

Multi-modal Time Series Analysis

— Methods, Datasets, and Applications

Survey Paper



Github



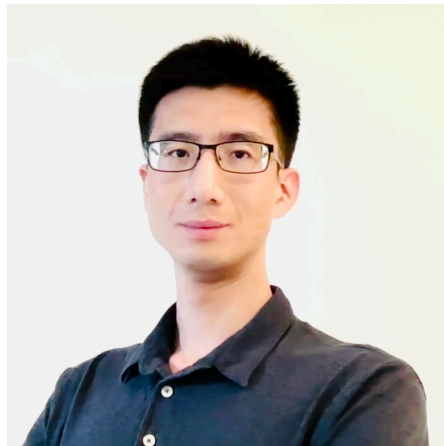
NEC
NEC Laboratories **America**

Morgan
Stanley

Presenters



Dongjin Song
Associate Professor
School of Computing
University of Connecticut



Jingchao Ni
Assistant Professor
Department of Computer Science
University of Houston

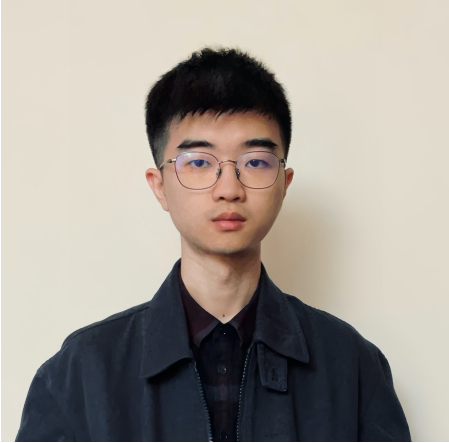


Zijie Pan
Ph.D. Student
School of Computing
University of Connecticut



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



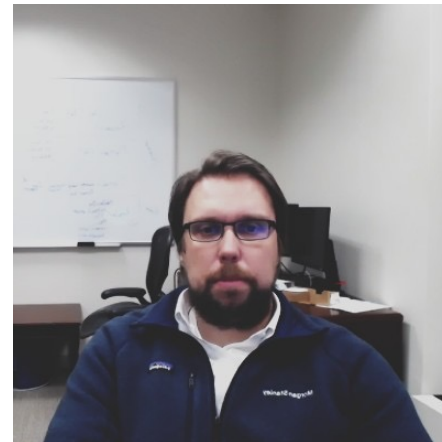
Xuyang Shen
Ph.D. Student
School of Computing
University of Connecticut



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 – Haifeng)
- **Part 2-1: Taxonomy of Multi-modal Time Series Methods** (30 min – Dongjin)
- ---

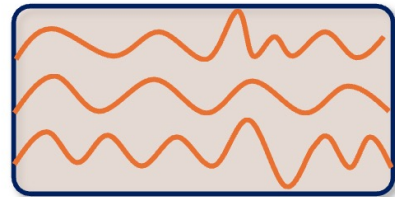
Break (20 min)

- **Part 2-2: Taxonomy of Multi-modal Time Series Methods** (30 min - Jinchao)
- **Part 3: Multi-modal Time Series Applications and Datasets** (40 min - Zijie)
- **Part 4: Future Directions** (10 min - Dongjin)
- **Part 5: Q&A**

Introduction to Multi-modal Time Series Analysis

Background –Time Series Analysis

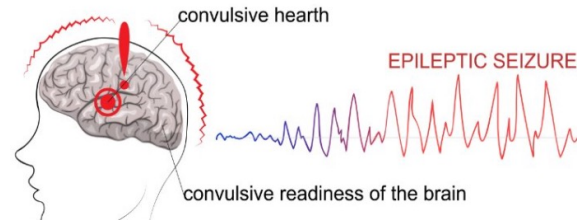
Time Series: Sequential data points indexed by time (e.g., Electricity Load, EEG, Traffic volume).



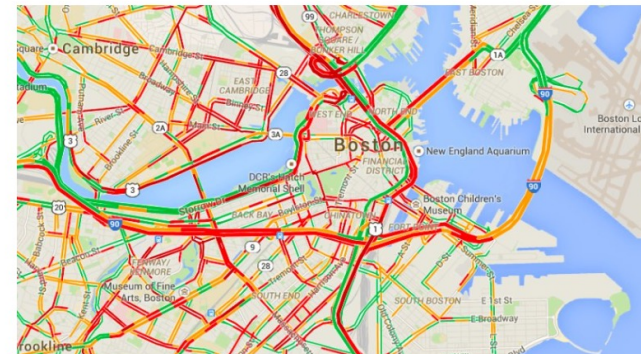
Time series data



Electricity load & Power consumption



Healthcare

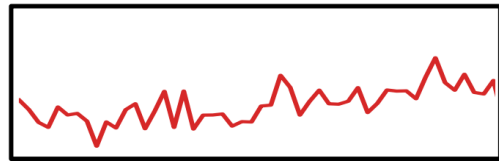


Traffic networks

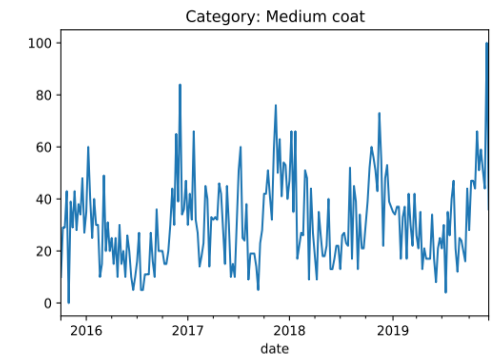
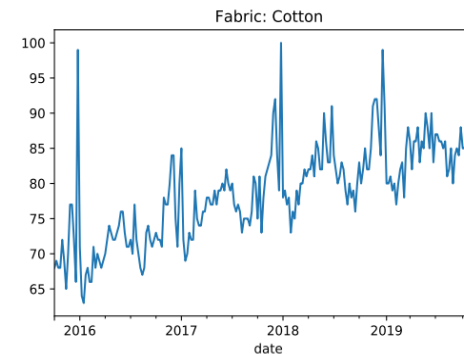
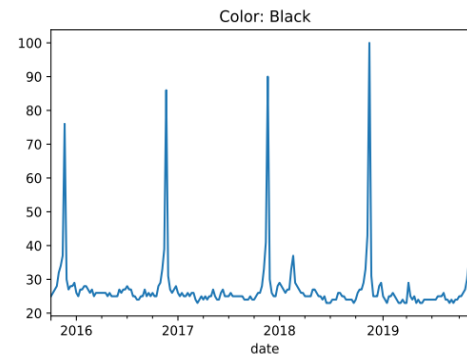
Background – Multi-modal Time Series Analysis

Multi-modal: Involves multiple data sources/modalities (e.g., Image, Text, Audio).

Multi-modal Time Series: Time series that associated with external contexts, which can carry rich semantic information for time series analysis.



Time Series



Major Oil-Producing Nations
Announce Supply Cut, Fuel
Prices Expected to Rise

Text



Background – Multi-modal Time Series Analysis

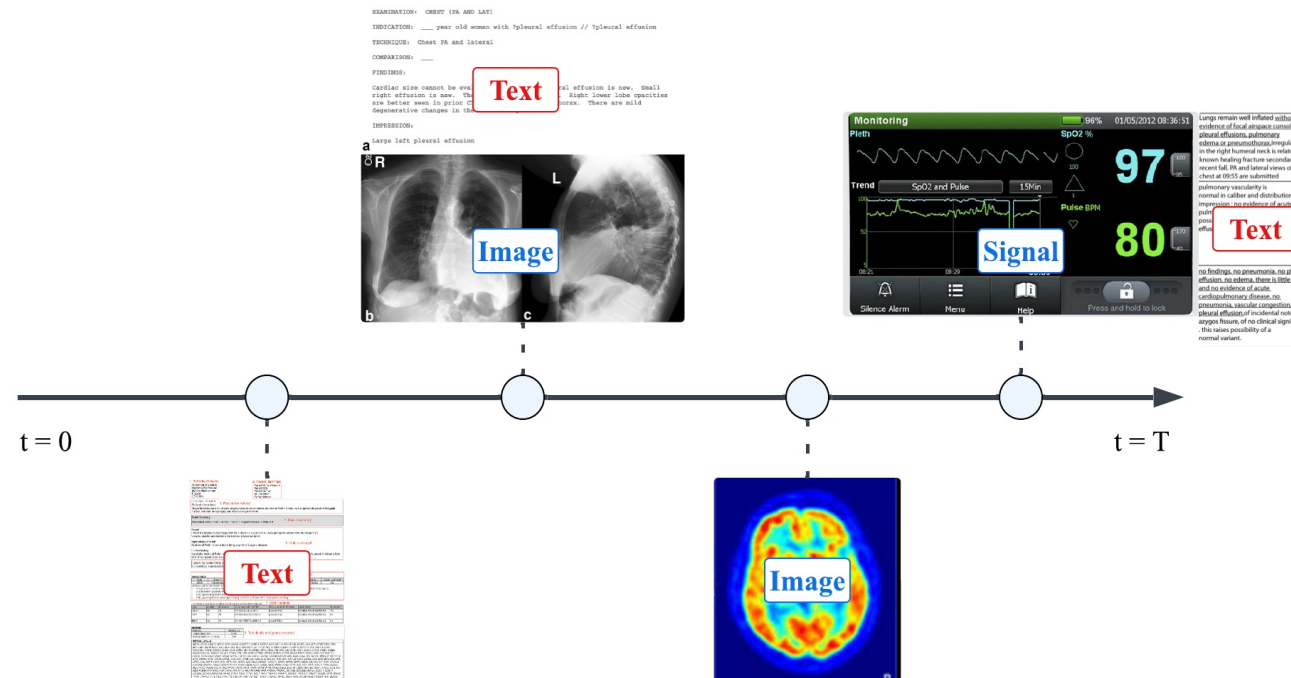
Why is Multi-modality Important?

Real-world systems are **heterogeneous**.

Combining multimodal signals leads to **richer understanding** and **better predictions**.

Examples:

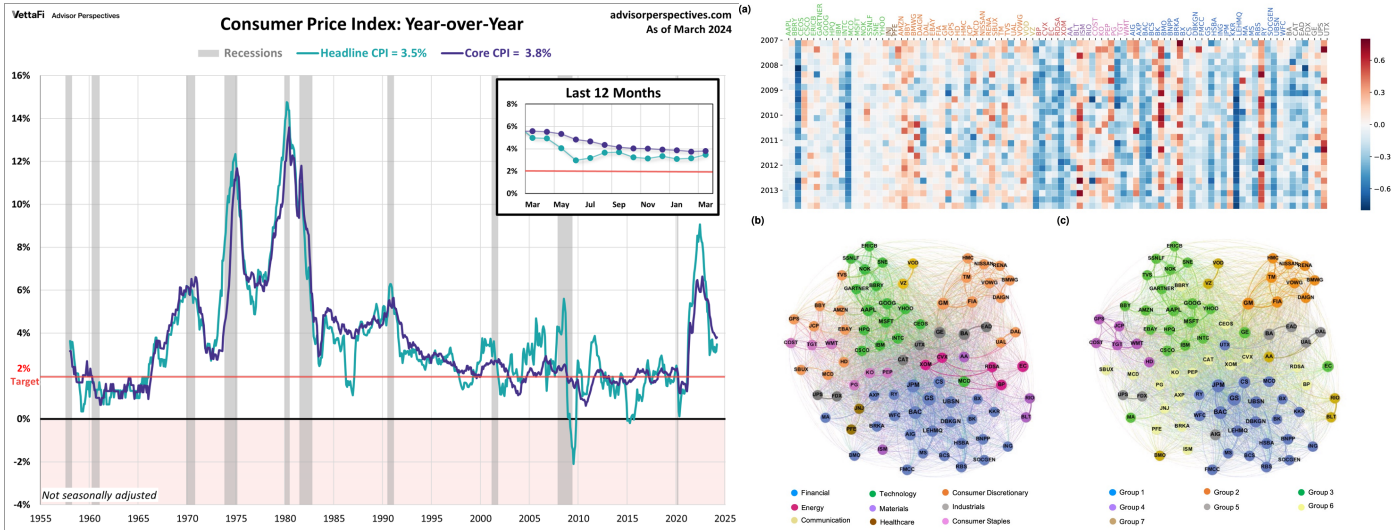
Electronic
Health
Records
(EHR)



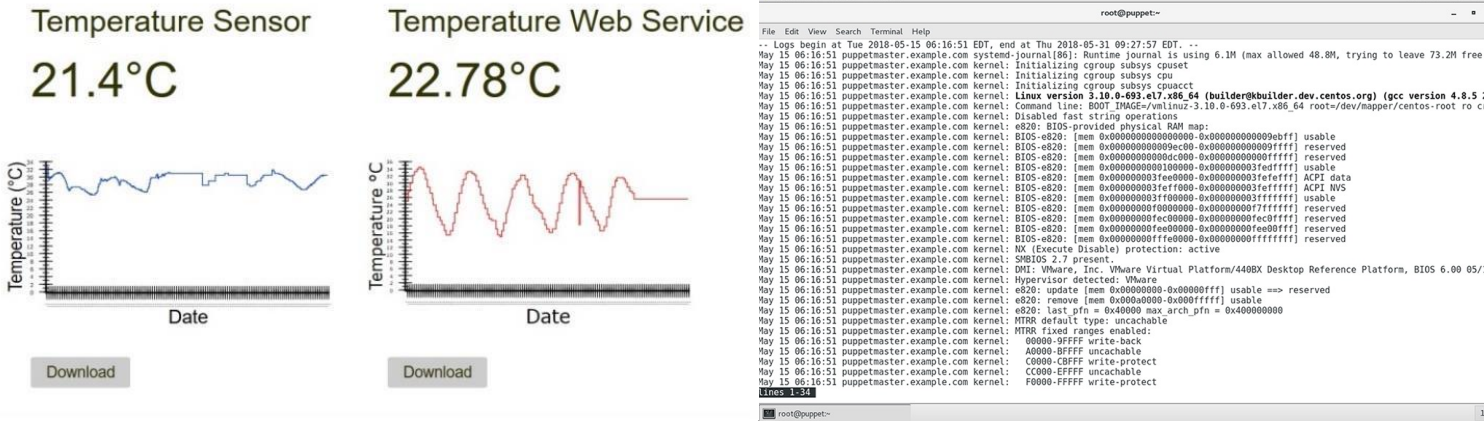
Background – Multi-modal Time Series Analysis

More Examples:

Finance: Price + News Sentiment



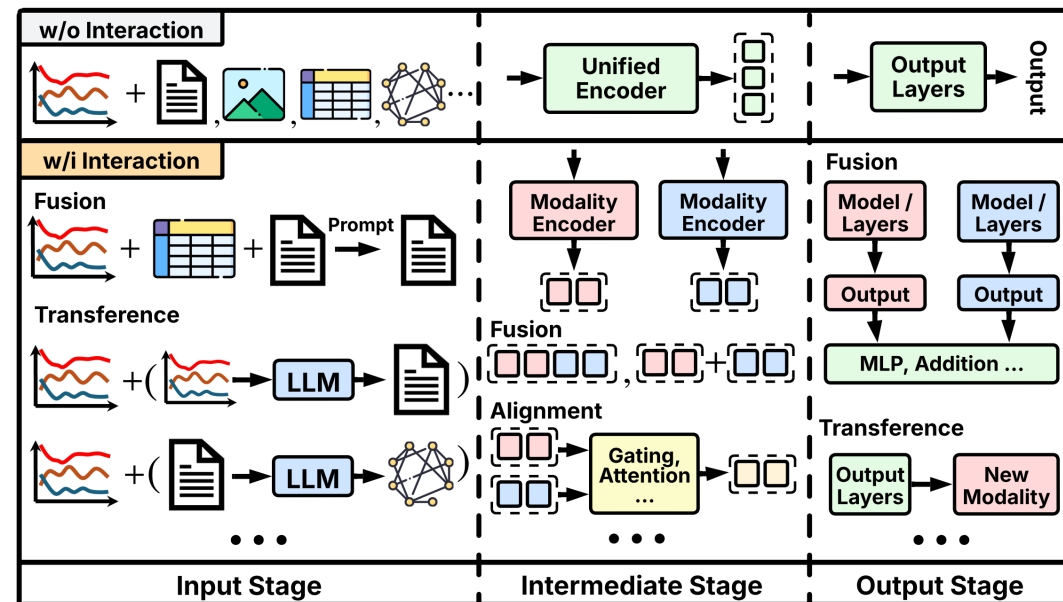
IoT Systems: Temperature + Logs



Background – Multi-modal Time Series Analysis

- **Problem Statement**

- Effective analysis of multi-modal time series is hindered by data heterogeneity, modality gap, misalignment, and inherent noise.
- We summarize the general pipeline and categorize existing methods through a unified cross-modal interaction framework encompassing fusion, alignment, and transference at different stages.

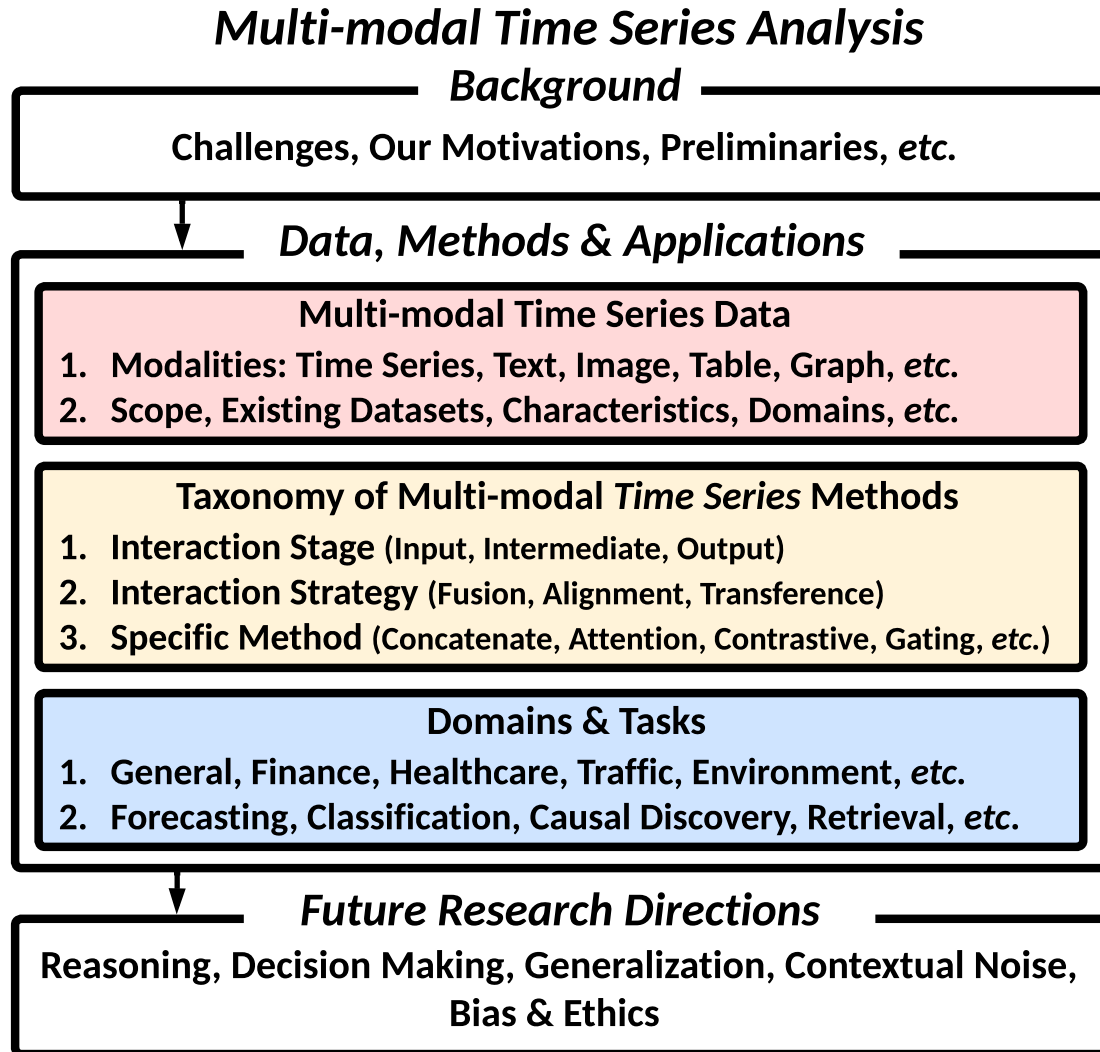


Background – Multi-modal Time Series Analysis

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 Part 1, the focus is to leverage **multi-modal inputs** from multiple sources in real-world contexts.
 - For Part 2, the focus is more on **transforming** the input modality to another output modality and leveraging **multimodal views** of time series.
3. We discuss the existing applications and available datasets for multi-modal time series analysis.

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.

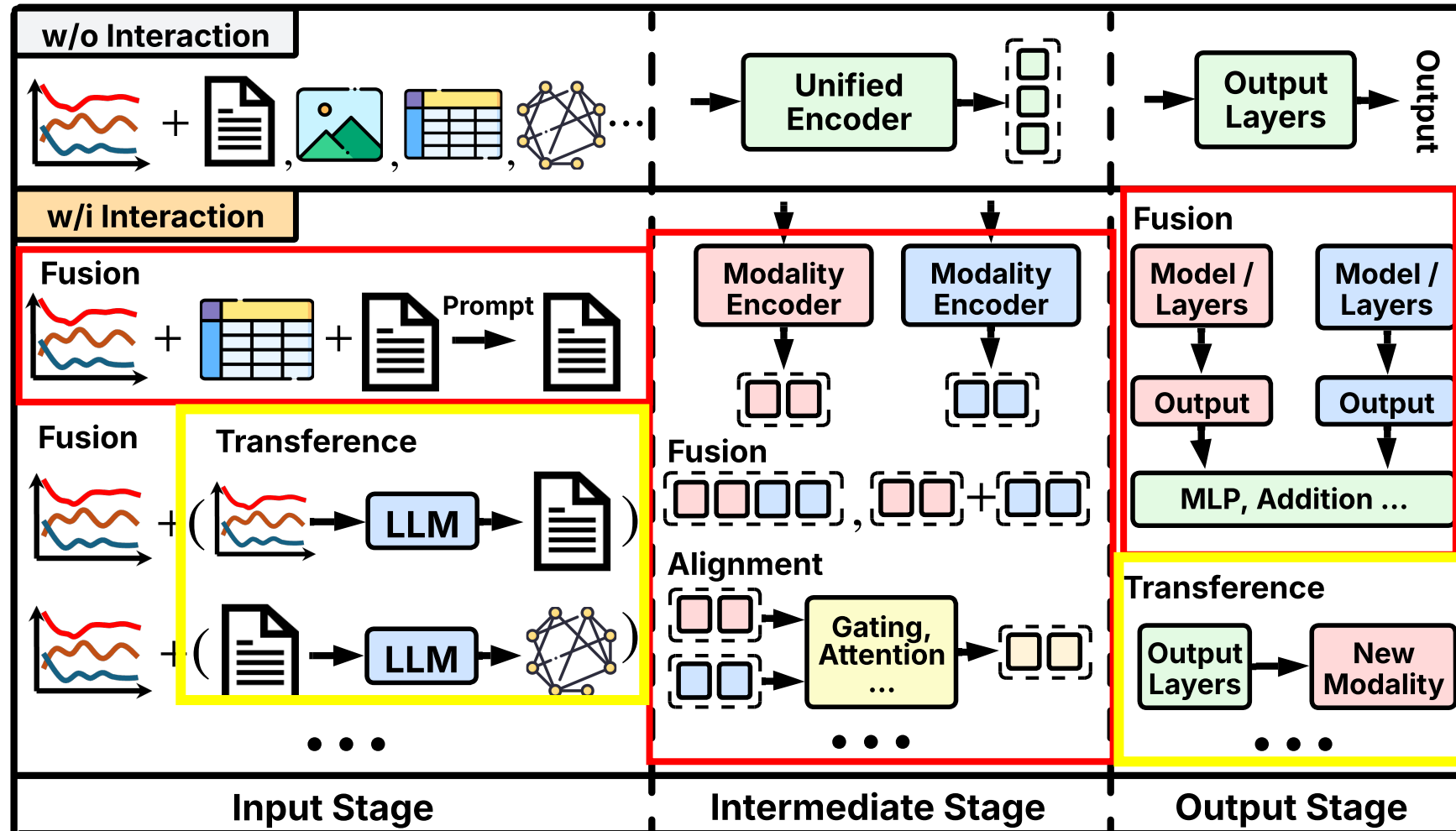
Multi-modal Time Series Methods

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 (Multimodal views of TS)**
- 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



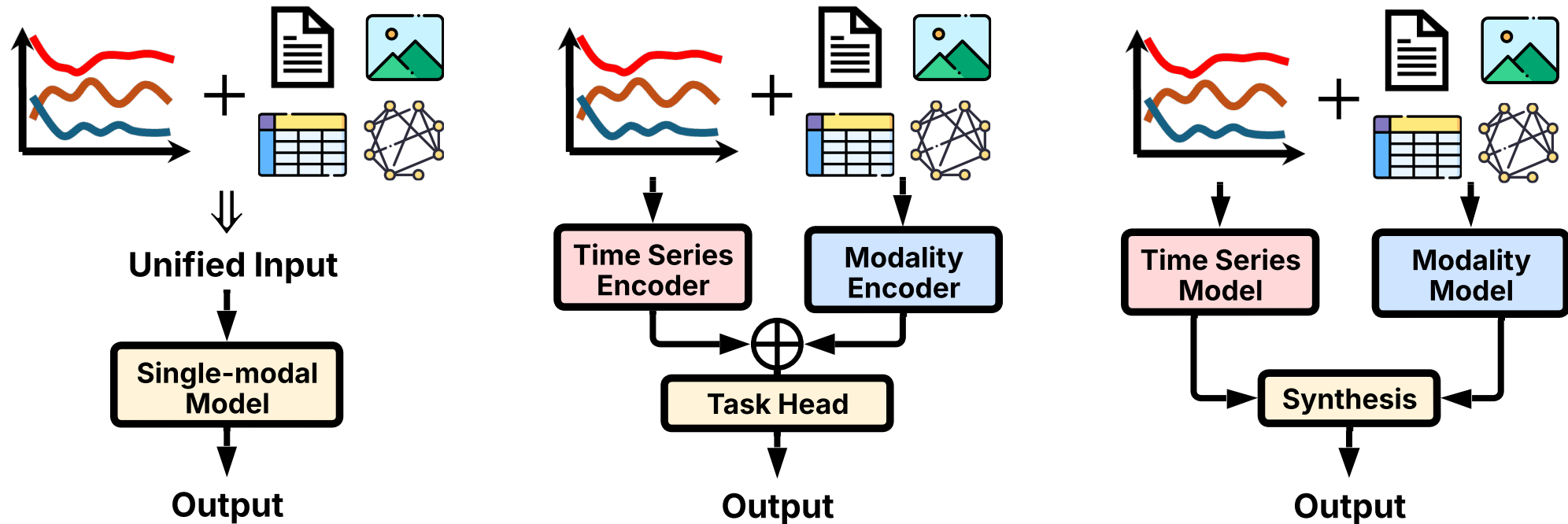
Overview and representative examples of cross-modal interactions

Multi-modal Time Series Methods

Part 1: Fusion and Alignment

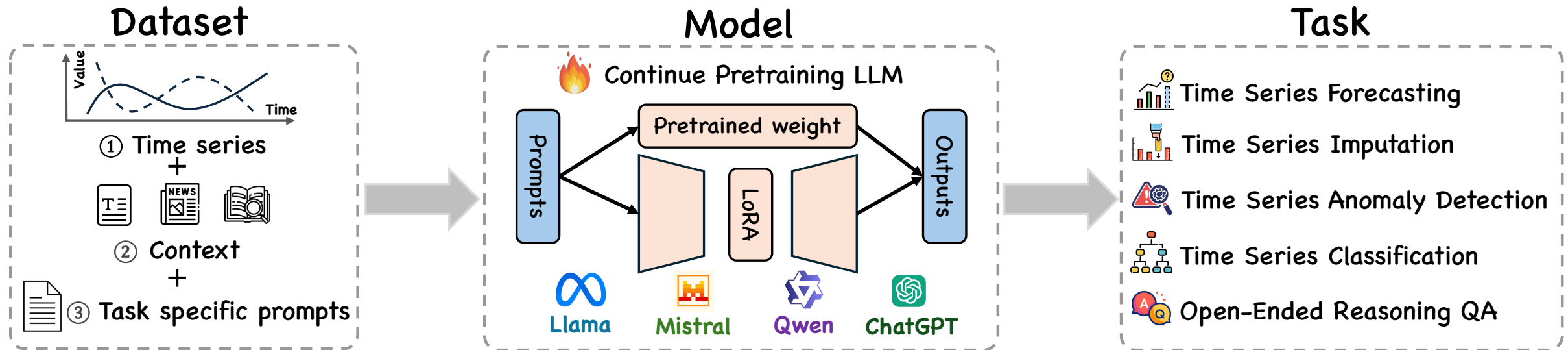
Cross-modal Interaction with Time Series: Fusion

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



Multi-modal Fusion with Time Series – Input level

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

(1) Forecasting

[Context] This dataset aims to estimate heart rate during physical exercise using wrist-worn PPG sensors and sampled at 125 Hz from subjects aged 18 to 35 ...

The input Time Series are:



Predict the next 24 time series point given information above.

Future Time Series

Why?

(2) Imputation

[Context] The Self-regulation of Slow Cortical Potentials dataset, provided by the University of Tuebingen, involves EEG recordings from ...

Please give full time series with missing value imputed.



Why?

(3) Anomaly Detection

[Context] The following data is derived from traffic systems, recording variations in traffic flow, such as ...

Please determine whether there are anomalies in this time series given information above.

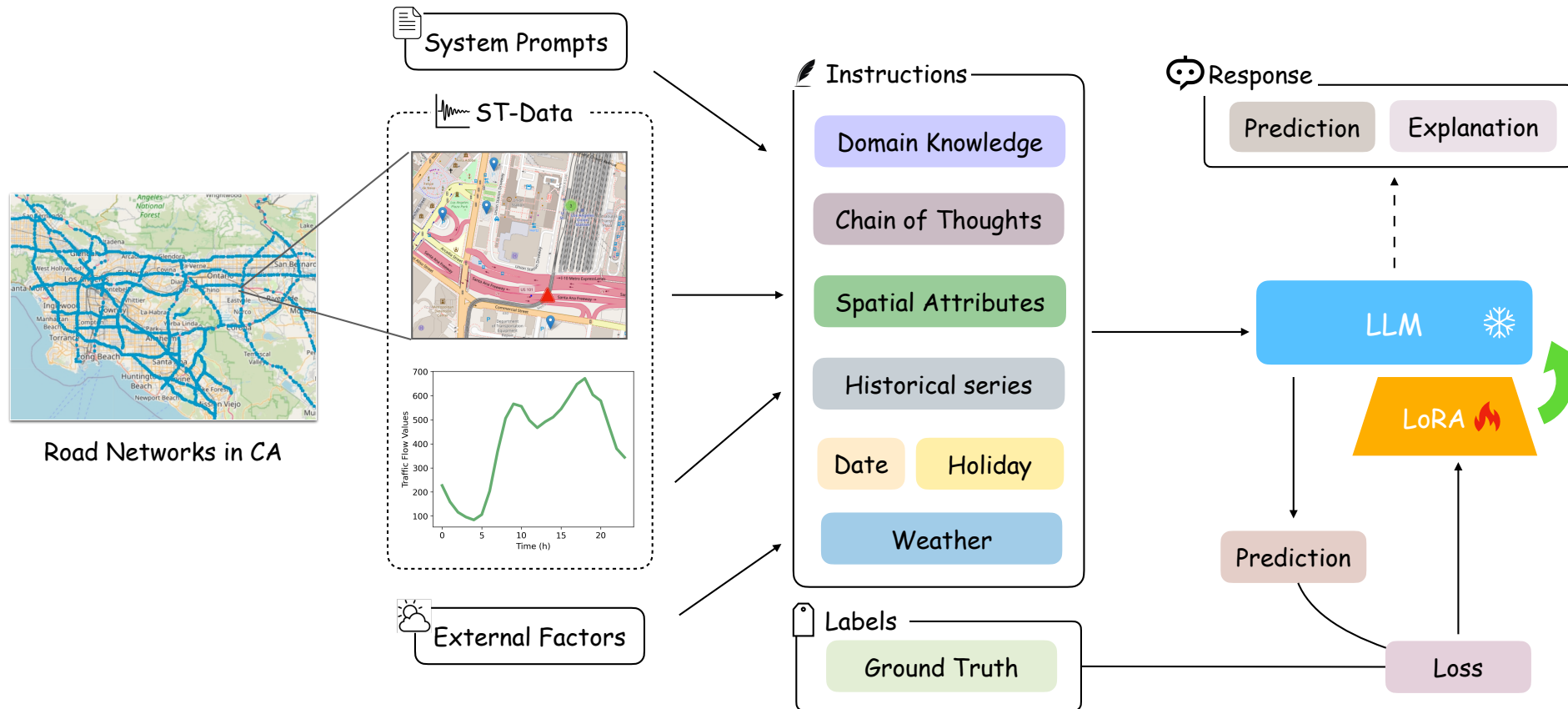


Why?

