Spatio-Temporal Electromagnetic Kernel Learning for Channel Prediction

Jinke Li, Jieao Zhu, and Linglong Dai, *Fellow, IEEE*

***Abstract*—Accurate channel prediction is essential for addressing channel aging caused by user mobility. However, the actual channel variations over time are highly complex in high-mobility scenarios, which makes it difficult for existing predictors to obtain future channels accurately. The low accuracy of channel predictors leads to difficulties in supporting reliable communication. To overcome this challenge, we propose a channel predictor based on spatio-temporal electromagnetic (EM) kernel learning (STEM-KL). Specifically, inspired by recent advancements in electromagnetic information theory (EIT), the STEM kernel function is derived. The velocity and the concentration kernel parameters are designed to reflect the time-varying propagation of the wireless signal. We obtain the parameters through kernel learning. Then, the future channels are predicted by computing their Bayesian posterior, with the STEM kernel acting as the prior. To further improve the stability and model expressibility, we propose a grid-based EM mixed kernel learning (GEM-KL) scheme. We design the mixed kernel to be a convex combination of multiple sub-kernels, where each sub-kernel corresponds to a grid point in the set of pre-selected parameters. This approach transforms the non-convex STEM kernel learning problem into a convex grid-based problem that can be easily solved by weight optimization. Finally, simulation results verify that the proposed STEM-KL and GEM-KL schemes can achieve more accurate channel prediction. This indicates that EIT can improve the performance of wireless systems efficiently.**

***Index Terms*—Channel prediction, electromagnetic information theory (EIT), spatio-temporal electromagnetic kernel learning (STEM-KL), grid-based electromagnetic mixed kernel learning (GEM-KL), and multi-input multi-output (MIMO).**

I. Introduction

In recent years, with the development of new applications such as digital twins and virtual reality, the demand for spectral efficiency is predicted to increase rapidly [1]. As a key technology in current wireless communication, massive multiple-input multiple-output (MIMO) can achieve significant improvements in spectral efficiency and system capacity [2]–[4].

The effective communication of massive MIMO systems highly relies on accurate and timely channel state information (CSI) [5]. However, dynamic environments, characterized by user mobility, complicate the acquisition of CSI [6], [7]. According to the current 5G standard [8], in time-division duplexing (TDD) mode, CSI acquisition, or channel estimation, is performed periodically. When user mobility is high, significant channel changes may occur within a single channel estimation period, leading to outdated CSI [9]. This phenomenon is termed as *channel aging* [10]. For example, when the user speed is $60\,\mathrm{km/h}$, channel aging could result in approximately 30% loss in achievable sum-rate performance [11].

To achieve higher spectral efficiency, XL-MIMO, which has many more antennas than massive MIMO, is considered a key technology for 6G [12]. In future 6G scenarios, as the number of antennas in MIMO systems increases significantly, the number of pilots required for channel estimation will also increase. Although pilot density can be increased to accommodate this growing demand, when the number of antennas increases several times, the pilot density cannot withstand the dramatic increase subsequently. Consequently, extending the channel estimation period becomes inevitable, leading to more severe channel aging. Therefore, addressing channel aging has become an urgent priority for XL-MIMO communication systems.

A. Prior Works

To address the challenges posed by channel aging, various channel prediction techniques have emerged that utilize the *temporal correlation* between historical CSI and future CSI. Existing channel prediction methods can be categorized into two main types: The model-based methods, and deep learning (DL)-based methods. The model-based methods can be divided into two categories, i.e., the sparsity-based methods and the autoregressive (AR)-based methods.

Sparsity-based methods typically exploit the Doppler domain sparse structure of channel responses to predict future channels. For instance, the sum-of-sinusoids model-based predictor [13] represents the channel response as a combination of sinusoidal waves. This scheme first identifies the dominant sinusoidal components and then uses the harmonic retrieval method [14] to obtain these components for channel prediction. To be more suitable for predicting massive MIMO channels with a larger number of vector elements, the authors of [15] proposed the Prony vector (PVEC) method, which fits a linear prediction model for the observed channel response. Specifically, PVEC applies to predicting uniformly sampled signals composed of damped sinusoidal components. It models the future channel as a linear combination of the past channels, where the combination weights are computed

This work was supported in part by the National Key Research and Development Program of China (Grant No. 2023YFB3811503), in part by the National Natural Science Foundation of China (Grant No. 62325106), and in part by the National Natural Science Foundation of China (Grant No. 62031019).

The authors are with the Department of Electronic Engineering, Tsinghua University, and the State Key Laboratory of Space Network and Communications, Tsinghua University, Beijing 100084, China (e-mails: lijk23@mails.tsinghua.edu.cn, zja21@mails.tsinghua.edu.cn, daill@tsinghua.edu.cn).

from the received pilot signals. The authors of [16] believe time-varying channels have sparsity in the Doppler frequency domain. Consequently, compressive sensing algorithms such as orthogonal matching pursuit (OMP) [17] can be used to obtain the dominant Doppler frequencies for predicting future channels.

AR-based methods use autoregressive principle to process channel time series [18]. The original AR prediction method models the future channel as a weighted sum of its past values, where the weights, i.e., the AR parameters, are obtained from the autocorrelation function of channels at different times [19]–[21]. The Wiener channel predictor and Kalman channel predictor are extensions of the AR prediction method [22]–[25]. The Wiener predictor enhances channel prediction by predicting an autoregressive multivariate random process using a Wiener linear filter [26]. Moreover, the authors of [27] and [28] explore the application of the Kalman predictor within a time-correlated channel aging model. This method implements channel prediction by modeling the channel as a linear dynamic system with state and observation equations. It predicts the next state based on the current estimate and the state transition model, then improves this prediction using new CSI to correct the estimate and reduce uncertainty.

DL-based methods utilize neural network architectures to learn complex patterns from historical channel data for prediction [28]–[31]. Specifically, fully connected neural networks (FCN) model the channel as a non-linear mapping from past observations to future states, where layers of interconnected neurons process input features to capture underlying dependencies [28]. However, training such FCNs may be challenging due to the high-dimensional input channels in previous frames. To avoid the high-dimensional input, recurrent neural networks (RNNs) extend this by incorporating feedback loops to handle sequential data, enabling the prediction of time-varying channels through hidden states that retain temporal information [32]. Long short-term memory (LSTM)-based methods address the vanishing gradient problem with gating mechanisms (input, forget, and output gates) to selectively remember long-term dependencies, making them effective for channel prediction in high-mobility scenarios [33]. More recently, transformer-based predictors utilize self-attention mechanisms to capture global temporal interactions across channel sequences, outperforming traditional RNNs by parallelizing computations and focusing on relevant dependencies [11], [34]. These DL approaches have been applied to various wireless systems, such as massive MIMO and vehicular communications, often achieving better performance in data-rich environments by training on large datasets of simulated or measured channels.

The existing two categories of channel prediction methods mentioned above can accomplish channel prediction for massive MIMO systems. However, DL-based methods require extensive historical datasets for training, which may be impractical in practical wireless deployments. Moreover, DL models typically function as black boxes [35], often lacking explicit physical interpretability and failing to incorporate underlying electromagnetic principles. Additionally, the training overhead is usually high, involving computationally intensive processes and significant time for optimization. As for the existing model-driven methods, simply modeling time-varying channels as sinusoidal or Gaussian random processes is inaccurate. Due to inaccurate channel modeling, these methods cannot accurately predict the channel. The low accuracy of channel predictors can lead to difficulties in supporting reliable communication in high-mobility scenarios. Therefore, it is essential to investigate a more accurate channel prediction method with physical interpretability.

B. *Our Contributions*

To design a high-accuracy channel predictor with physical interpretability for the XL-MIMO system, we propose a channel prediction scheme based on *electromagnetic kernel learning*, which simultaneously utilizes the spatio-temporal electromagnetic correlation characteristic of the channel from the perspective of electromagnetic information theory (EIT) [36], [37]. The contributions of this paper are summarized as follows:

- Unlike existing channel prediction schemes that do not utilize channel EM characteristics, the proposed scheme uses the EIT-based channel model. Inspired by the spatial correlation function based on electromagnetic (EM) physical principles [38], we consider the time-varying property of the channel and derive the spatio-temporal electromagnetic (STEM) correlation function, i.e., STEM kernel. Specifically, we introduce the velocity parameter in the correlation function to describe user mobility. This STEM kernel originates from EM physics, thus it is more suitable for modeling practical wireless propagation environments than other kernel functions.
- Since the proposed STEM kernel characterizes the channel temporal correlation, we utilize the STEM kernel to construct time-domain channel predictors. To get the STEM kernel parameter, we formulate a maximum likelihood (ML) problem, where the kernel parameters are optimized to fit the noisy channel observations. Furthermore, we design the velocity and concentration kernel parameters to reflect the time-varying propagation of the wireless signal. After determining the kernel parameters, future channels are predicted by computing their Bayesian posterior, with the STEM kernel acting as the prior. Therefore, we introduce EM information into the channel predictor in a physically interpretable way.
- To deal with the non-convexity of the ML problem, we convert it into a convex problem by introducing additional grid weight parameters, leading to a convex grid-based problem that can be easily solved by weight optimization. Specifically, the STEM kernel is approximated by a new grid-based EM mixed (GEM) kernel, which is composed of STEM sub-kernels. For each sub-kernel, parameters are fixed at a set of pre-selected grid points, leaving only the weights to be optimized. Thus, the original continuous parameter optimization problem is converted into a discrete weight optimization problem with favorable convexity and reliability.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2025.3635412

- Finally, through performance analysis and numerical experiments, it can be verified that the proposed GEM-KL channel predictor outperforms the PVEC and AR baselines, which demonstrates that EIT can benefit the performance of wireless communication systems.

C. Organization and Notation

The rest of the paper is organized as follows. Section II introduces the channel model and signal model. Section III formulates the channel prediction problem. In Section IV, we first introduce the spatio-temporal electromagnetic correlation function (STEM-CF). Then, Gaussian process regression (GPR) is introduced to solve the channel prediction problem. Kernel learning is considered to improve EM-based GPR channel prediction, and finally, the GEM-KL scheme is proposed. Simulation results are provided in Section V, and we conclude this paper in Section VI.

Notations: $\mathbf{X}$ and $\mathbf{x}$ denote matrices and vectors, respectively. $\mathbb{E}[X]$ denotes the expectation of random variable $X(\omega)$; $\mathbb{C}$ denotes the set of complex numbers and $\mathbb{R}$ denotes the set of real numbers; $(\cdot)^*$ denotes the conjugate operation; $[\cdot]^{-1}$, $[\cdot]^{\mathsf{T}}$, $[\cdot]^{\mathsf{H}}$ and $\mathrm{diag}(\cdot)$ denote the inverse, transpose, conjugate-transpose and diagonal operations, respectively; i denotes the imaginary unit; $\mathbf{I}_N$ is an $N \times N$ identity matrix; For $\mathbf{x} \in \mathbb{C}^n$ or $\mathbb{R}^n$, $|\mathbf{x}| = \sqrt{\mathbf{x}^{\mathsf{T}}\mathbf{x}} \in \mathbb{C}$ denotes the pseudonorm; $\|\mathbf{x}\|$ denotes the standard vector 2-norm $\sqrt{\mathbf{x}^{\mathsf{H}}\mathbf{x}} \in \mathbb{R}_{\geq 0}$; $\hat{\mathbf{x}}$ denotes $\mathbf{x}/\,|\mathbf{x}|$; $\Re\{\cdot\}$ and $\Im\{\cdot\}$ respectively represent the real and imaginary part of the arguments; $j_m(x)$ is the mth-order spherical Bessel function of the first kind.

II. System Model

In this section, we review the Gaussian random field (GRF)-based channel model and explain the signal model.

A. Channel Model

Traditional channel models express the channel matrix as a weighted Gaussian mixture of steering vectors, which is a discrete special case of a Gaussian random field. In this section, we model the channel with a complex symmetric Gaussian random field (CSGRF) to capture the continuously varying properties of the wireless channel [39], [40]. Let function $h(\boldsymbol{\rho}) : \mathbb{R}^4 \to \mathbb{C}$ represent a circularly symmetric Gaussian random field (CSGRF). The variable is $\boldsymbol{\rho} = (\mathbf{x}, t)$, where $\mathbf{x} = (x, y, z)$ represents the spatial location, t represents time indicator, and $(\mathbf{x}, t) \in \mathbb{R}^4$. For any Q points, the joint distribution of their function values $(h(\boldsymbol{\rho}_1), h(\boldsymbol{\rho}_2), \ldots, h(\boldsymbol{\rho}_Q))$ follows a multivariate Gaussian distribution, then the random field is a Gaussian random field, denoted as $h(\boldsymbol{\rho}) \sim \mathcal{GRF}(0, k(\boldsymbol{\rho}, \boldsymbol{\rho}'))$, and its probability measure is determined by their autocorrelation function

$$k(\boldsymbol{\rho}, \boldsymbol{\rho}') = \mathbb{E}\left[h(\boldsymbol{\rho})h^*(\boldsymbol{\rho}')\right]. \tag{1}$$

The autocorrelation function is usually called the kernel. Note that the kernel function of the GRF must be semi-positive definite. To enable CSGRF to represent the wireless channel, some restrictions should be imposed on $k(\boldsymbol{\rho}, \boldsymbol{\rho}')$ so that the $h(\boldsymbol{\rho})$ generated by it satisfies the EM propagation constraints. We use $h(\boldsymbol{\rho})$ to model the electric field distribution $\mathbf{E}(\boldsymbol{\rho}) : \mathbb{R}^4 \to \mathbb{C}^3$. Then, the autocorrelation function can be defined as $\mathbf{K}_{\mathbf{E}}(\boldsymbol{\rho}, \boldsymbol{\rho}') = \mathbb{E}\left[\mathbf{E}(\boldsymbol{\rho})\mathbf{E}(\boldsymbol{\rho}')^{\mathsf{H}}\right] \in \mathbb{C}^{3\times 3}$ [41]. Similarly, for a channel vector with N_{BS} components, it can also be modeled using CSGRF by constructing the autocorrelation function of $\boldsymbol{\rho}_n$ for $n = 1, 2, \ldots, N_{\mathrm{BS}}$.

B. Signal Model

For the signal model, the XL-MIMO system is considered, in which a single base station (BS) with N_{BS} antennas serves a single user with 1 antenna. We will try to solve the problem of uplink channel prediction in a narrowband system. Consider the simplest communication scenario, assuming we use an N_{BS}-antenna base station with fully digital precoding, where each antenna is connected to a dedicated radio frequency (RF) chain. The uplink signal model is

$$\mathbf{y}_t = \sqrt{p}\mathbf{h}_t + \mathbf{n}_t, \tag{2}$$

where $\mathbf{y}_t \in \mathbb{C}^{N_{\mathrm{BS}}\times 1}$ is the BS received pilots at time t, $\mathbf{h}_t \in \mathbb{C}^{N_{\mathrm{BS}}\times 1}$is the normalized channel vector satisfying $\mathbb{E}[\|\mathbf{h}_t\|^2] = N_{\mathrm{BS}}$, p denotes the pilot transmit power, and $\mathbf{n}_t$ is the complex-valued additive white Gaussian noise (AWGN) with zero mean and covariance $\sigma_n^2\mathbf{I}_{N_{\mathrm{BS}}}$.

