Towards Robust Travel Time Estimation: An Out-of-Distribution Generalization Approach

Xiwen Jiang
Renmin University of China
Beijing, China
jiangxw@live.cn

Chuan Zhou
Peking University
Beijing, China
zhouchuancn@pku.edu.cn

Xiaofeng Meng*
Renmin University of China
Beijing, China
xfmeng@ruc.edu.cn

Haoxuan Li*
Peking University
Beijing, China
hxli@stu.pku.edu.cn

Abstract

Travel is becoming increasingly convenient with the development of the Internet. Travel Time Estimation (TTE) serves as a fundamental task for online traffic services. However, it faces a pressing challenge due to the volatile nature of traffic: the out-of-distribution (OOD) problem. In this paper, we investigate the OOD generalization problem, with a specific focus on the TTE task. We analyze the underlying generative process of traffic data by constructing a relational Structural Causal Model (SCM). We reveal that the complex causal relations can be simplified through a technique we term selective blocking. Based on this simplification, we propose a two-step deconfounding procedure to eliminate the spurious correlation of the environment on the invariant representation. Specifically, we design an invariant representation model, **OOD** model for **T**ravel **T**ime **E**stimation (OOD4TTE). First, our model infers potential environments for data generation through an environmental inference module. Then, we implement the two-step deconfounding procedure by contrastive learning and an invariant gating network. Finally, we generate robust travel time estimates based on the invariant representations. We demonstrate the superior generalization capability of OOD4TTE through comprehensive experiments on two real-world datasets.

CCS Concepts

• **Information systems** → **Location based services**; **Information systems applications**; • **Applied computing** → *Transportation*; • **Computing methodologies** → *Neural networks*; Supervised learning by regression; Learning under covariate shift.

Keywords

Travel Time Estimation, OOD Generalization, Traffic Data Management

*Corresponding Authors.

KDD 2026, Jeju Island, Republic of Korea.

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3817802

ACM Reference Format:
Xiwen Jiang, Chuan Zhou, Xiaofeng Meng, and Haoxuan Li. 2026. Towards Robust Travel Time Estimation: An Out-of-Distribution Generalization Approach. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD 2026), August 9–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3770855.3817802

1 Introduction

Travel Time Estimation (TTE), as a core service of location-based services (LBS), refers to the task of predicting how long it will take to travel from a given origin to a destination along a specific route, using available traffic context such as departure time, weather, etc. For instance, in Fig. 1, traveling on the red route on a rainy Saturday at 4:13 p.m. is estimated to cost 33 minutes. The TTE task plays a vital role in various applications, including trip planning [12], transportation scheduling [33], and delivery services [26].

Despite extensive efforts in prior work [13, 14, 22, 34, 36], TTE models still suffer from poor out-of-distribution (OOD) generalization. As shown in Fig. 2, OOD primarily originates from covariate shifts induced by discrepancies in data generation environments between training and testing, while shortcut features further exacerbate this issue by failing to generalize to unseen environments. Most existing OOD generalization approaches [2, 3, 15, 37] assume access to **environment** labels—i.e., prior knowledge of the data generation mechanisms—and learn invariant representations across these environments. Unfortunately, such environment annotations are typically unavailable in real-world TTE scenarios due to the complexity of the underlying data generation process.

To address the OOD generalization problem in the TTE task, we face two *key challenges*: ❶ the data generation process is highly complex, leading to rich and intertwined spurious relations in the observed data; ❷ the notion of environment is implicit and difficult to identify in TTE scenarios. Therefore, we propose an **OOD** model for **T**ravel **T**ime **E**stimation (OOD4TTE), which aims to generate invariant representations for TTE. We construct a TTE-specific structural causal model (SCM) to visualize the data generation process, and selectively blocking relations to simplify the learning process for the invariant representation. The environment label is also inferred in a soft type to balance the risk across the traffic concepts. Our OOD4TTE is a two-stage invariant learning framework,

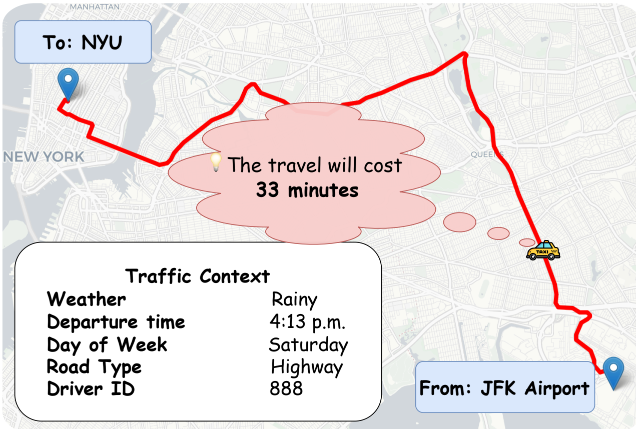

Figure 1: An example of the TTE: The travel time from JFK Airport to New York University is estimated to be 33 minutes.

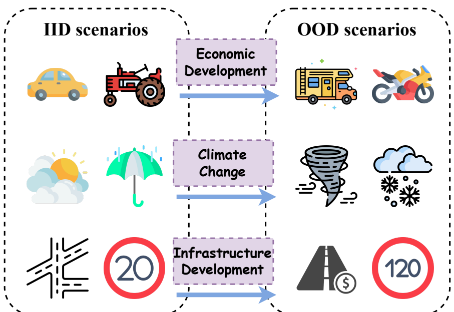

Figure 2: Illustration of the OOD problem in the TTE task: Training on IID and Test on OOD.

with a bi-level optimization on the environment inference module and the invariant representation model.

Our literature review reveals that no existing studies have addressed the OOD problem in the TTE task, despite its urgency and practical importance. In a nutshell, the main contributions of this paper are as follows:

- **New Problem.** To our best knowledge, we are the first to discuss the OOD problem in the TTE task. We use the causal theory and invariant learning to address the challenges.
- **Novel Method.** We propose an end-to-end model, OOD4TTE, to address the OOD generalization problem in the TTE task. To tackle the two key challenges, we introduce an *environment inference module* and a *two-step deconfounding procedure* for learning invariant representations. By leveraging *selective blocking*, we simplify the underlying causal structure, thereby learning invariant representation much easier and improving generalization performance. Based on this simplified structure, the *invariant gating network* effectively captures stable features across different environments.
- **Extensive Experiments.** We conduct extensive experiments to evaluate the performance of our method. OOD4TTE is evaluated on two real-world datasets. Our experimental results demonstrate that OOD4TTE achieves superior generalization performance under OOD scenarios.

2 Preliminary

2.1 Definition & Problem Formulation

Generally, the queries for travel time estimation are composed of two types of covariates, categorized by their underlying semantics.

Definition 1 (Route). *The route T is of sequential characteristics with identification information, which usually displays as a sequence $\{s_1, s_2, \dots, s_m\}$, such as road names or IDs.*

Definition 2 (Traffic context). *The traffic context C comprises traffic attributes and other factors that may affect travel time, including the departure time, weather conditions, and infrastructure information.*

Our core problem, Travel Time Estimation (TTE), is formally defined as follows:

Problem 1 (Travel Time Estimation). *Given the traffic context $c \sim \mathcal{D}(C)$ and the planned route $t \sim \mathcal{D}(T)$, the estimating function f aims to accurately estimate the travel time $\hat{Y} = f(c, t)$.*

We define the OOD generalization problem in the TTE as follows:

Definition 3 (Data-generating Environment). *The data-generating environment $e \in \mathcal{E}$ is defined as the root causal factor that determines the joint generative distribution of route and traffic context covariates $P^e(T, C)$.*

The traffic context is a snapshot of a single trip (e.g., the real-time weather, departure time of a trip), while the data-generating environment is the distribution rule of all trips (e.g., the rainy season, the peak hour of a specific region).

Problem 2 (OOD Generalization in TTE). *Given the environment $e \in \mathcal{E}$, we aim to find the estimation function $\hat{f}(\cdot)$ that satisfies:*

$$\hat{f} = \arg\min_{f} \sup_{e \in \mathcal{E}} \mathbb{E}_{P(T,C,Y|E=e)} \left[\mathcal{L}\left(f(T, C), Y \right) \right], \tag{1}$$

where $\mathcal{L}$ denotes the loss function. The objective is to optimize performance in the *worst-case* environment, thereby enhancing generalization to previously unseen environments.

3 A Causal Perspective for OOD

We aim to address the two problems using causal theory. Nevertheless, the provable identification of invariant features requires the imposition of certain general constraints.

Assumption 1 (Identifiability of Environment). *The environment is inferable from the joint information of traffic context C and route T.*

It is reasonable to make this assumption in traffic scenarios. For example, the regional peak hour can be inferred from the departure time and route selection preferences in the traffic context. The rainy season characteristics can be identified from the real-time weather and road avoidance patterns of the route.

