

Towards Cross-Modality Modeling for Time Series Analytics: A Survey in the LLM Era

Chenxi Liu¹, Shaowen Zhou¹, Qianxiong Xu^{1,*}, Hao Miao², Cheng Long^{1,*}, Ziyue Li^{3,*},
Rui Zhao⁴

¹S-Lab, Nanyang Technological University, Singapore

²Aalborg University, Denmark

³Technical University of Munich, Germany

⁴SenseTime Research, China

{chenxi.liu, qianxiong.xu, c.long}@ntu.edu.sg, s200061@e.ntu.edu.sg
haom@cs.aau.dk, zlibn@connect.ust.hk, zhaorui@sensetime.com

Abstract

The proliferation of edge devices has generated an unprecedented volume of time series data across different domains, motivating various well-customized methods. Recently, Large Language Models (LLMs) have emerged as a new paradigm for time series analytics by leveraging the shared sequential nature of textual data and time series. However, a fundamental cross-modality gap between time series and LLMs exists, as LLMs are pre-trained on textual corpora and are not inherently optimized for time series. Many recent proposals are designed to address this issue. In this survey, we provide an up-to-date overview of LLM-based cross-modality modeling for time series analytics. We first introduce a taxonomy that classifies existing approaches into four groups based on the type of textual data employed for time series modeling. We then summarize key cross-modality strategies, e.g., alignment and fusion, and discuss their applications across a range of downstream tasks. Furthermore, we conduct experiments on multimodal datasets from different application domains to investigate effective combinations of textual data and cross-modality strategies for enhancing time series analytics. Finally, we suggest several promising directions for future research. This survey is designed for a range of professionals, researchers, and practitioners interested in LLM-based time series modeling.

1 Introduction

With the proliferation of edge devices and the development of mobile sensing techniques, a large amount of time series data has been generated, enabling a variety of real-world applications [Liu *et al.*, 2025d; Pettersson *et al.*, 2023; Liu *et al.*, 2024b; Cai *et al.*, 2024; Liu *et al.*, 2021b]. Time series data typically take the format of sequential observations

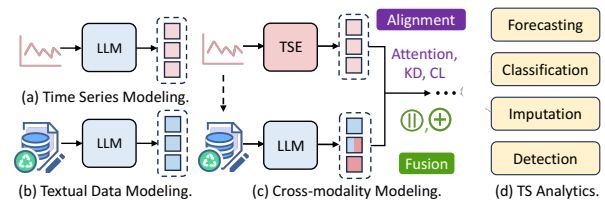


Figure 1: Cross-Modality Modeling for Time Series Analytics.

with varying features [Liu *et al.*, 2025c; Alnegheimish *et al.*, 2024; Liu *et al.*, 2024i; Liu *et al.*, 2022]. Considerable research efforts have been made to design time series modeling and analytics methods, which enables different downstream tasks, such as time series forecasting [Liu *et al.*, 2021c; Chen *et al.*, 2020; Jin *et al.*, 2022], imputation [Chen *et al.*, 2024; Xiao *et al.*, 2022], classification [Liu *et al.*, 2024d; Liu *et al.*, 2021a], and anomaly detection [Xu *et al.*, 2024].

Recently, large language model (LLM)-based methods [Touvron *et al.*, 2023; Radford *et al.*, 2019] have emerged as a new paradigm for time series modeling. These methods are inspired by time series and natural text exhibit similar formats (i.e., sequence) [Liu *et al.*, 2025a], and assume that the generic knowledge learned by LLMs can be easily transferred to time series [Xue and Salim, 2023]. Although existing surveys have introduced broad overviews of LLM-based time series methods [Jin *et al.*, 2024b; Zhang *et al.*, 2024; Jiang *et al.*, 2024], they overlook the critical challenge posed by the *cross-modality gap* [Liu *et al.*, 2024d] between time series and textual data. To be specific, LLMs are pre-trained on textual corpora and are not inherently designed for time series, there is a pressing need to develop cross-modality modeling strategies that effectively integrate textual knowledge into time series analytics.

This survey makes a unique contribution to the existing literature by addressing the cross-modality gap between time series and textual data, thereby enhancing LLM-based time series analytics. Figure 1 shows a general framework for LLM-based time series modeling. In this paper, we divide textual data into four types: *numerical prompt*, *statistical prompt*, *contextual prompt*, and *word token embedding*. To

*Corresponding author

contend with the cross-modality modeling, we summarize two overarching strategies according to recent studies [Jin *et al.*, 2024a; Liu *et al.*, 2025e], i.e., alignment and fusion, to integrate time series with different textual data. For alignment, we identify four key methods: *unidirectional retrieval*, *bidirectional retrieval*, *contrastive learning*, and *knowledge distillation*. In addition, the fusion strategy primarily relies on *concatenation* and/or *addition* to integrate textual information into time series embeddings.

Furthermore, this taxonomy covers forecasting, classification, anomaly detection, as well as multiple combinations across diverse application domains, including general analysis, healthcare, traffic, and event-oriented scenarios. Finally, we conduct experimental evaluations on multi-domain multimodal datasets to assess the effective combinations of textual data and cross-modality strategies for effective time series forecasting, providing practical insights for future research.

The major contributions are summarized as follows.

- We present a comprehensive catalog of literature on LLM-based cross-modality modeling for time series analytics, highlighting recent representative methods.
- We propose a taxonomy that classifies related studies into four groups based on the type of textual data. Additionally, we explore cross-modality modeling strategies, including alignment and fusion, and discuss their applications across various tasks and domains.
- We perform experimental evaluations on multi-domain multimodal datasets to explore effective combinations of additional textual data and strategies that facilitate time series analytics.

2 Formulation

2.1 Definitions

Time Series. We define a time series as an ordered sequence, denoted by $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_S\} \in \mathbb{R}^{S \times N}$, where S represents the sequence length, and N is the number of variables. Each observation \mathbf{x}_i is an N -dimensional vector at time step i . The scalar v_i refers to the numerical value of a specific variable in the time series at time step i .