The least squares (LS) and minimum mean square error (MMSE) channel estimation methods [42] can be used to estimate the channel. Let $\hat{\mathbf{h}}_t^{\mathrm{LS}}$ and $\hat{\mathbf{h}}_t^{\mathrm{MMSE}}$ represent the LS and MMSE estimation results of $\mathbf{h}_t$, respectively, and calculate them using the following two formulas:

$$\hat{\mathbf{h}}_t^{\mathrm{LS}} = \frac{1}{\sqrt{p}}\mathbf{y}_t, \tag{3}$$

$$\hat{\mathbf{h}}_t^{\mathrm{MMSE}} = \mathbb{E}\left[\mathbf{h}_t|\mathbf{y}_t\right] = \boldsymbol{\Sigma}_{\mathbf{h}_t}\left(\boldsymbol{\Sigma}_{\mathbf{h}_t} + \frac{1}{\mathsf{SNR}}\mathbf{I}_{N_{\mathrm{BS}}}\right)^{-1}\mathbf{y}_t, \tag{4}$$

where $\boldsymbol{\Sigma}_{\mathbf{h}_t} = \mathbb{E}\left\{\mathbf{h}_t\mathbf{h}_t^{\mathsf{H}}\right\}$ is the prior covariance matrix of channel. The symbol $\mathsf{SNR} = p/\sigma_n^2$ represents the received signal-to-noise ratio of the BS. In the subsequent analysis, for convenience, the pilot power p is set to 1.

III. Problem Formulation

In this section, the channel aging issue is illustrated, and the channel prediction problem is formulated to alleviate the channel aging.

As shown in Fig. 1, in the XL-MIMO communication system, the user moves at velocity $\mathbf{v}$, and the Doppler shift will cause significant differences in the channel at different times. We refer to the period of channel estimation as a frame, which contains N_{s} time slots. Channel estimation is only performed in the first slot. In mobile scenarios, because of the influence of the Doppler effect, except for the channel at the first slot, the actual time-varying channels of the follow-up slots may have significant differences from the channels obtained by the channel estimation, resulting in a decrease in the accuracy of the obtained CSI and thus affecting communication quality. Specifically, according to [43], the channel coherence time T_{c}

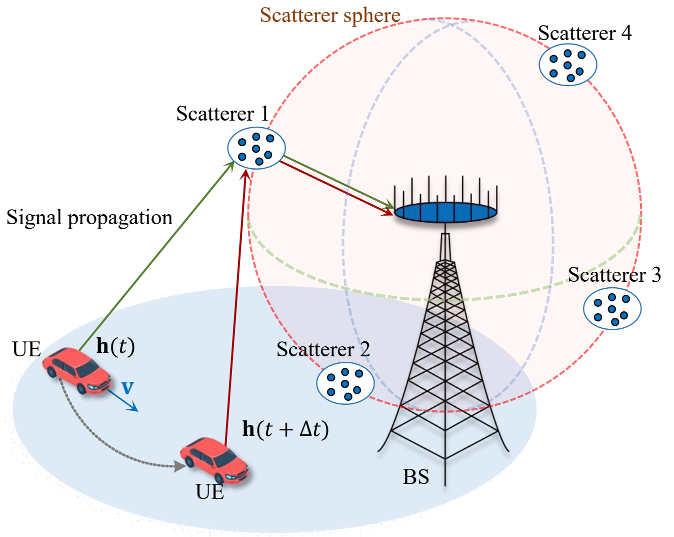

Fig. 1. The XL-MIMO communication system with scatterers distributed on the spherical surface surrounding the base station. User is in motion with velocity $\mathbf{v}$.

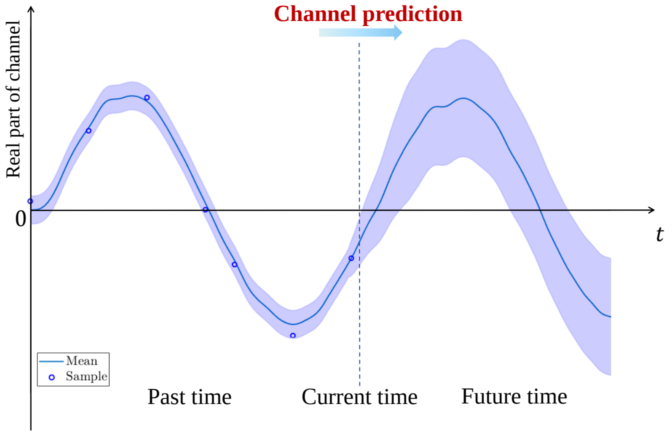

Fig. 2. An illustration of channel prediction: Taking a component of a channel vector as an example, represent the variation of the channel and its uncertainty over time

is defined as the time during which the channel can be well regarded as time-invariant, which is inversely proportional to the carrier frequency and user motion speed, i.e.,

$$T_{\mathrm{c}} \approx \frac{c}{2fv} = \frac{\lambda}{2v}, \tag{5}$$

where f is the carrier frequency, λ is carrier wavelength and v represents the user's moving speed. Channel coherence time is a rough estimate used to describe the time interval. Let $v_{\mathrm{r}} \leq v$ represent the radial velocity relative to the BS. The calculation of Doppler shift f_{d} is

$$f_{\mathrm{d}} = \frac{v_{\mathrm{r}}}{\lambda}. \tag{6}$$

The larger the Doppler frequency shift, the shorter the channel coherence time, and the more severe the channel aging. When the channel coherence time is shorter than the channel estimation period, using the channel estimation result of the first time slot for subsequent time slots will result in performance loss. The variations of the channel and its uncertainty due to imperfection of channel measurements over time are shown in Fig. 2. The solid curve represents the real part of a channel vector component, and the shadow area represents its uncertainty region. It can be observed that the channel uncertainty significantly increases at future time moments.

To solve the problem of severe channel aging mentioned above, some channel prediction methods have been proposed. The channel prediction is to obtain future channels through past channels. The existing channel prediction methods are typically based on sequential prediction. Specifically, it is to use the channels from frame 1 to frame T to predict the channel at frame $T+1$, and then use the channel at frame $T+1$ as known information to predict the channel at frame $T+2$ from frame 2 to frame $T+1$, and so on. However, due to errors in the channel prediction results at frame $T+1$, using it as a known channel to predict subsequent channels will bring errors to the subsequent predicted channels, which is the problem of error propagation.

To avoid performance loss caused by error propagation, unlike existing sequential channel prediction methods, we formulate the channel prediction problem in parallel form. That is, using the channel estimation results of the past L channels to predict the future channel of the next F channels. It should be noted that the channels of future F channels are predicted simultaneously. Considering the characteristics of the GRF channel, achieving accurate channel prediction requires an appropriate autocorrelation function, i.e., the kernel. We can then predict the future channel through inference based on this kernel. The appropriate kernel form will be discussed in the next section. Let $\boldsymbol{\omega} \in \boldsymbol{\Omega}$ denote model parameters of the kernel, and $\boldsymbol{\Omega}$ is the set of model parameters. $\mathbf{y} = (\mathbf{y}_1^{\mathsf{T}}, \mathbf{y}_2^{\mathsf{T}}, \ldots, \mathbf{y}_L^{\mathsf{T}})^{\mathsf{T}} \in \mathbb{C}^{N_{\mathrm{BS}}L\times 1}$ denotes the column vector composed of the received pilot sequences in the past L time frames. Let $\mathcal{L}$ denote the set of past channel indices and $\mathcal{F}$ denote the set of future channel indices. $\mathbf{h}_{\mathcal{L}} = (\mathbf{h}_1^{\mathsf{T}}, \mathbf{h}_2^{\mathsf{T}}, \ldots, \mathbf{h}_L^{\mathsf{T}})^{\mathsf{T}} \in \mathbb{C}^{N_{\mathrm{BS}}L\times 1}$ denotes the column vector composed of the previous L channels. $\mathbf{h}_{\mathcal{F}} = (\mathbf{h}_{L+1}^{\mathsf{T}}, \mathbf{h}_{L+2}^{\mathsf{T}}, \ldots, \mathbf{h}_{L+F}^{\mathsf{T}})^{\mathsf{T}} \in \mathbb{C}^{N_{\mathrm{BS}}F\times 1}$ denotes the column vector composed of F future channels that need to be predicted. By using the ML criterion to obtain kernel parameters and then using the MMSE criterion to predict future channels, the channel prediction problem can be formulated as

$$\begin{aligned} \hat{\boldsymbol{\omega}}(\mathbf{y}) &= \underset{\boldsymbol{\omega}\in\boldsymbol{\Omega}}{\arg\max} \left\{ \ln \int p(\mathbf{y}|\mathbf{h}_{\mathcal{L}}) p(\mathbf{h}_{\mathcal{L}}|\boldsymbol{\omega}) \mathrm{d}\mathbf{h}_{\mathcal{L}} \right\}, \\ \hat{\mathbf{h}}_{\mathcal{F}}(\mathbf{y}) &= \underset{\mathbf{h}_{\mathcal{F}}\in\mathbb{C}^{N_{\mathrm{BS}}F\times 1}}{\arg\max} \left\{ \ln p(\mathbf{y}|\mathbf{h}_{\mathcal{F}}) + \ln p(\mathbf{h}_{\mathcal{F}}|\hat{\boldsymbol{\omega}}(\mathbf{y})) \right\}. \end{aligned} \tag{7}$$

In (7), $\hat{\boldsymbol{\omega}}$ is the ML estimate of $\boldsymbol{\omega}$ and $\hat{\mathbf{h}}_{F}$ is the MMSE estimate of $\mathbf{h}_{F}$. Due to the characteristics of the GRF channel, MMSE estimation is equivalent to maximum a posteriori (MAP) estimation. $\mathbf{h}_t$ and $\mathbf{h}_{t+1}$ can be used to determine the channel of the n-th slot. For example, for the t-th frame, if $0 < n \leq N_{\mathrm{s}}/2$, then determine that the channel of time slot n is $\mathbf{h}_t$. Otherwise, it is determined as $\mathbf{h}_{t+1}$. In the following Section IV, we need to accurately solve the problem in (7).

IV. Proposed Spatio-Temporal Electromagnetic Kernel Learning Based Channel Prediction

In this section, we propose a parallel channel prediction scheme that simultaneously utilizes the temporal and spatial EM correlation between channels to improve the accuracy of channel prediction. Firstly, in Section IV-A, we introduce the construction of spatio-temporal EM kernel. Then, in Section IV-B, we introduce Gaussian process regression (GPR), it can be used to infer future channels. Moreover, the parameters of the spatio-temporal electromagnetic kernel need to be obtained through kernel learning as described in Section IV-C. In Section IV-D, we propose a grid-based electromagnetic mixed (GEM) kernel to further enhance reliability. Finally, in Section IV-E, the proposed GEM-KL channel prediction algorithm is elaborated.

A. Construction of STEM Correlation Function

To fully utilize the EM physical characteristics, it is essential to consider the fundamental physical principles behind the communication processes [44], including electromagnetics and information theory [45], [46]. The integration of these two theories could advance research in electromagnetic information theory (EIT), which provides insights into wireless communication issues from the perspective of electromagnetic wave propagation [47], [48]. We use the EIT-based channel model. Specifically, based on the channel model of the Gaussian random field described in subsection II-A, we analyze the characteristics of EM channels and their correlation. Electromagnetic information can be combined with the autocorrelation function of the channel [49]. We calculate the correlation integral of the electric field on the scatterer sphere S^2 shown in Fig. 1 to obtain the correlation function of the time-varying channel, i.e.,

$$\mathbf{K}(\mathbf{x},t;\mathbf{x}',t') \propto \int_{\hat{\boldsymbol{\kappa}}\in S^2} (\mathbf{I}-\hat{\boldsymbol{\kappa}}\hat{\boldsymbol{\kappa}}^{\mathsf{T}}) e^{\mathrm{i}k_0\hat{\boldsymbol{\kappa}}\cdot((\mathbf{x}-\mathbf{x}')+\mathbf{v}(t-t'))}\nu(\hat{\boldsymbol{\kappa}})\mathrm{d}S, \tag{8}$$

where the integration is carried out over the surface of the unit sphere S^2, $k_0 = 2\pi/\lambda_0$ is the wavenumber. $\hat{\boldsymbol{\kappa}}$ denotes the unit radial vector, and $\nu : S^2 \to \mathbb{R}_+$ denotes the angular power spectral density of the incident wave, with units of Watts per steradian per polarization. This function is also named as electromagnetic correlation function (EMCF) [49].

For time-varying channels, we incorporate the Doppler frequency shift into the EM correlation function by introducing the velocity vector $\mathbf{v}$, hence, this EMCF can also be referred to as the spatio-temporal kernel function (STEM-CF). To represent the incoming direction of electromagnetic waves, we use the von Mises-Fisher (vMF) distribution, i.e., $\nu(\hat{\boldsymbol{\kappa}}) = (\zeta^2/(8\pi))e^{\hat{\boldsymbol{\kappa}}\cdot\boldsymbol{\delta}}$. $\boldsymbol{\delta} \in \mathbb{C}^3$ is the concentration parameter, and its direction represents the direction in which the electromagnetic wave is concentrated. If the electromagnetic incidence is isotropic, $\nu(\hat{\boldsymbol{\kappa}})$ is a constant $\zeta^2/(8\pi)$. It should be noted that the larger the concentration, the stronger the channel sparsity.

Remark 1: In the construction of the EM kernel, the distribution of EM wave concentration represented by the parameter $\boldsymbol{\delta}$ on the spherical surface S^2 is called the von Mises Fisher (vMF) distribution [50], which is widely used for modeling wireless channels.

We can compute the closed-form expression [49] for STEM-CF as follows:

$$\begin{aligned}\mathbf{K}_{\mathrm{STEM}}(\boldsymbol{\rho},\boldsymbol{\rho}') &= \mathbb{E}\left[\mathbf{E}(\boldsymbol{\rho})\mathbf{E}(\boldsymbol{\rho}')^{\mathsf{H}}\right] \\ &= \frac{\zeta^2}{S(\|\boldsymbol{\delta}\|)}\boldsymbol{\Sigma}(\boldsymbol{\xi}),\end{aligned} \tag{9}$$

where $\mathbf{K}_{\mathrm{STEM}}$ is a 3×3 complex matrix, $\mathrm{tr}(\mathbf{K}_{\mathrm{EMCF}}(\boldsymbol{\rho},\boldsymbol{\rho}')) = \zeta^2$, $\boldsymbol{\xi} = k_0\mathbf{w} = k_0(\mathbf{x}-\mathbf{x}'+\mathbf{v}(t-t')) - \mathrm{i}\boldsymbol{\delta} \in \mathbb{C}^3$. $S(\delta) = \sinh(\delta)/\delta$ is an additional normalisation factor, where $\delta = \|\boldsymbol{\delta}\| \in \mathbb{R}_+$. We utilize the commonly used spherical Bessel functions $j_n(\xi)$ in 3D scenes to represent the correlation function $\boldsymbol{\Sigma}(\boldsymbol{\xi})$

$$\boldsymbol{\Sigma}(\boldsymbol{\xi}) = \frac{1}{6}(4j_0(\xi)-j_2(\xi))\mathbf{I}_3 + \frac{1}{2}(j_2(\xi)-2j_0(\xi))\hat{\boldsymbol{\xi}}\hat{\boldsymbol{\xi}}^{\mathsf{T}}, \tag{10}$$

where $\xi = |\boldsymbol{\xi}| = \sqrt{\boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{\xi}}$, and $\hat{\boldsymbol{\xi}} = \boldsymbol{\xi}/\xi$ denotes the normalized $\boldsymbol{\xi}$. The spherical Bessel function $j_n(\xi)$ is expressed as

$$j_n(\xi) = (-\xi)^n\left(\frac{1}{\xi}\frac{\mathrm{d}}{\mathrm{d}\xi}\right)^n\frac{\sin\xi}{\xi}, \tag{11}$$

It is important to note that $\mathbf{w} = \mathbf{x}-\mathbf{x}'+\mathbf{v}(t-t') - \mathrm{i}\boldsymbol{\delta}/k_0$ contains the spatial and temporal variables, which means that the correlation function we use is capable of describing the spatial and temporal correlation in an EM-consistent way. The proposed STEM-CF can be used as prior information in Gaussian process regression to address prediction problems, which will be discussed in the next subsection.

