DRIFT: Regime-Aware Self-Supervised Learning, Zero-Inflated Forecasting, and Inference Fusion for Metro OD Demand under Temporal Distribution Shift

Hsiang-Chuan Chang[1], Tzu-Chia Huang[2], Chih-Yung Chang[2], Wen-Hwa Liao[3], Diptendu Sinha Roy[4]*

[1] Department of Transportation Management, Tamkang University, Taiwan
[2] Department of Computer Science and Information Engineering, Tamkang University, Taiwan
[3] Institute of Information and Decision Sciences, National Taipei University of Business, Taiwan
[4] Department of Computer Science and Engineering, National Institute of Technology Meghalaya, India
149190@o365.tku.edu.tw, 813410015@o365.tku.edu.tw, cychang@mail.tku.edu.tw, whliao@ntub.edu.tw, diptendu.sr@nitm.ac.in

Abstract

Accurate origin–destination (OD) demand forecasting is critical for safe and efficient metro operations, yet remains challenging due to temporal distribution shift and structural sparsity. The former destabilizes learned representations across demand regimes, while the latter biases models toward inactive OD pairs. Existing methods typically address these issues in isolation, limiting robustness in real-world settings. This paper proposes DRIFT, a unified framework that jointly handles representation instability and zero-inflated demand. DRIFT integrates three components: (1) regime-consistent self-supervised pre-training for stable spatio-temporal representations; (2) a zero-inflated AutoGate predictor that separates zero occurrence from flow magnitude; and (3) an adaptive inference-time fusion strategy to improve robustness under distribution mismatch. Experiments on 109 months of Taipei Metro OD data (28 stations, three regimes) under a strict no-look-ahead protocol show that DRIFT achieves MAE = 23.69, RMSE = 38.60, MAPE = 11.70%, and Peak Recall = 0.993, outperforming the strongest baseline by up to 9.3%. Ablation studies confirm that each component contributes distinct functionality, and that the core representation and sparsity-aware designs alone, without the inference-time fusion, already surpass all baselines. These results indicate that jointly modeling regime consistency and sparsity is effective for long-horizon OD forecasting.

Keywords: OD demand forecasting, Temporal distribution shift, Regime-aware self-supervised learning, Zero-inflated prediction, Inference-time fusion

1 Introduction

In urban metro operations, inaccurate origin–destination (OD) forecasts directly affect both prediction reliability and system stability. Missed demand surges lead to platform crowding and transfer bottlenecks, whereas persistent overestimation of weak OD pairs results in inefficient dispatching and unnecessary standby capacity. These risks make OD forecasting a central problem in metro operations [1-5]. Compared with aggregate inflow prediction, OD forecasting provides finer-grained movement patterns but significantly enlarges the output space, thereby intensifying data sparsity and increasing modeling difficulty.

Recent advances in spatio-temporal deep learning, including graph neural networks and Transformer-based models, have improved traffic forecasting by capturing spatial interactions and temporal dependencies [6-14]. However, these models are typically developed under the assumption of stable training and deployment distributions, which rarely holds in real-world metro systems.

In practice, metro demand evolves under long-horizon structural changes such as network expansion, policy adjustments, and external disruptions. These changes alter both demand magnitude and distributional structure, often producing highly sparse and zero-inflated OD matrices. When models trained under one regime are applied to another, predictive performance deteriorates due to representation mismatch. This issue is particularly severe in OD forecasting, where the output space scales quadratically and many OD pairs remain inactive for extended periods. As a result, long-horizon OD forecasting faces a coupled challenge: temporal distribution shift destabilizes learned representations, while structural sparsity biases dense predictors toward unreliable estimates on inactive OD pairs.

The Taipei Metro Tamsui–Xinyi Line exemplifies this setting. During the COVID-19 period, mean daily OD flow dropped from 4,821 to 2,107 trips, while the proportion of zero-valued OD pairs increased from 31% to 67%. Although demand partially recovered, the post-disruption distribution remained substantially different from the pre-pandemic regime [15]. Under such conditions, standard supervised models often fail to generalize, indicating that

*Corresponding Author: Chih-Yung Chang; Email: cychang@mail.tku.edu.tw
DOI: https://doi.org/10.70003/160792642026072704009

the key difficulty lies in jointly handling representation shift and output sparsity.

Existing studies address these issues separately. Transfer learning and domain adaptation improve cross-domain generalization [16-18], while self-supervised learning enhances representation quality [19-21]. Conversely, zero-inflated models reduce bias on inactive OD pairs [22-24]. However, none of these approaches jointly address representation instability and sparsity under sustained regime variation, which limits robustness in realistic long-horizon settings.

To address this gap, this paper proposes DRIFT, a regime-consistent and sparsity-aware framework for metro OD forecasting under temporal distribution shift. The key insight is that, although demand magnitude varies across time, the underlying structural interactions of the metro network remain relatively stable. DRIFT therefore decouples representation learning from demand magnitude estimation and aligns them under regime variation.

DRIFT consists of three coordinated components. First, a self-supervised stage learns regime-consistent spatio-temporal representations. Second, a zero-inflated forecasting module introduces a differentiable AutoGate mechanism that separates zero-occurrence tendency from flow magnitude. Third, an inference-time fusion strategy improves robustness under distribution mismatch.

The main contributions of this work are summarized as follows:

(1) A unified problem formulation that explicitly characterizes metro OD forecasting as a joint challenge of temporal distribution shift and zero-inflated structural sparsity, establishing their interaction as the primary driver of performance degradation in long-horizon settings.

(2) A regime-consistent self-supervised pre-training framework that stabilizes spatio-temporal representations across evolving demand regimes, combining masked flow reconstruction, temporal contrastive learning, and cross-regime structural consistency.

(3) A zero-inflated AutoGate forecasting mechanism that explicitly decouples zero-occurrence tendency from conditional flow magnitude through a differentiable dual-branch design, improving estimation reliability on sparse and inactive OD pairs.

(4) An adaptive inference-time fusion strategy that combines the drift-aware predictor with a supervised branch, improving point accuracy under partial distribution overlap while maintaining strong peak-demand identification performance.

2 Related Work

This section reviews prior research related to OD demand forecasting under temporal distributional shift.

2.1 Spatio-Temporal Graph Neural Networks for Traffic Forecasting

Graph neural networks are widely used in traffic forecasting because transportation systems are naturally represented as graphs. Representative models, including DCRNN, STGCN, Graph WaveNet, AGCRN, MTGNN, and PDFormer, learn spatial interactions and temporal dependencies from dynamic graphs [6-11]. Recent Transformer-based forecasters—including Autoformer, PatchTST, and iTransformer—extend this trend toward long-horizon sequence modeling [12-14]. Foundation and pretrained models such as One Fits All and MOMENT further advance the direction of transferable, general-purpose time series representations [17-18].

Despite strong benchmark performance, most of these models still assume relatively stable train–test distributions. When demand evolves over long horizons, the learned representations can become tied to historical patterns, leading to reduced generalization under demand regime shifts in metro OD matrices.

2.2 Origin–Destination Demand Forecasting

OD demand forecasting is harder than link-level prediction because the output space is quadratic in the number of stations, directionally asymmetric, and often sparse. Existing metro studies use convolutional-recurrent, physical–virtual, and multi-resolution temporal models to improve in-distribution accuracy [1-4].

However, many still optimize dense regression objectives, whereas zero-inflated OD matrices require models that distinguish zero occurrence from flow magnitude [22-24]. Without this distinction, models may produce biased estimates on inactive OD pairs, especially when sparsity increases under changing demand conditions.

2.3 Self-Supervised Learning for Spatio-Temporal Data

Self-supervised learning has become an effective way to learn transferable spatio-temporal representations without extra labels. Contrastive and predictive pretext tasks such as SimCLR, MoCo, TS-TCC, TS2Vec, and ST-SSL improve downstream forecasting by exploiting unlabeled structure [19-21, 25-26].

Even so, existing SSL methods rarely model regime transitions explicitly. As a result, learned representations may still drift when the underlying demand distribution changes. In addition, these methods are typically not designed to couple representation learning with zero-inflated OD estimation.

2.4 Transfer Learning and Domain Adaptation

Transfer learning and domain adaptation aim to generalize across cities or datasets through adaptation techniques and pretrained models [16-18]. Their emphasis, however, is cross-domain portability rather than within-network temporal distribution shift.

In metro systems, distributional changes often occur within the same network over time and are accompanied by persistent OD sparsity. Existing approaches do not explicitly address this combination of temporal shift and sparse demand structure.

2.5 Distributional Drift and Non-Stationarity

Distributional drift violates the stationarity assumption by changing the joint distribution over time [27]. In transportation systems, pandemic-era disruptions provide a concrete example of abrupt and persistent regime change [15].

Although recent work on non-stationary or uncertainty-aware forecasting improves robustness [12, 28-29], these methods usually do not account for the directional asymmetry and zero inflation of metro OD matrices.

In summary, prior work addresses representation learning, sparsity modeling, and distributional robustness as largely separate problems. Metro OD forecasting under temporal distribution shift, however, involves both challenges simultaneously. No existing method explicitly models their interaction, leaving a clear gap in robustness when evolving demand regimes and zero-inflated sparsity co-occur.

3 Problem Definition and Notation

This section formalizes the metro OD forecasting problem under temporal distribution shift and introduces all symbols used in the sequel.

