electronics

MDPI

Article

Multi-Scale Spatio-Temporal Graph Transformer with xLSTM for Electric Vehicle Charging Demand Prediction

Yedan Li [1], Zhuoran Ye [1], Bo Li [1,2,*] and Zijun Chen [1]

[1] School of Electrical Engineering, Guangxi University, Nanning 530004, China; 2302110349@st.gxu.edu.cn (Y.L.)
[2] Guangxi Key Laboratory of Power System Optimization and Energy-Saving Technology, Guangxi University, Nanning 530004, China
* Correspondence: boli@gxu.edu.cn

Abstract

Accurate prediction of electric vehicle (EV) charging demand is critical for planning charging infrastructure and allocating resources efficiently. However, existing methods often fail to capture multi-scale spatial dependencies and struggle to model both long-range temporal dependencies and short-term fluctuations. To address these limitations, a deep learning framework, STGFormer, is developed for citywide EV charging demand forecasting. First, a temporal dilated convolution module (TDConv) is proposed to extract multi-scale local temporal patterns. Second, an adaptive spatial dilated graph attention module (ADGAT) is proposed to mine multi-hop spatial correlations between geographically adjacent and functionally similar regions. Third, a hybrid xLSTM-Transformer encoder captures global temporal dependencies while preserving local continuity. The performance of STGFormer is evaluated on a real-world dataset from Shenzhen. Extensive experiments demonstrate that STGFormer consistently outperforms fifteen representative baseline models. It achieves average improvements of 8.58% in RMSE, 13.95% in RAE, and 14.64% in MAE.

Keywords: electric vehicle charging demand forecasting; spatiotemporal graph transformer; graph attention network; dilated convolution; multi-scale spatiotemporal modeling

1. Introduction

The rapid transition toward sustainable transportation has driven the widespread adoption of various electric vehicle (EVs) [1,2]. With the rapid growth of EV adoption, the demand for public charging services is rising sharply, placing increasing pressure on urban charging infrastructure [3]. Accurate forecasting of charging demand at EV charging stations is essential for aligning infrastructure investments with actual usage patterns [4]. From an operational perspective, reliable forecasts help charging operators to allocate capacity efficiently and reduce congestion [5]. From the power-system perspective, citywide EV charging demand forecasting is key for proactive planning and coordination. Without accurate predictions, planners may under-provision in high-demand areas or overbuild in low-demand zones, leading to wasted investment and degraded service quality [6]. Consequently, citywide EV charging demand forecasting provides fundamental decision support for scalable, cost-effective, and user-centric EV infrastructure development.

Many efforts have been made to predict citywide electric vehicle charging demand [7]. Deep learning has shown a strong ability to model complex patterns in charging data. Recurrent neural networks, like LSTM [8], have proven effective at modeling temporal dependencies for individual charging stations. Convolutional neural networks, such as

Check for updates

Academic Editor: Herminio Martínez-García

Received: 20 June 2026
Revised: 8 August 2026
Accepted: 10 August 2026
Published: 14 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1D-CNNs and temporal convolutional networks (TCNs) [9], are widely used to extract local temporal patterns via receptive fields. On the spatial side, graph neural networks are widely adopted to capture spatial spillover effects between regions [10]. Recent studies have increasingly adopted attention mechanisms to better capture complex spatiotemporal dependencies. For instance, attention-based spatiotemporal graph convolutional networks (ASTGCN) [11] incorporate spatiotemporal attention into GCN [12]. This allows the model to focus on relevant regions and time steps adaptively. Transformers have been adopted for temporal modeling, while graph attention networks (GAT) are used for spatial aggregation [13,14]. These approaches offer greater flexibility in attention weighting and have been shown to enhance spatiotemporal forecasting performance [15].

However, existing methods often assign similar importance to all historical states, although their contributions to future demand are different. Capturing long-range spatial correlations and time-varying external influences effectively remains a key challenge for accurate charging demand prediction [16,17].

Charging demand is jointly shaped by multiple factors. First, charging demand is governed by the regional activity structure. Zones with distinct travel patterns and daily routines exhibit different load patterns at daily and weekly scales, as car charging patterns differ across regions [18]. Many studies have found that future EV charging demand is influenced by historical demand and spatial dependence [19]. Second, charging demand is driven by charging cost, including time-of-use electricity prices and service fees [20,21]. In addition, weather conditions such as temperature have been observed to influence the intensity of charging demand [22]. Incorporating multiple relevant factors into urban EV charging demand forecasting has become an important research topic. However, studies that can effectively account for multiple influencing factors simultaneously remain limited [21].

To address these issues, a deep learning framework termed STGFormer is proposed for citywide EV charging demand forecasting. Multi-scale temporal patterns, multi-hop spatial dependencies, and complementary global–local temporal dynamics are jointly captured within the framework. In addition, the effectiveness of STGFormer is evaluated using a real-world dataset collected in Shenzhen.

The main contributions of this study are as follows.

(1) Inspired by the atrous spatial pyramid, ADGAT constructs parallel multi-hop graph branches rather than stacking layers serially as in ChebNet and STGCN-LSTM, thereby preserving scale-specific features without cross-scale contamination. APPNP propagation further diffuses attentive features beyond fixed hop boundaries, unlike recent models such as GAT-Mamba that rely solely on single-hop graph attention.
(2) Convolutions are applied along the time dimension to capture key fluctuation patterns that help the model better identify important time windows. At the same time, to enhance the model's temporal awareness, position encoding is incorporated to inject time-step information into the feature representation, complementing the local patterns captured by convolutions.
(3) A hybrid temporal encoder (xLSTM–Transformer) is designed to capture both global patterns and local dynamics. Global self-attention handles long-range dependencies, while xLSTM focuses on short-term sequential changes. Together, they provide a better representation of periodic trends and peak fluctuations.

The rest of this paper is organized as follows. Section 2 reviews the related literature. Section 3 presents the proposed STGFormer method in detail. Section 4 evaluates the proposed method using a representative EV charging demand dataset. Finally, Section 5 concludes the paper and discusses future research directions.

2. Literature Review

Early charging demand forecasting relied mainly on statistical modeling. Autoregressive Integrated Moving Average (ARIMA) and seasonal decomposition techniques have been widely used to capture linear trends and periodic patterns in charging load series [23,24]. Grey forecasting models have been used for short-term prediction, especially with limited data. These studies primarily focused on single-station forecasting and established time series modeling baselines for charging demand [25]. However, these statistical approaches are limited when it comes to capturing nonlinear fluctuations and peak events. Moreover, because they fail to account for external factors and spatial dependencies between areas, these approaches are inadequate for scenarios such as multi-station prediction [26].

Deep learning improves the modeling capacity for charging-demand forecasting. They can not only model nonlinear variations in charging demand, but also learn interactions among exogenous factors [27]. RNN-based models, such as LSTM and GRU, are widely adopted due to their gating mechanisms, which help capture short-term temporal dependencies and periodic patterns from historical load [8].

Recent studies have incorporated exogenous variables such as weather, holidays, and electricity prices into deep learning-based forecasting models. In this way, exogenous factors are jointly represented. Convolutional models such as 1D-CNN and TCN are also widely used [9]. By hierarchical feature extraction, multi-scale periodic patterns are captured more directly, offering advantages in training efficiency and local pattern representation. To better model temporal relationships across the globe, Transformers have been widely used for time series forecasting [13,27]. Their self-attention mechanism captures long-range dependencies across time steps, making them suited for multi-step-ahead prediction and scenarios with strong non-stationary fluctuations. Temporal Fusion Transformers (TFT) are designed to integrate heterogeneous inputs, such as static metadata and exogenous variables, providing interpretable multi-horizon forecasting [28,29]. To better incorporate external factors, recent models such as TimeXer, use patch-wise and variate-wise attention mechanisms to effectively integrate temporal dependencies and exogenous variables [30]. However, distinctive temporal patterns in charging demand, such as sharp peaks and short-term fluctuations, are often overlooked. Few studies have specifically optimized network structures to capture both long-range dependencies and short-term fluctuations.

Urban charging networks also exhibit significant spatial correlations. These correlations arise not only from geographic proximity, but also from interactions between areas with similar functions [29]. To capture spatial dependencies, graph attention networks (GAT) have been introduced [14]. Learnable weights are assigned to neighboring nodes, enabling adaptive aggregation of spatial information. This helps reflect differences in spatial correlations and avoids over-smoothing. Furthermore, to model complex spatial interactions, a Heterogeneous Spatio-Temporal Graph Convolutional Network (HSTGCN) is developed to capture diverse spatial dependencies through dynamic heterogeneous graph structures [18,31]. However, it relies on local message-passing mechanisms, which make it difficult to capture complex, non-local spatial correlations effectively [32]. As a result, most existing GAT-based approaches are often only propagated within first-order neighborhoods, which limits the ability to capture non-local correlations between areas that are functionally similar but geographically dispersed.

In summary, deep learning and attention mechanisms provide powerful tools for modeling temporal and spatial patterns in charging demand. Researchers are increasingly turning to advanced attention mechanisms to better capture complex data dependencies. However, existing approaches still have limitations. They struggle to jointly capture non-

local correlations between functionally similar but geographically dispersed areas. They also fail to precisely model peak fluctuations and critical time windows within a unified framework. Therefore, spatiotemporal modeling mechanisms that can balance long-range dependencies and local short-term fluctuations still need to be explored.

3. Methods

3.1. Spatio-Temporal Prediction Framework

As shown in Figure 1, STGFormer adopts a Transformer backbone and augments the encoder with an xLSTM branch to enhance temporal encoding and capture critical temporal intervals. A spatial dilated graph attention is introduced to model multi-hop and multi-scale dependencies, with geographical proximity and functional similarity integrated as constraints. Functional similarity between zones is quantified via their Point of Interest (POI) category labels, where two zones sharing the same POI type are considered functionally similar. This functional graph complements the geographical adjacency matrix in guiding the ADGAT attention mechanism. Convolutional layers are further employed to extract temporal patterns, while positional encoding is incorporated to inject temporal order information into the feature representation. By jointly leveraging these mechanisms, the proposed framework captures multi-hop spatial couplings and temporal fluctuations in urban charging networks. The model parameters are initialized via a physics-informed meta-learning pre-training stage prior to fine-tuning.

