electronics

MDPI

Article

GGTr: An Innovative Framework for Accurate and Realistic Human Motion Prediction

Biaozhang Huang [1,2] and Xinde Li [1,2,*]

[1] Key Laboratory Measurement and Control of CSE, Ministry of Education, School of Automation, Southeast University, Nanjing 210096, China; bzhuang@seu.edu.cn
[2] Nanjing Center for Applied Mathematics, Nanjing 211135, China
* Correspondence: xindeli@seu.edu.cn

Abstract: Human motion prediction involves forecasting future movements based on past observations, which is a complex task due to the inherent spatial-temporal dynamics of human motion. In this paper, we introduced a novel framework, GGTr, which adeptly encapsulates these patterns by integrating positional graph convolutional network (GCN) layers, gated recurrent unit (GRU) network layers, and transformer layers. The proposed model utilizes an enhanced GCN layer equipped with a positional representation to aggregate information from body joints more effectively. To address temporal dependencies, we strategically combined GRU and transformer layers, enabling the model to capture both local and global temporal dependencies across body joints. Through extensive experiments conducted on Human3.6M and CMU-MoCap datasets, we demonstrated the superior performance of our proposed model. Notably, our framework shows significant improvements in predicting long-term movements, outperforming state-of-the-art methods substantially.

Keywords: human motion prediction; graph convolutional network; gated recurrent unit; transformers

Citation: Huang, B.; Li, X. GGTr: An Innovative Framework for Accurate and Realistic Human Motion Prediction. *Electronics* **2023**, *12*, 3305. https://doi.org/10.3390/electronics12153305

Academic Editor: Grzegorz Dudek

Received: 6 July 2023
Revised: 29 July 2023
Accepted: 30 July 2023
Published: 1 August 2023

1. Introduction

Human motion prediction, the task of forecasting future human movements based on past observations, plays a critical role in various domains such as robotics, computer vision, healthcare, and sports analysis. Accurately predicting human motion is instrumental for facilitating effective human-robot collaboration [1,2], ensuring system security [3,4], analyzing human behavior and emotions [5,6], and supporting sports performance analysis [7]. However, predicting human motion presents significant challenges due to the complexity and diversity of human behaviors. First, human motion exhibits high variability and uncertainty. This is evident at the 3D skeletal level, due to the diversity in human body sizes, and at the movement level, due to individual idiosyncrasies. In scenarios with rapid changes, such as sudden reactions or motions, it is exceedingly difficult for predictive models to adapt quickly. Second, the interplay between different body parts and their coordinated movements further complicates the task. For instance, a motion initiated by one body part can propagate to other body parts in complex and often non-intuitive ways.

So, it becomes essential to capture both spatial and temporal features, as illustrated in Figure 1. Numerous research efforts have been made to address the challenges of modeling human motion. Traditional methods and machine learning methods such as hidden Markov models (HMM) [8], Gaussian processes (GP) [9], and restricted Boltzmann machine [10]. However, these methods may not fully capture the complex interdependencies and non-linear dynamics present in human motion. More recently, deep learning approaches, such as convolutional neural networks (CNNs) [11], graph convolutional networks (GCN) [12–15], temporal modules such as recurrent neural networks (RNNs) [16–21], and transformers [22–24] have been used. While existing RNN and deep learning-based models have significantly improved the prediction performance, they still have limitations. These methods often struggle to capture the dynamic and complex interactions

between different body parts. The relations among body joints cannot be simply modeled by static spatial proximity; instead, they are influenced by various factors such as the individual's physical attributes, the current body state, and the motion context. Moreover, these models frequently encounter challenges in effectively capturing both local and global temporal dependencies.

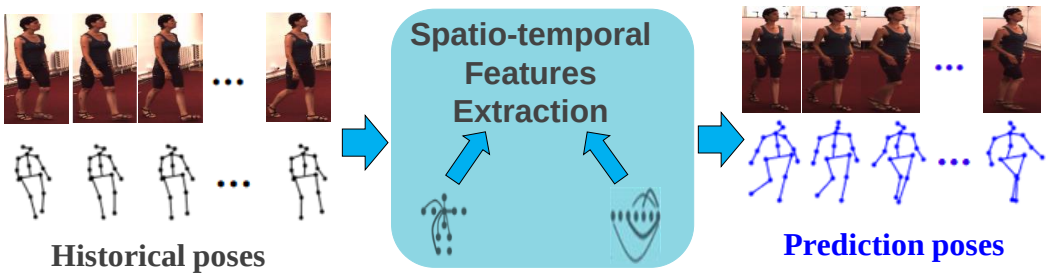

Figure 1. The process of human motion prediction. The historical sequences of human motion are subjected to feature extraction in terms of both space and time. This involves the analysis of the spatial configuration of the body, as well as the timing and sequence of these movements. The derived spatial and temporal features constitute the foundation for building a comprehensive understanding of the motion patterns.

To address these challenges, we proposed a novel approach for human motion prediction that combines positional graph convolutional network (GCN) layers, gated recurrent unit (GRU) network layers, and transformer layers, termed as GGTr. This combination allows us to better capture the complex spatial-temporal patterns in human motion data. In summary, the primary contributions of this paper include:

(1) The introduction of a novel GCN layer with positional representation, enabling better aggregation of information from adjacent body joints;
(2) The strategic combination of GRU and transformer layers to capture both local and global temporal dependencies across body joints;
(3) The conduction of extensive experiments on the Human3.6M and CMU-MoCap datasets, demonstrating the effectiveness and advantages of our proposed framework, our model shows significant improvements in predicting short-term movements. Experiments reveal that our model significantly outperforms state-of-the-art methods .

The remainder of this paper is organized as follows: Section 2 reviews related work. In Section 3, we first introduce some preliminary concepts and formalize the human motion prediction problem and the overall framework of the spatial temporal network model. Then, we discuss in detail the three model components to deal with the spatial and temporal dependencies, separately. In Section 4, experiments were conducted on two large-scale datasets, comparing the performance of the proposed method with baselines. Section 5 provides a summary and conclusion, as well as a discussion of future work.

2. Related Work

Human motion prediction is a challenging task due to the complexity and variability of human movements. This complexity arises from the intricate interplay of various factors such as the individual's physical characteristics, the environment, and the task at hand. Over the years, several approaches have been proposed to tackle this problem, each with its own strengths and limitations.

Traditional methods primarily rely on data statistical approaches or prior knowledge, such as Markov prediction models [8], Gaussian process dynamical models [9] and restricted Boltzmann machine [10], which can only tackle simple human motion patterns. Although these methods have achieved some success in certain scenarios, they still face challenges in capturing complex spatial and temporal dependencies.

Chiu et al. [25] used LSTM units to model the underlying structure of human motion hierarchically, but did not adequately utilize spatial information of human motion data. Martinez et al. [18] introduced a residual structure using GRU to model the velocity of human motion sequences, Their model focused on short-term temporal modeling, ignoring long-term dependencies and spatial structure. Jain et al. [17] combined LSTM and fully connected (FC) layers in a structural RNN model to encode high-level spatio-temporal structures in human motion sequences. Guo et al. [26] employed FC layers and GRU to model local structures and capture long-term temporal dependencies, but they did not take into account the interactions between different human body limbs. The transformer model with a self-attention mechanism has been proved to be more effective than recurrent networks in various domains [23,27,28]. It applies a multi-head attention mechanism to directly learn dependencies between each pair of input and output positions without any latency.

Graph convolutional networks (GCN) have become widely used for modeling the underlying relationships of non-Euclidean data. Kipf et al. [29] proposed a layer-wise propagation rule for nodes inspired by first-order approximation of spectral convolutions on graphs. However, this approach is limited by the characteristics of the graph. Yuan et al. [30] proposed a more flexible approach by learning node connectivity based on node neighborhood. Velickovic et al. [31] introduced self-attention to determine the neighborhood structure to be considered, providing more flexibility to the network. This approach has been applied to action recognition [32] by using a GCN to capture the temporal and spatial dependencies of human body joints via a graph defined on temporally connected kinematic trees. These techniques have been applied to human motion prediction by building the human pose as a graph and using GCN to encode the spatial connectivity of human joints. Ma et al. [33] proposed two variants of GCN to extract spatial and temporal features. They built a multi-stage structure where each stage contains an encoder and a decoder and, during training, the model is trained with intermediate supervision to learn to progressively refine the prediction. References [12,13,34] extended the graph of the human pose to multi-scale the version across the abstraction levels of the human pose. However, as these models aggregate body joint features based on an input adjacency matrix, the relation between body parts is fixed and may limit the model's ability to adapt to complex human motions.