[Industrial Maintenance]

Multi-modal Fusion with Time Series – Input level

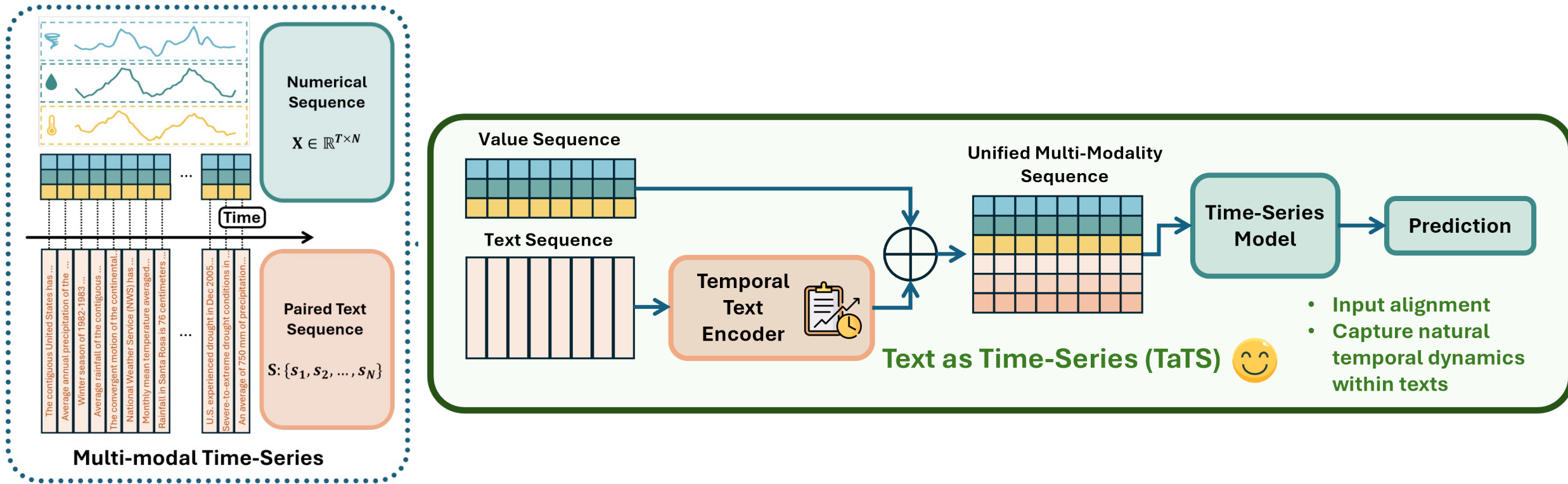
Integrate time series, tabular data and texts into a unified textual prompt



Guo et al. "Towards explainable traffic flow prediction with large language models", Communications in Transportation Research 2024

Multi-modal Fusion with Time Series – Input level

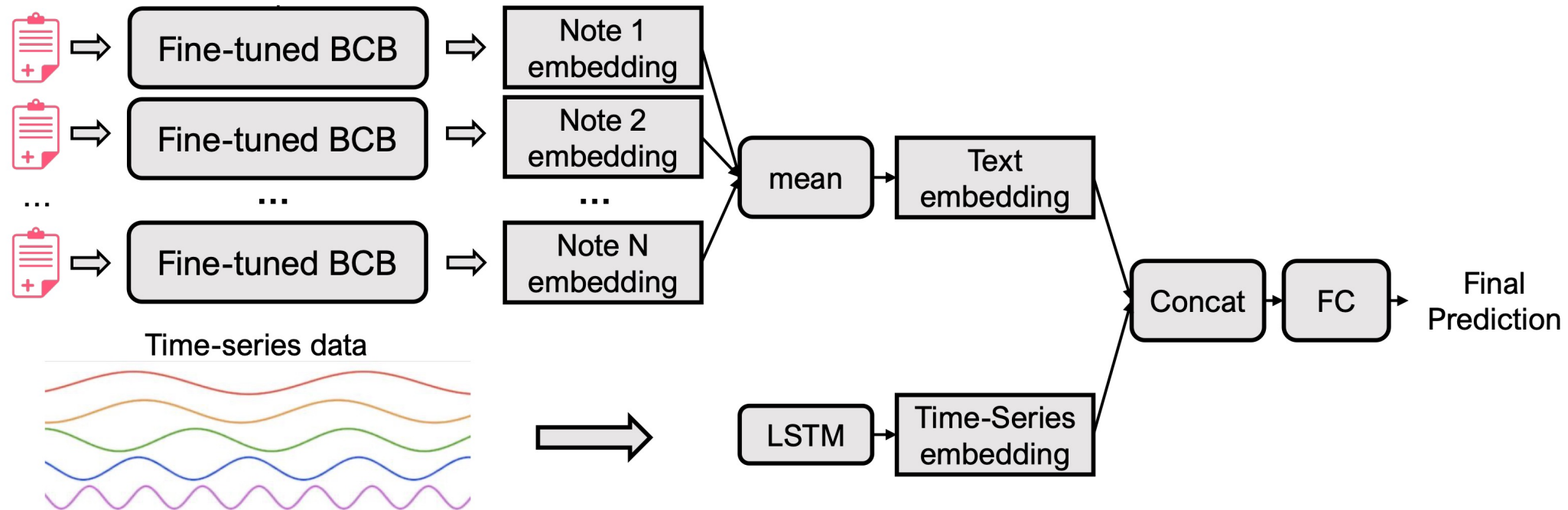
Integrate paired text embedding as an additional variable of time series



Li et al. "Language in the Flow of Time: Time-Series-Paired Texts Weaved into a Unified Temporal Narrative", CoRR 2025

Multi-modal Fusion with Time Series – Intermediate level

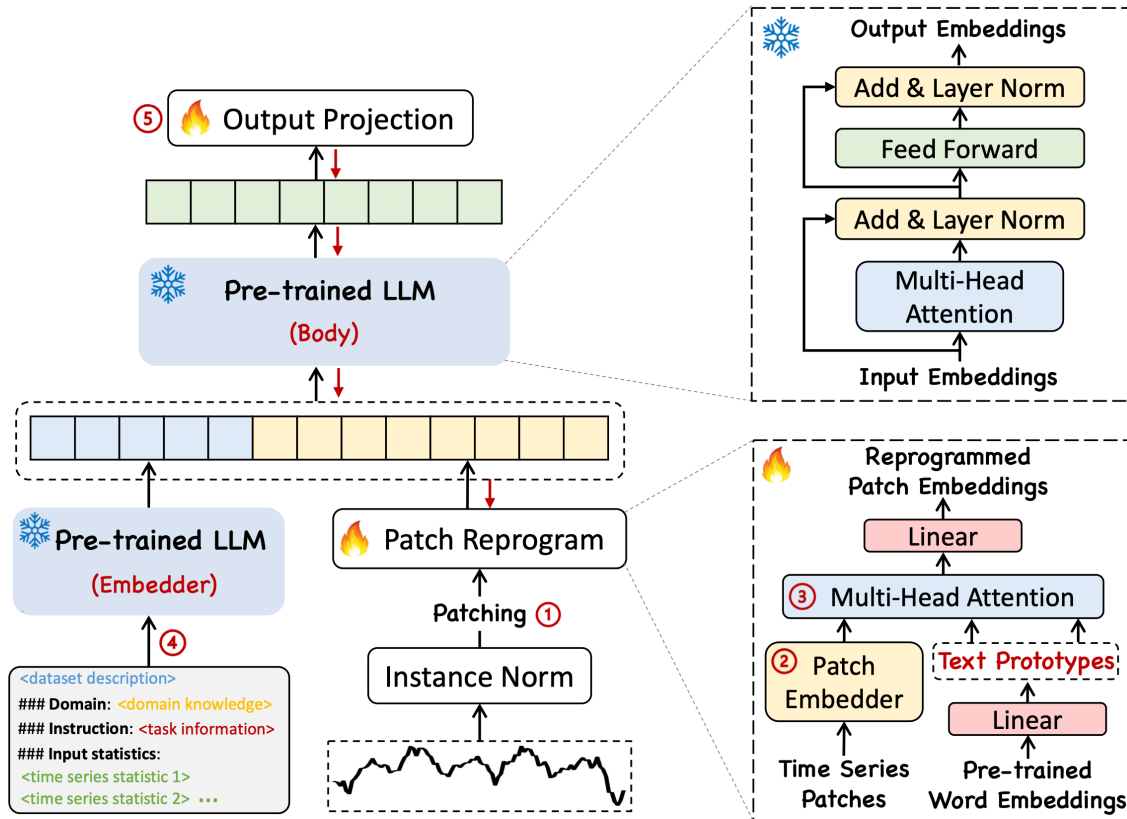
Simple aggregations (e.g., mean, addition, concatenation, etc.) of time series embedding and other modality embeddings



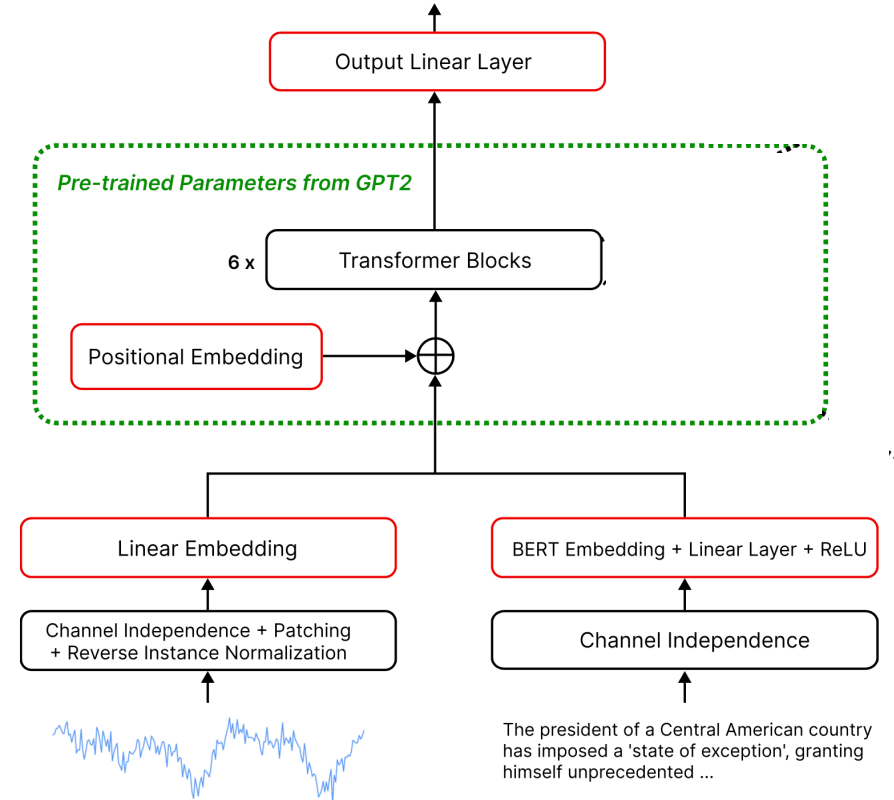
Deznabi et al. "Predicting In-hospital Mortality by Combining Clinical Notes with Time-series Data", ACL 2021.

Multi-modal Fusion with Time Series – Intermediate level

The fusion of modality embeddings is usually followed by alignments



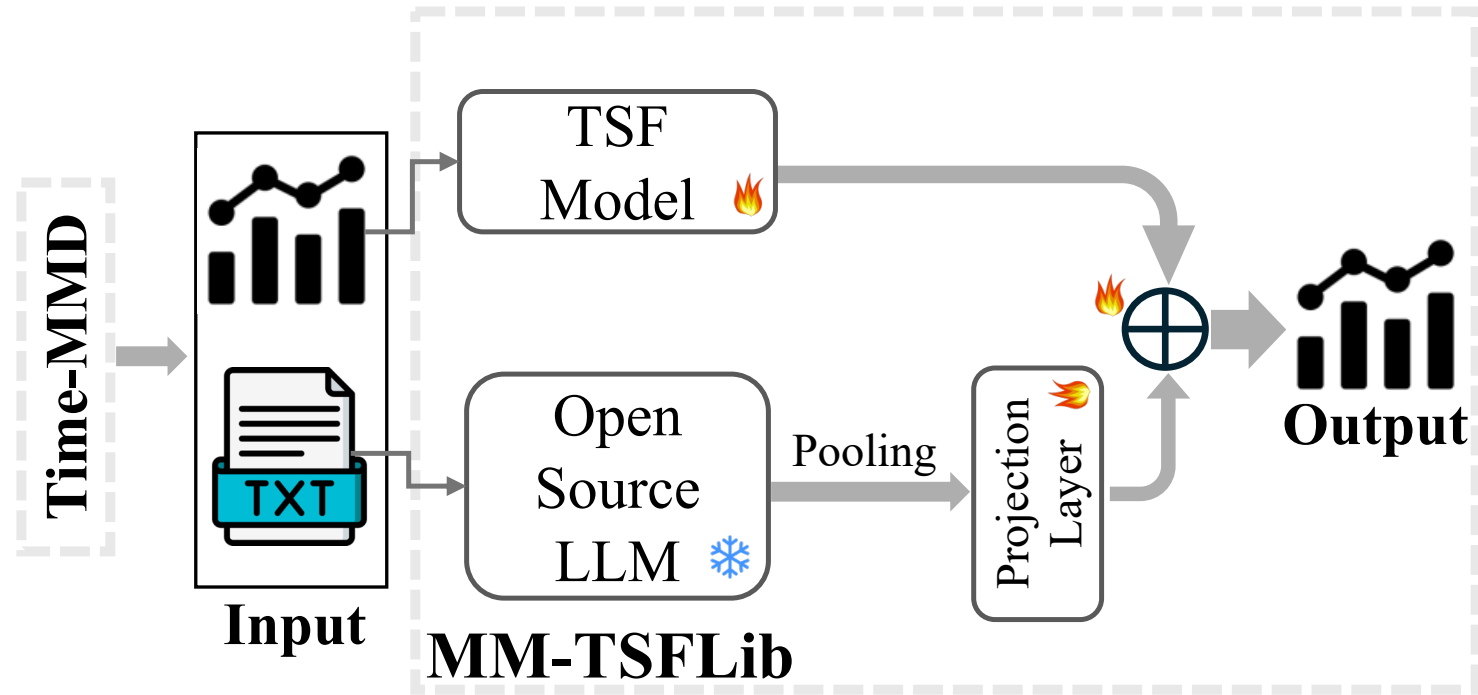
Jin et al. "Time-LLM: Time Series Forecasting by Reprogramming Large Language Models", ICLR 2024.



Jia et al. "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

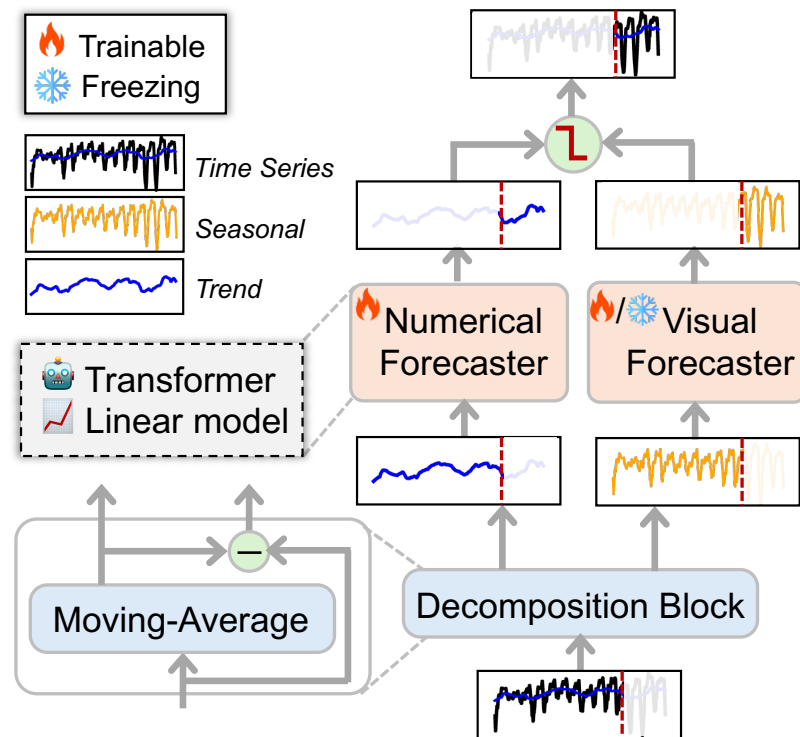


Liu et al. "Time-MMD: Multi-Domain Multimodal Dataset for Time Series Analysis", NeurIPS 2024.

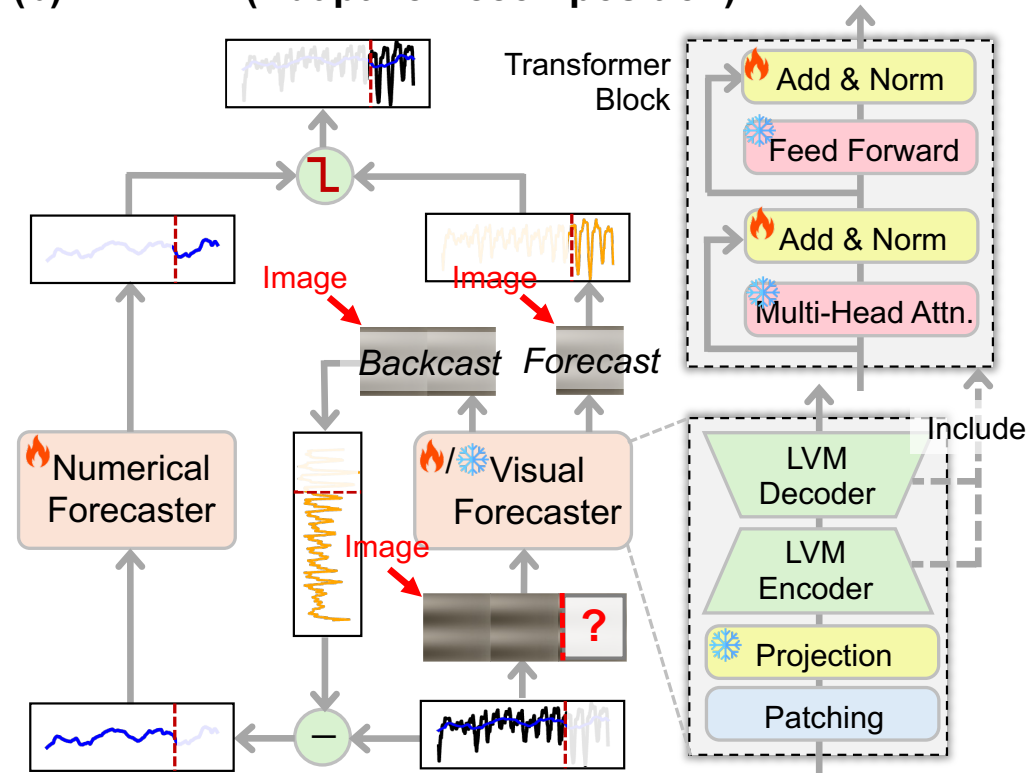
Multi-modal Fusion with Time Series – Output level

Assemble modality outputs as decomposed components of the final output

(a) DMMV-S (Simple Decomposition)



(b) DMMV-A (Adaptive Decomposition)



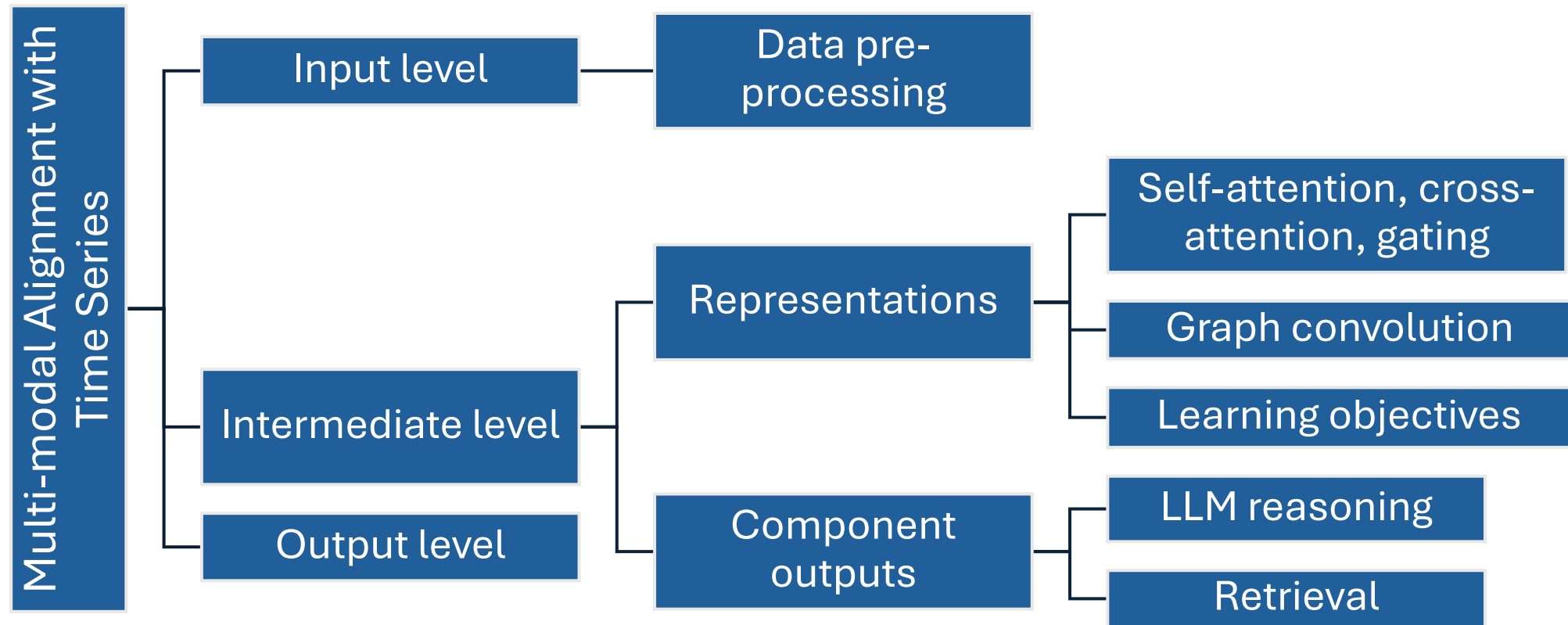
$$\hat{y}^i = g \circ \hat{y}_{\text{season}}^i + (1 - g) \circ \hat{y}_{\text{trend}}^i, \quad \text{where } \hat{y}_{\text{season}}^i = f_{\text{vis}}(\tilde{\mathbf{I}}^i), \quad \hat{y}_{\text{trend}}^i = f_{\text{num}}(\Delta \mathbf{x}^i)$$

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.
- 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.

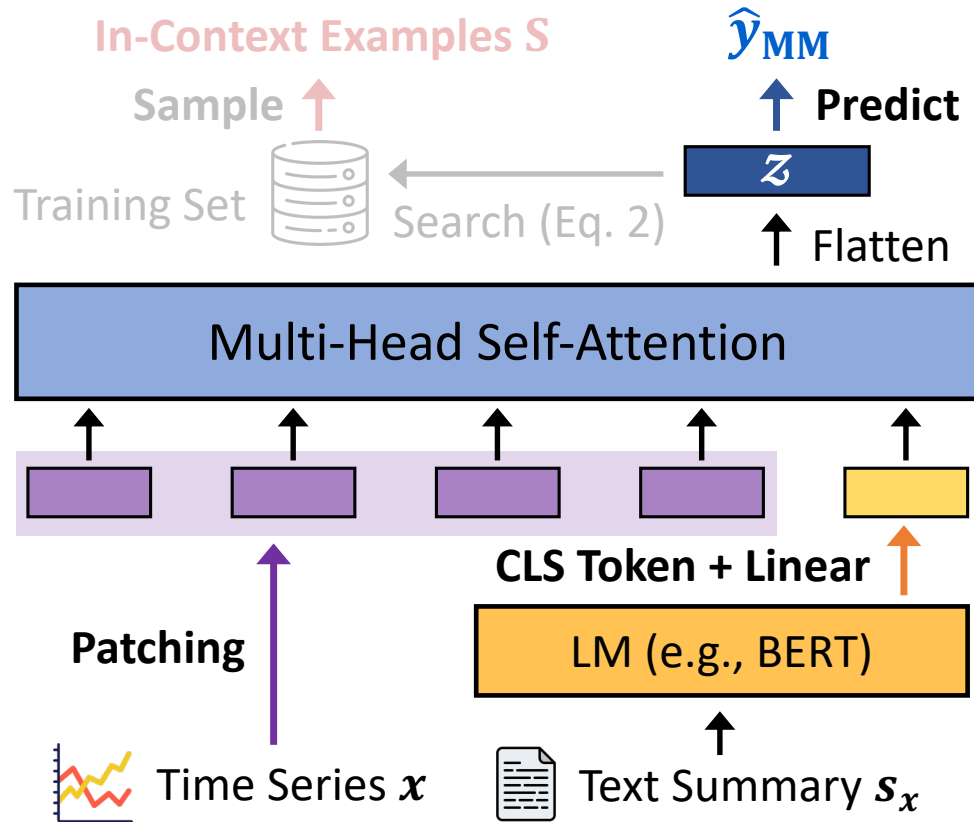
Given multi-modal embeddings $\mathbf{E}_{\text{mm}} \in \mathbb{R}^{n \times d}$, where n is the number of modality tokens and d is the embedding dimension:

$$\text{Attention}(\mathbf{E}_{\text{mm}}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}$$

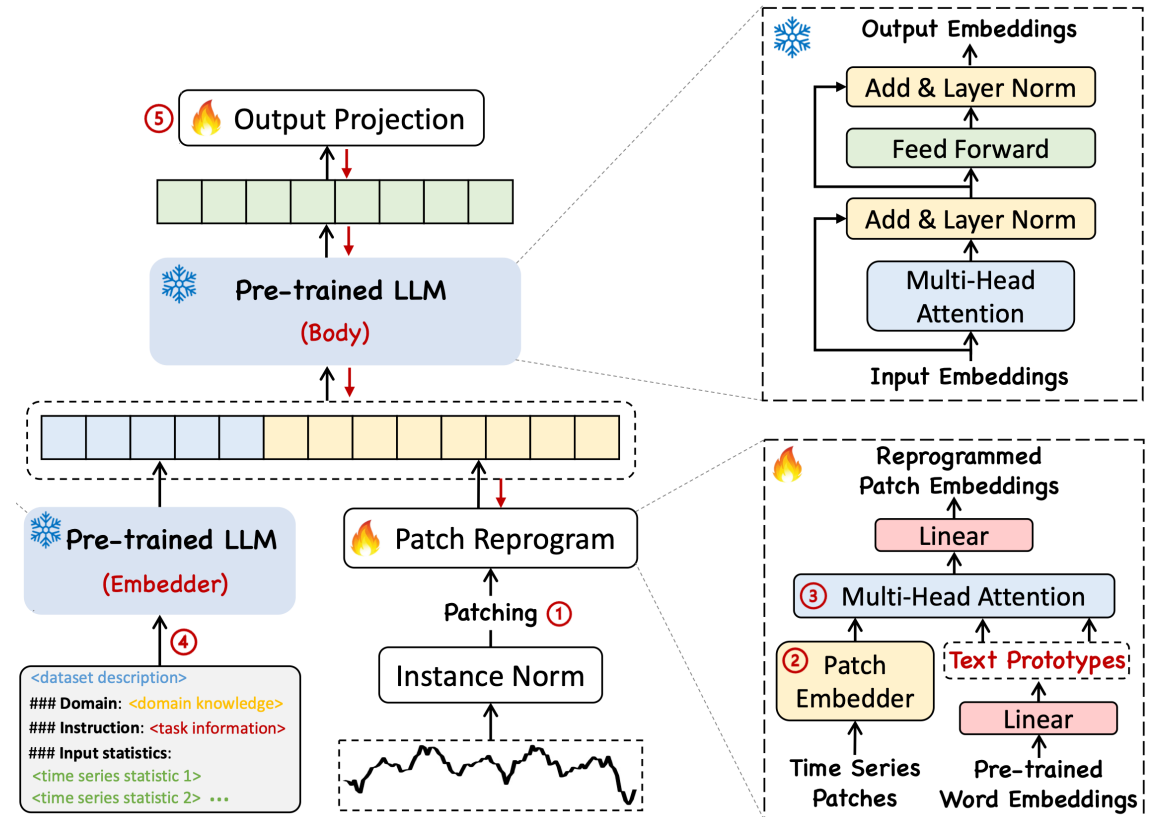
where the queries \mathbf{Q} , keys \mathbf{K} , and values \mathbf{V} are linear projections of \mathbf{E}_{mm} :

$\mathbf{Q} = \mathbf{E}_{\text{mm}}\mathbf{W}_Q$, $\mathbf{K} = \mathbf{E}_{\text{mm}}\mathbf{W}_K$, $\mathbf{V} = \mathbf{E}_{\text{mm}}\mathbf{W}_V$ with learnable weights $\mathbf{W}_{Q,K,V} \in \mathbb{R}^{d \times d_k}$

Multi-modal Alignment with Time Series - Representations

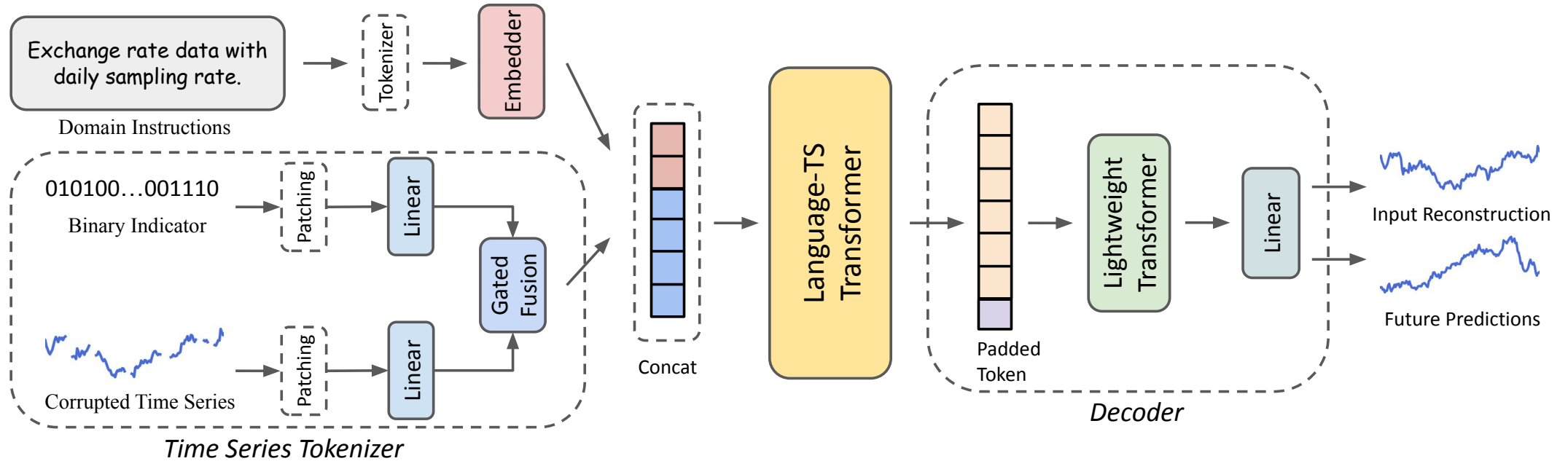


Lee et al, "TimeCAP: Learning to Contextualize, Augment, and Predict Time Series Events with Large Language Model Agents", AAAI 2025



Jin et al. "Time-LLM: Time Series Forecasting by Reprogramming Large Language Models", ICLR 2024

Multi-modal Alignment with Time Series - Representations



Liu et al. "UniTime: A Language-Empowered Unified Model for Cross-Domain Time Series Forecasting", WWW 2024

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.

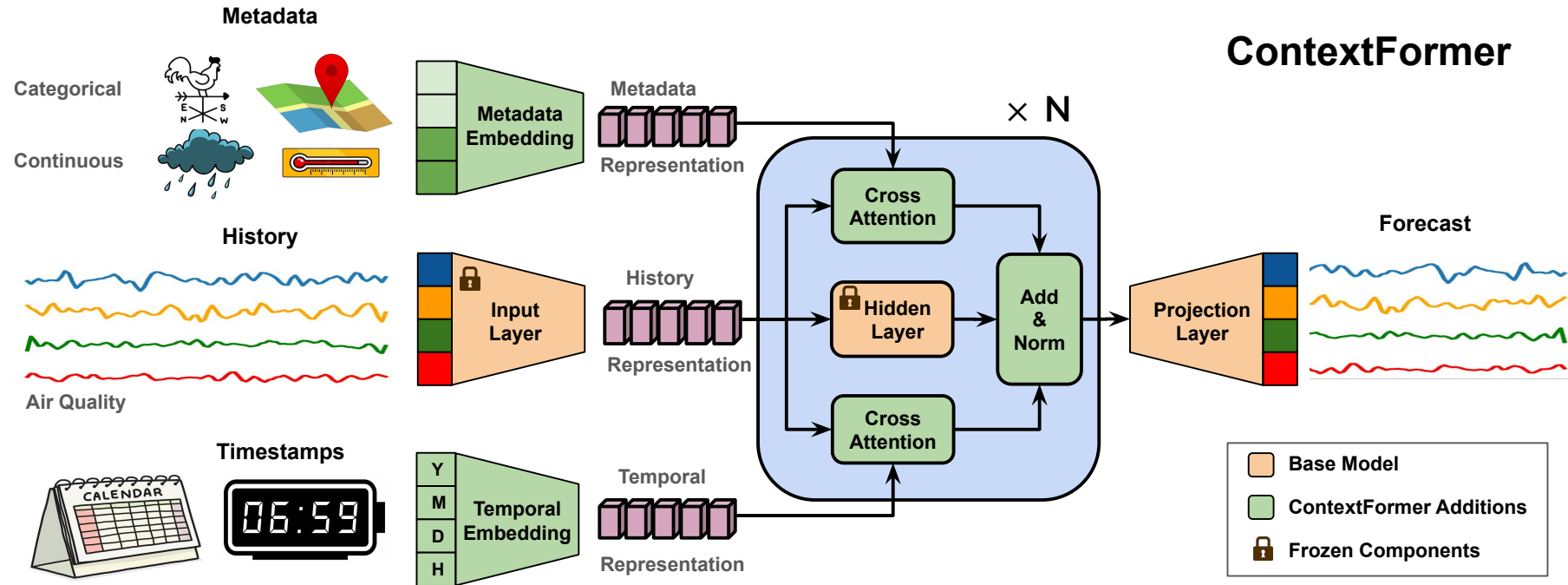
Given multi-modal embeddings $E_{ts} \in \mathbb{R}^{n \times d}$, where n is the number of modality tokens and d is the embedding dimension:

$$\text{CrossAttention}(E_{ts}, E_c) = \text{softmax}\left(\frac{Q_{ts}K_c^\top}{\sqrt{d_k}}\right)V_c$$

where the queries Q_{ts} , keys K_c , and values V_c are linear projections of E_{ts} :

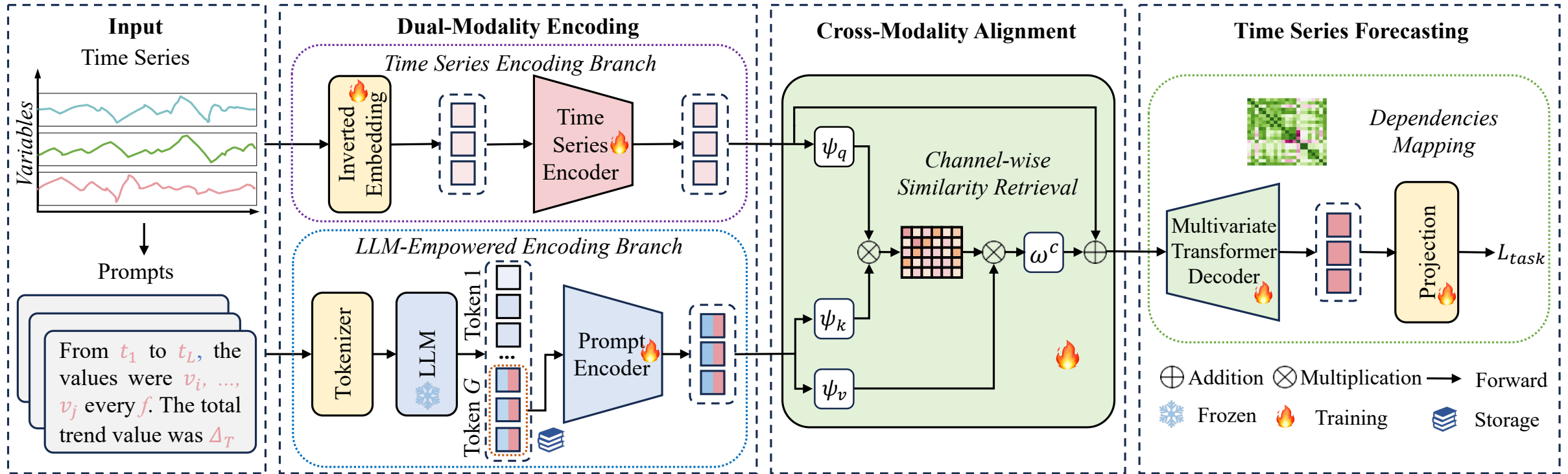
$Q_{ts} = E_{ts}W_Q$, $K_c = E_{ts}W_K$, $V_c = E_cW_V$ with learnable weights $W_{Q,K,V} \in \mathbb{R}^{d \times d_k}$

Multi-modal Alignment with Time Series - Representations



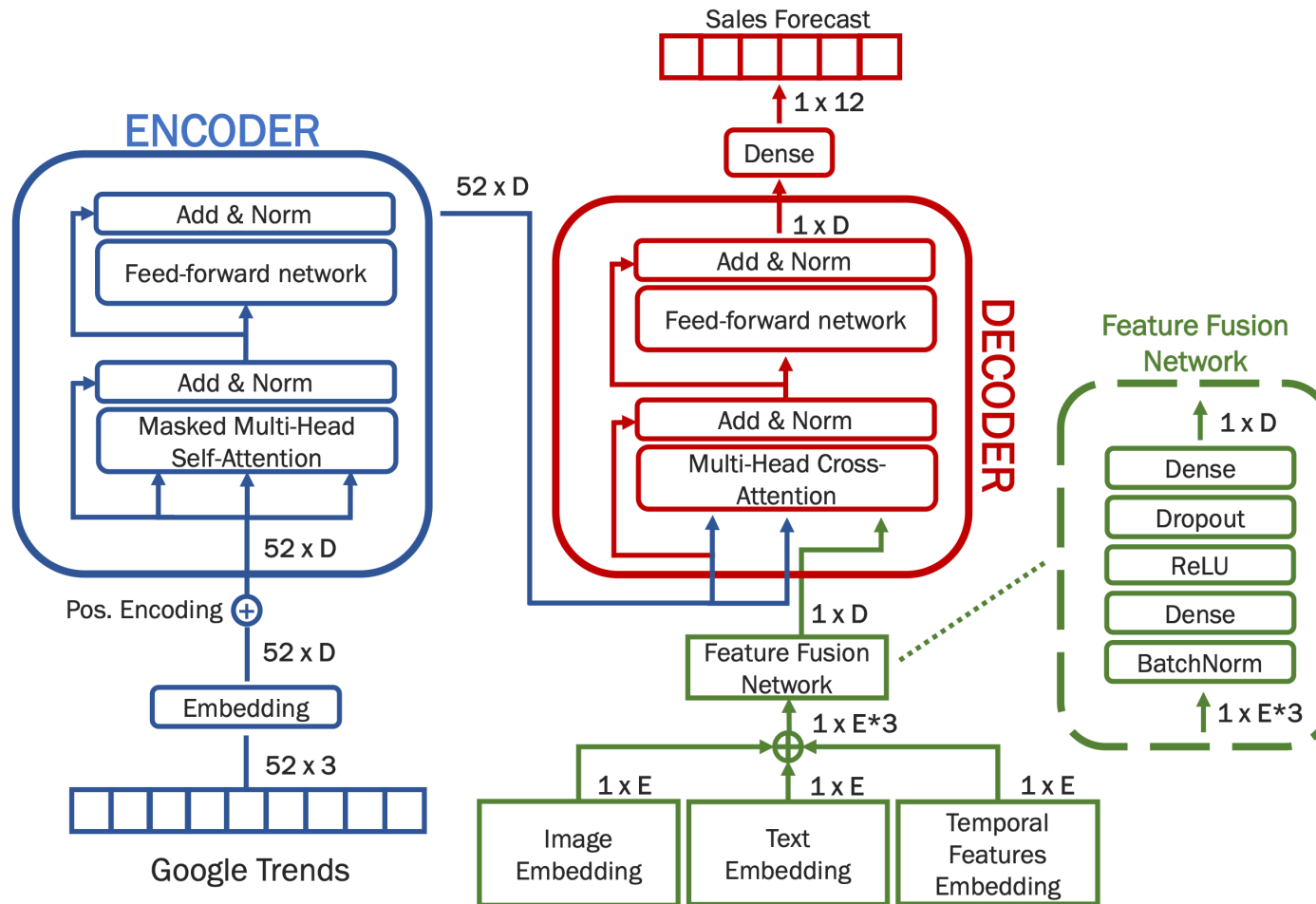
Chattopadhyay et al. "Context Matters: Leveraging Contextual Features for Time Series Forecasting" 2025

Multi-modal Alignment with Time Series - Representations



Liu et al. "TimeCMA: Towards LLM-Empowered Multivariate Time Series Forecasting via Cross-Modality Alignment", AAAI 2025

Multi-modal Alignment with Time Series - Representations



Skenderi et al. "Multimodal Forecasting of New Fashion Product Sales with Image-based Google Trends", Journal of Forecasting 2021

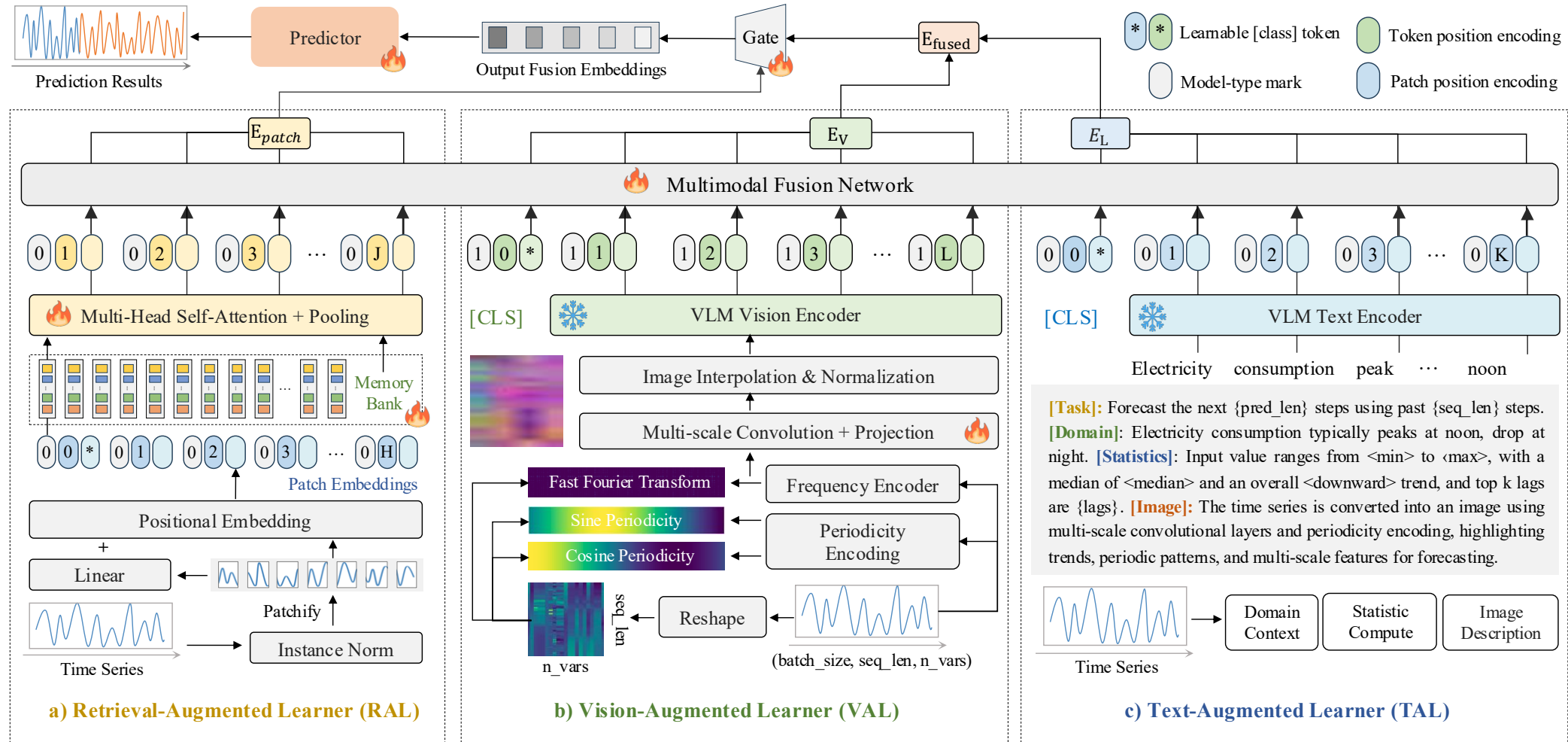
Multi-modal Alignment with Time Series - Representations

Gating: a parametric filtering operation that explicitly regulates the influence of time series and other modalities on the fused embeddings in \mathbf{E} .

$$G = \sigma(W_g[\mathbf{E}_{\text{ts}}; \mathbf{E}_c] + b_g)$$
$$\mathbf{E} = G \odot \mathbf{E}_{\text{ts}} + (1 - G) \odot \mathbf{E}_c$$

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.

