

Article

Calibration of SUMO Car-Following and Lane-Change Models for Heterogeneous Traffic on Egyptian Two-Lane Two-Way Roads

Usama Elrawy Shahdah ¹, Mahmoud Owais ^{2,3,*}, Sherif M. El-Badawy ¹, Ahmed Shoaeb ⁴ and Sania Elagamy ¹

¹ Public Works Engineering Department, Mansoura University, Mansoura 35516, Egypt; usama.shahdah@mans.edu.eg (U.E.S.); sbadawy@mans.edu.eg (S.M.E.-B.); saniareyad@mans.edu.eg (S.E.)

² Civil Engineering Department, Assiut University, Assiut 71515, Egypt

³ Civil Engineering Department, Sphinx University, New Assiut City 71684, Egypt

⁴ Civil Engineering Department, Nile Academy for Engineering and Technology, Mansoura 35511, Egypt; ahmed_soab@nilehi.edu.eg

* Correspondence: maowais@yahoo.com or maowais@aun.edu.eg; Tel.: +20-100-336-7388

Abstract

Traffic simulation models require calibration to capture local driving behavior, particularly in developing countries where conditions differ from Western defaults. This study calibrates three car-following models—Krauss, Wiedemann 99 (W99), and Intelligent Driver Model (IDM)—with lane-change parameters for heterogeneous traffic on six two-lane two-way roads in Egypt's Nile Delta. A sensitivity analysis reduced eight lane-change parameters to three critical ones: lcOpposite, lcSpeedGain, and lcStrategic. A Genetic Algorithm then optimized car-following and lane-change parameters jointly across multiple seeds and generations. The best configurations achieved mean absolute percentage errors of 8.79% (IDM), 8.94% (W99), and 10.56% (Krauss), indicating that IDM and W99 performed similarly under the original four-metric validation, while Krauss was less accurate overall and exceeded the overtaking-rate threshold. suggesting calibration methodology matters more than model choice. Minimum standstill gap (minGap) treatment varied by model: freeing it improved IDM and Krauss performance, while fixing it at 1.50 m benefited W99. Constraining minGap in IDM increased error by 2.06 percentage points and altered parameter sensitivity, with lcStrategic becoming dominant and tau showing reduced sensitivity. Notably, SUMO's uncalibrated Krauss defaults achieved 10.11% error under understaturated conditions, though this should not extend to congested scenarios. Simulated standard deviations underestimated observed variability, primarily due to aggregation mismatches rather than model inadequacy.

Keywords: SUMO calibration; genetic algorithm; car-following models; lane-change calibration; two-lane two-way roads; Egypt; microscopic traffic simulation; Wiedemann 99 (W99) Model; Krauss model; Intelligent Driver Model (IDM)



Academic Editor: Mariano Perneti

Received: 1 June 2026

Revised: 30 June 2026

Accepted: 6 July 2026

Published: 9 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Traffic simulation models are standard tools for traffic management, infrastructure planning, and operational evaluation, allowing researchers to predict traffic patterns without costly or disruptive field trials [1]. Among these, microscopic simulation models are particularly valuable because they replicate individual vehicle behaviors—such as acceleration, deceleration, and lane changing, with high temporal and spatial fidelity [2]. However, the predictive accuracy of such models depends on calibration, i.e., adjusting model parameters so that simulated outputs closely approximate observed

conditions [3]. Without proper calibration, simulation results may diverge substantially from field reality [1]. This problem is more pronounced in developing countries, where driver behavior, fleet composition, and enforcement practices differ from the Western conditions assumed by most simulation software [4].

1.1. Background and Motivation

In Egypt, two-lane two-way (TLTW) roads make up about 75% of the paved rural highway network [5]. Despite their prevalence, only a limited number of studies have examined traffic operations on these facilities [5–8]. The traffic stream on Egyptian TLTW roads is heterogeneous: pickup trucks account for 39–42% of the fleet, minibuses for 20–23%, private cars and taxis for 16–22%, trucks for 9–11%, motorcycles for 2–6%, and buses for 3–8% [9]. This composition—dominated by light commercial and paratransit vehicles rather than private automobiles—differs from what most microscopic simulators assume by default, so model calibration is necessary for credible simulation of Egyptian rural traffic.

The relevance of this work to sustainable transportation planning is threefold. First, accurate microscopic simulation of TLTW traffic enables evidence-based evaluation of traffic management strategies—such as overtaking-zone design, speed management, and shoulder utilization—that directly affect road safety, a critical component of sustainable mobility (SDG 3.9 and SDG 11.2). Second, by calibrating SUMO’s car-following and lane-change models for Egyptian conditions, the study removes a key barrier to using open-source simulation tools for infrastructure planning in developing countries, where costly proprietary software may be unavailable and where default parameters from Western contexts produce unreliable predictions. Third, the calibrated models support more accurate estimation of fuel consumption and emissions at the network level, since vehicle acceleration, deceleration, and lane-change behavior—directly governed by the parameters calibrated here—are primary determinants of tailpipe emissions in microscopic emission models. In summary, the calibrated simulation framework contributes to sustainable transportation planning by providing a validated, low-cost tool for evaluating interventions that improve safety, reduce environmental impact, and optimize the use of existing road infrastructure in data-scarce developing-country contexts.

1.2. SUMO Microscopic Traffic Simulation

SUMO (Simulation of Urban Mobility) is an open-source microscopic traffic simulator that has gained widespread international adoption [2]. It offers multiple car-following (CF) models, three of which were evaluated in the present study:

1. Krauss Model. The default CF model in SUMO, which computes a safe speed for each following vehicle at every simulation step to prevent collisions under emergency braking [10,11]. Its key parameters include desired speed, driver reaction time, and comfortable deceleration [12].
2. Wiedemann 99 (W99) Model. A psychological-perceptual model derived from Wiedemann’s [13] original framework and widely implemented in VISSIM. It governs standstill distance, desired headway, perception thresholds, and acceleration behavior through ten coefficients {cc0, . . . , cc9} [12,14].
3. Intelligent Driver Model (IDM). A model that formulates acceleration as a continuous, differentiable function of velocity, gap, and relative speed with respect to the leading vehicle [15,16]. Its parameters include desired velocity, minimum spacing, desired time headway, maximum acceleration, and comfortable deceleration [17].

1.3. Research Objectives

This study pursues the following objectives:

1. Review the literature on traffic simulation calibration, with emphasis on Genetic Algorithm (GA)–based approaches and SUMO-specific studies;
2. Document data-collection procedures and traffic characteristics at six Egyptian TLTW study sites;
3. Develop and apply a two-phase calibration framework comprising a lane-change (LC) parameter sensitivity sweep followed by joint CF–LC GA optimization;
4. Formulate practical guidelines for multi-metric calibration incorporating speed, travel time, headway, and overtaking rate;
5. Identify research gaps and provide recommendations for traffic simulation studies in developing-country contexts.

Beyond these objectives, the present study contributes four methodological elements that are tailored to TLTW roads in a data-scarce developing-country context. First, the calibration uses a two-phase procedure in which an LC2013 lane-change sensitivity sweep precedes joint CF–LC GA optimization, reducing the active lane-change parameter set before full calibration. Second, the study compares free versus fixed minGap treatment across three structurally different CF models under the same sites, metrics, and GA settings. Third, the overtaking-rate target is computed using a vehicle-ID-based round-trip protocol, which avoids double-counting the departure and return lane-change events that occur during one completed opposite-lane overtaking maneuver. Fourth, the calibration aggregates six field sites into a single multi-site fitness function, providing a practical framework for developing-region settings where detailed trajectory data are limited. These contributions are incremental rather than a new optimization algorithm, but they address measurement and calibration issues that are especially important for TLTW overtaking simulation.

1.4. Paper Organization

The paper proceeds as follows. Section 2 presents a comprehensive literature review. Section 3 describes the study area, data collection procedures, simulation framework, and GA calibration methodology. Section 4 presents the results and discussion. Section 5 concludes the study, discusses limitations, and proposes directions for future research.

2. Literature Review

Calibration of microscopic traffic simulation models has been an active research domain for more than three decades. The following review is organized thematically, with emphasis on methodological frameworks, GA applications, SUMO-specific studies, CF model calibration, and developing-country contexts.

2.1. Calibration Fundamentals and Methodological Frameworks

Calibration involves several interrelated decisions: which parameters to adjust, their permissible ranges, and the target outputs for comparison [18]. From a methodological standpoint, calibration tasks may be classified into three categories: capacity calibration, route/demand calibration, and trajectory-based calibration [12,19].

Capacity calibration aims to match aggregate flow characteristics—volume, speed, and density—using data from loop detectors, traffic counts, and travel-time surveys [20]. Although computationally efficient, this approach may fail to capture the individual-vehicle dynamics required for emissions or safety analyses [12]. Route and demand calibration targets route-choice and origin–destination accuracy. SUMO provides built-in calibrators that adjust flows and speeds using induction-loop data, as well as a REROUTER tool that modifies vehicle routes in response to prevailing traffic states [2]. Trajectory-based calibra-

tion represents the most detailed tier, reproducing individual acceleration, deceleration, and car-following profiles. This approach is essential for safety, energy, and emissions modeling [3,17].

Punzo and Ciuffo [21] demonstrated that parameter sensitivity depends on prevailing traffic conditions: coefficients that perform well under free-flow conditions may prove inadequate under congestion. Lownes and Machemehl [22] observed that specific VISSIM parameters exert disproportionate influence on capacity and advocated sensitivity screening prior to full calibration—as Liang et al. [23] later confirmed for safety-related outputs. Ge and Menendez [24] developed the quasi one-at-a-time elementary effects (quasi-OTEE) method, a qualitative variant of Morris Elementary Effects, to identify influential parameters in computationally expensive simulation models. Antoniou et al. [20] proposed a five-stage framework covering data collection, filtering, quality assessment, calibration, and validation that remains widely adopted. Yet the MULTITUDE project found that two-thirds of published simulation studies lack empirical data entirely [19].

2.2. Genetic Algorithm Applications in Traffic Simulation Calibration

Among the 52 calibration studies reviewed herein, 31 (60%) employed Genetic Algorithms (GAs). GAs are popular in this domain because of their ability to navigate nonlinear, multimodal response surfaces without gradient information—which suits traffic simulation problems where parameter-to-output relationships are nonlinear [25].

2.2.1. Early Applications

Cheu et al. [26] were among the first to apply a GA to microscopic simulation, calibrating 12 FRESIM parameters for Singapore expressways. Park and Schneeberger [27] formulated a nine-step GA procedure for VISSIM at isolated signalized intersections, and Park and Qi [28] extended this work by employing average travel time as the calibration criterion for eight parameters. Collectively, these studies established a canonical workflow: sensitivity analysis, followed by GA optimization, followed by independent validation.

2.2.2. Multi-Objective GA Calibration

Optimizing a single metric—for instance, minimizing travel-time error—can degrade accuracy on other performance dimensions [18]. Multi-objective GAs address this limitation by simultaneously optimizing several potentially conflicting objectives, yielding a Pareto front of non-dominated solutions. Cascan et al. [18] proposed a two-stage procedure in which parameters are first screened for sensitivity, after which a multi-objective GA operates exclusively on the influential subset, achieving satisfactory accuracy for both flow and safety metrics. Cobos et al. [29] adapted the Multi-Objective Global-Best Harmony Search algorithm, incorporating statistical tests to verify solution stability. Chiappone et al. [30] combined GA optimization with speed–density fundamental diagrams and demonstrated that automated calibration can accurately replicate observed macroscopic relationships.

2.2.3. Recent Advances

Afshari et al. [1] calibrated microscopic models using high-resolution trajectories from 338 connected vehicles (Wejo dataset) and reported GA-induced reductions in mean trajectory error of 14.19% during calibration and 32.68% during validation. Muhan et al. [25] calibrated four VISSIM parameters at a Beijing intersection using queue length and travel time, with weighting coefficients to balance the two objectives. Hale et al. [31] compared directed brute-force search with Simultaneous Perturbation Stochastic Approximation (SPSA) and found that the former yielded superior solutions at substantially higher computational cost.

2.2.4. Comparative Algorithmic Studies

Several studies have compared alternative optimization strategies. Lee and Ozbay [32] introduced enhanced SPSSA (E-SPSSA) for faster microscopic calibration convergence. Antoniou et al. [20] refined Weighted SPSSA (W-SPSSA) weight-matrix approximations, and Oh et al. [33] extended SPSSA to multimodal demand calibration for large-scale networks. Paz et al. [34] achieved superior convergence with a memetic algorithm (GA fused with local search). Hou and Lee [35] showed that parallel multi-threaded optimization reduces GA runtimes without sacrificing quality. Yu and Fan [36] found GA robust across test networks, though algorithmic ranking varied with complexity. Guastella et al. [37] proposed local optimization using traffic-count data, and Zhang et al. [38] applied a sequential multi-stage GA to multi-intersection networks with faster convergence than single-stage approaches.

2.2.5. Alternative Calibration Methods

Beyond GAs, several strategies address computational cost or expand calibration scope. Cipriani et al. [39] approximated gradients for time-varying origin–destination matrices; Cheng et al. [40] constructed surrogate models to reduce the required number of simulation runs; and Patwary et al. [41] extended metamodels to million-parameter networks. Bayesian optimization and Bayesian calibration approaches have also been applied to transportation simulation and car-following calibration [42–44] uncertainty-aware alternatives to single point estimates when formulated probabilistically. Machine-learning approaches have been used either to support simulator calibration with optimization [45] or to estimate microscopic simulation parameters from input–output data [46]. Owais et al. [47] provide a related simulation-based traffic-operation example from Saudi Arabia, illustrating the use of simulation for operational decision support in the region.

2.3. SUMO-Specific Calibration Studies

2.3.1. Calibration Fundamentals

Lopez et al. [2] provided a comprehensive overview of SUMO calibration requirements. Typical validation errors for microscopic behavior (gaps, speeds, accelerations) range between 10% and 20%, whereas at the network level, more than 85% of simulated counts should deviate less than 15% from observed values. Although SUMO offers built-in tools—such as a calibrator for dynamic flow and speed adaptation from induction-loop data and a REROUTER for route modification—Lopez et al. [2] emphasized that no universal calibration solution exists.

Although SUMO was originally developed as “Simulation of Urban MObility,” the CF and LC model implementations used here are not intrinsically urban-only. Krauss, W99, and IDM describe longitudinal following through gap, speed, and acceleration relationships that can be calibrated for different road contexts. Prior SUMO studies (refer to Table 1) have calibrated or applied SUMO in urban and regional contexts [4,19,48], but those applications do not remove the need for context-specific calibration on Egyptian rural TLTW roads. The present study, therefore, applies SUMO to rural TLTW segments with calibration targets adapted to the rural setting: speed, travel time, headway, and successful overtaking rate. SUMO’s urban origins remain relevant as a limitation because shoulder use and informal lateral offset are not automatically represented in the lane-based model.

2.3.2. Heterogeneous Traffic

Chowdhury and Chakraborty [4] calibrated SUMO for the Powerhouse intersection in Khulna, Bangladesh, minimizing the root-mean-square error (RMSE) of queue length and travel time. Their calibrated sigma (0.3) and tau (1.4) values, together with scenario analyses of channelization and signal management, are pertinent to the Egyptian con-

text, as both countries share mixed vehicle fleets, relaxed lane discipline, and aggressive driving behavior.

