Spatio-Temporal Graph Convolutional Network-Based Beam Training and Tracking for XL-MIMO Systems

Yanzhi Qian, *Student Member, IEEE*, Jing Jiang, *Member, IEEE*, Mengyan Yuan, Yinghui Ye, *Member, IEEE*, Shuaifei Chen, *Member, IEEE*, and Jiayi Zhang, *Senior Member, IEEE*

***Abstract*—In highly dynamic environments, the performance of extremely large-scale multiple-input multiple-output (XL-MIMO) is obviously degraded due to rapid and frequent beam switching. To address this challenge, this letter proposes a spatio-temporal graph convolutional network (STGCN) based beam training and tracking scheme. Initially, the spatio-temporal graphs are constructed to build the spatial relationships between base station (BS) and user equipments (UEs). And then, graph convolutional network (GCN) learns and updates the spatial relationship of all nodes, leveraging the adjacency matrix to enhance the robustness of beam selection in dynamic environments. Finally, gated recurrent unit (GRU) outputs the best beam at the subsequent times by learning the changes of the UEs' motion trajectory. Simulation results demonstrate that the proposed method achieves superior accuracy and robustness compared to conventional beam training approaches.**

***Index Terms*—Beam tracking, beam training, XL-MIMO, spatio-temporal graph convolutional network, near field.**

I. Introduction

EXTREMELY large-scale multiple-input multiple-output (XL-MIMO) is a key enabler for sixth-generation (6G) wireless communication. By deploying hundreds or even thousands of antenna elements at the base station (BS), it can generate numerous narrow, pencil-shaped beams to achieve significant beamforming and spatial multiplexing gains. As the antenna scale increases, the systems exhibit spatial non-stationarity in wireless channel, especially when user equipment (UE) locates in the near-field region [1]. Consequently, accurate channel state information (CSI) must capture both angular and distance information [2], [3]. Therefore, in high-mobility scenarios that UE moves rapidly or scatters change dynamically, the frequent and rapid beam switching poses major challenges to beam training for XL-MIMO systems.

To improve the training efficiency, it leveraged the near-field "rainbow" effect to generate multiple beams aligning different angles within a single training slot [3]. In further, a tree-search beam alignment scheme is designed that first focused on angle alignment and then subsequently on distance [4]. Another work utilized position-aware beam training, in which UE location information was used to narrow down the candidate beam set [5]. More recently, data-driven techniques have been introduced. A convolutional neural network (CNN) was used to predict the best near-field beam from pre-estimated CSI [6], while graph neural networks (GNNs) was used to combine UE position with the far-field wide-beam gains to improve beam-searching accuracy, especially for users at same directions [7]. To handle frequent beam switching induced by UE moving, an extended Kalman filter was applied by kinematic models to track UE positions and predict the best beam [8].

The above works have contributed to reducing beam training overhead and improving training efficiency. Although the work in [7] predicted the best beam according to the UE's trajectory, the performance of XL-MIMO systems is still difficult to guarantee in high dynamic environment since the mapping relationship between UE's positions and the best beams changes frequently. Thus, further investigation is needed to effectively learn the environmental dynamics. Notably, spatio-temporal graph convolutional networks (STGCN) have exhibited strong predictive capabilities by modeling complex spatio-temporal dependencies, e.g., the mobile network traffic forecasting [9]. Inspired by it, we utilize STGCN to learn, update and predict the mapping relationships among UE's positions and the best beam in highly dynamic environments.

In this letter, we propose a STGCN-based beam training and tracking method for XL-MIMO systems. It constructs a series of spatio-temporal graphs at multiple times to build the spatial relationships between BS and UEs. A graph convolutional network (GCN) is used to learn spatial dependencies and multi-user interference patterns, while a gated recurrent unit (GRU) captures temporal dynamics to predict the best beam in the subsequent times. Simulation results show that the proposed method outperforms the existing methods in both accuracy and robustness, particularly in highly dynamic environments.

