UNICO I+D Project
6G-SORUS-RAN

SORUS-RAN-A3.2-E1 (E16)

Initial performance assessment

Abstract

This work addresses the challenge of understanding user mobility in wireless networks, which is crucial for optimizing space usage, managing smart infrastructures, and improving network efficiency. Given the privacy risks associated with real mobility data, we propose DiWi, a Transformer-based model that generates realistic spatiotemporal mobility traces while preserving user privacy. We evaluate the quality and privacy guarantees of the synthetic data and show that it retains the key mobility patterns of real traces. These insights not only enable privacy-preserving mobility analysis but also provide a foundation for developing more efficient resource allocation and reconfiguration strategies in future vRAN systems.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia

Document properties

Document number	SORUS-RAN-A3.2-E1 (E16)
Document title	Initial Performance Assessment
Document responsible	Marco Gramaglia
Document editor	Juan Manuel Montes
Editorial team	Juan Manuel Montes, Marco Gramaglia
Target dissemination level	Public
Status of the document	Final
Version	1.0
Delivery date	31-12-2023
Actual delivery date	31-12-2023

Production properties

Reviewers	Pablo Serrano

Disclaimer

This document has been produced in the context of the 6G-SORUS Project. The research leading to these results has received funding from the Spanish Ministry of Economic Affairs and Digital Transformation and the European Union-NextGenerationEU through the UNICO 5G I+D programme.

Contents

List of Figures

List of Acronyms

5G: Fifth Generation Mobile Networks

5GC: 5G Core

AI: Artificial Intelligence

AMF: Access and Mobility Management Function

API: Application Programming Interface

AP: Access Point

B5G: Beyond 5G

CU: Centralized Unit (in vRAN architecture)

DL: Downlink

DT: Digital Twin

DU: Distributed Unit (in vRAN architecture)

ETSI: European Telecommunications Standards Institute

gNB: Next Generation NodeB (5G Base Station)

HVAC: Heating, Ventilation, and Air Conditioning

IoT: Internet of Things

KPI: Key Performance Indicator

LSTM: Long Short-Term Memory (Neural Network Architecture)

ML: Machine Learning

MLP: Multi-Layer Perceptron

QoS: Quality of Service

RAN: Radio Access Network

RIS: Reconfigurable Intelligent Surface

RU: Radio Unit

vRAN: Virtualized Radio Access Network

WLAN: Wireless Local Area Network

Resumen Ejecutivo

La gestión eficiente de las redes inalámbricas depende en gran medida de la comprensión de los patrones de movilidad de los usuarios. Los datos de movilidad son fundamentales no sólo para optimizar la distribución espacial de los recursos de red y mejorar el funcionamiento de las infraestructuras inteligentes, sino también para mejorar la eficiencia general de la red. Sin embargo, el uso de datos de movilidad reales plantea importantes riesgos para la privacidad, lo que limita su disponibilidad y aplicabilidad en estudios a gran escala y tareas de optimización en el mundo real. Este trabajo aborda este reto introduciendo DiWi, un modelo generativo basado en Transformer diseñado para producir trazas espacio-temporales de movilidad realistas al tiempo que garantiza la privacidad del usuario.

DiWi se entrena con conjuntos de datos de movilidad reales y es capaz de generar trazas sintéticas que preservan los patrones estadísticos y de comportamiento clave del movimiento del usuario sin exponer información sensible. Evaluamos a fondo el modelo comparando las propiedades estructurales y temporales de los datos reales y sintéticos, y valorando su resistencia a ataques a la privacidad como la inferencia de pertenencia y la fuga de información. Nuestros resultados demuestran que DiWi capta con éxito las características esenciales de la movilidad a la vez que ofrece sólidas garantías de privacidad, lo que lo convierte en una valiosa herramienta para el análisis de la movilidad con preservación de la privacidad.

Más allá de su contribución inmediata a la privacidad de los datos de movilidad, los conocimientos obtenidos a través de DiWi tienen implicaciones significativas para el desarrollo de futuras arquitecturas de redes móviles, en particular en el contexto de las redes Beyond 5G (B5G) y las redes de acceso radioeléctrico virtualizadas (vRAN). A medida que las redes B5G avanzan hacia infraestructuras más flexibles y definidas por software, la capacidad de asignar y reconfigurar dinámicamente los recursos de red en respuesta a la movilidad de los usuarios adquiere cada vez más importancia. Sin embargo, para desarrollar y probar estas estrategias avanzadas de gestión de recursos suele ser necesario acceder a datos de movilidad realistas, que suelen estar restringidos por motivos de privacidad.

Al proporcionar una fuente fiable de trazas de movilidad sintéticas pero realistas, DiWi cubre este vacío y permite diseñar y validar algoritmos de asignación adaptativa de recursos en entornos vRAN sin comprometer la privacidad del usuario. Estos conjuntos de datos sintéticos pueden utilizarse para simular diversos escenarios de movilidad, evaluar el rendimiento de las políticas de configuración dinámica y guiar el desarrollo de sistemas de orquestación inteligentes capaces de ajustar los recursos de red a escalas espaciales y temporales finas. De este modo, nuestro trabajo contribuye al esfuerzo más amplio de hacer que las redes B5G sean más eficientes, flexibles y respetuosas con la privacidad.

Executive Summary

The efficient management of wireless networks relies heavily on understanding the mobility patterns of users. Mobility data is fundamental not only for optimizing the spatial distribution of network resources and enhancing the operation of smart infrastructures but also for improving overall network efficiency. However, the use of real mobility data poses significant privacy risks, limiting its availability and applicability in large-scale studies and real-world optimization tasks. This work addresses this challenge by introducing DiWi, a Transformer-based generative model designed to produce realistic spatiotemporal mobility traces while ensuring user privacy.

DiWi is trained on real mobility datasets and is capable of generating synthetic traces that preserve the key statistical and behavioral patterns of user movement without exposing sensitive information. We thoroughly evaluate the model by comparing the structural and temporal properties of real and synthetic data, and by assessing its resilience to privacy attacks such as membership inference and information leakage. Our results demonstrate that DiWi successfully captures essential mobility characteristics while providing strong privacy guarantees, making it a valuable tool for privacy-preserving mobility analysis.

Beyond its immediate contribution to mobility data privacy, the insights obtained through DiWi have significant implications for the development of future mobile network architectures, particularly in the context of Beyond 5G (B5G) networks and virtualized Radio Access Networks (vRAN). As B5G networks move towards more flexible, software-defined infrastructures, the ability to dynamically allocate and reconfigure network resources in response to user mobility becomes increasingly important. However, developing and testing such advanced resource management strategies often require access to realistic mobility data, which is typically restricted due to privacy concerns.

By providing a reliable source of synthetic yet realistic mobility traces, DiWi bridges this gap and enables the design and validation of adaptive resource allocation algorithms in vRAN environments without compromising user privacy. These synthetic datasets can be used to simulate diverse mobility scenarios, assess the performance of dynamic configuration policies, and guide the development of intelligent orchestration systems capable of adjusting network resources at fine spatial and temporal scales. In this way, our work contributes to the broader effort of making B5G networks more efficient, flexible, and privacy-respecting.

1. Introduction

The rising demand for flexible, scalable, and efficient wireless infrastructure is driving a rapid evolution of Radio Access Networks (RAN), culminating in the concept of Virtualized RAN (vRAN). By decoupling traditional hardware-dependent RAN functions and virtualizing them on cloud-based or commodity platforms, vRAN promises improved network flexibility, centralized control, and dynamic resource management. To achieve these benefits, however, network operators must be able to predict network behaviors and user demands accurately, so that scheduling, resource allocation, and Quality of Service (QoS) can be proactively managed.

Building upon our previous analysis of the state of the art in AI-driven resource allocation and energy-aware vRAN (SORUS-RAN-A2.3-E1 (E11)), we identified mobility modeling and demand prediction as key enablers for intelligent, data-driven network management. Specifically, our SOTA study revealed that while many existing approaches focus on aggregated traffic modeling or reactive resource allocation, there is a critical need for fine-grained, user-level mobility insights to inform proactive decision-making in virtualized environments. Accurate modeling of device movements and demand fluctuations is fundamental to feeding the AI-based orchestration mechanisms reviewed earlier, enabling predictive control strategies and improving the efficiency of vRAN deployments.

