MM4ST: MM'25 TUTORIAL

MULTIMODAL LEARNING FOR SPATIO-TEMPORAL DATA MINING

- 11:00 AM – 12:30 PM, Monday, October 27th
- Swift 1 & Swift 2, Radisson
- Dublin Ireland

Organizers

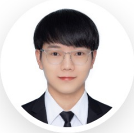
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

YUXUAN LIANG

Assistant Professor
Hong Kong University of Science and Technology (Guangzhou)

HAO MIAO

Research Assistant Professor
Hong Kong Polytechnic University



SIRU ZHONG

PhD Student
Hong Kong University of Science and Technology (Guangzhou)

XIXUAN HAO

PhD Student
Hong Kong University of Science and Technology (Guangzhou)

YAN ZHAO

Professor
University of Electronic Science and Technology of China

QINGSONG WEN

Head of AI Research & Chief Scientist
Squirrel Ai Learning

ROGER ZIMMERMANN

Professor
National University of Singapore

NUS National University of Singapore

Squirrel Ai Learning

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

UESTC 1956 电子科技大学 University of Electronic Science and Technology of China

THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學

Outline

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

1 **Background & Examples**

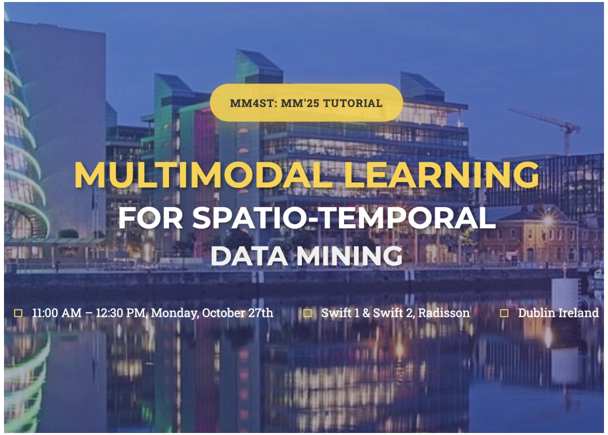
2 Foundation of ST Data

3 Why Multimodal ST Data Fusion

4 Principle of ST Multimodal Fusion

5 Visual/Language Knowledge Transfer

6 Conclusions



ACM multimedia

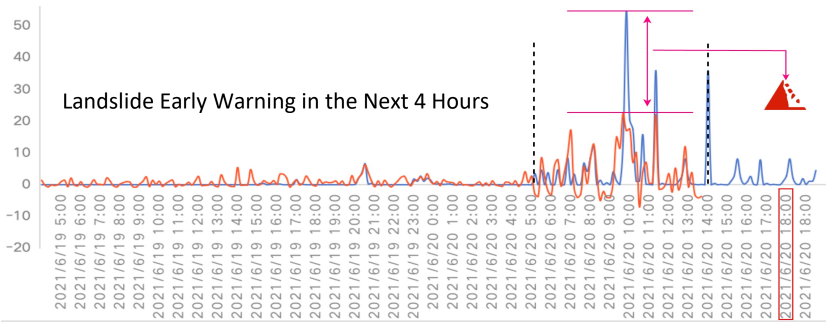
Dublin, Ireland 27-31.10.2025

Big Challenges in Big Cities

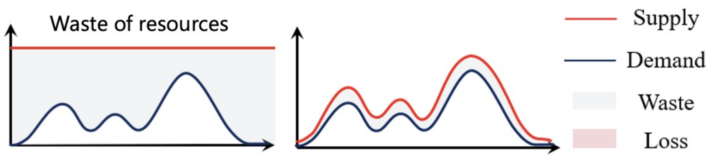
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Examples (Data / Analysis / Decision)

- **Geological disaster response**
  1. Current and historical rainfall
  2. Predicting whether geological disaster will happen in the future
  3. When to take what precautionary measures
- **Autoscaling of cloud resources**
  1. Current and historical user load
  2. Predicting future user load
  3. Decision on whether to scale up or scale down at future time, and by how much.



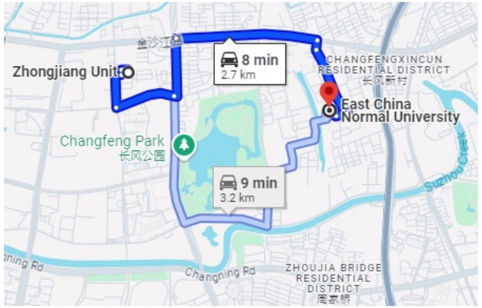
Time series: Rainfall



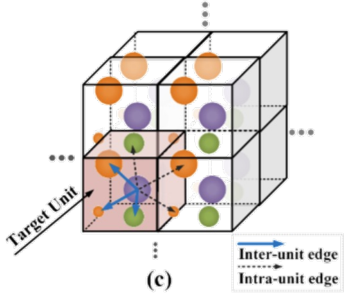
Time series: User Load

Examples (Data / Analysis / Decision)

- **Navigation from a source to a destination**
  - Current and historical traffic flow
  - Predicting traffic flows, identifying congested areas
  - Selecting fastest/greenest paths; whether using highways, bridges vs. tunnels
- **Design new materials/drugs**
  - Structure of existing materials and their properties
  - Predicting properties of unknown material structures
  - Potential candidates for new materials that satisfy specific properties



Spatio-temporal Data: Trajectories, Traffic Flows



Molecular structure spatial data: geometric coordinates and topological graphs

Outline

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

1 Background & Examples

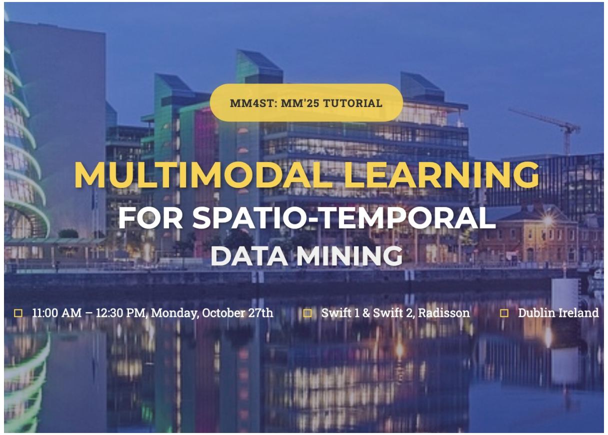
2 Foundation of ST Data

3 Why Multimodal ST Data Fusion

4 Principle of ST Multimodal Fusion

5 Visual/Language Knowledge Transfer

6 Conclusions

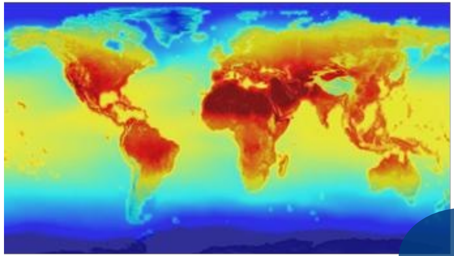


ACM multimedia

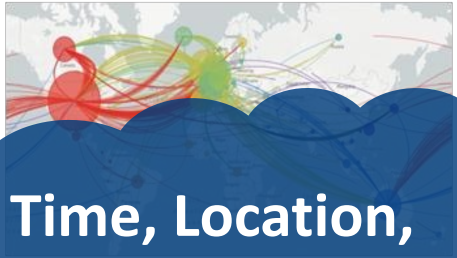
Dublin, Ireland 27-31.10.2025

What is Spatio-Temporal (ST) Data

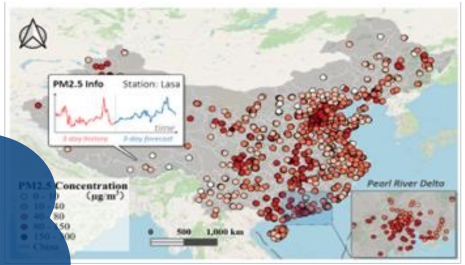
- ST data refers to data that integrates spatial (location), temporal (time), and event-related information, capturing how phenomena change across both space and time.



Climate



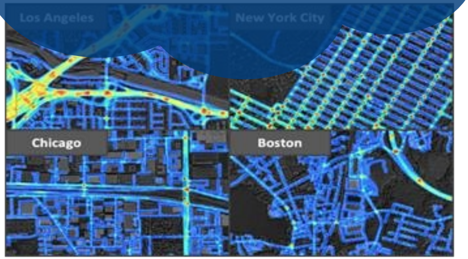
Epidemiology



Environment

Time, Location, Event

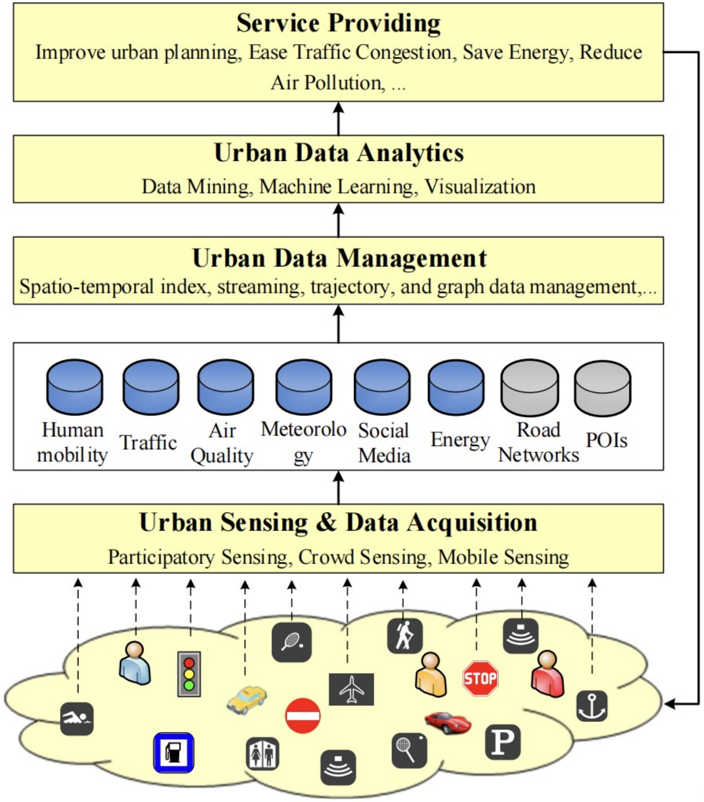
Social Science

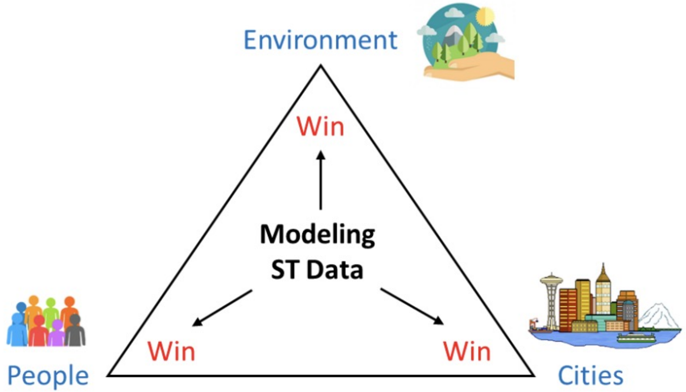


Transportation

Sports Analysis

ST Data Intelligence Framework





Tackle the Big challenges

in Big cities

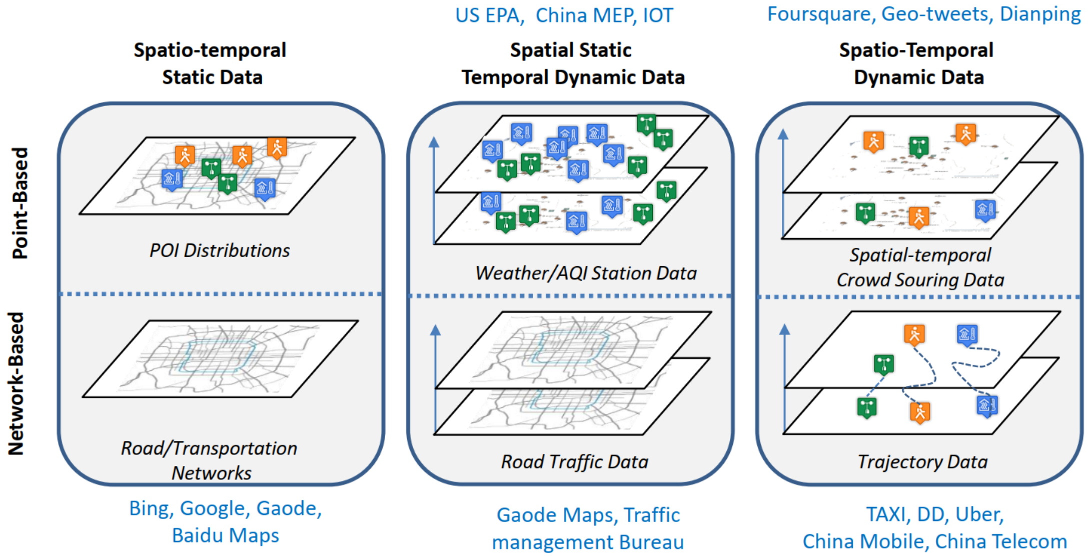
using Big data!

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

ST Data - Taxonomy

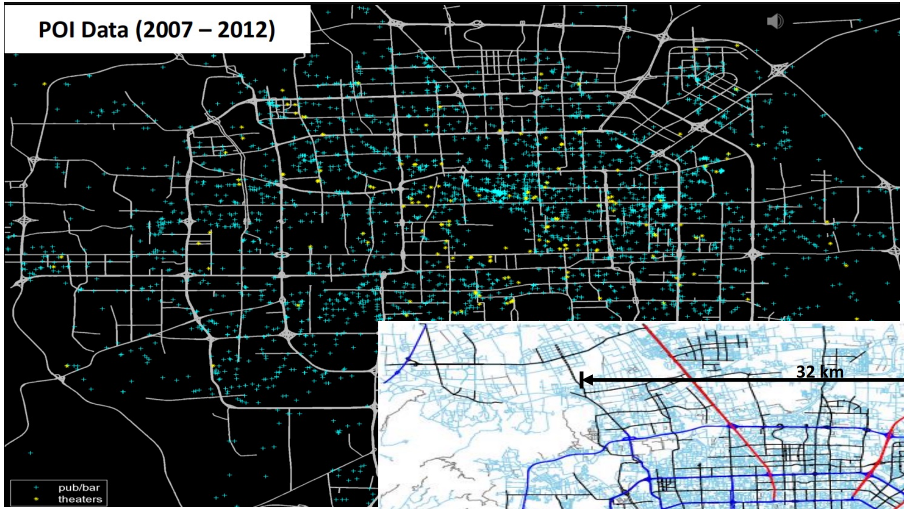
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

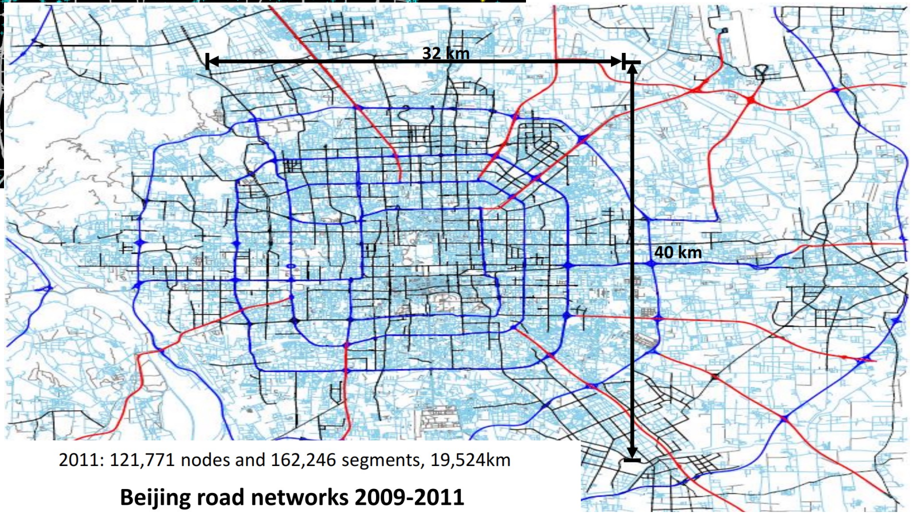
- Spatially and temporally static data
- Spatially static and temporally dynamic data
- Spatially and temporally dynamic data



Spatially and Temporally Static Data

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



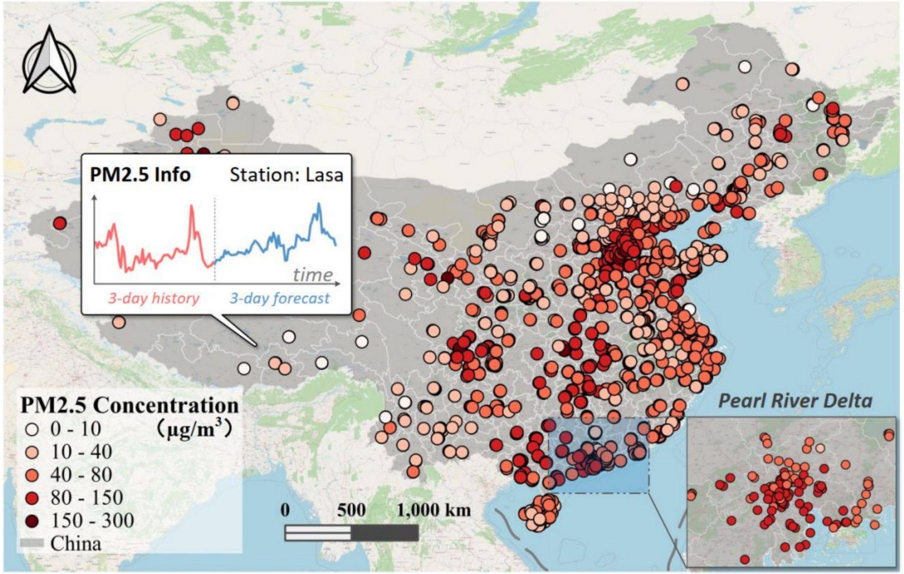


2011: 121,771 nodes and 162,246 segments, 19,524km

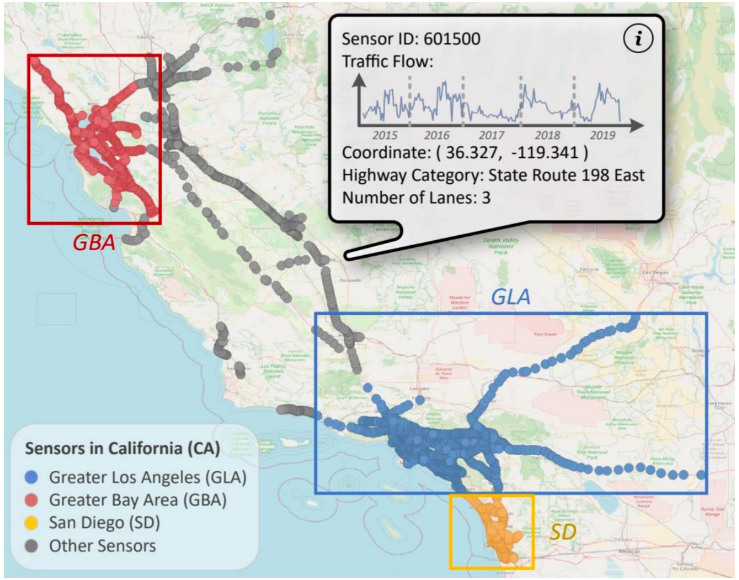
Beijing road networks 2009-2011

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 Concentration



Traffic flow

Spatially and Temporally Dynamic Data

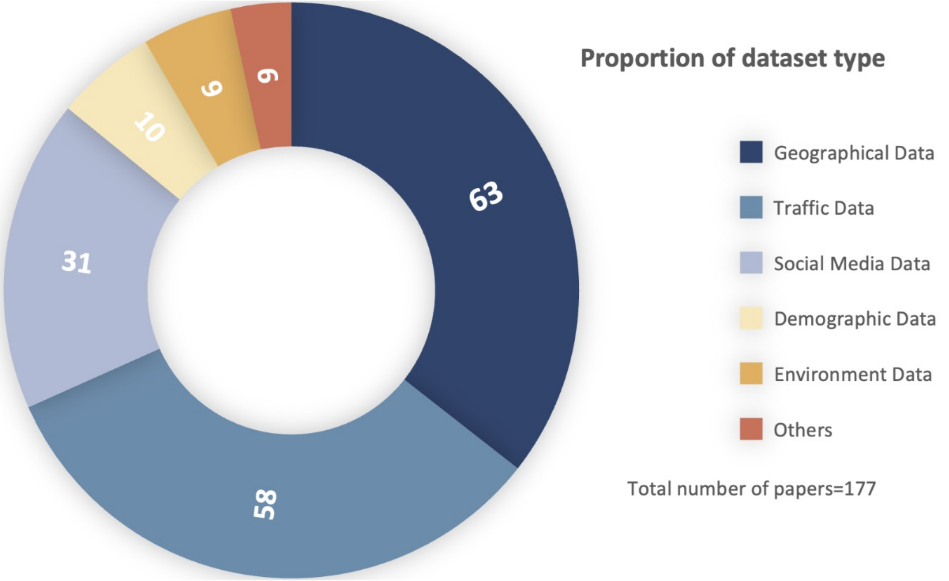
- Spatial and temporal values varying in time
  - Moving objects
  - Trajectories

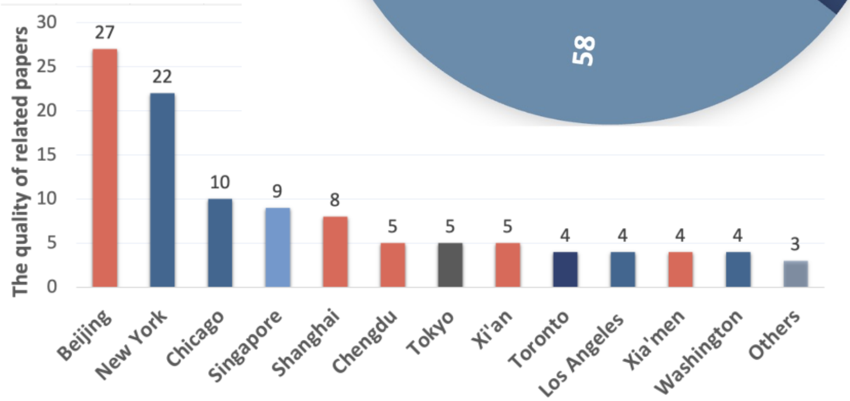
$$T = p_1 \rightarrow p_2 \rightarrow \cdots \rightarrow p_n, \quad p_i = (\underbrace{a_i, b_i}_{\text{Location (latitude \& longitude)}}, \overset{\text{Timestamp}}{\boxed{t_i}})$$

- E.g. Human mobility (travel logs, check-ins, credit card transactions, trajectories of taxis / airplanes / ferries, ...), Animals migration, Natural phenomena.

Data Types and Data Sources

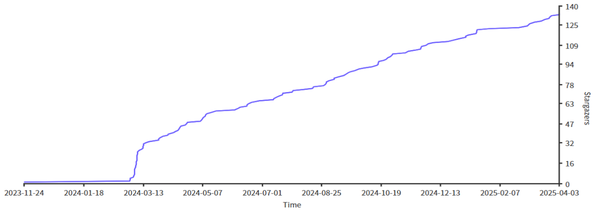
- Geographical data
- Traffic data
- Social media data
- Demographic data
- Environment data
- Others

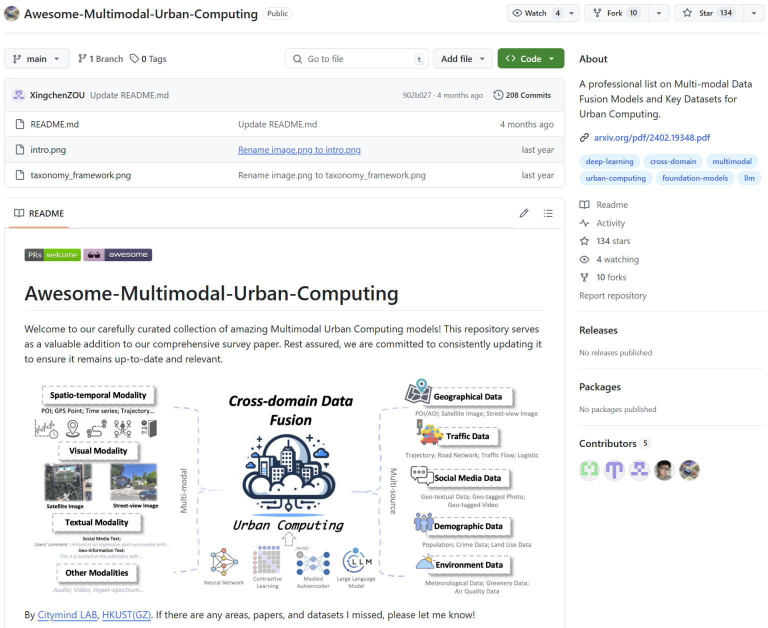




Data Types and Data Sources

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

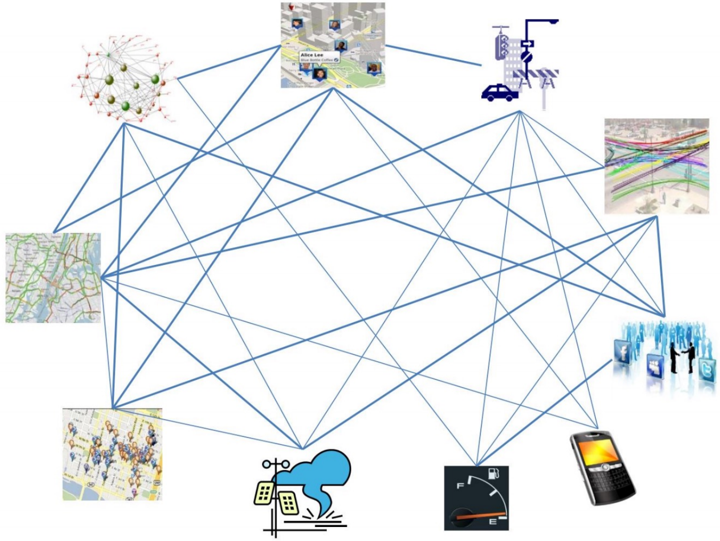


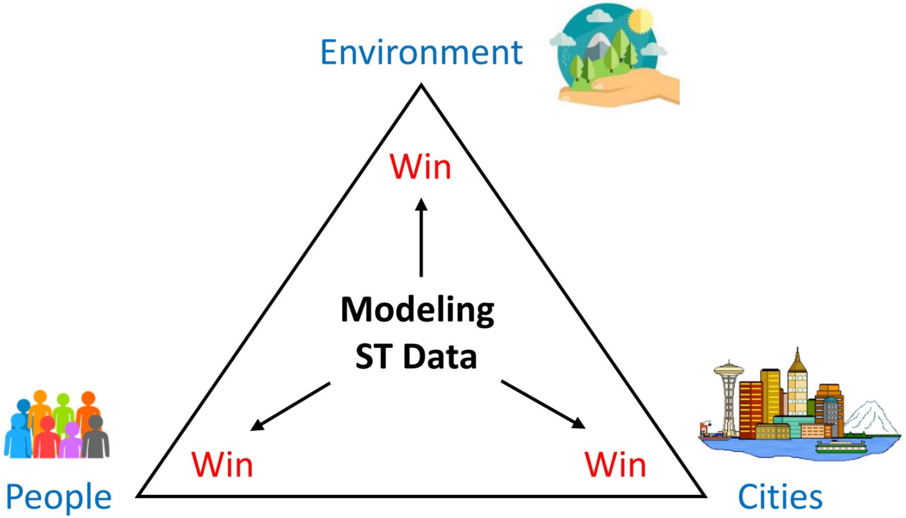


Category	Content	Format	Dataset	Link	Reference
Geographical Data	Satellite Image	Image	ArcGIS	https://developers.arcgis.com	[186]
			PlanetScope	https://developers.planet.com/docs/data/planetscope/	[154]
			Google Earth	https://developers.google.com/maps/documentation/	[116]
			OpenStreetMap	https://www.openstreetmap.org/	[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]

Why ST Data Mining

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





Outline

1 Background & Examples

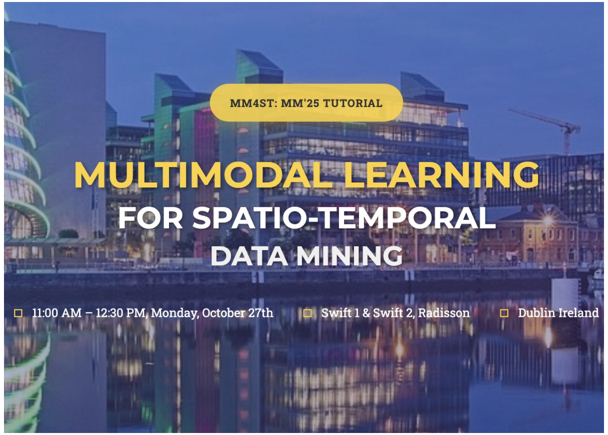
2 Foundation of ST Data

3 **Why Multimodal ST Data Fusion**

4 Principle of ST Multimodal Fusion

5 Visual/Language Knowledge Transfer

6 Conclusions



ACM multimedia

Dublin, Ireland 27-31.10.2025

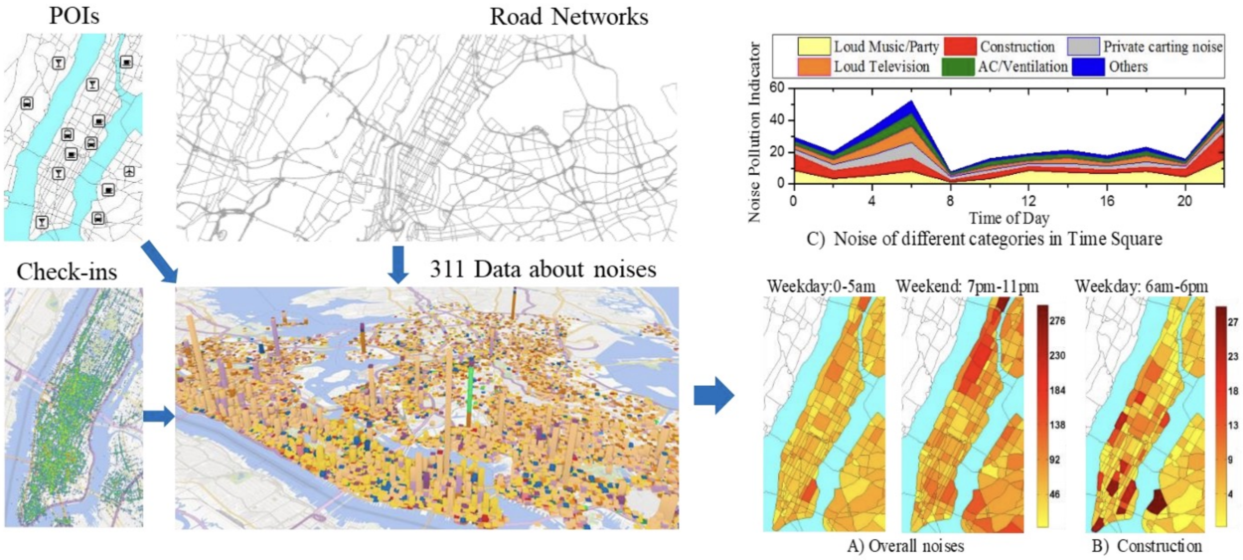
Why Multimodal ST Data Mining

- Single-modality information fails to address the complexity of real-world scenarios.
- Example 1: Air Quality Inference — Unlock the power from multiple (**sparse**) data across **different domains**

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

Why Multimodal ST Data Mining

- Single-modality information fails to address the complexity of real-world scenarios.
- <u>Example</u> 2: Noise Diagnosis — Unlock the power from multiple (**sparse**) data across **different domains**

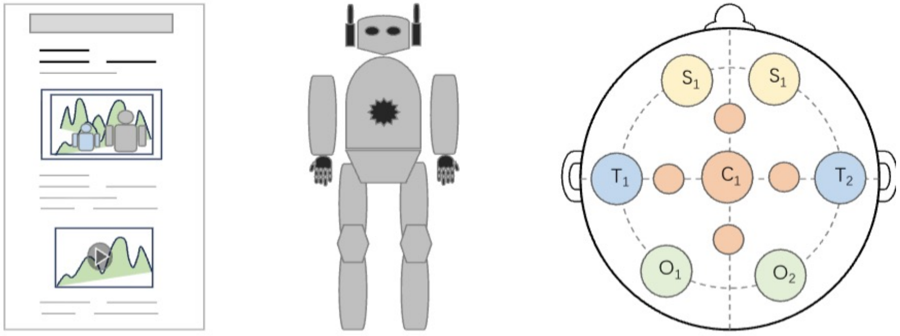


Diagnosing New York City's Noises with Ubiquitous Data, Ubicomp 2014

Research Gap

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Multimodal ST Data Mining vs Traditional Multimodal Learning
- Multimodal ST Data Mining ⇔ 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.