Remark 2: The kernel, as a function of the covariance matrix of the channel vector, contains the EM characteristics of the channels across spatial and temporal dimensions. By incorporating parameters such as the concentration $\boldsymbol{\delta}$ (indicating the direction of EM wave) and the user's motion velocity $\mathbf{v}$ (affecting the Doppler-induced temporal correlation), the kernel accurately reflects the channel's physical characteristics.

B. Gaussian Process Regression

Gaussian process regression (GPR) [51] can obtain predictions through prior information and observation data of GRF. Specifically, for the GRF $f(x) \sim \mathcal{GRF}(\mu(x), k(x,x'))$, GPR uses observation data $y_i = f(x_i) + n_i$, $n_i \sim \mathcal{CN}(0,\sigma_n^2), i = 1,2,\ldots,L_N$ to get a set of F-point prediction $\mathcal{F} = \{f(x_{L_N+1}), f(x_{L_N+2}),\ldots,f(x_{L_N+F_N})\}$. where $L_N = N_{\mathrm{BS}}L$ and $F_N = N_{\mathrm{BS}}F$.

The joint probability distribution of the observed and predicted joint vector $\mathbf{g} = [y_1, y_2, \ldots, y_L, f(x_{L_N+1}), f(x_{L_N+2}), \ldots, f(x_{L_N+F_N})]^{\mathsf{T}}$ satisfies

$$\mathbf{g} \sim \mathcal{CN}\left(\begin{bmatrix}\boldsymbol{\mu}_{\mathcal{L}}\\ \boldsymbol{\mu}_{\mathcal{F}}\end{bmatrix}, \begin{bmatrix}\mathbf{K}_{\mathcal{LL}}+\sigma_n^2\mathbf{I}_{L_N} & \mathbf{K}_{\mathcal{LF}}\\ \mathbf{K}_{\mathcal{FL}} & \mathbf{K}_{\mathcal{FF}}\end{bmatrix}\right), \tag{12}$$

where $\boldsymbol{\mu}_{\mathcal{L}} = [\mu(x_1), \mu(x_2), \ldots, \mu(x_{L_N})]^{\mathsf{T}}$ and $\boldsymbol{\mu}_{\mathcal{F}} = [\mu(x_{L_N+1}), \mu(x_{L_N+2}), \ldots, \mu(x_{L_N+F_N})]^{\mathsf{T}}$. The (m,n)-th entry of $\mathbf{K}_{\mathcal{LL}} \in \mathbb{C}^{L_N\times L_N}$ is $k(x_m,x_n)$, for all $m,n \in \{1,\ldots,L_N\}$. The (m,n)-th entry of $\mathbf{K}_{\mathcal{LF}} \in \mathbb{C}^{L_N\times F_N}$ is $k(x_m,x_n)$, for all $m \in \{1,\ldots,L_N\}$ and $n \in$

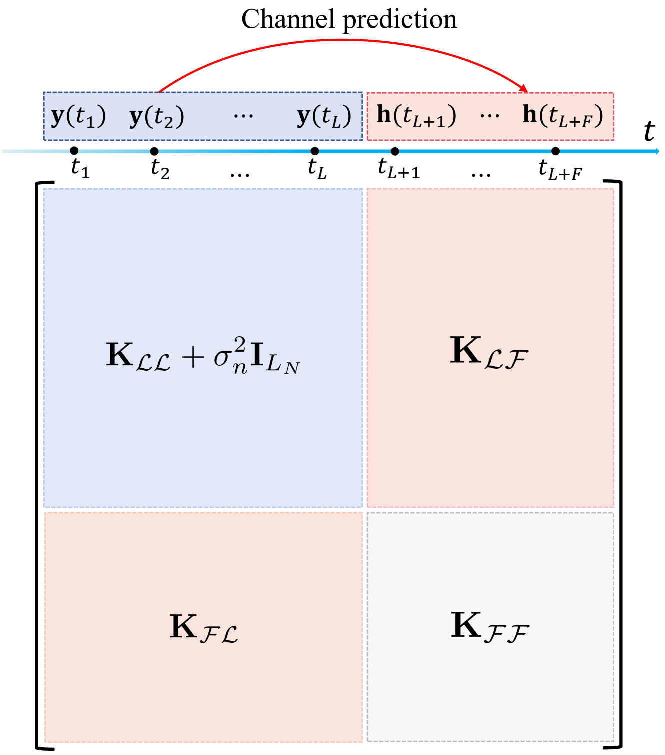

Fig. 3. Gaussian process regression for time domain channel prediction.

$\{L_N+1,\ldots,L_N+F_N\}$. $\mathbf{K}_{\mathcal{LF}} \in \mathbb{C}^{L_N\times F_N}$ and $\mathbf{K}_{\mathcal{FL}} = \mathbf{K}_{\mathcal{LF}}^{\mathsf{H}} \in \mathbb{C}^{F_N\times L_N}$. The (i,j)-th entry of $\mathbf{K}_{\mathcal{FF}} \in \mathbb{C}^{F_N\times F_N}$ is $k(x_i,x_j)$, for all $i,j\in\{L_N+1,\ldots,L_N+F_N\}$. We use $\mathbf{K}_{\mathbf{y}}$ to represent $\mathbf{K}_{\mathcal{LL}}+\sigma_n^2\mathbf{I}_{L_N}$. From the Gaussian posterior formula [52], we can obtain

$$\begin{aligned}\boldsymbol{\mu}_{\mathcal{F}|\mathcal{L}} &= \boldsymbol{\mu}_{\mathcal{F}} + \mathbf{K}_{\mathcal{LF}}^{\mathsf{H}}\mathbf{K}_{\mathbf{y}}^{-1}\mathbf{y},\\ \mathbf{K}_{\mathcal{F}|\mathcal{L}} &= \mathbf{K}_{\mathcal{FF}} - \mathbf{K}_{\mathcal{LF}}^{\mathsf{H}}\mathbf{K}_{\mathbf{y}}^{-1}\mathbf{K}_{\mathcal{LF}},\end{aligned} \tag{13}$$

The results of Bayesian regression are given by $\boldsymbol{\mu}_{\mathcal{F}|\mathcal{L}}$ and $\mathbf{K}_{\mathcal{F}|\mathcal{L}}$.

As shown in Fig. 3, the GPR-based channel predictor utilizes the covariance matrix of the channels from past frames (blue part) and the covariance matrix of the channels from between past and future frames (red part) to achieve parallel prediction of channels for multiple future frames. Since the prior distribution is a complex Gaussian distribution, the GPR predictor is consistent with the maximum a posteriori (MAP) predictor. Due to its Bayesian optimality and flexibility in adjusting kernel function parameters, GPR can be used for various estimation and prediction problems. Furthermore, the advantage of adjustable kernels makes GPR more widely used. The kernel adjustment measure will be introduced in subsection IV-C.

C. Kernel Learning

The kernel function $k(x,x')$ implicitly encodes the prior information of the Gaussian random field $f(x)$. This feature allows for more parameter configurations, thereby enhancing the model's ability to be adjusted. Choosing appropriate kernel parameters is an important step in constructing an effective regression model, which affects the accuracy of the kernel function in reconstructing Gaussian processes. The parameters that need to be adjusted in this process are usually referred to as hyperparameters. Assuming that the hyperparameters $\boldsymbol{\omega}\in\Omega\subset\mathbb{R}^{N_{\boldsymbol{\omega}}}$ of the adjustable kernel $k(x;x'|\boldsymbol{\omega})$ is also tunable. The process of finding the optimal hyperparameters for the STEM kernel is called kernel learning.

It is necessary to specify a criterion for evaluating whether hyperparameters are appropriate. The maximum likelihood (ML) criterion is a commonly used method, which can be expressed as

$$\hat{\boldsymbol{\omega}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\omega}\in\Omega} \ln p(\mathbf{y}|\boldsymbol{\omega}), \tag{14}$$

where the probability density function (PDF) of the pilot observation $\mathbf{y}$ under the condition of parameter $\boldsymbol{\omega}$ is expressed as

$$p(\mathbf{y}|\boldsymbol{\omega}) = \frac{1}{\pi^{L_N}\det\mathbf{K}_{\mathbf{y}}}\exp(-\mathbf{y}^{\mathsf{H}}\mathbf{K}_{\mathbf{y}}^{-1}\mathbf{y}). \tag{15}$$

The kernel $\mathbf{K}_{\mathbf{y}} = \mathbf{K}_{\mathbf{y}}(\boldsymbol{\omega}) = \mathbf{K}_{\mathcal{LL}}(\boldsymbol{\omega}) + \sigma_{\mathbf{h}}^2\mathbf{I}_{L_N}$ is a function of hyperparameter $\boldsymbol{\omega}$. Function $l(\boldsymbol{\omega}|\mathbf{y}) = \ln p(\mathbf{y}|\boldsymbol{\omega}) = -\ln\det\mathbf{K}_{\mathbf{y}} - (L_N+F_N)\ln\pi - \mathbf{y}^{\mathsf{H}}\mathbf{K}_{\mathbf{y}}^{-1}\mathbf{y}$ is the log-likelihood function. To obtain the maximum likelihood estimator of the hyperparameter $\boldsymbol{\omega}$, methods such as gradient descent, conjugate gradient descent, and Newton iteration can be used. All of these methods require the derivative of the log-likelihood function with respect to $\boldsymbol{\omega}$. The calculation result of this derivative is

$$\begin{aligned}\frac{\partial l(\boldsymbol{\omega}|\mathbf{y})}{\partial\omega_i} &= \frac{\partial}{\partial\omega_i}(-\ln\det\mathbf{K}_{\mathbf{y}} - \mathbf{y}^{\mathsf{H}}\mathbf{K}_{\mathbf{y}}^{-1}\mathbf{y})\\ &= \mathrm{tr}((\mathbf{g}\mathbf{g}^{\mathsf{H}} - \mathbf{K}_{\mathbf{y}}^{-1})\frac{\partial\mathbf{K}_{\mathbf{y}}}{\partial\omega_i}),\end{aligned} \tag{16}$$

where ω_i for $i=1,2,\ldots,N_{\boldsymbol{\omega}}$ represents each component of hyperparameter $\boldsymbol{\omega}$. For simplicity, let $\mathbf{g}=\mathbf{K}_{\mathbf{y}}^{-1}\mathbf{y}$. When the hyperparameter components are complex numbers, we need to consider the Wirtinger derivatives $(\partial/\partial\omega_{i,\mathrm{Re}} - \mathrm{i}\partial/\partial\omega_{i,\mathrm{Im}})/2$. Since $l(\boldsymbol{\omega}|\mathbf{y})$ is an analytic function of each elements of $\mathbf{K}_{\mathbf{y}}$, the derivative formula (16) remains unchanged.

For the STEM kernel capable of predicting four-dimensional spatio-temporal channels, we design its hyperparameters. According to Subsection IV-A, the concentration parameter $\boldsymbol{\delta}$ represents the direction and intensity of electromagnetic waves. Moreover, we introduce the velocity parameter $\mathbf{v}$ to describe the time-varying characteristics of the channel caused by user mobility. The channel energy is denoted as $\zeta_{\mathbf{h}}^2$. The Wirtinger derivatives of $\mathbf{K}_{\mathrm{STEM}}$ w.r.t. $\boldsymbol{\delta}(m)$, $\mathbf{v}(m)$ and $\zeta_{\mathbf{h}}^2$ are respectively expressed as

$$\begin{aligned}\frac{\partial\mathbf{K}_{\mathrm{STEM}}}{\partial\boldsymbol{\delta}(m)} &= -\frac{\zeta_{\mathbf{h}}^2}{S(\delta)}\left[\frac{S'(\delta)\boldsymbol{\delta}(m)}{S(\delta)\delta}\boldsymbol{\Sigma}(\boldsymbol{\xi}) + \mathrm{i}\frac{\partial\boldsymbol{\Sigma}(\boldsymbol{\xi})}{\partial\boldsymbol{\xi}(m)}\right],\\ \frac{\partial\mathbf{K}_{\mathrm{STEM}}}{\partial\mathbf{v}(m)} &= \frac{\zeta_{\mathbf{h}}^2 k_0(t_p-t_q)}{S(\delta)}\frac{\partial\boldsymbol{\Sigma}(\boldsymbol{\xi})}{\partial\boldsymbol{\xi}(m)},\\ \frac{\partial\mathbf{K}_{\mathrm{STEM}}}{\partial(\zeta_{\mathbf{h}}^2)} &= \frac{\boldsymbol{\Sigma}(\boldsymbol{\xi})}{S(\delta)},\end{aligned} \tag{17}$$

where $\delta=\|\boldsymbol{\delta}\|$, $\boldsymbol{\xi}=k_0\mathbf{w}$, $\mathbf{w}=\mathbf{x}_p-\mathbf{x}_q+\mathbf{v}(t_p-t_q)-\mathrm{i}\boldsymbol{\delta}/k_0$. The spherical Bessel functions of different orders have the following derivative relationship

$$\Big(\frac{1}{\xi}\frac{\mathrm{d}}{\mathrm{d}\xi}\Big)^a(\xi^{-b}j_b(\xi)) = (-1)^a\xi^{-b-a}j_{b+a}(\xi). \tag{18}$$

From the derivative property of spherical Bessel function (18), combined with correlation function formula (10), it can be inferred that

$$\begin{aligned}\frac{\partial\boldsymbol{\Sigma}(\boldsymbol{\xi})}{\partial\boldsymbol{\xi}(m)} =& \frac{1}{6}(-4j_1(\xi)-2\xi^{-1}j_2(\xi)+j_3(\xi))\hat{\boldsymbol{\xi}}(m)\mathbf{I}_3\\ &+\frac{1}{2}(2j_1(\xi)+2\xi^{-1}j_2(\xi)-j_3(\xi))\hat{\boldsymbol{\xi}}(m)\hat{\boldsymbol{\xi}}\hat{\boldsymbol{\xi}}^\mathsf{T}\\ &+\frac{1}{2}(-2j_0(\xi)+j_2(\xi))(\partial_m\hat{\boldsymbol{\xi}}\cdot\hat{\boldsymbol{\xi}}^\mathsf{T}+\hat{\boldsymbol{\xi}}\cdot\partial_m\hat{\boldsymbol{\xi}}^\mathsf{T}),\end{aligned} \tag{19}$$

where $\partial_m = \partial/\partial\boldsymbol{\xi}(m)$, $\xi = |\boldsymbol{\xi}|$ and $\hat{\boldsymbol{\xi}} = \boldsymbol{\xi}/\xi$. Moreover, $\partial_m\hat{\boldsymbol{\xi}} = \xi^{-1}(\hat{\mathbf{e}}_m - (\hat{\boldsymbol{\xi}}(m))\hat{\boldsymbol{\xi}})$ and $\hat{\mathbf{e}}_m$ denotes the unit vector which the only "1" is located at the m-th component. By combining (16) and (17), we can obtain the real-variable derivative which is expressed as

$$\frac{\partial l}{\partial\boldsymbol{\delta}(m)} = 2c\Re\left[\mathrm{tr}(\frac{\partial\mathbf{K}_{\mathcal{LL}}}{\partial\boldsymbol{\delta}(m)}(\mathbf{g}\mathbf{g}^\mathsf{H}-\mathbf{K}_\mathbf{y}^{-1}))\right], \tag{20}$$

and

$$\frac{\partial l}{\partial\mathbf{v}(m)} = 2c\Re\left[\mathrm{tr}(\frac{\partial\mathbf{K}_{\mathcal{LL}}}{\partial\mathbf{v}(m)}(\mathbf{g}\mathbf{g}^\mathsf{H}-\mathbf{K}_\mathbf{y}^{-1}))\right]. \tag{21}$$

$\mathbf{K}_{\mathcal{LL}}$ represents the channel correlation matrix constructed by the STEM method. Through gradient-based methods such as gradient ascent, these results can be used to obtain better $\boldsymbol{\omega}$ according to the ML criterion.