A key enabler of such prediction capabilities is Machine Learning (ML), which leverages historical and real-time network data to forecast user-level behavior and system-wide performance. This deliverable applies ML-based methods for time-series prediction in a wireless environment, focusing specifically on user-level trends and resource utilization. While the overarching objective is to demonstrate how these methods can be adopted and scaled in a vRAN deployment, the experiments draw upon an extensive dataset gathered from a campus-wide Wireless LAN (WLAN). Although WLANs and cellular-based vRANs operate under different protocols and frequency bands, both are predicated on radio access principles. Wireless connectivity within a campus environment can serve as a representative testbed, capturing essential similarities such as fluctuating user density, time-varying traffic patterns, mobility of client devices, and multi-AP (or multi-cell) interactions.

Wireless networks provide seamless connectivity by eliminating the need for physical connections, enabling broad access to digital services. As devices move through the network, they generate connectivity patterns that represent traces of their activity [Trivedi2020]. These traces provide insights into aspects such as high-traffic areas, peak activity times, and typical usage patterns [Zhang2014]. These insights into device behavior help operators allocate resources more efficiently, reduce energy use with adaptive management strategies [Planchy2016], and improve systems like heating optimization [Pan2014].

However, the pervasive use of wireless networks comes with significant privacy concerns. Effective network management, optimization, and security often require extensive monitoring, which involves collecting sensitive data about users' mobility, device activity, and personal habits. If unauthorized entities access this data, it could be exploited in a privacy-invasive manner, highlighting the need for robust privacy-preserving techniques in data analysis and synthetic trace generation [YData2023].

To address these concerns, it is essential to implement robust techniques that secure and protect sensitive mobility data even when encryption is breached, or data is leaked. Methods such as

pseudonymization are known to only mildly solve this issue as demonstrated in [Montjoye2013], where 93 % of the users were re-identified even when this technique is applied.

In contrast to pseudonymization, synthetic data offers a practical way to protect user privacy while preserving the statistical properties needed for analysis [Surendra2017]. By generating datasets that mimic real data without directly linking to individual users, it reduces the risk of re-identification. This allows researchers and organizations to study patterns, train machine learning models, and develop solutions without exposing sensitive information.

One clear advantage of synthetic data is that it does not have a one-to-one correspondence with real devices (unlike some pseudonymization techniques), but its utility may be questioned if it fails to capture real-life dynamics.

In this work, we propose DiWi, a transformer-based Digital Twin for Wireless mobility. DiWi generates synthetic data that captures the features of real-life device mobility in wireless networks, supporting the development of relevant use cases without leaking personal data. By situating this contribution within the framework established in our prior SOTA analysis, DiWi directly addresses the need for privacy-preserving, fine-grained mobility models to support AI-driven resource allocation and predictive vRAN orchestration.

More specifically, the key contributions of this deliverable are:

1. Transformer-based mobility modeling: We design and implement a Transformer-based model to predict device connectivity sequences in a campus WLAN. By decomposing temporal information during encoding, the model reduces the number of parameters while improving generalization, enabling it to generate realistic spatiotemporal traces. The model is trained on real traces collected from a campus environment with millions of devices, users, and access points, ensuring accurate and meaningful results.
2. Applications in network and space management: We explore the model's applicability in key use cases such as optimizing space usage, managing intelligent buildings, predicting user occupancy for energy-efficient HVAC systems, and forecasting mobility for network management. Additionally, we demonstrate its utility in developing and evaluating privacy metrics for wireless networks.
3. Privacy-preserving data generation: To ensure the generated data does not compromise user privacy, we assess risks such as direct leakage, similarity searches, and membership inference attacks [YData2023]. Our analysis confirms that the model effectively captures mobility patterns while preventing the exposure of identifiable user traces.

The rest of the Deliverable is organized as follows: In Section 2, we describe DiWi, providing an overview of the system, the dataset used, the sequence encoding, and the modeling and training stages. We evaluate DiWi in Section 3, both by analyzing device and Access Points (AP) statistics, and by exploring its applicability in key use cases related to sustainability, mobility management, and understanding anonymity schemes. We assess DiWi's privacy features in Section 4 by examining exact matches, dataset closeness, and membership inference risks. Finally, we conclude the Deliverable in Section 5.

2. DiWi: Digital Twin for wireless Mobility

Here we describe DiWi, a Transformer-based model designed for predicting the next location and generating synthetic mobility traces. We detail its key components and their roles in processing device connectivity sequences. Specifically, we explain how the model encodes spatial and temporal information, integrates these representations, and leverages a Transformer framework to predict the next connectivity state.

2.1. Dataset

We rely on a dataset of users connected to the WiFi network at in one Campus of Universidad Carlos III de Madrid (UC3M) The network is provided by 279 APs distributed across seven buildings. These buildings, most with at least three floors, serve various purposes such as classrooms, cafeterias, and study areas. Each AP provides coverage for approximately 100 m^2, while the entire campus spans 1.6 million m^2 and supports a community of around 10,000 individuals. The dataset spans four weeks of connectivity logs, with no holidays or special events in between.

We are provided the data by system administrators, which rely on a system that logs device connections and discretizes them every 5-minute intervals, according to the following rules:

- If a device remained connected to a single AP during the interval, that AP was recorded.
- If the device connected to multiple APs, the AP where it spent most of the time was assigned.
- If no connection was detected, the device was considered disconnected, and we introduce an artificial OUT status to represent periods without connectivity.

Furthermore, to following ethical and privacy guidelines, user identities and device MAC addresses are pseudonymized using an MD5 hash.

2.2. Model Overview

The architecture of DiWi is illustrated in Figure 1 which summarizes the entire workflow (represented from top to bottom): the orange box shows the decomposed encoding and creation of the spatiotemporal trace, the red box represents the transformer blocks and their computations, and the green box indicates the prediction of the next AP and the inference process.

Starting from the dataset, the system sequentially encodes spatial and temporal components of device connectivity traces (top part of the figure). These components are then merged into a unified spatiotemporal representation, capturing the full behavior of each device. Finally, the model predicts the next connectivity state: whether the device will remain connected to the same AP, transition to a different AP, or disconnect entirely (OUT). The training process is designed to optimize these predictions, ensuring accurate modeling of mobility behavior across the network.