3.1 OD Demand Sequence

Let $G = (V, E)$ denote a directed metro network, where $V = \{v_1, ..., v_N\}$ is the set of stations and $E \subseteq V \times V$ is the set of directed connections. Let $N = |V|$ denote the number of stations.

Let $T \in \mathbb{N}$ denote the total number of time steps, and let $\mathcal{T} = \{1, 2, ..., T\}$ denote the corresponding set of discrete time indices. Let $M \in \mathbb{N}$ denote the total number of months, and let $\mathcal{M} = \{1, 2, ..., M\}$ denote the corresponding set of monthly indices.

Let

$$m: \mathcal{T} \to \mathcal{M} \tag{1}$$

denote the mapping that assigns each time step $t \in \mathcal{T}$ to its corresponding month $m(t)$. For any ordered station pair $(i, j) \in \{1, ..., N\} \times \{1, ..., N\}$, let $x_t^{(i,j)} \in \mathbb{R} \geq 0$ denote the passenger flow from origin station v_i to destination station v_j at time step t.

Let

$$X_t = [x_t^{(i,j)})]_{i,j=1}^{N} \in \mathbb{R}^{N \times N} \tag{2}$$

denote the OD demand matrix at time step t. Let $\mathcal{X} = \{X_t\}_{t \in \mathcal{T}}$ denote the full OD demand sequence.

3.2 Demand Regimes

Long-horizon metro demand may evolve under external disruptions, policy changes, and gradual behavioral shifts. Let $K \in \mathbb{N}$ denote the number of latent demand regimes, and let $\mathcal{K} = \{1, ..., K\}$ denote the set of regime indices.

Let $\{\mathcal{T}_k\}_{k \in \mathcal{K}}$ denote a partition of $\mathcal{T}$ into K disjoint contiguous subsets such that

$$\mathcal{T} = \bigcup_{k=1}^{k} \mathcal{T}_k, \ \mathcal{T}_k \cap \mathcal{T}_\ell = \varnothing \quad \text{for all } k, \ell \in \mathcal{K} \text{ with } k \neq \ell. \tag{3}$$

Each subset $\mathcal{T}_k$ corresponds to one demand regime. Here, k and ℓ are regime indices in $\mathcal{K}$.

Let $\tilde{r}: \mathcal{M} \to \mathcal{K}$ denote the pseudo-regime assignment defined at the monthly level. Let $r: \mathcal{T} \to \mathcal{K}$ denote the induced time-step regime function, defined by

$$r(t) = \tilde{r}(m(t)). \tag{4}$$

Let

$$\mathcal{X}^{(k)} = \{X_t | \ t \in \mathcal{T}, \ r(t) = k\} \tag{5}$$

denote the OD subsequence associated with regime k. Equivalently, $\mathcal{X}^{(k)} = \{X_t | \ t \in \mathcal{T}_k\}$.

In this paper, a demand regime is operationally defined as a maximal contiguous time interval over which the monthly DDI remains relatively stable. Regime boundaries are identified by applying penalized change-point detection to the monthly distributional drift index sequence $\{\mathrm{DDI}_m\}_{m=1}^{M}$ defined in Section 3.3. Specifically, the PELT (Pruned Exact Linear Time) algorithm with an ℓ_2 cost function is used, and the number of regimes K is selected using a Bayesian information criterion (BIC) style penalty $\mathcal{C} + \lambda K \log M$, where $\mathcal{C}$ denotes the segmentation cost and M is the number of months.

Each month first receives a pseudo-regime label, and every time step within that month inherits the same label $r(t)$. The resulting labels are used only for conditioning, representation regularization, and evaluation stratification, and are not directly optimized in the forecasting objective or treated as prediction targets.

3.3 Sparsity and Drift Measures

Let $[\cdot]$ denote the Iverson bracket, which evaluates to 1 if the enclosed condition holds and 0 otherwise. The sparsity rate at time step $t \in \mathcal{T}$ is defined as

$$\rho_t = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left[x_t^{(i,j)} = 0 \right], \tag{6}$$

which measures the proportion of zero-valued OD entries in X_t.

Let

$$\mathcal{T}^{(m)} = \{t \in \mathcal{T} | \ m(t) = m\}, \ m \in \mathcal{M}, \tag{7}$$

denote the set of time steps associated with month m.

The monthly average OD flow is defined as

$$\bar{\mu}_m = \frac{1}{\left| \mathcal{T}^{(m)} \right| N^2} \sum_{t \in \mathcal{T}^{(m)}} \sum_{i=1}^{N} \sum_{j=1}^{N} x_t^{(i,j)}. \tag{8}$$

Let $\varepsilon > 0$ denote a small constant for numerical stability. The distributional drift index (DDI) is defined for $m \in \{2, ..., \mathcal{M}\}$ as

$$\mathrm{DDI}_m = \frac{\left|\bar{\mu}_m - \bar{\mu}_{m-1}\right|}{\bar{\mu}_{m-1} + \varepsilon} \tag{9}$$

The first month does not admit a previous reference and is therefore excluded from the DDI computation.

The distributional drift index (DDI) serves as a low-frequency proxy statistic for characterizing temporal distributional change at the monthly level. It is used for pseudo-regime identification and descriptive analysis rather than as a direct model input or a full estimate of the OD distribution.

3.4 Forecasting Task

Let $L \in \mathbb{N}$ denote the look-back window length.

$$\mathcal{T}_{\text{valid}} = \{L, L+1, ..., T-1\} \tag{10}$$

denote the set of valid prediction indices.
For each $t \in \mathcal{T}_{\text{valid}}$, let

$$X_{t-L+1:t} = \text{stack}(X_{t-L+1}, ..., X_t) \in \mathbb{R}^{L\times N\times N} \tag{11}$$

denote the input OD tensor constructed from the past L time steps.

Let

$$f_\theta : \mathbb{R}^{L\times N\times N} \to \mathbb{R}^{N\times N} \tag{12}$$

denote the forecasting function parameterized by θ. The prediction is given by

$$\hat{X}_{t+1} = f_\theta\left(X_{t-L+1:t}\right), t \in \mathcal{T}_{\text{valid}}. \tag{13}$$

The formulation follows a single-step forecasting setting. The structured OD tensor representation is preserved to retain spatial and temporal dependencies without flattening.

3.5 Learning Objective and Problem Requirements

Let

$$\mathcal{L}_{\text{pred}} : \mathbb{R}^{N\times N} \times \mathbb{R}^{N\times N} \to \mathbb{R} \geq 0 \tag{14}$$

denote a prediction loss defined between an estimated OD matrix and its ground truth.

The learning objective is defined as

$$\theta^* = \underset{\theta}{\text{argmin}}\, \mathbb{E}_{t\in\mathcal{T}_{\text{valid}}}\left[\mathcal{L}_{\text{pred}}\left(f_\theta\left(X_{t-L+1:t}\right), X_{t+1}\right)\right] \tag{15}$$

An effective predictor in this setting should satisfy three requirements:

(1) It produces accurate next-step OD predictions across all station pairs.
(2) It remains robust under temporal distribution shift.
(3) It accounts for the zero-inflated sparsity structure of OD matrices.

These requirements capture the coupled challenge of representation instability under regime variation and estimation bias induced by sparse observations. The proposed method in Section 4 is designed to address these aspects jointly.

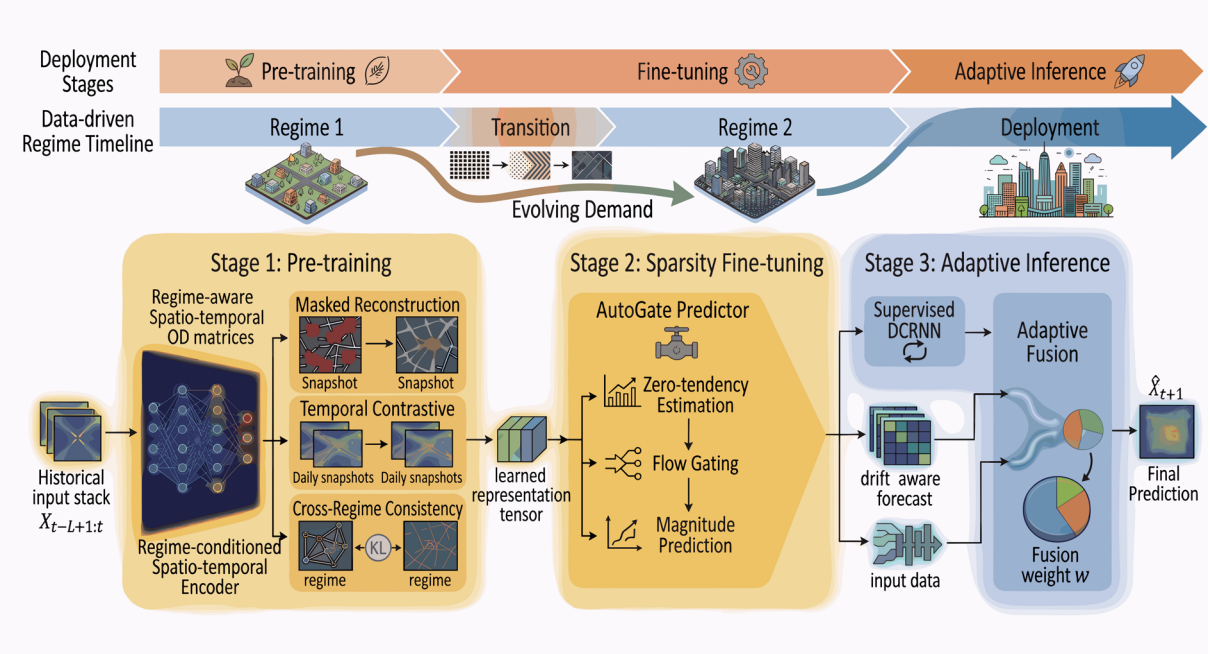

Figure 1. Framework overview of DRIFT (Top: regime timeline. Bottom: three-stage pipeline.)

