

# CausalMob: Causal Human Mobility Prediction with LLMs-derived Human Intentions toward Public Events

Xiaojie Yang

The University of Tokyo

Tokyo, Japan

xiaojieyang@g.ecc.u-tokyo.ac.jp

Zipei Fan\*

School of Artificial Intelligence

Jilin University†

Changchun, China

fanzipei@jlu.edu.cn

Hangli Ge

The University of Tokyo

Tokyo, Japan

hangli.ge@koshizuka-lab.org

Renhe Jiang

The University of Tokyo

Tokyo, Japan

jiangrh@csis.u-tokyo.ac.jp

Jiawei Wang

The University of Tokyo

Tokyo, Japan

jiawei@g.ecc.u-tokyo.ac.jp

Ryosuke Shibasaki

The University of Tokyo

Tokyo, Japan

shiba@csis.u-tokyo.ac.jp

Noboru Koshizuka

The University of Tokyo

Tokyo, Japan

noboru@koshizuka-lab.org

## Abstract

Large-scale human mobility exhibits spatial and temporal patterns that can assist policymakers in decision making. Although traditional prediction models attempt to capture these patterns, they often interfere by non-periodic public events, such as disasters and occasional celebrations. Since regular human mobility patterns are affected by these events, estimating their causal effects is critical to accurate mobility predictions. News articles provide unique perspectives on these events, though processing is a challenge. In this study, we propose a causality based prediction model, **CausalMob**, to analyze the causal effects of public events. We first utilize large language models (LLMs) to extract human intentions from news and transform them into features that act as causal treatments. Next, the model learns representations of spatio-temporal regional covariates from multiple data sources to serve as confounders for causal inference. Finally, we present a causal effect estimation framework to ensure that event features remain independent of confounders during prediction. Based on large-scale real-world data, the experimental results show that the proposed model excels in human mobility prediction, outperforming state-of-the-art models.

## CCS Concepts

- Information systems → Spatial-temporal systems; • Human-centered computing → Ubiquitous and mobile computing theory, concepts and paradigms.

\*Corresponding author. †First affiliation.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD '25, August 3–7, 2025, Toronto, ON, Canada

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-1245-6/25/08

<https://doi.org/10.1145/3690624.3709231>

## Keywords

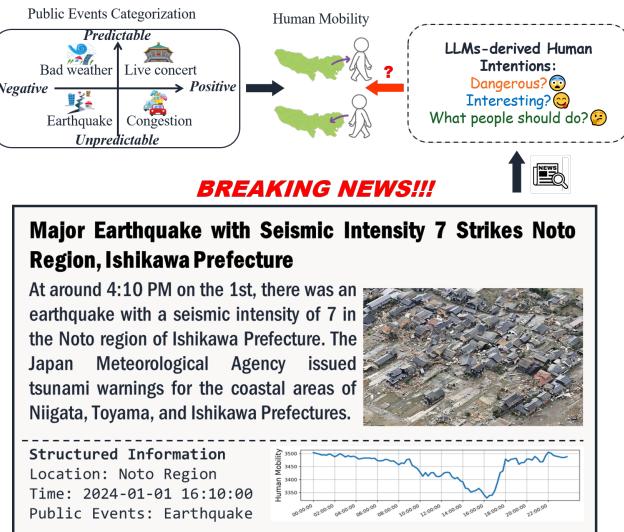
Human Mobility Prediction, Causal Inference, LLMs

### ACM Reference Format:

Xiaojie Yang, Hangli Ge, Jiawei Wang, Zipei Fan, Renhe Jiang, Ryosuke Shibasaki, and Noboru Koshizuka. 2025. CausalMob: Causal Human Mobility Prediction with LLMs-derived Human Intentions toward Public Events. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (KDD '25), August 3–7, 2025, Toronto, ON, Canada*. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3690624.3709231>

## 1 Introduction

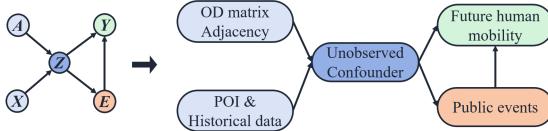
Human mobility exhibits spatio-temporal patterns that can aid policy-makers in urban management and decision-making. However, large-scale public events complicate the capture of these mobility patterns, resulting in less robust predictions [8, 14, 29]. These public events change human mobility in various ways, driven by corresponding human intentions [2, 12]. For example, as shown in the "Public Event Categorization" of Figure 1, typhoons can be predicted through weather reports, causing people to stay home. In contrast, emergencies like earthquakes have similar effects but are harder to forecast. Additionally, live concerts and traffic congestion can increase human volumes in specific regions, with the latter often links to unexpected traffic accidents. Based on the above cases, our research categorizes public events based on their predictability and their impact on human mobility (positive or negative to count increases), as shown in Figure 1 [38]. Understanding and analyzing human mobility and intentions toward public events using this categorization criteria can help decision-makers better prepare for potential opportunities and risks. However, current research often captures the statistical dependencies between human mobility and public events without analyzing the *causal relationship* between them [48]. This limitation is significant because human intentions play a crucial role in how people respond to public events. To support our point, we provide an example of an earthquake occurring on January 1, 2024, along with the corresponding human mobility in Figure 1. Therefore, to address this limitation, we



**Figure 1: Conceptual graph of this research**

leverage LLMs to simulate human intentions toward public events, enabling an explainable exploration of the causal relationship in human mobility [18, 34].

However, traditional causal inference methods that use randomized controlled trials to study causal effects [4] face difficulties in abstracting public event features and controlling their occurrence. Instead, researchers overcome this challenge by leveraging counterfactual analysis to study causal effects in different scenarios [21]. Given this background, we present a causal graph to discuss the causal relationships among variables in this research, as illustrated in Figure 2. We aim to analyze the causal effect of public events  $E$  (treatment) on human mobility  $Y$  (outcomes) by distinguishing the inference of spatio-temporal regional confounders (referred to as confounders in the following content)  $Z$  that we learn from observation data, including spatio-temporal covariates  $X$  and Origin-Destination (OD) matrix adjacency  $A$  [20, 48]. Benefiting



**Figure 2: Causal graph of this research**

from the developed media industry, we can obtain high-quality reports on public events through news articles and extract structured information from them [17, 35]. This is illustrated in Figure 1, which includes translated news articles and corresponding human mobility data that fluctuated significantly due to an earthquake. Furthermore, based on previous research, with the help of LLMs, we can extract structured information more efficiently from articles and simulate human intentions toward public events as features through prompt engineering [7, 24].

Accurate estimations of causal effects rely on the unconfoundedness assumption, which requires learning confounders with all features independent of treatments [36]. As a weak assumption, our research employs a causal effect calculation framework to remove dependencies between public events and confounders. We begin by extracting structured information from news articles and simulating mobility-related human intentions based on prompted LLMs. Next, we learn the representations of the confounders from observational data. In the causal inference part, an inference network predicts future human mobility, and a re-weighting network helps to remove event bias for each region [36]. Finally, we estimate the causal effect of public events on human mobility. We evaluated our proposed method on real-world data, including human mobility data and news articles, covering most major administrative regions in Japan. Our contributions are as follows:

- We introduce LLMs to process news articles, extracting structured information on public events and simulating human intentions as treatment features in causal inference.
- We develop a causal framework to learn confounder representations and remove their influence from the effects of public events represented with multidimensional features.
- We conducted various experiments to demonstrate the superior performance of our proposed method and provided interpretable results with thorough analysis.

## 2 Related Work

**Human Mobility Prediction.** In recent years, the development of geolocated datasets and data mining technologies has made human mobility increasingly important for urban management, transportation and disaster response [6, 30]. Benefiting from high-performance computational power, large-scale human mobility prediction based on deep learning has achieved good performance [47]. Researchers have focused on capturing periodical and semantic information to represent regional features for prediction. For example, STGCN [44] captured long-term periodical spatio-temporal patterns based on graph structures for traffic prediction, while DeepSTN+ [19] modeled long-range spatial dependence of different regions using POI features as semantic information. These models effectively capture regular patterns in human mobility. However, for irregular situations with low periodicity, such as disasters or festivals, previous research tends to model mobility under specific conditions, making predictions less generalizable [3, 13, 39].

