

Article

Causal–Semantic Spatiotemporal Traffic Flow Forecasting for Expressway UAV Pre-Deployment Using ETC Gantry Networks

Zeen Yang ^{1,†} , Zhuoer Wang ^{2,†}, Hongjuan Zhang ^{2,*}  and Bijun Li ² 

¹ School of Navigation, Wuhan University of Technology, Wuhan 430063, China; jw3194557044@whut.edu.cn

² State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430072, China; wangzhuoer@whu.edu.cn (Z.W.); lee@whu.edu.cn (B.L.)

* Correspondence: hongjuanzhang@whu.edu.cn

† These authors contributed equally to this work.

Abstract

Expressway unmanned aerial vehicle (UAV) pre-deployment is a geospatial decision-support task that requires reliable road-segment-level traffic flow prediction based on spatial sensing networks. However, existing spatiotemporal forecasting models remain limited in characterizing cross-segment propagation relationships, long-lag causal dependencies, and atypical traffic evolution patterns. In addition, complex models often fail to meet the computational requirements of edge-device deployment. Based on electronic toll collection (ETC) gantry data, this study proposes a causal–semantic spatiotemporal forecasting framework for long-term traffic flow prediction with a 24 h forecasting horizon. First, conditional Granger causality analysis is used to construct a directed causal prior graph that characterizes traffic propagation relationships among expressway segments. Second, scenario-semantic priors generated by a large language model are introduced to describe atypical traffic conditions. Then, causal structural priors and scenario-semantic priors are integrated into a teacher model and transferred to a lightweight student model through response-level and feature-level knowledge distillation. Experiments using expressway data from Hubei Province, China, show that the proposed model achieves the best overall performance in the typical scenario and competitive performance in the atypical scenario. The results indicate that the proposed framework can provide day-scale decision support for expressway law-enforcement UAV pre-deployment and enhance the spatial intelligence of traffic emergency management.

Keywords: spatiotemporal forecasting; expressway network; ETC gantry data; causal graph; scenario-semantic prior; large language model; knowledge distillation; UAV pre-deployment; edge deployment



Academic Editors: Wolfgang Kainz,
Min Hu and Vijayan Sugumaran

Received: 18 May 2026

Revised: 26 July 2026

Accepted: 4 August 2026

Published: 6 August 2026

Copyright: © 2026 by the authors.

Published by MDPI on behalf of the
International Society for

Photogrammetry and Remote Sensing.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Expressway unmanned aerial vehicle (UAV) pre-deployment requires road-segment-level traffic flow prediction based on spatial sensing networks, which is the key to improving the emergency response efficiency of intelligent transportation systems. In particular, emergency response demand increases substantially under atypical scenarios such as holiday travel surges, large-scale events, and unexpected incidents [1]. Once congestion or accidents occur, traditional ground-based rescue forces are constrained by road-network accessibility and response distance, which may lead to significant delays in emergency response. Compared with traditional methods, low-altitude UAVs used in intelligent transportation systems—thanks to their agility, rapid deployment, and wide field of view—can

not only reach the scene more quickly to patrol road sections and identify and monitor sudden traffic incidents, but also transmit real-time footage before ground rescue teams arrive, providing on-site guidance for vehicles on their next course of action [2]. Currently, some provinces in China have incorporated UAV systems into routine expressway management and emergency support frameworks and put them into practical operation. However, low-altitude UAV systems still face practical constraints, including limited flight range, short endurance, and restricted base-station coverage. When traffic incidents occur, if UAVs are not pre-deployed and rely solely on ad hoc dispatch, much of their limited endurance is consumed in transit, which shortens the effective operational window and lowers support efficiency.

Therefore, identifying road segments that are likely to experience high load and high risk within a future deployment cycle, and pre-deploying UAVs accordingly, has become a critical issue for improving the emergency support efficiency of UAVs on expressways. This creates an urgent need for road-segment-level long-term traffic flow prediction tailored to deployment planning. Compared with conventional traffic flow prediction tasks, road-segment-level long-term traffic flow prediction cannot be addressed by simply extending the single-step prediction horizon. Instead, it is a long-term multi-step forecasting task designed for the pre-deployment of law-enforcement UAVs. In this study, the model is updated once per day and directly predicts traffic flow for the following 24 h, and the data used must reflect the overall traffic state of the expressway network. Specifically, the model needs to predict traffic-load changes across road segments in the following 24 h deployment period based on historical observations, road-network propagation relationships, and atypical scenario information, thereby providing predictive support for downstream UAV deployment decisions. Complete multi-base-station scheduling, fleet allocation, endurance-constrained routing, and route optimization are outside the scope of this study. This requirement is consistent with previous traffic forecasting studies emphasizing that traffic prediction should account for spatial correlations, spatial heterogeneity, temporal fluctuations, and temporal regularity in road networks [3]. Achieving this goal further requires traffic flow observations that cover multiple road segments over a long time span. Since the removal of provincial toll stations in 2020, China has established an electronic toll collection (ETC) gantry system covering the entire expressway network. ETC gantries provide network-wide vehicle identification, stable all-weather data collection, and inherent path-topology associations, making them naturally suitable for long-term prediction tasks oriented toward deployment planning. In this context, the main road segment between a gantry and the adjacent gantry in the direction of travel is considered a single segment, and the traffic conditions—such as traffic volume and speed—recorded by that gantry are regarded as the traffic conditions for that segment. Moreover, ETC gantries share roadside expressway infrastructure with UAV base stations, eliminating the need for additional traffic-counting equipment. This physical co-location further supports the engineering feasibility of the proposed framework.

Existing traffic flow prediction research has evolved from early statistical modeling to deep learning and joint spatiotemporal modeling, achieving significant progress in traffic state prediction. Early classical statistical methods, such as autoregressive integrated moving average (ARIMA) and vector autoregression (VAR) [4], are constrained by linear assumptions and therefore struggle to capture the nonlinear dynamics of traffic systems. As research has advanced, deep time-series models have been widely used to learn complex traffic patterns [4,5]. Compared with traditional statistical methods, deep learning methods demonstrate stronger capabilities in feature extraction and nonlinear modeling, advancing traffic prediction from single-point estimation to multi-node joint modeling [5,6]. To further capture spatial dependencies in traffic systems, researchers

have introduced joint spatiotemporal modeling methods, especially graph-based traffic prediction models [6,7]. Recent studies have shown that graph neural networks are effective for modeling traffic forecasting problems over graph-structured road networks, where sensors or road segments are represented as nodes and traffic observations are treated as time-dependent node attributes [8,9]. In addition, directed graph convolution has been explored to represent directional traffic dependencies in road-network time series [10]. By representing road networks as graph structures, these methods partially improve the modeling of spatial correlations and enable traffic prediction to evolve from local state estimation to network-level joint prediction [7]. On this basis, research on multi-source information fusion and long-term prediction has also continued to advance, providing new insights for prediction in complex traffic scenarios [11,12]. However, for the specific scenario of UAV pre-deployment on expressways, existing traffic flow prediction studies still face several unresolved challenges.

In terms of spatial dependency modeling, this task requires a holistic view of the expressway network. However, most existing methods still rely mainly on fixed neighborhood relationships or static topological structures to describe spatial connections. Such approaches struggle to reflect the true direction of influence and time-lag characteristics between road segments and may misinterpret synchronous fluctuations or common upstream drivers as direct spatial dependencies. Although recent mixed-graph models combine graph convolution and attention mechanisms to capture both static and dynamic spatial correlations in traffic networks [13], their edge weights are still mainly learned from observed feature correlations rather than explicit causal relationships. To address this issue, recent studies have introduced causal analysis to reveal directional influences between nodes. Nevertheless, these methods mainly focus on short-range correlations between adjacent nodes in urban road networks, where time lags are typically measured in minutes [14,15]. In contrast, expressway traffic flow propagation across road segments often involves time lags of tens of minutes or longer and is strongly constrained by travel direction and physical distance [11,16]. The traffic-flow composition of expressways also differs substantially from that of urban road networks. This means that causal modeling methods designed for urban scenarios cannot be directly applied to expressway tasks. Therefore, designing a causal graph construction strategy tailored to the directionality and long-lag characteristics of expressways remains an important problem to be addressed.

The issue of knowledge constraints in complex traffic scenarios has attracted increasing attention. In real-world expressway systems, traffic flow evolution is influenced not only by historical temporal patterns but also by external factors such as holiday travel patterns, accident-induced disturbances, and the diffusion of severe congestion. As a result, models relying solely on historical observations often struggle to achieve accurate traffic flow prediction under atypical scenarios [12,17]. To address this limitation, researchers have begun to incorporate external knowledge and semantic priors into traffic prediction models. Large language models (LLMs) can help interpret the formation mechanisms of holiday travel peaks and the propagation effects of accident-induced disturbances on adjacent road segments [18]. The semantic knowledge embedded in LLMs can provide additional scenario understanding for time-series prediction [18–21]. However, effectively incorporating such knowledge into expressway traffic flow prediction models remains underexplored.

In addition to the above issues, UAV pre-deployment scenarios impose higher requirements on model computational efficiency. Unlike urban traffic prediction systems, which are typically deployed on cloud servers, the scenario considered in this paper requires the prediction model to operate independently on local devices at base stations along expressways. This ensures autonomous inference and day-scale rolling decision-making under unexpected conditions such as communication interruptions. However, directly incorporat-

ing complex teacher models or online LLM inference introduces substantial computational overhead, making it difficult to satisfy edge-deployment constraints on model size, inference latency, and computational resources. Therefore, constructing a lightweight prediction model suitable for edge deployment while preserving causal structural knowledge and scenario semantic knowledge is another critical challenge.

In summary, to address the practical needs of UAV pre-deployment, we propose a road-segment-level long-term traffic flow prediction framework that integrates a conditional Granger causal prior graph, LLM-generated scenario-semantic priors, and a multi-level knowledge distillation mechanism. The framework uses expressway ETC data as the basis for traffic flow prediction, thereby providing prediction support for proactive UAV deployment at base stations. First, conditional Granger causality tests are used to uncover directional and time-lagged influence relationships between road segments, based on which a directed causal prior graph is constructed. Then, for atypical scenarios such as holidays, traffic accidents, and severe congestion, scenario-specific semantic priors generated by LLMs were introduced to build expert knowledge and incorporate risk information into traffic flow predictions, thereby providing external knowledge constraints for the prediction models. Finally, through a multi-level knowledge distillation mechanism, the structured spatiotemporal knowledge learned by the teacher model and the scenario-semantic priors are jointly transferred to a lightweight student model. This enables the student model to perform high-accuracy prediction without explicitly relying on a complex teacher model or online LLM inference. Ultimately, the proposed method provides reliable road-segment-level traffic flow predictions to support the pre-deployment of UAVs at expressway base stations, thereby improving UAV allocation efficiency and emergency support capability.

The main contributions of this study are summarized as follows:

1. We propose an expressway cross-segment dependency modeling method based on a Granger causal prior graph. By mining causal influence relationships among road segments from multi-segment ETC historical time series, the proposed method characterizes the global propagation patterns of expressway traffic flow. In long-term full-network spatiotemporal traffic flow prediction, it effectively reduces the interference of spurious associations caused by synchronous fluctuations and common upstream drivers.
2. We construct an LLM-driven scenario-semantic prior generation mechanism. For representative scenarios such as holidays, traffic accidents, and severe congestion, semantic priors and expert knowledge are introduced to provide scenario constraints and knowledge guidance for the traffic flow prediction model, thereby enhancing its ability to identify atypical traffic evolution patterns.
3. We design a knowledge distillation framework for lightweight road-segment-level traffic flow prediction. The framework enables the teacher model to integrate causal structural knowledge, historical temporal knowledge, and scenario-semantic knowledge, and then transfers this knowledge to a lightweight student model. As a result, the model reduces parameter size and inference overhead while maintaining strong road-segment-level day-scale traffic flow prediction performance, thereby supporting day-scale rolling prediction and pre-deployment decisions for law-enforcement UAVs.