2.3.3. Diverse Data Sources

Flitsch et al. [19] integrated floating car data, Bluetooth traces, permanent counting stations, and test-ride travel times to calibrate SUMO in Upper Austria, proposing a data-selection framework based on quality, coverage, and update frequency. Harth et al. [48] automated the ingestion of traffic counts and signal timings into SUMO, enabling calibration for arbitrary road networks. Stang and Bogenberger [49] employed GPS traces to calibrate urban SUMO models, complementing loop-detector data in areas with sparse fixed-sensor infrastructure. Vishnoi et al. [50] demonstrated that road-speed data from probe vehicles can substitute for conventional origin–demand surveys in urban traffic simulators—an approach that works well when data are scarce. Across these studies, data availability remained the primary bottleneck, with many simulation projects continuing to rely on assumptions rather than empirical observations.

2.3.4. Evolutionary Algorithms for SUMO

Kappenberger and Stuckenschmidt [51] employed Covariance Matrix Adaptation Evolution Strategy (CMA-ES) to calibrate route probabilities in the Mannheim SUMO Traffic Model under data-scarce conditions. The method improved accuracy for both 3-h and 24-h scenarios relative to baseline heuristics. Siam et al. [52] introduced VISICAL, a heuristic calibration platform that supports multiple optimization algorithms and provides a graphical user interface for practitioners lacking programming expertise.

Table 1. Summary of selected prior SUMO calibration studies.

Study	Road Type	CF Model (s)	Calibration/ Optimization	Main Metrics	LC Parameters Calibrated?	Context
Chowdhury and Chakraborty [4]	Urban intersection	Krauss	Manual/calibrated scenario analysis	Queue length, travel time	No	Bangladesh
Flitsch et al. [19]	Regional/ highway network	SUMO scenario calibration/data integration	Data-driven calibration	Flow, speed, travel time	No	Austria
Schrader et al. [12]	Urban corridor	Krauss, W99, IDM	SUMO-in-the- loop calibration	Headway, speed, acceleration	No	USA
Kappenberger and Stuckenschmidt [51]	Urban scenario	SUMO route/scenario calibration under data scarcity	CMA- ES/evolutionary optimization	Counts, route probabilities	No	Germany
Present study	Rural TLTW segments	Krauss, W99, IDM	GA with LC sensitivity sweep	Speed, travel time, headway, overtaking	Yes	Egypt

2.3.5. SUMO-in-the-Loop Calibration

Schrader et al. [12] embedded SUMO within a real-time calibration loop, feeding it roadside radar trajectories from a 1.5 km urban corridor. They calibrated the Krauss, IDM, and W99 models and found that default SUMO parameters produced shorter headways than observed; calibration attenuated this bias. W99 performed best at the high end of energy consumption despite exhibiting unrealistic acceleration profiles, whereas IDM with

default parameters most closely matched observed acceleration. Their main conclusion, that calibrating within the simulator matters more than solving the mathematical model in isolation, directly motivates the SUMO-in-the-loop approach adopted in the present study.

2.3.6. Extended Car-Following Models

Salles et al. [53] extended SUMO's IDM to improve reproduction of acceleration from standstill, validating drive-off trajectories against aerial measurements. The improvement was most pronounced at signalized approaches. Kathes et al. [54] calibrated W99 for bicycles, yielding reduced standstill distances and distinct perception thresholds that facilitated more realistic bicycle–car interaction modeling.

2.4. Car-Following Model Calibration

2.4.1. Intelligent Driver Model

The IDM defines acceleration as a function of velocity, gap, and relative speed. Its parameters include desired velocity (v_0), minimum spacing (s_0), desired time headway (T), maximum acceleration (a), comfortable deceleration (b), and acceleration exponent (δ , default 4) [15,16]. Kesting and Treiber [17] calibrated the IDM using radar trajectory data and a GA, obtaining errors of 11–29%, consistent with prior work. They further noted that IDM parameters were more robust to changes in the optimization criterion than the Velocity Difference Model. C. Zhang and Sun [44] applied Bayesian calibration to the IDM, producing probabilistic parameter estimates that accommodate noisy or limited data more effectively than deterministic methods. Punzo et al. [3] conducted extensive multi-model experiments and recommended RMSE of spacing over RMSE of velocity, finding the latter inferior in 93.5% of cases.

2.4.2. Wiedemann 99

The W99 model employs ten coefficients $\{cc_0, \dots, cc_9\}$ governing standstill distance, headway, perception thresholds, and acceleration [12]. Chaudhari et al. [14] calibrated W99 for mixed Indian traffic and found that calibrated thresholds for closing, following, and emergency braking differ between two-wheelers and cars. They further observed that strict versus staggered car-following behavior required separate calibration. Durrani et al. [55] demonstrated that calibrating Wiedemann models with mixed vehicle-pair interactions improves accuracy relative to homogeneous-pair assumptions.

2.4.3. Krauss Model

The Krauss model is SUMO's default CF formulation. It calculates a safe speed from the net gap, reaction time, and braking capability [12]. Schrader et al. [12] calibrated Krauss parameters— τ , deceleration, and speed distribution—using roadside radar data; the calibration eliminated systematic bias in headway distributions.

2.4.4. Comparative Studies

Pourabdollah et al. [56] compared the IDM, Wiedemann, and Krauss models using real-world driving datasets and found that IDM performed best on average, although the optimal model depends on the specific application and data availability. Kong et al. [57] addressed scalable car-following model calibration using active simulation-based inference, illustrating another uncertainty-aware direction for calibration research.

2.5. Developing-Country Contexts

2.5.1. Heterogeneous Traffic

Chowdhury and Chakraborty [4] calibrated SUMO in Bangladesh—a setting comparable to Egypt since both countries exhibit mixed vehicle fleets and aggressive driving

norms. Hoque et al. [58] employed Latin Hypercube Sampling and k-means clustering to calibrate VISSIM for mixed traffic. Chaudhari et al. [14] demonstrated that vehicle-class-specific W99 calibration is essential for Indian mixed traffic, a conclusion that extends to the Egyptian fleet.

2.5.2. Data Scarcity

Kappenberger and Stuckenschmidt [51] calibrated SUMO under sparse-data conditions using CMA-ES—conditions common in developing countries where comprehensive traffic surveys are impractical. Flitsch et al. [19] provided a framework for selecting alternative data sources when conventional loop-detector data are unavailable.

2.5.3. Egyptian Traffic Studies

Hashim and Abdel-Wahed [5] analyzed seven performance measures across eight sites in Minoufiya Governorate. Hashim [6] reported that approximately 40% of drivers on Egyptian TLTW roads exceed the posted 60 km/h speed limit. Shoaeb [59] collected 20 h of video from six TLTW sites, documenting fleet composition, speeds, headways, and overtaking maneuvers—the dataset employed in the present study. Shoaeb et al. [9] utilized the same data to develop an overtaking model (96.45% accuracy for accepted overtakes, 95.90% for rejected), accounting for follower acceleration, leader deceleration, leader length, and speed changes during the maneuver. Previous work on two-lane rural-highway passing behavior has also emphasized that passing maneuvers are directly linked to risk and gap-acceptance decisions [60], supporting the recommendation that future calibration should include aborted or rejected overtakes when reliable field coding is available.

2.6. Goodness-of-Fit Measures

Punzo et al. [3] compared RMSE of spacing, velocity, acceleration, and composite measures across multiple models and datasets. Spacing-based RMSE consistently outperformed velocity-based metrics. Ciuffo et al. [61] compared the Geoffrey E. Havers (GEH) statistic, RMSE, and mean absolute percentage error (MAPE); their findings indicated that no single measure is universally superior, and selection should be guided by calibration objectives. Arafat et al. [62] employed degree of saturation at signalized intersections as a calibration target, validating with Kolmogorov–Smirnov tests on saturation-headway distributions, thereby demonstrating that nontraditional metrics can also serve effectively for validation.

2.7. Research Gaps

This review identifies five gaps that motivate the present study:

1. Few SUMO calibration studies address TLTW roads; the majority focus on urban networks or freeways.
2. Egyptian traffic simulation calibration is virtually nonexistent—only four traffic-performance studies have been published in 35 years, and none addressed simulation calibration.
3. Few studies have systematically compared the Krauss, W99, and IDM models under an identical calibration framework in a developing-country context, particularly for rural TLTW roads.
4. Data scarcity remains a persistent challenge; most calibration methods presuppose abundant data, an assumption that is unrealistic for many developing countries.
5. Multi-model comparisons are scarce for developing-country traffic streams.

This study addresses these gaps through a multi-model SUMO calibration framework applied to six Egyptian TLTW road sites. The framework is not proposed as a new

optimization algorithm; rather, its contribution lies in adapting GA-based calibration to TLTW overtaking conditions through a two-phase LC–CF calibration sequence, explicit comparison of minGap treatment, and a corrected vehicle-ID-based overtaking-count protocol. These elements are intended to improve the credibility of SUMO calibration where opposite-lane passing, mixed fleets, and limited field data make direct transfer of urban or freeway calibration practices problematic.

3. Methodology

3.1. Study Area and Data Collection

3.1.1. Study Area

TLTW roads represent approximately 75% of Egypt's paved rural highway network [5]. Although this proportion has declined somewhat since 2014 owing to national road-development programs, TLTW facilities remain the backbone of rural travel in Egypt.

Only four prior studies have examined traffic on Egyptian TLTW roads. The earliest [63] analyzed headway distributions on the Cairo–Alexandria road. Hashim [6] reported that approximately 40% of drivers exceed the 60 km/h speed limit. Hashim and Abdel-Wahed [5] investigated seven performance measures and three platooning variables, and Hashim et al. [7] examined operating speed on horizontal curves. In addition to Shoaeb's work [8,9,59].

3.1.2. Study Sites

Six TLTW sites in the Nile Delta were selected for data collection, as depicted in Figure 1 [59]. Video recordings were captured during midday hours (10:00–14:00), yielding a total of 20 h of observation. Two sites (S4 and S5) were restricted to two hours each owing to security clearance limitations.

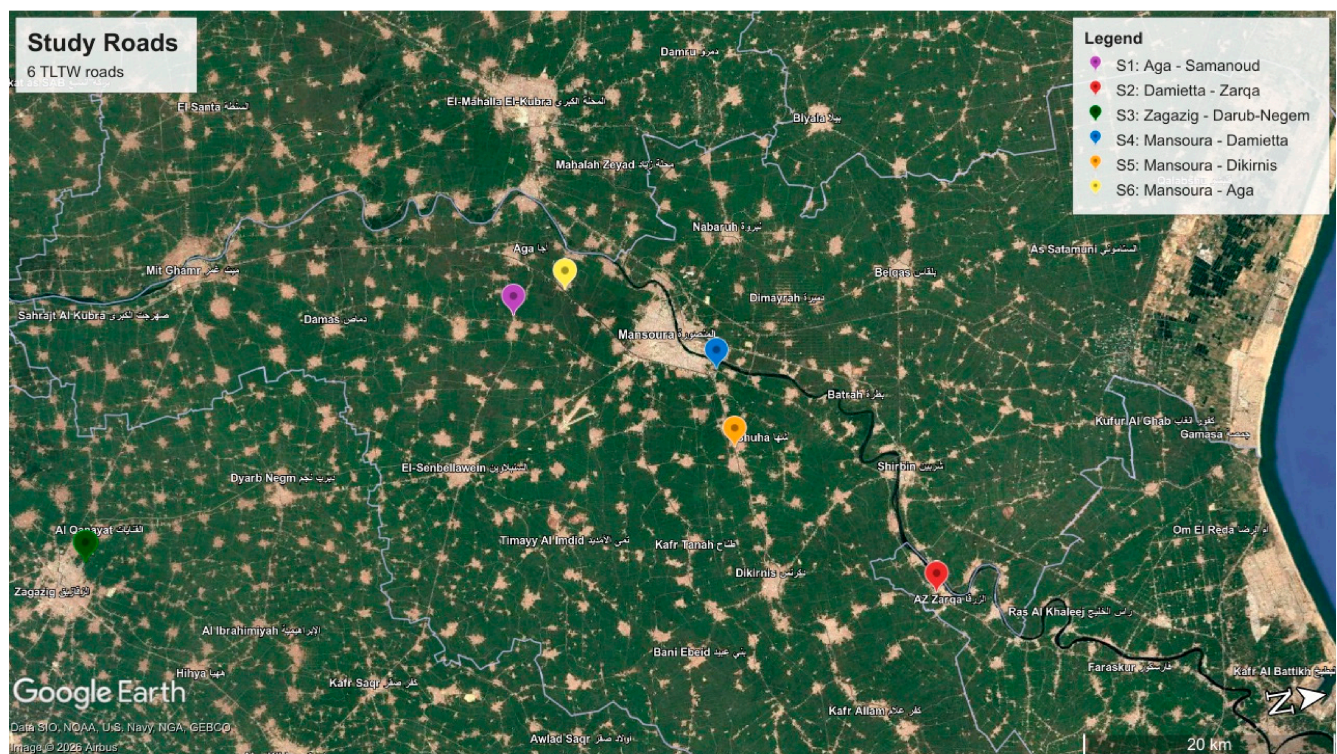


Figure 1. The location of the six selected sites.

1. S1: Aga–Samanoud road, Dakahlia Governorate,
2. S2: Damietta–Zarqa road, Damietta Governorate,

3. S3: Zagaziq–Darub Negem road, Sharqia Governorate,
4. S4: Mansoura–Damietta road, Dakahlia Governorate,
5. S5: Mansoura–Dikirnis road, Dakahlia Governorate,
6. S6: Mansoura–Aga road, Dakahlia Governorate.

All six sites are straight tangent sections with no horizontal or vertical curvature. Each roadway comprises a single 3.0 m lane per direction. Road widths and shoulder dimensions were field-measured with a tape measure and verified against Google Earth imagery. Shoulder dimensions, listed in Table 2, are reported descriptively and were not incorporated into the SUMO network models. Under the Highway Capacity Manual (HCM) 6th edition [64], the sites are classified as Class II TLTW highways; under the 7th edition [65], they are classified as Class I.

Table 2. Geometric properties and data collection details.

Property	S1	S2	S3	S4	S5	S6
Road Name	Aga–Samanoud	Damietta–Zarqa	Zagaziq–Darub Negem	Mansoura–Damietta	Mansoura–Dikirnis	Mansoura–Aga
Governorate	Dakahlia	Damietta	Sharqia	Dakahlia	Dakahlia	Dakahlia
Collection Date	17 June 2019	19 April 2019 & 15 March 2021	15 June 2021	27 June 2021	27 May 2021	7 July 2021
Section Length (m)	280.0	215.0	250.0	375.0	205.0	220.0
Duration (h)	4.0	4.0	4.0	2.0	2.0	4.0
Total Width (m)	9.0	8.0	9.0	8.5	8.0	8.5
Right Shoulder (m)	1.5	1.0	1.0	1.5	1.0	1.5
Left Shoulder (m)	1.0	1.0	1.0	1.0	1.0	1.0

The six study sites were selected to represent straight rural TLTW operating conditions in the Nile Delta while minimizing geometric confounding. All sites are tangent sections; based on field selection criteria and available roadway records, longitudinal grades are mild and were treated as approximately level for simulation purposes. The restriction to straight segments was deliberate because horizontal curvature and steep grades affect speed choice and overtaking behavior in ways that would require additional geometric modeling beyond the scope of this CF–LC calibration. This choice improves internal validity for calibrating following and lane-changing parameters, but it limits generalizability to curved or graded roads.