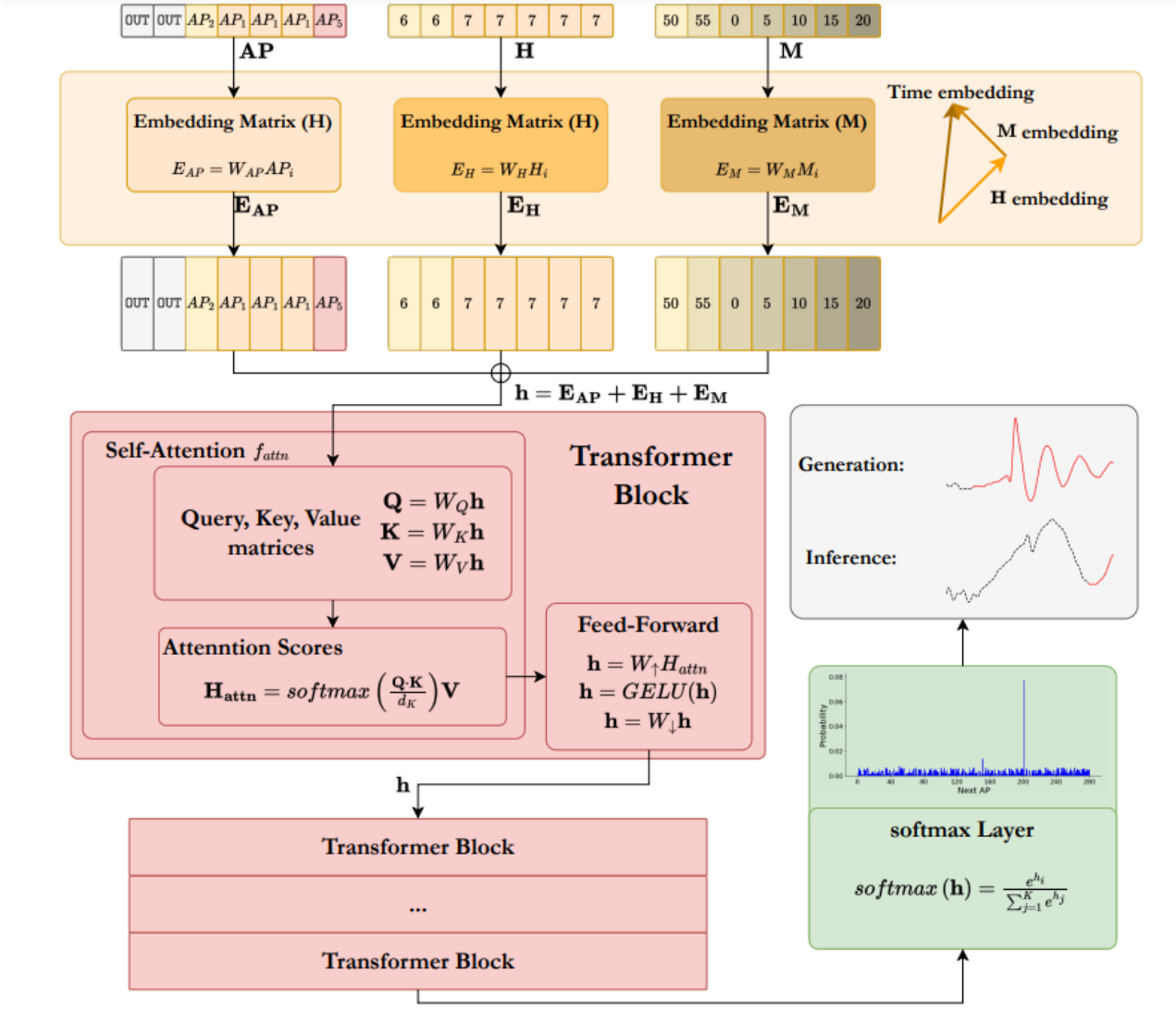

FIGURE 1: ARCHITECTURE OF DIWI

2.3. Sequence Encoder

We restrict ourselves to the interval between 6:00 AM and 10:00 PM. Given the 5-minute discretization, a trace consists of 192 tokens (i.e., 12 per hour). The model's vocabulary ($\mathcal{V}$) includes the 279 APs plus the OUT token, and therefore it captures all possible device locations, including disconnections.

To model device connectivity, we encode both spatial and temporal information into a unified vector representation. This combined encoding allows the model to capture the full context of a device's mobility, i.e., its location within the network and the timestamp.

For **spatial encoding**, we use an AP embedding matrix (W_{AP}), which maps one-hot vectors representing possible connectivity states—either a specific AP or the OUT state for disconnections—into high-dimensional vectors of size d. This process generates spatial embeddings ($\boldsymbol{E_{AP}}$), as shown in Eq 1, which captures the spatial properties of each connection:

EQ 1

$$\boldsymbol{E_{AP}} = W_{AP} \times \boldsymbol{AP_i}$$

For the **temporal encoding,** we use absolute positional embeddings to represent the exact timestamp of each connection. Unlike relative positional embeddings commonly used in natural language processing (e.g., to preserve word order in text), absolute embeddings are better suited for mobility data, since the absolute position of a token within the timeline provides critical contextual information. For example, being disconnected at 7:00 AM (typically before a user arrives) conveys different behavior than being disconnected at 2:00 PM (a common lunch hour in Spain). To encode absolute time, we decompose it into two components: hours **H** and minutes **M**. Each component is mapped into high-dimensional space using its own embedding matrix as follows:

The hour embedding $\boldsymbol{E_H}$ is computed using W_H, an embedding matrix of shape ($N_H \times d$), where N_H represents the number of possible hour values:

EQ 2

$$\boldsymbol{E_H} = W_H \times \boldsymbol{H_i}$$

• The minute embedding $\boldsymbol{E_M}$ is computed using W_M, an embedding matrix of shape ($N_M \times d$), where N_M represents the number of possible minute values:

EQ 3

$$\boldsymbol{E_M} = W_M \times \boldsymbol{M_i}$$

The final temporal embedding at each time step is obtained by adding both embeddings ($\boldsymbol{E_H} + \boldsymbol{E_M}$)**.** This segmented approach to temporal encoding is more efficient than using a single absolute positional encoding matrix of size ($L \times d$), where $L = 192$ represents the total number of time steps in a trace. By breaking time into smaller components (hours and minutes), we significantly reduce the number of parameters required, which improves memory efficiency while preserving the semantic meaning of absolute time, making the model both scalable and practical for processing large mobility traces.

Together, the spatial embedding $\boldsymbol{E_{AP}}$ and temporal embeddings ($\boldsymbol{E_H} + \boldsymbol{E_M}$) form the joint spatiotemporal h representation of device traces.

2.4. Sequence Modeling

After the joint spatiotemporal vector $\boldsymbol{h}$ is composed, it is processed through a series of Transformer blocks to capture dependencies and relationships between connectivity states. These blocks refine the representation of spatiotemporal patterns within the sequences, as described below.

The attention mechanism within the Transformer blocks computes attention scores by projecting the input embeddings $\boldsymbol{h}$ into the following vectors: query ($\boldsymbol{Q}$), key ($\boldsymbol{K}$), and value ($\boldsymbol{V}$). A scaled dot-product operation between $\boldsymbol{Q}$ and $\boldsymbol{K}$ captures relationships between elements in the sequence. These scores are then normalized using a $softmax$ function and applied to $\boldsymbol{V}$, producing the attention output $\boldsymbol{H_{attn}}$. This process is summarized by Eq. (4)

The attention output $\boldsymbol{H_{attn}}$ is computed as:

EQ 4

$$\boldsymbol{H_{attn}} = f_{attn}(\boldsymbol{E_{AP}} + \boldsymbol{E_H} + \boldsymbol{E_M})$$

where $\boldsymbol{f_{attn}}$ represents the self-attention mechanism applied to the combined embeddings of access points, hours, and minutes.

The output from the attention mechanism, $\boldsymbol{H_{attn}}$, is then passed through a feedforward network to further transform the representation. This network consists of two linear layers:

- The first layer, parameterized by $W_{\uparrow}$, expands the dimensionality of the representation.
- The second layer, parameterized by $W_{\downarrow}$, reduces it back to its original size.

A GELU activation function is applied between these layers to introduce non-linearity, enabling the model to learn complex transformations. The resulting output of this feedforward network is given by:

EQ 5

$$\boldsymbol{h} = W_{\downarrow}\, GELU(W_{\uparrow} \times \boldsymbol{H_{attn}})$$

Finally, the output from the Transformer blocks, **h**, is passed through a softmax layer to generate a probability distribution over the model's vocabulary **V**. This probability distribution represents the likelihood of each possible next connection state (e.g., AP or disconnection). The prediction process is formalized as:

EQ 6

$$\boldsymbol{P_{next}} = softmax(\boldsymbol{h})$$

Using this probability distribution, synthetic datasets are generated sequentially. At each step, the model predicts the next AP by sampling from the probabilities calculated based on the sequence of tokens observed so far. This approach assumes conditional independence, meaning that each prediction depends solely on the context provided by the preceding sequence.

2.5. Training

The model is trained in a supervised manner using real device connectivity traces, where the input consists of a sequence of AP connections and disconnections within a fixed **context length** (256 time steps). The expected output is the same sequence shifted by one step, allowing the model to learn temporal dependencies and predict the next connection state.

The hyperparameters were chosen following the principles outlined in [Kaplan2020], which analyze the relationship between model performance and hyperparameter choices. These parameters include:

- **Embedding dimension**: **d = 384**
- **Number of transformer layers**: **6**
- **Attention heads**: **6**
- **Dropout rate**: **0.15**, to balance performance and generalization

The model is trained using a cross-entropy loss function to compare predictions with actual states and optimized with the **Adam optimizer** ($\beta_1 = 0.9, \beta_2 = 0.999$) with a **learning rate** of 2×10^{-5}. Training runs for up to **50 epochs**, with early stopping if validation loss stagnates for **5 consecutive epochs** to prevent overfitting. A **batch size of 64** ensures stable and efficient convergence.

