On the Feasibility of Reference-Free Tracking via Motion Sensors in Public Transport

Anne Josiane Kouam
Inria & TU Berlin
Palaiseau, France
0000-0002-8803-6256

Stefan Czybik
BIFOLD & TU Berlin
Berlin, Germany
0009-0001-3342-7460

Konrad Rieck
BIFOLD & TU Berlin
Berlin, Germany
0000-0002-5054-8758

***Abstract*—Smartphone motion sensors have been identified as a potential source of privacy leakage regarding user mobility in public transport networks. However, prior trajectory inference attacks often rely on strong reference data, including prior recordings on the target transport network and fixed phone placement, limiting the understanding of real-world attack feasibility. In this work, we investigate the feasibility of *reference-free* trajectory inference from smartphone motion sensors under realistic sensing and attacker assumptions. We design an unsupervised inference pipeline in which the attacker relies only on publicly available transport information combined with structural features extracted from accelerometer and gyroscope signals in order to progressively reduce the set of possible trajectories and apply a plausibility-based ranking. We evaluate the proposed framework on 172 real-world motion sensor recordings collected across 21 transport lines in two German public transport networks. Our findings indicate that exact trajectory reconstruction without reference information remains challenging in practice. Nevertheless, coarse motion sensor observations still reveal meaningful mobility information and can substantially reduce trajectory uncertainty, particularly in structurally distinctive transport corridors.**

***Index Terms*—Trajectory Inference, Mobile Sensors, Privacy**

I. Introduction

Smartphones and wearable devices have become deeply embedded in our daily life, supporting essential activities such as communication, navigation, and financial transactions. To enable these functionalities, modern devices integrate a wide range of sensors that continuously capture information about the user environment. Among these, motion sensors such as accelerometers, gyroscopes, magnetometers, and barometers provide fine-grained measurements of device motion and physical surroundings. Originally designed to support functionalities such as activity tracking, gesture control, or device orientation detection [1], these sensors have progressively attracted attention due to their potential privacy implications.

Over the last decade, a growing body of work has shown that motion sensors can leak surprisingly sensitive information. Prior studies demonstrated that motion sensor data can enable the inference of human activities (e.g., walking, running, sitting [2]), transportation modes (e.g., car, metro, train, bus [3–5]), device colocation [5] or allow device fingerprinting [6].

Among these risks, the inference of user mobility is particularly critical due to its ability to disclose a wide range of personal information: where users live or work, their daily routines, or even their relationships. Prior work has demonstrated the feasibility of tracking user trajectories in public transport networks using motion sensor data [3, 7], thereby revealing the sequence of stations or segments traversed during a trip. In these approaches, the attacker typically collects reference sensor traces across the target transport network under controlled conditions and later matches victim recordings against these pre-collected patterns. We refer to such attacks as reference-based trajectory inference. For example, Hua et al. [3] report trajectory inference accuracies approaching 90% on the Nanjing metro system, while Nguyen et al. [7] achieve nearly perfect identification on London Underground trips containing more than four stops but using the length of the recording as a discriminative feature.

While these works suggest that motion sensor-based trajectory inference is feasible, we argue that the highlighted privacy risk may be difficult to translate to realistic attack scenarios. This is because reference-based approaches rely on two strong assumptions: First, due to their supervised setting, they assume that the attacker has previously collected reference traces across the targeted public transport network, often by extensively riding trains and recording motion sensor data beforehand. In practice, however, such large-scale data collection campaigns are costly, difficult to maintain over time, and inherently limited to networks that have already been surveyed by the attacker. Second, prior work relies on sensor recordings obtained under carefully controlled conditions, such as fixed device placement or stable phone orientation. Real-world usage conditions differ substantially: users handle their devices, place them in pockets or bags, and change orientation unpredictably during a trip, resulting in significantly noisier and less reproducible sensor observations.

As a result, it remains unclear to what extent the attack feasibility demonstrated in prior work persists under realistic sensing conditions and attacker capabilities. In this paper, *we revisit motion sensor-based trajectory inference in a reference-free setting.* We consider an attacker collecting motion sensor data from a victim's device. The attacker has no prior knowledge of the target network beyond publicly available information and instead combines the collected sensor signals with public transport network topology and schedules from sources such as OpenStreetMap (OSM) and General Transit Feed Specification (GTFS) feeds.

Under these realistic attack conditions, a fundamental privacy question emerges: *how much can noisy motion sensor observations reduce uncertainty over a user's trajectory in a complex public transport network?* To investigate this question, we model the transport network as a dynamic graph and formulate trajectory inference as a constrained search problem over plausible paths. We design a best-effort unsupervised inference pipeline that combines motion sensor-derived structural features with temporal and topological constraints extracted from public transport data. The approach is primarily evaluated on our real-world dataset collected in the Berlin public transport network and further validated on recordings from Stuttgart, comprising 172 annotated recordings across 21 transport lines in both cities.

Our results show that the high trajectory reconstruction accuracy reported in reference-based studies does not translate to realistic sensing conditions. Under reference-free and uncontrolled settings, exact trajectory reconstruction remains difficult due to sensing noise and structural ambiguity in dense transport networks. Nevertheless, motion sensor observations combined with publicly available transport information still enable substantial uncertainty reduction. On our Berlin dataset, approximately 25% of the recordings rank the ground-truth trajectory first, while the median remaining uncertainty is reduced to 25% of the initial candidate space. Inference performance further increases in structurally simpler networks such as Stuttgart, where the remaining uncertainty never exceeds 15%. Our explainability analysis reveals that the attack is most effective on structurally distinctive trajectories traversing weakly shared transport corridors and less central regions of the transport network. Overall, these findings suggest that realistic motion sensor attacks remain fundamentally constrained in practice, yet can still leak meaningful mobility information under favorable network conditions.

Contributions: The main contributions of this work are:

- We revisit motion sensor-based trajectory inference under realistic attacker assumptions without prior data collection, supervised training, or controlled sensing conditions.
- We propose a reference-free trajectory inference framework combining motion sensor observations with publicly available transport information.
- We collect an annotated dataset of 172 motion sensor recordings and nearly 24 hours across the transport networks of Berlin and Stuttgart under realistic passenger conditions.
- We conduct an extensive evaluation quantifying both the achievable uncertainty reduction and the practical limitations of realistic trajectory inference attacks.
- We show that inference performance is strongly influenced by structural ambiguity in the underlying transport network.

