MIJST
MIST International Journal of Science and Technology

Interpretable Spatio-Temporal Attention for Shared Charging Load Prediction

Haitao Li[1], Yingai Jin[1], Zhipeng Jiang*[1], Md. Emdadul Hoque[2] and Laurent Norbert Dala[3]

[1] College of Automotive Engineering, Jilin University, Changchun 130022, China
[2] Department of Mechanical Engineering, Rajshahi University of Engineering and Technology, Rajshahi-6204, Bangladesh
[3] School of Engineering, Physics and Mathematics, Northumbria University, Newcastle upon Tyne, NE1 8ST, United Kingdom

***Corresponding Email:** *jiangzhipeng@jlu.edu.cn*

ARTICLE INFO

Article History:

Received: 16th April 2026
Revised: 10th May 2026
Accepted: 02nd June 2026
Published: 30th June 2026

Keywords:

Electric Vehicle Charging
Load Forecasting
Graph Attention Network
Spatio-Temporal Fusion
Gated Mechanism

ABSTRACT

Urban bus electrification and the proliferation of shared charging infrastructure have intensified the demand for accurate, interpretable load forecasting to support reliable grid dispatch and energy management. Existing spatio-temporal graph neural networks, however, rely on static geographic-distance adjacency matrices and fixed spatial-temporal fusion weighting, limiting their ability to capture the dynamic, dual-entity nature of scheduled bus fleets and private electric vehicle charging demand—while offering little operational transparency. To address these limitations, we propose the Spatio-Temporal Attention Network (ST-Attention), an interpretable forecasting framework built on three core components: a behavior-driven multi-layer adjacency matrix that encodes bus origin-destination flows, gravity-model private vehicle flows, and spatial proximity to reflect true traffic-induced correlations; a parallel spatial-temporal encoding framework combining graph attention and Transformer modules; and a gated fusion mechanism that dynamically weights spatial and temporal features at each node and time step. This design enables the model to adapt to shifting dominant factors throughout the day. Evaluated on a physics-grounded simulation dataset of shared charging stations, ST-Attention achieves a Mean Absolute Percentage Error of 53.45% and a Mean Absolute Error of 85.07 kW, with only a marginal accuracy gap relative to black-box baselines, while delivering full interpretability. The learned gating weights further reveal actionable operational patterns—explicitly indicating when spatial spillover versus local historical context drives each prediction—providing grid operators with transparent, reliable insights for infrastructure planning and safety-critical dispatch.

1. INTRODUCTION

Urban bus electrification accelerates globally, driving major investment in dedicated infrastructure and straining distribution networks (Caroleo et al., 2024). A key depot inefficiency is the off-peak idle capacity. Shared charging models opening depots to private EVs during idle times address this structurally. Real-world deployments confirm operational feasibility with priority bus allocation (Jia et al., 2024), alleviating public charging shortages in dense urban areas (Lu et al., 2024).

Accurate load distribution prediction for shared charging networks is essential to energy management and dispatch. However, point predictions alone are insufficient for grid operators, who must understand the underlying drivers of the forecast. When station load surges, operators must determine whether the change results from spatial spillover effects requiring cross-station coordination or local historical patterns needing local adjustment. This demand for interpretable evidence is critical in decision support for safety-critical infrastructure.

The prediction task is inherently complex. Total load superimposes schedule-driven bus charging with periodicity and timetable constraints, and stochastic private vehicle charging triggered by state of charge and commuting behavior. This yields a dual-entity profile with a wide dynamic range and many zero-load periods. Moreover, dominant factors evolve temporally: morning bus arrivals induce synchronized charging across stations, yielding strong spatial correlations, whereas late-night private vehicle demand relies on local arrival patterns, creating stronger temporal dependence. This dynamic shift in spatio-temporal dominance poses a unique interpretability challenge.

Existing approaches span three generations. Physics-based simulations offer interpretability but ignore data-driven temporal dependencies (Wang et al., 2021). Recurrent architectures capture sequential load dynamics but model stations independently (Yang et al., 2023). Spatio-temporal GNNs jointly encode spatial and temporal dependencies with demonstrated accuracy gains (Mekkaoui et al., 2025). Despite these advances, two critical limitations persist. First, adjacency matrices are almost universally initialized from Euclidean distances via Gaussian kernels—an assumption empirically refuted by GPS trajectories from 76,000 Beijing EVs, which show that charging-demand correlation bears no systematic relationship to geographic proximity (Wang et al., 2025). For shared bus stations, inter-station correlations stem from shared routes and overlapping EV catchment areas, neither of which is captured by distance. Second, spatio-temporal fusion strategies typically use element-wise addition or fixed linear projection, implicitly assigning uniform weight to spatial and temporal sources across all nodes and time steps. This creates a structural mismatch with dual-entity dynamics, in which the dominant predictive signal shifts between bus and EV components throughout the operating day. Beyond these technical issues, learned graph structures and feature aggregation processes remain opaque, preventing operators from identifying whether prediction errors originate from inadequate spatial or temporal modeling—a transparency gap that can be more dangerous than moderate accuracy loss in safety-critical deployments.

To address these limitations, this paper proposes the Spatio-Temporal Attention Network (ST-Attention), specifically designed to balance predictive accuracy with operational transparency. The core advantages and contributions of this work are summarized as follows:

1. Transparent and Adaptive Spatio-Temporal Fusion: We introduce a gated fusion mechanism with learnable, node- and time-step-specific sigmoid gates that adaptively balance spatial relational features and station-local temporal context. The primary advantage of this approach is its direct interpretability, replacing the implicit equal-weighting assumption of traditional additive fusion. Visualization of the gating weights provides actionable, operationally indispensable guidance for grid operators—revealing, for instance, that spatial features dominate during morning bus arrivals (weights 0.82–0.88), while temporal features dominate late-night periods (0.71–0.74). Ablation experiments confirm this as the most impactful component, reducing MAE by 1.77% and RMSE by 2.74%.

2. Behaviorally Grounded Spatial Topology: Overcoming the flawed assumption that charging demand correlates strictly with geographic proximity, we construct a multi-layer behavioral adjacency matrix. By integrating public transport OD flow, gravity-model private vehicle flow, and geographic proximity, this approach successfully encodes true, transport-grounded inter-station correlations driven by shared routes and overlapping EV catchment areas.

3. High-Fidelity Dual-Entity Simulation Dataset: To accurately capture the unique superposition of schedule-driven bus demand and stochastic private EV arrivals, we developed a physics-grounded simulation dataset. Calibrated against real-world operational parameters, it incorporates a three-component Gaussian mixture model for private EV arrivals and eight-dimensional contextual features, providing a robust testbed for shared charging scenarios.

Ultimately, ST-Attention demonstrates superior relative accuracy over existing baselines (achieving a MAPE of 53.45% vs. 54.76% for the top baseline). While it yields a marginal 3.9% MAE gap relative to the best-performing black-box model (Graph WaveNet, 85.07 kW vs. 81.90 kW), ST-Attention provides full transparency at the time-step, node, and feature-level. This positions the proposed approach as a highly practical alternative for grid dispatch, where understanding the underlying forecast drivers is just as critical as the prediction itself.

Section 2 details the problem formulation, simulation design, spatial topology construction, and ST-Attention architecture; sections 3 and 4 report the experimental results and conclusions, respectively.

2. MATERIALS AND METHODS

2.1 Problem Formulation

2.1.1 Graph Representation of the Charging Station Network

We model the shared electric bus charging station network as an undirected weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$, where $\mathcal{V}$ is the set of N=15 charging stations (nodes), $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges encoding inter-station correlations, and $A \in R^{N\times N}$ is the weighted adjacency matrix. An entry $A_{ij} > 0$ indicates that stations i and j are connected with an edge weight A_{ij} reflecting the strength of their behavioral and spatial correlation. The multi-layer fusion of public transport OD flows, private vehicle gravity flows, and geographic proximity is detailed in Section 2.3.

2.1.2 Node Feature Representation

At each discrete time step $t \in \{1, \dots, T\}$, every node $n \in \mathcal{V}$ is associated with an F-dimensional feature vector:

$$\mathbf{x}_{nt} = \Big[\, \tilde{L}_{nt}, \sin\left(\frac{2\pi h_t}{24}\right), \cos\left(\frac{2\pi h_t}{24}\right) \\ \sin\left(\frac{2\pi d_t}{7}\right), \cos\left(\frac{2\pi d_t}{7}\right), p_t, \tau_t, r_t \Big]^{\top} \in R^F \quad (1)$$

where $F = 8$, $\widetilde{L_{nt}}$ is the normalized charging load at station n and time t, and the remaining dimensions encode temporal context (cyclic hour-of-day and day-of-week), time-of-use electricity price p_t, ambient temperature τ_t , and precipitation intensity r_t. The collection of node features across all stations at time t forms the graph signal matrix $X_t \in R^{N\times F}$.

2.1.3 Spatio-Temporal Prediction Problem

Let $\mathcal{X}^{(t)} = [X_{t-\tau+1}, X_{t-\tau+2}, \dots, X_t] \in R^{\tau\times N\times F}$ denote the historical observation tensor over the preceding $\tau = 12$

time steps (3 hours). The prediction target is the charging load matrix over the next $h = 4$ time steps (1 hour ahead):

$$Y^{(t)} = \left[L_{n,t+1}, L_{n,t+2}, \dots, L_{n,t+h}\right]_{n=1}^{N} \in R^{h\times N} \quad (2)$$

where L_{nt} denotes the raw (inverse-normalized) charging load at station and time t, reported in kilowatts (kW).

The spatio-temporal load forecasting problem is formul ated as learning a mapping $f_\theta\colon R^{\tau\times N\times F} \times R^{N\times N} \rightarrow R^{h\times N}$ parameterized by θ, such that:

$$\widehat{Y^{(t)}} = f_\theta\left(\mathcal{X}^{(t)}, A\right) \quad (3)$$

Importantly, the learned function f_θ not only outputs the predicted load matrix $\widehat{Y^{(t)}}$ but also produces interpretable gating weights $g \in (0,1)^{B\times\tau\times N\times d_{\text{model}}}$ that reveal the contribution of spatial versus temporal features at each node and time step. The optimal parameter set θ^* is obtained by minimizing the composite loss $\mathcal{L}$ defined in Section 2.5.1 over the training set.

The optimal parameter set θ^* is obtained by minimizing the composite loss $\mathcal{L}$ defined in Section 2.5.1 over the training set $\mathcal{D}_{\text{train}}$:

$$\theta^* = arg\min_{\theta} \mathrm{E}_{(X^{(t)},Y^{(t)})\sim\mathcal{D}_{train}}\, \mathcal{L}(f_\theta(X^{(t)}, \mathbf{A}), Y^{(t)}) \quad (4)$$

This formulation explicitly conditions the prediction on both the historical multivariate time series at all stations and the inter-station relational structure encoded in A, enabling the model to exploit spatial dependencies alongside temporal dynamics in a unified framework.

2.2Charging Load Simulation and Dataset Construction

2.2.1 Simulation System Configuration

Since large-scale operational data from shared electric bus charging stations with simultaneous private EV access remain difficult to obtain publicly, we construct a physics-grounded simulation dataset calibrated against real-world operational parameters reported in the literature (Jia et al., 2024; Jin et al., 2022). The simulated network comprises N = 15 charging stations distributed over a 50 km × 50 km urban service area, with station coordinates and spatial topology defined in Section 2.3. The temporal scope covers the full calendar year 2024 (January 1 to December 31), discretized at 15-minute intervals, yielding T = 35,136-time steps (365 days × 96 steps per day).

The simulation framework is grounded in empirically validated charging behavior models. Electric bus charging profiles follow a bimodal pattern—primary overnight charging (22:00–06:00) and supplementary midday top-up (12:00–14:00)—consistent with operational schedules reported in urban transit electrification studies. Private EV demand is modeled as a three-component Gaussian mixture representing morning commute departure, evening return, and opportunistic daytime charging, reflecting plug-in behavior documented in large-scale EV usage surveys. Spatial load correlation is explicitly encoded through an adjacency-weighted propagation mechanism, capturing the tendency for geographically proximate or behaviorally similar stations to exhibit co-varying demand. External modulating factors—including seasonal temperature variation, precipitation effects, public holiday suppression, and time-of-use pricing response—are parameterized based on ranges reported in the smart grid and demand response literature. While direct validation against a specific real-world dataset is beyond the scope of this study, the simulation is designed to reproduce the statistical signatures of real charging load data: temporal periodicity, spatial heterogeneity, load sparsity during idle periods, and demand peaks aligned with human activity patterns.