2. Related Work

2.1. Graph-Based Spatiotemporal Traffic Flow Prediction

Spatial dependencies in expressway traffic systems are a key factor affecting traffic flow prediction, and how to model them remains an important research issue. Graph neural networks, represented by the spectral convolution of graph convolutional networks (GCNs) [22] and the attention-based aggregation of graph attention networks (GATs) [23],

provide a general paradigm for road-network feature modeling. Within this paradigm, the construction of the adjacency matrix directly determines whether the model can accurately characterize dependencies among road segments. Accordingly, related studies have evolved from fixed graphs based on physical distance to data-driven adaptive graphs, and further to causal graphs based on causal analysis. Early graph construction methods defined fixed adjacency matrices based on road-network topology or sensor distance [24,25]. Although simple, these methods struggle to reflect dynamic changes in spatial dependencies. To alleviate this limitation, researchers have proposed data-driven adaptive graphs, which improve representational flexibility through learnable node embeddings or time-varying adjacency relationships [26–28]. However, the edge weights in these methods are still essentially based on feature correlations rather than causality. In expressway scenarios, fluctuations in upstream road segments may be misinterpreted as influences on unrelated downstream segments, thereby introducing spurious associations. To further characterize the true direction of influence and time-lag effects among road segments, causal analysis has been introduced into graph structure construction. Tank et al. [14] proposed Neural Granger Causality, which parameterizes causal relationships using neural networks and enables the modeling of nonlinear causal dependencies among time series. Runge et al. [29] proposed the PC-momentary conditional independence (PCMCI) method, providing an effective tool for causal discovery in high-dimensional time series. Building on these studies, researchers have begun to explicitly incorporate causal relationships into spatiotemporal prediction models. Yang et al. [15] proposed PSTCGCN, which extracts key road-network features through principal component analysis and captures causal chains of traffic events using a bidirectional causal learning module. Dong et al. [30] proposed CASTMGCN, a causality-aware multi-graph convolutional network that integrates physical road-network structure, time-lagged causal effects, and contemporaneous causal relationships into a multi-graph adjacency framework. Li et al. [31] designed a causal intervention encoder to identify and remove spurious spatial associations in adaptive graphs through causal inference. Yang et al. [32] proposed a causality-aware traffic prediction framework that independently estimates the causal effects of traffic accidents on future traffic states, thereby eliminating the interference of spatiotemporal confounders. Nevertheless, existing causal modeling studies mainly focus on urban road-network scenarios. Even though methods such as CASTMGCN explicitly introduce time-lagged causal effects, their lag scales are still designed primarily for urban settings. In contrast, expressway road segments are separated by longer physical distances, traffic flow exhibits stronger directionality, and cross-segment causal propagation often involves time lags of tens of minutes or longer. This pronounced scenario difference implies that causal modeling methods designed for urban road networks cannot be directly transferred to expressway tasks. Therefore, systematically constructing cross-segment and long-lag causal relationships in expressway networks remains an open problem.

2.2. Large Language Models for Traffic Flow Prediction

Large language models (LLMs), with their rich pretrained knowledge, contextual understanding, and semantic generalization capabilities, provide a new direction for introducing external knowledge into traffic flow prediction. In recent years, researchers have begun to explore how traffic time-series data and scenario information can be transformed into representations understandable by LLMs, and how LLMs can be used to enhance the representational capacity of conventional traffic prediction models. Guo et al. [18] proposed xTP-LLM, which converts multimodal traffic data into natural language descriptions and aligns an LLM with traffic flow prediction tasks through instruction tuning, thereby generating prediction results while providing interpretable analysis. Xu et al. [19]

proposed GPT4TFP, which maps traffic flow data into representations suitable for LLM processing through spatiotemporal embedding and spatiotemporal fusion layers, and then uses a pretrained LLM for traffic flow prediction, demonstrating the potential of LLMs in modeling spatiotemporal traffic patterns. Beyond directly using LLMs for traffic prediction, some studies have further focused on integrating LLMs with event semantics. Li et al. [31] designed Chain-of-Actions prompting in STCInterLLM to guide the LLM in decomposing the spatiotemporal prediction process, thereby enhancing causal representations and mitigating LLM hallucinations. Yang et al. [32] used low-rank adaptation (LoRA) to fine-tune a small-scale LLM in LLeCaT, extracting semantic embeddings from accident text records and injecting them into the prediction model. Jiang et al. [33] proposed the KIDCM framework, which uses LLM-generated traffic data to model general spatial dynamic relationships and distills LLM capabilities to achieve a balance between generalization ability and computational efficiency. These studies indicate that LLMs hold considerable potential for traffic flow prediction in complex scenarios. However, most existing studies focus on the direct prediction capability of LLMs or the injection of contextual information, while relatively little attention has been paid to incorporating LLM-generated scenario semantics into prediction models as structured constraints. Therefore, using LLMs to construct scenario-semantic priors for atypical scenarios—such as holidays, traffic accidents, and severe congestion—to improve model robustness remains an important open issue.

2.3. Knowledge Distillation for Traffic Flow Prediction

Knowledge distillation can transfer learned knowledge to lightweight networks, improving computational efficiency and supporting practical deployment. The concept of knowledge distillation was first proposed by Hinton et al. [34], who transferred knowledge from a teacher model to a student model by enabling the student to fit the soft-label distribution output by the teacher. Subsequent studies further extended distillation targets from the output layer to intermediate features and attention responses. Romero et al. [35] proposed FitNets, which introduce intermediate hint supervision so that the student model learns not only the final output of the teacher model but also its hidden representations, thereby improving the representation learning capability of lightweight models. In the field of traffic flow prediction, models such as graph neural networks and Transformers can effectively capture spatiotemporal patterns, but their high computational complexity hinders deployment on edge devices. Existing studies have begun to introduce knowledge distillation into spatiotemporal prediction tasks. Zhang et al. [36] proposed LightST, a spatiotemporal knowledge distillation framework for traffic flow prediction. It transfers spatial and temporal knowledge from a high-capacity teacher model to a lightweight student model, improving the scalability and computational efficiency of traffic prediction models in practical deployment. Jiang et al. [33] proposed KIDCM, which distills the prediction capability of LLMs into a lightweight GCN-GRU surrogate model, achieving a balance between generalization ability and computational efficiency. In addition, because multi-level distillation often involves multiple distillation losses, Liu et al. [37] proposed a dynamic weight averaging (DWA) mechanism, which adaptively adjusts loss weights according to the rate of change of each task loss and provides a training strategy for multi-objective joint optimization. However, existing studies have rarely explored the collaborative distillation of LLM-generated scenario-semantic priors and causal structural knowledge in expressway scenarios. Therefore, how to transfer complex cross-segment causal propagation patterns and scenario priors to lightweight models suitable for edge deployment through knowledge distillation still requires further investigation.

3. Problem Definition

To provide predictive support for expressway UAV pre-deployment, road-segment-level long-term traffic flow prediction for the pre-deployment of law-enforcement UAVs is formulated as a multi-road-segment time-series prediction problem.

Definition 1. *Road-segment traffic flow time series: The historical ETC observation sequence is denoted as $X \in \mathbb{R}^{N \times T \times F}$ where N is the number of ETC gantries deployed in the expressway network as observation nodes. T is the length of the historical time window, which is set to 48 time steps, corresponding to 24 h at the 30 min resolution, and F denotes the dimensionality of traffic features recorded at each node, such as traffic volume and speed.*

Definition 2. *Causal prior graph: The directed causal adjacency matrix among road segments is denoted as $A \in \mathbb{R}^{N \times N}$.*

Definition 3. *Scenario-semantic embedding: The scenario-semantic embeddings for holidays, traffic accidents, severe congestion, and other scenarios are denoted as $Z_{sem} \in \mathbb{R}^{M \times d}$, where M is the number of semantic tokens and d is the unified hidden dimension of the model.*

Considering the need for a lightweight framework suitable for edge deployment, our objective is to introduce a teacher model f_T during the training stage. The teacher model takes $\{X, A, Z_{sem}\}$ as joint inputs. Then, through knowledge distillation, the spatiotemporal causal structures and scenario-semantic prior knowledge extracted by f_T are transferred to the student model f_S . Finally, a lightweight mapping function f_S suitable for edge deployment is obtained:

$$\hat{Y} = f_S(X) \quad (1)$$

where $\hat{Y} \in \mathbb{R}^{N \times T_{pred}}$ denotes the prediction output.

4. Methodology

4.1. Framework Overview

As shown in Figure 1, we construct a road-segment-level long-term traffic flow prediction framework with a 24 h forecasting horizon that supports UAV pre-deployment while meeting lightweight edge-deployment requirements. The framework consists of three components. First, the teacher model takes ETC time series, the causal adjacency matrix, and LLM-generated scenario semantics as inputs. It extracts structural, temporal, and scenario features through spatial, temporal, and semantic branches, respectively, forming transferable comprehensive knowledge. The teacher prediction and high-level fused representation are then used as distillation supervision signals. Second, the knowledge distillation model uses the teacher prediction and high-level fused representation as distillation targets, enabling the structured spatiotemporal knowledge and scenario-semantic priors learned by the teacher model to be jointly transferred to the student model. Third, the student model consists of several simple neural network modules. The student substantially reduces parameters and inference latency while still capturing long-term traffic evolution patterns. The student model outputs road-segment-level day-scale predictions to support UAV pre-deployment decisions.

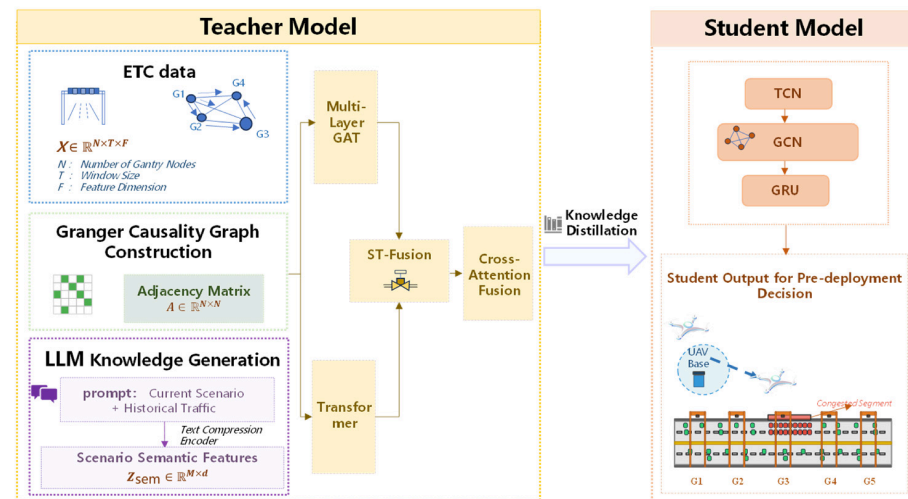


Figure 1. Overall architecture of the proposed method. Arrows indicate the flow of data and knowledge between modules; colors distinguish functional components. Abbreviations: ETC, electronic toll collection; LLM, large language model; GAT, graph attention network; GCN, graph convolutional network; GRU, gated recurrent unit; TCN, temporal convolutional network; UAV, unmanned aerial vehicle; ST, spatiotemporal.

4.2. Teacher Model

To collect and integrate important information from road segments for student learning, the teacher model must support full-input, multi-node, and joint spatiotemporal modeling. To this end, the teacher model takes three types of inputs: historical ETC time series, the conditional Granger causal adjacency matrix A , and the LLM-generated scenario semantics Z^{sem} . The constructed A enhances the teacher model's structural awareness along causal directions, while the generated Z^{sem} improves its scenario generalization ability under atypical conditions. During inference, the distilled student model uses the ETC time series and the precomputed fixed causal graph, which is stored locally and used by the GCN. Neither the teacher model nor online LLM inference is required.

The overall workflow is shown in Figure 2. This section introduces the teacher model from three aspects: causal dependency graph construction, LLM-based scenario-semantic embedding generation, and the causality- and semantics-enhanced spatiotemporal encoder.