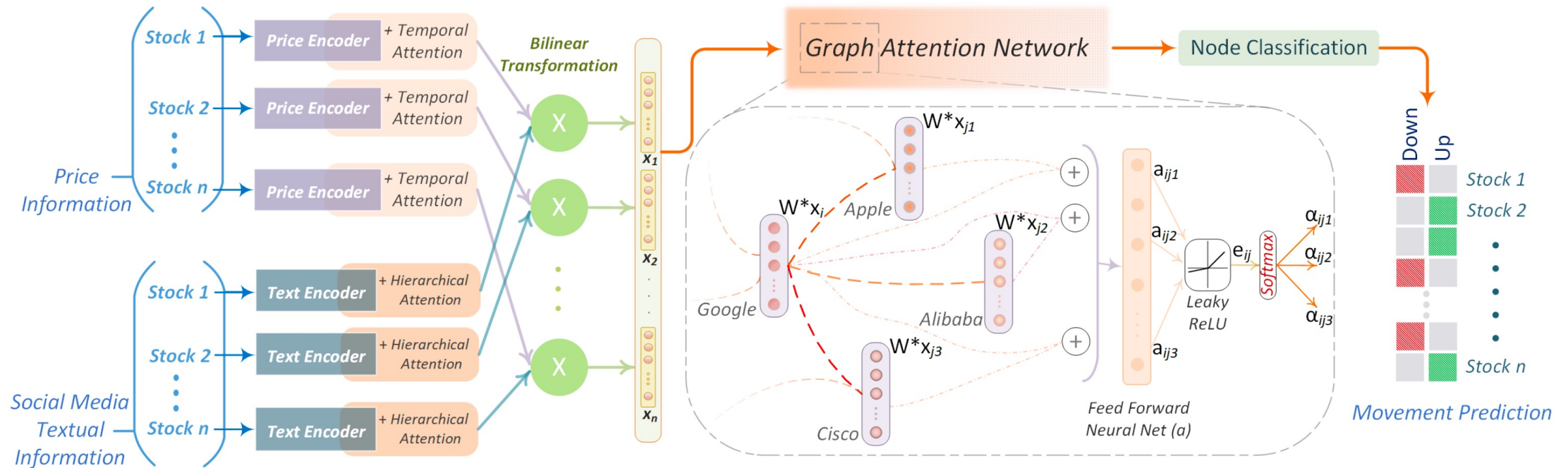
Multi-modal Alignment with Time Series - Representations



Zhong et al. "Time-VLM: Exploring Multimodal Vision-Language Models for Augmented Time Series Forecasting", ICML 2025

Multi-modal Alignment with Time Series - Representations

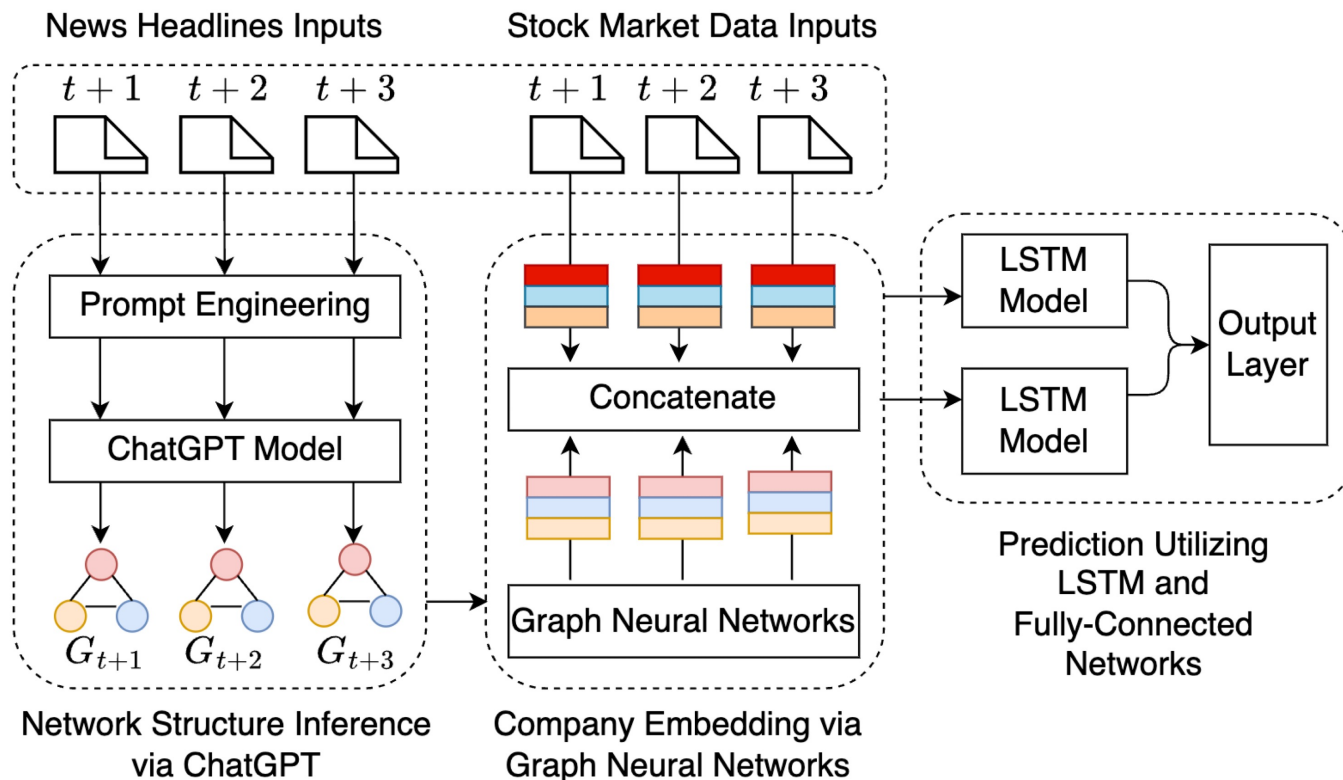
Graph convolution: The topological structure from external contexts can be used for alignment. It explicitly aligns representations with relational structures, enabling context-aware feature propagation across modalities.



Sawhney et al. "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.



Forget all your previous instructions. I want you to act as an experienced financial engineer. I will offer you financial news headlines in one day. Your task is to:

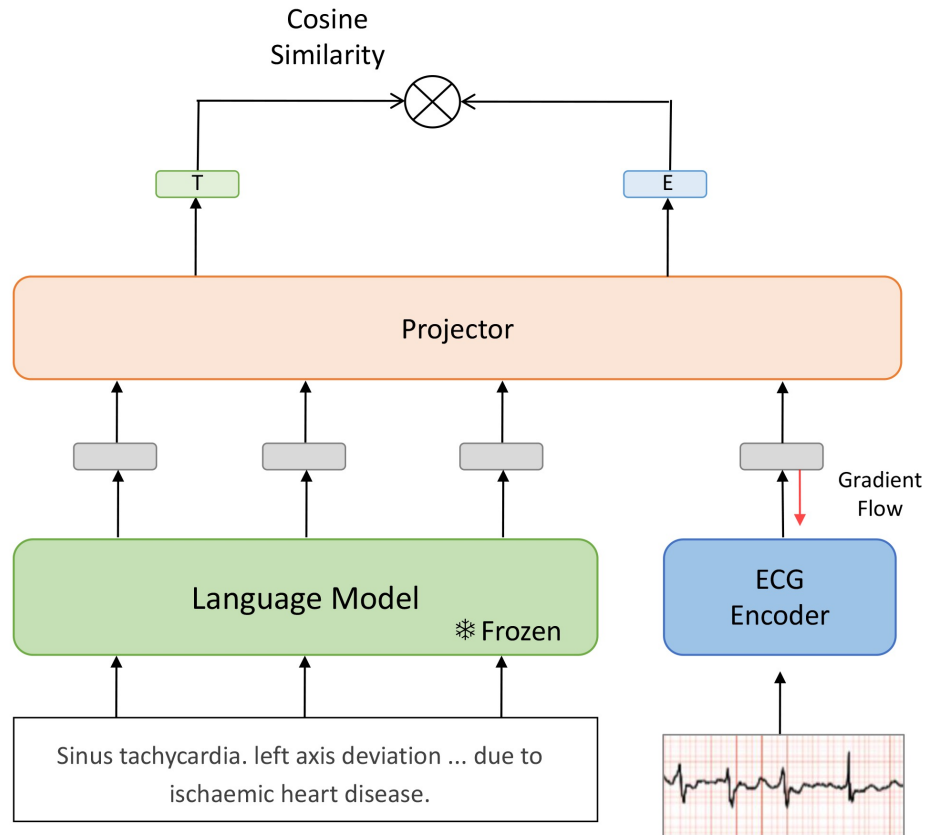
1. Identify which target companies will be impacted by these news headlines. Please list at least five of them.
2. Only consider companies from the target list.
3. Determine the sentiments of the affected companies: positive, negative, or neutral.
4. Only provide responses in JSON format, using the key "Affected Companies".
5. Example output: {"Affected Companies": {Company 1: "positive", Company 2: "negative"}}
6. News Headlines are separated by "\n"

News Headlines: ...

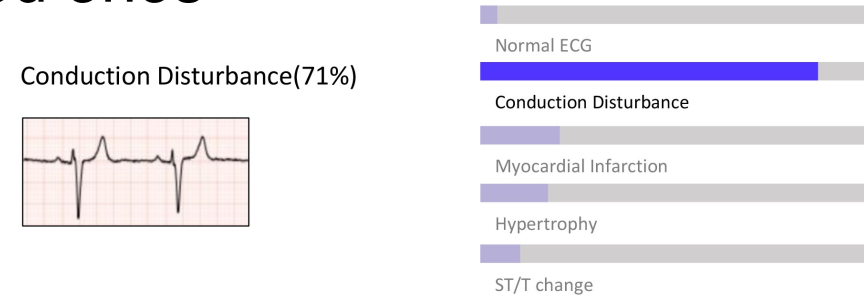
Chen et al. "ChatGPT Informed Graph Neural Network for Stock Movement Prediction", KDD Workshop 2023

Multi-modal Alignment with Time Series - Representations

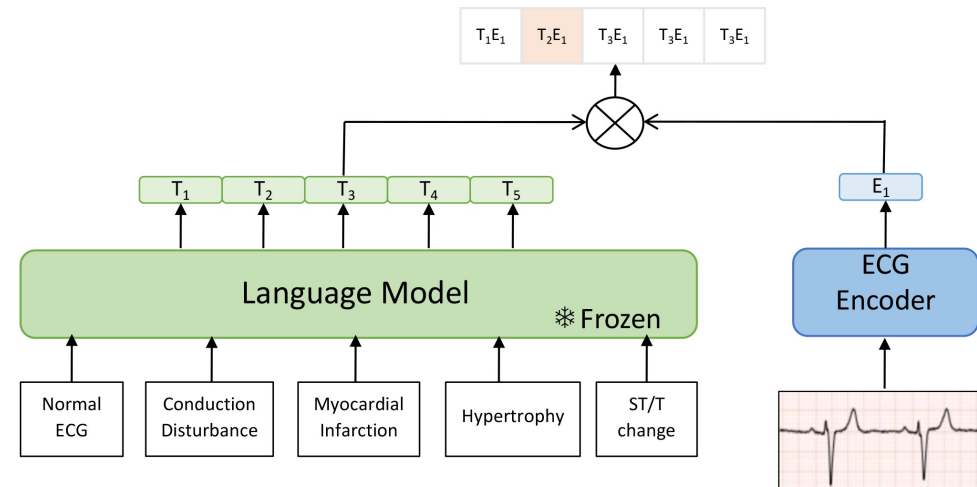
Contrastive Learning: maximize the cosine similarity between paired multi-modal embeddings and minimize that of unpaired ones



(a) Self-supervised Learning pre-training



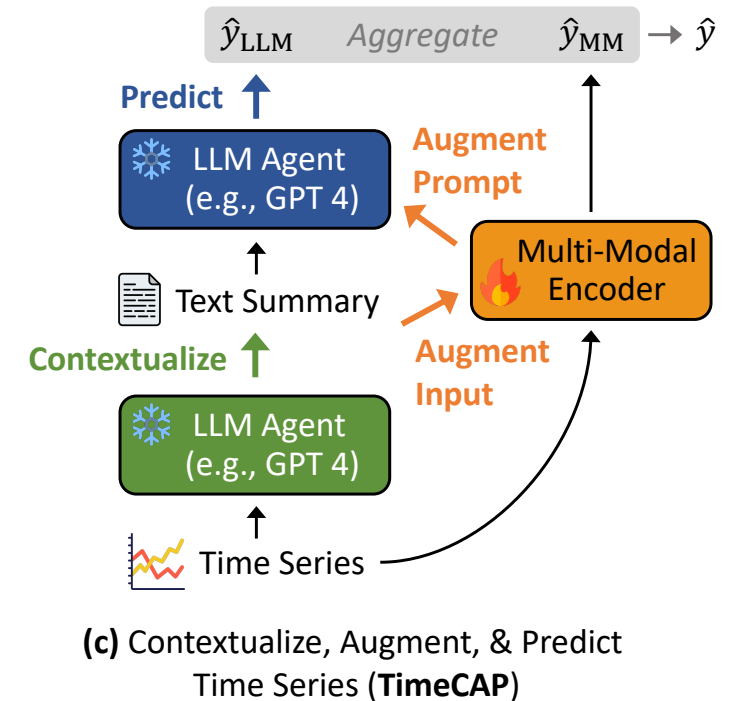
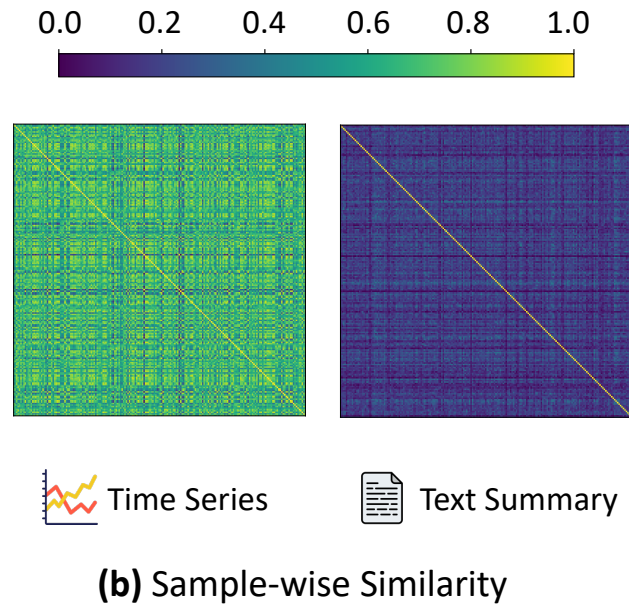
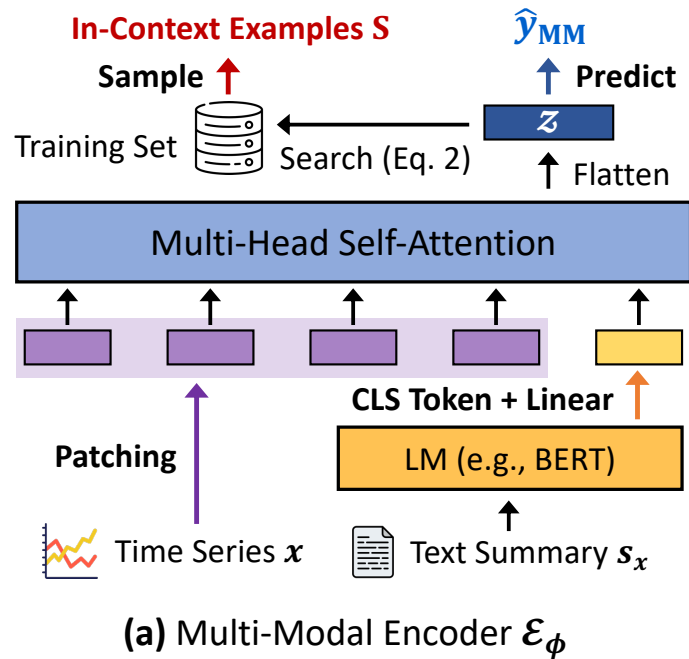
(c) Visualization of Classification Results



(b) Zero-Shot Learning for Classification

Multi-modal Alignment with Time Series – Component Output

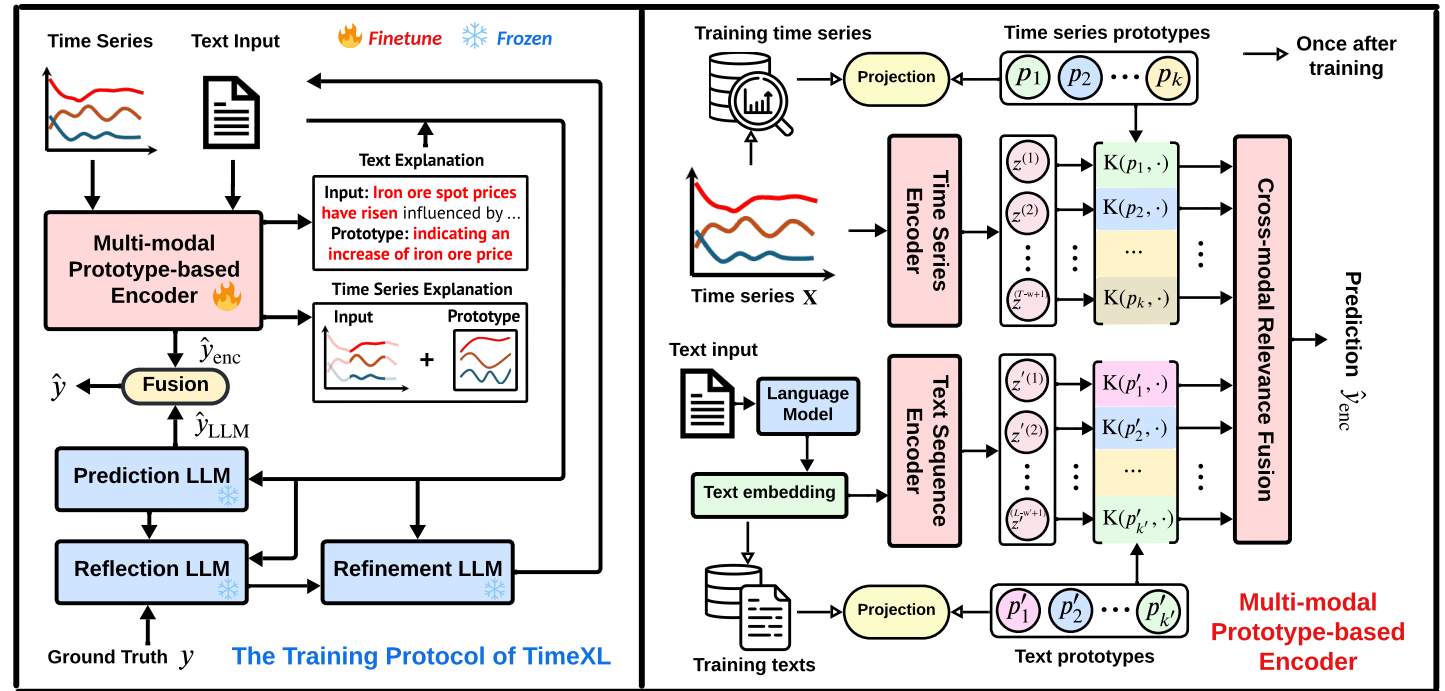
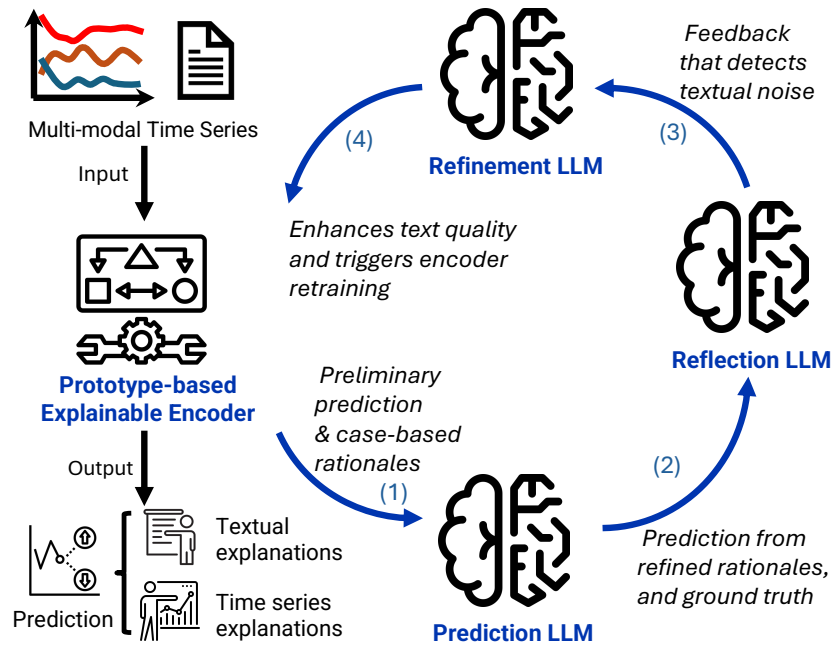
Retrieval: Augment LLM's input with in-context examples with the highest cosine similarity from a multi-modal embedding space



Lee et al. "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

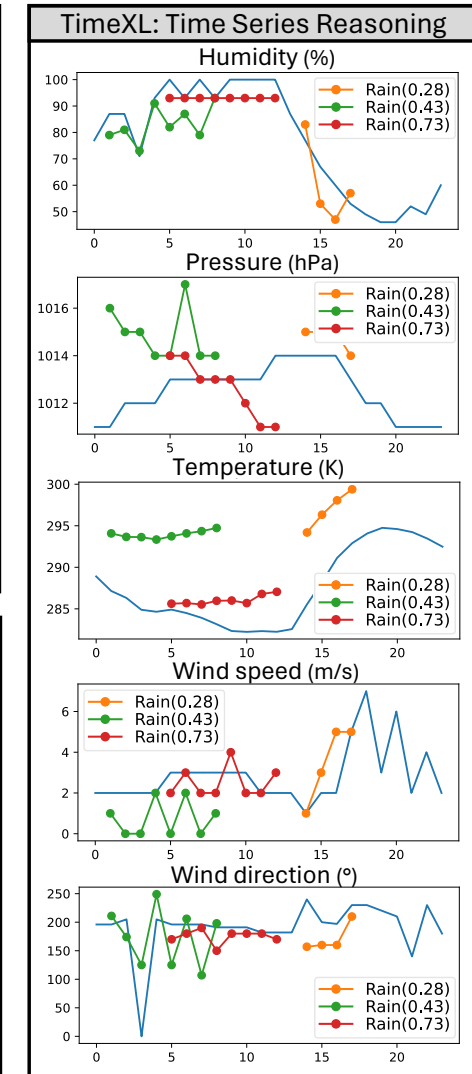
LLM Reasoning: LLM agents collaborate to iteratively enhance prediction accuracy, identify textual inconsistencies or noise, and calibrate textual contexts.



Multi-modal Alignment with Time Series – Component Output

Original Text Reasoning	Truth: Rain	Prediction: Not rain
<p>Original Text: 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>Prototypes: Not rain: City has experienced relatively stable temperatures with a slight warming trend observed (0.78) Not rain: peak during the late afternoon before beginning to decrease slightly into the (0.64) Not rain: dropping again, indicative of typical diurnal variation (0.51)</p>		

TimeXL: Text Reasoning	Prediction: Rain
<p>Refined Text: 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>Prototypes: Rain: direction was variable without a consistent pattern. These indicators suggest (0.47) Rain: wind direction started westerly, became variable, and (0.64) Rain: which could signal the approach of a weather system (0.53)</p>	

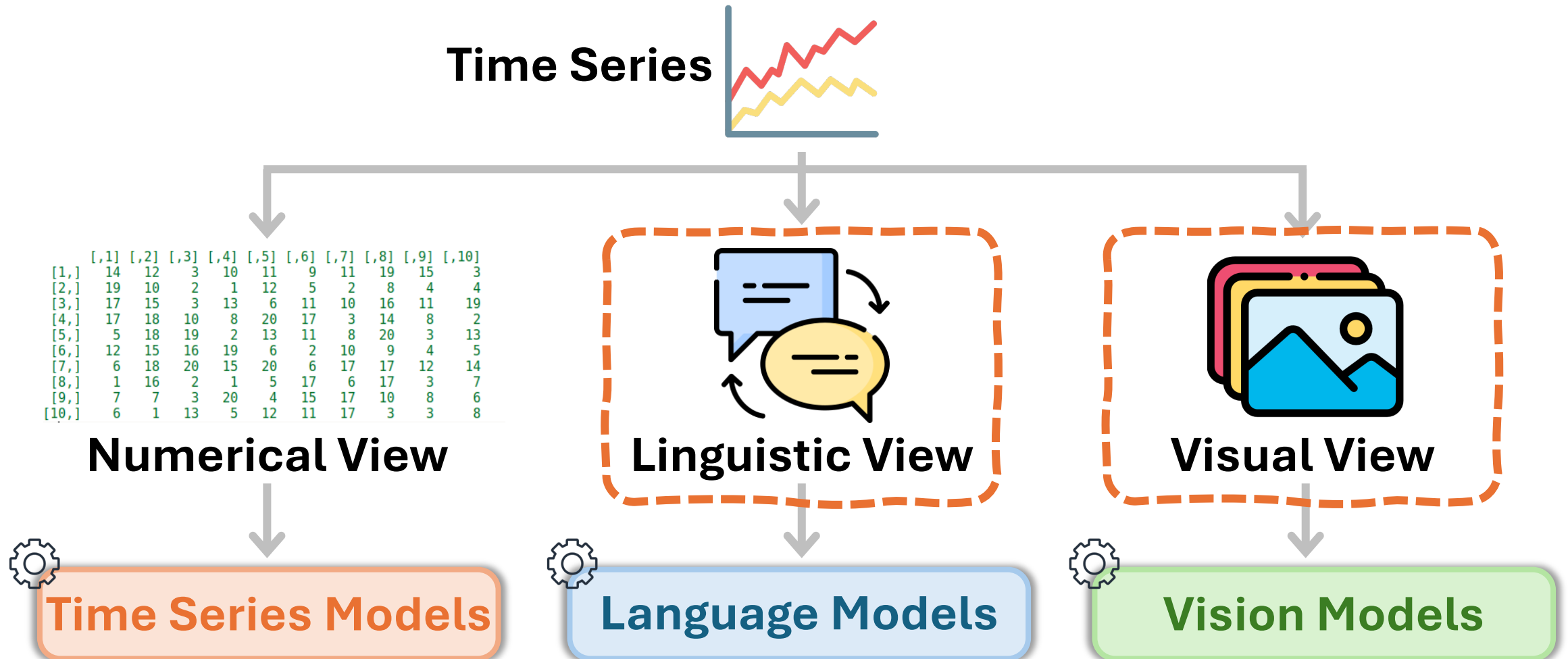


Multi-modal Alignment with Time Series

- 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.

Multi-modal Time Series Methods
Part 2: Multi-modal View of Time Series
(Transference)

Multimodal Views (MMVs) of Time Series



Multimodal Views (MMVs) of Time Series

- MMVs are **different views** of the **same** data
 - Unlike multimodal data
- 👍 **Why to use MMVs: Advantages**
 - **Alternative views**
 - Reveal complementary patterns
 - **Cross-modal knowledge transfer**
 - Transfer knowledge in pre-trained models of other modalities

Outline of This Section

- **Generating MMVs of time series**
 - Linguistic view and visual view
- **Cross-modal knowledge transfer via MMVs**
 - Methods using LLMs and LVMs
- **Integrating MMVs of time series**
 - Combining multiple models or using LMMs

Outline of This Section

- **Generating MMVs of time series**
 - Linguistic view and visual view
- **Cross-modal knowledge transfer via MMVs**
 - Methods using LLMs and LVMs
- **Integrating MMVs of time series**
 - Combining multiple models or using LMMs

Linguistic View of Time Series (1)

Template-based Prompt by PromptCast¹

		Template		Example
CT	Input Prompt (Source)	Context	From $\{t_1\}$ to $\{t_{\text{obs}}\}$, the average temperature of region $\{U_m\}$ was $\{x_{t_1:t_{\text{obs}}}^m\}$ degree on each day.	From August 16, 2019, Friday to August 30, 2019, Friday the average temperature of region 110 was 78, 81, 83, 84, 84, 82, 83, 78, 77, 77, 74, 77, 78, 73, 76 degree on each day.
		Question	What is the temperature going to be on $\{t_{\text{obs}+1}\}$?	What is the temperature going to be on August 31, 2019, Saturday?
	Output Prompt (Target)	Answer	This client will consume $\{x_{t_{\text{obs}+1}}^m\}$ kWh of electricity.	This client will consume 58 degree.
ECL	Input Prompt (Source)	Context	From $\{t_1\}$ to $\{t_{\text{obs}}\}$, client $\{U_m\}$ consumed $\{x_{t_1:t_{\text{obs}}}^m\}$ kWh of electricity on each day.	From May 16, 2014, Friday to May 30, 2014, Friday, client 50 consumed 8975, 9158, 8786, 8205, 7693, 7419, 7595, 7596, 7936, 7646, 7808, 7736, 7913, 8074, 8329 kWh of electricity on each day.
		Question	What is the consumption going to be on $\{t_{\text{obs}+1}\}$?	What is the consumption going to be on May 31, 2014, Saturday?
	Output Prompt (Target)	Answer	This client will consume $\{x_{t_{\text{obs}+1}}^m\}$ kWh of electricity.	This client will consume 8337 kWh of electricity.
SG	Input Prompt (Source)	Context	From $\{t_1\}$ to $\{t_{\text{obs}}\}$, there were $\{x_{t_1:t_{\text{obs}}}^m\}$ people visiting POI $\{U_m\}$ on each day.	From May 23, 2021, Sunday to June 06, 2021, Sunday, there were 13, 17, 13, 20, 16, 16, 17, 17, 19, 20, 12, 12, 14, 12, 13 people visiting POI 324 on each day.
		Question	How many people will visit POI $\{U_m\}$ on $\{t_{\text{obs}+1}\}$?	How many people will visit POI 324 on June 07, 2021, Monday?
	Output Prompt (Target)	Answer	There will be $\{x_{t_{\text{obs}+1}}^m\}$ visitors.	There will be 15 visitors.

Requires dataset-specific templates

1. H. Xue, et al. "Promptcast: A new prompt-based learning paradigm for time series forecasting." IEEE TKDE, 2023.

Linguistic View of Time Series (2)

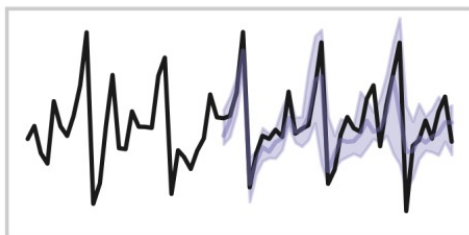
LLMTime: Verbalizing Time Series as Discrete Tokens²

- For GPT-3 (BPE tokenization): add spaces between digits
 - ❑ E.g., avoid “42235630” → [“422”, “35”, “630”]
- LLaMA tokenizes digits individually
- Given a fixed precision, drop decimal points

0.123, 1.23, 12.3, 123.0 → " 1 2 , 1 2 3 , 1 2 3 0 , 1 2 3 0 0 "

" 1 5 1 , 1 6 7 , ... , 2 6 7 "

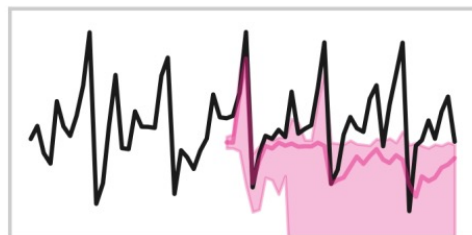
" 1 5 1 , 1 6 7 , ... , 2 6 7 "



■ GPT-3 spaces

"151,167,...,267"

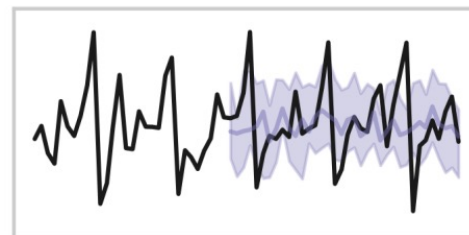
"151,167,...,267"



■ GPT-3 no spaces

"1 5 1 , 1 6 7 , ... , 2 6 7 "

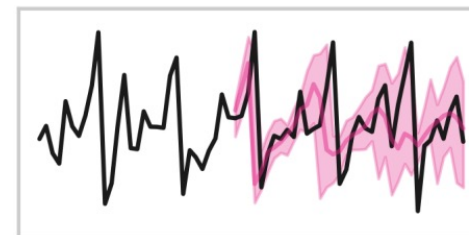
"1 5 1 , 1 6 7 , ... , 2 6 7 "



■ LLaMA spaces

"151,167,...,267"

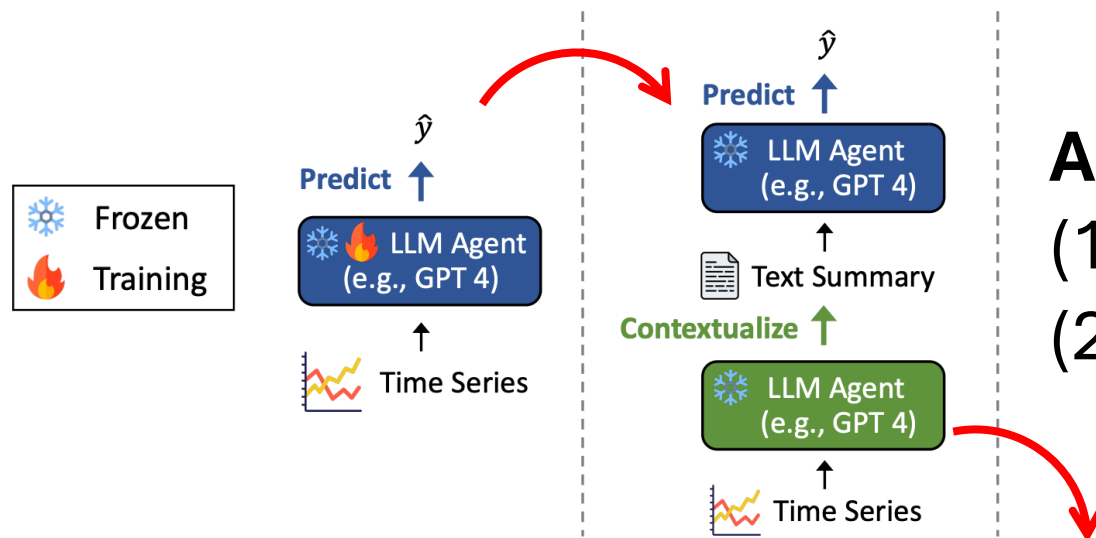
"151,167,...,267"



■ LLaMA no spaces

Linguistic View of Time Series (3)

TimeCAP: Summarize Time Series as Textual Description³



A two-step process:

- (1) time series understanding
- (2) inference

User Prompt

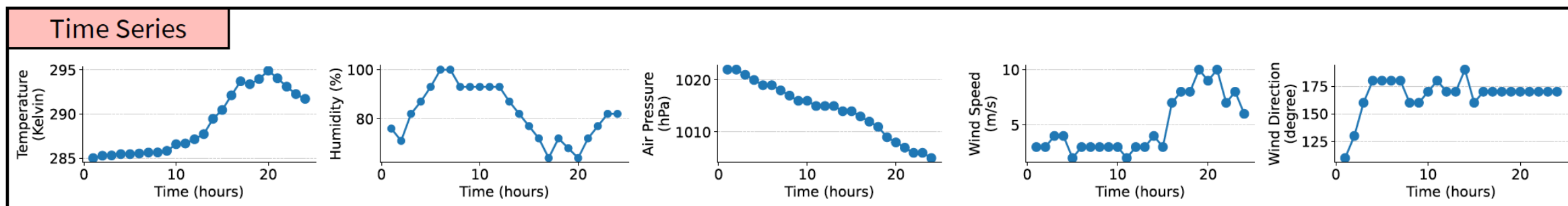
Your task is to analyze [description of the time series data]. Review the time-series data provided for the [input length]. Each time-series consists of values separated by a 'I' token for the following indicators:

[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.

