Learning Constrained Dynamic Correlations in Spatiotemporal Graphs for Motion Prediction

Jiajun Fu, Fuxing Yang, Yonghao Dang, Xiaoli Liu, and Jianqin Yin, Member, IEEE

Abstract—Human motion prediction is a challenging task that involves modeling dynamic spatiotemporal correlations between body joints. This task is becoming increasingly important in autonomous driving [1], human–machine interaction [2], [3], [4], and healthcare [5].

Traditional human motion prediction methods can handle simple and periodic scenarios with classical time series processing approaches [6], [7], but these methods make unrealistic predictions when human motion becomes complex and erratic. Recently, researchers have adopted different deep-learning techniques for human motion prediction, such as convolution neural network (CNN) [8], [9], recurrent neural network (RNN) [8], [10], [11], [12], [12], [13], [14], [15], [16], [17], [18], generative adversarial network (GAN) [19], [20], [21], [22], [23] and transformer [24], [25], [26]. Though these methods have achieved significant advancements over traditional approaches, they cannot explicitly represent connections among body joints, and thus, fail to model inherent body joint relations effectively.

With joints as vertices, bone, and trajectory connections as edges, a human motion sequence can be naturally represented as a spatiotemporal graph. Graph convolution networks (GCNs) have been applied to motion prediction [27], [28], [29], [30], [31], [32], [33], [34], [35]. Within the GCN, the spatiotemporal correlation guides the feature aggregation, and thus, is the key to achieving high predictability. For the spatiotemporal correlation, flexible and various motion sequences are represented by two crucial characteristics. The first one is spatiotemporal-unshared decomposition [Fig. 1(a)]. Since a joint is physically connected to its kinetic neighbors and virtually connected to its trajectory neighbors [33], [36], we can decompose a spatiotemporal motion graph into several spatial and temporal graphs. The spatial graphs model intrabody relations for each frame, and the temporal graphs model interframe relations for each joint. With this decomposition, spatiotemporal feature modeling is conducted by alternatively updating features on spatial and temporal graphs. This strategy reduces the complexity of spatiotemporal graph adjacency representation and has been widely adopted. For the spatiotemporal-unshared decomposition, the spatial graph correlations are varied at each frame to reflect changing motion patterns, while the temporal graph correlations are different at each joint to depict distinct trajectory patterns. We refer to these correlations as unshared spatial and temporal correlations. This precise representation contributes to predictive feature extraction. The second characteristic is sample-specific representation, where spatiotemporal graph adjacency is adjusted for different motion sequences. Since the motion patterns are varied for different motion sequences, the spatiotemporal

I. INTRODUCTION

HUMAN motion prediction is a challenging task that involves modeling dynamic spatiotemporal correlations

Manuscript received 12 September 2022; revised 13 February 2023; accepted 15 May 2023. Date of publication 31 May 2023; date of current version 8 October 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62173045 and Grant 62273054, in part by the Fundamental Research Funds for the Central Universities under Grant 2020XD-A04-3, and in part by the Natural Science Foundation of Hainan Province under Grant 622RC075. (Corresponding author: Jianqin Yin.)

The authors are with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: jjqin@bupt.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNNLS.2023.3277477, provided by the authors.

Digital Object Identifier 10.1109/TNNLS.2023.3277476
Fig. 1. (a) Spatiotemporal-unshared decomposition. A spatiotemporal motion graph with $T$ frames and $J$ joints can be decomposed into $J$ spatial graphs and $T$ temporal graphs. The spatial/temporal correlations are different from each frame/joint. The skeleton color brightness in spatial graphs indicates the corresponding frame and the joint color in temporal graphs indicates the corresponding joint. (b) Dynamic correlation modeling. We use the correlation for the red knee joint as an example. Correlations vary for different actions, and the correlations in walking (I) are different from those in directing traffic (II). These variations cannot be explicitly captured by the learned correlations from global optimization (III), thus introducing correlation amplification or reduction (the dashed line).

correlation should precisely reflect sample-specific patterns in the spatiotemporal graph.

For spatiotemporal-unshared decomposition, Sofianos et al. [33] proposed the only spatiotemporal-unshared method, where crosstalk between unshared spatial and temporal correlations depict the connections in a spatiotemporal graph. This method significantly improves the representation ability of GCs and outperforms previous methods with far fewer parameters. However, it still introduces redundant parameters for representing common prior knowledge, such as body connections and temporal context. There is still room to explore more parameter-saving representations of spatiotemporal correlations. For sample-specific representations, despite the plausible motive behind this design, previous works all [27], [28], [29], [30], [31], [32], [33] adopted generic graph correlations across input sequences. These sample-generic representations are learned through backpropagation and are optimal for all data, but they may disturb individual predictions by mistakenly enhancing unrelated correspondences or reducing critical connection strengths [Fig. 1(b)]. The correspondence disturbance amplifies the slight movement or the inability to infer future motion from tiny cues. Therefore, it is urgent to develop methods with sample-specific spatiotemporal correspondence.

To overcome the aforementioned challenges, we look into human motions’ constrained dynamic relational nature, a critical characteristic ignored by all previous GCN-based methods. For every sample, the relations between body joints are subject to common static constraints such as body connections and temporal context (e.g., a frame connects to its adjacent one). Meanwhile, the relation strengths between certain joints are dynamically enhanced or reduced according to different samples.

Inspired by the above-mentioned discussions, we introduce dynamic spatiotemporal decompose GC (DSTD-GC), which contains dynamic spatial GC (DS-GC) and dynamic temporal GC (DT-GC) for constrained dynamic spatial and temporal correlation modeling. The DS-GC/DT-GC is equipped with constrained dynamic correlation modeling, which decouples correlations into static constraints and dynamic adjustments. The static-constrained correlation is parameterized as a vanilla spatial/temporal adjacency matrix, and it is shared by all frames/joints across all samples. This sample-generic shared correlation serves as motion prior and indicates general relations between vertices, and it is completed by sample-specific unshared correspondence adjustments. These dynamic adjustments are extracted by an adjustment modeling function based on the input sample. According to this formulation, we represent spatiotemporal-unshared correlations in quarter space complexity, while the previous representation [33] is of cubic space complexity. Besides, DSTD-GC captures dynamic motion patterns between different samples and adjusts correlations between graph vertices, which precisely enhances correspondences for mutual motions and reduces correspondences for unrelated motions. Furthermore, we analyze different types of spatiotemporal GCs in a unified form and theoretically prove that DSTD-GC enhances feature representation by relaxing certain restrictions of spatiotemporal GCs. Finally, by combining DSTD-GC and prior knowledge, such as body connections and temporal context, we propose DSTD-GCN, which outperforms other state-of-the-art methods in prediction accuracy with the fewest parameters.

The contributions of this article are summarized as follows.

1) DSTD-GC is proposed, which utilizes constrained dynamic correlation modeling to represent spatiotemporal-unshared sample-specific correlations in the spatiotemporal graph, thus achieving lightweight motion pattern representation and accurate motion prediction.

2) By mathematically unifying all forms of GCs in human motion prediction, we show that DSTD-GCN improves the representation ability by overcoming the constraints of other GCs.

3) Extensive experiments show that our proposed DSTD-GCN outperforms other state-of-the-art methods in terms of prediction errors and parameter numbers on three benchmark datasets.

II. RELATED WORK

A. Human Motion Prediction

Researchers have applied various deep-learning techniques to human motion prediction. CNN-based methods mainly treat a human pose as a pseudoimage [8], [9]. For example, Liu et al. [9] stacked pose sequences along the channel dimension and extracted multilevel motion features in different CNN layers. RNN-based methods show their capability in modeling temporal motion features and make consistent predictions [8], [12], [13], [14], [15], [16], [17], [18], [29], [32], [37]. Martinez et al. [12] proposed a simple yet efficient Seq2Seq model and made predictions based on residual velocities. Wang et al. [18] extended this framework by introducing joint velocity and frame position. GAN-based methods make multiple future predictions based on data pattern similarities and produce realistic results [19], [30], [38], [39]. Lyu et al. [39] utilized stochastic differential equations and path integrals. Transformer-based methods directly model long-range spatial and temporal dependencies [24], [25]. Aksan et al. [25] designed spatial and temporal transformers to update spatiotemporal
representations simultaneously. Although these methods make remarkable advancements compared with traditional methods, they cannot directly model the natural connectivity between body joints. Natural connectivity is crucial for human motion prediction because human motion follows kinetic chains and trajectory paths, which are constrained by body connections and temporal context. To explicitly model body connections, researchers have turned to GCNs. GCs are suitable for processing nongrid and graph-structured data and have been successfully applied to social networks [40], point clouds [41], and traffic prediction [42]. By taking joints as vertices and bones and trajectory connections as edges, a human pose can be naturally represented as a spatiotemporal graph. Recently, many researchers have applied GCs to motion prediction [27], [28], [30], [31], [32], [33], [43]. In this article, we will only focus on validating our proposed method under conventional offline prediction settings, where the inputs and outputs are all motion sequences. With the success of deep learning in other studies, some work concentrated on extending the prediction settings [26], [34], [35]. For instance, Dong and Xu [35] introduced motion classification as an auxiliary task. Li et al. [34] looked into online prediction settings. Tevet et al. [26] combined natural language information with the transformer model.

