TRANSFERABLE SPATIO-TEMPORAL NODE IDENTITIES: ELEVATING MULTILAYER PERCEPTRONS TO LEADING CROSS-CITY ZERO-SHOT TRAFFIC FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatio-temporal forecasting models are conventionally trained and deployed per city, and the learnable node embeddings that empower state-of-the-art in-domain predictors are bound to the training network, precluding zero-shot transfer to unseen cities. We propose Transferable Spatio-Temporal Node Identities (TransSTNI), a paradigm that reconstructs node identity as a function computed from deployment-time priors rather than retrieved from dataset-bound parameters. TransSTNI derives node identity from three transferable sources: the Graph-Conditioned Structural Identity (GSI) encodes each node's structural role from the spectral coordinates of its own topology via a permutation-equivariant encoder; the Window-Conditioned Temporal-State Identity (WTI) captures node-level temporal states through multi-reference statistics computed over the current history window; and the Road-Attribute Semantic Embedding (RASE) provides functional context from de-geolocated road attributes through a frozen text encoder. These representations condition a residual MLP shared across all nodes, whose parameter count is independent of network size. To prevent memorization of the source topology, source-domain training interleaves full-graph learning with episodic optimization on variable-size induced subgraphs. At deployment, all parameters remain frozen: GSI and RASE are computed once and cached, and each window is predicted by a single forward pass without target-domain labels or adaptation. Experiments on LargeST demonstrate that TransSTNI achieves state-of-the-art zero-shot performance, while retaining the inference efficiency.The source code is available at repository: https://anonymous.4open.science/r/TSTNI.

1 INTRODUCTION

Traffic forecasting underpins a wide range of urban applications, yet the prevailing paradigm trains a dedicated model for each city. This per-dataset paradigm incurs a substantial cold-start cost: the deployment of a forecasting service in a new city requires months of sensor data collection followed by training from scratch, which is prohibitive for newly instrumented or data-scarce cities. Cross-city zero-shot traffic forecasting, i.e., transferring a model trained on source cities to an unseen target city without any target-domain training, therefore holds significant practical value Stickel et al. (2025b); Sun et al. (2026a).

Recent advances reveal that remarkably simple architectures can achieve state-of-the-art in-domain performance once equipped with learnable spatio-temporal embeddings. By assigning each node free parameters that encode its spatial-temporal identity, methods such as STID Shao et al. (2022a) and STAEformer Liu et al. (2023a) surpass sophisticated graph-based models with far fewer parameters and lower training cost. The underlying insight—that endowing the model with explicit spatio-temporal identity matters more than architectural complexity—is universal across cities. Its realization, however, is not. These identities are materialized as dataset-bound parameters indexed by node, which become undefined when the target city differs in node count or sensor layout. In effect, the embeddings memorize per-dataset patterns rather than capture transferable knowledge, reducing the universal insight to per-dataset memorization.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Efforts toward transferability have departed from this embedding paradigm in two directions. Meta-learning approaches Lu et al. (2022a) generate node-specific parameters via hypernetworks conditioned on node attributes, introducing considerable architectural complexity, while spatio-temporal foundation models Li et al. (2025); Chen et al. (2026a) attain zero-shot capability through large-scale multi-city pre-training at the cost of massive data and computation. Neither preserves the simplicity and efficiency that make the embedding family attractive. Neither preserves the simplicity and efficiency that make the embedding family attractive, and both operate under resource assumptions outside the strict zero-shot setting: the former requires target-domain samples for adaptation, while the latter presumes access to large-scale multi-city corpora for pre-training.

To bridge this gap, we propose **Transferable Spatio-Temporal Node Identities (TransSTNI)**, which abandons the index-bound identity paradigm and redefines node identity as a quantity computed at deployment time from available spatio-temporal priors rather than retrieved from fixed source-domain parameters. Concretely, TransSTNI models spatial structure and temporal dynamics through two computable identity embeddings. The Graph-conditioned Spatial Identity (GSI) encodes each node of the target topology with a permutation-equivariant encoder, deriving spatial representations purely from structural relational features instead of node indices, and thus generalizes to nodes and road networks unseen during training. The Window-conditioned Temporal Identity (WTI) estimates node-level temporal states from statistics of the real-time history window and maps them into identity features through pure forward inference, requiring no target-domain labels, parameter updates, or adaptation. Built upon a lightweight MLP predictor, the resulting architecture fully retains the inference efficiency of the embedding family while eliminating non-transferable lookup tables, constructing node identities solely from the target topology and real-time inputs. To further enhance generalization across diverse road networks, we introduce a subgraph rotation training strategy in the source domain: by continuously sampling multi-scale induced subgraphs with heterogeneous local topologies and alternately optimizing the spatial encoder, temporal mapper, and prediction decoder, the model is exposed to rich and varied local structures, which discourages memorization of the fixed source layout and fosters general, robust spatio-temporal mapping rules. This strategy requires no target-domain information and ensures stable forecasting performance on unseen target networks.

Our contributions are summarised as follows.

- We identify fixed index-based node embeddings as the core bottleneck limiting the cross-city generalization of traffic forecasting models with simple architectures. We abandon dataset-bound lookup embeddings and construct Graph-conditioned Spatial Identity (GSI) and Window-conditioned Temporal Identity (WTI) from spatial structure and temporal dynamics, addressing the non-transferable limitation of conventional node embeddings.
- We adopt residual MLP, a typical simple architecture, as the prediction backbone. We replace non-transferable node lookup tables with cross-domain spatio-temporal identities to maintain low inference overhead. We propose a subgraph rotation training strategy, which samples multi-scale heterogeneous local road networks in the source domain to mitigate the model's memorization of fixed source topology and boost generalization to unseen road networks.
- Experiments across diverse cross-domain scenarios validate that cross-domain node identities break the constraints imposed by source-domain node layouts. Stable zero-shot traffic forecasting is achieved on unseen cities without target-domain labels or parameter fine-tuning.

2 Related Work

2.1 Efficient Spatiotemporal Forecasting and Spatiotemporal Identity Embedding

Efficient spatiotemporal forecasting preserves spatiotemporal dependencies with limited parameters and computation. Lightweight models including DLinear Zeng et al. (2023), TSMixer Chen et al. (2023) and TimeMixer Wang et al. (2024) realize low-cost modeling via linear decomposition and feature mixing; iTransformer Liu et al. (2024), ModernTCN Luo & Wang (2024) and Pathformer Chen et al. (2024) strengthen long-range dependencies with specialized tokens, convolutions

and adaptive pathways Zhang et al. (2025a); Fang et al. (2025); Zhao et al. (2026); Wang et al. (2026); Huang et al. (2026); Yu et al. (2026). While such shared backbones perform major temporal transformations, they cannot handle node-set shifts and topology transfer.