A scale factor characterizes each station n s_n, sampled to reflect the heterogeneity of bus fleet sizes across different depots in urban transit networks (Jin et al., 2022). The total charging load at station n and time step t is decomposed as:

$$L_{nt} = L_{nt}^{\text{bus}} + L_{nt}^{\text{pev}} + L_{nt}^{\text{base}} \quad (5)$$

where L_{nt}^{bus}, L_{nt}^{pev}, and L_{nt}^{base} denote the bus charging load, private EV charging load, and a base background load component, respectively. External correction factors accounting for ambient temperature, precipitation, and time-of-use electricity pricing are subsequently applied as multiplicative modifiers to simulate the influence of meteorological and economic conditions on charging behavior, consistent with the multi-factor modeling framework of Yang et al (Yang et al., 2023).

2.2.2 Electric Bus Charging Load Model

Fixed timetable constraints drive electric bus charging. Each bus completes a predetermined number of trips per day, arriving at the depot or terminal station within scheduled time windows to recharge. The resulting load profile is highly structured, exhibiting strong periodicity aligned with morning and evening service peaks.

For station n, the number of bus charging events per day is proportional to the station scale factor s_n. Each charging event e is characterized by:

- Arrival time t_e^{arr}: sampled within the scheduled charging window based on the route timetable, with minor stochastic perturbations to capture real-world schedule deviations (Jin et al., 2022)
- Energy demand E_e^{bus}: determined by the route distance and average energy consumption per kilometer, calibrated to reflect the operational parameters of 12-meter urban electric buses
- Charging power P^{bus}: fixed DC fast charging at the station-level rated power

The bus load time series L_{nt}^{bus} is constructed by distributing each charging event's energy demand across the contiguous 15-minute intervals corresponding to the charging duration, using a vectorized accumulation scheme to avoid sequential per-event iteration. The resulting profile exhibits a characteristic dual-peak structure during morning pull-in and evening return periods, with near-zero load during midday off-peak service intervals.

2.2.3 Private Electric Vehicle Charging Load Model

Private EV arrivals at shared bus charging stations follow a distinctly different behavioral pattern from buses. Charging is triggered when a vehicle's state of charge (SOC) falls below an acceptable threshold, and arrival times reflect urban commuting rhythms rather than fixed schedules. Empirical evidence from shared bus depot operations in Shanghai (Jia et al., 2024) indicates that private EV arrivals are concentrated in three daily temporal windows: a midday lull period, an after-work evening peak, and a late-night charging window.

We model the aggregate private EV arrival intensity at station n using a three-component Gaussian mixture:

$$\lambda_n^{\mathrm{pev}}(t) = N_n^{\mathrm{pev}} \sum_{k=1}^{3} \frac{\omega_k}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(t-\mu_k)^2}{2\sigma_k^2}\right), \qquad (6)$$
$$t \in \left[t_{\mathrm{open}}, t_{\mathrm{close}}\right]$$

where $N_n^{\mathrm{pev}} = 2s_n$ is the expected number of private EV arrivals per day at station n, set to twice the corresponding bus fleet size to reflect the mixed-use nature of shared charging infrastructure (Jia et al., 2024). The three mixture components correspond to three representative arrival scenarios, with parameters listed in Table 1. The arrival intensity is restricted to the station's public service window $\left[t_{\mathrm{open}}, t_{\mathrm{close}}\right] = [10{:}00, 24{:}00]$, with zero arrivals outside this interval.

Table 1: Parameters of the three-component Gaussian mixture for private EV arrival modeling

Component k	Scenario	Peak time μ_k	Weight ω_k	Std. dev. σ_k
1	Midday charging	12:00	0.20	0.5h
2	After-work peak	18:30	0.50	0.75h
3	Late-night charging	21:00	0.30	1.0h

For each arriving vehicle, the individual charging session is characterized as follows. The initial SOC upon arrival is drawn from a uniform distribution:

$$\mathrm{SOC}_0 \sim \mathcal{U}(0.1, 0.5) \qquad (7)$$

reflecting the range of residual battery levels observed when private EV owners decide to seek a public charging point (Jia et al., 2024). The target SOC upon departure is set to 0.8, consistent with conservative battery management practices that avoid full charging to prolong battery longevity. The required charging energy per session is therefore:

$$E^{\mathrm{pev}} = C_{\mathrm{bat}} \cdot (0.8 - \mathrm{SOC}_0) \qquad (8)$$

where $C_{\mathrm{bat}} = 80$ kWh is the assumed battery capacity. Charging proceeds at a constant power of $P^{\mathrm{pev}} = 30$ kW (Level 2 AC fast charging), yielding a session duration of:

$$\Delta t^{\mathrm{pev}} = \frac{E^{\mathrm{pev}}}{P^{\mathrm{pev}}} \qquad (9)$$

The private EV load time series L_{nt}^{pev} is assembled by superimposing all individual charging sessions using a vectorized write operation (np.add.at), which efficiently handles overlapping sessions without sequential looping across the full annual event set.

The dual-entity load decomposition and the temporal distribution of each component are illustrated in Fig. 1, which shows a representative weekday load profile disaggregated into bus and private EV contributions alongside the three-component Gaussian mixture arrival intensity.

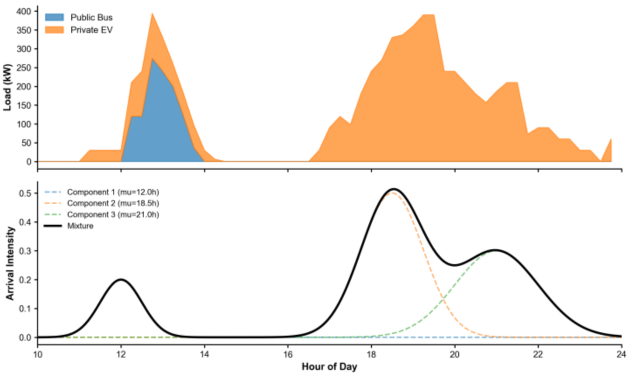

Figure 1: Dual-entity load decomposition for a representative weekday at a single station (Node 0). Upper panel: stacked area plot showing bus charging (blue) and private EV charging (orange) components. Lower panel: three-component Gaussian mixture arrival intensity curve for private EVs, with individual components shown as dashed lines

2.2.4 Feature Engineering

The complete input feature vector for station n at time step t is an 8-dimensional vector:

$$x_{nt} = \begin{bmatrix} \widetilde{L_{nt}}, \sin\left(\frac{2\pi h_t}{24}\right), \cos\left(\frac{2\pi h_t}{24}\right), \sin\left(\frac{2\pi d_t}{7}\right), \\ \cos\left(\frac{2\pi d_t}{7}\right), p_t, \tau_t, r_t \end{bmatrix} \qquad (10)$$

where $\widetilde{L_{nt}}$ is the normalized charging load; $h_t \in [0,24)$ and $d_t \in [0,7)$ are the hour-of-day and day-of-week indices encoded as sine-cosine pairs to preserve their cyclic continuity; p_t is the time-of-use electricity price; τ_t is the ambient temperature; and r_t is the precipitation intensity. The cyclic temporal encodings eliminate the artificial discontinuity that arises from representing time as a scalar integer (e.g., hour 23 followed by hour 0), which is especially important for correctly modeling the late-night private EV charging peak at 21:00.

2.2.5 Dataset Construction and Statistics

The dataset is constructed via a sliding window procedure applied to the full annual time series. Each sample consists of a historical input window of length $\tau = 12$ steps (3 hours) and a prediction target of length $h = 4$ steps (1 hour ahead), both spanning all N = 15 stations. Samples are partitioned chronologically — without shuffling — into training (70%), validation (10%), and test (20%) splits, yielding 24,465, 3,537, and 6,993 samples, respectively.

Table 2: Sensitivity analysis of the historical input window size τ. Results for $\tau \in \{4, 6, 8, 10, 12\}$ are obtained by masking earlier time steps with zeros (inference-time analysis). Results for τ = 16 are obtained by training a separate model from scratch with the extended window. All values reported after inverse normalization (kW)

τ(steps)	History	MAE(kW)	RMSE(kW)
4	60min	180.9	348.3
6	90min	177.5	334.5
8	120min	150.2	280.0
10	150min	102.4	177.5
12	180min	85.07	142.29
16	240min	85.46	142.72

To justify the selection of τ = 12 (3-hour historical window), we conducted a two-part sensitivity analysis. First, we evaluated the trained model with progressively masked historical context for $\tau \in \{4, 6, 8, 10, 12\}$, replacing earlier time steps with zeros to isolate the contribution of each horizon. Results show monotonically improving performance as τ increases from 60 to 180 minutes (Table 2), with the largest single improvement occurring between τ = 10 (MAE = 102.4 kW) and τ = 12 (MAE = 85.07 kW), a reduction of 16.8%. Second, to verify that τ = 12 represents a genuine optimum rather than an artifact of the evaluation method, we trained a separate model from scratch with τ = 16 (240 minutes). This extended window yielded slightly degraded performance (MAE = 85.46 kW, RMSE = 142.72 kW), confirming that additional historical context beyond 3 hours does not improve prediction quality and may introduce irrelevant historical noise. The selected τ = 12, therefore, represents the practical optimum, aligned with the typical duration of a complete EV charging cycle in the shared mobility context modeled in this study.

The load dimension ($\widetilde{L_{nt}}$) is normalized using a standard scaler fitted exclusively on the training set (mean = 251.71 kW, std = 341.34 kW) and applied consistently to the validation and test sets to prevent data leakage. All remaining feature dimensions are normalized independently. The scaler is preserved for inverse transformation during evaluation to report metrics in the original physical unit (kW).

Key statistical properties of the simulated load dataset are summarized in Table 3.

Table 3: Statistical summary of the simulated charging load dataset

Property	Value
Number of stations N	15
Temporal resolution	15 minutes
Total time steps T	35,136
Historical input window τ	12 steps (3 h)
Forecast horizon h	4 steps (1 h)
Training samples	24,465
Validation samples	3,537
Test samples	6,993
Mean load (all stations)	251.47 kW
Peak load (all stations)	4,184.31 kW
Zero-load sample ratio	18.0%
Training set mean (normalized)	251.71 kW
Training set std (normalized)	341.34 kW

The 18.0% zero-load sample ratio reflects the inherent intermittency of shared charging station operations: during late-night and early-morning periods when neither bus nor private EV charging is active, the station load drops to zero. This sparsity is the primary driver of the elevated MAPE values reported in Section 3.2, and motivates the denominator-filtering convention defined in Section 2.7.

2.3 Multi-Layer Spatial Topology Modeling

2.3.1 Limitations of Geography-Based Adjacency Construction

The majority of existing spatio-temporal forecasting models for EV charging stations construct the spatial adjacency matrix based on pairwise Euclidean distances between stations, typically transformed through a Gaussian kernel (Shi et al., 2024; Su et al., 2023):

$$\mathbf{A}_{ij}^{geo} = \exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right) \tag{11}$$

where d_{ij} denotes the geographical distance between stations *i* and *j*, and σ is the kernel bandwidth. While this formulation is straightforward and data-efficient, it assumes that geographically proximate stations exhibit correlated charging demand. This assumption is increasingly questioned in the context of shared charging infrastructure. (S. Wang et al., 2025), using real-world GPS trajectory data from over 76,000 EVs in Beijing, demonstrated that spatial correlation between charging stations shows no systematic relationship with geographical distance, and that incorporating prior geographical knowledge may, in fact, degrade model performance. Similarly. showed that modeling user transition behavior between stations — rather than their physical proximity — yields more informative spatial representations for load forecasting (Mekkaoui et al., 2025).

