Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.066944

ARTICLE

Human Motion Prediction Based on Multi-Level Spatial and Temporal Cues Learning

Jiayi Geng[1], Yuxuan Wu[1], Wenbo Lu[2], Pengxiang Su[1,*], Amel Ksibi[3], Wei Li[1], Zaffar Ahmed Shaikh[4,5] and Di Gai[6]

[1]School of Software, Nanchang University, Nanchang, 330000, China
[2]School of Queen Mary, Nanchang University, Nanchang, 330000, China
[3]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O.Box 84428, Riyadh, 11671, Saudi Arabia
[4]Department of Computer Science and Information Technology, Benazir Bhutto Shaheed University Lyari, Karachi, 75660, Pakistan
[5]School of Engineering, École Polytechnique Fédérale de Lausanne, Lausanne, 1015, Switzerland
[6]School of Mathematics and Computer Science, Nanchang University, Nanchang, 330000, China

*Corresponding Author: Pengxiang Su. Email: supengxiang@ncu.edu.cn

Received: 21 April 2025; Accepted: 24 July 2025; Published: 23 September 2025

ABSTRACT: Predicting human motion based on historical motion sequences is a fundamental problem in computer vision, which is at the core of many applications. Existing approaches primarily focus on encoding spatial dependencies among human joints while ignoring the temporal cues and the complex relationships across non-consecutive frames. These limitations hinder the model's ability to generate accurate predictions over longer time horizons and in scenarios with complex motion patterns. To address the above problems, we proposed a novel multi-level spatial and temporal learning model, which consists of a Cross Spatial Dependencies Encoding Module (CSM) and a Dynamic Temporal Connection Encoding Module (DTM). Specifically, the CSM is designed to capture complementary local and global spatial dependent information at both the joint level and the joint pair level. We further present DTM to encode diverse temporal evolution contexts and compress motion features to a deep level, enabling the model to capture both short-term and long-term dependencies efficiently. Extensive experiments conducted on the Human 3.6M and CMU Mocap datasets demonstrate that our model achieves state-of-the-art performance in both short-term and long-term predictions, outperforming existing methods by up to 20.3% in accuracy. Furthermore, ablation studies confirm the significant contributions of the CSM and DTM in enhancing prediction accuracy.

KEYWORDS: Human motion prediction; spatial dependencies learning; temporal context learning; graph convolutional networks; transformer

1 Introduction

Human motion prediction, the process of forecasting future pose sequences based on historical data, serves as a cornerstone for enabling intelligent and safe interactions between robots or machines and the physical world. Consequently, human motion prediction has attracted remarkable attention and is at the core of wide applications, such as autonomous driving [1,2], human tracking [3], motion generation [4], and human-robot interactions [5,6].

Early attempts in human motion prediction primarily employed latent variable models, such as the Hidden Markov Model (HMM) [7], Gaussian Processes (GP) [8], and Restricted Boltzmann Machine

(RBM) [9]. These models aimed to establish a correlation between past and future human motions. However, unlike inanimate physical movements governed by deterministic laws, human motion involves intricate dynamic contexts. Consequently, these initial mathematical models faced significant challenges in accurately capturing the complex dynamics of human motion, resulting in suboptimal forecasting outcomes.

Recently, with the improvement of computational power and the availability of large datasets, a large number of deep learning methods have been proposed and show improved results. The general human motion prediction algorithms encompass a sequence-to-sequence model whereby the historical pose sequence is encoded to a hidden temporal dynamic context that is decoded to obtain the future pose sequence. For example, references [10–12] employed recurrent neural networks (RNN), including long short term memory (LSTM), gated recurrent unit (GRU), to capture temporal information via a frame-wise reading method. However, these methods fail to encode long-term temporal dependencies, which only generate accurate predictions at short-term horizons. References [1,13,14] utilized convolutional neural networks (CNN) to encode the input pose sequence into a long-term latent variable, which can capture sufficient short- and long-term temporal correlations. In the aforementioned motion forecasting models, they directly extract motion information only from consecutive frames. Unfortunately, existing methods focus exclusively on consecutive frames and neglect the important temporal cues provided by non-adjacent frames, which are often crucial for accurate long-term predictions. Moreover, existing models treat all frames as equally important, which limits prediction accuracy by ignoring the varying relevance of different frames in the sequence. Embracing the above challenges, we propose a novel Temporal Cues Learning strategy and design a Dynamic Temporal Encoding Module that first utilizes a temporal relation matrix to flexibly establish diverse temporal dependencies from consecutive or non-consecutive frames, and then employs a temporal compression matrix to condense the captured shallow temporal information to deep level for improving information density and smoothness.

Moreover, early approaches only encode the motion information from the temporal dimension, which ignores the complementary motion features from related joints at the spatial dimension. Specifically, in the “Walking” and “Swimming” action, the arms and the legs usually show a mirror symmetry tendency and a side synchronization tendency. Some existing human motion prediction models have demonstrated that extracting the spatial coherence information from related joints and capturing the temporal dynamic context from the whole sequence simultaneously are positive for improving prediction performance. Several studies [15–18] have utilized a spatio-temporal graph to map known spatial dependencies among joints, while others [1,19–21] have employed graph convolutional networks (GCN) to reveal and aggregate intrinsic spatial relationships within a single frame. Nevertheless, these methods often struggle to accurately depict the complex and variable nature of human motion by focusing solely on establishing singular spatial dependencies across different joints. In contrast, transformer-based algorithms have gained prominence due to their ability to encode both local and global features effectively in various computer vision tasks, including human motion prediction [22,23]. Our study introduces a novel Cross Spatial Dependencies Encoding Module leveraging the transformer architecture, which adeptly extracts both local and global spatial dependency information across the entire human body, including joint level and joint pair level.

Inspired by the latest advancements in human motion prediction, we are dedicated to mastering the intricate spatial relationships among individual joints or joint pairs within frames, as well as the complex temporal connections spanning any distance within a pose sequence. The model handles the identification and connection of non-consecutive frames in a learned manner. Specifically, the Dynamic Temporal Encoding Module (DTM) employs a temporal relation matrix that is learned during training to capture temporal dependencies not only from consecutive frames but also from non-consecutive frames that are contextually relevant. This matrix enables the model to dynamically learn and establish connections between

discontinuous frames based on their temporal relevance to the overall motion sequence. As illustrated in Fig. 1, we have developed an advanced framework called the Multi-level Spatial and Temporal Learning Model (STM). This framework is structured around two fundamental modules: the Cross Spatial Dependencies Encoding Module (CSM) and the Dynamic Temporal Connection Encoding Module (DTM). (1) The CSM dissects each human body frame into distinct joints and joint pairs, which are then simultaneously processed through parallel self-attention layers designed to effectively integrate local and global associated features. This process is further enhanced by employing a multi-head mechanism capable of deriving various motion representations, thereby capturing a comprehensive range of spatial motion features at different spatial levels. (2) In contrast, the DTM treats each joint in every frame as an individual time step. It constructs an adjacency matrix that serves as a foundational tool for establishing temporal correlations between any two frames in the pose sequence. To achieve this goal, a strategic employment of a temporal compression matrix condenses and interprets temporal data efficiently, thus enhancing predictive capabilities.