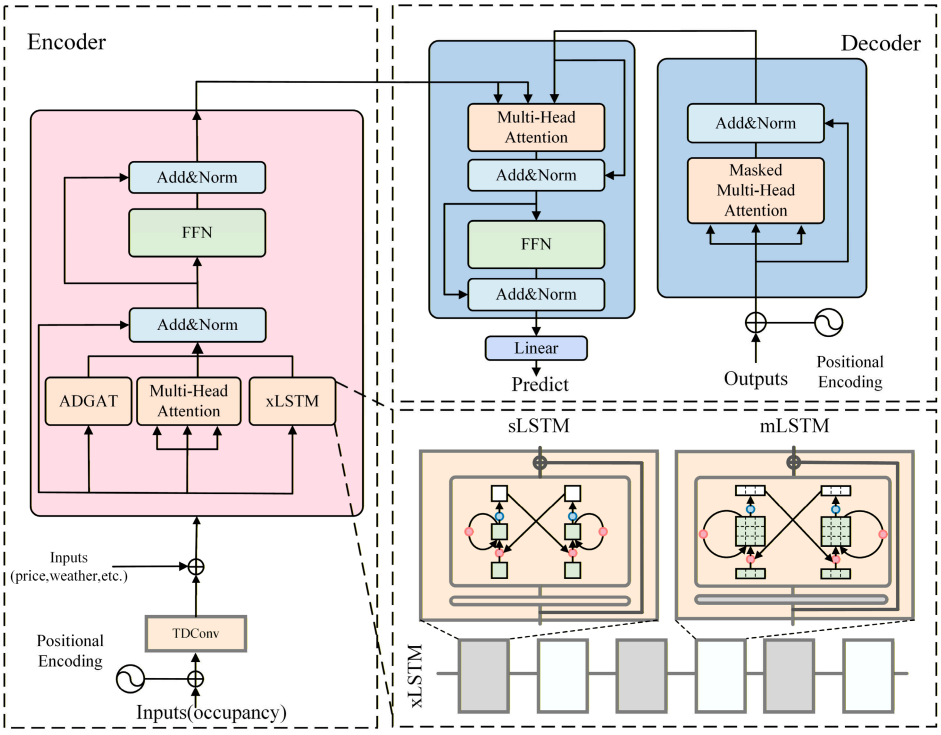


Figure 1. Architecture of STGFormer.

The input comprises a historical charging-load sequence spanning a fixed-length time window. Multi-dimensional exogenous features are also provided. The computation is formulated in (1).

$$\mathbf{L}_{t+1:t+T} = \mathrm{F}_\theta(\mathbf{L}_{t-H+1:t}, \mathbf{X}_{t-H+1:t}, \mathbf{T}_{t-H+1}, \mathbf{M}) \tag{1}$$

where $\mathbf{L}_{t+1:t+T} \in \mathbb{R}^T$ denotes the predicted charging load for the next T time steps. $\mathbf{X}_{t-H+1:t} = \{\mathbf{P}_{t-H+1:t}, \mathbf{S}_{t-H+1:t}\}$ denotes the economic factors, including the electricity price $\mathbf{P}_t$ and the service fee $\mathbf{S}_t$. $\mathbf{T}_{t-H+1:t}$ denotes the temperature factor. $\mathbf{M}$ denotes the

clustering labels, and θ denotes the trainable parameters. Equation (1) maps the historical charging load and exogenous features to future load predictions.

3.2. xLSTM-Based Temporal Encoding Branch

Self-attention does not work well for EV charging demand because charging demand often has sudden peaks. Self-attention looks at the whole sequence at once and averages everything out. This makes the peaks less noticeable. It also gives too much attention to irrelevant or noisy time steps. As a result, it becomes harder to accurately predict the peaks.

To solve these problems, we add a parallel xLSTM branch to work together with the Multi-Head Self-Attention (MHSA) module [29]. In this two-branch design, MHSA learns global repeating patterns, while the xLSTM branch focuses on short-term changes and reduces the influence of past noise. Specifically, the xLSTM branch uses sLSTM, which has scalar memory cells and better gating. It learns local patterns that change smoothly and stays stable when sudden changes happen. At the same time, mLSTM uses matrix memory to store a large amount of key–value pairs and selectively retrieve past states related to peaks, as shown in Figure 2.

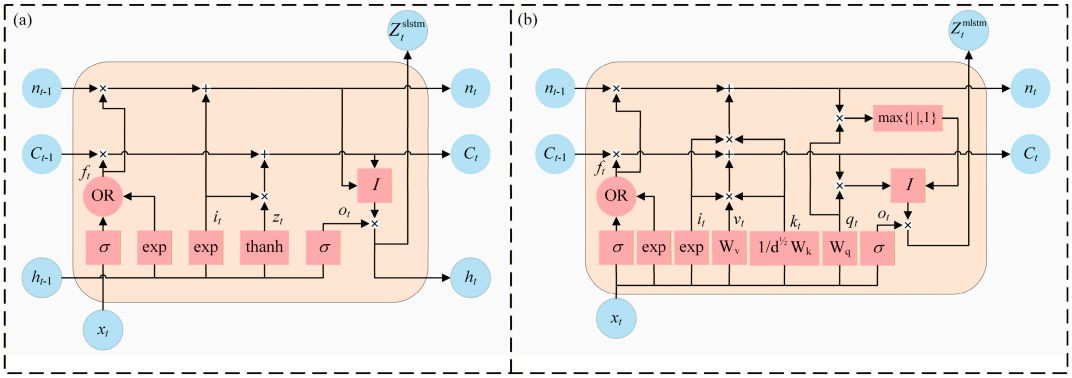


Figure 2. (**a**) sLSTM architecture. (**b**) mLSTM architecture.

The sLSTM update rules are formulated in (2)–(6).

$$\mathbf{i}_t = \sigma(\mathbf{W}_i \cdot \mathbf{x}_t + \mathbf{U}_i \cdot \mathbf{h}_{t-1} + \mathbf{b}_i) \tag{2}$$

$$\mathbf{f}_t = \sigma\left(\mathbf{W}_f \cdot \mathbf{x}_{t-1} + \mathbf{U}_f \cdot \mathbf{h}_{t-1} + \mathbf{b}_f\right) \tag{3}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_o \cdot \mathbf{x}_{t-1} + \mathbf{U}_o \cdot \mathbf{h}_{t-1} + \mathbf{b}_o) \tag{4}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tanh(\mathbf{W}_c \cdot \mathbf{x}_t + \mathbf{U}_c \cdot \mathbf{h}_{t-1} + \mathbf{b}_c) \tag{5}$$

$$\mathbf{h}_t^{\text{slstm}} = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \tag{6}$$

where $\mathbf{i}_t$, $\mathbf{f}_t$, $\mathbf{o}_t$ denote the input, forget, and output gates respectively. $\mathbf{c}_t$ is the scalar cell state. $\mathbf{h}_t^{\text{slstm}}$ denotes the hidden state generated by the sLSTM. σ is the sigmoid function. Equations (2)–(4) calculate the input, forget, and output gates, while Equations (5) and (6) update the sLSTM cell state and generate its hidden representation, respectively.

The mLSTM employs matrix memory $\mathbf{c}_t \in \mathbb{R}^{D \times D}$ and updates are formulated in (7)–(9).

$$\mathbf{q}_t = \mathbf{W}_t \cdot \mathbf{x}_t, \mathbf{k}_t = \mathbf{W}_k \cdot \mathbf{x}_t, \mathbf{v}_t = \mathbf{W}_v \cdot \mathbf{x}_t \tag{7}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \left(\mathbf{v}_t \mathbf{k}_t^T\right) \tag{8}$$

$$\mathbf{h}_t^{\text{mlstm}} = \mathbf{o}_t \odot \frac{\mathbf{c}_t \cdot \mathbf{q}_t}{\| \mathbf{c}_t \cdot \mathbf{q}_t \|^2} \tag{9}$$

where $\mathbf{q}_t$, $\mathbf{k}_t$, $\mathbf{v}_t$ are the query, key, and value vectors. $\mathbf{f}_t$, $\mathbf{i}_t$, $\mathbf{o}_t$ are gating scalars computed similarly to (7)–(9). $\mathbf{h}_t^{\text{mlstm}}$ denotes the hidden state generated by the mLSTM. Equation (7) generates the query, key, and value vectors; Equation (8) updates the matrix memory, while Equation (9) produces the mLSTM hidden representation.

The temporal representation from MHSA, denoted $\mathbf{H}^1 \in \mathbb{R}^{N\times T\times D}$, and the xLSTM output, denoted $\mathbf{H}^2 \in \mathbb{R}^{N\times T\times D}$, are fused through a learnable gate.

$$\gamma = \sigma(g) \tag{10}$$

$$\mathbf{H} = \gamma \cdot \mathbf{H}^1 + (1-\gamma)\mathbf{H}^2 \tag{11}$$

where $g \in \mathbb{R}$ is a learnable scalar and σ is the sigmoid function. This gating mechanism allows the model to dynamically balance global pattern recognition and local fluctuation modeling according to the input characteristics $\mathbf{S}_t$. Equation (10) calculates the fusion coefficient and Equation (11) fuses the MHSA and xLSTM representations.

