Data-Driven Last-Mile Shuttle Service Design: A Precision Matching and Service Optimization Model

1st Di Lu
Beijing Normal University
Beijing, China
1132589068@qq.com

2nd Hao Sun
University of Southampton
Southampton, United Kingdom
zhoupeitong321@gmail.com

Abstract*—With the acceleration of urbanization and the rise of large-scale industrial parks, commuter traffic pressure is mounting, particularly the "last-mile" problem from public transport hubs to final destinations, which has become a critical bottleneck constraining urban mobility efficiency and sustainable development. Existing last-mile solutions, such as fixed-route shuttles, commonly suffer from service rigidity, resource wastage, and supply-demand mismatch. To address this challenge, this study proposes a new paradigm for last-mile shuttle design based on dynamic data, aiming to enhance service efficiency and user satisfaction through precision matching and service optimization. Taking the Shenzhen High-Tech Industrial Park as an empirical case, this research constructs a three-stage integrated design framework utilizing large-scale ride-hailing order data, Points of Interest (POI) data, and built environment data. First, an improved DBSCAN clustering algorithm and a spatio-temporal analysis model are employed to accurately identify and predict dynamically changing travel demands. Second, a Passenger-Stop-Vehicle (PSV) three-level precision matching model is proposed to achieve effective alignment between personalized demands and service resources. Finally, a multi-objective optimization model is formulated with the goals of minimizing operating costs, minimizing total passenger travel time, and maximizing system service coverage. An improved Genetic Algorithm (GA) is used to coordinately optimize the shuttle stop layout, routes, and schedules. Simulation experiments and comparative analysis demonstrate that the proposed model, compared to the traditional fixed-route model, can reduce average waiting times by approximately 36.3%, decrease vehicle deadheading rates by 25.4%, and improve overall operational profit by 96.3%. This research not only provides a data-driven, intelligent solution for the last-mile transportation problem in industrial parks but also offers theoretical support and practical reference for building more efficient and resilient urban public transport systems.

Keywords—Last-Mile Transportation, Shuttle Service Design, Precision Matching, Data-Driven, Service Optimization

I. INTRODUCTION

In recent decades, the spatial evolution of cities — especially the interplay between suburbanization and employment decentralization — has reshaped commuting patterns and intensified peak-period travel demand in metropolitan areas [1].

Classical urban spatial structure theory further indicates that industrial clustering and functional integration can generate highly concentrated directional flows, placing persistent pressure on urban transport systems and their connecting links [2]. Within this context, last-mile mobility (i.e., the connection from major transit hubs to final activity locations such as workplaces or residences) is widely regarded as a critical determinant of public transport attractiveness and overall service quality. Inadequate performance in this segment often aggravates car dependence and related externalities, ultimately undermining urban livability [3].

The last-mile problem is particularly pronounced in large industrial parks, where extensive spatial scale and high employment density induce tidal commuter flows. Traditional solutions predominantly rely on fixed-route, fixed-schedule shuttle services. Although such services can meet basic commuting needs, their "one-size-fits-all" operation struggles to adapt to heterogeneous and time-varying demand, resulting in long waiting times, excessive walking distances, and unstable travel times for passengers. Meanwhile, operators lacking accurate demand awareness may experience low vehicle utilization and high deadheading, which increases operational costs and weakens service sustainability.

Motivated by these challenges, this study aims to construct a data-driven last-mile shuttle design and optimization framework with a focus on integrating precision matching and service optimization. Specifically, this research seeks: (1) to build an end-to-end design pipeline that transforms multi-source mobility data into a complete shuttle plan (stops, routes, and schedules); (2) to propose a Passenger – Stop – Vehicle (PSV) three-level precision matching mechanism to better align service resources with dynamic spatiotemporal demand; and (3) to develop a multi-objective coordinated optimization model that jointly improves passenger experience, operator cost-effectiveness, and service coverage. A case study in the Shenzhen High-Tech Industrial Park is conducted to validate the effectiveness and practical value of the proposed framework using anonymized ride-hailing travel data.

Corresponding Author: Hao Sun,University of Southampton, University Road, Highfield, Southampton, SO17 1BJ, Southampton, United Kingdom,zhoupeitong321@gmail.com

The remainder of this paper is organized as follows: Section 2 reviews relevant literature. Section 3 presents the proposed methodology. Section 4 introduces the case study and experimental design. Section 5 reports results and analyses. Section 6 discusses implications. Section 7 concludes the paper and outlines future research directions.

II. Literature Review

To establish a solid theoretical foundation, this section reviews the literature in three areas: (1) the challenges and solution pathways for last-mile transportation systems; (2) shuttle service design and optimization techniques; and (3) the application of data-driven methods in public transport planning and operations.

A. The Last-Mile Transportation System

Last-mile connectivity is widely recognized as a critical factor affecting the effectiveness and attractiveness of public transport systems, because it directly shapes passengers ' access/egress costs and perceived service quality [4]. From the perspective of sustainable mobility, deficiencies in last-mile service can reinforce private car dependency and its negative externalities, which undermines broader sustainability goals [5]. Structural incentives, such as parking-related factors, may further strengthen car-oriented travel choices when last-mile services are weak [6]. In response, shared collective transport — especially shuttle services — has been regarded as a promising approach to balance capacity, cost, and service quality for concentrated commuting demand [7].

B. Shuttle Service Design and Optimization

With the rise of large-scale mobility data, researchers have demonstrated that individual travel exhibits measurable regularities, enabling demand characterization beyond traditional surveys [8]. In public transport contexts, the development of passenger flow forecasting has been extensively surveyed, highlighting mainstream modeling paradigms and challenges for operational decision-making [9]. More broadly, data-intensive urban science emphasizes that massive datasets can reshape how cities are understood and managed, providing methodological support for data-driven transport planning [10].

C. Data-Driven Public Transport Planning

Shuttle service design is typically formulated as a multi-objective and multi-constraint combinatorial optimization problem, often involving stop location, routing, and scheduling. For stop location, classical accessibility and coverage-oriented formulations aim to improve system accessibility while expanding service coverage [11]. For routing, many shuttle planning problems can be framed as variants of the Vehicle Routing Problem (VRP), including time-window and pickup-and-delivery extensions, which are well known to be NP-hard and thus frequently solved using heuristic or metaheuristic methods [12].

