Title

AI for Cross-Domain Mobility Data Fusion, Human Activity Generation, and Travel Demand Modeling

Authors

Liao, Xishun
Liu, Yifan
Ma, Haoxuan
et al.

Publication Date

2026-08-01

DOI

10.82903/B23W22

CENTER OF EXCELLENCE
on New Mobility and Automated Vehicles

AI for Cross-Domain Mobility Data Fusion, Human Activity Generation, and Travel Demand Modeling

August 2026

About the Center of Excellence on New Mobility and Automated Vehicles

The Center of Excellence on New Mobility and Automated Vehicles focuses on the impacts of new mobility technologies and highly automated vehicles on the evolving transportation system when deployed at scale. Launched in 2023, the center studies the anticipated long-term impacts of increased new mobility technologies and services on land use, real estate and urban design; transportation system optimization; access to mobility and job participation; and municipal budgets and cost-effective allocation of public resources. Established through a cooperative agreement between the Federal Highway Administration and UCLA, the Mobility COE is made up of five core partners and a community of practice from academia and nonprofit organizations.

www.mobilitycoe.org

Acknowledgments

This material is based upon work supported by the Federal Highway Administration under Agreement No. 693JJ32350027. The Mobility Center of Excellence at UCLA acknowledges the Gabrielino/Tongva peoples as the traditional land caretakers of Tovaangar (the Los Angeles basin and So. Channel Islands). As a land grant institution, we pay our respects to the Honuukvetam (Ancestors), ‘Ahiihirom (Elders) and ‘Eyoohiinkem (our relatives/relations) past, present and emerging.

Disclaimer

Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the Author(s) and do not necessarily reflect the view of the Federal Highway Administration.

Citing This Report

Suggested Citation: Liao, X., Liu, Y., Ma, H., & Ma, J. (2026). AI for Cross-Domain Mobility Data Fusion, Human Activity Generation, and Travel Demand Modeling. *UCLA: Mobility Center of Excellence*. http://dx.doi.org/10.82903/B23W22

AI for Cross-Domain Mobility Data Fusion, Human Activity Generation, and Travel Demand Modeling

Xishun Liao, Ph.D., Assistant Professor, University of Central Florida

Yifan Liu, Graduate Student Researcher, University of California, Los Angeles,

Haoxuan Ma, Graduate Student Researcher, University of California, Los Angeles,

Jiaqi Ma, Ph.D., Professor, University of California, Los Angeles

August 2026

CENTER OF EXCELLENCE
on New Mobility and Automated Vehicles

Table of Contents

List of Tables

List of Figures

Executive Summary

Transportation agencies and researchers face significant challenges due to fragmented, incomplete, or inaccessible mobility data, which limits the ability to accurately model and plan complex transportation systems. Although many datasets exist, ranging from GPS trajectories and household travel surveys to traffic sensors and Points of Interest (POIs), they remain siloed, inconsistent in resolution, or difficult to access. This lack of integration hinders the development of robust models for demand prediction, network optimization, and mobility planning. It also affects automated vehicle (AV) deployment for operational design domain definition, infrastructure readiness assessment, safety oversight, and identification of where automated services can deliver clear public benefit. To address this gap, this project reviews key mobility data sources, develops a fusion pipeline that combines multiple cross-domain datasets, and demonstrates how these integrated resources can support the synthesis and modeling of travel behavior, demand, and trajectories in heterogeneous urban settings. The project delivers a survey of the mobility data landscape and a scalable, transferable framework for data fusion that agencies can apply to transportation research and planning.

The first component is a comprehensive survey and taxonomy of mobility data sources in the market. Rather than simply cataloging datasets, the project organizes them into a functional stack and five-dimensional classification system, enabling stakeholders to assess content, generation methods, spatial and temporal resolution, and relevance for mobility applications. This taxonomy can inform AV deployment, as it links travel demand, roadway context, operational conditions, connected and mobile data sources, and vehicle level intelligence into a single framework, making it easier for agencies to identify the combinations of data needed for planning, validation, and public oversight. This data taxonomy highlights integration opportunities while also exposing persistent data gaps, offering transportation agencies and researchers a structured reference for building reliable models. In practice, the taxonomy gives agencies a repeatable way to compare candidate datasets, identify integration gaps, and decide which sources are fit for planning, simulation, operations, and AV-related analysis.

Building on this foundation, the second component develops scalable data fusion frameworks that integrate cross-domain datasets to generate high-resolution synthetic travel demand datasets. Such datasets could be used for operations planning, as well as transportation and urban planning at large. The case study in Los Angeles County demonstrates the employment of large language models (LLMs) to annotate and enrich GPS data, deep learning methods to reconstruct incomplete activity chains, and transfer learning techniques to adapt models to regions with varying data availability. Using GPS data, survey information, and POIs, the synthetic travel pattern is closely aligned with observed behaviors. The framework's value-add is twofold: it improves key fusion steps relative to explicit baselines, and it produces analysis-ready synthetic travel-demand outputs from existing mobility-data inputs. This reduces the burden of conducting new large-scale household travel surveys as well as manual annotation.

When evaluated through traffic simulations, the synthetic data achieved strong accuracy compared with Caltrans PeMS sensor data on the I-405 corridor, with mean absolute percentage errors of 5.85% for traffic volume and 4.36% for speed. These behaviorally realistic and spatially grounded data products support AV mixed-traffic scenario testing, service planning, digital twin evaluation, and assessment of network effects before AV real world deployment.

The project outcomes include a comprehensive report on the mobility data landscape, synthetic data generation frameworks for multi-source data fusion and modeling, and a validated case study in real-world settings. The broader implications span intelligent transportation systems, urban planning, public policy, and infrastructure development. By integrating heterogeneous mobility datasets and producing demographically representative and spatially consistent mobility patterns, the framework supports more informed decision-making to improve traffic management, energy use and congestion, and infrastructure investment. Importantly, it reduces barriers to advanced mobility modeling in data-scarce regions, enabling a more universally applicable and transferable foundation for transportation and urban studies. In doing so, this work establishes a scalable and transferable digital mobility infrastructure to guide demand forecasting, congestion management, and long-term planning.

Introduction

As urban systems become more complex, accurate and scalable models of how people and vehicles move are essential for infrastructure planning, congestion mitigation, transit design, and emergency response. This human mobility modeling increasingly depends on the availability and integration of a variety of data sources. Human mobility data refers to observed or inferred records describing where, when, and how people, vehicles, or goods move through space. Trip data are a subset of mobility data that record discrete trips, typically including origins, destinations, departure and arrival times, modes, and sometimes purposes. Travel demand refers to the amount and distribution of trips that travelers wish to make across time, space, mode, and purpose, while travel demand models estimate these patterns for forecasting, planning, operations, and simulation. In this report, mobility data serve as the input evidence, and travel demand modeling is the primary application. These data are used by transportation agencies for planning and operations, metropolitan planning organizations for forecasting, researchers for behavior analysis, and mobility or AV service providers for service design, market sizing, and coverage analysis.

Currently, the core challenge is data: while many mobility-relevant datasets now exist—including GPS traces, household travel surveys, road networks, traffic sensors, and transit feeds—each captures only a limited view of travel behavior, often with incompatible structure, resolution, or coverage. For example, trajectory data (e.g., smartphone GPS) offers high-frequency movement signals but lacks information about activity purpose or user attributes. Surveys such as the National Household Travel Survey (NHTS) and American Time Use Survey (ATUS) provide semantically rich behavioral insights but are infrequent, expensive, and spatially sparse. POI and road network datasets describe the physical environment but do not reveal how it is used. These sources differ in spatial scale, temporal granularity, update frequency, and ownership—making unified modeling across them both technically and conceptually challenging.

Traditional modeling approaches rely on static assumptions and localized calibration, limiting their transferability across regions. In contrast, data-driven methods, including deep learning and large-scale representation learning, offer a path forward by learning behavioral patterns directly from real-world observations. However, their effectiveness depends critically on how well heterogeneous datasets are integrated. In this report, mobility data serves as evidence, and travel demand modeling is the primary application. A travel demand model that captures not just where and when people move, but also why and under what constraints, requires combining multiple complementary data sources. This report addresses that challenge by proposing a structured fusion of trajectory, survey, and contextual data to build semantically interpretable, generalizable models of travel demand.

To address this gap, the primary objective of this project is to develop a scalable and transferable data fusion framework that integrates heterogeneous travel demand data sources into a coherent, structured dataset suitable for modeling and simulation. By aligning geographical, temporal, demographic, and traffic-related data, the proposed framework aims to support the generation of realistic synthetic travel trajectories and the modeling of region-specific mobility patterns. The resulting unified framework will enhance the capacity of researchers and transportation agencies to conduct fine-grained travel demand analyses, improve transportation planning processes, and simulate travel behavior at both system-wide and individual levels.

This project was executed in three interlinked tasks:

Task 1: Survey and Assessment of Available Mobility Data
This task involved a systematic review of travel demand data sources in the United States. It identifies open-source and proprietary datasets, including GPS trajectories, Points of Interest (POIs), household travel surveys, demographic statistics, and traffic network data, and evaluates their availability, integrity, and relevance for modeling. The survey also mapped these datasets to typical stakeholders (e.g., government agencies, private mobility providers, academic institutions) and identified key data gaps and integration opportunities. The proposed taxonomy is therefore not only a descriptive catalog, but also a practical screening tool for selecting fit-for-purpose data inputs and identifying where additional collection or fusion is needed.

Task 2: Development of a Scalable Data Fusion Framework
Building upon the surveyed data, this task focused on designing and implementing a framework that harmonizes multi-source mobility data. The framework addressed technical challenges related to spatial-temporal alignment, resolution mismatches, and missing or noisy data. Techniques such as large language models (LLMs), deep learning-based imputation, and domain adaptation were employed to enhance data completeness and generalizability. The outcome was an engineered system that produces structured, high-quality synthetic travel demand datasets applicable across different regions.

Task 3: Case Study and Performance Evaluation
To validate the framework, it was applied to Los Angeles County as a real-world urban case study. Evaluation metrics were designed to assess the realism, consistency, and demographic accuracy of the generated travel-demand data. The framework's outputs were then tested through large-scale traffic simulations, comparing synthetic travel behavior with observed network performance. The insights from the case study can guide further refinement of the framework and contribute toward the creation of a digital mobility infrastructure.

This project addresses the longstanding challenge of fragmented and siloed mobility data by developing an integrative, generalizable data fusion framework. It bridges the gap between rich but limited traditional surveys and high-frequency but noisy passive data, enabling robust mobility modeling across diverse contexts and supporting the advancement of intelligent transportation systems.

The structure of the report is as follows. Section 2 presents a survey of currently available datasets for mobility modeling to inform transportation agencies and researchers. Section 3 presents the Los Angeles County case study and the resulting data-fusion workflow for synthetic travel-demand generation and validation.

Data Survey: Existing Mobility Data Sources in the U.S.

Accurate mobility modeling requires the integration of a variety of datasets capturing spatial trajectories, activity semantics, demographic contexts, transportation infrastructure, and traffic dynamics. To facilitate systematic integration and enhance data interoperability, we propose a structured taxonomy framework. This section surveys recent mobility datasets available in the United States, ranging from static road infrastructure to real-time sensor feeds, that plays a critical role in traffic management, safety improvements, and mobility optimization. We categorize these datasets by information type, assess their accessibility and integrity, identify typical data holders, and highlight integration opportunities and data gaps.

Mobility Dataset Taxonomy

The mobility dataset taxonomy addresses the objective of systematically classifying mobility data assets to enable better stakeholder collaboration, identify data gaps, and support large-scale implementation in smart mobility systems (Figure 1).

The taxonomy has two complementary components. The mobility functional stack organizes datasets by their role in the mobility evidence pipeline, from population-level behavior to individual-agent observations. It tells users where a dataset sits in the mobility evidence pipeline and what it is useful for. It gives users an immediate sense of why the data exists, what mobility-system level it represents, and how it can serve as an input to data discovery, screening, and fusion planning. The five-dimensional metadata attributes then describe each dataset along independent dimensions: content, generation method, source infrastructure, resolution, and intelligence processing level. It tags each dataset with precise technical attributes—what the content is, how it was produced, where it originates, at what spatial and temporal scales it operates, and what processing level it reflects—so researchers can filter, compare, and integrate sources with confidence.

In short, the functional stack explains where a dataset originates, while the metadata attributes explain the resolution, timeliness, security, processing, etc. of that dataset in order to assess if it is fit for purpose. This separation makes it easier for stakeholders to apply the taxonomy consistently across heterogeneous sources.

Moreover, this taxonomy provides a structured basis for automated vehicle (AV) deployment by clarifying what data is needed not only for vehicle level perception and control, but also for operational design domain definition, infrastructure readiness assessment, safety validation, and long-term service planning. For AV applications, these mobility-data inputs are especially important because safe and efficient deployment depends on not just onboard sensor logs, but also may be affected by population demand, roadway geometry, traffic control, environmental conditions, connected infrastructure, and the ability to link microscopic vehicle behavior with network level outcomes.

