Beyond Numbers: A Survey of Time Series Analysis in the Era of Multimodal LLMs

Xiongxiao Xu[1], **Yue Zhao**[2], **Philip S. Yu**[3], **Kai Shu**[*4]
[1]Illinois Institute of Technology, [2]University of Southern California,
[3]University of Illinois Chicago, [4]Emory University
xxu85@hawk.iit.edu, yzhao010@usc.edu, psyu@uic.edu, kai.shu@emory.edu

Project Website: https://mllm-ts.github.io

Abstract

The rapid advancements in Multimodal Large Language Models (MLLMs) have garnered significant research attention and revolutionized various domains, including time series analysis. Notably, time series data can be represented in diverse modalities, making it highly compatible with the progress of MLLMs. This survey provides a comprehensive overview of time series analysis in the era of multimodal LLMs. We systematically summarize existing work from two perspectives: data (*taxonomy of time series modalities*) and models (*taxonomy of multimodal LLMs*). From a data perspective, we emphasize that time series, traditionally represented as a sequence of *numbers* with temporal order, can also be expressed in modalities such as text, images, graphs, audios, and tables. From a model perspective, we explore MLLMs that are either applicable or hold potential for specific time series modalities. Finally, we identify future research directions and key challenges at the intersection of time series and MLLMs, including the video modality, reasoning, agents, interpretability, and hallucinations. We curate and maintain a GitHub repository to facilitate the latest developments in this rapidly evolving field at HERE.

1 Introduction

Multimodal Large Language Models (MLLMs) have demonstrated unprecedented capability across diverse domains, marking a stride toward Artificial General Intelligence (AGI). For example, GPT-4 (Achiam et al., 2023) has achieved human-level performance on multiple benchmarks, ranking the top 10% of test takers in an exam.

The success of MLLMs has opened vast opportunities in time series analysis. ***Time series, traditionally represented as a temporally ordered sequence of numbers, can be flexibly expressed across diverse modalities, including text, images, graphs, audios, and tables***. As shown in Figure 1, time series in various modalities can be fed into MLLMs to process. Below, we highlight key connections between time series and other modalities:

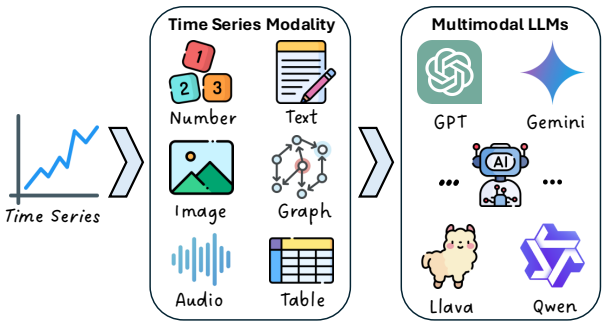

Figure 1: Time series, traditionally represented as a temporally ordered sequence of numbers, can be flexibly represented in number, text, image, graph, audio, and table formats, enabling multimodal LLMs to process them effectively.

Numbers. Time series is inherently a sequence of temporally ordered numerical values. The numerical sequences can be directly trained to develop large-scale time series foundation models (Garza et al., 2023), excelling in zero-shot or few-shot numerical tasks without requiring modality adaption. However, it is computationally expensive to build time series foundation models at scale and suffers from limited interpretability (Shi et al., 2024).

Text. Time series can be tokenized into textual representation, enabling the application of LLM methodologies (Gruver et al., 2024). This approach leverages the sequential nature shared between text (as ordered tokens) and time series (as ordered numerical values). However, the text format of time series may suffer from a modality gap between text and numbers. For instance, MLLMs make surprising errors in the basic numerical tasks, like incorrectly evaluating $9.11 > 9.9$ (Yang et al., 2024a). Furthermore, the limited input context of LLMs (Ding et al., 2024) poses challenges to fit high-dimensional multivariate time series (MTS).

*Corresponding author. Work updated on Mar. 27, 2025.

Table 1: Summary of the six time series modalities in the era of multimodal LLMs. TS: time series; UTS: univariate time series; MTS: multivariate time series.

Modality	TS Type	Advantage	Limitation	Domain
Number	UTS, MTS	Raw time series; Eliminate modality adaption	Computationally expensive; Limited interpretability	General
Text	UTS, MTS	Easily integrate with various LLM methodologies	Modality gap; Limited input context length	General, Urban, Finance, Healthcare
Image	UTS, MTS	Visual representation of TS; Robust for irregularly sampled TS	Sensitive to resolution; Challenging for MTS	General, Finance, Healthcare
Graph	MTS	Capture inter-variate dependencies	Not applicable for UTS; Non-trivial to build a graph	General, Urban, Finance, Healthcare
Audio	UTS, MTS	Easily integrate with audio processing techniques	Specialized audio preprocessing; Computational & memory overhead	Audio
Table	UTS, MTS	Preserve both temporal and channel-specific information	Lose sequential dependency without indicating the time dimension	General

Images. Time series can be visualized as images to enable pattern recognition via vision encoders (Li et al., 2024c). The image representation provides an intuitive and robust way to capture temporal characteristics, even in challenging scenarios like irregularly sampled data (Xu et al., 2025). However, the resolution of the generated images can significantly impact model performance, and effectively visualizing high-dimensional MTS as images remains a challenging open issue.

Graphs. Multivariate time series can be structured as graphs by treating channels (i.e., univariate time series) as nodes and defining edges based on inter-channel relationships (Chen et al., 2023c). Such representation is valuable for tasks requiring the understanding of inter-channel dynamics like spatiotemporal modeling. However, this approach is not applicable for univariate time series (UTS), and and constructing an effective graph is non-trivial.

Audios. Time series naturally aligns with audio data, as audio signals are fundamentally sequences of numerical amplitudes over time (Yang et al., 2021). Techniques such as spectrogram analysis (Wang et al., 2019) in audio processing share similarities with frequency-domain transformations in time series analysis. However, audio-based representation often requires specialized audio preprocessing approaches and incurs intensive computational and memory overhead.

Tables. Time series can be structured into a tabular format, where rows correspond to time steps and columns represent channels. This representation ensures the preservation of both temporal dynamics and channel-specific information (Wang et al., 2024b) of time series. However, tabular representation may lose sequential dependencies unless explicitly indicating the time dimension.

We summarize the characteristics of the six different time series modalities in Table 1. The versatility of time series across multiple modalities aligns with the evolution of MLLMs, which have expanded from text-based processing to multimodal capability. This motivates our survey to explore time series analysis in the age of MLLMs.

To the best of our knowledge, this is a very early survey to review recent advancements on time series analysis in the era of MLLMs. We begin by introducing the background of time series analysis and multimodal LLMs (Section 2). Next, we summarize the literature from the perspective of data and models: ***taxonomy of time series modalities*** (Section 3) and ***taxonomy of multimodal LLMs*** (Section 4). To facilitate comparison, we present an overview of time series modalities and representative MLLMs in Table 2 and Table 3, respectively. Finally, we explore future research directions and open challenges (Section 5). In summary, the key contributions of this survey are:

- We present a comprehensive survey of time series analysis in the era of MLLMs, systematically summarizing existing work from the view of data and models.
- From a data perspective, we categorize time series into six different modalities: *numbers, text, images, graphs, audios, and tables*. We highlight their unique advantages, limitations, and applications, drawing insights from relevant literature.
- From a model perspective, we introduce representative MLLMs that are either currently applicable or hold potential for specific time series modalities, offering a valuable resource for time series researchers.
- We outline future research directions and open challenges, including video-based time series, reasoning, agents, interpretability, and hallucinations, to advance time series analysis in the evolving MLLMs landscape.

2 Background

In this section, we introduce the background knowledge of time series analysis and multimodal LLMs.

2.1 Time Series Analysis

Time series analysis has long been a fundamental area of research (Hamilton, 2020), typically involving tasks such as classification, forecasting, anomaly detection, and imputation. Early approaches relied on statistical and traditional machine learning methods like ARIMA (Ahmed and Cook, 1979) and SVM (Jin et al., 2007). With the advent of deep learning, models such as CNNs (Wu et al., 2018; Yao et al., 2019), RNNs (Cheng et al., 2018; Madan and Mangipudi, 2018), LSTMs (Wang et al., 2022; Xu et al., 2023b), GNNs (Chai et al., 2018; Zhu et al., 2022), Transformers (Wen et al., 2022; Yıldız et al., 2022), and SSMs (Rangapuram et al., 2018; Xu et al., 2024) have demonstrated superior performance by effectively capturing complex temporal patterns.

Recently, the emergence of LLMs have revolutionized time series analysis. A straightforward approach is to treat time series as a form of natural language, harnessing the powerful capability of LLMs. Two recent surveys (Zhang et al., 2024c; Jiang et al., 2024) discuss this intersection. (Zhang et al., 2024c) examines various methodologies for applying LLMs to time series and offers an overview of multimodal datasets. (Jiang et al., 2024) reviews existing methods that employs LLMs for time series analysis and their domain-specific applications. However, the rapid evolution of MLLMs, which extend from text to diverse modalities, has opened new opportunities for time series analysis. In this survey, we emphasize time series work in the era of MLLMs and discuss existing work from data and model perspectives.