4 Methodology

This section presents the proposed DRIFT framework. Recall from Section 3 that the forecasting task is to learn a mapping from the historical OD tensor $X_{t-L+1:t}$ to the next-step OD matrix X_{t+1} under temporal distribution shift and zero-inflated sparsity. The proposed method instantiates this mapping through three coordinated stages: regime-consistent representation learning, zero-inflated forecasting, and regime-adaptive inference.

4.1 Framework Overview

Recall that $G = (V, E)$ denotes the metro network, $N = |V|$ denotes the number of stations, $X_t: \mathbb{R}^{N\times N}$ denotes the OD matrix at time step t, and $r(t)$ denotes the inferred regime label defined from the monthly pseudo-regime assignment in Section 3.2.

Let θ_e , θ_d, and ϕ denote the parameters of the regime-conditioned encoder, the drift-aware zero-inflated predictor, and the supervised forecasting branch, respectively. DRIFT decomposes the overall predictor into three stages.

In Stage 1, the encoder learns regime-consistent spatio-temporal representations from unlabeled OD observations through self-supervised objectives. In Stage 2, these representations are converted into a next-step OD forecast through a zero-inflated predictor that separates zero-occurrence tendency from conditional flow magnitude. In Stage 3, the drift-aware forecast is fused with the output of a supervised forecaster to improve deployment robustness when the test distribution partially overlaps with the dominant training regime.

Formally, the final prediction is written as

$$\hat{X}_{t+1} = f_\theta\,(X_{t-L+1:t}), \tag{16}$$

where f_θ is realized by the three-stage architecture described below.

4.2 Regime-Consistent Representation Learning

4.2.1 Spatio-Temporal Encoder

Let $\tilde{A} \in \mathbb{R}^{N\times N}$ denote the normalized graph propagation matrix associated with G. Let $d \in \mathbb{N}$ denote the latent dimension. Unless otherwise stated, all products in this subsection denote standard matrix multiplication.

For each prediction anchor $t \in \mathcal{T}_{\text{valid}}$, and for each historical time step $s \in \{t - L + 1, ..., t\}$, let

$$U_s = \left[\, X_s \,\|\, X_s^{\mathrm{T}} \,\right] \in \mathbb{R}^{N\times 2N} \tag{17}$$

denote the station-feature matrix obtained by concatenating outgoing and incoming OD profiles, where $\|$ denotes column-wise concatenation. Each row of U_s represents the outgoing and incoming flow signatures of one station.

Let $W_{\text{in}} \in \mathbb{R}^{2N\times d}$ denote the input projection matrix. The initial node representation at time step s is

$$H_s^{(0)} = U_s W_{\text{in}} \in \mathbb{R}^{N\times d} \tag{18}$$

Let $K_g \in \mathbb{N}$ denote the number of graph propagation layers. For $\ell = 0, ..., K_g - 1$, the spatial propagation is defined as

$$H_s^{(\ell+1)} = \tilde{A} H_s^{(\ell)} \tag{19}$$

Let the superscript p denote the spatially encoded representation. The spatial representation at time step s is defined by layer averaging:

$$H_s^{p} = \frac{1}{K_g + 1} \sum_{\ell=0}^{K_g} H_s^{(\ell)} \in \mathbb{R}^{N\times d} \tag{20}$$

Linear propagation, following the LightGCN design principle [30], is adopted to preserve topology-aware information flow without introducing additional nonlinear distortion in the graph aggregation step. For two regime-specific inputs a and b, the propagation is non-expansive under normalized adjacency:

$$\| H_a^{(\ell+1)} - H_b^{(\ell+1)} \|_F \le \| \tilde{A} \|_2 \| H_a^{(\ell)} - H_b^{(\ell)} \|_F . \tag{21}$$

For normalized propagation, $\|\tilde{A}\|_2 \le 1$. This gives an inductive bias toward controlled cross-regime representation shift before the explicit consistency term is imposed.

Let ψ_{temp} denote the temporal encoder, where the subscript temp indicates temporal encoding. In the current implementation, ψ_{temp} is instantiated as a standard Transformer encoder [31]. It maps the historical sequence

$$\left\{ H_{t-L+1}^{p}, \ldots, H_t^{p} \right\} \tag{22}$$

to the temporally contextualized representation

$$H_t = \Psi_{\text{temp}} \left(H_{t-L+1}^{p}, \ldots, H_t^{p} \right) \in \mathbb{R}^{N\times d} . \tag{23}$$

Let $d_r \in \mathbb{N}$ denote the regime-embedding dimension. Let $e_{r(t)} \in \mathbb{R}^{d_r}$ denote the trainable embedding associated with regime $r(t)$. Let $W_r \in \mathbb{R}^{d\times d_r}$ denote the regime projection matrix, and let $\mathbf{1}_N \in \mathbb{R}^N$ denote the all-ones vector.

Let the superscript cond denote regime-conditioned representation. Additive regime conditioning is defined as

$$H_t^{\text{cond}} = H_{\text{t}} + 1_N \left(W_r e_{r(t)} \right)^{\mathrm{T}} \in \mathbb{R}^{N\times d} . \tag{24}$$

Let $W_z \in \mathbb{R}^{d\times 2d}$ denote the pairwise projection matrix. For each ordered OD pair (i, j), the pair representation is

$$z_t^{(i,j)} = W_z\left[H_t^{\text{cond}}(i,:) \,\|\, H_t^{\text{cond}}(j,:)\right] \in \mathbb{R}^d \tag{25}$$

4.2.2 Self-Supervised Pre-Training Objectives

Stage 1 optimizes a three-term self-supervised objective:

$$\mathcal{L}_{\text{ssl}} = \alpha\mathcal{L}_{\text{mfr}} + \beta\mathcal{L}_{\text{tcl}} + \gamma\mathcal{L}_{\text{crc}}, \tag{26}$$

where $\alpha, \beta, \gamma \geq 0$ and $\alpha + \beta + \gamma = 1$.

Let the subscripts mfr, tcl, and crc denote masked flow reconstruction, temporal contrastive learning, and cross-regime consistency, respectively.

Masked Flow Reconstruction

Let Ω_{mask} denote the set of masked OD entries sampled from the input tensor. Let the superscript rec denote reconstructed quantity. For each masked entry $(s, i, j) \in \Omega_{\text{mask}}$, let $\hat{x}_{s,\text{rec}}^{(i,j)}$ denote the reconstructed value of $x_s^{(i,j)}$. The reconstruction loss is

$$\mathcal{L}_{\text{mfr}} = \frac{1}{|\Omega_{\text{mask}}|}\sum_{(s,i,j)\in\Omega_{\text{mask}}}\left(x_s^{(i,j)} - \hat{x}_{s,\text{rec}}^{(i,j)}\right)^2 \tag{27}$$

Temporal Contrastive Learning

Let $\mathcal{P}$ denote the set of anchor-positive pairs, and for each anchor a, let $\mathcal{N}(a)$ denote the associated negative set. Let $\text{sim}(\cdot,\cdot)$ denote cosine similarity, and let $\tau > 0$ denote the temperature parameter. The temporal contrastive loss is

$$\mathcal{L}_{\text{tcl}} = -\frac{1}{|\mathcal{P}|}\sum_{(a,p)\in\mathcal{P}}\log \frac{\exp\left(\frac{\text{sim}(z_a, z_p)}{\tau}\right)}{\exp\left(\frac{\text{sim}(z_a, z_p)}{\tau}\right) + \sum_{n\in\mathcal{N}(a)}\exp\left(\frac{\text{sim}(z_a, z_n)}{\tau}\right)} \tag{28}$$

Cross-Regime Consistency

Let $\text{softmax}_{\text{row}}(\cdot)$ denote row-wise softmax. For each time step t, define the relational response matrix

$$S_t = \text{softmax}_{\text{row}}\left(\frac{H_t^{\text{cond}}\left(H_t^{\text{cond}}\right)^{\text{T}}}{\sqrt{d}}\right) \in \mathbb{R}^{N\times N}. \tag{29}$$

Recall that $\mathcal{T}_k$ denotes the set of time steps associated with regime k. The regime-level relational response is

$$R^{(k)} = \frac{1}{|\mathcal{T}_k|}\sum_{t\in\mathcal{T}_k} S_t \in \mathbb{R}^{N\times N}. \tag{30}$$

Let $D_{\text{KL}}(\cdot\|\cdot)$ denote row-wise Kullback–Leibler divergence averaged over rows.

$$\mathcal{L}_{\text{crc}} = \frac{1}{K(K-1)}\sum_{k,\ell\in\mathcal{K},k\neq\ell} D_{\text{KL}}\left(R^{(k)} \,\|\, R^{(\ell)}\right). \tag{31}$$

This objective does not enforce identical latent embeddings across regimes. Instead, it encourages consistent relational organization across pseudo-regimes, which is more compatible with the assumption that topology remains stable while demand magnitude and activity patterns change.

