*electronics*

MDPI

*Article*

# 3D Human Motion Prediction via the Decoupled Spatiotemporal Clue

**Mingrui Xu [1,*], Zheming Gu [2] and Erping Li [2]**

[1] ZJU-UIUC Institute, Zhejiang University, HaiNing Campus, Haining 314400, China
[2] College of Information Science and Electronic Engineering, Zhejiang University, Yuquan Campus, Hangzhou 310027, China
* Correspondence: 22371202@zju.edu.cn

**Abstract**

Human motion exhibits high-dimensional and stochastic characteristics, posing significant challenges for modeling and prediction. Existing approaches typically employ coupled spatiotemporal frameworks to generate future poses. However, the intrinsic nonlinearity of joint interactions over time, compounded by high-dimensional noise, often obscures meaningful motion features. Notably, while adjacent joints demonstrate strong spatial correlations, their temporal trajectories frequently remain independent, adding further complexity to modeling efforts. To address these issues, we propose a novel framework for human motion prediction via the **decoupled spatiotemporal clue** (DSC), which explicitly disentangles and models spatial and temporal dependencies. Specifically, DSC comprises two core components: (i) a *spatiotemporal decoupling module* that dynamically identifies critical joints and their hierarchical relationships using graph attention combined with separable convolutions for efficient motion decomposition; and (ii) a *pose generation module* that integrates local motion denoising with global dynamics modeling through a spatiotemporal transformer that independently processes spatial and temporal correlations. Experiments on the widely used human motion datasets H3.6M and AMASS demonstrate the superiority of DSC, which achieves 13% average improvement in long-term prediction over state-of-the-art methods.

**Keywords:** 3D human motion prediction; decoupled spatiotemporal model; graph neuronal network; self-attention mechanisms

check for updates

Academic Editor: José Carlos Castillo

Received: 17 September 2025
Revised: 15 October 2025
Accepted: 22 October 2025
Published: 24 October 2025

**Citation:** Xu, M.; Gu, Z.; Li, E. 3D Human Motion Prediction via the Decoupled Spatiotemporal Clue. *Electronics* **2025**, *14*, 4162. https://doi.org/10.3390/electronics14214162

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

Human motion prediction is a complex spatiotemporal modeling issue that has been a longstanding and challenging research topic in the field of computer vision. Due to its importance in modeling and understanding human behavior, 3D human motion prediction has applications in areas such as animation [1], motion synthesis [2], and human–computer interaction [3–5]. Future extensions may enable deployment in more complex environments, including autonomous systems and augmented/virtual reality.

Accurate prediction of human motion relies on the effective modeling of spatiotemporal correlations, joint interactions, and temporal dynamics. Current approaches mainly employ strategies: temporal modeling through temporal convolutional networks (TCNs) [6,7], recurrent neural networks (RNNs) [8–11], gated recurrent units (GRUs) [12,13], long short-term memory networks (LSTMs) [14], or Transformer [15–17]; spatial modeling via graph-based methods, like graph convolutional networks (GCNs) [18,19], with advanced variants,

such as dynamic multi-scale graph neural networks (DMGNN) [20] and multi-scale residual GCN (MSR-GCN) [21]), capture multi-scale relationships. While such isolated spatial or temporal modeling reduces complexity, it inherently fails to capture the nuanced interplay of human kinematics. In reality, the high-dimensional stochasticity of human motion weakens idealized spatiotemporal correlations (as shown in Figure 1). Our analysis reveals a critical issue: while adjacent joints exhibit strong spatial coupling, their temporal trajectories evolve with surprising independence. This divergence implies that forcibly coupling spatiotemporal features, as in existing methods, risks obscuring the discriminative patterns within noise, demanding more principled architectural designs. To address the confusion caused by forced spatiotemporal feature coupling in existing models, we propose a novel prediction framework that can explicitly comprehend relationships via ***D**ecoupled **S**patiotemporal **C**lue* (**DSC**). DSC consists of two primary components that together enable precise future pose synthesis: a *spatiotemporal decoupling module* (**SDM**) for independent feature learning, and a *pose generation module* (**PGM**) for specialized temporal and spatial modeling.

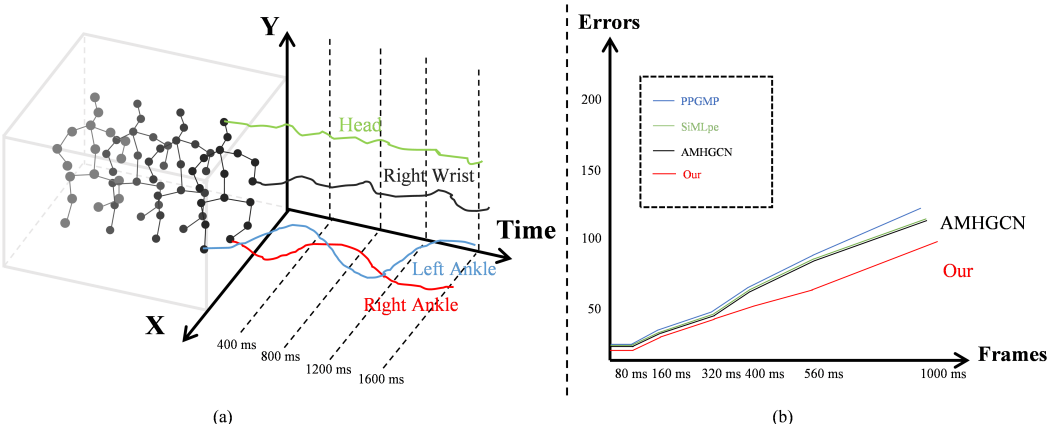



**Figure 1.** Spatial correlation and temporal independence in human motion. (**a**) When a person is walking, there exists intrinsic spatial correlations among body joints, yet each joint exhibits independent motion trajectories over time. (**b**) The performance comparison of different methods on the H3.6M [22] dataset based on average results under the Mean Per Joint Position Error (MPJPE) metric. PPGMP [23]: Pose Propagation with Graph-based Motion Prediction; SiMLpe [24]: Simplified Multi-Layer Pose Estimation; and AMHGCN [25]: Adaptive Multi-Hierarchy Graph Convolutional Network.

The **SDM** explicitly models the independence of the joint motion trajectories within human pose representations. Utilizing a decoupled spatiotemporal encoding scheme, this module effectively captures the distinct spatiotemporal feature relationships through two complementary components: (i) a *graph attention component* (GAC) that dynamically learns multi-level joint dependencies, and (ii) a *spatiotemporal decoupling encoding component* (SDEC) for hierarchical feature extraction. The *GAC* overcomes limitations of conventional graph convolutional networks by replacing fixed skeletal graphs with an attention mechanism. This enables the adaptive modeling of cross-joint dependencies across semantic levels, where critical joints automatically receive higher weights based on their topological and functional significance. The *SDEC* employs parallel temporal and spatial processing streams. Temporal convolutions capture the motion trajectory features along the time dimension, while spatially constrained convolutions preserve the kinematic relationships between joints. Through an interpretation of channel dimensions as temporal scales, the SDEC achieves efficient multi-resolution feature extraction while strictly decoupling the temporal dynamics from structural semantics. These components generate a disentangled representation that captures both the temporal evolution and spatial configuration of human motion.

The aim of the **PGM** is to accurately predict future human pose sequences through independent temporal and spatial dependencies. This module consists of two parts: a *temporal feature extraction* (TFE), and a *spatiotemporal decoupled Transformer* (STDT). The *TFE* stage utilizes two 1D convolutions with rectified linear unit (ReLU) activations, sliding solely along the temporal axis to capture the local motion patterns between adjacent frames without incorporating spatial dimensions. This processing attenuates high-frequency noise while standardizing local features, thereby providing cleaner, higher-order inputs for subsequent deep attention mechanisms. Within the *STDT*, we depart from conventional Transformer architectures by designing four distinct, independently parameterized attention modules: spatial, temporal, spatiotemporal crossover, and global. Each module learns its own projection matrices, preventing spatiotemporal feature coupling during parameter updates. This explicit decoupling enables the direct visualization of attention weights, clarifying the source of learned dependencies. Critically, because the spatial module explicitly models the joint topology, the temporal module is dedicated solely to dynamic evolution. We posit that single-head attention acts as an effective "temporal smoother" for spatial features, whereas multi-head attention may introduce noisy interactions. Consequently, the STDT employs a single-head design. Following TFE filtering, the STDT independently models the spatial dependencies among joints and their temporal evolution, while also capturing complex cross-spatiotemporal dependencies through its dedicated modules, ensuring accurate human motion modeling.

**Contributions**. Our main contributions are as follows:

- We propose a novel approach for human motion prediction that explicitly decouples the joint topological relationships from trajectory dynamics through the decomposition of spatial and temporal dependencies.
- We propose two co-designed modules—a spatiotemporal decoupling module that leverages multi-semantic graph attention, and hierarchical decomposition for isolated feature extraction. A pose generation module that integrates noise-robust temporal feature learning with a single-head Transformer, featuring four dedicated attention pathways to avoid feature contamination, was also utilized.
- Our method achieves state-of-the-art results on two challenging benchmark datasets, H3.6M [22] and AMASS [26].

## 2. Related Work

Human motion prediction has always been a challenging research topic. We discuss related work by distinguishing between traditional motion prediction methods and the recently emerging spatiotemporal decoupling approaches.