The transformer [35] has become the dominant approach in natural language processing (NLP). The key component of the transformer is a multi-head self-attention mechanism that captures long-range dependencies. Building upon the success of the transformer in various tasks [36–38], researchers have increasingly focused on exploring its potential applications in 3D human motion prediction [22–24]. Cai et al. [22] employed a transformer-based architecture with discrete cosine transform (DCT) to capture the long-range spatial correlations and temporal dependencies in human motion dynamics. Another notable advancement is the spatio-temporal transformer (ST-Trans) mechanism proposed by Aksan et al. [23], which effectively captures the spatio-temporal dependencies of decomposed 3D human motions. However, the ST-Trans method overlooks the importance of ensuring consistency between spatial and temporal information, which is a crucial factor when dealing with time-varying data. To address this limitation, a cross-transformer approach [24] has been developed to explore effective interaction between spatial and temporal branches. This approach is designed to learn the coherence of spatial and temporal information and simultaneously enhance the model's predictive capacity. Despite these advancements, transformer-based methods may overlook local information when dealing with human motion data, warranting further investigation into this aspect.

In summary, with the development of human motion prediction in recent years, GCN/GRU/transformer-based architectures have been well explored and results have significantly improved. In this paper, we proposed a new spatial-temporal graph convolutional network framework to address the human motion prediction problem. We used graph convolutional networks with a position-wise attention mechanism to capture the spatial dependencies of the human body joints. A gated recurrent neural network with

transformer layers was used to capture both local and global information of human motion sequences in the temporal dimension.

3. The Proposed Framework

Problem Fromulation. In the realm of human motion prediction, the objective is to forecast future human movements utilizing historical motion data. Specifically, these historical motion data can be expressed as a time series within a 3D skeletal structure. This structure can be denoted as $G = (V, E)$, where $V = \{v_1, v_2, \ldots, v_M\}$ represents a set of M joints in the human body, and E is a set of edges that represent the physical connections between these joints. The relationships between joints can be demonstrated by a symmetric adjacency matrix $A \in R^{M \times M}$, where the i, j-th element, A_{ij}, represents the biomechanical correlation between joints v_i and v_j. This correlation is typically gauged by the physical linkage or common motion patterns shared between two joints. Note that $A_{ij} = 0$ if there is no substantial correlation between the two respective joints. The historical motion data can be represented in a sequence $Z = (z_1, z_2, \ldots, z_\tau)$ on the 3D skeletal structure, where each $z_t \in R^{M \times 1}$ signifies the motion status of M joints at time t. With the above notations, we can formally state the problem as follows.

Given the 3D skeletal structure $G = (V, E)$ and the historical motion data $Z = (z_1, z_2, \ldots, z_\tau)$, the aim was to formulate a model f that can accept a new sequence $X = (x_1, x_2, \ldots, x_T)$ of length T as an input and predict the motion status for the subsequent T' time steps, denoted as:

$$X_{pred} = f(G; X|Z) = (x_{T+1}, \ldots, x_{T+T'}). \quad (1)$$

Moreover, during the training phase, we employed a sliding window of length $T + T'$ over the historical motion sequence Z to generate the training samples. These samples serve to train the model f.

Overview. The conceptual framework of the proposed GGTr network, illustrated in Figure 2, is primarily composed of three crucial components. The spatial graph convolutional network (GCN) layers were designed to model the spatial correlation between the human body joints graph. These layers were applied to the input representations of the gated recurrent unit (GRU), which modeled the sequential temporal relations, also referred to as local temporal dependencies. The framework also includes a transformer layer that is specifically designed to directly apprehend the long-range or global temporal dependencies within the motion sequence. The GRU and transformer layers work in tandem to capture the temporal dependencies for each joint independently, albeit from distinct perspectives. Subsequently, we delve into the spatial dependency modeling with the GCN, followed by an explanation of how the GRU layer and the transformer layer function to capture temporal dependencies, leading to a comprehensive summary of the complete framework.

3.1. Spatial Dependence Modeling

Acquiring the complex spatial dependence is a crucial problem in human motion prediction. The traditional convolutional network (CNN) can grasp local spatial features, but its application is primarily confined to Euclidean space, such as images or regular grids. However, the human body essentially forms a graph, not a two-dimensional grid. A graph is a structure composed of nodes and edges. In the context of human motion prediction, nodes can represent joints in the human body, and edges can represent connections between joints. Graphical models can effectively represent the complex relationships between human joints, which is useful for human motion prediction. This means the CNN model cannot reflect the complex topological structure of the human body and thus cannot accurately capture spatial dependence. Recently, generalizing the CNN to the graph convolutional network (GCN), which can handle arbitrary graph-structured data, has received widespread attention.

In the human motion prediction problem, if two body joints are connected or in close proximity, their movements are likely to mutually influence each other. Thus, to capture these spatial relationships, we employed the graph convolutional networks model proposed in [29,31]. This model is used to transform and propagate motion information through the graph structure. Specifically, given input motion information $X^l \in R^{M \times d^l}$ on the structure, where d^l represents the input dimension, the output $X^{l+1} \in R^{M \times d^{l+1}}$ can be computed, with d^{l+1} denoting the output dimension.

$$X^{l+1} = \sigma(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}X^l W^l), \tag{2}$$

where $\sigma(.)$ is a non-linear activation function. In our work, we adopted ReLU $(\cdot)$, standing for REctified Linear Unit, which is advantageous for its ability to speed up the convergence of stochastic gradient descent compared to sigmoid and tanh functions. $\tilde{A} = A + I_M$ is the adjusted adjacency matrix with I_M as the M-dimensional identity matrix, and $\tilde{D}$ is the modified degree matrix, with $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$. W^l are the parameters to be learned. For convenience, we represent the operation in Equation (2) as follows:

$$X^{l+1} = GCN_0(X^l, A). \tag{3}$$

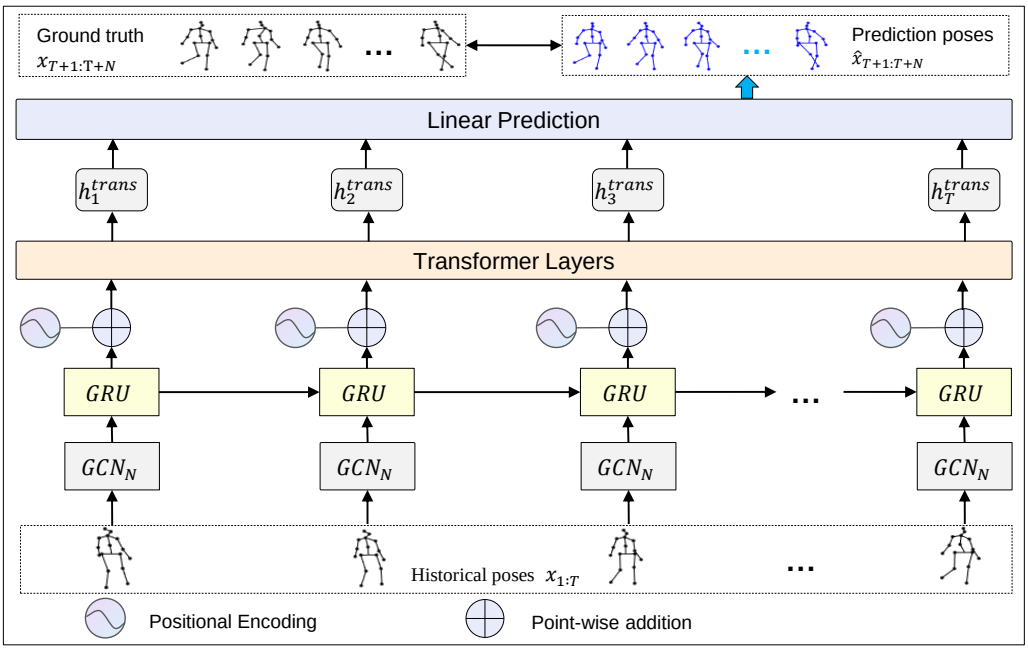

Figure 2. The whole architecture of our suggested solution for motion prediction, which employs an end-to-end framework. We encode human poses $x_{1:T}$ and feed them to GGTr. The GCN layer aids in understanding spatial relationships among body joints within the human motion network. Following the GRU is a transformer layer, designed to grasp global temporal dependencies. Ultimately, the transformer's output is harnessed to predict future motion states.

