			Walking together		
Frame No.	1	3	5	8	9	10
HistRep [12]	5.68	17.28	27.62	40.31	43.69	46.81
DeFT-Net [11]	5.45	16.78	26.50	38.41	41.69	44.78
Ours	**5.31**	**16.23**	**25.48**	**37.04**	**40.75**	**43.54**



(a) Loss Function (b) MPJPE

Fig. 4. Training loss (a) and MPJPE (b) over 50 epochs for HisRepFixedDCT, DualWindowDCT, and our N-WindowDCT (3 windows).

Attention mechanism

An attention mechanism is employed to weigh the importance of different frames in the motion sequence, enabling the model to focus on the most relevant information. The general attention-weighted representation of the motion sequence is given by:

$$\mathbf{X}_{\text{attention}} = \sum_{t=1}^{T} \alpha_t \mathbf{x}_t \tag{8}$$

where α_t are the normalized attention weights for each frame $\mathbf{x}_t$, computed as:

$$\alpha_t = \frac{\exp(a_t)}{\sum_{i=1}^{T} \exp(a_i)}, \quad a_t = \text{softmax}(\mathbf{q}^\top \mathbf{k}_t) \tag{9}$$

Here, $\mathbf{q}$ represents the query (the current motion), and $\mathbf{k}_t$ represents the key (motion history).

To incorporate contributions from multiple temporal windows in our n-windowDCT approach, the attention mechanism is extended to weigh both individual frames within each window and the windows themselves. The n-windowed attention-weighted representation is given by:

$$\mathbf{X}_{\text{weighted}} = \sum_{i=1}^{n} \alpha_i \sum_{t=1}^{T} \beta_t^{(i)} \mathbf{x}_t^{(i)} \tag{10}$$

Here:

- α_i: Attention weight for the ith temporal window.
- $\beta_t^{(i)}$: Attention weight for the tth frame in the ith window, normalized over frames within that window.
- $\mathbf{x}_t^{(i)}$: The tth frame in the ith temporal window.

Our N-WindowDCT attention ensures both per-frame and per-window relevances are captured, aligning with the intuition that certain frames within each temporal window may carry more importance for the prediction task. As observed from Fig. 4, transitioning from *HisRepDCT* to *OurDualwindow* yields an average 10% improvement in training loss, with a further 12% gain beyond Dual to *N-windows*. However, the observed trend suggests diminishing returns of 4 windows of observations would yield only approximately 6.7% total improvement beyond the 3-window case. Given the memory and processing overhead of tracking multiple windows, the 3-window configuration stands out as the most practical and effective choice. Similarly diminishing returns are reflected in MPJPE, reinforcing this balance between performance and efficiency.

Inverse DCT and final prediction

After applying the GCN, we transform the output back to the time domain using the Inverse DCT (IDCT):

$$\mathbf{X}_{\text{pred}} = \text{IDCT}(\mathbf{X}_{\text{GCN}}) \tag{11}$$

This produces the final predicted motion sequence, incorporating both temporal and spatial dependencies.

Algorithm 1: Multi-Window Frequency Attention Algorithm

Input: Motion sequence $\mathbf{x}_t$ for $t = 1, \ldots, T$
Output: Predicted motion sequence $\mathbf{X}_{\text{pred}}$

```
// Segment the Motion History into n Temporal Windows
for i = 1 to n do
    Extract i-th window X_{w_i} = {x_{T-Σ_{j=1}^{i} w_j+1}, ..., x_{T-Σ_{j=1}^{i-1} w_j}};
// Compute Per-Frame Attention Weights Within Each Window
for i = 1 to n do
    for t = 1 to T do
        Compute β_t^{(i)} for frames in window i using query-key attention;
// Apply Attention Mechanism to Each Frame and Window
for i = 1 to n do
    Compute X_{w_i}^{attention} = Σ_{t=1}^{T} β_t^{(i)} x_t^{(i)};
Combine weighted windows: X_weighted = Σ_{i=1}^{n} α_i X_{w_i}^{attention};
// Transform to Frequency Domain Using DCT
Transform X_weighted to the frequency domain;
// Model Spatial Dependencies Using GCN
Update joint relationships using GCN: X_GCN;
// Transform Back to Time Domain Using IDCT
Transform X_GCN back to the time domain: X_pred;
return X_pred
```

