

A Behavior-Guided Online Probabilistic Forecasting Method for Electric vehicle Charging Loads

Chenghan Li, *Graduate Student Member, IEEE*, Qingxiang Liu, *Graduate Student Member, IEEE*,
Yinliang Xu, *Senior Member, IEEE*, and Yuxuan Liang, *Member, IEEE*

Abstract—Electric vehicle (EV) charging loads exhibit strong behavioral heterogeneity and temporal variability, posing significant challenges for online probabilistic forecasting under evolving operating conditions. In particular, persistent charging patterns may differ substantially across stations, while recent behavioral changes can continuously alter the underlying load distributions. This paper proposes a behavior-guided online probabilistic forecasting framework that explicitly characterizes persistent station-specific patterns and recent behavioral changes. A dual-timescale behavior representation is constructed to distinguish long-term charging characteristics from recent behavioral states and quantify their deviations. These behavioral changes are further semantically encoded to guide drift-aware forecasting adaptation, while a delayed-feedback mechanism ensures temporally consistent online updates when observations become available across different forecasting horizons. Experiments on ten heterogeneous real-world charging stations demonstrate that the proposed method consistently outperforms conventional forecasting models and concept-drift-aware online baselines in forecasting accuracy and probabilistic reliability. For 1-h-ahead forecasting, the proposed method reduces MSE and Pinball loss by 15.3% and 17.8%, respectively, over the corresponding best baselines. For 4-h-ahead forecasting, the improvements further reach 16.8% and 22.6%, respectively, demonstrating consistent performance gains under evolving charging behaviors and extended forecasting horizons.

Index Terms—EV charging load forecasting, online forecasting, charging behavior, concept drift.

I. INTRODUCTION

The rapid penetration of EVs is reshaping electricity demand and introducing increasingly flexible yet uncertain loads into power systems [1], [2]. Accurate forecasting of EV charging loads is therefore essential for charging infrastructure operation, demand-side management, and reliable grid-operation [3], [4]. Unlike conventional loads, EV charging demand is strongly driven by individual mobility and charging decisions, resulting in substantial heterogeneity in charging time, duration, energy demand, and charging frequency across users and stations [5], [6]. Such behavioral heterogeneity can propagate to the grid level, producing markedly different temporal and spatial load profiles and increasing system-level load variability [7]. Moreover, charging behaviors are not necessarily stationary: changes in travel routines, charging preferences, infrastructure availability, and operating conditions can continuously reshape the resulting charging patterns [8], [9].

Consequently, future charging loads may deviate substantially from historically learned patterns, making them difficult to characterize using deterministic forecasts alone. Probabilistic forecasting, which represents predictive uncertainty through conditional distributions, quantiles, or prediction intervals, therefore provides a more informative characterization of future charging demand and is increasingly important for reliable EV charging management [10].

Existing studies have extensively investigated EV charging load forecasting from temporal, spatial, and probabilistic perspectives. Wu et al. [11] proposed a VMD-SSA-SVR framework that decomposes historical charging loads and incorporates real-time electricity prices and ambient temperature to improve short-term forecasting accuracy. However, its decomposition and regression pipeline relies primarily on predefined numerical features and does not explicitly characterize heterogeneous charging behaviors. Wang et al. [12] developed an LSTM-based model using large-scale EV trajectory data to capture temporal dependencies in station-level charging demand. Although effective for short-term prediction, the model mainly learns historical temporal regularities and provides limited capability to quantify predictive uncertainty. Cao et al. [13] proposed a feature-enhanced deep learning framework that incorporates correlations between charging demand and external factors into probabilistic forecasting. However, the learned feature relationships are mainly derived from historical data and do not explicitly describe how station-specific charging behaviors evolve over time. Li et al. [10] introduced a conditional diffusion model to approximate the predictive distribution of EV charging loads and generate reliable probabilistic forecasts under stochastic charging demand. Nevertheless, its focus is primarily on modeling predictive distributions rather than adapting the forecasting process to continuously evolving behavioral patterns. Dai and Shi [14] developed a multi-source spatio-temporal framework that integrates heterogeneous station attributes, historical observations, and inter-station dependencies for fine-grained charging demand forecasting. However, the model is primarily designed to learn spatio-temporal relationships from predefined inputs, while temporal changes in station-specific behavior are not explicitly modeled. More recently, Tu et al. [15] proposed a Kolmogorov–Arnold spatio-temporal attention recurrent network to jointly capture spatial dependencies and nonlinear temporal patterns for multi-target EV charging load forecasting. Despite its enhanced spatio-temporal representation, the forecasting mechanism still mainly relies on learned historical dependencies and does not explicitly distinguish persistent

Chenghan Li and Yinliang Xu are Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China (Corresponding author e-mail: xu.yinliang@sz.tsinghua.edu.cn)

Qingxiang Liu and Yuxuan Liang is with the Intelligent Transportation Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511458, China.

charging patterns from recent behavioral deviations. Overall, these studies have substantially advanced EV charging forecasting, but most focus on improving temporal, spatial, or uncertainty representations under historically learned patterns.

Beyond improving forecasting architectures under historically learned patterns, recent studies have increasingly investigated online forecasting to accommodate non-stationarity and concept drift. Pham et al. [16] proposed FSNet, which combines fast adaptation with associative memory to respond to new patterns while retaining recurring knowledge. However, its adaptation is driven primarily by changes in temporal representations rather than explicit behavioral information. Zhang et al. [17] introduced OneNet to dynamically ensemble complementary forecasters for improved adaptation to concept drift, but the detected distribution changes remain difficult to associate with interpretable behavioral variations. Lau et al. [18] developed a dual-stream framework that separates fast adaptation from slow learning and avoids information leakage in online forecasting. Nevertheless, its adaptation mainly focuses on balancing recent and historical temporal information. Cai et al. [19] disentangled long- and short-term latent states to improve adaptation under unknown interventions; however, these states are learned implicitly and do not explicitly characterize domain-specific behavioral changes. Zhao and Shen [20] proposed PROCEED to estimate concept drift and proactively translate it into model adaptation, while the drift is characterized mainly from statistical changes between training and test samples. More recently, Chen et al. [21] distinguished macro- and micro-drift and introduced a meta-learning framework to accommodate both long-term and short-term distribution changes. Despite these advances, existing online forecasting methods predominantly characterize non-stationarity from data distributions or latent representations, leaving the behavioral mechanisms underlying EV charging-load evolution largely unexplored.

Despite these advances, a critical gap remains between EV charging behavior analysis and online load forecasting. Existing behavioral studies have demonstrated that EV users exhibit substantial heterogeneity in charging timing, duration, frequency, and energy demand [5], [22]. Such heterogeneity is closely associated with vehicle usage patterns, charging environments, and user preferences [23], [24]. Moreover, recent evidence suggests that charging characteristics are not necessarily stationary, as user engagement and the effects of behavioral features may evolve over time [25]. However, these behavioral insights are typically used for offline user segmentation, charging-pattern interpretation, or static feature construction, rather than being explicitly incorporated into the online forecasting process. Conversely, existing concept-drift-aware forecasting methods can adapt model parameters to distributional changes, but generally represent such changes through prediction errors, latent states, or statistical distribution shifts, without describing *what charging behavior has changed* and *how it deviates from the station's persistent behavioral pattern*. Consequently, there remains a lack of a unified forecasting framework that explicitly distinguishes persistent station-specific charging behavior from recent behavioral evolution and translates their deviation into actionable

information for online probabilistic adaptation. Bridging this gap is essential for maintaining accurate and reliable forecasts under heterogeneous and continuously evolving EV charging behaviors.