Beyond fixed-route services, last-mile solutions also include non-motorized and individualized motorized options. Bike sharing and cycling can serve short-distance access/egress and may reduce car use under suitable conditions, but their effectiveness is limited by context and operational constraints [13]. Ride-hailing provides flexible door-to-door mobility, yet large-scale adoption may increase congestion and emissions and introduce broader policy concerns [14]. In practice, public transit planning and operation also involve behavioral and operational considerations that shape the feasibility of integrating last-mile services into the wider transit system [15].

To better respond to dynamic and heterogeneous demand, Demand Responsive Transit (DRT) has emerged as a flexible supplement between fixed-route buses and taxi-like services [16]. Classic dial-a-ride research has examined dynamic request handling and routing decisions using approaches such as dynamic programming, providing foundational insights for responsive dispatching problems [17]. Meanwhile, recent reviews summarize multi-objective optimization perspectives for smart-city public transport route planning, emphasizing integrated design considerations and practical constraints [18]. In addition, learning-augmented optimization has gained attention as a way to improve the control of local search and other heuristic processes in large-scale combinatorial problems [19]. For scheduling and frequency-setting, simulation-based approaches have been used to capture behavioral responses and mode substitution effects under frequency adjustments, offering more realistic evaluation for service planning [20].

D. Data-Driven Public Transport Planning in Practice

Data-driven public transport planning also benefits from empirical spatiotemporal analyses of shared mobility trajectories, which can help identify demand hotspots and service gaps relevant to last-mile design [21]. Deep learning methods have been proposed to improve passenger flow prediction accuracy by incorporating spatial features, supporting more informed operational adjustments [22]. Moreover, emerging ITS research discusses how sensing and edge computing can support responsive and scalable transport systems, while also outlining technical challenges for real-world deployment [23].

In summary, existing studies provide valuable foundations for last-mile solutions, shuttle optimization, and data-driven planning. However, an important research gap remains: a comprehensive end-to-end framework that integrates dynamic demand discovery, precision supply − demand matching, and systematic service optimization for last-mile shuttle services.

III. Problem Formulation and Modeling

The core of this research is to design an efficient and flexible last-mile shuttle service system for an industrial park. This system must be capable of responding to dynamically changing travel demands while ensuring service quality and operational sustainability. To this end, we construct a three-stage integrated design framework that covers the entire process from demand discovery to service optimization. This chapter will first formally describe the research problem and then detail the model formulation for each stage of the framework.

A. Problem Description

The last-mile shuttle service design problem can be defined as follows: given a specific service area (an industrial park), a major public transport hub (a metro station), a set of potential passenger travel demands, and a fleet of available shuttle resources (vehicles), how to determine an optimal set of shuttle service plans—including Stop Layout (S), Routes (R), and Schedules (T)—to achieve predefined optimization objectives.

This problem is characterized by:

- Demand Dynamism: Passenger travel demand exhibits high dynamism in both time and space, with pronounced tidal phenomena during morning and evening commuting peaks.
- Multi-Objectivity: The service design must balance multiple, often conflicting, objectives, including passenger interests (reducing waiting and in-vehicle time), operator interests (lowering operational costs), and social benefits (expanding service coverage).
- Complex Constraints: The design solution must satisfy a series of real-world constraints, such as vehicle capacity, passengers' maximum acceptable walking distance and waiting time, and the physical limitations of the road network.

Based on this, we formulate the problem as a multi-objective combinatorial optimization problem, defined as follows:

Inputs:

- Demand Point Set D: A set of all last-mile travel demand points, D = {d_1, d_2, ..., d_n}. Each demand point d_i includes its geographical coordinates, demand occurrence time t_i, and the number of people (usually 1).
- Transport Hub H: The geographical location of the main public transport hub within the service area.
- Road Network G: G = (V, E), a graph representing the road network in the service area, where V is the set of nodes (intersections) and E is the set of edges (road segments), with each edge associated with a weight representing travel time or distance.
- Vehicle Fleet K: K = {k_1, k_2, ..., k_m}, a set of available shuttle vehicles, each with the same capacity Q.

Decision Variables:

- Stop Set S: S = {s_1, s_2, ..., s_p}, a subset of actual operating shuttle stops selected from a set of candidate stops.
- Route Set R: R = {r_1, r_2, ..., r_q}, a set of shuttle routes. Each route r_j is an ordered sequence of nodes consisting of the hub H and a subset of stops s_i.
- Schedule Plan T: T = {t_1, t_2, ..., t_q}, the timetable of departure times corresponding to each route r_j.

Objectives:

- Minimize Total Operating Cost (Z1): Primarily proportional to the total travel distance or time of the vehicles.
- Minimize Total Passenger Travel Time (Z2): Includes passengers' access time walking from demand points to stops, waiting time at stops, and in-vehicle travel time.
- Maximize Service Coverage (Z3): The proportion of successfully served travel demand points to the total number of demand points.

B. Research Framework

To systematically address the problem described above, we propose a three-stage data-driven framework as shown in Figure 1. This framework decomposes the complex design problem into three logically sequential yet interconnected modules: Dynamic Demand Discovery, Precision Supply-Demand Matching, and Service Coordinated Optimization.

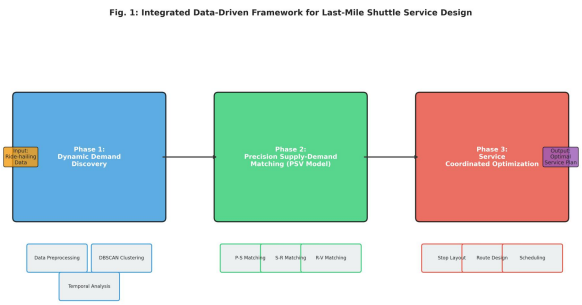

Fig. 1. Integrated Data-Driven Framework for Last-Mile Shuttle Service Design

- Stage 1: Dynamic Demand Discovery and Prediction. The goal of this stage is to extract and analyze the dynamic spatio-temporal characteristics of last-mile travel demand from raw, multi-source data. It begins with data cleaning, standardization, and spatio-temporal correlation. Then, it utilizes an improved clustering algorithm to identify high-density demand areas and analyzes their evolution patterns across different time scales (hourly, daily, weekly), providing precise demand inputs for the subsequent service design.
- Stage 2: Precision Supply-Demand Matching Model. This stage acts as a bridge connecting "demand" and "supply," aiming to establish an effective matching relationship among passengers, stops, and vehicles. We innovatively propose a "Passenger-Stop-Vehicle" (PSV) three-level matching model, which decomposes the macroscopic service optimization problem into microscopic matching decisions, ensuring that service resources can precisely respond to individualized travel needs.
- Stage 3: Service Coordinated Optimization Model. This is the core decision-making module of the framework. Based on the matching relationships established in the second stage, it constructs a multi-objective optimization model. It then uses an intelligent optimization algorithm (such as a Genetic Algorithm) to perform an integrated, coordinated optimization of the three key elements of the shuttle service — stop layout, routes, and schedules — to generate the final service plan.

