The 40th Annual AAAI Conference on Artificial Intelligence
JANUARY 20 – JANUARY 27, 2026 | SINGAPORE

# Multi-modal Time Series Analysis

## — Methods, Datasets, and Applications

Survey Paper

Github

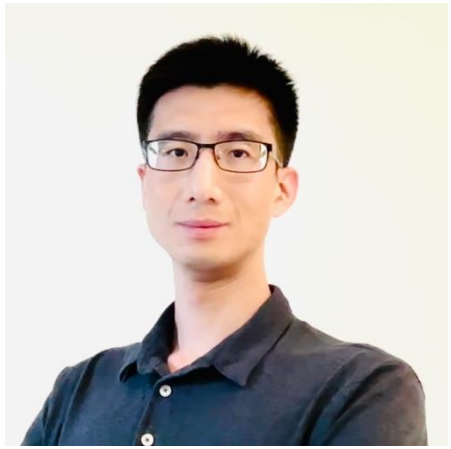
HKUST (GZ)

NEC Laboratories America

Morgan Stanley

# Presenters

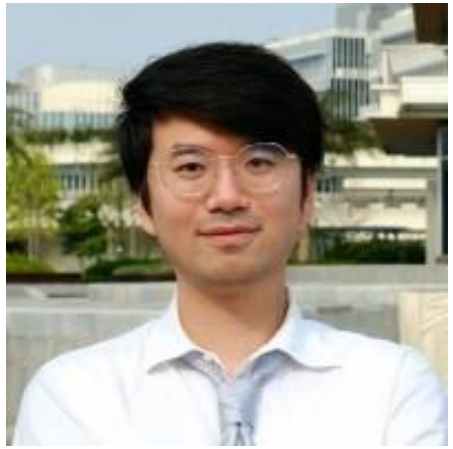
**Dongjin Song**
Associate Professor
School of Computing
University of Connecticut



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

**Siru Zhong**
Ph.D Student
Hong Kong University of Science
and Technology (GZ)

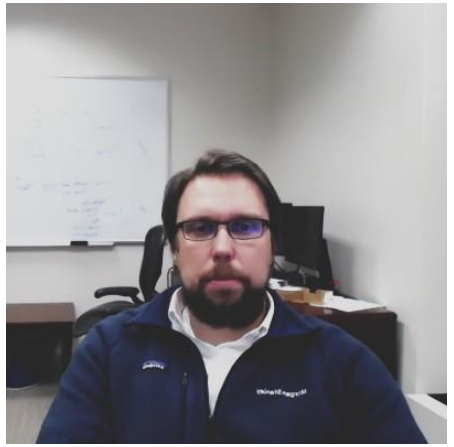
# Contributors



**Yuxuan Liang**
Assistant Professor
Hong Kong University of Science and Technology (GZ)

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

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



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

# Agenda

- **Part 1: Opening and Introduction** (10 min – Dongjin)
- **Part 2-1: Taxonomy of Multi-modal Time Series Methods** (40 min – Dongjin)
- **Part 2-2: Taxonomy of Multi-modal Time Series Methods** (40 min - Jinchao)

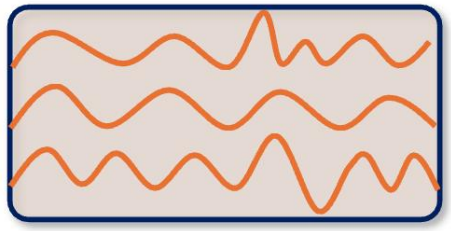
——————————————— Q&A + Break (30 min) ———————————————-

- **Part 3: Multi-model Learning for Spatial-temporal Data** (40 min - Siru)
- **Part 4: Multi-modal Time Series Applications and Datasets** (15 min - Dongjin)
- **Part 5: Future Directions** (10 min - Dongjin)
- **Part 6: Q&A**

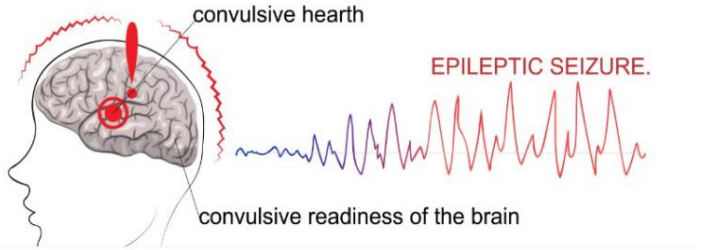
# *Introduction to Multi-modal Time Series Analysis*

# Background –Time Series Analysis

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

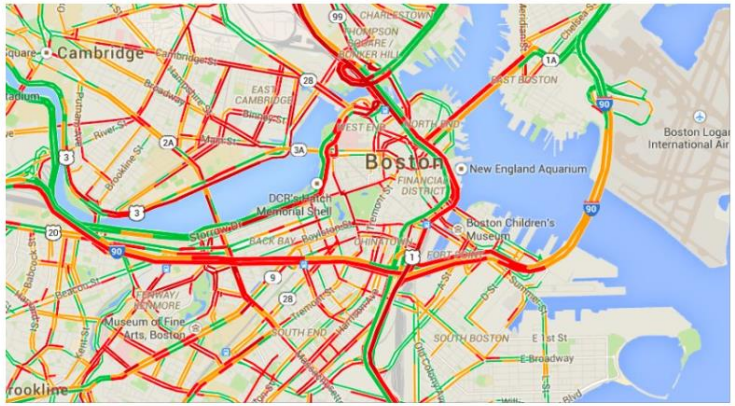


**Time series data**




**Healthcare**

**Electricity load & Power consumption**

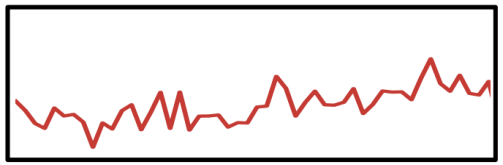



**Traffic networks**

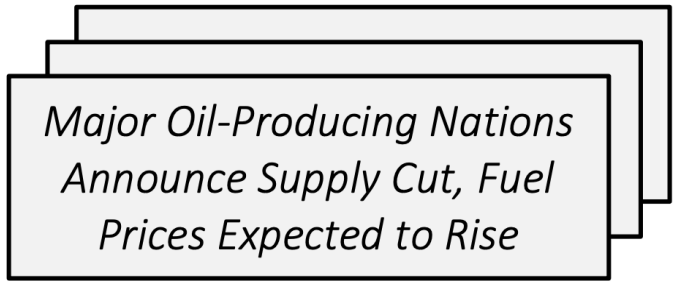
# Background – Multi-modal Time Series Analysis

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

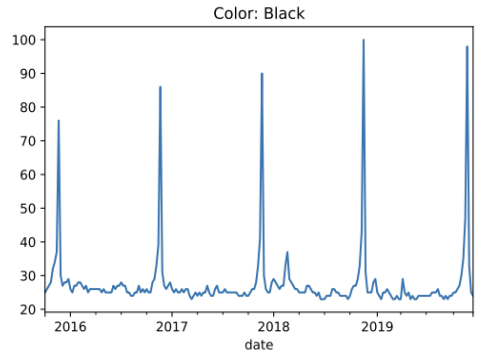
**Multi-modal Time Series:** Time series that associated with external contexts (knowledge), which can carry rich semantic information for time series analysis.

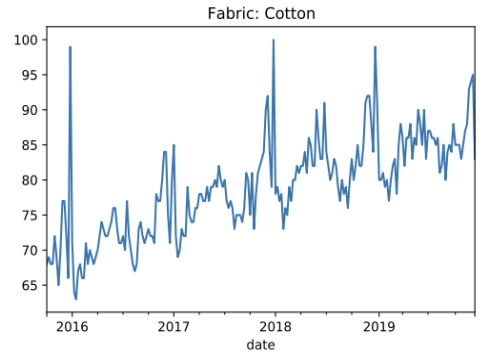


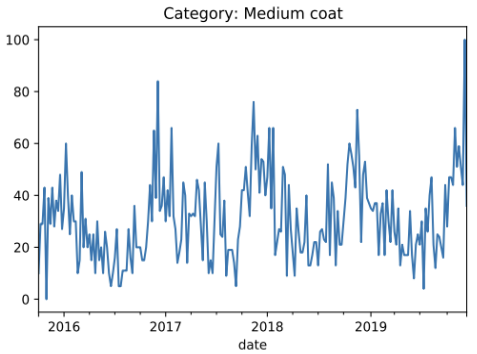
Time Series



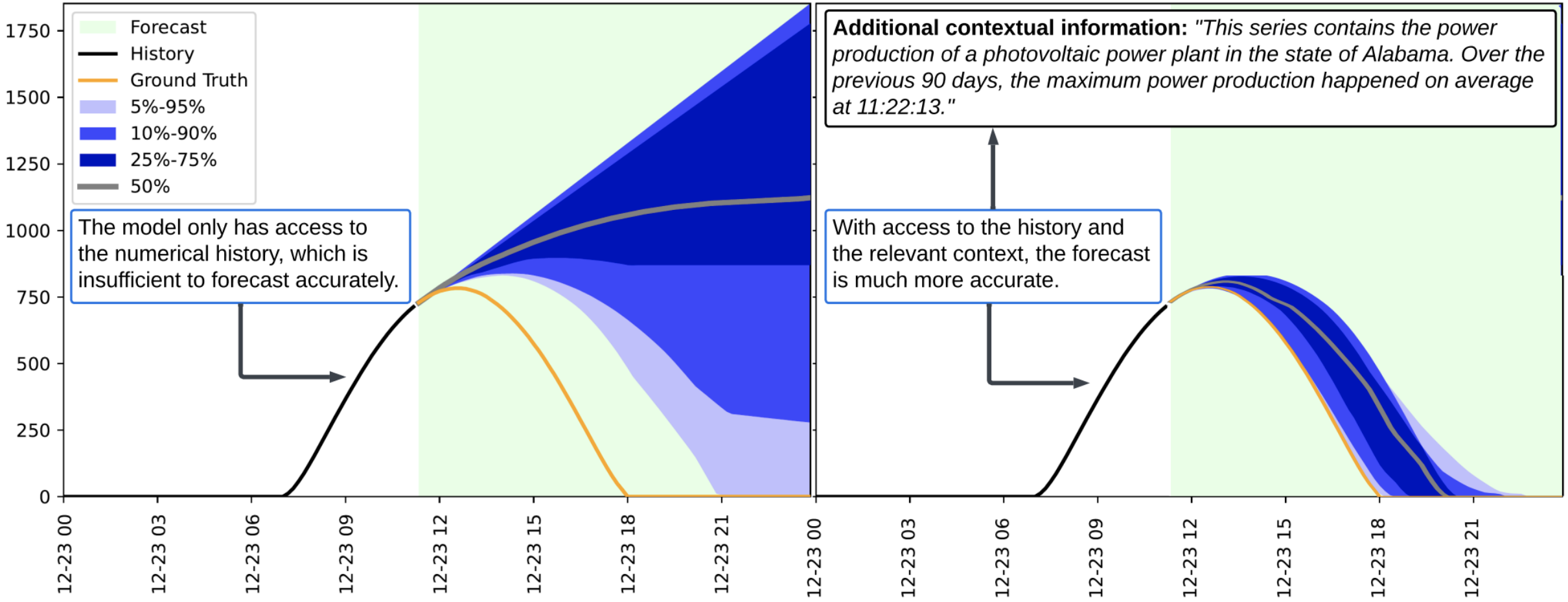

Text




# Background – Multi-modal Time Series Analysis




**Williams et al. "Context is Key: A Benchmark for Forecasting with Essential Textual Information", ICML 2025.**

# Background – Multi-modal Time Series Analysis

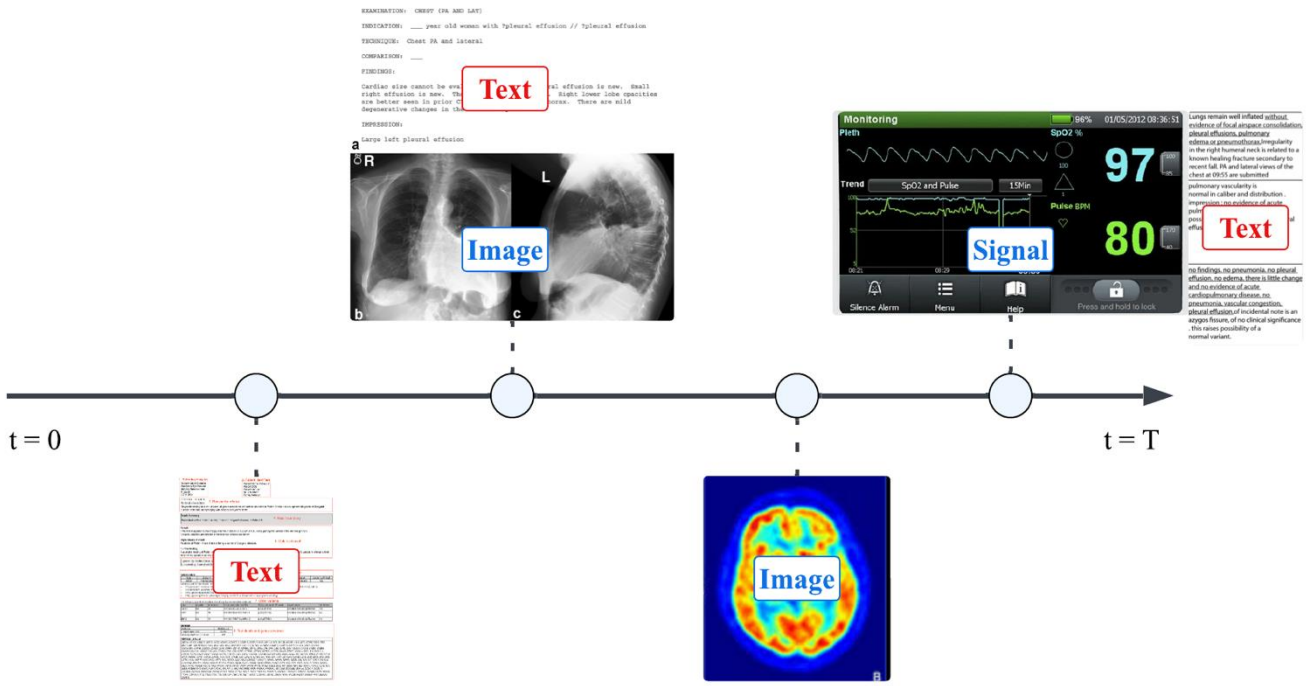
**Why is Multi-modality Important?**

Real-world systems are **heterogeneous**.

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

**Examples:**

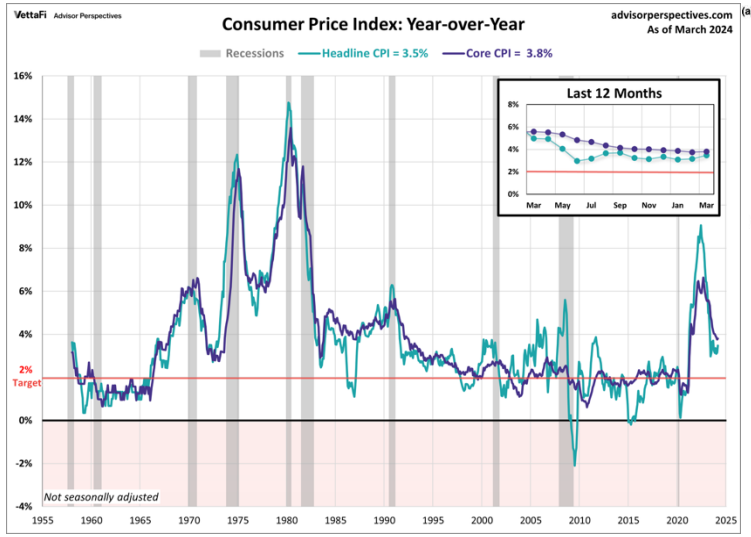
Electronic Health Records (EHR)

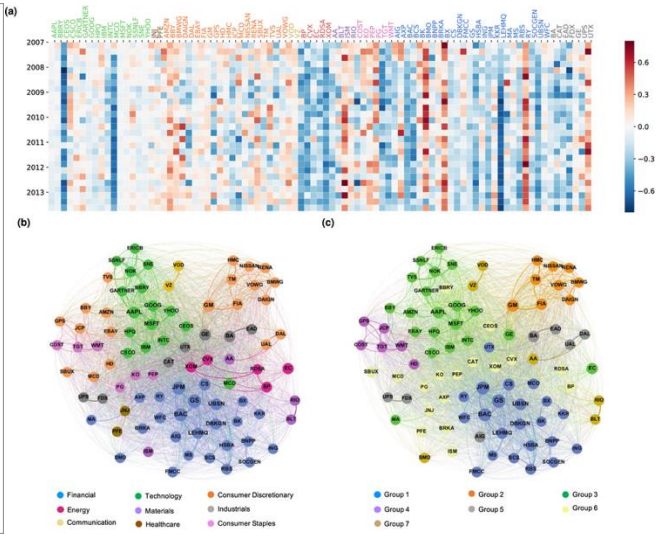


# Background – Multi-modal Time Series Analysis

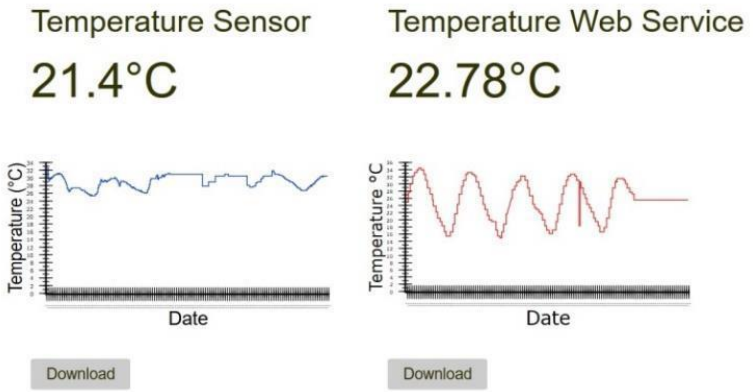
**More Examples:**

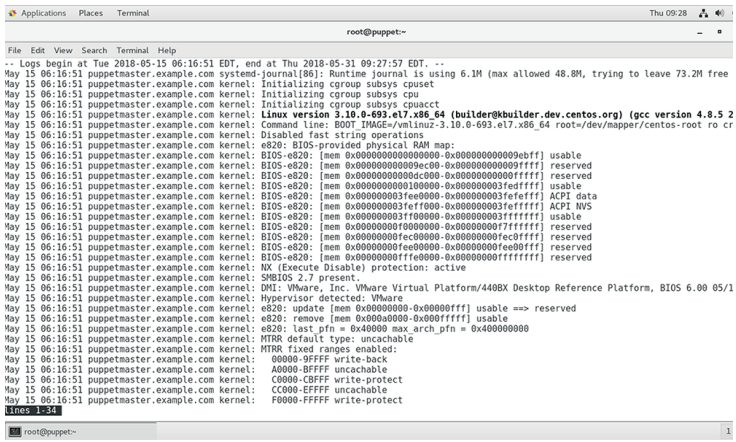
**Finance:** Price + News Sentiment



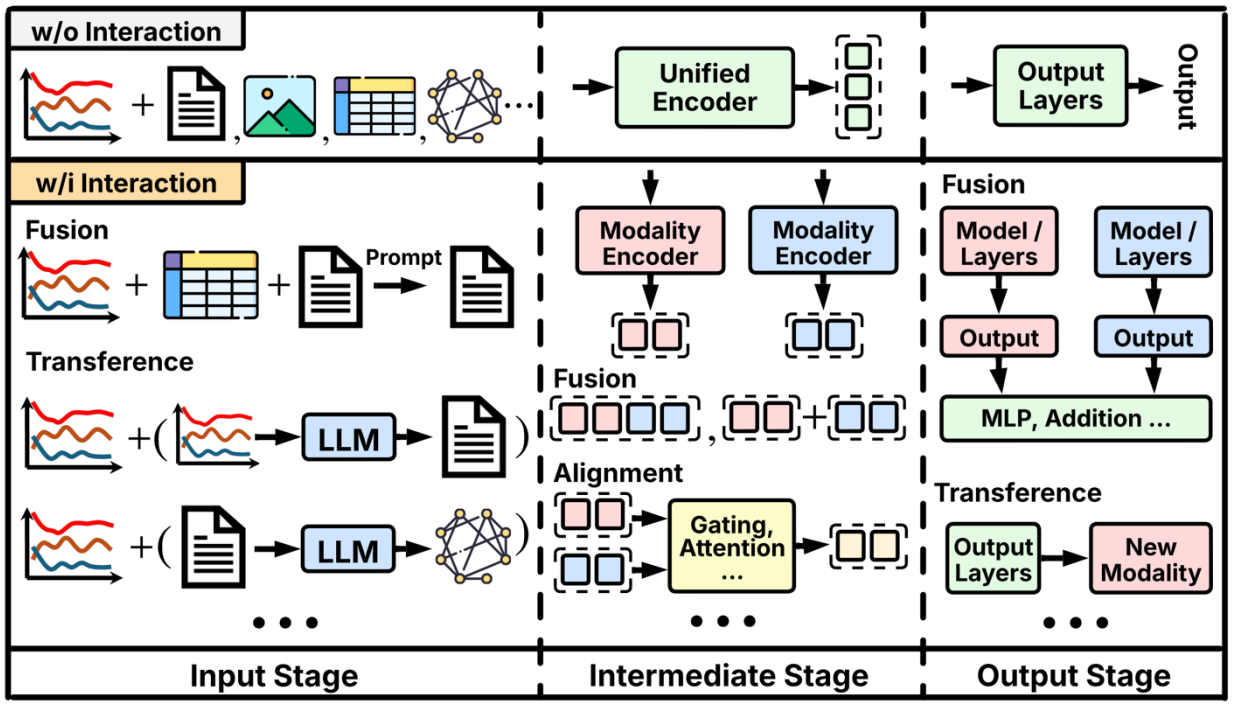

**IoT Systems:** Temperature + Logs



# 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.



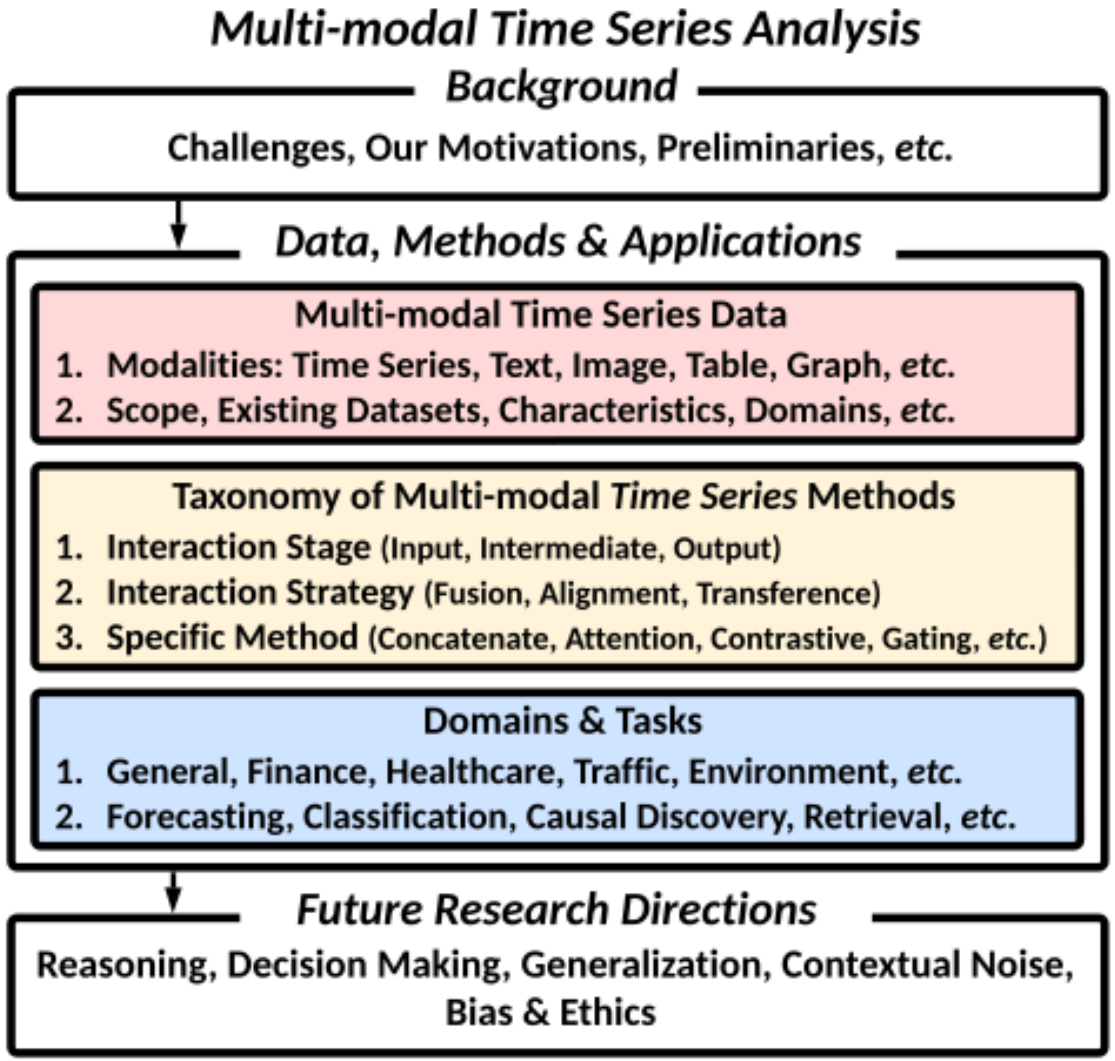

Jiang et al. Multi-modal Time Series Analysis: A Tutorial and Survey, KDD 2025

# 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.
   - For Part 3, the focus is on ***multi-modal spatial-temporal*** 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.

Jiang et al. Multi-modal Time Series Analysis: A Tutorial and Survey, KDD 2025

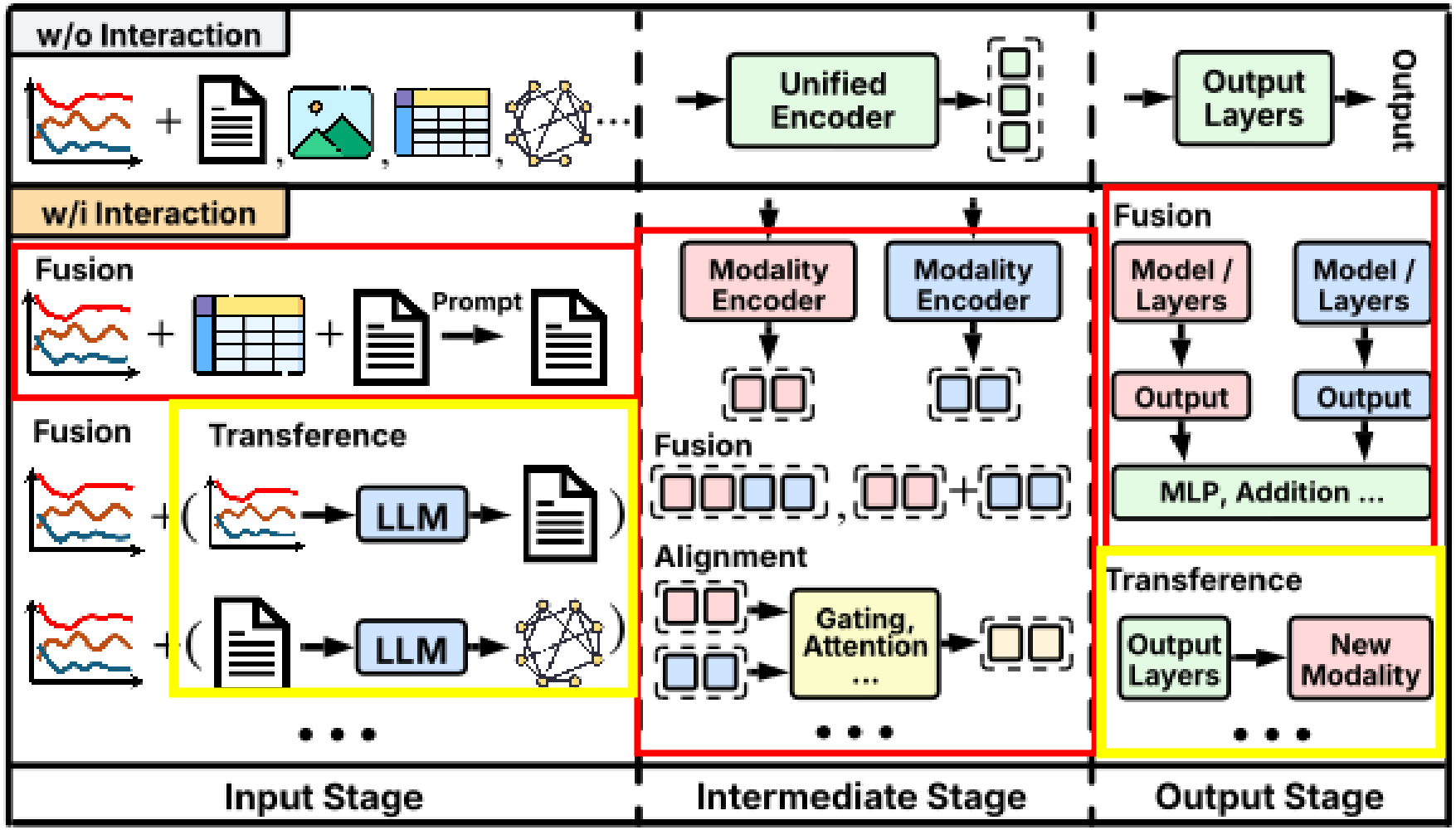
# *Multi-modal Time Series Methods*

# 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



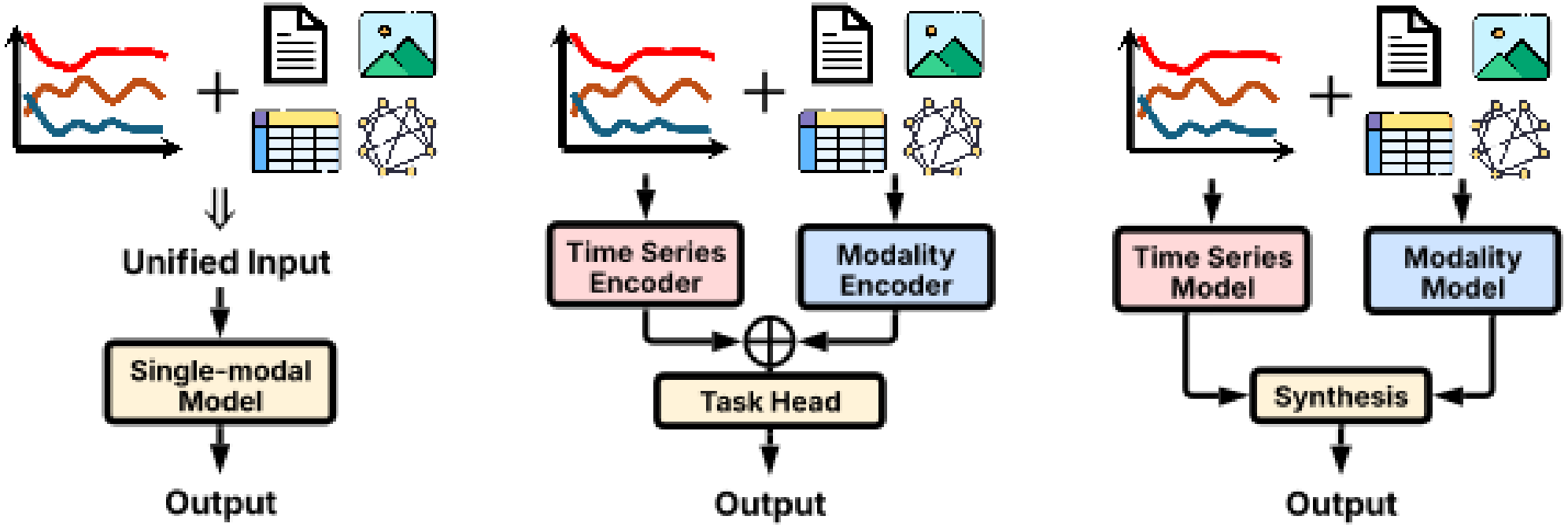

Overview and representative examples of cross-modal interactions

Jiang et al. Multi-modal Time Series Analysis: A Tutorial and Survey, KDD 2025

# *Multi-modal Time Series Methods*
# *Part 1: Fusion and Alignment*

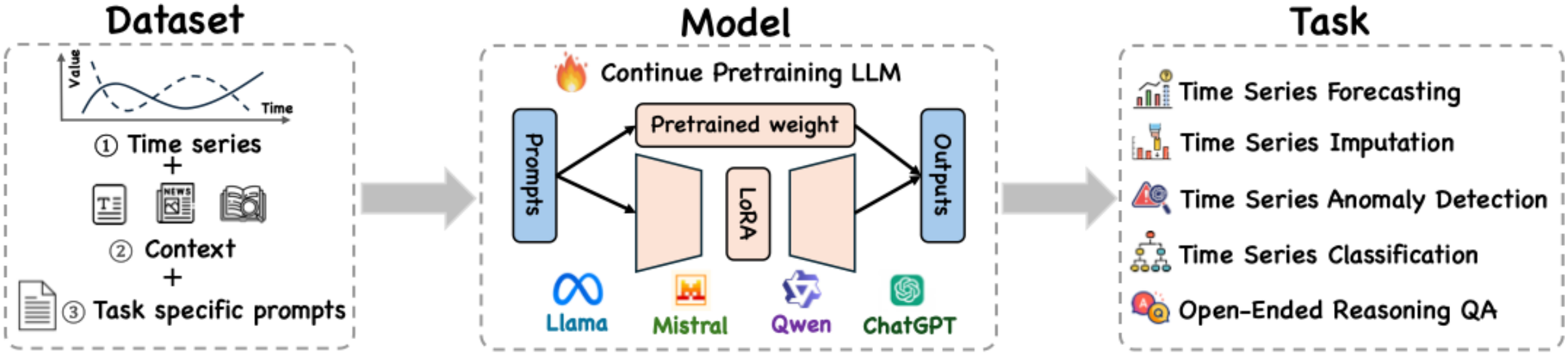
# Cross-modal Interaction with Time Series: Fusion

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

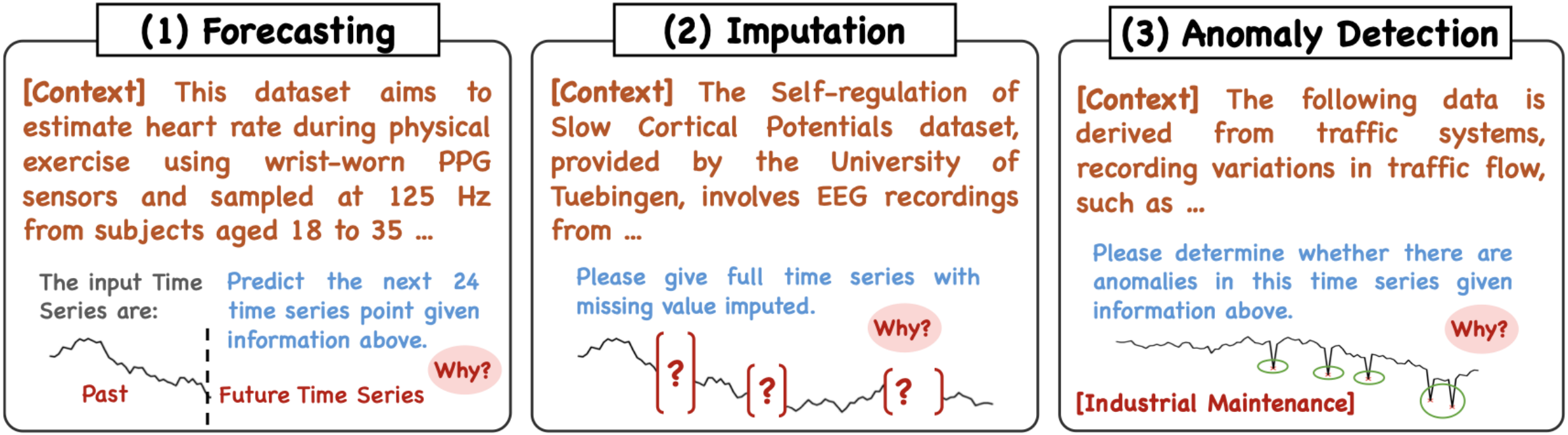
Integrate time series, tabular data and texts into a unified textual prompt




**Kong et al. "Time-MQA: Time Series Multi-Task Question Answering with Context Enhancement", ACL 2025**

# Multi-modal Fusion with Time Series – Input level

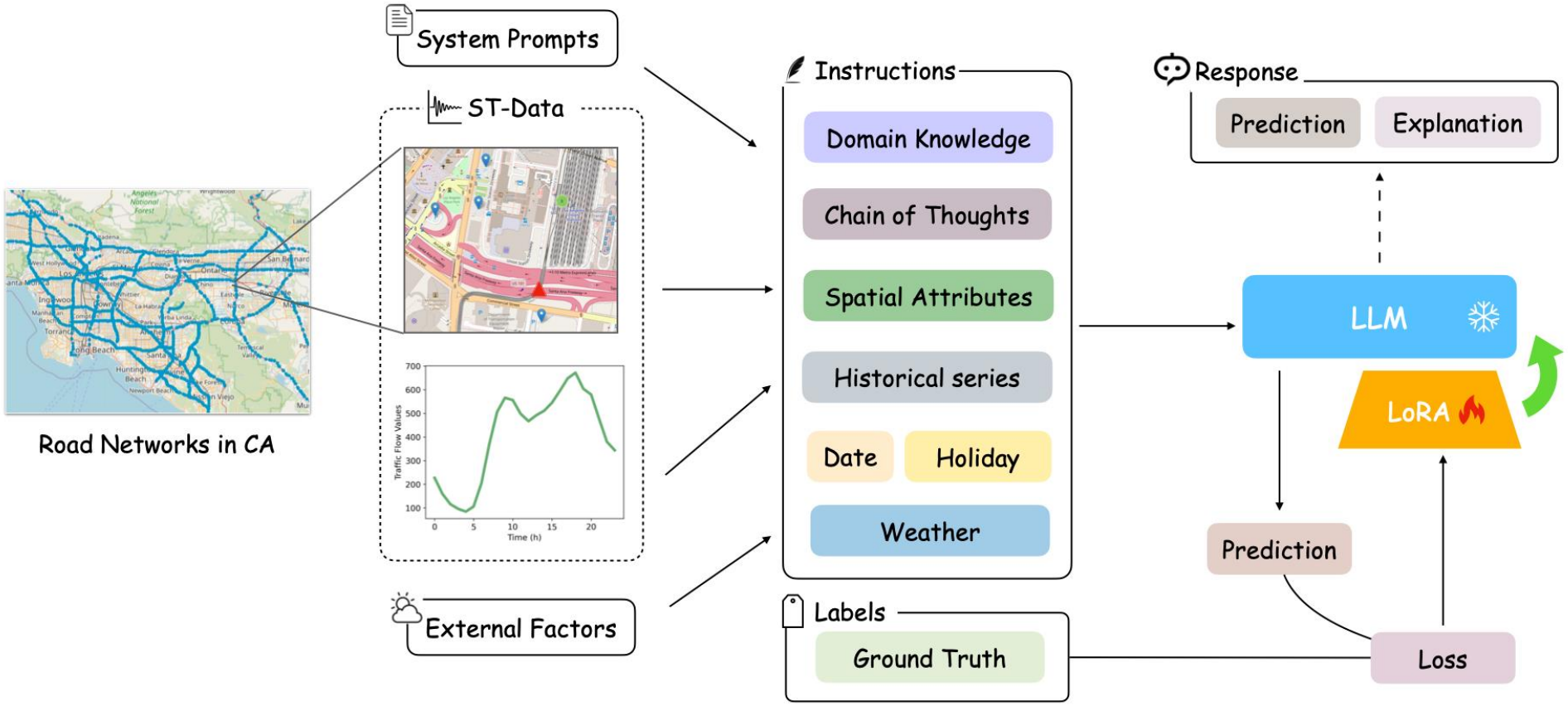
Integrate time series, tabular data and texts into a unified textual prompt




**Kong et al. "Time-MQA: Time Series Multi-Task Question Answering with Context Enhancement", ACL 2025**

# Multi-modal Fusion with Time Series – Input level

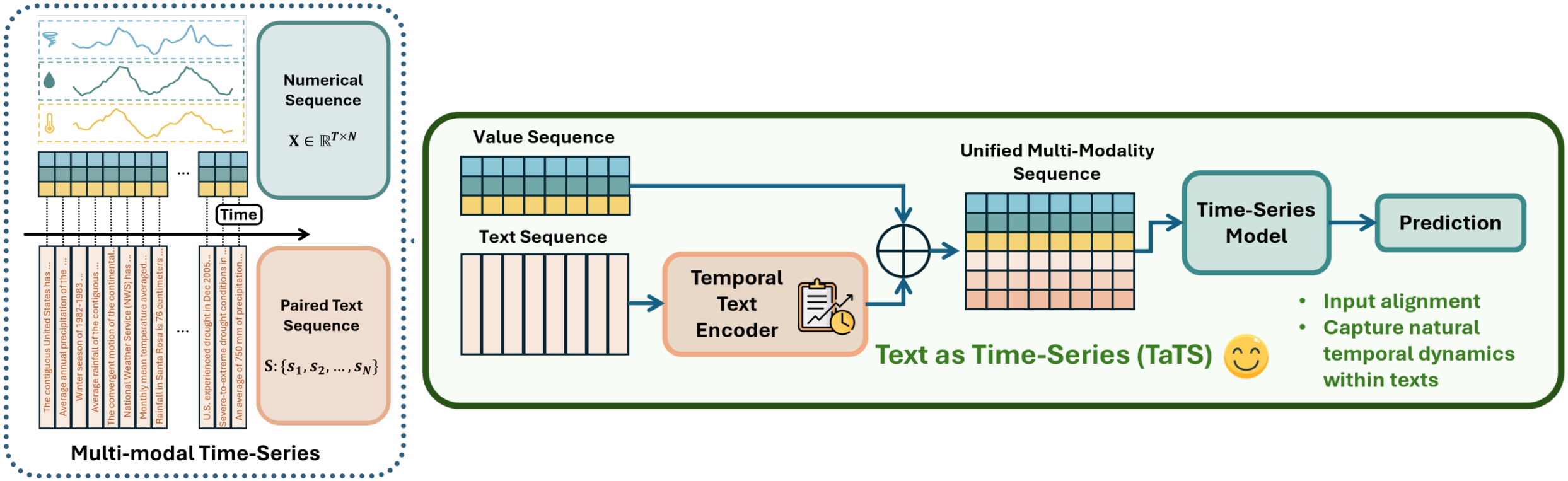
Integrate time series, tabular data and texts into a unified textual prompt




**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

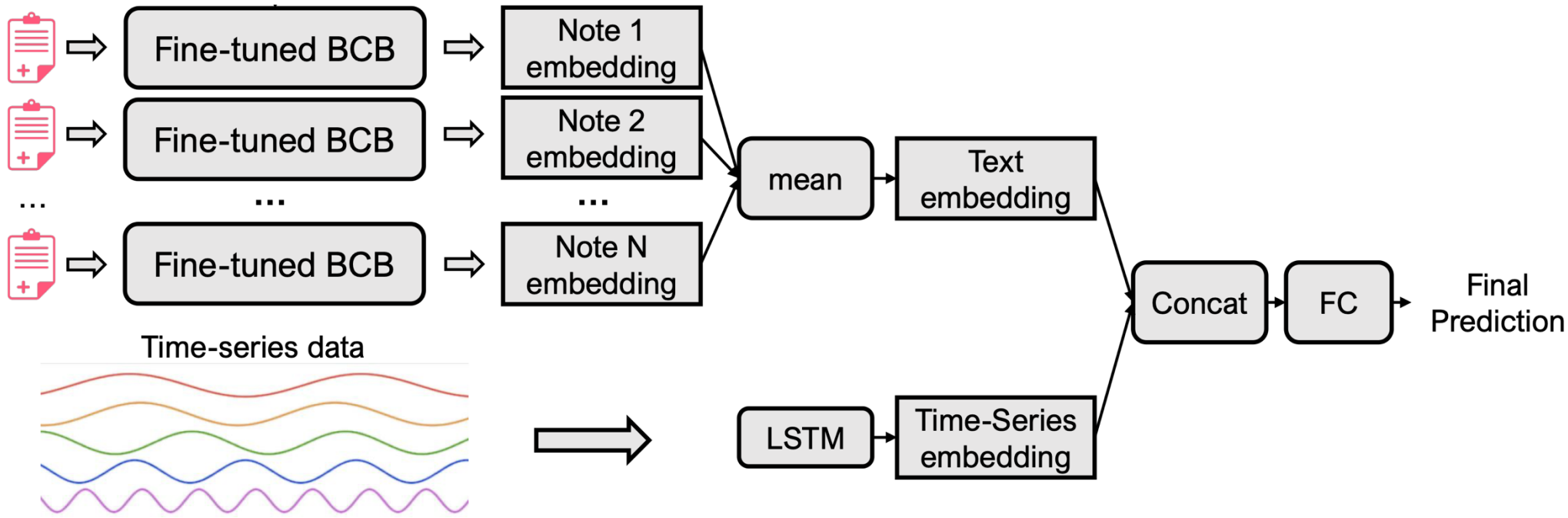
Integrate paired text embedding as an additional variable of time series




**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

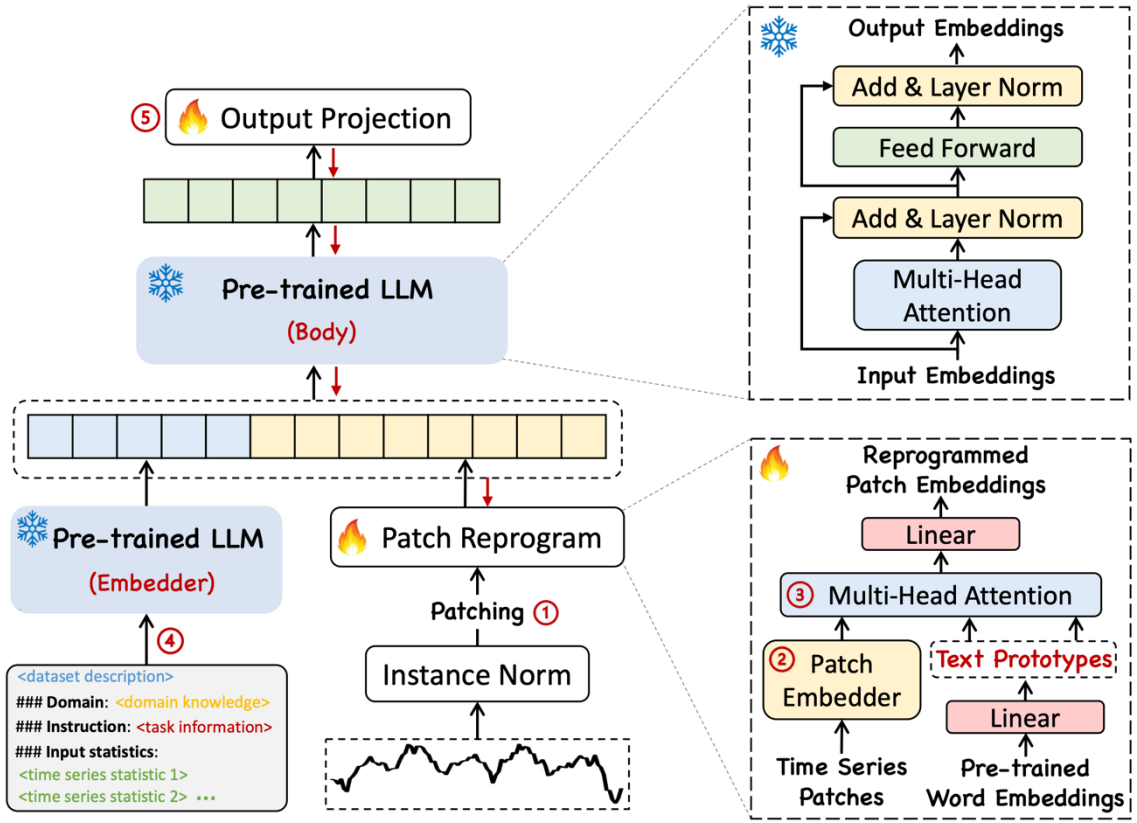
Simple aggregations (*e.g.*, mean, addition, concatenation, *etc.*) of time series embedding and other modality embeddings



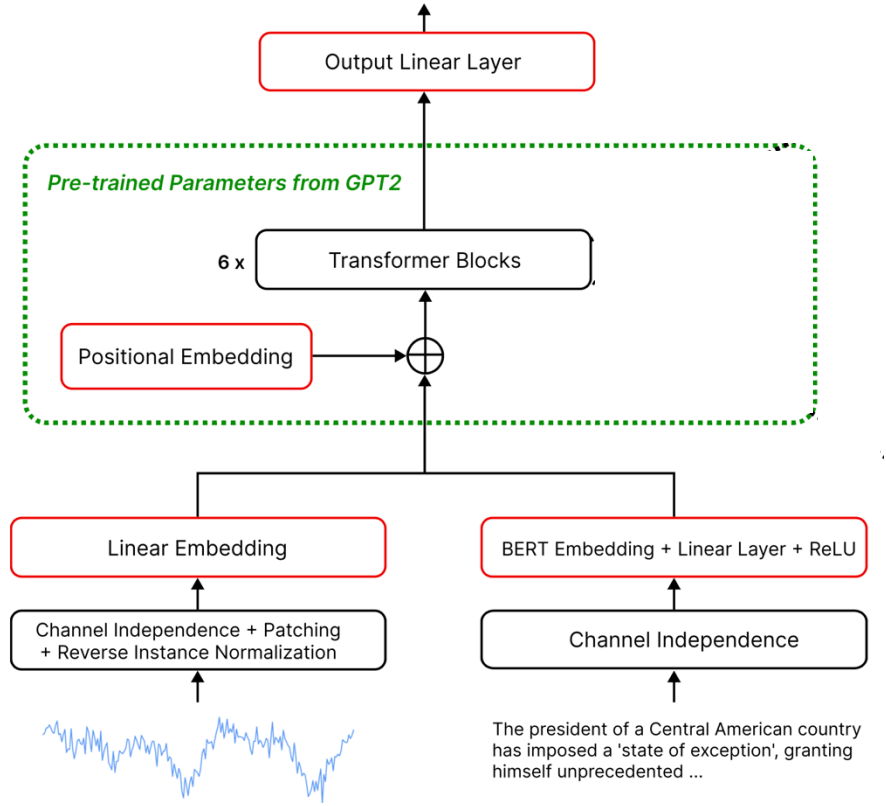

**Deznabi et al. "Predicting In-hospital Mortality by Combining Clinical Notes with Time-series Data", ACL 2021.**

# Multi-modal Fusion with Time Series – Intermediate level

The fusion of modality embeddings is usually followed by alignments



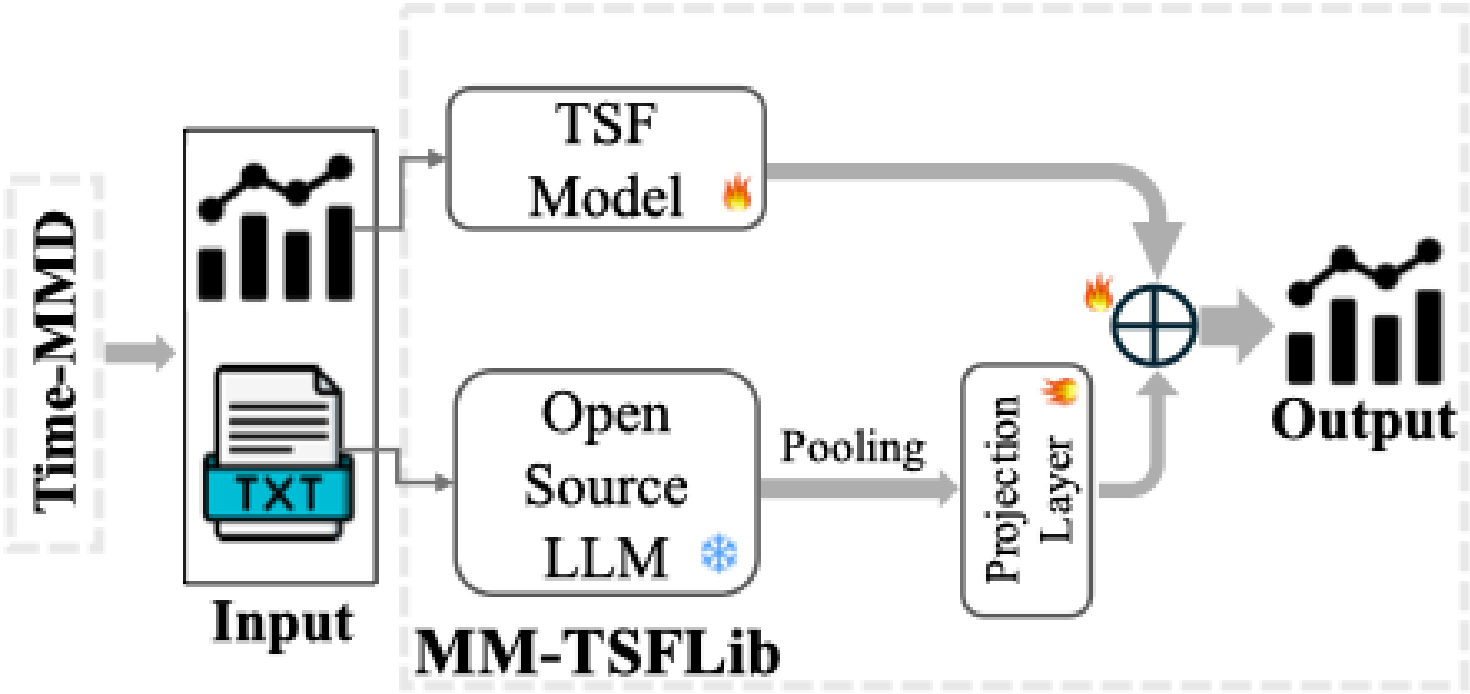

**Jin et al. "Time-LLM: Time Series Forecasting by Reprogramming Large Language Models", ICLR 2024.**




**Jia et al. "GPT4MTS: Prompt-Based Large Language Model for Multimodal Time Series Forecasting", AAAI 2024.**

# Multi-modal Fusion with Time Series – Output level

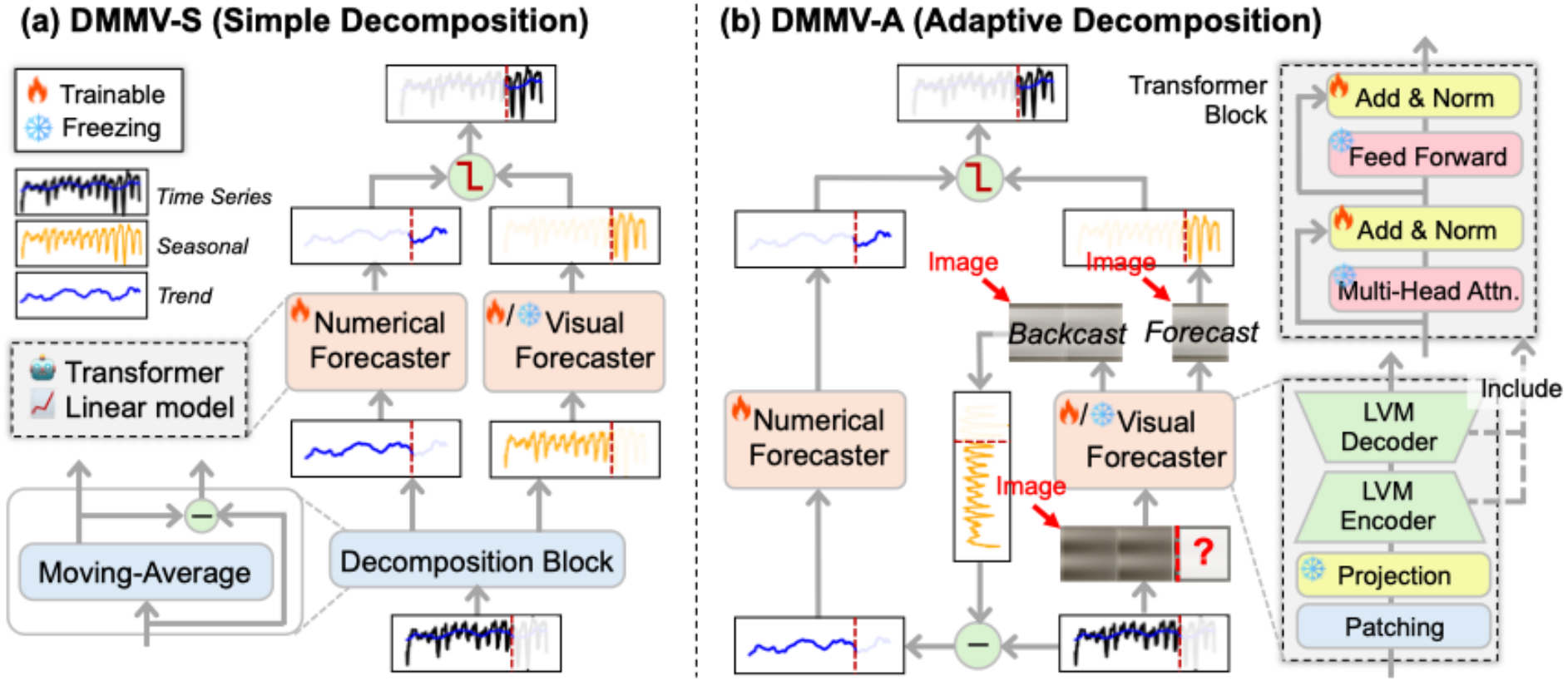
Project multiple modality outputs onto a unified space