3.3. Adaptive Spatial Dilated Graph Attention Module (ADGAT)

Urban EV charging demand is characterized by spatial dependencies that often extend beyond first-order neighbors. Charging variations observed in one region are affected not only by adjacent areas, but also by regions located several hops away in the spatial graph. To capture such non-local spatial interactions more effectively, an adaptive multi-hop graph attention module (ADGAT) is introduced. As shown in Figure 3, instead of enlarging the receptive field through repeated stacking of graph neural network layers, multi-hop graph branches are constructed directly from the original adjacency graph, and graph attention aggregation is performed over these branches in parallel. As a result, both short-range and long-range spatial correlations can be modeled within a single layer.

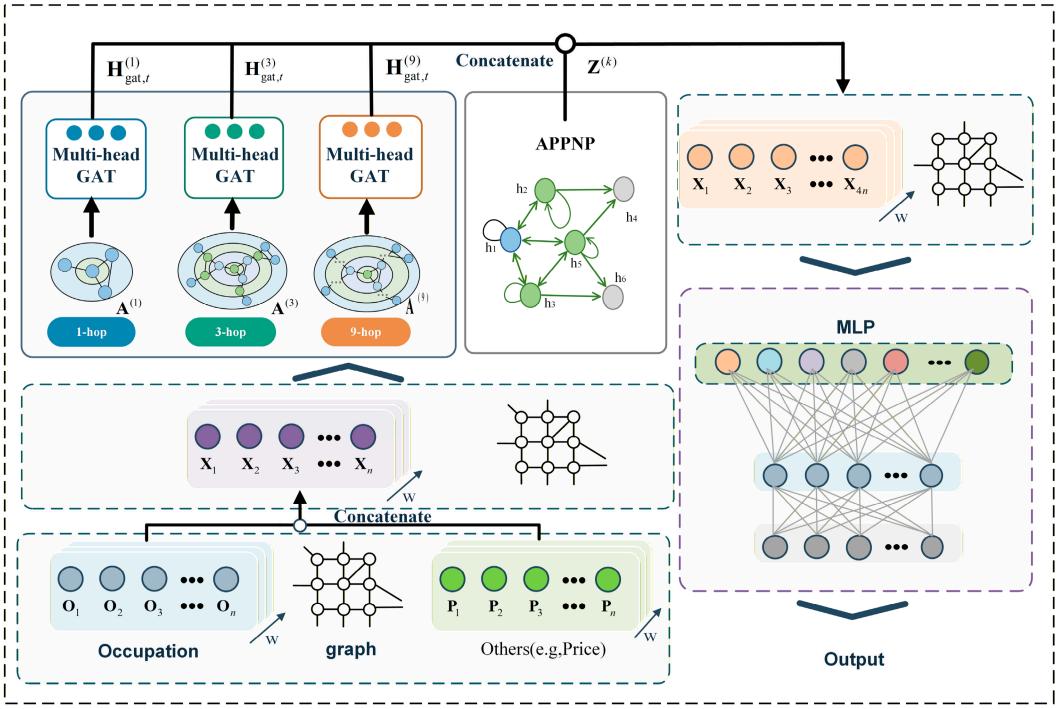


Figure 3. Adaptive dilated-graph attention (ADGAT) module.

The urban charging network is represented by a graph $\mathcal{G} = (V, \mathcal{E})$ with N nodes, where $\mathbf{X} \in \mathbb{R}^{N\times D}$ denotes the node feature matrix. Starting from the original adjacency matrix $\mathbf{A}$, a group of hop-specific adjacency matrices is constructed according to a prede-

fined scale set $S = \{s_1, s_2, \ldots, s_K\}$. For each hop scale $s \in S$, the corresponding hop-specific adjacency matrix is given in (12).

$$\mathbf{A}^{(s)} = \begin{cases} \mathbf{A} + \mathbf{I}, & s = 1 \\ \mathbb{I}(\mathbf{A}^s > 0) + \mathbf{I}, & s > 1 \end{cases} \tag{12}$$

where $\mathbf{I}$ is the identity matrix and $\mathbb{I}(\cdot)$ is the indicator function. Under this formulation, a separate graph branch is assigned to each spatial scale, such that node interactions within 1-hop, 3-hop, or 9-hop neighborhoods can be modeled independently. For each branch, graph attention is applied independently on the corresponding adjacency $\mathbf{A}^{(s)}$. In the h_{th} attention head, the attention coefficient between node i and node j is computed in (13).

$$e_{ij}^{(s,h)} = \text{LeakyReLU}\left(\mathbf{a}^{(s,h)^T}\left[\mathbf{W}^{(s,h)}\mathbf{x}_i \| \mathbf{W}^{(s,h)}\mathbf{x}_j\right]\right) \tag{13}$$

where $\mathbf{W}^{(s,h)}$ and $\mathbf{a}^{(s,h)}$ are learnable parameters. $||$ denotes concatenation. The corresponding normalized attention coefficient is given in (14).

$$\alpha_{ij}^{(s,h)} = \frac{\exp\left(e_{ij}^{(s,h)}\right)}{\sum_{k \in \mathcal{N}_i^{(s)}} \exp\left(e_{ik}^{(s,h)}\right)} \tag{14}$$

where $\mathcal{N}_i^{(s)}$ denotes the neighbors of node i under the s-hop adjacency. Based on these attention weights, the output of the h_{th} head is given in (15).

$$\mathbf{z}_i^{(s,h)} = \sum_{j \in \mathcal{N}_i^{(s)}} \alpha_{ij}^{(s,h)} \mathbf{W}^{(s,h)} \mathbf{x}_j \tag{15}$$

Through this process, the importance of neighboring nodes is no longer treated uniformly. Instead, informative neighbors are assigned larger attention weights under each hop scale, allowing scale-specific spatial dependencies to be captured more adaptively. The outputs of different heads are then concatenated, and the spatial representation of branch s is given in (16).

$$\mathbf{H}^{(s)} = \phi\left(\text{Concat}\left(\mathbf{z}_i^{(s,1)}, \ldots, \mathbf{z}_i^{(s,H)}\right)\right) \tag{16}$$

where $\phi(\cdot)$ denotes a linear projection followed by a nonlinear activation. In this way, each hop branch is provided with greater representational capacity, since spatial interactions can be encoded across multiple attention subspaces rather than a single aggregation path.

To further enhance information diffusion after graph attention aggregation, an Approximate Personalized Propagation of Neural Predictions (APPNP) process is introduced. Among the three hop GAT outputs $\mathbf{H}_{\text{gat},t}^{(1)}$, $\mathbf{H}_{\text{gat},t}^{(3)}$ and $\mathbf{H}_{\text{gat},t}^{(9)}$, the propagation process is defined in (17) and (18).

$$\mathbf{Z}^{(k)} = \mathbf{H}_{\text{gat},t}^{(1)} \tag{17}$$

$$\mathbf{Z}^{(k)} = (1 - \alpha)\mathbf{A}\mathbf{Z}^{(k-1)} + \alpha\mathbf{H}_{\text{gat},t}^{(1)}, \quad k = 1, \ldots, K \tag{18}$$

where K denotes the number of propagation steps, α is the restart probability, and $\mathbf{A} = \mathbf{D}^{-\frac{1}{2}}(\mathbf{A} + \mathbf{I})\mathbf{D}^{-\frac{1}{2}}$ is the symmetrically normalized adjacency matrix with self-loops.

The final propagated representation is denoted as $\mathbf{H}_{\text{appnp},t} = \mathbf{Z}^{(K)}$. Through this process, the 1-hop GAT features are diffused across the entire graph, allowing broader neighborhood context to be incorporated. The restart term $\alpha\mathbf{H}_{\text{gat},t}^{(1)}$ preserves the original local features, preventing excessive smoothing of the learned node representations. Mean-

while, the 3-hop and 9-hop GAT branches retain their expanded receptive fields without additional propagation, providing complementary medium and long-range spatial coverage.

The 1-hop, 3-hop, and 9-hop GAT outputs are concatenated with the APPNP-propagated features and fused through a learnable linear transformation, as defined in (19).

$$\mathbf{H}_t^{\text{spatial}} = \text{ReLU}\left(\left[\mathbf{H}_{\text{gat},t}^{(1)} \,||\, \mathbf{H}_{\text{gat},t}^{(3)} \,||\, \mathbf{H}_{\text{gat},t}^{(9)} \,||\, \mathbf{Z}^{(K)}\right]\mathbf{W}_{\text{fusion}}\right) \tag{19}$$

where $\mathbf{W}_{\text{fusion}} \in \mathbb{R}^{4D\times D}$ is a learnable linear transformation and $||$ denotes concatenation along the feature dimension.

These spatial representations are stacked across all time steps and then normalized via LayerNorm, yielding $\mathbf{X}_{\text{gat}} \in \mathbb{R}^{B\times N\times T\times D}$, which is then normalized via LayerNorm. Within each encoder layer, this spatial representation is combined with the temporal features from both the Transformer branch $\mathbf{X}_{\text{trans}}$ and the xLSTM branch $\mathbf{X}_{\text{xlstm}}$ through a three-way learnable gate.

$$[\gamma_1, \gamma_2, \gamma_3] = \text{Softmax}\left(\left[g_{\text{trans}}, g_{\text{gat}}, g_{\text{xlstm}}\right]\right) \tag{20}$$

$$\mathbf{X}_{\text{out}} = \gamma_1 \mathbf{X}_{\text{trans}} + \gamma_2 \mathbf{X}_{\text{gat}} + \gamma_3 \mathbf{X}_{\text{xlstm}} \tag{21}$$

where g_{trans}, g_{gat} and g_{xlstm} are learnable scalar parameters initialized to uniform values. The three weights are constrained by Softmax to sum to one.

3.4. Temporal Dilated Convolution Module (TDConv)

Unlike the Transformer and xLSTM branches within the encoder, which operate at the representation level by modeling dependencies among already-embedded time steps, TDConv operates at the signal level directly on the raw input sequences. The parallel dilated convolutions function as a multi-rate filter bank that extracts temporal patterns at distinct frequency scales before any spatial or cross-temporal interactions are introduced, providing a multi-scale temporal foundation upon which the encoder's higher-level attention and memory mechanisms can operate.