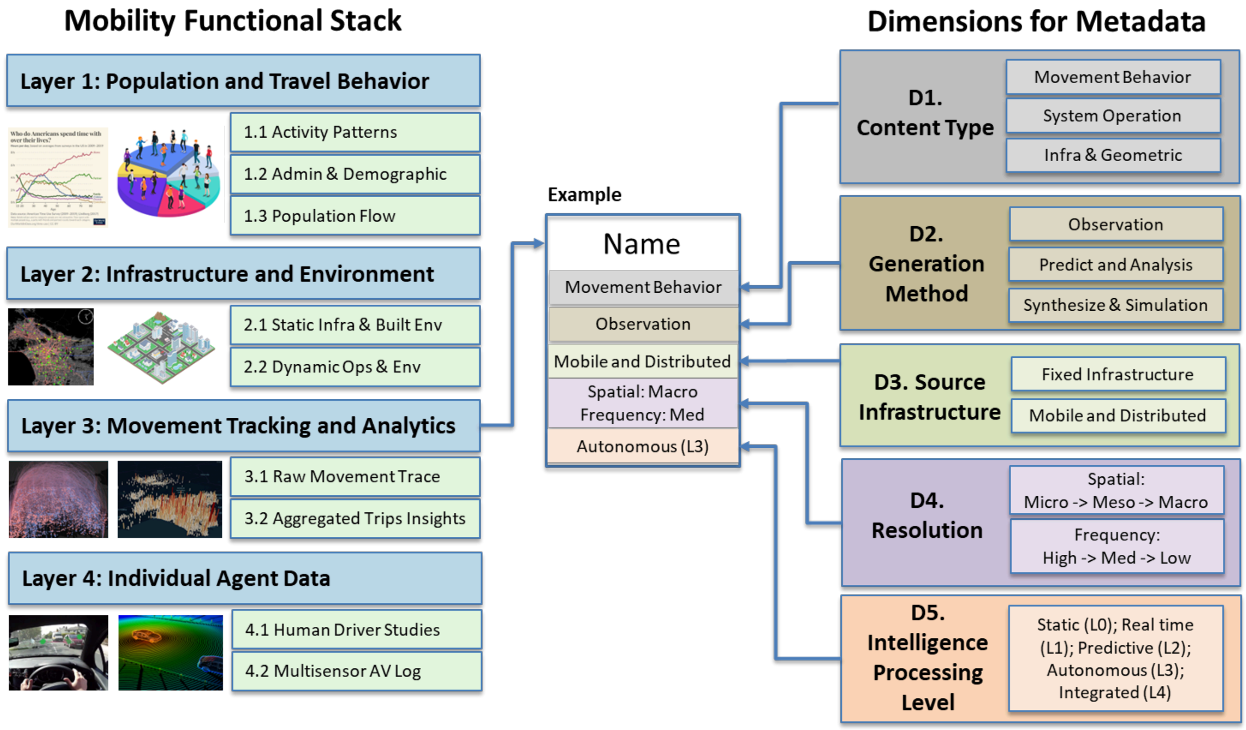

Figure 1: Taxonomy of mobility dataset

Mobility Functional Stack

The mobility functional stack, represented on the left side of the framework in Figure 1, presents a four-layer conceptual architecture that provides intuitive understanding of mobility data progression from broad population patterns to individual agent behaviors. This layered approach follows the fundamental principle that data granularity and scope naturally progress from broad coverage with long-term mass behavior characteristics toward specific location data with real-time individual agent focus.

The functional stack is intended as a data-discovery and fusion-planning tool. Users first locate each candidate dataset within one of the four system layers to understand the scale and role of the mobility phenomenon it observes. Datasets within the same layer are often candidates for comparison, substitution, or harmonization, while datasets across complementary layers can be combined to fill information gaps. For example, Layer 1 surveys provide behavioral intent and demographic priors, Layer 2 POIs and networks provide physical context, and Layer 3 GPS traces provide observed movement dynamics. The stack does not by itself define the final fused output; rather, it organizes input sources so users can select fit-for-purpose combinations for downstream travel-demand modeling, simulation, operations, or AV applications.

For AV deployment, this layered view makes clear that performance exists at multiple systems levels and not just at vehicle level: AVs must be evaluated against who needs the service, where the service operates, what infrastructure and environmental conditions exist, and how vehicle behavior interacts with the larger transportation network.

Layer 1: Population and Travel Behavior

Layer 1 is the most macroscopic and longest-time-constant layer in the stack. It is the foundational layer to describe population-level characteristics, strategic travel behavior, and long-term demand patterns, capturing who people are, why and how they travel, and how these patterns evolve over time. Its scope spans household travel surveys, census and employment records, demographic transitions, and tourist or visitor surveys. Together, these sources provide ground truth for mode share and trip purposes, reveal economic mobility patterns, and supply baselines for commuting flows and long-term demand forecasts. This layer operates at annual to generational time scales, with spatial resolution ranging from block groups to national coverage, and emphasizes population-level dynamics rather than individual trips. Its three sublayers organize this foundation. Layer 1.1 Activity Patterns and Travel Behavior focuses on detailed diaries of trip purpose, scheduling, and time use, ranging from daily activities to seasonal patterns. Layer 1.2 Administrative and Demographic Registers consolidate census and employment data to document residence, workplace, and commute patterns across scales. Layer 1.3 Population Flow Evolution highlights longer-term shifts in residential and labor mobility and economic transitions, anchored by demographic forecasts and strategic planning. In the AV context, these data are essential for estimating where automated mobility services are likely to generate demand, which geographic areas may benefit or not, how travel needs vary by purpose and time of day, and how AV deployment could affect access, affordability, and mode substitution.

Layer 2: Infrastructure and Built Environment

This layer provides the physical and environmental context within which mobility occurs. It covers transportation networks, built environment features, environmental conditions, and smart city systems. Data in this layer are used for operations management, safety monitoring, asset condition assessment, and accessibility analysis. Temporal scales range from static infrastructure inventories to real-time sensor feeds, while spatial scales extend from point-level sensors to corridors, full networks, and regional systems. Its two sublayers capture these different characteristics. 2.1 Static Infrastructure and Built Environment documents transportation networks, transit topology, buildings, points of interest, land use, and infrastructure asset inventories. 2.2 Dynamic Operations and Environmental Monitoring captures real-time conditions through traffic and air quality sensors, crash and incident reports, construction and work-zone alerts, and smart city coordination systems. Supported by this layer, AV deployment is able to consider roadway geometry, lane configuration, signage, signal timing, curb regulations, work zones, and weather conditions to determine whether an automated system remains within its defined operational design domain. It also provides the map and infrastructure context needed for localization, rule compliance, remote operations support, and digital twin testing.

Layer 3: High-Resolution Movement Tracking and Analytics

Layer 3 examines how people and vehicles move across space and time, progressing from raw GPS traces to sophisticated aggregated flow analysis. The scope encompasses GPS tracking systems, movement tag technologies, user-reported information, and behavioral analytics platforms. This layer demonstrates granularity progression from individual tracking records through corridor flow analysis to Traffic Analysis Zone (TAZ)-level patterns and point-of-interest connection mapping. Temporal scales range from sub-minute tracking intervals to daily aggregation periods, while spatial scales span from individual device positioning to comprehensive regional network analysis. This movement tracking layer incorporates two specialized sublayers that capture two different data types reflecting different levels of data processing. 3.1 Device-Level Raw Movement Traces provide anonymized device (e.g., smartphones) pings, GPS logs, and movement tag records at high resolution, with sub-minute temporal precision and fine-grained spatial positioning from device to point-of-interest level. 3.2 Movement Pattern Analytics offers pre-processed and privacy-protected outputs, including aggregated origin–destination flows, travel pattern recognition, and trip record analysis. These datasets operate at TAZ to regional scales, with temporal coverage from near real-time monitoring to long-term historical trend analysis.

For AV applications, layer 3 helps identify where demand concentrates, how speeds and route choices vary, where pickup and drop off frictions emerge, and where mixed traffic interactions or recurrent bottlenecks may challenge AV fleet operations. At the aggregated level, these data also support fleet sizing, repositioning minimization, service area design, and assessment of network impacts before field deployment.

Layer 4: Individual Agent Data

The most detailed layer represents single-agent, real-time sensor observations at individual vehicle and driver levels, operating with sub-second to trip-level temporal scales and maintaining vehicle-centric, lane-level spatial precision. This layer captures the highest resolution mobility data available, focusing on individual behavior and vehicle performance characteristics that inform both safety research and AV development initiatives. Individual Agent Data encompasses two distinct sublayers. 4.1 Human Driver Studies, also known as Naturalistic Driving Log, captures in-vehicle sensor data under real human driving conditions, supporting human-machine interaction research, risk modeling initiatives, and comprehensive behavior analysis studies. These studies operate with temporal scales from individual trip analysis to longitudinal study periods, maintaining individual vehicle level spatial focus for detailed behavioral understanding. 4.2 Multi-sensor Autonomous Vehicle Log provides comprehensive sensor data including LiDAR point clouds, radar measurements, and camera imagery specifically designed for object detection algorithm development, motion prediction model training, and autonomous vehicle stack development. These datasets are produced by instrumented or automated vehicles and used to develop, test, monitor, or validate AV functions, maintaining sub-second to trip-level temporal resolution with vehicle-centric, lane-level spatial precision requirements. However, they are often proprietary, safety-sensitive, and subject to privacy controls, they are typically accessed through restricted research partnerships, controlled testbeds, or regulated data-sharing arrangements rather than through unrestricted public release.

Five-Dimensional Metadata Attributes

The right-side column in Figure 1 provides technical specification through five independent dimensions following the taxonomic principle: Content, Generation Method, Source, Resolution, Intelligence. This approach enables comprehensive data characterization using independent, multi-selectable dimensions that maximize orthogonality, thus reducing the overlap problems inherent in traditional hierarchical taxonomies. For AV related datasets, these dimensions also clarify whether a source is reliable enough for perception, planning, simulation, monitoring, or operational support.

Dimension 1: Mobility Data Content Types

Dimension 1 defines what the data fundamentally represents and is divided into three core categories. Movement and Behavioral Data (D1.1) captures dynamic aspects of mobility at both individual and system levels. This includes detailed vehicle trajectories, trip characteristics, and activity patterns that reveal how people and goods move through space and time. It also includes collective traffic behaviors, such as flow metrics, congestion dynamics, and modal interactions, along with behavioral attributes like compliance, safety actions, and technology adoption patterns. System Operational Data (D1.2) focuses on the performance and control of mobility infrastructure. This includes signal operations, route guidance systems, infrastructure status monitoring, incident and event tracking, as well as external environmental conditions that shape system dynamics, such as weather, air quality, and energy consumption. Infrastructure and Geometric Data (D1.3) encompasses the physical backbone of mobility systems, including road networks, control devices, transit infrastructure, and land use characteristics. It extends to high-definition mapping, lane-level geometric detail, and infrastructure condition assessments, all of which are foundational for simulation and digital twin applications. In AV deployment, D1.1 is needed to understand interactions among AVs, human drivers, pedestrians, and cyclists; D1.2 is needed for real time operating awareness, incident handling, and coordination with traffic management systems; and D1.3 is needed for localization, geofencing, rule interpretation, and infrastructure readiness assessment.

Dimension 2: Data Generation Methods

This dimension captures how data is produced and processed. Observational Data (D2.1) includes direct measurements from sensors, surveys, and transactions, as well as derived data streams created through filtering, aggregation, inference, or fusion. Predictive and Analytical Data (D2.2) encompasses outputs from modeling and algorithmic processing, including traffic forecasting, risk assessment, and AI-driven decision support tools such as classification systems and optimization models. Synthetic and Simulation Data (D2.3) accounts for artificially generated data from traffic simulators, agent-based models, virtual testing platforms, and privacy-preserved synthetic datasets, which are especially critical for algorithm development, scenario testing, and methodological benchmarking. For AV applications, D2.1 provides ground truth from real world operations, D2.2 supports motion prediction, risk estimation, demand forecasting, and fleet management, and D2.3 enables scalable testing of rare, safety critical, and edge case scenarios that cannot be captured frequently enough through field data alone.

Dimension 3: Data Source Infrastructure

Dimension 3 details the physical and digital systems through which data is collected. Fixed Infrastructure Sources (D3.1) include roadway-based sensors, traffic management systems, air quality monitors, and the communications networks that support them. These systems offer data from non-mobile sources that can be structured for operational control. Mobile and Distributed Sources (D3.2) include connected vehicle telemetry, smartphone-based mobility data, transit card records, and crowdsourced inputs. These sources support scalable, user-centered insights and often capture real-time movement and decision-making behavior. Fixed sources can provide signal, weather, work zone, and roadside sensing information that extends beyond the onboard field of view, while mobile and distributed sources capture vehicle telemetry, user demand, and network conditions that help agencies and operators monitor AV performance in real time.

Dimension 4: Spatial and Temporal Resolution

This dimension classifies data according to scale and frequency. Spatial Resolution (D4.1) ranges from microscopic levels—individual users, lanes, intersections—to mesoscopic scales such as corridors and networks, and finally to macroscopic levels spanning cities, regions, or nations. Temporal Resolution (D4.2) similarly progresses from sub-second data used in real-time control to longer intervals—hourly, daily, monthly, or yearly—suitable for planning, evaluation, and trend analysis. Event-triggered data is also recognized, representing information generated in response to specific incidents, policies, or user actions. For AV systems, resolution is not just a descriptive attribute but a deployment constraint. Lane level and sub second data are necessary for perception validation, motion planning, and safety analysis, while corridor and region level summaries are necessary for service design, fleet supervision, and assessment of networkwide impacts. Event driven data are especially important for crashes, emergency response, work zones, severe weather, and other disruptions that can temporarily shrink the permissible service area of an AV operational design domain.

Dimension 5: Intelligence Processing Levels

Dimension 5 describes the level of analytic sophistication embedded in the data lifecycle. L0 corresponds to static data capture with minimal processing, used for reference or reporting. This includes baseline maps, infrastructure inventories, and archived design or regulatory information in AV settings. L1 includes real-time monitoring with basic rule-based alerts and system state feedback, supporting AV's live awareness of traffic conditions, incidents, work zones, and weather. L2 introduces forecasting and short-term prediction through pattern detection and statistical learning, which is relevant to AV motion prediction, near term demand estimation, and proactive fleet repositioning. L3 supports AV decision-making, including adaptive control systems and dynamic routing. L4 reflects integrated system coordination across agencies and domains, enabling strategic, multimodal, and interjurisdictional management, becoming significant for large scale AV deployment because it connects fleets with traffic management centers, transit operations, curb management, emergency response, and regional policy objectives.