B. Graph Convolution Networks in Motion Prediction

Nearly all GCN-based prediction approaches are developed based on the work of Kipf and Welling [44], where features are updated in two steps: 1) feature transformation with a simple linear transformation or a multilayer perceptron and 2) feature aggregation with graph correlations. The graph correlation explicitly depicts connections between body joints, so it is a critical component that distinguishes GCN from other deep learning techniques. Based on the spatiotemporal graph correlation, there are two classifications: 1) spatiotemporal-shared/spatiotemporal-unshared, which is defined for decomposed spatiotemporal representations and depends on whether the graph correlations are shared by joints in the temporal-unshared spatial graph or the spatial-unshared temporal graph. 2) sample-generic/sample-specific, which depends on whether graph correlations are adjusted dynamically to different samples.

1) Spatiotemporal-Shared/Spatiotemporal-Unshared Methods: For spatiotemporal-shared methods, the spatial/temporal correlations are shared across frames/joints. Most of the previous studies [27], [28], [30], [31], [32], [35], [43] belong to this type. A limitation of these methods is that they cannot explicitly model changing spatial relations in diverse motion stages and diverse temporal patterns in different joints. Thus, they generally need to stack many GC layers to model complex spatiotemporal features in human motion. For spatiotemporal-unshared methods, the spatial/temporal correlations vary from time/joints. Thus, these methods can explicitly depict joint relations in spatiotemporal graphs and precisely model complex spatiotemporal correlations in human motion with far fewer parameters [33].

2) Sample-Generic/Sample-Specific Methods: For sample-generic methods, graph correlations keep unchanged for all samples. Almost all previous studies applied this strategy. Mao et al. [27] and Dang et al. [31] directly learned graph correlations from data, while Li et al. [29] initialized the trainable correlation as a predefined graph. Cui et al. [30] and Liu et al. [32] combined the previous two correlations and proposed a semiconstrained graph correlation. For sample-specific methods, graph correlations can be adjusted to each sample. To the best of our knowledge, there is no application of sample-specific GCs to motion prediction. We are the first to model sample-specific spatiotemporal-unshared GCs in motion prediction.

III. METHODOLOGY

The proposed DSTD-GC is introduced in this section. First, human motion prediction and some graph-related definitions are presented. Then, all GCs in motion prediction are summarized in a unified form. Next, constrained dynamic correlation modeling and dynamic spatiotemporal GC are described. Finally, the detailed model architecture is introduced.

A. Preliminaries

1) Problem Definition: Human motion prediction is to predict $L$ future human pose frames with $K$ historical observations. The entire sequence length is $T = K + L$. As a human pose is represented by $J$ joints with $D$-dimensional spatial information, the human pose at time $t$ is denoted as $X_t \in \mathbb{R}^{J \times D}$. The historical observations are formulated as $X_{1:K} = \{X_{1}, \ldots, X_{K}\}$. Our goal is to predict the future human motion $\hat{X}_{K+1:K+L} = \{\hat{X}_{K+1}, \ldots, \hat{X}_{K+L}\}$, where the corresponding ground truth is denoted as $X_{K+1:K+L} = \{X_{K+1}, \ldots, X_{K+L}\}$.

2) Notations: The spatiotemporal graph and its decomposition are formally defined as follows.

a) Spatiotemporal graph representation: As shown in Fig. 1(a), a human motion sequence with $T$ frames is presented as a spatiotemporal graph $G^{st} = (V^{st}, E^{st})$, where $V^{st} \in \mathbb{R}^{J \times T}$ is the set of all joints across frames, and $E^{st}$ is the set of spatiotemporal edges. $A^{st} \in \mathbb{R}^{J \times T \times J \times T}$ is the spatiotemporal adjacency matrix, which represents the correlations between the vertices in the graph. However, it takes many parameters to directly store and learn spatiotemporal correspondences with space complexity of $O((JT)^2)$. Therefore, we further decompose a spatiotemporal graph into a unique pair of $T$ spatial graphs across frames and $J$ temporal graphs across joints. The spatial graphs are denoted as $G^{s} = (V^{s}, E^{s})$ where $V^{s} \in \mathbb{R}^{T \times J}$ is the set of joint vertices. $E^{s}$ is the set of spatial edges, and it is formulated as a temporal-unshared spatial adjacency matrix$^1$ $A^{s} \in \mathbb{R}^{T \times J \times J}$ when spatial correlations are shared across frames, the spatial adjacency matrix is degraded into a vanilla form and is represented by a matrix from $\mathbb{R}^{J \times J}$. Similarly, the temporal graphs are denoted as $G^{t} = (V^{t}, E^{t})$, where $V^{t} \in \mathbb{R}^{J \times T}$ is the set of trajectory vertices. $E^{t}$ is the set of temporal edges, and it is formulated as a spatial-unshared temporal adjacency matrix $A^{t} \in \mathbb{R}^{J \times T \times T}$.

$^1$For clarity, the third-order tensor is denoted as a matrix, and the second-order tensor is denoted as a vanilla matrix.
b) Spatiotemporal-equivalence: Spatiotemporal-equivalence is defined for two functions on spatial graphs and corresponding temporal graphs. Formally, function $F^s$ on $G^s$ and function $F^t$ on the corresponding $G^t$ are spatiotemporal-equivalent if and only if $F^s$ is equivalent to $F^t$ after switching all operations for the time and space dimensions, and vice versa. The GCs with spatiotemporal-equivalent operations are called spatiotemporal-equivalent GCs. In the following, we only introduce the function on spatial graphs for a spatiotemporal-equivalence pair.

3) Spatiotemporal Decompose Graph Convolutions: We introduce different GCs by taking the example of the feature updating process for joint $q$ at frame $n$. The spatial and temporal neighbor indices of the selected joint are denoted as $p$ and $m$, respectively. For a typical spatiotemporal GC (ST-GC), the output feature $y_{qn}$ is obtained by

$$y_{qn} = \sum_p \sum_m a^s_{(pm)(qn)} x_{pm} W$$

where $x_{pm} \in \mathbb{R}^C$ is the input feature from joint $p$ of frame $m$, $y_{qn} \in \mathbb{R}^C$ is the output feature, $a^s_{(pm)(qn)}$ is a correlation strength in the spatiotemporal adjacency matrix $A^s \in \mathbb{R}^{J \times J}$, and $W \in \mathbb{R}^{C \times C}$ is a trainable parameter for feature transformation.

With the graph decomposition introduced earlier, two graph GCs are defined: spatial GC (S-GC) and temporal GC (T-GC)

$$x^s_{qn} = \sum_p a^s_{npq} x_{pn} W_1$$

$$x^t_{qn} = \sum_m a^t_{qmn} x_{qm} W_2$$

where $W_1$ and $W_2$ are feature transformation matrices from $\mathbb{R}^{C \times C}$ and $\mathbb{R}^{C \times C}$.

Specifically, the S-GC models framewise spatial correlations, while T-GC models jointwise temporal correlations. Then, we propose STD-GC by alternatively stacking an S-GC (2) and a T-GC (3)

$$y_{qn} = \sum_m a^t_{qmn} \left( \sum_p a^s_{npq} x_{pm} W_1 \right) W_2$$

where $a^s_{npq}$ and $a^t_{qmn}$ are from a spatial adjacency matrix $A^s$ and a temporal adjacency matrix $A^t$. STD-GC decomposes ST-GC on a spatiotemporal graph into S-GC on spatial graphs and T-GC on temporal graphs. Then, spatiotemporal feature modeling is performed by stacking these two GCs. The capability of spatiotemporal modeling is not affected by the stacking order of S-GC and T-GC. Meanwhile, STD-GC is equivalent to space-time-separable (STS)-GC [33]. All of these will be demonstrated with experiments in Section IV-C. Furthermore, STD-GC becomes vanilla STD-GC (VSTD-GC) when the spatial and temporal adjacency matrices are in vanilla form.

In sample-generic methods, $a^s_{npq}$ and $a^t_{qmn}$ are generated by a model according to the input features.