To address the above research gaps, our core contribution lies in proposing a behavior-guided online probabilistic forecasting framework for EV charging loads under heterogeneous and evolving charging behaviors. Rather than treating temporal variations solely as statistical distribution shifts, the proposed framework explicitly characterizes persistent station-specific charging patterns and their recent behavioral deviations, and further translates such behavioral changes into actionable information for online adaptation. In this manner, behavior characterization and probabilistic forecasting are integrated within a unified online framework, allowing the forecasting model to continuously adapt to evolving charging patterns while maintaining reliable uncertainty estimation. The specific contributions are summarized as follows:

- This paper develops a dual-timescale behavior modeling method to explicitly characterize heterogeneous and evolving EV charging behaviors. Specifically, a long-term behavior anchor is constructed from historical charging observations to represent persistent station-specific patterns, while a rolling recent behavior state continuously captures short-term behavioral changes. The deviation between these two representations explicitly describes how current charging behavior differs from the station-specific historical baseline. Compared with existing concept-drift-aware online forecasting methods [16], [17], which primarily adapt to distribution shifts through temporal representations or model ensembling, the proposed method characterizes charging-load evolution from a behavioral perspective, thereby providing more informative guidance for station-specific online adaptation.
- Building upon the above dual-timescale behavior representation, this paper proposes a behavior-semantic guided probabilistic forecasting method to translate the identified behavioral deviations into actionable information for online adaptation. Specifically, descriptor-level prompts are constructed from the long-term behavior, recent behavior, and their deviations, which are further encoded by a frozen language model to obtain semantic representations of behavioral evolution. A lightweight adapter maps these representations into behavior-aware residual corrections, while a drift-aware gating mechanism dynamically regulates their contribution and the online adaptation rate according to the behavioral deviation. Compared with existing EV forecasting studies that primarily incorporate behavioral or contextual information as numerical features [5], [13], the proposed method further interprets the identified behavior changes at the semantic level and directly uses them to guide online probabilistic adaptation.
- The proposed framework is extensively evaluated on a real-world dataset containing ten heterogeneous EV charging stations. Compared with the best-performing baselines, it reduces MSE and Pinball loss by 15.3%

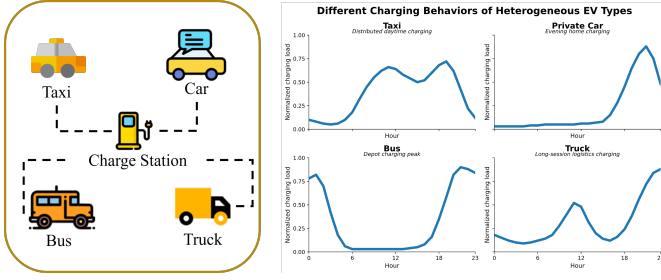


Fig. 1: Heterogeneous charging behaviors across different EV types.

and 17.8% for 1-h-ahead forecasting, and by 16.8% and 22.6% for 4-h-ahead forecasting, respectively, demonstrating consistent improvements across different forecasting horizons.

The remainder of this paper is organized as follows: Section II formulates the online probabilistic forecasting problem for EV charging loads, including the delayed-feedback mechanism and probabilistic forecasting objective. Section III details the proposed behavior-guided online forecasting framework, including dual-timescale behavior modeling, behavior-semantic prompt encoding, and behavior-guided probabilistic adaptation. Section IV presents case studies on ten heterogeneous EV charging stations, with numerical results covering benchmark comparisons, ablation studies, and computational efficiency. Section V concludes the paper and discusses future work.

II. PROBLEM FORMULATION

EV charging loads are inherently driven by users' charging behaviors, resulting in strong heterogeneity and temporal variability across stations and operating periods (see Figure 1). As these behaviors evolve over time, the underlying load patterns may deviate from historical observations, challenging forecasting models trained on fixed data. Based on this motivation, we first formulate the online EV charging load forecasting problem as follows.

Consider a set of EV charging stations $\mathcal{S} = \{1, \dots, S\}$. The hourly charging load of station $s \in \mathcal{S}$ at time t is denoted by $y_{s,t}$. At each forecast origin t , the historical input is given by

$$\mathbf{x}_{s,t} = [y_{s,t-L+1}, \dots, y_{s,t}]^\top \quad (1)$$

where L is the look-back length.

Let $\mathcal{H}_{s,t} = \{y_{s,\tau} : \tau \leq t\}$ denote the load history available for station s at online forecast origin t . Given the available historical information and the calendar information associated with the target time, the H -step-ahead online forecasting problem can be formulated as

$$y_{s,t+H} \sim p(y \mid \mathbf{x}_{s,t}, \mathbf{c}_{s,t+H}, \mathcal{H}_{s,t}; \boldsymbol{\theta}_t) \quad (2)$$

where $\mathbf{x}_{s,t} \subset \mathcal{H}_{s,t}$ contains the most recent L hourly load observations, $\mathbf{c}_{s,t+H}$ denotes the calendar covariates known for the target hour $t+H$, and $\boldsymbol{\theta}_t$ represents the model parameters available at forecast origin t .

Rather than explicitly specifying a parametric form for the conditional distribution $p(\cdot)$, we characterize the predictive

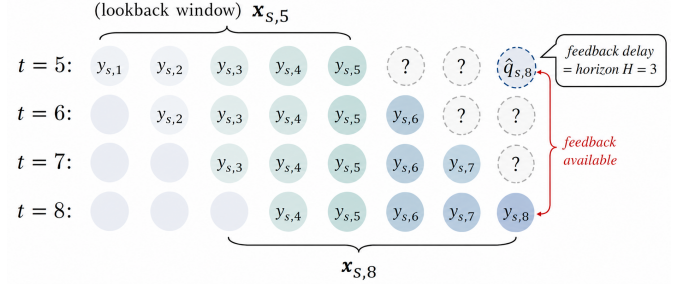


Fig. 2: Delayed-feedback online forecasting with horizon $H = 3$.

distribution through a set of conditional quantiles. Specifically, the model produces

$$\hat{\mathbf{q}}_{s,t+H} = f_{\boldsymbol{\theta}_t}(\mathbf{x}_{s,t}, \mathbf{c}_{s,t+H}) \quad (3)$$

where $\hat{\mathbf{q}}_{s,t+H} = [\hat{q}_{s,t+H}^\alpha]_{\alpha \in \mathcal{Q}}$ denotes the predicted quantiles at the prescribed probability levels $\mathcal{Q}$. Since the corresponding observation $y_{s,t+H}$ is not available when the forecast is issued at time t , the prediction cannot be immediately used for online adaptation. Instead, each forecast is retained until its target observation becomes available.