Received 1 December 2025; accepted 7 January 2026. Date of publication 12 January 2026; date of current version 26 January 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62371392, Grant 62201451, and Grant 62401643; in part by the Key Program for Key Industrial Chain Projects of Shaanxi Province under Grant 2023-ZDLGY-49 and Grant 2024GX-ZDCYL-05-01; in part by the Shaanxi Qinchuangyuan "Scientist and Engineer" Team Building Project under Grant 2024QCY-KXG-156; in part by the Xi'an Science and Technology Plan Project under Grant 23GXFW0084 and Grant 24LLRHZDZX0018; in part by the China Postdoctoral Science Foundation under Grant 2024M752439; and in part by the Xi'an University of Posts and Telecommunications 2024 Graduate Innovation Fund under Grant CXJJBDL2024006. The associate editor coordinating the review of this article and approving it for publication was I. Al-Nahhal. *(Corresponding author: Jing Jiang.)*

Yanzhi Qian, Jing Jiang, Mengyan Yuan, and Yinghui Ye are with the School of Communication and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710121, China (e-mail: qianstudent24@163.com; jiangjing@xupt.edu.cn; yuanmengyan0@163.com; connectyyh@126.com).

Shuaifei Chen is with Purple Mountain Laboratories, Nanjing 211111, China, and also with the National Mobile Communications Research Laboratory, School of Information Science and Engineering, Southeast University, Nanjing 211189, China (e-mail: shuaifeichen@seu.edu.cn).

Jiayi Zhang is with the State Key Laboratory of Advanced Rail Autonomous Operation, Beijing Jiaotong University, Beijing 100044, China (e-mail: zhangjiayi@bjtu.edu.cn).

Digital Object Identifier 10.1109/LWC.2026.3653554

II. System Model

We consider a downlink single-cell XL-MIMO communication system with a BS and K single-antenna UEs. The BS is equipped with a large uniform linear array (ULA) comprising N antenna elements. The BS utilizes an analog beamforming system, where N antennas are connected to N_{RF} radio frequency (RF) chains via a phase shifter network. The near-field channel between the UE k and the BS can be expressed as [10]

$$\mathbf{h}_k = \sum_{l=0}^{L-1} \sqrt{N}\beta_{k,l}\mathbf{b}\left(\theta_{k,l}, r_{k,l}\right) \in \mathbb{C}^{1\times N}, \tag{1}$$

where L is the number of paths that the line-of-sight (LoS) component of the channel corresponds to $l = 0$ and the non-LoS (NLoS) components to $l \geq 1$. $\beta_{k,l}$, $\theta_{k,l}$, and $r_{k,l}$ represent the path gain, the angle, and the propagation distance of the l-th path between the first antenna of the BS and the k-th UE, respectively. And, $\theta_{k,l} = \sin\varphi_{k,l}$, where $\varphi_{k,l}$ is the angle of arrival (AOA). $\mathbf{b}(\theta_{k,l}, r_{k,l})$ is the near-field beam steering vector and can be expressed as

$$\begin{aligned}&\mathbf{b}\left(\theta_{k,l}, r_{k,l}\right)\\ &= \sqrt{\frac{1}{N}}\left[1, e^{-j\frac{2\pi}{\lambda}\left(r_{k,l}^{(2)}-r_{k,l}\right)}, \cdots, e^{-j\frac{2\pi}{\lambda}\left(r_{k,l}^{(N)}-r_{k,l}\right)}\right],\end{aligned} \tag{2}$$

where $r_{k,l}^{(n)} = \sqrt{(r_{k,l})^2 + n^2d^2 - 2r_{k,l}\theta_{k,l}nd}$ denotes the distance from the n-th antenna at BS to the k-th UE, and $d = \frac{\lambda}{2}$ denotes the antenna spacing.

Then, the received signal for the k-th UE can be defined as

$$\mathbf{y}_k = \mathbf{h}_k\mathbf{w}_k^H\mathbf{s}_k + \sum\nolimits_{k'=1,k'\neq k}^{K}\mathbf{h}_k\mathbf{w}_{k'}^H\mathbf{s}_{k'} + \mathbf{n}_k, \tag{3}$$

where $\mathbf{s}_k$ denotes the data signal and satisfies $|\mathbf{s}_k|^2 = 1$, $\mathbf{w}_k \in \mathbb{C}^{1\times N}$ denotes the analog beamformer, and $\mathbf{n}_k \sim \mathcal{CN}(0, \sigma_k^2)$ represents the additive white Gaussian noise (AWGN).