Roadmap: The remainder of this paper is organized as follows. Section II reviews prior work on motion sensor-based tracking and transportation mode detection. Section III introduces the threat model. Section IV presents the proposed reference-free trajectory inference pipeline. Section V evaluates the approach on real-world datasets and analyzes its practical limitations. Finally, section VI discusses the implications of our findings and section VII concludes the paper.

II. Related Works

Prior work has shown that motion sensors can leak significant information about user activities and mobility, enabling applications ranging from transportation mode detection to trajectory inference. In this work, we primarily rely on accelerometers and gyroscopes, which capture translational and rotational movements and therefore provide coarse structural information about user mobility. In the following, we review prior work on transportation mode detection and motion sensor-based trajectory inference.

A. Transportation mode detection

Transportation Mode Detection (TMD) aims to classify user activities (e.g., walking, driving, or traveling by train) based on motion sensor data. In our setting, it serves as an enabling primitive: we assume that an attacker first isolates public transport segments before attempting trajectory inference.

Early approaches, such as Hemminki et al. [8], extract statistical and frequency-domain features from accelerometer signals and rely on hierarchical classification schemes to distinguish transportation modes, achieving up to 80% accuracy. Subsequently, Carpineti et al. [9] introduced the publicly available US-TMD dataset [10] and evaluated several machine-learning classifiers, reporting accuracies of up to 93%. Several public datasets have since become available, including Collecty [11], MobilitApp [12], and NOR-TMD [13]. Czybik et al. [5] detect train rides from the presence of the 16.7 Hz traction frequency in the smartphone magnetometer. Other recent work explores deep-learning approaches based on CNNs, LSTMs, and temporal convolutional networks [14–16], further confirming that train usage can be identified from smartphone motion sensors under different operating conditions.

Collectively, these works indicate that TMD is a mature and reliable primitive, supporting our assumption that train journeys can be isolated prior to trajectory inference.

B. Sensor-based location tracking

A growing body of work has shown that motion sensors can leak location information even in the absence of explicit positioning systems such as GPS. Early systems such as ACComplice [17] reconstruct driving trajectories from accelerometer-derived turns and map matching, while surveys by Kröger et al. [18] highlight motion sensors as privacy-relevant side channels capable of revealing mobility patterns. Closer to our setting, several works have investigated location inference in public transport networks using motion sensors. Table I summarizes these approaches, which can be grouped into two categories: *localization*, which estimates a user's current position within the network and is primarily designed as a navigation service, and *trajectory inference*, which aims to reconstruct all or part of a user's journey and is often studied from a privacy perspective.

TABLE I: Comparison of public transport tracking approaches.

Work	Goal	Ref.	Ctrl.	Public	Transf.
Stockx et al. [19]	Localization	✓	✓	✓	✗
Ye et al. [20]	Localization	✓	✗	✓	✗
Hua et al. [3]	Traj. inference	✓	✓	✗	✗
Nguyen et al. [7]	Traj. inference	✓	✓	✗	✗
This paper	Traj. inference	✗	✗	✓	✓

Ref. = reference traces collected on the target network, Ctrl. = controlled sensing conditions, Public = operation using publicly available transport data, Transf. = transferability across transport networks.

Localization approaches typically estimate the current user position within a transport network. SubwayPS [19] detects train movements and station stops from accelerometer and gyroscope observations and combines them with timetable information to interpolate the passenger's position along a known route, requiring network-specific calibration and prior knowledge of the trip's origin and destination. Similarly, M-Loc [20] localizes metro passengers by matching motion sensor observations against crowdsourced fingerprints collected on the target network. Both systems report high localization accuracy in metro environments, demonstrating that motion sensors contain sufficient structural information for location inference, but rely on reference information collected or calibrated for the target transport network.

Trajectory inference approaches aim to reconstruct all or part of a user's journey from motion sensor observations. Hua et al. [3] collect labeled accelerometer traces for individual metro station intervals and train supervised classifiers to recognize trajectory segments, reporting inference accuracies between 70% and 90% depending on trip length. Similarly, Nguyen et al. [7] construct a reference database covering the entire London Underground by recording accelerometer traces under fixed phone position and orientation and match passenger observations against these fingerprints, achieving route-identification accuracies of up to 90%.

In contrast to these approaches, this paper investigates trajectory inference in a fully *reference-free* setting. We assess the privacy risks of trajectory inference attacks that rely solely on publicly available transport information and assume realistic, uncontrolled sensing conditions. By eliminating the need for network-specific reference traces, this setting fundamentally changes the nature of the attack and extends its applicability to any transport system for which public transport data are available. Consequently, it remains unclear whether the privacy risks demonstrated by reference-based attacks persist and how much trajectory uncertainty can still be reduced in practice.

III. Threat model

We consider the problem of inferring a user's trajectory within a public transport network from motion sensor data, without relying on prior data collection or supervised training. This setting differs from prior work by assuming a fully unsupervised attacker operating under realistic constraints.

Our threat model assumes an attacker who gains access to motion sensor data through a seemingly benign mobile application. Such an application can collect motion sensor readings either without explicit permission, as is still largely the case on Android platforms [1], or under permissions that users routinely grant without careful consideration (e.g., motion and fitness access on iOS [21]). The application continuously records motion sensor data and transmits it to the attacker.

We assume that the attacker knows the geographic area in which the user operates (e.g., the city). Such information may originate from prior knowledge in a targeted attack scenario or from coarse application usage context in untargeted settings. The attacker does not possess any prior knowledge specific to the target transport network or labeled training data. Instead, the attack relies exclusively on publicly available information, such as transport network topology and schedules.

Leveraging the high accuracy and availability of public datasets and pre-trained models in prior work (cf. Section II-A), we assume that the attacker reliably identifies segments corresponding to train usage, thereby separating relevant portions of the signal from other activities.

