Generalising Traffic Forecasting to Regions Without Traffic Observations

Xinyu Su[1,2*], **Majid Sarvi**[1], **Feng Liu**[1], **Egemen Tanin**[1], **Jianzhong Qi**[1†]

[1]School of Computing and Information Systems, The University of Melbourne
[2]Artificial Intelligence Thrust, The Hong Kong University of Science and Technology (Guangzhou)
{suxs3@student., majid.sarvi@, feng.liu1@, etanin@, jianzhong.qi@}unimelb.edu.au

Abstract

Traffic forecasting is essential for intelligent transportation systems. Accurate forecasting relies on continuous observations collected by traffic sensors. However, due to high deployment and maintenance costs, not all regions are equipped with such sensors. This paper aims to forecast for regions without traffic sensors, where the lack of historical traffic observations challenges the generalisability of existing models. We propose a model named **GenCast**, the core idea of which is to exploit external knowledge to compensate for the missing observations and to enhance generalisation. We integrate physics-informed neural networks into GenCast, enabling physical principles to regularise the learning process. We introduce an external signal learning module to explore correlations between traffic states and external signals such as weather conditions, further improving model generalisability. Additionally, we design a spatial grouping module to filter localised features that hinder model generalisability. Extensive experiments show that GenCast consistently reduces forecasting errors on multiple real-world datasets.

Code — https://github.com/suzy0223/GenCast
Extended version — https://arxiv.org/abs/2508.08947

Introduction

Traffic forecasting is essential for intelligent transportation systems, enabling optimisations such as real-time route planning and transportation scheduling. Accurate traffic forecasting yields substantial social and economic benefits by improving travel efficiency, reducing congestion-related losses, and supporting sustainable urban development (Zheng et al. 2020; Qi et al. 2022). However, high deployment costs of traffic sensors often result in their sparse and limited spatial coverage (Wu et al. 2021), creating a gap between limited traffic observations and the need for fine-grained, wide-coverage forecasting.

To bridge this gap, recent works study *kriging and extrapolation*. Kriging models estimate current traffic conditions at locations of interest without sensors, i.e., *unobserved locations* (Zheng et al. 2023; Xu et al. 2025), while extrapolation models take a step further and forecast for such locations (Hu et al. 2023, 2024c). Although these models have produced promising results for scattered unobserved locations (Fig. 1a), they struggle when applied to large continuous regions without traffic sensors (Fig. 1b), i.e., traffic forecasting for *unobserved regions* (Su et al. 2024).

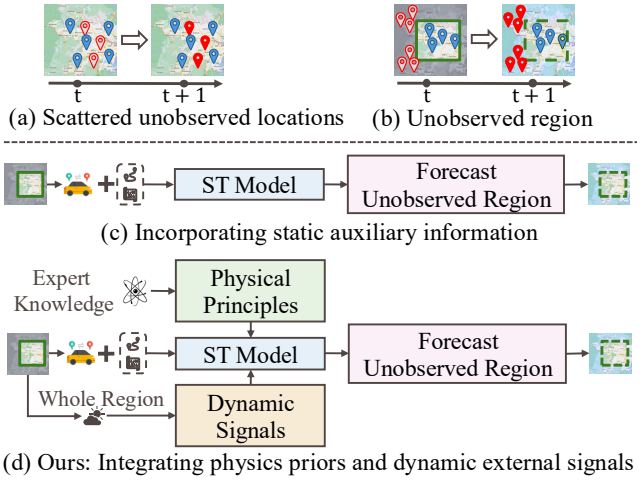


Figure 1: Illustration of the problem setting and modelling strategies. Upper: (a) Scattered unobserved locations vs. (b) Unobserved region (focus). Blue bubbles = observed locations; red hollow bubbles = unobserved ones; and red solid bubbles = targets. Lower: Comparison of different modelling strategies: (c) Static auxiliary features (e.g., POIs, geo-coordinates); (d) Physics priors and dynamic external signals (e.g., weather) for improving generalisation (ours).

We consider traffic forecasting for a region without traffic observations that is adjacent to (or enclosed by) observed regions. As Fig. 1b illustrates, the region with red hollow bubbles does not have traffic observations at current time t, which is adjacent to the region with observations as denoted by the blue bubbles. We aim to forecast for the unobserved region for a future time (the red bubbles at time $t + 1$). This setting is practical due to the staged deployment of sensors

*Work partially done after Xinyu Su joined The Hong Kong University of Science and Technology (Guangzhou).
†Corresponding author.

or unbalanced regional development (Su et al. 2024).

The state-of-the-art (SOTA) model for this setting, STSM (Su et al. 2024), masks locations in observed regions that are similar to the unobserved region and is trained to forecast for such locations, with the aim to generalise to unobserved regions. The similarity is defined based on static auxiliary features, i.e., POI categories and geo-coordinates (Fig. 1c), which, however, do *not* capture the dynamic nature of traffic patterns, thus limiting generalisation capacity.

Another work, KITS (Xu et al. 2025), creates virtual nodes at random locations and trains a model to forecast for such nodes, instead of masking the already-observed locations. It implicitly assumes scattered unobserved locations of a known density (the virtual nodes are created to match this density), thereby introducing a structural prior that limits its generalisability to large continuous unobserved regions.

To address these limitations, we propose to exploit more versatile forms of guidance to enhance the generalisation of traffic forecasting models to regions without traffic observations. We explore two guidance signals: (1) *physical principles* that encode inherent traffic dynamics generalisable across regions; and (2) *dynamic external signals* available across regions that closely correlate with traffic patterns. These guidance signals offer a principled way to bridge the gap between observed and unobserved regions. Accordingly, we propose a traffic fore*cast*ing model *gen*eralisable to unobserved regions named **GenCast**.

To incorporate physical principles, we exploit the Lighthill–Whitham–Richards (LWR) equation (Lighthill and Whitham 1955; Richards 1956) as a soft learning constraint (i.e., a loss term). LWR governs the relationships between traffic density and flow in a road network. Using it poses two technical challenges: (1) The LWR equation considers both traffic density and flow, while density data are typically unavailable. (2) Applying LWR in model learning requires partial derivative over space, while most traffic forecasting models consider a graph over locations of interest which are discrete and not directly differentiable.

To address these challenges, we reformulate the LWR constraint based on traffic speed, yielding a physical constraint without requiring explicit density/flow data. We introduce two continuous spatial embeddings that enable automatic differentiation: (i) SE-L, a large language model (LLM)-based embedding that encodes semantic attributes (e.g., POIs and road structure); and (ii) SE-H, a GeoHash-based embedding that preserves spatial locality. These embeddings serve as differentiable proxies for locations on a graph, enabling residual-based physical loss computation.