**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




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

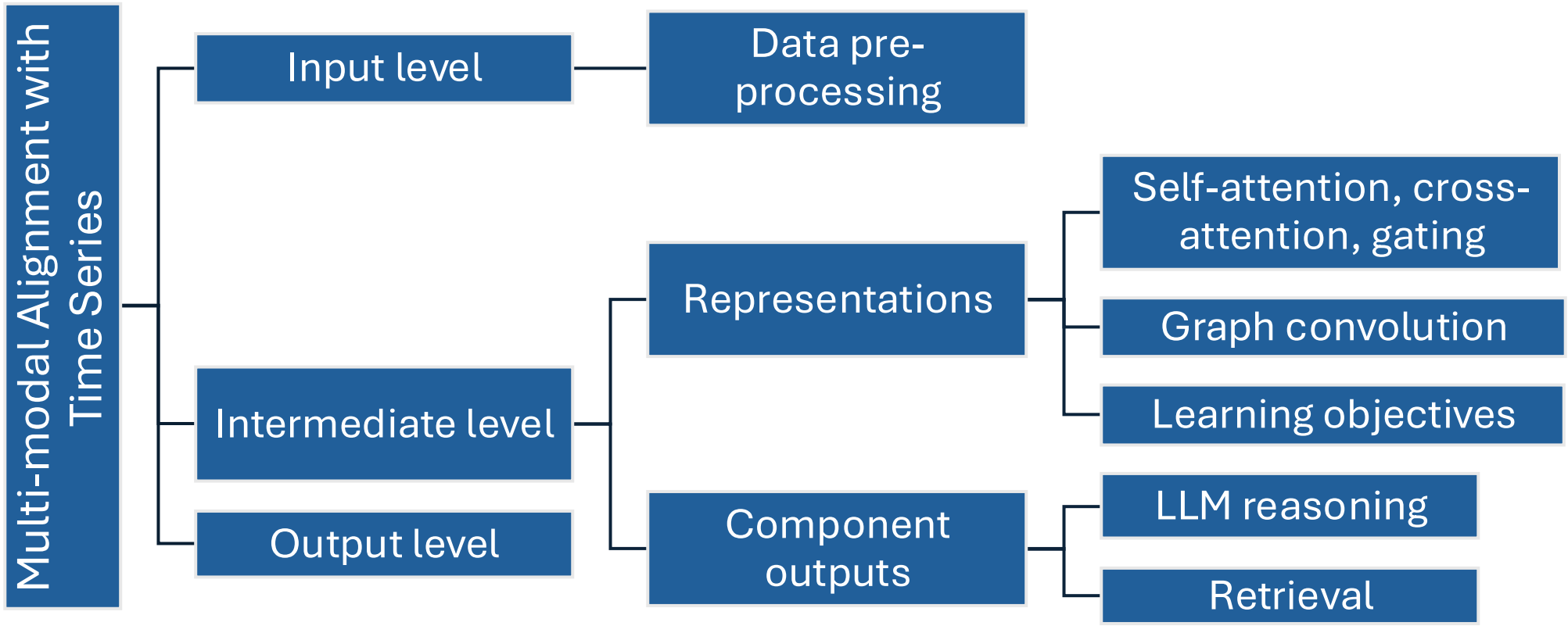
**Shen et al. "Multi-Modal View Enhanced Large Vision Models for Long-Term Time Series Forecasting", 2025**

# 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

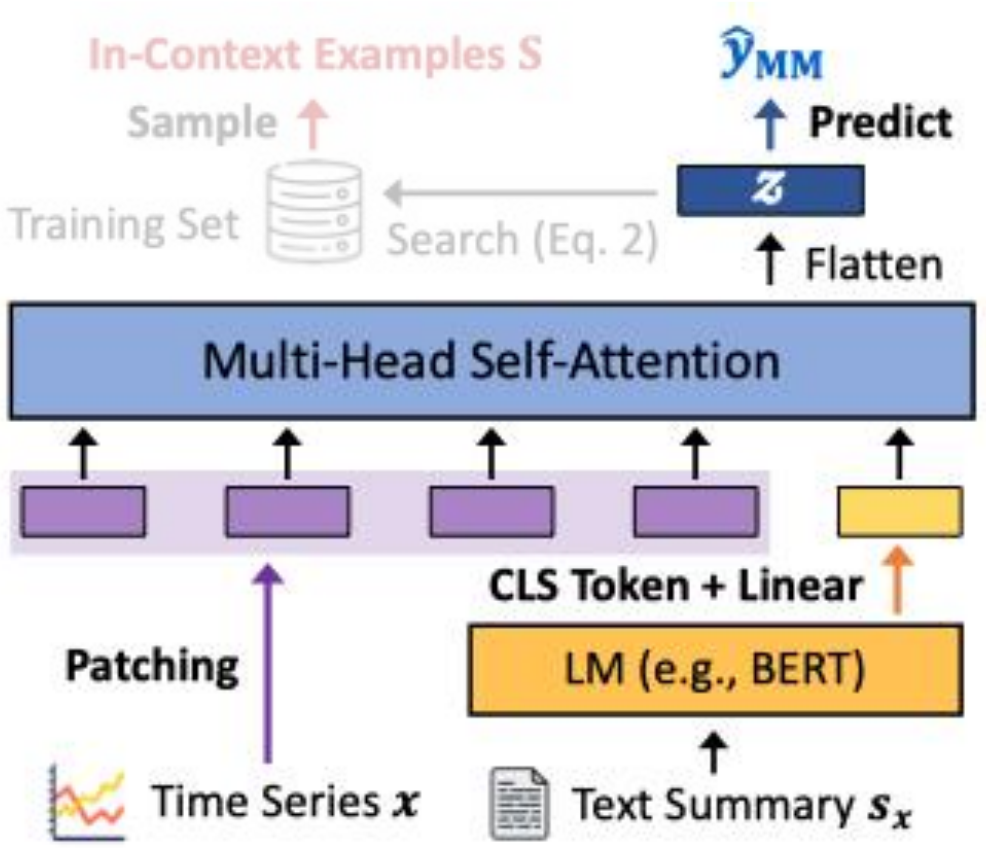
**Self-attention:** a joint and undirected alignment across all modalities by dynamically attending to important features.

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

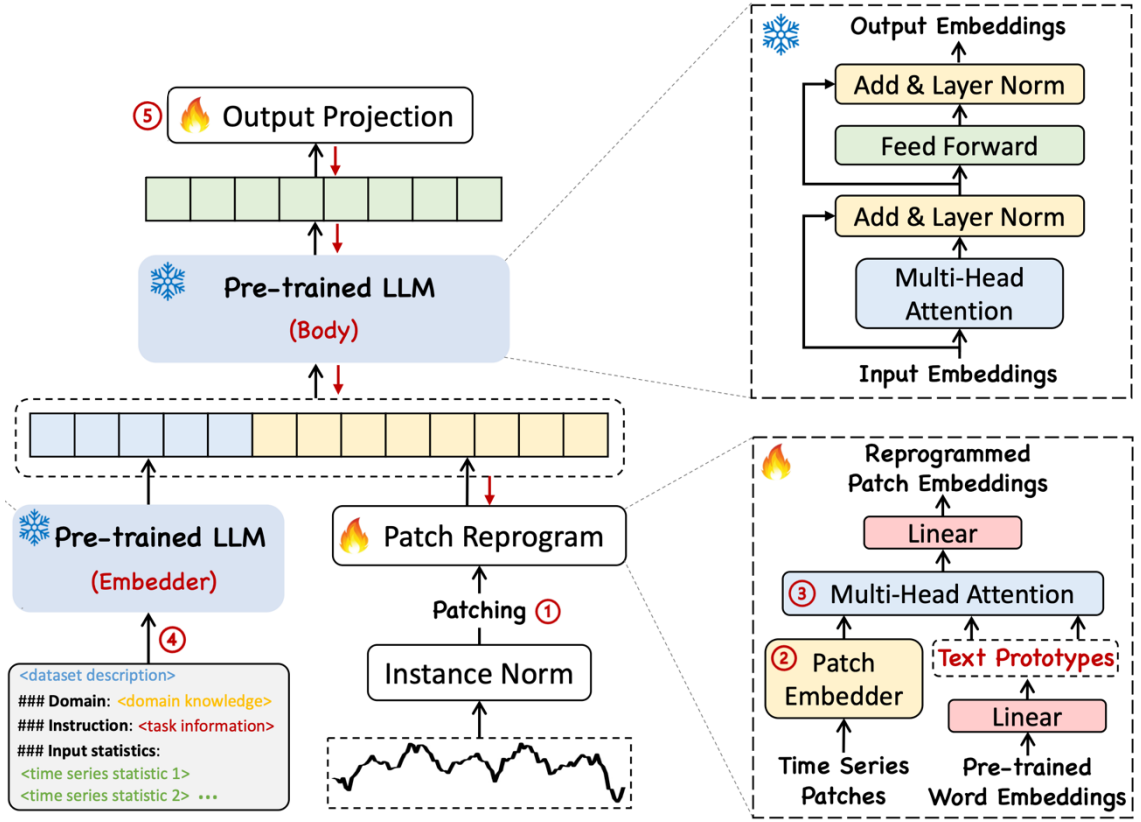
$$\mathrm{Attention}(\boldsymbol{E}_{\mathrm{mm}}) = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\top}}{\sqrt{\boldsymbol{d}_{\boldsymbol{k}}}}\right)\boldsymbol{V}$$

where the queries $\boldsymbol{Q}$, keys $\boldsymbol{K}$, and values $\boldsymbol{V}$ are linear projections of $\boldsymbol{E}_{\mathrm{mm}}$:
$\boldsymbol{Q} = \boldsymbol{E}_{\mathrm{mm}} W_Q$, $\boldsymbol{K} = \boldsymbol{E}_{\mathrm{mm}} W_K$, $\boldsymbol{V} = \boldsymbol{E}_{\mathrm{mm}} W_V$ with learnable weights $W_{Q,K,V} \in \mathbb{R}^{d \times d_k}$

# Multi-modal Alignment with Time Series - Representations

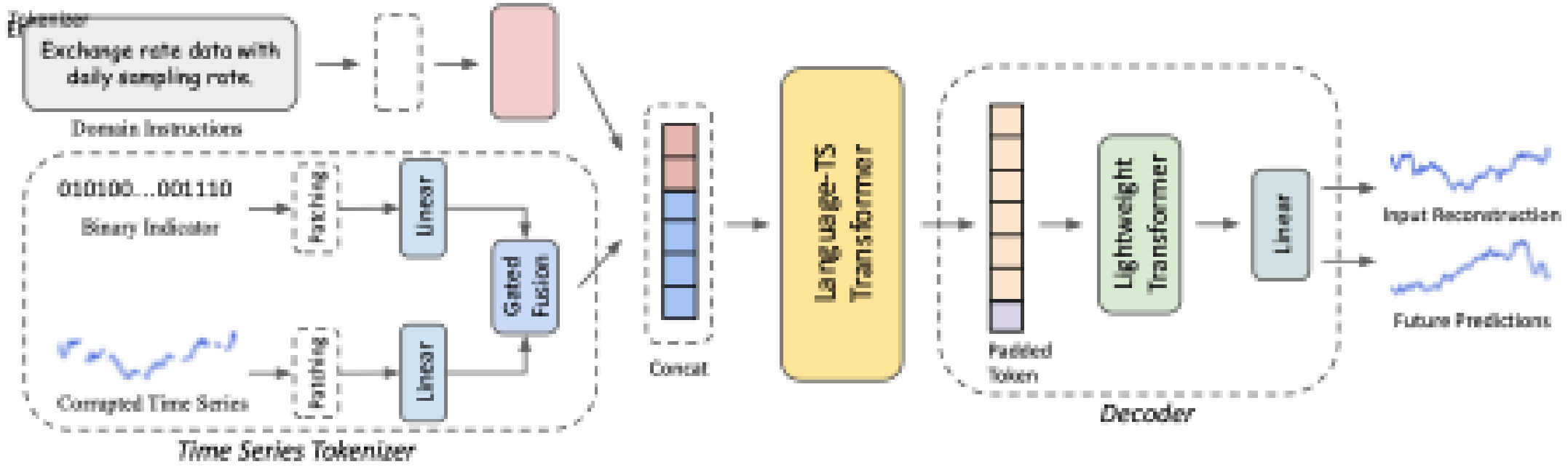



Lee et al, "TimeCAP: Learning to Contextualize, Augment, and Predict Time Series Events with Large Language Model Agents", AAAI 2025




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

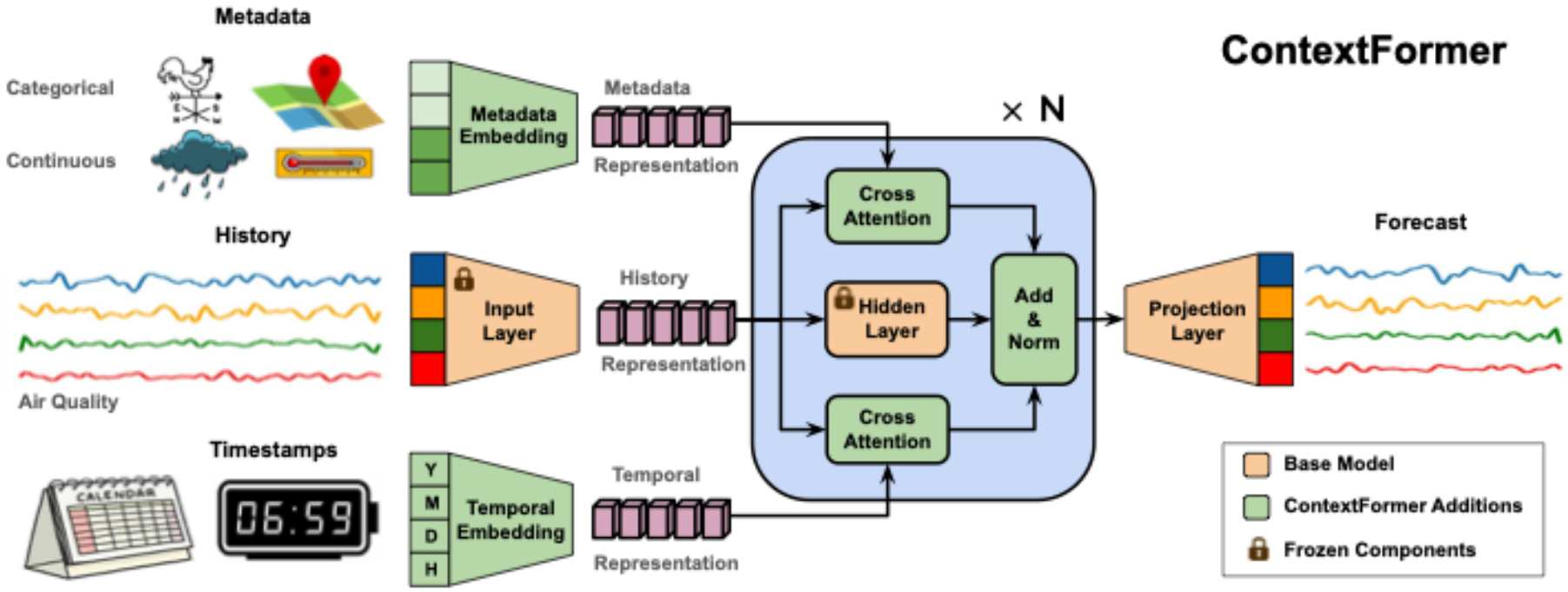
**Cross-attention**: **time series** acts as the query modality, while auxiliary modalities like **text or images,** provide context through their keys and values.

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

$$\mathrm{CrossAttention}(\boldsymbol{E}_{\mathrm{ts}}, \boldsymbol{E}_{\mathrm{c}}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}_{\mathrm{ts}}\boldsymbol{K}_{\mathrm{c}}{}^{\top}}{\sqrt{\boldsymbol{d}_{\boldsymbol{k}}}}\right)\boldsymbol{V}_{\mathrm{c}}$$

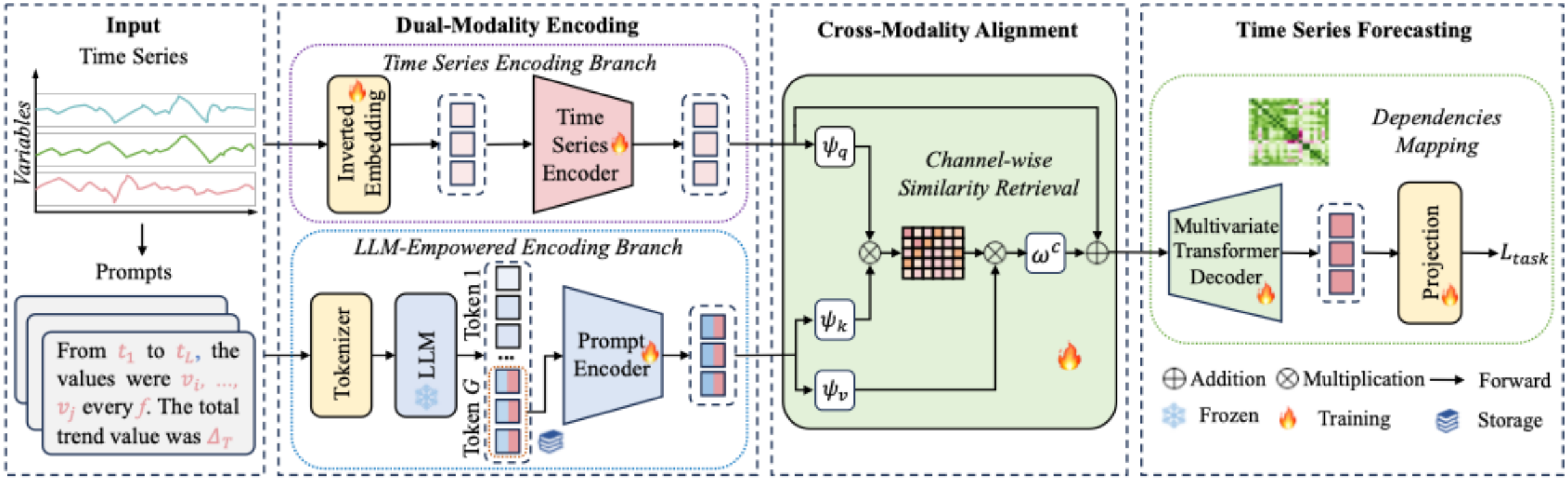
where the queries $\boldsymbol{Q}_{\mathrm{ts}}$, keys $\boldsymbol{K}_{\mathrm{c}}$, and values $\boldsymbol{V}_{\mathrm{c}}$ are linear projections of $\boldsymbol{E}_{\mathrm{ts}}$: $\boldsymbol{Q}_{\mathrm{ts}} = \boldsymbol{E}_{\mathrm{ts}}W_Q$, $\boldsymbol{K}_{\mathrm{c}} = \boldsymbol{E}_{\mathrm{ts}}W_K$, $\boldsymbol{V}_{\mathrm{c}} = \boldsymbol{E}_{\mathrm{c}}W_V$ with learnable weights $W_{Q,K,V} \in \mathbb{R}^{d \times d_k}$

# Multi-modal Alignment with Time Series – Representations



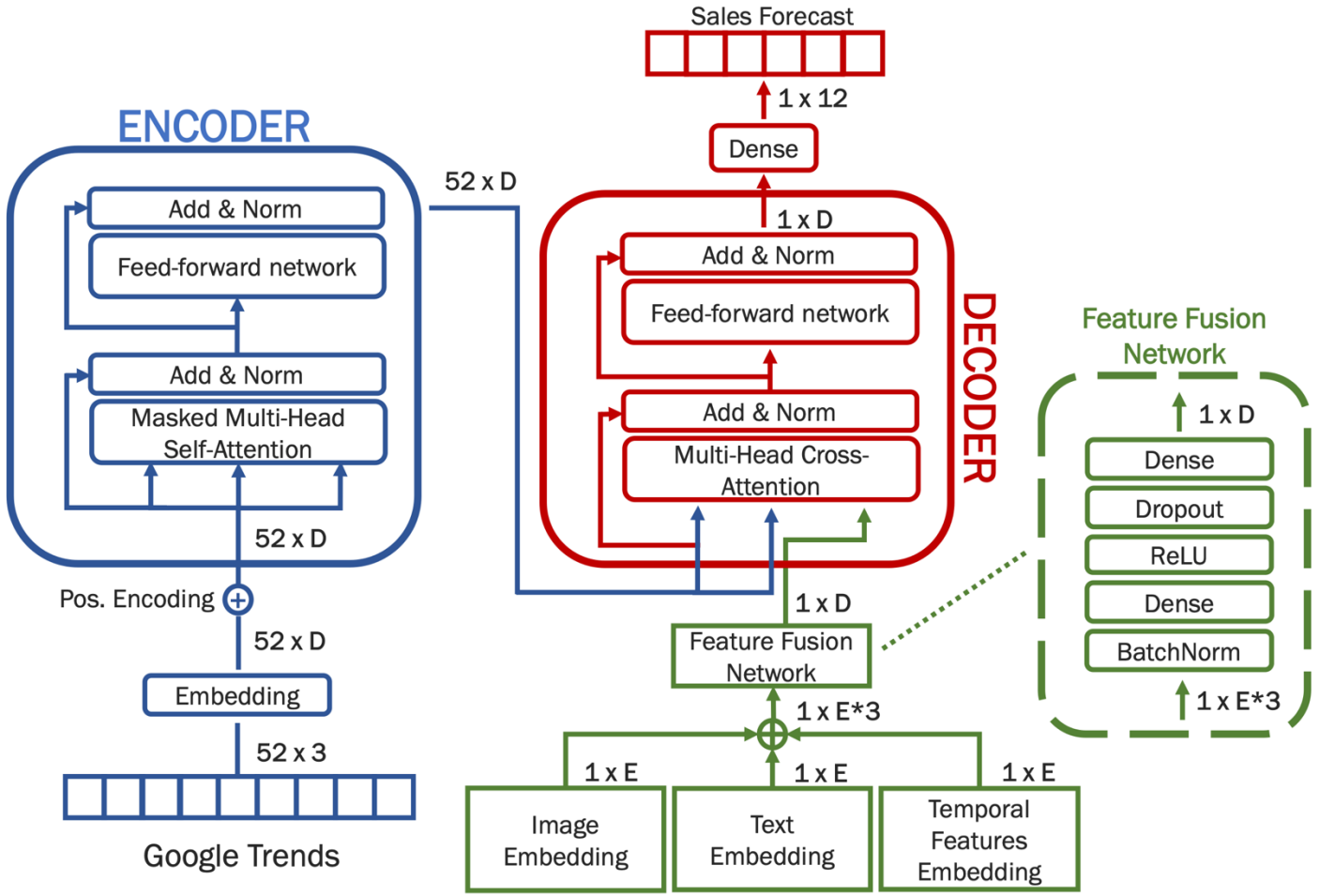

**Chattopadhyay et al. "Context Matters: Leveraging Contextual Features for Time Series Forecasting" 2025**

# 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

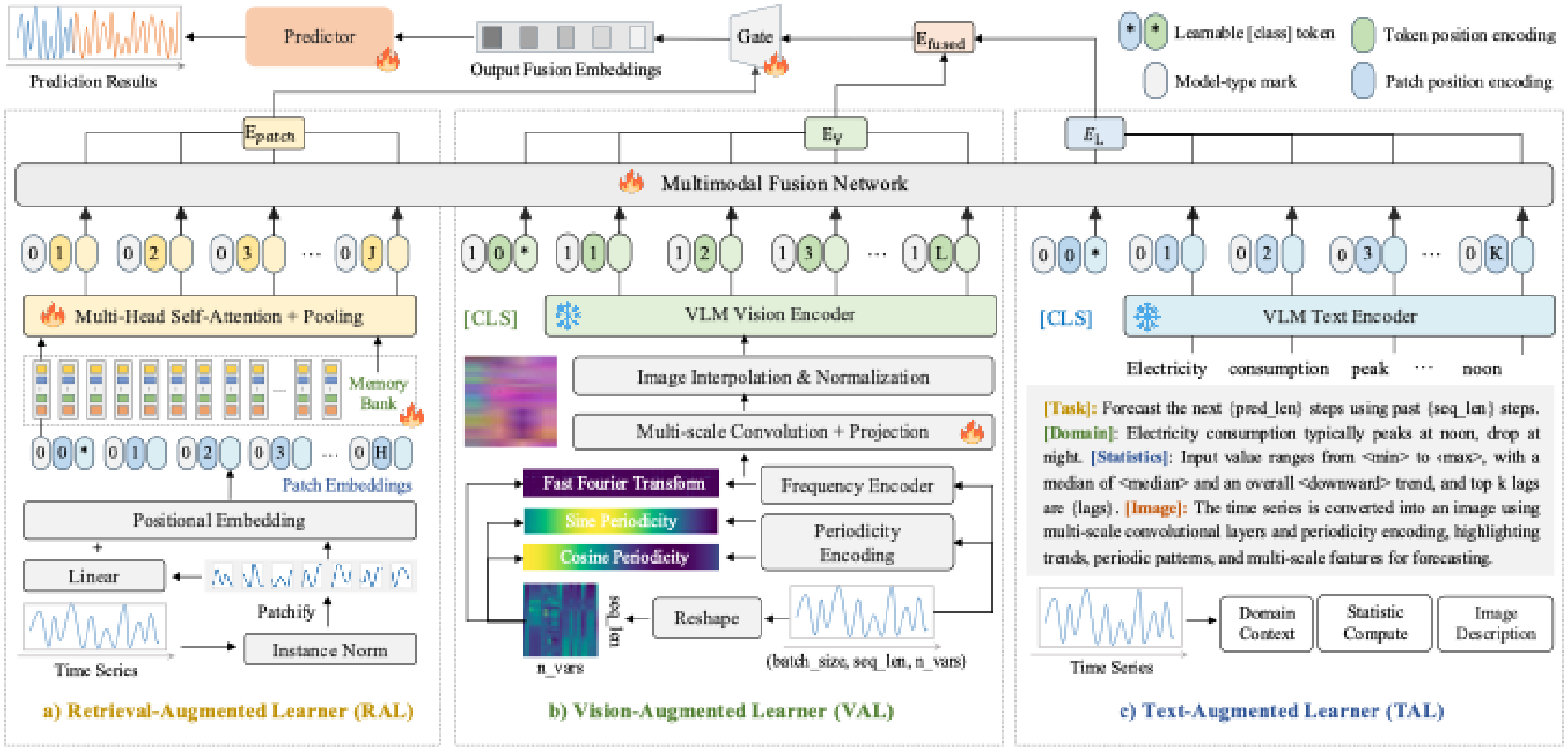
**Gating:** a parametric filtering operation that explicitly regulates the influence of time series and other modalities on the fused embeddings in $\boldsymbol{E}$.

$$G = \sigma\left(W_g[\boldsymbol{E}_{\text{ts}}; \boldsymbol{E}_c] + b_g\right)$$

$$\boldsymbol{E} = G \odot \boldsymbol{E}_{\text{ts}} + (1 - G) \odot \boldsymbol{E}_c$$

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.

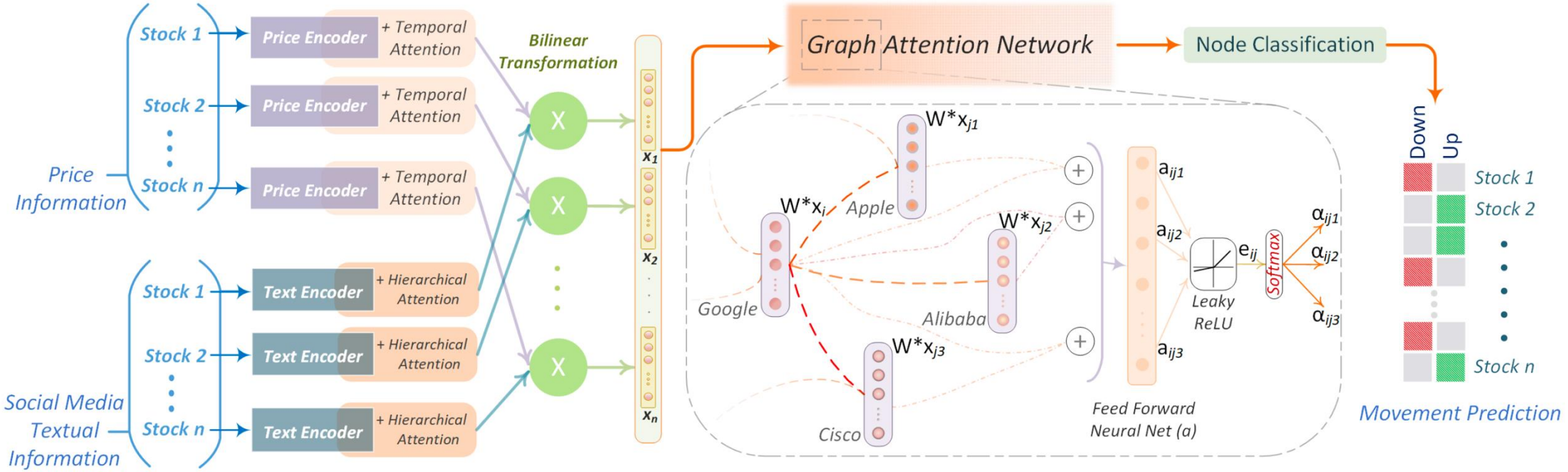
# Multi-modal Alignment with Time Series – Representations




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

# Multi-modal Alignment with Time Series – Representations

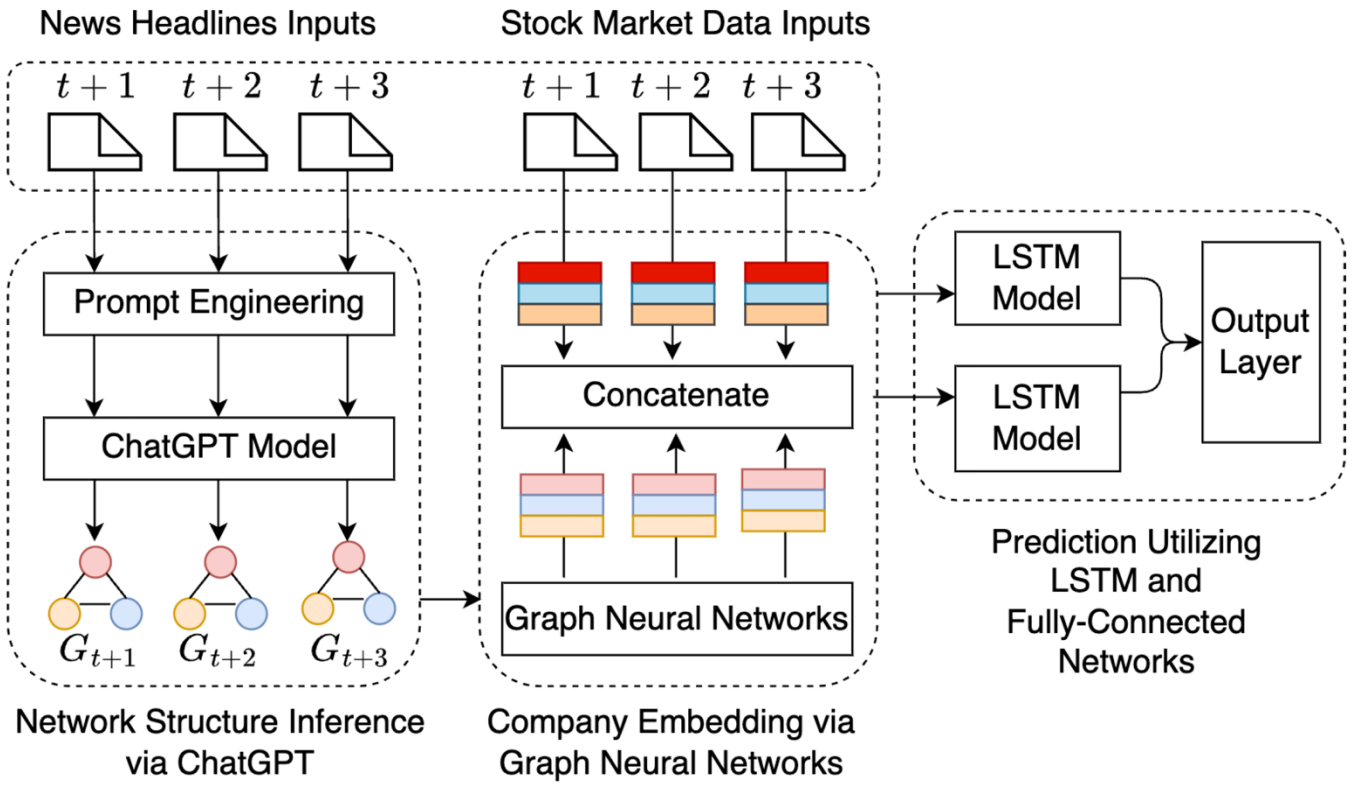
**Graph convolution:** The topological structure from external contexts can be used for alignment. It explicitly aligns representations with relational structures, enabling context-aware feature propagation across modalities.




**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:

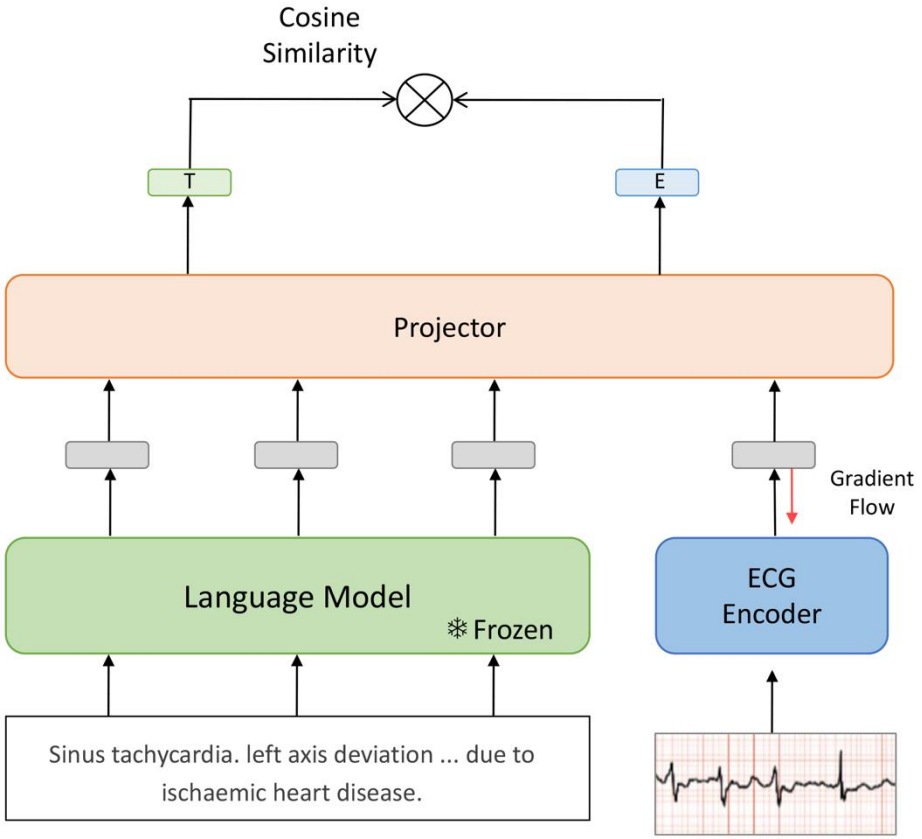
1. Identify which target companies will be impacted by these news headlines. Please list at least five of them.
2. Only consider companies from the target list.
3. Determine the sentiments of the affected companies: positive, negative, or neutral.
4. Only provide responses in JSON format, using the key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

News Headlines: ...

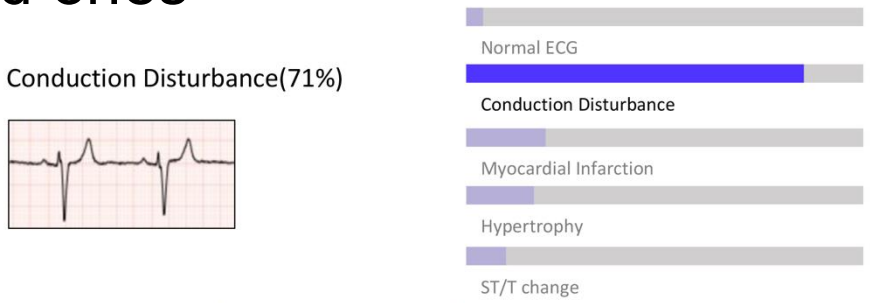
**Chen et al. "ChatGPT Informed Graph Neural Network for Stock Movement Prediction", KDD Workshop 2023**

# Multi-modal Alignment with Time Series – Representations

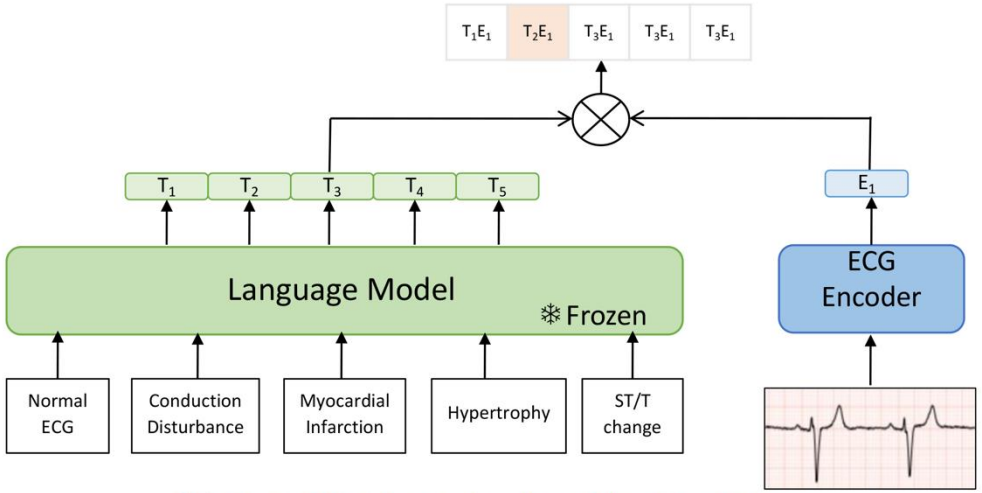
**Contrastive Learning:** maximize the cosine similarity between paired multi-modal embeddings and minimize that of unpaired ones




(a) Self-supervised Learning pre-training




(c) Visualization of Classification Results

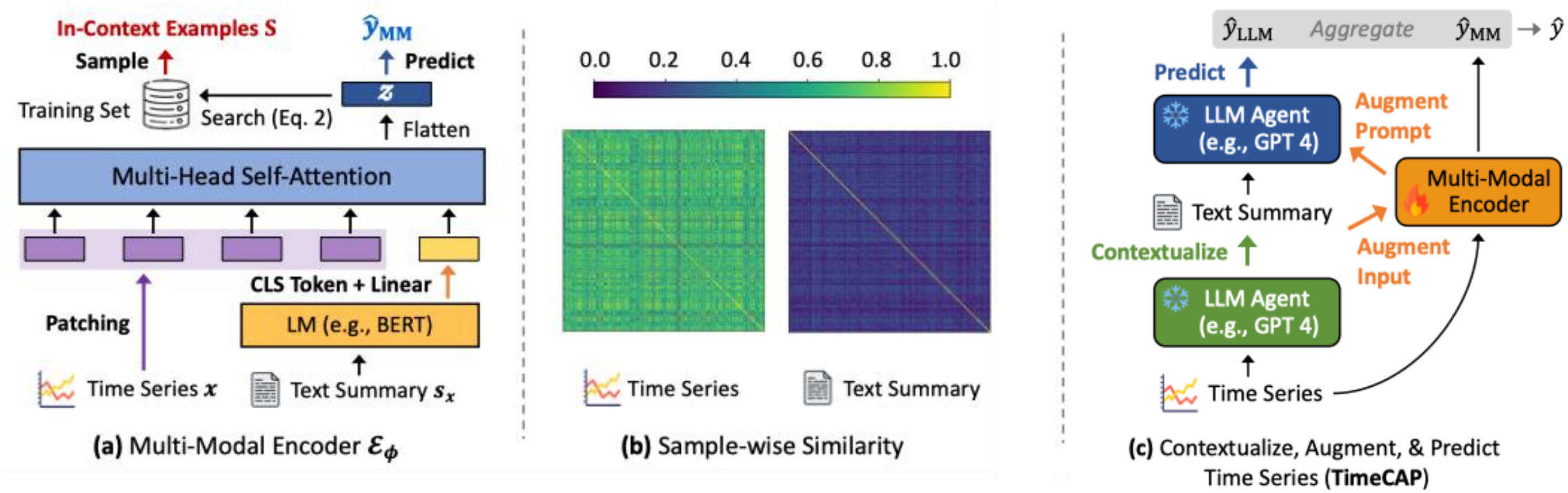



(b) Zero-Shot Learning for Classification

**Li et al. Frozen Language Model Helps ECG Zero-Shot Learning, MIDL 2023**

# Multi-modal Alignment with Time Series – Component Output

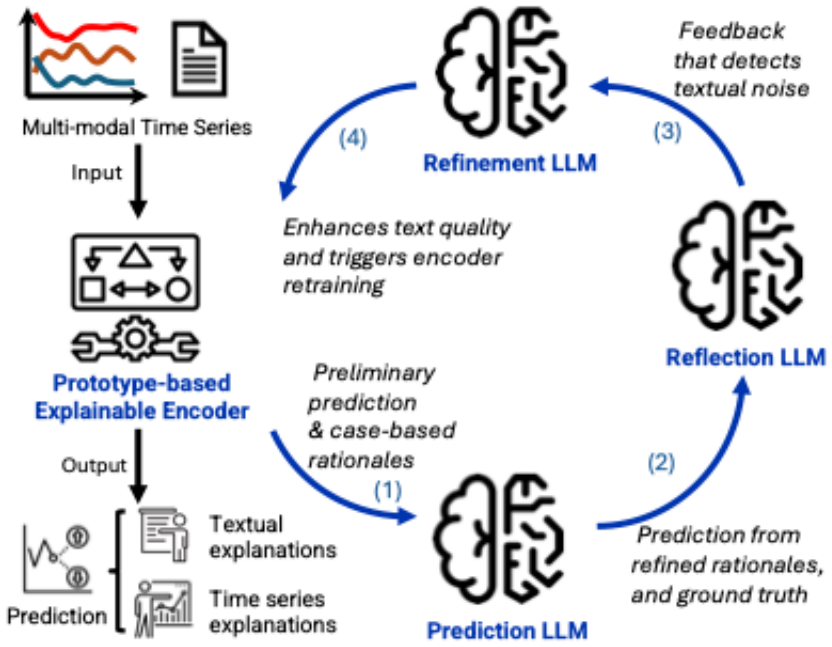
**Retrieval:** Augment LLM's input with in-context examples with the highest cosine similarity from a multi-modal embedding space

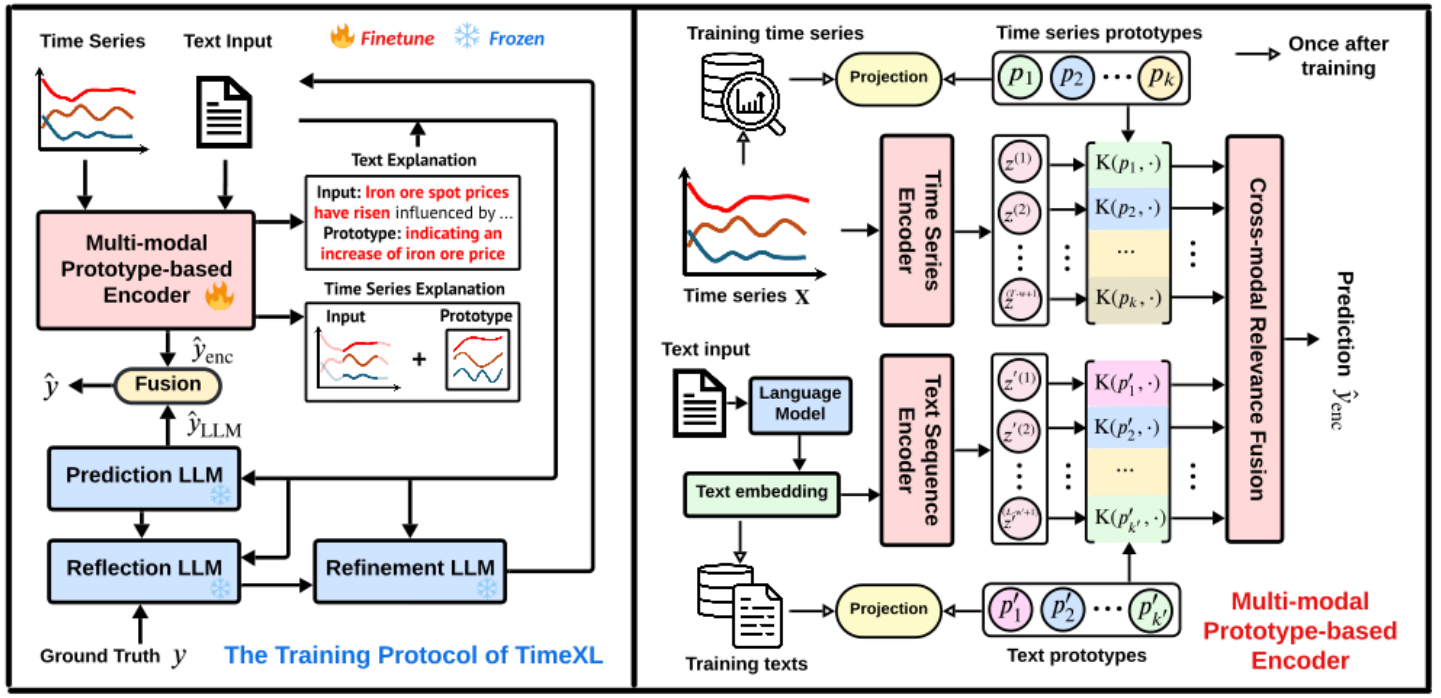



**Lee et al. "TimeCAP: Learning to Contextualize, Augment, and Predict Time Series Events with Large Language Model Agents", AAAI 2025**

# 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.




**Jiang et al, "TimeXL: Explainable Multi-modal Time Series Prediction with LLM-in-the-Loop", NeurIPS 2025**

# Multi-modal Alignment with Time Series – Component Output

**Original Text Reasoning** | **Truth: Rain** | **Prediction: Not rain**

**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.

**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)

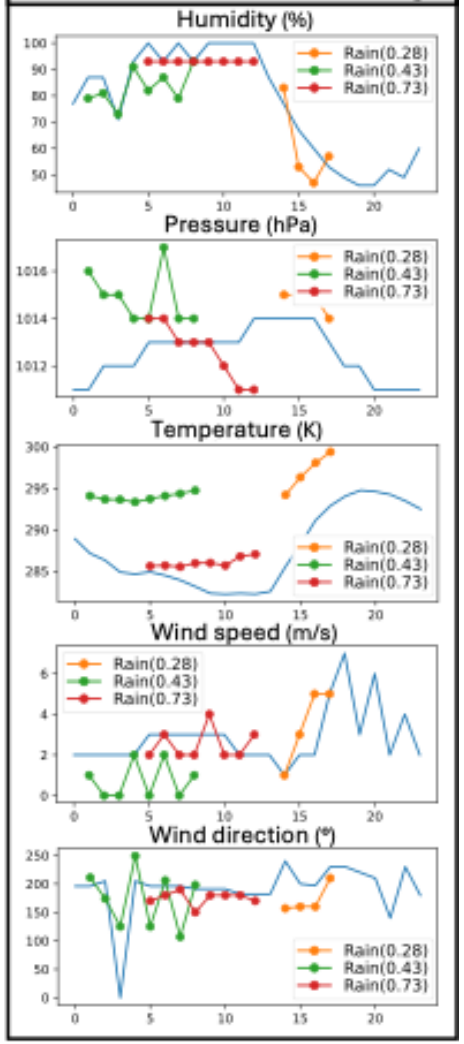
**TimeXL: Text Reasoning** | **Prediction: Rain**

**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.

**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)

**TimeXL: Time Series Reasoning**



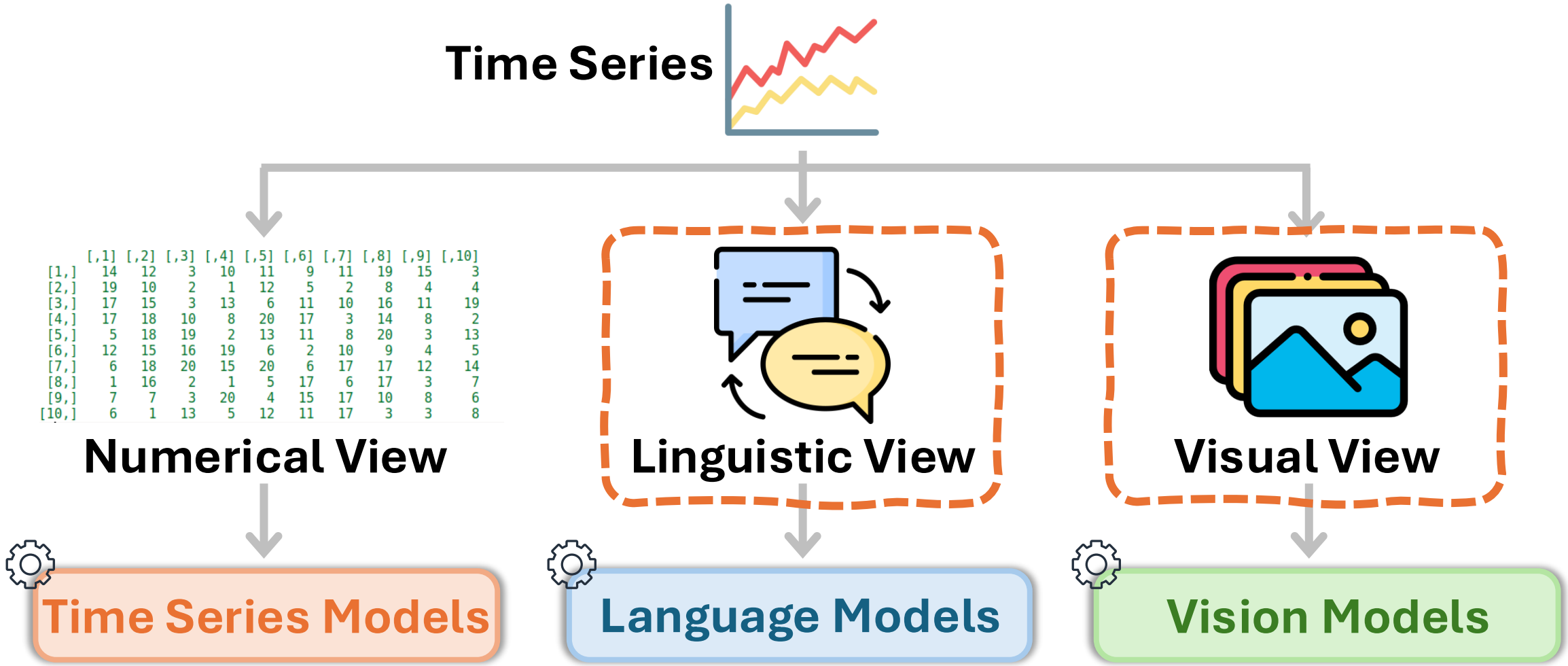

**Jiang et al, "Time XL: Explainable Multi-modal Time Series Prediction with LLM-in-the-Loop", NeurIPS 2025**

# 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 **PromptCast**[1]

| | | | Template | Example |
|---|---|---|---|---|
| CT | Input Prompt (Source) | Context | From $\{t_1\}$ to $\{t_{\text{obs}}\}$, the average temperature of region $\{U_m\}$ was $\{x^m_{t_1:t_{\text{obs}}}\}$ 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 | The temperature will be $\{x^m_{t_{\text{obs}+1}}\}$ degree. | The temperature will be 78 degree. |
| ECL | Input Prompt (Source) | Context | From $\{t_1\}$ to $\{t_{\text{obs}}\}$, client $\{U_m\}$ consumed $\{x^m_{t_1:t_{\text{obs}}}\}$ 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 $\{x^m_{t_{\text{obs}+1}}\}$ 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 $\{x^m_{t_1:t_{\text{obs}}}\}$ 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 $\{x^m_{t_{\text{obs}+1}}\}$ visitors. | There will be 15 visitors. |

**Requires dataset-specific templates**

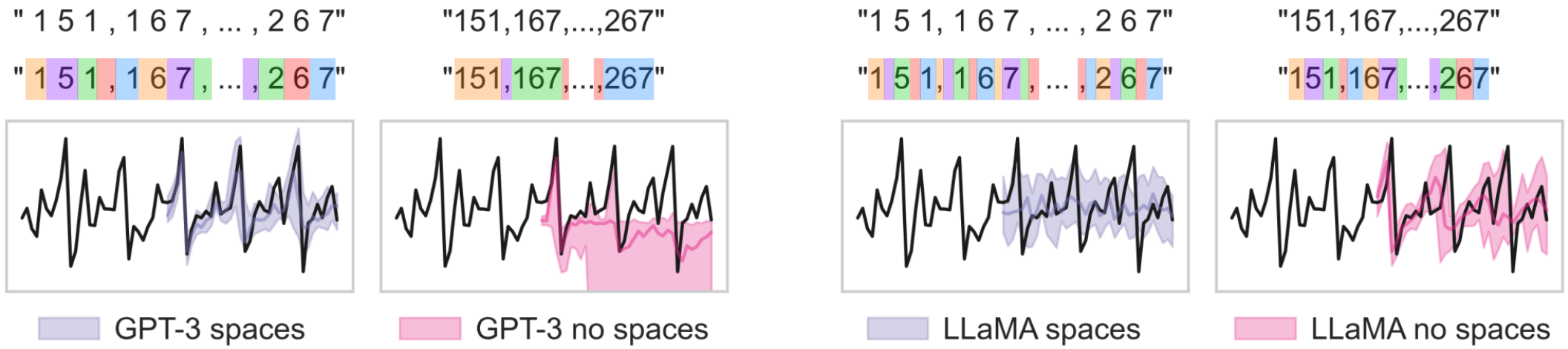
1. H. Xue, et al. "Promptcast: A new prompt-based learning paradigm for time series forecasting." IEEE TKDE, 2023.