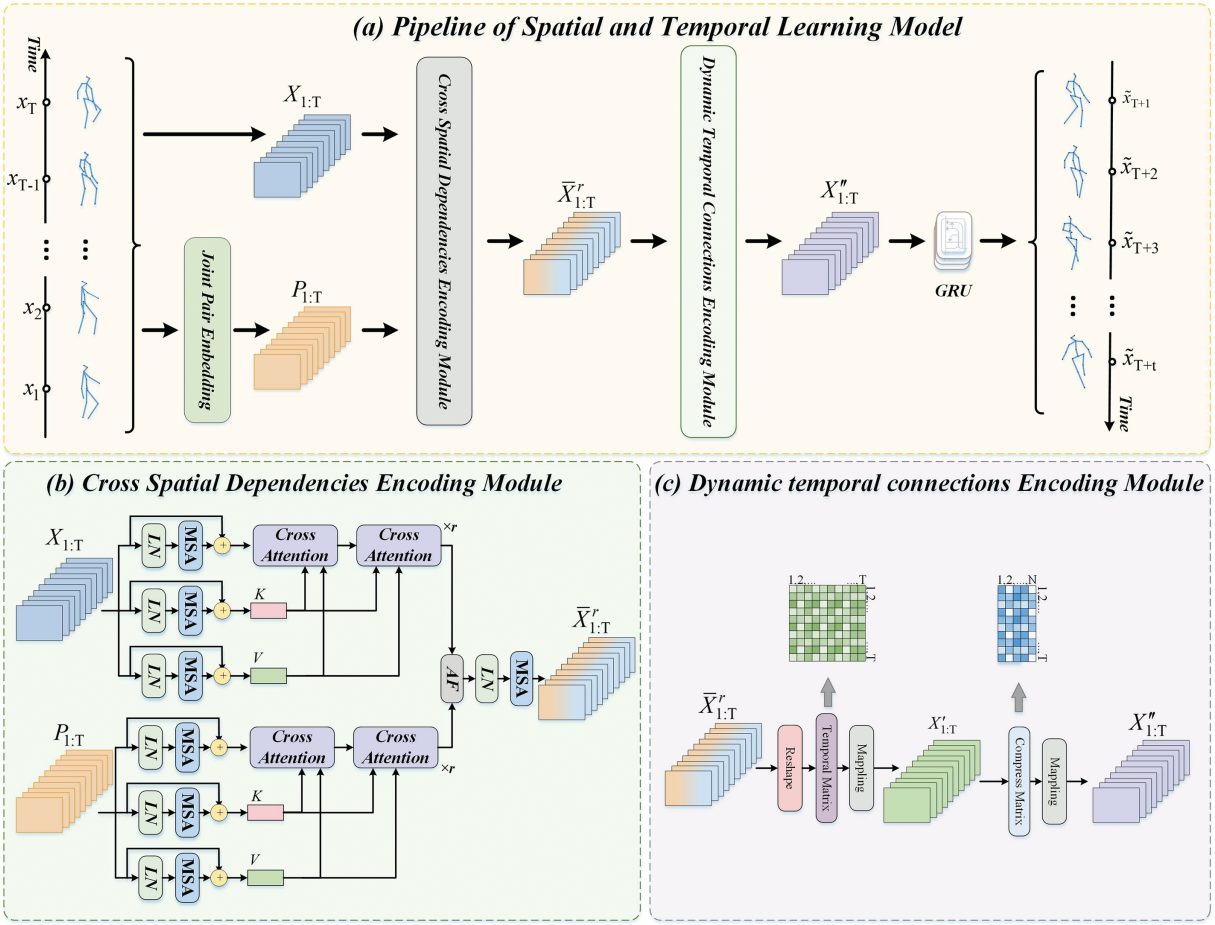

Figure 1: The architecture of our proposed method. **(a)** the pipline of multi-level spatial and temporal learning model, **(b)** the cross spatial dependencies encoding module, **(c)** the dynamic temporal connection encoding module

Overall, the key contributions of this paper are summarized as follows:

- We propose a novel cross spatial dependencies encoding module to simultaneously ensure the local and global spatial coherence from joint level and joint pair level.

- We introduce a dynamic temporal connection encoding module as the key realization of Temporal Cues Learning, which captures diverse and informative temporal evolution patterns, including non-consecutive dependencies, and compresses them to deeper levels to enhance motion smoothness.
- Our method achieves the state-of-the-art performance on both short- and long-term predictions. On the Human 3.6M dataset, our model improves MPJPE by 17.8% over DeepSSM and 25.3% over STSGCN for long-term predictions. On the CMU Mocap dataset, we achieve 3.9% improvement over STSGCN and 17.9% over TIM for long-term predictions at 1000 ms.

The rest of the paper is organized as follows. We review the related works in the field of human motion prediction in Section 2. We then introduce the details of the proposed method in Section 3. Thereafter, Section 4 presents the experimental setup and performance evaluation on the Human 3.6M and CMU Mocap datasets. Finally, Section 5 concludes the paper and outlines potential future work.

2 Related Work

In this section, we review the background of human motion prediction, reviewing key challenges in the field and summarizing relevant prior works that have laid the foundation for our approach.

2.1 Human Motion Prediction

In the field of human motion prediction, initial methodologies were categorized under latent variable models, including the Hidden Markov Model (HMM) [7], Gaussian Processes [8], and the Restricted Boltzmann Machine (RBM) [9]. The efficacy of deep learning models across a broad spectrum of computer vision tasks has been well-documented [24–26]. In recent years, researchers have developed a variety of deep learning strategies to effectively capture the temporal dynamics present in historical motion sequences [27–29], surpassing the performance of traditional methods in motion prediction. For instance, the ERD framework [11] employs an Encode-Decode architecture with recurrent layers to model motion information and generate future motion predictions. The Res-GRU [10] introduces a sequence-to-sequence model with a residual module before the GRU layer, encoding first-order motion directions. Despite their advancements, these models often fail to fully capture the structural intricacies of the human body, leading to inaccurate or distorted future pose predictions.

Addressing this gap, the S-RNN model [30] creates a spatio-temporal map to represent human structure, using LSTM to extract interactions among body parts. Similarly, the ConSeq2Seq model [13] employs a hierarchical CNN structure to simultaneously construct spatial and temporal correlated features. However, these methods tend to rely on manually predefined correlations, which can limit their robustness and comprehensive characterization. The LTD approach [1] treats the human body as a generic graph via GCN, automating the exploration of graph connections to learn long-range spatial relationships. The GAGCN method [15] utilizes an adaptive adjacency matrix to capture complex dependencies across diverse action types, while the DD-GCN [16] employs a dense graph with 4D adjacency modeling to illustrate temporal dynamics between frames. Nonetheless, these GCN-based approaches primarily focus on individual spatial relationships among joints and often struggle to encapsulate the full spectrum of complex and variable spatial dependencies. To address this, we propose a cross spatial dependencies encoding module, designed to learn both local and global spatial dependency information from a variety of motion representations at the joint and joint-pair levels.

2.2 Spatio-Temporal Dependencies Learning

Human motion is a coordinated movement involving multiple joints, exhibiting pronounced temporal evolution trends. A fundamental challenge in human motion prediction lies in effectively encoding the

spatial and temporal contextual information. Previous studies [30,31] have relied on predefined spatial and temporal relationships based on prior knowledge, such as kinematic chains. In contrast, S-RNN proposes a generic and principled spatio-temporal graph combined with an RNN to explicitly capture the interrelationships for accurate motion prediction [30]. This approach defines local spatial relationships among human joints, leading to satisfactory long-term prediction performance. HMR utilize Li algebra based the human global kinematic tree to represent the human spatial connection and utilizes a hierarchical recurrent network for learning temporal evolution [31]. However, this predefined approach for spatio-temporal relations lacks flexibility in modeling motion features effectively. Recently, researchers have proposed various data-driven models to flexibly capture relevant motion features from pose sequences. The STGCN method utilizes a signal adjacent matrix to effectively model the spatial relationships among joints in a pose and the temporal correlations of the same joint across neighboring frames [19]. The MPT method proposes a semi-constrained graph to comprehensively encode the relationships between joints, explicitly encoding prior skeletal connections and adaptively learning implicit dependencies between joints for a more comprehensive representation [32]. The STSGCN designs a signal-graph framework with a space-time separable module to divided the spatio-temporal connectivity into space and time affinity matrices to avoid the space-time cross-talk [5]. The MSR method combines and decodes motion features from fine to coarse scale, as well as from coarse to fine scale, in order to capture the evolutionary characteristics between historical and future pose sequences [33]. The SPGSN approach adopts an adaptive graph scattering between joints and fuses decomposed motion information from both graph spectrum and spatial dimensions [34]. The STGA employs a gating network to learn the cross dependencies of spatial and temporal information, effectively aggregating decoupled features from various action types [15]. Additionally, it has been demonstrated in [1] that compressing time information enhances the smoothness of temporal evolution, thereby assisting in performance improvement. Consequently, we propose a dynamic temporal connection encoding module designed to capture sufficient contextual dynamics from both continuous and discontinuous frames and subsequently compress these features to obtain more concise evolutionary patterns at deep level.

Table 1 shows the comparison of the existing methods, although the existing methods achieving good performance, but there still some problems, including 1) Limited spatial dependency modeling, where many models fail to capture complex relationships between joint pairs; 2) Rigid predefined relationships, restricting flexibility in modeling dynamic motion features; 3) Inadequate handling of non-consecutive temporal dependencies, which impacts long-term prediction accuracy.