C. Model Formulation

1) Dynamic Demand Discovery Model

Accurate demand discovery is the foundation for all subsequent optimization. We adopt the following steps to extract dynamic demand from raw data:

Data Preprocessing and Standardization: Raw ride-hailing order data is first cleaned to remove abnormal and missing records. The data is then converted into a standard format including order ID, pickup time, drop-off time,

pickup location (latitude/longitude), and drop-off location (latitude/longitude), as shown in Table I. Based on the research scenario, we filter orders that end within the industrial park during specific time windows (morning peak, 7:00-10:00) or start within the park during the evening peak (17:00-20:00) to form the candidate dataset for last-mile travel.

TABLE I. STANDARDIZED TRAVEL RECORD FIELDS USED IN THIS STUDY

Field Name	Type	Description
user id	String	Anonymous unique user identifier
start time	Datetime	Trip start time
end time	Datetime	Trip end time
origin _lon	Double	Origin longitude
origin_la t	Double	Origin latitude
dest_lon	Double	Destination longitude
dest lat	Double	Destination latitude
mode	String	Transportation mode (e.g. 'subway', 'bike')

Demand Hotspot Identification: To identify spatially clustered demand areas, we employ the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm. Compared to algorithms like K-Means, which require a predefined number of clusters, DBSCAN can discover clusters of arbitrary shape and effectively identify noise points (discrete demands), making it highly suitable for handling spatially uneven travel demand points . DBSCAN requires two parameters: eps and MinPts. We determine them using a k-distance elbow method with k = MinPts, and then fix the final values for the case study to eps = 250 m and MinPts = 60. Distances are computed in meters after projecting coordinates to a local metric coordinate system. We also conduct a sensitivity check by varying eps in [200 m, 300 m] and MinPts in [40, 80] and confirm that the set of major hotspots and downstream optimization conclusions remain stable.

After the algorithm is executed, spatially dense travel demand points are grouped into different clusters. The core points of each cluster represent a high-frequency demand "hotspot area." The centroids of these hotspots will serve as candidate locations for subsequent stop placement.

Temporal Feature Analysis: For each identified demand cluster, we further analyze its distribution patterns over time. By counting the demand volume within each cluster during different time slices (every 30 minutes), we can plot time-series curves for each demand area. These curves clearly reveal the periodicity, peak hours, and volatility of demand, providing a critical basis for designing a dynamic dispatching schedule.

2) Precision Supply-Demand Matching Model (PSV Matching)

The precision matching model aims to establish the connection that gets the "right people" (passengers) to the "right place" (stop) at the "right time" to board the "right vehicle" (trip). We decompose this into three levels(Figure 2):

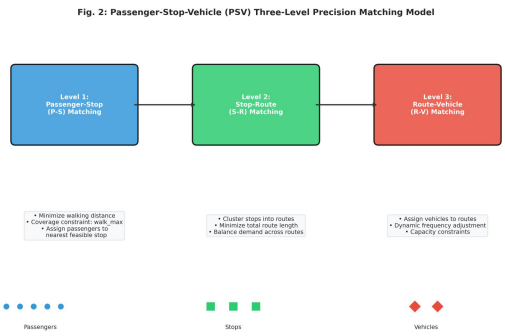

Fig. 2. Passenger-Stop-Vehicle (PSV) Three-Level Precision Matching Model

- Level 1: Passenger-Stop (P-S) Matching. This level matches each travel demand point d_i to one or more potential service stops. The matching principle is the shortest walking distance. We set a maximum acceptable walking distance, walk_max (300 meters). For each demand point d_i, we calculate its walking distance to all candidate stops s_j and assign it to the nearest stop within the walk_max threshold. This step aggregates discrete individual demands to candidate stops, forming the demand volume for each stop at different time slices.
- Level 2: Stop-Route (S-R) Matching. This level combines the demand-laden candidate stops into several shuttle routes. This is a variant of the classic Traveling Salesperson Problem (TSP) or Vehicle Routing Problem (VRP). The objective is to design a set of routes that can connect multiple stops with the highest efficiency. The quality of the match depends on the total route length, the number of stops covered, and the total demand served.
- Level 3: Route-Vehicle (R-V) Matching. This level assigns appropriate vehicles and departure schedules to the generated routes. The core of this match is to dynamically adjust the service frequency based on the demand at the stops along the route during different time periods. For example, during the morning peak, routes covering major residential areas to the park should have a significantly higher frequency than during off-peak hours.

The PSV matching model is not a standalone algorithm but a logical framework that guides the establishment of the subsequent optimization model. It deconstructs the complex system optimization problem into three interconnected matching links, making the optimization objectives clearer and the model formulation more targeted.

3) Service Coordinated Optimization Model

Guided by the PSV matching framework, we construct a multi-objective integer programming model to achieve the coordinated optimization of stops, routes, and schedules. The objective functions of the model are as follows:

a) Objective 1: Minimize Total Operating Cost (min Z1)

$$Z1 = c\_v * \sum(k \in K) T\_k + c\_d * \sum(k \in K) D\_k \quad (1)$$

where T_k and D_k are the total operating time and total travel distance of vehicle k, respectively. c_v and c_d are the variable costs per unit of time (fuel, maintenance) and per unit of distance, respectively.

b) Objective 2: Minimize Total Passenger Travel Time (min Z2)

$$Z2 = \sum (i \in D\_s) (T\_wait(i) + T\_in\_vehicle(i) + T\_walk(i)) \quad (2)$$

where D_s is the set of successfully served demand points. T_wait(i) is the waiting time of passenger i at the stop, T_in_vehicle(i) is the in-vehicle time of passenger i, and T_walk(i) is the walking time of passenger i from the demand point to the stop.

c) Objective 3: Maximize Service Coverage (max Z3)

$$Z3 = |D\_s| / |D| \quad (1)$$

where |D_s| is the number of served demand points, and |D| is the total number of demand points. A demand point is considered "served" if there is a shuttle stop within its maximum walking distance walk_max.