# Linguistic View of Time Series (2)

## **LLMTime**: Verbalizing Time Series as Discrete Tokens[2]

- 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

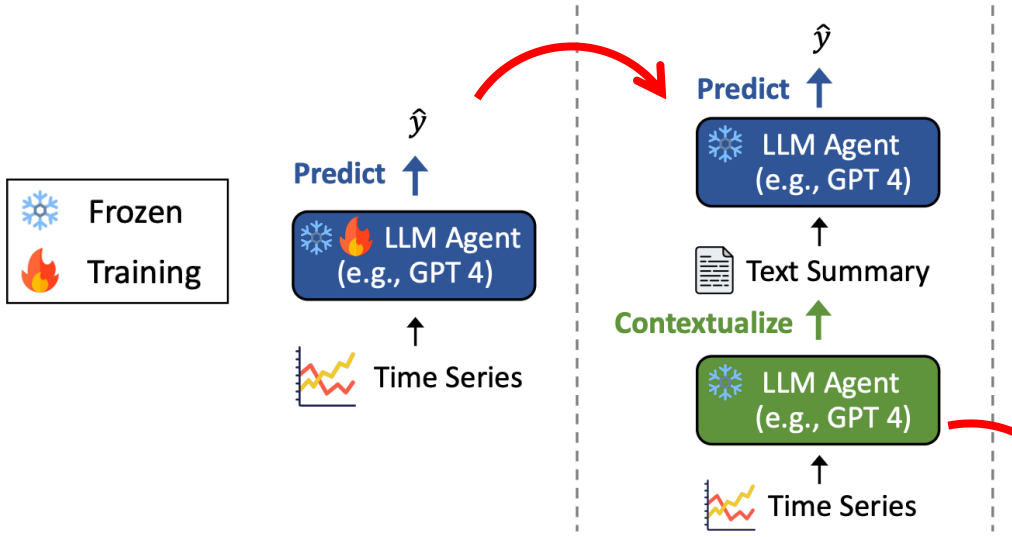
$$0.123, 1.23, 12.3, 123.0 \rightarrow \text{" 1 2 , 1 2 3 , 1 2 3 0 , 1 2 3 0 0"}$$




2. N. Gruver et al. "Large language models are zero-shot time series forecasters." In NeurIPS, 2023.

# Linguistic View of Time Series (3)

## **TimeCAP**: Summarize Time Series as Textual Description[3]




**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 '|' token for the following indicators:

[Time Series Data]

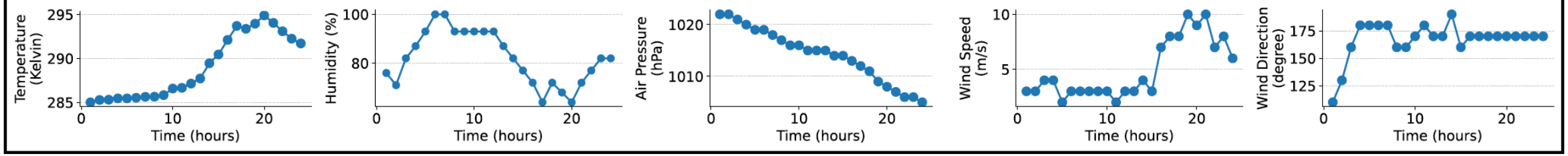
Based on this time-series data, write a concise report that provides insights crucial for understanding the current [domain] situation. Your report should be limited to five sentences, yet comprehensive, highlighting key trends and considering their potential impact on [background]. Do not write numerical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

3. G. Lee et al. "Timecap: Learning to contextualize, augment, and predict time series events with large language model agents." In AAAI, 2025.

# Linguistic View of Time Series (3)

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

**Time Series**




**Text Summary**

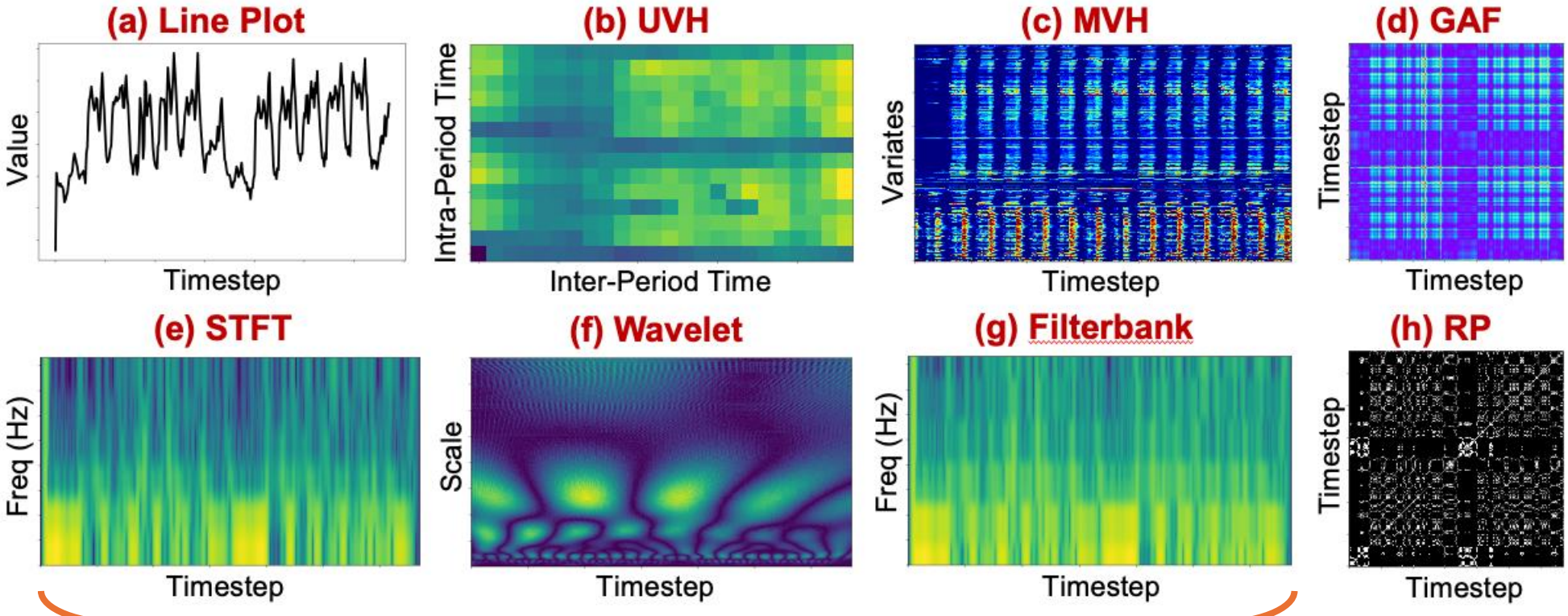
Over the past 24 hours, New York City has experienced a gradual increase in temperature, indicating a possible warming trend. Humidity levels have seen fluctuations, reaching full saturation at certain points, which could have led to foggy conditions or precipitation. Air pressure has been consistently 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.

3. G. Lee et al. "Timecap: Learning to contextualize, augment, and predict time series events with large language model agents." In AAAI, 2025.

# Visual View of Time Series

Code for the imaging methods

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

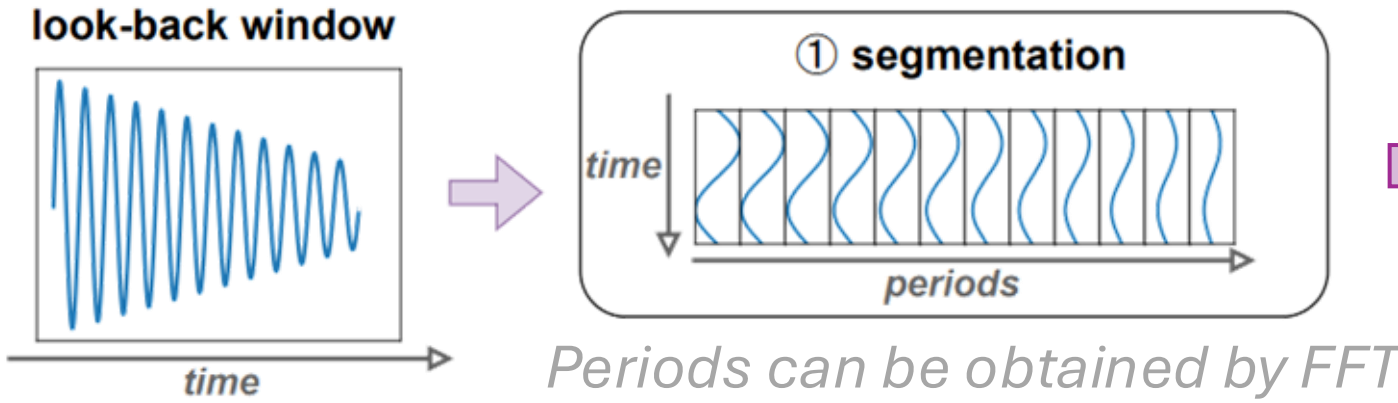



Spectrograms

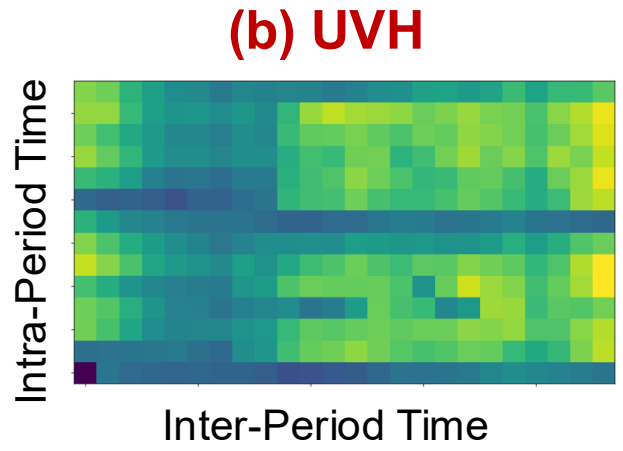
4. J. Ni, et al. "Harnessing vision models for time series analysis: A survey." In IJCAI, 2025.

# Visual View of Time Series

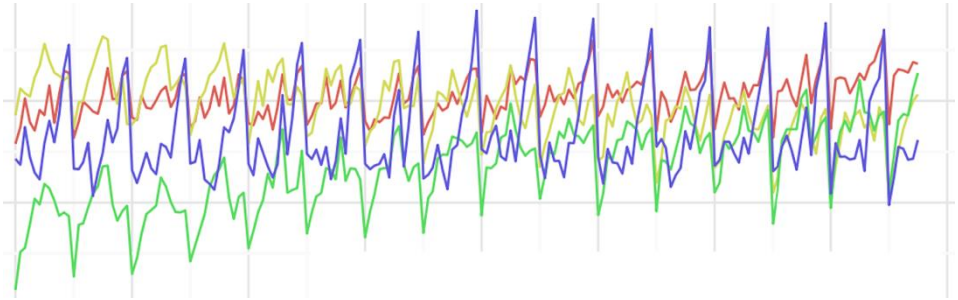
## (b) UVH – Univariate Heatmap[5,6]

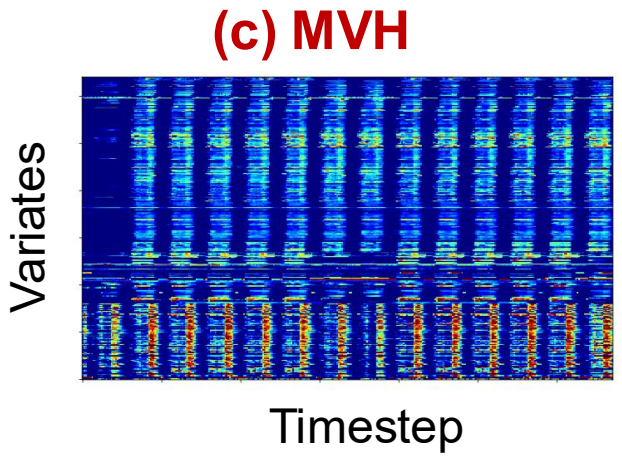



*Periods can be obtained by FFT*




## (c) MVH – Multivariate Heatmap

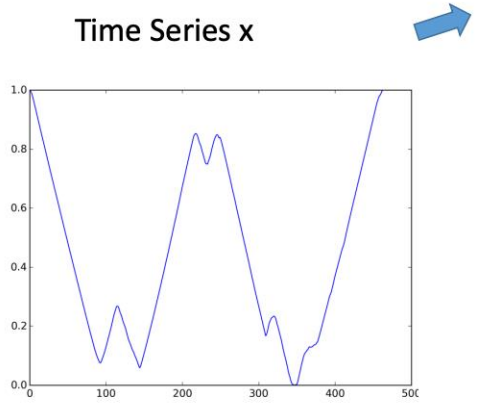


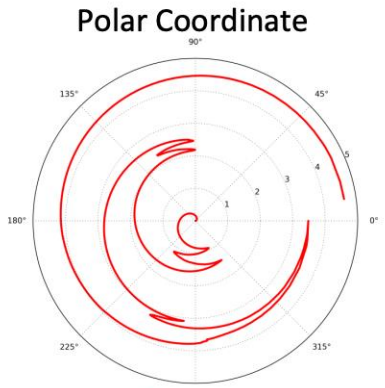



5. M. Chen, et al. "VisionTS: Visual Masked Autoencoders Are Free-Lunch Zero-Shot Time Series Forecasters." In ICML, 2025.
6. H. Wu et al. "TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis." In ICLR, 2023.

# Visual View of Time Series

## (d) GAF – Gramian Angular Field[7]




$$\begin{cases} \phi = \arccos(\tilde{x}_i), -1 \leq \tilde{x}_i \leq 1, \tilde{x}_i \in \tilde{X} \\ r = \frac{t_i}{N}, t_i \in \mathbb{N} \end{cases}$$



Gramian Angular **Summation** Field

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

GASF



GADF

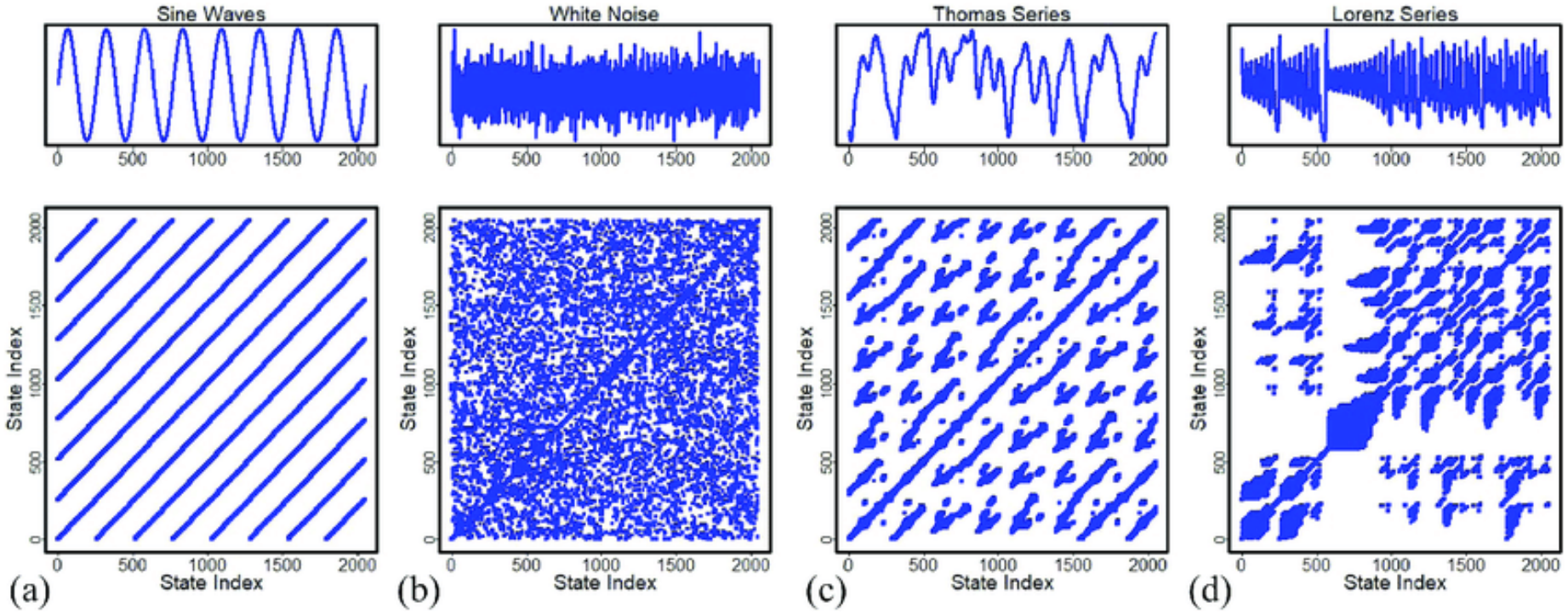
Gramian Angular **Difference** Field

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

7. Z. Wang, et al. "Imaging time-series to improve classification and imputation." In IJCAI, 2015.

# Visual View of Time Series

## (h) RP – Recurrence Plot[8]




📈 *Captures periodic patterns* 🧩 *Binary values → black-white images*

8. J. P. Eckmann et al. "Recurrence Plots of Dynamical System." Europhysics Letters, 1987.

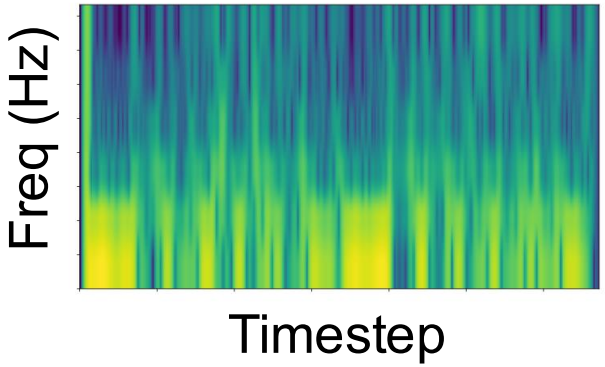
# Visual View of Time Series

## Spectrograms

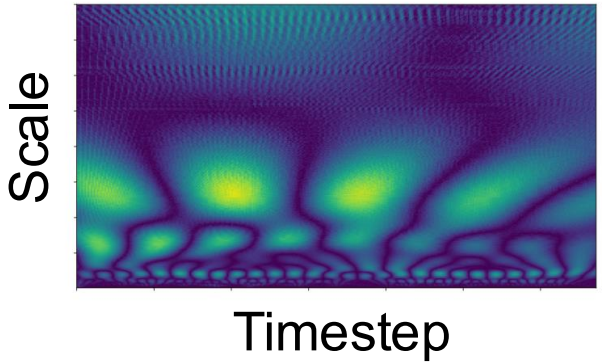
Fixed window size

Variable wavelet size

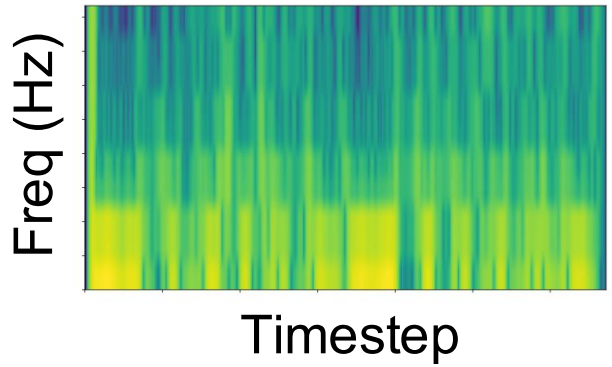
**(e) STFT[9]**




**(f) Wavelet[10]**




**(g) Filterbank[11]**




(Resembles SFTF, more suited audio signals)

✅ *Time-frequency space*

✅ *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

## Summary[4]

| 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 |
| **RP** | UTS | flexible image size | thresholding → information loss |
| **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 |

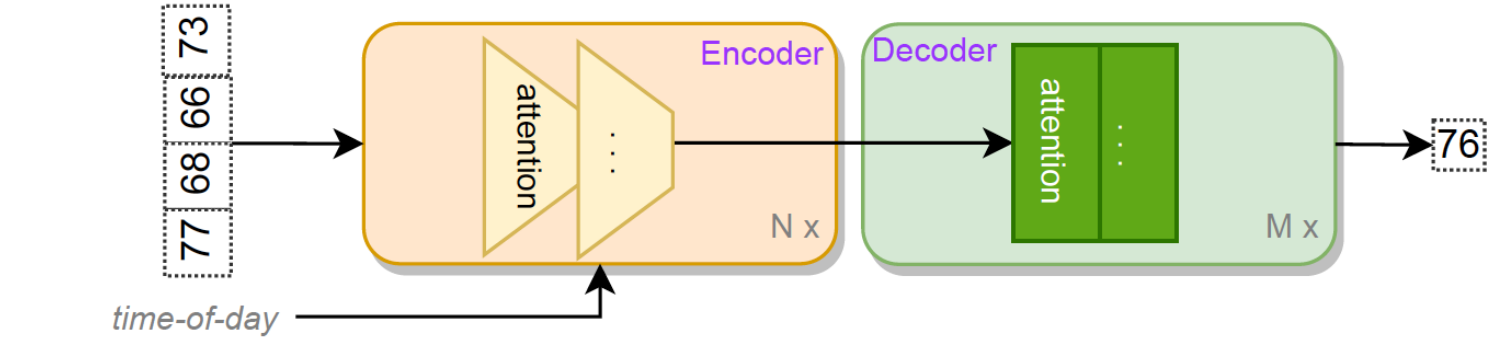
4. J. Ni, et al. "Harnessing vision models for time series analysis: A survey." In IJCAI, 2025.
12. Z. Zhao et al. "From Images to Signals: Are Large Vision Models Useful for Time Series Analysis?." arXiv preprint arXiv:2505.24030 (2025).

# Outline of This Section

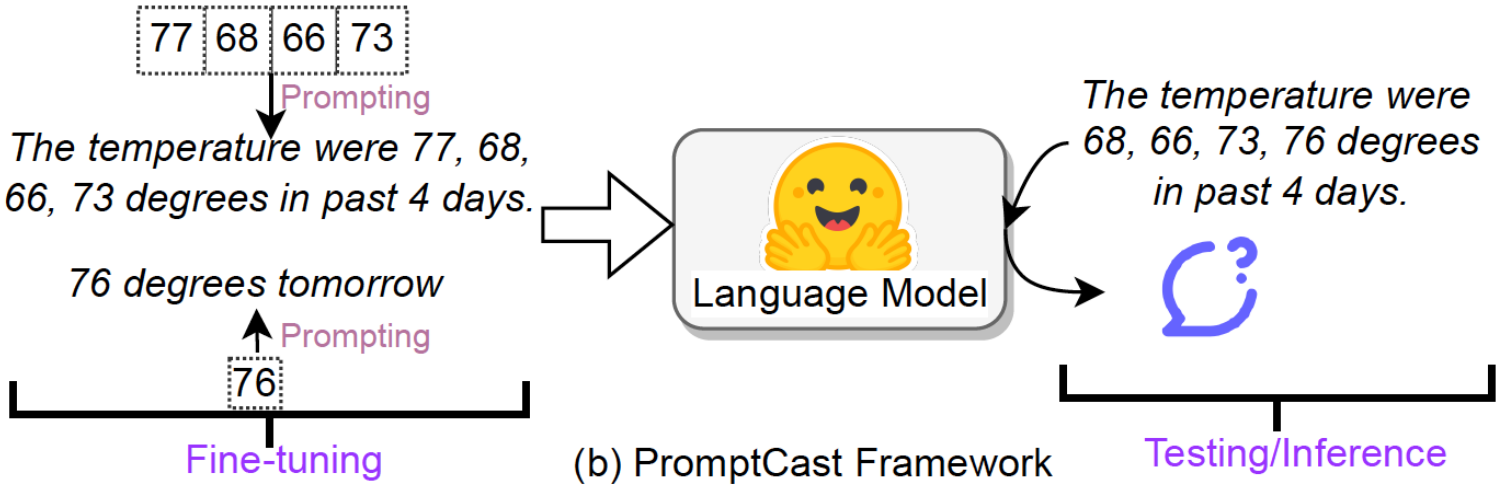
- ☑ **Generating MMVs of time series**
  - Linguistic view and visual view
- **Cross-modal knowledge transfer via MMVs**
  - Methods using LLMs and LVMs
- **Integrating MMVs of time series**
  - Combining multiple models or using LMMs

# Cross-Modal Knowledge Transfer via Linguistic View

## Forecasting as a QA problem with LLMs – **PromptCast**[1]




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

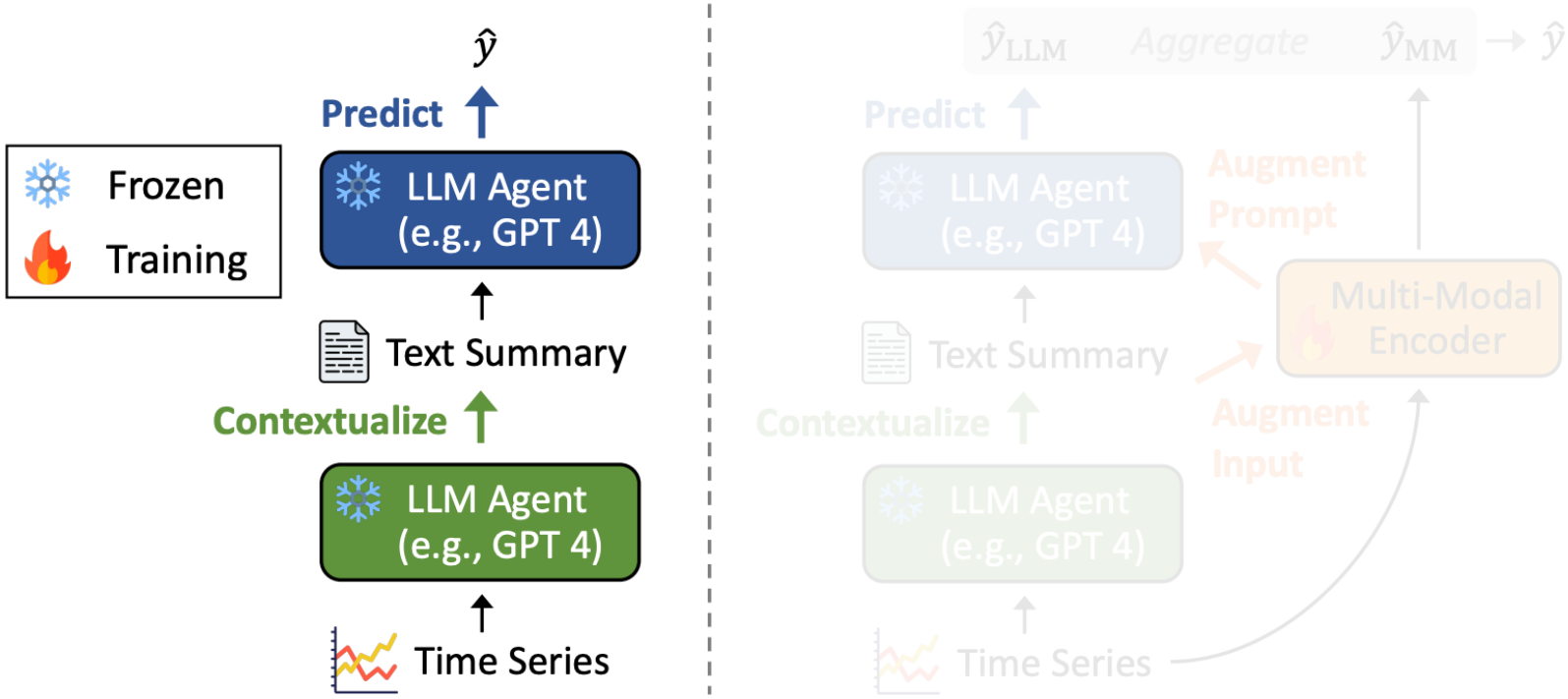



(b) PromptCast Framework

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

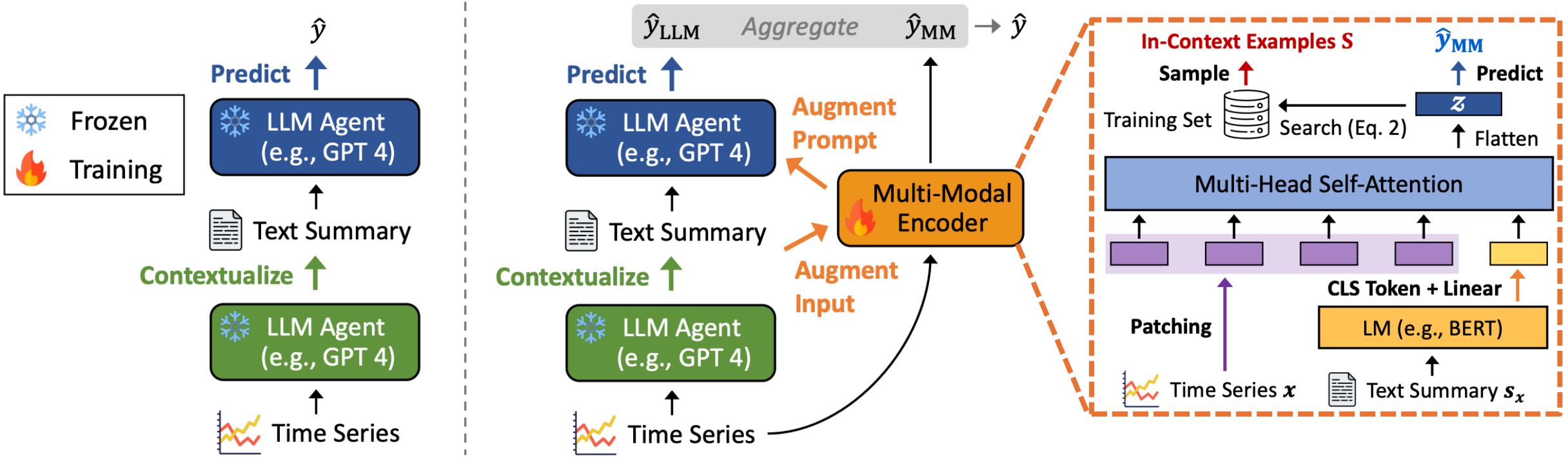
## Event detection (classification) with LLMs – **TimeCAP**[3]




3. G. Lee et al. "Timecap: Learning to contextualize, augment, and predict time series events with large language model agents." In AAAI, 2025.

# Cross-Modal Knowledge Transfer via Linguistic View

## Event detection (classification) with LLMs – **TimeCAP**[3]




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

3. G. Lee et al. "Timecap: Learning to contextualize, augment, and predict time series events with large language model agents." In AAAI, 2025.

# Summary of LLMs on Linguistic View of Time Series

✅ **Reasoning**: leveraging LLMs' reasoning capabilities

✅ **Context**: straightforward to integrate additional textual data

✅ **Explanation**: potential to provide explanation

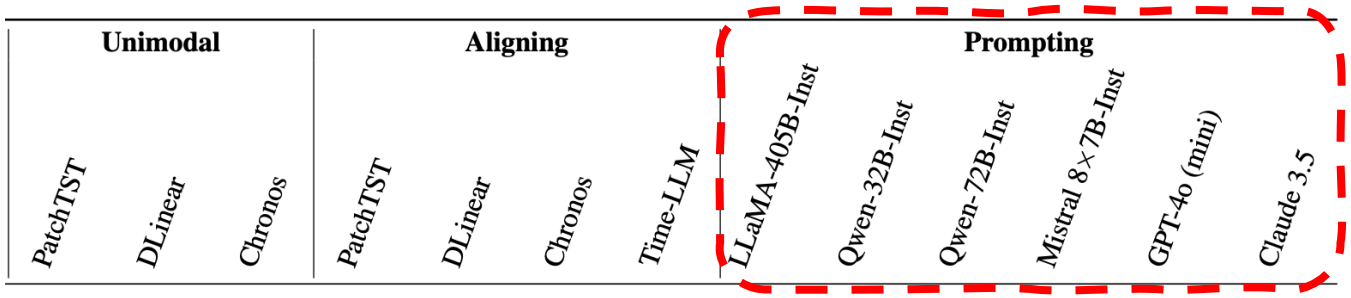
⛔ Model long time series

⛔ Model multivariate time series (e.g., spatiotemporal data)

⛔ Perform long-term forecasting

WHEN DOES MULTIMODALITY LEAD TO BETTER TIME SERIES FORECASTING?

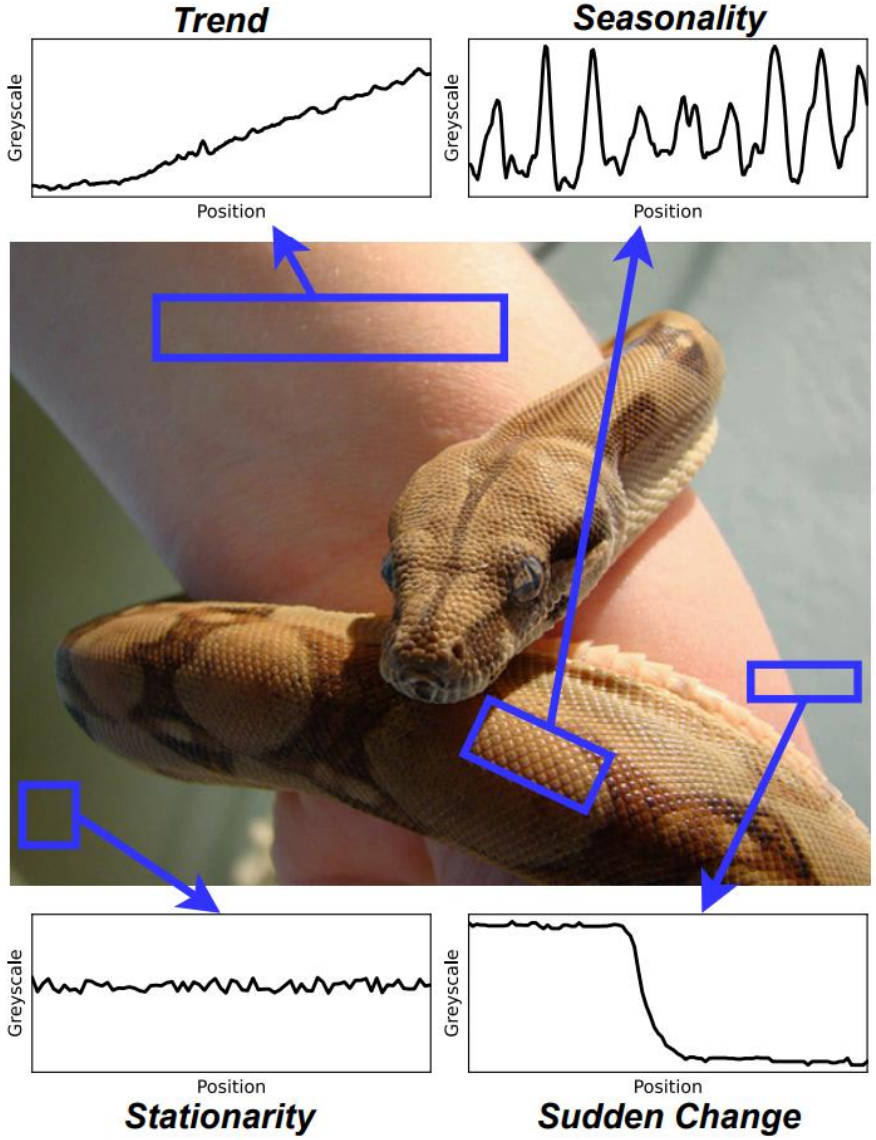
**Xiyuan Zhang*, Boran Han*, Haoyang Fang, Abdul Fatir Ansari, Shuai Zhang, Danielle C. Maddix, Cuixiong Hu, Andrew Gordon Wilson, Michael W. Mahoney, Hao Wang, Yan Liu, Huzefa Rangwala, George Karypis, Bernie Wang**
Amazon Web Services
{xiyuanz,boranhan,haoyfang,ansarnd,shuaizs,
dmmaddix,tonyhu,wilsmman,zmahmich,
howngz,yanliuyl,rhuzefa,gkarypis,yuyawang}@amazon.com



# Cross-Modal Knowledge Transfer via Visual View

*Why LVMs are potentially useful in cross-modal knowledge Transfer?*[4,5]

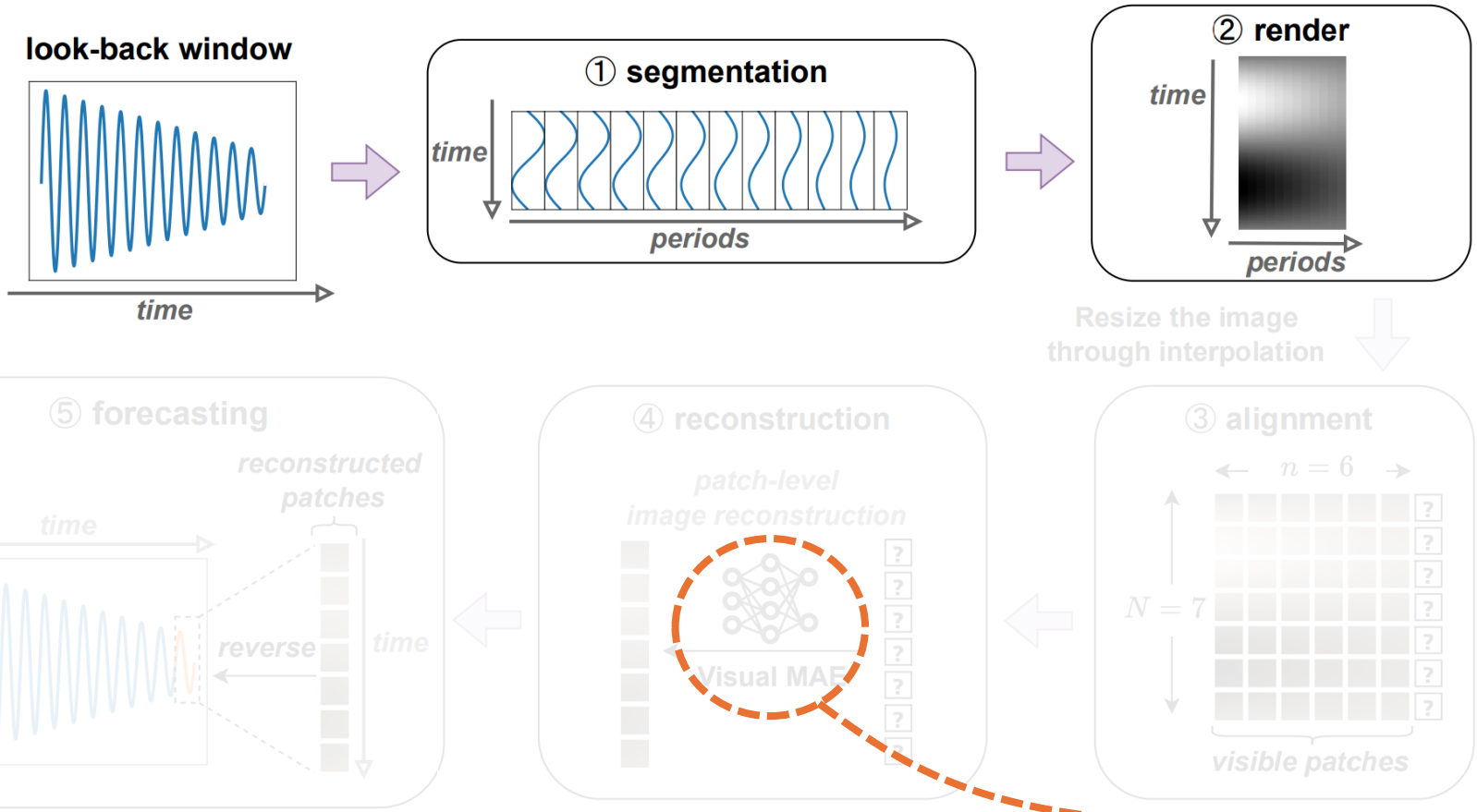
- ✅ Structural Similarity:
  - ❑ Images: continuous pixels
  - ❑ Time series: continuous values
- ✅ Large-scale imaged-based pre-training
- ✅ Multiple imaging methods
- ✅ Multivariate time series
- ✅ Long time series

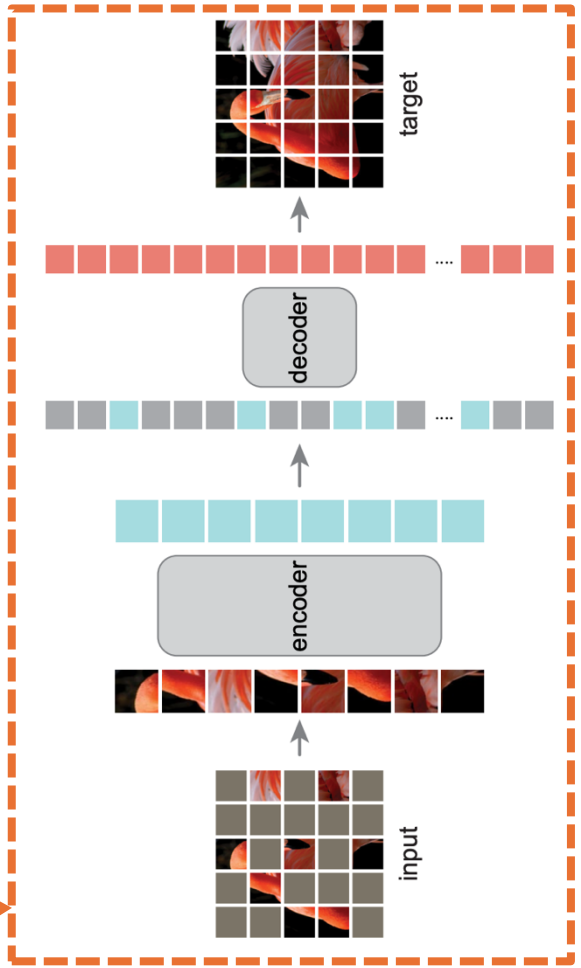



4. J. Ni, et al. "Harnessing Vision models for time series analysis: A survey." In IJCAI, 2025.
5. M. Chen, et al. "VisionTS: Visual Masked Autoencoders Are Free-Lunch Zero-Shot Time Series Forecasters." In ICML, 2025.

# Cross-Modal Knowledge Transfer via Visual View

## Time Series Forecasting with LVMs – **VisionTS**[5]





5. M. Chen, et al. "VisionTS: Visual Masked Autoencoders Are Free-Lunch Zero-Shot Time Series Forecasters." In ICML, 2025.
13. K. He et al. "Masked autoencoders are scalable vision learners." In CVPR, 2022.

# Cross-Modal Knowledge Transfer via Visual View

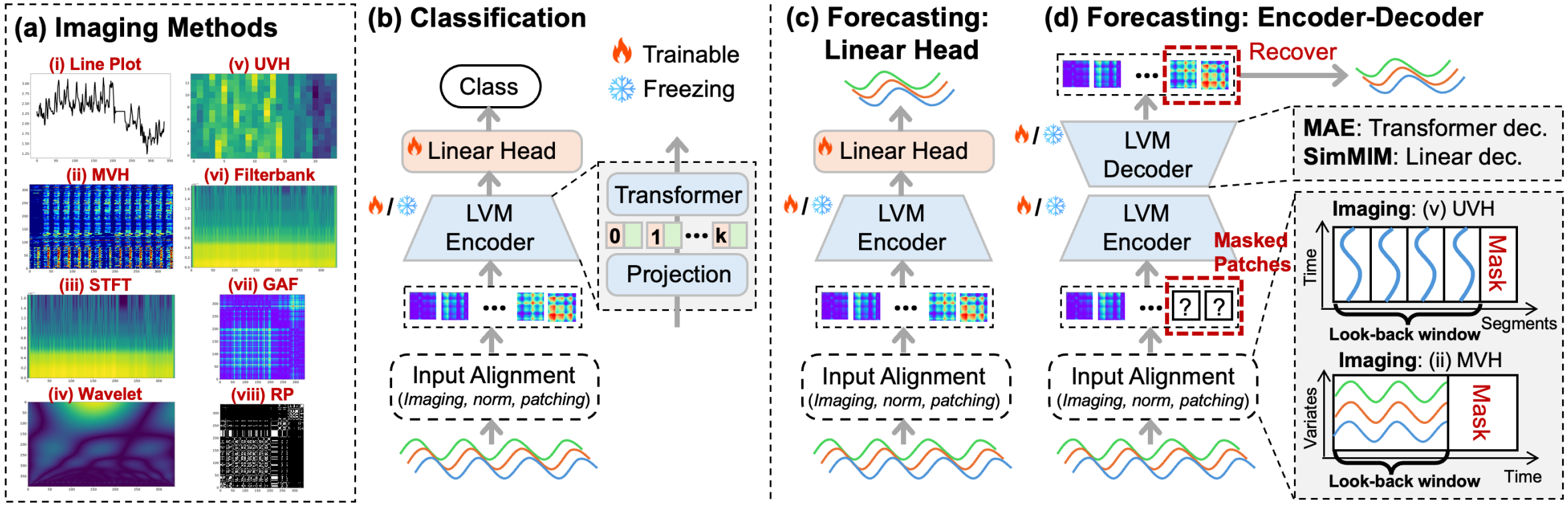
## VisionTS[5] – Zero-Shot Time Series Forecasting

| | | Zero-Shot | | | | Few-Shot (10% In-distribution Downstream Dataset) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pretrain | | *Images* | *Time series* | | | *Text* | | *No Pretrain* | | | | |
| Method | | 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 |
| **1st count** | | 7 | 0 | 3 | 1 | 2 | 1 | 0 | 0 | 0 | 0 | 0 |

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

# 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*)?[12]




12. Z. Zhao et al. "From Images to Signals: Are Large Vision Models Useful for Time Series Analysis?." arXiv preprint arXiv:2505.24030 (2025).

# Are LVMs Useful for Time Series Analysis?

## 📊 A Comprehensive Study[12]

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

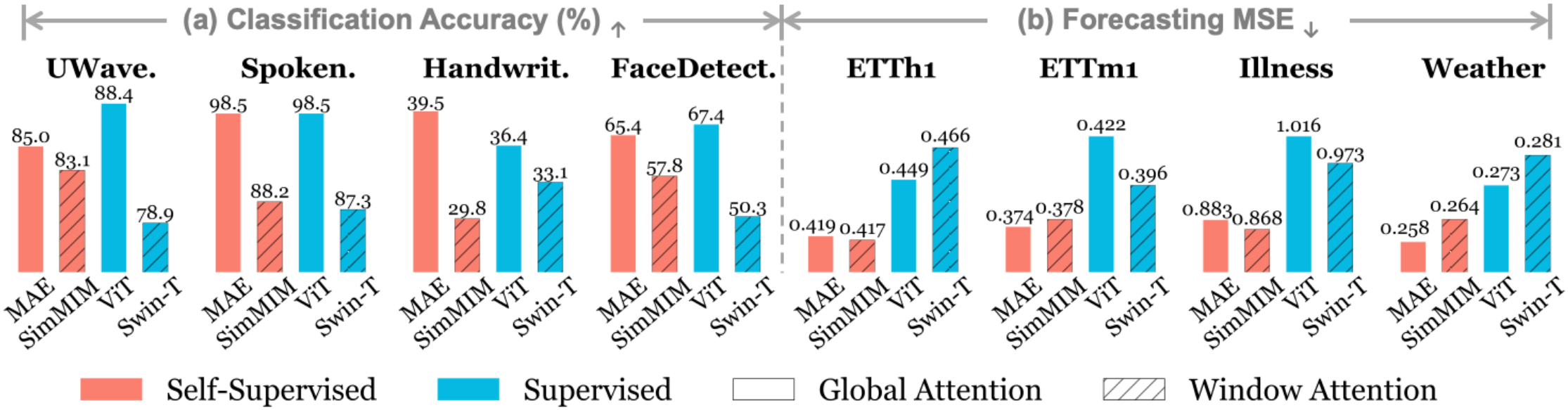
## 💡 Key Conclusions

- ❑ Generally useful for **classification**
- ❑ Challenging for **forecasting**
  - Limited to specific types of LVMs and imaging methods
  - Bias toward forecasting periods

12. Z. Zhao et al. "From Images to Signals: Are Large Vision Models Useful for Time Series Analysis?." arXiv preprint arXiv:2505.24030 (2025).

# Are LVMs Useful for Time Series Analysis?

**Insights**[12] – What type of LVM best fits classification (forecasting) task?




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

12. Z. Zhao et al. "From Images to Signals: Are Large Vision Models Useful for Time Series Analysis?." arXiv preprint arXiv:2505.24030 (2025).

# Are LVMs Useful for Time Series Analysis?

## Comparing *fine-tuned* LVMs with non-LVM baselines on forecasting[12]

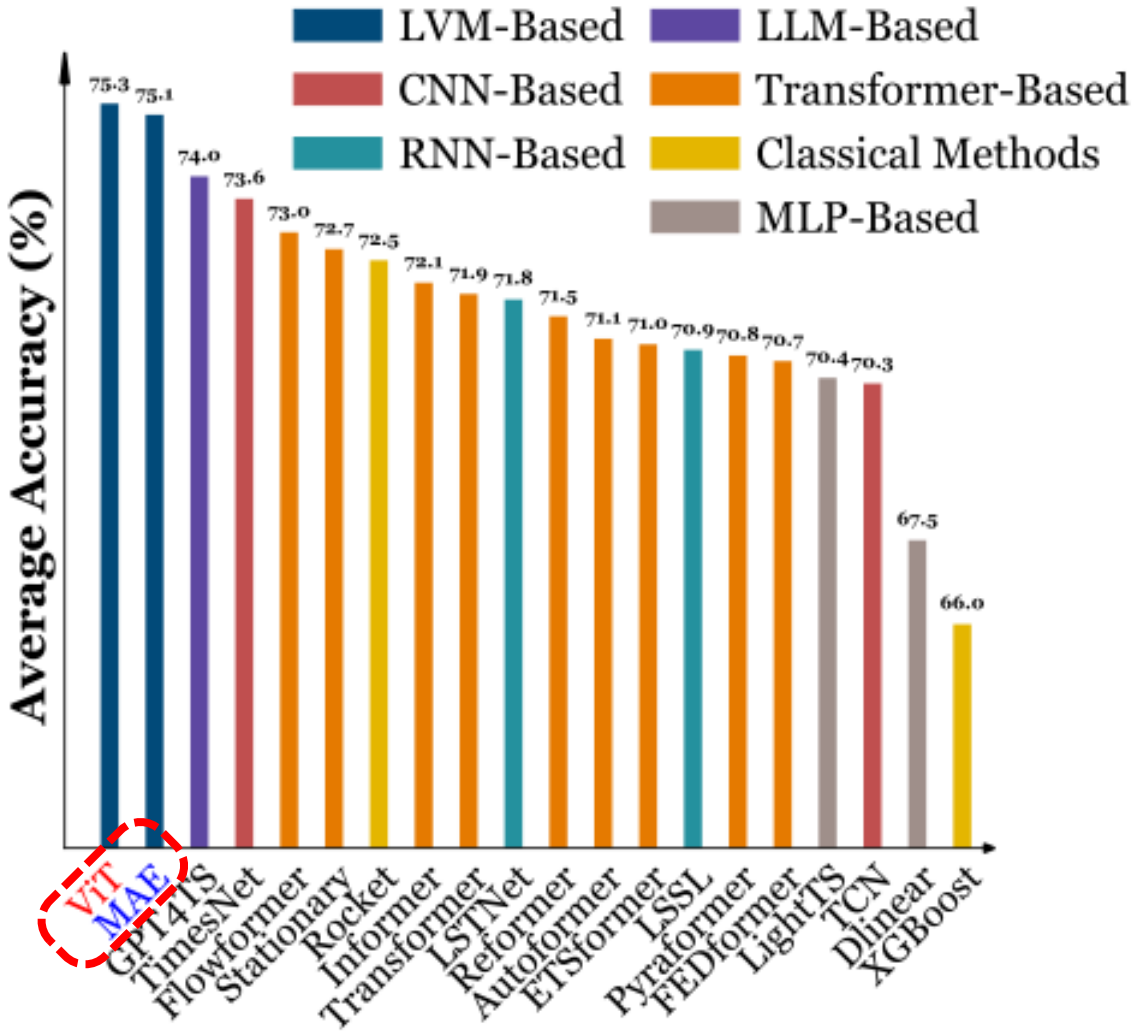
LVMs | LLMs | Linear | Transformer

| Method | MAE | | ViT | | Time-LLM | | GPT4TS | | CALF | | Dlinear | | PatchTST | | TimesNet | | FEDformer | | Autoformer | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Metrics** | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE |
| ETTh1 | 0.409 | 0.419 | 0.445 | 0.449 | 0.418 | 0.432 | 0.418 | 0.421 | 0.432 | 0.431 | 0.423 | 0.437 | 0.413 | 0.431 | 0.458 | 0.450 | 0.440 | 0.460 | 0.496 | 0.487 |
| ETTh2 | 0.357 | 0.390 | 0.389 | 0.411 | 0.361 | 0.396 | 0.354 | 0.389 | 0.351 | 0.384 | 0.431 | 0.447 | 0.330 | 0.379 | 0.414 | 0.427 | 0.437 | 0.449 | 0.450 | 0.459 |
| ETTm1 | 0.345 | 0.374 | 0.409 | 0.422 | 0.356 | 0.377 | 0.363 | 0.378 | 0.396 | 0.391 | 0.357 | 0.379 | 0.351 | 0.381 | 0.400 | 0.406 | 0.448 | 0.452 | 0.588 | 0.517 |
| ETTm2 | 0.268 | 0.327 | 0.300 | 0.337 | 0.261 | 0.316 | 0.254 | 0.311 | 0.283 | 0.323 | 0.267 | 0.334 | 0.255 | 0.315 | 0.291 | 0.333 | 0.305 | 0.349 | 0.327 | 0.371 |
| Weather | 0.225 | 0.258 | 0.234 | 0.273 | 0.244 | 0.270 | 0.227 | 0.255 | 0.251 | 0.274 | 0.249 | 0.300 | 0.226 | 0.264 | 0.259 | 0.287 | 0.309 | 0.360 | 0.338 | 0.382 |
| Illness | 1.837 | 0.883 | 2.179 | 1.016 | 2.018 | 0.894 | 1.871 | 0.852 | 1.700 | 0.869 | 2.169 | 1.041 | 1.443 | 0.798 | 2.139 | 0.931 | 2.847 | 1.144 | 3.006 | 1.161 |
| Traffic | 0.386 | 0.256 | 0.430 | 0.343 | 0.422 | 0.281 | 0.421 | 0.274 | 0.444 | 0.284 | 0.434 | 0.295 | 0.391 | 0.264 | 0.620 | 0.336 | 0.610 | 0.376 | 0.628 | 0.379 |
| Electricity | 0.159 | 0.250 | 0.173 | 0.266 | 0.165 | 0.259 | 0.170 | 0.263 | 0.176 | 0.266 | 0.166 | 0.264 | 0.162 | 0.253 | 0.193 | 0.295 | 0.214 | 0.327 | 0.227 | 0.338 |
| **# Wins** | **9** | | 0 | | 0 | | 3 | | 0 | | 0 | | **4** | | 0 | | 0 | | 0 | |

12. Z. Zhao et al. "From Images to Signals: Are Large Vision Models Useful for Time Series Analysis?." arXiv preprint arXiv:2505.24030 (2025).