Figure 4 illustrates the temporal dilated convolution module (TDConv). The embedded input $\mathbf{X} \in \mathbb{R}^{B\times P\times N\times d}$ is processed independently for each node along the temporal axis. Positional encoding $\mathbf{P}$ is integrated into this temporal dimension. This integration ensures that the temporal representation maintains correct time-step ordering, which is required for compatibility with the subsequent attention-based modules.

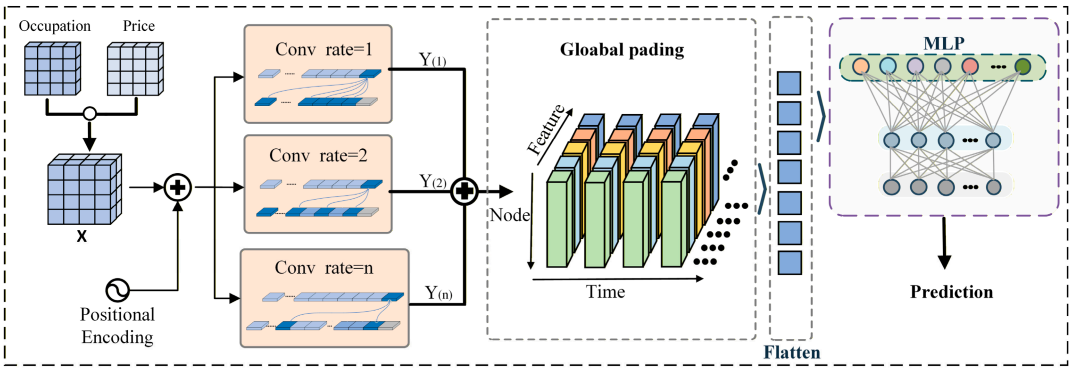


Figure 4. Temporal dilated convolution (TDConv) module.

TDConv module employs a parallel multi-branch architecture where each branch is assigned a specific dilation rate from the predefined set $\mathcal{D}$. To dynamically fuse multi-scale features, global average pooling is applied across the temporal dimension of each branch output to extract a branch-level descriptor. Subsequently, a lightweight gating

network generates adaptive weights to govern the fusion process. The comprehensive computational procedure is formulated in (22)–(26).

$$\mathbf{X}^{\mathrm{pe}} = \mathbf{X} + \mathbf{P} \tag{22}$$

$$\mathbf{Y}^{(r)} = \mathrm{Conv}^{(r)}(\mathbf{X}^{\mathrm{pe}}), \quad r \in \mathcal{D} \tag{23}$$

$$\mathbf{s}^{(r)} = \frac{1}{P}\sum_{t=1}^{P} \mathbf{Y}_t^{(r)}, \quad \mathbf{s}^{(r)} \in \mathbb{R}^{B\times N\times d} \tag{24}$$

$$g_r = \sigma\left(\mathbf{W}_g \mathbf{s}^{(r)} + \mathbf{b}_g\right) \tag{25}$$

$$\mathbf{X}_{\mathrm{TDconv}} = \sum_{r\in\mathcal{D}} g_r \cdot \mathbf{Y}^{(r)} \tag{26}$$

where $\mathbf{P}$ denotes the temporal positional encoding. The variables $\mathbf{W}_g$ and $\mathbf{b}_g$ denote the learnable weight matrix and bias vector of the gating network. The activation function σ maps the branch weight g_r to the $(0,1)$ interval. The resulting weight is used to construct the final dynamically fused temporal representation $\mathbf{X}_{\mathrm{TDConv}}$. Equation (22) adds the temporal positional encoding. Equation (23) extracts temporal features at dilation rate (r). Equation (24) calculates the descriptor of each dilation branch. Equation (25) calculates the adaptive branch weight. Equation (26) produces the fused TDConv representation.

3.5. Deep Feature Aggregation and Multi-Step Decoding for Prediction

To alleviate information degradation caused by deep network stacking, a cross-layer aggregation (CLA) mechanism is introduced into the encoder [33]. Furthermore, the decoder employs a causal masked attention mechanism together with a linear projection layer. This structure enables parallel multi-step decoding while strictly preserving temporal causality. The core computational logic is formulated in (27)–(29).

$$\mathbf{H}_{\mathrm{enc}} = \sum_{l=1}^{N} \beta_l \mathbf{X}^{(l)}, \quad \sum_{l=1}^{N} \beta_l = 1, \quad \beta_l \geq 0 \tag{27}$$

$$\mathrm{MaskedAttn}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}} + \mathbf{M}\right)\mathbf{V}, \quad \mathbf{M}_{ij} = \begin{cases} 0, & i \geq j, \\ -\infty, & i < j. \end{cases} \tag{28}$$

$$\mathbf{Y}_{t+1:t+T} = \mathbf{W}_{\mathrm{out}}\mathbf{H}_{\mathrm{dec}} + \mathbf{b}_{\mathrm{out}} \tag{29}$$

where (27) represents the cross-layer aggregation. The variable β_l is the learnable weight for the output $\mathbf{X}^{(l)}$ of the l_{th} layer. Equation (28) defines the causal masked attention. The matrix $\mathbf{M}$ is a mask matrix with its upper triangle set to negative infinity. Equation (29) formulates the linear prediction projection. The parameters $\mathbf{W}_{\mathrm{out}}$ and $\mathbf{b}_{\mathrm{out}}$ map the decoded representation $\mathbf{H}_{\mathrm{dec}}$ to the load predictions for the future T steps.

4. Experiments

This section presents the experimental evaluation of the proposed model. STGFormer is compared with state-of-the-art baselines to evaluate its performance. Furthermore, ablation experiments are conducted to show that each module contributes to improving prediction accuracy. Additionally, we analyze the spatial attention learned by ADGAT through heatmap visualization and conduct extensive hyperparameter studies to investigate the impact of key model components. All experiments were implemented in PyCharm (version 2024.3.2), while Mapize (https://mapize.com/, accessed on 12 August 2026) was used for spatial data processing.

4.1. Data Description

Empirical evaluations are conducted using real-world public charging station operation data from Shenzhen, China, covering major urban areas from 1 September 2023 to 28 February 2024. The study area is first partitioned into multiple functional zones, covering four categories of features: historical charging load, economic factors (electricity price and service fees), environmental factors (temperature), and spatial context (adjacency and POI-based functional similarity) [32,34]. Real-time traffic conditions and socio-demographic data, which may further improve prediction accuracy, are not included in the current feature set and are left as future extensions.

The spatial distribution of these charging stations across different functional zones is illustrated in Figure 5. Charging demand at each zone is aggregated at hourly intervals. All data are split into training (60%), validation (20%), and testing (20%) sets. Continuous variables are standardized using Z-scores to eliminate dimensional effects. The proposed STGFormer model takes a three-part input: (1) a multi-dimensional tensor constructed from historical load, electricity prices, service fees, and temperature as spatiotemporal feature representations; (2) positional encodings that inject temporal order information; and (3) cluster labels that serve as discrete features indicating zone categories.

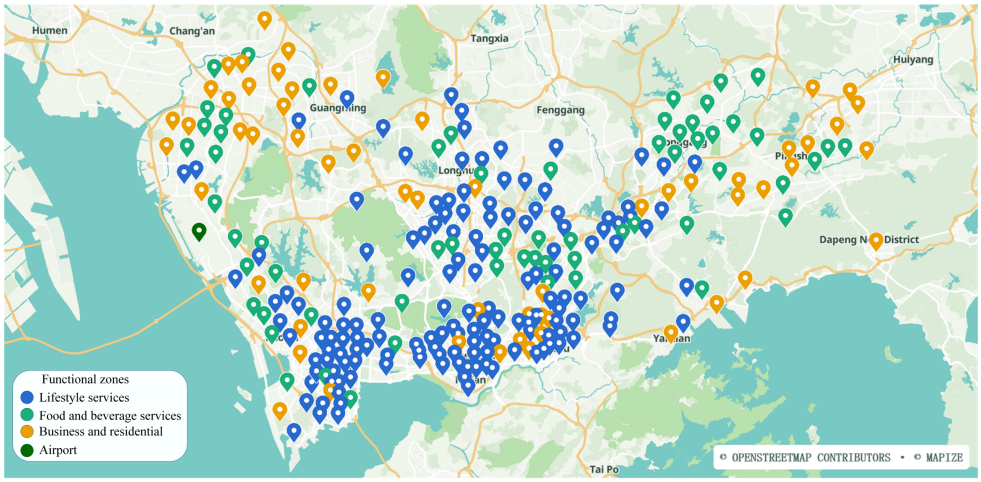


Figure 5. Spatial distribution of charging stations across functional zones in Shenzhen.

4.2. Model Comparison

This subsection evaluates the performance of STGFormer on citywide EV charging demand forecasting by comparing it with several representative baseline models. The models for comparison include one statistical model, two traditional machine learning models, four sequential neural networks, four graph-based spatial models, and three advanced spatiotemporal models. Table 1 presents the features considered in the compared models. A checkmark indicates that the corresponding information is used as an input in our implementation of the baseline, rather than necessarily being an intrinsic architectural component of the original model.

Table 1. Features considered by the compared models.