Textual Data. The textual data $\mathbf{T} = \{\mathbf{P}, \mathbf{W}\}$ in time series modeling can be categorized into four types: numerical prompt [Gruver *et al.*, 2023], statistical prompt [Liu *et al.*, 2024c], contextual prompt [Liu *et al.*, 2024f], and word token embedding [Pan *et al.*, 2024]. Some studies [Liu *et al.*, 2025e; Jin *et al.*, 2024a] utilize a combination of prompts, denoted as $\mathbf{P} = \{\mathbf{P}_N, \mathbf{P}_S, \mathbf{P}_C\}$, while others directly adopt word token embeddings \mathbf{W} [Pan *et al.*, 2024; Liu *et al.*, 2024e] extracted from LLMs. In this survey, we unify the definitions of these textual data types as follows:

- **Numerical Prompt** transforms the numerical data of \mathbf{X} into a textual format, denoted as \mathbf{P}_N . Each prompt consists of M words, primarily representing the numerical values of the time series.
- **Statistical Prompt** encodes statistical features of the time series, such as mean, maximum, minimum, median, top- k , or trend values. These statistics are typically expressed in textual format and denoted as \mathbf{P}_S .

- **Contextual Prompt** provides auxiliary descriptions, including dataset metadata, media news, or event-related information. We denote contextual instructions as \mathbf{P}_C .
- **Word Token Embedding** refers to the pre-trained weights within LLMs. Instead of using textual prompts, the textual representations can be directly captured from the word token embeddings, denoted as \mathbf{W} .

2.2 Cross-Modality Modeling for Time Series Analytics

Given time series \mathbf{X} and textual data \mathbf{T} , cross-modality modeling aims to learn a function that integrates \mathbf{X} with \mathbf{T} to generate the target output \mathbf{Y} for downstream tasks, such as long-term forecasting, short-term forecasting, classification, imputation, and anomaly detection. Formally, the objective is to learn a mapping function:

$$f : (\mathbf{X}, \mathbf{T}) \rightarrow \mathbf{Y}, \quad (1)$$

where $f(\cdot)$ is the method that aligns and fuses both modalities to enhance time series modeling.

3 Cross-Modality Alignment

We highlight three widely adopted alignment methods: retrieval, contrastive learning, and knowledge distillation.

3.1 Retrieval

Retrieval is the method of leveraging data from one modality to access relevant information in another. Based on the retrieval direction, we categorize it into two types: *unidirectional retrieval*, where information flows from one modality to another, and *bidirectional retrieval*, where both modalities can retrieve information from each other.

Unidirectional Retrieval

This method has been applied to forecasting tasks across general domains. For example, TimeCMA [Liu *et al.*, 2025e] introduces hybrid prompts that integrate numerical, statistical, and contextual information to improve time series forecasting. These hybrid prompts are processed by an LLM to generate prompt embeddings, which are then aligned with the original time series through unidirectional similarity-based retrieval. This retrieval leverages time series embeddings to extract disentangled and robust representations from the LLM-empowered prompt embeddings.

Similarly, Time-LLM [Jin *et al.*, 2024a], Time-FFM [Liu *et al.*, 2024e], S^2 IP-LLM [Pan *et al.*, 2024], and CALF [Liu *et al.*, 2024d] employ unidirectional retrieval by aligning time series embeddings with word token embeddings in pre-trained LLMs, using the former as queries. In contrast, TEMPO [Cao *et al.*, 2024] takes an inverse approach, utilizing prompt embeddings as queries to retrieve the top- K corresponding values from the patched time series input.

Overall, the unidirectional retrieval method typically involves using the time series embedding \mathbf{E}_X to retrieve relevant information from the LLM-enhanced textual embedding \mathbf{E}_T , implemented via cross-attention:

$$\mathbf{E}_A = \text{CrossAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}), \quad (2)$$