3.1.3. Data Extraction

Data were extracted manually from video recordings. Each road section was subdivided into 5 m segments. A transparent overlay with reference lines was placed on the video display, and the video-to-real-world scale was calibrated individually for each site.

For every vehicle, two timestamps were recorded per segment: the instant the front bumper entered (t_{start}) and exited (t_{end}) the segment. Temporal resolution was 0.25 s, corresponding to the traversal time of one 5 m segment at typical operating speeds. Segment speed was computed as $v = 5 \text{ m} / \Delta t$. Acceleration was derived from speed changes between consecutive segments. For each vehicle traveling in both directions, the following variables were extracted: travel time, headway, speed, acceleration, deceleration, and overtaking behavior (both successful and aborted).

The 0.25 s timestamp resolution and 5 m segmentation introduce uncertainty into computed speeds and derived accelerations. At a typical operating speed of 60 km/h (16.7 m/s), a vehicle traverses one 5 m segment in approximately 0.30 s; timing uncertainty therefore produces non-negligible speed uncertainty at the segment level. To reduce transcription errors, the same trained operator extracted all observations using the same

segmentation overlay and site-specific scale calibration. No formal inter-rater reliability check was performed, and this is acknowledged as a limitation because measurement noise from manual extraction cannot be fully separated from model error.

At each site, a video camera was positioned at an elevated roadside location where the full study section could be observed. The camera provided an oblique overhead view of the segment, and the 5 m reference segmentation was used to correct for perspective when extracting vehicle positions and timestamps (refer to Figure 2).

3.1.4. Traffic Characteristics

Table 3 presents the measured traffic characteristics across all six sites [59], reported as mean $\pm$ standard deviation (SD). Mean speeds ranged from 53 to 68 km/h. Directional flows, reported for eastbound (EB) and westbound (WB) directions, ranged from 175 to 387 veh/h. Mean travel times varied from 13 to 21 s. Mean headways varied from 9 to 17 s, with high SD values (comparable in magnitude to the means) attributable to fleet heterogeneity. The GA calibration employed four metrics per site: mean speed, mean travel time, mean headway, and successful overtakes per hour, each accompanied by its SD.

Table 3. Traffic characteristics by site (Mean $\pm$ SD).

Site	Mean Speed (km/h)	Mean TT (s)	Mean Headway (s)	Overtakes (/h)	EB Flow (veh/h)	WB Flow (veh/h)
S1	66.48 $\pm$ 11.29	15.09 $\pm$ 2.66	12.10 $\pm$ 9.25	15.50 $\pm$ 2.35	305.0 $\pm$ 6.3	309.0 $\pm$ 5.2
S2	54.34 $\pm$ 11.02	14.25 $\pm$ 3.43	10.44 $\pm$ 9.70	18.50 $\pm$ 3.16	303.0 $\pm$ 6.4	379.0 $\pm$ 8.5
S3	61.04 $\pm$ 12.60	14.75 $\pm$ 3.94	12.05 $\pm$ 9.32	18.63 $\pm$ 1.73	289.2 $\pm$ 9.8	310.2 $\pm$ 9.0
S4	64.40 $\pm$ 12.55	20.96 $\pm$ 3.75	15.13 $\pm$ 8.90	16.00 $\pm$ 1.41	182.0 $\pm$ 7.0	205.0 $\pm$ 7.0
S5	56.81 $\pm$ 8.30	13.00 $\pm$ 2.33	15.19 $\pm$ 6.77	14.00 $\pm$ 1.87	201.0 $\pm$ 5.0	278.0 $\pm$ 7.0
S6	54.13 $\pm$ 9.10	14.64 $\pm$ 3.03	10.18 $\pm$ 8.55	17.50 $\pm$ 1.87	309.2 $\pm$ 6.6	358.5 $\pm$ 6.1

The simulated fleet composition comprised six vehicle classes: pick-ups (39–42%), minibuses (20–23%), private cars and taxis (16–22%), trucks (9–11%), motorcycles (2–6%), and buses (3–8%). Taxis, constituting approximately 2% of the traffic stream, were merged into the private-car category for simulation purposes. This composition differs from the Western defaults built into SUMO.

Over the 20-h observation period, 685 successful and 555 aborted overtakes were recorded across all sites and both directions. The per-site overtaking rates in Table 2 (14.0–18.63 overtakes/h) are direction-specific; the overall average across both directions is approximately 34 overtakes/h, reflecting the sum of opposing-direction maneuvers. Of these, 482 successful and 384 aborted maneuvers were allocated to the calibration dataset (randomly, stratified by site); the remaining 203 successful and 171 aborted maneuvers were reserved for validation. The GA objective function employs the rate of successful overtakes per hour as one of four target metrics.

3.2. Simulation Framework

3.2.1. SUMO Network Development

Network development followed the standard SUMO workflow [2]. Each study site was modeled as a straight TLTW road section. Each direction comprises a single 3.0 m lane; the two opposing lanes are placed adjacently to form the TLTW cross-section. Roadway length corresponds to the measured section length for each site (205–375 m; see Table 1).



S1



S2



S3



S4



S5



S6

Figure 2. Snap shots of the six studied site.

Although the modeled facilities are segments rather than complete networks, microscopic simulation is useful here because the research objective is behavioral calibration rather than network assignment. The overtaking rate depends on stochastic interactions between vehicles traveling in opposite directions, following gaps, and lane-change decisions; these interactions cannot be represented adequately by a simple macroscopic flow relationship. The segment-level SUMO models, therefore, serve as controlled computational experiments for comparing CF and LC parameterizations against observed aggregate outputs. Traffic demand was synthesized from observed vehicle counts and composition data. The demand model incorporates (a) a vehicle-type distribution matching the observed composition, (b) a temporal distribution based on observed hourly volumes, (c) route assignment reflecting observed directional splits, and (d) speed distributions matching observed speed characteristics. Vehicle types were defined in SUMO with appropriate physical dimensions (length, width) and per-class speed restrictions based on Egyptian fleet data. However, the CF parameters (accel, decel, minGap) were applied uniformly across all vehicle classes—a simplifying assumption that departs from SUMO’s class-specific defaults (e.g., passenger cars: $\text{accel} = 2.6 \text{ m/s}^2$, $\text{decel} = 4.5 \text{ m/s}^2$; buses: $\text{accel} = 1.2 \text{ m/s}^2$, $\text{decel} = 4.0 \text{ m/s}^2$; trucks: $\text{accel} = 1.3 \text{ m/s}^2$, $\text{decel} = 4.0 \text{ m/s}^2$; trailers: $\text{accel} = 1.0 \text{ m/s}^2$, $\text{decel} = 4.0 \text{ m/s}^2$; motorcycles: $\text{accel} = 6.0 \text{ m/s}^2$, $\text{decel} = 10.0 \text{ m/s}^2$), with $\text{minGap} = 2.5 \text{ m}$ for all classes. The implications of employing a single parameter set are discussed as a limitation in Section 5.2.2.

The use of a single aggregate CF parameter set across vehicle classes is a simplifying assumption. It was adopted for three reasons. First, the main calibration targets in this study are aggregate site-level metrics—mean speed, travel time, headway, and overtaking rate—rather than class-specific trajectory profiles. Second, calibrating separate CF parameters for each vehicle class would substantially increase the dimensionality of the GA search space and computational burden, especially because each candidate solution is evaluated across six sites and multiple random seeds. Third, although CF parameters are unified, fleet heterogeneity is still represented in SUMO through vehicle-type proportions, vehicle lengths and widths, desired speed distributions, and class-specific maximum speeds. Therefore, the calibrated CF parameters should be interpreted as aggregate mixed-stream coefficients rather than class-specific behavioral parameters. The implications of this simplification are discussed in Section 5.2.2.

The field sites have shoulders of approximately 1.0–1.5 m, and such shoulders may be used informally for evasive maneuvers, slow-vehicle refuge, or partial passing. SUMO’s lane-based representation assumes lane-centered movement unless a sublane model is explicitly configured; therefore, the present simulation does not reproduce informal lateral positioning within a lane or shoulder-assisted passing. As a result, the calibrated LC parameters, especially *lcOpposite* and *lcStrategic*, may partly compensate for unmodeled lateral freedom. This limitation is relevant when interpreting overtaking-rate errors and is revisited in the discussion and limitations sections.

3.2.2. Simulation Parameters

Key global simulation parameters were configured as follows:

- Simulation step size: 0.1 s (recommended for CF calibration).
- Integration scheme: Ballistic (provides smoother trajectories than the Euler scheme).
- Action step length: Decoupled from the simulation step length to emulate realistic driver reaction times.
- Random seed: Multiple seeds were employed to account for stochastic variability.

3.3. Car-Following Models

3.3.1. Krauss Model

The Krauss model computes a safe speed at each time step to prevent collisions, even under emergency braking by the leading vehicle [10,11]. The formulation implemented in SUMO derives the maximum collision-free velocity from the braking-distance safety condition: the follower's braking distance must not exceed the leader's braking distance plus the current gap [11]. SUMO default parameters for the passenger-car class are $\text{accel} = 2.6 \text{ m/s}^2$, $\text{decel} = 4.5 \text{ m/s}^2$, $\text{sigma} = 0.5$ (driver imperfection), $\text{tau} = 1.0 \text{ s}$, and $\text{minGap} = 2.5 \text{ m}$ [66].

The safe speed in the SUMO implementation [11,66] is computed as:

$$v_{\text{safe}}(t) = -b\tau + \sqrt{b^2\tau^2 + v_l^2(t) + 2b \cdot g(t)} \quad (1)$$

The desired speed is then determined as the minimum of the maximum vehicle speed, the safe following speed, and the speed attainable given the vehicle's acceleration capability [11]:

$$v_{\text{des}}(t) = \min(v_{\text{max}}, v_{\text{safe}}(t), v(t - \Delta t) + a \cdot \Delta t) \quad (2)$$

Finally, a stochastic dawdling term is applied to model driver imperfection [11,12]:

$$v(t) = \max(0, v_{\text{des}}(t) - \epsilon \cdot a \cdot \sigma) \quad (3)$$

where v_l denotes the leader speed (m/s), g denotes the net gap (m), τ denotes the driver reaction time (tau, s), b denotes the comfortable deceleration (m/s^2), a denotes the maximum acceleration (m/s^2), σ denotes the driver imperfection parameter (sigma), and $\epsilon \in [0, 1]$ is a uniform random deviate. Equation (1) represents the continuous-time approximation implemented in SUMO, which simplifies the discrete-time braking-distance formulation presented in Krauss et al. [10] and Krauss [11] (Chapter 3) by assuming equal braking capabilities for leader and follower. The calibrated parameters are tau, accel, decel, minGap, sigma, and sigmaStep (which controls the resampling frequency of the imperfection random deviate). The actionStepLength is a global simulation parameter that governs driver decision frequency and was also included in the calibration (see Section 3.4.1).

3.3.2. Wiedemann 99 Model

The W99 model is a psychological-perceptual CF model defined by ten coefficients $\{\text{cc0}, \dots, \text{cc9}\}$. Table 4 compares SUMO default values (drawn from the SUMO source code) with VISSIM defaults reported by Chaudhari et al. [14] and Schrader et al. [12]. Although the two implementations share the same conceptual structure, they differ on several default values—notably cc_2 (8.0 vs. 4.0 m), cc_6 (6.0 vs. 11.44), and cc_8 (2.0 vs. 3.5 m/s^2).

Table 4. W99 default parameter values: SUMO versus VISSIM.

Parameter	Description	SUMO Default	VISSIM Default
cc0	Standstill distance (minGap)	2.50 m (equals vType minGap)	1.50 m
cc1	Desired time headway	1.30 s	0.90 s
cc2	Following variation	8.00 m	4.00 m
cc3	Threshold for entering car-following (s)	−12.00	−8.00
cc4	Negative following threshold (m/s)	−0.25	−0.35
cc5	Positive following threshold (m/s)	0.35	0.35
cc6	Speed dependency of oscillation ($\times 10^{-4} \text{ rad/s}$)	6.00	11.44

Table 4. Cont.

Parameter	Description	SUMO Default	VISSIM Default
cc7	Oscillation acceleration (m/s ²)	0.25	0.25
cc8	Standstill acceleration (m/s ²)	2.00	3.50
cc9	Acceleration at 80 km/h (m/s ²)	1.50	1.50

In SUMO's implementation, cc0 is not an independent parameter; it is permanently equal to the vehicle type's minGap attribute (SUMO source code, module MSCF-Model_W99). For this study, only cc1, cc3, and cc5 were calibrated, based on the pre-calibration sensitivity sweep and the findings of Chaudhari et al. [14]. The remaining seven parameters were held fixed at SUMO defaults.

3.3.3. Intelligent Driver Model

The IDM defines acceleration as a continuous, differentiable function of the ego vehicle's speed, the net gap to the leading vehicle, and the velocity difference (approaching rate) to the leading vehicle [15,16]. SUMO default parameters for the passenger-car class are accel = 2.6 m/s², decel = 4.5 m/s², delta = 4 (acceleration exponent), minGap = 2.5 m, tau = 1.0 s, and stepping = 0.25 s [66].

The IDM acceleration equation is:

$$\dot{v}_\alpha = a \left[1 - \left(\frac{v_\alpha}{v_0} \right)^\delta - \left(\frac{s^*(v_\alpha, \Delta v_\alpha)}{s_\alpha} \right)^2 \right] \quad (4)$$

where the desired minimum gap s^{**} is defined as (with $s_1 = 0$):

$$s^*(v, \Delta v) = s_0 + vT + \frac{v\Delta v}{2\sqrt{ab}} \quad (5)$$

The original formulation in Treiber et al. [15,16] includes an additional term $s_1 \sqrt{(v/v_0)}$ in the desired gap; however, Treiber et al. [15,16] recommended setting $s_1 = 0$ for simulations with identical vehicles, and SUMO's implementation adopts this simplification. In Equations (4) and (5), v_α denotes the speed of vehicle α (m/s), v_0 denotes the desired velocity (m/s), s_α denotes the actual net (bumper-to-bumper) gap to the leader (m), $\Delta v_\alpha = v_\alpha - v_\alpha - 1$ denotes the velocity difference with respect to the leading vehicle (positive when approaching), s_0 denotes the minimum spacing at standstill (minGap, m), T denotes the desired safe time headway (tau, s), a denotes the maximum acceleration (accel, m/s²), b denotes the comfortable deceleration (decel, m/s²), and δ denotes the acceleration exponent (delta). The calibrated parameters are accel (a), decel (b), delta (δ), minGap (s_0), tau (T), and stepping (internal step length). The stepping parameter controls numerical integration granularity; its consistently negligible sensitivity (see Section 4.5) renders it a candidate for removal from the calibrated set to reduce dimensionality.

3.4. GA Calibration Framework

The GA configuration follows the methodological approaches of Muhan et al. [25], Afshari et al. [1], and Cascan et al. [18], adapted for SUMO and Egyptian TLTW roads.

