MTP: Exploring Multimodal Urban Traffic Profiling with Modality Augmentation and Spectrum Fusion

Haolong Xiang[1, 2], Peisi Wang[1], Xiaolong Xu[1], Kun Yi[3], Xuyun Zhang[4], Quan Z. Sheng[4], Amin Beheshti[4], Wei Fan[5*]

[1] School of Software, Nanjing University of Information Science and Technology, Nanjing, P.R. China
[2]State Key Lab. for Novel Software Technology, Nanjing University, P.R. China
[3]State Information Center, P.R. China
[4]School of Computing, Macquarie University, NSW, Australia
[5]School of Computer Science, University of Auckland, Auckland, New Zealand
{hlxiang, 202512492293, xlxu}@nuist.edu.cn, yikun@bit.edu.cn, {xuyun.zhang, michael.sheng, amin.beheshti}@mq.edu.au, wei.fan@auckland.ac.nz

Abstract

With rapid urbanization in the modern era, traffic signals from various sensors have been playing a significant role in monitoring the states of cities, which provides a strong foundation in ensuring safe travel, reducing traffic congestion and optimizing urban mobility. Most existing methods for traffic signal modeling often rely on the original data modality, i.e., numerical direct readings from the sensors in cities. However, this unimodal approach overlooks the semantic information existing in multimodal heterogeneous urban data in different perspectives, which hinders a comprehensive understanding of traffic signals and limits the accurate prediction of complex traffic dynamics. To address this problem, we propose a novel *M*ultimodal framework, *MTP*, for urban *T*raffic *P*rofiling, which learns multimodal features through numeric, visual, and textual perspectives. The three branches drive for a multimodal perspective of urban traffic signal learning in the frequency domain, while the frequency learning strategies delicately refine the information for extraction. Specifically, we first conduct the visual augmentation for the traffic signals, which transforms the original modality into frequency images and periodicity images for visual learning. Also, we augment descriptive texts for the traffic signals based on the specific topic, background information and item description for textual learning. To complement the numeric information, we utilize frequency multilayer perceptrons for learning on the original modality. We design a hierarchical contrastive learning on the three branches to fuse the spectrum of three modalities. Finally, extensive experiments on six real-world datasets demonstrate superior performance compared with the state-of-the-art approaches.

Code — https://github.com/jorcy3/MTP

Introduction

With rapid urbanization, traffic volumes continue to rise, placing unprecedented pressure on transportation systems (Zhou et al. 2025; Wu et al. 2025). Persistent congestion during peak hours, delayed responses to traffic incidents, and imbalanced road network resource allocation not only significantly harm the travel efficiency and experience of citizens but also directly restrict the operational efficiency of urban economies and environmental sustainability (Liu, Zheng, and Yu 2025). As a core carrier reflecting how transportation systems operate, traffic time series data contains critical information such as traffic flow variations, the operation status of road segments, and early signs of abnormal events (Fang et al. 2025). Thorough profiling of these data enables real-time perception and scientific assessment of traffic conditions. It can quickly identify congested road segments and their congestion levels to provide precise guidance for traffic management authorities (Wang et al. 2024c). Additionally, it offers data support for urban road network planning, the optimization of public transportation routes, and the adjustment of traffic signal timing, thereby fundamentally enhancing the operational efficiency of transportation systems (Xiang et al. 2025). Currently, urban traffic profiling is a fundamental component for achieving intelligent traffic management and addressing urban traffic challenges.

Traditional traffic data processing methods mostly rely on static feature extraction, such as sliding window statistics and support vector machines, but they assume that the data distribution is stable and cannot adapt to the dynamic characteristics of actual traffic systems (Zerveas et al. 2021). In practice, traffic data has strong temporal dependence, i.e., the traffic flow characteristics of main roads during morning peak hours are significantly different from those during off-peak hours, and the traffic status under extreme weather may even deviate from the conventional distribution (Cheng et al. 2021). This temporal dynamics in traffic signals leads to a sharp decline in the accuracy of static methods in cross-time profiling (Zhang, Chen, and He 2023). For example, a model trained based on morning peak data will misjudge normal traffic flow during off-peak hours as abnormal. Therefore, designing a dedicated profiling method for the comprehensive temporal features has become a fundamental research issue for the accurate classification of traffic states.

Urban traffic profiling can be divided into two core tasks: one is state profiling, such as smooth, slow, and congested, and the other is event profiling, such as traffic accidents and

*Corresponding author.

road construction (Nie et al. 2023). Neural networks have become the mainstream method due to their time series modeling capabilities, e.g., SVP-T (Zuo et al. 2023) learns representations from both the shape-level and velocity-level of the time series for more robust feature capture. However, traffic data has expanded from single structured time series data to a multimodal form, such as surveillance images, text information, and social media feedback have been incorporated into the analysis (Wen, Ma et al. 2025). Although these data can supplement semantic information, traditional neural networks are designed for a single modality and cannot achieve deep correlation between spatial, visual and textual features, resulting in limited utilization of multimodal information.

To meet the needs of multimodality, large language models (LLMs) based methods have gradually developed for better urban traffic profiling. In terms of time series modeling, TRACK (Han et al. 2025) leverages transformer-based models to learn dynamic road network and trajectory representations for better capturing spatial-temporal dynamics. Besides, CAFO (Li, Wang, and Liu 2024) effectively combines the local feature extraction capabilities of convolutional layers with the ability of attention mechanisms to capture long-range dependencies. In terms of multimodal fusion, urban-level CLIP (Yan et al. 2024) realizes the associated classification of urban images and texts through visual-text pretraining. Although the above methods have made certain progress, they still face core challenges in actual traffic scenarios: existing LLMs are mostly optimized for a single modality. *Large large models* are good at processing image data but are hard to parse the dynamic changes of time series features (Gruver et al. 2023). *Text-Augmented Models* can understand traffic event descriptions but lack the ability to model the time dimension (Wang et al. 2024a). *Time series large models* cannot effectively integrate semantic information in images and texts (Wang et al. 2024b). Despite advancements in textual and visual large models, it has been less investigated in traffic classification by integrating multiple modalities.