Assumption 2 (Sufficient Condition). *The sufficient invariants for accurate and generalizable travel time estimation can be identified from traffic context C and route T.*

This assumption aligns with those made in many OOD generalization studies [15, 40]. Considering the success of incorporating traffic context and route as covariates in practical online travel time estimation [8, 35, 36], we believe that they contain sufficient invariant information. Then, the invariance principle for invariant representations is defined as:

Principle 1 (TTE Invariance Principle). *For the random variables: route T, traffic context C, and travel time Y, two distinct representation functions $\Phi(\cdot)$ and $\Psi(\cdot)$ capture invariants from C and T respectively, which satisfy:*

$$P^{e_i}(Y \mid \Phi(C), \Psi(T)) = P^{e_j}(Y \mid \Phi(C), \Psi(T)). \quad (2)$$

We aim to learn invariant representations for traffic contexts and routes simultaneously. This allows the model to avoid learning shortcut features and to generalize robustly from the training data when it properly captures the invariants. However, it remains a question to design an appropriate representation method. Next, we leverage causal theory to address this problem.

3.1 A Causal Perspective of TTE

To rigorously analyze the problem in a causal perspective, we design a Structural Causal Model (SCM) [25] for analyzing complex interactions in the travel time estimation. Specifically, in Fig. 3, each node represents a causal variable, and each edge denotes a causal relation between two variables. The SCM illustrates the causal relations among five key variables: environment E, route T, traffic context C, invariant representation R, and travel time Y. The interactions between these variables are described as follows:

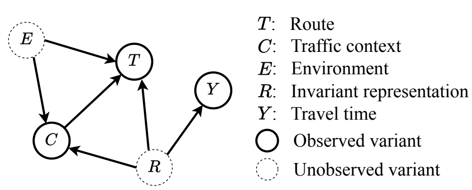

Figure 3: SCM for TTE.

- $C \leftarrow E \rightarrow T$: As defined in Def. 3, the environment is the root causal factor that determines the joint generative distribution of T and C. It is generally difficult to fully observe in practice. For example, the regional peak hour and inherent traffic patterns cannot be directly identified without prior domain knowledge.
- $T \leftarrow R \rightarrow C$: In addition to the confounder, invariant information is also contained in C and T. The key variable R is what we aim to learn. Thus we can make an accurate estimation with the observed C and T. For example, the accessibility of roads could be an invariant.
- $R \rightarrow Y$: Since the invariant representation has a stable relationship with travel time, R has sufficient information for estimation.
- $C \rightarrow T$: The route is usually influenced by the traffic context. For example, people may choose to take a detour to avoid congestion during peak hours.

From this SCM, we observe that the paths $E \rightarrow C \leftarrow R$ and $E \rightarrow T \leftarrow R$ form collider structures, which induce spurious correlation between E and R. For example, E = *rainy season* causes C = *rainy* and T = *expressway* (because small roads are blocked), which leads to a decrease in expected travel time. The spurious correlation may be learned: C = *rainy* $\rightarrow$ $Y \downarrow$. Thus, we need to adjust for C and T (i.e., by learning invariant representation $\Phi(C)$ and $\Psi(T)$) to block the spurious correlation between E and R and learn the pure R.

Traditional tools for causal theory like backdoor or front-door adjustment is hard to apply to the SCM due to the complex relations. Intuitively, removing the path $C \rightarrow T$ can simplify the causal relationships. Although directly removing the path is feasible, it inevitably leads to the loss of essential information. Hence, we propose **selective blocking**, which partially blocks $C \rightarrow T$ to eliminate confounding influence from E. Selective blocking offers two benefits: (1) Simplifying the causal relationships by reducing the number of paths from E to R from four to two, thereby weakening confounding; (2) Enhancing the learnability of double invariants: misidentifying invariants in C (or T) increases difficulty in identifying invariants in T (C). Therefore, we propose a **two-step deconfounding procedure**: first by selectively blocking $C \rightarrow T$, then blocking $E \rightarrow C$ and $E \rightarrow T$. Based on this analysis, we now present our method.

4 Methodology

In this section, we introduce the architecture of our proposed **OOD** model for **T**ravel **T**ime **E**stimation (OOD4TTE), as shown in Fig. 4.

Given routes with various numbers of steps, we first use a temporal method (e.g., LSTM [7] and Transformer [31]) to represent them into vectors $\mathbf{T} \in \mathbb{R}^{N \times d_{\text{model}}}$. Together with the traffic contexts $\mathbf{C} \in \mathbb{R}^{N \times d_{\text{model}}}$, OOD4TTE infers the environment for the invariant representation learning. Next, we implement the two-step deconfounding procedure to adjust for $\mathbf{C}$ and $\mathbf{T}$, yielding $\Phi(\mathbf{C})$ and $\Psi(\mathbf{T})$, which block spurious correlation between the environment and the invariant representation. Finally, the estimation head predicts the travel time vector $\hat{\mathbf{y}} \in \mathbb{R}^N$ from $\Phi(\mathbf{C})$ and $\Psi(\mathbf{T})$.

4.1 Environmental Inference Module

As the goal of invariant representation learning, we would like to have robust representations for each trip, which could be stable across environments. However, in traffic data, environment labels are unobservable. This module aims to infer the environment labels from the observable features for invariant identification.

Since traffic generation is not governed by a single aspect, we adopt multiple concepts to explain the underlying data-generating mechanism. We define $\mathbf{E} \in \mathbb{R}^{N \times V}$ as the soft labels for V traffic concepts. The risk of e-th concept is defined as the weighted loss of each instance in the e-th concept, as follows:

$$\mathcal{R}^e_{env}(f, \mathbf{E}) = \frac{1}{\sum_{i'} \mathbf{E}_{i',e}} \sum_i \mathbf{E}_{i,e}\, \mathcal{L}(f(\mathbf{C}_i, \mathbf{T}_i), y_i), \quad (3)$$

where f is an invariant representation model, and $\mathcal{L}$ is a loss function. Given that the representation model learns stable relations

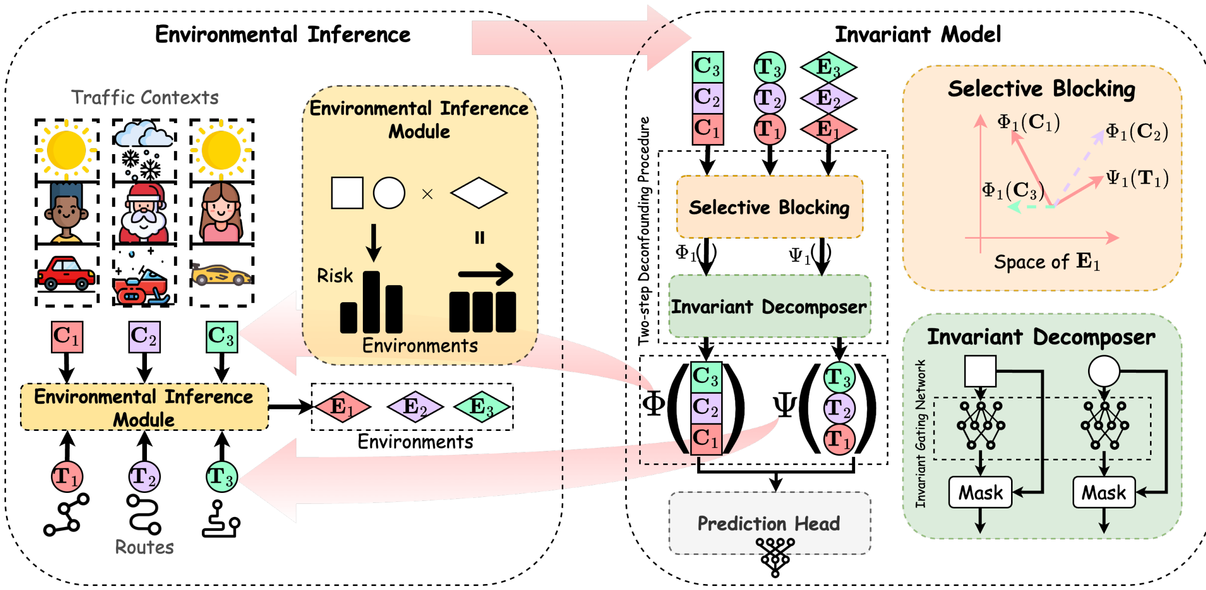

Figure 4: Architecture of OOD4TTE.

and achieves consistent performance across environments, the soft label $\mathbf{E}$ should define a proper environment partition that minimizes the risk gap among concepts. In other words, the module searches for a partition that is most consistent with the invariance already captured by the invariant model f. Here we use the variance of each concept risk to encourage equal risks [11, 37]. The invariant model and the environment labels are updated iteratively as follows:

$$\mathbf{E}^* = \arg\min_{\mathbf{E}} \operatorname{Var} \mathcal{R}^e_{env}(f, \mathbf{E}), \tag{4}$$

$$f^* = \arg\min_{f} \mathcal{R}_{inv}(f, \mathbf{E}), \tag{5}$$

where $\mathcal{R}_{inv}$ is the risk of the representation model. Furthermore, an entropy regularizer is suggested to be imposed on $\mathbf{E}^*$ to prevent collapse. Next, we will discuss f and $\mathcal{R}_{inv}$.