Table 1: Comparison of related works

Method	Spatial modeling	Temporal modeling	Key limitation
ERD [11]	Uses recurrent layers for encoding motion	Encode-Decode architecture	Struggles to capture body structure intricacies
Res-GRU [10]	Sequence-to-sequence model with residual module	GRU-based encoding of motion directions	Limited in capturing spatial dependencies
S-RNN [30]	Spatio-temporal map with LSTM for part interactions	LSTM for temporal evolution	Relies on manually predefined correlations
ConSeq2Seq [13]	Hierarchical CNN structure for spatial and temporal features	CNN-based temporal feature learning	Dependency on predefined correlations
LTD [1]	GCN-based graph for spatial relationships	Graph-based learning of spatial dependencies	Limited in capturing complex spatial dependencies
GAGCN [15]	Adaptive adjacency matrix for capturing dependencies	Uses GCN for temporal dynamics	Struggles with complex joint-pair relationships
DD-GCN [16]	Dense graph with 4D adjacency modeling	4D graph for temporal dynamics	Focuses more on spatial than temporal dependencies
STGCN [19]	Signal adjacency matrix for spatial relationships	Temporal correlations between neighboring frames	Not fully address complex temporal dependencies

(Continued)

Table 1 (continued)

Method	Spatial modeling	Temporal modeling	Key limitation
STSGCN [5]	Space-time separable module for spatial and temporal matrices	Divides space-time into separate matrices	Fail to encode complex spatial and temporal dependencies
MSR [33]	Fine-to-coarse motion feature decoding	Evolutionary characteristics between sequences	Overlook subtle temporal dependencies
STGA [15]	Gating network for cross-spatial and temporal dependencies	Aggregates features across action types	Struggle with long-term dependencies

To address these challenges, we propose a novel approach that integrates both spatial and temporal dependencies across multiple levels: The Cross Spatial Dependencies Encoding Module (CSM) captures both local and global multi-level spatial dependencies, addressing the limitations of current GCN-based methods; and the Dynamic Temporal Connection Encoding Module (DTM) is designed to learn consecutive and non-consecutive temporal dependencies, enabling our model to perform better on long-term prediction tasks.

3 Our Approach

In this section, we initially provide a concise overview of our approach, followed by a detailed description of the proposed cross-level spatial dependencies encoding module. Subsequently, we elaborate on the designed dynamic temporal connection encoding module and discuss the employed loss function for training our model.

3.1 Overview

Problem Formulation The human motion prediction model is to capture the spatial connected information and temporal evolution context to extrapolate the future pose sequence based the historical pose sequence. We denote the historical pose sequence of T frames as $X_{1:\mathrm{T}} = \langle x_1, x_2, x_3, ..., x_{\mathrm{T}} \rangle \in R^{\mathrm{T}\times \mathrm{J}\times 3}$, the predicted pose sequence of future t frames represented as $\tilde{X}_{1:\mathrm{T}} = \langle \tilde{x}_{\mathrm{T}+1}, \tilde{x}_{\mathrm{T}+2}, \tilde{x}_{\mathrm{T}+3}, ..., \tilde{x}_{\mathrm{T}+\mathrm{t}} \rangle \in R^{\mathrm{t}\times \mathrm{J}\times 3}$, where $x_i \in R^{\mathrm{J}\times 3}$ is a human pose at frame i that total consist of J joint.

Method Overview The process of Spatial and Temporal Learning Model is shown in Algorithm 1. We can see that the proposed method consists of two key components. (1) A module for encoding cross spatial dependencies is employed to establish local and global spatial correlations from joint level and joint pair level, thereby obtaining complementary and robust spatial related information. (2) These features are fed into a dynamic temporal connection encoding module to adaptively capture the temporal evolution context between any two frames in the entire sequence. Subsequently, the features are compressed to deep level for enhancing compactness and smoothness.

Algorithm 1: Multi-level spatial-temporal learning pipeline

Input: Input motion sequence X (size $T \times J \times 3$)
Output: Predicted future motion sequence $\hat{X}$
Step 1: Spatial Encoding (CSM)
for *each frame t in X* **do**
  for *each joint k in J* **do**
    Calculate local and global spatial dependencies.
    Apply self-attention to capture spatial features
    Update spatial features with attention weights.

(Continued)

Algorithm 1 (continued)

end
Aggregate spatial features to form x_k^r.
end
Spatially encoded features $\bar{x}_s$ are obtained.
Step 2: Temporal Encoding (DTM)
for *each pair of frames (i, j)* **do**
Compute temporal dependencies.
Capture motion evolution patterns over time.
end
Apply temporal compression to condense temporal information. Temporal evolution features $X''_{1:T}$ are obtained.
Step 3: Prediction
Using spatio-temporal features $X''_{1:T}$ to predict future poses.
return $\tilde{X}_{1:T}$ as the predicted future motion sequence.

3.2 Cross Spatial Dependencies Encoding Module

For the task of human motion prediction, an efficient model for capturing spatial dependencies is crucial in extracting complementary features that are spatially related, which enhance the performance of prediction. Hence, a multi-level spatial dependencies encoding module are explored to promote the various local-global spatial coherence form joint and joint pair level. The details of our module are presented as follows. Firstly, we constructs an adjacency matrix based on the human body's anatomical structure. This matrix represents the initial spatial relationships between joint pairs. In particular, we set p_k and x_k to denote the joint pairs and joints information of frame k and fed them to our module as two branches. Existing GCN or Transformer-based methods directly extract information from the related joints at each iteration are inappropriate. However, these methods ignore the fact that directly fusing information from spatial related joints may introduce noise. For instance, joint d exhibits spatial correlation with joint c, and joint c in turn demonstrates spatial correlation with joint e. Existing methodologies enable the transmission of information from joint d to joint e through joint c. Therefore, we propose to investigate implicit spatial correlations using a cross-spatial learning submodule (SLS) to capture complementary motion information. Taking the example of the joint branch, this submodule utilizes the updated search features (x_k^r) obtained after the r-th iteration as well as the original auxiliary information (x_k) derived from the joints to explore relevant motion cues.

$$\begin{aligned}
&\left[\mathrm{MSA}\left(\mathrm{LN}\left(x_k^r\right)\right) \oplus x_k^r\right] \bullet M_q^r \Rightarrow Q^r, \\
&\left[\mathrm{MSA}\left(\mathrm{LN}\left(x_k\right)\right) \oplus x_k\right] \bullet M_k \Rightarrow K, \\
&\left[\mathrm{MSA}\left(\mathrm{LN}\left(x_k\right)\right) \oplus x_k\right] \bullet M_v \Rightarrow V,
\end{aligned} \tag{1}$$

where $\mathrm{MSA}(\cdot)$ is several parallel self-attention layers. $\mathrm{LN}(\cdot)$ is layer normalization. $\oplus$ denotes residual connection. M_q^r, M_k, and M_v denote the linear mapping matrices. Q, K, and V represent the query matrix, key matrix, and value matrix, receptively. In the first iteration operation x_k^1 is the same as x_k.

The formula for SLS is expressed as follows,

$$x_k^r = \mathrm{softmax}\left(Q^r \bullet K/\sqrt{d}\right) \bullet V \tag{2}$$

where $\mathrm{softmax}(\cdot)$ denotes the softmax activation function, and d is the dimensions of matrices. x_k^r is joint feature of frame k after the r-th iteration.

Finally, we aggregate x_k^r and p_k^r via an adaptive fusion module and then input them into MSA to perform feature extraction. Formally,

$$\bar{x}_s = \mathrm{MSA}\left(\mathrm{LN}\left(\mathrm{AF}\left(x_k^r, p_k^r\right)\right)\right) \tag{3}$$

where $\bar{x}_s$ is a comprehensive matrix that aggregates information from spatial connected joints and joint pairs. $\mathrm{AF}(\cdot)$ denotes adaptive fusion module which selectively aggregates x_k^r and p_k^r according to their feature contributions for prediction.

3.3 Dynamic Temporal Connection Encoding Module

The essence of human motion prediction lies in capturing the underlying temporal evolution relationships to extrapolate motion information. To address this task, it is crucial to consider the temporal evolution by encoding interactions and interdependencies among poses (frames). Traditionally, researchers have employed RNNs and TCNs to capture the temporal evolution from pose sequences. However, these models, which utilize frame-by-frame and sequence-to-sequence methods, are limited in their ability to extract continuous information from rigidly connected adjacent frames and capture implicit patterns of temporal evaluation. Nevertheless, strong connections may also exist between temporally separated poses. For example, in periodic movements such as walking and eating, the hand motion may exhibit high similarity with historical movements. Disregarding these crucial non-consecutive temporal evolution features that are concealed within the inconsecutive pose sequence results in a significant degradation of accuracy. Hence, we propose a dynamic temporal connection encoding module comprising two key components. Firstly, we consider each pose as a time node and disregard the temporal order between poses, which construct a graph structure consisting of a series of time nodes. Consequently, we employ a temporal matrix to enable the model to flexibly encode the temporal relationship between any two time nodes. Formally, given a historical pose sequence which processed by our CSM generate $\bar{X}_{1:\mathrm{T}} = \langle \bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_{\mathrm{T}} \rangle$, the operation of a temporal adjacency matrix can be formulated as,

$$X'_{1:T} = \sigma(\mathrm{D}(\Gamma(\bar{X}_{1:T})\mathrm{A}^T)) \tag{4}$$

where $\Gamma(\cdot)$ denote reshape operation. $\mathrm{A}^{\mathrm{T}} \in R^{T\times T}$ is a temporal adjacency matrix. $\mathrm{D}(\cdot)$ is reshape and mapping operations. $\sigma(\cdot)$ is an activation function. $X'_{1:T} \in R^{J\times 3\times T}$ is a motion feature that aggregates the complementary temporal evolution contexts from time continuous or discontinuous associated poses.

