A Large-scale Dataset and Benchmark for Commuting Origin-Destination Flow Generation

Anonymous authors
Paper under double-blind review

Abstract

Commuting Origin-Destination (OD) flows are critical inputs for urban planning and transportation, providing crucial information about the population residing in one region and working in another within an interested area. Due to the high cost of data collection, researchers have developed physical and computational models to generate commuting OD flows using readily available urban attributes, such as sociodemographics and points of interest, for cities lacking historical OD flows —commuting OD flow generation. Existing works developed models based on different techniques and achieved improvement on different datasets with different evaluation metrics, which hinderes establishing a unified standard for comparing model performance. To bridge this gap, we introduce a large-scale dataset containing commuting OD flows for 3,333 areas including a wide range of urban environments around the United States. Based on that, we benchmark widely used models for commuting OD flow generation. We surprisingly find that the network-based generative models achieve the optimal performance in terms of both precision and generalization ability, which may inspire new research directions of graph generative modeling in this field. The dataset and benchmark are available at `https://anonymous.4open.science/r/CommutingODGen-Dataset-0D4C/`.

1 Introduction

Commuting refers to the daily round-trip movement of individuals from their homes to their workplaces, which is an important topic in fields like urban planning, transportation, environmental science, and economics (Batty, 2007; Gonzalez et al., 2008; Iqbal et al., 2014; Liu et al., 2020). These movement between all pair of origins and destinations within the interested area can be effectively recorded as Origin-Destination (OD) flows. All OD flows across the entire area named the commuting OD matrix, where each element represents the number of people reside in one region and work in another. The commuting OD matrix can be naturally modeled as a directed weighted graph, i.e, commuting OD network, where nodes represent regions and edges represent the commuting OD flows between regions (Saberi et al., 2017; 2018). Understanding commuting OD flows at both the pair-wise and network-level allows urban planners to analyze the structured mobility patterns, optimize the transportation system, and make informed decisions on urban development (Zeng et al., 2022; Imai et al., 2021; Zhong et al., 2014). However, collecting the data often costs a lot and raises privacy concerns. Thus, researchers have developed both classic physical models (Zipf, 1946; Simini et al., 2012) and more recent, promising data-driven approaches (Pourebrahim et al., 2019; Liu et al., 2020; Simini et al., 2021; Rong et al., 2023c;b;d) to model commuting OD flows and generate data for areas lacking historical flows. This task is named as commuting OD flow generation.

There are two main challenges lying on two aspects: the lack of a comprehensive dataset and the absence of a unified and systematic evaluation. In details, existing works can be categorized in three types: physical models, classic machine learning models, and graph neural network models. Physical models campare the OD flow to physical phenomenon, such as the gravity model (Zipf, 1946; Barbosa et al., 2018) and radiation model (Simini et al., 2012). The physical models utilize simple mathematical equations to capture the pair-wise relationships between origins and destinations, which have a strong theoretical basis but are limited by the underfitting of the complex human mobility. Recent popular data-driven models (Rodriguez-Rueda et al., 2021; Pourebrahim et al., 2019; 2018; Robinson & Dilkina, 2018; Simini et al., 2021; Liu et al., 2020; Rong et al., 2023c) can capture the complex relationships between urban attributes and commuting OD flows with sophisticated models. These works based on machine learning or deep learning techniques learning from only one single or a few areas, have shown poor generalizablility to distinct urban environments. Despite the significant practical value of commuting OD flow generation, it has not gained widespread attention from the deep learning community. One key reason is the lack of a unified benchmark based on a comprehensive dataset. Currently, studies use their own datasets from individual city scenarios for evaluation, making it difficult to compare and communicate insights between different model designs.

To address the above issue, we collect data from multiple sources and construct a **large**-scale dataset containing **commuting OD** matrices for 3,333 diverse areas around the whole United States (**LargeCommuingOD**). Thanks to the extensive spatial scale of the dataset, various urban environments are covered, including metropolitan areas, small cities, towns, and rural areas. For supporting better study of modeling, each area in the dataset has not only the commuting OD matrix but also regional sociodemographics and numbers of point-of-interests (POIs) within different categories for all regions in the area. Specifically, each area is profiled with its boundary and the boundaries of regions within it, which are represented as polygons with detailed geographic coordinates, i.e., latitude and longitude. The sociodemographics include the population of different genders and age groups, the number of househoulds, and income levels, etc. The point-of-interests are categorized into various types, such as restaurants, education, and shopping, etc. This dataset can be used to comprehensively study and evaluate the models for commuting OD flow generation.

Based on our dataset, we benchmark the existing widely used models for commuting OD flow generation in a common framework. We utilize randomly selected areas in the dataset as the test set, which covers diverse urban environments, to comprehensively evaluate the models in terms of both precision and generalizablility. The remaining areas are leveraged to train the models. Existing works including physical models, classic machine learning models, and graph neural network models are all benchmarked. Besides, the generative models trained on the large-scale dataset emerge powerful performance, which has been demonstrated not only in fields like natural language processing (Brown et al., 2020; Kaplan et al., 2020) and computer vision (Peebles & Xie, 2023) but also in spatial-temporal data modeling (Yuan et al., 2024; Jin et al., 2023). We introduce a preliminary adaptation of the graph diffusion model to **W**eighted **E**dges **D**iffusion condition on **A**ttributed **N**odes (WEDAN) into our benchmark. We surprisingly find that the network-based generative models perform the best in terms of both precision and generalization ability, which may call for a new paradigm of graph generative modeling in this field.

In summary, the contributions of this work are as follows:

- We construct a large-scale dataset (LargeCommuingOD) containing commuting OD flows for 3,333 diverse areas around the United States covering 9,372,610 km^2 including a wide range of urban environments. Each area also includes sociodemographics and point-of-interests totaly 131 features as urban attributes for regions within it.
- Based on the LargeCommuingOD, we benchmark the existing widely used models for commuting OD flow generation. With dataset containing distinct areas, we can comprehensively evaluate the models in terms of precision and generalizablility.
- We find that network-based modeling for commuting OD flow supported by our dataset gives a promising performance, which treats an area and the commuting OD flow within it as a network. Training on a large number of commuting OD networks, generative models can capture the universal and distinct mobility patterns at the city level, leading to better generalizablility.

2 Preliminaries

In this section, we introduce the definitions and problem formulation of commuting OD flow modeling, followed by the existing works of this field.

2.1 Definitions and Problem Formulation

Definition 1. Regions. We divide the interested area into non-overlapping regions, represented as $\mathcal{R} = \{r_i | i = 1, 2, ..., N\}$, with N being the total count of the regions. Each region fulfills unique functions, indicated by their urban attributes $\mathbf{X}_r$, which include sociodemographics and the distribution of points-of-interests in different categories.

Definition 2. Spatial Characteristics. The spaital characteristics of an area $\mathcal{C}_{\mathcal{R}}$ are composed of urban attributes of each region $\{\mathbf{X}_{r_i} | r_i \in \mathcal{R}\}$ and the interactions, such as distance, between all regions $\{d_{ij} | r_i \text{ and } r_j \in \mathcal{R}\}$.

Table 1: Comparison of the proposed dataset and other dataset utilized in existing works.

Dataset	#Area	Area Type	Cover Area (km^2)	Metropolitan	Town	Rural	Curated & Public
Karimi et al. (2020)	1	Central District	-	✓	✗	✗	✗
Pourebrahim et al. (2018; 2019)	1	Whole City	789	✓	✗	✗	✗
Liu et al. (2020)	1	Whole City	789	✓	✗	✗	✓
Yao et al. (2020)	1	Central District	900	✓	✗	✗	✗
Lenormand et al. (2015)	2	Whole City	15,755	✓	✗	✗	✗
Rong et al. (2023c;d;b)	8	Whole City	25,954	✓	✗	✗	✓
Simini et al. (2021)	2,911	National Gridding Coverage	686,983	✓	✓	✓	✗
Ours	3,333	Census Area Coverage	9,372,610	✓	✓	✓	✓

Definition 3. Commuting OD Flow. The term commuting OD flow refers to the population $\mathcal{F}_{r_{org},r_{dst}}$, residing in r_{org} and working at r_{dst}.

Definition 4. Commuting OD Matrix. Denoted by $\mathbf{F} \in \mathbb{R}^{N \times N}$, the commuting OD matrix includes commutings among all regions within the area. $F_{i,j}$ means the commuting from r_i to r_j.

Problem 1. ***Commuting OD Flow Modeling.*** *The problem aims to learn a model, given any area's spatial characteristics $\mathcal{C}_{\mathcal{R}}$, generating their corresponding commuting OD matrices* $\mathbf{F}$ *that closely resemble those in the real world without any historical information.*

2.2 Existing Works on Commuting OD Flow Modeling

Limitations of Dataset Used in Existing Works. As shown in Table 1, existing datasets used in commuting OD flow modeling have several major limitations. *First*, existing datasets utilized in the literature have a **limited spatial scale**, usually focusing on a single or few large cities, leading two very limited spatial coverage. For example, Karimi et al. (2020) and Yao et al. (2020) only consider a central district in a city, and Pourebrahim et al. (2018; 2019), Liu et al. (2020), Lenormand et al. (2015), and Rong et al. (2023c;d) only consider less than 8 large metropolitans, whose areas are less than 30,000 km^2. Although Simini et al. (2021) consider a national gridding coverage in the United Kingdom and Italy, the area is still limited to 686,983 km^2. Besides, they do not provide the curated dataset for public use, which cannot be used for further research. In contrast, our dataset covers 3,333 areas around the United States, a total area of 9,372,610 km^2, providing a much broader spatial scale. And our dataset is curated and publicly available, which can be found at https://anonymous.4open.science/r/CommutingODGen-Dataset-0D4C/. *Second,* with the limited spatial scale, existing datasets usually focus on **a single type of urban environments**, such as metropolitan areas, central districts, or whole cities, which cannot include a massive areas with high diversity in terms of size and structure. Models trained on such datasets may not be generalized to other areas with different characteristics, limiting their applicability on only similar areas. Our dataset covers metropolitan areas, towns, and rural areas around the United States, providing a more comprehensive dataset for training and evaluating models. With the diversity of areas, models trained on our dataset can be more generalizable.

OD Flow Modeling Approaches. Existing works can be categorized into three types. The **first** is *physical models*, such as the gravity model (Zipf, 1946) and the radiation model (Simini et al., 2012), which mimick the commuting OD flows as physical pheonomena and utilize simple mathematical equations to model the flows. Physicists dive into the mechanisms of individual mobility decisions and try to explain the phenomenon of commuting OD flows. The **second** is *statistical learning models*, such as tree-based models (Robinson & Dilkina, 2018; Pourebrahim et al., 2018; 2019), SVR (Rodriguez-Rueda et al., 2021), artificial neural networks (ANNs) (Sana et al., 2018; Lenormand et al., 2016; Simini et al., 2021), which predict the OD flows between pairs of regions in data-driven schemes. The **third** is *graph learning models*. Liu et al. (2020); Cai et al. (2022) utilized GATs to aggregate the neighbors' information to profile the regions better and improve the prediction accuracy. Yao et al. (2020) model the local spatial adjacenct structure of regions with graph convolutional networks and imputate the missing OD flows in a semi-supervised manner. Rong et al. (2023d;b) introduce adversarial and denoising diffusion generative methods with graph transformers to model the commuting OD matrix generation as graph generation problem. Many researchers from urban planning and transportation have shown interest in data-driven models because of the better performance Barbosa et al. (2018); Luca et al. (2021); Rong et al. (2023a). But there lacks a large-scale dataset containing a wide range of urban environments and unified benchmark for comparing the performance of different models, which hinders the development of more powerful models. Our dataset and benchmark can fill this gap and provide a common ground for evaluating the models.

