



August 3-7, 2025

# Multi-modal Time Series Analysis

## — Data, Methods, and Applications

Survey Paper



Github



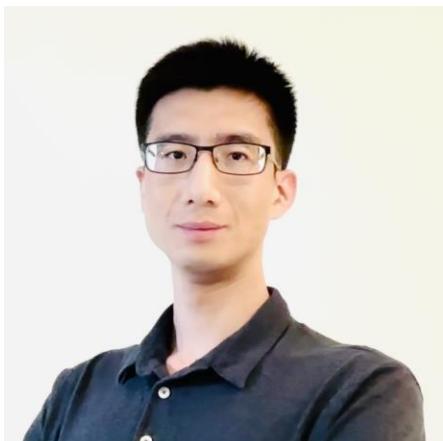
**NEC**  
NEC Laboratories America

Morgan  
Stanley

# Presenters



**Dongjin Song**  
Associate Professor  
School of Computing  
University of Connecticut



**Jingchao Ni**  
Assistant Professor  
Department of Computer Science  
University of Houston



**Zijie Pan**  
Ph.D. Student  
School of Computing  
University of Connecticut



**Haifeng Chen**  
Department Head  
Data Science & System Security  
NEC Labs America

# Contributors



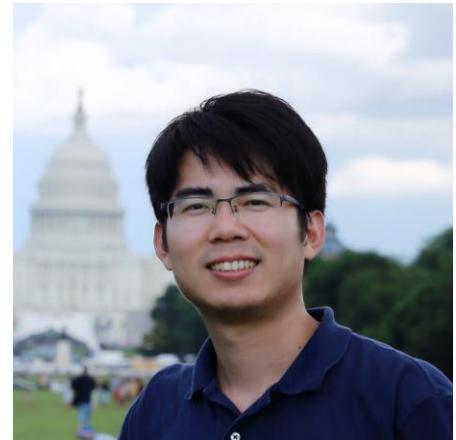
**Yushan Jiang**  
Ph.D. Student  
School of Computing  
University of Connecticut



**Kanghui Ning**  
Ph.D. Student  
School of Computing  
University of Connecticut



**Xuyang Shen**  
Ph.D. Student  
School of Computing  
University of Connecticut



**Wenchao Yu**  
Senior Researcher  
Data Science & System Security  
NEC Labs America



**Anderson Schneider**  
Executive Director  
Machine Learning Research  
Morgan Stanley



**Yuriy Nevmyvaka**  
Managing Director  
Machine Learning Research  
Morgan Stanley

# Agenda

- **Part 1: Opening and Introduction (10 min)**
- **Part 2: Multi-modal Time Series Datasets (30 min)**
- **Part 3: Taxonomy of Multi-modal Time Series Methods (60 min)**

---

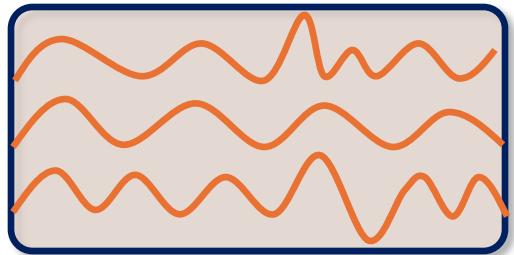
Break (20 min)

---

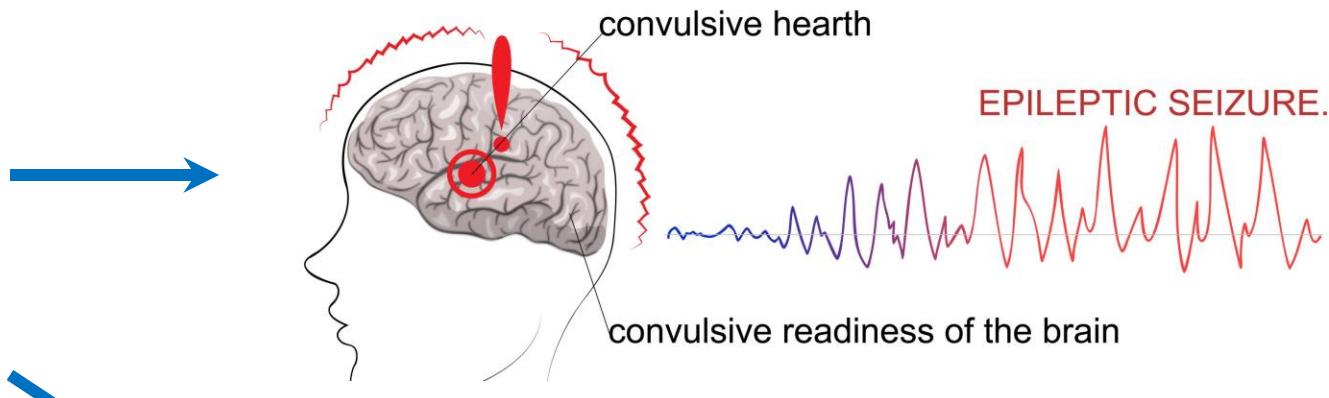
- **Part 4: Multi-modal Time Series Applications (40 min)**
- **Part 5: Future Directions and Q&A (20 min)**

# *Introduction to Multi-modal Time Series Analysis*

# Background – Time Series Analysis



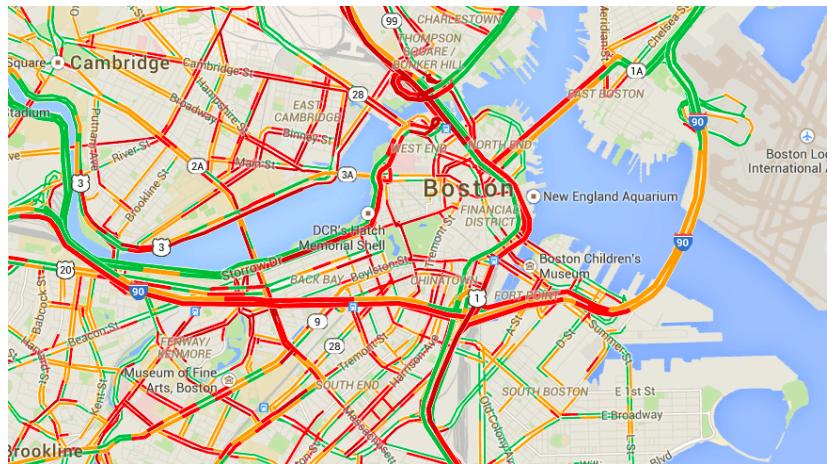
Time series data



Healthcare



Electricity load & Power consumption



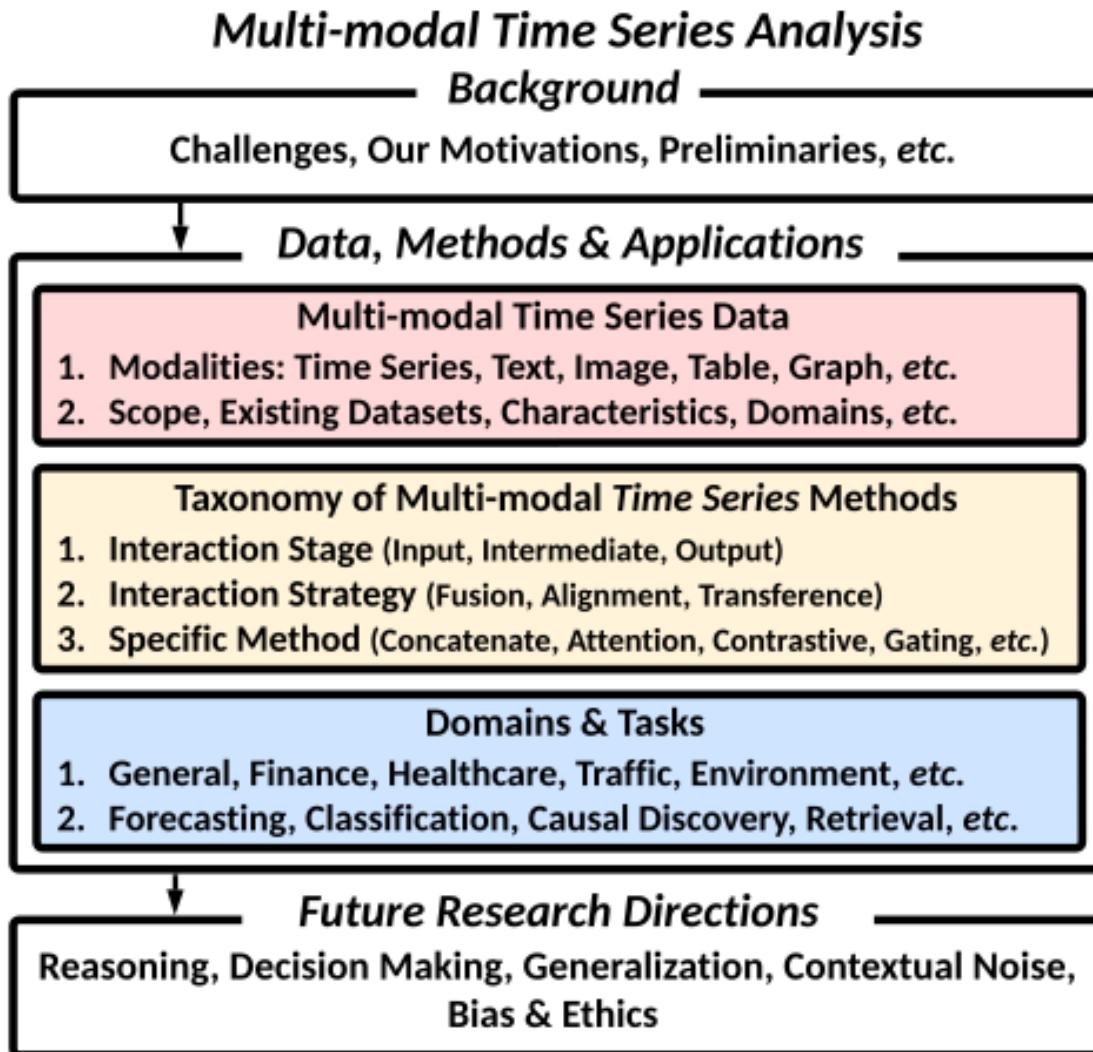
Traffic networks

# Background – Multi-modal Time Series Analysis

## Scope of our tutorial

1. We mainly consider standard time series and spatial time series.
  - Spatial structures (often represented as graphs) are inherently paired and **not treated** as a separate modality.
2. We focus on multi-modal methods for a spectrum of tasks:
  - For predictive tasks, the focus is to leverage **multi-modal inputs** from real-world contexts
  - For generation and retrieval tasks, the focus is more on **transforming** the input modality to another output modality.

# Background – Multi-modal Time Series Analysis



- We uniquely categorize the existing methods into a unified cross-modal interaction framework, highlighting fusion, alignment, and transference at the input/intermediate/output levels.
- We discuss real-world applications of multi-modal time series and identify promising future directions, encouraging researchers and practitioners to explore and exploit multi-modal time series.

# *Multi-modal Time Series Datasets*

# Multi-modal Time Series Datasets

We collect 28 open-source multi-modal time series datasets:

- 1) Include 5 modalities: TS (Time Series), Text, Table, Image, ST (Spatial Time Series)
- 2) Include 7 combinations:
  - TS, Text
  - TS, Image
  - TS, Text, Image
  - TS, Text , Table
  - TS, Image, Text, Table
  - ST, Text
  - ST, Text, Image

# Multi-modal Time Series Datasets – Overview

**Table 1: Representative open-source multi-modal time series datasets and across domains.**

Domain	Dataset (Superscripts include the URLs to the datasets)	Modalities
Healthcare	MIMIC-III [36] <sup>[1]</sup> , MIMIC-IV [35] <sup>[2]</sup>	TS, Text, Table
	ICBHI [67] <sup>[3]</sup> , Coswara [4] <sup>[4]</sup> , KAUH [22] <sup>[5]</sup> , PTB-XL [73] <sup>[6]</sup> , ZuCo [15, 27] <sup>[7]</sup>	TS, Text
	Image-EEG [23] <sup>[8]</sup>	TS, Image
Finance	FNSPID [18] <sup>[9]</sup> , AlphaFin [47] <sup>[10]</sup> , ACL18 [86] <sup>[11]</sup> , CIKM18 [80] <sup>[12]</sup> , DOW30 [12] <sup>[13]</sup>	TS, Text
	FinMultiTime [85] <sup>[14]</sup>	TS, Image, Text, Table
Retail	VISUELLE [72] <sup>[15]</sup>	TS, Image, Text
IoT	LEMMA-RCA [41] <sup>[16]</sup>	TS, Text
Speech	LRS3 [1] <sup>[17]</sup> , VoxCeleb2 [14] <sup>[18]</sup>	TS (Audio), Image
Traffic	NYC-taxi, NYC-bike [50] <sup>[19]</sup>	ST, Text
Environment	Terra [11] <sup>[20]</sup>	ST, Text
Multi-domain	Time-MMD [55] <sup>[21]</sup> , TimeCAP [43] <sup>[22]</sup> , NewsForecast [75] <sup>[23]</sup> , TTC [38] <sup>[24]</sup> , CiK [78] <sup>[25]</sup> , TSQA [39] <sup>[26]</sup> , MTBench [7] <sup>[27]</sup>	TS, Text
	MoTime [100] <sup>[28]</sup>	TS, Text, Image

# Multi-modal Time Series Datasets – TS, Text

ZuCo (Zurich Cognitive Language Processing Corpus) benchmark on cross-subject reading task classification with EEG and eye-tracking data

## TS:

- EEG
- Eye-tracking

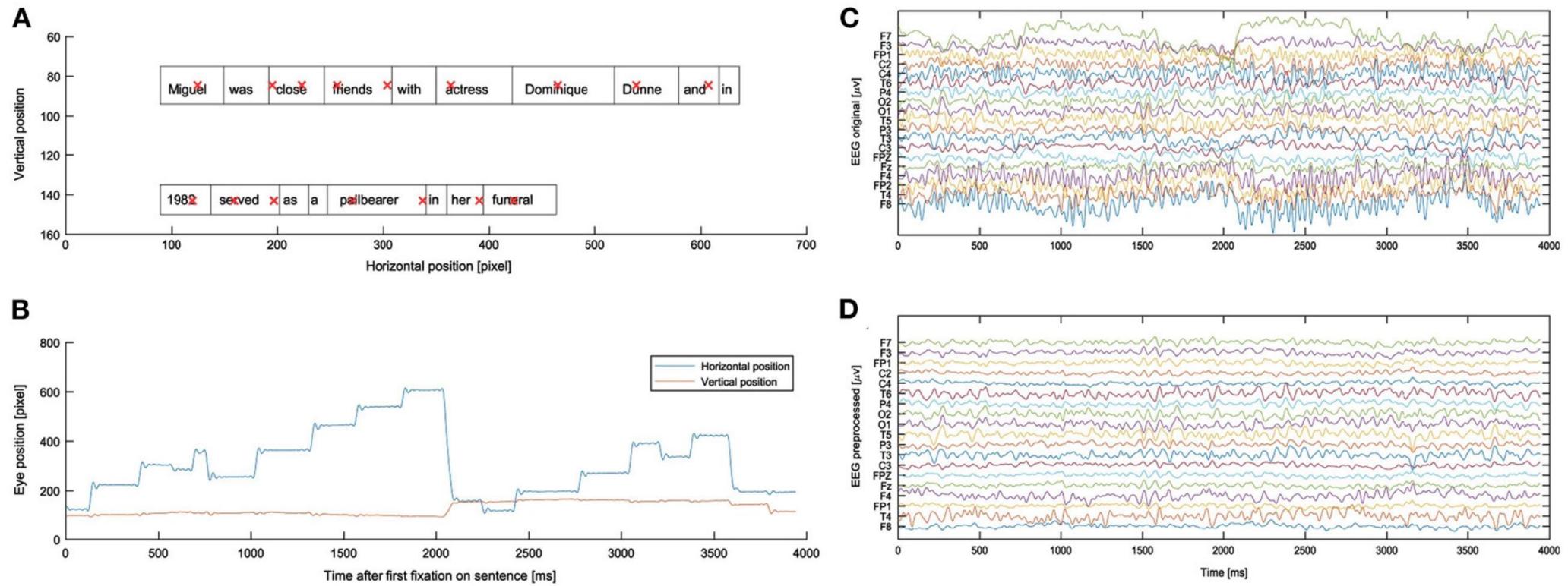
## Text: Reading materials

- 16 Participants, 10 female, 6 male
- 2 Task: Normal Reading & Task Specific Reading

TABLE 1 Descriptive statistics of reading materials (SD, standard deviation), including Flesch readability scores.

	NR	TSR
Sentences	349	390
Sent. length	Mean (SD), range	Mean (SD), range
	19.6 (8.8), 5–53	21.3 (9.5), 5–53
Total words	6,828	8,310
Word types	2,412	2,437
Word length	Mean (SD), range	Mean (SD), range
	4.9 (2.7), 1–29	4.9 (2.7), 1–21
Flesch score	55.38	50.76

# Multi-modal Time Series Datasets – TS, Text



# Multi-modal Time Series Datasets – TS, Text

FNSPID: A Comprehensive Financial News Dataset in Time Series

**TS:** Stock prices

**Text:** Financial news

- 29.7 million stock prices
- 15.7 million time-aligned financial news records
- 4,775 S&P500 companies, covering the period from 1999 to 2023

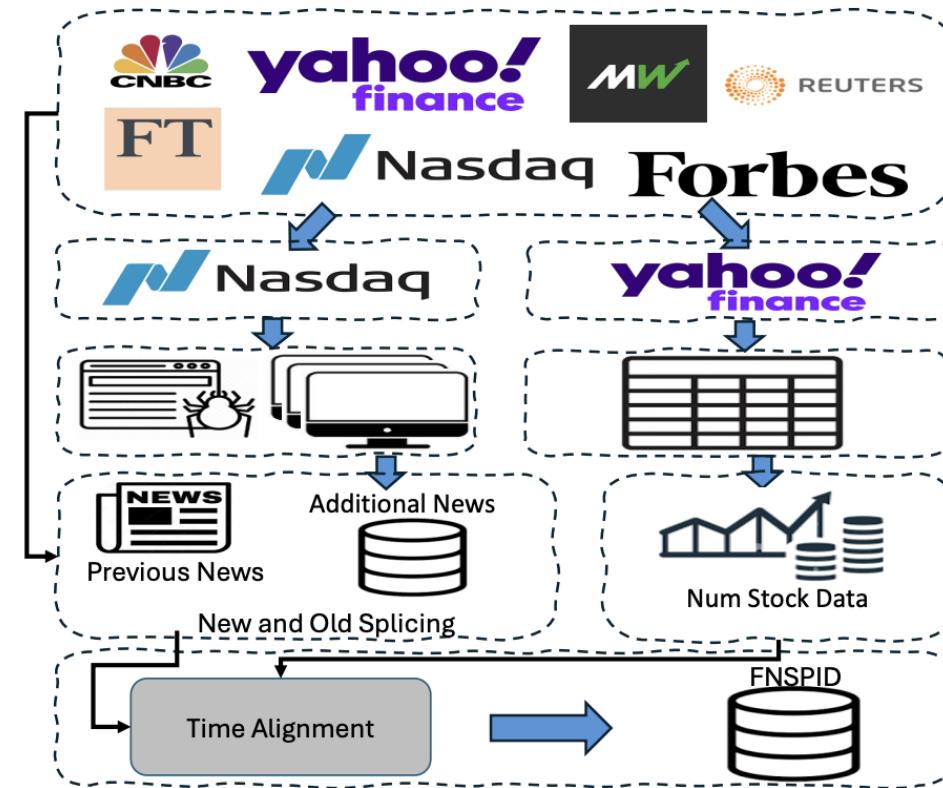
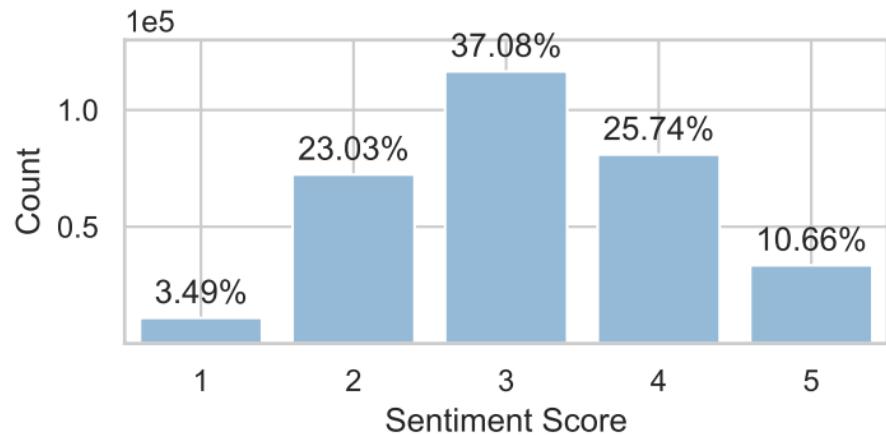


Figure 1: Data Collection Process from website selection in the first level box; data segmentation in second level boxes; data collection for web scraping on left and numerical data collection on right; data organization on fourth level boxes and final FNSPID build-up on the last level box.

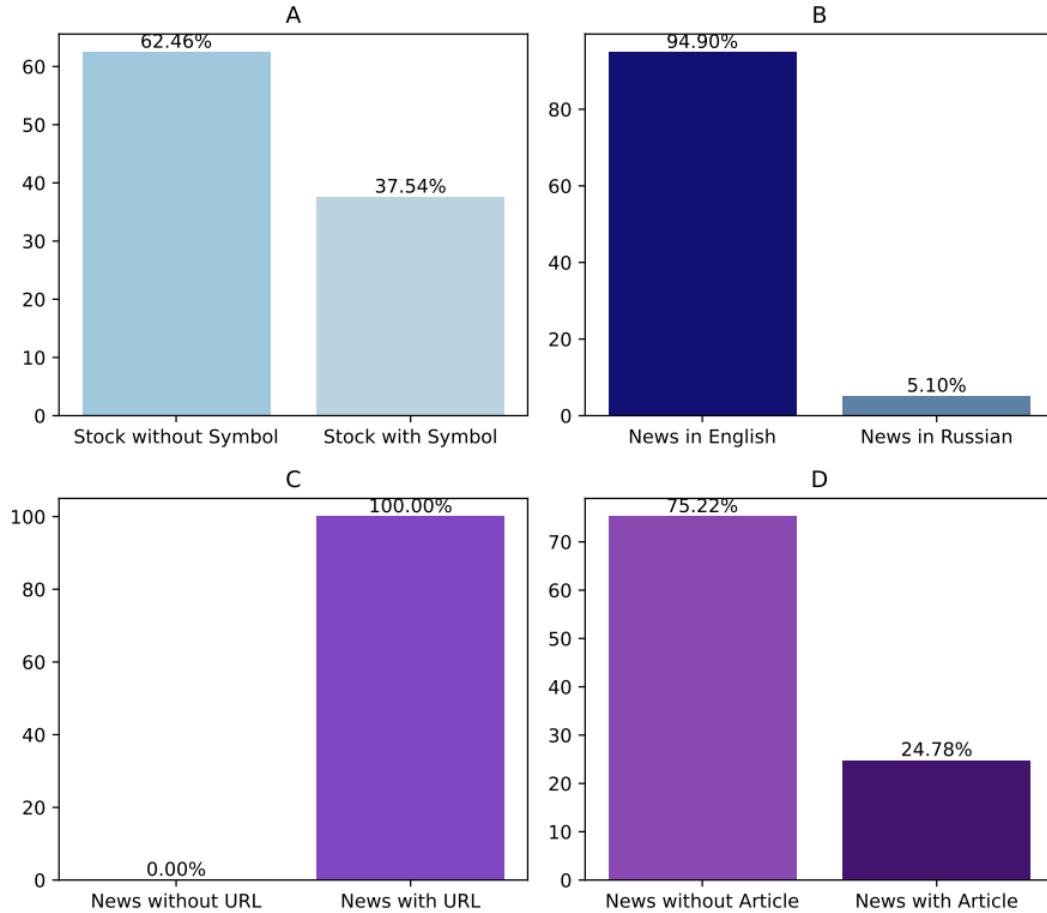
# Multi-modal Time Series Datasets – TS. Text

Date	Open	High	Low	Close	Adj.	Volume
2023-12-28 00:00:00	194.14	194.66	193.17	193.58	193.58	34014500
2023-12-27 00:00:00	192.49	193.50	191.09	193.15	193.15	48087700
2023-12-26 00:00:00	193.61	193.89	192.83	193.05	193.05	28919300
...	...	...	...	...	...	...

**Table 2: Stock Numerical Data:** 'Open' represents the opening stock price, 'High' indicates the highest price within the day, 'Low' signifies the lowest price within the day, 'Adj Close' represents the close price adjusted for dividends, and 'Volume' denotes the number of shares traded.



**Figure 4: Sentiment Distribution:** 1 is negative, 2 is somewhat negative, 3 is neutral, 4 is somewhat positive, 5 is positive



**Figure 5: Statistical Overview:** In A, we provide information on news articles that include the stock symbol. The B displays the language distribution, encompassing English and Russian. In C, a comparison of the included URLs is presented. Finally, in the D, details are provided on the news text already incorporated in the dataset, along with potential expansions into additional text data.

# Multi-modal Time Series Datasets – TS, Image

VoxCeleb: A large scale audio-visual dataset of human speech

**TS:** Audio

**Image:** Short clips of human speech

- VoxCeleb1: over 150,000 utterances from 1251 celebrities
- VoxCeleb2: over 1,000,000 utterances from 6112 celebrities

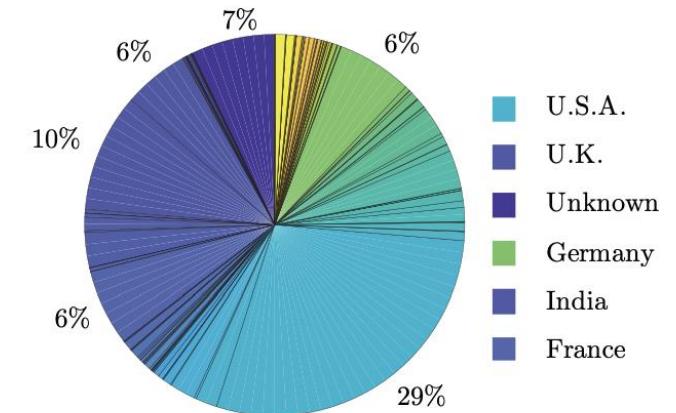
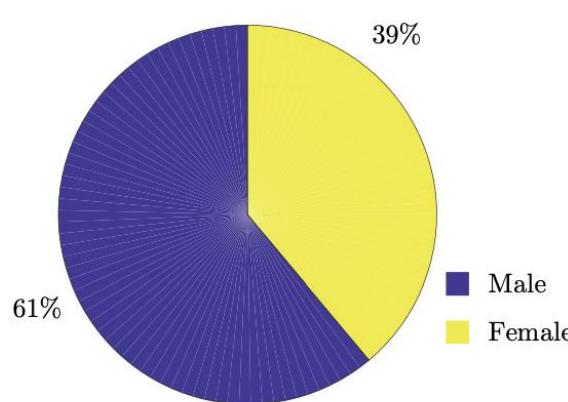
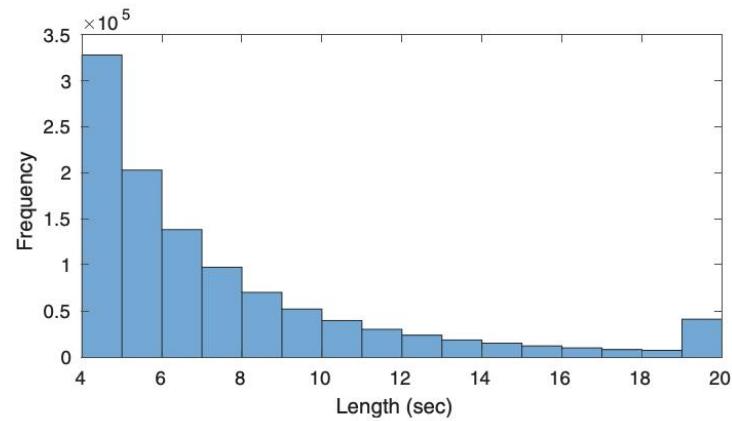
**Table 2**

Dataset statistics for both VoxCeleb1 and VoxCeleb2. Note VoxCeleb2 is more than 5 times larger than VoxCeleb1.

Dataset	VoxCeleb1	VoxCeleb2
# of speakers	1251	6112
# of male speakers	690	3761
# of videos	22,496	150,480
# of hours	352	2442
# of utterances	153,516	1,128,246
Avg # of videos per speaker	18	25
Avg # of utterances per speaker	116	185
Avg length of utterances (s)	8.2	7.8

# Multi-modal Time Series Datasets – TS, Image

VoxCeleb: A large scale audio-visual dataset of human speech



**Fig. 1.** Top row: Examples from the VoxCeleb2 dataset. We show cropped faces of some of the speakers in the dataset. Both audio and face detections are provided. Bottom row: (left) distribution of utterance lengths in the dataset – lengths shorter than 20s are binned in 1s intervals and all utterances of 20s+ are binned together; (middle) gender distribution and (right) nationality distribution of speakers. For readability, the percentage frequencies of only the top-5 nationalities are shown. Best viewed zoomed in and in colour.

# Multi-modal Time Series Datasets – TS, Text, Image

MoTime: A Dataset Suite for Multimodal Time Series Forecasting

Systematically re-purposing and transforming existing datasets.

Spanning e-commerce, web traffic, media, and user behavior domains

Table 2: Statistics of the eight multimodal time series datasets in MoTime.

Dataset	TS Shape	Density(%)	Text	Image	Metadata	Notes
PixelRec	4,865 × 43,082	4.41	✓	✓	✓	Long sparse multivariate TS
TaobaoFashion	365 × 890	68.01	–	✓	–	One image per item
MovieLens	10,505 × 84,518	1.66	✓	–	✓	Text scraped externally
AmazonReview	3,934 × 668,756	6.18	✓	–	✓	29 categories, sparse TS
Tianchi	365 × 36,397	53.15	✓	✓	–	E-commerce purchase logs
News	144 × 26,612	17.61	✓	–	✓	20-min interval resolution
WikiPeople	550 × 3,856	99.96	✓	–	✓	Multichannel access modes
VISUELLE	11 × 5,355	62.48	✓	✓	✓	Irregular time series

# Multi-modal Time Series Datasets – TS, Text, Image

Table 6: Descriptive statistics of time series values per dataset. The minimum value here refers to the smallest positive value that is neither zero nor -1.

<b>Dataset</b>	<b>Median</b>	<b>Mean</b>	<b>Min</b>	<b>Max</b>
PixelRec	2	13.87	1	3196
Tianchi	7	46.74	1	90472
MovieLens	1	1.08	1	549
News	1	11.63	1	13291
TaobaoFashion	3	5.36	1	966
WikiPeople	921.25	3165.67	1	5816910
AmazonReview	1	1.24	1	6311

# Multi-modal Time Series Datasets – TS, Text, Image

VISUELLE: A public dataset for new fashion product sales forecasting

**TS:** Sales time series; Google Trends

**Text:** Associated tags of the product

**Image:** The clothing item on a white background

Composed of 5577 new products and about 45M sales related to fashion seasons from 2016-2019



Figure 1: Sample images representing various product categories within the VISUELLE dataset.