**Causal Analysis for Human Mobility.** Causal analysis aims to reveal cause-effect relationships among variables [23]. Many previous studies have separated treatments from traditional inputs to analyze their effects on outcomes. These treatments are usually single variables, including discrete and continuous variables [43]. However, in the real world, treatment variables affecting human mobility are much more complex. Previously, causal analysis for human mobility has focused on specific situations. For example, Ma [20] introduced causal inference to assess the effectiveness of COVID-19-related policies by learning confounders from human mobility data and residents' vigilance from Google Trends. Similarly, Zhang [48] used these data to assess the effect of typhoons, represented by wind speed and precipitation, on human mobility.

### 3 Data and Analysis

In this section, we describe the observation data used to estimate the causal effects of public events on human mobility. We also include data analysis of public events and human mobility patterns to demonstrate the necessity of introducing causality analysis.

#### 3.1 Observational Data

This research aims to causally analyze the relationship between public events (treatment) and human mobility (outcomes) in Japan. We select 490 administrative regions (referred to as *regions* in the following content) across Japan, as shown in Figure 3(a). Some areas were excluded due to insufficient human mobility data (aggregated from GPS records provided by BlogWatcher Inc. [1]). We continuously filter these data within the time range of 2023/04/01 to 2024/03/31, corresponding to the official fiscal year in Japan.

**Treatment E:** Public events features extracted from news articles. We collaborate with Kyodo News, one of the largest international news agencies, which distributes news to almost all newspapers, radio, and television networks in Japan, reaching over 50 million subscribers [41]. With the same time range as the human mobility data, we filter out unrelated articles as data-cleaning process. We then align the news articles with the corresponding regions using manually labeled region codes. To further analyze the effects on human mobility, we introduce LLMs to extract structured information and human intentions as features from these filtered news articles. More details will be provided in the methodologies section.

**Outcome Y:** Future human mobility in selected regions. This study focuses on human mobility affected by the causal effects of public events. We aggregate human mobility data (GPS records) in selected regions as our prediction target. During this process, all users remain anonymous, and we take the hourly samples of *uids* aggregation within mentioned time ranges.

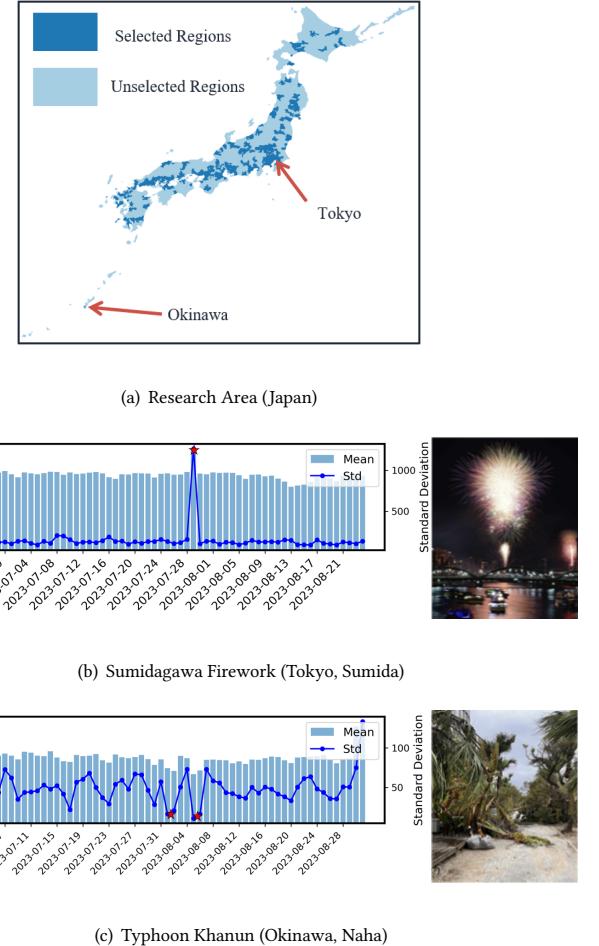
**Covariates X:** Spatio-temporal regional covariates. We introduce proxy variables as region covariates to capture the unobserved confounders *Z*. We collect Points of Interest (POI) from OpenStreetMap (OSM) and categorize them to represent region features [19]. Before calculating confounders *Z*, we concatenate POI features, time embeddings, region embeddings, and historical human mobility data. More details will be provided in the methodologies section.

**Networks A:** Re-normalized Mobility network. We use a human mobility network to model the inter-regional interactions and capture unobserved confounders in the selected regions. We aggregate the hourly Origin-Destination (OD) matrix adjacency within the regional network to represent the strength of each region's relationship. Each edge in this network contains a time-varying weight computed based on mobility volumes between two regions in the OD matrix. We re-normalize this mobility network, resulting in a weighted adjacency matrix  $A_t$  for each timestamp [16].

#### 3.2 Preliminary Data Analysis

To explore the mentioned causal relationship, we conduct preliminary data visualization analysis to illustrate the dependence between changes in human mobility and the occurrence of public events. We manually select two cases including a firework and a typhoon from our datasets and visualize them, respectively.

As illustrated in Figure 1, public events are divided into four



**Figure 3: Public events examples along with changes of daily human mobility. We calculate mean values and standard deviations of daily human mobility records (★ markers refer to the dates when public events happened).**

types based on their positive or negative effects on human mobility and predictability. While most cases are simple and straightforward, *positive unpredictable* events are much harder to capture. We assume some might be accompanied by *negative unpredictable* events or side effects of policies, such as the behavior of moving to low-cost areas due to remote working policies during COVID-19. Public events of this type require more complex reasoning processes, so we do not focus on them in this research. Furthermore, news articles are sometimes not temporally synchronized with public events, which is one of the main reasons we introduce LLMs to correct the exact occurrence time of public events, as part of structured information. Here, we provide cases as follows and related regions are marked in Figure 3(a) with daily human mobility patterns:

**Sumidagawa Firework Festival.** Held at 19:00 on June 29, 2023, the Sumidagawa Firework Festival was the largest fireworks festival in Japan, attracting over one million people including locals and visitors to the related regions in 2023. As shown in Figure 3(b),

human mobility maintained regular patterns except on the 29th. **Typhoon Khanun.** Hitting Okinawa on the 2nd and 4th of August 2023, the typhoon caused widespread power outages after its landfall. As shown in Figure 3(c), human mobility changed significantly during the typhoon and recovered one day later.

Based on the above cases, different public events have varying effects on human mobility, with both positive and negative impacts as we mentioned. In this research, we aim to capture these effects from a causal perspective. In this way, we can predict how human mobility will change if similar public events occur in other regions.

### 3.3 Problem Definition

In this research, we investigate the causal effects of public events on human mobility. Our methodology leverages large language models (LLMs) with chain-of-thought prompting to extract structured information, including human intentions, denoted as  $N = [n_1, n_2, \dots]$ , from news articles. Additionally, we process aggregated human mobility data, represented as  $\mathbf{X}$ , and mobility networks, denoted as  $A$ , to predict future human mobility patterns,  $Y$ , in selected regions over fixed time steps. In particular, we focus on crowd flow, measured as human volume in each region, treating it as the main mobility feature for analysis.

## 4 Methodology

In this section, we explain why and how to obtain human intentions towards public events extracted from news articles as treatments with LLMs and describe how we formulate the causal effect estimation problem in human mobility prediction.

### 4.1 Human Intentions Extraction with Semantic-Aware LLMs

In most situations, news articles report public events objectively, allowing us to estimate mobility-related human intentions. Our dataset contains over 130,000 news articles of one year, covering multiple genres as shown in Figure 4, making manual processing impractical. Taking advantage of the development of open source LLMs, we can extract the necessary structured information [11] and simulate human intentions related to human mobility using prompt engineering, starting with simple manual filtering [15].