Methodological constraints: Our threat model imposes important constraints on the design of the inference methodology: First, the absence of labeled data and pre-recorded traces prevents the use of supervised or network-specific learning approaches, such as classification models for stop or segment recognition. Consequently, inference must rely exclusively on unsupervised techniques or statistical patterns. Second, the attacker only has access to theoretical schedules and network descriptions obtained from public sources, while real circulation conditions may differ due to delays, irregular dwell times, traffic disturbances, or operational disruptions. Together with noisy motion observations, these effects make exact temporal alignment infeasible in practice and require trajectory inference to progressively reduce uncertainty by ranking plausible candidate trajectories.

Under these assumptions, the attacker observes a sequence of motion sensor signals corresponding to a user train trip and aims to infer the most plausible trajectory within the public transport network.

IV. Methodology

We propose an reference-free pipeline for trajectory inference in transport networks that combines public transport information with structural features extracted from motion sensors. Our objective is to design a best-effort inference methodology that realistically evaluates the privacy risk achievable under the attacker constraints introduced in Section III, instead of achieving perfect trajectory reconstruction.

A. Overview

Let r denote a motion sensor recording collected during a single public transport trip over a time interval $[t_s, t_e]$. The recording consists of accelerometer and gyroscope observations from which the proposed methodology first extracts coarse structural mobility features such as stops, motion segments, and railway curves using unsupervised signal-processing techniques (Section **IV-B**).

Using publicly available information from GTFS and OSM, we then construct a dynamic graph representation of the transport network and derive a set of feasible candidate trajectories $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$, where each candidate p_i corresponds to a plausible public transport run whose schedule overlaps the recording interval (Section **IV-C**).

As motion sensors do not directly encode spatial location information, multiple trajectories may remain consistent with the same recording. The methodology computes a plausibility score $D(r, p_i)$ measuring the consistency between sensor-derived features and candidate-derived ones, and ranks candidates accordingly (Section **IV-D**). This comparison relies on soft sequence alignment rather than exact matching.

A successful attack therefore corresponds to a significant reduction of uncertainty over the set of plausible trajectories. This reduction is quantified through the rank of the ground-truth trajectory and the remaining uncertainty within the candidate set. Lower ranks and lower residual uncertainty indicate stronger inference capability and higher privacy leakage.

B. Sensor-side processing

The first stage of the pipeline transforms the raw motion sensor recording r into a set of coarse mobility features. The input consists of timestamped accelerometer and gyroscope signals collected during the trip interval $[t_s, t_e]$. From these signals, we extract two complementary types of structural information. First, accelerometer signals are used to infer the approximate temporal structure of the trip through stop detection, yielding an estimated number of stops and coarse motion segments between stations. Second, gyroscope signals are used to identify railway curves and estimate their angular magnitudes. These features remain deliberately coarse and capture structural properties of the underlying trajectory that can later be matched against public transport candidates.

a) Preprocessing: Accelerometer and gyroscope measurements are affected by high-frequency oscillations caused by train vibrations, track friction, and passenger movements. Before feature extraction, we therefore apply low-pass filtering to smooth the raw sensor signals. This preprocessing step suppresses short-lived fluctuations that do not correspond to macroscopic train motion while preserving the low-frequency patterns associated with stops and railway curves.

Because motion sensors are measured in the smartphone coordinate system, sensor observations additionally depend on device orientation and user behavior. To reduce this dependency, we transform sensor measurements into a common world coordinate system using rotation matrices derived from accelerometer and magnetometer signals.

b) Stop detection from accelerometer data: We use accelerometer signals to infer approximate station stops without relying on labels or network-specific training data. For each recording, we compute the acceleration magnitude and identify station stops as periods of reduced motion intensity appearing as local minima in the filtered signal. Candidate stops are detected using prominence-based peak detection together with a minimum temporal separation constraint to avoid duplicate detections. Each retained peak at time t_i defines an approximate stop interval $[t_i - \delta, t_i + \delta]$, yielding an estimate $\hat{n}_{stop}$ of *the number of stops* together with *a normalized strength score* derived from peak prominence. Detected stops are further used to derive approximate *motion segments* between consecutive stations. Figure 1 illustrates a stop-detection example, showing both the effect of low-pass filtering and the resulting approximate stop intervals compared to ground-truth stops.

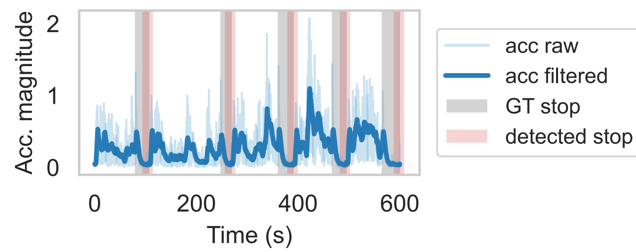


Fig. 1: Stop detection example from accelerometer signals.

c) Curve extraction from gyroscope data: Railway curves provide discriminative structural features reflecting the geometry of the underlying track. When a train traverses a curve, the smartphone undergoes a smooth horizontal rotation that appears in gyroscope observations as angular velocity around the vertical axis.

Curve extraction is performed on the gyroscope over the entire recording. Consecutive samples sharing the same rotation sign are grouped into candidate rotation intervals, and the angular variation of each interval is estimated by integrating angular velocity over time:

$$r_i \approx \sum_{t \in i} \omega(t) \, \Delta t,$$

where $\omega(t)$ denotes the gyroscope angular velocity around the vertical axis.

Candidate intervals are retained as railway curves only if their angular magnitude exceeds a threshold θ and their duration remains sufficiently long to filter out short device manipulations. For each retained curve, we store its *temporal interval*, and *angular magnitude*.

C. Transit graph modeling and candidate generation

The second stage of the pipeline models a public-transport representation from publicly available data sources and uses it to derive the set of feasible candidate trajectories $\mathcal{P}$. The approach of this stage is twofold. First, we model both the physical structure and the temporal operation of the transport network. Second, we extract structural descriptors from railway geometry that can later be compared with the sensor-derived features introduced in the previous subsection.