In the aforementioned formulation, the operation relies solely on the human body joints structure information, which is premised on the physical proximity between body parts. However, the interplay between body joints can be significantly more intricate. Various factors such as muscle constraints, physical capabilities, motion style, and ongoing actions can impact the motion of body parts. Therefore, the information from neighboring joints should not be aggregated equally to a given central joint when performing the aggregation in Equation (2). Recently, GaAN [39] attempted to employ the attention mechanism to model the complex relationships between graph nodes. Ideally, the aforementioned factors could be used to calculate the attention score, but these factors are not always available. Furthermore, there may be other factors influencing the relations between joints that we are

not aware of. Therefore, in this paper, we proposed learning a positional representation to capture these factors for each joint. Specifically, for joints v_i, v_j, we aimed to learn two latent positional representations $h_i, h_j \in R^M$. We then modeled the pairwise relations between any body joints as:

$$\begin{aligned} e_{ij} &= a([Wh_i||Wh_j]) \\ \alpha_{ij} &= \frac{\exp(LeakyReLU(a(h_i, h_j)))}{\sum_{k=1}^{M} \exp(LeakyReLU(a(h_i, h_j)))}, \end{aligned} \tag{4}$$

where $a(.)$ is a relation score function modeled with dot product as follows:

$$a(h_i, h_j) = (Wh_i)^T Wh_j, \tag{5}$$

where W is a transformation matrix to be learned. We further employed a mask to sparsify the relation matrix α in order to reduce computational complexity:

$$mask(\alpha) = \begin{cases} \alpha_{ij}, & \text{if } \tilde{A}_{ij} > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Subsequently, the GCN operation can be applied to the newly learned relation matrix $mask(\alpha)$:

$$X^{l+1} = \sigma(\tilde{D}_{\alpha}^{-1/2} \tilde{\alpha} \tilde{D}_{\alpha}^{-1/2} X^l W^l), \tag{7}$$

where $\tilde{\alpha} = mask(\alpha) + I_M$ and $\tilde{D}_{\alpha}$ is the degree matrix for $\tilde{\alpha}$. For simplicity, we represent the operation in Equation (7) as:

$$X^{l+1} = GCN_N(A, X^l). \tag{8}$$

Instead of the operation in (2), the operation in (8) was used to capture the spatial relations. This approach allowed us to learn a positional representation for each joint, capturing various factors that influence their relationships. By using these learned representations in combination with an attention mechanism and a sparsified relation matrix, we can effectively model complex spatial relationships between different body joints and improve the accuracy of human motion prediction.

3.2. Temporal Dependence Modeling

The gated recurrent unit (GRU) is a variant of a traditional RNN and can effectively capture the semantic association between long sequences and alleviate the problem of gradient vanishing or explosion. Its core structure can be divided into two parts for analysis: update gate and reset gate, illustrated in Figure 3. To capture the temporal dependency in human motion sequences, we adopted the gated recurrent unit (GRU) [40] to process the sequence information.

For each time step, we maintained a hidden representation that served as the output for the current step and influenced the information flow to the subsequent step. It is important to note that the GRU operation was applied individually to each joint in the body, and the parameters were shared for all joints. To incorporate spatial relations in human motion while processing the sequence, we applied a modified graph convolutional network (GCN) operation, as described in Equation (8), to the input representations of the GRU. Specifically, at each time step t, with input x_t and previous step's hidden representations h_{t-1}, we applied the modified GCN operation:

$$\begin{aligned} \tilde{x}_t &= GCN_N(A, x_t) \\ \tilde{h}_{t-1} &= GCN_N(A, h_{t-1}). \end{aligned} \tag{9}$$

For a body joint v_i at time step t, the history observation $x_t[i]$ contains the dynamic features for joint v_i at *t-th* time step. The update gate and reset gate are denoted as follows:

$$\begin{aligned} z_t &= \sigma(W_z \tilde{x}_t[i] + U_z \tilde{h}_{t-1}[i] + b_z) \\ r_t &= \sigma(W_r \tilde{x}_t[i] + U_r \tilde{h}_{t-1}[i] + b_r), \end{aligned} \tag{10}$$

where W_z, W_r, W_h, U_z, U_r, b_z, b_r are the trainable parameters.

We then calculated the candidate hidden state and the updated hidden state $h_t[i]$ as follows:

$$\begin{aligned} \tilde{h}_t[i] &= \tanh(W_h \tilde{x}_t[i] + r_t \odot U_h \tilde{h}_{t-1}[i] + b_h) \\ h_t[i] &= (1 - z_t)\tilde{h}_{t-1}[i] + z_t \odot \tilde{h}_t[i], \end{aligned} \tag{11}$$

where $\odot$ denotes the element-wise multiplication, U_h,b_h are the parameters to be learned, and $h_t[i]$ serves as extracted features of joint v_i at the next time step.

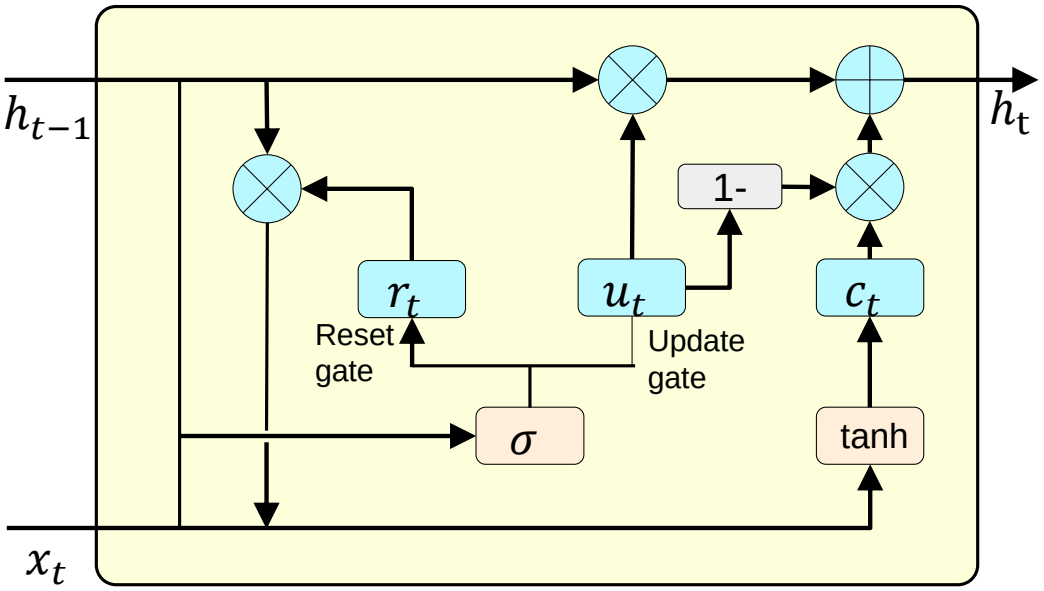

Figure 3. Architectural elements in a GRU layer.

The above are the fundamental equations and steps of GRU. By controlling the update and reset gates, GRU can dynamically update and adjust the hidden state based on different patterns in the sequence, enabling it to better capture the local temporal information. However, in the human body motion prediction problem, the temporal information may not only be sequentially dependent. Hence, it is important to capture the global temporal information for accurate human motion prediction. Thus, after processing with the GRU layer, we adopted a transformer layer [35] to capture the global dependencies following the GRU.