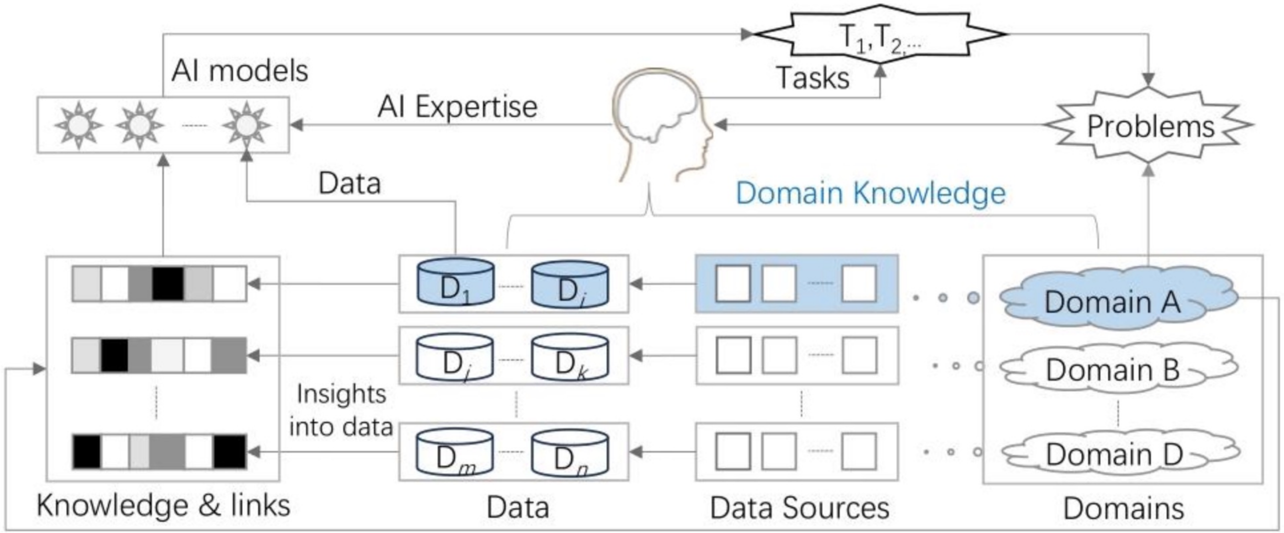


A) A Webpage B) A robot C) Sensors for Brains

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

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)

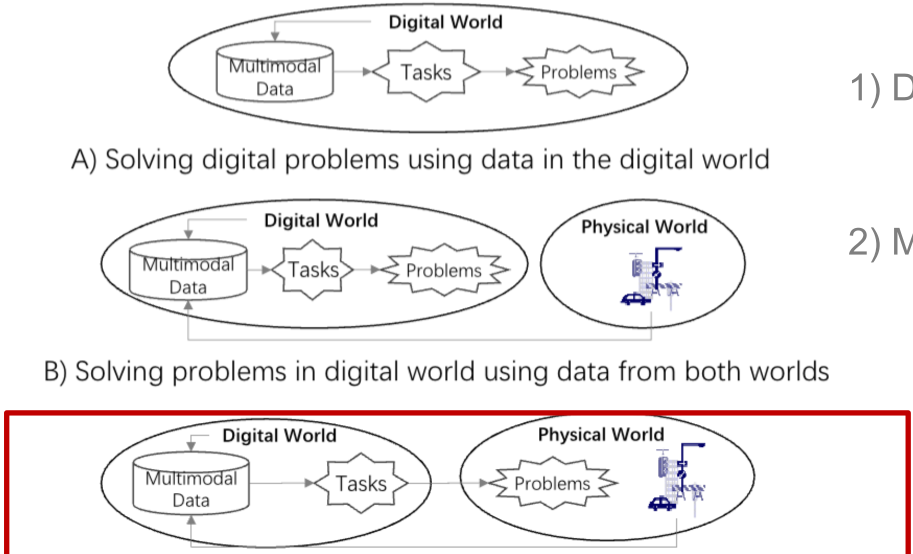


Fusing Cross-Domain Knowledge from Multimodal Data to Solve Problems in the Physical World. In ACM Transactions on Intelligent Systems and Technology, 2025.

ST Multimodal Learning is Future

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- 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

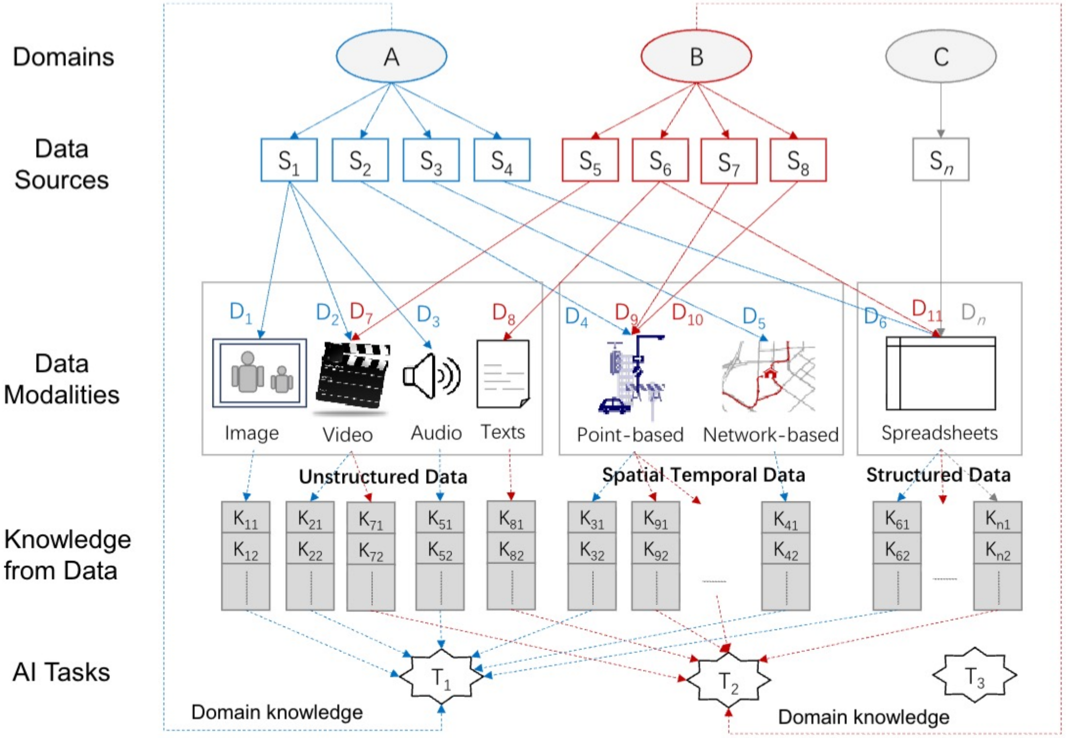
3) Real World Problems, e.g. AQI

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

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 Data Fusion System

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

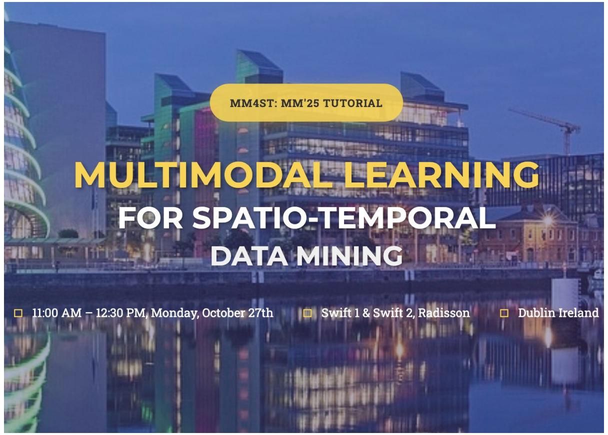


Fusing Cross-Domain Knowledge from Multimodal Data to Solve Problems in the Physical World. In ACM Transactions on Intelligent Systems and Technology, 2025.

Outline

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

1. Background & Examples
2. Foundation of ST Data
3. Why Multimodal ST Data Fusion
4. **Principle of ST Multimodal Fusion**
5. Visual/Language Knowledge Transfer
6. Conclusions



Machine Learning Era

Methodologies for Cross-Domain Data Fusion: An Overview

Publisher: IEEE Cite This PDF

Yu Zheng **All Authors**

500+ citations on Google Scholar!

Abstract

Document Sections

1 Introduction

2 Related Work

3 Stage-Based Data Fusion Methods

4 Feature-Level-Based Data Fusion

5 Semantic Meaning-Based Data Fusion

Show Full Outline ▾

Authors

Abstract:
Traditional data mining usually deals with data from a single domain. In the big data era, we face a diversity of datasets from different sources in different domains. These datasets consist of multiple modalities, each of which has a different representation, distribution, scale, and density. How to unlock the power of knowledge from multiple disparate (but potentially connected) datasets is paramount in big data research, essentially distinguishing big data from traditional data mining tasks. This calls for advanced techniques that can fuse knowledge from various datasets organically in a machine learning and data mining task. This paper summarizes the data fusion methodologies, classifying them into three categories: stage-based, feature level-based, and semantic meaning-based data fusion methods. The last category of data fusion methods is further divided into four groups: multi-view learning-based, similarity-based, probabilistic dependency-based, and transfer learning-based methods. These methods focus on knowledge fusion rather than schema mapping and data merging, significantly distinguishing between cross-domain data fusion and traditional data fusion studied in the database community. This paper does not only introduce high-level principles of each category of methods, but also give examples in which these techniques are used to handle real big data problems. In addition, this paper positions existing works in a framework, exploring the relationship and difference between different data fusion methods. This paper will help a wide range of communities find a solution for data fusion in big data projects.

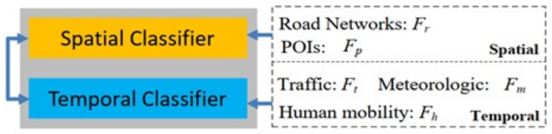
Published in: IEEE Transactions on Big Data (Volume: 1 , Issue: 1, 01 March 2015)

Methodologies for Cross-Domain Data Fusion: An Overview, IEEE Transaction on Big Data, 2015

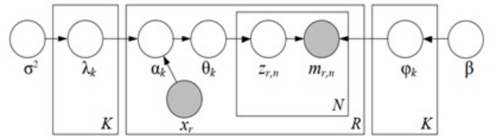
Principle of ST Multimodal Fusion

Machine Learning Era

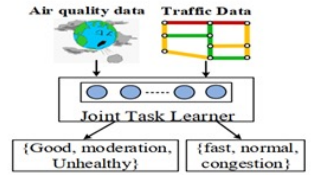
- **Stage-based data fusion**
- **Feature-level-based data fusion**
  - Feature concatenation + regularization
  - DNN-based
- **Semantic meaning-based fusion**
  - Multiple-view-based: like co-training
  - Similarity-based: Coupled matrix factorization
  - PGM-based
  - Transfer learning-based



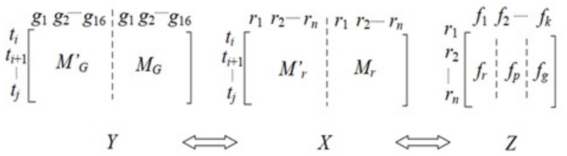
Multi-view learning (Co-training)



Pro. dependency-based (Topic Models)



Transfer Learning-based

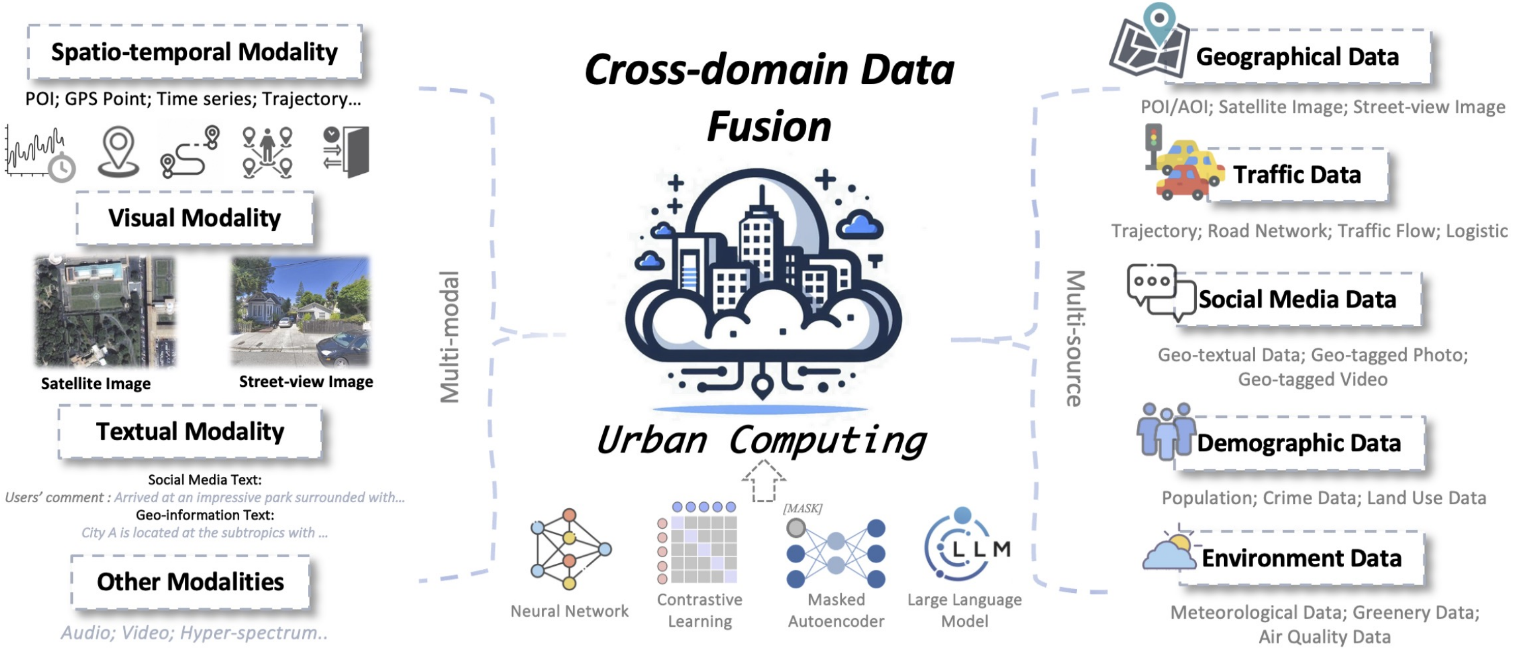


Similarity-based (matrix factorization)

Principle of ST Multimodal Fusion

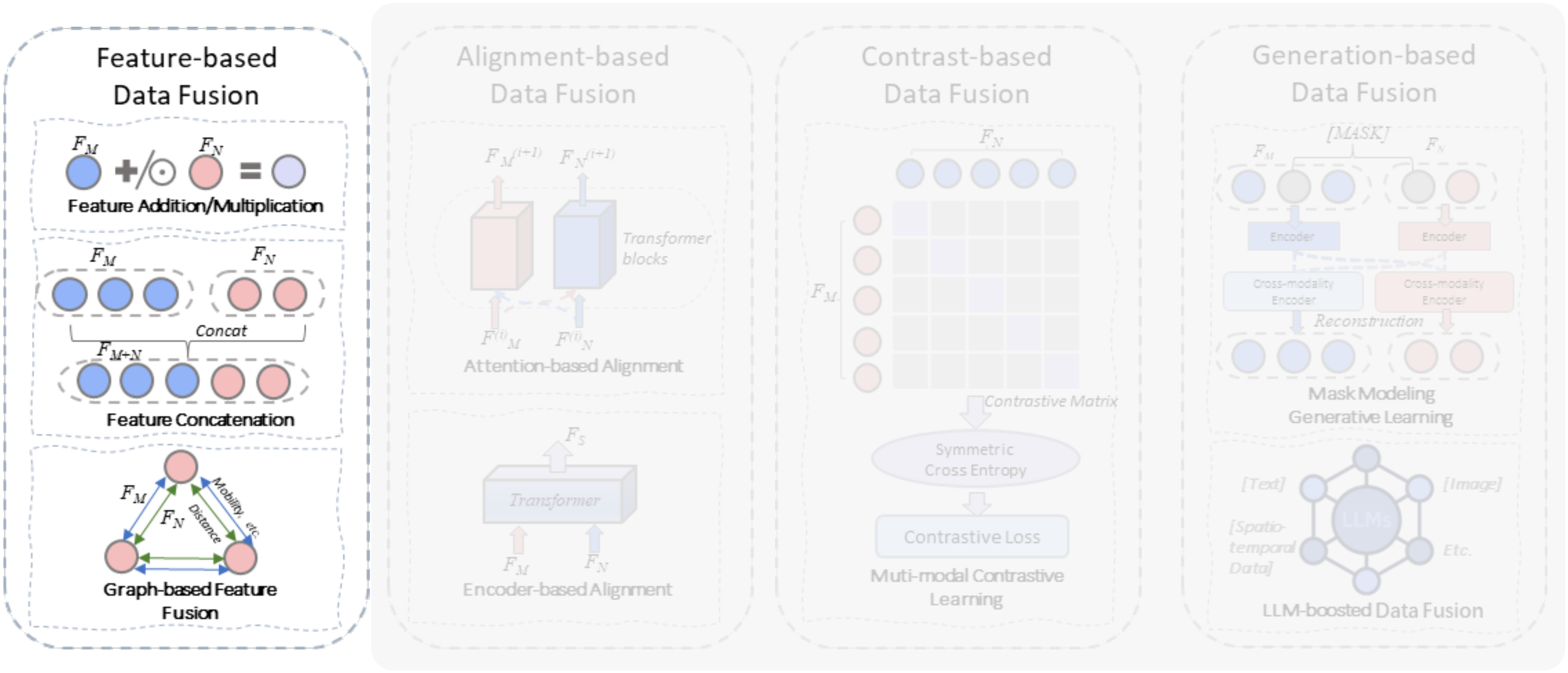
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Deep Learning Era



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

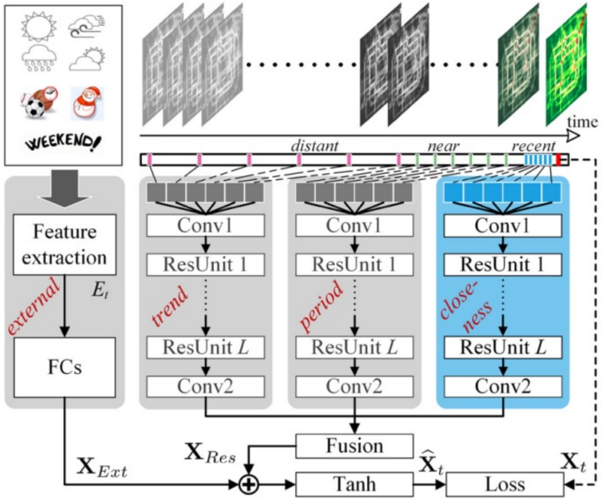
Deep Learning-based Fusion Methods



Feature-based Fusion (Simplest!)

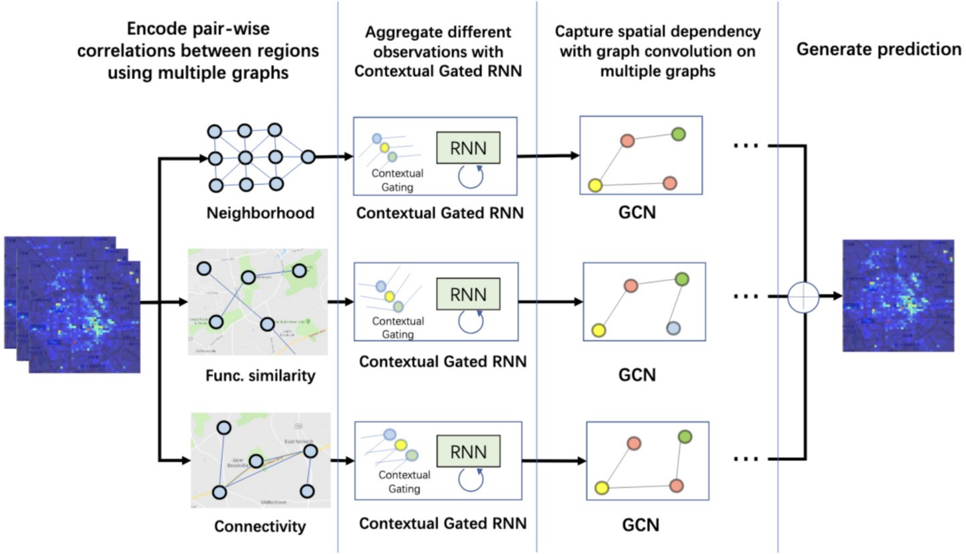
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Feature Addition/Multiplication
- Feature Concatenation



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

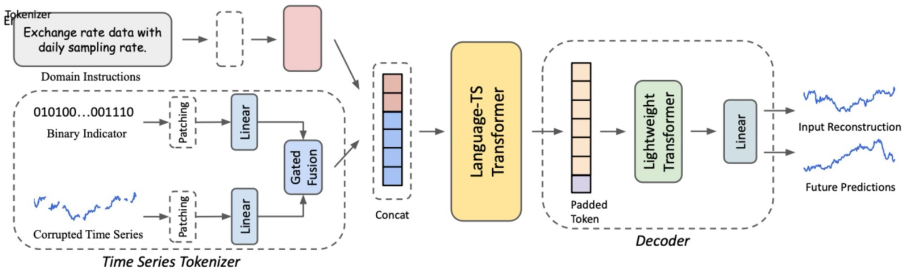
- Graph-based Data Fusion



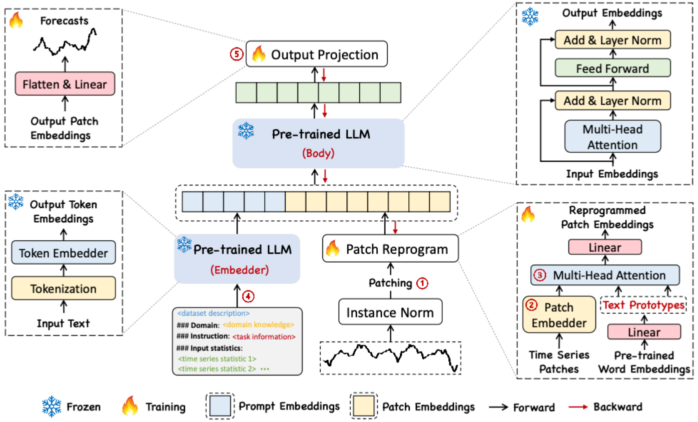
Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting, AAAI 2019

Feature-based Fusion (Simplest!)

- **UniTime** (Unified TS Modeling): Directly concatenates TS patch tokens with Text tokens and feeds them into the LLM.
- **Time-LLM** (end-to-end LLM4TS): Converts TS patch tokens into Text tokens, combines them with prompt embeddings, and feeds them into the LLM.



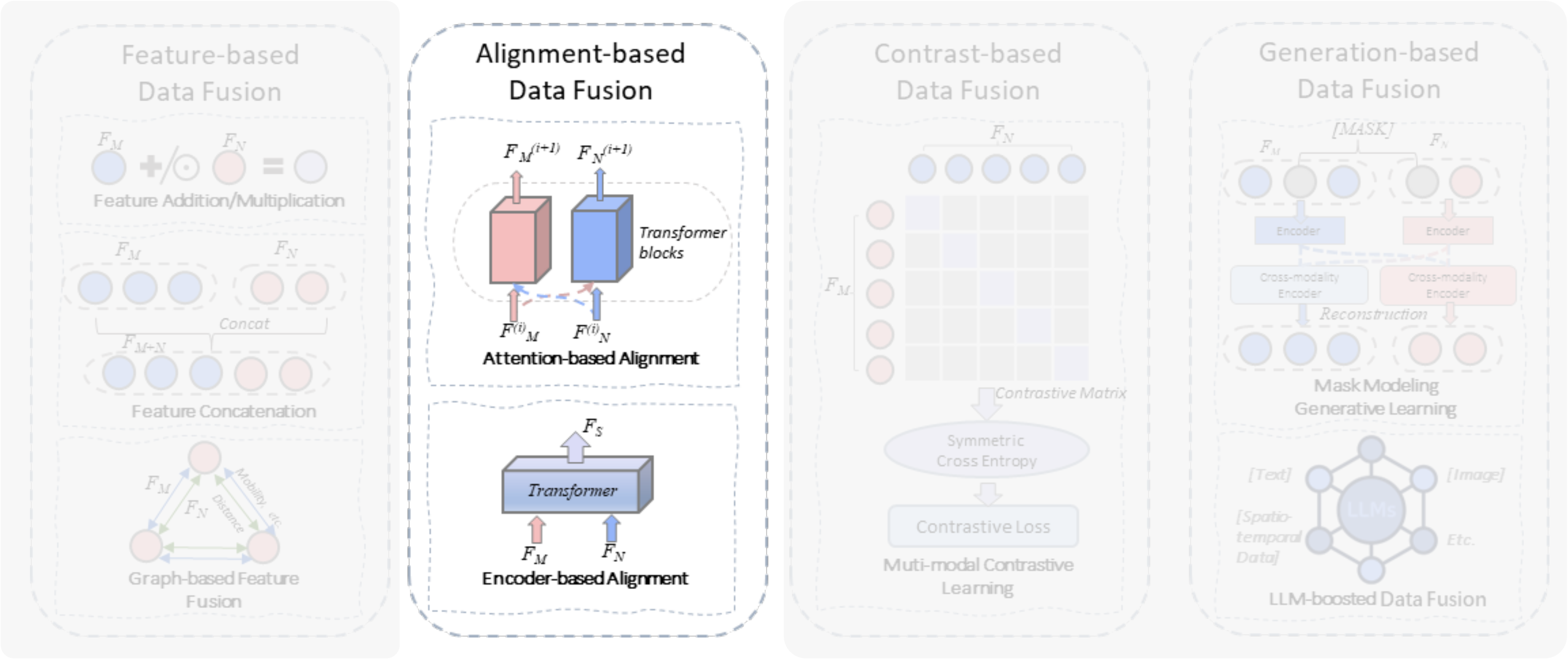
UniTime: A Lanquage-Empowered Unified Model for Cross-Domain Time Series Forecasting. WWW 2024



Time Series Forecasting by Reprogramming Large Language Models. ICLR 2024

Deep Learning-based Fusion Methods

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

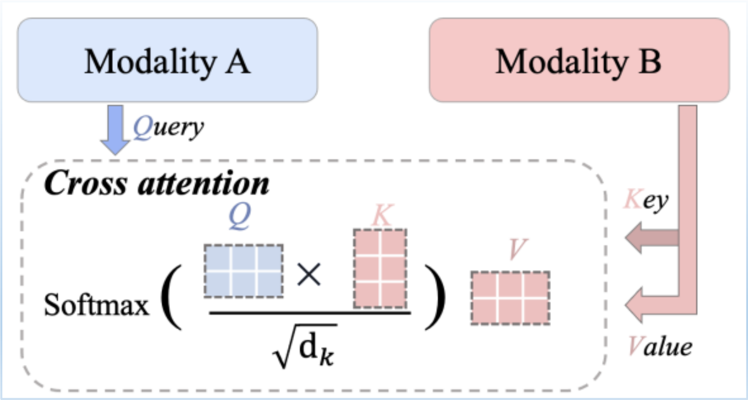


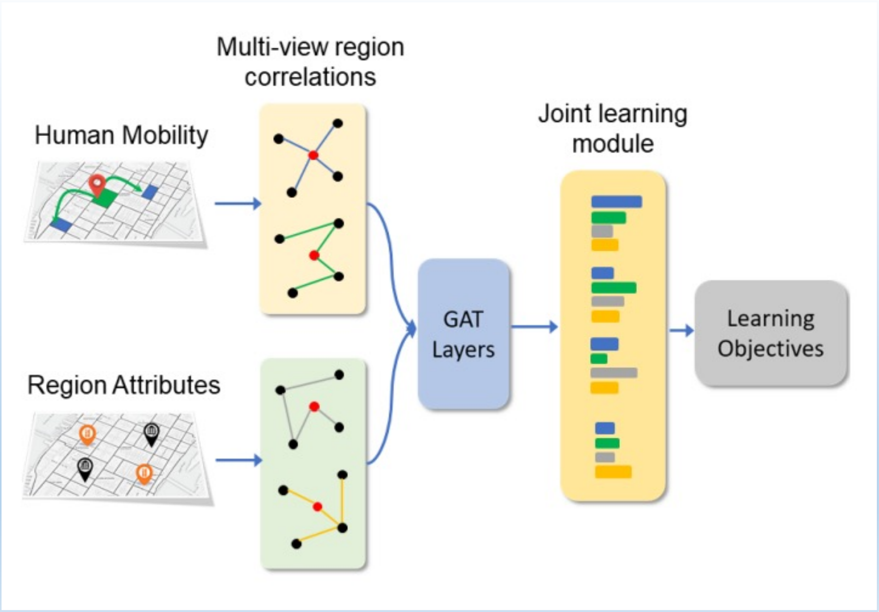
Deep Learning for Cross-Domain Data Fusion in Urban Computina:Taxonomy, Advances, and Outlook, Information Fusion, 2024.