4.3 Sparsity-Aware Forecast Generation

4.3.1 Dual-Branch Prediction and AutoGate

AutoGate follows a two-stage design analogous to zero-inflated models. The gate $g_v \in [0,1]$ estimates whether an OD pair is active, while the magnitude head μ_v predicts flow conditional on activity. The multiplicative form $\hat{y}_v = g_v \cdot \mu_v$ suppresses predictions when $g_v \to 0$, eliminating spurious outputs for inactive pairs, and reduces to the magnitude estimate when $g_v \to 1$, preserving accuracy on active flows.

Stage 2 converts the learned pair representation into a next-step OD forecast under zero-inflated sparsity. Recall from Section 3.3 that $[\cdot]$ denotes the Iverson bracket. Let $y_{t+1}^{(i,j)} = [x_{t+1}^{(i,j)} = 0]$ denote the zero indicator of the target entry, and let $a_{t+1}^{(i,j)} = [x_{t+1}^{(i,j)} > 0]$ denote the active-entry indicator.

Let the subscripts cls and reg denote classification and regression, respectively. Let

$$h_{\text{cls}} : \mathbb{R}^d \to \mathbb{R} \text{ and } h_{\text{reg}} : \mathbb{R}^d \to \mathbb{R} \geq 0 \tag{32}$$

denote the zero-classification head and the conditional-magnitude regression head. The zero-tendency probability is

$$p_{t+1,\text{zero}}^{(i,j)} = \sigma\left(h_{\text{cls}}\left(z_t^{(i,j)}\right)\right), \tag{33}$$

where $\sigma(\cdot)$ is the sigmoid function.

The conditional magnitude prediction is

$$x_{t+1,\text{val}}^{(i,j)} = h_{\text{reg}}\left(z_t^{(i,j)}\right). \tag{34}$$

Let $s \geq 0$ denote the gate-scaling coefficient. Let

$$\text{clip}(u,0,1) = \min\left\{1, \max\left\{0, u\right\}\right\}, \tag{35}$$

denote clipping to [0,1]. The gate value is defined as

$$g_{t+1}^{(i,j)} = \text{clip}\left(1 - s\cdot p_{t+1,\text{zero}}^{(i,j)}, 0, 1\right), \tag{36}$$

Let the superscript drift denote the drift-aware branch output. The final drift-aware pairwise prediction is obtained by multiplicative gating:

$$x_{t+1,\text{drift}}^{(i,j)} = g_{t+1}^{(i,j)} \cdot x_{t+1,\text{val}}^{(i,j)} \tag{37}$$

The matrix-level drift-aware forecast is

$$\hat{x}_{t+1}^{\text{drift}} = \left[x_{t+1,\text{drift}}^{(i,j)} \right]_{i,j=1}^{N} \in \mathbb{R}^{N\times N}. \tag{38}$$

The multiplicative gate in Eq. (37) follows a zero-inflated mixture interpretation. The classification branch estimates the zero-tendency probability p_{ij}, while the regression branch estimates the conditional magnitude $\hat{r}_{ij}$ assuming that the OD pair is active. The gate g_{ij} = clip(1 − sp_{ij}) acts as a continuous relaxation of the active-pair indicator. When p_{ij} is large, g_{ij} shrinks toward zero and suppresses spurious nonzero predictions on inactive OD pairs. When p_{ij} is small, g_{ij} approaches one and preserves the magnitude estimate for active flows. Unlike an additive correction, the multiplicative form does not inject a separate offset into $\hat{r}_{ij}$, making it better suited to OD flows whose magnitudes vary substantially across station pairs and regimes.

4.3.2 Fine-Tuning Objective

The encoder is not frozen in Stage 2. Instead, the encoder and the zero-inflated predictor are optimized jointly. The zero-classification loss is

$$\begin{aligned}\mathcal{L}_{\text{cls}} = &-\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_{t+1}^{(i,j)} \log p_{t+1,\text{zero}}^{(i,j)} \\ &+\left(1-y_{t+1}^{(i,j)}\right)\log\left(1-p_{t+1,\text{zero}}^{(i,j)}\right)\end{aligned} \tag{39}$$

The active-entry regression loss is

$$\mathcal{L}_{\text{reg}} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} a_{t+1}^{(i,j)}\left(\hat{x}_{t+1,\text{val}}^{(i,j)} - x_{t+1}^{(i,j)}\right)^2}{\sum_{i=1}^{N}\sum_{j=1}^{N} a_{t+1}^{(i,j)} + \varepsilon}, \tag{40}$$

where $\varepsilon > 0$ avoids division by zero.

To preserve temporal discrimination during supervised adaptation, the contrastive term is retained as a regularizer. The Stage 2 objective is

$$\mathcal{L}_{\text{ft}} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{tcl}}\mathcal{L}_{\text{tcl}}, \tag{41}$$

where $\lambda_{\text{cls}}, \lambda_{\text{reg}}, \lambda_{\text{tcl}} \geq 0$.

4.4 Adaptive Inference-Time Fusion

Although the drift-aware predictor improves robustness under temporal shift, it does not always achieve the best point accuracy when the deployment condition still shares substantial overlap with the dominant training regime.

Let the superscript sup denote the supervised branch output. Let $f_{\varnothing}^{\text{sup}}: \mathbb{R}^{L\times N\times N} \to \mathbb{R}^{N\times N}$ denote the supervised forecasting branch. In the current study, this branch is instantiated as DCRNN under the same chronological protocol. Its output is

$$\hat{X}_{t+1} = (1-w)\,\hat{X}_{t+1}^{\text{drift}} + w\hat{X}_{t+1}^{\text{sup}},\ w\in[0,1] \tag{42}$$

Here, w is an inference-time blending coefficient selected on the validation split ($w = 0$ recovers the standalone drift-aware predictor; $w = 1$ recovers pure supervised DCRNN). The selected value w = 0.85 reflects that the deployment test period partially overlaps with the dominant training regime, so the supervised branch contributes substantially to point accuracy while the drift-aware branch anchors cross-regime robustness. This calibration step is intentionally kept separate from Stage 1 representation learning and Stage 2 zero-inflated estimation. For fairness, the supervised DCRNN branch used in the fusion follows the same chronological split, look-back window L = 12, optimizer, gradient clipping, early-stopping criterion, and validation-based model-selection protocol as the standalone DCRNN baseline. The fusion weight w is selected only on the validation split by minimizing validation RMSE, and no test-period observations are used for branch training, model selection, or fusion calibration.

4.5 Stagewise Optimization and Design Rationale

The three-stage workflow is illustrated in Figure 1. Stage 1 trains the regime-conditioned encoder using the self-supervised objectives in Section 4.2.2, with pseudo-regime labels serving only for conditioning rather than as prediction targets. The goal is to stabilize representation geometry before supervised fine-tuning begins. Stage 2 jointly adapts the encoder and the AutoGate predictor, reducing bias on inactive OD pairs while maintaining accuracy on active flows. Stage 3 combines the drift-aware predictor with a supervised DCRNN branch via adaptive inference-time fusion, which accounts for the partial regime overlap between training and deployment conditions. These three stages address the three core challenges from Section 3.5: representation instability, zero-inflated bias, and inference-time calibration.

5 Results and Discussion

5.1 Experimental Setup

This subsection specifies the dataset, chronological split, baselines, implementation settings, and evaluation metrics used throughout Section 5.

For the Taipei Metro corpus, pseudo-regimes are obtained using the procedure defined in Section 3.2. The resulting labels are used only for training-time conditioning, stratified evaluation, and conditional analysis, and are not future supervision targets. The

detected regime boundaries are visualized in Figure 2(a) and supported by the month-on-month DDI pattern in Figure 2(d). Boundary perturbation tests with ±1 month changes result in less than 1.2% relative variation in held-out RMSE, indicating that DRIFT is not sensitive to fine-grained boundary placement.

Table 1 summarizes the dataset and evaluation protocol. Figure 2 summarizes the data characteristics and the detected regime boundaries. Panel (A) reports monthly mean daily OD flow. Panel (B) reports the sparsity rate ξ_t, defined in Section 3.3 as the fraction of zero OD entries. The adjacent heatmaps place destination stations on the horizontal axis and origin stations on the vertical axis, where darker cells indicate larger flows and lighter cells indicate near-zero demand. The substantially lighter COVID-period heatmap is consistent with the increase of ξ_t from 0.31 to 0.67. Panel (C) compares the regime-wise time-of-day demand profiles. Panel (D) plots the month-on-month distributional drift index, where elevated values around the detected boundaries indicate sustained temporal drift.

This paper compares ten representative baselines from four families. These include recurrent graph models, namely DCRNN [6] and AGCRN [9], convolutional graph models, namely STGCN [7], Graph WaveNet [8], and MTGNN [10], Transformer-based forecasters, namely STAEformer [32] and Autoformer [12], self-supervised representation-learning baselines, namely TS-TCC [19] and ST-SSL [21], and a transfer-learning baseline, namely CityTrans [16]. Unless otherwise noted, all methods use the same input horizon, chronological split, and validation-driven model-selection protocol. Table 2 summarizes the compared method families and the key settings of DRIFT. Note that Table 3 and Table 5 report performance on the single held-out chronological test split, while Table 4 reports means over 10 temporal cross-validation folds; absolute values therefore differ slightly between the two settings, though relative rankings remain consistent.