To further utilise dynamic external signals, we exploit global weather observations from ECMWF (Muñoz Sabater 2019), for their universal availability and strong correlation with dynamic traffic patterns (Nigam and Srivastava 2023). We introduce an attention-based fusion module to learn the correlations between weather and traffic patterns.

Beyond introducing external guidance signals, it is also crucial to filter patterns local to individual locations (e.g., induced by a traffic accident) that are non-generalisable to unobserved regions. We propose a spatial grouping module that dynamically learns to group locations based on their intrinsic spatial-temporal patterns. The resulting groups enable GenCast to learn shared patterns of a group and suppress disruptive signals local to individual locations (Lin et al. 2023).

Overall, our contributions are summarised as follows:

(1) We propose a traffic forecasting model, GenCast, that aims to generalise to regions without observations.

(2) We design (i) a physical principle-guided loss together with continuously differentiable spatial embeddings and (ii) a cross-domain fusion module to fuse dynamic weather signals with traffic observations. These modules guide GenCast to generalise to regions without traffic observations through external knowledge and signals.

(3) We propose a spatial grouping module that filters out noisy localised signals while preserving region-invariant patterns to further strengthen model generalisability.

(4) We conduct extensive experiments on real-world datasets. The results show that GenCast consistently excels SOTA baselines, reducing forecasting errors by up to 3.1%.

Related Work

Kriging and extrapolation aim to infer current or future observations at unobserved locations (Hu et al. 2023, 2024a,b,c). However, as discussed in the introduction, existing methods often struggle to generalise, particularly to regions without traffic observations. Next, we briefly review representative strategies to improve model generalisation.

Physics-guided approaches have been proposed to enhance generalisability of spatial-temporal models (Hwang et al. 2021; Hettige et al. 2024; Verma, Heinonen, and Garg 2024). In the context of traffic forecasting, these approaches fall into two categories: (1) simulating latent dynamics via neural ODEs or energy-based models (Ji et al. 2022; Wang et al. 2022), which typically require fully observed traffic data to initialise hidden model states; and (2) imposing traffic flow constraints (e.g., LWR) via physics-informed neural networks (PINNs) (Raissi, Perdikaris, and Karniadakis 2019; Shi, Mo, and Di 2021; Zhang et al. 2024), which often assume *continuous* spatial locations. Our model introduces differentiable spatial embeddings to enable PINNs on traffic graphs, which are discrete.

Other studies use external signals to enhance forecasting performance by capturing invariant spatial-temporal patterns from external datasets (Li et al. 2024) or environmental signals (e.g., weather, events) (Mystakidis and Tjortjis 2024; Zhou et al. 2024; Su et al. 2025a; Ruan et al. 2025) These studies use such signals to help detect irregular events or handle short-term missing data. The use of such signals to guide model generalisation to unobserved regions, combined with advanced spatial-temporal graph networks, remains underexplored. A full discussion is provided in Appendix A of our online technique report (Su et al. 2025b).

Preliminaries

Region and Region Graph. Following Su et al. (2024), we represent a region as a graph $G = (V, E)$, where V denotes N locations of interest and E represents their connections based on spatial proximity. Each location is associated

with a feature vector, forming a matrix $\mathbf{L} \in \mathbb{R}^{N \times F}$ that encodes static attributes such as geo-coordinates, road network information, or regional descriptors. The specific features, which can be raw attributes or embeddings, may vary across different methods. For each $v_i \in V$, we use $\mathbf{x}_i$ to denote the series of traffic observations at location v_i and $\mathbf{x}_i^t \in \mathbb{R}^C$ to denote the C different types of observations (e.g., speed and volume) at time t, if there are such observations collected.

Observed and Unobserved Regions. We consider two disjoint but adjacent regions (i.e., graphs): an *observed region* $G_o = (V_o, E_o)$ and an *unobserved region* $G_u = (V_u, E_u)$, where $V_o \cap V_u = \emptyset$. We denote $N_o = |V_o|$ and $N_u = |V_u|$. Graphs G_o and G_u together form the input region graph $G = (V, E)$, where $V = V_o \cup V_u$, $N = N_o + N_u$ and $E_o + E_u \subseteq E$.

Weather Data. We collect weather data from the ECMWF ERA5 dataset (Muñoz Sabater 2019), provided in a gridded format with 9 km × 9 km resolution. For each grid cell i at time t, the weather observation is denoted as $\mathbf{x}_{w,i}^t \in \mathbb{R}^{C_w}$, where $C_w = 4$ is the number of weather attributes (2-meter temperature, surface net solar radiation, surface runoff, and total precipitation). We denote the full weather observations at time t as $\mathbf{X}_w^t \in \mathbb{R}^{N_w \times C_w}$, where N_w is the number of grid cells overlapping the input region G.

Problem Statement. Given region $G = (V, E) = G_o \cup G_u$, location features $\mathbf{L}$ corresponding to V, traffic observations $\mathbf{X}_{G_o}^{t-T+1:t}$ for G_o over the past T time steps, and weather observations $\mathbf{X}_w^{t-T_w+1:t}$ for G over the past T_w time steps (T and T_w may vary and hence are denoted differently), we aim to learn a function f to forecast the traffic conditions $\hat{\mathbf{X}}_{G_u}^{t+1:t+T'}$ for G_u over the next T' time steps:

$$\hat{\mathbf{X}}_{G_u}^{t+1:t+T'} = f(\mathbf{X}_{G_o}^{t-T+1:t}; \mathbf{X}_w^{t-T_w+1:t}; G; \mathbf{L}). \quad (1)$$

We note that observed (or unobserved) regions (locations) refer to regions (locations) with (or without) *traffic* observations. Both types of regions (locations) have weather observations from the ECMWF ERA5 dataset.

Methodology

Model Overview

Fig. 2 overviews GenCast. The backbone (in *gray*) adopts a contrastive architecture with a pseudo-observation generator, following (Su et al. 2024). Each epoch constructs a masked graph G_o^m by randomly masking a subgraph of G_o, producing two input views: original $\mathbf{X}{G_o}^{t-T+1:t}$ and masked $\mathbf{X}{G_o^m}^{t-T+1:t}$. Pseudo-observations are generated for masked locations to compute spatial (geo-coordinate) and temporal (traffic-series) similarity matrices, which feed GCNs and TCNs to capture respective dependencies.

