Deep Generative Networks for Synthesizing Data on Electric Vehicle Driving and Charging Events

Zhi Li[1,2,3] Wei Ma[4] Monica Menendez[5] Zhibin Chen*[1,2] Minghui Zhong[6]

[1]Shanghai Frontiers Science Center of Artificial Intelligence and Deep Learning, NYU Shanghai, Shanghai, China
[2]Shanghai Key Laboratory of Urban Design and Urban Science, NYU Shanghai, Shanghai, China
[3]Department of Civil and Urban Engineering, New York University, Brooklyn, United States
[4]Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hongkong
[5]Division of Engineering, New York University Abu Dhabi,, Abu Dhabi, United Arab Emirates
[6]Shanghai Electric Vehicle Public Data Collecting, Monitoring, and Research Center, Shanghai, China

Keywords: Generative Networks, Electric Vehicle, Data Synthesis, Charging Behavior Simulation

1 INTRODUCTION

The electric vehicle (EV) market has seen rapid growth, with global sales hitting around 14 million units by 2023 (IEA, 2024). Equipped with advanced electronics, these vehicles collect crucial real-time data on driving and charging behaviors. The data is essential for formulating transportation policies and deploying charging infrastructure. Furthermore, by engaging in power market activities such as demand response, frequency regulation, and capacity market leasing, EVs contribute to enhancing grid stability and efficiency, as well as facilitating greater penetration of renewable energy sources (Li *et al.*, 2023). These capabilities underscore the critical role of EV data in advancing intelligent transportation systems and promoting the integration of transportation with the power grid.

However, privacy concerns have limited the sharing of EV data among entities such as power utilities, automakers, and vehicle owners (Wang *et al.*, 2023). Vehicle owners, for example, worry about their information being used for targeted marketing or unauthorized tracking (Bilousova, 2024). Given recent advancements in generative AI (i.e., diffusion models (Ho *et al.*, 2020), ChatGPT (Achiam *et al.*, 2023)), a promising approach to addressing these issues is to leverage deep generative models for large-scale EV data synthesis. However, generating high-quality synthetic time-series tabular data remains challenging due to its complex distributions, mixed data types, and strong temporal-feature correlations, which contribute to a persistent gap between synthetic and real data quality (Suh *et al.*, 2024). In this study, we propose training generative models on real-world driving and charging data to generate synthetic EV data that accurately reflects human behavior while preserving privacy.

2 METHODOLOGY

2.1 Problem setting

Let $V = \{v_1, v_2, \ldots, v_n\}$ denote the ensemble of EVs under study. We define a series of events indexed by $t \in \{0, 1, 2, \ldots, T\}$, for segmenting trip-based event. For each vehicle v_i at time t, an event $x_{v_i}^{(t)} = (x_{v_i,1}^{(t)}, x_{v_i,2}^{(t)}, \ldots, x_{v_i,M}^{(t)})$ is constructed, representing M various features. We classify the features into three distinct categories, each serving a unique role within our sequence data analysis:

a. **Static Inherent Features**: Constant attributes such as user labels (i.e., commercial or commuting drivers) and battery capacity that provide essential context for the dataset.

*Corresponding author. Email: zc23@nyu.edu

b. **Dynamic Iterative Features**: Features that update iteratively over time, such as the start state of charge (SoC), which is derived from the end SoC of the previous event, and the start location, which follows from the end location of the previous event. This category also includes the capability to iterate over time-related aspects like weekday and month for added temporal context.

c. **Dynamic Generative Features**: Features generated anew at each time step, reflecting the changing dynamics of the events. This includes event type (i.e., driving and charging), start time, end location, distance traveled, duration, and end SoC, capturing the dynamic nature of each event.

Based on these three categories, we model charging and driving events as sequential data, as illustrated in Table 1.

