arXiv:2508.08947v1 [cs.LG] 12 Aug 2025

Generalising Traffic Forecasting to Regions without Traffic Observations

Xinyu Su, Majid Sarvi, Feng Liu, Egemen Tanin, Jianzhong Qi*

School of Computing and Information Systems, The University of Melbourne
{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.

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. 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. 2024b).

*Corresponding author.

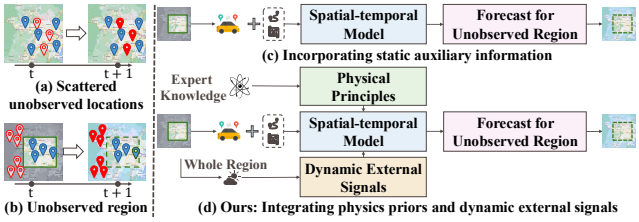


Figure 1: Illustration of the problem setting and modelling strategies. **Left**: (a) Scattered unobserved locations vs. (b) Unobserved region (**our focus**). Blue bubbles represent observed locations; red hollow bubbles denote unobserved ones; and red solid bubbles denote the target forecasts. **Right**: Comparison of different modelling strategies: (c) Incorporating static auxiliary features (e.g., learning from locations with similar POIs or geo-coordinates); (d) Integrates physics priors and dynamic external signals (e.g., weather) to improve 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 staged deployment of sensors or unbalanced regional development (Su et al. 2024b).

The state-of-the-art (SOTA) model for this setting, STSM (Su et al. 2024b), 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 outperforms SOTA baselines, reducing forecasting errors by up to 3.1% and improving R^2 scores by up to 125.6%.

Related Work

Kriging and extrapolation in traffic forecasting 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 when forecasting for 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) (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-domain datasets (Li et al. 2024) or environmental features (e.g., weather (Nigam and Srivastava 2023; Mystakidis and Tjortjis 2024) or events (Su et al. 2024a)). 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 of these works is included in Appendix A in the supplementary materials.

Preliminaries

Region and Region Graph. Following Su et al. (2024b), 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 $G = (V, E) = G_o + G_u$, 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 presents an overview of GenCast. The backbone of the model (components shown in *gray*) takes a contrastive learning architecture. At each training epoch, a masked graph view G_o^m is constructed by randomly masking a subgraph from the observed graph G_o, yielding two input sequences (i.e., two *views*): the original $\mathbf{X}_{G_o}^{t-T+1:t}$ and the masked $\mathbf{X}_{G_o^m}^{t-T+1:t}$. Following Su et al. (2024b), we generate pseudo-observations for the masked locations, to enable computing both spatial (i.e., geo-coordinate) proximity-based and temporal (i.e., traffic series) similarity-based adjacency matrices. These allow constructing graph convolution layers (GCNs) to capture spatial dependencies and temporal convolutional networks (TCNs) to capture temporal patterns.

The two views go through the GCNs and TCNs (i.e., a spatial-temporal (ST) model) separately to produce forecasts and graph representations $\mathbf{Z}_{G_o}^{t+T'}$ and $\mathbf{Z}_{G_o^m}^{t+T'}$, where the representation is 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. Since both views go through identical pipelines, we omit subscripts 'G_o' and 'G_o^m' in the subsequent discussion as long as the context is clear. More details about the backbone contrastive learning process are in Appendix B.

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. (2024b).

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.

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: $\mathbf{TE}_{enc} = \left[\sin\left(2\pi \cdot \frac{\mathbf{TE}}{T_d}\right), \cos\left(2\pi \cdot \frac{\mathbf{TE}}{T_d}\right)\right]$.

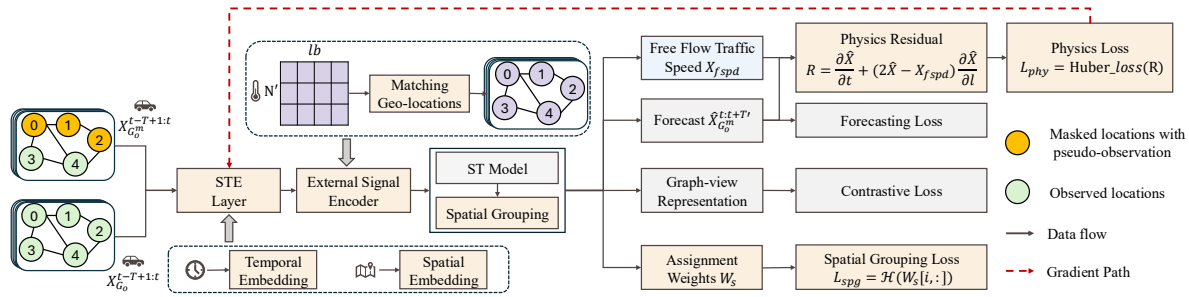


Figure 2: Overview of GenCast. Given observed traffic data G_o, a masked view G_o^m is generated by randomly masking a subgraph. Temporal and spatial embeddings $\mathbf{TE}_{enc}$ and $\mathbf{L}_{enc}$ are fused with inputs $\mathbf{X}_{G_o^m}^{t-T+1:t}$ and $\mathbf{X}_{G_o}^{t-T+1:t}$ via a spatial-temporal embedding (STE) layer to produce initial features $\mathbf{H}_m^0$ and $\mathbf{H}^0$, respectively. External signals (weather) are matched to graph nodes via geo-coordinates and are integrated with the node features using an external signal encoder, yielding $\mathbf{H}_{m,fuse}^0$ and $\mathbf{H}_{fuse}^0$.These feature matrices 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. Three loss terms are used: forecast loss L_{pred} measures forecast errors; contrastive loss L_{cl} measures node representation consistency across G_o^m and G_o; and group-aware loss L_{spg} measures group assignment confidence. The physics module computes a residual R based on LWR to form a fourth (i.e., physics) loss L_{phy}, to regularise GenCast by physical principles of traffic dynamics.

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 with geo-coordinates, geometric properties of the surrounding area, POI information and attributes of the nearest road segments (see Appendix B). 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, $\mathbf{L}_{llm} \in \mathbb{R}^{N_o \times d_{llm}}$ (d_{llm} is the embedding dimensionality).

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

To use 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: $\mathbf{H}_{wx}^t = \sum_{t'=1}^{T_w} \alpha_{t,t'} \mathbf{V}^{t'}$, 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:

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

where FC denotes linear layers, $\sigma(\cdot)$ is the sigmoid function, and $\odot$ denotes element-wise multiplication. The fused output $\mathbf{H}_{fuse}^0$ 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 (Fig. 3), enabling GenCast to capture shared patterns of a group and filtered 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 $\mathbf{H}'^l \in \mathbb{R}^{N_o \times D}$.

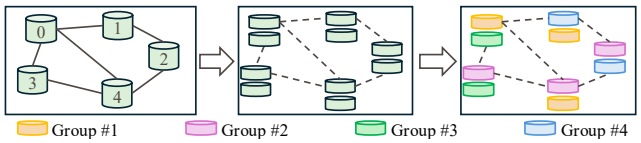


Figure 3: Spatial grouping module. The feature dimensionality D of each node is split into cg channel groups, producing $N_o \times cg$ samples. These samples are softly assigned to $cg \times sg$ spatial groups through assignment weights, with an entropy loss encouraging confident (i.e., one-hot-like) group assignments.