5. ControlNet with depth maps for motion attention visualization

A powerful recent development arises where Stable Diffusion can be enhanced with ControlNet [30] to provide greater control over image generation. ControlNet allows for the incorporation of additional conditions, such as human pose and depth maps, to guide the generation process. This capability is particularly useful for visualizing

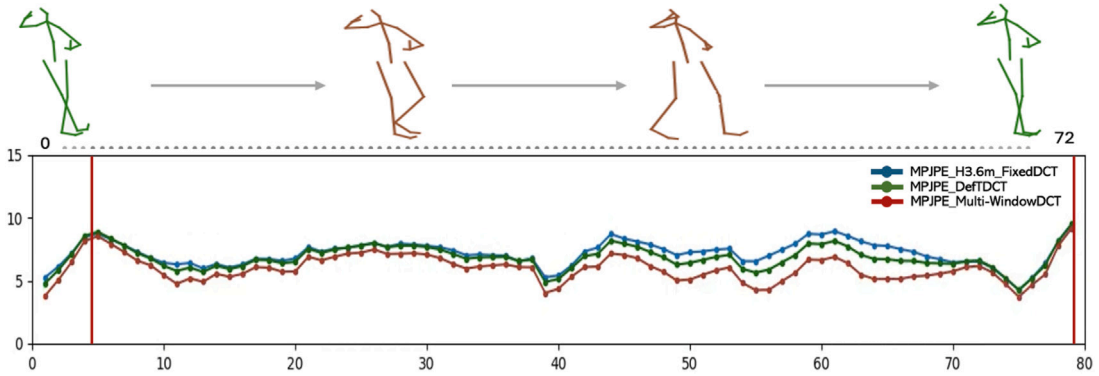


Fig. 5. From left to right, a plot visualization of the Mean Per Joint Position Error (MPJPE) across 72 frames for training on History Repeats Itself DCT, multi-window extended DCT, and *n*-window DCT encoded motion sequences. Note: The red vertical lines start and end of the foot placement cycle.

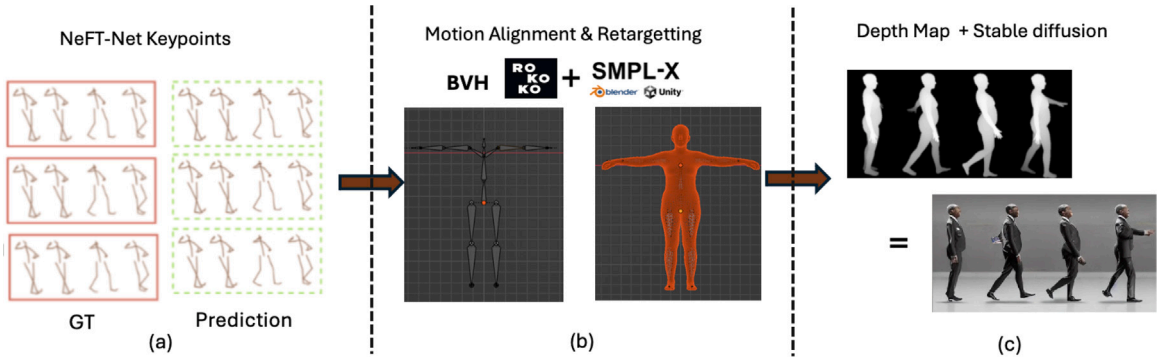


Fig. 6. From left to right: (a) NeFT-Net predicted keypoints, visualized alongside ground truth (GT) motion sequences in red and green outlines. (b) The predicted keypoints are aligned and retargeted, starting with BVH format (left) and mapped onto an SMPL-X mesh (right) using Rokoko Studio and Blender (c) Depth maps (top) are refined using stable diffusion to produce photorealistic rendered motion sequences (bottom).

motion, as depth maps can capture the spatial relationships between different body parts and their environment. ComfyUI,[1] a GUI-based Stable Diffusion interface, provides a user-friendly environment for composing images with this approach.