2.2 Multimodal Large Language Models

The evolution of MLLMs originates from text-only foundational models such as BERT (Devlin et al., 2018), GPT (Radford et al., 2018), and T5 (Raffel et al., 2020). Subsequent advancements in LLMs are driven by scaling laws (Kaplan et al., 2020) and the integration of multimodal inputs, leading to increasingly larger and stronger models like GPT-4 (Achiam et al., 2023), Gemini-1.5 (Team et al., 2024), and Llama-3.2 (Grattafiori et al., 2024). Modern MLLMs are able to effectively process diverse modalities and develop various applications,

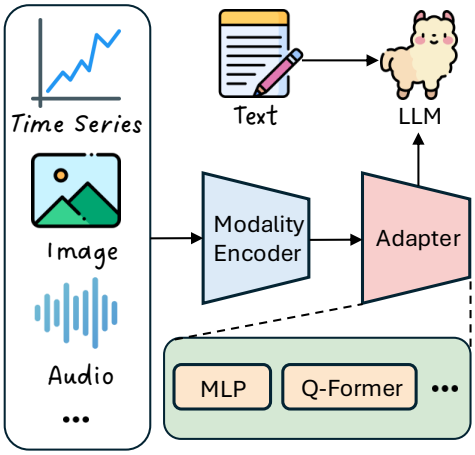

Figure 2: The architecture of multimodal LLMs comprises of three core components: a modality encoder, an adapter, and a LLM. The modality encoder processes input from a specific modality (e.g., time series, image, audio, etc.); the adapter bridges the gap between the modality-specific representation and textual embedding; and the LLM receives, processes, and reasons over both the textual and modality-aligned information.

including text (Liang et al., 2024; Huang et al., 2024a), images (Liu et al., 2024b; Hu et al., 2024), audios (Zhang et al., 2023a; Huang et al., 2024b).

As shown in Figure 2, a MLLM typically consists of three key components: a modality encoder, a LLM, and an adapter that bridges them (Yin et al., 2023; Caffagni et al., 2024; Zhang et al., 2024a). The modality encoder processes non-text inputs, such as images and audios, analogous to human sensory organs that capture visual or auditory information. The LLM handles both text inputs and processed modality-specific embeddings, functioning as a cognitive center to interpret and reason over the received data. The adapter plays a crucial role in aligning textual and modality-specific representation. Its design is particularly significant, as the modality encoder and LLM are generally frozen during training, leaving only the adapter trainable. The architecture of adapters ranges from lightweight MLP layers to more complex implementations, such as Q-Former (Li et al., 2023) and P-Former (Jian et al., 2024).

3 Taxonomy of Time Series Modalities

In this section, we discuss time series modalities, including numbers, text, images, graphs, audios, and tables. The taxonomy is provided in Table 2.

3.1 Time Series as Numbers

The raw time series is a sequence of numbers ordered chronologically:

Table 2: Taxonomy of time series modalities.

Method	Modality	Domain	Task	Modality Alignment Strategy	LLM
TimeGPT (Garza et al., 2023)	Number	General	Forecasting	None	Transformer
Lag-Llama (Rasul et al., 2023)	Number	General	Forecasting	None	Llama
TimesFM (Das et al., 2023)	Number	General	Forecasting	None	Decoder-only Transformer
Timer (Liu et al., 2024i)	Number	General	General	None	Decoder-only Transformer
MOIRAI (Woo et al., 2024)	Number	General	Forecasting	None	Encoder-only Transformer
MOMENT (Goswami et al., 2024)	Number	General	General	None	Encoder-only T5
Chronos (Ansari et al., 2024)	Number	General	Forecasting	None	T5
TIME-MOE (Shi et al., 2024)	Number	General	Forecasting	None	Decoder-only Transformer
Mantis (Feofanov et al., 2025)	Number	General	Classification	None	Vision Transformer
TimesBERT (Zhang et al., 2025)	Number	General	General	None	BERT
PromptCast (Xue and Salim, 2023)	Text	General	Forecasting	Direct Querying	GPT-3.5, T5, etc.
OFA (Zhou et al., 2023)	Text	General	General	Fine-Tuning	GPT-2, BERT, etc.
LLM4TS (Chang et al., 2023)	Text	General	Forecasting	Fine-Tuning	GPT-2
TEST (Sun et al., 2023)	Text	General	General	Text-Prototype-Aligned Contrast	GPT-2, BERT, etc.
TIME-LLM (Jin et al., 2023)	Text	General	Forecasting	Reprogramming	Llama
TEMPO (Cao et al., 2023)	Text	General	Forecasting	Semi-Soft Prompting	GPT-2
LLMTIME (Gruver et al., 2024)	Text	General	Forecasting	Tokenization	GPT-3, Llama-2
UniTime (Liu et al., 2024g)	Text	General	Forecasting	Tokenization	GPT-2
AutoTimes (Liu et al., 2024h)	Text	General	Forecasting	In-Context Forecasting	Llama
(Tang et al., 2025)	Text	General	Forecasting	Tokenization	GPT-4, Gemini-1.0, etc.
LSTPrompt (Liu et al., 2024d)	Text	General	Forecasting	Prompting	GPT-4, GPT-3.5
S^2IP-LLM (Pan et al., 2024)	Text	General	Forecasting	Semantic Space Informed Prompting	GPT-2
InstructTime (Cheng et al., 2024a)	Text	General	Classification	Fine-Tuning	GPT-2
CrossTimeNet (Cheng et al., 2024b)	Text	General	General	Fine-Tuning	BERT
GPT4MTS (Jia et al., 2024)	Text	General	Forecasting	Fine-Tuning	GPT-2
Time-MMD (Liu et al., 2024c)	Text	General	Forecasting	Fine-Tuning	GPT-2
ChatTime (Wang et al., 2024a)	Text	General	General	Fine-Tuning	Llama-2
TimeRAG (Yang et al., 2024c)	Text	General	Forecasting	Reprogramming	Llama-3
DECA (Hu et al., 2025)	Text	General	General	Context-Alignment	GPT-2
TaTS (Li et al., 2025)	Text	General	General	Tokenization	GPT-2 Encoder
AuxMobLCast (Xue et al., 2022)	Text	Urban	Forecasting	Fine-Tuning	BERT, GPT-2, etc.
LLM-Mob (Wang et al., 2023)	Text	Urban	Classification	Context-Inclusive Prompting	GPT-3.5
xTP-LLM (Guo et al., 2024)	Text	Urban	Forecasting	Fine-Tuning	Llama-2
(Xie et al., 2023)	Text	Finance	Classification	Direct Querying	ChatGPT
(Lopez-Lira and Tang, 2023)	Text	Finance	Forecasting	Direct Querying	ChatGPT
(Yu et al., 2023)	Text	Finance	Forecasting	Direct Querying	GPT-4, Llama
FinSeer (Xiao et al., 2025)	Text	Finance	Forecasting	Fine-Tuning	Llama-3.2
(Liu et al., 2023)	Text	Healthcare	General	Direct Querying	PaLM
MedTsLLM (Chan et al., 2024)	Text	Healthcare	General	Reprogramming	Llama-2
Insight Miner (Zhang et al., 2023d)	Image	General	Trend Description	Llava	Llava
(Daswani et al., 2024)	Image	General	Understanding	GPT-4o, Gemini-1.5	GPT-4o, Gemini-1.5
AnomLLM (Zhou and Yu, 2024)	Image	General	Anomaly Detection	GPT-4o, Gemini-1.5, etc.	GPT-4o, Gemini-1.5, etc.
TimeSeriesExam (Cai et al., 2024)	Image	General	Question Answering	GPT-4o, Gemini-1.5, etc.	GPT-4o, Gemini-1.5, etc.
TAMA (Zhuang et al., 2024)	Image	General	Anomaly Detection	GPT-4o	GPT-4o
VLM-TSC (Prithyani et al., 2024)	Image	General	Classification	Llava	Llava
Time-VLM (Zhong et al., 2025)	Image	General	Forecasting	ViLT	ViLT
VisualTimeAnomaly (Xu et al., 2025)	Image	General	Anomaly Detection	GPT-4o, Gemini-1.5, etc.	GPT-4o, Gemini-1.5, etc.
Agent Trading Arena (Ma et al., 2025)	Image	Finance	Simulation	GPT-4o, Gemini-1.5	GPT-4o, Gemini-1.5
ViTST (Li et al., 2024c)	Image	Healthcare	Classification	Swin Transformer	RoBERTa
METS (Li et al., 2024a)	Image	Healthcare	Classification	ResNet1d-18	ClinicalBert
HeLM (Belyaeva et al., 2023)	Image	Healthcare	Classification	ResNet18	Flan-PaLMChilla
GATGPT (Chen et al., 2023a)	Graph	General	Imputation	GAT	GPT-2
LLM-OSR (Yan et al., 2024)	Graph	General	Imputation	Graph Signal Processing	GPT-4o, GPT-3.5
STLLM (Zhang et al., 2023b)	Graph	Urban	Forecasting	GCN	GPT-3.5
ST-LLM (Liu et al., 2024a)	Graph	Urban	Forecasting	Spatial-Temporal Embedding	GPT-2, Llama-2
STG-LLM (Liu et al., 2024f)	Graph	Urban	Forecasting	STG-Tokenizer/Adapter	GPT-2
STGCN-L (Li et al., 2024b)	Graph	Urban	Forecasting	STGCN	GPT-4
(Qin et al., 2024)	Graph	Urban	Imputation	None	GPT-3.5
Strada-LLM (Moghadas et al., 2024)	Graph	Urban	Forecasting	Hierarchical Feature Extractor	Mistral
(Chen et al., 2023c)	Graph	Finance	Classification	GNN	ChatGPT
LA-GCN (Xu et al., 2023a)	Graph	Healthcare	Classification	GCN	BERT
Voice2Series (Yang et al., 2021)	Audio	Audio	Classification	Reprogramming	Transformer
TableTime (Wang et al., 2024b)	Table	General	Classification	Table Encoding	Llama-3.1
TabPFN-TS (Hoo et al., 2025)	Table	General	Forecasting	Feature Engineering	TabPFN

Formulation. A time series can be denoted as $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\} \in \mathbb{R}^{T \times M}$ with T time steps and M variates, where $\mathbf{x}_t \in \mathbb{R}^M$ carries M numerical values at the t^{th} time steps.