Accordingly, a prediction generated at forecast origin t is stored as

$$\pi_{s,t}^{(H)} = (s, t, t+H, \mathbf{x}_{s,t}, \mathbf{c}_{s,t+H}, \hat{\mathbf{q}}_{s,t+H}). \quad (4)$$

At clock time u , the matured sample set is defined as $\mathcal{M}_u^{(H)} = \{\pi_{s,t}^{(H)} : t+H = u, s \in \mathcal{S}\}$, where only samples whose corresponding observations are available at time u are eligible for online adaptation. The stored prediction $\hat{\mathbf{q}}_{s,t+H}$ is retained for evaluating the forecast originally issued at time t , while $\mathbf{x}_{s,t}$ and $\mathbf{c}_{s,t+H}$ are used with the newly available observation for subsequent online adaptation.

Once a pending sample matures, the corresponding observation becomes available and its forecasting loss can be evaluated. Specifically, for a sample issued at time t , the gradient is computed at its target time $u = t+H$ using the current model state and the pinball loss as

$$\begin{aligned} \mathbf{g}_{s,t \rightarrow u}^{(H)} &= \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{PB}}(\tilde{y}_{s,u}, f_{\boldsymbol{\theta}}(\mathbf{x}_{s,t}, \mathbf{c}_{s,u}))|_{\boldsymbol{\theta}=\boldsymbol{\theta}_u} \\ \mathcal{L}_{\text{PB}}(\tilde{y}_{s,u}, \hat{\mathbf{q}}_{s,u}) &= \frac{1}{|\mathcal{Q}|} \sum_{\alpha \in \mathcal{Q}} \left[\alpha(\tilde{y}_{s,u} - \hat{q}_{s,u}^\alpha)_+ + (1-\alpha)(\hat{q}_{s,u}^\alpha - \tilde{y}_{s,u})_+ \right] \end{aligned} \quad (5)$$

where $(z)_+ = \max(z, 0)$, $\mathcal{Q}$ denotes the set of prescribed quantile levels, and $\mathbf{g}_{s,t \rightarrow u}^{(H)}$ represents the gradient of the matured sample evaluated using the current model state. The gradient can only be evaluated when the corresponding target observation becomes available at $u = t+H$ (see Figure 2).

Accordingly, once the gradients associated with the newly matured samples are evaluated, they are aggregated to update the current model parameters at clock time u as

$$\boldsymbol{\theta}_{u^+} = \boldsymbol{\theta}_u - \sum_{s \in \mathcal{S}} \eta_{s,u} \sum_{\pi_{s,t}^{(H)} \in \mathcal{M}_u^{(H)}} \mathbf{g}_{s,t \rightarrow u}^{(H)} \quad (6)$$

where $\eta_{s,u} > 0$ denotes the station-specific online learning rate determined by the current behavior drift, and θ_{u+} represents the parameter state after incorporating all newly matured samples at time u .

Through the above delayed-feedback updates, the model sequentially minimizes the probabilistic forecasting loss over the online stream, yielding the overall objective

$$\min_{\{\theta_t\}_{t \geq 1}} \sum_{s \in \mathcal{S}} \sum_{t \in \mathcal{T}_s} \mathcal{L}_{s,t}^{(H)} \quad (7)$$

where $\mathcal{T}_s$ denotes the sequence of online forecast origins for station s and $\mathcal{L}_{s,t}^{(H)}$ is the pinball loss defined above. In addition to minimizing the forecasting loss, the resulting prediction intervals are expected to maintain their nominal coverage over the online stream. For the 90% prediction interval $\hat{\mathcal{I}}_{s,t+H}^{0.90}$, define the miscoverage indicator as $m_{s,t}^{(H)} = \mathbb{I}\{y_{s,t+H} \notin \hat{\mathcal{I}}_{s,t+H}^{0.90}\}$. The desired long-run calibration condition is therefore $\left| \frac{1}{N} \sum_{t \in \mathcal{T}_s(N)} m_{s,t}^{(H)} - 0.10 \right| \rightarrow 0$. Equivalently, the empirical coverage of $\hat{\mathcal{I}}_{s,t+H}^{0.90}$ converges to its nominal level of 0.90.

III. METHOD

To address the above challenges, we propose a behavior-guided online forecasting framework for EV charging loads (see Figure 3). The framework integrates persistent station-specific patterns with recent behavioral changes. Specifically, historical charging observations are used to establish a long-term behavioral representation, while a rolling window continuously captures recent charging dynamics and their deviations from this representation. The resulting changes are semantically encoded to guide drift-aware model adaptation, followed by quantile-based forecasting and online calibration to produce reliable probabilistic predictions under evolving charging behaviors.

A. Dual-Timescale Behavior Modeling

EV charging behavior exhibits different temporal characteristics: persistent patterns characterize the regular charging preference of a station, whereas recent observations reflect short-term behavioral changes. To capture these complementary characteristics, we construct a dual-timescale behavior representation consisting of a long-term behavior anchor and a recent behavior state. Specifically, a behavior operator $\phi(\cdot)$ is introduced to extract charging characteristics from a load window:

$$\phi(\mathcal{W}) = [\mu, \sigma, z, h^{\text{pk}}, r, p^{\text{hi}}, v^{\text{day}}]^\top \quad (8)$$

where μ , σ , z , h^{pk} , r , p^{hi} , and v^{day} denote the mean activity, load volatility, zero-load ratio, peak-hour index, ramping intensity, high-load fraction, and day-to-day variability, respectively.

The long-term behavior anchor of station s is obtained from its initial D_L -day historical observations as $\mathbf{b}_s^L = \phi(\mathcal{H}_s^{D_L})$ and remains fixed during online forecasting. In contrast, the

recent behavior state is continuously extracted from a rolling window of length D_R :

$$\mathbf{b}_{s,t}^R = \phi(\{y_{s,\tau}\}_{\tau=t-D_R+1}^t). \quad (9)$$

Their difference $\Delta \mathbf{b}_{s,t} = \mathbf{b}_{s,t}^R - \mathbf{b}_s^L$ therefore provides an explicit characterization of how the current charging behavior deviates from the station-specific historical baseline, which subsequently guides the online adaptation process.

Meanwhile, the historical load sequence and target-time calendar information are encoded to capture the temporal dynamics underlying charging demand. Specifically, the normalized historical input $\tilde{\mathbf{x}}_{s,t}$ and calendar covariates $\mathbf{c}_{s,t+H}$ are mapped into a temporal representation as

$$\mathbf{h}_{s,t}^{\text{TS}} = \mathcal{F}(\tilde{\mathbf{x}}_{s,t}; \mathbf{c}_{s,t+H}) \quad (10)$$

where $\mathcal{F}$ denotes the temporal encoder.

B. Behavior-Semantic Prompt Encoding

Although the numerical behavior descriptors quantify recent charging changes, they provide limited semantic interpretation of evolving behavior patterns. To leverage the semantic understanding capability of language models, we construct a descriptor-specific prompt for each behavioral characteristic based on its long-term and recent states:

$$\mathcal{P}_{s,t} = \{P_j(b_{s,j}^L, b_{s,t,j}^R, \Delta b_{s,t,j})\}_{j=1}^J \quad (11)$$

where J denotes the number of behavior descriptors and $P_j(\cdot)$ describes the recent state and its change relative to the corresponding long-term behavior (see Figure 4).

Behavior prompt is encoded by a frozen language model $E_{\text{LM}}(\cdot)$ and projected into a compact semantic space:

$$\tilde{\mathbf{z}}_{s,t} = [\mathbf{W} E_{\text{LM}}(\mathcal{P}_{s,t})] \quad (12)$$

where $\mathbf{W}$ is a shared trainable projection matrix. The language model remains frozen throughout forecasting and adaptation and the resulting $\tilde{\mathbf{z}}_{s,t}$ characterizes behavior changes and is subsequently used for behavior-guided adaptation.