1) Transit graph modeling: We define a directed graph

$$G = (V, E),$$

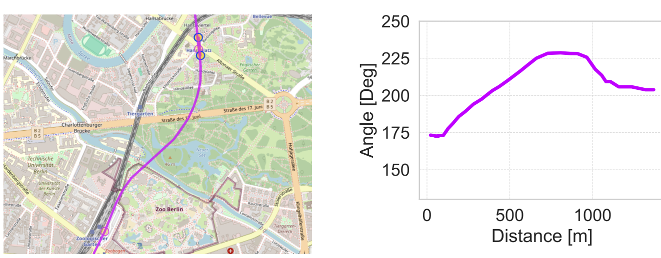
representing the physical transport infrastructure. Each node $v \in V$ corresponds to a station, while each directed edge

$e = (u, v) \in E$ represents a track segment connecting two consecutive stations in a specific travel direction.

The graph is initialized from publicly available GTFS transport feeds and route information. Ordered station sequences are extracted for each transport line and converted into directed station-to-station connections. Opposing directions are represented by distinct directed edges, since the corresponding physical tracks may exhibit different geometric properties. To enrich this topological representation with physical track information, we integrate OSM railway geometries. Using OSM railway relations and track nodes, each edge is associated with an ordered sequence of geographic coordinates describing the underlying railway segment.

From these geometries, we derive structural descriptors on a per-edge basis, including *segment length* and ordered *railway curves* characterized by their angular magnitude and rotation direction. Similar to the sensor-side processing, curves are extracted by identifying monotonic direction changes along the railway geometry exceeding a configurable angular threshold. Figure 2 illustrates an example of railway curve extraction from OSM geometries together with the corresponding angular representation used by the inference pipeline.

To incorporate temporal transport information, we additionally construct a time-dependent operational layer from GTFS schedules, stop times, routes, and service calendars. Each scheduled train run is mapped to a path in the graph together with its arrival and departure times. This representation enables efficient retrieval of trajectories whose operating times remain compatible with a recording interval $[t_s, t_e]$.



(a) Real-world railway segment geometry.

(b) Corresponding angular profile extracted along the segment.

Fig. 2: Illustration of railway curve extraction of the segment Hansaplatz-Zoologischer Garten from OSM geometries.

2) Candidate generation and filtering: Given a recording interval $[t_s, t_e]$, the attacker first retrieves all scheduled trips whose operating times remain compatible with the observation window under a temporal tolerance margin. Concretely, candidate trips are selected if their scheduled departure and arrival times fall within a configurable slack interval around the sensor recording window.

This produces an initial candidate set $\mathcal{P}$, where each candidate trajectory corresponds to a feasible scheduled transport run represented as a path in the time-dependent transit model. To further reduce the search space, we apply lightweight consistency constraints designed to eliminate clearly incompatible candidates while remaining robust to sensing uncertainty and schedule imperfections: First, candidates whose number of scheduled stops differs substantially from the detected stop count $\hat{n}_{stop}$ are discarded. Second, candidate departure and arrival times must remain compatible with the recording interval $[t_s, t_e]$ under relaxed temporal tolerances.

This filtering stage intentionally favors recall over precision. As stop detection remains approximate and public schedules may deviate from real circulation conditions, aggressive early filtering could incorrectly eliminate the true trajectory. Consequently, all trajectories that remain temporally and structurally plausible are preserved for the subsequent ranking stage.

D. Trajectory ranking via sequence alignment

After candidate generation and filtering, the remaining trajectories are compared against the sensor recording using a soft sequence-alignment strategy. To this end, we first derive comparable structural feature sequences from both sensor observations and candidate trajectories, then compute feature-wise similarities between these sequences, and finally combine them into a global plausibility score used for ranking.

1) Candidate feature sequences: Let $S^{(r)} = (s_1, \ldots, s_m)$ denote the sequence of motion segments extracted from recording r in Section IV-B, and $p = (e_1, \ldots, e_n)$ a candidate trajectory represented as an ordered sequence of graph edges derived in Section IV-C. For both representations, we build comparable feature sequences:

- *Curve count.* For each sensor motion segment s_i and each candidate edge e_j, we count the number of detected curves whose angular magnitude exceeds the curve threshold θ. This yields one sequence over sensor segments and one sequence over candidate edges.
- *Cumulative curve amplitude.* For each segment or edge, we sum the absolute angular amplitudes of all curves above the threshold θ. This captures how strongly the trajectory bends within each portion.
- *Normalized duration/length pattern.* Sensor recordings provide the duration Δt_i of each motion segment, while the public graph provides the physical length ℓ_j of each candidate edge. To compare relative movement patterns independently of absolute timing, we normalize both quantities:

$$T_i^{(s)} = \frac{\Delta t_i}{\sum_{k=1}^{m} \Delta t_k}, \qquad L_j^{(p)} = \frac{\ell_j}{\sum_{k=1}^{n} \ell_k}.$$

  The resulting sequences capture how movement is distributed across the trip on both the sensor and network sides.
- *Turn-shape sequence.* Beyond aggregated curve statistics, we additionally preserve finer directional variations along the trajectory. On the sensor side, these sequences are derived from gyroscope-based yaw increments, while on the candidate side they are extracted from bearing changes along OSM geometries. This representation captures the local shape of turns and curves along the trip.

2) Dynamic Time Warping alignment: For each feature, similarity between the sensor-derived sequence and the candidate-derived sequence is computed using Dynamic Time Warping (DTW). DTW is well suited to our setting because the two sequences may have different lengths: detected motion segments may not perfectly correspond to graph edges, and stop detection may miss or merge stops. Instead of enforcing a strict one-to-one correspondence, DTW finds a minimum-cost alignment path between the two sequences. We use a normalized DTW distance, defined as the total alignment cost divided by the length of the alignment path. This avoids systematically favoring shorter candidates and makes distances more comparable across trajectories.

The local DTW cost depends on the feature type. For angular features (curve amplitudes and turn-shape values), we use normalized angular differences (δ_{angle}), while normalized duration/length patterns rely on absolute differences. For curve counts, differences are normalized by the largest observed count using δ_{count}, where $S = \max(\text{count values}, 1)$ avoids scale bias across recordings:

$$\delta_{\text{angle}}(a, b) = \frac{|a - b|}{180}, \quad \delta_{\text{count}}(a, b) = \frac{|a - b|}{S}.$$

3) Final ranking: For a candidate trajectory p, the final plausibility score is computed as a weighted combination of the feature-specific DTW distances:

$$D(p) = \sum_k w_k \cdot \hat{d}_k(p),$$

where $\hat{d}_k(p)$ denotes the normalized DTW distance associated with feature k, and w_k its adaptive weight.