Remark 3: Kernel methods provide a powerful framework for capturing the spatio-temporal correlation of wireless channels [49]. The proposed spatio-temporal electromagnetic kernel is designed to encode physical insights from electromagnetic information theory into the channel's correlation function, enabling precise modeling of channel responses. Through kernel learning, these parameters are tuned to optimize the kernel, enhancing the accuracy of channel predictions via Gaussian process regression.

D. Proposed Grid-Based Electromagnetic Mixed Kernel

The gradient-based hyperparameter optimization method may get stuck in local optima. Fortunately, the grid-based electromagnetic mixed kernel (GEM) proposed in this subsection can achieve more global learning results.

Firstly, we analyze the objective function $l(\boldsymbol{\omega}|\mathbf{y})$, which can be intuitively represented as a function of the kernel $\mathbf{K}_\mathbf{y}$. However, $l(\boldsymbol{\omega}|\mathbf{y})$ is neither a convex nor a concave function of $\mathbf{K}_\mathbf{y}$. Therefore, gradient-based optimization methods are difficult to find the maximum value of $l(\boldsymbol{\omega}|\mathbf{y})$. Moreover, the kernel $\mathbf{K}_\mathbf{y}$ can be expressed as a function of the hyperparameter $\boldsymbol{\omega}$. Unfortunately, the components $\boldsymbol{\delta}, \mathbf{v}$ of $\boldsymbol{\omega}$ are not linearly related to $\mathbf{K}_\mathbf{y}$, making it difficult to directly characterize the relationship between $\boldsymbol{\omega}$ and $l(\boldsymbol{\omega}|\mathbf{y})$.

To avoid the inconvenience caused by the non-convexity/concavity of functions, the grid-based method [53] can be used in the parameter learning of the STEM kernel. We define $\mathbf{K}_{\mathrm{GEM}}$ as a combination of sub-kernels, and each of the sub-kernels corresponds to a grid point in the parameter space. In particular, several fixed values of $\boldsymbol{\delta}$ and $\mathbf{v}$ are taken as the selection values for the grid. By introducing the idea of the mixed kernel, we define k_{GEM} to be a combination of multiple sub-STEM kernels. We assume that there are N_k sub-correlation kernels and each of them has a weight of $c_n \in \mathbb{R}_+$, $n = 1, 2, \ldots, N_k$. Specifically, the GEM kernel function is designed as

$$\begin{aligned}&k_{\mathrm{GEM}}(\mathbf{x}_p, t_p; \mathbf{x}_q, t_q|\boldsymbol{\omega})\\ &\qquad= \mathbf{u}_p^\mathsf{T}\big(\sum_{n=1}^{N_k} c_n\mathbf{K}_{\mathrm{STEM}}(\mathbf{x}_p, t_p; \mathbf{x}_q, t_q|\boldsymbol{\omega}_n)\big)\mathbf{u}_q,\end{aligned} \tag{22}$$

where the value of each $k_{\mathrm{GEM}}(\mathbf{x}_p, t_p; \mathbf{x}_q, t_q|\boldsymbol{\omega}_n)$ is on the grid $(\boldsymbol{\delta}_n, \mathbf{v}_n)$. The pre-selected hyperparameters satisfy $\boldsymbol{\delta}_n \in \boldsymbol{\Delta}$ and $\mathbf{v}_n \in \mathbf{V}$, where $\boldsymbol{\Delta} \subset \mathbb{R}^3$ represents the set of concentration parameters on the grid points and $\mathbf{V} \subset \mathbb{R}^3$ represents the set of velocity parameters on the grid points. $\boldsymbol{\omega}_n \in \{\boldsymbol{\delta}_n, \mathbf{v}_n, c_n\}_{n=1}^{N_k} \subset \boldsymbol{\Omega}$ is the collection of all the hyperparameters $\boldsymbol{\omega}_n \in \boldsymbol{\Omega}$. The unit vector $\mathbf{u} \in \mathbb{R}^{3\times 1}$ denotes antenna polarization direction. Correspondingly, the components of the mixed correlation kernel matrix can be represented as

$$(\mathbf{K}_{\mathcal{LL},\mathrm{Mix}})_{p,q} = k_{\mathrm{GEM}}(x_p, t_p; x_q, t_q|\boldsymbol{\omega}). \tag{23}$$

The weight c_n is linearly related to the kernel $k_{\mathrm{STEM}}(\mathbf{x}_p, t_p; \mathbf{x}_q, t_q|\boldsymbol{\omega}_n)$ in the objective function $l(\boldsymbol{\omega}|\mathbf{y})$, so optimizing the weights $\{c_n\}_{n=1}^{N_k}$ corresponding to different $\boldsymbol{\delta}_n$ and $\mathbf{v}_n$ is sufficient to obtain the nearly optimal hyperparameters on the grid. We combine these weights into vector $\mathbf{c} = (c_1, c_2, \ldots, c_{N_k}) \in \mathcal{C} \subset \mathbb{R}^{N_k}$, where $\mathcal{C}$ denotes the set of N_k-dimensional non-negative vectors with the sum of elements equal to 1. The channel correlation matrix considering noise is represented as

$$\begin{aligned}\mathbf{K}_{\mathbf{y},\mathrm{Mix}} &= \mathbf{K}_{\mathcal{LL},\mathrm{Mix}} + \sigma_\mathbf{h}^2\mathbf{I}_{L_N}\\ &= \textstyle\sum_{n=1}^{N_k} c_n\mathbf{K}_{\mathcal{LL},n} + \sigma_\mathbf{h}^2\mathbf{I}_{L_N}.\end{aligned} \tag{24}$$

The mixed and grid-based kernel can improve the fitting ability of Gaussian random fields defined by STEM functions to channel observation data. Theoretically, a mixed kernel composed of a finite number of sub-correlation functions can represent the angular power spectrum of any incident electromagnetic field. The ML problem is simplified as

$$\hat{\mathbf{c}}_{\mathrm{ML}} = \arg\max_{\mathbf{c}\in\mathcal{C}} \ln p(\mathbf{y}|\mathbf{c}). \tag{25}$$

The log likelihood function is

$$\begin{aligned}l(\{c_n\}_{n=1}^{N_k}, \zeta^2|\mathbf{y}) =& \ln p(\mathbf{y}|\,\{c_n\}_{n=1}^{N_k})\\ =& -\ln\det\mathbf{K}_{\mathbf{y},\mathrm{Mix}} - \mathbf{y}^\mathsf{H}\mathbf{K}_{\mathbf{y},\mathrm{Mix}}^{-1}\mathbf{y}\\ &+\mathrm{const},\end{aligned} \tag{26}$$

where $l(\{\boldsymbol{\delta}_n, \mathbf{v}_n, c_n\}_{n=1}^{N_k}, \zeta^2|\mathbf{y})$ is the objective function. It is the fuction of $\mathbf{K}_{\mathbf{y},\mathrm{Mix}}$. Let $l_r(\{c_n\}_{n=1}^{N_k}, \zeta^2|\mathbf{y}) = \ln\det\mathbf{K}_{\mathbf{y},\mathrm{Mix}} + \mathbf{y}^\mathsf{H}\mathbf{K}_{\mathbf{y},\mathrm{Mix}}^{-1}\mathbf{y}$, we transform ML problems into finding the minimum value of the objective function to elim-

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2025.3635412

Algorithm 1 Proposed GEM Kernel Parameter Learning Algorithm.

```
Input: Number of sub-kernels $N_k$; grid hyperparameters $\{\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \ldots, \boldsymbol{\delta}_{N_k}\}$ and $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{N_k}\}$; Received pilots $\{y_1, y_2, \ldots, y_{L_N}\}$; Noise variance $\sigma_{\mathbf{h}}^2$; Maximum iteration number $M_{\text{iter}}$.
Output: Hyperparameters learning results $\{\boldsymbol{\delta}_n, \mathbf{v}_n, \hat{c}_n\}_{n=1}^{N_k}$; $\hat{\zeta}^2$.
Initialization: $\left\{c^{(0)}\right\}_{n=1}^{N_k}$, learning rates of Armijo-Goldstein's optimizer.
Set $m \leftarrow 0$.
Let $\mathbf{y} \in \mathbb{C}^{L_N \times 1}$ containing received pilots from $\{y_1, y_2, \ldots, y_{L_N}\}$.
for $m = 1, 2, \ldots, M_{\text{iter}}$ do
    Construct the GEM kernel $\mathbf{K}_{\mathbf{y},\text{Mix}}$ from hyperparameters $\left\{\boldsymbol{\delta}_n^{(m-1)}, \mathbf{v}_n^{(m-1)}, {c_n}^{(m-1)}\right\}_{n=1}^{N_k}$ by (22), (23) and (24).
    $\mathbf{g} \leftarrow \mathbf{K}_{\mathbf{y},\text{Mix}}^{-1}\mathbf{y}$
    for $n = 1, 2, \ldots, N_k$ do
        Construct surrogate function $l_s(c_n|c_n^{(m)})$ by (31).
        Compute $\frac{\partial l_s}{\partial c_n}$ from (33).
        Update $c_n^{(m)}$ from (31). by Armijo-Goldstein's optimizer.
        Update $\left\{c^{(m)}\right\}_{n=1}^{N_k}$ from $c_n^{(m)}$.
        Update $\mathbf{K}_{\mathbf{y},\text{Mix}}$ from $\left\{c^{(m)}\right\}_{n=1}^{N_k}$.
    end for
end for
$\hat{\zeta}^2 \leftarrow 2\sum_{\ell=1}^{L_N} |y_\ell|^2 / \left(L_N \cdot (1+\sigma_{\mathbf{h}}^2)\right)$
return Hyperparameter learning results $\{\boldsymbol{\delta}_n, \mathbf{v}_n, \hat{c}_n\}_{n=1}^{N_k}$, and $\hat{\zeta}^2$.
```

inate negative signs.

$$\hat{\mathbf{c}}_{\text{ML}} = \underset{\mathbf{c}\in\mathcal{C}}{\arg\min}(\ln\det\mathbf{K}_{\mathbf{y},\text{Mix}} + \mathbf{y}^{\mathsf{H}}\mathbf{K}_{\mathbf{y},\text{Mix}}^{-1}\mathbf{y}), \tag{27}$$

where $\mathbf{y}^{\mathsf{H}}\mathbf{K}_{\mathbf{y},\text{Mix}}^{-1}\mathbf{y}$ is a convex function of $\mathbf{K}_{\mathbf{y},\text{Mix}}$ and $\ln\det\mathbf{K}_{\mathbf{y},\text{Mix}}$ is a concave function of $\mathbf{K}_{\mathbf{y},\text{Mix}}$. The majorization-minimization (MM) algorithm [54] can be used to solve the optimal hyperparameters with non-convex and non-concave objective functions through an iterative scheme. Each iteration must minimize the designed surrogate function.

In the majorization step, we use the first-order Taylor expansion to design the surrogate function, which approximates the upper bound of the concave part of the function. Linearization of $\ln\det\mathbf{K}_{\mathbf{y},\text{Mix}}$ at $\mathbf{K}_{\mathbf{y},\text{Mix}} = \mathbf{K}_{\mathbf{y},\text{Mix}}^{(m)}$, i.e., $\mathbf{c} = \mathbf{c}^{(m)}$, yields the following inequality:

$$\begin{aligned} l_r(\mathbf{K}_{\mathbf{y},\text{Mix}}) \leq & \mathbf{y}^{\mathsf{H}}\mathbf{K}_{\mathbf{y},\text{Mix}}^{-1}\mathbf{y} + l_{\text{CCV}}(\mathbf{K}_{\mathbf{y},\text{Mix}}^{(m)}) \\ & + \operatorname{tr}\left(\nabla l_{\text{CCV}}(\mathbf{K}_{\mathbf{y},\text{Mix}}^{(m)})^{\mathsf{T}}(\mathbf{K}_{\mathbf{y},\text{Mix}} - \mathbf{K}_{\mathbf{y},\text{Mix}}^{(m)})\right), \end{aligned} \tag{28}$$

where $l_{\text{CCV}}(\mathbf{K}_{\mathbf{y},\text{Mix}}^{(m)}) = \ln\det\mathbf{K}_{\mathbf{y},\text{Mix}}^{(m)}$ and $\left(\nabla l(\mathbf{K})\right)_{ij} = \partial l / \partial \mathbf{K}_{ij}$. The Wirtinger derivative of l_{CCV} w.r.t. $\mathbf{K}_{\mathbf{y},\text{Mix}}$ is given by the following formula

$$\frac{\partial l_{\text{CCV}}}{\partial \mathbf{K}_{\mathbf{y},\text{Mix}}} = (\mathbf{K}_{\mathbf{y},\text{Mix}}^{-1})^*, \tag{29}$$

where $\mathbf{g} = \mathbf{K}_{\mathbf{y},\text{Mix}}^{-1}\mathbf{y}$. The real-variable derivative of l_{CCV} with respect to c_n is expressed as

$$\frac{\partial l_{\text{CCV}}}{\partial c_n} = 2\Re\left[\operatorname{tr}(\mathbf{K}_{\mathcal{LL},n}(\boldsymbol{\omega}_n)(\mathbf{K}_{\mathbf{y},\text{Mix}}^{-1}))\right], \tag{30}$$

Using formulas (28) and (30), the surrogate function l_s of the MM algorithm is written as

$$\begin{aligned} l_s(\mathbf{c}|\mathbf{c}^{(m)}) = & \mathbf{y}^{\mathsf{H}}\mathbf{K}_{\mathbf{y},\text{Mix}}^{-1}\mathbf{y} + \ln\det\mathbf{K}_{\mathbf{y},\text{Mix}}^{(m)} \\ & + 2\Re\left\{\operatorname{tr}\left[\left((\mathbf{K}_{\mathbf{y},\text{Mix}}^{(m)})^{-1}\right)(\mathbf{K}_{\mathbf{y},\text{Mix}} - \mathbf{K}_{\mathbf{y},\text{Mix}}^{(m)})\right]\right\}. \end{aligned} \tag{31}$$

Proof. The proof is provided in **Appendix B**. ■

Due to the high computational complexity of formula (31), which requires matrix inversion and determinant calculation, we perform the Cholesky decomposition on matrix $\mathbf{K}_{\mathbf{y},\text{Mix}}^{(m)}$ and $\mathbf{K}_{\mathbf{y},\text{Mix}}$. The resulting lower triangular matrix can be used for matrix inversion and determinant calculation, which can significantly reduce computational complexity.