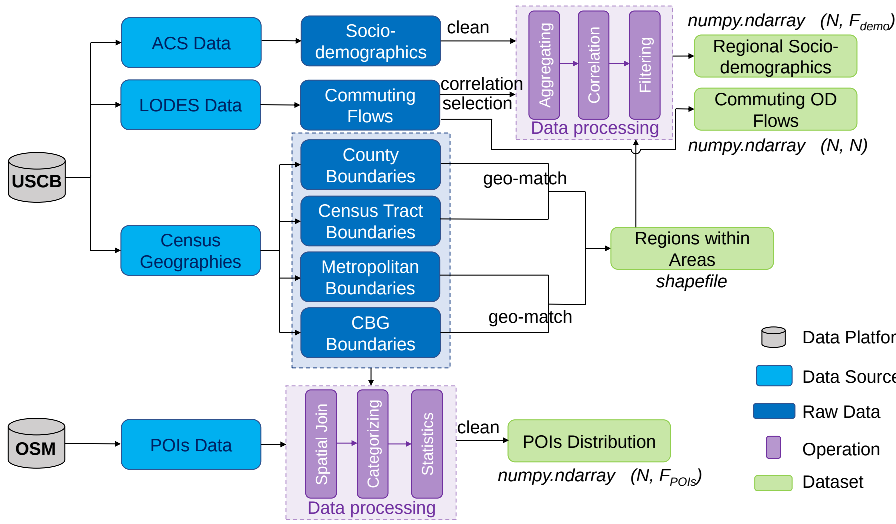


Figure 1: Discreption of the pipeline constructing our datasest.

3 LargeCommuingOD: A Large-scale Commuting OD Flow Dataset

3.1 Data Collection and Curation

The pipeline for constructing our dataset is shown in Figure 1. As shown in the figure, the dataset contains four main componets: 1) boundaries of areas and regions 2) sociodemographics, 3) POIs distributions, 4) commuting OD flows. First of all, the boundaries of areas and regions are download from the U.S. Census Bureau, which include all counties, metropolitans, census tracts, and census block groups (CBGs). And we set the counties as the areas and census tracts as the regions for the county areas, and set the metropolitans as the areas and CBGs as the regions for the metropolitan areas. The counties can be related to the census tracts by code of Federal Information Processing Standards (FIPS). The CBGs belong to the metropolitans, which is detected by the spatial relationship between the boundaries of CBGs and metropolitans, i.e., whether the CBG is inside the metropolitan. Then, the sociodemographics for each region can be accessed from the American Community Survey (ACS) on the website of the U.S. Census Bureau. For each indicators, we use regression analysis on the indicator and flow intensity to decide whether to choose the indicator into the urban attributes. The information not related to human mobility is excluded. And for each region, we use API of OpenStreetMap to get the number POIs in different categories. The POIs are divided into 36 categories, including restaurants, schools, hospitals, etc. The commuting OD flow is provided by the 2018 Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics (LODES) dataset on the website of the U.S. Census Bureau. The data is orgagized in form of tables. Each table contains the commuting information of one state. Each row in the table represents the commuting flow between two specific census blocks. We aggregate the flow into census tract level and construct the OD matrix.

3.2 Data Description

We have collected data from a total count of 3,333 areas around the United States. There are two kind of spatial divisions in LargeCommuingOD: 1) 3,233 counties as the areas and census tracts inside each county as the regions, 2) 100 metropolitans, where the population is more than 1 million, as the areas and census block groups CBGs inside each metropolitan as the regions. LargeCommuingOD includes the following information: 1) regional urban attributes, including sociodemographics and POIs, 2) commuting OD flows, represented by OD matrices, which are aggregated commuting flows within areas. The counties are defined by the U.S. Census Bureau. Each county is a local government unit in the United States, and the counties should cover a similar number of households and population. The metropolitan boundaries are obtained from Topologically Integrated Geographic Encoding and Referencing (TIGER) dataset. The metropolitan areas exclude the rural areas, which do not have population and urban functionalities.

Regional Urban Attributes. Each region is characterized by sociodemographics and urban functionalities, derived from American Community Survey (ACS) (U.S. Census Bureau, 2012) by the

Table 2: Summary of urban attributes used to profile a region.

Categories	Centents	#Features
Sociodemographics	Total population	1
	Population with different genders and ages	56
	Median age of people with different genders	3
	Median earnings	1
	Ratio of different classes of jobs	5
	Vehicle ownership	4
	The number of households with different types	4
	Population with different education levels	21
	Poverty with different genders	2
POIs	The number of POIs in different kind.	34
Total		131

U.S. Census Bureau and the distribution of POIs from OpenStreetMap (OpenStreetMap contributors, 2017), as shown in Table 2. Demographics include the population structure of a region based on age, gender, income, education, and other factors, encompassing a total of 97 dimensions. POIs are divided into 36 different categories. The distances between regions are calculated using the planar Euclidean distance between their centroids.

OD matrices. We construct commuting OD matrices for all areas using data on commuting patterns from the 2018 Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics (LODES) dataset. These matrices represent aggregated commuting flows within areas. Each entry in an OD matrix denotes the count of individuals residing in one region and working in another, effectively mapping commuting patterns of workers across different regions. The LODES dataset is widely used in existing works (Liu et al., 2020; Pourebrahim et al., 2019; 2018). In LargeCommuingOD, the commuting information is aggregated by the cooperation and other kind of work units, which is more reliable and accurate than the individual commuting data. Therefore, in the data collection process, information has been ensured to be representative at a national scale, thus eliminating sampling errors. The raw data provided is at the census block level, which is then aggregated to the census tract level for the county areas and to the CBG level for the metropolitan areas.

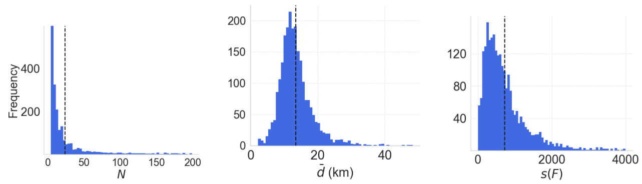
It is worth noting that the commuting OD flows within the 3,233 counties cannot carry the mobility across different counties, while the flows within metropolitans can. So LargeCommuingOD include both intra-county and inter-county flows.

3.3 Data Statistics

We provide a statistical analysis of LargeCommuingOD to illustrate the diversity of the dataset. We analysis the dataset from two perspectives: area characteristics and mobility patterns. From Figure 2, we can see that the number of regions in each area varies significantly, which shows the heterogeneity of the areas in LargeCommuingOD. Furthermore, cases in Figure 3 reveal the diverse structure of the areas, including monocentric, polycentric, and evenly distributed spreading. For analyzing the mobility patterns, we measure the average trip distances and the variance of the regional mobility intensity. The travel distances tend to be shorter but there are still long-distance trips, make the mobility patterns complex. The variance of the regional mobility intensity is also diverse in a wide range, which indicates the heterogeneity of the mobility patterns. For commonalities among areas, we analyze the distribution of OD flows and outflows in areas of different scales, as shown in Figure 4. We can observe that the heterogeneity exsits between different scales of areas. Yet, the commonalities also exist, i.e., the scaling behaviors are the same among areas. This demonstrates that LargeCommuingOD is a comprehensive dataset that covers a wide range of urban environments with diverse mobility patterns. To further intuitively understand the dataset, we provide the Visualization of the OD flows via heatmaps in Appendix A.1.

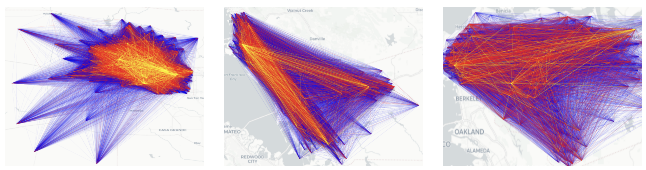
3.4 Discussion

Superiority. From the statistical analysis, we can see that LargeCommuingOD is large-scale and comprehensive, covering a wide range of areas of different scales and mobility patterns, i.e., diverse urban environments. For **learning**, the sufficient scenarios in LargeCommuingOD can support the modeling research to capture the distinctness and commonalities of the mobility patterns in different areas, as shown in Figure 5. For **evaluation**, the diverse urban environments in LargeCommuingOD



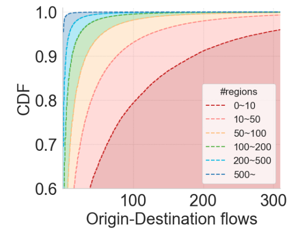
(a) #Regions (b) Distances. (c) Variance.

Figure 2: Statistical analysis of LargeCommuingOD, including the distribution of a) the number of regions in each area, b) the average trip distance in each area, c) the variance of the in/out flow of each region in each area.

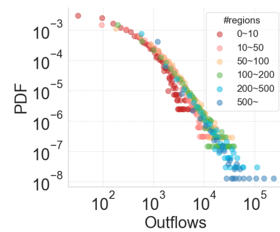


(a) Maricopa (b) Alameda (c) Contra Costa

Figure 3: Visualization of the OD matrices of three areas with different mobility structure, a) monocentric (Maricopa in Arizona), b) polycentric (Alameda in California), and c) smoothly distributed (Contra Costa in California).



(a) Edge weights.



(b) Node degrees.

Figure 4: Distributions of OD flows and outflows in areas of different scales. a) cumulative distribution function of edge weights, and b) probabilistic density function at log scale of node degrees.

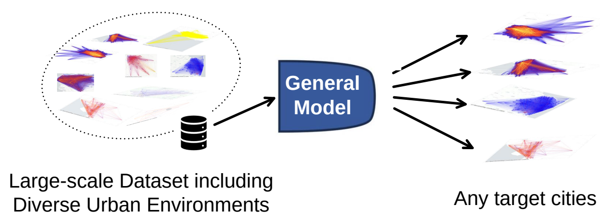


Figure 5: Superiority of LargeCommuingOD. The models trained on LargeCommuingOD can capture the distinctness and commonalities of the mobility patterns in different areas.

can support the comprehensive evaluation of the models in terms of both precision and generalizability, which cannot be achieved by existing datasets.

Limitations. Despite the comprehensiveness of LargeCommuingOD, there are several limitations. First, the data is collected from a single year, which may not fully capture the temporal changes of commuting patterns. Second, the data is limited to the U.S., which may not be generalizable to other countries with different characteristics and cultures. Third, the data only includes commuting OD flows, which may not fully represent the human mobility patterns in urban areas. This may limit the utility of our dataset in more fine-grained mobility analysis tasks.

4 Benchmark

In this section, we investigate the **precision** and **generalizability** of benchmark models by answering the following questions:

- For **precision**, how realistic are the models in generating the commuting OD flows in terms of the flow values and network properties? (Section 4.2)
- For **generalizability**, can the models capture the distinctness and commonalities of the mobility pattern across different urban areas? (Section 4.3 and 4.4)

We further evaluated and discussed the models from three perspectives: interpretability, robustness, and fairness. For details, please refer to Appendix B.1, B.2, and B.3.

Table 3: A comprehensive comparison of the benchmark models.