Taken together, these five dimensions allow for multi-faceted classification of any mobility dataset. This enables not only granular technical specification, but also supports interoperability, discoverability, and reuse across research, operations, and policy domains. The multi-dimensional structure allows datasets to be described in contextually appropriate ways for a variety of users—planners, engineers, data scientists—while maintaining consistency in cataloging, assessment, and system integration efforts. Viewed from an AV perspective, the taxonomy also supports deployment: it helps stakeholders identify which data are already available, which remain fragmented, and where new investment may be helpful for mapping, operations feeds, infrastructure communications, safety monitoring, validation workflows, and privacy preserving data governance before deployment can scale responsibly.

Existing Mobility Data Sources within the Taxonomy

Building on the taxonomy framework detailed in Section 2.1, this section surveys and classifies prominent mobility data sources available in the United States. The following table organizes these datasets using the four-layer Mobility Functional Stack, which progresses from broad population-level patterns down to specific, individual agent data. The layer structure helps users sort datasets by their role in the mobility evidence pipeline before evaluating technical compatibility. Layer 1 sources provide population, demographic, and behavioral priors; Layer 2 sources provide built-environment and operational constraints; Layer 3 sources provide observed movement patterns; and Layer 4 sources provide vehicle- or agent-level operational logs. This organization helps users identify which datasets are comparable within a layer and which datasets are complementary across layers. The five-dimensional taxonomy tags are then used to assess whether selected sources have compatible content, resolution, generation methods, and processing levels for fusion.

For each entry, the table describes the dataset and its use case, its geographical coverage, the primary stakeholders involved, and its level of accessibility. The final column applies the five-dimensional metadata classification as in Figure 1 (D1-D5) to systematically tag each source with its precise technical attributes. This practical application of the taxonomy is designed to help stakeholders compare sources, identify integration opportunities, and recognize data gaps within the current mobility data landscape.

Table 1: Existing Primary Mobility Data Sources. Layer 1 - Population and Travel Behavior

Dataset	Description and Use Case	Coverage	Stakeholder	Access	Taxonomy Tags
American Community Survey (ACS [1]) / Longitudinal Employer-Household Dynamics (LEHD [2])	Provides commute flows, population, and employment distribution, supporting commute analysis.	National (US)	U.S. Census Bureau	Public / Open Data	D1.1 | D2.1 | D3.1 | D4.1 Mesoscopic | D4.2 Low-frequency | D5: L0
American Time Use Survey (ATUS) [3]	Provides diaries of individual time allocation for activities and travel, supporting activity modeling.	National (US)	Federal Agency (BLS)	Public Use Microdata	D1.1 | D2.1 | D3.2 | D4.1 Microscopic | D4.2 Low-frequency | D5: L0
Household Travel Surveys (NHTS [4], CHTS [5])	Describes national and state surveys on travel behavior, trip purpose, and demographics, supporting mobility planning.	National (US)	Federal & State Agencies	Public Use Files	D1.1 | D2.1 | D3.2 | D4.1 Mesoscopic | D4.2 Low-frequency | D5: L0

Note: The "Taxonomy Tags" classify each dataset according to the five-dimensional framework as: D1. Content Type: 1.1 Movement & Behavioral; 1.2 System Operational; 1.3 Infrastructure & Geometric. D2. Generation Method: 2.1 Observational; 2.2 Predictive & Analytical; 2.3 Synthetic & Simulation. D3. Source Infrastructure: 3.1 Fixed Infrastructure; 3.2 Mobile & Distributed. D4. Resolution: 4.1 Spatial (Micro -> Macro); 4.2 Frequency (High -> Low, Event-driven). D5. Intelligence Level: L0 Static -> L4 Integrated.

Table 2: Existing Primary Mobility Data Sources. Layer 2 - Infrastructure and Built Environment

Dataset	Description and Use Case	Coverage	Stakeholder	Access	Taxonomy Tags
OpenStreetMap [6]/Foursquare Places [7] /SafeGraph Building Footprints [8]	Describes millions of points of interest and building footprints, supporting travel pattern analysis.	Global	Community / Private Firm	Public / Proprietary / Subscription	D1.2 | D2.1 | D3.1 | D4.1 Microscopic | D4.2 Low-frequency | D5: L0
National Transit Map (NTM) [9]	Catalogs fixed-route and fixed-guideway transit services nationwide, support transit mapping and planning.	National (US)	USDOT	Public	D1.3 | D2.1 | D3.1 | D4.1 Mesoscopic | D4.2 Low-frequency | D5: L0
National Bridge Inventory [10]	Describes the Nation's bridges on public roads, supporting bridge safety and maintenance.	National (US)	USDOT	Public	D1.3 | D2.1 | D3.1 | D4.1 Microscopic | D4.2 Low-frequency | D5: L0
National Zoning and Land Use Database [11]	Describes zoning codes and restrictions for municipalities, supporting accessibility and mobility planning.	National (US)	Eviction Lab	Public	D1.3 | D2.1 | D3.1 | D4.1 Mesoscopic | D4.2 Low-frequency | D5: L0
Traffic Signal [12]	Representative municipal traffic-signal inventory contains traffic signal locations and attributes in	District of Columbia (US)	District Department of Transportation (DDOT)	Public	D1.3 | D2.1 | D3.1 | D4.1 Microscopic | D4.2 Low-frequency | D5: L0

Dataset	Description and Use Case	Coverage	Stakeholder	Access	Taxonomy Tags
	Washington, DC, supporting traffic operations.				
Traffic Signal Network Device Status Log [13]	Logs communications with devices in traffic signal networks, supporting real-time monitoring.	City of Austin (US)	City of Austin (Texas)	Public	D1.3 | D2.1 | D3.1 | D4.1 Microscopic | D4.2 High-frequency | D5: L1
Model Inventory of Roadway Elements [14]	Lists roadway and traffic inventory elements critical to safety, supporting safety data systems.	National (US)	USDOT	Public	D1.3 | D2.1 | D3.1 | D4.1 Microscopic | D4.2 Low-frequency | D5: L0
PeMS Traffic Online [15]	Provides archived and real-time traffic volume, speed, and occupancy, support roadway operations and safety.	State-specific (CA, US)	State DOTs	Public / Restricted	D1.2 | D2.1 | D3.1 | D4.1 Microscopic | D4.2 High-frequency | D5: L1
HSIS (Highway Safety Information System) [16]	Links crash records with roadway and traffic data, support safety analysis.	National (US)	FHWA / State DOTs (e.g., Caltrans)	Restricted; researchers can request via HSIS portal	D1.2 | D2.1 | D3.1 | D4.1 Microscopic | D4.2 Event-driven | D5: L1
National Performance Management Research Data Set (NPMRDS) [17]	Provides probe-based travel time data for freight and passenger vehicles, supporting performance monitoring.	National (US Highway System)	Federal Agency (FHWA)	Restricted to agencies	D1.2 | D2.2| D3.2 | D4.1 Mesoscopic | D4.2 Medium-frequency | D5: L2

Dataset	Description and Use Case	Coverage	Stakeholder	Access	Taxonomy Tags
WZDx (Work Zone Data Exchange) [18]	Standardizes data on work zone activity, support safety and operations.	National (US)	USDOT / State DOTs	Public feeds (GeoJSON)	D1.2 | D2.1 | D3.1 | D4.1 Mesoscopic | D4.2 Event-driven | D5: L1
RITIS (Regional Integrated Transportation Info System) [19]	Aggregates real-time data from sensors, probes, and incidents, supporting traffic operations.	National (US)	UMD CATT Lab / Participating DOTs	Restricted to agencies	D1.2 | D2.1 | D3.1 | D4.1 Macroscopic | D4.2 High-frequency | D5: L2
MIT AVT [35]	Provides gaze and attention monitoring data, supporting driver monitoring.	Regional (Boston area, US)	MIT AgeLab / AVT Consortium	Application required; access for research partners	D1.1 | D2.1 | D3.2 | D4.1 Microscopic | D4.2 Medium-frequency | D5: L2
SHRP 2 NDS [36]	Provides high-resolution crash and near-crash driver behavior data, supporting safety research.	National (US) (US, across six states)	TRB / VTTI	Restricted; access via InSight data portal	D1.1 | D2.1 | D3.2 | D4.1 Microscopic | D4.2 Medium-frequency | D5: L1

Table 3: Existing Primary Mobility Data Sources. Layer 3 - Movement Tracking and Analytics

Dataset	Description and Use Case	Coverage	Stakeholder	Access	Taxonomy Tags
Veraset [20], Cuebiq [21], Spectus [22]	Provides raw pseudonymous mobile app traces, supports mobility analytics.	Primarily US	Private data firms	Proprietary / License	D1.1 | D2.1 | D3.2 | D4.1 Microscopic | D4.2 High-frequency | D5: L1
StreetLight Data [23]	Aggregates origin–destination flows and route choices from mobile data,	US & Canada	Private analytics firm	Proprietary / Subscription	D1.1 | D2.2 | D3.2 | D4.1 Mesoscopic | D4.2 Medium-freq

Dataset	Description and Use Case	Coverage	Stakeholder	Access	Taxonomy Tags
	support mobility analytics.				uency | D5: L2
NYC TLC Trip Data [24]	Provides trip records from taxis and for-hire vehicles, supporting urban mobility analysis.	New York City	Municipal Agency (NYC TLC)	Public / Open Data	D1.1 | D2.1 | D3.2 | D4.1 Microscopic | D4.2 High-freque ncy | D5: L1
INRIX Global Traffic Data [25]	Provides real-time and historical traffic speeds and congestion, supporting traffic analytics.	Global	INRIX, Inc.	Commercial license; used by agencies and firms	D1.2 | D2.1 | D3.2 | D4.1 Mesoscopic | D4.2 High-freque ncy | D5: L2
Waze for Cities (Connected Citizens Program) [26]	Collects user-reported incidents, congestion, and hazards, supporting real-time operations.	Global	Waze / Google (Alphabet)	Free for public agencies via partnership	D1.2 | D2.1 | D3.2 | D4.1 Mesoscopic | D4.2 Event-driven | D5: L1
HERE Location Services [27]	Provides real-time and historical probe data, support logistics and routing.	Global	HERE Technologies	Commercial license; widely used in OEM systems	D1.2 | D2.1 | D3.2 | D4.1 Mesoscopic | D4.2 High-freque ncy | D5: L2
TomTom Traffic Index [28]	Provides probe-based traffic data, supporting benchmarking.	Global	TomTom International BV	Summary index is public; detailed data under license	D1.2 | D2.1 | D3.2 | D4.1 Macroscopi c | D4.2 Low-frequen cy | D5: L2

Table 4: Existing Primary Mobility Data Sources. Layer 4 - Individual Agent Data & Shared Mobility

Dataset	Description and Use Case	Coverage	Stakeholder	Access	Taxonomy Tags
Waymo Open Dataset [29]	Provides sensor data for simulation platforms like CARLA and LGSVL, supporting perception and planning.	City-specific (multiple US cities)	Waymo (Alphabet)	Public; license required	D1.1 | D2.1 | D3.2 | D4.1 Microscopic | D4.2 High-frequency | D5: L3
nuScenes [30]	Provides radar and 360° camera datasets with annotations, support perception and prediction.	City-specific (Boston, USA & Singapore)	Motional	Public for academic use	D1.1 | D2.1 | D3.2 | D4.1 Microscopic | D4.2 High-frequency | D5: L3
Argoverse [31]	Provides annotated data for motion forecasting and HD maps, supporting trajectory prediction.	City-specific (Pittsburgh, PA & Miami, FL)	Argo AI	Public	D1.1 | D2.1 | D3.2 | D4.1 Microscopic | D4.2 High-frequency | D5: L3
Lyft Level 5 [32]	Provides LiDAR and camera datasets with HD urban maps, support perception and simulation.	City-specific (San Francisco, CA)	Lyft Level 5	Public despite division shutdown	D1.1 | D2.1 | D3.2 | D4.1 Microscopic | D4.2 High-frequency | D5: L3
Comma2k19 [33]	Records continuous driving logs with video, location, and CAN data, supporting	City/Regional (San Francisco Bay Area, CA)	comma.ai	Public (MIT License)	D1.1 | D2.1 | D3.2 | D4.1 Microscopic | D4.2 High-frequency | D5: L1

Dataset	Description and Use Case	Coverage	Stakeholder	Access	Taxonomy Tags
	driver state modeling.				
100-Car Naturalistic Driving Study [34]	Links driver behavior and crash risk with in-vehicle data, supporting safety research.	Regional/National (US)	Virginia Tech Transportation Institute (VTTI)	Application required; privacy protected	D1.1 | D2.1 | D3.2 | D4.1 Microscopic | D4.2 Medium-frequency | D5: L1
MIT AVT [35]	Provides gaze and attention monitoring data, supporting driver monitoring.	Regional (Boston area, US)	MIT AgeLab / AVT Consortium	Application required; access for research partners	D1.1 | D2.1 | D3.2 | D4.1 Microscopic | D4.2 Medium-frequency | D5: L2
SHRP 2 NDS [36]	Provides high-resolution crash and near-crash driver behavior data, supporting safety research.	National (US) (US, across six states)	TRB / VTTI	Restricted; access via InSight data portal	D1.1 | D2.1 | D3.2 | D4.1 Microscopic | D4.2 Medium-frequency | D5: L1

Case Study in Los Angeles County

As established in the survey of mobility data, human mobility modeling is a critical field of research that underpins effective urban planning and transportation management ([37], [38]). However, a primary challenge limiting progress is the difficulty of integrating the vast number of available datasets, which are often heterogeneous, incomplete, and sparse. This data fragmentation, illustrated in Figure 2(a), means that existing data sources typically offer only a partial view of mobility, hindering the development of comprehensive and robust models.

