SEEING THE UNSEEN NODES: ZERO-SHOT AND FEW-SHOT TRAFFIC FLOW PREDICTION VIA SEMANTIC PROTOTYPE-BASED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Traffic flow forecasting is a fundamental task in intelligent transportation systems. However, most existing methods assume a fixed set of sensors and operate under a transductive setting, limiting their ability to handle continuously expanding traffic networks. While previous studies have explored traffic expansion forecasting, they typically rely on sufficient observations from newly deployed sensors. In contrast, we study a more challenging setting in which new sensors come with only a few or even no historical observations, referred to as few-shot and zero-shot Traffic Expansion Flow Forecasting (TEFF). To tackle this problem, we propose a Semantic-driven Prototype-based Spatio-Temporal framework (SemProtoST). Specifically, we first introduce a metadata-driven node representation to enable inductive generalization to unseen nodes. We then develop dual spatio-temporal pattern prototypes, where node metadata and temporal features query shared prototypes via a cross-attention mechanism to model traffic patterns. Furthermore, we incorporate a semantic consistency constraint to enhance generalization to unseen nodes. To support few-shot adaptation, we design a knowledge distillation strategy that preserves performance on seen nodes while enabling effective adaptation to unseen nodes. Extensive experiments on four real-world datasets demonstrate that SemProtoST achieves average improvements of 3.58% and 14.22% on unseen nodes in few-shot and zero-shot settings, respectively. Our code is available at https://anonymous.4open.science/r/SemProtoST.

1 INTRODUCTION

Traffic flow forecasting is a fundamental task in Intelligent Transportation Systems (ITS), supporting a wide range of smart-city applications (Shao et al., 2022b; Chen et al., 2018; Wang et al., 2021; Lv et al., 2015). Recent Spatio-temporal Graph Neural Networks (STGNNs) have become the dominant approach for traffic forecasting(Shao et al., 2025). These methods represent sensors in urban networks as nodes and their connections as edges. To capture spatio-temporal dependencies, STGNNs combine graph neural networks with temporal sequence learning, where graph neural networks are employed to model spatial correlations(Yu et al., 2018; Bai et al., 2020; Shao et al., 2022b), while sequential models are used to extract temporal dynamics(Li et al., 2018; Wu et al., 2019; Xu et al., 2020). Additionally, Shao et al. (2022a) and Dong et al. (2024) has been shown that incorporating specific learnable parameters into STGNNs can enhance predictive performance. However, recent advancements have focused on designing novel architectures to improve accuracy, often overlooking the practical demands arising from the expansion of urban road networks in real-world scenarios.

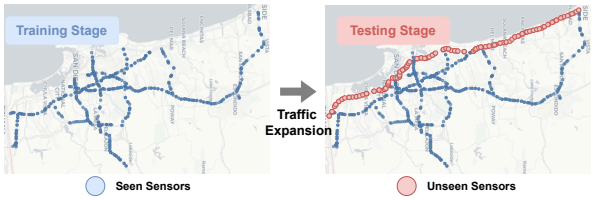


Figure 1: Illustration of the Traffic Expansion.

In this paper, we study *Traffic Expansion Flow Forecasting* (TEFF). As illustrated in Fig. 1, the traffic network expands beyond its original coverage with the addition of newly deployed sensors. We define the initially deployed sensors as *seen sensors* and the newly added sensors as

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.

unseen sensors. Unlike existing traffic expansion studies that typically assume sufficient observations from newly deployed sensors (Chen et al., 2021; Chen & Liang, 2025), we focus on a more challenging setting where unseen sensors have only a few or even no historical observations. This enables traffic forecasting immediately after sensor deployment, which is important for rapidly integrating newly expanded areas into traffic management systems(Ma et al., 2025).

As shown in Fig. 2, few-shot and zero-shot TEFF introduce three challenges beyond traditional traffic forecasting: **Challenge 1: Identity Dependency and Scalability.** Typical traffic forecasting models follow a transductive paradigm, relying on learnable Node ID embeddings to characterize spatial heterogeneity. Such representations are tied to the sensors observed during training and cannot be directly constructed for newly deployed sensors, making these models difficult to generalize to dynamically expanding networks. **Challenge 2: Spatio-Temporal Inductive Inference.** Removing Node ID dependency alone does not guarantee effective generalization. Unseen sensors may differ substantially from seen sensors in their metadata, such as geographic locations, road attributes, and surrounding POI information, resulting in different traffic dynamics. The key challenge is therefore to learn universal spatio-temporal regularities from seen sensors and infer the latent traffic patterns of unseen sensors from their available metadata. **Challenge 3: The Trade-off between Stability and Plasticity.** When a few observations from newly deployed sensors become available, fine-tuning the model can improve its performance on unseen sensors but may degrade its previously learned forecasting capability on seen sensors due to catastrophic forgetting (Chen et al., 2021; Wang et al., 2023b). Therefore, effective few-shot fine-tuning requires a balance between the plasticity on unseen nodes and stability on seen nodes.

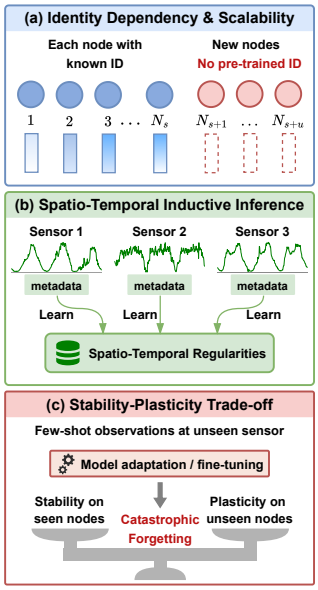


Figure 2: TEFF Challenges.

To address these challenges, we propose a Semantic-driven Prototype-based Spatio-Temporal framework (SemProtoST) for few-shot and zero-shot Traffic Expansion Flow Forecasting. For **Challenge 1**, we introduce a metadata-driven node representation (MetaNR) that replaces conventional node ID embeddings with static node attributes. This design enables the model to generate meaningful representations for unseen nodes without relying on historical observations, thereby equipping it with inductive capability for the TEFF task. For **Challenge 2**, we propose dual spatio-temporal pattern prototypes (DSTPP) to learn a set of representative spatio-temporal prototypes. Using a cross-attention mechanism, node metadata and periodic temporal features are utilized to query these prototypes, thereby modeling spatio-temporal patterns. We further introduce a semantic consistency constraint (SCC) that aligns spatial-prototype usage between semantically similar sensors, which enhances the model's inference and generalization capabilities for unseen nodes. For **Challenge 3**, we develop a knowledge distillation-based few-shot adaptation strategy to ensure that the model preserves previously learned knowledge while adapting to unseen nodes. This effectively mitigates catastrophic forgetting and achieves a balance between stability and plasticity.

Our contributions are summarized as follows:

❶ We study a more practical and challenging traffic expansion forecasting setting, namely few-shot and zero-shot Traffic Expansion Flow Forecasting (TEFF), where newly deployed sensors contain limited or even no historical observations.

❷ We propose SemProtoST, which combines MetaNR and DSTPP for few-shot and zero-shot Traffic Expansion Flow Forecasting. In SemProtoST, we design SCC and a knowledge distillation strategy to enhance generalization and alleviate catastrophic forgetting.

❸ Extensive experiments on four real-world datasets demonstrate that SemProtoST maintains competitive performance on seen sensors while improving forecasting performance on unseen sensors by an average of 3.58% and 14.22% under the few-shot and zero-shot settings, respectively.

2 Related Work

Traffic Flow Forecasting. Spatio-temporal graph neural networks (STGNNs) have become the dominant paradigm for traffic forecasting, including representative models such as STGCN(Yu et al., 2018), DCRNN(Li et al., 2018), and Graph WaveNet(Wu et al., 2019). These methods combine graph convolution networks(GCNs)(Kipf & Welling, 2017) with sequence modeling(Cho et al., 2014; Yu & Koltun, 2016) to effectively capture spatial dependencies and temporal dynamics in traffic data. Building upon these pioneering efforts, numerous improvements and extensions have been proposed in recent years(Bai et al., 2020; Li et al., 2023; Wu et al., 2020; Xia et al., 2023; Shang et al., 2021; Lan et al., 2022). Recent research highlights that modeling spatial heterogeneity is crucial for further enhancing predictive performance. These approaches integrate learnable mechanisms, such as spatial embeddings(Shao et al., 2022a), parameter pools(Liu et al., 2023a), and contextual pattern banks(Dong et al., 2024), to more precisely characterize complex spatial dynamics, thereby improving both forecasting accuracy and model adaptability.