# Are LVMs Useful for Time Series Analysis?

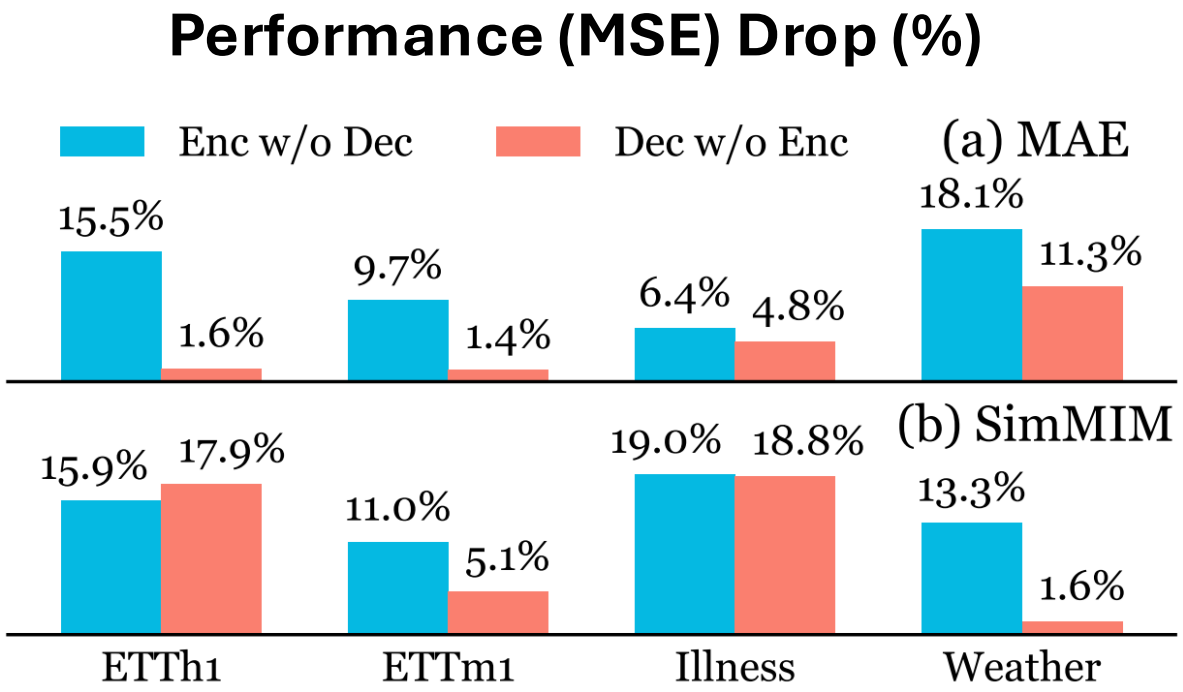
**Comparing *fine-tuned* LVMs with non-LVM baselines on classification**[12]




12. Z. Zhao et al. "From Images to Signals: Are Large Vision Models Useful for Time Series Analysis?." arXiv preprint arXiv:2505.24030 (2025).

# Are LVMs Useful for Time Series Analysis?

**Insights**[12] – Why self-supervised LVMs are useful for **forecasting**?




💡 Decoder contributes more than Encoder

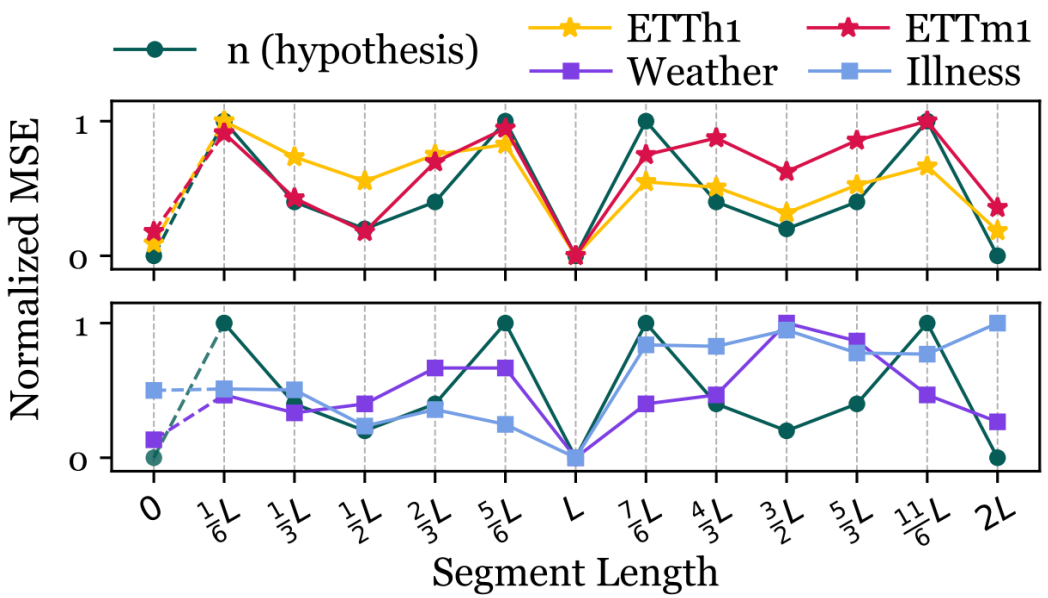
- ❑ SimMIM's decoder: only 3.8% of all parameters

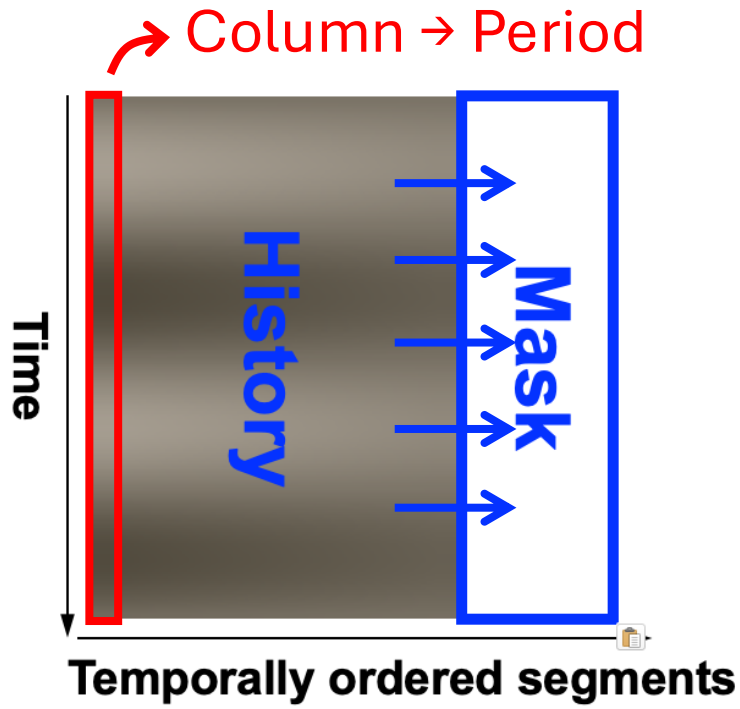
12. Z. Zhao et al. "From Images to Signals: Are Large Vision Models Useful for Time Series Analysis?." arXiv preprint arXiv:2505.24030 (2025).

# Are LVMs Useful for Time Series Analysis?

**Insights**[12] – 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

12. Z. Zhao et al. "From Images to Signals: Are Large Vision Models Useful for Time Series Analysis?." arXiv preprint arXiv:2505.24030 (2025).

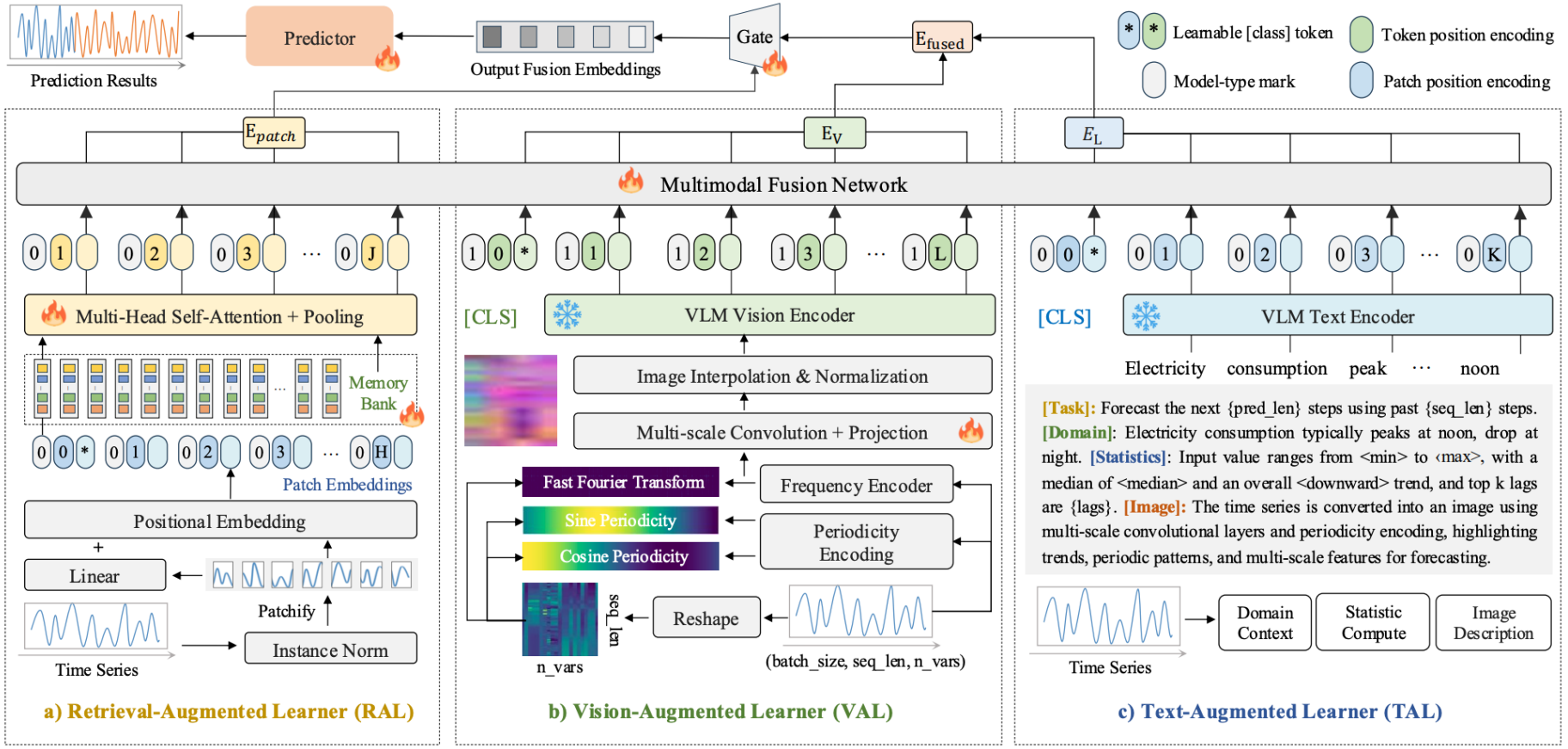
# Outline of This Section

- ☑ **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

# Integrating MMVs of Time Series

## Integrating **numerical**, **visual** views and **contexts** – TimeVLM[14]

- **Vision-Language Model (VLM)**
  - ViLT
- **Imaging**
  - Frequency-periodicity encoding
- **Contexts**
  - Not a linguistic view

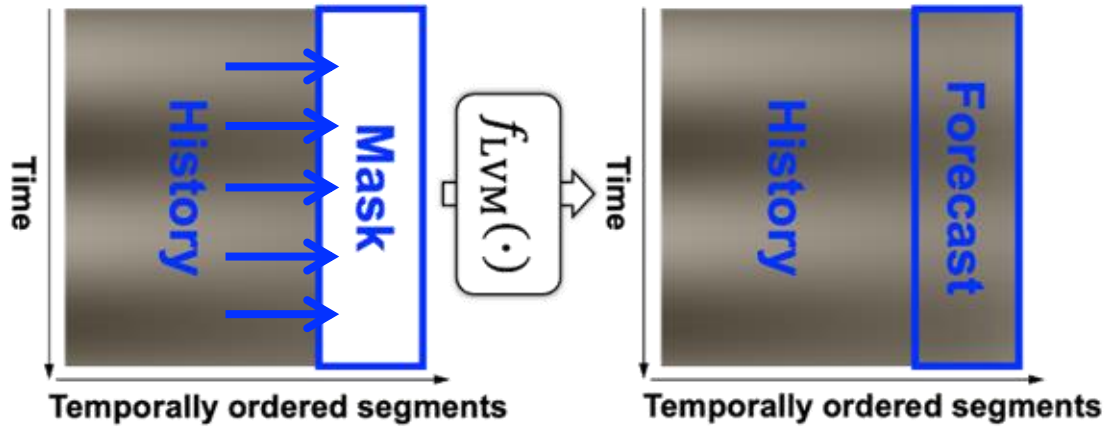


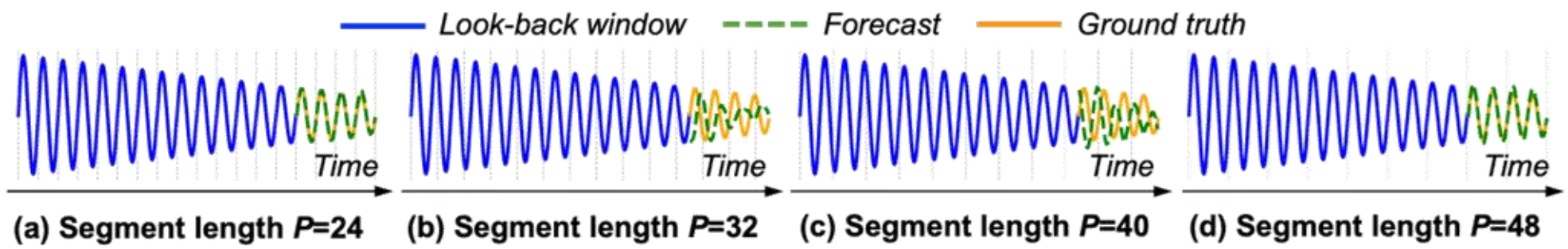

14. S. Zhong et al. "Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting." In ICML, 2025.

# Integrating MMVs of Time Series

## Integrating **numerical** and **visual** views – **DMMV**[15]

*Leveraging a bias of LVM forecaster*

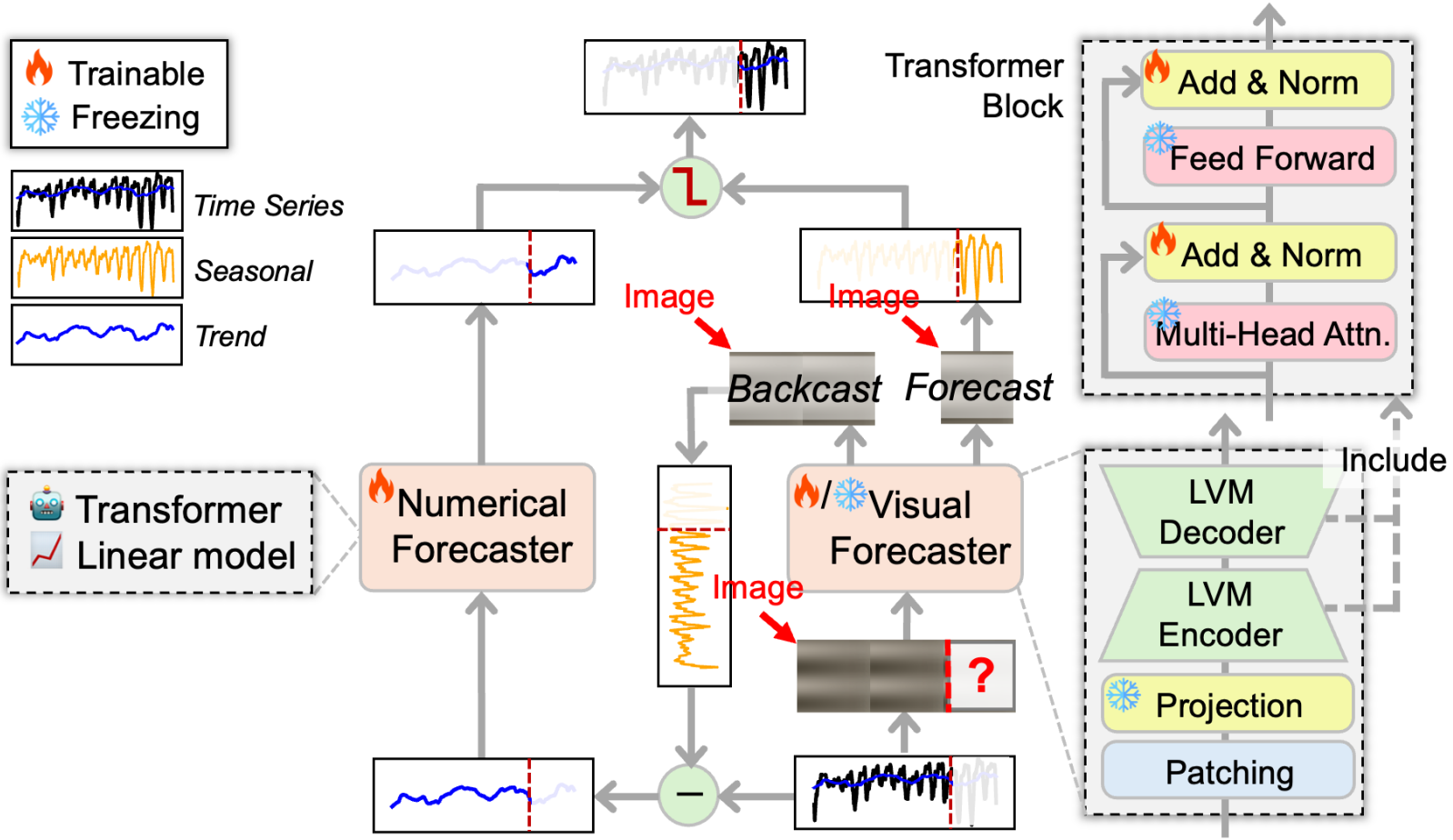



15. C. Shen et al. "Multi-Modal View Enhanced Large Vision Models for Long-Term Time Series Forecasting." In NeurIPS, 2025.

# Integrating MMVs of Time Series

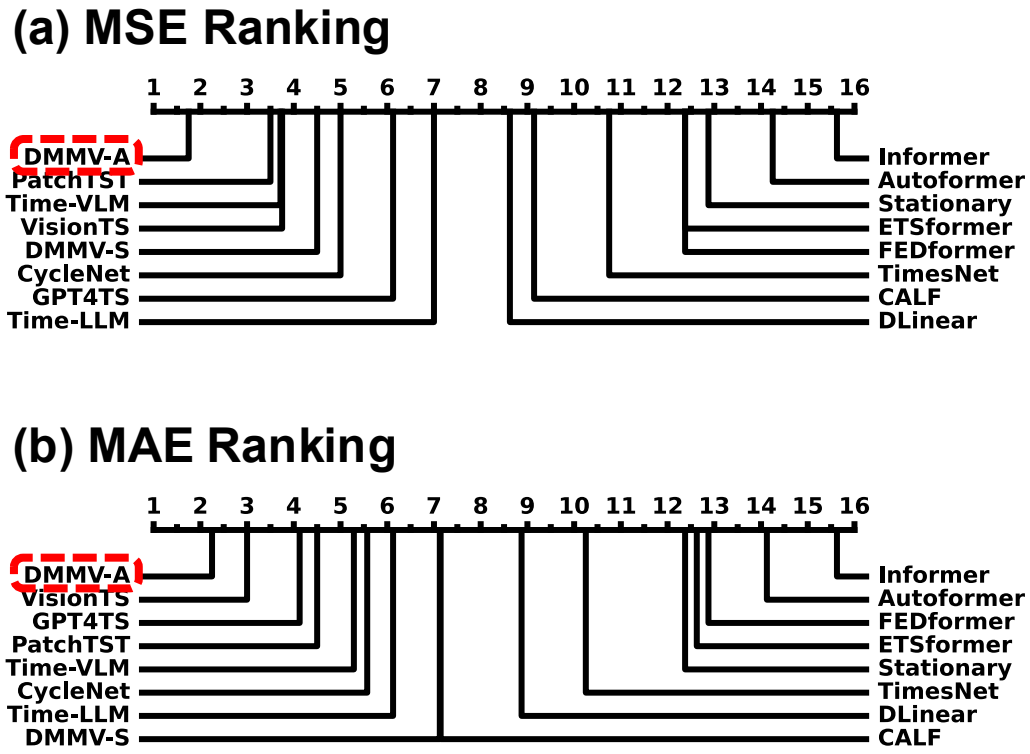
## Integrating numerical and visual views – DMMV[15]




15. C. Shen et al. "Multi-Modal View Enhanced Large Vision Models for Long-Term Time Series Forecasting." In NeurIPS, 2025.

# Integrating MMVs of Time Series

## DMMV[15] – Long-Term Time Series Forecasting

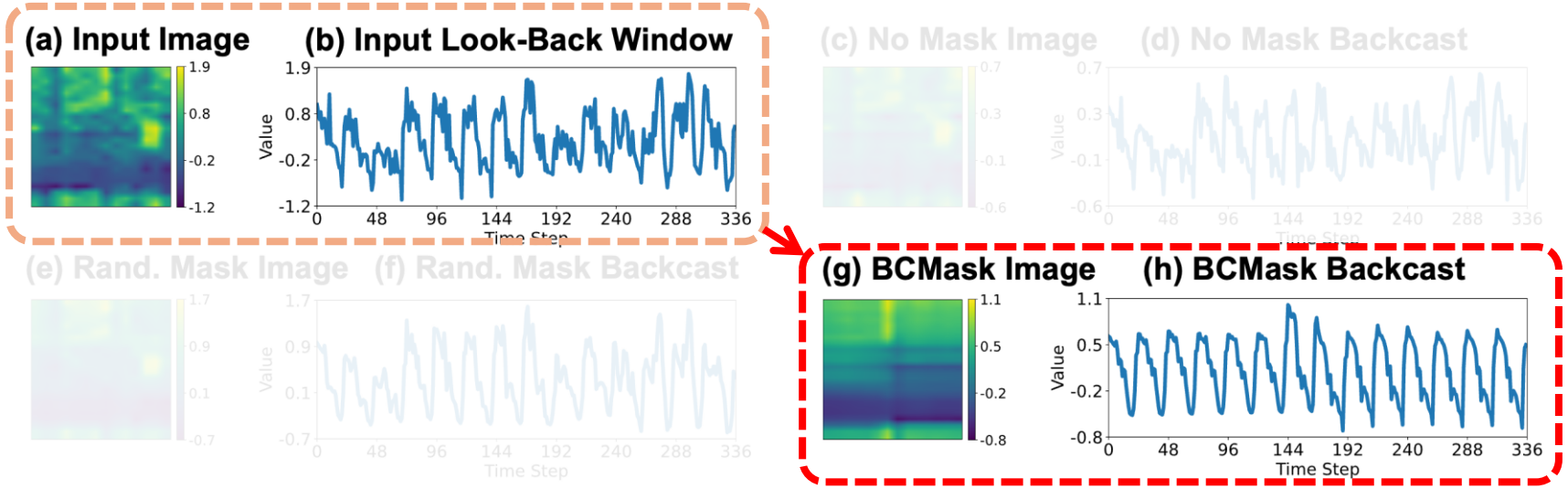



| View | | Multi-Modal | | | | Visual | | Language | | | | Numerical | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | | DMMV-A | | Time-VLM | | VisionTS | | GPT4TS | | Time-LLM | | PatchTST | | CycleNet | | TimesNet | | DLinear | | 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 | |

15. C. Shen et al. "Multi-Modal View Enhanced Large Vision Models for Long-Term Time Series Forecasting." In NeurIPS, 2025.

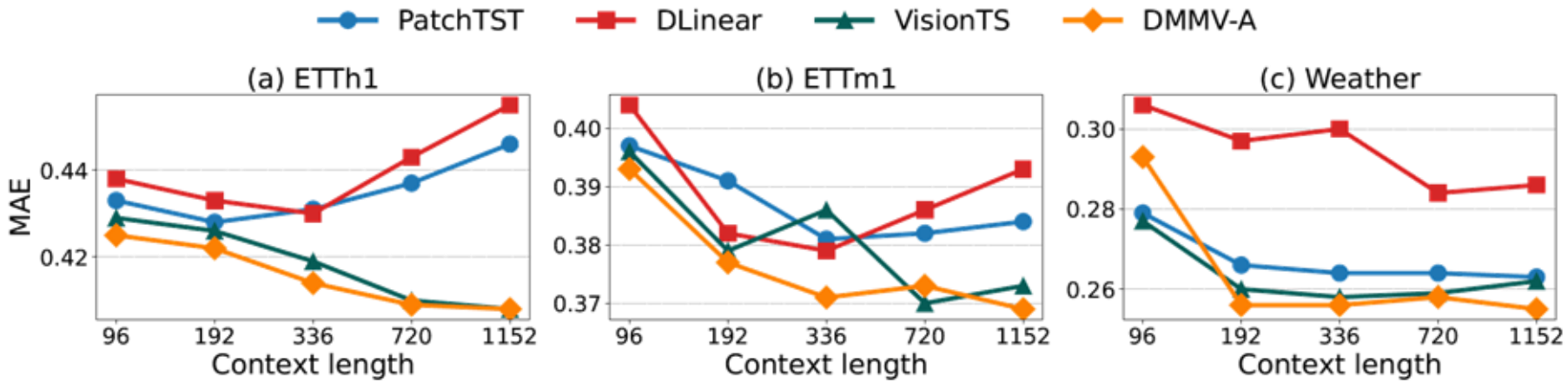
# Integrating MMVs of Time Series

## **DMMV** – Effective Extraction of Periodic Component




**Proposed Method**

## **DMMV** – Impact of Look-Back Window



# Outline of This Section

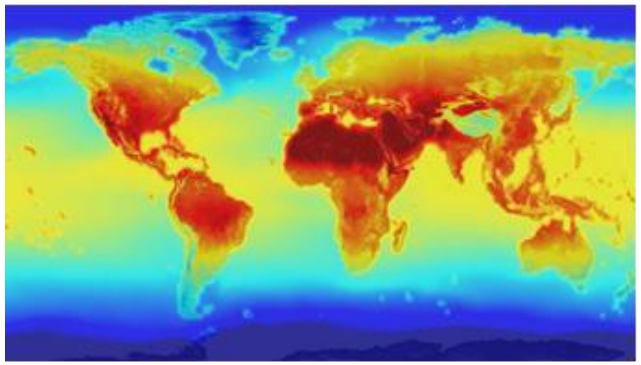
- ☑ **Generating MMVs of time series**
  - Linguistic view and visual view
- ☑ **Cross-modal knowledge transfer via MMVs**
  - Methods using LLMs and LVMs
- ☑ **Integrating MMVs of time series**
  - Combining multiple models or using LMMs

# *Multimodal Learning for Spatio-Temporal Data*

# Table of Contents

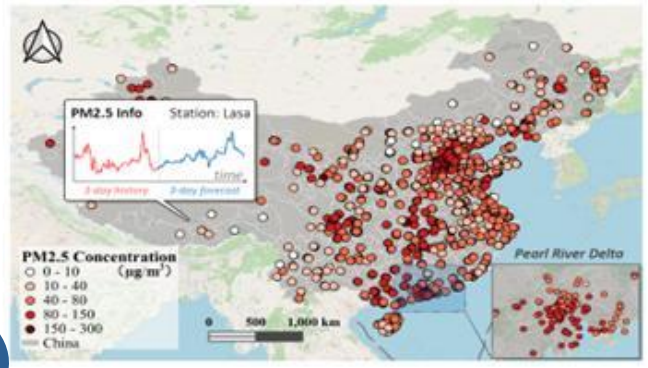
# What is Spatio-Temporal (ST) Data

- Data that integrates **spatial** (location), **temporal** (time), and **event-related** information, capturing how phenomena change across both space and time.



Climate

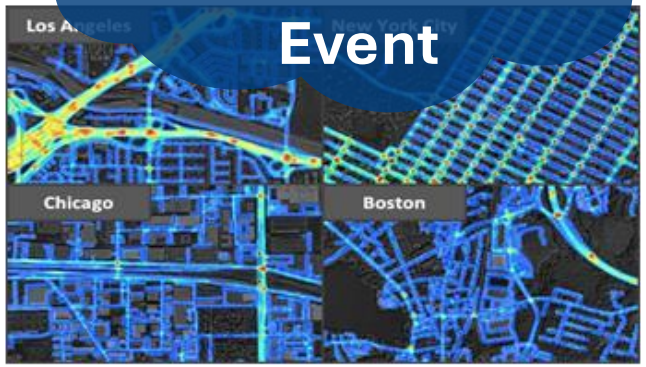
Epidemiology




Environment

**Time, Location, Event**

Social Science

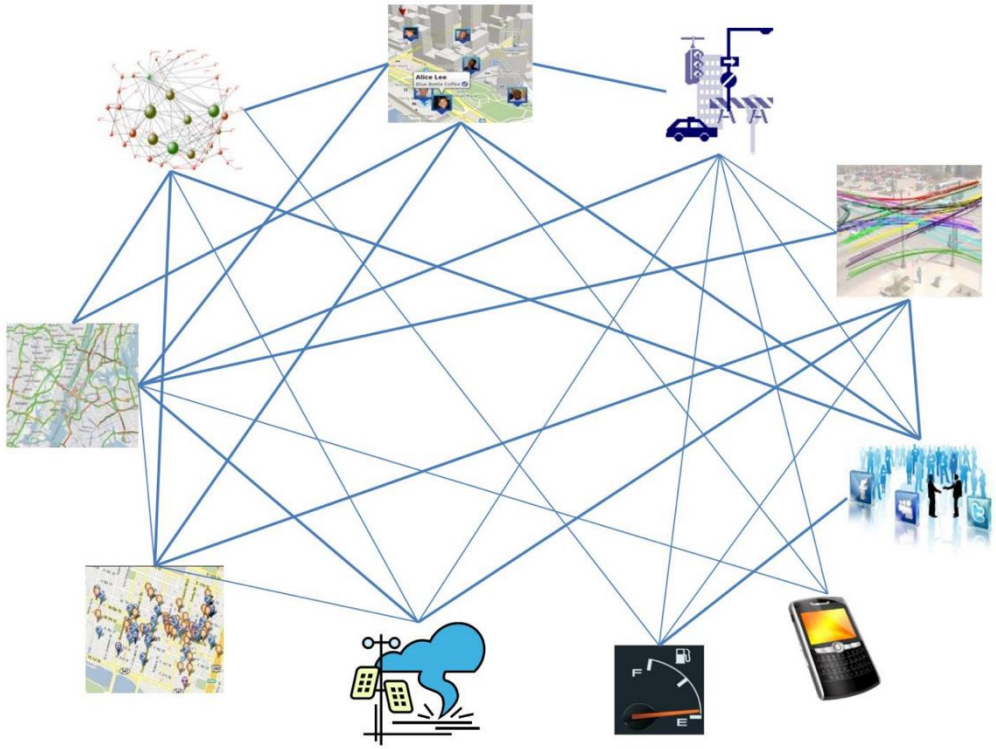


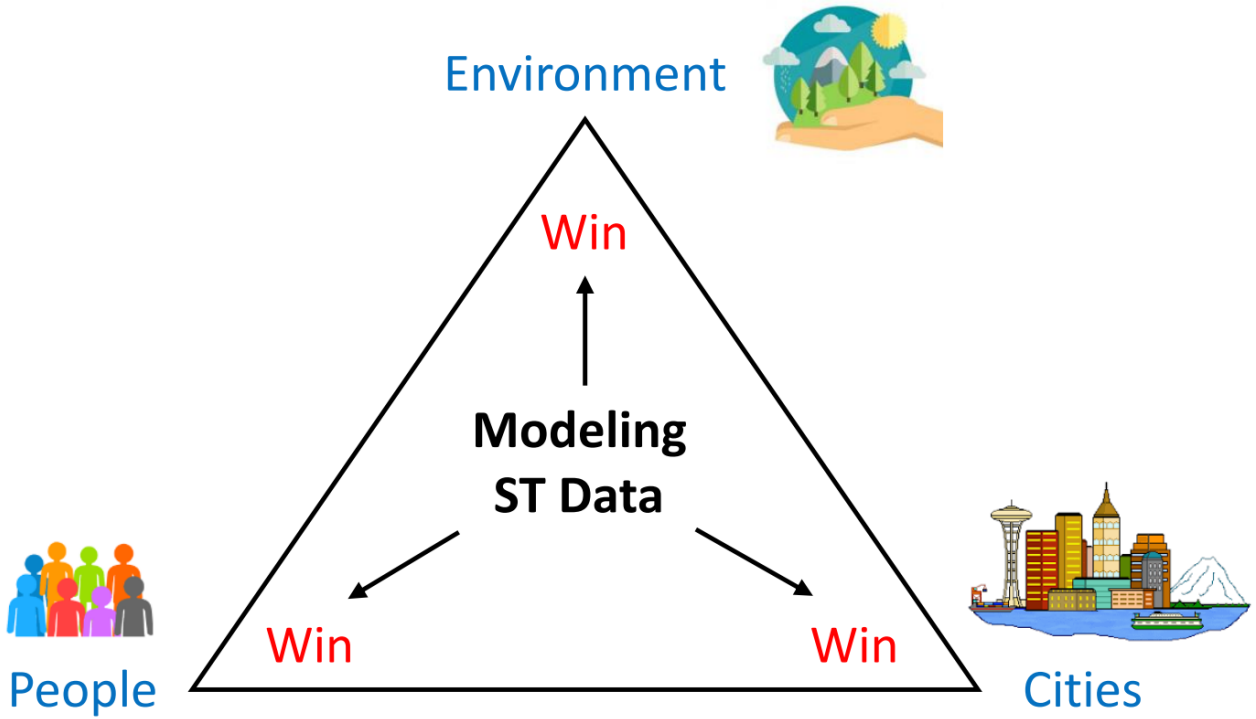

Transportation

Sports Analysis

# What is Spatio-Temporal (ST) Data

- Modeling ST data is the **foundation** of real-world **applications**, creating win-win-win solutions improving the environment, human life quality, and city operation systems.
- *ST data are anywhere, connecting with each other.*

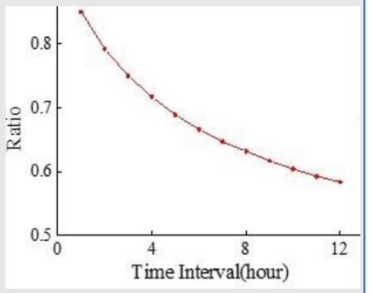




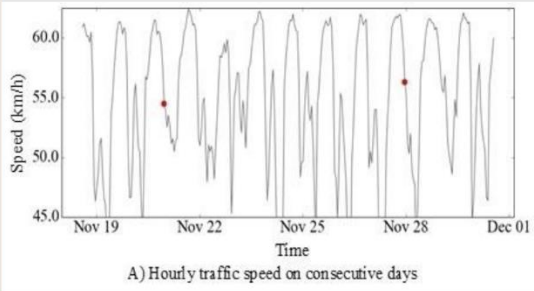
# Why Space is Special?

- Space is not just another feature channel. It carry unique structural priors (geometric laws) more complex than a 1D sequence.

**Closeness**

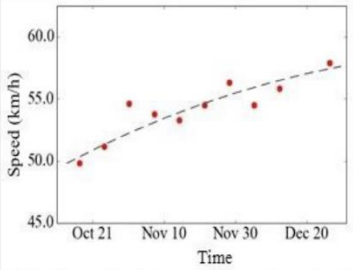



**Periodicity**



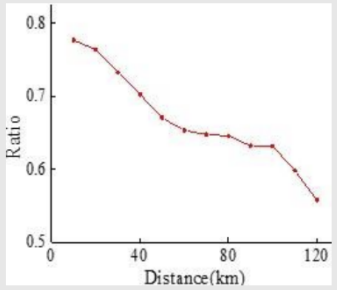

A) Hourly traffic speed on consecutive days

**Trend**



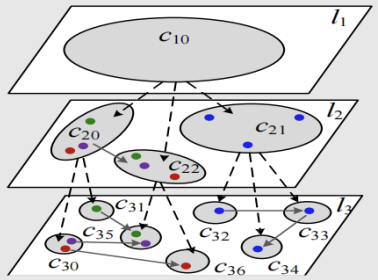

B) Traffic speed at 9-10am on consecutive weekends

**Spatial closeness**




Describing correlations

**Spatial hierarchy**



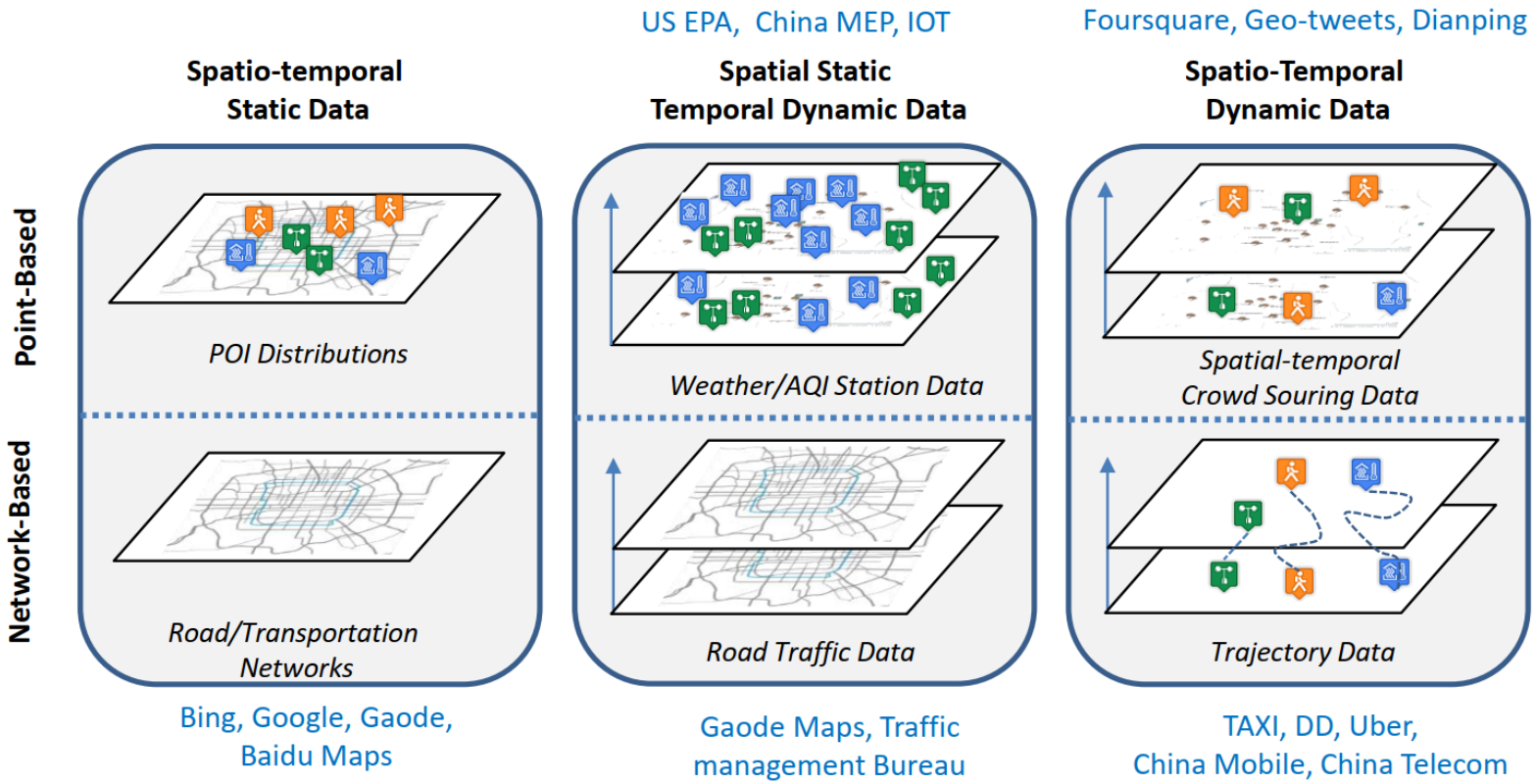

Structural constraints between different spatial granularity

**Spatial distance**


Triangle inequality:
$|d_1 - d_2| \leq d_3 \leq |d_1 + d_2|$

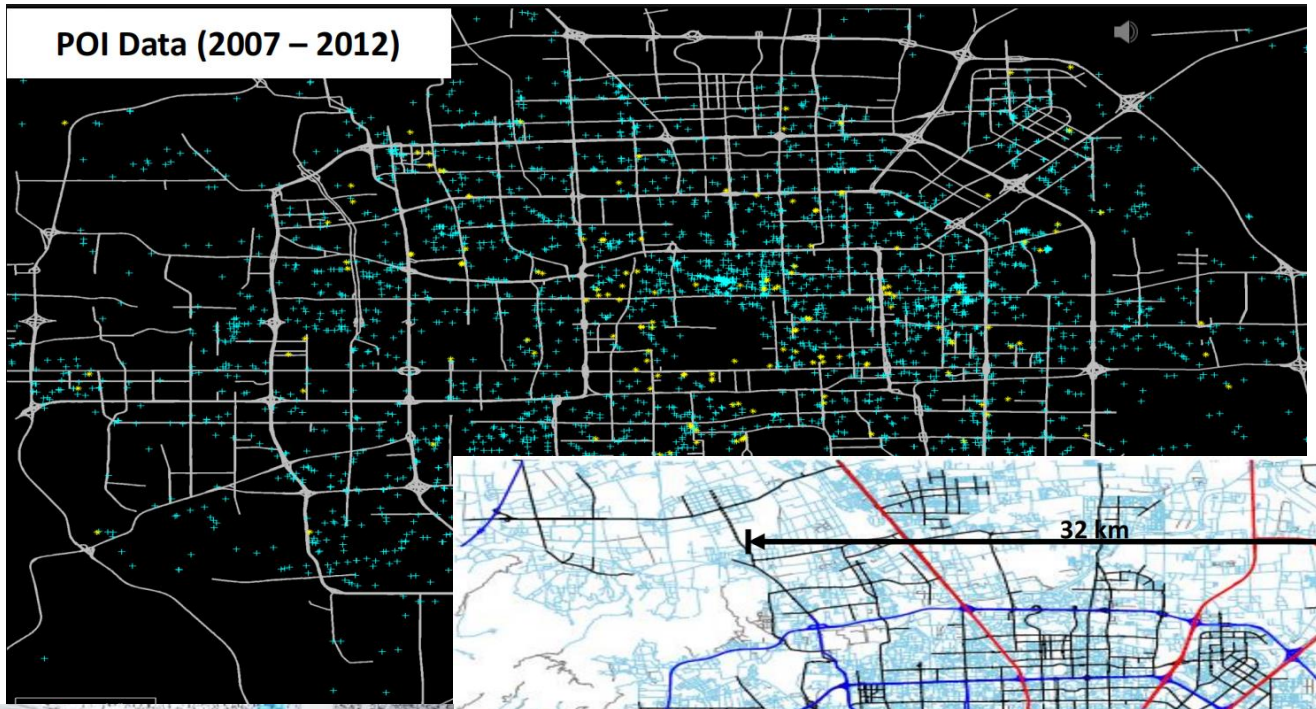
# ST Data – Taxonomy

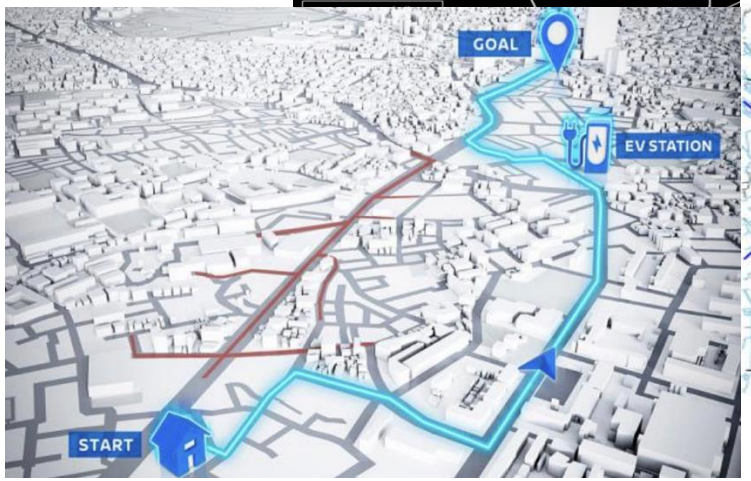
- Spatially and temporally static data
- Spatially static and temporally dynamic data
- Spatially and temporally dynamic data

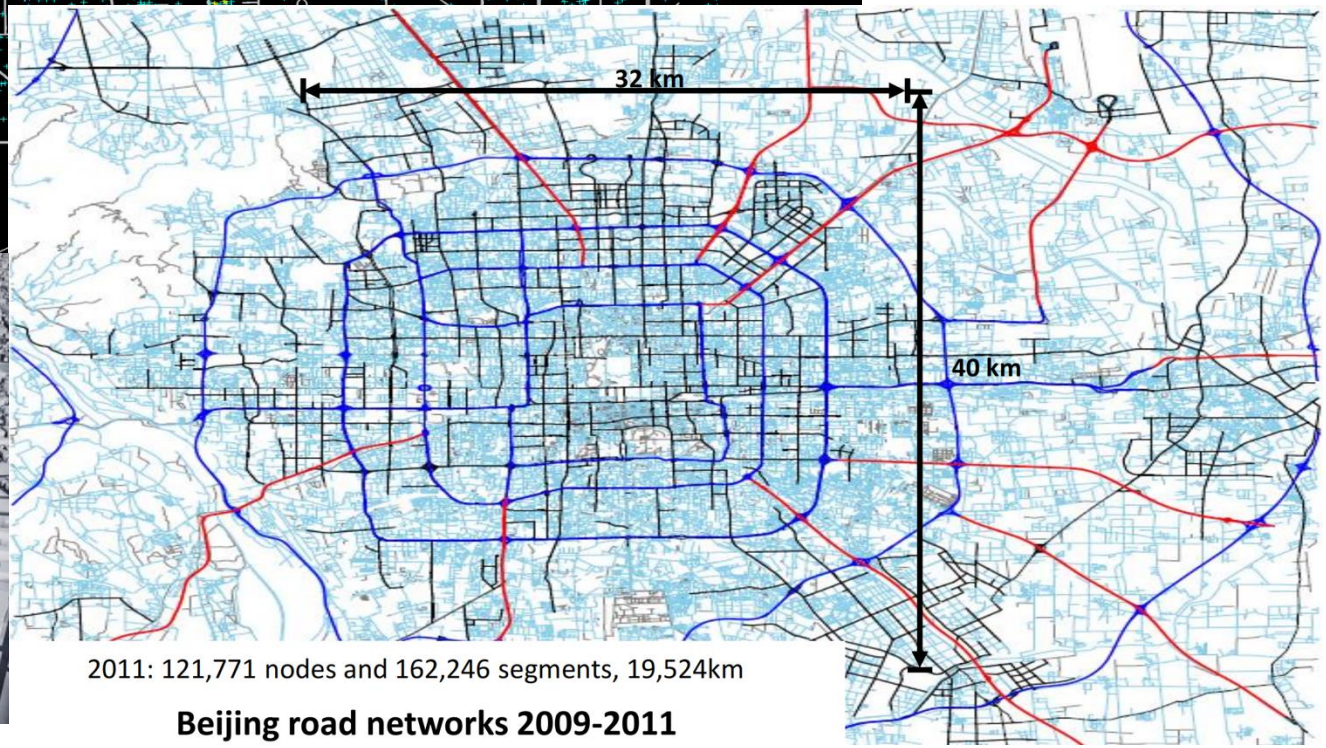


# Spatially and Temporally Static Data

- Points & Locations
- Lines
  - Route, pipeline,
  - Rivers, coast,...
- Graphs
  - Road networks
  - Air lines



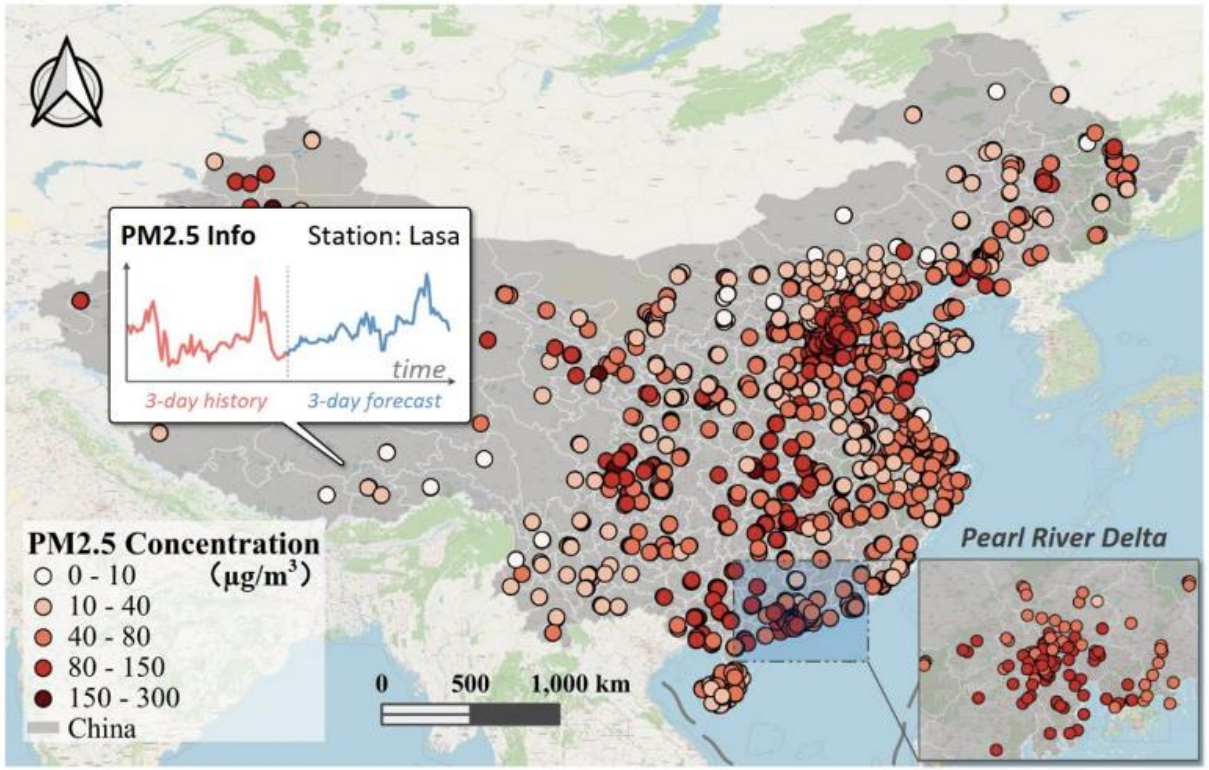




2011: 121,771 nodes and 162,246 segments, 19,524km

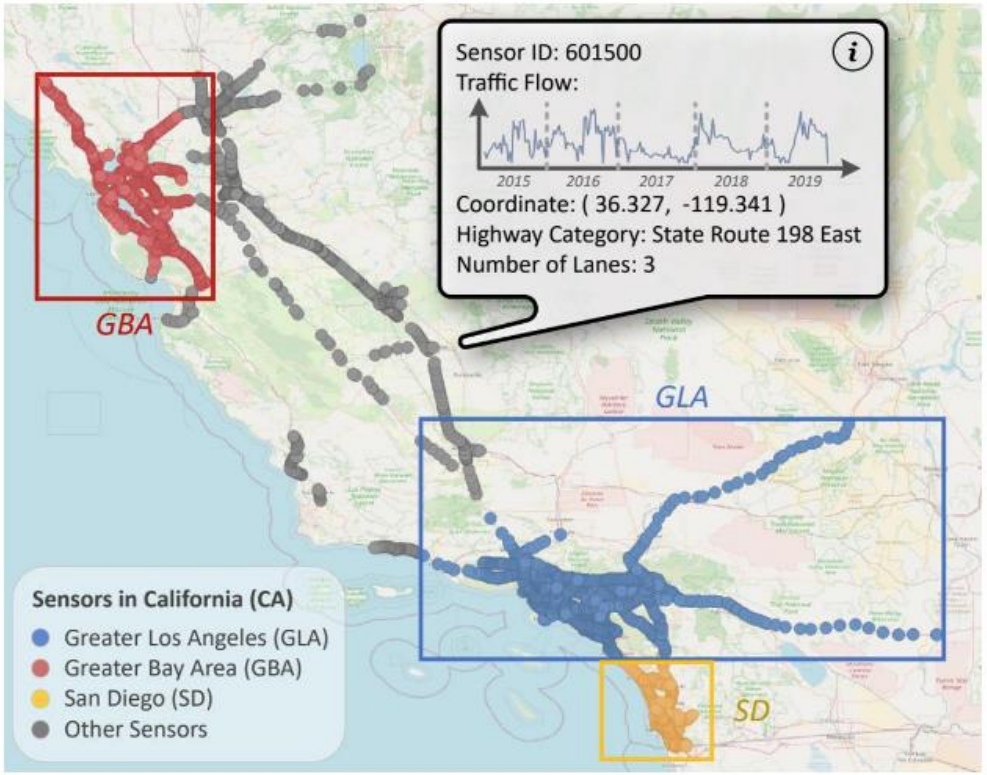
**Beijing road networks 2009-2011**

# Spatially Static and Temporally Dynamic Data

- Usually derived from sensors deployed in different locations.
- Also can be called standard time series and spatial time series.