4.2 Two-step Deconfounding Procedure

We perform deconfounding of the influence of environment labels on invariant representations in two steps: (1) selectively blocking the confounding path of traffic contexts on routes, where Φ_1 and Ψ_1 are the mediate representing functions for C and T, respectively; (2) based on the updated causal relations, we block the remaining spurious relations from the environment to invariant representations, denoted as $\Phi(C)$ and $\Psi(T)$.

4.2.1 Selective Blocking. In Fig. 3, path $C \to T$ increases the complexity of causality due to introducing more paths from E to R, and the model finds it much harder to learn invariants due to misinterpretation arising from the misidentification of T and C. To achieve the goal of selective blocking, it is necessary to ensure that the $\Phi_1(C)$ and $\Psi_1(T)$ are independent conditioned on the environment E (i.e., $\Phi_1(C) \perp\!\!\!\perp \Psi_1(T) \mid E$). To this end, we design a contrastive module as follows.

Initially, we shuffle the traffic contexts in data to generate negative samples. With slight abuse of notation, we define the shuffled $\mathbf{C}'$ to distinguish it from the original $\mathbf{C}$, where $\mathbf{C}$ is regarded as positive samples and $\mathbf{C}'$ as negative ones.

Without loss of generality, we adopt partial distance measure [30, 43] to quantify the conditional dependency between $\Phi_1(C)$ and $\Psi_1(T)$. The partial distance measure is a scalar quantity that quantifies dependence with the conditional correlation in Gaussian settings. In non-Gaussian situation, a partial distance of zero does not imply conditional independence. However, values closer to zero suggest weaker correlation [30].

We first define the *double centered pair-wise distance* of the inferred environments $\mathbf{E}$, denoted as $\pi(\mathbf{E}) \in \mathbb{R}^{N\times N}$, as follows :

$$\begin{aligned}[\pi(\mathbf{E})]_{i,j} = \|\mathbf{E}_i - \mathbf{E}_j\|_2 - \frac{1}{N}\sum_{k=1}^{N}\|\mathbf{E}_k - \mathbf{E}_j\|_2 - \\ \frac{1}{N}\sum_{l=1}^{N}\|\mathbf{E}_i - \mathbf{E}_l\|_2 + \frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}\|\mathbf{E}_k - \mathbf{E}_l\|_2,\end{aligned} \tag{6}$$

where $\|\cdot\|_2$ is the l_2 norm. The $\pi(\mathbf{C})$ and $\pi(\mathbf{T})$ can be similarly obtained. The *inner product operator* $\otimes$ for $\pi(\mathbf{C})$ and $\pi(\mathbf{T})$ is defined as:

$$\pi(\mathbf{C}) \otimes \pi(\mathbf{T}) := \frac{\sum_{i\neq j}[\pi(\mathbf{C})]_{i,j}\cdot[\pi(\mathbf{T})]_{i,j}}{N\times(N-3)}. \tag{7}$$

Others are similarly calculated. The *orthogonal projection operator* $\mathcal{P}$ of $\pi(\mathbf{C})$ on $\pi(\mathbf{E})$ is defined as:

$$\mathcal{P}_{\mathbf{E}}(\mathbf{C}) = \pi(\mathbf{C}) - \frac{\pi(\mathbf{C})\otimes\pi(\mathbf{E})}{\pi(\mathbf{E})\otimes\pi(\mathbf{E})}\pi(\mathbf{E}). \tag{8}$$

The projection $\mathcal{P}_{\mathbf{E}}(\mathbf{T})$ can be similarly obtained. The **partial distance measure** of C and T to E is defined as:

$$D_{\mathbf{E}}(\mathbf{C},\mathbf{T}) = \frac{|\,\mathcal{P}_{\mathbf{E}}(\mathbf{C})\otimes\mathcal{P}_{\mathbf{E}}(\mathbf{T})\,|}{\|\mathcal{P}_{\mathbf{E}}(\mathbf{C})\|\cdot\|\mathcal{P}_{\mathbf{E}}(\mathbf{T})\|}, \tag{9}$$

where $|\cdot|$ denotes the absolute value, $\|\mathcal{P}_{\mathbf{E}}(\mathbf{C})\| = (\mathcal{P}_{\mathbf{E}}(\mathbf{C}) \otimes \mathcal{P}_{\mathbf{E}}(\mathbf{C}))^{1/2}$, and $\|\mathcal{P}_{\mathbf{E}}(\mathbf{T})\|$ is similarly defined.

To achieve the goal of $\Phi_1(C) \perp\!\!\!\perp \Psi_1(T) \mid E$, we hope the $D_{\mathbf{E}}(\Phi_1(\mathbf{C}), \Psi_1(\mathbf{T})) \ll D_{\mathbf{E}}(\Phi_1(\mathbf{C}'), \Psi_1(\mathbf{T}))$. That means, when projected onto the space of environments, the positive representations $\Phi_1(\mathbf{C})$ are orthogonal to $\Psi_1(\mathbf{T})$, while the negative representations $\Phi_1(\mathbf{C}')$ are not. Accordingly, we define the contrastive loss $\mathcal{L}_{con}$ as follows:

$$\mathcal{L}_{con} = \mathrm{SoftPlus}(D_{\mathbf{E}}(\Phi_1(\mathbf{C}), \Psi_1(\mathbf{T})) - D_{\mathbf{E}}(\Phi_1(\mathbf{C}'), \Psi_1(\mathbf{T}))). \tag{10}$$

Both Φ_1 and Ψ_1 are two-layer MLPs. This contrastive design ensures that the learned representations $\Phi_1(\mathbf{C})$ and $\Psi_1(\mathbf{T})$ become conditionally independent given the environment, thereby selectively blocking the confounding influence through $C \to T$ from E.

4.2.2 Invariant Decomposer. Selective blocking initially mitigates spurious correlation from $C \to T$. However, the environment E still has spurious correlation with the invariant representation R through T and C. To further eliminate the spurious correlation, we design the **invariant gating network** for the invariant representation extraction.

Selectively blocked representations $\Phi_1(\mathbf{C})$ are shuffled, denoted as $\Phi_1'(\mathbf{C})$, and then, the gating weight $\mathbf{G}$ is generated as follows:

$$\mathbf{G} = \sigma\left(topk\left(\mathbf{W}_C\Phi_1(\mathbf{C}) \mid \mathrm{Var}\left[\mathbf{W}_C\Phi_1(\mathbf{C}), \mathbf{W}_C\Phi_1'(\mathbf{C})\right]\right)\right), \tag{11}$$

where $topk$ selects the smallest $k\%$ of elements according to the given condition, while setting all remaining entries to zero. σ is an activation function. $\mathbf{W}_C$ is a learnable parameter for linear projection. The gating weight is applied to $\Phi_1(\mathbf{C})$, and the invariant representation function Φ is defined as:

$$\Phi(\mathbf{C}) = \mathbf{G} \odot \Phi_1(\mathbf{C}). \tag{12}$$

Similarly, routes $\mathbf{T}$ can be represented by Ψ with a different linear projection weight $\mathbf{W}_T$.

4.3 Estimation Head

Under the sufficient condition, we use the invariant representations $\Phi(\mathbf{C})$ and $\Psi(\mathbf{T})$ as inputs to the estimation head. The estimation head is implemented as a three-layer MLP, with the sum of the invariant representations as the input. The estimation is as follows:

$$\hat{\mathbf{y}} = \mathrm{MLP}(\Phi(\mathbf{C}) + \Psi(\mathbf{T})). \tag{13}$$

4.4 Model Optimization

The quality of estimated travel time $\hat{\mathbf{y}}$ is evaluated using Mean Absolute Percentage Error (MAPE). Additionally, to encourage the model to learn invariant representations and enhance generalization, we introduce the variance of the loss. Note that the variance term used here serves the purpose of invariant representation learning by optimizing the invariant model f, whereas the one in $\mathcal{R}_{env}$ is used for environment inference by optimizing $\mathbf{E}$. The risk of invariant learning $\mathcal{R}_{inv}$ is defined as:

$$\mathcal{R}_{inv} = \underbrace{\mathbb{E}_E\left[\mathcal{L}_{tte}\right]}_{first\ moment} + \alpha \underbrace{\mathbb{E}_E\left[\mathrm{Var}\mathcal{L}_{tte}\right]}_{second\ central\ moment} + \beta\mathbb{E}\left[\mathcal{L}_{con}\right], \tag{14}$$

where α and β are tunable hyperparameters that balance the weights between different risk terms. Eq. (4) and (5) are optimized iteratively. During training, we adopt Adam optimizer [10] for the invariant representation model and the environmental inference module.