Paper	Model	Type	Perspective	Output
Barbosa et al. (2018)	GM	Physical Model	Pair-wise	OD flow
Rodriguez-Rueda et al. (2021)	SVR	Kernel-based Model	Pair-wise	OD flow
Pourebrahim et al. (2019)	RF	Tree-based Model	Pair-wise	OD flow
Robinson & Dilkina (2018)	GBRT	Tree-based Model	Pair-wise	OD flow
Simini et al. (2021)	DGM	MLP-based Model	Pair-wise for one origin	Outflows to all destinations
Luo et al. (2024)	TransFlower	Transformer -based Model	Pair-wise for one origin	Outflows to all destinations
Liu et al. (2020)	GMEL	GNN-based Model	Pair-wise	OD flow
Rong et al. (2023d)	NetGAN	GAN-based Model	Network-wise	OD matrix
(Rong et al., 2023b)	DiffODGen	Diffusion-based Model	Network-wise	OD matrix
-	WEDAN	Diffusion-based Model	Network-wise	OD matrix

4.1 Experimental Setup

Benchmark Models. We utilize our propose dataset to benchmark 9 existing models. The existing models are in three categories: physical models, classical statistical learning approaches, and graph learning models. Besides, we also explore the feasibility of utilizing the graph generative modeing, which construct the fourth category (generative models).

Within the first category are two physical models:

- **Gravity Model with Power-law Decay (GM-P)** (Zipf, 1946; Barbosa et al., 2018) is inspired by the gravitation in physics, positing that the OD flow is directly proportional to the populations of the origin and the destination, and inversely proportional to the distance between them. The power-law decay is used to model the distance decay effect.
- **Gravity Model with Exponential Decay (GM-E)** (Zipf, 1946; Barbosa et al., 2018) is almost identical to GM-P, except that it uses an exponential decay function to model the distance decay effect.

The second category encompasses classical statistical learning approaches tailored for OD flow modeling:

- **Support Vector Regression (SVR)** (Rodriguez-Rueda et al., 2021) is a kernel-based machine learning algorithm that has been widely used in regression tasks. It is employed to predict the OD flow between two regions based on the urban attributes of the regions by Rodriguez-Rueda et al. (2021).
- **Random Forest (RF)** (Pourebrahim et al., 2019) stands out as a tree-based machine learning algorithm known for its robustness, demonstrating commendable results in generating OD flows.
- **Gradient Boosting Regression Tree (GBRT)** (Robinson & Dilkina, 2018) use boosting techniques to enhance the performance of decision trees, which has been applied to predict the OD flow in the city.
- **Deep Gravity Model (DGM)** (Simini et al., 2021) use multi-layer perceptrons (MLPs) inspired by gravity models to calculate flows to different destinations by estimating the distribution probabilities. We have adapted this model to generate OD flow volumes directly.
- **TransFlower** (Luo et al., 2024) is a transformer-based model under the framework of DGM, which utilizes the transformer to model the spatial dependency of all destinations for each origin rather than MLPs. The model is also adapted to generate OD flows directly.

The third category includes approaches based on graph neural networks, which model the urban space or commuting OD networks as graphs:

- **Geo-contextual Multitask Embedding Learning (GMEL)** (Liu et al., 2020) leverages graph neural networks (GNNs) to aggregate neighboring information for each region. This process enhances the spatial characteristic representation of the regions in a city, which contributes to the refinement of regional embeddings and augments precision.
- **NetGAN** (Bojchevski et al., 2018) is a GAN-style framework that recreates realistic network architectures by generating random walks that mirror the distribution of walks extracted from real networks. We have tailored it to construct directed and weighted graphs, i.e., OD matrices.

The last but not least category includes the generative models based on transformer-backbone models:

- **DiffODGen** (Rong et al., 2023b) employs a cascaded diffusion model specifically for large cities, leveraging the sparsity of the mobility network to separately model the topology of the graph and the weights given edges, achieved SOTA results in large cities.
- **WEDAN** is a preliminary try to adapt graph diffusion models to model the joint distribution of all elements in OD matrices conditioned on urban attributes, which named WEDAN (**W**eighted **E**dges **D**iffusion condition on **A**ttributed **N**odes).. We use this model to explore the new paradigm for commuting OD flow generation. The details of the model are introduced in Appendix C.

Parameter Settings The graph transformer in diffusion models employs 4 layers with each having 32 hidden dimensions. We utilize 250 diffusion steps in diffusion models, following a cosine noise scheduler as suggested by Nichol & Dhariwal (2021). Denoising networks are optimized using AdamW optimizer (Loshchilov & Hutter, 2017), with a learning rate set at 1e-3. Our method and DiffODGen both sample 50 times during generation and take the average as final generated results.

For the gravity model, we adopt the approach outlined by Barbosa et al. (2018), which involves four fitting parameters. In the random forest algorithm, the number of estimators is set to 100. The DGM (Simini et al., 2021) is stacked by 10 layers with 64 hidden dimensions in each layer, while GNN-based models are designed with 3 layers and 64 channels all. TransFlower is stacked by 3 transformer encoder with 8 heads and 64 hidden dimensions in each head. The hyper-parameters for the denoising networks in two cascaded diffusion models of DiffODGen are aligned with our methodology.

All the selection of hyper-parameters is based on the validation set and trade off between the performance and computational resources.

Evaluation Metrics. We uniformly evaluate the performance based on widely adopted metrics from two perspectives: the error between the generated OD matrices and the corresponding real ones, and the distribution deviation in graph properties between the generation and the real data. The error metrics include Root Mean Square Error (RMSE), Normalized Root Mean Square Error (NRMSE) and Common Part of Commuting (CPC), while the distribution difference metrics include Jensen-Shannon Divergence (JSD) for inflow, outflow, and OD flow. These metrics are calculated for each area and then averaged across all. The calculation formulas are shown as follows.

$$RMSE = \sqrt{\frac{1}{|\mathbf{F}|} \sum_{r_i, r_j \in \mathcal{R}} ||\mathbf{F}_{ij} - \hat{\mathbf{F}}_{ij}||_2^2}, \tag{1}$$

$$NRMSE = \frac{RMSE}{\sqrt{\frac{1}{N^2} \sum_{r_i, r_j \in \mathcal{R}} ||F_{ij} - \bar{F}_{ij}||_2^2}}, \tag{2}$$

$$CPC = \frac{2 \sum_{r_i, r_j \in \mathcal{R}} \min(\mathbf{F}_{ij}, \hat{\mathbf{F}}_{ij})}{\sum_{r_i, r_j \in \mathcal{R}} \mathbf{F}_{ij} + \sum_{r_i, r_j \in \mathcal{R}} \hat{\mathbf{F}}_{ij}}, \tag{3}$$

$$JSD = \frac{\mathbf{KL}(\mathbf{P}_{\mathbf{F}} || \mathbf{P}_{\hat{\mathbf{F}}}) + \mathbf{KL}(\mathbf{P}_{\hat{\mathbf{F}}} || \mathbf{P}_{\mathbf{F}})}{2}. \tag{4}$$

where the $\bar{\mathbf{F}}$ denotes the mean of elements in OD matrix $\mathbf{F}$, $\mathbf{KL}$ means Kullback–Leibler divergence, and $\mathbf{P}$ denotes the empirical probability distribution. The inflow is determined by totaling all flows entering each region, while the outflow is calculated by summing up all flows leaving each region.

4.2 Performance Comparison

The results are shown in Table 4. All models utilize the ratio of 8:1:1 for dividing the data into training, validation, and test sets. We conducted experiments five times and averaged the results.

The exploration of the graph generative modeling, WEDAN, achieves the best performance. The OD matrix generated by WEDAN demonstrates superior realism, from both flow value and property distribution deviation perspectives. Notably, in comparison to the top-performing baseline, WEDAN reduces RMSE/NRMSE by more than 8.0% and improves the CPC over 11.5%. Furthermore, the property distribution of the generated OD matrices closely matches the real ones, as evidenced by the lowest JSD from all the perspectives.

Table 4: Performance comparison on all existing models.

Model		Flow Value			Property Distribution (JSD)		
		CPC↑	RMSE↓	NRMSE↓	inflow↓	outflow↓	ODflow↓
Pair-wise	GM-P	0.321	174.0	2.222	0.668	0.656	0.409
	GM-E	0.329	162.9	2.080	0.652	0.637	0.422
	SVR	0.420	95.4	1.218	0.417	0.555	0.410
	RF	0.458	100.4	1.282	0.424	0.503	0.219
	GBRT	0.461	91.0	1.620	0.424	0.491	0.233
	DGM	0.431	92.9	1.186	0.469	0.561	0.230
	TransFlower	0.488	97.8	1.249	0.356	0.337	0.269
	GMEL	0.440	94.3	1.204	0.445	0.355	0.207
Network-based	NetGAN	0.487	89.1	1.138	0.429	0.354	0.191
	DiffODGen	0.532	74.6	0.953	0.324	0.270	0.149
	WEDAN	**0.593** (+11.5%)	**68.6** (+8.04%)	**0.876** (+8.04%)	**0.291** (+10.2%)	**0.269** (+0.96%)	**0.147** (+1.34%)

The performance of data-driven approaches significantly outperforms the physical model. The Gravity Model, using only four parameters, attempts to fit the complex human mobility, leading to inevitably underfitting. On the contrary, data-driven approaches, employing models with a multitude of parameters, go beyond by incorporating rich information such as demographics and POIs. Therefore, they have shown significantly better performance.

Modeling the joint distribution of all elements in OD matrices from the graph perspective hold advantages. Modeling the dependency between the area's spatial space and the OD matrix globally, as opposed to merely modeling human flows between two regions (i.e., origin and destination), results in a more effective capture of the properties of the mobility networks, i.e., OD matrices.

Utilizing training data from various massive areas can enhance the performance. Existing models based on graph generation have been designed only for large graphs, such as NetGAN and DiffODGen. In contrast, WEDAN is more versatile, capable of adapting to areas/graphs, of various sizes, from small to large. Consequently, it achieves more outstanding results.

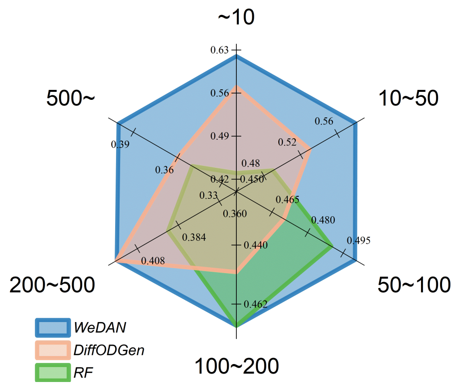
4.3 Performance Analysis on Heterogeneous Areas

To further explore the heterogeneity handled by the models and the applicability in different urban scenarios, we conducted comparative experiments on the model's performance across areas with various sizes and structures. Typically, developed areas are often larger and imply a stronger attraction to populations. Conversely, underdeveloped areas are usually small in size. Areas of different sizes also exhibit distinct mobility patterns, especially in terms of the skewness of OD flow distribution. Larger areas typically indicate stronger heterogeneity in mobility patterns from both node and edge perspectives, with a more pronounced long-tail effect in flow distribution.