To address the above issues, we propose a new Multimodal framework, *MTP*, for urban Traffic Profiling. Specifically, MTP first augments visual and text traffic profiles using the original traffic signals and then incorporates multiple features, including temporal, visual, and textual information for learning. The main contributions are:

- We propose a novel multimodal framework for urban traffic profiling, which firstly augments multimodal features on traffic signals and learns through numeric, visual, and textual perspectives in the frequency domain.
- We design a hierarchical contrastive learning on the augmented image, text, and numerical value to optimize the multimodal learning and fuse the three representations.
- Extensive experiments are conducted on six real-world datasets, which validates the effectiveness of the proposed framework compared with the state-of-the-art baselines. We also design several ablation studies to show the influence of three different modalities and conduct qualitative analysis with visualization for our framework.

Related Work

Traditional Traffic Time Series Profiling. Existing methods for road traffic condition analysis mostly rely on single-modality data. In the field of time series analysis (Fan et al. 2022, 2023), deep learning techniques such as Convolutional Neural Networks (CNNs) (He et al. 2016; Alam et al. 2023), Recurrent Neural Networks (RNNs) (Jin et al. 2017; Zheng et al. 2020), Graph Neural Networks (GNNs) (Zhang et al. 2023; Deng, Wang, and Xue 2024), and Transformer-based methods (Lin et al. 2022a; Zerveas et al. 2021) have been widely used to analyze various traffic conditions. These methods excel at processing structured time series data, driver profiling (Cura et al. 2020), and assessing driving risks (Abdelrahman, Hassanein, and Abu-Ali 2020). However, their core limitation lies in their unimodal nature. Merely analyzing time series data or isolatedly analyzing image and text information is insufficient to capture dynamic real-world traffic conditions.

Traffic Profiling with LLMs. LLMs have powerful capabilities in processing multimodal data, especially in text understanding and generalization (Khattar et al. 2019; Feng et al. 2024), which offer new avenues to address the problem of single-modality information loss. In recent years, researchers have begun to apply LLMs to the field of intelligent transportation. For example, the multimodal framework proposed by Qian et al. (2021) combines BERT and ResNet to jointly capture contextual information; Chen et al. (2024) utilize an LLM-driven framework to optimize vehicle dispatching and navigation; and Yan et al. (2024) use LLMs to enhance textual information and fuse it with images via contrastive learning to generate multimodal representations for urban region profiling. Although these methods demonstrate the potential of LLMs, their applications are often task-specific, which partially hinders their exploration in more general road traffic profiling research.

Traffic Profiling with VLMs. VLMs have made significant breakthroughs in jointly processing and understanding visual and textual information. Many recent works, such as BLIVA (Hu et al. 2024), EMMA (Yang et al. 2024), and OmniActions (Li et al. 2024), have demonstrated the powerful capabilities of VLMs in handling complex visual question answering and multimodal interaction tasks. However, these methods have not fully combined multimodal data to generate powerful representations for road traffic profiling. These methods indicate that VLMs can serve as a powerful "bridge" to transform visual information into high-quality textual information, laying the foundation for subsequent multimodal fusion.

Despite significant progress in the fields mentioned above, a key research gap remains: The joint modeling and fusion of numerical, textual and visual modalities have not been explored in urban traffic profiling, which largely hinders the accurate prediction or classification of traffic conditions.

Problem Definition

Definition 1 (Urban Area) *Given an urban area $\mathbb{U}$, we can divide it into M traffic jurisdictions. For different time intervals T, a corresponding traffic status profiling is conducted.*

Definition 2 (Numerical representation) *By using devices such as sensors or cameras, we can collect data information within an urban area $\mathbb{U}$. Then, the data of urban traffic will be stored in numerical representation, denoted as $\boldsymbol{v}_{\mathbb{U}}$, which contains information such as the traffic background, vehicle position, environment, and item description.*

Definition 3 (Image Augmented representation) *The image data is generated from the original traffic time series data, denoted as $\boldsymbol{g}_{\mathbb{U}}$, which enables the model to capture spatial features from the original temporal data.*

Definition 4 (Text Augmented representation) *The text data is generated from the original urban traffic data, denoted as $\boldsymbol{t}_{\mathbb{U}}$, which enables the model to capture semantic information and from the original temporal data.*

This paper mainly deals with urban traffic analysis that focuses on traffic road conditions and vehicle flow, which we define as a classification task. The state information of the data includes three modalities: original numerical values v, images g, and texts t. Given a traffic time series dataset $\boldsymbol{X} = (\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), ..., (\boldsymbol{x_n}, y_n)$, where each data instance $(\boldsymbol{x_i}, y_i)$ contains the feature representation $\boldsymbol{x_i}$ and the types of transportation y_i. Specifically, the feature representation can be optimized through our multi-modal feature fusion, which can be calculated by:

$$\boldsymbol{x}'_{\mathbb{U}} = \mathbb{H}(\boldsymbol{v}_{\mathbb{U}}, \boldsymbol{g}_{\mathbb{U}}, \boldsymbol{t}_{\mathbb{U}} | \boldsymbol{x}). \tag{1}$$

Finally, we can use the augmented representation to predict the traffic condition through $\boldsymbol{x}'_{\mathbb{U}} \rightarrow y_i$.

Methodology

We propose a novel multimodal framework (MTP) for urban traffic profiling, which learns multimodal features through numeric, visual, and textual perspectives in the frequency domain. As shown in Figure 1, MTP consists of three modal encoder branches and a feature fusion scheme: a) time series modality encoder, b) vision modality encoder and c) text modality encoder. The fused features of our framework can simultaneously retain the temporal patterns of the numerical modality, the intuitive patterns of the visual modality, and the semantic information of the text modality. The following sections will elaborate on the design of each module of our framework.

Time Series Modality Encoder

This module mainly processes the original traffic time series data modality with spectrum conversion technologies (Yi et al. 2023), as shown in "a" part of Figure 1. Time series modality encoder mainly involves the semantic embedding, Fast Fourier Transform (FFT), frequency-domain multi-layer perceptions (MLPs), and inverse Fast Fourier Transform (IFFT). Inspired by word embeddings (Mikolov et al. 2013), we mapped the input data $\boldsymbol{X} \in \mathbb{R}^{n \times l}$ into a hidden representation $\boldsymbol{D} \in \mathbb{R}^{n \times l \times m}$ to introduce richer semantic information, which is realized by a learnable weight vector $\boldsymbol{\psi} \in \mathbb{R}^{1 \times m}$. The process can be denoted as $\boldsymbol{D} = \boldsymbol{X} \times \boldsymbol{\psi}$.

