DiffPLF: A Conditional Diffusion Model for Probabilistic Forecasting of EV Charging Load

Siyang Li, Hui Xiong, and Yize Chen
Hong Kong University of Science and Technology (Guangzhou), Guangzhou
sli572@connect.hkust-gz.edu.cn, xionghui@hkust-gz.edu.cn, yizechen@ust.hk

***Abstract*—Due to the vast electric vehicle (EV) penetration to distribution grid, charging load forecasting is essential to promote charging station operation and demand-side management. However, the stochastic charging behaviors and associated exogenous factors render future charging load patterns quite volatile and hard to predict. Accordingly, we devise a novel Diffusion model termed DiffPLF for Probabilistic Load Forecasting of EV charging, which can explicitly approximate the predictive load distribution conditioned on historical data and related covariates. Specifically, we leverage a denoising diffusion model, which can progressively convert the Gaussian prior to real time-series data by learning a reversal of the diffusion process. Besides, we couple such diffusion model with a cross-attention-based conditioning mechanism to execute conditional generation for possible charging demand profiles. We also propose a task-informed fine-tuning technique to better adapt DiffPLF to the probabilistic time-series forecasting task and acquire more accurate and reliable predicted intervals. Finally, we conduct multiple experiments to validate the superiority of DiffPLF to predict complex temporal patterns of erratic charging load and carry out controllable generation based on certain covariate. Results demonstrate that we can attain a notable rise of 39.58% and 49.87% on MAE and CRPS respectively compared to the conventional method.**

***Index Terms*—EV charging, probabilistic forecasting, diffusion models, deep learning.**

I. Introduction

Rapid advancements of energy storage, fast-charging infrastructure, and carbon-reduction blueprints [1] facilitate recent proliferation of electric vehicles (EVs). For instance, there will be 30-42 million EVs supported by 26-35 million charging piles in United States by 2030 [2]. Such substantial EV penetration exerts additional large-scale, highly stochastic load to power networks, imposing unprecedented challenges on grid operation [3]. EV charging load forecasting is crucial to host such vast charging demand, which can benefit operators of both distribution network and charging stations. For instance, grid operators are able to design the optimal coordinated dispatch of EVs and renewables in light of predicted charging power outcomes [4]. Station managers can reduce the electricity procurement cost and optimize the real-time charging scheduling aided by the demand forecast information [5].

Point forecast is a classical way to procure future charging demand. In [6], a stacked recurrent neural network is forged by reinforcement learning to execute demand prediction for different charging scenarios. In [7], charging demand is estimated by the joint statistical analysis for battery behaviors, traffic flow and weather data. However, since the real-world charging demand is extremely stochastic and volatile due to a set of uncertain factors (e.g. battery dynamics, traveller charging patterns and weather conditions), single deterministic charging load forecasts can not give station and power grid operators the most effective tools for load management and operations. Inaccurate prediction can raise operational costs and jeopardize power quality [5]. Besides, it is more crucial for stakeholders to gain a group of reliable forecasts which are beneficial for stochastic optimization and robust decision making [8]. To this end, probabilistic forecasting is a promising approach to model the forecast uncertainties, which can be achieved by generating a host of plausible charging load profiles. Operators can exploit such probabilistic predictions to lessen charging energy deviation costs [5]. For instance, [9] proposes to forecast the uncertainty of EV parking duration.

In this work, we are interested in probing an accurate and reliable probabilistic forecasting model to tackle unknown charging load uncertainties. Forecasting both the values and uncertainties of EV charging scenarios is an emergent topic, while several works focus on predicting probable charging demand in urban areas. Quantile regression like [10] and [11] is a typical way to explicitly construct the prediction interval (PI) by learning a group of quantiles of different degrees. They utilize disparate neural networks to learn the relationships between historical charging data and these quantiles supervised by certain PI evaluation metrics. Nevertheless, such approach falls short in modeling complex temporal EV charging processes which are further complicated by conditional information such as weather and user behaviors. Another feasible approach is to directly quantify the point forecast uncertainties, which stem from both predictive model misspecifications and charging mode variability. For example, hidden state variations in the LSTM-based point predictor are captured by proximal policy optimization in [12]. A novel queuing model is proposed in [13] to link mobile EV charging load to time-varying traffic flow, while a meta-learning method is introduced in [14] to address the data scarcity issue. However, these methods depend on both well-trained deterministic prediction model and assumption of the forecast error distribution.

Essentially, the principal objective of probabilistic charging load forecasting is to obtain the predictive distribution of future charging demand profiles based on observed information. The currently heated generative models [15] are able to approximate the complex distribution of high-dimensional data and generate various samples of high quality, which are promising

to explicitly model the predictive distribution of charging time-series and yield a host of plausible future trajectories. Similar work has been done on the generic time-series analysis, like multivariate time-series forecasting [16], [17] and imputation [18], all of which leverage generative models to estimate the target conditional distribution. However, these work do not explicitly devise a conditioning scheme to entangle the predicted charging load with historical data and relevant covariates.

In this paper, we aim to develop a data-driven generative model to directly learn the joint distribution of future charging load, which can be also conditioned on historical load data and associated covariates including weather forecasts, calendar variables and EV number. These informative covariates are non-negligible to achieve high-quality PIs. More importantly, EV charging station operators and grid operators are also interested in analyzing the impacts of certain variables on EV charging sessions [19]. We adopt the denoising diffusion model proposed in [20], which has demonstrated expressive capacity to produce diverse high-fidelity images and perform controllable generation guided by text prompts [21]. Diffusion models can transform the prior Gaussian noise to target charging load time-series by learning a parameterized reversal of diffusion process. Besides, diffusion models are quite efficient in training and inference, which also evade the mode collapse and training instability issues in other generative model such as generative adversarial networks (GANs) [22].