3. Performance Evaluation

In this section, we evaluate DiWi's ability to generate realistic data by comparing the statistical relationships in the synthetic dataset to those in the real dataset. This analysis ensures that the model not only produces plausible individual samples but also captures the broader patterns and dependencies present in the original data. Additionally, we highlight potential applications of DiWi, particularly in sustainability and user behavior analysis within wireless networks.

We first evaluate DiWi's ability to predict the next connection state by measuring its accuracy in identifying the next AP connection or disconnection. Specifically, we compute the prediction accuracy across a test set of real connectivity traces, using the ground truth next states as a reference. To provide context, we compare the model's performance against two benchmarks: an LSTM network, and the baseline GPT-2 model without our proposed temporal encoding. Aside from the Markov model, all models have a comparable number of parameters, ensuring a fair comparison. TablE 1 presents the accuracy of different models based on these key evaluation metrics.

Model	Accuracy [%]
LSTM Network	84.3
GPT-2	88.1
DiWi (ours)	90.3

TABLE 1: COMPARISON OF MODEL ACCURACY IN GENERATING REALISTIC MOBILITY TRACES.

The results in TablE 1 demonstrate that DiWi achieves the highest accuracy, outperforming prior models.

3.1. Device and AP statistics

Here, we compare different statistics from the data generated by DiWi and the real dataset. We begin by analyzing the number of unique APs a device connects to in a single day, as illustrated in Figure 2: Distribution of AP visited.Figure 2.This figure presents the Empirical Cumulative Distribution Function (ECDF) of this metric, providing insights into mobility patterns. The blue line represents the real dataset, while the dashed orange line corresponds to DiWi. The results confirm the similarity of statistics between both datasets, with 80% of devices connecting to fewer than 10 unique APs per day, suggesting a localized mobility pattern where users typically connect to only one or two buildings.

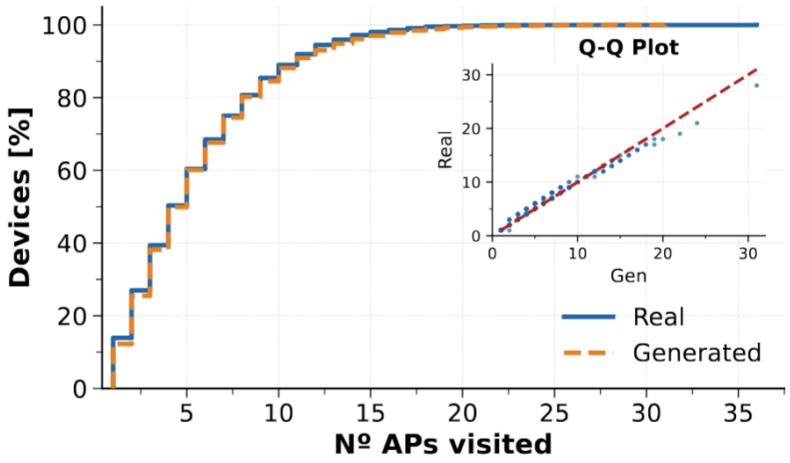

FIGURE 2: DISTRIBUTION OF AP VISITED.

To further validate this similarity, we compare both distributions using a Q-Q plot (subplot within Figure 2 that shows a strong alignment along the reference line (red dashed line), indicating that the synthetic dataset closely replicates real mobility patterns (R^2 = 0.98).

Next, we analyze how many devices connect to a given AP. Specifically, we compute the rank of APs for both datasets and compare them in Figure 3. This figure displays the proportion of devices connected to each AP, ranked from most visited to least. As in the previous case, we depict the real dataset in blue and DiWi in orange. The results reveal that in both datasets, a small number of APs handle most connections, indicating that activity is clustered in key locations (e.g., building halls, the library, and cafeterias). Both the real and synthetic datasets exhibit the same trend, with the largest difference being less than 1%, demonstrating that the synthetic dataset effectively captures key mobility behaviors.

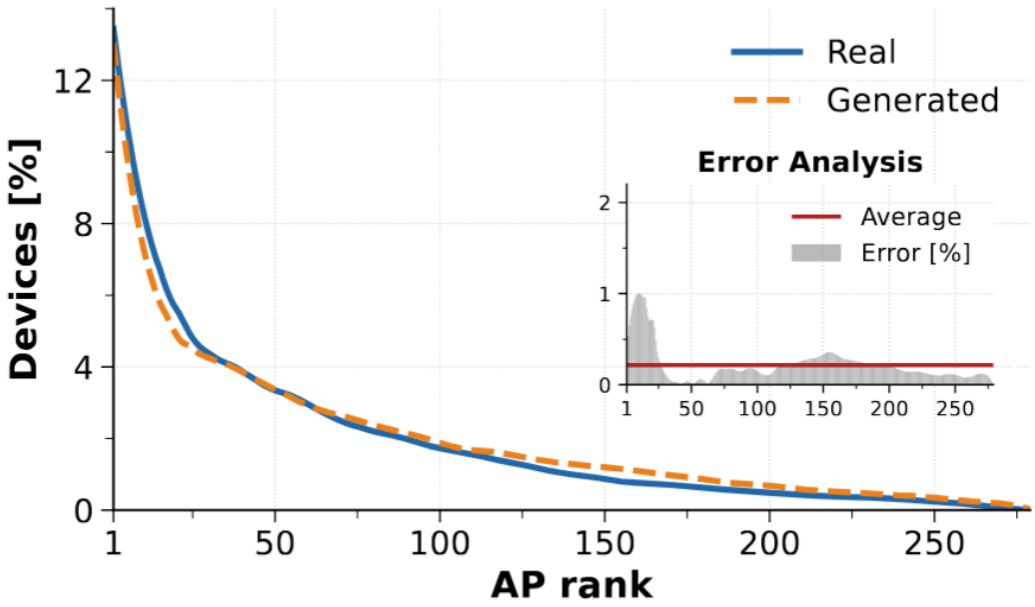

FIGURE 3: DEVICE DISTRIBUTION ACROSS APS

We next analyze temporal patterns by comparing the time of arrival and time spent in the network for each dataset, as shown in Figure 4. These figures confirm a close alignment between the real dataset and DiWi, while also capturing both short-term interactions (e.g., brief check-ins) and extended periods of connectivity (e.g., attending classes or studying in the library).