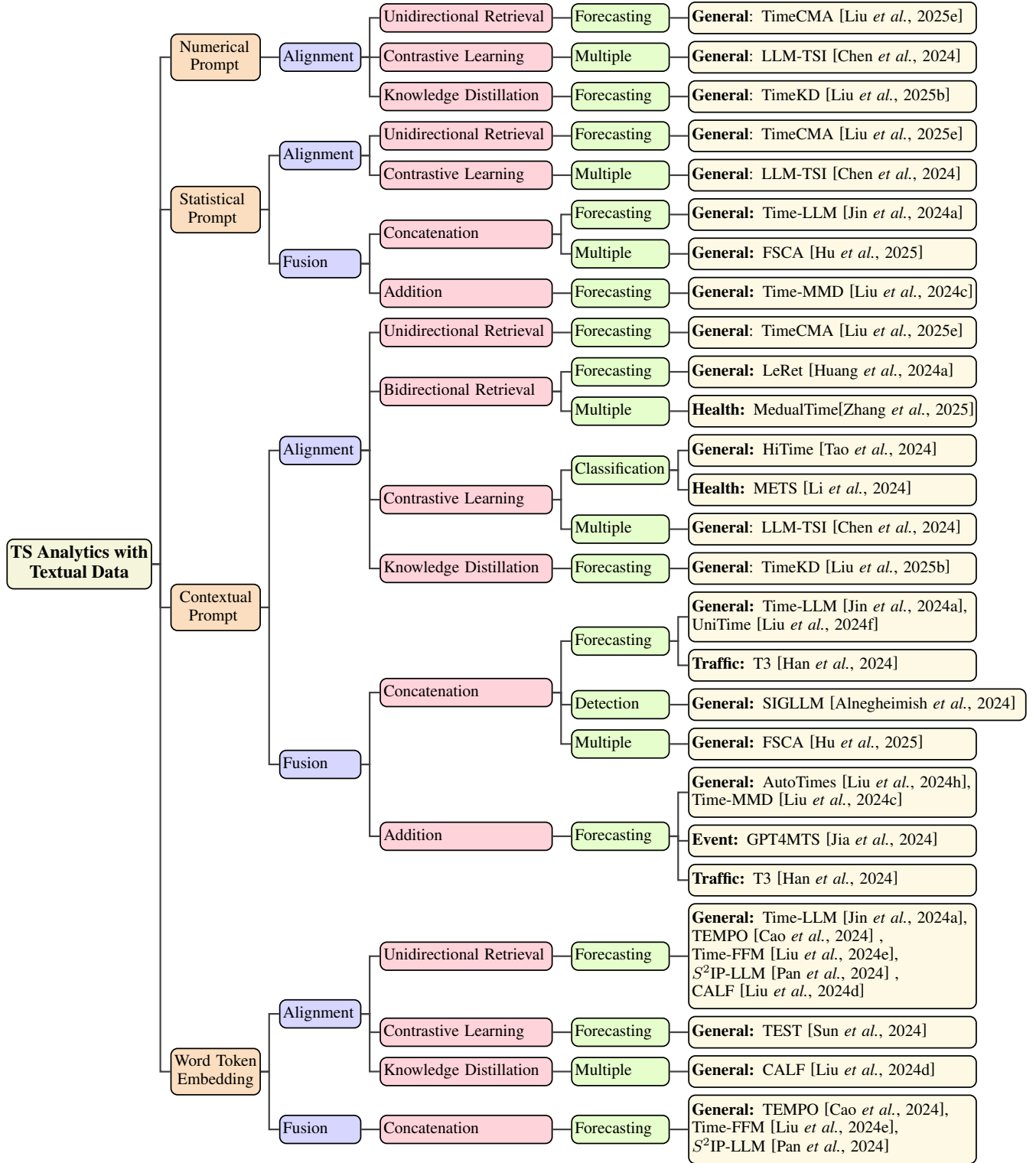


Figure 2: Taxonomy of cross-modality modeling for time series (TS) analytics incorporating textual data, including numerical prompt, statistical prompt, contextual prompt, and word token embedding. The textual data is processed by LLMs.

where \mathbf{E}_A is the aligned time series embedding.

$$\mathbf{Q} = \mathbf{E}_X \mathbf{W}_Q, \quad \mathbf{K} = \mathbf{E}_T \mathbf{W}_K, \quad \mathbf{V} = \mathbf{E}_T \mathbf{W}_V. \quad (3)$$

Conversely, if the prompt embeddings act as the query:

$$\mathbf{Q} = \mathbf{E}_T \mathbf{W}_Q, \quad \mathbf{K} = \mathbf{E}_X \mathbf{W}_K, \quad \mathbf{V} = \mathbf{E}_X \mathbf{W}_V, \quad (4)$$

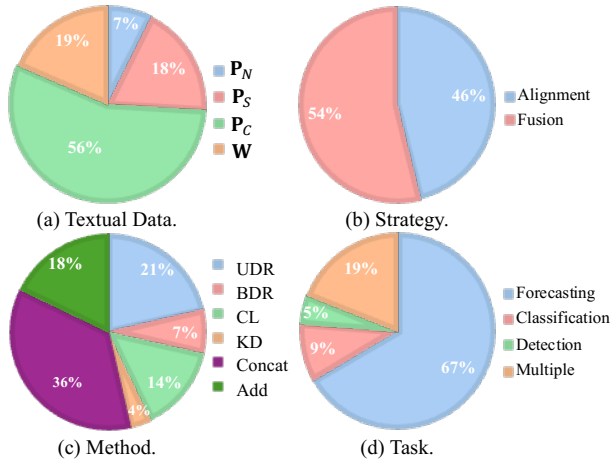


Figure 3: Distribution of taxonomies in cross-modality time series modeling. (a) Textual data types: numerical prompts P_N , statistical prompts P_S , contextual prompts P_C , and word token embeddings W . (b) Strategy: Alignment vs. Fusion. (c) Method categories: unidirectional retrieval (UDR), bidirectional retrieval (BDR), contrastive learning (CL), knowledge distillation (KD), concatenation (Concat), and addition (Add). (d) Task types: forecasting, classification, anomaly detection, and multiple tasks.

Here, W_Q, W_K, W_V are learnable projection matrices that transform the features into query, key, and value spaces.

Bidirectional Retrieval

This method extends unidirectional retrieval by allowing both time series and textual embeddings to retrieve information from each other. For instance, LeRet [Huang *et al.*, 2024a] introduces a bidirectional retrieval method for time series forecasting. Instead of relying solely on time series embeddings as queries, LeRet allows textual embeddings to retrieve relevant time series features, creating a dynamic exchange between the two modalities. This bidirectional retrieval strategy improves forecasting accuracy by leveraging the strengths of both data sources.

In the healthcare domain, MedualTime [Zhang *et al.*, 2025] employs bidirectional retrieval for forecasting and classification, integrating textual and time series embeddings to enhance predictive modeling in clinical applications. Formally, bidirectional retrieval can be expressed as an extension of unidirectional retrieval, where either modality can act as the query. For example, textual knowledge is mapped to the time series feature space in the first stage:

$$\mathbf{E}'_T = \text{CrossAttention}(\mathbf{E}_T, \mathbf{E}_X, \mathbf{E}_X). \quad (5)$$

Second stage aims to integrate this aligned textual knowledge with time series features:

$$\mathbf{E}_A = \text{CrossAttention}(\mathbf{E}_X, \mathbf{E}'_T, \mathbf{E}'_T), \quad (6)$$

where \mathbf{E}'_T denotes the aligned textual embeddings.

3.2 Contrastive Learning

Contrastive learning aims to establish a shared representation by maximizing the agreement between corresponding time

series and textual embeddings while minimizing the similarity between non-corresponding pairs [Özyurt *et al.*, 2023]. For example, Chen *et al.* [Chen *et al.*, 2024] propose a contrastive module to align time series and textual prompts by maximizing the mutual information between a small model's time series representation and LLM's textual representation.

Similarly, METS [Li *et al.*, 2024] utilizes the auto-generated clinical reports to guide electrocardiogram (ECG) self-supervised pre-training. The contrastive strategy aims to maximize the similarity between paired and report while minimizing the similarity between ECG and other reports. TEST [Sun *et al.*, 2024] builds an encoder to embed TS via instance-wise, feature-wise, and text-prototype-aligned contrast, where the TS embedding space is aligned to LLM's embedding layer space.