For the probabilistic forecasting of EV charging load task, we design a specific conditional denoising diffusion model entitled DiffPLF, to generate an array of plausible charging load profiles given historical charging demand and a group of informative covariates. We also incorporate the cross-attention mechanism [21] into the denoising network, which can condition each perturbed load time-series in the diffusion process on input conditional terms. In order to further gain more accurate and reliable probabilistic forecasts, our training objective is to make the 50%-quantile (i.e. median) of produced outcomes be close to ground-truth charging load at each prediction step. In light of it, we propose a fine-tuning block over the pre-trained diffusion model via a 50%-quantile deviation minimization (QDM) loss. Such QDM loss imposes a task-informed inductive bias on the diffusion model, and improves forecasting performance by around 40% in contrast to standard quantile regression. We verify that DiffPLF can output more accurate predictive distribution and point forecast. Besides, it can be adapted to various prediction horizons and execute controllable generation conditioned on different EV numbers. Our code is publicly available at https://github.com/LSY-Cython/DiffPLF for better study of EV grid integration.

II. Problem Formulation

We start from describing the probabilistic charging load prediction task. We aim at capturing the forecast value along with uncertainties induced by stochastic and variable charging behaviors. The key is to explicitly model the conditional distribution of predicted charging load profiles given historical observations and correlated covariates. Specifically, at time

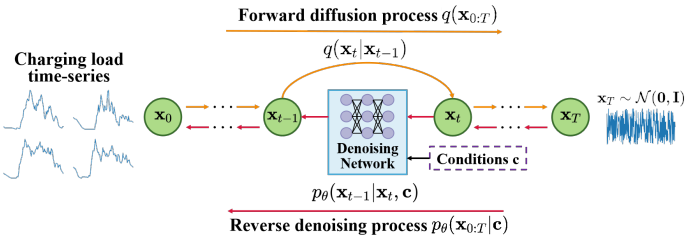

Fig. 1. The diagram of the conditional diffusion model.

s, we collect a series of past charging demand $p_i \in \mathbb{R}, i \in [s-\omega, s]$, where ω is the length of the context window and i is the time step. We also prescribe a covariate set $\mathbf{r}$ which will be described shortly, and utilizing $\mathbf{r}$ is beneficial for accurate forecasts. Note that $\mathbf{r}$ shall be known in advance for the prediction horizon, which can be regarded as external information to constrain unwarranted forecasts. Based on the prior $\mathbf{p} \in \mathbb{R}^\omega$ and $\mathbf{r}$, we want to procure possible outcomes of future charging load $x_0^j \in \mathbb{R},\ j \in [s+1, s+\tau]$, where τ stands for the length of the prediction horizon and we use subscript 0 to align with the original data notation for diffusion models introduced in latter sections. Thereby, our goal is to approximate the conditional predictive distribution $q(\mathbf{x}_0|\mathbf{p}, \mathbf{r})$ of future charging load profiles $\mathbf{x}_0 \in \mathbb{R}^\tau$. The main challenge of this problem is how to derive diverse temporal patterns in $\mathbf{x}_0$ that are consistent with conditions $\mathbf{p}$ and $\mathbf{r}$.

Covariates Selection: Choosing informative covariates is critical for probabilistic time-series forecasting [16]. In this paper, we find three types of covariates are most helpful: 1) *Weather forecasts*. Weather conditions are non-negligible for battery charging dynamics and EV travel behaviors, which can elicit various temporal modes of charging load [23]. We utilize temperature forecast $\mathbf{u} \in \mathbb{R}^\tau$ and humidity forecast $\mathbf{v} \in \mathbb{R}^\tau$, which have been shown to be the most two influential weather factors for EV charging load [7]. 2) *Calendar variables*. EV mobility can be distinct both temporarily and spatially in terms of the day type (e.g. weekdays or weekends) [24]. In our setting, we use a one-hot vector $\mathbf{d} \in \mathbb{R}^7$ to signify seven days within a week. 3) *EV number*. The total number of charged EV e in the forecast window can affect both shape and peak value of predicted charging load time-series. We employ EV number as an unique condition to attest how its variation takes effect on ultimate forecasted profiles. In summary, the covariate set can be denoted as $\mathbf{r} = \{\mathbf{u}, \mathbf{v}, \mathbf{d}, e\}$.

III. Denoising Diffusion Models

In this section, we explicate how the denoising diffusion model derive the complex distribution of charging load time-series, and how to extend it to perform conditional generation.

Due to the eminent high-resolution synthesis capacity and efficient training advantage, denoising diffusion models have been actively applied to various multi-modal content creation fields [21]. The fundamental principle is inspired by the non-equilibrium thermodynamics, which implies that it is feasible to restore the true data distribution by simulating a physically

reversible diffusion process [25]. Firstly, we portray the *forward diffusion process*, where real charging load profiles $\mathbf{x}_0$ are progressively transformed into standard Gaussian noise $\mathbf{x}_T$ after T-step diffusion procedures in total. This forward process is fixed to a Markov chain $q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})$, which suggests step-wise Gaussian noise is gradually imposed on original $\mathbf{x}_0$. The forward transition $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ has the form of Gaussian distribution as follows:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{1}$$

where β_t stands for the magnitude of Gaussian noise added at each diffusion step t, and can be determined by the variance scheduling scheme in [18]. $\mathbf{x}_t$ denotes the perturbed charging load profile by the Gaussian noise with variance β_t. A critical property [20] of this forward process is that we can obtain arbitrary $\mathbf{x}_t$ based on $\mathbf{x}_0$ in closed form:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_0, (1-\alpha_t)\mathbf{I}); \tag{2a}$$

$$\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{2b}$$

where $\alpha_t = \prod_{s=1}^{t}(1-\beta_s)$. This helps significantly accelerate the forward noising procedure and improve the training efficiency of diffusion models.

Once the forward diffusion process is fixed, we can restore the genuine distribution of charging demand time-series $q(\mathbf{x}_0)$ from the standard Gaussian $\mathbf{x}_T$ by learning a *reverse denoising process*. Such reverse process can be defined as a learnable Markov chain $p_\theta(\mathbf{x}_{0:T-1}|\mathbf{x}_T) = \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ parameterized by θ. In light of the physical properties of the invertible diffusion process [20], if β_t is small enough, the diffusion process is continuous and the reverse transition $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ will hold the same function form as the forward transition $q(\mathbf{x}_t|\mathbf{x}_{t-1})$. Hence, we acquire a Gaussian reverse transition:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)). \tag{3}$$

The main target of (3) is to eliminate the noise added at step t from the perturbed $\mathbf{x}_t$. Once the parameters θ in (3) are determined, we can transform the prior Gaussian $\mathbf{x}_T$ into initial charging load data $\mathbf{x}_0$ through this reverse process.