To enable LLMs to reason about human mobility-related inten-

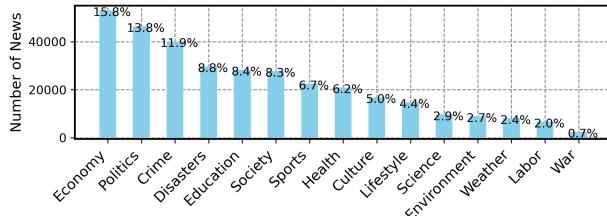


Figure 4: Number of News labeled with categories

tions, we design a step-by-step dialog framework as shown in Figure 5 inspired by chain of thought [33, 40]. With carefully designed prompts, we aim for LLMs to extract necessary information from

news articles on the following aspects: 1) Most influential events, 2) Time of the event, 3) Four basic questions related to human mobility, i.e., where, when, who, and how (3W1H), 4) Predictability of events, and 5) Human intentions of the public events with scores. Due to space limitations, prompt details can be found in the Appendix E. Historical answers will inform LLMs' future reasoning. To ensure the robustness of the generated answers for each prompt, we use one of the most powerful open-source LLMs, *Llama3-70b* [31] based on a library called *Ollama* [22]. Finally, we obtain structured information of public events as shown in the top-right of Figure 5. Specifically, human intentions toward public events will be used as causal treatments on human mobility in our prediction model. Furthermore, news cases are listed in Appendix C.

### 4.2 Causal Effect Inference through Counterfactual Estimation

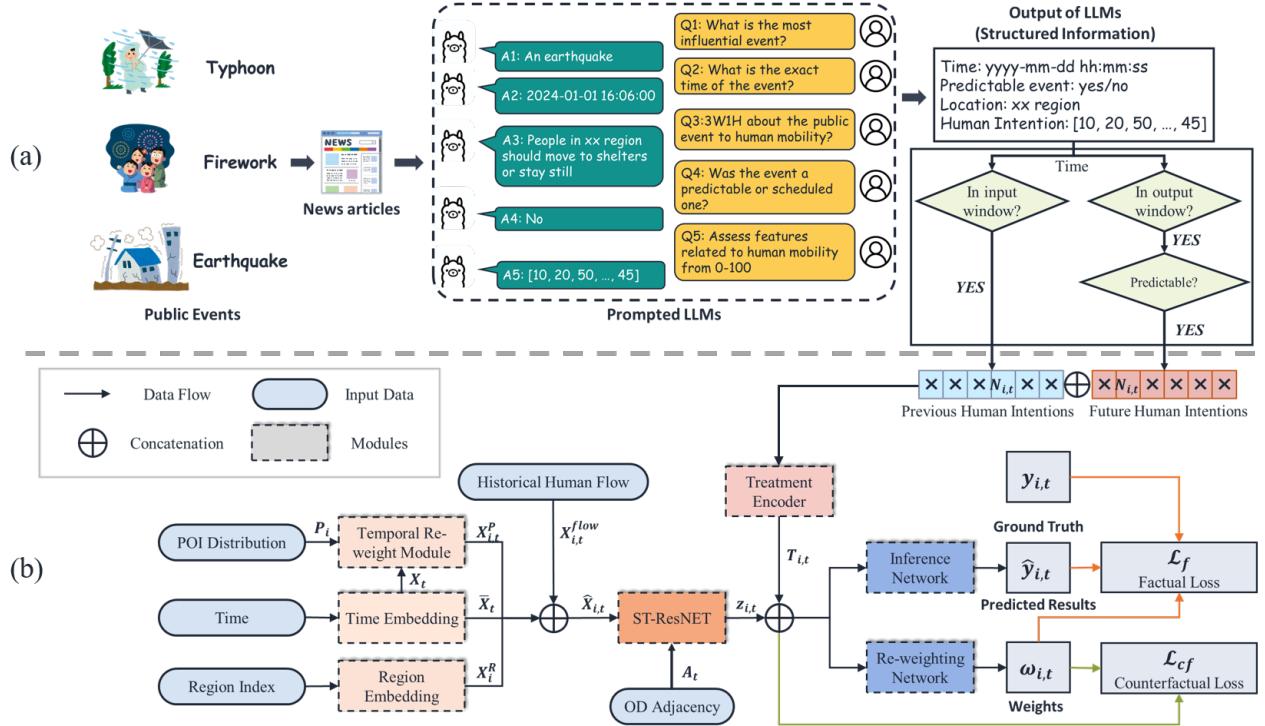
For features of  $n$  selected regions, we collect POI information and divide them into  $c$  specific categories according to their labels. We denote this feature as  $P_i = \{p_i^j\}_{j=1}^c$ , where  $p_i^j$  refers to the  $j$ -th category of the  $i$ -th region. Using human mobility data consisting of GPS coordinate records, we aggregate anonymous *uids* in one region with a fixed time interval. This can be represented as  $X_t^{flow} = \{x_{i,t}^{flow}\}_{i=1}^n$ . Furthermore, considering that human mobility is highly dependent on regional interactions, we used a  $n \times n$  normalization adjacency matrix  $A_t$  at timestamp  $t$  to represent time-varying relations between regions. In Figure 5, we calculate confounder  $Z_t = \{z_{i,t}\}_{i=1}^n$  with observation data, where  $X_i^{Reg}$  refers to natural region representations,  $X_t$  is the embedded timestamp, and  $X_{i,t}^n$  is temporally re-weighted POI feature.

As for treatments for causal inference, we obtain structured information including information about time, location and human intentions toward public events of each region  $N_{i,t}$  at time step  $t$  with prompted LLMs, and we conclude observed events in the input window and scheduled or predictable ones in the output window, as the flow chart in Figure 5. When it comes to evaluating treatment effects, one of the challenges is that previous studies focused on binary or continuous covariates in one dimension, while in our research we are facing complicated event features depicted by news articles over multiple time steps, and we need to convert them to a single feature  $T_{i,t}$ . To calculate the causal effect of different public events on human mobility, we need to calculate the average treatment effect (ATE) at timestamp  $t$  of  $i$ -th selected region in our research area:

$$\tau_{ATE}^{i,t} = \mathbb{E} \left[ y_{i,t}^1 - y_{i,t}^0 | z_{i,t}, T_{i,t} \right], \quad (1)$$

where  $y_{i,t}^1$  and  $y_{i,t}^0$  refer to potential outcomes of human mobility with and without causal effects of public events under the same spatio-temporal condition, corresponding to the outcomes  $Y$ .

Generally, it is impossible to observe the effects of different public events in one region simultaneously. Furthermore, in our observation data, public events show biases because of confounder imbalance and non-randomized treatments among regions. It means that some events tend to happen in certain regions, which will lead to unreliable inference. Processing public event features into a one-dimensional feature will result in significant information loss. Fortunately, we can quantify the *similarity* or *strength* of public events



**Figure 5: Framework of CausalMob:** We first process news articles to obtain structured information of public events in part (a). Then in part (b) we use ST-ResNET to learn the representation of the confounder  $z$ . Finally, an inference network and a re-weighting network are built to calculate factual loss and counterfactual loss as causal inference.

at a semantic level. Therefore, we introduce the concept of *average dose-response function* (ADRF) [10], denoted as  $\mathbb{E}(Y_t|Z_t, T_t = \mathcal{T})$ , to apply causal effect evaluation with multidimensional treatments. Previous research stressed that ADRF loss is bounded by the marginal loss comprising factual loss and counterfactual loss. As we cannot easily observe and define a counterfactual, we alternatively use a re-weighting schema to reduce confounder imbalance and calculate the upper bound of ADRF loss with Integral Probability Metrics (IPM) distance as our counterfactual loss (also an auxiliary loss) [36]. As IPM distance represents a group of metrics to measure distribution differences between re-weighted fact and counterfactual, optimizing the upper bound allows us to encompass most ADRF estimation methods used in previous research.

Based on the above formulations, we propose a causal effects calculation framework, as shown in part (b) of Figure 5, to evaluate the causal effect of public events on human mobility using LLMs-derived human intentions. In our framework, we process news articles with LLMs to extract structured information, including human intentions. We then calculate spatio-temporal regional confounders using observation data within a certain input window. After aggregating the characteristics of the public event at the predicted timestamp  $t$  as  $T_{i,t}$  for each region, we develop a causal inference module to calculate the reweighted factual loss and the upper bound of the counterfactual loss using the IPM distance. Finally, we output regional human mobility within an output window. More details will be provided in the following section.