Traffic Network Expansion Forecasting. Recent studies(Wang et al., 2019; Chen et al., 2021) have begun to consider traffic forecasting under evolving sensor networks. One line of work formulates network expansion as a continual learning problem and adapts forecasting models to newly added sensors using mechanisms such as historical replay (Chen et al., 2021; Lee & Park, 2024), prototype-based consolidation (Wang et al., 2023b;a), and dynamic pattern banks (Chen & Liang, 2025). Another line transfers knowledge from existing networks to target sensors through domain adaptation (Wang et al., 2022; Fang et al., 2022), knowledge distillation (Jin et al., 2023), shared pattern representations (Liu et al., 2023c), or prompt learning (Yuan et al., 2024). However, these approaches generally require newly deployed sensors to accumulate sufficient historical observations before effective adaptation can be performed. STEV (Ma et al., 2025) further studies few-shot traffic expansion forecasting and improves generalization through a flattening strategy and contrastive learning. Nevertheless, it still relies on limited observations from target sensors and does not consider the more challenging zero-shot traffic expansion forecasting scenario.

3 Preliminaries

Definition 1 (Traffic Expansion Flow Dataset). Let N_{seen} and N_{unseen} denote the number of existing and newly deployed sensors, respectively. Let $\mathcal{D}_1 \in \mathbb{R}^{N_{seen} \times P_1}$ denote the pre-expansion flow sequence dataset and $\mathcal{D}_2 \in \mathbb{R}^{(N_{seen}+N_{unseen}) \times P_2}$ denote the post-expansion flow sequence dataset, where typically $P_2 \ll P_1$. In the few-shot setting, only a small P_2 is available for adaptation, whereas in the zero-shot setting $P_2 = 0$.

Definition 2 (Sensor Metadata). Each sensor is associated with metadata $\mathbf{M}_i = [\mathbf{M}_i^{cat}, \mathbf{M}_i^{num}]$, including categorical and numerical features. Detailed metadata are provided in Appendix A.2.1. We denote the metadata of all sensors as $\mathbf{M} \in \mathbb{R}^{N \times C}$, where $N = (N_{seen}, N_{unseen})$.

Problem Statement (Traffic Expansion Flow Forecasting). Given historical traffic flow $\mathbf{X} \in \mathbb{R}^{N \times L}$ and sensors metadata $\mathbf{M} \in \mathbb{R}^{N \times C}$, Traffic Expansion Flow Forecasting (TEFF) aims to learn a function $\mathcal{F}$ that predicts future traffic flow for both seen and unseen sensors:

$$\hat{\mathbf{Y}}_{seen}, \hat{\mathbf{Y}}_{unseen} = \mathcal{F}\left(\mathbf{X}, \mathbf{M}\right), \tag{1}$$

where $\hat{\mathbf{Y}}_{seen} \in \mathbb{R}^{N_{seen} \times Q}$ and $\hat{\mathbf{Y}}_{unseen} \in \mathbb{R}^{N_{unseen} \times Q}$ denote the predictions for seen and unseen sensors, respectively. L and Q represent the historical and forecasting horizons, respectively.

4 Our Proposed Method SemProtoST

Fig. 3 presents the overall architecture of SemProtoST, which consists of three components. ❶ **Metadata-driven node representation** constructs node embeddings from static sensor attributes, enabling representations for unseen nodes without relying on fixed node identities. ❷ **Dual Spatio-Temporal Pattern Prototypes** maintain a set of spatio-temporal prototypes, while a semantic consistency constraint encourages semantically similar sensors to exhibit consistent prototype usage. ❷ **Knowledge-Distillation-based Few-shot Adaptation** strategy preserves performance on seen nodes while adapting to unseen nodes, achieving a balance between stability and plasticity.

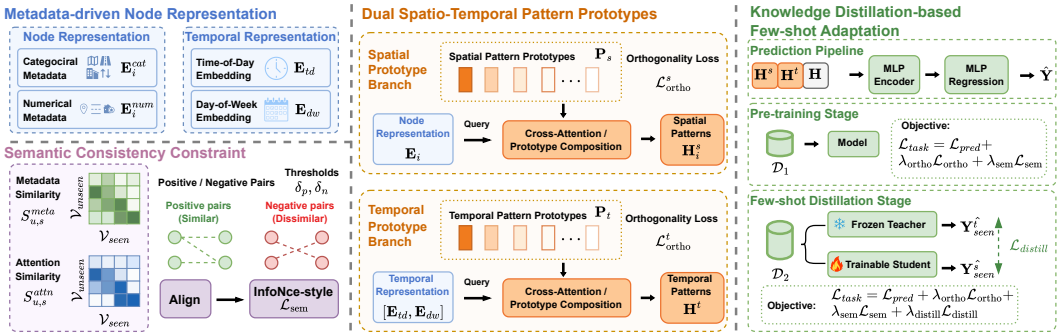


Figure 3: The framework of SemProtoST.

4.1 Metadata-driven Node Representation

Conventional node-specific embeddings are tied to sensors observed during training and cannot be directly constructed for newly deployed sensors. To enable inductive forecasting, we instead represent each sensor using its static metadata. For node i, let $\mathbf{M}_i = (\mathbf{M}_i^{cat}, \mathbf{M}_i^{num})$ denote its categorical and numerical metadata. The categorical features include district, county, freeway identifier, and traffic direction, while the numerical features consist of geographical coordinates, the number of lanes, and surrounding POI features.

For each categorical attribute k, we use a learnable embedding matrix $\mathrm{Emb}^{(k)} \in \mathbb{R}^{C_k \times d_k}$, where C_k is the number of category k, while normalized numerical attributes are projected through an MLP. The resulting node representation is

$$\begin{aligned} \mathbf{E}_i^{cat} = \texttt{Concat}_{k=1}^{K} \left[\mathrm{Emb}^{(k)} \left[\mathbf{M}_i^{cat_k} \right] \right], \quad \mathbf{E}_i^{num} = \texttt{MLP}\left(\texttt{Norm}(\mathbf{M}_i^{num}) \right), \\ \mathbf{E}_i = \texttt{Concat}\left[\mathbf{E}_i^{cat}, \mathbf{E}_i^{num} \right], \end{aligned} \tag{2}$$

where K denotes the number of categorical metadata fields with $C_k > 1$. With this design, the model is no longer dependent on a fixed set of nodes and can naturally generalize to unseen nodes, thereby adapting to the dynamic expansion of traffic networks.

4.2 Dual Spatio-Temporal Pattern Prototypes

Traffic dynamics often exhibit recurring patterns shared across sensors with similar spatial attributes or temporal contexts. Motivated by this observation, we propose a prototype-based spatio-temporal pattern modeling. Instead of learning node-specific patterns, the model captures a set of representative and reusable pattern prototypes. Through a cross-attention mechanism, node metadata and periodic temporal features are used as queries to dynamically compose spatio-temporal representations from spatial and temporal pattern prototypes.

Prototype-based Spatio-temporal Pattern Modeling. We maintain a spatial prototype bank $\mathbf{P}_s \in \mathbb{R}^{K_s \times d_s}$ and a temporal prototype bank $\mathbf{P}_t \in \mathbb{R}^{K_t \times d_t}$, where K_r and d_r denote the number and dimension of prototypes for branch $r \in \{s, t\}$. For an input representation $\mathbf{z}$, we define prototype composition as:

$$\mathbf{A}_r(\mathbf{z}) = \texttt{Softmax}\left(\frac{(\mathbf{z}W_q^r)(\mathbf{P}_r W_k^r)^\top}{\sqrt{d_r}} \right), \quad \Phi_r(\mathbf{z}) = \mathbf{A}_r(\mathbf{z})(\mathbf{P}_r W_v^r), r \in \{s, t\}, \tag{3}$$

where W_q^r, W_k^r, and W_v^r are learnable projections. For spatial pattern modeling, the metadata-driven node representation $\mathbf{E}_i$ queries the spatial prototype bank:

$$\mathbf{a}_i = \mathbf{A}_s(\mathbf{E}_i), \qquad \mathbf{H}_i^s = \Phi_s(\mathbf{E}_i). \tag{4}$$

For temporal pattern modeling, we obtain time-of-day and day-of-week embeddings, $\mathbf{E}_{td}$ and $\mathbf{E}_{dw}$, from the sample timestamp and use$[\mathbf{E}_{td}, \mathbf{E}_{dw}]$ to query the temporal prototypes:

$$\mathbf{H}^t = \Phi_t\left([\mathbf{E}_{td}, \mathbf{E}_{dw}] \right). \tag{5}$$

To encourage prototype diversity, we further impose a weak orthogonality regularization:

$$\mathcal{L}_{\text{ortho}} = \sum_{r \in \{s,t\}} \left\| \hat{\mathbf{P}}_r \hat{\mathbf{P}}_r^\top - I \right\|_F^2, \tag{6}$$

where $\hat{\mathbf{P}}_r$ denotes the row-wise normalized prototype bank.

Semantic Consistency Constraint. In the prototype-based modeling paradigm, unseen nodes lack historical observations, and their underlying spatio-temporal patterns need to be inferred solely from metadata. However, due to the inherent complexity of traffic systems, there often exists a significant gap between the static attribute space and the dynamic pattern space. When optimized solely with prediction loss, the model cannot guarantee semantically consistent attention mappings, which may result in nodes with similar physical attributes being assigned to substantially different prototype combinations. Such inconsistency further weakens the model's generalization capability to unseen nodes. To address this issue, we introduce a Semantic Consistency Constraint (SCC), aiming to require nodes with similar metadata should exhibit consistent attention distributions in the prototypes.