Model	Point of Interest	Adjacency	Charging Price	Temperature
ARIMA [23]				
XGBoost [35]	✓		✓	✓
LightGBM [36]	✓		✓	✓

Table 1. *Cont.*

Model	Point of Interest	Adjacency	Charging Price	Temperature
GCN [14]	✓	✓		
LSTM [8]			✓	✓
TCN [9]			✓	✓
Transformer [13]			✓	✓
TimesNet [24]			✓	✓
STGCN [26]	✓	✓		
ASTGCN [11]	✓	✓		
TFT [28]	✓		✓	✓
HSTGCN [18]	✓	✓		
TimeXer [30]			✓	✓
GAT-Mamba [33]	✓	✓	✓	✓
STGCN-LSTM [37]	✓	✓	✓	✓
STGFormer	✓	✓	✓	✓

In Table 2, we present the performance comparison between the proposed STGFormer and fifteen representative baseline models under the 1-h forecast horizon. Four commonly used regression metrics are adopted to evaluate the overall prediction error, including the root mean square error (RMSE), mean absolute error (MAE), relative absolute error (RAE), and mean absolute percentage error (MAPE). The proposed model achieves the best results across all evaluation metrics, further demonstrating its modeling advantages. Compared to the deep learning baselines, STGFormer achieves average improvements of 8.58% in RMSE, 13.95% in RAE, 14.64% in MAE, and 19.11% in MAPE.

Table 2. Performance of the compared models.

Model	RMSE ($\times 10^{-2}$)		MAPE (%)		RAE ($\times 10^{-2}$)		MAE ($\times 10^{-2}$)	
	6	12	6	12	6	12	6	12
ARIMA	13.02	15.36	58.14	65.20	68.50	76.04	9.26	10.74
XGBoost	12.81	14.73	56.52	62.74	63.73	73.29	8.94	10.28
LightGBM	13.27	15.25	57.55	63.86	69.18	79.56	9.28	10.67
GCN	11.84	13.26	49.40	54.83	54.34	60.86	7.90	8.85
LSTM	10.78	11.94	45.99	50.51	51.55	57.22	7.09	7.82
TCN	10.43	11.58	43.25	47.58	49.73	55.23	6.82	7.63
Transformer	10.14	11.46	42.23	46.30	49.20	54.59	6.71	7.58
TimesNet	9.73	10.60	41.50	44.82	48.97	53.38	6.58	7.19
STGCN	9.61	10.57	39.84	43.43	47.56	52.32	6.43	7.01
TFT	9.42	10.27	35.68	38.54	44.21	48.19	5.75	6.27
ASTGCN	9.54	10.49	37.20	40.55	46.74	51.41	5.99	6.59
HSTGCN	9.37	10.21	33.18	35.82	42.53	46.36	5.72	6.23
TimeXer	9.24	10.03	32.74	35.03	41.95	45.31	5.61	6.06
GAT-Mamba	9.43	10.38	33.72	36.75	45.38	49.46	5.81	6.36
STGCN-LSTM	9.29	10.13	32.95	35.59	42.16	45.95	5.74	6.25
STGFormer	9.02	10.01	31.28	34.63	40.84	44.07	5.42	5.96

Among the spatio-temporal baselines, STGCN-LSTM and GAT-Mamba both incorporate graph structures and exogenous features, making them the most relevant models for comparison. Averaged across the 6-h and 12-h forecast horizons, STGFormer reduces RMSE and MAE by 2.0% and 5.1%, respectively, compared with STGCN-LSTM. This result suggests that the parallel multi-hop attention mechanism in ADGAT captures spatial dependencies more effectively than the serial GCN-LSTM architecture. Compared with

GAT-Mamba, STGFormer achieves average reductions of 4.0% in RMSE and 6.5% in MAE, indicating that ADGAT's multi-scale spatial attention and three-way gating mechanism provide more flexible feature fusion than GAT-Mamba's sequential spatial-to-temporal processing. Overall, these results demonstrate the advantages of STGFormer's parallel encoder design over both convolution-based and sequential spatio-temporal alternatives.

Among the comparison models, several observations can be made. First, traditional statistical and machine learning models (ARIMA, XGBoost, LightGBM) perform the worst, highlighting their limitations in capturing complex spatiotemporal dependencies in charging demand. Second, sequential neural networks (LSTM, TCN, Transformer, TFT, TimesNet) show improved performance by effectively modeling temporal patterns, but they lack spatial awareness, limiting their ability to capture cross-regional correlations. Third, graph-based models (GCN, STGCN, ASTGCN, HSTGCN) incorporate spatial dependencies via graph structures but often lack explicit mechanisms for mining multi-hop spatial correlations. Fourth, TimeXer achieves strong performance by integrating exogenous variables, yet it still lacks dedicated modules for multi-scale temporal extraction and multi-hop spatial modeling. In contrast, STGFormer addresses these limitations through its TDConv module for multi-scale local temporal patterns, ADGAT module for multi-hop spatial correlations, and hybrid xLSTM-Transformer encoder for complementary temporal modeling. The consistent superiority of STGFormer across all metrics confirms that effectively capturing multi-scale spatiotemporal patterns is key to improving charging demand prediction accuracy.

To provide a clearer comparison of model performance, Figure 6 presents the forecasting results for two representative areas. The outputs of STGFormer and TimeXer are compared with their respective ground-truth charging demands. Although STGFormer does not match the true values perfectly at every time step, it achieves better overall performance than TimeXer. In addition, the prediction errors of the two models are reported for a more intuitive comparison.

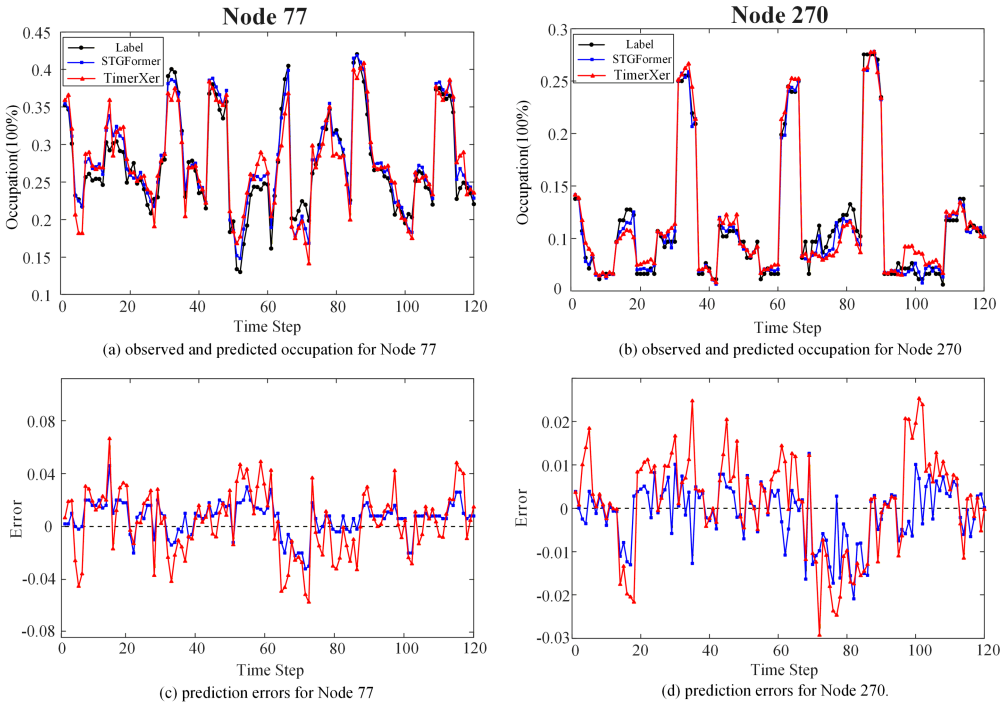


Figure 6. The EV charging demand prediction results and errors for two example areas.

Figure 6 presents the forecasting results for two representative areas, comparing STGFormer and TimeXer against ground-truth charging demand. Both areas exhibit distinct

demand patterns: Node 77 is a commercial zone with sharp midday and late-afternoon peaks, while Node 270 is a transit hub near Shenzhen East Station with irregular passenger-driven demand. STGFormer demonstrates greater responsiveness to abrupt fluctuations in both cases, while TimeXer exhibits noticeable prediction lag during peak transition periods. This improvement stems from the TDConv module's multi-scale temporal extraction and the hybrid xLSTM-Transformer encoder's complementary modeling of long-range trends and short-term dynamics, which TimeXer's standard self-attention cannot match.

To further quantify performance across functional zone types, Table 3 reports per-category error metrics based on POI labels. The model performs best on Lifestyle services zones (RMSE 8.76×10^{-2}), followed by Food and beverage services (RMSE 9.24×10^{-2}) and Business and residential zones (RMSE 9.87×10^{-2}). The larger error on mixed-use zones is consistent with their more complex demand patterns combining residential overnight peaks with daytime business activity.

Table 3. Per-category error metrics across different functional zone types.

Zone Type	RMSE ($\times 10^{-2}$)	MAE ($\times 10^{-2}$)
Business and residential	9.87	5.98
Food and beverage services	9.24	5.61
Lifestyle services	8.76	5.21

4.3. Ablation Experiments

Ablation experiments are conducted to evaluate the contribution of each module in the proposed STGFormer framework. Five ablation settings are considered: Module (a) replacing the hybrid xLSTM-Transformer encoder with a standard Transformer encoder; Module (b) removing the adaptive spatial dilated graph attention module (ADGAT); Module (c) removing the temporal dilated convolution module (TDConv); Module (d) removing temperature information from the inputs; Module (e) removing charging price from the inputs. Results are reported in Table 4 and visualized in Figure 7.

Table 4. Statistical results of ablation experiments.

Model	RMSE ($\times 10^{-2}$)	MAE ($\times 10^{-2}$)	MAPE (%)	ΔRMSE (%)
STGFormer	9.02	5.42	31.28	baseline
Module (a)	9.39	5.88	34.14	+4.1%
Module (b)	9.34	5.74	33.25	+3.6%
Module (c)	9.27	5.62	32.46	+2.8%
Temperature	9.20	5.57	31.92	+2.0%
Charging Price	9.17	5.51	31.76	+1.7%

As shown in Table 4, STGFormer achieves the best performance, and removing any component results in a performance drop. This confirms that all modules contribute positively to citywide EV charging demand forecasting.