C. Behavior-Guided Online Probabilistic Forecasting

The behavior representation, its deviation from the long-term pattern, and the semantic features are integrated to construct a behavior-aware residual correction. Its contribution is further controlled by the degree of behavior drift, allowing the model to adaptively respond to evolving charging patterns. The overall process is formulated as

$$\mathbf{u}_{s,t} = [\mathbf{b}_{s,t}^R; \Delta \mathbf{b}_{s,t}; \tilde{\mathbf{z}}_{s,t}] \quad (13a)$$

$$\mathbf{r}_{s,t} = \mathbf{W}_{A,2} \text{GELU}(\mathbf{W}_{A,1} \mathbf{u}_{s,t}) \quad (13b)$$

$$d_{s,t} = \|\mathbf{b}_{s,t}^R - \mathbf{b}_s^L\|_2 \quad (13c)$$

$$g_{s,t} = \text{sigmoid}\left(\frac{d_{s,t} - \mu_{d,t}}{\sigma_{d,t} + \epsilon}\right) \quad (13d)$$

$$\eta_{s,t} = \eta_0 [\lambda_{\min} + (\lambda_{\max} - \lambda_{\min}) g_{s,t}] \quad (13e)$$

where $\mathbf{u}_{s,t}$ denotes the integrated behavior representation, $\mathbf{W}_{A,1}$ and $\mathbf{W}_{A,2}$ are the trainable adapter matrices, and $\mathbf{r}_{s,t}$ is the resulting behavior-aware residual. The drift magnitude $d_{s,t}$

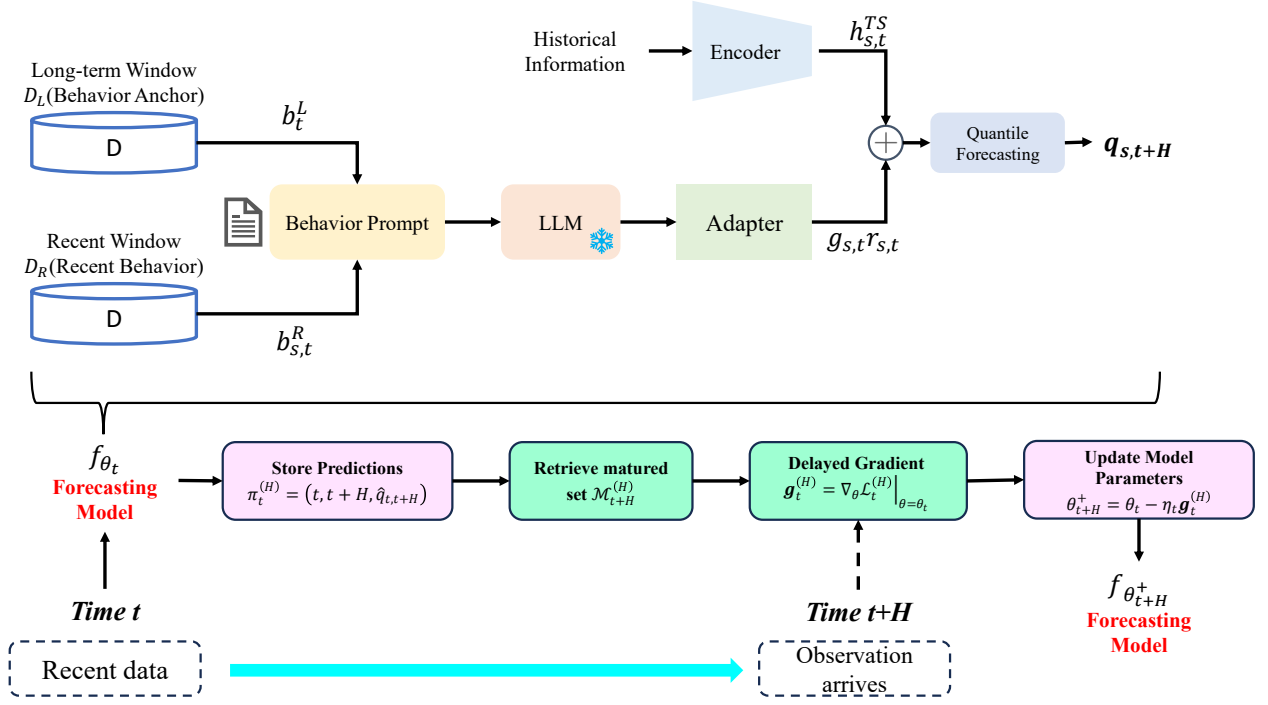


Fig. 3: Framework of the proposed behavior-guided online probabilistic forecasting.

This EV charging station has a long-term operating pattern. Its load volatility is $\{\text{volatility_level}\}$, the zero-load frequency is $\{\text{zero_load_level}\}$, and its dominant charging peak occurs in the $\{\text{period}\}$ near hour $\{\text{peak_hour}\}$. The load ramping intensity is $\{\text{ramping_level}\}$, the high-load frequency is $\{\text{high_load_level}\}$, and day-to-day variability is $\{\text{daily_variability_level}\}$.

This EV charging station exhibits a recent seven-day operating pattern. Its mean charging activity is $\{\text{mean_level}\}$ and is $\{\text{mean_trend}\}$ relative to the long-term pattern. Its load volatility is $\{\text{volatility_level}\}$ and is $\{\text{volatility_trend}\}$, while its zero-load frequency is $\{\text{zero_level}\}$ and is $\{\text{zero_trend}\}$. The charging demand peaks near hour $\{\text{peak_hour}\}$, with a $\{\text{peak_displacement_level}\}$ peak-time displacement from the long-term behavior. Its load ramping intensity is $\{\text{ramping_level}\}$ and is $\{\text{ramping_trend}\}$, whereas its high-load frequency is $\{\text{high_load_level}\}$ and is $\{\text{high_load_trend}\}$. Its day-to-day variability is $\{\text{daily_variability_level}\}$, and the overall behavior concept drift is $\{\text{drift_level}\}$.

Fig. 4: Semantic prompts for long-term and recent charging behaviors.

measures the deviation of recent behavior from the long-term anchor, while $\mu_{d,t}$ and $\sigma_{d,t}$ denote the historical mean and standard deviation of the drift, respectively, and ϵ is a small constant for numerical stability. The resulting $g_{s,t} \in (0, 1)$ represents the drift intensity and jointly controls the residual

contribution and online learning rate $\eta_{s,t}$, where η_0 is the base learning rate and $\lambda_{\min}$ and $\lambda_{\max}$ specify its lower and upper scaling factors.

The behavior-aware residual is then incorporated into the temporal representation:

$$\mathbf{h}_{s,t} = \mathbf{h}_{s,t}^{\text{TS}} + g_{s,t} \mathbf{r}_{s,t} \quad (14)$$

where $\mathbf{h}_{s,t}$ represents the behavior-adapted forecasting representation. In this manner, the temporal encoder provides the primary forecasting representation, while the behavior branch introduces an adaptive correction according to the observed charging-pattern changes.