The second step is to convert the spatial domain to the frequency domain, so that our model can extract multi-scale features and periodic features of the traffic time series data. Given the converted input $\boldsymbol{D}$, the Fourier transform of the original time series embedding is defined as:

$$\boldsymbol{\mathcal{D}}^v[k] = \sum_{i=0}^{n-1} \boldsymbol{D}^v[i] e^{-j\frac{2\pi k i}{n}}, \tag{2}$$

where i represents the integral variable, j represents the imaginary unit, and $e^{-j\frac{2\pi k i}{n}} = \cos(\frac{2\pi k i}{n}) - j\sin(\frac{2\pi k i}{n})$. Through the above process, we can obtain the numerical spectrum at the frequency $2\pi k i/n$.

The obtained frequency components are input into the frequency-domain MLPs, and operations are performed through the set complex weight matrix $\boldsymbol{W}$ and bias $\boldsymbol{B}$ to obtain the frequency-domain output result:

$$\boldsymbol{\mathcal{H}}_i = FMLP(\boldsymbol{\mathcal{D}}^v, \boldsymbol{W}, \boldsymbol{B}). \tag{3}$$

The concrete process of frequency-domain MLPs is shown in the green box of the framework figure. The core function of frequency-domain MLPs is to perform nonlinear mapping and feature extraction on the frequency domain features after FFT conversion, enhancing the expression ability of the frequency domain features to meet the requirements for capturing periodic and abnormal patterns in traffic time series analysis. Specifically, frequency-domain MLPs can be calculated by:

$$\boldsymbol{\mathcal{Z}} = ReLU(\mathcal{H}\boldsymbol{W} + \boldsymbol{B}). \tag{4}$$

If the MLPs consists of l layers, then the input of each layer is the output ($\boldsymbol{\mathcal{Z}^l}$) of the frequency-domain MLPs of the previous layer. The complex weight matrix $\boldsymbol{W}$ fulfill the condition: $\boldsymbol{W} = \boldsymbol{W}_i + \eta \boldsymbol{W}_j$, and bias $\boldsymbol{B}$ fulfill the condition: $\boldsymbol{B} = \boldsymbol{B}_i + \eta \boldsymbol{B}_j$. According to the rule of multiplication of complex numbers, we can derive the following condition from Equation (4):

$$\begin{aligned} \boldsymbol{\mathcal{Z}}^l = & ReLU(O(\mathcal{Z}^{l-1})\boldsymbol{W}_i^l - I(\mathcal{Z}^{l-1})\boldsymbol{W}_j^l + \boldsymbol{B}_i^l) \\ & + \eta ReLU(O(\mathcal{Z}^{l-1})\boldsymbol{W}_j^l - I(\mathcal{Z}^{l-1})\boldsymbol{W}_i^l + \boldsymbol{B}_j^l). \end{aligned} \tag{5}$$

where O represents the original parts of frequency components, and I represents the imaginary parts of frequency components.

We use Inverse Fast Fourier Transform (IFFT) to inverse the optimized frequency-domain features to the spatial domain, which provides the features with frequency-domain information. This process provide feature support in the spatial domain for the subsequent feature concatenation (Concat), similarity calculation, and traffic series fusion. The calculation of IFFT is formularized as:

$$\boldsymbol{D}^v[i] = \sum_{i=0}^{n-1} \boldsymbol{\mathcal{D}}^v[k] e^{j\frac{2\pi k i}{n}}, \tag{6}$$

Vision Modality Encoder

The "b" part in Figure 1 is a feature extraction module for the visual modality. The process of the vision modality encoder is to first convert the traffic time series data into an image, and then perform frequency-domain processing on the image to extract visual features. In our approach, the essence

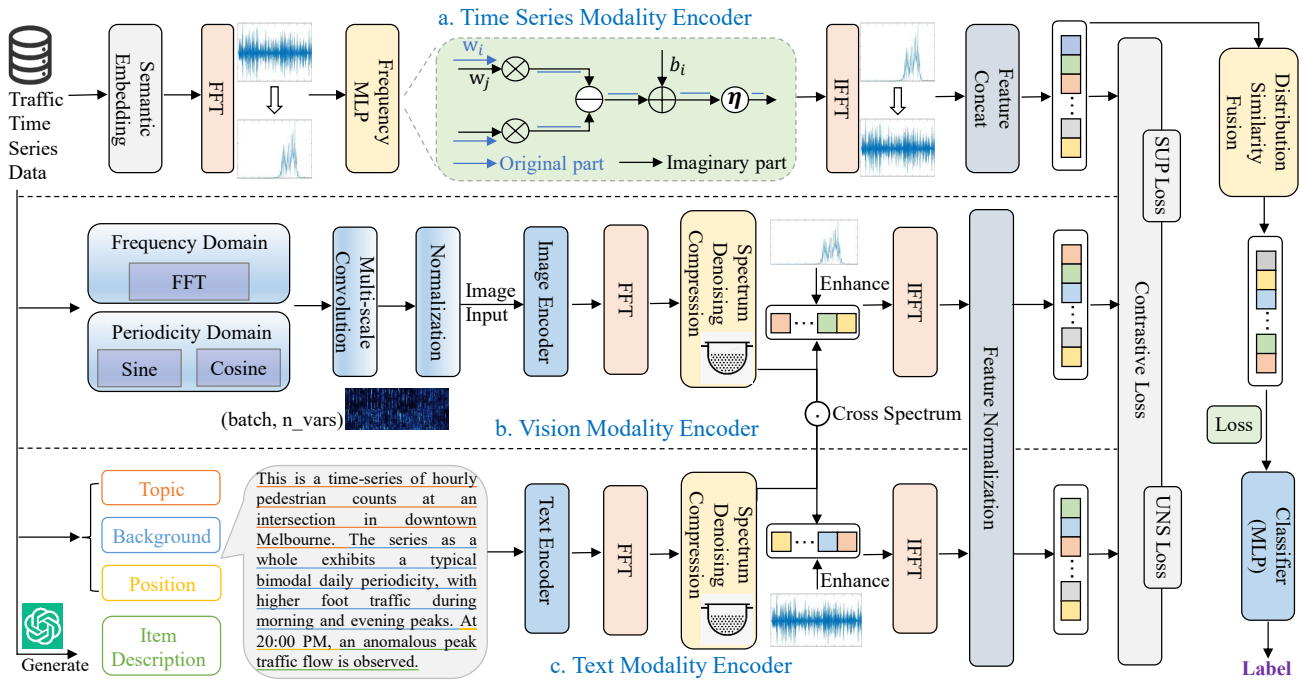


Figure 1: The overview of our framework. MTP learns multimodal features in the frequency domain from three perspectives: numerical, visual, and textual. These modalities are fused to provide more comprehensive features for urban traffic profiling.