Inspired by the remarkable development of foundation models in NLP, time series researchers are working toward building "time series LLMs." These foundation models are trained on vast amounts of time series data from scratch and are competent in zero-shot or few-shot numerical tasks without additional modality adaptation (Garza et al., 2023; Shi et al., 2024; Feofanov et al., 2025).

Similar to LLMs, time series foundation models follow three primary architectural paradigms: encoder-decoder, encoder-only, and decoder-only. The encoder-decoder architecture is from vanilla Transformer (Vaswani, 2017) where the encoder transforms an input sentence into representation and the decoder generates the output sequence token by token, based on the encoded representation and previously decoded tokens. TimeGPT (Garza et al., 2023) and Chronos (Ansari et al., 2024) adopt this approach. Encoder-only frameworks generate multi-step predictions in a single forward pass, eliminating the need for autoregressive decoding and mitigating error accumulation (Zeng et al., 2023), including MOIRAI (Woo et al., 2024), MOMENT (Goswami et al., 2024), and TimesBERT (Zhang et al., 2025). Mantis (Feofanov et al.,

2025) employs the adaption of Vision Transformer (ViT) (Dosovitskiy, 2020) architecture, which is an encoder-only framework. In contrast, decoder-only architectures such as Lag-Llama (Rasul et al., 2023), TimesFM (Das et al., 2023), Timer (Liu et al., 2024i), and TIME-MOE (Shi et al., 2024) emphasize sequential dependencies and employ autoregressive prediction. Notably, (Rasul et al., 2023) and (Liu et al., 2024i) observe that decoder-only models are more effective with larger training datasets compared to other architectures.

The key advantage of time series foundation models is their ability to perform numerical forecasting without requiring modality adaptation. Unlike LLM-based approaches that often require alignment between textual and numerical representation, these models inherently generate accurate numerical forecasts on unseen datasets, eliminating the need for modality adaptation. However, they often face computation challenges. To mitigate the issue, TTMs (Ekambaram et al., 2024) and Time-MOE (Shi et al., 2024) leverage lightweight backbones and Mixture-of-Experts (MoEs) (Fedus et al., 2022) architectures to improve efficiency. Built on the TTM, TimeRAF (Zhang et al., 2024b) enhances zero-shot time series forecasting by Retrieval-Augmented Generation (RAG) techniques (Lewis et al., 2020). Additionally, these methods are limited in their ability to generate human-readable outputs, which hinders interpretability and poses significant challenges for high-stakes applications, like the healthcare domain (Shaheen, 2021) and natural disaster prediction (Tan et al., 2021).

3.2 Time Series as Text

The time series can be encoded into textual representation:

Formulation. A time series $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\} \in \mathbb{R}^{T \times M}$ can be serialized into a string U using preprocessing methods such as delimiter-based separation (e.g., spaces or commas). The string U is then tokenized into a sequence of textual tokens $\mathbf{S} = \{s_1, \ldots, s_N\}$, where N represents the tokenized length of the sequence $\mathbf{S}$. This representation enables time series data to be processed by language models.

The integration of text-based time series with LLMs mainly involves four methodologies (Jiang et al., 2024): direct querying, tokenization, prompting, and fine-tuning techniques.

Direct querying is a straightforward method to leverage LLMs. By framing time series forecasting as a sentence-to-sentence generation task, PromptCast (Xue and Salim, 2023) directly apply LLMs to forecast numerical data. Direct querying can be enhanced by specialized tokenization and prompting techniques, like Chain-of-Thought (CoT) (Lightman et al., 2023). For example, LSTPrompt (Liu et al., 2024d) integrates CoT into prompts to form reasoning path for predictions. The querying can also obtain benefits from a external time series knowledge base (Yang et al., 2024c; Xiao et al., 2025) by Retrieval-Augmented Generation (RAG). In the real-world scenarios, direct querying is often facilitated by integrating with domain expertise (Lopez-Lira and Tang, 2023; Yu et al., 2023; Wang et al., 2023; Liu et al., 2023).

Simple tokenization methods like LLM-Time (Gruver et al., 2024) represent time series by adding spaces and commas. More tokenization designs integrate with time series inductive biases. The STL decomposition (Seasonal and Trend decomposition using Loess) (Cleveland et al., 1990) is a classical decomposition technique in time series studies (Tang et al., 2025; Liu et al., 2024d). For example, TEMPO (Cao et al., 2023) and S^2IP-LLM (Pan et al., 2024) learn distinct time series embeddings for trend, seasonal, and residual components. Other common strategies include reversible instance normalization (Kim et al., 2021a), channel independence (Zeng et al., 2023), and patching (Nie et al., 2022), which help mitigate distribution shifts and preserve local information in time series data. These methods are widely integral to time series models (Sun et al., 2023; Jin et al., 2023; Cheng et al., 2024b; Jia et al., 2024). Additionally, ChatTime (Wang et al., 2024a) regards time series as a foreign language. TaTS (Li et al., 2025) treats text as auxiliary variables associated with the time series.

The advanced prompting techniques incorporate contextual information (Jin et al., 2023; Liu et al., 2024g; Tang et al., 2025; Cheng et al., 2024a), which is crucial for domain-specific applications (Xie et al., 2023; Wang et al., 2023; Guo et al., 2024). For example, (Yu et al., 2023) feeds the latest news into LLMs to predict stock movements. MedTsLLM (Chan et al., 2024) incorporates patient profiles and dataset statistics for medical analysis. AuxMobLCast (Xue et al., 2022) appends Place-of-Interest (POI) data for customer flow prediction. Another key technique is CoT

prompting, which enables LLMs to reason through predictions step by step (Liu et al., 2024d; Xie et al., 2023; Guo et al., 2024).

Fine-tuning pre-trained LLMs is essential for adapting them to downstream time series tasks (Chang et al., 2023; Jia et al., 2024). Many fine-tuning approaches aligns textual prompts with numerical values within a shared representation space to effectively activate LLMs' capability (Cao et al., 2023; Wang et al., 2024a). For example, DECA (Hu et al., 2025) proposes Context-Alignment paradigm to achieve alignment between time series and a linguistic component. However, fine-tuning all layers is computationally expensive and may lead to catastrophic forgetting (Kirkpatrick et al., 2017). A common solution is to update partial layers, where models such as OFA (Zhou et al., 2023) and GPT4MTS (Jia et al., 2024) freeze attention and feed-forward layers to retain majority knowledge of LLMs. Another efficient approach is Low-Rank Adaptation (LoRA) (Hu et al., 2021), which attaches a trainable low-rank matrix to the attention mechanism (Cao et al., 2023; Wang et al., 2024a; Xiao et al., 2025). Instead of fine-tuning LLMs, Time-MMD (Liu et al., 2024c) fine-tunes unimodal time series models and a projection layer to bridge numerical and textual representation spaces.

The primary advantage of text-based time series is their seamless compatibility with LLM methodologies, such as CoT prompting and fine-tuning. However, their performance is limited by two key challenges: the inherent representation gap between textual and numerical data (Gruver et al., 2024) and the constrained input context length of LLMs (Prithyani et al., 2024). For example, the GPT-3 tokenizer tends to break a single number into tokens that do not align with its digits (Gruver et al., 2024), where one token may represent multiple digits. This can lead to incorrect evaluations of basic arithmetic understanding, such as $9.11 > 9.9$ (Yang et al., 2024a).

3.3 Time Series as Images

The time series can be visualized as images:

Formulation. A time series $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\} \in \mathbb{R}^{T \times M}$ can be converted into an image format $\mathbf{I}$ like a line chart. This transformation allows time series data to be analyzed by MLLMs with a vision encoder.

Image-based time series, termed time series images (Xu et al., 2025), provides an intuitive way to understand their patterns. This approach has proven superiority over text- or number-based representation across classification, forecasting, and anomaly detection tasks (Prithyani et al., 2024; Chen et al., 2024a; Zhou and Yu, 2024; Ni et al., 2025), even for irregularly sampled time series data (Li et al., 2024c; Xu et al., 2025).