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 reshap 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:

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

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.

To guide GenCast to follow the physical principles, we use the Huber loss (Huber 1992) to minimise the physics residual (i.e., to penalise violations of the physical law):

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

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 datasets (PEMS-Bay, PEMS07, PEMS08, and METR-LA) and one urban dataset (Melbourne), with 5-min or 15-min intervals. Dataset details, including statistics, processing, and visualisations, are in Appendix C.1. 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 LLM-based embeddings are from OpenStreetMap (2018).

Following prior work (Zheng et al. 2023; Su et al. 2024b), 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. 2024b).

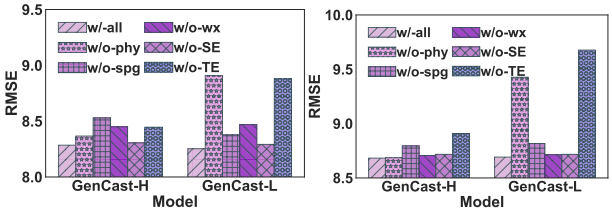
Implementation Details. We use the default settings of all baselines. For imputation-based models, we adapt their objective to forecast future values. Our model is trained with Adam (initial learning rate: 0.01), batch size 32, and masking ratio $\sigma_m = 0.5$, with hyperparameters selected on the 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.

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 $p \ll 10^{-8}$.

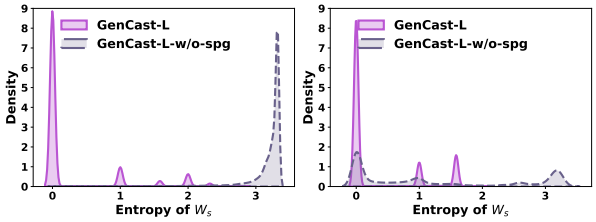
STSM, the SOTA model, is the best baseline in most cases (16 out of 20). GE-GAN and IGNNK underperform due to limited information flow. GE-GAN relies on static similarity, while IGNNK struggles with message propagation path construction. INCREASE iteratively masks and reconstructs nodes using partial 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. The data for PEMS-Bay and METR-LA were collected a long time ago (see Table 3 in Appendix C.1), whereas the regional information used for SE-L is retrieved from the latest OpenStreetMap data. 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 covers a small, homogeneous area concentrated in Melbourne CBD, 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.



(a) PEMS07 (b) PEMS-Bay

Figure 4: Ablation study results. We include results on the other datasets in the appendix. Same below.



(a) $\mathbf{W}_s$ of Layer 1 (b) $\mathbf{W}_s$ of Layer 2

Figure 5: SPG entropy: GenCast-L vs. `w/o-spg` (PEMS07).

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. 4 shows, all variants lead to higher errors, confirming the effectiveness of the modules.

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.

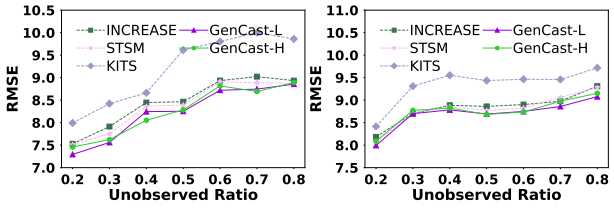
In contrast, w/o-spg has similar impact across datasets, showing its generalise applicability. We further compare the entropy of spatial grouping weights $\mathbf{W}_s$ between GenCast-L and its w/o-spg variant across all splits on PEMS07 (Fig. 5). GenCast-L has sharp peaks near zero entropy, suggesting 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. The results (see Appendix C) 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 unobserved ratio (i.e., percentage of unobserved nodes in G) from 0.2 to 0.8 on all datasets. As before, we split each dataset ei-

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. "↓" (and "↑") indicates that lower (and larger) values are 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 6: Model performance vs. unobserved ratio.

ther horizontally or vertically and report the performance averaged over four setups. We plot the top-3 best baselines with our models in Fig. 6. GenCast consistently performs the 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. Appendix C presents further results on the other datasets, model running time, and generalisability of the external signal encoder.

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: Performance on PEMS-Bay (Ring Split) and NREL.

Conclusion

We proposed a model named GenCast to address the challenges in traffic forecasting for unobserved regions. Unlike purely data-driven models, GenCast uses physics knowledge and external spatial-temporal data (i.e., weather) to achieve generalisability over regions without traffic observations. To support physics-informed learning, we designed two continuously differentiable spatial embeddings. We further introduced a spatial grouping module to filter out localised features that are not transferable to unobserved regions. 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.

References

Berndt, D. J.; and Clifford, J. 1994. Using Dynamic Time Warping to Find Patterns in Time Series. In *KDD workshop*, 359–370.

CalTrans. 2001. California Department of Transportation Performance Measurement System (PeMS). https://pems.dot.ca.gov.

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.

Li, Y.; Yu, R.; Shahabi, C.; and Liu, Y. 2018. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. In *ICLR*.

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

Muñoz Sabater, J. 2019. ERA5-Land Hourly Data from 1950 to Present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS).

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.

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.

OpenStreetMap. 2018. OpenStreetMap US Northeast Data Dump. https://download.geofabrik.de/.

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.

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. 2024a. DualCast: Disentangling Aperiodic Events from Traffic Series with a Dual-Branch Model. *arXiv preprint arXiv:2411.18286*.

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

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*, 1907–1913.

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.

A Related Work

Extrapolation, kriging, and forecasting are core spatial-temporal tasks that often share backbone spatial-temporal learning models (Hu et al. 2024b,c), with recent approaches combining graph and sequence models to capture spatial-temporal dependencies (Wu et al. 2019; Zheng et al. 2020; Hu et al. 2023, 2024a). Kriging and extrapolation are particularly challenging due to the lack of ground truth at unobserved target locations. To simulate such settings during training, existing models typically mask observed locations (Wu et al. 2021; Zheng et al. 2023) or interpolate between them (Xu et al. 2025), and are trained to reconstruct the masked values. While being effective for scattered missing observations, these approaches often fail in more realistic scenarios involving large, continuous unobserved regions. To address this issue, STSM (Su et al. 2024b) introduces a selective masking strategy based on location similarity, encouraging the model to learn from locations that resemble those in the unobserved region. However, because the similarity is derived from static auxiliary features (e.g., POI categories and geographic coordinates), it fails to capture dynamic traffic patterns, thereby limiting the STSM's generalisability in complex real-world environments.

To improve generalisability, recent efforts have explored (1) *contrastive learning*, (2)*physics-guided models*, and (3) *external data signals*.