of image generation is to convert the spatial-domain traffic time series data into visual images, achieving the transformation from numerical modality to visual modality. Specifically, we applied the FFT to extract frequency information from the input data as the frequency domain encoder. The extracted frequency representations are concatenated with the original input. Besides, we also design the periodicity domain encoder to extract the temporal dependencies. For each time stamp t, we use the following equation to get the new encoder:

$$\boldsymbol{P}_t = [\sin(2\pi t/\phi), \cos(2\pi t/\phi)], \tag{7}$$

where ϕ represents the periodicity hyperparameter. These encodings are also concatenated with the original input, which constitutes a group of new representations $\boldsymbol{X}^g$.

Next, we employ multi-scale convolution to extract hierarchical temporal patterns. Specifically, we first use a 1D convolutional layer to capture local dependencies. Among the subsequent two 2D convolutional layers, one halves the channel dimension, and another maps features to multiple output channels, thereby capturing global temporal structures. The output features are resized to the desired image dimensions via bilinear interpolation and then subjected to normalization.

After generating the image dimensions, we use the image encoder to obtain the numerical representations. These features are also converted into the frequency domain by FFT, which can be formulated as:

$$\boldsymbol{\mathcal{X}}^g[k] = \sum_{i=0}^{n-1} \boldsymbol{X}^g[i] e^{-j\frac{2\pi k i}{n}}. \tag{8}$$

In order to reduce the noise contained in the augmented images and focus on the core information, we introduced the finite impulse response (FIR) filter to process the features in the frequency domain. The FIR filter is constructed based on the "window" technique, specifically using the Hamming window. The characteristics of the Hamming window allow the filter to naturally aggregate the main information in traffic data while smoothing spectrum fluctuations. Given the filter length s, we can generate window function parameters through the Hamming window function by:

$$\omega_i = 0.54 - 0.46\cos(z\pi i/s - 1). \tag{9}$$

Then, we can get the actual impulse response $r^i = \omega[i] \cdot r'$ by multiplying the window function with the filter's ideal impulse response r'. These impulse responses form a filter bank $\boldsymbol{R} = [r^1, r^2, ..., r^s]$ with s filters. The filter bank can divide the input spectrum into multiple sub-bands. Through the impulse response r^i of each filter, key features within the corresponding frequency range are filtered out to avoid interference from irrelevant frequencies. Through the spectrum compression, we can calculate the spectrum by:

$$\boldsymbol{\mathcal{X}}^g_{spe} = \sum_{i=1}^{s} \frac{1}{c} |\boldsymbol{\mathcal{X}}^g|^2 \odot r^i, \tag{10}$$

where c represents the length of image modality, and $\odot$ represents the element-wise multiplication. Essentially, the filter is used to weighted the spectrum through this operation, which retains important frequency components and weakens redundant information, ultimately achieving efficient spectrum compression.

To address the limitation of fixed spectrum compression in being unable to remove high-frequency noise in traffic image processing, average pooling is introduced. It reduces high-frequency noise and random fluctuations by smoothing the spectrum, preserves the overall trend to make the spectrum more regular, and thereby improves the efficiency of compression algorithms and the effect of traffic image feature recognition. The average pooling process can be calculated by:

$$\boldsymbol{\mathcal{X}}^g_{pool} = Average(\boldsymbol{\mathcal{X}}^g_{spe} \odot \boldsymbol{\delta}^g), \tag{11}$$

where $\boldsymbol{\delta}^g$ represents a matrix that holds the corresponding dimension with $\boldsymbol{X}^g_{spe}$. In terms of cross-modal fusion, the enhanced spectrum of images is generated with the help of text modal information, and the formula is:

$$\boldsymbol{\mathcal{X}}^g_{out} = \boldsymbol{\mathcal{X}}^g_{spe} \odot \boldsymbol{\mathcal{X}}^t_{pool}, \tag{12}$$

where $\boldsymbol{X}^t_{pool}$ represents the output of text modality with pooling enhancement.

After pooling enhancement and spectrum cross in the frequency domain, we apply the IFFT to invert the features to the spatial representations. The IFFT of image features can be calculated by:

$$\boldsymbol{X}^g[i] = \sum_{i=0}^{n-1} \boldsymbol{\mathcal{X}}^g_{out}[k] e^{j\frac{2\pi ki}{n}}. \tag{13}$$

Text Modality Encoder

In this module, text can be pre-defined in the original traffic time series data or generated from the input data. If a generated textual description is required, we design text generation standards shown in "c" part of Figure 1. First, we can use LLMs (e.g., ChatGPT) to generate some item descriptions, which can enhance the semantic information for textual feature extraction. Then, more contextual information, such as topic, background, and vehicle position, can be extracted directly from the input data, thereby facilitating complete textual information for traffic profiling. If the input data already contains complete textual information, it can be directly fed into the text encoder to generate the vector features for subsequent processing.

Similar to the previous vision modality encoder, we use spectrum transformation technology to convert the vector generated by the text encoder into the frequency domain, followed by denoising and cross-modal spectrum processing with the image modality. Given the vector $\boldsymbol{X}^t$ generated by the text encoder, the representation in the frequency domain can be formulated as:

$$\boldsymbol{\mathcal{X}}^t[k] = \sum_{i=0}^{n-1} \boldsymbol{X}^t[i] e^{-j\frac{2\pi ki}{n}}. \tag{14}$$

Through FIR filter, average pooling, and cross-modal enhancement processing, we can calculate new spectral representations by:

$$\boldsymbol{\mathcal{X}}^t_{out} = \boldsymbol{\mathcal{X}}^t_{spe} \odot Average(\boldsymbol{\mathcal{X}}^g_{spe} \odot \boldsymbol{\delta}^g). \tag{15}$$

Finally, IFFT is applied to invert the frequency-domain features into the spatial-domain features $\boldsymbol{\mathcal{X}}^t_{out}$ for further cross-modal fusion.