Intuitively, MLLMs are more effective at coarse-grained tasks like time series classification and anomaly detection compared to fine-grained time series forecasting. AnomLLM (Zhou and Yu, 2024) evaluates LLMs in anomaly detection by testing several pre-defined hypotheses. TAMA (Zhuang et al., 2024) is a MLLM-based framework that enhances time series classification and anomaly detection by analyzing time series images. VLM-TSC (Prithyani et al., 2024) demonstrates that Llava can produce competitive time series classification results in two epochs of finetuning. Visual-TimeAnomaly (Xu et al., 2025) finds that MLLMs detect range- and variate-wise anomalies more effectively than point-wise anomalies like humans.

Although less straightforward, image-based time series representation also benefit time series forecasting tasks. ViTime (Yang et al., 2024b) performs time series forecasting in the binary image space. VisionTS (Chen et al., 2024a) reformulates time series forecasting as an image reconstruction task. CLIP-LSTM (Wimmer and Rekabsaz, 2023) employs CLIP (Radford et al., 2021) to extract features from stock market line charts, which are then fed into an LSTM (Hochreiter and Schmidhuber, 1997) for stock movement prediction. While these methods do not directly harness LLMs as their backbones, they underscore the potential of image-based time series representation. Different from the above, Time-VLM (Zhong et al., 2025), built on a vision-language model ViLT (Kim et al., 2021b), augments time series forecasting by integrating text and vision modalities.

The emergence of MLLMs unlocks novel tasks such as time series understanding, reasoning, and simulation (Daswani et al., 2024; Kong et al., 2025; Ma et al., 2025), extending beyond traditional time series tasks, including classification, forecasting, anomaly detection, and imputation. By answering questions related to time series images, MLLMs can understand and reason over time series data. InsightMiner (Zhang et al., 2023d) queries MLLMs to describe different stages of temporal trends. (Daswani et al., 2024) demonstrates additional ben-

efits of visual time series representation over the textual format. TimeSeriesExam (Cai et al., 2024) designs a series of multiple-choice questions to examine MLLMs' capability to understand time series data. More interestingly, MLLMs can also simulate complex real-world scenarios, such as stock markets. Agent Trading Arena (Ma et al., 2025) simulates a zero-sum stock market environment where MLLMs act as agents to make investment decisions for stock portfolios.

A core advantage of image-based time series representation is their robustness to irregular time series (Xu et al., 2025) and adoption in the healthcare domain. Irregularly sampling is common in healthcare applications (Sun et al., 2020); for example, wearable devices often record sensor readings at irregular intervals due to patient activity or connectivity issues. The irregularity challenge can be effectively addressed by time series images, where missing values are left blank. ViTST (Li et al., 2024c) converts irregular time series into images and fine-tunes a Swin Transformer (Liu et al., 2021) for medical time series classification. Moreover, physiological indicators (e.g., ECG and spirogram) are often stored as images. METS (Li et al., 2024a) and HeLM (Belyaeva et al., 2023) feed the ECG and spirogram into language models with a vision encoder for multimodal learning. However, image-based representation is sensitive to resolution and present challenges in visualizing high-dimensional multivariate time series. VisualTimeAnomaly (Xu et al., 2025) highlights that MLLMs struggle with high-dimensional multivariate time series images due to the increased information density and reduced resolution per variate.

3.4 Time Series as Graphs

The time series can be structured into a graph:

Formulation. A time series $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\} \in \mathbb{R}^{T \times M}$ can be transformed into a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_1, \cdots, v_M\}$ is the set of nodes, corresponding to variates, and $\mathcal{E} = \{e_1, \cdots, e_E\}$ is the set of edges, defining inter-variate dependencies. This graph representation enables MLLMs with a graph encoder to model structure among variates.

Representing multivariate time series as graphs is an effective approach to model inter-dependencies among multiple variates. This is important for real-world applications with highly complex relationships like urban analytics (Zhang et al., 2023b).

Graph-based time series representation is impactful in spatio-temporal problems, including network traffic analysis, human mobility prediction, and intelligent transportation systems (Wang et al., 2020). GATGPT (Chen et al., 2023a) employs GATs (Veličković et al., 2017) to capture spatial dependencies while leveraging GPT-2 to model temporal relationships for spatio-temporal imputation. Similarly, STLLM (Zhang et al., 2023b), ST-LLM (Liu et al., 2024a), STG-LLM (Liu et al., 2024f), and STGCN-L (Li et al., 2024b) utilize LLMs to extract additional urban information from text-based data, such as POI tags and spatio-temporal context. Strada-LLM (Moghadas et al., 2024) advances traffic prediction by integrating proximal traffic information into a graph-aware LLM. Additionally, LLMs can online infer missing values in the spatial-temporal graph (Qin et al., 2024; Yan et al., 2024).

The benefits of graph-based time series extend to domains like finance and healthcare, where understanding intricate relationships between entities is critical. (Chen et al., 2023c) leverages ChatGPT's graph inference capability for stock movement prediction. By inferring dynamic network structures among companies from textual data, the approach enhances graph embeddings, leading to more accurate forecasts for the next trading day. In healthcare, LA-GCN (Xu et al., 2023a) integrates the prior knowledge of LLMs to augment GCNs for skeleton-based action recognition. Specifically, BERT's knowledge simulates brain regions involved in action reasoning, aiding GCNs in making more precise forecasts. However, graph-based time series methods face limitations: they are inapplicable to univariate time series, and it remains challenging to construct an appropriate graph for multivariate time series, such as determining the directionality and weight of edges.

3.5 Time Series as Audios

The time series can be regarded as audio data:

Formulation. A time series $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\} \in \mathbb{R}^{T \times M}$ can be treated as an audio signal, where each $\mathbf{x}_t \in \mathbb{R}^M$ represents the amplitude of M channels at the t^{th} time step. This representation allows MLLMs with an audio encoder to handle.

Time series naturally aligns with the audio signal, as audio data is fundamentally a sequence of numerical amplitudes over time (Tzanetakis and Cook,

2002). This correspondence enables the application of audio foundation models and audio-specific techniques (Yang et al., 2021).

Inspired by the similarity between voice data and univariate temporal signals, Voice2Series (Yang et al., 2021) reprograms pre-trained acoustic foundation models for time series classification. To harness the power of well-trained acoustic models, the approach introduces a trainable reprogramming layer and a label mapping function, ensuring alignment between the time series and audio signals. Despite the inherent alignment between audio and time series data, treating time series as audio introduces additional complexity as it necessitates audio preprocessing techniques. Moreover, audio-based processing can be significantly more resource-intensive in terms of both computation and memory compared to directly handling numerical time series data.

To the best of our knowledge, no more work explicitly connects time series tasks and audio data within the context of MLLMs, suggesting a promising research direction. For instance, audio processing techniques like spectrogram analysis (Wang et al., 2019) and wavelet transforms (Zhang and Zhang, 2019) could be adapted for univariate time series to capture frequency-domain features and multi-resolution patterns. Similarly, techniques like beamforming (Xu et al., 2017) and source separation (Makino, 2018), common in multi-channel audio processing, hold potential for advancing multivariate time series analysis.

3.6 Time Series as Tables

The time series can be formatted as a table:

Formulation. A time series $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\} \in \mathbb{R}^{T \times M}$ can be explicitly represented as a table where each row corresponds to a time step t, and each column represents a variate m.

The tabular format of time series preserves both temporal and channel-specific information better than serializing them into a textual format when processed by MLLMs (Wang et al., 2024b).

TableTime (Wang et al., 2024b) reformulates multivariate time series classification as a table understanding task. Specifically, it converts time series into a tabular format and feed them into Llama for zero-shot classification. TabPFN-TS (Hoo et al., 2025) harnesses the tabular foundation model TabPFN (Hollmann et al., 2022) for time series forecasting. Despite its compact size of only 11M parameters and a simple feature engineering approach, TabPFN-TS outperforms Chronos-Mini (Ansari et al., 2024), a model of comparable scale, and matches or even slightly exceeds Chronos-Large, which is 65 times larger. While table-based time series representations can preserve both temporal and channel-specific information, they may lose sequential dependencies if the time dimension is not explicitly indicated.

4 Taxonomy of Multimodal LLMs

While numerous LLMs are dedicated to text and images modalities, other modalities remain largely unexplored. In this section, we briefly outline text & image-compatible LLMs, and discuss LLMs which are compatible with less-explored modalities. Note that the comprehensive discussion of MLLMs is a broad and extensive topic beyond the scope of this survey. Instead, we focus on representative MLLMs to offer a resource for future research in time series. The taxonomy is provided in Table 3.

4.1 Text & Image-Compatible LLMs

Among various modalities, the integration of text and images has received the most attention, as language and vision are two fundamental ways humans perceive and interact with the world. Text & image-compatible LLMs can be categorized based on code accessibility into proprietary and open-source models. Proprietary MLLMs, such as GPT-4 (Achiam et al., 2023), Gemini-1.5 (Team et al., 2024), and Claude-3 (Anthropic, 2024), are not publicly available but can be accessed through APIs provided by their respective companies. In contrast, open-source MLLMs, including Llama-3.2 (Grattafiori et al., 2024), Qwen2.5-VL (Bai et al., 2025), and InternVL-2.5 (Chen et al., 2024c), allow researchers and developers access code or weights. An extensive discussion of text & image-compatible LLMs is beyond the scope of this work; we refer readers to existing surveys (Yin et al., 2023; Caffagni et al., 2024; Zhang et al., 2024a).