Table 1 – *Examples of EV event sequences*

Event type	End index	Start time	Distance	Duration	End SoC	Post-event duration	Start index	Start SoC	User label	Battery	Weekday	Month
1	15	10:10	2.1 km	0.1 hr	77%	1 day	16	79%	0	35 kWh	Mon	Sep
1	13	11:20	22.2 km	0.5 hr	50%	0 day	15	77%	0	35 kWh	Mon	Sep
0	13	13:00	0.0 km	2.2 hr	100%	1 day	13	50%	0	35 kWh	Mon	Sep

2.2 Model structure

Given a vehicle v_i and its sequence of feature vectors $s_{v_i} = \{x_{v_i}^{(0)}, x_{v_i}^{(1)}, \ldots, x_{v_i}^{(T)}\}$, we define a joint probability estimator $P_\Phi(s_{v_i})$ parameterized by Φ, expressed as:

$$P_\Phi(s_{v_i}) = P_{\phi_0}(x_{v_i}^{(0)}) \prod_{t=1}^{T} P_\phi(x_{v_i}^{(t)} \mid x_{v_i}^{(t-1)}, \ldots, x_{v_i}^{(0)}) \tag{1}$$

Note that $P_{\phi_0}(x_{v_i}^{(0)})$ represents the initial state of the vehicle. This initial state can be defined using either the original data or user-defined values. To further protect privacy, we use a simple Variational Autoencoder (VAE) model to generate the initial state. Our comparison between the VAE-generated initial state and the original data shows minimal differences between the two. The objective is to find the parameter set Φ that maximizes the likelihood of the observed sequences for all vehicles. Formally, it can be obtained by solving the following optimization problem:

$$\Phi^* = \arg\max_{\Phi} \prod_{v_i \in V} P_\Phi(s_{v_i}) \tag{2}$$

Here, Φ^* denotes the optimal set of parameters for the probability model. We employ a decoder only Transformer (DoT) to develop the sequence model. The model operation is defined as:

$$(b_{v_i}^{(1)}, \ldots, b_{v_i}^{(T)}) = \mathrm{DoT}(x_{v_i}^{(0)}, \ldots, x_{v_i}^{(T-1)}) \tag{3}$$

where $\mathrm{DoT}(\cdot)$ denotes a specialized variant of the Transformer architecture, primarily comprising encoder blocks with Masked MultiHead Self-Attention layer. Notably, the architecture is recognized as the base blocks of GPT-3 structure (Brown, 2020). It processes the sequence of feature vectors $x_{v_i}^{(0)}, \ldots, x_{v_i}^{(T)}$ to capture the dependencies and relationships within the sequence. The output $(b_{v_i}^{(1)}, \ldots, b_{v_i}^{(T)})$ represents the transformed sequence where each $b_{v_i}^{(t)}$ is a feature vector that encapsulates the learned contextual information up to the t-th event. Eq.(1) can then be formulated as:

$$P_\Phi(s_{v_i}) = P_{\phi_0}(x_{v_i}^{(0)}) \prod_{t=1}^{T} P_\phi(x_{v_i}^{(t)} \mid b_{v_i}^{(t)}) \tag{4}$$

Given that a charging/driving event comprises M interrelated discrete or continuous mixed variables, the expression can be further formulated as follows:

$$P_\Phi(s_{v_i}) = P_{\phi_0}(x_{v_i}^{(0)}) \prod_{t=1}^{T} P_\phi(x_{v_i,1}^{(t)}, x_{v_i,2}^{(t)}, \ldots, x_{v_i,M}^{(t)} \mid b_{v_i}^{(t)}) \tag{5}$$

To model dynamic generative features, we utilize a deep Gibbs sampler for its ability to effectively generate joint variable distributions. For a random variable $x_{v_i,j}^{(t)}$ at t-th event., the conditional probability can be modeled as follows:

$$\tilde{x}_{v_i,j}^{(t)} \sim P_{\phi_j}(x_{v_i,j}^{(t)} \mid x_{v_i,-j}^{(t)}, b_{v_i}^{(t)}) \tag{6}$$

where, P_{ϕ_j} represents the conditional probability distribution defined by the model parameters ϕ_j. $\tilde{x}_{v_i,-j}^{(t)}$ denotes all variables except j at t-th event. Discrete variables are modeled using a Softmax layer, while continuous variables are handled through a Gaussian mixture model.