Contrastive Learning (CL)-Based Models. CL-based models learn transferable representations by aligning similar inputs while distinguishing dissimilar ones through positive and negative sample pairs. For spatial-temporal forecasting, models often construct different graph views by perturbing nodes or edges, helping to uncover invariant patterns across space and time (Liu et al. 2022). To enable generalisation to unobserved regions, STSM (Su et al. 2024b) contrasts the forecasts learned for the full input graph with those learned for selectively masked graphs, encouraging consistency under simulated unobserved scenarios. We also follow a contrastive learning backbone. Instead of selective masking, we adopt a more flexible random subgraph masking strategy, which does not require knowledge about the similarity between the locations in the observed and the unobserved regions as mentioned above.

Physics-Guided Models. Physics-guided models have shown strong potential for enhancing generalisation in scientific domains such as fluid dynamics and climate modelling, by embedding known physical laws into deep learning. Recent traffic forecasting studies have explored physics guidance to introduce inductive biases and improve model accuracy given sparse or irregular observations.

Physics-guided forecasting models generally follow two directions: (1) using neural networks to solve or approximate Partial Differential Equations (PDEs) or (2) incorporating physical constraints into the training loss.

Models in the first category leverage neural Ordinary Differential Equations (ODEs) or diffusion-inspired dynamics to simulate latent physical processes (Hwang et al. 2021; Ji et al. 2022; Wang et al. 2022; Verma, Heinonen, and Garg 2024). Such models embed physics priors implicitly through model design, capturing continuous-time evolution aligned with domain dynamics. They require fully observed spatial domains to initialise physical states, limiting their applicability to settings with unobserved regions.

Models in the second category explicitly incorporate physical constraints into the training objective. Physics-Informed Neural Networks (PINNs) (Raissi, Perdikaris, and Karniadakis 2019) extend this idea by introducing residuals of governing equations directly into the loss function, enabling models to learn solutions consistent with physical laws. PINNs have been successfully applied to domains such as air quality estimation (Hettige et al. 2024) and climate dynamics (Hwang et al. 2021), where robustness and generalisation with sparse or noisy observations are critical. For traffic forecasting, classical physical models such as the Lighthill-Whitham-Richards (Lighthill and Whitham 1955; Richards 1956) and Greenshields (Greenshields 1935) equations define macroscopic relationships among speed, density, and flow, and have been adopted as soft constraints (Shi, Mo, and Di 2021; Zhang et al. 2024).

Despite these advancements, existing physics-guided models are often grid-based or assume continuous spatial coordinates, posing challenges for graph-based traffic modelling. Adapting PINN-style techniques to discrete graphs remains underexplored for improving spatial-temporal generalisation to unobserved regions.

Exploiting External Data Signals. Beyond traffic observations, a variety of external (auxiliary) signals have been explored to enhance robustness and generalisation in spatial-temporal forecasting tasks. These signals can be broadly grouped into: (1) multi-sourced spatial-temporal data e.g., traffic flow data, taxi demand data, and traffic index statics (Li et al. 2024); (2) external contextual features, e.g., weather conditions, major events, and holidays (Su et al. 2024a). Most existing work focuses on using these signals to enhance sequence-based models to detect irregular events or handle short-term missing data. The use of such signals for guiding model generalisation to unobserved regions, especially when combined with advanced spatial-temporal graph networks, remains underexplored.

B Additional Model Details

B.1 Details of GenCast Backbone

Following Su et al. (2024b), our model backbone takes a contrastive learning architecture (Fig. 7). At each training epoch, a masked graph view G_o^m is constructed by randomly masking a subgraph from the observed graph G_o, yielding two input sequences (i.e., two *views*): the original $\mathbf{X}_{G_o}^{t-T+1:t}$ and the masked $\mathbf{X}_{G_o^m}^{t-T+1:t}$. Then, we generate pseudo-observations for the masked locations, allowing the construction of both spatial (i.e., geo-coordinate) proximity-based and temporal (i.e., traffic series) similarity-based adjacency matrices.

The spatial adjacency matrix $\mathbf{A}_{sg}$ is defined by Eq. 13, where ϵ_{sg} is a hyperparameter (value set following Su et al. (2024b)), and $dist(c_i, c_j)$ denotes the Euclidean distance be-

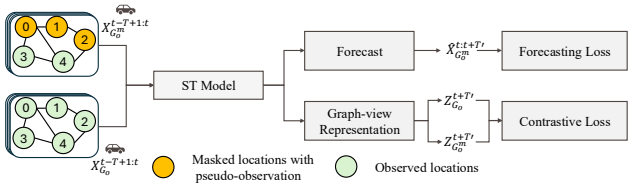


Figure 7: Framework of GenCast backbone.

tween locations i and j (c_i and c_j are their geo-coordinates.)

$$A_{sg,ij} = \begin{cases} 1 & \exp(-\frac{\text{dist}(c_i,c_j)^2}{\sigma^2}) \geq \epsilon_{sg}, \\ 0 & \text{otherwise.} \end{cases} \quad (13)$$

GenCast uses Dynamic Time Warping (DTW) (Berndt and Clifford 1994) to compute the temporal similarity-based (non-symmetric) adjacency matrix, which links q_{kk} most similar observed–observed location pairs and q_{ku} most similar observed–unobserved (masked) pairs. GenCast constructs directed edges from observed to unobserved nodes only, to prevent unobserved representations from affecting observed node embeddings (by message passing). The temporal adjacency matrix for model training is denoted as $\mathbf{A}_{dtw}^{train} \in \mathbb{R}^{N_o \times N_o}$, and that for model testing is $\mathbf{A}_{dtw} \in \mathbb{R}^{N \times N}$. As masking is dynamic across epochs, $\mathbf{A}_{dtw}^{train}$ is updated at each training epoch.

Then, GenCast feeds $\mathbf{X}_{G_o}^{t-T+1:t}$ and $\mathbf{X}_{G_o^m}^{t-T+1:t}$ into a spatial-temporal (ST) model composed of stacked layers, each containing a temporal block and a spatial block operating in parallel. The temporal block uses 1-D temporal convolution networks (TCNs) to extract temporal dependencies, while the spatial block applies two graph convolution layers (GCNs) based on $\mathbf{A}_{dtw}^{train}$ and $\mathbf{A}_{sg}^{train}$, respectively. Their outputs are aggregated via element-wise maximum to form the spatial representation. The outputs of the TCN and GCN blocks are summed to produce the output of the l-th layer.

The final output of the ST model is passed through two linear layers with activation functions to generate a forecast $\hat{\mathbf{X}}_{G_o^m}^{t+1:t+T'}$. To obtain graph representations, we extract the output at the last predicted time step and apply a linear projection over all nodes, resulting in $\mathbf{Z}_{G_o}^{t+T'}$ and $\mathbf{Z}_{G_o^m}^{t+T'}$. For a batch size of $|B|$, this yields $2|B|$ representations.