We utilize the maximum likelihood estimation to learn the parameterized reverse transition in (3) and approximate the real charging load distribution $q(\mathbf{x}_0)$. We opt to minimize the negative log-likelihood of real demand data $-\log p_\theta(\mathbf{x}_0)$ via its variational upper bound below:

$$-\log p_\theta(\mathbf{x}_0) \leq \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}]. \tag{4}$$

At each diffusion step t, we use a denoising network to predict the noise added on $\mathbf{x}_0$. By decomposing (4) into $T+1$ closed-form items, we get the following training objective for $\boldsymbol{\epsilon}_\theta$, while we refer the detailed derivations to [20]:

$$\mathcal{L}_{t-1} = \mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon},t}[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}, t)\right\|_2^2]. \tag{5}$$

Until now, we mainly cover generating EV charging load time-series $\mathbf{x}_0$ without any restrictions. In practice, many side features are collected together with charging load profiles, such as weather and past charging demand data, which help us better inform the generated curves. We adopt such features as conditional information, and develop the conditional diffusion model depicted in Fig. 1. A straightforward way to achieve conditional generation is to update the reverse transition in (3) into the conditional normal distribution below [21]:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c}) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, \mathbf{c}, t), \tilde{\beta}_t\mathbf{I}); \tag{6}$$

where $\mathbf{c}$ represents generic condition terms and in our problem setup, $\mathbf{c}$ consists of historical charging load and covariate set. For simplicity, $\boldsymbol{\Sigma}_\theta$ in (3) is fixed to $\tilde{\beta}_t$ and $\tilde{\beta}_t = \frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t$ [20], which alleviates the burden to learn $\boldsymbol{\Sigma}_\theta$ separately. Then, we can naturally condition the training objective in (5) on $\mathbf{c}$:

$$\mathcal{L}_{t-1} = \mathbb{E}_{\mathbf{x}_0,\mathbf{c},\boldsymbol{\epsilon},t}[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}, \mathbf{c}, t)\right\|_2^2]. \tag{7}$$

Once we train a conditional denoising network $\boldsymbol{\epsilon}_\theta$, we can sample $\mathbf{x}_{t-1}$ via the step-wise denoising operation below:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{1-\beta_t}}(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}, t)) + \sqrt{\tilde{\beta}_t}\mathbf{z}, \tag{8}$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This is actually a stochastic sampling procedure, because at each step t in the reverse process, we also need to sample Gaussian noise $\mathbf{z}$.

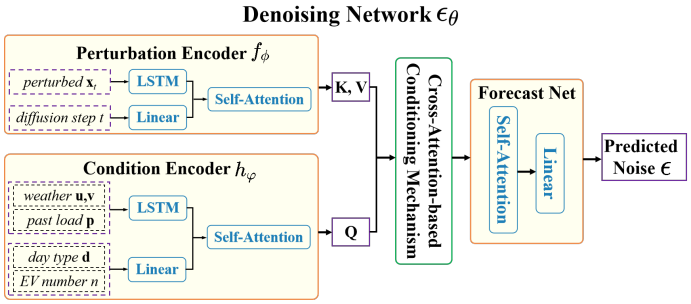

Fig. 2. The architecture of the proposed denoising network dedicated for the conditional diffusion model.

IV. Probabilistic Forecasting Framework

In this section, we detail how to construct DiffPLF by appling the diffusion model for probabilistic charging load forecasts, which includes a denoising network with cross-attention conditioning mechanism and task-specific fine-tuning. Holistic implementation of DiffPLF is depicted in Algorithm 1.

A. Denoising Network

To model the target conditional distribution $q(\mathbf{x}_0|\mathbf{p}, \mathbf{r})$, we should train an effective denoising network $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{p}, \mathbf{r}, t)$ by $\mathcal{L}_{t-1}$ in (7), whose input is the perturbed charging load profile $\mathbf{x}_t$, diffusion step t and conditional terms $\mathbf{c} = \{\mathbf{p}, \mathbf{r}\}$, and output is the noise $\boldsymbol{\epsilon}$ added on $\mathbf{x}_t$. How to incorporate the supplemental $\mathbf{c}$ into $\boldsymbol{\epsilon}_\theta$, i.e., to condition $\boldsymbol{\epsilon}_\theta$ on $\mathbf{c}$ is a vital issue. For instance, TimeGrad [16] simply concatenates $\mathbf{x}_t$ and $\mathbf{c}$ into joint input vectors for LSTM units and does not

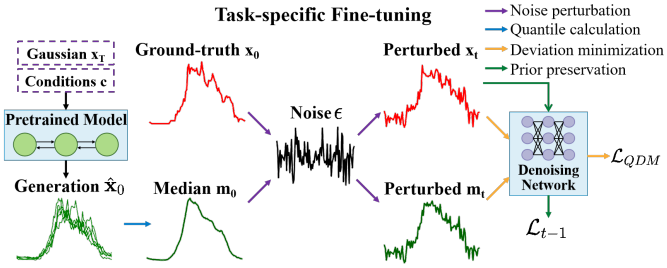

Fig. 3. The illustration of the fine-tuning procedure.