**Human Motion Prediction.** Traditional methods include Gaussian Process Dynamical Models [27], Restricted Boltzmann Machines [28], Markov Models [29], Nonlinear Markov Models [30], and other nonlinear frameworks [31]. With the rise of deep learning, significant progress has been made in this field. Deep learning-based approaches can be roughly divided into three paradigms: those based on either recurring neural networks (RNNs), convolutional neural networks (CNNs), or graph convolutional networks (GCNs). Additionally, Generative Adversarial Networks (GANs) and Transformers have also contributed. Early RNN-based methods [32,33] modeled temporal dependencies by encoding skeleton sequences as vectors but neglected the spatial correlations between joints. The widely used encoder–decoder architecture within RNN frameworks [32–34] introduced error accumulation issues [9] and is prone to converge toward static average poses. Despite some improvements, RNN-based methods still suffer from motion discontinuities in long-term predictions [35] and static pose bias problems [36]. CNN-based methods [3,36–38] have emerged as an alternative but have limited performance due to

their inability to model the intrinsic dependencies between joints. Temporal convolutional networks (TCNs) [39] offer beneficial robustness in time modeling [40,41] and excel in capturing local temporal features. Advanced motion prediction methods increasingly adopt graph convolutional networks (GCNs) to model the dependencies between skeleton joints [18–21,42,43]. Research on spatiotemporal graphs [7,8,20,40,42,44–46] has become mainstream, explicitly encoding skeletal topology and motion dynamics through joint spatiotemporal representations. Moreover, GAN-based methods [35,47] show potential but still face optimization challenges [21] and are sensitive to noise [36]. After achieving success in Natural Language Processing (NLP) and computer vision, Transformers have been adapted for 3D human motion prediction [15,48–51]. For example, HisRep [8] combined graph convolution with discrete cosine transforms to encode spatiotemporal information, and they used self-attention mechanisms to align historical motion segments.

**Spatiotemporal Decoupling for Human Motion Prediction.** The adjacent joints of a human pose inherently exhibit spatial correlations, whereas joint movement trajectories over time are relatively independent. Previous efforts have attempted to process temporal and spatial features independently [7,15,46,50], achieving promising results. In the first spatiotemporal graph convolutional network (ST-GCN) [46] designed for human action recognition using an encoder–decoder architecture, joints are treated as nodes, the natural topologies (i.e., skeletal connections) between adjacent joints are treated as spatial edges, and the identical node connections across adjacent frames are treated as temporal edges. The generated spatiotemporal graph embedding human action features is then fed into ST-GCN to extract the relevant human action features, spatial dimensions (through spatial GCN layers), and temporal dimensions (through TCN layers). Based on ST-GCN, a new spatiotemporal separable graph convolutional network (STS-GCN) [7] has been proposed to predict human motion, utilizing graph adjacency matrix decomposition to separately extract the temporal and spatial features. These are examples of attempts to decouple spatiotemporal features using GCNs. Previous works have also tried using Transformers for spatiotemporal decoupling. The spatiotemporal Transformer (ST-Trans) [15] and the spatiotemporal deformable transformer adaptive network (STDTA) [50] include the design of two attention modules to decompose the spatial and temporal dependencies, achieving effective modeling. Inspired by previous work, our approach includes designing a spatiotemporal decoupling module and a pose generation module. The spatiotemporal decoupling module employs spatial and temporal convolutions to decouple spatiotemporal features, while the pose generation module designs four attention modules for more accurate human pose modeling.

## 3. Methodology

**Problem Formulation.** In modeling a human pose, we focus on decoupling extraction and analyses of joint information. The observed human poses over $T$ frames are represented as $X_{\text{in}} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T]$, and the predicted poses for the subsequent $N$ frames are denoted as $X\text{out} = [\mathbf{x}_{T+1}, \mathbf{x}_{T+2}, \ldots, \mathbf{x}_{T+N}]$, where $\mathbf{x}_i$ represents the 3D coordinates ($x$, $y$, $z$) of $V$ body joints. Most existing methods adopt spatiotemporal coupling strategies for human motion encoding and modeling. However, while adjacent joints exhibit strong spatial correlations due to the anatomical structure of the human body, their temporal trajectories tend to evolve independently. The combination of these two modalities can introduce redundant noise, increase model complexity, and limit performance.

**Method Overview.** The overall architecture of the proposed framework is illustrated in Figure 2, consisting of two key components: the **spatiotemporal decoupling module** (Section 3.1), and the **pose generation module** (Section 3.2). Specifically, the input pose sequence is first processed by the spatiotemporal decoupling encoding module, which

separates spatial and temporal features to more effectively capture the relationships across different dimensions. Subsequently, within the pose generation module, the *temporal feature extractor* (TFE) captures the local motion patterns between consecutive frames and filters out high-frequency noise. Finally, the *spatiotemporal decoupled transformer* (STDT) models the spatial dependencies among joints, as well as their temporal evolution, while simultaneously capturing complex cross-dimensional spatiotemporal dependencies, thereby enhancing the accuracy of human pose modeling. The following sections provide detailed descriptions of these two modules.

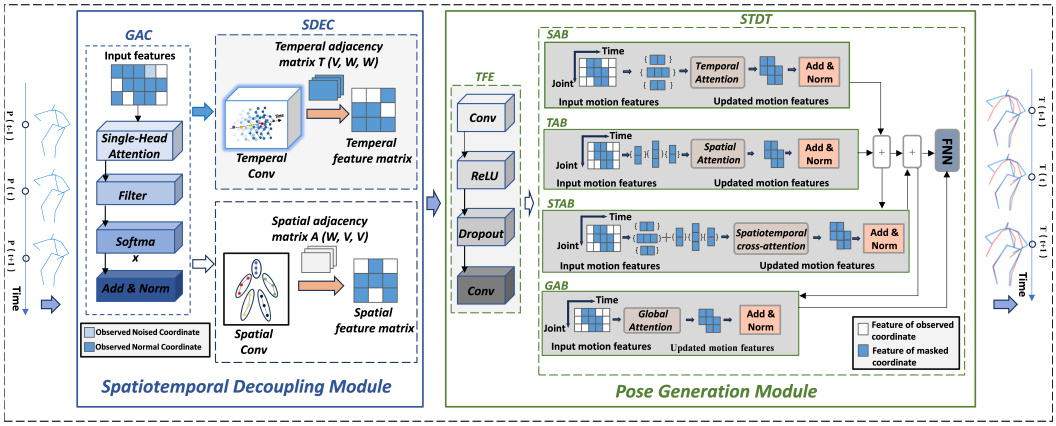



**Figure 2.** Our proposed framework via the decoupled spatiotemporal clue: the input historical pose sequence is first processed by the graph attention component (GAC) to dynamically model the spatial dependencies among joints, generating enhanced features. Then, the spatiotemporal decoupling encoding component (SDEC) employs parallel temporal and spatial convolutional branches to separately extract the temporal evolution patterns and spatial structural features, effectively decoupling spatial and temporal information. Subsequently, the temporal feature extractor (TFE) captures local motion dynamics, while the spatiotemporal decoupled transformer (STDT) models the spatial, temporal, cross-dimensional, and global dependencies through four dedicated attention branches (SAB, TAB, STAB, and GAB). Finally, the fused features are passed through a feed-forward Network (FFN) to generate the predicted future pose sequence.

### *3.1. Spatiotemporal Decoupling Module*

As previously discussed, most existing approaches adopt spatiotemporally coupled strategies for modeling human motion encoding. However, the temporal trajectories of different joints often evolve independently, suggesting that coupling spatial and temporal information may result in independent features being overwhelmed by noise, thereby limiting model expressiveness.

To address this issue, we propose the **spatiotemporal decoupling module**, which aims to model the independence of joint motion trajectories in human pose representations. The module consists of two complementary components: the *graph attention component* (GAC), and the *spatiotemporal decoupling encoding component* (SDEC).

**Graph Attention Component.** Our GAC dynamically learns multi-semantic joint dependencies by introducing an attention mechanism that replaces the fixed skeleton graph used in traditional graph convolutional networks (GCNs), thereby overcoming their limitations in modeling expressive and context-aware features. Unlike static graphs that assume rigid connectivity, our approach enables the adaptive modeling of skeletal topology, where key joints automatically receive higher attention weights based on their topological centrality and functional significance in specific actions. To further improve robustness and physical plausibility, we constructed a dynamic spatial adjacency matrix $A_{\text{space}}$ that facilitates more accurate message passing while suppressing spurious or noisy connections in the graph structure.

Specifically, the dynamic matrix $A_{\text{space}}$ models joint relationships by integrating three components: the static skeleton adjacency matrix $A_{\text{ske}}$, a predefined semantic adjacency matrix $A_{\text{pre_sem}}$ , and a learnable semantic adjacency matrix $A_{\text{sem}}$ that captures data-driven dependencies. Their combination is nonlinearly transformed via the Sigmoid function to ensure stable gradient propagation:

$$A_{\text{space}} = \sigma(A_{\text{ske}} + A_{\text{pre_sem}} + \alpha A_{\text{sem}}). \tag{1}$$

Here, $\alpha$ is a hyperparameter that balances fixed anatomical priors with learned, action-specific relationships (typically set to $\alpha = 1$). This hybrid design ensures that the learned graph respects fundamental biomechanical constraints while adapting to dynamic motions. After processing the input with GAC, we proceeded to the SDEC for further encoding.

**Spatiotemporal Decoupling Encoding Component**. The *SDEC* employs parallel temporal and spatial processing streams. *Temporal convolution* captures the motion trajectory features along the time dimension, while *spatial convolution* preserves the kinematic relationships between joints.