The loss function for training is composed of two parts. The first part is naturally the negative log-likelihood, formulated as:

$$\mathcal{L}_{\text{gibbs},LL} = -\sum_{t=1}^{T}\sum_{j=1}^{M} \log P_{\phi_j}(x_{v_i,j}^{(t)} \mid x_{v_i,-j}^{(t)}, b_{v_i}^{(t)}) \tag{7}$$

The second component involves constraints for continuous, distance, and end state of charge (SoC) variables, enforced through penalties.

2.3 Evaluation metric

We evaluate the model's performance using three metrics:

a) $\boldsymbol{\rho_1}$: Assessing univariate distributions by Jensen-Shannon Divergence for discrete variables and Wasserstein distance for continuous ones.

b) $\boldsymbol{\rho_2}$: Evaluating multivariate distributions by grouping discrete variables (e.g., user label, geographic index, battery capacity) and comparing continuous variables like driving speed (distance/duration) and charging power (SoC difference/duration) within these groups.

c) $\boldsymbol{\rho_3}$: As proposed by van den Burg & Williams (2021), this metric checks if the model has memorized the dataset. A value of $\rho_3 > 1$ indicates memorization, while $\rho_3 < 1$ suggests underfitting. An ideal ρ_3 is close to 1, showing a balance between generalization and fitting.

3 RESULTS

Our dataset contains approximately 7,019,876 driving and charging records from May 2020 to August 2021, involving 3,777 EVs in Shanghai. We compared four models with variations in the joint variable distribution modeling component, specifically using Gibbs sampling, VAE, WGAN, or Bayes decomposition, across three metrics (see Table 2). We found that our proposed Transformer+Gibbs model performs the best overall in terms of univariate (ρ_1) and multivariate (ρ_2) distributions, as well as privacy protection (ρ_3). Although it performs slightly worse than the Transformer+WGAN model on the univariate distribution metric, it significantly outperforms all other models on the multivariate distribution, demonstrating its strength in capturing complex relationships.

Table 2 – *Comparative analysis of different model performances*

Model	Layer	d_{model}	ρ_1	ρ_2	ρ_3
Transformer+Gibbs	48	128	0.18	**0.47**	**0.97**
Transformer+WGAN	48	128	**0.13**	0.81	1.10
Transformer+Bayes	48	128	1.18	0.88	1.55
Transformer+VAE	48	128	0.82	2.04	1.17

We also compare the impact of different Transformer widths (i.e., the input embedding size, d_{model}) and depths (the number of stacked layers) on model performance. The results are shown in Table 3. Increasing model depth by stacking more Transformer layers improves performance more effectively than increasing the model width. The optimal configuration was achieved with 48 layers and $d_{\text{model}} = 128$.

In Figure 1, we present a comparison between the generated model and the original data in terms of the distribution of trip origin-destination (OD) points, showing that both exhibit very similar patterns. Overall, our model effectively generates EV driving and charging sequences with distributions similar to the original dataset while ensuring privacy. This approach also applies to other traffic-related time-series data.

Table 3 – *Comparison of Transformer parameter variations*

Layer	d_{model}	ρ_1	ρ_2	ρ_3	Parameter Size (million)
16	64	0.46	0.66	0.87	1.15
32	64	0.41	0.65	0.90	1.95
48	64	0.21	0.52	0.95	2.75
64	64	0.19	0.50	0.96	3.55
16	128	0.47	0.67	0.82	3.91
32	128	0.23	0.51	0.96	7.08
48	128	**0.18**	**0.47**	**0.97**	10.26
64	128	0.22	0.49	0.95	13.43