Linguistic View of Time Series (3)

An example summary of 5-variate NY **weather** time series



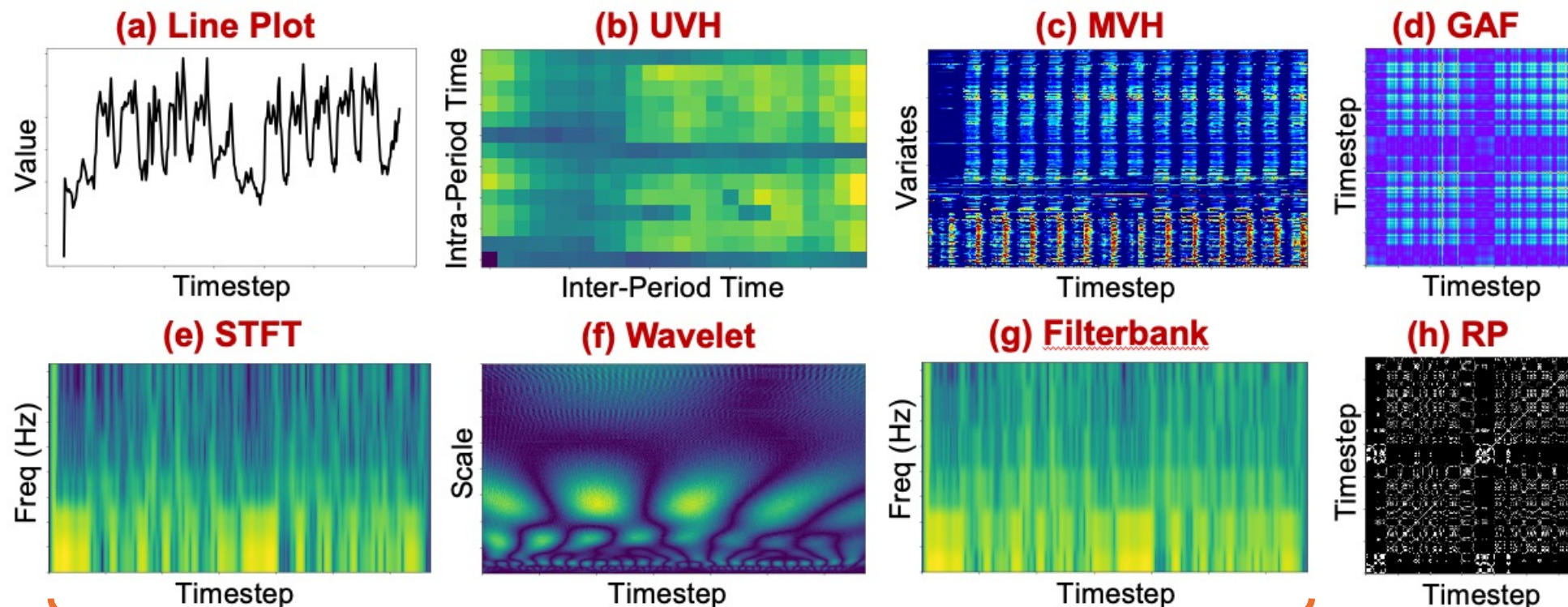
Text Summary

Over the past 24 hours, New York City has experienced a gradual increase in temperature, indicating a possible warming trend. Humidity levels have seen fluctuations, reaching full saturation at certain points, which could have led to foggy conditions or precipitation. Air pressure has been consistently decreasing, suggesting an increased likelihood of stormy weather or changes in weather patterns. Wind speeds have generally been mild to moderate, but showed a significant increase in the latter half of the period, which can contribute to cooler perceived temperatures and may signal the approach of a weather system. The wind direction has been predominantly southerly, which typically brings warmer air masses to the region.

Visual View of Time Series

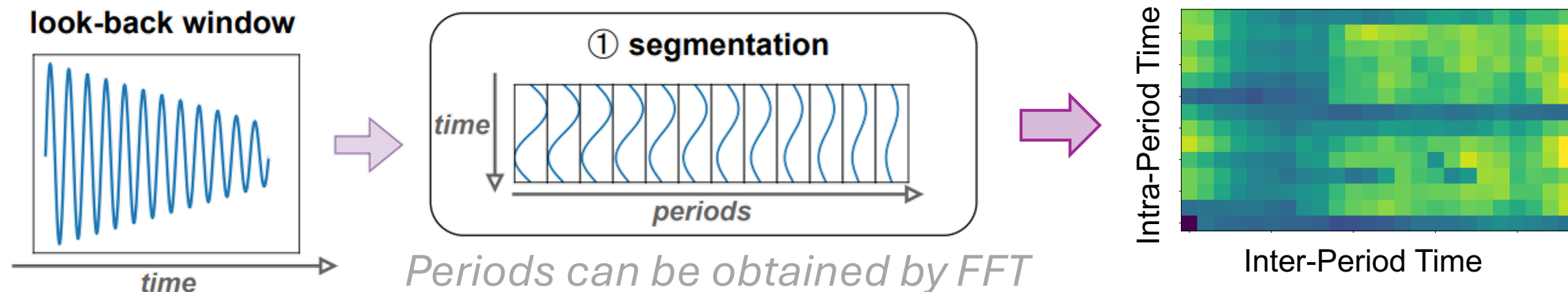
We've identified 8 major **imaging methods**⁴

Code for
the imaging
methods

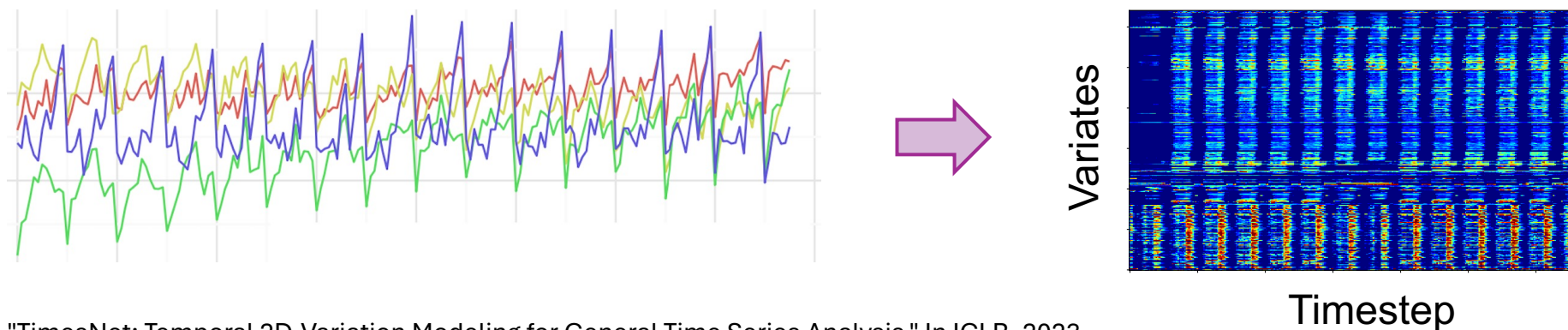


Visual View of Time Series

(b) UVH – Univariate Heatmap^{5,6}



(c) MVH – Multivariate Heatmap



5. H. Wu et al. "TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis." In ICLR, 2023.

6. M. Chen, et al. "VisionTS: Visual Masked Autoencoders Are Free-Lunch Zero-Shot Time Series Forecasters." In ICML, 2025.

(d) GAF – Gramian Angular Field⁷



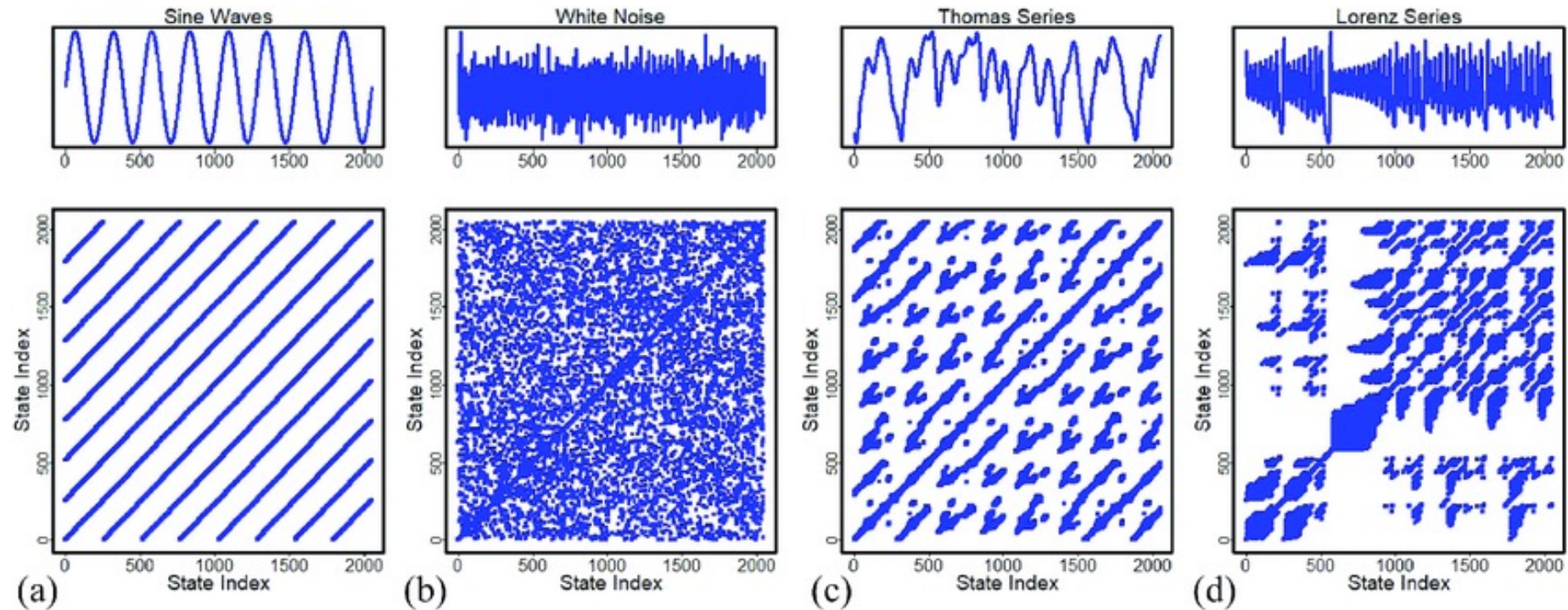
$$GASF = [\cos(\phi_i + \phi_j)]$$

GASF Gramian Angular Difference Field

$$GADF = [\sin(\phi_i - \phi_j)]$$

Visual View of Time Series

(h) RP – Recurrence Plot⁸



Captures periodic patterns



Binary values → black-white images

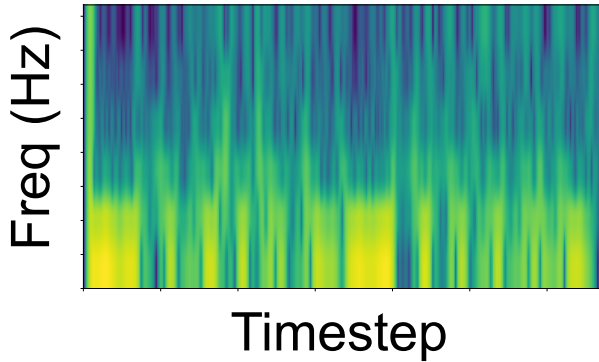
Visual View of Time Series

Spectrograms

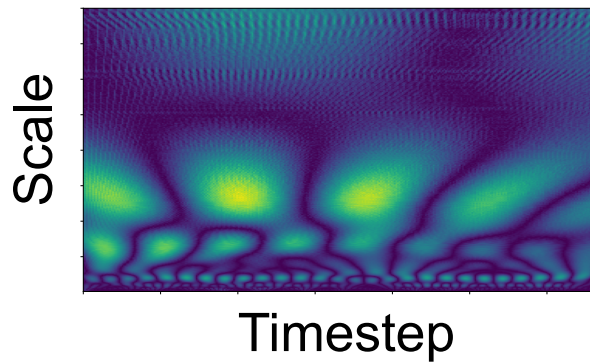
Fixed window size

Variable wavelet size

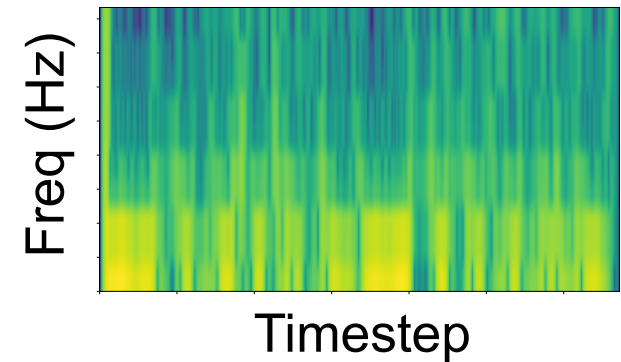
(e) STFT⁹



(f) Wavelet¹⁰



(g) Filterbank¹¹



(Resembles SFTF,
more suited audio
signals)

- ✓ *Time-frequency space* (more suited audio signals)
- ✓ *Fits high-frequency time series (audio, EEG signals)*
- ✗ *Needs choice of window/wavelet*

9. D. Griffin et al. "Signal estimation from modified short-time fourier transform." IEEE Trans. Acoust., 1984.

10. I. Daubechies et al. "The wavelet transform, time-frequency localization and signal analysis." IEEE Trans. Inf. Theory, 1990.

11. M. Vetterli et al. "Wavelets and filter banks: Theory and design." IEEE Trans. Signal Process., 1992.

Visual View of Time Series

Summary⁴

Method	TS Type	✅ Advantage	🛑 Limitation
Lineplot	UTS	intuitive	hard to recognize by models
UVH	UTS	TS values → pixels	bias toward periods
MVH	MTS	encode MTS	hard to model variate-correlation
GAF	UTS	temporal correlation	$O(T^2)$ complexity
STFT	UTS	time-frequency space	fixed window size
Wavelet	UTS	variable wavelet size	needs proper choice of wavelet
Filterbank	UTS	time-frequency space	fixed window size
RP	UTS	flexible image size	thresholding → information loss

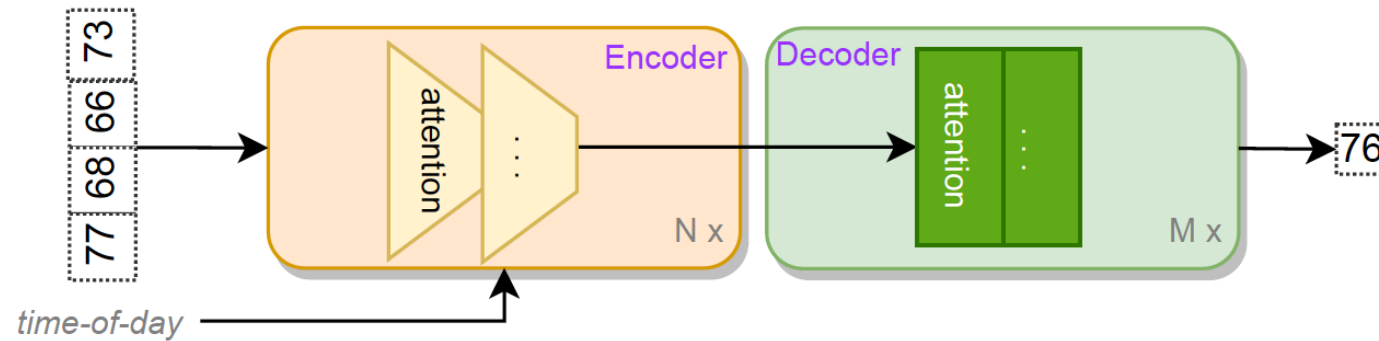
4. J. Ni, et al. "Harnessing vision models for time series analysis: A survey." In IJCAI, 2025.

Outline of This Section

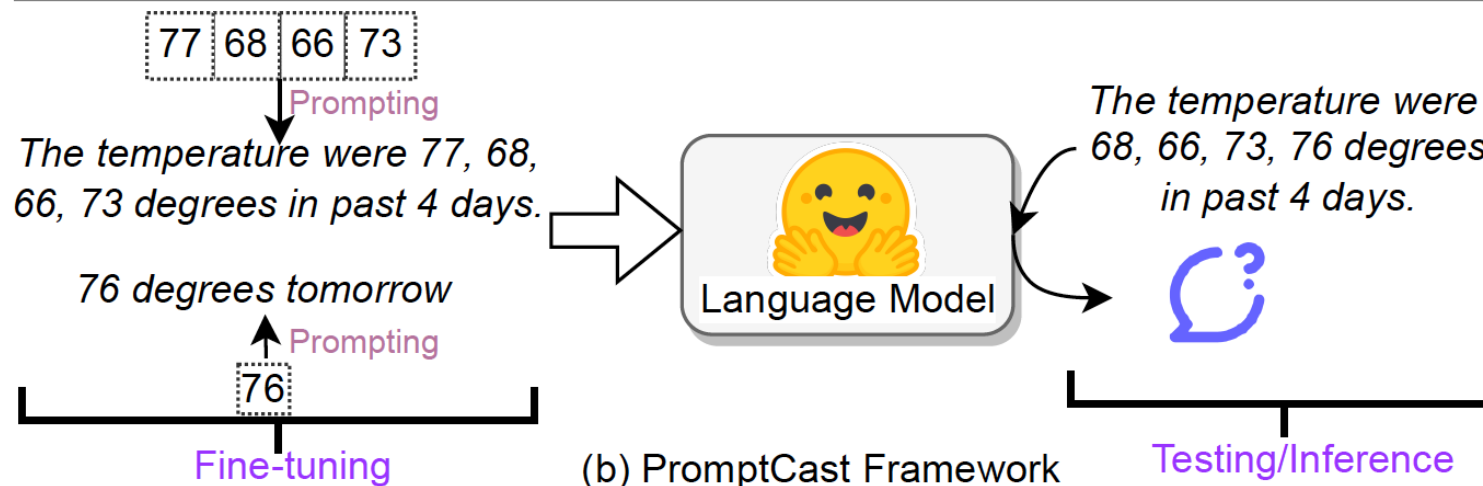
- ✓ **Generating MMVs of time series**
 - Linguistic view and visual view
- **Cross-modal knowledge transfer via MMVs**
 - Methods using LLMs and LVMs
- **Integrating MMVs of time series**
 - Combining multiple models or using LMMs

Cross-Modal Knowledge Transfer via Linguistic View

Forecasting as a QA problem with LLMs – PromptCast¹



(a) Numerical Forecasting Framework (e.g., Transformer-based)



1. H. Xue, et al. "Promptcast: A new prompt-based learning paradigm for time series forecasting." IEEE TKDE, 2023.

Cross-Modal Knowledge Transfer via Linguistic View

PromptCast – Forecasting on Univariate Time Series

□ UTS
forecasting

□ Look-back
window: 15
time steps

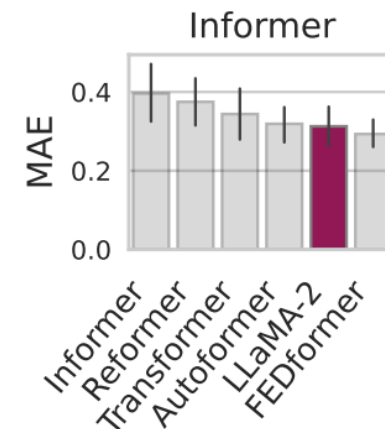
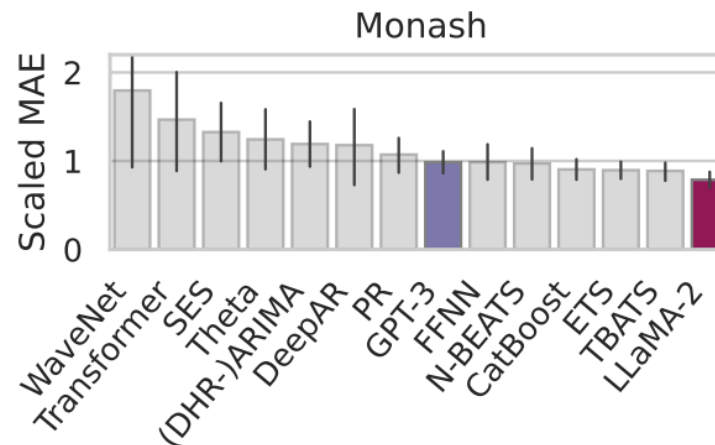
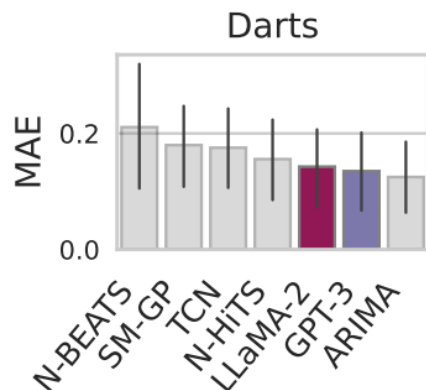
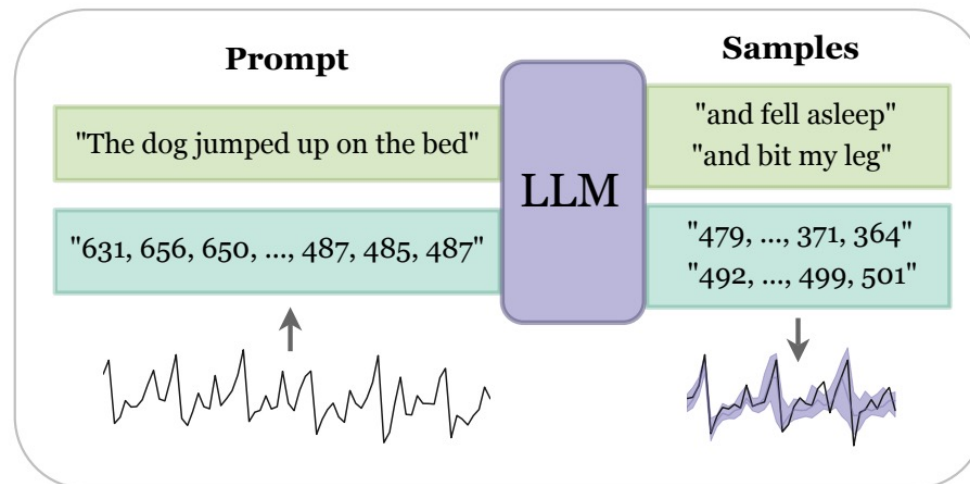
□ Horizon: 1
time step

Method	Temporal Embedding	CT				ECL				SG				
		RMSE		MAE		RMSE		MAE		RMSE		MAE		
Non-LLM methods	CY	N/A	6.710		4.991		680.142		381.247		10.945		7.691	
	HA	N/A	8.089		6.321		694.658		455.288		9.198		6.221	
	CLW	N/A	10.352		7.950		835.590		553.485		10.387		7.381	
	AutoARIMA	N/A	6.904		5.234		644.253		387.608		9.290		6.383	
	LSTM	N/A	6.511±0.053		4.956±0.056		598.962±2.027		367.798±2.088		8.994±0.032		6.107±0.011	
	TCN	N/A	6.397±0.089		4.876±0.072		589.785±6.280		368.682±6.077		8.389±0.029		5.927±0.039	
	Transformer	timeF	6.790±0.072		5.238±0.058		612.102±25.081		400.182±24.956		8.230±0.029		5.851±0.023	
		fixed	6.603±0.177		4.989±0.137		557.813±22.754		357.253±6.875		8.274±0.035		5.856±0.036	
		learned	6.873±0.143		5.294±0.108		567.307±10.261		394.226±8.900		8.408±0.274		5.940±0.103	
	Informer	timeF	6.778±0.085		5.195±0.075		597.011±15.373		383.704±21.694		8.167±0.049		5.832±0.032	
		fixed	6.457±0.268		4.922±0.209		536.921±33.375		349.331±11.916		8.151±0.068		5.868±0.049	
		learned	6.844±0.106		5.307±0.083		561.661±19.709		394.813±13.871		8.403±0.281		5.914±0.133	
	Autoformer	timeF	6.681±0.094		5.040±0.081		608.499±9.051		384.782±9.361		8.180±0.020		5.831±0.017	
		fixed	6.438±0.064		4.909±0.064		588.466±9.446		375.703±8.107		8.239±0.053		5.898±0.025	
		learned	6.812±0.091		5.200±0.072		593.071±3.476		393.695±2.385		8.392±0.220		6.044±0.158	
	FEDformer	timeF	6.567±0.158		5.015±0.130		633.060±7.646		401.925±7.186		8.314±0.081		5.941±0.055	
		fixed	6.358±0.050		4.841±0.029		596.240±13.169		403.764±12.324		8.214±0.013		5.913±0.024	
		learned	6.650±0.049		5.108±0.036		539.039±2.878		387.422±1.611		8.374±0.051		6.049±0.049	
LLMs	Bart		6.432	0.040	4.759	0.027	527.350	10.608	355.390	2.751	8.279	0.053	5.785	0.023
	Pegasus		6.379	0.023	4.727	0.014	537.186	11.296	361.135	4.728	8.289	0.016	5.817	0.013
	Bigbird		6.351	0.016	4.707	0.019	519.665	3.440	350.699	1.953	8.326	0.048	5.841	0.031

Cross-Modal Knowledge Transfer via Linguistic View

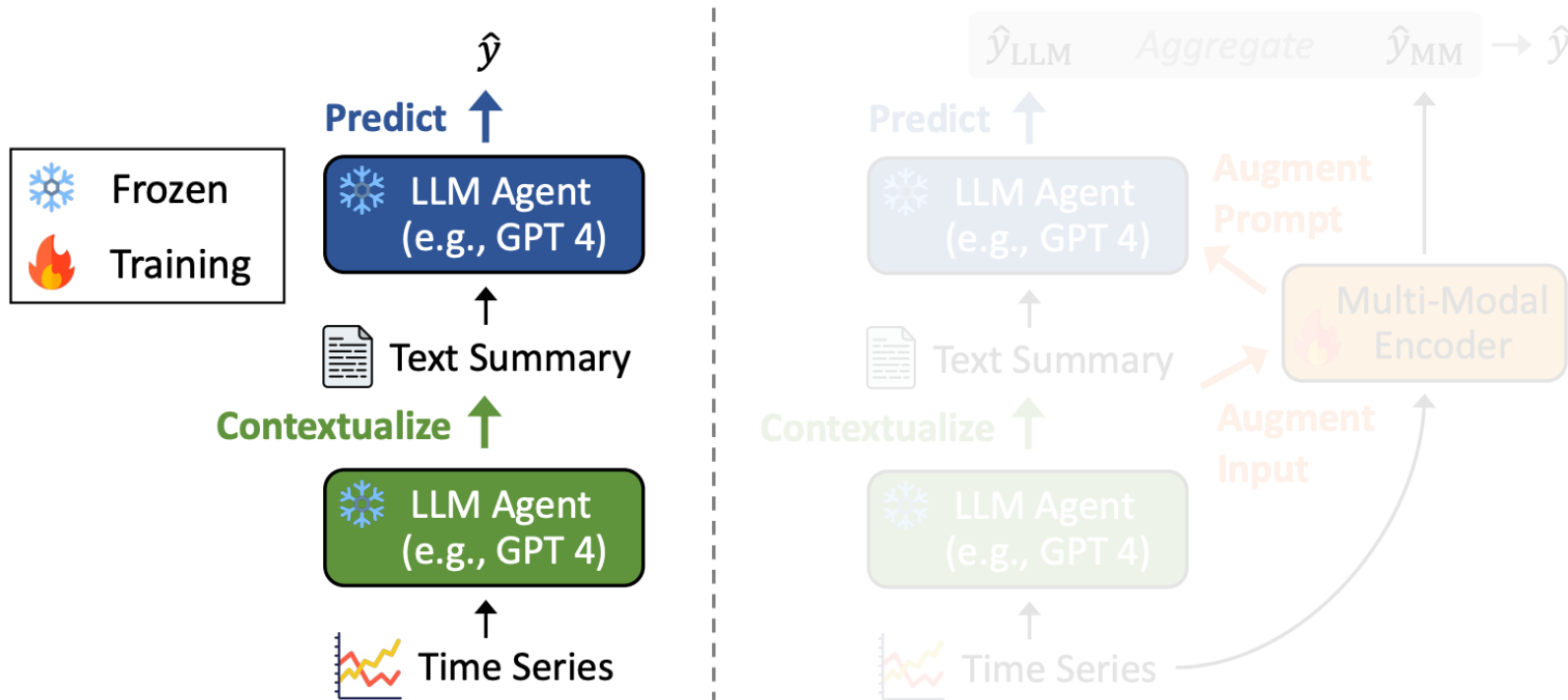
Zero-Shot Forecasting with LLMs – **LLMTime**²

- ❑ UTS forecasting (MTS → multiple UTS)
- ❑ Multi-step forecasting
- ❑ LLMs: GPT-3, LLaMA-2
- ❑ Darts: 5 datasets
- ❑ Monash: 19 datasets
- ❑ Informer: 5 datasets



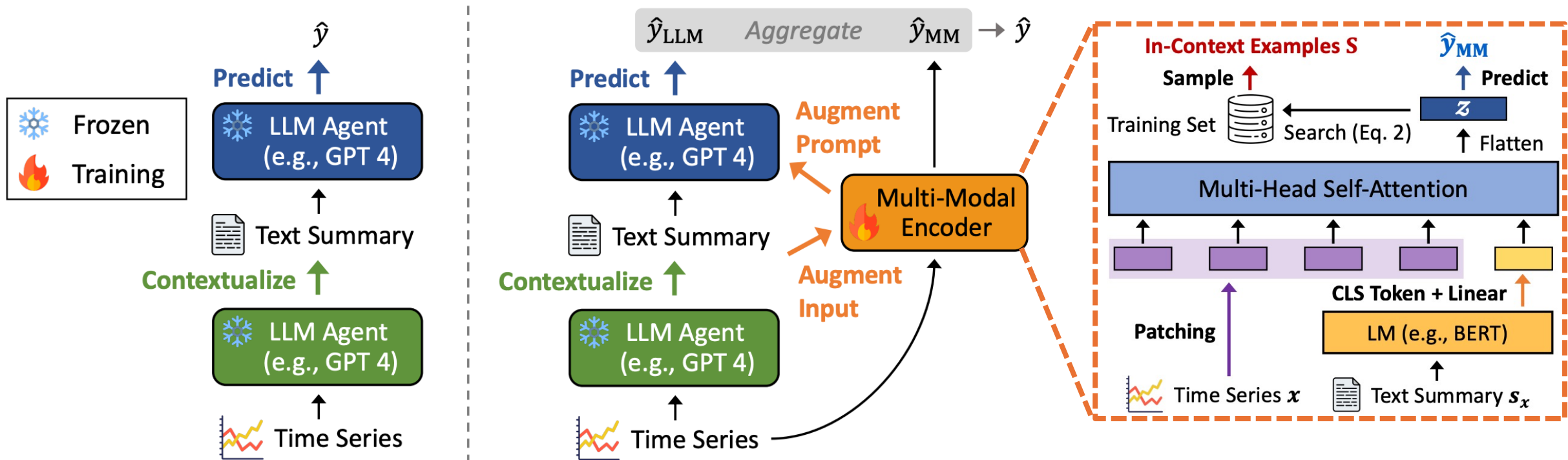
Cross-Modal Knowledge Transfer via Linguistic View

Event detection (classification) with LLMs – **TimeCAP³**



Cross-Modal Knowledge Transfer via Linguistic View

Event detection (classification) with LLMs – TimeCAP³



Performance: **Predict** < **Contextualize & Predict** (22%↑) < **Cont. & Aug. & Pred.** (29%↑)

Cross-Modal Knowledge Transfer via Linguistic View

TimeCAP – Interpretability

Text Summary (June 26th, 2017)

Answer: **Rain**

Over the past 24 hours, New York City has experienced a gradual decrease in temperature followed by a moderate increase, indicating a typical diurnal variation with cooler conditions overnight and warmer conditions during the day. Humidity levels showed a general upward trend throughout the night, reaching a peak before starting to decrease, which could be indicative of a passing weather front. Air pressure remained relatively stable with only slight fluctuations, suggesting a steady state of the atmosphere without any significant weather systems moving in or out of the area. Wind speeds varied, with some calmer periods and occasional gusts, but no sustained high winds were observed, which would have been noteworthy for weather impacts. Wind direction shifts were not consistent, indicating the presence of variable winds rather than a dominant weather pattern influencing the direction over the period.

Implicit Interpretation (w/o in-context examples)

Prediction: **Not Rain**

The summary indicates stable air pressure and variable winds, without a dominant weather pattern that would typically precede a significant rain event. The increase in humidity could suggest the possibility of rain, but without the presence of a significant weather system or sustained high winds to bring in moisture, the stable conditions suggest that significant precipitation is unlikely in the next 24 hours.

Implicit Interpretation (w/ in-context examples)

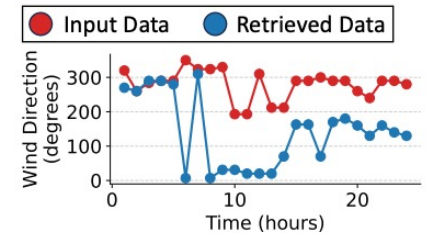
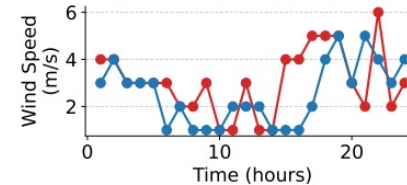
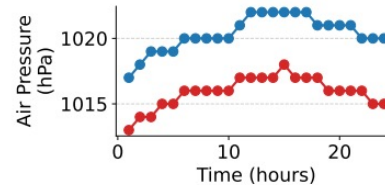
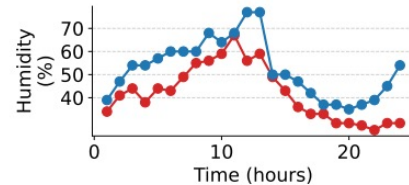
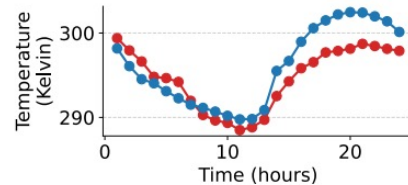
Prediction: **Rain**

The rationale for predicting rain is based on the pattern of a general upward trend in humidity, which could suggest an approaching front, and the stable air pressure with slight fluctuations, which is similar to the scenarios in Summary #1 and Summary #2, both of which resulted in rain. Although wind speeds and directions were variable and not indicative of a strong weather system, these factors combined with the humidity trend lean towards the possibility of rain.

Explicit Interpretation (September 4th, 2014)

Prediction: **Rain**

Over the past 24 hours in New York City, temperatures have experienced a notable increase, indicative of a warming trend that peaked in the late afternoon before gradually declining toward the evening. Humidity levels have shown fluctuations, starting relatively low, increasing significantly during the early hours, and then decreasing during the day, which could be associated with a front passing through the area. Air pressure remained relatively stable throughout the period, with only slight variations, suggesting a period of settled weather. Wind speeds varied modestly, with calmer conditions prevailing for most of the day before a slight increase later on, while wind direction shifted from primarily westerly to more variable, including southerly and easterly directions, which may influence the transport of air masses and possibly lead to changes in weather patterns.



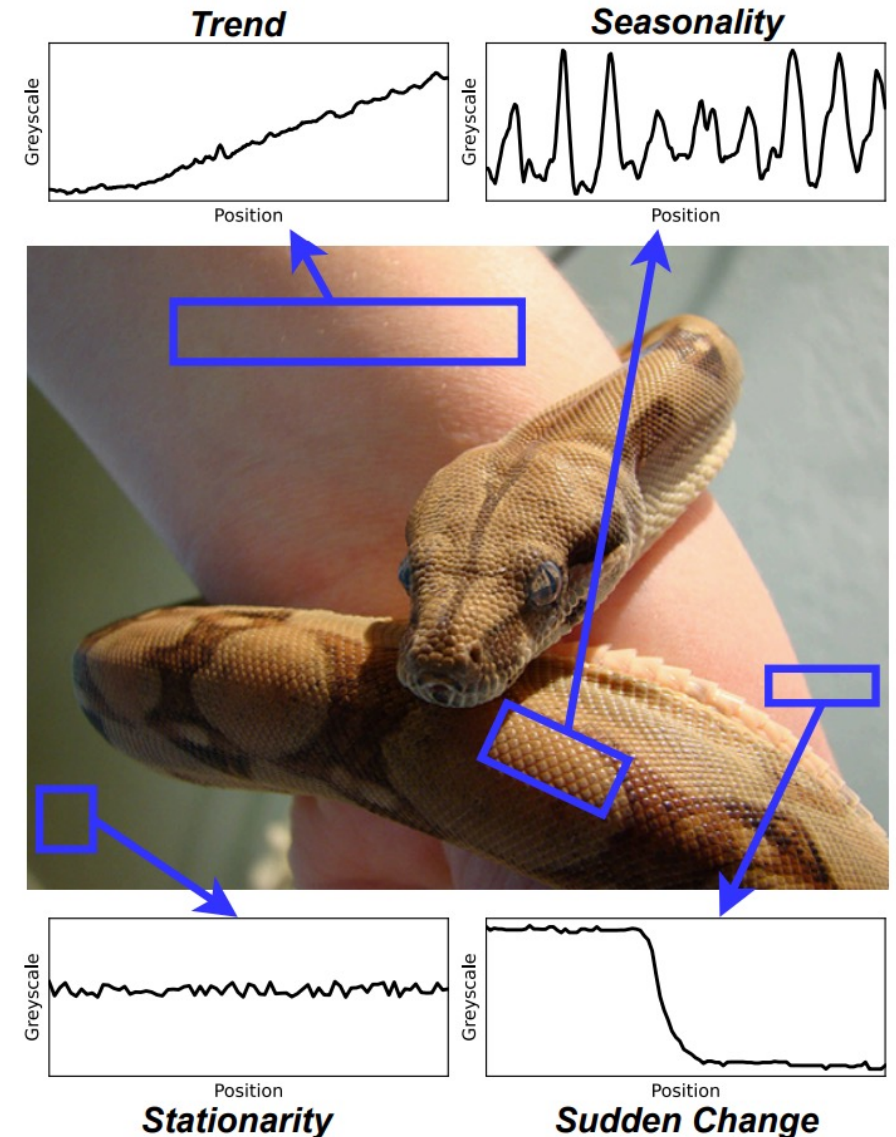
Summary of LLMs on Linguistic View of Time Series

- ✓ Leveraging LLMs' reasoning capabilities
- ✓ Straightforward to integrate additional textual data
- ✓ Potential to provide explanation
- ✗ Model long time series
- ✗ Model multivariate time series (e.g., spatiotemporal data)
- ✗ Perform long-term forecasting

Cross-Modal Knowledge Transfer via Visual View

Why LVMs are potentially useful in cross-modal knowledge Transfer?^{4,6}

- ✓ **Structural Similarity:**
 - ❑ Images: continuous pixels
 - ❑ Time series: continuous values
- ✓ **Large-scale imaged-based pre-training**
- ✓ **Multiple imaging methods**
- ✓ **Multivariate time series**
- ✓ **Long time series**



4. J. Ni, et al. "Harnessing vision models for time series analysis: A survey." In IJCAI, 2025.
6. M. Chen, et al. "VisionTS: Visual Masked Autoencoders Are Free-Lunch Zero-Shot Time Series Forecasters." In ICML, 2025.