Further, we compresses motion features into a deeper latent representation for promoting the continuation of the temporal information and assist the model in obtaining smoother temporal evolution features. The formula is expressed as,

$$X''_{1:T=}\sigma\left(\mathrm{D}\left('_{1:T}\mathrm{A}^{\mathrm{C}}\right)\right) \tag{5}$$

where $\mathrm{A}^{\mathrm{C}} \in R^{J\times N}$ is a compresses matrix. $X''_{1:T} \in R^{N\times J\times 3}$ is the compressed motion information.

3.4 Loss Functions

Following existing [32,35], we utilize the weighted position loss and bone length loss as our cost functions. Training aims to minimize the L_2 distance between the ground truth and the prediction position for all joints in the future pose sequence.

(1) The weighted position loss ensures the consistency between the forecasted joint position and the ground truth, enhancing the accuracy of our predictions. The formulation of this loss function is as follows:

$$\xi_p = \frac{1}{t \times J} \sum_{k=T+1}^{T+t} \sum_{j=1}^{J} \left\| \left(x_k^j - \tilde{x}_k^j \right) \circ \alpha_k^j \right\|_2 \tag{6}$$

where x_k^j, $\tilde{x}_k^j$, and α_k^j respectively denote the ground truth, prediction, and associated weight of the joint j in frame k.

(2) The weighted loss for bone length is employed to supervise the model in generating consistent predicted bone lengths across different poses, thereby mitigating the occurrence of unnatural movements and inter-frame discontinuities. Formally,

$$\xi_b = \frac{1}{t \times L} \sum_{k=T+1}^{T+t} \sum_{l=1}^{L} \left\| \left(B_k^b - \tilde{B}_k^b \right) \circ \beta_k^b \right\|_2 \tag{7}$$

where B_k^b and $\tilde{B}_k^b$ respectively represent the ground truth and prediction bone length of bone b in frame k. β_k^b denotes the corresponding weight. L is the total number of bones in a pose.

4 Experiments

In this section, we present the results of our experiments, comparing the performance of our model with existing methods. We analyze the key findings, highlighting the advantages of our approach.

4.1 Experimental Settings

4.1.1 Datasets

Human 3.6M is the largest available human motion dataset, includes 3.6 million 3D skeleton poses with 15 actions performed by 7 subjects (e.g., "Directions", "Eating", "Purchases", and "Walking"). Each skeleton pose is represented by the 3D coordinates of 32 joints. This dataset was recorded using a Vicon motion capture system. Following the preprocessing protocol employed in previous studies [32,34,36], we downsampled the frame rate from 50 to 25 frames per second (FPS) and eliminated duplicate and constant joints. We employ 5 subjects (S1, S5, S6, S7, S8) for training, S5 and S11 utilized for testing and validation.

CMU Mocap is a challenging and publicly available collection of human motion data, consisting of 6 categories and 23 subcategories (e.g., "Human Interaction", "Locomotion"). Each human subject is represented by a skeleton comprising 31 joints. Following the methodology employed in previous works [1,12], we select 25 representative joints for each pose and adopt the same training/test split approach. To evaluate our proposed model, we employ a set of 8 actions including "Basketball", "Directing Traffic", "Jumping", and "Washing Window".

4.1.2 Parameter Settings

We implemented our model using PyTorch on an Nvidia Tesla A40 GPU. The Adam optimizer was employed for training the model. The learning rate was initialized as 0.001 with a decay of 0.95 every 2 epochs. The batch size was set to 16, and the gradient clipping threshold was set to 5. The model was trained for 50 epochs. During the training stage, we utilized a sequence of historical poses consisting of 10 frames (400 ms) as input to predict future poses comprising of 25 frames (1000 ms). In the MSM module, there were four spatial encode heads incorporated. The adjacent matrix size for the MSM module was 10×10.

4.1.3 Performance Metric

We adopt the Mean Per Joint Position Error (MPJPE) in millimeter to evaluate the performance of all approaches. MPJPE calculates the spatial distance between the ground truth and predicted joint locations. The formula for MPJPE is given by:

$$\text{MPJPE} = \frac{1}{J}\sum_{i=1}^{J} \|\tilde{x}_i - x_i\|_2 \tag{8}$$

where $\tilde{x}_i$ is the predicted position of the i-th joint, x_i is the ground truth position, and J is the total number of joints in the dataset. As suggested by [1,35,37], we divide the forecast results into short-term (0–400 ms) and long-term predictions (400–1000 ms).

4.2 Comparison with Existing Methods

4.2.1 Human 3.6M

We present a comprehensive summary of the Mean Percentage of Joint Position Error (MPJPR) achieved by various methods on 15 human actions using the Human 3.6M dataset, as shown in Table 2. In total, we compare our method with 11 state-of-the-art models, namely LSTM3LR [11], Res-GRU [10], HP-GAN [38], Bi-GAN [39], ConSeq2Seq [13], HMR [31], LTD [1], DM-GNN [12], DeepSSM [40], Traj-Net [41], and STSGCN [5]. Based on empirical evidence, STM outperforms the alternatives, particularly DeepSSM and STSGCN, in terms of performance. Notably, STM demonstrates a significant average accuracy improvement of 19.7% and 21.3% compared to DeepSSM and STSGCN, respectively. For periodic actions such as "Eating" and "Walking", all algorithms deliver satisfactory prediction results. However, for non-periodic actions like "Photo" and "Purchases", models that encode spatial relationships (e.g., GCN-based models [1,5]) show significant improvements compared to those that solely rely on temporal context (e.g., [10,11]). Unfortunately, even these GCN-based models [1,5,12] fail to capture sufficient spatial information effectively. In contrast, our proposed CSM focuses on exploring multi-level local and global spatial dependencies which consistently enhance performance in complex and challenging motions. So, the main reasons STM outperforms existing models is its multi-level spatial-temporal learning framework. While many traditional models focus on either spatial dependencies or temporal dependencies, STM integrates both at multiple levels, capturing complex relationships that others miss. The Cross Spatial Dependencies Encoding Module (CSM) enables STM to capture fine-grained spatial dependencies between joint pairs, allowing it to better understand human motion, especially in challenging scenarios like non-periodic actions (e.g., "Photo" and "Purchases"). Additionally, STM's ability to learn non-consecutive temporal dependencies through the Temporal Cues Learning strategy improves its long-term prediction accuracy. This is particularly useful for motion sequences where dependencies between frames are not strictly sequential.

Table 2: Comparisons of MPJPE for predictions on the Human 3.6M datase, and the bold entries represent the best results