Formally, contrastive learning for cross-modality alignment can be defined as follows. Given a time series embeddings \mathbf{E}_X and textual embeddings \mathbf{E}_T , the contrastive loss function is formulated as:

$$\mathcal{L}_{\text{contrast}} = -\log \frac{\exp(\text{sim}(\mathbf{E}_X, \mathbf{E}_T)/\tau)}{\sum_{\bar{\mathbf{E}}_T \in \mathcal{N}} \exp(\text{sim}(\mathbf{E}_X, \bar{\mathbf{E}}_T)/\tau)}, \quad (7)$$

where $\text{sim}(\cdot, \cdot)$ denotes the similarity function (e.g., cosine similarity), τ is a temperature hyperparameter, and \mathcal{N} represents a set of negative samples (i.e., unrelated textual embeddings). The objective is to maximize the similarity between aligned pairs $(\mathbf{E}_X, \mathbf{E}_T)$ while minimizing the similarity between mismatched pairs $(\mathbf{E}_X, \bar{\mathbf{E}}_T)$.

3.3 Knowledge Distillation

The LLM-based knowledge distillation (KD) achieves a small student model from an LLM, enabling efficient inference solely on the distilled student model. Recent works have been proposed to address the cross-modal misalignment problem with knowledge distillation [Liu and Zhang, 2025], which can generally be categorized into black-box distillation [Liu *et al.*, 2024a] and white-box distillation [Liu *et al.*, 2025b] based on the accessibility of the teacher model's internal information during the distillation process.

CALF [Liu *et al.*, 2024d] is a black-box KD method that aligns LLMs for time series forecasting via cross-modal fine-tuning. To adapt the word token embeddings to time series data, they align the outputs of each intermediate layer l in time series-based LLM with those of the textual LLM, also align the output consistency between these two modalities to maintain a coherent semantic representation:

$$\begin{aligned} \mathcal{L}_{\text{feature}} &= \sum_{i=1}^L \gamma^{(L-i)} \text{sim}(\mathbf{F}_X^l, \mathbf{F}_T^l), \\ \mathcal{L}_{\text{output}} &= \text{sim}(\mathbf{E}_X, \mathbf{E}_T), \end{aligned} \quad (8)$$

where \mathbf{F}_X^l and \mathbf{F}_T^l are the outputs of the l -th Transformer block in time series-based and textual LLMs, respectively. L is the total number of layers in the LLM. γ is the hyperparameter that controls the loss scale from different layers.

In contrast, TimeKD [Liu *et al.*, 2025b] is a white-box KD method that benefits from the design of privileged correlation distillation, the student model explicitly aligns its internal attention maps with those of the teacher model to mimic their behavior.

4 Cross-Modality Fusion

Fusion strategy allows models to leverage complementary information from different modalities [Zhao *et al.*, 2024], enhancing their ability to capture richer contextual dependencies. We summarize two common fusion methods: concatenation and addition of embeddings. Unlike alignment strategies, fusion-based methods often introduce data entanglement issues [Liu *et al.*, 2025e], which may lead to suboptimal performance compared to alignment-based methods.

4.1 Addition-based Fusion

Addition-based fusion integrates textual embeddings with time series representations by summing their feature vectors. This method allows models to incorporate textual information without significantly increasing the dimensionality of the feature space, making it a computationally efficient alternative to concatenation. Unlike concatenation, addition-based fusion maintains a compact representation, ensuring that the model does not introduce unnecessary complexity while still leveraging multimodal information.

Several studies have adopted addition-based fusion for time series analysis. Time-MMD [Liu *et al.*, 2024c], GPT4MTS [Jia *et al.*, 2024], AutoTimes [Liu *et al.*, 2024h], and T3 [Han *et al.*, 2024] add the textual embedding with time series embedding for time series analysis, respectively. Formally, the addition can be expressed as follows:

$$\mathbf{E}_F = \mathbf{E}_X + \mathbf{E}_T, \quad (9)$$

where \mathbf{E}_F is the fused embeddings and $+$ denotes addition.

4.2 Concatenation-based Fusion

Concatenation-based fusion directly merges textual embeddings with time series features to create a joint representation. This method enables models to incorporate textual information alongside time series data, allowing for a more comprehensive feature space. While concatenation provides a straightforward way to multimodal integration, it can increase the dimensionality of the feature space, leading to greater computational complexity. Moreover, the lack of explicit alignment mechanisms between modalities may introduce noise, reducing the effectiveness of downstream tasks.

Some studies directly concatenate time series and textual embeddings. For instance, UniTime [Liu *et al.*, 2024f] concatenates the contextual prompt embedding with time series embedding to retain a LLM-based unified model for cross-domain time series forecasting. SIGLLM [Alnegheimish *et al.*, 2024] concatenates the contextual prompt embedding with time series embedding for the zero-shot anomaly detection task. TEMPO [Cao *et al.*, 2024] concatenates the word token embedding with different time series features, such as trend, seasonal, and residual, for time series forecasting. Time-FFM [Liu *et al.*, 2024e] concatenates word token embedding with time series embedding for time series forecasting. The concatenation can be formulated as follows:

$$\mathbf{E}_F = \mathbf{E}_X \parallel \mathbf{E}_T. \quad (10)$$

where \parallel denotes concatenation.

Other works utilize multiple strategies to integrate data embeddings. Beyond retrieval-based alignment, Time-LLM [Jin

Domain	Dim	Frequency	Samples	Timespan
Agriculture	1	Monthly	496	1980 - 2024
Climate	5	Monthly	496	2000 - 2024
Economy	3	Monthly	423	1987 - 2024
Energy	9	Weekly	1479	1993 - 2024
Health	11	Weekly	1389	2002 - 2024

Table 1: Overview of datasets.

et al., 2024a] further enhances the adaptability of LLMs for time series forecasting by concatenating the textual prompt embedding as a prefix to the general time series embedding. S^2 IP-LLM [Pan *et al.*, 2024] concatenates time series embedding and retrieved embedding to avoid the data entanglement issue. FCSA [Hu *et al.*, 2025] concatenates the time series embeddings and textual prompt embeddings to extract fine-grained features for further alignment. After addition-based fusion, T3 [Han *et al.*, 2024] concatenates the contextual prompt and traffic data embeddings to maximize the utilization of the training data for the traffic forecasting task.