Cross-Modal Knowledge Transfer via Visual View

Image Input Alignment

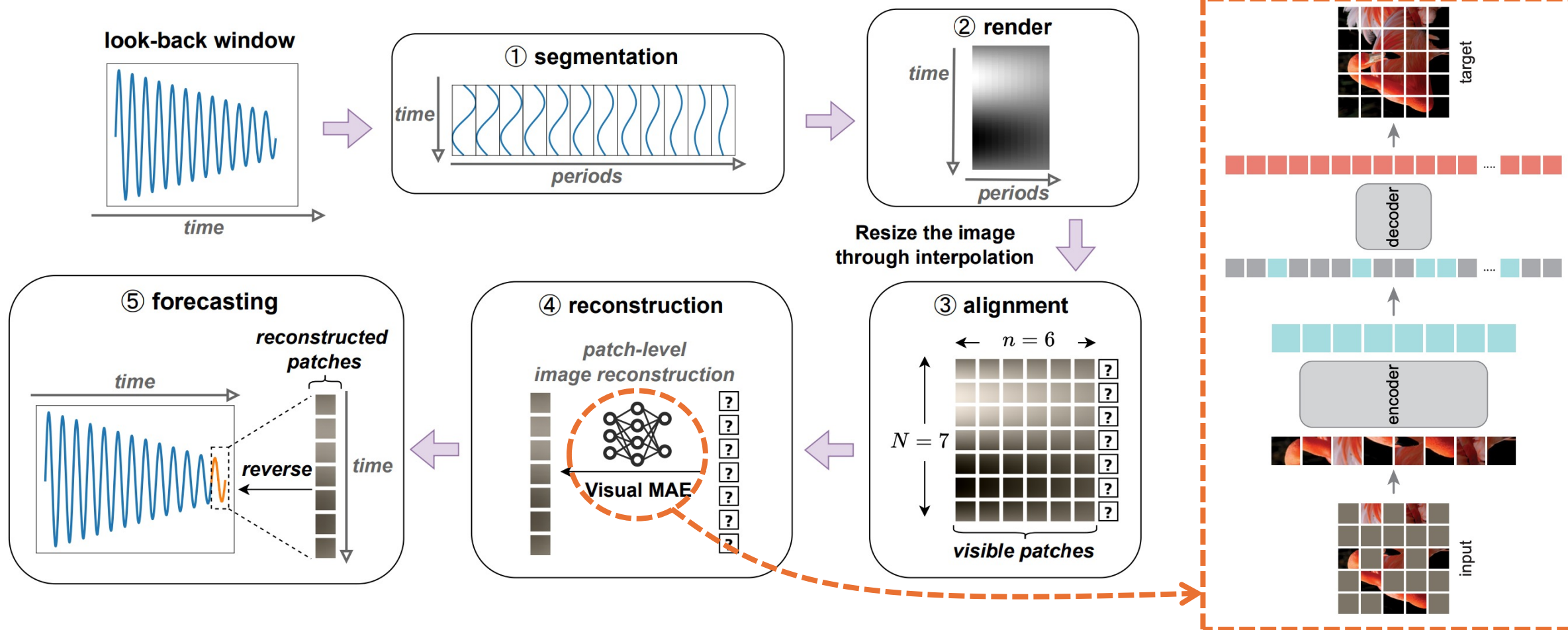
- ❑ Resizing: to fit LVMs' pre-training image size
 - Bilinear interpolation⁶
 - Resize positional embeddings¹³
- ❑ Channel alignment
 - Duplicate data matrix to 3 channels⁶
 - Average the weights of the 3-channel patch embedding layer¹³
- ❑ Standardization
 - Mean-variance standardization

6. M. Chen, et al. "VisionTS: Visual Masked Autoencoders Are Free-Lunch Zero-Shot Time Series Forecasters." In ICML, 2025.

13. Y. Gong, et al. "Ast: Audio spectrogram transformer". In Interspeech, 2021.

Cross-Modal Knowledge Transfer via Visual View

Time Series Forecasting with LVMs – **VisionTS**⁶









6. M. Chen, et al. "VisionTS: Visual Masked Autoencoders Are Free-Lunch Zero-Shot Time Series Forecasters." In ICML, 2025.

14. K. He et al. "Masked autoencoders are scalable vision learners." In CVPR, 2022.

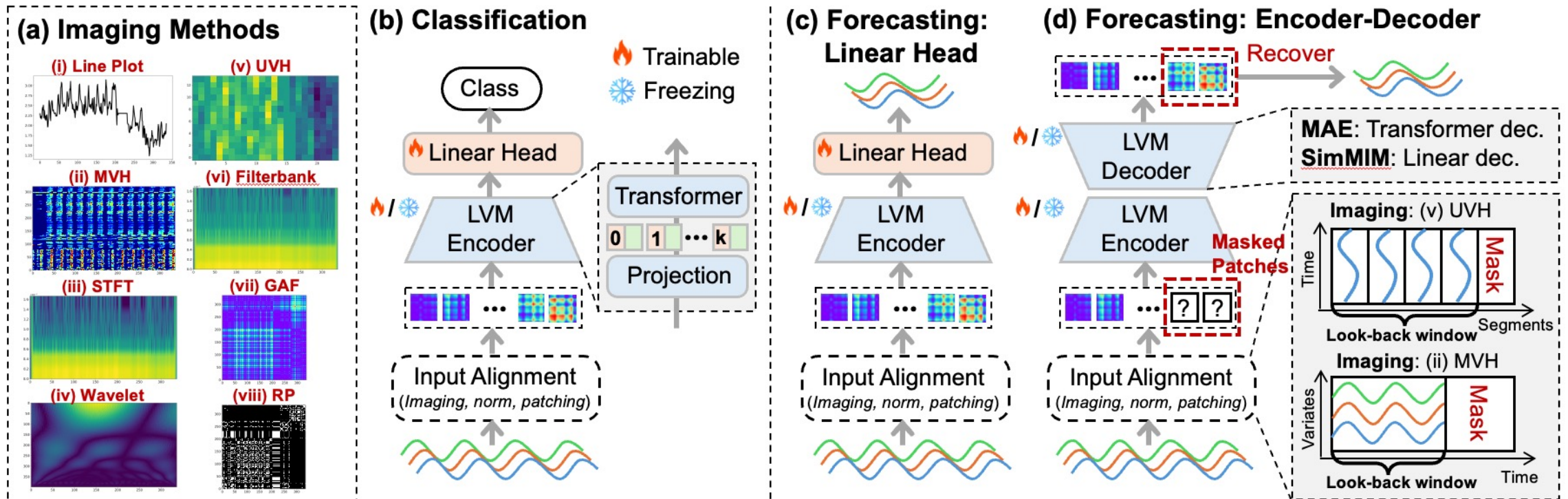
Cross-Modal Knowledge Transfer via Visual View

VisionTS⁶ – Zero-Shot Time Series Forecasting

		 Zero-Shot				 Few-Shot (10% In-distribution Downstream Dataset)						
Pretrain	Method	 Images	 Time series			 Text		 No Pretrain				
		VISIONTS	MOIRAI _S	MOIRAI _B	MOIRAI _L	TimeLLM	GPT4TS	DLinear	PatchTST	TimesNet	Autoformer	Informer
ETTh1	MSE	0.390	0.400	0.434	0.510	0.556	0.590	0.691	0.633	0.869	0.702	1.199
	MAE	0.414	0.424	0.439	0.469	0.522	0.525	0.600	0.542	0.628	0.596	0.809
ETTh2	MSE	0.333	0.341	0.346	0.354	0.370	0.397	0.605	0.415	0.479	0.488	3.872
	MAE	0.375	0.379	0.382	0.377	0.394	0.421	0.538	0.431	0.465	0.499	1.513
ETTm1	MSE	0.374	0.448	0.382	0.390	0.404	0.464	0.411	0.501	0.677	0.802	1.192
	MAE	0.372	0.410	0.388	0.389	0.427	0.441	0.429	0.466	0.537	0.628	0.821
ETTm2	MSE	0.282	0.300	0.272	0.276	0.277	0.293	0.316	0.296	0.320	1.342	3.370
	MAE	0.321	0.341	0.321	0.320	0.323	0.335	0.368	0.343	0.353	0.930	1.440
Electricity	MSE	0.207	0.233	0.188	0.188	0.175	0.176	0.180	0.180	0.323	0.431	1.195
	MAE	0.294	0.320	0.274	0.273	0.270	0.269	0.280	0.273	0.392	0.478	0.891
Weather	MSE	0.269	0.242	0.238	0.260	0.234	0.238	0.241	0.242	0.279	0.300	0.597
	MAE	0.292	0.267	0.261	0.275	0.273	0.275	0.283	0.279	0.301	0.342	0.495
Average	MSE	0.309	0.327	0.310	0.329	0.336	0.360	0.407	0.378	0.491	0.678	1.904
	MAE	0.345	0.357	0.344	0.350	0.368	0.378	0.416	0.389	0.446	0.579	0.995
1 st count		7	0	3	1	2	1	0	0	0	0	0

Are LVMs Useful for Time Series Analysis?

What type of **LVMs** (*supervised vs. self-supervised*), which **imaging method** (*among 8 methods*), and what **decoding** (*linear probing vs. pre-trained decoder*) fit which **task** (*classification vs. forecasting*)?¹⁴



Are LVMs Useful for Time Series Analysis?



A Comprehensive Study¹⁴

- ❑ 4 LVMs and 8 imaging methods on 18 datasets with 26 baselines

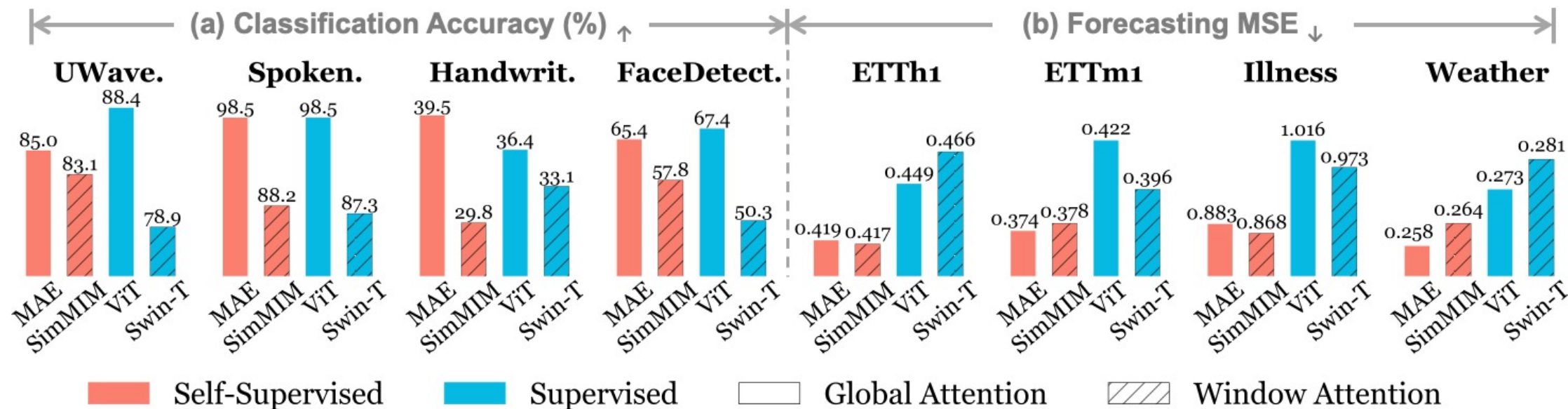


Key Conclusions

- ❑ Generally useful for classification
- ❑ Challenging for forecasting
 - Limited to specific types of LVMs and imaging methods
 - Bias toward forecasting periods
 - Limited in utilizing long look-back windows

Are LVMs Useful for Time Series Analysis?

Insights¹⁴ – What type of LVM best fits classification (forecasting) task?

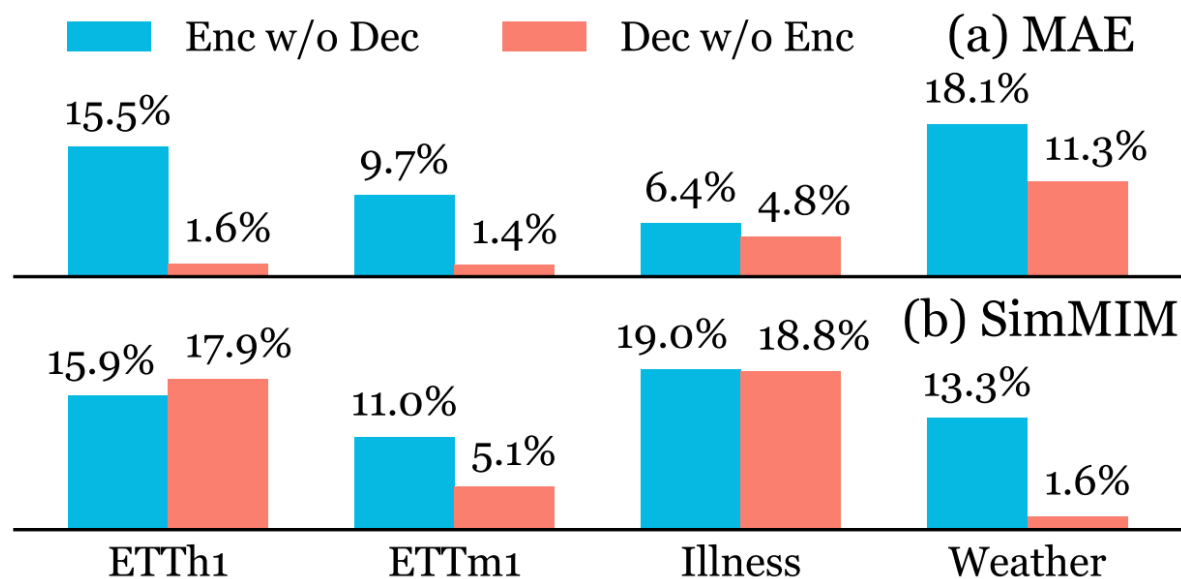


- 💡 LVMs with *global attention* fit classification
- 💡 LVMs that were *self-supervisedly pre-trained (masking)* fit forecasting

Are LVMs Useful for Time Series Analysis?

Insights¹⁴ – Why self-supervised LVMs are useful for **forecasting**?

Performance (MSE) Drop (%)



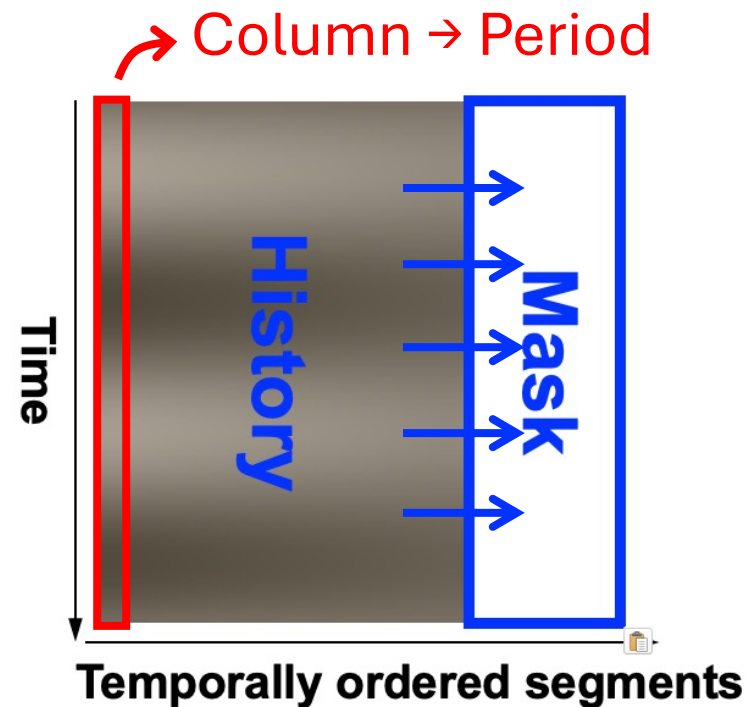
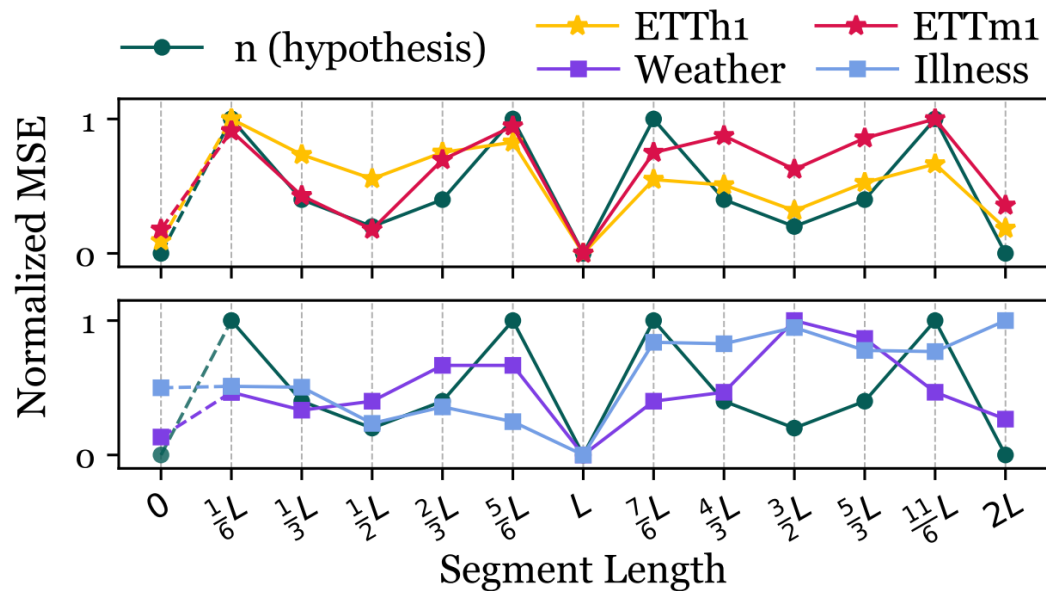
Decoder contributes more than Encoder

SimMIM's decoder: only 3.8% of all parameters

Are LVMs Useful for Time Series Analysis?

Insights¹⁴ – Limitation of self-supervised LVM forecasters

MSE change w. varying segment length



Performance is best when segment length equals period



UVH imaging leads to a bias toward forecasting periods

Outline of This Section

- ✓ **Generating MMVs of time series**
 - Linguistic view and visual view
- ✓ **Cross-modal knowledge transfer via MMVs**
 - Methods using LLMs and LVMs
- **Integrating MMVs of time series**
 - Combining multiple models or using LMMs

Integrating MMVs of Time Series

Integrating **numerical**, **visual** views and **contexts** – TimeVLM¹⁵

□ Vision-Language Model (VLM)

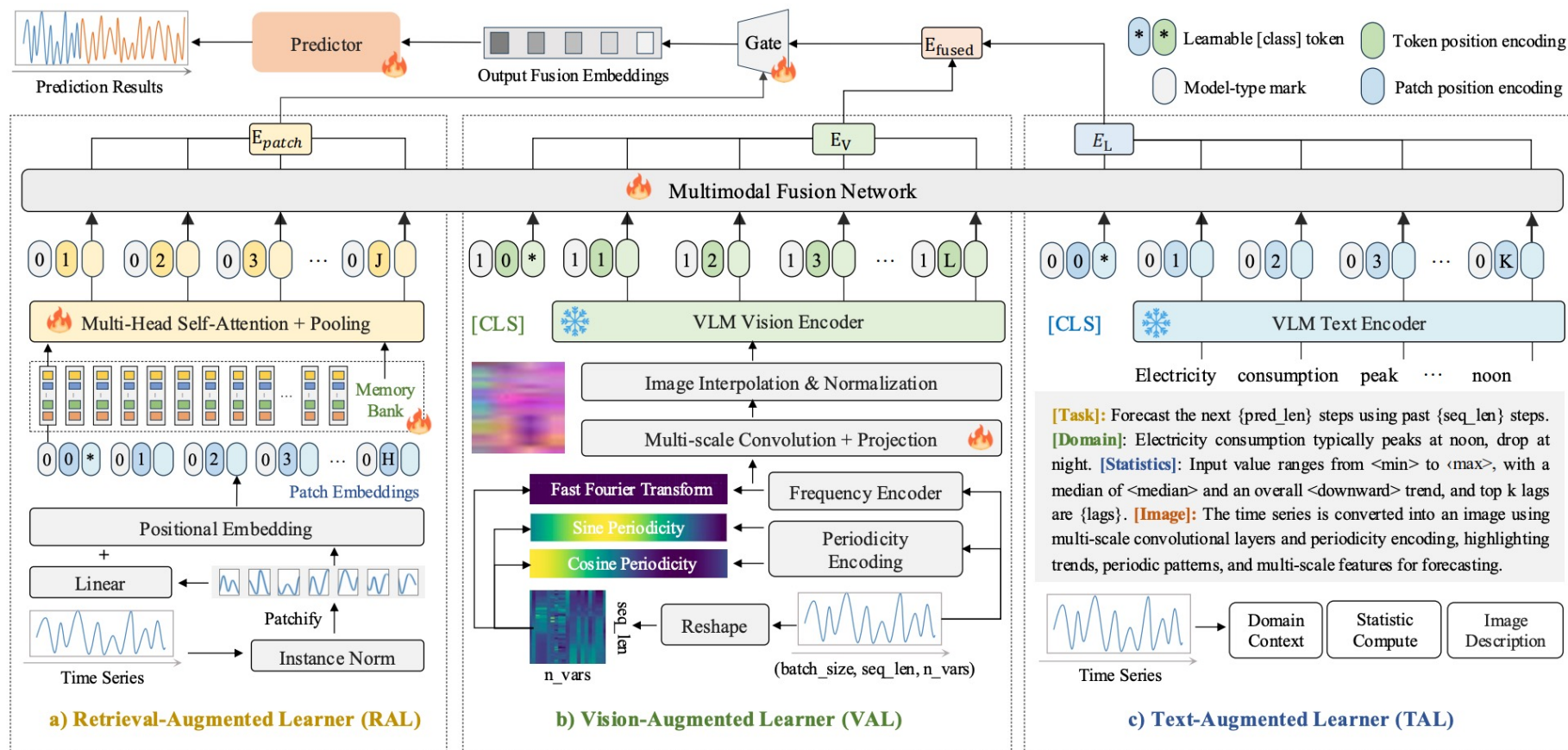
- ViLT

□ Imaging

- Frequency-periodicity encoding

□ Contexts

- Not a linguistic view



Integrating MMVs of Time Series

TimeVLM – Few-shot forecasting results

Methods	Time-VLM _{143M} (Ours)		Time-LLM _{3405M} (2024)		GPT4TS (2023)		DLinear (2023)		PatchTST (2023)		TimesNet (2023a)		FEDformer (2022)		Autoformer (2021)		Stationary (2022b)		ETSformer (2022)		LightTS (2022)		Informer (2021)		Reformer (2020)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
<i>ETTh1</i>	0.442	0.453	0.627	0.543	0.681	<u>0.560</u>	0.750	0.611	0.694	0.569	0.925	0.647	<u>0.658</u>	0.562	0.722	0.598	0.943	0.646	1.189	0.839	1.451	0.903	1.225	0.817	1.241	0.835
<i>ETTh2</i>	0.354	0.402	<u>0.382</u>	<u>0.418</u>	0.400	0.433	0.694	0.577	0.827	0.615	0.439	0.448	0.463	0.454	0.441	0.457	0.470	0.489	0.809	0.681	3.206	1.268	3.922	1.653	3.527	1.472
<i>ETTm1</i>	0.364	0.385	0.425	0.434	0.472	0.450	<u>0.400</u>	<u>0.417</u>	0.526	0.476	0.717	0.561	0.730	0.592	0.796	0.620	0.857	0.598	1.125	0.782	1.123	0.765	1.163	0.791	1.264	0.826
<i>ETTm2</i>	0.262	0.323	<u>0.274</u>	0.323	0.308	<u>0.346</u>	0.399	0.426	0.314	0.352	0.344	0.372	0.381	0.404	0.388	0.433	0.341	0.372	0.534	0.547	1.415	0.871	3.658	1.489	3.581	1.487
<i>Weather</i>	0.240	0.280	<u>0.260</u>	0.309	0.263	<u>0.301</u>	0.263	0.308	0.269	0.303	0.298	0.318	0.309	0.353	0.310	0.353	0.327	0.328	0.333	0.371	0.305	0.345	0.584	0.527	0.447	0.453
<i>ECL</i>	0.218	0.315	<u>0.179</u>	0.268	0.178	<u>0.273</u>	0.176	0.275	0.181	0.277	0.402	0.453	0.266	0.353	0.346	0.404	0.627	0.603	0.800	0.685	0.878	0.725	1.281	0.929	1.289	0.904
<i>Traffic</i>	0.558	0.410	<u>0.423</u>	<u>0.298</u>	0.434	0.305	0.450	0.317	0.418	0.296	0.867	0.493	0.676	0.423	0.833	0.502	1.526	0.839	1.859	0.927	1.557	0.795	1.591	0.832	1.618	0.851

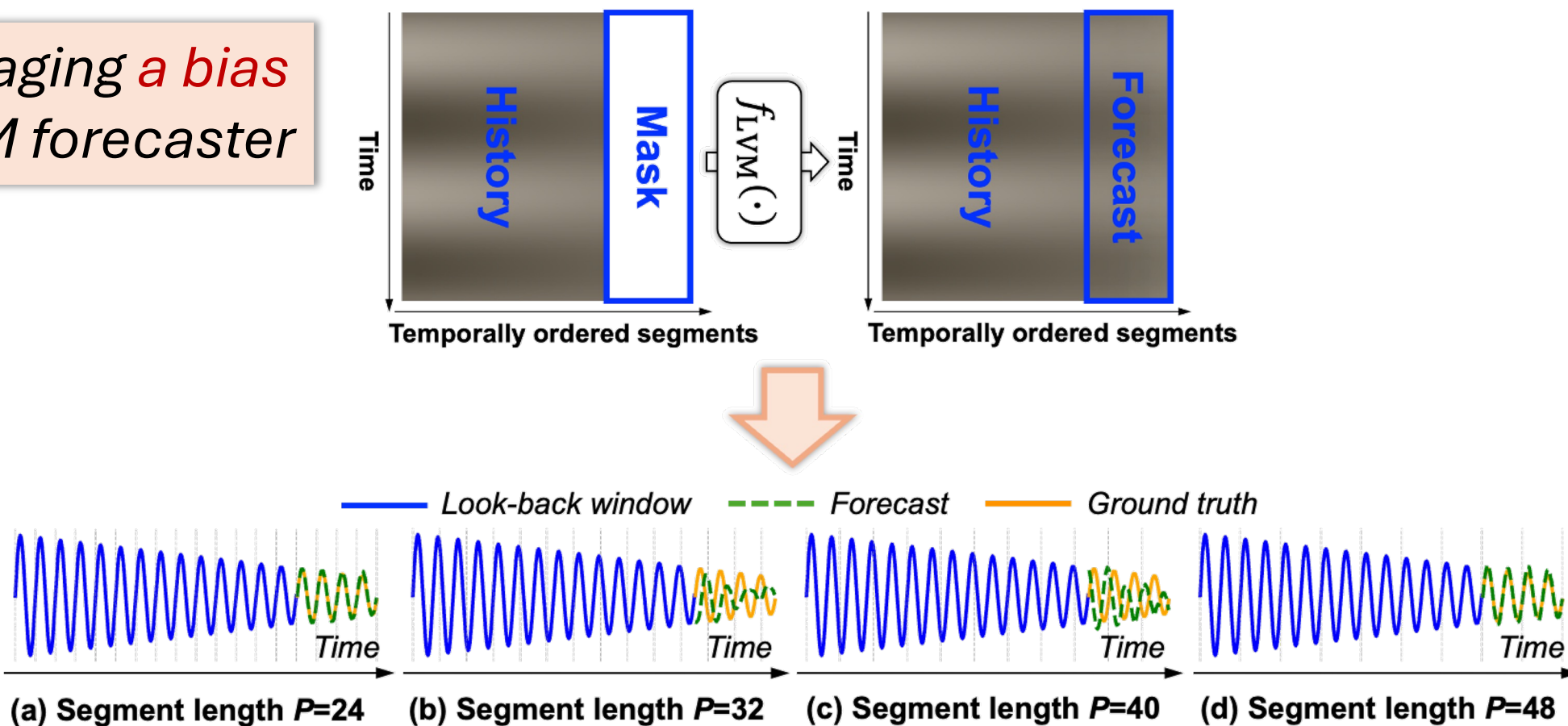
TimeVLM – Full long-term forecasting results

Methods	Time-VLM _{143M} (Ours)		Time-LLM _{3405M} (2024)		GPT4TS (2023)		DLinear (2023)		PatchTST (2023)		TimesNet (2023a)		FEDformer (2022)		Autoformer (2021)		Stationary (2022b)		ETSformer (2022)		LightTS (2022)		Informer (2021)		Reformer (2020)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
<i>ETTh1</i>	0.405	0.420	<u>0.408</u>	<u>0.423</u>	0.465	0.455	0.422	0.437	0.413	0.430	0.458	0.450	0.440	0.460	0.496	0.487	0.570	0.537	0.542	0.510	0.491	0.479	1.040	0.795	1.029	0.805
<i>ETTh2</i>	0.341	0.391	<u>0.334</u>	<u>0.383</u>	0.381	0.412	0.431	0.446	0.330	0.379	0.414	0.427	0.437	0.449	0.450	0.459	0.526	0.516	0.439	0.452	0.602	0.543	4.431	1.729	6.736	2.191
<i>ETTm1</i>	<u>0.347</u>	<u>0.377</u>	0.329	0.372	0.388	0.403	0.357	0.378	0.351	0.380	0.400	0.406	0.448	0.452	0.588	0.517	0.481	0.456	0.429	0.425	0.435	0.437	0.961	0.734	0.799	0.671
<i>ETTm2</i>	0.248	0.311	<u>0.251</u>	<u>0.313</u>	0.284	0.339	0.267	0.333	0.255	0.315	0.291	0.333	0.305	0.349	0.327	0.371	0.306	0.347	0.293	0.342	0.409	0.436	1.410	0.810	1.479	0.915
<i>Weather</i>	0.224	<u>0.263</u>	<u>0.225</u>	0.257	0.237	0.270	0.248	0.300	0.225	0.264	0.259	0.287	0.309	0.360	0.338	0.382	0.288	0.314	0.271	0.334	0.261	0.312	0.634	0.548	0.803	0.656
<i>Electricity</i>	0.172	0.273	0.158	0.252	0.167	<u>0.263</u>	0.166	<u>0.263</u>	<u>0.161</u>	0.252	0.192	0.295	0.214	0.327	0.227	0.338	0.193	0.296	0.208	0.323	0.229	0.329	0.311	0.397	0.338	0.422
<i>Traffic</i>	0.419	0.303	0.388	<u>0.264</u>	0.414	0.294	0.433	0.295	<u>0.390</u>	0.263	0.620	0.336	0.610	0.376	0.628	0.379	0.624	0.340	0.621	0.396	0.622	0.392	0.764	0.416	0.741	0.422

Integrating MMVs of Time Series

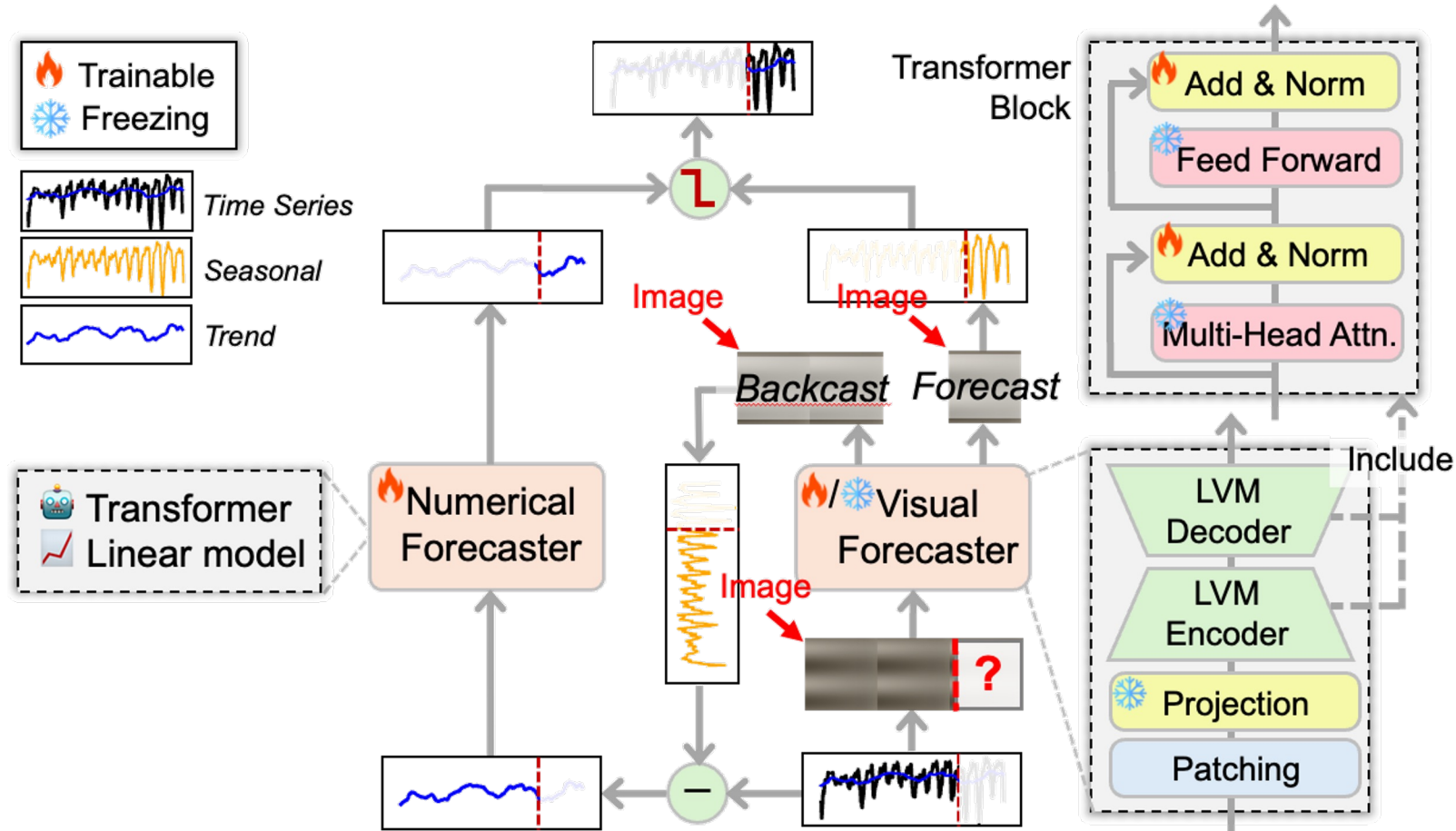
Integrating **numerical** and **visual** views – **DMMV**¹⁶

Leveraging *a bias*
of LVM forecaster



Integrating MMVs of Time Series

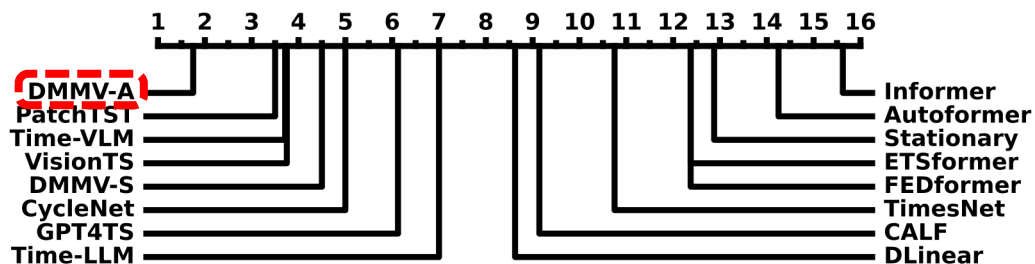
Integrating **numerical** and **visual** views – **DMMV**¹⁶



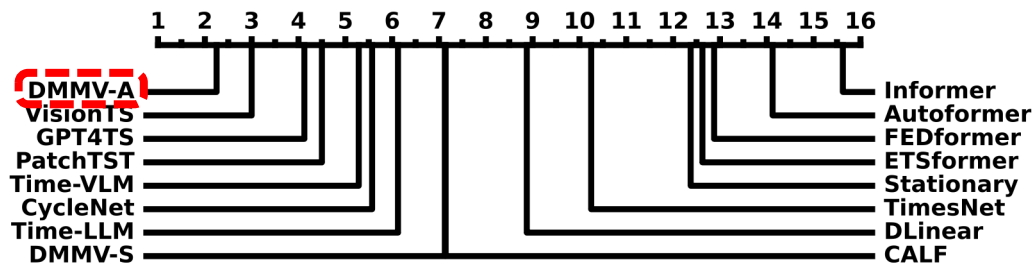
Integrating MMTVs of Time Series

DMMV¹⁶ – Long-Term Time Series Forecasting

(a) MSE Ranking



(b) MAE Ranking

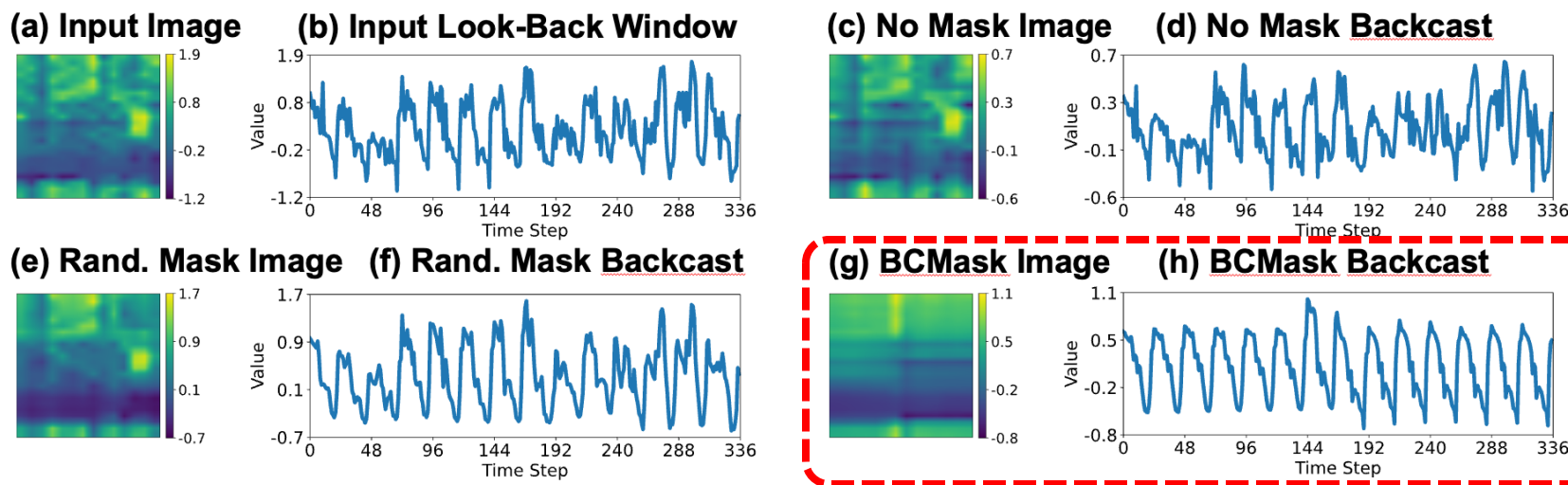


16. C. Shen et al. "Multi-Modal View Enhanced Large Vision Models for Long-Term Time Series Forecasting." arXiv preprint arXiv:2505.24003 (2025).