3.4.1. Parameter Selection and Bounds

Before running the joint CF–LC optimization, a screening sweep was conducted to determine which of the eight LC2013 parameters actually influence overtaking behavior. Each parameter was swept individually (5–6 values, with CF parameters locked at defaults)

across all six sites using five seeds per value. Only three parameters produced meaningful changes in overtaking rate:

1. lcOpposite (200% maximum change)—bilateral controller of opposite-lane overtaking willingness.
2. lcSpeedGain (100%)—speed incentive for discretionary lane changes.
3. lcStrategic (91–114% across models)—strategic route-based lane-change tendency.

The remaining five parameters produced negligible effects (<32%) and were locked at their default values. This screening reduced the LC dimensionality from eight to three. Table 5 summarizes the sweep results.

Table 5. LC2013 parameter sensitivity sweep results (W99 Baseline CF).

LC Parameter	Default	Max Change	Impact	Selected for GA?
lcOpposite	1.0	200%	Bilateral overtaking control	Yes (critical)
lcSpeedGain	1.0	100%	Speed incentive to change lanes	Yes (important)
lcStrategic	1.0	91–114%	Strategic lane-change tendency	Yes (model-dependent)
lcAssertive	1.0	31.5%	Marginal—only at extreme values	No
lcCooperative	0.0	4.7%	Negligible—yielding not a bottleneck	No
lcOvertakeDeltaSpeedFactor	0.0	0%	Inert—controls abort threshold only	No
lcSpeedGainUrgency	0.0	0.8%	Inert at default lcSpeedGain	No
lcSpeedGainLookahead	5.0	3.1%	Negligible—horizon not limiting	No
lcSigma	0.5	0%	Inactive—deterministic logic dominates	No

Based on the screening above, the GA optimized the three retained LC parameters plus actionStepLength jointly with the CF parameters for each model. Search bounds were expanded beyond the typical literature recommendations, particularly for LC parameters, for which prior work in contexts similar to Egypt [9,14] suggested the need for wider ranges. Table 6 lists the bounds for the three retained LC parameters and actionStepLength, which are common to all models. Tables 7–9 present the CF-specific bounds for the Krauss, W99, and IDM models, respectively. For the complete calibration process, see the Supplementary Materials section.

Table 6. Common lane-change parameters (LC2013)—all models.

Parameter	Lower	Upper	Description
lcOpposite	0.0	5.0	Opposite-lane overtaking willingness
lcSpeedGain	0.0	5.0	Speed incentive to change lanes
lcStrategic	0.0	10.0	Strategic lane-change factor
actionStepLength	0.1	1.0	Driver decision frequency (s)—rounded to 0.1 s

Table 7. Krauss model parameter search bounds (10 free parameters).

Parameter	Default	Lower	Upper	Description
accel	2.6	0.1	7.0	Maximum acceleration (m/s^2)
decel	4.5	0.1	7.0	Comfortable deceleration (m/s^2)
sigma	0.5	0.1	1.0	Driver imperfection
sigmaStep	0.1	0.1	1.0	Imperfection update interval (s)
tau	1.0	0.1	5.0	Reaction time (s)
minGap	2.5	0.1	10.0	Minimum standstill gap (m)

Table 7. Cont.

Parameter	Default	Lower	Upper	Description
+3 LC params	—	—	—	lcOpposite, lcSpeedGain, lcStrategic
+actionStepLength	—	—	—	Action step length

Table 8. Wiedemann 99 model parameter search bounds (3 or 4 free CF parameters).

Parameter	Default	Lower	Upper	Description	Scenarios
cc1	1.30	0.5	10.0	Desired headway time (s)	Both
cc3	−12.00	−20.0	0.0	Threshold for entering car-following	Both
cc5	0.35	0.1	5.0	Positive following threshold (m/s)	Both
minGap (cc0)	1.50	0.1	10.0	Standstill gap (m)	Locked/Free

Note. Locked W99 parameters: cc2 = 8.0, cc4 = −0.25, cc6 = 6.0, cc7 = 0.25, cc8 = 2.0, cc9 = 1.5 (SUMO source defaults).

Table 9. IDM parameter search bounds (6 free CF parameters).

Parameter	Default	Lower	Upper	Description
accel	2.6	0.1	7.0	Maximum acceleration (m/s^2)
decel	4.5	0.1	7.0	Comfortable deceleration (m/s^2)
delta	4.0	1.0	10.0	Acceleration exponent
minGap	2.5	0.1	10.0	Minimum standstill gap (m)
stepping	0.25	0.1	1.0	Internal step length (s)
tau	1.0	0.1	5.0	Desired time headway (s)

3.4.2. GA Configuration

The GA configuration is summarized in Table 10. Each generation entails $15 \times 6 \times 5 = 450$ simulations. With ten parallel workers and progressive simulation durations, a single generation requires 2–17 min. The GA was implemented in Python (3.11) using the NumPy library with parallel execution via the multiprocessing module. Model states were checkpointed after each generation using Python’s pickle serialization, enabling resumption in the event of process failure.

Table 10. Genetic algorithm configuration.

Parameter	Value	Justification
Population size	15	Smaller population for computational efficiency given 6 sites $\times$ 5 seeds = 450 simulations per generation
Maximum generations	25	Sufficient for convergence; early termination after 5 generations without improvement
Parallel workers	10	Maximizes CPU utilization while avoiding Traffic Control Interface (TraCI) port contention
Evaluation seeds	5	Balances stochastic robustness with computational cost
Validation seeds	10	Independent seeds for overfitting assessment
Selection	Tournament ($k = 3$)	Standard pressure–diversity balance [25]
Crossover	BLX- α ($\alpha = 0.5$)	Blend crossover extends offspring beyond the convex hull of parents
Mutation	Adaptive Gaussian	σ starts at 15% of range, decays to 3% by final generation
Elitism	1 best individual	Monotonic convergence guarantee
Seeding	2 known individuals	Seed 1 = model defaults; Seed 2 = conservative values
Progressive simulation end	Short to full	Early generations truncated; full 3800 s from ~66% progress onward

3.4.3. Fitness Function

The objective function was defined as an equally weighted MAPE across four metrics per site, averaged over five evaluation seeds, following [18]:

$$F = \frac{1}{N} \sum_{s \in S} \sum_{m \in M} \sum_{k=1}^{S_{\text{seeds}}} w_m \cdot \frac{|\text{sim}_{s,m,k} - \text{obs}_{s,m}|}{S_{\text{seeds}} \cdot \text{obs}_{s,m}} \quad (6)$$

where $S = \{\text{six sites}\}$, $M = \{\text{speed, travel time, headway, overtakes/h}\}$, $w = 0.25$ is the per-metric weight, $S_{\text{seeds}} = 5$ (the number of evaluation seeds), and N is the number of valid site–metric pairs (typically 24). Equivalently, $F = w \times \text{MAPE}_{\text{frac}}$, where $\text{MAPE}_{\text{frac}}$ is the unweighted average absolute fraction error across all valid site–metric pairs and seeds. MAPE was selected because percentage errors are interpretable across metrics with disparate units (Punzo et al. [3], provide a comprehensive comparison of goodness-of-fit measures for CF calibration).

Equal weighting ($w = 0.25$) assumes that each metric contributes equally to total error. Speed and travel time are algebraically related ($\text{TT} = L/v$); including both, therefore, increases the effective influence of speed-related behavior relative to the nominal 25% speed weight. Headway and overtaking rate also exhibit higher noise-to-signal ratios than speed and travel time. To examine the speed–travel-time redundancy, a dedicated ablation analysis was conducted in which travel time was removed from the fitness function and the GA was re-run for the three leading calibrated configurations (Section 4.15). The ablation changed the model ranking, indicating that metric selection can affect comparative conclusions. Therefore, the four-metric weighted-sum formulation is retained because it reflects the pre-specified practitioner metrics used in this study, but the influence of metric selection is acknowledged as a methodological limitation. A Pareto-based multi-objective formulation, such as NSGA-II [67], is recommended for future work because it would show trade-offs among speed, travel time, headway, and overtaking more transparently.

The overtaking-rate metric is based on a modest sample of 685 successful overtakes across all sites and directions, and its MAPE is higher than the speed and travel-time MAPE values. A descriptive post-hoc reweighting of the per-metric validation errors showed that reducing the overtaking weight narrows the difference among leading configurations, while inverse-error weighting can favor models that perform best on speed and travel time. Because these alternative weights were not used to re-run the GA, the calculation is descriptive only. It nevertheless shows that model ranking is sensitive to metric weighting, and the equal-weighted results should be interpreted in that context.

3.4.4. Selection, Crossover, and Mutation

Selection: Tournament selection ($k = 3$) was employed to balance selection pressure and population diversity [25].

Crossover: BLX- α ($\alpha = 0.5$) with probability pc was used to generate offspring that extend beyond the convex hull of parent solutions.

Mutation: Adaptive Gaussian mutation was applied, in which the standard deviation σ decays from 15% to 3% of each parameter's range, applied with probability pm (where n is the number of parameters).

3.4.5. Convergence Criterion

Each generation proceeds through five steps: (1) fitness evaluation, (2) parent selection, (3) crossover, (4) mutation, and (5) elitism (retention of the best individual). The algorithm terminates when 25 generations have elapsed or when five consecutive generations fail to improve the best fitness.

3.4.6. Calibration Workflow

The calibration procedure comprised the following sequential steps:

1. Extract observed trajectory data from video recordings;
2. Construct SUMO network models for each study site;
3. Generate traffic demand from observed volumes and fleet composition;
4. Initialize the GA population;
5. Execute SUMO simulations for each parameter set;
6. Extract simulated trajectories and aggregate metrics;
7. Compute fitness against observed values;
8. Apply selection, crossover, and mutation;
9. Evaluate convergence; and
10. Validate the final solution on 10 random seeds.

3.5. Performance Measures

Following Ciuffo et al. [61] and Punzo et al. [3], the mean absolute percentage error (MAPE) was selected as the primary performance measure for both calibration and validation. MAPE was chosen because percentage errors are interpretable across metrics with disparate units (speed in m/s, travel time in s, headway in s, and overtaking rate in veh/h), enabling direct comparison of model accuracy across the four calibration targets.

3.5.1. Mean Absolute Percentage Error

Mean absolute percentage error (MAPE):

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i} \times 100 \quad (7)$$

where y denotes the observed value, $\hat{y}$ denotes the simulated value, and n is the number of observations. Additional measures—including root-mean-square error (RMSE), the Geoffrey E. Havers (GEH) statistic, and volume-adjusted RMSE—were considered but not employed in the present study because MAPE provided sufficient discrimination among model configurations and is directly reported in the GA fitness function (Section 3.4.3).

3.5.2. Acceptance Criteria

The following acceptance thresholds were adopted, based on the expected accuracy range for microscopic behavior calibration (10–20%, [2]) and adapted for heterogeneous traffic conditions:

1. MAPE < 15% for speed, travel time, and headway;
2. MAPE < 20% for overtaking rate (reflecting its higher stochastic variability);
3. Overall MAPE < 12% across all metrics and sites.

These thresholds are author-defined rather than prescribed by a standard; they were selected to ensure that no individual metric error exceeds the typical validation range reported in the SUMO literature, while accounting for the additional variability introduced by mixed vehicle classes and multi-site aggregation.

3.6. Computational Implementation

The computational implementation comprised the following components:

1. SUMO interface: TraCI v1.27.1;
2. GA engine: Custom Python implementation (NumPy), without external GA libraries;
3. Parallelization: Python multiprocessing. Pool with spawn context;
4. Checkpointing: Python pickle serialization after each generation;

5. Overtaking detection: Custom XML parser (round-trip per vehicle ID);
6. Hardware: 10-core CPU, 16 GB RAM, with temporary outputs directed to a RAM-backed filesystem when available;
7. Typical runtimes: IDM: 157 min; W99: ~170–250 min; Krauss: ~330 min.

3.7. Calibration and Validation Workflow

The calibration and validation workflow proceeds in three stages. Stage 1 is GA calibration, in which each candidate parameter set is evaluated across six sites and five random seeds to average stochastic variation during optimization. Stage 2 is validation with ten independent random seeds not used during GA optimization, which assesses robustness to stochastic simulation variation. Stage 3 is a limited multi-replicate GA check for the top-performing IDM free-minGap configuration, used to estimate optimization noise and reproducibility. The final goal is to obtain aggregate parameter sets that reproduce observed speed, travel time, headway, and successful-overtaking rates for the six study sites while transparently reporting the limits of stochastic and spatial validation.

Validation Methodology

All six study sites were included in the original GA calibration; there was no site-level holdout in the main calibration. The validation conducted in the main analysis is therefore stochastic and behavioral rather than a full out-of-sample spatial validation. It assesses robustness to: (1) seed independence—ten independent random seeds distinct from the five used during GA optimization; (2) a successful-overtaking behavioral holdout—203 successful overtakes that were not used during calibration; and (3) per-site diagnostics—MAPE reported individually for all six sites to identify site-specific biases. Although 171 aborted overtakes were also reserved in the holdout dataset, they were not simulated or used as validation targets in the present study; this omission is acknowledged as a limitation. A leave-one-site-out sensitivity analysis for the IDM free-minGap configuration is reported in Section 4.16 to quantify sensitivity to excluding individual sites. cross-site consistency. Transferability was further examined by comparing per-site error patterns across the Nile Delta region.

4. Results and Discussion

4.1. GA Convergence

Table 11 summarizes the convergence results across all six model configurations. The IDM with free minGap converged in 157 min with a fitness of 0.02068—the best fitness achieved across all configurations.

Table 11. GA convergence summary.

Model	Free Params	Generations	Runtime (min)	Best Fitness	MAPE (%)
IDM (free minGap)	10	13 (early stop)	157	0.02068	8.79
Krauss (free minGap)	10	21 (early stop)	330	0.02209	10.56
W99 (locked minGap = 1.50)	7	19 (early stop)	250	0.02076	8.94
IDM (locked minGap = 1.50)	9	13 (early stop)	180	0.02675	10.85
W99 (free minGap)	8	14 (early stop)	170	0.02534	10.51
Krauss (locked minGap = 1.50)	9	16 (early stop)	246	0.03195	12.49

The W99 model with locked minGap matched this performance at 0.02076 but required 19 generations (250 min). The Krauss model with free minGap required 21 generations and 330 min to reach 0.02209. The progressive simulation-end strategy (see Section 3.4.2) shortened early generations to approximately half the full 3800 s duration, reducing per-

generation runtime during the first ~16 generations (approximately 66% of progress) without affecting the final converged fitness, as the last full-duration generations determined the reported minima.

4.2. Calibrated Parameters

Tables 12–15 present the calibrated parameter values for all model configurations. For each parameter, the calibrated value, the default (or GA starting value), and the status relative to the search bounds are reported. Table 13 presents the IDM results when minGap was locked at 1.50 m.

Table 12. IDM calibrated parameters—free minGap (MAPE: 8.79%).

Parameter	Calibrated	Default	Status
accel (m/s^2)	3.32	2.6	Interior
decel (m/s^2)	3.56	4.5	Below default—gentler braking
delta	3.14	4.0	Below default—lower acceleration exponent
minGap (m)	3.10	2.5	Slightly above default
stepping (s)	0.39	0.25	Above default—coarser internal stepping
tau (s)	0.86	1.0	Below default—shorter desired headway
lcOpposite	2.34	1.0	Aggressive opposite-lane overtaking
lcSpeedGain	5.00	1.0	Hit upper bound
lcStrategic	0.84	1.0	Slightly below default
actionStepLength (s)	0.85	0.4 *	Coarser driver decision frequency

Note: * The “default” column reports the GA search starting value (0.4 s). SUMO’s built-in actionStepLength default equals the simulation step length (0.1 s in this study). The starting value is reported to ensure the calibrated result is interpretable relative to the search range.