5 Experiments

We perform extensive experimental evaluations on multi-domain, multimodal datasets. We employ four types of textual data in Figure 3 (a) as well as cross-modality alignment and fusion strategies in Figure 3 (b). Specifically, we select the three most common methods from the literature: unidirectional retrieval-based alignment (21%), concatenation-based fusion (36%), and addition-based fusion (18%) in Figure 3 (c). We focus on the time series forecasting task, accounting for 67% of reported tasks across multiple domains in Figure 3 (d). We also implement a single-modality model without textual inputs. Our code and datasets are available¹.

5.1 Dataset Description

We utilize datasets from five domains [Liu *et al.*, 2024c], spanning agriculture, climate, economy, energy, and health. Each dataset consists of time series with relevant textual data. The textual data includes expert reports and news summaries, each annotated with timestamps corresponding to the periods they describe.

Time Series Data is summarized in Table 1, covering five domains: agriculture, climate, economy, energy, and health. The datasets are recorded at varying temporal resolutions, including weekly and monthly frequencies, with records spanning from the 1980s to 2024. Each dataset consists of univariate or multivariate time series, with the number of dimensions ranging from 1 to 11.

Textual Data consists of expert reports and news summaries. Expert reports are categorized as statistical prompts, as they provide insights into averages and trends within specific timeframes. News summaries can serve as contextual prompts, offering cues about future trends. We follow TimeCMA [Liu *et al.*, 2025e] to wrap time series values into numerical prompts. The word token embeddings are obtained from the pre-trained GPT-2 [Pan *et al.*, 2024].

¹<https://github.com/ChenxiLiu-HNU/CM2TS>

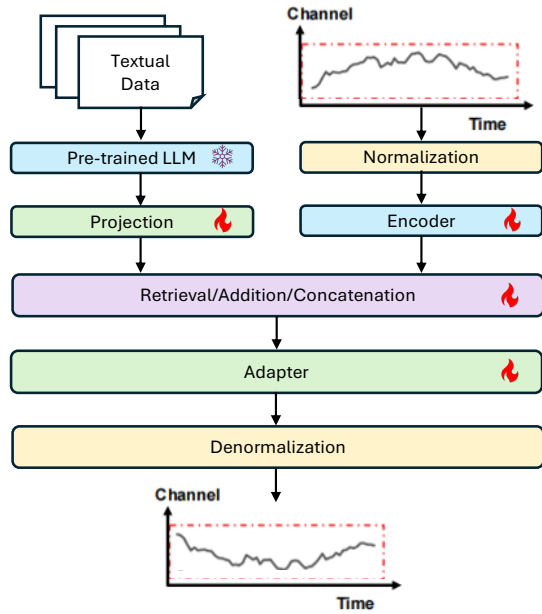


Figure 4: Overall Framework.

5.2 Implementation Details

Figure 4 presents an overview of our cross-modality modeling framework for time series analysis, comprising three key components: a pre-trained LLM, an alignment or fusion layer, and an adapter, detailed below.

Pre-trained LLM. This module includes a tokenizer and a pre-trained GPT-2 model with frozen weights, efficiently embeds textual data.

Encoder. It can be a Transformer-based encoder than embeds time series data and captures temporal dynamics [Liu *et al.*, 2024g], with reversible instance normalization applied to normalize time series values.

Alignment or Fusion. The alignment strategy employs a unidirectional retrieval method. The fusion strategy includes concatenation and addition methods.

Task Adapter. The adapter is designed by the downstream task. In this section, we conduct the forecasting task. We use a linear projection layer to predict the future time series. We also de-normalize the values to restore the actual prediction.

5.3 Experiment Settings

We configure the lookback window size to 36 for weekly data and 8 for monthly data, while setting the prediction horizon to 24 time steps across all datasets. We evaluate models’ performance using Mean Squared Error (MSE) and Mean Absolute Error (MAE). The model is optimized using the Adam optimizer, with Cosine Annealing as the learning rate scheduler. Training is conducted on NVIDIA A100 GPUs, with 20 epochs for weekly data and 50 epochs for monthly data.

5.4 Results and Discussion

Table 2 presents the results of cross-modality time series forecasting across five domains. The findings highlight the impact

of incorporating textual data and provide insights into the effectiveness of different alignment and fusion methods.

Textual Data Enhances Forecasting. Overall, textual data significantly enhance time series forecasting performance. For example, in the climate domain, the retrieval-based numerical prompt reduces MSE by 22.6% compared to the time series-only baseline. Among the different types of textual information, numerical prompts and statistical prompts lead to the most notable improvements. These text types contain numerical values that directly correlate with time series patterns, providing structured signals for forecasting. In contrast, contextual prompts and word token embeddings show relatively weaker performance, as they contain less structured information, which may not align as closely with numerical trends.

Numerical Prompts Perform Better. Numerical prompts consistently deliver the best performance across most domains. This is particularly evident in climate, energy, and health forecasting. For example, in the climate and health domains, numerical prompts reduce MSE by 21.15% and 17.95%, respectively, compared to contextual prompts. Statistical prompts show strong results in economy forecasting. Word token embeddings perform the worst, indicating that general semantic representations may not effectively capture time series-relevant information.

Alignment Outperforms Fusion. We evaluate three methods for each textual data type: retrieval-based alignment, addition-based fusion, and concatenation-based fusion. Retrieval-based alignment consistently performs well, particularly for numerical and statistical prompts. In the the economy domain, when using numerical prompts as textual data, retrieval reduces MSE by 16.96% compared to concatenation. Addition-based fusion often outperforms concatenation with the three textual prompts but underperforms it when using word token embeddings.