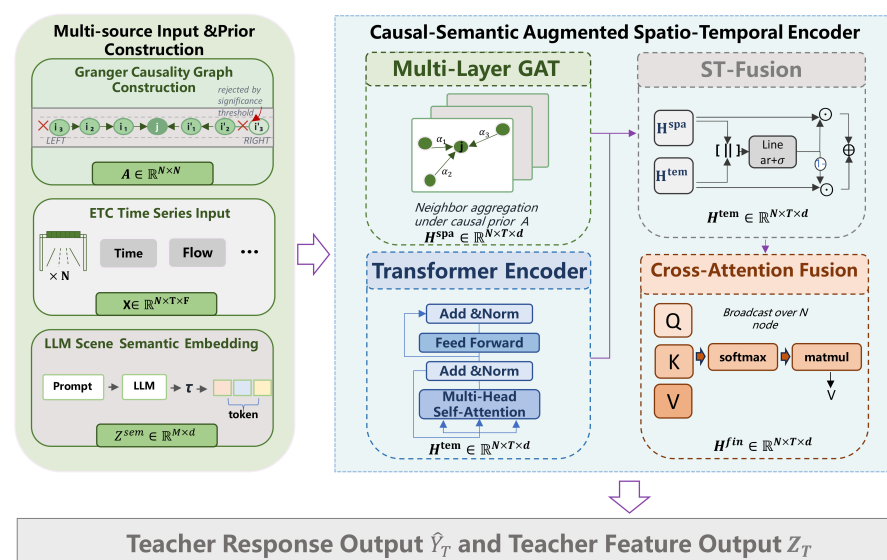


Figure 2. Architecture of the teacher model. Arrows indicate information flow, and colors distinguish the input/prior-construction, spatiotemporal-encoding, and fusion modules.

4.2.1. Conditional Granger-Based Causal Prior Graph Construction

The target task is UAV pre-deployment in expressway scenarios. In this setting, the influence relationships among road segments exhibit strong directionality and long-lag characteristics. A fixed adjacency matrix defined solely by physical distance, or a data-driven adaptive adjacency matrix, cannot adequately characterize such global propagation patterns. To address this limitation, this study constructs a directed causal prior graph using conditional Granger causality tests. The adjacency matrix $A \in \mathbb{R}^{N \times N}$ represents the directed influence strength between nodes. Considering the large scale of expressway networks and the distance-decay effect of cross-segment influence along the driving direction, A is constructed using a near-to-far iterative expansion strategy.

Before the causal test, each node-level traffic-flow series is standardized using the mean and standard deviation of the training set, and its stationarity is examined by the Augmented Dickey–Fuller test at the 0.05 significance level. For any series that fails to satisfy the stationarity requirement, first-order differencing is applied before the causal test, whereas the original standardized series is retained as the input of the prediction model to preserve the original traffic-flow scale and temporal patterns. Such preprocessing ensures that the series entering each conditional Granger test is stationary, which is required for standard Granger-causality inference.

Specifically, each node is treated as a target node j , and candidate source nodes i are sequentially incorporated from near to far along both upstream and downstream directions. Let $y_i(t)$ and $y_j(t)$ denote the target traffic flow sequences of node i and node j , respectively, and let p be the maximum lag order. Because the ETC data are sampled at a 30 min interval, the maximum lag order p is set to 12, so that the causal-propagation window reaches up to six hours and is able to accommodate both short-term traffic disturbances and the long-lag propagation effects that are characteristic of expressway segments. For each node pair, the specific lag order within the range from 1 to p is further determined by the Bayesian information criterion, which avoids the spurious dependencies that an excessively long lag may introduce. The currently accepted significant candidate set is used as the conditioning set C_j . If the lagged terms of node i still significantly improve the prediction of node j after controlling for C_j , then node i is considered to have a conditional Granger causal influence on node j . Accordingly, the causal prior adjacency matrix is defined as follows:

$$A_{i \rightarrow j} = \begin{cases} \phi(F_{i \rightarrow j|C_j}), & q_{i \rightarrow j|C_j} < \alpha, \\ 0, & \text{otherwise,} \end{cases} \quad (2)$$

where $F_{i \rightarrow j|C_j}$ denotes the corresponding conditional Granger test statistic, $\phi(\cdot)$ is a normalization function, $q_{i \rightarrow j|C_j}$ is the significance test result, and α is the significance threshold. The significance threshold α is set to 0.05. Considering that a large number of node-pair tests are performed over the whole expressway network, the significance test result $q_{i \rightarrow j|C_j}$ is taken as the value adjusted by the Benjamini–Hochberg procedure, which controls the false discovery rate across all tests. Only the edges that remain significant after this correction are retained in the adjacency matrix, so that spurious causal edges induced by multiple comparisons are effectively suppressed.

The resulting A preserves driving direction and time-lag decay, while the conditioning set C_j removes spurious associations caused by common upstream factors and synchronous fluctuations. To keep the discovered causal graph consistent with the physical topology of the expressway, candidate edges are further restricted to node pairs that are physically reachable along the driving direction, and any edge whose implied lag is incompatible with the corresponding travel distance is pruned. Consequently, A is not a purely statistical construct but a directed causal graph constrained by the physical road network. In the

subsequent spatial branch, A is incorporated into a graph attention network. Neighborhood aggregation is performed along the edges specified by A , which filters out interference from irrelevant road segments and enables the model to focus on upstream and downstream segments with genuine causal relationships. Compared with existing causal-graph-based traffic prediction methods such as CASTMGCN [30], which calibrate time-lagged causal graphs primarily for short-delay urban road networks, the lag order, significance correction, and topology constraints adopted above are instead configured for the long-lag, strongly directional propagation characteristic of expressways. This design is therefore not a direct application of urban causal-graph methods to a new dataset, but a scenario-specific reconstruction of the causal discovery procedure for expressway-scale propagation. In this way, the constructed graph characterizes the global propagation patterns and latent causal dependencies of expressway traffic flow, providing accurate spatial guidance for UAV pre-deployment.

4.2.2. LLM-Based Scenario-Semantic Encoding

Expressway traffic flow is often affected by external factors such as holidays, traffic accidents, and severe congestion. The causal graph A and ETC data X alone cannot characterize these scenario patterns. Therefore, this section introduces an LLM to provide semantic priors, thereby improving the robustness of the model under atypical scenarios. The offline LLM is instantiated using Qwen2.5-7B-Instruct. Specifically, before each daily 24 h forecasting cycle, the system constructs a current scenario description s and a recent historical traffic flow summary $hist$ based on next-day calendar information, recent atypical event records in the road network, and historical traffic summaries. The calendar information is obtained from the public statutory-holiday calendar, the atypical event records are retrieved from the event log of the expressway management platform, and the historical traffic summaries are computed directly from the ETC observations.

A scenario prompt is then generated using a predefined template and fed into a pretrained LLM to produce semantic text τ , which describes traffic evolution trends and potential risks. The prompt follows a fixed template that supplies the date and holiday flag, the recent event records, and the historical traffic summary as structured fields, and instructs the model to output the trend, peak period, risk segments, and congestion level for the next prediction cycle using only the provided facts. Under a typical scenario, the prompt fields indicate an ordinary weekday with no holiday and no recorded incident, and the LLM outputs a normal-traffic prior with a low risk level. Under an atypical scenario, the prompt fields indicate a public holiday together with a recorded accident on a specific segment, and the LLM outputs a congestion-related scenario type with an elevated risk level and the affected segment identified. The LLM itself does not participate in end-to-end training; instead, it is invoked offline as an external knowledge source. This decouples the inference overhead of the LLM from each training step. To avoid unstable or incorrect priors, the LLM is queried with greedy decoding to keep the output deterministic, the response is required to follow a fixed structured schema; otherwise, it is rejected and replaced by a neutral prior or regenerated using a schema-repair prompt. The generated τ remains fixed during the corresponding 24 h forecasting cycle and is cached for reuse, and is regenerated once per day, consistent with the daily forecasting schedule described in Section 4.2.3. Unlike methods that compress text into a single global vector, this study preserves the token-level semantic representations output by the text encoder, enabling fine-grained interaction with spatiotemporal features through cross-attention. Let the output of the text encoder be:

$$Z^{\text{sem}} = \text{Proj}(\text{Enc}_{\text{tok}}(\tau)) \in \mathbb{R}^{M \times d} \quad (3)$$

where M denotes the number of semantic tokens, d denotes the unified hidden dimension of the teacher model, $\text{Enc}_{\text{tok}}(\cdot)$ denotes the token-level text encoder, which is instantiated as a frozen BERT-base model with a maximum length of 128 tokens and a hidden size of 768, and $\text{Proj}(\cdot)$ denotes the linear projection layer for dimensional alignment. The text encoder is kept frozen during training to prevent the semantic representation from overfitting to the limited atypical-scenario samples. Unlike LLeCaT [32], which fine-tunes a small-scale LLM via LoRA, or STCInterLLM [31], which incorporates Chain-of-Actions prompting into the LLM reasoning process, the LLM in this study serves only as an offline structured scenario-prior generator. It is neither fine-tuned nor involved in sample-level spatiotemporal prediction; instead, it is invoked once before each daily 24 h forecasting cycle to generate a cached structured prior, which is subsequently injected through the pathway described above.

The generated Z^{sem} describes network-level evolutionary contexts, such as overall holiday patterns or large-scale congestion diffusion. Injecting this information into the model enables UAVs to be pre-deployed toward potentially risky road segments under complex scenarios. However, Z^{sem} differs from the node-level structural and temporal dependencies represented by A and X . If it were directly incorporated into GAT neighborhood aggregation or Transformer temporal self-attention, the same global context would be forcibly replicated at every node and time position during low-level aggregation. This would dilute its semantic properties and interfere with the independent modeling of spatial structure and temporal dynamics. Therefore, this study introduces Z^{sem} through cross-attention. Z^{sem} is broadcast and shared along the node dimension, while inter-node differences are independently handled by the spatial branch constrained by A . In this design, A acts on the low-level information propagation pathway, whereas Z^{sem} acts on high-level feature selection. The two types of priors complement each other at different levels and jointly provide prior constraints for the subsequent spatiotemporal encoder.

4.2.3. Causality- and Semantics-Enhanced Spatiotemporal Encoder

To integrate the causal prior graph A , the scenario-semantic embedding Z^{sem} , and the raw traffic time series X , this section designs a joint spatiotemporal encoder with spatial, temporal, and semantic branches. This design enables comprehensive modeling of structural, temporal, and scenario features.

First, the raw input is linearly projected and augmented with temporal positional encoding, mapping it into a unified hidden space:

$$\tilde{\mathcal{X}} = \mathcal{X}W_{\text{in}} + P^{\text{pos}}, \tilde{\mathcal{X}} \in \mathbb{R}^{N \times T \times d} \quad (4)$$

where $W_{\text{in}} \in \mathbb{R}^{F \times d}$ and $P^{\text{pos}} \in \mathbb{R}^{1 \times T \times d}$. The projected representation $\tilde{\mathcal{X}}$ maps traffic volume, speed, and other features into a unified d -dimensional hidden space, while assigning distinguishable positional information to each time step. It serves as the shared input for subsequent spatial and temporal modeling.

In the spatial branch, the edge weights in A reflect directed causal strengths under conditional Granger testing. Since A is a weighted, sparse, and explicitly directional topology, this study adopts GAT as the spatial aggregation operator. Compared with GCN, which directly uses A as a fixed convolution kernel, GAT does not treat A as constant aggregation weights. Instead, within the neighborhood specified by A , it dynamically reassesses the contribution of each neighbor according to node features at each time step, thereby characterizing the actual influence strength at each moment:

$$H^{\text{spa}} = f_{\text{gat}}(\tilde{\mathcal{X}}, A), H^{\text{spa}} \in \mathbb{R}^{N \times T \times d} \quad (5)$$

In the temporal branch, the traffic flow evolution of each expressway node spans multiple temporal scales. When recurrent structures are used, information propagates step by step along the temporal axis, and signals from more distant time steps are diluted more easily. However, this task requires the model to simultaneously capture nearby local changes and long-range trends. Therefore, this study applies a Transformer encoder independently along the temporal dimension for each node, allowing any two time steps to establish direct dependencies through self-attention:

$$H^{\text{tem}} = f_{\text{trm}}(\tilde{\mathcal{X}}), H^{\text{tem}} \in \mathbb{R}^{N \times T \times d} \quad (6)$$

However, H^{tem} and H^{spa} are not always equally important. For upstream and downstream nodes affected by atypical events, information propagated through the spatial path is more critical. In contrast, during relatively stable periods, a node's own historical pattern may be more informative. Direct concatenation or equal-weight summation would obscure this difference. Therefore, this study adopts a gated fusion mechanism to integrate spatial and temporal representations while avoiding the adverse effects of direct concatenation:

$$g = \sigma([H^{\text{spa}} \parallel H^{\text{tem}}]W_g + b_g) \quad (7)$$

$$H^{\text{st}} = g \odot H^{\text{spa}} + (1 - g) \odot H^{\text{tem}} \quad (8)$$

The fused representation H^{st} simultaneously carries cross-segment dependencies under the causal structure and the temporal patterns of each node.