	Directions					Discussion					Eating					Greeting				
Millisecond (ms)	**80**	**160**	**320**	**400**	**1000**	**80**	**160**	**320**	**400**	**1000**	**80**	**160**	**320**	**400**	**1000**	**80**	**160**	**320**	**400**	**1000**
LSTM3LR [11]	46.6	52.3	87.0	101.6	135.1	29.9	49.4	85.5	105.8	135.1	21.3	38.8	78.1	91.4	121.1	21.0	61.5	73.1	78.2	177.6
Res-GRU [10]	21.6	41.3	72.1	84.1	129.1	25.7	47.8	80.0	91.3	131.8	16.8	31.5	53.5	61.7	98.0	31.2	58.4	96.3	108.8	153.9
HP-GAN [38]	80.9	101.3	148.6	168.8	234.6	71.4	91.3	105.2	129.7	150.4	64.1	78.4	99.9	113.7	136.2	81.5	118.8	178.4	200.1	258.6
Bi-GAN [39]	22.0	37.5	58.9	72.0	114.7	19.2	39.0	67.7	75.3	122.5	13.6	26.1	51.4	63.1	74.1	24.6	45.8	89.9	103.0	148.1
ConSeq2Seq [13]	13.5	29.0	57.6	69.7	115.8	17.1	34.5	64.8	77.6	129.3	11.0	22.4	40.7	48.4	87.1	22.0	45.0	82.0	96.0	147.3
HMR [31]	23.3	25.0	47.2	61.5	116.9	19.0	38.8	67.3	75.0	121.5	13.2	26.0	51.1	62.6	74.0	12.9	31.9	55.6	82.5	123.2
LTD [1]	9.2	20.6	46.9	58.8	108.8	12.2	25.8	53.9	66.7	118.6	9.2	19.5	40.3	48.9	79.5	16.7	33.9	67.5	81.6	140.2
DM-GNN [12]	12.3	23.8	46.2	55.5	90.3	9.7	21.9	39.5	40.0	**68.3**	8.7	18.7	39.5	47.1	57.0	14.0	29.8	74.0	89.1	140.2
DeepSSM [40]	9.7	21.8	47.1	57.5	104.5	8.6	21.3	37.9	43.4	109.2	7.8	15.9	33.9	42.5	71.4	12.5	27.4	68.0	82.8	89.5
Traj-Net [41]	9.7	22.3	50.2	61.7	104.2	7.5	20.0	41.3	47.8	103.0	8.5	18.4	37.0	44.8	71.5	12.6	28.1	67.3	80.1	84.3
STSGCN [5]	14.2	24.3	44.2	53.2	109.9	14.3	24.3	52.6	68.5	107.7	7.3	14.0	37.9	45.0	75.1	15.0	30.7	67.1	87.6	103.8
Our	**6.9**	**15.0**	**33.4**	**41.7**	**79.4**	**6.3**	**16.1**	**35.1**	**41.3**	76.1	**6.5**	**12.9**	**25.8**	**31.7**	**51.5**	**10.6**	**22.2**	**54.0**	**65.4**	**77.9**
	Phoning					**Posing**					**Purchases**					**Smoking**				
Millisecond (ms)	**80**	**160**	**320**	**400**	**1000**	**80**	**160**	**160**	**160**	**1000**	**80**	**160**	**160**	**160**	**1000**	**80**	**160**	**160**	**160**	**1000**
LSTM3LR [11]	39.8	56.4	68.8	70.1	133.3	46.1	72.0	130.1	156.7	176.5	39.1	78.6	88.0	104.1	142.9	26.5	42.3	88.9	99.2	129.2
Res-GRU [10]	21.1	38.9	66.0	76.4	126.4	29.3	56.1	98.3	114.3	183.2	28.7	52.4	86.9	100.7	154.0	18.9	34.7	57.5	65.4	102.1
HP-GAN [38]	78.8	100.3	152.7	179.0	244.2	75.5	107.4	168.3	178.0	250.1	42.4	88.9	95.0	120.2	170.2	67.2	88.6	100.1	123.9	140.4
Bi-GAN [39]	17.0	29.7	54.1	62.1	112.0	16.8	35.0	86.4	105.6	187.0	29.0	54.1	82.2	92.4	139.0	11.0	21.0	33.1	38.2	50.1
ConSeq2Seq [13]	13.5	26.6	49.9	59.9	114.0	16.9	26.7	75.7	92.9	187.4	20.3	41.8	76.5	89.9	151.5	11.6	22.8	41.3	48.9	81.7
HMR [31]	12.5	21.3	39.3	58.6	112.8	13.6	23.5	62.5	114.1	143.6	15.3	30.6	64.7	73.9	122.7	10.3	20.5	33.0	37.2	49.1
LTD [1]	10.2	20.2	40.9	50.9	105.1	12.5	27.5	62.5	79.6	171.7	15.5	32.3	63.6	77.3	135.9	8.4	16.8	32.5	39.5	73.6
DM-GNN [12]	10.2	14.0	32.8	40.0	104.1	9.2	23.5	65.0	82.8	170.2	19.3	38.0	64.2	72.1	112.3	8.2	14.5	25.1	58.8	56.8
DeepSSM [40]	10.6	19.2	35.7	42.5	113.7	7.3	21.3	63.9	80.1	210.3	17.9	39.2	64.8	75.8	120.5	6.4	13.1	24.2	29.6	57.3
Traj-Net [41]	10.7	18.8	37.0	43.1	113.5	6.9	21.3	62.9	78.8	210.9	17.1	36.1	64.3	75.1	115.5	6.3	12.8	23.7	27.8	58.7
STSGCN [5]	14.9	21.4	46.6	52.0	109.9	15.0	25.7	58.4	73.1	**107.6**	15.3	**26.3**	63.5	74.3	119.3	13.1	20.2	37.1	44.7	74.1
Our	**9.7**	**16.6**	**30.4**	**35.2**	**88.7**	**6.5**	**19.3**	**47.3**	**59.8**	164.5	**12.7**	26.4	**44.6**	**52.8**	**98.8**	**4.5**	**9.2**	**19.2**	**24.9**	**50.1**
	Sitting					**Sitting down**					**Taking photo**					**Waiting**				
Millisecond (ms)	**80**	**160**	**160**	**160**	**1000**	**80**	**160**	**160**	**160**	**1000**	**80**	**160**	**160**	**160**	**1000**	**80**	**160**	**160**	**160**	**1000**
LSTM3LR [11]	34.0	57.1	115.0	111.3	142.1	36.9	63.0	88.1	121.3	198.6	35.4	47.5	71.1	74.0	155.6	41.1	57.1	100.1	120.4	159.0
Res-GRU [10]	23.8	44.7	78.0	91.2	152.6	31.7	58.3	96.7	112.0	187.4	21.9	41.4	74.0	87.6	153.9	23.8	44.2	75.8	87.7	135.4
HP-GAN [38]	36.3	60.0	120.0	123.1	168.2	39.9	65.9	92.1	130.0	200.2	38.0	49.3	79.9	83.8	160.4	65.2	98.1	148.3	168.2	199.9
Bi-GAN [39]	19.9	41.0	76.3	88.2	120.5	17.0	34.8	66.5	76.9	152.0	14.2	27.1	53.5	66.1	128.0	17.8	36.4	74.4	90.2	188.9
ConSeq2Seq [13]	13.5	27.0	52.0	63.1	120.7	20.7	40.6	70.4	82.7	150.3	12.7	26.0	52.1	63.6	128.1	14.6	29.7	58.1	69.7	117.7
HMR [31]	12.6	25.6	44.7	60.7	118.4	9.6	**18.6**	**41.1**	57.7	148.3	7.9	19.0	**31.5**	57.3	108.5	12.8	24.5	45.2	85.1	121.9
LTD [1]	10.0	24.6	50.6	62.0	114.9	17.0	33.4	61.6	74.4	144.1	9.9	20.5	43.8	55.2	120.2	10.5	21.6	45.9	57.1	106.9
DM-GNN [12]	10.6	24.4	50.3	61.8	115.3	11.2	27.5	56.1	67.7	143.3	7.1	15.0	38.1	49.5	116.4	9.6	21.8	56.9	71.9	106.9
DeepSSM [40]	9.7	23.3	48.2	61.6	116.4	10.3	26.2	51.7	61.3	131.2	5.2	14.2	38.9	49.9	86.8	8.1	20.3	52.3	67.0	162.7
Traj-Net [41]	9.0	22.0	49.4	62.6	116.3	10.7	28.8	55.1	62.9	123.8	5.4	13.4	36.2	47.0	86.6	8.2	21.0	53.4	68.9	165.9
STSGCN [5]	15.2	23.0	46.8	58.3	119.3	16.7	28.1	56.2	72.0	129.7	16.6	24.8	46.0	61.8	119.8	16.3	27.3	48.1	59.8	118.0
Our	**8.2**	**20.4**	**42.3**	**51.2**	**93.0**	**9.3**	23.2	46.4	**55.2**	**109.0**	**4.8**	**13.2**	35.6	**46.3**	**73.3**	**7.3**	**16.2**	**38.6**	**49.4**	**110.4**

(Continued)

Table 2 (continued)

	Walking					Walking dog					Walking together					Average				
Millisecond (ms)	**80**	**160**	**160**	**160**	**1000**	**80**	**160**	**160**	**160**	**1000**	**80**	**160**	**160**	**160**	**1000**	**80**	**160**	**160**	**160**	**1000**
LSTM3LR [11]	27.5	42.8	80.2	95.5	129.0	77.1	80.1	160.1	190.0	230.3	40.1	55.2	80.3	99.9	180.2	37.5	56.9	88.0	202.4	156.3
Res-GRU [10]	23.2	40.9	61.0	66.1	79.1	36.4	64.8	99.1	110.6	164.5	20.4	37.1	59.4	67.3	98.2	25.0	46.2	77.0	88.3	136.6
HP-GAN [38]	70.1	89.6	98.2	121.0	145.2	83.1	92.1	170.0	198.4	238.8	68.6	79.9	95.3	108.4	188.1	64.2	87.3	123.5	143.1	192.4
Bi-GAN [39]	17.5	31.3	53.9	61.4	89.5	41.2	78.3	116.2	130.1	210.5	14.8	30.1	54.2	65.1	150.2	19.7	37.8	67.9	79.3	132.5
ConSeq2Seq [13]	17.7	33.5	56.3	63.6	82.3	27.7	53.6	90.7	103.3	162.4	15.3	30.4	53.1	61.2	87.4	16.6	33.3	61.4	72.7	124.2
HMR [31]	17.2	31.4	53.5	61.1	89.0	38.2	63.6	109.3	125.6	190.0	12.2	25.2	46.2	50.2	134.1	15.4	28.4	52.8	70.9	118.3
LTD [1]	12.6	23.6	39.4	44.5	60.9	32.2	58.0	102.2	122.7	174.3	10.8	21.7	39.6	47.0	69.6	13.1	26.7	52.7	64.4	115.0
DM-GNN [12]	8.9	14.9	29.0	33.1	50.2	31.8	58.3	101.9	122.4	184.3	8.8	18.0	35.5	44.2	102.5	12.0	24.3	50.3	62.4	107.9
DeepSSM [40]	7.6	15.6	30.2	34.9	48.1	21.9	48.9	89.8	105.4	191.8	**6.8**	**15.7**	31.7	41.1	77.7	10.0	22.9	47.9	58.4	112.7
Traj-Net [41]	8.2	14.9	30.0	35.4	46.4	23.6	52.0	98.1	116.9	181.3	8.5	18.5	33.9	43.4	77.3	10.2	23.2	49.3	59.7	110.6
STSGCN [5]	16.3	24.6	40.1	45.9	64.7	**16.5**	**37.6**	**70.6**	**86.3**	**118.3**	11.4	22.4	39.9	47.5	95.8	14.5	25.0	50.8	62.0	104.9
Our	**7.4**	**13.0**	**25.6**	**29.2**	**37.5**	22.0	43.8	81.9	98.0	143.3	7.4	16.6	**29.4**	**34.3**	**60.7**	**8.7**	**18.9**	**39.3**	**47.8**	**87.6**