Finally, the adapted representation $\mathbf{h}_{s,t}$ is mapped to the probabilistic forecasting output through a monotonic quantile decoder:

$$[a_{s,t}, w_{s,t}^-, w_{s,t}^+] = F_Q(\mathbf{h}_{s,t}) \quad (15a)$$

$$\hat{q}_{s,t+H}^{0.50} = a_{s,t} \quad (15b)$$

$$\hat{q}_{s,t+H}^{0.05} = a_{s,t} - \text{softplus}(w_{s,t}^-) \quad (15c)$$

$$\hat{q}_{s,t+H}^{0.95} = a_{s,t} + \text{softplus}(w_{s,t}^+) \quad (15d)$$

where the resulting quantiles satisfy $\hat{q}_{s,t+H}^{0.05} \leq \hat{q}_{s,t+H}^{0.50} \leq \hat{q}_{s,t+H}^{0.95}$ by construction, providing the median forecast and the corresponding 90% prediction interval. The complete algorithm process can be seen in Algorithm 1.

IV. CASE STUDY

A. Setting

In this section, we evaluate the proposed approach using the real-world multi-prototype EV charging dataset introduced

Algorithm 1 Behavior-Guided Online Probabilistic Forecasting

Require: Charging stream $\mathcal{H}_{s,t}$, horizon H , look-back L , windows D_L and D_R , model f_θ

Ensure: Probabilistic forecasts $\hat{\mathbf{q}}_{s,t+H}^L$

- 1: Initialize long-term behavior $\mathbf{b}_s^L = \phi(\mathcal{H}_s^{D_L})$
- 2: **for** each forecast origin t **do**
- 3: Encode temporal representation $\mathbf{h}_{s,t}^{\text{TS}} = F_{\text{TS}}(\tilde{\mathbf{x}}_{s,t}, \mathbf{c}_{s,t+H})$
- 4: Extract recent behavior $\mathbf{b}_{s,t}^R$ and deviation $\Delta \mathbf{b}_{s,t} = \mathbf{b}_{s,t}^R - \mathbf{b}_s^L$
- 5: Encode behavior prompts into semantic feature $\tilde{\mathbf{z}}_{s,t}$
- 6: Compute residual $\mathbf{r}_{s,t}$, drift intensity $g_{s,t}$, and learning rate $\eta_{s,t}$
- 7: Adapt temporal representation $\mathbf{h}_{s,t} = \mathbf{h}_{s,t}^{\text{TS}} + g_{s,t} \mathbf{r}_{s,t}$
- 8: Generate quantile forecast $\hat{\mathbf{q}}_{s,t+H} = F_Q(\mathbf{h}_{s,t})$
- 9: Store pending sample $\pi_{s,t}^{(H)} = (s, t, t + H, \mathbf{x}_{s,t}, \mathbf{c}_{s,t+H}, \hat{\mathbf{q}}_{s,t+H})$
- 10: Retrieve matured set $\mathcal{M}_t^{(H)}$
- 11: **if** $\mathcal{M}_t^{(H)} \neq \emptyset$ **then**
- 12: Evaluate gradients using matured observations $\mathbf{g}_{s,\tau \rightarrow t}^{(H)} = \nabla_{\theta} \mathcal{L}_{\text{PB}}|_{\theta=\theta_t}$
- 13: Update current model parameters $\theta_{t+} = \theta_t - \frac{\sum_{s \in \mathcal{S}} \eta_{s,t} \sum_{\pi_{s,\tau}^{(H)} \in \mathcal{M}_t^{(H)}} \mathbf{g}_{s,\tau \rightarrow t}^{(H)}}{\sum_{s \in \mathcal{S}} \eta_{s,t}}$
- 14: **end if**
- 15: **end for**

in [26], which contains ten heterogeneous charging stations with diverse charging behaviors and operating characteristics. The proposed method is compared against five representative baselines: TCN, PatchTST [27], LSTM, FSNet [16], and OneNet [17] (Details in Appendix). Since this study focuses on online forecasting, the first 60 days of observations are used as the initial historical period, after which the remaining data are processed sequentially in a rolling online manner. At each forecast origin, probabilistic forecasts are generated using only the information available at that time, and model updates are performed only when the corresponding observations become available. Both 1-h-ahead and 4-h-ahead forecasting tasks are considered. Deep learning models are implemented in PyTorch with CUDA support (PyTorch version 2.4.1 on an NVIDIA GeForce RTX 3090 GPU), and all experiments are conducted on a workstation with an Intel Core i7-13900HX CPU and 16 GB RAM. The performance is assessed using five metrics: MSE, Pinball loss, PICP@90, PINAW@90, and CD@90 (Details in Appendix).

B. Dataset

Fig. 5 provides an overview of the heterogeneous and evolving charging behaviors across the ten selected stations. As shown in Fig. 5(a), the stations exhibit distinct long-term behavior compositions in terms of mean activity, volatility, zero-load ratio, ramping intensity, and day-to-day variability. Fig. 5(b) further shows substantial differences in their daily charging profiles, including the timing and magnitude of charging peaks as well as within-day variability. Beyond

these persistent station-specific patterns, Fig. 5(c) reveals that the behavioral states continue to evolve during the online forecasting period. In particular, the seven-day drift trajectories vary considerably across stations and over time, with several stations exhibiting pronounced deviations from their historical behavior. These observations demonstrate that EV charging loads are characterized by both cross-station behavioral heterogeneity and time-varying behavioral drift, motivating the proposed behavior modeling and online adaptation.

C. Result Analysis

Fig. 6 presents the online probabilistic forecasting results across the ten heterogeneous charging stations, showing that the proposed method can effectively track the observed charging dynamics despite substantial differences in load magnitude, peak characteristics, and temporal patterns. The quantitative results in Table I further compare the forecasting performance over the 1-h and 4-h horizons. Overall, TCN, PatchTST, and LSTM exhibit relatively limited performance under the online setting. For the 1-h-ahead task, their MSE values are 0.653, 0.678, and 0.674, respectively, while the corresponding PICP@90 values are only 0.751, 0.812, and 0.751. This indicates that conventional temporal representations combined with online updates remain insufficient to simultaneously accommodate evolving charging patterns and maintain reliable probabilistic coverage. Although PatchTST improves PICP@90 to 0.812, its PINAW@90 increases to 0.357, indicating that the improved coverage is accompanied by relatively wider prediction intervals. In comparison, FSNet and OneNet, which are specifically designed to address non-stationarity and concept drift in online time-series forecasting, achieve substantially better overall performance. Their 1-h Pinball losses decrease to 0.124 and 0.118, respectively, while PICP@90 improves to 0.848 and 0.873, demonstrating the importance of adapting forecasting models to evolving temporal distributions.