View	Multi-Modal				Visual		Language				Numerical										
Model	DMMV-A		Time-VLM		VisionTS		GPT4TS		Time-LLM		PatchTST		CycleNet		TimesNet		DLinear		FEDformer		
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	
ETTh1	96	0.354	0.389	0.361	0.386	0.355	0.386	0.370	0.389	0.376	0.402	0.370	0.399	0.374	0.396	0.384	0.402	0.375	0.399	0.376	0.419
	192	0.393	0.405	0.397	0.415	0.395	0.407	0.412	0.413	0.407	0.421	0.413	0.421	0.406	0.415	0.436	0.429	0.405	0.416	0.420	0.448
	336	0.387	0.413	0.420	0.421	0.419	0.421	0.448	0.431	0.430	0.438	0.422	0.436	0.431	0.430	0.491	0.469	0.439	0.416	0.459	0.465
	720	0.445	0.450	0.441	0.458	0.458	0.460	0.441	0.449	0.457	0.468	0.447	0.466	0.450	0.464	0.521	0.500	0.472	0.490	0.506	0.507
	Avg.	0.395	0.414	0.405	0.420	0.407	0.419	0.418	0.421	0.418	0.432	0.413	0.431	0.415	0.426	0.458	0.450	0.423	0.430	0.440	0.460
ETTh2	96	0.294	0.349	0.267	0.335	0.288	0.334	0.280	0.335	0.286	0.346	0.274	0.336	0.279	0.341	0.340	0.374	0.289	0.353	0.358	0.397
	192	0.339	0.395	0.326	0.373	0.349	0.380	0.348	0.380	0.361	0.391	0.339	0.379	0.342	0.385	0.402	0.414	0.383	0.418	0.429	0.439
	336	0.322	0.384	0.357	0.406	0.364	0.398	0.380	0.405	0.390	0.414	0.329	0.380	0.371	0.413	0.452	0.452	0.448	0.465	0.496	0.487
	720	0.392	0.425	0.412	0.449	0.403	0.431	0.406	0.436	0.405	0.434	0.379	0.422	0.426	0.451	0.462	0.468	0.605	0.551	0.463	0.474
	Avg.	0.337	0.388	0.341	0.391	0.351	0.386	0.354	0.389	0.361	0.396	0.330	0.379	0.355	0.398	0.414	0.427	0.431	0.447	0.437	0.449
ETTm1	96	0.279	0.329	0.304	0.346	0.284	0.332	0.300	0.340	0.291	0.341	0.290	0.342	0.299	0.348	0.338	0.375	0.299	0.343	0.379	0.419
	192	0.317	0.357	0.332	0.366	0.327	0.362	0.343	0.368	0.341	0.369	0.332	0.369	0.334	0.367	0.374	0.387	0.335	0.365	0.426	0.441
	336	0.351	0.381	0.364	0.383	0.354	0.382	0.376	0.386	0.359	0.379	0.366	0.392	0.368	0.386	0.410	0.411	0.369	0.386	0.445	0.459
	720	0.411	0.415	0.402	0.410	0.411	0.415	0.431	0.416	0.433	0.419	0.416	0.420	0.417	0.414	0.478	0.450	0.425	0.421	0.543	0.490
	Avg.	0.340	0.371	0.351	0.376	0.344	0.373	0.363	0.378	0.356	0.377	0.351	0.381	0.355	0.379	0.400	0.406	0.357	0.379	0.448	0.452
ETTm2	96	0.172	0.260	0.160	0.250	0.174	0.262	0.163	0.249	0.162	0.248	0.165	0.255	0.159	0.247	0.187	0.267	0.167	0.260	0.203	0.287
	192	0.227	0.298	0.215	0.291	0.228	0.297	0.222	0.291	0.235	0.304	0.220	0.292	0.214	0.286	0.249	0.309	0.224	0.303	0.269	0.328
	336	0.272	0.327	0.270	0.325	0.281	0.337	0.273	0.327	0.280	0.329	0.274	0.329	0.269	0.322	0.321	0.351	0.281	0.342	0.325	0.366
	720	0.351	0.381	0.348	0.378	0.384	0.410	0.357	0.376	0.366	0.382	0.362	0.385	0.363	0.382	0.408	0.403	0.397	0.421	0.421	0.415
	Avg.	0.256	0.317	0.248	0.311	0.267	0.327	0.254	0.311	0.261	0.316	0.255	0.315	0.251	0.309	0.291	0.333	0.267	0.332	0.305	0.349
Illness	24	1.409	0.754	-	-	1.613	0.834	1.869	0.823	1.792	0.807	1.319	0.754	2.255	1.017	2.317	0.934	2.215	1.081	3.228	1.260
	36	1.290	0.745	-	-	1.316	0.750	1.853	0.854	1.833	0.833	1.430	0.834	2.121	0.950	1.972	0.920	1.963	0.963	2.679	1.080
	48	1.499	0.810	-	-	1.548	0.818	1.886	0.855	2.269	1.012	1.553	0.815	2.187	1.007	2.238	0.940	2.130	1.024	2.622	1.078
	60	1.428	0.773	-	-	1.450	0.783	1.877	0.877	2.177	0.925	1.470	0.788	2.185	0.997	2.027	0.928	2.368	1.096	2.857	1.157
	Avg.	1.407	0.771	-	-	1.482	0.796	1.871	0.852	2.018	0.894	1.443	0.798	2.187	0.992	2.139	0.931	2.169	1.041	2.847	1.144
Electricity	96	0.126	0.213	0.142	0.245	0.127	0.217	0.141	0.239	0.137	0.233	0.129	0.222	0.128	0.223	0.168	0.272	0.140	0.237	0.193	0.308
	192	0.145	0.237	0.157	0.260	0.148	0.237	0.158	0.253	0.152	0.247	0.157	0.240	0.144	0.237	0.184	0.289	0.153	0.249	0.201	0.315
	336	0.162	0.254	0.174	0.276	0.163	0.253	0.172	0.266	0.169	0.267	0.163	0.259	0.160	0.254	0.198	0.300	0.169	0.267	0.214	0.329
	720	0.197	0.286	0.214	0.308	0.199	0.293	0.207	0.293	0.200	0.290	0.197	0.290	0.198	0.287	0.220	0.320	0.203	0.301	0.246	0.355
	Avg.	0.158	0.248	0.172	0.272	0.159	0.250	0.170	0.263	0.165	0.259	0.162	0.253	0.158	0.250	0.193	0.295	0.166	0.264	0.214	0.327
Weather	96	0.143	0.195	0.148	0.200	0.146	0.191	0.148	0.188	0.155	0.199	0.149	0.198	0.167	0.221	0.172	0.220	0.176	0.237	0.217	0.296
	192	0.187	0.242	0.193	0.240	0.194	0.238	0.192	0.230	0.223	0.261	0.194	0.241	0.212	0.258	0.219	0.261	0.220	0.282	0.276	0.336
	336	0.237	0.273	0.243	0.281	0.243	0.275	0.246	0.273	0.251	0.279	0.245	0.282	0.260	0.293	0.280	0.306	0.265	0.319	0.339	0.380
	720	0.302	0.315	0.312	0.332	0.318	0.328	0.320	0.328	0.345	0.342	0.314	0.334	0.328	0.339	0.365	0.359	0.333	0.362	0.403	0.428
	Avg.	0.217	0.256	0.224	0.263	0.225	0.258	0.227	0.255	0.244	0.270	0.226	0.264	0.242	0.278	0.259	0.287	0.249	0.300	0.309	0.360
Traffic	96	0.344	0.237	0.393	0.290	0.346	0.232	0.396	0.264	0.392	0.267	0.360	0.249	0.397	0.278	0.593	0.321	0.410	0.282	0.587	0.366
	192	0.363	0.249	0.405	0.296	0.376	0.245	0.412	0.268	0.409	0.271	0.379	0.256	0.411	0.283	0.617	0.336	0.423	0.287	0.604	0.373
	336	0.387	0.256	0.420	0.305	0.389	0.252	0.421	0.273	0.434	0.296	0.392	0.264	0.424	0.289	0.629	0.336	0.436	0.296	0.621	0.383
	720	0.433	0.284	0.459	0.323	0.432	0.293	0.455	0.291	0.451	0.291	0.432	0.286	0.450	0.305	0.640	0.350	0.466	0.315	0.626	0.382
	Avg.	0.382	0.257	0.419	0.304	0.386	0.256	0.421	0.274	0.422	0.281	0.391	0.264	0.421	0.289	0.620	0.336	0.434	0.295	0.610	0.376
# Wins	43		9		9		7		1		9		11		0		0		0		

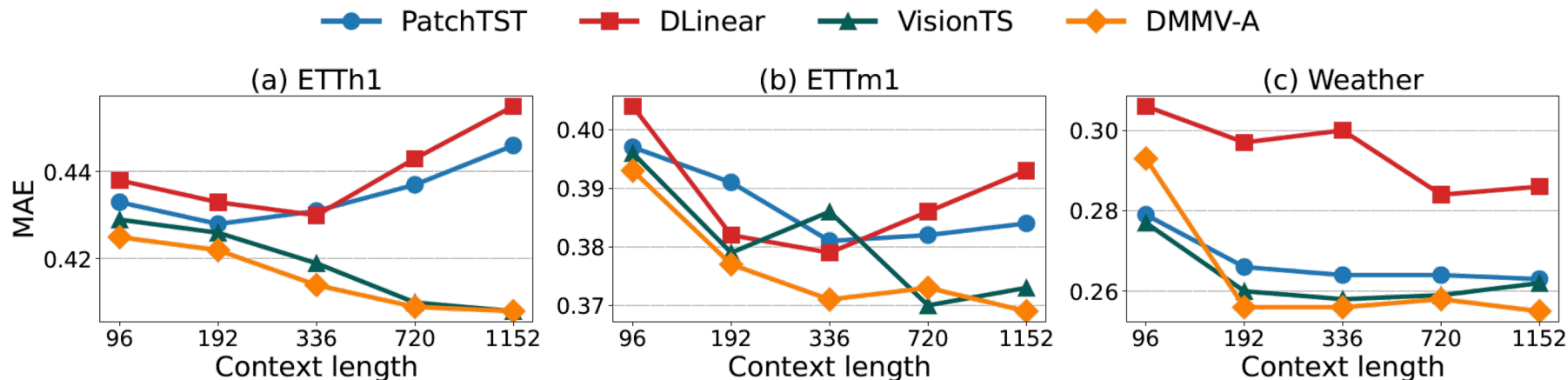
Integrating MMVs of Time Series

DMMV – Effective Extraction of Periodic Component



**Proposed
Method**

DMMV – Impact of Look-Back Window

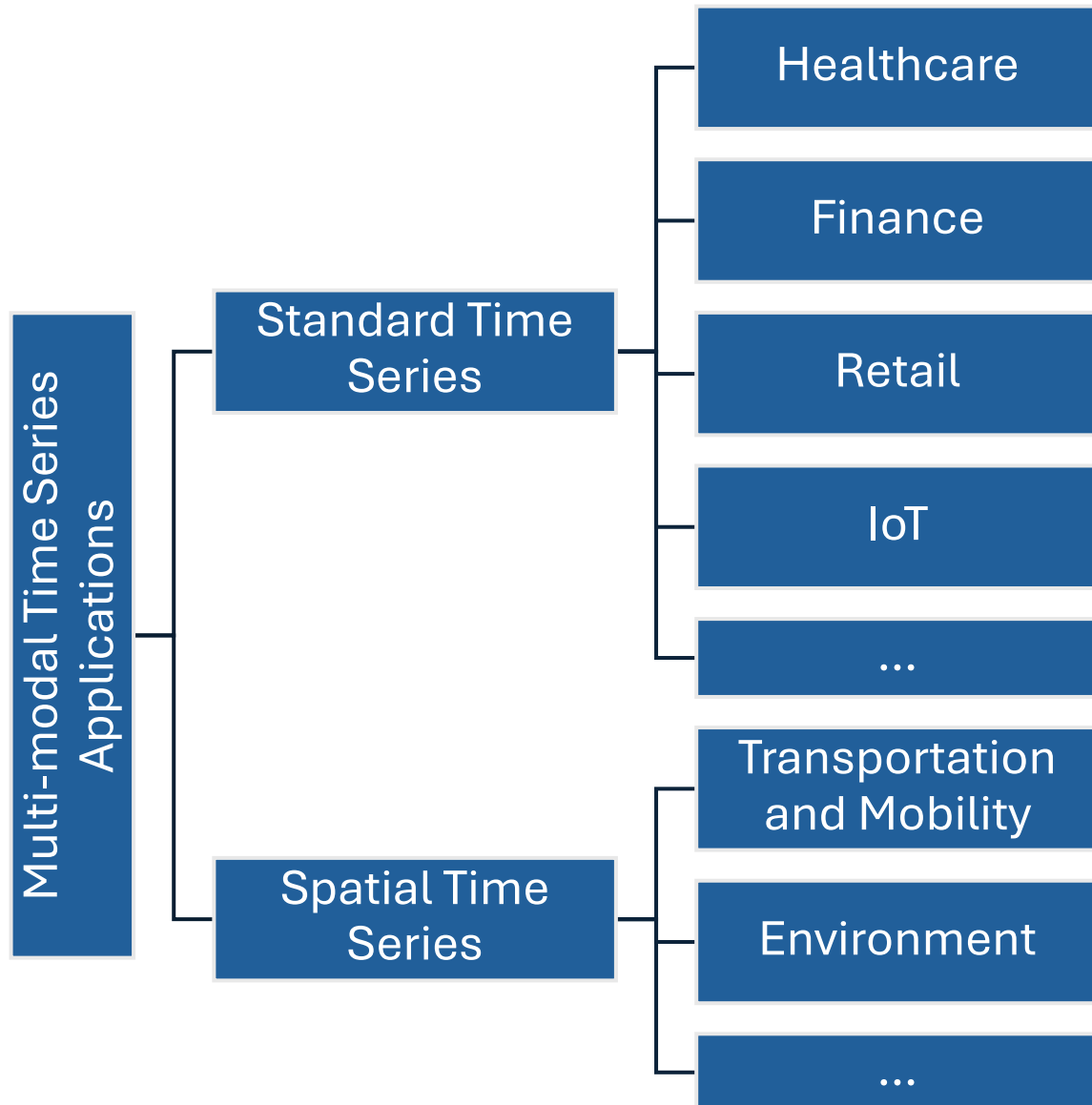


Outline of This Section

- ✓ **Generating MMVs of time series**
 - Linguistic view and visual view
- ✓ **Cross-modal knowledge transfer via MMVs**
 - Methods using LLMs and LVMs
- ✓ **Integrating MMVs of time series**
 - Combining multiple models or using LMMs

Multi-modal Time Series Application and Datasets

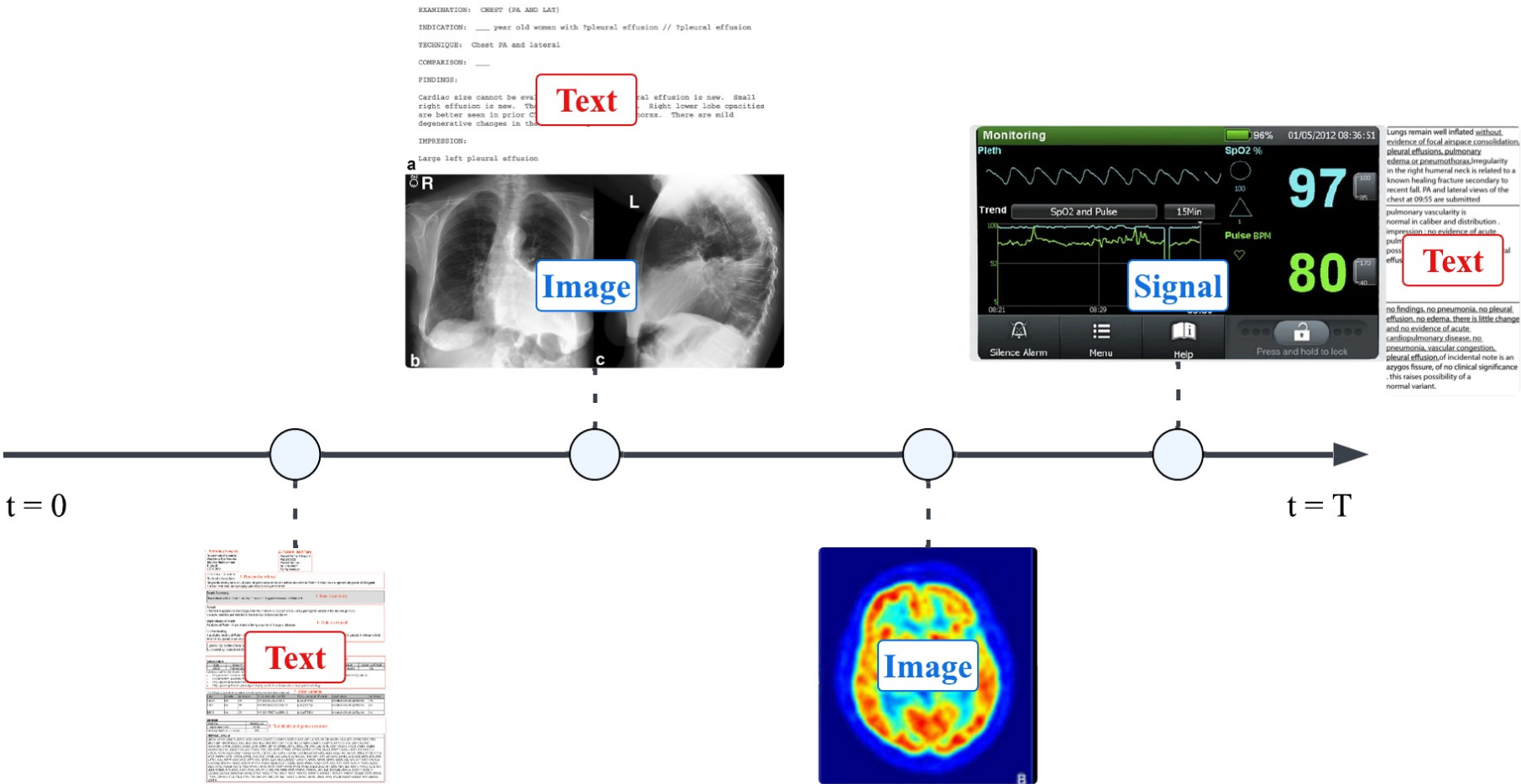
Multi-modal Time Series Applications



- 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, generation...*

Healthcare - EHR

- Electronic Health Records (EHR)



Healthcare - EHR

- **In-hospital Mortality Prediction**

- Predicting patient death during hospital stay

- **Readmission Risk Prediction**

- Forecasting the likelihood of patient re-hospitalization within 30 days

- **Clinical Event Forecasting**



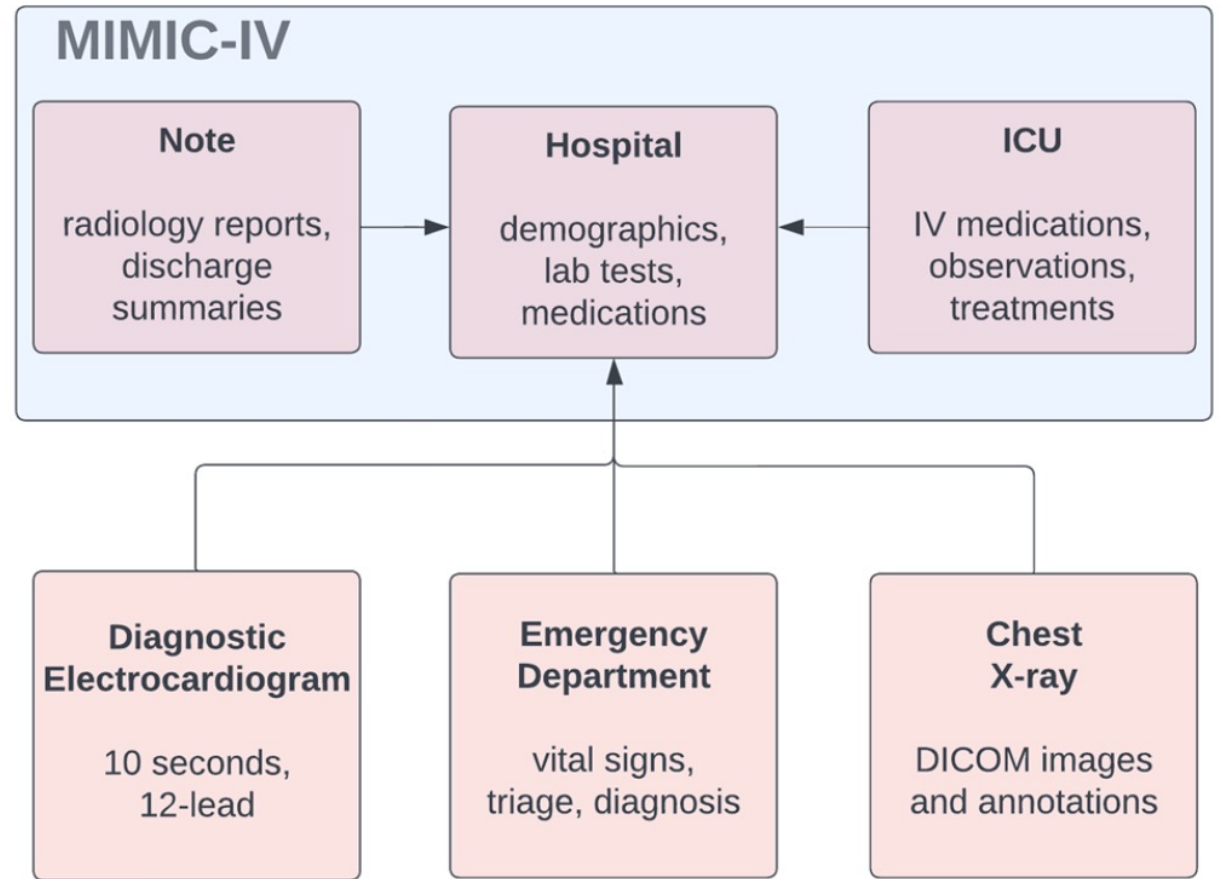
Healthcare - EHR Datasets

MIMIC-III & 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.

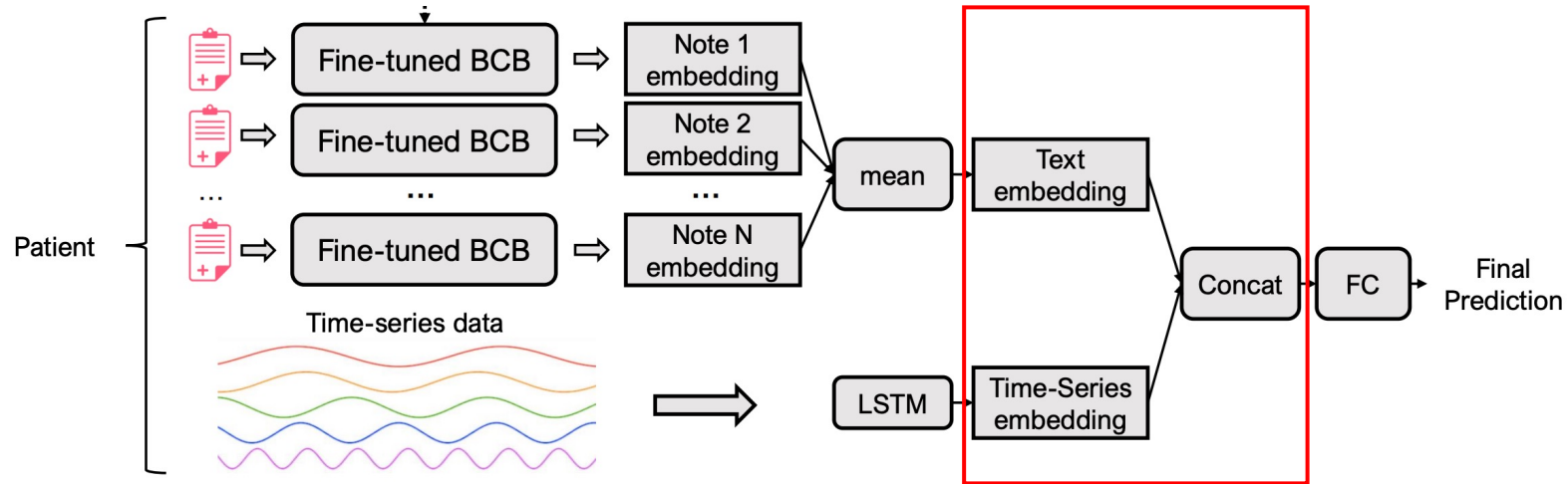
Healthcare – EHR Datasets

	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.

Healthcare - EHR Modeling

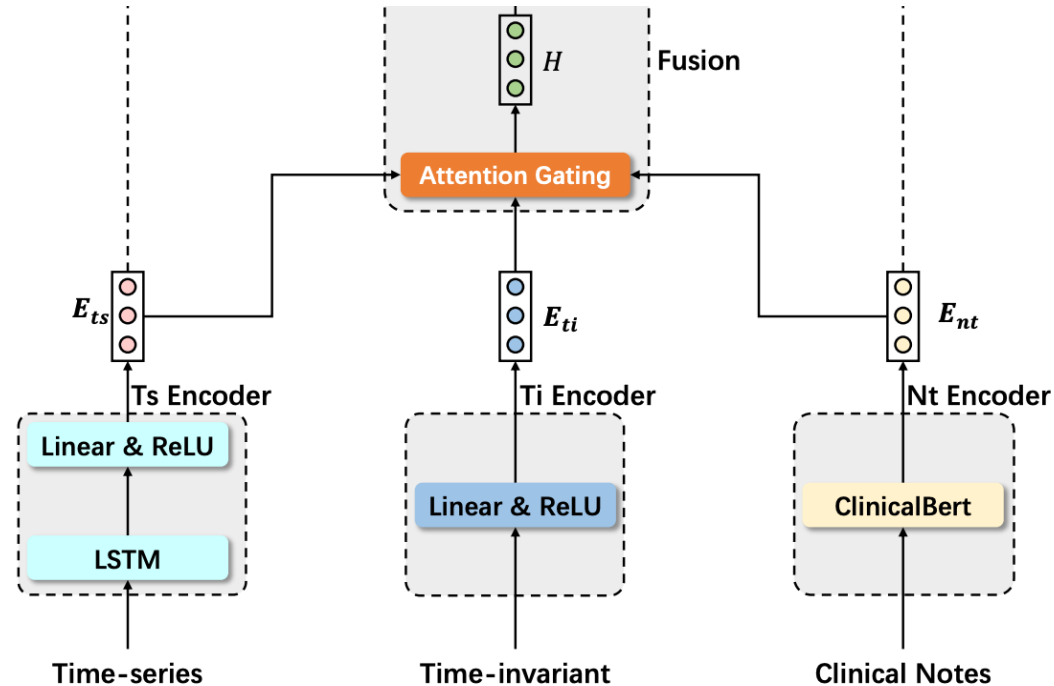
- Leverage multi-modality data lab values and clinical reports
 - Concatenation



Deznabi et al. "Predicting in-hospital mortality by combining clinical notes with time-series data", ACL Findings 2021

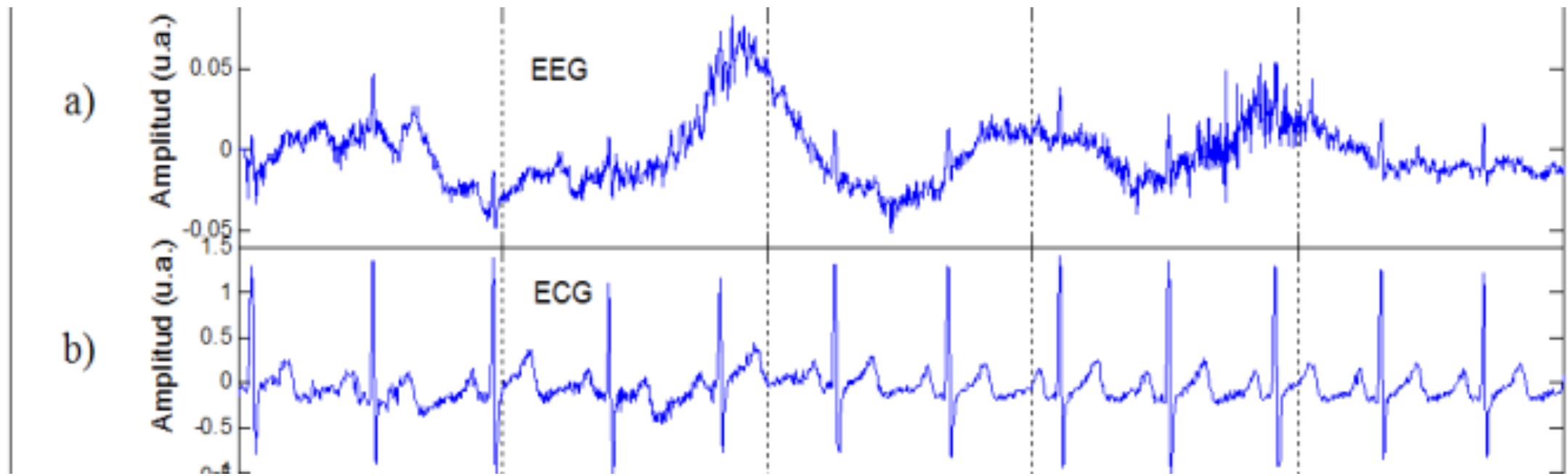
Healthcare – EHR Modeling

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



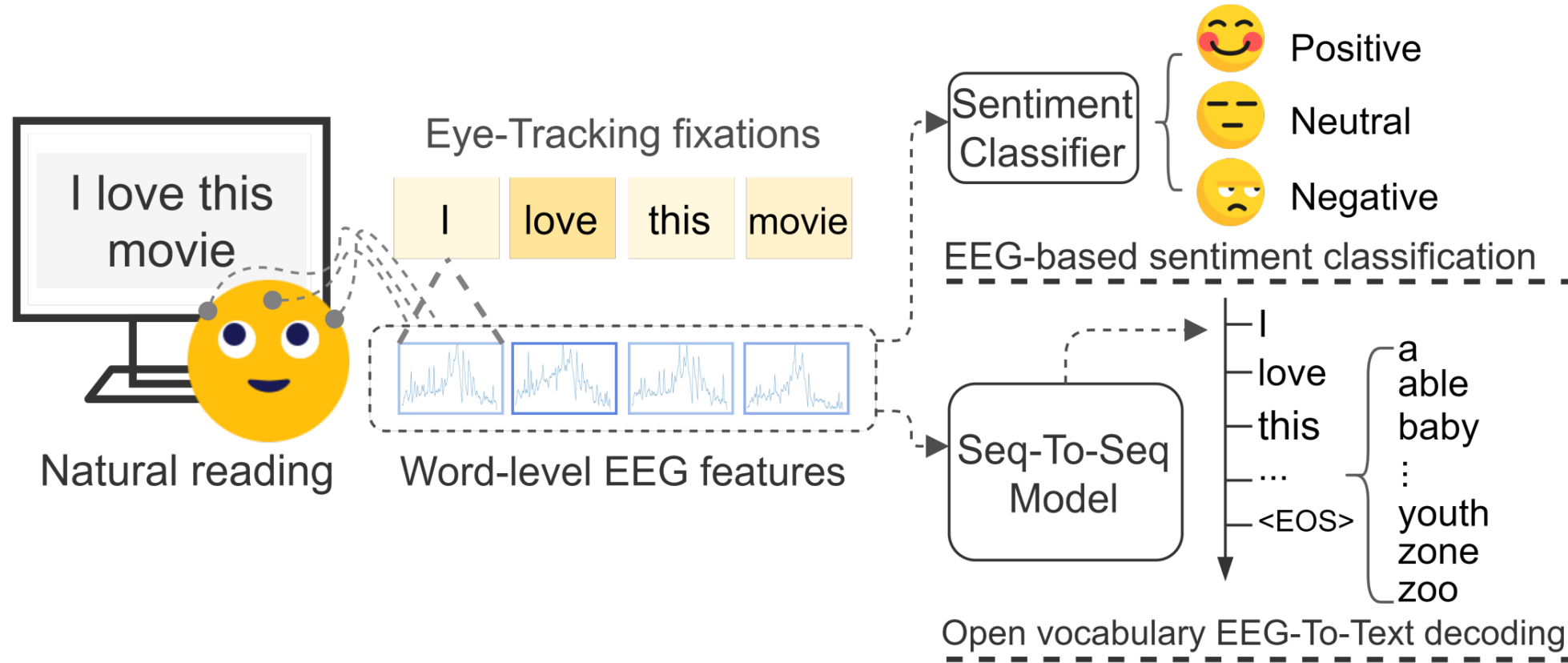
Yang et al. "How to leverage multimodal EHR data for better medical predictions", EMNLP 2021

Healthcare – ECG/EEG



Healthcare – ECG/EEG

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



Wang et al. "Open Vocabulary Electroencephalography-To-Text Decoding and Zero-shot Sentiment Classification", AAAI 2022

Healthcare – ECG/EEG

- Text decoding

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

Wang et al. "Open Vocabulary Electroencephalography-To-Text Decoding and Zero-shot Sentiment Classification", AAAI 2022

Healthcare – ECG/EEG Datasets

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

Healthcare – ECG/EEG Datasets

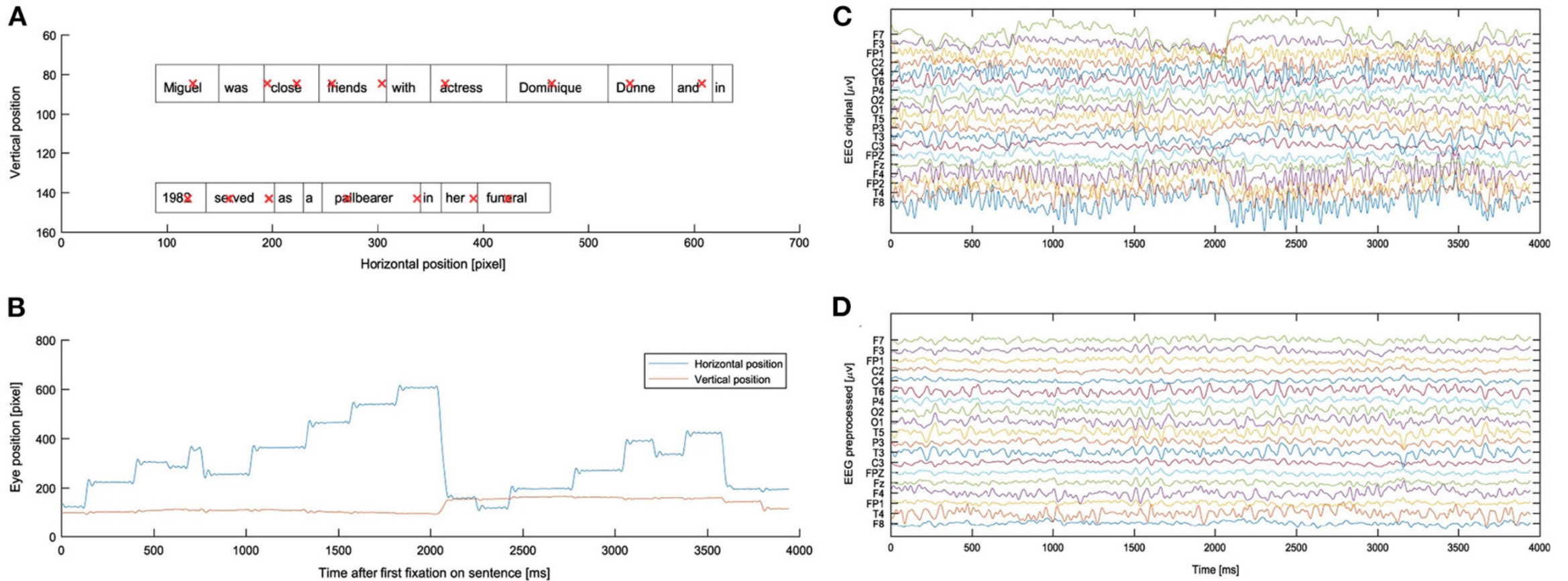
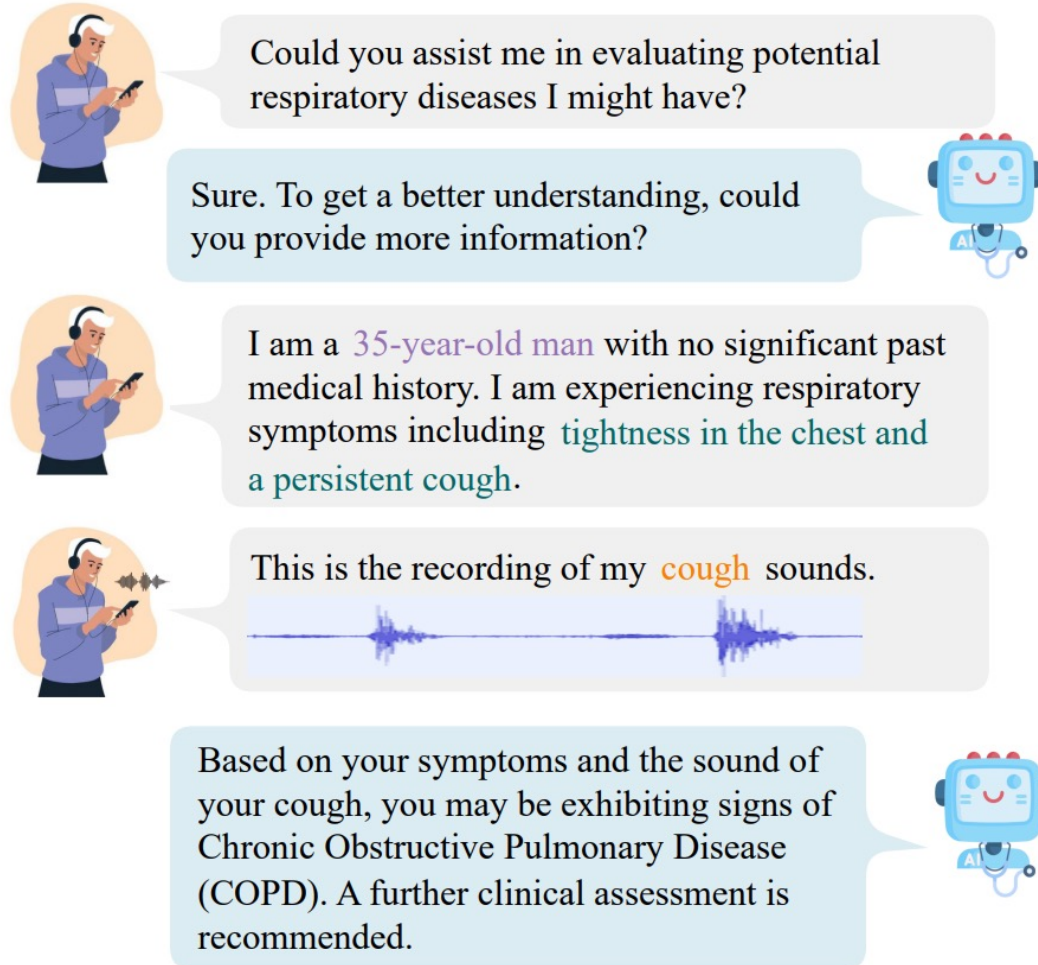


FIGURE 3


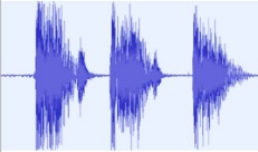



Visualization of eye-tracking and EEG data for a single sentence. **(A)** Prototypical sentence fixation data. Red crosses indicate fixations; boxes around the words indicate the wordbounds. **(B)** Fixation data plotted over time. **(C)** Raw EEG data during a single sentence. **(D)** Same data as in **(C)** after preprocessing.