We also present visualizations of the prediction results obtained from HMR [31], LTD [1], DM-GNN [12], STSGCN [5], and our proposed method. Fig. 2 illustrates the forecasted outcomes for the actions "Smoking" and "Walking". Our model exhibits superior performance compared to existing methods, generating future pose sequences that closely align with ground truth data. Notably, in the case of the "Smoking" action, our model successfully captures both hand movements and stationary leg states, which are not effectively captured by other models. Consequently, these models exhibit significant errors in forecasting leg positions.

Figure 2: The visual results on Human 3.6M dataset

4.2.2 CMU Mocap

To further evaluate our model, we conducted extensive experiments on the CMU Mocap dataset. Our Spatial and Temporal Learning Model (STM) is compared with six previous methods, namely Res-GRU [10], LTD [1], DM-GNN [12], STSGCN [5], LPJP [42], and TIM [4]. The short-term and long-term prediction performance under the MPJPE metric is presented in Table 3. The proposed STM consistently outperforms the alternatives for both periodic and non-periodic actions. Additionally, our model demonstrates competitive long-term prediction performance with an average improvement of 4.0% over STSGCN and 17.9% over TIM at 1000 ms. These results highlight the benefits of extracting diverse temporal evolution contexts from motion sequences to generate accurate long-term predictions.

Table 3: Comparisons of MPJPE for predictions on the CMU Mocap dataset, and the bold entries represent the best results

	Basketball					Basketball_signal					Directing traffic					Jumping				
Millisecond (ms)	**80**	**160**	**320**	**400**	**1000**	**80**	**160**	**320**	**400**	**1000**	**80**	**160**	**320**	**400**	**1000**	**80**	**160**	**320**	**400**	**1000**
LTD [1]	14	25.4	49.6	61.4	104.8	3.5	6.1	11.7	15.2	45.1	7.4	15.1	31.7	42.2	142.1	16.9	34.4	76.3	96.8	164.6
Res-GRU [10]	15.5	26.9	43.5	**49.2**	**88.73**	20.2	33.0	42.8	44.7	60.6	20.5	40.6	75.4	90.4	153.1	26.9	48.1	93.5	108.9	162.8
DMGNN [12]	15.6	28.7	59.0	73.1	138.6	5.0	9.3	20.2	26.2	52.0	10.21	20.9	41.6	52.3	111.2	32.0	54.3	96.7	119.9	224.6
STSGCN [5]	12.6	23.0	**41.92**	50.3	94.2	4.7	6.7	14.5	17.9	49.5	6.4	12.6	29.1	38.9	109.4	17.5	31.5	**58.74**	**72.06**	**127.4**
LPJP [42]	11.6	21.7	44.4	57.3	90.9	2.6	4.9	12.7	18.7	75.8	6.2	12.7	29.1	39.6	149.1	12.9	27.6	73.5	92.2	176.6
TIM [4]	12.7	22.6	44.6	55.6	102	3	5.6	11.6	15.5	55.1	7.1	14.1	31.1	41.4	138.3	14.8	31.1	71.2	91.3	163.5
Our	**9.2**	**19.8**	43.8	55.6	110.2	**1.9**	**3.5**	**6.5**	**14.0**	**26.1**	**4.3**	**8.3**	**20.1**	**26.8**	**85.5**	**9.4**	**25.3**	66.2	83.2	143.5

	Soccer					Walking					Washwindow					Average				
Millisecond (ms)	**80**	**160**	**320**	**400**	**1000**	**80**	**160**	**160**	**160**	**1000**	**80**	**160**	**160**	**160**	**1000**	**80**	**160**	**160**	**160**	**1000**
LTD [1]	11.3	21.5	44.2	55.8	117.5	7.7	11.8	19.4	23.1	40.2	5.9	11.9	30.3	40	79.3	11.5	20.4	37.8	46.8	96.5
Res-GRU [10]	17.8	31.3	52.6	61.4	107.4	44.4	76.7	126.8	151.4	194.3	22.8	44.7	86.8	104.7	202.7	24.2	43.8	76.2	88.9	139.0
DMGNN [12]	14.9	25.3	52.2	65.4	111.9	9.6	15.5	26.0	30.4	67.0	7.9	14.7	33.3	44.2	82.8	14.07	24.4	45.9	55.5	104.3
STSGCN [5]	13.5	25.2	39.9	51.6	109.6	7.2	11.0	17.8	22.6	44.1	6.8	12.1	24.9	36.7	69.5	10.8	18.2	31.2	41.1	81.8
LPJP [42]	9.2	18.4	39.2	49.5	**93.9**	6.7	10.7	21.7	27.5	37.4	5.4	11.3	29.2	39.6	79.1	9.8	17.6	35.7	45.1	93.2
TIM [4]	11.2	22.1	45.1	58.1	122.1	10.1	11.1	19.9	22.8	39.3	5.9	12.3	32.1	42.6	80.4	10.8	19.5	36.9	46.2	95.7
Our	**7.9**	**16.6**	**36.4**	**45.2**	98.96	**5.7**	**9.6**	**15.7**	**17.6**	**29.4**	**4.1**	**8.6**	**22.4**	**28.8**	**56.5**	**6.1**	**13.1**	**30.2**	**38.8**	**78.6**

The advantage of STM lies in its ability to model both spatial and temporal dependencies across multiple levels. The CSM captures fine-grained spatial dependencies between joint pairs, allowing the model to better handle complex human motion patterns. Unlike models like LTD and DM-GNN, which focus on spatial dependencies but struggle with non-periodic actions, STM excels in these cases due to its integrated temporal learning capabilities. The DTM, through Temporal Cues Learning, captures both consecutive and non-consecutive temporal dependencies, enabling STM to make accurate long-term predictions, especially for non-periodic actions such as "Jumping" and "Direction Traffic." This ability to capture both short- and long-term dependencies contributes to STM's superior performance.

As shown in Fig. 3, we qualitatively compare our STM with HMR [31], LTD [1], DM-GNN [12], and STSGCN [5] on the CMU Mocap dataset. These results demonstrate that our STM produces more authentic future motions for the "Jumping" and "Direction Traffic" actions, exhibiting a higher similarity to the ground truth. In comparison, LTD, DM-GNN, and STSGCN exhibit inaccuracies in certain joints such as the left ankle and right hand during long-term predictions.