# Multi-modal Time Series Datasets - TS, Text, Image

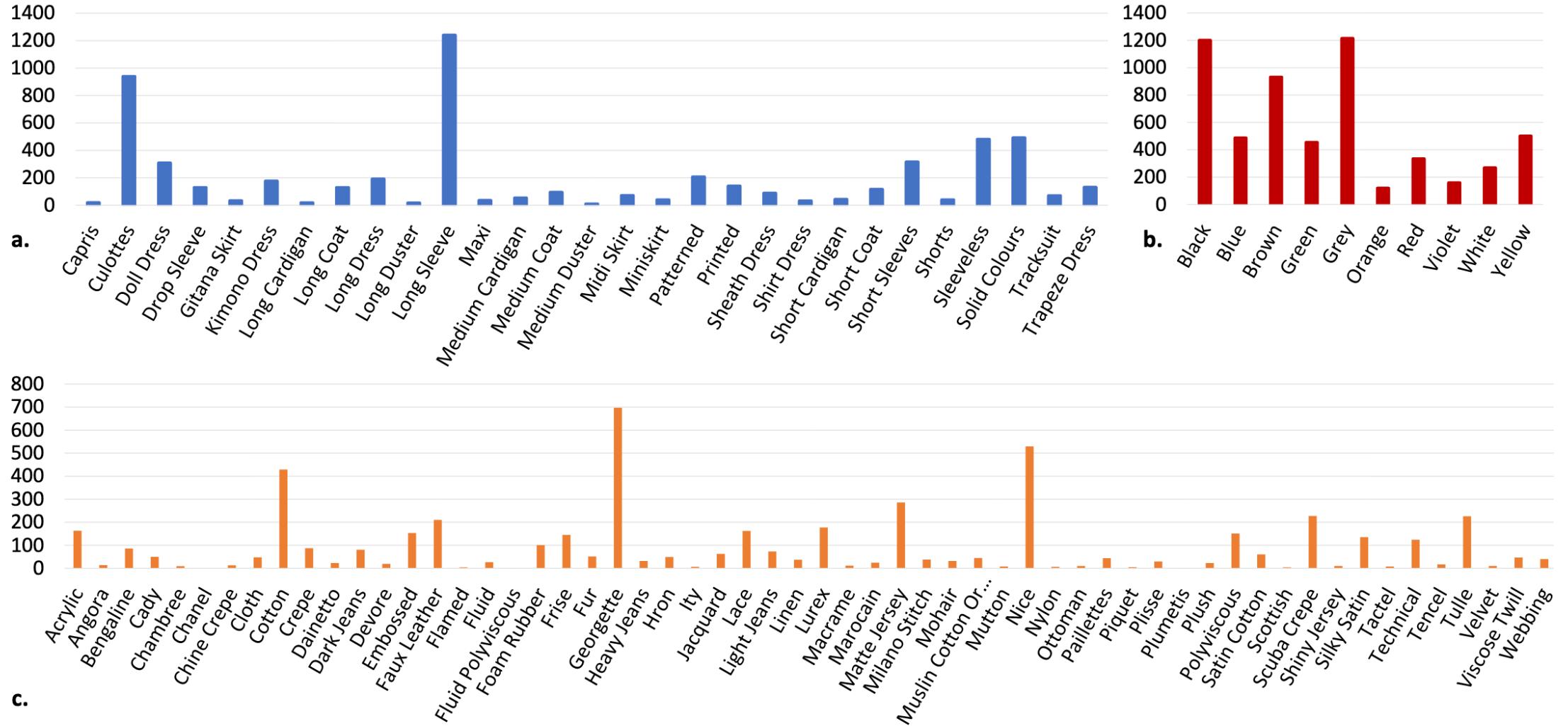


Figure 2: Cardinalities of the dataset for clothing categories (a), color (b) and fabric (c).

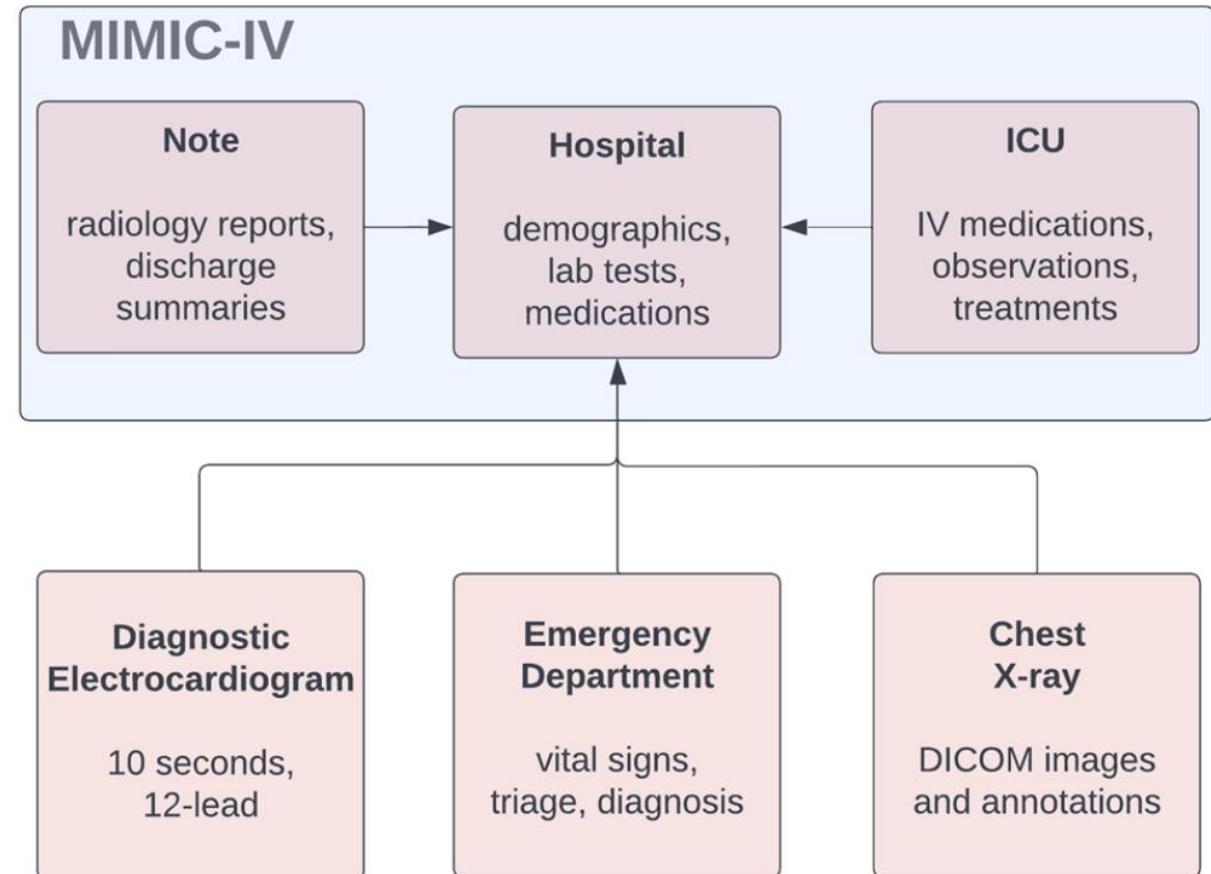
# Multi-modal Time Series Datasets – TS, Text, Table

MIMIC-IV: A freely accessible electronic health record dataset

**TS:** Dynamic, timestamped physiological or treatment data such as heart rate and blood pressure

**Text:** Unstructured free-text clinical narratives

**Table:** Static or low-frequency structured data such as Patient demographics and medication prescription



MIMIC-IV follows a modular structure. Modules can be linked by identifiers including subject\_id, hadm\_id, and deidentified date and time.

# Multi-modal Time Series Datasets – TS, Text, Table

	Hospital admissions	ICU admissions
Number of stays	431,231	73,181
Unique patients	180,733	50,920
Age, mean (SD)	58.8 (19.2)	64.7 (16.9)
Female Administrative Gender, n (%)	224,990 (52.2)	32,363 (44.2)
Insurance, n (%)		
Medicaid	41,330 (9.6)	5,528 (7.6)
Medicare	160,560 (37.2)	33,091 (45.2)
Other	229,341 (53.2)	34,562 (47.2)
Hospital length of stay, mean (SD)	4.5 (6.6)	11.0 (13.3)
In-hospital mortality, n (%)	8,974 (2.1)	8,519 (11.6)
One year mortality, n (%)	106,218 (24.6)	28,274 (38.6)

**Table 1.** Demographics for patients admitted to an intensive care unit (ICU) in MIMIC-IV v2.2.

# Multi-modal Time Series Datasets – TS, Image, Text, Table

## FinMultiTime: A Four-Modal Bilingual Dataset for Financial Time-Series Analysis

**TS:** Stock price time series

**Image:** K-line technical charts

**Text:** Financial news

**Table:** Structured financial tables

- Across both the S&P 500 and HS 300 universes
- Covering 5,105 stocks from 2009 to 2025 in the United States and China

Table 2: Overview of Bilingual Financial Dataset Specifications for the HS300 (Chinese) and S&P 500 (English) Indices

Bilingual Dataset	Type	Size	Format	Stocks	Records	Frequency
HS300 (Chinese)	Image	2.43 GB	PNG	810	52,914	Semi-Annual
	Table	568 MB	JSON/JSONL	810	2,430	Quarterly/Annual
	Time series	345 MB	CSV	810	810	Daily
	Text	652.53 MB	JSONL	892	1,420,362	Minute-Level
	All	3.96 GB	–	–	1,476,516	–
SP500 (English)	Image	8.67 GB	PNG	4,213	195,347	Semi-Annual
	Table	84.04 GB	JSON/JSONL	2,676	8,028	Quarterly/Annual
	Time series	1.83 GB	CSV	4,213	4,213	Daily
	Text	14.1 GB	JSONL	4,694	3,351,852	Minute-Level
	All	108.64 GB	–	–	3,559,440	–

# Multi-modal Time Series Datasets – TS, Image, Text, Table

Table 6: HS300 vs. S&P 500 — Multimodal Record Counts (35 stocks each)

	Semi-annual trend images	Quarterly / annual tables	Daily time-series points	News-sentiment scores
HS300	299,923	1,749	299,923	26,467
S&P 500	299,923	2,104	299,923	51,235
<b>Total</b>	<b>599,846</b>	<b>3,853</b>	<b>599,846</b>	<b>77,702</b>

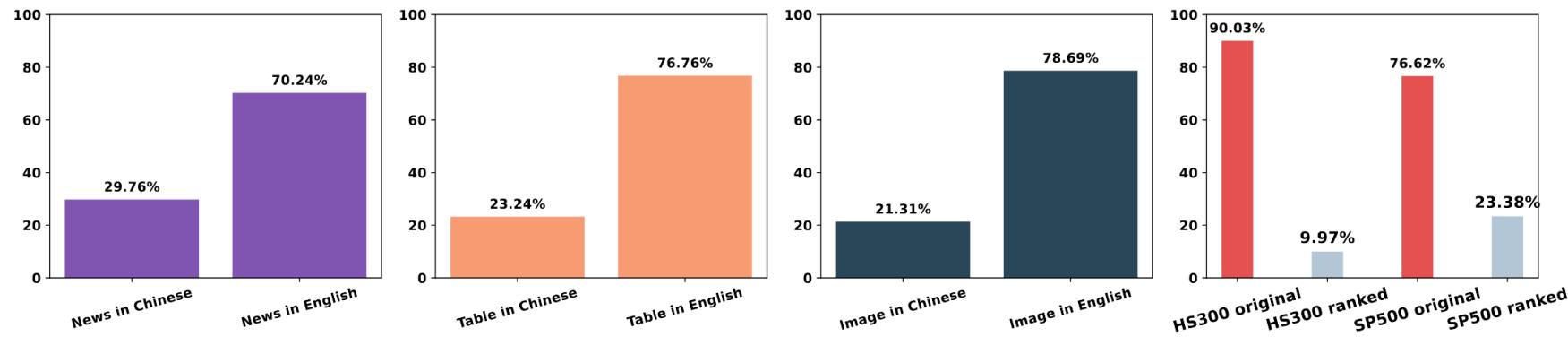
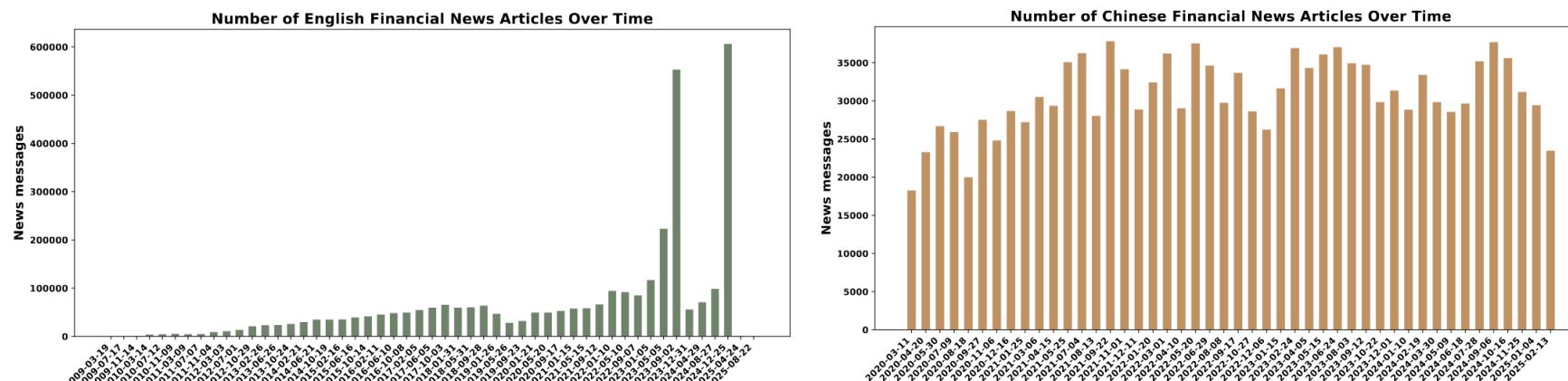


Figure 6: Proportions of Chinese vs. English Modalities (News, Tables, Images) and Coverage Ratios of Ranked vs. Original Daily News for HS300 and S&P 500.

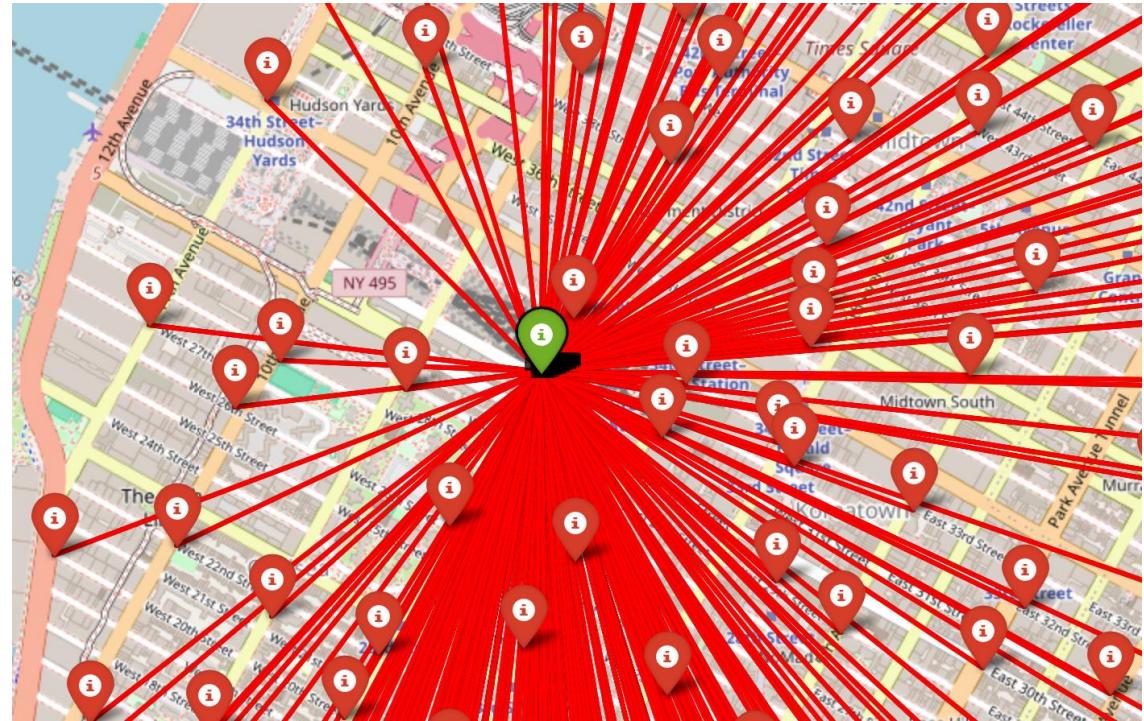


# Multi-modal Time Series Datasets – ST, Text

NYC Bike Sharing Network: Time-Series Enhanced Nodes and Edges Dataset

**ST:** Time-series data of bike availability and trip flows across spatially distributed stations

**Text:** Station-level static information such as ID, name, and capacity



An illustration of start-end trip flows originating from 8th Ave & W 31st St

# Multi-modal Time Series Datasets – ST, Text, Image

## Terra: A Multimodal Spatio-Temporal Dataset Spanning the Earth

**ST:** Multi-variable spatio-temporal data

**Text:** LLM-Derived text description

**Image:** Geo-Image and satellite image

Encompasses hourly time series data from 6,480,000 grid areas worldwide over the past 45 years

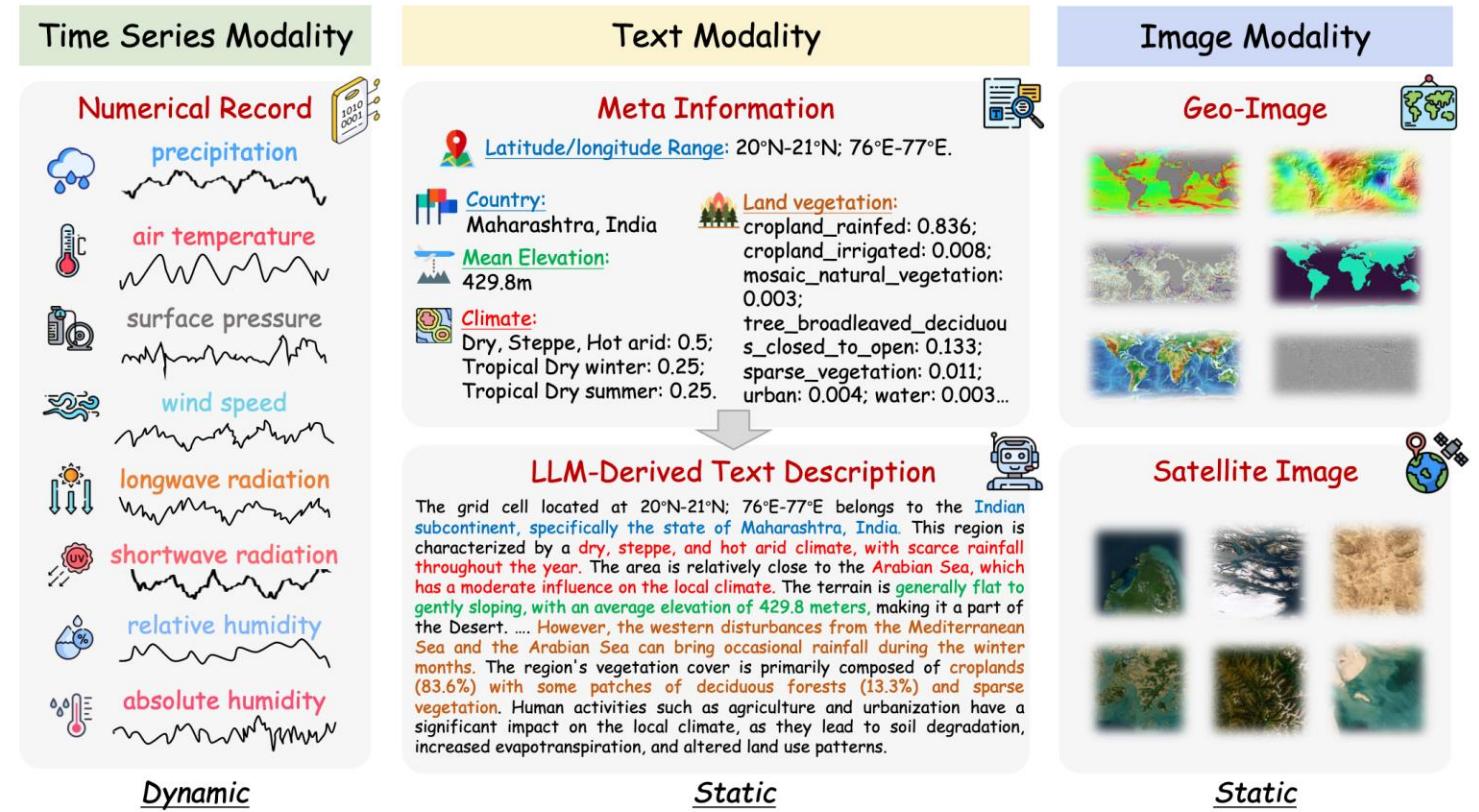


Figure 2: Different modality components of Terra. We provide the data with three temporal scales (3 hourly / daily / monthly), and three spatial scale ( $0.1^\circ$  /  $0.5^\circ$  /  $1^\circ$ ).

# Multi-modal Time Series Datasets - ST, Text, Image

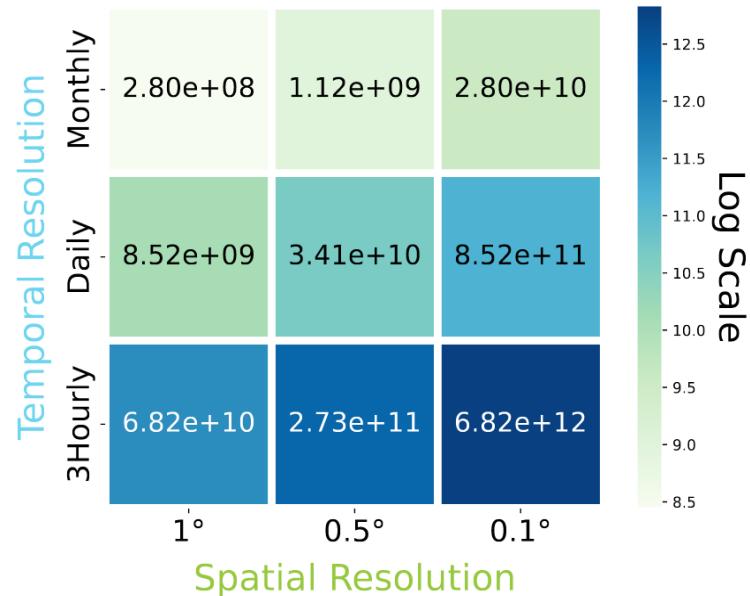


Figure 3: Dataset volume comparison.

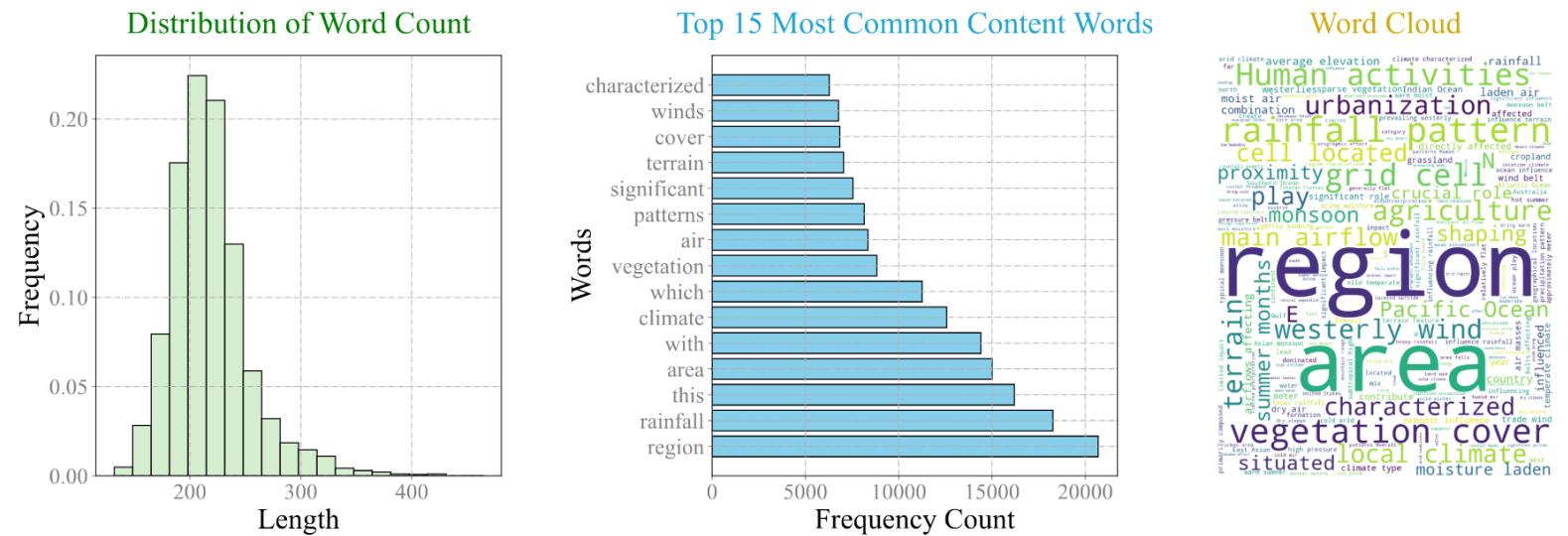


Figure 5: Statistical and visual insights of text modality data.

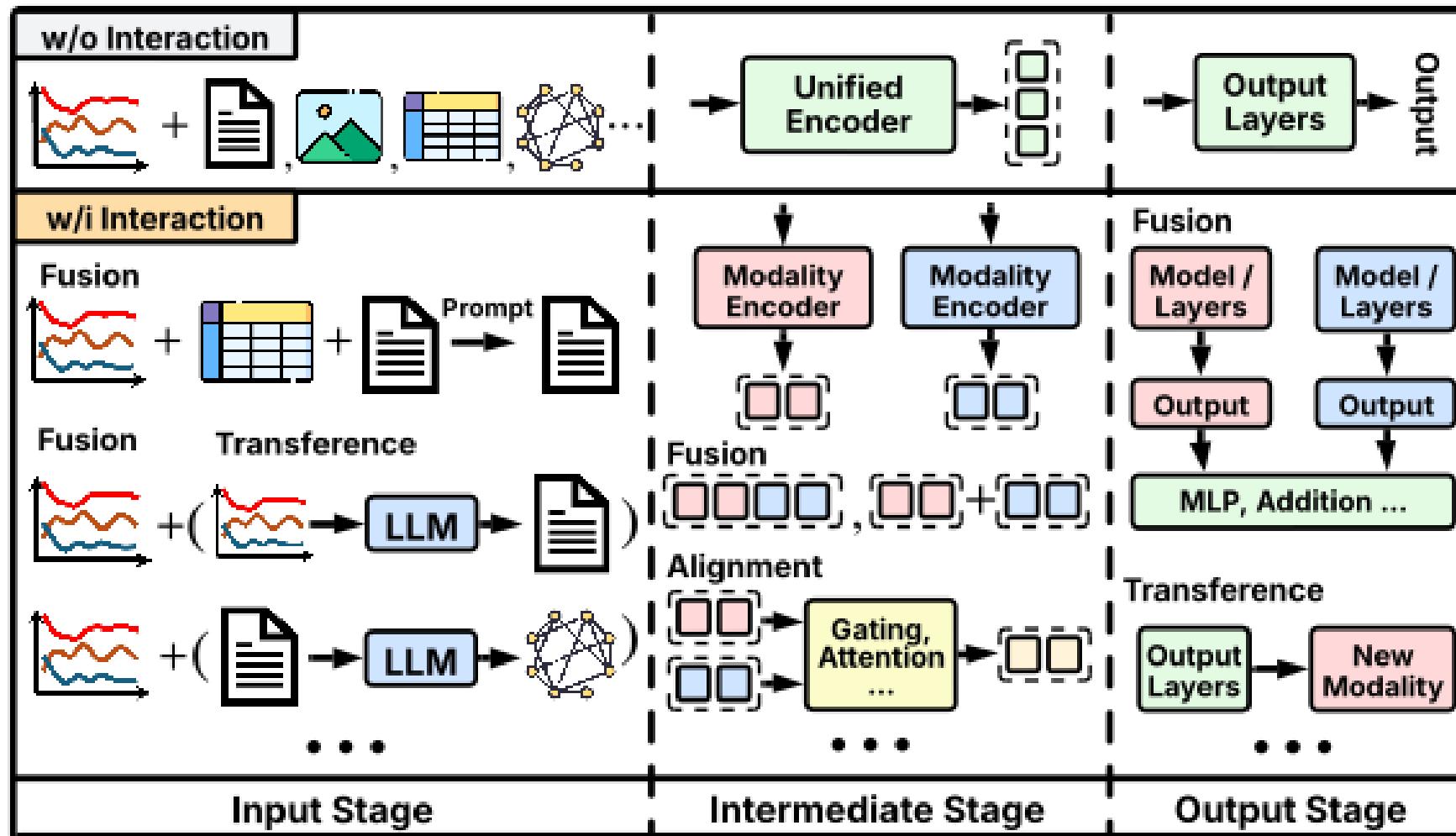
# *Multi-modal Time Series Methods*

# Taxonomy of Multi-modal Time Series Methods

We categorize over 40 multi-modal time series methods and define:

- 1) Three fundamental types of cross-modal interactions
  - **Fusion, Alignment, Transference**
- 2) Occurring at three levels within a framework
  - **Input, Intermediate, Output**
  - **Intermediate: representation or midpoint output** (not end-to-end)
- 3) An interaction can occur at one or more levels
- 4) Multiple interactions can co-occur at the same level

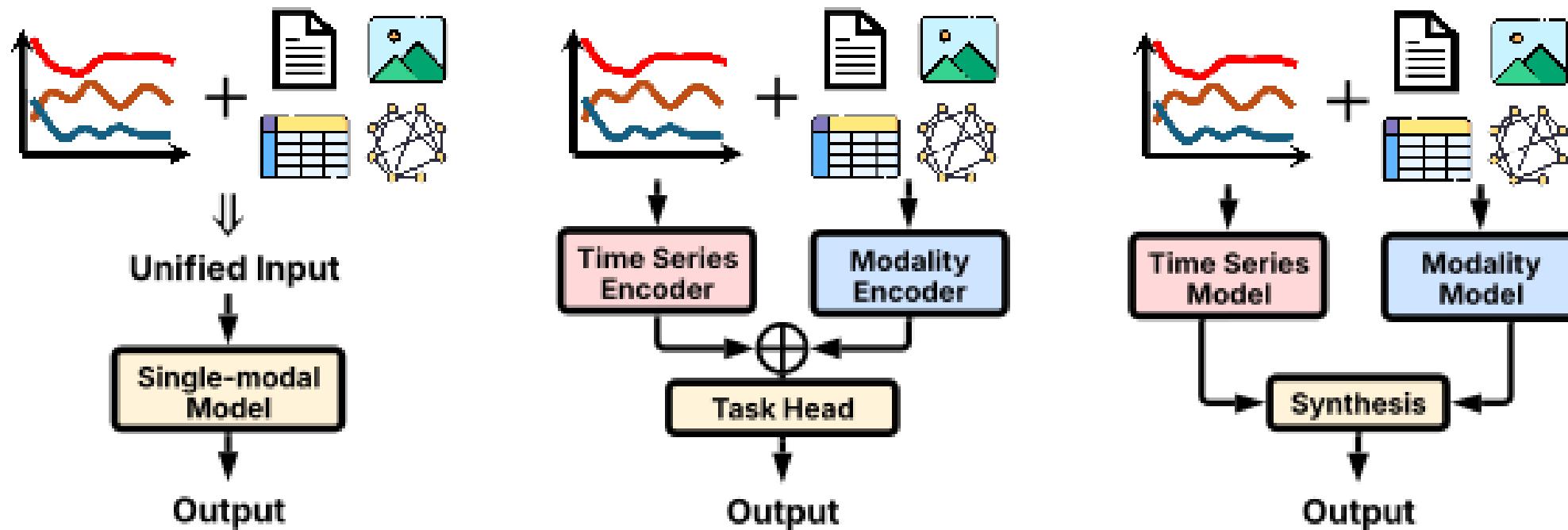
# Taxonomy of Multi-modal Time Series Methods



Overview and representative examples of cross-modal interactions

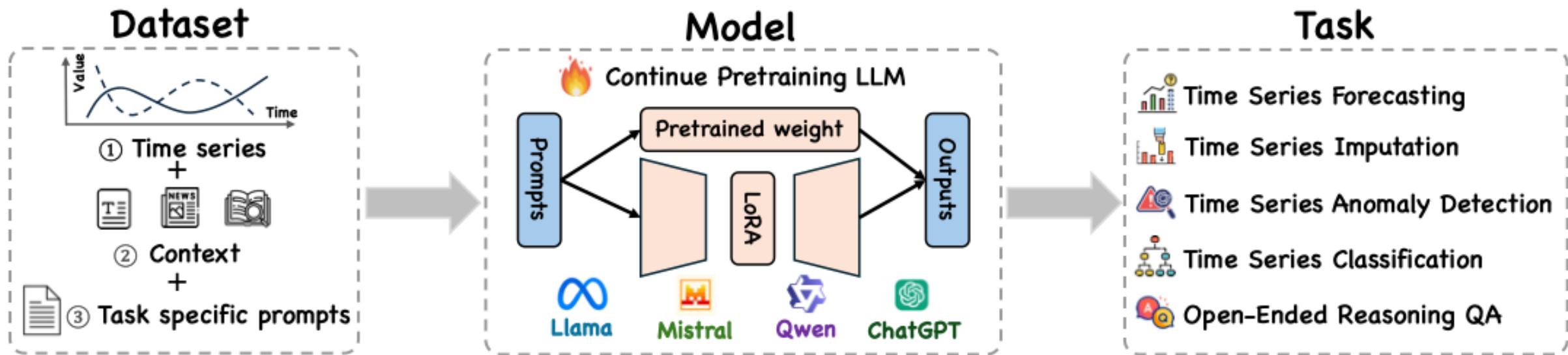
# Cross-modal Interaction with Time Series: Fusion