To ensure that the representations are robust and invariant to masking, we apply contrastive learning between the original graph G_o and its masked variant G_o^m. For each time window t, the pair $(\mathbf{Z}_{G_o}^{t+T'}, \mathbf{Z}_{G_o^m}^{t+T'})$ is treated as a positive pair, encouraging consistency under masking. In contrast, pairs from different time windows ($t' \neq t$) are treated as negative samples. Accordingly, the contrastive loss is defined as:

$$L_{cl} = -\frac{1}{|B|} \sum_t \log \frac{\exp(\text{sim}(\mathbf{Z}_{G_o}^{t+T'}, \mathbf{Z}_{G_o^m}^{t+T'})/\omega)}{\sum_{t' \neq t} \exp(\text{sim}(\mathbf{Z}_{G_o}^{t+T'}, \mathbf{Z}_{G_o^m}^{t'+T'})/\omega)}, \quad (14)$$

where $\omega = 0.5$ is a temperature parameter and $\text{sim}(\cdot)$ denotes a similarity function (e.g., cosine similarity).

The final loss combines forecast error (RMSE) and representation alignment: $L = L_{pred} + \lambda L_{cl}$, where λ is a weighting factor. We follow the default hyperparameter from Su et al. (2024b) without fine-tuning in GenCast.

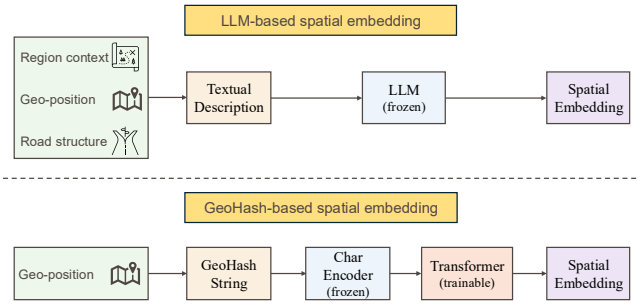


Figure 8: Comparison of two spatial embedding strategies. Upper: LLM-Based spatial embedding uses regional and road network information for spatial embedding. Lower: GeoHash-based spatial embedding only uses geo-location.

Sub-graph and Random Sub-graph Masking. We define the sub-graph of an observed location as its 1-hop neighbours. Given a masking ratio δ_m, we aim to mask approximately $N_o \cdot \delta_m$ observed locations. Since sub-graph sizes vary, GenCast randomly and iteratively selects observed nodes and masks both the node and its 1-hop neighbours, until the desired number of masked nodes is reached.

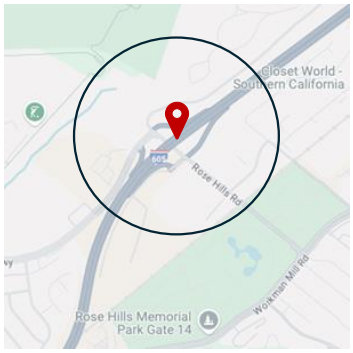
B.2 Spatial Embeddings of GenCast

Generating Process. We introduce two methods for generating continuously differentiable spatial embeddings: (1) *LLM-based spatial embedding* (SE-L), and (2) *GeoHash-based spatial embedding* (SE-H). These two methods differ in both their generation processes and update mechanisms. As illustrated in Fig. 8, SE-L is produced using a frozen large language model (LLM), ensuring that the geographic knowledge acquired during pre-training is retained. These embeddings remain fixed during training. In contrast, SE-H is generated using a pre-trained character-level encoder, and such embeddings are updated during training via a Transformer model.

Location Description Generation for SE-L. Generating location descriptions is the first step in our LLM-based spatial embedding module. We construct a textual description for each location by incorporating four categories of information: (i) the location's geo-coordinates, including latitude, longitude, and address; (ii) geometric properties describing the shape and extent of the surrounding area, particularly the spatial coverage of nearby POIs; (iii) POI information (Su et al. 2024b), such as the categories and counts of POIs within the region; and (iv) attributes of the closest road segments, including road type, maximum speed, number of lanes, and whether the road is one-way. Fig. 9 shows an example of a location description used as a prompt for generating spatial embeddings via a large language model (LLaMA 3 8B Instruct (Meta 2024)). The dimension of each spatial embedding generated by the LLM is 4,096.

B.3 Physics-informed Module in GenCast

The physics-informed module introduces a constraint based on physical principles of traffic systems formulated by the



The coordinates of the road is (37.25330, -121.94544), and the region near the road is a circle with radius of about 200m:
Included POIs:
"The region contains 1.0 Healthcare and Medical Facilities (25.0%), 3.0 Parking Facilities (75.0%), the value in bracket indicate the proportion of different POI categories, and the total number of POIs in this region is 4.0

The region is located near a motorway and trunk road. The maximum speed limit is 104.60736 km/h. The road is a oneway road.

The number of lane is 4."

Please infer the functionality of this region used for traffic forecasting, to enhance model generalisability.

Figure 9: An example of location description used as a prompt for generating LLM-based spatial embeddings.

Lighthill–Whitham–Richards (LWR) model (Lighthill and Whitham 1955; Richards 1956). LWR is a macroscopic first-order traffic flow model that describes the evolution of traffic density over space and time using a conservation law:

$$\frac{\partial \rho}{\partial t} + \frac{\partial (\rho x)}{\partial l} = 0, \tag{15}$$

where $\rho = \rho(l,t)$ denotes traffic density, $x = x(l,t)$ denotes traffic speed (i.e., traffic observation in our study), and t and l denote time and geo-location, respectively.

As benchmark datasets often lack traffic density observations, we rewrite the formulation using velocity, assuming a closed system with a functional density–velocity relationship (Greenshields 1935):

$$x = x_{fspd}\left(1 - \frac{\rho}{\rho_{max}}\right) \quad \Rightarrow \quad \rho = \rho_{max}\left(1 - \frac{x}{x_{fspd}}\right). \tag{16}$$

Here, x_{fspd} denotes Free-Flow Traffic Speed, which refers to the speed at which vehicles travel under low traffic density (i.e., no delays caused by congestion, signals, or other disruptions). Following Xiong, Zhou, and Bennett (2023), we define the 85th percentile of traffic speed during non-peak hours (i.e., 9:00am–4:00pm and 10:00pm–6:00am) as the free flow speed for each location, i.e., $[\mathbf{X}_{i,fspd} = \text{Percentile}_{85}\left(\mathbf{X}_i^t \mid t \in \mathcal{T}_{non-peak}\right), \forall i \in G_o]$. We estimate $\mathbf{X}_{i,fspd}$ for all observed locations from the training set. The spatial and temporal derivatives hence can be rewritten as:

$$\frac{\partial (\rho x)}{\partial l} = \rho_{max}\frac{\partial}{\partial l}\left(x - \frac{x^2}{x_{fspd}}\right) = \rho_{max}\left(1 - \frac{2x}{x_{fspd}}\right)\cdot\frac{\partial x}{\partial l}. \tag{17}$$

Further computing the temporal derivative yields:

$$\frac{\partial \rho}{\partial t} = \frac{\partial}{\partial t}\left[\rho_{max}\left(1 - \frac{x}{x_{fspd}}\right)\right] = -\frac{\rho_{max}}{x_{fspd}}\cdot\frac{\partial x}{\partial t}. \tag{18}$$

Then, this partial derivative is substituted into the conservation equation, Eq. 15:

$$-\frac{\rho_{max}}{x_{fspd}}\frac{\partial x}{\partial t} + \rho_{max}\left(1 - \frac{2x}{x_{fspd}}\right)\frac{\partial x}{\partial l} = 0. \tag{19}$$

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:

$$R = \frac{\partial x}{\partial t} + (2x - x_{fspd})\frac{\partial x}{\partial l} \tag{20}$$

To guide GenCast to follow the physical principles, we use the Huber loss (Huber 1992) to minimise the physics residual (i.e., to penalise violations of the physical law):

$$\mathcal{L}_{phy} = \text{Huber}(R, \delta) \tag{21}$$

which penalises violations of the physical principles and is jointly optimised with the main forecasting objective.