Main Constraints:

- Stop Assignment Constraint: Each served demand point must be assigned to one and only one stop.
- Route Continuity Constraint: Each route must start from and return to the hub H.
- Vehicle Capacity Constraint: The number of passengers on a vehicle must not exceed its maximum capacity Q on any segment of the route.
- Maximum Waiting Time Constraint: The average or maximum waiting time for passengers at a stop must not exceed a preset threshold wait_max.
- Service Time Window Constraint: Vehicles must arrive at the respective stops to provide service within the time windows of passenger demand.

Since this model is an NP-hard multi-objective optimization problem, it is difficult to solve for an exact solution within a reasonable time. Therefore, in the next chapter, we will design a heuristic solution approach based on an improved genetic algorithm to obtain high-quality approximate optimal solutions.

IV. Algorithm Design and Solution Approach

The multi-objective coordinated optimization model constructed in the previous chapter is an NP-hard problem, making it difficult to find an exact solution for large-scale instances. Therefore, this chapter designs an efficient heuristic algorithm to solve this problem. The overall solution strategy follows a divide-and-conquer approach, decomposing the complex coordinated optimization problem into three sequential steps: Stop Layout Planning, Route and Schedule Optimization, and Cost-Benefit Analysis. The algorithm is carefully designed to ensure organic linkage between these steps.

A. Overall Solution Procedure

The solution procedure is illustrated in Figure 3 and is divided into the following three stages:

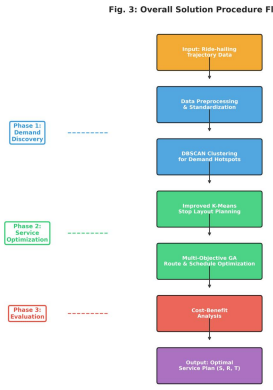

Fig. 3. Overall Solution Procedure Flowchart

Stop Layout Planning: The objective of this stage is to determine the optimal set of shuttle stops from a vast number of potential demand points. We first use the DBSCAN algorithm mentioned in the previous chapter to identify demand hotspots and use their centroids as candidate stops. Then, we employ an improved K-Means clustering algorithm with dynamic "split" and "merge" operations to iteratively optimize the candidate stops, ultimately determining the optimal number and location of the stops.

Route and Schedule Optimization: After determining the stop layout, this stage aims to design the optimal driving routes and dynamic departure timetables for the stops. We design a Multi-Objective Genetic Algorithm (MOGA) based on time-of-day demand. This algorithm divides the day into multiple periods (morning peak, off-peak, evening peak) and, based on the demand characteristics of each period, simultaneously optimizes the vehicle routes and service frequencies to minimize operating costs and passenger travel time.

Cost-Benefit Analysis: Once a complete service plan (stops, routes, schedules) is obtained, this stage constructs a quantitative cost-benefit model to evaluate the economic feasibility of the plan. Through simulated operations, we calculate Key Performance Indicators (KPIs) such as total revenue, total cost, and profit per passenger, providing data support for final decision-making.

B. Stop Layout Planning Algorithm

The quality of the stop layout directly affects the service coverage and passenger access convenience. Our proposed improved K-Means algorithm aims to overcome the drawback of traditional K-Means, which requires a preset K value. Through an adaptive split-and-merge mechanism, it finds a stop plan that best matches the spatial distribution of demand. The algorithm flowchart is shown in Figure 4.

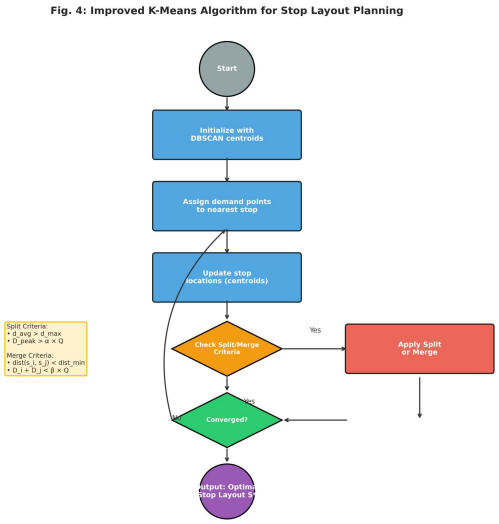

Fig. 4. Improved K-Means Algorithm for Stop Layout Planning

Initialization: The centroids of the demand hotspots identified by DBSCAN clustering are used as the initial set of candidate stops. An initial number of clusters, k, is set (estimated based on total demand and average vehicle capacity).

Iterative Optimization: The algorithm enters an iterative loop until no more stops can be split or merged.

- Assignment: All travel demand points are assigned to their nearest stop, forming k service clusters.
- Update: The centroid of each cluster is recalculated and becomes the new stop location.
- Split and Merge Check: All stops (clusters) are checked to determine if they meet the criteria for splitting or merging.

Splitting Criterion: When the demand served by a stop is too dispersed or the demand volume is too large, it should be split to improve service precision and avoid stop congestion. Splitting is triggered if either of the following conditions is met:

- High Dispersion: The average walking distance d_avg from all demand points in a cluster to its centroid exceeds a preset maximum acceptable average walking distance d_max.
- Demand Overload: The total demand D_peak of a cluster during peak hours exceeds the service capacity threshold of a single stop, α Q (where Q is vehicle capacity and α is an overload factor, 1.5). If the splitting condition is met, ak=2 K-Means algorithm is performed on the demand points within that cluster, and the two new centroids replace the original stop centroid.

Merging Criterion: When two stops are geographically too close and their combined demand is within a manageable range, they should be merged to reduce redundant construction and operational costs. Merging is triggered if all of the following conditions are met:

- Proximity: The distance dist(s_i, s_j) between two stop centroids s_i and s_j is less than a preset minimum inter-stop distance dist_min.
- Unsaturated Demand: The sum of the total demand of the two clusters during peak hours, D_peak(i) + D_peak(j), does not exceed the service capacity threshold β Q (β is a merging factor, 1.2). If the merging condition is met, the two stops are merged, and the new stop location is the weighted average center of the two clusters, with demand as the weight.