B. Analysis of Graph Convolutions on Spatiotemporal Graphs

By reformulating several GCs on spatiotemporal graphs into a uniform form, we evaluate their representation capabilities in human motion prediction. First, the spatiotemporal graph decomposition constraint is introduced, and then several GCs are evaluated based on their spatial and temporal modeling capability. We focus exclusively on the correlation between distinct GCs because the correlation dominates feature aggregation and is critical to obtaining representative spatiotemporal features. Based on the formulation in Section III-A3, we add (i) in the superscript to indicate different samples.

By comparing ST-GC (1) and STD-GC (4), it can be seen that the spatiotemporal correlations can be represented by a combination of spatial and temporal correlations with the following constraints.

Constraint 1: $a^{s(i)}_{(pm)(qn)}$ equals to the product of $a^{s(i)}_{npq}$ and $a^{t(i)}_{(pm)(qn)}$, where $a^{t(i)}_{(pm)(qn)}$ is the spatiotemporal correlation strength, $a^{s(i)}_{npq}$ is the corresponding spatial correlation strength, and $a^{t(i)}_{qmn}$ is the corresponding temporal correlation strength.

With the decomposition constraint, we further reformulate decomposed ST-GC from the spatial and temporal graph perspectives. These two views are presented as follows:

$$y^{(i)}_{qn} = \sum_m a^{s(i)}_{npq} x^{(i)}_{pm} W$$

$$y^{(i)}_{qn} = \sum_p a^{t(i)}_{qmn} \sum_m x^{(i)}_{pm} W$$

where $a^{s(i)}_{npq}$ and $a^{t(i)}_{qmn}$ are the spatial and temporal correlation weights for the input sample, respectively. Meanwhile, we combine correlation strengths and feature transformation weights into a generalized weight matrix $a^{s(i)}_{npq} W(a^{t(i)}_{npq} W)$.

Thus, the feature updating process consists of feature transformation with the generalized weight matrix and feature aggregation with framewise spatial or jointwise temporal correlations. With these two formulations, we analyze different graph GCs based on the spatial and temporal correlations for feature aggregation.

1) Sample-Generic Spatiotemporal-Shared GCs: Most recent studies [27], [28], [30], [31], [36], [43] adopted spatiotemporal-shared GCs, where joint vertices from all frames share one spatial correlation and trajectory vertices for all joints share one temporal correlation. The whole GC process can be formulated as

$$y^{(i)}_{qn} = \sum_m a^{s(i)}_{npq} a^{s(i)}_{mn} x^{(i)}_{pm} W$$

$$y^{(i)}_{qn} = \sum_p a^{t(i)}_{qmn} \sum_m x^{(i)}_{pm} W$$

$$y^{(i)}_{qn} = \sum_p a^{t(i)}_{qmn} \sum_m x^{(i)}_{pm} (a^{t(i)}_{npq} W)$$
where \( a'_{pq} \) is the shared spatial correlation strength and \( a'_{mn} \) is the shared temporal correlation strength. They are from the vanilla form of the spatial adjacency matrix \( A' \) and the temporal adjacency matrix \( A'' \), respectively. Equations (8) and (9) are the reformulations of (7) in the spatial graph’s view and temporal graph’s view. Besides, these two correlation strengths keep unchanged across input samples. Therefore, the sample-generic spatiotemporal-shared GCs are subject to the following constraints.

**Constraint 2:** \( a'_{n1p} \) and \( a'_{n2p} \) are forced to be the same, and \( a'_{qmn} \) and \( a'_{qnm} \) are forced to be the same.

**Constraint 3:** \( a'_{n1p} \) and \( a'_{n2p} \) are forced to be the same, and \( a'_{qmn} \) and \( a'_{qnm} \) are forced to be the same.

Note that \( i_1 \) and \( i_2 \) are indices of different input samples, \( n_1 \) and \( n_2 \) are indices of different joints, and \( q_1 \) and \( q_2 \) are indices of different frames.

2) **Sample-Generic Spatiotemporal-Unshared GCs:** The only difference between sample-generic spatiotemporal-shared GCs and sample-generic spatiotemporal-unshared GCs is that the spatiotemporal-unshared ones adopt unshared spatial and temporal correlations.

The sample-spatiotemporal-unshared GCs can be formulated as

\[
\begin{align*}
\mathbf{y}_{qn}^{(i)} &= \sum_{p}^{J} \sum_{m}^{T} a'_{npq} a'_{qmn} x_{pnm} \mathbf{W} \\
\mathbf{y}_{qn}^{(i)} &= \sum_{p}^{J} \sum_{m}^{T} a'_{npq} \sum_{j}^{T} x_{pmj} a'_{qmn} \mathbf{W} \\
\mathbf{y}_{qn}^{(i)} &= \sum_{p}^{J} \sum_{m}^{T} x_{pmj} (a'_{npq} \mathbf{W})
\end{align*}
\]

where \( a'_{npq} \) is framewise spatial correlation strength and \( a'_{qmn} \) is jointwise temporal correlation strength. They are from the spatial adjacency matrix and the temporal adjacency matrix, respectively. With the two unshared correlations, sample-generic spatiotemporal-unshared GCs capture evolving spatial relations between joints at different action stages and diverse temporal relations for different body parts. Thus, they generally outperform spatiotemporal-shared GCs. However, these correlations are optimized across all data and may not be optimized for individual action sequences. From the formula’s view, sample-generic spatiotemporal-unshared GCs still suffer from **Constraint 3** but relax **Constraint 2** into the following constraint.

**Constraint 4:** \( a'_{n1p} \) and \( a'_{n2p} \) differ by a scaling factor, and \( a'_{qmn} \) and \( a'_{qnm} \) differ by a scaling factor.

3) **Sample-Specific Spatiotemporal-Unshared GCs:** In comparison with sample-generic spatiotemporal-unshared GCs, the sample-specific ones infer dynamic correlations between vertices, and thus, have higher representation capacity. The formulation for these GCs is as follows:

\[
\begin{align*}
\mathbf{y}_{qn}^{(i)} &= \sum_{p}^{J} \sum_{m}^{T} a'_{npq} a'_{qmn} x_{pm} \mathbf{W} \\
\mathbf{y}_{qn}^{(i)} &= \sum_{p}^{J} \sum_{m}^{T} x_{pm} (a'_{npq} \mathbf{W})
\end{align*}
\]

where both \( a'_{npq} \) and \( a'_{qmn} \) are adjustable to the \( i \)-th input sample. Based on these formulations, sample-specific spatiotemporal-unshared GCs relax both **Constraint 2** and **Constraint 3**. Specifically, they relax **Constraint 2** into **Constraint 4** and relax **Constraint 3** into the following constraint.

**Constraint 5:** \( a'_{n1p} \) and \( a'_{n2p} \) differ by a scaling factor, and \( a'_{qmn} \) and \( a'_{qnm} \) differ by a scaling factor.

From the above-mentioned analysis, it can be seen that sample-specific spatiotemporal-unshared GCs are the least constrained. They are also the theoretically strongest decomposed GCs for spatiotemporal feature representation. This will be demonstrated in Section IV-C by comparing different GCs.

### C. Dynamic SpatioTemporal Decompose Graph Convolutions

In the following, our proposed DSTD-GC is introduced, which extends STD-GC with DS-GCs and its spatiotemporal-equivalent counterpart DT-GCs. The main difference between DSTD-GC and STD-GC is that the former GC equips constrained dynamic correlation modeling, which is the key component for powerful and lightweight representations. In this section, constrained dynamic correlation modeling is introduced, and then the details of DS-GC are presented.

1) **Constrained Dynamic Correlation Modeling:** The motivation behind constrained dynamic correlation modeling is that the correlations between joints are both constrained and dynamic. Specifically, the correlations between joints are constrained by the inherent human body structure, consistent trajectory connections, and learned motion prior. Meanwhile, these correlations are dynamically changed based on different motion sequences. To show these two characteristics, we utilize the correlations of the elbow and knee joints in the walking and directing traffic scenarios as an example (Fig. 2). The elbow and knee movements generally follow the motion chain of the human body and common trajectory patterns, so their spatial/temporal correlation presents static constraints, which are shared by all frames/joints and stay fixed under these two scenarios. Meanwhile, compared with the dynamic spatial and temporal adjustments in the directing traffic scenario, the spatial adjustment strengths between the elbow joint and the knee joint and temporal adjustment strengths in the knee joint...
are generally stronger in the walking scenario. These dynamic adjustments refine static spatial/temporal constraints at each frame/joint and vary from different inputs.

Inspired by this, we present constrained dynamic correlation modeling to directly represent the aforementioned static constraints and dynamic adjustments. Taking the spatial modeling from the purple box in Fig. 3(a) as an example, the static constraints are parameterized as a shared adjacency matrix, while the dynamic adjustments are extracted by the dynamic spatial correlation adjustment modeling function. This function takes the spatiotemporal motion feature as inputs and outputs framewise spatial correspondence adjustments that update static constraints for framewise feature aggregation.