Cross-modal Fusion

After each modality undergoes spectral transformation and frequency domain processing, the feature fusion is achieved through two schemes: contrastive learning and distribution similarity fusion.

Contrastive Learning. In our framework, the significance of contrastive loss lies in achieving semantic alignment of cross-modal features by reducing the distance between different modal features of the same traffic scene, while increasing the distance between irrelevant modal features, thereby enhancing the consistency of multi-modal features. For the labeled data, we can first conduct supervised learning to learn the supervised loss $\mathcal{L}(SUP)$. Given a data instance x_i, we can get the pairwise $(\boldsymbol{x'_i}, \boldsymbol{s_i})$ to calculate the supervised loss, where $\boldsymbol{x'_i}$ corresponds to the encoding feature and $\boldsymbol{s_i}$ corresponds to the real feature. Given a dataset with m categories, we can divide all instances into these m types $\boldsymbol{\mathcal{Y}} = \{\boldsymbol{\mathcal{M}}_1, \boldsymbol{\mathcal{M}}_2, ..., \boldsymbol{\mathcal{M}}_m\}$. For each instance, we can define a supervised loss as $\mathcal{L}_i(\boldsymbol{x^{prime}}, \boldsymbol{s_i})$ (Lin et al. 2022b). Next, the whole supervised loss is calculated by:

$$\begin{aligned}\mathcal{L}(SUP) = \sum_{\boldsymbol{X}} \sum_{\boldsymbol{\mathcal{Y}}} (\sum_{\boldsymbol{x'} \in \boldsymbol{\mathcal{M}}_i} \frac{1}{|\boldsymbol{\mathcal{M}}_i|} \sum_{\boldsymbol{s} \in \boldsymbol{\mathcal{M}}_i, \boldsymbol{x'} \neq \boldsymbol{s}} [\mathcal{L}_i(\boldsymbol{x'^v}, \boldsymbol{s}^v) \\ + \mathcal{L}_i(\boldsymbol{x'^g}, \boldsymbol{s}^g) + \mathcal{L}_i(\boldsymbol{x'^t}, \boldsymbol{s}^t)]).\end{aligned} \tag{16}$$

Unsupervised learning mainly captures the differences between modalities by aligning the features of different modalities. We introduce the InfoNCE loss (He et al. 2020) to calculate the similarity, which is defined as follows:

$$\begin{aligned}\mathcal{L}(UNS) = \frac{1}{3|\boldsymbol{X}|} \sum_{i=1}^{|\boldsymbol{X}|} [\mathcal{L}_v(\boldsymbol{x}^v_i, \boldsymbol{x}^g_i, \boldsymbol{x}^t_i) + \mathcal{L}_g(\boldsymbol{x}^g_i, \boldsymbol{x}^t_i, \boldsymbol{x}^v_i) \\ + \mathcal{L}_t(\boldsymbol{x}^t_i, \boldsymbol{x}^v_i, \boldsymbol{x}^g_i)].\end{aligned} \tag{17}$$

Fusion Loss. To ensure the semantic consistency of cross-modal features, we design a distribution similarity fusion scheme to assess the similarity between different modal features. Specifically, we apply Jensen-Shannon (JS) divergence between any two modalities to calculate the distribution similarity. Given a data instance x, its posterior probability in numerical modality can be defined as $\mathbb{I}(\boldsymbol{\alpha}^v|\boldsymbol{x}^v)$. After distribution similarity fusion, the JS divergence can be calculated by:

$$\begin{aligned}\Delta = (JS(\mathbb{I}(\boldsymbol{\alpha}^v|\boldsymbol{x}^v)||\mathbb{I}(\boldsymbol{\alpha}^g|\boldsymbol{x}^g)) + JS(\mathbb{I}(\boldsymbol{\alpha}^v|\boldsymbol{x}^v)||\mathbb{I}(\boldsymbol{\alpha}^t|\boldsymbol{x}^t)) \\ + JS(\mathbb{I}(\boldsymbol{\alpha}^g|\boldsymbol{x}^g)||\mathbb{I}(\boldsymbol{\alpha}^t|\boldsymbol{x}^t)))/3,\end{aligned} \tag{18}$$

Then, new features after distribution similarity fusion can be obtained through the similarity measure results, defined as: $\hat{\boldsymbol{x}} = (1-\Delta)(\boldsymbol{K}^v\boldsymbol{x}^v + \boldsymbol{K}^g\boldsymbol{x}^g + \boldsymbol{K}^t\boldsymbol{x}^t) + \Delta\boldsymbol{x}^v + \Delta\boldsymbol{x}^g + \Delta\boldsymbol{x}^t$ ($\boldsymbol{K}$ represents the training metric of a instance $\boldsymbol{x}$). Finally, we use the Multi-Layer Perceptron (MLP) classifier to predict the label of each data, which is realized by minimizing the fusion loss. Considering that urban traffic profiling is a multi-classification problem, we introduce multi-class cross-entropy loss to calculate the fusion loss, defined as:

$$\mathcal{L}(CE) = -\mathbb{E}_{y \sim \hat{Y}} \sum_{i=1}^{m} y_i log(y'_i), \tag{19}$$

Dataset	ShapeNet		TST		PatchTST		SVP-T		LightTS		ModernTCN		CAFO		InterpGN		**MTP**	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
Chinatown	0.7206	0.7259	0.9472	0.9563	0.9714	0.9767	0.9456	0.9592	0.9680	0.9708	0.9712	0.9767	0.9784	0.9825	0.9541	0.9659	**0.9820**	**0.9839**
Melbourne	0.7186	0.7314	0.8426	0.8421	0.8897	0.8877	0.8030	0.8065	0.8670	0.8655	0.8732	0.8786	0.8876	0.8860	0.8392	0.8364	**0.9669**	**0.9635**
PEMS-BAY	0.6365	0.6790	0.6712	0.6882	0.6838	0.6929	0.6573	0.6844	0.6736	0.6860	0.6950	0.7055	0.6637	0.6840	0.6770	0.6989	**0.7091**	**0.7200**
METR-LA	0.7186	0.7314	0.7143	0.7224	0.7295	0.7425	0.7158	0.7269	0.7113	0.7229	0.7483	0.7562	0.7158	0.7266	0.7262	0.7385	**0.7590**	**0.7684**
DodgerLoop	0.1500	0.2153	0.3529	0.4125	0.5435	0.5750	0.3817	0.4250	0.5156	0.5625	0.2442	0.3750	0.3607	0.4500	0.1519	0.2250	**0.5676**	**0.6000**
PEMS-SF	0.6373	0.6503	0.7900	0.7919	0.7468	0.7446	0.8215	**0.8266**	0.7384	0.7514	0.7594	0.7630	0.7857	0.7919	0.6246	0.6705	**0.8310**	0.8227

Table 1: Overall performance comparison on all datasets. Our proposed model (MTP) is compared with state-of-the-art baselines on metrics F1-score (F1) and accuracy (Acc). The best result is in bold, and the second-best is underlined.

where y_i is the real label and y_i' represents the probability that the prediction label belongs to category i.