Termination: When no stops are split or merged in a full iteration, the algorithm converges and outputs the final stop layout plan.

C. Route and Schedule Optimization Algorithm

Once the stop layout is determined, the route and schedule optimization problem can be modeled as a Multi-Depot Vehicle Routing Problem with Time Windows and Capacity Constraints (MDVRPTW). Given its complexity, we use a Genetic Algorithm for the solution.

- Time-of-Day Strategy: To handle the dynamic demand, the operating day is divided into several time slices T_slice (30-minute intervals). The GA is run independently for each time slice, using the demand data specific to that period. This approach allows the service frequency and routes to adapt to demand fluctuations throughout the day.
- Chromosome Encoding: An integer-based encoding scheme based on a sequence of stops is used. A chromosome represents a complete vehicle routing plan. It is a sequence of stop numbers and delimiters (0). For example, for 8 stops and 2 vehicles, a chromosome could be [H, 3, 5, 1, H, 0, H, 2, 8, 6, 4, 7, H]. Here, H represents the transport hub, and 0 is the vehicle separator, indicating that the first vehicle serves the route H-3-5-1-H, and the second vehicle serves H-2-8-6-4-7-H.
- Fitness Function: Our objective is multi-faceted (minimize cost Z1, minimize passenger time Z2). In the GA, we convert this into a single-objective fitness function F using a weighted sum:

Minimize F = w_1 Z1' + w_2 Z2'

where Z1' and Z2' are normalized to [0, 1] using min − max scaling over the population in each generation to maintain comparability. Unless otherwise stated, we set w1 = 0.5 and w2 = 0.5. Because GA is stochastic, we run 10 independent trials with different random seeds for each scenario and report mean ± standard deviation of KPIs to avoid seed-specific conclusions.

Constraint Handling: A penalty function is introduced into the fitness calculation. Solutions that violate vehicle capacity constraints or maximum passenger waiting-time constraints are assigned a large penalty value, which effectively eliminates infeasible solutions during evolution.

The algorithm evolves over multiple generations, eventually converging to a high-quality solution, which represents the optimal route and service frequency plan for the current time slice. By repeating this process for all time

slices, a fully dynamic operational plan for the entire day is obtained.

D. Cost-Benefit Analysis Model

To evaluate the economic feasibility of the final plan, we establish the following cost-benefit analysis model to calculate the daily total profit P:

P = Rev - Cost_op - Cost_fixed

- Total Revenue (Rev): Rev = price_ticket N_passenger where price_ticket is the fare per trip, and N_passenger is the total number of passengers served per day.
- Total Operating Cost (Cost_op): Cost_op = c_dist Σ D_k where c_dist is the fuel and maintenance cost per unit distance, and D_k is the total daily travel distance of vehicle k.
- Total Fixed Cost (Cost_fixed): Cost_fixed = c_driver N_driver + c_vehicle N_vehicle where c_driver is the daily salary of a driver, c_vehicle is the daily depreciation cost of a vehicle, and N_driver and N_vehicle are the total number of drivers and vehicles required, respectively.

Using this model, we can quantitatively assess the profitability of the service plan under different parameter settings (ticket price, vehicle size) and compare it with traditional fixed-route shuttles, thereby providing a scientific basis for operators' investment decisions.

V. Numerical Experiments and Case Study

To validate the effectiveness and practicality of the proposed model and algorithms, this chapter conducts a case study based on the Shenzhen High-Tech Industrial Park. We first introduce the study area and data sources, then detail the experimental setup, and finally present and analyze the results of the numerical experiments.

A. Case Study Background

1) Study Area

The Shenzhen High-Tech Industrial Park, located in the Nanshan District of Shenzhen, is one of China's most important high-tech industry clusters. It covers an area of approximately 11.5 square kilometers and is home to thousands of technology companies, including renowned enterprises like Tencent, Huawei, and ZTE. The park has an employment population of over 500,000, generating enormous daily commuting demand. The park is served by several metro lines, with stations like Shenda, Hi-Tech Park, and Houhai acting as major public transport hubs. The last-mile connection from these metro stations to the various office buildings within the park is a prominent issue, making it an ideal scenario for this research(Figure 5).

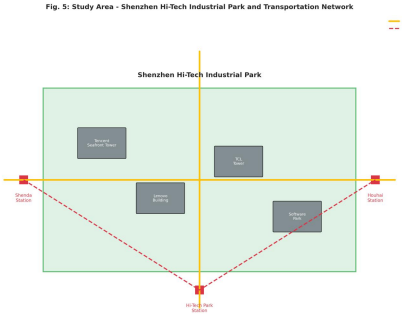

Fig. 5. Study Area - Shenzhen Hi-Tech Industrial Park and Transportation Network

2) Data Sources

This study utilizes the following multi-source data:

Ride-hailing Order Data: We use an anonymized ride-hailing order dataset obtained under a data-use agreement for Shenzhen over a continuous one-month period (22 weekdays). Informed consent was obtained from all subjects involved in the study.Each record contains an irreversible hashed trip identifier, pickup/drop-off timestamps, and pickup/drop-off coordinates (latitude/longitude). No personal identifiers are included. Due to licensing and privacy constraints, raw trip-level data cannot be publicly released; however, we provide (i) the full data dictionary, (ii) the preprocessing scripts, and (iii) aggregated demand grids and hotspot centroids sufficient to reproduce all reported figures and optimization results.

Road Network Data: The road network is extracted from OpenStreetMap (OSM) using a fixed download date to ensure version consistency. We retain drivable roads and intersections, and compute shortest-path travel distances/times between stops using posted speed limits where available; otherwise, we adopt road-class default speeds and report these defaults in the parameter settings.

POI Data: POIs (office buildings, metro stations, residential compounds, and major entrances) are obtained from a commercial map provider or open POI source with a fixed snapshot date. POIs are used only for functional interpretation and auxiliary validation of hotspots; the optimization is driven primarily by observed trip demand points.

3) Data Preprocessing

We focus on the morning peak hours (7:00 AM to 10:00 AM) on weekdays. Ride-hailing orders with drop-off locations within the Hi-Tech Park during this period are extracted as the raw dataset for last-mile demand. After preprocessing, we obtained approximately 85,000 valid last-mile records (about 3,800 per weekday). The preprocessing includes: removing records with missing timestamps/coordinates; filtering out implausible speeds (e.g., trips implying > 120 km/h); snapping coordinates to a consistent projection for distance calculation; and extracting weekday morning-peak trips (7:00 – 10:00) whose drop-off locations fall within the industrial park boundary polygon. We report the retained record counts after each step in the supplementary material to facilitate auditing.