2) Dynamic Spatial Graph Convolutions: The general framework of DS-GC is illustrated in Fig. 3(b). Specifically, DS-GC consists of three parts: 1) feature transformation; 2) spatial constrained dynamic correlation modeling; and 3) framewise feature aggregation. DS-GC takes spatiotemporal motion feature $X \in \mathbb{R}^{J \times T \times C}$ and constrained spatial adjacency matrix $A^s \in \mathbb{R}^{J \times J}$ as inputs, and then it outputs $Y \in \mathbb{R}^{J \times T \times C}$. This section will only introduce the functions in DS-GC, and the implementations will be presented in Section III-D.

a) Feature transformation: As shown in the yellow block of Fig. 3(b), feature transformation is accomplished by transformation function $T$. Here, we set the function as a linear transformation function for clarity, while other functions, such as a multilayer perceptron, can also be used. The function is defined as

$$F' = T(X) \triangleq XW$$

where $F' \in \mathbb{R}^{J \times T \times C}$ is the high-level representation.

b) Spatial constrained dynamic correlation modeling: The spatial constrained dynamic correlation modeling is shown in the purple part of Fig. 3(b). Here, the static constraints are parameterized as a vanilla adjacency matrix $C^s \in \mathbb{R}^{J \times J}$, which is set as a model parameter and optimized across all data. Meanwhile, the dynamic correspondence adjustments $M' \in \mathbb{R}^{T \times J \times J}$ are inferred from the input sample to capture framewise spatial correspondence adjustments. Specifically, $M'$ is extracted by the dynamic spatial correspondence adjustment modeling function $M^d$, which traverses all pairs of joints and extracts mutual relations. The formulation of this function for a joint pair $(p, q)$ is presented in the following:

$$M^d = M^d(x_p, x_q) \triangleq \text{MLP}(\theta(x_p) || \phi(x_q))$$

where $(x_p, x_q)$ from the input feature $X$ of the joint pair $(p, q)$, and $||$ is the pairwise concatenation along the joint dimension. $\theta$ and $\phi$ are two linear transformation functions that project the graph vertex features into low-dimensional representations to reduce computational cost. Multi-layer perceptron (MLP) is a multilayer perceptron, we utilize MLP because it can model complex spatial correlations. Note that $M^d$ is not required to be symmetric, which benefits correlation modeling by increasing analysis flexibility and enhancing representative ability. Finally, the output spatial constrained dynamic correlation, $A^s$, is obtained by a correlation updating function $U^t$. This function adjusts the static constraints $C^s$ with dynamic adjustments $M^d$.

$$A^s = U^t(A^s, M^d) \triangleq C^s + \alpha \cdot M^d$$

where $\alpha$ is a learnable parameter for controlling the adjustment intensity.

c) Framewise feature aggregation: As shown in the pink block of Fig. 3(b), given the spatial-constrained dynamic correlations $A^s$ and high-level features $F'$, the output is obtained with the framewise aggregation function $A^f$. Specifically, $A^f$ constructs a spatial graph along the temporal dimension for the updated correlation $A^s$ and feature $F'$, where $A^s_t$ and $F^t_{:, :, :}$ are from the $t$th frame of $A^s$ and $F'$, respectively. The output feature $Y \in \mathbb{R}^{J \times T \times C}$ is obtained by

$$Y = A^f(F', M^d) \triangleq [F^t_{:, :, :}, M^t_{:, :, :}, \ldots, M^T_{:, :, :}]$$

where $||$ is a concatenation function along the frame dimension.

Based on the above-mentioned three steps, DS-GC is formulated as

$$Y = A^f(T(X), U^t(A^s, M^d(X)))$$.
Since DT-GC is spatiotemporal-equivalent to DS-GC, DT-GC can be directly obtained by replacing all operations in DS-GC with their spatiotemporal-equivalent counterparts

\[ Y = A'(T(X), U'(C', M'(X))). \]  

Note that in DS-GC, the spatial-constrained dynamic correlations are sample-specific and unshared in (17). Similarly, the temporally constrained dynamic correlations from DT-GC are sample-specific and unshared. Therefore, the DSTD-GC, which combines DS-GC and DT-GC for spatiotemporal modeling, belongs to the sample-specific spatiotemporal-unshared GC. The detailed derivation is presented in Appendix A.

3) Discussions: Fig. 3(a) shows the comparison between STS-GC/S-GC and DS-GC in spatial modeling. It is found that DSTD-GC has three advantages: 1) parameter-saving spatiotemporal-unshared representation— to accomplish unshared spatial correlation modeling, STS-GC/S-GC directly parameterizes the correspondence as a spatial adjacency matrix \( A_s \) in the GC, where the space complexity is \( O(JT^2) \). Conversely, DS-GC accomplishes unshared spatial modeling with space complexity of \( O(J^2) \) by learning a vanilla adjacency matrix and a function with negligible parameters. Similarly, the space complexity of spatial-unshared temporal correlation modeling is \( O(J^2T) \) for STS-GC/T-GC and \( O(T^2) \) for DT-GC. As the joint number is always proportional to the sequence length, the overall space complexity for STS-GC and STD-GC is \( O(T^3) \) and that for DSTD-GC is only \( O(T^2) \). 2) Joint optimization—DSTD-GC learns common motion patterns in a vanilla matrix and a function to extract framewise/jointwise correlation variations. With fewer parameters to learn, this mechanism has less optimization difficulty than directly learning optimal spatial/temporal correlations through an adjacency matrix. 3) Sample-specific modeling—DSTD-GC models dynamic spatiotemporal correspondence adjustments in different motion inputs. These adjustments reflect sample-specific motion patterns and contribute to accurate motion predictions. In addition to these advantages, our DSTD-GC introduces little computation overhead. The analysis of the time complexity of DSTD-GC and STS-GC is presented in Appendix B. Both GCs have the same time complexity of \( O(CT^2(C + T)) \).

D. Model Architecture

Based on DSTD-GC, we propose a lightweight and powerful model called DSTD-GCN, which is built upon the prediction framework proposed by [27]. In this section, the prediction framework is introduced first, and then the implementation details of DSTD-GCN are given.

1) Prediction Framework: We duplicate the last input human pose for \( L \) times and formulate a new input motion sequence \( X_{1:L} \). Then, the original prediction task is reformulated into predicting the residual motion sequence between \( X_{1:L} \) and the corresponding ground truth \( \tilde{X}_{1:L} \). Based on the problem reformulation, the prediction framework is proposed, as shown in Fig. 4. The framework contains a basic conv unit to encode motion input to high-level representations, \( n_b \) basic blocks for spatiotemporal representation modeling, and a basic conv unit to decode predictions from motion features. The basic conv unit is a single-convolution layer like CNN and ST-GC. It stacks \( n_b \) layers to constitute a basic block. Besides, the global residual connection enforces the residual learning, while the local residual connection stabilizes the training process. Recent studies [27], [30], [31] prove the effectiveness of this framework. With this framework, the input and output human poses can be modeled as a whole sequence, and multilevel associated spatiotemporal correlations can be modeled, while the previous frameworks [29], [33] restrict high-level feature interactions between the input and output feature. This framework will be used to build our model and compare different GCs.

2) Model Implementation: The entire network adopts the prediction framework and stacks five basic blocks. Each basic block contains one basic conv unit. The basic conv unit of DSTD-GCN is shown in Fig. 5(a), which contains two parallel DS-GCs for spatial modeling and a DT-GC for temporal modeling. Specifically, for DS-GC, the modeling function \( M' \) takes \( X \in \mathbb{R}^{J \times T \times C} \) as input. Two linear transformation functions \( \theta \) and \( \phi \) with a reduction rate of \( r \) are used to
transform $X$ into a neatly compact representation and then squeeze the temporal dimension into the channel dimension. After this, a pairwise concatenation and a multilayer perceptron are adopted to extract spatial adjustments for each frame based on (17). These dynamic spatial adjustments update the static constraints $C'$ to obtain the sample-specific framewise spatial correlation. Finally, batch matrix multiplication is used for framewise feature aggregation to obtain the output $Y$.

We further consider prior knowledge for the static-constrained correlation. As Cui et al. [30] pointed out that a restricted prior graph correlation will reduce flexibility, we consider a simple solution here. We adopt a trainable constrained adjacency matrix and initialize it according to prior knowledge, where the weight is set to 1 for correlated joint pairs and 0 for other pairs. For spatial correspondence, we consider two types of joint relations. It is assumed that these two relations have independent influence and should be modeled separately. The first relation is the body connection, where a joint is correlated with its physically connected neighbors. The second relation is the semantic correlation. Since the arms and legs have a large variability and are hard to predict, we connect vertices in the same arm/leg. Meanwhile, body symmetry is considered, and the arm or leg is connected to its mirror counterpart. For temporal constraints, we only consider temporal context and connect every frame to its adjacent one.