### 4.3 Confounder Representation Learning by Spatio-Temporal Data Modeling

In Figure 5, to learn spatio-temporal regional confounders, we use ST-ResNET [45] based on Graph Convolution Networks (GCNs) to capture hidden states independent of treatments from historical observation data. First, we capture time-varying regional covariates using POI features and time information [19]. For each region, we collect POI data with  $c$  categories, represented as  $P_i \in \mathbb{R}^{n \times c}$ . After normalization, we fuse POI features with embedded time information  $X_t \in \mathbb{W}^{T \times c}$  within a fixed input window. Finally, we re-weight the static POI features with average time embedding and calculate the mean value as time-varying regional features:

$$\mathbf{X}_{i,t}^P = \bar{\mathbf{X}}_t \odot \mathbf{P}_i, \quad (2)$$

where  $\odot$  refers to element wise multiplication and  $\bar{\mathbf{X}}_t$  represents mean value of temporal embedding of input window.

Next, we concatenate multiple time-step ground truth observations  $\mathbf{X}_{i,t}^{flow}$  with other input variables before being passed into the GCN-based ST-ResNET, just as described below [9, 16]:

$$\hat{\mathbf{X}}_{i,t} = \mathbf{X}_{i,t}^{flow} \oplus \mathbf{X}_{i,t}^P \oplus \bar{\mathbf{X}}_t \oplus \mathbf{X}_i^{Reg}, \quad (3)$$

where  $\oplus$  refers to concatenation operations. As for  $f_{GCN}$ , we use a two-layer GCN to fuse and capture the hidden state as follows:

$$f_{GCN}(\mathbf{X}, \mathbf{A}) = \mathbf{A}[(\mathbf{A}\mathbf{X})\mathbf{W}_1]\mathbf{W}_2, \quad (4)$$

where  $\mathbf{W}_1$  and  $\mathbf{W}_2$  are learnable parameters of two GCN layers. Finally, the output of  $f_{GCN}$  will be added with input to build up a ST-ResNET [45] structure to get the confounder  $\mathbf{z}_{i,t}$ :

$$\mathbf{z}_{i,t} = f_{GCN}(\hat{\mathbf{X}}_{i,t}, A_t) + \hat{\mathbf{X}}_{i,t}. \quad (5)$$

#### 4.4 Human Mobility Prediction Enhanced by Causality Modeling

With the confounder  $\mathbf{z}_{i,t}$ , the model starts calculating the factual loss and counterfactual loss. We encode the treatment  $T_{i,t}$  with human intentions  $N_{i,t}$  in structured information. Considering that the impact of events diminishes over time, we chose GRU to process these features and use the output  $T_{i,t}$  as representations of treatments. As mentioned in Section 4.2, we need to calculate re-weighted factual loss and use the upper bound of IPM distance as counterfactual loss. We use two multilayer perceptron (MLP) layers, denoted as the inference network  $f_{inf}$  and the re-weighting network  $f_{rwt}$ , to realize the causal inference process:

$$w_{i,t} = f_{rwt}(\hat{\mathbf{z}}_{i,t}, T_{i,t}), \quad (6)$$

$$\hat{y}_{i,t} = f_{inf}(\hat{\mathbf{z}}_{i,t}, T_{i,t}), \quad (7)$$

where  $w_i$  is the weight to remove regional selection bias of treatments, and  $\hat{y}_i$  refers to predicted outcomes of future human mobility. Then, the factual loss can be represented as:

$$\mathcal{L}_f = \frac{1}{n} \sum_{i=1}^n \left[ (w_{i,t} \cdot (\hat{y}_{i,t} - y_{i,t})^2) \right]. \quad (8)$$

On the other hand, as we aim to calculate IPM distance, we need to adjust the representations of treatments and confounders to ensure their original distributions and conditional distributions, denoted as  $\mathbf{z}$  and  $\mathbf{z}^T$ , are independent of each other [25]. To calculate reasonable IPM distance, we create a manually created baseline (i.e., all zero scores for human intentions) and input it into the same GRU modules to get a baseline treatment  $T_b$ , simulating regular patterns without any events. We then calculate cosine similarities between the baseline treatment and all other treatments, sorting all treatments according to the results. We use  $p^j(z)$  and  $p^j(z|T)$  to represent original distributions and conditional distributions for the  $j$ -th similarity interval after dividing  $[-1, 1]$  into  $J$  intervals. Therefore, we can calculate the upper bound of IPM distance with maximum inter-interval as counterfactual loss:

$$\mathcal{L}_{cf} = \frac{1}{n} \sum_{i=1}^J \max_{j \in \{1, 2, \dots, J\}, j \neq i} \left\{ IPM_{\mathcal{G}}(p^j(z), p^j(z|T)) \right\}, \quad (9)$$

where  $IPM_{\mathcal{G}}$  refers to the Maximum Mean Discrepancy (MMD) metric [5], which is used to measure the difference between two representation distributions. Based on the above content, our final prediction target and loss function for model training is:

$$\mathcal{L} = \mathcal{L}_f + \alpha \mathcal{L}_{cf}, \quad (10)$$

where  $\alpha$  controls the weight of auxiliary counterfactual loss  $\mathcal{L}_{cf}$ .

## 5 Experiment

In this section, we conduct experiments to introduce different analyses of our model based on Figure 5. In Section 5.2, we show overall performance by comparing with baseline models and present the

results of the ablation study. In Section 5.3, we provide distributions of LLMs' output for human intentions toward public events in news articles. In Sections 5.5 and 5.6, we provide analyses of average treatment effects estimation and the causal response of human mobility to two public events as counterfactuals. As a case study in Section 5.7, we show prediction results in our test process comparing predictions without human intentions.

### 5.1 Experimental Settings

Our model is developed using PyTorch 1.8.2 and Python 3.8, and is trained on a server equipped with four NVIDIA RTX A6000 GPUs (48GB memory) running the Ubuntu 20.04 operating system. The raw GPS data is interpolated at 15-minute intervals and aggregated into one-hour intervals. For training, we randomly select 80% of the data for training, 10% for validation, and 10% for testing. For POI data categorization, we aggregate the data into 17 categories based on their labels. The input and output windows for the experiments are configured differently, as discussed in the following section. We use the Adam optimizer with an initial learning rate of 0.001 and select the model with the best validation performance using an early stopping strategy before testing. The weight  $\alpha$  is set to 1, and we repeat our experiments five times with fixed random seeds. For LLMs, we used LLama3-70b with Ollama, benefiting from its optimizations. This configuration allows the LLM to run with approximately 44GB of GPU memory usage, evenly distributed across the four graphics cards.

### 5.2 Overall Performance

To evaluate the prediction performance of our proposed model, we select several state-of-the-art models in human mobility prediction. It's important to note that it is difficult to implement the event-based treatments into traditional causal-based models, as they are not one-dimensional and cannot be easily classified into binary treatments as in previous studies. Therefore, we mainly focus on the following human mobility prediction models:

**3DGCN** [42]: A crowd flow prediction model based on a three-dimensional GCN to capture spatio-temporal dependencies.

**GTS** [26]: A time-series prediction model that explores correlation and causation among variables based on GNNs by learning a probabilistic graph by optimizing performance over graph distribution.

**PromptST** [46]: A spatio-temporal prediction model combining transformers and spatio-temporal prompts to balance common knowledge learning with efficient adaptation to specific tasks.

**STGNN** [37]: A traffic flow prediction model based on GCN, GRU, and transformer structure to learn the latent patterns and capture spatial dependency in traffic nodes of transportation networks.

**DeepST** [45]: A crowd flow prediction model based on ST-ResNET by simulating close, period and trend properties of crowd traffic and further combined with external factors during prediction.

**DeepSTN+** [19]: A crowd flow prediction model by modeling the long-range spatial dependence among crowd flows in different regions and fusing POI semantic features simultaneously.

**STID** [27]: A simple MLP-based baseline model for multivariate time series forecasting using spatial and temporal identity embeddings to improve efficiency and accuracy.