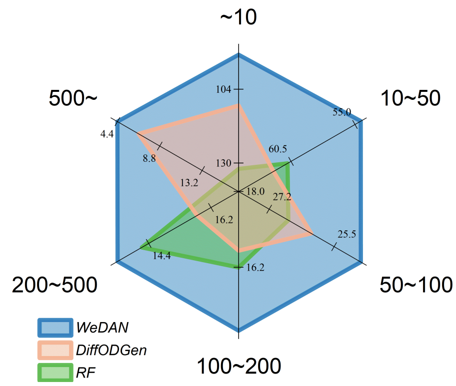
We divided the test areas into six groups based on the number of regions and into three groups based on structure, and the results are shown in Figure 6. We find that when trained under the new paradigm, WEDAN can consistently achieve optimal performance across areas of all sizes and structures. Polycentric areas often have a larger size and more complex pattern, as they develop satellite towns based on the original monocentric structure. Therefore, polycentric areas are more challenging to deal with. However, our model still achieved optimal performance in CPC. Larger areas tend to have more structured layouts, so smaller areas mostly fall into the 'others' category, resulting in better metrics for this category. While DiffODGen is specifically designed for large areas, our method can also enhance its performance by 11.1% on CPC and 33.3% on RMSE thanks to the various massive training data. Generative models demonstrate better adaptability to different structures of areas. And our method averagely improves the performance by 34.9% on RMSE on the monocentric and polycentric areas. Further analysis is conducted in Appendix D.5.

4.4 Analysis the Commonalities Captured Across Various Areas

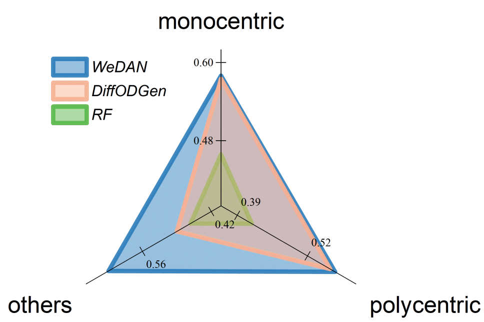
We conducted in-depth analysis of the dependencies captured across areas of different sizes and structures (Xu et al., 2023) in the new paradigm. Specifically, we utilize areas with varying sizes and structures to mutually serve as training and testing sets, thereby validating the capture of commonalities across areas. The results, as illustrated in Figure 7, we find that there are commonalities across areas with different sizes and structures, allowing for a certain degree of mutual transfer between them. Experiments have shown that a performance of 89.7% can be achieved solely through cross-type



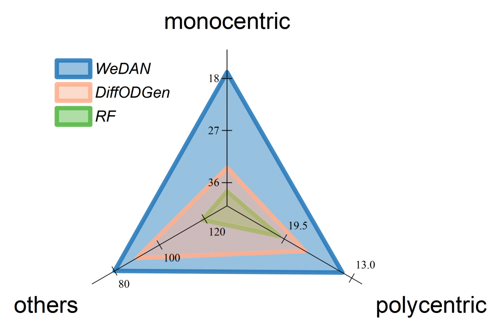
(a) Areas of different sizes on CPC.



(b) Areas of different sizes on RMSE.

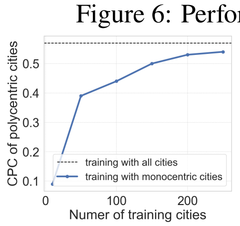


(c) Areas of different structures on CPC.

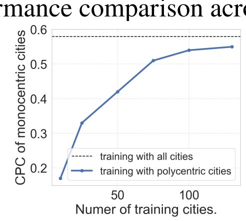


(d) Areas of different structures on RMSE.

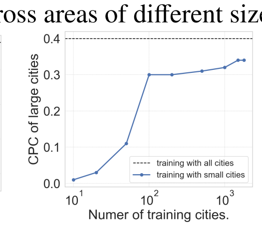
Figure 6: Performance comparison across areas of different sizes and structures.



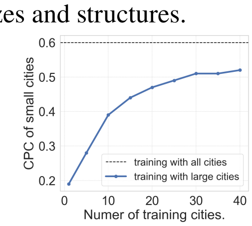
(a) From monocentric to polycentric areas.



(b) From polycentric to monocentric areas.



(c) From small to large areas.



(d) From large to small areas.

Figure 7: Analysis the dependencies captured across areas with different sizes and structures. The small areas consist of less than 100 regions, and the large areas consist of more than 500 regions. The black dash line represents the performance of training with all types of areas.

transfer learning and applications. Large areas contains more information about flows. Therefore, achieving a performance of 86.7% can be accomplished with only a small number of training large areas. Training the model with a diverse range of areas can enhance its generalizability, allowing it to achieve good performance across various areas. This indicates the validity of the novel paradigm. Extended analysis is conducted in Appendix D.6. To further explore the transferability of the model across even different countries, we conduct generation experiments on the United Kingdom, and the results are shown in Appendix D.4.

5 Conclusion

In this work, we introduce a large-scale commuting OD flow dataset (LargeCommuingOD) to support sysmatical comparison of existing studies and to facilitate the development of more powerful models. The dataset contains 3,333 areas around the United States including diverse urban environments. Besides, regions with each area are profiled with urban attributes, such as sociodemographics and POIs. Based on this dataset, we benchmark existing works with a common evaluation and find that network-based generative models may be a promising direction for future research, which could utilize the data collected from distinct areas to learn a more generalizable model. The model should capture the universal and distinct mobility patterns at the city level.

References

Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J Ramasco, Filippo Simini, and Marcello Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1–74, 2018.

Michael Batty. *Cities and complexity: understanding cities with cellular automata, agent-based models, and fractals*. The MIT press, 2007.

Aleksandar Bojchevski, Oleksandr Shchur, Daniel Zügner, and Stephan Günnemann. Netgan: Generating graphs via random walks. In *International conference on machine learning*, pp. 610–619. PMLR, 2018.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Mingfei Cai, Yanbo Pang, and Yoshihide Sekimoto. Spatial attention based grid representation learning for predicting origin–destination flow. In *2022 IEEE International Conference on Big Data (Big Data)*, pp. 485–494. IEEE, 2022.

Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs. *arXiv preprint arXiv:2012.09699*, 2020.

Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *nature*, 453(7196):779–782, 2008.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Ryuichi Imai, Daizo Ikeda, Hiroyasu Shingai, Tomohiro Nagata, and Koichi Shigetaka. Origin-destination trips generated from operational data of a mobile network for urban transportation planning. *Journal of Urban Planning and Development*, 147(1):04020049, 2021.

Md Shahadat Iqbal, Charisma F Choudhury, Pu Wang, and Marta C González. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40:63–74, 2014.

Shan Jiang, Yingxiang Yang, Siddharth Gupta, Daniele Veneziano, Shounak Athavale, and Marta C González. The timegeo modeling framework for urban mobility without travel surveys. *Proceedings of the National Academy of Sciences*, 113(37):E5370–E5378, 2016.

Ming Jin, Qingsong Wen, Yuxuan Liang, Chaoli Zhang, Siqiao Xue, Xue Wang, James Zhang, Yi Wang, Haifeng Chen, Xiaoli Li, et al. Large models for time series and spatio-temporal data: A survey and outlook. *arXiv preprint arXiv:2310.10196*, 2023.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Hadi Karimi, Sayed Nader Shetab Boushehri, and Ramin Nasiri. Origin-destination matrix estimation using socio-economic information and traffic counts on uncongested networks. *International Journal of Transportation Engineering*, 8(2):165–183, 2020.

Maxime Lenormand, Thomas Louail, Oliva G Cantú-Ros, Miguel Picornell, Ricardo Herranz, Juan Murillo Arias, Marc Barthelemy, Maxi San Miguel, and José J Ramasco. Influence of sociodemographic characteristics on human mobility. *Scientific reports*, 5(1):10075, 2015.

Maxime Lenormand, Aleix Bassolas, and José J Ramasco. Systematic comparison of trip distribution laws and models. *Journal of Transport Geography*, 51:158–169, 2016.

Mufei Li, Eleonora Kreačić, Vamsi K Potluru, and Pan Li. Graphmaker: Can diffusion models generate large attributed graphs? *arXiv preprint arXiv:2310.13833*, 2023.

Zhicheng Liu, Fabio Miranda, Weiting Xiong, Junyan Yang, Qiao Wang, and Claudio Silva. Learning geo-contextual embeddings for commuting flow prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 808–816, 2020.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Massimiliano Luca, Gianni Barlacchi, Bruno Lepri, and Luca Pappalardo. A survey on deep learning for human mobility. *ACM Computing Surveys (CSUR)*, 55(1):1–44, 2021.

Yan Luo, Zhuoyue Wan, Yuzhong Chen, Gengchen Mai, Fu-lai Chung, and Kent Larson. Transflower: An explainable transformer-based model with flow-to-flow attention for commuting flow prediction. *arXiv preprint arXiv:2402.15398*, 2024.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . `https://www.openstreetmap.org`, 2017.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Nastaran Pourebrahim, Selima Sultana, Jean-Claude Thill, and Somya Mohanty. Enhancing trip distribution prediction with twitter data: comparison of neural network and gravity models. In *Proceedings of the 2nd acm sigspatial international workshop on ai for geographic knowledge discovery*, pp. 5–8, 2018.

Nastaran Pourebrahim, Selima Sultana, Amirreza Niakanlahiji, and Jean-Claude Thill. Trip distribution modeling with twitter data. *Computers, Environment and Urban Systems*, 77:101354, 2019.

Caleb Robinson and Bistra Dilkina. A machine learning approach to modeling human migration. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, pp. 1–8, 2018.

PJ Rodriguez-Rueda, JJ Ruiz-Aguilar, J Gonzalez-Enrique, and I Turias. Origin–destination matrix estimation and prediction from socioeconomic variables using automatic feature selection procedure-based machine learning model. *Journal of Urban Planning and Development*, 147(4): 04021056, 2021.

Can Rong, Jingtao Ding, and Yong Li. An interdisciplinary survey on origin-destination flows modeling: Theory and techniques. *arXiv preprint arXiv:2306.10048*, 2023a.

Can Rong, Jingtao Ding, Zhicheng Liu, and Yong Li. Complexity-aware large scale origin-destination network generation via diffusion model. *arXiv preprint arXiv:2306.04873*, 2023b.

Can Rong, Jie Feng, and Jingtao Ding. Goddag: generating origin-destination flow for new cities via domain adversarial training. *IEEE Transactions on Knowledge and Data Engineering*, 2023c.

Can Rong, Huandong Wang, and Yong Li. Origin-destination network generation via gravity-guided gan. *arXiv preprint arXiv:2306.03390*, 2023d.

Meead Saberi, Hani S Mahmassani, Dirk Brockmann, and Amir Hosseini. A complex network perspective for characterizing urban travel demand patterns: graph theoretical analysis of large-scale origin–destination demand networks. *Transportation*, 44:1383–1402, 2017.

Meead Saberi, Taha H Rashidi, Milad Ghasri, and Kenneth Ewe. A complex network methodology for travel demand model evaluation and validation. *Networks and Spatial Economics*, 18:1051–1073, 2018.

Bhargava Sana, Joe Castiglione, Drew Cooper, and Dan Tischler. Using google's passive data and machine learning for origin-destination demand estimation. *Transportation Research Record*, 2672 (46):73–82, 2018.

Filippo Simini, Marta C González, Amos Maritan, and Albert-László Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, 2012.

Filippo Simini, Gianni Barlacchi, Massimilano Luca, and Luca Pappalardo. A deep gravity model for mobility flows generation. *Nature communications*, 12(1):6576, 2021.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

U.S. Census Bureau. 2009-2011 american community survey 3-year public use microdata samples [sas data file]. `https://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t`, 2012.

Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation. *arXiv preprint arXiv:2209.14734*, 2022.