PM2.5 Data




Traffic Volume Data

# Spatially and Temporally Dynamic Data

- Spatial and temporal values varying in time
  - Moving objects
  - Trajectories

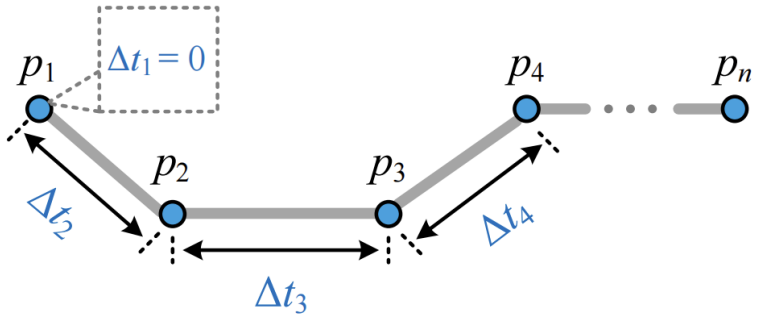
$$T = p_1 \rightarrow p_2 \rightarrow \cdots \rightarrow p_n, \quad p_i = (a_i, b_i, t_i)$$

Timestamp

Location (latitude & longitude)

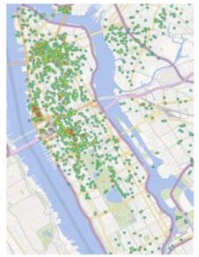
- A spatial trajectory is a sequence derived from a moving object in geographical spaces, formulated by a series of chronologically ordered points

| Latitude, Longitude, Time | | | |
|---|---|---|---|
| $p_1$: | Lat1, | Lng1, | $t_1$ |
| $p_2$: | Lat2, | Lng2, | $t_2$ |
| | …… | …… | …… |
| $p_n$: | Latn, | Lngn, | $t_n$ |



# Spatially and Temporally Dynamic Data

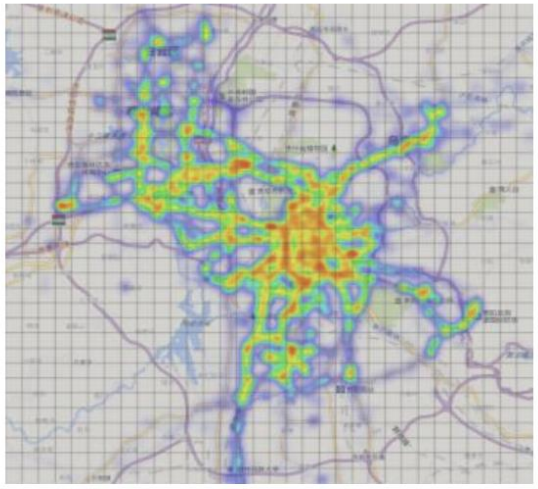
- **Human mobility**
  - Travel logs
  - Check-ins
  - Credit card transactions
  - Phone signal, Wi-Fi...



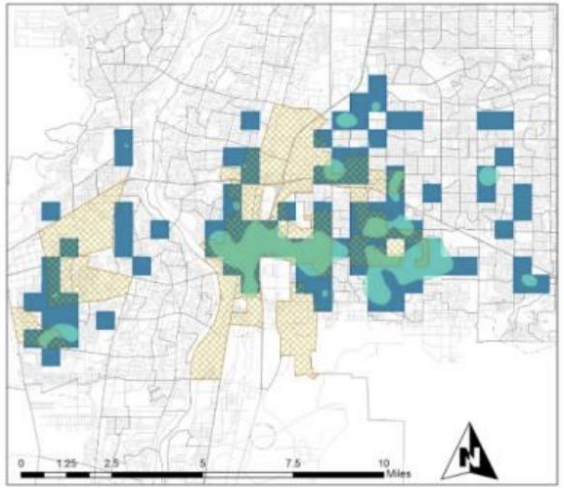

- **Human mobility**
  - Taxis, buses, truck trajectories
  - Airplanes, ferries, cruise, ...
- **Animals migration**
- **Natural phenomena**

# ST Raster Data

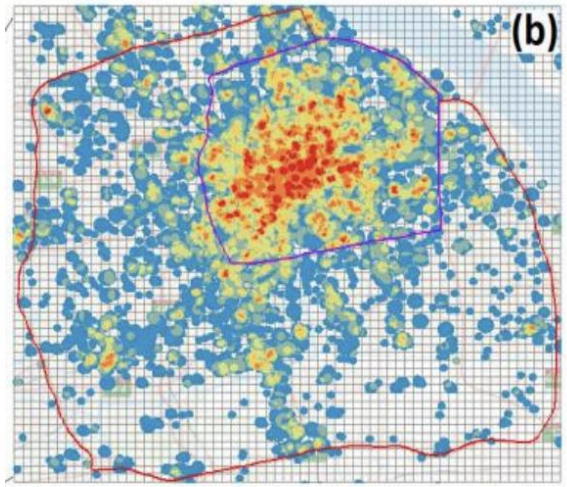
- We partition an area of interest (e.g., a metropolitan) evenly into grid cells, leading to an image-like data format called ST raster data.
  - A pixel → A region
  - RGB → Observations / Attributes



Taxi flows

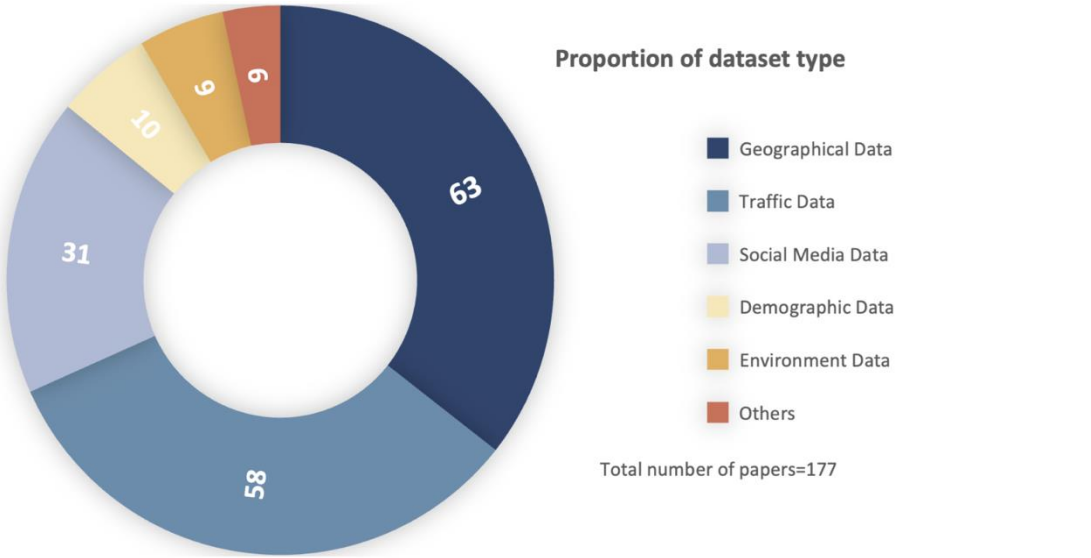


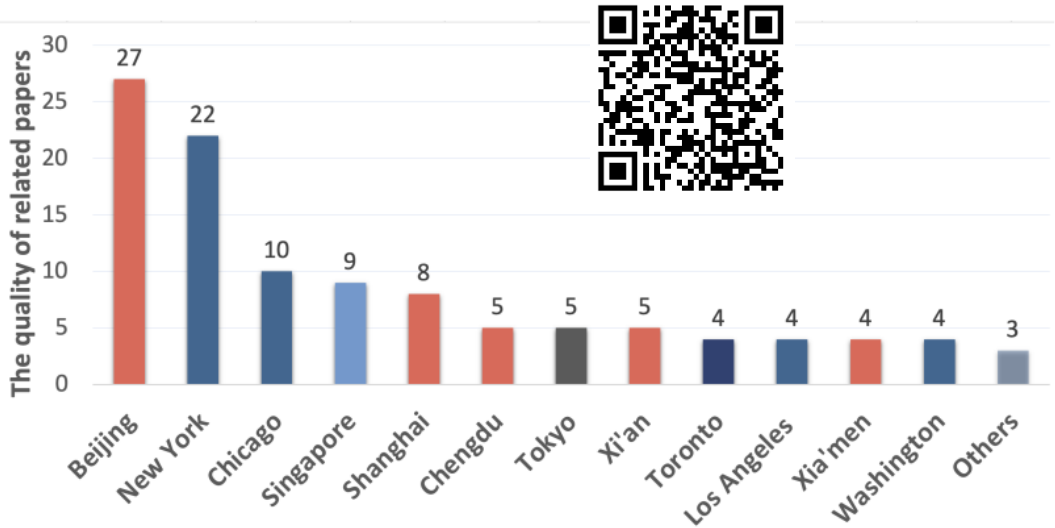
Crime hotspots




Bike-sharing demands

# Data Types and Data Sources



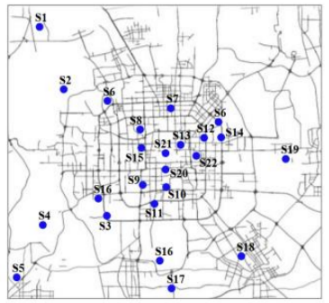

| Category | Content | Format | Dataset | Link | Reference |
|---|---|---|---|---|---|
| Geographical Data | Satellite Image | Image | ArcGIS | https://developers.arcgis.com | [186] |
| | | | PlanetScope | https://developers.planet.com/docs/data/planetscope/ | [154] |
| | | | Google Earth | https://developers.google.com/maps/documentation/ | [116] |
| | | | OpenStreetMap | https://www.openstreetmap.org/ | [337] |
| | | | Baidu Maps | https://lbsyun.baidu.com | [324, 313] |
| | Street-View Image | Image | Baidu Map | https://lbsyun.baidu.com | [186, 124] |
| | | | Google Street | https://developers.google.com/maps/ | [186, 4] |
| | | | Tencent Map | https://lbs.qq.com/tool/streetview/index.html | [112] |
| | POIs | Point Vector | Tencent Map Service | https://lbs.qq.com/getPoint/ | [309, 235] |
| | | | WeChat POIs | https://open.weixin.qq.com | [277] |
| | | | Baidu Map POIs | https://lbsyun.baidu.com | [154, 172, 175, 110, 313] |
| | | | NYC Open POIs | https://opendata.cityofnewyork.us/ | [170, 272, 20, 366, 288] |
| | | | Foursquare | https://developer.foursquare.com/docs/checkins/checkins | [20, 381, 13, 42, 107, 116] |
| | | | Wikipedia POIs | https://www.wikipedia.org | [386] |
| | | | AMap Service | https://lbs.amap.com | [10] |
| | | | Yelp POIs | https://www.yelp.com/developers | [13, 380, 383] |
| | | | Dianping POIs | https://api.dianping.com/ | [33, 63] |
| | | | Weibo POIs | https://open.weibo.com/wiki/API | [33, 134, 77] |
| | | | Flickr POIs | https://www.flickr.com/services/developer/api/ | [99] |
| | | | Bing Map POIs | https://www.bingmapsportal.com | [37] |
| Traffic Data | Traffic Trajectory | Spatio-temporal Trajectory | Shenzhou UCar | https://bit.ly/2MG47xz | [93] |
| | | | Chicago Transportation | https://data.cityofchicago.org/ | [272, 288, 116] |
| | | | VED | https://github.com/gsoh/VED | [209, 372] |
| | | | Taxi Shenzhen | https://github.com/cbdog94/STL | [113, 302] |
| | | | NYC Open Taxi Data | https://opendata.cityofnewyork.us/how-to/ | [368, 366] |
| | | | GeoLife | http://urban-computing.com/index-893.htm | [96, 398, 400, 394, 347] |
| | | | T-Drive Taxi | http://urban-computing.com/index-58.htm | [350, 351, 217, 191] |
| | | | DiDi Traffic | https://outreach.didichuxing.com/research/opendata/ | [349, 188, 228, 328, 261] |
| | | | Xiamen Taxi | https://data.mendeley.com/datasets/6xg39x9vgd/1 | [342, 40, 124, 39] |
| | | | Grab-Posisi | https://goo.su/W3yD5m | [337, 339] |
| | Taffic Flow | Spatio-temporal Graph | California-PEMS | http://pems.dot.ca.gov | [9, 254] |
| | | | METR-LA | https://www.metro.net | [143, 171] |
| | | | Large-ST | https://github.com/liuxu77/LargeST | [182] |
| | | | MobileBJ | https://github.com/FIBLAB/DeepSTN/issues/4 | [170, 134, 33] |
| | | | TaxiBJ | https://goo.su/aQyjTAz | [164, 11, 226, 120, 368, 74] |
| | | | BikeNYC | https://citibikenyc.com/ | [170, 11, 226, 120] |
| | Road Network | Spatial Graph | OpenStreetMap | https://www.openstreetmap.org | [339, 13, 188, 349, 84] |
| | | | US Census Bureau | https://www.census.gov/data.html | [366] |
| | Logistics | Spatio-temporal Trajectory | LaDe | https://cainiaotechai.github.io/LaDe-website/ | [305] |
| | | | JD Logistics | https://corporate.jd.com/ourBusiness#jdLogistics | [235] |
| Social Media Data | Text | Text | Twitter | https://developer.twitter.com/en/docs | [20, 381, 383, 352, 270, 301, 240] [289, 283, 285, 284, 200, 184] |
| | | | Common Crawl | https://registry.opendata.aws/commoncrawl/ | [380] |
| | | | Yelp Reviews | https://www.yelp.com/dataset | [380, 383] |
| | | | Weibo Traffic Police | http://open.weibo.com/developers/ | [342] |
| | Geo-tagged Image & Video | Image&Video | YFCC100M | https://goo.su/jzaDU | [386, 340, 99] |
| | | | NUS-WIDE | https://goo.su/dWPQZcD | [340, 338] |
| | | | GeoUGV | https://qualinet.github.io/databases/video/ | [187] |
| | Users' Info | Time Series | Jiepang User Check-in | https://jiepang.app/ | [74] |
| | | | Gowalla User Location | http://konect.cc/networks/loc-gowalla_edges/ | [42, 352] |
| | | | WeChat Mobility | https://open.weixin.qq.com/ | [277] |
| Demographic Data | Crime | Time Series | NYC Crime | https://opendata.cityofnewyork.us/ | [368] |
| | Land Use | Time Series | Land Use SG | https://www.ura.gov.sg/Corporate/Planning/Master-Plan | [156] |
| | | | Land Use NYC | https://goo.su/puTuG | [156] |
| | Population | Time Series | WorldPop | https://www.worldpop.org/ | [309, 154, 10] |
| Environment Data | Meteorology | Time Series | TipDM China Weather | https://www.tipdm.org/ | [178] |
| | | | DarkSky Weather | https://support.apple.com/en-us/102594 | [349] |
| | | | WeatherNY | https://opendata.cityofnewyork.us/ | [272] |
| | | | WeatherChicago | https://data.cityofchicago.org/ | [272] |
| | | | Weather Underground | https://www.wunderground.com/ | [342] |
| | | | DidiSY | https://www.didiglobal.com/ | [12] |
| | | | WD_BJ weather | https://goo.su/DmHFHd | [192] |
| | | | WD_USA weather | https://goo.su/RVhBA | [192] |
| | Greenery | Time Series | Google Earth | https://earth.google.com/ | [342] |
| | Air Quality | Time Series | UrbanAir | https://goo.su/hfzNB53 | [399, 396, 392] |
| | | | KnowAir | https://github.com/shuowang-ai/PM2.5-GNN | [286, 346, 370, 318] |

Deep Learning for Cross-Domain Data Fusion in Urban Computina:Taxonomy, Advances, and Outlook, Information Fusion, 2024.

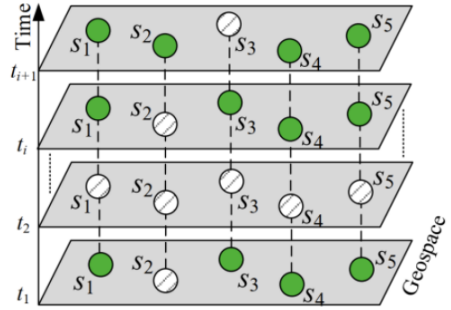
# Why Multimodal ST Data Mining

- Real-world ST data is often "broken" or "incomplete."
- Unlock the power from multiple (sparse) data across different domains

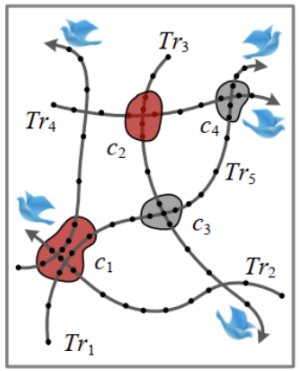

**Biased distribution**

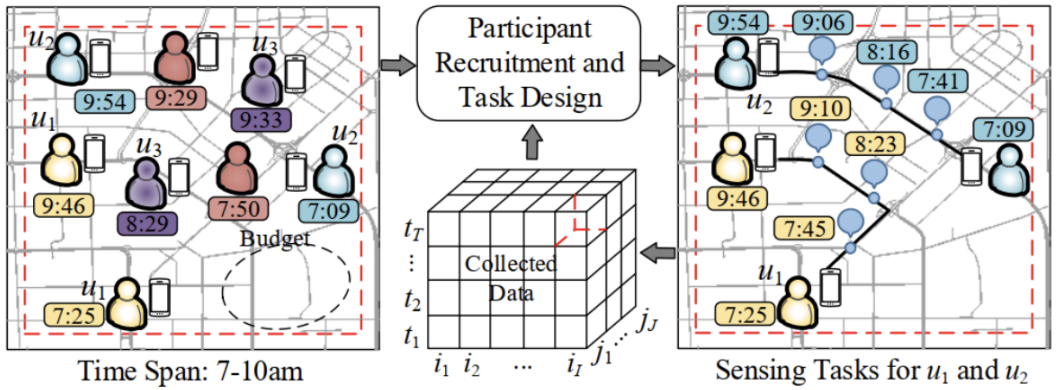


**Data sparsity**




**Data missing**






**Resource deployment**

# Example 1 – Air Quality Inference

- **Challenge**: AQI stations are expensive and sparse => massive spatial data gaps.
- **Solution**: Fusing sparse meteorology with dense, heterogeneous data (Traffic, POIs, Mobility) to infer fine-grained air quality everywhere.

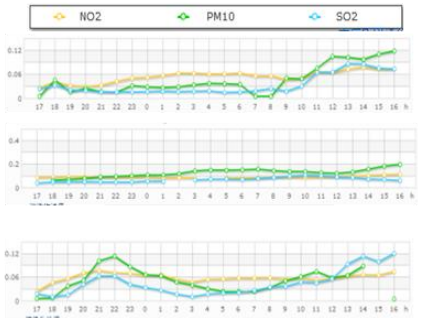
Meteorology

Traffic

Human Mobility

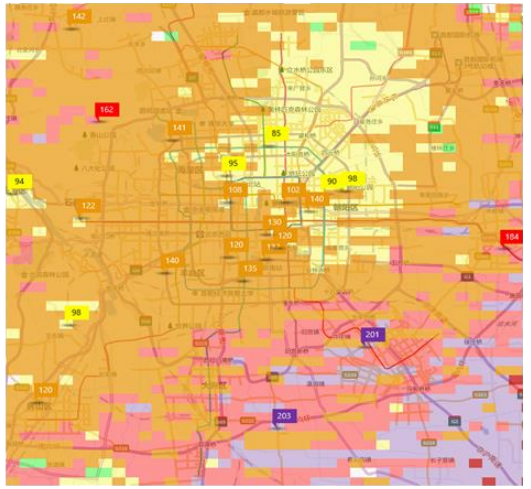
POIs

Road networks




Historical air quality data

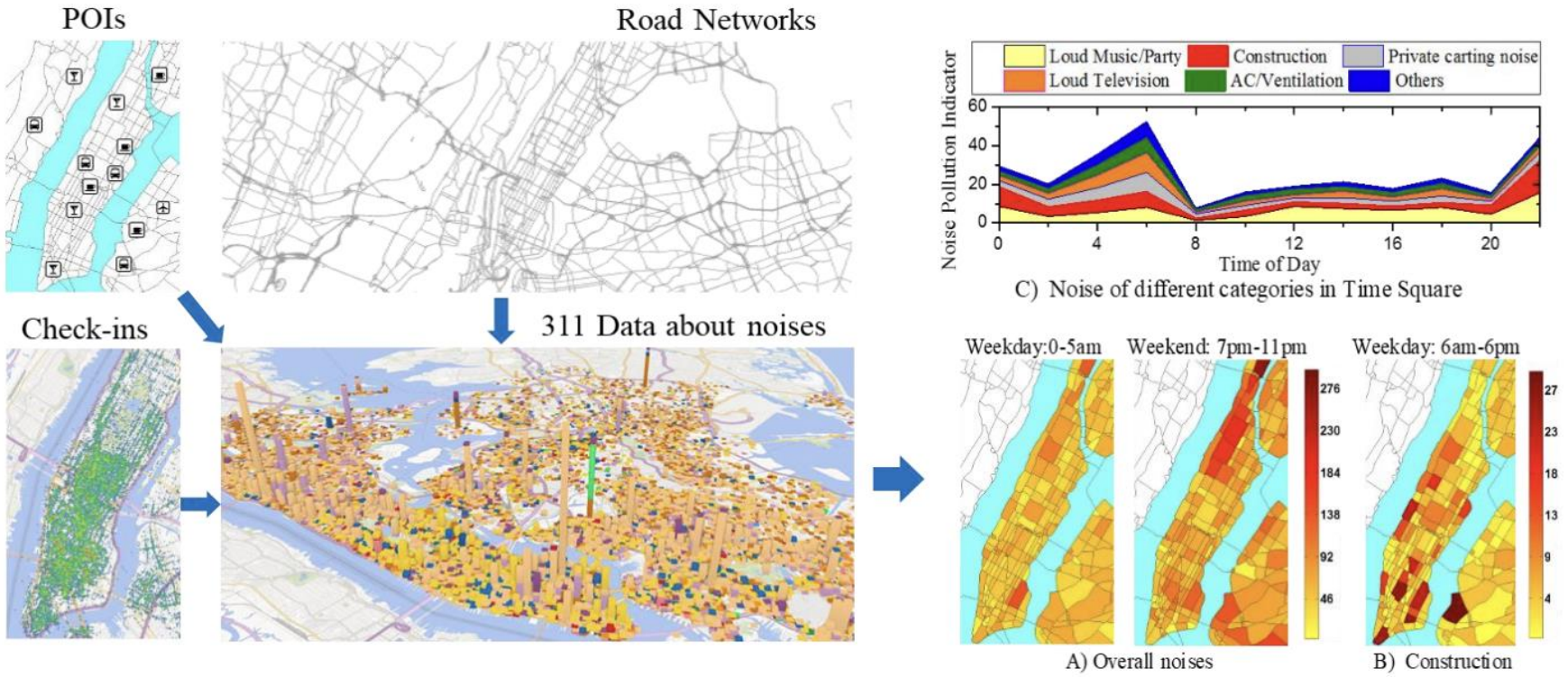
Real-time air quality reports



U-Air: When Urban Air Quality Inference Meets Big Data, KDD 2013

# Example 2 – NYC Noise Inference

- **Challenge**: Noise complaints are passive and lagged, no noise data for manage.
- **Solution**: Aligning social media (Check-ins) and POIs with road structures to categorize and predict noise sources (e.g., Construction vs. Nightlife).

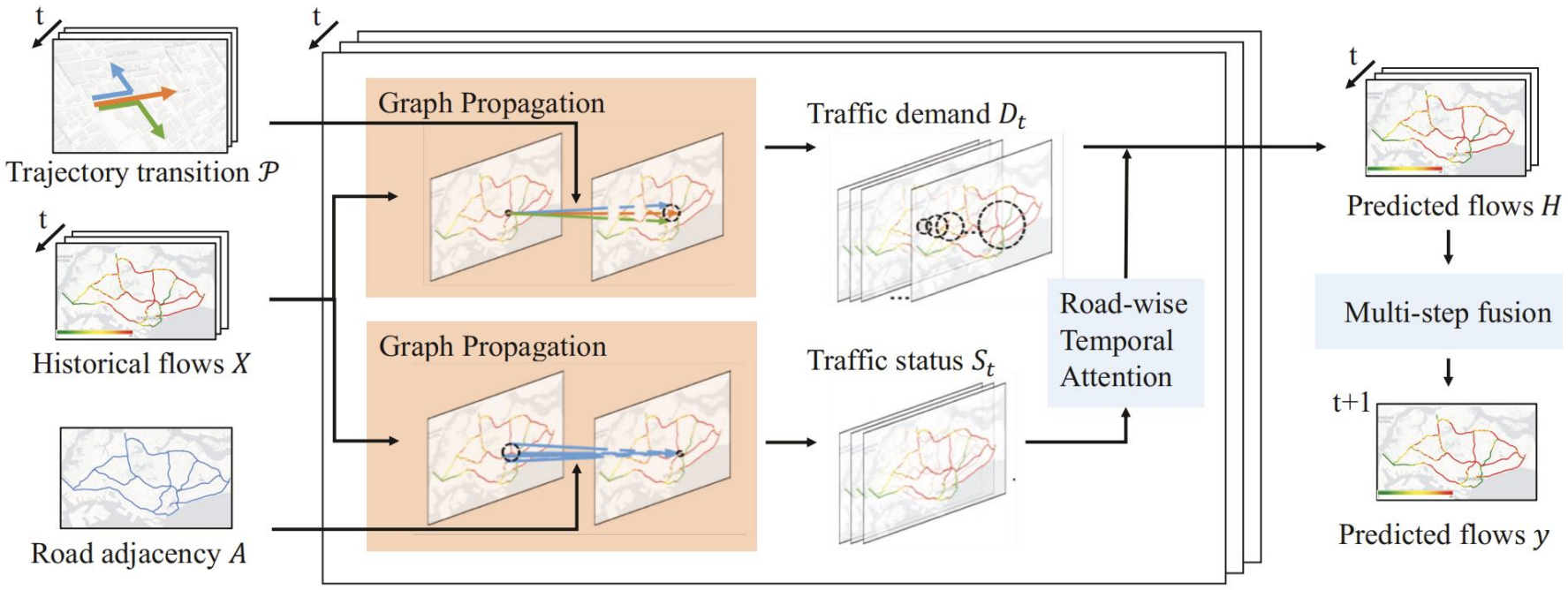



Diagnosing New York City's Noises with Ubiquitous Data, Ubicomp 2014

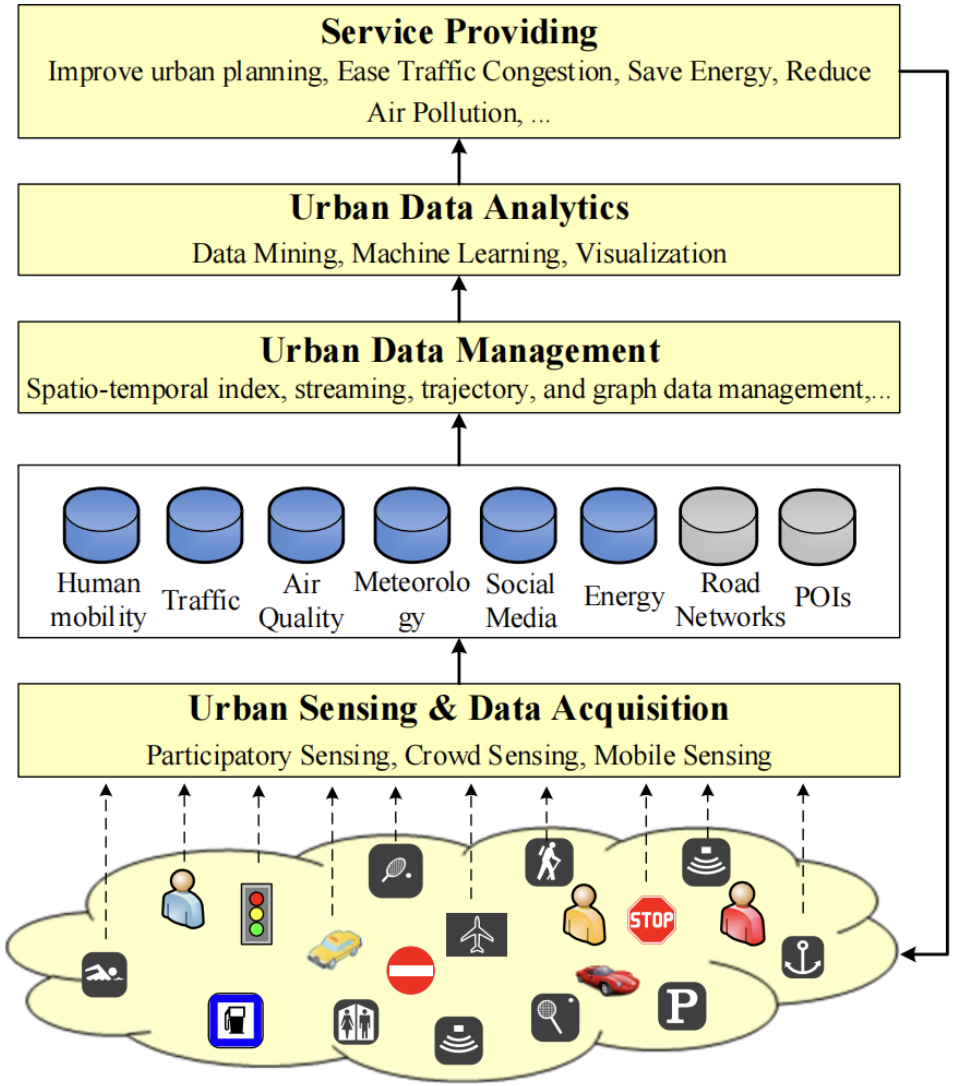
# Example 3 — Traffic Flow Prediction

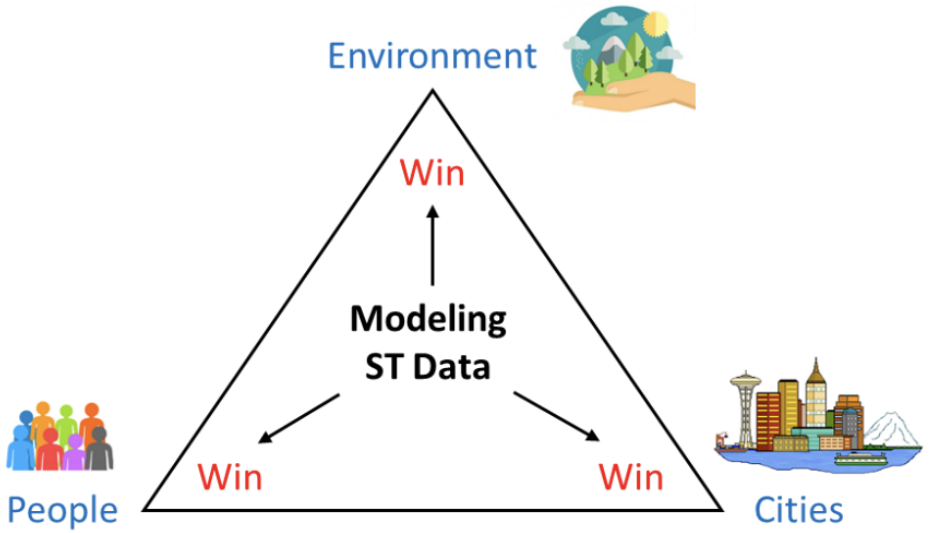
- **Challenge**: Traditional models treat traffic as numbers, ignoring road network constraints and vehicle transition logic.

- **Solution**: Integrating road graphs (Static) with vehicle trajectories (Dynamic) to model how traffic physically flows across the urban network.



# ST Data Intelligence Framework




*Tackle the Big challenges*

*in Big cities*

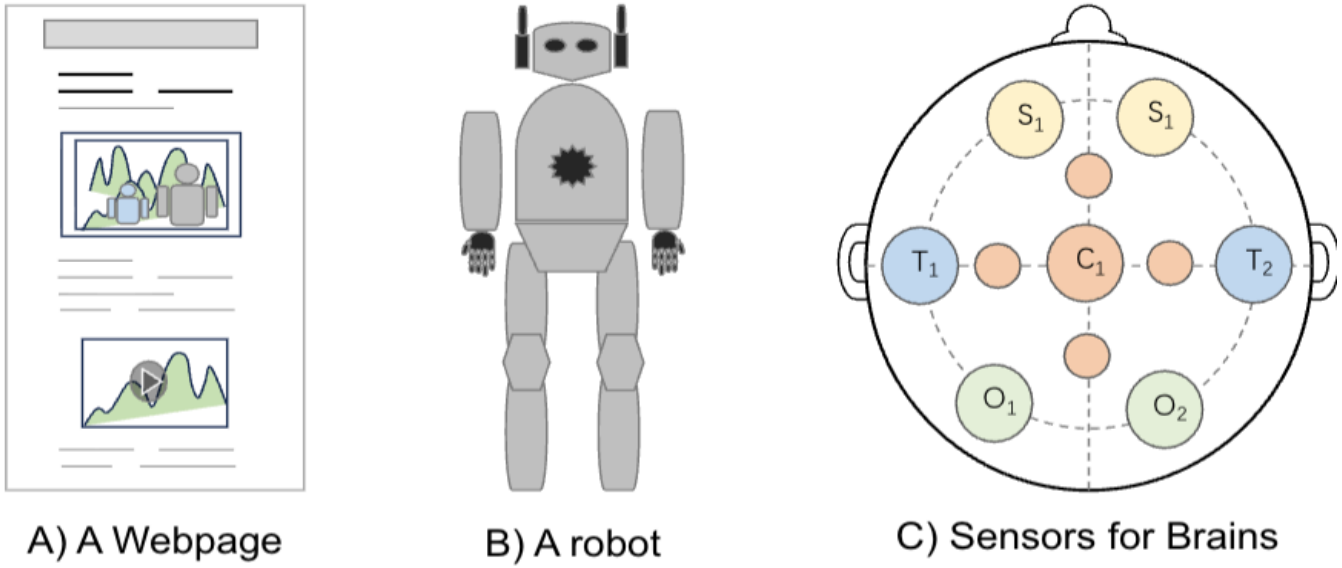
*using Big data!*

**Urban Computing: concepts, methodologies, and applications**.
Zheng, Y., et al. *ACM transactions on Intelligent Systems and Technology*.

# Multimodal STDM vs Traditional MM ?

- Research Gap: **In Domain** vs **Cross Domain** Knowledge Fusion

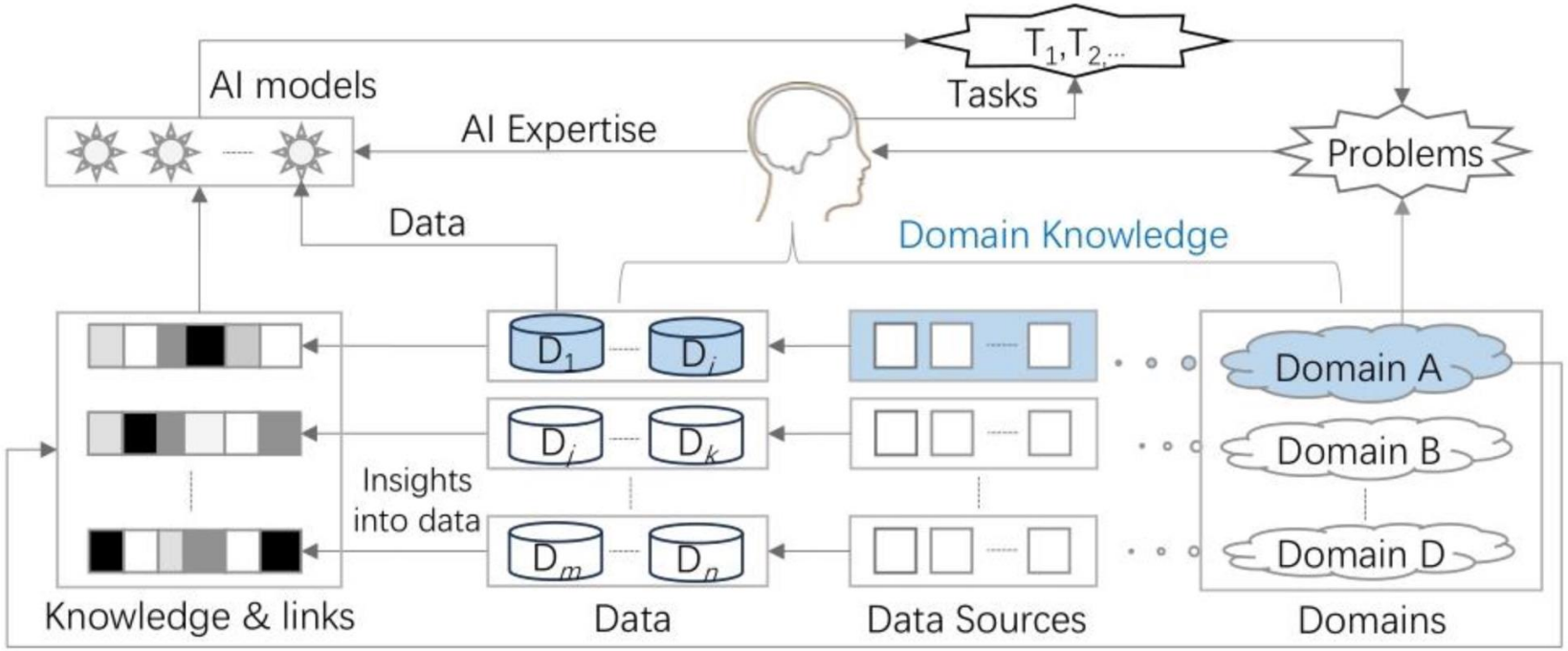
- Existing research focuses on **single-domain** multimodal fusion, data are originally aligned (collected for same problem), which fails in **cross-domain** ST scenarios.




A) A Webpage B) A robot C) Sensors for Brains

# What is Cross-domain Data Fusion

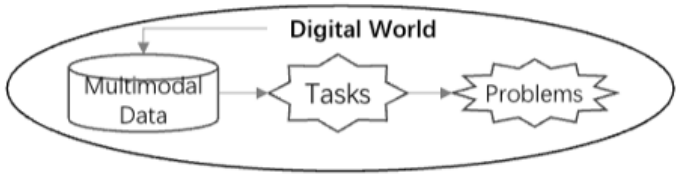
- Data from different domains, collected for different problems, originally not aligned.
- E.g. Air Quality Inference (history AQI, traffic, land uses, meteorology data)



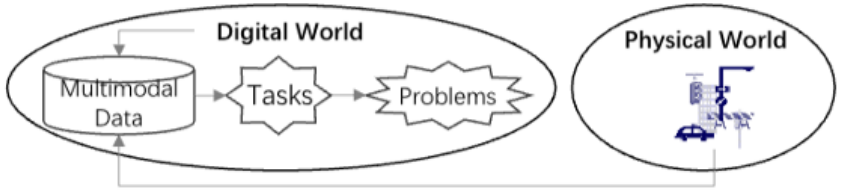

Fusing Cross-Domain Knowledge from Multimodal Data to Solve Problems in the Physical World. In ACM Transactions on Intelligent Systems and Technology, 2025.

# ST Multimodal Learning is Future

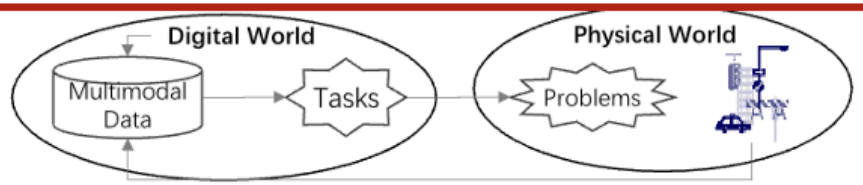
- Current research on multimodal learning is mainly focus on solving problems in digital world (stage a & b), rarely stepping into the physical world (stage c).




A) Solving digital problems using data in the digital world




B) Solving problems in digital world using data from both worlds




C) Solving problems in the physical world using data from both worlds

1) Daily Multimodal Apps, Image/Video Generation

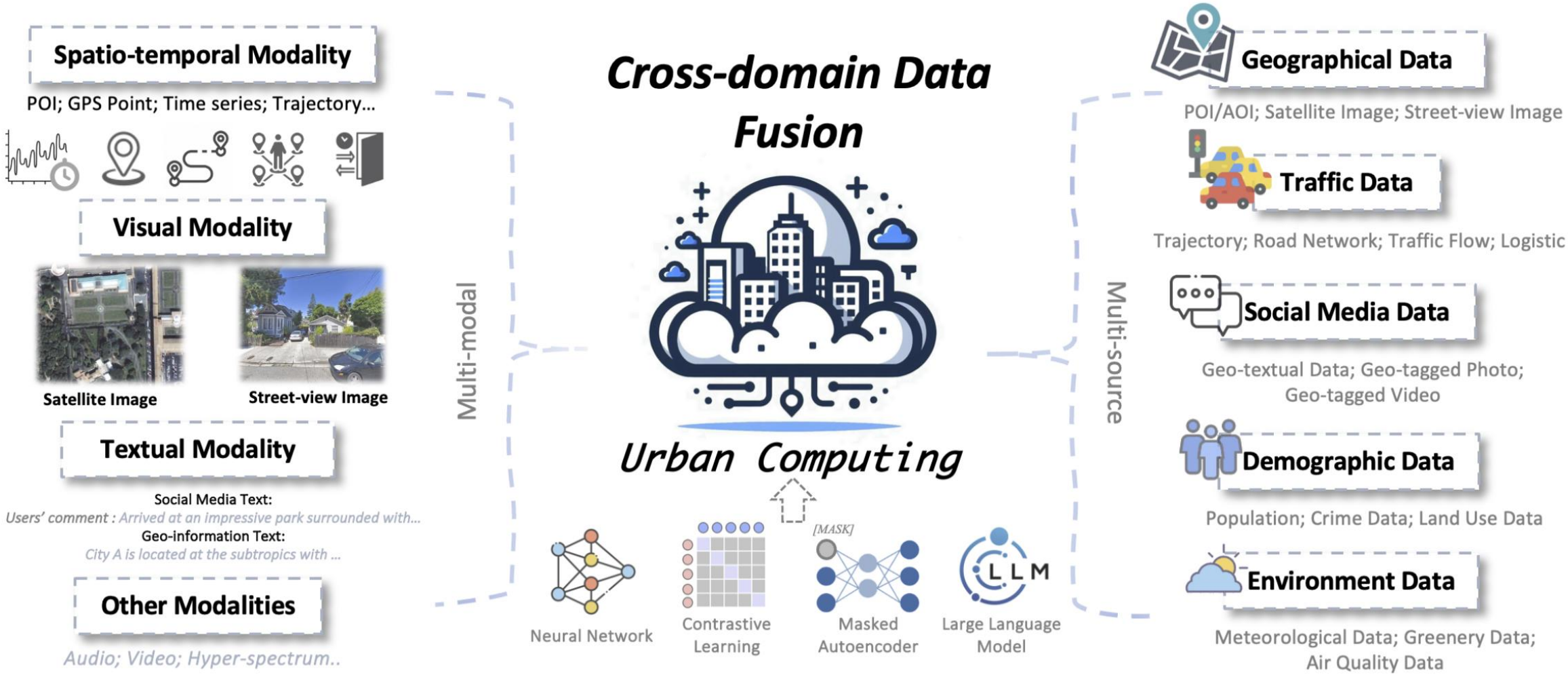
2) Motion-sensing Game, e.g. Switch

3) Real World Problems, e.g. AQI

Essential difference between multimodal ML in ST compared to the common MM.

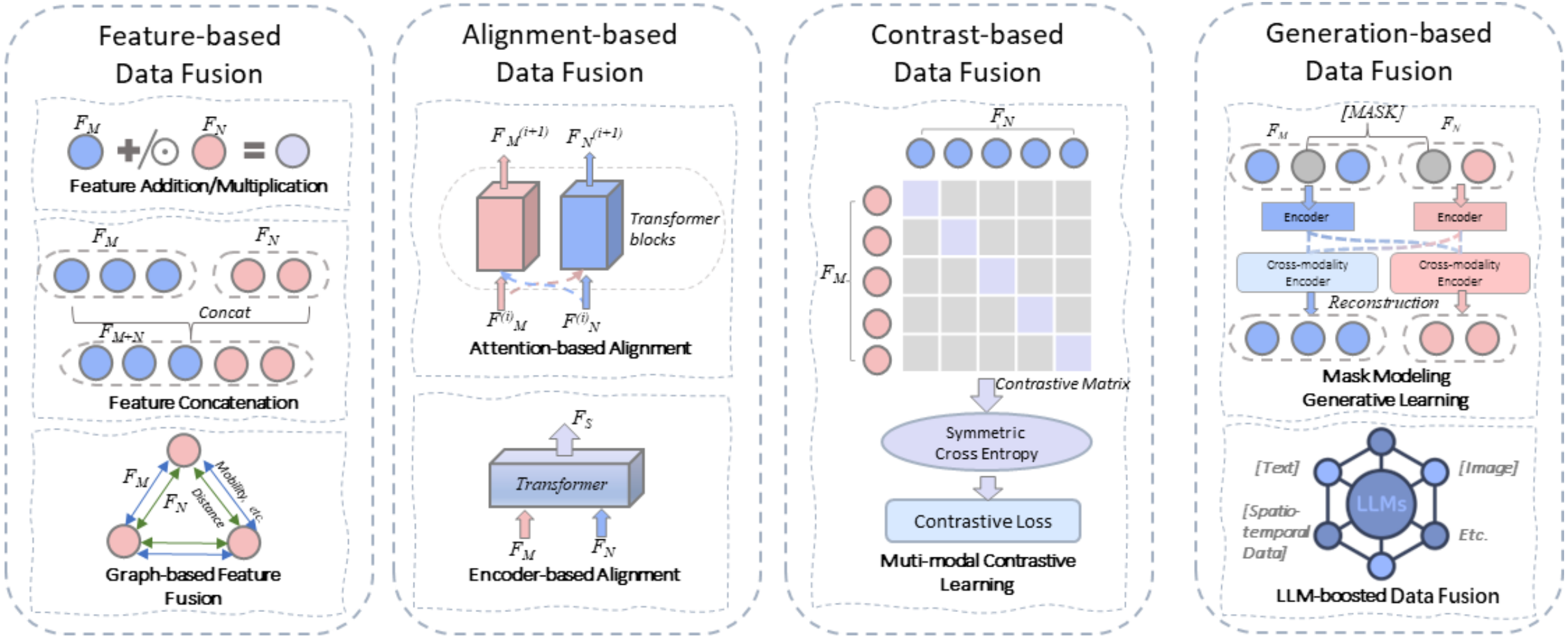
Fusing Cross-Domain Knowledge from Multimodal Data to Solve Problems in the Physical World. In ACM Transactions on Intelligent Systems and Technology, 2025.

# Principle of ST Multimodal Fusion



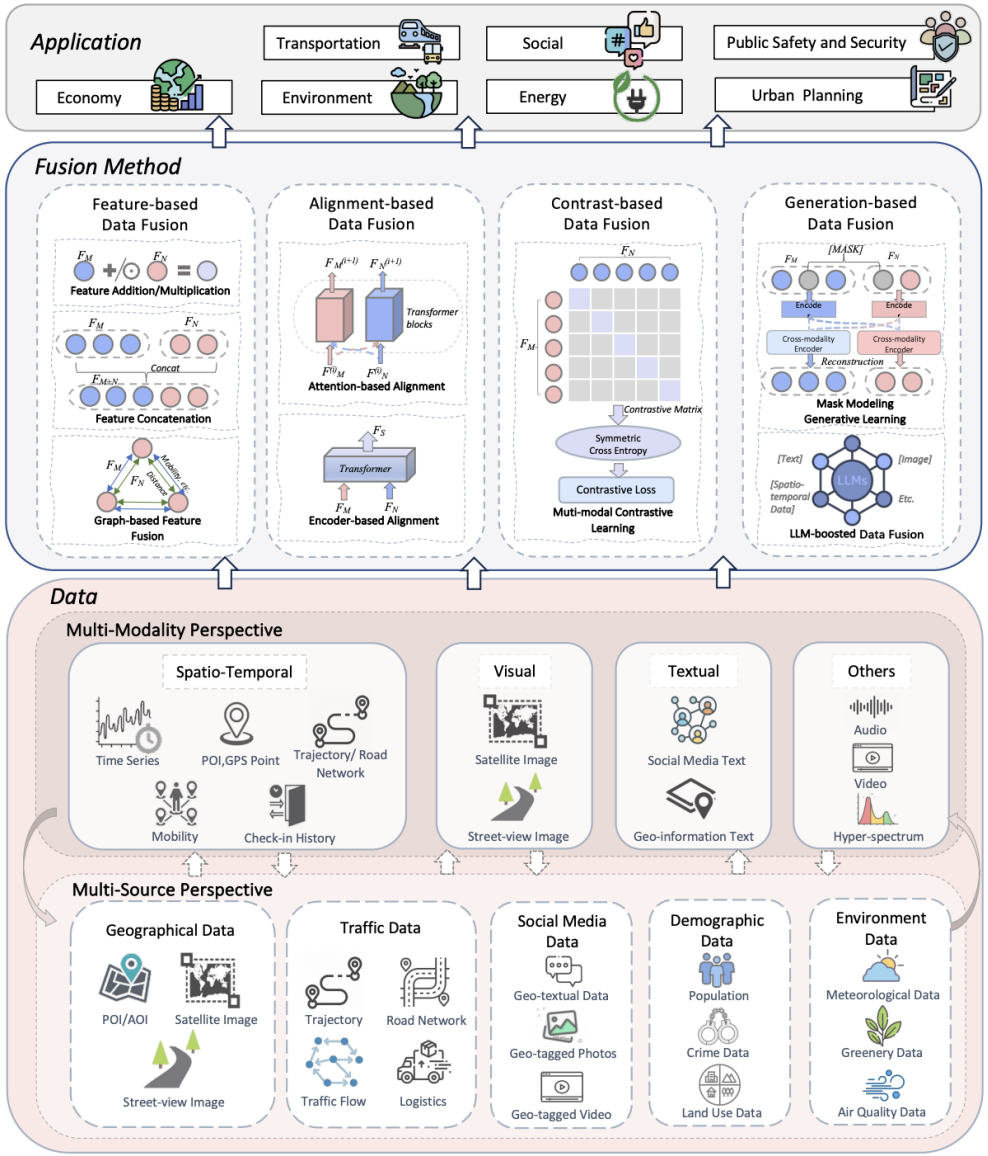

Deep Learning for Cross-Domain Data Fusion in Urban Computing: Taxonomy, Advances, and Outlook, Information Fusion, 2024.

# Deep Learning-based Fusion Methods



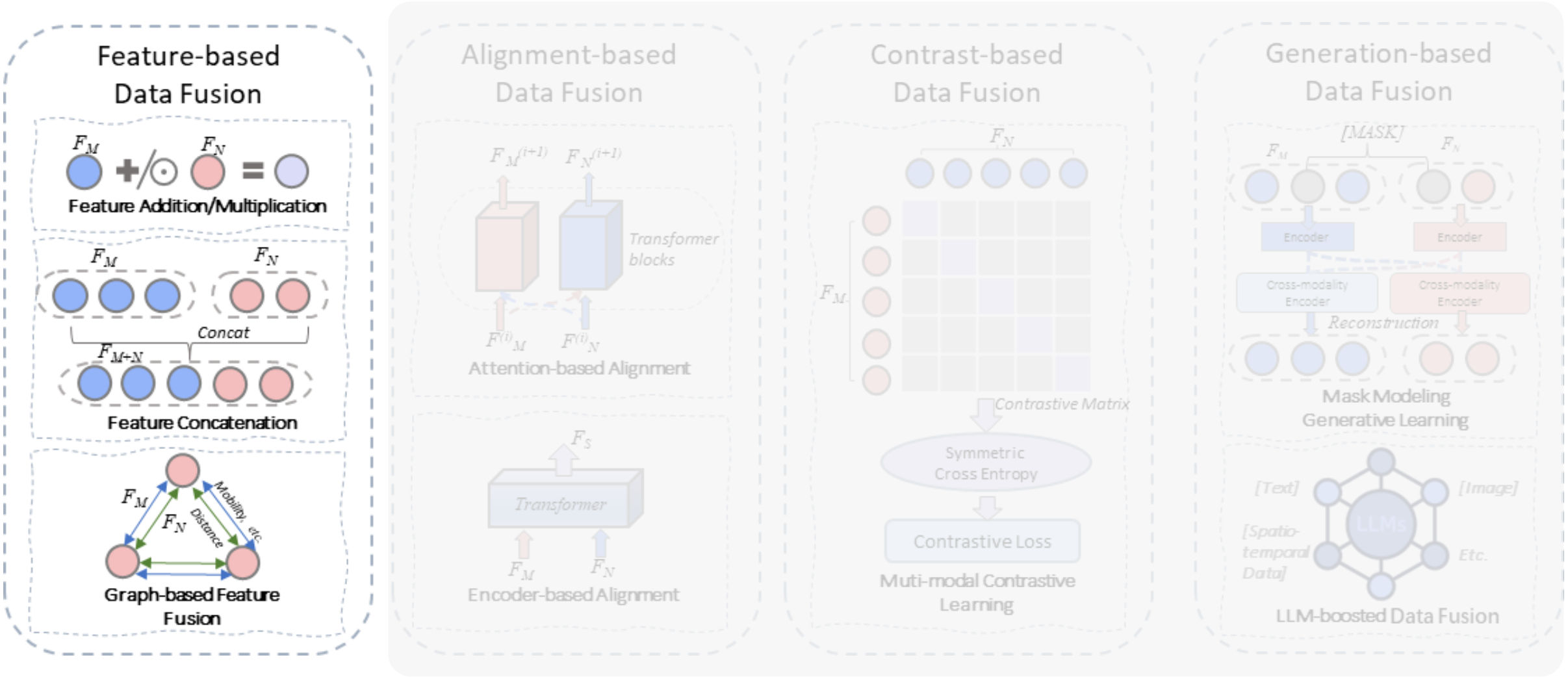

Deep Learning for Cross-Domain Data Fusion in Urban Computing: Taxonomy, Advances, and Outlook, Information Fusion, 2024.