The transformer layer, similar to the GRU layer, was applied to each joint individually. For joint v_i, the output sequence $(h_1[i], \ldots, h_T[i])$ from GRU was taken as the input for the transformer. Figure 4 illustrates that a transformer layer consists of a multi-head attention layer, a shared feed-forward neural work layer, and normalization layers between them.

Inputs of model. The input of the model is the spatial dependency embedding sequence of each joint. Let us consider the *i-th* joint's spatial dependency embedding sequence $(h_1[i], h_2[i], \ldots, h_T[i])$. Although the self-attention mechanism is effective in capturing hidden dependency relationships in the sequence, it fails to maintain location information. Hence, we incorporated position encoding e_t as proposed by [35], between GRU and the transformer layer. This choice was motivated by the ability of this approach to facilitate the

model's learning to attend to relative positions in a sequence. The new representations $h'_t[i]$ were computed as follows:

$$h'_t[i] = h_t[i] + e_t, \tag{12}$$

where e_t is defined as

$$e_t = \begin{cases} sin(t/10000^{2i/d_{model}}), & \text{if } i = 0, 2, 4 \ldots \\ cos(t/10000^{2i/d_{model}}), & \text{otherwise.} \end{cases} \tag{13}$$

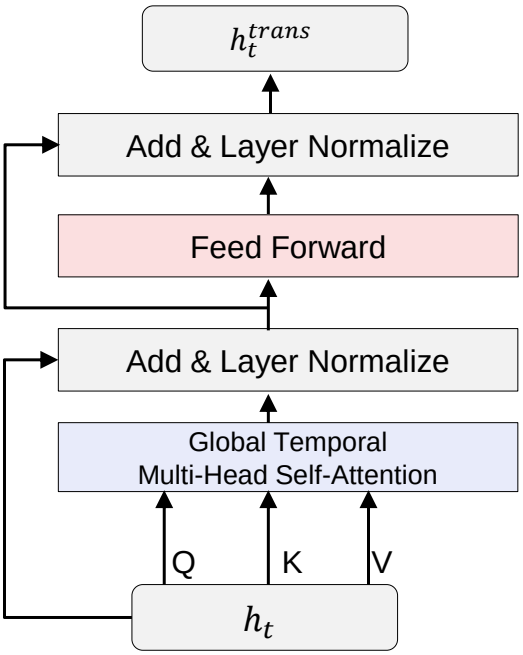

Figure 4. Architectural elements in a transformer layer.

Temporal multi-head attention. In our model, we defined a matrix $h^{v_i} \in R^{T \times d}$ by arranging $(h_1[i], h_2[i], \ldots, h_T[i])$ in a row-wise manner across the temporal axis. The multi-head attention mechanism is composed of several attention heads, which process the input in parallel and whose outputs are then combined. This concept can be represented mathematically as follows:

$$MultiHead(h^{v_i}) = Concat(head_1, \ldots, head_s)W^O. \tag{14}$$

This signifies that the output of multi-head attention is the concatenation of s single head attention blocks, each one being projected with the matrix W^O. Each single head attention block, $head_s$, can be given by:

$$\begin{aligned} head_s &= Attention(h^{v_i}) \\ &= softmax(\frac{Q_s K_s^T}{\sqrt{d_k}})V_s, \end{aligned} \tag{15}$$

where W_s^Q, W_s^K, and W_s^V denote the queries, keys, and values of the *s-th* single head attention for joint v_i, respectively. We obtained the Q_s, K_s, and V_s by a linear projection with W^Q, W^K, and W^V:

$$Q_s = h^{v_i} W_s^Q,\ K_s = h^{v_i} W_s^K,\ V_s = h^{v_i} W_s^V, \tag{16}$$

where $W_s^Q \in \mathbb{R}^{d \times d_k}$, $W_s^K \in \mathbb{R}^{d \times d_k}$, and $W_s^V \in \mathbb{R}^{d \times d_v}$ are the trainable projection matrices for the *s-th* attention head and are shared by all the joints.

Feed-forward Layer. After the multi-head attention layer, the output states are then processed by a feedforward neural network layer. This layer is identical for all joints, which includes two linear transformations and a ReLU activation function in between:

$$FFN(X) = max(0, XW_1 + b_1)W_2 + b_2, \tag{17}$$

where W_1, W_2 are the trainable parameters of the feed-forward neural network and b_1 and b_2 are bias terms.

Residual connection and normalization. As shown in Figure 4, the residual connection and normalization operators appear in each layer. The residual connection was introduced to tackle the difficulty in training deep network optimization algorithms, and normalization was applied to prevent overfitting. Given the embedded feature Z_{in}, the process of the residual connection and normalization layer is denoted as follows:

$$\begin{aligned} Z_{in}^{'} &= LayerNorm(Z_{in} + MultiHead(Z_{in})) \\ h_{out} &= LayerNorm(Z_{in}^{'} + FFN(Z_{in}^{'})), \end{aligned} \tag{18}$$

where $LayerNorm(.)$ denotes the normalization function.

Lastly, after each sub-layer in the transformer, such as the temporal multi-head attention layer and the feed-forward network layer, a layer normalization operation and a residual connection were applied. The final output of the transformer layer for joint v_i can be denoted as $h_{out}^{v_i} \in \mathbb{R}^{T \times d}$.

In summary, this approach combines both the local and global temporal information through the GRU and transformer layers, which can be highly beneficial for accurate human motion prediction. The individual sequence of movements and the overall pattern of movements are both considered, making this method a versatile and robust choice for this task.

3.3. Loss Function

To train our model, we employed an end-to-end training technique. The mean position per joint error (MPJPE) loss [41] function between the anticipated motion sequence and the ground truth motion sequence was used to analyse the difference between the predicted outcomes and the true pose, which is defined as follows:

$$L = \frac{1}{M \times T} \sum_{i=1}^{M} \sum_{t=1}^{T} ||\hat{x}_t[i] - x_t[i]||_2^2, \tag{19}$$

where M is the number of joints in human pose, T is the number of time steps in the future series, $\hat{x}_t[i]$ is the predicted joint position at the *t-th* time step of the *i-th* joint, and $x_t[i]$ is the corresponding ground truth.

We optimised the loss function using the improved Adam method (AdamW [42]), which mitigates the overfitting problem by adding a weight decay term and can significantly improve the robustness of the model.

4. Experiments

In this section, we present experiments on two large-scale human motion capture benchmark datasets (Human3.6M and CMU-MoCap) to demonstrate the effectiveness of the GGTr network for human motion prediction. We first introduced the experimental settings, including datasets, baselines, and parameter settings. Then, we conducted experiments to compare the performance of the GGTr with other baselines. Finally, we designed comprehensive ablation studies to evaluate the impact of the essential architectural components.

4.1. Datasets

Human3.6M [41] is the largest existing human motion analysis database, consisting of seven actors (S1, S5, S6, S7, S8, S9, and S11) performing 15 actions such as walking, eating, smoking, discussing, and directions. Each pose includes 32 joints, represented in the form of an exponential map. Following the data processing of [18], by converting these into 3D coordinates, eliminating redundant joints, global rotation, and translation, the resulting skeleton retains 17 joints that provide sufficient human motion details. These joints include key ones that locate major body parts (e.g., shoulders, knees, and elbows). We downsampled the frame rate to 25 fps and used S5 and S11 for testing and validation, while the remaining five actors were used for training.

CMU-MoCap (Available at http://mocap.cs.cmu.edu/ accessed on 5 July 2023) is a 3D human motion dataset, released by Carnegie Mellon University, that used 12 Vicon infrared MX-40 cameras to record the positions of 41 sensors attached to the human body, describing human motion. The dataset can be divided into six motion themes, including human interaction, interaction with environment, locomotion, physical activities and sports, situations and scenarios, and test motions. We adopted the same data preprocessing method as described in the literature [43], simplifying each human body and reducing the motion rate to 25 frames per second. Furthermore, seven actions were selected from the dataset to evaluate the model's performance. No hyperparameters were adjusted on this dataset. We used only the training and testing sets, with the splitting method consistent with common practice in the literature [43].