Next, the generated scenario-semantic prior Z^{sem} needs to be incorporated into the spatiotemporal representation. Since Z^{sem} is not a single global vector but a sequence composed of M tokens, each token corresponds to a different aspect, such as holidays, accidents, or congestion diffusion. Different nodes and time steps may require attention to different semantic aspects. Therefore, cross-attention fusion is adopted. Each node and time step uses its own spatiotemporal state as a cue to select the most relevant semantic information from the M tokens. Specifically, H^{st} is treated as the Query, while Z^{sem} is treated as the Key and Value, yielding the semantically enhanced representation:

$$H^{\text{fin}} = f_{\text{ca}}(H^{\text{st}}, Z^{\text{sem}}), H^{\text{fin}} \in \mathbb{R}^{N \times T \times d} \quad (9)$$

where $f_{\text{ca}}(\cdot)$ denotes the cross-attention-based fusion function. H^{fin} is the comprehensive representation produced by the teacher model after integrating structural, temporal, and scenario information. However, the prediction of 24 h traffic flow for the next day depends only on the node states summarized at the current time. Therefore, the slice corresponding to the last historical time step is extracted from H^{fin} along the temporal dimension, yielding the current state representation of all nodes, $Z \in \mathbb{R}^{N \times d}$. This representation is then mapped by a multilayer perceptron (MLP) prediction head to the traffic flow prediction for the future T_{pred} steps:

$$\hat{Y}_T = \text{MLP}_o(Z) \in \mathbb{R}^{N \times T_{\text{pred}}} \quad (10)$$

Through the above three-branch joint encoding, the teacher model integrates causal propagation patterns, long-range temporal evolution, and scenario-semantic priors under atypical scenarios into a unified high-level spatiotemporal representation, providing the prediction foundation for UAV pre-deployment. This comprehensive perception capability is condensed into two transferable knowledge carriers: the teacher response output $\hat{Y}_T$ and the last-time-step slice Z_T of the teacher feature output. These two outputs serve as response-level and feature-level distillation signals, respectively, for the subsequent knowledge distillation process.

4.3. Knowledge Distillation Model

The prediction model needs to run locally at the sparsely distributed UAV base stations and infer autonomously under conditions such as communication interruptions. Although the teacher model can conduct complex traffic flow prediction by leveraging multi-source inputs, its large number of parameters and high computational cost make it unsuitable for direct deployment on edge devices. Knowledge distillation provides an effective solution by transferring the knowledge learned by the teacher model to a lightweight student model, thereby satisfying both deployment constraints and prediction accuracy requirements. As shown in Figure 3, this section proposes a multi-level knowledge distillation model. Unlike conventional distillation methods, the proposed model is jointly driven by response-level and feature-level distillation, considering the complex spatiotemporal dependencies in traffic flow prediction. Meanwhile, hard supervision from ground-truth labels is incorporated to prevent the student model from falling into suboptimal local minima.

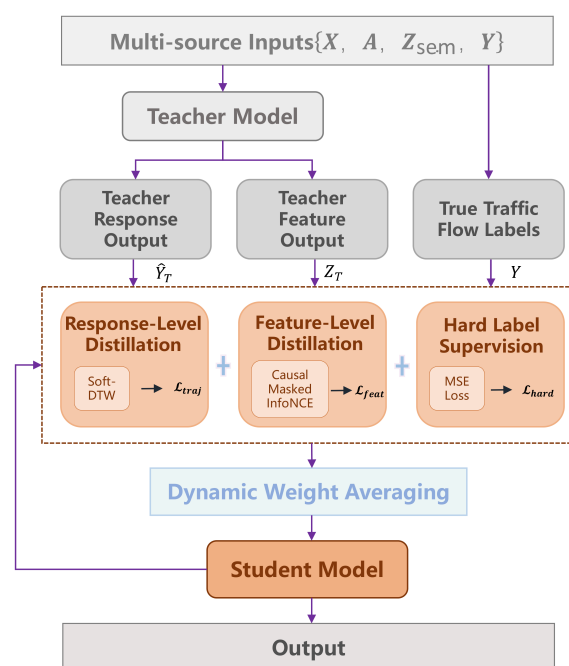


Figure 3. Knowledge distillation process. Arrows indicate the flow of supervision and knowledge transfer.

Response-level distillation is used to encourage the student model to approximate the final decisions of the teacher model in the output space. The teacher response $\hat{Y}_T$ is generated under the joint influence of X , A , and Z^{sem} . It contains cross-segment spatial propagation structures, the historical temporal evolution patterns of each node, and the teacher model's understanding of complex scenario semantics, such as holidays, accident impacts, and congestion. By aligning this soft target in the output space, the student model can implicitly inherit cross-segment spatiotemporal propagation priors. To enable the student model to accurately capture the time-lagged fluctuations caused by the spatial propagation of congestion peaks and preserve the true evolution trend, Soft dynamic time warping (Soft-DTW) is introduced to replace point-wise alignment with optimal temporal path matching:

$$\mathcal{L}_{\text{traj}} = \frac{1}{N} \sum_{i=1}^N f_{\text{sdwtw}}(\hat{Y}_S^{(i)}, \hat{Y}_T^{(i)}; \gamma) \quad (11)$$

where $\gamma > 0$ controls the smoothness of the alignment path.

However, because the teacher model itself may contain prediction bias, its output cannot be fully identical to the actual traffic conditions. Directly fitting the teacher output may cause the student model to converge to a suboptimal solution. Therefore, hard-label supervision is introduced to ensure that the student model does not deviate from the true data distribution. The ground-truth traffic flow label $Y \in \mathbb{R}^{N \times T_{\text{pred}}}$ is used as the hard supervision anchor, and the corresponding loss is defined as:

$$\mathcal{L}_{\text{hard}} = \frac{1}{N \cdot T_{\text{pred}}} \|\hat{Y}_S - Y\|_F^2 \quad (12)$$

Nevertheless, aligning only $\hat{Y}_T$ would lose the intermediate information through which the teacher model internally fuses A and Z^{sem} . Therefore, this section further introduces feature-level distillation to align the teacher intermediate representation Z_T before the prediction head, so that causal propagation patterns and scenario semantics can be jointly injected into the hidden space of the student model. Directly aligning feature vectors with Euclidean distance can easily cause feature collapse. To address this issue, an information noise-contrastive estimation (InfoNCE)-based contrastive loss is introduced to transfer features by maximizing the lower bound of mutual information between teacher and student representations. However, in feature-level contrastive distillation, conventional methods usually treat all other road segments in the same batch as negative samples and push them apart. In a strongly connected expressway network, this may cause serious false-negative problems. Therefore, this paper designs a causal-masked contrastive loss mechanism. Specifically, a continuous soft gate based on a causal threshold is constructed to output smooth soft weights and adaptively adjust the penalty strength. The maximum conditional Granger test strength between road segments i and j is defined as $c_{i,j} = \max(A_{ij}, A_{ji})$ where θ denotes the causal significance threshold and η denotes the sensitivity parameter. When the causal strength $c_{i,j}$ is greater than the threshold θ , the gate is smoothly closed, thereby suppressing the repulsion between causally related road segments in the latent space. When the causal strength is weak, the gate is opened, allowing the feature distance to be enlarged normally. The causal mask is defined as:

$$I_{i,j} = f_{\text{gate}}(c_{i,j}) \quad (13)$$

After aligning the dimensions of the teacher and student representations through the projection matrix P_S , we introduce a temperature-constrained similarity metric $E(u, v)$. The feature-level distillation loss is defined as:

$$\mathcal{L}_{\text{feat}} = \frac{1}{N} \sum_{i=1}^N -\log \frac{E(\tilde{z}_{S,i}, z_{T,i})}{E(\tilde{z}_{S,i}, z_{T,i}) + \sum_{j \neq i}^N I_{i,j} \cdot E(\tilde{z}_{S,i}, z_{T,j})} \quad (14)$$

This causal-masked design distinguishes the proposed distillation scheme from existing spatiotemporal or LLM-related distillation methods: LightST [36] transfers knowledge without any causal guidance, treating all other segments as equally valid negatives, while KIDCM [33] distills the LLM's own predictive output as a surrogate target rather than leveraging causal structure. The mechanism above instead couples the causal graph directly into the distillation objective, rather than treating distillation as an independent module. Through the above design, multi-source priors are decoupled into three types of supervision signals: $\mathcal{L}_{\text{traj}}$, $\mathcal{L}_{\text{hard}}$, and $\mathcal{L}_{\text{feat}}$. However, these losses differ in gradient magnitude and convergence speed. If static constants are used for linear weighting, the large feature-contrastive gradients in the early training stage may overwhelm the subtle evolution signals from the response level. To coordinate these three supervision signals, this paper introduces the dynamic weight averaging (DWA) mechanism [37], which allows

the network to adaptively allocate weights according to the real-time learning progress of each task. First, based on the loss states of the previous two epochs, the relative descent rate at epoch $e - 1$ is calculated to quantify the convergence lag of each constraint:

$$r_k^{(e1)} = f_{\text{rate}}\left(\mathcal{L}_k^{(e1)}, \mathcal{L}_k^{(e2)}\right) \quad (15)$$

A larger ratio indicates slower convergence of the corresponding loss, and therefore a larger loss weight should be assigned to it in the current epoch to accelerate optimization. Then, a Softmax function with a smoothing temperature factor Γ is used to redistribute the weights:

$$w_k^{(e)} = f_{\text{dwa}}\left(r_k^{(e1)}; \Gamma\right) \quad (16)$$

Finally, the global joint optimization objective at epoch e is defined as the dynamically weighted sum of all constraint signals:

$$\mathcal{L}_{\text{final}}^{(e)} = \sum_{k \in K} w_k^{(e)} \mathcal{L}_k^{(e)} \quad (17)$$

The joint objective in Equation (17) governs how the student is trained under teacher supervision; the resulting change in online inference cost can be characterized quantitatively. Let N be the number of ETC nodes, L the input-window length, $|E|$ the number of causal edges in A , M the number of semantic tokens, d the teacher's hidden dimension, and d_s the student's channel dimension, with $d_s \ll d$. The teacher's dominant cost is $\mathcal{O}(L |E| d + NT^2d + NTMd)$, where the temporal self-attention term NL^2d grows quadratically with L and the semantic cross-attention term $NTMd$ scales with the number of LLM-generated tokens. By removing both the temporal self-attention and the semantic cross-attention branches, the student model—composed of a gated temporal convolutional network (TCN), a fixed-graph GCN, and a GRU, as detailed in Section 4.4—reduces this to $\mathcal{O}(NLkd_s^2 + L |E| d_s + NLd_s^2)$, where k is the convolution kernel size. This cost is linear in L and contains no token-level interaction, so that once trained under the losses above, the student requires only the ETC time series and the fixed causal graph A at inference, with both the LLM and the teacher network entirely removed from the online pipeline.

Through the multi-level knowledge distillation model, complex cross-segment causal propagation patterns and scenario priors are successfully transferred to the student model. This removes the dependence on online inference by cloud-based large models and enables the student to approximate the teacher's prediction performance, thereby providing accurate and robust decision support for UAV pre-deployment.

4.4. Student Model

A lightweight student model is still required to carry the transferred knowledge while satisfying the edge-deployment constraints of UAV base stations. Therefore, as shown in Figure 4, this paper designs a student model composed of several basic neural network modules. Specifically, TCN, GCN, and GRU are adopted to cover three dimensions: local temporal perception, spatial causal propagation, and long-range dynamic evolution.