This limitation arises from transductive, static designs of existing spatiotemporal embeddings. Temporal priors such as ToD/DoW and periodic encodings only offer global phases instead of per-node real-time traffic dynamics Guo et al. (2019); Zheng et al. (2020); Shao et al. (2022b). Learnable time slots and $T \times N$ adaptive tensors lock the model to fixed sampling settings and node scales Shao et al. (2022b); Liu et al. (2023b). Spatially, index-dependent node embeddings and adaptive graph components fail under node reordering, addition, deletion or topology replacement Wu et al. (2019); Bai et al. (2020); Wu et al. (2020); Shao et al. (2022b); Liu et al. (2023b).

To address this gap, we adopt a shared residual MLP and resort to explicit conditional modeling. GSI and WTI serve as computable identity functions: GSI produces spatial identities from target topology, and WTI extracts node-wise temporal states from historical windows for subsequent MLP prediction. Our framework maintains low overhead and enables frozen-parameter cross-network zero-shot inference with variable node counts.

2.2 Graph-Domain Generalization and Zero-Shot Spatiotemporal Forecasting

Most existing traffic transfer approaches follow a target-aware adaptation paradigm: they learn transferable patterns on the source domain and calibrate city- or node-level components with a small number of target-domain observations Yao et al. (2019); Lu et al. (2022b); Li et al. (2024); Zhang et al. (2025b). Further lines of work improve cross-dataset generalization via invariant representations and multi-city pre-training Li et al. (2022); Ji et al. (2025); Hao et al. (2026); Chen et al. (2026b). While these methods mitigate cross-domain distribution shifts, they typically require target observations, adaptation samples, or comprehensive pre-training coverage. When the node set, output dimension, and road topology are simultaneously replaced, they lack valid node-level references for previously unseen road networks.

Recent cross-domain mobility prediction explores stricter zero-shot protocols, where inference is performed directly on unseen cities without target trajectories or parameter fine-tuning Stickel et al. (2025a); Sun et al. (2026b). The key insight is to construct input-output references that remain valid across domains Stickel et al. (2025a); Sun et al. (2026b). Nevertheless, these approaches focus on individual-level user trajectory and next-location prediction and cannot be readily applied to synchronous continuous-value forecasting over all nodes of road-sensor graphs Stickel et al. (2025a); Sun et al. (2026b).

In contrast, under a frozen-parameter protocol, our method only uses target topology and current historical windows. GSI and WTI construct structural and dynamic conditions for each target node, enabling cross-network zero-shot traffic forecasting without target labels or adaptation-oriented optimization.

3 Preliminaries

Let the traffic sensor network be denoted as a graph $G = (V, E, A)$, where $V = \{v_i\}_{i=1}^{N}$ is the set of nodes, E represents the edge set that describes the topological connectivity between sensors, and $A \in \mathbb{R}^{N \times N}$ is the weighted adjacency matrix.

Given a historical observation of length T ending at time step t, the spatio-temporal forecasting model learns the mapping

$$[\boldsymbol{X}_{t-T+1}, \ldots, \boldsymbol{X}_t; G] \xrightarrow{\mathcal{F}(\cdot)} [\boldsymbol{X}_{t+1}, \ldots, \boldsymbol{X}_{t+H}], \tag{1}$$

where $[\boldsymbol{X}_{t-T+1}, \ldots, \boldsymbol{X}_t] \in \mathbb{R}^{T \times N \times C}$ denotes the historical input sequence, and the predicted output $[\boldsymbol{X}_{t+1}, \ldots, \boldsymbol{X}_{t+H}] \in \mathbb{R}^{H \times N \times C}$ represents future node states over the next H time steps.

This work studies zero-shot cross-city forecasting. The source city and target city are represented by $G^S = (V^S, E^S, A^S)$ and $G^T = (V^T, E^T, A^T)$, respectively. They are not required to share the same number of nodes, node-to-node correspondences, or topological structures. All trainable parameters for the forecasting task are learned solely from the source-domain training set $\mathcal{D}^S$ by

minimizing the empirical loss:

$$\theta^\star = \arg\min_\theta \mathbb{E}_{(\boldsymbol{X}_t^S, \boldsymbol{Y}_t^S)\sim\mathcal{D}^S}\Big[\mathcal{L}\big(f_\theta(\boldsymbol{X}_t^S; G^S), \boldsymbol{Y}_t^S\big)\Big]. \tag{2}$$

During deployment on the target city, the optimized parameters $\theta^\star$ remain frozen. The target-domain historical window $\boldsymbol{X}_t^T$ serves only as the immediate input for the forecasting mapping defined in Eq. equation 1. Accordingly, target-domain predictions take the form

$$\widehat{Y}_t^T = f_{\theta^\star}(\boldsymbol{X}_t^T; \boldsymbol{G}^T), \tag{3}$$

where the parameter set $\theta = \theta^\star$ is fixed during inference. That is, the model constructs required representations on an unseen graph and produces predictions via a single frozen forward pass following Eq. equation 3.

4 METHODOLOGY

This section introduces the TransSTNI framework for cross-city zero-shot traffic forecasting. As illustrated in Figure 1, we first present its overall architecture and three core transferable representation modules, then describe the training strategy for cross-domain generalization.

4.1 GSI: GRAPH-CONDITIONED STRUCTURAL IDENTITY

The spatial structural identity of a node is determined by its topological position within the road network. Although source and target cities do not share any identical physical nodes, functional structural roles such as hubs, corridor nodes, connection junctions, and peripheral nodes recur across different urban topologies and can be extracted purely from graph connectivity. GSI treats structural identity as a permutation-equivariant function of the input graph adjacency matrix, rather than an index-bound lookup table tied to fixed node identifiers.