Alignment-based Fusion

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

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

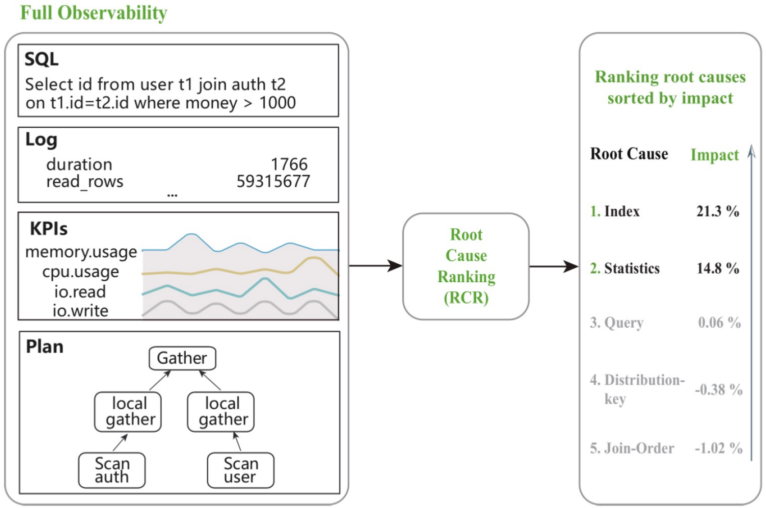


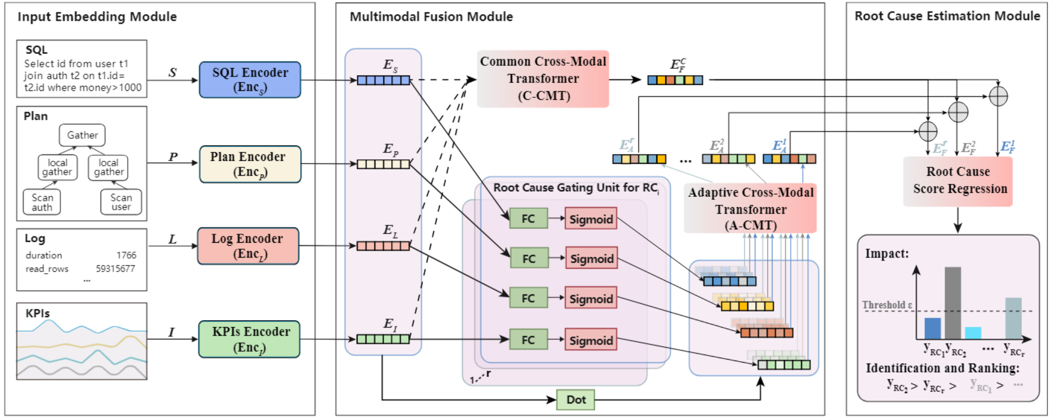


Multi-view joint graph representation learning for urban region embedding, IJCAI 2021

Alignment-based Fusion

- RCRank: Ranking of root causes of slow SQL queries in cloud databases
  - SQL statements, logs, KPIs, and query plans
  - Ranking of potential root causes that result in slow queries





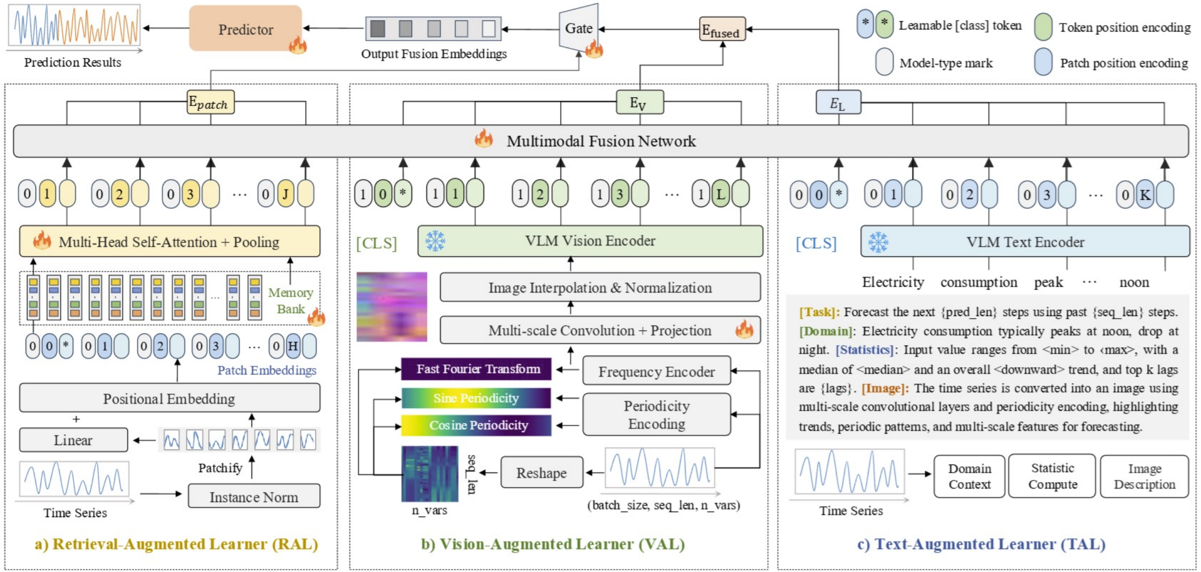
Pretraining an encoder for each modality

Cross modal fusion
Commonality vs. Specificity

Root cause ranking

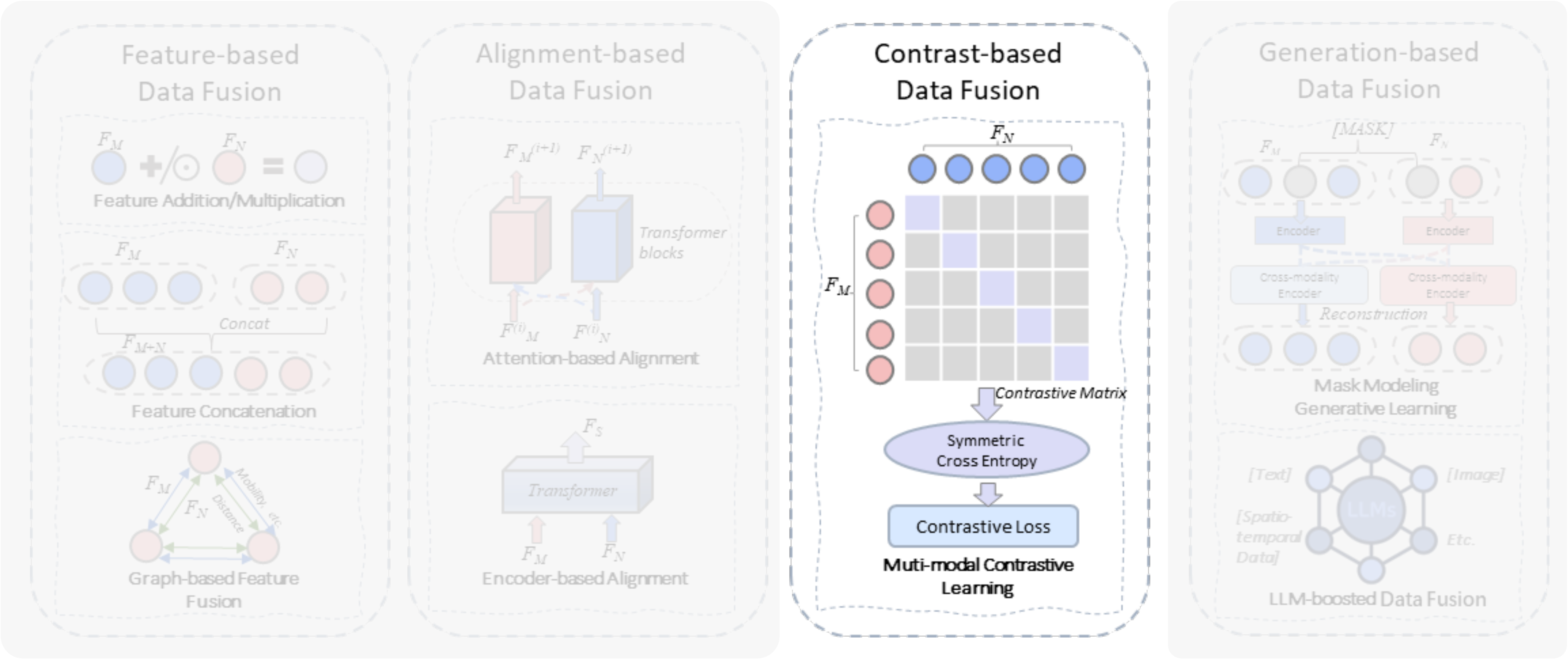
Encoder-based Fusion

- Token-level concatenation: Unified representations across modalities;
- Usually based on Self-Attention



Deep Learning-based Fusion Methods

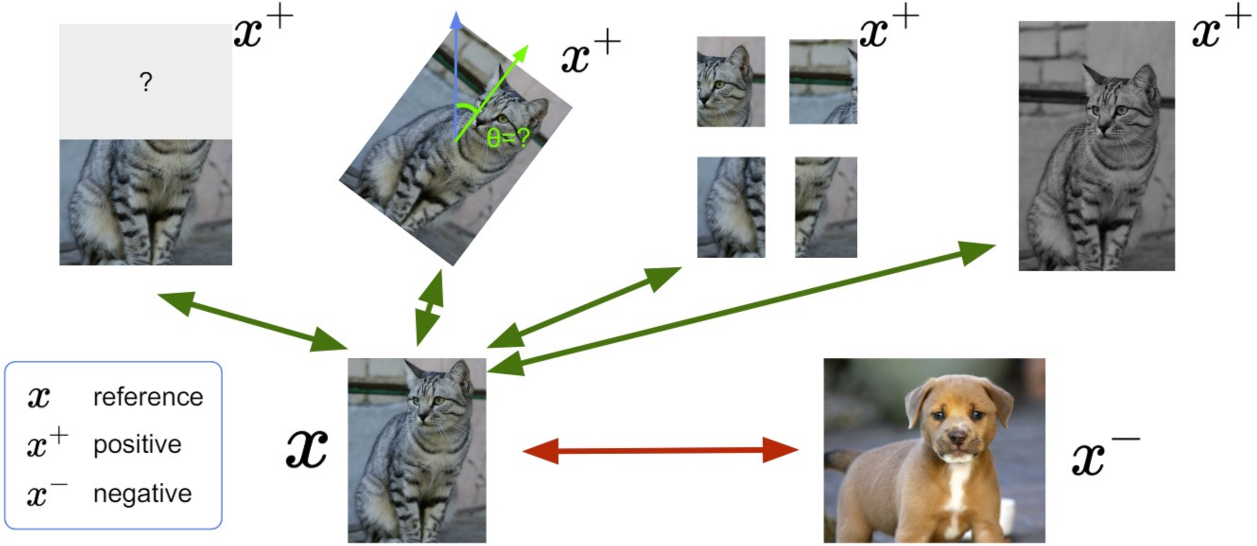
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)



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

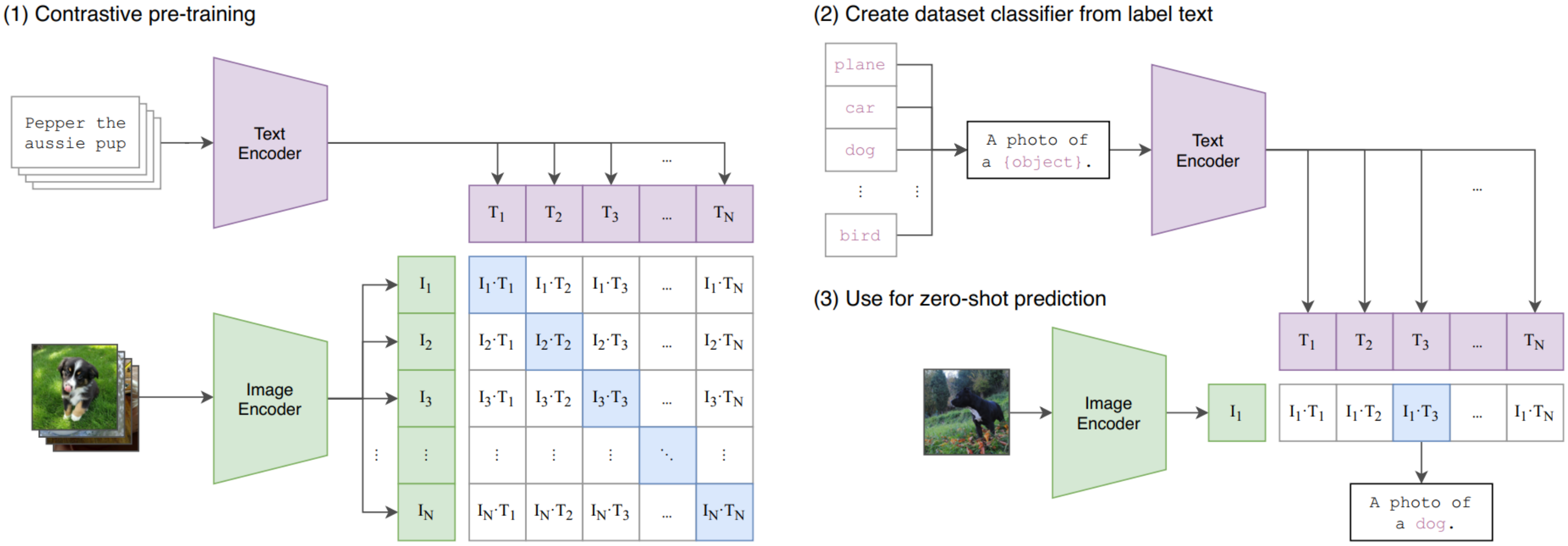
Contrastive Learning

- A representative class of self-supervised learning
- Building negative and positive samples to provide supervision signals



Contrastive Learning

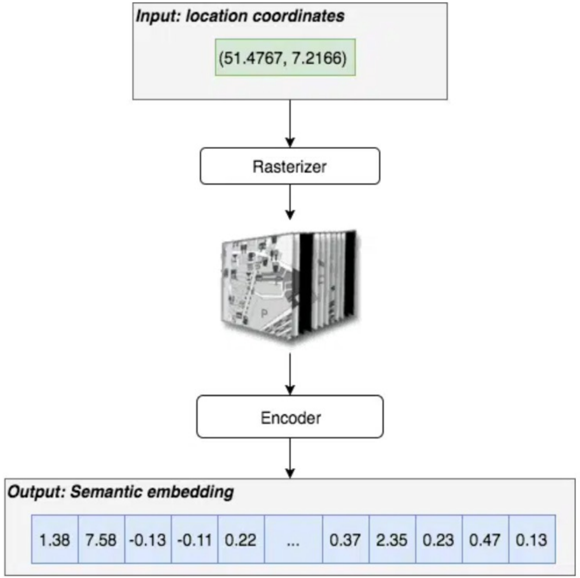
- How to fuse two modalities using contrastive learning?
- The answer is CLIP!

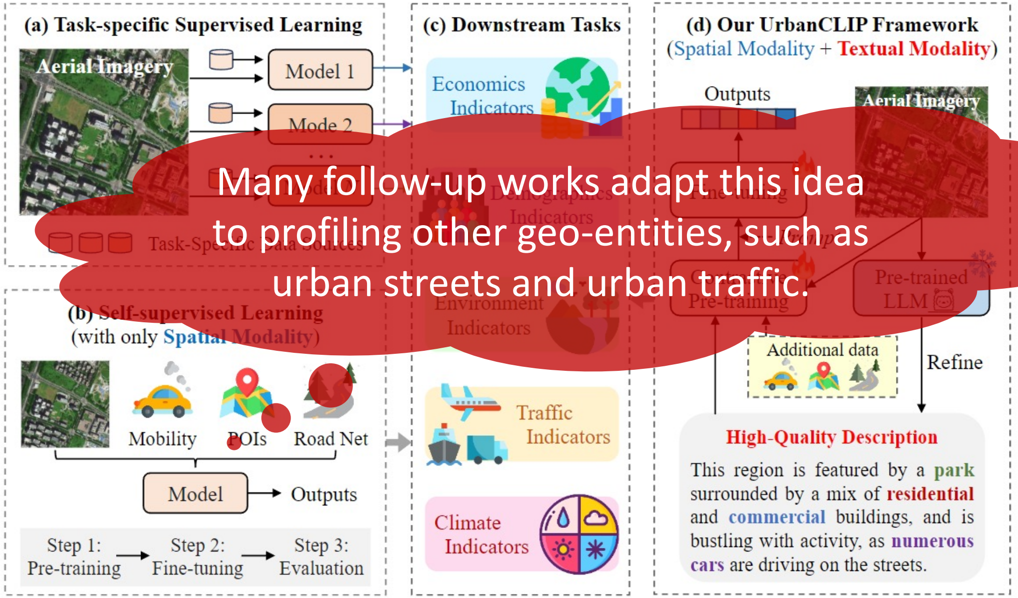


A. Radford et al., Learning Transferable Visual Models From Natural Language Supervision. ICML 2021

Contrast-based Fusion

- **Urban Contrastive Language-Image Pre-training** (UrbanCLIP) is the first framework that integrates the knowledge of text modality into urban region profiling.



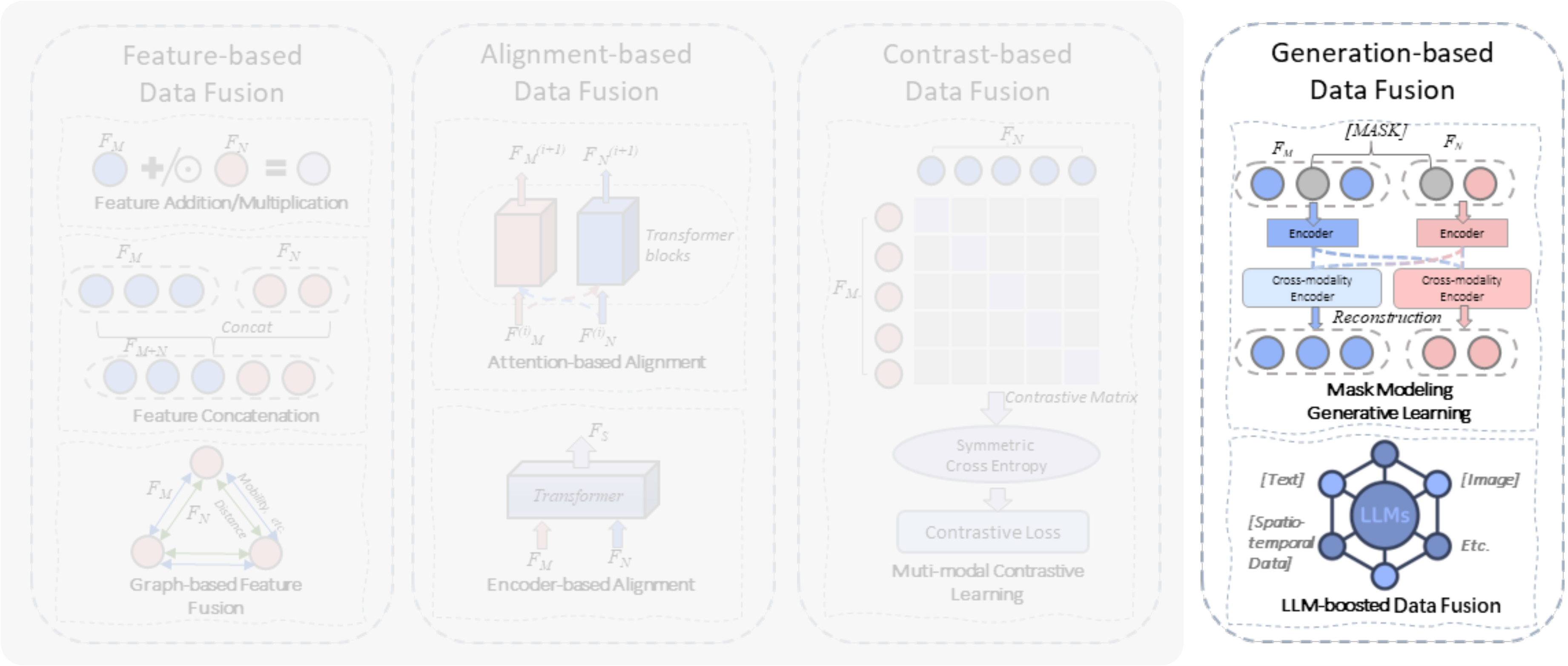


Many follow-up works adapt this idea to profiling other geo-entities, such as urban streets and urban traffic.

UrbanCLIP: Learning Text-enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web. WWW 2024

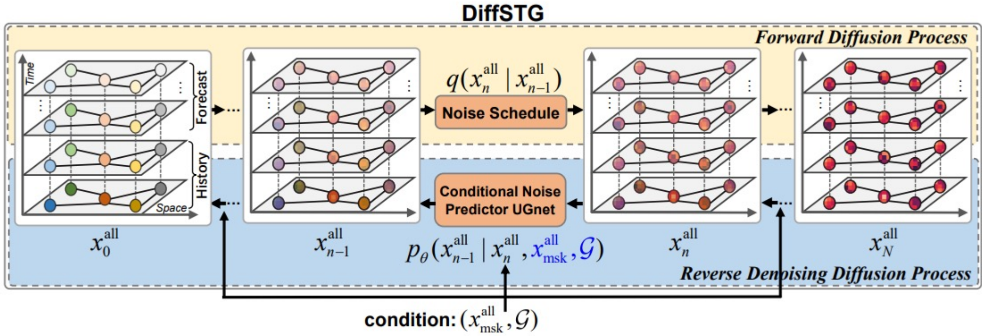
Deep Learning-based Fusion Methods

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

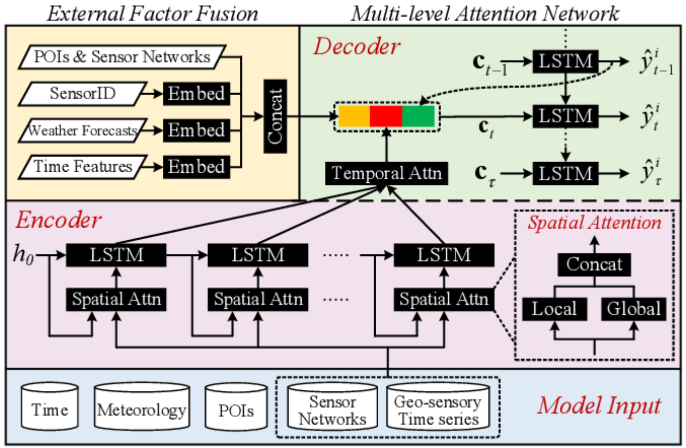


Generation-based Data Fusion

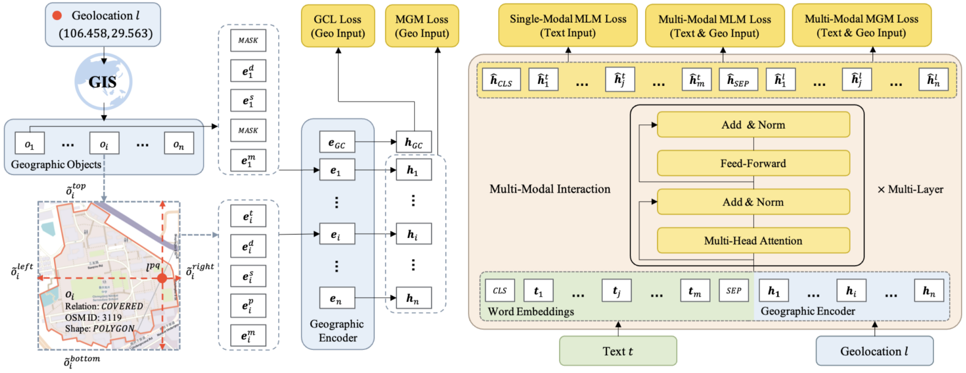
- Autoregression-based fusion
- Masked modeling-based fusion
- Diffusion-based fusion



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



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



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

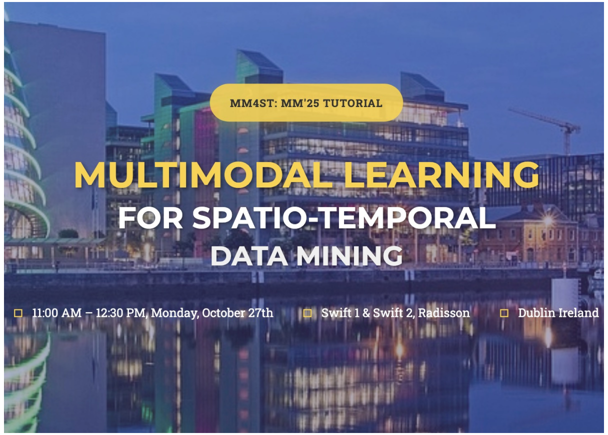
More Works

Summary Table

Category	Method	Data Source	Modality									Application	Institution	Year
			General Spatio-temporal					Visual		Textual				
			Time series	POI / Location	Trajectory/ Road network	Mobility	ST events	Satellite image	Street-view image	Social media text	Geo-information text			
Feature Based Data Fusion	DeepST [363]	✚▲❖	●		●							Transportation	Microsoft	2016
	ST-ResNet [364]	✚▲❖✱	●	●	●							Transportation	Microsoft	2018
	ST-MetaNet+ [218]	✚▲	●	●	●							Transportation	JD Research	2020
	DeepCrime [108]	▲◆❖	●	●	●							Social	JD Research	2021
	STUKG [278]	▲◆		●	●							Transportation	THU	2021
	DeepSTN+ [170]	✚▲	●	●								Transportation	THU	2019
	DeepTP [349]	✚▲		●		●						Transportation	THU	2021
	Guo et al. [94]	✚◆	●	●	●							Transportation	BUAA	2019
	Photo2Trip [386]	✚▲◆		●					●			Transportation	SU/RU/UCA	2017
	ST-SHN [311]	▲❖	●	●								Public Safety	SCUT/HKU	2021
	GeoMAN [162]	▲✱	●	●								General	XDU	2018
	Huang et al. [112]	✚▲				●		●	●			Urban planning	PKU	2023
	Liang et al. [164]	✚	●	●	●							Transportation	NUS	2021
	Balsebre et al. [13]	▲		●							●	Urban Planning	NTU	2022
	Ruan et al. [235]	▲✚	●	●								Transportation	NTU	2022
	Liu et al. [172]	▲◆		●			●					Economy	HKUST(GZ)	2023
	PANDA [342]	✚◆❖✱	●		●							Public Safety	XMU	2022
	UVLens [40]	✚▲❖			●	●		●				Urban Planning	XMU	2021
	Miyazawa et al. [200]	▲◆		●	●					●		Transportation	SUSTech	2019
	NodeSense2Vec [33]	✚▲		●			●			●		Social	UCF	2021
	Keerthi Chandra et al. [134]	✚▲◆		●		●						Urban Planning	UCF	2020
	Fu et al. [74]	✚▲◆		●			●					General	UCF	2019
	Liu et al. [184]	◆		●						●		Social	Gatech	2022
	Yuan et al. [349]	▲	●	●								Transportation	RMIT	2021
	Bai et al. [12]	✚✱	●									Transportation	Shanghai AI Lab	2019
	Ke et al. [132]	✚	●	●								Transportation	Alibaba	2021
	Geng et al. [84]	✚		●	●							Transportation	Alibaba	2019
	Yao et al. [329]	✚	●	●								Transportation	PSU	2018
	Gao et al. [79]	✚	●	●	●	●	●					Transportation	SWJTU	2022
	DeepMob [249]	✚❖	●	●	●	●					●	Public Safety	SUSTech	2017
	Geng et al. [82]	✚▲		●	●	●						Transportation	HKUST	2019
Alignment Based Data Fusion	Xi et al. [309]	▲		●				●				General	THU	2022
	Zhang et al. [368]	✚▲❖		●		●						General	THU	2021
	Yuan et al. [349]	✚▲✱		●		●						Transportation	THU	2021
	Yin et al. [339]	✚▲❖✱			●							Urban Planning	NUS	2020
	GSNet [272]	✚▲✱	●	●	●							Public Safety	BJTU	2021
	Hashem et al. [98]	▲❖		●								General	NTU	2023
	TrajGAT [328]	✚		●	●							Transportation	NTU	2022
	RADAR [39]	✚	●		●			●				General	XMU	2018
	Wang et al. [287]	✚▲✱		●	●							General	CSU	2021
	Tedjopurnomo et al. [261]	✚	●	●								Transportation	RMIT	2021
	ERNIE-GeoL [110]	▲		●							●	General	Baidu	2022
	SAInf [193]	✚▲✱	●		●		●					Transportation	JD Research	2023
	Gao et al. [77]	✚▲	●	●		●	●					Transportation	SWJTU	2023
Contrast Baed Data Fusion	KnowCL [186]	▲◆❖		●				●	●			Economy	THU	2023
	Li et al. [154]	▲◆❖			●			●	●			Economy	THU	2022
	MMGR [10]	▲◆❖		●				●				General	NTU	2023
	ReMVC [366]	✚▲❖		●		●						Urban Planning	NTU	2022
	HMTRL [173]	✚	●		●							Transportation	UCF	2023
	Mao et al. [196]	✚	●		●							Transportation	Shanghai AI Lab	2022
	UrbanSTC [226]	✚	●			●						Transportation	JD Research	2022
	UrbanCLIP [323]	▲◆						●			●	General	HKUST(GZ)	2023
Generation Based Data Fusion	SG-GAN [378]	▲		●				●				Urban Planning	NUS	2020
	ActSTD [354]	▲◆	●		●							Transportation	THU	2022
	DiffSTG [299]	✚▲	●		●							General	BJTU	2023
	CP-Route [297]	✚▲	●		●							Transportation	BJTU	2023
	G2PTL [304]	✚▲◆		●	●						●	Transportation	Cainiao	2023
	DiffUFlow [401]	✚▲		●	●							Transportation	CSU	2023
	DP-TFI [320]	✚▲✱	●		●							Transportation	UESTC	2023
	Wang et al. [273]	✚▲❖		●		●	●					Urban Planning	UCF	2021
	Chattraffc[358]	✚				●					●	Transportation	BJUT	2023
	MGEO [61]	▲		●							●	General	Alibaba	2023

Outline

1. Background & Examples
2. Foundation of ST Data
3. Why Multimodal ST Data Fusion
4. Principle of ST Multimodal Fusion
5. **Visual/Language Knowledge Transfer**
6. Conclusions

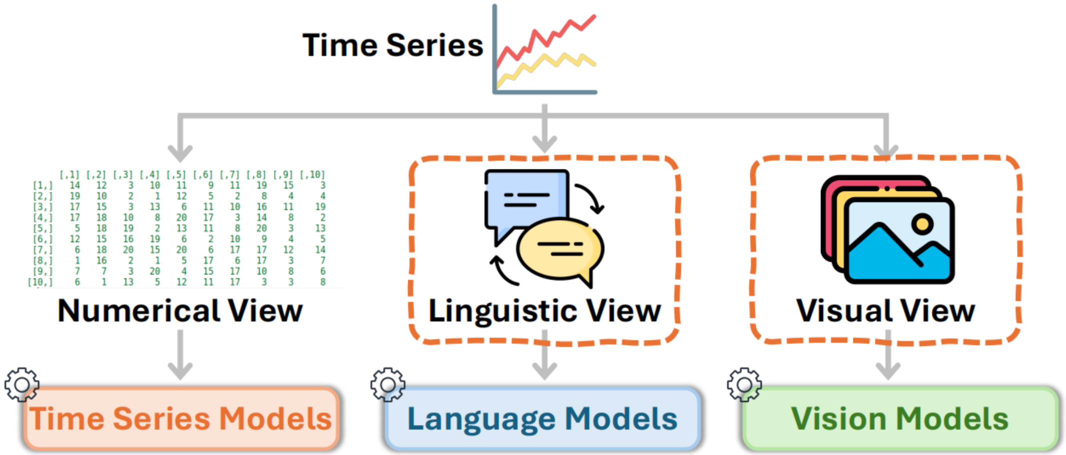


Scope

- We focus on spatially static and temporally dynamic data.

  i.e. standard time series and spatial time series data (e.g. traffic flow, air quality).