In the invariance principle (Eq. 2), we aim to align the loss distributions across different environments, i.e., $\forall e_i, e_j \in \mathcal{E}$, we have $P^{e_i}(\mathcal{L}_{tte}) \approx P^{e_j}(\mathcal{L}_{tte})$. The three terms in our invariant risk can then be interpreted as the first moment, the second central moment, and a regularization term. Minimizing the risk can thus be seen as performing moment matching, which approximates the loss distributions. In this way, our method adheres to the invariance principle and can capture the invariance.

5 Experiments

In this section, we report our experimental results. The implemented code and data are available at our repository[1].

5.1 Experimental Setting

5.1.1 Datasets. The experiments are conducted on two real-world datasets. The **Shenzhen Dataset**[2] contains 30-day routes from August 1st, 2020 to August 30th, 2020 in Shenzhen, China, collected from the Didi Platform. We selected the first 5 days of data, which includes 636,373 routes and their corresponding traffic contexts. The traffic contexts include features such as weather, departure time, and driver ID, etc. The **Porto Dataset**[3] contains 1,710,671 routes and corresponding traffic contexts from 442 taxis in Porto, Portugal, spanning from July 1st, 2013 to June 30th, 2014. Since the routes consist of GPS positions, we mapped them onto road segments using the Valhalla[4] engine. The traffic contexts here include features such as road width, speed limit, departure time, and driver ID, etc.

5.1.2 Baselines. To comprehensively evaluate the performance of our approach under distribution shifts, we compare it against a range of baselines. These include a standard training strategy (**ERM**), four SOTA OOD generalization methods, and four representative TTE backbone models.

OOD Generalization Methods. We apply the following methods to enhance the robustness of the base predictors:

- **ERM** (Empirical Risk Minimization): The standard training objective that minimizes the average loss over training data without any distributional robustness or OOD constraints.
- **EIIL** [3]: Environment Inference for Invariant Learning, which infers worst-case environment partitions by maximizing an invariance-violation objective, enabling invariant learning without explicit environment labels.
- **RevIN** [9]: Reversible Instance Normalization, a symmetric module that normalizes input sequences and denormalizes the corresponding outputs to mitigate distribution shifts in time-series forecasting.
- **FOIL** [17]: an invariant learning framework that infers latent environments via a surrogate loss to mitigate the impact of unobserved confounders in time-series forecasting.
- **HRM** [18]: Heterogeneous Risk Minimization, which jointly clusters latent environments and learns invariant predictors to filter out spurious correlations and ensure robustness under distribution shifts.

[1]https://github.com/CrazySpy/OOD4TTE_public
[2]https://sigspatial2021.sigspatial.org/sigspatial-cup/
[3]http://www.geolink.pt/ecmlpkdd2015-challenge/dataset.html
[4]https://github.com/valhalla/valhalla

TTE Backbones. To adapt the above OOD methods to TTE tasks, we implement them across four distinct backbone architectures:

- **WDR** [36]: a classic architecture that simultaneously trains wide linear models and deep neural networks to capture both memorization and generalization features.
- **WDDRA** [6]: Wide-Deep-Double-Recurrent model with an auxiliary loss, which couples wide & deep components with a double-recurrent structure and an auxiliary link-status prediction loss to capture complex spatial-temporal dependencies in trajectory data.
- **ProbTTE** [16]: A probabilistic framework that estimates the probability distribution of travel times rather than point estimates, focusing on uncertainty quantification.
- **Informer** [41]: An efficient Transformer-based model for long-sequence time-series forecasting, employing ProbSparse self-attention, self-attention distilling, and a generative-style decoder to reduce complexity.

5.1.3 Training Protocols. We follow the protocol that splits the datasets into training, validation, and test sets with ratios of 75%, 5%, and 20%. Experimental results are reported with three evaluation metrics: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE) for all three datasets consistently. All these metrics are better when their values are lower. They evaluate different aspects of performance: MAE and MAPE are sensitive to small values (i.e., short routes in TTE), while RMSE is sensitive to large values (i.e., long routes in TTE). We repeat each experiment five times independently, and additionally perform a statistical significance test on the aggregated experimental results. OOD4TTE is implemented with PyTorch. The model is trained on a server equipped with three pieces of NVIDIA A40 GPU and one piece of Intel(R) Xeon(R) Platinum 8268 CPU.

5.2 Experimental Results

To better reveal the effectiveness of our design, we conclude with four questions and try to answer them with experimental results.

5.2.1 ***Q1: What is the performance on Travel Time Estimation in OOD scenarios?*** The OOD performance is evaluated under two scenarios: variable traffic contexts and variable routes. For variable traffic contexts, we divide the departure times into four folds, select three as training data, and use the remaining one as testing data. For variable routes, we divide the routes into 2 folds based on their lengths to simulate out-of-distribution conditions.

Table 1 reports the results on the Shenzhen and Porto datasets under the traffic context OOD setting. In particular, OOD4TTE achieves a 7.50% improvement in the worst-case performance on the Shenzhen dataset during the 6:00–11:59 time window, demonstrating its robustness to context distribution shifts. Results for the route OOD scenario are provided in the Appendix.

5.2.2 ***Q2: Are the components for optimization effective?*** We conduct an ablation study to validate the effectiveness of each key component in OOD4TTE, with the experimental settings consistent with the traffic context OOD scenario. Specifically, we ablate the core modules and loss terms by removing or replacing them with naive alternatives: ❶ setting $\alpha = 0$ to ablate the term $\mathbb{E}_E\left[\mathrm{Var}\mathcal{L}_{tte}\right]$, thus verifying the effect of the invariant modeling for risk equalization across environments; ❷ setting $\beta = 0$ to remove the contrastive loss $\mathcal{L}_{con}$, so as to evaluate the contribution of the selective blocking strategy; ❸ replacing the Invariant Gating Network with a simple MLP (denoted as w/o Gat) to test the effectiveness of the invariant representation decomposition; ❹ replacing the Environmental Inference Module with all-one vectors (denoted as w/o Env) to assess the importance of inferring soft environmental labels.

The results in the traffic context scenario have been included in Table. 2. The results of the route OOD scenario are in the Appendix. All the components prove effective: α minimizes cross-environment loss variance to drive the model toward environment-independent invariant representations and avoid spurious correlations, while β strengthens selective blocking via contrastive loss to achieve conditional independence of traffic context and route, simplifying causal structures for robust invariant representation learning.

5.2.3 ***Q3: How sensitive is the model to the hyperparameters?*** The sensitivity experiments are conducted on the Porto dataset for four hyperparameters: the coefficients α and β, the number of environments V and the invariant ratio k. As shown in Fig. 5, we observe that:

1. **α and β.** Both the hyperparameters vary over {0.1, 0.5, 1, 5}. We find that two hyperparameters are usually fit between 0.5 and 1. This also indicates that α and β are two trade-off parameters for the invariance risk.
2. **Number of traffic concept** V. The V is adjusted in {4, 8, 16, 32}. A few traffic concepts reduce the representational ability, whereas a greater number of concepts fail to extract common factors of environments.
3. **Invariant Ratio** k. The invariant ratio k controls the ratio of shortcuts to invariant representations in embeddings. The values of k are tuned in {0.2, 0.4, 0.6, 0.8}. A lower value could lose invariant information, whereas a larger value could introduce shortcuts. The results suggest a value around 0.5.

5.2.4 ***Q4: Does our method introduce significant overhead during training and inference?*** We compare OOD4TTE against four baselines, including WDR, a lightweight industrial solution deployed on the Didi platform. Train time and memory are measured during training, while inference latency and throughput are measured during inference.

As shown in Table 3, our method achieves efficiency comparable to WDR on the Shenzhen dataset, using a single NVIDIA A40 GPU. Compared with WDR, OOD4TTE incurs a 54.1% increase in memory consumption (323.57 vs. 210.00 MB). This overhead primarily originates from the pairwise distance computation (Eq. 6–9), which is required by our method.

5.3 Case Study: A Typhoon day in Shenzhen

To further validate the robustness of OOD4TTE in extreme traffic scenarios, we conduct a case study on the Shenzhen dataset focusing on August 16, 2020, when a typhoon hit Shenzhen, causing severe traffic disruptions, e.g., road waterlogging, reduced speed limits, and route deviations.

Table 4 presents the TTE performance of all baseline methods and OOD4TTE under this extreme scenario. The results indicate

Table 1: Performance in the traffic context OOD scenarios on the Shenzhen and Porto datasets. The first row indicates the testing period. A star (*) denotes statistical significance with $p < 0.05$. Bold font indicates the best results, and underlined font indicates the runner-up.