4.2 Graph-Compatible LLMs

Graph-compatible LLMs include models designed specifically for spatio-temporal problems or general graph tasks. UrbanGPT (Li et al., 2024d) tailors a spatio-temporal LLM for urban applications by incorporating a multi-level temporal convolutional network (Lea et al., 2016) to capture complex dependencies. It aligns textual and spatio-

Table 3: Taxonomy of representative multimodal LLMs.

Method	Modality	Task	Modality Encoder	Adapter	LLM
GPT-4 (Achiam et al., 2023)	Text & Image	General	Close-Source	Close-Source	Close-Source
Gemini-1.5 (Team et al., 2024)	Text & Image	General	Close-Source	Close-Source	Close-Source
Claude-3 (Anthropic, 2024)	Text & Image	General	Close-Source	Close-Source	Close-Source
Llama-3.2 (Grattafiori et al., 2024)	Text & Image	General	ViT	Cross-Attention	Llama-3.1
Qwen2.5-VL (Bai et al., 2025)	Text & Image	General	ViT	MLP	Qwen2.5
InternVL-2.5 (Chen et al., 2024c)	Text & Image	General	InternViT	MLP	InternLM-2.5, Qwen2.5
UrbanGPT (Li et al., 2024d)	Graph	General	Temporal Convolutional Network	MLP	Vicuna
STD-PLM (Huang et al., 2024c)	Graph	General	Spatial-Temporal Tokenizer	None	GPT-2
LLaGA (Chen et al., 2024b)	Graph	General	Node-Level Template	MLP	Vicuna
SpeechGPT (Zhang et al., 2023a)	Audio	General	HuBERT	None	Llama
AudioGPT (Huang et al., 2024b)	Audio	General	Whisper, GenerSpeech, etc.	None	GPT-3.5
MinMo (Chen et al., 2025)	Audio	General	SenseVoice	Transformer + CNN	Qwen2.5
TabPFN (Hollmann et al., 2022)	Table	General	None	None	PFN
TableLlama (Zhang et al., 2023c)	Table	General	None	None	Llama
TableGPT2 (Su et al., 2024a)	Table	General	Bi-Dimensional Attention	Q-Former	Qwen2.5

temporal data through instruction-tuning on Vicuna (Zheng et al., 2024) through a lightweight MLP. STD-PLM (Huang et al., 2024c) utilizes LoRA to fine-tune multi-head attention and positional embeddings. It enhances effectiveness with a spatial-temporal tokenizer and a sandglass attention module that captures higher-order region-level dependencies. Different from the two approaches which focus on spatio-temporal modeling, LLaGA (Chen et al., 2024b) is universally applicable for general graph tasks, including node classification, link prediction, and node description. It reorganizes graph data into node sequences with a node-level template and aligns graph and token spaces by a versatile projector MLP.

4.3 Audio-Compatible LLMs

Audio-compatible LLMs have significant potential for time series tasks, as both audio data and time series share inherent sequential patterns, temporal dependencies, and dynamic variations over time. SpeechGPT (Zhang et al., 2023a) is a speech-language model capable of perceiving and generating both speech and text. Trained on speech-text cross-modal and chain-of-modality instruction fine-tuning datasets, SpeechGPT demonstrates a strong ability to follow cross-modal instructions from humans. Rather than training from scratch, AudioGPT (Huang et al., 2024b) integrates various audio foundation models with GPT-3.5-Turbo for a wide range of audio tasks, including speech, music, sound processing, and talking head synthesis. For instance, AudioGPT employs Whisper (Radford et al., 2023) to perform speech recognition. Minmo (Chen et al., 2025), an end-to-end aligned MLLM, achieves SOTA performance on several open-source audio benchmarks, including spoken dialogue, multilingual speech recognition, and speech translation.

4.4 Table-Compatible LLMs

Table-compatible LLMs have emerged as a powerful paradigm for tackling tabular data challenges. Tabular data can be directly processed by LLMs or LLMs with a specialized table encoder. TabPFN (Hollmann et al., 2022) is a tabular foundation Transformer pre-trained to capture complex feature dependencies and causal mechanisms. It can be adapted (Hoo et al., 2025) to outperform specialized time series forecasters with minimal feature engineering. TableLlama (Zhang et al., 2023c) is a generalist model to tackle diverse table tasks. It mitigates the long-context challenge by fine-tuning Llama-2 with LongLoRA (Chen et al., 2023b) on the newly constructed TableInstruct dataset. TableGPT2 (Su et al., 2024a) further advances performance across 23 benchmarks by training on 593.8K tables and 2.36M query-table-output tuples. Its key innovation lies in a semantic table encoder with bi-dimensional attention, enabling schema- and cell-level representation.

5 Future Directions and Challenges

In this section, we point out several promising future research directions and their challenges.

Time Series as Videos. Video is a fundamental modality representing a sequence of temporal image frames. The temporal dependencies between frames suggest that time series can be modeled as video. (Zeng et al., 2021) formulates financial time series forecasting as a video prediction task by visualizing historical stock data of 9 assets in a 3×3 heatmap, where each pixel represents a relative percentage change. They adapt a video prediction network SRVP (Franceschi et al., 2020) to capture pixel changes for asset movement prediction. Although the above work highlights the potential of video-based time series modeling, it does not leverage powerful capability of MLLMs.

However, the paradigm of treating time series as video faces two key challenges. First, numerical forecasting is inherently a regression task requiring precise continuous outputs, whereas video-based models may struggle with fine-grained numerical accuracy. Instead, the approach may be more suited for high-level tasks like clarification and anomaly detection. Second, video-compatible LLMs typically incur high computational and memory overhead (Zolfaghari et al., 2018), making them impractical for time-sensitive applications, such as real-time traffic prediction (Zhang et al., 2020) and high-frequency trading (Levendovszky and Kia, 2012).

Time Series Reasoning. Reasoning is a hallmark of advanced intelligence and has recently garnered significant attention, like DeepSeek-R1 (Guo et al., 2025) and OpenAI-o1 (Jaech et al., 2024). It is observed that LLMs show surprisingly limited zero-shot reasoning capability over time series data (Merrill et al., 2024). Specifically, they score marginally above random on etiological reasoning and question answering tasks (up to 30% points worse than humans). To bridge the gap, fine-tuning LLMs demonstrates a promising path with various techniques, such as a self-critic temporal optimization method (Su et al., 2024b) and chain-of-thought (Chow et al., 2024).

However, multimodal reasoning capability of MLLMs, such as reasoning over image-based time series and audio-based time series, remains largely untapped. The key challenge in multimodal reasoning over time series lies in the seamless integration of various modalities. For example, in spatio-temporal problems, reasoning must account for both unstructured graph dependencies and temporal dynamics. The incorporation of additional modalities complicates the reasoning process. Moreover, time series data lacks inherent semantic meaning, making it difficult to design reasoning tasks (Kong et al., 2025; Chow et al., 2024).

Time Series Agents. Agents represent a promising step toward AGI (Wang et al., 2024c), enabling self-planning and autonomous task execution. Recent research has demonstrated the preliminary application of time series agents. (Ravuru et al., 2024) demonstrates the effectiveness of a hierarchical, multi-agent approach for time series analysis with RAG (Lewis et al., 2020). TimeCAP (Lee et al., 2025) employs two independent LLM agents for time series event prediction: one generates contextual information, while the other utilizes the enriched summary to make more informed predictions. Agent Trading Arena (Ma et al., 2025) simulates stock trading environments where agents discuss market trends, analyze stock data, and engage in trading activities.

While these studies emphasize the potential of agent-based approaches, developing MLLM-based agents for modalities beyond text remains an open challenge. A key issue in designing MLLM-based agents for time series analysis is the effective integration of heterogeneous modalities. It requires robust mechanisms to align and fuse information across diverse domains. Furthermore, the complexity of coordinating multiple agents increases with multimodal inputs, demanding advanced synchronization and communication protocols.

Interpretability and Hallucinations. A key advantage of MLLMs is their ability to generate human-readable language, enabling them to provide explanations for predictions. For example, LLMs can not only identify the location of an anomaly but also indicate its severity and offer a textual explanation (Liu et al., 2024e).

However, the human-readable output also introduces hallucinations (Xu et al., 2025), where MLLMs generate plausible-sounding responses. Notably, GPT-4 has been observed to produce hallucinations in over 21% of time series segments (Dong et al., 2024). Exploring and developing techniques to mitigate these hallucinations remains an important and open research direction.

6 Conclusion

In this survey, we explore the transformative impact of multimodal LLMs on time series analysis, emphasizing diverse modalities of time series. By systematically categorizing existing work from both data and model perspectives, we highlight the versatility of time series representations, ranging from traditional numerical values to text, image, graph, audio, and table modalities. We also focus on representative multimodal LLMs which are applicable or potential to handle these varied modalities of time series. We finally identify several promising research directions and their challenges, including the video modality, reasoning, agent, interpretability, and hallucination. We believe this survey provides valuable insights and advances time series analysis in the era of multimodal LLMs.