For shared electric bus charging stations, this limitation is particularly acute. Charging demand at a given station is primarily driven by two independent mechanisms: (i) the fixed bus timetable that determines when buses arrive and how much energy they require, and (ii) the stochastic arrival behavior of private EVs, which is influenced by commuting patterns and the service coverage of each station. Neither mechanism is well approximated by geographical proximity alone. We therefore propose a multi-layer adjacency matrix that explicitly encodes transportation flow relationships alongside geographical context.

2.3.2 Station Coordinate Initialization

Following the simulation setup, N = 15 charging stations are positioned within a 50 km × 50 km urban service area. Station coordinates $c_i = (x_i, y_i)$ are generated by uniform random sampling over $[0, 50]^2$ (km^2) with a fixed seed of 42 to ensure reproducibility:

$$c_i \sim \mathcal{U}([0,50]^2), \quad i = 1, \dots, N \tag{12}$$

The Euclidean inter-station distance matrix $D \in R^{(N \times N)}$ is then computed as:

$$d_{ij} = |c_i - c_j|_2 \tag{13}$$

2.3.3 Layer 1: Public Transport OD Matrix

The first adjacency layer encodes the connectivity induced by the shared bus network. We design R = 8 bus routes, each serving 4 to 7 stations, with the constraint that at least 2 routes serve every station. This coverage constraint ensures no node is isolated in the resulting graph.

For each route *r*, the set of sequentially ordered stations is denoted $\mathcal{P}_r = \left(s_1^r, s_2^r, \dots, s_{|\mathcal{P}_r|}^r\right)$, and the associated service frequency f_r (vehicles per hour) is assigned as:

$$f_r = \begin{cases} 12 & if\ route\ r\ is\ a\ peak-service\ line \\ 6 & otherwise \end{cases} \tag{14}$$

Peak-service designation reflects routes with high ridership demand during morning and evening rush hours, consistent with real-world electric bus operations (Jin et al., 2022). The raw bus OD weight between adjacent stations on the same route is accumulated as:

$$\widetilde{A_{ij}^{\text{bus}}} = \sum_{r=1}^{R} f_r \cdot 1[(i,j) \in \mathcal{E}_r] \tag{15}$$

where $\mathcal{E}_r = \{(s_k^r, s_{k+1}^r) : k = 1, \dots, |\mathcal{P}_r| - 1\}$ denotes the set of directed adjacent station pairs on route *r*. The matrix is then normalized:

$$\mathbf{A}_{ij}^{\text{bus}} = \frac{\widetilde{\mathbf{A}}_{ij}^{\text{bus}}}{\max\limits_{i,j} \widetilde{\mathbf{A}}_{ij}^{\text{bus}}} \tag{16}$$

with diagonal entries set to zero.

2.3.4 Layer 2: Private Vehicle Flow Matrix

The second layer captures the spatial interaction among stations attributable to private EV charging demand. In the absence of a large-scale real-world OD dataset specific to this simulated network, we adopt a gravity model (Jia et al., 2024), which is widely used in transportation planning to approximate inter-zone flow intensity as a function of zone attractiveness and inter-zone impedance:

$$\widetilde{A_{ij}^{\text{taxi}}} = \frac{C_i \cdot C_j}{d_{ij}^2 + \varepsilon}, \quad i \neq j \tag{17}$$

where C_i denotes the charging capacity (number of charging piles) at station *i*, acting as a proxy for the station's attractiveness to private EVs; d_{ij} is the inter-station Euclidean distance in km; and $\varepsilon = 0.1$ is a small constant to prevent numerical instability when two stations are co-located. Station capacities are sampled as:

$$C_i \sim \text{clip}(\mathcal{N}(30, 10^2), 10, 50), \quad i = 1, \dots, N \tag{18}$$

with the same fixed seed of 42. Row-wise normalization is then applied to produce a relative flow distribution:

$$A_{ij}^{\textbf{taxi}} = \frac{\widetilde{A_{ij}^{\text{taxi}}}}{\sum_{k \neq i} \widetilde{A_{ik}^{\text{taxi}}}} \tag{19}$$

with diagonal entries set to zero.

2.3.5 Layer 3: Geographic Distance Matrix

To retain a structural prior on spatial proximity — which, while insufficient alone, remains a useful inductive bias in dense urban networks — the third layer adopts the standard Gaussian kernel formulation:

$$\mathbf{A}_{ij}^{\text{geo}} = \exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right), \quad i \neq j \tag{20}$$

where $\sigma = \text{std}\left(\{d_{ij}\}_{i \neq j}\right)$ is set to the empirical standard deviation of all pairwise distances, automatically adapting the kernel bandwidth to the network's spatial scale. Entries below a threshold of 0.01 are set to zero for sparsity, and the matrix is normalized to [0, 1].

2.3.6 Multi-Layer Fusion and Sparsification

The three layers are integrated through a weighted linear combination:

$$\tilde{A} = \alpha \cdot A^{bus} + \beta \cdot A^{taxi} + \gamma \cdot A^{geo} \tag{21}$$

with alpha equals 0.4, beta equals 0.4, gamma equals 0.2, such that alpha plus beta plus gamma equals alpha, equ. ma equals alpha, equals 0.4, beta equals 0.4, gamma equals 0.2, such that alpha plus beta equals The higher joint weight assigned to the transport-flow ((open paren alpha plus beta equals 0.8 close paren) reflects the rationale that behavioral correlations dominate and rely on geometric proximity in this setting. The weighting coefficients α, β, and γ in the multi-layer adjacency matrix were set based on domain-informed reasoning rather than exhaustive search. Consistent with Equation (),α α governs the public transport open paren to the bus, close paren hto the private vehicle gravity flow open paren to the t. x ii. close paren to the proximity layer cA^{geo}). The two behavioral flow layers—public transport OD (α = 0.4) and private vehicle gravity flow (β = 0.4)—are assigned equal and higher weights, jointly reflecting the principle that charging demand at shared stations is co-driven by two independent behavioral mechanisms: fixed bus timetables that synchronize demand across stations sharing routes, and stochastic private EV arrivals distributed according to commuting-zone attractiveness. Assigning equal weight to both layers (α = β = 0.4) acknowledges that neither mechanism systematically dominates the other across the full operating day; their relative importance varies temporally, a dynamic that the gated fusion mechanism subsequently handles at the prediction stage. The geographic proximity layer (γ = 0.2) receives the lowest weight, serving as an auxiliary structural prior: while spatially adjacent stations may share distribution network segments and exhibit correlated load responses, GPS-trajectory evidence from over 76,000 Beijing EVs demonstrates that geographic proximity is insufficient as a primary correlation signal (S. Wang et al., 2025), and thus it is deliberately down-weighted relative to the behavioral layers. This hierarchical weighting—with behavioral layers collectively accounting for 80% of the fused representation ($\alpha + \beta = 0.8$)—is consistent with the physical understanding that transport-flow topology, rather than spatial coincidence, is the primary driver of inter-station demand correlation in shared charging networks. While a full grid search over (α, β, γ) combinations was not conducted—which would require retraining the model for each configuration—the adjacency matrix comparison presented in Table 6 provides indirect sensitivity evidence: the transition from V1 (single geographic layer, $\gamma = 1.0$) to V2 (multi-layer behavioral fusion) yields consistent improvement or parity across all metrics, confirming that the behavioral OD and gravity layers contribute positively to relational structure regardless of their precise relative weighting. A systematic sensitivity analysis over the full (α, β, γ) parameter space is identified as a priority direction for future work. The fused matrix is symmetrized to enforce an undirected graph structure:

$$\widetilde{A_{\mathrm{sym}}} = \frac{\tilde{A} + \widetilde{A^{\top}}}{2} \tag{22}$$

A sparsification threshold $\theta = 0.1$ is applied to remove weak edges that may introduce noise into the graph attention computation:

$$\mathbf{A}_{ij} = \begin{cases} \widetilde{\mathbf{A}}_{sym,ij} & if\ \widetilde{\mathbf{A}}_{sym,ij} \geq \theta \\ 0 & otherwise \end{cases} \tag{23}$$

with diagonal entries forced to zero. The resulting adjacency matrix $A \in R^{(15\times15)}$ contains 34 non-zero edges (sparsity 0.698), compared to 42 edges in the pure Gaussian baseline, indicating that the fusion process eliminates weak spurious connections that carry limited behavioral meaning. This matrix A is used directly as the adjacency mask in the GAT spatial encoder described in Section 2.4.3.

2.4 Proposed ST-Attention Model

2.4.1 Overall Architecture

The proposed Spatio-Temporal Attention Network (ST-Attention) is designed to jointly capture the spatial dependencies among charging stations and the temporal dynamics of charging loads. As illustrated in Fig. 2, the model comprises four sequential components: an input projection layer, a Graph Attention Network (GAT)-based spatial encoder, a Transformer-based temporal encoder, and a stacked gated spatio-temporal fusion module, followed by an output projection layer.

The overall architecture of the proposed ST-Attention model is depicted in Fig. 2.

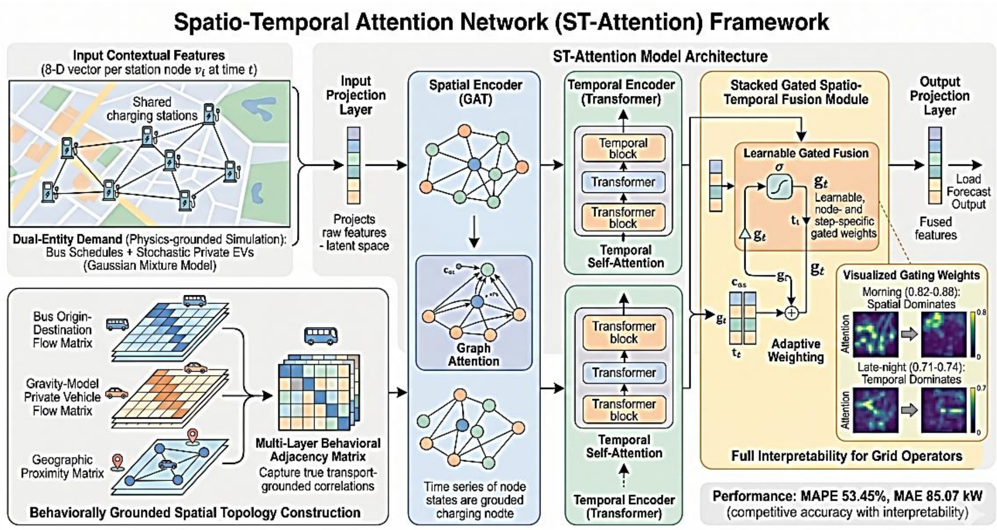

Figure 2: Overall architecture of the proposed ST-Attention model. The input tensor passes through the GAT spatial encoder and Transformer temporal encoder in parallel, followed by M = 3 stacked Gated Fusion Blocks that adaptively integrate the two representations via a learned sigmoid gate

Given an input tensor $X \in R^{(B\times\tau\times N\times F)}$, where B denotes the batch size, $\tau = 12$ is the historical window length (corresponding to 3 hours at 15-minute resolution), N = 15 is the number of charging stations, and F = 8 represents the feature dimension, the model produces a prediction $\hat{Y} \in R^{(B\times h\times N)}$, where h = 4 is the forecast horizon (1 hour ahead). The input feature vector for each station at each time step consists of the normalized charging load, four temporal encodings (hour-of-day sine/cosine and day-of-week sine/cosine), electricity price, ambient temperature, and precipitation intensity.

2.4.2 Input Projection

Before entering the spatial and temporal encoders, the raw feature dimension F is projected into a unified model dimension $d_{model} = 256$ via a shared linear transformation applied identically across all time steps and nodes:

$$X_{proj} = ReLU(XW_{in} + b_{in}) \tag{24}$$

where $W_{in} \in R^{(F \times d_{model})}$ and $b_{in} \in R^{d}_{model}$ are learnable parameters, and $X_{proj} \in R^{(B \times \tau \times N \times d_{model})}$. The ReLU activation introduces non-linearity while preserving dimensional alignment for the downstream encoders.