Dataset	Backbone	Method	0:00–5:59			6:00–11:59			12:00–17:59			18:00–23:59		
			MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE
Shenzhen	WDR	ERM	119.40	16.41	176.56	153.18	17.48	229.53	105.71	14.97	145.57	132.54	18.96	215.84
		+ EIIL	114.20	16.10	170.50	145.20	16.90	215.40	103.80	15.05	150.10	120.50	17.80	195.20
		+ RevIN	110.19	15.73	163.68	138.73	16.42	208.87	101.50	15.00	149.16	114.42	17.19	184.31
		+ FOIL	134.47	24.88	174.65	131.61	18.92	189.37	143.88	27.60	178.21	158.95	32.44	201.50
		+ HRM	196.49	24.53	272.29	250.75	27.65	342.85	149.35	19.79	212.39	131.97	18.69	213.88
	WDDRA	ERM	113.77	15.95	166.47	149.86	17.23	222.12	104.30	14.99	145.05	115.09	17.70	171.24
		+ EIIL	113.20	15.85	165.10	129.50	15.80	202.40	102.80	14.90	145.80	113.50	17.40	169.80
		+ RevIN	113.02	15.97	166.85	131.60	15.73	198.69	104.94	14.98	144.51	113.71	17.38	177.09
		+ FOIL	118.51	19.29	163.87	132.96	15.82	206.07	112.71	16.18	164.17	209.85	27.59	302.34
		+ HRM	112.06	16.05	164.38	132.54	15.89	200.15	101.14	**14.82**	147.55	112.13	17.24	171.40
	ProbTTE	ERM	119.80	16.64	193.15	147.73	16.21	240.95	108.01	15.35	145.21	119.24	17.94	192.63
		+ EIIL	115.50	16.10	175.20	140.50	16.35	220.10	104.20	15.20	147.50	116.80	17.60	185.50
		+ RevIN	111.43	15.66	164.42	137.73	16.23	209.54	101.13	15.13	149.19	112.16	16.93	175.17
		+ FOIL	118.20	17.48	169.76	148.76	19.17	216.72	167.21	27.20	219.08	117.00	17.52	174.10
		+ HRM	118.43	16.29	175.59	133.25	15.93	201.38	103.77	14.96	145.89	115.43	17.33	188.08
	Informer	ERM	161.71	22.25	241.27	226.58	25.23	332.76	153.51	21.77	222.18	182.46	25.33	277.21
		+ EIIL	155.40	20.80	245.10	195.20	22.50	298.50	148.50	21.20	215.50	165.20	24.10	260.50
		+ RevIN	150.26	19.42	242.93	174.04	20.49	267.68	143.06	20.85	208.63	153.03	22.66	240.86
		+ FOIL	252.99	45.47	366.16	229.35	31.08	337.97	207.12	36.31	285.12	223.74	44.37	307.97
		+ HRM	182.28	23.70	286.28	305.58	32.30	427.61	182.88	24.33	269.59	182.98	27.93	277.71
	OOD4TTE		**109.30***	**15.28***	**163.62***	**119.79***	**14.75***	**183.30***	**99.10***	14.95	**144.35**	**108.33***	**16.88**	**166.11***
Porto	WDR	ERM	196.39	23.03	383.93	137.44	24.26	239.71	166.18	21.21	406.88	195.50	21.84	479.32
		+ EIIL	198.50	23.35	379.40	138.10	24.40	241.50	165.12	21.75	409.20	198.40	22.30	475.10
		+ RevIN	211.45	25.19	378.59	137.74	24.79	239.97	177.99	23.46	411.91	203.79	23.46	481.20
		+ FOIL	208.49	23.94	394.18	143.13	25.19	268.02	181.12	23.74	480.04	231.16	24.12	470.36
		+ HRM	207.55	23.98	375.34	136.28	24.08	242.03	172.63	21.33	441.26	211.63	22.95	468.27
	WDDRA	ERM	195.53	22.92	360.63	140.27	24.99	241.22	168.35	21.36	385.63	194.56	21.74	436.11
		+ EIIL	197.10	23.15	375.20	141.50	24.60	240.80	170.20	21.50	388.50	195.80	21.90	442.10
		+ RevIN	206.90	24.27	407.14	138.10	24.32	251.18	178.65	23.43	399.21	206.85	23.85	441.75
		+ FOIL	329.72	77.52	429.75	238.26	47.95	410.20	321.76	78.32	475.21	291.18	58.66	560.69
		+ HRM	201.97	23.71	382.57	145.48	24.14	236.79	168.00	**21.14***	**382.66***	189.51	21.64	452.39
	ProbTTE	ERM	207.72	24.19	383.92	166.69	26.87	273.45	186.65	23.92	382.09	195.79	22.05	461.09
		+ EIIL	204.50	24.10	372.10	144.20	24.50	250.30	182.50	23.50	395.20	200.10	22.50	450.50
		+ RevIN	203.17	24.96	362.15	145.19	24.08	259.82	212.31	26.26	456.62	220.16	26.96	435.11
		+ FOIL	266.29	50.49	424.93	304.21	83.70	390.98	277.73	50.16	485.49	254.50	43.98	475.57
		+ HRM	197.78	23.43	364.30	144.84	24.44	237.85	173.34	22.16	391.96	191.78	21.89	438.36
	Informer	ERM	258.81	29.65	434.79	172.34	26.10	284.92	229.57	29.11	426.41	240.16	27.83	476.37
		+ EIIL	194.20	23.10	363.50	139.50	24.80	240.20	169.10	22.30	401.50	199.50	22.40	460.20
		+ RevIN	196.85	23.24	364.91	140.27	24.99	241.22	170.49	22.47	402.97	200.27	22.54	463.98
		+ FOIL	275.93	30.76	453.37	210.11	48.90	476.03	304.21	83.70	390.98	316.03	34.38	509.83
		+ HRM	200.25	23.80	365.59	139.24	24.43	246.83	177.11	23.17	442.97	199.19	22.51	455.91
	OOD4TTE		**189.71***	**22.58***	**358.91***	**132.24***	**23.43***	**235.84***	**164.29***	21.84	377.34	**187.48***	**21.44***	**432.49***

that OOD4TTE's predictions are more consistent and less sensitive to extreme traffic fluctuations caused by typhoons. Most baseline methods struggle to generalize to typhoon-induced OOD scenarios. They over-rely on spurious correlations (e.g., normal weather traffic patterns) that break down under extreme conditions.

Table 2: Ablation study in the traffic context OOD scenarios on the Shenzhen and Porto datasets. The first row indicates the testing period. Bold font indicates the best results.

Dataset	Method	0:00–5:59			6:00–11:59			12:00–17:59			18:00–23:59		
		MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE
Shenzhen	OOD4TTE	**109.30**	**15.28**	**163.62**	**119.79**	**14.75**	**183.30**	**99.10**	**14.95**	**144.35**	**108.33**	**16.88**	**166.11**
	$\alpha = 0$	113.11	15.62	169.81	143.00	16.64	217.45	100.18	15.18	145.53	109.84	17.72	168.92
	$\beta = 0$	110.77	15.33	167.54	124.59	15.98	189.78	103.28	15.27	149.96	109.75	17.21	168.17
	w/o Gat	111.20	15.42	166.90	128.85	16.22	196.40	99.85	15.05	146.10	108.95	17.05	167.20
	w/o Env	112.25	15.78	168.95	125.40	16.05	190.90	101.90	15.40	148.85	110.60	17.55	170.85
Porto	OOD4TTE	**189.71**	**22.58**	**358.91**	**132.24**	**23.43**	**235.84**	**164.29**	**21.84**	**377.34**	**187.48**	**21.44**	**432.49**
	$\alpha = 0$	194.66	23.43	360.96	136.53	24.18	239.13	177.41	22.87	431.73	188.02	21.78	435.66
	$\beta = 0$	194.31	23.30	362.01	135.19	23.99	237.77	172.07	22.49	412.84	189.70	21.82	436.10
	w/o Gat	204.80	25.50	392.60	138.70	25.80	243.40	176.40	24.15	415.80	190.90	22.15	440.20
	w/o Env	201.50	24.85	381.40	140.90	26.35	244.60	182.10	25.10	428.90	196.30	23.55	452.00

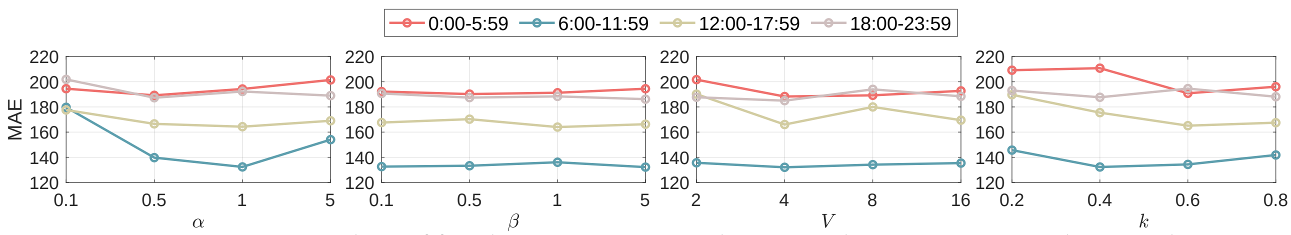

Figure 5: Sensitivity analysis of four hyperparameters in the temporal OOD scenarios on the Porto dataset.