According to (3), the achievable rate of the k-th UE can be expressed as

$$R_k = \log_2\left(1 + \frac{|\mathbf{h}_k\mathbf{w}_k^H|^2}{\sum_{k'=1,k'\neq k}^{K}\left|\mathbf{h}_k\mathbf{w}_{k'}^H\right|^2 + \sigma_k^2}\right). \tag{4}$$

To reduce the CSI feedback overhead, the beamforming vectors are selected from a predefined near-field codebook $\mathcal{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_i, \ldots, \mathbf{c}_Q]$, where $Q = M \times S$ with M and S being the sampled number of angles and distances, respectively. Each codeword is defined as $\mathbf{c}_i = \mathbf{b}(\bar{\theta}_m, r_s^m)$, $\quad i = (s-1)M + m$. The angle sampling is uniformly divided and given by

$$\bar{\theta}_m = \frac{2m - M + 1}{M}, \quad m = 1, 2, \ldots, M. \tag{5}$$

In distance dimension, the sampling is non-uniform due to the distance ring and can be calculated as [3]

$$r_m^s = \frac{M^2d^2}{2s\beta_\Delta^2\lambda}\left(1 - \bar{\theta}_m^2\right), s = 1, 2, 3, \ldots, S, \tag{6}$$

where β_Δ denotes the correlation of the near-field beam steering vectors.

In XL-MIMO communication systems, beam training selects the best codeword $\mathbf{w}_k^{opt}$ from the predefined codebook $\mathcal{C}$ to align the beam to the objective user and obtain the maximum data rate. Therefore, the XL-MIMO beam training problem can be formulated as

$$\{\mathbf{w}_k^{\mathrm{opt}}\}_{k=1}^K = \underset{\{\mathbf{w}_k\in\mathcal{C}\}_{k=1}^K}{\arg\max} \sum_{k=1}^{K} R_k. \tag{7}$$

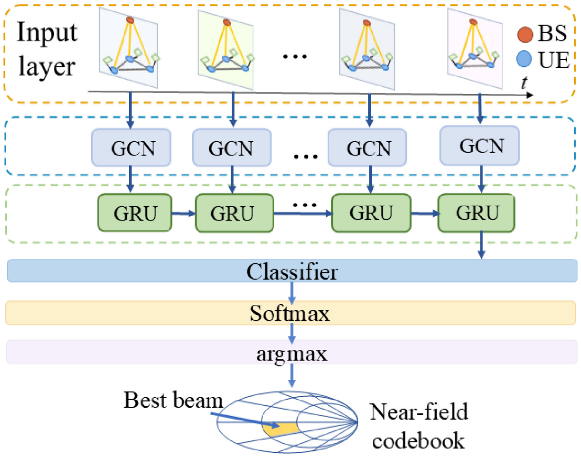

Fig. 1. The STGCN-based Beam Training and Tracking Algorithm.

III. The STGCN-Based Beam Training and Tracking Scheme

We propose a STGCN-based beam training and tracking scheme as shown in Fig. 1. It consists of three phases. Firstly, we build a series of spatio-temporal graphs to capture the spatio-temporal relationships between the BS and all UEs across multiple times. Secondly, GCN constructs a global spatial representation through the adjacency matrix, leveraging the correlations between adjacent nodes to enhance the robustness of beam selection in dynamic environments. Finally, GRU learns the temporal dynamics to predict the best beam for the upcoming time slots.

A. Spatio-Temporal Graph Modeling for XL-MIMO Systems

XL-MIMO system is formulated as a spatio-temporal graph, which could visually describe the relationship of BS, UEs and their best beams. Specifically, it is expressed as a sequence of spatial graphs $\{\mathcal{G}_t(\mathcal{V}_t, \mathcal{E}_t)\}$, where $\mathcal{V}_t$ and $\mathcal{E}_t$ are the sets of nodes and edges at time t, respectively. Firstly, the set of nodes consist of BS and UEs, which can be expressed as $\mathcal{V}_t = \{\mathbf{v}_{0,t}, \cdots, \mathbf{v}_{k,t}, \cdots, \mathbf{v}_{K,t}\}$. In specific, $\mathbf{v}_{k,t} = [r_{k,t}, \psi_{k,t}]$ denotes the location of the k-th UE node, where $r_{k,t}$ and $\psi_{k,t}$ represent the polar radius and polar angle in the polar coordinate, respectively. In addition, $\mathbf{v}_{0,t} = [0, 0]$ denotes the location of BS. On the other hand, $\mathcal{E}_t$ is the edge set and expressed as $\mathcal{E}_t = \{\mathbf{e}_{0,t}, \ldots, \mathbf{e}_{k,t}, \ldots, \mathbf{e}_{K,t}\}$, where $\mathbf{e}_{k,t}$ is the edge connecting the BS to the k-th UE at time t.