Since unseen nodes are not available during training, we adopt a virtual induction strategy. In each mini-batch, nodes are randomly partitioned into $\mathcal{V} = \mathcal{V}_{seen} \cup \mathcal{V}_{unseen}$, where $\mathcal{V}_{unseen}$ denotes pseudo unseen nodes, $|\mathcal{V}_{\text{seen}}| = \alpha N$, $|\mathcal{V}_{\text{unseen}}| = (1-\alpha)N$, α is a hyperparameter. For any selected unseen node $u \in \mathcal{V}_{unseen}$ and seen node $s \in \mathcal{V}_{seen}$, we calculate the similarity between their metadata and attention weights, respectively:

$$S_{u,s}^{meta} = \frac{\mathbf{M}_u \cdot \mathbf{M}_s^\top}{\|\mathbf{M}_u\| \, \|\mathbf{M}_s\|}, \quad S_{u,s}^{attn} = \frac{\mathbf{a}_u \cdot \mathbf{a}_s^\top}{\|\mathbf{a}_u\| \, \|\mathbf{a}_s\|}, \tag{7}$$

where $S_{u,s}^{meta}$ denotes metadata similarity, $S_{u,s}^{attn}$ denotes prototypes attention similarity. Based on the metadata similarity, we formulate a threshold-based contrastive learning task to regularize the latent space. We define the positive and negative sample pairs as follows:

$$\mathcal{P}(u) = \{s \mid S_{u,s}^{\text{meta}} > \delta_p\}, \quad \mathcal{N}(u) = \{s \mid S_{u,s}^{\text{meta}} < \delta_n\}, \tag{8}$$

where $\delta_p = 0.7$, $\delta_n = 0.3$. Positive sample pairs are considered semantically highly consistent and should exhibit analogous spatial patterns. Negative sample pairs are considered semantically distinct, and their inferred patterns should be pushed apart. The semantic consistency loss $\mathcal{L}_{\text{sem}}$ is defined using an InfoNCE-style:

$$\ell_{u,s} = \frac{S_{u,s}^{attn}}{\tau}, \quad \mathcal{L}_{\text{sem}} = -\frac{1}{|\mathcal{V}_{\text{unseen}}|} \sum_{u \in \mathcal{V}_{\text{unseen}}} \log \frac{\sum_{s \in \mathcal{P}(u)} \exp(\ell_{u,s})}{\sum_{j \in \mathcal{P}(u) \cup \mathcal{N}(u)} \exp(\ell_{u,j})}, \tag{9}$$

where τ is the temperature hyperparameter. By enforcing the model to minimize the discrepancy between the attention distributions of physically similar nodes, SCC successfully injects inductive bias (i.e., "similar causes lead to similar effects") into the representation learning process.

4.3 Knowledge Distillation-based Few-shot Adaptation

In traffic expansion scenarios, Few-shot adaptation to newly deployed sensors improves plasticity on unseen nodes but may degrade previously learned knowledge on seen nodes. To mitigate such catastrophic forgetting, we adopt a teacher–student distillation strategy that constrains the adapted model to preserve the predictions of the pre-trained model on seen sensors.

Pre-training Stage. We concatenate the traffic representation $\mathbf{H}$ with the spatial and temporal prototype representations $\mathbf{H}^s$ and $\mathbf{H}^t$, and feed them into a residual MLP followed by a linear prediction head. More detailed are provided in Appendix A.1.2 The model is pre-trained on the pre-expansion dataset $\mathcal{D}_1$ using:

$$\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{pred}} + \lambda_{\text{ortho}} \mathcal{L}_{\text{ortho}} + \lambda_{\text{sem}} \mathcal{L}_{\text{sem}}, \tag{10}$$

where $\mathcal{L}_{\text{pred}} = \left| \hat{\mathbf{Y}} - \mathbf{Y} \right|$, $\lambda_{\text{ortho}}, \lambda_{\text{sem}}$ are hyperparameters that control the contribution of prototype orthogonality and semantic consistency, respectively. The resulting parameters are denoted by θ_{pre}.

Few-shot Distillation Stage. For adaptation on the post-expansion data $\mathcal{D}_2$, both the student and teacher are initialized with θ_{pre}, while the teacher remains frozen. To preserve the forecasting be-

havior learned on existing sensors, we minimize the discrepancy between teacher and student predictions on the seen node set:

$$\mathcal{L}_{\text{distill}} = \left\| \hat{\mathbf{Y}}^{s}_{\text{seen}} - \hat{\mathbf{Y}}^{t}_{\text{seen}} \right\|^2, \tag{11}$$

The final joint objective function minimized by the student model during the few-shot distillation stage is defined as:

$$\mathcal{L} = \left| \hat{\mathbf{Y}}^{s} - \mathbf{Y} \right| + \lambda_{\text{ortho}} \mathcal{L}_{\text{ortho}} + \lambda_{\text{sem}} \mathcal{L}_{\text{sem}} + \lambda_{\text{distill}} \mathcal{L}_{\text{distill}}, \tag{12}$$

where λ_{distill} is the hyperparameter that control the contribution of knowledge distillation.

5 Experiments

5.1 Experimental Setup

Datasets. Following the LargeST (Liu et al., 2023b) benchmark, we evaluate SemProtoST on four large-scale traffic datasets with richer contextual information and larger sensor networks. Due to the lack of publicly available traffic expansion flow datasets, we customize these datasets to fit the TEFF task: **(i) ESD:** We select 95 sensors located on highway I5-N as the N_{unseen}, while the remaining 621 sensors are treated as the N_{seen}. **(ii) EGBA:** We select 103 sensors located on highway I880-N as the N_{unseen}, while the remaining 2249 sensors are treated as the N_{seen}. **(iii) EGLA:** We select 388 sensors located on highways I10-W and I15-S as the N_{unseen}, while the remaining 3446 sensors are treated as the N_{seen}. **(iv) ECA:** We select 863 sensors located on highways US101-N, I405-S, and I210-E as the N_{unseen}, while the remaining 7737 sensors are treated as the N_{seen}. Detailed dataset statistics are provided in Appendix A.2.1

Baselines and Metrics. We compare the proposed SemProtoST with 12 baselines. DCRNN(Li et al., 2018) and AGCRN(Bai et al., 2020) are RNN-based STGNNs. STGCN(Yu et al., 2018) and GWNet(Wu et al., 2019) are TCN-based STGNNs. D^{2}STGNN (Shao et al., 2022b) is a prior-graph-based method. STDN(Cao et al., 2025) and STAEformer(Liu et al., 2023a) are models based on the self-attention mechanism. STWave(Fang et al., 2023) and STID(Shao et al., 2022a) are embedding-based models. TrafficStream(Chen et al., 2021) and EAC(Chen & Liang, 2025) are continual learning-based models. STEV(Ma et al., 2025) is an expanding-variate model. Detail baseline introduction are provided in Appendix A.2.2 For conventional traffic forecasting models that are not designed for the TEFF task, we first standardize the input format by applying padding operations to align data shapes and then use a masking matrix to filter out padding loss. We use Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Detail implementation of our method are provided in Appendix A.2.3

5.2 Main Results

Few-shot. Table 1 reports the results under the few-shot setting. Overall, SemProtoST achieves the best or second-best performance across most datasets and metrics, with an average improvement of 3.58% on unseen sensors while maintaining competitive performance on seen sensors.

Compared with conventional traffic forecasting models, SemProtoST shows a clear advantage on unseen sensors. Although models such as AGCRN, D^{2}STGNN, and STDN achieve strong performance on seen sensors, their errors increase substantially on unseen sensors, revealing the difficulty of transferring node-specific representations learned from a fixed network. SemProtoST alleviates this limitation through MetaNR. For embedding-based methods, such as STID and STWave, competitive performance can still be observed on ESD and EGBA, but their generalization becomes less stable on the larger EGLA and ECA datasets. In contrast, DSTPP and SCC allow SemProtoST to model reusable spatio-temporal patterns shared across sensors and infer appropriate pattern combinations from sensor semantics, making the learned knowledge more transferable to unseen nodes.

Compared with traffic expansion methods, TrafficStream and EAC perform considerably worse when only limited post-expansion observations are available, while the limited scalability of STEV on larger datasets leads to OOM issues. SemProtoST instead learns transferable representations and pattern prototypes, reducing its dependence on newly accumulated observations. Meanwhile, its performance on seen sensors indicates that adaptation to unseen nodes does not severely disrupt

Table 1: Few-shot TEFF experimental results. **bold**: best, underlined: second best. The OOM indicates that the models incur **out-of-memory** issue even when we set batch size to 1.