*Temporal Convolution.* Temporal convolution aggregates the motion trajectory features along the time dimension. By treating the channel dimension as a temporal scale, it enables efficient multi-resolution feature extraction over time. Specifically, we define a temporal convolution kernel $\mathbf{T} \in \mathbb{R}^{V \times T_{\text{in}} \times T_{\text{out}}}$, where $V$ corresponds to the number of joints and $T_{\text{in}}$ and $T_{\text{out}}$ denote the input and output time steps, respectively. Using Einstein summation notation (Einsum), we combine the input features $\mathbf{X}$ with the temporal kernel $\mathbf{T}$ to generate temporal features $\mathbf{X}_{\text{t}}$:

$$\mathbf{x}_{\text{t}}^{(n,c,q,v)} = \text{Einsum}\left(\sum_{t=1}^{T_{\text{in}}} \mathbf{x}^{(n,c,t,v)} \cdot \mathbf{T}^{(v,t,q)}\right). \tag{2}$$

This operation aggregates historical information across time for each joint $v$, focusing on motion trajectory patterns and effectively encoding temporal dependencies.

*Spatial Convolution.* The spatial convolution, on the other hand, models the spatial relationships among joints, preserving the biomechanical structure of the human body. Specifically, we define a spatial convolution kernel $\mathbf{A} \in \mathbb{R}^{T_{\text{out}} \times V \times V}$, which is used to model the spatial relationships between joints. This operation is implemented as a learnable graph convolution that explicitly respects the skeletal topology through a masked adjacency matrix, ensuring that information exchange only occurs between anatomically connected joints.

The effective kernel size is $V \times V = 25 \times 25$, but it remains sparse due to the structural mask. Padding is not applicable (as this is a graph convolution, i.e., not grid-based). The stride is 1, with full connectivity in the spatial dimension. The key constraint is that only adjacent joints in the skeleton graph are allowed to interact, which enforces kinematic plausibility.

Similarly, using Einsum, we combine the temporally convolved features $\mathbf{X}_{\text{temp}}$ with the spatial kernel $\mathbf{A}$ to extract spatial features:

$$\mathbf{x}_{\text{s}}^{(n,c,t,w)} = \text{Einsum}\left(\sum_{v=1}^{V} \mathbf{x}_{\text{temp}}^{(n,c,t,v)} \cdot \mathbf{A}^{(t,v,w)}\right). \tag{3}$$

Through this process, we effectively model the spatial topology of the human skeleton and extract meaningful spatial features.

In summary, the SDEC effectively captures diverse spatiotemporal feature relationships through its two complementary components, laying a solid foundation for accurate future pose prediction.

### *3.2. Pose Generation Module*

We propose the pose generation module aims to accurately predict future human pose sequences by separately modeling temporal and spatial dependencies. The module adopts a hierarchical architecture, consisting of two core components: the *temporal feature extraction component* (TFE), and the *spatiotemporally decoupled transformer* (STDT).

**Temporal Feature Extraction Component.** The TFE employs two one-dimensional convolutional kernels with ReLU activation functions, which slide only along the temporal axis without involving the spatial dimension. This design effectively captures the local motion patterns between consecutive frames. Each 1D convolution has a kernel size of 3, a stride of 1, and a padding of 1, ensuring temporal resolution is preserved and local dynamics (e.g., velocity and acceleration) are captured within a short context. The number of input channels is $C_{\text{in}}$ (e.g., 3 for 3D coordinates), and the hidden dimension is set to 64. In this formulation, $h_{n,t}$ denotes the hidden state of joint $n$ at time step $t$, $W_\tau$ represents the weight matrix associated with time offset $\tau$, and $x_{n,t+\tau}$ is the input feature of joint $n$ at time step $t+\tau$. The output feature after the first convolution is denoted as $y_n$, where $W'$ and $b'$ are additional learnable weights and bias, respectively. The original input used in the skip connection is represented by $x_n$. The computational process is defined as follows:

$$h_{n,t} = \text{ReLU}(W_\tau x_{n,t+\tau} + b_\tau), \tag{4}$$

$$y_n = W' h_{n,t} + b' + x_n. \tag{5}$$

Our design not only filters out high-frequency noise during temporal feature extraction, but it also normalizes local features, providing clean and expressive high-level inputs for the deeper attention mechanisms that follow.

**Spatiotemporally Decoupled Transformer.** Our STDT is capable of separately modeling the spatial dependencies among joints and their temporal evolution, while simultaneously capturing complex cross-dimensional spatiotemporal dependencies. In contrast to traditional Transformers, we designed four independent attention branches: the *spatial attention branch* (SAB), the *temporal attention branch* (TAB), the *spatiotemporal cross-Attention branch* (STAB), and the *global attention branch* (GAB). Each attention branch learns its own projection matrices independently, effectively preventing the coupling of spatial and temporal features during parameter updates. Moreover, since the spatial module explicitly models the joint topology, the temporal module can focus solely on dynamic evolution. We argue that single-head attention can act as a "temporal smoother" for spatial features, whereas multi-head attention may introduce redundant or noisy interactions. Therefore, our STDT adopts a single-head attention mechanism. Finally, the fused features are further processed by a feed-forward network (FFN) to produce the final output. We will now provide a detailed description of each component.

*Spatial attention branch.* In SAB, we compute the self-attention across spatial dimensions within the same time step $t$. During processing, we divide the TFE outputs into time slices, $X_t = X[:, t, :, :] \in \mathbb{R}^{B \times S \times C}$, and we then compute the query, key, and value as follows:

$$Q_t = X_t W_Q^{\text{space}}, K_t = X_t W_K^{\text{space}}, V_t = X_t W_V^{\text{space}}. \tag{6}$$

We share projection weights for keys and values across joints. Notably, in graph-structured data, such as for human skeletons, non-adjacent joints (e.g., left and right feet) typically lack

direct mechanical interaction. Hence, we introduce a spatial mask $M_s$ to restrict attention to connected nodes only:

$$M_s[i,j] = \begin{cases} 1 & \text{if node } i \text{ is connected to node } j, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

$$H_t^{\text{space}} = \text{Softmax}\left(\frac{Q_t K_t^T}{\sqrt{d_h}} \odot M_s\right) V. \tag{8}$$

Note that $Q_t K_t^T$ models point-to-point relationships in the spatial dimension $S$, resulting in a shape in the following form: $B \times S \times S$. Finally, we concatenate all time steps:

$$Z^{\text{space}} = \text{Concat}_{t=1}^{T} H_t^{\text{space}} \in \mathbb{R}^{B \times T \times S \times d_k}. \tag{9}$$

*Temporal Attention Branch.* In TAB, we consider the self-attention across all time steps $T$ within the same spatial point $s$. Similar to the spatial attention branch, we partition the data by spatial points. Here, a temporal mask $M_t$ ensures that past information does not affect future predictions.

*Spatiotemporal cross-attention branch.* In STAB, we use spatial attention outputs as queries and temporal attention outputs as keys and values to compute the cross-attention:

$$Q = Z^{\text{space}} W_Q^{\text{cross}}, K = Z^{\text{temp}} W_K^{\text{cross}}, V = Z^{\text{temp}} W_V^{\text{cross}}. \tag{10}$$

For computational efficiency, we reshape the temporal dimension into a single dimension: $Q', K', V' \in \mathbb{R}^{B \times (T \times S) \times d_k}$. To reduce complexity, we apply the spatiotemporal $M_{st}$ during attention computation:

$$M_{st}[(v_i, t_i), (v_j, t_j)] = M_s[v_i, v_j] \cdot M_t[t_i, t_j]. \tag{11}$$

$$Z^{\text{cross}} = \text{Softmax}\left(\frac{Q' K'^T}{\sqrt{d_k}} \odot M_{st}\right) V' \in \mathbb{R}^{B \times (T \times S) \times d_k}. \tag{12}$$

We reshape the result back to its original form as follows: $Z^{\text{cross}} \in \mathbb{R}^{B \times T \times S \times d_k}$.

*Global attention branch.* The GAB flattens all spatiotemporal points to compute the global self-attention, capturing long-range dependencies effectively. After computing all four attention branches, we fuse the features via summation. This ensures that the STDT independently models the spatial dependencies among joints and the temporal evolution of each joint, while capturing complex spatiotemporal interactions. Finally, an FFN produces accurate future human pose sequences.

Our pose generation module is designed not only from the perspective of spatiotemporal decoupling, but it also takes into account both local and global feature modeling. As a shallow network, TFE mainly extracts local features, while the STDT focuses on capturing global features, enabling accurate prediction of future human pose sequences.

### *3.3. Loss Function*

Our loss functions aim to minimize the error between predicted states $s_p$ and target states $s_t$, and this is followed by SINN (Symplectic INtegral Network) [52]. The loss function $L_s$ is calculated as follows:

$$L_s = \frac{1}{N} \sum_{i=1}^{N} \left\| (s_p - s_t) \cdot W_f \cdot W_n \right\|_2, \tag{13}$$

where $W_f$ and $W_n$ are the frame-level and node-level weighting factors, which are introduced to prioritize specific frames and nodes during training.

## 4. Experimental Results

In this section, we comprehensively evaluate our proposed method on two large-scale benchmarks: Human3.6M [22] and AMASS [26].

### *4.1. Datasets*

Human3.6M [22] is the largest benchmark dataset used for evaluating human motion prediction performance. It consists of 15 actions performed by 7 subjects, namely S1, S5, S6, S7, S8, S9, and S11. In this dataset, each human pose is represented by 32 joints to depict the skeletal structure. To ensure fair comparisons with previous works [9,20,21,24,33,53], we follow their protocols by discarding 10 redundant joints (i.e., duplicated joints and global translation points) and retaining 22 joints to represent the human skeleton. Moreover, the frame rate is downsampled from 50 FPS to 25 FPS. Following the testing protocol of Dang et al., Subjects S1, S6, S7, S8, and S9 are used for training; Subject S11 for validation; and Subject S5 for testing.