Limitations

Several limitations should be acknowledged for this survey. First, despite our rigorous efforts to comprehensively review related work, particularly in the taxonomy of time series modalities, some relevant studies may have been inadvertently overlooked. Second, while we define the scope of this survey, it does not provide an exhaustive collection of multimodal LLMs; we refer readers to related surveys in the main content. Lastly, although we briefly explore new tasks in the future direction, our primary focus remains on established tasks such as forecasting, classification, imputation, and anomaly detection. Investigating novel tasks in the era of multimodal LLMs presents a promising and exciting avenue for future research.

Ethical and Broader Impacts

Our survey provides a novel perspective on bridging time series analysis with multimodal LLMs. We do not foresee any ethical concerns requiring special attention. We believe that this survey will significantly contribute to advancing time series analysis in the era of multimodal LLMs.

References

Josh Achiam et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mohammed S Ahmed and Allen R Cook. 1979. *Analysis of freeway traffic time-series data by using Box-Jenkins techniques*. 722.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. 2024. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*.

Anthropic. 2024. Claude 3 haiku: our fastest model yet. Available at: https://www.anthropic.com/news/claude-3-haiku.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Anastasiya Belyaeva, Justin Cosentino, Farhad Hormozdiari, Krish Eswaran, Shravya Shetty, Greg Corrado, Andrew Carroll, Cory Y McLean, and Nicholas A Furlotte. 2023. Multimodal llms for health grounded in individual-specific data. In *Workshop on Machine Learning for Multimodal Healthcare Data*, pages 86–102. Springer.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The (r) evolution of multimodal large language models: A survey. *arXiv preprint arXiv:2402.12451*.

Yifu Cai, Arjun Choudhry, Mononito Goswami, and Artur Dubrawski. 2024. Timeseriesexam: A time series understanding exam. *arXiv preprint arXiv:2410.14752*.

Defu Cao, Furong Jia, Sercan O Arik, Tomas Pfister, Yixiang Zheng, Wen Ye, and Yan Liu. 2023. Tempo: Prompt-based generative pre-trained transformer for time series forecasting. *arXiv preprint arXiv:2310.04948*.

Di Chai, Leye Wang, and Qiang Yang. 2018. Bike flow prediction with multi-graph convolutional networks. In *Proceedings of the 26th ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 397–400.

Nimeesha Chan, Felix Parker, William Bennett, Tianyi Wu, Mung Yao Jia, James Fackler, and Kimia Ghobadi. 2024. Medtsllm: Leveraging llms for multimodal medical time series analysis. *arXiv preprint arXiv:2408.07773*.

Ching Chang, Wen-Chih Peng, and Tien-Fu Chen. 2023. Llm4ts: Two-stage fine-tuning for time-series forecasting with pre-trained llms. *arXiv preprint arXiv:2308.08469*.

Mouxiang Chen, Lefei Shen, Zhuo Li, Xiaoyun Joy Wang, Jianling Sun, and Chenghao Liu. 2024a. Visionts: Visual masked autoencoders are free-lunch zero-shot time series forecasters. *arXiv preprint arXiv:2408.17253*.

Qian Chen, Yafeng Chen, Yanni Chen, Mengzhe Chen, Yingda Chen, Chong Deng, Zhihao Du, Ruize Gao, Changfeng Gao, Zhifu Gao, et al. 2025. Minmo: A multimodal large language model for seamless voice interaction. *arXiv preprint arXiv:2501.06282*.

Runjin Chen, Tong Zhao, Ajay Jaiswal, Neil Shah, and Zhangyang Wang. 2024b. Llaga: Large language and graph assistant. *arXiv preprint arXiv:2402.08170*.

Yakun Chen, Xianzhi Wang, and Guandong Xu. 2023a. Gatgpt: A pre-trained large language model with graph attention network for spatiotemporal imputation. *arXiv preprint arXiv:2311.14332*.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2023b. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zihan Chen, Lei Nico Zheng, Cheng Lu, Jialu Yuan, and Di Zhu. 2023c. Chatgpt informed graph neural network for stock movement prediction. *arXiv preprint arXiv:2306.03763*.

Mingyue Cheng, Yiheng Chen, Qi Liu, Zhiding Liu, and Yucong Luo. 2024a. Advancing time series classification with multimodal language modeling. *arXiv preprint arXiv:2403.12371*.

Mingyue Cheng, Xiaoyu Tao, Qi Liu, Hao Zhang, Yiheng Chen, and Defu Lian. 2024b. Cross-domain pre-training with language models for transferable time series representations. *arXiv preprint arXiv:2403.12372*.

Xingyi Cheng, Ruiqing Zhang, Jie Zhou, and Wei Xu. 2018. Deeptransport: Learning spatial-temporal dependency for traffic condition forecasting. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Winnie Chow, Lauren Gardiner, Haraldur T Hallgrímsson, Maxwell A Xu, and Shirley You Ren. 2024. Towards time series reasoning with llms. *arXiv preprint arXiv:2409.11376*.

Robert B Cleveland, William S Cleveland, Jean E McRae, Irma Terpenning, et al. 1990. Stl: A seasonal-trend decomposition. *J. off. Stat*, 6(1):3–73.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. 2023. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*.

Mayank Daswani, Mathias MJ Bellaiche, Marc Wilson, Desislav Ivanov, Mikhail Papkov, Eva Schnider, Jing Tang, Kay Lamerigts, Gabriela Botea, Michael A Sanchez, et al. 2024. Plots unlock time-series understanding in multimodal models. *arXiv preprint arXiv:2410.02637*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. 2024. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*.

Manqing Dong, Hao Huang, and Longbing Cao. 2024. Can llms serve as time series anomaly detectors? *arXiv preprint arXiv:2408.03475*.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Vijay Ekambaram, Arindam Jati, Nam H Nguyen, Pankaj Dayama, Chandra Reddy, Wesley M Gifford, and Jayant Kalagnanam. 2024. Ttms: Fast multi-level tiny time mixers for improved zero-shot and few-shot forecasting of multivariate time series. *arXiv preprint arXiv:2401.03955*.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Vasilii Feofanov, Songkang Wen, Marius Alonso, Romain Ilbert, Hongbo Guo, Malik Tiomoko, Lujia Pan, Jianfeng Zhang, and Ievgen Redko. 2025. Mantis: Lightweight calibrated foundation model for user-friendly time series classification. *arXiv preprint arXiv:2502.15637*.

Jean-Yves Franceschi, Edouard Delasalles, Mickaël Chen, Sylvain Lamprier, and Patrick Gallinari. 2020. Stochastic latent residual video prediction. In *International Conference on Machine Learning*, pages 3233–3246. PMLR.

Azul Garza et al. 2023. Timegpt-1. *arXiv preprint arXiv:2310.03589*.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. 2024. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. 2024. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Xusen Guo, Qiming Zhang, Mingxing Peng, Meixin Zhua, et al. 2024. Explainable traffic flow prediction with large language models. *arXiv preprint arXiv:2404.02937*.

James D Hamilton. 2020. *Time series analysis*. Princeton university press.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. 2022. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*.

Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. 2025. The tabular foundation model tabpfn outperforms specialized time series forecasting models based on simple features. *arXiv preprint arXiv:2501.02945*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. 2024. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2256–2264.

Yuxiao Hu, Qian Li, Dongxiao Zhang, Jinyue Yan, and Yuntian Chen. 2025. Context-alignment: Activating and enhancing llm capabilities in time series. *arXiv preprint arXiv:2501.03747*.

Baixiang Huang, Canyu Chen, Xiongxiao Xu, Ali Payani, and Kai Shu. 2024a. Can knowledge editing really correct hallucinations? *arXiv preprint arXiv:2410.16251*.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. 2024b. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23802–23804.

Yiheng Huang, Xiaowei Mao, Shengnan Guo, Yubin Chen, Junfeng Shen, Tiankuo Li, Youfang Lin, and Huaiyu Wan. 2024c. Std-plm: Understanding both spatial and temporal properties of spatial-temporal data with plm. *arXiv preprint arXiv:2407.09096*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Furong Jia, Kevin Wang, Yixiang Zheng, Defu Cao, and Yan Liu. 2024. Gpt4mts: Prompt-based large language model for multimodal time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23343–23351.

Yiren Jian, Chongyang Gao, and Soroush Vosoughi. 2024. Bootstrapping vision-language learning with decoupled language pre-training. *Advances in Neural Information Processing Systems*, 36.

Yushan Jiang, Zijie Pan, Xikun Zhang, Sahil Garg, Anderson Schneider, Yuriy Nevmyvaka, and Dongjin Song. 2024. Empowering time series analysis with large language models: A survey. *arXiv preprint arXiv:2402.03182*.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. 2023. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*.