- We focus on how vision and language can enhance ST forecasting.



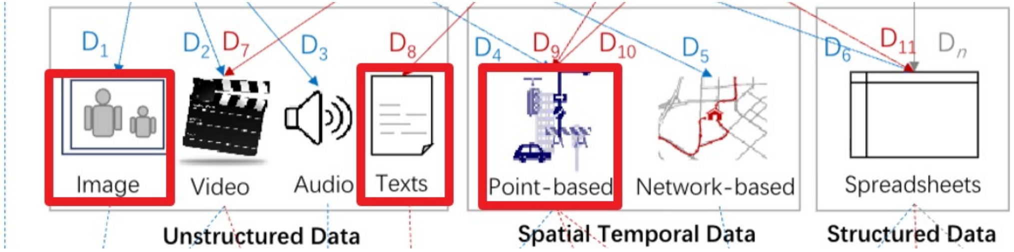
Scope

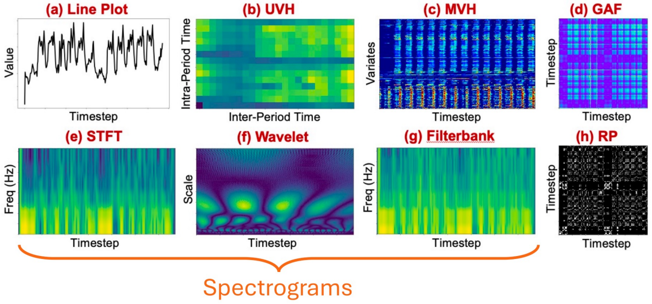
- We focus on spatially static and temporally dynamic data.

  i.e. standard time series and spatial time series data (e.g. traffic flow, air quality).

- We focus on how vision and language can enhance ST forecasting.

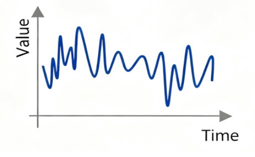
Data Modalities



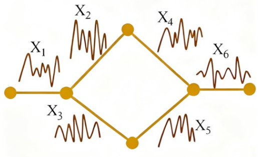


The Electricity Transformer Temperature (ETT) indicates the electric power long-term deployment. Each data point consists of the target oil temperature and 6 power load features ...
Below is the information about the input time series:

[BEGIN DATA]
***
[Domain]: We usually observe that electricity consumption peaks at noon, with a significant increase in transformer load
***
[Instruction]: Predict the next <H> steps given the previous <T> steps information attached
***
[Statistics]: The input has a minimum of <min_val>, a maximum of <max_val>, and a median of <median_val>. The overall trend is <upward or downward>. The top five lags are <lag_val>.
[END DATA]



Time Series



Spatial Time Series

Part 1

Language-enhanced Spatio-Temporal Analysis

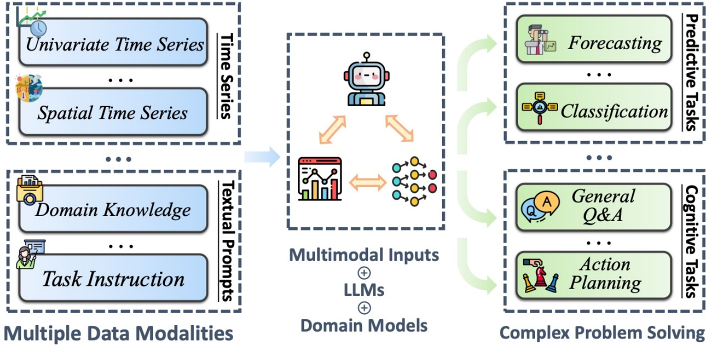
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

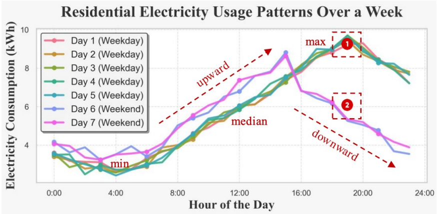
Why Language for ST

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Limitations of traditional methods
  - incomplete information
  - lack of causality
  - poor response to shocks

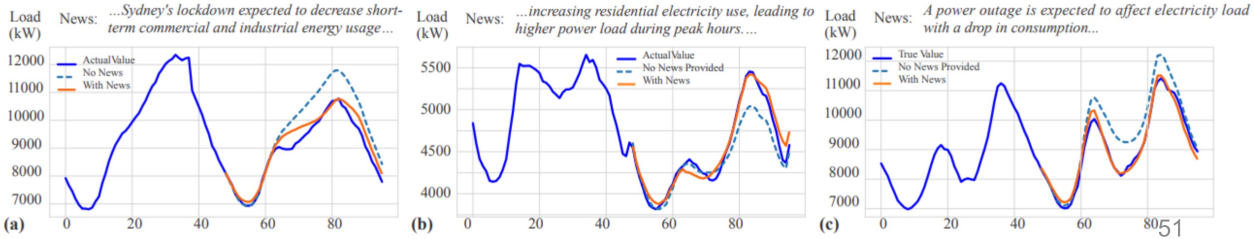
- Advantages brought by language
  - context provided
  - interpretability
  - robustness





Late Down Early Up … Steady Short Long Increase

[Context]: These are residential electricity usage patterns.
[Dataset]: Electricity consumption peaks on weekday evenings, drops during daytime, and reverses on rest days.
[Statistic]: The input range from <min> to <max>, with a median of <median>. The overall trend is <upward>.



Challenge

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Data heterogeneity

  Time-series data are orderly continuous numerical signals, while text is a high-dimensional, discrete symbolic expression.

- Temporal alignment

  There exists an uncertain lag effect or asynchrony between textual events and peaks in numerical sequences.

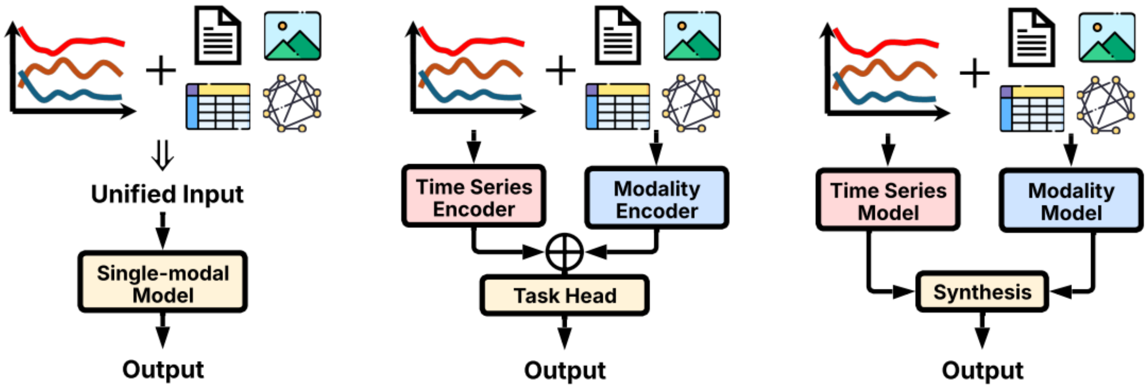
- Noise and irrelevant context

  Compared to the vast amount of daily text, event descriptions truly related to temporal dynamics are extremely sparse and often implicitly expressed.

Cross-modal Fusion

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- The process of Integrating heterogeneous modalities (Text, ST Data) in a way that captures complementary information across diverse sources.
- Three stages of fusion: input level, intermediate level, output level



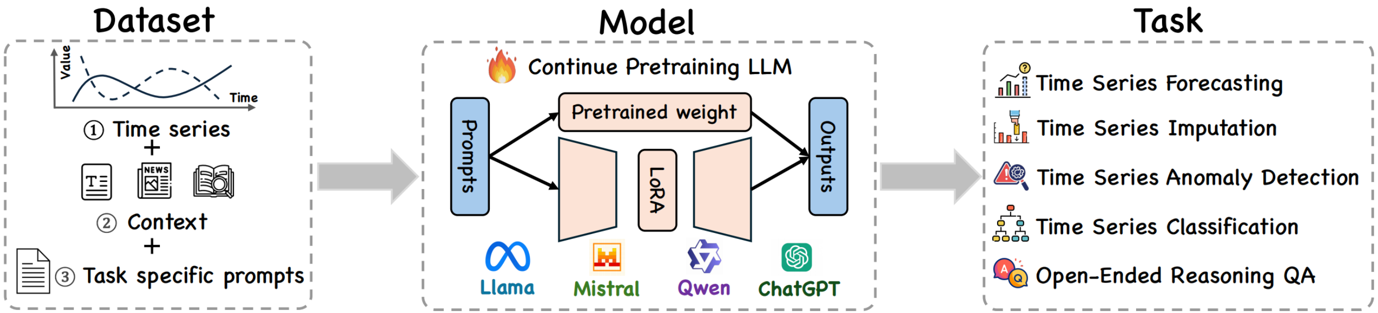
High parameter efficiency
Strong deep feature interaction
Potential modality imbalance
limited flexibility

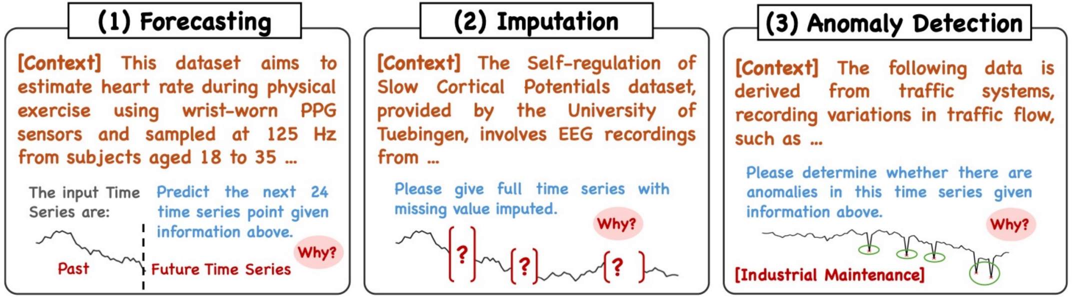
High flexibility
Superior performance
Complex model design High
computational cost

Simple implementation
Strong robustness, Flexibility
Low performance ceiling
Requires modality independence.

Input Level Fusion

- Integrate time series and texts into a unified textual prompt.

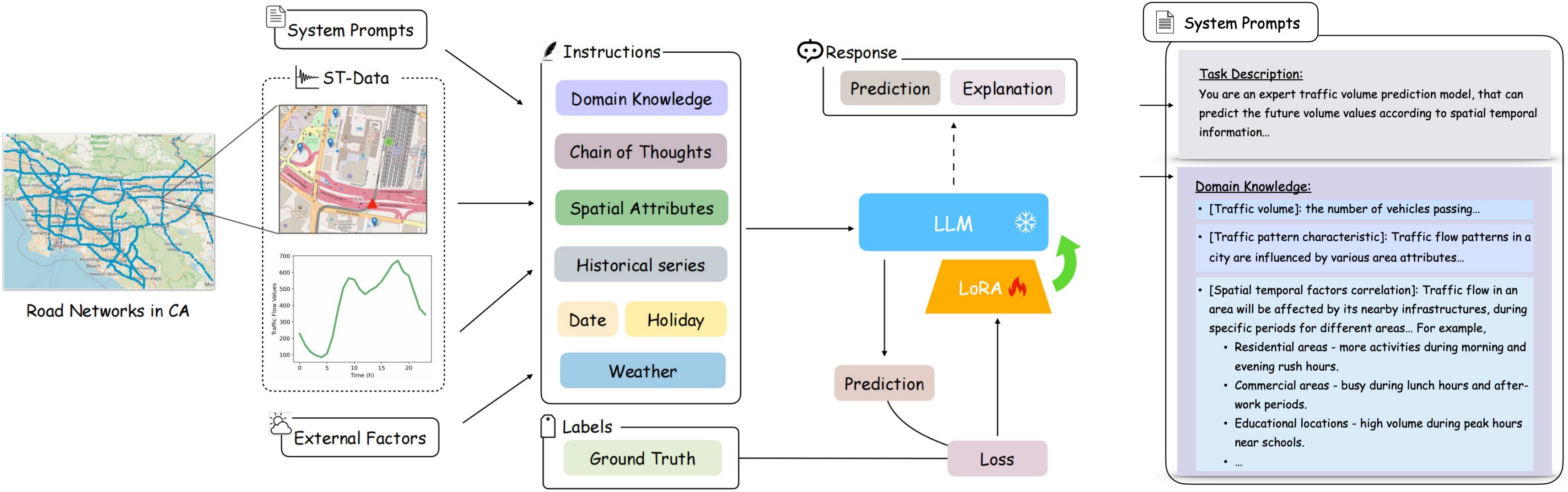




Time-MQA: Time Series Multi-Task Question Answering with Context Enhancement. In ACL, 2025.

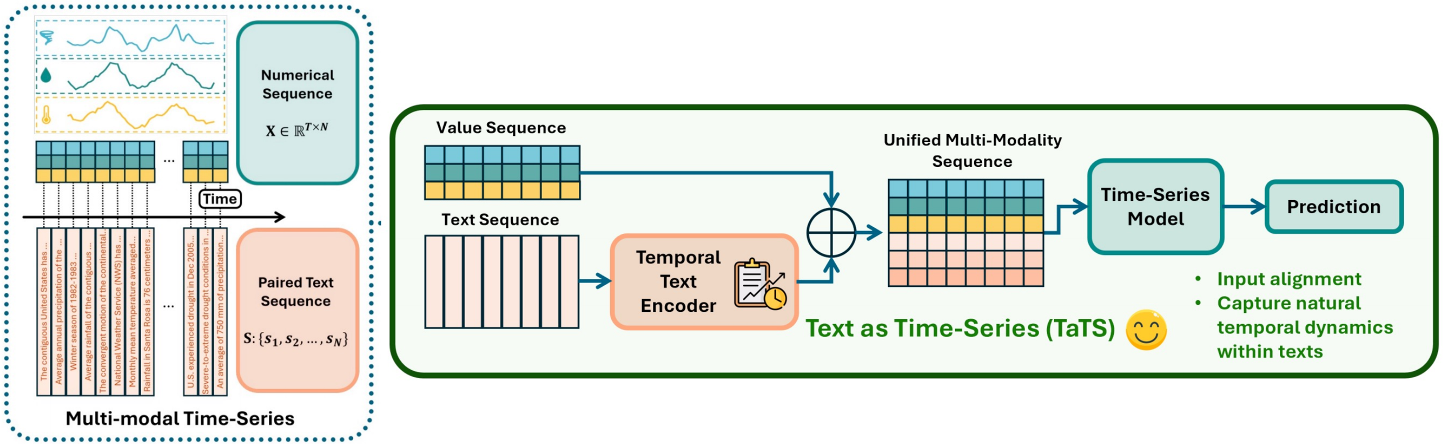
Input Level Fusion

- Integrate time series and texts into a unified textual prompt.



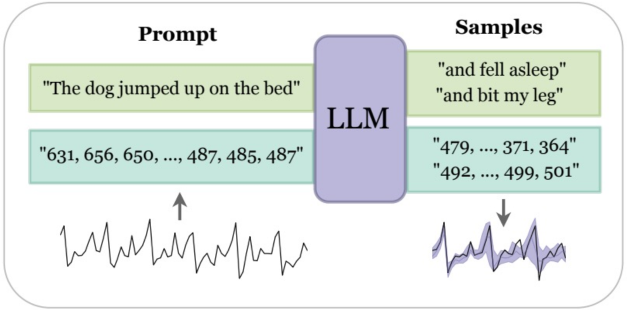
Input Level Fusion

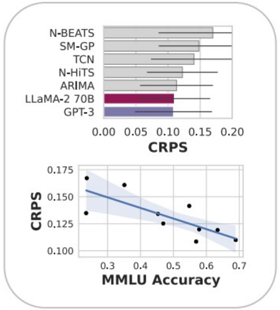
- Integrate paired text embedding as an additional variable of time series.

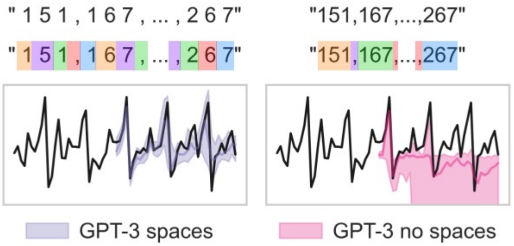


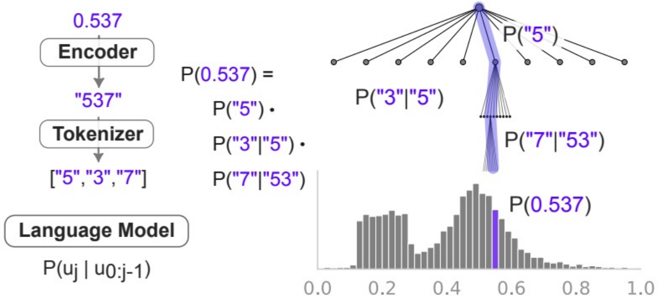
Input Level Fusion

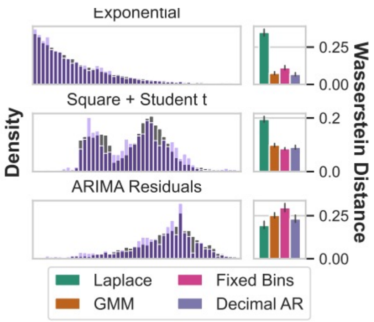
- Describe time series as discrete marks, using LLM's autoregressive generation ability

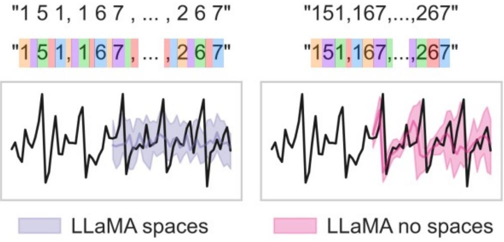








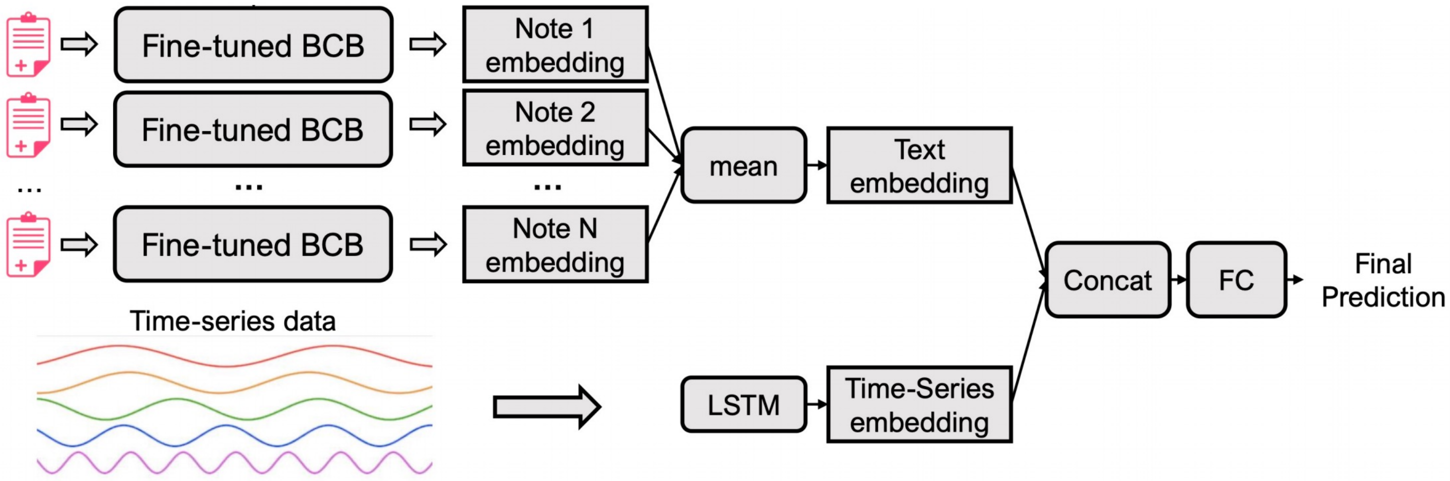




Large language models are zero-shot time series forecasters, In NeurIPS, 2023.

Intermediate Level Fusion

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



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

Intermediate Level Fusion

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- The fusion of modality embeddings is usually followed by alignments.

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

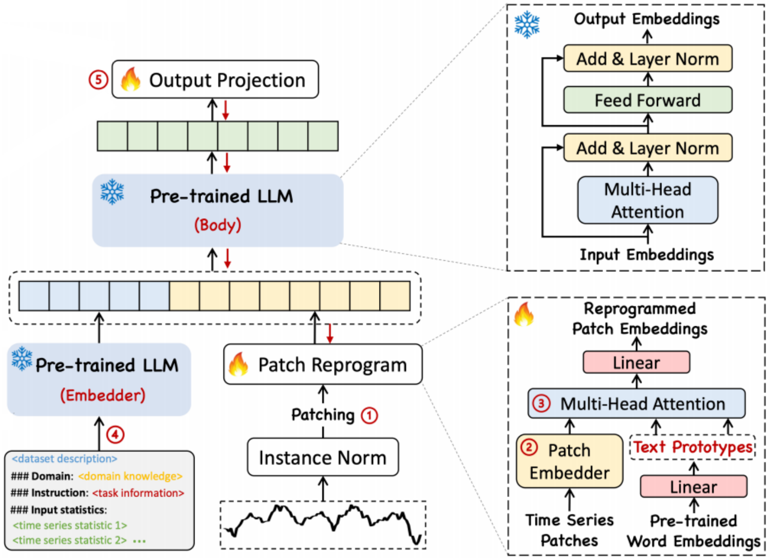
  - self-attention, cross-attention, gating

  - graph convolution

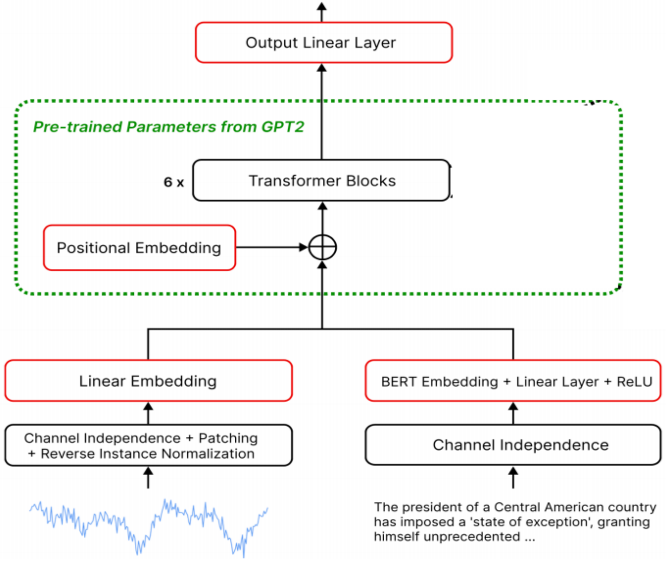
  - learning objectives

Intermediate Level Fusion

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



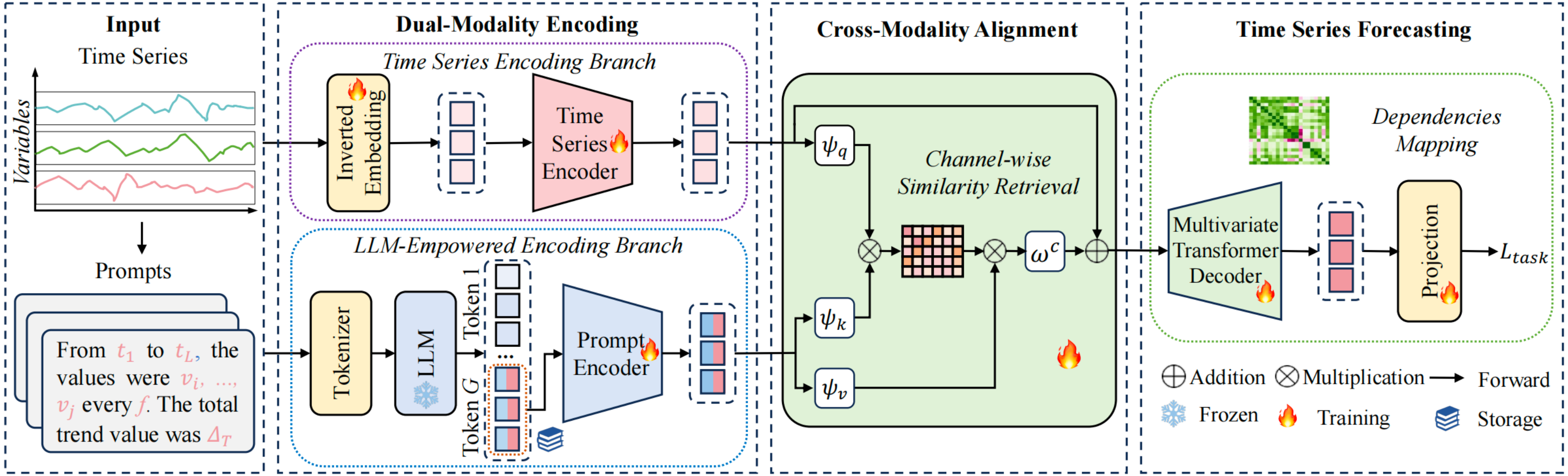
Time-LLM: Time Series Forecasting by Reprogramming Large Language Models, In ICLR, 2024.



GPT4MTS: Prompt-Based Large Language Model for Multimodal Time Series Forecasting, In AAAI, 2024.

Intermediate Level Fusion

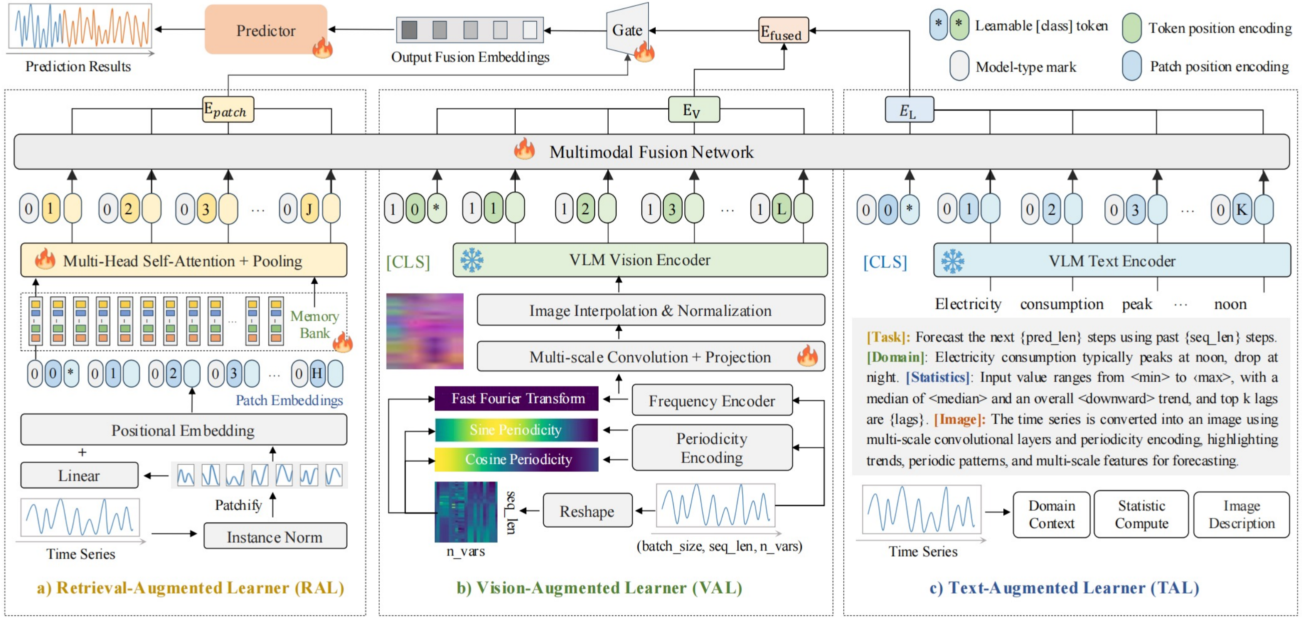
- Cross-attention: time series serves as the query modality to get contextualized by other modalities, providing a directed alignment that ensure auxiliary modalities contribute relevant contexts while preserving the temporal structure of time series.



TimeCMA: Towards LLM-Empowered Multivariate Time Series Forecasting via Cross-Modality Alignment, In AAAI, 2025.

Intermediate Level Fusion

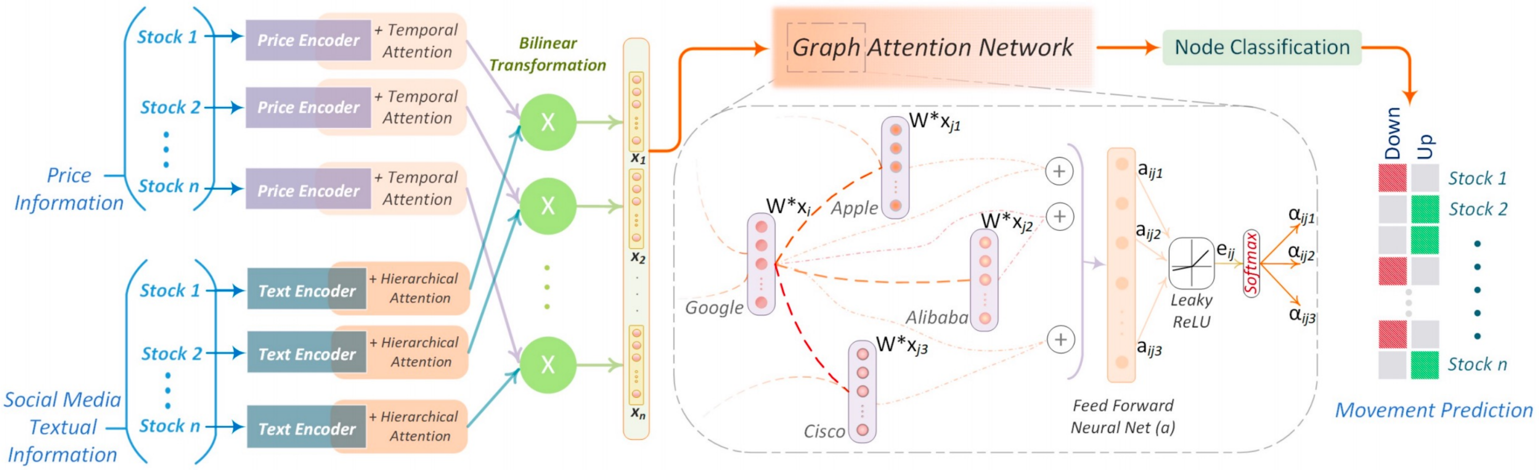
- Gating: a parametric filtering operation that explicitly regulates the influence of time series and other modalities on the fused embeddings.