AMASS [26] is a recently released human motion dataset that aggregates 18 motion capture datasets, including CMU (Carnegie Mellon University motion database), KIT (Karlsruhe Institute of Technology motion dataset), and BMLrub (BML Walking and Running Database from the Max Planck Institute for Intelligent Systems). We selected eight action categories (i.e., playing basketball, basketball signals, directing traffic, jumping, running, playing soccer, walking, and washing windows) to evaluate human motion prediction performance. To ensure fair comparisons with previous works [9,24,53], we removed the global translation joint and redundant joints from the dataset, and we also adopted a 25-joint representation for the human skeleton.

### *4.2. Comparative Setting*

***Metrics.*** We adopted the Mean Per Joint Position Error (MPJPE) as our primary evaluation metric for human motion prediction. MPJPE measures the average Euclidean distance between the predicted and ground-truth joint positions over all $J$ joints and $T$ time steps. A lower MPJPE indicates more accurate predictions. Formally, MPJPE is defined as follows:

$$\text{MPJPE} = \frac{1}{J \times T} \sum_{t=1}^{T} \sum_{j=1}^{J} \left\| \tilde{X}_{t,j} - X_{t,j} \right\|_2, \tag{14}$$

where $\tilde{X}_{t,j} \in \mathbb{R}^3$ denotes the predicted 3D location of joint $j$ at time $t$, and $X_{t,j} \in \mathbb{R}^3$ is its corresponding ground-truth position.

***Baselines.*** We compared our proposed method with six state-of-the-art (SOTA) methods to comprehensively assess its performance. HisRepItself [8] introduced the innovative concept that future human poses often resemble historical poses using historical information to reconstruct future movements. ProGenIntial [43] proposed a two-stage framework that incrementally refines prediction outputs to align them more closely with the ground truth. CGHMP [53] incorporated additional semantic information to assist the model in predicting future human motion. siMLPe [24] demonstrated that a network consisting solely of fully connected layers could achieve SOTA performance. Lastly, AMHGCN [25] employed four levels of hypergraph representation—joint level, part level, component level, and global level—to comprehensively describe the human skeleton and achieve superior prediction accuracy.

***Implementation Details.*** All experiments were conducted on a single NVIDIA GeForce RTX 4060 GPU with 8 GB of memory. The model was implemented using `PyTorch(v2.3.1)` [54], and it was trained on an Intel Core i9-13900HX CPU @ 2.20 GHz with 32 GB RAM. We used the Adam optimizer with a batch size of 16 and an initial learning rate of $1 \times 10^{-3}$, which decayed by a factor of 0.95 every epoch. Training each model on the Human3.6M dataset took approximately 1.5 h.

#### 4.2.1. Results on Human3.6M

We evaluated our method on Human3.6M for both short-term ($\leq$400 ms) and long-term (>400 ms) motion prediction, and we compared it with six SOTA baselines: HisRepItself [8], ProGenIntial [43], CGHMP [53], PPGMP [23], siMLPe [24], and AMHGCN [25]. As shown in Table 1, our method achieved the best performance in 85 out of 90 prediction scenarios, with near-optimal results in the remaining five (e.g., slight under performance in 'Walking' at 80 ms, 'TakingPhoto' at 80 ms, and 'Purchases' at 160 ms). Crucially, our approach yielded consistent improvements across all actions, as shown in Table 2, achieving an average MPJPE of 94.8 mm at 1000 ms—13.4% lower than the strongest baseline (AMHGCN, 109.5 mm)—validating the claim of a 13% average improvement over the SOTA methods mentioned in the abstract.

Figure 3 visualizes the predictions for the 'Eating' sequence, with ground truth in magenta and predictions in blue–green, and it is annotated with MPJPE values at 80 ms, 400 ms, and 1000 ms. Red dashed rectangles highlight the significant errors. While the baselines performed well in the short term, their predictions drifted over time. In contrast, our method maintained high accuracy throughout, demonstrating superior long-term stability—consistent with the quantitative results.

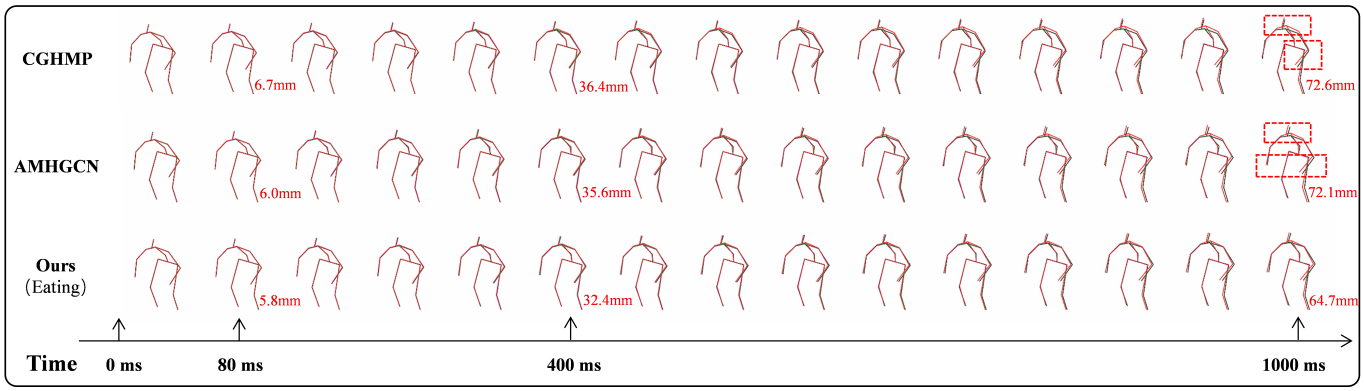



**Figure 3.** The qualitative analysis between our proposed DSC and the baselines when applied to the scenario 'Eating'. From top to bottom, we present the ground truth, as well as the prediction results of CGHMP [53], AMHGCN [25], and our proposed DSC. From left to right, the first pose corresponds to the prediction at 80 ms, the sixth pose is the prediction at 400 ms, and the last pose represents the prediction at 1000 ms. The ground truth is shown in magenta with a solid line, while the comparative results are displayed in blue–green with a solid line. To provide a more intuitive evaluation, we overlaid the corresponding ground truth at each predicted pose location and annotated the MPJPE values at key time points: 80 ms, 400 ms, and 1000 ms. Obvious errors are marked with red dashed rectangles.

To further analyze the performance across motion types, we categorized actions into *periodic* (e.g., *Walking* and *Eating*) and *non-periodic/complex* (e.g., *Discussion*, *SittingDown*, and *TakingPhoto*) following [36]. As shown in Table 3, our model reduces MPJPE by 23.1% (periodic) and 12.6% (non-periodic) at 1000 ms compared to AMHGCN. This highlights its robustness not only in regular motions, but also in complex, irregular behaviors, indicating strong generalization across diverse temporal and structural patterns.

In summary, both quantitative (Table 1) and qualitative (Figure 3) analyses confirm that our method establishes a new SOTA on the H3.6M benchmark.