B. Experimental Setup

To evaluate the performance of our proposed Dynamic Optimized Service (DOS) model, we designed a Traditional Fixed Service (TFS) model as a baseline for comparison.

- DOS Model: The service plan generated using the framework and algorithms proposed in this paper. Stop layout, routes, and schedules are all dynamically optimized based on data.
- TFS Model (baseline): The Traditional Fixed Service uses fixed routes and fixed headways (15 minutes) during operating hours. To ensure a fair comparison, TFS is constrained to use the same fleet size and vehicle capacity as DOS in each time slice. Fixed routes are constructed to connect the metro hub to

major park corridors and are tuned to minimize total route length while maintaining comparable stop coverage (using the same candidate stop set where applicable). This prevents performance gaps from being driven purely by mismatched resource assumptions.

Key parameters for the experiment are summarized below to ensure reproducibility (see “Parameter Settings” in this section)(Table II).

TABLE II. KEY PARAMETER SETTINGS FOR THE EXPERIMENT

User Group	Proportion	Main Travel Time	Travel Frequency	Activity Range	Typical Profile
Regular Commuters	45%	8:00-9:30,18:30-21:00	High (on workdays)	Relatively fixed	Young white-collar workers who work in the science parklive along subway lines, and have highly regular travelpatterns.
Flexible Business Travelers	20%	10:00-17:00	Medium	Relatively extensive	Business professionals visiting the science park formeetings or projects, with flexible but purposeful traveltimes.
Local Residents	35%	Distributed throughout the day, more active on weekends	LoW	Concentrated around residential communities	Residents living near Gaoxinyuan Station, whose travelpurposes are mostly local life consumption such asshopping, leisure, and dining.

C. *Experimental Results and Analysis*

1) Demand Discovery Results

Using the DBSCAN algorithm, we identified 4 major demand hotspot clusters within the park. The spatial distribution of these clusters is shown in the heatmap in Figure 6. It is evident that the demand is highly concentrated in areas with a high density of office buildings, such as the areas around Tencent Seafront Towers and the Software Park.

Fig. 6. Spatial Distribution Heatmap of Last-Mile Demand (Morning Peak)

Figure 7 shows the temporal variation curves of demand for these four clusters. All clusters exhibit a clear morning peak phenomenon, but their peak times and demand volumes differ. For example, Cluster 1 (Tencent Area) has the highest peak demand, occurring around 8:30 AM, while Cluster 4 (Software Park) has a relatively later and more prolonged peak. These spatio-temporal differences in demand underscore the necessity and potential of dynamic service design.

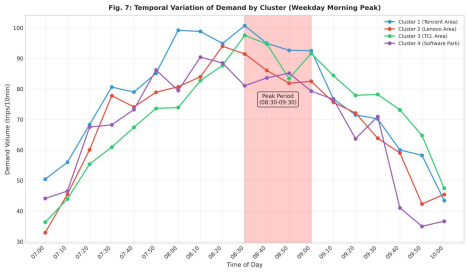

Fig. 7. Temporal Variation of Demand by Cluster (Weekday Morning Peak)

2) Stop Layout and Route Optimization Results

Applying the improved K-Means algorithm, we obtained an optimized stop layout plan consisting of 18 stops, as shown in Figure 8. These stops are strategically located near the centroids of demand clusters, effectively covering the main office areas while ensuring that the walking distance for most passengers is within 300 meters.

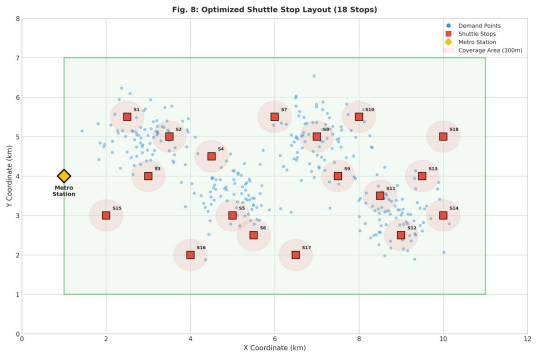

Fig. 8. Optimized Shuttle Stop Layout (18 Stops)

Based on the stop layout, the MOGA algorithm generated 4 optimized shuttle routes. Figure 9 shows one of the sample routes and its dynamic schedule. This route connects the metro hub with several key stops. Its service frequency is dynamically adjusted according to the demand forecast for different time periods: 15-minute intervals during the early morning, shortened to 8-minute intervals during the peak period (8:00-9:30), and then extended to 12-minute intervals after the peak. This dynamic scheduling strategy allows service capacity to precisely match demand fluctuations.

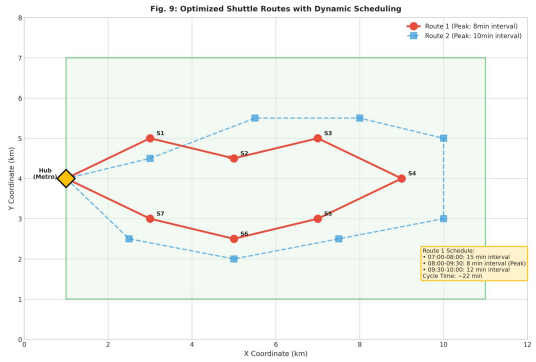

Fig. 9. Optimized Shuttle Routes with Dynamic Scheduling

3) Performance Comparison: DOS vs. TFS

To quantitatively evaluate the performance of the two models, We simulate one full workweek (5 weekdays) using observed demand patterns. For stochastic components (e.g., GA), each scenario is repeated for 10 independent seeds. KPIs are computed per day and then aggregated over the week; we report the weekly mean ± standard deviation(Figure 10).