Definition: the process of Integrating heterogeneous modalities in a way that captures **complementary information** across diverse sources



# Multi-modal Fusion with Time Series – Input level

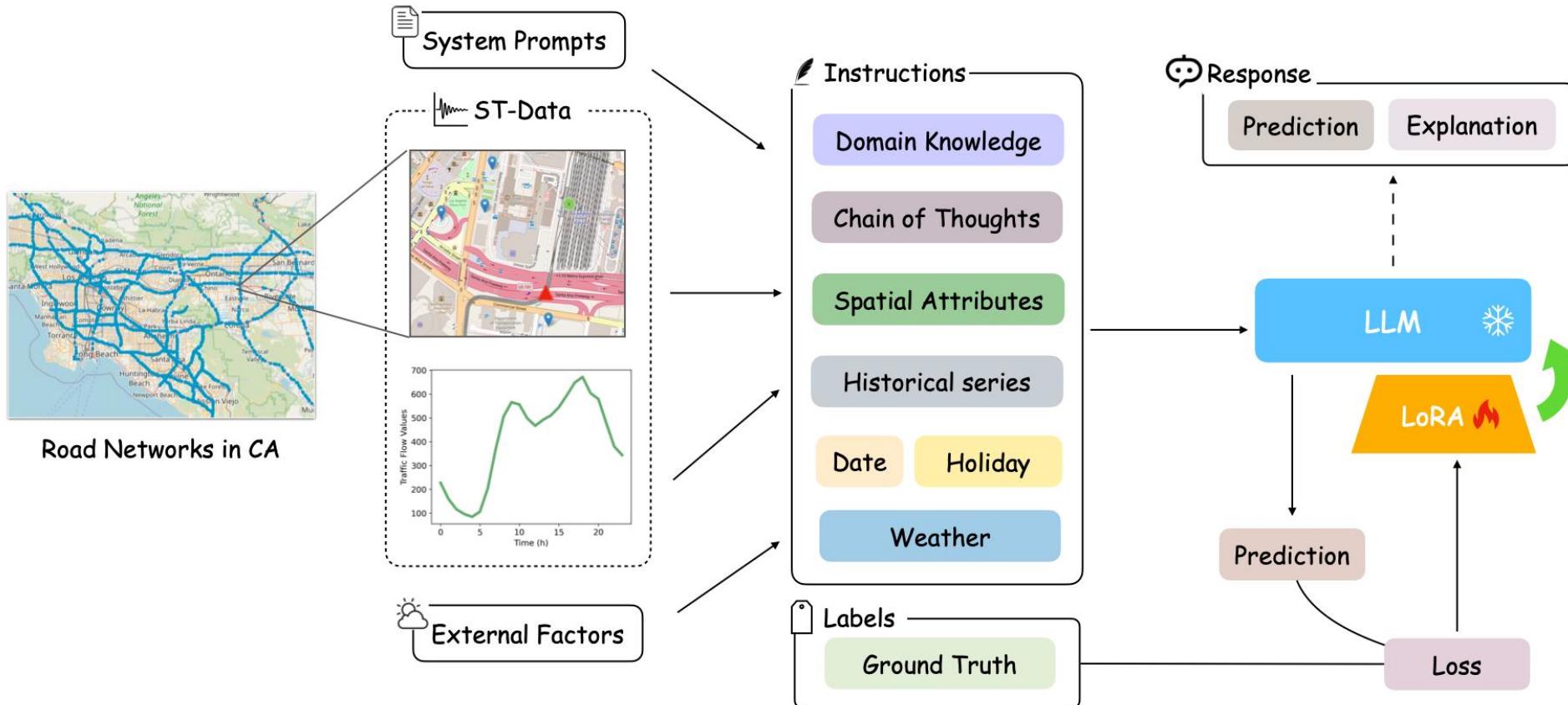
Integrate time series, tabular data and texts into a unified textual prompt



**Time-MQA: Time Series Multi-Task Question Answering with Context Enhancement**

# Multi-modal Fusion with Time Series – Input level

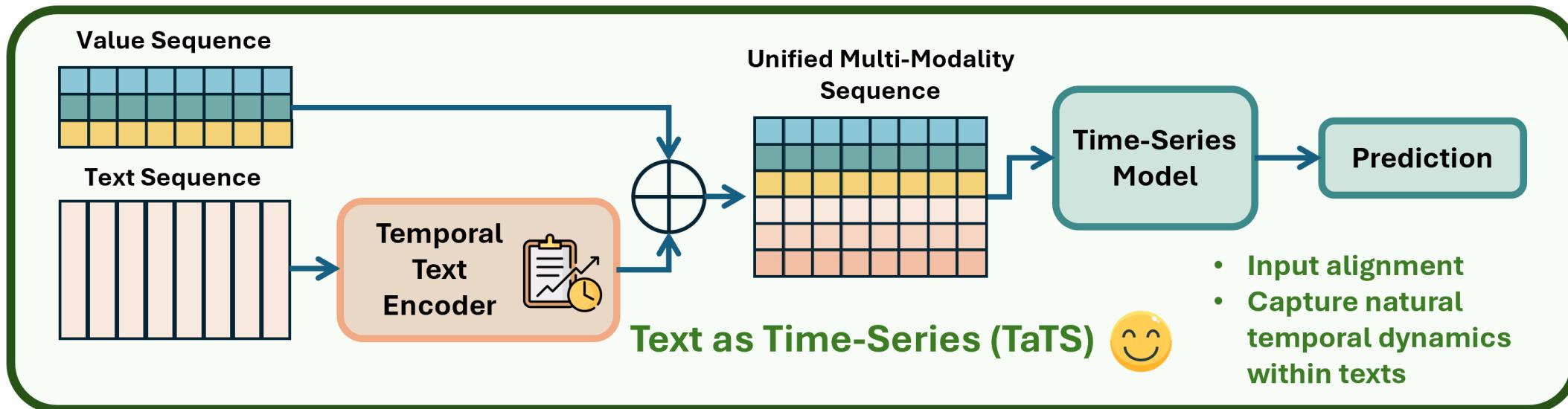
Integrate time series, tabular data and texts into a unified textual prompt



Towards explainable traffic flow prediction with large language models

# Multi-modal Fusion with Time Series – Input level

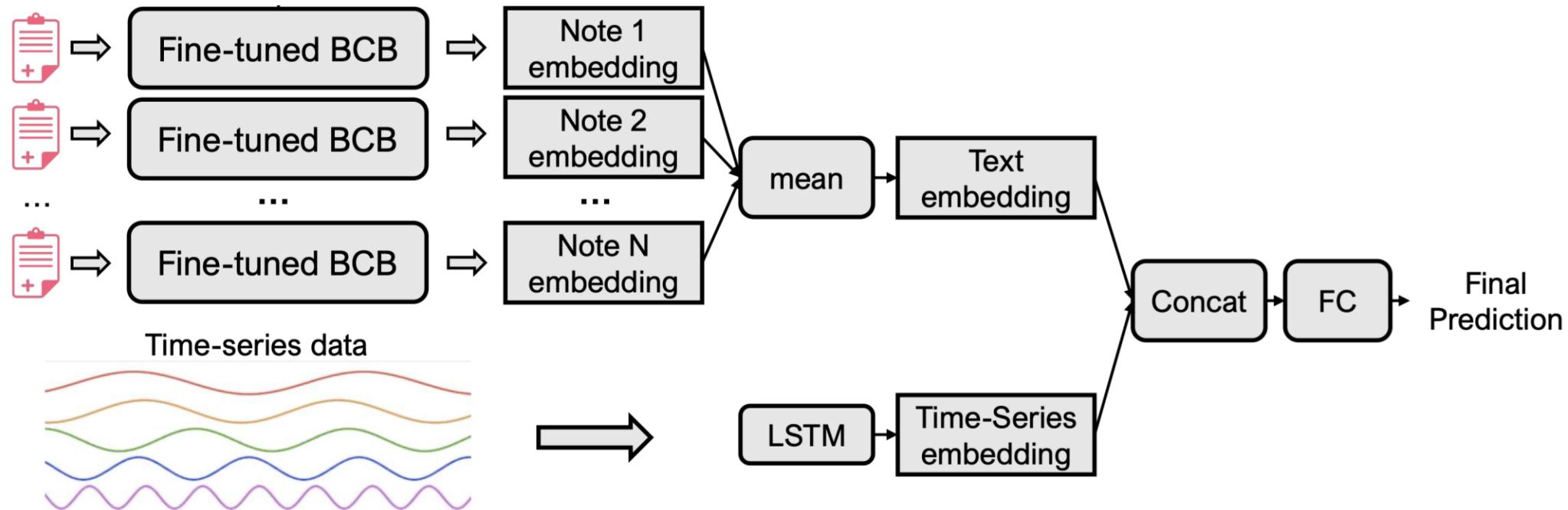
Integrate paired text embedding as an additional variable of time series



**Language in the Flow of Time: Time-Series-Paired Texts Weaved into a Unified Temporal Narrative**

# Multi-modal Fusion with Time Series – Intermediate level

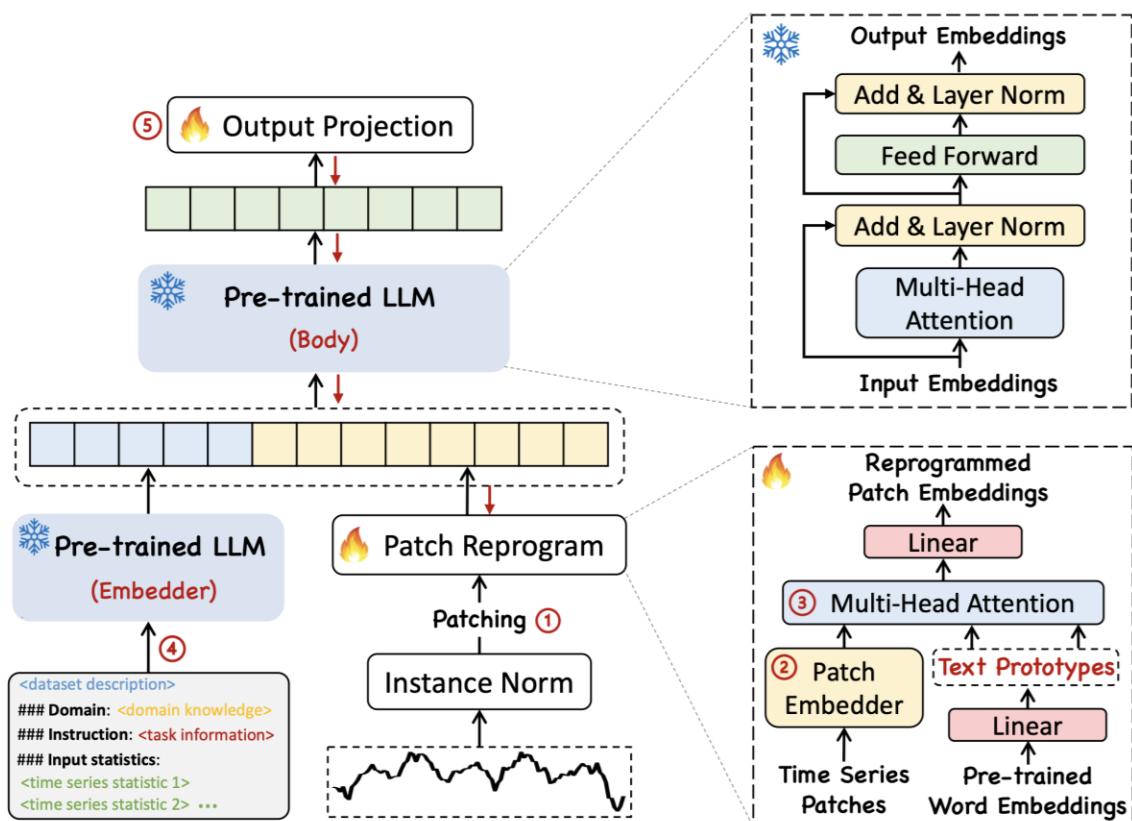
Simple aggregations (e.g., mean, addition, concatenation, etc.) of time series embedding and other modality embeddings



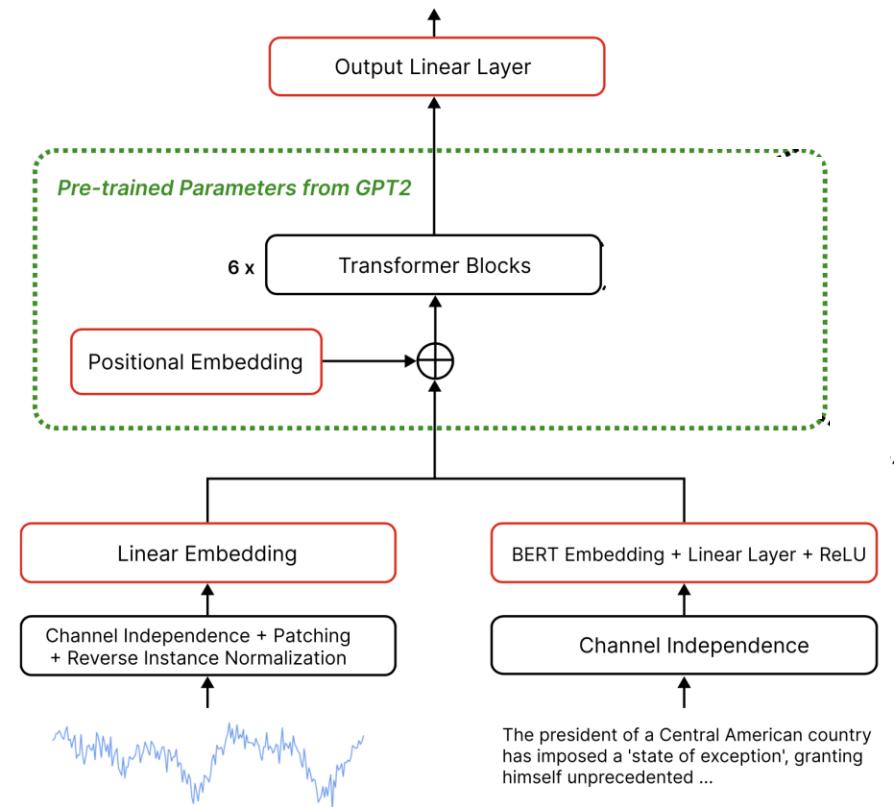
Predicting In-hospital Mortality by Combining Clinical Notes with Time-series Data (ACL 2021)

# Multi-modal Fusion with Time Series – Intermediate level

The fusion of modality embeddings is usually followed by alignments



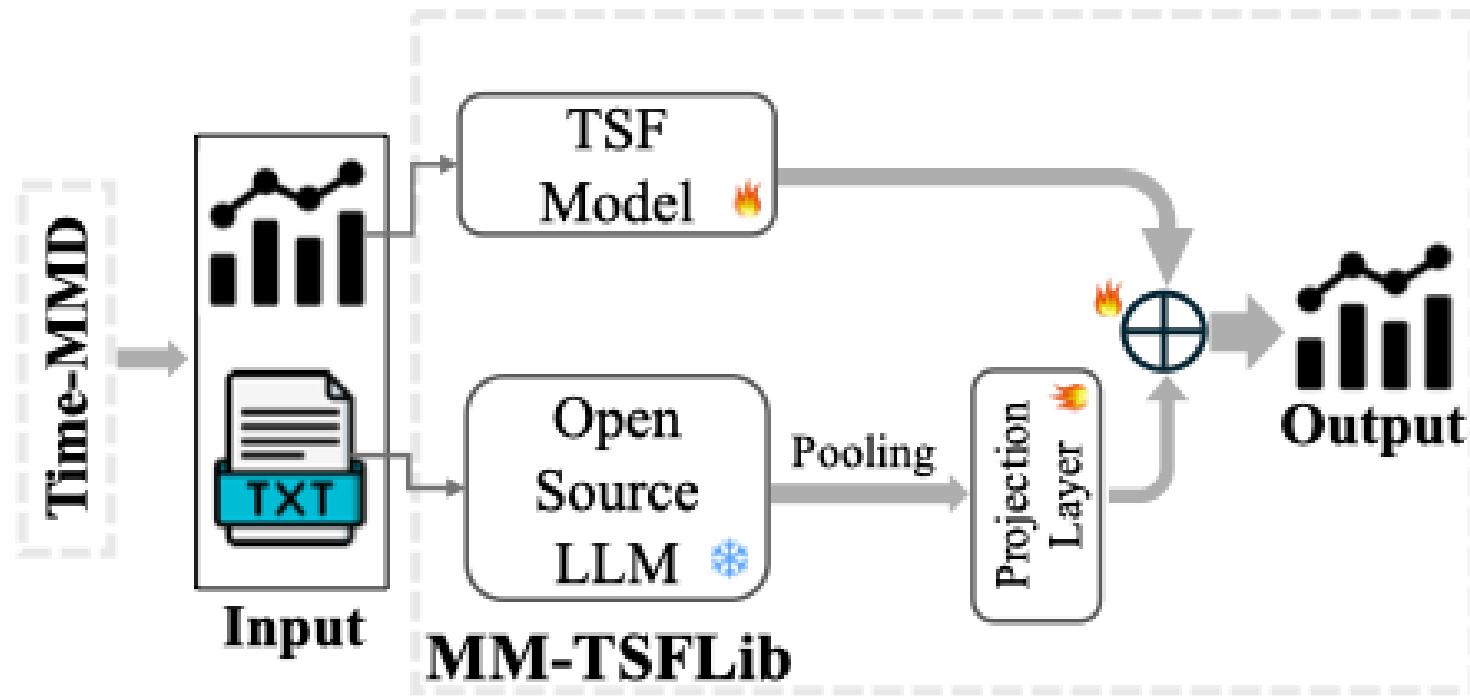
**Time-LLM: Time Series Forecasting by Reprogramming Large Language Models (ICLR 2024)**



**GPT4MTS: Prompt-Based Large Language Model for Multimodal Time Series Forecasting (AAAI 2024)**

# Multi-modal Fusion with Time Series – Output level

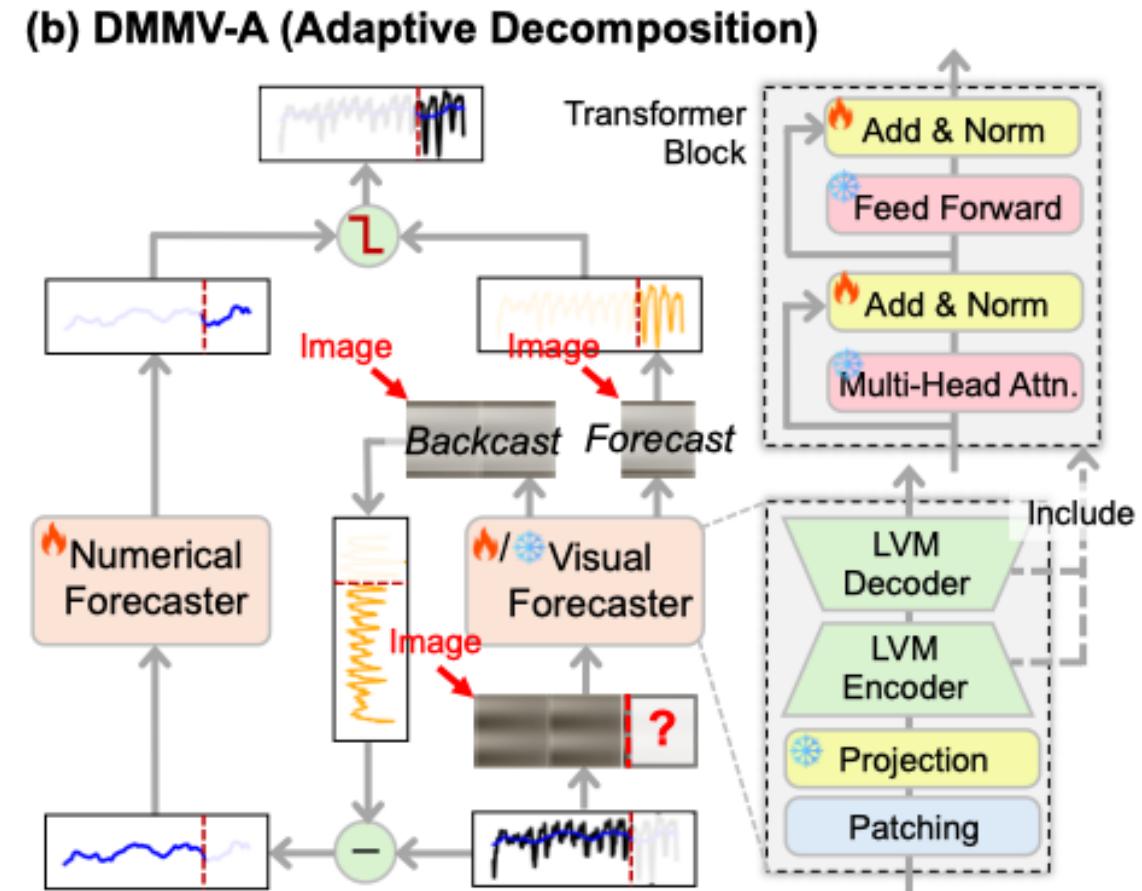
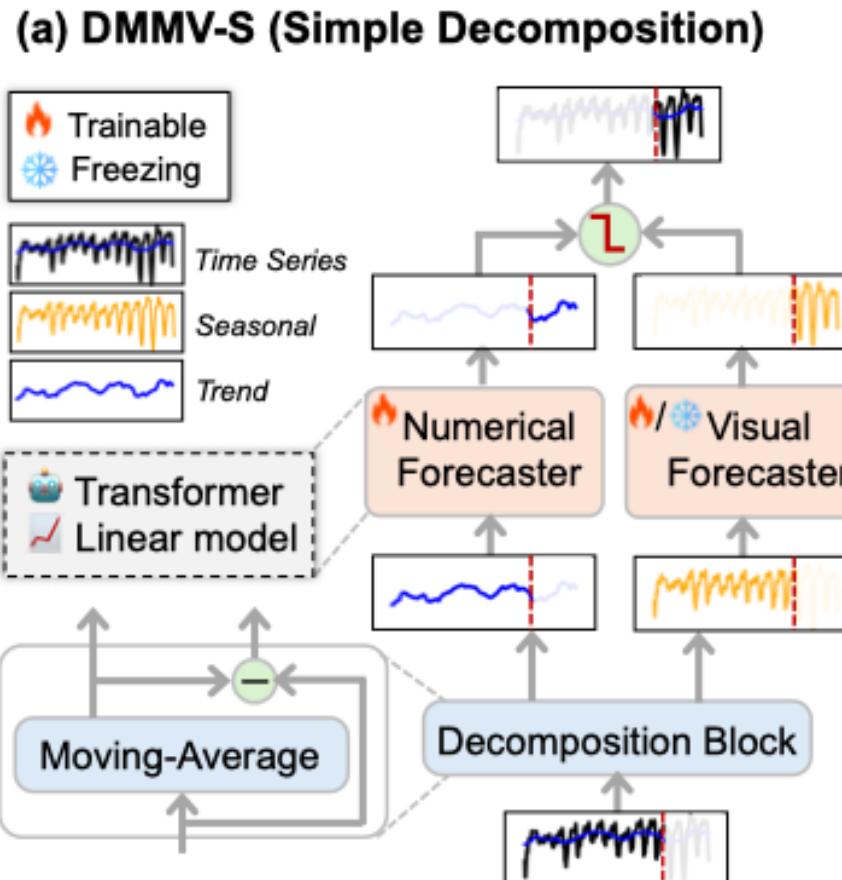
Project multiple modality outputs onto a unified space



**Time-MMD: Multi-Domain Multimodal Dataset for Time Series Analysis (NeurIPS 2024)**

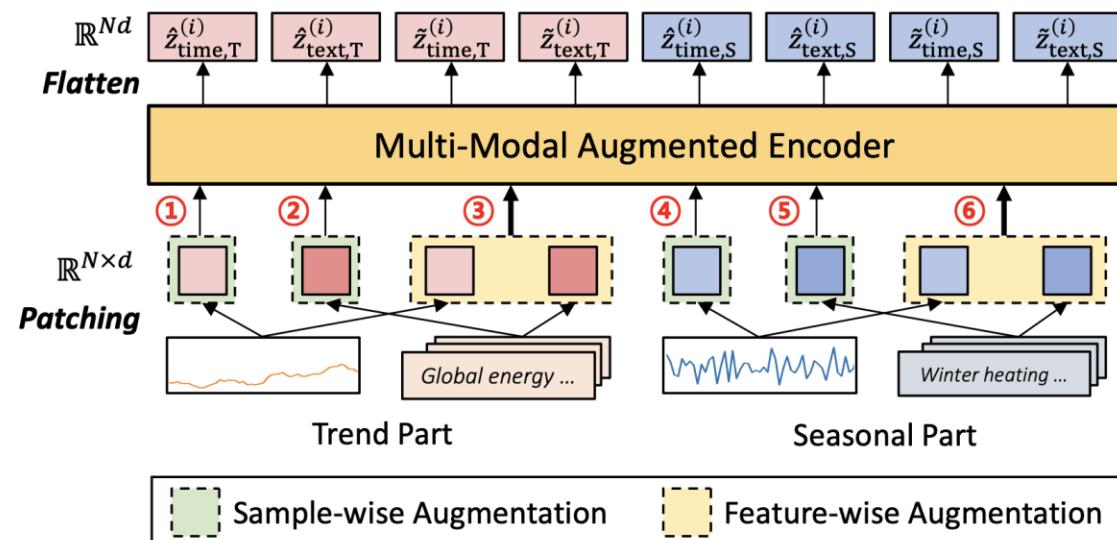
# Multi-modal Fusion with Time Series – Output level

Assemble modality outputs as decomposed components of the final output

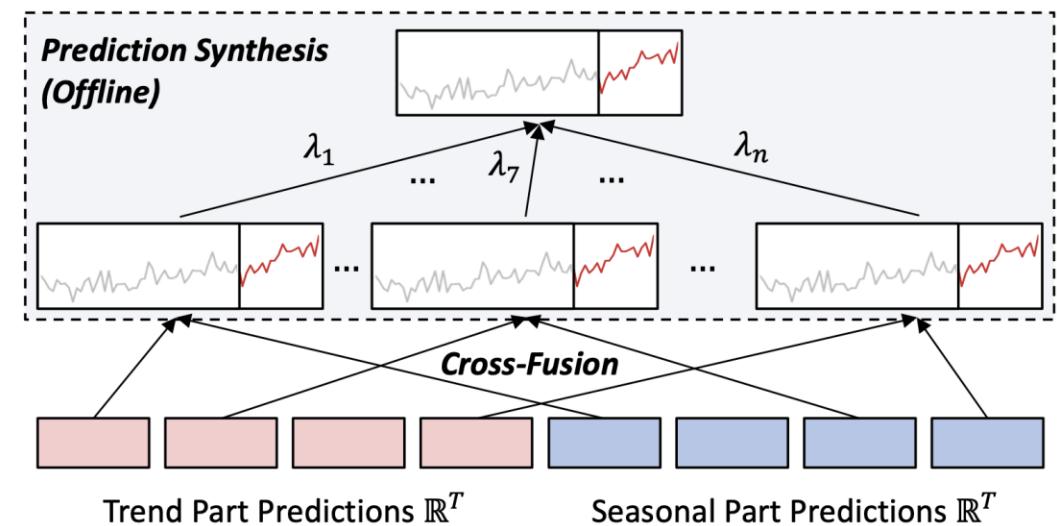


# Multi-modal Fusion with Time Series – Output level

Offline synthesis of modality outputs to assemble the final output



(a) Multi-modal augmented encoder



(b) Cross-modal fusion and offline synthesis

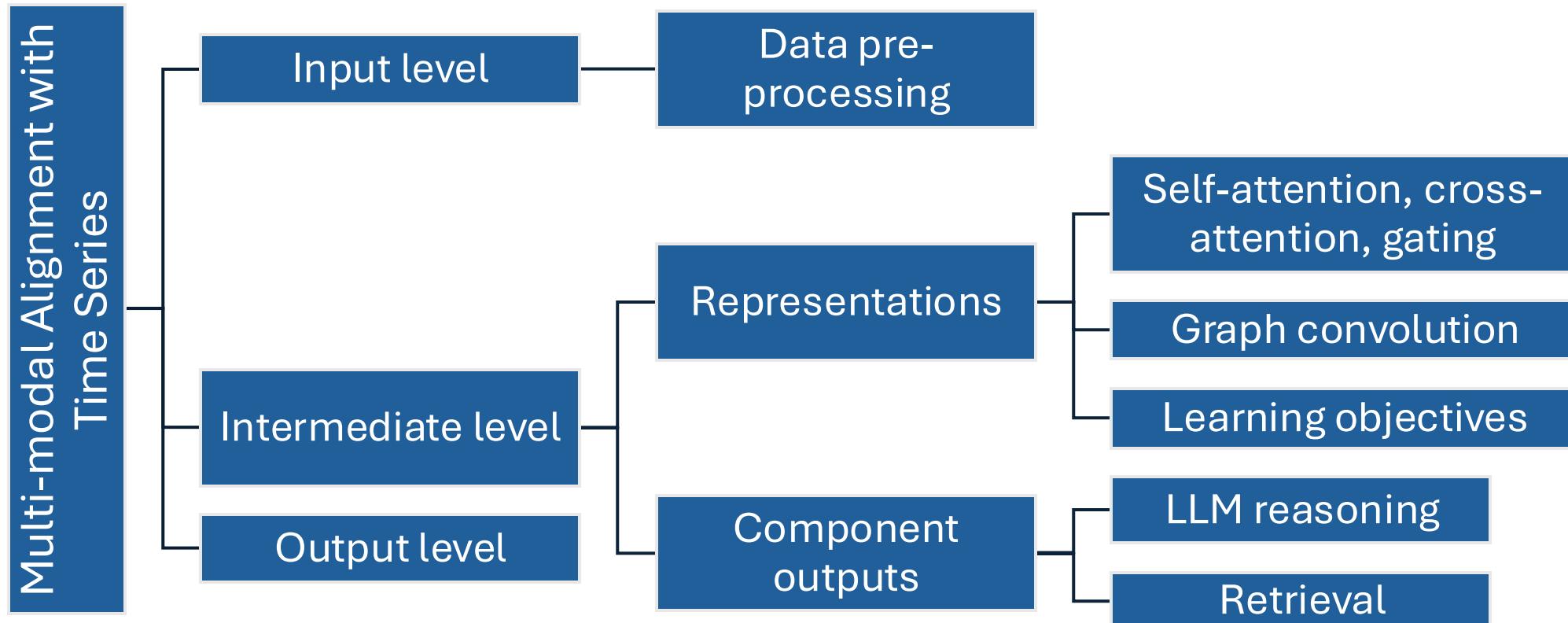
**MoAT: Multi-Modal Augmented Time Series Forecasting**

# Multi-modal Fusion with Time Series

- Fusion relies on well-aligned multi-modal data for effective exploitation of the contextual information.
- However, ideally-aligned data may not be given in real-world scenarios.
- As such, existing methods also leverage alignment mechanisms to mitigate the challenge

# Cross-modal Interaction with Time Series: Alignment

Definition: the process of preserving inter-modal relationships and ensuring semantic coherence when integrating different modalities into a unified framework



# Multi-modal Alignment with Time Series - Representations

**Self-attention:** a joint and undirected alignment across all modalities by dynamically attending to important features.