Table 13. IDM calibrated parameters—locked minGap = 1.50 m (MAPE: 10.85%).

Parameter	Calibrated	Default	Status
accel (m/s^2)	1.90	2.6	Below default—gentler acceleration
decel (m/s^2)	3.51	4.5	Below default—gentler braking
delta	9.07	4.0	Near upper bound—high acceleration exponent
minGap (m)	1.50 (locked)	2.5	Fixed
stepping (s)	0.42	0.25	Above default—coarser internal stepping
tau (s)	0.77	1.0	Below default—shorter desired headway
lcOpposite	4.80	1.0	Near upper bound—aggressive overtaking
lcSpeedGain	1.11	1.0	Interior
lcStrategic	0.93	1.0	Interior
actionStepLength (s)	1.00	0.4	Hit upper bound

Table 14. Krauss calibrated parameters—free minGap (MAPE: 10.56%).

Parameter	Calibrated	Default	Status
accel (m/s^2)	3.75	2.6	Interior
decel (m/s^2)	2.42	4.5	Below default—gentler braking
sigma	0.25	0.5	Below default—less driver imperfection
sigmaStep (s)	0.10	step-length	Near lower bound
tau (s)	0.64	1.0	Below default—shorter reaction time
minGap (m)	0.10	2.5	Near lower bound
lcOpposite	1.82	1.0	Interior
lcSpeedGain	0.87	1.0	Interior
lcStrategic	0.91	1.0	Interior
actionStepLength (s)	0.50	0.4	Interior

Table 15. W99 calibrated parameters—locked minGap (MAPE: 8.94%).

Parameter	Calibrated	Default	Status
cc1 (s)	1.41	1.30	Interior
cc3 (s)	0.00	−12.00	Hit upper bound
cc5 (m/s)	1.81	0.35	Interior
minGap (m)	1.50 (locked)	2.50	Fixed (below default)
lcOpposite	2.87	1.0	Interior
lcSpeedGain	1.68	1.0	Interior
lcStrategic	0.44	1.0	Below default
actionStepLength (s)	1.00	0.4	Hit upper bound

Locking minGap to 1.50 m substantially altered the IDM solution. With minGap unavailable for headway tuning, the GA compensated by pushing delta to 9.07 (near the upper bound), actionStepLength to 1.0 s (upper bound), and lcOpposite to 4.80 (near upper bound), while accel decreased from 3.32 to 1.90 m/s². MAPE degraded from 8.79% to 9.35% (GA internal), confirming that free minGap served as a valuable optimization channel for the IDM. Table 14 lists the Krauss calibrated parameters with free minGap. Table 15 shows the W99 calibrated parameters with minGap locked at 1.50 m.

It is worth noting that for Tables 11–14, the “default” column reports each parameter’s GA search starting value. For actionStepLength, the starting value is 0.4 s; SUMO’s built-in default equals the simulation step length (0.1 s in this study). The minGap default value of 2.50 m corresponds to the SUMO passenger-class default; experimental configurations using locked minGap = 1.50 m are noted accordingly.

4.3. Validation Results

Validation was performed with ten independent seeds (distinct from the five employed during GA optimization). Table 16 presents per-site and overall MAPE values.

Table 16. Ten-seed validation comparison across all models.

Site	IDM Free	IDM Locked *	Krauss (Free)	W99 (Locked)	W99 (Free)	Krauss Defaults
Site 01	14.38%	18.74%	14.50%	13.86%	18.65%	10.31%
Site 02	1.75%	7.71%	2.26%	3.09%	4.96%	10.29%
Site 03	7.50%	5.85%	11.18%	11.70%	9.09%	4.58%
Site 04	13.22%	14.89%	11.46%	11.04%	14.53%	17.03%
Site 05	12.96%	6.13%	9.96%	10.97%	7.54%	7.75%
Site 06	2.93%	11.79%	4.11%	3.00%	8.26%	10.71%
Overall	8.79%	10.85%	10.56%	8.94%	10.51%	10.11%

Note. * The IDM locked configuration’s MAPE exceeds its GA internal fitness (9.35%) because the constrained search space forced the GA toward extreme parameter values (delta = 9.07, lcOpposite = 4.80) that generalized poorly to independent seeds.

To estimate optimization noise and assess reproducibility for the top-performing configuration, three independent GA replicates were executed for the IDM (free minGap) configuration using different GA random seeds; each replicate was then validated with ten SUMO seeds. The replicate MAPE values were 9.65%, 9.35%, and 8.32%, with a mean of 9.11% and a standard deviation of 0.69 percentage points. The original single-run IDM free-minGap validation result (8.79%) lies within this replicate range. Because replicate GA runs were performed only for IDM free minGap, these results are used only as an estimate of optimization noise for that configuration and are not treated as a formal cross-model significance test. The 0.15 pp difference between IDM free minGap (8.79%) and W99 locked

minGap (8.94%) is smaller than this IDM noise estimate, so these two configurations should be interpreted as practically equivalent rather than strictly ranked.

Table 17 presents the per-metric MAPE for the three best configurations, revealing how error is distributed across the four calibration targets. Per-metric MAPE values are computed as the unweighted average of absolute percentage errors across six sites.

Table 17. Per-metric MAPE for best configurations (%).

Metric	IDM Free	Krauss Free	W99 Locked
Speed (m/s)	4.95	4.50	5.39
Travel time (s)	4.05	3.26	6.44
Headway (s)	11.52	12.99	10.17
Overtaking (/h)	14.64	21.35	13.71
Overall	8.79	10.56	8.94

The MAPE (%) values reported in Tables 16 and 17 are computed as the unweighted average of percentage errors across ten independent validation seeds, which differs from the five-seed GA fitness due to additional stochastic variation.

The overall MAPE is from the per-site computation (Table 16); the mean of the four per-metric values may differ by up to 0.03 pp due to rounding of intermediate values. Speed and travel time achieve the lowest errors (3–6%), while headway and overtaking rate exhibit higher errors (10–21%), reflecting their greater stochastic variability.

With respect to the acceptance thresholds defined in Section 3.5.2, the IDM (free minGap) and W99 (locked minGap) configurations satisfy all four individual-metric criteria (MAPE < 15% for speed, travel time, and headway; MAPE < 20% for overtaking rate) as well as the overall criterion (MAPE < 12%). The Krauss (free minGap) configuration satisfies the overall criterion (10.56% < 12%) but exceeds the overtaking-rate threshold (21.35% > 20%). Thus, Krauss free minGap is classified as partially compliant: it may be acceptable where aggregate accuracy is the primary concern, but it should be used cautiously for analyses in which overtaking rate is a critical output.

The three top-performing models—IDM with free minGap, W99 with locked minGap, and Krauss with free minGap—achieved overall MAPE values of 8.79%, 8.94%, and 10.56%, respectively. The IDM configuration achieved the lowest overall error and performed best at three of the six sites. Locking the IDM minGap degraded overall MAPE by 2.06 percentage points (pp) to 10.85%, a result inferior to both the free-minGap W99 configuration (10.51%) and the uncalibrated Krauss defaults (10.11%). The penalty was most severe at Site 01 (+4.36 pp) and Site 06 (+8.86 pp). The locked configuration, however, improved accuracy at Site 03 (−1.65 pp) and Site 05 (−6.83 pp), indicating that the constraint redistributes error across sites rather than eliminating it.

4.4. Overtaking Comparison

Table 18 compares simulated and observed successful overtaking rates (overtakes per hour). The calibration and validation in this study used only successful overtaking rates. The 171 aborted overtaking maneuvers reserved in the holdout dataset were not incorporated into the SUMO simulations or used in the fitness function. This limitation is important because aborted maneuvers are directly related to overtaking risk and gap acceptance; future work should include aborted-overtake rate as a separate objective or constraint if suitable field coding and model representation are available. Most calibrated models slightly underpredicted the mean overtaking rate (15.2–15.5 vs. 16.7 observed). Notable exceptions include the IDM with locked minGap (18.13), the free-minGap W99 (16.78), and the uncalibrated Krauss defaults (18.38), all of which overpredicted. The IDM

locked configuration overestimated overtaking at Site 06 by 41.7% and at Site 02 by 25.4%, confirming that constraining minGap forces more aggressive overtaking behavior at those locations. Site 05 exhibited the largest absolute prediction error across all models (−1.8 to −5.4 overtakes/h), suggesting the presence of a site-specific factor not captured by CF or LC parameters. Only successful overtakes were used as a calibration target; the 555 aborted overtakes recorded in the field were excluded. Including the aborted overtake rate as a separate objective could improve safety realism without requiring additional data collection.

Table 18. Overtaking rate comparison (overtakes/h).

Site	Observed	DM Free	IDM Locked	Krauss (Free)	W99 (Locked)	W99 (Free)	Krauss Defaults
Site 01	15.5	12.7	10.2	12.4	13.4	10.8	16.0
Site 02	18.5	18.6	23.2	18.8	20.1	20.5	22.3
Site 03	18.6	16.5	19.2	13.4	13.6	16.4	18.1
Site 04	16.0	17.8	18.9	17.0	16.6	18.4	20.2
Site 05	14.0	8.6	12.5	10.0	9.8	12.2	11.9
Site 06	17.5	18.9	24.8	19.4	18.5	22.4	21.8
Average	16.69	15.52	18.13	15.17	15.33	16.78	18.38

4.5. Sensitivity Analysis

Two sensitivity stages were conducted. The pre-calibration LC sweep (see Table 5, Section 3.4.1) evaluated all eight LC2013 parameters individually with CF parameters locked, thereby reducing the active set to three parameters.

The post-calibration one-at-a-time (OAT) analysis applied $\pm 10\%$ perturbations to each calibrated parameter at the converged solution (not at default values). Table 19 compares overtaking sensitivity across the top-performing models. It should be noted that OAT captures only local main effects; interactions between LC and CF parameters—visible when minGap is locked—may be underestimated.

Table 19. Sensitivity ranking comparison (overtaking sensitivity percentage).

Rank	IDM Free (8.79%)	Sens%	IDM Locked (10.85%)	Sens%	Krauss Free (10.56%)	Sens. %	W99 Locked (8.94%)	Sens. %
1	lcOpposite	199.1	lcStrategic	232.7	lcSpeedGain	116.5	lcOpposite	164.8
2	lcStrategic	113.9	accel	137.8	lcOpposite	111.5	actionStepLength	130.5
3	accel	104.9	decel	111.2	decel	109.5	cc1	129.5
4	actionStepLength	104.5	lcSpeedGain	109.5	accel	108.9	lcSpeedGain	101.9
5	lcSpeedGain	98.5	lcOpposite	105.5	lcStrategic	104.2	lcStrategic	91.9
6	decel	94.9	actionStepLength	89.2	tau	96.2	-----	---
7	tau	85.6	delta	73.6	sigmaStep	88.6	-----	---
8	delta	71.9	tau	9.7	sigma	70.9	-----	---
9	minGap	38.3	stepping	0.0	minGap	47.6	-----	---
10	stepping	0.0	minGap	0.0	actionStepLength	36.0	-----	---

Key findings from the sensitivity analysis include the following:

- lcOpposite ranked first or second in overtaking influence across all free-minGap models. In the IDM locked configuration, it dropped to fifth (105.5%), as the constraint shifted influence to other parameters.
- minGap exhibited moderate sensitivity (38–48%) when free—not negligible, but secondary to LC parameters. Sensitivity was trivially 0% when locked.
- Locking minGap for the IDM fundamentally altered which parameters controlled overtaking. The sensitivity of lcStrategic doubled from 113.9% to 232.7%, becoming

the dominant parameter, whereas tau sensitivity decreased from 85.6% to 9.7%. This finding indicates that parameter sensitivity depends on which parameters remain available for simultaneous optimization.

- actionStepLength exhibited model-dependent sensitivity: high for W99 (130.5%) and IDM free (104.5%), but low for Krauss (36.0%).

cc3, cc5, and stepping were negligible across all configurations, rendering them candidates for removal from the calibrated set. The cc5 parameter was calibrated to 1.81 m/s ($5.2 \times$ the SUMO default of 0.35) despite its low sensitivity, likely because collinearity with cc1 or actionStepLength allows many (cc1, cc5) pairs to yield similar fitness; the GA converged to one extreme without a clear fitness penalty.

4.6. Model Performance Comparison

The IDM with free minGap achieved the lowest MAPE (8.79%), followed by W99 with locked minGap (8.94%) and Krauss with free minGap (10.56%). The models differed primarily in parsimony and computational efficiency:

IDM with free minGap (Table 12): lowest MAPE (8.79%) and fastest runtime (157 min). It offers the greatest CF flexibility through the delta parameter.

IDM with locked minGap (Table 13): MAPE of 10.85% (+2.06 pp), inferior to both the free-minGap W99 (10.51%) and the uncalibrated Krauss defaults (10.11%). Free minGap is therefore critical for the IDM.

W99 with locked minGap (Table 15): matched top-tier accuracy with only three CF parameters, compared to six for the IDM and Krauss models. It is the preferred choice when parsimony is important.

W99 with free minGap: MAPE of 10.51%—freeing minGap degraded W99 performance because the GA converged to unrealistic values of cc1 = 0.50 s and minGap = 0.51 m. Krauss with free minGap (Table 14): MAPE of 10.56% but required 21 generations and 330 min—the slowest convergence among all models tested.

Krauss defaults: MAPE of 10.11% with zero calibration effort, a result that modestly outperformed several calibrated configurations under the straight, undersaturated conditions examined. This should not be taken as a general endorsement: trial-and-error (22.20%, [59]) is not a systematic baseline, and the defaults were not tested on congested, curved, or urban geometries.

4.7. Parameter Interpretation

All three calibrated models converged to gentler deceleration values than the SUMO defaults (IDM: 3.56 vs. 4.5 m/s²; Krauss: 2.42 vs. 4.5 m/s²). The W99 cc3 value at the upper bound implies minimal braking intervention under normal following conditions. These results suggest that Egyptian drivers on TLTW roads brake less aggressively than the Western-calibrated defaults assume.

Tau values (Krauss: 0.64 s; IDM: 0.86 s) and cc1 (W99: 1.41 s) fell below their respective defaults, indicating shorter following headways. This outcome is consistent with the aggressive driving behavior documented by [6,9].

lcOpposite values (IDM: 2.34; Krauss: 1.82; W99: 2.87) all exceeded the SUMO default of 1.0, which is consistent with the high frequency of opposite-lane overtaking observed on TLTW roads. lcSpeedGain reached the upper bound for the IDM (5.0) but remained moderate for Krauss (0.87) and W99 (1.68). lcStrategic converged at or below default values (0.44–0.91); on short sections, immediate speed gain appears to be more consequential than route-based strategy. For the IDM with locked minGap, lcOpposite shifted to 4.80 (near the upper bound), as the constraint forced more aggressive overtaking to compensate for the fixed standstill gap.

All three top-performing models converged to coarser action step lengths (0.50–1.00 s) than the search starting value (0.4 s). This pattern may indicate that Egyptian drivers on TLTW roads make fewer, more deliberate decisions—a behavior consistent with the high-risk nature of opposite-lane overtaking.