IV. EXPERIMENTS

To evaluate the effectiveness of our proposed model, experiments are conducted on three standard benchmark motion capture datasets, including Human3.6M [45], Carnegie Mellon University (CMU) Mocap, and 3D Poses in the Wild (3DPW) [46] datasets. Meanwhile, experiments are designed to answer the following questions: 1) what is the performance of DSTD-GC as compared with other GCs? 2) What is the performance of our proposed DSTD-GCN as compared with other state-of-art motion prediction methods? 3) What is the influence of different components and what findings can be obtained from the experimental results?

A. Dataset Settings

1) Human3.6M: Human3.6M [45] is a widespread benchmark dataset for evaluating motion prediction. It contains 15 actions performed by seven actors. Following the data preprocessing procedure of [12] and [27], the original data in the exponential mapping format is transformed into 3-D joint coordinate space and a single pose is represented by 22 body joints. Then, each data sequence is down-sampled to 25 FPS. The data from S5 and S11 are used as test and validation datasets, while the remaining five subjects are used for training.

2) CMU Mocap: CMU Mocap $^3$ is another well-known benchmark dataset. Each pose is represented as 25 body joints. The preprocessing procedure is the same as that of Human3.6M, and the dataset split in [27] and [31] is adopted.

$^3$http://mocap.cs.cmu.edu/. We only report results on eight actions due to page limits. The results of “running” are presented in the Supplementary Material.

3) 3DPW: 3DPW [46] is a more challenging dataset and contains human motion sequences from indoor and outdoor scenarios. Like other baselines, a pose is represented as 23 body joints, and the official split is utilized [9], [27], [31].

B. Comparison Settings

1) Metrics: The mean per joint position error (MPJPE) is a standard evaluation metric used in previous studies [8], [12], [31]. With the predicted motion sequence $\hat{X}_{K+1}$ and corresponding ground truth $X_{K+1}$, the MPJPE loss is defined as

$$L_{MPJPE} = \frac{1}{J} \sum_{t=K+1}^{K+L} \sum_{j=1}^{J} \|\hat{p}_{j,t} - p_{j,t}\|_2$$

where $\hat{p}_{j,t} \in \mathbb{R}^D$ is the predicted $j$th joint position in frame $t$, while $p_{j,t} \in \mathbb{R}^D$ is the corresponding ground truth.

2) Implementation Details: All experiments were conducted on one RTX 3080 Ti GPU with the PyTorch framework [47]. For the state-of-the-art baseline models, their official implementations were adopted, and they were trained with default settings. For the comparison of DSTD-GCN and the baseline models for GCs, PReLU [48] was adopted as the activation function, and the channel number was set to 64. Meanwhile, all the models were trained with the Adam optimizer [49] with an initial learning rate of $3e-3$, which was decayed by 0.9 for every five epochs; MPJPE was used as the loss function, and the batch size was set to 32; the reduction rate $r$ was set to 32 in the DSTD-GC. Besides, the test set and sequence length settings followed the experiment setting of Dang et al. [31]. Moreover, the full test set was utilized, and the input length was set to 10, while the output length was set to 30 on the 3DPW dataset and 25 on the other two datasets. The detailed model architecture is provided in the Supplementary Material.

C. RQ1: Comparison With Other GCs

We compare the parameter number and prediction error of DSTD-GC against other GCs in Table I. Specifically, three key aspects of DSTD-GCN are investigated: spatiotemporal-unshared decomposition, sample-specific correlation, and stacking order invariance. For a fair comparison, the prediction framework is kept unchanged, and only the GC in the basic conv unit is changed. Moreover, the linear transformation is changed to a two-layer MLP for ST-GC,

| GCs | Params | Classification | Constraints | MPJPE |
|-----|--------|----------------|-------------|-------|
| ST  | 144M  | ✓              | ✓           | 90.7  |
| VSTD| 0.65M | X              | ✓           | 86.48|
| RC (27) | 0.20M | ✓              | ✓           | 73.11|
| STS (33) | 0.45M | ✓              | ✓           | 83.15|
| TSD | 0.46M | ✓              | ✓           | 78.84|
| STD | 0.46M | ✓              | ✓           | 81.14|
| DSTD | 0.13M | ✓              | ✓           | 72.31|
| DSTD | 0.13M | ✓              | ✓           | 73.56|

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
VST-GC, FC-GC, and STS-GC. For our DSTD-GC, the basic conv unit in Fig. 5(a) is adopted, and one DS-GC is removed in the spatial modeling. Also, TSD-GC and DTSD-GC are introduced by switching the order of spatial and temporal GCs in STD-GC and DSTD-GC, respectively. For a fair comparison, we follow previous approaches [27], [31], [33] and initialize adjacency matrices randomly for all GCs.

1) Spatiotemporal-Shared GCs Versus Spatiotemporal-Unshared GCs: Since spatiotemporal-shared and spatiotemporal-unshared GCs are defined for decomposing ST-GC, they are compared with the ST-GC. It can be observed that spatiotemporal-shared GCs (STS, TSD, and STD) have fewer parameters than spatiotemporal-unshared GCs but perform worse than ST-GC. Meanwhile, the spatiotemporal-unshared GCs obtain similar errors as ST-GC, which indicates that unshared spatial and temporal correlations are more suitable forms for describing spatiotemporal correlations because the unshared correlations can explicitly describe the varying spatiotemporal correlations in different body parts and different motion stages. Although spatiotemporal-shared GCs take fewer parameters than spatiotemporal-unshared ones, the corresponding models are not lightweight because these models need to stack more layers to achieve comparable performance. This will be shown in Section IV-D.4. Our DSTD-GC utilizes a shared vanilla adjacency matrix to store general correlations and a function to extract unshared correlation variances. Thus, our DSTD-GC is parametersaving and explicitly models spatiotemporal correlations.

2) Sample-Generic GCs Versus Sample-Specific GCs: We compare the two sample-generic GCs (TSGC and STD-GC) with their sample-specific counterparts (DTSD-GC and DSTD-GC). It can be found that sample-specific GCs make consistent improvements in accuracy. This is because the learned sample-generic correlations cannot precisely depict the correlations in individual sequences and the correlation balance between various input samples. This global optimal strategy may introduce false correspondence that disturbs individual predictions. More investigations on this are shown in Section IV-D.3.

3) Spatial–Temporal Order Versus Temporal–Spatial Order: As STD-GC and TSD-GC, DSTD-GC and DTSD-GC perform similarly, the modeling capability of spatiotemporal decomposed GC is invariant to the stacking order of spatial and temporal GCs.

D. RQ2: Comparison With the State-of-the-Art

To evaluate the performance of DSTD-GCN, the quantitative and qualitative results on the Human3.6M, CMU Mocap, and 3DPW datasets are presented. First, quantitative results are provided by comparing DSTD-GCN with state-of-the-art methods, and then qualitative results are provided. Following previous studies [27], [31], the results are divided into short-term (<500 ms) and long-term (>500 ms) predictions.

1) Baselines: We select six state-of-the-art baselines, including the RNN-based model (residual sup [12]), CNN-based model (Traj-CNN [9]), and GCN-based model (FC-GCN [27], DMGNN [29], MSR-GCN [31], and STS-GCN [33]). For GCN-based models, FC-GCN, DMGNN, and MSR-GCN belong to the sample-generic spatiotemporal-shared method, and STS-GCN is the only sample-generic spatiotemporal-unshared method. For the graph correlations, DMGNN adopts fixed correlations from body connections, while the other three GCN-based methods initialize the adjacency matrix randomly and then optimize it through backpropagation. Among all the baseline methods, FC-GCN, Traj-CNN, MSR-GCN, STS-GCN, and our approach utilize the prediction framework demonstrated in Fig. 5(a).

2) Results: The detailed quantitative comparisons for both short-term and long-term prediction results are presented in Tables II and III. Meanwhile, we also present the average results of selected frames in Tables IV and V. Apparently, methods with the prediction framework make more accurate predictions with fewer parameters. Meanwhile, GCN-based methods generally outperform the RNN-based method (residual sup.) and CNN-based methods (Traj-CNN). For GCN-based methods, spatiotemporal-unshared methods outperform spatiotemporal-shared methods in terms of prediction accuracy and parameter numbers. Note that MSR-GCN contains 68 GC layers and combines DCT temporal encoding [27] with multiscale modeling, while STS-GCN and DSTD-GCN only contain seven GC layers with raw 3-D inputs. With constrained dynamic correlation modeling, DSTD-GCN also models sample-specific correlations and achieves the best prediction accuracy with the fewest parameters. Even when the motion sequences become more challenging in the 3DPW dataset, our model still makes consistent improvements over the other approaches. For qualitative comparison, our method is compared with two state-of-the-art approaches on three benchmark datasets, and the examples are shown, as Figs. 6–8. In each of these examples, our method makes more accurate predictions. Specifically, we analyze the examples from the CMU Mocap dataset in detail. In the basketball scenario, spatiotemporal-unshared methods generally work better than spatiotemporal-shared ones because spatiotemporal-unshared
methods explicitly model individual joints’ spatiotemporal correlations, and thus, can infer a large movement from tiny motion cues. Here, both STS-GCN and our method can capture the backward trend of the right leg. However, sample-generic modeling of spatiotemporal correlations may disturb individual motion predictions, especially in some static action scenarios. As shown in the directing traffic scenario, most joints stay still, but STS-GCN amplifies the motion of the right arm and infers leg movement mistakenly. With constrained dynamic correlation modeling, our method can dynamically model sample-specific motion patterns and thus make more accurate predictions under different scenarios. Due to page limits, we present more visual comparison results in the Supplementary Material.