When the latest ETC observations are available at each daily forecasting cycle, the edge device must efficiently generate the following 24 h traffic flow forecasts. Temporal convolution-based modeling has been widely used in traffic flow prediction tasks. Compared with recurrent structures that perform step-by-step propagation, TCN rapidly extracts short-term temporal features from time series, such as local congestion bursts or traffic surges, with extremely low inference latency, owing to the parallel computing capability of one-dimensional dilated convolution. However, dilated convolution alone has difficulty selectively retaining key signals such as abrupt congestion changes. Therefore,

this study adopts a gated temporal convolution structure. By combining dilated causal convolution with a gating mechanism, the model can expand its receptive field while adaptively determining which temporal signals at each time step should be propagated forward:

$$X' = XW_x, X' \in \mathbb{R}^{N \times T \times d} \quad (18)$$

$$H_{\text{tcn}}^S = \tanh(\Theta_1 \star X' + b) \odot \sigma(\Theta_2 \star X' + c), H_{\text{tcn}}^S \in \mathbb{R}^{N \times T \times d} \quad (19)$$

where $\star$ denotes dilated causal convolution, Θ_1 denotes the convolution parameters of the content extraction branch, and Θ_2 denotes the convolution parameters of the gating branch. The gradient of $\mathcal{L}_{\text{traj}}$ in knowledge distillation is propagated to Θ_1 and Θ_2 , enabling the gating parameters to continuously adjust the filtering of local temporal patterns under the teacher model's supervision of traffic trend judgment.

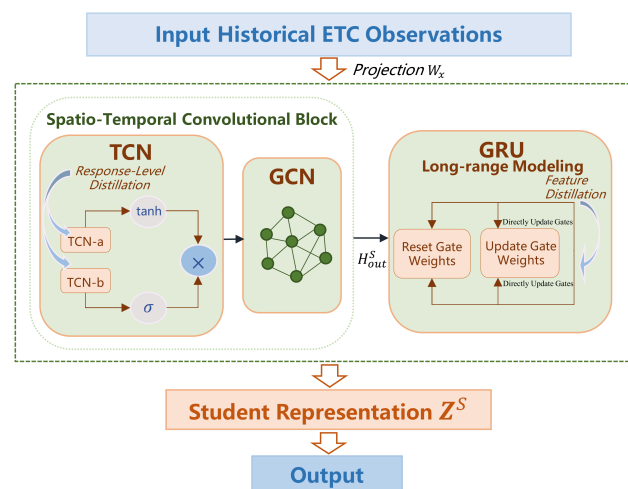


Figure 4. Architecture of the student model. Arrows indicate forward feature propagation.

After local temporal features are obtained, the model must further characterize spatial dependencies among different road segments. Existing graph neural network methods have shown that GCN can effectively model node association structures in traffic networks. The causal prior graph A is constructed and fixed before training. GCN performs one-step neighborhood aggregation on the fixed graph structure, which injects causal directionality constraints into the feature space while avoiding the additional inference overhead introduced by dynamic attention mechanisms such as GAT on edge devices. To preserve node-specific features, this paper adds self-loops to $\tilde{A} = A + I$. The graph convolution operation is then defined as:

$$H_{\text{gcn}}^S = \sigma(\tilde{A}H_{\text{tcn}}^S W_{\text{gcn}}), H_{\text{gcn}}^S \in \mathbb{R}^{N \times T \times d} \quad (20)$$

where W_{gcn} denotes the trainable graph convolution feature transformation parameter, and $\sigma(\cdot)$ is a nonlinear activation function. To preserve low-level features and ensure gradient backpropagation, the gated TCN and GCN are treated as a complete spatiotemporal feature extraction unit, and a residual connection across this unit is introduced:

$$H_{\text{out}}^S = H_{\text{gcn}}^S + X', H_{\text{out}}^S \in \mathbb{R}^{N \times T \times d} \quad (21)$$

Single-step spatial aggregation is performed independently at each time step and cannot capture cross-period evolution trends. Therefore, after obtaining the spatially enhanced features, GRU is introduced to recurrently process the N nodes independently

along the temporal axis. Through the selective memory mechanism of the update gate and reset gate, GRU models dynamic changes over a longer time range:

$$H_{\text{gru}}^S = f_{\text{gru}}(H_{\text{out}}^S), H_{\text{gru}}^S \in \mathbb{R}^{N \times T \times d} \quad (22)$$

$$Z^S = H_{\text{gru},T}^S, Z^S \in \mathbb{R}^{N \times d} \quad (23)$$

Here, Z^S is the compressed representation of the current road-network state learned by the student model. It carries both the local temporal patterns extracted by TCN and the spatial causal information propagated by GCN. The teacher's last-time-step fused representation Z^T contains causal structural knowledge and scenario-semantic priors related to holidays, accidents, and other atypical conditions. The gradient of the feature-level distillation loss $\mathcal{L}_{\text{feat}}$ directly affects the update-gate and reset-gate parameters of GRU, allowing Z^T to guide the hidden-state recurrence of the student GRU. To align the hidden dimensions of the teacher and student representations for causal-masked contrastive distillation, a projection layer is introduced to map Z^S into the teacher feature space:

$$\tilde{Z}^S = Z^S W_p, \tilde{Z}^S \in \mathbb{R}^{N \times d_T} \quad (24)$$

where $W_p \in \mathbb{R}^{d_S \times d_T}$ denotes the feature-alignment projection matrix. The projected student representation $\tilde{Z}^S$ is then used together with the teacher feature Z^T for causal-masked contrastive distillation. The overall training objective is consistent with the knowledge distillation formulation described above and consists of the dynamically weighted sum of $\mathcal{L}_{\text{hard}}$, $\mathcal{L}_{\text{traj}}$, and $\mathcal{L}_{\text{feat}}$ through the DWA mechanism.

Through the above collaborative design, the lightweight student model is independent of both the LLM and the teacher network during inference. With the ETC time series and the precomputed fixed causal graph as inputs, it can approximate the prediction performance of the teacher model. In practice, the traffic-flow predictions of the student model are translated into a pre-deployment plan at each base station. Since a UAV can only cover the segments within its flight range on a single sortie, each base station first narrows its attention to the segments it can actually reach; among these reachable segments, those with the highest predicted near-term load are prioritized for UAV deployment. It should also be clarified that the segments identified through this predicted-load ranking are not identified in a risk-agnostic manner. Because the semantic branch described in Section 4.2.2 already conditions the teacher's, and hence the distilled student's, prediction on recorded accident, control, and congestion events together with an explicit risk assessment, the predicted load used for ranking already reflects risk-relevant scenario context to some extent. The current deployment rule, however, still ranks segments by predicted load alone, without separately re-weighting this risk signal. Explicitly fusing the risk-related output of the semantic prior with the predicted load into a joint, separately weighted priority score is left for future work. This provides a lightweight implementation path for localized rolling prediction and pre-deployment decision-making at expressway UAV base stations. A full multi-station scheduling and routing optimization that additionally considers endurance limits, inter-station coordination, and fleet balancing is beyond the scope of the present forecasting-focused study and is left for future work.

5. Experiments

5.1. Dataset and Evaluation Metrics

ETC gantry data from Hubei Province are used to evaluate road-segment-level traffic flow forecasting across the provincial expressway network. Currently, ETC has achieved full coverage on expressways in Hubei Province. ETC gantries located at expressway

ramps record information such as vehicle passing time, location, travel direction, and vehicle type, enabling the calculation of traffic volume and average speed for all road segments. Aggregated ETC data can reflect the day-scale traffic flow state of the entire expressway network.

Figure 5 shows the spatial distribution of the expressway network and ETC gantries in Hubei Province. The ETC gantries cover major traffic corridors and surrounding connecting routes in Wuhan, Xiangyang, Yichang, Huanggang, Shiyan, and other areas, reflecting the multi-center and multi-corridor traffic operation characteristics of the expressway network in Hubei Province. The ETC node data used in the experiments mainly include two types of features: traffic volume and average speed. The sampling interval is 30 min. This temporal granularity can capture the evolution of traffic flow during peak emergence, persistence, and dissipation, while reducing random disturbances caused by excessively high-frequency sampling. This study uses records from all $N = 1713$ gantry nodes over $T = 4946$ time steps at the 30 min sampling interval described above. The data are partitioned strictly in chronological order into training, validation, and test sets at a 70%:15%:15% ratio (3462, 742, and 742 time steps, respectively), without shuffling, so that the test period is strictly posterior to the training period and no future information leaks into model fitting. Consistent with the day-scale rolling formulation described in Section 4.2.3, the model predicts the traffic flow of all N road segments over the next 24 h from the current historical window, corresponding to a prediction horizon of 48 steps at the 30 min resolution.

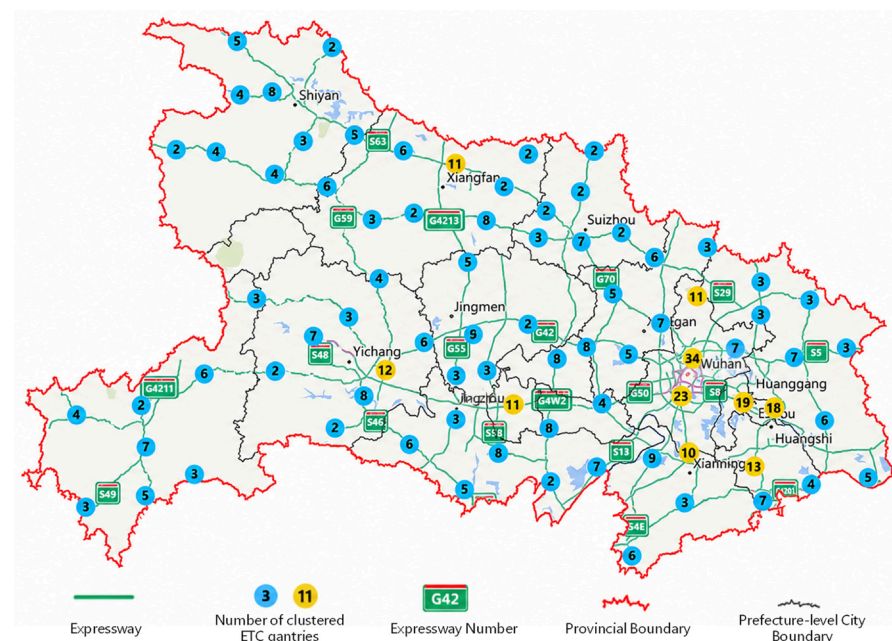


Figure 5. Spatial distribution of the expressway network and ETC gantries in Hubei Province.

To quantitatively evaluate the prediction performance of different models and analyze their applicability for guiding UAV pre-deployment, five metrics are adopted: MAE, RMSE, MAPE, R^2 , and KGE. MAE reflects the mean absolute error, RMSE is more sensitive to large errors, MAPE measures the prediction error relative to the ground truth, and R^2 reflects the model's ability to explain real traffic flow variations. KGE is used as a comprehensive consistency metric to evaluate the agreement between predicted and observed sequences in terms of correlation, variability, and mean bias. Both R^2 and KGE are reported as percentages in the tables.

Let y_i and $\hat{y}_i$ denote the ground-truth value and predicted value of the i -th sample, respectively, and let M denote the number of test samples. The metrics are defined as follows:

1. Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{M} \sum_{i=1}^M |y_i - \hat{y}_i| \quad (25)$$

2. Root Mean Square Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{M} \sum_{i=1}^M (y_i - \hat{y}_i)^2} \quad (26)$$

3. Mean Absolute Percentage Error (MAPE)

$$\text{MAPE} = \frac{1}{M} \sum_{i=1}^M \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \quad (27)$$

4. Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^M (y_i - \hat{y}_i)^2}{\sum_{i=1}^M (y_i - \bar{y})^2} \quad (28)$$

5. Kling–Gupta Efficiency (KGE)

$$\text{KGE} = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \quad (29)$$

where r is the Pearson correlation coefficient between the predicted and observed sequences, $\alpha = \sigma_{\hat{y}}/\sigma_y$ denotes the variability ratio, and $\beta = \mu_{\hat{y}}/\mu_y$ denotes the mean bias ratio.