1) Feature Matrix: To establish the relationship between UEs and their best beams, We consider the coordinates, path gain, motion direction, and velocity components of the nodes at time slot t to be each node's feature. While the feature of the node changes, its best beam varies accordingly. We construct the feature matrix as $\mathbf{X}_t = [\mathbf{x}_t^0; \mathbf{x}_t^1; \ldots; \mathbf{x}_t^K] \in \mathbb{R}^{(K+1)\times P}$, where the first row corresponds to the BS and the remaining

K rows represent the UEs. Here, P denotes the dimension of the feature vector for each node. Notably, the BS is considered as an anchor node at the origin, with its feature vector set to zero. This real-valued transformation enables efficient neural network processing while preserving the essential channel and geometry characteristics.

2) Adjacency Matrix: In XL-MIMO beam training and tracking, the edge weight matrix $\mathbf{A}_t \in \mathbb{R}^{(K+1)\times(K+1)}$ of the graph $\mathcal{G}_t(\mathcal{V}_t, \mathcal{E}_t)$ represents the spatial relationships of different nodes. Let $\mathbf{w}_{k,t} \in \mathbb{C}^{1\times N}$ denote the beamforming vector selected for the k-th UE at time t. $\mathbf{a}_t^{kk'} \in \mathbf{A}_t$ is the weighted relationship between the k-th node and the k'-th node, which can be denoted as

$$\mathbf{a}_t^{kk'} = \begin{cases} |\mathbf{h}_{k,t}\mathbf{w}_{k,t}^H|, & \text{if } k' = 0, \\ |\mathbf{h}_{k,t}\mathbf{w}_{k',t}^H|, & \text{if } k' \neq k \text{ and } k' \neq 0, \\ 1, & \text{if } k' = k. \end{cases} \tag{8}$$

Specifically, $\mathbf{a}_t^{kk'}$ represents the beamforming gain between BS and the k-th UE while $k' = 0$, the inter-user interference while $k' \neq k$, and the autocorrelation characteristics of the k-th node $(k' = k)$.

During the offline training phase, $\mathbf{A}_t$ is computed using the beamforming vectors $\mathbf{w}_{k,t}$, which are obtained through the exhaustive search. In the subsequent online stage, the beam vectors selected by the trained STGCN model are used to iteratively update $\mathbf{A}_t$. It establishes a closed-loop interaction between the predicted beam vectors and the spatio-temporal graph. Each user can get the feature information from other nodes by the adjacency matrix to improve the robustness of beam selection in the dynamically changing environments.

3) Problem Definition: STGCN will output the best beam for every UE by learning and updating the spatial and temporal features of high dynamic wireless environment. Accordingly, Problem (7) can be reformulated as

$$\left\{\mathcal{W}_{t+1}^{opt}\right\} = \mathrm{STGCN}\Big(\{\mathcal{G}_{t-i}, \mathbf{A}_{t-i}, \mathbf{X}_{t-i}\}_{i=0}^{T-1}\Big), \tag{9}$$

where $\mathcal{W}_{t+1}^{opt} = \{\mathbf{w}_{1,t+1}^{opt}, \ldots, \mathbf{w}_{k,t+1}^{opt}, \ldots, \mathbf{w}_{K,t+1}^{opt}\}$ denote the set of best beamforming vectors for all K UEs at time $t+1$. Here, $\mathbf{w}_{k,t+1}^{opt}$ represents the best beam for the k-th UE at time $t+1$, which is then utilized to construct the adjacency matrix for the subsequent prediction step.

B. The STGCN Beam Training and Tracking Method

1) Spatial Learning Network Based GCN: In the spatial learning stage, we employ a two-layer GCN to extract the spatial features from network nodes [11]. It builds the global spatial mapping through the adjacency matrix, leveraging the mapping correlations between the adjacent nodes to improve the robustness of beam selection in dynamic environments. Specifically, at time t, each GCN layer performs message passing to propagate the information of the adjacent nodes and feature aggregation to update each node by combining its own features with those of its neighbors. In the message transformation phase, each node encodes its local features into a latent representation. This feature preprocessing step is formulated as:

$$\hat{\mathbf{X}}_t = \mathrm{ReLU}(\mathbf{U}_1 \cdot \mathbf{X}_t + \mathbf{B}_1), \tag{10}$$

where the ReLU activation introduces crucial nonlinearities, enabling the learning of complex feature interactions. Here, $\mathbf{U}_1$ denotes the weight matrix and $\mathbf{B}_1$ represents the bias matrix.