The objective loss consists of two parts: the contrastive loss and the fusion loss. The full loss can be calculated by:

$$\mathcal{L} = \alpha\mathcal{L}(SUP) + \beta\mathcal{L}(UNS) + \gamma\mathcal{L}(CE), \qquad (20)$$

where α, β, and γ are hyperparameters for balancing the influence of different modules.

Experiments

To comprehensively evaluate the performance of our proposed MTP framework, we conduct extensive experiments on six public time series classification datasets. This section aims to answer the following core research questions (RQs):

- **RQ1:** How is the performance of MTP compared against state-of-the-art baselines?
- **RQ2:** What are the contributions of the core components within MTP to the final performance?
- **RQ3:** How are the multimodal features learned by our model distributed and separated in the feature space?

Experimental Setting

Baselines. Our framework is compared against 8 state-of-the-art time series models, including TST (Zerveas et al. 2021), ShapeNet (Cheng et al. 2021), PatchTST (Nie et al. 2023), SVP-T (Zuo et al. 2023), LightTS (Zhang, Chen, and He 2023), ModernTCN (Wang et al. 2024c), CAFO (Li, Wang, and Liu 2024), and InterpGN (Wen, Ma et al. 2025). These models cover a range of techniques from Transformer-based architectures to shapelet-based methods and pre-training frameworks. For detailed descriptions of these baselines, please refer to Appendix.

Datasets. Experiments are conducted on six widely-used public benchmarks for time series classification: Chinatown, MelbournePedestrian (Melbourne), PEMS-BAY, METR-LA, DodgerLoopDay (DodgerLoop), and PEMS-SF. Detailed descriptions of all datasets refer to Appendix.

Metrics. We adopt a set of classification metrics to measure the performance (Xu et al. 2025), including: Accuracy, Macro-Precision, Macro-Recall, and Macro F1-Score.

Implementation Details. All experiments are implemented using the PyTorch framework and conducted on a single NVIDIA RTX series GPU. To ensure a fair comparison, we follow the optimal parameter settings for each baseline.

Variant	Melbourne				DodgerLoop			
	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
MTP	**0.9672**	**0.9671**	**0.9669**	**0.9669**	**0.6000**	**0.6978**	**0.5923**	**0.5848**
w/o Visual	0.7593	0.7617	0.7595	0.7584	0.2375	0.0674	0.2348	0.1048
w/o Textual	0.9659	0.9660	0.9657	0.9657	0.5375	0.6268	0.5247	0.5315
w/o TS	0.6839	0.6845	0.6833	0.6766	0.5875	0.6015	0.5788	0.5676

Table 2: Ablation study on two representative datasets. Best results are in bold. MTP is the full model; w/o V: without Visual; w/o T: without Textual; w/o TS: without Timeseries.

The detailed parameter configurations for our framework are available in Appendix. All experiments are run 15 times, and we report the average results.

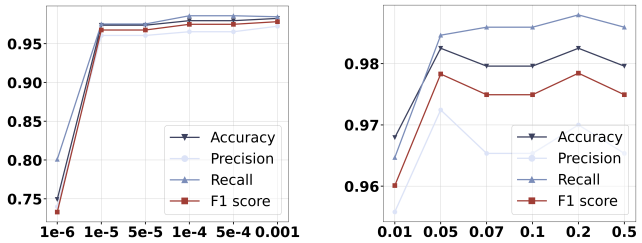
RQ1: Performance Comparison

We first compare the performance of our proposed framework against eight state-of-the-art baselines. The detailed experimental results are presented in Table 1. MTP consistently achieves state-of-the-art results, outperforming all baselines in the majority of cases. Specifically, MTP's strength is evident on datasets with clear, classifiable patterns. On the Melbourne dataset, it secures the best performance in both F1-score (0.9669) and Accuracy (0.9635). On the Chinatown dataset, our MTP framework achieves the highest F1-score of 0.9820 and a comparable Accuracy of 0.9839, while PatchTST achieves the second-best results.

MTP continues to excel in large-scale traffic datasets. It achieves the highest F1-score and Accuracy on PEMS-BAY (0.7091 / 0.7200) and METR-LA (0.7590 / 0.7684). Furthermore, on the highly volatile DodgerLoopDay dataset, MTP again ranks first with an F1-score of 0.5676 and an Accuracy of 0.6000, significantly outperforming most baselines. Even on PEMS-SF, MTP's performance is highly competitive, obtaining the second-best Accuracy. These results strongly validate that our framework can learn more comprehensive feature representations, leading to state-of-the-art performance across diverse tasks.

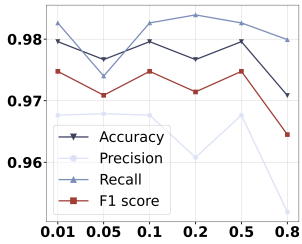
RQ2: Ablation Study

We conduct a rigorous ablation study of each component and a sensitivity analysis of key hyperparameters to show the impact of all design components on MTP.