The normalized distances $\hat{d}_k(p)$ are obtained through min–max normalization across all candidates for the current recording. This normalization makes distances comparable across heterogeneous features whose raw DTW costs may operate on different numerical scales. On the other hand, feature weights w_k cannot be learned through supervised optimization. Instead, we estimate them adaptively from the candidate distribution associated with the current recording. Concretely, the weight of feature k is chosen proportional to the interquartile range (IQR) of its normalized distance distribution:

$$w_k \propto \text{IQR}\big(\hat{d}_k(p)\big).$$

Intuitively, features producing similar distances for all candidates contribute little to discrimination and are automatically downweighted, whereas features producing stronger separation between candidates receive larger weights.

Candidates are finally ranked by increasing plausibility score. Rather than deterministically reconstructing a unique path, the methodology progressively reduces uncertainty by prioritizing trajectories whose structural characteristics remain most consistent with the observed sensor recording.

V. Evaluation

We now evaluate the proposed reference-free trajectory inference methodology under realistic sensing and attacker assumptions using datasets collected in the Berlin and Stuttgart public transport networks. The analysis focuses on inference performance and on understanding the factors that influence successful trajectory discrimination. More specifically, we address the following research questions:

- **RQ1 – Performance:** *To what extent can reference-free trajectory inference reduce trajectory uncertainty under realistic attacker assumptions?*
- **RQ2 – Explainability:** *Which structural and operational factors influence inference performance and explain observed successes or failures?*
- **RQ3 – Generalization:** *How do the different features contribute to inference performance, and how robust is the methodology across different transport networks?*

We first describe the datasets collected for the evaluation. We then evaluate the attack performance and analyze its explainability on our Berlin dataset. Finally, we conduct an ablation study and discuss the generalization of the methodology to the Stuttgart transport network.

A. Data collection

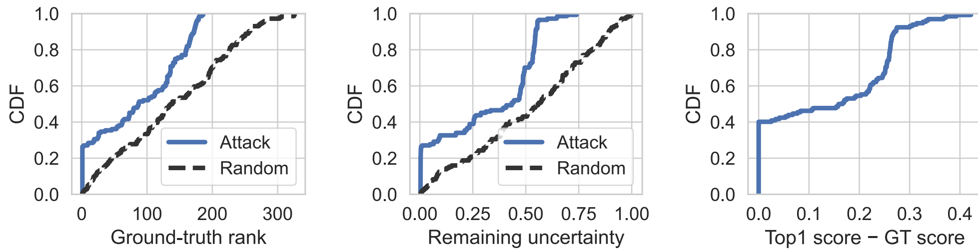
We collect smartphone sensor recordings in the Berlin and Stuttgart public transport networks under realistic passenger conditions. Recordings were acquired using commodity smartphones and the *Sensor Logger* Choi [22] application during regular train journeys across multiple transport lines, stations, recording days, and operational conditions. Participants received a short set of instructions for recording and labeling data but were otherwise free to use their devices naturally throughout the trips, including holding the phone in the hand, placing it in pockets, or interacting with applications during travel. Consequently, the dataset captures realistic sensing variability rather than controlled placement conditions.

Each journey was manually annotated with station names together with approximate arrival and departure timestamps recorded during the trip. These annotations enable the reconstruction of the ground-truth trajectory used exclusively for evaluation. Journeys involving line changes were decomposed into independent recording segments and analyzed separately. Table II summarizes the characteristics of both datasets.

TABLE II: Overview of the datasets used in the evaluation.

	Berlin	**Stuttgart**
Unique recordings	160	12
Total duration [h]	20.17	3.63
Distinct collection days	45	6
Unique station pairs	138	22
Unique stations	86	12
Transport lines	17	4
Collection period	Feb.–Sep. 2024	Feb.–Aug. 2024

Our Berlin dataset constitutes the primary evaluation corpus. It contains 160 recordings with a cumulative duration of more than 20 hours, collected over 45 distinct days between February and September 2024. The recordings cover 138 distinct station pairs and 86 unique stations distributed across



(a) GT rank (b) Uncertainty (c) Score gap

Fig. 3: Distribution (CDF) of attack performance on our Berlin dataset. Lower ground-truth ranks, lower remaining uncertainty, and smaller Top-1 score gaps indicate stronger trajectory discrimination.

17 transport lines, including suburban rail (S-Bahn), underground metro (U-Bahn), and regional rail services. Our smaller Stuttgart dataset is used as an additional cross-city validation scenario to evaluate the transferability of the methodology

B. Attack performance

We evaluate the attack performance on our Berlin dataset by analyzing how effectively the proposed pipeline reduces trajectory uncertainty and prioritizes the ground-truth trajectory among the remaining plausible candidates.

Figure 3a shows the cumulative distribution of the ground-truth rank after the soft-ranking stage together with a theoretical random baseline that assigns a uniformly random ordering to the same candidate set. The results reveal a highly heterogeneous inference performance across recordings. In approximately 25% of the recordings, the ground-truth trajectory is ranked first. Furthermore, the attack places the ground-truth trajectory within the top-10 and top-25 candidates in 28.5% and 31.9% of the recordings, respectively. The corresponding aggregate performance metrics are summarized in Table III. Although the methodology does not consistently recover the exact trajectory, it substantially improves over random ranking, reducing the median ground-truth rank from 137.5 to 87. This indicates that the extracted motion features provide meaningful information beyond chance-level candidate ordering.

To better quantify this reduction, Figure 3b reports the remaining uncertainty ratio after ranking, defined as the normalized position of the ground-truth trajectory within the set of time-compatible candidate trajectories. A value close to zero indicates strong uncertainty reduction, whereas values closer to one correspond to weak discrimination. The median remaining uncertainty is approximately 0.25, meaning that in half of the recordings the ranking stage eliminates around 75% of the initially plausible trajectories. In comparison, the random baseline yields a median remaining uncertainty of approximately 0.56. Overall, the proposed methodology reduces the median uncertainty by more than a factor of two, demonstrating that coarse motion sensor observations provide substantial trajectory information despite realistic sensing noise and the absence of explicit localization signals.