C Additional Experimental Details

C.1 Detailed Experimental Setup

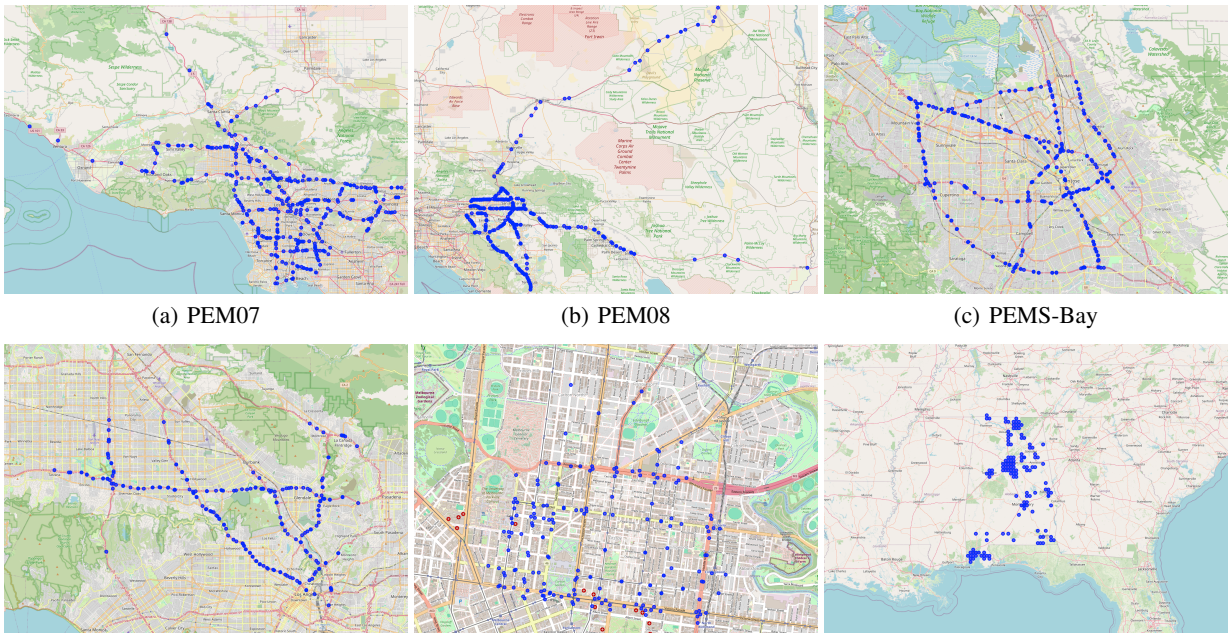
Datasets. We conduct experiments on four highway traffic datasets (PEMS-Bay, PEMS07, PEMS08 (CalTrans 2001) and METR-LA) and an urban traffic dataset (Melbourne). **PEMS-Bay**, **PEMS07**, and **PEMS08** contain traffic speed data collected by 358, 400 and 400 sensors in California, **METR-LA** (Li et al. 2018) contains traffic speed data collected by 207 sensors in Los Angeles County. All highway traffic datasets are collected at 5-minute intervals, i.e., 288 time slots per day. The urban traffic dataset Melbourne is collected by 182 sensors in Melbourne at 15-minute intervals, i.e., 96 time slots per day. Additionally, we generalise our proposed model GenCast to a solar power dataset, named NREL (NREL 2018), which contains solar power data collected by 137 sensors in Alabama. Table 3 lists the dataset statistics, and Fig. 10 visualises the sensor distribution among all datasets.

Implementation Details. For datasets with temporally missing values, we first apply interpolation to obtain complete data, which is then used for training, validation, and testing. Since STSM does not report results on METR-LA and NREL, we follow the default hyperparameter settings from their paper. For NREL, due to missing regional information, we replace STSM's selective masking with random subgraph masking. For shared model components, we adopt the same architectures and reuse STSM's reported hyperparameters to ensure fair comparison.

For our own hyperparameters, $sp = 5$, $cg = 2$ and $T_w = 12$ are determined using the LLM-based variant on PEMS08, and *the same values are applied across all datasets and model variants* without further tuning. We set $\mu = 1$ for all models on traffic datasets, and set $\mu = 0.1$

Dataset	Data Type	Time period	Interval	#Sensors
PEMS-07	Traffic Speed	01/09/2022 - 31/12/2022	5 min	400
PEMS-08	Traffic Speed	01/09/2022 - 31/12/2022	5 min	400
PEMS-Bay	Traffic Speed	01/01/2017 - 30/06/2017	5 min	325
METR-LA	Traffic Speed	01/03/2012 - 27/06/2012	5 min	207
Melbourne	Traffic Speed	01/07/2022 - 30/09/2022	15 min	182
NREL	Solar Power	01/01/2006 - 31/12/2006	5 min	137

Table 3: Dataset Statistics



(a) PEM07 (b) PEM08 (c) PEMS-Bay

(d) METR-LA (e) Melbourne (f) NREL

Figure 10: Visualisations of sensor distribution.

on the solar power dataset. We set $\theta = 5$ for the LLM-based model variant (GenCast-L) on all datasets, except for PEMS08 and Melbourne, where $\theta = 1$. For the GeoHash-based model variant (GenCast-H), we use $\theta = 0.01$ on PEMS07 and METR-LA, $\theta = 0.05$ on PEMS-Bay and PEMS08, and $\theta = 0.1$ on Melbourne. The observed differences in θ between GenCast-L and GenCast-H can be attributed to their distinct spatial embeddings, which yield different scales of physics residuals, e.g., on PEMS07, the 75th percentile is around 2 for GenCast-L and 84 for GenCast-H at the first training epoch. We set $\tau = 1.0$ on PEMS08 for both model variants and for the LLM-based variant on all datasets, except for PEMS-Bay and PEMS07, where $\tau = 0.85$. For the GeoHash-based variant, we use $\tau = 0.8$ on PEMS-Bay, $\tau = 0.9$ on PEMS07 and Melbourne, and $\tau = 0.95$ on MTRE-LA. A GeoHash string length of 8 is used for generating SE-H, except on Melbourne (which has a smaller region and a more dense sensor distribution), where the length is set to 9.