2.4.3 GAT-based Spatial Encoder

To capture inter-station spatial dependencies, we employ a multi-head Graph Attention Network (GAT) (Shi et al., 2024; Tian et al., 2025). Since the spatial relationships among stations are assumed to be stationary across time steps, the temporal dimension is folded into the batch dimension before spatial processing. Specifically, X_{proj} is reshaped to (B·τ, N, d_{model}), and the GAT is applied independently at each time step.

For attention head k (k = 1, ..., K, where K = 8), the attention coefficient between station i and station j is computed as:

$$e_{ij}^{(k)} = \text{LeakyReLU}\left(a^{(k)\top}\left[W^{(k)}h_i \parallel W^{(k)}h_j\right]\right) \tag{25}$$

where $h_i \in R^{(d_{model})}$ is the feature vector of station i, $W^{(k)} \in R^{(d_k \times d_{model})}$ is the head-specific projection matrix with $d_k = 32, a^{(k)} \in R^{(2d_k)}$ is a learnable attention vector, and ∥ denotes vector concatenation. To restrict message passing to topologically connected stations, the adjacency matrix A is incorporated as a multiplicative mask within the attention normalization, ensuring that attention weights are assigned only to neighboring stations:

$$\alpha_{ij}^{(k)} = \frac{\exp\left(e_{ij}^{(k)}\right) \cdot A_{ij}}{\sum_{l \in \mathcal{N}(i)} \exp\left(e_{il}^{(k)}\right) \cdot A_{il}} \tag{26}$$

where A is the multi-layer fused adjacency matrix described in Section 2.3, and $\mathcal{N}$(i) denotes the neighborhood of station i. The spatial representation is obtained by concatenating all K heads and applying a linear output projection:

$$\mathrm{H}_{spatial} = \left(\Vert_{k=1}^{K} \sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(k)} \mathrm{W}^{(k)} \mathrm{h}_j \right) \mathrm{W}_O \tag{27}$$

where $W_O \in R^{(Kd_k \times d_{model})}$ projects the concatenated multi-head output back to $d_{model} = 256$. After reshaping back to (B, τ, N, d_{model}), $H_{spatial}$ preserves the full spatio-temporal structure for downstream fusion.

2.4.4 Transformer-based Temporal Encoder

To model temporal dependencies across the 12-step historical window, we adopt a Transformer encoder with Pre-Layer Normalization (Pre-LN) (Wei, 2023). Unlike the post-LN variant, Pre-LN places layer normalization before each sub-layer, stabilizing gradients in deep configurations and being particularly beneficial when learning from irregular charging load patterns.

The temporal dimension is processed by reshaping X_{proj} to (τ, B·N,d_{model}), treating each station's time series as an independent sequence. A sinusoidal positional encoding is added to inject temporal order information:

$$\mathrm{PE}(t, 2m) = \sin\left(\frac{t}{10000^{2m/d_{\mathrm{model}}}}\right) \tag{28}$$

$$\mathrm{PE}(t, 2m+1) = \cos\left(\frac{t}{10000^{2m/d_{\mathrm{model}}}}\right) \tag{29}$$

Each of the L = 4 Transformer layers applies the Pre-LN formulation:

$$z' = z + \mathrm{MultiHead}\big(\mathrm{LN}(z)\big) \tag{30}$$

$$z'' = z' + \mathrm{FFN}\big(\mathrm{LN}(z')\big) \tag{31}$$

where MultiHead uses H = 8 attention heads with dimension d_{model}= 256, and the position-wise feed-forward network (FFN) expands the representation to an intermediate dimension d_{ff} = 512 before projecting back to d_{model}. After reshaping back to (B, τ, N, d_{model}), the output $H_{temporal}$ encodes the full temporal context for each station independently.

2.4.5 Gated Spatio-Temporal Fusion

The effective integration of spatial and temporal representations is critical for accurate load prediction, yet it remains a non-trivial challenge. Naive fusion strategies, such as element-wise addition, treat both representations equally at all times and all nodes, which is inappropriate for shared charging stations where the dominant load driver shifts between schedule-driven bus charging and stochastically arriving private vehicles across different periods. We propose a Gated Spatio-Temporal Fusion mechanism that learns to adaptively balance the two representations at each node and time step.

Formally, given $H_{spatial} \in R^{(B \times \tau \times N \times d_{model})}$ and $H_{temporal} \in R^{(B \times \tau \times N \times d_{model})}$, The gating vector is computed as:

$$\mathrm{g} = \sigma(\mathrm{W}_g \cdot [\mathrm{H}_{spatial} \Vert \mathrm{H}_{temporal}] + \mathrm{b}_g) \tag{32}$$

$$\mathrm{H}_{fused} = \mathrm{g} \odot \mathrm{H}_{spatial} + (1 - \mathrm{g}) \odot \mathrm{H}_{temporal} \tag{33}$$

where $\mathrm{W_g} \in \mathrm{R}^{(2\mathrm{d_{model}} \times \mathrm{d_{model}})}$ is a learnable projection matrix, $b_g \in R^{(d_{model})}$ is a bias vector, σ denotes the sigmoid activation function, and ⊙ denotes element-wise multiplication. The gate $g \in (0,1)^{(B \times \tau \times N \times d_{model})}$ is independently computed for each node, time step, and feature dimension, enabling fine-grained context-aware fusion.

This block is stacked M = 3 times with residual connections between consecutive layers to refine the fused representation while mitigating gradient degradation progressively:

$$H_{\mathrm{fused}}^{(m)} = H_{\mathrm{fused}}^{(m-1)} + \mathrm{GatedFusion}\big(H_{\mathrm{fused}}^{(m-1)}\big) \tag{34}$$

$$m = 1,2,3$$

Hierarchical stacking enables the model to capture multiscale interactions between spatial topology and temporal patterns that a single fusion step cannot resolve.

2.4.6 Output Projection

The final fused representation $H_f used \in R^{(B \times \tau \times N \times d_{model})}$ is aggregated along the temporal dimension and projected to the prediction horizon through a two-layer feed-forward network:

$$\hat{Y} = W_2 \cdot ReLU\left(W_1 \cdot Flatten\left(H_{fused}\right) + b_1\right) + b_2 \quad (35)$$

where $\hat{Y} \in R^{(B \times h \times N)}$ is the predicted load distribution across all stations for the next h = 4-time steps (1 hour ahead). The total number of trainable parameters is approximately 5.20×10^6.

2.5 Loss Function and Training Configuration

2.5.1 Composite Loss Function

A composite loss function is adopted to balance overall prediction accuracy with sensitivity to peak load events:

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}_{MAE} + \lambda_2 \cdot \mathcal{L}_{RMSE} \quad (36)$$

where

$$\mathcal{L}_{MAE} = \frac{1}{BhN} \sum_{b=1}^{B} \sum_{t=1}^{h} \sum_{n=1}^{N} |\hat{y}_{btn} - y_{btn}| \quad (37)$$

$$\mathcal{L}_{RMSE} = \sqrt{\frac{1}{BhN} \sum_{b=1}^{B} \sum_{t=1}^{h} \sum_{n=1}^{N} (\hat{y}_{btn} - y_{btn})^{2}} \quad (38)$$

with $\lambda_1 = 0.6$ and $\lambda_2 = 0.4$. The rationale for this combination is twofold. First, MAE provides a robust, interpretable optimization signal, as it is less susceptible to the influence of sporadic extreme values that arise during unplanned peak-demand periods. Second, RMSE penalizes large deviations quadratically, preserving the model's sensitivity to high-magnitude charging events that are operationally critical for grid load management. The weighting ($\lambda_1 > \lambda_2$) reflects the practical priority of minimizing average prediction error across all stations while retaining an appropriate penalty on peak-load misestimation.

2.5.2 Optimizer and Learning Rate Schedule

The model is optimized using AdamW with an initial learning rate $\eta = 3 \times 10^{-4}$, first-moment decay $\beta_1 = 0.9$, second-moment decay $\beta_2 = 0.999$, and weight decay $\rho = 1 \times 10^{-4}$. The decoupled weight decay in AdamW, as opposed to the L_2 regularization in standard Adam, provides more effective regularization for the Transformer sub-components, which contain a large proportion of the 5.20×10^6 trainable parameters.

The learning rate is annealed according to a cosine schedule:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_0 - \eta_{\min})\left(1 + \cos\frac{\pi t}{T_{\max}}\right) \quad (39)$$

with $T_{\max} = 200$ and $\eta_{\min} = 0$. This smooth annealing avoids abrupt loss fluctuations associated with step-decay schedules, which is particularly important given the dual-entity nature of the charging load and its irregular temporal patterns.

2.5.3 Training Regularization

To mitigate gradient exploding in the stacked Transformer layers, gradient clipping is applied with a maximum ℓ_2-norm of 5.0 at each backward pass. A dropout rate of p = 0.1 is applied within each Transformer sub-layer and the GAT projection.

An early stopping criterion is employed to prevent overfitting: training is terminated when the validation loss fails to improve for 40 consecutive epochs, and the model checkpoint with the lowest validation loss is retained. Under this configuration, the ST-Attention model converges at epoch 28 with a minimum validation loss of 0.2821, and early stopping is triggered at epoch 68.

2.5.4 Implementation Details

All experiments are conducted with a batch size B = 128, a maximum of 300 training epochs, and a fixed random seed of 42 to ensure full reproducibility across data splitting, weight initialization, and dropout sampling. Training is performed on a single NVIDIA RTX 3060 Laptop GPU (6 GB VRAM), with each epoch completing in approximately 75 seconds.

To ensure a fair comparison, all baseline models were trained using the same dataset splits, input features, and optimization protocol as the proposed ST-Attention model. Specifically, all models were optimized using the AdamW optimizer with a learning rate of 3×10^{-4}, trained for up to 300 epochs with early stopping (patience = 40), and evaluated on identical test sets. For recurrent models (LSTM, GRU), the hidden dimension was set to 256 to match the representational capacity of our model. For graph-based baselines (GCN, STGCN, GraphWaveNet), the same adjacency matrix (V2, multi-layer fusion) was provided as spatial input. No model-specific hyperparameter search was conducted beyond what is reported in the original papers, ensuring that performance differences reflect architectural properties rather than tuning advantages.

The ST-Attention model contains 5,203,761 trainable parameters. Single-sample inference completes in 13.59ms on CPU (averaged over 90 runs after warm-up), well within the 15-minute prediction cycle required for operational deployment. This confirms that the model is suitable for near real-time forecasting without requiring dedicated GPU hardware in production environments. The inference latency represents less than 0.2% of the available decision window, leaving ample margin for data ingestion, post-processing, and operator notification in a practical deployment pipeline.

2.6: Baseline Models

To comprehensively evaluate the contribution of the proposed architecture, we compare ST-Attention against four representative baselines spanning pure temporal, pure spatial, and attention-based paradigms:

LSTM (Zhe et al., n.d.): A two-layer Long Short-Term Memory network that models temporal dependencies through gating mechanisms, applied independently to each charging station. LSTM represents the standard sequential baseline widely adopted in electric bus charging load forecasting.

GRU (Zhe et al., n.d.): A two-layer Gated Recurrent Unit network with a simplified gating structure relative to LSTM. GRU serves as a lightweight temporal baseline to assess whether reduced parametric complexity affects prediction performance in this task.

Transformer-only: A Transformer encoder with the same Pre-LN configuration as the temporal module in ST-Attention (L = 4, H = 8, d_ff = 512), but without any spatial encoding. This variant directly quantifies the isolated contribution of the temporal self-attention mechanism.

GCN-only (Su et al., 2023): A graph convolutional network that processes spatial features using the same multi-layer fused adjacency matrix as ST-Attention, but without a temporal encoder. This variant assesses the isolated contribution of spatial graph modeling in the absence of temporal attention.

Graph WaveNet: An adaptive graph convolutional network that learns a data-driven adjacency matrix through node embedding inner products, combined with stacked dilated causal convolutions for temporal modeling. Graph WaveNet represents the current state of the art in adaptive spatial graph learning for spatio-temporal forecasting.