The two views are processed by GCNs and TCNs (i.e., a spatial-temporal model (Wu et al. 2019)) to produce forecasts and graph representations $\mathbf{Z}^{t+T'}G_o$ and $\mathbf{Z}^{t+T'}G_o^m$, with representations taken from the final time step $t + T'$ following Liu et al. (2022). A contrastive loss, L_{cl}, is applied over $\mathbf{Z}_{G_o}^{t+T'}$ and $\mathbf{Z}_{G_o^m}^{t+T'}$ to promote consistent forecasts with and without masked (i.e., unobserved) locations, thereby achieving generalisation to the unobserved region. As both views share identical pipelines, we omit subscripts G_o and G_o^m hereafter. Further backbone details are in Appendix B (Su et al. 2025b).

Our Proposed Modules. We power GenCast with four modules to achieve high generalisability to unobserved regions: **spatial and temporal encoder**, **external signal encoder**, **spatial grouping module**, and **physics-informed module** (coloured in *almond*).

We design a *spatial-temporal encoder* to embed node geo-locations and time indices into differentiable representations ($\mathbf{L}_{enc}$ and $\mathbf{TE}_{enc}$), enabling back-propagation for model optimisation guided by a physical principle-based loss. These embeddings are fused with the input $\mathbf{X}^{t-T+1:t}$ through an STE layer to produce the initial representation $\mathbf{H}^0 \in \mathbb{R}^{T \times N \times D}$, where D is the hidden dimensionality.

To further utilise dynamic external signals, we match the nodes with weather data by their geo-locations. We denote the matched weather data by $\mathbf{X}_{wx}^{t-T_w+1:t}$. An *external signal encoder* (i.e., a cross-attention module) fuses these signals with $\mathbf{H}^0$, producing an enriched representation $\mathbf{H}_{fuse}^0$.

Then, $\mathbf{H}_{fuse}^0$ is fed into the ST model as part of the contrastive learning process. To improve model generalisability, we encourage the learning of essential patterns by filtering out localised signals. We introduce a spatial grouping module that softly assigns each node to a small number of spatial groups via learnable weights. To avoid group representations being impacted by ad hoc local features, we apply an entropy loss L_{spg} to promote confident, near one-hot assignments. This allows each node to primarily contribute to one representative group, supporting clearer group-level representations and better generalisation to unobserved regions.

Besides outputting learned graph representations as mentioned earlier, the ST model also produces forecasts $\hat{\mathbf{X}}^{t+1:t+T'}$, which are fed into the *physics-informed module* with an automatic differentiation step to compute spatial and temporal derivatives. A residual R is computed based on LWR. The physics loss L_{phy} minimises this residual to encourage confinement to physical laws of traffic dynamics.

Model Training. GenCast is optimised with a loss function of four terms, a forecast loss L_{pred} that encourages accurate forecasts at the *masked* locations, plus the contrastive loss L_{cl}, spatial grouping loss L_{spg}, and physics loss L_{phy} as mentioned above. Note that, except for L_{cl}, all other losses are computed based on the masked graph G_o^m to simulate unobserved locations and enhance generalisation.

$$L = L_{pred} + \lambda L_{cl} + \mu L_{spg} + \theta L_{phy} \quad (2)$$

Here, θ and μ are hyper-parameters. The contrastive learning loss weighting follows Su et al. (2024).

Model Inference. For inference, we first compute pseudo-observations for the unobserved locations, and let graph $G = G_o + G_u$ with pseudo-observations (for G_u) be G_m. Then, the model forward process described above is run to generate forecasts for G_m, from which forecasts for the unobserved locations can be extracted.

Next, we detail the four proposed modules of GenCast.

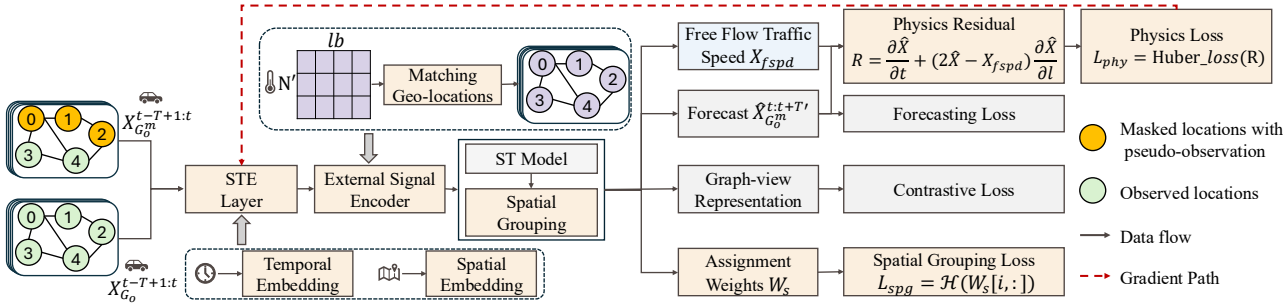


Figure 2: Overview of GenCast. From the observed graph G_o, a masked view G_o^m is formed by randomly masking a subgraph. Temporal and spatial embeddings are fused with both view inputs via a spatial-temporal embedding (STE) layer to produce initial features H_m^0 and H^0, respectively. External signals (weather) are aligned to nodes by geo-coordinates and fused with node features through an external signal encoder, yielding fused representations. These are passed to a spatial-temporal (ST) model with a spatial grouping module, producing forecasts, graph representations, and learnable soft grouping scores for each node. The training objective combines: (1) forecast loss (L_{pred}), (2) contrastive loss for cross-view consistency (L_{cl}), (3) group-aware regularisation (L_{spg}), and (4) a physics loss (L_{phy}) computed from LWR residuals to enforce traffic-dynamic constraints.

Spatial and Temporal Encoder

We encode temporal and spatial features into differentiable embeddings to enable physics residual computation.

Temporal Embedding. We construct a time embedding $\mathbf{TE}$ that captures daily traffic cycles. Each observed time step is assigned a position index within the day ($\mathbf{TE}[i] = i \mod T_d$), where T_d is the number of time intervals per day. The time embedding $\mathbf{TE}$ is encoded using sinusoidal functions to produce a smooth and continuous representation: $\mathbf{TE}_{enc} = \left[\sin\left(2\pi \cdot \frac{\mathbf{TE}}{T_d}\right), \cos\left(2\pi \cdot \frac{\mathbf{TE}}{T_d}\right)\right]$.