Among the core modules, Module (a) has the largest impact. Replacing the hybrid xLSTM-Transformer encoder with a standard Transformer increases RMSE by 4.1%, indicating that both global temporal dependencies and short-term fluctuations are important for accurate prediction. This result verifies the effectiveness of the hybrid temporal encoder in capturing complex temporal patterns. Module (b) is the second-most-important component. Removing ADGAT increases RMSE by 3.6%. ADGAT helps capture dependencies between geographically adjacent and functionally similar regions, as well as non-local correlations that cannot be modeled using only one-hop neighbor information. Module (c) also provides a clear improvement. Removing TDConv increases RMSE by 2.8%, showing

that multi-scale local temporal patterns are beneficial for forecasting. Exogenous variables further improve prediction accuracy. Removing temperature and charging price increases RMSE by 2.0% and 1.7%, respectively. Thus, environmental and economic factors provide useful auxiliary information for modeling charging demand.

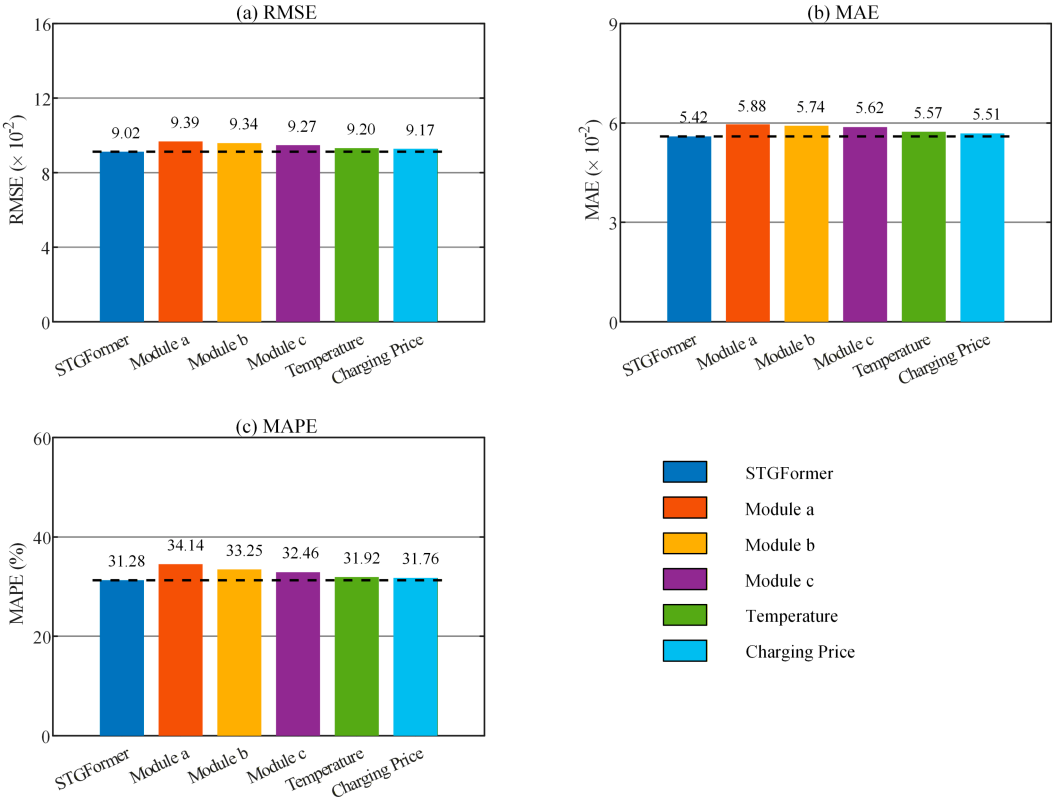


Figure 7. Ablation study on the STGFormer.

To further verify that xLSTM is the optimal choice for the hybrid encoder rather than any recurrent module, we conduct an additional experiment replacing the xLSTM branch with standard LSTM and GRU while keeping the Transformer and ADGAT branches unchanged.

As shown in Table 5, both LSTM and GRU improve over the standard Transformer, confirming the value of adding a recurrent branch for complementary temporal modeling. xLSTM further improves upon LSTM by 1.7% in RMSE and upon GRU by 2.5% in RMSE. This additional gain stems from xLSTM's dual-memory design: sLSTM tracks smooth diurnal trends via scalar gating, while mLSTM retrieves historical peak patterns through matrix memory—a capability that standard LSTM and GRU, with only scalar memory cells, cannot provide. We acknowledge the improvement over LSTM and GRU is modest, and note that Transformer + LSTM offers a reasonable lightweight alternative when computational resources are constrained, with 22.1 s/epoch training time compared to 28.6 s/epoch for the full xLSTM variant.

Table 5. Comparison of different temporal encoders in the hybrid architecture.

Encoder Variant	RMSE ($\times 10^{-2}$)	MAE ($\times 10^{-2}$)	Training Time (s/epoch)
Transformer	10.14	6.71	18.3
Transformer + LSTM	9.85	6.67	22.1
Transformer + GRU	9.93	6.75	20.4
Transformer + xLSTM	9.68	6.22	28.6

Overall, the ablation results show that the superior performance of STGFormer comes from the joint effects of hybrid temporal encoding, multi-hop spatial correlation modeling, multi-scale temporal feature extraction, and exogenous variable integration. The consistent drop in performance across all ablated variants confirms the importance of each component in the proposed framework.

4.4. Hyperparameter Analysis

4.4.1. Effect of Dilation Rate Set

Figure 8 presents the 48-h predictions for a representative commercial zone, under four dilation rate configurations: $I = \{1\}$, $I = \{1,3\}$, $I = \{1,3,9\}$, $I = \{1,3,9,16\}$. The model with only dilation rate 1 ($I = \{1\}$) performs the worst, particularly during peak periods. At midday around 12:00 and late afternoon around 18:00, it significantly underestimates the demand and exhibits noticeable lag in capturing the rising and falling edges. This indicates that simply sampling all hops uniformly is insufficient for modeling demand in commercial zones, where distant functionally similar areas require specialized attention.

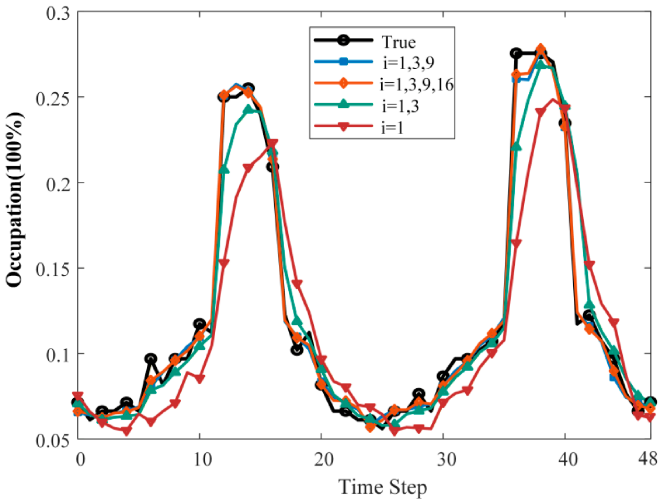


Figure 8. Prediction results under different dilation rate configurations.

When expanding the dilation set to $I = \{1,3\}$, the prediction improves, especially during the midday peak. The addition of dilation rate 3 enables the model to emphasize medium-range spatial correlations, strengthening dependencies with functionally similar commercial clusters and reducing the underestimation error. However, it still struggles with the late-afternoon surge, suggesting that dilation rate 3 alone cannot fully capture the remote yet functionally similar regions that contribute to demand peaks.

The model with $I = \{1,3,9\}$ achieves significant improvement, closely following the ground truth throughout the day. By incorporating dilation rate 9, the model successfully attends to distant commercial zones, which exhibit synchronized demand patterns despite being geographically separated by multiple hops.

Interestingly, further expanding the dilation set to $I = \{1,3,9,16\}$ yields only marginal improvements over $I = \{1,3,9\}$. The prediction curves of the two configurations are almost overlapping, indicating that adding dilation rate 16 introduces redundant information without a significant performance gain. This suggests that the spatial correlations captured through dilation rate 9 are sufficient for modeling long-range dependencies in urban charging networks, while larger dilation rates increase computational complexity without commensurate accuracy improvements.

4.4.2. Attention Weight Analysis

As shown in Figure 9, compared with the single-scale configuration (dilation rate = 1), the proposed multi-scale ADGAT (dilation rates = 1, 3, 9) produces clearer and more distinct attention patterns. The single-scale model produces smoothed, blurry structures, with attention weights evenly distributed across nodes. In contrast, the multi-scale design generates sharper and more localized patterns, which better reflect spatial differences. This indicates that multi-scale dilation helps capture more informative spatial relationships and reduces the over-smoothing problem to some extent.