3) Visualization of Constrained Dynamic Correlations: In this section, we analyze different aspects of constrained modeling. The results are shown in the Supplementary Material.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Fig. 8. Visualization of the predicted results of the state-of-the-art methods on two action examples on the CMU Mocap dataset. The blue and red poses indicate ground truths and predictions, respectively. (a) Basketball. (b) Directing Traffic.

Fig. 9. Visualization of constrained correlation. (a) and (b) indicate spatial correlations and (c) indicates temporal correlations. (I) is the initialized mask, where blue colors indicate predefined connections. For spatial correlation, we also show the filtered points and their natural connections in black. Optimized constrained correlations are shown in (II), where color brightness indicates connection strength.

TABLE IV

| Model            | Human3.6M | CMU Mocap |
|------------------|-----------|-----------|
|                  | MPJPE     | Params    | Time     | MPJPE     | Params    | Time     |
| DMGNN [29]       | 63.27     | 46.90M    | 422ms    | 50.84     | 46.94M    | 590ms    |
| PC-GCN [27]      | 58.32     | 2.55M     | 46ms     | 40.30     | 2.70M     | 47ms     |
| Traj-CNN [9]     | 57.55     | 1.20M     | 95ms     | 38.62     | 1.20M     | 110ms    |
| STS-GCN [33]     | 57.47     | 0.40M     | 90ms     | 38.64     | 0.46M     | 41ms     |
| MSR-GCN [31]     | 57.88     | 6.30M     | 70ms     | 38.10     | 6.37M     | 82ms     |
| Ours             | 55.19     | 0.18M     | 48ms     | 35.93     | 0.20M     | 50ms     |

TABLE V

| Model            | 200  | 400  | 600  | 800  | 1000 | Average |
|------------------|------|------|------|------|------|---------|
|                  | MPJPE | Params | Time      | MPJPE | Params | Time      |
| DMGNN [29]       | 37.3 | 67.8 | 94.5 | 109.7 | 123.6 | 86.6 | 46.9M | 470ms |
| PC-GCN [27]      | 35.6 | 67.8 | 90.6 | 106.9 | 117.8 | 83.7 | 2.60M | 46ms  |
| Traj-CNN [9]     | 30.0 | 59.7 | 85.3 | 99.1 | 107.7 | 76.4 | 1.20M | 103ms |
| STS-GCN [33]     | 27.2 | 52.6 | 75.1 | 94.1 | 107.9 | 71.4 | 0.49M | 41ms  |
| MSR-GCN [31]     | 37.4 | 65.4 | 83.1 | 96.3 | 106.2 | 77.7 | 6.31M | 73ms  |
| Ours             | 24.5 | 49.9 | 69.9 | 85.2 | 96.3 | 65.2 | 0.19M | 49ms  |

Fig. 10. Visualization of spatial correlation adjustments. (a)–(c) are three sample poses (black, orange, and purple boxes in Fig. 8). (I)–(III) are the ground truth pose, the predicted pose with constrained correlation, and the prediction with constrained dynamic correlation. For (II) and (III), we show certain pairs of constrained dynamic correlations for clarity.

Fig. 11. Visualization of temporal correlation adjustments. (a) and (b) are from the right foot, while (c) is from the head. (a) and (c) are from “Basketball,” while (b) is from “directing traffic.” Color brightness indicates connection strength.

After the constrained correlation analysis, we use two example sequences in Fig. 8 to better understand why dynamic adjustments take effect. First, we remove dynamic adjustments in DSTD-GC and compare the visualization sequences (ours w/o dynamic adjustments) with the results from the full model. Then, we visualize the sample spatial and temporal adjustments from the two motion sequences in Figs. 10 and Fig. 11. Combining the example comparison with the correlation visualization, we obtain two observations: 1) sample-specific correspondence adjustments contribute to accurate motion prediction. As shown in Fig. 10(a), the adjustment strengths from the arms to the right leg are three strongest connections for each joint. We find that most optimized connections follow the predefined masks, which demonstrates the importance of the proposed prior knowledge initialization. Besides, some strong connections outside predefined masks intuitively contribute to accurate motion prediction, such as the hands to the legs and the torso and long-term frame links.

After the constrained correlation analysis, we use two example sequences in Fig. 8 to better understand why dynamic adjustments take effect. First, we remove dynamic adjustments in DSTD-GC and compare the visualization sequences (ours w/o dynamic adjustments) with the results from the full model. Then, we visualize the sample spatial and temporal adjustments from the two motion sequences in Figs. 10 and Fig. 11. Combining the example comparison with the correlation visualization, we obtain two observations: 1) sample-specific correspondence adjustments contribute to accurate motion prediction. As shown in Fig. 10(a), the adjustment strengths from the arms to the right leg are three strongest connections for each joint. We find that most optimized connections follow the predefined masks, which demonstrates the importance of the proposed prior knowledge initialization. Besides, some strong connections outside predefined masks intuitively contribute to accurate motion prediction, such as the hands to the legs and the torso and long-term frame links.
enhanced in “Basketball.” With these enhancement cues, the model predicts more accurate leg movement. Similarly, the connection strengths between the two arms are reduced in “directing traffic” in Fig. 10(b). Without these adjustments, the constrained correlation tends to infer wrong correspondences between the arms and amplify their movement mistakenly. Moreover, we also present temporal correspondence adjustments from the right ankle joint in Fig. 11(a) and (b), where we find the adjustment strengths are generally stronger in the basketball case. As the right foot moves backward in the basketball example, it correlates to other vertices of the trajectory more intensely. 2) Spatiotemporal-unshared correspondence adjustments reflect vital correlation variations in different movement stages/joints. For spatial modeling, the left arm moves in the late stage of the “Basketball” sequence. As shown in Fig. 11(a) and (c), the connection strengths between the left arm and the right leg are larger than those in the early movement stages. For temporal modeling, the knee joint moves continuously, so its temporal correlations are stronger than the ones of the head joint [Fig. 11(a) and (c)]. Although temporal adjustment strengths are much smaller than constrained temporal correlation, they introduce critical correlation refinements for accurate motion prediction. As shown in Table VI, the performance drops sharply when we retrain the model without temporal adjustments.

4) Effectiveness Analysis: The model parameters and inference time of a single iteration are presented in Tables IV and V. Generally, spatiotemporal-unshared methods outperform other methods in prediction accuracy and parameter number. For the models with the prediction framework, we also report the basic block number and number of convolutions in all basic blocks in Table VII. Although the spatiotemporal-shared GC has fewer parameters than spatiotemporal-unshared GC as a single-convolution unit, it needs to stack more layers to achieve comparable performance. The overall stacking overhead outweighs their lightweight benefits at an individual level. Meanwhile, our method outperforms other methods in prediction accuracy with the fewest parameters. The parameter number of our method is half of the most lightweight model STS-GCN and only 3% of the best baseline MSR-GCN on the CMU Mocap dataset. Besides, our method is slightly slower than STS-GCN due to the additional computation of correspondence adjustments.

E. RQ3: Ablation Study

In this section, some major components of our method are investigated by comparing DSTD-GCN to its variants and other state-of-the-art methods.

1) Constrained Dynamic Correlation Modeling: First, the performance of the constrained and dynamic correlation is investigated by only using the constrained vanilla adjacency matrix and dynamic adjacency matrix for feature aggregation, respectively. The results are shown as A and B in Table VIII, and two findings can be obtained by comparing the average errors: 1) constrained correlations cannot express erratic patterns in human motions, and there should be an adjustment function to complete the vertex correspondences for each motion sequence, as H outperforms A by 3.33. 2) With respect to the constrained dynamic correlation characteristic, explicit separate modeling is better than implicit united modeling, as variant B extracts both correlations simultaneously and performs worse than H by 6.84. Interestingly, A outperforms B by 3.51. Since the constrained correspondences are presented across all samples, the dynamic correspondences are presented for each individual, these correlations are optimized toward two distinct directions, and neither of them can reach the optimum under the united optimization scenario. Meanwhile, the updating order is investigated by only using the constrained feature to adjust the dynamic feature (C). C performs worse than A by 2.29, which indicates that dynamic correlations adjust constrained correlations but not vice versa.