5.2. Comparative Experiments

5.2.1. Baseline Models

To verify the effectiveness of the proposed method, eight representative baselines related to traffic state prediction are selected. These baselines cover different technical paradigms, including recurrent neural networks, frequency-domain learning, segmented recurrent modeling, and convolution-based long-term prediction. A brief introduction to each baseline is provided below.

LSTM [38]: This model uses LSTM as the core temporal modeling unit and combines convolutional feature extraction, attention mechanisms, and parameter optimization strategies to enhance the capture of local fluctuation features and long-term temporal dependencies in traffic flow sequences.

BiGRU [39]: This model uses BiGRU as the core bidirectional temporal modeling unit and combines CNN-based spatial feature extraction with an attention mechanism to fuse forward and backward contextual information in traffic flow sequences.

FreTS [40]: FreTS is a frequency-domain MLP prediction model. It transforms time-series signals into the frequency domain and learns cross-time and cross-variable dependencies to enhance global pattern modeling in long-term prediction.

SegRNN [41]: SegRNN replaces point-wise iteration with segmented iteration and replaces recursive prediction with parallel multi-step prediction, alleviating the efficiency

and accuracy degradation caused by excessive recurrent steps in traditional RNNs for long-term prediction tasks.

ModernTCN [42]: ModernTCN is a modern pure-convolutional time-series model. It improves the traditional TCN structure by expanding the effective receptive field, thereby enhancing its ability to model long-range dependencies.

ConvTimeNet [43]: ConvTimeNet proposes a deformable patching layer to adaptively capture local temporal patterns in a data-driven manner. It also designs multi-scale hierarchical convolution blocks to establish dependencies among basic unit representations at different scales.

STGCN [24]: STGCN is a graph-based spatiotemporal prediction model. It uses graph convolution to capture spatial dependencies among road-network nodes and gated temporal convolution to model temporal dynamics, enabling efficient end-to-end traffic flow prediction without relying on recurrent structures.

DCRNN [25]: DCRNN models traffic flow propagation as a diffusion process on a directed road graph. It integrates bidirectional diffusion convolution with an encoder–decoder recurrent architecture and employs scheduled sampling for multi-step prediction, thereby capturing directional spatial dependencies and nonlinear temporal evolution patterns.

5.2.2. Results and Analysis

The comparative experiments in this section are conducted on two test subsets constructed according to an explicit typical/atypical criterion, consistent with the scenario distinction introduced in Section 4.2.2. A period is classified as a typical scenario when it contains no statutory holiday and no recorded incident cluster, so that traffic flow predominantly follows ordinary workday and weekend patterns; it is classified as an atypical scenario when it coincides with sustained holiday or seasonal travel demand, producing a persistently elevated, peak-plateau flow level across the network. The typical-scenario subset was selected from January, whereas the atypical-scenario subset was selected from July, corresponding to the sustained seasonal travel peak in this network.

To evaluate prediction performance at a finer temporal granularity than a single aggregated value, the test set of each scenario is divided into 12 consecutive temporal blocks, and the MAE and RMSE of every method are computed within each block. Figures 6 and 7 present the resulting block-wise error distributions and the corresponding average ranks of all nine methods in the typical and atypical scenarios, respectively.

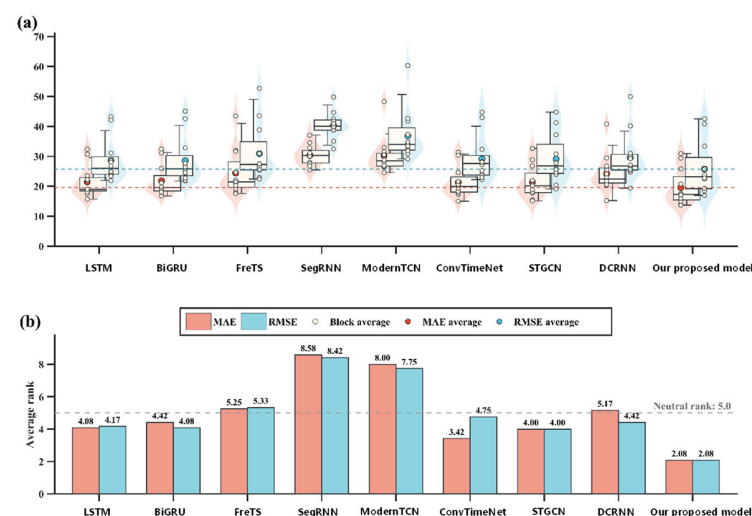


Figure 6. Comparison experiments in typical scenarios: (a) block-wise MAE and RMSE distributions; (b) average ranks of the compared methods.

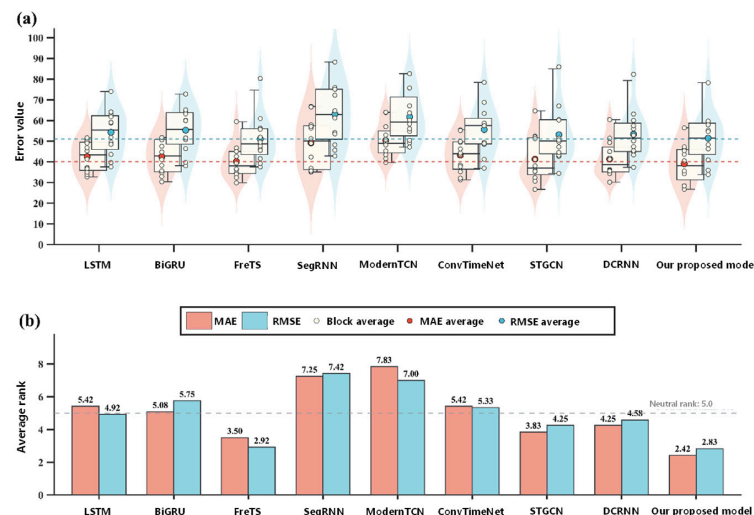


Figure 7. Comparison experiments in atypical scenarios: (a) block-wise MAE and RMSE distributions; (b) average ranks of the compared methods.

Figure 6 presents the resulting block-wise error distribution and average ranks of all nine methods in the typical scenario. In Figure 6a, the per-block MAE and RMSE are visualized as combined violin and box plots, with the two dashed lines marking the global MAE and RMSE averages of the proposed method. The proposed method shows the most compact distribution among all nine methods, with the narrowest spread and no block deviating far above these two lines, indicating that its accuracy remains stable across the test period rather than being driven by a few favorable blocks. SegRNN and ModernTCN show the widest spreads and the highest medians, consistent with their weak performance in Table 1. Figure 6b summarizes this block-wise behavior as an average rank across the 12 blocks, separately for MAE and RMSE, where a lower rank indicates more consistently accurate prediction and the dashed line marks the neutral rank of 5.0 among nine methods. The proposed method attains the best average rank on both metrics, 2.08 and 2.08, clearly ahead of every baseline. ConvTimeNet and STGCN follow as the most consistent baselines, while FreTS, despite its moderate aggregate accuracy in Table 1, falls slightly below the neutral rank on both metrics, indicating that its block-to-block performance is less stable than several other baselines.

Table 1. Performance comparison of different methods in typical scenarios.

Method	MAE	RMSE	MAPE (%)	R ² (%)	KGE (%)
LSTM	21.4266	29.3574	26.7862	90.0321	92.6342
BiGRU	21.7773	29.4454	24.5763	89.9542	92.6995
FreTS	24.3738	32.1253	37.4084	87.4714	90.5069
SegRNN	30.3752	40.5065	44.3796	78.3121	83.6123
ModernTCN	30.2475	37.7274	51.8575	81.5834	71.4492
ConvTimeNet	21.2584	29.9321	24.6137	89.5194	91.8688
STGCN	21.3594	30.0813	24.7960	89.3846	92.0059
DCRNN	24.1765	30.5188	36.2782	88.9858	90.1352
Our proposed model	19.5246	26.9771	22.6109	92.0497	94.2295

As shown in Table 1, none of the eight baselines leads across every metric: ConvTimeNet achieves the lowest MAE, LSTM the lowest RMSE and highest R², and BiGRU the lowest MAPE and highest KGE, a spread that fits the typical scenario's fairly regular workday and weekend rhythm, which purely temporal models can capture reasonably well even without spatial or scenario information. SegRNN and ModernTCN perform worst overall, likely because their architectures are better suited to short-horizon prediction

than to the long, network-scale horizon used here. STGCN and DCRNN, which introduce graph convolution over spatial adjacency, perform comparably to the stronger temporal baselines but do not surpass them, indicating that graph convolution alone does not confer an advantage without an explicit causal structure. Table 1 shows that, in the typical scenario, the proposed method outperforms all eight baselines across all five metrics, namely MAE, RMSE, MAPE, R^2 , and KGE. Relative to the strongest baseline on each metric, it reduces MAE by approximately 8.2%, RMSE by approximately 8.1%, and MAPE by approximately 8.0%, and improves R^2 , by approximately 2.2% and KGE by approximately 1.7%.

Similarly, Figure 7 shows the same block-wise analysis for the atypical scenario. The proposed method again has the most compact distribution and the best average rank, 2.42 on MAE and 2.83 on RMSE, while FreTS ranks second at 3.50/2.92 despite its close performance on the aggregated metrics in Table 2.

Table 2. Performance comparison of different methods in the atypical scenario.

Method	MAE	RMSE	MAPE (%)	R^2 (%)	KGE (%)
LSTM	42.5198	55.2470	12.1345	91.7775	93.4712
BiGRU	42.4698	56.2261	11.2354	91.4835	94.7152
FreTS	40.1379	52.1859	11.0054	92.6635	95.5363
SegRNN	49.1884	64.5422	12.8936	88.7779	92.9659
ModernTCN	50.3723	62.6741	14.5854	89.4181	83.4754
ConvTimeNet	43.1990	56.6618	11.6310	91.3510	94.7681
STGCN	41.2399	54.5903	11.1323	91.9718	93.4917
DCRNN	41.2931	54.4459	12.0431	92.0142	93.6251
Our proposed model	39.0477	52.6934	10.5258	92.5200	94.8080

Table 2 shows that STGCN and DCRNN once again lag, while SegRNN and ModernTCN remain the weakest; the reasons are the same as in the typical case. The proposed method leads in MAE and MAPE, outperforming FreTS by approximately 2.7% and 4.4%, respectively. Although FreTS has a slight lead in RMSE, R^2 , and KGE, the margins are all less than 1%.

The comparative results across the two scenarios show that the proposed road-segment-level traffic flow prediction framework, which integrates a conditional Granger causal graph and LLM-generated scenario priors, achieves stable and competitive prediction performance under both low-flow, high-volatility conditions and high-flow sustained-peak conditions. These results comprehensively validate the effectiveness and robustness of the proposed framework for ETC-data-driven expressway traffic flow prediction.

5.2.3. Computational Cost and Inference Efficiency

In addition to predictive accuracy, we further compare the computational cost of the proposed student model against the teacher model and all baseline methods, in order to substantiate the edge-deployment motivation stated in Section 1. All models are evaluated on the same platform, an NVIDIA GeForce RTX 4060 Laptop GPU, with a fixed random seed of 42, so that the comparison reflects relative computational cost under identical conditions rather than an on-board measurement on a specific UAV-mounted device. The teacher model adopts a high-capacity configuration of five layers with 256 hidden units, whereas the student model uses a single layer with 128 hidden units. Model size is reported as the FP32 storage of the trainable parameters, and inference time is reported as the mean and standard deviation of the per-sample forward-pass latency over the entire test set. All latencies measure the network forward pass only: as described in Section 4.2.2, the LLM-generated semantic prior is produced offline once per day-scale rolling cycle and

cached, so it does not enter the per-sample online inference path of either the teacher or the student.

As shown in Table 3, the student model attains the lowest cost on all four indicators. Relative to the teacher, it reduces the trainable parameters from 65.7039×10^6 to 4.4126×10^6 and the model size from 250.64 MB to 16.83 MB (an approximately $14.9\times$ reduction), while shortening the per-sample inference time from 3.0352 ms to 0.8294 ms (an approximately $3.7\times$ speedup). It is also lighter than every baseline, including the smallest one, LSTM. More importantly, after distillation the student’s online inference depends only on the ETC time series and the fixed causal graph A, with both the LLM and the teacher network entirely removed from the deployed pipeline—whereas the teacher still relies on the cached semantic prior and its cross-attention branch—which is the essential sense in which the student, and not the teacher, is deployable at resource-constrained expressway UAV base stations.