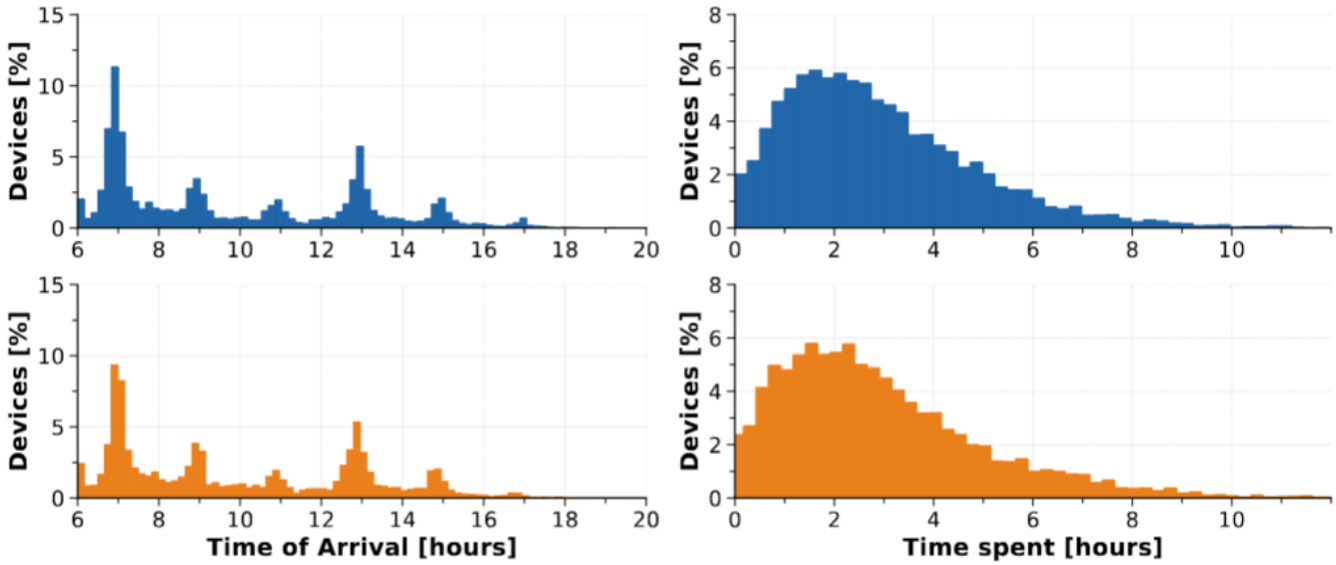

FIGURE 4: DISTRIBUTION OF ARRIVAL TIMES, AND TIME SPENT

To quantify the similarity between DiWi and the real dataset, we use the Kullback–Leibler (KL) divergence [Bench], where lower values indicate a closer match between distributions. The results of these comparisons are summarized in TablE 2. The KL divergence values demonstrate that DiWi closely aligns with the real dataset across both spatial and temporal dimensions, further validating its ability to replicate real-world mobility patterns.

Metric	KL Divergence
N° visited Aps	0.031
AP rank	0.026
Time spent	0.017
Time of arrival	0.032

TABLE 2: KULLBACK–LEIBLER (KL) DIVERGENCE BETWEEN THE REAL DATASET AND THE

3.2. Use Case 1: Sustainability

In this section, we demonstrate how DiWi can identify space occupancy patterns, such as student arrival times and duration of stay, to help optimize HVAC schedules [Trivedi2021], reduce energy consumption, and improve building management, all while maintaining privacy. Prior research [Aftad2017, Ardivanto2018] has shown that HVAC systems can significantly reduce energy use by adjusting settings based on occupancy patterns. However, traditional methods, such as camera-based systems [Aftad2017], pose privacy risks. In contrast, synthetic traces generated by DiWi

simulate device mobility without linking real and generated data (see Section 4), providing a privacy-friendly alternative.

To validate the utility of DiWi, we compare its campus space occupancy patterns with those from the real dataset. Figure 5 illustrates the total number of connected devices on campus every 5 minutes, with the blue line representing real data and the orange dashed line representing DiWi. The results confirm that DiWi accurately captures overall trends, such as:

- Minimal network activity at 6:00 AM
- Peak activity around 9:00 AM
- A gradual decrease in occupation from 5:00 PM

These findings suggest that DiWi effectively replicates real occupancy dynamics, supporting its potential application in optimizing building operations while ensuring user privacy.

In addition to per-Campus occupation, we also analyze the ocuppation in a couple of representative buildings: a classroom building and a cafeteria building in Figure 6. According to the results, both datasets result in the same activity patterns, with sharp peaks at the start times of lectures in the case of the classroom building, and at times of coffee breaks and lunch times in the case of the cafeteria.

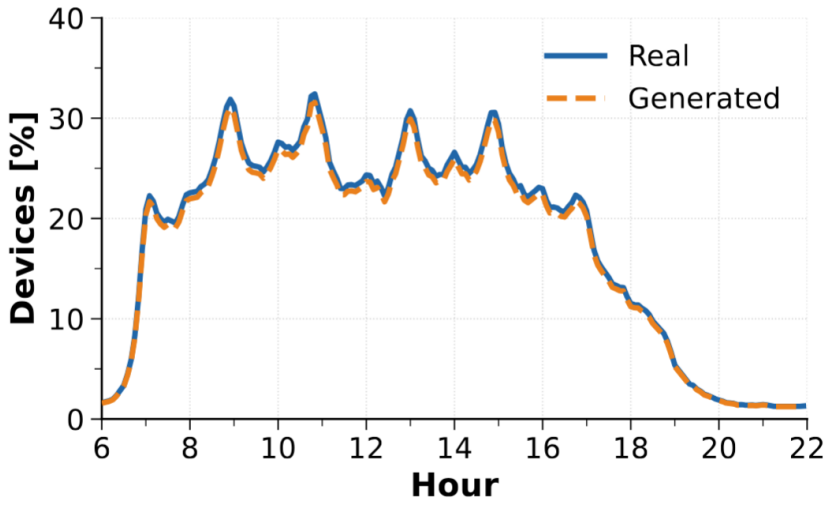

FIGURE 5: OCCUPANCY OF THE CAMPUS.

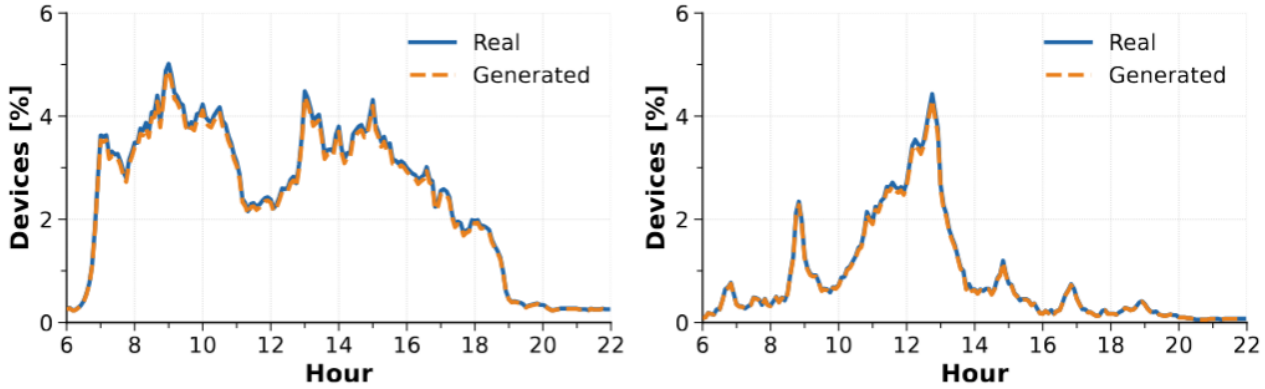

FIGURE 6: OCCUPANCY OF CLASSROOMS AND CAFETERIA

These results confirm that DiWi can be used to design intelligent HVAC systems without compromising user privacy. For instance, it is clear that e.g. the system could be switched off at 7:00 AM, as fewer than 5% of devices remain connected across the university at that time. Furthermore, more sophisticated approaches could be designed, with a proper isolation of zones, e.g., the HVAC system in the the cafeteria between 3:00 PM and 5:00 PM could be switched to a low power mode (the actual desing of these policies is outside the scope of this work).

3.3. Use Case 2: Mobility Management

Here we evaluate Diwi's ability to support the design of advanced mobility management schemes, i.e., mechanisms forecast the movement of devices, to support the design of e.g. better mobility schemes or the design of infrastructure on demand approaches by activating APs as needed. More specifically, here we focus on the ability of Diwi to analyze the connection history of a device, which we refer to as context, and then forecast its next connection point. To assess the model's predictive performance, we evaluate how its accuracy varies with the amount of context provided for inference. To compute this, we first select the **context length** used by the model for inference. Then, we sample a sequence of the chosen length for each device and use it to predict the next connection. This process is repeated across different context lengths, allowing us to assess how the model performs under varying amounts of historical data.