**Table 1.** Short-term and long-term prediction results on Human3.6M (only selected methods).

| Method | Walking | | | | | | Eating | | | | | | Smoking | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 80 | 160 | 320 | 400 | 560 | 1000 | 80 | 160 | 320 | 400 | 560 | 1000 | 80 | 160 | 320 | 400 | 560 | 1000 |
| HisRepItself [8] | 10.1 | 19.9 | 34.3 | 40.3 | 46.6 | 56.3 | 6.9 | 14.5 | 29.4 | 40.1 | 49.2 | 73.6 | 6.7 | 14.1 | 30.2 | 35.2 | 48.6 | 69.2 |
| CGHMP [53] | 9.9 | 18.9 | 31.8 | 37.0 | 44.9 | 53.7 | 6.7 | 14.4 | 30.1 | 36.4 | 48.4 | 72.6 | 6.5 | 14.2 | 28.7 | 34.8 | 46.2 | 68.6 |
| ProGenIntial [43] | 10.5 | 20.2 | 35.1 | 39.9 | 48.1 | 56.4 | 6.9 | 16.0 | 30.5 | 37.6 | 51.1 | 76.0 | 6.5 | 14.0 | 27.6 | 34.7 | 46.5 | 69.5 |
| PPGMP [23] | 9.6 | 18.5 | 38.5 | 39.8 | 46.6 | 56.7 | 6.5 | 14.3 | 28.6 | 35.5 | 47.4 | 72.2 | 6.4 | 14.1 | 28.5 | 34.0 | 45.5 | 68.0 |
| siMLpe [24] | 9.5 | 18.3 | 33.4 | 37.1 | **43.7** | **51.9** | 6.2 | 13.8 | 29.3 | 35.9 | 48.0 | 73.1 | 6.2 | 13.6 | 28.1 | 34.6 | 45.6 | 68.0 |
| AMHGCN [25] | **9.3** | 18.2 | 31.8 | 36.9 | 44.5 | 52.4 | 6.0 | 13.6 | 28.5 | 35.6 | 47.4 | 72.1 | 6.1 | 13.5 | 27.4 | 34.0 | 44.9 | 66.8 |
| Our(DSC) | 9.5 | **18.0** | **30.4** | **35.9** | 44.2 | 54.1 | **5.8** | **12.8** | **25.9** | **32.4** | **43.5** | **64.7** | **5.5** | **12.9** | **25.1** | **31.7** | **40.9** | **63.8** |
| Method | Discussion | | | | | | Directions | | | | | | Greeting | | | | | |
| | 80 | 160 | 320 | 400 | 560 | 1000 | 80 | 160 | 320 | 400 | 560 | 1000 | 80 | 160 | 320 | 400 | 560 | 1000 |
| HisRepItself [8] | 10.3 | 24.6 | 56.4 | 68.6 | 90.6 | 123.1 | 8.6 | 19.9 | 43.3 | 53.3 | 70.4 | 103.2 | 16.9 | 37.5 | 77.4 | 93.1 | 117.6 | 152.5 |
| CGHMP [53] | 10.1 | 25.3 | 55.3 | 66.3 | 89.4 | 120.6 | 6.9 | 16.9 | 39.8 | 50.0 | 68.6 | 100.7 | 13.7 | 33.4 | 72.1 | 88.4 | 112.4 | 147.9 |
| ProGenIntial [43] | 10.0 | 24.2 | 53.6 | 66.5 | 87.1 | 118.2 | 7.5 | 17.2 | 40.1 | 51.9 | 69.3 | 100.4 | 15.2 | 34.0 | 71.5 | 87.6 | 110.2 | 143.5 |
| PPGMP [23] | 9.5 | 23.1 | 53.5 | 66.8 | 87.2 | 120.0 | 6.6 | 16.5 | 39.5 | 49.8 | 67.5 | 99.3 | 14.1 | 32.3 | 70.1 | 85.5 | 109.6 | 147.0 |
| siMLpe [24] | 9.5 | 23.4 | 54.7 | 67.2 | 87.4 | 119.4 | 6.5 | 16.5 | 39.7 | 50.1 | 68.0 | 100.4 | 13.4 | 32.2 | 70.6 | 86.9 | 109.8 | 146.7 |
| AMHGCN [25] | 9.4 | 23.1 | 53.6 | 66.3 | 86.7 | 118.7 | 6.3 | 16.2 | 39.4 | 49.9 | 67.6 | 100.7 | 13.4 | 23.2 | 70.3 | 85.6 | 109.0 | 146.0 |
| Our(DSC) | **9.0** | **22.7** | **50.8** | **64.5** | **72.7** | **101.4** | **6.0** | **15.4** | **35.4** | **45.9** | **63.5** | **89.7** | **11.8** | **21.8** | **57.9** | **70.1** | **89.1** | **123.3** |
| Method | Phoning | | | | | | Posing | | | | | | Purchases | | | | | |
| | 80 | 160 | 320 | 400 | 560 | 1000 | 80 | 160 | 320 | 400 | 560 | 1000 | 80 | 160 | 320 | 400 | 560 | 1000 |
| HisRepItself [8] | 10.7 | 20.1 | 41.2 | 51.2 | 66.2 | 102.6 | 12.3 | 29.6 | 67.4 | 85.6 | 116.3 | 174.7 | 12.7 | 29.9 | 64.3 | 76.3 | 99.4 | 139.8 |
| CGHMP [53] | 8.5 | 18.2 | 39.7 | 49.8 | 66.2 | 101.8 | 10.6 | 26.3 | 63.3 | 82.3 | 97.4 | 135.8 | 12.4 | 28.9 | 61.8 | 75.0 | 97.4 | 135.8 |
| ProGenIntial [43] | 8.5 | 18.0 | 38.6 | 48.4 | 65.9 | 102.7 | 10.7 | 25.2 | 59.6 | 76.5 | 95.3 | 133.3 | 12.5 | 28.2 | 60.1 | 73.9 | 95.3 | 133.3 |
| PPGMP [23] | 8.4 | 18.1 | 38.5 | 47.8 | 64.4 | 99.3 | 9.9 | 25.0 | 61.5 | 79.2 | 94.4 | 134.0 | 11.6 | 27.5 | 59.5 | 72.8 | 94.4 | 134.0 |
| siMLpe [24] | 8.2 | 18.3 | 39.1 | 48.6 | 65.5 | 101.2 | 9.7 | 24.9 | 61.4 | 78.6 | 93.4 | 134.1 | 11.8 | 27.3 | 58.4 | 71.1 | 93.4 | 134.1 |
| AMHGCN [25] | 8.1 | 18.5 | 39.1 | 48.0 | 64.9 | 99.8 | 9.7 | 24.6 | 60.6 | 77.9 | 93.1 | 134.6 | 11.6 | **27.2** | 58.8 | 70.9 | 93.1 | 134.6 |
| Our(DSC) | **7.5** | **16.9** | **35.1** | **45.7** | **60.4** | **86.7** | **8.3** | **16.6** | **34.8** | **59.5** | **76.6** | **131.5** | **11.5** | 27.4 | **57.4** | **68.9** | **84.2** | **114.8** |
| Method | Sitting | | | | | | Sittingdown | | | | | | Takingphoto | | | | | |
| | 80 | 160 | 320 | 400 | 560 | 1000 | 80 | 160 | 320 | 400 | 560 | 1000 | 80 | 160 | 320 | 400 | 560 | 1000 |
| HisRepItself [8] | 9.9 | 19.5 | 44.4 | 56.1 | 78.0 | 123.7 | 14.7 | 28.1 | 60.2 | 75.2 | 100.6 | 155.2 | 8.7 | 19.6 | 44.4 | 56.6 | 78.9 | 123.2 |
| CGHMP [53] | 9.1 | 19.4 | 43.1 | 55.4 | 76.3 | 119.3 | 14.5 | 28.2 | 58.7 | 74.8 | 98.6 | 152.5 | 8.6 | 19.2 | 43.3 | 55.0 | 76.4 | 122.2 |
| ProGenIntial [43] | 8.9 | 19.0 | 42.5 | 53.6 | 74.4 | 116.1 | 13.9 | 27.9 | 57.6 | 71.7 | 96.7 | 147.8 | 8.4 | 18.9 | 42.6 | 53.5 | 74.3 | 118.6 |
| PPGMP [23] | 8.5 | 19.3 | 42.6 | 54.5 | 74.1 | 115.9 | 13.4 | 27.1 | 56.5 | 70.0 | 94.7 | 146.4 | 8.2 | 18.1 | 41.5 | 52.8 | 72.0 | 115.3 |
| siMLpe [24] | 8.5 | 18.8 | 42.3 | 53.9 | 73.9 | 115.5 | 13.4 | 26.6 | 55.1 | 69.6 | 93.8 | 146.0 | **8.1** | 18.3 | 41.7 | 53.2 | 72.6 | 116.4 |
| AMHGCN [25] | 8.5 | 18.6 | 42.5 | 53.6 | 73.9 | 115.8 | 13.1 | 26.5 | 55.4 | 69.0 | 93.9 | 146.7 | **8.1** | 18.3 | 41.2 | 52.3 | 71.5 | 114.2 |
| Our(DSC) | **8.4** | **17.9** | **41.9** | **51.4** | **61.1** | **97.4** | **12.5** | **25.9** | **53.1** | **65.7** | **83.6** | **119.8** | 8.2 | **17.7** | **38.8** | **49.5** | **66.5** | **105.7** |
| Method | Waiting | | | | | | Walkingdog | | | | | | Walkingtogether | | | | | |
| | 80 | 160 | 320 | 400 | 560 | 1000 | 80 | 160 | 320 | 400 | 560 | 1000 | 80 | 160 | 320 | 400 | 560 | 1000 |
| HisRepItself [8] | 9.1 | 19.9 | 45.3 | 57.3 | 76.0 | 111.0 | 19.9 | 41.5 | 79.4 | 92.1 | 111.8 | 147.5 | 8.7 | 18.1 | 35.2 | 41.9 | 93.8 | 146.0 |
| CGHMP [53] | 8.9 | 20.9 | 44.8 | 55.0 | 73.7 | 107.0 | 18.7 | 40.4 | 76.1 | 89.4 | 108.1 | 142.3 | 8.8 | 18.6 | 33.7 | 39.8 | 83.6 | 119.8 |
| ProGenIntial [43] | 8.9 | 20.2 | 43.1 | 54.9 | 72.2 | 103.4 | 18.9 | 39.0 | 73.5 | 86.6 | 104.7 | 139.8 | 8.7 | 18.0 | 34.6 | 41.7 | 96.7 | 147.8 |
| PPGMP [23] | 8.4 | 19.4 | 43.5 | 54.8 | 73.5 | 106.2 | 18.0 | 38.3 | 74.4 | 87.9 | 107.6 | 139.8 | 8.5 | 17.9 | 33.5 | 39.9 | 94.7 | 146.4 |
| siMLpe [24] | 8.4 | 19.5 | 44.4 | 55.1 | 74.0 | 106.7 | 18.2 | 39.8 | 74.3 | 88.9 | 107.1 | 139.6 | 8.2 | 17.6 | 33.1 | 39.6 | 93.8 | 146.0 |
| AMHGCN [25] | 8.3 | 19.3 | 43.8 | 54.9 | 73.3 | 105.3 | 18.0 | 39.6 | 75.5 | 89.6 | 108.4 | 142.4 | 8.1 | 17.5 | 33.4 | 39.0 | 93.9 | 146.7 |
| Our(DSC) | **7.9** | **16.4** | **35.4** | **48.9** | **67.5** | **92.36** | **16.8** | **36.8** | **69.9** | **84.4** | **99.4** | **122.9** | **7.9** | **15.9** | **31.1** | **35.7** | **42.3** | **53.7** |

MPJPE (mm): lower is better. Best results in each column are in bold. Values are the mean results over eight independent runs with 95% confidence intervals.

**Table 2.** Comparison of the average prediction errors (MPJPE) on Human3.6M.

| **Millisecond (ms)** | **80 ms** | **160 ms** | **320 ms** | **400 ms** | **560 ms** | **1000 ms** |
|---|---|---|---|---|---|---|
| HisRepItself [8] | 11.1 | 23.8 | 50.2 | 61.5 | 80.2 | 114.9 |
| CGHMP [53] | 10.5 | 24.0 | 49.6 | 60.9 | 77.9 | 112.0 |
| ProGenIntial [43] | 11.4 | 27.1 | 55.5 | 68.9 | 76.9 | 110.3 |
| PPGMP [23] | 10.2 | 27.0 | 53.4 | 65.4 | 76.3 | 110.1 |
| siMLPe [24] | 9.9 | 25.8 | 52.6 | 64.2 | 75.9 | 109.8 |
| AMHGCN [25] | 9.6 | 24.7 | 51.2 | 62.6 | 75.5 | 109.5 |
| **Ours** | **9.1** | **21.6** | **40.8** | **54.0** | **66.4** | **94.8** |

MPJPE (mm) results; lower is better. Best results per column are in bold.