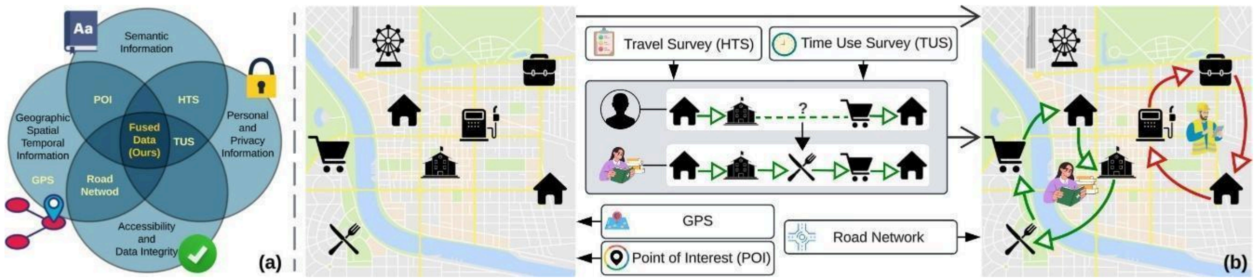

Figure 2: Integration of cross-domain multi-modal mobility data with human mobility modeling. (a) The need for data fusion and (b) the fused data

To address this gap, this project introduces a foundation model framework designed for universal human mobility patterns. As shown in the conceptual diagram in Figure 2(b), our approach leverages cross-domain data fusion to synthesize a complete and coherent mobility dataset. The framework integrates multi-modal data of distinct natures and spatio-temporal resolutions, including geographical (GPS, POI), mobility (Household Travel Surveys), demographic (Time Use Surveys), and traffic information—to construct a privacy-preserving and semantically enriched dataset of human travel trajectories. This is achieved using large language models (LLMs) for semantic enrichment and advanced transfer learning techniques to ensure the model is adaptable across a variety of contexts.

The practical utility and effectiveness of this foundation model were validated through a large-scale case study in Los Angeles County. When the generated synthetic data was used in a traffic simulation, the results aligned closely with real-world observations. On California's I-405 corridor, a notoriously complex traffic environment, the simulation yielded a Mean Absolute Percentage Error (MAPE) of just 5.85% for traffic volume and 4.36% for speed when compared to Caltrans PeMS sensor data. This high level of accuracy illustrates the framework's significant potential for improving intelligent transportation systems and real-world urban mobility applications.

This case study makes several significant contributions:

- We developed a novel foundation model that processes and integrates cross-domain multi-modal data sources to create comprehensive synthetic datasets, while balancing usability and privacy considerations. The framework harmonizes heterogeneous data types and demonstrates generalization capabilities through validation with large-scale traffic simulation.
- We incorporated LLMs as a core component for semantic enrichment of trajectory data, fusing the realism of GPS data with the knowledge of survey data and reconstructing missing activities with a survey-trained model. This LLM-powered approach significantly improves the completeness and interpretability of mobility trajectories.
- We implemented transfer learning mechanisms within our foundation model to address two key data challenges: enabling knowledge and pattern adaptation in data-sparse regions through semi-supervised learning, leveraging limited GPS trajectories to adapt source domain patterns without requiring complete ground truth data, and incorporating demographic variations into anonymized trajectories through unsupervised distribution adaptation, thus aligning mobility patterns and demographic characteristics without pairwise individual labels.

Problem Formulation

To address the multi-modal and cross-domain challenges, we propose a three-stage fusion framework that advances beyond existing approaches typically limited to combining similar data sources. As shown in Figure 3, our framework integrates a variety of data types to generate realistic human mobility patterns.

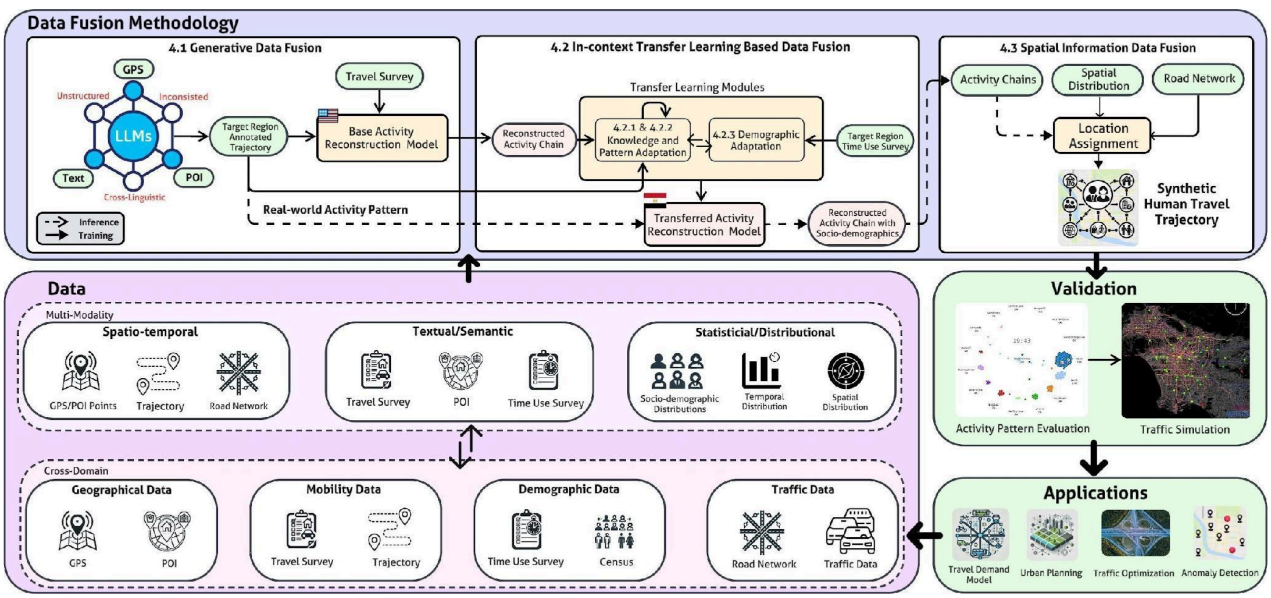

Figure 3: An overview of the proposed data fusion framework. The captions within methodology indicate their corresponding sections in the paper for reference

The generative module first employs LLMs to fuse unstructured POI data with GPS stay points, creating semantically annotated but temporally incomplete trajectories due to GPS fragmentation. These trajectories are then completed through an activity reconstruction model incorporating mobility patterns learned from travel surveys. The transfer learning stage is designed to capture real-world mobility dynamics while enabling deployment in data-scarce regions. Through knowledge and pattern adaptation, the model can be adapted to different regions and cultures and learns to balance the generic knowledge learned from surveys with the realism in GPS trajectories, including region-specific patterns. Our demographic adaptation module further enhances realism by incorporating demographic variations using Time Use Survey data, which provides rich activity patterns across different population groups, ensuring our synthetic trajectories reflect both demographic-specific behaviors and real-world mobility dynamics while maintaining privacy.

Finally, the third stage grounds these behaviorally-realistic patterns in physical space by fusing the generated activity chains with spatial distributions and traffic data to assign feasible locations, enabling validation through large-scale traffic simulation. This complete integration of dynamic human behavior with static infrastructure data creates a comprehensive framework that generates mobility patterns reflecting both the complexity of human decision-making and the constraints of physical infrastructure.

Table 5: Aggregated activity types

ID	Type	ID	Type	ID	Type
1	Home	6	Buy services (dry cleaners, banking, service a car)	11	Visit friends or relatives
2	Work	7	Buy meals (go out for a meal, snack, carry-out)	12	Health care visit (medical, dental, therapy)
3	School	8	Other general errands (post office, library)	13	Religious or other community activities
4	Child/Adult Care	9	Recreational activities (visit parks, movies, bars)	14	Something else
5	Buy goods (groceries, clothes, appliances, gas)	10	Exercise (go for a jog, walk, walk the dog)	15	Drop off /pick up someone

Data Sources

Our modeling framework is designed to capture the full behavioral and contextual richness of human mobility. To meet this objective, we select datasets based on their ability to address specific modeling needs: understanding why and who travels, capturing where and when movement occurs, and interpreting what contextual features shape those decisions. Because no single dataset fulfills all of these requirements, we integrate complementary sources from three functional layers of the mobility data landscape.

To model behavioral intent and activity structure, we use the American Time Use Survey (ATUS) and the National Household Travel Survey (NHTS), both from Layer 1: Population and Travel Behavior. ATUS offers detailed time-allocation data that reveals how individuals structure daily activities across home, work, and discretionary domains. NHTS provides trip-level records tied to demographic profiles and travel modes, making it highly suitable for learning interpretable and policy-relevant behavioral patterns. These surveys serve as the foundation for training, enabling the model to learn the semantics of human travel, though their limited geographic coverage necessitates transfer learning.

To capture fine-grained, real-world movement dynamics, we incorporate Veraset GPS traces from Layer 3: Movement Tracking and Analytics. Veraset data provide anonymized, high-frequency device-level trajectories that allow us to observe where and when individuals travel at scale, across diverse regional contexts. While precise, this data lacks information on activity purpose or individual characteristics.

To enrich these trajectories with semantic meaning and spatial context, we integrate SafeGraph POI data and OpenStreetMap (OSM) features from Layer 2: Infrastructure and Environment. For the Los Angeles case study, these sources are aligned to the same metropolitan context. SafeGraph offers detailed, regularly updated information on building footprints and commercial activity locations, which is critical for inferring plausible activities from GPS stay-points. OSM complements this with extensive coverage of public amenities, road networks, and non-commercial points of interest, offering broad contextual grounding for both urban and non-urban settings. By fusing (Layer 1) American Time Use Survey and NHTS to learn activity intent and demographic structure, (Layer 3) Veraset GPS stay-points to model high-resolution mobility patterns, and (Layer 2) SafeGraph and OSM to interpret movement in physical and semantic space, our framework constructs a behaviorally interpretable, context-aware mobility model that generalizes across space.

Methodology

The detailed methodology and validation have been published in [40]. Readers are referred to that paper for a full description. This report provides only a high-level summary and key findings

LLM-Informed Trajectory Semantic Annotation

Human trajectory data collected from GPS devices include geo-location and time-stamped stay points that elucidate individual mobility patterns. However, these trajectories suffer from both semantic sparsity and temporal incompleteness. We develop the generation-based data fusion module, as Figure 3 (4.1), to address these two challenges. By adopting the methodology from our previous work [39], we use LLMs to annotate stay points using the POI dataset. Each POI is then categorized into one of 15 activity codes as described in Table 5. Then a Bayesian-based algorithm is used to associate stay points with the most appropriate POI and activity type, as $T = argmax_{T_i}\left\{P\left(T_i|POI, S\right)\right\}$, where T is the activity at the stay point, given POIs and start time S.

This yields semantically enriched trajectories suitable for mobility modeling.

Data Augmentation by Activity Reconstruction

Although LLM-based annotation adds semantics, trajectories remain incomplete and unsuitable for direct use. To address this, we apply an activity reconstruction model trained on survey data to generate full daily activity chains from partial GPS stay points, leveraging the similarity of human activity patterns across regions [40][41]. One-day activity chains are encoded into 96 time slots and the model uses masking strategies to simulate missing data. A Transformer-based sequence model then reconstructs complete activity chains, trained with loss functions that balance activity accuracy, temporal coherence, and sequence similarity (see Figure 4).

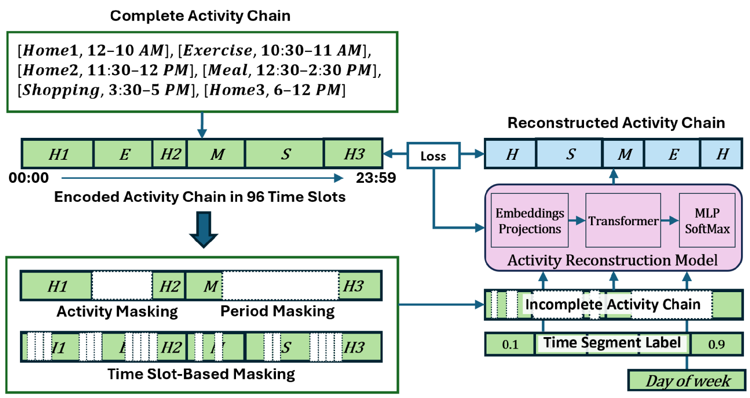

Figure 4: Network architecture of the activity reconstruction model

While this LLM-based annotation provides semantic context to the trajectories, their inherent incompleteness limits their direct use for mobility analysis. Therefore, our framework continues with an activity reconstruction process that fuses these annotated but incomplete trajectories with comprehensive mobility data from travel surveys. A base Activity Reconstruction Model $M0$ is trained on survey data to learn generic activity patterns and generate complete activity chains based on limited observed GPS stay points. It can be adapted to other regions with a small dataset, supported by the strong similarity in human activities across regions [40], [42]. The adaptation of this base model to realistic geographical and cultural contexts is detailed in In-context Transfer Learning Section.

One-day activity chains are encoded into 96 time slots and the model uses masking strategies to simulate missing data. A Transformer-based sequence model then reconstructs complete activity chains, trained with loss functions that balance activity accuracy, temporal coherence, and sequence similarity, as shown in Figure 4.