2) Prior Constrained Adjacency: The improvement in average MPJPE is compared for prior constrained adjacency and constrained dynamic correlation modeling. To show the effectiveness of the prior constrained adjacency, we designed a model with random constrained adjacency initialization (D). For constrained dynamic correlation modeling, DS-GC and DT-GC are replaced by S-GC and T-GC, respectively (E). As shown in Table VIII, the reduction of prior connections (0.86) is much less than the reduction of constrained dynamic correlation modeling (2.06). Besides, other state-of-the-art baselines do not obtain prior connections, and D still outperforms the best baseline MSR-GCN by 1.31 in average MPJPE, indicating the effectiveness of our proposed constrained dynamic correlation modeling.

3) Spatial and Temporal GCs: To investigate the contribution of spatial and temporal GCs, the spatial and temporal
that temporal correlations are more important than spatial modeling and 2) H outperforms G by 1.74, demonstrating and emphasizing the importance of spatiotemporal correlation. In the full model, all spatial or temporal GCs cannot make accurate predictions by 5.19 and 3.45, respectively, indicating that individual spatial or temporal GCs are replaced with their spatiotemporal-equivalent variants.

V. DISCUSSIONS

A. Limitations

We only design a multilayer perceptron based on pairwise concatenation to extract frame-wise/jointwise spatial/temporal correlation adjustments. This exhaustive correlation modeling strategy calculates correlations between some unrelated vertices. A more powerful and efficient correlation modeling strategy can be designed. Another limitation is the scalability of the prediction framework, where input and output sequence length is fixed. This limitation hinders us from directly adopting the model to motion sequences with arbitrary lengths. This is a common problem for several state-of-the-art models [9], [27], [30], [31].

B. Broader Impacts

Apart from human motion prediction, our proposed DSTD-GCN has two broader impacts.

First, spatiotemporal graphs are presented in a wide range of prediction applications, including point clouds [50], [51], traffic flows [42], [52], trajectory [53], [54], and so on [55], [56], where spatiotemporal-unsampled decomposition and sample-specific correlation are crucial for feature representation. DSTD-GC can be used as a powerful basic convolution unit for model design in the above-mentioned applications.

Second, the key idea of constrained dynamic correlation modeling is to decouple graph correlations into common patterns presented in all samples and unique patterns presented in each instance. The sample-generic and sample-specific patterns are intuitively optimized in distinct directions and should be modeled differently. Explicitly modeling these two patterns with constrained dynamic correlation modeling can reduce the optimization difficulty and might inspire adjacency modeling for other graph-based visual learning tasks [54], [57], [58], [59], [60]. In many of these works, the graph correlations are set either as model parameters, which cannot model sample-specific graph correlations, or as attention weights from input features, where sample-generic correlations are ignored. Their GC representation abilities can potentially be enhanced if they can separately model sample-generic and sample-specific correlations.

For example, our method can potentially be adopted in trajectory prediction. It is obvious that cars move in similar patterns, which are different from the ones from the pedestrians. Thus, the sample-generic correlation can be utilized here to represent inherent trajectory pattern similarities within cars/pedestrians. On the other hand, a car’s movement pattern depends on surrounding cars and pedestrians under different circumstances, which could be depicted by the sample-specific correlation adjustments. Due to the page limits, we cannot conduct experiments to show this. We plan to test the effectiveness of DSTD-GC in future work.

VI. Conclusion

In this work, a novel DSTD-GC is proposed for human motion prediction. DSTD-GC employs constrained dynamic correlation modeling, which extends the conventional GC by combining constrained correlations (from training or prior knowledge) and dynamic correlations from input motion sequences. Mathematical analyses and extensive experiments are conducted to illustrate the powerful spatiotemporal modeling capability of DSTD-GC. The results indicate that DSTD-GC breaks certain constraints of state-of-the-art GCs on spatiotemporal graphs. With this strategy, we propose DSTD-GCN, which outperforms other state-of-the-art methods in prediction accuracy with the fewest parameters.
APPENDIX B
TIME COMPLEXITY OF DSTD-GC

The time complexity analysis of STS-GC and DSTD-GC is presented here. As a GC consists of feature transformation and feature aggregation [44], we first show the time complexity of these two operations is the same for both GCs. Comparing (13) with (10), we find the time complexity is both $O(JTC^2)$ for feature transformation and $O((J + T)JTC)$ for feature aggregation. For clarity, we utilize the frame-joint proximity assumption in Section III-C3 and assume that $C = C'$. Under these assumptions, the overall time complexity of the basic GC operations for both GCs is $O((J + T)JTC^2(C + T))$.

In addition to basic GC operations, DSTD-GC conducts constrained dynamic correlation modeling to extract dynamic correspondence adjustments. The modeling process does not increase the time complexity. Taking spatial modeling as an example, the most time-consuming operations are pairwise concatenation and its following MLP, as illustrated in Fig. 5(b). The time complexity for these two operations is $O(JTC^2/C/r)$. In our implementation, a very large $r$ is used, so $O(TC/r)$ can be approximated to $O(C)$. Thus, the overall time complexity of spatial-constrained dynamic correlation modeling is $O(JTC)$. Combining the assumptions in the previous paragraph, the overall time complexity for the constrained dynamic correlation modeling is $O((J + T)JTC^2)$, which is less or equal to the time complexity of the basic GC operations. Therefore, the overall time complexity of DSTD-GC is $O((J + T)JTC^2(C + T))$.

With the above-mentioned analysis, it can be concluded that STS-GC and DSTD-GC have the same time complexity, and the extra operations in DSTD-GC do not increase the time complexity.