# Deep Learning-based Fusion Methods




Deep Learning for Cross-Domain Data Fusion in Urban Computing: Taxonomy, Advances, and Outlook, Information Fusion, 2024.

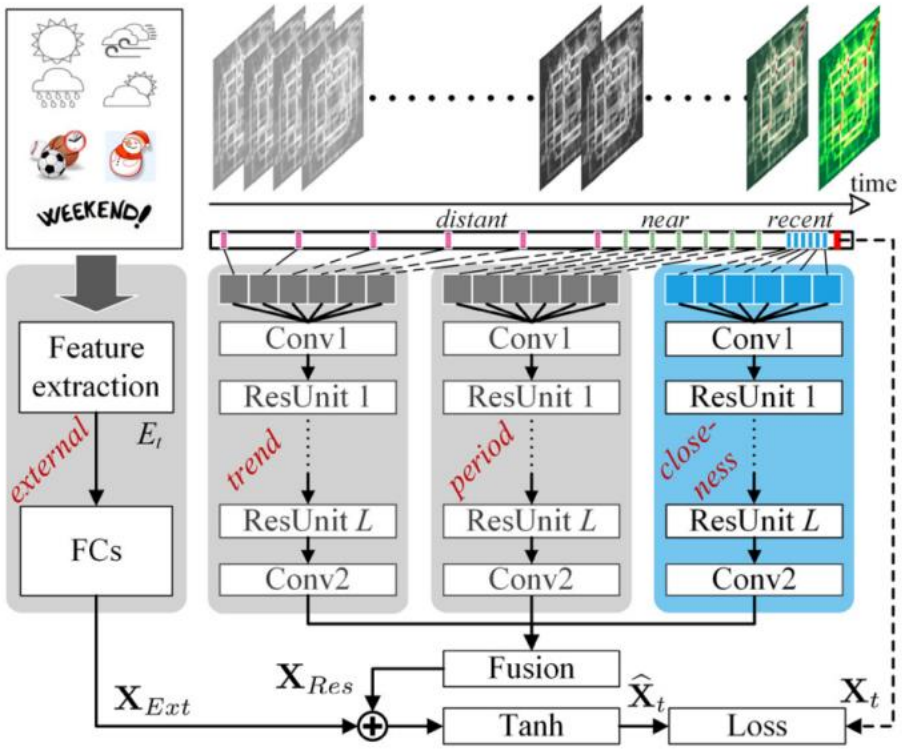
# Deep Learning-based Fusion Methods




Deep Learning for Cross-Domain Data Fusion in Urban Computing: Taxonomy, Advances, and Outlook, Information Fusion, 2024.

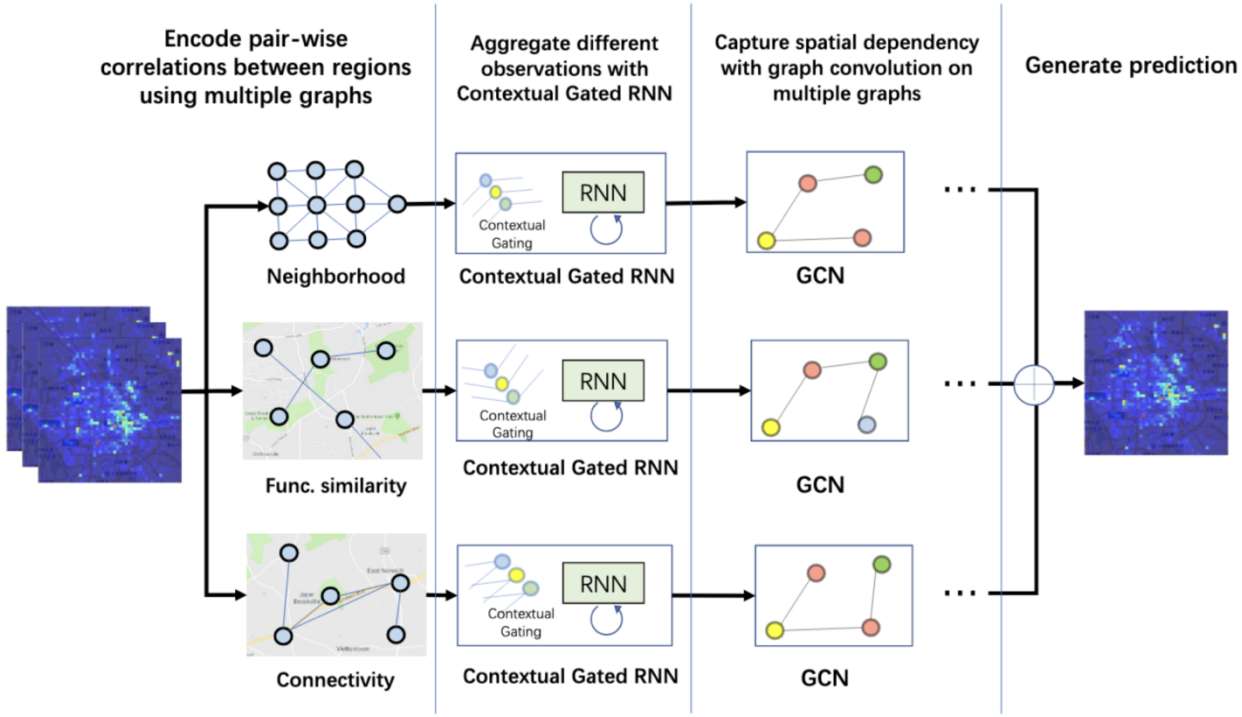
# Feature-based Fusion (Simplest!)

- Feature Addition/Multiplication
- Feature Concatenation



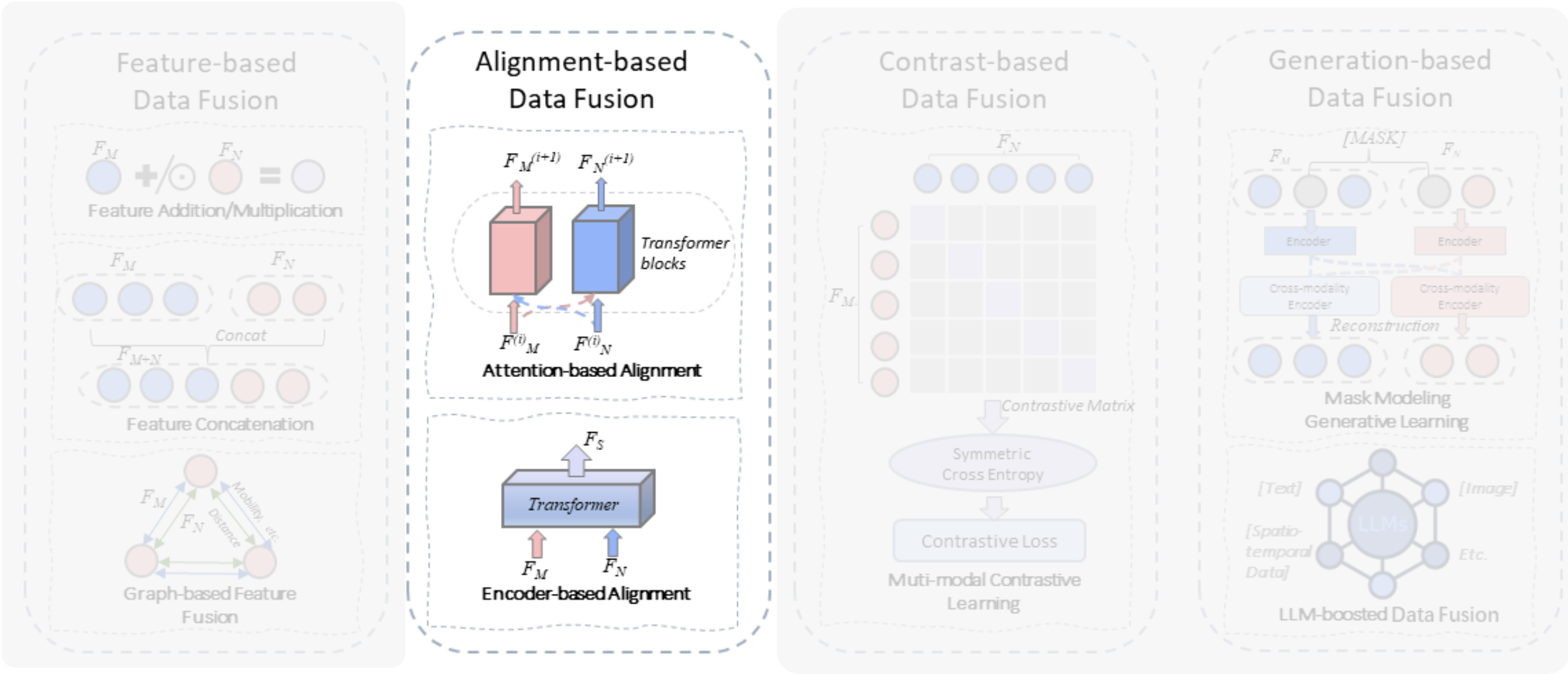

Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction, AAAI 2017

- Graph-based Data Fusion




Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting, AAAI 2019

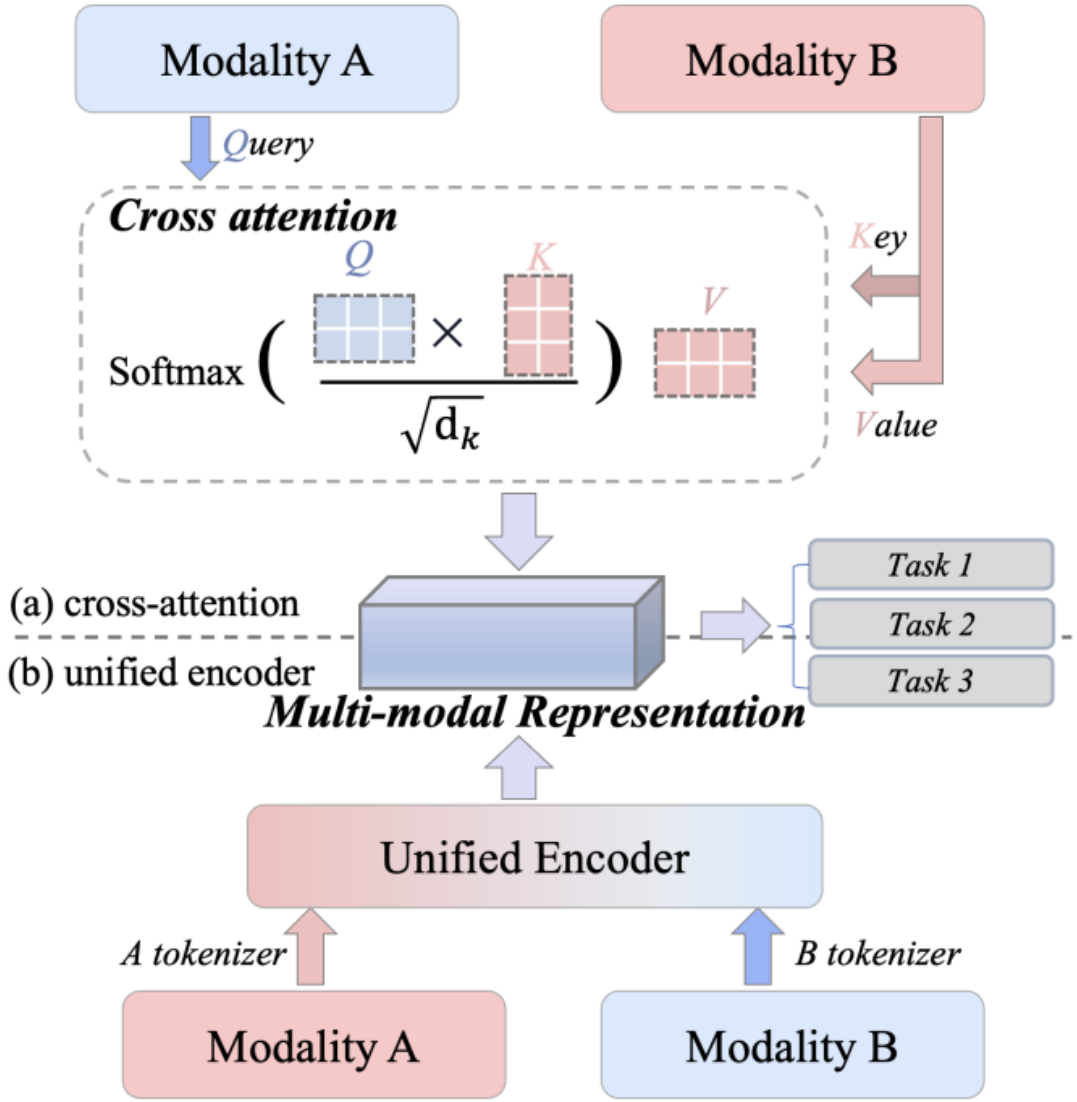
# Deep Learning-based Fusion Methods




Deep Learning for Cross-Domain Data Fusion in Urban Computing: Taxonomy, Advances, and Outlook, Information Fusion, 2024.

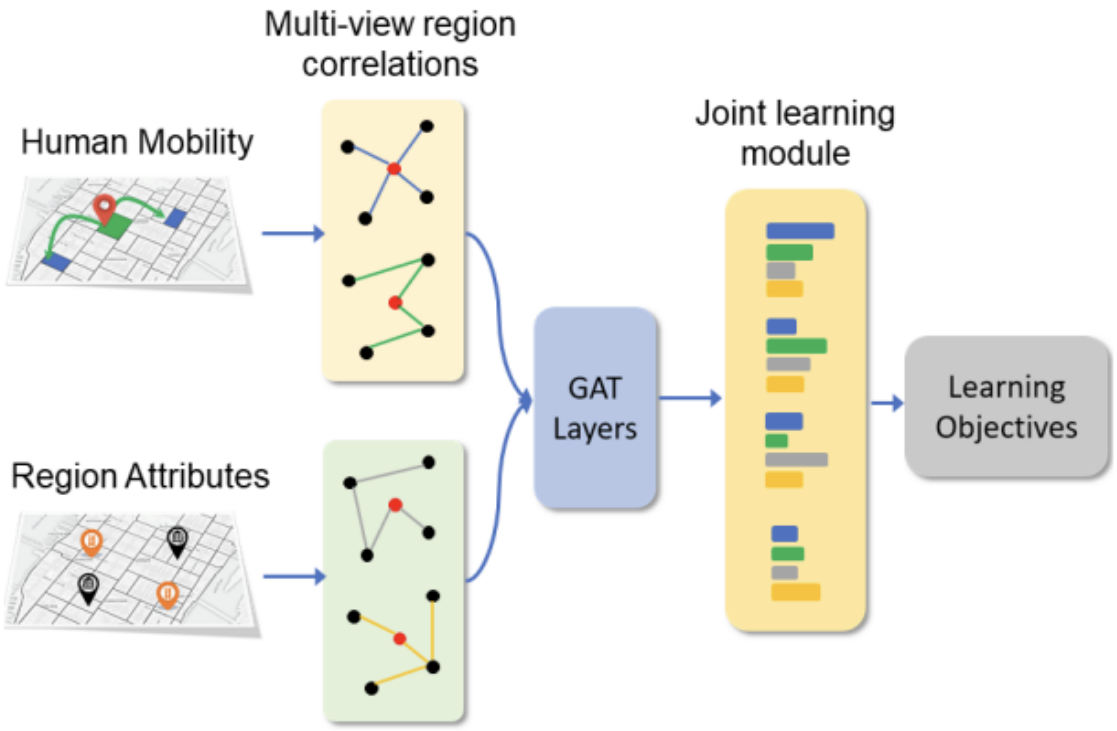
# Alignment-based Fusion

- Attention-based
- Encoder-based



# Alignment-based Fusion

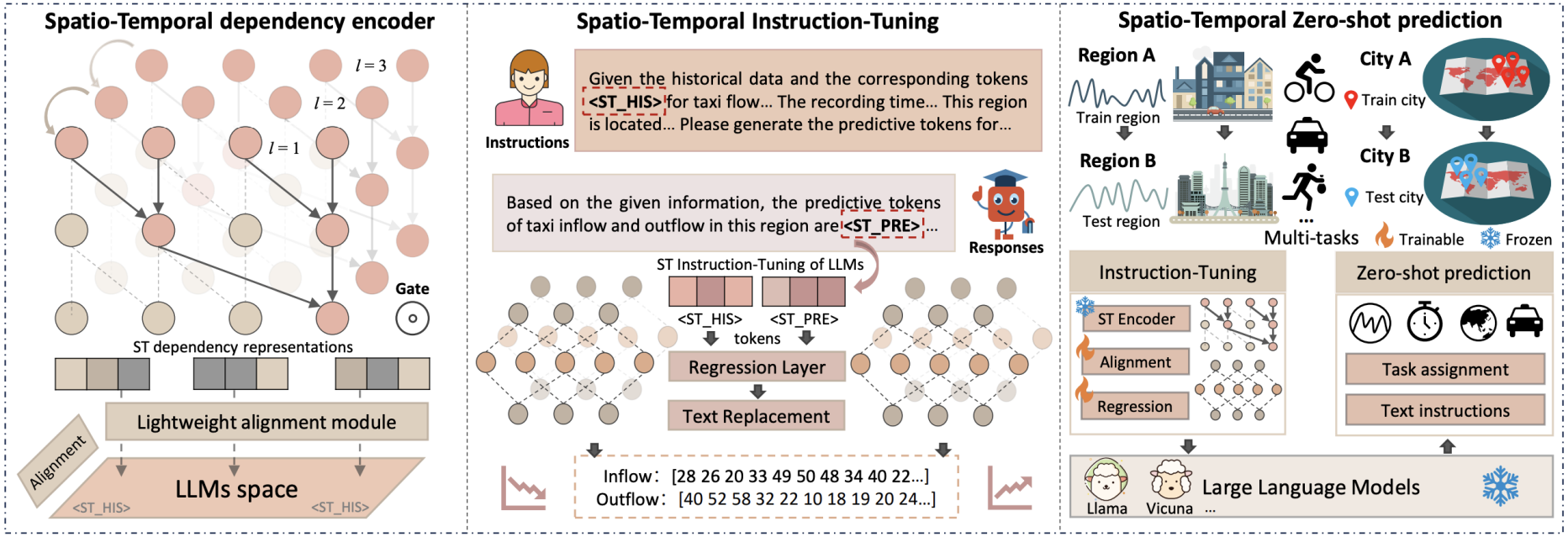
- **Attention-based**
  - Based on Cross-Attention mechanism
  - Query and Keys (Values) are from different modalities



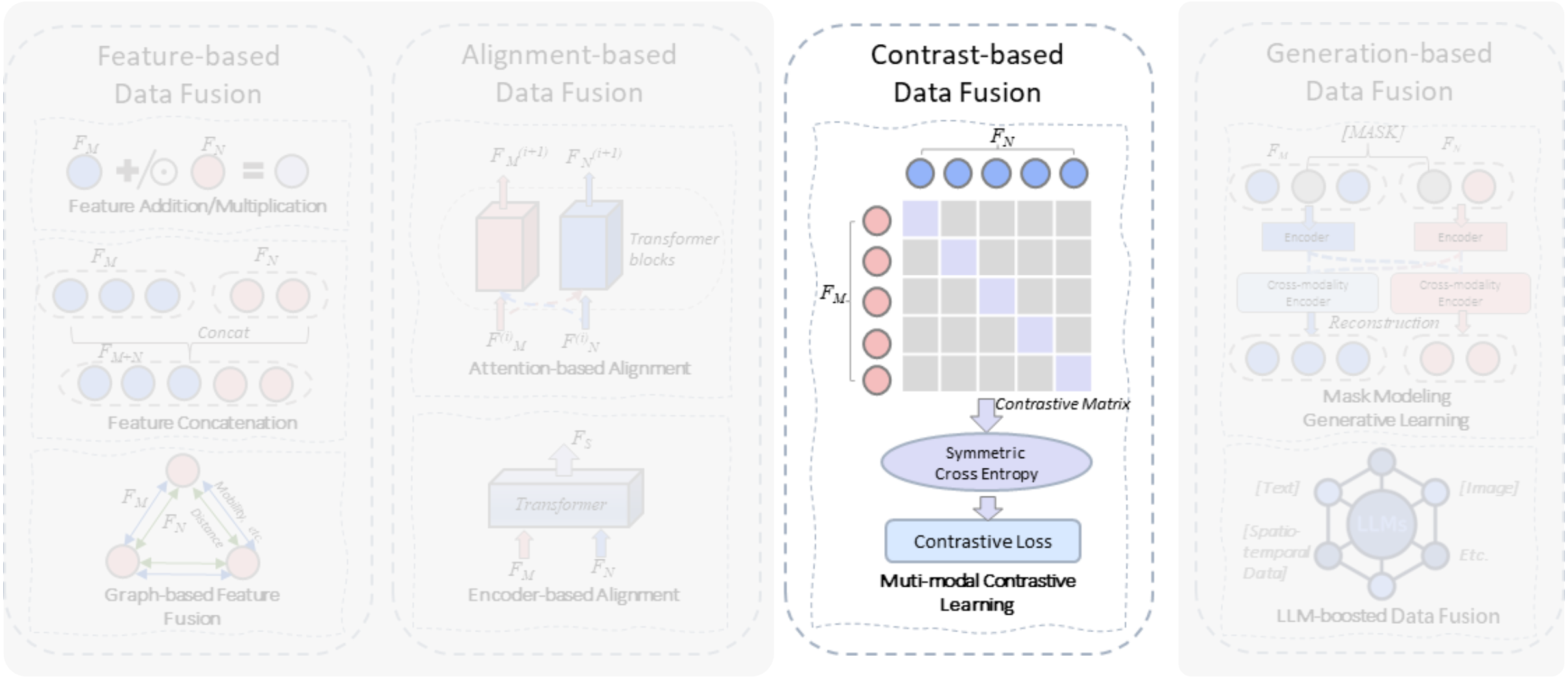

Multi-view joint graph representation learning for urban region embedding, IJCAI 2021

# Alignment-based Fusion

- **Encoder-based**
  - Token-level concatenation
  - Unified representations across modalities
  - Usually based on Self-Attention mechanism

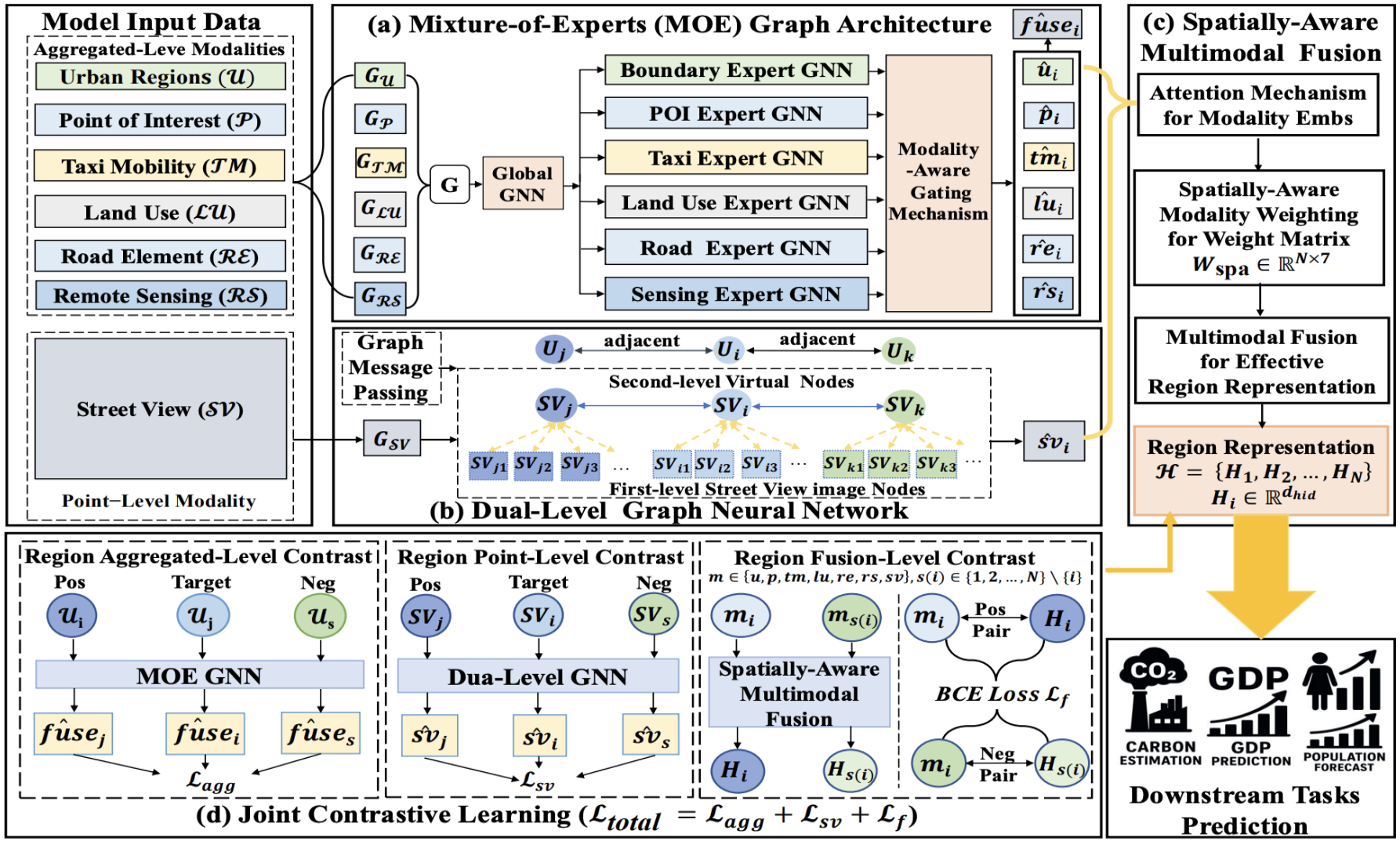


# Deep Learning-based Fusion Methods



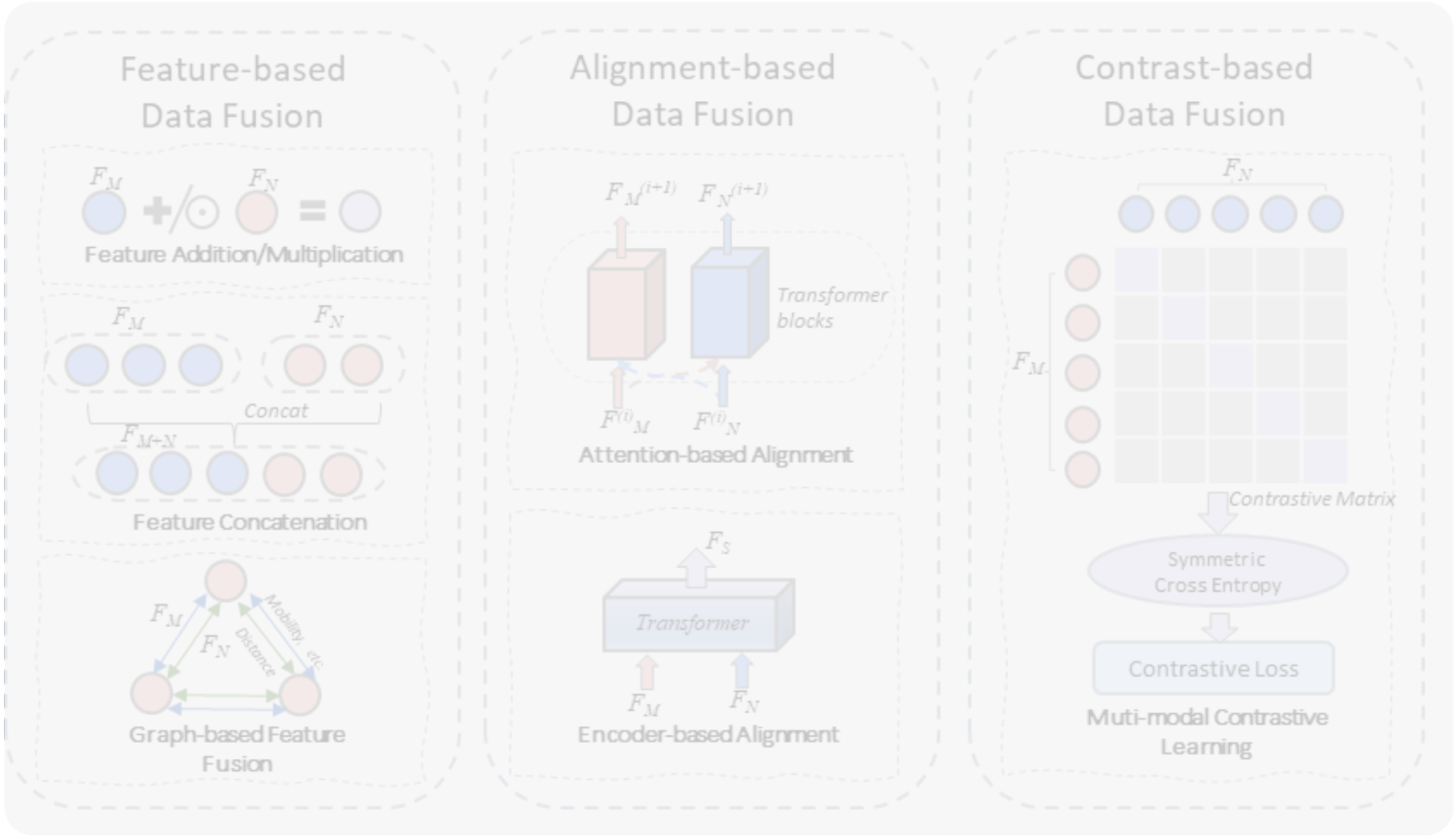

Deep Learning for Cross-Domain Data Fusion in Urban Computing: Taxonomy, Advances, and Outlook, Information Fusion, 2024.

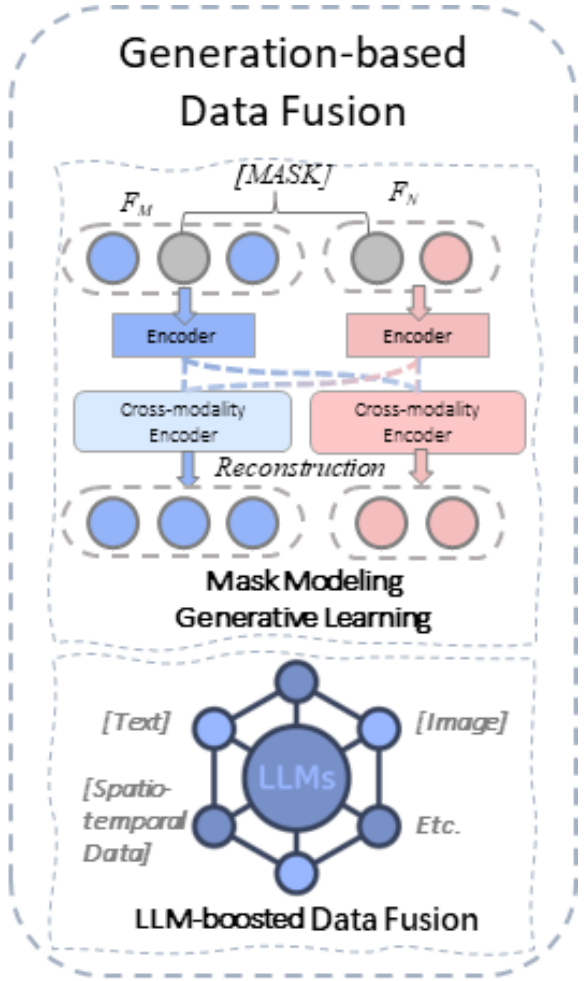
# Contrast-based Fusion




A Modality-Tailored Graph Modeling Framework for Urban Region Representation via Contrastive Learning. arxiv 2025

# Deep Learning-based Fusion Methods

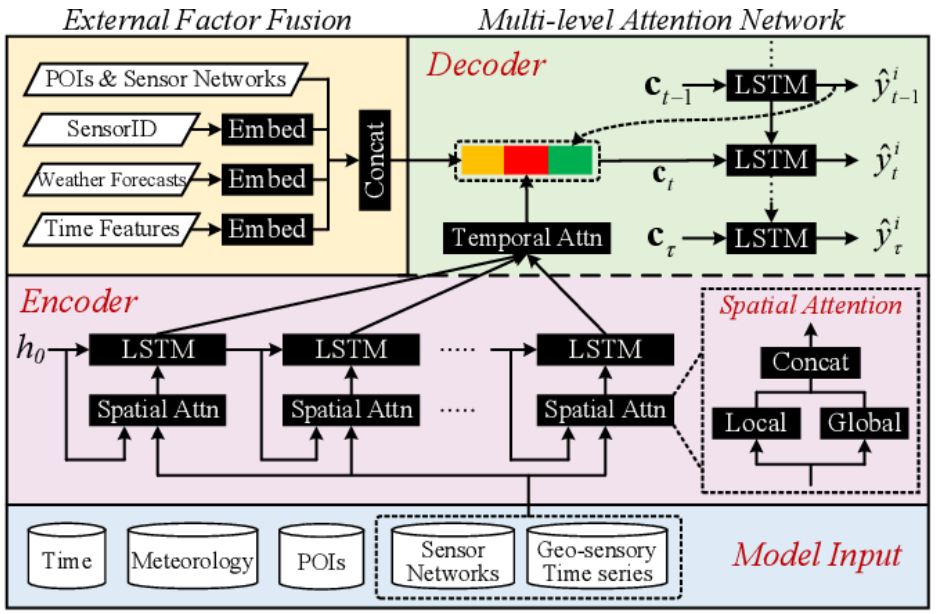


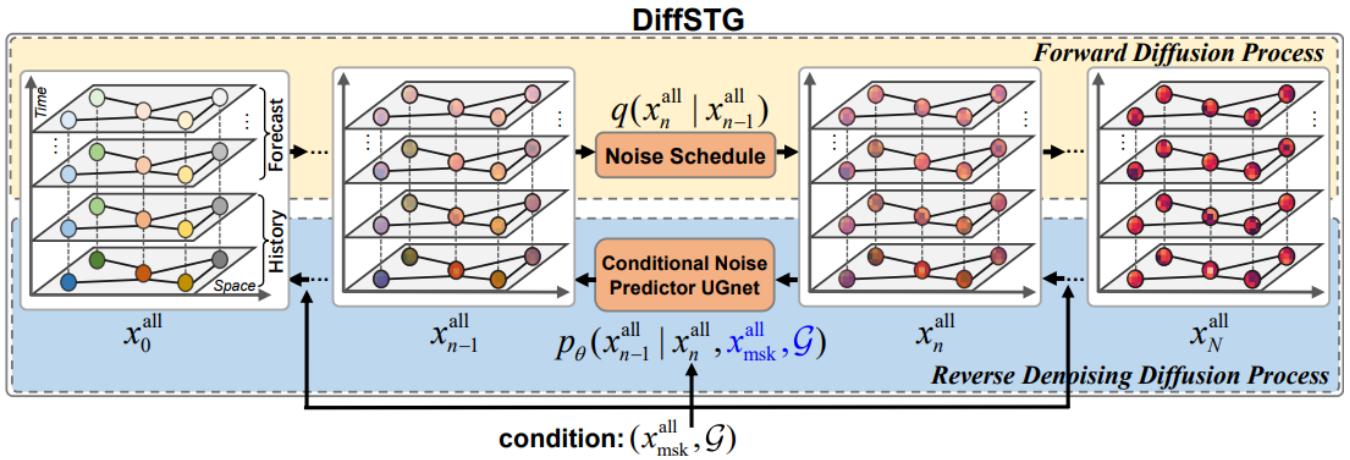

Deep Learning for Cross-Domain Data Fusion in Urban Computing: Taxonomy, Advances, and Outlook, Information Fusion, 2024.

# Generation-based Data Fusion

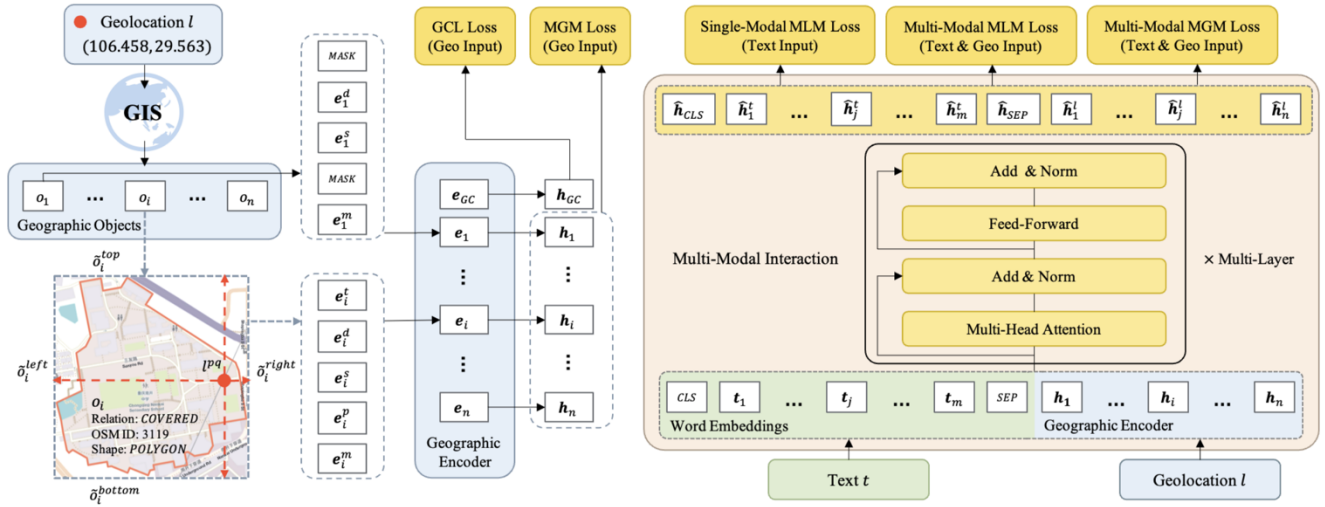
- Autoregression-based fusion
- Masked modeling-based fusion
- Diffusion-based fusion




GeoMAN: Multi-Level Attention Networks for Geo-Sensory Time Series Prediction. IJCAI 2018




DiffSTG: Probabilistic Spatio-Temporal Graph Forecasting with Denoising Diffusion Models. SIGSPATIAL 2023




MGeo: Multi-Modal Geographic Language Model Pre-Training. SIGIR 2023

# LLMs Open New Opportunities

- LLMs are good at processing **multi-modal data**
  - Multi-modal data understanding
    - e.g., vision, texts, time series
  - Multi-format data parsing
    - tabular, json, images
  - Example: Qwen-VL
- **Tool calling capability** (LLM Agents)

Who is in <box>(750,0),(999,999)</box> in the picture

It is Bill Gates, one of the founders of Microsoft.

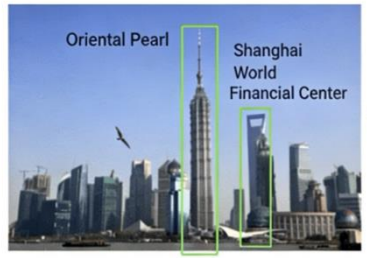
Describe his outfit

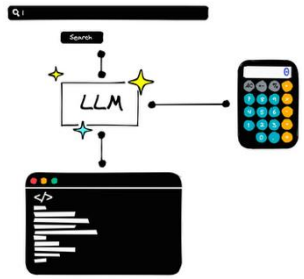
It is Bill Gates himself. He is wearing a black sweater and looks quite casual.

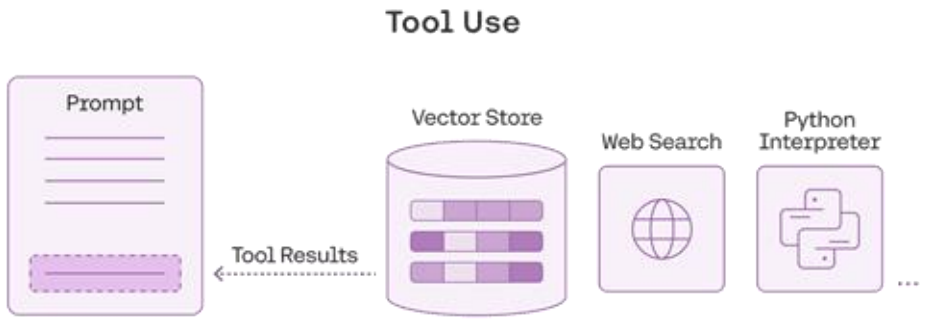
Where is this?

This is the Lujiazui Financial District in Shanghai, where you can see famous buildings such as the Oriental Pearl TV Tower and the Shanghai World Financial Center in the photo.

Find the Oriental Pearl and the Shanghai World Financial Center in the picture



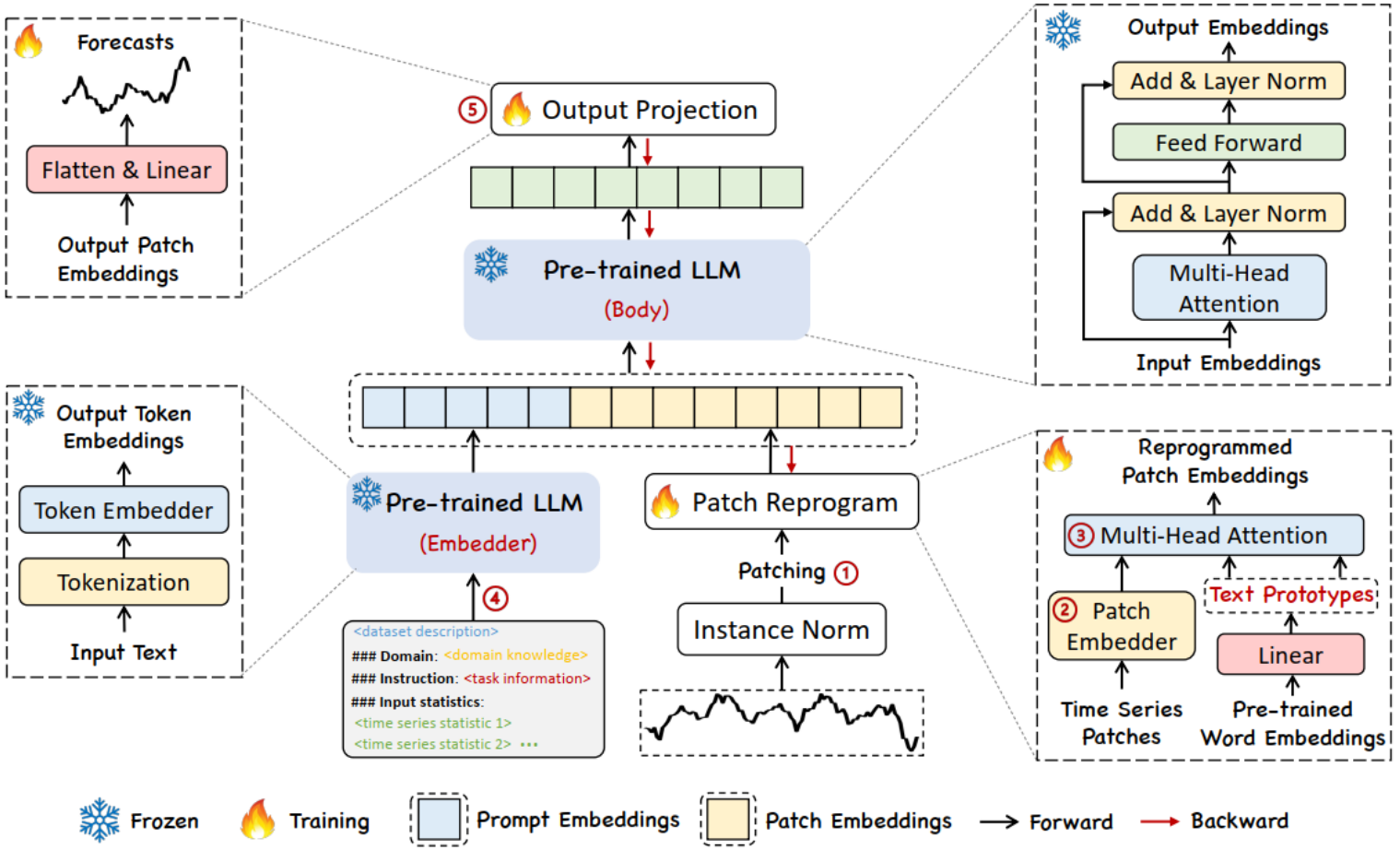



# LLMs for ST

- Our standpoint is that LLMs can serve as the central hub for **understanding** and **advancing** urban data fusion in two principal ways:
  - **LLM-as-Encoders**: utilizing their extensive internal knowledge and multi-modal understanding abilities to process multi-modal data
  - **LLM-as-Agents**: transcending conventional roles to actively engage in and transform spatio-temporal data mining
    - Tool-use agents
    - Multi-modal foundation agents

# LLM as Encoders

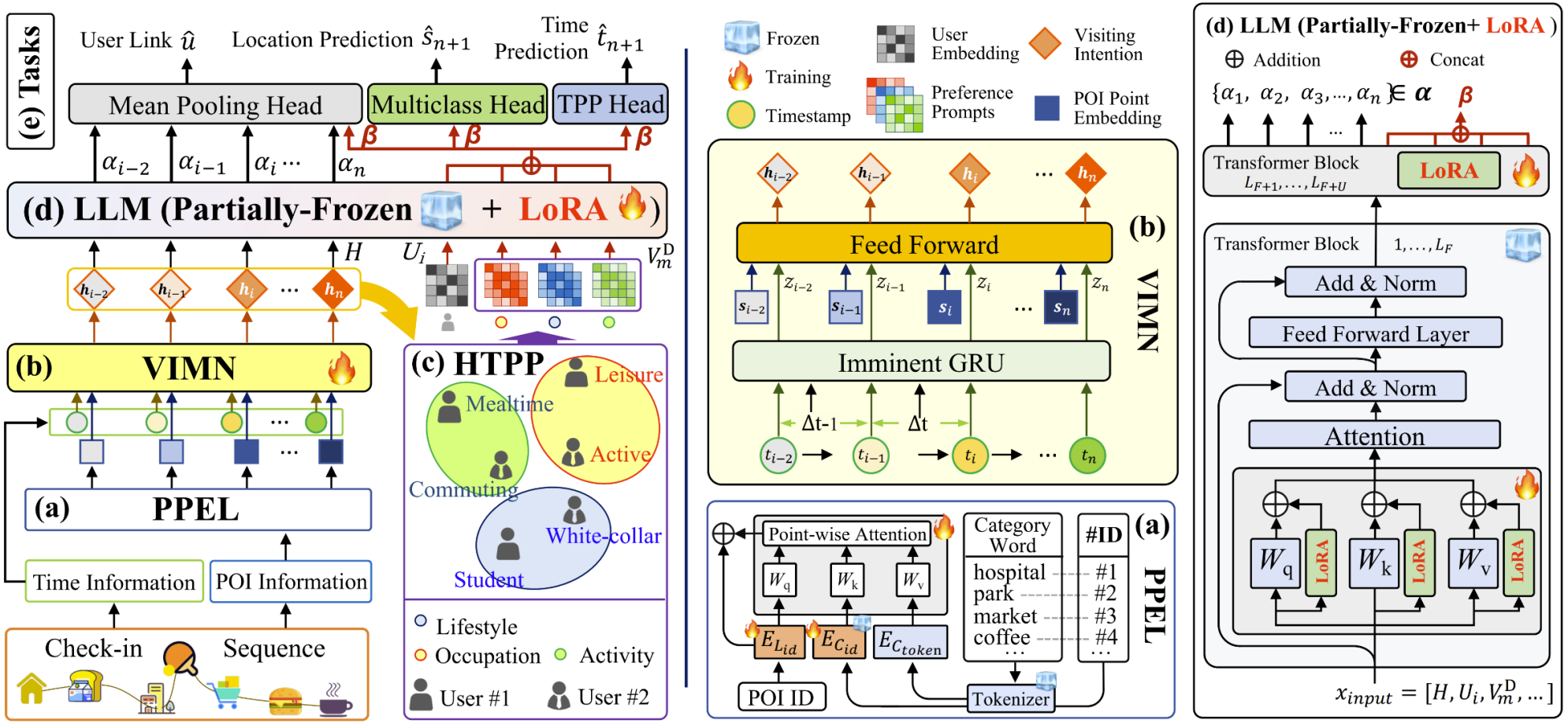
- Time-LLM: A pioneering work on LLM + Spatio-Temporal Data




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

# LLM as Encoders

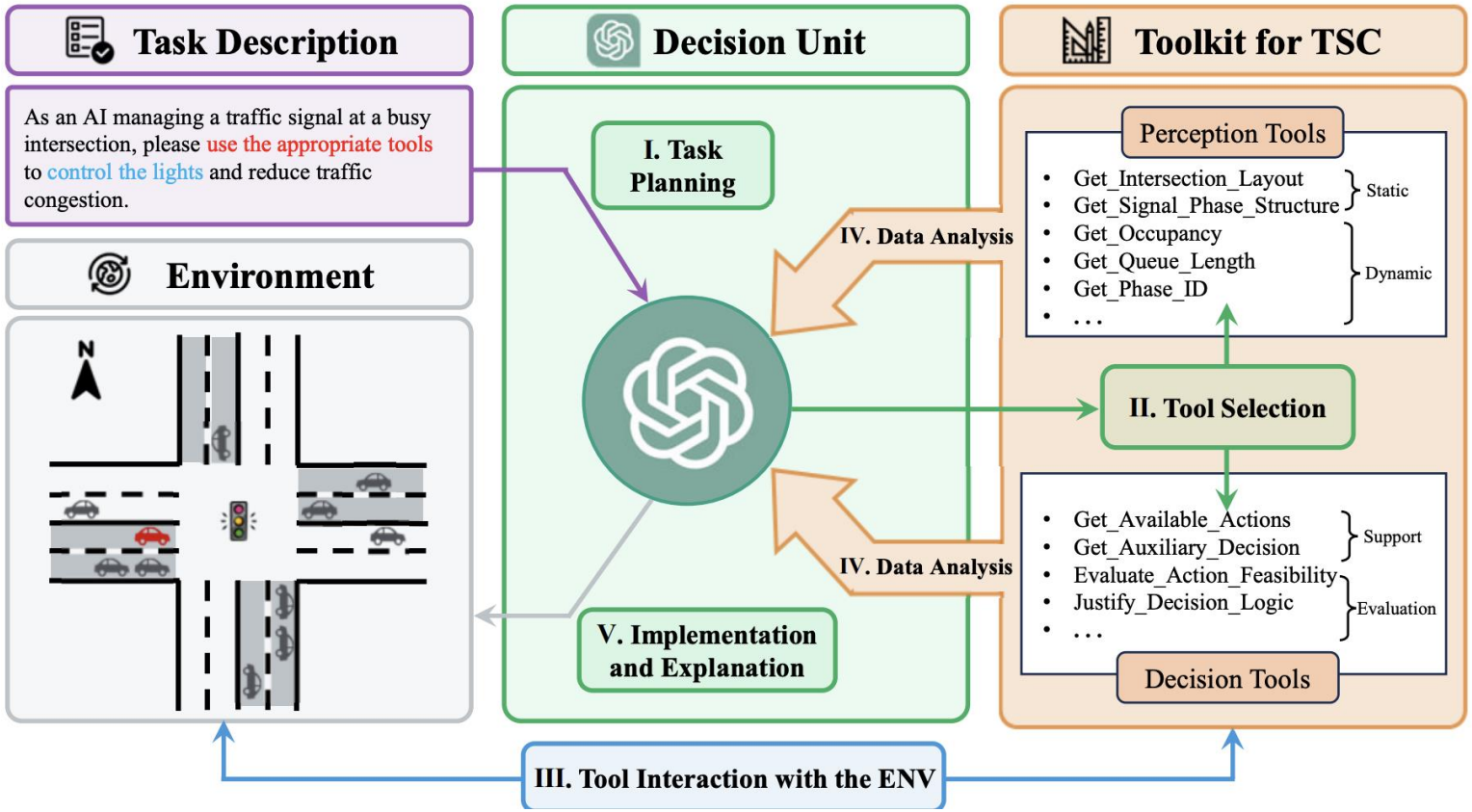
- Mobility-LLM: A pioneering work LLM + Trajectory Data




Mobility-LLM: Learning Visiting Intentions and Travel Preferences from Human Mobility Data with Large Language Models. NeurIPS 2024

# LLM as Agents — Traffic Signal Control

- By systematically calling various tools to integrate multi-source and multi-modal urban data, specific urban computing tasks can be accomplished.




LLM-assisted light: Leveraging large language model capabilities for human-mimetic traffic signal control in complex urban environments. arxiv 2024

# LLM as Agents — Reimaging Urban Data Science

What if we created **a team of agents** to simulate an **urban research lab**?