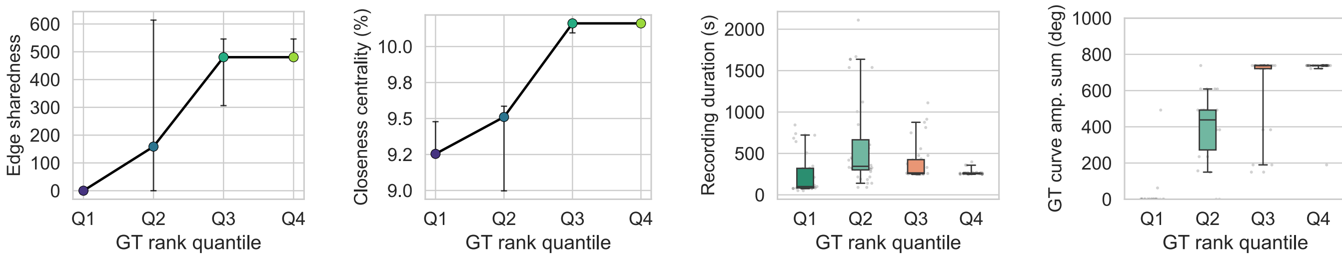
Finally, Figure 3c analyzes the score separation between the top-ranked candidate and the best-ranked ground-truth candidate. Overall, the observed score gap remains relatively small, with approximately 40% of the recordings exhibiting a zero score gap and a median gap of only 0.16. This indicates that the ground-truth trajectory often remains highly competitive with respect to the top-ranked candidate, even when it is not ranked first. At the same time, the presence of small but nonzero gaps reveals structurally similar alternative trajectories that, according to the extracted motion and geometric features, appear equally or sometimes more plausible than the true trajectory. These results suggest that many inference failures arise from intrinsic structural ambiguity within the transport network rather than purely random ranking behavior. The following subsection therefore investigates which structural and operational factors contribute to these ambiguities and influence inference performance.

RQ1: *Under realistic reference-free sensing conditions, exact trajectory reconstruction becomes significantly harder than reported in prior supervised work. Nevertheless, motion sensor observations still enable substantial uncertainty reduction and frequently narrow the set of plausible transport trajectories, even when several candidates remain highly competitive.*

C. Factors influencing inference performance

To better understand which factors influence inference quality, we partition the recordings into four quartiles according to the rank position of the ground-truth trajectory after the ranking stage. Quartile Q1 corresponds to the recordings with the best inference performance (lowest ground-truth ranks), while Q4 contains the hardest trajectories to discriminate. We then analyze how several recording-level and graph-level characteristics evolve across these quartiles.

a) Structural ambiguity in the transport graph: The clearest relationship with inference performance is observed for graph-structural ambiguity metrics. Figure 4a shows the evolution of edge sharedness across quartiles. This metric measures how frequently the traversed graph edges are reused by different time-compatible candidate trajectories in the public transport network. A strong monotonic increase is observed from Q1 to Q4. While Q1 recordings traverse almost



(a) Sharedness (b) Closeness (c) Duration (d) Curve amplitude

Fig. 4: Influence of recording-level and graph-structural characteristics on inference performance across ground-truth rank quartiles. Q1 corresponds to the easiest trajectories to infer, while Q4 contains the hardest trajectories.

exclusively unique or weakly shared edges (median sharedness $= 0$), Q4 trajectories are dominated by highly reused corridors with median sharedness values exceeding 480.

From a privacy perspective, this result suggests that journeys traversing structurally distinctive or weakly shared parts of the transport network are substantially more vulnerable to trajectory inference attacks than trajectories occurring on dense and highly reused urban corridors. Conversely, highly interconnected regions naturally create ambiguity between multiple plausible candidate paths and therefore provide a degree of structural privacy protection.

A similar trend is observed for node closeness centrality in Figure 4b. Recordings belonging to Q4 tend to traverse stations that are more centrally located within the transport graph and, on average, lie at shorter network distances from the rest of the system, whereas Q1 trajectories are associated with slightly more peripheral areas. Although the effect is less pronounced than for edge sharedness, it further supports the hypothesis that inference performance is governed primarily by the structural uniqueness of traversed transport corridors.

b) Recording characteristics: Figure 4c analyzes the relationship between recording duration and inference performance. Although longer recordings may intuitively be expected to leak more trajectory information, the results do not reveal a clear monotonic relationship between recording duration and ranking quality. While Q2 recordings exhibit the largest durations and variability, Q1 and Q4 both contain comparatively shorter recordings. In particular, Q1 recordings have a median duration of only $100\,\mathrm{s}$, whereas Q4 recordings exhibit a median duration of $261\,\mathrm{s}$. Overall, the results suggest that recording length alone is insufficient to explain successful trajectory inference or increased privacy leakage.

Figure 4d reports the cumulative railway curve amplitude derived from the public transport graph along the ground-truth trajectories. Counter-intuitively, trajectories with larger cumulative curve amplitudes do not lead to improved inference performance. Indeed, Q3 and Q4 recordings exhibit substantially larger cumulative curve amplitudes than Q1 recordings. One possible explanation is that trajectories containing many turns and curves create more opportunities for disagreement between sensor-derived and public-data-derived curve sequences. Therefore, richer geometric structure may amplify alignment errors rather than improve trajectory discrimination. These results suggest that recording-level characteristics become less reliable for trajectory inference in realistic sensing conditions.

RQ2: *Inference performance is governed primarily by the structural ambiguity of the transport network rather than by imperfect recording-level characteristics. Weakly shared and structurally distinctive corridors are easier to infer, whereas highly interconnected urban regions naturally create ambiguity between plausible trajectories.*

D. Ablation study and cross-city generalization

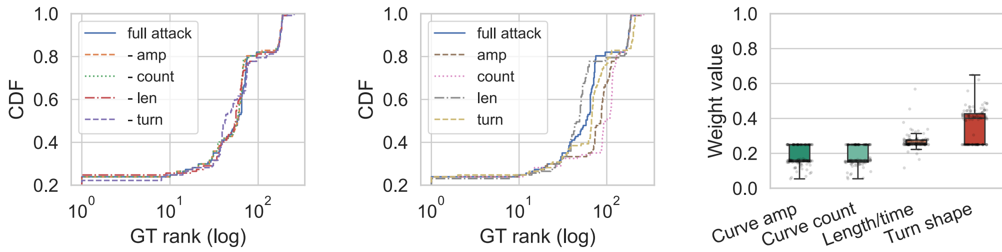
Finally, we analyze the contribution of the different plausibility scoring features through an ablation study and investigate the transferability of the proposed methodology to a different public transport network.

a) Feature ablation: Figure 5a compares the ranking performance obtained with the complete framework against variants where individual feature channels are removed, while Figure 5b evaluates configurations relying on a single feature.