Dataset		ESD			EGBA			EGLA			ECA		
Model	Metric	Unseen	Seen	Overall	Unseen	Seen	Overall	Unseen	Seen	Overall	Unseen	Seen	Overall
Conventional Traffic Forecasting Models													
DCRNN	MAE	39.03	20.42	22.93	35.21	20.47	21.12	30.03	24.14	24.64	OOM	OOM	OOM
	RMSE	59.18	32.52	37.25	50.43	32.43	33.44	47.10	37.33	38.32	OOM	OOM	OOM
AGCRN	MAE	47.20	17.70	21.68	36.29	19.43	20.18	37.13	20.32	22.02	40.06	21.25	23.16
	RMSE	70.07	28.07	36.66	52.37	31.83	33.02	61.68	33.22	37.15	58.66	**34.38**	37.57
STGCN	MAE	56.80	25.17	29.44	64.11	26.98	28.63	43.20	27.73	29.32	51.96	23.68	26.55
	RMSE	73.95	39.18	45.45	80.32	41.26	43.75	61.76	42.58	44.93	73.66	38.41	43.32
GWNET	MAE	51.26	20.04	24.25	36.99	21.80	22.48	30.69	23.03	23.81	33.23	22.95	23.95
	RMSE	81.40	31.56	41.90	54.75	34.34	35.50	51.88	36.16	38.07	51.47	37.08	38.97
D^2STGNN	MAE	44.21	17.67	21.25	39.33	18.41	19.34	36.40	20.02	21.70	OOM	OOM	OOM
	RMSE	70.39	28.30	36.97	57.09	30.73	32.36	59.88	31.72	35.65	OOM	OOM	OOM
STDN	MAE	49.95	17.01	21.54	47.64	19.93	21.16	31.87	19.74	21.12	36.83	22.53	23.97
	RMSE	73.66	27.30	37.10	63.64	32.31	36.44	52.03	32.60	35.21	58.64	37.31	39.67
STAEformer	MAE	39.85	17.17	20.22	33.63	18.96	19.62	29.95	19.82	20.86	38.64	22.56	24.19
	RMSE	64.18	27.59	34.85	49.23	31.43	32.43	48.96	32.46	34.52	58.83	40.96	43.11
STWave	MAE	39.22	**17.00**	20.01	37.22	19.41	20.21	30.42	20.89	21.87	35.11	22.45	23.73
	RMSE	61.98	**27.28**	34.09	54.82	31.74	33.11	51.72	33.13	35.49	57.68	37.23	39.79
STID	MAE	36.83	17.56	20.15	32.32	18.27	18.90	31.97	19.02	20.35	35.35	21.49	22.90
	RMSE	56.29	28.52	33.63	54.67	**29.37**	30.94	52.92	31.69	34.48	55.02	38.99	40.91
Traffic Expansion Models													
TrafficStream	MAE	49.64	45.71	46.24	41.84	38.60	38.75	44.04	50.86	50.16	55.61	49.15	49.80
	RMSE	71.13	68.40	68.82	59.76	59.20	59.25	67.62	76.95	76.04	85.16	76.79	77.68
EAC	MAE	47.93	31.46	33.68	50.82	32.35	33.18	50.36	34.95	36.53	42.70	31.61	32.73
	RMSE	69.84	48.68	33.68	69.24	47.66	48.83	73.80	51.55	54.26	64.22	48.18	50.04
STEV	MAE	64.00	25.32	30.54	75.29	44.86	46.22	OOM	OOM	OOM	OOM	OOM	OOM
	RMSE	88.80	42.02	50.90	126.5	77.62	80.44	OOM	OOM	OOM	OOM	OOM	OOM
SemProtoST	MAE	**35.88**	17.34	**19.89**	**31.97**	**18.23**	**18.84**	**27.57**	**19.01**	**19.89**	**31.49**	**20.94**	**22.01**
	RMSE	**54.52**	27.95	**32.81**	**47.45**	29.41	**30.45**	**44.44**	**30.52**	**32.23**	**49.73**	35.86	**37.51**

Table 2: Zero-shot forecasting performance on TEFF datasets. **bold**: best, underline: second best.

Dataset		ESD			EGBA			EGLA			ECA		
Model	Metric	Unseen	Seen	Overall	Unseen	Seen	Overall	Unseen	Seen	Overall	Unseen	Seen	Overall
GWNET	MAE	55.85	20.39	25.17	48.15	21.66	22.84	30.31	23.49	24.19	42.35	21.81	24.95
	RMSE	85.06	32.42	43.42	71.40	34.10	36.58	48.49	36.60	37.99	65.14	38.07	46.97
STAEformer	MAE	54.35	17.42	22.40	55.76	18.97	20.61	44.31	19.63	22.16	51.13	21.75	24.75
	RMSE	82.68	28.07	40.04	75.69	31.16	34.40	63.14	32.17	36.58	76.52	39.18	44.52
STWave	MAE	50.63	17.76	22.20	46.06	19.79	20.96	32.73	21.03	22.23	38.40	22.86	24.44
	RMSE	75.34	28.37	38.24	67.49	32.07	34.43	54.63	33.36	36.13	63.80	38.28	41.60
STID	MAE	47.88	17.11	21.26	52.46	18.21	19.74	45.37	18.96	21.67	54.55	21.27	24.65
	RMSE	70.90	27.83	36.72	69.70	30.35	33.12	66.30	31.53	36.65	81.86	38.35	44.74
TrafficStream	MAE	42.99	37.94	38.62	40.61	37.73	37.86	35.92	40.98	40.46	36.79	34.51	34.75
	RMSE	64.40	58.62	59.46	57.60	57.28	57.29	55.45	62.43	61.75	56.27	53.41	53.71
SemProtoST	MAE	**39.37**	**17.00**	**20.02**	**37.49**	**18.07**	**18.93**	**28.71**	**18.76**	**19.78**	**33.96**	**20.93**	**22.25**
	RMSE	**58.46**	**27.61**	**33.47**	**53.93**	**29.34**	**30.86**	**45.99**	**30.27**	**32.24**	**52.79**	**36.55**	**38.52**

previously learned forecasting capability, which is facilitated by the knowledge-distillation-based adaptation strategy that balances plasticity on unseen nodes and stability on seen nodes.

Zero-shot. We further evaluate the effectiveness of SemProtoST under the zero-shot setting, where newly deployed sensors have no historical observations during training. As shown in Table 2, SemProtoST consistently achieves the best performance on unseen sensors across all four datasets, with an average improvement of 14.22%, while maintaining competitive performance on seen sensors.

The performance gap becomes more pronounced in the zero-shot setting because conventional transductive models cannot adapt their node-specific representations using target-sensor observations. In contrast, MetaNR provides unseen sensors with ID-independent representations, while DSTPP and SCC exploit their metadata to infer transferable spatio-temporal patterns from knowledge learned on seen sensors. The stronger gains under zero-shot TEFF therefore further demonstrate the inductive generalization capability of SemProtoST.

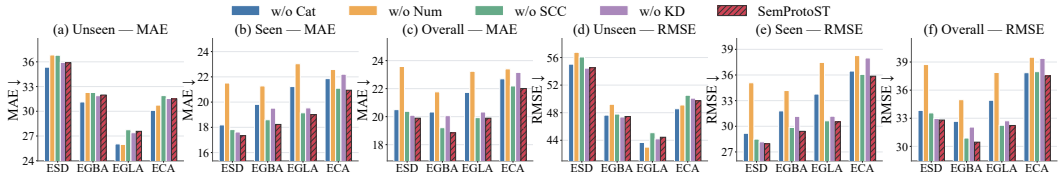


Figure 4: Results of ablation experiments.

5.3 Ablation Study

Module Ablation. We design several variants to evaluate the contribution of the key components in SemProtoST: (1) **w/o Cat**: removes the categorical features in the node metadata. (2) **w/o Num**: removes the numerical features in the node metadata. (3) **w/o SCC**: removes the semantic consistency loss $\mathcal{L}_{\text{sem}}$. (4) **w/o KD**: removes the knowledge distillation loss $\mathcal{L}_{\text{distill}}$.

As shown in Fig. 4, removing either categorical or numerical metadata generally degrades overall performance, with w/o Num causing a larger deterioration, indicating the importance of numerical attributes for characterizing traffic dynamics. Interestingly, removing part of the metadata occasionally improves performance on unseen nodes, while substantially degrading seen node accuracy. This suggests fine-grained spatial and road attributes provide strong discriminative cues for seen nodes, but may also encode distribution-specific information that is less transferable to unseen nodes. Removing SCC weakens unseen node forecasting, supporting its role in promoting transferable pattern inference across semantically related sensors. Furthermore, removing the knowledge distillation module (w/o KD) consistently causes noticeable degradation on seen nodes across all datasets, demonstrating that the teacher–student distillation strategy effectively alleviates catastrophic forgetting. Meanwhile, without the distillation constraint, the model can adapt more freely to the distribution of unseen nodes. However, this also makes the model more prone to forgetting previously learned knowledge from seen nodes, thereby degrading the overall performance.

Metadata Ablation. We further study the robustness of zero-shot TEFF to three metadata groups: **Geographic** (district, county, latitude, and longitude), **Road** (freeway, direction, and lanes), and **POI** (residential, workplace, commercial, and transport). We consider the removal and two perturbations on unseen sensors: **Removed**, where the metadata groups are removed. **Missingness**, where 50% of the corresponding attributes are replaced by statistics from seen sensors (mode for categorical and mean for numerical features), and **Noise**, where categorical attributes are randomly replaced and Gaussian noise ($\sigma = 0.3$) is added to numerical features.