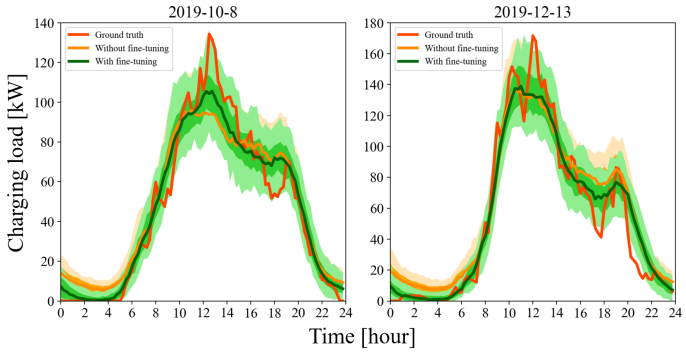

Fig. 4. Forecasting sample comparisons between the fine-tuned model and model without fine-tuning.

explicitly use any conditioning schemes. Whereas in text-to-image generation, several conditioning ways such as classifier-based guidance [15], classifier-free guidance [26] and cross-attention mechanism [21] are proposed to yield proper images which are highly aligned with text semantics. We employ the cross-attention mechanism to entangle $\mathbf{x}_t$ with $\mathbf{c}$. The cross-attention in our method aims to discover latent information of conditions $\mathbf{c}$ that are correlated with predicted profiles $\mathbf{x}_0$.

As shown in Fig. 2, the denoising network $\boldsymbol{\epsilon}_\theta$ contains four components: 1) *Perturbation encoder* f_ϕ, which feeds perturbed time-series $\mathbf{x}_t$ along with t to a LSTM and linear layer respectively, and then utilizes a self-attention module to integrate the latent features of $\mathbf{x}_t$ and t. It actually amounts to the unconditional noise prediction manner defined in (5). 2) *Condition encoder* h_φ, which helps represent the condition set $\mathbf{c} = \{\mathbf{p}, \mathbf{u}, \mathbf{v}, \mathbf{d}, e\}$. We concatenate the temporal data $\{\mathbf{p}, \mathbf{u}, \mathbf{v}\}$, and use LSTM to characterize their time dependencies. The discrete calendar vector $\mathbf{d}$ and EV number e are handled by a linear layer. Then we also employ self-attention to fuse their latent features. 3) *Cross-attention mechanism*, which is to condition the latent encoding of $\mathbf{x}_t$ and t based on conditions $\mathbf{c}$, and this module is conducive to capture the conditional predictive distribution by (7). Denote $\mathbf{Q} = h_\varphi(\mathbf{p}, \mathbf{r}) \cdot \mathbf{W}^Q$, $\mathbf{K} = f_\phi(\mathbf{x}_t, t) \cdot \mathbf{W}^K$, $\mathbf{V} = f_\phi(\mathbf{x}_t, t) \cdot \mathbf{W}^V$, then the cross-attention mechanism is formulated as follows:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V}, \tag{9}$$

where $\mathbf{W}^Q$, $\mathbf{W}^K$, $\mathbf{W}^V$ are three linear transformation matrix weights to be optimized, and d is their hidden dimension. 4) *Forecast network*, which finally uses a self-attention module coupled with a linear projection to transform output features of cross-attention to the predicted noise $\boldsymbol{\epsilon}$ added on $\mathbf{x}_t$.

B. Fine-tuning Technique

Indeed, we can train the former denoising network using (7) and employ it to generate N profiles $\{\hat{\mathbf{x}}_0^n\}_{n=1}^N$ via (8), where $\hat{\mathbf{x}}_0$ denotes the synthetic time-series. But these produced profiles may not form a high-quality PI [16] that is required by probabilistic forecasting. To this end, we expect the 50%-quantile $m_{0,i}$ of N generated charging load $\{\hat{x}_{0,i}^n\}_{n=1}^N$ at each step i should be close to the actual value $x_{0,i}$ as much as possible. In this work, we can treat $\mathbf{m}_0 \in \mathbb{R}^\tau$ as point forecasts. Then our goal is to minimize the 50%-quantile deviation $\|\mathbf{m}_0 - \mathbf{x}_0\|_2^2$. Such term is an inductive bias and a *task-specific refinement*. We leverage it to fine-tune the former model pre-trained by (7), where we explicitly enforce the conditional diffusion model to attain a particular objective. Besides, since $\mathbf{m}_0$ is constituted by the predicted outcomes of the pre-trained diffusion model, this fine-tuning procedure is realized *by its own generated samples*. As shown in Fig. 4, we can obtain sharper PI as well as more accurate deterministic forecasts after the task-informed fine-tuning stage. As $\hat{\mathbf{x}}_0^n$ is generated by a stochastic sampling process which contains multiple times of iterations for $\boldsymbol{\epsilon}_\theta$, it is infeasible to directly use the divergence between $\mathbf{m}_0$ and $\mathbf{x}_0$ to refine the parameters of $\boldsymbol{\epsilon}_\theta$. To this end, we propose to leverage an alternative fine-tuning 50%-quantile deviation minimization (QDM) objective:

$$\mathcal{L}_{QDM} = \|\boldsymbol{\epsilon}_\theta(\mathbf{m}_t, \mathbf{c}, t) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}, t)\|_2^2, \tag{10}$$

where $\mathbf{m}_t$ and $\mathbf{x}_t$ indicate we adpot (2b) to corrupt $\mathbf{m}_0$ and $\mathbf{x}_0$ by the same noise $\boldsymbol{\epsilon}$ at diffusion step t, and c belongs to the ground-truth $\mathbf{x}_0$. Intuitively, minimizing $\|\mathbf{m}_0 - \mathbf{x}_0\|_2^2$ can be equivalent to the goal of $\mathcal{L}_{QDM}$, which is also consistent with the training paradigm of the noise predictor. Moreover, we wish the conditional predictive distribution learned by the pre-trained diffusion model can not be impaired by $\mathcal{L}_{QDM}$, thus we also incorporate the $\boldsymbol{\epsilon}$-prediction loss of (7) in this fine-tuning stage, then the total loss for $\boldsymbol{\epsilon}_\theta$ refinement is:

$$\mathcal{L}_{ref} = \mathcal{L}_{t-1} + \lambda \mathcal{L}_{QDM}, \tag{11}$$

where λ is the weight of QDM loss. The role $\mathcal{L}_{t-1}$ plays in $\mathcal{L}_{ref}$ can be comprehended as a *prior preservation* item, which indicates that when the pre-trained diffusion model is carrying out the task-specific fine-tuning, it is able to retain its prior generative capability simultaneously. Our fine-tuning operation is illustrated in Fig. 3.