Xuexiang Jin, Yi Zhang, and Danya Yao. 2007. Simultaneously prediction of network traffic flow based on pca-svr. In *Advances in Neural Networks–ISNN 2007: 4th International Symposium on Neural Networks, ISNN 2007, Nanjing, China, June 3-7, 2007, Proceedings, Part II 4*, pages 1022–1031. Springer.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. 2021a. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021b. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Yaxuan Kong, Yiyuan Yang, Shiyu Wang, Chenghao Liu, Yuxuan Liang, Ming Jin, Stefan Zohren, Dan Pei, Yan Liu, and Qingsong Wen. 2025. Position: Empowering time series reasoning with multimodal llms. *arXiv preprint arXiv:2502.01477*.

Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager. 2016. Temporal convolutional networks: A unified approach to action segmentation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 47–54. Springer.

Geon Lee, Wenchao Yu, Kijung Shin, Wei Cheng, and Haifeng Chen. 2025. Timecap: Learning to contextualize, augment, and predict time series events with large language model agents. *arXiv preprint arXiv:2502.11418*.

János Levendovszky and Farhad Kia. 2012. Prediction based–high frequency trading on financial time series. *Periodica Polytechnica Electrical Engineering and Computer Science*, 56(1):29–34.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation

for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jun Li, Che Liu, Sibo Cheng, Rossella Arcucci, and Shenda Hong. 2024a. Frozen language model helps ecg zero-shot learning. In *Medical Imaging with Deep Learning*, pages 402–415. PMLR.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Peisen Li, Yizhe Pang, and Junyu Ren. 2024b. Spatio-temporal graph convolutional network combined large language model: A deep learning framework for bike demand forecasting. *arXiv preprint arXiv:2403.15733*.

Zekun Li, Shiyang Li, and Xifeng Yan. 2024c. Time series as images: Vision transformer for irregularly sampled time series. *Advances in Neural Information Processing Systems*, 36.

Zhonghang Li, Lianghao Xia, Jiabin Tang, Yong Xu, Lei Shi, Long Xia, Dawei Yin, and Chao Huang. 2024d. Urbangpt: Spatio-temporal large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5351–5362.

Zihao Li, Xiao Lin, Zhining Liu, Jiaru Zou, Ziwei Wu, Lecheng Zheng, Dongqi Fu, Yada Zhu, Hendrik Hamann, Hanghang Tong, et al. 2025. Language in the flow of time: Time-series-paired texts weaved into a unified temporal narrative. *arXiv preprint arXiv:2502.08942*.

Yueqing Liang, Liangwei Yang, Chen Wang, Xiongxiao Xu, Philip S Yu, and Kai Shu. 2024. Taxonomy-guided zero-shot recommendations with llms. *arXiv preprint arXiv:2406.14043*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's Verify Step by Step. *arXiv preprint arXiv:2305.20050*.

Chenxi Liu, Sun Yang, Qianxiong Xu, Zhishuai Li, Cheng Long, Ziyue Li, and Rui Zhao. 2024a. Spatial-temporal large language model for traffic prediction. In *2024 25th IEEE International Conference on Mobile Data Management (MDM)*, pages 31–40. IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Kamarthi, Aditya B Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, et al. 2024c. Time-mmd: A new multi-domain multimodal dataset for time series analysis. *arXiv preprint arXiv:2406.08627*.

Haoxin Liu, Zhiyuan Zhao, Jindong Wang, Harshavardhan Kamarthi, and B Aditya Prakash. 2024d. Lst-prompt: Large language models as zero-shot time series forecasters by long-short-term prompting. *arXiv preprint arXiv:2402.16132*.

Jun Liu, Chaoyun Zhang, Jiaxu Qian, Minghua Ma, Si Qin, Chetan Bansal, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. 2024e. Large language models can deliver accurate and interpretable time series anomaly detection. *arXiv preprint arXiv:2405.15370*.

Lei Liu, Shuo Yu, Runze Wang, Zhenxun Ma, and Yanming Shen. 2024f. How can large language models understand spatial-temporal data? *arXiv preprint arXiv:2401.14192*.

Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Shun Liao, Paolo Di Achille, and Shwetak Patel. 2023. Large language models are few-shot health learners. *arXiv preprint arXiv:2305.15525*.

Xu Liu, Junfeng Hu, Yuan Li, Shizhe Diao, Yuxuan Liang, Bryan Hooi, and Roger Zimmermann. 2024g. Unitime: A language-empowered unified model for cross-domain time series forecasting. In *Proceedings of the ACM on Web Conference 2024*, pages 4095–4106.

Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. 2024h. Autotimes: Autoregressive time series forecasters via large language models. *Advances in Neural Information Processing Systems*, 37:122154–122184.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. 2024i. Timer: Transformers for time series analysis at scale. *arXiv preprint arXiv:2402.02368*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Alejandro Lopez-Lira and Yuehua Tang. 2023. Can chatgpt forecast stock price movements? return predictability and large language models. *arXiv preprint arXiv:2304.07619*.

Tianmi Ma, Jiawei Du, Wenxin Huang, Wenjie Wang, Liang Xie, Xian Zhong, and Joey Tianyi Zhou. 2025. Llm knows geometry better than algebra: Numerical understanding of llm-based agents in a trading arena. *arXiv preprint arXiv:2502.17967*.

Rishabh Madan and Partha Sarathi Mangipudi. 2018. Predicting computer network traffic: a time series forecasting approach using dwt, arima and rnn. In *2018 Eleventh International Conference on Contemporary Computing (IC3)*, pages 1–5. IEEE.

Shoji Makino. 2018. *Audio source separation*, volume 433. Springer.

Mike A Merrill, Mingtian Tan, Vinayak Gupta, Tom Hartvigsen, and Tim Althoff. 2024. Language models still struggle to zero-shot reason about time series. *arXiv preprint arXiv:2404.11757*.

Seyed Mohamad Moghadas, Yangxintong Lyu, Bruno Cornelis, Alexandre Alahi, and Adrian Munteanu. 2024. Strada-llm: Graph llm for traffic prediction. *arXiv preprint arXiv:2410.20856*.

Jingchao Ni, Ziming Zhao, ChengAo Shen, Hanghang Tong, Dongjin Song, Wei Cheng, Dongsheng Luo, and Haifeng Chen. 2025. Harnessing vision models for time series analysis: A survey. *arXiv preprint arXiv:2502.08869*.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2022. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*.

Zijie Pan, Yushan Jiang, Sahil Garg, Anderson Schneider, Yuriy Nevmyvaka, and Dongjin Song. 2024. s^2 ip-llm: Semantic space informed prompt learning with llm for time series forecasting. In *Forty-first International Conference on Machine Learning*.

Vinay Prithyani, Mohsin Mohammed, Richa Gadgil, Ricardo Buitrago, Vinija Jain, and Aman Chadha. 2024. On the feasibility of vision-language models for time-series classification. *arXiv preprint arXiv:2412.17304*.

Dayu Qin, Yi Yan, and Ercan Engin Kuruoglu. 2024. Llm-based online prediction of time-varying graph signals. *arXiv preprint arXiv:2410.18718*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI blog*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Syama Sundar Rangapuram, Matthias W Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski. 2018. Deep state space models for time series forecasting. *Advances in neural information processing systems*, 31.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Biloš, Hena Ghonia, Nadhir Hassen, Anderson Schneider, et al. 2023. Lag-llama: Towards foundation models for time series forecasting. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

Chidaksh Ravuru, Sagar Srinivas Sakhinana, and Venkataramana Runkana. 2024. Agentic retrieval-augmented generation for time series analysis. *arXiv preprint arXiv:2408.14484*.

Mohammed Yousef Shaheen. 2021. Applications of artificial intelligence (ai) in healthcare: A review. *ScienceOpen Preprints*.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. 2024. Time-moe: Billion-scale time series foundation models with mixture of experts. *arXiv preprint arXiv:2409.16040*.

Aofeng Su, Aowen Wang, Chao Ye, Chen Zhou, Ga Zhang, Guangcheng Zhu, Haobo Wang, Haokai Xu, Hao Chen, Haoze Li, et al. 2024a. Tablegpt2: A large multimodal model with tabular data integration. *arXiv preprint arXiv:2411.02059*.

Zhaochen Su, Jun Zhang, Tong Zhu, Xiaoye Qu, Juntao Li, Min Zhang, and Yu Cheng. 2024b. Timo: Towards better temporal reasoning for language models. *arXiv preprint arXiv:2406.14192*.

Chenxi Sun, Shenda Hong, Moxian Song, and Hongyan Li. 2020. A review of deep learning methods for irregularly sampled medical time series data. *arXiv preprint arXiv:2010.12493*.

Chenxi Sun, Hongyan Li, Yaliang Li, and Shenda Hong. 2023. Test: Text prototype aligned embedding to activate llm's ability for time series. *arXiv preprint arXiv:2308.08241*.

Ling Tan, Ji Guo, Selvarajah Mohanarajah, and Kun Zhou. 2021. Can we detect trends in natural disaster management with artificial intelligence? a review of modeling practices. *Natural Hazards*, 107:2389–2417.

Hua Tang, Chong Zhang, Mingyu Jin, Qinkai Yu, Zhenting Wang, Xiaobo Jin, Yongfeng Zhang, and Mengnan Du. 2025. Time series forecasting with llms: Understanding and enhancing model capabilities. *ACM SIGKDD Explorations Newsletter*, 26(2):109–118.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

George Tzanetakis and Perry Cook. 2002. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Chengsen Wang, Qi Qi, Jingyu Wang, Haifeng Sun, Zirui Zhuang, Jinming Wu, Lei Zhang, and Jianxin Liao. 2024a. Chattime: A unified multimodal time series foundation model bridging numerical and textual data. *arXiv preprint arXiv:2412.11376*.