Table 3. Comparison of computational cost and inference efficiency of different models.

Model	Trainable Parameters	GFLOPs	Model Size (MB)	Inference Time (Ms/Sample)
LSTM	5.3420×10^6	0.513	20.3838	2.0445 ± 1.4186
BiGRU	11.1654×10^6	1.072	42.6000	2.6646 ± 1.5516
FreTS	11.5443×10^6	1.108	44.0439	3.2576 ± 2.5045
SegRNN	25.2219×10^6	2.421	96.2204	5.5274 ± 0.7435
ModernTCN	53.5586×10^6	5.142	204.3104	4.6269 ± 0.1009
ConvTimeNet	50.5352×10^6	4.851	192.8922	3.1623 ± 1.8396
STGCN	79.7645×10^6	7.657	304.2775	7.7098 ± 1.3391
DCRNN	8.9293×10^6	0.857	34.0626	3.7015 ± 0.1763
Teacher	65.7039×10^6	6.308	250.6407	3.0352 ± 0.1218
Student	4.4126×10^6	0.424	16.8328	0.8294 ± 0.1545

5.3. Ablation Study

To verify the synergistic effects of the components in the proposed framework, three groups of ablation experiments are designed to investigate the effects of LLM-generated scenario-semantic priors, the conditional Granger causal graph, and teacher-model distillation supervision on prediction performance. Figure 8a presents the network-wide mean observed and predicted traffic flows over a representative 24 h test window in the typical scenario, obtained by averaging all 1713 ETC gantry nodes at each 30 min time step, while Figure 8b shows the corresponding network-wide mean signed prediction errors. These curves are provided only for qualitative visualization; all metrics in Table 4 are calculated over the complete typical-scenario test subset. The ablation results in the typical scenario show that all three modules make clear contributions to prediction performance.

Table 4. Ablation study of different components in typical scenarios.

Method	MAE	RMSE	MAPE (%)	R^2 (%)	KGE (%)
Our proposed model	13.4477	18.6074	13.8982	95.6036	96.5624
<i>w/o</i> LLM	14.1009	19.4177	19.3940	94.4821	92.0184
<i>w/o</i> Granger graph	19.6836	24.3339	33.8281	92.4812	93.8114
<i>w/o</i> teacher guidance	36.2493	46.9113	42.9031	72.0564	82.5705
<i>w/o</i> response distillation	26.8411	35.5448	39.3905	81.9573	88.0422
<i>w/o</i> feature distillation	20.2239	25.0590	36.3672	90.0264	91.4895
fixed loss weights	25.4020	31.8950	28.0062	85.0827	90.8144

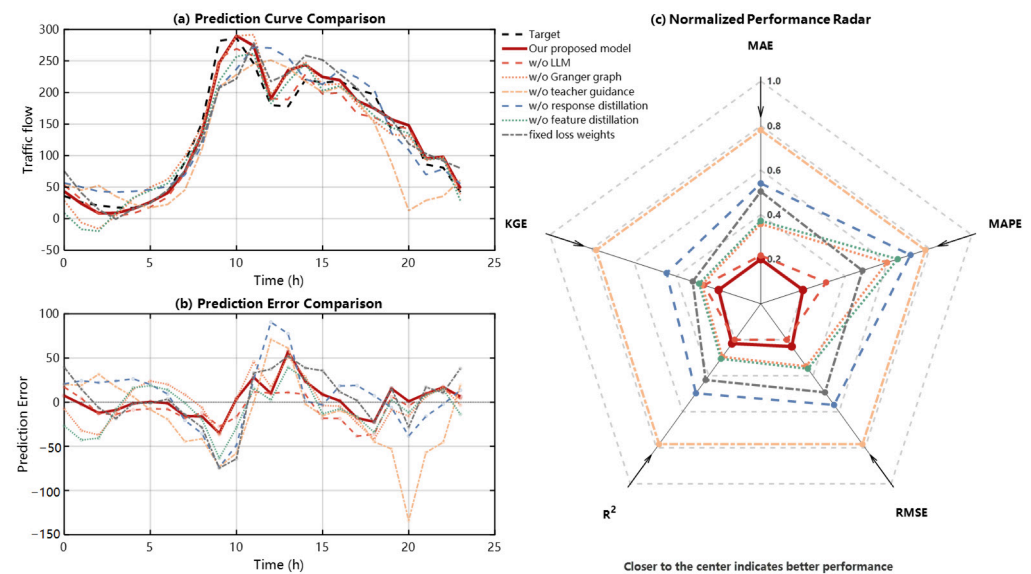


Figure 8. Ablation study in typical scenarios. Arrows indicate the direction of the normalized metric axes; values closer to the center indicate better performance.

As shown in Table 4, after removing the LLM-generated scenario knowledge, MAPE increases from 13.8982% to 19.3940%. In addition, the prediction curve in Figure 8a clearly shows that this variant exhibits obvious sustained drift during the decline period from 14:00 to 20:00. After removing the conditional Granger causal graph, MAE and MAPE increase to 19.6836 and 33.8281%, respectively. This indicates that relying only on implicit correlations cannot accurately identify the propagation direction and long-lag effects of expressway traffic flow, and may easily misinterpret synchronous fluctuations or common upstream influences as true spatial dependencies. The most significant performance degradation occurs when teacher-model distillation supervision is removed. In this case, MAE, RMSE, and MAPE deteriorate to 36.2493, 46.9113, and 42.9031%, respectively, while R^2 decreases to 72.0564%. Breaking distillation down further, removing response-level distillation alone raises MAE and MAPE to 26.8411 and 39.3905%, respectively, whereas removing feature-level distillation alone raises them only to 20.2239 and 36.3672%, indicating that response-level supervision contributes more than feature-level alignment to the student's overall accuracy, though both are needed to approach the full model. Replacing the dynamic loss weighting with fixed weights degrades MAE to 25.4020 and R^2 to 85.0827, a drop close in magnitude to removing response distillation, showing that adaptively balancing the loss terms during training is not a minor implementation detail but a meaningful contributor to final accuracy. As shown in Figure 8a,b, the *w/o* teacher guidance variant shows severe deviation lasting several hours after the traffic peak, with extremely large error peaks. Its polygon area in the radar chart in Figure 8c is also much larger than those of the other variants. These results indicate that the causal structural knowledge and scenario-semantic knowledge integrated by the teacher model during training cannot be fully learned automatically by the lightweight student model through hard-label supervision alone. Feature-level distillation is therefore critical for maintaining high accuracy in edge-deployed models.

In the atypical scenario, as shown in Table 5, compared with the *w/o* LLM variant, the complete model reduces MAE, RMSE, and MAPE by 8.34%, 13.50%, and 20.18%, respectively. The curves in Figure 9 also show that the complete model achieves better calibration of the prediction distribution. This indicates that LLM-generated scenario priors are effective not only in the typical scenario but also in the atypical scenario, improving accuracy during peak and sustained-congestion periods. With semantic priors related

to holidays, congestion, and external events, the model can significantly reduce relative errors. Notably, removing response-level distillation in the atypical scenario raises MAE to 43.0036, the largest degradation among all variants and even higher than removing teacher guidance entirely, which raises MAE to 41.2878; this suggests that under sustained high-flow conditions, direct output-level supervision from the teacher is especially important for calibrating the student's predictions. Removing feature-level distillation alone has a comparatively smaller effect, raising MAE only to 37.6376, and fixed loss weights again produce a substantial drop, to 40.4610, consistent with the typical-scenario results.

Table 5. Ablation study of different components in atypical scenarios.

Method	MAE	RMSE	MAPE (%)	R^2 (%)	KGE (%)
Our proposed model	32.3107	42.1787	7.0943	94.7757	94.9435
<i>w/o</i> LLM	35.2525	48.7605	8.8882	93.0181	94.0424
<i>w/o</i> Granger graph	39.2264	51.8421	9.1093	92.1077	91.1225
<i>w/o</i> teacher guidance	41.2878	55.3911	9.7009	90.9901	92.0992
<i>w/o</i> response distillation	43.0036	56.7335	10.0142	90.5481	93.9910
<i>w/o</i> feature distillation	37.6376	48.8828	9.0290	92.9830	93.0219
fixed loss weights	40.4610	49.0837	12.4331	92.9252	88.0951

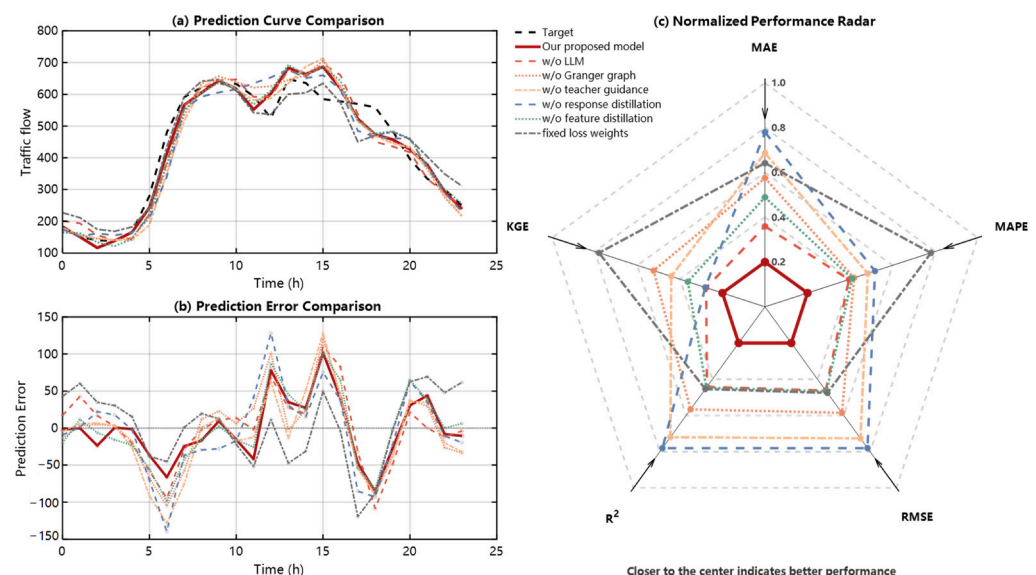


Figure 9. Ablation study in atypical scenarios.

Compared with the *w/o* Granger graph variant, the complete model reduces MAE and RMSE by 17.63% and 18.64%, respectively, indicating that an explicit causal graph is more important for modeling cross-segment propagation under high-flow conditions. Compared with the *w/o* teacher guidance variant, the complete model reduces MAE, RMSE, and MAPE by 21.74%, 23.85%, and 26.87%, respectively. In Figure 9, this is reflected by the effective suppression of severe error oscillations, further verifying the necessity of teacher guidance for transferring complex knowledge to the lightweight student model.

The above results demonstrate that the three modules are complementary and indispensable. Together, they support the prediction accuracy and robustness of the proposed framework under extreme traffic scenarios.

6. Conclusions

This paper proposes a road-segment-level long-term traffic-flow prediction framework with a 24 h forecasting horizon for expressway law-enforcement UAV pre-deployment. Through three collaborative modules, the framework achieves accurate multi-segment traffic flow prediction under typical and atypical traffic scenarios. Specifically, the framework constructs a directed causal prior graph to characterize global propagation patterns. It then introduces LLM-generated scenario-semantic priors to enhance model robustness under atypical scenarios. Finally, through a multi-level knowledge distillation mechanism, the knowledge learned by the teacher model is collaboratively transferred to a lightweight student model, satisfying the computational constraints of localized edge deployment at UAV base stations.

Compared with the baseline models, the proposed method achieves the best overall performance in the typical scenario and competitive performance in the atypical scenario, exhibiting lower overall prediction errors and the lowest average MAE and RMSE ranks across different temporal blocks. Specifically, the causal graph captures directional, long-lag cross-segment propagation that purely temporal or frequency-domain baselines cannot represent, which is most evident during high-flow peak and rapid-decline periods; the LLM-generated semantic prior keeps prediction error from drifting under atypical, holiday-driven traffic patterns where baselines lacking scenario context degrade most; furthermore, the distilled student model can be deployed on resource-constrained edge devices while maintaining this level of accuracy. The ablation experiments further demonstrate the functionality of the three core modules.