We adopt mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE) and R-Square (R^2) for evaluation:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - x_i)^2}, \quad \text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{x}_i - x_i|$$
$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{x}_i - x_i}{x_i}\right|, \quad R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{x}_i - x_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{22}$$

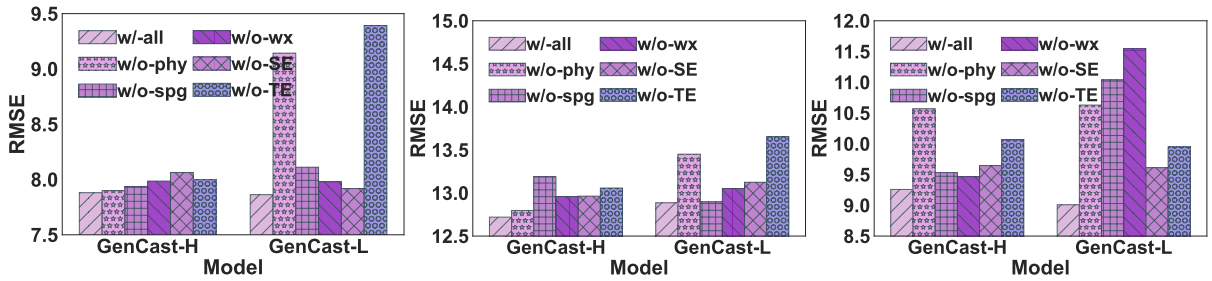
where, $\hat{x}$ and x denote the forecasted values and ground-truth values, respectively; $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and n denotes the number of samples.

C.2 Running Time Results

We report training and inference times across all traffic datasets. While our model incurs slightly longer training time due to its richer architecture, all models remain within a comparable time scale. Importantly, inference speed of GenCast, which matters most in deployment, is on par with other methods. Given that training is conducted offline, the small increase in training time is a reasonable cost. Considering the consistent improvements in generalisation performance, this trade-off is practically justified.

Model	Time	PEMS-Bay	PEMS07	PEMS08	Melbourne	METR
GE-GAN	Train (h)	4.4	4.1	4.1	0.3	3.4
	Test (s)	0.9	0.7	0.8	0.1	0.3
IGNNK	Train (h)	0.3	0.2	0.2	0.1	0.2
	Test (s)	8.3	7.8	8.8	2.4	2.6
INCREASE	Train (h)	0.3	0.2	0.2	0.2	0.1
	Test (s)	9.5	7.3	7.5	4.0	5.2
STSM	Train (h)	1.1	1.9	2.2	0.3	1.0
	Test (s)	1.6	1.3	1.2	0.3	2.0
KITS	Train (h)	1.0	1.5	1.2	0.2	1.1
	Test (s)	1.4	1.4	1.4	0.2	0.6
GenCast-H	Train (h)	3.0	2.6	2.2	2.9	1.4
	Test (s)	3.6	2.7	2.6	1.3	2.3
GenCast-L	Train (h)	3.0	2.1	1.9	2.1	1.3
	Test (s)	3.4	2.4	2.3	1.2	2.1

Table 4: Training and inference time comparison.



(a) PEMS08 (b) METR-LA (c) Melbourne

Figure 11: Ablation study results.

C.3 Ablation Study

Effectiveness of all components. We compare GenCast with five variants: **w/o-phy**, **w/o-spg**, and **w/o-wx** remove the physics constraint, spatial grouping loss, and external signal (i.e., weather) encoder, respectively; **w/o-SE** and **w/o-TE** remove spatial and temporal embeddings, respectively, together with the physics constraint. Fig. 11 presents ablation results on PEMS08, METR-LA, and Melbourne. All modules in GenCast contribute to its performance gains, as verified by the higher errors of all model variants. As observed on PEMS07 and PEMS-Bay (Fig. 4 in the main paper), the physics constraint is especially important for GenCast-L. Notably, it is also crucial for GenCast-H on the Melbourne dataset. This is because Melbourne is an urban dataset with densely located sensors (see Fig. 10 (e)), making it more difficult to differentiate by the spatial embeddings, even with SE-H. In such settings, the physics-guided loss becomes essential for helping the model effectively utilise spatial embeddings.

Additionally, w/o-spg and w/o-wx have similar impacts across datasets, showing their general applicability. On Melbourne, GenCast-L is more sensitive to the removal of the key modules, due to the high sensor density and urban spatial layout as mentioned above.

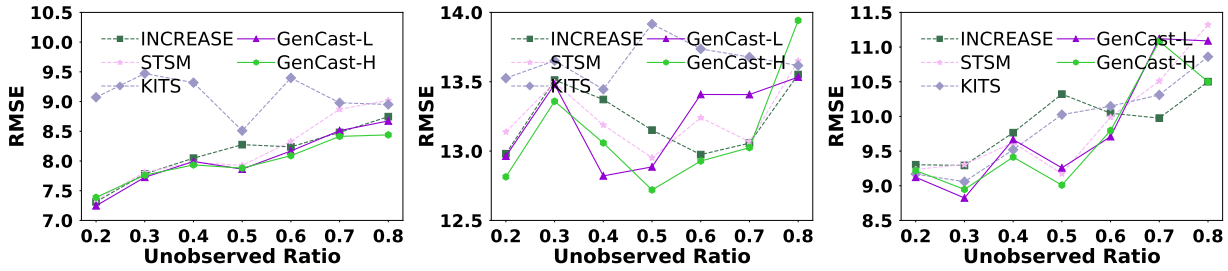
Effectiveness of Huber Loss. Table 5 compares model performance using dynamic Huber loss and fixed L2 loss (i.e., Huber with δ set to the 100%-quantile) on the physics residual (Eq. 10). When δ is large, Huber loss behaves like the L2 loss, treating all residual values equally. We observe that the distribution of physics residuals varies across datasets – some may contain more outliers or heavier tails. This motivates our adaptive calibration strategy, where δ is set based on a quantile of the initial error distribution during a warm-up stage. This lightweight adjustment improves robustness to outliers while retaining sensitivity to typical errors. Empirically, this strategy yields consistent (even though somewhat small) improvements across datasets and metrics. While simple, the method enhances training stability and generalisation with negligible overhead.

C.4 Parameter and Case Study

Impact of Unobserved Ratio. Fig. 12 presents model performance when we vary the unobserved ratio (i.e., percentage of unobserved nodes in G) from 0.2 to 0.8 on the PEMS08, METR-LA and Melbourne datasets. As before, we split each dataset either horizontally or vertically and report the performance averaged over four setups. We plot the top-3 best baselines with our models. Our GenCast, with either variant GenCast-H (using GeoHash embeddings) or GenCast-L (using LLM embeddings), performs the best in most cases (20 out of 21), confirming our proposed model GenCast's robustness against the unobserved ratio.