Given raw adjacency matrix A, we first construct symmetrized augmented adjacency and symmetric normalized Laplacian as shown in Eq. equation 4:

$$\widetilde{A} = \frac{A + A^\top}{2} + I, \qquad L_{\text{sym}} = I - D^{-1/2}\widetilde{A}D^{-1/2}, \tag{4}$$

where $D = \text{diag}(\widetilde{A}\mathbf{1})$ is the degree matrix of $\widetilde{A}$. We select the smallest d_f non-trivial eigenvectors of L_{sym} to form spectral topological coordinates in Eq. equation 5:

$$F(A) = [u_2, \ldots, u_{d_f+1}] \in \mathbb{R}^{N\times d_f}. \tag{5}$$

Different from transductive node-indexed embedding tables, the spectral feature matrix $F(A)$ is computed on-the-fly purely from graph topology. It does not depend on predefined node indices and can be obtained for arbitrary-size road graphs. We feed these spectral topological coordinates into a shared structural encoder to derive final structural identities. Initialize layer input $H^{(0)} = F(A)$. The feature update at layer l is given in Eq. equation 6:

$$H_i^{(l+1)} = \sigma\Big(W_l\Big[H_i^{(l)} \parallel \text{MEAN}_{j\in\mathcal{N}(i)} H_j^{(l)}\Big]\Big), \tag{6}$$

where each layer shares weight W_l for all graphs. The encoder concatenates the node's own feature and the mean-aggregated features of its direct neighbors, and applies non-linear transformation. After L_s layers of local structural refinement, we apply layer normalization to produce the output structural identity matrix Z^{GSI}:

$$Z^{\text{GSI}} = \text{LN}(H^{(L_s)}) \in \mathbb{R}^{N\times d_z}. \tag{7}$$

The whole encoding mapping $\alpha_\theta(\cdot)$ satisfies the permutation-equivariant property formulated in Eq. equation 8:

$$\alpha_\theta(PAP^\top, PF) = P\,\alpha_\theta(A, F), \tag{8}$$

for any permutation matrix P. That is, if nodes are relabeled by permutation P, only the rows of output representations are permuted accordingly, while the encoding logic and the semantics

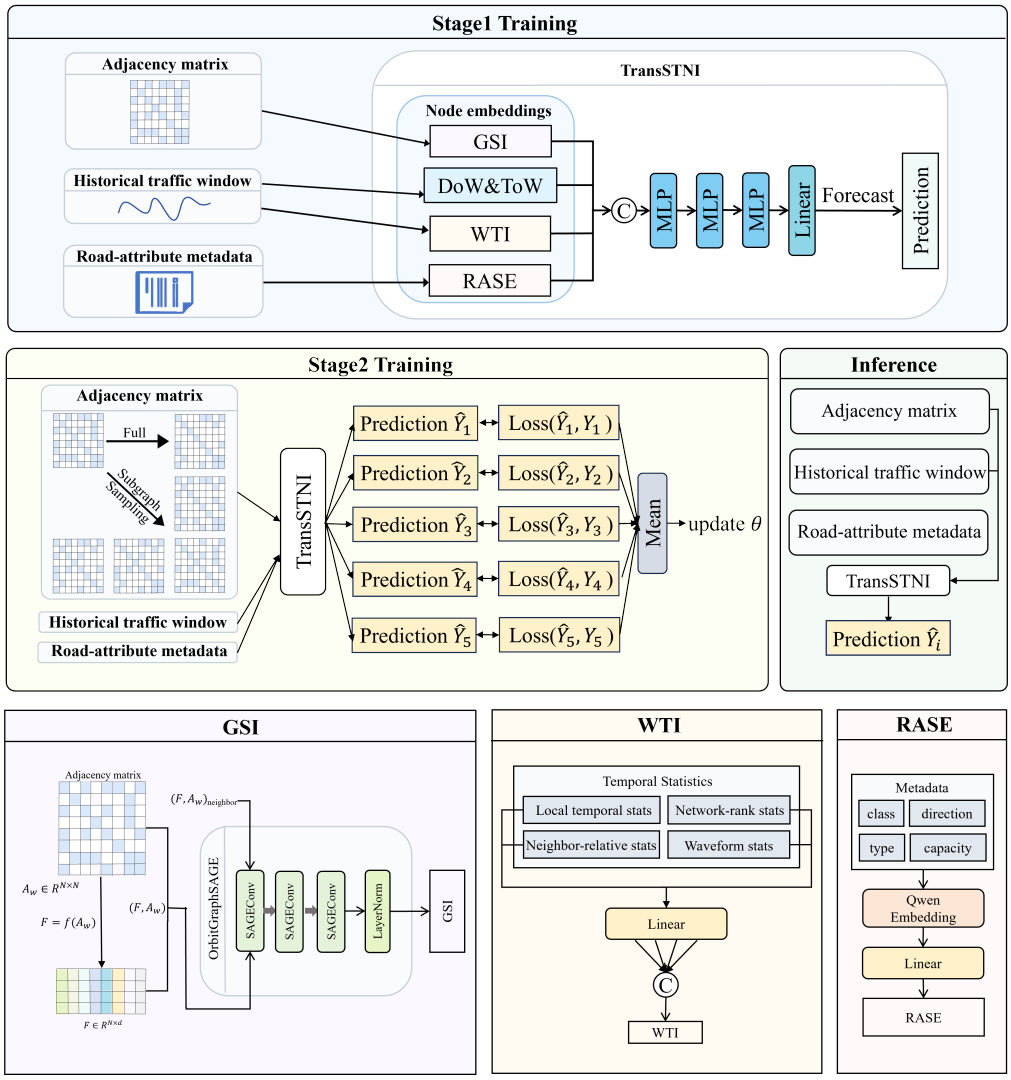

Figure 1: **Overview of the proposed TransSTNI framework.** **(a)** The framework comprises a training backbone and a zero-shot inference pipeline. The training backbone accepts weighted adjacency matrices and historical traffic sequences as inputs and constructs graph samples through subgraph sampling. It learns three distinct node representation branches, including GSI, WTI, and RASE, which are further complemented by conventional Time-of-Day (ToD) and Day-of-Week (DoW) temporal encodings. All feature representations are concatenated and subsequently processed by stacked MLPs and a linear projection layer for traffic prediction. The zero-shot inference pipeline follows a dedicated two-stage inference paradigm to support cross-domain generalization. **(b)** The subplots at the bottom elaborate on the detailed internal computation workflows of the GSI, RASE, and WTI modules.

of each node's structural representation remain unchanged. Critically, the parameter set of α_θ is agnostic to node number N. Once trained on the source graph, the identical encoder weights can be directly applied to target-city graphs with completely different node count and connectivity, without additional parameter fine-tuning.

Instead of seeking exact one-to-one node correspondence across cities, GSI learns a mapping from local graph substructures to spatial conditioning representations required for forecasting. Nodes with analogous local connectivity patterns yield similar structural identities, even across disjoint source and target graphs. For a static road topology, $\boldsymbol{Z}^{\text{GSI}}$ is invariant over different time windows;

thus it can be pre-computed and cached once during inference, avoiding repeated recomputation for each forecasting step.