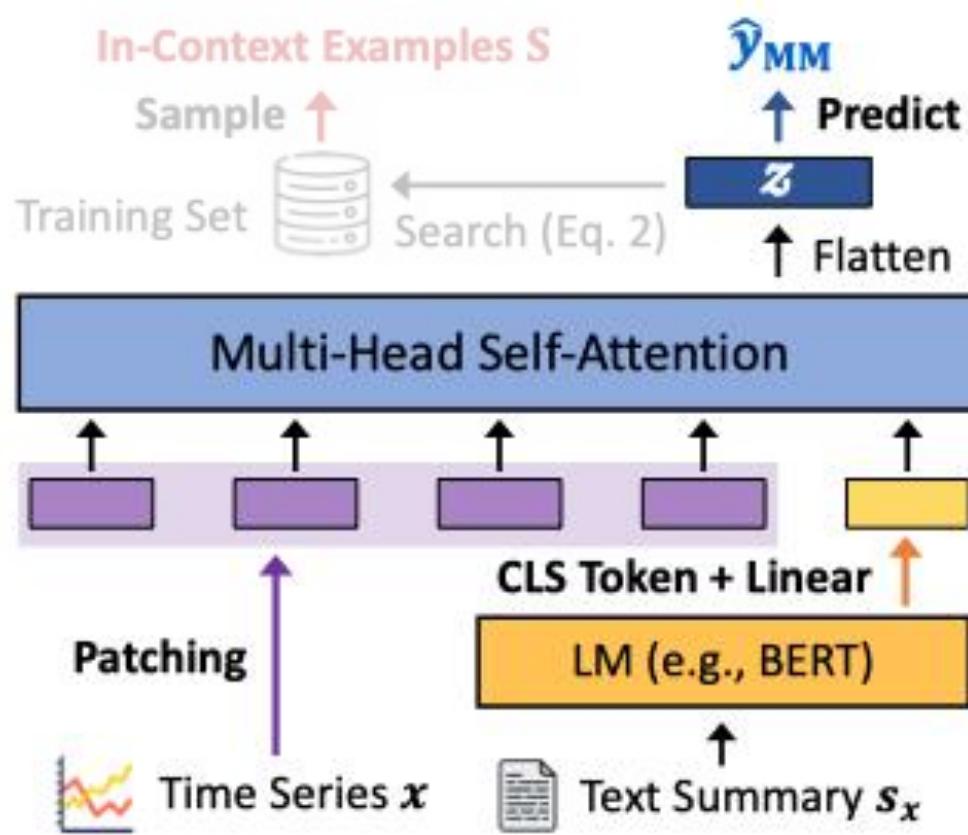
Given multi-modal embeddings  $E_{\text{mm}} \in \mathbb{R}^{n \times d}$ , where  $n$  is the number of modality tokens and  $d$  is the embedding dimension:

$$\text{Attention}(E_{\text{mm}}) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

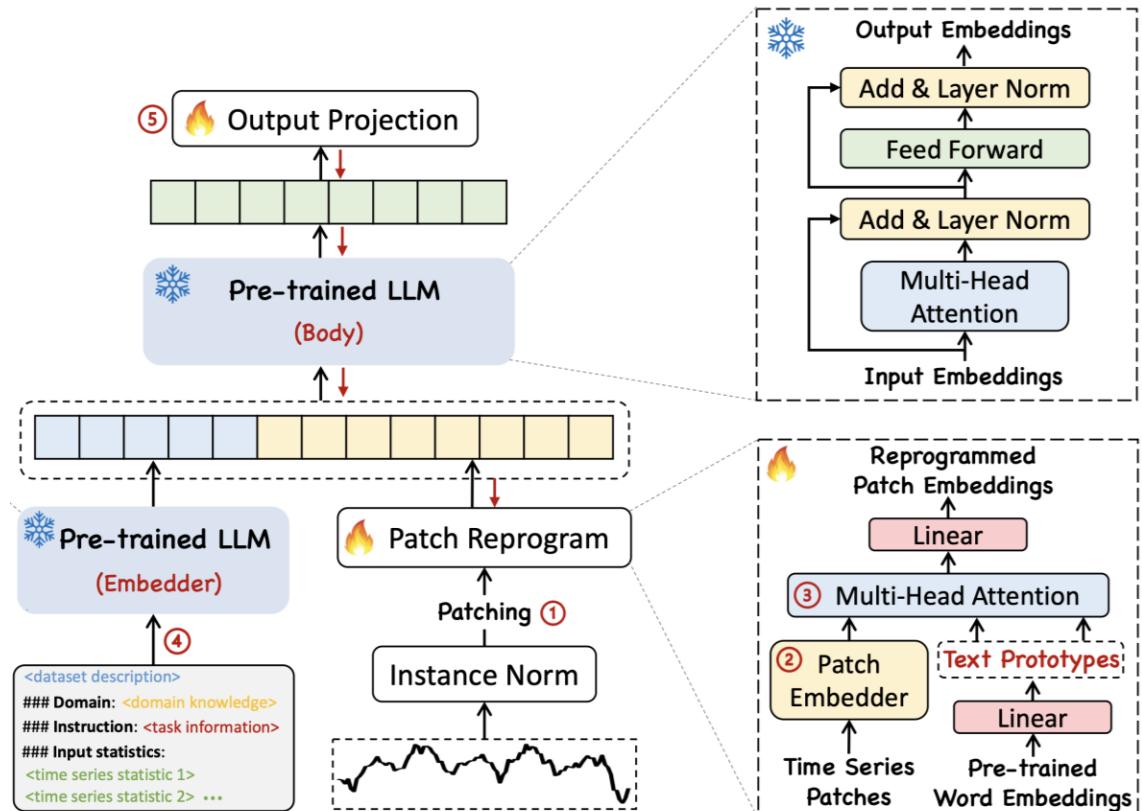
where the queries  $Q$ , keys  $K$ , and values  $V$  are linear projections of  $E_{\text{mm}}$ :

$$Q = E_{\text{mm}}W_Q, K = E_{\text{mm}}W_K, V = E_{\text{mm}}W_V \text{ with learnable weights } W_{Q,K,V} \in \mathbb{R}^{d \times d_k}$$

# Multi-modal Alignment with Time Series - Representations

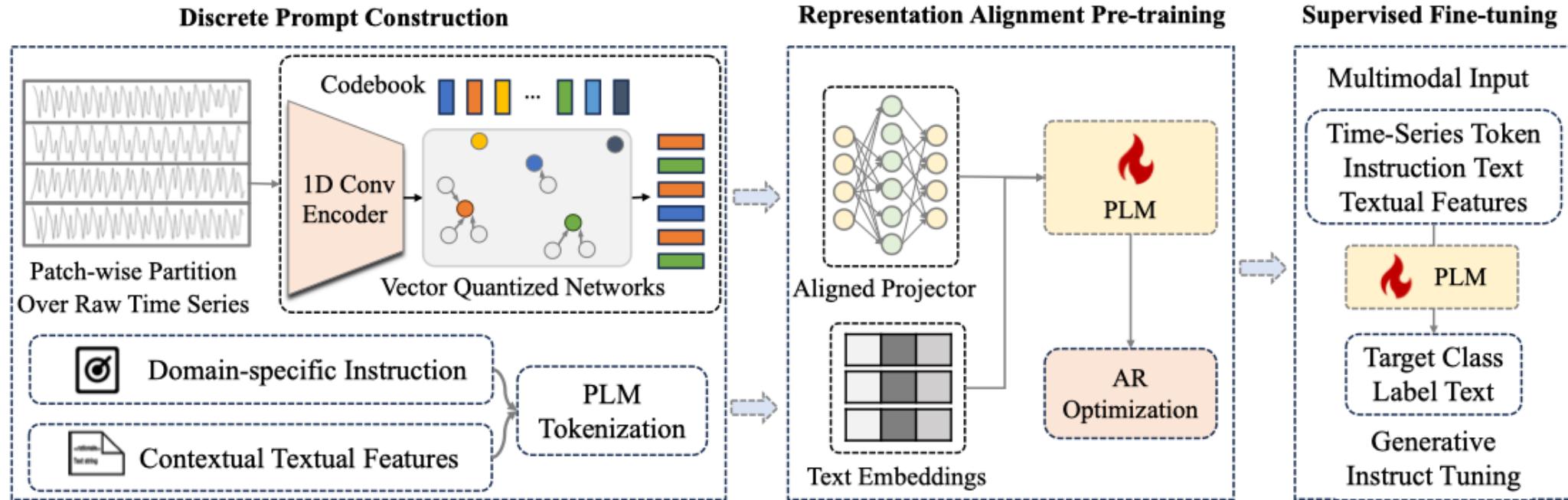


TimeCAP: Learning to Contextualize, Augment, and Predict Time Series Events with Large Language Model Agents (AAAI 2025)



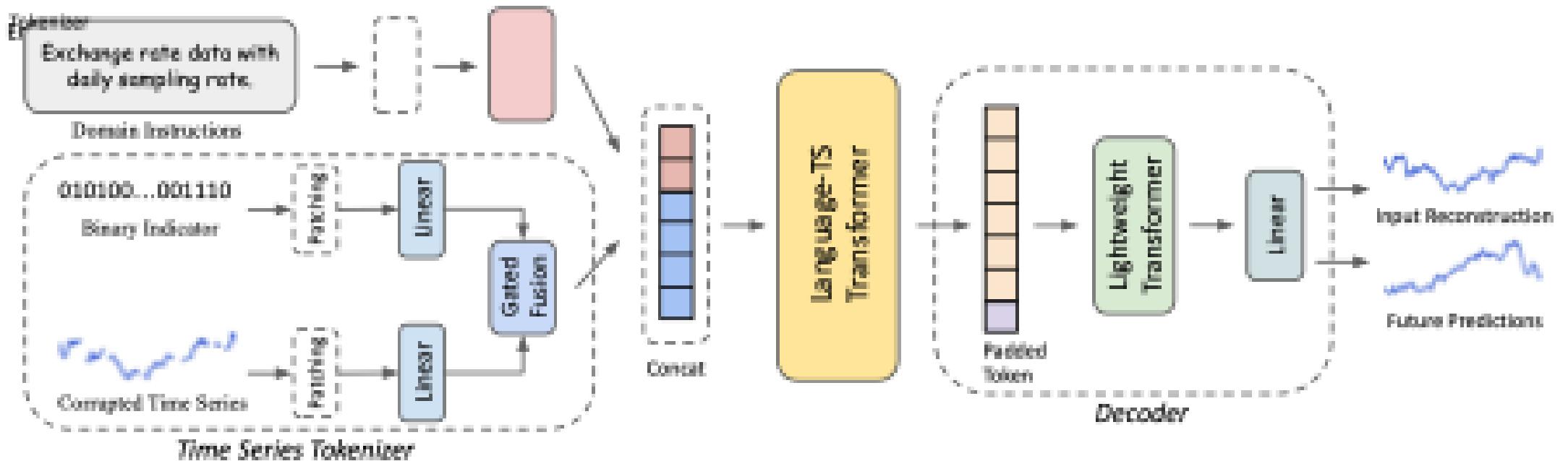
Time-LLM: Time Series Forecasting by Reprogramming Large Language Models (ICLR 2024)

# Multi-modal Alignment with Time Series - Representations



**Advancing Time Series Classification with Multimodal Language Modeling (WSDM 2025)**

# Multi-modal Alignment with Time Series - Representations



UniTime: A Language-Empowered Unified Model for Cross-Domain Time Series Forecasting (WWW'24)

# Multi-modal Alignment with Time Series - Representations

**Cross-attention:** time series serves as the query modality to get contextualized by other modalities, providing a directed alignment that ensure auxiliary modalities contribute relevant contexts while preserving the temporal structure of time series.

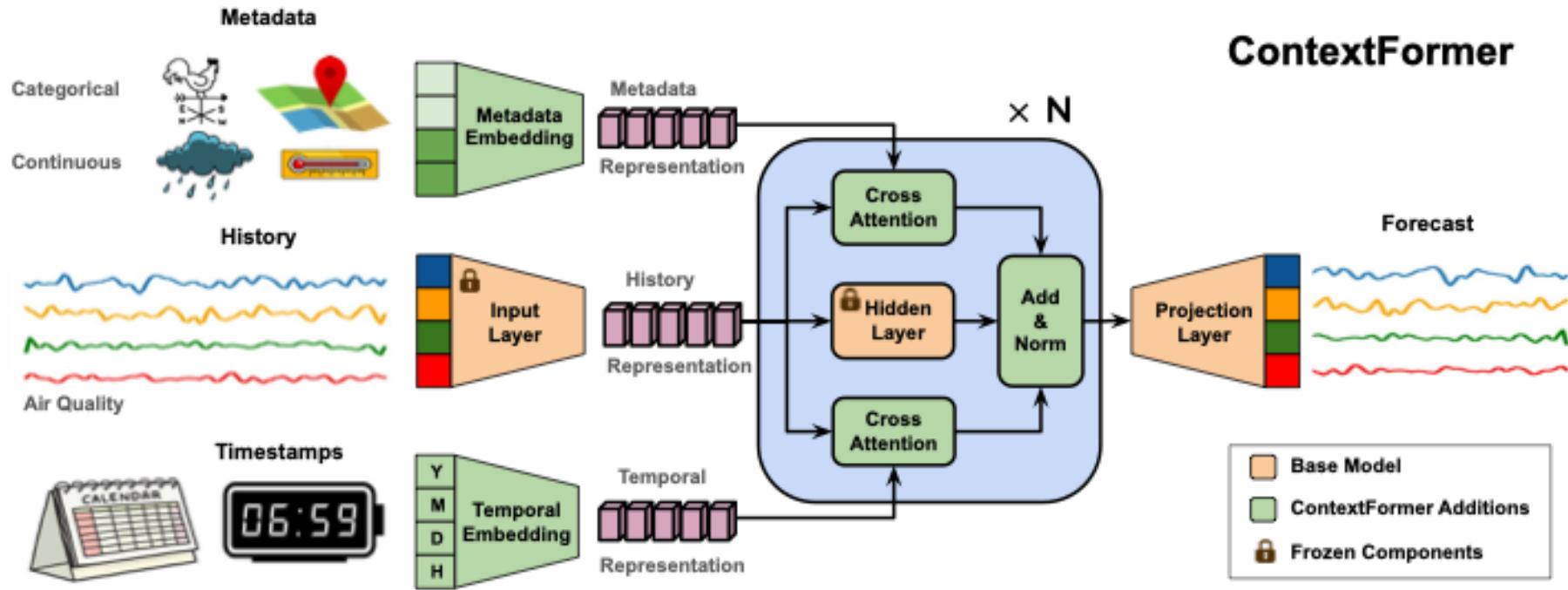
Given multi-modal embeddings  $E_{\text{ts}} \in \mathbb{R}^{n \times d}$ , where  $n$  is the number of modality tokens and  $d$  is the embedding dimension:

$$\text{CrossAttention}(E_{\text{ts}}, E_{\text{c}}) = \text{softmax}\left(\frac{Q_{\text{ts}} K_{\text{c}}^T}{\sqrt{d_k}}\right) V_{\text{c}}$$

where the queries  $Q_{\text{ts}}$ , keys  $K_{\text{c}}$ , and values  $V_{\text{c}}$  are linear projections of  $E_{\text{ts}}$ :

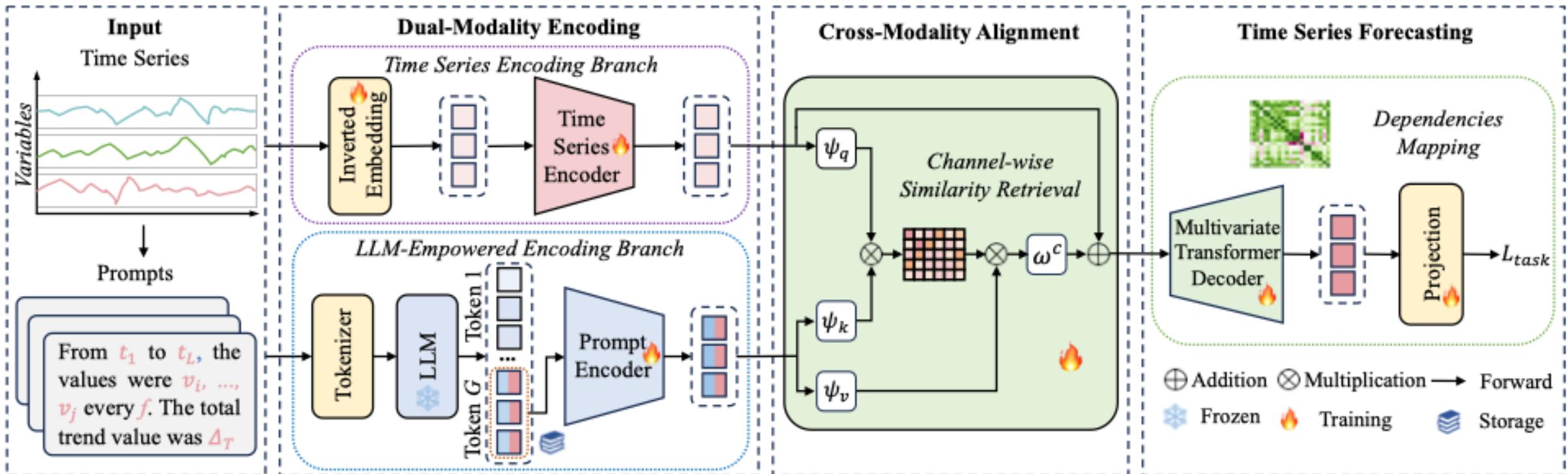
$Q_{\text{ts}} = E_{\text{ts}} W_Q$ ,  $K_{\text{c}} = E_{\text{ts}} W_K$ ,  $V_{\text{c}} = E_{\text{c}} W_V$  with learnable weights  $W_{Q,K,V} \in \mathbb{R}^{d \times d_k}$

# Multi-modal Alignment with Time Series - Representations



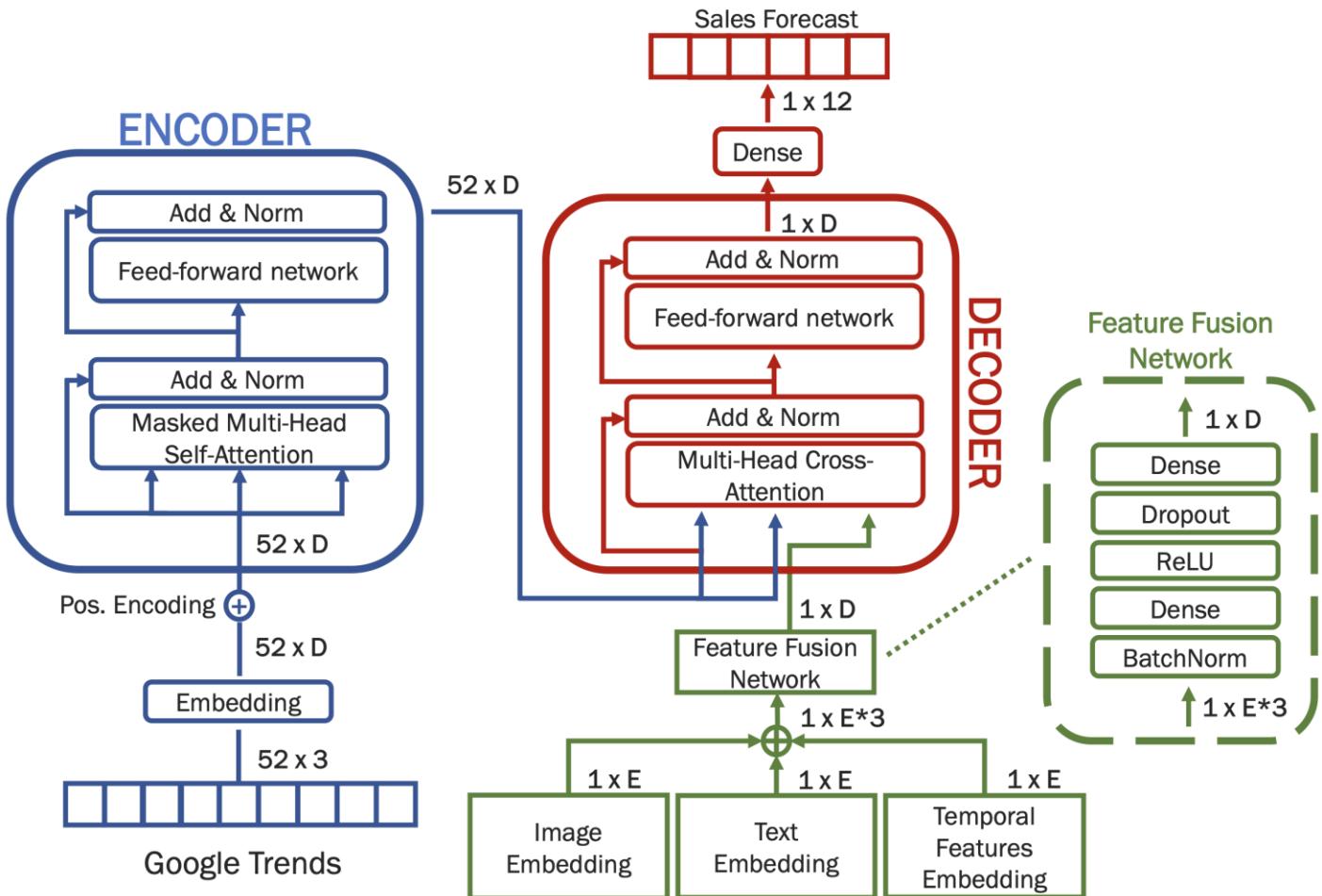
**Context Matters: Leveraging Contextual Features for Time Series Forecasting**

# Multi-modal Alignment with Time Series - Representations



TimeCMA: Towards LLM-Empowered Multivariate Time Series Forecasting via Cross-Modality Alignment (AAAI 2025)

# Multi-modal Alignment with Time Series - Representations



**Multimodal Forecasting of New Fashion Product Sales with Image-based Google Trends (Journal of Forecasting)**

# Multi-modal Alignment with Time Series - Representations

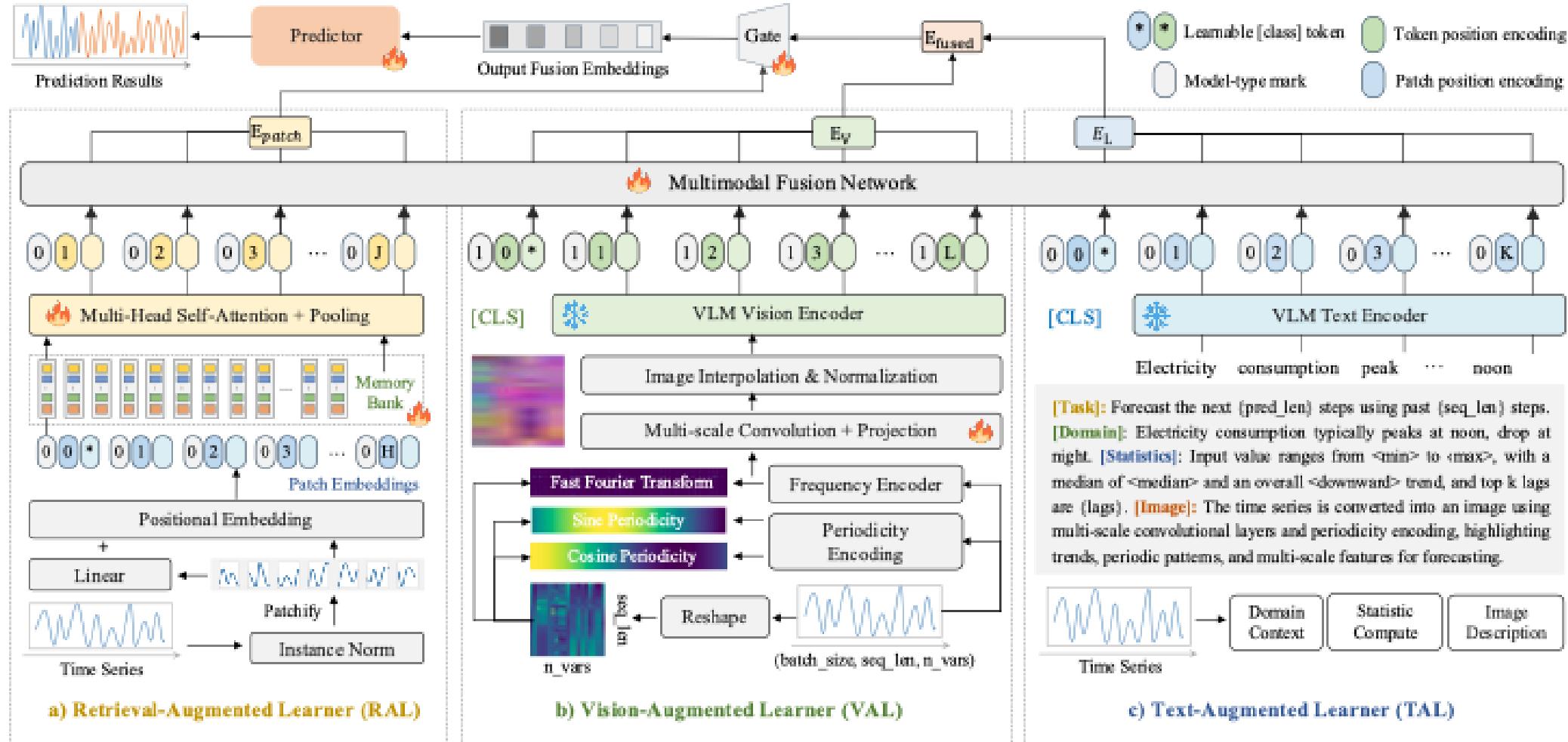
**Gating:** a parametric filtering operation that explicitly regulates the influence of time series and other modalities on the fused embeddings in  $E$ .

$$G = \sigma(W_g [E_{\text{ts}}; E_c] + b_g)$$

$$E = G \odot E_{\text{ts}} + (1 - G) \odot E_c$$

where  $\sigma(\cdot)$  denotes the sigmoid function, the learnable weight and bias are denoted as  $W_g \in \mathbb{R}^{2d \times d}$  and  $b_g \in \mathbb{R}^d$ , respectively.

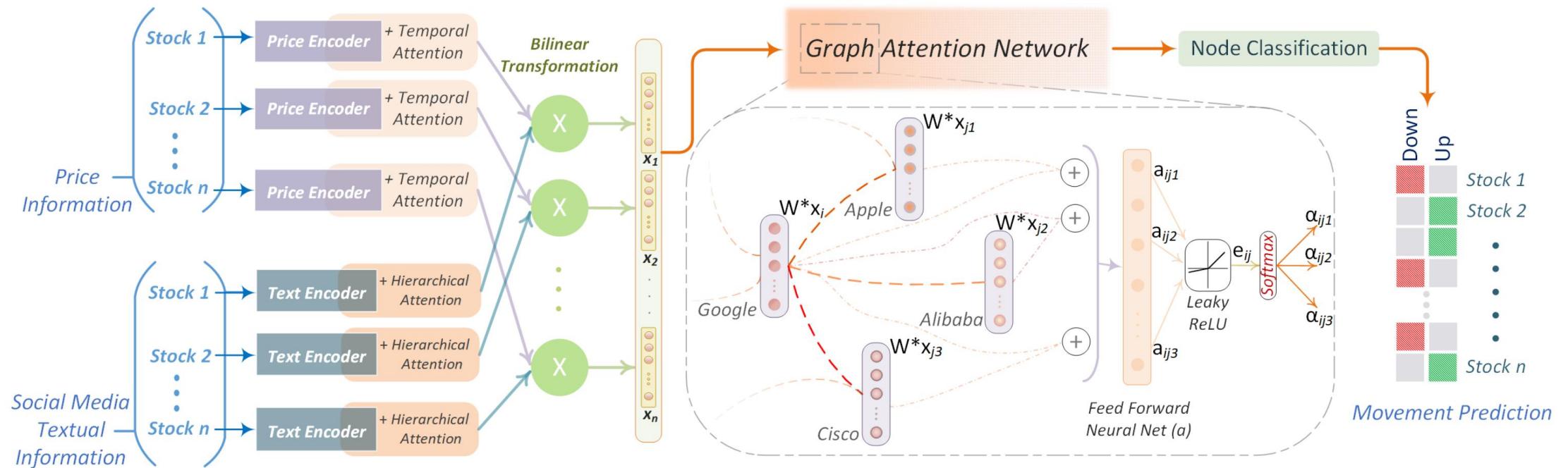
# Multi-modal Alignment with Time Series - Representations



**Time-VLM: Exploring Multimodal Vision-Language Models for Augmented Time Series Forecasting (ICML 2025)**

# Multi-modal Alignment with Time Series - Representations

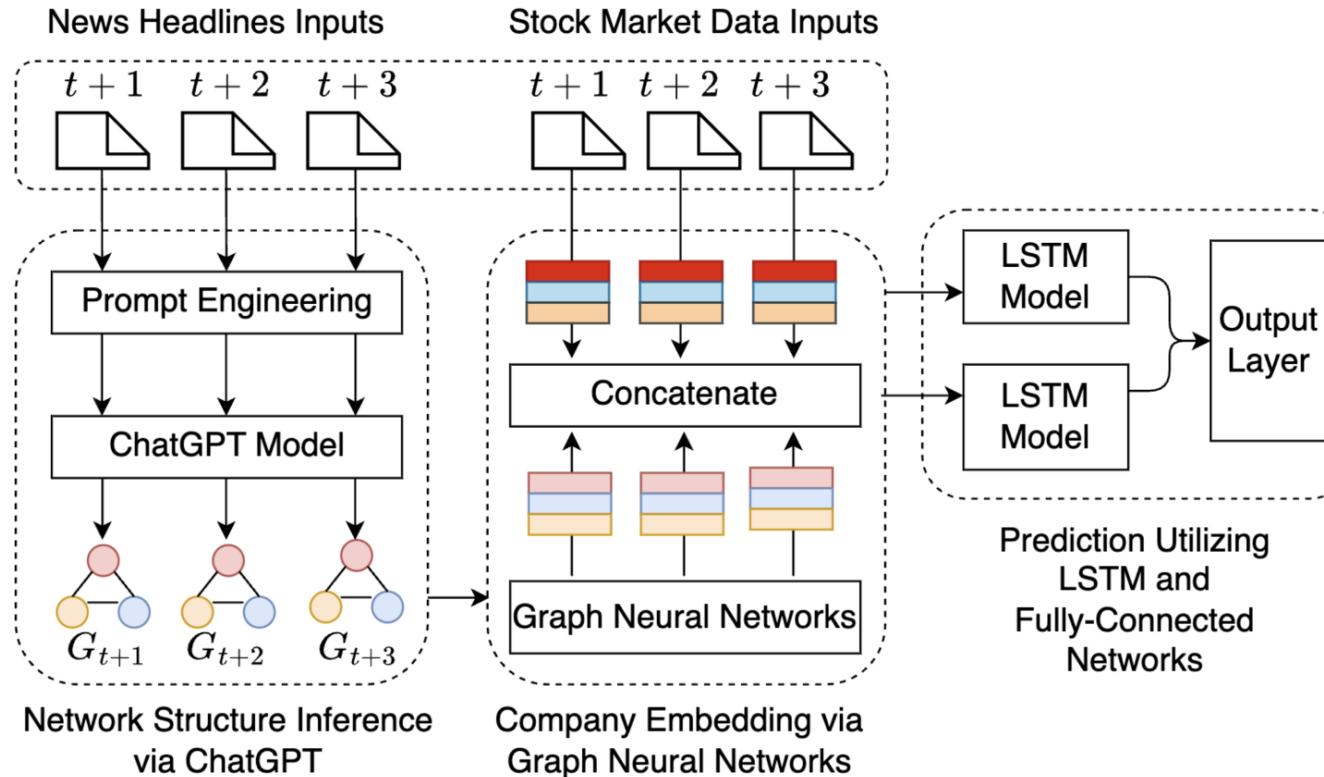
**Graph convolution:** The topological structure from external contexts can be used for alignment. It explicitly aligns representations with relational structures, enabling context-aware feature propagation across modalities.