The proposed method further improves upon these strong online baselines. For the 1-h-ahead task, it achieves an MSE of 0.360 and a Pinball loss of 0.097, representing reductions of 15.3% and 17.8% over the corresponding best baseline results of FSNet (0.425) and OneNet (0.118), respectively. More importantly, the improved coverage is not achieved by simply widening the prediction intervals. The proposed method obtains a PICP@90 of 0.897, close to the nominal coverage level of 0.900, while reducing PINAW@90 to 0.201, compared with 0.228 for OneNet and 0.240 for FSNet, resulting in a CD@90 of only 0.002. This demonstrates a better balance between probabilistic reliability and interval sharpness. When the forecasting horizon increases from 1 h to 4 h, all baseline methods exhibit varying degrees of performance degradation. In particular, the MSE of TCN increases from 0.653 to 0.984, while its PICP@90 decreases from 0.751 to 0.703; the coverage levels of PatchTST and LSTM also decrease to 0.768 and 0.743, respectively. FSNet and OneNet show greater robustness to the increased forecasting horizon, but their PICP@90 values remain at 0.852 and 0.844. In contrast, the proposed method maintains an MSE of 0.412, a Pinball

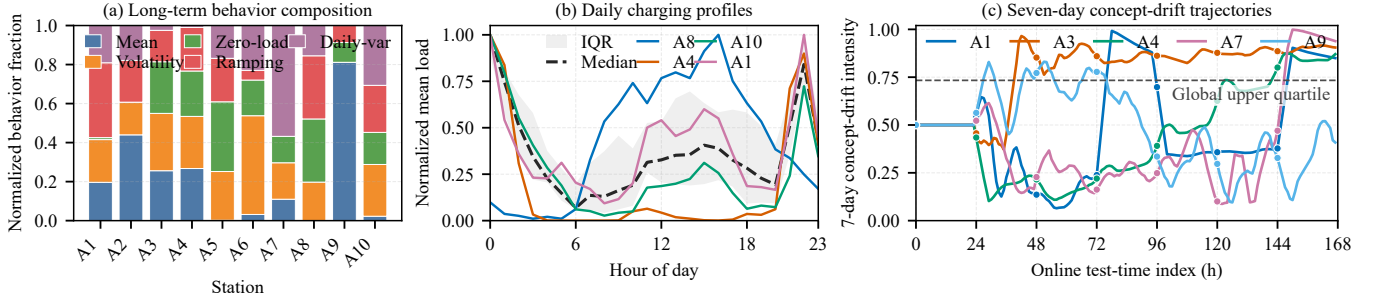


Fig. 5: Heterogeneous and evolving EV charging behaviors across stations.

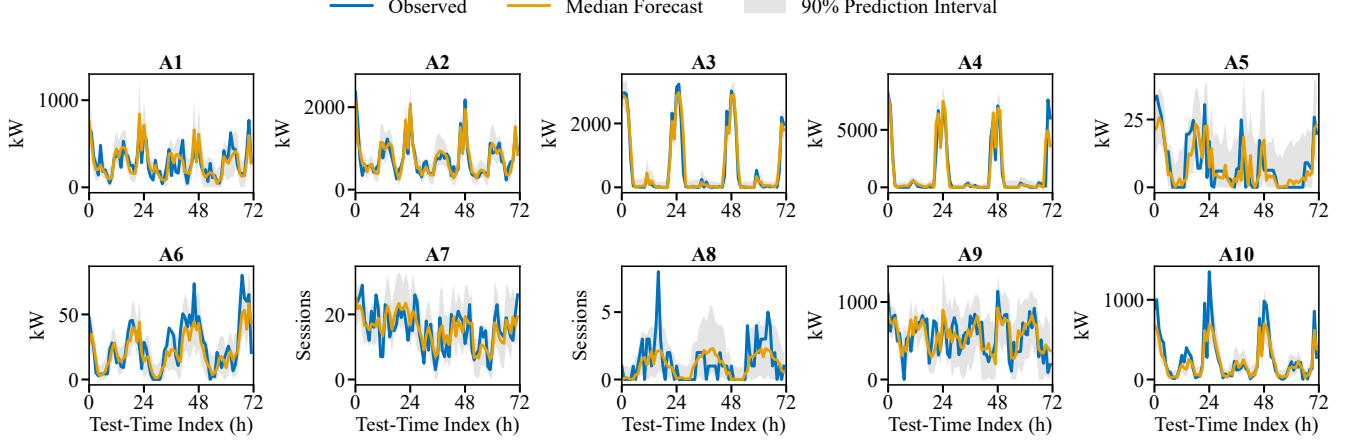


Fig. 6: Online probabilistic forecasts across heterogeneous EV charging stations.

loss of 0.106, a PINAW@90 of 0.223, and a CD@90 of 0.003, while preserving a PICP@90 of 0.897. These results indicate that, compared with conventional online updating and general concept-drift adaptation, explicitly characterizing persistent station-specific behaviors and their recent deviations provides more informative guidance for online adaptation, thereby maintaining forecasting accuracy, interval sharpness, and probabilistic calibration under evolving charging behaviors and extended forecasting horizons.

Figs. 7 and 8 further investigate the contributions of the proposed behavior-guided components. As shown in Fig. 7(a), the full model generally outperforms the TS-only, long-term, and 7-day-prompt variants, confirming the effectiveness of incorporating behavioral information into online probabilistic forecasting. Fig. 7(b) further reveals a clear complementarity between the two behavior components, as most stations benefit from jointly incorporating long-term behavioral patterns and recent behavioral changes. The station-wise results in Fig. 8 show that the magnitude of these improvements varies considerably across heterogeneous charging stations. In particular, some stations benefit more from the long-term representation, whereas others exhibit stronger dependence on recent behavioral information, reflecting differences in their operating regularity and temporal variability. A few stations show limited or negative gains from individual components, indicating that behavioral information is not equally informative under all operating patterns. Nevertheless, the combined framework

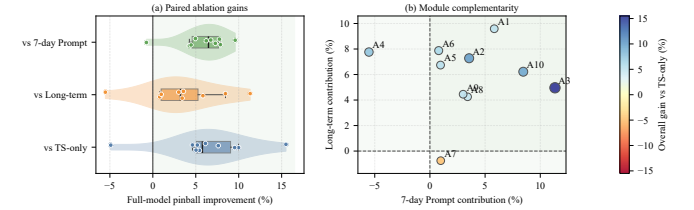


Fig. 7: Ablation analysis of behavior-guided forecasting components.



Fig. 8: Station-wise ablation analysis of behavior-guided components.

provides consistent improvements for the majority of stations.

Table II evaluates the computational efficiency of different forecasting methods. Although the proposed method contains

TABLE I: Performance comparison over 1 h and 4 h forecasting horizons. The best result is highlighted in bold red, while the second-best result is underlined in blue.

Model	1 h ahead					4 h ahead				
	MSE↓	Pinball↓	PICP@90	PINAW@90↓	CD@90↓	MSE↓	Pinball↓	PICP@90	PINAW@90↓	CD@90↓
TCN	0.653	0.173	0.751	0.329	0.149	0.984	0.239	0.703	0.430	0.197
PatchTST	0.678	0.179	0.812	0.357	0.088	0.923	0.212	0.768	0.417	0.132
LSTM	0.674	0.164	0.751	0.362	0.149	0.896	0.225	0.743	0.391	0.157
FSNet	<u>0.425</u>	0.124	0.848	0.240	0.052	0.516	0.142	<u>0.852</u>	0.283	0.048
OneNet	0.453	<u>0.118</u>	<u>0.873</u>	<u>0.228</u>	<u>0.027</u>	<u>0.495</u>	<u>0.137</u>	0.844	<u>0.278</u>	<u>0.056</u>
Proposed	0.360	0.097	0.898	0.201	0.002	0.412	0.106	0.897	0.223	0.003

Note: Lower values are better for MSE, Pinball, PINAW@90, and CD@90. For PICP@90, values closer to the nominal coverage level of 0.900 are better.