Table 3: Computational efficiency comparison on the Shenzhen dataset. All experiments are conducted on a single NVIDIA A40 GPU.

Method	Train Time (h)	Inference Latency (ms)	Throughput (samples/s)	Memory (MB)
WDR	2.13	4.12	248,543	210.00
WDDRA	3.25	4.58	231,206	267.43
ProbTTE	3.41	4.71	224,817	281.65
Informer	5.67	6.89	158,432	412.38
OOD4TTE	3.82	4.33	236,473	323.57

6 Related work

6.1 Travel Time Estimation

Route-based Travel Time Estimation. Early segment-based TTE methods [1, 35] often suffer from error accumulation. To mitigate this issue, WDR [36] introduces holistic route modeling. Subsequent multi-task frameworks, such as DeepTTE [33] and WDDRA [6], incorporate auxiliary objectives to capture complex dependencies. ProbTTE [16] further advances this line by modeling predictive uncertainty.

Origin-Destination Travel Time Estimation. OD-TTE is a branch of TTE. Unlike route-dependent methods, it estimates duration using only origin and destination information, which is important for pre-dispatch scenarios. While early works [13, 39] focused on static geo-representations, recent studies address sparsity with graph-based oracles. DOT [14], for example, uses a diffusion probabilistic model to infer latent pixelated trajectories as map-based priors.

Table 4: Results on the Shenzhen dataset on August 16, 2020, a typhoon day. Bold font indicates the best results, and underlined font indicates the runner-up.

Backbone	Method	MAE	MAPE	RMSE
WDR	ERM	108.78	0.1433	159.92
	+ EIIL	106.50	0.1420	154.60
	+ RevIN	104.23	0.1410	151.88
	+ FOIL	132.76	0.2247	170.96
	+ HRM	156.23	0.1897	215.95
WDDRA	ERM	107.43	0.1430	157.78
	+ EIIL	101.50	0.1365	<u>147.50</u>
	+ RevIN	<u>100.84</u>	0.1358	148.03
	+ FOIL	128.70	0.2057	171.16
	+ HRM	105.82	0.1376	154.48
ProbTTE	ERM	108.55	0.1373	184.81
	+ EIIL	101.20	<u>0.1345</u>	150.80
	+ RevIN	102.01	0.1357	152.16
	+ FOIL	122.38	0.1707	187.04
	+ HRM	105.84	0.1421	153.33
Informer	ERM	173.72	0.2313	250.73
	+ EIIL	146.50	0.1920	212.50
	+ RevIN	144.93	0.1903	210.03
	+ FOIL	149.74	0.1959	217.93
	+ HRM	158.88	0.2037	236.94
OOD4TTE		**95.28**	**0.1289**	**140.99**

En-route Travel Time Estimation. En-route TTE requires real-time adaptability, where balancing accuracy and latency remains challenging. Traditional recurrent models such as CompactETA [5] incur high overhead due to continuous state updates. Recent methods leverage uncertainty estimation, as exemplified by DutyTTE [23]. This motivates uncertainty-guided frameworks [28] that replace fixed-frequency updates with dynamic triggers and recompute only when trajectory deviations exceed confidence bounds.

6.2 OOD Generalization

Existing OOD generalization methods can be broadly categorized based on the accessibility of environment annotations and the mechanism of invariance learning.

Supervised Domain Generalization. Methods in this category rely on explicit environment labels or prior knowledge (e.g., spatial coordinates) to learn invariant representations. Classical approaches like **IRM** [2] and **REx** [11] enforce invariance across annotated training environments via penalty terms, while **Group-DRO** [27] minimizes the worst-group risk. In the context of complex series data, **DIVERSIFY** [21] and **ITSR** [20] extend these principles to segment-level classification tasks, requiring fine-grained environment labels. Similarly, spatio-temporal methods [32, 38, 42] typically utilize explicit spatial correlations to define environments. However, since our task focuses on non-spatial attributes, and obtaining precise environment labels in traffic scenarios is often impractical, these supervised methods are not applicable to our setting.

Latent Environment Inference. To address label scarcity, recent approaches aim to infer latent environments or adapt to shifts directly from data. From a debiasing perspective, **LfF** [24] identifies bias-conflicting samples without supervision. Optimization frameworks like **EIIL** [3] and **HRM** [18] infer worst-case environment partitions to maximize risk variance or capture latent heterogeneity. Specific to temporal shifts, **AdaRNN** [4] dynamically adjusts recurrent states to capture evolving dependencies, while **CauSTG** [42] segments time into latent environments based on correlation situations. Notably, **RevIN** [9] employs a simple instance normalization strategy to mitigate non-stationarity, and **FOIL** [17] introduces a novel label-free framework to infer temporal environments. These methods align with our goal of handling distribution shifts in the absence of explicit environment labels. Beyond these optimization-centric methods, recent theoretical advancements in Large Language Models offer a mechanistic perspective. Specifically, [29] reveals that the OOD generalization capability of Transformers in LLMs is largely driven by induction heads via compositionality.

7 Conclusion & Limitation

In this paper, we first study the OOD generalization problem in the travel time estimation task. We reveal the complex causal relationships in the TTE task. Our model, OOD4TTE uses an environmental inference module to infer the missed environment labels in the TTE task and a two-step deconfounding procedure for better invariant representation learning. All the components are proven to be effective in the ablation study. Experimental results show that OOD4TTE can effectively solve the OOD generalization problem.

A key limitation of OOD4TTE is its lack of consideration for graph-modality-specific OOD issues, such as the distribution shift of road network topology across cities. The model cannot be directly transferred to new urban environments with distinct spatial connectivity and traffic patterns. This prevents OOD4TTE from serving as a general solution to cross-city OOD problems.

Acknowledgments

X. Meng was supported by the National Natural Science Foundation of China (No. 62172423). H. Li was supported by the Beijing Major Science and Technology Project (No. Z251100008425006). X. Jiang was supported by the Outstanding Innovative Talents Cultivation Funded Programs 2025 of Renmin University of China.

References

[1] Pouria Amirian, Anahid Basiri, and Jeremy Morley. 2016. Predictive Analytics for Enhancing Travel Time Estimation in Navigation Apps of Apple, Google, and Microsoft. In *Proceedings of the 9th ACM SIGSPATIAL International Workshop on Computational Transportation Science.* ACM, 31–36.
[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant Risk Minimization. arXiv:1907.02893
[3] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. 2021. Environment Inference for Invariant Learning. In *Proceedings of the 38th International Conference on Machine Learning.* PMLR, 2189–2200.
[4] Yuntao Du, Jindong Wang, Wenjie Feng, Sinno Pan, Tao Qin, Renjun Xu, and Chongjun Wang. 2021. AdaRNN: Adaptive Learning and Forecasting of Time Series. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management.* ACM, 402–411.
[5] Kun Fu, Fanlin Meng, Jieping Ye, and Zheng Wang. 2020. CompactETA: A Fast Inference System for Travel Time Prediction. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 3337–3345.
[6] Yunchong Gan, Haoyu Zhang, and Mingjie Wang. 2021. Travel Time Estimation Based on Neural Network with Auxiliary Loss. In *Proceedings of the 29th International Conference on Advances in Geographic Information Systems.* ACM, 642–645.
[7] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.
[8] Jizhou Huang, Zhengjie Huang, Xiaomin Fang, Shikun Feng, Xuyi Chen, Jiaxiang Liu, Haitao Yuan, and Haifeng Wang. 2022. DuETA: Traffic Congestion Propagation Pattern Modeling via Efficient Graph Learning for ETA Prediction at Baidu Maps. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management.* ACM, 3172–3181.
[9] Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. 2022. Reversible Instance Normalization for Accurate Time-Series Forecasting Against Distribution Shift. In *Proceedings of the 10th International Conference on Learning Representations.*
[10] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations.*
[11] David Krueger, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. 2021. Out-of-Distribution Generalization via Risk Extrapolation (REx). In *Proceedings of the 38th International Conference on Machine Learning.* PMLR, 5815–5826.
[12] Yaguang Li, Dingxiong Deng, Ugur Demiryurek, Cyrus Shahabi, and Siva Ravada. 2015. Towards Fast and Accurate Solutions to Vehicle Routing in a Large-Scale and Dynamic Environment. In *Proceedings of the 14th International Symposium on Advances in Spatial and Temporal Databases (Lecture Notes in Computer Science, Vol. 9239).* Springer, 119–136.
[13] Yaguang Li, Kun Fu, Zheng Wang, Cyrus Shahabi, Jieping Ye, and Yan Liu. 2018. Multi-Task Representation Learning for Travel Time Estimation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 1695–1704.
[14] Yan Lin, Huaiyu Wan, Jilin Hu, Shengnan Guo, Bin Yang, Youfang Lin, and Christian S. Jensen. 2023. Origin-Destination Travel Time Oracle for Map-Based Services. *Proceedings of the ACM on Management of Data* 1, 3 (2023), 1–27.
[15] Yong Lin, Shengyu Zhu, Lu Tan, and Peng Cui. 2022. ZIN: When and How to Learn Invariance Without Environment Partition?. In *Advances in Neural Information Processing Systems*, Vol. 35. 24529–24542.
[16] Hao Liu, Wenzhao Jiang, Shui Liu, and Xi Chen. 2023. Uncertainty-Aware Probabilistic Travel Time Prediction for On-Demand Ride-Hailing at DiDi. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* ACM, 4516–4526.
[17] Haoxin Liu, Harshavardhan Kamarthi, Lingkai Kong, Zhiyuan Zhao, Chao Zhang, and B. Aditya Prakash. 2024. Time-Series Forecasting for Out-of-Distribution