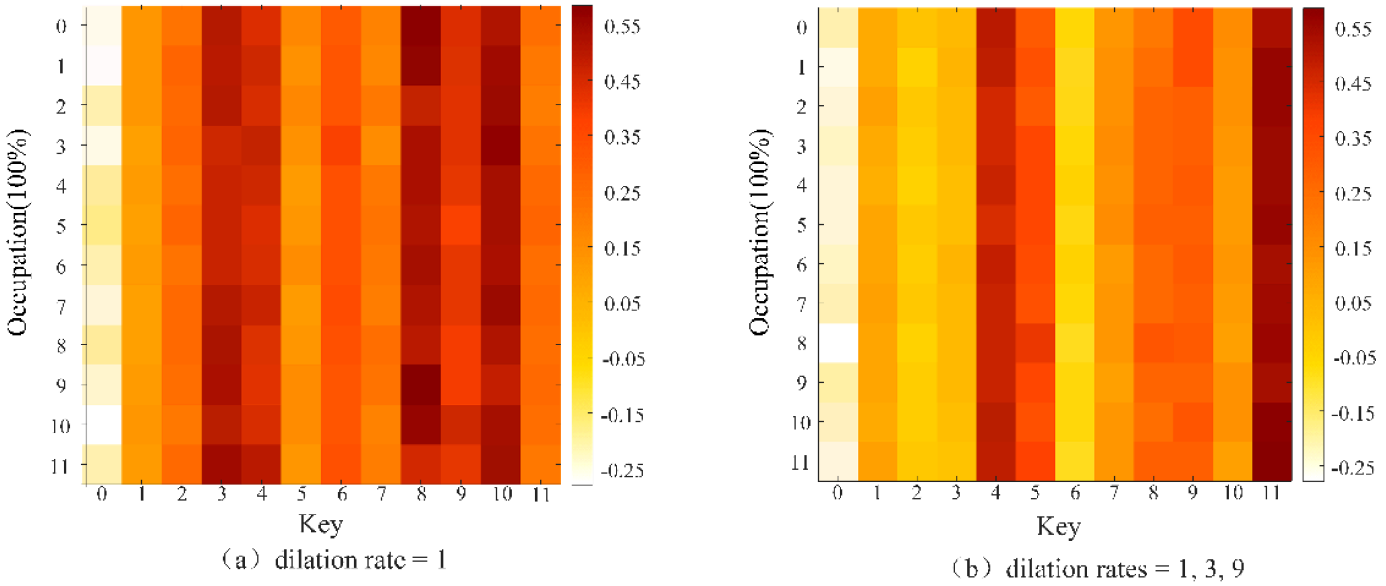


Figure 9. Visualization of Attention Weights under Single-Scale and Multi-Scale Settings.

To further investigate the attention allocation mechanism of ADGAT across different spatial scales, a more detailed analysis is conducted. Figure 10 shows the average attention-weight heatmaps across different spatial scales for Node 80, a commercial zone.

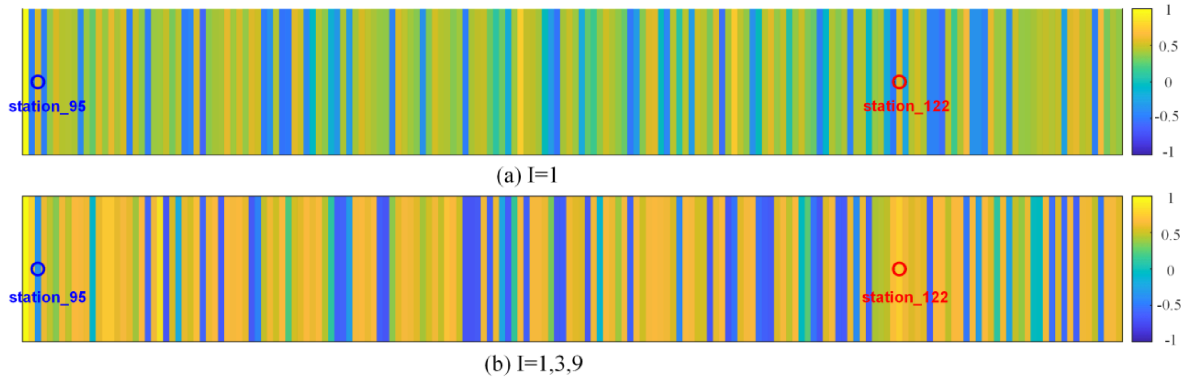


Figure 10. Attention weight visualization for Node 80 under (**a**) ($I = \{1\}$) and (**b**) ($I = \{1, 3, 9\}$).

The comparison reveals fundamental differences in attention allocation patterns. Under the single-scale configuration $I = \{1\}$, the attention weights exhibit a relatively diffuse distribution. Although the model captures certain spatial correlations, its ability to precisely discriminate between functionally relevant and irrelevant regions remains constrained. At Node 95, a residential area merely two hops away, whose demand pattern is inversely correlated with Node 80, the attention weight remains moderately elevated, suggesting insufficient suppression of this functionally dissimilar region. At Node 122, a geographically

distant commercial zone that demonstrates strong demand correlation with Node 80, the attention weight is not adequately amplified, failing to fully exploit this valuable non-local spatial dependency. Figure 11 presents the demand patterns of these representative nodes for reference.

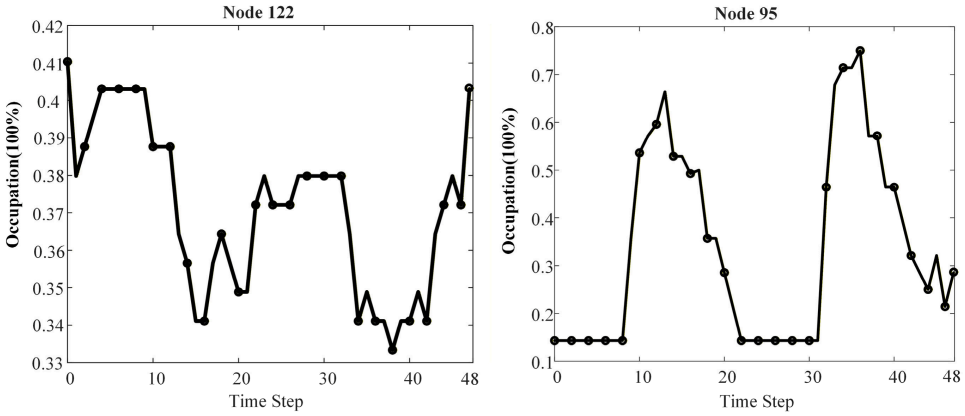


Figure 11. Demand patterns of Node 95 and Node 122.

Under the full multi-scale configuration $I = \{1, 3, 9\}$, the attention pattern becomes substantially sharper and more discriminative. At Node 95, the attention weight is markedly reduced, indicating that the multi-scale model successfully attenuates functionally dissimilar regions despite their geographical proximity. At Node 122, the attention weight is significantly enhanced, demonstrating that the model effectively identifies and prioritizes functionally similar regions regardless of spatial distance. Under the multi-scale configuration, weights for dissimilar regions are effectively suppressed, while weights for similar regions are significantly amplified. This clear contrast provides empirical evidence that ADGAT distinguishes regions based on functional similarity rather than mere geographic adjacency.

These findings confirm that integrating multiple dilation rates enables ADGAT to move beyond geographic adjacency and learn attention based on functional relevance. This enhanced representational capacity plays an important role in the forecasting performance of STGFormer.

4.5. Convergence Analysis

To evaluate the convergence behavior of STGFormer, Figure 12 presents the validation loss curves of STGFormer and TimeXer over 20 fine-tuning epochs. Both models converge smoothly without oscillation. STGFormer converges faster than TimeXer, reaching a stable validation loss approximately 5 epochs earlier, attributed to the physics-informed meta-initialization providing a well-conditioned starting point. The validation loss of STGFormer stabilizes around epoch 12 and shows no upward trend thereafter, indicating no overfitting. Note that the pre-training loss curves are excluded because the pre-training stage employs a meta-learning objective with multiple inner-loop updates per epoch involving support-query splits and task-specific gradient updates, making the pre-training loss values not directly comparable to the standard MSE loss used in fine-tuning. Only the fine-tuning validation loss is shown to provide a fair and interpretable convergence comparison.

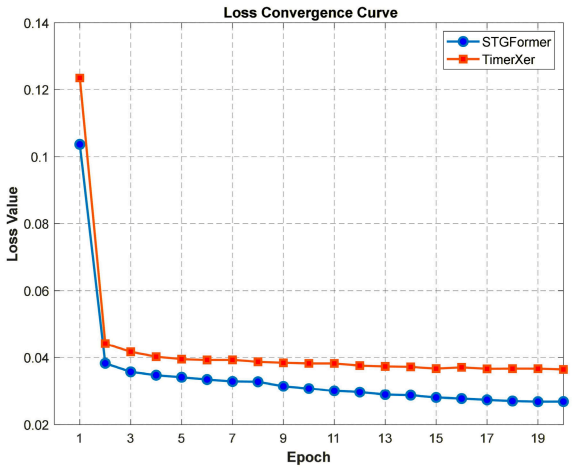


Figure 12. Validation loss curves of STGFormer and TimeXer during fine-tuning.

5. Conclusions

Accurate electric vehicle charging demand prediction is essential for grid stability and intelligent energy management. The complex spatiotemporal dependencies and heterogeneous effects of multi-source external factors make this task particularly challenging. In this study, STGFormer is proposed as a deep learning framework that integrates three complementary modules: a temporal dilated convolution module (TDConv) for extracting multi-scale local temporal patterns, an adaptive spatial dilated graph attention module (ADGAT) for mining multi-hop spatial correlations between geographically adjacent and functionally similar regions, and a hybrid xLSTM-Transformer encoder that captures both long-range dependencies and short-term fluctuations in a complementary manner.

Experimental results on a real-world dataset from Shenzhen demonstrate that STGFormer outperforms thirteen representative baseline models across all evaluation metrics, achieving average improvements of 8.58% in RMSE, 13.95% in RAE, and 14.64% in MAE. Ablation studies further validate the contribution of each proposed module, with the hybrid encoder, spatial attention module, and temporal convolution module contributing 4.1%, 3.6%, and 2.8% performance gains, respectively.

The attention weights learned by ADGAT reveal its ability to distinguish regions based on functional relevance rather than geographic proximity alone. Under the multi-scale dilation configuration, the model assigns substantially closer attention to functionally similar regions even when they are geographically distant. This discriminative attention pattern demonstrates that ADGAT captures meaningful multi-hop spatial dependencies that contribute directly to the forecasting accuracy of STGFormer.