The same chronological protocol is used for all models, and DRIFT is optimized with Adam [33]. Performance is evaluated using MAE, RMSE, MAPE, and Peak Recall. MAE and RMSE quantify absolute forecasting errors, with RMSE treated as the primary metric because large OD deviations are operationally more costly. MAPE is computed only on active OD pairs to avoid division-by-zero artifacts under sparse OD matrices. Peak Recall is computed per time step as the fraction of correctly identified top-K_p OD pairs, where K_p is fixed to 10 according to the 95th-percentile peak active-pair count on the validation split. The notation K_p is used to distinguish the peak-count parameter from the regime number K. For cross-regime robustness, later sections additionally report Drift Gap, defined as $(RMSE_{R2}-RMSE_{R1})/RMSE_{R1}$, and Recovery Gap, defined as $(RMSE_{R3}-RMSE_{R1})/RMSE_{R1}$.

Table 1. Dataset and evaluation protocol (OD = origin—destination)

Dataset	Period	Spatio dimension	Protocol	Role
Taipei Metro Tamsui-Xinyi Line	2017/01 -2026/01	28 stations; 784 directed OD pairs	Strict chronological split at the R1/R2 boundary. Training and validation use R1 only, while R2 and R3 are used for held-out testing. No future observations are used for fitting or model selection, which avoids temporal leakage.	Main benchmark

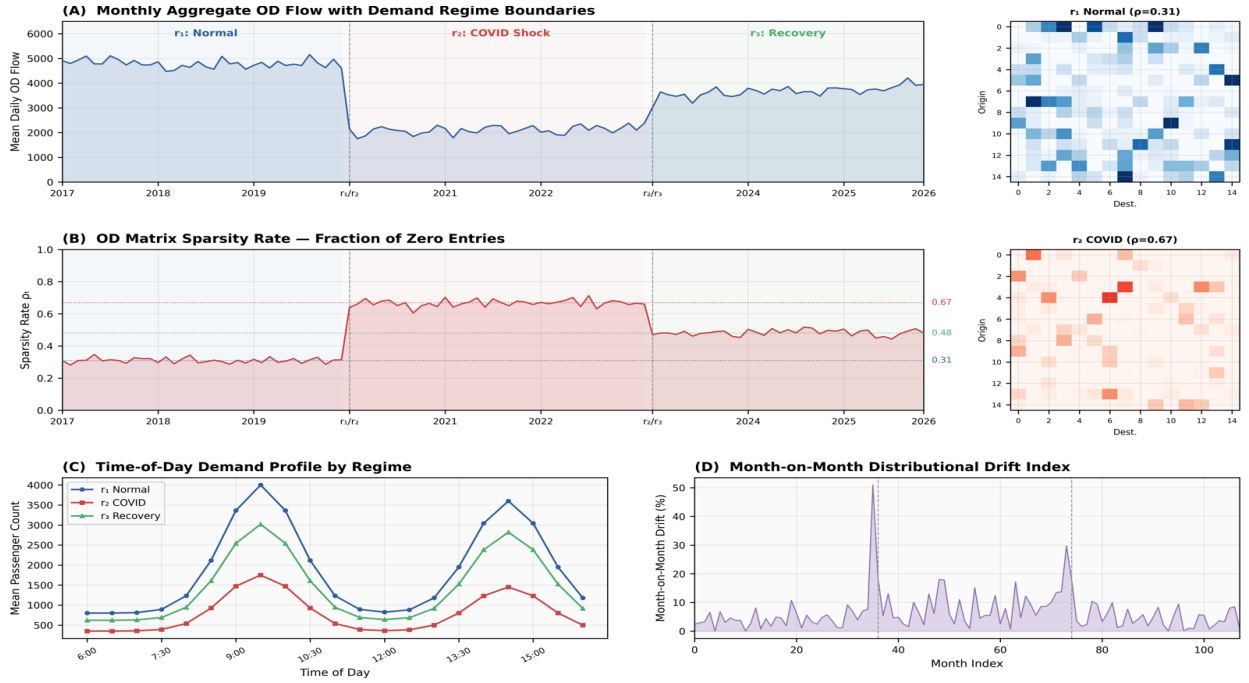

Figure 2. Data characteristics of the Taipei Metro corpus

Table 2. Compared baselines and key implementation settings of DRIFT

Item	Specification
Compared baselines	DCRNN [6], STGCN [7], Graph WaveNet [8], AGCRN [9], MTGNN [10], STAEformer [32], Autoformer [12], TS-TCC [19], ST-SSL [21], and CityTrans [16]
Baseline coverage	Recurrent graph, convolutional graph, Transformer-based, self-supervised, and transfer-learning paradigms
Look-back window	$L = 12$ time steps
Model input	Stacked OD tensor $X_{t-L+1:t} \in \boldsymbol{R}^{L \times N \times N}$; no global flattening before encoding
Forecast target/output	Single-step OD forecasting with target X_{t+1} and output X_{t+1}
Spatial module	LightGCN-style propagation with 2 layers and latent dimension $d = 64$
Temporal module	Transformer encoder with 4 attention heads
Optimization	Adam with gradient clipping at norm 5.0
Model selection	Early stopping based on validation RMSE under the same chronological protocol
Inference settings	AutoGate scale $s = 1.0$ and fusion weight $w = 0.85$
Hardware	NVIDIA RTX 4090 32GB

Table 3. Overall performance on the Taipei Metro chronological test set (Best in **bold**; second-best in *italics*.)

Method	MAE↓	RMSE↓	MAPE (%)↓	Peak Recall↑
DCRNN [6]	29.14	48.66	14.76	0.944
STGCN [7]	30.69	50.90	15.50	0.938
Graph WaveNet [8]	28.26	47.20	14.29	0.948
AGCRN [9]	27.66	46.01	13.97	0.951
MTGNN [10]	26.67	44.31	13.48	0.954
STAEformer [32]	26.24	43.72	13.25	0.956
Autoformer [12]	28.40	47.36	14.37	0.949
TS-TCC [19]	27.18	45.24	13.70	0.952
ST-SSL [21]	*25.52*	*42.56*	*12.90*	*0.961*
CityTrans [16]	26.19	43.64	13.26	0.957
DRIFT	**23.69**	**38.60**	**11.70**	**0.993**
Improv.over ST-SSL	-7.2%	-9.3%	-9.3%	+3.3%

Table 4. Significance test of DRIFT vs. ST-SSL over 10 chronological folds

Metric	DRIFT mean ± std	ST-SSL mean ± std	$p_{t\text{-}test}$	$p_{Wilcoxon}$	Cohen's d	Effect
MAE	23.95±1.18	27.63±1.52	1.8×10^{-5}	0.002	2.59	Large
RMSE	39.03±1.94	46.08±2.68	5.9×10^{-6}	0.002	2.93	Large
MAPE	11.83±0.58	13.97±0.81	6.4×10^{-6}	0.002	2.89	Large

5.2 Overall Performance and Event-Shock Evaluation

5.2.1 Overall Performance

Table 3 reports the aggregate results on the full chronological test set. DRIFT ranks first on all four metrics, with a 9.3% RMSE reduction and a 3.3% relative Peak Recall improvement over ST-SSL.

The simultaneous improvement in RMSE and Peak Recall confirms that error reduction is not obtained at the expense of surge detection—DRIFT improves both average accuracy and peak-demand identification.

5.2.2 Statistical Significance

Since ST-SSL is the strongest baseline in Table 3, the significance analysis focuses on the DRIFT vs. ST-SSL comparison. Table 3 reports results on a single held-out chronological split, while Table 4 summarizes mean ± std over 10 temporal cross-validation folds under the same no-look-ahead protocol. Absolute values differ slightly, but relative rankings remain consistent.

Table 4 reports the results of the paired t-test and the Wilcoxon signed-rank test. All three error metrics are significant under both tests ($p < 0.01$), with Cohen's d values ranging from 2.59 to 2.93, indicating large effect sizes. These results confirm that the performance gain of DRIFT is statistically robust across repeated chronological evaluations.

5.2.3 Event-Shock Case Study

Two unseen event-shock days are examined: New Year's Eve (2023-12-31) and a secondary shock (2023-08-10), representing a destination-concentrated surge and a hub-centered redistribution, respectively. Figure 3 and Figure 4 present station-level temporal curves, peak-hour Top-K_p OD flows, and spillover heatmaps.

For New Year's Eve, demand concentrates toward Taipei 101/WTC. DRIFT recovers 8 of the top-10 peak-hour OD pairs and preserves the main destination hierarchy and long-range spillover structure. For the 2023-08-10 event, centered on Taipei Main, DRIFT retrieves 7 of the top-10 OD pairs despite a different spatial pattern.

Across both cases, DRIFT tracks peak timing accurately and maintains consistent OD-level ranking and spatial redistribution under distinct shock morphologies, indicating that its advantage extends beyond average error reduction to robust peak-demand structure preservation.