Time-VLM: Exploring multimodal vision-language models for augmented time series forecasting. In ICML, 2025.

Intermediate Level Fusion

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

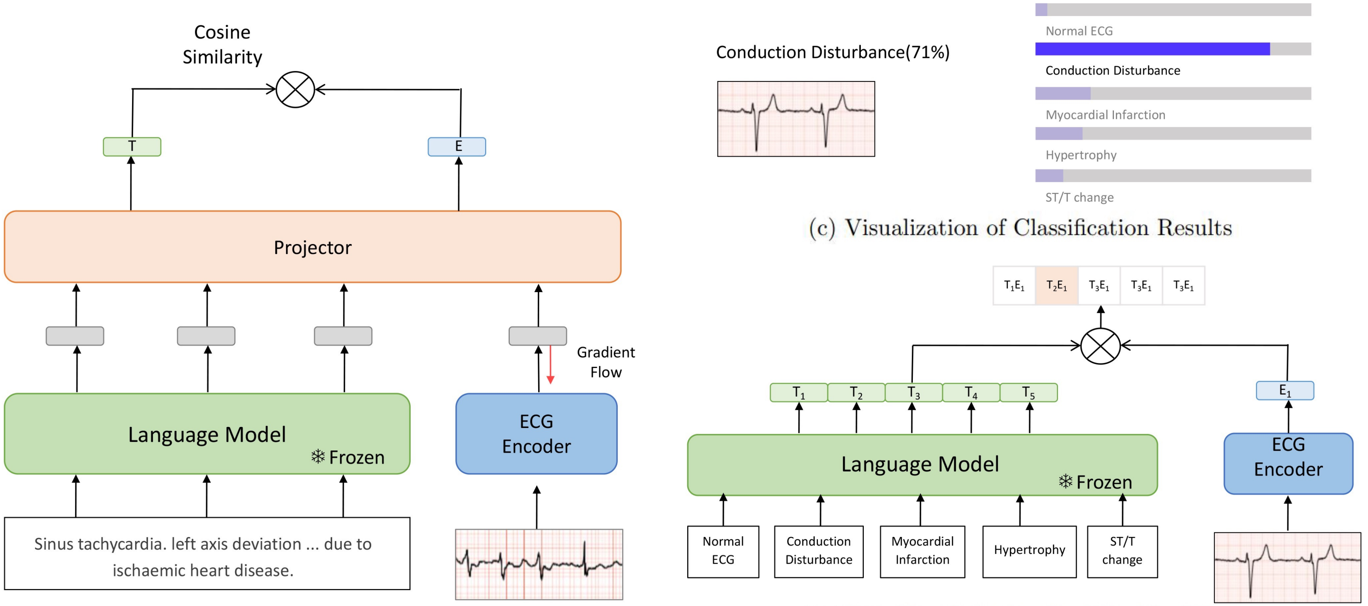


Deep Attentive Learning for Stock Movement Prediction from Social Media Text and Company Correlations, In EMNLP, 2020

Intermediate Level Fusion

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

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



(a) Self-supervised Learning pre-training

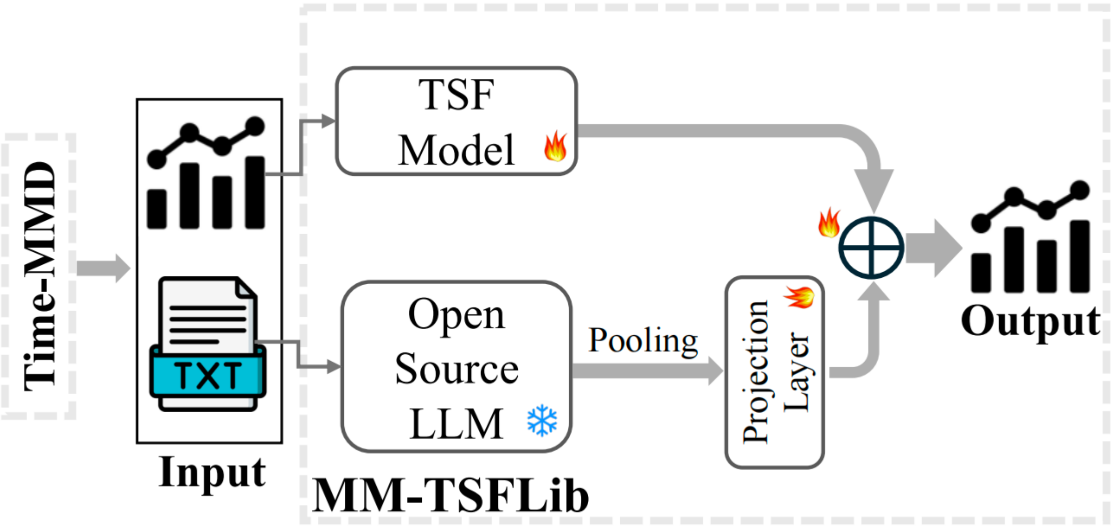
(c) Visualization of Classification Results

(b) Zero-Shot Learning for Classification

Frozen Language Model Helps ECG Zero-Shot Learning, In MIDL, 2023

Output Level Fusion

- Project multiple modality outputs onto a unified space.



Time-MMD: Multi-Domain Multimodal Dataset for Time Series Analysis, In NeurIPS, 2024.

Summary

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

✅ Leveraging LLMs' reasoning capabilities

✅ Straightforward to integrate additional textual data

✅ Potential to provide explanation

⛔ Model long time series

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

⛔ Perform long-term forecasting

Part 2

Vision-enhanced Spatio-Temporal Analysis

Motivation of Vision for TS

- Compared to LLM, **vision model** has more advantages:

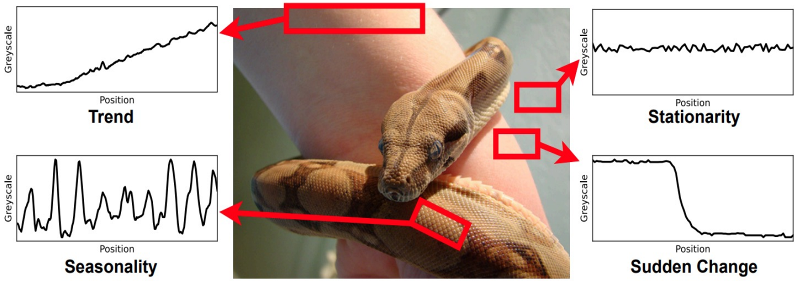
  - Using continuous pixel sequences (vs. text's discrete tokens).

  - Supporting multivariate time series (vs. LLM follows channel independence).

  - Compactly encoding long time series (vs. LLM's context length/precision limits).

  - Enabling more intuitive human/system understanding.

- Multimodal LLM perhaps simultaneously integrate the advantages of both.

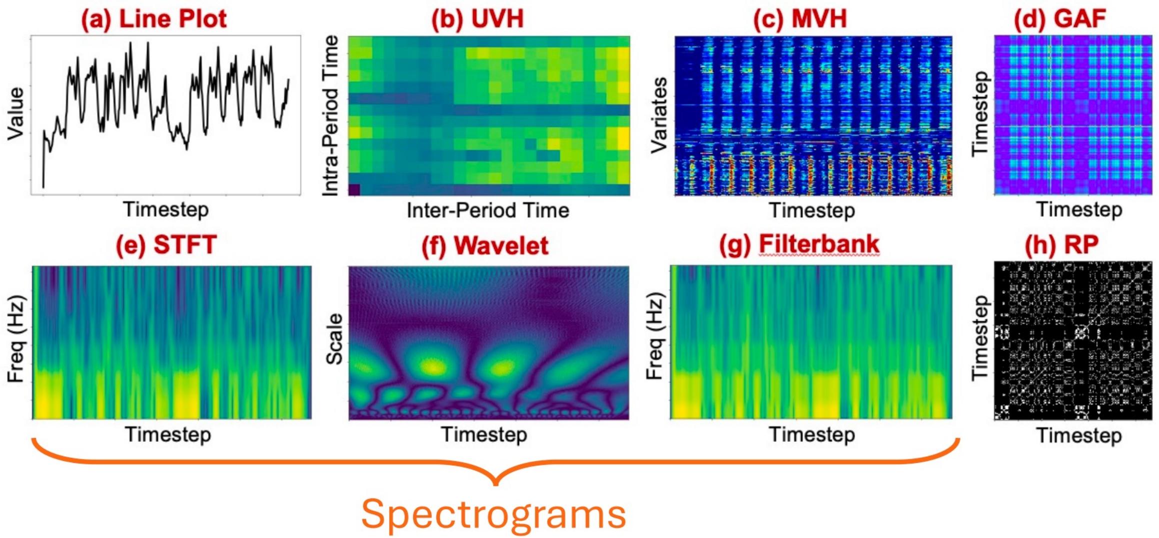


	Characteristics	Origin	Information
Time series	continuous	physical systems	high redundancy
Image	continuous	physical systems	high redundancy
Text	discrete	human cognitive construct	semantically dense

TS-to-Vision Transformation

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- There are 8 major time-series imaging methods:



Method	TS-Type	Advantages	Limitations
Line Plot (§3.1)	UTS, MTS	matches human perception of time series	limited to MTS with a small number of variates
Heatmap (§3.2)	UTS, MTS	straightforward for both UTS and MTS	the order of variates may affect their correlation learning
Spectrogram (§3.3)	UTS	encodes the time-frequency space	limited to UTS; needs a proper choice of window/wavelet
GAF (§3.4)	UTS	encodes the temporal correlations in a UTS	limited to UTS; $O(T^2)$ time and space complexity
RP (§3.5)	UTS	flexibility in image size by tuning m and τ	limited to UTS; information loss after thresholding

Harnessing Vision Models for Time Series Analysis: A Survey. In IJCAI, 2025.

Line Plot — Intro

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Line Plot is a 2D image with time on x-axis, values on y-axis, and a line connecting points.

- **Ex.1**: Line Plot Imaging for Financial Univariate Time Series Classification.

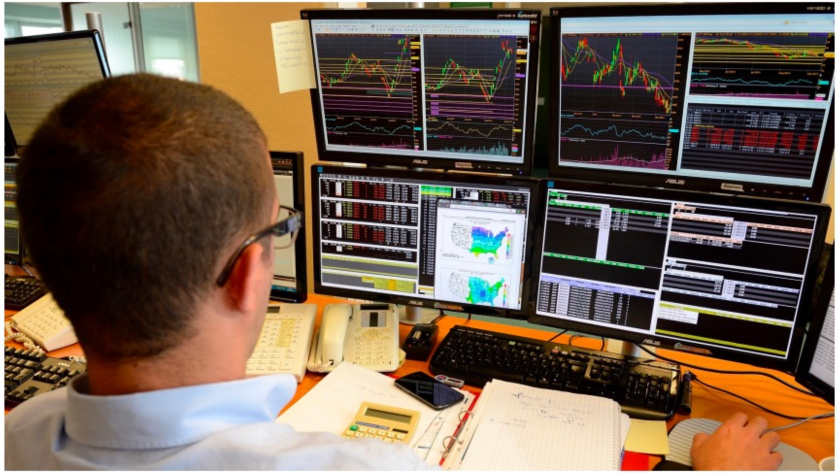
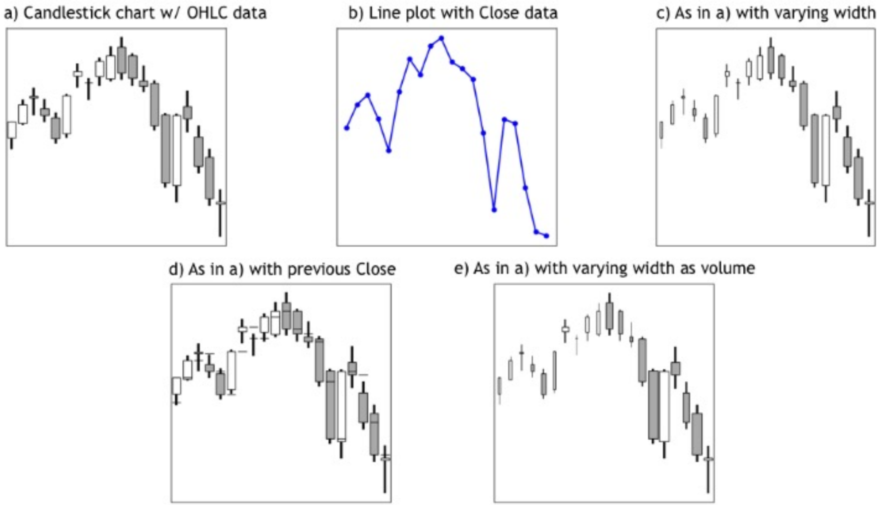


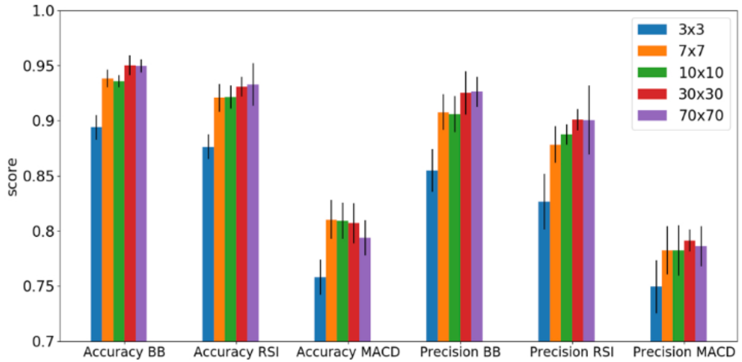
Figure 1: Typical workstation of a professional trader.

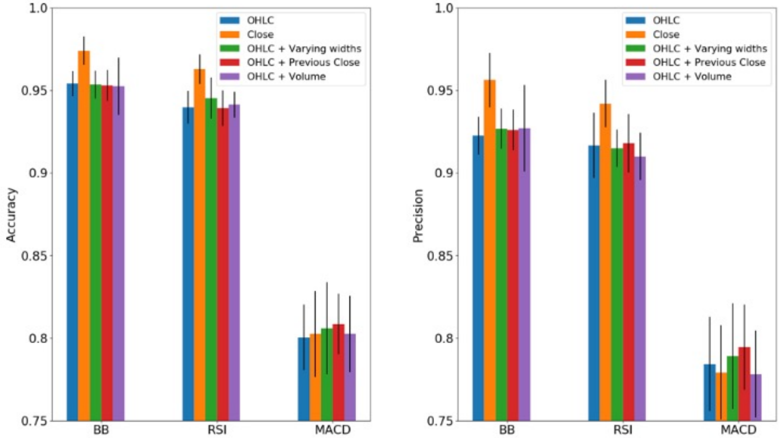
Figure 2: Converting continuous time series to images.

Line Plot — e.g. Univariate TSC

- Ablations on Imaging Details and Resolution.







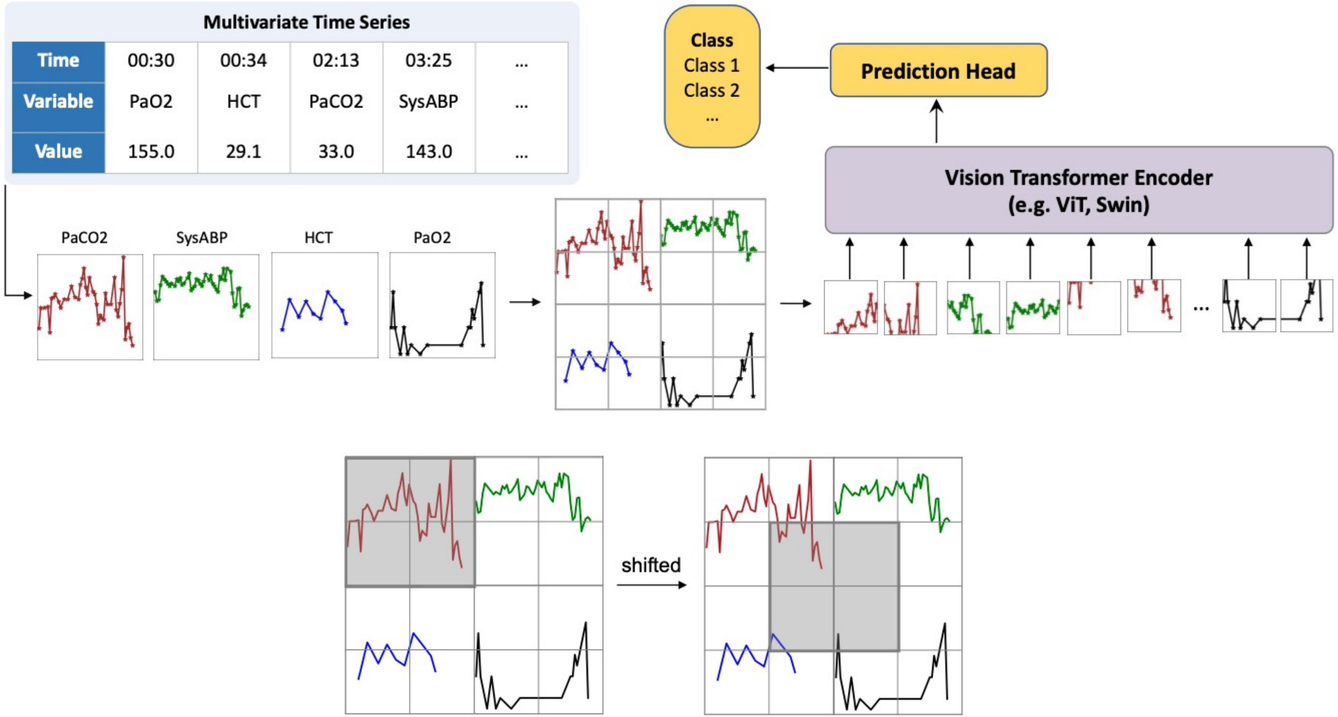
1. Simple line plot is better
2. Sufficient resolution is important

Trading via image classification. In ICAIF, 2020.

Line Plot — e.g. Multivariate TSC

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- **Ex.2**: Line Plot Imaging for irregularly Multivariate Time Series Classification.



Time series as images: Vision transformer for irregularly sampled time series. NeurIPS, 2023.

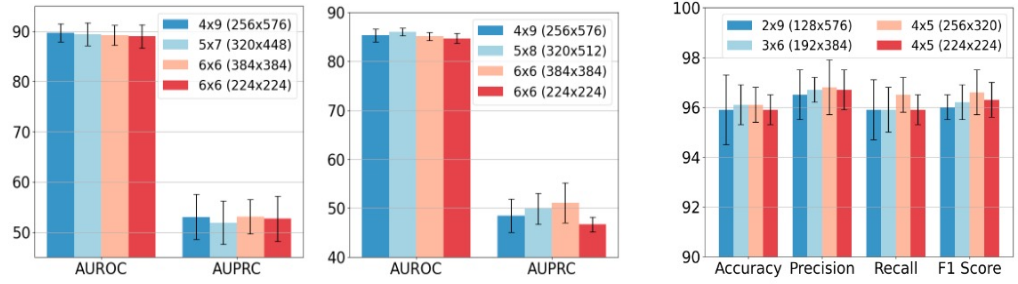
Line Plot — e.g. Multivariate TSC

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Ablations on Time Series Imaging Strategies and Details.

Table 3: Ablation studies on different strategies of time series-to-image transformation.

Methods	P19 AUROC	P19 AUPRC	P12 AUROC	P12 AUPRC	PAM Accuracy	PAM Precision	PAM Recall	PAM F1 score
Default	89.2 ± 2.0	53.1 ± 3.4	85.1 ± 0.8	51.1 ± 4.1	95.8 ± 1.3	96.2 ± 1.1	96.2 ± 1.3	96.5 ± 1.2
w/o interpolation	89.6 ± 2.1	52.9 ± 3.4	85.7 ± 1.0	51.9 ± 3.4	95.6 ± 1.1	96.6 ± 0.9	95.9 ± 1.0	96.2 ± 1.0
w/o markers	89.0 ± 2.1	51.7 ± 2.5	85.3 ± 0.9	50.3 ± 3.2	95.8 ± 1.1	96.9 ± 0.7	96.0 ± 1.0	96.4 ± 0.9
w/o colors	88.8 ± 1.8	51.4 ± 4.1	84.4 ± 0.7	47.0 ± 2.9	95.0 ± 1.0	96.2 ± 0.7	95.3 ± 1.0	95.7 ± 0.9
w/o order	89.3 ± 2.3	52.7 ± 4.5	84.0 ± 1.8	47.8 ± 4.6	-	-	-	-



(a) P19 (b) P12 (c) PAM

Figure 5: Ablation study of the influence of grid layouts and image sizes. For instance, 4x9 (256x576) denotes a grid layout of 4×9 with an image size of 256×576 pixels.

Table 4: Robustness regarding the style and size of lines and markers. In the brackets, the first element denotes style, and the second represents size.

Line	Marker	AUROC	AUPRC
(solid,1)	($*$,2)	89.2 ± 2.0	53.1 ± 3.4
(dashed,1)	($*$,2)	89.2 ± 2.1	53.7 ± 4.1
(dotted,1)	($*$,2)	89.2 ± 2.1	52.8 ± 4.0
(solid,0.5)	($*$,2)	88.6 ± 1.7	53.0 ± 3.6
(solid,1)	($*$,2)	89.2 ± 2.0	53.1 ± 3.4
(solid,2)	($*$,2)	88.5 ± 2.3	53.6 ± 3.1
(solid,1)	($*$,2)	89.2 ± 2.0	53.1 ± 3.4
(solid,1)	($\wedge$,2)	89.3 ± 1.9	52.6 ± 4.0
(solid,1)	($\circ$,2)	89.1 ± 1.9	51.3 ± 4.2
(solid,1)	($*$,1)	88.2 ± 1.4	52.1 ± 4.5
(solid,1)	($*$,2)	89.2 ± 2.0	53.1 ± 3.4
(solid,1)	($*$,3)	88.9 ± 1.9	52.8 ± 3.2

1. Linear interpolation of two nodes is useless.
2. The style and size of marks and lines are robust.
3. Color differentiation is very important for MTS.
4. The order of multivariate subgraphs is robust.

Time series as images: Vision transformer for irregularly sampled time series. NeurIPS, 2023.

Line Plot — e.g. Multivariate TSC

- Vision backbone analysis

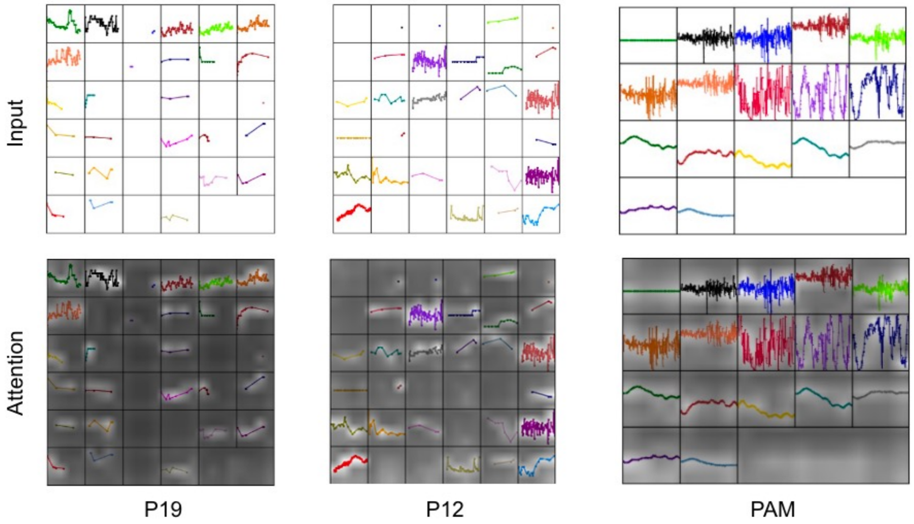


Figure 6: Illustration of the averaged attention map of ViTST.

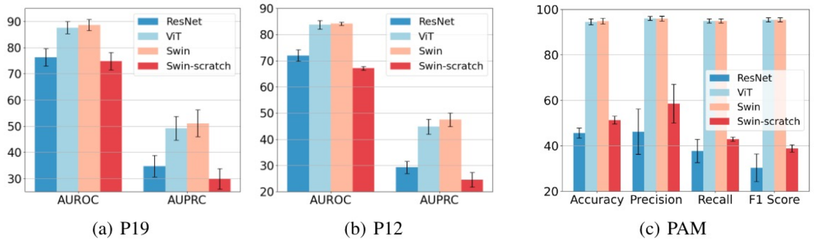
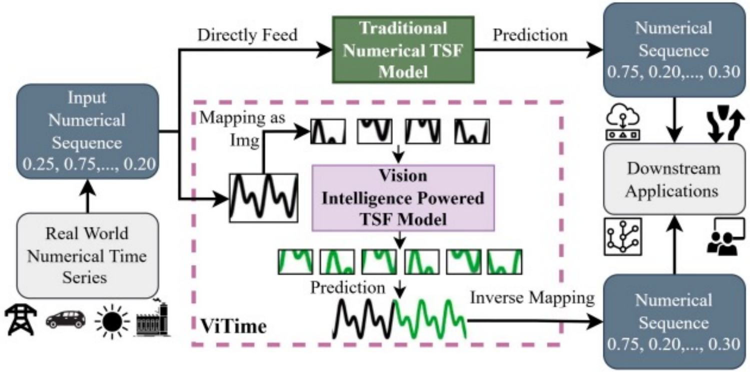


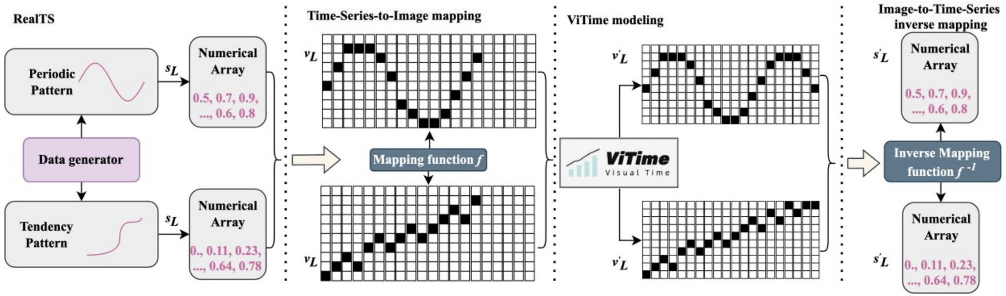
Figure 4: Performance of different backbone vision models on P19, P12, and PAM datasets. We do not use static features for our approach here to exclude their influence.

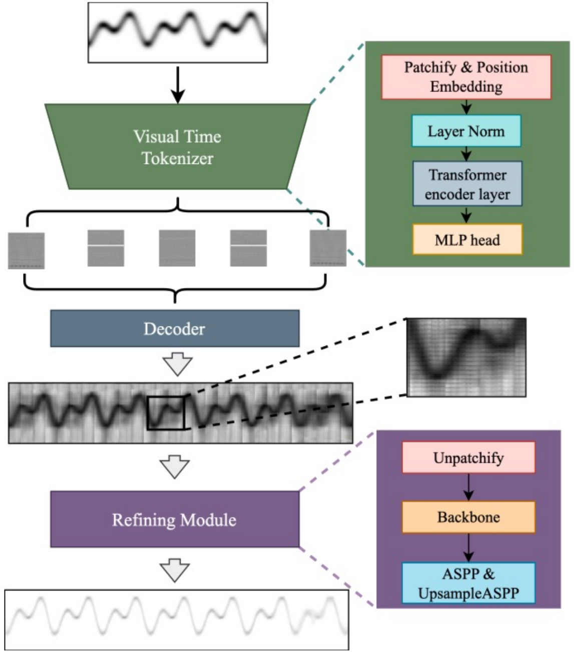
1. Transformer (ViT) better captures spatial correlations compared to CNN (ResNet).
2. it can focus on the meaningful parts of TS images.
3. The pretrained vision knowledge is useful.

Line Plot — e.g. TSFM (ViTime)

- **Ex.3**: First Vision-based Foundation Model for TSF.







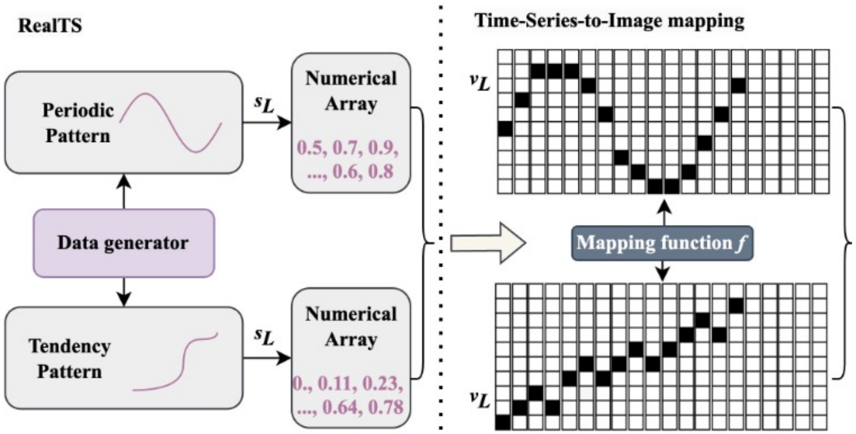
$$\mathcal{L} = d\left(\mathbf{v}'_{\mathbf{L}}, \mathbf{v}_{\mathbf{L}}\right) + \alpha \mathrm{KLD}\left(\mathbf{v}'_{\mathbf{L}}, \mathbf{v}_{\mathbf{L}}\right)$$

Eearth Moving Distance + KL Divergence

Line Plot — e.g. TSFM (ViTime)

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Theoretical Advantages of Visual Intelligence for TSF.



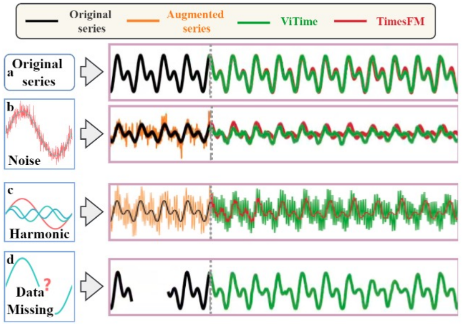


Figure 6: Performance comparison of ViTime versus TimesFM on TSF tasks under various data perturbations: a. Original time series. b. Time series with noises injected. c. Time series with harmonic added. d. Time series with missing data.