4.2 WTI: WINDOW-CONDITIONED TEMPORAL-STATE IDENTITY

Structural roles alone cannot fully determine prediction conditions for nodes at a given time step. A single road segment can exhibit diverse states including stable flow, growth, decay, or anomalous fluctuations. Moreover, identical absolute traffic readings may carry different meanings across cities and neighborhoods. Conventional time-of-day and day-of-week embeddings only encode "what time it is", yet cannot differentiate "what state each node is in at this moment". To address this limitation, we propose WTI, which adopts Multi-Reference Temporal Statistics (MRTS) as its state extraction mechanism to produce window-aware temporal identities for each node. WTI applies deterministic MRTS statistical operators to the currently observable window X_t, where X_t denotes the observed traffic tensor of the current historical window and A represents the graph adjacency matrix. For node i, MRTS characterizes its instantaneous temporal state from four complementary perspectives, as formulated in Eq. 9:

$$s_{i,t}^{\text{local}} = \phi_{\text{local}}(X_{t,:,i}), \quad s_{i,t}^{\text{nei}} = \phi_{\text{nei}}(X_t, A)_i, \quad s_{i,t}^{\text{net}} = \phi_{\text{net}}(X_t)_i, \quad s_{i,t}^{\text{wave}} = \phi_{\text{wave}}(X_{t,:,i}). \tag{9}$$

where s^{local} summarizes window end value, mean, standard deviation, OLS trend, extreme-value range, and the offset of the end value relative to the window mean. s^{nei} takes temporal states of directly connected neighbors as references and describes relative deviations of the node trajectory in current level, window-averaged level, and trend. s^{net} treats the end values and window means of all nodes in the network as a global population and computes their empirical percentile ranks, so as to quantify relative magnitude among all sensors. s^{wave} summarizes low- and high-frequency energy proportions, the energy proportion of the dominant non-DC frequency component, and level/trend variations between the near and far sub-segments of the input window. These four descriptors capture local dynamics, neighborhood-relative dynamics, network-relative dynamics, and waveform dynamics for temporal identity modeling. Statistics from different groups differ in scale and semantics. We map them into latent spaces via learnable affine projections and assemble WTI along the channel dimension, where $\|$ denotes channel-wise concatenation, as shown in Eq. 10:

$$B_{i,t}^{\text{WTI}} = \left[\ell_{\text{local}}(s_{i,t}^{\text{local}}) \parallel \ell_{\text{nei}}(s_{i,t}^{\text{nei}}) \parallel \ell_{\text{net}}(s_{i,t}^{\text{net}}) \parallel \ell_{\text{wave}}(s_{i,t}^{\text{wave}})\right]. \tag{10}$$

WTI serves a distinct role from vanilla temporal encodings. Temporal encodings preserve fine-grained numeric sequences required for future prediction; WTI projects scale, trend, relative-magnitude, and waveform descriptors extracted by MRTS into low-dimensional conditions, enabling the shared MLP to recognize the instantaneous temporal state of each node. As such, WTI represents node-specific temporal conditions as node-wise embeddings that can be instantly generated for unseen cities.

4.3 AUXILIARY SEMANTIC CONTEXT AND IDENTITY-CONDITIONED MLP PREDICTOR

TransSTNI uses lightweight MLPs as the prediction backbone. It aggregates multi-source contextual information to build node embeddings. We extract transferable road attributes (road grade, travel direction, capacity bins, detector type) from source and target domain priors to form road-attribute semantic embeddings, with details in A.3. $H_{i,t}$ is the concatenated feature of node i at window t. It consists of historical traffic representations, structural identity embeddings, road-attribute semantics, temporal encodings, and ToD/DoW embeddingsas shown in Eq. equation 11.

$$\begin{aligned} H_{i,t} = \big[&E_{i,t}^{\text{hist}} \| Z_i^{\text{GSI}} \| e_t^{\text{tod}} \| e_t^{\text{dow}} \| Q_i^{\text{RASE}} \\ &\| \ell_{\text{local}}(s_{i,t}^{\text{local}}) \| \ell_{\text{nei}}(s_{i,t}^{\text{nei}}) \| \ell_{\text{net}}(s_{i,t}^{\text{net}}) \| \ell_{\text{wave}}(s_{i,t}^{\text{wave}})\big]. \end{aligned} \tag{11}$$

All spatial nodes share a unified residual MLP prediction backbone. The prediction output is defined in Eq. equation 12:

$$\widehat{Y}_t = \boldsymbol{F}(U^{(L_p)}). \tag{12}$$

We implement a three-layer residual MLP. The feature dimensions of each input branch in Eq. equation 11 are set to $[32, 32, 32, 32, 8, 16, 8, 8, 8]$, yielding a total input dimension of 176.

4.4 Two-stage Subgraph Training and Zero-shot Inference

Stage 1. We first perform fully supervised pre-training of the shared parameters on the full source graph G^S. The training objective of Stage 1 is denoted as $\mathcal{L}_{\text{s1}}$, defined in Eq. equation 13:

$$\mathcal{L}_{\text{s1}} = \mathcal{L}_{\text{Huber}}\big(\widehat{Y}^S,\, Y^S;\, G^S\big), \tag{13}$$

It computes the Huber loss between the prediction $\widehat{Y}^S$ and ground truth Y^S on the full source graph, to learn initial weights for the transferable identity generators and predictor.

Stage 2. Training solely on the complete source graph may cause representation generators to over-adapt to specific graph sizes and connectivity patterns. To enable the model to handle varying graph scales and local connectivity, we construct induced subgraphs of diverse sizes on the source domain. The sampled induced subgraph on the source domain is denoted as $\widetilde{G}^S$, following the subgraph sampling distribution $\mathcal{S}(G^S)$. GSI is recomputed from the induced adjacency matrix of this subgraph, and WTI is regenerated using the time series of corresponding nodes within the sampled windows. The training objective of Stage 2 is denoted as $\mathcal{L}_{\text{src}}$, defined in Eq. equation 14:

$$\mathcal{L}_{\text{src}} = \mathcal{L}_{\text{full}} + \mathbb{E}_{\widetilde{G}^S \sim \mathcal{S}(G^S)} \mathcal{L}_{\text{sub}}(\widetilde{G}^S), \tag{14}$$

where $\mathcal{L}_{\text{full}}$ denotes the loss term evaluated on the full source graph, and $\mathcal{L}_{\text{sub}}(\widetilde{G}^S)$ represents the loss term computed on the sampled subgraph $\widetilde{G}^S$. Both loss terms adopt the Huber loss function.