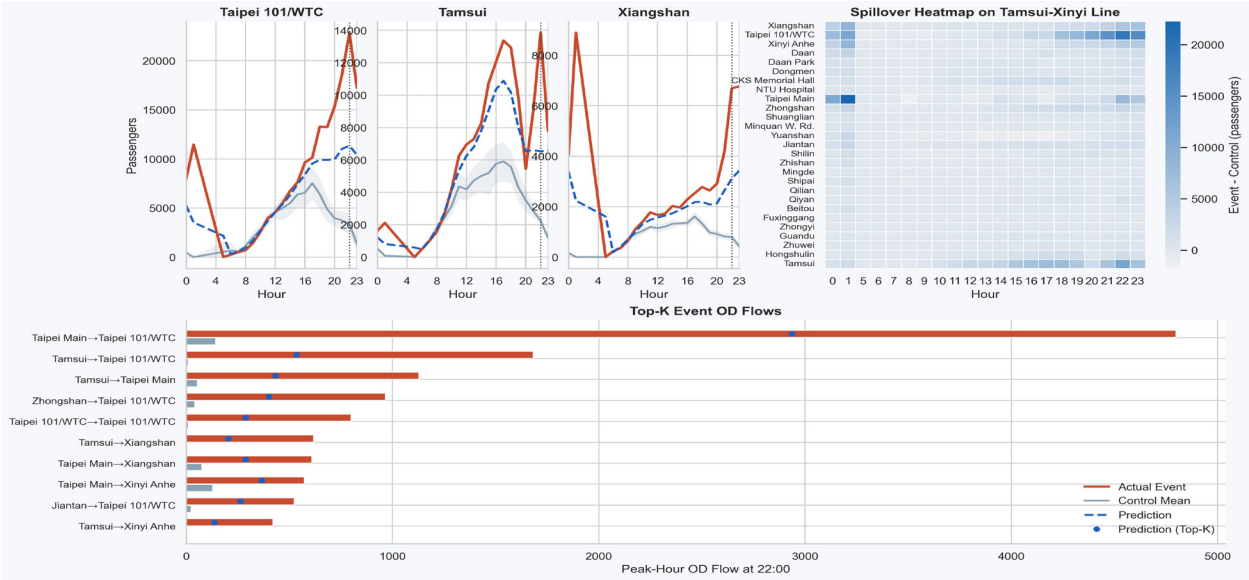

Figure 3. Held-out New Year's Eve case on 2023-12-31, including station-level temporal curves, peak-hour Top-K_p OD flows, and a spillover heatmap

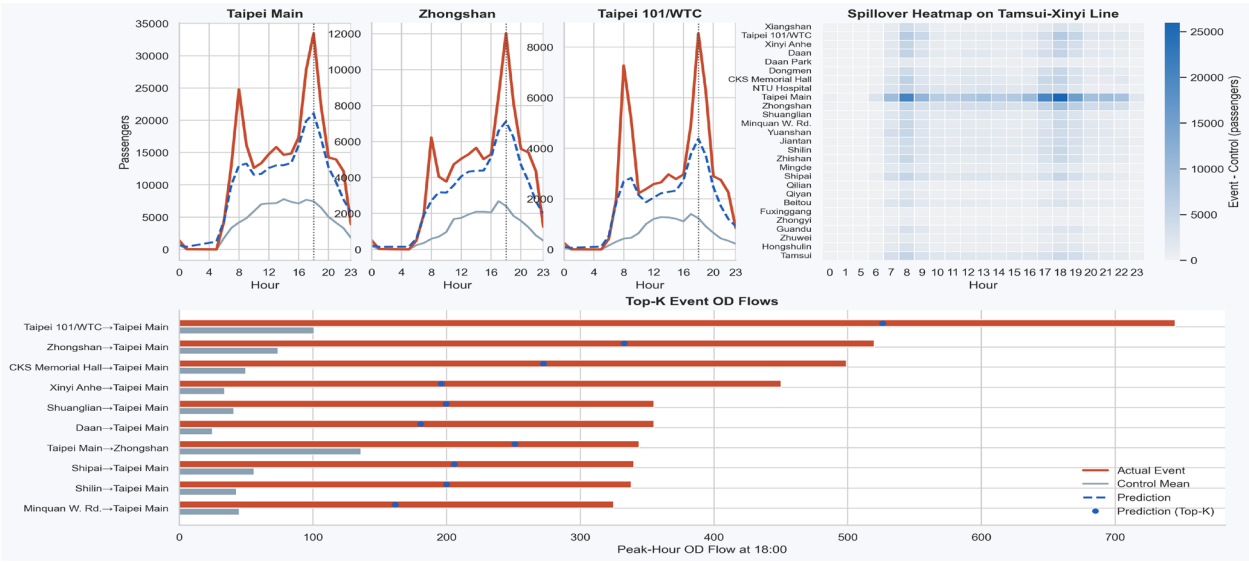

Figure 4. Held-out event-shock case on 2023-08-10, including station-level temporal curves, peak-hour Top-*Kp* OD flows, and a spillover heatmap

Table 5. Per-regime RMSE, Drift Gap, and Recovery Gap [Drift Gap = $(RMSE_{R2}-RMSE_{R1})/RMSE_{R1}$; Recovery Gap = $(RMSE_{R3}-RMSE_{R1})/RMSE_{R1}$.]

Method	R1	R2	R3	Drift Gap (%)	Recovery Gap (%)
DCRNN [6]	38.73	52.36	47.18	35.2	21.8
STGCN [7]	40.27	54.74	49.43	35.9	22.7
Graph WaveNet [8]	37.84	50.54	45.97	33.6	21.5
AGCRN [9]	37.21	49.04	44.98	31.8	20.9
MTGNN [10]	36.58	46.50	43.92	27.1	20.1
STAEformer [32]	36.10	46.00	43.21	27.4	19.7
Autoformer [12]	37.63	51.07	45.82	35.7	21.8
TS-TCC [19]	37.04	47.91	44.47	29.4	20.1
ST-SSL [21]	35.21	44.97	41.86	27.7	18.9
CityTrans [16]	35.89	46.32	42.74	29.1	19.1
DRIFT	**32.14**	**41.45**	**37.21**	29.0	**15.8**

5.3 Robustness Under Regime Shift and Sparsity

The event-shock analysis verifies model behavior on extreme days, but robustness should also hold under sustained regime transition. DRIFT is therefore further evaluated through per-regime RMSE, Drift Gap, and sparsity-conditioned error.

Table 5 shows that R2 is the most challenging regime for all methods. Despite this shift, DRIFT achieves the lowest absolute RMSE in all three regimes, with values of 32.14 in R1, 41.45 in R2, and 37.21 in R3. Relative to ST-SSL, the RMSE is reduced by 8.7% in R1, 7.8% in R2, and 11.1% in R3.

The relative Drift Gap for DRIFT (29.0%) appears higher than MTGNN (27.1%) and STAEformer (27.4%), but this reflects its lower R1 starting point rather than greater regime sensitivity. In absolute terms, DRIFT's RMSE increases by only 9.31 points from R1 to R2, compared with 9.92 for MTGNN and 9.90 for STAEformer. DRIFT also achieves the best Recovery Gap at 15.8%, indicating the fastest return toward pre-disruption accuracy. This pattern is further supported by the ablation in Table 6: removing SSL pre-training raises the Drift Gap from 29.0% to 33.8%, confirming that Stage 1 is the primary driver of cross-regime robustness.

Figure 5 is a regime-anchored visualization derived from the demand statistics in Figure 2 and the regime-wise RMSE values in Table 5. Its role is interpretive rather than inferential: it visualizes transition severity and recovery behavior across regimes, rather than introducing additional numerical estimates. Under this view, baseline errors rise sharply at the R1/R2 boundary, whereas DRIFT shows a milder elevation and a lower R3 error level.

Figure 6 relates forecasting error to the observed OD zero ratio while making sample support explicit. RMSE increases for all models as sparsity becomes stronger, but the DRIFT curve remains the shallowest over the observed range. This pattern is consistent with the conditional results in Figure 8: the improvement is not limited to zero-dominant regions, but is accompanied by the lowest active-pair MAE across all regimes. The widening margin in the high-sparsity region therefore suggests that the gain is structural rather than event-specific.

Figure 5. Regime-anchored chronological robustness view: (a) Regime-aligned monthly demand path and DDI; (b) Regime-anchored RMSE profile over time (Vertical lines mark regime boundaries identified by data-driven change-point detection on aggregated demand statistics.)

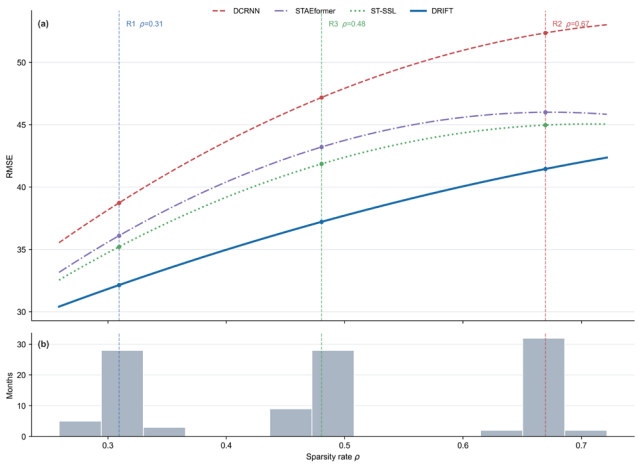

Figure 6. Sparsity-conditioned RMSE (Curves pass through regime-specific RMSE anchors at the observed regime mean sparsities; bars show monthly sample density.)