Generalization Using Invariant Learning. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 31312–31325.
[18] Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. 2021. Heterogeneous Risk Minimization. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 6804–6814.
[19] David Lopez-Paz and Maxime Oquab. 2017. Revisiting Classifier Two-Sample Tests. In *Proceedings of the 5th International Conference on Learning Representations*.
[20] Wang Lu, Jindong Wang, Xinwei Sun, Yiqiang Chen, Xiangyang Ji, Qiang Yang, and Xing Xie. 2024. Diversify: A General Framework for Time Series Out-of-Distribution Detection and Generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 6 (2024), 4534–4550.
[21] Wang Lu, Jindong Wang, Xinwei Sun, Yiqiang Chen, and Xing Xie. 2023. Out-of-Distribution Representation Learning for Time Series Classification. In *Proceedings of the 11th International Conference on Learning Representations*.
[22] Xiaowei Mao, Tianyue Cai, Wenchuang Peng, and Huaiyu Wan. 2021. Estimated Time of Arrival Prediction via Modeling the Spatial-Temporal Interactions Between Links and Crosses. In *Proceedings of the 29th International Conference on Advances in Geographic Information Systems*. ACM, 658–661.
[23] Xiaowei Mao, Yan Lin, Shengnan Guo, Yubin Chen, Xingyu Xian, Haomin Wen, Qisen Xu, Youfang Lin, and Huaiyu Wan. 2025. DutyTTE: Deciphering Uncertainty in Origin-Destination Travel Time Estimation. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, Vol. 39. AAAI Press, 12390–12398.
[24] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. 2020. Learning from Failure: De-Biasing Classifier from Biased Classifier. In *Advances in Neural Information Processing Systems*, Vol. 33. 20673–20684.
[25] Judea Pearl. 2009. *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press.
[26] Sijie Ruan, Zi Xiong, Cheng Long, Yiheng Chen, Jie Bao, Tianfu He, Ruiyuan Li, Shengnan Wu, Zhongyuan Jiang, and Yu Zheng. 2020. Doing in One Go: Delivery Time Inference Based on Couriers' Trajectories. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2813–2821.
[27] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. In *Proceedings of the 8th International Conference on Learning Representations*.
[28] Zekai Shen, Haitao Yuan, Xiaowei Mao, Congkang Lv, Shengnan Guo, Youfang Lin, and Huaiyu Wan. 2025. Towards an Efficient and Effective En Route Travel Time Estimation Framework. In *Proceedings of the 30th International Conference on Database Systems for Advanced Applications (Lecture Notes in Computer Science, Vol. 15987)*. Springer, 604–619.
[29] Jiajun Song, Zhuoyan Xu, and Yiqiao Zhong. 2025. Out-of-Distribution Generalization via Composition: A Lens Through Induction Heads in Transformers. *Proceedings of the National Academy of Sciences* 122, 6 (2025), e2417182122.
[30] Gábor J. Székely and Maria L. Rizzo. 2014. Partial Distance Correlation with Methods for Dissimilarities. *The Annals of Statistics* 42, 6 (2014), 2382–2412.
[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 5998–6008.
[32] Binwu Wang, Jiaming Ma, Pengkun Wang, Xu Wang, Yudong Zhang, Zhengyang Zhou, and Yang Wang. 2024. STONE: A Spatio-Temporal OOD Learning Framework Kills Both Spatial and Temporal Shifts. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 2948–2959.
[33] Dong Wang, Junbo Zhang, Wei Cao, Jian Li, and Yu Zheng. 2018. When Will You Arrive? Estimating Travel Time Based on Deep Neural Networks. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, Vol. 32. AAAI Press, 2500–2507.
[34] Hongjian Wang, Xianfeng Tang, Yu-Hsuan Kuo, Daniel Kifer, and Zhenhui Li. 2019. A Simple Baseline for Travel Time Estimation Using Large-Scale Trip Data. *ACM Transactions on Intelligent Systems and Technology* 10, 2 (2019), 1–22.
[35] Yilun Wang, Yu Zheng, and Yexiang Xue. 2014. Travel Time Estimation of a Path Using Sparse Trajectories. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 25–34.
[36] Zheng Wang, Kun Fu, and Jieping Ye. 2018. Learning to Estimate the Travel Time. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 858–866.
[37] Yuli Wu, Niklas Walter, Georgy Antonov, Sandeep Pant, Marcel Vu, and Dorit Merhof. 2022. Handling Distribution Shift in Tire Design. *NeurIPS 2022 Workshop on Distribution Shifts* (2022).
[38] Yutong Xia, Yuxuan Liang, Haomin Wen, Xu Liu, Kun Wang, Zhengyang Zhou, and Roger Zimmermann. 2023. Deciphering Spatio-Temporal Graph Forecasting: A Causal Lens and Treatment. In *Advances in Neural Information Processing Systems*, Vol. 36. 37068–37088.
[39] Haitao Yuan, Guoliang Li, Zhifeng Bao, and Ling Feng. 2020. Effective Travel Time Estimation: When Historical Trajectories Over Road Networks Matter. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. ACM, 2135–2149.
[40] Zeyang Zhang, Xin Wang, Ziwei Zhang, Zhou Qin, Weigao Wen, Hui Xue, Haoyang Li, and Wenwu Zhu. 2023. Spectral Invariant Learning for Dynamic Graphs under Distribution Shifts. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*.
[41] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, Vol. 35. AAAI Press, 11106–11115.
[42] Zhengyang Zhou, Qihe Huang, Kuo Yang, Kun Wang, Xu Wang, Yuxuan Liang, and Yang Wang. 2023. Maintaining the Status Quo: Capturing Invariant Relations for OOD Spatiotemporal Learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 3603–3614.
[43] Minqin Zhu, Anpeng Wu, Haoxuan Li, Ruoxuan Xiong, Bo Li, Xiaoqing Yang, Xuan Qin, Peng Zhen, Jiecheng Guo, Fei Wu, et al. 2024. Contrastive Balancing Representation Learning for Heterogeneous Dose-Response Curves Estimation. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, Vol. 38. AAAI Press, 17175–17183.

A Experimental Reproducibility

A.1 Data Preparation

Since the routes in the Porto Dataset consist of GPS positions, we first used the Valhalla engine to map the GPS positions onto the road network, obtaining additional road information such as road width and speed limit. This was done to align it with the Shenzhen dataset, which is composed of anonymous segment IDs. Valhalla ensures consistent results due to its reliance on OpenStreetMap.

A.2 Hyperparameter Settings

The OOD4TTE model includes 4 hyperparameters: two coefficients (α and β) in the risk function, the invariant ratio k, and the number of traffic concept V. The coefficients α and β are searched over {1e-1, 5e-1, 1, 5}. k is searched over {0.2, 0.4, 0.6, 0.8}, and V is searched over {4, 8, 16, 32}. Since the combinations of the four are very large (128 combinations), we adopt a group grid search strategy to find the suboptimal hyperparameters. Parameters in one group are grid searched first, and then sequentially searched across groups. Specifically, α and β are grouped together, and k and V form the other group. All the reported results are the average of five runs. The learning rate is fixed to 0.0005 with the batch size of 256.

B Discussion on OOD Evaluation Protocol

To construct a temporal OOD scenario, we partition each day into four time periods. A natural concern is whether such a partition genuinely induces distribution shift. To verify this, we employ the Classifier Two-Sample Test (C2ST) [19] to quantify the discrepancy between the training and test distributions within each period. Table 5 reports the resulting AUC scores. All values exceed 0.6 (reaching up to 0.787), confirming that the partition induces a substantial covariate shift well beyond what a simple random hold-out would yield.