We focus on leveraging depth maps rendered from Blender as a guiding input to ControlNet, enabling precise and realistic depictions of both ground truth and predicted motions. By combining depth-based conditioning with the generative power of stable diffusion, this approach bridges the gap between data-driven motion prediction and its compelling visual representation, as shown in Fig. 6, offering a unique perspective on how AI can translate abstract motion data into vivid, interpretable renders.

The core technique is to use depth maps as the control input for ControlNet as seen in Fig. 1. By feeding a sequence of depth maps extracted from a video or generated from a simulated environment into ControlNet, we can guide the generation of a corresponding sequence of images that visualize the motion depicted in the depth maps. This approach offers several potential advantages:

- Enhanced Realism: The generative AI imagery is effortlessly realistic. Our prompting approach simply described the style of dress and context of walking, marching, etc. in a graphical depiction. Some orientation terms for example, from left to right assisted the success rate of more oriented diagrammatic results, but were not as influential as combining all pose frames side-by-side in producing coherent outcomes.
- Precise Control: ControlNet's ability to precisely control the generation process allows for fine-tuning the visualization based on the depth information. Our experiments supplying an alternative 2D bone hierarchy stick representation directly to ControlNet proved to be far less controllable than the more information rich depth representation.
- Novel Visualizations: The combination of ControlNet and depth maps opens up possibilities for creating novel and abstract visualizations of motion.

To visualize the motion of walking of our attention mechanism, we used Blender to generate depth maps of a character model at different stages of the walking cycle. These depth maps can then be used as input to ControlNet, along with text prompts describing the desired motion, to generate images that accurately depict the character's movement. The prompting strategy used in this study was deliberately minimal, primarily to maintain consistent orientation and scene composition (e.g., "person walking forward, side view, consistent lightning"). This simplicity ensured camera alignment across various frames but limited the generative detail in body articulation, clothing variation, and scene interaction.

All frames shown in Fig. 1 were generated together in a combined single diffusion pass, with depth maps concatenated to reinforce temporal coherence across the motion sequence. This batch conditioning approach helped maintain the consistency of lightening, background, and carbon appearance, which are often challenges in frame-by-frame generation.

In experimentation of this approach we naturally also applied a ControlNet model steered from Open Pose[2] derived bone hierarchy images, but found the information of such skeletal wire frame pose

[1] https://github.com/comfyanonymous/ComfyUI

[2] https://github.com/CMU-Perceptual-Computing-Lab/openpose

images led to excessive ambiguity and relatively poor posed generative image results when compared with the more information rich depth image controlled generations.

6. Conclusion

In this paper, we introduced an N-window-based motion attention model that leverages historical pose information based on the similarity between the current pose context and cyclic sub-sequences in the motion history. Our approach achieves state-of-the-art performance in predicting rhythmic motion by re-timing the Human3.6M dataset using foot anchor placements. Experimental results show strong generalization to previously unseen *walking* and *walking-together* sequences, as indicated by the training loss in Fig. 4(a) and MPJPE in Fig. 4(b), demonstrating improved joint pose accuracy. Our N-Window extended frequency transformer model aligns ideally upon three historical windows, arrived at due to the observation of a clear trend of diminishing returns in both training loss and predictive accuracy. Quantitative power regression analysis of results in Fig. 4(b) indicate that while the shift from a fixed representation to a dual-window model provides a substantial performance boost, and adding a third window slice contributes an even more meaningful improvement. Beyond this, however, predicted gains taper off progressively: estimating improvements less than 0.1% on successive windows beyond the 10th window. These reducing potential gains come at the cost of increased memory computation, storage and runtime complexity. This is coupled with the lower practical consideration of a motion pattern 10 cycles ago being as relevant to the current cycle in all but a regimented repeated march. We therefore consider three-windows both effective and efficient—avoiding unnecessary overhead while retaining strong predictive accuracy.