In-context Transfer Learning for Domain Adaptation

While our generation module effectively reconstructs activity chains, adapting to diverse regional contexts and demographic characteristics is crucial for realistic mobility patterns. As shown in Figure 3 (4.2), our transfer learning approach addresses this through three adaptations: 1) knowledge adaptation for region-specific knowledge when regional survey data is available, 2) mobility pattern adaptation integrating survey data with GPS data while retaining universal activity motifs, and 3) demographic-aware adaptation for demographic-specific patterns.

When regional survey data are available, 1) Knowledge Adaptation is achieved by applying supervised transfer learning using a progressive unfreezing strategy. This approach fine-tunes embedding layers, Transformer layers, and the output module in stages, ensuring that general knowledge from the source domain is preserved while adapting to new regional characteristics. By gradually unfreezing layers, the model retains core behavioral structures while learning region-specific activity relationships.

In regions where only partial GPS data are available, 2) Mobility Pattern Adaptation is implemented with a semi-supervised adaptation strategy. Starting from a survey-trained base model, the system iteratively generates synthetic datasets that are refined with real GPS inputs until the synthetic and observed patterns converge. This process balances survey-derived behavioral logic with the realism of trajectory data, allowing adaptation even in data-scarce settings while preventing overfitting through dataset retention across iterations.

To incorporate demographic variation, we use aggregated Time Use Survey distributions as guidance for the 3) Demographic Adaptation. A learnable adapter modifies reconstructed activity chains so that generated trajectories align with demographic-specific activity distributions, such as those by age, sex, or employment status. The optimization process ensures both global alignment with survey distributions and local consistency in activity timing, producing demographically representative mobility patterns without requiring paired GPS-demographic datasets.

Activity Location Assignment

To translate our generated activity chains into real-world use, we developed a location assignment module that merges spatial limits and traffic data, as in Figure 3, which assigns each activity to specific TAZs through a three-layer approach. The process begins by fixing mandatory activities such as home, work, and school, using commute distances sampled from demographic distributions and filtered by land use and commuting patterns to establish realistic anchors. Then non-mandatory activities like shopping and leisure are assigned between these anchors with consideration of distance, direction, and network travel times to maintain feasible trip chains. Finally, the overall spatial distribution is refined by comparing activity frequencies with observed data and adjusting parameters until regional patterns converge, ensuring that the resulting trajectories are both behaviorally realistic and spatially consistent with real-world contexts.

Results

This LA study demonstrates our complete pipeline capabilities and transfer learning in data-scarce regions. We compare the output synthetic data against multiple sources, such as HTS, Time Use Survey, and raw GPS points, as our baseline, which helps reflect the disagreement on the same measurements across different data sources.

Data preprocessing

For our Los Angeles County case study, we used half-year GPS points from Veraset [20]. Raw trajectories were processed into daily stay points with at least 10-minute dwell times. After filtering, we retained about 1 million high-quality daily stay points with a minimum of 55% daily coverage and more than one activity per trajectory. We built a synthetic population using a rule-based classification approach. Around 70% of annotated trajectories included work-related activities, consistent with county employment rates [47]. Trajectories with work or school activities were classified as workers or students, while uncertain cases remained unclassified to avoid misattribution. For training, we used the Southern California Association of Governments (SCAG) Activity-Based Model (ABM) dataset, which generates detailed activity chains for the region [38]. We chose SCAG over NHTS because SCAG provides a much larger Southern California training corpus and finer regional activity-chain detail, with 700,000 training, 200,000 validation, and 100,000 test samples. This combination of larger sample size and region-specific detail makes SCAG more suitable than the smaller national NHTS sample for transformer-based training in the Los Angeles case study. Evaluation employed Jensen-Shannon Divergence (JSD) for activity distributions (length, duration, type, temporal patterns), cosine similarity for Origin-Destination matrices, and Mean Absolute Percentage Error (MAPE) for traffic metrics.

LLM-based POI classification and assignment validation

We fused daily stay points with POI data using the LLM-based semantic enrichment method from [39] and validated against Los Angeles County ground truth. Figure 5 and Figure 6 show examples of POI labeling and annotated trajectories. For evaluation, we sampled 500 POIs per region and obtained ground truth from five independent annotators. Using GPT-3.5, our LLM achieved 91.4% accuracy, 92.9% F1-score, and 75.7% Hit@1, showing robust performance across geographic and linguistic contexts. We further validated POI assignment with a Bayesian method combining LLM classifications and temporal activity patterns. Using the NHTS California Add-on dataset as ground truth, we tested on 362 individuals with 2007 activities (1724 mandatory and 283 non-mandatory).

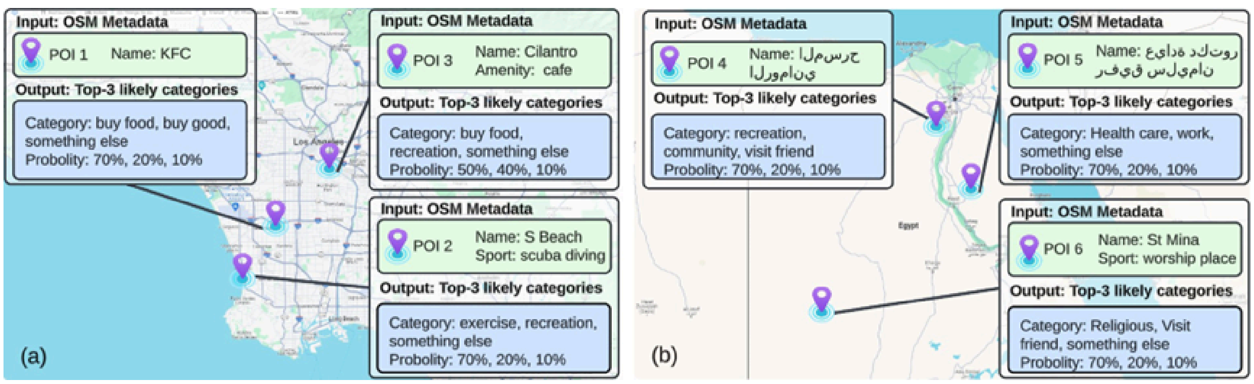

Figure 5: POI labeling using LLM by associating three activity types with likelihood estimation (a) Los Angeles, USA (b) Cairo, Egypt as examples

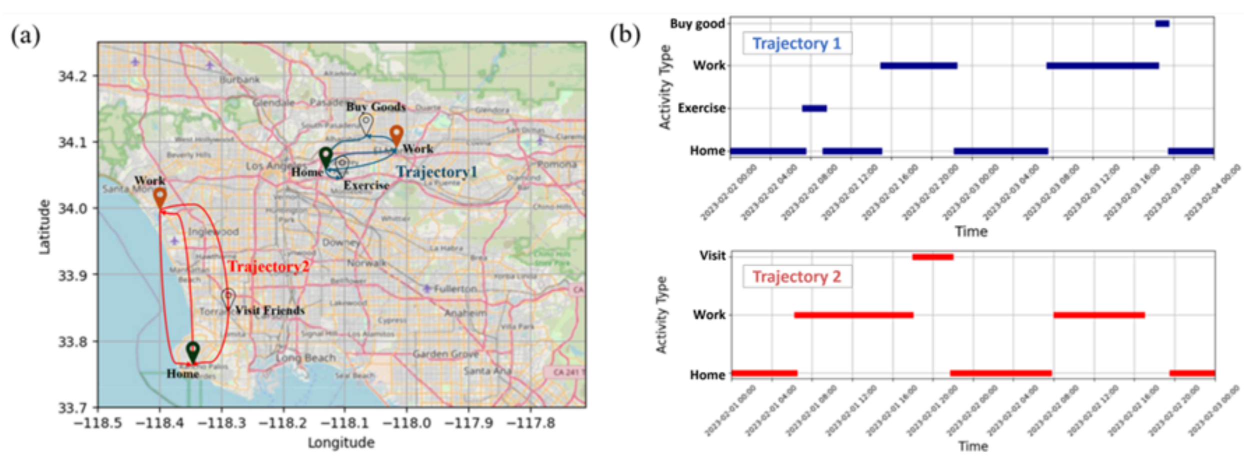

Figure 6: Semantic annotated trajectories with activity type

Table 6: Performance metrics for POI classification with different approaches

Metric	LA (gpt3.5)	LA (gpt4)	LA (Heuristic)
Accuracy@1	82.5%	91.4%	49.4%
Accuracy@2	15.4%	8.7%	0.4%
Accuracy@3	1.9%	7.0%	0.0%
F-1 Score	82.1%	92.9%	51.7%

Our activity inference validation achieved 91.7% accuracy and 92.3% F1-score. Mandatory activities such as home, work, and school reached 98.3%, 87.2%, and 90.0% accuracy, while non-mandatory activities varied, with healthcare and religious activities at 100% and more ambiguous cases like visiting friends at 52.2%.

Noise tolerance tests with Gaussian noise (5 m, 10 m, 20 m) showed stable performance, with Accuracy@3 remaining above 84% (88.4%, 85.3%, 84.2% respectively), confirming robustness to GPS errors. The Bayesian inference component combined LLM-based POI probabilities with temporal likelihoods from survey data, supporting context-aware classification and validating the full trajectory annotation pipeline from raw GPS to activity sequences.

As a baseline, we implemented a rule-based POI classification using OSM tags mapped to 15 activity codes. It achieved 49.4% Accuracy@1 and 51.7% F1-score for LA County, far below the LLM approaches (over 90%), as shown in Table 6. Accuracy@2 and @3 also lagged, reflecting limited handling of ambiguous or multi-purpose POIs. These results underscore the significant advantage of LLM-based methods over heuristic baselines.

Base model evaluation on survey data

The performance of the base model M^N in reconstructing 70% masked NHTS data and its transferability to the SCAG datasets is quantified in Table 7. The model achieves strong performance on NHTS data and is evaluated by Jensen-Shannon Distribution (JSD) values ranging from 0.001 to 0.017 across all metrics. JSD values capture the similarity between two or more probability distributions, and lower JSD values indicate closer agreement between the generated and reference distributions, with 0 indicating identical distributions and 1 indicating completely distinct distributions. Here this result (0.001–0.017 on NHTS) suggests that the NHTS and synthetically derived datasets are highly similar. However, performance degradation is observed when evaluating the NHTS-trained base model M^N on SCAG reference data without transfer learning, revealing significant regional differences in mobility patterns.

The JSD values between NHTS and SCAG, which range from 0.009 to 0.059 across various activity attributes, quantify these differences and are visualized in Figure 7. Supervised transfer learning is applied to M^N to obtain M^S, yielding low JSD values when reconstructing the SCAG dataset. This result shows the algorithm effectively adapts to different geographical contexts and data formats, despite initial disparities between SCAG and NHTS datasets. Beyond its commendable performance in system-level JSD evaluations, the proposed model exhibits proficiency in capturing patterns of common activities, with the transfer learning successfully adapting the model from regional to national contexts. As illustrated in Figure 8, the model demonstrates remarkable accuracy in reconstructing the temporal patterns of activity starting times across both datasets.

Table 7: JSD for distributional differences across source datasets and model outputs (lower is better)

JSD values	Length	Duration	Type	Start time	End time
NHTS vs SCAG	0.017	0.009	0.059	0.018	0.013
M^N **on NHTS**	0.017	0.002	0.001	0.001	0.005
M^N on SCAG	0.021	0.003	0.017	0.004	0.007
M^S **on SCAG**	0.011	0.004	0.004	0.001	0.001

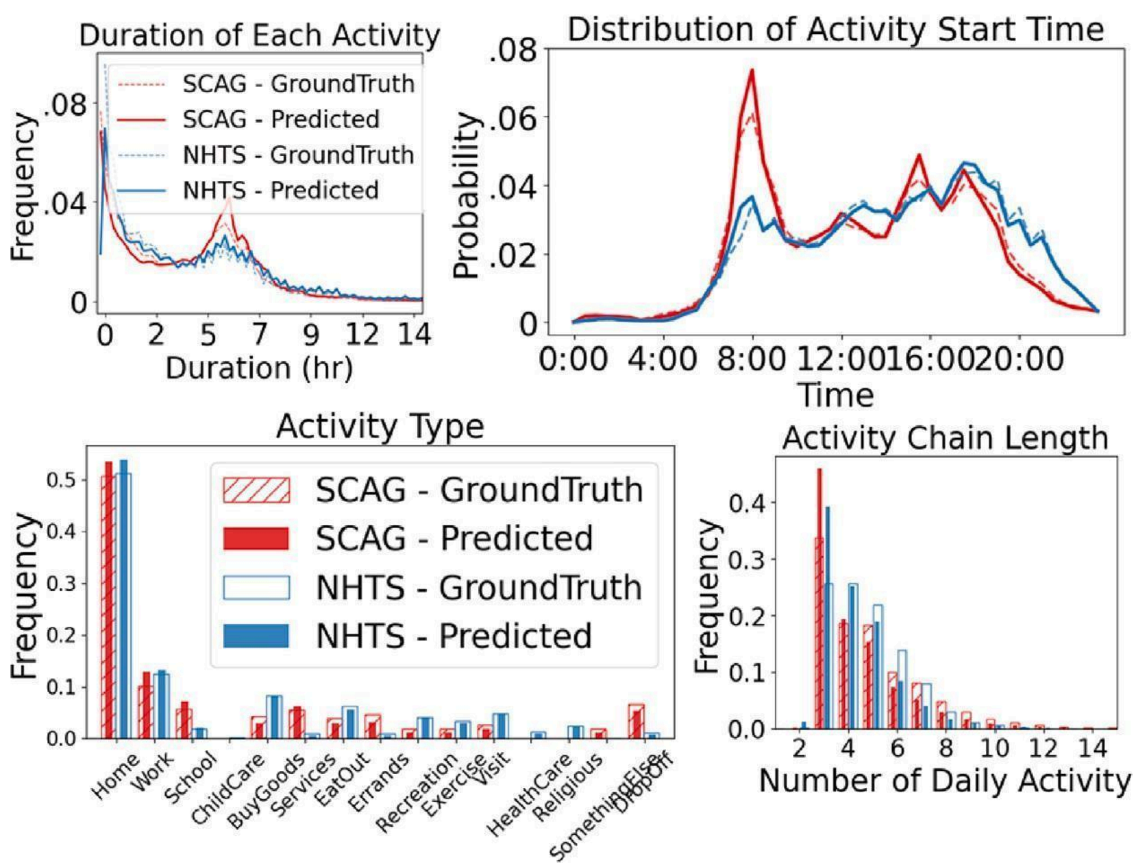

Figure 7: Performance evaluation on NHTS and SCAG with 70% masked activity chain

Activities show closely aligned patterns between the two datasets, evidenced by low JSD values in Table 7, highlighting the universality of these routines. Conversely, regional variations are apparent in other activities. Shopping patterns, for example, differ significantly, with SCAG data showing a midday peak (JSD 0.054) versus NHTS's more even distribution, suggesting distinct regional shopping behaviors. Exercise activities also vary, with NHTS indicating a preference for early morning sessions (JSD 0.036). These differences likely mirror varying lifestyle choices between urban Los Angeles and the broader national context. The model's consistently low M^S and M^N values, typically below 0.01, confirm its efficacy in capturing these nuances, underscoring its utility for urban planning and policy-making by identifying both universal and regional behavioral patterns.