Spatial Embedding. We introduce two spatial feature encoding strategies: (1) *LLM-based spatial embedding* (SE-L) and (2) *GeoHash-based spatial embedding* (SE-H). They differ in their input features (and hence processing mechanisms) to suit different location feature availability settings.

LLM-based spatial embedding. SE-L is computed with two steps: (i) location description generation and (ii) embedding generation. Location description generation constructs textual descriptions for each location by prompting an LLM (Sun et al. 2025) with geo-coordinates, geometric properties of the surrounding area, POI information and attributes of the nearest road segments (see Appendix B (Su et al. 2025b)). The location description of S tokens, $\mathbf{L}_t \in \mathbb{R}^{N_o \times S}$, is fed into a frozen LLaMA3 8B Instruct (Meta 2024). The final token embedding from the last hidden layer of the model is SE-L, $\mathbf{L}_{llm} \in \mathbb{R}^{N_o \times d_{llm}}$.

GeoHash-based spatial embeddings. When location features are unavailable beyond the geo-coordinates, we use GeoHash (Niemeyer 2008) to compute spatial embeddings SE-H. There are two main steps: (i) GeoHash string generation and (ii) embedding generation. GeoHash string generation applies GeoHash coding on the geo-coordinates of a location to convert them into a fixed-length alphanumeric string, the length of which decides spatial precision. The embedding generation step then embeds the GeoHash string using a pre-trained character-BERT (Li 2023). The last hidden layer output, discarding the special tokens [CLS] and [SEP], produces $\mathbf{L}'_{hash} \in \mathbb{R}^{N_o \times S \times d_{bert}}$. Here, S is reused to denote the GeoHash string length, and d_{bert} is the embedding dimensionality. To capture semantic dependencies within $\mathbf{L}'_{hash}$, we feed it into a multi-layer Transformer encoder, and we apply mean pooling along the character dimension of the output to obtain the SE-H, $\mathbf{L}_{hash} \in \mathbb{R}^{N_o \times d_{hash}}$.

We use $\mathbf{L}_{enc}$ to denote spatial embeddings ($\mathbf{L}_{llm}$ or $\mathbf{L}_{hash}$) when the context is clear. We add $\mathbf{TE}_{enc}$ and $\mathbf{L}_{enc}$ to obtain $\mathbf{STE}$, and concatenate it with $\mathbf{X}^{t-T+1:t}$ as $\mathbf{H}^0$.

External Signal Encoder

For external weather signals, each traffic location v_i is matched to its nearest weather station s_j based on proximity:

$$\mathbf{X}_{wx,i} = \mathbf{X}_{w,j^*(i)}, \text{where } j^*(i) = \arg\min_{s_j \in \mathcal{S}_w} \text{dist}(v_i, s_j), \quad (3)$$

where $\mathcal{S}_w$ denotes the set of all weather stations, and $\mathbf{X}_{wx,i}$ is the external signal assigned to location v_i. After matching, we obtain an external signal tensor $\mathbf{X}_{wx} \in \mathbb{R}^{T_w \times N_o \times C_w}$, where T_w is the weather window length and C_w is the number of weather features. Empirically, we associate each traffic observation with a 12-hour weather context in the past, i.e., T_w=12, to account for the lasting impact of weather.

To capture traffic–weather interdependencies, we apply cross-attention by projecting $\mathbf{H}^0$ into queries $\mathbf{Q}$, and $\mathbf{X}_{wx}$ into keys $\mathbf{K}$ and values $\mathbf{V}$. Temporal attention scores $\alpha_{t,t'}$ are used to aggregate weather signals: $\mathbf{H}^t_{wx} = \sum_{t'=1}^{T_w} \alpha_{t,t'} \mathbf{V}^{t'}$, yielding $\mathbf{H}_{wx} \in \mathbb{R}^{T \times N_o \times D}$. Recall that D is the hidden dimensionality.

We use gated fusion to fuse weather and traffic signals:

$$\begin{aligned} \mathbf{H}^0_{fuse} &= \text{ReLU}\left(\text{FC}_h\left(z \odot \mathbf{H}^0 + (1-z) \odot \mathbf{H}_{wx}\right)\right), \\ z &= \sigma\left(\text{FC}_s(\mathbf{H}^0) + \text{FC}_t(\mathbf{H}_{wx})\right), \end{aligned} \quad (4)$$

where FC denotes linear layers, $\sigma(\cdot)$ is the sigmoid function, and $\odot$ denotes element-wise multiplication. The fused output $\mathbf{H}^0_{fuse}$ is then passed through the ST model to generate node-level forecasts $\hat{\mathbf{X}}_{G_o^m}^{t+1:t+T'}$ and graph representations $\mathbf{Z}_{G_o}^{t+T'}$ and $\mathbf{Z}_{G_o^m}^{t+T'}$ for loss computation.

Spatial Grouping Module

We adopt spatial grouping to softly cluster locations into latent groups (see Fig. 9 in Appendix B (Su et al. 2025b)), enabling GenCast to capture shared group-level patterns and filter out ad hoc patterns at individual locations, with the help of an entropy regularisation term.

We add a spatial grouping module to each layer of the ST model. For the output feature map $\mathbf{H}^l \in \mathbb{R}^{N_o \times T \times D}$ from the l-th layer, we first perform temporal average pooling to obtain a static spatial representation $\mathbf{H'}^l \in \mathbb{R}^{N_o \times D}$. We then divide the D channels into cg (a hyperparameter) channel groups and reshape the representation into $\mathbf{Z}^l \in \mathbb{R}^{(N_o \cdot cg) \times d'}$, where $d' = D/cg$, producing $N_o \times cg$ samples. This transformation enables the model to capture fine-grained features across channel partitions.

We also project $\mathbf{H'}^l \in \mathbb{R}^{N_o \times D}$ to obtain a learnable $\mathbf{W}_c \in \mathbb{R}^{N_o \times (sg \cdot cg \cdot cg)}$, where sg is the number of spatial groups (a hyperparameter). Then, we reshape it to obtain $\mathbf{W}_c \in \mathbb{R}^{(sg \cdot cg) \times (N_o \cdot cg)}$, and we map $\mathbf{W}_c$ to obtain the representations of group centres as: $\mathbf{C} = \mathrm{Softmax}(\mathbf{W}_c)\mathbf{Z}^l \in \mathbb{R}^{(sg \cdot cg) \times d'}$, i.e., the representation of each group centre is determined by all input samples.