**Table 3.** Comparison of the periodic and non-periodic motion prediction results on Human3.6M (where the MPJPE is in mm, and the lower the value, the better).

| Method | Periodic Motions (Avg. over 6 Actions) | | | | | | Complex Motions (Avg. over 9 Actions) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 80 | 160 | 320 | 400 | 560 | 1000 | 80 | 160 | 320 | 400 | 560 | 1000 |
| HisRepItself [8] | 10.50 | 21.37 | 41.62 | 50.13 | 69.37 | 99.20 | 11.47 | 25.40 | 55.90 | 69.12 | 91.98 | 134.04 |
| CGHMP [53] | 9.85 | 20.78 | 40.02 | 47.87 | 66.23 | 93.13 | 10.53 | 24.28 | 53.58 | 66.91 | 87.80 | 126.87 |
| ProGenIntial [43] | 10.00 | 20.87 | 39.98 | 48.15 | 68.83 | 98.70 | 10.67 | 23.87 | 52.30 | 65.57 | 86.09 | 123.84 |
| PPGMP [23] | 9.57 | 20.20 | 40.33 | 47.48 | 67.70 | 97.07 | 10.02 | 23.14 | 52.02 | 65.13 | 85.27 | 124.23 |
| siMLpe [24] | 9.42 | 20.23 | 39.55 | 47.45 | 67.28 | 96.63 | 9.92 | 23.06 | 52.03 | 65.08 | 85.14 | 124.37 |
| AMHGCN [25] | 9.27 | 20.15 | 39.28 | 47.18 | 67.33 | 96.70 | 9.82 | 21.89 | 51.73 | 64.49 | 84.68 | 124.07 |
| Our (DSC) | **8.83** | **18.88** | **36.25** | **44.30** | **55.12** | **74.32** | **9.29** | **20.20** | **45.06** | **58.27** | **73.87** | **108.44** |

MPJPE (mm): lower is better. Periodic motions: average over 6 actions (Walking, Eating, Smoking, Phoning, WalkingDog, and WalkingTogether). Non-periodic/complex motions: average over 9 actions (Discussion, Directions, Greeting, Sitting, SittingDown, TakingPhoto, Waiting, Posing, and Purchases) [36].

#### 4.2.2. Results on AMASS

To further assess the generality and accuracy of our method, we evaluated its performance on the AMASS dataset. Table 4 reports the prediction errors for both short-term ($\leq$400 ms) and long-term (>400 ms) horizons. Our proposed method consistently outperforms all baselines across both prediction lengths, with the performance gap widening in long-term forecasts. These results confirm that our network effectively disentangles the spatial and temporal motion features from the input historical sequence, enabling the precise prediction of future human poses.

**Table 4.** Comparison of the average prediction errors (MPJPE) on AMASS.

| Millisecond (ms) | 80 ms | 160 ms | 320 ms | 400 ms | 560 ms | 1000 ms |
|---|---|---|---|---|---|---|
| HisRepItself [8] | 6.9 | 13.3 | 27.7 | 36.2 | 52.0 | 88.1 |
| CGHMP [53] | 6.8 | 13.0 | 28.9 | 37.7 | 54.8 | 92.2 |
| ProGenIntial [43] | 7.8 | 14.5 | 28.2 | 36.9 | 50.4 | 80.9 |
| PPGMP [23] | 6.1 | 11.7 | 25.7 | 33.8 | 50.3 | 87.2 |
| siMLpe [24] | 6.0 | 11.5 | 26.0 | 34.2 | 50.2 | 88.1 |
| AMHGCN [25] | 5.9 | 11.7 | 25.8 | 33.9 | 50.4 | 87.8 |
| our(DSC) | **5.5** | **10.5** | **24.2** | **30.3** | **45.6** | **76.9** |

MPJPE (mm): lower is better. Best results in each column are in bold. Values are the mean results over eight independent runs with 95% confidence intervals.

#### 4.2.3. Computational Efficiency

We evaluated the computational efficiency of our method to assess its suitability for real-time applications. We compared our approach with key baselines: CGHMP [53], siMLPe [24], and AMHGCN [25]. All experiments were conducted on a single NVIDIA GeForce RTX 4060 GPU (8 GB VRAM) using a 25-frame input sequence (625 ms at 40 Hz). Results were averaged over 1000 forward passes on the Human3.6M test set.

Table 5 reports the model parameters (in millions), FLOPs (in G), inference latency (in ms), GPU memory usage (in MB), and approximate training time per run. Our method achieved a latency of **13.9 ms**, or about **72 FPS**, meeting the requirements for real-time use in AR/VR, robotics, and animation. This is faster than AMHGCN (13.4 ms) and only slightly slower than siMLPe (9.1 ms), which has lower computational demands but weaker accuracy (see Section 4.2.1).

Our model has **2 *M* parameters**, making it more compact than CGHMP (4.9 M) and reasonably lightweight overall. While our FLOPs (0.7 G) are higher than siMLPe (0.1 G) and AMHGCN (0.2 G), this reflects the cost of explicit temporal modeling through disentangled spatiotemporal pathways. Despite this, our GPU memory footprint (1110 MB) is moderate and lower than CGHMP (1320 MB), indicating efficient memory utilization. Training converges in approximately 1.5 h, faster than CGHMP (1.8 h) and comparable to siMLPe

(1.2 h) and AMHGCN (1.3 h). The efficient training is attributed to stable gradient flow and reduced feature entanglement due to the decoupled architecture design.

In summary, our method achieves a strong balance between accuracy and efficiency. It offers competitive inference speed, moderate resource usage, and fast training, while maintaining a compact model size. These characteristics make it practical for both research and real-world deployment.

**Table 5.** Computational efficiency comparison on Human3.6M. Lower values are better for all metrics except FPS (where higher is better).

| Method | Params (M) | FLOPs (G) | Inference Latency (ms) | FPS | GPU Mem (MB) | Training Time |
|---|---|---|---|---|---|---|
| CGHMP [53] | 4.9 | 0.3 | 15.2 | 65 | 1320 | ∼1.8 h |
| siMLPe [24] | 0.8 | 0.1 | 9.1 | 110 | 540 | ∼1.1 h |
| AMHGCN [25] | 1.4 | 0.2 | 13.4 | 75 | 843 | ∼1.3 h |
| Our (DSC) | 2.0 | 0.7 | 13.9 | 72 | 1110 | ∼1.5 h |

All metrics measured on the same hardware: RTX 4060 GPU and i9-13900HX CPU. Input: 25-frame sequence. FPS computed as 1000/Latency (ms).

### *4.3. Decoupling Spatiotemporal Experiment*

To empirically evaluate the effectiveness of explicit spatiotemporal decoupling, we designed four configurations that selectively remove or weaken the decoupling mechanism: (1) **No-Decoupling Encoder**, which replaces the disentangled encoder with a standard joint spatiotemporal model, forcing spatial and temporal features to be learned jointly; (2) **Multi-Head Generator**, which employs multi-head self-attention in the generation module—instead of the proposed single-head temporal smoother—and may introduce redundant cross-joint interactions; (3) **Fully Coupled Baseline**, which removes spatiotemporal separation in both the encoder and generator, representing a conventional end-to-end architecture; and (4) **Decoupled Full Model**, which retains both the disentangled encoder and the single-head attention generator, corresponding to the complete proposed framework. This comparative study enables us to assess how varying levels of decoupling influence the overall prediction performance.

As shown in Table 6, all models without full decoupling suffer performance degradation, especially in long-term predictions. The fully coupled baseline performed the worst, with the average MPJPE increasing from 47.8 mm to 69.7 mm, which represents an increase of over 45%. This sharp decline suggests that entangled modeling obscures the critical kinematic patterns under complex motion dynamics. Even partial removal of decoupling, such as using multi-head attention in the generation module, leads to noticeable error growth at longer horizons. For example, the prediction error at 1000 ms rises significantly compared to the full model, indicating that excessive interaction modeling harms temporal consistency. In contrast, the fully decoupled model exhibits the most stable error progression, demonstrating superior generalization and reduced error accumulation over time.

**Table 6.** Comparison of spatiotemporal decoupling in the pose encoding and generation modules.

| Configuration | 80 ms | 160 ms | 320 ms | 400 ms | 560 ms | 1000 ms | Average |
|---|---|---|---|---|---|---|---|
| Pose Encoding–No Decoupling | 9.8 | 28.4 | 57.8 | 67.7 | 79.4 | 110.2 | 58.9 |
| Pose Generation–Multihead Attention | 9.7 | 25.6 | 52.3 | 64.0 | 75.3 | 108.8 | 56.0 |
| No Decoupling in Both Modules | 13.2 | 29.3 | 57.2 | 80.9 | 103.4 | 134.1 | 69.7 |
| Full Spatiotemporal Decoupling | **9.1** | **21.6** | **40.8** | **54.0** | **66.4** | **94.8** | **47.8** |

MPJPE (mm): lower is better. Best results in each column are in bold. Values are the mean results over eight independent runs with 95% confidence intervals.

These results demonstrate that explicit spatiotemporal decoupling is not merely beneficial but essential for accurate and stable motion prediction. By preventing feature entan-

glement and allowing specialized modeling of spatial topology and temporal dynamics, our design enables more reliable long-term forecasting—particularly under autoregressive settings where error accumulation is a major challenge.