**Deep Attentive Learning for Stock Movement Prediction from Social Media Text and Company Correlations (EMNLP-2020)**

# Multi-modal Alignment with Time Series - Representations

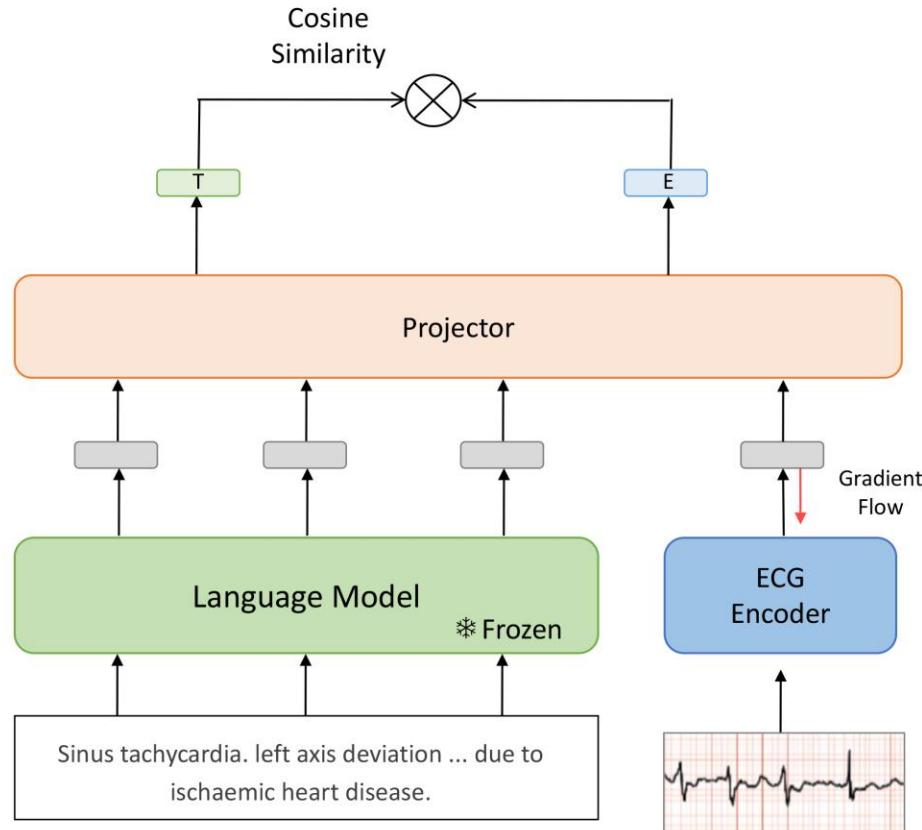
**Graph convolution:** The topological structure from external contexts can be used for alignment. It explicitly aligns representations with relational structures, enabling context-aware feature propagation across modalities.



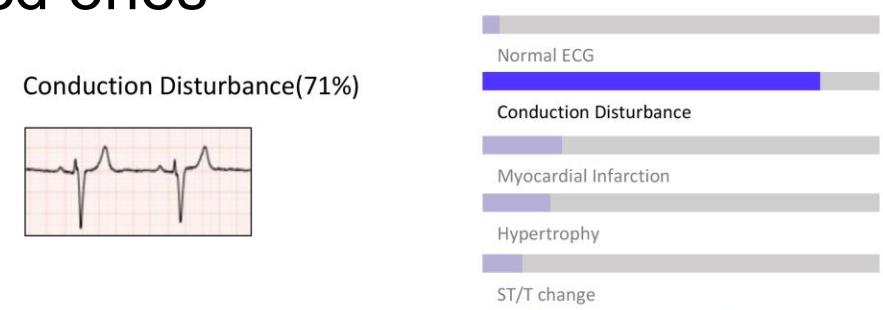
**ChatGPT Informed Graph Neural Network for Stock Movement Prediction (KDD23 Workshop)**

# Multi-modal Alignment with Time Series - Representations

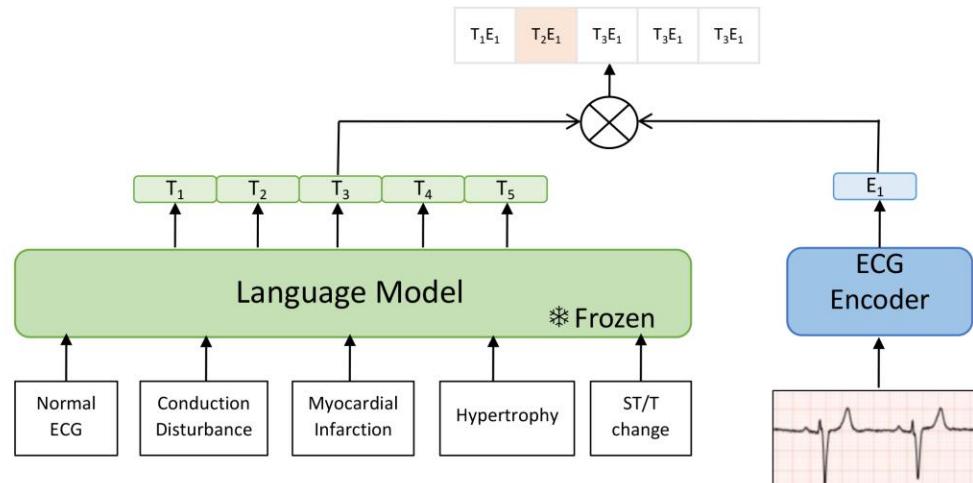
**Contrastive Learning:** maximize the cosine similarity between paired multi-modal embeddings and minimize that of unpaired ones



(a) Self-supervised Learning pre-training



(c) Visualization of Classification Results

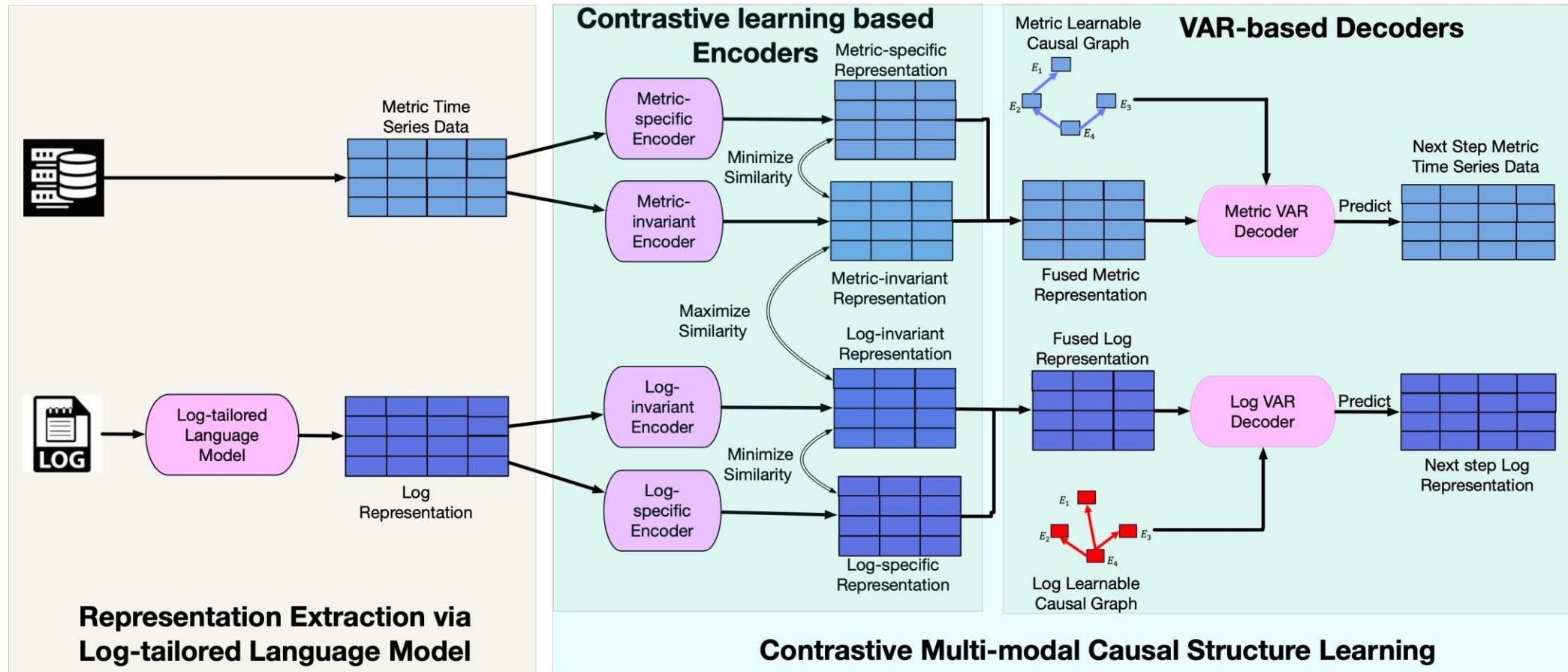


(b) Zero-Shot Learning for Classification

**Frozen Language Model Helps ECG Zero-Shot Learning (MIDL 2023)**

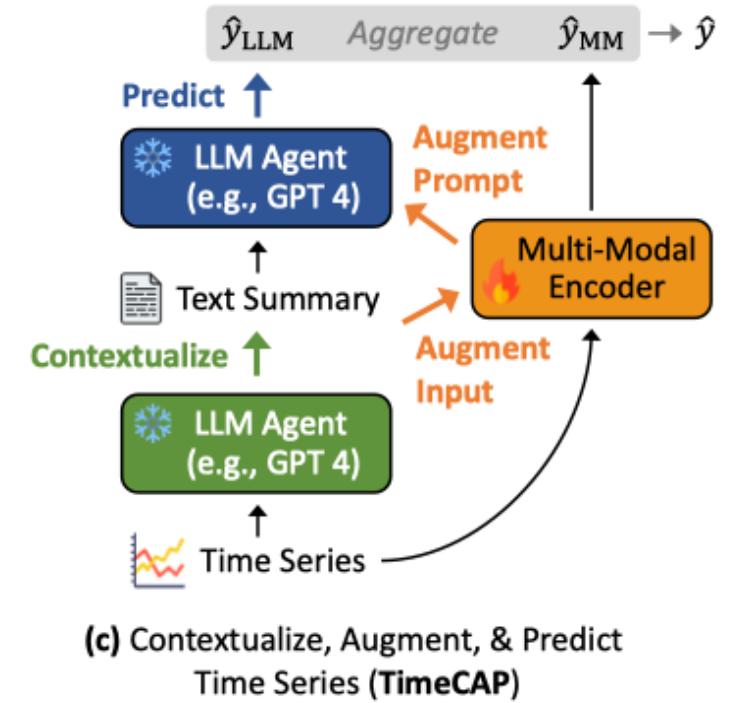
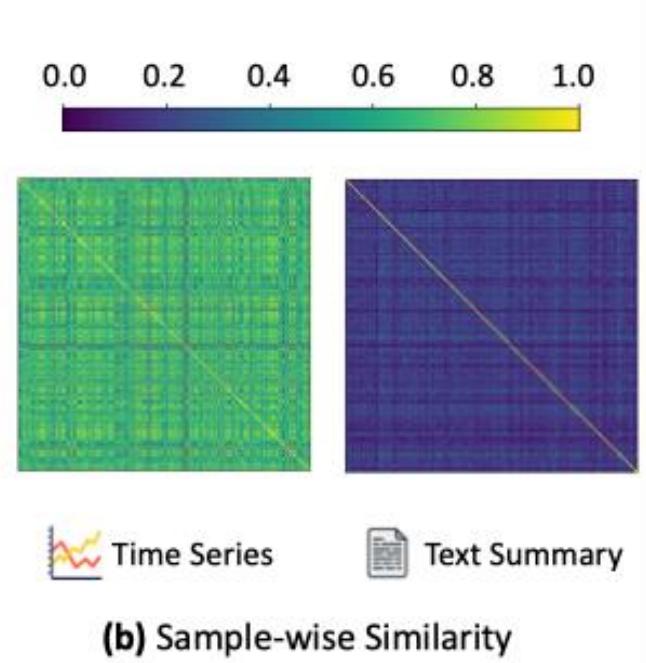
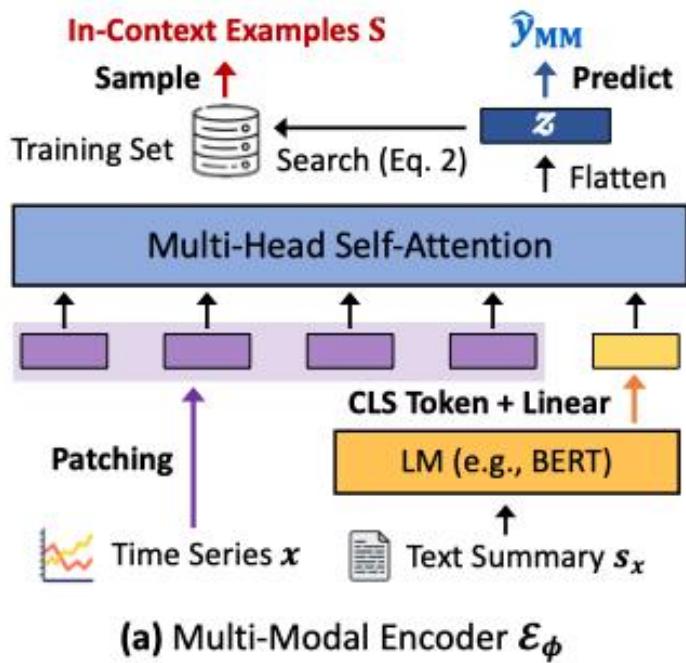
# Multi-modal Alignment with Time Series - Representations

**Contrastive Learning:** capture and regularize modal-invariant and modal-specific representations via mutual information



# Multi-modal Alignment with Time Series – Component Output

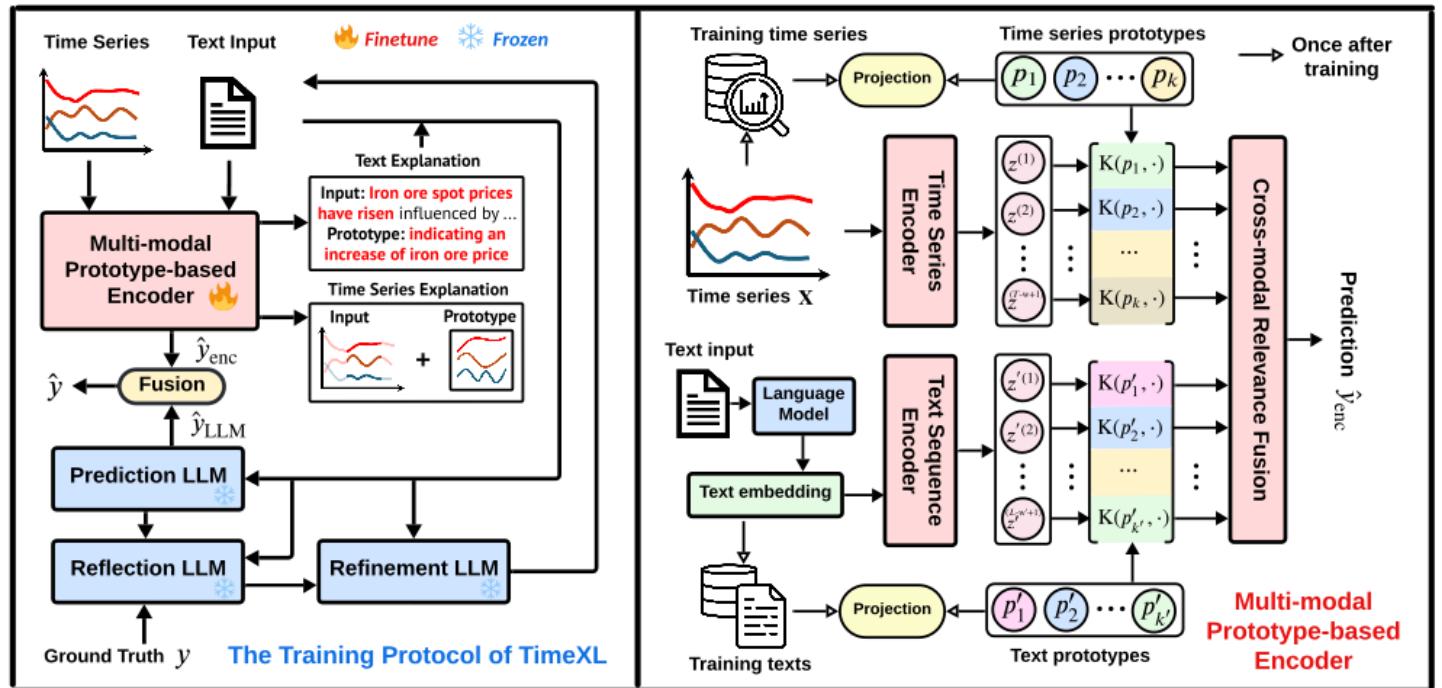
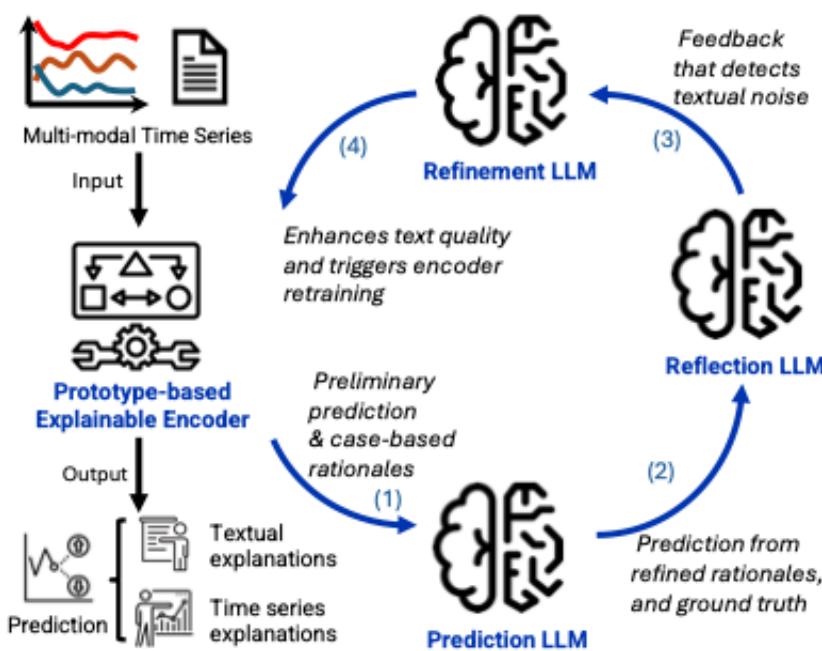
**Retrieval:** Augment LLM's input with in-context examples with the highest cosine similarity from a multi-modal embedding space



**TimeCAP: Learning to Contextualize, Augment, and Predict Time Series Events with Large Language Model Agents (AAAI 2025)**

# Multi-modal Alignment with Time Series – Component Output

**LLM Reasoning:** LLM agents collaborate to iteratively enhance prediction accuracy, identify textual inconsistencies or noise, and calibrate textual contexts.

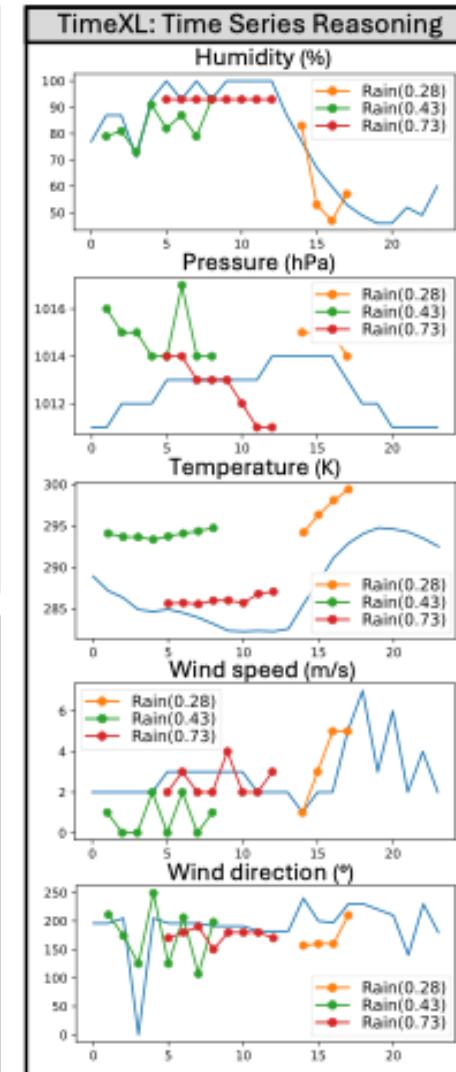


Explainable Multi-modal Time Series Prediction with LLM-in-the-Loop

# Multi-modal Alignment with Time Series – Component Output

Original Text Reasoning	Truth: Rain	Prediction: Not rain
<p><b>Original Text:</b> Over the past 24 hours, New York City experienced a gradual increase in temperature throughout the day, peaking in the late afternoon before a modest decline towards the evening. Humidity levels started high in the early hours, dropped significantly during the day, and then showed a rising trend at the end of the period, indicating a possible increase in moisture content in the air. Air pressure remained relatively stable with only minor fluctuations, suggesting a period of stable weather conditions. Wind speeds varied, with a noticeable increase during the late afternoon, which could have contributed to a brief period of cooler and more turbulent conditions. Wind direction was predominantly from the south-southwest, shifting slightly to a more south-southeast orientation later in the day, which is typical for the region's weather patterns during this time.</p> <p><b>Prototypes:</b></p> <p>Not rain: City has experienced relatively stable temperatures with a slight warming trend observed (0.78)</p> <p>Not rain: peak during the late afternoon before beginning to decrease slightly into the (0.64)</p> <p>Not rain: dropping again, indicative of typical diurnal variation (0.51)</p>		

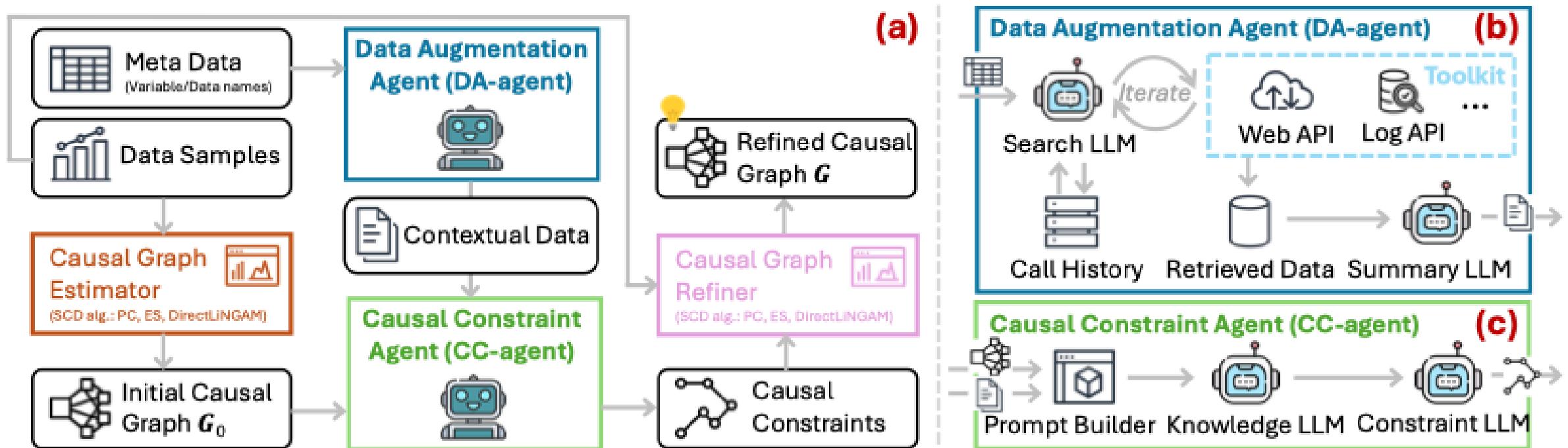
TimeXL: Text Reasoning	Prediction: Rain
<p><b>Refined Text:</b> Over the past 24 hours, New York City experienced a stable air pressure pattern with minor fluctuations, indicating stable weather conditions. The day saw a gradual increase in temperature, peaking in the late afternoon before declining in the evening. Humidity levels were high early on, dropped significantly during the day, and rose again later, suggesting increased moisture content. Wind direction shifted from south - southwest to south - southeast, bringing moisture-laden air, which could increase the likelihood of rain.</p> <p><b>Prototypes:</b></p> <p>Rain: direction was variable without a consistent pattern. These indicators suggest (0.47)</p> <p>Rain: wind direction started westerly, became variable, and (0.64)</p> <p>Rain: which could signal the approach of a weather system (0.53)</p>	



Explainable Multi-modal Time Series Prediction with LLM-in-the-Loop

# Multi-modal Alignment with Time Series – Component Output

**LLM Reasoning:** Align statistical causal discovery with textual reasoning by using iterative, self-reflective tool-calling to structure context and explain causal constraints



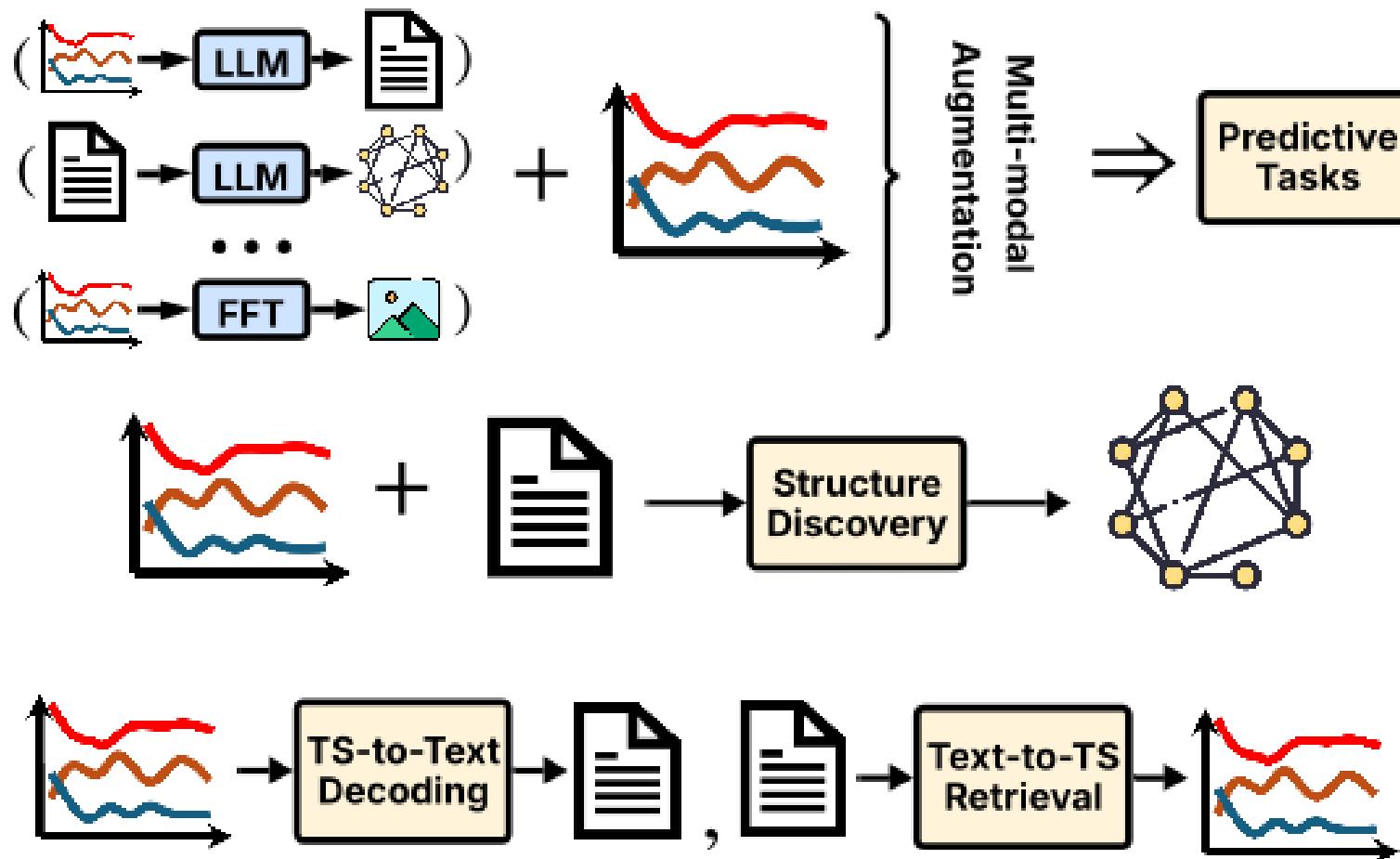
Exploring Multi-Modal Data with Tool-Augmented LLM Agents for Precise Causal Discovery (ACL 2025 Findings)

# Multi-modal Alignment with Time Series

- Alignment plays a crucial role in multi-modal interactions.
- It aims to calibrate and effectively capture relevant multi-modal elements for a semantically coherent modeling
- It enhances task performance, robustness and explanation, ensuring that models leverage meaningful contextual information for improved decision-making.

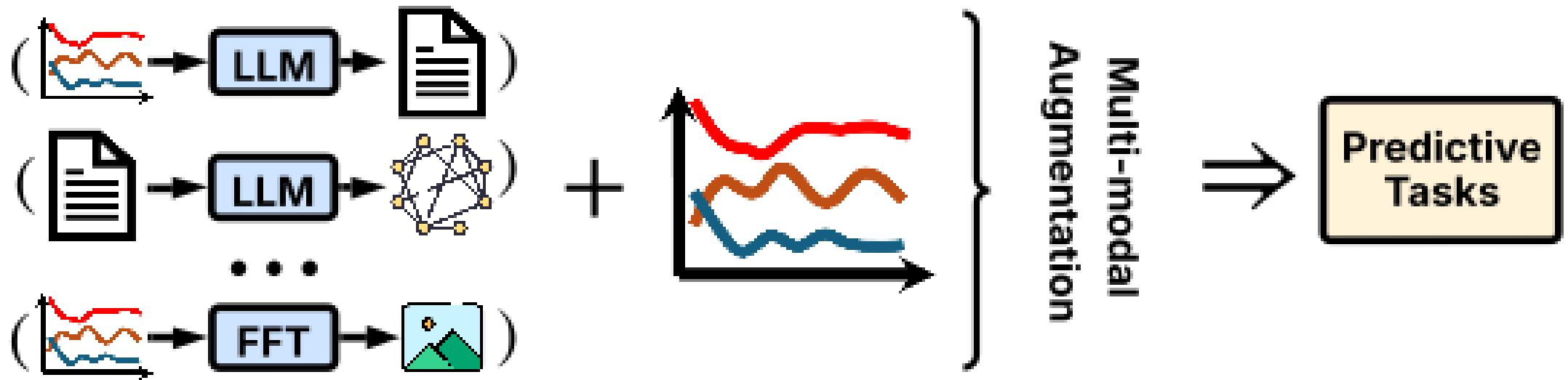
# Cross-modal Interaction with Time Series: Transference

Definition: the process of mapping between different modalities. It allows one modality to be inferred, translated, or synthesized from another.