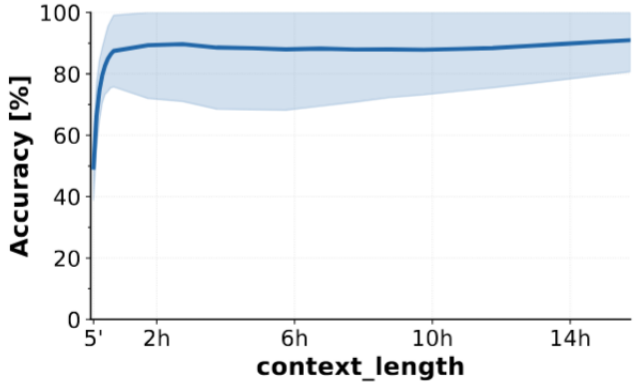

FIGURE 7: ACCURACY AS A FUNCTION OF THE CONTEXT LENGTH

We represent in Figure 7 the accuracy as a function of **context_length**, measured in hours, where each hour corresponds to 12 tokens (with each token representing a 5-minute interval). The shaded regions around the curve represent twice the standard deviation (2σ), indicating the variability in accuracy across samples. The results illustrate that accuracy improves as context length increases but gains level off beyond 2 hours. This suggests that while additional context initially enhances predictions, there is a point of diminishing returns. We note that there is a slight increase in accuracy after 10 hours of context, which we conjecture that may be caused by the ability of DiWi to leverage context from the previous day to predict the behavior on a given next day.

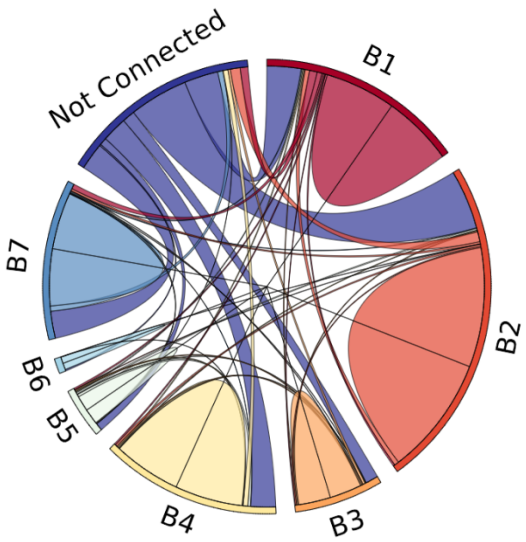

FIGURE 8: TRANSITIONS BETWEEN 6:00-7:00

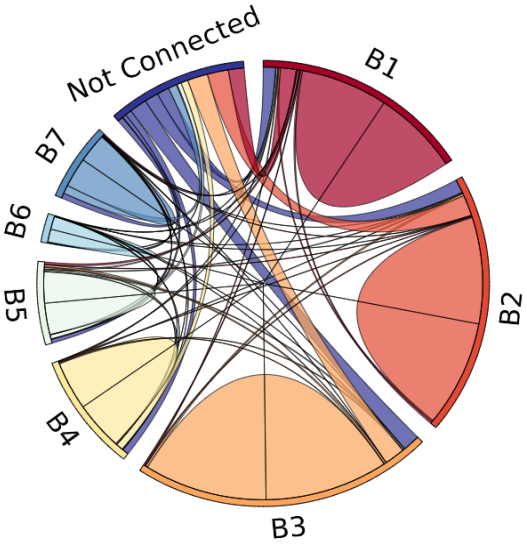

FIGURE 9: TRANSITIONS BETWEEN 11:00-12:00

Figure 8 and illustrate the mobility patterns of devices across different buildings and states (e.g., "Not Connected") between two distinct ranges, 6:00-7:00 and hour 11:00-12:00. Both diagrams illustrate the most frequent types of connections as percentages. Fig shows that between 6:00 and 7:00, most connections come from devices arriving at the university or moving within the same building, indicating localized movement. In contrast, Figure 9 reveals that by 11:00 to 12:00, most devices are already connected to the network, so transitions primarily occur within or between buildings.

These visualizations provide insight into device behavior at different times, revealing patterns in building occupancy and disconnection trends. For example, the number of devices connected to APs in building 3 is low around 6:00 but increases significantly by 11:00. Additionally, the highest disconnection rates are observed in buildings 1, 2, and 3.

3.4. Use Case 3: Understanding Anonymity schemes

In this section we describe how Diwi enables the development of privacy-preserving schemes. This is motivated by previous studies, such as [Montjoye2013] and [Cunche2014], which have identified some variables related to network activity that can be used to unequivocally identify a device (without using any device ID). In this section, we briefly introduce two different metrics regarding the identifiability of a device, and illustrate how the synthetic trace provided by Diwi shares the same results as the real-life dataset.

First, we define spatiotemporal unicity as the number of randomly chosen spatiotemporal points in a device's trace that make it uniquely identifiable. Following this definition, the uniqueness of a dataset is the average unicity across all devices in that dataset, which summarizes how identifiable the devices are. We represent in Figure 10 the unicity for an increasing number of randomly chosen spatiotemporal points. According to the figures, both the real dataset and Diwi produce vary similar results, e.g., with four spatiotemporal points, 93% of devices in the real dataset are unique, compared to 91% in the synthetic dataset. These results confirm that Diwi preserves the privacy dynamics of the real data, making it a useful tool for evaluating privacy risks in mobility networks and designing new Randomized and Changing MAC (RCM) schemes.

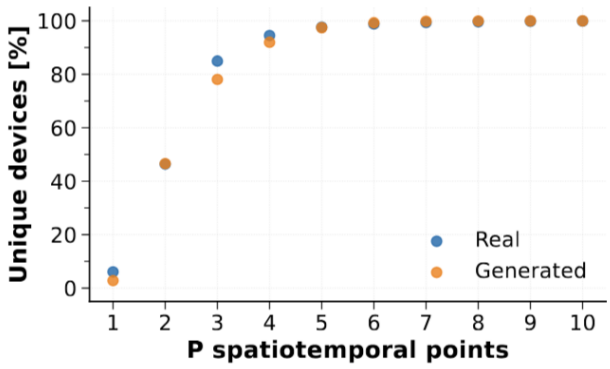

FIGURE 10: SPATIOTEMPORAL UNIQUENESS.

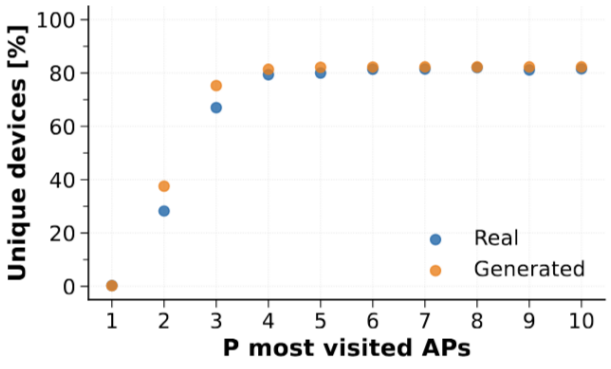

FIGURE 11: FINGERPRINT UNIQUENESS

Second, we can also defined a device's uniqueness based on its most frequently visited APs. To do this, we create a fingerprint for each device using the list of APs where it spends most time. Following this, a device is considered unique if no other device has the same list of top p APs in that order. We represent in Figure 11 the relative number of unique devices in a trace as p increases. As in the previous case, both the real dataset and Diwi follow a similar pattern, e.g., with fingerprints of 4 APs, 77% of devices are unique in the real dataset, compared to 81% in the synthetic dataset. Beyond p = 5, uniqueness levels off, suggesting that adding more APs to the fingerprint does not significantly increase a device's distinctiveness.

4. Privacy Assessment

In this section, we assess the potential risks of real-life data from the use of Diwi. This is essential to confirm that the generated data can mimic mobility patterns (as illustrated above) while preventing the exposure of sensitive information, making it suitable for the applications exposed in Section 3.

4.1. Exact matches

In this section, we assess the potential risks of real-life data from the use of Diwi. This is essential to confirm that the generated data can mimic mobility patterns (as illustrated above) while preventing the exposure of sensitive information, making it suitable for the applications exposed in the previous section.

First, we assess the risk of the model leaking complete real traces, i.e., the possibility of generating a trace that is identical to one from the training dataset. This scenario represents a significant concern as it would indicate two problems: (1) the model is overfitting and failing to generalize, and (2) the behavior of a real device is being exposed in the synthetic dataset. Such outcomes would compromise the purpose of using synthetic data by diminishing its privacy-preserving benefits.