V. Numerical Experiments

A. Experimental Setup

1) *Data description*: We harness a real-world dataset in the city of Palo Alto, California termed EV Charging Station Usage Open Data[1]. It elaborates daily charging session details of individual stations, including charging durations and delivered

[1]https://www.kaggle.com/datasets/venkatsairo4899/ev-charging-station-usage-of-california-city

Algorithm 1: Implementation of DiffPLF

Input: training data $\mathcal{X}$, testing data $\mathcal{Y}$, diffusion step T

Stage 1: Pretrain the denoising network $\boldsymbol{\epsilon}_\theta$

1. Sample $t \sim \text{Uniform}(\{1, ..., T\})$
2. Sample $\mathbf{x}_0, \mathbf{c} \sim \mathcal{X}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
3. Optimize θ using $\mathcal{L}_{t-1}$ in (7)

Stage 2: Task-specific fine-tuning

1. Sample $\mathbf{x}_0, \mathbf{c} \sim \mathcal{X}$
2. Generate $\{\hat{\mathbf{x}}_0^n\}_{n=1}^N$ via (8) and pretrained model
3. Calculate median $\mathbf{m}_0$ for generated $\{\hat{\mathbf{x}}_0^n\}_{n=1}^N$
4. Obtain perturbed $\mathbf{x}_t$ and $\mathbf{m}_t$ using (2a)
5. Refine θ using $\mathcal{L}_{ref}$ in (11)

Stage 3: Forecasting via inference

1. Sample $\mathbf{c} \sim \mathcal{Y}$, $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2. Predict $\mathbf{x}_0$ by iterating (8) for T times

energy for each EV. We resort to the transformation method in [7] to aggregate single battery charging curves to total charging load profiles with 15-min resolution. We fetch corresponding weather forecasts for Palo Alto from Meteostat platform[2]. For all experiments involved in this section, we use historical EV charging demand and weather recording from 2016 to 2018 for both training and fine-tuning, while data in 2019 for testing.

2) *Implementation details*: Regarding the historical data utilization, we exploit the aggregate charging demand of past 5 days to guide future load forecasts. In terms of the model architecture, we unify the hidden dimensions of LSTM, cross-attention and self-attention as 32, and the head number of two attention modules is set to 4. As for the noise scheduling, we follow the common quadratic scheme adopted in [18], where the start variance $\beta_1 = 0.0001$, the eventual variance $\beta_T = 0.5$, and the number of diffusion steps T is 200.

We compile the whole DiffPLF using Pytorch library and implement it on a Linux service machine with a 48GB Nvidia A40 GPU. We use Adam optimizer to carry out the stochastic gradient descent with batch size of 16 for the noise predictor. During the pre-training and fine-tuning stage, the initial learning rate and total training epochs are 0.001/0.0002 and 200/100 respectively. Meanwhile, we find that we can achieve the best results when the weight of QDM loss λ is 0.001 for model refinement. For every testing scenario, we randomly generate 1,000 possible trajectories of future charging load to constitute the target PI. After the whole implementation of the proposed model, we find that the average training time of every epoch is 2.0174s and 2.5514s for the pre-training and fine-tuning stage respectively. Such discrepancy results from the fact that the denoising network will be executed only once during each standard diffusion training epoch, whilst being operated twice within each fine-tuning session. Besides, the mean inference time over each test case is 5.7511s, and note that when running DiffPLF on every test sample, we generate 1000 future charging load profiles in parallel.

[2]https://dev.meteostat.net/

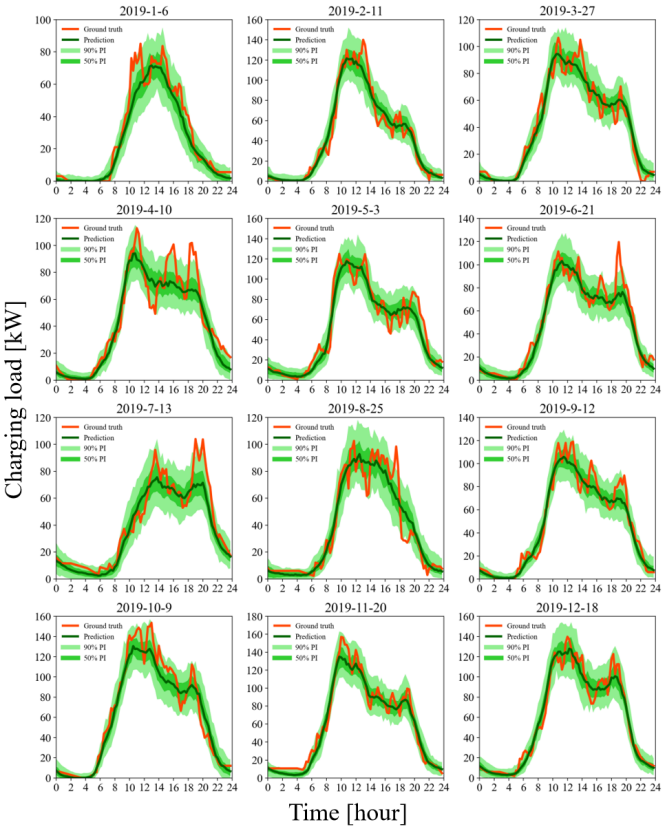

Fig. 5. Randomly selected testing samples. In each subplot, we depict the real charging load versus the generated day-ahead PI and point forecast.

B. Simulation Results

We focus on the day-ahead charging demand forecasting, where we predict charging load in the next day based on historical observations of past five days. We adopt two common metrics to evaluate the holistic performance of the proposed DiffPLF framework: 1) Mean Absolute Error (MAE) [14]. This indicator is aimed at assessing the point forecasting ability of the probabilistic forecasting model. We treat the median of the generated PI as the deterministic prediction. 2) Continuous Ranked Probability Score (CRPS) [16]. This index is utilized to judge the quality of the predictive distribution which is supposed to encompass the true realization. We compute the CRPS value for each prediction step and average that over all time steps as final CRPS for one testing sample.