Urban Scientist

Data Scientist

Reader

Data Engineer

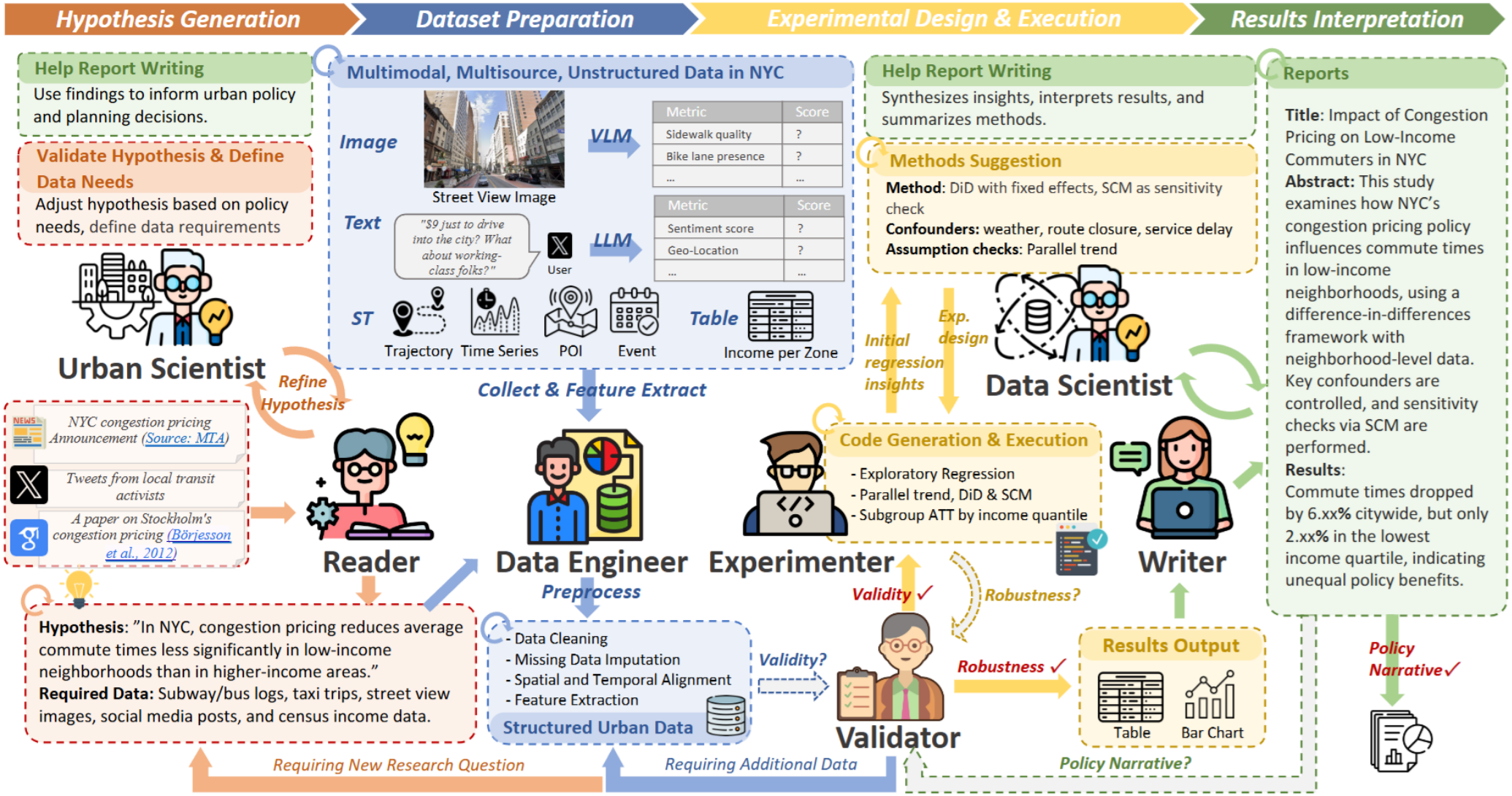
Experimenter

Writer

Validator

Reimagining Urban Science: Scaling Causal Inference with Large Language Models. arXiv 2025

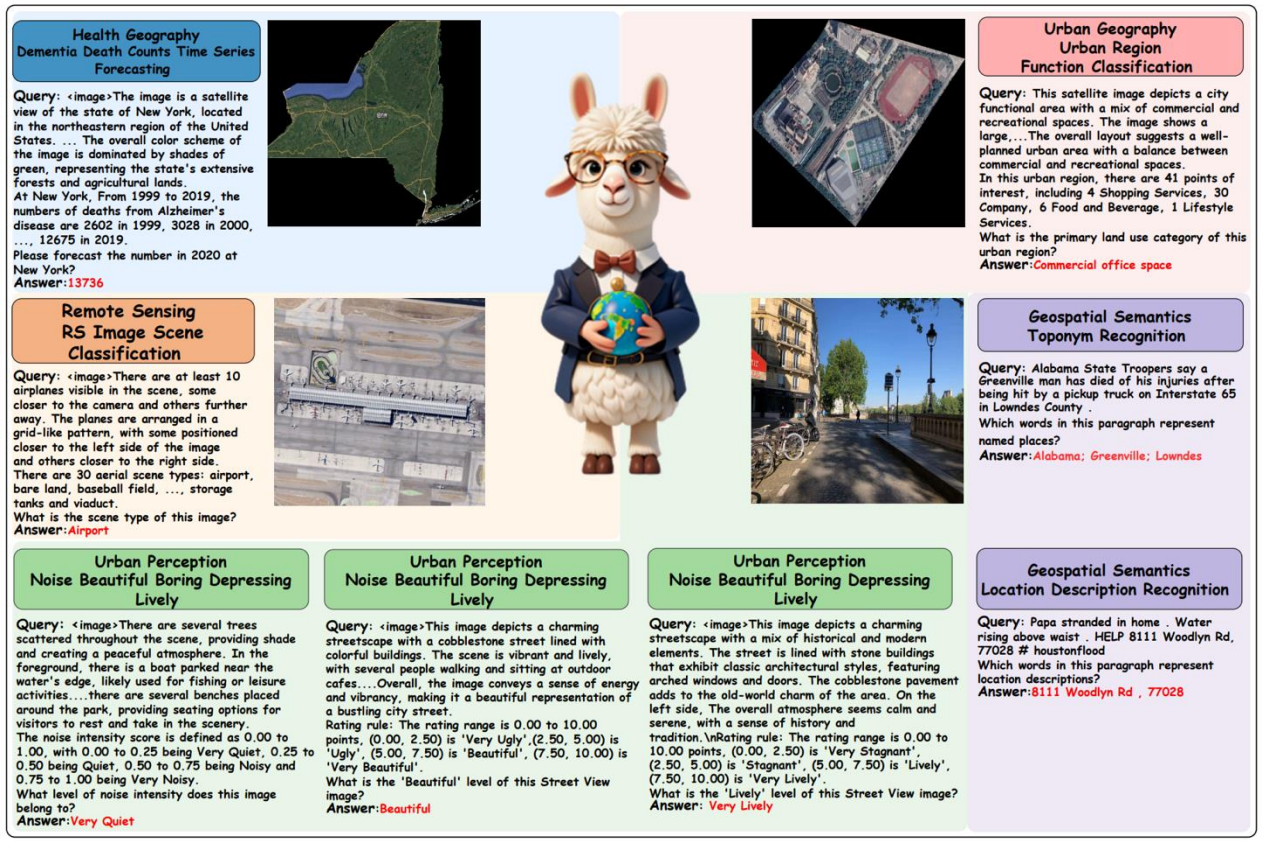
# LLM as Agents — Reimaging Urban Data Science




Reimagining Urban Science: Scaling Causal Inference with Large Language Models. arXiv 2025

# Multimodal Foundation Agents

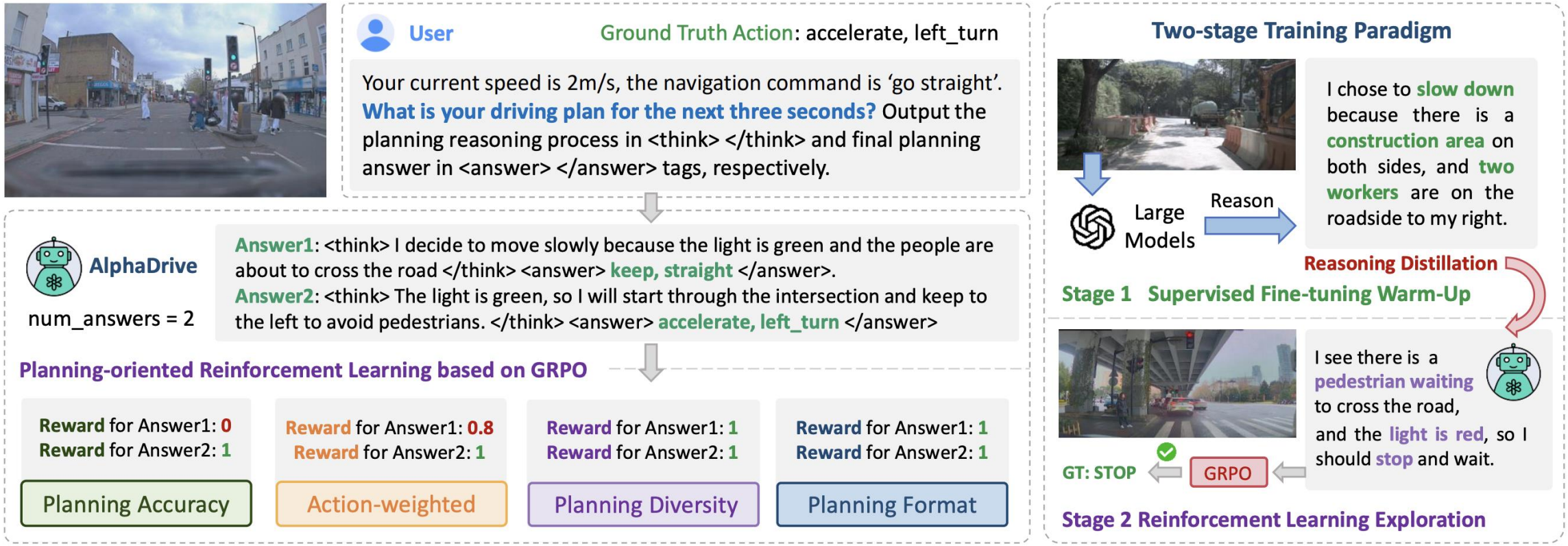
- **OmniGeo:** Unifies diverse geospatial tasks (e.g. health geography forecasting, remote sensing scene classification, urban perception, and geospatial semantic understanding) by seamlessly integrating textual and visual inputs.




OmniGeo: Towards a Multimodal Large Language Models for Geospatial Artificial Intelligence. arXiv 2025

# Multimodal Foundation Agents

- **AlphaDrive**: Vehicle speed / orientation / task information / images => MLLMs => GRPO => stable multimodal inference in autonomous driving scenarios.



# Earth AI — From Prediction to Decision

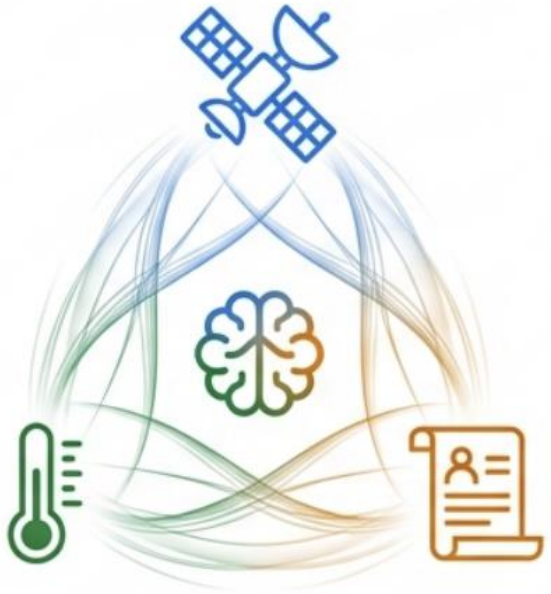
## The Old Paradigm: Siloed Prediction

**Siloed Data:** Imagery, demographics, and weather exist in isolation.

**Single-Task Models:** Focus on raw value prediction (e.g., 'Classify this pixel').

**High Friction:** Requires manual cross-referencing and deep domain expertise.

## The Earth AI Paradigm: Holistic Decision



**Multi-Modal Foundations:** Interoperable models for Imagery, Population, and Environment.

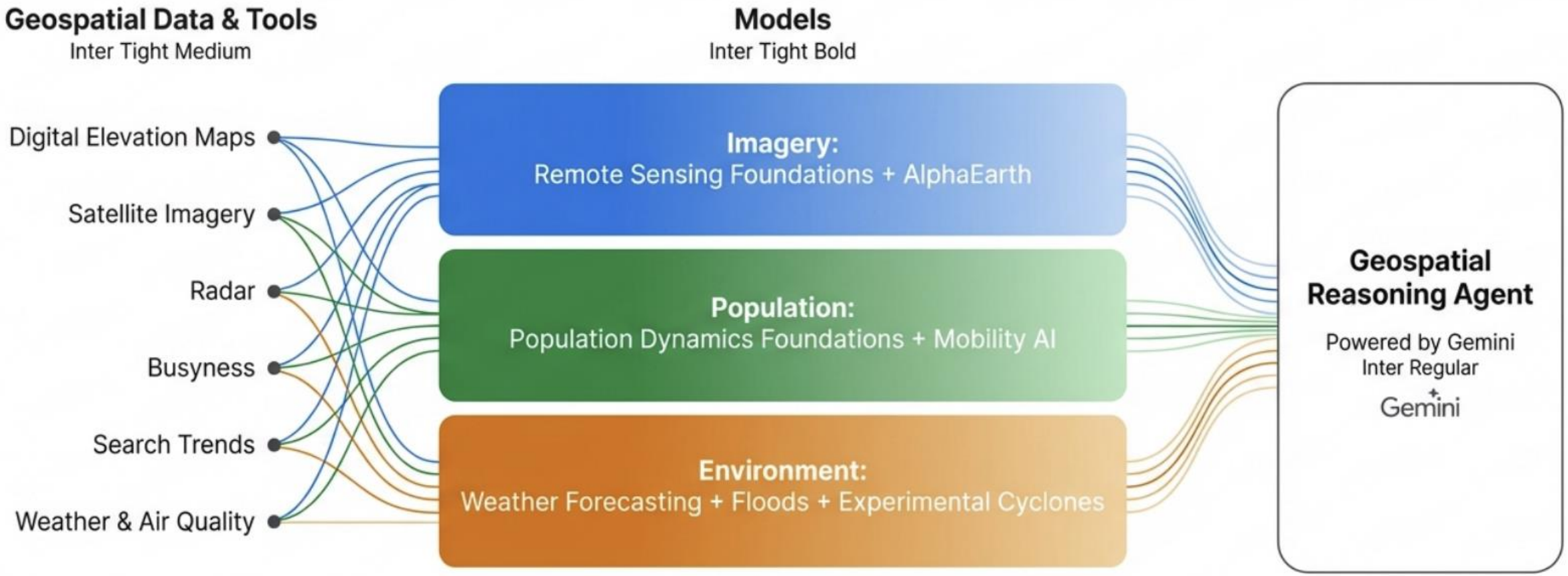
**Synergistic Inference:** Combined modalities yield superior accuracy.

**Agentic Reasoning:** Answers "What should we do?" using Gemini-powered agents.

*"The goal is no longer just observing the Earth, but reasoning about it to solve complex real-world problems."*

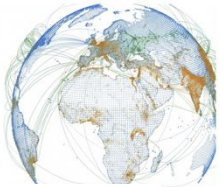
Earth AI: Unlocking Geospatial Insights with Foundation Models and Cross-Modal Reasoning, 2025.

# Agents Pipeline




Earth AI: Unlocking Geospatial Insights with Foundation Models and Cross-Modal Reasoning, 2025.

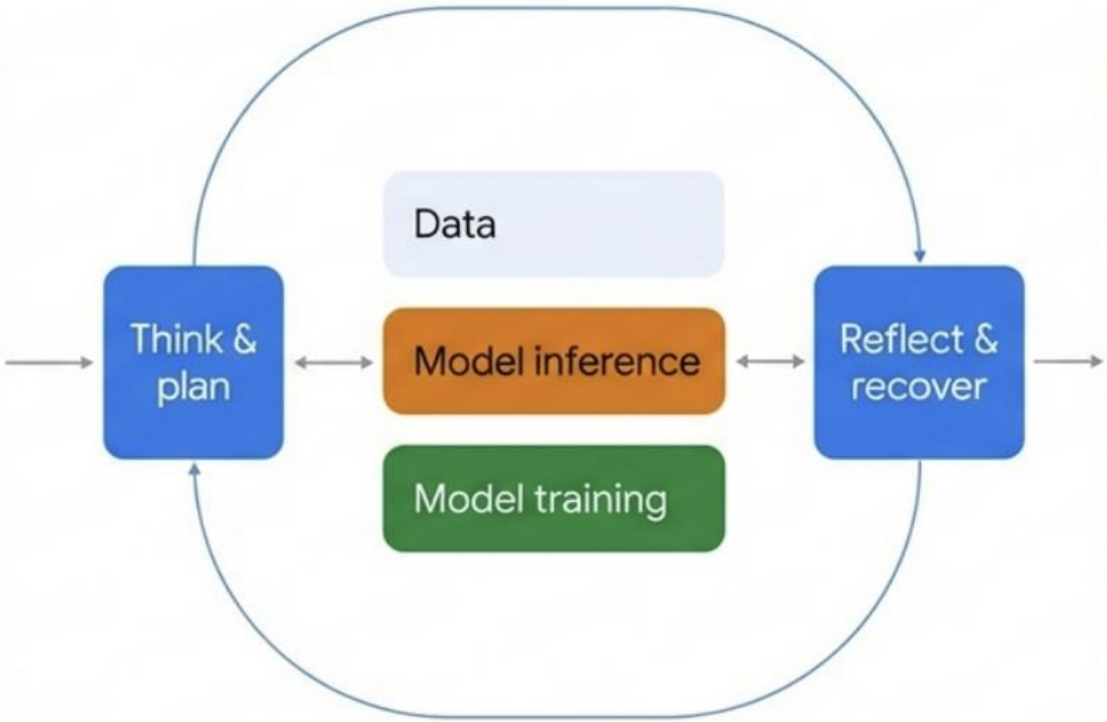
# Geospatial Reasoning Agent



## Analytical Queries

Standard models excel at retrieval ('What is the temperature?').

They struggle with inferential logic: 'Where should we send aid based on flood risk *and* elderly population?'




## Gemini-Powered Reasoning

Acts as an intelligent intermediary.

Deconstructs natural language queries into sub-tasks.

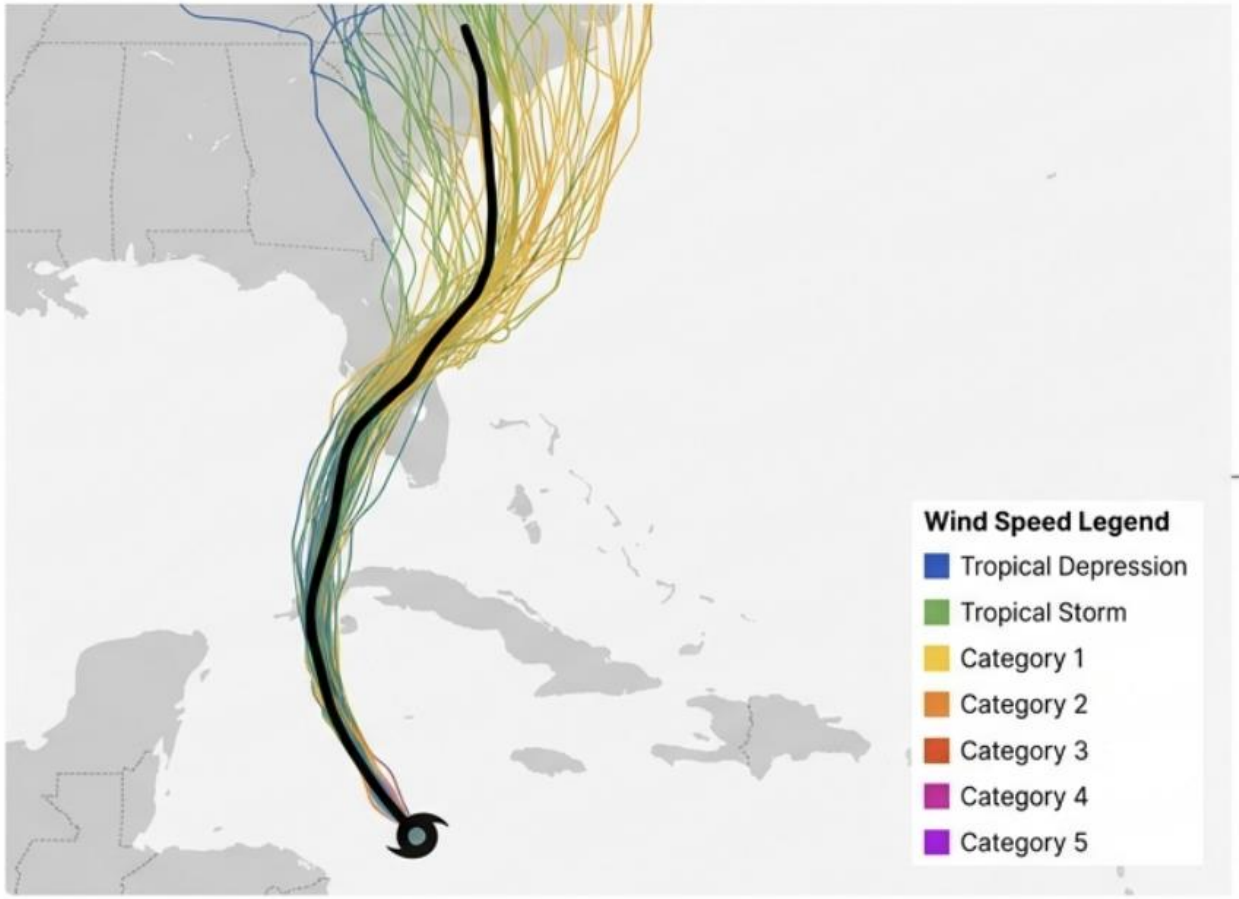
Selects the right 'Expert' tool (Imagery, Population, Weather).

Synthesizes outputs into actionable answers.

Earth AI: Unlocking Geospatial Insights with Foundation Models and Cross-Modal Reasoning, 2025.

# Model Combination Improves Performance

Integrating diverse viewpoints yields higher accuracy than single-modality analysis.




## Key Metrics

**FEMA Risk Scores:** Combining **Population Dynamics + AlphaEarth** (Imagery) = 11% increase in $R^2$.

**Public Health:** Combined embeddings improved prediction of CDC health stats (e.g., diabetes) by 7–43%.

**Cyclone Damage:** Predicting wind damage to buildings with 97% accuracy (Hurricane Ian case study).

# Case Study: Cholera Forecasting

Proactive healthcare through environmental integration.

## Methodology

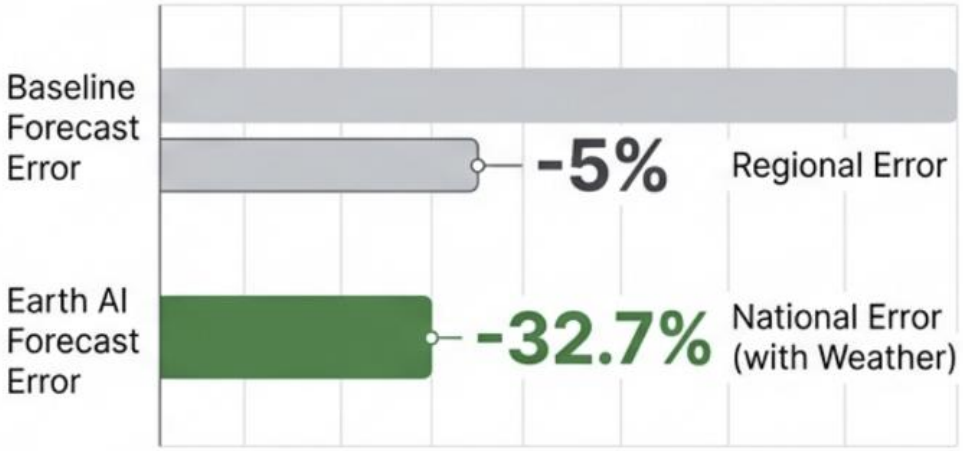
- **Challenge:** Endemic cholera requires proactive resource allocation.
- **Input:** Historical Case Counts + Population Dynamics (Socio-economic) + Weather Forecasts (Precipitation).

Historical Case Counts — Population Dynamics (Socio-economic) — Weather Forecasts (Precipitation)

- **Model**
  TimesFM (Time-series Foundation Model).

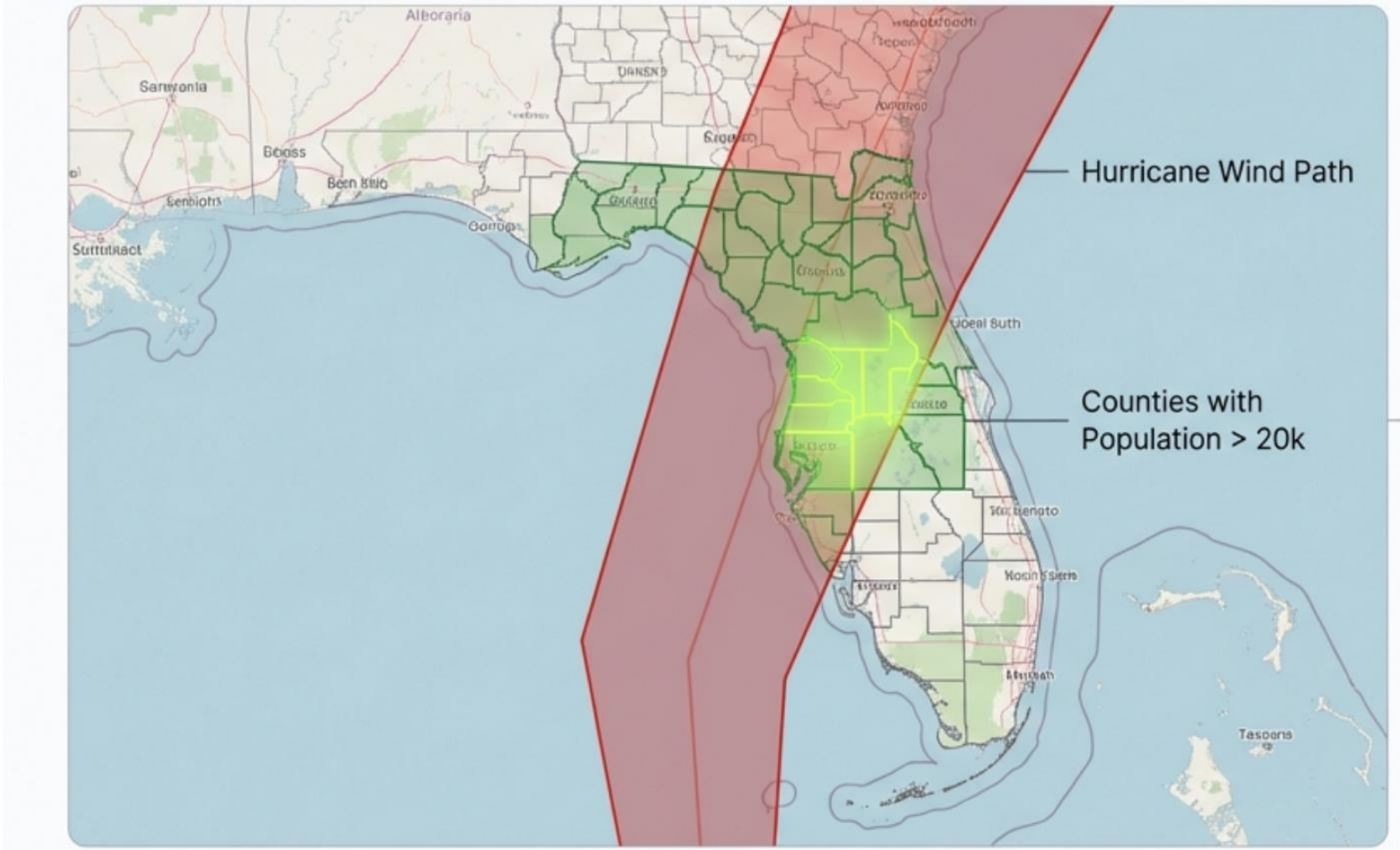
## Results

**Forecast Error Reduction (%)**




Augmenting time-series models with weather and population data significantly reduces forecast error, enabling better preparation.

# Case Study: Crisis Response (Hurricane Helene)




**Geospatial Reasoning**

**User:** Identify counties in Florida with population >20k predicted to face hurricane-force winds.

**Agent Thinking**

1. **Fetch Forecast:** Query Cyclone model for wind speed polygons.
2. **Fetch Data:** Query Data Commons for FL county populations.
3. **Spatial Join:** Intersect wind polygons with county geometries.
4. **Filter:** Keep counties where pop > 20,000.

**Result:** Here are the actionable targets...

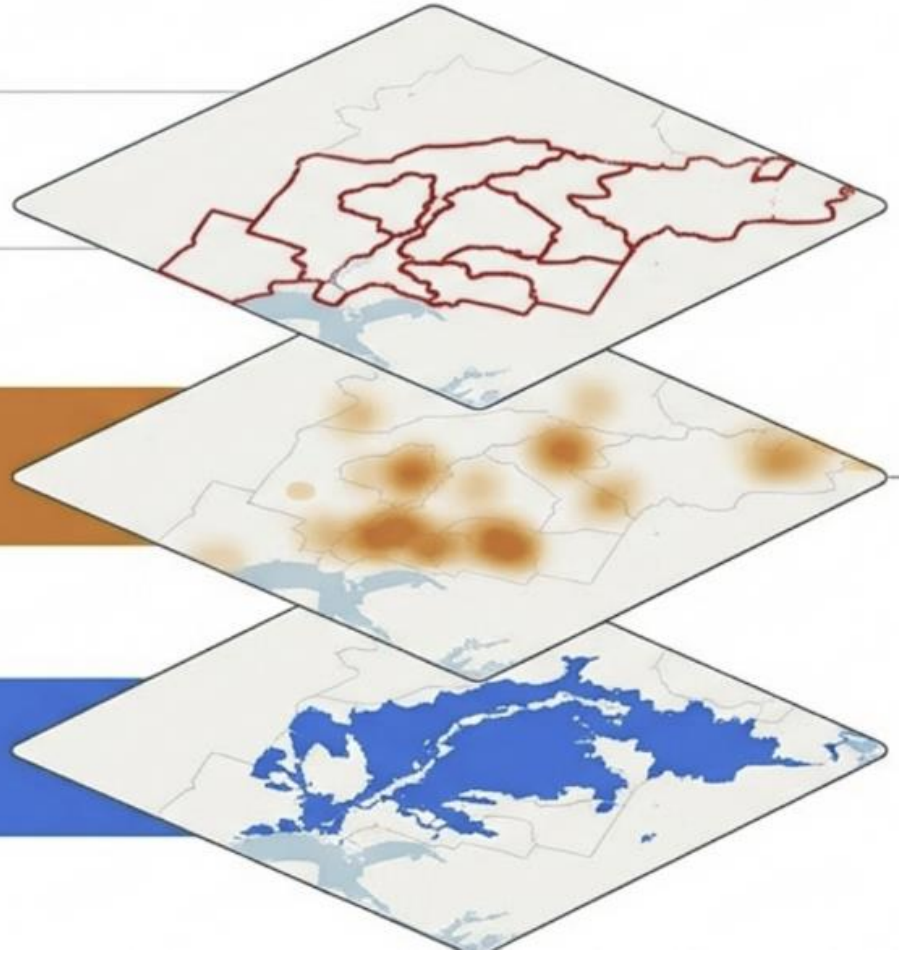
# Case Study: Flood Risk & Social Vulnerability

**Query:** Find zip codes in Matanuska-Susitna Borough with high flood risk and high percentage of vulnerable populations (age > 60).

**Decision Support:**
Priority List

**Population Model:**
Demographics (Age > 60)

**Flood Model:**
Inundation Zones



1. **Flood Expert** identifies inundation.

2. **Demographics Expert** retrieves age distribution.

3. **Agent** cross-references to prioritize evacuation.

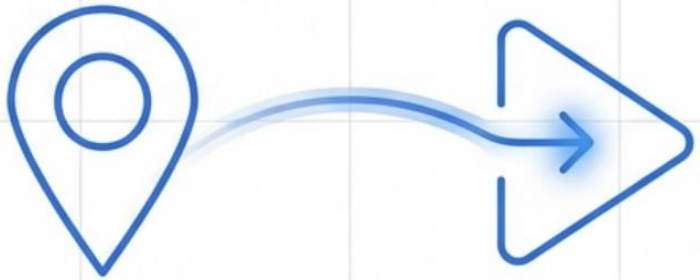
# Future Research

## From Observation to Action

Earth AI bridges the physical, human, and environmental worlds.

By combining the '**Eyes**' of Foundation Models with the '**Brain**' of Agentic Reasoning, we unlock the ability to answer the most critical question of all:

## What should we do next?



Earth AI: Unlocking Geospatial Insights with Foundation Models and Cross-Modal Reasoning, 2025.

# Summary

- **Integration into the Physical World**

  Cross-domain ST multimodal fusion acts as a critical link to the physical world. It goes beyond simple data combination, enabling in-depth understanding of complex real-world dynamics.

- **Revolution by Foundation Models**

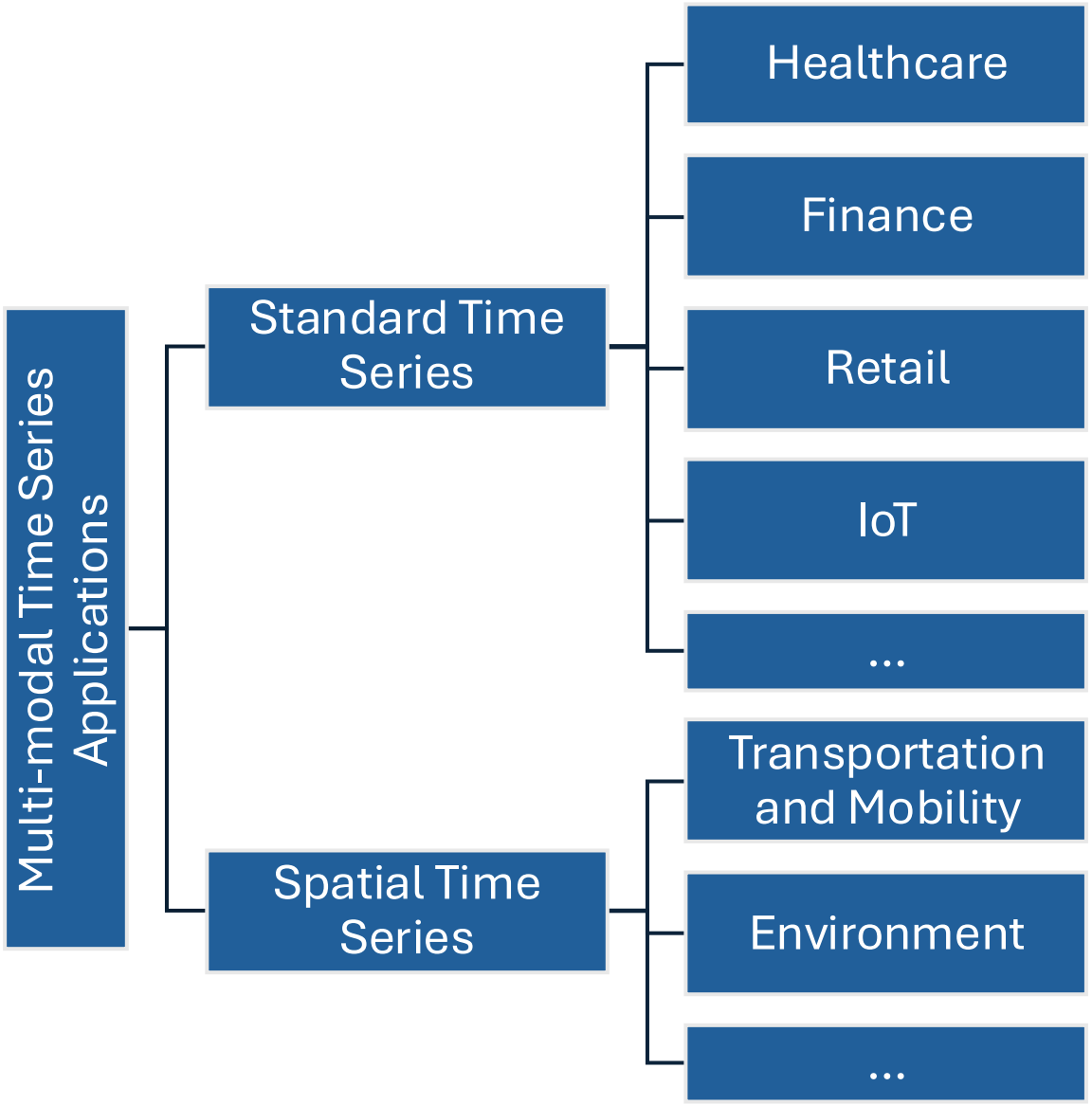
  FMs became core cognitive engine for ST learning, drive the entire STDM workflow—from autonomous hypothesis generation and reasoning to complex data engineering.

- **From Prediction to Decision**

  The paradigm of ST computing is shifting from passive numerical analysis to active decision intelligence, to generate actionable insights for guiding urban operations and policy-making.

# *Multi-modal Time Series Application and Datasets*

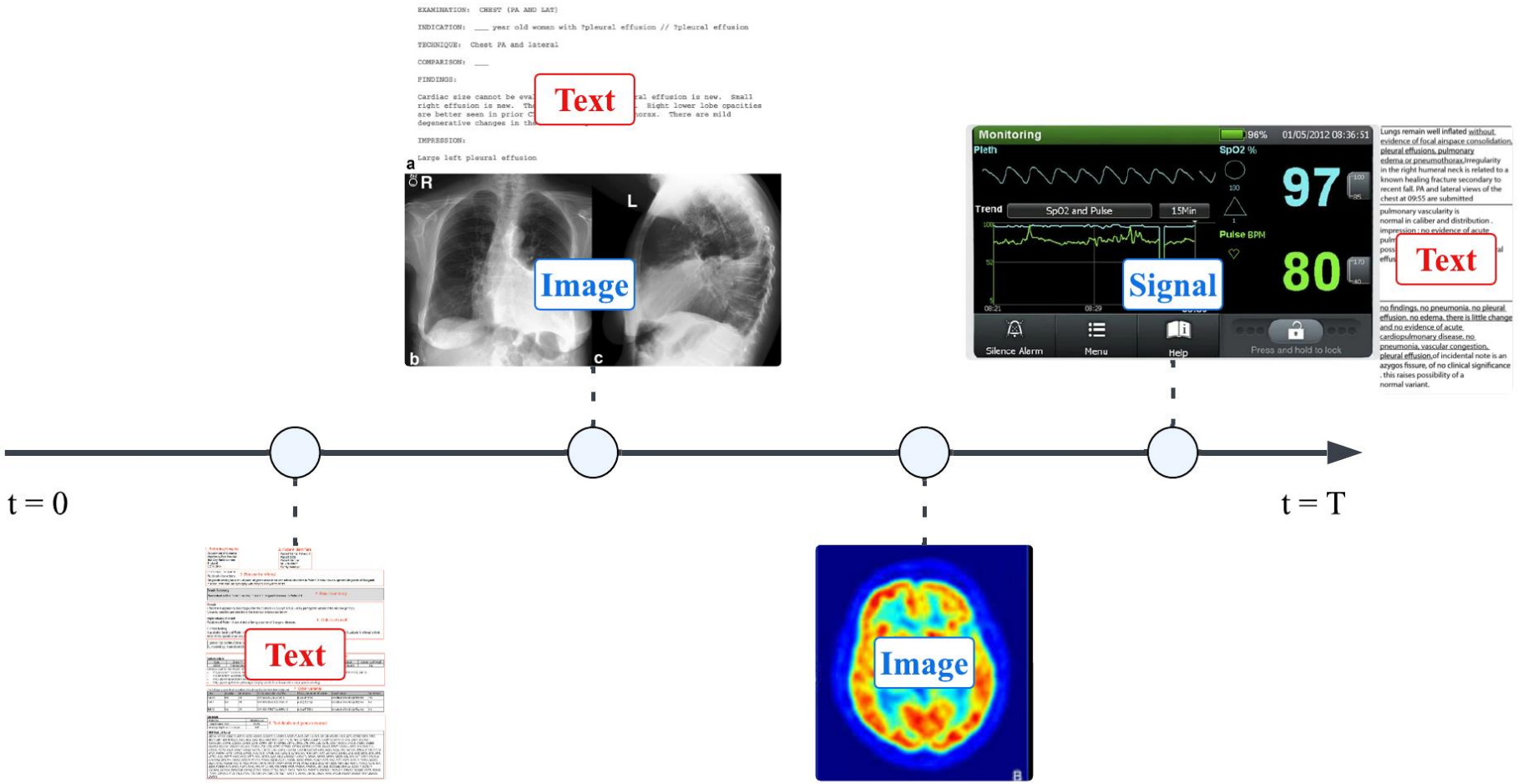
# Multi-modal Time Series Applications




- 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…*

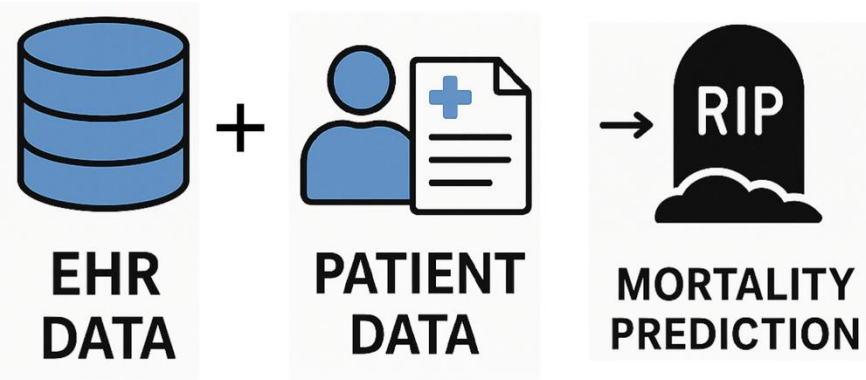
# Healthcare - EHR

- **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**



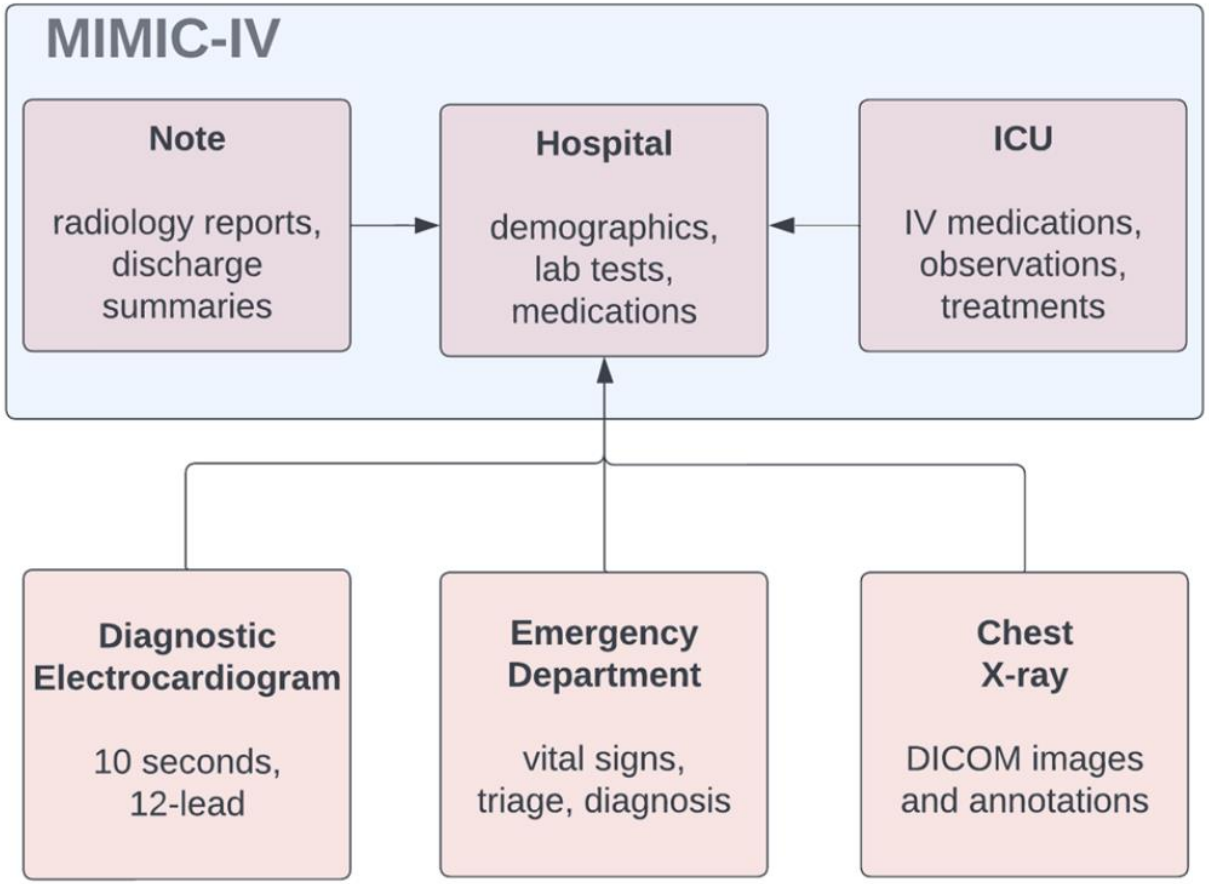
# Healthcare – EHR Datasets

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.

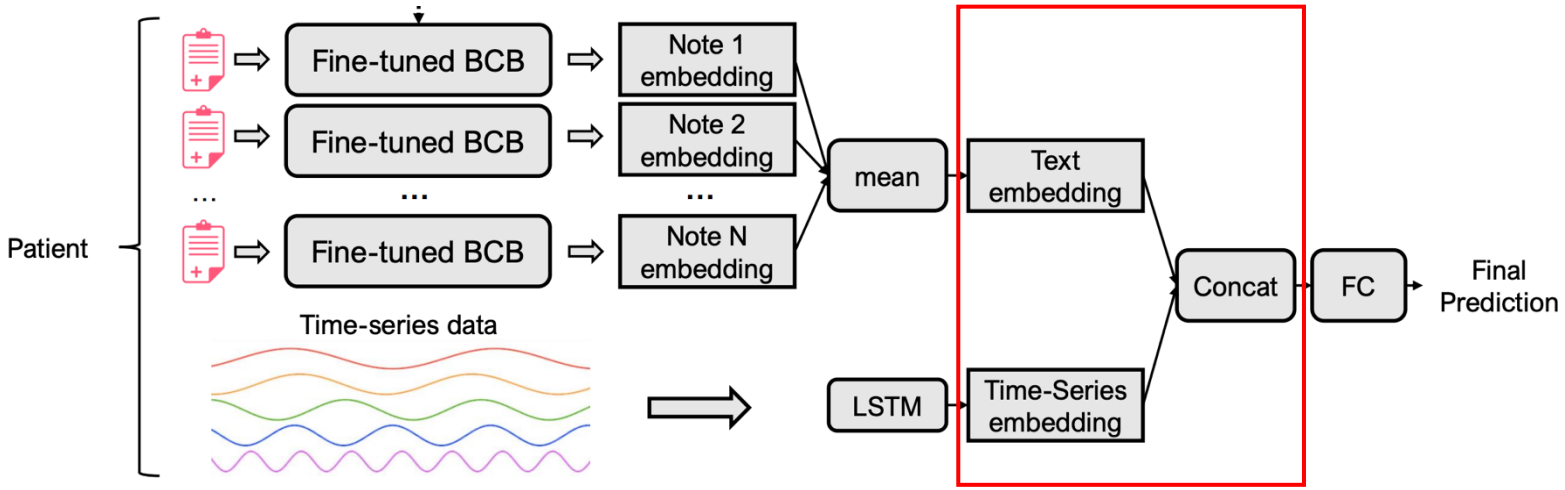
# Healthcare – EHR Datasets

| | 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

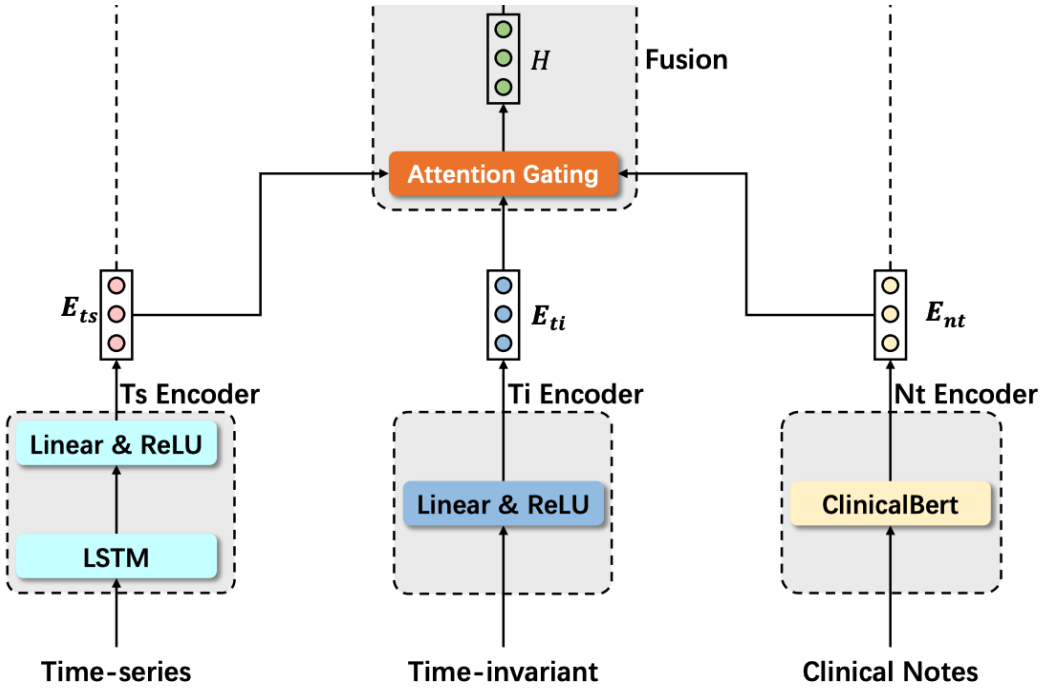
- **Leverage multi-modality data lab values and clinical reports**
  - Concatenation




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

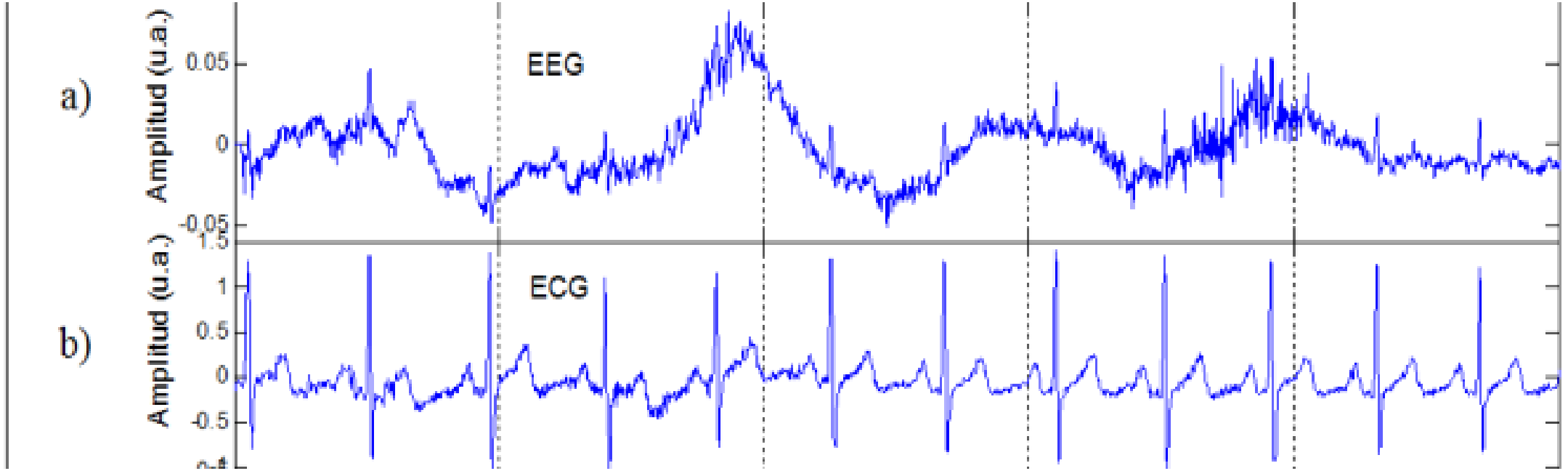
# Healthcare – EHR Modeling

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



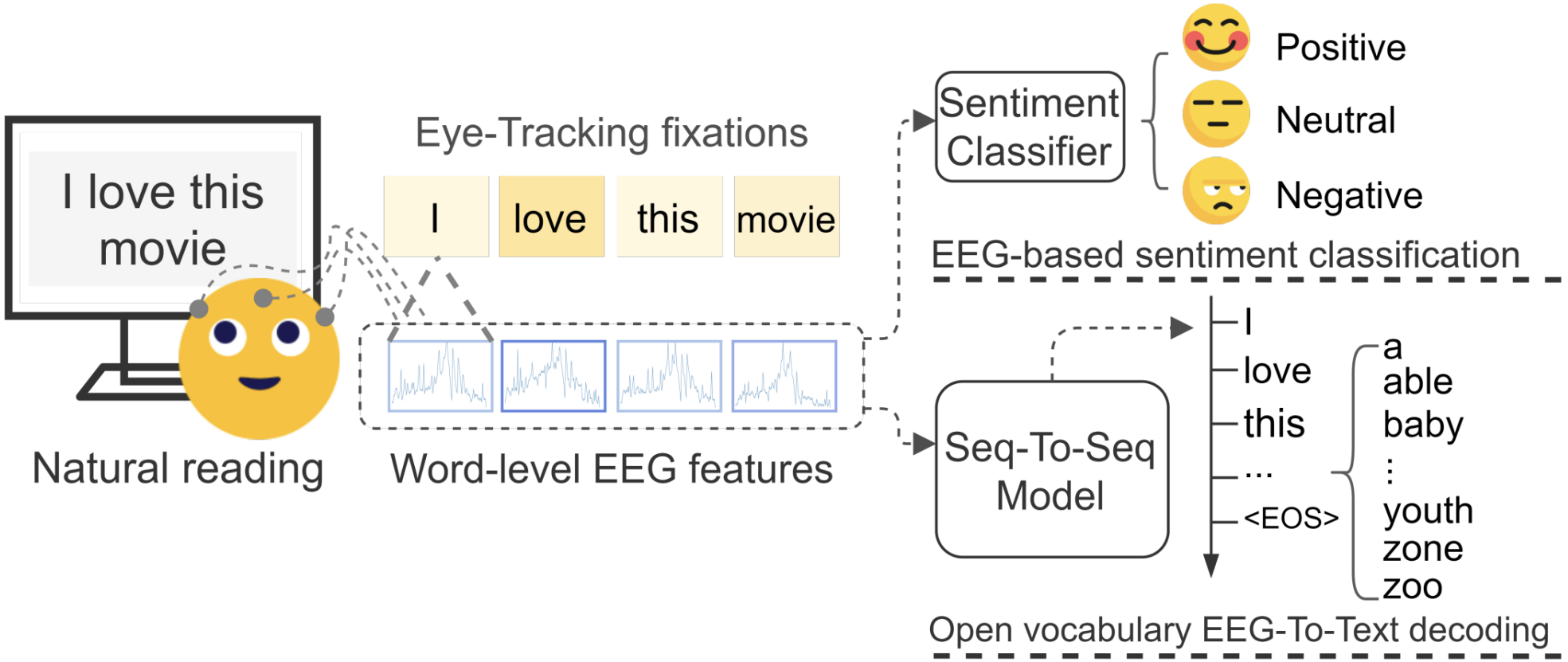

**Yang et al. "How to leverage multimodal EHR data for better medical predictions", EMNLP 2021**

# 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 *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. |

**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 readibility 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