Then, the feature aggregation phase propagates and combines these transformed messages across the graph topology. To stabilize feature aggregation, the normalized adjacency matrix is defined as $\widehat{\mathbf{A}}_t = \tilde{D}_t^{-\frac{1}{2}}\mathbf{A}_t\tilde{D}_t^{-\frac{1}{2}}$, where $\tilde{D}_t = \sum_k \mathbf{a}_t^{kk'}$ is a diagonal degree matrix. Then the output of two-layer GCN model can be formulated as

$$\mathbf{F}_t = \sigma\Big(\widehat{\mathbf{A}}_t\sigma\Big(\widehat{\mathbf{A}}_t\hat{\mathbf{X}}_t\mathbf{U}_2\Big)\mathbf{U}_3\Big), \tag{11}$$

where $\mathbf{F}_t \in \mathbb{R}^{(K+1)\times H}$ denotes the output feature matrix. $\mathbf{U}_2 \in \mathbb{R}^{P\times H}$ and $\mathbf{U}_3 \in \mathbb{R}^{H\times H}$ are the weight matrices, H denotes the hidden dimension of the GCN layers and $\sigma(x)$ is the sigmoid function used for nonlinear modeling.

2) Temporal Learning Network Based GRU: Utilizing the historical spatial graph features from time $t-T+1$ to time t, the temporal learning phase employs a GRU to model the dynamic evolution of wireless channels and predict the best beams for time $t+1$. Specifically, the input sequence to the GRU at step t is given by $\{\mathbf{F}_{t-i}\}_{i=0}^{T-1}$. The GRU includes a reset gate $\mathbf{r}_t$ and an update gate $\mathbf{z}_t$, which are computed as

$$\mathbf{r}_t = \sigma(\mathbf{U}_r[\mathbf{F}_t, \mathbf{g}_{t-1}] + \mathbf{B}_r), \tag{12}$$

$$\mathbf{z}_t = \sigma(\mathbf{U}_z[\mathbf{F}_t, \mathbf{g}_{t-1}] + \mathbf{B}_z), \tag{13}$$

where $\mathbf{U}_r$ and $\mathbf{U}_z$ are the weight matrices, and $\mathbf{B}_r$ and $\mathbf{B}_z$ are the bias vectors. $\mathbf{F}_t$ represents the spatial graph feature at time t and is defined in (12). $\mathbf{g}_{t-1}$ denotes the hidden state at time $t-1$, which stores spatio-temporal features from time $t-T+1$ to time $t-1$.

From the values of the update and reset gates, the unit state is computed as

$$\mathbf{c}_t = \tanh\big(\mathbf{U}_c\big[\mathbf{F}_t, (\mathbf{r}_t * \mathbf{g}_{t-1})\big] + \mathbf{B}_c\big), \tag{14}$$

where $\mathbf{U}_c$ is the weight matrix and $\mathbf{B}_c$ is the bias vector. $\tanh(x) = (1-e^{-2x})/(1+e^{-2x})$ is the activation function. The hidden state is then updated as

$$\mathbf{g}_t = \mathbf{z}_t * \mathbf{g}_{t-1} + (1-\mathbf{z}_t) * \mathbf{c}_t. \tag{15}$$

After processing $\mathbf{F}_t$, we use the hidden states corresponding to the K UEs to predict the best beams at time $t+1$. The output layer is given by

$$\mathbf{q}_{t+1} = \mathbf{g}_t\mathbf{U}_q + \mathbf{B}_q \in \mathbb{R}^{K\times Q}, \tag{16}$$

where $\mathbf{U}_q$ is the output layer weight matrix and $\mathbf{B}_q$ is the output layer bias vector.

For the k-th UE, its spatio-temporal features encapsulated in $\mathbf{q}_{k,t+1} \in \mathbb{R}^{1\times Q}$ are decoded into a probability distribution over the near-field codebook $\mathcal{C}$ via the softmax function, which is given by

$$\mathbf{p}_{k,t+1} = \mathrm{softmax}\big(\mathbf{q}_{k,t+1}\big), \tag{17}$$

Finally, for the k-th UE at time $t+1$, the best beam is selected as the codeword with the largest predicted probability. Let $\mathbf{p}_{k,t+1} = [p_{k,t+1}^{(1)}, p_{k,t+1}^{(2)}, \ldots, p_{k,t+1}^{(Q)}]$ denote the probability vector over the Q codewords. The corresponding beamforming vector is given by

$$\mathbf{w}_{k,t+1}^{opt} = \mathrm{argmax}\,\mathbf{p}_{k,t+1}. \tag{18}$$

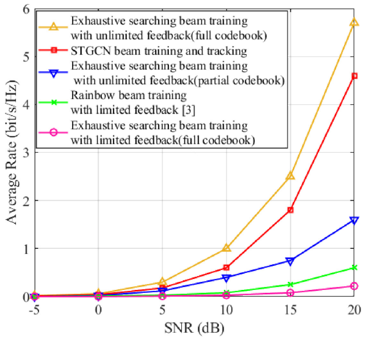

Fig. 2. Sum-rate performance for different SNRs.