Daniel Wang, Jack McFarland, Afra Mashhadi, and Ekin Ugurel. Comparing fairness of generative mobility models. *arXiv preprint arXiv:2411.04453*, 2024.

Yanyan Xu, Luis E Olmos, David Mateo, Alberto Hernando, Xiaokang Yang, and Marta C Gonzalez. Urban dynamics through the lens of human mobility. *Nat Comput Sci*, 3:611–620, 2023.

Xin Yao, Yong Gao, Di Zhu, Ed Manley, Jiaoe Wang, and Yu Liu. Spatial origin-destination flow imputation using graph convolutional networks. *IEEE Transactions on Intelligent Transportation Systems*, 22(12):7474–7484, 2020.

Yuan Yuan, Jingtao Ding, Jie Feng, Depeng Jin, and Yong Li. Unist: a prompt-empowered universal model for urban spatio-temporal prediction. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4095–4106, 2024.

Jinwei Zeng, Guozhen Zhang, Can Rong, Jingtao Ding, Jian Yuan, and Yong Li. Causal learning empowered od prediction for urban planning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 2455–2464, 2022.

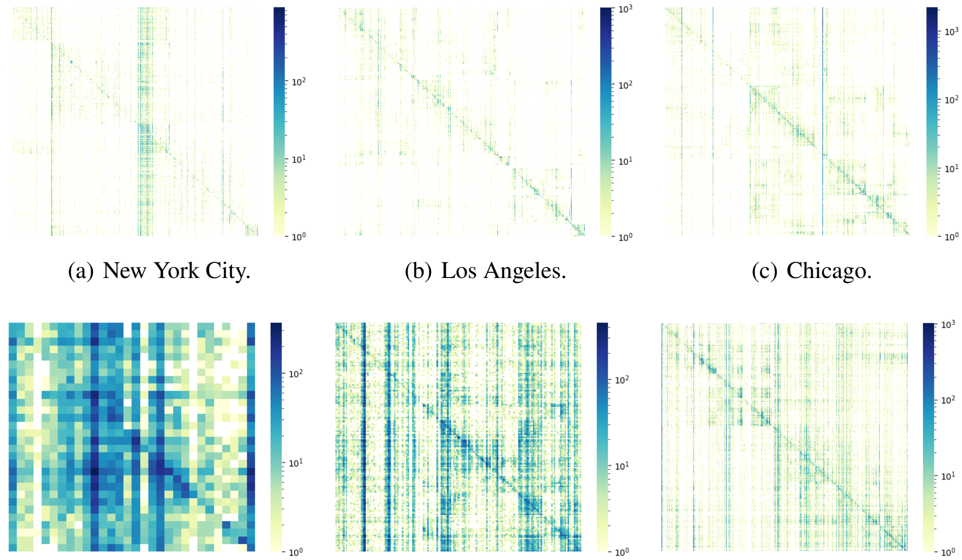
Chen Zhong, Stefan Müller Arisona, Xianfeng Huang, Michael Batty, and Gerhard Schmitt. Detecting the dynamics of urban structure through spatial network analysis. *International Journal of Geographical Information Science*, 28(11):2178–2199, 2014.

George Kingsley Zipf. The p 1 p 2/d hypothesis: on the intercity movement of persons. *American sociological review*, 11(6):677–686, 1946.

A Details of the Dataset

A.1 Visualization of the Commuting OD Flows via Heatmaps

We plot the heatmaps of the commuting OD flows for three large metropolitans and three counties in our dataset in Figure 8. The heatmaps show that the larger the city, the sparser the commuting OD flows are. This is because people tend to conduct closer commutes while the large city has more regions that are far away from each other. In small cities, the regions are all close to each other, leading to denser commuting OD flows.



(a) New York City. (b) Los Angeles. (c) Chicago.

(d) Etowah County. (e) Jefferson County. (f) Harris County.

Figure 8: Heatmaps of commuting OD flows for three large metropolitans and three counties in our dataset.

B Comprehensive Evaluation in the Benchmark: Interpretability, Robustness, and Fairness

B.1 Discussion on the interpretability of Benchmark Models

We discuss the interpretability of the benchmark models in this part. Because different models have different structures and mechanisms, their interpretability varies. We dive into the details one by one.

- Physical Models: Physical models are rooted in fundamental principles, providing strong interpretability through their clear and well-defined mathematical formulations.
  - Gravity Model: The interpretability of the gravity model can be summarized from two key aspects: (1) it highlights the production at the origin and attraction at the destination, both modeled using population size, and (2) it incorporates travel costs between regions, represented through distance decay functions (commonly power-law or exponential). The model includes three core variables: origin population, destination population, and distance, along with four parameters that control production, attraction, distance decay, and an overall scaling factor. This design makes the model intuitive and transparent in explaining how population and distance influence mobility. The formula of the gravity model is shown below.
$$F_{ij} = \lambda f_i(\mathbf{P}_i) f_j(\mathbf{P}_j) f_d(d_{ij}), \tag{5}$$
    where F_{ij} is the flow from region i to region j, λ is the scaling factor, f_i and f_j are the production and attraction functions, and f_d is the distance decay function.
  - Radiation Model: The radiation model models OD (Origin-Destination) flow by mimicking the physical process where particles are released from the origin region and absorbed by the destination region. The release of particles in the origin region is a function of the population, typically calculated as the total population multiplied by the proportion of the working rate. Whether the particles are absorbed by a destination region depends on the distance and the availability of its job opportunities. Specifically, the income associated with a job opportunity in a region is sampled independently from a probability distribution $p(z)$. The attractiveness of a region is quantified by the maximum job income available there, which determines its capacity to attract workers from other regions. Each individual also has an expected income threshold, which is defined as the maximum income they can earn in their home region. The decision-making process for job selection involves two steps: (a) individuals (analogous to

particles) are released from their home region; (b) they are absorbed by the nearest region offering a job income higher than their expected income. This process mirrors the radiation mechanism, where particles move and are absorbed based on specific criteria. As such, the model leads to the following derived macro formula.

$$< F_{ij} >= T_i P(1|m_i, n_j, S_{ij}) = T_i \frac{m_i n_j}{(m_i + S_{ij})(m_i + n_j + S_{ij})}, \tag{6}$$

where $< F_{ij} >$ is the OD flow from region i to region j, T_i is the outflow of region i, m_i is the number of jobs in region i, n_j is the total population of region j, and S_{ij} is the number of jobs in the circle region between i and j.

- Statistical Models: Statistical models are inherently data-driven, which may reduce their interpretability compared to theoretically derived physical models. However, they still offer strong interpretability by providing comprehensive insights at the input level, even if a full explanation at the parameter level is not achievable.
  - SVR leverages all support vectors from the training process as references, enabling strong interpretability by assessing the similarity between the new prediction target and each support vector. For instance, if a prediction sample has high similarity to a specific support vector, as determined by the kernel function, its predicted target will be closer to the target of that support vector. The contribution of this support vector to the prediction value will thus be more significant. In such cases, the features of the corresponding training sample can be referenced to explain the prediction target. For example, if certain features of a prediction sample closely resemble those of a support vector, it can be inferred that these features have a similar influence on the predicted value. As shown in Figure 9, the training samples are surrounded and profiled by the support vectors, which means that the distribution of the training samples is well captured by the support vectors.

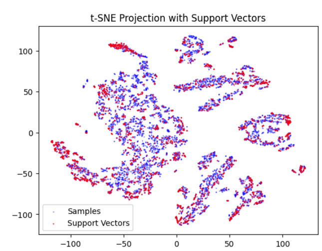
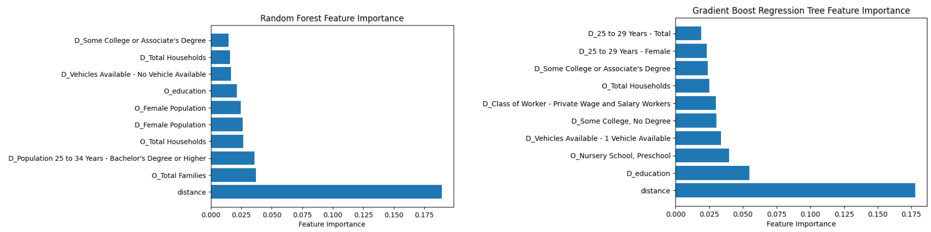


Figure 9: Visualization of the training samples and the support vectors.

  - Tree-based regression models (e.g., RF and GBRT) assess the importance of each feature by analyzing the conditions at each split and the proportion of data in the resulting subtrees. This allows the model to quantify the influence of each feature on the prediction results, providing a degree of global interpretability. Specifically, Random Forest quantifies feature importance by calculating each feature's contribution to reducing impurity (e.g., Gini index or information gain) during splits. Features with greater importance are frequently used as split conditions across multiple decision trees, highlighting their global impact on the predictions. As shown in Figure 10, the feature importance of each feature is visualized, providing a clear understanding of the model's interpretability. Specifically, the distance provides the most significant contribution to the prediction, followed by the population and features that can denote job opportunities.
- NN-based predictive models generally have weak interpretability due to their complex architectures and a large number of parameters, making it challenging to understand the specific meaning of individual parameters. However, certain techniques, such as SHapley Additive exPlanations (SHAP) and feature visualization, can provide a certain degree of interpretability. We combine these two approaches to discuss the interpretability of neural network-based predictive models in the benchmark.
  - DeepGravity: We utilize SHAP to obtain the interpretability of this model according to Simini et al. (2021). As shown in Figure 11, the global SHAP values are visualized to provide an overview of the feature importance. The population at working age and the economic activity index are the most influential features. It is interesting that the significant features are different



(a) Feature importance of random forest.

(b) Feature importance of gradient boosting decision trees.

Figure 10: Visualization of feature importance in tree-based models to access the global interpretability of the models.

from those in the tree-based models. This may indicate that DGM can capture more complex patterns from another perspective.

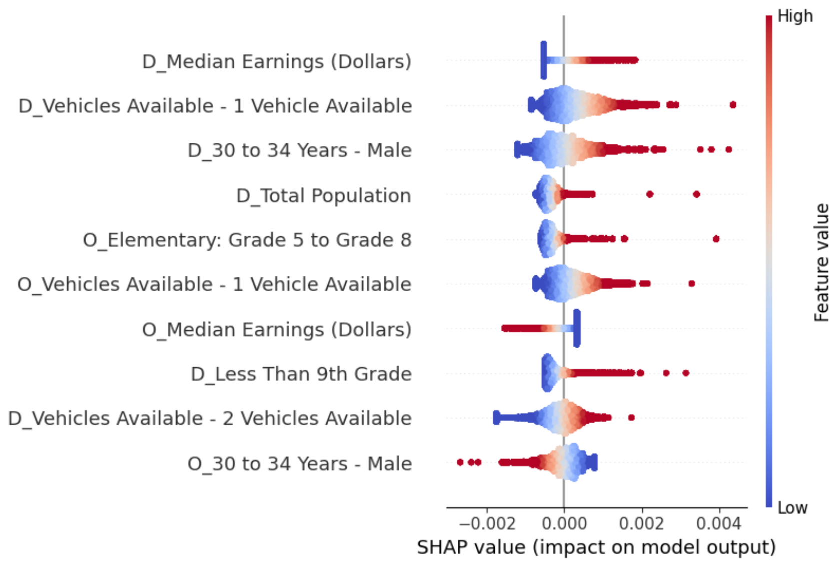
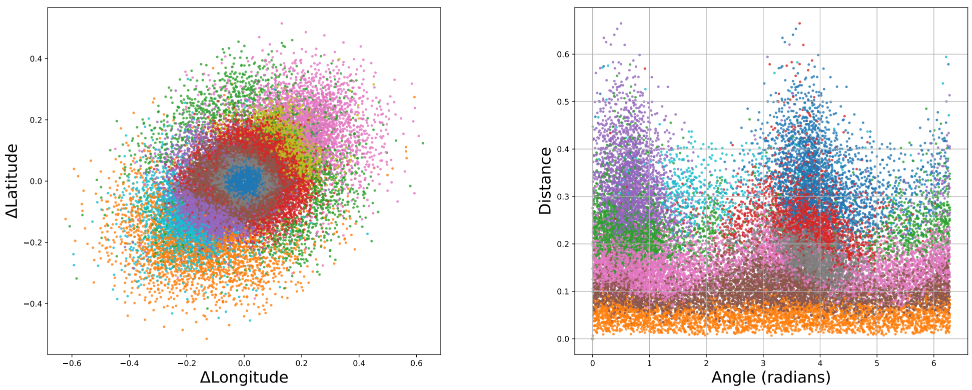


Figure 11: Visualization of the global SHAP values for the DeepGravity model.