Overall, the proposed model provides a feasible traffic flow prediction support scheme for the pre-deployment of law-enforcement UAVs on expressways. By identifying potential high-load road segments in future periods, the framework supports UAV pre-deployment decisions. Nevertheless, the construction of the causal graph relies on offline historical data, and its adaptability to changes in causal relationships under real-time event disturbances still needs to be improved. In addition, the LLM-generated scenario semantics rely on manually designed prompt templates, and their coverage of different atypical-event types requires further validation. Future work will explore online updating mechanisms for dynamic causal graphs and further investigate adaptive scenario knowledge generation strategies for broader atypical scenarios, with the aim of continuously improving the engineering applicability of the framework in complex expressway environments.

Author Contributions: Conceptualization, Zeen Yang, Zhuoer Wang and Hongjuan Zhang; methodology, Zeen Yang and Zhuoer Wang; software, Zeen Yang and Zhuoer Wang; validation, Zeen Yang, Zhuoer Wang and Bijun Li; formal analysis, Zeen Yang and Zhuoer Wang; investigation, Zeen Yang, Zhuoer Wang and Bijun Li; resources, Hongjuan Zhang and Bijun Li; data curation, Zeen Yang and Zhuoer Wang; writing—original draft preparation, Zeen Yang and Zhuoer Wang; writing—review and editing, Hongjuan Zhang, Bijun Li, Zeen Yang and Zhuoer Wang; visualization, Zeen Yang and Zhuoer Wang; supervision, Hongjuan Zhang; project administration, Hongjuan Zhang; funding acquisition, Hongjuan Zhang. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key R&D Program of China (Grant No. 2023YFB3907102).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: We also gratefully acknowledge the support from the Hubei Traffic Research Institute.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

UAV	unmanned aerial vehicle
DWA	dynamic weight averaging
ETC	electronic toll collection
GAT	graph attention network
CNN	convolutional neural network
TCN	temporal convolutional network
GCN	graph convolutional network
GRU	gated recurrent unit
LLM	large language model

References

1. Pulugurtha, S.S.; Duddu, V.R.; Venigalla, M. Evaluating spatial and temporal effects of planned special events on travel time performance measures. *Transp. Res. Interdiscip. Perspect.* **2020**, *6*, 100168. [\[CrossRef\]](#)
2. Butilă, E.V.; Boboc, R.G. Urban Traffic Monitoring and Analysis Using Unmanned Aerial Vehicles (UAVs): A Systematic Literature Review. *Remote Sens.* **2022**, *14*, 620. [\[CrossRef\]](#)
3. Cheng, S.; Lu, F.; Peng, P.; Wu, S. A spatiotemporal multi-view-based learning method for short-term traffic forecasting. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 218. [\[CrossRef\]](#)
4. Tedjopurnomo, D.A.; Bao, Z.; Zheng, B.; Choudhury, F.M.; Qin, A.K. A Survey on Modern Deep Neural Network for Traffic Prediction: Trends, Methods and Challenges. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 1544–1561.
5. Du, K.; Guo, X.; Li, L.; Song, J.; Shi, Q.; Hu, M.; Fang, J. Traffic prediction in time series, spatialtemporal, and OD data: A systematic survey. *J. Traffic Transp. Eng. Engl. Ed.* **2025**, *12*, 666–700. [\[CrossRef\]](#)
6. Liu, X.; Qin, L.; Xu, M.; Zhou, Y.; Wang, B.; Yu, W.; Xiong, W. A comprehensive review of traffic flow forecasting based on deep learning. *Neurocomputing* **2026**, *668*, 132269. [\[CrossRef\]](#)
7. Fan, G.; Sabri, A.Q.M.; Rahman, S.S.A.; Pan, L.; Rahardja, S. Emerging Trends in Graph Neural Networks for Traffic Flow Prediction: A Survey. *Arch. Comput. Methods Eng.* **2025**, *32*, 4811–4855. [\[CrossRef\]](#)
8. Jiang, W.; Luo, J.; He, M.; Gu, W. Graph neural network for traffic forecasting: The research progress. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 100. [\[CrossRef\]](#)
9. Bai, J.; Zhu, J.; Song, Y.; Zhao, L.; Hou, Z.; Du, R.; Li, H. A3t-gcn: Attention temporal graph convolutional network for traffic forecasting. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 485. [\[CrossRef\]](#)
10. Chen, K.; Deng, M.; Shi, Y. A temporal directed graph convolution network for traffic forecasting using taxi trajectory data. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 624. [\[CrossRef\]](#)
11. Zong, X.; Yan, H.; Qi, Y. Recent Advances in Multi-source Data Fusion for Traffic Flow Prediction: A Review. *Arch. Comput. Methods Eng.* **2026**, *33*, 1181–1203.
12. Kou, F.; Zhang, Z.; Yao, Y.; Zhu, Y.; Wang, J.; Yuan, R.; Zhu, Y. A survey on long-term traffic prediction from the information fusion perspective: Requirements, methods, applications, and outlooks. *Inf. Fusion* **2026**, *127*, 103677. [\[CrossRef\]](#)
13. Su, X.; Li, P.; Cai, Z.; Guo, L.; Zhang, B. Mixed-Graph Neural Network for Traffic Flow Prediction by Capturing Dynamic Spatiotemporal Correlations. *ISPRS Int. J. Geo-Inf.* **2025**, *14*, 379. [\[CrossRef\]](#)
14. Tank, A.; Covert, I.; Foti, N.; Shojaie, A.; Fox, E.B. Neural Granger Causality. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 4267–4279. [\[PubMed\]](#)
15. Yang, S.; Wu, Q.; Li, Z.; Wang, K. PSTCGCN: Principal spatio-temporal causal graph convolutional network for traffic flow prediction. *Neural Comput. Appl.* **2025**, *37*, 14751–14764. [\[CrossRef\]](#)
16. Simonelli, F.; Centofanti, F.; Lepore, A.; Pariota, L.; Marzano, V.; Palumbo, B. Data-driven investigation of traffic spatial correlation through functional data analysis. *Transp. Res. Part C Emerg. Technol.* **2025**, *174*, 105088. [\[CrossRef\]](#)
17. Méndez, M.; Merayo, M.G.; Núñez, M. Long-term traffic flow forecasting using a hybrid CNN-BiLSTM model. *Eng. Appl. Artif. Intell.* **2023**, *121*, 106041. [\[CrossRef\]](#)
18. Guo, X.; Zhang, Q.; Jiang, J.; Peng, M.; Zhu, M.; Yang, H.F. Towards explainable traffic flow prediction with large language models. *Commun. Transp. Res.* **2024**, *4*, 100150. [\[CrossRef\]](#)
19. Xu, Y.; Liu, M. GPT4TFP: Spatio-temporal fusion large language model for traffic flow prediction. *Neurocomputing* **2025**, *625*, 129562. [\[CrossRef\]](#)
20. Maksoud, N.; AlJassmi, H.; Ali, L.; Masoud, A.R. Applications of large language models and generative AI in transportation: A systematic review and bibliometric analysis. *Transp. Res. Interdiscip. Perspect.* **2025**, *34*, 101699. [\[CrossRef\]](#)

21. Nie, T.; Sun, J.; Ma, W. Exploring the roles of large language models in reshaping transportation systems: A survey, framework, and roadmap. *Artif. Intell. Transp.* **2025**, *1*, 100003. [\[CrossRef\]](#)
22. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. In Proceedings of the 5th International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
23. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph attention networks. In Proceedings of the 6th International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
24. Yu, B.; Yin, H.; Zhu, Z. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 3634–3640.
25. Li, Y.; Yu, R.; Shahabi, C.; Liu, Y. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In Proceedings of the 6th International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
26. Wu, Z.; Pan, S.; Long, G.; Jiang, J.; Zhang, C. Graph WaveNet for deep spatial-temporal graph modeling. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019; pp. 1907–1913.
27. Bai, L.; Yao, L.; Li, C.; Wang, X.; Wang, C. Adaptive graph convolutional recurrent network for traffic forecasting. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 17804–17815.
28. Lan, S.; Ma, Y.; Huang, W.; Wang, W.; Yang, H.; Li, P. DSTAGNN: Dynamic Spatial-Temporal Aware Graph Neural Network for Traffic Flow Forecasting. In Proceedings of the 39th International Conference on Machine Learning, Proceedings of Machine Learning Research, Baltimore, MD, USA, 17–23 July 2022; pp. 11906–11917.
29. Runge, J.; Nowack, P.; Kretschmer, M.; Flaxman, S.; Sejdinovic, D. Detecting and quantifying causal associations in large nonlinear time series datasets. *Sci. Adv.* **2019**, *5*, eaau4996. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Dong, P.; Wang, X.-L.; Bose, I.; Ng, K.K.; Zhang, X.; Zhang, X. Causally aware spatiotemporal multigraph convolutional network for accurate and reliable traffic prediction. *Inf. J. Comput.* **2025**. [\[CrossRef\]](#)
31. Li, S.; Li, H.; Li, X.; Xu, Y.; Lin, Z.; Jiang, H. Causal Intervention Is What Large Language Models Need for Spatio-Temporal Forecasting. *IEEE Trans. Cybern.* **2025**, *55*, 3825–3837. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Yang, X.; Tao, Y.; Ge, H.; Fan, Z.; Akerkar, R.; Koshizuka, N. LLeCaT: LLM Enhanced Causality-Aware Traffic Accidents Post-Effects Prediction. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 19256–19269. [\[CrossRef\]](#)
33. Jiang, R.; Wang, S.; Ma, W.; Zhang, Y.; Fan, P.; Jia, D. A knowledge-informed dynamic correlation modeling framework for lane-level traffic flow prediction. *Inf. Fusion* **2025**, *124*, 103327. [\[CrossRef\]](#)
34. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
35. Romero, A.; Ballas, N.; Kahou, S.E.; Chassang, A.; Gatta, C.; Bengio, Y. FitNets: Hints for thin deep nets. In Proceedings of the 3rd International Conference on Learning Representations ICLR, San Diego, CA, USA, 7–9 May 2015.
36. Zhang, Q.; Gao, X.; Wang, H.; Yiu, S.M.; Yin, H. Efficient traffic prediction through spatio-temporal distillation. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; pp. 1093–1101.
37. Liu, S.; Johns, E.; Davison, A.J. End-to-end multi-task learning with attention. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1871–1880.
38. Topilin, I.; Jiang, J.; Feofilova, A.; Beskopylny, N. Traffic Flow Prediction via a Hybrid CPO-CNN-LSTM-Attention Architecture. *Smart Cities* **2025**, *8*, 148. [\[CrossRef\]](#)
39. Liu, S.; Lin, W.; Wang, Y.; Yu, D.Z.; Peng, Y.; Ma, X. Convolutional neural network-based bidirectional gated recurrent unit-additive attention mechanism hybrid deep neural networks for short-term traffic flow prediction. *Sustainability* **2024**, *16*, 1986. [\[CrossRef\]](#)
40. Yi, K.; Zhang, Q.; Fan, W.; Wang, S.; Wang, P.; He, H.; An, N.; Lian, D.; Cao, L.; Niu, Z. Frequency-domain mlps are more effective learners in time series forecasting. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 76656–76679. [\[CrossRef\]](#)
41. Lin, S.; Lin, W.; Wu, W.; Zhao, F.; Mo, R.; Zhang, H. Segrrnn: Segment recurrent neural network for long-term time series forecasting. *IEEE Internet Things J.* **2026**, *13*, 9861–9871. [\[CrossRef\]](#)
42. Luo, D.; Wang, X. ModernTCN: A Modern Pure Convolution Structure for General Time Series Analysis. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024.
43. Cheng, M.; Yang, J.; Pan, T.; Liu, Q.; Li, Z.; Wang, S. Convtimenet: A deep hierarchical fully convolutional model for multivariate time series analysis. In Proceedings of the Companion Proceedings of the ACM Web Conference, Sydney, Australia, 28 April–2 May 2025; pp. 171–180.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.