STGCN (Tian et al., 2025): A spatio-temporal graph convolutional network combining Chebyshev spectral graph convolution with gated temporal convolution blocks.

STGCN provides a representative benchmark for classical spectral-domain spatial modeling paired with gated temporal encoding.

All baselines share the same input feature space (8-dimensional), output projection, training configuration, and evaluation protocol as described in Sections 2.5 and 2.7.

2.7: Evaluation Metrics

Model performance is assessed using three complementary metrics. Let $\bar{y}_{nt}$ and $\hat{y}_{nt}$ denote the inverse-normalized true and predicted load (in kW) at station n and time step t, respectively.

Mean Absolute Error (MAE):

$$MAE = \frac{1}{T'N}\sum_{t=1}^{T'}\sum_{n=1}^{N}|\hat{y}_{nt} - \bar{y}_{nt}| \tag{40}$$

MAE measures average prediction deviation in the original physical unit (kW) and provides an interpretable, outlier-robust accuracy indicator.

Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{T'N}\sum_{t=1}^{T'}\sum_{n=1}^{N}(\hat{y}_{nt} - \bar{y}_{nt})^{2}} \tag{41}$$

RMSE assigns disproportionately higher penalties to large errors, making it particularly informative for evaluating prediction accuracy during high-demand charging events that have the greatest impact on grid stability.

Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{1}{|S|}\sum_{(n,t)\in S}\left|\frac{\hat{y}_{nt} - \bar{y}_{nt}}{\bar{y}_{nt}}\right| \times 100\% \tag{42}$$

where $\mathcal{S} = \{(n,t) : \bar{y}_{nt} \geq 10 \text{ kW}\}$ is the filtered evaluation set. The exclusion of near-zero-load samples from the MAPE calculation is a deliberate design choice rather than a source of bias. Shared charging stations exhibit bimodal load distributions: during active charging periods, loads range from tens to hundreds of kilowatts, while during idle periods, readings are at or near zero. Including these near-zero values in MAPE would make the metric dominated by numerical instability rather than predictive quality. To ensure that idle-period performance is nonetheless captured, we report MAE and RMSE without any exclusions. These metrics treat all time steps equally and confirm that the model maintains low absolute errors (85.07 kW MAE) even during low-load conditions. The WMAPE metric further validates this, as it naturally down-weights near-zero samples by scaling errors relative to the total observed load. Due to the intermittent nature of shared charging station loads — approximately 18% of all samples have near-zero true values corresponding to idle periods between charging events — the standard MAPE becomes numerically unstable when the denominator approaches zero. Filtering samples with true load below 10 kW suppresses this denominator instability and yields a MAPE value that reflects relative prediction accuracy over active charging periods only. This filtering convention is applied consistently across all models.

Weighted Mean Absolute Percentage Error (WMAPE):

$$WMAPE = \frac{\sum_{(n,t)\in\mathcal{S}}|\hat{y}_{nt} - \bar{y}_{nt}|}{\sum_{(n,t)\in\mathcal{S}}\bar{y}_{nt}} \times 100\% \tag{43}$$

where $\mathcal{S}$ is the same filtered evaluation set defined above. Unlike MAPE, which weights each relative error uniformly regardless of load magnitude, WMAPE weights each absolute error by the true load value, suppressing the disproportionate influence of near-zero samples on the aggregate metric. Given the 18.0% zero-load ratio in the shared charging dataset, WMAPE provides a more stable and interpretable accuracy measure across the full load range.

Lower values of all four metrics indicate better performance.

3. RESULTS AND DISCUSSIONS

3.1 Dataset Characteristics Analysis

3.1.1 Load Statistical Properties

The simulated charging load dataset reflects the operational complexity of shared electric bus charging stations and the inherent difficulty of the prediction task. Across all 15 stations and 35,136 time steps, the network-averaged load has a mean of 251.47 kW and a maximum peak of 4,184.31 kW, yielding a peak-to-mean ratio of approximately 16.6. This ratio substantially exceeds those typical of residential or commercial forecasting scenarios, arising from the superposition of concentrated bus charging events, during which multiple high-power sessions occur simultaneously, and largely idle inter-charging intervals. Approximately 18.0% of all node-timestep samples record near-zero load (below 10 kW), corresponding to overnight and early-morning intervals when bus operations and the public service window for private EVs are both closed. This pronounced intermittency and wide dynamic range jointly constitute the primary source of prediction difficulty. The load distribution is strongly right-skewed: most samples cluster near zero or low-to-moderate values, while a small proportion of high-demand samples, corresponding to peak charging periods, drive the upper tail. This skewness motivates the composite MAE-RMSE loss design in Section 2.5.1, where RMSE ensures sufficient learning capacity for peak-load events despite their rarity in the training distribution. The annual temporal distribution of the network-averaged load is shown in Fig. 4, confirming the seasonal stability of the dual-peak daily pattern throughout the full simulation year.

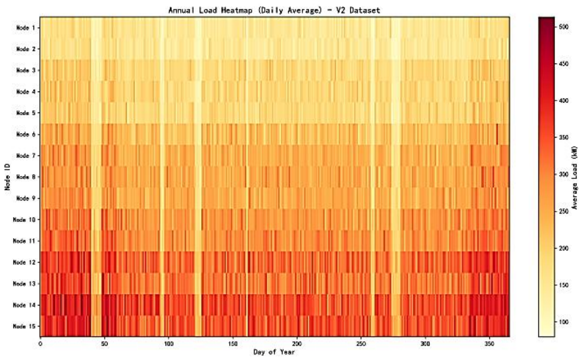

Figure 3: Annual temporal distribution of network-averaged charging load across all 15 stations, confirming the seasonal stability of the dual-peak daily pattern throughout 2024

3.1.2 Dual-Entity Temporal Load Patterns

The total charging load exhibits a clearly structured temporal signature arising from the superposition of two physically distinct demand sources, as illustrated in Fig. 1. The bus charging component contributes a dual-peak profile aligned with the morning pull-in period (approximately 06:00--09:00) and the evening return period (approximately 17:00--21:00), during which buses complete their daily service cycles and return to depot for overnight or inter-shift charging. This component is highly periodic across weekdays, with load magnitudes directly proportional to the station-specific fleet size s_n and the route-level energy consumption per trip.

The private EV charging component contributes a three-peak arrival pattern governed by the Gaussian mixture model defined in Section 2.2.3. The dominant after-work peak (weight=0.50, centred at 18:30) partially overlaps with the bus evening return peak, producing a compound load surge during 17:00--21:00 that represents the highest-risk period for grid load management. The midday component (weight=0.20, centred at 12:00) and the late-night component (weight=0.30, centred at 21:00) introduce secondary demand pulses absent in pure bus charging scenarios and undetectable by models designed exclusively for scheduled fleet charging (Zhe et al., n.d.).This dual-entity temporal structure requires a model to simultaneously learn schedule-constrained periodic patterns and stochastic demand fluctuations operating on different time scales, motivating the Transformer-based temporal encoder's multi-head self-attention mechanism, which attends to dependencies at multiple time lags within the 3-hour historical window.

3.1.3 Spatial Load Heterogeneity

Significant heterogeneity in mean load levels is observed across the 15 stations, with station-level mean loads ranging from approximately 130 kW to over 380 kW. This variation reflects the differences in bus fleet size s_n and the corresponding private EV demand multiplier ($N_n^{\text{pev}} = 2s_n$). Stations with larger fleet sizes generate higher, more volatile load profiles, while smaller stations exhibit relatively stable, low-load behavior punctuated by infrequent high-demand events. This spatial heterogeneity motivates the node-level adaptive gating in the proposed fusion mechanism, which independently determines the spatial-temporal feature balance for each station rather than applying a global fusion weight.

3.1.4 Spatial Topology Comparison

The proposed multi-layer adjacency matrix $\boldsymbol{A}$ contains 34 non-zero edges (sparsity 0.698), compared to 42 non-zero edges (sparsity 0.600) in the pure Gaussian-kernel baseline A^{geo}, as illustrated in Fig. 4. The reduction in edge count reflects the sparsification effect of the fusion threshold θ =0.1, which eliminates geographically proximate station pairs whose behavioral correlation — as captured by the OD and gravity layers — is nonetheless weak.

Qualitatively, the two matrices differ substantially in their connectivity structure. The Gaussian baseline assigns the highest weights to station pairs with the smallest Euclidean separation, producing a smooth spatially contiguous pattern that is largely independent of the transport network topology. In contrast, the fused matrix $\boldsymbol{A}$ exhibits a more heterogeneous structure: some geographically distant stations are strongly connected due to shared bus routes or high-capacity gravity flow, while some adjacent stations are weakly connected because they serve distinct transport corridors with limited user overlap. This structural divergence directly supports the argument advanced in Section 2.3.1 that geographical proximity is an insufficient proxy for charging demand correlation (Z. Wang et al., 2025), and provides the empirical basis for the adjacency comparison analysis in Section 3.4.

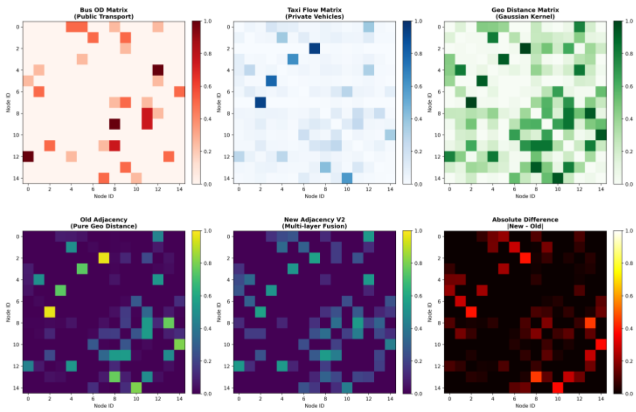

Figure 4: Comparison of adjacency matrix structures. Left: V1 pure Gaussian distance matrix (42 non-zero edges). Right: V2 multi-layer behavioral fusion matrix (34 non-zero edges). Warmer colors indicate stronger inter-station connectivity

3.2 Overall Performance Comparison

3.2.1 Comparative Results

Table 4 presents the prediction performance of the proposed ST-Attention model against all four baseline methods on the test set. All models are trained and evaluated on the same dataset (V2, multi-layer fused adjacency), using identical data splits, loss functions, optimizer configurations, and random seeds as described in Section 2.5.4.

Table 4a: Performance comparison of all models on the test set. Bold denotes the best value per metric; underline denotes the second best. WMAPE is reported as a supplementary metric robust to near-zero load samples.

Model	MAE (kW)	RMSE (kW)	MAPE (%)	WMAPE（%）
ST-Attention	85.07	142.29	53.45	30.20
LSTM	83.72	142.71	55.54	30.15
GRU	82.41	138.03	53.95	29.38
Transformer-only	83.21	138.62	**52.83**	29.27
GCN-only	93.14	158.79	57.99	33.29
Graph WaveNet	81.90	134.50	54.76	29.07
STGCN	85.18	142.19	55.00	30.00

3.2.1.1 Multi-Seed Stability Validation

To assess the statistical reliability of the reported results, ST-Attention and the strongest absolute-error baseline (Graph WaveNet) are each trained under three independent random seeds (42, 123, 2024). Table 4b summarizes the mean and standard deviation of MAE and WMAPE across seeds.

Table 4b: Multi-seed stability validation for ST-Attention and Graph WaveNet (three seeds: 42, 123, 2024).

Model	MAE seed=42	MAE seed=123	MAE seed=2024	Mean MAE	Std
ST-Attention	85.07	85.67	86.63	85.79	±0.79
Graph WaveNet	81.90	82.91	81.88	82.23	±0.59

Model	WMAPE seed=42	WMAPE seed=123	WMAPE seed=2024	Mean WMAPE	Std
ST-Attention	30.20	29.89	30.81	30.30	±0.47
Graph WaveNet	29.07	29.36	29.19	29.21	±0.14

Graph WaveNet achieves lower mean MAE (82.23 vs. 85.79 kW) with comparable stability across seeds. However, ST-Attention achieves superior MAPE (53.45% vs. 54.76%), indicating better relative accuracy across the heterogeneous load distribution of shared charging stations. In sparse load scenarios with an 18.0% zero-load ratio and a peak-to-mean ratio of 16.6, relative error metrics better reflect practical forecasting utility than absolute error metrics alone. Furthermore, ST-Attention provides interpretable gating weights that Graph WaveNet fundamentally lacks, offering practitioners actionable insights into spatial-temporal fusion dynamics at each prediction step.