4.2. Implementation Details

All experiments in this paper were implemented using the PyTorch deep learning framework. The experimental environment was Ubuntu 20.04, utilizing an NVIDIA A100 GPU. During the training process, a batch size of 16 was used and the number of attention heads was set to 4. The number of GRU and transformer layers were both set to 1. The model was optimized using the AdamW optimizer. The initial learning rate was set to 0.003, with a 5% decay every 5 epochs. Training was conducted for 800 epochs, and each experiment was repeated three times to obtain an average result, ensuring a more robust evaluation of the model's performance. For the input motion prediction, a length of 25 frames (1000 ms) was considered, and the prediction generated 25 frames (1000 ms). The choice and configuration of related hyperparameters are summarized in Table 1.

Table 1. The choice and configuration of related hyperparameters.

Hyperparameter/Config	Value
Optimizer	AdamW
Base learning rate	5×10^{-3}
Weight decay	10^{-2}
Batch size	16
Warmup epochs	5
Epochs	800

4.3. Evaluation Metrics and Baselines

The evaluation metrics employed in our study were consistent with those utilized in existing algorithms [33,43]. The mean per joint position error (MPJPE) [41] was adopted as the standard measurement, which computed the average Euclidean distance (in millimeters, mm) between the predicted joint 3D coordinates and the ground truth. In addition, to further illustrate the advantages of the proposed method, a comparative analysis was conducted with several state-of-the-art approaches [15,18,22,23,33,34,43].

Res. sup. [18] is an early RNN-based method. LTD [43], MSR [34], ST-DGCN [33], and LCDC [15] utilize GCN-based methodologies. Meanwhile, LPJP [22] and STCT [23] employ transformer-based approaches. We compared the proposed method with all the above approaches on the Human3.6M dataset, and with approaches [15,18,22,34,43] on CMU-MoCap. By comparing the proposed method with these existing approaches, this study

aimed to demonstrate its effectiveness and highlight its advantages in terms of accuracy and performance.

4.4. Experimental Results and Analysis

Human3.6M. Table 2 presents quantitative comparisons of our method and other approaches in predicting short-term (80 ms, 160 ms, 320 ms, 400 ms) and long-term movements (560 ms, 1000 ms) on the Human3.6M dataset. The final column provides the average performance across all tested time intervals for 15 actions.

Table 2. Prediction of 3D joint positions on Human3.6M for all actions. The best results are marked in bold.

Time(ms)	80	160	320	400	560	1000	80	160	320	400	560	1000
Action	**Walking**						**Eating**					
Res. sup. [18]	29.4	50.8	76.0	81.5	81.7	100.7	16.8	30.6	56.9	68.7	79.9	100.2
LTD [43]	12.3	23.0	39.8	46.1	54.1	59.8	8.4	16.9	33.2	40.7	53.4	77.8
ST-Trans [23]	**8.9**	**15.5**	**32.1**	**38.5**	-	-	9.4	21.1	36.4	42.3	-	-
MSR [34]	12.2	22.7	38.6	45.2	52.7	63.0	8.4	17.1	33.0	40.4	52.5	77.1
ST-DGCN [33]	10.2	19.8	34.5	40.3	48.1	56.4	7.0	15.1	**30.6**	38.1	51.1	76.0
LCDC [15]	11.1	22.4	38.8	45.2	52.7	59.8	7.0	15.5	31.7	39.2	51.9	76.2
Ours	10.3	19.5	34.9	41.8	**51.1**	**54.9**	**6.9**	**15.0**	31.6	**36.3**	**50.2**	**71.7**
Action	**Smoking**						**Discussion**					
Res. sup. [18]	23.0	42.6	70.1	82.7	94.8	137.4	32.9	61.2	90.9	96.2	121.3	161.7
LTD [43]	7.9	16.2	31.9	38.9	50.7	72.6	12.5	27.4	58.5	71.7	91.6	121.5
ST-Trans [23]	8.8	15.2	**25.1**	**24.5**	-	-	**7.9**	25.7	**39.9**	**47.5**	-	-
MSR [34]	8.0	16.3	31.3	38.2	49.5	71.6	12.0	26.8	57.1	69.7	88.6	117.6
ST-DGCN [33]	6.6	14.1	28.2	34.7	46.5	69.5	10.0	23.8	53.6	66.7	87.1	118.2
LCDC [15]	6.6	14.8	29.8	36.7	48.1	71.2	10.0	24.4	54.5	67.4	87.0	116.3
Ours	**6.4**	**14.1**	28.4	33.1	**45.3**	**67.3**	9.8	**23.2**	49.5	58.4	**83.9**	**106.5**
Action	**Directions**						**Greeting**					
Res. sup. [18]	35.4	57.3	76.3	87.7	110.1	152.5	34.5	63.4	124.6	142.5	156.1	166.5
LTD [43]	9.0	19.9	43.4	53.7	71.0	101.8	18.7	38.7	77.7	93.4	115.4	148.8
ST-Trans [23]	10.2	17.8	42.5	48.6	-	-	13.5	**26.1**	**54.0**	**73.2**	-	-
MSR [34]	8.6	19.7	43.3	53.8	71.2	100.6	16.5	37.0	77.3	93.4	116.3	147.2
ST-DGCN [33]	7.2	17.6	40.9	51.5	69.3	100.4	15.2	34.1	71.6	87.1	110.2	143.5
LCDC [15]	6.9	17.4	41.0	51.7	69.1	99.1	14.3	33.5	72.2	87.3	108.7	142.3
Ours	**6.9**	**17.0**	**39.2**	**48.4**	**66.4**	**94.6**	**13.4**	39.3	69.4	81.6	**103.3**	**137.6**
Action	**Phoning**						**Posing**					
Res. sup. [18]	38.0	69.3	115.0	126.7	141.2	131.5	36.1	69.1	130.5	157.1	194.7	240.2
LTD [43]	10.2	21.0	42.5	52.3	69.2	103.1	13.7	29.9	66.6	84.1	114.5	173.0
ST-Trans [23]	15.3	20.4	**31.4**	**38.8**	-	-	10.6	22.8	57.6	73.7	-	-
MSR [34]	10.1	20.7	41.5	51.3	68.3	104.4	12.8	29.4	67.0	85.0	116.3	174.3
ST-DGCN [33]	**8.3**	18.3	38.7	48.4	65.9	102.7	10.7	25.7	60.0	76.6	106.1	164.8
LCDC [15]	8.5	19.2	40.3	49.9	66.7	102.2	10.1	25.4	60.6	77.3	106.5	163.3
Ours	8.6	**18.4**	39.9	46.5	**63.8**	**99.1**	**9.9**	**22.6**	**57.6**	**73.5**	**103.7**	**158.4**
Action	**Purchases**						**Sitting**					
Res. sup. [18]	36.3	60.3	86.5	95.9	122.7	160.3	42.6	81.4	134.7	151.8	167.4	201.5
LTD [43]	15.6	32.8	65.7	79.3	102.0	143.5	10.6	21.9	46.3	57.9	78.3	119.7
ST-Trans [23]	17.3	32.5	**60.0**	68.3	-	-	8.5	22.9	47.8	66.8	-	-
MSR [34]	14.8	32.4	66.1	79.6	101.6	139.2	10.5	22.0	46.3	57.8	78.2	120.0
ST-DGCN [33]	12.5	28.7	60.1	73.3	95.3	133.3	8.8	19.2	42.4	53.8	74.4	116.1
LCDC [15]	12.7	29.7	62.3	75.8	97.5	137.8	8.8	19.3	42.9	54.3	74.9	117.8
Ours	**12.3**	**28.5**	61.7	**67.9**	**91.1**	**126.1**	**8.5**	**19.1**	**40.1**	**49.8**	**69.6**	**115.2**