### *4.4. Ablation Studies*

In this section, we present a systematic ablation study that was conducted on the two core modules of our proposed method: the *Spatiotemporal Decoupling Module* (SDM), and the *Pose Generation Module* (PGM). The **SDM** performs disentangled encoding of human poses by separately modeling spatial and temporal feature relationships. The **PGM** accurately predicts future pose sequences by independently modeling temporal and spatial dependencies. The spatiotemporal decoupling module consists of two components: the *graph attention component* (GAC), and the *spatiotemporal decoupling encoding component* (SDEC). The PGM also comprises two key components: the *temporal feature extraction* (TFE) component, and the *spatiotemporally decoupled transformer* (STDT).

To evaluate the contribution of each key component to the overall performance, we designed four plug-and-play ablation settings: (1) removing the GAC; (2) replacing the TFE module with a standard TCN; (3) substituting the STDT with a conventional Transformer architecture; and (4) replacing the SDEC encoder with a standard GCN. These ablation studies help us better understand the role and effectiveness of each individual component in the proposed framework.

The results in Table 7 show that all ablation settings lead to performance degradation, confirming that each component contributes meaningfully to the final accuracy. Notably, replacing SDEC with a standard GCN results in the largest performance drop, with the average MPJPE increasing from 47.8 mm to 61.5 mm. This highlights the limitations of joint spatiotemporal modeling, which tends to introduce feature entanglement and impair prediction accuracy. Similarly, substituting STDT with a standard Transformer leads to a substantial increase in long-term prediction errors, particularly at the 1000 ms horizon. This suggests that entangled attention mechanisms are prone to error accumulation over time. Removing the GAC component results in an average MPJPE of 59.5 mm, with noticeable degradation in mid-term predictions. This indicates that the GAC not only enhances spatial feature learning through adaptive attention, but that it also helps preserve biomechanical plausibility by reinforcing anatomically meaningful joint dependencies and suppressing physically implausible connections during message passing.

The full model achieves the lowest MPJPE across all horizons, validating the effectiveness of our decoupled design. These results collectively demonstrate that explicit separation of spatial and temporal modeling is not only beneficial, but also essential for accurate and stable long-term human motion prediction.

**Table 7.** Ablation study on the components of human motion prediction via the decoupled spatiotemporal clue network.

| Component Replaced/Removed | 80 ms | 160 ms | 320 ms | 400 ms | 560 ms | 1000 ms | Average |
|---|---|---|---|---|---|---|---|
| Replace TFE with a TCN | 9.7 | 23.8 | 49.4 | 58.8 | 75.0 | 112.8 | 54.9 |
| Replace the STDT with a standard Transformer | 11.6 | 26.5 | 52.1 | 61.9 | 76.8 | 110.5 | 56.6 |
| Remove the GAC component. | 9.6 | 25.1 | 60.9 | 70.0 | 82.9 | 108.3 | 59.5 |
| Replace the SDEC encoder with a GCN | 13.1 | 29.0 | 57.3 | 66.5 | 81.3 | 121.7 | 61.5 |
| Full Model | **9.1** | **21.6** | **40.8** | **54.0** | **66.4** | **94.8** | **47.8** |

MPJPE (mm): lower is better. Best results in each column are in bold. Values are the mean results over eight independent runs with 95% confidence intervals.

## 5. Discussion

Our method achieves SOTA performance in 3D human motion prediction on the Human3.6M and AMASS benchmarks. The proposed spatiotemporal decoupling enables

independent yet complementary modeling of joint kinematics (spatial topology) and motion dynamics (temporal evolution), validating our hypothesis that disentangled representation learning mitigates error propagation and improves long-term stability. Compared to recent SOTA methods, such as siMLPe and AMHGCN, our model shows consistent reductions in MPJPE, especially at longer horizons (e.g., 600–1000 ms), indicating effective suppression of error accumulation in autoregressive generation. The explicit disentanglement of spatial and temporal factors not only boosts prediction accuracy, but also enhances model interpretability—enabling the diagnosis of spatial versus temporal error sources—and it supports modular designs for downstream applications.

Despite these advances, several limitations remain. First, the model relies on a fixed skeletal graph learned from a canonical skeleton, limiting generalization across subjects with varying body proportions (e.g., limb lengths). This constraint may hinder deployment in personalized domains, such as clinical gait analysis or prosthetic control, where subject-specific biomechanics are critical. Second, while the decoupled architecture improves accuracy, it increases model complexity and parameter count compared to unified end-to-end designs, potentially affecting inference speed and memory footprint—challenging real-time use in mobile robotics or AR/VR. Moreover, training on studio-collected datasets (e.g., Human3.6M) under controlled conditions may compromise robustness to real-world variations, such as clothing, occlusion, or outdoor motion patterns.

Future work will address these challenges through several principled directions. (1) **Adaptive graph learning**: Exploring dynamic graph neural networks (e.g., EvolveGCN) or attention-based adjacency adaptation to model subject-specific skeletons; (2) **Physical plausibility**: Integrating biomechanical constraints (e.g., joint limits and self-collision avoidance) and dynamic smoothness priors (e.g., jerk minimization) into the loss or inference pipeline; (3) **Controllable generation**: Extending the framework to support multimodal prediction (via stochastic latent variables) and interactive synthesis (conditioning on goals or environmental cues); (4) **Cross-dataset generalization**: Investigating domain-invariant representations and unsupervised alignment techniques (e.g., via contrastive learning or adversarial adaptation) to enable transfer across datasets with different capture protocols, sensor types, or annotation styles; (5) **Few-shot and zero-shot adaptation**: Designing meta-learning or prompt-based mechanisms to rapidly adapt the model to new subjects or actions with minimal labeled data, leveraging the disentangled structure for efficient parameter reuse. (6) **Physics-aware interactive generation via reinforcement learning**: Integrate physics engines (e.g., MuJoCo and Bullet) with reinforcement learning to enable physically plausible and goal-driven motion synthesis. By training RL agents within a simulated environment governed by physical laws (e.g., gravity and contact dynamics), our model can generate motions that are not only kinematically accurate, but also dynamically realistic and responsive to user commands or environmental constraints (e.g., obstacle avoidance and target reaching).

These extensions aim to build more realistic, personalized, and adaptive motion prediction systems—advancing toward a systems science perspective that emphasizes modularity, transferability, and robust co-design across diverse environments and users. Such capabilities are essential for next-generation applications in assistive robotics, digital avatars, and human-AI collaboration.

## 6. Conclusions

In this paper, we present a novel framework for 3D human motion prediction based on explicit spatiotemporal disentanglement. The method consists of two core modules: (1) a **spatiotemporal decoupling module** that separately models spatial kinematics (bone lengths, joint connectivity, etc.) and temporal dynamics (motion trajectories), and (2) a

**pose generation module** that synthesizes future poses by integrating decoupled spatial and temporal features. This design enables more stable long-term prediction by reducing error propagation across entangled dimensions.

Our work presents a principled framework for disentangled spatiotemporal modeling, offering new insights into the design of interpretable and modular architectures for human motion prediction. By decoupling spatial and temporal factors, we achieve more accurate, stable, and reproducible long-term forecasting. This approach lays the foundation for future systems that are not only predictive, but also physically plausible, personalized, and controllable key requirements for real-world deployment in robotics, virtual avatars, and human-centered AI.

**Author Contributions:** Conceptualization, M.X. and E.L.; methodology, M.X.; software, M.X.; validation, M.X.; formal analysis, M.X.; investigation, M.X.; data curation, M.X.; writing—original draft preparation, M.X.; writing—review and editing, Z.G.; visualization, M.X.; supervision, E.L.; project administration, E.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are available in two publicly accessible datasets: the Human3.6M dataset (http://vision.imar.ro/human3.6m/description.php, accessed on 24 November 2024), and the AMASS dataset (https://amass.is.tue.mpg.de/, accessed on 15 March 2025)). These datasets were used for training and evaluating the proposed 3D human motion prediction model. No new data were generated in this study.

**Acknowledgments:** We wish to thank the teams behind the Human3.6M and AMASS datasets for their publicly available motion data. Computational resources were provided by the High-Performance Computing Platform at Tsinghua University.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Lyu, K.; Chen, H.; Liu, Z.; Zhang, B.; Wang, R. 3D Human Motion Prediction: A Survey. *Neurocomputing* **2022**, *489*, 345–365. [CrossRef]
2. Troje, N.F. Decomposing Biological Motion: A Framework for Analysis and Synthesis of Human Gait Patterns. *J. Vis.* **2002**, *2*, 2. [CrossRef] [PubMed]
3. Liu, X.; Yin, J.; Liu, 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*, 2133–2146. [CrossRef]
4. Koppula, H.S.; Saxena, A. Anticipating Human Activities for Reactive Robotic Response. In Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Tokyo, Japan, 3–7 November 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 4476–4483.
5. Unhelkar, V.V.; Lasota, P.A.; Tyroller, Q.; Buhai, R.-D.; Marceau, L.; Deml, B.; Shah, J.A. Human-Aware Robotic Assistant for Collaborative Assembly: Integrating Human Motion Prediction with Planning in Time. *IEEE Robot. Autom. Lett.* **2018**, *3*, 2394–2401. [CrossRef]
6. Cao, Z.; Gao, H.; Mangalam, K.; Cai, Q.-Z.; Vo, M.; Malik, J. Long-Term Human Motion Prediction with Scene Context. In *Computer Vision—ECCV 2020, Proceedings of the 16th European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020*; Proceedings, Part I; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Springer: Cham, Switzerland, 2020; pp. 387–404.
7. Sofianos, T.; Sampieri, A.; Franco, L.; Galasso, F. Space-Time-Separable Graph Convolutional Network for Pose Forecasting. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 11209–11218.
8. Mao, W.; Liu, M.; Salzmann, M. History Repeats Itself: Human Motion Prediction via Motion Attention. In *Computer Vision—ECCV 2020, Proceedings of the 16th European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020*; Proceedings, Part XIV; Bartoli, A., Fusiello, A., Eds.; Springer: Cham, Switzerland, 2020; pp. 474–489.