A separate sweep of actionStepLength across the full calibration range (0.1–1.0 s in 0.1 s steps, 6 sites $\times$ 10 seeds = 600 simulations) using the trial-and-error W99 configuration from Shoaeb [59] revealed minimal impact on car-following metrics. Speed varied by 0.02 m/s (0.1% of observed), travel time by 0.05 s (0.3%), and headway by 0.06 s (0.5%). Overtaking rate was more responsive (1.7 overtakes/h, 10.0% range), decreasing with larger actionStepLength as coarser temporal resolution reduces detected lane-change opportunities. MAPE ranged from approximately 20.6% (actionStepLength = 0.1 s) to 23.1% (1.0 s); the 0.4 s value used by Shoaeb [59] yielded 21.7%, near the lower end. This relatively flat MAPE surface explains why the GA converged to different actionStepLength values (0.50–1.00 s) across the three best configurations without meaningful CF differences—the parameter predominantly affects overtaking dynamics rather than CF fidelity.

4.8. The Role of minGap

The appropriate treatment of minGap proved model-dependent, and incorrect specification imposed measurable accuracy penalties:

For the IDM, the minGap parameter should remain free (calibrated to 3.10 m; 38.3% sensitivity). Locking it at 1.50 m degraded MAPE by 2.06 pp (to 10.85%)—the largest penalty observed across all experiments. In the locked configuration, the GA compensated with extreme parameter values ($\delta \rightarrow 9.07$, $lcOpposite \rightarrow 4.80$, actionStepLength $\rightarrow 1.0$), while accel decreased to 1.90 m/s². The sensitivity structure shifted substantially: lcStrategic increased from 113.9% to 232.7%, whereas tau sensitivity decreased from 85.6% to 9.7%. Free minGap is integral to the IDM acceleration function (Equations (4) and (5)). Krauss. The minGap parameter should also remain free (calibrated to 0.10 m; 47.6% sensitivity). Locking it imposed a penalty of +3.58 pp (to 12.49%). The Krauss model relies primarily on tau for spacing control; explicit standstill gap is a secondary tuning aid.

For the W99, the minGap parameter (equivalently cc0, which SUMO forces to equal the vType minGap attribute) should be locked at 1.50 m. Freeing it degraded MAPE by 1.57 pp (to 10.51%) because the GA drove cc1 to 0.50 s and minGap to 0.51 m—an unrealistic regime.

The Krauss model incurred the largest penalty from locking minGap (+3.58 pp), followed by IDM (+2.06 pp), whereas W99 degraded when minGap was freed (+1.57 pp). The parameter serves a structurally different role in each model: as a tunable offset interacting with δ in the IDM, as a secondary spacing aid mediated by tau in Krauss, and as a fixed reference against cc1 in W99. Consequently, increasing degrees of freedom does not guarantee superior calibration outcomes.

Tau and minGap are known to be collinear in both the IDM and Krauss models [3]. The GA may, therefore, converge to different tau–minGap combinations with equivalent fitness, meaning that individual parameter values cannot be treated as uniquely identified. This collinearity does not impair predictive accuracy—the fitness landscape is flat along the collinear ridge—but it limits the physical interpretability of specific coefficients.

4.9. Physical Plausibility of Calibrated Values

Several calibrated values exceeded physically intuitive ranges and required explicit qualification:

Krauss minGap (0.10 m). A standstill gap of 10 cm is physically implausible on a rural road. This value indicates that the Krauss model delegates spacing control to tau rather

than to minGap; the two parameters are collinear [3], and the GA drove minGap to its lower bound while tau (0.64 s) provides effective headway control. In practice, the Krauss minGap should be interpreted as a numerical offset in the safe-speed calculation rather than as a physical standstill distance.

IDM lcSpeedGain (5.00, upper bound). Reaching the upper bound indicates that the search range is restrictive rather than that 5.0 represents the true optimum. For TLTW roads with frequent opposite-lane overtaking, the speed-gain incentive likely saturates above SUMO's default range; extending the bound to [0, 10] is recommended (see Section 5.3).

W99 cc3 (0.00 vs. default −12.00). The cc3 parameter governs the threshold for entering car-following mode. A calibrated value of 0.0 implies no perceptual delay before entering following mode, which is unlikely to reflect actual driving behavior on rural roads. This outcome likely reflects short-section bias (see Section 5.2.1): on 205–375 m sections, car-following episodes are too brief for the cc3 threshold to be empirically identifiable, causing the GA to push the parameter to its bound.

IDM locked delta (9.07). With delta approaching 10, the IDM acceleration function becomes nearly discontinuous near the desired speed. This is not a physically meaningful driver-following regime; rather, it represents a compensatory response to the locked minGap constraint (see Section 4.8). The free-minGap configuration yielded delta = 3.14, a physically plausible value.

Krauss decel (2.42 m/s²). A comfortable deceleration below 2.5 m/s² is relatively low for straight rural roads. It may reflect gentler braking behavior in Egyptian traffic (consistent with the cc3 value at the upper bound in W99) or a site-specific adaptation to the short-section geometry.

These examples illustrate a broader principle: calibrated parameters that reach search bounds or adopt extreme values should not be interpreted as direct measurements of driver behavior. They represent effective tuning coefficients that may compensate for model structural limitations, data constraints, or restrictive search ranges. Transparent reporting (as provided in Tables 11–14) is essential, but physical interpretation requires the caveats articulated above.

4.10. Site-Specific Patterns

The same sites were consistently well or poorly predicted across all models, indicating that site-level factors—rather than model deficiencies—dominate the residual error:

Best predicted: Sites 02 and 06 ($\leq 3\%$ error for IDM free)—both feature moderate section lengths (215–220 m) and balanced directional fleets.

Worst predicted overall: Sites 01 and 04—both have longer sections (280–375 m) and exhibit the highest MAPE across most models. Longer sections amplify cumulative speed and headway errors. Moreover, the short section lengths (205–375 m) across all sites may limit the identifiability of steady-state CF parameters, biasing calibration toward free-flow and lane-change initiation dynamics.

Worst metric: Site 05 overtaking, underpredicted by 28.6–38.6% across all models. The omission of shoulder-assisted and lateral-offset behavior may contribute to this discrepancy. This explanation should be interpreted cautiously because the present data do not isolate shoulder use as a separate measured variable. However, the field geometry indicates that shoulders were available at all sites, and such lateral behavior is common on Egyptian TLTW roads. In SUMO's lane-centered representation, partial evasive maneuvers and shoulder-assisted passing opportunities are not directly represented. Therefore, some overtaking underprediction may reflect geometric and lateral-behavior simplification rather than only incorrect CF or LC parameter values. Future work should evaluate SUMO's sublane model or an explicit shoulder-use representation using field observations that separately classify

full opposite-lane passes, partial passes, and shoulder-assisted maneuvers. Cross-site trade-offs: Locking the IDM minGap improved Site 03 (−1.65 pp) and Site 05 (−6.83 pp) while degrading Site 01 (+4.36 pp) and Site 06 (+8.86 pp). The constraint redistributed error across sites rather than eliminating it.

Because only six sites are available, formal regression with multiple predictors would not be statistically reliable. The site-characteristic analysis is, therefore, treated as descriptive. Site 04, which has the longest section length (375 m), the longest observed travel time, and the lowest total flow among the six sites, produced the largest LOSO error when excluded from calibration (Section 4.16). This suggests that longer sections and distinctive traffic conditions may amplify cumulative speed, headway, and overtaking errors, but the pattern cannot be generalized from six sites alone.

4.11. Under-Dispersion

A consistent pattern across all models was that simulated standard deviations were smaller than observed values for all four metrics (simulated SD: 0.1–5.3 vs. observed: 1.4–9.7). The ratio of simulated-to-observed standard deviation ranged from approximately 0.2 (headway) to 0.55 (speed).

This apparent under-dispersion is primarily a measurement artifact arising from mismatched aggregation levels. For example, at Site 01 the observed speed SD was 11.29 km/h (Table 2), computed from individual vehicle speeds. The simulated SDs reported here are computed from ten run-mean speeds, each already averaging many individual vehicles. These two quantities measure different forms of variability: observed SD reflects within-stream vehicle-to-vehicle variability, whereas simulated SD reflects variability of run means across stochastic seeds. By the central limit theorem, the SD of run means is approximately $\sigma/\sqrt{n}$. If the simulated per-vehicle speed SD were comparable to the observed 11.29 km/h and a run contained about 250 vehicles, the expected SD of run means would be approximately $11.29/\sqrt{250} \approx 0.71$ km/h. Thus, comparing observed across-vehicle SD with simulated run-mean SD naturally compresses the simulated dispersion. A fair distributional comparison requires simulated per-vehicle standard deviations from individual trajectories within a representative run, not only run-mean values. The uniform CF parameter set (Section 5.2.2) may contribute secondarily by reducing behavioral spread across vehicle classes, but the dominant effect is the aggregation-level mismatch. For a fair comparison, simulated per-vehicle standard deviations should be computed from individual vehicle trajectories within a single representative run, rather than from run-mean values. This is recommended as a direction for future work (Section 5.3). At present, the models are best suited to mean-value comparisons; statements about absolute distributional spread should be interpreted with the caveat above.

4.12. Validation and Transferability

The 10-seed validation confirmed that IDM free minGap, W99 locked minGap, and Krauss free minGap were the three leading GA-calibrated configurations in the original four-metric validation. However, per-site validation MAPE varied considerably (1.75–14.38% for IDM free), indicating that a single parameter set optimized across six sites cannot achieve uniform accuracy at all locations. Validation based on independent random seeds should be interpreted as stochastic validation, not spatial transferability to unseen sites. The LOSO analysis in Section 4.16 provides a more direct sensitivity check for within-region transferability. Overall, the results support reasonable aggregate prediction within the Nile Delta study context but not uniform site-level accuracy. Site-specific calibration would likely improve per-site accuracy but at the cost of generalizability.

4.13. Comparison with Literature

The MAPE range of 8.79–10.85% compares favorably with prior work. However, a direct comparison requires caution: the aggregate MAPE reported herein (averaged across sites, metrics, and seeds) is not equivalent to trajectory-level errors reported in the literature. Aggregate metrics cancel individual-vehicle deviations and will, therefore, always be lower than pointwise trajectory errors. Lopez et al. [2] expected validation errors of 10–20% for microscopic behavior. The present multi-site, multi-seed approach achieved values at the lower end of this range. Chaudhari et al. [14]: Vehicle-class-specific W99 variations for Indian traffic. The single-parameter-set assumption adopted in the present study limits accuracy but simplifies practical application. Kesting and Treiber [17]: IDM errors of 11–29% using trajectory-level RMSE. The lower 8.79% reported herein is expected because aggregate MAPE averages out individual-vehicle deviations and is not directly comparable to microscopic trajectory metrics. This caveat applies throughout the comparison. Krauss defaults: the uncalibrated MAPE of 10.11% is comparable to many calibrated results in the literature. SUMO's default parameters provide a competitive baseline for straight, undersaturated sections. The finding that cc3 and cc5 exert negligible influence on overtaking is specific to the TLTW context. Schrader et al. [12] calibrated all ten W99 parameters on a signalized urban corridor without sensitivity screening, explicitly noting this as a direction for future work. In their study, cc3 remained near default (−11.3 vs. −12.0), whereas cc5 shifted substantially (1.97–2.31 vs. 0.35 default)—confirming that parameter importance is context-dependent. The LC pre-sweep and post-calibration OAT analyses reported in the present study respond to the sensitivity-analysis gap identified by Schrader et al. [12].

4.14. Methodological Contributions, TLTW Specificity, and Parameter Identifiability

The calibration results should be interpreted not only as point estimates of driver behavior but also as effective simulation coefficients that compensate for unmodeled features of Egyptian TLTW traffic, including shoulder use, mixed vehicle dynamics, and lane-centered representation in SUMO.

The TLTW setting creates a specific measurement issue for overtaking. A completed opposite-lane pass consists of two lane-change events: departure into the opposing lane and return to the original lane. Counting raw lane-change events would, therefore, overestimate successful overtakes. For this reason, the present study counts completed overtakes using a vehicle-ID-based round-trip protocol, in which one completed departure–return sequence by the same vehicle is counted as one successful overtake.

Parameter identifiability was assessed using the final GA population and replicate/LOSO parameter comparisons where available. For IDM free minGap, the strongest observed correlations included tau–minGap ($r = +0.40$), delta–lcStrategic ($r = +0.54$), and decel–minGap ($r = +0.58$). For IDM locked minGap, the strongest observed relationship was delta–lcSpeedGain ($r = -0.79$). Correlations involving locked parameters were not computed because fixed parameters have zero variance and Pearson correlation is undefined. These diagnostics indicate moderate-to-strong compensation among some parameters, meaning that individual calibrated values should not be interpreted as uniquely identified behavioral constants.

This conclusion is reinforced by the IDM replicate and LOSO results. Several parameters varied substantially across runs while producing comparable MAPE values, indicating a near-optimal region rather than a single unique optimum. Thus, the calibrated parameters are best understood as effective coefficients for reproducing aggregate speed, travel time, headway, and overtaking-rate patterns under the observed undersaturated TLTW conditions.

When minGap is fixed in the IDM, the optimizer loses a direct spacing-adjustment channel in the dynamic-gap term. The calibrated solution appears to compensate by increasing delta, which changes the free-road acceleration response, increasing lcOpposite and lcStrategic, which increases overtaking propensity, and increasing actionStepLength, which changes the frequency of driver decision updates. The reduced sensitivity of tau under locked minGap is, therefore, interpreted as an empirical result of parameter compensation, not as proof of a single causal mechanism.

4.15. Ablation Study

Speed–Travel-Time Redundancy Because travel time is algebraically related to speed for a fixed section length, including both metrics in the weighted fitness function increases the influence of speed-related behavior. To test whether this redundancy affects the comparative conclusions, travel time was removed from the fitness function and the GA was re-run for the three leading calibrated configurations, with weights redistributed equally across speed, headway, and overtaking rate.

The ablation runs produced the following validation MAPE values: Krauss free minGap = 9.82%, W99 locked minGap = 10.00%, and IDM free minGap = 10.70%. These results do not preserve the original four-metric ranking, which was IDM free minGap (8.79%) < W99 locked minGap (8.94%) < Krauss free minGap (10.56%). Therefore, the speed–travel-time redundancy can influence model ranking, and the original ranking should be interpreted as conditional on the four-metric fitness definition used in this study. This finding supports the recommendation that future work use Pareto-based or otherwise multi-objective calibration, such as NSGA-II [67], to reveal trade-offs among speed, travel time, headway, and overtaking rather than relying only on a single weighted-sum score.

4.16. Leave-One-Site-Out Sensitivity Analysis

To quantify sensitivity to the inclusion of individual sites, a leave-one-site-out (LOSO) analysis was conducted for the IDM free-minGap configuration. In each of six runs, one site was excluded from the GA fitness function, the remaining five sites were used for calibration, and the resulting parameter set was evaluated on the excluded site and across all six sites. This analysis was limited to the top-performing IDM free-minGap configuration; LOSO analyses for W99 and Krauss were not performed.