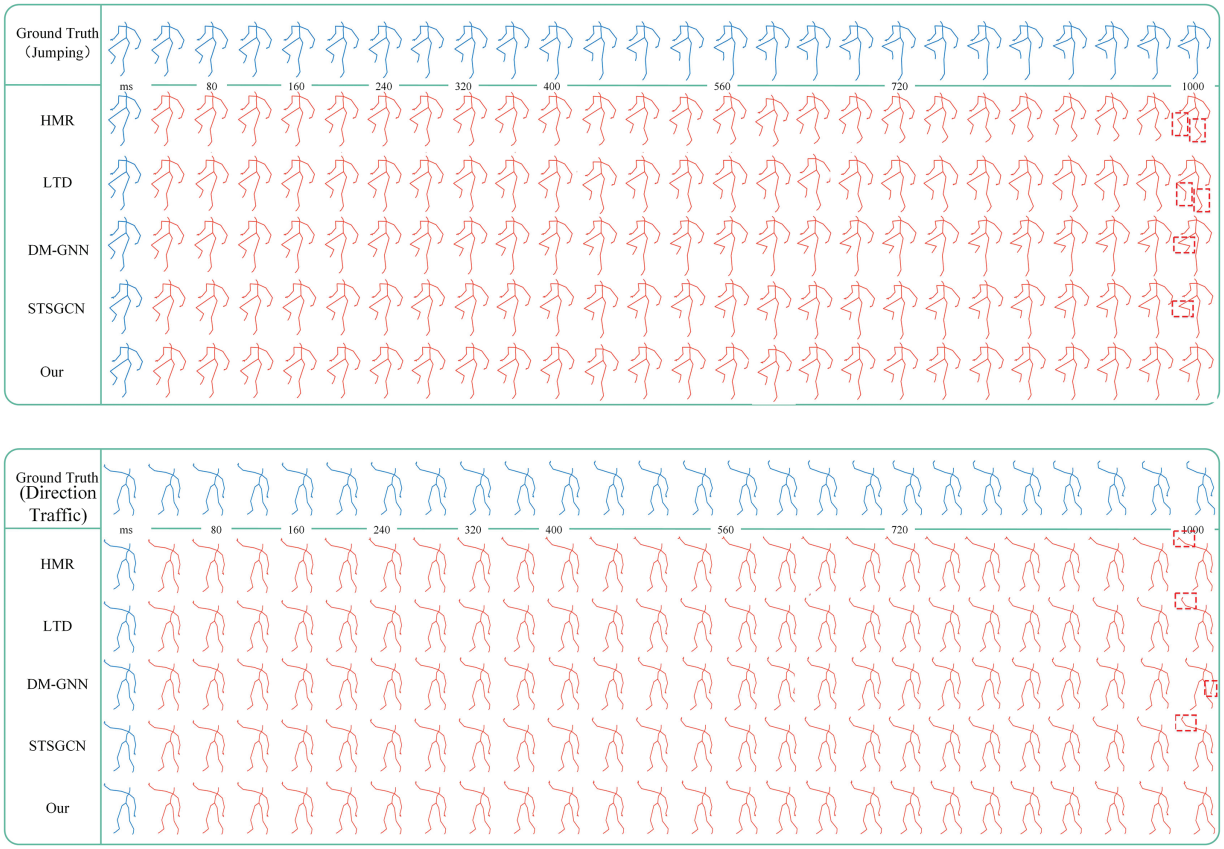

Figure 3: The visual results on CMU Mocap dataset

4.3 Ablation Study

We further investigate the efficacy of individual components comprising the proposed Multi-level Spatial-Temporal Memory (STM) through a series of ablation studies. In this section, we initially analyze the impact of the cross-spatial dependency encoding module and subsequently evaluate the dynamic temporal connection encoding module using Human 3.6M dataset.

4.3.1 Cross Spatial Dependencies Encoding Module

The cross-spatial dependencies encoding block in our model is designed to capture both local and global spatial coherence at joint level and joint pair level. As shown in Table 4, removing this block leads to a significant decrease in short-term prediction accuracy, while only slightly affecting long-term prediction performance. Our focus on capturing spatially correlated features solely from joints or joint pairs results in reduced accuracy, indicating a failure to leverage sufficient and complementary multilevel spatial information.

Table 4: Performance comparision of encoding diverse spatial dependencies

Joint	Joint pair	80	160	320	400	1000
		13.4	27.5	46.2	59.4	98.7
✓		10.5	23.7	44.2	57.4	95.1
	✓	10.1	22.8	45.6	55.7	96.4
✓	✓	8.7	18.9	39.3	47.8	87.6

4.3.2 Dynamic Temporal Connection Encoding Module

The module is specifically designed to capture temporal connections from consecutive or inconsecutive frames and condense the motion sequence to deep level, thereby intensifying the temporal context. As demonstrated in Table 5, removing this module leads to a significant degradation in performance. This confirms that solely exploring temporal evolution from connected frames fails to provide sufficient motion contexts. Additionally, as shown in Table 6, we further investigate the effects of capturing diverse temporal correlations and compressing motion sequences separately. It is observed that both approaches result in an improvement in prediction accuracy.

Table 5: Performance comparision of using DTM

DTM	80	160	320	400	1000
	12.3	25.4	51.1	60.4	97.8
✓	8.7	18.9	39.3	47.8	87.6

Table 6: Comparison of the number of different compress frame

Compress frame	80	160	320	400	1000
10	11.2	23.7	51.2	60.8	99.4
8	10.7	23.8	49.5	58.3	98.5
6	10.3	22.4	46.7	58.4	96.8
5	8.7	18.9	39.3	47.8	87.6
4	9.8	21.8	45.8	55.9	94.6

5 Conclusion

In this paper, we propose a novel algorithm termed Multi-level Spatial and Temporal Learning Model for motion prediction, which consists of a Cross Spatial Dependencies Encoding Module (CSM) and a Dynamic Temporal Connection Encoding Module (DTM). The CSM captures spatial correlation information at the

joint level and joint-pair level, respectively, and then adaptively fuses them, especially when extracting correlation information from the original motion features to avoid noise introduced. The DTM encodes temporal context from temporal continuous or discontinuous frames and compresses motion sequences to a deep level that allows the model to acquire supplementary and smooth temporal evolution patterns. Overall, our approach addresses key limitations of existing methods by effectively capturing both spatial and temporal dependencies across diverse frame sequences. The extensive experiments conducted on the Human 3.6M and CMU Mocap datasets demonstrate that our model achieves state-of-the-art performance, outperforming existing methods by up to 20.3% in accuracy. Furthermore, ablation studies confirm the significant contributions of the CSM and DTM in enhancing prediction accuracy. Our approach not only improves motion prediction accuracy but also provides a robust framework for modeling diverse human actions in dynamic environments.

Acknowledgement: Not applicable.

Funding Statement: This work was supported by the Urgent Need for Overseas Talent Project of Jiangxi Province (Grant No. 20223BCJ25040), the Thousand Talents Plan of Jiangxi Province (Grant No. jxsg2023101085), the National Natural Science Foundation of China (Grant No. 62106093), the Natural Science Foundation of Jiangxi (Grant Nos. 20224BAB212011, 20232BAB212008, 20242BAB25078, and 20232BAB202051), The Youth Talent Cultivation Innovation Fund Project of Nanchang University (Grant No. XX202506030015). This study was funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R759), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Jiayi Geng and Yuxuan Wu; Methodology, Jiayi Geng and Yuxuan Wu; Software, Wenbo Lu, Di Gai and Amel Ksibi; Validation, Jiayi Geng, Yuxuan Wu and Wei Li; Formal Analysis, Jiayi Geng, Zaffar Ahmed Shaikh and Wei Li; Investigation, Yuxuan Wu and Amel Ksibi; Resources, Amel Ksibi and Wei Li; Data Curation, Yuxuan Wu and Amel Ksibi; Writing—Original Draft Preparation, Yuxuan Wu, Di Gai and Wei Li; Writing—Review and Editing, Jiayi Geng; Visualization, Yuxuan Wu and Amel Ksibi; Supervision, Pengxiang Su; Project Administration, Pengxiang Su. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The sources of all datasets are cited in the paper, and can be accessed through the links or GitHub repositories provided in their corresponding papers. Human 3.6M: http://vision.imar.ro/human3.6m (accessed on 23 July 2025), CMU Mocap: http://mocap.cs.cmu.edu (accessed on 23 July 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Mao W, Liu M, Salzmann M, Li H. Learning trajectory dependencies for human motion prediction. In: IEEE/CVF International Conference on Computer Vision, ICCV 2019. Seoul, Republic of Korea; 2019. p. 9488–96. doi:10.1109/ICCV.2019.00958.
2. Xu J, Zhao C, Yang J, Huang Y, Yee L. FDDSGCN: fractional decoupling dynamic spatiotemporal graph convolutional network for traffic forecasting. In: IEEE International Conference on Acoustics, Speech and Signal Processing. Hyderabad, India; 2025. p. 1–5. doi:10.1109/ICASSP49660.2025.10888084.
3. Ding R, Qu KH, Tang JKSOF. Leveraging kinematics and spatio-temporal optimal fusion for human motion prediction. Pattern Recognit. 2025;161(8):111206. doi:10.1016/J.PATCOG.2024.111206.
4. Lebailly T, Kiciroglu S, Salzmann M, Fua P, Wang W. Motion prediction using temporal inception module. In: Computer Vision-ACCV 2020-15th Asian Conference on Computer Vision. Kyoto, Japan; 2020. p. 651–65. doi:10.1007/978-3-030-69532-3_39.