C. Loss Function

During the training phase, our objective is to learn a beam selection strategy that maximizes the achievable sum-rate. Since the beamforming vectors are selected from a predefined codebook, the beam selection problem is formulated as a supervised multi-class classification task. Specifically, for the k-th UE at time $t+1$, we first perform an exhaustive search over the codebook to select the beam index that maximizes the achievable sum-rate. The resulting beam index is encoded as a one-hot label vector $\mathbf{u}_{k,t+1} \in \{0,1\}^Q$. Let $\mathbf{p}_{k,t+1}$ denotes the output probability distribution from the softmax layer in (18).

To train the STGCN to imitate these oracle beam-selection decisions, we minimize the cross-entropy loss between the oracle one-hot labels and the predicted probabilities. Accordingly, the loss function for STGCN training at time $t+1$ is given by

$$\mathcal{L} = -\sum_{k=1}^{K}\sum_{q=1}^{Q} u_{k,t+1}^{(q)} \log p_{k,t+1}^{(q)}, \tag{19}$$

where $u_{k,t+1}^{(q)}$ is the q-th element of the one-hot label vector $\mathbf{u}_{k,t+1}$ for user k at time $t+1$, and $p_{k,t+1}^{(q)}$ is the predicted probability for the q-th beam. In practice, this loss is accumulated over all training samples and time slots during training.

IV. Simulation Results

In this section, the performance of the proposed STGCN-based beam training and tracking scheme is evaluated. During the offline phase, the ground-truth beam labels are generated via exhaustive search across the entire codebook. Once trained, STGCN will forecast the beamforming set for the subsequent time slots, preventing the need for further exhaustive searching. STGCN's learning ability enables autonomous update the best beam coordinated with the dynamic environment.

The channel between each user and the BS is set up with $L = 3$ channel paths with one LoS path and two NLoS paths, where the channel gain of the LoS path obeys $\beta_{k,0} \sim \mathcal{CN}(0,1)$ and the NLoS paths obey $\beta_{k,1} \sim \mathcal{CN}(0,0.01)$ and $\beta_{k,2} \sim \mathcal{CN}(0,0.01)$ [8]. The simulation parameters are shown in Table I.

The STGCN consists of a GRU layer and two GCN layers. The hidden size of each GCN or GRU is 64. The number of training epochs is 500, and the batch size is set to 100. A dataset of 50,000 simulated samples with a fixed number of K single-antenna UEs is divided into 80% for training and 20% for testing. In each sample, we employ a Gauss–Markov process for speed evolution and a correlated random-walk model for trajectory generation. Let v_t denote the UE's scalar speed (in m/s) and $\mathbf{p}_t = [x_t, y_t]^\top$ its 2D position of the UE (in m) at time *t*. At the start of each simulation run, the initial position $\mathbf{p}_0$ is sampled uniformly over the cell area, while the initial speed follows $v_0 \sim \mathcal{N}(18,2)$. At each time slot, the heading angle α_t is independently drawn from $\alpha_t \sim \mathcal{U}[-\pi,\pi]$, which determines the unit direction vector $\mathbf{d}_t$. The frame duration is set to $T_{\text{frame}} = 0.1\,\text{ms}$. The resulting trajectory $\{\mathbf{p}_t\}_{t=0}^{T-1}$ is generated via the following recursive updates:

$$\begin{cases} \mathbf{p}_t &= \mathbf{p}_{t-1} + T_{\text{frame}}\, v_{t-1}\mathbf{d}_{t-1}, \\ v_t &= \gamma_m v_{t-1} + (1-\gamma_m)\mu_m + \sigma_m\sqrt{1-\gamma_m^2}\, n_{t-1} + a_t, \\ \mathbf{d}_t &= [\cos\alpha_t, \sin\alpha_t]^\top, \end{cases} \tag{20}$$