Whilst our analysis of varying windows of attention is a form of ablation study itself, we also compared re-timed and non-re-timed data preparations between the non-re-timed HistRepeatDCT method, and the Dual and N-Windowed approaches. Ablations of replacement or simplification of GCN and DCT elements could further indicate the relative importance of each of these measures, including further dissection of the model pipeline–such as isolating the roles of DCT encoding, the re-timing strategy, and window-based attention–as well as exploring the impact of window size for varying periodic motions and offset through hyperparameter sensitivity analysis.

We introduced the use of generative AI techniques to visualize predicted motions, which revealed the model's strong temporal consistency, particularly in sequential foot for placements—a core feature of rhythmic motion. While articulation of hands and facial expressions remain limited by the generative pipeline used, higher-fidelity synthesis approaches may offer future improvements.

Although real-time performance was not the primary target, our approach demonstrates inference speeds that are compatible with near-interactive rates and strong opportunities for optimization. Profiling and benchmarking will be key to validating deployment in time-sensitive scenarios. Moving forward, enhancing the model's real-time capabilities will be a priority—especially for interactive applications such as dance [2] and performance animation, where timing and rhythm are crucial.

CRediT authorship contribution statement

Adeyemi Ademola: Writing – original draft, Visualization, Software, Resources, Methodology, Investigation, Data curation. **David Sinclair:** Visualization, Software, Investigation. **Babis Koniaris:** Supervision, Methodology, Funding acquisition. **Samantha Hannah:** Visualization, Software. **Kenny Mitchell:** Writing – review & editing, Visualization, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

All authors disclosed no relevant relationships.

Acknowledgements

This article is an output of the CAROUSEL project which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101017779.

Data availability

The data is available at https://github.com/CarouselDancing/NeFT-net.

References

[1] Koniaris B, Sinclair D, Mitchell K. DanceMark: An open telemetry framework for latency-sensitive real-time networked immersive experiences. In: 2024 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops. IEEE; 2024, p. 462–3.

[2] Sinclair D, Ademola AV, Koniaris B, Mitchell K. DanceGraph: A complementary architecture for synchronous dancing online. In: 36th international computer animation social agents. 2023.

[3] Butepage J, Black MJ, Kragic D, Kjellstrom H. Deep representation learning for human motion prediction and classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, p. 6158–66.

[4] Cui Q, Sun H, Yang F. Learning dynamic relationships for 3D human motion prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 6519–27.

[5] Jain A, Zamir AR, Savarese S, Saxena A. Structural-rnn: Deep learning on spatio-temporal graphs. In: Proceedings of the ieee conference on computer vision and pattern recognition. 2016, p. 5308–17.

[6] Fragkiadaki K, Levine S, Felsen P, Malik J. Recurrent network models for human dynamics. In: Proceedings of the IEEE international conference on computer vision. 2015, p. 4346–54.

[7] Li C, Zhang Z, Lee WS, Lee GH. Convolutional sequence to sequence model for human dynamics. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 5226–34.

[8] Ionescu C, Papava D, Olaru V, Sminchisescu C. Human3. 6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans Pattern Anal Mach Intell 2013;36(7):1325–39.

[9] Mao W, Liu M, Salzmann M, Li H. Learning trajectory dependencies for human motion prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, p. 9489–97.