Domain-Specific Observations. The impact of textual data varies across different domains. Economy and health forecasting benefit considerably from numerical and statistical prompts, where structured reports align well with economic indicators and public health trends. Climate and agriculture forecasting show more moderate improvements. The reason is that these domains may rely on more external factors that are not always well captured by textual descriptions. Energy forecasting achieves the better results with addition-based fusion of numerical prompts.

6 Future Directions

Multi-Modality Modeling. Expanding beyond the integration of time series and textual data, future research could delve into additional modalities such as images [Huang *et al.*, 2024b], video [Wang *et al.*, 2024a], and audio [Huang *et al.*, 2021]. In this context, it is essential to explore the capability of LLMs for enhancing multimodal representation. Recent advancements in multi-modal LLMs exemplify the potential of such integrations. For instance, Meta’s LLaMa 3.2 processes both images and textual data, enabling applications ranging from augmented reality to document summarization. These developments underscore the importance of

Data	Method	Agriculture		Climate		Energy		Economy		Health	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Single Modality											
Time Series	-	2.68	1.35	0.371	0.473	0.183	0.315	0.0244	0.128	1.12	0.774
Cross Modality											
Time Series & Numerical Prompt	Retrieval	2.83	1.38	0.287	0.425	0.186	0.315	0.0247	0.129	0.96	0.665
	Addition	2.77	1.34	0.297	0.434	0.180	0.309	0.0252	0.132	1.05	0.754
	Concatenation	2.87	1.41	<u>0.296</u>	<u>0.425</u>	0.224	0.350	0.0267	0.132	1.16	0.797
Time Series & Statistical Prompt	Retrieval	<u>2.67</u>	<u>1.33</u>	0.386	0.483	0.196	0.333	0.0232	0.125	<u>0.97</u>	<u>0.667</u>
	Addition	2.64	1.29	0.380	0.478	0.183	0.313	<u>0.0244</u>	<u>0.126</u>	1.08	0.771
	Concatenation	2.75	1.38	0.386	0.488	0.190	0.325	0.0254	0.130	1.16	0.802
Time Series & Contextual Prompt	Retrieval	2.88	1.35	0.389	0.493	0.185	0.317	0.0261	0.131	1.11	0.661
	Addition	2.85	1.32	0.364	0.467	<u>0.182</u>	<u>0.313</u>	0.0263	0.132	1.17	0.713
	Concatenation	2.93	1.41	0.387	0.485	0.193	0.327	0.0272	0.134	1.28	0.776
Time Series & Word Token Embedding	Retrieval	2.90	1.37	0.388	0.490	0.181	0.312	0.0271	0.135	1.18	0.721
	Addition	2.95	1.45	0.393	0.479	0.187	0.319	0.0265	0.133	1.22	0.768
	Concatenation	2.92	1.42	0.389	0.484	0.188	0.320	0.0274	0.138	1.29	0.779

Table 2: Time series forecasting performance across multiple domains using diverse textual data and cross-modality modeling methods.

investigating how LLMs can be leveraged to create effective multimodal modeling.

Improving Effectiveness. While LLM-based cross-modality methods have demonstrated strong capabilities, they do not always surpass smaller, task-specific models [Wang *et al.*, 2024c]. In some cases, employing an LLM with an excessive number of parameters can lead to overfitting, particularly on specialized tasks across several domains. Future research could focus on techniques such as dynamic model selection, meta-learning, and continual adaptation that can help improve model effectiveness by allowing models to adjust to changing data distribution.

Efficient Optimization. Despite their success, existing studies still meet the challenge of high computational costs, particularly when processing long sequences, more tokens, or handling multivariate data. This is due to the high dimensionality of multivariate time series (i.e., multiple variables over timestamps) and the multi-head attention mechanism within LLMs. Recent advancements have explored strategies to mitigate these challenges, such as last token storage [Liu *et al.*, 2025e], knowledge distillation [Gu *et al.*, 2024]. Future research could focus on developing lightweight architectures, efficient attention mechanisms, and adaptive computation frameworks to optimize efficiency and scalability.

Transparency of LLMs. LLMs have demonstrated remarkable performance in textual-time series analytics [Wang *et al.*, 2024b], yet they often operate as “black-box” systems, raising concerns about their reasoning processes and overall transparency. Much of the current research primarily applies or fine-tunes LLMs without an explicit focus on exposing their internal reasoning processes. This lack of interpretability can hinder trust, particularly in high-stakes applications such as healthcare and finance. Moreover, LLMs are prone to generating hallucinations, seemingly plausible but incorrect

outputs, which further complicates their deployment in real-world scenarios. Future research on textual-time series analysis could prioritize enhancing the transparency of LLMs, ensuring that these models operate more reliably during subsequent alignment or fusion processes.

7 Conclusion

This paper aims to highlight the importance of cross-modality modeling for time series analytics in the LLM era. We propose a novel taxonomy from a textual data-centric perspective, categorizing existing studies by key data types, namely numerical prompts, statistical prompts, contextual prompts, and word token embeddings. Our main premise is that through cross-modality alignment and fusion, textual data can significantly enhance time series analytics tasks across diverse domains. To validate this viewpoint, we perform multi-domain multimodal experiments to systematically evaluate the effectiveness of various alignment and fusion strategies in key time series tasks. Finally, we explore open challenges and promising directions for future research.

Acknowledgments

This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contributions from the industry partner(s).

References

[Alnegheimish *et al.*, 2024] Sarah Alnegheimish, Linh Nguyen, Laure Berti-Équille, and Kalyan Veeramachaneni. Can large language models be anomaly detectors for time series? In *DSAA*, pages 1–10, 2024.