where $\gamma_m = 0.9$ controls the temporal correlation of speed, $\mu_m = 18\text{m/s}$ is the long-term mean speed, and $\sigma_m = 1$ controls the variance of the Gauss-Markov process. The sequence $\{n_t\}$ consists of i.i.d. standard Gaussian variables $n_t \sim \mathcal{N}(0,1)$, modeling inherent stochasticity in speed variation. Additionally, the term $\{a_t\}$ captures short-term random accelerations and is modeled as an i.i.d. Gaussian process with $a_t \sim \mathcal{N}(0,0.5)$, enriching the model's ability to emulate realistic speed fluctuations. In this section, we evaluate the sum-rate performance versus signal-to-noise ratio (SNR) of the proposed STGCN scheme against four traditional methods: the exhaustive searching beam training with unlimited feedback/limited feedback/the partial codebook and the rainbow beam training with limited feedback [3]. For the unlimited and limited feedback, the beam training select and feedback the best beam at every time slot or every T time slot, respectively. The method with partial codebook updates the beam at every

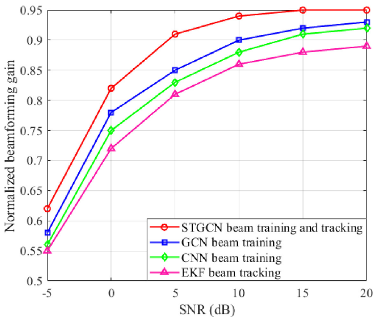

Fig. 3. Normalized beam gain for different SNRs.

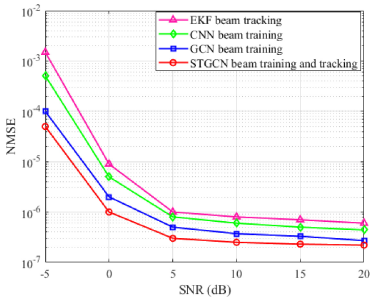

Fig. 4. NMSE performance comparison for different SNRs.

TABLE I
SIMULATION PARAMETERS

Parameters	Value
Number of antennas N	512
Number of UEs K	10
Time window length T	8 time slots
The number of node feature P	8
Carrier frequency f	60GHz
Bandwidth	3GHz
Near-field distance	$(3\text{m}, 200\text{m})$
The number of sampling angles M	256
The number of sampling distance S	8
The number of codewords Q	2048
Size of partial codebook $\lvert\mathcal{C}_p\rvert$	256

time slot, but the search space of the exhaustive search is restricted to a pre-selected subset $\mathcal{C}_p \subset \mathcal{C}$ with $|\mathcal{C}_p| \ll |\mathcal{C}|$ to reduce the training time.

As shown in Fig. 2, the sum-rates of all schemes increase with SNR. Among them, the exhaustive search-based beam training scheme with unlimited feedback establishes the performance upper bound since it performs an exhaustive search in each time slot. It can be observed that the proposed STGCN scheme achieves the performance close to this upper bound. This is attributed to its ability to capture the variations of spatio-temporal graph and update the mapping relationship between the best beam and UE's positions. In contrast, both the partial codebook and limited feedback schemes suffer from performance degradation in highly dynamic environments. Specifically, the partial codebook scheme may fail because the best beam at the current times falls outside the pre-defined codebook due to UE moves or scattering changes. Similarly, the limited feedback schemes are unable to track rapid beam switching as they are constrained by the feedback interval.

In Fig. 3 and Fig. 4, we evaluate the performance of the proposed STGCN scheme against another three schemes that could predict the best beam: the extended Kalman filter (EKF) [8], the GCN-based beam training scheme and the CNN-based beam training scheme. In specific, CNN-based scheme predicts the best beam by capturing local correlations through convolutional operations on 2-dimension feature maps. The GCN-based scheme predicts the best beam by leveraging global spatial dependencies at a single time slot, but it suffers from performance degradation due to its inability to exploit temporal dynamics. In Fig. 3, we evaluate the normalized beamforming gain for different SNRs, which is defined as

$$G = \frac{1}{KT}\sum_{k=1}^{K}\sum_{t=1}^{T}\frac{\left|\hat{\mathbf{w}}_{k,t}^{H}\mathbf{h}_{k,t}\right|^2}{\left|\mathbf{w}_{k,t}^{H}\mathbf{h}_{k,t}\right|^2}, \tag{21}$$