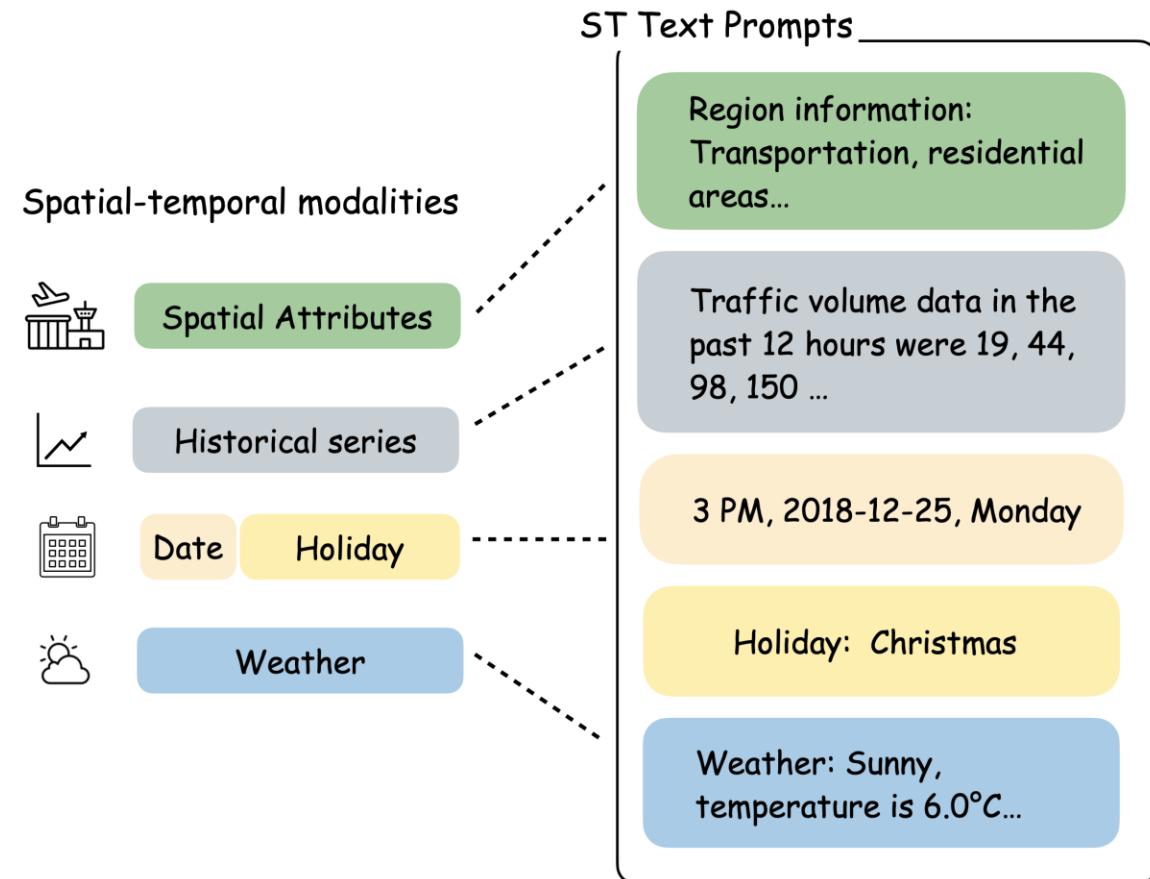
# Cross-modal Transference with Time Series: Input level

- The input-level transference typically serves for modality augmentation
- Introduce priors, enrich training samples, and provide alternative representations
- Particularly useful in scenarios of data scarcity and imbalance

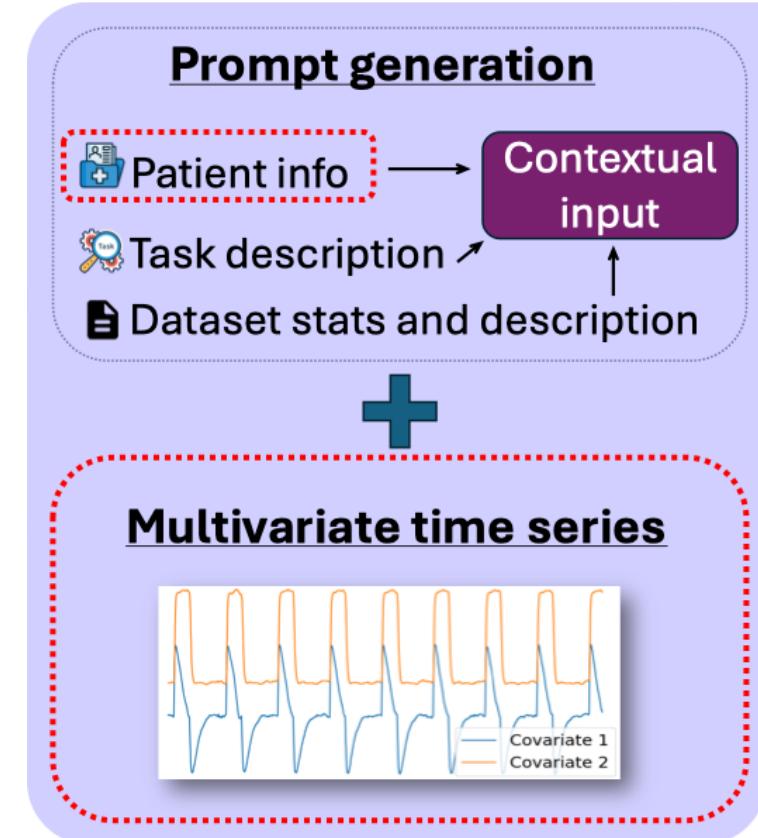


# Cross-modal Transference with Time Series: Input level

Use meta information to describe the narrative of real-world contexts (e.g., domain, data statistics and granularity, variable descriptions, other co-variates, etc.)



Towards Explainable Traffic Flow Prediction with Large Language Models



MedTsLLM: Leveraging LLMs for Multimodal Medical Time Series Analysis (MLHC 2024)

# Cross-modal Transference with Time Series: Input level

Generate a textual summary that offers contextual understanding of the input time series, leveraging the LLM's domain knowledge.

**System Prompt**

Your job is to act as [specific role/job]. You will write a high-quality report that is informative and helps in understanding the current [domain] situation.

**User Prompt**

Your task is to analyze [description of the time series data]. Review the time-series data provided for the [input length]. Each time-series consists of values separated by a 'l' token for the following indicators:

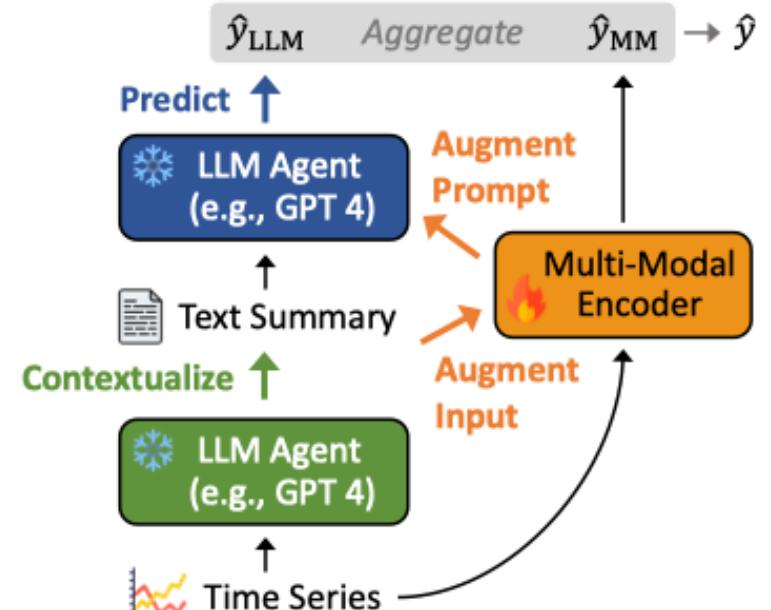
[Time Series Data]

Based on this time-series data, write a concise report that provides insights crucial for understanding the current [domain] situation. Your report should be limited to five sentences, yet comprehensive, highlighting key trends and considering their potential impact on [background]. Do not write numerical values while writing the report.

**Time Series**

**Text Summary**

Over the last 20 weeks, there has been a significant fluctuation in mortality due to Influenza or Pneumonia in relation to total deaths. Initially, the ratio of mortality from these causes was below the 20-week average of 7.84%, with a gentle increase from 6.15% to 6.83% by the 10th week. However, a notable spike occurred from the 11th week onwards, reaching a peak at the 14th week with a ratio of 10.55%, which correlates with an absolute increase in deaths attributed to Influenza or Pneumonia, particularly Influenza with a dramatic rise from 157 to 1536 deaths between the 11th and 14th weeks. Following this peak, there has been a gradual decline, yet the ratio remains above the average, settling at 9.20% in the 20th week. This trend suggests a potential strain on healthcare resources due to a surge in severe respiratory infections, necessitating increased vigilance and potentially enhanced public health measures.

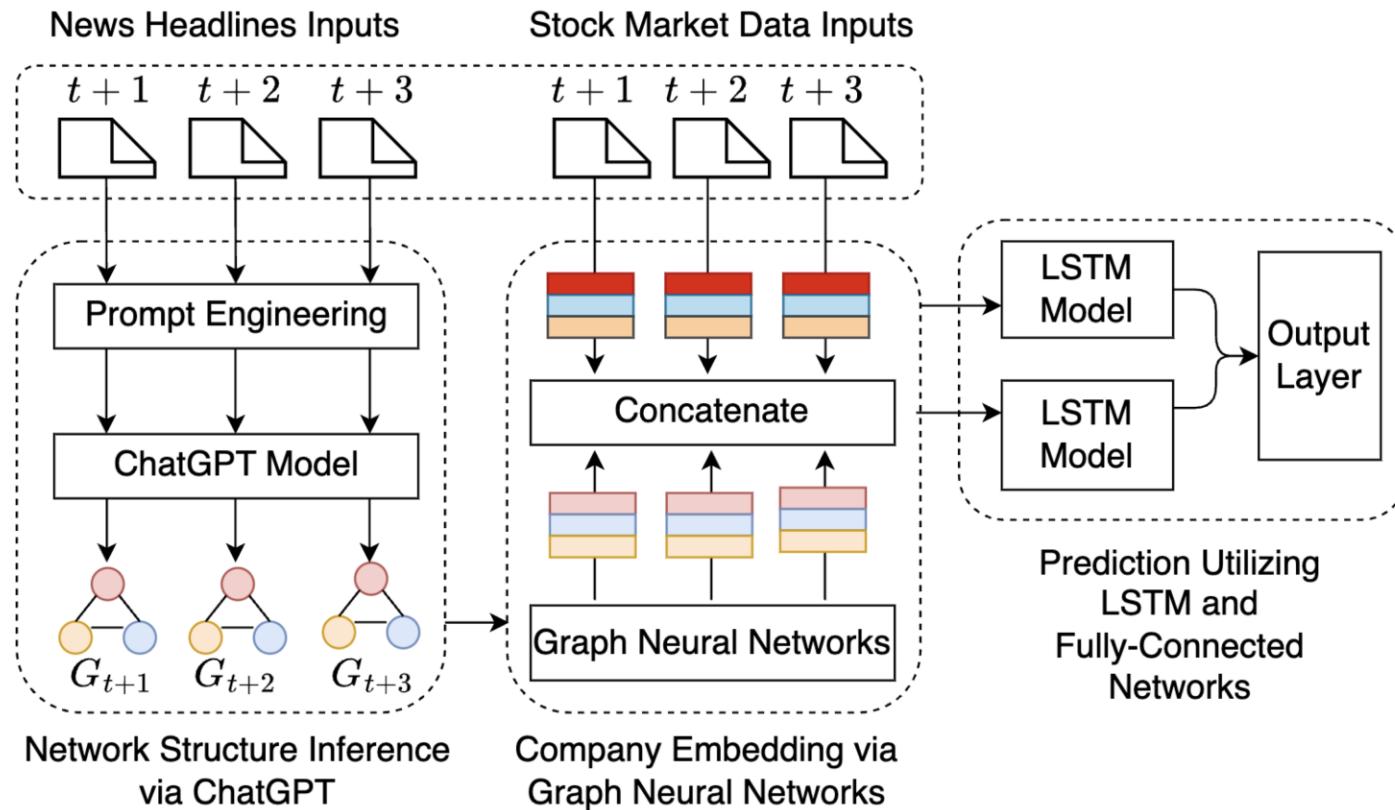


**(c) Contextualize, Augment, & Predict Time Series (TimeCAP)**

**TimeCAP: Learning to Contextualize, Augment, and Predict Time Series Events with Large Language Model Agents (AAAI 2025)**

# Cross-modal Transference with Time Series: Input level

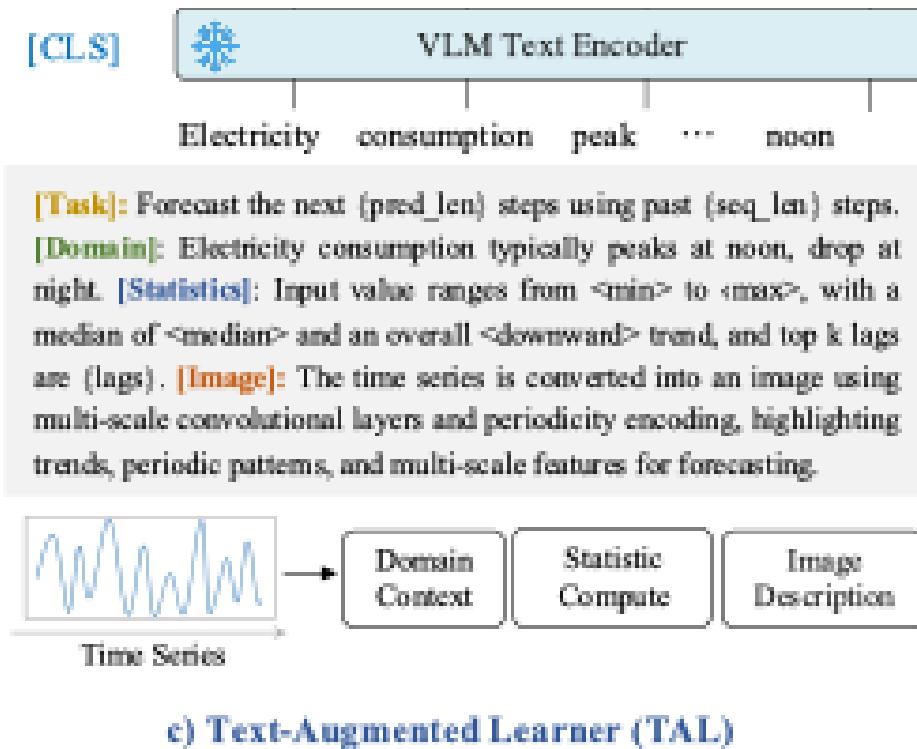
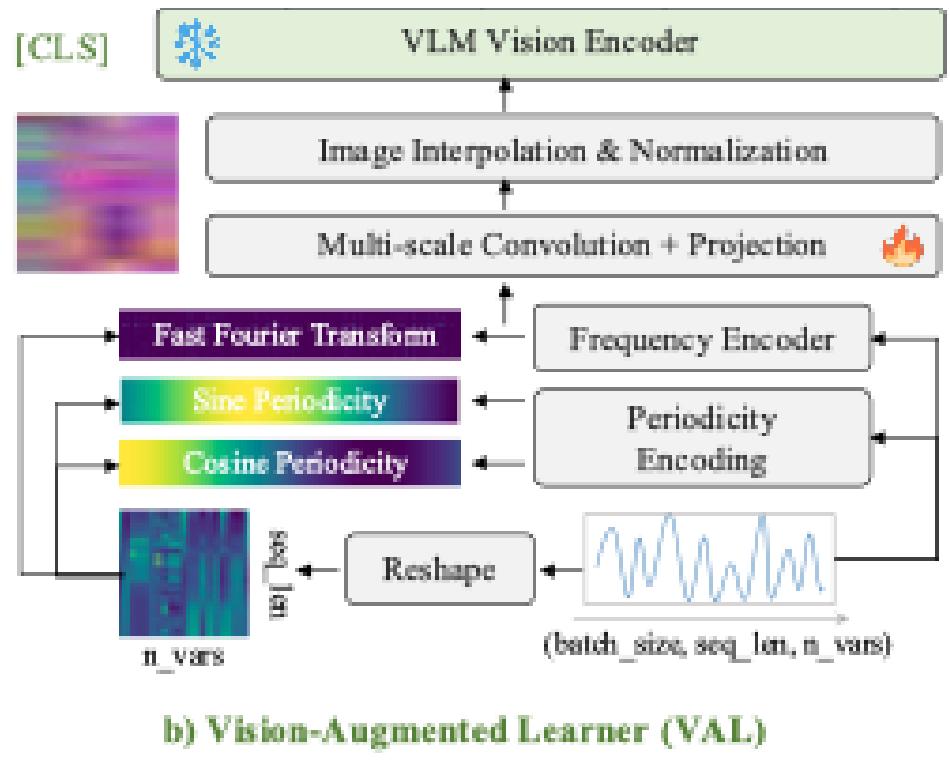
Generate a graph that represents real-world context of the input time series, leveraging the LLM's reasoning capability on news headlines.



**ChatGPT Informed Graph Neural Network for Stock Movement Prediction  
(KDD23 Workshop)**

# Cross-modal Transference with Time Series: Input level

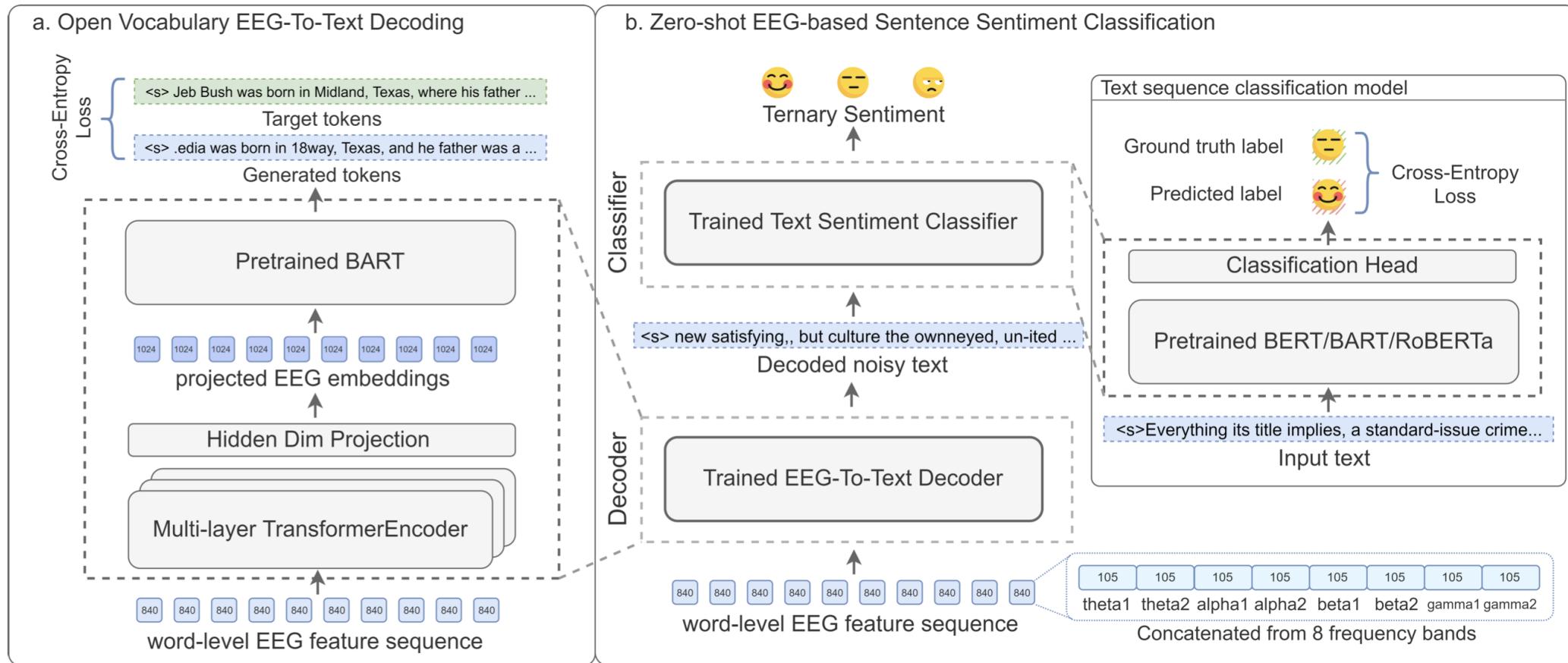
Generate image and texts that represents time series and its real-world contexts



**Time-VLM: Exploring Multimodal Vision-Language Models for Augmented Time Series Forecasting (ICML 2025)**

# Cross-modal Transference with Time Series: Intermediate

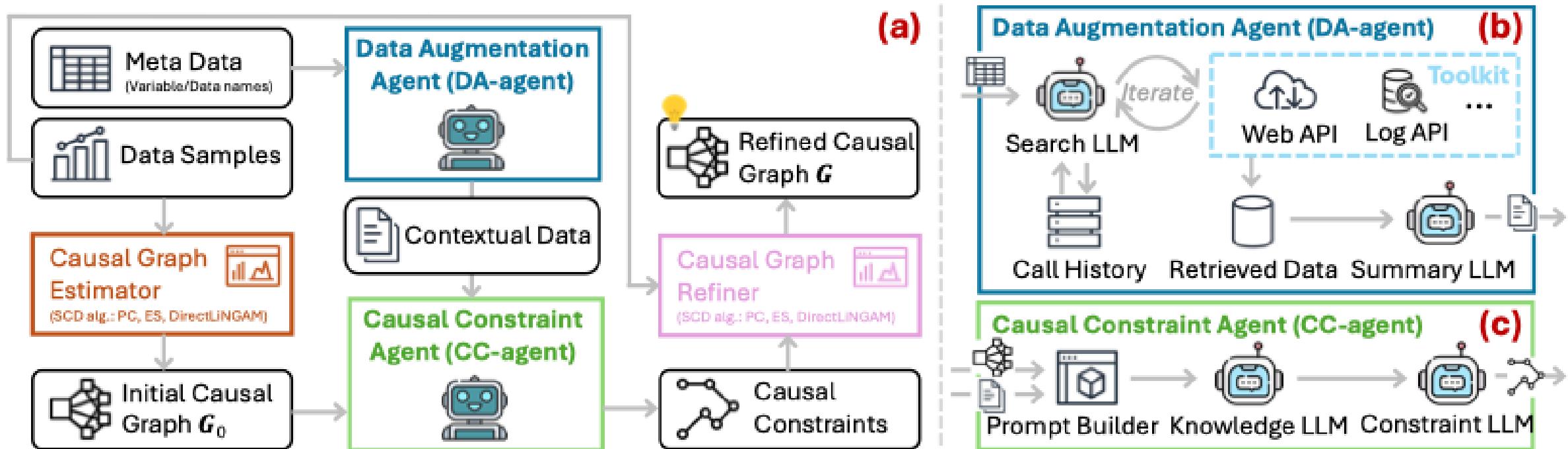
First converts EEG signals into reading texts, then employs a pretrained LLM for classification tasks



**Open Vocabulary Electroencephalography-To-Text Decoding and Zero-shot Sentiment Classification (AAAI 2022)**

# Cross-modal Transference with Time Series: Intermediate

Generate an initial causal graph from time series, followed by iterative refinements using textual context to improve alignment and interpretability.



# Cross-modal Transference with Time Series: Output

End-to-end generation of new modalities including text-based and image-based time series retrieval, where users provide textual descriptions or sketched trends to query relevant time series data

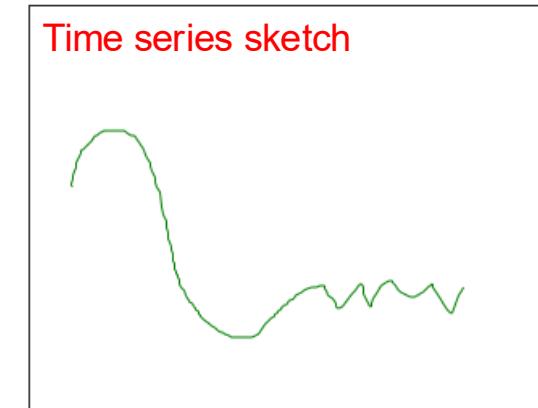
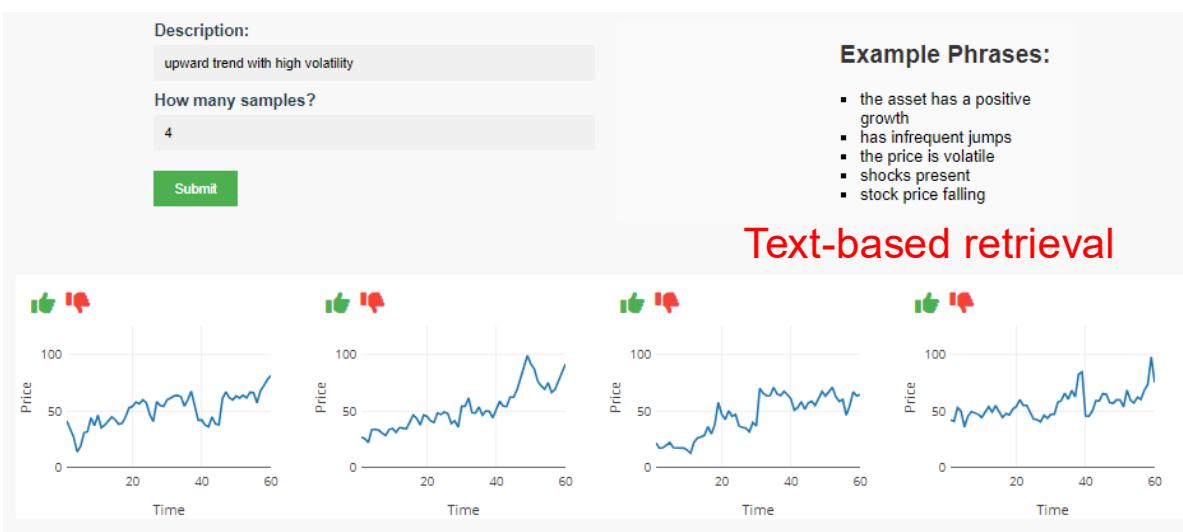
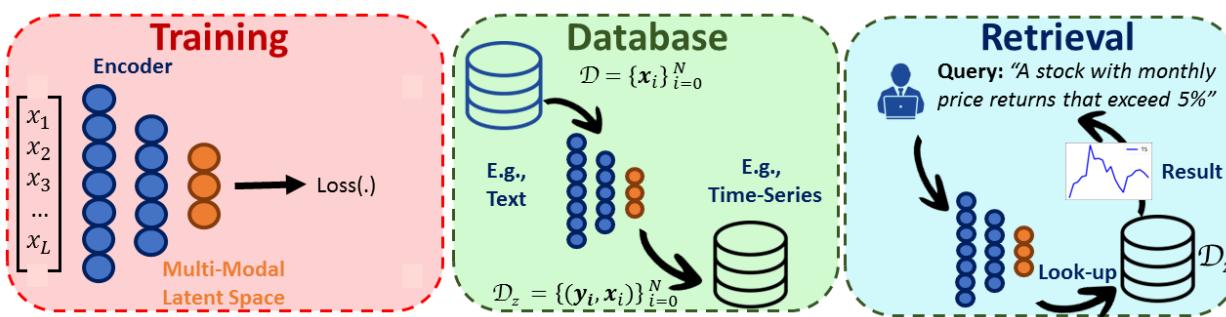
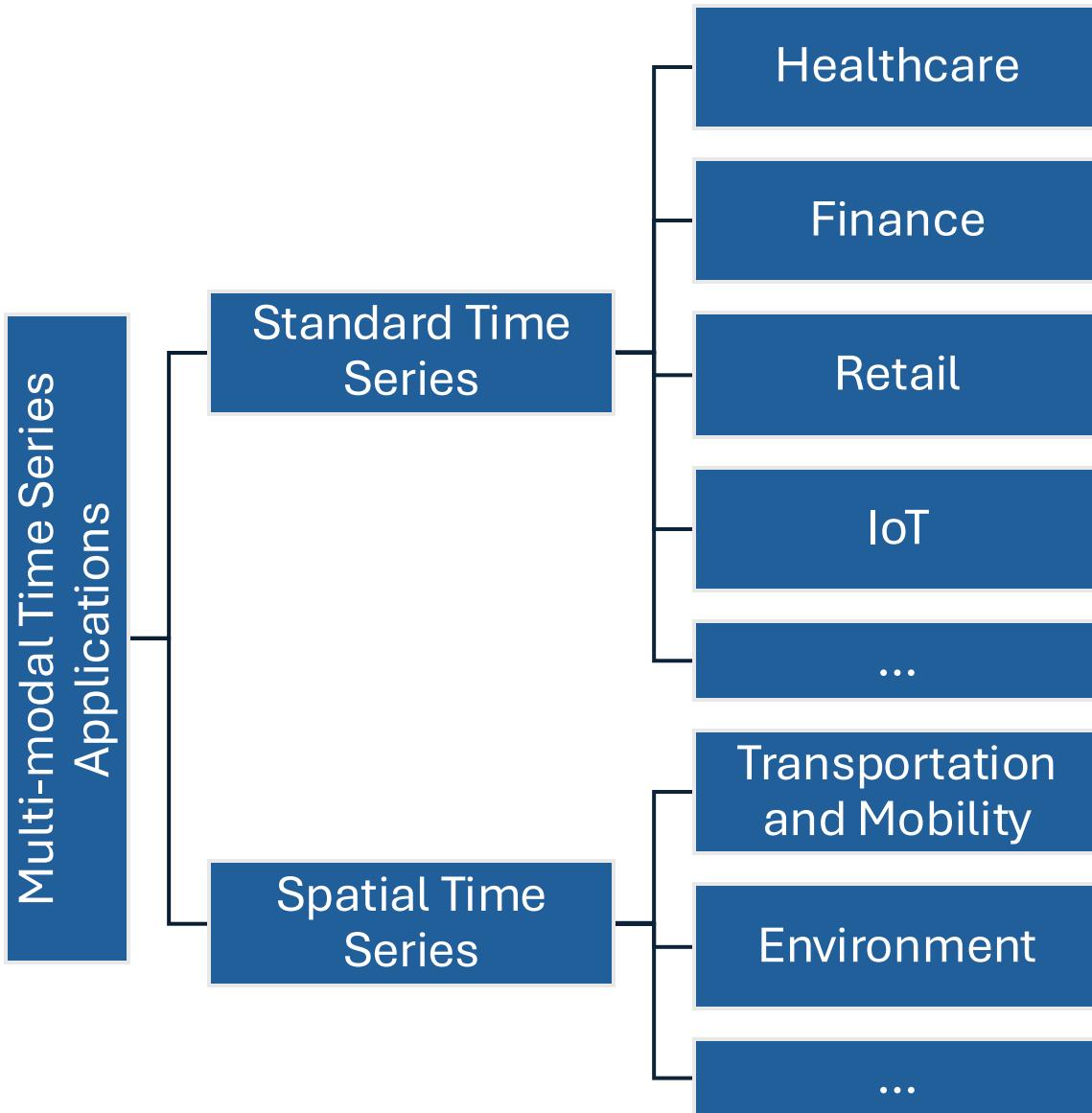


Image-based retrieval

# *Multi-modal Time Series Applications*

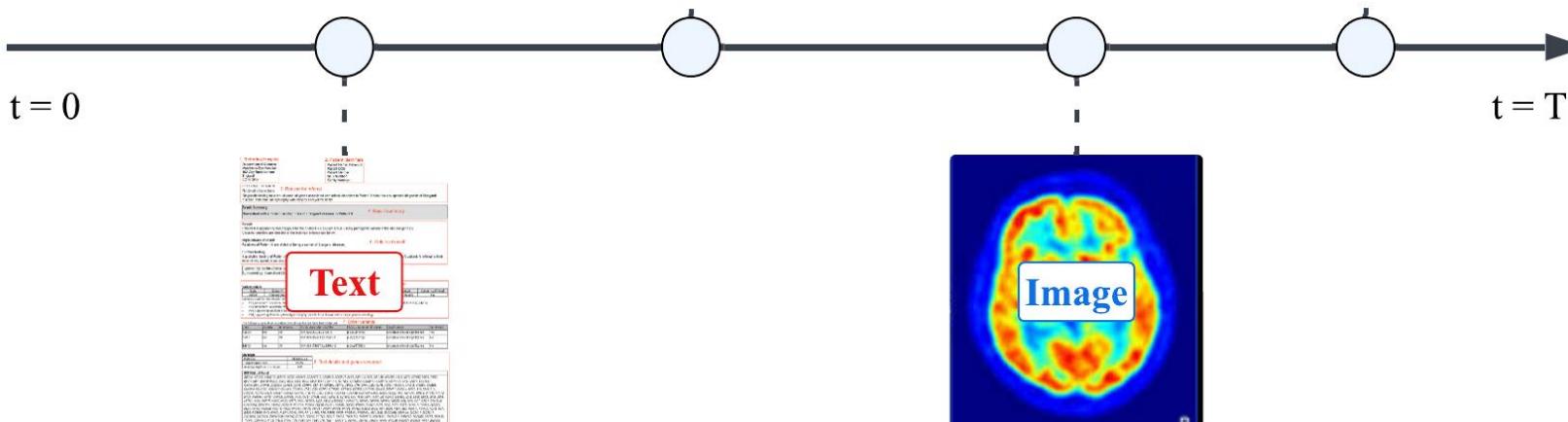
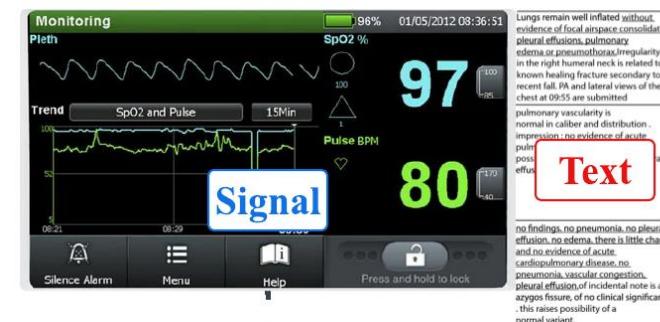
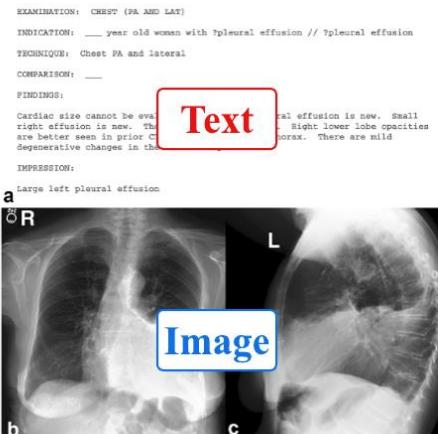
# Multi-modal Time Series Applications



- Covers real-world use cases of multi-modal time series
- **Domains:** Healthcare, Finance, Retail, IoT, Traffic, Environment, Speech
- **Types:** Standard Time Series vs Spatial Time Series
- **Task types:** *prediction, classification, retrieval, imputation*