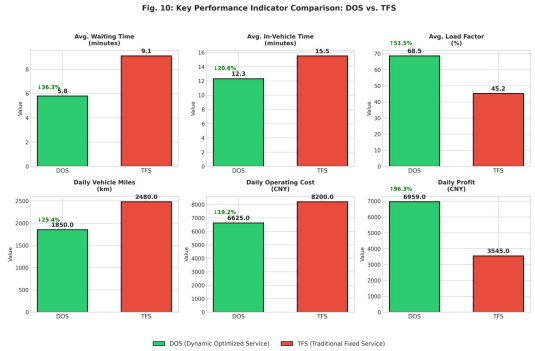

Fig. 10. Key Performance Indicator Comparison: DOS vs. TFS

- Passenger Service Level: DOS improves passenger experience compared with TFS. Over the simulated week, the average waiting time decreases from 9.1 to 5.8 minutes (mean values; variability across days/seeds is reported as mean ± standard deviation), corresponding to an average reduction of about 36.3%. In-vehicle time is also reduced because optimized routes are more direct under the same fleet constraints.
- Operational Efficiency: DOS improves seat utilization and reduces unnecessary vehicle mileage under matched fleet constraints. The average load factor increases from 45.2% to 68.5%, and vehicle deadheading/VMT is reduced by about 25.4% on average, which translates into lower operating costs in the cost model.
- Economic Benefits: Under the assumed fare and cost parameters (reported explicitly in the parameter settings), DOS yields higher estimated daily profit than TFS while serving more passengers. With a ticket price of 3 CNY per trip, the estimated daily profit is about 6,959 CNY for DOS versus 3,545 CNY for TFS in the case study week. We emphasize that absolute profit values depend on cost assumptions; therefore, we additionally provide a sensitivity analysis over key parameters (fare, per-km cost, driver cost, and depreciation).

Overall, the numerical experiments suggest that the proposed DOS framework can outperform a fixed-route baseline in both service quality and operational efficiency under the study assumptions. We report full parameter settings and repeated-run statistics to support reproducibility and to clarify the conditions under which these gains hold.

VI. Discussion and Conclusion

This study proposed and validated a data-driven framework for last-mile shuttle service design, centered on precision matching and service optimization. The empirical results from the Shenzhen High-Tech Industrial Park case study demonstrate the significant potential of this framework in enhancing service efficiency, improving passenger experience, and increasing operational profitability. This chapter will further discuss the theoretical contributions and practical implications of the research, acknowledge its limitations, and suggest directions for future work.

A. Discussion of Findings

The superiority of the Dynamic Optimized Service (DOS) model over the Traditional Fixed Service (TFS) model can be attributed to several key innovations:

- From Static to Dynamic: The core advantage of our model lies in its ability to capture and adapt to the dynamic nature of travel demand. By analyzing high-frequency ride-hailing data, we moved beyond the static, aggregated demand assumptions of traditional planning methods. The time-of-day scheduling strategy allows service capacity to be deployed precisely when and where it is needed most, effectively resolving the classic supply-demand mismatch problem that plagues fixed-route systems.
- The Power of Precision Matching: The proposed Passenger-Stop-Vehicle (PSV) three-level matching model serves as a crucial conceptual bridge. It deconstructs the complex system optimization problem into a series of clear, targeted matching relationships. This ensures that the optimization process is not a black box but is guided by the principle of aligning service resources with individual needs at every level, from a passenger choosing a stop to a vehicle being assigned to a route.
- Integrated Optimization: Unlike previous studies that often optimized stops, routes, or schedules in isolation, our framework emphasizes their coordinated optimization. The improved K-Means algorithm for stop layout considers not just spatial coverage but also demand volume, providing a more rational input for the subsequent routing algorithm. The Multi-Objective Genetic Algorithm (MOGA) then simultaneously considers multiple performance metrics, finding a balanced solution that would be difficult to achieve through sequential, single-objective optimization.

B. Theoretical Contributions and Practical Implications

Theoretical Contributions:

- This research enriches the literature on public transport planning by providing a comprehensive,

end-to-end data-driven framework for designing demand-responsive transit systems. It systematically integrates methodologies from data mining (DBSCAN), machine learning (K-Means), and operations research (MOGA).

- The introduction of the PSV precision matching concept offers a new theoretical lens for analyzing and modeling the intricate relationships within shared mobility systems, which can be extended to other shared transport services like carpooling or on-demand buses.

Practical Implications:

- For transport operators, this study provides a complete, actionable toolkit for designing and operating modern shuttle services. Adopting this model can help them break away from the low-efficiency, low-profitability dilemma, significantly improving their market competitiveness and financial sustainability.
- For urban planners and policymakers, this research demonstrates how leveraging existing big data resources can lead to more efficient and user-centric public transport solutions. It provides a strong case for promoting data sharing and investing in intelligent transportation systems to alleviate urban congestion and advance sustainable mobility goals.
- For commuters, the implementation of such a service would mean shorter waiting times, less walking, and a more reliable and comfortable last-mile travel experience, which could incentivize a shift from private cars to public transit.

C. *Limitations and Future Research*

Despite the promising results, this study has several limitations that open avenues for future research:

- Data Source Limitations: This study primarily relied on ride-hailing data, which may not capture the full spectrum of last-mile travel demand (those who walk, bike, or use other modes). Future research could integrate multi-source data, such as public transit smart card data, mobile signaling data, and shared-bike data, to create a more holistic demand profile.
- Real-time Dynamics: The current model, while dynamic, operates on an offline, day-ahead planning basis. It does not yet incorporate real-time adjustments based on live traffic conditions or sudden demand surges. Future work could focus on developing a real-time dynamic dispatching module that can adjust routes and schedules on the fly, further enhancing system responsiveness.
- Behavioral Considerations: The model assumes that passengers will always choose the service based on rational factors like time and cost. It does not account for more complex behavioral factors, such as comfort, safety perceptions, or brand loyalty. Incorporating discrete choice models or agent-based simulations could lead to more realistic demand predictions and service designs.
- Scalability and Transferability: While the framework is designed to be general, its specific parameters and performance were validated in the context of a single industrial park. Further studies are needed to test its scalability and transferability to other urban contexts, such as university campuses, large residential communities, or tourist areas, which may have different demand characteristics.

D. *Reproducibility and Data/Code Availability*

Reproducibility and Data/Code Availability: To facilitate reproducibility, we provide the complete list of parameter settings, algorithmic pseudocode, and the preprocessing pipeline (cleaning rules, boundary filtering, coordinate projection, and demand aggregation). Due to privacy and licensing restrictions, raw trip-level ride-hailing records cannot be publicly shared; however, we release derived and non-identifying artifacts, including aggregated demand grids, hotspot centroids, candidate stop sets, and the processed road-network graph used for routing. These artifacts, together with the optimization code and random seeds, are sufficient to reproduce all figures, KPIs, and comparative conclusions reported in this paper.