Transfer learning evaluation on GPS and survey data

JSD values across iterations for the evaluation metrics are shown in Table 8, showing how the iterative transfer learning process improves alignment with the target region. In this table, only three metrics are included because the lack of ground truth in incomplete data renders activity duration and chain length unreliable for comparison, whereas the selected metrics are more accurate at the system level and can better reflect regional mobility patterns. All values reported for Iteration 1 through Iteration 6 are the mean ± standard deviation over three random seeds. Activity type JSD drops from 0.0508 at baseline to 0.0028 by Iteration 3.

Table 8: Model evolution during iterative transfer learning for LA case

JSD values	Activity type	Start time	End time
NHTS vs LA GPS	0.0508	0.0289	0.0367
Iteration_1	0.0131 ± 0.0033	0.0018 ± 0.0006	0.0044 ± 0.0007
Iteration_2	0.0041 ± 0.0011	0.0015 ± 0.0004	0.0042 ± 0.0006
Iteration_3	**0.0028** ± **0.0004**	0.0010 ± 0.0002	0.0030 ± 0.0003
Iteration_4	0.0030 ± 0.0003	**0.0008** ± **0.0001**	**0.0022** ± **0.0003**
Iteration_5	0.0052 ± 0.0008	0.0027 ± 0.0005	0.0054 ± 0.0007
Iteration_6	0.0081 ± 0.0012	0.0050 ± 0.0009	0.0085 ± 0.0011

Start and end time JSDs reached their lowest values (0.0008 and 0.0022) at Iteration 4, after which performance declined, consistent with prior findings that recursive training on synthetic data erodes rare activity patterns [48]. Figure 9 illustrates how the NHTS-based model progressively adapts to the LA context via Semi-Supervised Transfer Learning with GPS data.

In the second step, demographic adaptation was performed using time-use survey data. We trained 12 demographic adapters by sex, age, and employment status. Workers achieved higher reward values (0.95–0.98) and faster convergence, reflecting structured daily patterns. Mid-aged and elderly workers converged within 400–500 epochs, while young workers required >1000 epochs but achieved comparable rewards. Non-workers showed greater variability, converging more slowly with slightly lower rewards (0.82–0.93).

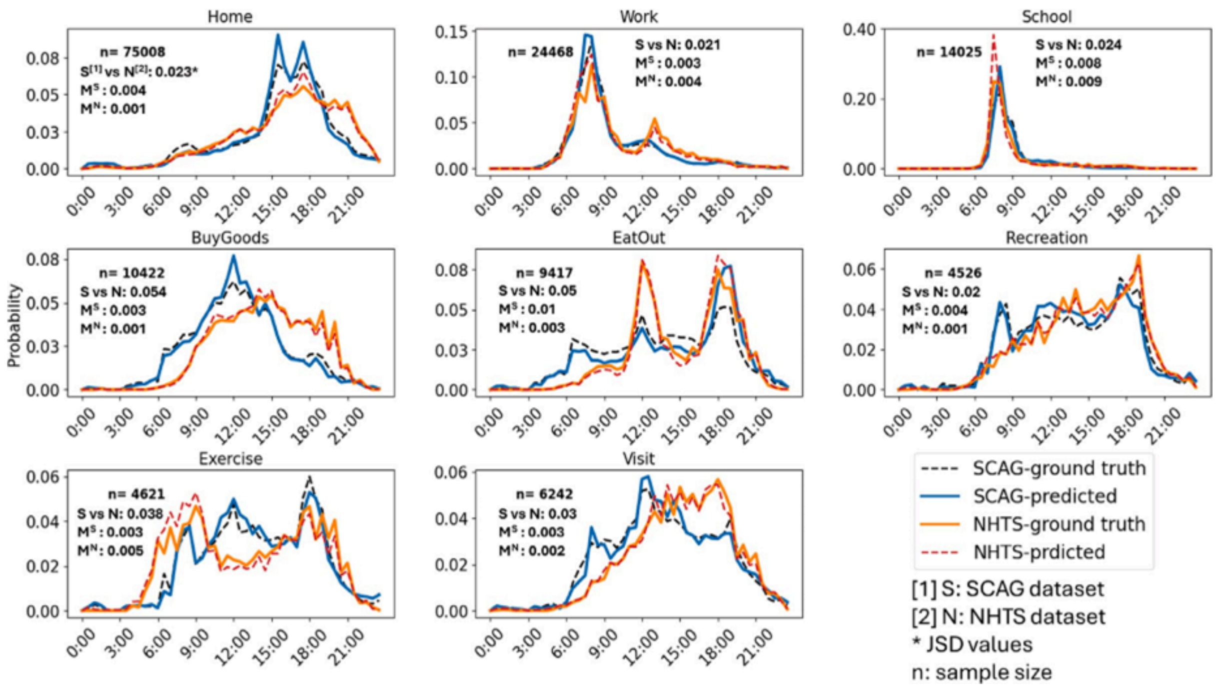

Figure 8: The start time of common activities reconstructed by base and its transferred model using NHTS and SCAG

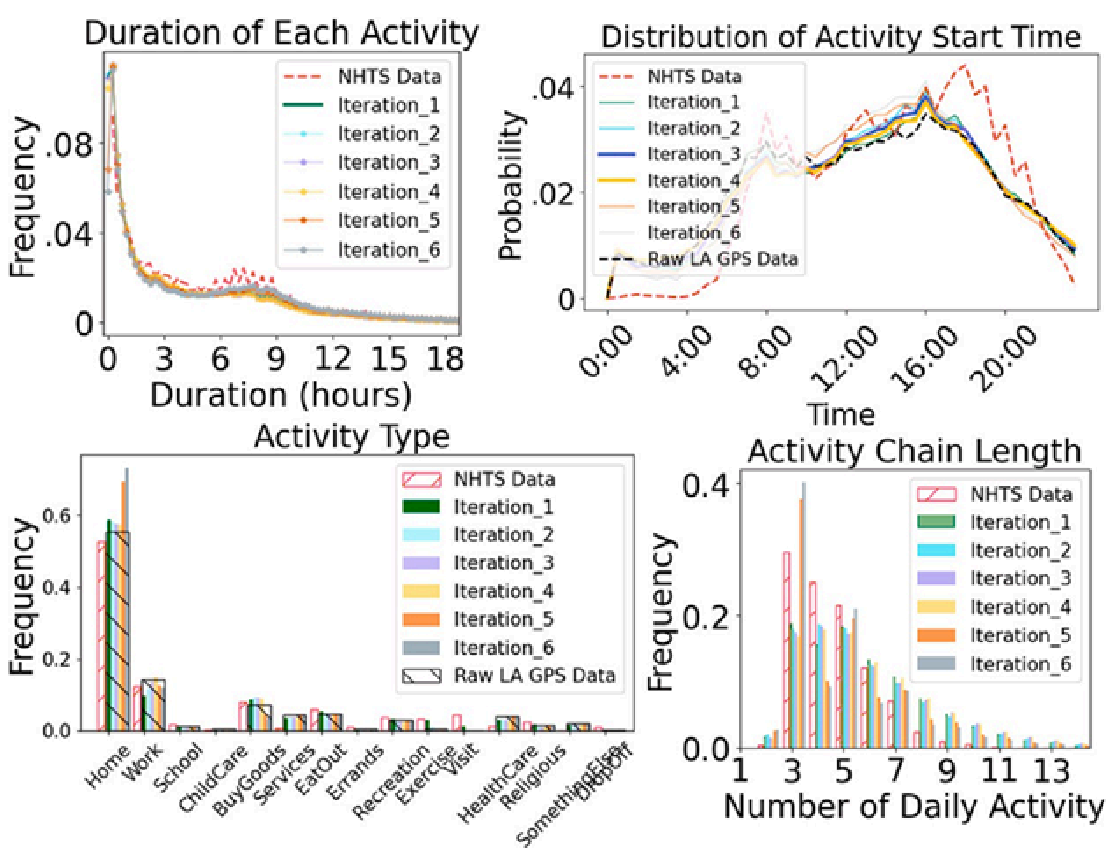

Figure 9: Iterative adaptation from NHTS to LA via semi-supervised transfer learning with LA GPS data

Comparison between generated patterns with time use survey and HTS data is presented in Table 9. Alignment with time use survey was strong across all demographic groups (JSD: 0.006–0.018), validating the adapters' ability to capture demographic-specific activity patterns. Larger JSDs relative to HTS (0.048–0.094) are observed and should be interpreted as cross-source differences, as time use survey and HTS themselves differ (0.011–0.063). This is expected because the demographic-adaptation step is guided by Time Use Survey distributions, whereas HTS serves as an external regional validation benchmark rather than a direct training target in this stage. This approach relies more on Time Use Survey, because compared with HTS, time-use surveys can be less costly and less burdensome to field, and they are often collected on a regular annual basis in national statistical systems, whereas household travel surveys are typically periodic [49][50][51].

Table 9: JSD value for similarity measurement of activity duration distribution across demographic groups

Group	HTS vs Output	Time use surveys vs Output	Time use survey vs HTS
Young M	.083/.086	.008/.009	.014/.014
Young F	.094/.085	.015/.013	.016/.011
Mid-aged M	.048/.067	.018/.006	.016/.054
Mid-aged F	.055/.066	.018/.007	.022/.059
Elder M	.070/.066	.015/.009	.020/.055
Elder F	.077/.071	.013/.009	.029/.063

Note: Values shown as (Worker/Non-worker); M: Male F: Female

The generation demonstrates strong alignment with the Time Use Survey across all demographic groups. Quantitative evaluation using JSD shows excellent performance (detailed results in Table 10), with non-worker groups achieving particularly strong alignment (JSD 0.0064–0.0103) and worker groups showing comparable performance (JSD 0.0085–0.0189). Mid-aged male non-workers achieve the best performance (JSD = 0.0064) while maintaining realistic activity patterns, such as work duration differences across age groups and sex-specific activity preferences.

Beyond activity durations, we evaluate temporal patterns to ensure the demographic adapter preserves learned temporal dynamics while incorporating demographic variations. The temporal evaluation spans demographic groups (Figure 10) and activity- specific distributions (Figure 11). At the demographic level, mid-aged and elderly groups show consistent patterns across both worker and non-worker categories. The young group shows some discrepancies between time use survey and HTS because the young worker group is insignificant in HTS (around 2% of the total population).

At the aggregate level, the model maintains strong agreement with both raw GPS data (JSD: 0.004) and NHTS patterns (JSD: 0.014), capturing key features like morning (8:00–10:00) and afternoon (16:00–18:00) peaks. At the activity level, as shown in Figure 11, major activities like work and school are well captured and the patterns learned from the base model were maintained throughout the complex transfer learning process.