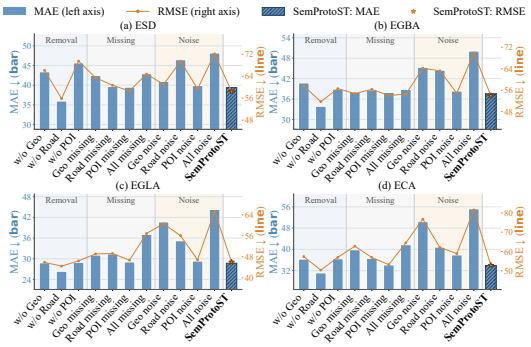


Figure 5: Results of metadata ablation.

Similar to the case of (**w/o Num**), excluding the Road Group slightly improves forecasting accuracy on unseen nodes but causes a noticeable degradation in performance on seen nodes. Fig. 5 shows that perturbing metadata at inference time consistently degrades performance. Although the complete model can tolerate moderate missingness in an individual metadata group, simultaneously missing all metadata groups results in substantial performance degradation, particularly on EGLA and ECA. The degradation is more pronounced under noisy metadata. These results indicate that SemProtoST relies on meaningful metadata for zero-shot inference, while retaining a certain degree of robustness when only part of the metadata is unavailable. In contrast, noisy metadata is generally more detrimental than missing metadata.

5.4 Different Traffic Expansion Scenarios Exploration

To evaluate the robustness of SemProtoST to different spatial configurations of newly deployed sensors, as shown in Fig 6, we construct three traffic expansion scenarios on the ESD dataset while keeping the number of unseen sensors fixed at 95: (1) Region-based Expansion, (2) Fwy-based

Table 3: Performance on Different Traffic Expansion Datasets.

Model		AGCRN			GWNET			STAEformer			STWave			STID			**SemProtoST**		
Dataset	Metric	Unseen	Seen	Overall	Unseen	Seen	Overall	Unseen	Seen	Overall	Unseen	Seen	Overall	Unseen	Seen	Overall	Unseen	Seen	Overall
Region ESD	MAE	32.99	19.81	21.50	33.52	22.46	23.88	**29.21**	17.90	19.73	29.49	19.04	20.35	36.70	18.05	20.43	30.83	**17.69**	**19.46**
	RMSE	47.65	31.74	34.19	55.24	37.18	39.95	**43.21**	29.96	31.84	47.02	31.79	33.95	54.67	28.88	33.31	45.61	**28.62**	**31.45**
Fwy ESD	MAE	47.20	17.70	21.68	51.26	20.04	24.25	39.85	17.17	20.22	39.22	**17.00**	20.01	36.83	17.56	20.15	**35.88**	17.34	**19.89**
	RMSE	70.07	28.07	36.66	81.40	31.56	41.90	64.18	27.59	34.85	61.98	**27.28**	34.09	56.29	28.52	33.63	**54.52**	27.95	**32.81**
Distance ESD	MAE	49.93	18.01	22.09	35.04	22.20	23.84	33.35	**17.67**	19.60	35.84	18.44	20.66	29.52	18.54	19.94	**27.73**	17.79	**19.12**
	RMSE	70.20	28.96	36.65	55.25	34.72	37.97	50.23	**28.56**	32.08	54.25	29.43	33.64	45.50	29.76	32.20	**41.59**	28.86	**30.78**

Expansion and (3) Distance-based Expansion. Detailed traffic expansion scenarios are provided in Appendix A.2.6.

Table 3 reports the corresponding forecasting results. Overall, SemProtoST achieves the best overall performance across the three scenarios, showing that its performance is not restricted to a single expansion pattern. The three settings exhibit different levels of difficulty. Region-based expansion generally yields lower unseen-node errors for most methods, whereas freeway- and distance-based expansion are more challenging. In the freeway-based setting, unseen sensors are concentrated on a distinct freeway corridor, introducing a stronger shift in road-specific traffic characteristics. In the distance-based setting, unseen sensors are spatially separated from the central area, increasing the spatial discrepancy between seen and unseen sensors. Across these settings, several baselines perform well in individual scenarios but do not maintain consistent performance across all three. In contrast, SemProtoST remains comparatively stable, suggesting that metadata-driven representations and shared prototypes provide robustness to different spatial expansion patterns.

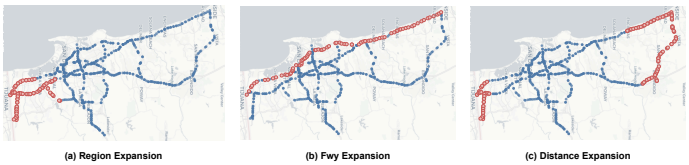


Figure 6: Visualization of different traffic expansion scenarios.

5.5 CASE STUDY

To further investigate the representation capability of SemProtoST, we visualize the learned node representations using t-SNE(Maaten & Hinton, 2008), as shown in Fig. 7. We select the San Diego (ESD) dataset as a case study. Specifically, we compare the metadata-driven node representations $\mathbf{E}_i$ and the prototype-enhanced representations $\mathbf{H}_i^s$ under different visualization settings.

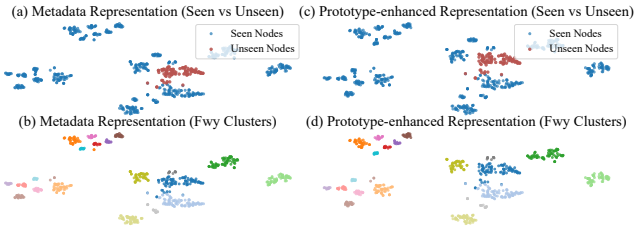


Figure 7: T-SNE visualization of node representations.

Fig. 7(a) and (c) show that unseen nodes are naturally distributed within the representation space of seen nodes rather than forming isolated clusters, indicating that MetaNR provides meaningful inductive representations for newly deployed sensors. Fig. 7(b) and (d) further show distinguishable clusters across freeway categories, suggesting that sensor metadata captures transferable traffic-related semantics. Additional visualizations and detailed analyses are provided in Appendix A.2.7

6 CONCLUSION

We propose SemProtoST for few-shot and zero-shot Traffic Expansion Flow Forecasting. By combining metadata-driven representations, shared spatio-temporal prototypes, semantic consistency regularization, and knowledge-distillation-based adaptation, SemProtoST enables inductive forecasting for newly deployed sensors while preserving performance on existing ones. Experiments on four large-scale traffic datasets demonstrate its effectiveness and generalization capability.

AI USE STATEMENT

In this work, we used generative AI tools (ChatGPT) solely for improving the grammar, clarity, and fluency of the manuscript. We have not used generative AI tools for research ideation, experimental design, data analysis, technical development, interpretation of results, or the generation of scientific claims, and other disclosure categories are not applicable to this work. The content, structure, technical contributions, experiments, analyses, and scientific conclusions were conceived, developed, and validated by the human authors. We have reviewed and edited all AI-assisted text to ensure its accuracy, consistency, and correctness. We take responsibility for the final content of this work, including any text produced with the aid of generative AI.

REFERENCES

Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. Adaptive graph convolutional recurrent network for traffic forecasting. In *NeurIPS*, 2020.

Lingxiao Cao, Bin Wang, Guiyuan Jiang, Yanwei Yu, and Junyu Dong. Spatiotemporal-aware trend-seasonality decomposition network for traffic flow forecasting. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *AAAI*, pp. 11463–11471, 2025.

Lu Chen, Qilu Zhong, Xiaokui Xiao, Yunjun Gao, Pengfei Jin, and Christian S. Jensen. Price-and-time-aware dynamic ridesharing. In *ICDE*, pp. 1061–1072, 2018.

Wei Chen and Yuxuan Liang. Expand and compress: Exploring tuning principles for continual spatio-temporal graph forecasting. In *ICLR*, 2025.

Xu Chen, Junshan Wang, and Kunqing Xie. Trafficstream: A streaming traffic flow forecasting framework based on graph neural networks and continual learning. In Zhi-Hua Zhou (ed.), *IJCAI*, pp. 3620–3626, 2021.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *EMNLP*, pp. 103–111, 2014.

Zheng Dong, Renhe Jiang, Haotian Gao, Hangchen Liu, Jinliang Deng, Qingsong Wen, and Xuan Song. Heterogeneity-informed meta-parameter learning for spatiotemporal time series forecasting. In *KDD*, pp. 631–641, 2024.

Yuchen Fang, Yanjun Qin, Haiyong Luo, Fang Zhao, Bingbing Xu, Liang Zeng, and Chenxing Wang. When spatio-temporal meet wavelets: Disentangled traffic forecasting via efficient spectral graph attention networks. In *ICDE*, pp. 517–529, 2023.

Ziquan Fang, Dongen Wu, Lu Pan, Lu Chen, and Yunjun Gao. When transfer learning meets cross-city urban flow prediction: Spatio-temporal adaptation matters. In Luc De Raedt (ed.), *IJCAI*, pp. 2030–2036, 2022.