The out-of-sample MAPE values for the excluded sites were 12.22% (Site 01), 8.78% (Site 02), 9.82% (Site 03), 25.86% (Site 04), 15.44% (Site 05), and 4.86% (Site 06). The mean out-of-sample MAPE was 12.83% with a standard deviation of 6.66 percentage points. Overall MAPE across all six sites ranged from 10.14% to 16.57% across folds. These results indicate that aggregate performance is reasonably stable for most folds, but site-level transferability is not uniform. Site 04 is a clear outlier: when it was excluded, the out-of-sample error was 25.86% and the overall MAPE increased to 16.57%. Site 04 is distinctive because it has the longest section, longest observed travel time, and lowest total flow among the six sites. Thus, a single regional parameter set can provide useful aggregate prediction but should not be assumed to deliver uniform accuracy at every location.

5. Conclusions and Recommendations

5.1. Principal Conclusions

The GA-based calibration of three SUMO CF models across six TLTW sites in the Nile Delta yielded nine principal findings spanning model performance, parameter sensitivity, and methodological implications. Overall MAPE values for the three leading GA-calibrated configurations under the original four-metric validation were 8.79% (IDM with free minGap), 8.94% (W99 with locked minGap), and 10.56% (Krauss with free minGap). The

0.15 pp difference between IDM and W99 is smaller than the estimated IDM optimization noise from the replicate analysis, so these two configurations should be interpreted as practically equivalent rather than strictly ranked. Krauss free minGap satisfies the overall MAPE threshold but exceeds the overtaking-rate threshold. Because several calibrated parameters are correlated or near search bounds, the reported values should be interpreted as effective SUMO coefficients that reproduce aggregate field metrics under the observed conditions, not as uniquely identified physical constants of Egyptian driver behavior.

The treatment of the minimum gap parameter proved both model-dependent and structurally critical. For the IDM and Krauss models, minGap should remain free (calibrated to 3.10 m and 0.10 m, respectively). For the W99 model, minGap should be locked at 1.50 m. Locking the IDM minGap imposed the largest penalty, raising MAPE by 2.06 pp to 10.85%, and forced compensatory extreme values in delta (9.07), lcOpposite (4.80), and actionStepLength (1.0 s). Freeing the W99 minGap degraded performance by 1.57 pp to 10.51%, as the GA converged to unrealistic cc1–minGap combinations. Additional degrees of freedom do not guarantee superior calibration outcomes. Locking minGap also fundamentally altered which parameters governed overtaking: for the IDM, lcStrategic doubled in sensitivity from 113.9% to 232.7% and became the dominant parameter, whereas tau sensitivity decreased from 85.6% to 9.7%. Parameter sensitivity, therefore, depends on which parameters remain available for simultaneous optimization. Lane-change parameters, especially lcOpposite, were the most influential controls for overtaking behavior across the free-minGap models in the sensitivity analysis. The CF parameters cc3, cc5, and stepping had low overtaking sensitivity in the tested configurations. The pre-calibration LC sweep reduced the LC search space from eight parameters to three before joint CF–LC GA calibration. The travel-time ablation analysis showed that the model ranking is conditional on the selected fitness metrics; when travel time was removed and the GA was re-run, the ranking changed. Therefore, the comparative model conclusions should be interpreted together with the stated fitness-function design. SUMO's default Krauss parameters produced a competitive MAPE of 10.11% under the undersaturated, straight-section conditions examined—a result consistent with the observation that defaults work reasonably well when traffic is free-flowing and geometry is simple. This should not be extrapolated to congested or geometrically complex settings without additional validation. Nevertheless, the default MAPE substantially outperformed the prior trial-and-error result of 22.20% [59]; trial-and-error, however, is not a systematic baseline.

Site-level factors dominated over model-level differences. The same sites were consistently well predicted (Sites 02 and 06: $\leq 3\%$ error) or poorly predicted (Sites 01 and 04: highest MAPE) across all models. Site 05 chronically underpredicted overtakes by 28.6–38.6%, suggesting unmodeled factors such as localized driving norms. Cross-site transferability within the Nile Delta was reasonable but not uniform: per-site MAPE varied substantially (1.75–14.38% for IDM free), yet no site exhibited errors systematically higher than the acceptance thresholds. A single parameter set optimized across six sites cannot fit all locations equally.

All calibrated models exhibited systematic under-dispersion of run-mean standard deviations relative to observed across-vehicle standard deviations. This discrepancy reflects primarily a mismatched aggregation level: the simulated SDs are standard errors of the mean across runs, whereas observed SDs reflect across-vehicle variability within a single observation period. The uniform CF parameter set may contribute secondarily. The multi-site, multi-seed GA framework presented herein—incorporating joint CF–LC calibration, progressive simulation duration, and corrected overtake counting—provides a reproducible methodology for developing-country applications where data are scarce and traffic behavior differs from Western defaults.

5.2. Limitations

5.2.1. Data Limitations

Data were collected exclusively during midday hours (10:00–14:00) under undersaturated conditions on straight tangent sections. The calibrated parameters and relative model rankings should therefore not be extrapolated to peak periods, nighttime conditions, congested operations, horizontal curves, grades, intersections, or adverse-weather conditions without additional validation. The section lengths used in this study, 205–375 m, are sufficient for comparing aggregate speed, travel time, headway, and overtaking rates over the observed segments, but they may be too short for fully developed steady-state car-following behavior to emerge. Consequently, the calibration may emphasize free-flow, lane-change initiation, and short-duration following interactions more than sustained platoon following. All sites are in the Nile Delta; generalizability to other Egyptian regions with different topography or fleet composition is untested. Timestamping at 0.25 s resolution from manual video extraction is labor-intensive; no inter-rater reliability metric was computed, so measurement noise cannot be fully disentangled from model error. Overtaking counts are modest for a stochastic behavior, limiting statistical power when used as a calibration target.

5.2.2. Methodological Limitations

Acceleration, deceleration, and minGap were held uniform across all six vehicle types, whereas SUMO's per-class defaults vary widely (accel: 1.0–6.0 m/s², decel: 4.0–10.0 m/s²). This simplification collapses the behavioral range of the fleet into a single regime: the calibrated parameters effectively overweight the most numerous class (pickups, 39–42%) while forcing motorcycles (2–6%) and trucks (9–11%) into unrealistic spacing and acceleration regimes. The likely bias is that simulated headway variance is artificially compressed and that overtaking frequencies for fast vehicles (motorcycles) are underpredicted. Vehicle-class-specific calibration [14] is the most direct remedy but would increase dimensionality by a factor of 5–6.

The use of one CF parameter set for all vehicle classes is defensible for aggregate calibration but cannot capture class-specific dynamics such as motorcycle filtering, truck acceleration constraints, or bus/microbus stopping and merging behavior. The likely effect is that the model reproduces average site-level metrics better than the full distribution of class-specific trajectories. Future work should test at least a two-class calibration, for example light vehicles versus heavy vehicles, before attempting a full vehicle-class-specific parameter set.

Shoulder dimensions were measured but not modeled. On Egyptian TLTW roads, shoulders may be used for evasive maneuvers, slow-vehicle refuge, or partial passing. SUMO's lane-based model does not reproduce these lateral-offset behaviors unless a sublane or shoulder-use representation is explicitly configured. Therefore, calibrated LC parameters may partly compensate for unmodeled lateral freedom. Shoulder dimensions (1.0–1.5 m) were measured but not modeled. On Egyptian TLTW roads, shoulders may be used for overtaking-related evasive maneuvers, slow-vehicle refuge, or partial passing, and their omission may cause the GA to compensate through LC parameters. The LC2013 model does not handle lateral positioning within a lane; in lane-undisciplined traffic, lateral offsets affect effective capacity and spacing, and the model cannot reproduce the staggered following patterns observed in the field. Speed and travel time are algebraically related ($TT = L/v$); including both metrics in the fitness function doubles the effective weight on speed-related behavior. Although a decorrelated composite would address this, both metrics were retained because practitioners routinely measure and report them. No safety surrogate (TTC, DRAC, PET) was computed despite the head-on collision risk on TLTW

roads. Population size (15) and generations (25) were chosen for computational expediency; larger populations might discover superior solutions.

5.2.3. Model and Generalizability Limitations

Simulated standard deviations of run-mean values were smaller than observed across-vehicle standard deviations (Section 4.11), primarily due to mismatched aggregation levels rather than a model deficiency. Tau and minGap are collinear in the IDM and Krauss models [3]; individual parameter values should not be treated as uniquely identified. The parameters were calibrated on six straight tangent sections under midday, undersaturated conditions in the Nile Delta and have not been validated for curves, grades, urban settings, peak congestion, or different fleet compositions.

5.3. Recommendations for Future Research

Based on the findings and limitations of this study, the following directions are recommended for future research on traffic simulation calibration in developing-country contexts:

- Weather effects: Incorporate weather conditions into the simulation to address the anomalous overtaking underprediction observed at Site 05.
- Vehicle-class-specific calibration: Calibrate separate parameter sets per vehicle class [14]; at minimum, motorcycles and trucks should be assigned distinct driver classes to better capture the behavioral heterogeneity underlying observed variance. Vehicle-class-specific calibration and lateral behavior: Future studies should evaluate at least a two-class CF calibration, separating light and heavy vehicles, and should test whether SUMO's sublane model or an explicit shoulder-use representation improves overtaking prediction on Egyptian TLTW roads. These extensions require field data that classify maneuvers by vehicle class and by overtaking type, including full opposite-lane passing, partial passing, and shoulder-assisted maneuvers. Expanded data collection: Extend data collection to peak and nighttime periods, adverse weather conditions, and sites outside the Nile Delta that include horizontal curves and grades. Trajectory-level validation: Extend validation to individual trajectories (speed profiles, acceleration patterns, time–space diagrams). Safety surrogate measures: Compute safety surrogates (TTC, DRAC, PET) from simulated trajectories to verify that calibrated parameters do not produce unrealistically safe or unsafe interactions on TLTW roads. Alternative optimization algorithms: Compare the GA with multi-objective Pareto approaches, Bayesian optimization, particle swarm optimization, and surrogate methods. A Pareto-based analysis, such as NSGA-II [67], would reveal speed–overtaking trade-offs masked by the current weighted-sum fitness. Extended search ranges: Expand LC parameter bounds: lcSpeedGain reached the upper bound (5.0) for the IDM, and lcOpposite reached the upper bound (5.0) in the free-minGap W99 configuration; extending the search range to [0, 10] is warranted.
- Aborted overtake rate: Include aborted overtake rate as a separate calibration objective or constraint to improve safety realism.
- Parameter identifiability analysis: Report parameter correlation matrices or pair plots from the GA population to assess tau–minGap identifiability and guide parameter reduction.
- Minimum section length criterion: Adopt a minimum section length of at least 500 m for future calibration studies to ensure that CF parameters are identifiable from steady-state following episodes rather than exclusively from free-flow and lane-change transitions.

- Uncertainty quantification: Report calibrated parameter distributions (not only point estimates) from the GA population to communicate uncertainty and identifiability.
- Per-vehicle variability: Compute simulated standard deviations from individual vehicle trajectories within a single representative run (rather than from run-mean values) to enable a fair comparison with observed across-vehicle variability and to properly quantify the models' ability to reproduce distributional spread.
- Develop and release an open-source Python package for SUMO GA calibration, including the GA configuration, SUMO network files, and extracted aggregate metrics per site.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/su18147025/s1>, IDM FREE (minGap free, 10 params, 13 gens, early stop); IDM LOCKED (minGap = 1.50, 9 params, 14 gens, early stop); KRAUSS FREE (minGap free, 10 params, 21 gens, early stop); KRAUSS LOCKED (minGap=1.50, 9 params, 16 gens, early stop); W99 LOCKED (minGap = 1.50, 7 params, 19 gens, early stop); W99 FREE (minGap free, 8 params, 14 gens, early stop).

Author Contributions: Conceptualization, U.E.S., M.O. and S.M.E.-B.; methodology, U.E.S., M.O. and A.S.; software, U.E.S.; validation, U.E.S., M.O. and S.M.E.-B.; formal analysis, U.E.S.; investigation, U.E.S.; resources, A.S. and S.E.; data curation, U.E.S., A.S. and S.E.; writing—original draft preparation, U.E.S.; writing—review and editing, M.O., S.M.E.-B., A.S. and S.E.; visualization, U.E.S.; supervision, M.O. and S.M.E.-B.; project administration, M.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and codes used in this analysis are available from the corresponding authors upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Afshari, A.; Lee, J.; Besenski, D.; Dimitrijevic, B.; Spasovic, L. Calibrating Microscopic Traffic Simulation Model Using Connected Vehicle Data and Genetic Algorithm. *Appl. Sci.* **2025**, *15*, 1496. [\[CrossRef\]](#)
2. Lopez, P.A.; Wiessner, E.; Behrisch, M.; Bieker-Walz, L.; Erdmann, J.; Flotterod, Y.-P.; Hilbrich, R.; Lucken, L.; Rummel, J.; Wagner, P. Microscopic Traffic Simulation Using SUMO. In *Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC)*; IEEE: Maui, HI, USA, 2018; pp. 2575–2582.
3. Punzo, V.; Zheng, Z.; Montanino, M. About Calibration of Car-Following Dynamics of Automated and Human-Driven Vehicles: Methodology, Guidelines and Codes. *Transp. Res. Part C Emerg. Technol.* **2021**, *128*, 103165. [\[CrossRef\]](#)
4. Chowdhury, M.M.H.; Chakraborty, T. Calibration of SUMO Microscopic Simulation for Heterogeneous Traffic Condition: The Case of the City of Khulna, Bangladesh. *Transp. Eng.* **2024**, *18*, 100281. [\[CrossRef\]](#)
5. Hashim, I.H.; Abdel-Wahed, T.A. Evaluation of Performance Measures for Rural Two-Lane Roads in Egypt. *Alex. Eng. J.* **2011**, *50*, 245–255. [\[CrossRef\]](#)
6. Hashim, I.H. Analysis of Speed Characteristics for Rural Two-Lane Roads: A Field Study from Minoufiya Governorate, Egypt. *Ain Shams Eng. J.* **2011**, *2*, 43–52. [\[CrossRef\]](#)
7. Hashim, I.H.; Abdel-Wahed, T.A.; Moustafa, Y. Toward an Operating Speed Profile Model for Rural Two-Lane Roads in Egypt. *J. Traffic Transp. Eng. Engl. Ed.* **2016**, *3*, 82–88. [\[CrossRef\]](#)
8. Shoaeb, A.; El-Badawy, S.; Shawly, S.; Shahdah, U.E. Time Headway Distributions for Two-Lane Two-Way Roads: Case Study from Dakahlia Governorate, Egypt. *Innov. Infrastruct. Solut.* **2021**, *6*, 165. [\[CrossRef\]](#)
9. Shoaeb, A.; El-Badawy, S.M.; Shawly, S.; Shahdah, U.E. A Proposed Overtaking Model for Egyptian Two-Lane Two-Way Roads. *IATSS Res.* **2024**, *48*, 275–287. [\[CrossRef\]](#)