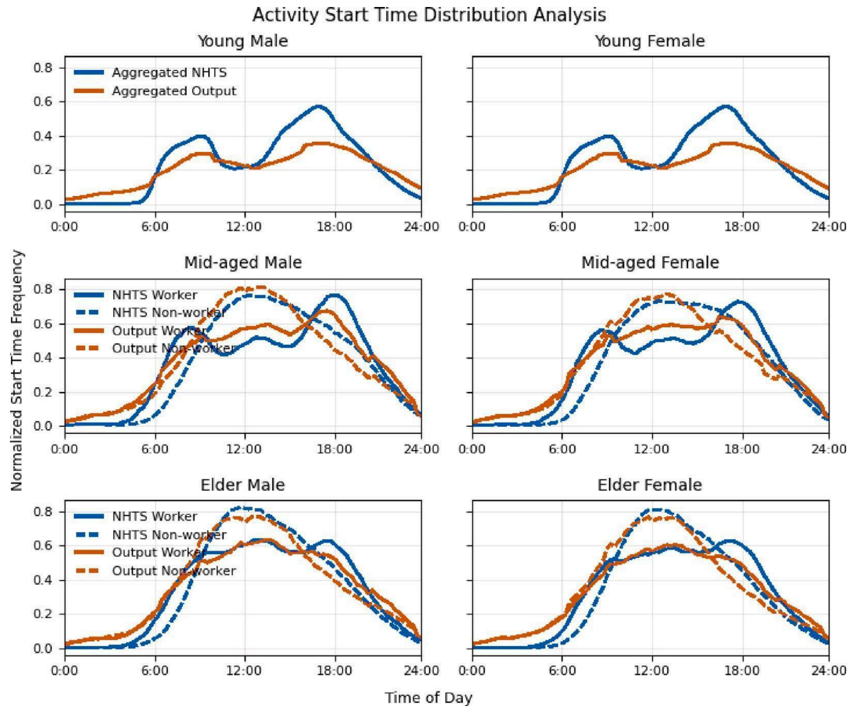

Figure 10: Temporal performance evaluation per demographic group

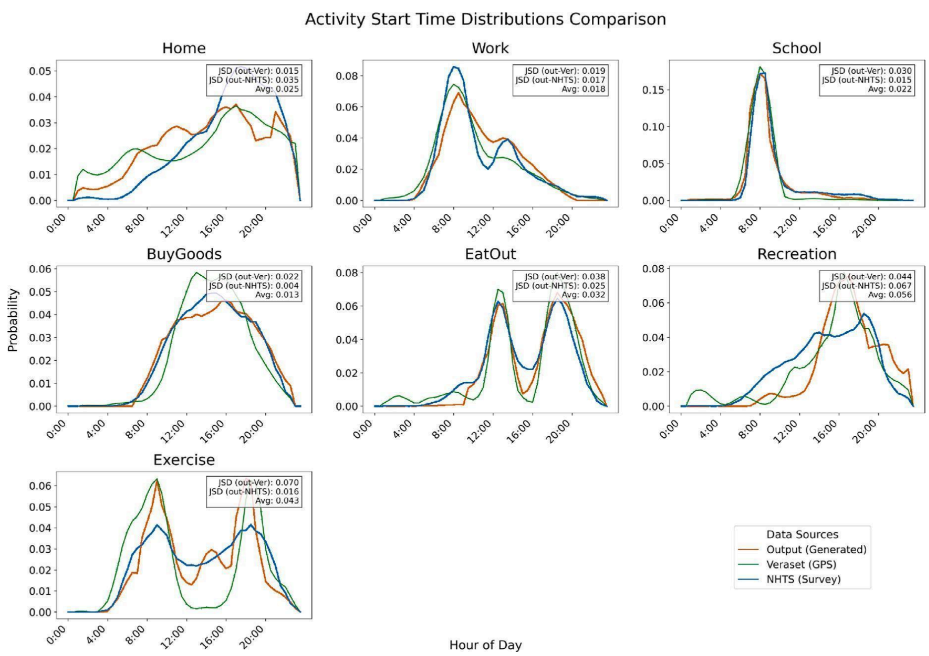

Figure 11: Activity-specific temporal pattern comparison: start time distributions for major activities across different data sources

Table 10: L.A. County major activity duration (minutes/day) comparison with JSD

Demographic group (JSD)	Activity duration (minutes)							
	Home	Work	School	Buy goods	Meals	Recreation	Exercise	Visit
	(1)	(2)	(3)	(5)	(7)	(9)	(10)	(11)
Young Male W. (0.0085)	843.7/ 837.5	209.5/ 224.5	117.7/ 134.7	19.6/ 15.0	52.5/ 38.5	100.7/ 53.3	23.8/ 18.6	61.5/ 53.4
Young Female W. (0.0159)	820.2/ 785.3	176.2/ 204.7	131.5/ 134.8	39.8/ 20.8	57.3/ 34.3	70.2/ 34.6	23.5/ 18.8	75.8/ 88.8
Mid-Aged Male W. (0.0189)	798.4/ 800.4	380.8/ 382.9	0.0/ 0.0	37.9/ 19.1	61.7/ 77.0	76.2/ 62.2	23.9/ 4.7	42.4/ 1.0
Mid-Aged Female W. (0.0185)	839.0/ 825.6	333.0/ 330.8	0.0/ 0.0	42.4/ 22.7	57.0/ 74.3	63.2/ 58.7	16.8/ 4.4	46.1/ 2.1
Elderly Male W. (0.0157)	875.0/ 886.0	295.6/ 284.5	0.0/ 0.0	34.8/ 30.0	69.2/ 77.7	83.2/ 61.2	20.5/ 4.4	35.6/ 0.8
Elderly Female W. (0.0136)	913.2/ 902.4	245.7/ 260.2	0.0/ 0.0	41.4/ 35.6	62.3/ 67.6	84.7/ 60.8	15.0/ 4.0	37.0/ 0.8
Young Male N. (0.0090)	903.8/ 960.4	0.0/ 0.0	170.7/ 203.1	21.3/ 14.5	54.0/ 68.7	97.7/ 84.6	57.9/ 41.1	52.1/ 44.7
Young Female N. (0.0103)	899.9/ 987.7	0.0/ 0.0	170.5/ 149.2	36.0/ 21.8	59.9/ 75.2	79.7/ 76.9	31.5/ 39.9	49.2/ 40.0
Mid-Aged Male N. (0.0064)	1013./ 1054.	0.0/ 0.0	18.5/ 3.8	40.6/ 43.0	57.2/ 69.0	127.3/ 118.2	27.0/ 37.2	45.6/ 32.7

Demographic group (JSD)	Activity duration (minutes)							
	Home	Work	School	Buy goods	Meals	Recreation	Exercise	Visit
	(1)	(2)	(3)	(5)	(7)	(9)	(10)	(11)
Mid-Aged Female N. (0.0078)	994.0/ 1083.5	0.0/ 0.0	11.9/ 3.9	44.2/ 22.5	58.3/ 64.7	90.6/ 67.7	16.9/ 11.3	54.4/ 57.2
Elderly Male N. (0.0092)	1022/ 1101	0.0/ 0.0	1.6/ 3.3	42.1/ 43.8	70.3/ 80.5	144.4/ 150.7	25.0/ 5.8	41.6/ 22.4
Elderly Female N. (0.0099)	1037/ 1115	0.0/ 0.0	1.6/ 3.3	37.4/ 41.4	63.5/ 88.9	120.8/ 128.9	15.1/ 4.7	46.9/ 31.5

Note: Values shown as time use survey/ Output. W: Worker, N: Non-worker. JSD values indicate distribution similarity (lower JSD means better result).

Notably, our generated patterns and GPS data capture early morning activities (before 6 AM) that are absent in HTS data, addressing known survey limitations such as underreporting and rounding effects, for example, people typically fail to report some emergency travel activities and tend to make the pattern more routine [52]. This demonstrates our framework's ability to balance GPS and survey data characteristics while maintaining demographic representativeness.

Spatial validation

We utilized the generated output from the prior pipeline to assign locations in the context of LA County, with a spatial division of eight sub-regions (Figure 12(a)). The model was first calibrated using 100,000 agents from the SCAG Activity-Based Model dataset, then scaled to generate mobility patterns for 1 million agents.

To ensure the reliability of our validation results, we implemented a rigorous three-stage calibration process following the methodology established in [53]. First, we calibrated the MATSim-based transportation network using an iterative proportional adjustment algorithm that minimizes the sum of squared differences between simulated and observed traffic volumes across 100+ Caltrans PeMS detector stations on the top 10 highest-traffic freeways in LA County. The calibration algorithm iteratively adjusts link capacity factors starting from initial values of 0.8, continuing until convergence criteria are met (typically 6 iterations achieving < 5% relative difference across time periods). Second, we validated spatial trip distributions against SCAG ABM benchmarks, achieving cosine similarity of 0.997 for origin–destination matrices. Third, we conducted temporal validation using real-world traffic patterns, confirming realistic peak-hour dynamics and congestion onset/ dissipation patterns.

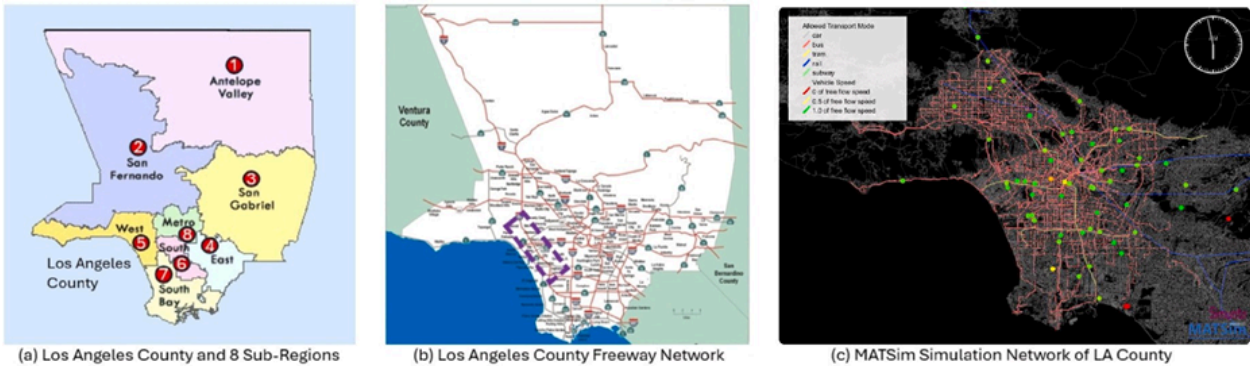

Figure 12: LA county map and freeway network

The spatial distribution of activities across sub-regions demonstrates the effectiveness of our location assignment method (Figure 13(a)). The high cosine similarity (0.997) between the origin–destination matrices of our model and SCAG ABM validates the preservation of regional flow patterns. Traffic simulation on the LA transportation network (Figure 12(b)) reveals strong system-level performance. The hourly VMT and traffic speed patterns closely match the benchmark SCAG ABM results, with MAPEs of 4.97 and 1.16 respectively (Figure 13(b)). These metrics indicate accurate reproduction of aggregate travel patterns.

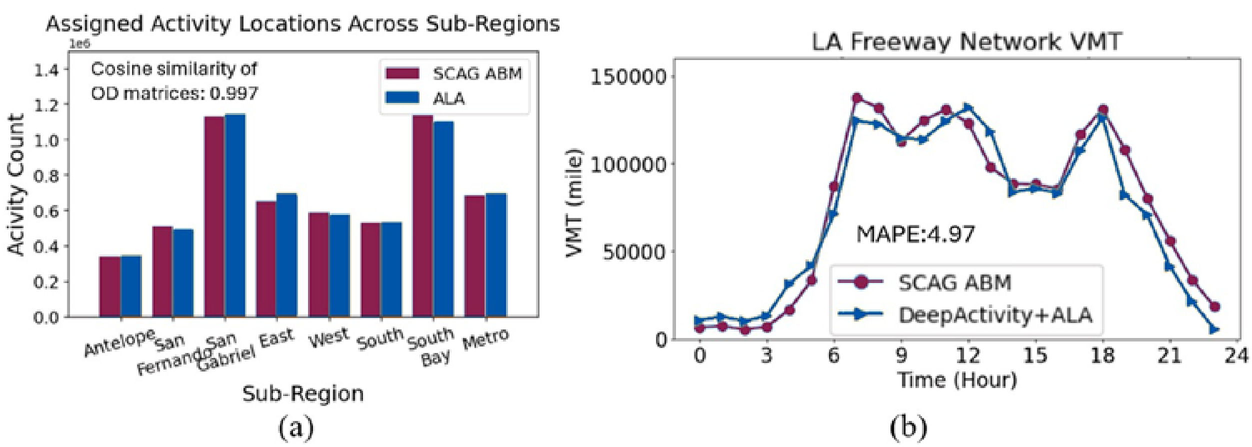

Figure 13: Validation for location assignment and traffic loading at network level

We conducted detailed analysis on a major segment of Interstate 405, comparing our results with both SCAG ABM predictions and PeMS real-world observations. The analysis reveals that directional variations in congestion patterns are well-captured. For northbound traffic, we observed a MAPE of 5.85% for volume and 4.45% for speed. Similarly, for southbound traffic, we measured a MAPE of 9.32% for volume and 4.36% for speed. The corridor-level results (Figure 14) demonstrate the model's capability to reproduce fine-grained traffic dynamics, including peak-hour congestion and directional flow patterns.

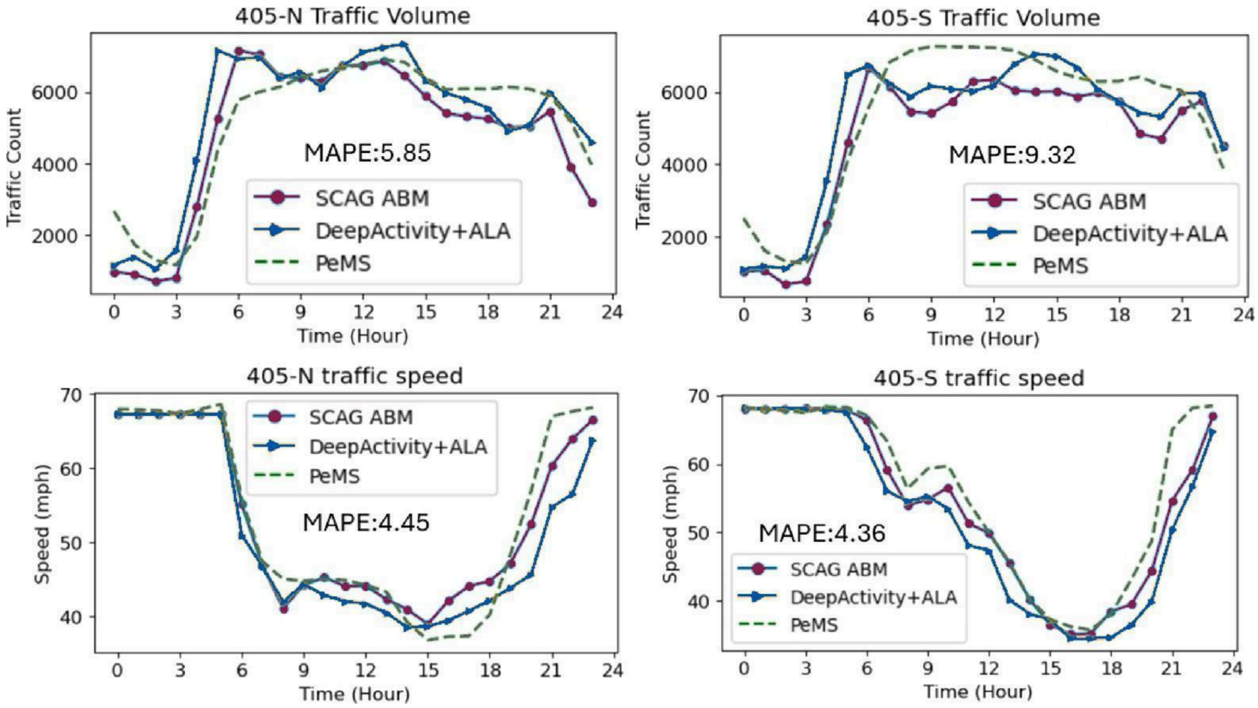

Figure 14: Validation for location assignment and traffic loading at corridor level

Conclusions and Future Work

This report addressed the persistent challenge of data fragmentation in human mobility modeling, a critical barrier to effective urban planning and transportation management. This challenge is central to AV and AI deployment, because safe and scalable automation depends on connecting vehicle level sensing with travel demand, roadway context, environmental conditions, and system operations. We developed a systematic data taxonomy as well as mobility data fusion functional stack to classify, interpret, and fuse existing data sources. The data taxonomy enables a methodical way to compare candidate datasets and decide which are fit for purpose. The functional stack supports integration and fusion of cross-domain datasets to generate high-resolution synthetic travel demand datasets.