Then, in the minimization step, the weight $\{c_n\}_{n=1}^{N_k}$ is updated through

$$\hat{\mathbf{c}}^{(m+1)} = \underset{\mathbf{c}\in\mathcal{C}}{\arg\min}(l_s(\mathbf{c}|\mathbf{c}^{(m)})), \tag{32}$$

the minimization step can be solved by finding the minimum value point of the convex function $l_s(\mathbf{c}|\mathbf{c}^{(m)})$, which requires the real-variable derivative of the surrogate function to c_n

$$\frac{\partial l_s}{\partial c_n} = 2\Re\left[\operatorname{tr}\left(\mathbf{K}_{\mathcal{LL},n}(\boldsymbol{\omega}_n)\left((\mathbf{K}_{\mathbf{y},\text{Mix}}^{(m)})^{-1} - \mathbf{g}\mathbf{g}^{\mathsf{H}}\right)\right)\right]. \tag{33}$$

These can be used for iteratively solving the optimal weight $\{c_n\}_{n=1}^{N_k}$ in the MM algorithm. The sequence $\left(l_r(\mathbf{c}^{(m)})\right)_{m\in\mathbb{N}}$ is non-increasing since

$$l_r(\mathbf{c}^{(m+1)}) \leq l_s(\mathbf{c}^{(m+1)}|\mathbf{c}^{(m)}) \leq l_s(\mathbf{c}^{(m)}|\mathbf{c}^{(m)}) = l_r(\mathbf{c}^{(m)}). \tag{34}$$

By iteratively executing the maximization and minimization steps, the MM algorithm ensures monotonic improvement of the objective function while avoiding non-convex optimization that leads to obtaining local optimal solutions. This approach transforms a non-convex optimization into a convex weight learning problem, ensuring stability while preserving EM physics. The intuition is similar to approximating a complex signal with a dictionary of basis functions—each sub-kernel acts as a "basis" for spatio-temporal correlations, and GEM-KL learns their optimal mixture. Therefore, GEM-KL performs better than STEM-KL.

The first term in the objective function (26) represents model complexity, while the second term represents data fitness. Kernel learning needs to balance these two factors. The process of maximizing the objective function l is capable of automatically balancing model complexity and data fitness. The GEM kernel parameter learning algorithm is summarized in **Algorithm 1**, and in the next subsection, we will summarize the overall GEM channel prediction algorithm.

It has been established in [55] that any continuous probability density function defined on the m-dimensional hypersphere S^m can be $\mathcal{L}^\infty-$approximated by finite mixtures of $m-$dimensional von Mises-Fisher (vMF) distributions, with

Algorithm 2 Channels Correlation Matrix Design.

```
Input: GEM hyperparameters $\boldsymbol{\omega} = \{\boldsymbol{\delta}_n, \mathbf{v}_n, \hat{c}_n\}_{n=1}^{N_k}$, channel indices $p \in \mathcal{P}$, $q \in \mathcal{Q}$, $p_{\min}$, $p_{\max}$, $q_{\min}$, $q_{\max}$.
Output: The correlation matrix between the channels in set $\mathcal{P}$ and the channels in set $\mathcal{Q}$: $\mathbf{K}_{\mathcal{PQ}}$.
Let $\mathbf{K}_{\mathcal{PQ}} \in \mathbb{C}^{|\mathcal{P}|\times|\mathcal{Q}|}$, $p = p_{\min}$, $q = q_{\min}$.
for $p = p_{\min}, p_{\min}+1, \ldots, p_{\max}$ do
    for $q = q_{\min}, q_{\min}+1, \ldots, q_{\max}$ do
        Calculate the GEM function: $\mathbf{K}_{pq} \leftarrow \mathbf{u}_p^{\mathsf{T}}\mathbf{K}_{\mathrm{GEM}}(\mathbf{x}_p, t_p; \mathbf{x}_q, t_q|\boldsymbol{\omega})\mathbf{u}_q$ according to (9) and (22).
    end for
end for
return The correlation matrix $\mathbf{K}_{\mathcal{PQ}}$.
```

Algorithm 3 Proposed EIT-GEM Channel Predictor.

```
Input: Past channel indices $l \in \mathcal{L}$; future channel indices $f \in \mathcal{F}$; Received pilots $y_l, l \in \mathcal{L}$; GEM hyperparameters $\boldsymbol{\omega}$; Noise variance $\sigma_n^2$.
Output: Channel prediction result $\hat{\mathbf{h}}_{\mathcal{F}}$.
Obtain GEM hyperparameters $\{\boldsymbol{\delta}_n, \mathbf{v}_n, \hat{c}_n\}_{n=1}^{N_k}$ according to Algorithm 1.
Compute the correlation matrix of past channels $\mathbf{K}_{\mathcal{LL}}$ and the correlation matrix between the past channels and the future channels $\mathbf{K}_{\mathcal{FL}}$ according to Algorithm 2.
$\mathbf{K}_{\mathbf{y},\mathrm{Mix}} = \mathbf{K}_{\mathcal{LL},\mathrm{Mix}} + \sigma_n^2\mathbf{I}_{L_N}$.
$\mathbf{g} \leftarrow \mathbf{K}_{\mathbf{y},\mathrm{Mix}}^{-1}\mathbf{y}$.
Reconstruct the predicted futrue channels $\hat{\mathbf{h}}_{\mathcal{F}} \leftarrow \mathbf{K}_{\mathcal{FL}}\mathbf{g}$ according to (37).
return The prediction result of vectorized future channels $\hat{\mathbf{h}}_{\mathcal{F}}$.
```

approximation accuracy arbitrarily constrained to $\epsilon > 0$. Therefore, for any EM incident density $\nu(S)$, the density can be approximated by a finite number of STEM-CFs to achieve any specified accuracy ϵ. This finite approximation corresponds to the mixed kernel method. Therefore, the design of the grid-based electromagnetic mixed (GEM) kernel can achieve any small approximation error.

Remark 4: In the GEM kernel learning, the acquisition of hyperparameters is mapped to the optimization of weights. The mapping utilizes the physical interpretability of the grid points $\boldsymbol{\omega}$, which encode EM propagation characteristics such as user mobility and the concentration direction of EM waves. Each weight c_n quantifies the relevance of the n-th sub-kernel to the observed data, effectively acting as a probabilistic measure of how closely $\boldsymbol{\omega}_n = (\mathbf{v}_n, \boldsymbol{\delta}_n)$ aligns with the true hyperparameters. The mixed kernel k_{GEM} thus aggregates contributions from all sub-kernels, weighted by their relevance to the observed data.

E. Proposed GEM-KL Channel Prediction Algorithm

We set the number of base station antennas to N_{BS}, assuming that these antennas are located at $\{\mathbf{x}_n\}_{n=1}^{N_{\mathrm{BS}}} \subset \mathbb{R}^3$. We consider the spacetime correlation tensor between the m-th polarization of antenna a at time t_i and the n-th polarization of antenna b at time t_j. Let $p = (a, m, i)$ and $q = (b, n, j)$, the correlation tensor can be expressed as

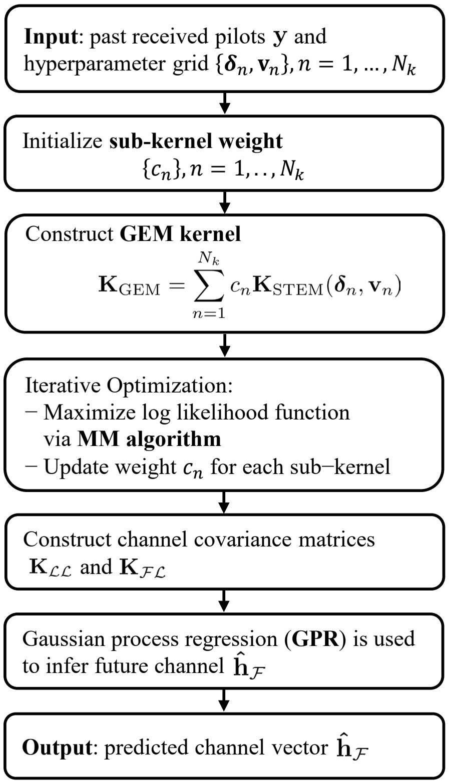
Input: past received pilots $\mathbf{y}$ and hyperparameter grid $\{\boldsymbol{\delta}_n, \mathbf{v}_n\}, n = 1, \ldots, N_k$
Initialize sub-kernel weight $\{c_n\}, n = 1, .., N_k$
Construct GEM kernel $\mathbf{K}_{\mathrm{GEM}} = \sum_{n=1}^{N_k} c_n \mathbf{K}_{\mathrm{STEM}}(\boldsymbol{\delta}_n, \mathbf{v}_n)$
Iterative Optimization:
− Maximize log likelihood function via MM algorithm
− Update weight c_n for each sub−kernel
Construct channel covariance matrices $\mathbf{K}_{\mathcal{LL}}$ and $\mathbf{K}_{\mathcal{FL}}$
Gaussian process regression (GPR) is used to infer future channel $\hat{\mathbf{h}}_{\mathcal{F}}$
Output: predicted channel vector $\hat{\mathbf{h}}_{\mathcal{F}}$

Fig. 4. The flowchart of the proposed GEM-KL channel prediction scheme.

$$\mathbf{K}_{p,q} = \mathbf{u}_p^{\mathsf{T}}\left[\mathbf{K}_{\mathrm{GEM}}(\mathbf{x}_p, t_p; \mathbf{x}_q, t_q)\right]\mathbf{u}_q, \tag{35}$$

where $\mathbf{u}_p$ represents the unit vector of antenna polarization direction. Based on formula (35), the correlation matrix between several channels in different time and space can be calculated, and the specific scheme is given by **Algorithm 2**. The proposed EIT-based GPR channel prediction method is summarized in **Algorithm 3**. Specifically, the BS receives noisy observations at any spatio-temporal coordinate at past times and predicts the channel at future times. In this algorithm, the unknown channels in the future or past time are modeled as a Gaussian random field. We need to first use GEM-KL method to STEM-CF to calculate the autocorrelation matrix $\mathbf{K}_{\mathbf{y}} = \mathbf{K}_{\mathcal{LL}} + \sigma_n^2\mathbf{I}_L$ of the channels at past times. And then calculate the correlation matrix between the past and future channels. Finally, we use (37) to obtain the future channels. For the convenience of understanding, the flowchart of the proposed channel prediction scheme is provided in Fig. 4.

While the proposed method is formulated for a single-antenna user equipment in a single-user scenario, it remains applicable to multi-antenna users and multi-user systems through straightforward extensions. In the case of multi-antenna users, forming a multiple-input multiple-output setup, the channel matrix can be reshaped into a vector, enabling the same Gaussian process regression framework with the spatio-temporal electromagnetic kernel to be used without altering its fundamental electromagnetic principles. For multi-user

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2025.3635412

systems, if users employ orthogonal pilots, the method allows for independent channel predictions per user. In scenarios with prominent interference from non-orthogonal access, the approach can jointly model channel correlation across users, capturing the correlations based on shared environmental factors like positions and velocities.

Furthermore, we provide a brief explanation of channel prediction in variable speed scenarios. we can handle variable speeds via periodic pilot updates. Both schemes support variable speed scenarios by employing a time-windowed prediction approach. Within each time window, the user's velocity is modeled as approximately constant, since abrupt velocity changes are infrequent in typical wireless communication scenarios, such as vehicular networks. At the end of each prediction window, the transmitter sends a new set of pilot signals, enabling the receiver to update the velocity parameter $\mathbf{v}$ in STEM-KL or adjust the sub-kernel weights $\mathbf{c}$ in GEM-KL, as described in Section IV-C and IV-D, respectively. This periodic pilot-based update mechanism ensures robust tracking of time-varying channels while balancing prediction accuracy and pilot overhead. By utilizing sparse pilot insertion, the proposed framework maintains accuracy and efficiency, making it well-suited for dynamic environments with varying user velocities.

Remark 5: While the proposed method is applicable to various MIMO configurations, it is particularly suitable for XL-MIMO systems. The large number of antennas increases the complexity of channel prediction due to the expansive spatial domain and rapid channel evolution under user mobility. Our proposed STEM and GEM kernel learning, using spatio-temporal correlations parameterized by velocity and concentration, offers a physics-informed solution to accurately predict the channel.

F. Computational Complexity Analysis

We analyze the computational complexity of the proposed STEM-KL, GEM-KL, and GPR-based channel prediction methods. The complexities are summarized in Table I, where N_{BS} denotes the number of antennas at the base station, M_{Siter} represents the number of iterations for the STEM-KL algorithm, M_{Giter} denotes the number of iterations for the GEM-KL algorithm. N_k represents the number of sub-kernels or basis functions, L is the number of past time slots used for training, and F is the number of future time slots for prediction. The explanation of the complexity formulas of the proposed methods is provided in **Appendix B**. As for the impact caused by parameters, STEM-KL and GEM-KL can converge with very few iterations, so there is no need for large M_{Siter} and M_{Giter}. The proposed methods only require a small number of historical channels, so L has little impact on complexity.

To further reduce the computational complexity of the proposed kernel learning-based channel prediction method, particularly the matrix inversion operations in kernel learning, several optimization strategies can be utilized:

TABLE I
COMPUTATIONAL COMPLEXITY

Algorithm	Complexity
AR	$\mathcal{O}(N_{\text{BS}}L^3F)$
PVEC	$\mathcal{O}(N_{\text{BS}}^3(L-1)^3)$
STEM-KL	$\mathcal{O}(M_{S\text{iter}}N_{\text{BS}}^3L^3)$
GEM-KL	$\mathcal{O}(M_{G\text{iter}}N_kN_{\text{BS}}^3L^3)$
GPR channel prediction	$\mathcal{O}(N_{\text{BS}}^2LF)$

- For channel correlation matrices, since it is positive semi-definite, *Cholesky decomposition* can be used to compute matrix inverses more efficiently. This method factorizes the matrix as $\mathbf{K} = \mathbf{L}\mathbf{L}^{\mathsf{H}}$, where $\mathbf{L}$ is lower triangular. The resulting lower triangular matrix $\mathbf{L}$ can be used for matrix inversion and determinant calculation, which can significantly reduce computational complexity.
- The matrix inversion and multiplication processes can be parallelized using algorithms such as divide-and-conquer [56] or block-based methods. Additionally, using graphics processing units (GPUs) can accelerate these operations through massive parallelism [57]. It is hopeful to accelerate the computation of large-scale matrices in practical applications.
- To approximate or avoid direct cubic-complexity inversions, iterative techniques can be adopted, including the Neumann series expansion and Newton iteration for refining approximations. Furthermore, diagonal band Newton iteration (DBNI) [58] can reduce the complexity from cubic to square by utilizing the diagonal dominance observed in $\mathbf{K}_{\mathbf{y},\text{Mix}}$.