Future research can be extended in several directions. (1) Exploring finer-grained prediction intervals, such as 15-min or real-time forecasts, to better support dynamic grid balancing and real-time pricing applications. (2) Incorporating dynamic graph structures that evolve with real-time traffic conditions or temporal changes in regional functionality to capture better time-varying spatial dependencies. (3) Developing transfer learning mechanisms that enable knowledge transfer between regions with limited data, facilitating accurate charging demand prediction in emerging urban areas.

Author Contributions: Y.L.: Writing—original draft, Visualization, Validation, Software, Methodology, Investigation. Z.Y.: Writing—original draft, Visualization, Validation, Methodology. B.L.: Writing—review & editing, Supervision, Methodology, Resources. Z.C.: Data curation, Resources, Writing—review & editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data is from the UrbanEV GitHub repository https://github.com/IntelligentSystemsLab/UrbanEV (accessed on 15 January 2026).

Conflicts of Interest: The authors declare that there is no conflict of interest in this paper.

References

1. Jiao, Z.; Li, K.; Meng, H.; Guo, Y.; Zhou, J.; Zhang, C.; Huang, Z. A methodology for lithium-ion battery state of health estimation using random constraints of state of charge. *J. Energy Storage* **2026**, *154*, 121349. [CrossRef]
2. Jia, C.; He, H.; Zhou, J.; Li, J.; Wei, Z.; Li, K.; Li, M. A novel deep reinforcement learning-based predictive energy management for fuel cell buses integrating speed and passenger prediction. *Int. J. Hydrogen Energy* **2025**, *100*, 456–465. [CrossRef]
3. Yang, Y.; Wen, Y.; Zhang, Z.; Dong, Y.; Shen, C.; Liu, Y. Toward 100% Renewable Power Grids: A Review. *IEEE Access* **2025**, *13*, 69690–69704. [CrossRef]
4. Yin, W.; Ji, J.; Wen, T.; Zhang, C. Study on orderly charging strategy of EV with load forecasting. *Energy* **2023**, *278*, 127818. [CrossRef]
5. Chen, Y.; Hu, S.; Xie, S.; Zheng, Y.; Hu, Q.; Yang, Q. Optimal dynamic pricing of fast charging stations considering bounded rationality of users and market regulation. *IEEE Trans. Smart Grid* **2024**, *15*, 3950–3965. [CrossRef]
6. Mu, Y.; Zhou, R.; Zhao, K.; Jia, H.; Zu, G.; Yang, Y. A charging demand prediction method for individual electric vehicle users based on dual-layer multi-source data clustering and a LightGBM. *Energy AI* **2025**, *22*, 100644. [CrossRef]
7. Tolun, Ö.C.; Zor, K.; Tutsoy, O. A comprehensive benchmark of machine learning-based algorithms for medium-term electric vehicle charging demand prediction. *J. Supercomput.* **2025**, *81*, 475. [CrossRef]
8. Wang, S.; Zhuge, C.; Shao, C.; Wang, P.; Yang, X.; Wang, S. Short-term electric vehicle charging demand prediction: A deep learning approach. *Appl. Energy* **2023**, *340*, 121032. [CrossRef]
9. Tian, J.; Liu, H.; Gan, W.; Zhou, Y.; Wang, N.; Ma, S. Short-term electric vehicle charging load forecasting based on TCN-LSTM network with comprehensive similar day identification. *Appl. Energy* **2025**, *381*, 125174. [CrossRef]
10. Kuang, H.; Qu, H.; Deng, K.; Li, J. A physics-informed graph learning approach for citywide electric vehicle charging demand prediction and pricing. *Appl. Energy* **2024**, *363*, 123059. [CrossRef]
11. Wang, S.; Li, Y.; Shao, C.; Wang, P.; Wang, A.; Zhuge, C. An adaptive spatio-temporal graph recurrent network for short-term electric vehicle charging demand prediction. *Appl. Energy* **2025**, *383*, 125320. [CrossRef]
12. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In Proceedings of the 5th International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
13. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
14. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. In Proceedings of the 6th International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
15. Qu, H.; Li, H.; You, L.; Zhu, R.; Yan, J.; Santi, P.; Ratti, C.; Yuen, C. ChatEV: Predicting electric vehicle charging demand as natural language processing. *Transp. Res. Part D Transp. Environ.* **2024**, *136*, 104470. [CrossRef]
16. Guo, H.; Chen, L.; Wang, Z.; Li, L. Day-ahead prediction of electric vehicle charging demand based on quadratic decomposition and dual attention mechanisms. *Appl. Energy* **2025**, *381*, 125198. [CrossRef]
17. Fang, W.; Lin, X.; Zhang, J.; Hu, J.; Huang, L.; Yuan, G. Spatiotemporal charging demand forecasting for EV stations via cross-attention fusion. *Appl. Soft Comput.* **2026**, *189*, 114475. [CrossRef]
18. Wang, S.; Chen, A.; Wang, P.; Zhuge, C. Predicting electric vehicle charging demand using a heterogeneous spatio-temporal graph convolutional network. *Transp. Res. Part C Emerg. Technol.* **2023**, *153*, 104205. [CrossRef]
19. Tang, M.; Wang, C.; Li, H.; Guo, X.; Zhang, K.; Wang, M.; Liang, X. Electric vehicle charging demand forecasting taking into account multiple time scales and dynamic road network information. *Electr. Power Syst. Res.* **2026**, *253*, 112499. [CrossRef]
20. Yang, S.Y.; Woo, J.; Lee, W. Assessing optimized time-of-use pricing for electric vehicle charging in deep vehicle-grid integration system. *Energy Econ.* **2024**, *138*, 107852. [CrossRef]
21. Tang, Q.; Guo, H.; Kirschen, D.S.; Kang, C. Understanding bidding strategies of intermittent renewables in negative price environments: A theoretical and empirical analysis. *Adv. Appl. Energy* **2025**, *17*, 100209. [CrossRef]
22. Wang, W.; Tang, A.; Wei, F.; Yang, H.; Xinran, L.; Peng, J. Electric vehicle charging load forecasting considering weather impact. *Appl. Energy* **2025**, *383*, 125337. [CrossRef]
23. Amini, M.H.; Kargarian, A.; Karabasoglu, O. ARIMA-based decoupled time series forecasting of electric vehicle charging demand for stochastic power system operation. *Electr. Power Syst. Res.* **2016**, *140*, 378–390. [CrossRef]

24. Wu, S.; Xiao, Y.; Fu, S.; Choi, J.; Zheng, C. A hybrid deep learning model for load forecasting of electric vehicle charging stations using time series decomposition. *J. Power Sources* **2025**, *655*, 237882. [CrossRef]
25. Tang, M.; Sheng, W.; Zhang, K.; Li, H.; Wang, M. Research on EV ownership and load forecasting methods based on improved Gray Markov models. *Int. J. Electr. Power Energy Syst.* **2025**, *173*, 111323. [CrossRef]
26. Kim, H.J.; Kim, M.K. Spatial-Temporal Graph Convolutional-Based Recurrent Network for Electric Vehicle Charging Stations Demand Forecasting in Energy Market. *IEEE Trans. Smart Grid* **2024**, *15*, 3979–3993. [CrossRef]
27. Jia, C.; He, H.; Zhou, J.; Li, K.; Li, J.; Wei, Z. A performance degradation prediction model for PEMFC based on bi-directional long short-term memory and multi-head self-attention mechanism. *Int. J. Hydrogen Energy* **2024**, *60*, 133–146. [CrossRef]
28. Lim, B.; Arık, S.Ö.; Loeff, N.; Pfister, T. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *Int. J. Forecast.* **2021**, *37*, 1748–1764. [CrossRef]
29. Zhou, J.; Shu, X.; Zhang, J.; Yi, F.; Jia, C.; Zhang, C.; Kong, X.; Zhang, J.; Wu, G. A deep learning method based on CNN-BiGRU and attention mechanism for proton exchange membrane fuel cell performance degradation prediction. *Int. J. Hydrogen Energy* **2024**, *94*, 394–405. [CrossRef]
30. Wang, Y.; Wu, H.; Dong, J.; Qin, G.; Zhang, H.; Liu, Y.; Qiu, Y.; Wang, J.; Long, M. TimeXer: Empowering Transformers for Time Series Forecasting with Exogenous Variables. In Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 10–15 December 2024.
31. Yang, H.; Chen, Q.; Tang, K.; Zhang, D.; Shen, Y. Flexibility aggregation and cooperative scheduling for distributed resources using a virtual battery equivalence technique. *Energy* **2025**, *334*, 137770. [CrossRef]
32. Kuang, H.; Deng, K.; You, L.; Li, J. Citywide electric vehicle charging demand prediction approach considering urban region and dynamic influences. *Energy* **2025**, *320*, 135170. [CrossRef]
33. Peng, H.; Zheng, W.; Yu, Y.; Luo, Y. Rapid seismic response prediction of steel frames based on Graph Attention Network (GAT) with Mamba blocks. *Structures* **2025**, *79*, 109390. [CrossRef]
34. Li, H.; Qu, H.; Tan, X.; You, L.; Zhu, R.; Fan, W. UrbanEV: An Open Benchmark Dataset for Urban Electric Vehicle Charging Demand Prediction. *Sci. Data* **2025**, *12*, 523. [CrossRef] [PubMed]
35. Mahmud, M.; Abedin, T.; Rahman, M.M.; Shoishob, S.A.; Sieh Kiong, T.; Nur-E-Alam, M. Integrating demand forecasting and deep reinforcement learning for real-time electric vehicle charging price optimization. *Util. Policy* **2025**, *96*, 102038. [CrossRef]
36. Yin, W.; Ji, J. Research on EV charging load forecasting and orderly charging scheduling based on model fusion. *Energy* **2024**, *290*, 130126. [CrossRef]
37. Lin, T.; Lin, R. Smart city traffic flow and signal optimization using STGCN-LSTM and PPO algorithms. *IEEE Access* **2025**, *13*, 15062–15078. [CrossRef]

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