To assess this risk, we analyze the trace generation process described, where traces are generated sequentially based on the model's learned probabilities. Assuming independence, the likelihood of reproducing a training trace is computed as the product of probabilities at each step (or equivalently, the sum of log probabilities). During synthetic dataset generation, we typically select the most likely token at each step. However, to evaluate the probability of reproducing a training trace, we take a different approach: instead of selecting the most likely AP, we use the actual AP observed in the training trace. By summing the log probabilities of these events across the sequence, we estimate the likelihood of generating an exact replica of a training trace.

Figure 12 presents the distribution of log probabilities for generating traces from the training set (blue) and the synthetic dataset (orange). The x-axis represents the log probability, while the y-axis shows the percentage of training traces with a given probability. The results indicate that the average probability of exactly reproducing a training trace is approximately 10^{-120}. This suggests that, on average, the model would need to generate 10^{-120} traces to replicate a single training trace, reinforcing that the model generalizes patterns rather than memorizing them.

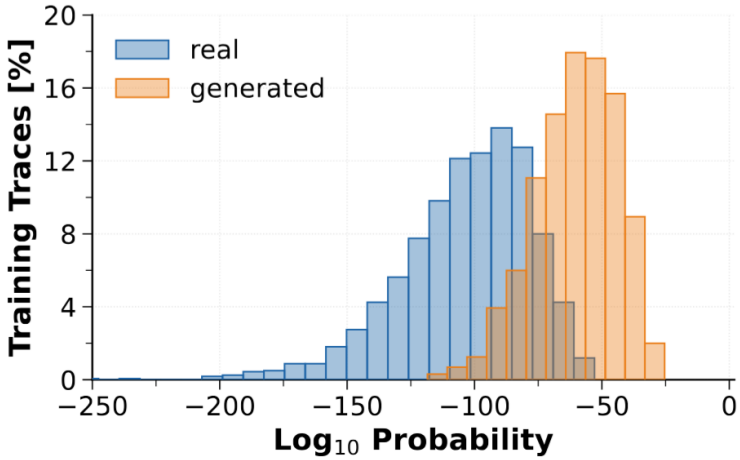

FIGURE 12: DISTRIBUTION OF THE LIKELIHOOD OF GENERATING A TRACE.

The orange distribution represents the probability of generating traces from the synthetic dataset used in this study. While some synthetic traces exhibit probability values close to those in the training set, the overlap remains limited. The overall probability of generating new, distinct traces remains dominant, indicating that the model primarily creates new patterns rather than reproducing previously seen data.

Interestingly, some traces are more likely to be replicated than others. This variation, discussed further in the next section, may result from certain AP sequences appearing more frequently in the training dataset. While this imbalance reflects the real-world distribution of user activity and preferences on campus, it is also necessary for the model to accurately capture device behavior. By representing these natural patterns, the synthetic dataset maintains the fidelity of user behavior without directly exposing sensitive information.

4.2. Closeness between datasets

Above we have analyzed the probability of the Diwi generating an exact replica of a real device's trace, which is practically negligible. However, the model may still produce synthetic traces that closely resemble real ones. For instance, consider two devices with similar mobility patterns: both connect to a sequence of the same APs but differ at one or two points. While these traces are not identical, their strong similarities may result in the synthetic data inadvertently revealing aspects of the real device's mobility pattern.

To assess this risk, we adapt inter- and intra-similarity analysis [Wolf2024], originally used for benchmarking dataset quality, to evaluate privacy leakage. Specifically, we measure the "closeness" between real-to-real traces and real-to-synthetic traces. By comparing the inter- and intra-similarity scores between pairs of real devices and pairs of real and synthetic devices, we evaluate whether the synthetic data retains identifiable patterns from the real data. If the synthetic traces are too similar to the real ones, privacy risks may emerge. Conversely, traces that are too dissimilar could make it obvious which traces are synthetic. Striking a balance ensures that the synthetic data preserves the utility of the real data without compromising privacy.

To quantify similarity, we use the Hamming distance between two traces. This metric counts the number of matching APs in the same order, excluding instances of the OUT state. This exclusion is justified because the OUT state does not provide meaningful information about a device's mobility. Using this metric, we calculate the percentage of a real trace that is "replicated" in a synthetic trace, providing a concrete measure of how much real mobility data is reflected in the synthetic dataset.

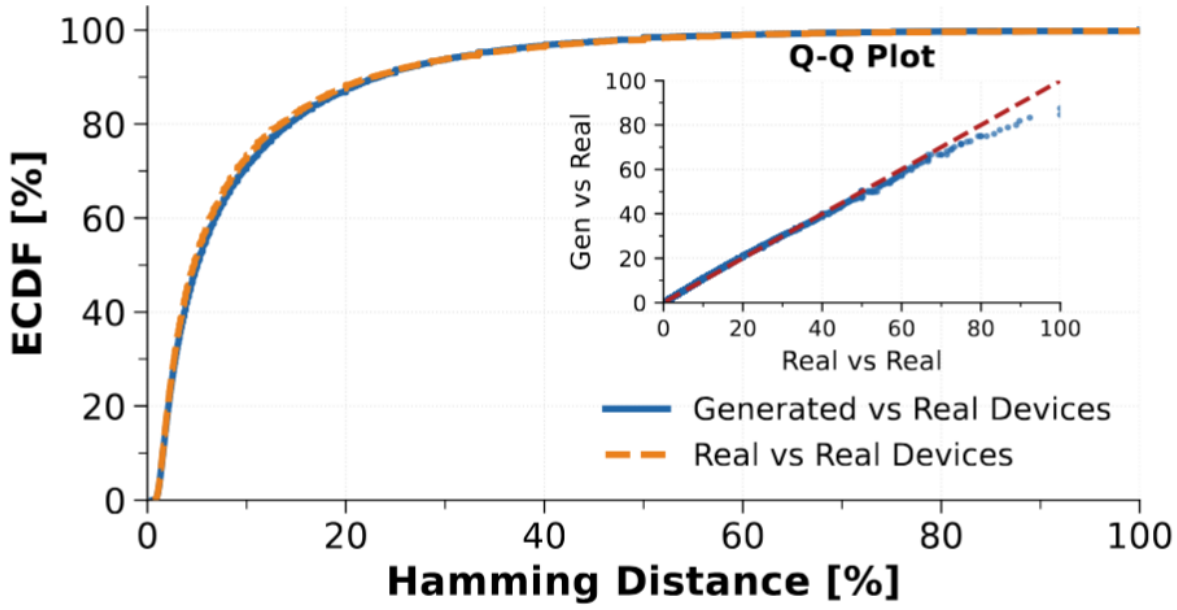

FIGURE 13: HAMMING DISTANCE BETWEEN TRACES.

Figure 13 presents the ECDF of the proportion of training traces that overlap with other training traces and with generated traces. In both cases, 90% of training traces intersect with at most 20% of others, indicating that generated traces are as similar to real traces as real traces are to each other. The QQ plot further confirms a strong alignment between the distributions, suggesting that the generated traces maintain a balance between similarity and diversity, reducing the risk of privacy breaches.

4.3. Membership Inference

We now examine a scenario in which an attacker has partial access to real device traces and the ability to query the model to generate synthetic traces.

In this attack, the goal is to train a classifier that determines whether a given trace comes from the real training dataset D_{tr} or the synthetic dataset D_{synth}.The attacker has access to both synthetic traces and a portion of the training data but relies solely on the information within the traces to distinguish between them.

To make this distinction, the attacker trains a classifier to assign a label of 1 to traces from the training set and 0 to synthetic traces. The classification is based on device behavior profiles, which include features such as time of arrival, duration of connection, most frequently used AP, and the number of APs visited.

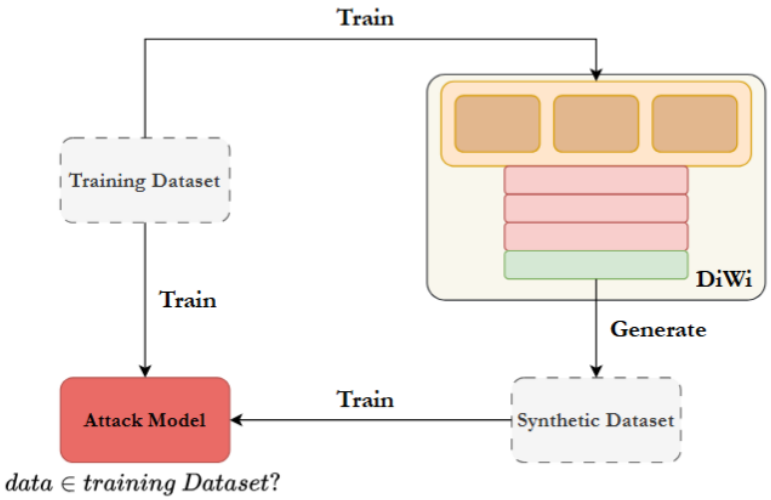

FIGURE 14: MEMBERSHIP INFERENCE ATTACK SCHEME.