**Table 1: Prediction Performance Comparison of Human Mobility Prediction Models with multiple experiments (Mean/Standard deviation). The best results are bold and the best baselines are marked with underline, where we calculate the improvements. Ablation studies are marked with  $\dagger$ . The time interval for each term is one hour.**

Models	Short Term (In 6, Out 1)			Medium Term (In 12, Out 6)			Long Term (In 24, Out 24)		
	RMSE	MAE	MAPE(%)	RMSE	MAE	MAPE(%)	RMSE	MAE	MAPE(%)
3DGCN	91.50/0.47	38.57/0.39	1.72/0.03	275.79/5.47	119.97/2.56	5.81/0.07	401.63/11.41	175.76/2.72	7.98/0.20
GTS	88.10/10.84	39.41/3.61	2.02/0.18	344.68/29.99	181.20/6.52	8.87/0.12	427.31/29.27	208.65/4.77	9.60/0.06
PromptST	64.29/1.25	26.72/0.39	1.65/0.02	105.87/2.20	47.51/0.51	2.60/0.02	233.92/21.12	81.98/4.60	3.98/0.17
STGNN	67.43/0.67	31.81/0.21	1.87/0.02	299.30/5.54	156.54/0.97	8.04/0.04	383.93/6.68	184.28/3.59	8.95/0.15
DeepST	55.11/1.25	28.91/0.56	1.53/0.03	101.75/1.80	54.88/0.79	2.80/0.05	163.80/1.47	81.44/1.01	3.95/0.06
DeepSTN+	39.45/0.45	22.25/0.19	1.21/0.01	87.57/2.26	48.28/1.24	2.51/0.07	145.92/1.32	70.07/1.15	3.43/0.05
STID	<u>37.19/0.96</u>	<u>20.76/0.50</u>	<u>1.10/0.02</u>	<u>73.15/1.63</u>	<u>38.42/0.62</u>	<u>1.93/0.02</u>	<u>133.80/3.15</u>	<u>54.76/1.73</u>	<u>2.76/0.12</u>
CIDER (Naïve)	56.95/2.66	24.92/0.68	1.42/0.03	79.89/1.94	36.99/0.43	2.06/0.02	139.40/1.72	54.98/0.51	2.85/0.03
CIDER (wo $\mathcal{L}_{cf}$ )	57.26/3.43	24.69/0.78	1.40/0.04	80.65/2.34	36.88/0.44	2.03/0.01	141.14/5.78	55.48/1.93	2.85/0.12
CIDER	50.56/6.13	23.15/1.75	1.27/0.09	77.62/1.39	<u>35.72/0.58</u>	<u>1.86/0.05</u>	138.53/11.93	54.85/3.67	2.72/0.14
Ours (Naïve) $\dagger$	35.87/1.05	18.57/0.30	1.00/0.01	72.68/0.76	34.47/0.37	1.76/0.03	132.28/2.82	52.96/0.78	2.58/0.05
Ours (wo $\mathcal{LLM}$ ) $\dagger$	35.14/0.94	18.56/0.29	1.00/0.02	72.96/3.05	34.29/1.03	1.73/0.04	130.34/2.77	<b>52.70/0.88</b>	2.62/0.06
Ours (wo $\mathcal{R}$ ) $\dagger$	35.17/0.88	18.75/0.34	1.00/0.02	72.01/2.67	34.77/0.87	1.75/0.04	130.47/2.35	53.07/1.34	<b>2.56/0.07</b>
Ours (wo $\mathcal{L}_{cf}$ ) $\dagger$	33.72/0.82	18.02/0.22	<b>0.99/0.01</b>	72.74/2.15	34.45/0.48	1.77/0.02	137.50/1.65	55.54/0.65	2.74/0.05
Ours	<b>33.31/0.53</b>	<b>18.01/0.29</b>	<b>0.99/0.01</b>	<b>68.91/3.34</b>	<b>33.12/0.91</b>	<b>1.70/0.04</b>	<b>128.92/6.95</b>	54.42/1.24	2.74/0.07
Improvements	10.43%	13.25%	10.00%	5.80%	7.28%	9.41%	3.65%	3.76%	7.25%

**CIDER** [20]: A causal effect estimation framework for COVID-19 policy in the US based on GRU and GCN. Also, Zhang [48] used the same backbone model for assessing causal effects of typhoons.

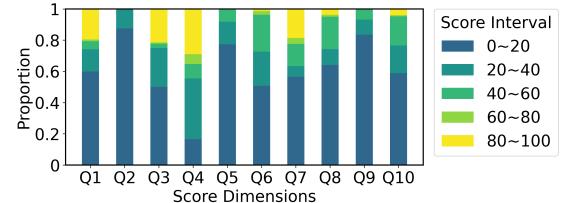
For prediction settings, we assume that public events have different potential effects at multiple time scales, so we design three tasks with different input (In) and output window (Out). We choose Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) as metrics for comparison.

Table 1 illustrates the prediction performance of baseline models and our proposed CausalMob. Overall, our model achieves the best performance in three prediction tasks with repeated experiments using different random seed initializations (see Appendix B). As the prediction window increases, there is a slight decrease in robustness in standard deviation, but this is acceptable given the excellent prediction accuracy. Specifically, for prediction performances of long-term prediction tasks, although the results show the weakness of our model in both the MAE and MAPE metrics, we believe that the strength of the RMSE is more representative of the outperforming of our model since we use weighted MSE, denoted as  $\mathcal{L}_f$  in our target function. Additionally, compared with the CIDER which has a different backbone, the results show that our method improves the prediction performances by effectively learning confounders.

**Ablation study** is shown in Table 1 in the last five lines. We evaluate the performances of the naïve model (Naïve) with simply output  $\hat{y}$ , the model that replaces all LLMs-derived human intentions with all-zero scores (wo  $\mathcal{LLM}$ ), the model without reweighting mechanism (wo  $\mathcal{R}$ ), the model without counterfactual (wo  $\mathcal{L}_{cf}$ ) loss. We prove that 1) combining public event-based causal inference for human mobility prediction improves prediction performance, and 2) a loss function based on counterfactual inference can further optimize this performance. Similarly, we conduct this ablation

study with the CIDER backbone based on Naïve and wo  $\mathcal{L}_{cf}$ , and the results also support the two conclusions we summarized.

### 5.3 Analysis of LLMs-derived Human Intentions

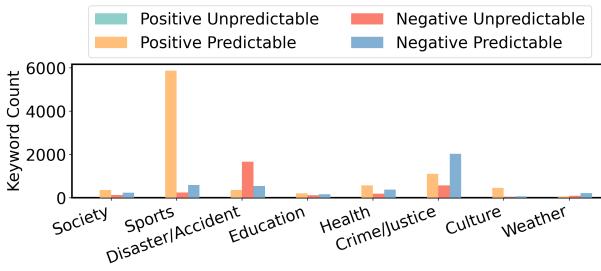


**Figure 6: Distribution of LLMs-derived human intentions based on scores with prompt engineering. Q1-Q10 refer to 10 questions listed in Prompt 5 of Appendix E.**

In this section, we analyze LLMs-derived human intentions extracted from structured information. In the previous sections, we set multiple scores to evaluate news articles with Prompt 5 in Appendix E. In Figure 6, we show the distributions of scores for all news articles. The score of Q1 describes the degree of danger of public events, with almost 60% having no negative impact on human mobility. The third score for Q3 refers to the level of interest, with almost half of the events not attracting people to certain areas. The concepts of danger and interest are not mutually exclusive, allowing an event to receive non-zero scores in both dimensions simultaneously. The remaining scores represent different aspects such as staying in Q2, impact on daily life in Q4, impact on business activities in Q5, transportation in Q6, public health in Q7, government intervention in Q8, and public services in Q9. The last score,

Q10, indicates the duration of the public events. With Prompt 4 in Appendix E, we expect LLMs to judge the predictability of public events. Combined with scores in *danger* and *interest*, we categorize news articles into four quadrants based on public events categorization in Figure 1. Finally, public events in news articles labeled with "Yes" are treated as unpredictable cases.