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 (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 9489–9497.
10. Wang, B.; Adeli, E.; Chiu, H.-K.; Huang, D.-A.; Niebles, J.C. Imitation Learning for Human Pose Prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 7124–7133.
11. Chiu, H.-K.; Adeli, E.; Wang, B.; Huang, D.-A.; Niebles, J.C. Action-Agnostic Human Pose Forecasting. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 7–11 January 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1423–1432.
12. Adeli, V.; Adeli, E.; Reid, I.; Niebles, J.C.; Rezatofighi, H. Socially and Contextually Aware Human Motion and Pose Forecasting. *IEEE Robot. Autom. Lett.* **2020**, *5*, 6033–6040. [CrossRef]
13. Yuan, Y.; Kitani, K.M. DLow: Diversifying Latent Flows for Diverse Human Motion Prediction. In *Computer Vision—ECCV 2020, Proceedings of the 16th European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020*; Proceedings, Part IX; Bolec, G., Corso, G., Fusiello, A., Eds.; Springer: Cham, Switzerland, 2020; pp. 346–364.
14. Yuan, Y.; Kitani, K.M. Ego-Pose Estimation and Forecasting as Real-Time PD Control. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 10082–10092.
15. Aksan, E.; Kaufmann, M.; Cao, P.; Hilliges, O. A Spatio-Temporal Transformer for 3D Human Motion Prediction. In Proceedings of the 2021 International Conference on 3D Vision (3DV), Montréal, QC, Canada, 1–3 November 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 565–574.
16. Xu, C.; Tan, R.T.; Tan, Y.; Chen, S.; Wang, X.; Wang, Y. Auxiliary Tasks Benefit 3D Skeleton-Based Human Motion Prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 9509–9520.
17. Cao, W.; Li, S.; Zhong, J. A Dual Attention Model Based on Probabilistically Mask for 3D Human Motion Prediction. *Neurocomputing* **2022**, *493*, 106–118. [CrossRef]
18. von Marcard, T.; Henschel, R.; Black, M.J.; Rosenhahn, B.; Pons-Moll, G. Recovering Accurate 3D Human Pose in the Wild Using IMUs and a Moving Camera. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Springer: Cham, Switzerland, 2018; pp. 601–617.
19. Ding, P.; Yin, J. Towards More Realistic Human Motion Prediction with Attention to Motion Coordination. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 5846–5858. [CrossRef]
20. Li, M.; Chen, S.; Zhao, Y.; Zhang, Y.; Wang, Y.; Tian, Q. Dynamic Multiscale Graph Neural Networks for 3D Skeleton Based Human Motion Prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 214–223.
21. 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 (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 11467–11476.
22. 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*, 1325–1339. [CrossRef] [PubMed]
23. Li, J.; Wang, J.; Kuang, C.; Wu, L.; Wang, X.; Xu, Y. A Human-Like Action Learning Process: Progressive Pose Generation for Motion Prediction. *Knowl.-Based Syst.* **2023**, *280*, 110948. [CrossRef]
24. Guo, W.; Du, Y.; Shen, X.; Lepetit, V.; Alameda-Pineda, X.; Moreno-Noguer, F. Back to MLP: A Simple Baseline for Human Motion Prediction. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 4–8 January 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 4809–4819.
25. Li, J.; Wang, J.; Wu, L.; Wang, X.; Luo, X.; Xu, Y. AMHGCN: Adaptive Multi-Level Hypergraph Convolution Network for Human Motion Prediction. *Neural Netw.* **2024**, *172*, 110948. [CrossRef] [PubMed]
26. Mahmood, N.; Ghorbani, N.; Troje, N.F.; Pons-Moll, G.; Black, M.J. AMASS: Archive of Motion Capture as Surface Shapes. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 5442–5451.
27. Wang, J.; Hertzmann, A.; Fleet, D.J. Gaussian Process Dynamical Models. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2005), Vancouver, BC, Canada, 5–8 December 2005; MIT Press: Cambridge, MA, USA, 2005; Volume 18.
28. Taylor, G.W.; Hinton, G.E.; Roweis, S. Modeling Human Motion Using Binary Latent Variables. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2006), Vancouver, BC, Canada, 4–7 December 2006; MIT Press: Cambridge, MA, USA, 2006; Volume 19.

29. Brand, M.; Hertzmann, A. Style Machines. In Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 2000), New Orleans, LA, USA, 23–28 July 2000; ACM: New York, NY, USA; Addison-Wesley: Boston, MA, USA, 2000; pp. 183–192.
30. Lehrmann, A.M.; Gehler, P.V.; Nowozin, S. Efficient Nonlinear Markov Models for Human Motion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 1314–1321.
31. Pavlovic, V.; Rehg, J.M.; MacCormick, J. Learning Switching Linear Models of Human Motion. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2000), Denver, CO, USA, 30 November–3 December 2000; MIT Press: Cambridge, MA, USA, 2000; Volume 13.
32. Fragkiadaki, K.; Levine, S.; Felsen, P.; Malik, J. Recurrent Network Models for Human Dynamics. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 4346–4354.
33. Martinez, J.; Black, M.J.; Romero, J. On Human Motion Prediction Using Recurrent Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 2891–2900.
34. Cui, Q.; Sun, H.; Kong, Y.; Zhang, X.; Li, Y. Efficient Human Motion Prediction Using Temporal Convolutional Generative Adversarial Network. *Inf. Sci.* **2021**, *545*, 427–447. [CrossRef]
35. Gui, L.-Y.; Wang, Y.-X.; Liang, X.; Moura, J.M.F. Adversarial Geometry-Aware Human Motion Prediction. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Springer: Cham, Switzerland, 2018; pp. 786–803.
36. Li, C.; Zhang, Z.; Lee, W.S.; Lee, G.H. Convolutional Sequence to Sequence Model for Human Dynamics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 5226–5234.
37. Butepage, J.; Black, M.J.; Kragic, D.; Kjellström, H. Deep Representation Learning for Human Motion Prediction and Classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 6158–6166.
38. Yang, H.; Yuan, C.; Zhang, L.; Sun, Y.; Hu, W.; Maybank, S.J. STA-CNN: Convolutional Spatial-Temporal Attention Learning for Action Recognition. *IEEE Trans. Image Process.* **2020**, *29*, 5783–5793. [CrossRef] [PubMed]
39. Lea, C.; Flynn, M.D.; Vidal, R.; Reiter, A.; Hager, G.D. Temporal Convolutional Networks for Action Segmentation and Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 156–165.
40. Bai, S.; Kolter, J.Z.; Koltun, V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv* **2018**, arXiv:1803.01271. [CrossRef]
41. Huang, J.; Kang, H. 3D Skeleton-Based Human Motion Prediction Using Spatial–Temporal Graph Convolutional Network. *Int. J. Multimed. Inf. Retr.* **2024**, *13*, 33. [CrossRef]
42. 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 (CVPR), Virtual, 13–19 June 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 6519–6527.
43. Ma, T.; Nie, Y.; Long, C.; Zhang, Q.; Li, G. Progressively Generating Better Initial Guesses towards Next Stages for High-Quality Human Motion Prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 6437–6446.
44. Paden, B.; Čáp, M.; Yong, S.Z.; Yershov, D.; Frazzoli, E. A Survey of Motion Planning and Control Techniques for Self-Driving Urban Vehicles. *IEEE Trans. Intell. Veh.* **2016**, *1*, 33–55. [CrossRef]
45. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv* **2014**, arXiv:1406.1078. [CrossRef]
46. Yan, S.; Xiong, Y.; Lin, D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-18), New Orleans, LA, USA, 2–7 February 2018; AAAI Press: Palo Alto, CA, USA, 2018; Volume 32.
47. Zhao, R.; Su, H.; Ji, Q. Bayesian Adversarial Human Motion Synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 13–19 June 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 6225–6234.
48. Cui, Q.; Sun, H. Towards Accurate 3D Human Motion Prediction from Incomplete Observations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 10–17 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 4801–4810.

49. Martínez-González, A.; Villamizar, M.; Odobez, J.-M. Pose Transformers (PoTR): Human Motion Prediction with Non-Autoregressive Transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 2276–2284.
50. Hua, Y.; Xuanzhe, F.; Yaqing, H.; Liu, Y.; Cai, K.; Dongsheng, Z.; Qiang, Z. Towards Efficient 3D Human Motion Prediction Using Deformable Transformer-Based Adversarial Network. In Proceedings of the 2022 International Conference on Robotics and Automation (ICRA), Philadelphia, PA, USA, 30 May–5 June 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 861–867.
51. Cai, Y.; Huang, L.; Wang, Y.; Cham, T.-J.; Cai, J.; Yuan, J.; Liu, J.; Yang, X.; Zhu, Y.; Shen, X.; et al. Learning Progressive Joint Propagation for Human Motion Prediction. In *Computer Vision—ECCV 2020, Proceedings of the 16th European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020*; Proceedings, Part VII; Springer: Cham, Switzerland, 2020; pp. 226–242.
52. Chen, H.; Lyu, K.; Liu, Z.; Yin, Y.; Yang, X.; Lyu, Y. Rethinking Human Motion Prediction with Symplectic Integral. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 2134–2143.
53. Li, J.; Pan, H.; Wu, L.; Huang, C.; Luo, X.; Xu, Y. Class-Guided Human Motion Prediction via Multi-Spatial-Temporal Supervision. *Neural Comput. Appl.* **2023**, *35*, 9463–9479. [CrossRef]
54. Paszke, A.; Gross, S.; Massa, F.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Proceedings of the Advances in Neural Information Processing Systems 32 (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; pp. 8026–8037.


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.