Yilun Jin, Kai Chen, and Qiang Yang. Transferable graph structure learning for graph-based traffic forecasting across cities. In Ambuj K. Singh, Yizhou Sun, Leman Akoglu, Dimitrios Gunopulos, Xifeng Yan, Ravi Kumar, Fatma Ozcan, and Jieping Ye (eds.), *SIGKDD*, pp. 1032–1043, 2023.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*. OpenReview.net, 2017.

Shiyong Lan, Yitong Ma, Weikang Huang, Wenwu Wang, Hongyu Yang, and Pyang Li. DSTAGNN: dynamic spatial-temporal aware graph neural network for traffic flow forecasting. In *ICML*, volume 162, pp. 11906–11917, 2022.

Sanghyun Lee and Chanyoung Park. Continual traffic forecasting via mixture of experts. *CoRR*, abs/2406.03140, 2024. doi: 10.48550/ARXIV.2406.03140.

Fuxian Li, Jie Feng, Huan Yan, Guangyin Jin, Fan Yang, Funing Sun, Depeng Jin, and Yong Li. Dynamic graph convolutional recurrent network for traffic prediction: Benchmark and solution. *ACM Trans. Knowl. Discov. Data*, 17(1):9:1–9:21, 2023.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *ICLR*, 2018.

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 *CIKM*, pp. 4125–4129, 2023a.

Xu Liu, Yutong Xia, Yuxuan Liang, Junfeng Hu, Yiwei Wang, Lei Bai, Chao Huang, Zhenguang Liu, Bryan Hooi, and Roger Zimmermann. Largest: A benchmark dataset for large-scale traffic forecasting. In *NeurIPS*, 2023b.

Zhanyu Liu, Guanjie Zheng, and Yanwei Yu. Cross-city few-shot traffic forecasting via traffic pattern bank. In Ingo Frommholz, Frank Hopfgartner, Mark Lee, Michael Oakes, Mounia Lalmas, Min Zhang, and Rodrygo L. T. Santos (eds.), *CIKM*, pp. 1451–1460, 2023c.

Yisheng Lv, Yanjie Duan, Wenwen Kang, Zhengxi Li, and Fei-Yue Wang. Traffic flow prediction with big data: A deep learning approach. *IEEE Trans. Intell. Transp. Syst.*, 16(2):865–873, 2015.

Minbo Ma, Kai Tang, Huan Li, Fei Teng, Dalin Zhang, and Tianrui Li. Beyond fixed variables: Expanding-variate time series forecasting via flat scheme and spatio-temporal focal learning. In Luiza Antonie, Jian Pei, Xiaohui Yu, Flavio Chierichetti, Hady W. Lauw, Yizhou Sun, and Srinivasan Parthasarathy (eds.), *SIGKDD*, pp. 2054–2065, 2025.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Chao Shang, Jie Chen, and Jinbo Bi. Discrete graph structure learning for forecasting multiple time series. In *ICLR*, 2021.

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 *CIKM*, pp. 4454–4458, 2022a.

Zezhi Shao, Zhao Zhang, Wei Wei, Fei Wang, Yongjun Xu, Xin Cao, and Christian S. Jensen. Decoupled dynamic spatial-temporal graph neural network for traffic forecasting. *Proc. VLDB Endow.*, 15(11):2733–2746, 2022b.

Zezhi Shao, Fei Wang, Yongjun Xu, Wei Wei, Chengqing Yu, Zhao Zhang, Di Yao, Tao Sun, Guangyin Jin, Xin Cao, Gao Cong, Christian S. Jensen, and Xueqi Cheng. Exploring progress in multivariate time series forecasting: Comprehensive benchmarking and heterogeneity analysis. *IEEE Trans. Knowl. Data Eng.*, 37(1):291–305, 2025.

Binwu Wang, Yudong Zhang, Jiahao Shi, Pengkun Wang, Xu Wang, Lei Bai, and Yang Wang. Knowledge expansion and consolidation for continual traffic prediction with expanding graphs. *IEEE Trans. Intell. Transp. Syst.*, 24(7):7190–7201, 2023a.

Binwu Wang, Yudong Zhang, Xu Wang, Pengkun Wang, Zhengyang Zhou, Lei Bai, and Yang Wang. Pattern expansion and consolidation on evolving graphs for continual traffic prediction. In Ambuj K. Singh, Yizhou Sun, Leman Akoglu, Dimitrios Gunopulos, Xifeng Yan, Ravi Kumar, Fatma Ozcan, and Jieping Ye (eds.), *SIGKDD*, pp. 2223–2232, 2023b.

Jingyuan Wang, Jiawei Jiang, Wenjun Jiang, Chao Li, and Wayne Xin Zhao. Libcity: An open library for traffic prediction. In Xiaofeng Meng, Fusheng Wang, Chang-Tien Lu, Yan Huang, Shashi Shekhar, and Xing Xie (eds.), *SIGSPATIAL*, pp. 145–148, 2021.

Leye Wang, Xu Geng, Xiaojuan Ma, Feng Liu, and Qiang Yang. Cross-city transfer learning for deep spatio-temporal prediction. In Sarit Kraus (ed.), *IJCAI*, pp. 1893–1899, 2019.

Senzhang Wang, Hao Miao, Jiyue Li, and Jiannong Cao. Spatio-temporal knowledge transfer for urban crowd flow prediction via deep attentive adaptation networks. *IEEE Trans. Intell. Transp. Syst.*, 23(5):4695–4705, 2022.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. In *IJCAI*, 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 *KDD*, pp. 753–763, 2020.

Yutong Xia, Yuxuan Liang, Haomin Wen, Xu Liu, Kun Wang, Zhengyang Zhou, and Roger Zimmermann. Deciphering spatio-temporal graph forecasting: A causal lens and treatment. In *NeurIPS*, 2023.

Mingxing Xu, Wenrui Dai, Chunmiao Liu, Xing Gao, Weiyao Lin, Guo-Jun Qi, and Hongkai Xiong. Spatial-temporal transformer networks for traffic flow forecasting. In *arXiv:abs/2001.02908*, 2020.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *IJCAI*, pp. 3634–3640, 2018.

Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016.

Yuan Yuan, Jingtao Ding, Jie Feng, Depeng Jin, and Yong Li. Unist: A prompt-empowered universal model for urban spatio-temporal prediction. In Ricardo Baeza-Yates and Francesco Bonchi (eds.), *SIGKDD*, pp. 4095–4106, 2024.

A Appendix

A.1 Additional Method Details

A.1.1 Semantic Consistency Constraint

Semantic consistency constraint (SCC) computes metadata and prototype-attention similarities between pseudo-unseen and pseudo-seen nodes, resulting in a dense pairwise comparison whose training cost grows quadratically with the number of nodes. To improve scalability, we sample at most $A = 512$ pseudo-unseen nodes as anchors and $C = 256$ pseudo-seen nodes as a shared candidate pool. We retain at most $K_p = 8$ positive samples and $K_n = 32$ negative samples for the contrastive objective. Let $\widetilde{\mathcal{P}}(u) \subseteq \mathcal{P}(u)$ and$\widetilde{\mathcal{N}}(u) \subseteq \mathcal{N}(u)$ denote the sampled positive and negative sets, respectively. The sampled semantic consistency loss is defined as

$$\mathcal{L}_{\text{sem}}^{\text{samp}} = -\frac{1}{|\mathcal{V}_a|} \sum_{u \in \mathcal{V}_a} \log \frac{\sum_{s \in \widetilde{\mathcal{P}}(u)} \exp(S_{u,s}^{attn}/\tau)}{\sum_{j \in \widetilde{\mathcal{P}}(u) \cup \widetilde{\mathcal{N}}(u)} \exp(S_{u,j}^{attn}/\tau)}, \tag{13}$$

where $\mathcal{V}_a$ denotes the sampled anchor set. All similarity definitions and positive/negative selection criteria remain identical to those used in the original SCC.

The original SCC requires pairwise comparison between pseudo-unseen and pseudo-seen nodes, with a complexity of $\mathcal{O}(N_u N_s(d + K_s))$. In contrast, the sampled variant requires only

$$\mathcal{O}(AC(d + K_s)), \tag{14}$$

where $A \leq 512$ and $C \leq 256$ are fixed sampling budgets. Moreover, the final contrastive objective involves at most $A(K_p + K_n)$ pairs. Therefore, with a fixed candidate budget, the SCC computation grows linearly with the number of sampled anchors rather than requiring a dense all-pairs similarity matrix. Since the number of anchors is also capped, its per-iteration overhead remains bounded as the traffic network grows.

A.1.2 Prediction Head

Given an input traffic sequence $\mathbf{X}$, the traffic encoder first produces a hidden traffic representation $\mathbf{H}$. The spatial and temporal prototype modules further generate the spatial representation $\mathbf{H}^s$ and temporal representation $\mathbf{H}^t$, respectively. We concatenate these representations and employ a residual MLP to obtain the fused representation:

$$\mathbf{Z} = \left[\mathbf{H}, \mathbf{H}^s, \mathbf{H}^t\right], \tag{15}$$

Class	Attribute	Description
Geographic	**Lat**	The latitude of a sensor
	Lng	The longitude of a sensor
	District	The district of a sensor in PeMS
	County	The county of a sensor in California
Road	**Fwy**	The highway where a sensor is located
	Lane	The number of lanes where a sensor is located
	Type	The type of a sensor
	Direction	The direction of the highway
POI	**POI**	The POI information near a sensor

Table 4: comprehensive metadata of sensors.