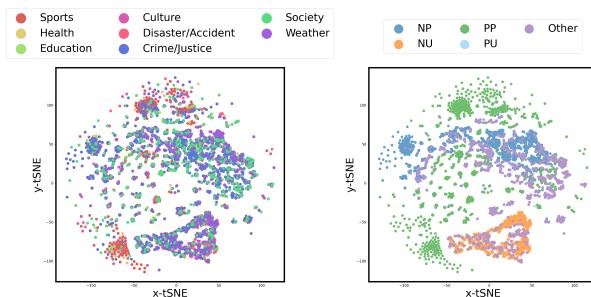
In Figure 7, we show the relationships between categorizations



**Figure 7: Count of news articles labels with public events categorization based on LLMs-derived human intentions**

based on LLM-derived intentions and labels that the original articles carried. We find that positive unpredictable public events are difficult to extract from our dataset, which is consistent with our considerations for not considering them. Second, positive predictable events comprise the majority of our dataset, especially Sports, Crime/Justice and Cultural related events. Negative events are less represented but dominate the category of Disaster/Accident and are mostly unpredictable. Additionally, negative predictable events are also found in the Crime/Justice. Overall, LLMs can effectively summarize human intentions toward public events from news articles, which is helpful in the analysis of causal inference.

#### 5.4 Analysis of Encoded Treatment Interpretability



**Figure 8: Visualization of encoded treatments with t-SNE.** Two figures share the same features, the left figure labeled with news articles types and the right figure is categorized based on human intentions (1st letter N/P: Negative/Positive, 2nd letter P/U:Predictable/Unpredictable).

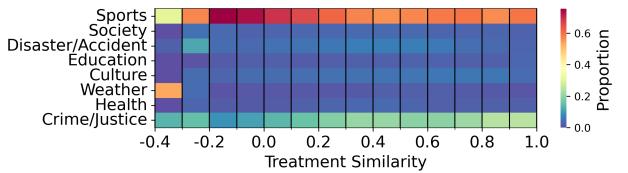
In this section, we present the results of encoded human intentions. In our CausalMob, human intentions toward public events

in the input and output window are considered simultaneously. All intentions in the input window are included, but only those of predictable events are considered and put into the encoder during training. Here, we choose the encoder in the long-term prediction task to process features obtained from LLMs. In Figure 8, we visualize encoded features of different public events using t-SNE [32].

In Figure 8, human intentions features related to Sports are generally distinguished from events with other labels, and the corresponding points in the right figure are categorized into positive predictable events. In contrast, Crime/Justice, Disaster/Accident events have similar distributions. Besides, those corresponding points are always classified into negative events in the right figure. The above conclusions are consistent with the information revealed in Figure 7, which also demonstrates the ability of the treatment encoder to identify and learn the characteristics of human intentions.

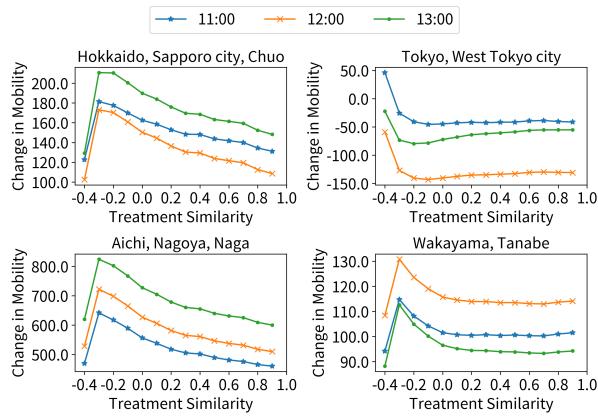
#### 5.5 Analysis of Regional Average Treatment Effects

In this section, we analyze the average treatment effects of different public events under the same spatio-temporal conditions. We collect all treatment representations of human intentions during the training process and sorted them according to their cosine similarities to the treatment baseline [28]. We use 0.1 as the interval to split similarities, resulting in 14 intervals of treatment similarities. In Figure 9, we visualize the distributions of the original news labels in each interval, and we can see that public events related to Sport and Crime/Justice occupy the majority of all treatments intervals. In the most dissimilar treatment interval of  $[-0.4, -0.3]$ , Weather shares a higher proportion, which is quite different from other intervals.



**Figure 9: Distribution of news articles labels based on treatment similarity compared with treatment baseline.**

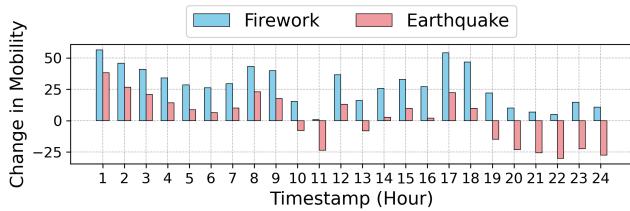
Then, we choose a specific weekday in the test dataset to evaluate the performance with different treatments. In Figure 10, we present the prediction results in selected regions along with the treatment similarity intervals. When treatment is in the most dissimilar interval, as mentioned in Section 5.5, the results show a significant decline in predicted human mobility. As treatment similarity increases towards 1, the change in human mobility gradually decreases (in West Tokyo City, it slightly increases to 0). As we show a similar distribution that constantly changes in Figure 9 from -0.3 to 1.0, we also observe a continuous change in human mobility in Figure 10. This demonstrates that our model can control prediction results according to the similarity between encoded treatments of public events and treatment baseline (regular patterns).



**Figure 10: Regional average treatment effects of human mobility in multiple prediction timestamps. We selected several regions and predicted average changes of human mobility with treatments in each similarity interval.**

## 5.6 Analysis of Causal Response with Counterfactual

In this section, we provide an analysis of the causal response of human mobility to different public events as counterfactuals. We select news articles about a firework and an earthquake, then extract structured information to get human intentions as treatments and input them into the long-term prediction process as a predictable event and an unpredictable one, respectively, for one region (Aichi, Matsuyama). In Figure 11, the event feature of the firework festival presents a positive effect on human mobility at all prediction timestamps. On the contrary, the feature of the earthquake tends to control the increase of human mobility compared to the firework, with our results presenting negative effects on human mobility at several timestamps in the mid-noon and night. This demonstrates that our proposed model can simulate the positive and negative effects of different public events on human mobility prediction.



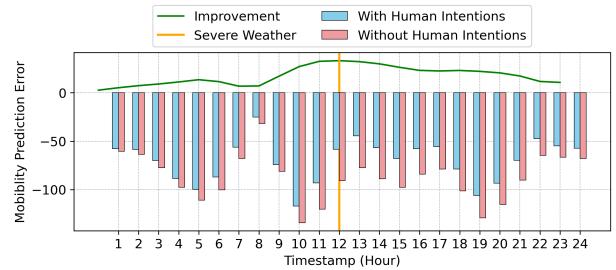
**Figure 11: Causal Responses with Counterfactual. We show causal responses of human mobility to two public events.**

## 5.7 Case Study of Prediction with Human Intentions

In this section, we provide a case study in test dataset to show the effect of our model. In Figure 12, a predictable public event of severe

weather is reported on 2024-03-12 08:25:00, and in the structured information processed with LLMs, the event time is reset to noon the next day. We use LLMs-derived human intentions, as shown in the title of Figure 12, to predict human mobility on the day following the 12th. For comparison, we use all-zero intentions to mask this event and also predict future human mobility. Then, we calculate the prediction errors for each timestamp and the improvement of our model with the green line, along with the occurrence of the severe weather at noon, indicated by the orange line.

In Figure 12, we can see that the prediction results are similar in



**Figure 12: Human mobility prediction errors with and without human intentions. The title of the news is as follows: Strong winds and heavy rain expected on the Pacific side. Unstable atmospheric conditions due to low pressure system. Human Intentions: [80, 20, 0, 90, 40, 60, 70, 80, 30, 60]**

the morning hours of the day. However, as the predictable severe weather approaches, our model starts to show more accurate predictions, which is evident from the change in the Improvement line. In this case, our model achieved the best results around the predictable public event (i.e., 12:00 p.m.). The improved performance of predictions with and without the causal influence of human intentions decreases gradually after that point. Overall, our model consistently maintained more accurate predictions. This result shows that our model can use structured information and human intentions from news articles to enhance human mobility prediction during predictable non-periodical public events.