where $\hat{\mathbf{w}}_{k,t}$ and $\mathbf{w}_{k,t}$ are the predicted and actual best beam, respectively, and $\mathbf{h}_{k,t}$ is the channel of the k-th UE at time slot t. It is observed that the normalized gain will increase with the SNR, the STGCN scheme is higher than the other schemes. This shows that STGCN is able to better adapt to spatio-temporal dynamics and achieve the efficient beam selection with limited beam training overhead. Finally, the normalized mean squared error (NMSE) is used to evaluate the prediction performance of the scheme, which is defined as

$$\text{NMSE} = \frac{1}{KT}\sum_{k=1}^{K}\sum_{t=1}^{T}\frac{\mathbb{E}\left\{\left\|\mathbf{w}_{k,t} - \hat{\mathbf{w}}_{k,t}\right\|^2\right\}}{\mathbb{E}\left\{\left\|\mathbf{w}_{k,t}\right\|^2\right\}}, \tag{22}$$

where $\mathbf{w}_{k,t}$ and $\hat{\mathbf{w}}_{k,t}$ are the ideal and predicted codeword, respectively. The results in Fig. 4 show that as the SNR increase, the NMSE of the three methods decreases, and the proposed STGCN scheme can effectively predict the best beam. The superior NMSE performance of STGCN is attributed to its ability to model both spatial and temporal dependencies, allowing it to adapt to dynamic UEs.

V. CONCLUSION

In this letter, we proposed a novel STGCN-based beam training and tracking framework for XL-MIMO systems. It constructed a serial of temporal graph snapshots that comprehensively capture historical channel state information. Through the combination of GCN and GRU, the proposed approach effectively adapts to the varying wireless environments. Simulation results proved that the proposed method achieved significant performance gains over the traditional beam training since it could track the variation of wireless channel.

REFERENCES

[1] H. Lu et al., "A tutorial on near-field XL-MIMO communications toward 6G," *IEEE Commun. Surveys Tut.*, vol. 26, no. 4, pp. 2213–2257, 4th Quart., 2024.

[2] M. Cui, Z. Wu, Y. Lu, X. Wei, and L. Dai, "Near-field MIMO communications for 6G: Fundamentals, challenges, potentials, and future directions," *IEEE Commun. Mag.*, vol. 61, no. 1, pp. 40–46, Jan. 2023.

[3] M. Cui, L. Dai, Z. Wang, S. Zhou, and N. Ge, "Near-field rainbow: Wideband beam training for XL-MIMO," *IEEE Trans. Wireless Commun.*, vol. 22, no. 6, pp. 3899–3912, Jun. 2023.

[4] X. Zhang, H. Zhang, J. Zhang, C. Li, Y. Huang, and L. Yang, "Codebook design for extremely large-scale MIMO systems: Near-field and far-field," *IEEE Trans. Commun.*, vol. 72, no. 2, pp. 1191–1206, Feb. 2024.

[5] Y. Liu, W. Deng, M. Li, and M. J. Zhao, "Position-aware beam training for near-field milimeter-wave XL-MIMO communications," in *Proc. IEEE 99th Veh. Technol. Conf. (VTC)*, 2024, pp. 1–6.

[6] J. Nie, Y. Cui, Z. Yang, W. Yuan, and X. Jing, "Near-field beam training for extremely large-scale MIMO based on deep learning," *IEEE Trans. Mobile Comput.*, vol. 24, no. 1, pp. 352–362, Jan. 2025.

[7] W. Liu, C. Pan, H. Ren, J. Wang, and R. Schober, "Near-field multiuser beam-training for extremely large-scale MIMO systems," *IEEE Trans. Commun.*, vol. 73, no. 4, pp. 2663–2679, Apr. 2024.

[8] S. H. Hyun, J. Song, K. Kim, J. H. Lee, and S. C. Kim, "Adaptive beam design for V2I communications using vehicle tracking with extended Kalman filter," *IEEE Trans. Veh. Technol.*, vol. 71, no. 1, pp. 489–502, Jan. 2022.

[9] H. Wan, S. Guo, Y. Lin, and G. Cong, "Spatio-temporal synchronous graph convolutional networks: A new framework for spatio-temporal network data forecasting," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2020, pp. 1234–1241.

[10] Z. Xu, Z. Zhang, and L. Dai, "Near-optimal near-field beam training: From searching to inference," *IEEE Trans. Wireless Commun.*, vol. 24, no. 11, pp. 9173–9185, Nov. 2025.

[11] X. Wang et al., "Adaptive multi-receptive field spatio-temporal graph convolutional network for traffic forecasting," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, 2021, pp. 1–7.