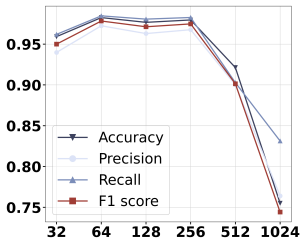


(a) Learning Rate

(b) Temperature

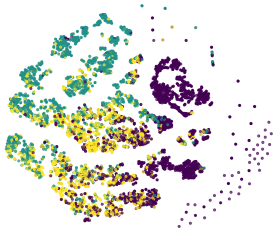


(c) Alpha weight



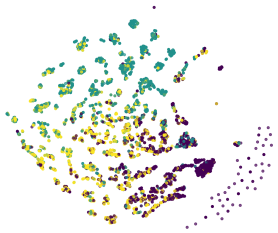
(d) Embedding dimension

Figure 2: Hyperparameter sensitivity analysis on four key parameters: (a) Learning Rate, (b) Temperature, (c) Alpha weight, and (d) Embedding dimension.



(a) Final fused features

(b) Image features



(c) Text features

(d) Time series features

Figure 3: Comparative t-SNE visualizations on the METR-LA dataset, which contains three types of labels.

Ablation Study. We remove all components, including removing the visual branch ('w/o V'), the textual branch ('w/o T'), and the time-series branch ('w/o TS'), from the full framework to evaluate the impact. The results on two representative datasets are shown in Table 2, with full results in the Appendix. Our complete framework consistently outperforms all of its ablated variants. The most pronounced impact is observed on the DodgerLoop dataset upon ablating the visual branch ('w/o V'), where the F1-score drops from 0.5848 to a mere 0.1048. Results on the Melbourne dataset further underscore the contribution of each modality: removing the textual branch ('w/o T') induces a measurable decline in F1-score (from 0.9669 to 0.9657), whereas exclusive reliance on visual and textual modalities without the original time-series ('w/o TS') leads to a substantial performance degradation (F1-score of 0.6766). Thus, it is approved that our proposed modality augmentation and fusion are the core drivers of the model's superior performance.

Hyperparameter Sensitivity Analysis. We investigate the sensitivity of MTP to four key hyperparameters, shown in Figure 2. The results indicate that our framework exhibits robustness across different hyperparameters. For the learning rate, performance reaches its peak around $1e-4$. Regarding the temperature parameter, the optimal range is between 0.05 and 0.1. MTP demonstrates a distinct preference for a smaller alpha value in contrastive loss, with the best performance achieved at 0.1. Finally, for the embedding dimension, performance plateaus once the dimension reaches 128. These findings collectively validate that MTP is stable and does not necessitate extensive hyperparameter tuning.

RQ3: Qualitative Analysis

To intuitively understand the effectiveness of our framework, we use t-SNE to visualize the feature distributions on the METR-LA test set, as shown in Figure 3. The final fused features learned by our complete framework form highly cohesive and clearly separated clusters in the 2D space. Samples from different classes represented by different colors are distinctly separated with minimal overlap. This provides strong visual evidence for the high classification performance reported in Table 1. In contrast, the feature distributions from single modalities are more diffuse and intermingled. This qualitative result is consistent with our conclusions from the ablation study, demonstrating that our fusion module successfully integrates complementary information to produce a more powerful and discriminative To facilitate a comprehensive understanding of our method's performance across multiple datasets, additional visualization analyses are provided in the Appendix.

Conclusion

In this paper, we propose a novel multi-modal urban traffic profiling framework, called MTP, which addresses the issue that existing traffic profiling methods rely on a single numerical modality and overlook the semantic information in multi-modal heterogeneous data. MTP learns multimodal features in the frequency domain from three perspectives: numerical, visual, and textual. Specifically, MTP processes numerical information using frequency multi-layer perceptions, performs visual augmentation by converting raw data into periodic images and frequency images, and generates descriptive texts based on information such as topics and backgrounds for textual augmentation. Besides, MTP designs hierarchical contrastive learning to fuse the three modalities. Experiments on six real-world datasets demonstrate that our framework significantly outperforms state-of-the-art methods. Future work will involve integrating more types of urban modal data and exploring fine-grained modeling mechanisms for cross-modal correlations.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62372242, Jiangsu Provincial Major Project on Basic Research of Cutting-edge and Leading Technologies, under grant no. BK20232032, and Open Research Projects of the State Key Laboratory for New Technologies of Computer Software, Nanjing University, under grant no. KFKT2025B64.

References

Abdelrahman, A. E.; Hassanein, H. S.; and Abu-Ali, N. 2020. Robust data-driven framework for driver behavior profiling using supervised machine learning. *IEEE Transactions on Intelligent Transportation Systems*, 23(4): 3336–3350.

Alam, M. G. R.; Haque, M.; Hassan, M. R.; Huda, S.; Hassan, M. M.; Strickland, F. L.; and AlQahtani, S. A. 2023. Feature cloning and feature fusion based transportation mode detection using convolutional neural network. *IEEE Transactions on Intelligent Transportation Systems*, 24(4): 4671–4681.

Chen, R.; Song, W.; Zu, W.; Dong, Z.; Guo, Z.; Sun, F.; Tian, Z.; and Wang, J. 2024. An LLM-driven framework for multiple-vehicle dispatching and navigation in smart city landscapes. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2147–2153.

Cheng, Z.; Yang, Y.; Wang, W.; Hu, W.; Zhuang, Y.; and Song, G. 2021. ShapeNet: A shapelet-neural network for multivariate time series classification. In *Proceedings of the ACM International Conference on Information & Knowledge Management (CIKM)*.

Cura, A.; Küçük, H.; Ergen, E.; and Öksüzoğlu, İ. B. 2020. Driver profiling using long short term memory (LSTM) and convolutional neural network (CNN) methods. *IEEE Transactions on Intelligent Transportation Systems*, 22(10): 6572–6582.

Deng, X.; Wang, Y.; and Xue, Z. 2024. AN-NET: an anti-noise network for anonymous traffic classification. In *Proceedings of the ACM Web Conference (WWW)*, 4417–4428.

Fan, W.; Wang, P.; Wang, D.; Wang, D.; Zhou, Y.; and Fu, Y. 2023. Dish-ts: a general paradigm for alleviating distribution shift in time series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 7522–7529.

Fan, W.; Zheng, S.; Yi, X.; Cao, W.; Fu, Y.; Bian, J.; and Liu, T.-Y. 2022. DEPTS: Deep Expansion Learning for Periodic Time Series Forecasting. In *International Conference on Learning Representations*.

Fang, Y.; Liang, Y.; Hui, B.; Shao, Z.; Deng, L.; Liu, X.; Jiang, X.; and Zheng, K. 2025. Efficient large-scale traffic forecasting with transformers: A spatial data management perspective. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 307–317.