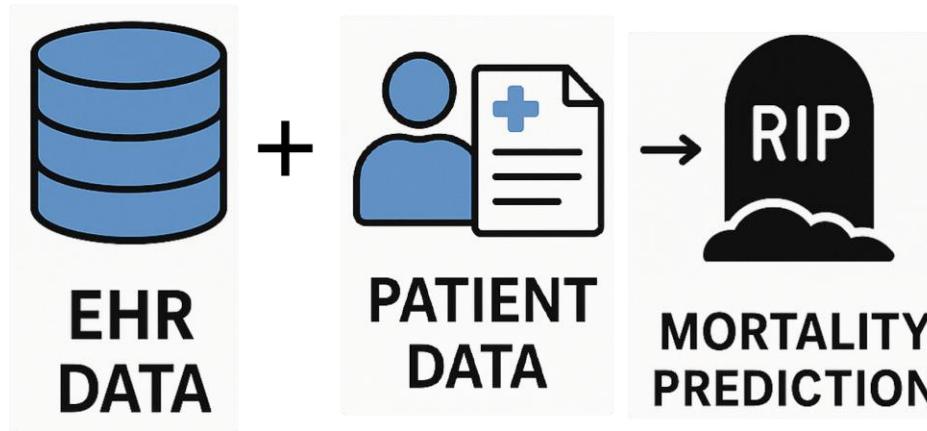
# Applications – Standard Time Series – Healthcare

- Healthcare Multi-modal Time Series Examples:
  - Electronic Health Records (EHR)



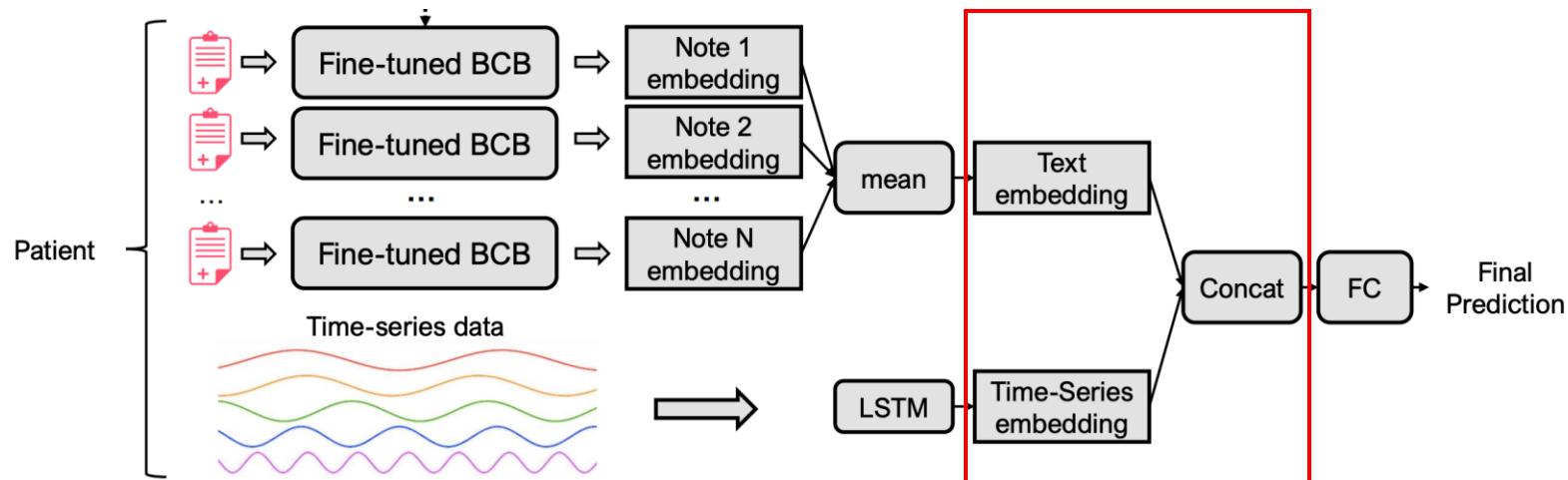
# Applications – Application on EHR

- In-hospital Mortality Prediction



# Applications – Application on EHR

- Leverage multi-modality data lab values and clinical reports
  - Concatenation

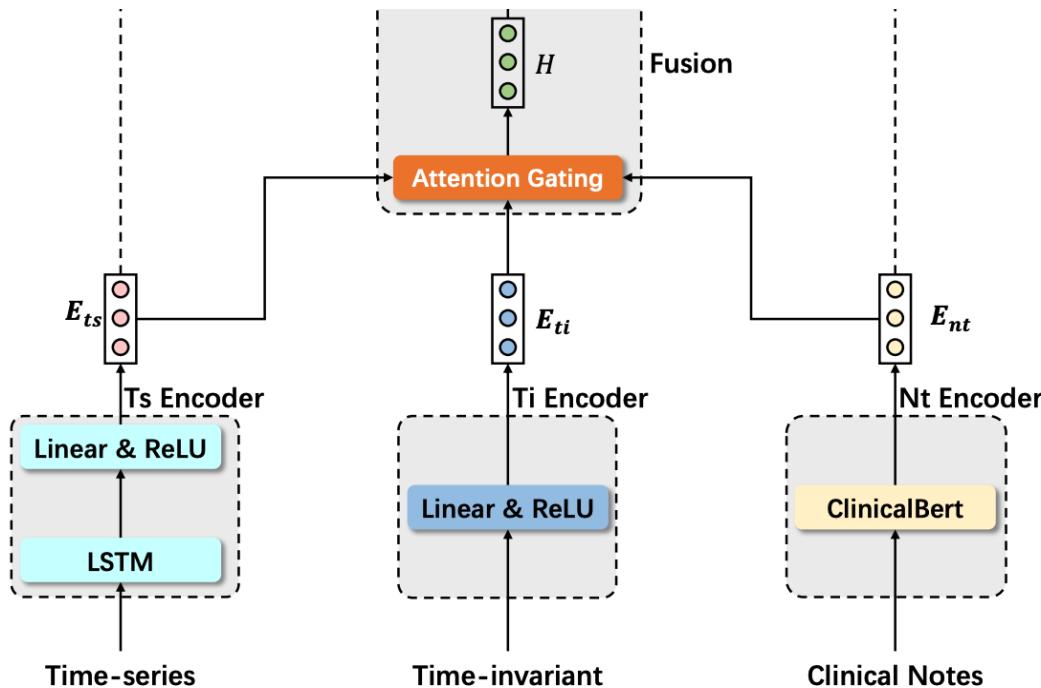


Predicting in-hospital mortality by combining  
clinical notes with time-series data (ACL-IJCNLP 2021)

# Applications – Analysis on EHR

- Leverage multi-modality data lab values and clinical reports

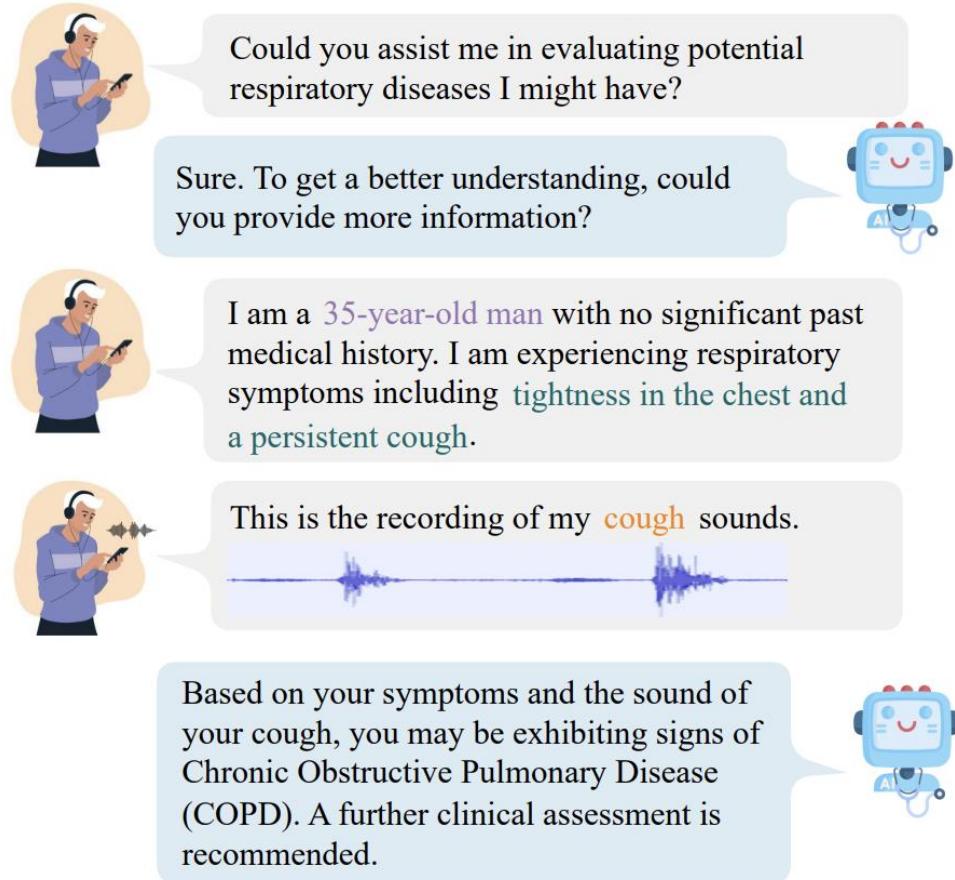
- Attention



**How to leverage multimodal EHR data for better medical predictions (2021)**

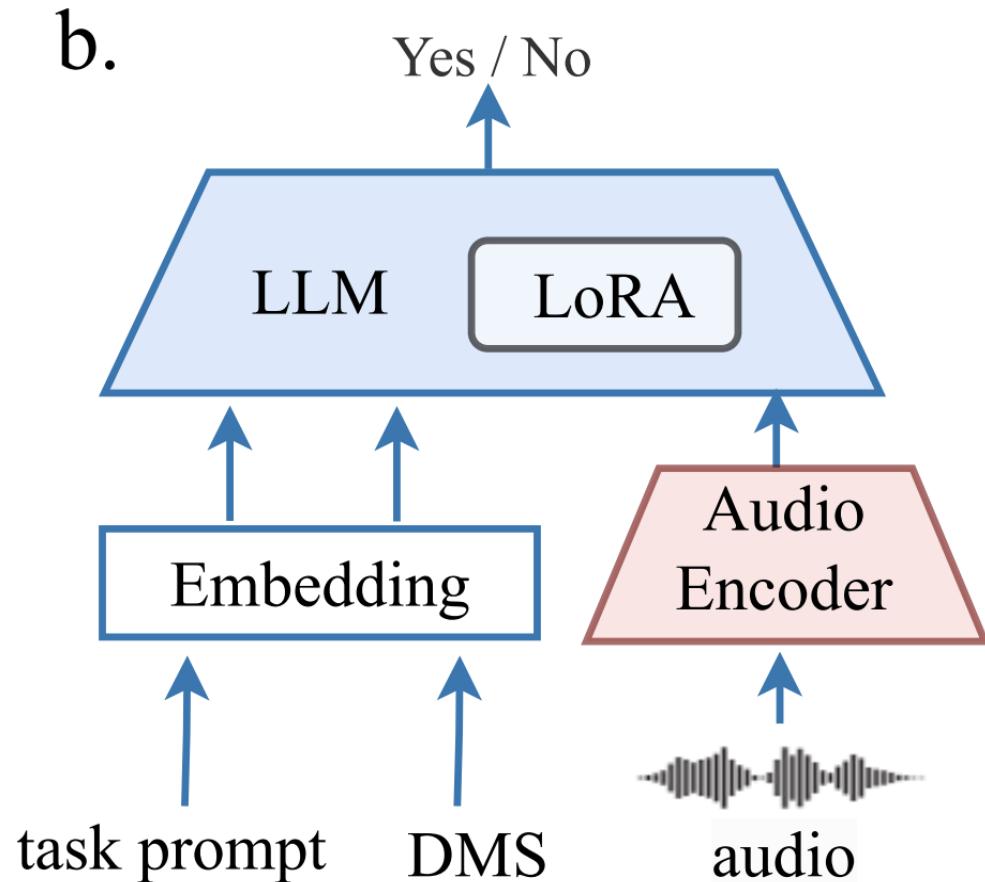
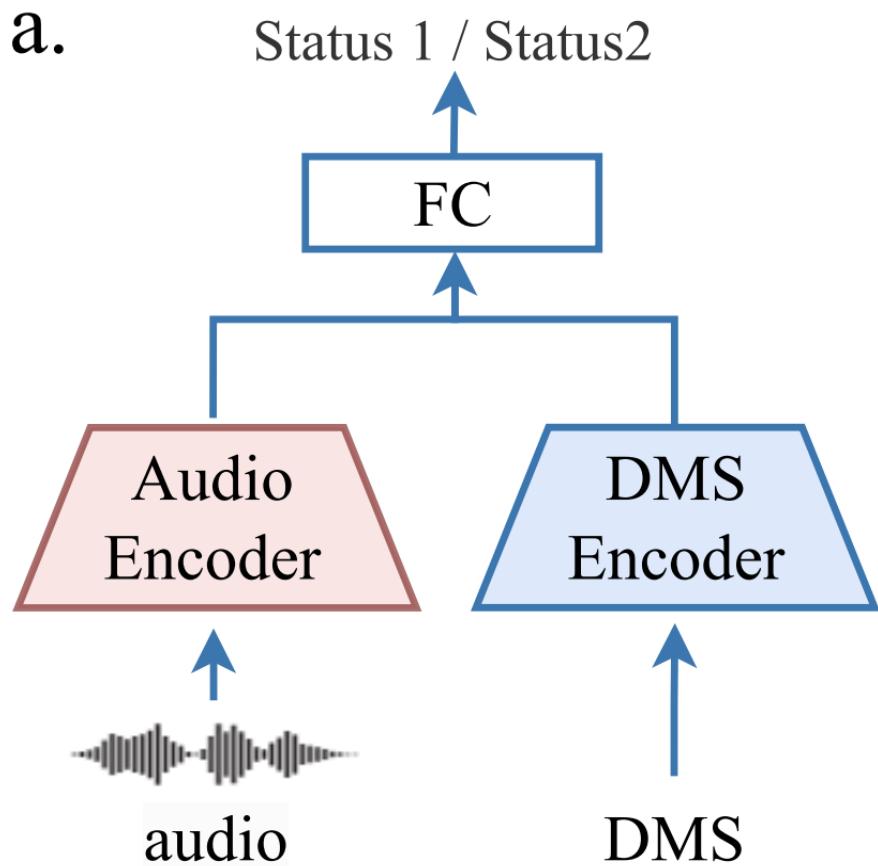
# Applications – Standard Time Series – Healthcare

- Healthcare Multi-modal Time Series Examples:
  - Incorporating with audio data for respiratory health screen



# Applications – Standard Time Series – Healthcare

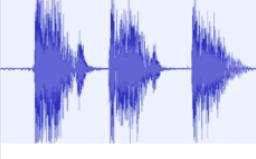
- Methods for respiratory health prediction



(a) Concatenation-based fusion method.

(b) LLM-based fusion method.

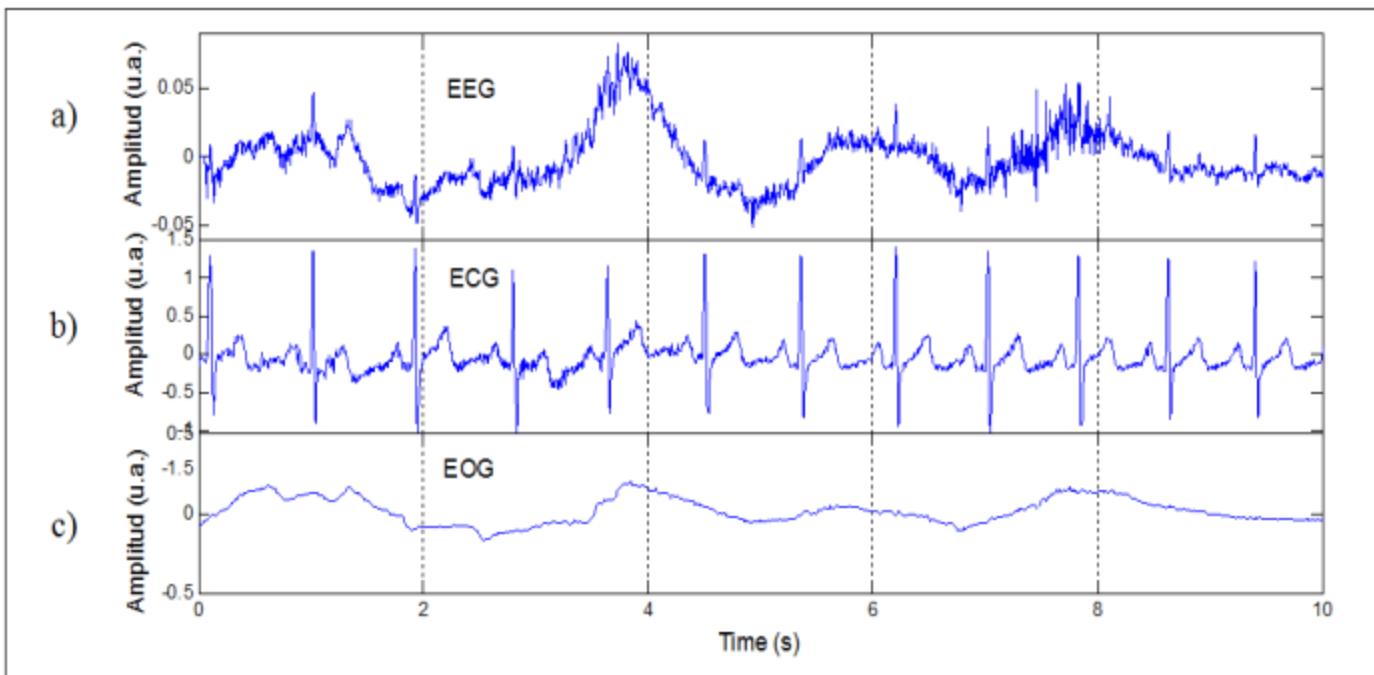
# Applications – Standard Time Series – Healthcare

Task	Text	Audio	Answer
S1 (Training)	<p><b>Task prompt:</b> Dataset description: This data comes from the UK COVID-19 Vocal Audio Dataset. Task description: classify whether the participant has COVID-19 given the following information and audio of the person's exhalation sounds. Please output 1 for COVID19, and 0 for non-COVID19.</p> <p><b>DMS text:</b> Gender: Female. Age: 45-64. Patient presents with the following medical history conditions: asthma. Patient presents with the following respiratory symptoms: cough, fatigue, headache.</p>		1
S6 (Training)	<p><b>Task prompt:</b> Dataset description: This data comes from the COVID-19 Sounds dataset. Task description: classify whether the person is a smoker or not given the following information and audio of the person's cough sounds. Please output 1 for smoker, and 0 for non-smoker.</p> <p><b>DMS text:</b> Gender: Female. Age: 50-59. Patient presents with no medical history conditions. Patient presents with no obvious respiratory symptoms.</p>		0
S7 (Training)	<p><b>Task prompt:</b> Dataset description: This data comes from the ICBHI Respiratory Sound Database Dataset. Task description: classify whether the person has Chronic obstructive pulmonary disease (COPD) given the following information and audio of the person's lung sounds. Please output 1 for COPD, and 0 for healthy.</p> <p><b>DMS text:</b> Gender: M. Age: 65. Record location: right posterior chest.</p>		1
T4 (Testing)	<p><b>Task prompt:</b> This data comes from the Coswara Covid-19 dataset. Task description: classify whether the participant has COVID-19 given the following information and audio of the person's breathing-deep sounds. Please output 1 for COVID19, and 0 for non-COVID19.</p> <p><b>DMS text:</b> Gender: male. Age: 35. Patient presents with the following respiratory symptoms: cold.</p>		0
T6 (Testing)	<p><b>Task prompt:</b> Dataset description: This data comes from the KAUH lung sound dataset, containing lung sounds recorded from the chest wall using an electronic stethoscope. Task description: classify whether the person has asthma given the following information and audio of the person's lung sounds. Please output 1 for asthma, and 0 for healthy.</p> <p><b>DMS text:</b> Gender: F. Record location: posterior right upper.</p>		1

**RespLLM: Unifying Audio and Text with Multimodal LLMs for Generalized Respiratory Health Prediction (2024)**

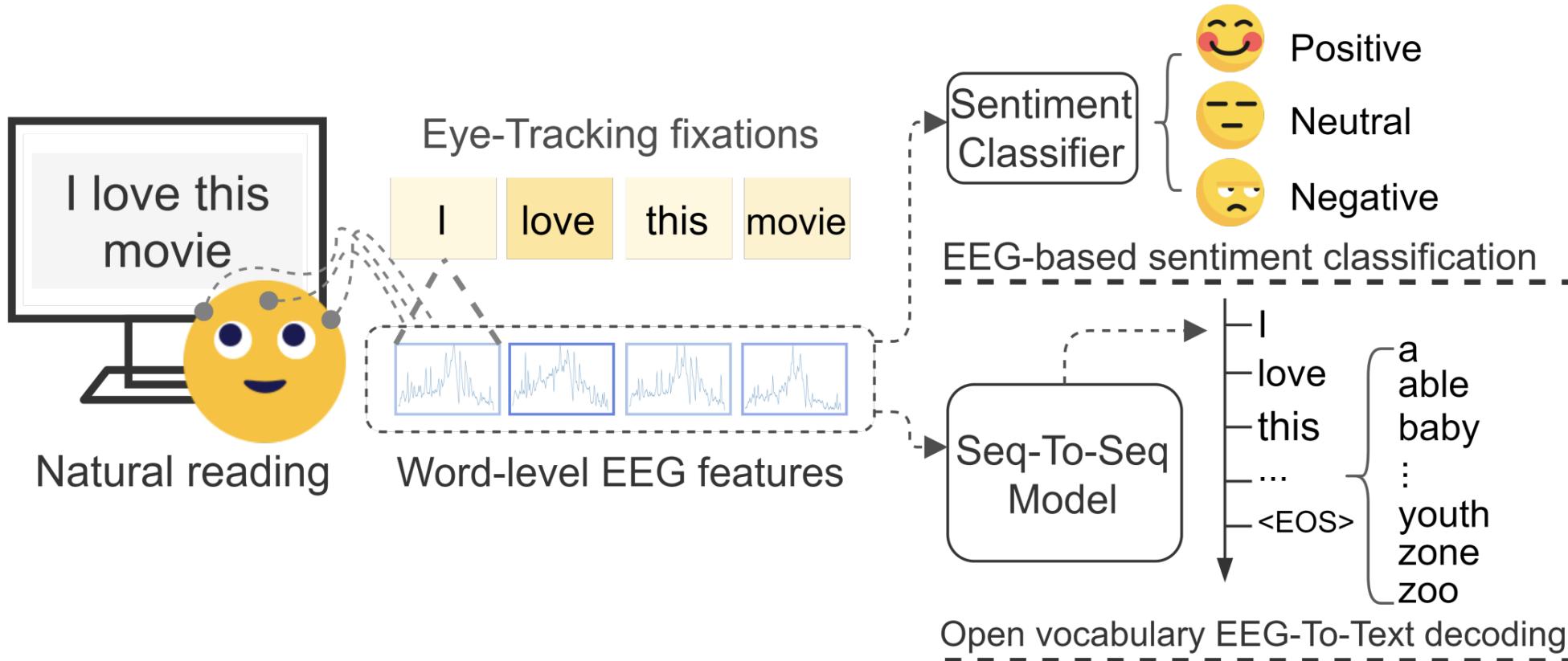
# Applications – Standard Time Series – Healthcare

- Healthcare Multi-modal Time Series Examples:
  - ECG/EGG



# Applications – Standard Time Series – Healthcare

- EEG data application: To-text decoding and sentiment analysis



**Open Vocabulary Electroencephalography-To-Text Decoding and Zero-shot Sentiment Classification (AAAI 2022)**

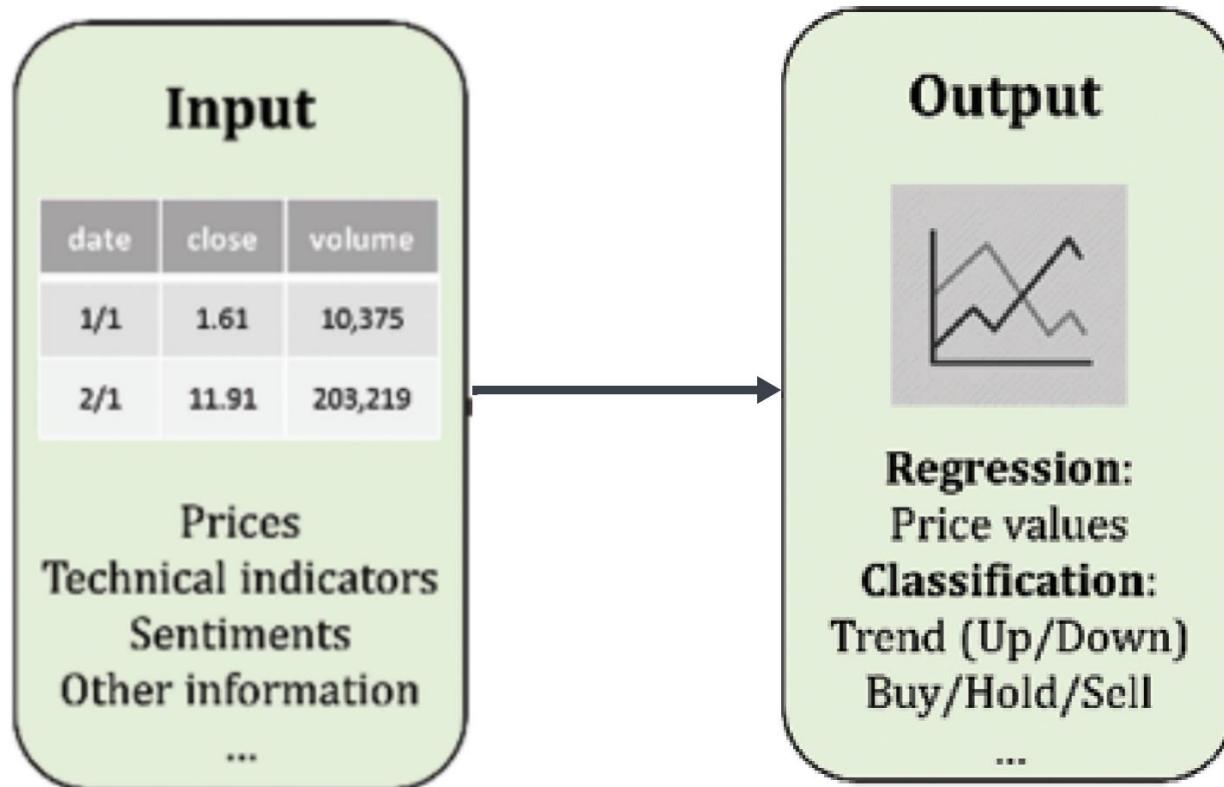
# Applications – Standard Time Series – Healthcare

- **Text decoding**

- 
- (1) | Ground Truth: He is a prominent **member of the Bush family**, the younger brother of **President George W. Bush**...  
| Model Output: was a former **member of the American family**, and son brother of **President George W. Bush**...
- 
- (2) | Ground Truth: Raymond Arrieta (born March 26, 1965 in San Juan, Puerto Rico) is considered by many to be one of **Puerto Rico's greatest comedians**.  
| Model Output: mond wasaga,19 in 17, 18) New Francisco, Puerto Rico) is a one many to be the of the **Rico's greatest poets**.
- 
- (3) | Ground Truth: He was first *appointed* to fill the Senate **seat** of Ernest Lundeen who had **died** in office.  
| Model Output: was a *elected* to the the position **seat** in the Hemy in **died** died in 18 in
- 
- (4) | Ground Truth: Adolf Otto Reinhold Windaus (December 25, 1876 - June 9, 1959) was a significant *German chemist*.  
| Model Output: rian Hitler,hardt,eren18 18, 1885 – January 3, 18) was a *German figure*- and
- 
- (5) | Ground Truth: It's *not a particularly good* film, but neither is it a *monsterous* one.  
| Model Output: was a a *bad good* story, but it is it *bad bad*. one.
-

# Applications – Standard Time Series – Finance

- **Data Modalities:** Stock prices, news, social media, company profiles
- **Tasks:** Stock return prediction, stock movement classification



# Applications – Standard Time Series – Finance

- Combine financial time series and text.