Remark 6: It is worth noting that although kernel learning requires cubic complexity, it does not need to be performed frequently, meaning that not every channel prediction operation requires kernel learning. As a result, the complexity of the proposed method can be reduced in the average time sense.

V. SIMULATION RESULTS

The simulation results of STEM-KL and GEM-KL channel predictors are provided in this section. We evaluate the statistical learning performance of the proposed GEM covariance predictor by comparing it to the traditional methods.

A. Simulation Setup

In the following channel prediction simulation, to ensure the realism of the channel, we evaluated the performance of various prediction algorithms using the standard 3GPP TR 38.901 CDL model and ray tracing channel model [59], respectively. For the CDL channel model, the standard CDL-A delay profile is adopted.

The system parameter settings are as follows: The 256-element array is considered in simulations. The center of the antenna array is located at $(0, 0, 0)$, the array is located on the x-axis, and the user moves in the xoz plane. The carrier frequency is set to $f_c = 3.5$ GHz. The array is half-wavelength space. We set the period of transmitting pilot signals to 0.625 ms. The unit vector of antenna polarization direction is $\mathbf{u} = (0, 1, 0)^{\mathsf{T}}$.

All channel prediction algorithms are evaluated using normalized mean square error (NMSE) performance, which is defined in (36).

$$\text{NMSE} = \mathbb{E}\left[\frac{\|\hat{\mathbf{h}}_t - \mathbf{h}_t\|^2}{\|\mathbf{h}_t\|^2}\right]. \tag{36}$$

We also evaluated the achievable sum-rate performance of the proposed channel prediction method and baseline algorithm. The calculation of the achievable sum-rate is as follows

$$R = \log_2(1 + \frac{\left\|\hat{\mathbf{w}}^{\mathsf{H}}\mathbf{h}\right\|^2}{\sigma_{\mathbf{h}}^2 \left\|\hat{\mathbf{w}}\right\|^2}), \tag{37}$$

where $\hat{\mathbf{w}} = \hat{\mathbf{h}}/||\hat{\mathbf{h}}||$ represents the combiner.

Initialization for STEM-KL. In the STEM-KL algorithm, the concentration parameter $\boldsymbol{\delta}$ is set to $(0, 0, 0)$, representing an isotropic angular power spectrum with no directional preference. This choice avoids imposing prior assumptions on scattering geometry. The velocity parameter $\mathbf{v}$ is initialized as $(0, 0, 0)$. While simplistic, this initialization ensures reliable learning of velocity from observed Doppler shifts in the channel time series.

Initialization for GEM-KL. The GEM-KL algorithm employs a grid-based strategy with 15 sub-kernels, where hyperparameters are predefined to cover reasonable propagation scenarios. For concentration grids $\boldsymbol{\delta}_n$, five directions uniformly spanning angles between $-\pi/3$ and $\pi/3$ relative to the z-axis (user movement plane), each with a fixed magnitude $||\boldsymbol{\delta}_n|| = 10$. This design ensures coverage of concentration grids aligned with typical scattering environments. The velocity directions are set as $+x, -x, +z, -z$. The appropriate direction is selected by calculating the likelihood function, and the speeds are selected as 54km/h, 27km/h, and 0km/h. Therefore, there are a total of three velocity grid points. By optimizing the sub-kernel weights, the weighted summation of sub-kernels corresponding to different velocity parameters can obtain a covariance function containing appropriate velocity information. To ensure fairness, the initial values of all sub-kernel weights are the same, i.e., $c_n = 1/15$.

Baseline algorithms. The no-prediction NMSE is obtained by comparing the current channel with the future channel. The AR predictor is given by the autoregressive modeling [21]. The PVEC predictor is reproduced from the prony vector prediction method proposed in [15]. We also compared the performance of deep learning predictors and the proposed channel predictors, including the LSTM-based method [32] and the transformer-based method [11] in Fig. 7. The training and validation data samples are generated using the CDL channel model from 3GPP. To enhance generalization, user speeds v are randomly set between $72\,\mathrm{km/h}$ and $108\,\mathrm{km/h}$, producing 30,000 training samples, where SNR is randomly set between $0\,\mathrm{dB}$ and $10\,\mathrm{dB}$.

B. Simulation Results on Multipath CDL Channel

In this subsection, we compare the performance of traditional channel prediction schemes with the proposed STEM-based and GEM-based channel prediction schemes using the CDL-A channel model generated by Matlab 5G Toolbox.

First, we compare the NMSE performance of different methods for using the channels of the past two frames to predict the channel of the next frame, i.e., $L = 2$ and $F = 1$. The NMSE is plotted in Fig. 5 and Fig. 6 as a function of SNR. We set the maximum Doppler speeds to $36\,\mathrm{km/h}$ in Fig. 5 and $72\,\mathrm{km/h}$ in Fig. 6 (i.e. Doppler shifts of approximately $117\,\mathrm{Hz}$ and $233\,\mathrm{Hz}$). From simulation results, it can be seen that the channel prediction method based on kernel learning proposed in this article is significantly better than traditional methods across an SNR range of $-10 \sim 15$ dB, especially in low signal-to-noise ratio situations. The grid-based electromagnetic (GEM) kernel learning method can achieve the lowest NMSE among them. For example, when $\mathsf{SNR} = 2.5\,\mathrm{dB}$, compared with the AR channel prediction method, the GEM kernel learning channel prediction scheme can achieve NMSE performance gains of $5\,\mathrm{dB}$ and $4.5\,\mathrm{dB}$ for the next channel prediction at $v = 36\,\mathrm{km/h}$ and $v = 72\,\mathrm{km/h}$, where scalar $v = \|\mathbf{v}\|$ is the user's moving speed.

Regarding the baseline algorithms, the PVEC algorithm does not have an advantage in computational complexity. The prediction accuracy of PVEC is relatively low with an NMSE of approximately -5.2 when $\mathsf{SNR} = 2.5\,\mathrm{dB}$ and $v = 36\,\mathrm{km/h}$. Therefore, the proposed algorithms outperform the PVEC algorithm in terms of complexity and prediction accuracy. The AR algorithm has relatively low complexity because it does not involve spatial correlation of channels, but its prediction accuracy is also low, basically close to no prediction method. Such low prediction accuracy is generally unacceptable in wireless communication systems. Therefore, considering the trade-off between prediction accuracy and computational complexity, our proposed algorithm is more suitable for channel prediction problems compared to baseline algorithms.

Compared to other channel prediction algorithms, the reason why the electromagnetic kernel-based scheme performs better is mainly because the electromagnetic prior information is successfully embedded in the STEM-CF covariance model used, so the prior information provided by the electromagnetic kernel is more accurate, thus enabling more accurate channel prediction. The performance of EM channel prediction methods with kernel learning is superior to all baseline methods. Because kernel learning-based channel prediction methods can obtain more accurate model hyperparameters through learning, allowing EM kernels to better fit the direct covariance function of the channel and provide more accurate prior information. The GEM kernel learning scheme outperforms all rivals mainly because it solves the problem of EM kernel learning methods falling into local optima during hyperparameter learning. It transforms the optimization of concentration $\boldsymbol{\delta}$ and user speed $\mathbf{v}$ into the optimization of weights for kernels composed of different $\boldsymbol{\delta}$ and $\mathbf{v}$. The mixed kernel approach can better adapt to multipath channels and has a strong ability

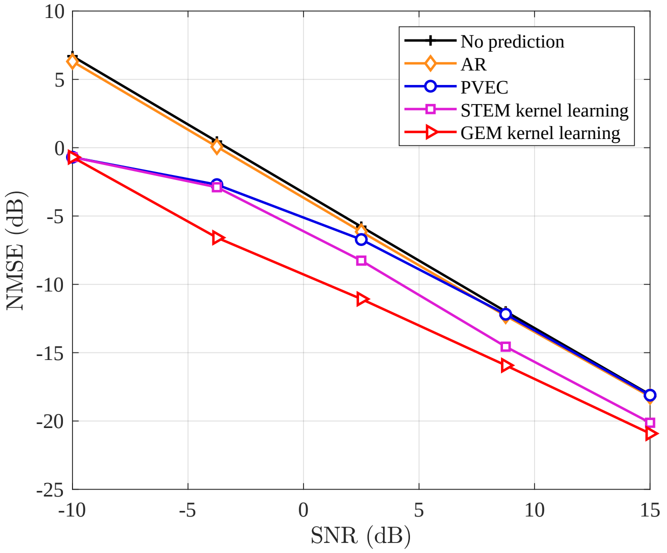

Fig. 5. Comparison of the NMSE performance versus SNR between the proposed EM kernel learning channel prediction method and traditional channel prediction schemes in CDL channel scenarios at the maximum Doppler velocity of $36\,\mathrm{km/h}$.

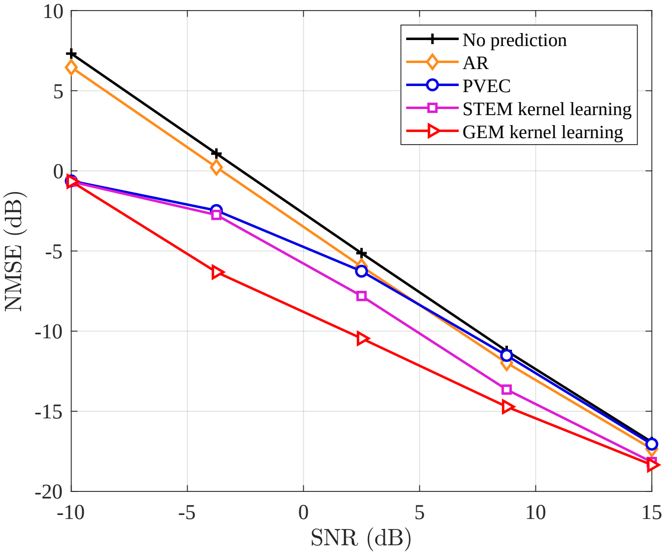

Fig. 6. Comparison of the NMSE performance versus SNR between the proposed EM kernel learning channel prediction method and traditional channel prediction schemes in CDL channel scenario at the maximum Doppler velocity of $72\,\mathrm{km/h}$.

to match electromagnetic correlation patterns in received pilots in the past. Therefore, the prior information of GEM is more accurate, resulting in more stable and precise performance.

To investigate the performance changes of the algorithm over time, we use the channels of the past two frames to predict the channels of the next five frames, that is, $L = 2$ and $F = 5$. We observe the NMSE performance of the channels predicted by different schemes at different frames through simulation. When $\mathsf{SNR} = 5\,\mathrm{dB}$ and $v = 72\,\mathrm{km/h}$, the corresponding performance comparison simulation results are shown in Fig. 7. The different simulation points represent the NMSE of channel prediction for different future frames. From the simulation results, we can observe that when predicting several future channels using a small number of past time channels, the NMSE performance of the EM kernel-based channel prediction algorithm is far superior to the AR and PVEC algorithms. Although the transformer-based method exceeds the STEM-KL method, GEM-KL's most accurate channel prediction result compensates for this. Among them, the proposed GEM kernel learning method performs the best. Taking the prediction of the channel in the second future frame as an example, the GEM-KL scheme proposed in this paper improves the NMSE performance by $3.4\,\mathrm{dB}$ respectively compared to the AR channel prediction method. Furthermore, the GEM-KL scheme outperforms the transformer-based method by $0.74\,\mathrm{dB}$.

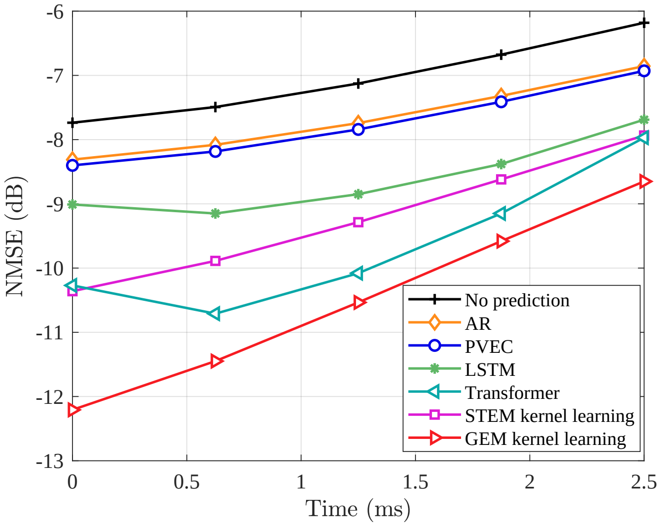

Fig. 7. The NMSE performance versus time in CDL channel scenario at the maximum Doppler velocity of $72\,\mathrm{km/h}$.

We also simulate the achievable sum-rate performance of the channel predictor over time to evaluate the effectiveness of the channel prediction algorithm. The upper bound of achievable sum-rate performance refers to the assumption that the accurate channel state information is known, which is the perfect CSI in the figures. When $\mathsf{SNR} = 5\,\mathrm{dB}$ and maximum Doppler speed is set to $72\,\mathrm{km/h}$, as shown in Fig. 8, at all times, the achievable sum-rate of the STEM-KL and GEM-KL channel prediction schemes is higher than that of other channel prediction algorithms, which indicates the effectiveness of the proposed channel prediction schemes.

In addition, we evaluate the performance of the proposed channel prediction scheme at different user movement speeds. The user speed ranges from $20\,\mathrm{m/s}$ $(72\,\mathrm{km/h})$ to $100\,\mathrm{m/s}$ $(360\,\mathrm{km/h})$. When $\mathsf{SNR} = 5\,\mathrm{dB}$, we select different channel prediction schemes to predict the channel of the next frame using the channels of the past two frames. The simulation results are shown in Fig. 9. It is easy to observe that the proposed STEM-KL and GEM-KL channel prediction schemes have significant NMSE performance advantages compared to the baseline algorithms. Among them, the GEM-KL channel prediction scheme has the lowest NMSE in all speed scenarios. Taking the scenario in which the user speed is $216\,\mathrm{km/h}$ as an example, the proposed GEM-KL channel prediction method has a $2.1\,\mathrm{dB}$ NMSE performance advantage compared to the AR channel prediction algorithm.

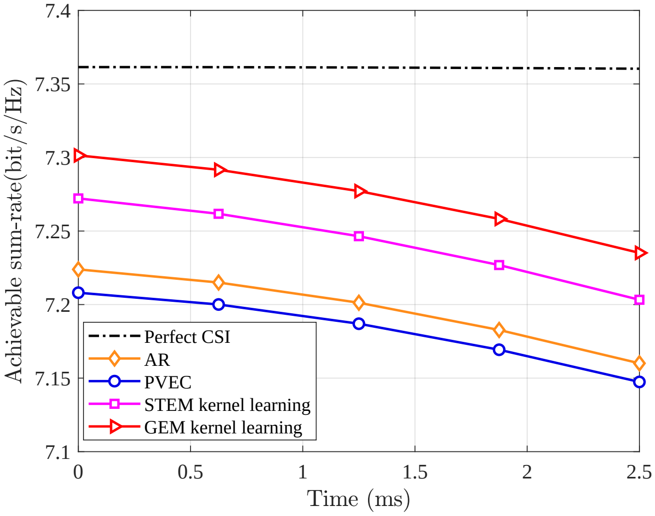

Fig. 8. The achievable sum-rate performance versus time in CDL channel scenario at the maximum Doppler velocity of $72\,\mathrm{km/h}$.