5 Experiment

To answer the following research questions(RQs), we conducts extensive empirical analyses.

- **RQ1:** How does TransSTNI compare to SOTA baselines on diverse datasets?
- **RQ2:** Is TransSTNI sensitive to identity dimension, predictor depth, and subgraph training intensity?
- **RQ3:** While attaining higher prediction accuracy, does TransSTNI retain the computational efficiency inherent to MLP architectures?

5.1 Experiment Settings

Dataset. We conduct experiments on three non-overlapping regions SD, GBA, and GLA from the LargeST dataset. These regions differ substantially in sensor scale, graph connectivity, and traffic-flow distribution, creating representative cross-domain distribution shifts. We split the dataset into training, validation, and test sets following a 6:2:2 ratio. Under standard traffic-forecasting window settings, we use $T = 12$ historical time steps to predict the subsequent $H = 12$ steps. Complete dataset statistics are provided in A.1.

Metrics. We evaluate our model from two dimensions: prediction accuracy and computational overhead. Three widely-adopted metrics including MAE, RMSE and MAPE are employed to quantify forecasting performance. Wall-clock time is reported to record training and inference runtime.

5.2 Cross-City Zero-Shot Forecasting (RQ1)

To verify TransSTNI's cross-city generalization, we conduct zero-shot experiments on SD, GBA and GLA. Under a unified protocol, all baselines are trained on one dataset and directly tested on the other two unseen datasets. We select seven mainstream spatio-temporal forecasting baselines: univariate models DLinear, PatchTST and STID, and multivariate models iTransformer, STGCN, STWave, PatchSTG and ST-GFSL. For fair comparison, we remove node spatial embeddings for graph-free baselines. Tables 1 and summarize the zero-shot results.

As shown in Tables 1, TransSTNI outperforms all baselines across all settings. Existing methods have clear transfer limitations: pure time-series models ignore road topology; graph-based models entangle spatio-temporal computation and overfit source topology; patch-based methods only generate region-level representations without node topological information; lookup embeddings are tied to source node indices. In contrast, TransSTNI leverages transferable node embeddings for robust zero-shot cross-city forecasting.

Table 1: Zero-shot cross-city traffic forecasting results on LargeST dataset. Bold indicates the best performance.

Methods	SD→GBA			SD→GLA			GBA→SD			GBA→GLA			GLA→SD			GLA→GBA		
	MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)
DLinear	47.82	75.17	42.49	51.71	81.86	44.74	50.06	81.02	40.20	51.81	81.36	46.84	50.45	82.32	36.80	48.23	75.94	42.32
PatchTST	48.54	77.93	40.27	50.61	83.64	30.72	50.13	83.28	31.47	49.22	81.68	30.55	50.58	82.93	31.99	47.29	74.96	39.88
STID	38.88	62.12	30.21	42.49	68.39	27.28	42.04	69.07	28.25	42.65	68.94	32.96	42.57	69.27	30.79	38.85	63.45	35.12
ST-GFSL	47.21	67.08	49.79	48.17	70.06	45.01	47.69	69.49	48.63	48.17	70.06	45.01	47.69	69.49	48.63	47.21	67.08	49.79
iTransformer	43.33	77.43	40.28	41.29	74.88	26.63	43.15	72.40	26.49	38.30	67.13	24.82	44.64	74.59	32.22	38.80	63.46	34.37
STGCN	38.58	60.37	34.88	43.17	69.97	37.71	34.62	52.20	26.40	35.57	53.56	28.53	33.46	51.74	23.94	32.88	49.81	27.60
STWave	45.56	69.46	46.42	50.80	76.53	66.37	48.33	73.56	37.27	53.85	82.62	48.67	49.01	77.19	38.65	45.49	69.67	40.43
PatchSTG	37.98	59.35	30.70	42.76	65.59	35.50	36.70	61.51	25.18	40.98	63.35	35.23	34.61	57.80	24.71	33.40	52.89	28.82
STUNet	34.46	53.90	26.88	36.77	57.98	24.55	32.71	52.16	21.90	33.06	50.96	24.42	32.19	49.44	22.41	32.39	47.52	26.41
TransSTNI	**32.02**	**50.35**	**21.75**	**34.66**	**55.43**	**19.41**	**30.95**	**48.57**	**19.72**	**32.46**	**50.84**	**20.20**	**31.10**	**49.35**	**20.63**	**30.84**	**47.36**	**23.58**

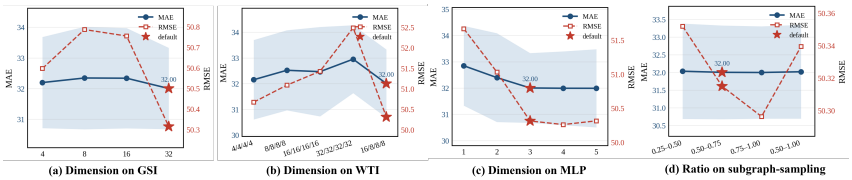

Figure 2: Hyperparameter sensitivity for TransSTNI zero-shot cross-city forecasting. Test set MAE and RMSE under varied hyperparameter settings: GSI embedding dimension, WTI embedding dimension, MLP predictor depth, and subgraph sampling ratio. The red star marks our default configuration. TransSTNI maintains stable performance within reasonable ranges of these hyperparameters; performance degrades when hyperparameters deviate substantially from the default values.

5.3 Hyperparameter Analysis (RQ2)

We conduct hyperparameter analysis of GSI dimension, WTI projection width, residual MLP depth and subgraph sampling ratio under SD→GLA/GBA, GLA→SD/GBA, GBA→SD/GLA transfer settings. The default configuration of the full model is: GSI dimension of 32, asymmetric WTI width of [16,8,8,8], residual MLP depth $L = 3$, and subgraph sampling ratio of 50%–75%. Figure 2 illustrates the impact of hyperparameters on the zero-shot transfer performance of TransSTNI.

GSI dimension. As shown in Figure 2(a), the performance of cross-city transfer tasks drops markedly as the GSI dimension decreases from 32. Higher-dimensional spaces can better encode transferable topological identities. To maintain stable and sufficient representation capacity across all datasets, we set the GSI dimension to 32.