Table 5: C2ST AUC across time periods. Higher values indicate stronger distribution shift between the training and test partitions.

Dataset	0:00–5:59	6:00–11:59	12:00–17:59	18:00–23:59
Shenzhen	0.651	0.677	0.646	0.787
Porto	0.610	0.759	0.603	0.644

C Additional Experiments

C.1 Performance on Route OOD

The OOD generalization performance in the route OOD scenario is shown in Table 6. We observe that the MAPE value on the shorter-half route subset is significantly higher than that of Informer + RevIN, despite our method achieving lower MAE. This is because MAPE is highly sensitive to small ground truth values, which are common in shorter routes. As a result, although our model yields smaller absolute errors, the relative errors measured by MAPE appear much larger.

C.2 Ablation study on Route OOD

Table 7 presents the results of the ablation study on route OOD. The introduced components improve the model's generalization performance.

Table 7: Ablation study in the route OOD scenarios on the Shenzhen and Porto datasets. The first row indicates the range of testing route length. Bold font indicates the best results.

Dataset	Method	0%–50%			50%–100%		
		MAE	MAPE	RMSE	MAE	MAPE	RMSE
Shenzhen	**OOD4TTE**	**110.23**	**24.93**	150.53	**287.50**	**17.84**	**491.40**
	$\alpha = 0$	110.36	25.33	**149.37**	329.08	20.12	547.02
	$\beta = 0$	123.63	30.42	160.50	337.73	20.54	563.90
	w/o Gat	114.30	26.80	153.10	300.60	18.65	510.80
	w/o Env	119.40	28.90	158.20	320.40	19.85	535.60
Porto	**OOD4TTE**	**112.09**	**43.48**	**196.70**	**285.98**	**18.83**	**746.57**
	$\alpha = 0$	112.89	46.70	209.66	289.98	19.46	755.80
	$\beta = 0$	149.93	63.66	226.40	292.45	19.98	764.00
	w/o Gat	120.60	52.20	205.10	292.40	19.90	770.30
	w/o Env	136.80	61.10	232.40	305.50	21.05	792.10

D Theoretical Justification of Selective Blocking

This section provides a formal explanation of why the selective blocking objective in Sec. 4.2.1 suppresses the confounding effect of path $C \to T$.

THEOREM 1 (PARTIAL-DISTANCE CRITERION AND CONDITIONAL DECORRELATION). *Let*

$$U = \mathcal{P}_{\mathbf{E}}(\Phi_1(\mathbf{C})), \quad V = \mathcal{P}_{\mathbf{E}}(\Psi_1(\mathbf{T})),$$

where $\mathcal{P}_{\mathbf{E}}(\cdot)$ is defined in Eq. 8. Then

$$D_{\mathbf{E}}(\Phi_1(\mathbf{C}), \Psi_1(\mathbf{T})) = \frac{|U \otimes V|}{\|U\|\|V\|} \in [0, 1].$$

In particular, $D_{\mathbf{E}}(\Phi_1(\mathbf{C}), \Psi_1(\mathbf{T})) = 0$ implies $U \otimes V = 0$, i.e., the residual dependency after conditioning on $\mathbf{E}$ vanishes in the partial-distance sense. Under Gaussian assumptions, this is equivalent to $\Phi_1(C) \perp\!\!\!\perp \Psi_1(T) \mid E$.

PROOF. The range $[0, 1]$ follows from the Cauchy–Schwarz inequality on the inner-product operator $\otimes$:

$$|U \otimes V| \le \sqrt{(U \otimes U)(V \otimes V)} = \|U\|\|V\|.$$

Hence the normalized quantity is bounded in $[0, 1]$. By construction, U and V are the residual components after removing the linear effect of $\mathbf{E}$, so $U \otimes V = 0$ means zero residual correlation under this metric. The Gaussian-equivalence statement follows from the standard result for partial correlation/partial distance [30]. □

THEOREM 2 (CONTRASTIVE OBJECTIVE ENFORCES SELECTIVE BLOCKING). *Define*

$$d^+ = D_{\mathbf{E}}(\Phi_1(\mathbf{C}), \Psi_1(\mathbf{T})), \quad d^- = D_{\mathbf{E}}(\Phi_1(\mathbf{C}'), \Psi_1(\mathbf{T})),$$

and $\mathcal{L}_{con} = \mathit{SoftPlus}(d^+ - d^-)$ in Eq. 10. For any $\epsilon > 0$, if $\mathcal{L}_{con} \le \epsilon$, then

$$d^+ - d^- \le \log(e^\epsilon - 1).$$

Therefore, minimizing $\mathcal{L}_{con}$ forces d^+ to be no larger than d^- up to a shrinking slack term, which suppresses conditional dependency of positive pairs and realizes selective blocking of $C \to T$.

PROOF. Since $\mathrm{SoftPlus}(x) = \log(1 + e^x)$ is strictly increasing,

$$\mathcal{L}_{con} \le \epsilon \implies \log(1 + e^{d^+ - d^-}) \le \epsilon \implies e^{d^+ - d^-} \le e^\epsilon - 1.$$

Taking logarithms gives

$$d^+ - d^- \le \log(e^\epsilon - 1).$$

As training decreases ϵ, the right-hand side decreases and eventually becomes negative, yielding $d^+ < d^-$. Thus, relative to shuffled negatives, the model is driven to reduce conditional dependence in the original $(\mathbf{C}, \mathbf{T})$ pair, corresponding to selective suppression of the confounding interaction path. □

Table 6: Performance under route-length OOD scenarios on the Shenzhen and Porto datasets. The column header indicates the percentile range of testing route length. All metrics are the lower the better. The best results are in bold and the second-best are underlined. $*$ denotes statistically significant improvement ($p < 0.05$) over the strongest baseline.

Backbone	Method	Shenzhen						Porto					
		0%–50%			50%–100%			0%–50%			50%–100%		
		MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE
WDR	ERM	116.81	25.04	154.46	345.04	22.06	516.74	128.68	48.93	208.95	353.45	24.11	814.48
	+ EIIL	115.20	25.40	155.10	320.50	20.50	505.20	125.80	40.20	209.20	340.50	23.50	810.10
	+ RevIN	114.66	25.71	156.38	299.81	18.97	499.73	124.45	31.55	209.66	333.93	21.96	807.83
	+ FOIL	126.14	26.95	159.72	357.65	22.31	529.86	255.78	85.77	316.86	356.64	36.89	781.60
	+ HRM	152.03	29.91	202.70	353.48	22.17	551.82	120.86	41.22	208.97	362.63	23.99	837.49
WDDRA	ERM	118.43	25.73	159.40	346.77	22.22	518.59	121.60	52.91	206.68	414.78	29.22	871.08
	+ EIIL	115.50	25.20	155.80	335.20	21.90	512.50	120.50	45.20	205.50	360.20	23.80	820.50
	+ RevIN	116.68	25.74	157.92	345.42	22.19	510.53	126.25	29.42	213.82	379.89	25.74	790.76
	+ FOIL	121.86	25.32	153.19	330.99	21.80	503.17	348.83	133.17	441.85	311.03	25.28	780.26
	+ HRM	114.15	25.01	152.88	345.86	22.03	509.05	117.91	38.99	204.77	341.01	21.96	778.47
ProbTTE	ERM	118.09	26.39	174.27	295.62	18.94	496.98	118.29	39.62	201.27	513.22	31.44	1034.85
	+ EIIL	117.50	26.50	170.20	295.10	18.80	495.50	120.10	35.50	205.10	450.20	29.50	950.50
	+ RevIN	158.30	32.79	218.08	294.74	18.27	493.70	130.83	26.09	225.06	504.38	31.00	2059.83
	+ FOIL	174.22	33.79	239.61	292.32	18.69	**489.17**	211.14	74.42	311.00	331.00	28.13	842.55
	+ HRM	116.90	26.99	168.70	293.94	18.45	494.15	122.92	27.90	209.00	320.98	26.61	830.26
Informer	ERM	234.16	48.82	313.76	610.55	44.36	751.03	225.02	72.55	311.10	635.29	53.67	929.61
	+ EIIL	245.50	45.20	330.10	450.20	30.50	650.50	160.20	40.50	250.20	600.50	50.20	910.50
	+ RevIN	261.68	43.19	347.64	380.98	24.68	583.55	131.66	**25.47**	221.66	661.14	56.78	973.15
	+ FOIL	263.38	62.24	337.06	500.26	32.91	692.86	291.83	80.67	356.53	525.96	40.39	825.24
	+ HRM	332.14	68.49	392.55	628.58	43.97	797.14	224.63	61.84	320.56	618.90	51.08	896.00
OOD4TTE		**110.23***	**24.93***	**150.53***	**287.50***	**17.84***	491.40	**112.09***	43.48	**196.70**	**285.98***	**18.83***	**746.57***