[10] Fu J, Yang F, Dang Y, Liu X, Yin J. Learning constrained dynamic correlations in spatiotemporal graphs for motion prediction. IEEE Trans Neural Netw Learn Syst 2023.

[11] Ademola A, Sinclair D, Koniaris B, Hannah S, Mitchell K. DeFT-Net: Dual-window extended frequency transformer for rhythmic motion prediction. In: The 42nd eurographics UK conference on computer graphics & visual computing conference. 2024.

[12] Mao W, Liu M, Salzmann M. 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. Springer; 2020, p. 474–89.

[13] Taylor GW, Hinton GE, Roweis S. Modeling human motion using binary latent variables. In: Advances in neural information processing systems, vol. 19, MIT Press; 2006, URL: https://proceedings.neurips.cc/paper/2006/hash/1091660f3dff84fd648efe31391c5524-Abstract.html.

[14] Brand M, Hertzmann A. Style machines. In: Proceedings of the 27th annual conference on computer graphics and interactive techniques. USA: ACM Press/Addison-Wesley Publishing Co.; 2000, p. 183–92. http://dx.doi.org/10.1145/344779.344865, URL: https://dl.acm.org/doi/10.1145/344779.344865.

[15] Sidenbladh H, Black MJ, Sigal L. Implicit probabilistic models of human motion for synthesis and tracking. In: Computer vision—ECCV 2002: 7th European conference on computer vision Copenhagen, Denmark, May 28–31, 2002 proceedings, part I 7. Springer; 2002, p. 784–800.

[16] Chiu Hk, Adeli E, Wang B, Huang DA, Niebles JC. Action-agnostic human pose forecasting. In: 2019 IEEE winter conference on applications of computer vision. IEEE; 2019, p. 1423–32.

[17] Aksan E, Kaufmann M, Hilliges O. Structured prediction helps 3D human motion modelling. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, p. 7144–53.

[18] Ghosh P, Song J, Aksan E, Hilliges O. Learning human motion models for long-term predictions. In: 2017 international conference on 3D vision. IEEE; 2017, p. 458–66.
[19] Martinez J, Black MJ, Romero J. On human motion prediction using recurrent neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, p. 2891–900.
[20] Pavllo D, Grangier D, Auli M. Quaternet: A quaternion-based recurrent model for human motion. In: British machine vision conference. 2018.
[21] Gui LY, Wang YX, Liang X, Moura JM. Adversarial geometry-aware human motion prediction. In: Proceedings of the European conference on computer vision. 2018, p. 786–803.
[22] Hernandez A, Gall J, Moreno-Noguer F. Human motion prediction via spatio-temporal inpainting. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, p. 7134–43.
[23] Aksan E, Kaufmann M, Cao P, Hilliges O. A spatio-temporal transformer for 3D human motion prediction. In: 2021 international conference on 3D vision. IEEE; 2021, p. 565–74.
[24] Cai Y, Huang L, Wang Y, Cham TJ, Cai J, Yuan J, et al. 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. Springer; 2020, p. 226–42.
[25] Guinot L, Matsumoto R, Iwata H. Stacked dual attention for joint dependency awareness in pose reconstruction and motion prediction. In: ICAT-EGVE 2023 - international conference on artificial reality and telexistence and eurographics symposium on virtual environments. 2023.
[26] Cao Z, Gao H, Mangalam K, Cai QZ, Vo M, Malik J. Long-term human motion prediction with scene context. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part I 16. Springer; 2020, p. 387–404.
[27] Kapoulkine A. Approximating slerp. 2015, URL: https://zeux.io/2015/07/23/approximating-slerp/.
[28] Chen LH, Zhang J, Li Y, Pang Y, Xia X, Liu T. Humanmac: Masked motion completion for human motion prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. 2023, p. 9544–55.
[29] Dang L, Nie Y, Long C, Zhang Q, Li G. Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 11467–76.
[30] Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. 2023, p. 3836–47.