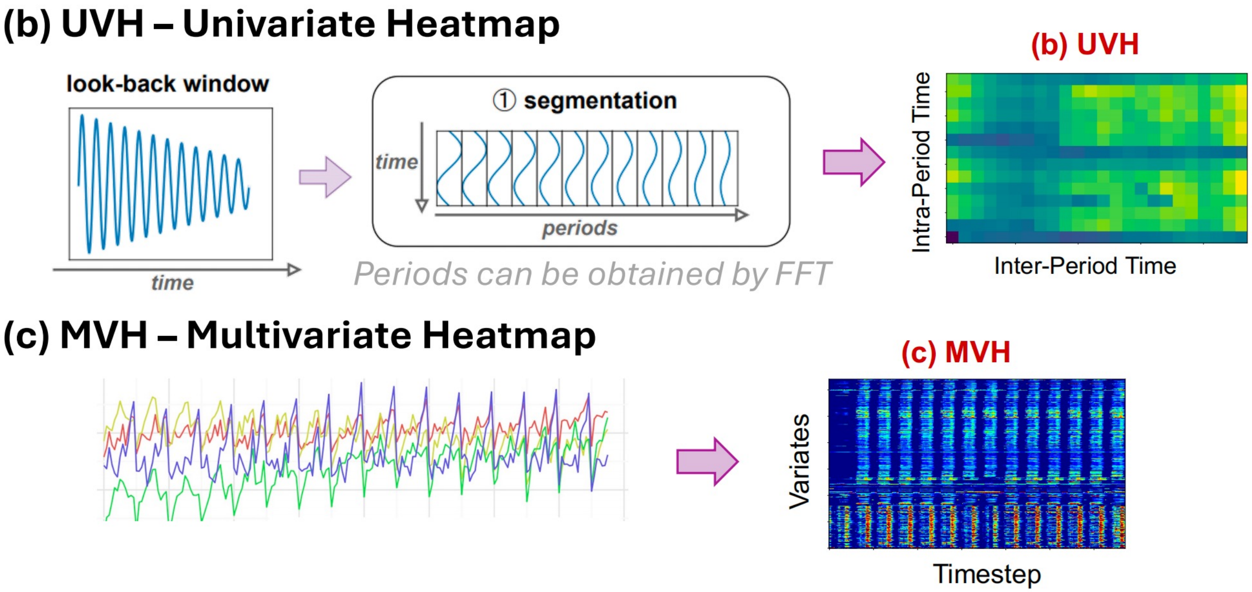
1. **Spatiotemporal Isometry**: value changes in TS is proportional to pixel variations in images.
2. **Pattern Preservation**: Visual Fourier spectra matching original time series.
3. **Geometric Regularization**: limited resolution of the image resists disturbance, small disturbance in TS only causes bitty changes in visual embedding.

ViTime: A visual intelligence-based foundation model for time series forecasting. arXiv 2024.

Heatmap — Intro

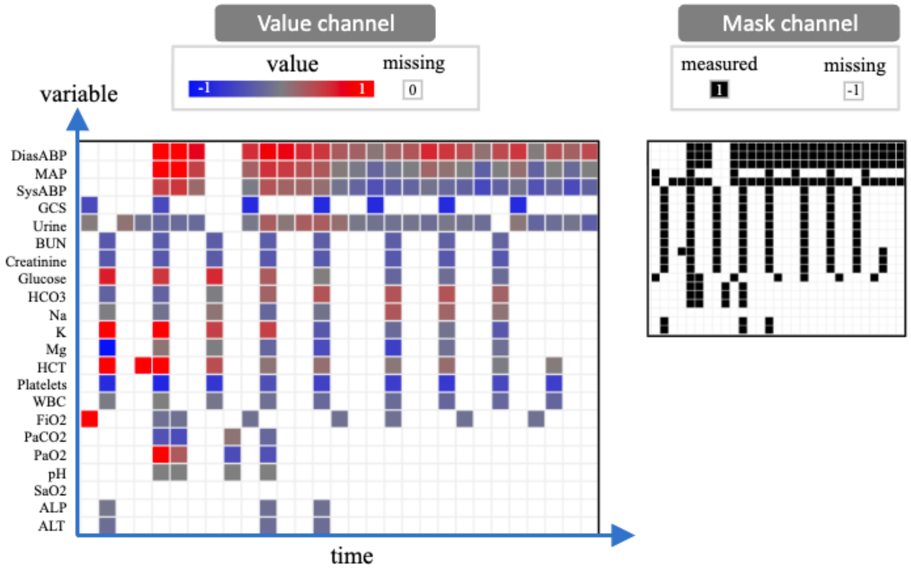
Heatmap visualizes the magnitudes of the values in matrix using color.

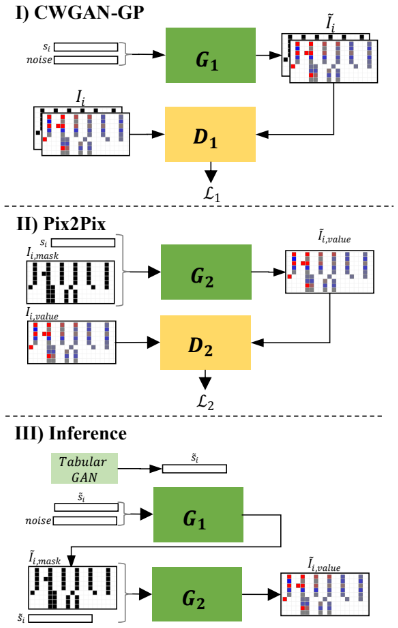
- Naturally supports MTS.



Harnessing Vision Models for Time Series Analysis: A Survey. In IJCAI, 2025.

- **Ex.1**: a GAN-based model for Electronic Health Records (EHR) time series generation, it aims to solve the Irregular sampling, missing value and high dimensional challenges.





$$\mathcal{L}_{L1} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

$$\mathcal{L}_{L2} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

TimEHR: Image-based time series generation for electronic health records. arXiv 2024.

Heatmap — e.g. MVH TSF

- **Ex.2**: Video Prediction Model for TSF

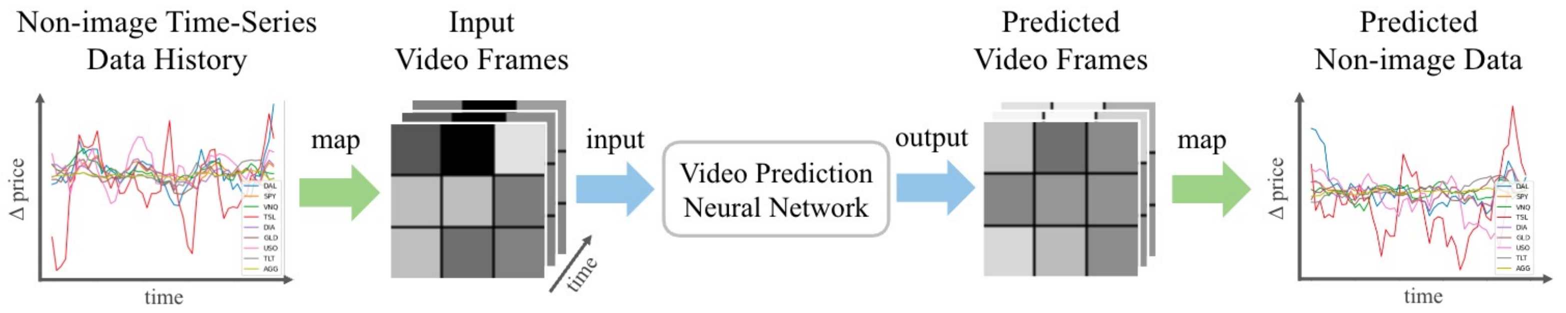


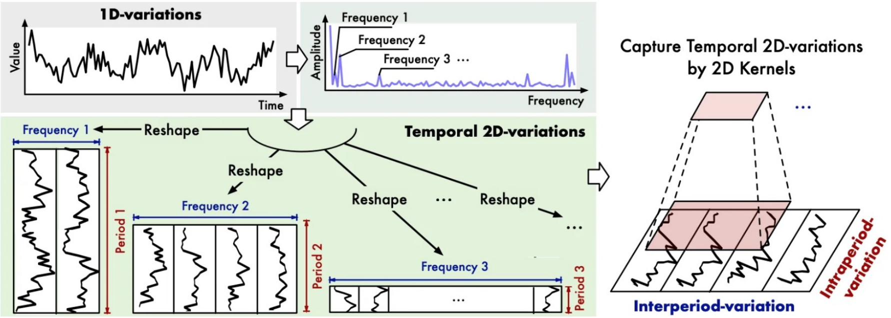
Figure 2: Method overview. First, we turn non-image time-series data history into a video frame at each time stamp. Then, we use a video prediction neural network to predict future video frames. Finally, we map the predicted video frames back to the numerical data space.

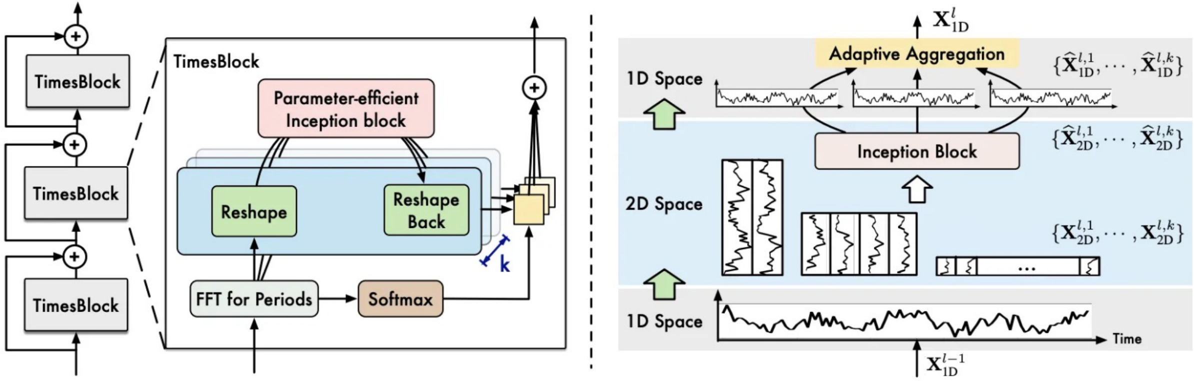
- Based on domain knowledge, variables with strong relevance are arranged spatially adjacent, facilitating the extraction of local correlation features by CNNs.

Deep video prediction for time series forecasting. In ICAIF, 2021.

Heatmap — e.g. UVH TSF (TimesNet)

- **Ex.3**: Perform periodic folding to capture inter-period and intra-period patterns via 2D CNNs.





TimesNet: Temporal 2d-variation modeling for general time series analysis. In ICLR, 2023.

Heatmap — e.g. UVH TSF (TimesNet)

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Sota Performance in Forecasting, Imputation, Classification and Anomaly Detection

Table 2: Long-term forecasting task. The past sequence length is set as 36 for ILI and 96 for the others. All the results are averaged from 4 different prediction lengths, that is $\{24, 36, 48, 60\}$ for ILI and $\{96, 192, 336, 720\}$ for the others. See Table 13 in Appendix for the full results.

Models	**TimesNet (Ours)**		ETSformer (2022)		LightTS (2022)		DLinear (2023)		FEDformer (2022)		Stationary (2022a)		Autoformer (2021)		Pyraformer (2021a)		Informer (2021)		LogTrans (2019)		Reformer (2020)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	**0.400**	**0.406**	0.429	0.425	0.435	0.437	0.403	0.407	0.448	0.452	0.481	0.456	0.588	0.517	0.691	0.607	0.961	0.734	0.929	0.725	0.799	0.671
ETTm2	**0.291**	**0.333**	0.293	0.342	0.409	0.436	0.350	0.401	0.305	0.349	0.306	0.347	0.327	0.371	1.498	0.869	1.410	0.810	1.535	0.900	1.479	0.915
ETTh1	0.458	**0.450**	0.542	0.510	0.491	0.479	0.456	0.452	**0.440**	0.460	0.570	0.537	0.496	0.487	0.827	0.703	1.040	0.795	1.072	0.837	1.029	0.805
ETTh2	**0.414**	**0.427**	0.439	0.452	0.602	0.543	0.559	0.515	0.437	0.449	0.526	0.516	0.450	0.459	0.826	0.703	4.431	1.729	2.686	1.494	6.736	2.191
Electricity	**0.192**	**0.295**	0.208	0.323	0.229	0.329	0.212	0.300	0.214	0.327	0.193	0.296	0.227	0.338	0.379	0.445	0.311	0.397	0.272	0.370	0.338	0.422
Traffic	0.620	**0.336**	0.621	0.396	0.622	0.392	0.625	0.383	**0.610**	0.376	0.624	0.340	0.628	0.379	0.878	0.469	0.764	0.416	0.705	0.395	0.741	0.422
Weather	**0.259**	**0.287**	0.271	0.334	0.261	0.312	0.265	0.317	0.309	0.360	0.288	0.314	0.338	0.382	0.946	0.717	0.634	0.548	0.696	0.602	0.803	0.656
Exchange	0.416	0.443	0.410	0.427	0.385	0.447	**0.354**	**0.414**	0.519	0.500	0.461	0.454	0.613	0.539	1.913	1.159	1.550	0.998	1.402	0.968	1.280	0.932
ILI	2.139	0.931	2.497	1.004	7.382	2.003	2.616	1.090	2.847	1.144	**2.077**	**0.914**	3.006	1.161	7.635	2.050	5.137	1.544	4.839	1.485	4.724	1.445

Table 3: Short-term forecasting task on M4. The prediction lengths are in $[6, 48]$ and results are weighted averaged from several datasets under different sample intervals. See Table 14 for full results.

Models	**TimesNet (Ours)**	N-HiTS (2022)	N-BEATS (2019)	ETSformer (2022)	LightTS (2022)	DLinear (2023)	FEDformer (2022)	Stationary (2022a)	Autoformer (2021)	Pyraformer (2021a)	Informer (2021)	LogTrans (2019)	Reformer (2020)
SMAPE	**11.829**	11.927	11.851	14.718	13.525	13.639	12.840	12.780	12.909	16.987	14.086	16.018	18.200
MASE	**1.585**	1.613	1.599	2.408	2.111	2.095	1.701	1.756	1.771	3.265	2.718	3.010	4.223
OWA	**0.851**	0.861	0.855	1.172	1.051	1.051	0.918	0.930	0.939	1.480	1.230	1.378	1.775

Table 4: Imputation task. We randomly mask $\{12.5\%, 25\%, 37.5\%, 50\%\}$ time points in length-96 time series. The results are averaged from 4 different mask ratios. See Table 16 for full results.

Models	**TimesNet (Ours)**		ETSformer (2022)		LightTS (2022)		DLinear (2023)		FEDformer (2022)		Stationary (2022a)		Autoformer (2021)		Pyraformer (2021a)		Informer (2021)		LogTrans (2019)		Reformer (2020)	
Mask Ratio	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	**0.027**	**0.107**	0.120	0.253	0.104	0.218	0.093	0.206	0.062	0.177	0.036	0.126	0.051	0.150	0.717	0.570	0.071	0.188	0.050	0.154	0.055	0.166
ETTm2	**0.022**	**0.088**	0.208	0.327	0.046	0.151	0.096	0.208	0.101	0.215	0.026	0.099	0.029	0.105	0.465	0.508	0.156	0.292	0.119	0.246	0.157	0.280
ETTh1	**0.078**	**0.187**	0.202	0.329	0.284	0.373	0.201	0.306	0.117	0.246	0.094	0.201	0.103	0.214	0.842	0.682	0.161	0.279	0.219	0.332	0.122	0.245
ETTh2	**0.049**	**0.146**	0.367	0.436	0.119	0.250	0.142	0.259	0.163	0.279	0.053	0.152	0.055	0.156	1.079	0.792	0.337	0.452	0.186	0.318	0.234	0.352
Electricity	**0.092**	**0.210**	0.214	0.339	0.131	0.262	0.132	0.260	0.130	0.259	0.100	0.218	0.101	0.225	0.297	0.382	0.222	0.328	0.175	0.303	0.200	0.313
Weather	**0.030**	**0.054**	0.076	0.171	0.055	0.117	0.052	0.110	0.099	0.203	0.032	0.059	0.031	0.057	0.152	0.235	0.045	0.104	0.039	0.076	0.038	0.087

Table 5: Anomaly detection task. We calculate the F1-score (as %) for each dataset. *. means the *former. A higher value of F1-score indicates a better performance. See Table 15 for full results.

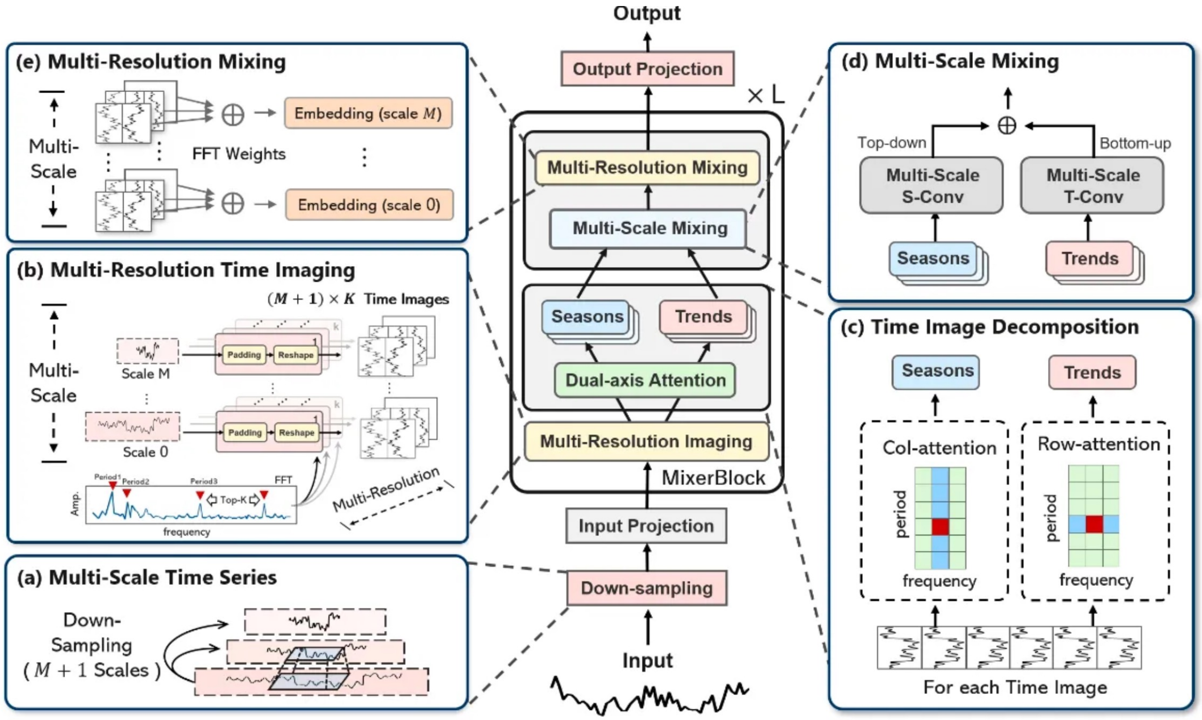
Models	**TimesNet (ResNeXt)**	**TimesNet (Inception)**	ETS. (2022)	FED. (2022)	LightTS (2022)	DLinear (2023)	Stationary (2022a)	Auto. (2021)	Pyra. (2021a)	Anomaly* (2021)	In. (2021)	Re. (2020)	LogTrans (2019)	Trans. (2017)
SMD	**85.81**	85.12	83.13	85.08	82.53	77.10	84.72	85.11	83.04	85.49	81.65	75.32	76.21	79.56
MSL	**85.15**	84.18	85.03	78.57	78.95	84.88	77.50	79.05	84.86	83.31	84.06	84.40	79.57	78.68
SMAP	**71.52**	70.85	69.50	70.76	69.21	69.26	71.09	71.12	71.09	71.18	69.92	70.40	69.97	69.70
SWaT	91.74	92.10	84.91	93.19	**93.33**	87.52	79.88	92.74	91.78	83.10	81.43	82.80	80.52	80.37
PSM	**97.47**	95.21	91.76	97.23	97.15	93.55	97.29	93.29	82.08	79.40	77.10	73.61	76.74	76.07
Avg F1	**86.34**	85.49	82.87	84.97	84.23	82.46	82.08	84.26	82.57	80.50	78.83	77.31	76.60	76.88

TimesNet: Temporal 2d-variation modeling for general time series analysis. In ICLR, 2023.

Heatmap — e.g. UVH TSF (TimeMixer++)

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- **Ex.4**: Periodic folding in frequency domain and multi-scale down-sapling in time domain.



TimeMixer++: A general time series pattern machine for universal predictive analysis. In ICLR, 2025.

Heatmap — e.g. UVH TSF (TimeMixer++)

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Sota in long/short/few/zero Forecasting, Imputation, Classification and Anomaly Detection

Table 1: Long-term forecasting results. We average the results across 4 prediction lengths: $\{96, 192, 336, 720\}$. The best performance is highlighted in **red**, and the second-best is underlined. Full results can be found in Appendix H.

Models	TimeMixer++ (Ours)		TimeMixer (2024b)		iTransformer (2024)		PatchTST (2023)		Crossformer (2023)		TiDE (2023a)		TimesNet (2023)		DLinear (2023)		SCINet (2022a)		FEDformer (2022b)		Stationary (2022c)		Autoformer (2021)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Electricity	**0.165**	**0.253**	0.182	0.272	0.178	0.270	0.205	0.290	0.244	0.334	0.251	0.344	0.192	0.295	0.212	0.300	0.268	0.365	0.214	0.327	0.193	0.296	0.227	0.338
ETT (Avg)	**0.349**	0.399	0.367	0.388	0.383	**0.377**	0.381	0.397	0.685	0.578	0.482	0.470	0.391	0.404	0.442	0.444	0.689	0.597	0.408	0.428	0.471	0.464	0.465	0.459
Exchange	0.357	0.391	0.391	0.453	0.378	**0.360**	0.403	0.404	0.940	0.707	0.370	0.413	0.416	0.443	**0.354**	0.414	0.750	0.626	0.519	0.429	0.461	0.454	0.613	0.539
Traffic	**0.416**	**0.264**	0.484	0.297	0.428	0.282	0.481	0.304	0.550	0.304	0.760	0.473	0.620	0.336	0.625	0.383	0.804	0.509	0.610	0.376	0.624	0.340	0.628	0.379
Weather	**0.226**	**0.262**	0.240	0.271	0.258	0.278	0.259	0.281	0.259	0.315	0.271	0.320	0.259	0.287	0.265	0.317	0.292	0.363	0.309	0.360	0.288	0.314	0.338	0.382
Solar-Energy	**0.203**	**0.238**	0.216	0.280	0.233	0.262	0.270	0.307	0.641	0.639	0.347	0.417	0.301	0.319	0.330	0.401	0.282	0.375	0.291	0.381	0.261	0.381	0.885	0.711

Table 2: Univariate short-term forecasting results, averaged across all M4 subsets. Full results are available in Appendix H

Models	TimeMixer++ (Ours)	TimeMixer (2024b)	iTransformer (2024)	TiDE (2023a)	TimesNet (2023)	N-HiTS (2023)	N-BEATS (2019)	PatchTST (2023)	MICN (2023a)	FiLM (2022a)	LightTS (2022a)	DLinear (2023)	FED. (2022b)	Stationary (2022c)	Auto. (2021)
SMAPE	**11.448**	11.723	12.684	13.950	11.829	11.927	11.851	13.152	19.638	14.863	13.525	13.639	12.840	12.780	12.909
MASE	**1.487**	1.559	1.764	1.940	1.585	1.613	1.559	1.945	5.947	2.207	2.111	2.095	1.701	1.756	1.771
OWA	**0.821**	0.840	0.929	1.020	0.851	0.861	0.855	0.998	2.279	1.125	1.051	1.051	0.918	0.930	0.939

Table 3: Results of multivariate short-term forecasting, averaged across all PEMS datasets. Full results can be found in Table 18 of Appendix H.

Models	TimeMixer++ (Ours)	TimeMixer (2024b)	iTransformer (2024)	TiDE (2023a)	SCINet (2022a)	Crossformer (2023)	PatchTST (2023)	TimesNet (2023)	MICN (2023a)	DLinear (2023)	FEDformer (2022b)	Stationary (2022c)	Autoformer (2021)
MAE	**15.91**	17.41	19.87	21.86	19.12	19.03	23.01	20.54	19.34	23.31	23.50	21.32	22.62
MAPE	**10.08**	10.59	12.55	13.80	12.24	12.22	14.95	12.69	12.38	14.68	15.01	14.09	14.89
RMSE	**27.06**	28.01	31.29	34.42	30.12	30.17	36.05	33.25	30.40	37.32	36.78	36.20	34.49

Table 5: Few-shot learning on 10% training data. All results are averaged from 4 prediction lengths: $\{96, 192, 336, 720\}$.

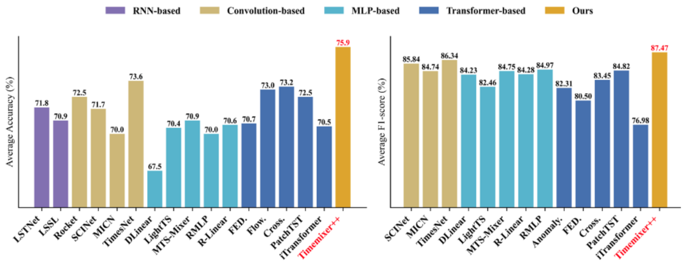
Models	TimeMixer++ (Ours)		TimeMixer (2024b)		iTransformer (2024)		TiDE (2023a)		Crossformer (2023)		DLinear (2023)		PatchTST (2023)		TimesNet (2023)		FEDformer (2022b)		Autoformer (2021)		Stationary (2022c)		ETSformer (2022)		LightTS (2022b)		Informer (2021b)		Reformer (2020)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETT(Avg)	**0.396**	**0.421**	0.453	0.445	0.458	0.497	0.432	0.444	0.470	0.471	0.506	0.484	0.461	0.446	0.586	0.496	0.573	0.532	0.834	0.663	0.627	0.510	0.875	0.687	1.497	0.875	2.408	1.146	2.535	1.191
Weather	**0.241**	**0.271**	0.242	0.281	0.291	0.331	0.249	0.291	0.267	0.306	0.241	0.283	0.242	0.279	0.279	0.301	0.284	0.324	0.300	0.342	0.318	0.323	0.318	0.360	0.289	0.322	0.597	0.495	0.546	0.469
ECL	**0.168**	**0.271**	0.187	0.277	0.241	0.337	0.196	0.289	0.214	0.308	0.180	0.280	0.180	0.273	0.323	0.392	0.346	0.427	0.431	0.478	0.444	0.480	0.660	0.617	0.441	0.489	1.195	0.891	0.965	0.768

Table 4: Results of imputation task across six datasets. To evaluate our model performance, we randomly mask $\{12.5\%, 25\%, 37.5\%, 50\%\}$ of the time points in time series of length 1024. The final results are averaged across these 4 different masking ratios.

Models	TimeMixer++ (Ours)		TimeMixer (2024b)		iTransformer (2024)		PatchTST (2023)		Crossformer (2023)		FEDformer (2022b)		TIDE (2023a)		DLinear (2023)		TimesNet (2023)		MICN (2023a)		Autoformer (2021)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETT(Avg)	**0.055**	**0.154**	0.097	0.220	0.096	0.205	0.120	0.225	0.150	0.258	0.124	0.230	0.314	0.366	0.115	0.229	0.079	0.182	0.119	0.234	0.104	0.215
ECL	**0.109**	**0.197**	0.142	0.261	0.140	0.223	0.129	0.198	0.125	0.204	0.181	0.314	0.182	0.202	0.080	0.200	0.135	0.255	0.138	0.246	0.141	0.234
Weather	**0.049**	**0.078**	0.091	0.114	0.095	0.102	0.082	0.149	0.150	0.111	0.064	0.139	0.063	0.131	0.071	0.107	0.061	0.098	0.075	0.126	0.066	0.107

Table 6: Zero-shot learning results. The results are averaged from 4 different prediction lengths: $\{96, 192, 336, 720\}$.

Methods	TimeMixer++ (Ours)		TimeMixer (2024b)		LLMTime (2023)		DLinear (2023)		PatchTST (2023)		TimesNet (2023)		iTransformer (2024)		Crossformer (2023)		Fedformer (2022b)		Autoformer (2021)		TiDE (2023a)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
$ETTh1 \rightarrow ETTh2$	**0.367**	**0.391**	0.427	0.424	0.992	0.708	0.493	0.488	0.380	0.405	0.421	0.431	0.481	0.474	0.555	0.574	0.712	0.693	0.634	0.651	0.593	0.582
$ETTh1 \rightarrow ETTm2$	**0.301**	**0.357**	0.361	0.397	1.867	0.869	0.415	0.452	0.314	0.360	0.327	0.361	0.311	0.361	0.613	0.629	0.681	0.588	0.647	0.609	0.563	0.547
$ETTh2 \rightarrow ETTh1$	**0.511**	**0.498**	0.679	0.577	1.961	0.981	0.703	0.574	0.565	0.513	0.865	0.621	0.552	0.511	0.587	0.518	0.612	0.624	0.599	0.571	0.588	0.556
$ETTm1 \rightarrow ETTh2$	**0.417**	**0.422**	0.452	0.441	0.992	0.708	0.464	0.475	0.439	0.438	0.457	0.454	0.434	0.438	0.624	0.541	0.533	0.594	0.579	0.568	0.543	0.535
$ETTm1 \rightarrow ETTm2$	**0.291**	**0.331**	0.329	0.357	1.867	0.869	0.335	0.389	0.296	0.334	0.322	0.354	0.324	**0.331**	0.595	0.572	0.612	0.611	0.603	0.592	0.534	0.527
$ETTm2 \rightarrow ETTm1$	**0.427**	**0.448**	0.554	0.478	1.933	0.984	0.649	0.537	0.568	0.492	0.769	0.567	0.559	0.491	0.611	0.593	0.577	0.601	0.594	0.597	0.585	0.571



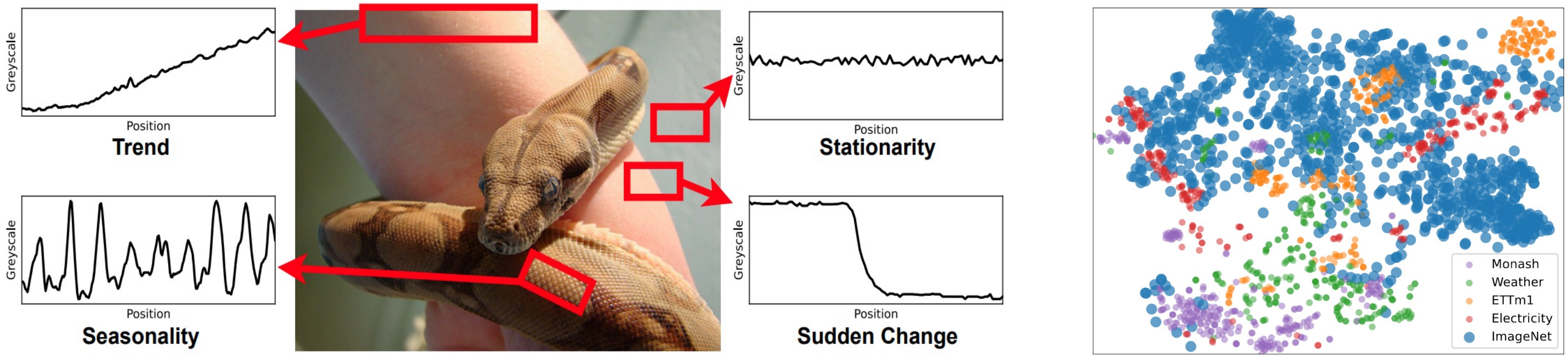
(a) Classification Results
(b) Anomaly Detection Results

TimesNet: Temporal 2d-variation modeling for general time series analysis. In ICLR, 2023.