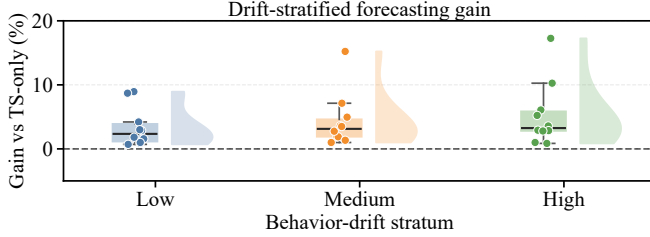


Fig. 9: Forecasting gains across different behavior-drift intensities.

substantially more parameters due to the behavior-semantic module, its inference and online update times remain at 1.85 ms/sample and 8.67 ms/step, respectively, demonstrating practical computational efficiency for online forecasting. Its inference latency is lower than that of PatchTST and remains comparable to other online baselines, while the P95 latency is maintained at 20.62 ms. The memory consumption of 17.4 MB also remains moderate under the experimental platform. These results indicate that the additional behavior-guided modeling introduces manageable computational overhead while preserving efficient online forecasting and adaptation.

TABLE II: Computational efficiency and resource consumption.

Model	Params. (M)	Inference (ms/sample)	Update (ms/step)	P95 (ms)	Memory (MB)
TCN	0.06	0.61	3.48	4.58	18.2
PatchTST	0.60	2.15	9.24	14.79	11.4
LSTM	0.05	0.21	2.97	4.21	1.8
FSNet	0.07	0.84	4.51	6.70	2.1
OneNet	0.13	1.42	6.99	11.30	3.0
Proposed	81.93	1.85	8.67	20.62	17.4

To further validate the effectiveness of behavior-guided adaptation and semantic encoding, additional experiments are conducted. As shown in Fig. 9, forecasting gains generally increase with behavior-drift intensity, indicating that behavioral information becomes more valuable as recent charging patterns deviate from persistent behaviors. The larger gains under high-drift conditions further demonstrate that the proposed adaptation mechanism can effectively respond to pronounced behavioral changes rather than providing uniform corrections. Fig. 10 shows that semantic encoding outperforms direct numerical encoding at most stations, demonstrating its ability to extract additional information from behavioral descriptors. The

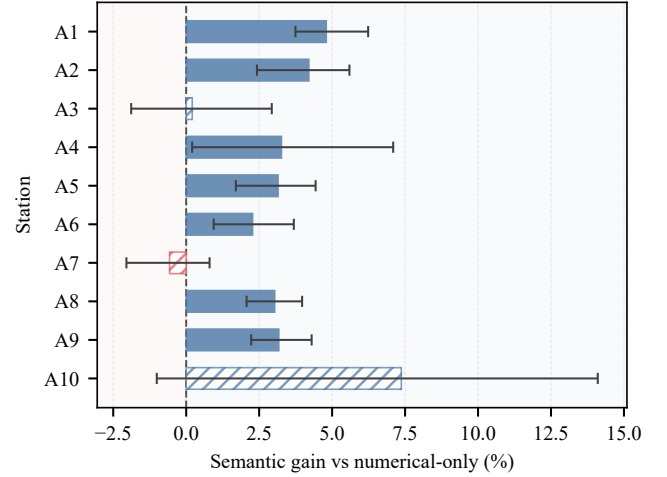


Fig. 10: Station-wise gains of semantic over numerical behavior encoding.

station-dependent gains also suggest that semantic information is particularly beneficial when charging behaviors exhibit complex temporal variations that cannot be fully represented by numerical descriptors alone. Overall, these results confirm that explicitly characterizing behavioral evolution and semantically interpreting such changes provide complementary guidance for online forecasting adaptation.

V. CONCLUSION

This paper proposed a behavior-guided online probabilistic forecasting framework for EV charging loads under heterogeneous and evolving charging behaviors. The framework integrates persistent station-specific patterns with recent behavioral changes through dual-timescale behavior modeling and semantic encoding, enabling behavior-aware adaptation as new observations become available. Experiments on ten heterogeneous charging stations demonstrated that the proposed method consistently outperformed conventional forecasting models and concept-drift-aware online baselines in both 1-h and 4-h forecasting. The results further showed improved probabilistic accuracy, interval sharpness, and calibration, while the ablation analysis confirmed the complementary roles of long-term and recent behavioral information. Despite the additional behavior modeling, the computational overhead remained manageable for online implementation. Overall, the

results highlight the value of explicitly incorporating evolving charging behaviors into online EV load forecasting. Future work will incorporate external behavioral factors and investigate the generalizability of the framework across broader charging environments.

APPENDIX

A. Metric

The forecasting performance is evaluated from the perspectives of point accuracy, probabilistic accuracy, interval coverage, and calibration. Let y_i denote the observed charging load, $\hat{y}_i$ the median forecast, and $\hat{q}_i^\tau$ the predicted quantile at quantile level $\tau \in \mathcal{Q}$, where N denotes the total number of online forecasting samples.

The point forecasting accuracy is evaluated using the MSE:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (16)$$

The probabilistic forecasting accuracy is evaluated using the average pinball loss:

$$\text{Pinball} = \frac{1}{N|\mathcal{Q}|} \sum_{i=1}^N \sum_{\tau \in \mathcal{Q}} \rho_\tau(y_i - \hat{q}_i^\tau) \quad (17)$$

where $\rho_\tau(e) = \max\{\tau e, (\tau - 1)e\}$ denotes the pinball loss at quantile level τ , with a lower value indicating better probabilistic forecasting performance.

For the nominal 90% prediction interval $\hat{I}_i^{0.90} = [\hat{q}_i^{0.05}, \hat{q}_i^{0.95}]$, the prediction interval coverage probability (PICP@90) is defined as

$$\text{PICP@90} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}\{\hat{q}_i^{0.05} \leq y_i \leq \hat{q}_i^{0.95}\} \quad (18)$$

where $\mathbb{I}\{\cdot\}$ denotes the indicator function. A well-calibrated 90% prediction interval should yield a PICP close to 0.90.

The sharpness of the prediction interval is measured by the prediction interval normalized average width (PINAW@90):

$$\text{PINAW@90} = \frac{1}{NR_y} \sum_{i=1}^N (\hat{q}_i^{0.95} - \hat{q}_i^{0.05}) \quad (19)$$

where $R_y = y_{\max} - y_{\min}$ denotes the range of the observed charging load over the evaluation set, with a lower PINAW indicating a sharper prediction interval.

To explicitly quantify the deviation between empirical and nominal coverage, the coverage deviation (CD@90) is defined as

$$\text{CD@90} = |\text{PICP@90} - 0.90|. \quad (20)$$

B. Baselines

TCN: TCN employs temporal convolutional operations to capture local and long-range dependencies in charging load sequences. It is implemented with online updates to provide a convolution-based baseline under the same forecasting setting.

PatchTST: PatchTST divides time-series observations into patches and employs a Transformer architecture to model temporal dependencies. It serves as a representative Transformer-based forecasting baseline.