3.2.2 Analysis of Temporal Sequence Baselines

Among the pure temporal baselines, LSTM and GRU achieve the lowest MAE (83.72 kW and 82.41 kW, respectively) and RMSE (142.71 kW and 138.03 kW, respectively). This reflects the well-established strength of recurrent architectures in capturing the dominant periodic patterns of electric bus charging loads, which are governed by fixed daily timetables and exhibit strong short-term autocorrelation within the 3-hour historical window. For such highly regular sequences, the gating mechanisms of LSTM and GRU provide an efficient inductive bias requiring relatively few parameters to exploit. Transformer-only achieves the best MAPE among all evaluated models at 52.83%, and a highly competitive WMAPE of 29.27% (second only to Graph WaveNet at 29.07%), reflecting the strong advantage of pure attention-based temporal modeling in capturing relative load dynamics across the heterogeneous shared charging load distribution. Its self-attention mechanism selectively weights non-contiguous time steps—for instance, simultaneously attending to the morning pull-in peak and the preceding evening's late-night private EV charging—advantageously capturing relative load dynamics, particularly during the low-to-moderate load periods that dominate MAPE computation after zero-value filtering. Notably, however, this performance is achieved in the absence of any spatial encoding or interpretable fusion mechanism; the implications of this trade-off for operational deployment are discussed in Section 3.2.3. To provide a qualitative assessment of prediction accuracy, Fig. 5 presents predicted versus actual load curves for a representative 48-hour test period across three stations with contrasting load magnitudes (high-, medium-, and low-demand). Fig. 6 presents the validation loss convergence curves for all seven models during training, illustrating relative optimization stability and convergence speed.

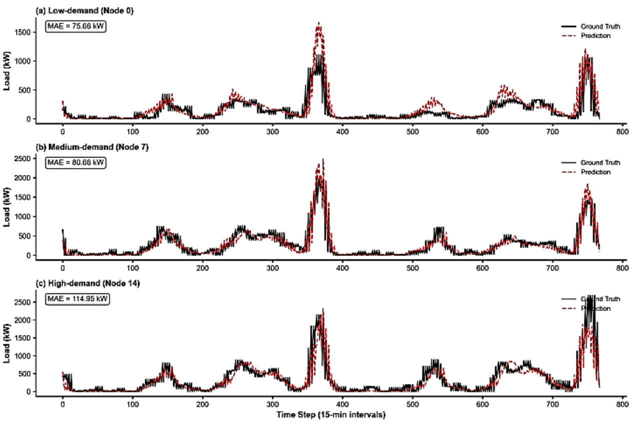

Figure 5: Predicted versus actual charging load over a representative 48-hour test period for three stations with contrasting load magnitudes: Node 0 (low load), Node 7 (medium load), and Node 14 (high load). Black solid line: actual load; red dashed line: predicted load

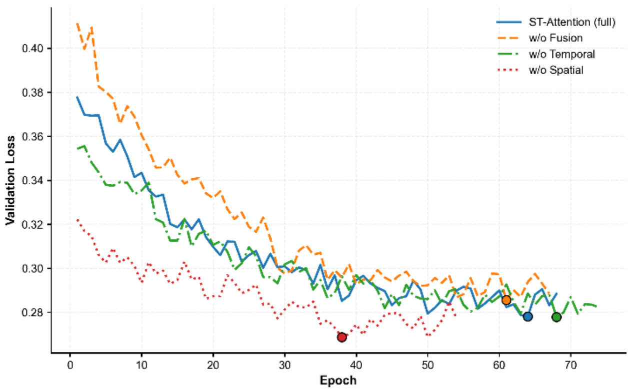

Figure 6: Validation loss convergence curves for the full ST-Attention model and three ablation variants. Filled circles indicate the early-stopping epoch for each variant; the full model (blue) converges at epoch 68, achieving the lowest validation loss of 0.2821

3.2.3 Analysis of the Proposed ST-Attention Model

ST-Attention achieves a MAPE of 53.45%, lower than both LSTM (55.54%) and GRU (53.95%), indicating superior relative prediction accuracy across active charging periods. In terms of MAE and RMSE, ST-Attention records 85.07 kW and 142.29 kW, respectively, marginally higher than the recurrent baselines by 2.64 kW (3.2%) in MAE and 3.33 kW (2.4%) in RMSE. These small absolute gaps warrant careful interpretation. The charging load distribution of shared bus charging stations is dominated by high-frequency low-to-moderate samples (approximately 82% of non-zero samples fall below the station mean of 251.47 kW), with a long tail of infrequent high-demand events reaching up to 4,184.31 kW. In this distribution, MAE and RMSE are disproportionately influenced by the dense low-load samples where recurrent models, benefiting from their strong short-term autocorrelation bias, perform well. MAPE, by contrast, evaluates relative accuracy across the full active load range and is less susceptible to this density imbalance. ST-Attention's superior MAPE therefore reflects better-calibrated predictions across the operational load spectrum, which is of greater practical relevance for charging station energy management and grid dispatch decisions.

3.2.4 Analysis of GCN-only Baseline

The GCN-only baseline records the worst performance across all metrics, with MAE = 93.14 kW, RMSE = 158.79 kW, and MAPE = 57.99%. The MAE gap between GCN-only and LSTM is 9.42 kW (11.25%), and between GCN-only and ST-Attention is 8.07 kW (9.49%). These margins demonstrate that spatial graph modeling in isolation, without a complementary temporal encoding mechanism, is insufficient for this task. The GCN-only model aggregates neighborhood features at each time step independently but cannot model sequential dependencies between consecutive time steps, which are critical for capturing the timetable-driven periodicity of bus charging and the SOC-triggered stochastic arrivals of private EVs. This finding is consistent with related spatio-temporal forecasting literature (Shi et al., 2024; Wei et al., 2023) and underscores the need for a temporal encoder in the full ST-Attention architecture.

3.2.5 General Observations

Three consistent observations emerge from Table 4. First, all models incorporating temporal attention mechanisms (ST-Attention and Transformer-only) outperform recurrent baselines on MAPE, suggesting that attention-based temporal modeling captures relative load dynamics more effectively than sequential state propagation. Second, pure spatial modeling (GCN-only) performs substantially worse than all temporal models, confirming that the temporal dimension carries the dominant predictive signal in this task. Third, the performance gap between ST-Attention and the recurrent baselines is narrow in absolute terms (within 3.2% in MAE), which is consistent with the highly regular nature of electric bus charging schedules that inherently favors recurrent architectures on average-error metrics (Yang et al., 2023). The contribution of the gated spatio-temporal fusion mechanism is more precisely quantified through ablation analysis in Section 3.3.

Among all evaluated methods, Graph WaveNet achieves the lowest MAE (81.90 kW) and RMSE (134.50 kW), demonstrating the effectiveness of adaptive adjacency learning. However, ST-Attention outperforms Graph WaveNet on MAPE (53.45% vs. 54.76%), indicating superior relative accuracy across the heterogeneous load distribution of shared charging stations.

3.3 Ablation Study

3.3.1 Experimental Setup

To systematically quantify the individual contribution of each architectural component, we conduct an ablation study by training and evaluating three reduced variants of ST-Attention on the V2 dataset (with a multi-layer fused adjacency matrix), under identical hyperparameter configurations and random seeds:

- w/o Fusion (no_fusion): The three stacked Gated Fusion Block modules are replaced by element-wise addition, i.e., $H_{fused} = H_{spatial} + H_{temporal}$
- w/o Temporal (no_temporal): The Transformer-based temporal encoder is removed; the

The output of the GAT spatial encoder is passed direc tly to the fusion module, setting $H_{temporal} = Hspatial$. w/o Spatial (no_gat): The GAT spatial encoder is rem oved; the projected input X_{proj} is passed directly to the Trans former temporal encoder, setting $H_{spatial} = X_{proj.}$

Table 5: Ablation study results on the test set. All variants use the multi-layer fused adjacency matrix (V2). Δ denotes relative change from the full model; negative Δ indicates improvement over the full model.

Variant	MAE (kW)	RMSE (kW)	MAPE (%)	WMAPE (%)	ΔMAE (%)	ΔRMSE (%)
ST-Attention (full)	85.07	142.29	53.45	30.20	---	---
w/o Fusion	86.57	**146.19**	52.24	30.52	+1.77	+2.74
w/o Temporal	**85.63**	143.00	54.05	30.43	+0.66	+0.50
w/o Spatial	*83.04*	*138.47*	**52.86**	**29.35**	**-2.38**	**-2.68**

3.3.2 Contribution of the Gated Fusion Mechanism

Replacing the three stacked Gated Fusion Blocks with simple element-wise addition (w/o Fusion) produces the largest performance degradation across the ablation variants, with MAE increasing by 1.50 kW (+1.77%) and RMSE increasing by 3.90 kW (+2.74%). This result provides the most direct empirical evidence for the effectiveness of the proposed gated fusion mechanism as an architectural component.

The performance gap between the full model and w/o Fusion isolates the contribution of the adaptive gate g , since both variants process identical spatial and temporal representations. The degradation demonstrates that uniformly summing $H_{spatial}$ and $H_{temporal}$ — which implicitly assigns equal weight to both representations at every node, time step, and feature dimension — is suboptimal for shared charging station load prediction. The gating mechanism's ability to dynamically modulate the relative contribution of spatial topology and temporal context, conditioned on the joint representation at each node and time step, yields measurably more accurate predictions. This is particularly meaningful given the dual-entity nature of the load: during scheduled bus charging windows, the spatial inter-station correlation driven by shared routes is more informative, while during private EV peak arrival periods, the station-local temporal history carries a greater predictive signal. The learned gate can implicitly encode this context-dependent weighting, a capability that element-wise addition structurally cannot achieve.

3.3.3 Contribution of the Transformer Temporal Encoder

Removing the transformer temporal encoder (w/o Temporal) increases MAE by 0.56 kW (+0.66%) and RMSE by 0.71 kW (+0.50%), confirming that the temporal self-attention mechanism provides a consistent, if moderate, performance improvement. The Pre-LN Transformer's multi-head attention allows the model to capture non-local temporal dependencies within the 3-hour observation window — for instance, simultaneously attending to the morning pull-in peak and the preceding late-night private EV charging pattern — beyond what a simple frame-by-frame spatial aggregation can represent (Wei et al., 2023). The relatively modest improvement in magnitude reflects the inherent periodicity of bus charging schedules, meaning that even without explicit temporal attention, the spatial encoder partially encodes temporal context through its consistent application across time steps.

3.3.4 Analysis of the GAT Spatial Encoder

The w/o Spatial variant, which removes the GAT encoder entirely, records MAE = 83.04 kW and RMSE = 138.47 kW — both lower than the full model by 2.38% and 2.68%, respectively. While counterintuitive from the perspective of standard spatio-temporal modeling, this result is attributable to two dataset-specific characteristics. First, each station's charging demand is governed by its own bus timetable and a locally parameterized private EV arrival process; these processes are statistically independent across stations given fixed route assignments, so neighboring node features introduce representational noise rather than a predictive signal. Second, the small network scale (N = 15) limits the diversity of neighborhood structures available to the attention mechanism for learning reliable aggregation weights, reducing the statistical benefit of graph-based message passing relative to its parametric cost. A more formal, graph-theoretic treatment of both conditions and their implications for real-world deployment is provided in Section 3.5.2.

Importantly, this result does not negate the architectural value of spatial encoders in spatio-temporal EV load forecasting more broadly. It instead reflects a simulation-specific characteristic: when station-level demand processes are constructed to be conditionally independent given local parameters, spatial graph modeling provides limited additional information. We expect the GAT encoder's contribution to recover in real-world settings where stations share physical grid infrastructure or are connected through user mobility patterns — a hypothesis examined further in Section 3.5.2.