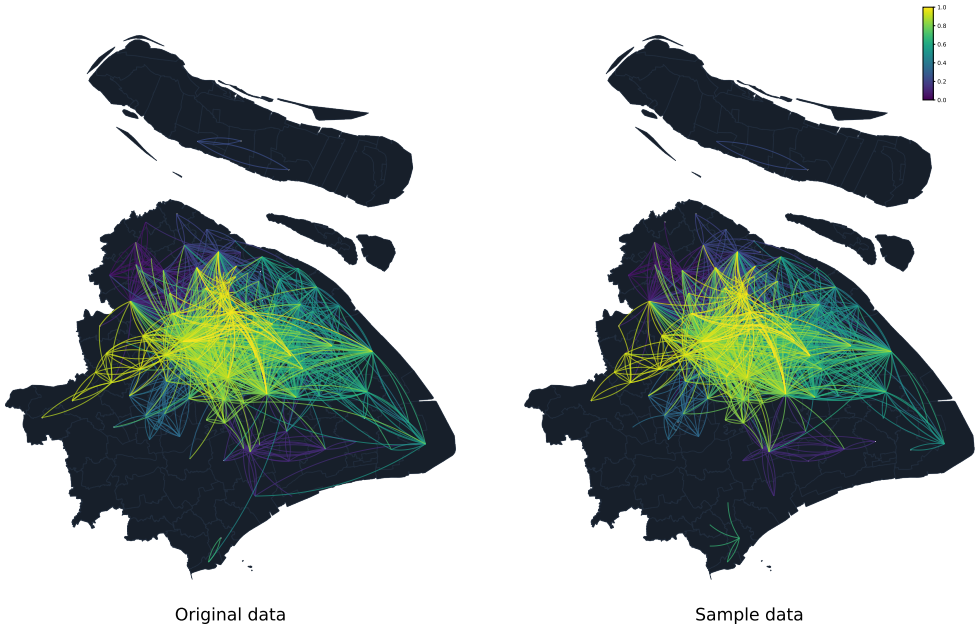

Figure 1 – *Comparison of the OD distribution of driving events*

4 Acknowledgment

This work was partially supported by the National Natural Science Foundation of China [72101153, 72431007], the NYUAD Center for Interacting Urban NEtworks (CITIES) funded by Tamkeen under the NYUAD Research Institute Award CG001, the Shanghai Frontiers Science Center of Artificial Intelligence and Deep Learning at NYU Shanghai, and the NYU Shanghai Doctoral Fellowships.

References

Achiam, Josh, Adler, Steven, Agarwal, Sandhini, Ahmad, Lama, Akkaya, Ilge, Aleman, Florencia Leoni, Almeida, Diogo, Altenschmidt, Janko, Altman, Sam, Anadkat, Shyamal, *et al.* 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bilousova, Mariia. 2024. To Share or Not to Share: Data Exchange Preferences of Vehicle-to-Grid Users.

Brown, Tom B. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Ho, Jonathan, Jain, Ajay, & Abbeel, Pieter. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, **33**, 6840–6851.

IEA. 2024. *Global EV Outlook 2024*. https://www.iea.org/reports/global-ev-outlook-2024. Licence: CC BY 4.0.

Li, Xiaohui, Wang, Zhenpo, Zhang, Lei, Sun, Fengchun, Cui, Dingsong, Hecht, Christopher, Figgener, Jan, & Sauer, Dirk Uwe. 2023. Electric vehicle behavior modeling and applications in vehicle-grid integration: An overview. *Energy*, **268**, 126647.

Suh, Namjoon, Yang, Yuning, Hsieh, Din-Yin, Luan, Qitong, Xu, Shirong, Zhu, Shixiang, & Cheng, Guang. 2024. TimeAutoDiff: Combining Autoencoder and Diffusion model for time series tabular data synthesizing. *arXiv preprint arXiv:2406.16028*.

van den Burg, Gerrit, & Williams, Chris. 2021. On memorization in probabilistic deep generative models. *Advances in Neural Information Processing Systems*, **34**, 27916–27928.

Wang, Jianxiao, Gao, Feng, Zhou, Yangze, Guo, Qinglai, Tan, Chin-Woo, Song, Jie, & Wang, Yi. 2023. Data sharing in energy systems. *Advances in Applied Energy*, **10**, 100132.