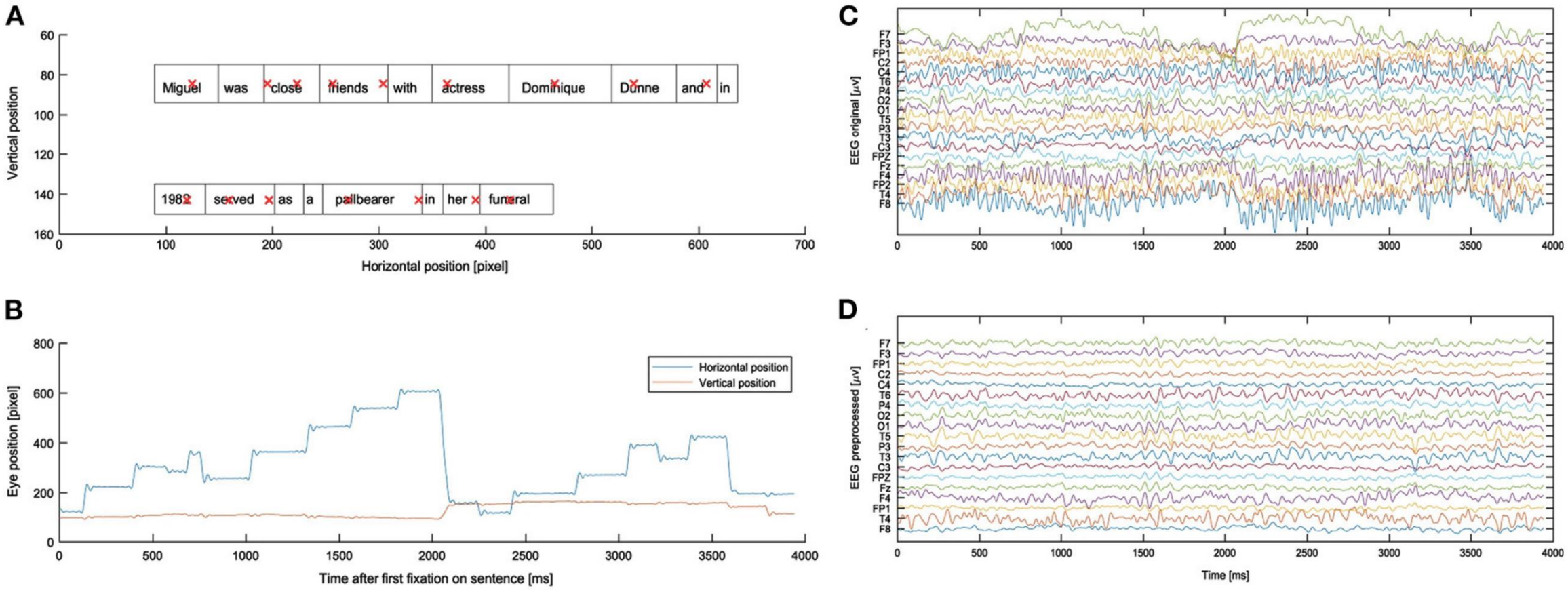
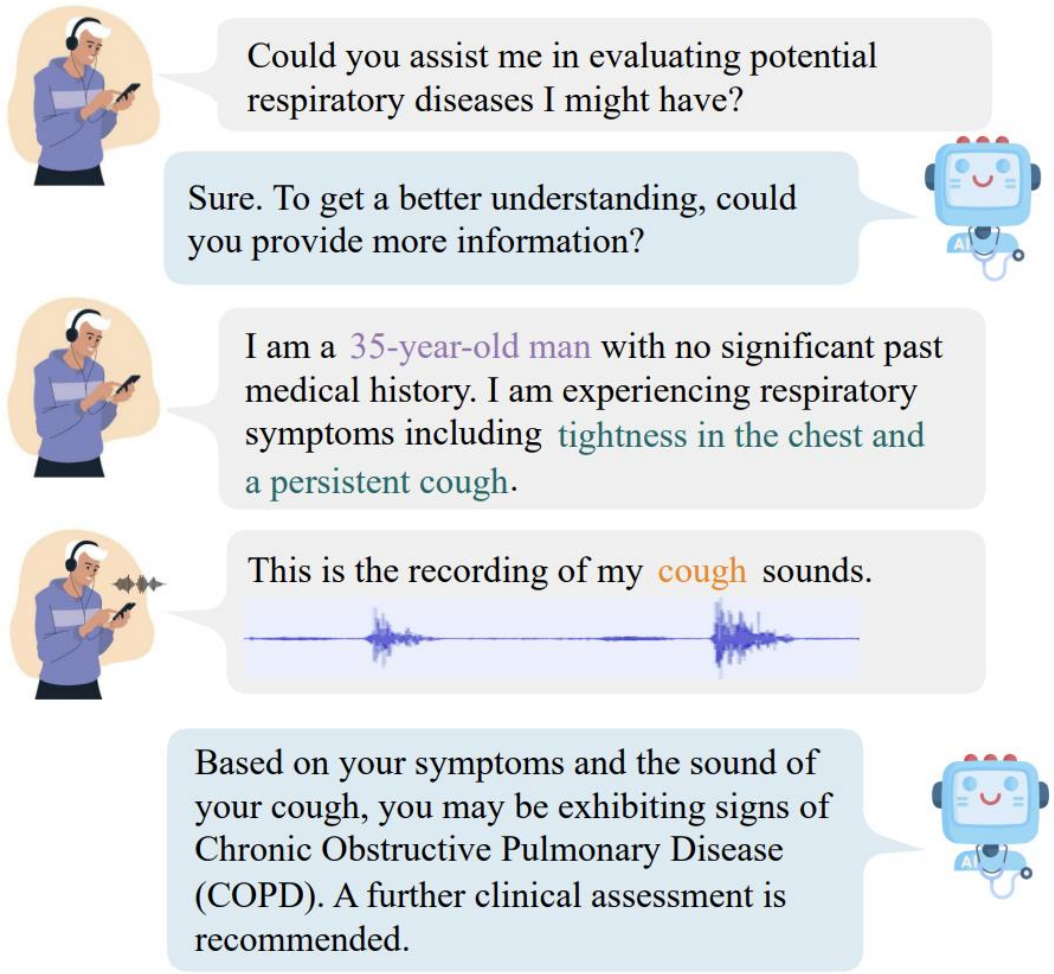



FIGURE 3

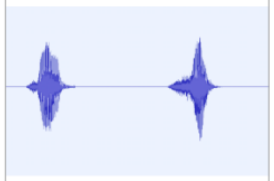
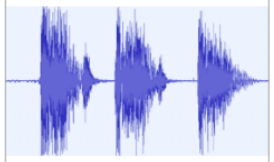
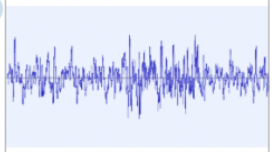
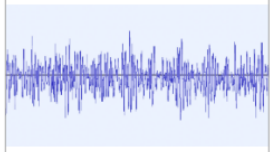
Visualization of eye-tracking and EEG data for a single sentence. **(A)** Prototypical sentence fixation data. Red crosses indicate fixations; boxes around the words indicate the wordbounds. **(B)** Fixation data plotted over time. **(C)** Raw EEG data during a single sentence. **(D)** Same data as in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

# Healthcare – Audio data

- **Incorporating with audio data for respiratory health screen**



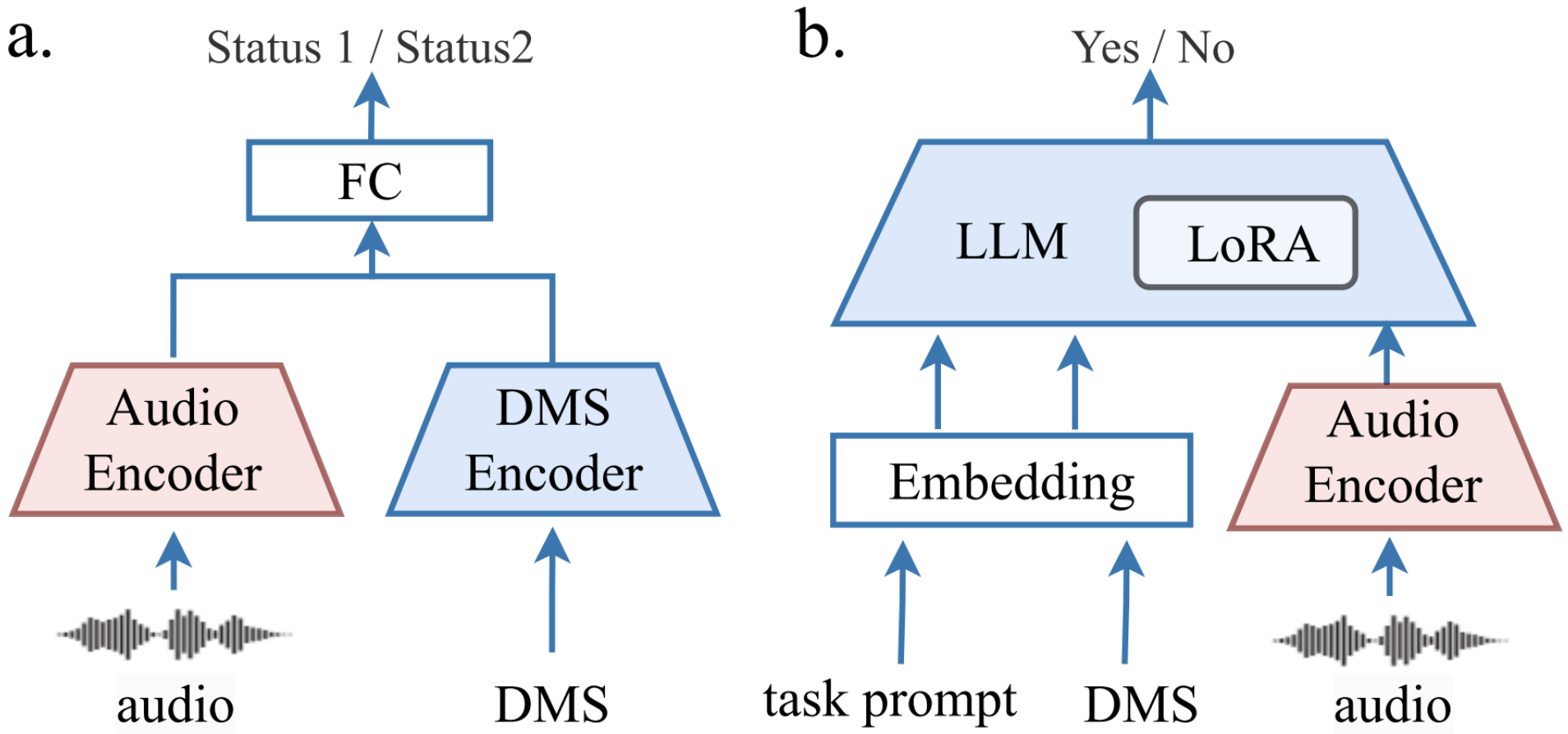
# Healthcare – Audio data

| Task | Text | Audio | Answer |
|---|---|---|---|
| S1 (Training) | **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.<br>**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. |  | 1 |
| S6 (Training) | **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.<br>**DMS text**: Gender: Female. Age: 50-59. Patient presents with no medical history conditions. Patient presents with no obvious respiratory symptoms. |  | 0 |
| S7 (Training) | **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.<br>**DMS text**: Gender: M. Age: 65. Record location: right posterior chest. |  | 1 |
| T4 (Testing) | **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.<br>**DMS text**: Gender: male. Age: 35. Patient presents with the following respiratory symptoms: cold. |  | 0 |
| T6 (Testing) | **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.<br>**DMS text**: Gender: F. Record location: posterior right upper. |  | 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**

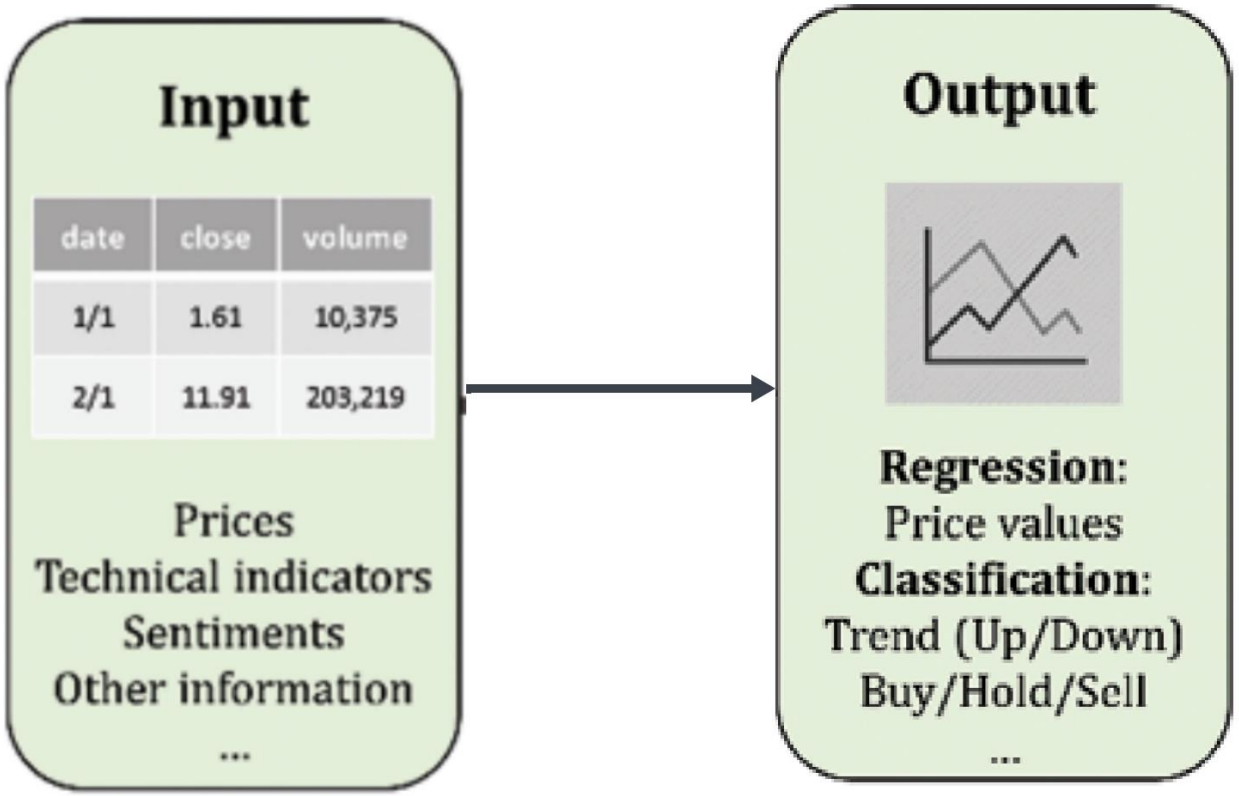



**(a) Concatenation-based fusion method.**

**(b) LLM-based fusion method.**

# 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

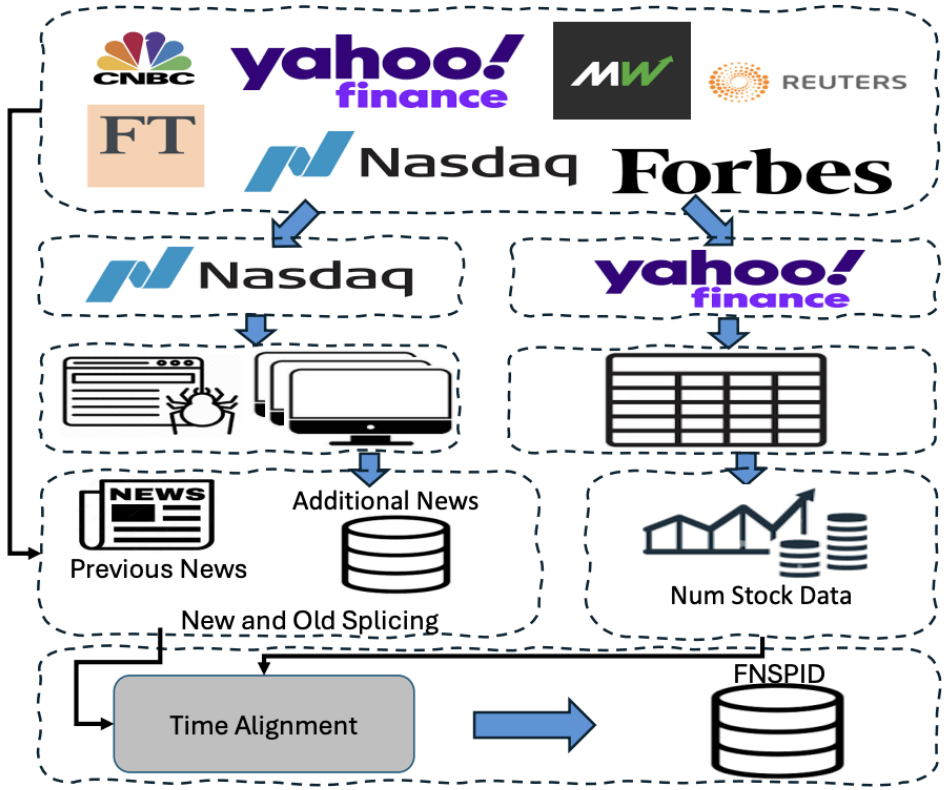



**Figure 1: Data Collection Process from website selection in the first level box; data segmentation in second level boxes; data collection for web scraping on left and numerical data collection on right; data organization on fourth level boxes and final FNSPID build-up on the last level box.**

# 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.**

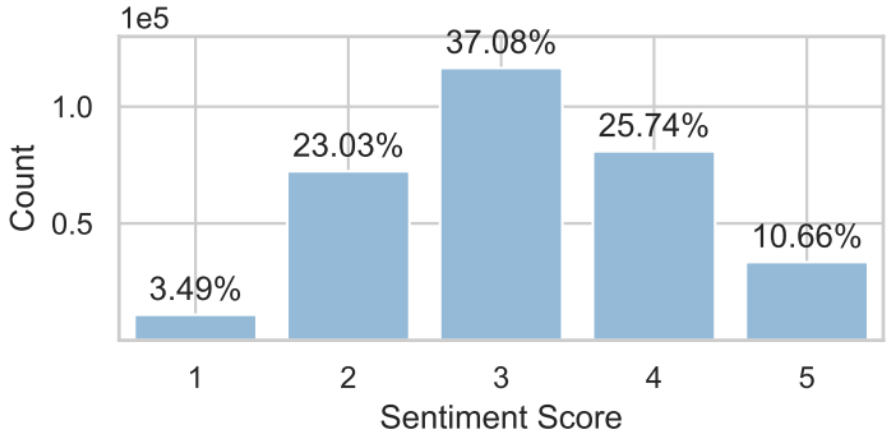



**Figure 4: Sentiment Distribution: 1 is negative, 2 is somewhat negative, 3 is neutral, 4 is somewhat positive, 5 is positive**

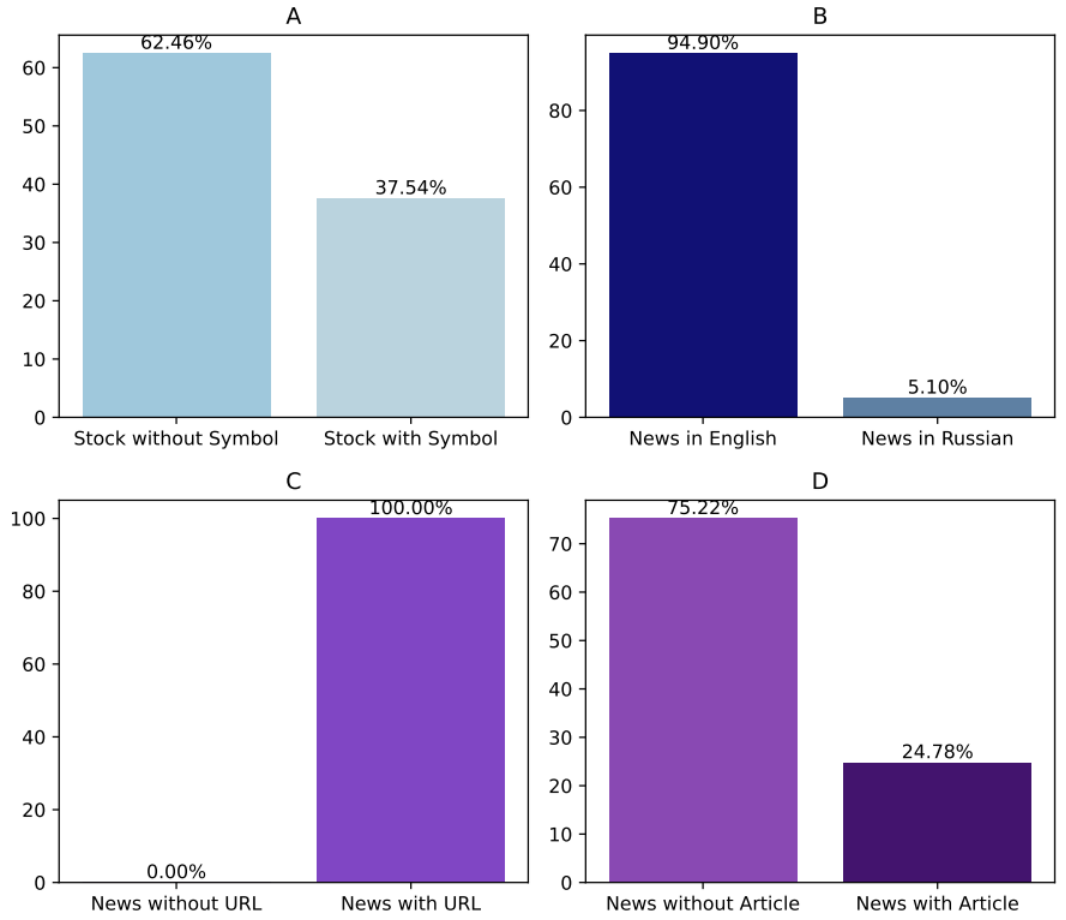



**Figure 5: Statistical Overview: In A, we provide information on news articles that include the stock symbol. The B displays the language distribution, encompassing English and Russian. In C, a comparison of the included URLs is presented. Finally, in the D, details are provided on the news text already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

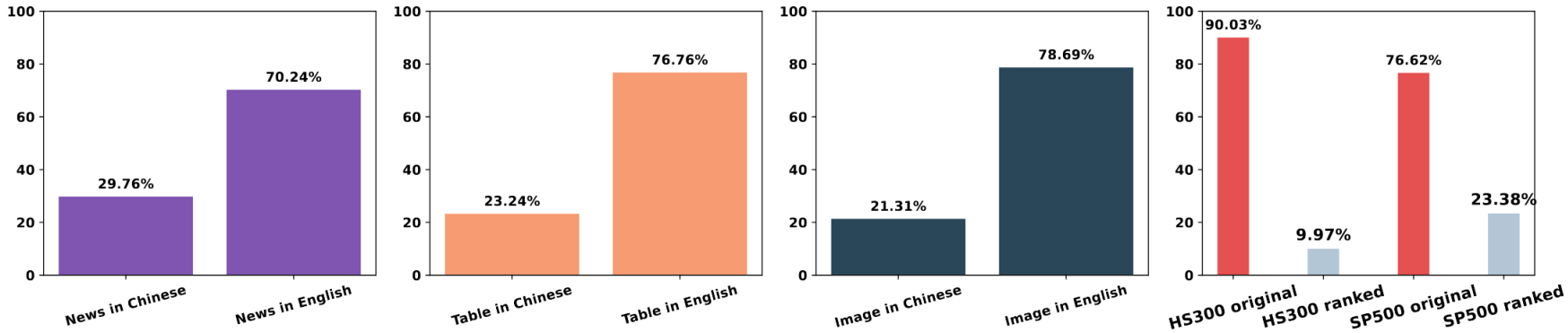
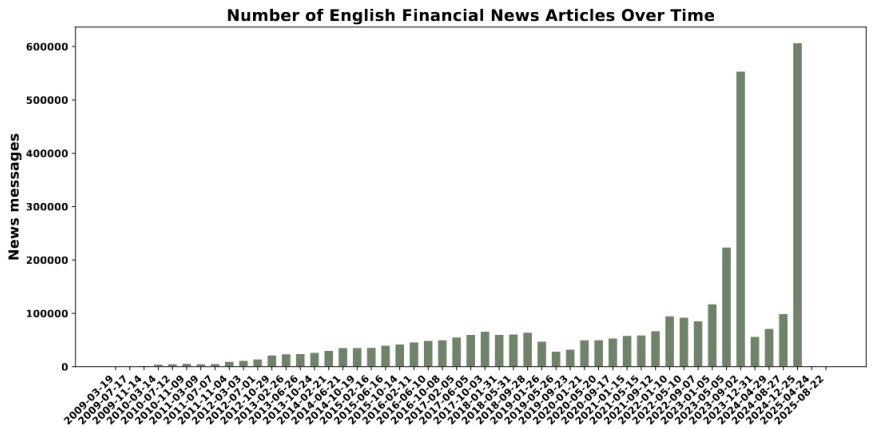
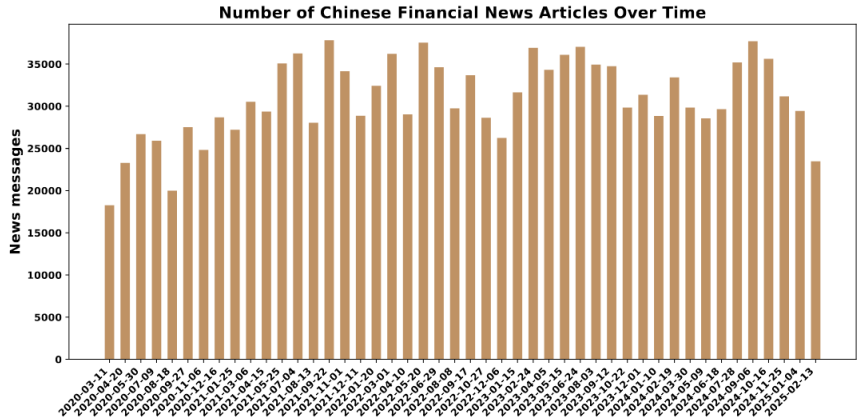



Figure 6: Proportions of Chinese vs. English Modalities (News, Tables, Images) and Coverage Ratios of Ranked vs. Original Daily News for HS300 and S&P 500.



# Finance

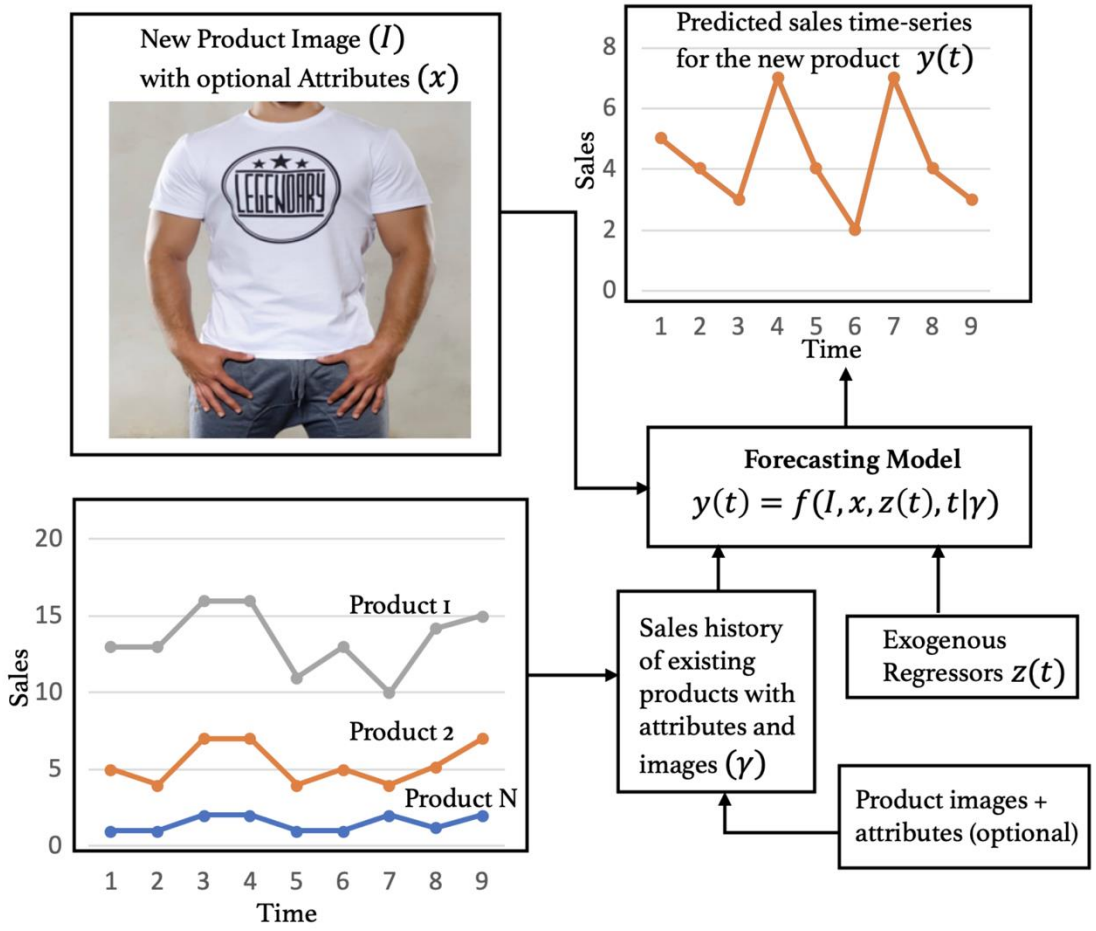
- Combine financial time series and text.

| Type | Content |
|---|---|
| Prompt | data:<br>date,open,high,low,close,adjusted-close,increase-in-5,10,15,20,25,30<br>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<br>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<br>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<br>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<br>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<br>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<br>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<br>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<br>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<br>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<br>tweets:<br>2015-12-23: fxi ishares ftse china 25 index fund ask$fxi $gpro $uco $unh #fxi #finance #stocksgbsn great basin scientific,...<br>2015-12-24: $unh:us looking for value? why it might be time to try unitedhealth group (unh) #unitedhealth group inc #sp500<br>2015-12-25: #dow #stocks $unh unitedhealth group daily:-0.29% weekly:-0.74% ytd:17.46% trend:73 / 100<br>2015-12-26: vz verizon communications company info$vz $aapl $unh $bib #vz #stock #stocks| #unh $unh #spx $spx $dja #dja #stocks #trading #traders #markets #investing #money more ...<br>2015-12-28: buzzers of healthcare sector: unitedhealth group inc $unh merck<br>2015-12-30: uvxy proshares ultra vix short-term futures etf investor opinions $uvxy $dust $ddd $unh #uvxy #pennystocks...<br>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.<br>Start with Rise or Fall, then explain your predictions step by step. |
| Response | Fall. Here are the reasons:<br>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.<br>2. The adjusted-close price is also decreasing consistently throughout this period.<br>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.<br>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.<br>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.




**Ekambaram et al. "Attention based Multi-Modal New Product Sales Time-series Forecasting", KDD 2020**

# 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

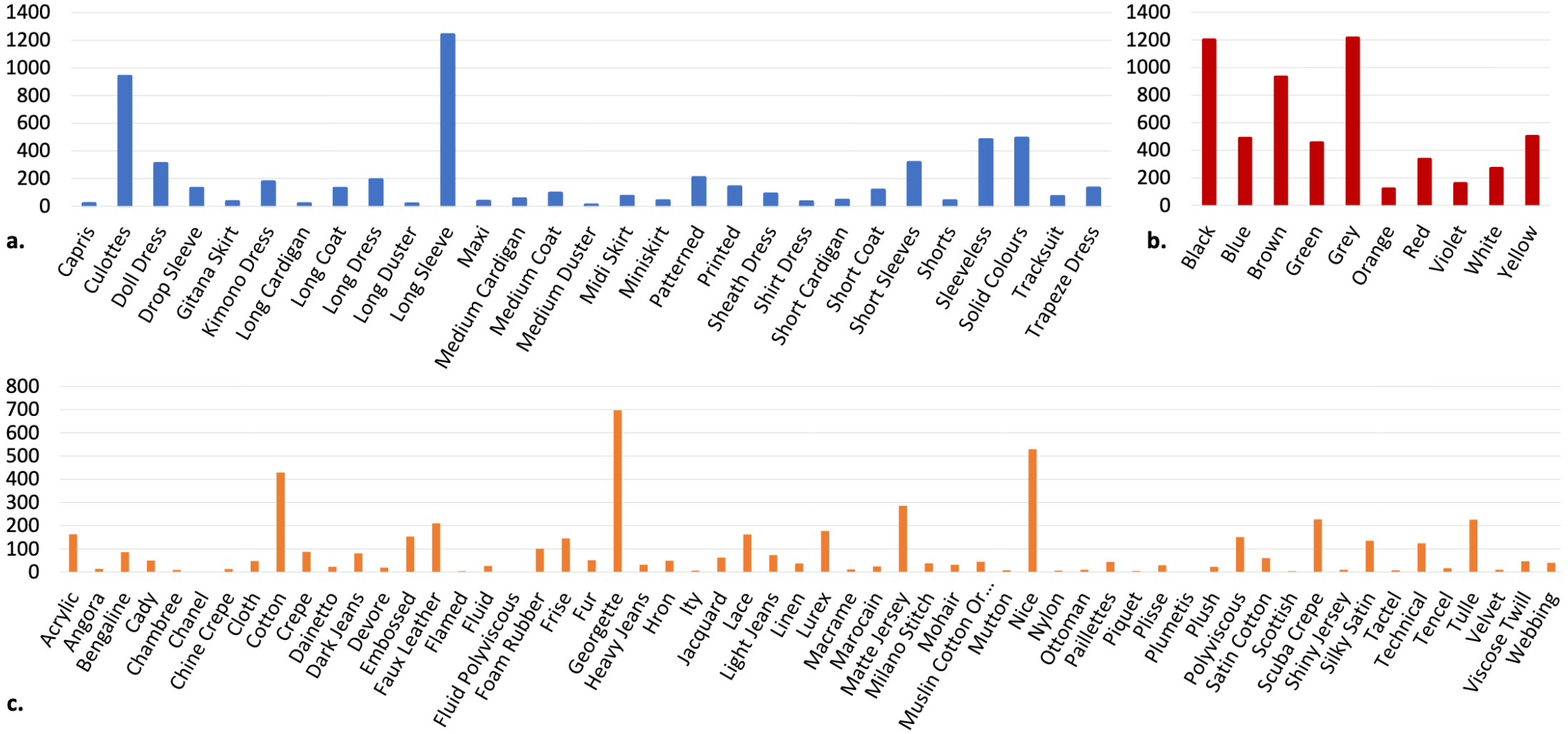
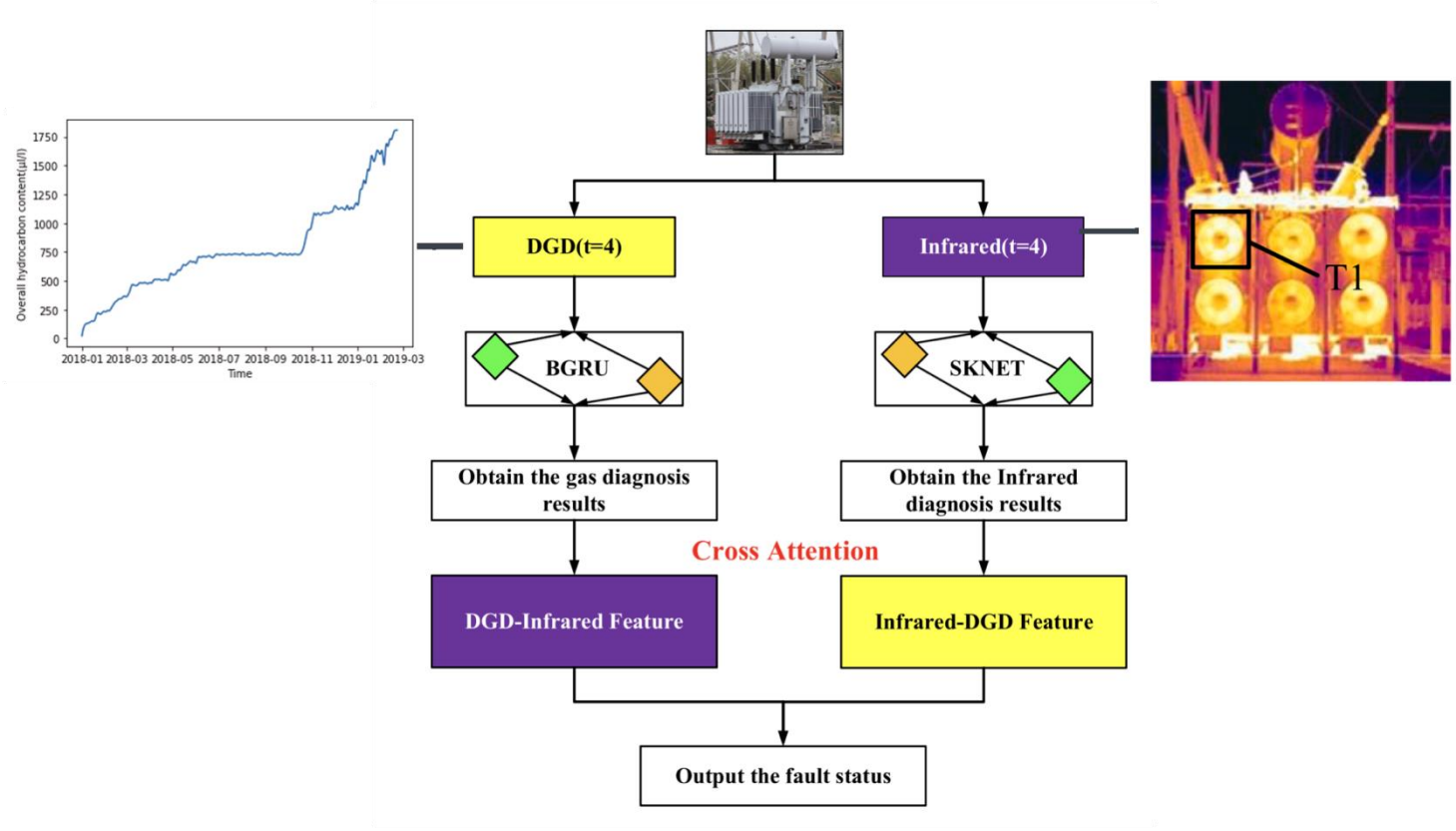



Figure 2: Cardinalities of the dataset for clothing categories (a), color (b) and fabric (c).

# 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**

# 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

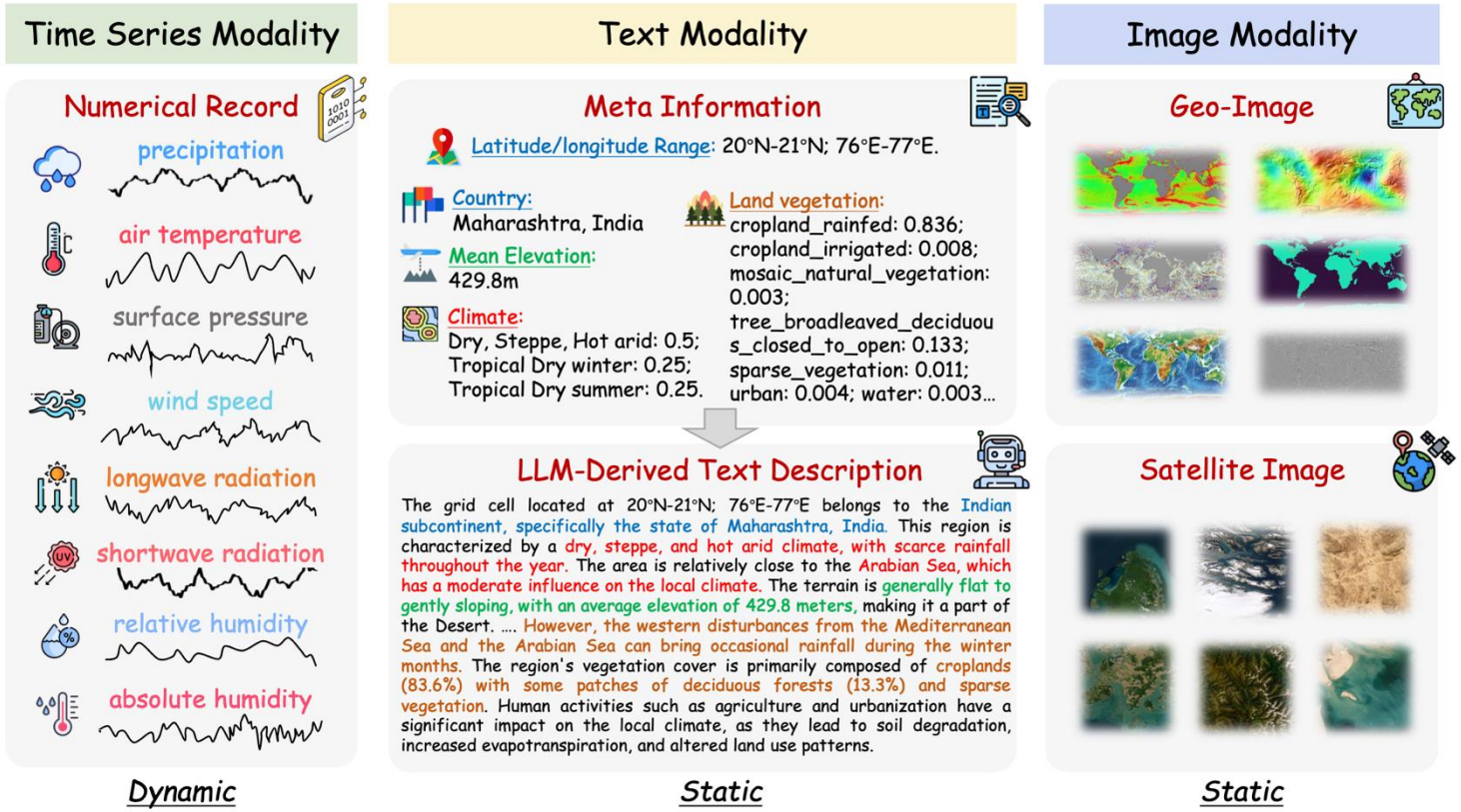



Figure 2: Different modality components of Terra. We provide the data with three temporal scales (3 hourly / daily / monthly), and three spatial scale (0.1° / 0.5° / 1°).

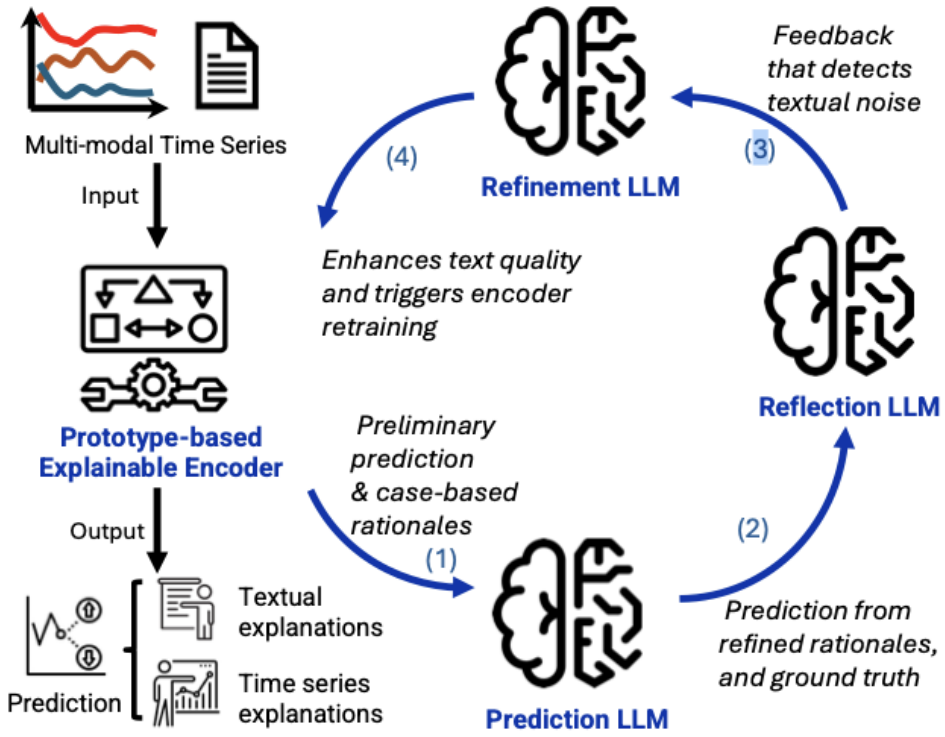
**Chen et al. Terra: A Multimodal Spatio-Temporal Dataset Spanning the Earth, NeurIPS 2024**

# *Future Research Directions*

# Future Research Directions

- **Robustness to imperfect Data:**

Handle missing or noisy real-world context effectively.

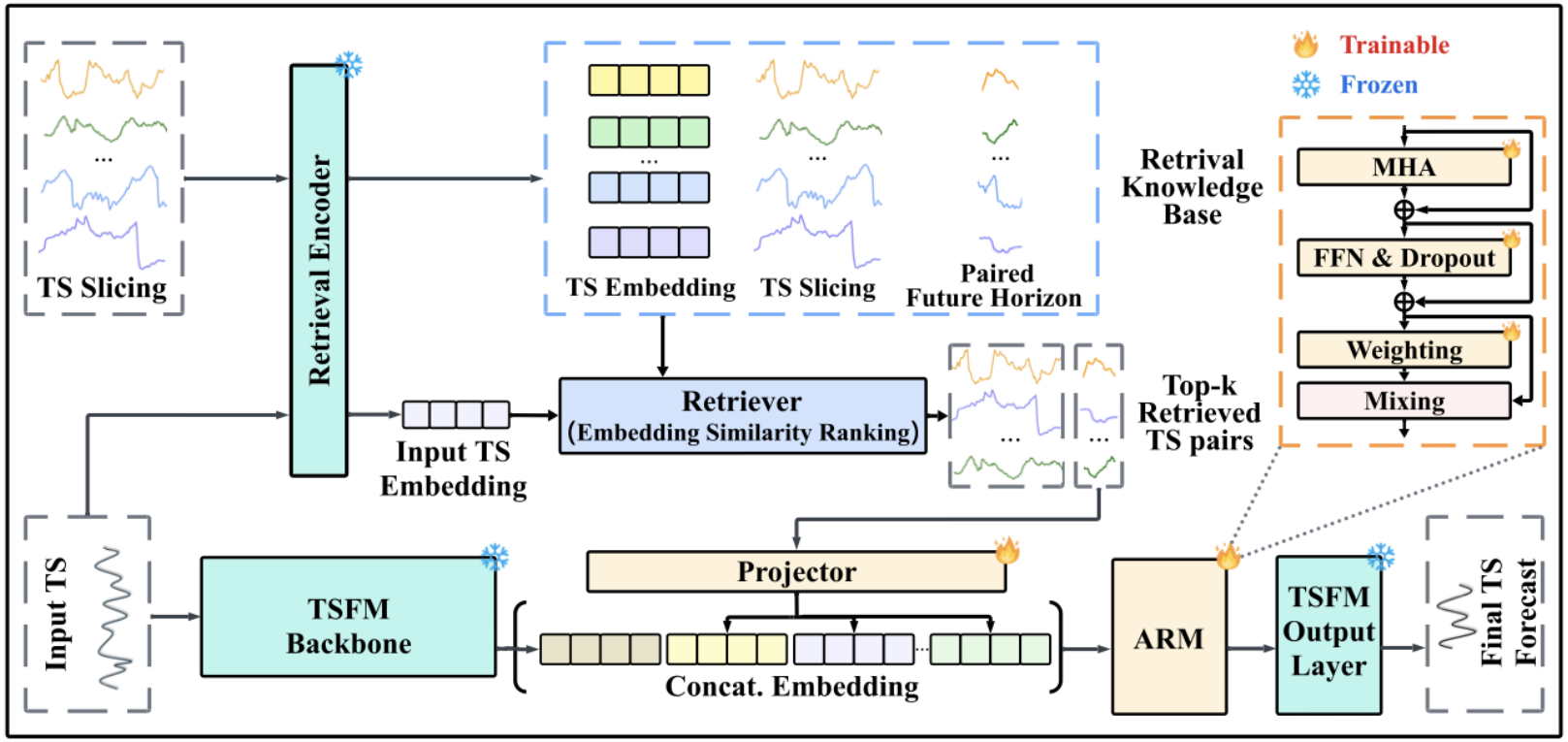



**Jiang et al. "TimeXL: Explainable Multi-modal Time Series Prediction with LLM-in-the-Loop", NeurIPS 2025**

# Future Research Directions

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

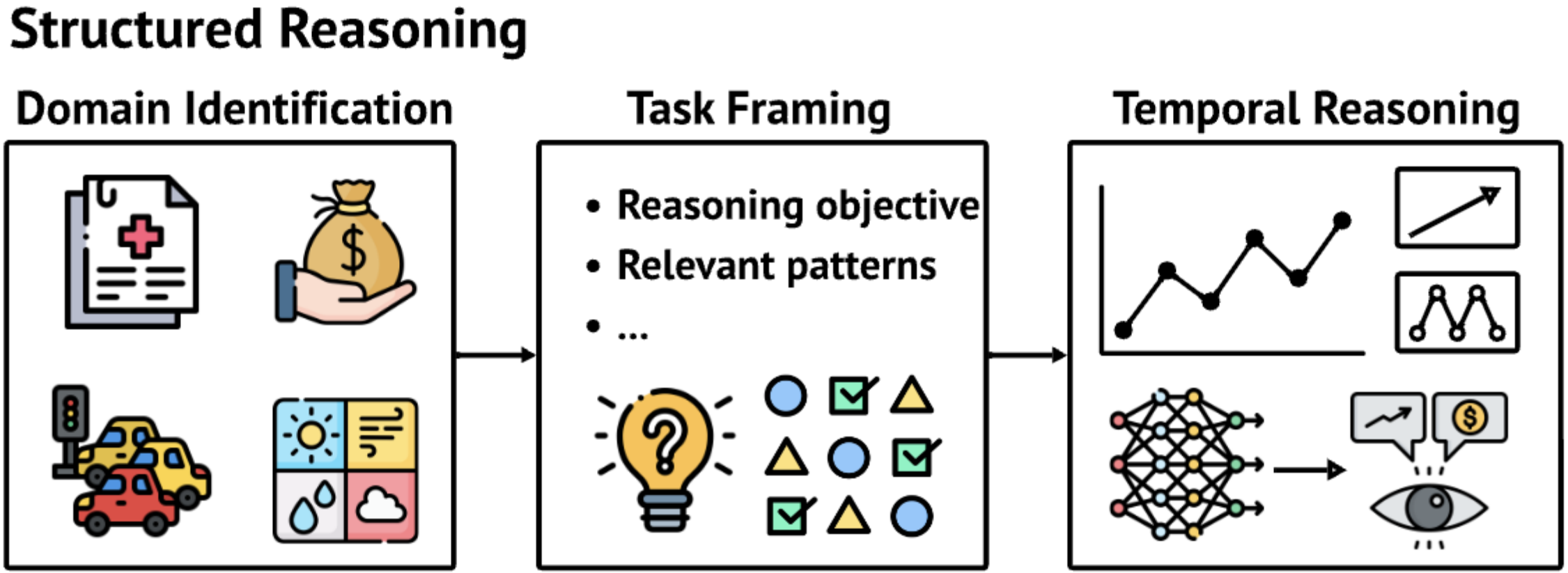
Combine temporal reasoning with context understanding for interpretable inference.




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

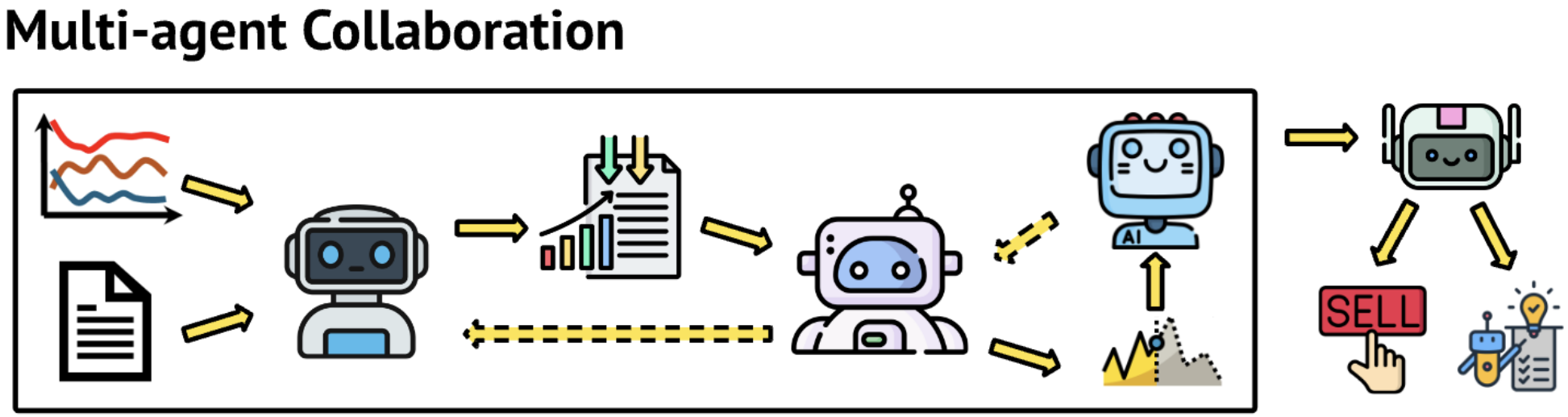
# Future Research Directions

- **Towards structured reasoning with multi-modal data**



# Future Research Directions

- **Multi-agent system for decision making.**



# Future Research Directions

- **Decision-making Systems:**

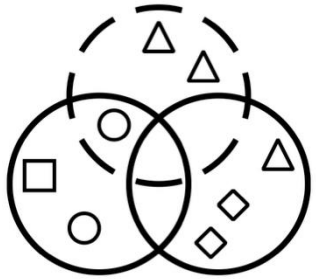
Develop adaptive decision-support systems using multi-modal data to facilitate downstream tasks.

- **Domain Generalization:**

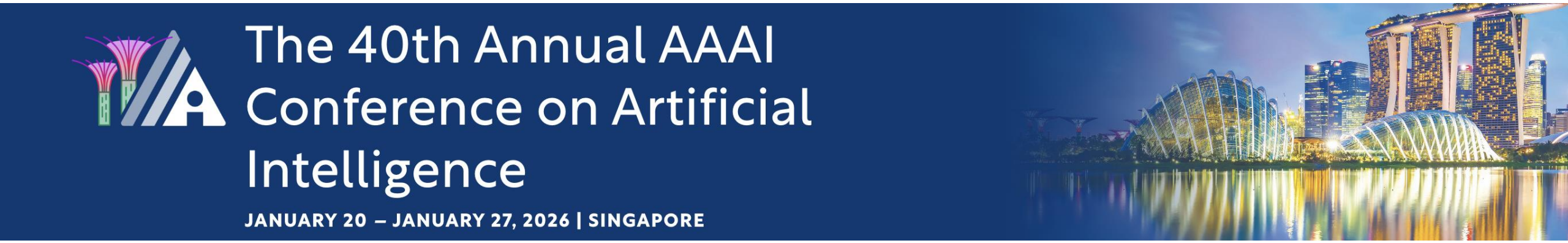
Address the challenges such as domain shifts, modality-specific variations, and temporal dynamics. Improve generalization across unseen domains.

- **Ethics and fairness:**

Address biases to promote equitable outcomes.



...




# *Thank you!*

# *Q & A*

**Survey Paper**

**Github**

HKUST (GZ)

NEC Laboratories America

Morgan Stanley