Healthcare – Audio data

- Incorporating with audio data for respiratory health screen



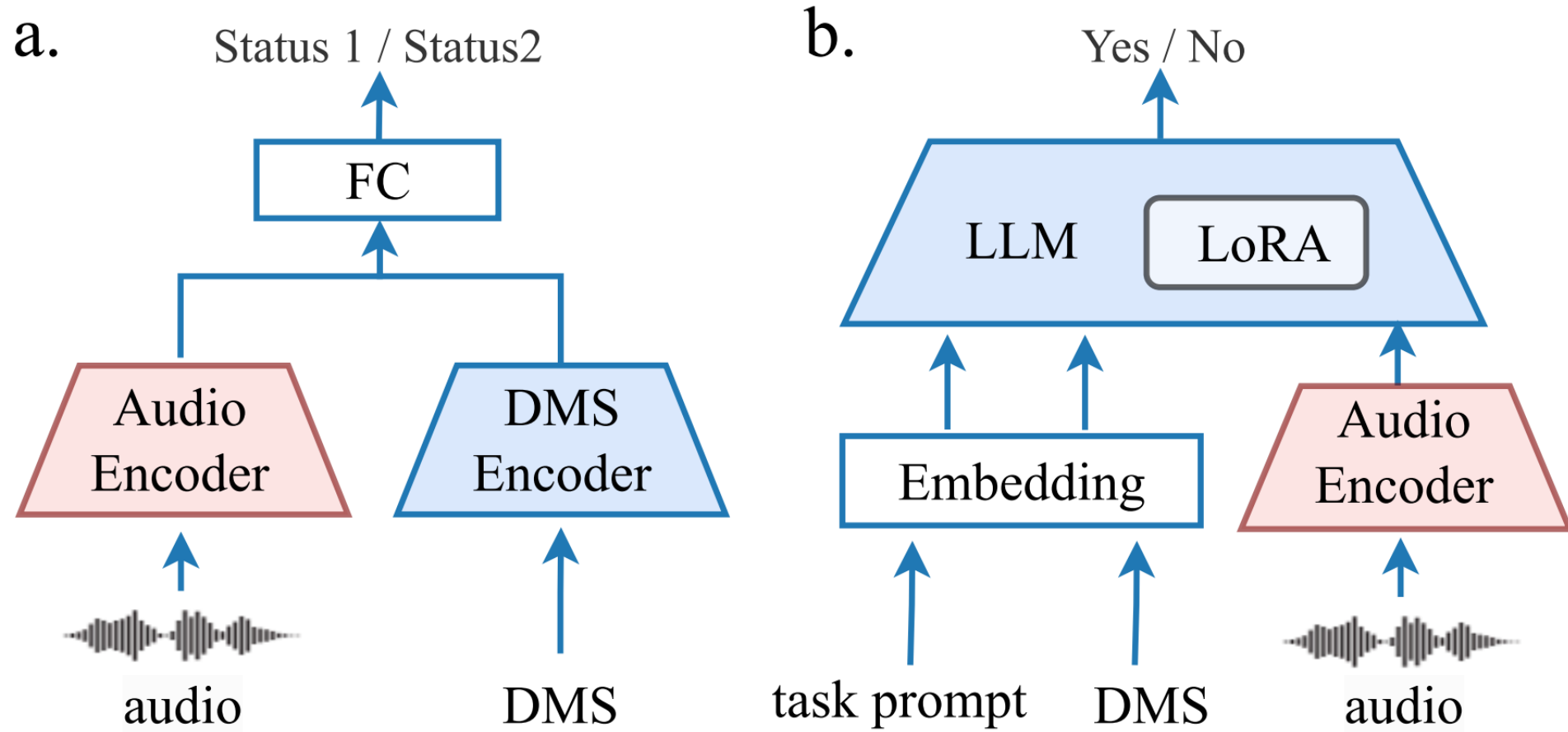
Healthcare – Audio data

Task	Text	Audio	Answer
S1 (Training)	<p>Task prompt: 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>DMS text: 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>Task prompt: 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>DMS text: Gender: Female. Age: 50-59. Patient presents with no medical history conditions. Patient presents with no obvious respiratory symptoms.</p>		0
S7 (Training)	<p>Task prompt: 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>DMS text: Gender: M. Age: 65. Record location: right posterior chest.</p>		1
T4 (Testing)	<p>Task prompt: 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>DMS text: Gender: male. Age: 35. Patient presents with the following respiratory symptoms: cold.</p>		0
T6 (Testing)	<p>Task prompt: 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>DMS text: Gender: F. Record location: posterior right upper.</p>		1

Zhang et al. RespLLM: Unifying Audio and Text with Multimodal LLMs for Generalized Respiratory Health Prediction, 2024

Healthcare – Audio data

- **Methods for respiratory health prediction**



(a) Concatenation-based fusion method.

(b) LLM-based fusion method.

Audio – 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

Audio - TS, Image

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

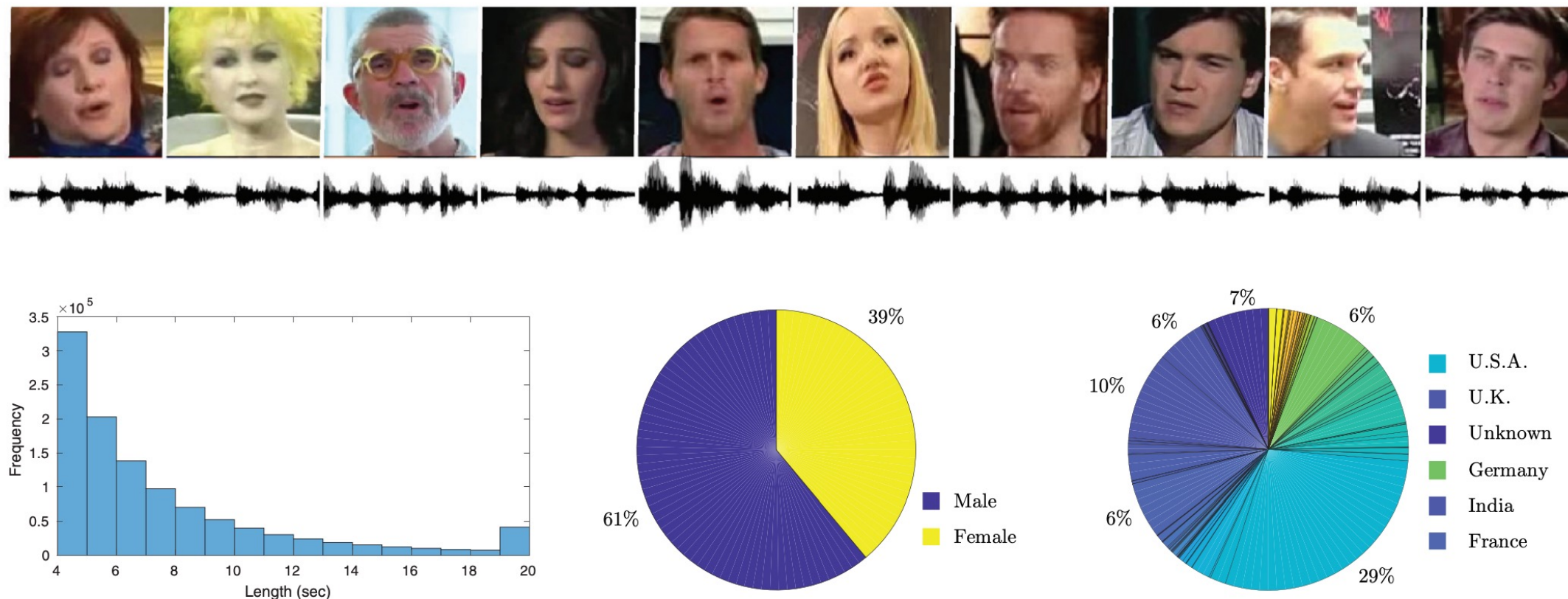
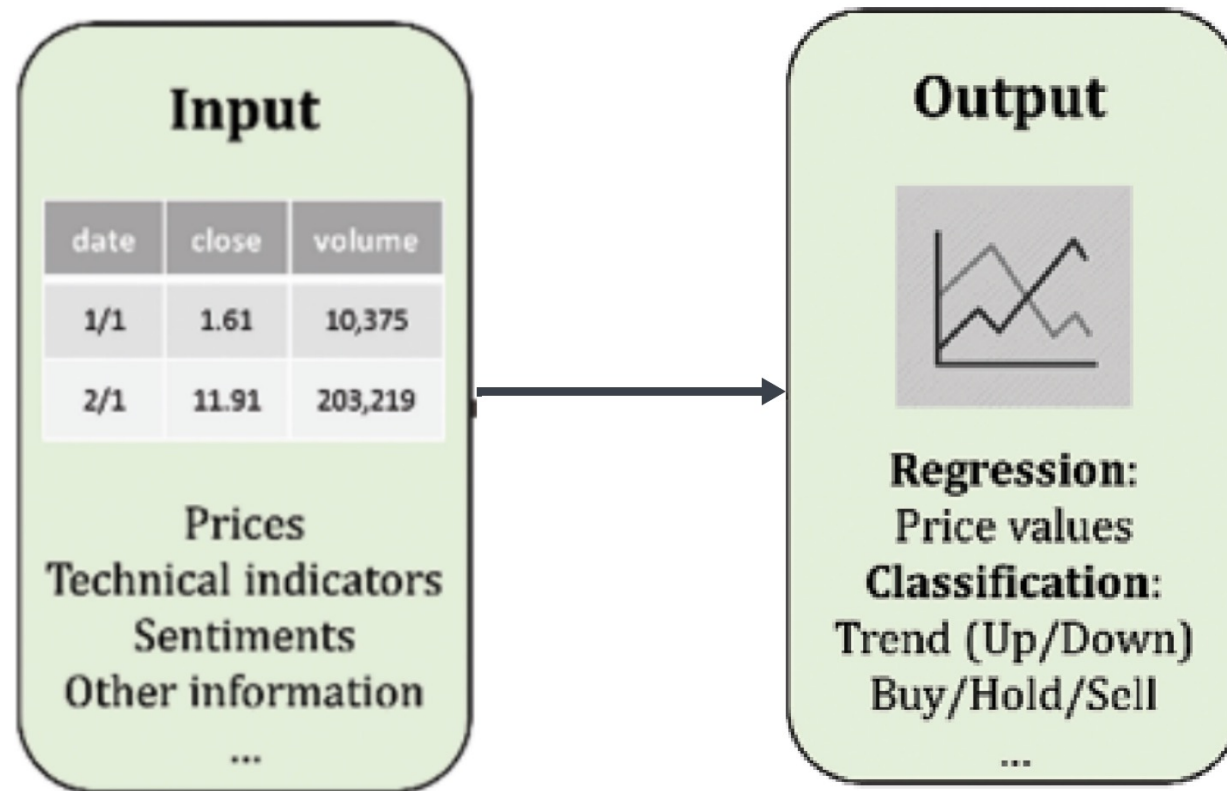


Fig. 1. *Top row:* Examples from the VoxCeleb2 dataset. We show cropped faces of some of the speakers in the dataset. Both audio and face detections are provided. *Bottom row:* (left) distribution of utterance lengths in the dataset – lengths shorter than 20s are binned in 1s intervals and all utterances 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.

Finance

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



Finance – TS&Text Dataset

- 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

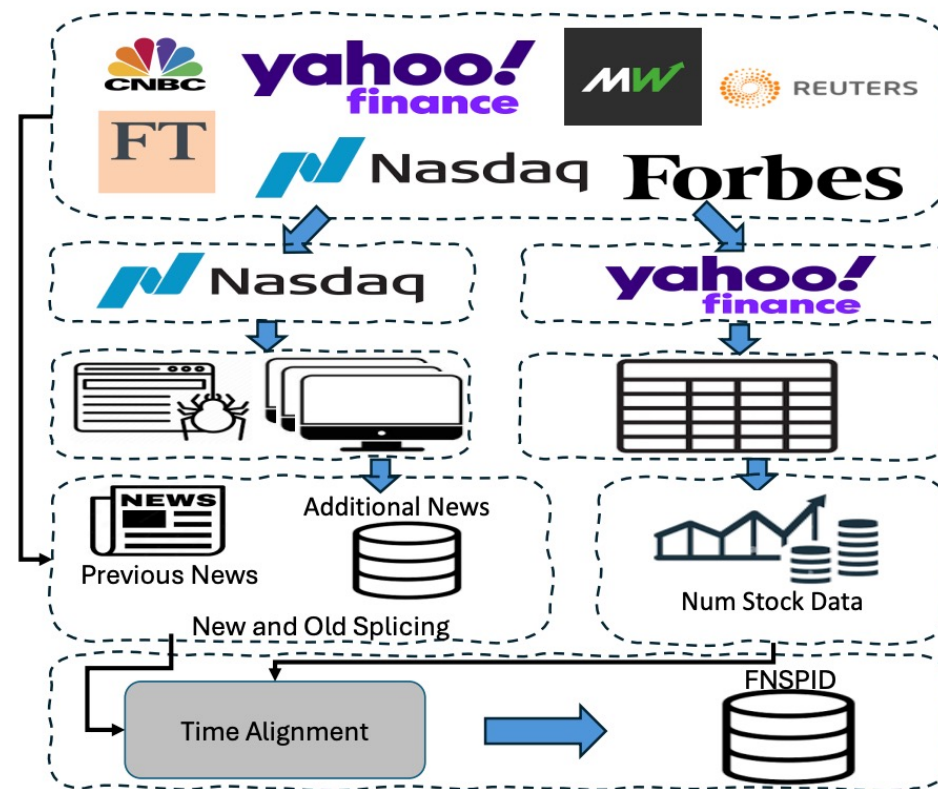


Figure 1: Data Collection Process from website selection in the first level box; data segmentation in second level boxes; data collection for web scraping on left and numerical data collection on right; data organization on fourth level boxes and final FNSPID build-up on the last level box.

Finance – TS&Text Dataset

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.

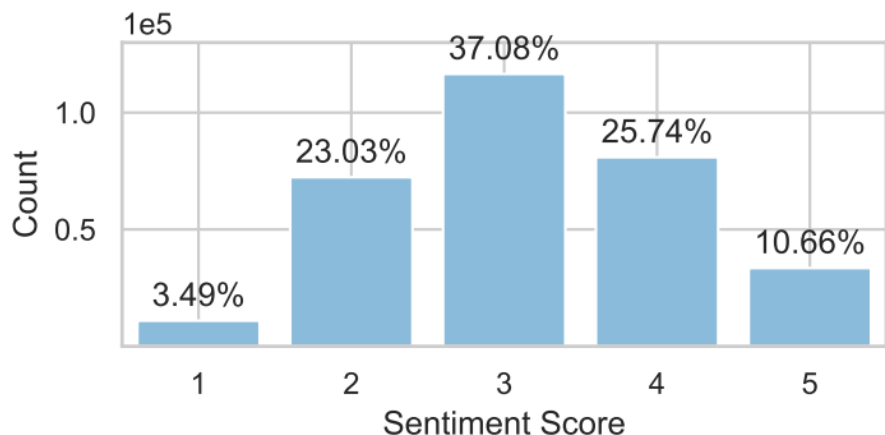


Figure 4: Sentiment Distribution: 1 is negative, 2 is somewhat negative, 3 is neutral, 4 is somewhat positive, 5 is positive

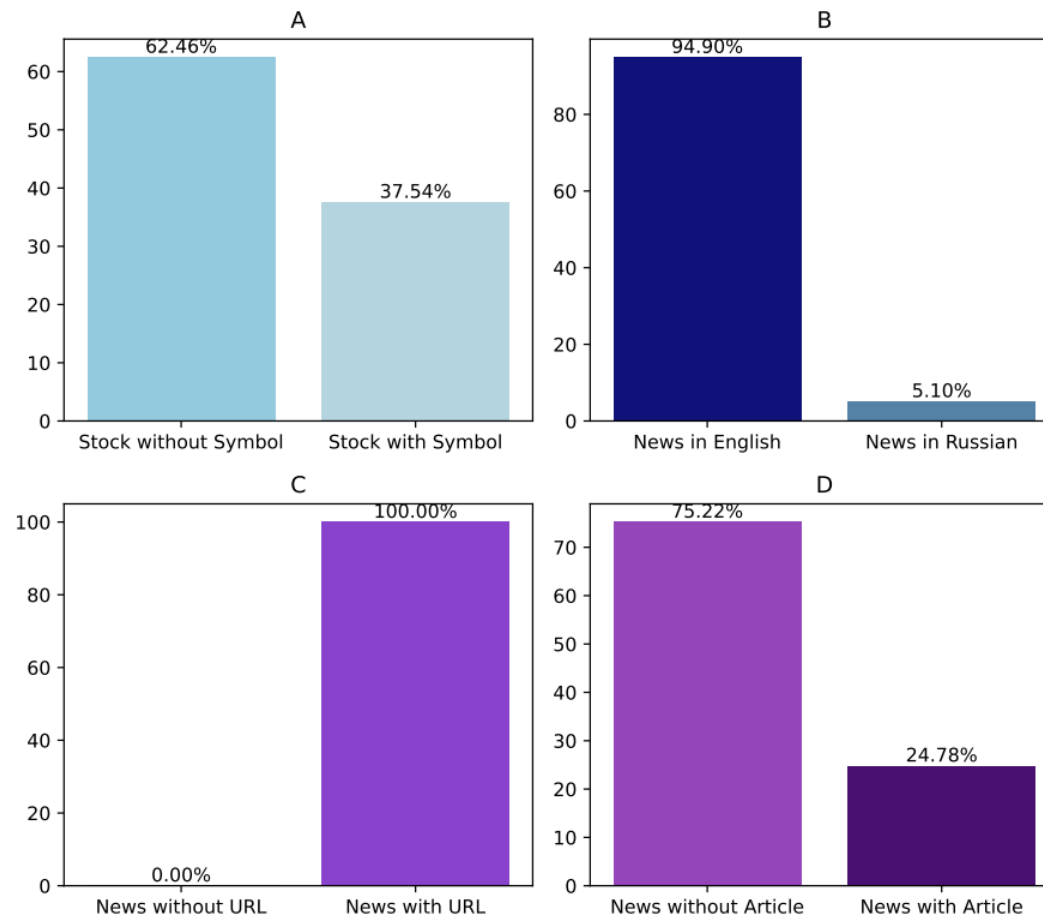


Figure 5: Statistical Overview: In A, we provide information on news articles that include the stock symbol. The B displays the language distribution, encompassing English and Russian. In C, a comparison of the included URLs is presented. Finally, in the D, details are provided on the news text already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

Finance – TS, Text, Image & Table Dataset

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
Total	599,846	3,853	599,846	77,702

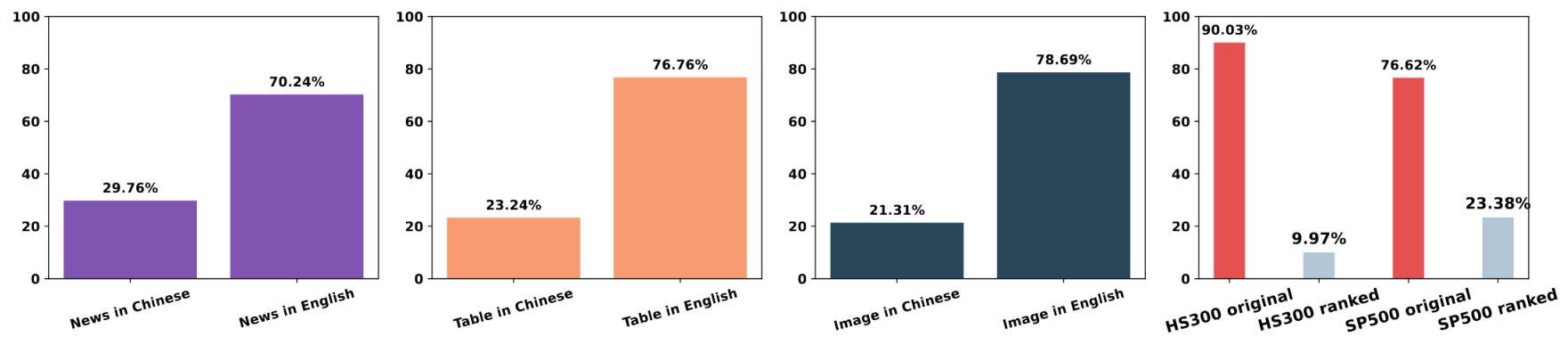
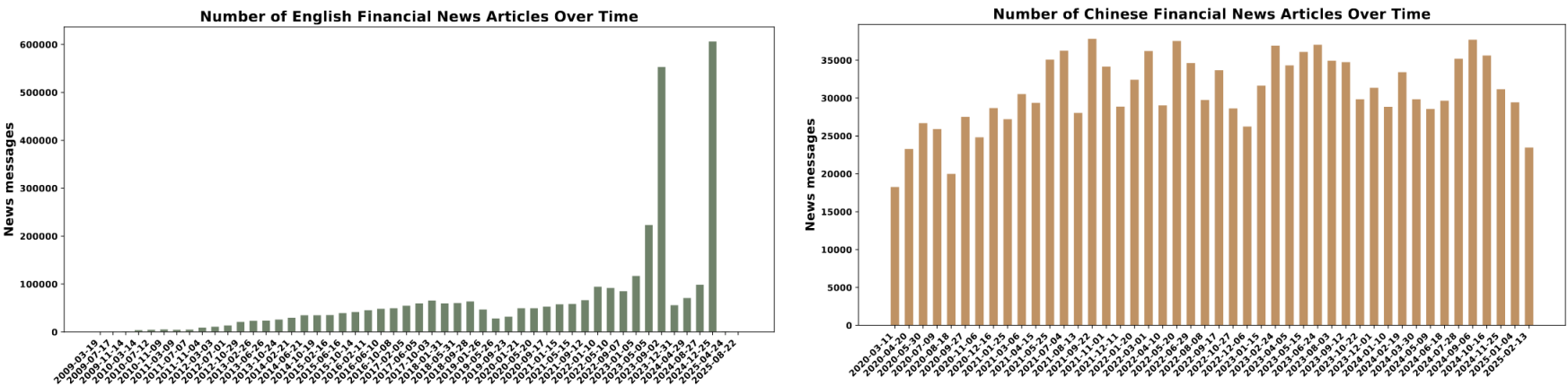


Figure 6: Proportions of Chinese vs. English Modalities (News, Tables, Images) and Coverage Ratios of Ranked vs. Original Daily News for HS300 and S&P 500.



Finance

- Combine financial time series and text.

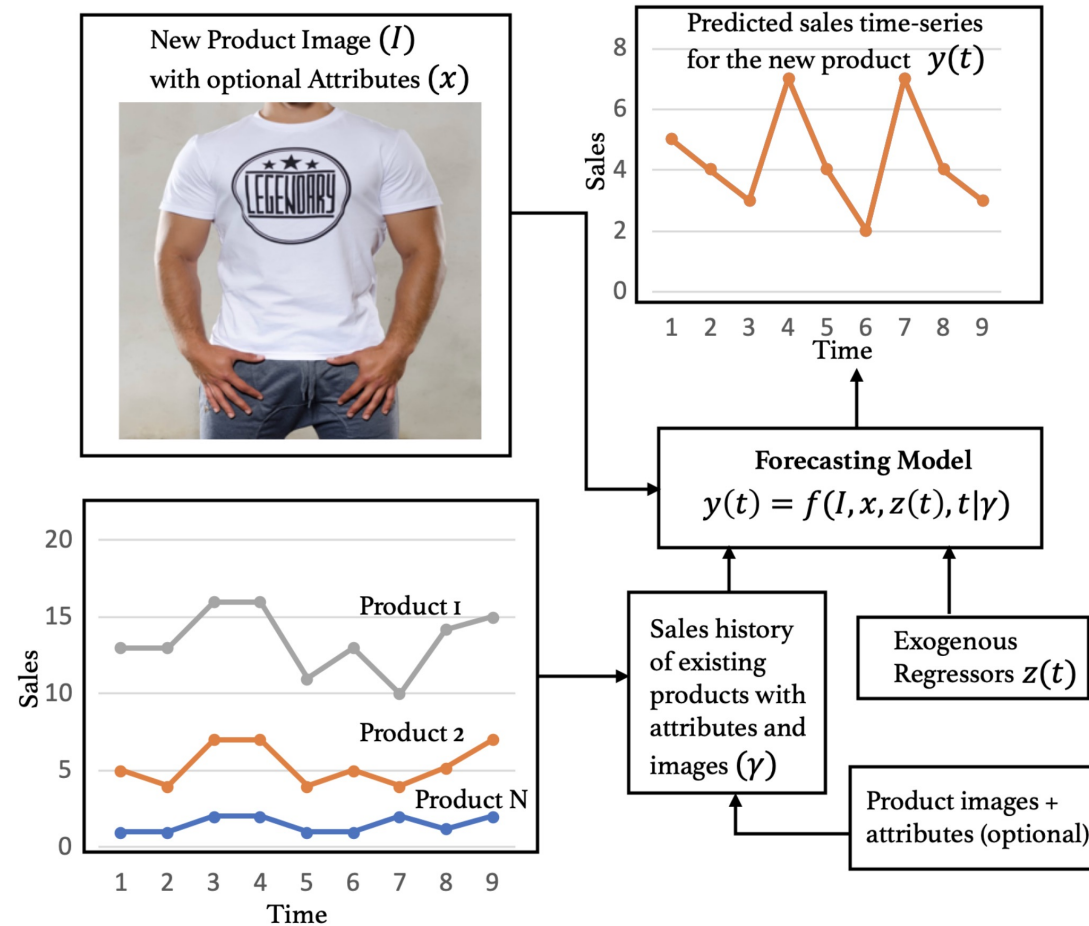
Type	Content
Prompt	data: date,open,high,low,close,adjusted-close,increase-in-5,10,15,20,25,30 2015-12-16,-0.45,0.78,-1.62,1.04,1.04,-1.63,-2.04,-2.52,-3.17,-3.53,-3.53 2015-12-17,-0.33,1.57,-0.49,0.33,0.33,-1.44,-2.01,-2.55,-3.38,-3.68,-3.70 2015-12-18,2.41,2.62,0.00,-2.85,-2.85,1.42,0.70,0.43,-0.30,-0.73,-0.87 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 #stocksgbsn 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\$zv \$aapl \$unh \$bib #vz #stock #stocksl \$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.

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

Xie et al. "The Wall Street Neophyte: A Zero-Shot Analysis of ChatGPT Over Multimodal Stock Movement Prediction Challenges", 2023

Retails

- Sales forecasting using product images and historical TS.



Retails – TS, Text & Image Datasets

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.

Retails – TS, Text & Image Datasets

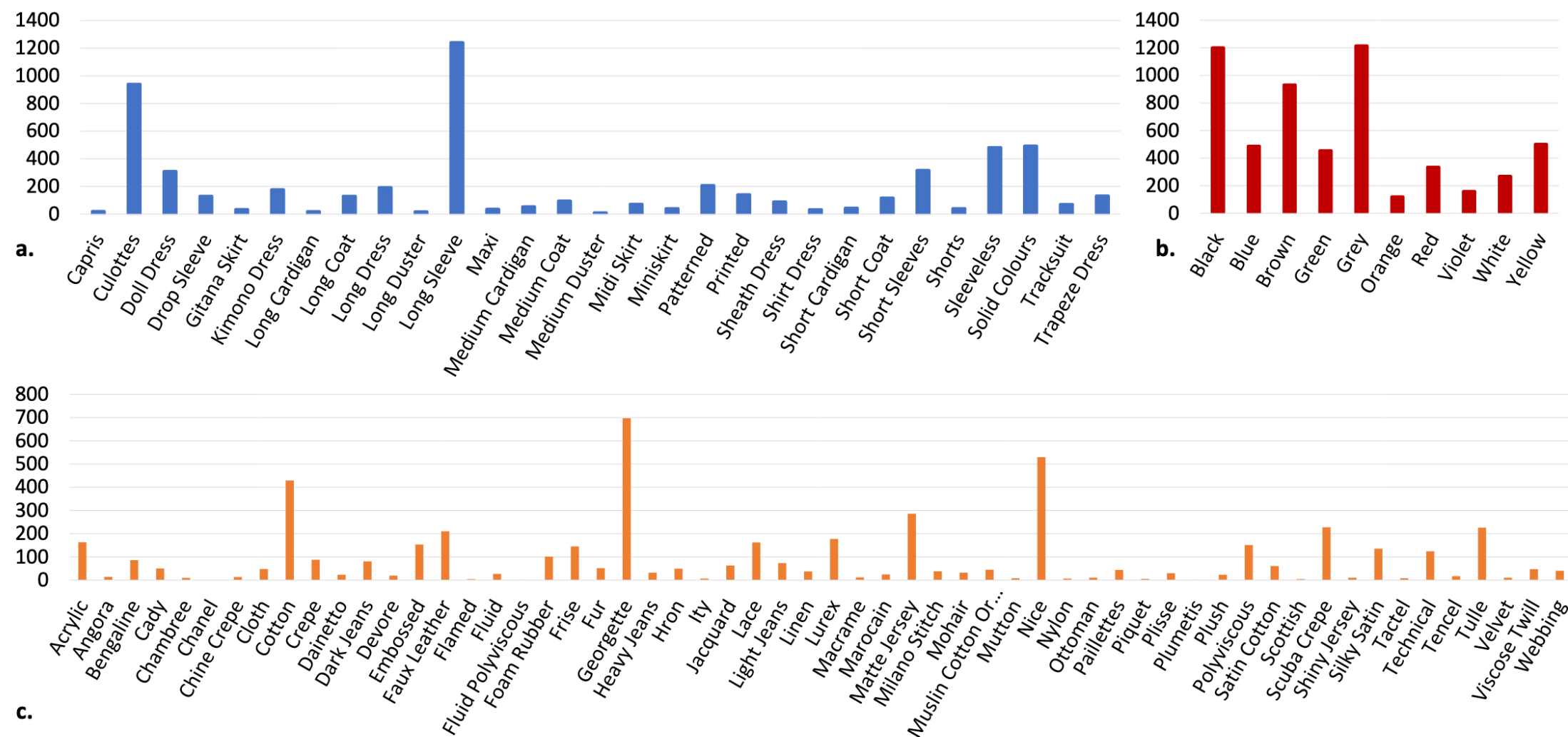
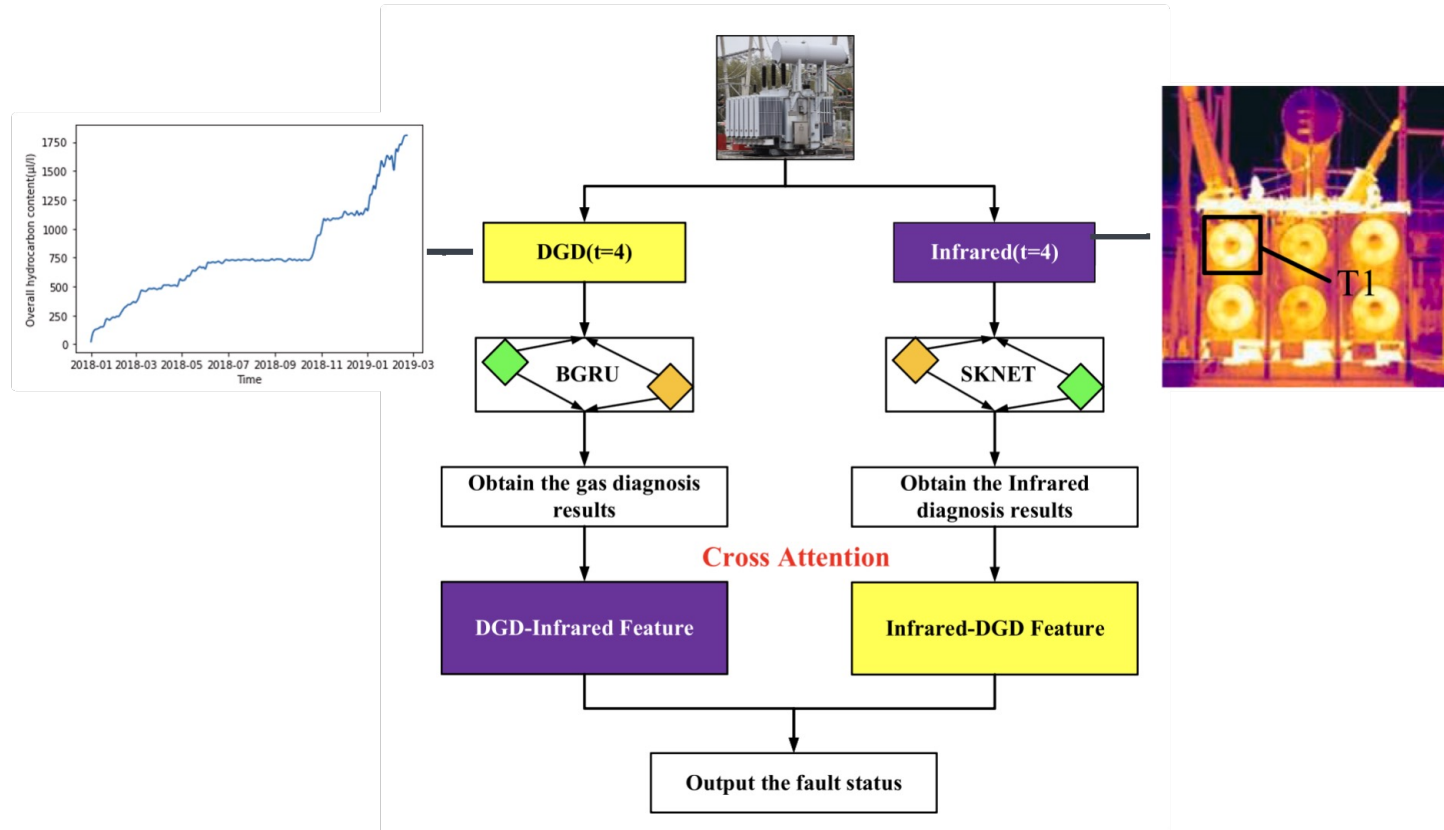


Figure 2: Cardinalities of the dataset for clothing categories (a), color (b) and fabric (c).