The group assignment score is computed via a distance-based softmax, where $\mathrm{cdist}(\cdot, \cdot)$ computes the pairwise Euclidean distances between all location samples and centres:

$$\mathbf{W}_s = \text{Softmax}\left(-\text{cdist}(\mathbf{Z}^l, \mathbf{C})\right) \in \mathbb{R}^{(N_o \cdot cg) \times (sg \cdot cg)}. \quad (5)$$

GenCast encourages each sample to be confidently assigned to a representative group, suppressing the influence of ad hoc features at individual locations that introduce inconsistent signals and obscure generalisable group-level patterns. To this end, we apply an entropy minimisation loss on the soft assignment weights:

$$\mathcal{L}_{spg} = \frac{1}{N \cdot c_g} \sum_{i=1}^{N \cdot c_g} \mathcal{H}(\mathbf{W}_s[i, :]), \quad (6)$$

where $\mathbf{W}_s[i, :]$ denotes the soft assignment weights of the i-th sample to all $sg \cdot cg$ latent groups, forming a probability distribution over group assignments. $\mathcal{H}(\mathbf{p}) = -\sum_j p_j \log(p_j + \epsilon)$, where p_j denotes the soft assignment probability of a sample to the j-th latent group. Lower entropy will encourage sharper group membership.

Physics-informed Module

The physics-informed module introduces a constraint with the LWR model (Lighthill and Whitham 1955; Richards 1956). LWR describes the evolution of traffic density over space (location l) and time (t) using a conservation law formulated as Eq. 7, where $\rho = \rho(l, t)$ denotes traffic density, and$x = x(l, t)$ represents traffic velocity (i.e., speed).

$$\frac{\partial \rho}{\partial t} + \frac{\partial (\rho x)}{\partial l} = 0. \quad (7)$$

As traffic density observations are *not* commonly available, we rewrite the equation using velocity, assuming a closed system with a functional density-velocity relationship (Greenshields 1935; Xiong, Zhou, and Bennett 2023): $x = x_{fspd}\left(1 - \frac{\rho}{\rho_{max}}\right)$, where x_{fpsd} denotes the free flow speed – the speed at which vehicles travel under low traffic density, and ρ_{max} denotes the maximum traffic density, where vehicles are fully packed. This can also be written as:

$$\rho = \rho_{max}\left(1 - \frac{x}{x_{fspd}}\right) \quad (8)$$

Putting it into Eq. 7 yields:

$$-\frac{\rho_{max}}{x_{fspd}}\frac{\partial x}{\partial t} + \rho_{max}\left(1 - \frac{2x}{x_{fspd}}\right)\frac{\partial x}{\partial l} = 0 \quad (9)$$

Multiplying both sides by $-\frac{x_{fspd}}{\rho_{max}}$ gives: $\frac{\partial x}{\partial t} + (2x - x_{fspd})\frac{\partial x}{\partial l} = 0$. The left-hand side is the physics residual R:

$$R = \frac{\partial x}{\partial t} + (2x - x_{fspd})\frac{\partial x}{\partial l} \quad (10)$$

Further details on the derivation of the physics residual are in Appendix B (Su et al. 2025b).

As the system is not assumed to be closed, the model enforces local flow conservation across connected segments consistent with LWR principles. The Huber loss (Huber 1992) penalises violations of flow conservation to enforce these physical constraints

$$\mathcal{L}_{phy} = \mathrm{Huber}(R, \delta). \quad (11)$$

Here, δ denotes the Huber threshold. To account for dataset-specific error scales, we adopt an adaptive strategy. After a warm-up run (an epoch using the RMSE loss), we compute the τ-quantile of the physical residuals to set δ, where τ is a tunable hyperparameter. Notably, when $\tau = 100\%$, the Huber loss reduces to RMSE. The model is then re-initialised and trained using the Huber loss with this fixed δ.

During model training, Eq. 10 becomes:

$$R = \frac{\partial \hat{\mathbf{X}}}{\partial \mathbf{TE}_{enc}} + \left(2\hat{\mathbf{X}} - \mathbf{X}_{fspd}\right) \cdot \frac{\partial \hat{\mathbf{X}}}{\partial \mathbf{L}_{enc}}, \quad (12)$$

where $\hat{\mathbf{X}} = \hat{\mathbf{X}}_{G_o^m}^{t+1:t+T'}$ denotes model forecasts based on the masked graph, and $\mathbf{X}_{fspd} \in \mathbb{R}^{N \times 1}$ represents estimated free-flow speed at each location (Xiong, Zhou, and Bennett 2023). The partial derivatives with respect to $\mathbf{TE}_{enc}$ and $\mathbf{L}_{enc}$ are computed via automatic differentiation.

Experiments

Experimental Setup

Datasets. We evaluate GenCast on four highway (PEMS-Bay, PEMS07, PEMS08, METR-LA) datasets and one urban (Melbourne) dataset, sampled every 5 or 15 minutes (details in Appendix C (Su et al. 2025b)). We use **ERA5-Land weather data** (Muñoz Sabater 2019) (hourly, 9km×9km) with four traffic-related variables: temperature, solar radiation, precipitation, and runoff. **Region and road network data** for GenCast-L are from OpenStreetMap (2018).

Following prior work (Zheng et al. 2023; Su et al. 2024), we use traffic records from the past two hours to forecast for the next two hours, i.e., $T = T' = 2$ hours. Each dataset is split into training, validation, and test sets in a 4:1:5 spatial ratio, ensuring spatial adjacency within each split. Locations in the training and validation sets are treated as observed, while test locations are unobserved. The space-based split is performed horizontally or vertically based on geo-coordinates. We generate four spatial splits per dataset and report results on average. Temporally, the first 70% of data is used for training, and the remaining 30% for testing.

Competitors. We compare with a transductive Kriging model **GE-GAN** (Xu et al. 2020), inductive Kriging models **IGNNK** (Wu et al. 2021), **INCREASE** (Zheng et al. 2023) and **KITS** (Xu et al. 2025), and the SOTA model for unobserved region forecasting **STSM** (Su et al. 2024).

Implementation Details. All baselines use their default settings. For imputation-based models, we adapt their objective to forecast future values. Our model is trained with Adam (initial $lr = 0.01$), batch size 32, and masking ratio $\sigma_m = 0.5$, tuning hyperparameters on the validation set. All experiments are run on an NVIDIA A100 (80GB) GPU. We report RMSE, MAE, MAPE, and R^2 – the first three measure errors and R^2 reflects improvement over historical averages. More details are in Appendix C.1 (Su et al. 2025b).