5. Sofianos TS, Sampieri A, Franco L, Galasso F. Space-time-separable graph convolutional network for pose forecasting. In: IEEE/CVF International Conference on Computer Vision. Montreal, QC, Canada; 2021. p. 11189–98. doi:10.1109/ICCV48922.2021.01102.
6. Yu T, Lin Y, Yu J, Lou J, Gui Q. Vision-guided action: human motion prediction with gaze-informed affordance in 3D scenes. In: Proceedings of the Computer Vision and Pattern Recognition Conference; 2025. p. 12335–46.
7. Lehrmann AM, Gehler PV, Nowozin S. Efficient nonlinear markov models for human motion. In: IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA; 2014. p. 1314–21. doi:10.1109/CVPR.2014.171.
8. Wang JM, Fleet DJ, Hertzmann A. Gaussian process dynamical models. In: Advances in Neural Information Processing Systems; 2005 Dec 5–8; Vancouver, BC, Canada; 2005. p. 1441–8. doi:10.1109/TPAMI.2007.1167.
9. Taylor GW, Hinton GE, Roweis ST. Modeling human motion using binary latent variables. In: Advances in Neural Information Processing Systems. Vancouver, BC, Canada: MIT Press; 2006. p. 1345–52. doi:10.7551/mitpress/7503.003.0173.
10. Martinez J, Black MJ, Romero J. On human motion prediction using recurrent neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA; 2017. p. 4674–83. doi:10.1109/CVPR.2017.497.
11. Fragkiadaki K, Levine S, Felsen P, Malik J. Recurrent network models for human dynamics. In: IEEE International Conference on Computer Vision; ICCV 2015. Santiago, Chile; 2015. p. 4346–54. doi:10.1109/ICCV.2015.494.
12. Li M, Chen S, Zhao Y, Zhang Y, Wang Y, Tian Q. Dynamic multiscale graph neural networks for 3D skeleton based human motion prediction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA; 2020. p. 211–20. doi:10.1109/CVPR42600.2020.00029.
13. Li C, Zhang Z, Lee WS, Lee GH. Convolutional sequence to sequence model for human dynamics. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018). Salt Lake City, UT, USA; 2018. p. 5226–34. doi:10.1109/CVPR.2018.00548.
14. 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; 2020. p. 474–89. doi:10.1007/978-3-030-58568-6_28.
15. Zhong C, Hu L, Zhang Z, Ye Y, Xia S. Spatio-temporal gating-adjacency gcn for human motion prediction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022). New Orleans, LA, USA; 2022. p. 6437–46. doi:10.1109/CVPR52688.2022.00634.
16. Wang X, Zhang W, Wang C, Gao Y, Liu M. Dynamic dense graph convolutional network for skeleton-based human motion prediction. IEEE Trans Image Process. 2023;33:1–15. doi:10.1109/TIP.2023.3334954.
17. Chen H, Hu J, Zhang W, Su P. Spatiotemporal consistency learning from momentum cues for human motion prediction. IEEE Trans Circuits Syst Video Technol. 2023;33(9):4577–87. doi:10.1109/TCSVT.2023.3284013.
18. Yang J, Tian K, Zhao H, Zheng F, Sami B, Sami D, et al. Wastewater treatment monitoring: fault detection in sensors using transductive learning and improved reinforcement learning. Expert Syst Appl. 2025;264(5):125805. doi:10.1016/j.eswa.2024.125805.
19. Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans, LA, USA; 2018. p. 7444–52.
20. Cui Q, Sun H, Yang F. Learning dynamic relationships for 3D human motion prediction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020). Seattle, WA, USA; 2020. p. 6518–26. doi:10.1109/CVPR42600.2020.00655.
21. Yang J, Wu Y, Yuan Y, Xue H, Abdel-Salam M, Prajapat S, et al. Web attack detection using a large language model with autoencoder and multilayer perceptron. Expert Syst Appl. 2025;274(4):126982. doi:10.1016/j.eswa.2025.126982.
22. Zhao M, Tang H, Xie P, Dai S, Sebe N, Wang W. Bidirectional transformer gan for long-term human motion prediction. ACM Trans Multim Comput Commun Appl. 2023;19(5):1–19. doi:10.1145/3579359.
23. Yu H, Fan X, Hou T, Pei W, Ge H, Yang X, et al. Towards realistic 3D human motion prediction with a spatio-temporal cross-transformer approach. IEEE Trans Circuits Syst Video Technol. 2023;33(10):5707–20. doi:10.1109/TCSVT.2023.3255186.

24. Zhu L, Lu X, Cheng Z, Li J, Zhang H. Deep collaborative multi-view hashing for large-scale image search. IEEE Trans Image Process. 2020;29:4643–55. doi:10.1109/TIP.2020.2974065.
25. Yang X, Zhou P, Wang M. Person reidentification via structural deep metric learning. IEEE Trans Neural Networks Learn Syst. 2019;30(10):2987–98. doi:10.1109/TNNLS.2018.2861991.
26. Feng F, He X, Tang J, Chua T. Graph adversarial training: dynamically regularizing based on graph structure. IEEE Trans Knowl Data Eng. 2021;33(6):2493–504. doi:10.1109/TKDE.2019.2957786.
27. You S, Zhou Y. Optimization driven cellular automata for traffic flow prediction at signalized intersections. J Intell Fuzzy Syst. 2021;40(1):1547–66. doi:10.3233/JIFS-192099.
28. Liang C. Prediction and analysis of sphere motion trajectory based on deep learning algorithm optimization. J Intell Fuzzy Syst. 2019;37(5):6275–85. doi:10.3233/JIFS-179209.
29. German B, Sergio E, Cristina P. Belfusion: latent diffusion for behavior-driven human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2023). Paris, France; 2023. p. 2317–27. doi:10.1109/ICCV51070.2023.00220.
30. Jain A, Zamir AR, Savarese S, Saxena A. Structural-RNN: deep learning on spatio-temporal graphs. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016). Las Vegas, NV, USA; 2016. p. 5308–17. doi:10.1109/CVPR.2016.573.
31. Liu Z, Wu S, Jin S, Liu Q, Lu S, Zimmermann R, et al. Towards natural and accurate future motion prediction of humans and animals. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019). Long Beach, CA, USA; 2019. p. 10004–12. doi:10.1109/CVPR.2019.01024.
32. Liu Z, Su P, Wu S, Shen X, Chen H, Hao Y, et al. Motion prediction using trajectory cues. In: IEEE/CVF International Conference on Computer Vision (ICCV 2021). Montreal, QC, Canada; 2021. p. 13279–88. doi:10.1109/ICCV48922.2021.01305.
33. Dang L, Nie Y, Long C, Zhang Q, Li G. MSR-GCN: multi-scale residual graph convolution networks for human motion prediction. In: IEEE/CVF International Conference on Computer Vision (ICCV 2021). Montreal, QC, Canada; 2021. p. 11447–56. doi:10.1109/ICCV48922.2021.01127.
34. Li M, Chen S, Zhang Z, Xie L, Tian Q, Zhang Y. Skeleton-parted graph scattering networks for 3D human motion prediction. In: Computer Vision-ECCV 2022-17th European Conference; 2022 Oct 23–27. Tel Aviv, Israel; 2022. p. 18–36. doi:10.1007/978-3-031-20068-7_2.
35. Su P, Shen X, Shi Z, Liu W. Adaptive multi-order graph neural networks for human motion prediction. In: IEEE International Conference on Multimedia and Expo (ICME 2022). Taipei, Taiwan; 2022. p. 1–6. doi:10.1109/ICME52920.2022.9859980.
36. Tseng KW, Kawakami R, Ikehata S. CST: character state transformer for object-conditioned human motion prediction. In: Proceedings of the Winter Conference on Applications of Computer Vision. Tucson, AZ, USA; 2025. p. 63–72. doi:10.1109/WACVW65960.2025.00012.
37. Zhang Y, Kephart JO, Ji Q. Incorporating physics principles for precise human motion prediction. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, HI, USA; 2024. p. 6164–74. doi:10.1109/WACV57701.2024.00605.
38. Barsoum E, Kender JR, Liu Z. HP-GAN: probabilistic 3D human motion prediction via GAN. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops 2018). Salt Lake City, UT, USA; 2018. p. 1418–27. doi:10.1109/CVPRW.2018.00191.
39. Kundu JN, Gor M, Babu RV. BiHMP-GAN: bidirectional 3D human motion prediction GAN. In: AAAI Conference on Artificial Intelligence. Honolulu, HI, USA; 2019. p. 8553–60. doi:10.1609/AAAI.V33I01.33018553.
40. Liu X, Yin J, Liu J. AGVNet: attention guided velocity learning for 3D human motion prediction. arxiv:2005.12155. 2020. doi:10.48550/arxiv.2005.12155.
41. Liu X, Yin J, Li J, Ding P, Liu J, Liu H. TrajectoryCNN: a new spatio-temporal feature learning network for human motion prediction. IEEE Trans Circuits Syst Video Technol. 2021;31(6):2133–46. doi:10.1109/TCSVT.2020.3021409.
42. Cai Y, Huang L, Wang Y, Cham T, Cai J, Yuan J, et al. Learning progressive joint propagation for human motion prediction. In: Computer Vision-ECCV 2020-16th European Conference; 2020 Aug 23–28. Glasgow, UK; 2020. p. 226–42. doi:10.1007/978-3-030-58571-6_14.