- [Cai *et al.*, 2024] Jiawei Cai, Dong Wang, Hongyang Chen, Chenxi Liu, and Zhu Xiao. Modeling dynamic spatiotemporal user preference for location prediction: a mutually enhanced method. *World Wide Web*, 27(2):14, 2024.
- [Cao *et al.*, 2024] Defu Cao, Furong Jia, Sercan O Arik, Tomas Pfister, Yixiang Zheng, Wen Ye, and Yan Liu. Tempo: Prompt-based generative pre-trained transformer for time series forecasting. In *ICLR*, 2024.
- [Chen *et al.*, 2020] Huiling Chen, Dong Wang, and Chenxi Liu. Towards semantic travel behavior prediction for private car users. In *HPCC*, pages 950–957, 2020.
- [Chen *et al.*, 2024] Can Chen, Gabriel Oliveira, Hossein Sharifi Noghabi, and Tristan Sylvain. LLM-TS Integrator: Integrating llm for enhanced time series modeling. *arXiv*, 2024.
- [Gruver *et al.*, 2023] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters. In *NeurIPS*, 2023.
- [Gu *et al.*, 2024] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *ICLR*, 2024.
- [Han *et al.*, 2024] Xiao Han, Zhenduo Zhang, Yiling Wu, Xinfeng Zhang, and Zhe Wu. Event traffic forecasting with sparse multimodal data. In *MM*, pages 8855–8864, 2024.
- [Hu *et al.*, 2025] Yuxiao Hu, Qian Li, Dongxiao Zhang, Jinyue Yan, and Yuntian Chen. Context-Alignment: Activating and enhancing llm capabilities in time series. In *ICLR*, 2025.
- [Huang *et al.*, 2021] Zhiqi Huang, Fenglin Liu, Xian Wu, Shen Ge, Helin Wang, Wei Fan, and Yuexian Zou. Audio-oriented multimodal machine comprehension via dynamic inter-and intra-modality attention. In *AAAI*, volume 35, pages 13098–13106, 2021.
- [Huang *et al.*, 2024a] Qihe Huang, Zhengyang Zhou, Kuo Yang, Gengyu Lin, Zhongchao Yi, and Yang Wang. Leret: Language-empowered retentive network for time series forecasting. In *IJCAI*, 2024.
- [Huang *et al.*, 2024b] Yipo Huang, Leida Li, Pengfei Chen, Haoning Wu, Weisi Lin, and Guangming Shi. Multi-modality multi-attribute contrastive pre-training for image aesthetics computing. *TPAMI*, 2024.
- [Jia *et al.*, 2024] Furong Jia, Kevin Wang, Yixiang Zheng, Defu Cao, and Yan Liu. GPT4MTS: Prompt-based large language model for multimodal time-series forecasting. In *AAAI*, volume 38, pages 23343–23351, 2024.
- [Jiang *et al.*, 2024] Yushan Jiang, Zijie Pan, Xikun Zhang, Sahil Garg, Anderson Schneider, Yuriy Nevmyvaka, and Dongjin Song. Empowering time series analysis with large language models: A survey. In *IJCAI*, 2024.
- [Jin *et al.*, 2022] Guangyin Jin, Chenxi Liu, Zhexu Xi, Hengyu Sha, Yanyun Liu, and Jincan Huang. Adaptive dual-view wavenet for urban spatial-temporal event prediction. *Information Sciences*, 588:315–330, 2022.
- [Jin *et al.*, 2024a] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-LLM: Time series forecasting by reprogramming large language models. In *ICLR*, 2024.
- [Jin *et al.*, 2024b] Ming Jin, Yifan Zhang, Wei Chen, Kexin Zhang, Yuxuan Liang, Bin Yang, Jindong Wang, Shirui Pan, and Qingsong Wen. Position: What can large language models tell us about time series analysis. In *ICML*, 2024.
- [Li *et al.*, 2024] Jun Li, Che Liu, Sibao Cheng, Rossella Arcucci, and Shenda Hong. Frozen language model helps ecg zero-shot learning. In *Medical Imaging with Deep Learning*, pages 402–415, 2024.
- [Liu and Zhang, 2025] Aoyu Liu and Yaying Zhang. CrossST: An efficient pre-training framework for cross-district pattern generalization in urban spatio-temporal forecasting. In *ICDE*, 2025.
- [Liu *et al.*, 2021a] Chenxi Liu, Jiawei Cai, Dong Wang, Jiaxin Tang, Lei Wang, Huiling Chen, and Zhu Xiao. Understanding the regular travel behavior of private vehicles: an empirical evaluation and a semi-supervised model. *IEEE Sensors Journal*, 21(17):19078–19090, 2021.
- [Liu *et al.*, 2021b] Chenxi Liu, Dong Wang, Huiling Chen, and Renfa Li. Study of forecasting urban private car volumes based on multi-source heterogeneous data fusion. *Journal on Communication*, 42(3), 2021.
- [Liu *et al.*, 2021c] Chenxi Liu, Zhu Xiao, Dong Wang, lei Wang, Hongbo Jiang, Hongyang Chen, and Jiangxia Yu. Exploiting spatiotemporal correlations of arrive-stay-leave behaviors for private car flow prediction. *TNSE*, 9(2):834–847, 2021.
- [Liu *et al.*, 2022] Chenxi Liu, Zhu Xiao, Dong Wang, Minhao Cheng, Hongyang Chen, and Jiawei Cai. Foreseeing private car transfer between urban regions with multiple graph-based generative adversarial networks. *World Wide Web*, 25(6):2515–2534, 2022.
- [Liu *et al.*, 2024a] Chen Liu, Shibo He, Qihang Zhou, Shizhong Li, and Wenchao Meng. Large language model guided knowledge distillation for time series anomaly detection. In *IJCAI*, 2024.
- [Liu *et al.*, 2024b] Chenxi Liu, Sun Yang, Qianxiong Xu, Zhishuai Li, Cheng Long, Ziyue Li, and Rui Zhao. Spatial-temporal large language model for traffic prediction. In *MDM*, 2024.
- [Liu *et al.*, 2024c] Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Ling kai Kong, Harshavardhan Kamarthi, Aditya B. Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, and B. Aditya Prakash. Time-MMD: A new multi-domain multimodal dataset for time series analysis. In *NeurIPS*, 2024.
- [Liu *et al.*, 2024d] Peiyuan Liu, Hang Guo, Tao Dai, Naiqi Li, Jigang Bao, Xudong Ren, Yong Jiang, and Shu-Tao Xia. CALF: Aligning llms for time series forecasting via cross-modal fine-tuning. In *AAAI*, 2024.