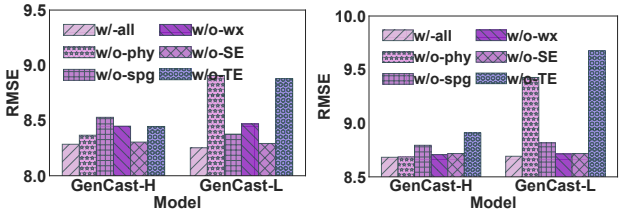
Results

Overall Results. Table 1 reports forecast errors averaged over two hours (24 time steps for highway datasets, and 8 time steps for urban traffic datasets). Our model, with either variant GenCast-H (using GeoHash embeddings) or GenCast-L (using LLM embeddings), consistently outperforms all competitors. Our model reduces forecast errors by up to 3.1% and increases R^2 by up to 125.6% on the Melbourne dataset. We further conducted paired t-tests and Wilcoxon signed-rank tests between our model and the best baseline across all datasets. The results show that our model consistently outperforms the baselines with statistically significant improvements $p \ll 10^{-8}$.

STSM, the SOTA model, is the best baseline in most cases (16 out of 20). GE-GAN and IGNNK underperform due to limited information flow. GE-GAN relies on static similarity, while IGNNK struggles with message propagation path construction. INCREASE iteratively masks and reconstructs nodes using partial observations and static similarities, overlooking dynamic dependencies. KITS creates virtual nodes inside the observed region. This setup limits generalisability to outside, unobserved regions, which is our setting.

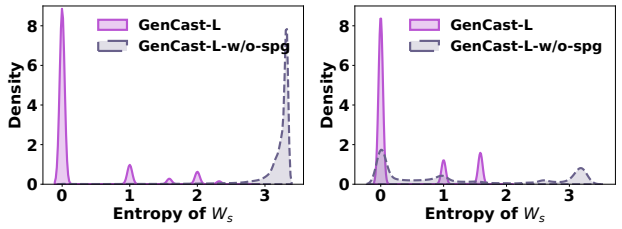
Comparison between Spatial Embedding Strategies. As shown in Table 1, GenCast-L achieves better performance on PEMS07 and PEMS08, while GenCast-H performs better on PEMS-Bay, METR-LA, and Melbourne. We attribute this discrepancy to the quality of SE-L. PEMS-Bay and METR-LA data were collected long ago (Table 3, Appendix C.1 (Su et al. 2025b)), whereas SE-L uses up-to-date OpenStreetMap information. Changes in the physical environment over time may result in a mismatch between the generated embeddings and the conditions at the data collection time. In addition, the Melbourne dataset spans a small, homogeneous CBD area, making it difficult for SE-L to capture distinctive spatial features or meaningful propagation patterns. In contrast, SE-H is generated from geo-coordinates and is updated during training, making it more adaptable and robust to different environments.

Ablation Study. We compare GenCast with five variants: **w/o-phy**, **w/o-spg**, and **w/o-wx** remove the physics constraint, spatial grouping loss, and cross-domain encoder (i.e., weather), respectively; **w/o-SE** and **w/o-TE** remove spatial and temporal embeddings, respectively, together with the physics constraint. As Fig. 3 shows, all variants lead to higher errors, confirming the effectiveness of the modules.



(a) PEMS07 (b) PEMS-Bay

Figure 3: Ablation study results. We include results on the other datasets in the appendix. Same below.



(a) W_s of Layer 1 (b) W_s of Layer 2

Figure 4: SPG entropy: GenCast-L vs. w/o-spg (PEMS07).

For GenCast-L, the physics constraint is particularly important, as the frozen spatial embeddings are less effective without additional guidance. This is evidenced by the high errors of w/o-TE, where only spatial embeddings are used. In contrast, GenCast-H uses simpler, trainable spatial embeddings, which function without physical constraints.

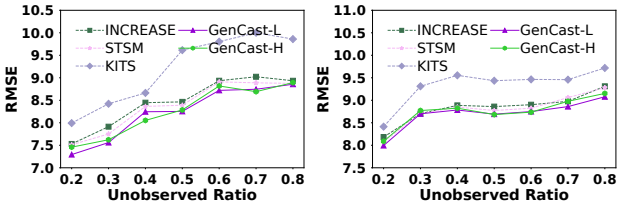
In contrast, w/o-spg has similar impact across datasets, showing its generalise applicability. We further compare the entropy of spatial grouping weights $\mathbf{W}_s$ between GenCast-L and its w/o-spg variant across all splits on PEMS07 (Fig. 4). GenCast-L has sharp peaks near zero entropy, suggesting confident and sparse (i.e., close to one-hot) assignments. The variant w/o-spg has more flat distributions and higher entropy, reflecting more diffuse and ambiguous groupings. These results confirm that our spatial grouping module effectively encourages confident group selection, filtering out localised features that could otherwise harm generalisation.

Parameter Study. We study the impact of key hyperparameters, including the number of spatial/channel groups in the grouping module, loss weights θ, μ, τ, and weather window length T_w. Results in Appendix C.4 (Su et al. 2025b) show that GenCast performs well under consistent settings across datasets, i.e., GenCast does *not* need heavy tuning.

Impact of Unobserved Ratio. We vary the ratio of unobserved nodes in G from 0.2 to 0.8 on all datasets. Each dataset is split horizontally or vertically, and results are averaged over four setups. Fig. 5 compares the top three baselines with our model. GenCast consistently performs the