Overall, no individual feature alone is sufficient for robust trajectory inference under realistic sensing conditions. Instead, inference performance emerges from the combination of weak but complementary structural signals. Removing individual features only moderately degrades performance, indicating a methodology relatively robust to partial feature loss.

Among the standalone channels, the normalized duration/length pattern achieves the strongest individual performance, suggesting that coarse temporal structure and interstation geometry already leak substantial trajectory information. In contrast, curve-count features provide the weakest discrimination, while cumulative curve amplitudes and turn-shape sequences yield intermediate performance.

Interestingly, removing geometric motion features from the complete model only slightly affects the ranking distribution. This observation is consistent with the explainability analysis of Section V-C: gyroscope-derived features remain noisy and partially ambiguous under realistic sensing conditions, particularly in dense urban corridors where multiple trajectories share similar geometric patterns. Nevertheless, combining multiple



(a) Feature removal (b) Single features (c) Feature weights

Fig. 5: Ablation study and adaptive feature weighting on our Berlin dataset.

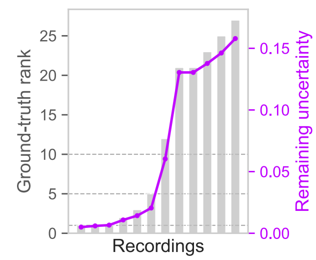


Fig. 6: Inference performance in Stuttgart.

partially informative features consistently improves robustness and reduces extreme ranking failures compared to single-feature configurations.

Figure 5c further reports the adaptive feature weights estimated across recordings. Overall, the turn-shape and normalized duration/length channels receive the largest average weights, whereas curve-count and cumulative curve-amplitude features are consistently downweighted. This confirms that the discriminative value of motion features strongly depends on the structural properties of the underlying trajectory.

b) Cross-city generalization: To evaluate transferability, we additionally apply the complete inference pipeline to the Stuttgart public transport network without modifying the feature extraction or ranking procedures.

Table III summarizes both the candidate-space reduction and the trajectory inference performance obtained in Stuttgart compared to our Berlin dataset. While schedule constraints substantially reduce the number of feasible trajectories in both networks, inference performance is considerably stronger in Stuttgart. For example, the median number of time-compatible candidates decreases from 329 in Berlin to 171 in Stuttgart, while the median ground-truth rank decreases from 87 to 8.5 and the median remaining uncertainty from 25.1% to 4.0%. Similarly, Top-10 accuracy increases from 28.5% to 50.0%, and Top-25 accuracy rises from 31.9% to 91.7%.

TABLE III: Summary of evaluation results.

	Berlin		Stuttgart
Metric	**Attack**	**Random**	**Attack**
Median same-day candidates	7989		4089
Median time-compatible candidates	329		171
Top-1 accuracy	26.4%	0.7%	25.0%
Top-10 accuracy	28.5%	4.9%	50.0%
Top-25 accuracy	31.9%	12.5%	91.7%
Median GT rank	87.0	137.5	8.5
Median uncertainty	25.1%	40.3%	4.0%

Figure 6 further illustrates this behavior by relating ground-truth rank to remaining trajectory uncertainty across Stuttgart recordings. Ground-truth ranks remain below 25 for most recordings, whereas Berlin recordings occasionally exceed ranks above 150. Likewise, the remaining uncertainty never exceeds approximately 15% in Stuttgart, while substantially larger values are observed in Berlin.

These results are consistent with the explainability analysis of Section V-C. While Stuttgart exhibits a smaller candidate space after schedule filtering, this reduction alone does not explain the substantial improvement in inference performance. Rather, Stuttgart constitutes a smaller and structurally less complex transport network with fewer highly shared corridors and fewer structurally similar candidate trajectories. Consequently, motion sensor observations become more discriminative and allow stronger uncertainty reduction. Beyond demonstrating transferability across transport systems, these findings suggest that privacy risks associated with motion sensor observations may substantially increase in smaller or less structurally ambiguous public transport networks.

RQ3: *Trajectory inference performance emerges from the combination of multiple weak structural signals and generalizes across transport networks without reference-based calibration. Moreover, smaller and less structurally ambiguous networks substantially increase inference capability and potentially amplify mobility privacy risks.*

VI. Practical Implications

Our results suggest that the privacy risks associated with motion sensor observations are more nuanced than previously suggested by supervised trajectory inference studies. Under realistic sensing conditions without network-specific calibration, exact trajectory reconstruction becomes substantially more difficult due to sensing noise, uncontrolled device handling, and uncertainty in public transport information.

Nevertheless, the proposed methodology still achieves substantial uncertainty reduction using only coarse motion observations and publicly available transport data. Our analysis shows that inference performance is governed primarily by the structural properties of the transport network: weakly shared and structurally distinctive corridors facilitate trajectory inference, whereas dense and highly interconnected regions naturally create ambiguity. Moreover, unlike prior supervised attacks tailored to the considered target network, the proposed methodology remains naturally transferable across transport systems since it relies exclusively on publicly available transport information. The cross-city evaluation further suggests that smaller and less structurally ambiguous networks may amplify mobility privacy leakage.

These observations have practical implications for both users and transport-system designers. From a user perspective, they highlight that even coarse motion sensor observations may leak meaningful mobility information under favorable network conditions despite the absence of explicit location permissions. From a system-design perspective, our findings suggest that the structural organization of transport networks itself influences the difficulty of trajectory inference and therefore contributes to mobility privacy exposure.