REFERENCES
[1] B. Paden, M. Cáp, S. Z. Yong, D. Yershov, and E. Frazzoli, “A survey of motion planning and control techniques for self-driving urban vehicles,” IEEE Trans. Intell. Vehicles, vol. 1, no. 1, pp. 33–55, Mar. 2016.
[2] Y. Kong and Y. Fu, “Human action recognition and prediction: A survey,” 2018, arXiv:1806.11230.
[3] V. V. Unhelkar et al., “Human-aware robotic assistant for collaborative assembly: Integrating human motion prediction with planning in time,” IEEE Robot. Autom., Lett., vol. 3, no. 3, pp. 2394–2401, Jul. 2018.
[4] Z. Yan, W. He, Y. Yang, L. Sun, and X. Yu, “Probabilistic motion prediction and skill learning for human-to-cobot dual-arm handover control,” IEEE Trans. Neural Netw. Learn. Syst., Early access, Jun. 30, 2022, doi: 10.1109/TNNLS.2022.3182973.
[5] N. F. Troje, “Decomposing biological motion: A framework for analysis and synthesis of human gait patterns,” J. Vis., vol. 2, no. 5, p. 2, Sep. 2002.
[6] G. W. Taylor, L. Sigal, D. J. Fleet, and G. E. Hinton, “Dynamical binary latent variable models for 3D human pose tracking,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. 2010, pp. 631–638.
[7] A. M. Lehrmann, P. V. Gehlbruck, and S. Nowozin, “Efficient nonlinear Markov models for human motion,” in Proc. IEEE Comput. Vis. Pattern Recognit., Jun. 2014, pp. 1314–1321.
[8] C. Li, Z. Zhang, W. S. Lee, and G. H. Lee, “Convolutional sequence to sequence model for human dynamics,” in Proc. IEEE/ACM Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 5226–5234.
[9] X. Liu, J. Yin, J. Liu, P. Ding, J. Liu, and H. Liu, “TrajectoryCNN: A new spatio-temporal feature learning network for human motion prediction,” IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 6, pp. 2133–2146, Jun. 2021.
[10] A. F. Al-qaqel and M. Ali Khan, “Attention mechanism for human motion prediction,” in Proc. 3rd Int. Conf. Comput. Appl. Inf. Secur. (ICCAIS), Mar. 2020, pp. 1–6.
[11] X. Guo and J. Choi, “Human motion prediction via learning local structure representations and temporal dependencies,” in Proc. AAAI Conf. Artif. Intell., 2019, pp. 2580–2587.
[12] J. Martinez, M. J. Black, and J. Romero, “On human motion prediction using recurrent neural networks,” in Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2891–2900.
[13] K. Fragiadakis, S. Levine, P. Felsen, and J. Malik, “Recurrent network models for human dynamics,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 4346–4354.
[14] Y. Tang, L. Ma, W. Liu, and W. Zheng, “Long-term human motion prediction by modeling motion context and enhancing motion dynamic,” 2018, arXiv:1805.02513.
[15] Z. Liu et al., “Towards natural and accurate future motion prediction of humans and animals,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 10004–10012.
[16] D. Pavlov, C. Feichtenhofer, M. Auli, and D. Grangier, “Modeling human motion with quaternion-based neural networks,” Int. J. Comput. Vis., vol. 128, no. 4, pp. 855–872, Apr. 2020.
[17] X. Shu, L. Zhang, G. Qi, W. Liu, and J. Tang, “Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 6, pp. 3300–3315, Jun. 2022.
[18] H. Wang, J. Dong, B. Cheng, and J. Feng, “PVRED: A position-velocity recurrent encoder-decoder for human motion prediction,” IEEE Trans. Image Process., vol. 30, pp. 6096–6106, 2021.
[19] L.-Y. Gui, Y.-X. Wang, X. Liang, and J. M. Moura, “Adversarial geometry-aware human motion prediction,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 786–803.
[20] A. Hernandez, J. Gali, and F. Moreno-Noguer, “Human motion prediction via spatio-temporal inpainting,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 7134–7143.
[21] Q. Cui, H. Sun, Y. Kong, X. Zhang, and Y. Li, “Efficient human motion prediction using temporal convolutional generative adversarial network,” Inf. Sci., vol. 545, pp. 427–447, Feb. 2021.
[22] Q. Ke, M. Bennamoun, H. Rahmani, S. An, F. Sohel, and F. Boussaid, “Learning latent global network for skeleton-based action prediction,” IEEE Trans. Image Process., vol. 29, pp. 959–970, 2020.
[23] J. N. Kundu, M. Gor, and R. V. Babu, “BiHMP-GAN: Bidirectional 3D human motion prediction GAN,” in Proc. AAAI Conf. Artif. Intell., 2019, pp. 8553–8560.
[24] Y. Cai et al., “Learning progressive joint propagation for human motion prediction,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 226–242.
[25] E. Aksan, M. Kaufmann, P. Cao, and O. Hilliges, “A spatio-temporal transformer for 3D human motion prediction,” in Proc. Int. Conf. 3D Vis. (3DV), Dec. 2021, pp. 565–574.
[26] G. Tevet, S. Raab, B. Gordon, Y. Shafir, A. H. Bermano, and D. Cohen-Or, “Human motion diffusion model,” in Proc. Int. Conf. Learn. Represent., 2023, pp. 1–12.
[27] W. Mao, M. Liu, M. Salzmann, and H. Li, “Learning trajectory dependencies for human motion prediction,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 9489–9497.
[28] W. Mao, M. Liu, and M. Salzmann, “History repeats itself: Human motion prediction via motion attention,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 474–489.
[29] M. Li, S. Chen, Y. Zhao, Y. Zhang, Y. Wang, and Q. Tian, “Dynamic multiscale graph neural networks for 3D skeleton based human motion prediction,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 214–223.
[30] Q. Cui, H. Sun, and F. Yang, “Learning dynamic relationships for 3D human motion prediction,” in Proc. IEEE/CVF Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 6519–6527.
[31] L. Dang, Y. Nie, C. Long, Q. Zhang, and G. Li, “MSR-GCN: Multi-scale residual graph convolution networks for human motion prediction,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 11467–11476.
[32] Z. Liu et al., “Motion prediction using trajectory cues,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 13299–13308.
[33] T. Sofofian, A. Sampieri, L. Franco, and F. Galasso, “Space-time-separable graph convolutional network for pose forecasting,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 11209–11218.
[34] M. Li, S. Chen, Y. Shen, G. Liu, I. W. Tsang, and Y. Zhang, “Online multi-agent forecasting with interpretable collaborative graph neural networks,” *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–15, 2022.

[35] M. Dong and C. Xu, “Skeleton-based human motion prediction with privileged supervision,” *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Apr. 21, 2022, doi: 10.1109/TNNLS.2022.3168681.

[36] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 1–9.

[37] P. Su, Z. Liu, S. Wu, L. Zhu, Y. Yin, and X. Shen, “Motion prediction via joint dependency modeling in phase space,” in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 713–721.

[38] Z. Liu, K. Lyu, S. Wu, H. Chen, Y. Hao, and S. Ji, “Aggregated multi-GANs for controlled 3D human motion prediction,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 2225–2232.

[39] K. N. Li, Z. Liu, S. Wu, H. Chen, X. Zhang, and Y. Yin, “Learning human motion prediction via stochastic differential equations,” in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4976–4984.

[40] T. Deng, Y. Ma, and J. Tang, “Signed graph convolutional networks,” in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2018, pp. 929–934.

[41] G. Qian, A. Abualshour, G. Li, A. Thabet, and B. Ghanem, “PU-GCN: Point cloud upsampling using graph convolutional networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11683–11692.

[42] B. Yu, H. Yin, and Z. Zhu, “Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting,” 2017, arXiv:1709.04875.

[43] M. Li, S. Chen, Y. Zhao, Y. Zhang, Y. Wang, and Q. Tian, “Multiscale spatio-temporal graph neural networks for 3D skeleton-based motion prediction,” *IEEE Trans. Image Process.*, vol. 30, pp. 7760–7775, 2021.

[44] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” 2016, arXiv:1609.02907.

[45] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, “Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, Jul. 2014.

[46] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, “Recovering accurate 3D human pose in the wild using IMUs and a moving camera,” in *Proc. EuC. Conf. Comput. Vis.*, 2018, pp. 601–617.

[47] A. Paszke et al., “PyTorch: An imperative style, high-performance deep learning library,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

[48] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[49] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–11.

[50] B. Mersch, X. Chen, J. Behley, and C. Stachniss, “Self-supervised point cloud prediction using 3D spatio-temporal convolutional networks,” 2021, arXiv:2110.04076.

[51] P. Gomes, S. Rossi, and L. Toni, “Spatio-temporal graph-RNN for point cloud prediction,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 3428–3432.

[52] H.-F. Yang, T. S. Dillon, and Y.-P.-P. Chen, “Optimized structure of the traffic flow forecasting model with a deep learning approach,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2371–2381, Oct. 2017.

[53] B. Yang, G. Yan, P. Wang, C. Chan, X. Song, and Y. Chen, “A novel graph-based trajectory predictor with pseudo-oracle,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 7064–7078, Dec. 2022.

[54] L. Shi et al., “SGCN: Sparse graph convolution network for pedestrian trajectory prediction,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 8994–9003.

[55] Y. Kong et al., “Spatio-temporal graph convolutional network for diagnosis and treatment response prediction of major depressive disorder from functional connectivity,” *Hum. Brain Mapping*, vol. 42, no. 12, pp. 3922–3933, Aug. 2021.

[56] L. Yu, B. Du, X. Hu, L. Sun, L. Han, and W. Lv, “Deep spatio-temporal graph convolutional networks for traffic accident prediction,” *Neurocomputing*, vol. 423, pp. 135–147, Jan. 2021.

[57] N. Liu, W. Zhao, D. Zhang, J. Han, and L. Shao, “Light field saliency detection with dual local graph learning and reciprocal guidance,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4712–4721.

[58] N. Liu, L. Li, W. Zhao, J. Han, and L. Shao, “Instance-level relative saliency ranking with graph reasoning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8321–8337, Nov. 2022.

[59] D. Zhang, W. Zeng, J. Yao, and J. Han, “Weakly supervised object detection using proposal- and semantic-level relationships,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3349–3363, Jun. 2022.

[60] K. Liu, R. Ding, Z. Zou, L. Wang, and W. Tang, “A comprehensive study of weight sharing in graph networks for 3D human pose estimation,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 318–334.

---

**Jiajun Fu** received the B.E. degree in computer science and technology from the Beijing University of Posts and Telecommunications, Beijing, China, in 2020, where he is currently pursuing the M.S. degree with the School of Artificial Intelligence. His research interests include spatiotemporal multimedia applications, such as human motion prediction and traffic flow prediction.

**Fuxing Yang** received the Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 1996. He is currently a Professor with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include computer vision, machine learning, image processing, and deep learning.

**Yonghao Dang** received the bachelor’s degree in computer science and technology from the University of Jinan, Jinan, China, in 2018. He is currently pursuing the Ph.D. degree with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include computer vision, machine learning, image processing, and deep learning.

**Xiaoli Liu** received the bachelor’s and M.Eng. degrees from the University of Jinan, Jinan, China, in 2015 and 2018, respectively, the Ph.D. degree from the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China. Her research interests include computer vision, machine learning, image processing, and deep learning.

**Jianqin Yin** (Member, IEEE) received the Ph.D. degree from Shandong University, Jinan, China, in 2013. She is currently a Professor with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China. Her research interests include service robots, pattern recognition, machine learning, and image processing.