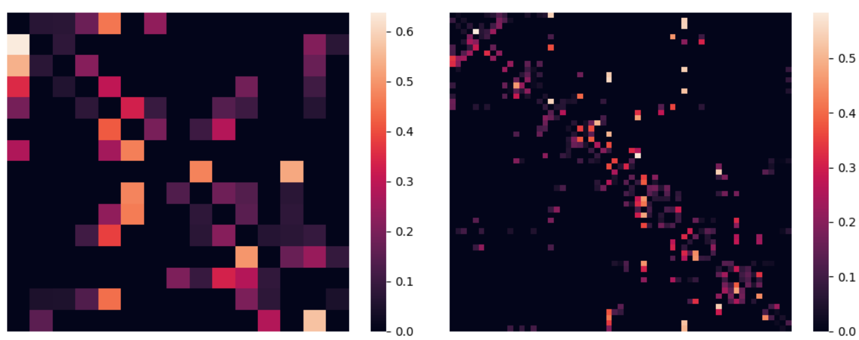
- TransFlower: We visualize the relative position embedding following Luo et al. (2024). It is important to note that TransFlower was originally designed for the setting where the outflow of each region is already given, generating OD flows toward a fixed number of destinations (256 in the work of Luo et al. (2024)). However, in our problem, the outflows are unknown, and we aim to model OD flows between all pairs of regions in the city. Limiting the number of destinations is therefore not applicable. As a result, we cannot use a model that predicts the probability distribution of flows from a given origin to a fixed number of destinations. We adapt this model to generate flows directly, thus the attention to destinations cannot be obtained. The visualization of the relative position embedding is shown in Figure 12. The clustering of the relative position embedding indicates that the model can capture the spatial relationships between regions even under unregularized division of the urban area. We can see that the embedding under the Cartesian coordinate system exhibits a clear circling patterns while the embedding under the polar coordinate system shows a fringe layer-like pattern. The regularity may not be as strong as Figure 3 in the original paper (Luo et al., 2024) because the original paper uses a grid-like division of the urban area, which is more regular than the unregularized division in our dataset. This demonstrates the strong interpretability of TransFlower in capturing the spatial relationships between regions.



(a) Catesian Coordinate System (b) Polar Coordinate System

Figure 12: Location embedding clustering of the relative location encoder in *Harris County* from the trained TransFlower.

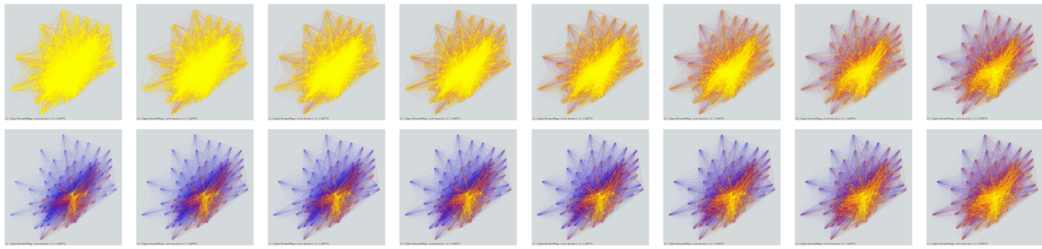
– GMEL: We visualize the attention weights in the graph attention networks within the model. These attention weights capture the similarity between regions, enhancing a region's representation by aggregating information from its similar neighbors. This provides a degree of interpretability aligned with the *First Law of Geography*, as shown in Figure 13(a).



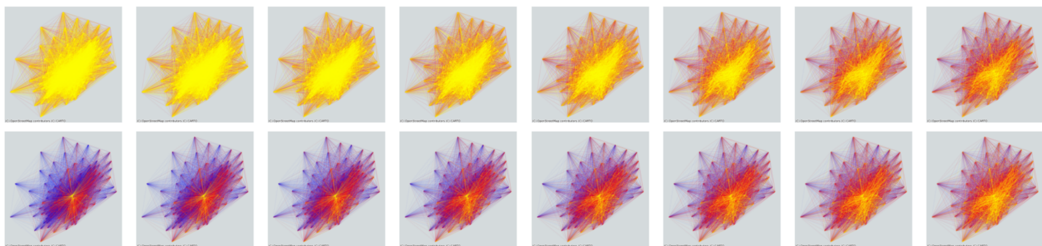
(a) Attention map of GMEL. (b) Attntion map of NetGAN.

Figure 13: Visualization of the attention relationships in graph attention networks of GMEL and NetGAN.

- Graph Generative Models: Generative models typically aim to fit the probability distribution of data, a challenging task that often results in highly complex model structures. As a result, they are generally the least interpretable class of methods. Additionally, generative models explicitly or implicitly handle randomness and noise in the data, using probabilities to generate nodes and edges—probabilities that are driven by random patterns in the data. Models like GANs and diffusion models inherently involve noise modeling: GAN generators often take Gaussian noise as input, while diffusion models explicitly model small noise in the diffusion process. This reliance on randomness and noise further reduces their intuitive interpretability. Moreover, generative methods are not well-suited for feature-level interpretability analysis using SHAP due to the high dimensionality of the conditional control variables, which scale as $N \times f$ (where N is the number of nodes and f is the dimention of node features). The number of features can also vary across samples with changes in N. Therefore, we use visualizations of attention mechanisms and denoising diffusion processes to discuss interpretability for NetGAN, DiffODGen, and WEDAN. Specifically, the attention map in NetGAN is shown in Figure 13(b).



(a) Desnoising diffusion process in DiffODGen.



(b) Desnoising diffusion process in WEDAN.

Figure 14: Visualization of the denoising diffusion process in DiffODGen and WEDAN.

In summary, as model complexity increases, performance continues to improve, but interpretability declines correspondingly. This presents a critical challenge: how to trade off the performance gains brought by complexity with the need for interpretability. Alternatively, exploring techniques that enhance the interpretability of complex models is a crucial direction for future research, as demonstrated by approaches like TransFlower.

B.2 Discussion on the Robustness of Benchmark Models on Edge Cases

To evaluate whether the models in our benchmark demonstrate good robustness on edge cases, we designed experiments to assess their performance under extreme large OD flows. Specifically, we measured the percentage of CPC on the top 5% largest OD flows relative to the overall CPC reported in Table 4. The results are presented in Figure 15 From the results, we observe that as model complexity increases, the ability to handle edge cases also improves, likely due to stronger nonlinear fitting capabilities. These models, during generation, continuously and smoothly model the distribution of commuting OD flows in urban spaces within the latent space. However, for edge cases, performance degradation is still observed to some extent. This is partly due to the strong long-tailed distribution of OD flows, where only a small number of extremely large flows are present, making it difficult to collect sufficient training data for these cases. Therefore, robustness on edge cases remains a challenge for such continuous modeling approaches in this field.

B.3 Discussion on the Fairness of Benchmark Models

We utilize the median earnings of the regions as a proxy for the economic status of the regions. We then divide the regions into equal-sized two groups: low-income regions and high-income regions. We then adopt Demographic Parity (PD) for OD flow modeling (Wang et al., 2024) to evaluate the fairness of the benchmark models. Specifically, we calculate the CPC for every region in each group and compare the difference between the distributions of the CPC values for the two groups. The results are shown in Figure 16. As we can see, the tree-based models exhibit the best fairness performance, with the smallest difference in the distributions of the CPC values between low-income and high-income regions. Graph diffusion-based models show a slightly higher performance for the high-income regions. The remaining models exhibit large DP values, but it seems like there is no obvious trend of modeling which group better. From the distribution differences shown in Figure 16,

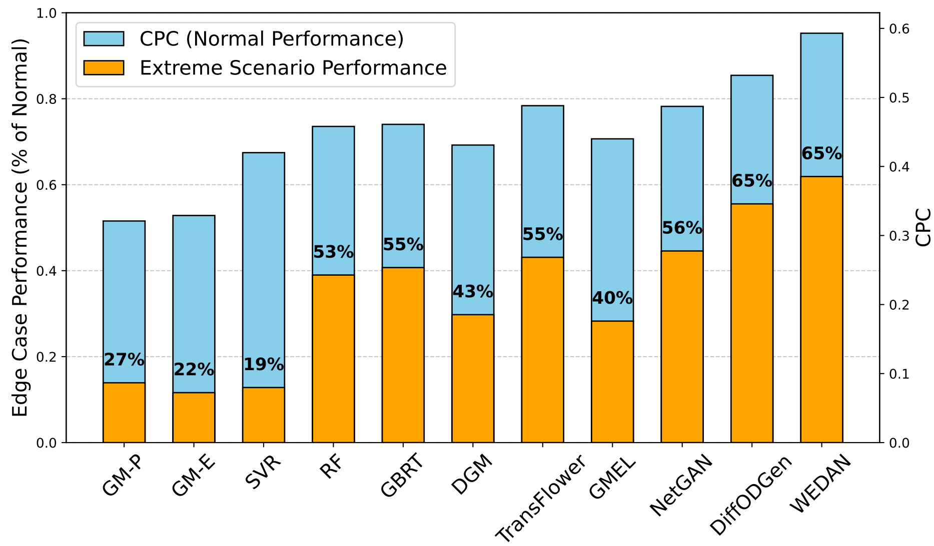
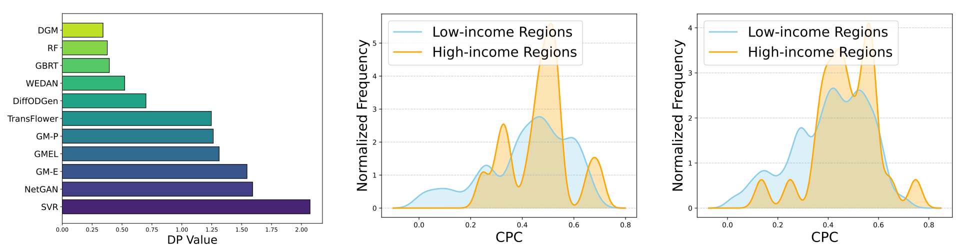


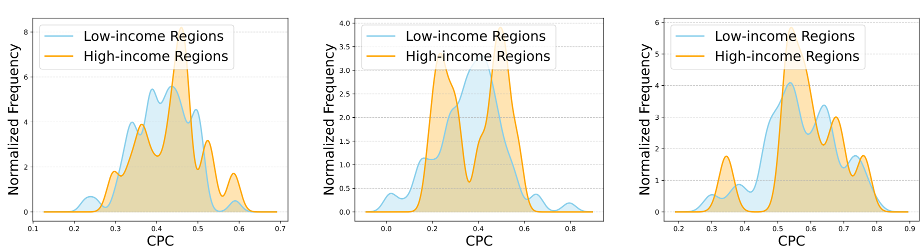
Figure 15: Performance of benchmark models on the top 5% largest OD flows. The left y-axis represents the percentage of CPC on the top 5% largest OD flows, while the right y-axis shows the overall CPC.