TABLE III: Training and Online Adaptation Settings

Name	Setting
<i>Offline pretraining</i>	
Initial training period	60 days
Pretraining epochs	30
Batch size	128
Optimizer	AdamW
Initial learning rate	3×10^{-4}
Weight decay	1×10^{-5}
Quantile levels	{0.05, 0.50, 0.95}
Training objective	Pinball loss
Gradient clipping norm	1.0
<i>Online adaptation</i>	
Recent behavior window	7 days (168 h)
Online batch size	1
Online updates per arrived label	1
Online optimizer	AdamW
Base online learning rate	3×10^{-4}
Learning-rate factor range	[0.25, 2.0]
Effective learning-rate range	$[7.5 \times 10^{-5}, 6 \times 10^{-4}]$
Drift-history window	168 samples
Drift warm-up period	24 samples
Initial drift intensity	0.5
Online objective	Pinball loss
Gradient clipping norm	1.0
Label delay	$H \in \{1, 4\}$ h
Frozen semantic encoder	DistilGPT2

LSTM: LSTM captures sequential dependencies through recurrent hidden states and gated memory mechanisms. It is adopted as a representative recurrent baseline for online EV charging load forecasting.

FSNet: FSNet combines fast and slow learning mechanisms to adapt forecasting models to evolving time-series distributions. It is included as a dedicated online forecasting baseline for handling temporal distribution shifts.

OneNet: OneNet employs online ensembling to dynamically combine forecasting models under concept drift. It provides a strong baseline for evaluating adaptation to non-stationary charging load patterns.

C. Setting Details

The detailed training and online adaptation settings are summarized in Table III.

REFERENCES

- [1] M. Shariatzadeh, M. A. R. Lopes, and C. Henggeler Antunes, "Electric vehicle users' charging behavior: A review of influential factors, methods and modeling approaches," *Applied Energy*, vol. 396, p. 126167, 2025.
- [2] S. Wang, A. Chen, P. Wang, and C. Zhuge, "Predicting electric vehicle charging demand using a heterogeneous spatio-temporal graph convolutional network," *Transportation Research Part C: Emerging Technologies*, vol. 153, p. 104205, 2023.
- [3] X. Zhang, K. W. Chan, H. Li, H. Wang, J. Qiu, and G. Wang, "Deep-learning-based probabilistic forecasting of electric vehicle charging load with a novel queuing model," *IEEE Transactions on Cybernetics*, vol. 51, no. 6, pp. 3157–3170, 2021.
- [4] R. Chemudupati, R. Bahmani, G. Fridgen, H. Marxen, and I. Pavić, "Uncertain electric vehicle charging flexibility, its value on spot markets, and the impact of user behaviour," *Applied Energy*, vol. 394, 2025.
- [5] M. Kreft, T. Brudermueller, E. Fleisch, and T. Staake, "Predictability of electric vehicle charging: Explaining extensive user behavior-specific heterogeneity," *Applied Energy*, vol. 370, p. 123544, 2024.
- [6] "Insights into household electric vehicle charging behavior: Analysis and predictive modeling," *Energies*, vol. 17, no. 4, p. 925, 2024.

- [7] “Behavioral uncertainty in ev charging drives heterogeneous grid load variability under climate goals,” *Nature Communications*, vol. 17, p. 43, 2026.
- [8] “Forecasting electric vehicles’ charging behavior at charging stations: A data science-based approach,” *Energies*, vol. 17, no. 14, p. 3396, 2024.
- [9] “Behavioral and infrastructure influences on electric vehicle charging and grid impact,” *Transportation Research Part D: Transport and Environment*, vol. 149, p. 105011, 2025.
- [10] S. Li, H. Xiong, and Y. Chen, “Diffplf: A conditional diffusion model for probabilistic forecasting of ev charging load,” *Electric Power Systems Research*, vol. 234, p. 110723, 2024.
- [11] Y. Wu, P. Cong, and Y. Wang, “Charging load forecasting of electric vehicles based on vmd-ssa-svr,” *IEEE Transactions on Transportation Electrification*, vol. 10, no. 2, pp. 3349–3362, 2024.
- [12] S. Wang, C. Zhuge, C. Shao, P. Wang, X. Yang, and S. Wang, “Short-term electric vehicle charging demand prediction: A deep learning approach,” *Applied Energy*, vol. 340, p. 121032, 2023.
- [13] T. Cao, Y. Xu, G. Liu, S. Tao, W. Tang, and H. Sun, “Feature-enhanced deep learning method for electric vehicle charging demand probabilistic forecasting of charging station,” *Applied Energy*, vol. 371, p. 123751, 2024.
- [14] Y. Dai and B. Shi, “Multi-source spatio-temporal model for station-level electric vehicle charging demand forecasting,” *Pervasive and Mobile Computing*, p. 102251, 2026.
- [15] M. Tu, J. Liu, Z. Tang, T. Su, and P. Zeng, “A short-term ev charging load forecast method based on kolmogorov–arnold spatiotemporal attention recurrent network,” *IEEE Transactions on Transportation Electrification*, 2026.
- [16] Q. Pham, C. Liu, D. Sahoo, and S. C. H. Hoi, “Learning fast and slow for online time series forecasting,” in *International Conference on Learning Representations*, 2023.
- [17] Y.-F. Zhang, Q. Wen, X. Wang, W. Chen, L. Sun, Z. Zhang, L. Wang, R. Jin, and T. Tan, “Onenet: Enhancing time series forecasting models under concept drift by online ensembling,” in *Advances in Neural Information Processing Systems*, vol. 36, 2023.
- [18] Y.-y. A. Lau, Z. Shao, and D.-Y. Yeung, “Fast and slow streams for online time series forecasting without information leakage,” in *International Conference on Learning Representations*, 2025.
- [19] R. Cai, H. Huang, Z. Jiang, Z. Li, C. Zhou, Y. Liu, Y. Liu, and Z. Hao, “Disentangling long-short term state under unknown interventions for online time series forecasting,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 15, 2025, pp. 15 641–15 649.
- [20] L. Zhao and Y. Shen, “Proactive model adaptation against concept drift for online time series forecasting,” in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2025, pp. 2020–2031.
- [21] W. Chen, Z. Zhu, Y. Zhang, L. Shen, L. Yang, Q. Wen, and L. Sun, “Learning to extrapolate and adjust: Two-stage meta-learning for concept drift in online time series forecasting,” in *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 2025, pp. 4869–4877.
- [22] “Identifying electric vehicle charging styles among consumers: A latent class cluster analysis,” *Transportation Research Interdisciplinary Perspectives*, vol. 27, p. 101198, 2024.
- [23] “Revealing charging patterns of electric vehicles with intrinsic usage heterogeneity,” *Transportation Research Part D: Transport and Environment*, vol. 150, p. 105101, 2026.
- [24] “Electric vehicle users’ charging patterns and selection considerations at public charging stations,” *Transportation Research Part D: Transport and Environment*, vol. 149, p. 105081, 2025.
- [25] M. Meža, G. Strle, and M. Meža, “Survival analysis of electric vehicle charging behavior and the temporal evolution of feature effects,” *Scientific Reports*, vol. 15, p. 34897, 2025.
- [26] R. Liu, Y. Li, N. Guo, D. Li, H. Qu, Z. Zhou, Y. Guo, Z. Yan, and J. Liu, “An ai-augmented dataset of multi-prototype electric vehicle charging load profiles in china,” *Scientific Data*, 2026.
- [27] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, “A time series is worth 64 words: Long-term forecasting with transformers,” in *International Conference on Learning Representations*, 2023.