## 6 Conclusion

In this research, we predict human mobility based on LLM-derived human intentions toward public events using a causal inference framework and evaluate the response of human mobility to different public events. We utilize news articles to extract structured information about public events and generate human intentions with LLMs through prompt engineering. We propose a prediction model called **CausalMob** to predict human mobility and estimate the causal effect of public events using multiple data sources (including POI data and human mobility data from GPS trajectories) in the main administrative regions of Japan. Compared with several state-of-the-art models, our proposed model performs better on multiple prediction tasks. Furthermore, our approach provides interpretable results for investigating the relationship between human mobility and public events. Our findings can help decision-makers estimate mobility changes in different regions and make informed decisions in advance when faced with various public events.

## References

- [1] Blogwatcher. 2024. Blogwatcher, Inc. <https://www.blogwatcher.co.jp/> (Japanese Only).
- [2] Quanjun Chen, Renhe Jiang, Chuang Yang, Zekun Cai, Zipei Fan, Kota Tsubouchi, Ryosuke Shibasaki, and Xuan Song. 2020. Dualsim: Dual sequential interaction network for human intentional mobility prediction. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems* 283–292.
- [3] Zipei Fan, Xuan Song, Tianqi Xia, Renhe Jiang, Ryosuke Shibasaki, and Ritsu Sakuramachi. 2018. Online deep ensemble learning for predicting citywide human mobility. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–21.
- [4] Ruth E Farmer, Daphne Kounali, A Sarah Walker, Jelena Savović, Alison Richards, Margaret T May, and Deborah Ford. 2018. Application of causal inference methods in the analyses of randomised controlled trials: a systematic review. *Trials* 19 (2018), 1–14.
- [5] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The Journal of Machine Learning Research* 13, 1 (2012), 723–773.
- [6] Ge Hangli, Lifeng Lin, Renhe Jiang, Takashi Michikata, and Noboru Koshizuka. 2022. Multi-Weighted Graphs Learning for Passenger Count Prediction on Railway Network. In *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, 374–382.
- [7] Md Rakibul Hasan, Md Zakir Hossain, Tom Gedeon, and Shafin Rahman. 2024. LLM-GEm: Large language model-guided prediction of people's empathy levels towards newspaper article. In *Findings of the Association for Computational Linguistics: EACL 2024*, 2215–2231.
- [8] Samiul Hasan, Christian M Schneider, Satish V Ukkusuri, and Marta C González. 2013. Spatiotemporal patterns of urban human mobility. *Journal of Statistical Physics* 151 (2013), 304–318.
- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- [10] Guido W Imbens. 2000. The role of the propensity score in estimating dose-response functions. *Biometrika* 87, 3 (2000), 706–710.
- [11] Jinzhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Structgpt: A general framework for large language model to reason over structured data. *arXiv preprint arXiv:2305.09645* (2023).
- [12] Renhe Jiang, Quanjun Chen, Zekun Cai, Zipei Fan, Xuan Song, Kota Tsubouchi, and Ryosuke Shibasaki. 2022. Will you go where you search? A deep learning framework for estimating user search-and-go behavior. *Neurocomputing* 472 (2022), 338–348.
- [13] Renhe Jiang, Xuan Song, Dou Huang, Xiaoya Song, Tianqi Xia, Zekun Cai, Zhao-nan Wang, Kyoung-Sook Kim, and Ryosuke Shibasaki. 2019. Deepurbanevent: A system for predicting citywide crowd dynamics at big events. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2114–2122.
- [14] Renhe Jiang, Zhao-nan Wang, Yudong Tao, Chuang Yang, Xuan Song, Ryosuke Shibasaki, Shu-Ching Chen, and Mei-Ling Shyu. 2023. Learning social meta-knowledge for nowcasting human mobility in disaster. In *Proceedings of the ACM Web Conference 2023*, 2655–2665.
- [15] Hanlei Jin, Yang Zhang, Dan Meng, Jun Wang, and Jinghua Tan. 2024. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv preprint arXiv:2403.02901* (2024).
- [16] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
- [17] Shuqi Li, Yuebo Sun, Yuxin Lin, Xin Gao, Shuo Shang, and Rui Yan. 2024. Causal-Stock: Deep End-to-end Causal Discovery for News-driven Stock Movement Prediction. *arXiv preprint arXiv:2411.06391* (2024).
- [18] Yuan Li, Yixuan Zhang, and Lichao Sun. 2023. Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents. *arXiv preprint arXiv:2310.06500* (2023).
- [19] Ziqian Lin, Jie Feng, Ziyang Lu, Yong Li, and Depeng Jin. 2019. Deepstn+: Context-aware spatial-temporal neural network for crowd flow prediction in metropolis. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33, 1020–1027.
- [20] Jing Ma, Yushun Dong, Zheng Huang, Daniel Mietchen, and Jundong Li. 2022. Assessing the causal impact of COVID-19 related policies on outbreak dynamics: A case study in the US. In *Proceedings of the ACM Web Conference 2022*, 2678–2686.
- [21] Stephen L Morgan and Christopher Winship. 2015. *Counterfactuals and causal inference*. Cambridge University Press.
- [22] ollama. 2024. ollama. <https://github.com/ollama/ollama>
- [23] Judea Pearl. 2009. Causal inference in statistics: An overview. (2009).
- [24] Xie Runfeng, Cui Xiangyang, Yan Zhou, Wang Xin, Xuan Zhanwei, Zhang Kai, et al. 2023. Lkpn: Llm and kg for personalized news recommendation framework. *arXiv preprint arXiv:2308.12028* (2023).
- [25] Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M Buhmann, and Walter Karlen. 2020. Learning counterfactual representations for estimating individual dose-response curves. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, 5612–5619.
- [26] Chao Shang, Jie Chen, and Jinbo Bi. 2021. Discrete graph structure learning for forecasting multiple time series. *arXiv preprint arXiv:2101.06861* (2021).
- [27] Zezhi Shao, Zhao Zhang, Fei Wang, Wei Wei, and Yongjun Xu. 2022. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 4454–4458.
- [28] Ritika Singh and Satwinder Singh. 2021. Text similarity measures in news articles by vector space model using NLP. *Journal of The Institution of Engineers (India): Series B* 102 (2021), 329–338.
- [29] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. Limits of predictability in human mobility. *Science* 327, 5968 (2010), 1018–1021.
- [30] Xuan Song, Quanshi Zhang, Yoshihide Sekimoto, Ryosuke Shibasaki, Nicholas Jing Yuan, and Xing Xie. 2016. Prediction and simulation of human mobility following natural disasters. *ACM Transactions on Intelligent Systems and Technology (TIST)* 8, 2 (2016), 1–23.
- [31] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llma: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [32] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
- [33] Aishwarya Vijayan. 2023. A prompt engineering approach for structured data extraction from unstructured text using conversational LLMs. In *Proceedings of the 2023 6th International Conference on Algorithms, Computing and Artificial Intelligence*, 183–189.
- [34] Xinglei Wang, Meng Fang, Zichao Zeng, and Tao Cheng. 2023. Where would i go next? large language models as human mobility predictors. *arXiv preprint arXiv:2308.15197* (2023).
- [35] Xinlei Wang, Maike Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. 2024. From news to forecast: Integrating event analysis in llm-based time series forecasting with reflection. *arXiv preprint arXiv:2409.17515* (2024).
- [36] Xin Wang, Shengfei Lyu, Xingyu Wu, Tianhao Wu, and Huanhuan Chen. 2022. Generalization bounds for estimating causal effects of continuous treatments. *Advances in Neural Information Processing Systems* 35 (2022), 8605–8617.
- [37] Xiaoyang Wang, Yao Ma, Yiqi Wang, Wei Jin, Xin Wang, Jiliang Tang, Caiyan Jia, and Jian Yu. 2020. Traffic flow prediction via spatial temporal graph neural network. In *Proceedings of the web conference 2020*, 1082–1092.
- [38] Yingzi Wang, Nicholas Jing Yuan, Defu Lian, Linli Xu, Xing Xie, Enhong Chen, and Yong Rui. 2015. Regularity and conformity: Location prediction using heterogeneous mobility data. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 1275–1284.
- [39] Zhaoan Wang, Renhe Jiang, Hao Xue, Flora D Salim, Xuan Song, and Ryosuke Shibasaki. 2022. Event-aware multimodal mobility nowcasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, 4228–4236.
- [40] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.
- [41] Wikipedia contributors. 2023. Kyodo News — Wikipedia, The Free Encyclopedia. [https://en.wikipedia.org/w/index.php?title=Kyodo\\_News&oldid=1187452593](https://en.wikipedia.org/w/index.php?title=Kyodo_News&oldid=1187452593) [Online; accessed 8-July-2024].
- [42] Tong Xia, Junjie Lin, Yong Li, Jie Feng, Pan Hui, Funing Sun, Diansheng Guo, and Depeng Jin. 2021. 3dgcnn: 3-dimensional dynamic graph convolutional network for citywide crowd flow prediction. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15, 6 (2021), 1–21.
- [43] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2021. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15, 5 (2021), 1–46.
- [44] Bing Yu, Haoteng Yin, and Zhanxing Zhu. 2017. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875* (2017).
- [45] Junbo Zhang, Yu Zheng, and Dekang Qi. 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.
- [46] Zijian Zhang, Xiangyu Zhao, Qidong Liu, Chunxu Zhang, Qian Ma, Wanyu Wang, Hongwei Zhao, Yiqi Wang, and Zitao Liu. 2023. Promptst: Prompt-enhanced spatio-temporal multi-attribute prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 3195–3205.
- [47] Kai Zhao, Sasu Tarkoma, Siyuan Liu, and Huy Vo. 2016. Urban human mobility data mining: An overview. In *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 1911–1920.
- [48] Zhang Zhiwen, Hongjun Wang, Zipei Fan, Ryosuke Shibasaki, and Xuan Song. 2023. Assessing the Continuous Causal Responses of Typhoon-related Weather on Human Mobility: An Empirical Study in Japan. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 3524–3533.