Feng, Y.; Tian, Z.; Zhu, Y.; Han, Z.; Luo, H.; Zhang, G.; and Song, M. 2024. CP-Prompt: Composition-based cross-modal prompting for domain-incremental continual learning. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2729–2738.

Gruver, N.; Finzi, M.; Qiu, S.; and Wilson, A. G. 2023. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems (NeurIPS)*, 36: 19622–19635.

Han, C.; Wang, J.; Wang, Y.; Yu, X.; Lin, H.; Li, C.; and Wu, J. 2025. Bridging traffic state and trajectory for dynamic road network and trajectory representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, 11763–11771.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hu, W.; Xu, Y.; Li, Y.; Li, W.; Chen, Z.; and Tu, Z. 2024. BLIVA: A simple multi-modal LLM for better handling of text-rich visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 2256–2264.

Jin, Z.; Cao, J.; Guo, H.; Zhang, Y.; and Luo, J. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 795–816.

Khattar, D.; Goud, J. S.; Gupta, M.; and Varma, V. 2019. MVAE: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference (WWW)*, 2915–2921.

Li, J. N.; Xu, Y.; Grossman, T.; Santosa, S.; and Li, M. 2024. OmniActions: Predicting digital actions in response to real-world multi-modal sensory inputs with LLMs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI)*, 1–22.

Li, Y.; Wang, S.; and Liu, Z. 2024. Convolutional attention for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Lin, X.; Xiong, G.; Gou, G.; Li, Z.; Shi, J.; and Yu, J. 2022a. ET-BERT: A contextualized datagram representation with pre-training transformers for encrypted traffic classification. In *Proceedings of the ACM Web Conference (WWW)*, 633–642.

Lin, Z.; Liang, B.; Long, Y.; Dang, Y.; Yang, M.; Zhang, M.; and Xu, R. 2022b. Modeling intra-and inter-modal relations: Hierarchical graph contrastive learning for multimodal sentiment analysis. In *Proceedings of the International Conference on Computational Linguistics (ACL)*, volume 29, 7124–7135.

Liu, Z.; Zheng, G.; and Yu, Y. 2025. Multi-scale traffic pattern bank for cross-city few-shot traffic forecasting. *ACM Transactions on Knowledge Discovery from Data*, 19(4): 1–24.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *International Conference on Learning Representations (ICLR Workshop)*.

Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2023. A time series is worth 64 words: long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR)*.

Qian, S.; Wang, J.; Hu, J.; Fang, Q.; and Xu, C. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR)*, 153–162.

Wang, C.; Finamore, A.; Michiardi, P.; Gallo, M.; and Rossi, D. 2024a. Data augmentation for traffic classification. In *International Conference on Passive and Active Network Measurement*, 159–186. Springer.

Wang, Y.; Wu, H.; Dong, J.; Liu, Y.; Long, M.; and Wang, J. 2024b. Deep time series models: A comprehensive survey and benchmark. *arXiv preprint arXiv:2407.13278*.

Wang, Z.; Zhang, Y.; Li, J.; and Long, M. 2024c. ModernTCN: A modernized and scalable time-convolutional network. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Wen, Y.; Ma, T.; et al. 2025. Shedding light on time series classification using interpretability gated networks. In *International Conference on Learning Representations (ICLR)*.

Wu, K.; Ding, J.; Lin, J.; Zheng, G.; Sun, Y.; Fang, J.; Xu, T.; Zhu, Y.; and Gu, B. 2025. Big-data empowered traffic signal control could reduce urban carbon emission. *Nature Communications*, 16(1): 2013.

Xiang, H.; Xu, X.; Wang, G.; Zhang, X.; Li, X.; Zhang, Q.; Beheshti, A.; and Fan, W. 2025. Empowering multimodal road traffic profiling with vision language models and frequency spectrum fusion. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Xu, X.; Zhou, Y.; Xiang, H.; Li, X.; Zhang, X.; Qi, L.; and Dou, W. 2025. NLGT: Neighborhood-based and Label-enhanced Graph Transformer Framework for Node Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, 12954–12962.

Yan, Y.; Wen, H.; Zhong, S.; Chen, W.; Chen, H.; Wen, Q.; Zimmermann, R.; and Liang, Y. 2024. UrbanCLIP: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM Web Conference (WWW)*, 4006–4017.

Yang, Y.; Zhou, T.; Li, K.; Tao, D.; Li, L.; Shen, L.; He, X.; Jiang, J.; and Shi, Y. 2024. Embodied multi-modal agent trained by an LLM from a parallel textworld. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 26275–26285.

Yi, K.; Zhang, Q.; Fan, W.; Wang, S.; Wang, P.; He, H.; An, N.; Lian, D.; Cao, L.; and Niu, Z. 2023. Frequency-domain MLPs are more effective learners in time series forecasting. *Advances in Neural Information Processing Systems (NeurIPS)*, 36: 76656–76679.

Zerveas, G.; Jayaraman, S.; Patel, D.; Bhamidipaty, A.; and Eickhoff, C. 2021. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery & Data Mining (SIGKDD)*.

Zhang, H.; Yu, L.; Xiao, X.; Li, Q.; Mercaldo, F.; Luo, X.; and Liu, Q. 2023. TFE-GNN: A temporal fusion encoder using graph neural networks for fine-grained encrypted traffic classification. In *Proceedings of the ACM Web Conference (WWW)*, 2066–2075.

Zhang, S.-A.; Chen, D.; and He, Y. 2023. LightTS: A lightweight framework for time series forecasting. *ACM Transactions on Knowledge Discovery from Data*.

Zheng, W.; Gou, C.; Yan, L.; and Mo, S. 2020. Learning to classify: A flow-based relation network for encrypted traffic classification. In *Proceedings of The Web Conference (WWW)*, 13–22.

Zhou, G.; Guo, X.; Liu, Z.; Li, T.; Li, Q.; and Xu, K. 2025. Trafficformer: an efficient pre-trained model for traffic data. In *2025 IEEE Symposium on Security and Privacy (S&P)*, 1844–1860. IEEE.

Zuo, R.; Li, G.; Choi, B.; Bhowmick, S. S.; Mah, D. N.-y.; and Wong, G. L. 2023. SVP-T: A shape-level variable-position transformer for multivariate time series classification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, 11497–11505.