Heatmap — e.g. UVH TSF (VisionTS)

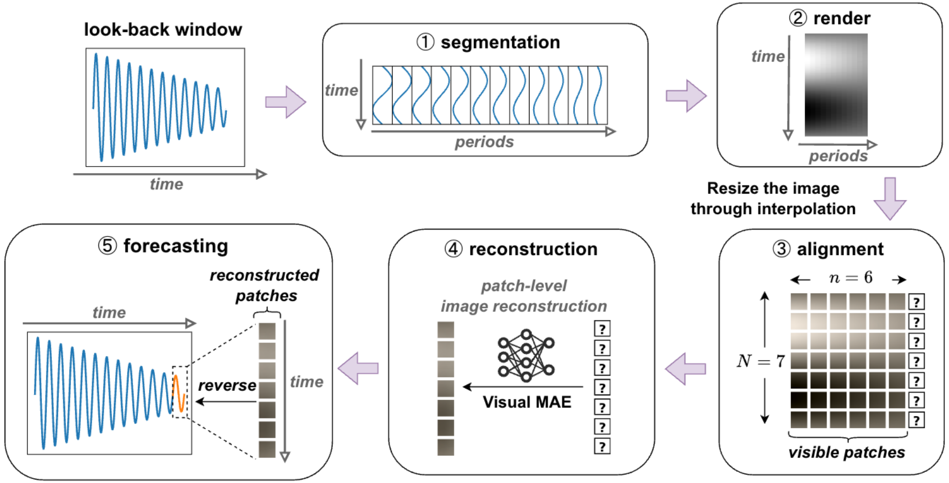
- **Ex.5**: Also adopts the periodic folding imaging, leveraging vision models for TSF.
- VisionTS reformulates TSF into an image reconstruction task via MAE.

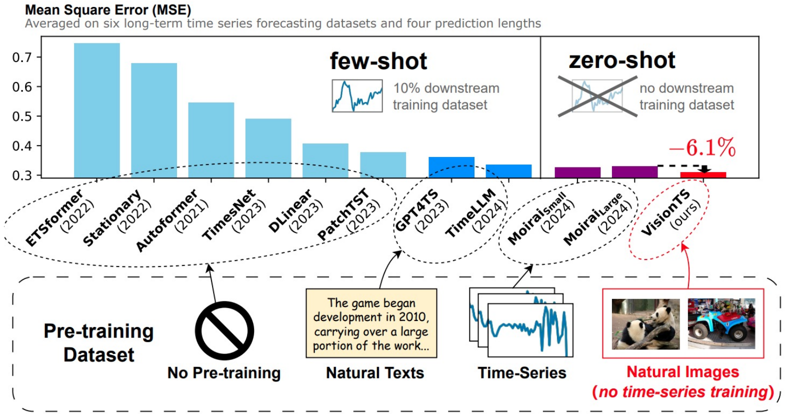
	Characteristics	Origin	Information
Time series	continuous	physical systems	high redundancy
Image	continuous	physical systems	high redundancy
Text	discrete	human cognitive construct	semantically dense



VisionTS: Visual masked autoencoders are free-lunch zero-shot time series forecasters. In ICML, 2025.

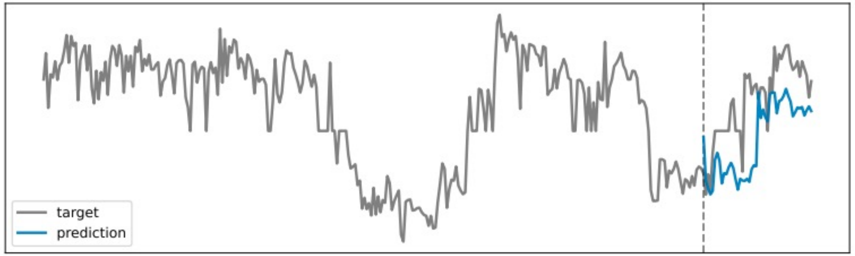
Heatmap — e.g. UVH TSF (VisionTS)





(a) Input Image

(b) Reconstructed Image



(c) VisionTS (MAE = 0.821)

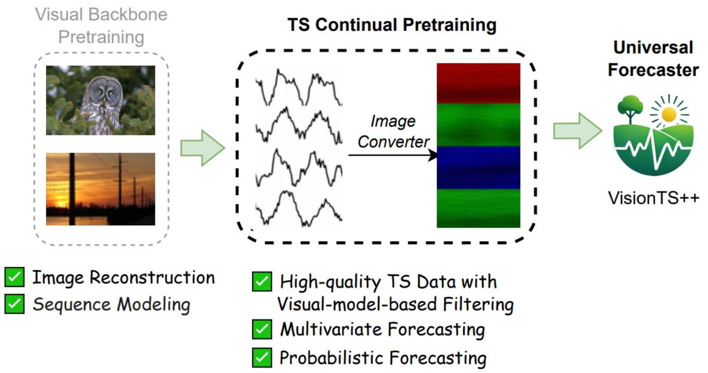
VisionTS: Visual masked autoencoders are free-lunch zero-shot time series forecasters. In ICML, 2025.

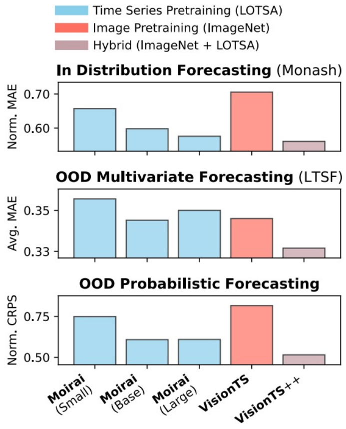
Heatmap — e.g. MVH TSFM (VisionTS++)

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Three gaps between TS and Image:

	Modality Gap	Dimensional Gap	Probabilistic-forecasting Gap
TS	unbounded, heterogeneous	arbitrary numbers of variates	need uncertainty-aware probabilistic predictions
Image	structured, bounded	3 channels (RGB)	deterministic output of most vision models

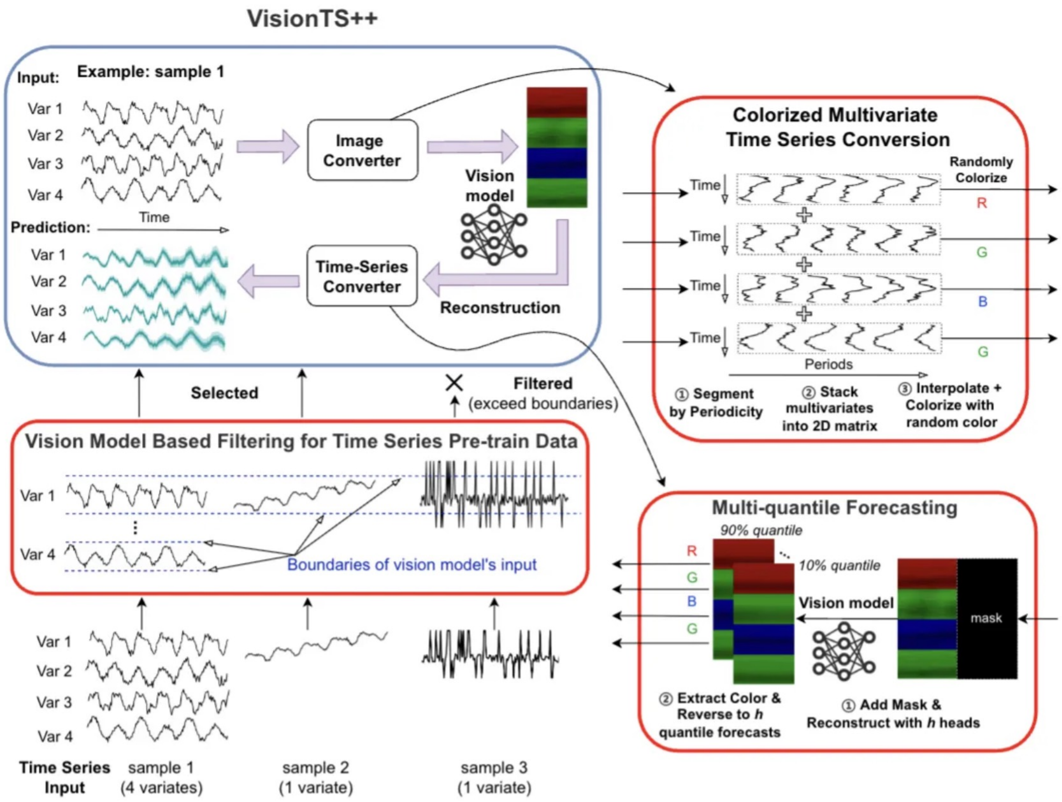




VisionTS++: Cross-Modal Time Series Foundation Model with Continual Pre-trained Visual Backbones. arxiv 2025.

Heatmap — e.g. MVH TSFM (VisionTS++)

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)



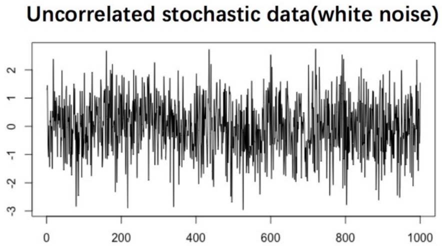
VisionTS++: Cross-Modal Time Series Foundation Model with Continual Pre-trained Visual Backbones. arxiv 2025.

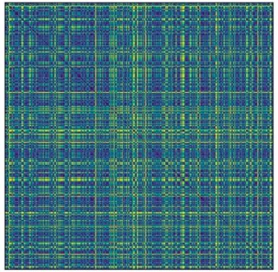
RP — Intro

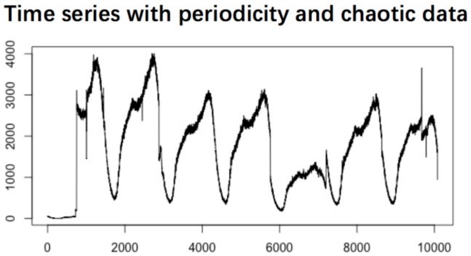
RP (Recurrent Plot) capture periodicity, chaos, and other dynamic patterns of the sequence.

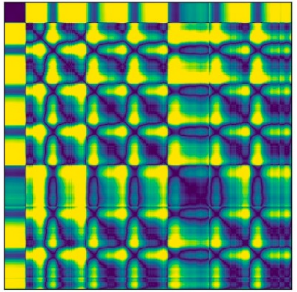
$\mathbf{x} \in \mathbb{R}^{1\times T} \quad \mathbf{v}_t = [x_t, x_{t+\tau}, x_{t+2\tau}, \ldots, x_{t+(m-1)\tau}] \in \mathbb{R}^{m\tau}, \quad 1 \le t \le l$

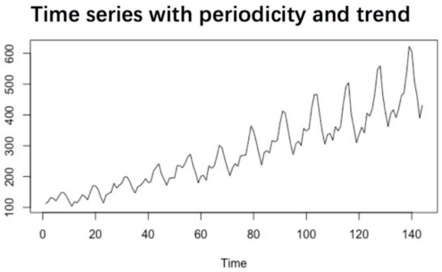
$\mathrm{RP}_{i,j} = \Theta(\varepsilon - \|\mathbf{v}_i - \mathbf{v}_j\|), \quad 1 \le i,j \le l$

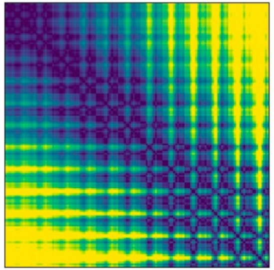








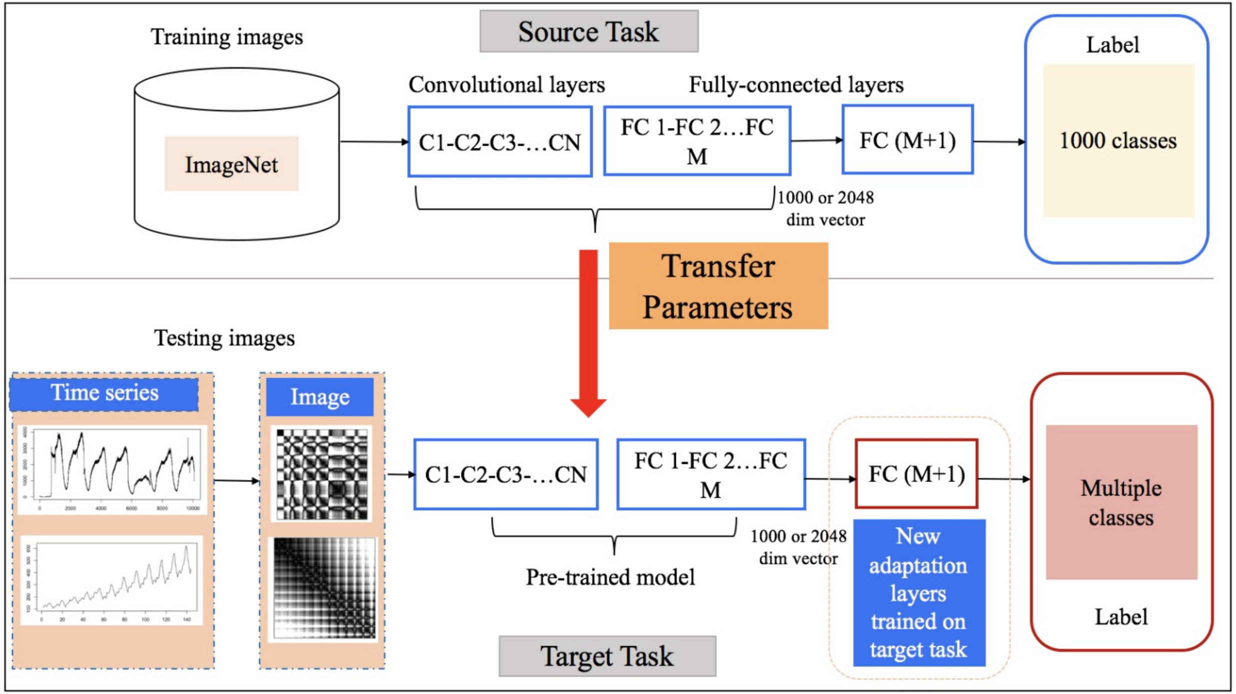




Forecasting with time series imaging. Expert Syst. Appl., 2020.

RP — Intro

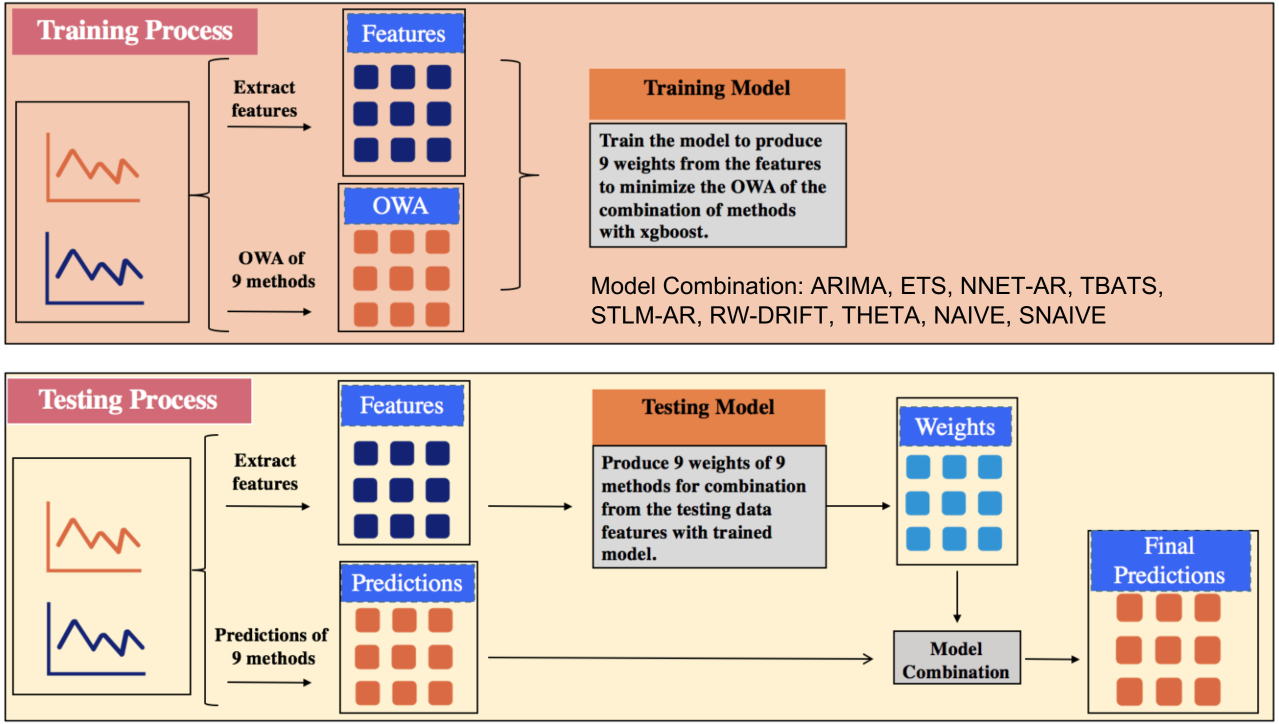
- **Ex.1**: Vision Models for TS Classification.



Forecasting with time series imaging. Expert Syst. Appl., 2020.

RP — Intro

- **Ex.2**: Vision Models for TS Forecasting.

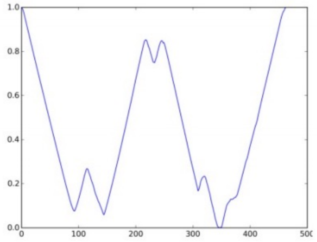


Forecasting with time series imaging. Expert Syst. Appl., 2020.

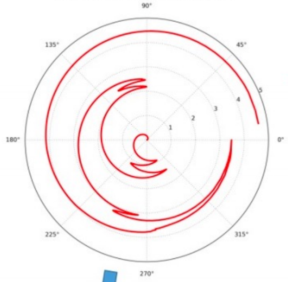
GAF — Intro

GAF (Gramian Angular Field) encodes the correlation of time series at different time steps.

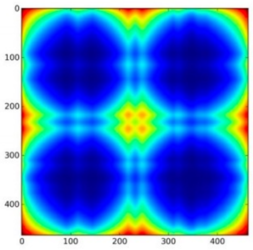
Time Series x



Polar Coordinate



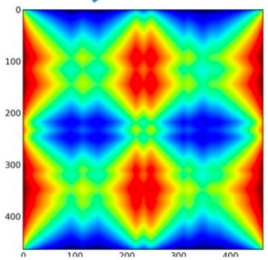
$$\begin{cases} \phi = \arccos(\tilde{x}_i), -1 \le \tilde{x}_i \le 1, \tilde{x}_i \in \tilde{X} \\ r = \frac{t_i}{N}, t_i \in \mathbb{N} \end{cases}$$



GASF

Gramian Angular **Summation** Field

$$\boxed{\text{GASF: } \cos(\phi_t + \phi_{t'}) = x_t x_{t'} - \sqrt{1 - x_t^2}\sqrt{1 - x_{t'}^2}}$$



GADF

Gramian Angular **Difference** Field

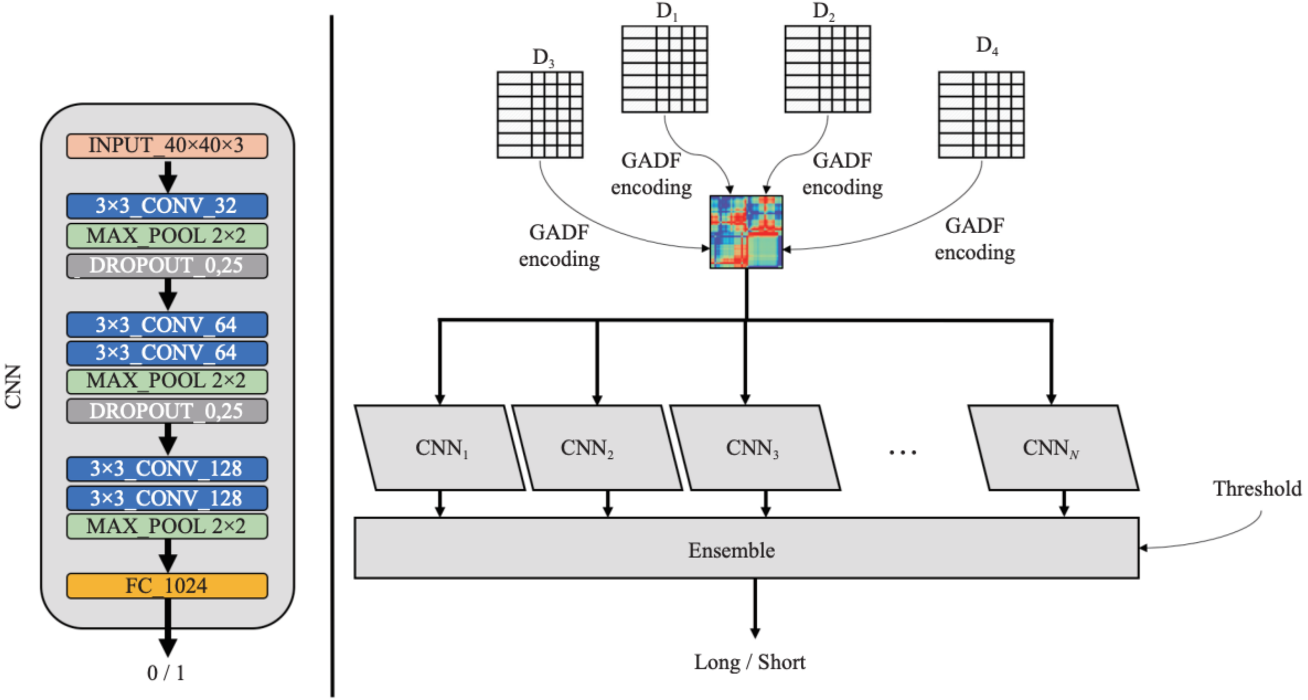
$$\boxed{\text{GADF: } \sin(\phi_t - \phi_{t'}) = x_{t'}\sqrt{1 - x_t^2} - x_t\sqrt{1 - x_{t'}^2}}$$

Encoding time series as images for visual inspection and classification using tiled convolutional neural networks. In AAAI Workshop, 2015

GAF — e.g. TSF

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- **Ex.1**: CNNs for TS Anomaly Detection

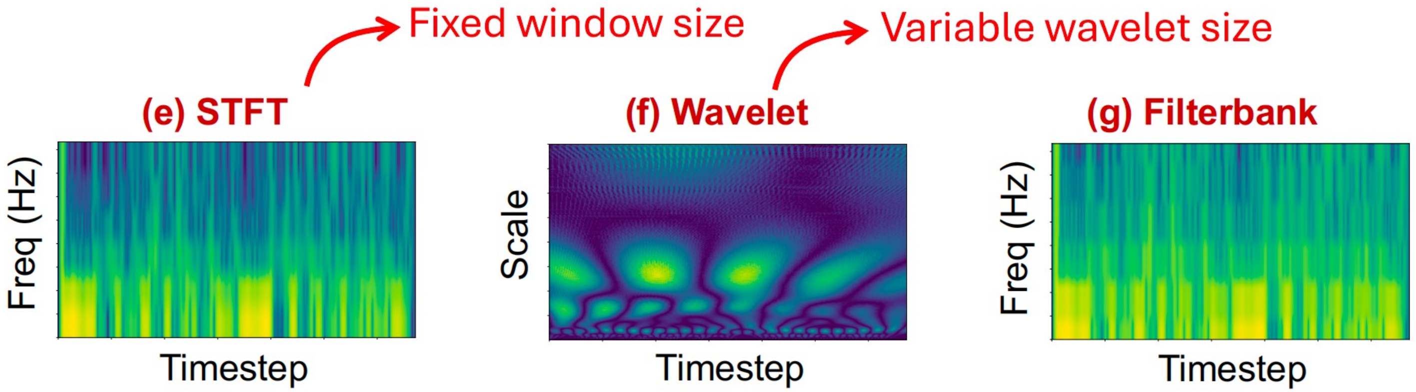


Deep learning and time series-to-image encoding for financial forecasting. IEEE/CAA J. Autom. Sin., 2020.

Spectrogram — Intro

Spectrogram is a visual representation of frequencies of a signal as it varies with time.

- Extensively used for audio signals analysis, type of UTS.

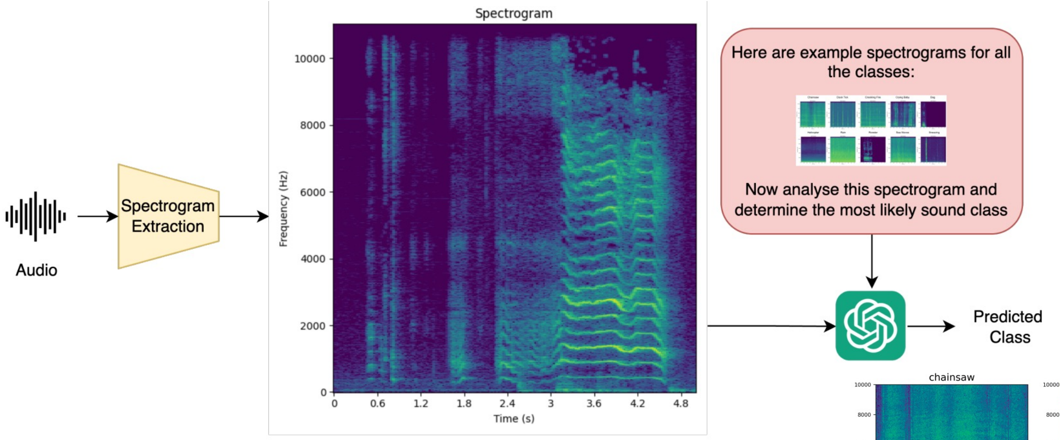


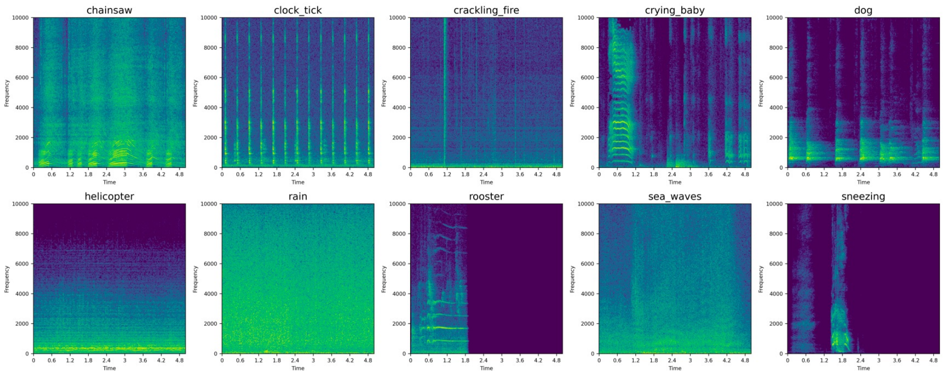
"Signal estimation from modified short-time fourier transform." IEEE Trans. Acoust., 1984.
"The wavelet transform, time-frequency localization and signal analysis." IEEE Trans. Inf. Theory, 1990.
"Wavelets and filter banks: Theory and design." IEEE Trans. Signal Process., 1992.

Spectrogram — e.g. TSC

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- **Ex.1**: VLMs for Few-shot Audio Spectrogram Classification.

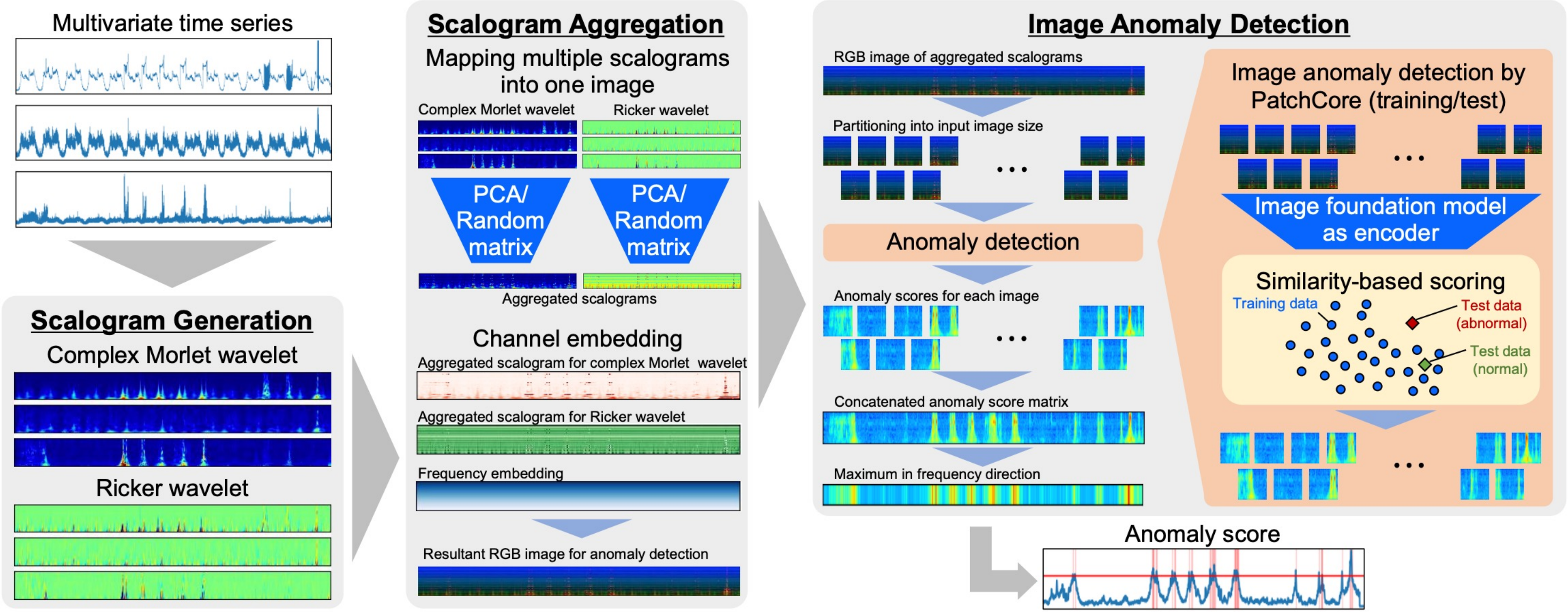




Vision language models are few-shot audio spectrogram classifiers. In NeurIPS Workshop, 2024.

Spectrogram — e.g. TSAD

- **Ex.2**: Vision Models for TS Anomaly Detection



Training-free time-series anomaly detection: Leveraging image foundation models. arXiv 2024.