As depicted in Figure 14, the attacker first accesses real traces from Diwi 's training dataset, constructs profiles for these traces, and assigns them the label 1. The attacker then queries the model to generate synthetic traces, computes profiles for these traces, and assigns them the label 0. Using this labeled dataset, the attacker trains a classifier to conduct the membership inference attack. By employing an explainable model, the attacker can identify which features the model uses to make its decisions, uncovering correlations and patterns that may reveal potential vulnerabilities in the synthetic dataset.

Metric	%
Accuracy	56
Precision	54
Recall	54
F1-Score	54

TABLA 3: MIA SCORES.

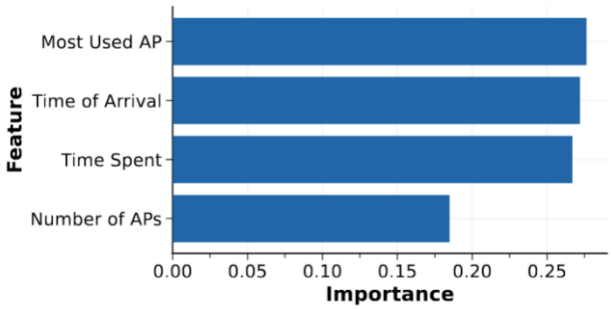

FIGURE 155: FEATURE IMPORTANCE.

In table 15 we present the results for decision trees and random forests. The observed accuracies are 52% and 56%, respectively, which are only slightly better than random guessing. In a binary classification task like distinguishing between real and synthetic traces, a random classifier would achieve approximately 50% accuracy. The minimal improvement over this baseline suggests that the models struggle to identify meaningful patterns or signals to reliably distinguish between the two classes. This indicates that the synthetic dataset closely replicates the distribution of connectivity patterns observed in the real dataset.

Figure 15 presents the feature importance of the Random Forest model used for the membership inference attack, which determines whether a given trace was part of the training dataset. The results show that Most Used AP is the most influential feature, indicating that access point usage patterns contribute to distinguishing between real (training) and synthetic traces. However, its importance is

not significantly higher than other features, and the overall distribution of feature importance is relatively uniform.

These results suggest that the membership inference attack struggles to distinguish between real and synthetic traces, with accuracy barely exceeding random guessing. The lack of a strong signal in feature importance further indicates that the generated dataset closely mirrors real connectivity patterns without exposing identifiable information. This reinforces the idea that Diwi effectively preserves privacy while maintaining realistic mobility patterns.

5. Summary and Conclusions

In this work, we have analyzed a dataset capturing the mobility patterns of users connected to the campus WLAN, using it as a representative testbed to study device behavior in dense wireless environments. Building on our previous analysis of the state of the art in AI-driven resource allocation for vRAN, our objective here is to model device activity at a fine-grained level using Transformer models, leveraging their proven ability to process sequential data and capture long-range dependencies effectively.

Our proposed model employs a Transformer architecture to analyze and predict device connectivity sequences while incorporating temporal encoding techniques that reduce the number of parameters and improve generalization. This design enables the generation of realistic spatiotemporal sequences that accurately mimic real user behavior, providing a powerful tool for modeling mobility-driven demand in virtualized wireless networks. We validate its performance through a comparative analysis with real traces and demonstrate its applicability to several use cases relevant to vRAN operations, including space optimization, intelligent building management, user occupancy prediction, network mobility forecasting, and privacy-aware resource planning.

Recognizing the sensitivity of mobility data, we also conducted a comprehensive privacy risk assessment to ensure that the generated synthetic data does not compromise user confidentiality. This evaluation examined potential risks such as direct leakage, similarity searches between real and synthetic traces, and membership inference attacks. Across all metrics, our results confirm that the model effectively generalizes behavioral patterns without exposing identifiable information or traceable device-level signatures.

By combining accurate mobility modeling with strong privacy guarantees, this work establishes a foundation for integrating synthetic data generation into AI-driven vRAN resource orchestration. These results provide the basis for future studies on privacy-preserving mobility analysis, performance evaluation of dynamic reconfiguration strategies, and predictive resource allocation in virtualized wireless networks.

References

[Trivedi2020] A. Trivedi, J. Gummeson, P. Shenoy, Empirical characterization of mobility of multi-device internet users (2020). URL https://arxiv.org/abs/2003.08512

[Zhang2014] Y. Zhang, User mobility from the view of cellular data networks, in:IEEE INFOCOM 2014-IEEE Conference on Computer Communica-tions, 2014, pp. 1348–1356. doi:10.1109/INFOCOM.2014.6848068.

[Planchy2016] J. Plachy, Z. Becvar, E. C. Strinati, Dynamic resource allocation exploiting mobility prediction in mobile edge computing, in: 2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), 2016, pp. 1–6. doi:10.1109/PIMRC.2016.7794955.

[Pan2014] J. Pan, R. Jain, S. Paul, A survey of energy efficiency in buildings and microgrids using networking technologies, IEEE Communications Surveys & Tutorials 16 (2014) 1709–1731.

[YData2023] YData, How is diversity preserved while ensuring privacy in synthetic data? (2023). URL https://ydata.ai/resources/how-is-diversity-preserved-while-ensuring-privacy-in-synthetic-data

[Montjoye2013] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, V. D. Blondel, Unique in the Crowd: The privacy bounds of human mobility, Scientific Reports 3 (1) (2013) 1376. doi:10.1038/srep01376. URL https://doi.org/10.1038/srep01376.

[Surendra2017] H. Surendra, S. MohanH, A review of synthetic data generation methods for privacy preserving data publishing, International Journal of Scientific & Technology Research 6 (2017) 95–101. URL https://api.semanticscholar.org/CorpusID:67051890.

[Kaplan2020] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling laws for neural language models, ArXiv abs/2001.08361 (2020). URL https://api.semanticscholar.org/CorpusID:210861095.

[Trivedi2021] A. Trivedi, K. Silverstein, E. Strubell, P. Shenoy, M. Iyyer, Wifimod: Transformer-based indoor human mobility modeling using passive sensing, in: Proceedings of the 4th ACM SIGCAS Conference on Computing and Sustainable Societies, COMPASS '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 126–137. doi:10.1145/3460112.3471951.

[Aftad2017] M. Aftab, C. Chen, C.-K. Chau, T. Rahwan, Automatic hvac control with real-time occupancy recognition and simulation-guided model predictive control in low-cost embedded system, Energy and Buildings 154 (2017) 141–156. doi: https://doi.org/10.1016/j.enbuild.2017.07.077. URL https://www.sciencedirect.com/science/article/pii/S0378778817305091

[Ardivanto2018] D. Ardiyanto, M. Pipattanasomporn, S. Rahman, N. Hariyanto, Suwarno, Occupant-based hvac set point interventions for energy savings in buildings, in: 2018 International

Conference and Utility Exhibition on Green Energy for Sustainable Development (ICUE), 2018, pp. 1–6. doi: 10.23919/ICUE-GESD.2018.8635595.

[Cunche2014] Cunche, M. I know your MAC address: targeted tracking of individual using Wi-Fi. *J Comput Virol Hack Tech* **10**, 219–227 (2014). https://doi.org/10.1007/s11416-013-0196-1

[Wolf2024] Wolf M., Tritscher J., Landes D., Hotho A., Schlör D. Benchmarking of synthetic network data: Reviewing challenges and approaches, Computers & Security, Volume 145, 2024, 103993, ISSN 0167-4048, https://doi.org/10.1016/j.cose.2024.103993.