WTI projection width. Figure 2(b) shows the default asymmetric [16,8,8,8] setup outperforms uniform-width alternatives with similar parameter counts. Narrow widths hinder multi-view temporal statistics modeling. Uniformly expanding all four branches to 32 causes severe parameter redundancy, brings no consistent performance improvement and hurts cross-city transfer in some cases. We thus adopt this asymmetric [16,8,8,8] configuration.

Residual MLP depth. As shown in Figure 2(c), the shallow configuration $L = 1$ causes severe underfitting and degrades transfer performance. Model performance improves with deeper layers and saturates at $L = 3$. Further increasing the depth yields negligible accuracy gains but increases computational overhead. We therefore set the default MLP depth to $L = 3$ to balance prediction accuracy and efficiency.

Stage 2 subgraph sampling ratio. Figure 2(d) shows that a sampling range below 50% degrades model performance significantly. After the sampling ratio exceeds 50%, further expanding the range yields no obvious performance improvement but increases computational overhead. We therefore select 50%–75% as the default setting.

Overall, the hyperparameter sensitivity experiments validate the optimal configuration of our full model: GSI dimension of 32, asymmetric WTI width of [16,8,8,8], residual MLP depth $L = 3$, and subgraph sampling ratio of 50%–75%.

Table 2: Comparison of the runtime efficiency of models across four datasets. BS indicates batch size, Train and Infer respectively denote per-epoch training and inference time in seconds, while Total stands for the overall training duration in hours. The symbol "-" denotes out-of-memory (OOM) errors even with the minimum batch size of 1.

Methods	SD				GBA				GLA			
	BS	Train	Infer	Total	BS	Train	Infer	Total	BS	Train	Infer	Total
DLinear	64	23	3	0.3	64	75	9	1	64	123	16	2
DCRNN	64	867	150	28	64	1816	319	59	43	2491	435	81
AGCRN	64	92	15	3	64	536	83	17	45	1413	245	46
STGCN	64	53	16	2	64	160	54	6	64	268	86	10
GWNET	64	97	14	3	64	483	66	15	64	1028	139	32
ASTGCN	64	128	19	4	45	1126	147	35	17	3060	393	77
STGODE	64	188	26	6	49	710	103	23	30	1305	192	42
DSTAGNN	64	240	23	7	27	1959	171	53	10	5241	467	120
DGCRN	64	430	76	14	12	4461	605	138	–	–	–	–
D^2STGNN	45	563	69	14	4	5885	796	148	–	–	–	–
STID	64	34	5	0.5	64	111	16	2	64	181	27	3
PatchSTG	64	64	6	1	64	262	30	4	64	295	27	4
STUNet	64	150	46	4	64	680	180	8	64	1140	312	12
TransSTNI	64	14	0.8	0.13	64	35	1.5	0.3	64	55	2.5	0.6

5.4 Efficiency Comparison (RQ3)

We empirically compare the runtime overhead of TransSTNI with other spatio-temporal forecasting models by measuring training wall-clock time and inference latency. Table 2 summarizes the results.

Our measurements reveal distinct trends. As node scale grows, several GNN-based baselines suffer out-of-memory (OOM) errors, with their computational cost rising super-linearly. In contrast, TransSTNI exhibits much slower growth in training and inference time under identical settings.

This benefit stems from the model pipeline. Static topology-aware representations like GSI and RASE can be pre-cached, avoiding repeated costly full-graph message passing during inference. This reduces computational burden for large road networks and becomes more prominent with more sensors.

6 Conclusion

In this paper, we present Transferable Spatio-Temporal Node Identities (TransSTNI), a lightweight framework for zero-shot cross-city traffic forecasting. Instead of static identities tied to source-domain node indices, TransSTNI computes transferable node representations on-the-fly from target-graph topology and historical observations. It extracts node-wise conditional signals via Graph-Conditioned Structural Identity (GSI) and Window-Conditioned Temporal-State Identity (WTI), enabling forecasting for unseen road networks without target-domain labels, adaptation data, or parameter fine-tuning. We further introduce a subgraph-driven training strategy to enhance robustness against varying graph sizes and topological shifts by training on diverse source-domain subgraph fragments.Experiments over multiple cross-domain forecasting settings verify that TransSTNI achieves effective zero-shot cross-network generalization with low parameter overhead and fast inference. We believe that this work is the first to realize cross-city zero-shot forecasting with fine-grained node discrimination and efficient inference under a lightweight MLP backbone. Our computable node-identity paradigm provides a promising path toward scalable, high-efficiency intelligent zero-shot traffic forecasting systems.

REFERENCES

Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. Adaptive graph convolutional recurrent network for traffic forecasting. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17804–17815, 2020.

Peng Chen, Yingying Zhang, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang, and Chenjuan Guo. Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting. In *International Conference on Learning Representations*, 2024.

Si-An Chen, Chun-Liang Li, Nate Yoder, Sercan O. Arik, and Tomas Pfister. Tsmixer: An all-mlp architecture for time series forecasting. *Transactions on Machine Learning Research*, 2023.

Wei Chen, Yuqian Wu, Junle Chen, Xiaofang Zhou, and Yuxuan Liang. Urbanfm: Scaling urban spatio-temporal foundation models. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 10635–10645, 2026a.

Yujun Chen, Shihao Tu, Wenyue Ding, Yicheng Lu, Qingkai Ren, Yangjie Zheng, and Yang Yang. Unified spatio-temporal tokens are bases for generalizable traffic forecasting. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 508–519, 2026b.

Yuchen Fang, Yuxuan Liang, Bo Hui, Zezhi Shao, Liwei Deng, Xu Liu, Xinke Jiang, and Kai Zheng. Efficient large-scale traffic forecasting with transformers: A spatial data management perspective. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 307–317, 2025.

Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, volume 33, pp. 922–929, 2019.

Xinyan Hao, Huaiyu Wan, Shengnan Guo, Shaojiang Wang, and Youfang Lin. Tracking topological shifts: How can dynamic graph invariant learning enable reliable out-of-time spatio-temporal prediction? In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*, pp. 2662–2670, 2026.

Sheng Huang, Guanjun Wang, Jiaming Ma, Binwu Wang, and Yang Wang. Minist: Unlocking input window length in traffic flow forecasting with compact parameters. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*, pp. 2680–2688, 2026.

Jiahao Ji, Wentao Zhang, Jingyuan Wang, and Chao Huang. Seeing the unseen: Learning basis confounder representations for robust traffic prediction. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 577–588, 2025.