Table 2. *Cont.*

Action			Sitting Down						Taking Photo			
Res. sup. [18]	47.3	86.0	145.8	168.9	205.3	277.6	26.1	47.6	81.4	94.7	117.0	143.2
LTD [43]	16.1	31.1	61.5	75.5	100.0	150.2	9.9	20.9	45.0	56.6	77.4	119.8
ST-Trans [23]	**9.2**	32.7	58.8	65.9	-	-	**6.8**	**16.5**	**37.9**	**48.5**	-	-
MSR [34]	16.1	31.6	62.5	76.8	102.8	155.5	9.9	21.0	44.6	56.3	77.9	121.9
ST-DGCN [33]	13.9	27.9	57.4	71.5	96.7	147.8	8.4	18.9	42.0	53.3	74.3	118.6
LCDC [15]	14.1	28.0	57.3	71.2	96.1	147.3	8.4	18.8	42.0	53.5	74.5	117.9
Ours	13.8	**26.1**	**55.3**	**65.3**	**94.2**	**141.9**	8.1	17.6	38.3	48.7	**72.8**	**110.8**
Action			**Waiting**						**Walking Dog**			
Res. sup. [18]	30.6	57.8	106.2	121.5	146.2	196.2	64.2	102.1	141.1	164.4	191.3	209.0
LTD [43]	11.4	24.0	50.1	61.5	79.4	108.1	23.4	46.2	83.5	96.0	111.9	148.9
ST-Trans [23]	9.2	20.5	59.8	62.2	-	-	22.3	67.8	103.8	122.5	-	-
MSR [34]	10.7	23.1	48.3	59.2	76.3	106.3	20.7	42.9	80.4	93.3	111.9	148.2
ST-DGCN [33]	8.9	20.1	**43.6**	54.3	72.2	103.4	18.8	39.3	73.7	86.4	104.7	139.8
LCDC [15]	8.7	20.2	44.3	55.3	73.2	105.7	19.6	41.8	77.6	90.2	109.8	147.7
Ours	**8.5**	**19.7**	44.6	**50.0**	**68.6**	**101.5**	**17.6**	**38.3**	**73.5**	**84.1**	**103.5**	**135.8**
Action			**Waiting Together**						**Average**			
Res. sup. [18]	26.8	50.1	80.2	92.2	107.6	131.1	34.7	62.0	101.1	115.5	97.6	130.5
LTD [43]	10.5	21.0	38.5	45.2	55.0	65.6	12.7	26.1	52.3	63.5	81.6	114.3
ST-Trans [23]	**7.5**	**13.7**	**27.7**	46.2	-	-	11.0	24.7	47.7	57.8	-	-
MSR [34]	10.6	20.9	37.4	43.9	52.9	65.9	12.1	25.6	51.6	62.9	81.1	114.2
ST-DGCN [33]	8.7	18.6	34.4	41.0	51.9	64.3	10.3	22.7	47.4	58.5	76.9	110.3
LCDC [15]	9.1	19.8	36.3	42.7	50.5	61.2	10.4	23.3	48.8	59.8	77.8	111.0
Ours	9.1	19.2	36.1	**39.8**	**48.4**	**57.7**	**10.0**	**21.8**	**46.8**	**55.1**	**74.4**	**105.4**

Among the comparison methods, RNN-based techniques exhibited the poorest performance, while transformer-based approaches outperformed GCN-based methods. Above all, our method emerged as the top performer. The most significant improvements in the MPJPE metric were observed with transformer-based techniques, highlighting their aptitude for modeling 3D human motion dynamics and capturing global dependencies.

Specifically, existing methods generally performed well when predicting periodic and simpler movements, such as "walking" and "eating". However, their performance dropped significantly when tasked with predicting more unpredictable and irregular movements like "directions", "posing", and "purchases". This indicates that these methods have difficulty managing the dynamic changes and local–global dependencies inherent in human motion.

On the other hand, the algorithm proposed in our study demonstrated a high prediction accuracy, even with highly complex, non-periodic, and irregular movements. Our experimental results revealed that our proposed method surpasses most baseline methods in short-term motion prediction, with even greater improvements noted in long-term prediction. Our method delivered a superior performance in the 560 ms and 1000 ms MPJPE metrics, an accomplishment attributable to GGTr's ability to fully capture spatial correlation and local–global temporal features, thereby bolstering the model's prediction accuracy.

While our predictions for movements such as "walking" , "smoking" , "greeting" and "waiting together" fell short compared to those of MSR [34], this nonetheless underscores the sophisticated nature of transformer-based approaches. Looking ahead, we are committed to further refining our model to enhance its performance.

Overall, our proposed method outperformed all the baseline models on average, proving its superior performance in motion prediction. The outstanding performance across both short-term and long-term human motion prediction highlights our model's effective capacity to capture both local and global temporal dependencies.

CMU-MoCap. To further validate the generalization of the proposed method, we compared its performance with five existing algorithms [15,18,22,34,43] on the CMU-MoCap dataset. The mean per joint position error was calculated for short-term and long-term predictions. The experimental results are shown in Table 3, which includes the actions "basketball", "basketball signal", "directing traffic", "jumping", "soccer", "walking", and "wash window", as well as the average prediction error across all actions.

Table 3. Prediction of 3D joint positions on CMU-MoCap for all actions. The best results are marked in bold.

Time(ms)	80	160	320	400	560	1000	80	160	320	400	560	1000
Action	**Basketball**						**Basketball Signal**					
Res. sup. [18]	29.5	53.1	91.2	106.0	128.7	157.4	14.6	22.1	39.1	46.6	60.0	89.9
LTD [43]	11.7	21.1	40.7	50.6	68.0	95.7	3.4	6.2	13.5	17.9	27.3	51.9
LPJP [22]	11.6	21.7	44.4	57.3	-	90.9	2.6	4.9	12.7	18.7	-	75.8
MSR [34]	10.3	18.9	37.7	47.0	62.0	86.3	3.0	5.6	12.5	16.6	25.5	50.0
LCDC [15]	9.6	17.6	35.4	44.4	60.0	**88.4**	2.6	4.7	**10.4**	13.9	21.9	46.2
Ours	**9.5**	**17.5**	**32.5**	**41.5**	**56.8**	88.4	**2.5**	**4.6**	11.5	**13.1**	**20.6**	**45.8**
Action	**Directing Traffic**						**Jumping**					
Res. sup. [18]	21.8	38.8	70.5	85.3	110.3	165.1	30.2	53.0	89.4	103.9	125.6	160.5
LTD [43]	6.8	13.4	29.6	39.1	59.6	112.8	17.1	32.1	59.8	72.5	94.3	127.2
LPJP [22]	6.2	12.7	29.1	39.6	-	149.1	12.9	27.6	73.5	92.2	-	176.6
MSR [34]	6.1	12.6	29.4	39.2	50.5	114.6	15.2	28.9	56.0	69.1	92.4	126.2
LCDC [15]	**5.0**	10.0	23.4	31.4	49.3	99.6	**12.8**	**26.1**	54.6	68.5	91.8	126.1
Ours	5.0	**9.9**	**22.4**	**30.9**	**48.6**	**93.9**	13.1	27.2	**53.6**	**67.8**	**89.8**	**116.9**
Action	**Soccer**						**Walking**					
Res. sup. [18]	26.5	47.0	81.5	96.2	117.9	139.1	14.6	22.9	36.1	40.9	51.1	69.5
LTD [43]	13.6	24.3	44.4	54.3	73.1	111.6	6.7	11.1	18.1	21.0	25.2	32.4
LPJP [22]	**9.2**	**18.4**	39.2	49.5	-	93.9	6.7	10.7	21.7	27.5	-	37.4
MSR [34]	10.9	19.4	37.4	47.0	65.3	101.9	6.4	10.3	16.9	20.1	25.5	36.8
LCDC [15]	10.3	19.0	36.8	45.7	62.3	96.9	6.3	10.4	16.1	18.6	23.3	33.6
Ours	9.8	19.0	**35.3**	**44.6**	**59.7**	**92.7**	**5.8**	**10.2**	**15.4**	**17.3**	**22.9**	**33.2**
Action	**Wash Window**						**Average**					
Res. sup. [18]	19.3	31.8	56.1	66.0	83.6	125.9	22.4	38.4	66.2	77.8	96.7	129.6
LTD [43]	5.9	11.3	24.1	31.0	43.4	66.9	9.3	17.1	32.9	40.9	55.9	85.5
LPJP [22]	5.4	11.3	29.2	39.6	-	79.1	7.8	15.3	35.7	46.3	-	100.4
MSR [34]	5.4	10.9	24.5	31.8	45.1	70.2	8.2	15.2	30.5	38.7	52.3	83.7
LCDC [15]	**4.8**	**9.5**	22.0	29.0	42.5	68.9	7.3	**13.9**	28.4	35.9	50.1	80.0
Ours	4.8	9.5	**20.3**	**28.0**	**41.3**	**67.4**	**7.2**	14.0	**27.3**	**34.7**	**48.5**	**76.9**

Through quantitative evaluation, we clearly observe that our method effectively handles various types of actions and consistently achieves a superior performance across all of them. These empirical findings reinforce the superiority of our approach in human motion prediction, both for short-term and long-term predictions. The consistent and significant performance improvement over state-of-the-art methods on the dataset demonstrates the robustness of our method.