**Table 2: Notation used in methodology**

$\mathbf{P}_i$	POI distributions in $i$ -th selected region, we categorized POI into 17 types
$\mathbf{X}_{i,t}^{flow}$	Historical human mobility information of $i$ -th region at $t$ -th time step within input window
$\mathbf{X}_t$	Temporal information with a input window
$A_t$	Normalized human mobility network among selected regions
$\mathbf{z}_{i,t}$	The representations of learned spatio-temporal confounders of $i$ -th region at $t$ -th time step
$T_{i,t}$	The encoded treatment variables of $i$ -th region at $t$ -th time step, fusing all historical events in input window and predictable ones in the output window
$w_{i,t}$	The output of re-weighting network the balance the selection bias of $i$ -th region at $t$ -th time step based on $\mathbf{z}$ and $T$
$N_{i,t}$	Human intentions toward public events in $i$ -th region at $t$ -th time step, all-zero scores will be used in there is no event

## A Table for Notations

In this section, we will provide a table of notation used in our methodology in Table 2.

## B Hyper-Parameter Settings

In this section, we provide other hyper-parameters for model training that we did not mention in the main content because of space limitation.

**Table 3: Additional hyper-parameter information**

learning rate decay rate	0.5
early stop threshold	9e-6
patience step	5
hidden state dimension	64
maximum training epoch	300
learning rate scheduler metric	MSE
random seeds	[1111, 2222, 3333, 4444, 5555]
batch size	24

## C News Articles with LLMs Generated Answers

In this section, we provide some cases of news articles and their structured information generated with LLMs. Due to copyright and other policies, we cannot share original articles, but we can provide links to the corresponding news articles (English version, not original Japanese news article).

### Typhoon Khanun:

**Answer 1:** I've identified the most influential events in the news content: Typhoon No. 6 is expected to approach Okinawa on August 3, (predicted closest approach).

**Answer 2:** 2023-08-03 06:00:00

**Answer 3:** Here are my evaluations:

Where: People in Okinawa and Amami regions may need to move to safer areas or evacuation centers, away from coastal areas prone to storm surges and flooding.

Who: Residents, commuters, and tourists in Okinawa and Amami regions, especially those living in low-lying areas or near the coast, will be affected by the typhoon.

When: The event is ongoing, with the typhoon expected to approach Okinawa on August 3. (Note: The news release time was August 2, 07:17:28)

How: If necessary, people should move to safer areas or evacuation centers using safe routes and modes of transportation, such as official evacuation buses or private vehicles, while following instructions from local authorities and staying informed about weather updates.

**Answer 4:** No.

**Answer 5:** [80, 20, 0, 0, 40, 70, 80, 60, 30, 50]

### Sumidagawa Firework:

**Answer 1:** The most influential event identified is the resumption of the "Sumida River Fireworks Festival" after a 4-year hiatus due to the COVID-19 pandemic. This event will take place on July 29th and is expected to attract approximately 950,000 people.

**Answer 2:** 2023-07-29 19:07:04

**Answer 3:** Here are my evaluations:

Where: If necessary, people may need to move away from the Sumida River area in Tokyo to avoid congestion and large crowds.

Who: People who plan to travel or move around in the Sumida River area in Tokyo, particularly those who live or work nearby, tourists, and festival-goers will be affected by the event.

When: The event happened on July 29th (the day of the news release).

How: If necessary, people should consider using public transportation, rescheduling their travel plans, and planning alternative routes to avoid congested roads and traffic jams.

**Answer 4:** No.

**Answer 5:** [0, 20, 85, 20, 0, 0, 0, 0, 0, 60]

## D About publicity of source code and data

We have already open-sourced the code for this research on GitHub, which can be accessed at [this github repository](#). However, due to data privacy policies, we cannot share human mobility data and the original text-based news articles (though the news articles are publicly available). To ensure that our source code can be implemented smoothly, we synthesized a portion of the data with additional noise, enabling our source code to be run without issues.

## E Prompts for LLMs

In this section, we list all the prompts used to extract structured information and human intentions from news articles. Specifically, we first input the release time of the news (to assist LLMs in inferring the timing of public events), along with the title and content of the news. We then engage the LLMs in a dialogue-based chain-of-thought paradigm by sequentially asking the following questions. This process ultimately generates human intentions, which serve as treatment variables in our causal inference framework.

**Prompt 1: Public Events**

*You are an AI assistant that notifies affected users and makes suggestions to change their mobility based on news text. First, identify the most influential events in the news content, including scheduled and unpredictable events.*

**Prompt 2: Time Information**

*Next, estimate the exact time of the most important event mentioned in the news in the following JSON format: "event time": "yyyy-mm-dd hh:mm:ss". If the exact time is unknown, use the news release time. Provide only the JSON string without any additional text or explanations.*

**Prompt 3: 3W1H about Human Mobility**

*Based on the news text and our chat, evaluate the 3W1H related to human mobility.*

- Where: Where should people move if necessary?
- Who: What kind of people will be affected?
- When: When did the event happen?
- How: How should people move if necessary?

**Prompt 4: Predictability**

*Is the content in the news more like an unpredictable event, such as an earthquake? Your answer can only be "Yes" or "No".*

**Prompt 5: Human Intentions**

*Background: Most news related to economic, politics, culture, and history issues usually have no effect on human mobility because they do not relate to people's daily life.*

*- Slight disasters, such as light earthquakes and tsunamis, or some local events (like political events) may also have no effect on human mobility in Japan.*

*- Only events that happened close to the release time (within several hours) may have an influence on human mobility.*

*Task: Score the news text based on the following aspects (0-100, where a higher number means higher agreement):*

*Q1. To what extent do the events described in the news make people leave the area because they are dangerous?*

*Q2. To what extent do the events described in the news make people stay in the area because it is better not to move?*

*Q3. To what extent do the events described in the news make people visit the area because they are interesting events?*

*Q4. To what extent do the events described in the news make people keep their daily routine as these events are not important to daily life?*

*Q5. To what extent do the events described in the news lead to interruption of economic activities, such as business closures or work stoppages?*

*Q6. To what extent do the events described in the news affect transportation conditions, such as traffic congestion or road closures?*

*Q7. To what extent do the events described in the news impact public health and safety, leading to decisions to leave or avoid certain areas?*

*Q8. To what extent do the events described in the news involve government or official instructions that influence people's movements?*

*Q9. To what extent do the events described in the news affect the availability of public services, such as school closures or interruptions in medical services?*

*Q10. To what extent do the events described in the news last a long time (like one day)?*

*Expected response: A list of 10 numbers between 0 and 100.*