Dataset	Metric	GE-GAN	IGNNK	INCREASE	STSM	KITS	GenCast-H	GenCast-L	Improve
PEMS07	RMSE↓	20.772	11.398	8.399	8.390	9.574	8.285	**8.253**	1.64%
	MAE↓	15.436	9.016	5.396	5.111	5.150	5.116	**5.073**	0.74%
	MAPE↓	0.270	0.179	0.124	0.123	0.135	0.122	**0.121**	1.63%
	R^2↑	-4.174	-0.618	0.168	0.169	0.094	0.193	**0.197**	16.57%
PEMS08	RMSE↓	23.405	10.646	8.375	7.925	8.182	7.880	**7.863**	0.79%
	MAE↓	17.613	8.138	5.097	4.899	4.863	4.776	**4.728**	2.78%
	MAPE↓	0.298	0.160	0.118	0.114	0.115	0.113	**0.112**	1.75%
	R^2↑	-6.531	-0.642	0.031	0.136	0.083	0.146	**0.150**	10.51%
PEMS-Bay	RMSE↓	25.801	10.051	8.860	8.773	9.435	**8.683**	8.692	1.03%
	MAE↓	24.822	6.596	5.339	5.390	5.270	5.192	**5.139**	2.49%
	MAPE↓	0.407	0.160	0.134	0.134	0.138	**0.131**	**0.131**	2.10%
	R^2↑	-5.856	0.042	0.196	0.210	0.094	**0.228**	0.225	8.43%
METR-LA	RMSE↓	32.303	14.825	13.151	12.952	13.916	**12.720**	12.886	1.79%
	MAE↓	26.371	12.119	9.062	9.010	8.910	**8.792**	8.799	1.33%
	MAPE↓	0.507	0.311	0.272	0.270	0.293	**0.265**	0.267	1.85%
	R^2↑	-4.901	-0.258	0.025	0.048	-0.086	**0.086**	0.063	79.58%
Melbourne	RMSE↓	10.233	14.262	9.579	9.175	10.026	**9.009**	9.258	1.81%
	MAE↓	7.891	12.296	7.627	7.308	7.971	**7.083**	7.253	3.08%
	MAPE↓	0.374	0.939	0.408	0.388	0.415	**0.366**	0.370	2.27%
	R^2↑	-0.213	-1.810	-0.042	0.027	-0.165	**0.061**	0.012	125.56%

Table 1: Overall model performance. "↓"/"↑" indicate lower/larger is better. The best baseline results are underlined, and the best results are in bold. "Improve" means the errors reduced by GenCast-L compared with the best baseline model.



(a) PEMS07 (b) PEMS-Bay

Figure 5: Model performance vs. unobserved ratio.

Model	PEMS-Bay (Ring Split)			
	RMSE↓	MAE↓	MAPE↓	R^2↑
GE-GAN	26.073	25.147	0.411	-6.395
IGNNK	12.881	10.056	0.198	-0.808
INCREASE	8.662	5.126	0.126	0.178
STSM	8.599	5.052	0.129	0.189
KITS	9.087	4.942	0.134	0.098
GenCast-H	**8.323**	4.929	**0.123**	**0.241**
GenCast-L	8.583	**4.734**	0.125	0.191
Improve	3.21%	4.21%	2.38%	27.51%
Model	NREL			
	RMSE↓	MAE↓	MAPE↓	R^2↑
GE-GAN	12.142	9.169	7.444	-0.358
IGNNK	13.732	10.444	2.409	-0.697
INCREASE	8.534	6.045	2.730	0.177
STSM	7.733	5.050	1.789	0.326
KITS	8.704	5.513	1.776	0.314
GenCast-w/wx	**7.620**	**4.770**	**1.750**	**0.345**
Improve	1.46%	5.54%	1.46%	5.83%

Table 2: Results on PEMS-Bay (Ring Split) and NREL.

best, confirming its robustness against the unobserved ratio.

Impact of Space Splits. The relative position of the observed and unobserved regions may also impact model performance. We test model robustness with a ring split on PEMS-Bay, reflecting city layouts. Experiments show that GenCast again consistently outperform all baseline models, achieving up to a 27.5% improvement in R^2 (Table 2).

Domain Generalisability of GenCast. We further adapt GenCast to the solar power NREL dataset (NREL 2018), using GenCast-w/wx (i.e., without spatial embeddings or physics constraints) due to domain complexity. As shown in Table 2, it still outperforms all baselines, demonstrating strong generalisability.

Additional Results. See Appendix C (Su et al. 2025b).

Conclusion

We proposed a model named GenCast to address the challenges in traffic forecasting for unobserved regions. Unlike purely data-driven approaches, GenCast uses physics knowledge and external spatial-temporal data (e.g., weather) to enhance generalisation to unobserved regions. It employs two continuously differentiable spatial embeddings to support physics-informed learning. A spatial grouping module filters localised, non-transferable features. We evaluated GenCast on real-world traffic datasets. The results show that it consistently outperforms SOTA models across different settings, achieving up to 3.1% reduction in forecast error and up to 125.6% improvement in R^2. Although GenCast achieves strong generalisation across regions, the relatively low R^2 indicates room for improvement. Future work will focus on balancing cross-region generalisability and local fidelity through adaptive or hybrid modelling strategies.

Ethical Statement

All datasets used in this study are publicly available and do not contain any personally identifiable information. There are no ethical concerns associated with this work.

Acknowledgments

This work is in part supported by the Australian Research Council (ARC) via Discovery Projects DP230101534 and DP240101006. Jianzhong Qi is supported by ARC Future Fellowship FT240100170. Feng Liu is supported by the ARC with grant numbers DE240101089, LP240100101, DP230101540 and the NSF&CSIRO Responsible AI program with grant number 2303037.

References

Greenshields, B. D. 1935. A Study of Traffic Capacity. *Proceedings of the Highway Research Board*, 448–477.

Hettige, K. H.; Ji, J.; Xiang, S.; Long, C.; Cong, G.; and Wang, J. 2024. AirPhyNet: Harnessing Physics-Guided Neural Networks for Air Quality Prediction. In *ICLR*.

Hu, J.; Liang, Y.; Fan, Z.; Chen, H.; Zheng, Y.; and Zimmermann, R. 2023. Graph Neural Processes for Spatio-temporal Extrapolation. In *KDD*, 752–763.

Hu, J.; Liang, Y.; Fan, Z.; Liu, L.; Yin, Y.; and Zimmermann, R. 2024a. Decoupling Long-and Short-term Patterns in Spatiotemporal Inference. *IEEE Transactions on Neural Networks and Learning Systems*, 16328–16340.

Hu, J.; Liu, X.; Fan, Z.; Liang, Y.; and Zimmermann, R. 2024b. Towards Unifying Diffusion Models for Probabilistic Spatio-temporal Graph Learning. In *SIGSPATIAL*, 135–146.

Hu, J.; Liu, X.; Fan, Z.; Yin, Y.; Xiang, S.; Ramasamy, S.; and Zimmermann, R. 2024c. Prompt-Based Spatio-Temporal Graph Transfer Learning. In *CIKM*, 890–899.

Huber, P. J. 1992. Robust Estimation of a Location Parameter. In *Breakthroughs in Statistics: Methodology and distribution*, 492–518.

Hwang, J.; Choi, J.; Choi, H.; Lee, K.; Lee, D.; and Park, N. 2021. Climate Modeling with Neural Diffusion Equations. In *ICDM*, 230–239.

Ji, J.; Wang, J.; Jiang, Z.; Jiang, J.; and Zhang, H. 2022. STDEN: Towards Physics-guided Neural Networks for Traffic Flow Prediction. In *AAAI*, 4048–4056.