10. Krauss, S.; Wagner, P.; Gawron, C. Metastable States in a Microscopic Model of Traffic Flow. *Phys. Rev. E* **1997**, *55*, 5597–5602. [\[CrossRef\]](#)
11. Krauß, S. Microscopic Modeling of Traffic Flow: Investigation of Collision Free Vehicle Dynamics. Doctoral Dissertation, University of Cologne, Cologne, Germany, 1998.
12. Schrader, M.; Karnik, A.; Hainen, A.; Bittle, J. Calibrating Car-Following Models Using SUMO-in-the-Loop and Vehicle Trajectories From Roadside Radar: Calibrating CF Model Parameters. *SUMO Conf. Proc.* **2024**, *5*, 209–233. [\[CrossRef\]](#)
13. Wiedemann, R. *Simulation des Strassenverkehrsflusses*; Schriftenreihe des Instituts für Verkehrswesen der Universität Karlsruhe; Institut für Verkehrswesen, Universität Karlsruhe: Karlsruhe, Germany, 1974.
14. Chaudhari, A.A.; Srinivasan, K.K.; Chilukuri, R.B.; Treiber, M.; Okhrin, O. Calibrating Wiedemann-99 Model Parameters to Trajectory Data of Mixed Vehicular Traffic. *Transp. Res. Rec. J. Transp. Res. Board* **2022**, 2676, 718–735. [\[CrossRef\]](#)
15. Treiber, M.; Hennecke, A.; Helbing, D. Congested Traffic States in Empirical Observations and Microscopic Simulations. *Phys. Rev. E* **2000**, *62*, 1805–1824. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Treiber, M.; Hennecke, A.; Helbing, D. Microscopic Simulation of Congested Traffic. In *Traffic and Granular Flow '99*; Helbing, D., Herrmann, H.J., Schreckenberg, M., Wolf, D.E., Eds.; Springer: Berlin/Heidelberg, Germany, 2000; pp. 365–376. ISBN 978-3-642-64109-1.
17. Kesting, A.; Treiber, M. Calibrating Car-Following Models by Using Trajectory Data: Methodological Study. *Transp. Res. Rec. J. Transp. Res. Board* **2008**, 2088, 148–156. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Cascan, E.T.; Ivanchev, J.; Eckhoff, D.; Sangiovanni-Vincentelli, A.; Knoll, A. Multi-Objective Calibration of Microscopic Traffic Simulation for Highway Traffic Safety. In *Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference (ITSC)*; IEEE: Auckland, New Zealand, 2019; pp. 4548–4555.
19. Flitsch, C.; Kastner, K.-H.; Bósa, K.; Neubauer, M. Calibrating Traffic Simulation Models in SUMO Based upon Diverse Historical Real-Time Traffic Data—Lessons Learned in ITS Upper Austria. In *Proceedings of the EPiC Series in Engineering*; German Aerospace Center (DLR): Berlin, Germany, 2018; Volume 2, pp. 25–42.
20. Antoniou, C.; Lima Azevedo, C.; Lu, L.; Pereira, F.; Ben-Akiva, M. W-SPSA in Practice: Approximation of Weight Matrices and Calibration of Traffic Simulation Models. *Transp. Res. Part C Emerg. Technol.* **2015**, *59*, 129–146. [\[CrossRef\]](#)
21. Punzo, V.; Ciuffo, B. How Parameters of Microscopic Traffic Flow Models Relate to Traffic Dynamics in Simulation: Implications for Model Calibration. *Transp. Res. Rec. J. Transp. Res. Board* **2009**, 2124, 249–256. [\[CrossRef\]](#)
22. Lownes, N.E.; Machemehl, R.B. Sensitivity of Simulated Capacity to Modification of VISSIM Driver Behavior Parameters. *Transp. Res. Rec. J. Transp. Res. Board* **2006**, 1988, 102–110. [\[CrossRef\]](#)
23. Liang, Q.; Wan, Q.; Bai, L.; Yu, H.; Lv, L.; Li, D. Sensitivity of Simulated Conflicts to VISSIM Driver Behavior Parameter Modification. In *Green, Smart and Connected Transportation Systems*; Wang, W., Baumann, M., Jiang, X., Eds.; Lecture Notes in Electrical Engineering; Springer: Singapore, 2020; Volume 617, pp. 113–122, ISBN 978-981-15-0643-7.
24. Ge, Q.; Menendez, M. An Efficient Sensitivity Analysis Approach for Computationally Expensive Microscopic Traffic Simulation Models. *Int. J. Transp.* **2014**, *2*, 49–64. [\[CrossRef\]](#)
25. Muhan, N.; Qin, Y.; Zhang, Q.; Yang, Y.; Zhang, Z. Parameter Calibration of VISSIM Simulation Model Based on Genetic Algorithm. In *Proceedings of the 2013 International Conference on Advanced Computer Science and Electronics Information*; Atlantis Press: Beijing, China, 2013.
26. Cheu, R.-L.; Jin, X.; Ng, K.-C.; Ng, Y.-L.; Srinivasan, D. Calibration of FRESIM for Singapore Expressway Using Genetic Algorithm. *J. Transp. Eng.* **1998**, *124*, 526–535. [\[CrossRef\]](#)
27. Park, B.; Schneeberger, J.D. Microscopic Simulation Model Calibration and Validation: Case Study of VISSIM Simulation Model for a Coordinated Actuated Signal System. *Transp. Res. Rec. J. Transp. Res. Board* **2003**, 1856, 185–192. [\[CrossRef\]](#)
28. Park, B.; Qi, H. Development and Evaluation of a Procedure for the Calibration of Simulation Models. *Transp. Res. Rec. J. Transp. Res. Board* **2005**, 1934, 208–217. [\[CrossRef\]](#)
29. Cobos, C.; Paz, A.; Luna, J.; Erazo, C.; Mendoza, M. A Multi-Objective Approach for the Calibration of Microscopic Traffic Flow Simulation Models. *IEEE Access* **2020**, *8*, 103124–103140. [\[CrossRef\]](#)
30. Chiappone, S.; Giuffrè, O.; Granà, A.; Mauro, R.; Sferlazza, A. Traffic Simulation Models Calibration Using Speed–Density Relationship: An Automated Procedure Based on Genetic Algorithm. *Expert Syst. Appl.* **2016**, *44*, 147–155. [\[CrossRef\]](#)
31. Hale, D.K.; Antoniou, C.; Brackstone, M.; Michalaka, D.; Moreno, A.T.; Parikh, K. Optimization-Based Assisted Calibration of Traffic Simulation Models. *Transp. Res. Part C Emerg. Technol.* **2015**, *55*, 100–115. [\[CrossRef\]](#)
32. Lee, J.-B.; Ozbay, K. New Calibration Methodology for Microscopic Traffic Simulation Using Enhanced Simultaneous Perturbation Stochastic Approximation Approach. *Transp. Res. Rec. J. Transp. Res. Board* **2009**, 2124, 233–240. [\[CrossRef\]](#)
33. Oh, S.; Seshadri, R.; Azevedo, C.L.; Ben-Akiva, M.E. Demand Calibration of Multimodal Microscopic Traffic Simulation Using Weighted Discrete SPSA. *Transp. Res. Rec. J. Transp. Res. Board* **2019**, 2673, 503–514. [\[CrossRef\]](#)
34. Paz, A.; Molano, V.; Martinez, E.; Gaviria, C.; Arteaga, C. Calibration of Traffic Flow Models Using a Memetic Algorithm. *Transp. Res. Part C Emerg. Technol.* **2015**, *55*, 432–443. [\[CrossRef\]](#)

35. Hou, Z.; Lee, J. Multi-Thread Optimization for the Calibration of Microscopic Traffic Simulation Model. *Transp. Res. Rec. J. Transp. Res. Board* **2018**, 2672, 98–109. [[CrossRef](#)]
36. Yu, M.; Fan, W. Calibration of Microscopic Traffic Simulation Models Using Metaheuristic Algorithms. *Int. J. Transp. Sci. Technol.* **2017**, 6, 63–77. [[CrossRef](#)]
37. Guastella, D.A.; Morales-Hernández, A.; Cornelis, B.; Bontempi, G. Calibration of Vehicular Traffic Simulation Models by Local Optimization. *Transportation* **2025**. [[CrossRef](#)]
38. Zhang, J.; Lin, Z.; Wang, C. Parameter Correction of VISSIM Multi-Intersection Simulation Model Based on Combined Genetic Algorithm. *IOP Conf. Ser. Mater. Sci. Eng.* **2019**, 688, 033018. [[CrossRef](#)]
39. Cipriani, E.; Florian, M.; Mahut, M.; Nigro, M. A Gradient Approximation Approach for Adjusting Temporal Origin–Destination Matrices. *Transp. Res. Part C Emerg. Technol.* **2011**, 19, 270–282. [[CrossRef](#)]
40. Cheng, Q.; Wang, S.; Liu, Z.; Yuan, Y. Surrogate-Based Simulation Optimization Approach for Day-to-Day Dynamics Model Calibration with Real Data. *Transp. Res. Part C Emerg. Technol.* **2019**, 105, 422–438. [[CrossRef](#)]
41. Patwary, A.U.Z.; Huang, W.; Lo, H.K. Metamodel-Based Calibration of Large-Scale Multimodal Microscopic Traffic Simulation. *Transp. Res. Part C Emerg. Technol.* **2021**, 124, 102859. [[CrossRef](#)]
42. Sha, D.; Ozbay, K.; Ding, Y. Applying Bayesian Optimization for Calibration of Transportation Simulation Models. *Transp. Res. Rec. J. Transp. Res. Board* **2020**, 2674, 215–228. [[CrossRef](#)]
43. Yousefi, M.; Ghoreishi, S.F. Bayesian Calibration of Simulation-Based Analysis of Transportation Systems. In *Proceedings of the International Conference on Transportation and Development 2022*; American Society of Civil Engineers: Seattle, WA, USA, 2022; pp. 168–179.
44. Zhang, C.; Sun, L. Bayesian Calibration of the Intelligent Driver Model. *IEEE Trans. Intell. Transp. Syst.* **2024**, 25, 9308–9320. [[CrossRef](#)]
45. Liu, Y.; Zou, B.; Ni, A.; Gao, L.; Zhang, C. Calibrating Microscopic Traffic Simulators Using Machine Learning and Particle Swarm Optimization. *Transp. Lett.* **2021**, 13, 295–307. [[CrossRef](#)]
46. Daguano, R.F.; Yoshioka, L.R.; Netto, M.L.; Marte, C.L.; Isler, C.A.; Santos, M.M.D.; Justo, J.F. Automatic Calibration of Microscopic Traffic Simulation Models Using Artificial Neural Networks. *Sensors* **2023**, 23, 8798. [[CrossRef](#)] [[PubMed](#)]
47. Owais, M.; Abulwafa, O.; Abbas, Y.A. When to Decide to Convert a Roundabout to a Signalized Intersection: Simulation Approach for Case Studies in Jeddah and Al-Madinah. *Arab. J. Sci. Eng.* **2020**, 45, 7897–7914. [[CrossRef](#)]
48. Harth, M.; Langer, M.; Bogenberger, K. Automated Calibration of Traffic Demand and Traffic Lights in SUMO Using Real-World Observations. *SUMO Conf. Proc.* **2022**, 2, 133–148. [[CrossRef](#)]
49. Stang, C.; Bogenberger, K. Calibration of Microscopic Traffic Simulation in an Urban Environment Using GPS-Data. *SUMO Conf. Proc.* **2024**, 5, 71–78. [[CrossRef](#)]
50. Vishnoi, S.; Shetty, A.; Tsogsuren, I.; Arora, N.; Osorio, C. On the Use of Abundant Road Speed Data for Travel Demand Calibration of Urban Traffic Simulators. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*; ACM: Hamburg, Germany, 2023; pp. 1–4.
51. Kappenberger, J.; Stuckenschmidt, H. Overcoming Data Scarcity in Calibrating SUMO Scenarios With Evolutionary Algorithms. *SUMO Conf. Proc.* **2025**, 6, 133–148. [[CrossRef](#)]
52. Siam, M.R.K.; Nasrin, S.; Hadiuzzaman, M.; Muniruzzaman, S.M.; Haque, N. VISCAL: Heuristic Algorithm Based Application Tool to Calibrate Microscopic Simulation Parameters. *J. Traffic Transp. Eng. Engl. Ed.* **2018**, 5, 28–43. [[CrossRef](#)]
53. Salles, D.; Kaufmann, S.; Reuss, H.-C. Extending the Intelligent Driver Model in SUMO and Verifying the Drive Off Trajectories with Aerial Measurements. *SUMO Conf. Proc.* **2022**, 1, 1–25. [[CrossRef](#)]
54. Kathis, H.; Keler, A.; Bogenberger, K. Calibrating the Wiedemann 99 Car-Following Model for Bicycle Traffic. *Sustainability* **2021**, 13, 3487. [[CrossRef](#)]
55. Durrani, U.; Lee, C.; Maoh, H. Calibrating the Wiedemann’s Vehicle-Following Model Using Mixed Vehicle-Pair Interactions. *Transp. Res. Part C Emerg. Technol.* **2016**, 67, 227–242. [[CrossRef](#)]
56. Pourabdollah, M.; Bjarkvik, E.; Furer, F.; Lindenberg, B.; Burgdorf, K. Calibration and Evaluation of Car Following Models Using Real-World Driving Data. In *Proceedings of the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*; IEEE: Yokohama, Japan, 2017; pp. 1–6.
57. Kong, M.; Zhang, C.; Sun, L. Active Simulation-Based Inference for Scalable Car-Following Model Calibration. *arXiv* **2026**, arXiv:2602.05246.
58. Hoque, I.; Ananda, T.N.; Anowar, P.; Naz, A.; Murshed, M.N. Machine Learning Approach in Calibrating VISSIM Microsimulation Model for Mixed Traffic Conditions. *J. Eng. Sci.* **2025**, 16, 21–30. [[CrossRef](#)]
59. Shoaeb, A. Traffic Operation and Safety of Two-Lane Two-Way Roads in Egypt. Ph.D. Thesis, Mansoura University, Mansoura, Egypt, 2024.
60. Farah, H.; Bekhor, S.; Polus, A. Risk Evaluation by Modeling of Passing Behavior on Two-Lane Rural Highways. *Accid. Anal. Prev.* **2009**, 41, 887–894. [[CrossRef](#)] [[PubMed](#)]

61. Ciuffo, B.; Punzo, V.; Montanino, M. *The Calibration of Traffic Simulation Models—Report on the Assessment of Different Goodness of Fit Measures and Optimization Algorithms*; Publications Office of the European Union: Luxembourg, 2012; p. 92.
62. Arafat, M.; Nafis, S.R.; Sadeghvaziri, E.; Tousif, F. A Data-Driven Approach to Calibrate Microsimulation Models Based on the Degree of Saturation at Signalized Intersections. *Transp. Res. Interdiscip. Perspect.* **2020**, *8*, 100231. [[CrossRef](#)]
63. Sabry, M. *Headway Distribution Model And Interrelationships Between Headway And Fundamental Traffic Flow Characteristics*; Ain Shams University: Cairo, Egypt, 1989.
64. Transportation Research Board (TRB). *Highway Capacity Manual 6th Edition: A Guide for Multimodal Mobility Analysis*; Elefteriadou, L.A., Ed.; The National Academies Press: Washington, DC, USA, 2016; ISBN 978-0-309-36997-8.
65. Transportation Research Board (TRB); National Academies of Sciences, Engineering, and Medicine (NASEM). *Highway Capacity Manual 7th Edition: A Guide for Multimodal Mobility Analysis*; The National Academies Press: Washington, DC, USA, 2022; ISBN 978-0-309-08766-7.
66. German Aerospace Center (DLR). SUMO Documentation. Available online: <https://sumo.dlr.de/docs/index.html> (accessed on 22 May 2026).
67. Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **2002**, *6*, 182–197. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.