E. *Conclusion*

The last-mile problem is a persistent challenge in urban transportation, but the proliferation of big data and advanced analytics offers a powerful new arsenal to tackle it. This study proposed a data-driven framework for last-mile shuttle service design, integrating dynamic demand discovery, a novel PSV precision matching model, and a coordinated optimization algorithm. Through a case study in Shenzhen, we demonstrated that this approach can lead to a service that is significantly more efficient, user-friendly, and profitable than traditional fixed-route systems.

As cities continue to grow and smart technologies become more ubiquitous, the future of urban mobility lies in systems that are adaptive, responsive, and personalized. The principles and methods developed in this research represent a firm step in that direction, offering a viable pathway to transform the last-mile journey from a daily frustration into a seamless and sustainable experience.

REFERENCES

[1] Redding, S. J. (2022). Suburbanization in the USA, 1970 – 2010. Economica, 89, S110–S136. https://doi.org/10.1111/ecca.12476

[2] Anas, A., Arnott, R., & Small, K. A. (1998). Urban spatial structure. Journal of Economic Literature, 36(3), 1426 – 1464. https://doi.org/10.1257/jel.36.3.1426

[3] Gärling, T., & Steg, L. (2007). Threats from car traffic to the quality of urban life: Problems, causes, solutions. Elsevier Science. https://doi.org/10.1016/B978-0-08-044638-1.X5000-1

[4] Vuchic, V. R. (2017). Urban transit: Operations, planning, and economics. John Wiley & Sons. https://doi.org/10.1002/9781119125352

[5] Banister, D. (2008). The sustainable mobility paradigm. Transport Policy, 15(2), 73–80. https://doi.org/10.1016/j.tranpol.2007.05.004

[6] Shoup, D. (2021). High cost of free parking. Routledge. https://doi.org/10.4324/9781003150250

[7] Shu, P., Sun, Y., Xie, B., Xu, S. X., & Xu, G. (2021). Data-driven shuttle service design for sustainable last mile transportation. Advanced Engineering Informatics, 49, 101344. https://doi.org/10.1016/j.aei.2021.101344

[8] Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. Nature, 453(7196), 779–782. https://doi.org/10.1038/nature06958

[9] Xue, Q., Zhang, W., Ding, M., Yang, X., Wu, J., & Gao, Z. (2023). Passenger flow forecasting approaches for urban rail transit: A survey.

International Journal of General Systems, 52(8), 919 – 947. https://doi.org/10.1080/03081079.2022.2161024

[10] Batty, M. (2013). The new science of cities. MIT Press. https://doi.org/10.7551/mitpress/9606.001.0001

[11] Murray, A. T. (2003). A coverage model for improving public transit system accessibility and expanding access. Annals of Operations Research, 123(1), 143 – 156. https://doi.org/10.1023/A:1024669607245

[12] Toth, P., & Vigo, D. (Eds.). (2014). Vehicle routing: Problems, methods, and applications. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611973594

[13] Fishman, E., Washington, S., & Haworth, N. (2014). Bike share's impact on car use: Evidence from the United States, Great Britain, and Australia. Transportation Research Part D: Transport and Environment, 31, 13–20. https://doi.org/10.1016/j.trd.2014.05.004

[14] Schaller, B. (2018). The new automobility: Lyft, Uber and the future of American cities. Schaller Consulting. https://schallerconsult.com/wp-content/uploads/2018/11/New-Automobility-2018.pdf

[15] Ceder, A. (2016). Public transit planning and operation: Modeling, practice and behavior. CRC Press. https://doi.org/10.1201/b19724

[16] Deka, U., Varshini, V., & Dilip, D. M. (2023). The journey of demand responsive transportation: Towards sustainable services. Frontiers in Built Environment, 8, 942651. https://doi.org/10.3389/fbuil.2022.942651

[17] Psaraftis, H. N. (1980). A dynamic programming solution to the single vehicle many-to-many immediate request dial-a-ride problem. Transportation Science, 14(2), 130 – 154. https://doi.org/10.1287/trsc.14.2.130

[18] Xiao, M., Chen, L., Feng, H., Peng, Z., & Long, Q. (2024). Smart city public transportation route planning based on multi-objective optimization: A review. Archives of Computational Methods in Engineering, 31(6), 3351–3375. https://doi.org/10.1007/s11831-023-09909-8

[19] Falkner, J. K., Thyssens, D., Bdeir, A., & Schmidt-Thieme, L. (2022). Learning to control local search for combinatorial optimization. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 361–376). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-26409-2_22

[20] Ma, Z., & Chow, J. Y. (2022). Transit network frequency setting with multi-agent simulation to capture activity-based mode substitution. Transportation Research Record, 2676(4), 41 – 57. https://doi.org/10.1177/03611981221089964

[21] Li, W., Wang, S., Zhang, X., Jia, Q., & Tian, Y. (2020). Understanding intra-urban human mobility through an exploratory spatiotemporal analysis of bike-sharing trajectories. International Journal of Geographical Information Science, 34(12), 2451 – 2474. https://doi.org/10.1080/13658816.2020.1722647

[22] Li, S., Liang, X., Zheng, M., Chen, J., Chen, T., & Guo, X. (2024). How spatial features affect urban rail transit prediction accuracy: A deep learning based passenger flow prediction method. Journal of Intelligent Transportation Systems, 28(6), 1032 – 1043. https://doi.org/10.1080/15472450.2023.2202560

[23] Zhou, X., Ke, R., Yang, H., & Liu, C. (2021). When intelligent transportation systems sensing meets edge computing: Vision and challenges. Applied Sciences, 11(20), 9680. https://doi.org/10.3390/app11209680.

ACKNOWLEDGEMENTS

The authors thank the organizations that supported this study, including the provider of anonymized mobility data and the open-source communities whose geographic resources enabled the analysis.

FUNDING

None.

AVAILABILITY OF DATA

Not applicable.

AUTHOR CONTRIBUTIONS

Di Lu: Conceptualization; Methodology; Data curation; Software; Formal analysis; Validation; Visualization; Writing – original draft.

Hao Sun: Conceptualization; Methodology; Supervision; Writing – review & editing; Project administration; Funding acquisition.

COMPETING INTERESTS

The authors declare no competing interests.