5.4 Mechanism Analysis

The results above motivate a mechanism-level examination of whether the performance gain can be attributed to the designs in Sections 4.2 and 4.3.

5.4.1 Representation Stability and Embedding Drift

Figure 7 compares latent representation trajectories across R1, R2, and R3. With SSL pre-training, trajectories are smoother near regime boundaries and less fragmented when transitioning between regimes. This visual pattern is consistent with Stage 1's design goal: stabilizing the relational structure of representations rather than forcing identical embeddings across regimes.

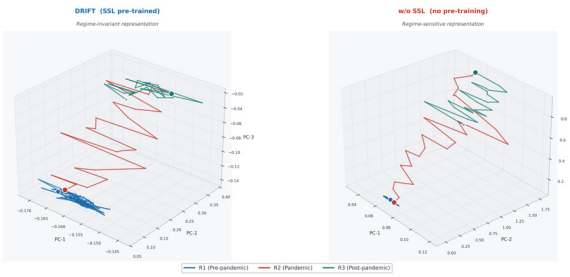

Figure 7. Regime-aligned latent trajectories across R1–R3: DRIFT versus without SSL

5.4.2 AutoGate Behavior

Figure 8 provides complementary evidence for AutoGate. On zero-valued pairs, DRIFT reduces MAE to 1.52, 2.31, and 1.84 in R1–R3 (vs. 4.38, 7.24, and 5.89 for ST-SSL), while active-pair MAE remains lowest in every regime (28.35, 42.17, 35.62). Classification AUC values of 0.981, 0.962, and 0.974 confirm that the gate stays discriminative across all regimes, and the gated-to-ungated loss ratio falls from 1.08 at epoch 1 to 0.68 by convergence.

5.5 Ablation and Sensitivity Analysis

Table 6 shows that the performance gain of DRIFT does not arise from simple module accumulation: each

component contributes a distinct and non-redundant function. Removing SSL pre-training leads to the largest degradation across all metrics, with RMSE increasing from 39.03 to 44.82, Drift Gap worsening from 29.0% to 33.8%, and Peak Recall dropping from 0.993 to 0.971. This confirms that Stage 1 is the primary driver of regime-shift robustness and peak-demand identification. Removing AutoGate produces the second-largest RMSE increase, from 39.03 to 42.57, and reduces Peak Recall to 0.978, indicating that sparse-pair suppression is essential for reliable active-pair estimation.

Removing regime fusion weakens cross-regime calibration and increases the Drift Gap to 31.6%. This corresponds to setting $w = 0$ in Eq. (42), which isolates the standalone drift-aware predictor without inference-time fusion.

Even without fusion, the model achieves RMSE = 41.96 and MAPE = 12.67%, outperforming ST-SSL (RMSE = 42.56, MAPE = 12.90%) and maintaining comparable MAE (25.84 vs. 25.52). This indicates that the performance gain primarily comes from the proposed representation and sparsity-aware design rather than from an ensemble effect.

The supervised branch in the fusion module is identical to the DCRNN baseline in Table 3, so the fusion is best viewed as a refinement layered on an existing baseline component rather than an external ensemble effect. Because the standalone variant already surpasses all baselines, applying analogous fusion strategies on the baseline side is not necessary to identify the source of the gain. Overall, representation learning and AutoGate provide the main performance gain, while fusion contributes further improvement under partial regime overlap.

Figure 8. AutoGate evidence (Top: conditional MAE on zero and active OD pairs. Bottom: precision, recall, F1, and AUC of the zero-tendency classifier.)

Table 6. Ablation study under a controlled component-analysis sweep

Variant	MAE ↓	RMSE ↓	MAPE (%)	Peak Recall ↑	Drift Gap (%) ↓
DRIFT (Full)	23.95	39.03	11.83	0.993	29.0
Without SSL Pretraining	27.43	44.82	13.52	0.971	33.8
Without AutoGate	26.18	42.57	12.91	0.978	32.4
Without Regime Fusion	25.84	41.96	12.67	0.981	31.6
Without L_{crc}	25.37	41.24	12.43	0.983	31.2
Without L_{tcl}	25.61	41.73	12.57	0.982	31.5
Without L_{mfr}	26.04	42.35	12.84	0.979	32.1

5.6 Discussion

DRIFT's performance advantage is primarily structural rather than purely numerical. Its most consistent gains arise from preserving demand organization across regime transitions, including the relative ranking of high-demand corridors, dominant sinks, and spillover patterns, even when absolute demand levels shift substantially. This structural preservation helps explain why DRIFT achieves the highest Peak Recall (0.993) while remaining slightly conservative at extreme amplitudes. Zero-inflation regularization introduces a deliberate trade-off between peak magnitude accuracy and reliable suppression of inactive OD pairs. Peak Recall is largely insensitive to this trade-off because metro operations depend more on identifying which OD pairs will surge than on estimating the exact peak magnitude. Under this criterion, DRIFT consistently delivers the strongest performance.

Two inherent constraints define the scope within which the framework is expected to hold. For extremely low-flow OD pairs that remain near zero for extended periods, the signal-to-noise ratio is fundamentally too low for any history-based model to reliably distinguish persistent inactivity from sporadic demand. DRIFT reduces this uncertainty but cannot eliminate it. For completely novel disruptions with no historical precedent, accurate anticipation requires exogenous information, such as event schedules, weather alerts, or policy announcements, which lie outside the current input space. These constraints should be interpreted as boundary conditions rather than design limitations. Within this scope, DRIFT provides reliable and well-calibrated forecasts; beyond it, no purely data-driven, regime-consistent model can be expected to generalize reliably.

Interpretability is a natural extension of DRIFT because the learned regime structure itself is informative. Unlike conventional black-box forecasters, DRIFT's regime-conditioned representations make it possible to identify which regime a prediction is associated with and how the forecast would change under an alternative regime assignment. Recent advances in Shapley-value attribution and interpretable monitoring [34-36] provide appropriate analytical tools. Integrating these methods with DRIFT's regime structure represents a promising direction toward operationally interpretable metro OD forecasting.

6 Conclusion

This paper presented DRIFT, a regime-consistent and sparsity-aware framework for metro OD demand forecasting under temporal distribution shift. The framework is grounded in the observation that metro network topology remains structurally stable across demand regimes, even when flow magnitudes and sparsity patterns vary substantially. By separating representation learning from magnitude estimation and explicitly modeling zero-inflated sparsity, DRIFT addresses two sources of degradation that prior methods typically handle in isolation: representation drift under regime shift and estimation bias on inactive OD pairs.

Evaluated on 109 months of Taipei Metro OD data spanning three demand regimes, DRIFT achieves superior overall performance among ten strong baselines across all four evaluation metrics, including a 9.3% reduction in RMSE and a 3.3 percentage-point gain in Peak Recall. It also yields the lowest absolute error during the COVID-era disruption and recovery periods. Ablation results further indicate that each component addresses a distinct failure mode: self-supervised pre-training stabilizes representations under regime shift, AutoGate reduces bias on zero-inflated OD pairs, and inference-time fusion balances drift robustness with in-regime prediction accuracy.

The current framework is limited to single-step forecasting, has been validated on a single metro system, and does not explicitly model predictive uncertainty. These remain important directions for future work. More fundamentally, this study shows that regime consistency and zero-inflated sparsity are not independent challenges but structurally coupled sources of degradation in long-horizon OD forecasting. Addressing them jointly is the central insight of this work.

References

[1] X. Ma, J. Zhang, B. Du, C. Ding, L. Sun, Parallel Architecture of Convolutional Bi-Directional LSTM Neural Networks for Network-Wide Metro Ridership Prediction, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 20, No. 6, pp. 2278-2288, June, 2019. https://doi.org/10.1109/TITS.2018.2867042

[2] L. Liu, J. Chen, H. Wu, J. Zhen, G. Li, L. Lin, Physical-Virtual Collaboration Modeling for Intra- and Inter-Station Metro Ridership Prediction, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 23, No. 4, pp. 3377-3391, April, 2022. https://doi.org/10.1109/TITS.2020.3036057

[3] J. Ye, L. Sun, B. Du, Y. Fu, H. Xiong, Coupled Layer-wise Graph Convolution for Transportation Demand Prediction, *Proceedings of the AAAI Conference on Artificial Intelligence*, Virtual, 2021, pp. 4617-4625.

[4] P. Noursalehi, H. N. Koutsopoulos, J. Zhao, Dynamic Origin-Destination Prediction in Urban Rail Systems: A Multi-Resolution Spatio-Temporal Deep Learning Approach, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 23, No. 6, pp. 5106-5115, June, 2022. https://doi.org/10.1109/TITS.2020.3047047

[5] D. A. Tedjopurnomo, Z. Bao, B. Zheng, F. M. Choudhury, A. K. Qin, A Survey on Modern Deep Neural Network for Traffic Prediction: Trends, Methods and Challenges, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 34, No. 4, pp. 1544-1561, April, 2022. https://doi.org/10.1109/TKDE.2020.3001195

[6] Y. Li, R. Yu, C. Shahabi, Y. Liu, Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting, *International Conference on Learning Representations*, Vancouver, Canada, 2018, pp. 1-16. https://openreview.net/forum?id=SJiHXGWAZ

[7] B. Yu, H. Yin, Z. Zhu, Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting, *Proceedings of the International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, 2018, pp. 3634-3640.