Table 5: Statistical information of real-world datasets.

Dataset	N_{seen}	N_{unseen}	P_1	P_2	Time Range
ESD	621	95	5,136	48	01/2019 - 03/2019
EGBA	2249	103	5,193	48	04/2019 - 06/2019
EGLA	3446	388	5,251	48	07/2019 - 09/2019
ECA	7737	863	5,251	48	10/2019 - 12/2019

$$\mathbf{H}' = \mathbf{Z} + \texttt{MLP}(\mathbf{Z}). \tag{16}$$

The final Q-step traffic prediction is produced through a linear prediction head:

$$\hat{\mathbf{Y}} = \texttt{Linear}(\mathbf{H}'). \tag{17}$$

This prediction head is shared by both the pre-training and few-shot adaptation stages.

A.1.3 Zero-shot Inference

In the zero-shot setting, no observations from unseen sensors are used for model adaptation. Therefore, the few-shot distillation stage is skipped, and the model obtained after pre-training is directly used to forecast unseen sensors based on their metadata-driven representations and the shared spatio-temporal prototypes. Zero-shot denotes no post-expansion samples for model adaptation. Each test instance uses $L = 12$ real observations, temporal features and static metadata $\mathbf{M}$ for all nodes to predict the next $Q = 12$ steps. No padding, masking, or estimation is applied to the input window.

A.2 Additional Experiment Details

A.2.1 Dataset Details

We evaluate the proposed SemProtoST on four real-world traffic datasets. Due to the lack of publicly available traffic expansion flow datasets, we customize these datasets to fit the TEFF task. Fig. 8 illustrates the data splits for training, validation, and testing, while Table 5 shows each dataset's statistics and data splits. The details are as follows: **(i) ESD:** The SD dataset is collected from San Diego County, California, USA. We select 95 sensors located on highway I5-N as the N_{unseen}, while the remaining 621 sensors are treated as the N_{seen}. **(ii) EGBA:** The GBA dataset is collected from the Greater Bay Area in California, USA. We select 103 sensors located on highway I880-N as the N_{unseen}, while the remaining 2249 sensors are treated as the N_{seen}. **(iii) EGLA:** The GLA dataset is collected from the Greater Los Angeles area in California, USA. We select 388 sensors located on highways I10-W and I15-S as the N_{unseen}, while the remaining 3446 sensors are treated as the

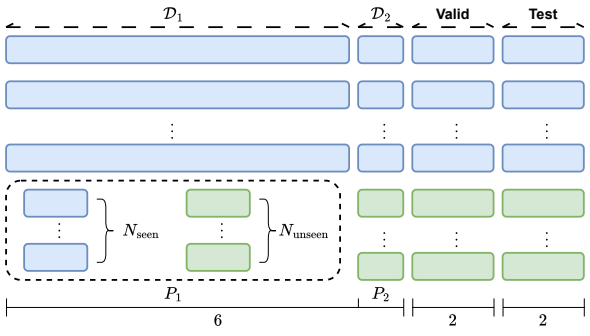


Figure 8: Illustration of data partition.

N_{seen}. **(iv) ECA:** The CA dataset covers traffic flow data across the state of California, USA. We select 863 sensors located on highways US101-N, I405-S, and I210-E as the N_{unseen}, while the remaining 7737 sensors are treated as the N_{seen}. The metadata is exemplified in Table 4. The POIs are obtained by OpenStreetMap.

Following BasicTS (Shao et al., 2025), we first applied z-score normalization to the raw data. We then set $N_{td} = 96$ and $N_{dw} = 7$. Each dataset was divided into training, validation, and test sets with a ratio of [0.6, 0.2, 0.2].

A.2.2 BASELINES DETAILS

Conventional traffic forecasting baselines:

- **DCRNN**(Li et al., 2018) replaces the fully connected layer within GRU with a variant of GCN to capture spatio-temporal correlations.
- **AGCRN**(Bai et al., 2020) merges RNNs with adaptive graphs to capture dynamic spatio-temporal patterns.
- **STGCN**(Yu et al., 2018) combines graph and temporal convolutions to model spatio-temporal relationships.
- **GWNET**(Wu et al., 2019) combines diffusion graph convolution and dilated convolution network for processing time .
- **D^{2}STGNN**(Shao et al., 2022b) proposes a decoupled framework to capture dynamic spatio-temporal correlations.
- **STDN**(Cao et al., 2025) enables the disentangled representation learning for traffic flow via a spatio-temporal-aware trend-seasonality decomposition.
- **STAEformer**(Liu et al., 2023a) introduces a spatio-temporal adaptive embedding to capture intrinsic spatial relationships and temporal patterns.
- **STWave**(Fang et al., 2023) utilizes the discrete wavelet transform to obtain the low and high-frequency components of traffic sequences.
- **STID**(Shao et al., 2022a) enhances predictions using spatial and temporal identity information with simple MLPs.

Traffic expansion baselines:

- **TrafficStream**(Chen et al., 2021) leverages spatio-temporal graph neural networks and continual learning to predict traffic flow in expanding and evolving networks.
- **EAC**(Chen & Liang, 2025) integrates a base STGNN with a continual prompt pool to addresses incremental learning and catastrophic forgetting in streaming data.
- **STEV**(Ma et al., 2025) studies few-shot traffic expansion forecasting and improves generalization through a flattening strategy and contrastive learning.

A.2.3 IMPLEMENTATION DETAILS

All experiments are conducted on a server equipped with an NVIDIA GeForce RTX 3090 GPU, an Intel Xeon Gold 5218R CPU (2.10 GHz), 251 GB RAM, and Ubuntu 22.04.5 LTS. SemProtoST uses input / output length 12, hidden dimension 128 on ESD and 196 otherwise, 5 MLP blocks, K_s=32 on ESD and 64 otherwise, and K_t=32. Pretraining uses Adam for 100 epochs (lr=0.002, weight decay=1e-4); orthogonality / SCC loss weights are 1000 / 100, and distillation weight is 0.3 on ESD and 0.1 otherwise.

A.2.4 ADDITIONAL METADATA ABLATION

To further investigate the role of sensor metadata in zero-shot TEFF, we conduct additional experiments on three groups of metadata attributes: (1) **Geographic** metadata, including district, county, latitude, and longitude; (2) **Road** metadata, including freeway identifier, traffic direction, and the number of lanes; and (3) **POI** metadata, including residential, workplace, commercial, and transport-related POI features.

We consider the removal, where the *w/o* variants remove one metadata group during both training and inference, and the model is retrained from scratch. These experiments evaluate the contribution of each metadata group to the learned forecasting model.

We consider two experiments: (1) **Metadata Missingness.** For each metadata group, we randomly select 50% of unseen sensors and replace the corresponding metadata fields with statistics computed from the seen sensors in $\mathcal{D}_1$. Categorical attributes are replaced by their modal values, while normalized numerical attributes are replaced by their feature-wise means. (2) **Metadata Noise.** To simulate noisy sensor metadata, we perturb the corresponding attributes of all unseen sensors. Categorical attributes are randomly replaced by valid categories observed from the seen sensors in $\mathcal{D}_1$, while Gaussian noise with standard deviation 0.3 is added to normalized numerical attributes.

A.2.5 Additional Spatial Transfer Baselines

To investigate whether zero-shot forecasting for unseen sensors can be trivially solved by transferring predictions from geographically nearby seen sensors, we introduce three additional forecast-transfer baselines: Nearest-Sensor, KNN-IDW ($K = 3$), and KNN-IDW ($K = 5$).

All three baselines follow the same zero-shot data split as SemProtoST. The forecasting model is trained only on seen sensors from $\mathcal{D}_1$, without any adaptation using $\mathcal{D}_2$. At inference time, the model first produces 12-step forecasts for seen sensors from their observed historical inputs. Predictions for an unseen sensor are then constructed solely by transferring the forecasts of geographically nearby seen sensors. Geographic proximity is measured using the Haversine distance computed from the original sensor coordinates. No freeway or direction restriction is imposed when selecting neighbors. Importantly, ground-truth traffic values within the forecasting horizon are used only for evaluation and are never involved in constructing the transferred predictions.