Dataset	Metric	GenCast-H	GenCast-H (L2)	GenCast-L	GenCast-L (L2)
PEMS-07	RMSE	8.285	8.425	8.253	8.311
	MAE	5.116	5.286	5.073	5.024
	MAPE	0.122	0.125	0.121	0.122
	R^2	0.193	0.166	0.197	0.189
PEMS-08	RMSE	7.880	7.880	7.863	7.863
	MAE	4.776	4.776	4.728	4.728
	MAPE	0.113	0.113	0.112	0.112
	R^2	0.146	0.146	0.150	0.150
METR-LA	RMSE	12.720	12.953	12.886	12.886
	MAE	8.792	8.983	8.799	8.799
	MAPE	0.265	0.272	0.267	0.267
	R^2	0.086	0.051	0.063	0.063
Melbourne	RMSE	9.009	9.530	9.258	9.258
	MAE	7.083	7.496	7.253	7.253
	MAPE	0.366	0.387	0.370	0.370
	R^2	0.061	-0.052	0.012	0.012
PEMS-Bay	RMSE	8.683	8.799	8.692	8.752
	MAE	5.192	5.193	5.139	5.143
	MAPE	0.131	0.132	0.131	0.132
	R^2	0.228	0.207	0.225	0.215

Table 5: Model Performance: dynamic Huber Loss vs. L2 loss on the physics residual.



(a) PEM08 (b) METR-LA (c) Melbourne

Figure 12: Model performance vs. unobserved ratio.

Impact of Hyperparameters. We study the impact of key hyperparameters, including the number of spatial/channel groups in the spatial grouping module, loss weights θ, μ, τ, and weather data input window length T_w.

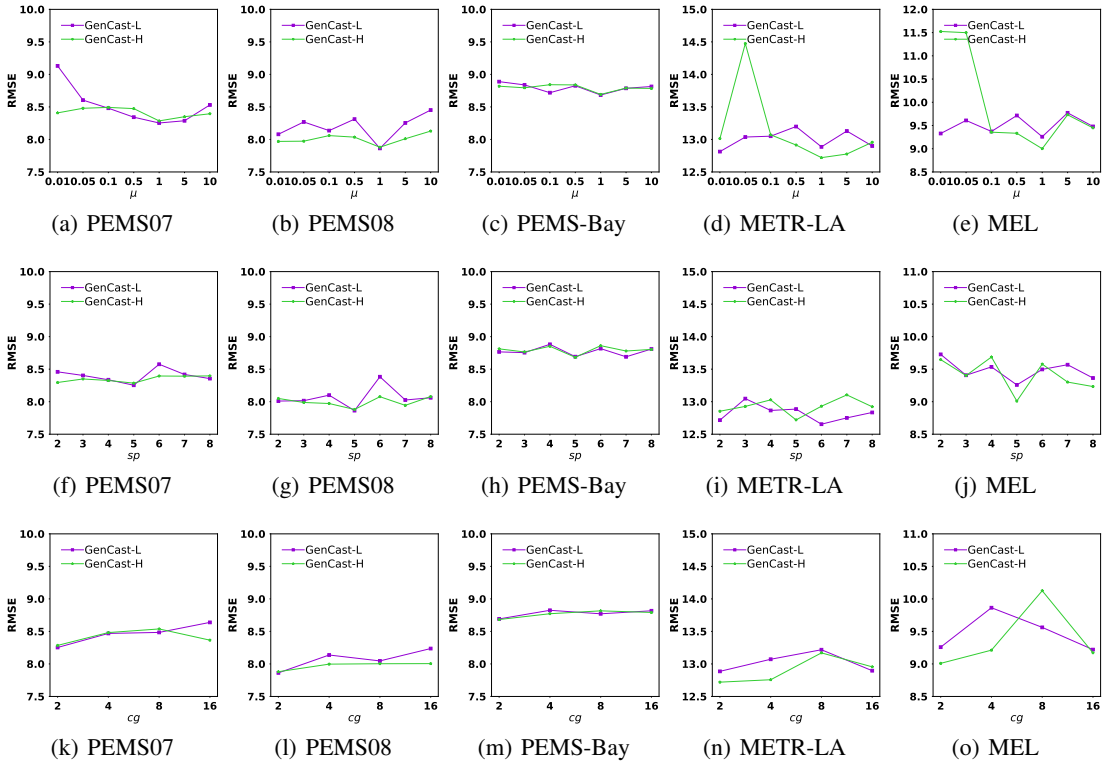
(1) *Impact of the weight μ of the spatial grouping loss.* This parameter controls the strength of the spatial grouping loss. As Fig. 13a to Fig. 13e show, $\mu = 1$ consistently produces lower errors across datasets and variants, indicating the robustness of our framework to this hyperparameter.

(2) *Impact of the number of spatial groups sp.* We present the impact of varying the number of spatial groups sp in Fig. 13f to Fig. 13j. We observe that $sp = 5$ consistently produces the best performance across all datasets and model variants, except for GenCast-L on METR-LA where $sp = 6$ is slightly better (though $sp = 5$ still achieves competitive results). Notably, this value was selected based solely on GenCast-L over PEMS08 without further tuning.

The forecast errors show a U-shaped pattern as sp increases: overly small number of groups fails to capture transferable spatial patterns, while overly large ones may capture overly specific, localised features that do not generalise well to unobserved regions. A moderate setting (e.g., $sp = 5$ or $sp = 6$) strikes a balance, effectively capturing shared structures while mitigating local noise.

(3) *Impact of the number of channel groups cg.* As shown in Fig. 13k to Fig. 13o, $cg = 2$ produces consistently low errors across datasets and model variants. This may be attributed to the fact that the number of feature channels D is typically much smaller than the number of spatial nodes N, and cg is required to divide D evenly. Using a small cg such as 2 offers a good trade-off between preserving feature coherence and enabling effective grouping. In contrast, overly large values of cg may fragment the features, making it more difficult to align with spatial groupings and weakening model generalisability.

(4) *Impact of th weight θ of the physics loss.* This parameter controls the contribution of the physics loss. As shown in Fig. 14a to Fig. 14e, GenCast-L has lower errors when θ has larger value (i.e., 1 or above), while GenCast-H has lower errors when θ has smaller values (i.e., 0.1 or below). These are expected, as the underlying spatial embeddings used in GenCast-L and GenCast-H differ, resulting in large



(a) PEMS07 (b) PEMS08 (c) PEMS-Bay (d) METR-LA (e) MEL

(f) PEMS07 (g) PEMS08 (h) PEMS-Bay (i) METR-LA (j) MEL

(k) PEMS07 (l) PEMS08 (m) PEMS-Bay (n) METR-LA (o) MEL

Figure 13: Impact of hyperparameters (I).

variations in the magnitude of the physics loss (e.g., 95th percentile: 21 vs. 992 on PEMS07). Consequently, the appropriate value of θ depends on the scale of the physical residuals induced by each embedding type.

(5) *Impact of the quantile used in dynamic Huber loss* τ. This parameter is used during the warm-up stage to optimise threshold δ for the Huber loss applied to physics residuals. It specifies the percentile of RMSE loss values, determining what proportion of residuals is treated as outliers in the adaptive calibration. As shown in Fig. 14f to Fig. 14j, GenCast-L shows greater robustness to this hyperparameter. We set $\tau = 100\%$ for it across most datasets, except for PEMS-Bay and PEMS07 where $\tau = 85\%$. This may be attributed to the use of frozen LLM-based spatial embeddings in GenCast-L, which leads to more stable physics residuals during training. GenCast-H relies on trainable GeoHash embeddings, making it more susceptible to outliers. Thus, suppressing the impact of large residuals via an adaptive Huber loss is particularly beneficial for GenCast-H.