(a) Comparison of Demographic Parity Values for all models in the benchmark.

(b) Distributions of CPC values for low-income and high-income regions based on Support Vector Regression (SVR).

(c) Distributions of CPC values for low-income and high-income regions based on Random Forest (RF).



(d) Distributions of CPC values for low-income and high-income regions based on Deep-Gravity (DGM).

(e) Distributions of CPC values for low-income and high-income regions based on TransFlower.

(f) Distributions of CPC values for low-income and high-income regions based on WEDAN.

Figure 16: Analysis of fairness performance of benchmark models on regions with different income levels.

we can conclude that the distribution of the CPC values for low-income regions is more concentrated than that for high-income regions.

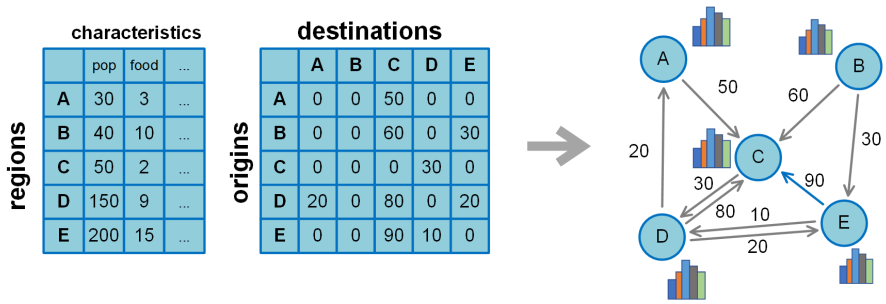


Figure 17: An example of construction of an attributed directed weighted graph formated by the spatial characteristics and commuting OD matrix of the corresponding area consisting of 5 regions.

The fairness performance is important but rarely studied in the field of commuting OD flow generation. Our analysis is a primary exploration of this topic, and we hope to inspire more research in this direction in the future.

C Additional Information About the New Paradigm

In this section, we give a detailed introduction to a new paradigm to solve the commuting OD flow generation supported by our comprehensive dataset. In the new paradigm, we consider the whole area combined with its commuting OD matrix as an attributed directed weighted graph. Thus, the commuting OD flow generation problem can be formulated as generating the weighted edges based on the attributed nodes. In this regard, we primarily adapt the graph generation model to the OD flow modeling task. And LargeCommuingOD containing diverse urban environments can support training on a large number of commuting OD networks, which can capture the universal and distinct mobility patterns at the city level, leading to better generalizablility. The comparison of the traditional transfer paradigm and our novel generative paradigm is shown in Figure 5.

To achieve better performance, we adopt the advanced diffusion-based graph generation model to generate the weighted edges condition on the attributed nodes, which named WEDAN (**W**eighted **E**dges **D**iffusion condition on **A**ttributed **N**odes). The framework of WEDAN is shown in Figure 18. We will introduce the relevant the graph construction, diffusion process, denoising network, and the training and generation process in detail next. The novelty of WEDAN shows in Appendix C.3.

Graph Construction. As shown in Figure 17, we model an whole area as a graph $G = (\mathcal{V}, \mathcal{E})$. Specifically, each node $v \in \mathcal{V}$ on the graph represents a region r within that area, and the directed edges $e_{ij} \in \mathcal{E}$ signify the commuting OD flows $\mathcal{F}_{r_i,r_j}$ between regions. Herein, we let $N = |\mathcal{V}|$ be the number of nodes in the graph, representing the number of regions, where $||$ denotes the cardinality of a set. Each edge corresponds to its unique origin node and destination node. The weight of each edge $w_{e_{ij}}$ is the OD flow volume F_{ij}. The graph is attributed with the spatial characteristics of each region r, which are represented as the node features $\mathbf{X}_v$ of each node. The graph construction process is illustrated in Figure 17. Thus, the spatial characteristics of an area $\mathcal{C}_{\mathcal{R}}$ can be represented by a feature matrix $\mathbf{X}_{\mathcal{R}}$ composed of the attributes of all nodes $\{v_r | r \in \mathcal{R}\}$ on the corresponding graph G, combined with the distances $\{d_{ij} | r_i \text{ and } r_j \in \mathcal{R}\}$ between all pairs of regions. Meanwhile, the commuting OD matrix $\mathbf{F}$ is equivalent to the set of all edges $\{e | e \in \mathcal{E}\}$ and their weights $\{w_e | e \in \mathcal{E}\}$ on its graph G.

By constructing a conditional generative model $\mathcal{P}_\theta(\mathcal{E}, \{w_e | e \in \mathcal{E}\} | \mathcal{V}, \mathbf{X}_{\mathcal{R}})$ that, given all nodes $\mathcal{V}$ and their attributes $\mathbf{X}_{\mathcal{V}}$ of a graph, generates all edges $\mathcal{E}$ and the corresponding weights $\{w_e | e \in \mathcal{E}\}$ on those edges, we can build an OD flow modeling model θ. The conditional distribution $\mathcal{P}_\theta(\mathcal{E}, \{w_e | e \in \mathcal{E}\} | \mathcal{V}, \mathbf{X}_{\mathcal{R}})$ mirrors $\mathcal{P}_\theta(\mathbf{F} | \mathcal{C})$.

Weighted Edges Diffusion Condition on Attributed Nodes. We will give a detailed introduction to the framework of the weighted edges diffusion process, which models the conditional distribution $\mathcal{P}_\theta(\mathcal{E}, \{w_e | e \in \mathcal{E}\} | \mathcal{V}, \mathbf{X}_{\mathcal{R}})$. As shown in Figure 18, the diffusion framework is composed of two parts: the forward diffusion process q and the reverse denoising process p_θ. Both processes take

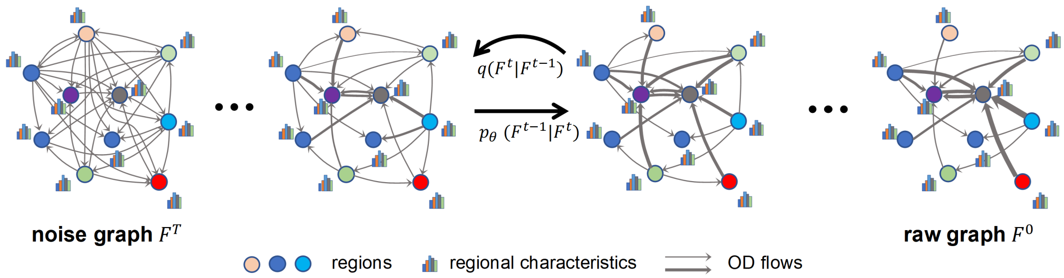


Figure 18: The framework of WEDAN for network-based generative commuting OD flow modeling.

place within the space of the edges $\mathcal{E}$ and the corresponding weights $\{w_e|e \in \mathcal{E}\}$ belong to the constructed attributed directed weighted graph.

Since OD matrices $\mathbf{F} \in \mathbb{R}^{N\times N}$ contain continuous flow values, our forward diffusion process utilizes Gaussian noise to perform the diffusion process. The forward diffusion process is shown in Figure 18 from the right to the left. It is important to note that the noise perturbations applied to all edges are independent. So the forward diffusion process can be described at the individual OD flow level by the following computational formula.

$$\begin{aligned} q(F_{ij}^t|F_{ij}^{t-1}) &= \mathcal{N}(F_{ij}^t; \sqrt{1-\beta_t}F_{ij}^{t-1}, \beta_t\mathbf{I}), \\ q(F_{ij}^1, ..., F_{ij}^T|F_{ij}^0) &= \prod_{t=1}^{T} q(F_{ij}^t|q^{t-1}). \end{aligned} \tag{7}$$

The reverse denoising process is the inverse of the forward diffusion process. In this context, the denoising process is facilitated by a denoising neural network θ, which predicts the small noise ϵ to be removed based on the latent state of the noise space at step t, aiming to reach the noise state of step $t-1$, in an iteratively manner. Unlike the forward diffusion process, to ensure the modeling of the joint distribution of all elements in the OD matrix $\mathbf{F}$, the noise to be removed for each edge needs to be determined based on the entire state of the corresponding noisy data $\mathbf{F}^t$. Furthermore, to ensure the generation of OD matrices for new cities with given their spatial characteristics, we have designed the denoising process of OD matrices to be guided by the spatial characteristics of the corresponding cities, i.e., the nodes and their features. Therefore, the denoising step in reverse process can be represented as follows.

$$p_\theta(\mathbf{F}^{t-1}|\mathbf{F}^t, \mathcal{C}_\mathcal{R}) = \mathcal{N}(\mathbf{F}^{t-1}; \mu_\theta(\mathbf{F}^t, t, \mathcal{C}_\mathcal{R}), (1-\bar{\alpha}^t)\mathbf{I}), \tag{8}$$

where

$$\mu_\theta(\mathbf{F}^t, t, \mathcal{C}_\mathcal{R}) = \frac{1}{\sqrt{\alpha_t}}(\mathbf{F}^t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{F}^t, t, \mathcal{C}_\mathcal{R})), \tag{9}$$

$\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$. Here, $\epsilon_\theta(\mathbf{F}^t, t, \mathcal{C}_\mathcal{R})$ is the noise predicted by θ based on the noisy state $\mathbf{F}^t$, diffusion step t and the spatial characteristics $\mathcal{C}_\mathcal{R}$ of the corresponding city.

The denoising network θ is trained to predict the noise $\epsilon_\theta(\mathbf{F}^t, t, \mathcal{C}_\mathcal{R})$ by minimizing the predictive errors. The parameterization and other detailed information of WEDAN is introduced in Appendix C, such as architecture of the denoising network, algorithms of training and generation processes.

Distance-based guidance. To fully utilize the association between spatial interactions $\{d_{ij}|r_i \text{ and } r_j \in \mathcal{R}\}$ and the OD matrix $\mathbf{F}$, we have designed node and edge levels distance-based conditional guidance to direct the denoising generation. As shown in Figure 19, we perform spectral decomposition on the distance matrix to obtain N Laplacian eigenvectors, which named distance-based Laplacian position encodings (d-LaPEs) are used to encode the specific position of each region in the planar urban space. Subsequently, the node features and edge features, before being inputted into each graph transformer layer, are combined with the corresponding d-LaPEs and distances.

Log-Transform. Existing theoretical works have discovered scaling behaviors in human mobility (Jiang et al., 2016; Saberi et al., 2017; 2018), namely that many properties follow the power law distribution. To enable our model to better capture the heterogeneity of OD flow distributions across different cities, we use log-transform to preprocess and post-process OD flows. The calculations are as follows.

$$\begin{aligned} \dot{F_{ij}} &= \log(F_{ij} + 1), \\ F_{ij} &= \exp(\dot{F_{ij}}) - 1. \end{aligned} \tag{10}$$

where $\dot{F_{ij}}$ is the log-transformed OD flow, which is used to train the denoising network. The generated $\dot{F_{ij}}$, after inverse transformation, yields the real size of OD flows F_{ij}.