IoT

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



Xing et al. "Multi-modal information analysis for fault diagnosis with time-series data from power transformer", JEPE 2023

Multi-Domain

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

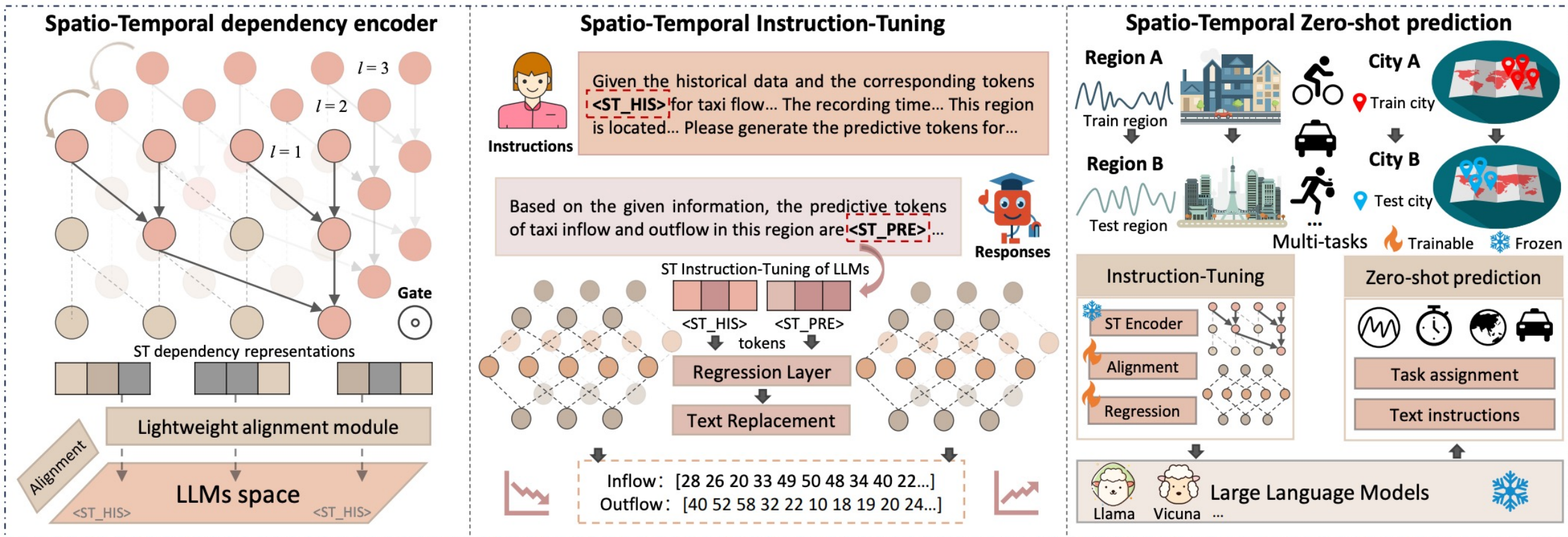
Multi-Domain

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.

Dataset	Median	Mean	Min	Max
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

Spatial Time Series – Transportation

- Prompting LLMs with structured traffic data for traffic prediction

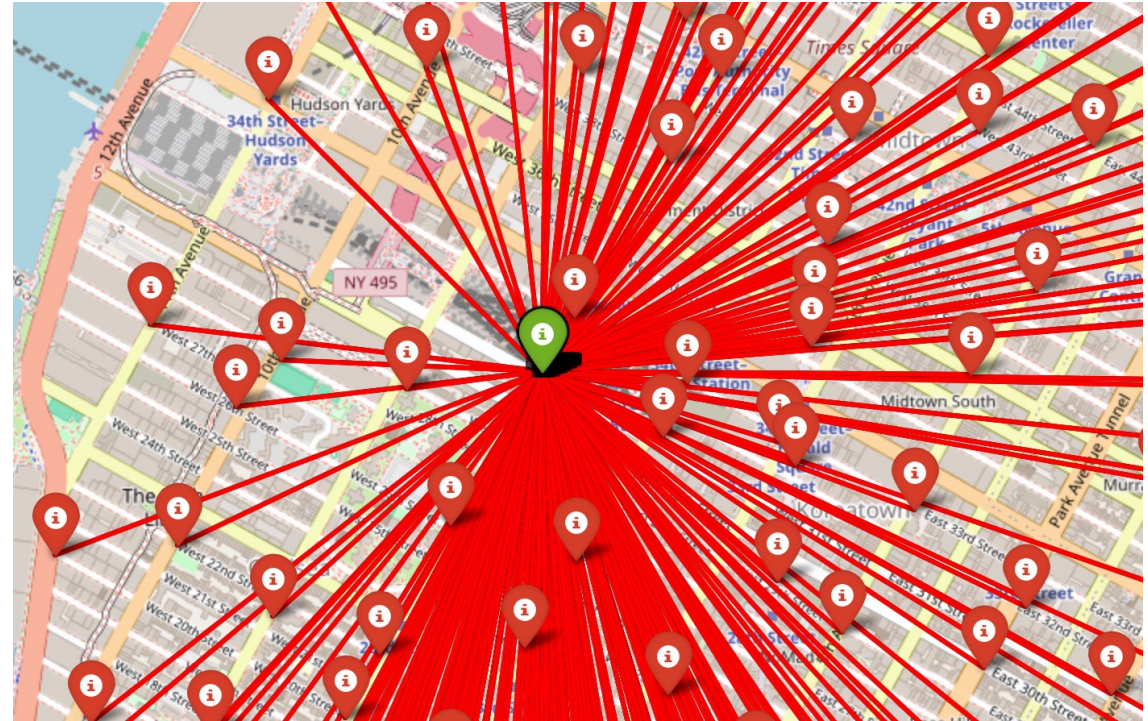


Spatial Time Series – Transportation – ST & Text Datasets

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

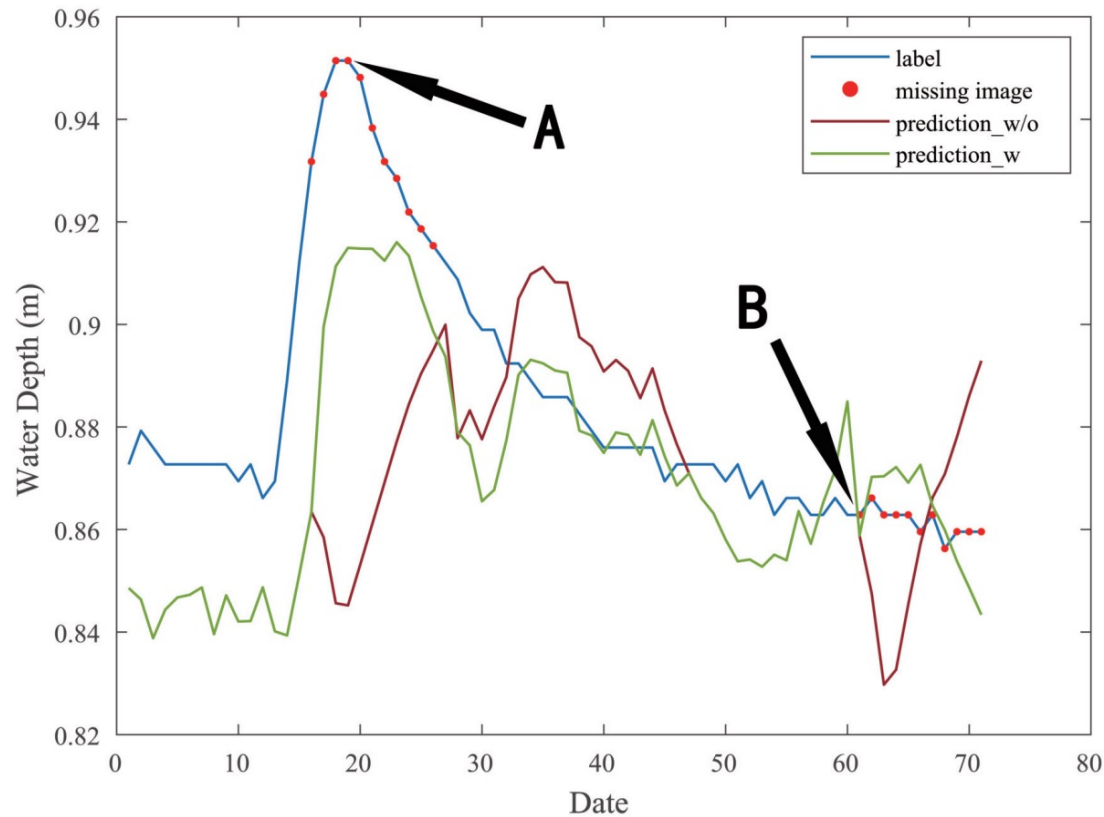
Text: Station-level static information such as ID, name, and capacity



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

Spatial Time Series – Environment

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



(a) Missing image at peak A.

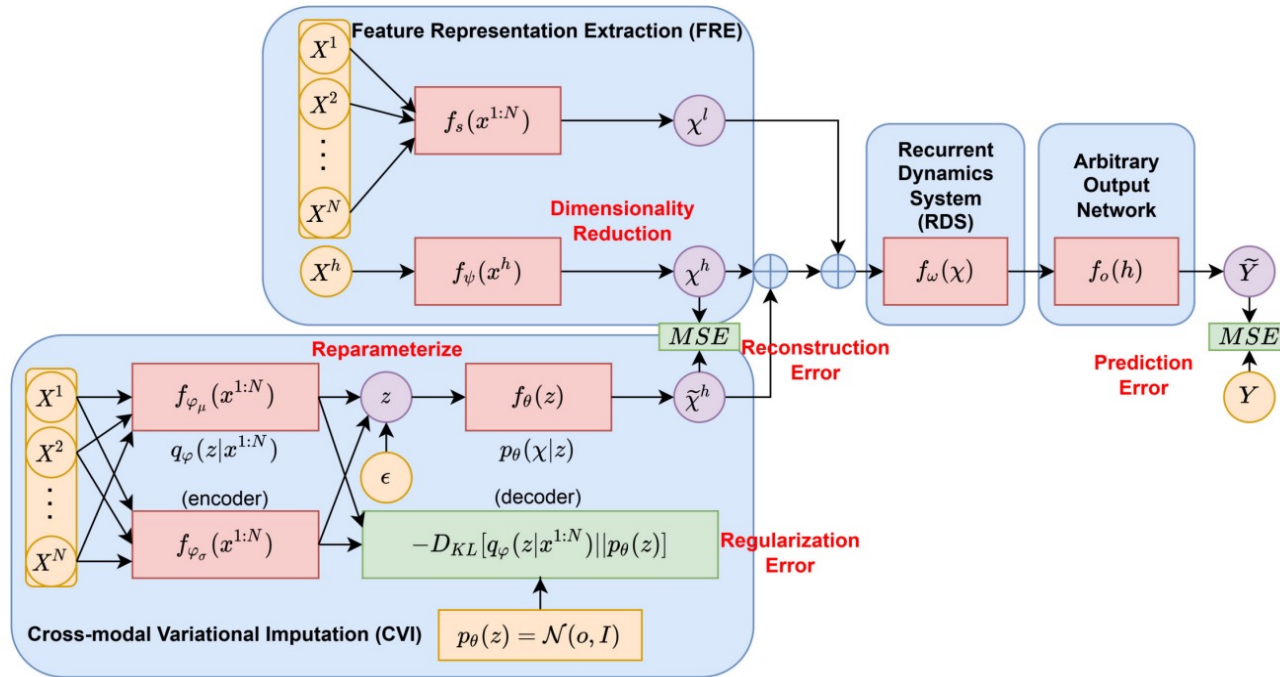


(b) Missing image at recession B.

Zhao et al. "VIMTS: Variational-based Imputation for Multi-modal Time Series", IEEE BigData 2022

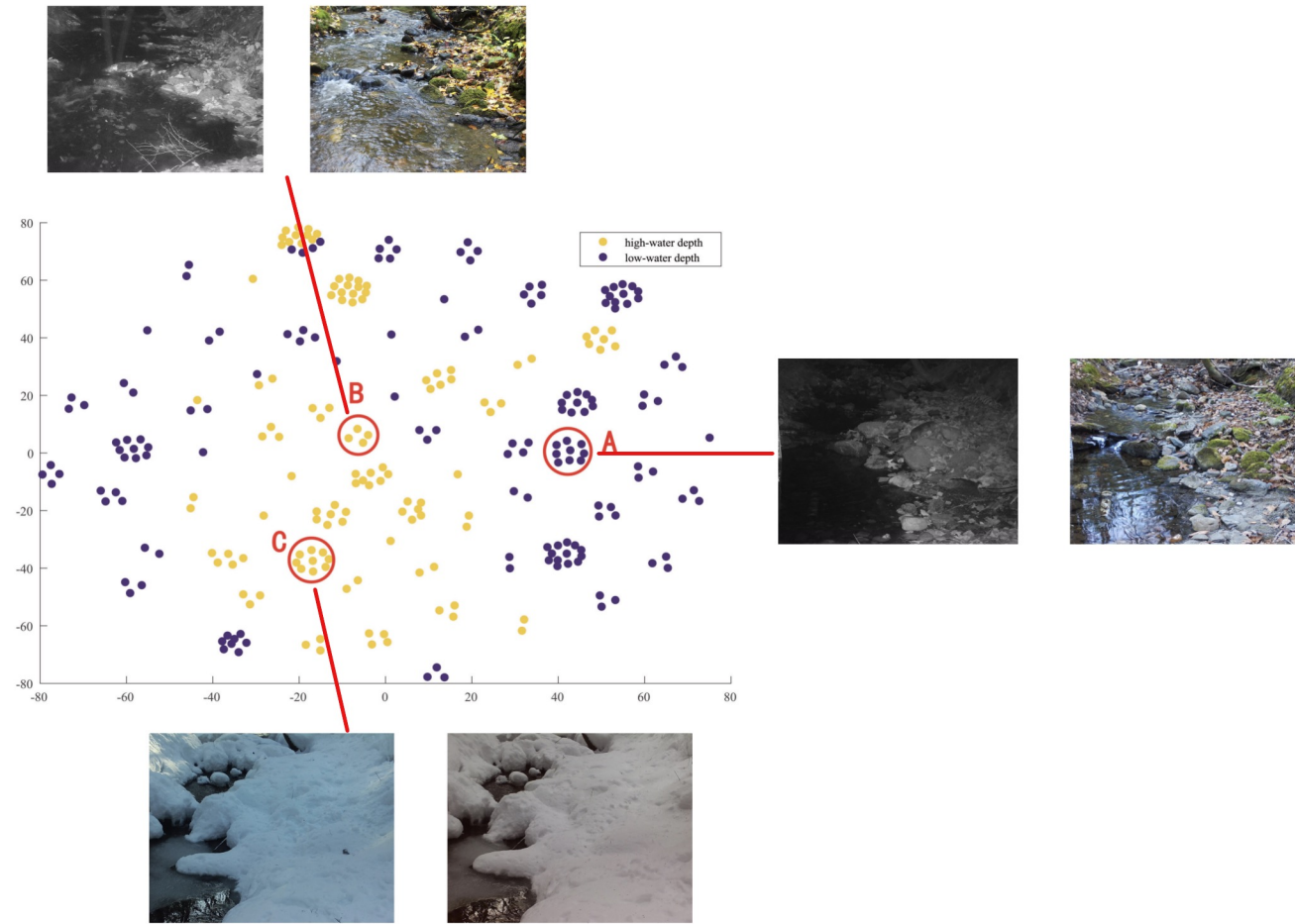
Spatial Time Series – Environment

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



Spatial Time Series – Environment

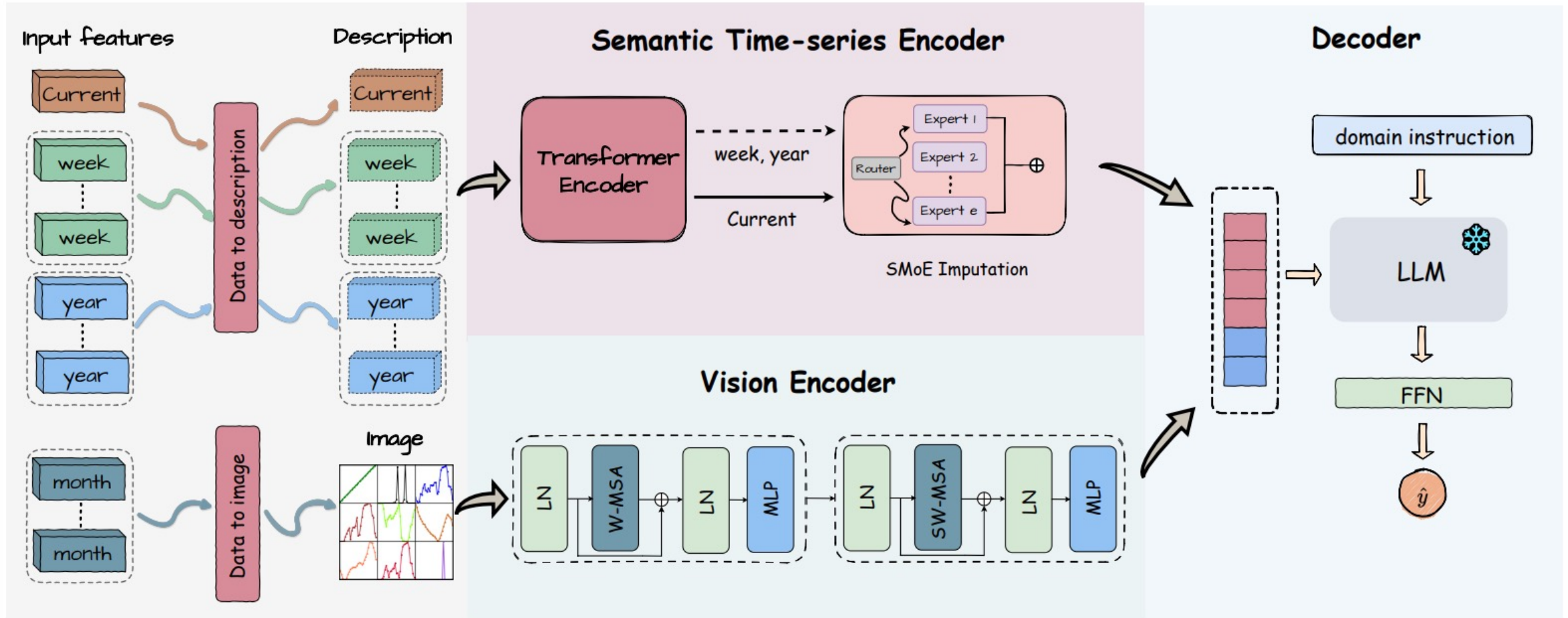
- Cross-modal imputation via variational approximation from low-dim features



Zhao et al. "VIMTS: Variational-based Imputation for Multi-modal Time Series", IEEE BigData 2022

Spatial Time Series - Environment

- Sparse mixture of experts + instruction-tuned LLM



Spatial Time Series – ST, Text, Image Datasets

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

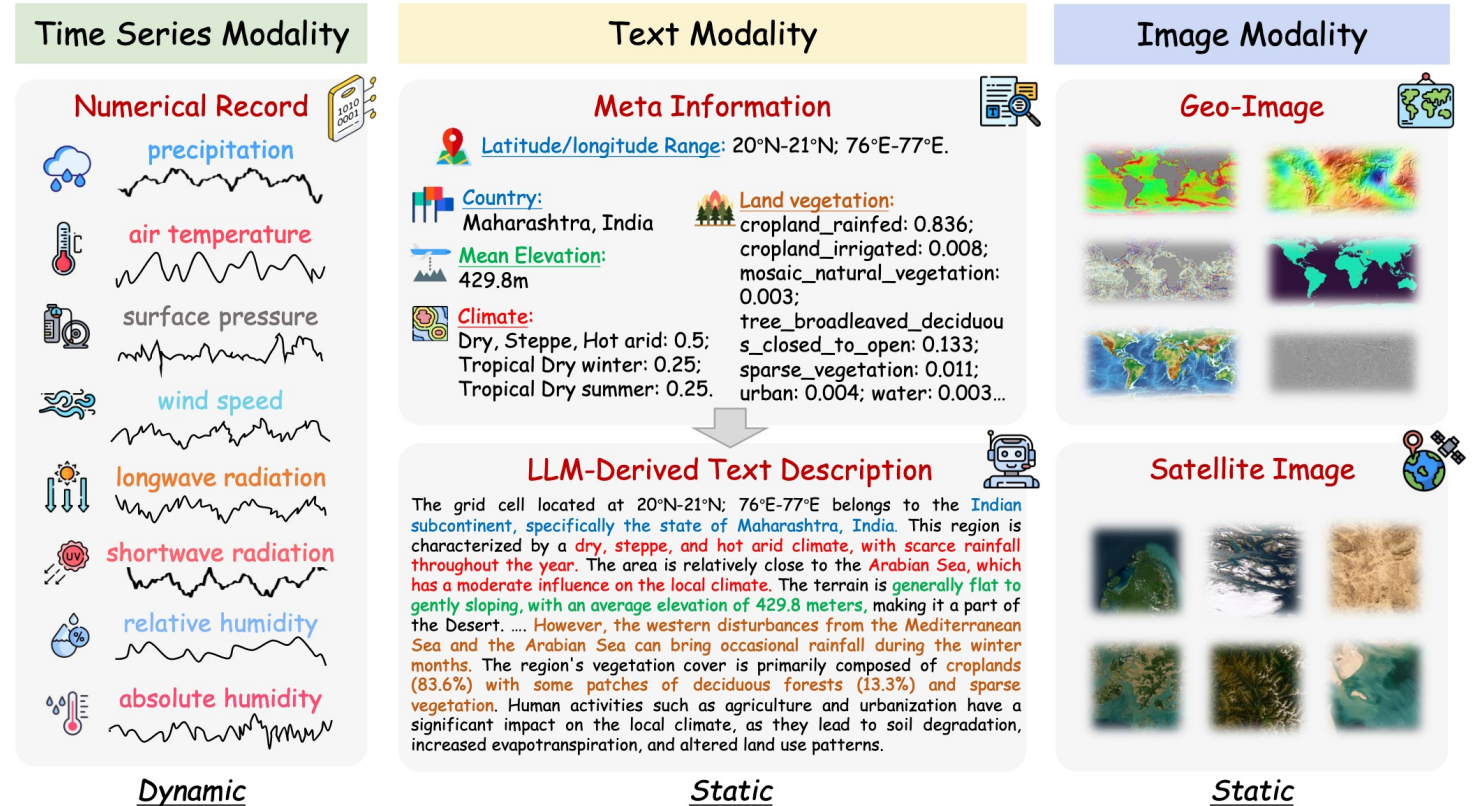


Figure 2: Different modality components of Terra. We provide the data with three temporal scales (3 hourly / daily / monthly), and three spatial scale (0.1° / 0.5° / 1°).

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

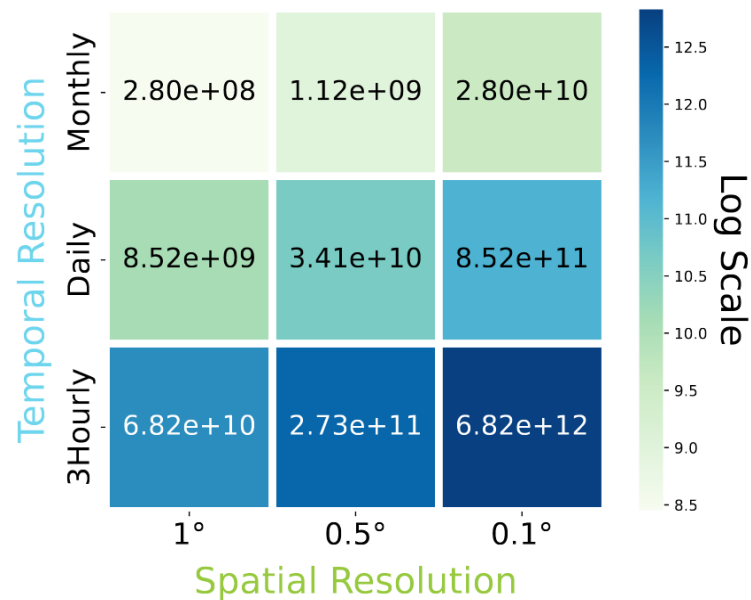


Figure 3: Dataset volume comparison.

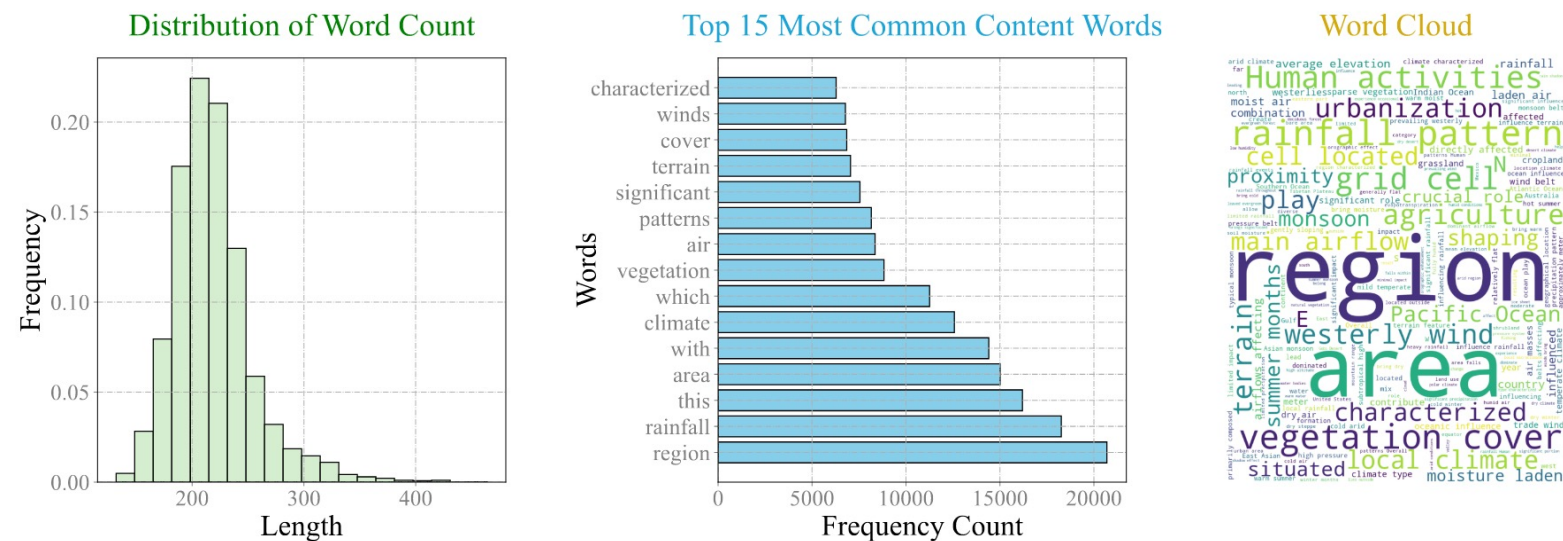
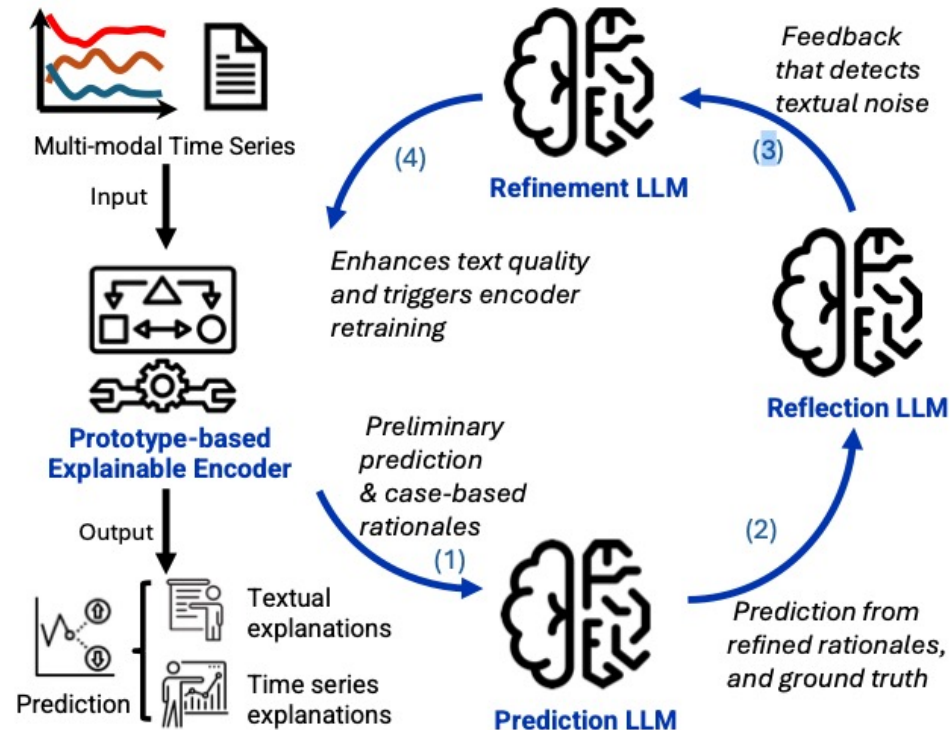


Figure 5: Statistical and visual insights of text modality data.

Future Research Directions

Future Research Directions

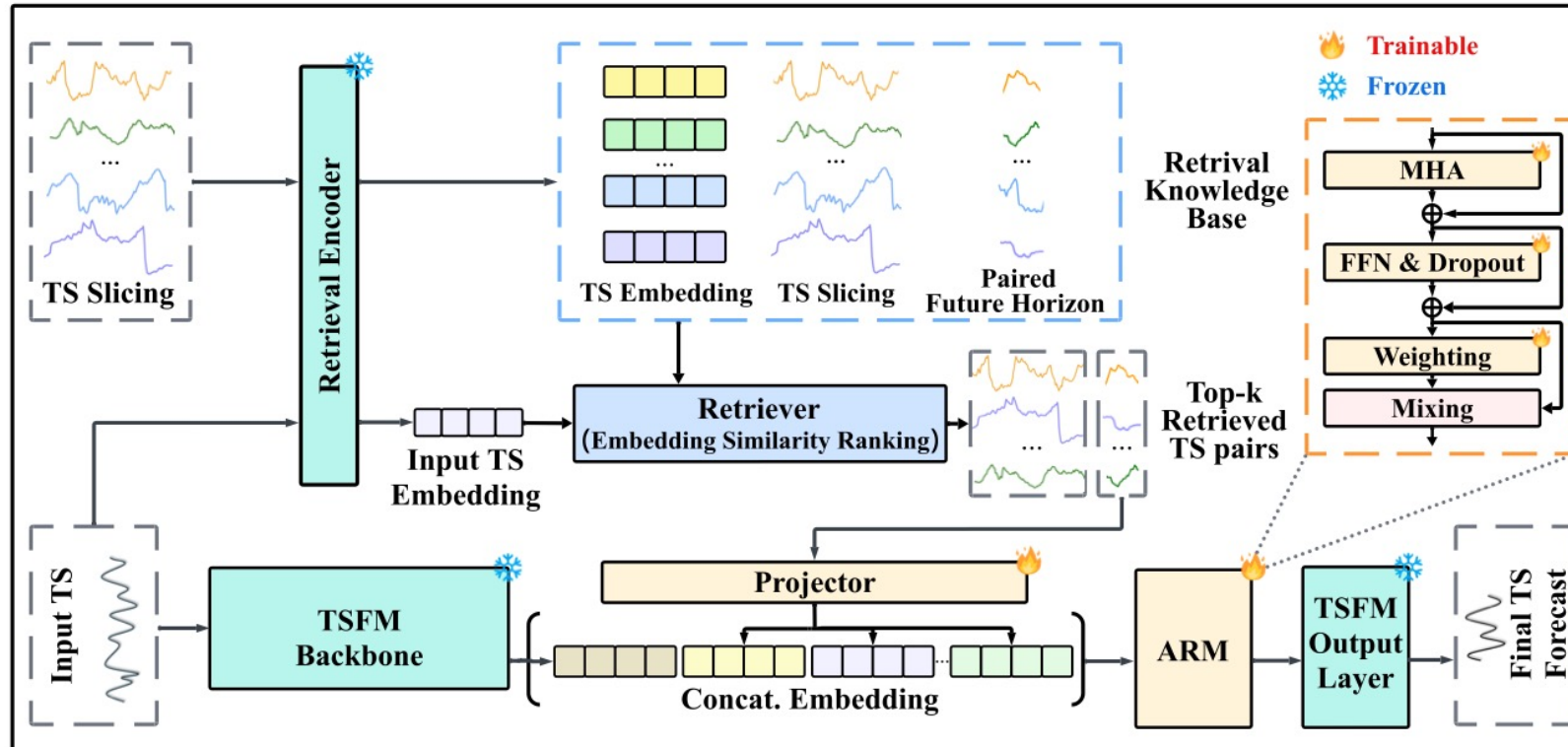
- **Robustness to imperfect Data:**
Handle missing or noisy real-world context effectively.



Future Research Directions

- **Enhanced reasoning with Multi-modal Time Series:**

Combine temporal reasoning with context understanding for interpretable inference.



Ning et al. "TS-RAG: Retrieval-Augmented Generation based Time Series Foundation Models are Stronger Zero-Shot Forecaster", 2025

Future Research Directions

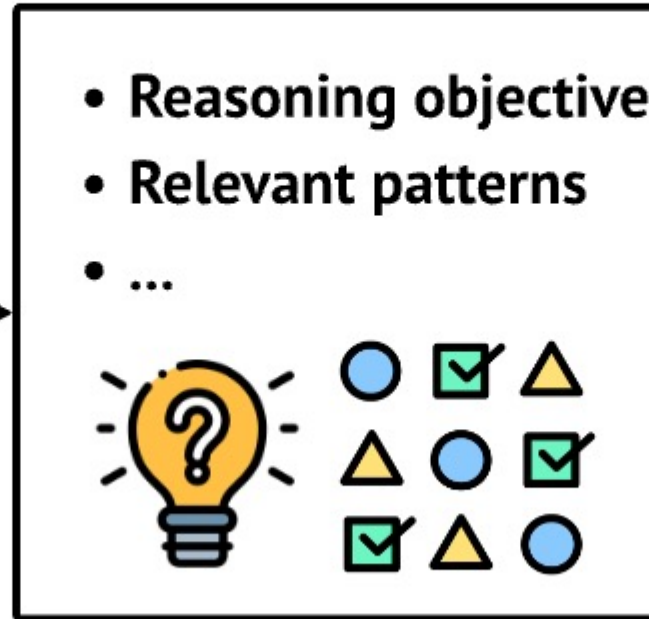
- Towards structured reasoning with multi-modal data

Structured Reasoning

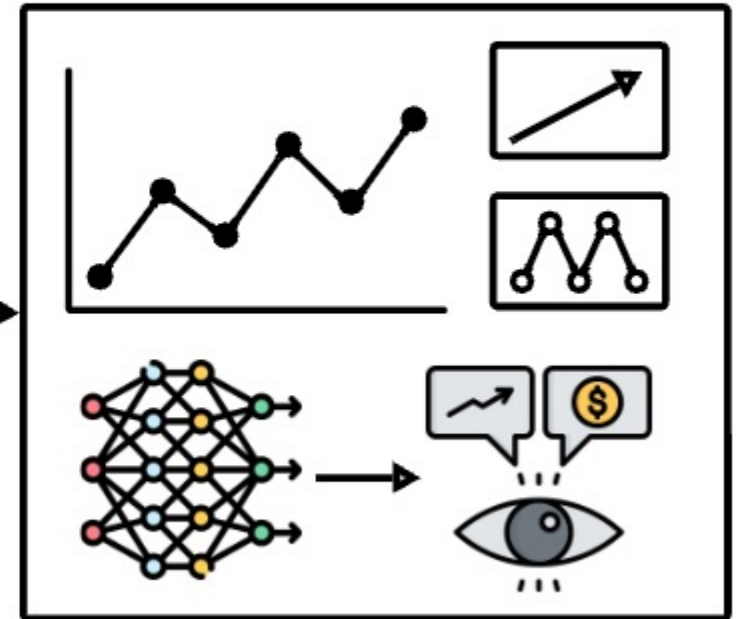
Domain Identification



Task Framing



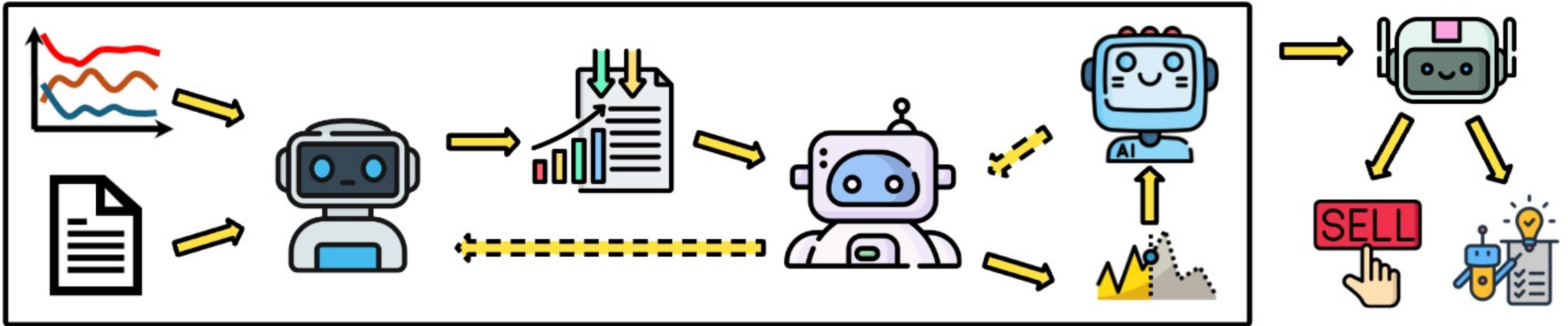
Temporal Reasoning



Future Research Directions

- Multi-agent system for decision making.

Multi-agent Collaboration



Future Research Directions

- **Decision-making Systems:**

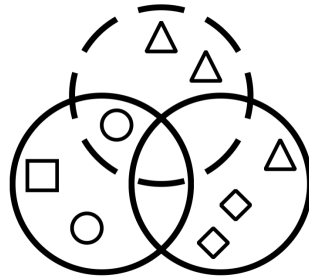
Develop adaptive decision-support systems using multi-modal data to facilitate downstream tasks.

- **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.



...

Thank you!

Q & A

Survey Paper



Github



NEC
NEC Laboratories **America**

Morgan
Stanley