- [Liu *et al.*, 2024e] Qingxiang Liu, Xu Liu, Chenghao Liu, Qingsong Wen, and Yuxuan Liang. Time-FFM: Towards llm-empowered federated foundation model for time series forecasting. In *NeurIPS*, 2024.
- [Liu *et al.*, 2024f] Xu Liu, Junfeng Hu, Yuan Li, Shizhe Diao, Yuxuan Liang, Bryan Hooi, and Roger Zimmermann. UniTime: A language-empowered unified model for cross-domain time series forecasting. In *WWW*, pages 4095–4106, 2024.
- [Liu *et al.*, 2024g] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *ICLR*, 2024.
- [Liu *et al.*, 2024h] Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. AutoTimes: Autoregressive time series forecasters via large language models. In *NeurIPS*, 2024.
- [Liu *et al.*, 2024i] Ziqiao Liu, Hao Miao, Yan Zhao, Chenxi Liu, Kai Zheng, and Huan Li. LightTR: A lightweight framework for federated trajectory recovery. In *ICDE*, pages 4422–4434, 2024.
- [Liu *et al.*, 2025a] Chenxi Liu, Hettige Kethmi Hirushini, Qianxiong Xu, Cheng Long, Ziyue Li, Shili Xiang, Rui Zhao, and Gao Cong. ST-LLM+: Graph enhanced spatial-temporal large language model for traffic prediction. *TKDE*, pages 1–14, 2025.
- [Liu *et al.*, 2025b] Chenxi Liu, Hao Miao, Qianxiong Xu, Shaowen Zhou, Cheng Long, Yan Zhao, Ziyue Li, and Rui Zhao. Efficient multivariate time series forecasting via calibrated language models with privileged knowledge distillation. In *ICDE*, 2025.
- [Liu *et al.*, 2025c] Chenxi Liu, Zhu Xiao, Cheng Long, Dong Wang, Tao Li, and Hongbo Jiang. MVCAR: Multi-view collaborative graph network for private car carbon emission prediction. *TITS*, 26(1):472–483, 2025.
- [Liu *et al.*, 2025d] Chenxi Liu, Zhu Xiao, Wangchen Long, Tong Li, Hongbo Jiang, and Keqin Li. Vehicle trajectory data processing, analytics, and applications: A survey. *ACM ACM Comput. Surv.*, 57(9), 2025.
- [Liu *et al.*, 2025e] Chenxi Liu, Qianxiong Xu, Hao Miao, Sun Yang, Lingzheng Zhang, Cheng Long, Ziyue Li, and Rui Zhao. TimeCMA: Towards llm-empowered multivariate time series forecasting via cross-modality alignment. In *AAAI*, volume 39, pages 18780–18788, 2025.
- [Özyurt *et al.*, 2023] Yilmazcan Özyurt, Stefan Feuerriegel, and Ce Zhang. Contrastive learning for unsupervised domain adaptation of time series. In *ICLR*, 2023.
- [Pan *et al.*, 2024] Zijie Pan, Yushan Jiang, Sahil Garg, Anderson Schneider, Yuriy Nevmyvaka, and Dongjin Song. S²IP-LLM: Semantic space informed prompt learning with llm for time series forecasting. In *ICML*, 2024.
- [Pettersson *et al.*, 2023] Markus B Pettersson, Mohammad Kakooei, Julia Ortheden, Fredrik D Johansson, and Adel Daoud. Time series of satellite imagery improve deep learning estimates of neighborhood-level poverty in africa. In *IJCAI*, pages 6165–6173, 2023.
- [Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [Sun *et al.*, 2024] Chenxi Sun, Hongyan Li, Yaliang Li, and Shenda Hong. TEST: text prototype aligned embedding to activate llm’s ability for time series. In *ICLR*, 2024.
- [Tao *et al.*, 2024] Xiaoyu Tao, Tingyue Pan, Mingyue Cheng, and Yucong Luo. Hierarchical multimodal llms with semantic space alignment for enhanced time series classification. *arXiv*, 2024.
- [Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv*, 2023.
- [Wang *et al.*, 2024a] Binwu Wang, Pengkun Wang, Yudong Zhang, Xu Wang, Zhengyang Zhou, and Yang Wang. Condition-guided urban traffic co-prediction with multiple sparse surveillance data. *TVT*, 2024.
- [Wang *et al.*, 2024b] Xinlei Wang, Maike Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. From news to forecast: Integrating event analysis in llm-based time series forecasting with reflection. *arXiv*, 2024.
- [Wang *et al.*, 2024c] Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. In *NeurIPS*, 2024.
- [Xiao *et al.*, 2022] Jianhua Xiao, Zhu Xiao, Dong Wang, Vincent Havyarimana, Chenxi Liu, Chengming Zou, and Di Wu. Vehicle trajectory interpolation based on ensemble transfer regression. *TITS*, 23(7):7680–7691, 2022.
- [Xu *et al.*, 2024] Ronghui Xu, Hao Miao, Senzhang Wang, Philip S Yu, and Jianxin Wang. Pefad: A parameter-efficient federated framework for time series anomaly detection. In *SIGKDD*, pages 3621–3632, 2024.
- [Xue and Salim, 2023] Hao Xue and Flora D Salim. Prompt-cast: A new prompt-based learning paradigm for time series forecasting. *TKDE*, 2023.
- [Zhang *et al.*, 2024] Xiyuan Zhang, Ranak Roy Chowdhury, Rajesh K. Gupta, and Jingbo Shang. Large language models for time series: A survey. In *IJCAI*, pages 8335–8343, 2024.
- [Zhang *et al.*, 2025] Weiqi Zhang, Jiexia Ye, Ziyue Li, Jia Li, and Fugee Tsung. MedualTime: A dual-adaptor language model for medical time series-text multimodal learning. In *IJCAI*, 2025.
- [Zhao *et al.*, 2024] Fei Zhao, Chengcui Zhang, and Baocheng Geng. Deep multimodal data fusion. *ACM Comput. Surv.*, 56(9):216:1–216:36, 2024.