[8] Z. Wu, S. Pan, G. Long, J. Jiang, C. Zhang, Graph WaveNet for Deep Spatial-Temporal Graph Modeling, *Proceedings of the International Joint Conference on Artificial Intelligence*, Macao, China, 2019, pp. 1907-1913.

[9] L. Bai, L. Yao, C. Li, X. Wang, C. Wang, Adaptive Graph Convolutional Recurrent Network for Traffic Forecasting, *Advances in Neural Information Processing Systems*, Virtual, 2020, pp. 17804-17815.

[10] Z. Wu, S. Pan, G. Long, J. Jiang, X. Chang, C. Zhang, Connecting the Dots: Multivariate Time Series Forecasting with Graph Neural Networks, *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Virtual, 2020, pp. 753-763. https://doi.org/10.1145/3394486.3403118

[11] J. Jiang, C. Han, W. X. Zhao, J. Wang, PDFormer: Propagation Delay-Aware Dynamic Long-Range Transformer for Traffic Flow Prediction, *Proceedings of the AAAI Conference on Artificial Intelligence*, Washington, DC, USA, 2023, pp. 4365-4373. https://doi.org/10.1609/aaai.v37i4.25556

[12] H. Wu, J. Xu, J. Wang, M. Long, Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting, *Advances in Neural Information Processing Systems*, Virtual, 2021, pp. 22419-22430.

[13] Y. Nie, N. H. Nguyen, P. Sinthong, J. Kalagnanam, A Time Series is Worth 64 Words: Long-term Forecasting with Transformers, *International Conference on Learning Representations*, Kigali, Rwanda, 2023, pp. 1-24. https://openreview.net/forum?id=Jbdc0vTOcol

[14] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, M. Long, iTransformer: Inverted Transformers Are Effective for Time Series Forecasting, *International Conference on Learning Representations*, Vienna, Austria, 2024, pp. 1-25. https://openreview.net/forum?id=JePfAI8fah

[15] A. Tirachini, O. Cats, COVID-19 and Public Transportation: Current Assessment, Prospects, and Research Needs, *Journal of Public Transportation*, Vol. 22, No. 1, pp. 1-21, January, 2020. https://doi.org/10.5038/2375-0901.22.1.1

[16] X. Ouyang, Y. Yang, W. Zhou, Y. Zhang, H. Wang, W. Huang, CityTrans: Domain-Adversarial Training With Knowledge Transfer for Spatio-Temporal Prediction Across Cities, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 36, No. 1, pp. 62-76, January, 2024. https://doi.org/10.1109/TKDE.2023.3283520

[17] T. Zhou, P. Niu, X. Wang, L. Sun, R. Jin, One Fits All: Power General Time Series Analysis by Pretrained LM, *Advances in Neural Information Processing Systems*, New Orleans, USA, 2023, pp. 43322-43355.

[18] M. Goswami, K. Szafer, A. Choudhry, Y. Cai, S. Li, A. Dubrawski, MOMENT: A Family of Open Time-series Foundation Models, *Proceedings of the International Conference on Machine Learning*, Vienna, Austria, 2024, pp. 16115-16152.

[19] E. Eldele, M. Ragab, Z. Chen, M. Wu, C. K. Kwoh, X. Li, C. Guan, Time-Series Representation Learning via Temporal and Contextual Contrasting, *Proceedings of the International Joint Conference on Artificial Intelligence*, Virtual, 2021, pp. 2352-2359.

[20] Z. Yue, Y. Wang, J. Duan, T. Yang, C. Huang, Y. Tong, B. Xu, TS2Vec: Towards Universal Representation of Time Series, *Proceedings of the AAAI Conference on Artificial Intelligence*, Virtual, 2022, pp. 8980-8987.

[21] J. Ji, J. Wang, C. Huang, J. Wu, B. Xu, Z. Xu, J. Zhang, Y. Zheng, Spatio-Temporal Self-Supervised Learning for Traffic Flow Prediction, *Proceedings of the AAAI Conference on Artificial Intelligence*, Washington, DC, USA, 2023, pp. 4356-4364.

[22] D. Lambert, Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing, *Technometrics*, Vol. 34, No. 1, pp. 1-14, February, 1992. https://doi.org/10.2307/1269547

[23] D. B. Hall, Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study, *Biometrics*, Vol. 56, No. 4, pp. 1030-1039, December, 2000. https://www.jstor.org/stable/2677034

[24] T. G. Martin, B. A. Wintle, J. R. Rhodes, P. M. Kuhnert, S. A. Field, S. J. Low-Choy, A. J. Tyre, H. P. Possingham, Zero Tolerance Ecology: Improving Ecological Inference by Modelling the Source of Zero Observations, *Ecology Letters*, Vol. 8, No. 11, pp. 1235-1246, November, 2005. https://doi.org/10.1111/j.1461-0248.2005.00826.x

[25] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A Simple Framework for Contrastive Learning of Visual Representations, *Proceedings of the International Conference on Machine Learning*, Virtual, 2020, pp. 1597-1607.

[26] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum Contrast for Unsupervised Visual Representation Learning, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Virtual, 2020, pp. 9726-9735. https://doi.org/10.1109/CVPR42600.2020.00975

[27] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, G. Zhang, Learning under Concept Drift: A Review, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 31, No. 12, pp. 2346-2363, December, 2019.

https://doi.org/10.1109/TKDE.2018.2876857

[28] W. Qian, D. Zhang, Y. Zhao, K. Zheng, J. J. Q. Yu, Uncertainty Quantification for Traffic Forecasting: A Unified Approach, *Proceedings of the IEEE International Conference on Data Engineering*, Anaheim, CA, USA, 2023, pp. 992-1004.
https://doi.org/10.1109/ICDE55515.2023.00081

[29] Y. Zhang, Q. Wen, X. Wang, W. Chen, L. Sun, Z. Zhang, L. Wang, R. Jin, T. Tan, OneNet: Enhancing Time Series Forecasting Models under Concept Drift by Online Ensembling, *Advances in Neural Information Processing Systems*, New Orleans, LA, USA, 2023, pp. 69949-69980.

[30] X. He, K. Deng, X. Wang, Y. Li, Y. D. Zhang, M. Wang, LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation, *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, Virtual, 2020, pp. 639-648.
https://doi.org/10.1145/3397271.3401063

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, *Advances in Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 5998-6008.

[32] H. Liu, Z. Dong, R. Jiang, J. Deng, J. Deng, Q. Chen, X. Song, Spatio-Temporal Adaptive Embedding Makes Vanilla Transformer SOTA for Traffic Forecasting, *Proceedings of the ACM International Conference on Information and Knowledge Management*, Birmingham, United Kingdom, 2023, pp. 4125-4129.
https://doi.org/10.1145/3583780.3615160

[33] D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, *International Conference on Learning Representations*, San Diego, USA, 2015, pp. 1-15.

[34] Y. S. Lee, S. J. Yen, W. D. Jiang, J. Chen, C. Y. Chang, Illuminating the Black Box: An Interpretable Machine Learning Based on Ensemble Trees, *Expert Systems with Applications*, Vol. 272, Article No. 126720, May, 2025.
https://doi.org/10.1016/j.eswa.2025.126720

[35] W. D. Jiang, C. Y. Chang, S. J. Yen, S. J. Wu, D. Sinha Roy, RealExp: Decoupling Correlation Bias in Shapley Values for Faithful Model Interpretations, *Information Processing & Management*, Vol. 62, No. 4, Article No. 104153, July, 2025.
https://doi.org/10.1016/j.ipm.2025.104153

[36] W. D. Jiang, C. Y. Chang, M. Y. Su, Y. S. Lee, D. Sinha Roy, Toward Interpretable Multimodal Violence Detection With Knowledge Distillation and Modality-Aligned Preprocessing, *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, Vol. 55, No. 9, pp. 6215-6228, September, 2025.
https://doi.org/10.1109/TSMC.2025.3576390

Biographies

Hsiang-Chuan Chang is an Assistant Professor in the Department of Transportation Management at Tamkang University, Taiwan. His research interests include light rail transit systems and artificial intelligence, with emphasis on intelligent applications in transportation management and urban transit systems.

Tzu-Chia Huang is a Ph.D. candidate and lecturer in Computer Science and Information Engineering at Tamkang University, Taiwan. Her research interests include multimodal learning, large language models, and multi-agent systems, focusing on intelligent industrial systems, generative AI deployment, and AI consulting for academic and corporate sectors.

Chih-Yung Chang (Member, IEEE) received the Ph.D. degree in computer science from National Central University, Taiwan, in 1995. He is a Distinguished Professor at Tamkang University. His research includes artificial intelligence, deep learning, IoT, and wireless sensor networks. He is Co-Chair of ACM SIGMOBILE Taiwan.

Wen-Hwa Liao received the Ph.D. degree in computer science and information engineering from National Central University, Taiwan, in 2002. He is a Professor at National Taipei University of Business. His research interests include IoT, wireless sensor networks, and artificial intelligence. He has served as an associate editor for international journals.

Diptendu Sinha Roy (Senior Member, IEEE) received the Ph.D. degree in engineering from Birla Institute of Technology, Mesra, India, in 2010. He is an Associate Professor and Chair of Computer Science and Engineering at NIT Meghalaya. His research interests include software reliability, cloud computing, IoT, and intelligent systems.