Haoyang Li, Ziwei Zhang, Xin Wang, and Wenwu Zhu. Learning invariant graph representations for out-of-distribution generalization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11828–11841, 2022.

Zhonghang Li, Lianghao Xia, Yong Xu, and Chao Huang. Flashst: A simple and universal prompt-tuning framework for traffic prediction. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 28978–28988. PMLR, 2024.

Zhonghang Li, Long Xia, Lei Shi, Yong Xu, Dawei Yin, and Chao Huang. Open spatio-temporal foundation models for traffic prediction. *ACM Transactions on Intelligent Systems and Technology*, 2025.

Hangchen Liu, Zheng Dong, Renhe Jiang, Jiewen Deng, Jinliang Deng, Quanjun Chen, and Xuan Song. Spatio-temporal adaptive embedding makes vanilla transformer sota for traffic forecasting. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pp. 4125–4129, 2023a.

Hangchen Liu, Zheng Dong, Renhe Jiang, Jiewen Deng, Jinliang Deng, Quanjun Chen, and Xuan Song. Spatio-temporal adaptive embedding makes vanilla transformer sota for traffic forecasting. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 4125–4129, 2023b.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, pp. 11116–11140, 2024.

Bin Lu, Xiaoying Gan, Weinan Zhang, Huaxiu Yao, Luoyi Fu, and Xinbing Wang. Spatio-temporal graph few-shot learning with cross-city knowledge transfer. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 1162–1172, 2022a.

Bin Lu, Xiaoying Gan, Weinan Zhang, Huaxiu Yao, Luoyi Fu, and Xinbing Wang. Spatio-temporal graph few-shot learning with cross-city knowledge transfer. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1162–1172, 2022b.

DongHao Luo and Xue Wang. Moderntcn: A modern pure convolution structure for general time series analysis. In *International Conference on Learning Representations*, pp. 31728–31770, 2024.

Zezhi Shao, Zhao Zhang, Fei Wang, Wei Wei, and Yongjun Xu. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pp. 4454–4458, 2022a.

Zezhi Shao, Zhao Zhang, Fei Wang, Wei Wei, and Yongjun Xu. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, pp. 4454–4458, 2022b.

Jonas Stickel, Andrea Tundis, and Max Mühlhäuser. Zero-shot cross-city trajectory prediction using hypernetworks. In *2025 IEEE International Conference on Data Mining*, pp. 307–316, 2025a.

Jonas Stickel, Andrea Tundis, and Max Mühlhäuser. Zero-shot cross-city trajectory prediction using hypernetworks. In *2025 IEEE International Conference on Data Mining (ICDM)*, pp. 307–316. IEEE, 2025b.

Tianao Sun, Kai Zhao, Weiming Huang, Ziqiang Yu, and Meng Chen. Where to go next: Enhancing zero-shot capability for cross-city mobility prediction. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 4520–4531, 2026a.

Tianao Sun, Kai Zhao, Weiming Huang, Ziqiang Yu, and Meng Chen. Where to go next: Enhancing zero-shot capability for cross-city mobility prediction. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4520–4531, 2026b.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations*, pp. 38626–38652, 2024.

Yongyao Wang, Xie Yu, Jingyuan Wang, Jiahao Ji, and Li Chao. Hieramix: A hierarchical mlp-mixer for large-scale traffic forecasting. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*, pp. 6555–6563, 2026.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 1907–1913, 2019.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 753–763, 2020.

Huaxiu Yao, Yiding Liu, Ying Wei, Xianfeng Tang, and Zhenhui Li. Learning from multiple cities: A meta-learning approach for spatial-temporal prediction. In *Proceedings of The Web Conference 2019*, pp. 2181–2191, 2019.

JiaJun Yu, Fang Yuan, Guang-Yong Chen, and Min Gan. Active arbitration: Decoupling spatio-temporal duality for efficient traffic forecasting. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*, pp. 5236–5244, 2026.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 11121–11128, 2023.

Qianru Zhang, Xinyi Gao, Haixin Wang, Siu-Ming Yiu, and Hongzhi Yin. Efficient traffic prediction through spatio-temporal distillation. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence*, volume 39, pp. 1093–1101, 2025a.

Yudong Zhang, Xu Wang, Xuan Yu, Zhaoyang Sun, Kai Wang, and Yang Wang. Drawing informative gradients from sources: A one-stage transfer learning framework for cross-city spatiotemporal forecasting. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence*, volume 39, pp. 1147–1155, 2025b.

Yiji Zhao, Zihao Zhong, Ao Wang, Haomin Wen, Ming Jin, Yuxuan Liang, Huaiyu Wan, and Hao Wu. Fast: Efficient and effective long-horizon forecasting for large-scale spatial-temporal graphs via mixture-of-experts. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 1975–1986, 2026.

Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianzhong Qi. Gman: A graph multi-attention network for traffic prediction. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, volume 34, pp. 1234–1241, 2020.

A Appendix

A.1 Additional Dataset Details

We conduct experiments on three non-overlapping subsets (SD, GBA, and GLA) of the LargeST benchmark dataset. This dataset contains five consecutive years (2017–2021) of highway traffic flow time-series records, with the raw data sampled at a 5-minute interval. And the statistics of the number of sensors, topological edges, average node degree, and time range for each regional sub-dataset are shown in Table X. To standardize experimental settings and mitigate temporal redundancy, we retain only the data from the year 2019. The original 5-minute sampling granularity is downsampled to a 15-minute resolution. After preprocessing, each traffic sensor yields 35040 valid time-series samples. This preprocessing pipeline is widely adopted in related traffic forecasting literature.

Table 3: Dataset statistics.

Datasets	Sensors	Edges	Degree	Time Range
SD	716	17,319	24.2	01/01/2019-12/31/2019
GBA	2,352	61,246	26.0	01/01/2019-12/31/2019
GLA	3,834	98,703	25.7	01/01/2019-12/31/2019

For data splitting, we partition the processed time-series into training, validation, and test sets following a 6:2:2 split ratio. The forecasting task is unified: given 12 historical time steps, the model predicts traffic conditions for the subsequent 12 steps. Three standard evaluation metrics are adopted to quantitatively assess the prediction accuracy and generalization performance: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE).