In Fig. 5, we randomly draw and show 12 samples out of all testing samples in 2019. Evidently, the ground truth profile can not only be covered by shaded areas of either 50% or 90% PI, but also keep close to the produced point prediction to a large margin. It validates that our DiffPLF is able to yield both accurate and reliable probabilistic forecasts. To further verify the superiority of the diffusion-driven generative paradigm to model the predictive distribution, we alter the original training manner (i.e. noise prediction) of DiffPLF to quantile regression [10]. Such method is trained to explicitly predict multiple probabilistic intervals and treated as a classical technique on probabilistic time-series forecasting. Specifically,

TABLE I
QUANTITATIVE EVALUATION FOR DIFFERENT VARIANTS OF DIFFPLF.

Method	MAE	CRPS
Quantile Regression	11.852±3.936	10.107±2.124
w/o covariates	7.842±2.065	5.592±1.570
w/o cross-attention	7.333±1.559	5.192±1.099
w/o fine-tuning	7.227±1.489	5.111±1.041
whole fine-tuning	7.200±1.607	5.089±1.122
DiffPLF	**7.161**±1.557	**5.067**±1.094

TABLE II
DIFFPLF PERFORMANCE UNDER DIFFERENT DIFFUSION STEPS.

Diffusion step T	MAE	CRPS
100	9.127±1.894	6.458±1.362
150	7.313±1.555	5.177±1.106
200	7.227±1.489	5.111±1.041
250	9.569±2.096	6.780±1.492
300	9.503±1.973	6.742±1.415

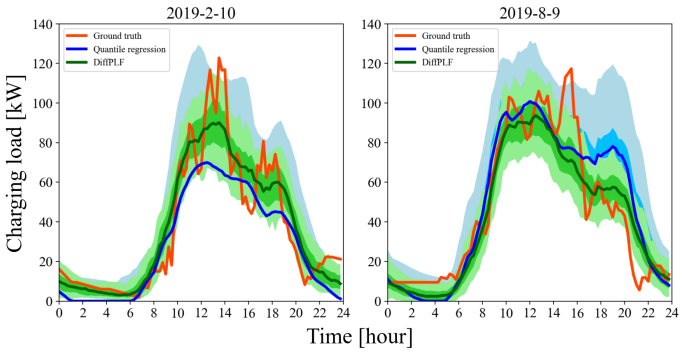

Fig. 6. Two forecasts of our generative DiffPLF versus quantile regression.

we employ the developed noise predictor to directly estimate several quantiles (i.e. 5%, 25%, 75%, 95%) of future charging demand using the pinball loss [27] and exhibit its results in both the first row of Table I and Fig. 6. We can observe that DiffPLF can achieve considerably more accurate and sharper PI than the quantile regression method, because the extreme volatility of EV charging load renders it quite tough for the neural network to stably optimize the quantile loss. In Table I we summarize the method comparisons, where w/o means model without the investigated component.

Cross-attention Mechanism: Previous works also bring up conditioning mechanisms like feature fusion in the latent space [16], [18]. For comparison, we utilize the element-wise addition to blend the latent embedding of both perturbation and condition encoder. From the third line in Table I, a moderate drop more than 0.12 occurs on two metrics, indicating that cross-attention is a more effective way compared to latent fusion when modeling the conditional distribution of temporal charging load data.

Supplementary Covariates: Many previous methods fetch charging demand forecasts merely based on historical observations, ignoring the benefit of certain covariates which can be known for the prediction horizon in advance. Actually, the possible charging load distribution ought to comply with the covariate set, containing weather forecasts, day type and EV number, which are actually a kind of conducive constraints for the predictive model to produce more realistic and accurate forecasts. In order to investigate how DiffPLF benefits from such additional covariates, we purposely discard the input set $\mathbf{r}$ for the condition encoder and write its testing results in the second row of Table I. After the covariate set is eradicated, we observe that there exists a salient decrease of 9.5% and 10.4% on MAE and CRPS respectively. It reflects that the appended covariates described in Section II are crucial for our approach to generate satisfactory probabilistic forecasts.

Task-specific Fine-tuning: In order to render the denoising diffusion method more amenable to the probabilistic time-series prediction task, we propose to fine-tune the pretrained diffusion model using the loss function $\mathcal{L}_{ref}$ defined in (11). $\mathcal{L}_{ref}$ consists of a prior preservation term $\mathcal{L}_{t-1}$ and a weighted item $\mathcal{L}_{QDM}$ to minimize the discrepancy between the median of generated PI and the ground-truth signal. To validate the efficacy of this fine-tuning trick, we compare the performance of entire DiffPLF and its pre-trained version without fine-tuning in Table I. Apparently, DiffPLF outcomes degrade to a mild extent after removing its task-specific refinement, which indicates that the proposed fine-tuning technique is able to improve the effect of diffusion-based generative modeling on probabilistic forecasting for EV charging load.

Furthermore, we look into how different fine-tuning methods affect the final outcomes. In the last two rows of Table I, we show that merely fine-tuning the former part of the noise predictor (while latter weights of cross-attention and parallel encoders are frozen) is better than refining the whole model. The performing gap between such two types of fine-tuning may arise from the implicit adversity of $\mathcal{L}_{QDM}$ for optimizing the latent encoding of heterogeneous input data and conditioning mechanism, which instead prefer to be optimized by the $\boldsymbol{\epsilon}$-prediction manner.

Varying diffusion step T: We investigate the model performance with respect to different settings of diffusion step T, since T is one of the key factors that can determine the ultimate generation outcomes of discrete-time diffusion models [28]. We conduct this sensitivity analysis just in the pre-training context, and forecasting outcomes of default $T = 200$ is shown in the fourth row of Table I, while evaluation results for other T values are exhibited in Table II. Totally, DiffPLF can achieve the best results on $T = 200$ but be relatively less effective on other four T settings. In light of [28] and [25], larger T will give rise to heavy tails in the noise schedule which can deteriorate the learning efficiency of denoising network. Smaller T can increase the discretization errors of continuous stochastic diffusion equations and render the Gaussian form presumption on the reverse transition (3) less valid. How to determine the best T for different forecasting scenarios more properly is of great significance, we leave it for future work.