Notably, as can be seen from Tables 2 and 3, our model performed well on the Human3.6m and CMU-MoCap datasets for prediction tasks up to 320 ms, and even better for tasks above 320 ms. This intriguing performance pivot at 320 ms may represent a transition point where prognostic difficulty shifts from short to long-term. We believe this is due to our model's capability to capture local and global time dependencies across body joints through a strategic combination of GRU and transformer layers, effectively handling this transition. Furthermore, this phenomenon could be tied to the inherent complexity of human movement, especially when the prediction time exceeds 320 ms. This complexity poses a challenge for models that rely on short-term prognosis. However, our model, thanks to its specific structure and training procedure, seems to cope with this challenge more effectively. In future work, we plan to delve deeper into this "prognostic barrier" at 320 ms, aiming to better understand its underlying causes and how we can further improve our model's performance at this critical transition point. This understanding will potentially enable us to optimize the dynamics of 3D joint position prediction.

4.5. Qualitative Comparison

To enhance our intuitive understanding and facilitate the evaluation of our model, we provided a visualization of the motion prediction results. Figure 5 illustrates three examples of predicted poses using our proposed method, alongside three other baseline methods. The visualization results for the "phoning", "discussion", and "purchases" actions of the Human3.6M dataset are presented in Figure 5. The first row in each subplot displays the

ground truth pose sequences (in black), followed by the predicted poses (in blue). In other words, each row presents the prediction results from one model.

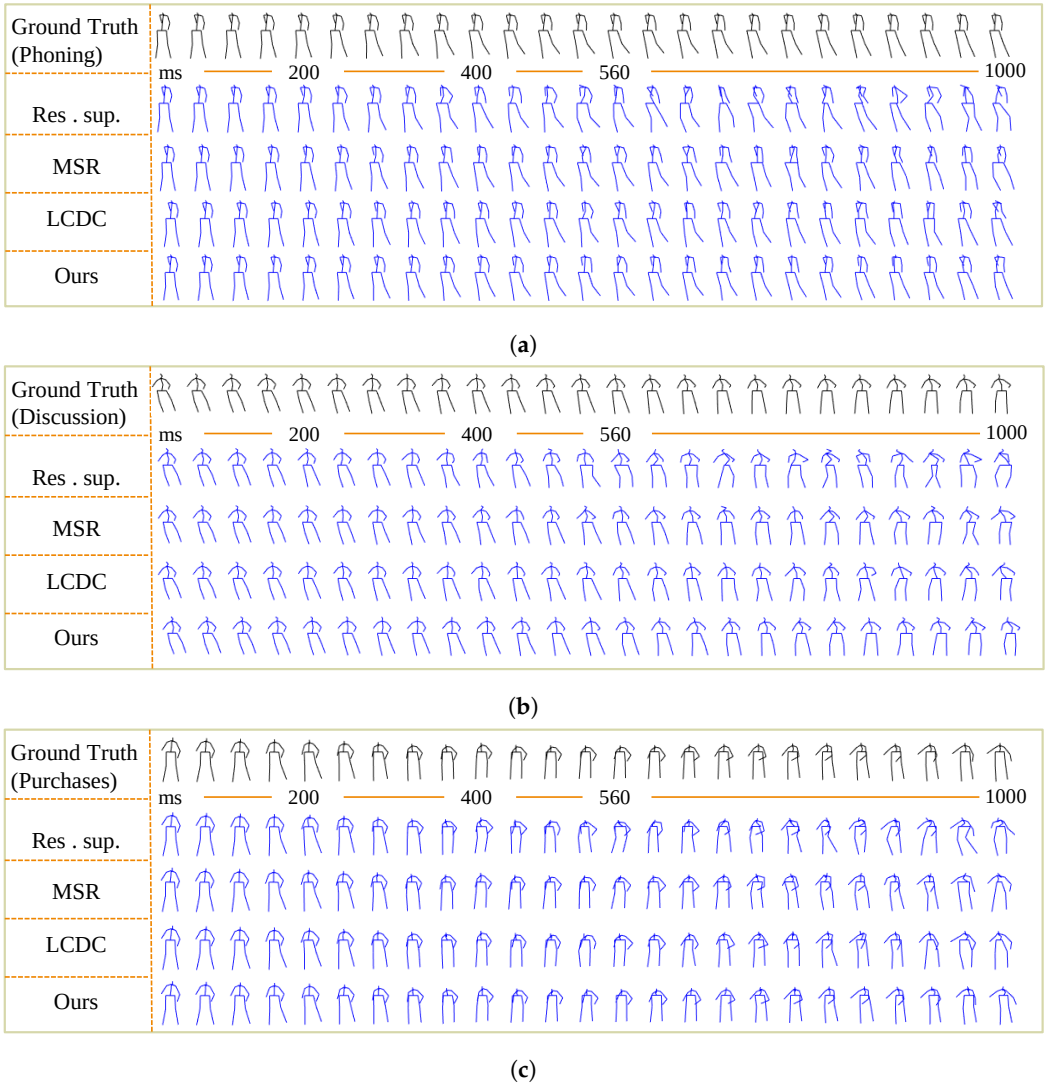

Figure 5. Qualitative analysis of the Human3.6M dataset: comparing phoning, discussion, and purchases scenarios. (**a**) Phoning; (**b**) Discussion; (**c**) Purchases.

The visualization results indicate that our model is capable of adequately capturing spatial dependencies and local–global temporal dependencies. It is noticeable that the predictions generated by our method show higher similarity to the actual sequences and better continuity between frames. For instance, in the case of subtler movements such as "phoning", our model successfully captures the long-distance temporal dependencies concealed within the movement sequence, yielding superior long-term predictions. Moreover, in the "purchases" motion visualization, the movements between hands are more coordinated. This illustrates our model's proficiency in forecasting highly complicated irregular movements and complex periodic motions.

4.6. Ablation Study

To critically assess the contribution of each component in our model, we performed a set of ablation studies on the Human3.6M dataset. All architecture parameters were kept constant to ensure a fair comparison of each module's impact. These experiments focused on assessing the impact of the graph convolutional network (GCN_N), gated recurrent units (GRU), and transformer layers (Tr) on the model's performance. The experimental results

are presented in Table 4. The optimal performance was achieved by integrating these three components.

Table 4. The influence of the GCN_N, GRU, and transformer layers (Tr) on the Human3.6M dataset, on average, is notable. These three components of our model significantly contribute to its overall accuracy. The best results are marked in bold.

			Human3.6M,MPJPE (mm)					
GCN_N	GRU	Tr	80	160	320	400	560	1000
	√	√	10.3	23.2	48.1	57.0	75.7	107.4
√		√	10.2	22.7	47.5	56.2	75.3	107.0
√	√		10.2	23.5	47.7	57.8	77.2	111.0
√	√	√	**10.0**	**21.8**	**46.8**	**55.1**	**74.4**	**105.4**

GCN_N: We used the original graph convolution module in place of the proposed module. As shown in the first and third rows of Table 4, when this module is replaced, prediction performance decreases. This result clearly indicates that there is critical spatially relevant information hidden in the adjacent pose. When this information is ignored, it is difficult for the model to capture time evolution trends, resulting in degraded performance.