A.2 Training Settings

In our model, the Laplacian feature dimension is set to 8, the output dimension of Graph Spatial Identity (GSI) embedding is set to 32 (with a 64-dimensional hidden layer in GraphSAGE). A three-layer GraphSAGE is adopted to model the transferable spatial identity. The four projection widths of Window Temporal Identity (WTI) are 16/8/8/8, respectively. The road semantic embedding (RASE) has an input dimension of 1024 and a projection dimension of 8. Both the TOD and DOW time embeddings are set to a dimension of 32. The Adam optimizer is used for model optimization

with a learning rate of 2×10^{-3} and weight decay of 1×10^{-4}. The backbone follows the Multi-Layer Perceptron (MLP) architecture, where the backbone embedding dimension is set to 32. In the subgraph sampling training stage, the task number is set to 4 and the meta learning rate is 1×10^{-4}. We implement TransSTNI using PyTorch 2.6.0, and all experiments are conducted on an NVIDIA A800 GPU with 80 GB of VRAM.

A.3 RASE Implementation Details

The Road-Attribute Semantic Embedding (RASE) module transforms raw attribute metadata of road-network detectors into standardized, transferable semantic text features, providing stable road-network semantic priors for the model. All attributes are directly sourced from the native metadata of the LargeST dataset without additional annotations or external data.

We assemble road-network attributes following a fixed structured text template, where road-network features are denoted by alphabetical variables:

> Highway traffic detector. Road class C. Travel direction D. N lanes. Detector type T. Capacity bin: Ca. S. Do not use city, county, administrative district, road-number or geographic identifiers.

Each variable corresponds to a native metadata field in the LargeST dataset, with concise mappings defined below:

- C: road-class attribute that distinguishes different highway hierarchies.
- D: travel direction of the road segment.
- N: actual number of lanes on the road segment.
- T: hardware type of the traffic detector.
- Ca: standardized capacity level mapped from lane counts.
- S: supplementary semantic description for road-network traffic characteristics.

To filter non-transferable geographic redundancy and unify the semantic encoding paradigm, the concatenated semantic text is programmatically wrapped into an instruction prompt following the official input format of Qwen3-Embedding. Only general transferable functional road-network attributes are retained, while all geography-specific information is discarded. The prompt template is shown below:

> Instruction: Characterize this highway traffic detector using only transferable functional attributes (road class, travel direction, lane count, detector type). Ignore city, county, administrative district, road number, and geographic identifier information.
> Query: [concatenated road-network semantic text]

Taking an arterial-road detector sample from the SD subset of LargeST as an example, the complete semantic text fed into the embedding model is:

> Highway traffic detector. Road class: interstate highway. Travel direction: northbound. 6 lanes. Detector type: mainline. Capacity bin: high throughput arterial. Restricted-throughput arterial traffic-flow sensor. Do not use city, county, administrative district, road-number or geographic identifiers.

Built upon native dataset metadata, RASE semantic features exhibit strong cross-scene generalization. They support pre-computation and caching, and effectively improve the inference efficiency of our model for cross-city zero-shot forecasting.

A.4 Comparison with Target-domain Supervised Upper Bound

To quantify the performance degradation under strict zero-shot constraints, we retrain TransSTNI on target-city data to build a fully supervised upper bound named TransSTNI-Sup. As shown in

Table 4: In-domain supervised upper-bound results. The models are trained and evaluated on the same target dataset with target-domain supervision. Bold indicates the best performance.

City	Stage 1			Stage 2		
	MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)
SD	25.70	42.32	16.95	**25.62**	**42.15**	**16.57**
GBA	27.46	43.95	18.51	**27.11**	**43.44**	**18.03**
GLA	27.58	45.14	16.18	**26.76**	**43.82**	**15.19**
Average	26.91	43.80	17.21	**26.50**	**43.14**	**16.60**

Table 4, TransSTNI-Sup achieves the lowest prediction errors. We further compare source-domain performance before and after two-stage subgraph training under full supervision to check whether subgraph training hurts source fitting. Table 4 shows Stage 2 consistently reduces MAE, RMSE and MAPE across SD, GBA and GLA. The results indicate variable-scale subgraph training does not sacrifice source accuracy for cross-domain generalization. Instead, it enables the model to learn diverse local structures, strengthens graph heterogeneity modeling, and improves both source prediction and cross-city generalization.

A.5 Ablation Study

We evaluate the contribution of each component by ablating one of the following four modules while keeping all other experimental settings identical to the full model:

- "w/o RASE": TransSTNI removes the road semantic embedding.
- "w/o GSI": TransSTNI removes the graph spatial identity embedding.
- "w/o WTI": TransSTNI removes the window temporal identity.
- "w/o Stage2": TransSTNI skips the source-domain subgraph Empirical Risk Minimization adaptation, and only uses the pre-trained weights from Stage 1 for cross-domain zero-shot evaluation.

Table 5: Ablation study of TransSTNI. Bold indicates the best performance.

Dataset	SD			GBA			GLA		
Metric	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)
w/o RASE	34.00	54.59	23.07	31.85	50.50	20.21	31.91	50.27	25.53
w/o WTI	44.03	89.89	30.38	38.71	67.13	24.67	36.95	60.54	28.67
w/o GSI	33.39	53.01	23.01	31.97	49.95	20.88	32.28	48.77	25.21
w/o Stage2	33.59	53.18	23.73	33.18	49.81	22.67	31.95	49.39	26.09
TransSTNI	**33.34**	**52.89**	**20.58**	**31.70**	**49.70**	**19.96**	**31.90**	**48.35**	**22.11**

Table 5 shows the quantitative ablation results.

Role of Core Representation Modules. A clear performance drop is observed across SD, GBA and GLA when RASE is removed. Without transferable road-semantic priors from RASE, spatial and temporal identities alone cannot fully capture the intrinsic functional properties of road nodes. Meanwhile, "w/o WTI" and "w/o GSI" also suffer large errors, since missing either temporal or spatial component breaks the complete transferable node-representation system for zero-shot cross-city forecasting.

Effectiveness of two-stage subgraph training. The "w/o Stage2" variant yields degraded accuracy on all datasets. Trained only with Stage1 full-graph pre-training, the model cannot benefit from subgraph-aware source-domain adaptation and diverse local topological patterns. Stage2 subgraph optimization strengthens graph-heterogeneity modeling, boosting cross-city generalization without hurting in-domain performance.

Completeness of the node identity representation system. Removing any core module raises prediction errors, where full TransSTNI obtains the best overall results. Among all components, ablating WTI causes the most severe performance degradation across three datasets. Despite divergent

topologies among cities, sensor temporal dynamics universally depend on window-level statistics. Static structural and semantic signals from GSI and RASE are inadequate to represent time-varying instantaneous traffic states, making WTI the most vital module for zero-shot cross-city transfer.