Relative to conventional survey- and ABM-centered workflows, the framework's practical value is that it can generate regionally adapted, analysis-ready travel-demand outputs by reusing existing GPS traces, POI/context data, and survey prior. This reduces the need for new large-scale household travel surveys and manual semantic annotation. This was demonstrated in our case study of a Los Angeles County data-fusion workflow using semantic enrichment, synthetic travel demand generation, transfer learning, and traffic validation. This workflow produced synthetic travel-demand outputs from heterogeneous mobility-data inputs and demonstrated measurable improvements over defined baselines. For semantic annotation, the LLM-based POI classifier improved accuracy from 49.4% under the heuristic OSM-tag baseline to 91.4%. For domain adaptation, semi-supervised transfer learning reduced activity-type JSD from 0.0508 to 0.0028, while start-time and end-time JSDs reached 0.0008 and 0.0022. For spatial validation, the generated outputs achieved 0.997 cosine similarity against SCAG ABM OD matrices and corridor-level I-405 MAPE values of 4.36% for traffic flow against PeMS observations. Together, these components show how existing, heterogeneous data can be cost-effectively and accurately fused into analysis-ready inputs for mobility behavior modeling, planning, simulation, and AV-related applications.

While this research marks a significant advancement, several promising avenues for future work could extend its capabilities even further. The next logical step is to enhance the model to capture more complex social dynamics, such as coordinated activities between individuals, the formation of crowds at large events, and the influence of social networks on travel patterns. Additionally, future iterations could focus on integrating real-time data streams from traffic sensors or public transit systems to create dynamic, operational models for short-term forecasting and traffic management. For AV related applications, these travel-demand outputs are also useful for AV service providers because they can support market sizing, coverage-area design, fleet deployment strategy, and business-case evaluation before service launch. More future work could also integrate signal phase and timing data, work zone and incident feeds, weather and visibility information, curb activity, and connected vehicle telemetry so that deployment conditions can be modeled more explicitly. Finally, the framework could be extended to explicitly model transportation mode choice and to simulate the impacts of proposed infrastructure projects or policy interventions, such as dynamic road pricing, thereby providing a direct link between mobility modeling and tangible urban improvements. This extension would support evaluation of mixed-traffic interactions among AVs, human driven vehicles, transit, pedestrians, and cyclists, which is necessary for realistic pre-deployment testing and policy assessment.

References

[1] U.S. Census Bureau. (n.d.). American Community Survey (ACS) [Data set]. Retrieved September 1, 2025, from https://www.census.gov/programs-surveys/acs/

[2] U.S. Census Bureau, Center for Economic Studies. (n.d.). Longitudinal Employer–Household Dynamics (LEHD) program data [Data set]. Retrieved September 1, 2025, from https://lehd.ces.census.gov/data/

[3] U.S. Bureau of Labor Statistics. (n.d.). American Time Use Survey (ATUS) [Data set]. Retrieved September 1, 2025, from https://www.bls.gov/tus/

[4] Federal Highway Administration. (n.d.). National Household Travel Survey (NHTS) [Data set]. Retrieved September 1, 2025, from https://nhts.ornl.gov/

[5] California Department of Transportation. (n.d.). 2010–2012 California Household Travel Survey (CHTS) [Data set]. Transportation Secure Data Center (NREL). Retrieved September 1, 2025, from https://www.nrel.gov/transportation/secure-transportation-data/tsdc-california-travel-survey

[6] OpenStreetMap contributors. (n.d.). Planet dump (planet.osm) [Data set]. Retrieved September 1, 2025, from https://planet.openstreetmap.org

[7] Foursquare Labs, Inc. (n.d.). Foursquare Places API and data [Data set]. Retrieved September 1, 2025, from https://docs.foursquare.com/developer/

[8] SafeGraph, Inc. (n.d.). Geometry (building footprint polygons) [Data set]. Retrieved September 1, 2025, from https://www.safegraph.com/free-data/polygon-data

[9] U.S. Department of Transportation, Bureau of Transportation Statistics. (2025). National Transit Map (NTM): Routes [Data set]. Retrieved September 1, 2025, from https://geodata.bts.gov/datasets/usdot::national-transit-map-routes/about

[10] Federal Highway Administration. (2025). National Bridge Inventory (NBI) [Data set]. Retrieved September 1, 2025, from https://www.fhwa.dot.gov/bridge/nbi/ascii.cfm

[11] Eviction Lab. (2023). National Zoning and Land Use Database (NZLUD) [Data set]. Retrieved September 1, 2025, from https://evictionlab.org/national-zoning-and-land-use-database/

[12] District Department of Transportation. (2025). Traffic Signal [Data set]. Retrieved September 1, 2025, from https://catalog.data.gov/dataset/traffic-signal-a46dd

[13] City of Austin, Austin Transportation Department. (2024). Traffic Signal Network Device Status Log [Data set]. Retrieved September 1, 2025, from https://catalog.data.gov/dataset/traffic-signal-network-device-status-log

[14] Federal Highway Administration. (2024). Model Inventory of Roadway Elements (MIRE) Version 2.1 [Data dictionary]. Retrieved September 1, 2025, from https://highways.dot.gov/sites/fhwa.dot.gov/files/2024-08/MIRE_2.1_FINAL_508v3.pdf

[15] California Department of Transportation. (n.d.). Performance Measurement System (PeMS) [Data set]. Retrieved September 1, 2025, from https://dot.ca.gov/programs/traffic-operations/mpr/pems-source

[16] Federal Highway Administration. (n.d.). Highway Safety Information System (HSIS) [Data set]. Retrieved September 1, 2025, from https://highways.dot.gov/research/safety/hsis

[17] Federal Highway Administration. (n.d.). National Performance Management Research Data Set (NPMRDS) [Data set]. Retrieved September 1, 2025, from https://data.transportationops.org/national-performance-management-research-data-set-npmrds

[18] U.S. Department of Transportation. (n.d.). Work Zone Data Exchange (WZDx) [Data spec and data feeds]. Retrieved September 1, 2025, from https://www.transportation.gov/av/data/wzdx

[19] CATT Laboratory, University of Maryland. (n.d.). Regional Integrated Transportation Information System (RITIS) [Data platform]. Retrieved September 1, 2025, from https://www.cattlab.umd.edu/ritis/

[20] Veraset. (n.d.). Movement dataset [Data set]. Retrieved September 1, 2025, from https://www.veraset.com/datasets/movement

[21] Cuebiq. (n.d.). Data for Good mobility data [Data set]. Retrieved September 1, 2025, from https://cuebiq.com/social-impact/

[22] Spectus. (n.d.). Spectus data clean room for mobility analysis [Data platform]. Retrieved September 1, 2025, from https://docs.spectus.ai/

[23] StreetLight Data, Inc. (n.d.). StreetLight InSight transportation analytics [Data product]. Retrieved September 1, 2025, from https://www.streetlightdata.com/

[24] New York City Taxi & Limousine Commission. (n.d.). TLC trip record data [Data set]. Retrieved September 1, 2025, from https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

[25] INRIX. (2024). Global Traffic Scorecard [Data set]. Retrieved September 1, 2025, from https://inrix.com/scorecard/

[26] Waze. (n.d.). Waze for Cities Data (formerly Connected Citizens Program) [Data feed]. Retrieved September 1, 2025, from https://www.waze.com/wazeforcities

[27] HERE Global B.V. (n.d.). HERE Location Services [Data/platform]. Retrieved September 1, 2025, from https://www.here.com/platform

[28] TomTom. (2025). TomTom Traffic Index [Data set]. Retrieved September 1, 2025, from https://www.tomtom.com/traffic-index/

[29] Waymo LLC. (n.d.). Waymo Open Dataset [Data set]. Retrieved September 1, 2025, from https://waymo.com/open/

[30] Motional. (2021). nuScenes [Data set]. Retrieved September 1, 2025, from https://www.nuscenes.org/

[31] Argo AI. (2021). Argoverse 2 [Data set]. Retrieved September 1, 2025, from https://www.argoverse.org/

[32] Lyft. (2019). Lyft Level 5 Dataset [Data set]. Retrieved September 1, 2025, from https://level5.lyft.com/dataset/

[33] comma.ai. (2018). comma2k19 [Data set]. Retrieved September 1, 2025, from https://github.com/commaai/comma2k19

[34] Virginia Tech Transportation Institute. (2006). 100-Car Naturalistic Driving Study [Data set]. Retrieved September 1, 2025, from https://rosap.ntl.bts.gov/view/dot/37370

[35] MIT Advanced Vehicle Technology Consortium. (n.d.). MIT-AVT naturalistic driving datasets [Data set]. Retrieved September 1, 2025, from https://avt.mit.edu/

[36] Strategic Highway Research Program 2. (n.d.). SHRP 2 Naturalistic Driving Study (NDS) [Data set]. Retrieved September 1, 2025, from https://insight.shrp2nds.us/

[37] M. C. González, C. A. Hidalgo, A. L. Barabási, Understanding individual human mobility patterns, Nature 453 (2008) 779–782.

[38] X. Liao, Q. Jiang, Y. Liu, B. Y. He, C. Kuai, J. Ma, Deep activity model: A generative approach for human mobility pattern synthesis, arXiv preprint arXiv:2405.17468 (2024).

[39] Liu, Y., Kuai, C., Liao, X., Ma, H., He, B. Y., & Ma, J. (2024, September). Semantic trajectory data mining with llm-informed poi classification. In 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC) (pp. 207-213). IEEE.

[40] Ma, H., Liao, X., Liu, Y., Jiang, Q., Stanford, C., Cao, S., & Ma, J. (2025). Learning universal human mobility patterns with a foundation model for cross-domain data fusion. *Transportation Research Part C: Emerging Technologies*, *180*, 105311.

[41] C. M. Schneider, V. Belik, T. Couronné, Z. Smoreda, M. C. González, Unravelling daily human mobility motifs, Journal of The Royal Society Interface 10 (2013) 20130246.

[42] J. Cao, Q. Li, W. Tu, F. Wang, Characterizing preferred motif choices and distance impacts, Plos one 14 (2019) e0215242.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[44] M. Müller, Dynamic time warping, Information retrieval for music and motion (2007) 69–84.

[45] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks?, Advances in neural information processing systems 27 (2014).

[46] M. E. Peters, S. Ruder, N. A. Smith, To tune or not to tune? adapting pretrained representations to diverse tasks, arXiv preprint arXiv:1903.05987 (2019).

[47] US Bureau of Labor Statistics, County employment and wages in California — fourth quarter 2023, 2023. Accessed: 2025-01-26.

[48] Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., Gal, Y., 2024. AI models collapse when trained on recursively generated data. Nature 631 (8022), 755–759.

[49] U.S. Census Bureau, 2026. American time use survey. U.S. Census Bureau. Retrieved April 19, 2026, from https://www.census.gov/programs-surveys/atus.html

[50] Beegle, K., 2022. Collecting good time use data with less cost and complexity. World Bank Blogs. Retrieved April 19, 2026, from https://blogs.worldbank.org/en/impactevaluations/collecting-good-time-use-data-less-cost-and-complexity

[51] Federal Highway Administration, 2026. National household travel survey. Federal Highway Administration. Retrieved April 19, 2026, from https://www.fhwa.dot.gov/policyinformation/nhts.cfm

[52] Bricka, S., Paleti, R., Bhat, C.R., 2012. An analysis of the factors influencing differences in survey-reported and GPS-recorded trips. Transp. Res. Part C: Emerg. Technol. 21, 67–78.

[53] He, B.Y., Jiang, Q., Ma, H., Ma, J., 2024. Multi-agent multimodal transportation simulation for mega-cities: Application of los angeles. In: The 13th International Workshop on Agent-Based Mobility, Traffic and Transportation Models, Methodologies and Applications. ABMTRANS.

420 Westwood Plaza, 7702 Boelter Hall
Los Angeles, CA 90095-1600

info@mobilitycoe.org
mobilitycoe.org

CENTER OF EXCELLENCE
on New Mobility and Automated Vehicles