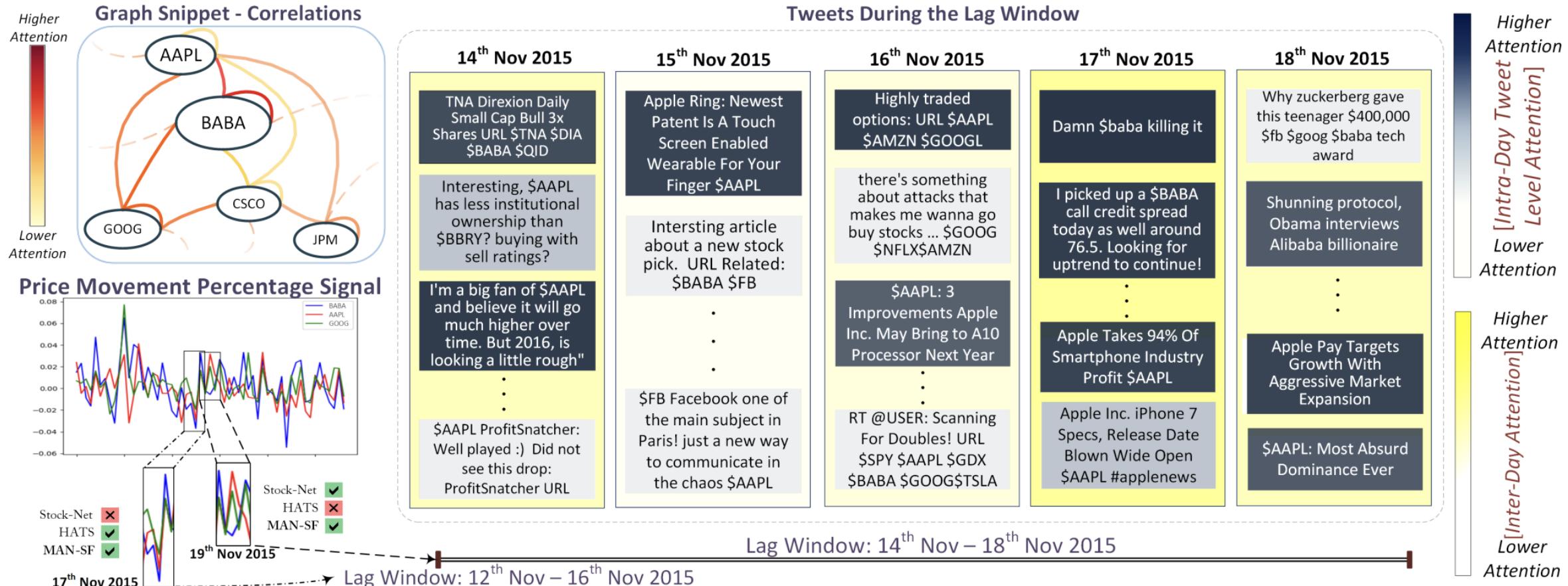
Type	Content
	<pre>data: date,open,high,low,close,adjusted-close,increase-in-5,10,15,20,25,30 2015-12-16,-0.45,0.78,-1.62,1.04,1.04,-1.63,-2.04,-2.52,-3.17,-3.53,-3.53 2015-12-17,-0.33,1.57,-0.49,0.33,0.33,-1.44,-2.01,-2.55,-3.38,-3.68,-3.70 2015-12-18,2.41,2.62,0.00,-2.85,-2.85,1.42,0.70,0.43,-0.30,-0.73,-0.87 2015-12-21,-0.72,0.31,-1.20,1.37,1.37,0.31,-0.53,-0.64,-1.44,-1.85,-2.13 2015-12-22,0.64,0.77,-1.05,0.03,0.03,0.26,-0.42,-0.57,-1.22,-1.74,-2.05 2015-12-23,-0.67,0.12,-0.96,1.06,1.06,-0.82,-1.17,-1.56,-2.01,-2.61,-2.99 2015-12-24,0.16,0.71,-0.04,-0.29,-0.29,-0.68,-0.69,-1.08,-1.54,-2.27,-2.58 2015-12-28,-0.06,0.24,-0.80,-0.01,-0.01,-0.24,-0.49,-1.04,-1.34,-1.98,-2.40 2015-12-29,-0.79,0.49,-0.93,1.26,1.26,-1.08,-1.39,-2.05,-2.25,-2.96,-3.37 2015-12-30,0.93,1.00,-0.22,-0.75,-0.75,-0.08,-0.54,-1.14,-1.38,-1.98,-2.48 tweets: 2015-12-23: fxi ishares ftse china 25 index fund ask\$fxi \$gpro \$uco \$unh #fxi #finance #stocks gbsn great basin scientific,... 2015-12-24: \$unh:us looking for value? why it might be time to try unitedhealth group (unh) #unitedhealth group inc #sp500 2015-12-25: #dow #stocks \$unh unitedhealth group daily:-0.29% weekly:-0.74% ytd:17.46% trend:73 / 100 2015-12-26: vz verizon communications company info\$vz \$apl \$unh \$bib #vz #stock #stocks \$unh \$unh #spx \$spx \$dja #dja #stocks #trading #traders #markets #investing #money more ... 2015-12-28: buzzers of healthcare sector: unitedhealth group inc \$unh merck 2015-12-30: uvxy proshares ultra vix short-term futures etf investor opinions \$uvxy \$dust \$ddd \$unh #uvxy #pennystocks... Consider the data and the investor sentiment in tweets to predict whether the close price movement of unh will rise or fall at 2015-12-31. Start with Rise or Fall, then explain your predictions step by step.</pre>
Prompt	

Response	Fall. Here are the reasons: 1. Looking at the historical data, \$ppl has been showing a downward trend for several days until 2015-12-30, with negative values for open, high, low and close prices. 2. The adjusted-close price is also decreasing consistently throughout this period. 3. In terms of the "increase-in-x" columns in the dataset, we can see that all indicators show negative values as well, indicating an overall bearish sentiment towards \$ppl during this time frame. 4. Moving on to investor sentiment expressed through tweets during this period: while there was some positive news about KBC group acquiring shares of \$ppl on 2015-12-21, it seems that there were no significant positive sentiments being shared in other tweets around that time. 5. Moreover, on December 22nd and December 28th we can see two negative tweets mentioning PPL Corp specifically - one highlighting a downgrade by Zacks Investment Research and another
Prediction	Fall
Truth	Fall

## The Wall Street Neophyte: A Zero-Shot Analysis of ChatGPT Over Multimodal Stock Movement Prediction Challenges (2023)

# Applications – Standard Time Series – Finance

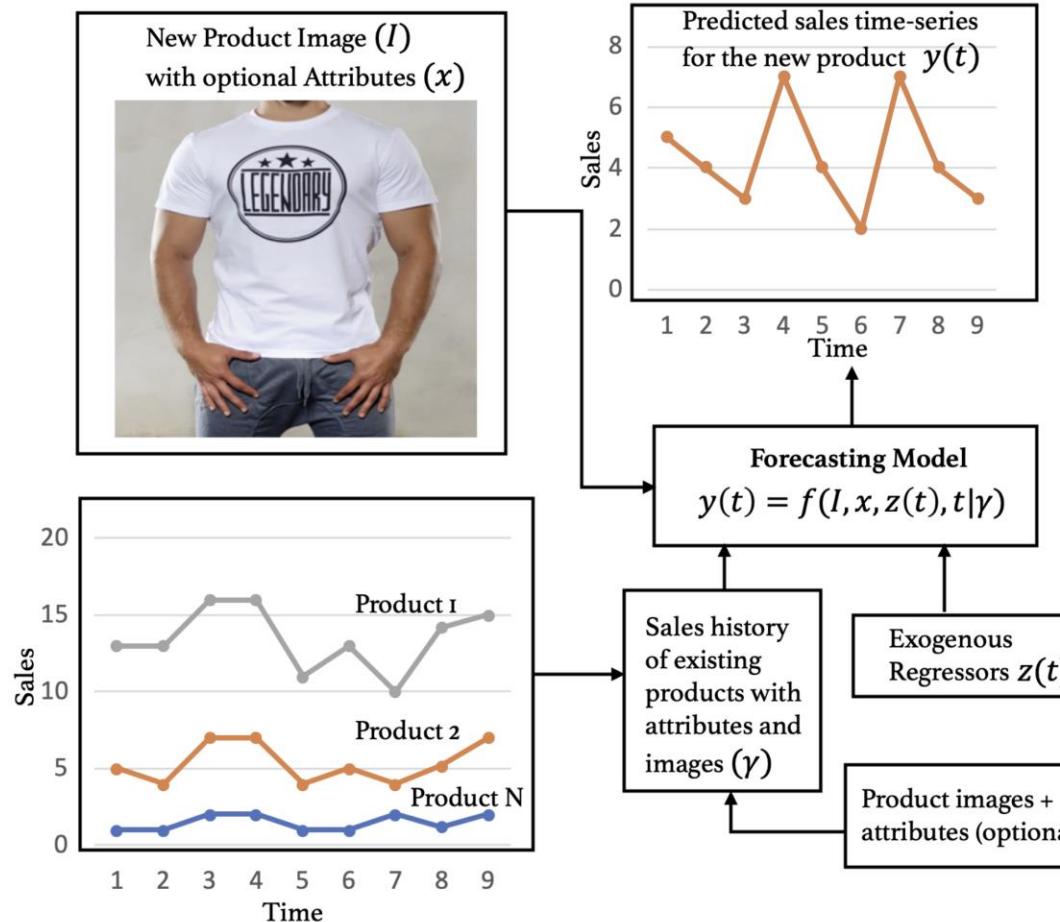
- Financial time series, text and graph



# Deep Attentive Learning for Stock Movement Prediction From Social Media Text and Company Correlations (ACL 2020)

# Applications – Standard Time Series – Retails

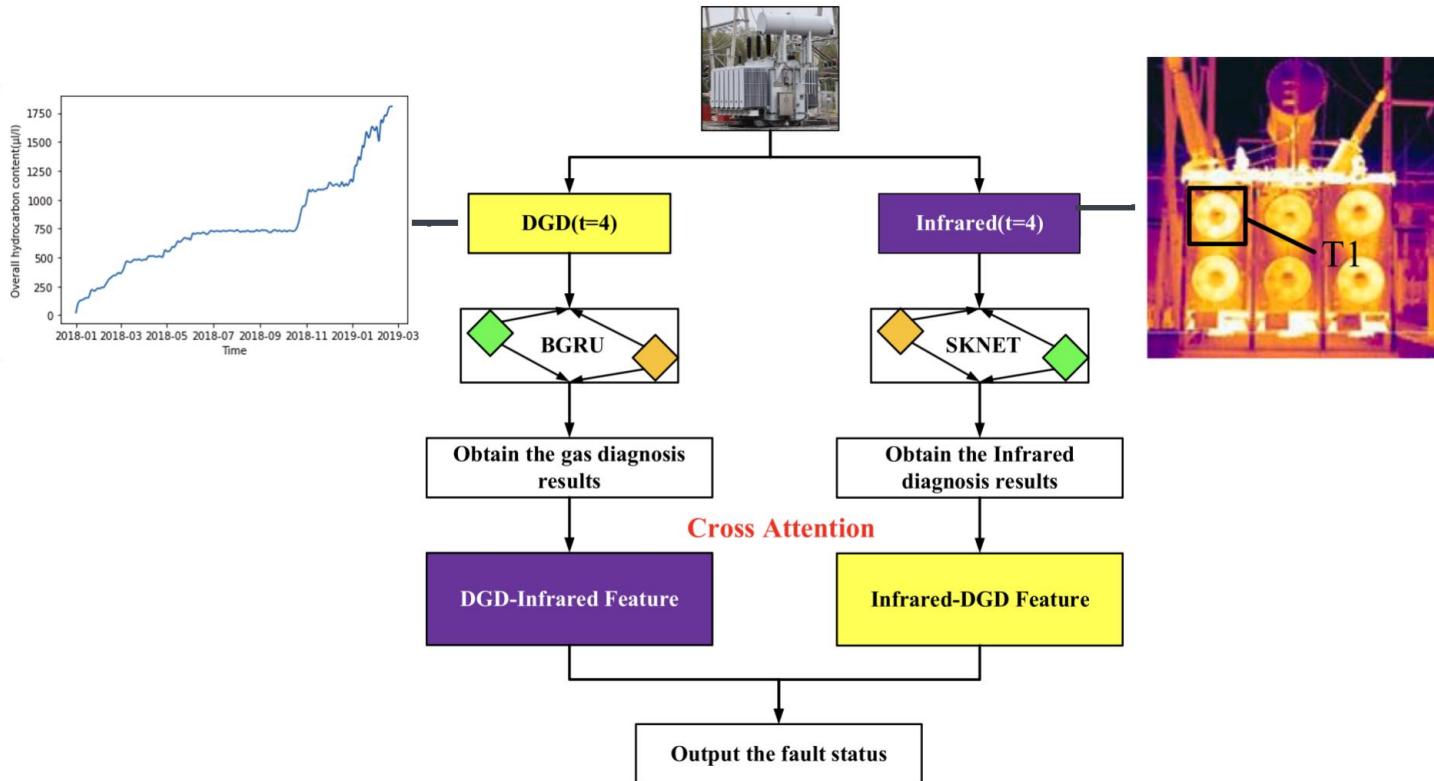
- Sales forecasting using product images and historical TS.



Attention based Multi-Modal New Product Sales Time-series Forecasting (KDD 2020)

# Applications – Standard Time Series – IoT

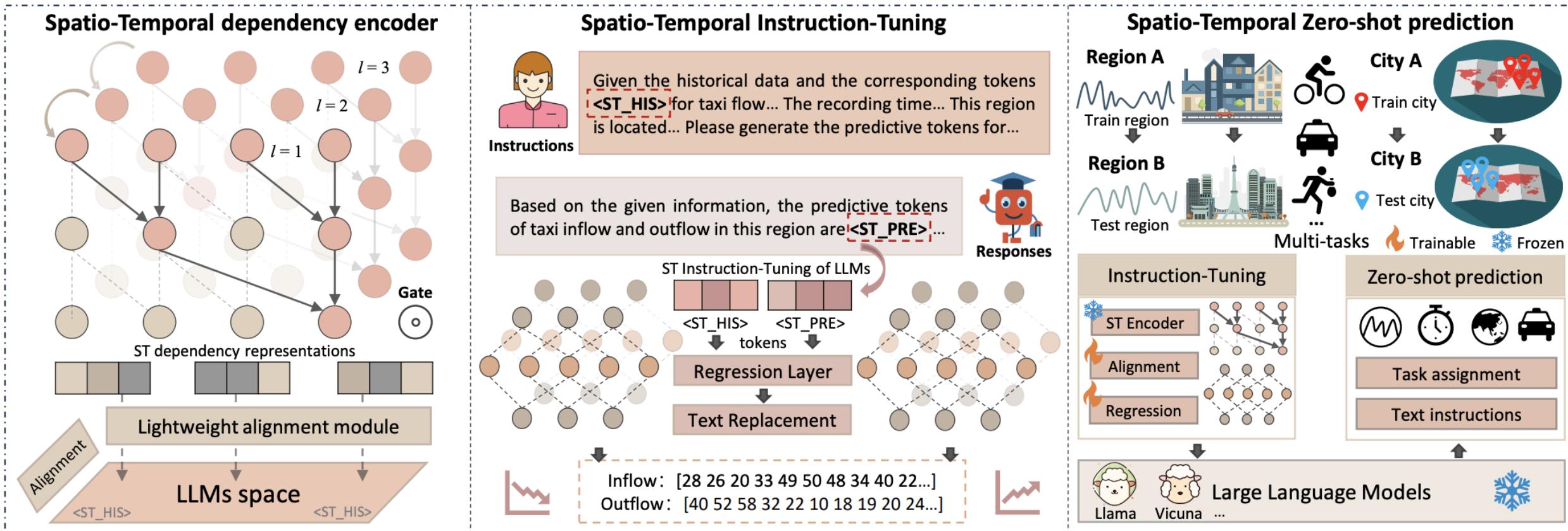
- Power transformer fault diagnosis using dissolved gas analysis (TS) and infrared images.



**Multi-modal information analysis for fault diagnosis with time-series data from power transformer (JEPE 2023)**

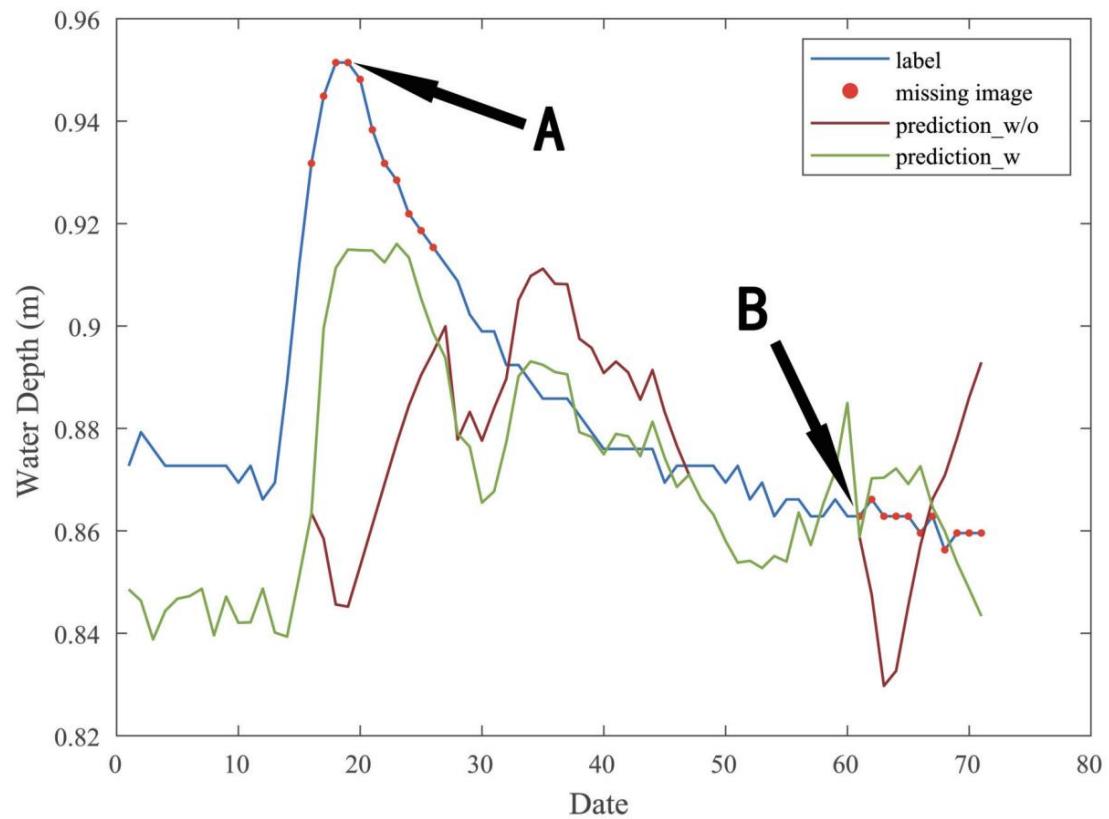
# Applications – Spatial Time Series – Transportation

- Prompting LLMs with structured traffic data for traffic prediction



# Applications – Spatial Time Series – Environment

- **Data Modalities:** Satellite imagery, meteorological time series, domain metadata
- **Challenges:** Missing features, high dimensionality



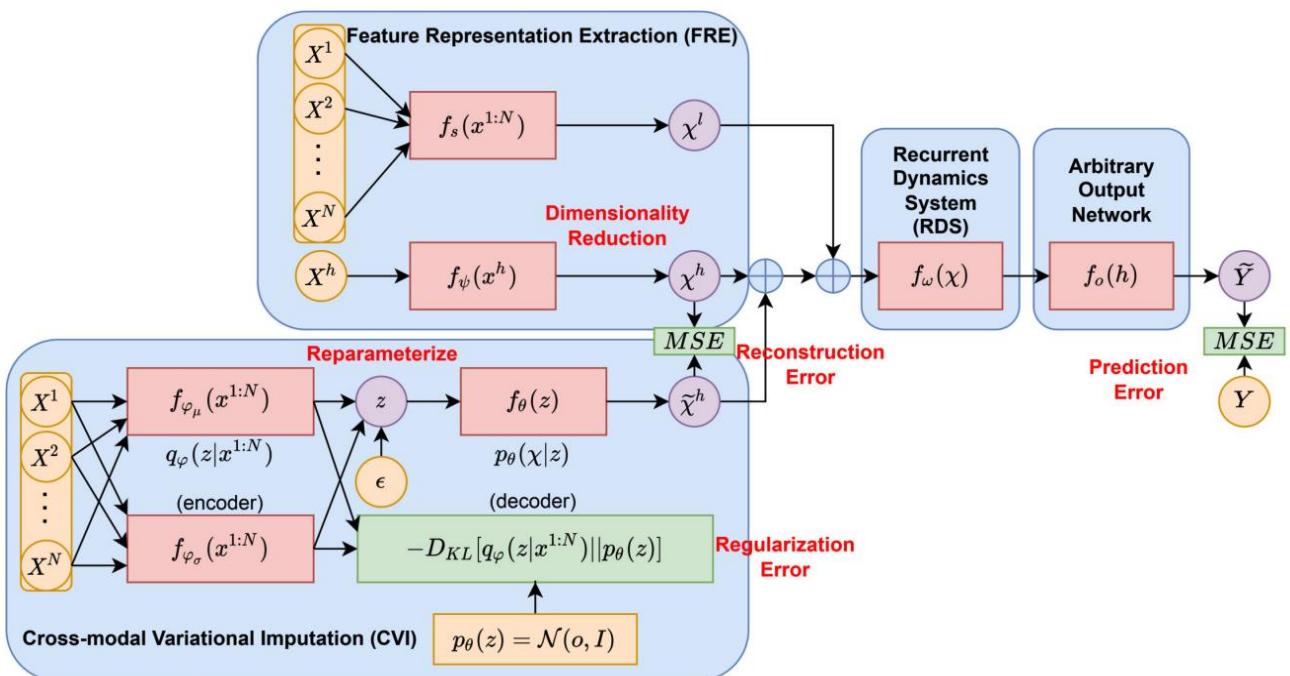
(a) Missing image at peak A.



(b) Missing image at recession B.

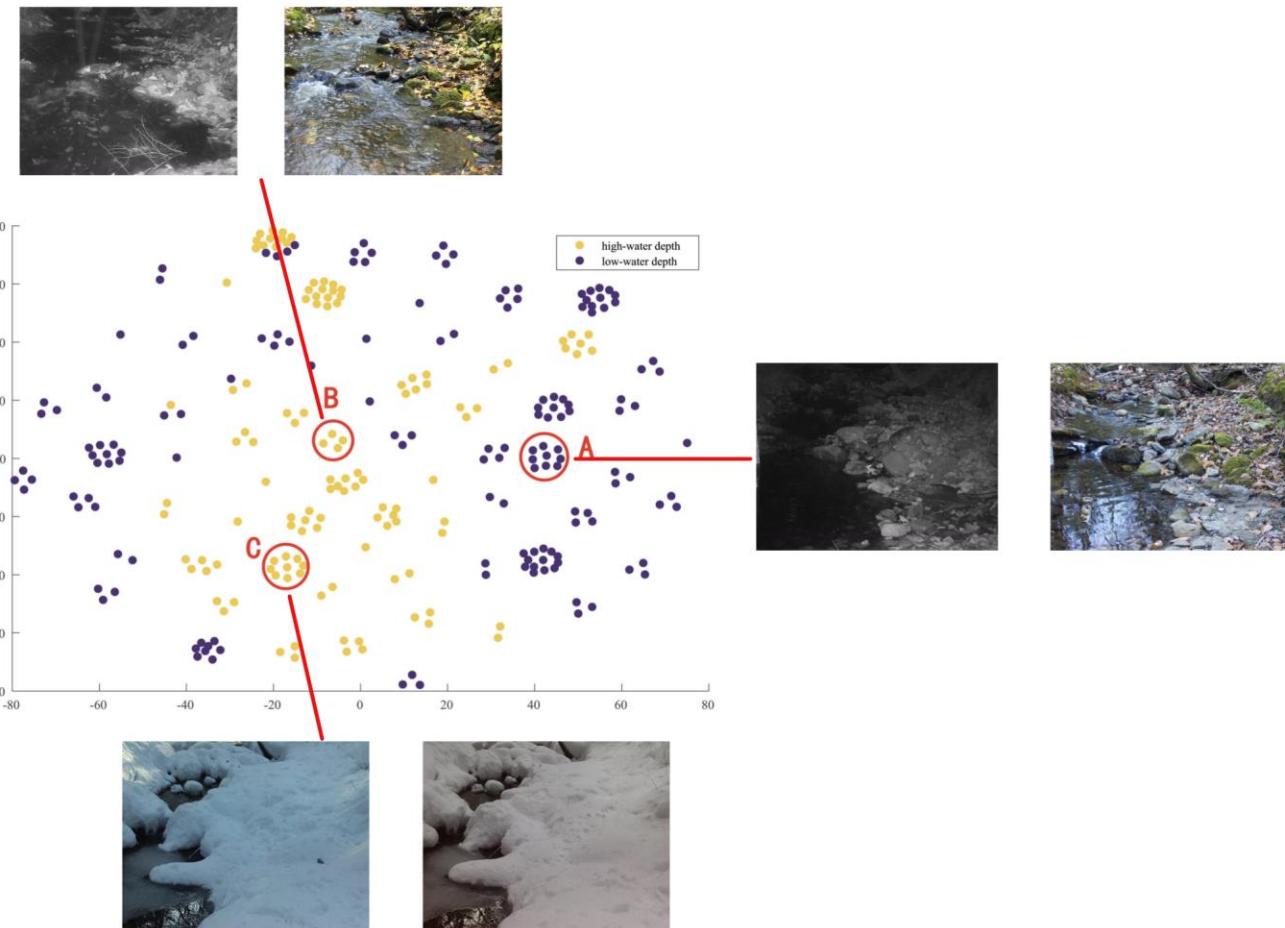
# Applications – Spatial Time Series – Environment

- Cross-modal imputation via variational approximation from low-dim features



# Applications – Spatial Time Series – Environment

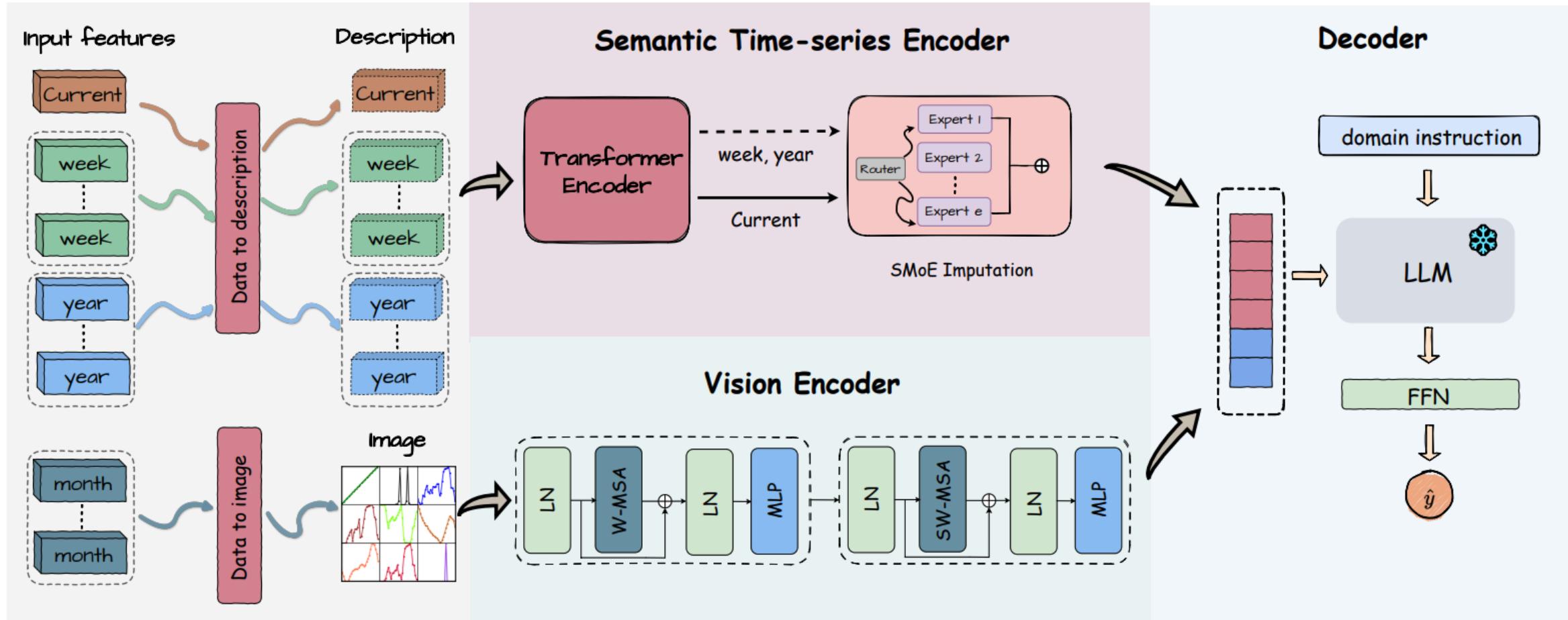
- Cross-modal imputation via variational approximation from low-dim features



VIMTS: Variational-based Imputation for Multi-modal Time Series (IEEE BigData 2022)

# Applications – Spatial Time Series – Environment

- Sparse mixture of experts + instruction-tuned LLM

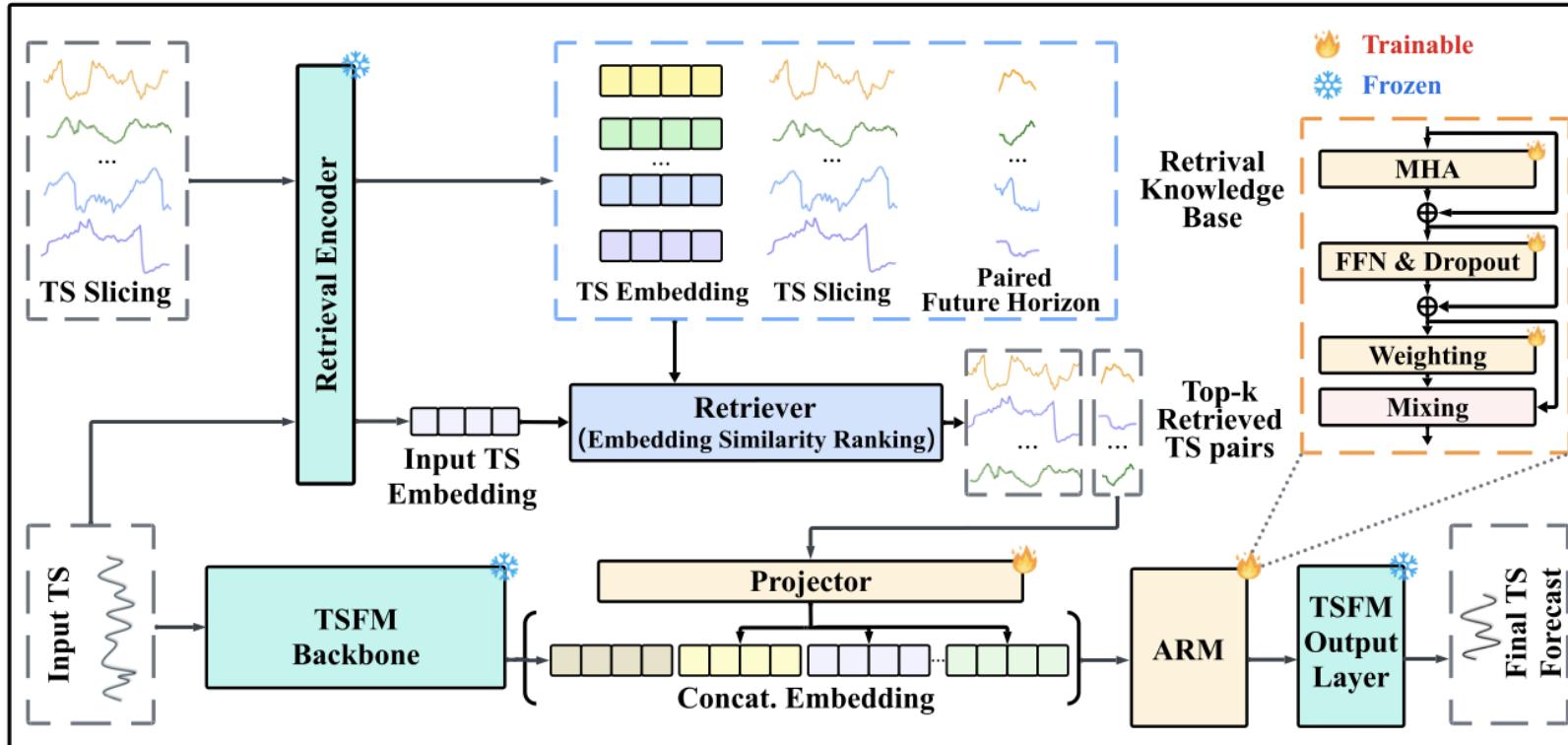


# *Future Research Directions*

# Future Research Directions

- Enhanced reasoning with Multi-modal Time Series:

Combine temporal reasoning with context understanding for interpretable inference.

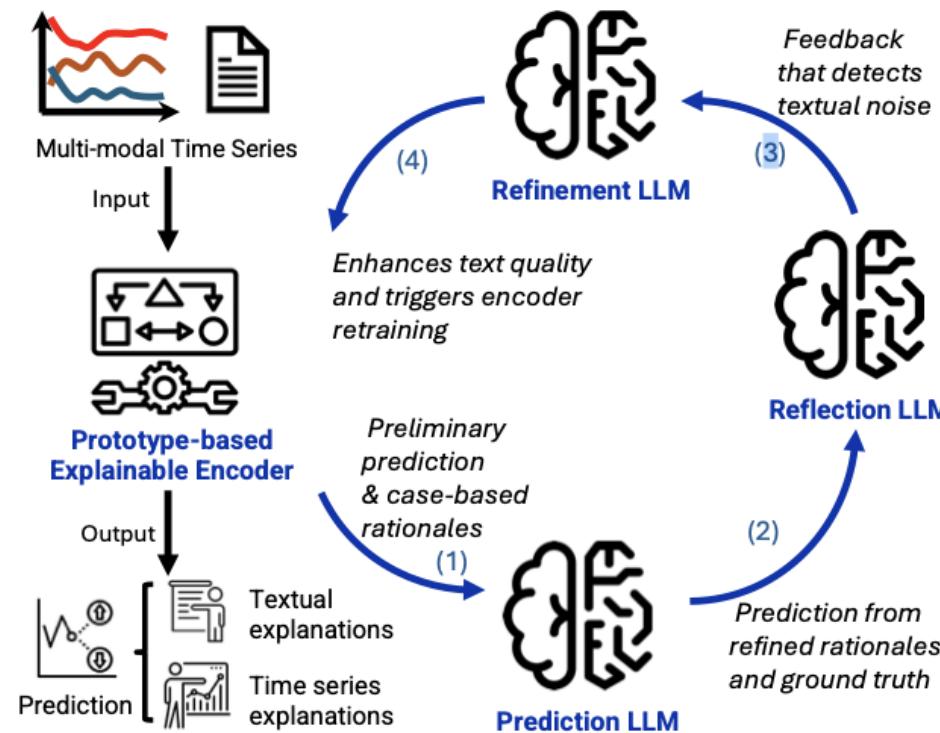


**TS-RAG: Retrieval-Augmented Generation based Time Series Foundation Models are Stronger Zero-Shot Forecaster**

# Future Research Directions

- **Robustness to imperfect Data:**

Handle missing or noisy real-world context effectively.



Explainable Multi-modal Time Series Prediction with LLM-in-the-Loop

# Future Research Directions

- **Decision-making Systems:**

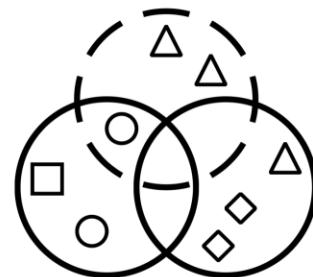
Develop adaptive decision-support systems using multi-modal data to facilitate downstream tasks.

- **Domain Generalization:**

Address the challenges such as domain shifts, modality-specific variations, and temporal dynamics. Improve generalization across unseen domains.

- **Ethics and fairness:**

Address biases to promote equitable outcomes.



...



August 3-7, 2025

# *Thank you!*

# Q & A

Survey Paper



Github



**NEC**  
NEC Laboratories America

Morgan  
Stanley