Varying prediction horizons: Here, we illustrate DiffPLF

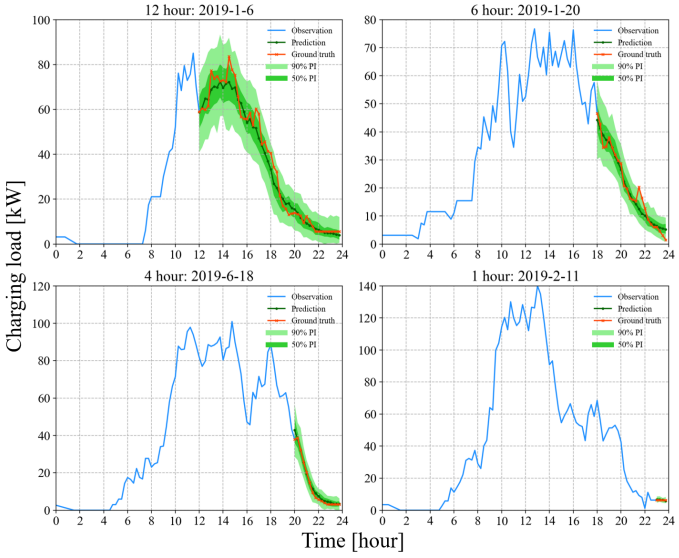

Fig. 7. Examples of produced PI with different prediction horizons.

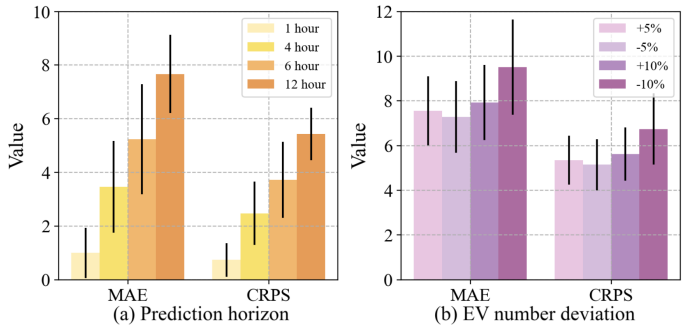

Fig. 8. Mean and variance of MAE and CRPS for two additional analysis: (a) Varying prediction horizons. (b) Different deviations of EV number.

can be seamlessly scaled to different prediction horizons on top of precedent day-ahead forecasting. For the original 24h horizon $[s+1, s+\tau]$, we assume that charging demand in $[s+1, s+\eta], 0 < \eta < \tau$ has been measured and that in $[s+\eta+1, s+\tau]$ should be forecasted. Then, the prediction length is changed from τ to $\tau-\eta$. We can achieve this target by just modifying the input data for $\mathcal{L}_{QDM}$, without overriding the pre-trained diffusion model. Specifically, we simply fix $\mathbf{m}_t^{s+1:s+\eta} = \mathbf{x}_t^{s+1:s+\eta}$, leaving the output demand in $[s+\eta+1, s+\tau]$ to be determined. We consider a set of 12h, 6h, 4h and 1h forecasting horizons with examples shown in Fig. 7, where we can find that DiffPLF can also yield sharp and reliable PI for various prediction horizons. As is shown in Fig. 8 (a), the fine-tuned model holds consistent performance under varying forecasting lengths under both MAE and CRPS metrics.

Varying EV numbers: In practice, operators may be interested to analyze the charging load under different number of EV customers. To this end, we wish to investigate how the variable e affects forecasting outcomes both visually and quantitatively. We fix the already trained DiffPLF and only adjust the input e for the condition encoder. We select a testing sample on June 8th and give five different e, and the resulting profiles under each e are depicted in Fig. 9

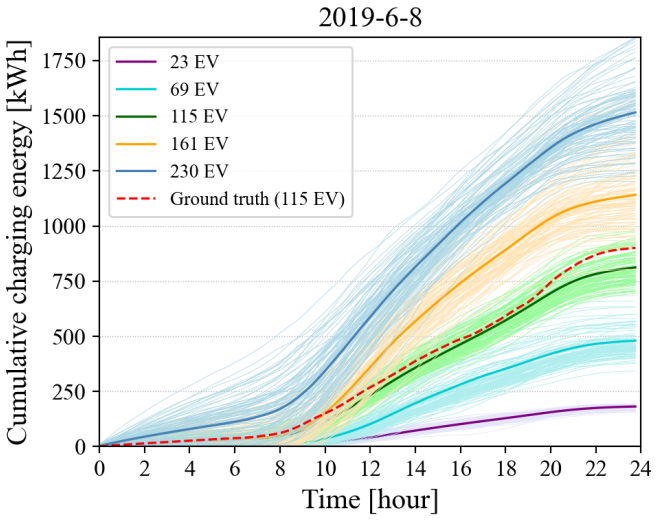

Fig. 9. Predicted daily cumulative charging energy profiles under various EV numbers. The bold line depicts the mean value of generated profiles (plotted by light lines) for each group with a fixed number of EVs.

respectively. Note that we showcase the cumulative charging energy, which are integration of instantaneous charging load. This shows DiffPLF can generate charging energy curves with unique magnitudes and ascending trends. It also suggests that our diffusion model is able to attain controllable generation conditioned on various EV numbers. Moreover, we also want to study the impact of EV quantity deviations on the accuracy of probabilistic forecasts. In Fig. 8(b), we exhibit numeric assessments when e diverges from its ground-truth value by $\pm 5\%$ and $\pm 10\%$. We find that DiffPLF is robust to errors of EV number in total, except for the -10% group, where our model shows a modest drop on prediction results. These results together verify that DiffPLF can be generalizable to EV numbers, forecasting horizons and data samples. In future work, it is also intriguing to benchmark diffusion model's efficacy across different patterns of EV charging datasets.