Jiahao Wang, Mingyue Cheng, Qingyang Mao, Qi Liu, Feiyang Xu, Xin Li, and Enhong Chen. 2024b. Tabletime: Reformulating time series classification as zero-shot table understanding via large language models. *arXiv preprint arXiv:2411.15737*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024c. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Senzhang Wang, Jiannong Cao, and S Yu Philip. 2020. Deep learning for spatio-temporal data mining: A survey. *IEEE transactions on knowledge and data engineering*, 34(8):3681–3700.

Weijie Wang, Gaopeng Zhang, Luming Yang, VS Balaji, V Elamaran, and N Arunkumar. 2019. Revisiting signal processing with spectrogram analysis on eeg, ecg and speech signals. *Future Generation Computer Systems*, 98:227–232.

Xinglei Wang, Meng Fang, Zichao Zeng, and Tao Cheng. 2023. Where would i go next? large language models as human mobility predictors. *arXiv preprint arXiv:2308.15197*.

Yujie Wang, Xin Du, Zhihui Lu, Qiang Duan, and Jie Wu. 2022. Improved lstm-based time-series anomaly detection in rail transit operation environments. *IEEE Transactions on Industrial Informatics*, 18(12):9027–9036.

Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. 2022. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*.

Christopher Wimmer and Navid Rekabsaz. 2023. Leveraging vision-language models for granular market change prediction. *arXiv preprint arXiv:2301.10166*.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. 2024. Unified training of universal time series forecasting transformers. *arXiv preprint arXiv:2402.02592*.

Yuankai Wu, Huachun Tan, Lingqiao Qin, Bin Ran, and Zhuxi Jiang. 2018. A hybrid deep learning based traffic flow prediction method and its understanding. *Transportation Research Part C: Emerging Technologies*, 90:166–180.

Mengxi Xiao, Zihao Jiang, Lingfei Qian, Zhengyu Chen, Yueru He, Yijing Xu, Yuecheng Jiang, Dong Li, Ruey-Ling Weng, Min Peng, et al. 2025. Retrieval-augmented large language models for financial time series forecasting. *arXiv preprint arXiv:2502.05878*.

Qianqian Xie, Weiguang Han, Yanzhao Lai, Min Peng, and Jimin Huang. 2023. The wall street neophyte: A zero-shot analysis of chatgpt over multimodal stock movement prediction challenges. *arXiv preprint arXiv:2304.05351*.

Haojun Xu, Yan Gao, Zheng Hui, Jie Li, and Xinbo Gao. 2023a. Language knowledge-assisted representation learning for skeleton-based action recognition. *arXiv preprint arXiv:2305.12398*.

Qinyi Xu, Chunxiao Jiang, Yi Han, Beibei Wang, and KJ Ray Liu. 2017. Waveforming: An overview with beamforming. *IEEE Communications Surveys & Tutorials*, 20(1):132–149.

Xiongxiao Xu, Canyu Chen, Yueqing Liang, Baixiang Huang, Guangji Bai, Liang Zhao, and Kai Shu. 2024. Sst: Multi-scale hybrid mamba-transformer experts for long-short range time series forecasting. *arXiv preprint arXiv:2404.14757*.

Xiongxiao Xu, Haoran Wang, Yueqing Liang, Philip S Yu, Yue Zhao, and Kai Shu. 2025. Can multimodal llms perform time series anomaly detection? *arXiv preprint arXiv:2502.17812*.

Xiongxiao Xu, Xin Wang, Elkin Cruz-Camacho, Christopher D. Carothers, Kevin A. Brown, Robert B. Ross, Zhiling Lan, and Kai Shu. 2023b. Machine learning for interconnect network traffic forecasting: Investigation and exploitation. In *Proceedings of the 2023 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, pages 133–137.

Hao Xue and Flora D Salim. 2023. Promptcast: A new prompt-based learning paradigm for time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*.

Hao Xue, Bhanu Prakash Voutharoja, and Flora D Salim. 2022. Leveraging language foundation models for human mobility forecasting. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, pages 1–9.

Yi Yan, Dayu Qin, and Ercan Engin Kuruoglu. 2024. Llm online spatial-temporal signal reconstruction under noise. *arXiv preprint arXiv:2411.15764*.

Chao-Han Huck Yang, Yun-Yun Tsai, and Pin-Yu Chen. 2021. Voice2series: Reprogramming acoustic models for time series classification. In *International conference on machine learning*, pages 11808–11819. PMLR.

Haotong Yang, Yi Hu, Shijia Kang, Zhouchen Lin, and Muhan Zhang. 2024a. Number cookbook: Number understanding of language models and how to improve it. *arXiv preprint arXiv:2411.03766*.

Luoxiao Yang, Yun Wang, Xinqi Fan, Israel Cohen, Jingdong Chen, Yue Zhao, and Zijun Zhang. 2024b. Vitime: A visual intelligence-based foundation model for time series forecasting. *arXiv preprint arXiv:2407.07311*.

Silin Yang, Dong Wang, Haoqi Zheng, and Ruochun Jin. 2024c. Timerag: Boosting llm time series forecasting via retrieval-augmented generation. *arXiv preprint arXiv:2412.16643*.

Huaxiu Yao, Xianfeng Tang, Hua Wei, Guanjie Zheng, and Zhenhui Li. 2019. Revisiting spatial-temporal similarity: A deep learning framework for traffic prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5668–5675.

A Yarkın Yıldız, Emirhan Koç, and Aykut Koç. 2022. Multivariate time series imputation with transformers. *IEEE Signal Processing Letters*, 29:2517–2521.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Xinli Yu, Zheng Chen, Yuan Ling, Shujing Dong, Zongyi Liu, and Yanbin Lu. 2023. Temporal data meets llm–explainable financial time series forecasting. *arXiv preprint arXiv:2306.11025*.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. 2023. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128.

Zhen Zeng, Tucker Balch, and Manuela Veloso. 2021. Deep video prediction for time series forecasting. In *Proceedings of the Second ACM International Conference on AI in Finance*, pages 1–7.

Dengsheng Zhang and Dengsheng Zhang. 2019. Wavelet transform. *Fundamentals of image data mining: Analysis, Features, Classification and Retrieval*, pages 35–44.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024a. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Haoran Zhang, Yong Liu, Yunzhong Qiu, Haixuan Liu, Zhongyi Pei, Jianmin Wang, and Mingsheng Long. 2025. Timesbert: A bert-style foundation model for time series understanding. *arXiv preprint arXiv:2502.21245*.

Huanyu Zhang, Chang Xu, Yi-Fan Zhang, Zhang Zhang, Liang Wang, Jiang Bian, and Tieniu Tan. 2024b. Timeraf: Retrieval-augmented foundation model for zero-shot time series forecasting. *arXiv preprint arXiv:2412.20810*.

Lianming Zhang, Huan Zhang, Qian Tang, Pingping Dong, Zhen Zhao, Yehua Wei, Jing Mei, and Kaiping Xue. 2020. Lntp: An end-to-end online prediction model for network traffic. *IEEE Network*, 35(1):226–233.

Qianru Zhang, Xubin Ren, Lianghao Xia, Siu Ming Yiu, and Chao Huang. 2023b. Spatio-temporal graph learning with large language model.

Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun. 2023c. Tablellama: Towards open large generalist models for tables. *arXiv preprint arXiv:2311.09206*.

Xiyuan Zhang, Ranak Roy Chowdhury, Rajesh K Gupta, and Jingbo Shang. 2024c. Large language models for time series: A survey. *arXiv preprint arXiv:2402.01801*.

Yunkai Zhang, Yawen Zhang, Ming Zheng, Kezhen Chen, Chongyang Gao, Ruian Ge, Siyuan Teng, Amine Jelloul, Jinmeng Rao, Xiaoyuan Guo, et al. 2023d. Insight miner: A time series analysis dataset for cross-domain alignment with natural language. In *NeurIPS 2023 AI for Science Workshop*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qingsong Wen, and Yuxuan Liang. 2025. Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting. *arXiv preprint arXiv:2502.04395*.

Tian Zhou, Peisong Niu, Liang Sun, Rong Jin, et al. 2023. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems*, 36:43322–43355.

Zihao Zhou and Rose Yu. 2024. Can llms understand time series anomalies? *arXiv preprint arXiv:2410.05440*.

Jiawei Zhu, Xing Han, Hanhan Deng, Chao Tao, Ling Zhao, Pu Wang, Tao Lin, and Haifeng Li. 2022. Kst-gcn: A knowledge-driven spatial-temporal graph convolutional network for traffic forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):15055–15065.

Jiaxin Zhuang, Leon Yan, Zhenwei Zhang, Ruiqi Wang, Jiawei Zhang, and Yuantao Gu. 2024. See it, think it, sorted: Large multimodal models are few-shot time series anomaly analyzers. *arXiv preprint arXiv:2411.02465*.

Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. 2018. Eco: Efficient convolutional network for online video understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 695–712.