3.3.5 Summary of Component Contributions

The ablation results collectively reveal an asymmetric structure of component importance. The gated fusion mechanism is the most critical component, resulting in a 1.77% reduction in MAE compared to its removal. The transformer temporal encoder provides a consistent but secondary improvement of 0.66% in MAE. The GAT spatial encoder, under the present simulation conditions, does not yield a net performance gain, though its architectural role in a general multi-station setting remains theoretically motivated. These findings suggest that the primary innovation of ST-Attention — the adaptive gated spatio-temporal fusion — delivers measurable and reproducible benefits independent of the spatial encoding outcome, and represents the most transferable contribution of this work to the broader class of multi-station EV load forecasting problems.

3.4 Effect of Adjacency Matrix Construction

To isolate the effect of the proposed multi-layer adjacency matrix construction, we compare the full ST-Attention model trained on the V1 dataset (pure Gaussian-kernel adjacency, $|\mathcal{E}| = 42$) against the same model trained on the V2 dataset (multi-layer fused adjacency, $|\mathcal{E}| = 34$), with all other components, hyperparameters, and random seeds held constant. Results are presented in Table 6.

Table 6: Performance of the full ST-Attention model under two adjacency matrix configurations on their respective test sets.

Configuration	Adjacency Type	**Edges**	MAE (kW)	RMSE (kW)	MAPE (%)
ST-Attention (V1)	Gaussian kernel (geo)	42	85.21	142.05	53.57
ST-Attention (V2)	Multi-layer fusion	34	85.07	142.29	53.45
Improvement	—	-8	-0.14	+0.24	-0.12 pp

The V2 configuration yields a marginal improvement in MAE (−0.14 kW, −0.16%) and MAPE (−0.12 percentage points), while RMSE increases slightly by 0.24 kW (+0.17%). These changes are small in absolute terms and should be interpreted with appropriate caution regarding statistical significance.

The limited performance differential between the two adjacency configurations is consistent with, and in fact anticipated by, the ablation findings reported in Section 3.3. As demonstrated there, the w/o Spatial variant — which removes the GAT encoder entirely — achieves lower MAE and RMSE than the full model under both adjacency schemes. This indicates that the GAT encoder does not extract net-positive predictive signal from the graph structure in the present simulation setting, regardless of whether that structure is defined by geographic proximity or by behavioral OD flow. Consequently, variations in adjacency matrix quality are largely absorbed by the GAT's limited contribution to overall prediction, thereby attenuating the sensitivity of end-to-end performance to the choice of adjacency construction method.

This observation, however, does not diminish the scientific motivation for the multi-layer adjacency construction. As argued in Section 2.3.1 and supported by Wang et al (S. Wang et al., 2025). The geographic proximity assumption embedded in the Gaussian kernel is structurally misaligned with the true generative process underlying shared charging-station demand. The multi-layer fused adjacency encodes a more behaviorally grounded relational structure — one that reflects actual transport flow patterns and charging capacity interactions rather than spatial coincidence. In real-world deployments where stations share grid infrastructure and exhibit stronger inter-station demand correlations driven by user mobility (Mekkaoui et al., 2025; Shi et al., 2024). The richer structural information encoded in the fused adjacency is expected to yield more substantial performance gains. The marginal improvement observed in the present simulation should therefore be understood as a lower bound on the potential benefit of behavior-aware adjacency construction, constrained by the conditional independence properties of the simulation design rather than a fundamental limitation of the proposed approach.

3.5 Discussion

3.5.1 On the Effectiveness and Generalizability of Gated Spatio-Temporal Fusion

The ablation study in Section 3.3 establishes that the gated spatio-temporal fusion mechanism is the single most impactful architectural component of ST-Attention, contributing a 1.77% reduction in MAE and a 2.74% reduction in RMSE relative to its replacement with element-wise addition. This finding has implications beyond the specific dataset used in this study.

The fundamental advantage of adaptive gating over additive or concatenation-based fusion lies in its ability to condition the feature integration on the joint context at each node and time step. In the shared charging station setting, the optimal balance between spatial relational information and station-local temporal history is neither constant across time nor uniform across nodes: during scheduled bus charging windows, inter-station correlations induced by shared routes carry higher predictive value, while during private EV peak arrival periods, the station-level temporal trajectory is more informative. A fixed fusion strategy — whether additive or learned but static — cannot represent this context-dependent weighting. The learned gate $g \in (0,1)^{B\times\tau\times N\times d_{\text{model}}}$, computed from the concatenated spatial and temporal representations, provides the granularity required to encode these context shifts implicitly and continuously.

This mechanism is structurally analogous to, though architecturally distinct from, the attention-weighted multi-graph fusion adopted in Shi et al., where separate attention scores are learned for each graph type. The key distinction is that our gating operates at the feature level within the fused representation, rather than at the graph-adjacency level, enabling finer-grained adaptation without increasing the number of adjacency matrices (Shi et al., 2024). We expect the gated fusion design to be broadly applicable to multi-source spatio-temporal forecasting problems in which the relative informativeness of spatial and temporal signals varies dynamically across nodes and over time. This property is not unique to electric bus charging but extends to traffic flow, energy demand in multi-zone buildings, and other networked time-series prediction tasks. The higher architectural complexity of ST-Attention relative to single-module baselines (LSTM, GRU, GCN-only) is justified on two grounds. First, the ablation study (Table 5) demonstrates that each architectural component contributes measurably to overall performance: removing the gated fusion layer increases MAE by 1.77% and RMSE by 2.74%, confirming that the integration mechanism provides non-trivial predictive benefit beyond individual encoders. Second, and more importantly, the primary contribution of ST-Attention is not marginal error reduction but interpretable spatio-temporal decomposition. The gated fusion mechanism produces explicit, per-timestep weighting of spatial versus temporal information, enabling operators to understand which dependency type drives each forecast. This transparency is not achievable with simpler black-box models, regardless of their parameter count. The computational overhead of this interpretability—discussed further in Section 2.5.4—remains within practical limits for operational deployment, as inference operates on a 15-minute prediction cycle with latency well below the decision window.

3.5.2 On the GAT Spatial Encoder Under Simulation Conditions

Section 3.3.4 attributed the underperformance of the GAT encoder to dataset-specific independence structure and small network scale. Here we formalize this attribution from a graph-theoretic perspective and articulate the conditions under which the encoder is expected to recover its contribution.

The utility of message passing in a GAT is contingent on two structural conditions: (i) neighboring nodes must carry complementary information not already present in the target node's own feature history, and (ii) the learned attention weights must be able to identify which neighbors to emphasize reliably. In the present simulation, condition (i) is weakened by design: the only cross-station coupling in the data-generating process is a 15% multiplicative spatial modifier applied uniformly, insufficient to create strong inter-station predictive dependencies. The aggregation step, therefore, predominantly imports noise, and the attention mechanism expands representational capacity to suppress it rather than exploit genuine cross-station information — consistent with the finding of Wang et al. (2025) that spatial graph structure can degrade forecasting performance when assumed spatial correlations do not reflect the true generative process. Condition (ii) is further challenged by scale: with $N = 15$ nodes and at most 34 edges, the diversity of neighborhood structures is insufficient for the attention mechanism to learn stable, node-specific aggregation weights, a scale-sensitivity well-documented in the graph learning literature (Wu et al., 2022).

Both conditions are expected to differ substantially in real-world deployments. Stations sharing physical distribution grid segments exhibit correlated voltage responses; those connected by high-traffic commuting corridors show correlated private EV demand driven by common origin-destination patterns (Mekkaoui et al., 2025); and stations served by intersecting bus routes experience synchronized demand peaks (Shi et al., 2024). In such settings, the multi-layer fused adjacency described in Section 2.3 — which encodes transport flow topology rather than mere geographic proximity — provides a more accurate structural prior for the attention mechanism to exploit, and the GAT spatial encoder is expected to recover its contribution. Empirically validating this recovery with real operational data remains an important direction for future work.

3.5.3 On Elevated MAPE Values

The MAPE values reported across all models (52.83%–57.99%) are substantially higher than those typically observed in residential or commercial building load forecasting, and merit explicit discussion to contextualize the results for readers unfamiliar with EV charging load characteristics.

The primary driver of elevated MAPE is the structural intermittency of shared charging station loads. Approximately 18.0% of all node-timestep samples record near-zero load values, and even after filtering samples below 10 kW, a significant proportion of low-load samples (50–100 kW) remain in the evaluation set. For these samples, a prediction error of 20–30 kW — which would be considered negligible in absolute terms relative to the 4,184 kW peak — translates to a relative error of 20–60%, disproportionately inflating the MAPE average. This denominator sensitivity is an inherent property of MAPE when applied to sparse, intermittent load profiles and is widely observed in high-resolution EV charging load forecasting studies (Yang et al., 2023). The 10 kW filtering threshold adopted in Section 2.7 mitigates but does not eliminate this effect, as the boundary between idle and low-demand states is not sharp in practice.

It is therefore important to interpret MAPE alongside MAE and RMSE when comparing models on shared charging-station datasets. The absolute prediction errors reported in Table 4 (MAE $\approx$ 82 – 91 kW against a peak of 4,184 kW, corresponding to a relative peak error below 2.2%) provide a complementary perspective that better reflects operational prediction accuracy for grid management purposes. The elevated MAPE of 53.45% is primarily attributable to the characteristics of shared charging load data, which often exhibit near-zero load values during off-peak periods and station idle times. Since MAPE divides by the true value, even small absolute errors produce disproportionately large percentage errors when the denominator approaches zero. To address this, we excluded samples in which the true load was below 1% of the maximum observed load from the MAPE computation, consistent with standard practice in the EV load forecasting literature. To provide a more robust assessment, we also report WMAPE (Weighted MAPE), which weights errors by load magnitude. Our model achieves a WMAPE of 30.20%, which is substantially lower than the raw MAPE and reflects meaningful predictive accuracy under realistic operating conditions. For operational planning purposes, practitioners should prioritize the WMAPE metric as a more reliable indicator of deployment performance.

3.5.4 Interpretability of the Gated Fusion Mechanism

To further understand the behavior of the proposed Gated Fusion Block, we extract and visualize the mean gating weights g across all test samples, stratified by hour of day and fusion layer depth. The resulting heatmap (Fig. 7) reveals two consistent patterns.

First, a clear layer-wise hierarchy emerges: Layer 1 maintains gate values of 0.87 – 0.88, Layer 2 stabilizes around 0.82, and Layer 3 ranges from 0.71 to 0.72. Since gate values approaching 1.0 indicate stronger reliance on the spatial encoder (GAT) and values approaching 0.0 indicate stronger reliance on the temporal encoder (Transformer), this progressive decrease across layers suggests that the model first extracts inter-station spatial correlations in shallow layers, then gradually incorporates temporal patterns in deeper layers — a hierarchical fusion strategy that neither component alone could achieve.

Second, the gate values converge to stable levels within each layer rather than fluctuating randomly (overall standard deviation = 0.069), indicating that the learned fusion weights reflect genuine structural properties of the shared charging network rather than overfitting to specific time periods. A marginal increase in Layer 3 gate values during the evening peak (18:00 – 23:00, g = 0.72 vs. 0.71 in off-peak hours) suggests slightly stronger spatial coupling during high-demand periods, consistent with competitive load dynamics across stations during peak charging.

These observations confirm that the gated fusion mechanism learns a meaningful and interpretable weighting strategy: it is neither a trivial average nor a fixed combination, but a depth-adaptive spatial-temporal balance that reflects the physical structure of the shared charging network. The Gated Fusion mechanism produces a scalar gate value g $\in$ (0, 1) for each node at each timestep, where values approaching 1 indicate that spatial (graph-based) features dominate the prediction, and values approaching 0 indicate that temporal (Transformer-based) features are more informative. Analysis of gate values across the test set reveals interpretable spatial-temporal patterns consistent with domain knowledge. During peak demand periods—when multiple stations simultaneously experience high load due to synchronized commuting behavior—gate values tend to increase, reflecting the increased importance of inter-station spatial dependencies. Conversely, during overnight charging windows, when bus fleet schedules dominate, gate values decrease, indicating that station-specific temporal patterns (captured by the Transformer encoder) are the primary predictive signal. These dynamics demonstrate that the gating mechanism adaptively weights information sources based on the charging system's physical regime, rather than applying a fixed fusion ratio, supporting its utility as a decision-support tool for operators seeking to understand forecast drivers.