GRU and transformer layers(Tr): The GRU unit, designed to capture local temporal dependence, shows a remarkable performance, enhancing the accuracy of short-term predictions. The transformer layer exhibits exceptional ability in handling global temporal dependence, which improves the accuracy of long-term predictions. We removed the transformer layer directly from the proposed method. It can be seen that the long-term prediction performance of the model was significantly reduced, and the short-term prediction performance was also slightly reduced.

5. Conclusions

In this paper, we have proposed a novel framework for human motion prediction that leverages the power of position-wise enhanced graph convolutional networks, gated recurrent unit networks, and transformer layers. By strategically combining these networks, our model effectively captures spatial information across body joints and temporally aligns these dependencies, both locally and globally. The proposed framework has shown significant improvements in predicting long-term movements, surpassing existing state-of-the-art methods by a substantial margin. Experiments on the Human3.6M and CMU-MoCap datasets provide evidence supporting the effectiveness of our proposed model. The efficacy of our approach accentuates the potential of integrating advanced neural network architectures for improved understanding and prediction of complex human motion dynamics. Future work will explore the integration of more sophisticated attention mechanisms and deep learning architectures to further enhance prediction accuracy and efficiency.

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

Funding: This work was supported in part by the National Natural Science Foundation of China under grants 62233003 and 62073072, the Key Projects of the Key R&D Program of Jiangsu Province under grants BE2020006 and BE2020006-1, and Shenzhen Natural Science Foundation under grants JCYJ20210324132202005 and JCYJ20220818101206014.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated and/or analyzed during the current study are publicly available. The Human3.6M dataset can be accessed through the reference in [41]. The CMU-MoCap dataset is publicly available and can be accessed online at http://mocap.cs.cmu.edu/ accessed on 5 July 2023. The use of these datasets is governed by their respective usage policies.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GGTr	Graph Convolutional, Gated Recurrent Unit, and Transformer layers
GCN	Graph convolutional networks
GRU	Gated recurrent unit
CNNs	Convolutional neural networks
RNN	Recurrent neural networks
GNNs	Graph neural networks

References

1. Koppula, H.S.; Saxena, A. Anticipating human activities using object affordances for reactive robotic response. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 14–29.
2. Gui, L.Y.; Zhang, K.; Wang, Y.X.; Liang, X.; Moura, J.M.; Veloso, M. Teaching robots to predict human motion. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 562–567.
3. Li, H.; Li, X.; Zhang, Z.; Hu, C.; Dunkin, F.; Ge, S.S. ESUAV-NI: Endogenous Security Framework for UAV Perception System Based on Neural Immunity. *IEEE Trans. Ind. Inform.* **2023**. [CrossRef]
4. Choi, S.H.; Park, K.B.; Roh, D.H.; Lee, J.Y.; Mohammed, M.; Ghasemi, Y.; Jeong, H. An integrated mixed reality system for safety-aware human-robot collaboration using deep learning and digital twin generation. *Robot. Comput.-Integr. Manuf.* **2022**, *73*, 102258. [CrossRef]
5. Dong, Y.; Li, X.; Dezert, J.; Zhou, R.; Zhu, C.; Wei, L.; Ge, S.S. Evidential reasoning with hesitant fuzzy belief structures for human activity recognition. *IEEE Trans. Fuzzy Syst.* **2021**, *29*, 3607–3619.
6. Sheng, W.; Li, X. Multi-task learning for gait-based identity recognition and emotion recognition using attention enhanced temporal graph convolutional network. *Pattern Recognit.* **2021**, *114*, 107868.
7. Kong, Y.; Wei, Z.; Huang, S. Automatic analysis of complex athlete techniques in broadcast taekwondo video. *Multimed. Tools Appl.* **2018**, *77*, 13643–13660. [CrossRef]
8. Lehrmann, A.M.; Gehler, P.V.; Nowozin, S. Efficient nonlinear Markov models for human motion. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
9. Wang, J.M.; Fleet, D.J.; Hertzmann, A. Gaussian Process Dynamical Models for Human Motion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *30*, 283–298.
10. 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 19 (NIPS 2006), Cambridge, MA, USA, 4–7 December 2006; Volume 19.
11. 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, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2275–2284.
12. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Symbiotic Graph Neural Networks for 3D Skeleton-Based Human Action Recognition and Motion Prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3316–3333. [CrossRef]
13. 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, Seattle, WA, USA, 13–19 June 2020; pp. 214–223.
14. Zhong, C.; Hu, L.; Zhang, Z.; Ye, Y.; Xia, S. Spatio-Temporal Gating-Adjacency GCN For Human Motion Prediction. In Proceedings of the Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
15. Fu, J.; Yang, F.; Dang, Y.; Liu, X.; Yin, J. Learning Constrained Dynamic Correlations in Spatiotemporal Graphs for Motion Prediction. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**. [CrossRef]
16. Fragkiadaki, K.; Levine, S.; Felsen, P.; Malik, J. Recurrent network models for human dynamics. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 4346–4354.
17. Jain, A.; Zamir, A.R.; Savarese, S.; Saxena, A. Structural-rnn: Deep learning on spatio-temporal graphs. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5308–5317.
18. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 2891–2900.

19. Liu, Z.; Wu, S.; Jin, S.; Liu, Q.; Lu, S.; Zimmermann, R.; Cheng, L. Towards natural and accurate future motion prediction of humans and animals. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
20. Shu, X.; Zhang, L.; Qi, G.J.; Liu, W.; Tang, J. Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3300–3315.
21. Liu, Z.; Wu, S.; Jin, S.; Ji, S.; Liu, Q.; Lu, S.; Cheng, L. Investigating pose representations and motion contexts modeling for 3D motion prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 681–697. [CrossRef]
22. 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 Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 226–242.
23. 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), London, UK, 1–3 December 2021; pp. 565–574.
24. Yu, H.; Fan, X.; Hou, Y.; Pei, W.; Ge, H.; Yang, X.; Zhou, D.; Zhang, Q.; Zhang, M. Towards Realistic 3D Human Motion Prediction with A Spatio-temporal Cross-transformer Approach. *IEEE Trans. Circuits Syst. Video Technol.* **2023**. [CrossRef]
25. 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 Village, HI, USA, 7–11 January 2019; pp. 1423–1432.
26. Guo, X.; Choi, J. Human motion prediction via learning local structure representations and temporal dependencies. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 2580–2587.
27. Zhou, L.; Zhou, Y.; Corso, J.J.; Socher, R.; Xiong, C. End-to-end dense video captioning with masked transformer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8739–8748.
28. Liu, P.J.; Saleh, M.; Pot, E.; Goodrich, B.; Sepassi, R.; Kaiser, L.; Shazeer, N. Generating wikipedia by summarizing long sequences. *arXiv* **2018**, arXiv:1801.10198.
29. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
30. Yuan, J.; Cao, M.; Cheng, H.; Yu, H.; Xie, J.; Wang, C. A unified structure learning framework for graph attention networks. *Neurocomputing* **2022**, *495*, 194–204. [CrossRef]
31. Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *Stat* **2017**, *1050*, 10–48550.
32. 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, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
33. 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 Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
34. 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 International Conference on Computer Vision, New Orleans, LA, USA, 18–24 June 2021.
35. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017 ; Volume 30.
36. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 213–229.
37. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
38. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
39. Zhang, J.; Shi, X.; Xie, J.; Ma, H.; King, I.; Yeung, D.Y. GaAN: Gated Attention Networks for Learning on Large and Spatiotemporal Graphs. In Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018, Monterey, CA, USA, 6–10 August 2018.
40. 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.
41. 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]
42. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
43. Mao, W.; Liu, M.; Salzmann, M.; Li, H. Learning trajectory dependencies for human motion prediction. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4317–4326.