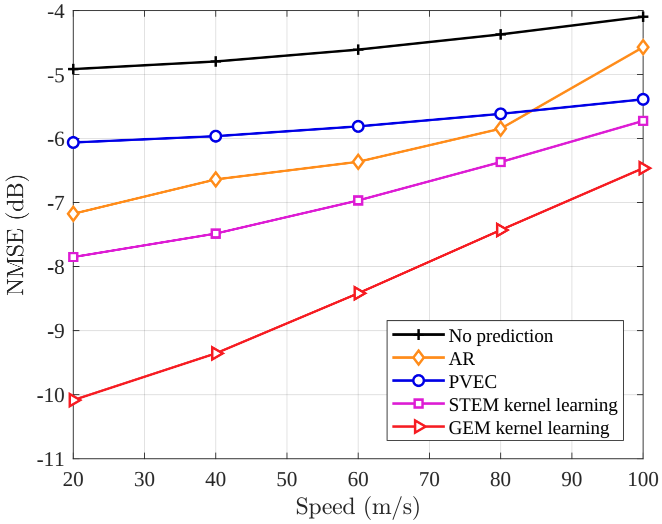

Fig. 9. The NMSE performance versus speed in CDL channel scenario.

It is worth noting that when predicting multiple future time channels, as time increases, the NMSE growth of the STEM kernel-based methods is slower compared to the baseline method, indicating more stable performance. This is because the channel prediction methods based on the STEM kernel can achieve parallel prediction of channels at multiple time points, avoiding the propagation of prediction errors.

By summarizing the simulation results of Fig. 5∼9, it can be concluded that the proposed GEM kernel learning method can achieve higher accuracy in predicting future channels. In addition, this scheme can effectively alleviate the negative impact of user mobility on wireless communication.

C. Simulation Results on the Ray Tracing Channel

In order to evaluate the performance of the proposed channel prediction scheme in more practical scenarios, we choose the ray tracing channel [59] for simulation. Both the base station and the users are located in Hong Kong. The path between the base station and the users adopts the ray tracing scheme.

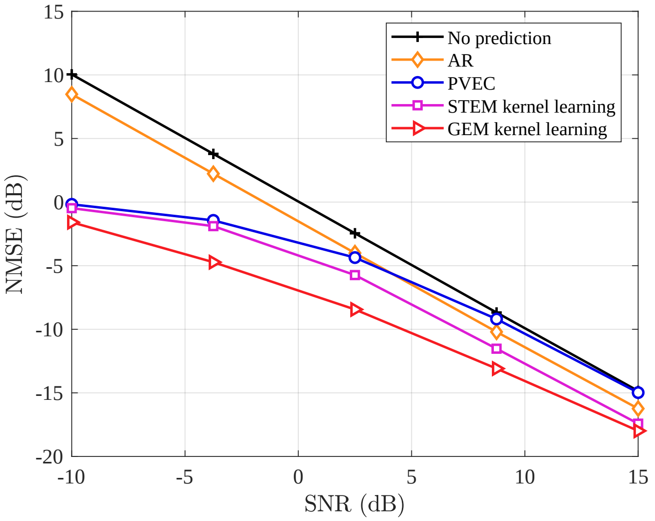

Fig. 10. Comparison of the NMSE performance versus SNR between the proposed EM kernel learning channel prediction method and traditional channel prediction schemes in the ray tracing channel scenario at the maximum Doppler velocity of $72\,\mathrm{km/h}$.

The trends of NMSE versus SNR for different channel prediction schemes are plotted in Fig. 10. Let $L = 2$ and $F = 1$. From the simulation results, it can be observed that several STEM-based channel prediction schemes perform better than no prediction scheme, AR scheme, and PVEC scheme in ray tracing channel scenarios with maximum Doppler velocities of $72\,\mathrm{km/h}$ (i.e. Doppler shifts of approximately $233\,\mathrm{Hz}$), respectively. It can be seen that the GEM-KL method can achieve the best performance.

For example, at $\mathsf{SNR} = 2.5\,\mathrm{dB}$, compared to the AR scheme, the GEM kernel learning scheme can achieve NMSE performance gains of approximately $4.4\,\mathrm{dB}$ at $72\,\mathrm{km/h}$, respectively.

In addition, we demonstrate the temporal variation of NMSE performance corresponding to different channel prediction schemes in Fig. 11. When $\mathsf{SNR} = 5\,\mathrm{dB}$ and the duration of one frame is $0.75\,\mathrm{ms}$, the channels from the previous two frames are used to predict the channels for the next five frames. We can observe that the proposed parallel channel prediction scheme based on GEM kernel learning also has the best NMSE performance in predicting the channels of subsequent time frames. Taking the prediction of the channel for the second future frames as an example, compared with the AR channel prediction scheme, the proposed GEM-KL method improves NMSE performance by $4.2\,\mathrm{dB}$ in the scenario of maximum Doppler velocity $72\,\mathrm{km/h}$.

The achievable sum-rate performance of the channel predictor over time is simulated in Fig. 12. When $\mathsf{SNR} = 5\,\mathrm{dB}$, the duration of one frame is $0.75\,\mathrm{ms}$ and maximum Doppler speed is set to $72\,\mathrm{km/h}$, For the next five frames, the achievable sum-rate of the STEM-KL and GEM-KL channel prediction schemes is higher than that of other channel prediction algorithms, which indicates the effectiveness of the proposed channel prediction schemes.

Moreover, the performance of the proposed channel prediction scheme at different user movement speeds is evaluated.

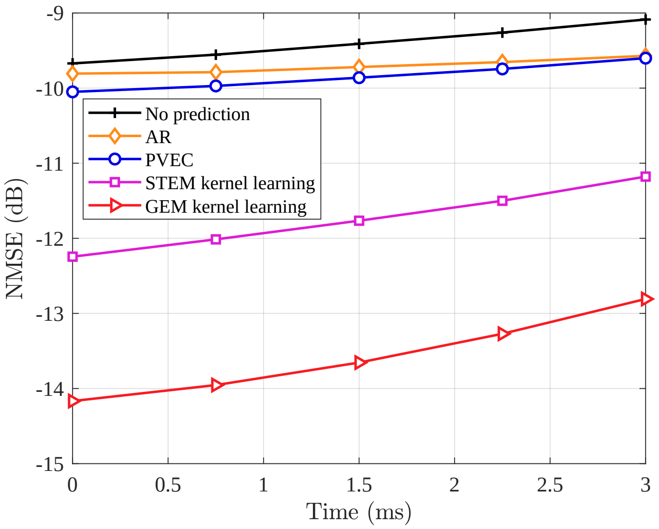

Fig. 11. The NMSE performance versus time in ray tracing channel scenario at the maximum Doppler velocity of $72\,\mathrm{km/h}$.

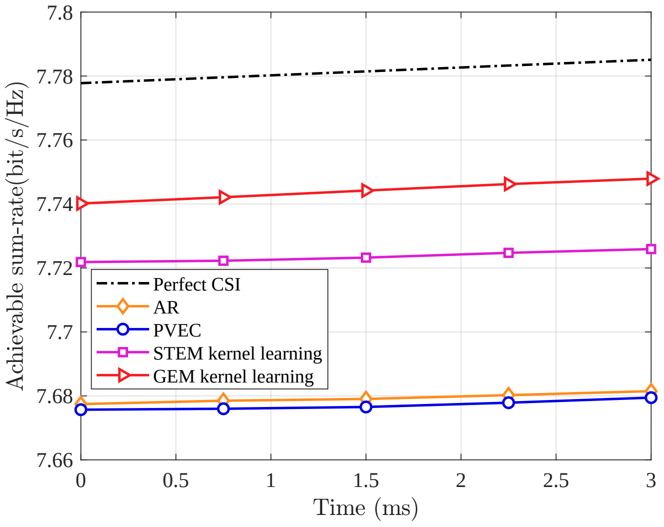

Fig. 12. The achievable sum-rate performance versus time in ray tracing channel scenario at the maximum Doppler velocity of $72\,\mathrm{km/h}$.

The user speed also ranges from $72\,\mathrm{km/h}$ to $360\,\mathrm{km/h}$. We set $\mathrm{SNR} = 5\,\mathrm{dB}$, different channel prediction schemes are used to predict the channel of the next frame using the channels of the past two frames. As plotted in Fig. 13, the simulation result shows that the proposed STEM-KL and GEM-KL channel prediction schemes have significant NMSE performance advantages compared to baseline algorithms. Among them, the GEM-KL channel prediction scheme has the lowest NMSE in all speed scenarios. Taking the scenario where the user speed is $60\,\mathrm{m/s}$ ($216\,\mathrm{km/h}$) as an example, the proposed GEM-KL channel prediction method has a $2.7\,\mathrm{dB}$ NMSE performance advantage compared to the AR channel prediction algorithm.

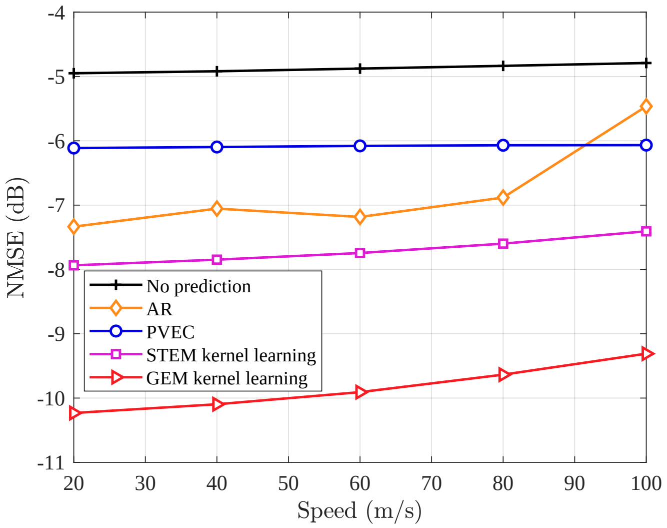

Fig. 13. The NMSE performance versus speed in ray tracing channel scenario.

The above simulation results have demonstrated that, on the ray tracing channel, the channel prediction schemes based on STEM-KL can achieve better NMSE performance. It has two advantages over traditional channel prediction algorithms. On the one hand, compared to other representations of channel correlation, the EM kernel can better describe the spatio-temporal correlation of the channel. On the other hand, the STEM-KL channel prediction method can predict multiple future channels in parallel, avoiding the accumulation of errors caused by sequential prediction. The advantage of STEM with kernel learning is that it can find better hyperparameters concentration $\boldsymbol{\delta}$ and user motion velocity $\mathbf{v}$ for the EM kernel, which makes the STEM kernel more accurate in reflecting the spatio-temporal correlation of the channel, and therefore performs better than baseline methods. However, using gradient descent-based learning methods to obtain hyperparameters relies heavily on initial values. If the initial values are not good, the learning results may be locally optimal hyperparameters. Fortunately, the proposed GEM-KL scheme solves this problem by combining the kernels of different $\boldsymbol{\delta}$ and $\mathbf{v}$ grid points according to the learned optimal weights, which can avoid the problem of hyperparameter local optima and make channel prediction performance more stable. Therefore, the GEM kernel learning GPR channel predictor performs best among all compared schemes.

VI. Conclusions

In this paper, we designed a high-accuracy channel predictor by STEM kernel learning, for XL-MIMIO scenarios. The STEM correlation function can capture the fundamental propagation characteristics of the wireless channel, making it suitable as a kernel function that incorporates prior information. We designed the hyperparameters of the STEM kernel, including user velocity and concentration to fit time-varying channels. The hyperparameters are obtained through kernel learning. Then, the future channels are predicted through GPR. To further improve the stability of channel prediction, we proposed a GEM-KL channel predictor. The STEM kernel is approximated by a grid-based EM mixed (GEM) kernel, which is composed of STEM sub-kernels. Moreover, multi-kernel schemes are more suitable for multipath channel prediction. Finally, we conducted numerical tests on the proposed schemes using the CDL channel model and the ray tracing channel model. The STEM-KL methods achieve improved performance over other baseline methods, and the GEM-KL method outperforms all compared methods.

In future research, we will use the GEM-KL scheme to investigate other, more complex channel prediction problems, such as frequency-domain wideband channel prediction.

Appendix A
Proof of the surrogate function for MM algorithm

The upper bound for the concave $\ln\det\mathbf{K}_{\mathbf{y},\mathrm{Mix}}(\mathbf{c})$ is:

$$\begin{aligned}\ln\det\mathbf{K}_{\mathbf{y},\mathrm{Mix}}(\mathbf{c}) \leq & \ln\det\mathbf{K}_{\mathbf{y},\mathrm{Mix}}^{(m)} \\ & + \sum_{n=1}^{N_k}(c_n - c_n^{(m)})\cdot\left.\frac{\partial(\ln\det\mathbf{K}_{\mathbf{y},\mathrm{Mix}})}{\partial c_n}\right|_{\mathbf{c}^{(m)}},\end{aligned} \tag{38}$$

using the gradient from (30), the upper bound can be expressed as

$$\begin{aligned}\ln\det\mathbf{K}_{\mathbf{y},\mathrm{Mix}}(\mathbf{c}) \leq & \ln\det\mathbf{K}_{\mathbf{y},\mathrm{Mix}}^{(m)} \\ & + 2\Re\left[\sum_{n=1}^{N_k}(c_n - c_n^{(m)})\mathrm{tr}\left(\mathbf{K}_{\mathcal{LL},\mathrm{n}}(\mathbf{K}_{\mathbf{y},\mathrm{Mix}}^{(\mathrm{m})})^{-1}\right)\right].\end{aligned} \tag{39}$$

Since $\sum_{n=1}^{N_k}(c_n - c_n^{(m)})\mathbf{K}_{\mathcal{LL},n} = \mathbf{K}_{\mathbf{y},\mathrm{Mix}} - \mathbf{K}_{\mathbf{y},\mathrm{Mix}}^{(m)}$, this simplifies to

$$\begin{aligned}\ln\det\mathbf{K}_{\mathbf{y},\mathrm{Mix}}(\mathbf{c}) \leq & \ln\det\mathbf{K}_{\mathbf{y},\mathrm{Mix}}^{(m)} \\ & + 2\Re\left[\mathrm{tr}\left((\mathbf{K}_{\mathbf{y},\mathrm{Mix}}^{(\mathrm{m})})^{-1}(\mathbf{K}_{\mathbf{y},\mathrm{Mix}} - \mathbf{K}_{\mathbf{y},\mathrm{Mix}}^{(\mathrm{m})})\right)\right].\end{aligned} \tag{40}$$

Substituting the upper bound into l_r, we can get the surrogate function (31). This completes the proof.