VI. Conclusion and Outlook

In this paper, we focus on forecasting EV charging load in a probabilistic way by proposing a novel conditional diffusion model DiffPLF. DiffPLF combines the denoising diffusion-driven generation and cross-attention mechanism to capture the predictive distribution conditioned on past demand and complementary covariates. A task-specific fine-tuning approach is devised to further ameliorate the quality of produced prediction intervals. Numerical experiments verify DiffPLF can achieve satisfactory probabilistic demand forecasting and controllable charging profile prediction under flexible look-ahead horizons. Since the current method requires an additional task-informed fine-tuning operation to improve the prediction accuracy, we look forward to developing a more efficient end-to-end diffusion model which can be specialized in probabilistic load forecasting in future work. Besides, we hope to extend our model to predict long-term charging load which is beneficial

for charging infrastructure planning. We also intend to explore multivariate diffusion-based generative model to handle EV charging and more general energy time-series.

REFERENCES

[1] H. Tu, H. Feng, S. Srdic, and S. Lukic, "Extreme fast charging of electric vehicles: A technology overview," *IEEE Trans. Transp. Electrif.*, vol. 5, no. 4, pp. 861–878, 2019.

[2] "The 2030 national charging network," 2023, https://www.nrel.gov/docs/fy23osti/85970.pdf.

[3] Z. Jia, J. Li, X.-P. Zhang, and R. Zhang, "Review on optimization of forecasting and coordination strategies for electric vehicle charging," *J. Mod. Power Syst. Clean Energy*, vol. 11, no. 2, pp. 389–400, 2023.

[4] Z. Wan, H. Li, H. He, and D. Prokhorov, "Model-free real-time ev charging scheduling based on deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5246–5257, 2019.

[5] H. Wang, Y. Jia, M. Shi, C. S. Lai, and K. Li, "A mutually beneficial operation framework for virtual power plants and electric vehicle charging stations," *IEEE Trans. Smart Grid*, pp. 1–1, 2023.

[6] M. Dabbaghjamanesh, A. Moeini, and A. Kavousi-Fard, "Reinforcement learning-based load forecasting of electric vehicle charging station using q-learning technique," *IEEE Trans. Ind. Inform.*, vol. 17, no. 6, 2021.

[7] M. B. Arias and S. Bae, "Electric vehicle charging demand forecasting model based on big data technologies," *Appl. Energy*, vol. 183, 2016.

[8] F. Wu and R. Sioshansi, "A two-stage stochastic optimization model for scheduling electric vehicle charging loads to relieve distribution-system constraints," *Transportation Research Part B: Methodological*, vol. 102, pp. 55–82, 2017.

[9] K. Phipps, K. Schwenk, B. Briegel, R. Mikut, and V. Hagenmeyer, "Customized uncertainty quantification of parking duration predictions for ev smart charging," *IEEE Internet of Things Journal*, 2023.

[10] T. Hu, H. Ma, H. Liu, H. Sun, and K. Liu, "Self-attention-based machine theory of mind for electric vehicle charging demand forecast," *IEEE Trans. Ind. Inform.*, vol. 18, no. 11, pp. 8191–8202, 2022.

[11] L. Buzna, P. De Falco, G. Ferruzzi, S. Khormali, D. Proto, N. Refa, M. Straka, and G. van der Poel, "An ensemble methodology for hierarchical probabilistic electric vehicle load forecasting at regular charging stations," *Appl. Energy*, vol. 283, p. 116337, 2021.

[12] Y. Li, S. He, Y. Li, L. Ge, S. Lou, and Z. Zeng, "Probabilistic charging power forecast of evcs: Reinforcement learning assisted deep learning approach," *IEEE T. Intell. Veh.*, vol. 8, no. 1, pp. 344–357, 2023.

[13] 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 T. Cybern.*, vol. 51, no. 6, 2021.

[14] D. W. X. Huang and B. Boulet, "Metaprobformer for charging load probabilistic forecasting of electric vehicle charging stations," *IEEE Trans. Intell. Transp. Syst.*, pp. 1–11, 2023.

[15] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[16] K. Rasul, C. Seward, I. Schuster, and R. Vollgraf, "Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, 2021, pp. 8857–8868.

[17] K. Rasul, A.-S. Sheikh, I. Schuster, U. Bergmann, and R. Vollgraf, "Multivariate Probabilistic Time Series Forecasting via Conditioned Normalizing Flows," in *International Conference on Learning Representations 2021*, 2021.

[18] Y. Tashiro, J. Song, Y. Song, and S. Ermon, "Csdi: Conditional score-based diffusion models for probabilistic time series imputation," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 24 804–24 816.

[19] J. H. Lee, D. Chakraborty, S. J. Hardman, and G. Tal, "Exploring electric vehicle charging patterns: Mixed usage of charging infrastructure," *Transportation Research Part D: Transport and Environment*, vol. 79, p. 102249, 2020.

[20] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 6840–6851.

[21] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 674–10 685.

[22] R. Yuan, B. Wang, Y. Sun, X. Song, and J. Watada, "Conditional style-based generative adversarial networks for renewable scenario generation," *IEEE Trans. Power Syst.*, vol. 38, no. 2, pp. 1281–1296, 2023.

[23] S. Shahriar, A. R. Al-Ali, A. H. Osman, S. Dhou, and M. Nijim, "Prediction of ev charging behavior using machine learning," *IEEE Access*, vol. 9, pp. 111 576–111 586, 2021.

[24] M. Neaimeh, R. Wardle, A. M. Jenkins, J. Yi, G. Hill, P. F. Lyons, Y. Hübner, P. T. Blythe, and P. C. Taylor, "A probabilistic approach to combining smart meter and electric vehicle charging data to investigate distribution network impacts," *Appl. Energy*, vol. 157, 2015.

[25] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[26] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[27] S. M. J. Jalali, P. Arora, B. Panigrahi, A. Khosravi, S. Nahavandi, G. J. Osório, and J. P. Catalão, "An advanced deep neuroevolution model for probabilistic load forecasting," *Electric Power Systems Research*, vol. 211, p. 108351, 2022.

[28] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8162–8171.