Li, H. 2023. Char-BERT: Character-level BERT Model. https://huggingface.co/lhy/char-bert-base-uncased. Accessed: 2025-03-01.

Li, Z.; Xia, L.; Shi, L.; Xu, Y.; Yin, D.; and Huang, C. 2024. OpenCity: Open Spatio-temporal Foundation Models for Traffic Prediction. arXiv:2408.10269.

Lighthill, M. J.; and Whitham, G. B. 1955. On Kinematic Waves. II. A Theory of Traffic Flow on Long Crowded Roads. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 317–345.

Lin, K.-Y.; Du, J.-R.; Gao, Y.; Zhou, J.; and Zheng, W.-S. 2023. Diversifying Spatial-temporal Perception for Video Domain Generalization. *NeurIPS*, 56012–56026.

Liu, X.; Liang, Y.; Huang, C.; Zheng, Y.; Hooi, B.; and Zimmermann, R. 2022. When Do Contrastive Learning Signals Help Spatio-Temporal Graph Forecasting? In *SIGSPATIAL*, 5:1–5:12.

Meta. 2024. LLaMA Models. https://www.llama.com/. Accessed: 2025-03-01.

Muñoz Sabater, J. 2019. ERA5-Land Hourly Data from 1950 to Present. https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land. Accessed: 2025-07-30.

Mystakidis, A.; and Tjortjis, C. 2024. Traffic Congestion Prediction and Missing Data: A Classification Approach Using Weather Information. *International Journal of Data Science and Analytics*, 1–20.

Niemeyer, G. 2008. GeoHash. https://github.com/davetroy/geohash. Accessed: 2025-03-01.

Nigam, A.; and Srivastava, S. 2023. Hybrid Deep Learning Models for Traffic Stream Variables Prediction during Rainfall. *Multimodal Transportation*, 100052.

NREL. 2018. Solar Power Data for Integration Studies. https://www.nrel.gov/grid/solar-power-data.html. Accessed: 2025-03-01.

OpenStreetMap. 2018. OpenStreetMap US Northeast Data Dump. https://download.geofabrik.de/. Accessed: 2023-03-01.

Qi, J.; Zhao, Z.; Tanin, E.; Cui, T.; Nassir, N.; and Sarvi, M. 2022. A Graph and Attentive Multi-Path Convolutional Network for Traffic Prediction. *IEEE Transactions on Knowledge and Data Engineering*, 6548–6560.

Raissi, M.; Perdikaris, P.; and Karniadakis, G. E. 2019. Physics-informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations. *Journal of Computational physics*, 686–707.

Richards, P. I. 1956. Shock Waves on the Highway. *Operations Research*, 42–51.

Ruan, W.; Dang, X.; Zhou, Z.; Lyu, S.; and Liang, Y. 2025. A Retrieval Augmented Spatio-Temporal Framework for Traffic Prediction. arXiv:2508.16623.

Shi, R.; Mo, Z.; and Di, X. 2021. Physics-informed Deep Learning for Traffic State Estimation: A Hybrid Paradigm Informed by Second-order Traffic Models. In *AAAI*, 540–547.

Su, X.; Liu, F.; Chang, Y.; Tanin, E.; Sarvi, M.; and Qi, J. 2025a. Dualcast: A model to Disentangle Aperiodic Events from Traffic Series. In *IJCAI*.

Su, X.; Qi, J.; Tanin, E.; Chang, Y.; and Sarvi, M. 2024. Spatial-temporal Forecasting for Regions without Observations. In *EDBT*, 488–500.

Su, X.; Sarvi, M.; Liu, F.; Tanin, E.; and Qi, J. 2025b. Generalising Traffic Forecasting to Regions without Traffic Observations. arXiv:2508.08947.

Sun, F.; Chang, Y.; Tanin, E.; Karunasekera, S.; and Qi, J. 2025. FlexiReg: Flexible Urban Region Representation Learning. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2702–2713.

Verma, Y.; Heinonen, M.; and Garg, V. 2024. ClimODE: Climate and Weather Forecasting with Physics-informed Neural ODEs. In *ICLR*.

Wang, J.; Ji, J.; Jiang, Z.; and Sun, L. 2022. Traffic Flow Prediction based on Spatiotemporal Potential Energy Fields. *IEEE Transactions on Knowledge and Data Engineering*, 9073–9087.

Wu, Y.; Zhuang, D.; Labbe, A.; and Sun, L. 2021. Inductive Graph Neural Networks for Spatiotemporal Kriging. In *AAAI*, 4478–4485.

Wu, Z.; Pan, S.; Long, G.; Jiang, J.; and Zhang, C. 2019. Graph WaveNet for Deep Spatial-Temporal Graph Modeling. In *IJCAI*.

Xiong, H.; Zhou, X.; and Bennett, D. A. 2023. Detecting Spatiotemporal Propagation Patterns of Traffic Congestion from Fine-grained Vehicle Trajectory Data. *International Journal of Geographical Information Science*, 1157–1179.

Xu, D.; Wei, C.; Peng, P.; Xuan, Q.; and Guo, H. 2020. GE-GAN: A Novel Deep Learning Framework for Road Traffic State Estimation. *Transportation Research Part C: Emerging Technologies*, 102635.

Xu, Q.; Long, C.; Li, Z.; Ruan, S.; Zhao, R.; and Li, Z. 2025. KITS: Inductive Spatio-temporal Kriging with Increment Training Strategy. In *AAAI*, 12945–12953.

Zhang, J.; Mao, S.; Yang, L.; Ma, W.; Li, S.; and Gao, Z. 2024. Physics-informed Deep Learning for Traffic State Estimation based on the Traffic Flow Model and Computational Graph Method. *Information Fusion*, 101971.

Zheng, C.; Fan, X.; Wang, C.; and Qi, J. 2020. GMAN: A Graph Multi-Attention Network for Traffic Prediction. In *AAAI*, 1234–1241.

Zheng, C.; Fan, X.; Wang, C.; Qi, J.; Chen, C.; and Chen, L. 2023. INCREASE: Inductive Graph Representation Learning for Spatio-Temporal Kriging. In *WWW*, 673–683.

Zhou, Z.; Lyu, G.; Huang, Y.; Wang, Z.; Jia, Z.; and Yang, Z. 2024. SDformer: Transformer with Spectral Filter and Dynamic Attention for Multivariate Time Series Long-term Forecasting. In *IJCAI*.