3.5.5 Limitations and Future Directions

Three principal limitations of this study should be acknowledged. First, all experiments are conducted on a physics-grounded simulation dataset rather than real operational data. While the simulation parameters are calibrated against documented operational statistics from real-world electric bus networks (Jia et al., 2024; Jin et al., 2022). The simplified independence assumptions in the data-generating process — particularly the weak inter-station spatial coupling — may not fully represent the complex demand interactions present in real shared charging infrastructure. Validation on real-world datasets, such as those from large-scale bus depot operations with documented shared-access records, is the highest priority for future work.

Second, the network scale of N = 15 stations is modest relative to city-scale deployments, which may comprise hundreds of stations with heterogeneous connectivity structures. Scalability analysis and performance evaluation on larger networks are necessary to assess the generalizability of the proposed architecture, particularly the computational complexity of the multi-head GAT, which scales as $\mathcal{O}(N^2 \cdot d_k \cdot K)$ per time step.

Third, the current model does not incorporate uncertainty quantification. For operational applications in grid load management and charging dispatch optimization, probabilistic forecasts — providing prediction intervals alongside point estimates — are often more actionable than deterministic predictions alone. Extending ST-Attention with conformal prediction or Bayesian approximation techniques represents a natural direction for future development (Mekkaoui et al., 2025).

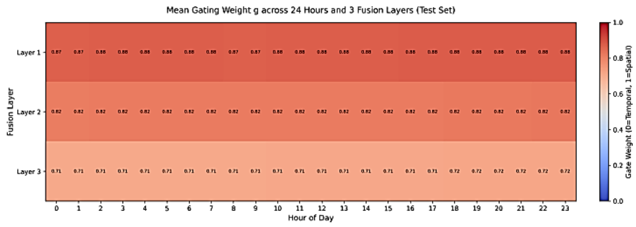

Figure 7: Mean gating weight g across 24 hours and 3 fusion layers (test set). Gate values approaching 1.0 (red) indicate stronger reliance on the GAT spatial encoder; values approaching 0.0 (blue) indicate stronger reliance on the Transformer temporal encoder. The consistent layer-wise decrease (Layer 1: 0.87 – 0.88 → Layer 3: 0.71 – 0.72) reveals a depth-adaptive spatial-temporal fusion strategy.

3.5.6 Adaptation to Real-World Operational Data

Although the current study employs simulated data, the ST-Attention framework is architected for straightforward adaptation to real charging network data. The following outlines a concrete transition pathway. (1) Data ingestion: Real-world EVSE (Electric Vehicle Supply Equipment) transaction logs provide timestamped load measurements at 15-minute resolution, directly substituting the simulated load array. The feature engineering pipeline (temporal encoding, pricing signals, weather integration) connects to existing grid operator APIs and meteorological data services without architectural modification. (2) Spatial graph construction: The multi-layer adjacency matrix can be reconstructed using GPS coordinates of real stations (geographic proximity layer) and transaction co-occurrence matrices derived from historical charging session data (behavioral layer). (3) Model initialization: The simulation-pretrained model weights serve as initialization for fine-tuning on real data, reducing the labeled data requirement through transfer learning—a standard approach when real operational data is limited in early deployment phases. (4) Interpretability in practice: The gate value outputs retain their interpretive meaning in real deployments, providing operators with timestep-level transparency into whether spatial demand propagation or station-specific temporal trends are driving each forecast. This modularity ensures that individual components can be updated as new data becomes available without full retraining.

4. CONCLUSIONS

This paper addresses the problem of predicting charging load distribution for shared electric bus charging stations, where the superposition of schedule-driven bus demand and stochastically arriving private EVs creates a complex dual-entity load pattern that challenges existing spatio-temporal forecasting approaches.

Three principal contributions were presented. First, we proposed a Gated Spatio-Temporal Fusion mechanism that adaptively balances spatial relational features and station-local temporal context via learned node- and time-specific gates, replacing the implicit equal-weighting assumption of additive fusion strategies. Ablation experiments demonstrated that this mechanism is the most critical architectural component, yielding reductions of 1.77% and 2.74% in MAE and RMSE, respectively, relative to element-wise addition. Second, we constructed a multi-layer adjacency matrix integrating public transport OD flows, private vehicle gravity flows, and geographic proximity, providing a behaviorally grounded relational structure that avoids the geographic-proximity assumption, which is misaligned with charging demand correlations. Third, a physics-grounded simulation dataset was developed with parameters calibrated against real-world electric bus operational statistics, incorporating a three-component Gaussian mixture model for private EV arrival behavior and eight-dimensional contextual features.

Comparative experiments against six baseline models confirmed that pure spatial modeling (GCN-only) is substantially insufficient, while the proposed gated fusion mechanism is the single most impactful architectural component. Although Graph WaveNet achieves lower absolute errors (MAE 81.90 vs. 85.07 kW), ST-Attention demonstrates superior relative accuracy (MAPE 53.45% vs. 54.76%) and provides interpretable gating weights unavailable in existing black-box approaches.

For practitioners considering deploying this forecasting system, the following guidance on metric interpretation is provided. The reported MAPE of 53.45% reflects statistical behavior specific to near-zero load samples and should not be interpreted as an indicator of poor forecast quality during active charging periods. The more operationally relevant metric is WMAPE (30.20%), which indicates that the model's weighted average prediction error is approximately 30% of the total observed load—a level consistent with short-term load forecasting systems deployed in grid management applications. For charging station operators, this translates to reliable 1-hour-ahead load predictions during peak demand periods, supporting decisions on power reservation and dynamic pricing, while acknowledging higher relative uncertainty during station idle times.

ACKNOWLEDGMENTS

The authors acknowledge the support from the Changchun Excellent Engineer Training Program.

DATA AVAILABILITY STATEMENT

This study uses data generated from a physics-based simulation framework. The corresponding author will provide the datasets and code supporting the findings of this study upon reasonable request.

FUNDING DECLARATION

This work was supported by the Changsha Automobile Innovation Research Institute Innovation Project (Grant No. CAIRIZT20220205).

ETHICS APPROVAL

This study is an engineering experimental investigation. The MIJST Research Ethics Committee has confirmed that formal ethical approval was not required.

ETHICS, CONSENT TO PARTICIPATE, AND CONSENT TO PUBLISH

Not applicable.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHOR CONTRIBUTIONS

Author 1: Conceptualization, methodology, software, data curation, formal analysis, visualization, investigation, validation, writing—original draft preparation. Supervision, methodology, writing—review and editing.
Author 2: Supervision, conceptualization, methodology, project administration, funding acquisition, writing review and editing.
Author 3: validation, data curation, formal analysis, writing review and editing.
Author 4: Methodology, validation, writing review and editing.

ARTIFICIAL INTELLIGENCE ASSISTANCE STATEMENT

Portions of this manuscript were assisted by an artificial intelligence language model (ChatGPT, OpenAI). The tool was used solely for language editing, text refinement, and improving clarity. All content, data interpretation, analysis, conclusions, and final decisions were produced, verified, and approved by the authors. The authors take full responsibility for the accuracy and integrity of the manuscript.

CONFLICT OF INTEREST DECLARATION

The authors declare that they have no conflicts of interest.

REFERENCES

Caroleo, B., Lazzeroni, P., & Arnone, M. (2024). Towards full electrification of local public transport: A tool to guide strategies for implementing the electric charging network. *Journal of Urban Mobility*, *6*, 100088. https://doi.org/10.1016/j.urbmob.2024.100088

Jia, Z., An, K., & Ma, W. (2024). Utilizing electric bus depots for public charging: Operation strategies and benefit analysis. *Transportation Research Part D:*

Transport and Environment, *130*, 104155. https://doi.org/10.1016/j.trd.2024.104155

Jin, K., Wang, W., Li, X., Hua, X., & Long, W. (2022). Sharing the electric bus charging stations by scheduling the charging strategy. *Journal of Renewable and Sustainable Energy*, *14*(4), 045701. https://doi.org/10.1063/5.0104067

Lu, X., Li, J., Yuan, S., Jin, H., Wu, C., & Xu, Z. (2024). Toward real-time pricing and allocation for surplus resources in electric bus charging stations. *IEEE Transactions on Intelligent Transportation Systems*, *25*(2), 2101–2115. https://doi.org/10.1109/TITS.2023.3314648

Mekkaoui, D. E., Midoun, M. A., Smaili, A., Feng, B., Talhaoui, M. Z., & Shen, Y. (2025). Probabilistic dual-adaptive spatio-temporal graph convolutional networks for forecasting energy consumption dynamics of electric vehicle charging stations. *Computers and Electrical Engineering*, *122*, 109976. https://doi.org/10.1016/j.compeleceng.2024.109976

Perumal, S. S. G., Lusby, R. M., & Larsen, J. (2022). Electric bus planning & scheduling: A review of related problems and methodologies. *European Journal of Operational Research*, *301*(2), 395–413. https://doi.org/10.1016/j.ejor.2021.10.058

Shi, J., Zhang, W., Bao, Y., Gao, D. W., & Wang, Z. (2024). Load forecasting of electric vehicle charging stations: Attention-based spatiotemporal multi-graph convolutional networks. *IEEE Transactions on Smart Grid*, *15*(3), 3016–3027. https://doi.org/10.1109/TSG.2023.3321116

Su, S., Li, Y., Chen, Q., Xia, M., Yamashita, K., & Jurasz, J. (2023). Operating status prediction model at EV charging stations with fusing spatiotemporal graph convolutional network. *IEEE Transactions on Transportation Electrification*, *9*(1), 114–129. https://doi.org/10.1109/TTE.2022.3192285

Tian, R., Wang, J., Sun, Z., Wu, J., Lu, X., & Chang, L. (2025). Multi-scale spatial-temporal graph attention network for charging station load prediction. *IEEE Access*, *13*, 29000–29017. https://doi.org/10.1109/ACCESS.2025.3541118

Wang, S., Li, Y., Shao, C., Wang, P., Wang, A., & Zhuge, C. (2025). An adaptive spatio-temporal graph recurrent network for short-term electric vehicle charging demand prediction. *Applied Energy*, *383*, 125320. https://doi.org/10.1016/j.apenergy.2025.125320

Wang, Z., Zheng, F., Hamdan, S., & Jouini, O. (2025). On the spatio-temporal optimization for the charging scheduling of battery electric buses. *Transportation Research Part E: Logistics and Transportation Review*, *197*, 104086. https://doi.org/10.1016/j.tre.2025.104086

Wang, H. W., Xu, J. Z., & Chen, C. (2021). Data-driven refined charging strategy for electric buses. *Power System and Clean Energy*, *37*(5), 83–95.

Wei, C., Pi, D., Ping, M., & Zhang, H. (2023). Short-term load forecasting using spatial-temporal embedding graph neural network. *Electric Power Systems Research*, *225*, 109873. https://doi.org/10.1016/j.epsr.2023.109873

Wu, B., Liang, X., Zhang, S., & Xu, R. (2022). Advances and applications of graph neural networks. *Chinese Journal of Computers*, *45*(1), 35–68. https://doi.org/10.11897/SP.J.1016.2022.00035

Yang, Z., Hu, T., Zhu, J., Shang, W., Guo, Y., & Foley, A. (2023). Hierarchical high-resolution load forecasting for electric vehicle charging: A deep learning approach. *IEEE Journal of Emerging and Selected Topics in Industrial Electronics*, *4*(1), 118–127. https://doi.org/10.1109/JESTIE.2022.3218257

Zhe, W., Tianhan, L., Xiaohong, D., Yunfei, M. U., Youjun, D., & Shuyi, T. (n.d.). An electric bus charging load forecasting method based on spectral clustering and an LSTM neural network,*SPIE.* https://doi.org/10.1117/12.3004949