Nearest-Sensor. For each unseen sensor u, we identify its geographically closest seen sensor:

$$v^* = \arg \min_{v \in \mathcal{V}_{seen}} d(u, v), \tag{18}$$

where $d(u, v)$ denotes the Haversine distance between sensors u and v. The forecast of the nearest seen sensor is directly transferred to the unseen sensor:

$$\hat{\mathbf{Y}}_u = \hat{\mathbf{Y}}_{v^*}. \tag{19}$$

KNN-IDW. We further consider inverse-distance-weighted K-nearest-neighbor transfer with $K \in \{3, 5\}$. Let $\mathcal{N}_K(u)$ denote the K geographically closest seen sensors to unseen sensor u. The transferred prediction is

$$\hat{\mathbf{Y}}_u = \sum_{v \in \mathcal{N}_K(u)} \alpha_{uv} \hat{\mathbf{Y}}_v, \tag{20}$$

where the inverse-distance weights are

$$\alpha_{uv} = \frac{(d(u, v) + \epsilon)^{-1}}{\sum_{j \in \mathcal{N}_K(u)} (d(u, j) + \epsilon)^{-1}}. \tag{21}$$

We set $\epsilon = 10^{-6}$ km for numerical stability. The weights are normalized independently for each target sensor and shared across all 12 forecasting horizons.

Results. Table 6 shows that incorporating more neighboring sensors consistently improves the forecast-transfer baselines: KNN-IDW ($K = 5$) outperforms KNN-IDW ($K = 3$), which in turn generally outperforms direct nearest-sensor transfer. Nevertheless, all three spatial-transfer baselines remain substantially worse than SemProtoST on every dataset.

Compared with the strongest KNN-IDW ($K = 5$) baseline, SemProtoST reduces unseen-node MAE and RMSE by 46.7% and 44.3% on average, respectively. This large gap indicates that zero-shot forecasting cannot be explained by simply copying or spatially interpolating forecasts from nearby seen sensors. Instead, successful transfer requires modeling relationships between sensor characteristics and traffic dynamics beyond geographic proximity.

These results also provide additional evidence regarding the potential ambiguity of metadata-similar sensors. Geographic proximity alone is insufficient to infer unseen-node traffic patterns, and the

Table 6: Zero-shot forecasting performance of spatial forecast-transfer baselines on unseen sensors

Method	ESD		EGBA		EGLA		ECA	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Nearest-Sensor	83.66	124.44	65.17	90.80	64.33	97.53	60.95	92.00
KNN-IDW ($K = 3$)	78.84	114.60	65.14	89.72	62.70	94.24	59.04	88.10
KNN-IDW ($K = 5$)	77.27	111.96	64.34	88.88	62.31	93.39	58.55	87.19
SemProtoST	39.37	58.46	37.49	53.93	28.71	45.99	33.96	52.79

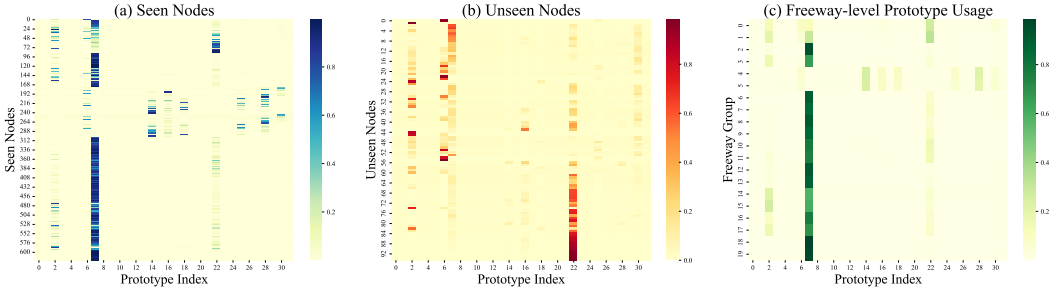


Figure 9: (a) Visualization of seen nodes prototype attention distribution. (b) Visualization of unseen nodes prototype attention distribution. (c) Visualization of freeway-level prototype attention distribution.

semantic prototype mechanism provides substantially stronger transfer than nearest-neighbor forecasting. However, we acknowledge that metadata may still be insufficient to distinguish sensors with nearly identical attributes but different functional or topological roles. Explicit functional topology is therefore complementary to our current formulation and represents a promising direction for future extensions.

A.2.6 DIFFERENT TRAFFIC EXPANSION SCENARIOS

To further evaluate the generalization capability of SemProtoST in dynamic traffic networks, we conduct experiments under different traffic expansion scenarios on the ESD dataset to investigate the impact of various expansion patterns on inductive forecasting performance. Considering that newly deployed sensors in real-world urban systems may follow different deployment strategies, we design three representative traffic expansion scenarios: (1) Region-based Expansion: This setting simulates sensor deployment in newly developed urban regions. As shown in Fig. 6(a), we select 95 sensors located in the southeast area of the map as the N_{unseen}. (2) Freeway-based Expansion: This setting simulates sensor expansion along a specific freeway. As shown in Fig. 6(b), we select 95 sensors located on the I5-N freeway as the N_{unseen}. (3) Distance-based Expansion: This setting simulates gradual urban expansion from the city center toward peripheral regions. As shown in Fig. 6(c), we use the city center as an anchor point and select the 95 sensors farthest from the city center as the N_{unseen}.

A.2.7 ADDITIONAL CASE STUDY

Spatial Prototype Attention. Fig. 9 visualizes the spatial prototype attention distributions learned by SemProtoST for both seen and unseen nodes. Specifically, Fig. 9(a) and Fig. 9(b) illustrate the prototype attention maps of seen and unseen nodes, respectively, while Fig. 9(c) presents the average prototype usage across different freeway. As shown in Fig. 9(a) and Fig. 9(b), both seen and unseen nodes exhibit highly concentrated attention distributions over a limited number of prototypes, rather than uniformly attending to all prototypes. In particular, several prototypes (e.g., Prototype 7 and Prototype 22) are repeatedly activated by a large number of nodes. This phenomenon indicates that SemProtoST successfully learns a set of transferable and reusable traffic pattern prototypes instead of relying solely on node-specific memorization.

Table 7: Time Complexity Comparison among Different Models

Model	Model Category	Time Complexity
AGCRN	GCN+RNN	$\mathcal{O}(BN^2Td)$
GWNET	GCN+TCN	$\mathcal{O}(BN^2Td)$
STAEformer	Attention	$\mathcal{O}(B(N^2+T^2)d)$
STNorm	Normalization	$\mathcal{O}(BNTd^2)$
STID	Embedding	$\mathcal{O}(BNTd^2)$
SemProtoST	Embedding	$\mathcal{O}\left(BN\left(T+d\right)d\right)$

Furthermore, unseen nodes exhibit prototype activation patterns highly consistent with those of seen nodes. Although unseen nodes contain extremely limited or even no historical observations, they are still able to query meaningful prototypes through metadata semantics. This shows that the proposed metadata-driven inductive mechanism effectively transfers traffic knowledge from seen nodes to unseen nodes.

As shown in Fig. 9(c), the freeway-level prototype visualization reveals that semantically related freeway groups tend to share similar prototype usage patterns. Multiple freeway groups consistently activate the same prototypes, suggesting that the learned prototypes capture shared traffic semantics across different road segments, such as highway traffic patterns, commuting dynamics, or regional congestion behaviors. This observation validates the effectiveness of the proposed semantic consistency constraint in aligning metadata semantics with latent traffic patterns.

In addition, several prototypes exhibit very low activation frequencies across all freeway groups. This suggests that urban traffic dynamics are dominated by a limited number of major shared traffic patterns, while only a few prototypes are required to model rare or specialized traffic behaviors.

Overall, these visualizations demonstrate that SemProtoST can learn transferable spatio-temporal representations and possesses robust inductive generalization capability for traffic expansion flow forecasting.

A.3 Time Complexity Analysis

We analyze the theoretical time complexity of SemProtoST and compare it with the other models.

Theoretical Complexity of SemProtoST. The computational overhead of SemProtoST comes from three components: ❶**Metadata-driven Node Representation.** Categorical metadata are transformed through embedding lookup operations, whose computational cost is linear. Numerical metadata are projected into the latent space via MLP. The total time complexity of this module is $\mathcal{O}(BNd)$. ❷**Dual Spatio-Temporal Pattern Prototypes.** For both spatial and temporal branches, the pattern modeling process consists of linear projections and cross-attention operations, with a complexity of $\mathcal{O}\left(B\left(Nd+Kd+NK\right)d\right)$. In addition, the orthogonality constraint computes prototype correlations, with a complexity of $\mathcal{O}(K^2d)$. The SCC computes similarity between pseudo unseen nodes and seen nodes, with a complexity of $\mathcal{O}\left(N^2\left(K+d\right)\right)$. Although SCC introduces quadratic complexity during training, it is only employed as an auxiliary term and is completely removed during inference. ❸**Pre-training and Few-shot Distillation.** The prediction involves only an MLP, with a complexity of $\mathcal{O}\left(BNTd\right)$. During pre-training, the overall complexity is dominated by the pattern modeling and prediction. During few-shot distillation, the teacher model only performs forward reasoning, doubling the computational cost. Since $K \ll N$, the overall time complexity of SemProtoST can be approximated as $\mathcal{O}\left(BN\left(T+d\right)d\right)$.

Comparative analysis with SOTA. Table 7 summarizes the time complexity of different models. Compared with GNN-based methods, SemProtoST avoids graph propagation over the adjacency matrix. As a result, its inference complexity scales approximately linearly with the number of nodes, making SemProtoST suitable for traffic expansion scenarios.