C.1 Denoising Network

During each step in the reverse denoising process, the denoising network predicts the small Gaussian noise ϵ to be removed, based on the current noisy state. We adopt the transformer-based neural network structure as the backbone, which has been proven to have strong learning and generalization capabilities across various domains.

As illustrated in Figure 19, the backbone of our denoising network is the graph transformer (Dwivedi & Bresson, 2020). It accepts inputs at both the node and edge levels, captures graph features, and then outputs noise predictions at the edge level. The characteristics of each region serve as node inputs, and the noisy OD matrix at the current state provides the edge inputs. They are processed separately through their respective Multilayer Perceptrons (MLPs) and then fed into the graph transformer. The graph transformer consists of a series of layers. In each layer, every node computes attention weights with all other nodes through the self-attention mechanism and aggregates information from all other nodes based on these weights. To model the dependencies between nodes and edges, the weights computed through self-attention are fused with edge features using Feature-wise Linear Modulation (FiLM) (Perez et al., 2018), resulting in the final attention weights. Simultaneously, the calculated attention information is also used to combine with the original edge features, serving as the new edge features for subsequent computations in the next layer of denoising network. Moreover, after the aggregation of node and edge information, the data passes through a feed-forward network. The computations within each graph transformer layer can be described by the following formula.

$$\begin{aligned} h_i^{l+1} &= O_h^l \|_{k=1}^K \Big(\sum_{r_j \in \mathcal{N}_{r_i}} \alpha_{ij}^{k,l} V^{k,l} h_j^l \Big), \\ e_{ij}^{l+1} &= O_e^l \|_{k=1}^K (a_{ij}^{k,l}), \\ \alpha_{ij}^{k,l} &= softmax_j (a_{ij}^{k,l}), \\ a_{ij}^{k,l} &= \Big(\frac{Q^{k,l} h_i^l \cdot K^{k,l} h_j^l}{\sqrt{d_k}} \Big) + W^{k,l} e_{ij}^l, \end{aligned} \tag{11}$$

where h_i^l and e_{ij}^l are the node and edge features at the l-th layer, respectively. $Q^{k,l}$, $K^{k,l}$, and $V^{k,l}$ are the query, key, and value matrices of the k-th attention head at the l-th layer. $W^{k,l}$ is the weight matrix of the k-th attention head at the l-th layer. O_h^l and O_e^l are the output MLPs of the node and edge features at the l-th layer. d_k is the dimension of the query and key vectors. K is the number of attention heads. $\mathcal{N}_{v_i}$ is the set of neighbor nodes that are connected to node v_i.

After the layers, the final edge features are fed into the fully-connected layer to predict the noise.

C.2 Training and Generation

We use the simple loss from DDPM (Ho et al., 2020) to train the denoising networks in our attributed graph diffusion model. This involves minimizing the Mean Squared Error (MSE) between the noise predicted by the denoising network and the noise from the forward diffusion process. The calculation of this loss is as follows.

$$\mathcal{L} = \mathbb{E}_{t, \epsilon \sim \mathcal{N}(0, \mathbf{I})} \left[\| \epsilon - \epsilon_\theta (\mathbf{F}^t, t, \mathcal{C}_\mathcal{R}) \|_2^2 \right] \tag{12}$$

where $\|\dot{\|}$ denotes the $L-2$ norm. The training algorithm is shown in Algorithm 1. The training and sampling methods are detailed in Appendix D.1.

C.3 Extended Discussion on Related Works of WEDAN

WEDAN is a novel and original model that applies denoising diffusion-based graph generation models from a network perspective to commuting OD flow generation. To our knowledge, this model is unique and has not been proposed elsewhere. The key novelties of WEDAN lies in two aspects:

- It models all OD flows within a city as a directed weighted network, considering the entire OD network as a single data sample.
- It utilizes the features of all regions (nodes) in the OD network as guidance for the diffusion model, enabling the generation of all edges and their corresponding weights.

It is worth noting that GraphMaker Li et al. (2023) also generates attributed graphs, but they differ significantly: WEDAN is specifically designed for the commuting OD flow generation task, emphasizing that each OD flow is influenced by the attributes of its origin and destination nodes, resulting in continuous flow volumes. In contrast, GraphMaker focuses on generating large, sparse graphs by determining the existence of edges between nodes. Additionally, other works **?**Vignac et al. (2022) generate both nodes and edge weights simultaneously, emphasizing the coupling between nodes and edges rather than using node attributes to guide edge generation.

D Additional Experimental Details

D.1 Training Algorithm of Graph Denoising Diffusion

The trained denoising network can be utilized in conjunction with the reverse denoising process to generate the OD matrix for new cities, which lack any OD flow information, using their spatial characteristics. We adopt the sampling algorithm from Denoising Diffusion Implicit Models (DDIM) (Song et al., 2020) to facilitate more efficient data generation. The sampling algorithm is shown in Algorithm 2.

Algorithm 1 Training of the Graph Diffusion Model

Input:
Graphs $\mathcal{G}_{train}$ that constructed from the data collected from the cities in training set
Output:
Learned noise prediction neural networks θ.

```
Sample a graph G from G_train
Sample t ~ U(1, 2, ..., T)
Sample ϵ ~ N(0, I)
loss ⟸ || [ϵ − ϵ_θ(√(ᾱ^t) F + √(1 − ᾱ^t) ϵ, t, C_R)] ||^2
optimizer.step(loss)
```

1: Sample a graph G from $\mathcal{G}_{train}$
2: Sample $t \sim \mathcal{U}(1, 2, ..., T)$
3: Sample $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
4: $loss \Longleftarrow \left\| \left[\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}^t} F + \sqrt{1 - \bar{\alpha}^t}\epsilon, t, \mathcal{C}_\mathcal{R}) \right] \right\|^2$
5: optimizer.step($loss$)

Algorithm 2 OD Matrix Generation through Trained Graph Diffusion Model

Input:
Spatial characteristics $\mathcal{C}_\mathcal{R}$ of a new city
Trained denoising network θ
Length τ of sub-sequence in DDIM sampling
Output:
OD matrix $\mathbf{F}$ of that new city.

1: Sample $\mathbf{F}^T \sim \mathcal{N}(0, \mathbf{I})$
2: $\Delta t = \frac{T}{\tau}$
3: **for** $t = T, T - \Delta t, ..., 1$ **do**
4: $\quad \mathbf{F}^{t-\Delta t} \leftarrow \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{F}^t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(\mathbf{F}^t, t, \mathcal{C}_\mathcal{R}) \right)$
5: **end for**
6: **return** $\mathbf{F}^0$

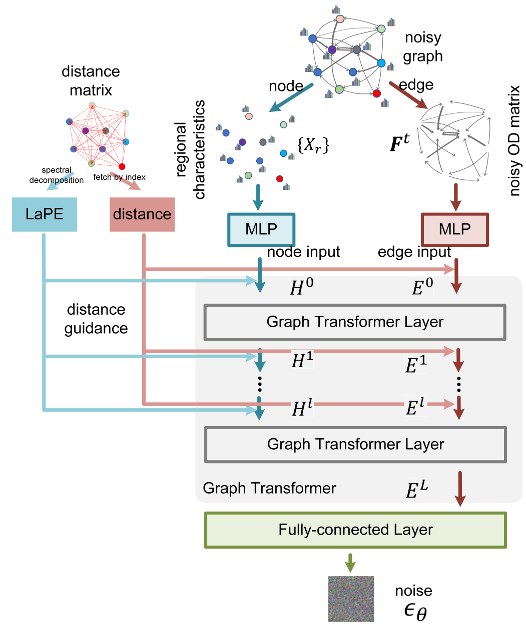


Figure 19: The architecture of denoising neural network θ of our graph diffusion model.

D.2 Architecture of Denoising Graph Transformer

D.3 Details for Reproducibility

The computational resources we used to conduct the experiments are as follows: a server with a Intel(R) Xeon(R) Platinum 8358P CPU @ 2.60GHz with 128 cores. The server is equipped with 1TB of RAM and 8 NVIDIA A100 GPUs. For all the experiments in the paper, we run 5 trials and report the average results.

D.4 Additional Results on Transferability to Other Countries

Model	CPC↑	RMSE↓	NRMSE↓
GM-P	0.240	101.6	1.752
RF	0.334	223.2	3.847
DGM	0.359	157.0	2.706
GMEL	0.362	149.1	2.570
NetGAN	0.331	198.9	3.429
WEDAN	**0.485**	**72.68**	**1.253**

Table 5: Transferability experiments of training models on the United States and generating OD flows for the United Kingdom.

We generated commuting OD flows for 326 Local Authority Districts (LADs) in the United Kingdom, where the flows among Middle Layer Super Output Areas (MSOA) within each LADs. The ground truth was obtained from the Office for National Statistics (ONS) of the UK. It is very difficult to obtain regional features in the UK with the same format and semantics as in the US. Therefore, we used satellite images of regions to represent the input features consistently across these two countries. The experimental results are shown in Table 5. The results show that WEDAN outperforms other models in terms of CPC, RMSE, and NRMSE, demonstrating its strong generalization ability to other countries. This indicates that models trained on the US dataset exhibit some transferability to other countries, particularly to developed countries like the UK. WEDAN benefits from graph generative modeling, achieving the best performance. However, the transferability cannot always be guaranteed, as there may be significant differences between countries. We aim to explore this direction in future work.

D.5 Details of Performance on the Heterogeneity of Urban Areas

From Figure 6(a), we observe that all models tend to perform better in smaller cities in terms of CPC, with performance declining as city size increases. This trend can be attributed to the increasing heterogeneity in OD flow distributions in larger cities. Smaller cities often have more homogeneous region-pairs with short-distance flows, making predictions relatively easier. In contrast, larger cities have both short-distance and long-distance commuting, leading to a long-tailed distribution of OD flows and higher heterogeneity, which increases prediction difficulty. Figure 6(b) further illustrates that smaller cities tend to have higher RMSE values. This is because smaller cities typically exhibit higher flow volumes due to a prevalence of short-distance commuting, which increases the absolute prediction error. Conversely, larger cities often have sparser OD flows between many distant regions, with certain extreme flows contributing large values but overall lower flow volumes, resulting in smaller RMSE. Figures 6(c) and 6(d) support similar conclusions for cities of varying structures. For larger monocentric and polycentric cities, models like DiffODGen, which incorporate hierarchical designs for large cities, perform well. However, DiffODGen struggles with the ”others” category, typically smaller cities, where its performance is less reliable. In contrast, WEDAN, benefiting from large-scale training data, demonstrates robust performance across all city sizes and structures.

D.6 Detailed Analysis on Commonalities Captured across Urban Areas

Figure 7 reveals that cities of different types share certain common human mobility patterns, supporting the feasibility of using a unified model to learn mobility patterns across diverse cities. Modeling both commonalities and distinctions between cities helps enhance the model's generalization capability. Figures 7(a) and 7(b) show that both monocentric and polycentric cities achieve high performance during training, likely because these city types cover a wide range of human mobility patterns. However, Figures 7(c) and 7(d) highlight that training solely on small or large cities fails to achieve strong transferability across each other. This result demonstrates the existence of differences in human mobility patterns across city types while also highlighting the value of training on a diverse set of city types. Such diverse training data enables the model to effectively capture both the differences and the shared mobility patterns between cities.