Limitations: Our evaluation assumes perfect transportation mode detection, meaning that public transport segments are identified without error before trajectory inference. Although modern transportation mode detection systems can achieve high accuracy, they remain imperfect in practice and may introduce additional errors that propagate to subsequent inference stages. Consequently, the results presented here should be interpreted as an upper bound on the privacy risk achievable under a reference-free attacker model. Moreover, our evaluation is limited to rail-based transport in two German cities with different network sizes and structures; further studies across cities of varying sizes and complexity, as well as buses and other transport systems, are needed to assess broader generalizability. Finally, several inference parameters remain fixed throughout the evaluation, and the computational scalability of candidate generation and matching on substantially larger networks remains to be systematically assessed.

Recommendations: The first countermeasure to mitigate the threat is to protect access to motion sensors with permissions, making it easier for users to identify suspicious apps. Second, reducing the sampling rate of the sensors makes it more difficult to extract detailed features. However, a balance needs to be found to allow benign applications.

VII. Conclusion

This work revisits motion-sensor-based trajectory inference in public transport under realistic assumptions: no prior data collection, supervised training, or controlled sensing conditions. We propose a reference-free framework combining smartphone motion-sensor observations with publicly available transport information and evaluate it on real-world datasets collected in Berlin and Stuttgart.

Our results show that the high reconstruction accuracy reported in prior reference-based studies does not carry over to realistic sensing conditions. Although exact reconstruction remains limited by sensing noise and structural ambiguity, motion-sensor observations still substantially reduce uncertainty and frequently isolate a small set of plausible trajectories. Inference performance is primarily governed by transport-network structure, with smaller and less ambiguous networks increasing privacy-leakage risk.

Overall, these findings provide a more realistic assessment of motion-sensor privacy risks and highlight the role of transport-network structure in mobility-privacy exposure.

Acknowledgments

This work was supported by the European Research Council (ERC) under the consolidator grant MALFOY (101043410). The work of Anne Josiane Kouam was supported through the Mob Sci-Dat Factory project (ANR-23-PEMO-0004) under the France 2030 program.

References

[1] Android, "Motion sensors," https://developer.android.com/develop/sensors-and-location/sensors/sensors_motion, 2024, accessed: 2026-04.

[2] S. Dernbach, B. Das, N. Krishnan, B. Thomas, and D. Cook, "Simple and complex activity recognition through smart phones," in *8th International Conference on Intelligent Environments, IE 2012*, pp. 214–221.

[3] J. Hua, Z. Shen, and S. Zhong, "We can track you if you take the metro: Tracking metro riders using accelerometers on smartphones," *Trans. Info. For. Sec.*, vol. 12, no. 2, Feb. 2017.

[4] K. Sankaran, M. Zhu, X. F. Guo, A. L. Ananda, M. C. Chan, and L.-S. Peh, "Using mobile phone barometer for low-power transportation context detection," in *SenSys*. ACM, 2014.

[5] S. Czybik, P. Hanfeld, and K. Rieck, "On the same track: Privacy leaks in electric rail transport via magnetic fields," in *33rd ACM Conference on Computer and Communications Security (CCS)*, 2026.

[6] C. S. Fandino, A. J. Kouam, and K. Rieck, "Tiny sensors, big threats: Assessing motion sensor-based fingerprinting in mobile systems," in *MSWiM*, 2025, pp. 286–295.

[7] K. A. Nguyen, Y. Wang, G. Li, Z. Luo, and C. Watkins, "Realtime tracking of passengers on the london underground transport by matching smartphone accelerometer footprints," *Sensors*, vol. 19, no. 19, 2019.

[8] S. Hemminki, P. Nurmi, and S. Tarkoma, "Accelerometer-based transportation mode detection on smartphones," in *SenSys '13*. ACM, 2013.

[9] C. Carpineti, V. Lomonaco, L. Bedogni, M. D. Felice, and L. Bononi, "Custom dual transportation mode detection by smartphone devices exploiting sensor diversity," in *IEEE PerCom Workshops*, 2018.

[10] ——, "US-TMD: A Multi-Sensor Dataset for Transportation Mode Detection," https://cs.unibo.it/projects/us-tm2017/index.html, 2017, accessed: 2026-04-28.

[11] M. Erdelić, T. Erdelić, and T. Carić, "Dataset for multimodal transport analytics of smartphone users – collecty," *Data in Brief*, vol. 50, p. 109481, 2023.

[12] M. A. Igartua, G. C. Ibáñez, and M. G. Estañol, "Mobilitapp – mobility sensor dataset," 2024, dataset.

[13] A. Skretting, T.-M. Grønli, and R. R. Mukkamala, "Feature importance from novel dataset for on-device transport mode detection," *Journal of Intelligent Transportation Systems*, 2025.

[14] P. Wang and Y. Jiang, "Dftrans: Dual frequency temporal attention mechanism-based transportation mode detection," *Sensors*, vol. 22, no. 21, p. 8499, 2022.

[15] ——, "Transportation mode detection using temporal convolutional networks based on sensors integrated into smartphones," *Sensors*, vol. 22, no. 17, p. 6712, 2022.

[16] X. Xu, "Automatic classification of transportation modes using smartphone sensors: Addressing imbalanced data and enhancing training with focal loss and artificial bee colony algorithm," *Journal of Optics*, vol. 53, no. 5, pp. 4656–4670, 2024.

[17] J. Han, E. Owusu, L. T. Nguyen, A. Perrig, and J. Zhang, "Accomplice: Location inference using accelerometers on smartphones," in *IEEE COMSNETS*, 2012.

[18] J. L. Kröger, P. Raschke, and T. R. Bhuiyan, "Privacy implications of accelerometer data: a review of possible inferences," in *ACM ICCSP*, 2019.

[19] T. Stockx, B. Hecht, and J. Schöning, "Subwayps: towards smartphone positioning in underground public transportation systems," in *SIGSPATIAL '14*. ACM, 2014, p. 93–102.

[20] H. Ye, T. Gu, X. Tao, and J. Lu, "Crowdsourced smartphone sensing for localization in metro trains," in *WoWMoM*, 2014.

[21] Apple Inc., "Control access to hardware features on iphone," https://support.apple.com/en-gb/guide/iphone/iph168c4bbd5/26/ios/26, 2026, accessed: 2026-04.

[22] K. T. H. Choi, "Sensor logger," Android. [Online]. Available: https://www.tszheichoi.com/sensorlogger