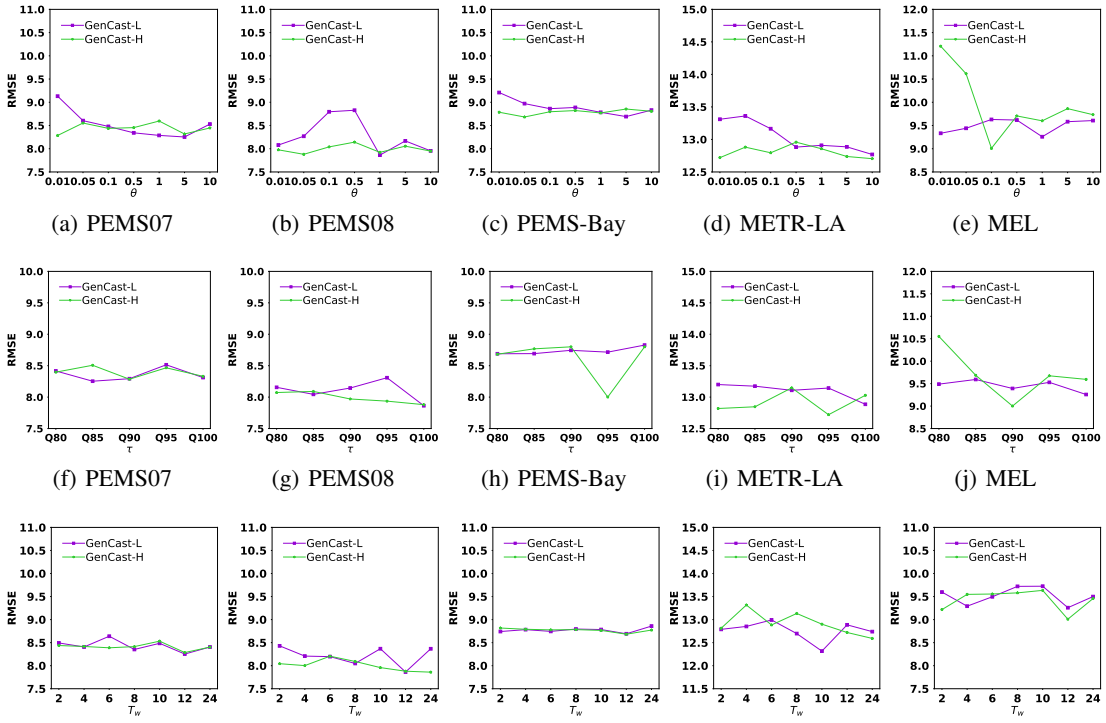
(6) *Impact of the weather look-back window length* T_w. As shown in Fig. 14k to Fig. 14o, our model forecast performance varies with T_w, which controls how many previous hours of weather data are used. We observe that $T_w = 12$ generally yields the lowest errors. This suggests that a 12-hour window provides a good temporal context to capture the delayed or cumulative impacts of weather on traffic dynamics. In contrast, short windows (e.g., $T_w = 2$ or 4) may miss such dependencies, while overly long windows (e.g., $T_w = 24$) may introduce irrelevant or outdated signals, thus diluting the more useful signals.

Generalisability of the External Signal Encoder. As shown in Table 6, incorporating our external signal encoder (**-w/wx**) consistently improves the performance of both INCREASE and KITS across all metrics. This confirms that our design, leveraging dynamic external signals to guide learning and enhance model generalisability to unobserved regions, is applicable beyond our model GenCast. The improvements are particularly pronounced in MAPE and R^2, indicating that external signals help reduce magnitude errors and better capture large-scale patterns. These results highlight the potential of integrating weather and other exogenous data as general-purpose enhancements to existing advanced models.

Models	RMSE↓	MAE↓	MAPE↓	R^2↑
INCREASE	8.534	6.045	2.730	0.177
INCREASE-w/wx	8.493	6.027	2.693	0.321
Improve	0.48%	0.30%	1.37%	81.36%
KITS	8.704	5.513	1.776	0.314
KITS-w/wx	7.715	4.582	0.746	0.496
Improve	12.82%	20.32%	138.07%	57.96%

Table 6: Generalisability of the external signal encoder.

Visualisation Weather-Traffic Attention. To understand the overall temporal alignment between traffic forecasts and weather input, we compute an average attention heatmap

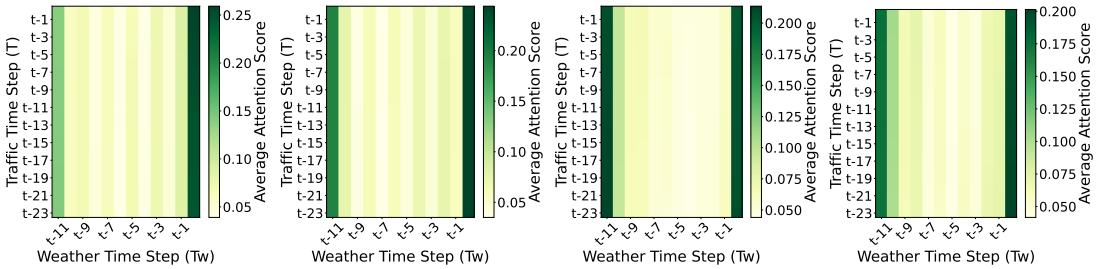


(a) PEMS07 (b) PEMS08 (c) PEMS-Bay (d) METR-LA (e) MEL

(f) PEMS07 (g) PEMS08 (h) PEMS-Bay (i) METR-LA (j) MEL

(k) PEMS07 (l) PEMS08 (m) PEMS-Bay (n) METR-LA (o) MEL

Figure 14: Impact of hyperparameters (II).



(a) Split type 1 (b) Split type 2 (c) Split type 3 (d) Split type 4

Figure 15: Average temporal attention to weather on the PEMS07 dataset.

over the PEMS07 dataset. We aggregate attention scores across all nodes and all test samples, resulting in a $T \times T_w$ matrix, where each entry indicates the average attention weight assigned from a given weather time step to a traffic prediction step. This global view reveals how GenCast utilises weather information at different temporal offsets.

The heatmap exhibits a bi-modal attention distribution. GenCast consistently focuses on both immediate weather observations (e.g., at t) and longer-range inputs (notably, at $t-11$ and $t-10$). This pattern suggests a reliance on both short-term conditions and broader temporal contexts. The periodicity in the distribution is likely a by-product of the 2-hour sliding window used during data generation, which introduces recurring emphasis on specific time steps. Furthermore, we observe slight variations in the distribution across different data splits, indicating that weather–traffic correlations may vary subtly between regions.