Appendix B
Explanation of the complexity formulas

The STEM-KL algorithm involves iterative gradient-based optimization of velocity parameter $\mathbf{v}$ and concentration parameter $\boldsymbol{\delta}$. For N_{BS} antennas and L historical time frames, in the process of calculating the likelihood function as shown below, we need to perform inverse operation on the matrix $\mathbf{K}_{\mathbf{y}} \in \mathbb{C}^{N_{\mathrm{BS}}L\times N_{\mathrm{BS}}L}$

$$\begin{aligned}l(\mathbf{v},\boldsymbol{\delta}|\mathbf{y}) =& \ln p(\mathbf{y}|\mathbf{v},\boldsymbol{\delta}) \\ =& -\ln\det\mathbf{K}_{\mathbf{y}} - \mathbf{y}^{\mathsf{H}}\mathbf{K}_{\mathbf{y},}^{-1}\mathbf{y} + \mathrm{const},\end{aligned} \tag{41}$$

each iteration requires a likelihood function calculation, and the complexity of other calculations is not higher than that of the inverse operation. The number of iterations is $M_{S\mathrm{iter}}$, so the total computational complexity is $\mathcal{O}(M_{S\mathrm{iter}}N_{\mathrm{BS}}^3L^3)$.

GEM-KL does not require alternating iterations to optimize speed and concentration parameters. However, it requires iterative optimization of the weights of N_k sub-kernels. Each iteration requires computing surrogate functions (31) N_k times, with the computational complexity mainly derived from the cubic complexity caused by matrix inversion. Compared to it, other computational complexities can be ignored. Therefore, the overall complexity calculation is $\mathcal{O}(M_{G\mathrm{iter}}N_kN_{\mathrm{BS}}^3L^3)$.

After obtaining the channel correlation matrix through STEM-KL or GEM-KL, we use the Gaussian posterior formula $\hat{\mathbf{h}}_{\mathcal{F}} = \mathbf{K}_{\mathcal{FL}}\mathbf{K}_{\mathbf{y}}^{-1}\mathbf{y}$ to infer the future channels. The operation $\mathbf{K}_{\mathbf{y}}^{-1}\mathbf{y}$ has completed during the process of kernel learning. Since $\mathbf{K}_{\mathcal{FL}} \in \mathbb{C}^{N_{\mathrm{BS}}F\times N_{\mathrm{BS}}L}$ and $\mathbf{g} = \mathbf{K}_{\mathbf{y}}^{-1}\mathbf{y} \in \mathbb{C}^{N_{\mathrm{BS}}L\times 1}$. Therefore, the complexity of channel prediction based on GPR is the complexity of multiplication calculation $\mathbf{K}_{\mathcal{FL}}\mathbf{g}$, i.e., $\mathcal{O}(N_{\mathrm{BS}}^2LF)$.

References

[1] Z. Wang, J. Zhang, H. Du, D. Niyato, S. Cui, B. Ai, M. Debbah, K. B. Letaief, and H. V. Poor, "A tutorial on extremely large-scale MIMO for 6G: Fundamentals, signal processing, and applications," *IEEE Commun. Surv. Tutorials*, Jan. 2024.

[2] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2012.

[3] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[4] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[5] Z. Xiao, Z. Zhang, Z. Chen, Z. Yang, C. Huang, and X. Chen, "From data-driven learning to physics-inspired inferring: A novel mobile MIMO channel prediction scheme based on neural ODE," *IEEE Trans. Wireless Commun.*, Dec. 2023.

[6] Y. Zhang, J. Zhang, and L. Yu, "Cluster-based fast time-varying MIMO channel fading prediction in the high-speed scenario," *IEEE Access*, vol. 7, pp. 148 692–148 705, Oct. 2019.

[7] Z. Zhang, Y. Zhang, J. Zhang, and F. Gao, "Adversarial training-aided time-varying channel prediction for TDD/FDD systems," *China Commun.*, vol. 20, no. 6, pp. 100–115, Jun. 2023.

[8] *Radio Resource Control (RCC) Protocol Specification*, 3GPP Std., Jun. 2019, 3GPP TS 38.331, Version 16.5.0.

[9] L. Yuan, F. Zhou, Q. Wu, and D. W. K. Ng, "Channel prediction-enhanced intelligent resource allocation for dynamic spectrum-sharing networks," in *Proc. 2024 IEEE Int. Conf. Commun.* IEEE, Aug. 2024, pp. 2767–2772.

[10] J. Zheng, J. Zhang, E. Björnson, and B. Ai, "Impact of channel aging on cell-free massive MIMO over spatially correlated channels," *IEEE Trans. Wireless Commun.*, vol. 20, no. 10, pp. 6451–6466, Apr. 2021.

[11] H. Jiang, M. Cui, D. W. K. Ng, and L. Dai, "Accurate channel prediction based on transformer: Making mobility negligible," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2717–2732, Jul. 2022.

[12] H. Lu, Y. Zeng, C. You, Y. Han, J. Zhang, Z. Wang, Z. Dong, S. Jin, C.-X. Wang, T. Jiang *et al.*, "A tutorial on near-field XL-MIMO communications toward 6G," *IEEE Comm. Surveys Tut.*, vol. 26, no. 4, pp. 2213–2257, Apr. 2024.

[13] F. Pena-Campos, R. Carrasco-Alvarez, O. Longoria-Gandara, and R. Parra-Michel, "Estimation of fast time-varying channels in OFDM systems using two-dimensional prolate," *IEEE Trans. Wireless Commun.*, vol. 12, no. 2, pp. 898–907, Jan. 2013.

[14] I. C. Wong and B. L. Evans, "Sinusoidal modeling and adaptive channel prediction in mobile OFDM systems," *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1601–1615, Mar. 2008.

[15] H. Yin, H. Wang, Y. Liu, and D. Gesbert, "Addressing the curse of mobility in massive MIMO with prony-based angular-delay domain channel predictions," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 12, pp. 2903–2917, Jun. 2020.

[16] Y. Shi, Z. Jiang, Y. Liu, Y. Wang, and S. Xu, "A compressive sensing based channel prediction scheme with uneven pilot design in mobile massive MIMO systems," in *Proc. IEEE 13th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Dec. 2021, pp. 1–6.

[17] J. Lee, G.-T. Gil, and Y. H. Lee, "Channel estimation via orthogonal matching pursuit for hybrid MIMO systems in millimeter wave communications," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2370–2386, Apr. 2016.

[18] K. E. Baddour and N. C. Beaulieu, "Autoregressive modeling for fading channel simulation," *IEEE Trans. Wireless Commun.*, vol. 4, no. 4, pp. 1650–1662, Sep. 2005.

[19] J. Yuan, H. Q. Ngo, and M. Matthaiou, "Machine learning-based channel prediction in massive MIMO with channel aging," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 2960–2973, Feb. 2020.

[20] L. Chen, M. Loschonsky, and L. M. Reindl, "Autoregressive modeling of mobile radio propagation channel in building ruins," *IEEE Trans. Microwave Theory Tech.*, vol. 60, no. 5, pp. 1478–1489, Mar. 2012.

[21] M. Yusuf, E. Tanghe, F. Challita, P. Laly, L. Martens, D. P. Gaillot, M. Lienard, and W. Joseph, "Autoregressive modeling approach for non-stationary vehicular channel simulation," *IEEE Trans. Veh. Technol.*, vol. 71, no. 2, pp. 1124–1131, Dec. 2021.

[22] D. Löschenbrand, M. Hofer, and T. Zemen, "Spectral efficiency of time-variant massive MIMO using wiener prediction," *IEEE Commun. Lett.*, vol. 27, no. 4, pp. 1225–1229, Feb. 2023.

[23] A. K. Papazafeiropoulos and T. Ratnarajah, "Deterministic equivalent performance analysis of time-varying massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 10, pp. 5795–5809, Jun. 2015.

[24] L. Lindbom, M. Sternad, and A. Ahlén, "Tracking of time-varying mobile radio channels. 1. the Wiener LMS algorithm," *IEEE Trans. Commun.*, vol. 49, no. 12, pp. 2207–2217, Aug. 2001.

[25] Z. Qin, H. Yin, and W. Li, "Eigenvector prediction-based precoding for massive MIMO with mobility," *arXiv preprint arXiv:2308.12619*, Aug. 2023.

[26] K. T. Truong and R. W. Heath, "Effects of channel aging in massive MIMO systems," *J. Commun. Netw.*, vol. 15, no. 4, pp. 338–351, Sep. 2013.

[27] S. Kashyap, C. Mollén, E. Björnson, and E. G. Larsson, "Performance analysis of (TDD) massive MIMO with Kalman channel prediction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 3554–3558.

[28] H. Kim, S. Kim, H. Lee, C. Jang, Y. Choi, and J. Choi, "Massive MIMO channel prediction: Kalman filtering vs. machine learning," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 518–528, Sep. 2020.

[29] W. Jiang and H. D. Schotten, "Neural network-based fading channel prediction: A comprehensive overview," *IEEE Access*, vol. 7, pp. 118 112–118 124, Aug. 2019.

[30] J. M. Huttunen, D. Korpi, and M. Honkala, "Deeptx: Deep learning beamforming with channel prediction," *IEEE Trans. Wireless Commun.*, vol. 22, no. 3, pp. 1855–1867, Sep. 2022.

[31] Y. Zhang, Y. Wu, A. Liu, X. Xia, T. Pan, and X. Liu, "Deep learning-based channel prediction for LEO satellite massive MIMO communication system," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, pp. 1835–1839, May 2021.

[32] W. Jiang and H. D. Schotten, "Deep learning for fading channel prediction," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 320–332, Mar. 2020.

[33] I. Helmy, P. Tarafder, and W. Choi, "LSTM-GRU model-based channel prediction for one-bit massive MIMO system," *IEEE Trans. Veh. Technol.*, vol. 72, no. 8, pp. 11 053–11 057, Mar. 2023.

[34] T. Zhou, X. Liu, Z. Xiang, H. Zhang, B. Ai, L. Liu, and X. Jing, "Transformer network based channel prediction for CSI feedback enhancement in AI-native air interface," *IEEE Trans. Wireless Commun.*, vol. 23, no. 9, pp. 11 154–11 167, Mar. 2024.

[35] M. Bodini, M. W. Rivolta, and R. Sassi, "Opening the black box: interpretability of machine learning algorithms in electrocardiography," *Phil. Trans.*, vol. 379, no. 2212, p. 20200253, Oct. 2021.

[36] M. D. Migliore, "Horse (electromagnetics) is more important than horseman (information) for wireless transmission," *IEEE Trans. Antenna Propagat.*, vol. 67, no. 4, pp. 2046–2055, Dec. 2018.

[37] R. Li, D. Li, J. Ma, Z. Feng, L. Zhang, S. Tan, E. Wei, H. Chen, and E.-P. Li, "An electromagnetic information theory based model for efficient characterization of MIMO systems in complex space," *IEEE Trans. Antenna Propagat.*, vol. 71, no. 4, pp. 3497–3508, Jan. 2023.

[38] J. Zhu, X. Su, Z. Wan, L. Dai, and T. J. Cui, "The benefits of electromagnetic information theory for channel estimation," in *Proc. 2024 IEEE Int. Conf. Commun.* IEEE, Jun. 2024, pp. 4869–4874.

[39] A. Pizzo, T. L. Marzetta, and L. Sanguinetti, "Spatially-stationary model for holographic MIMO small-scale fading," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 9, pp. 1964–1979, Jun. 2020.

[40] A. Pizzo, L. Sanguinetti, and T. L. Marzetta, "Fourier plane-wave series expansion for holographic MIMO communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 6890–6905, Mar. 2022.

[41] J. Zhu, Z. Wan, L. Dai, M. Debbah, and H. V. Poor, "Electromagnetic information theory: Fundamentals, modeling, applications, and open problems," *IEEE Wireless Commun.*, Jan. 2024.

[42] S. Mumtaz, J. Rodriguez, and L. Dai, *mmWave massive MIMO: A paradigm for 5G*. New York, NY, USA: Academic, 2016.

[43] L. Cheng, B. Henty, F. Bai, and D. D. Stancil, "Doppler spread and coherence time of rural and highway vehicle-to-vehicle channels at 5.9 GHz," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Sep. 2008, pp. 1–6.

[44] T. Gong, C. Huang, J. He, M. Di Renzo, M. Debbah, and C. Yuen, "A transmit-receive parameter separable electromagnetic channel model for LoS holographic MIMO," in *Proc. 2023 IEEE Global Commun. Conf. (GLOBECOM)*. IEEE, Feb. 2023, pp. 5701–5706.

[45] L. Wei, T. Gong, C. Huang, Z. Zhang, W. E. Sha, Z. N. Chen, L. Dai, M. Debbah, and C. Yuen, "Electromagnetic information theory for holographic mimo communications," *arXiv preprint arXiv:2405.10496*, Dec. 2024.

[46] Z. Wan, J. Zhu, and L. Dai, "Near-field channel modeling for electromagnetic information theory," *arXiv preprint arXiv:2403.12268*, Mar. 2024.

[47] M. Chafii, L. Bariah, S. Muhaidat, and M. Debbah, "Twelve scientific challenges for 6G: Rethinking the foundations of communications theory," *IEEE Comm. Surveys Tut.*, vol. 25, no. 2, pp. 868–904, Feb. 2023.

[48] Z. Wan, J. Zhu, and L. Dai, "Electromagnetic information theory motivated near-field channel model," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*. IEEE, Jul. 2024, pp. 1800–1805.

[49] J. Zhu, Z. Wan, L. Dai, and T. Jun Cui, "Electromagnetic information theory-based statistical channel model for improved channel estimation," *IEEE Trans. Inf. Theory*, vol. 71, no. 3, pp. 1777–1793, Jan. 2025.

[50] K. V. Mardia and S. El-Atoum, "Bayesian inference for the von Mises-Fisher distribution," *Biometrika*, vol. 63, no. 1, pp. 203–206, Apr. 1976.

[51] E. Schulz, M. Speekenbrink, and A. Krause, "A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions," *J. Math. Psychol.*, vol. 85, pp. 1–16, Aug. 2018.

[52] M. Wytock and Z. Kolter, "Sparse Gaussian conditional random fields: Algorithms, theory, and application to energy forecasting," in *Proc. Int. Conf. Machine Learning.* PMLR, Mar. 2013, pp. 1265–1273.

[53] F. Yin, L. Pan, T. Chen, S. Theodoridis, Z.-Q. T. Luo, and A. M. Zoubir, "Linear multiple low-rank kernel based stationary Gaussian processes regression for time series," *IEEE Trans. Signal Process.*, vol. 68, pp. 5260–5275, Sep. 2020.

[54] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794–816, Aug. 2016.

[55] T. L. J. Ng and K.-K. Kwong, "Universal approximation on the hypersphere," *Commun. Stat.-Theory and Methods*, vol. 51, no. 24, pp. 8694–8704, Mar. 2022.

[56] C.-j. Hsieh, A. Banerjee, I. Dhillon, and P. Ravikumar, "A divide-and-conquer method for sparse inverse covariance estimation," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 25, Jan. 2012.

[57] E. A. Fattah, H. Ltaief, H. Rue, and D. Keyes, "GPU-accelerated parallel selected inversion for structured matrices using stiles," *arXiv preprint arXiv:2504.19171*, Sep. 2025.

[58] C. Tang, C. Liu, L. Yuan, and Z. Xing, "High precision low complexity matrix inversion based on Newton iteration for data detection in the massive MIMO," *IEEE Commun. Lett.*, vol. 20, no. 3, pp. 490–493, Jan. 2016.

[59] K. R. Schaubach, N. J. Davis, and T. S. Rappaport, "A ray tracing method for predicting path loss and delay spread in microcellular environments," in *Proc. Veh. Technol. Soc. 42nd VTS Conf. Front. Technol.*, vol. 2. IEEE, Jun. 1992, pp. 932–935.