TS2Vision Summary

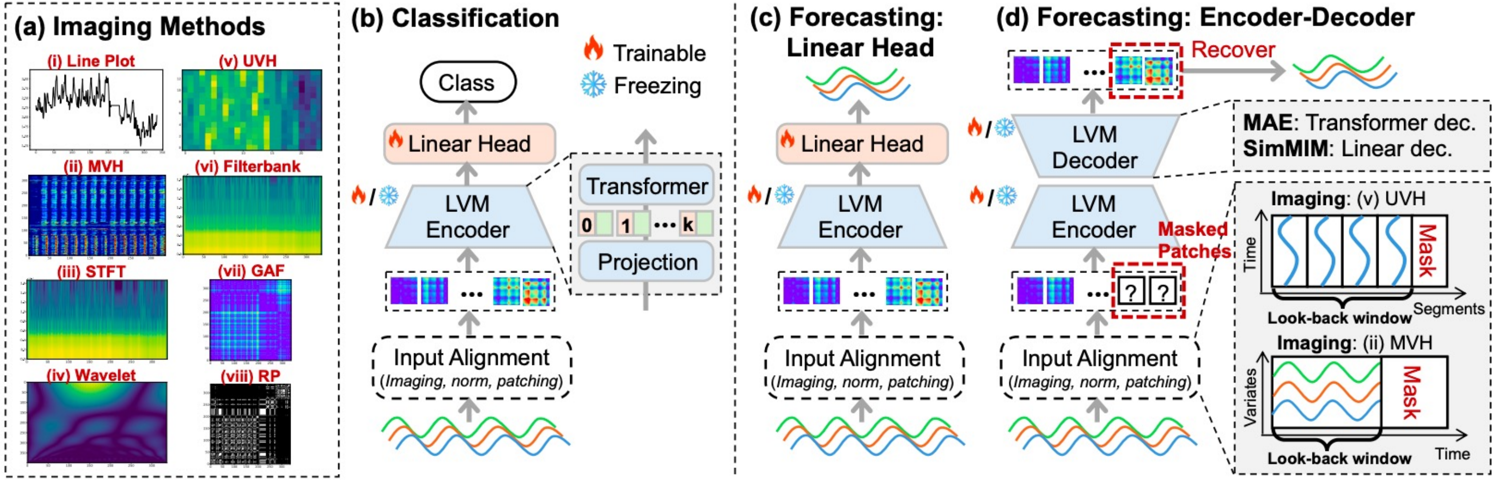
Method	TS-Type	Advantages	Limitations
Line Plot (§3.1)	UTS, MTS	matches human perception of time series	limited to MTS with a small number of variates
Heatmap (§3.2)	UTS, MTS	straightforward for both UTS and MTS	the order of variates may affect their correlation learning
Spectrogram (§3.3)	UTS	encodes the time-frequency space	limited to UTS; needs a proper choice of window/wavelet
GAF (§3.4)	UTS	encodes the temporal correlations in a UTS	limited to UTS; $O(T^2)$ time and space complexity
RP (§3.5)	UTS	flexibility in image size by tuning m and τ	limited to UTS; information loss after thresholding

Method	TS-Type	Imaging	Imaged Time Series Modeling					TS-Recover	Task	Domain	Code
			Multimodal	Model	Pre-trained	Fine-tune	Prompt				
[Silva *et al.*, 2013]	UTS	RP	✗	`K-NN`	✗	✗	✗	✗	Classification	General	✗
[Wang and Oates, 2015a]	UTS	GAF	✗	`CNN`	✗	✓$^\flat$	✗	✓	Classification	General	✗
[Wang and Oates, 2015b]	UTS	GAF	✗	`CNN`	✗	✓$^\flat$	✗	✓	Multiple	General	✗
[Ma *et al.*, 2017]	MTS	Heatmap	✗	`CNN`	✗	✓$^\flat$	✗	✓	Forecasting	Traffic	✗
[Hatami *et al.*, 2018]	UTS	RP	✗	`CNN`	✗	✓$^\flat$	✗	✗	Classification	General	✗
[Yazdanbakhsh and Dick, 2019]	MTS	Heatmap	✗	`CNN`	✗	✓$^\flat$	✗	✗	Classification	General	✓[1]
MSCRED [Zhang *et al.*, 2019]	MTS	Other (§3.6)	✗	`ConvLSTM`	✗	✓$^\flat$	✗	✗	Anomaly	General	✓[2]
[Li *et al.*, 2020]	UTS	RP	✗	`CNN`	✓	✓	✗	✗	Forecasting	General	✓[3]
[Cohen *et al.*, 2020]	UTS	LinePlot	✗	`Ensemble`	✗	✓$^\flat$	✗	✗	Classification	Finance	✗
[Barra *et al.*, 2020]	UTS	GAF	✗	`CNN`	✗	✓$^\flat$	✗	✗	Classification	Finance	✗
VisualAE [Sood *et al.*, 2021]	UTS	LinePlot	✗	`CNN`	✗	✓$^\flat$	✗	✓	Forecasting	Finance	✗
[Zeng *et al.*, 2021]	MTS	Heatmap	✗	`CNN,LSTM`	✗	✓$^\flat$	✗	✓	Forecasting	Finance	✗
AST [Gong *et al.*, 2021]	UTS	Spectrogram	✗	`DeiT`	✓	✓	✗	✗	Classification	Audio	✓[4]
TTS-GAN [Li *et al.*, 2022]	MTS	Heatmap	✗	`ViT`	✗	✓$^\flat$	✗	✓	Ts-Generation	Health	✓[5]
SSAST [Gong *et al.*, 2022]	UTS	Spectrogram	✗	`ViT`	✓$^\natural$	✓	✗	✗	Classification	Audio	✓[6]
MAE-AST [Baade *et al.*, 2022]	UTS	Spectrogram	✗	`MAE`	✓$^\natural$	✓	✗	✗	Classification	Audio	✓[7]
AST-SED [Li *et al.*, 2023a]	UTS	Spectrogram	✗	`SSAST,GRU`	✓	✓	✗	✗	EventDetection	Audio	✗
ForCNN [Semenoglou *et al.*, 2023]	UTS	LinePlot	✗	`CNN`	✗	✓$^\flat$	✗	✗	Forecasting	General	✗
Vit-num-spec [Zeng *et al.*, 2023]	UTS	Spectrogram	✗	`ViT`	✗	✓$^\flat$	✗	✗	Forecasting	Finance	✗
ViTST [Li *et al.*, 2023b]	MTS	LinePlot	✗	`Swin`	✓	✓	✗	✗	Classification	General	✓[8]
MV-DTSA [Yang *et al.*, 2023]	UTS*	LinePlot	✗	`CNN`	✗	✓$^\flat$	✗	✓	Forecasting	General	✓[9]
TimesNet [Wu *et al.*, 2023]	MTS	Heatmap	✗	`CNN`	✗	✓$^\flat$	✗	✓	Multiple	General	✓[10]
ITF-TAD [Namura *et al.*, 2024]	UTS	Spectrogram	✗	`CNN`	✓	✗	✗	✗	Anomaly	General	✗
[Kaewrakmuk *et al.*, 2024]	UTS	GAF	✗	`CNN`	✓	✓	✗	✗	Classification	Sensing	✗
HCR-AdaAD [Lin *et al.*, 2024]	MTS	RP	✗	`CNN,GNN`	✗	✓$^\flat$	✗	✗	Anomaly	General	✗
FIRTS [Costa *et al.*, 2024]	UTS	Other (§3.6)	✗	`CNN`	✗	✓$^\flat$	✗	✗	Classification	General	✓[11]
CAFO [Kim *et al.*, 2024]	MTS	RP	✗	`CNN,ViT`	✗	✓$^\flat$	✗	✗	Explanation	General	✓[12]
ViTime [Yang *et al.*, 2024]	UTS*	LinePlot	✗	`ViT`	✓$^\natural$	✓	✗	✓	Forecasting	General	✓[13]
ImagenTime [Naiman *et al.*, 2024]	MTS	Other (§3.6)	✗	`CNN`	✗	✓$^\flat$	✗	✓	Ts-Generation	General	✓[14]
TimEHR [Karami *et al.*, 2024]	MTS	Heatmap	✗	`CNN`	✗	✓$^\flat$	✗	✓	Ts-Generation	Health	✓[15]
VisionTS [Chen *et al.*, 2024]	UTS*	Heatmap	✗	`MAE`	✓	✓	✗	✓	Forecasting	General	✓[16]
TimeMixer++ [Wang *et al.*, 2025]	MTS	Heatmap	✗	`CNN`	✗	✓$^\flat$	✗	✓	Multiple	General	✓[17]
InsightMiner [Zhang *et al.*, 2023]	UTS	LinePlot	✓	`LLaVA`	✓	✓	✓	✗	Txt-Generation	General	✗
[Wimmer and Rekabsaz, 2023]	MTS	LinePlot	✓	`CLIP,LSTM`	✓	✓	✗	✗	Classification	Finance	✗
[Dixit *et al.*, 2024]	UTS	Spectrogram	✓	`GPT4o,Gemini` & `Claude3`	✓	✗	✓	✗	Classification	Audio	✗
[Daswani *et al.*, 2024]	MTS	LinePlot	✓	`GPT4o,Gemini`	✓	✗	✓	✗	Multiple	General	✗
TAMA [Zhuang *et al.*, 2024]	UTS	LinePlot	✓	`GPT4o`	✓	✗	✓	✗	Anomaly	General	✗
[Prithyani *et al.*, 2024]	MTS	LinePlot	✓	`LLaVA`	✓	✓	✓	✗	Classification	General	✓[18]

1. More research focuses on Line Plot and Heatmap, as they support MTS, more common in reality.
2. TS2Vision enables a wide range of tasks, mainly classification, forecasting, anomaly detection.

Are LVM useful for TSF?

- Select two supervised (ViT/Swin) and two self-supervised pre-trained LVMs (MAE/SimMIM).
- Employ 8 common time series visualization methods.
- Analyze the effects on 10 TSC datasets and 8 TSF datasets.



From Images to Signals: Are Large Vision Models Useful for Time Series Analysis?. arxiv 2025.

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Comparison results between LVM and non-LVM methods.

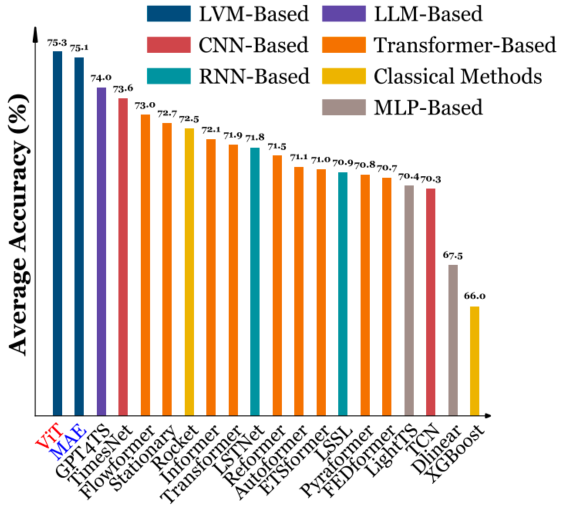


Figure 2: Model comparison in TSC. The results are averaged over 10 UEA datasets. See Table 9 in Appendix B.1 for full results.

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	

Table 2: Model comparison in TSF. The results are averaged over different prediction lengths. See Table 11 in Appendix B.2 for full results. Red and Blue numbers are the the best and second best results. # Wins is the number of times the method performed best.

Pre-trained LVMs are useful in TSC !

But pose challenges when used for TSF !

From Images to Signals: Are Large Vision Models Useful for Time Series Analysis?. arxiv 2025.

Are LVM useful for TSF?

- **RQ1**: What type of LVM best fits TSC (TSF) task?

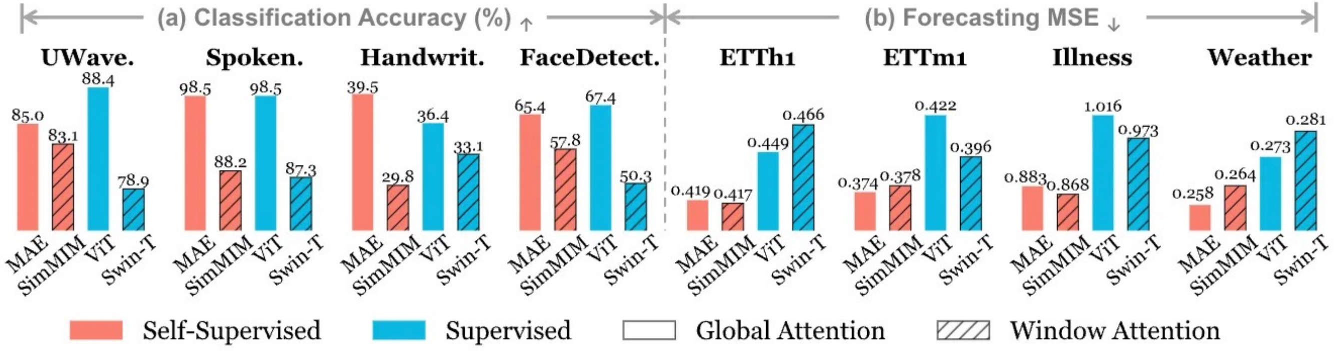


Figure 3: Comparison of 4 LVMs on TSC (accuracy) and TSF (MSE). ↑ (↓) indicates a higher (lower) value is better. Two taxonomies of the LVMs: (1) supervised (ViT, Swin) *vs.* self-supervised (MAE, SimMIM), (2) using global attention (ViT, MAE) *vs.* window-based attention (Swin, SimMIM).

Self-supervised LVM outperforms supervised LVM in TSF !

Global attention is more suitable for time series than window attention !

From Images to Signals: Are Large Vision Models Useful for Time Series Analysis?. arxiv 2025.

Are LVM useful for TSF?

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- **RQ2**: Which imaging method best fits TSC (TSF) task?

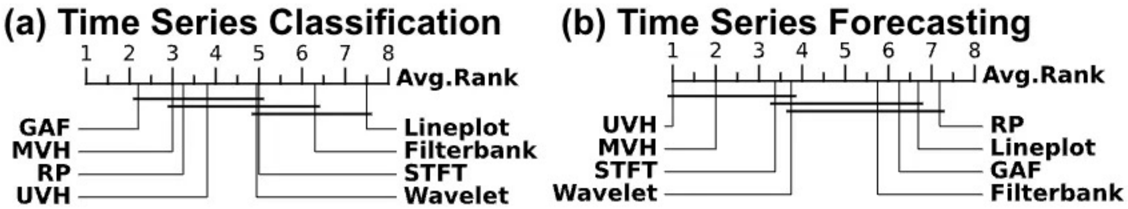


Figure 4: Average rank of different imaging methods in (a) TSC task, and (b) TSF task. Lower rank is better.

In TSC, GAF and MVH have the best effect !

In TSF, heatmap are more suitable for reconstruction frameworks due to retaining the original values !

From Images to Signals: Are Large Vision Models Useful for Time Series Analysis?. arxiv 2025.

Are LVM useful for TSF?

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- **RQ3**: Are the pre-trained parameters in LVMs useful in time series tasks?
- **RQ4**: How useful are LVMs' architectures

Task		TSC Task (accuracy (%)$_\uparrow$)				TSF Task (MSE$_\downarrow$)			
Dataset		UWave.	Spoken.	Handwrit.	FaceDetect.	ETTh1	ETTm1	Illiness	Weather
RQ3	(a) All parameters	88.4	**98.5**	**36.4**	**67.4**	0.558	0.399	1.781	0.273
	(b) All but CLS & Mask	87.5	98.2	35.2	66.3	0.530	0.408	1.783	0.275
	(c) MLP & norm	**88.7**	98.4	35.5	67.1	0.532	0.396	1.737	0.264
	(d) Norm	81.6	98.0	28.5	65.2	**0.409**	**0.345**	1.837	**0.225**
	(e) Zero-shot	84.0	**98.5**	27.8	66.7	0.452	0.420	2.037	0.308
	(f) Train from scratch	73.4	97.0	24.3	65.0	0.475	0.372	**1.723**	0.241
RQ4	w/o-LVM	78.6	96.4	22.4	64.1	0.423	0.376	2.291	0.255
	LVM2ATTN	80.1	96.5	20.7	66.2	0.428	0.357	2.108	0.254

Table 3: Ablation analysis of LVMs. For classification, higher accuracy indicates better performance. For forecasting, lower MSE is preferred. Full results are in Appendices B.5 and B.6.

Fine-tuning all parameters in TSF is best, only fine-tuning the norm layer in TSF can improve performance !

LVM may two complicated for TS, but its pretrained knowledge is useful !

From Images to Signals: Are Large Vision Models Useful for Time Series Analysis?. arxiv 2025.

Are LVM useful for TSF?

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- **RQ5**: Do LVMs capture temporal order of time series?

Task		Classification				Forecasting			
Dataset		UWave.	Spoken.	Handwrit.	FaceDetect.	ETTh1	ETTm1	Illiness	Weather
Sf-All	w/o-LVM	78.2%	49.7%	81.7%	19.3%	76.2%	98.4%	116.4%	24.1%
	LVM2Attn	86.4%	50.6%	89.9%	22.4%	79.7%	117.1%	109.1%	24.4%
	LVM	80.7%	84.7%	91.5%	29.2%	83.8%	118.4%	162.8%	44.5%
Sf-Half	w/o-LVM	6.6%	12.4%	74.6%	10.8%	14.4%	28.3%	41.6%	2.4%
	LVM2Attn	8.7%	11.6%	83.6%	11.3%	19.5%	44.8%	69.3%	2.4%
	LVM	36.4%	30.2%	86.5%	9.3%	14.5%	48.2%	21.3%	9.6%
Ex-Half	w/o-LVM	98.8%	82.2%	83.5%	22.8%	13.0%	145.3%	11.0%	34.0%
	LVM2Attn	98.9%	82.3%	87.0%	24.6%	9.1%	158.3%	27.9%	35.5%
	LVM	59.4%	89.9%	97.0%	9.2%	14.2%	242.3%	23.0%	67.2%
Masking	w/o-LVM	-1.0%	3.1%	22.3%	-1.2%	47.3%	58.5%	94.1%	33.4%
	LVM2Attn	1.0%	3.6%	20.3%	2.7%	46.0%	70.3%	127.8%	33.6%
	LVM	29.0%	41.8%	56.0%	7.4%	47.5%	58.4%	128.9%	49.6%

Table 4: Performance drop of the compared models under different temporal perturbations. Red color marks the largest drop for each perturbation strategy. Full results are in Appendix B.7.

LVM is sensitive to time disturbance, proving its effective utilization of temporal patterns.

From Images to Signals: Are Large Vision Models Useful for Time Series Analysis?. arxiv 2025.

Are LVM useful for TSF?

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- **RQ6**: What are the computational costs of LVMs?

Method		LVM			1st Baseline (task specific)			2nd Baseline (task specific)		
Task	Dataset	# Param (M)	Train (min)	Inference(ms)	# Param (M)	Train (min)	Inference(ms)	# Param (M)	Time (min)	Inference(ms)
TSC	UWave.	89.43	2.83	11.52	82.23	1.19	57.61	2.42	0.39	1.69
	Handwrit.	97.59	5.18	23.72	83.62	1.33	50.51	2.47	0.51	0.78
TSF	ETTh1	111.91	9.99	4.32	3.75	0.52	0.18	85.02	10.46	0.50
	Weather	111.91	207.83	1.50	6.90	16.97	0.10	86.64	94.10	0.35

Table 5: Computational costs of LVMs and two best baselines in TSC (GPT4TS, TimesNet) and TSF (PatchTST, GPT4TS). The forecasting costs are measured with prediction length 96.

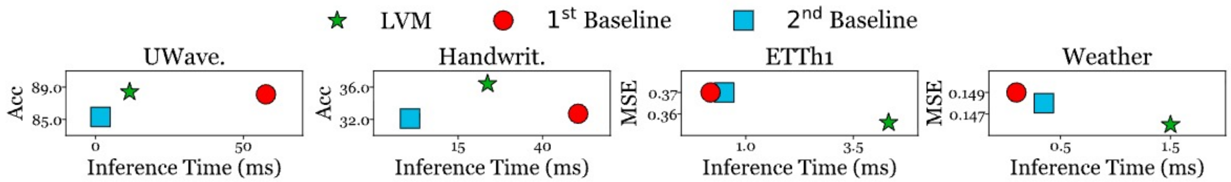


Figure 6: Inference time *vs.* performance of compared methods on TSC (accuracy) using UWaveGesture, SpokenArabicDigits, and TSF (MSE) using ETTh1, Weather. Full results are in Appendix B.8.

Although the calculation cost is higher, it has potential.

From Images to Signals: Are Large Vision Models Useful for Time Series Analysis?. arxiv 2025.

Are LVM useful for TSF?

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- **RQ7**: Which component of LVMs contributes more to forecasting?

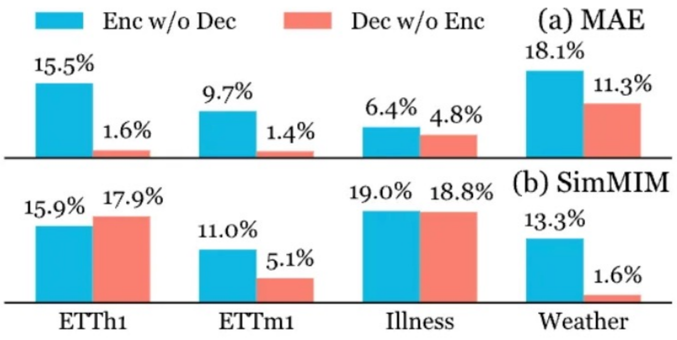


Figure 7: Forecasting performance drop (%) of (a) MAE and (b) SimMIM when only using encoder (blue) and decoder (red).

Decoder of SimMIM is a linear layer accounting for only 3.8% of all parameters

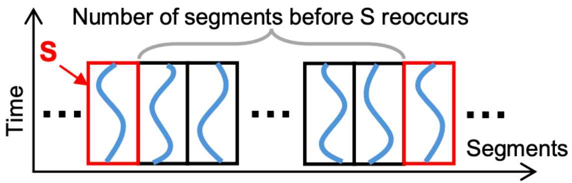
The decoder of the self-supervised LVM is more critical in prediction than the encoder.

From Images to Signals: Are Large Vision Models Useful for Time Series Analysis?. arxiv 2025.

Are LVM useful for TSF?

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- **RQ8**: Will period-based imaging method induce any bias?



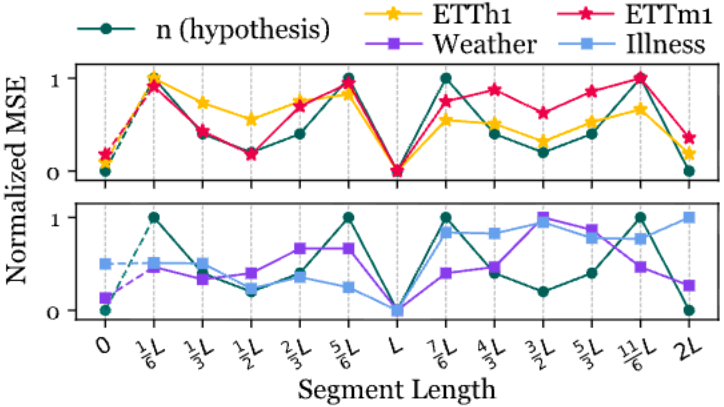


Figure 8: Forecasting performance of MAE *w.r.t.* varying segment length used in UVH imaging. n (green) estimates the difficulty of forecasting.

UVH imaging leads to LVM tending to "combine past cycles" prediction

From Images to Signals: Are Large Vision Models Useful for Time Series Analysis?. arxiv 2025.

Are LVM useful for TSF?

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- **RQ9**: Can LVMs make effective use of look-back windows?

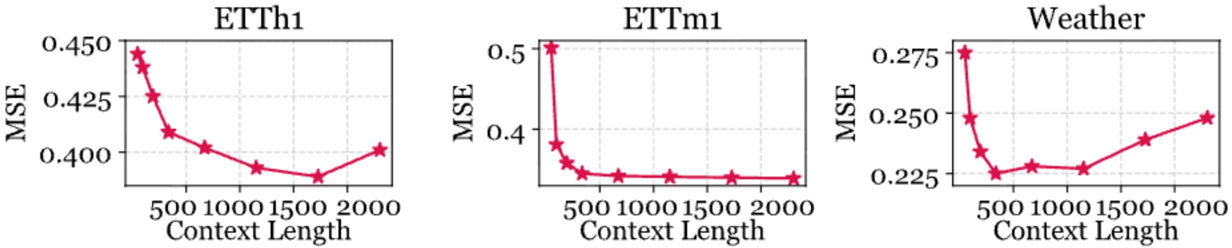


Figure 10: TSF performance (MSE) of MAE with varying look-back window (or context) lengths.

MAE prediction performance tends to stabilize as the window length increases to 1000,

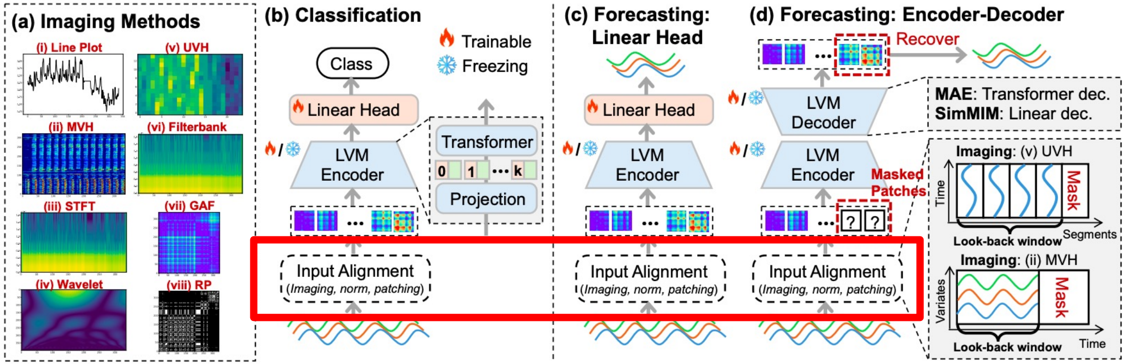
but too long windows may result in information loss due to image compression.

From Images to Signals: Are Large Vision Models Useful for Time Series Analysis?. arxiv 2025.

Summary Notes

When using vison models for time series analysis, several things are important:

- **Normalization**: Targeted processing (controlling mean/std, instance normalization, removing outliers) is needed to fit visual model training characteristics.
- **Image alignment**: Adjust channels (1→3 via duplication/weight averaging) and size (interpolation) for pre-trained models, risking information loss.

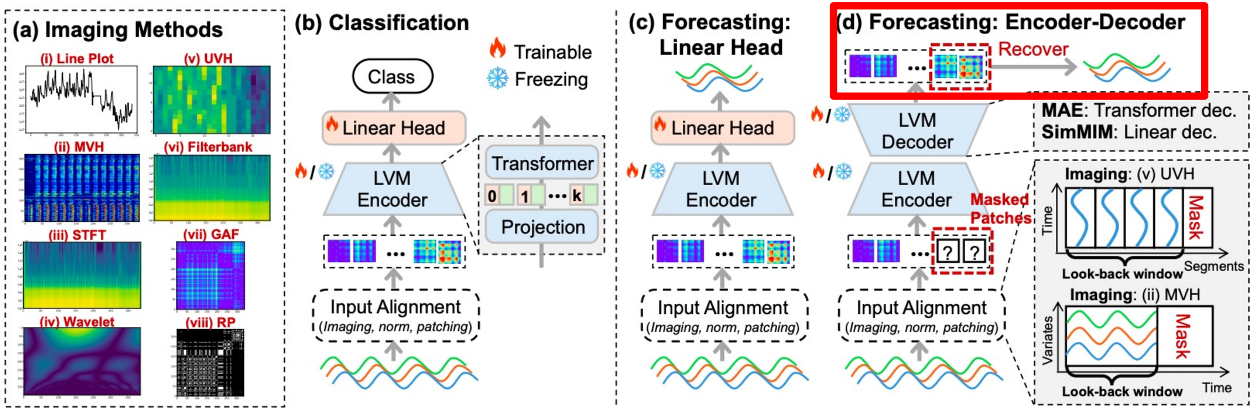


Harnessing Vision Models for Time Series Analysis: A Survey. In IJCAI, 2025.

Summary

When using vison models for time series analysis, several things are important:

- **Temporal recovery**: Recovering raw time series from predicted images: heatmaps and GAFs enable simple/accurate recovery; line plots require dedicated functions; spectrograms are underexplored; RPs are unsuitable due to information loss.



Future Work

- Enhance vision encoder for TSF (e.g., distillation), as decoders dominate TSF performance.
- Mitigate inductive bias from period-based imaging (e.g., UVH) for non-periodic data.
- Optimize time series imaging to resolve information density misalignment when mapping varying input steps to fixed-resolution images.
- Improve TSF performance via tailored components or new training paradigms.
- Reduce computational costs via compression or efficient attention.
- Explore multimodal TS analysis by integrating VLM Agents.

Outline

1 Background & Examples

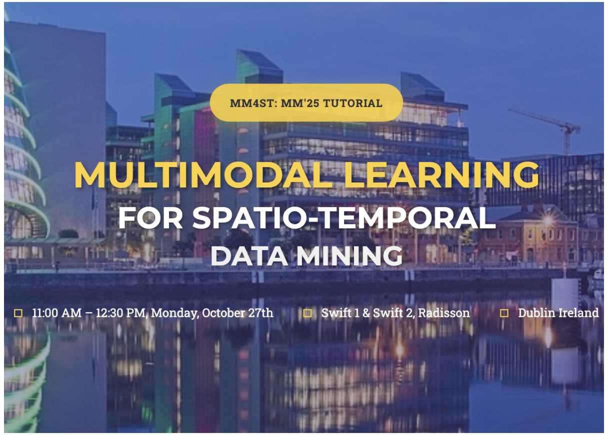
2 Foundation of ST Data

3 Why Multimodal ST Data Fusion

4 Principle of ST Multimodal Fusion

5 Visual/Language Knowledge Transfer

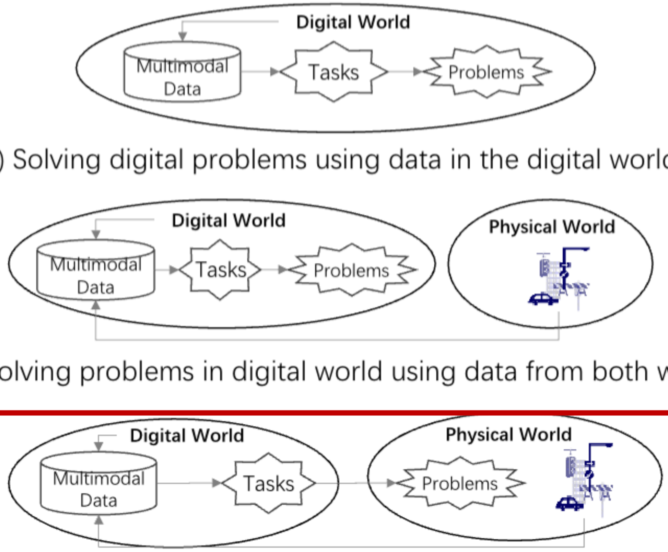
6 **Conclusions**



ST Multimodal Learning is Future

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- 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

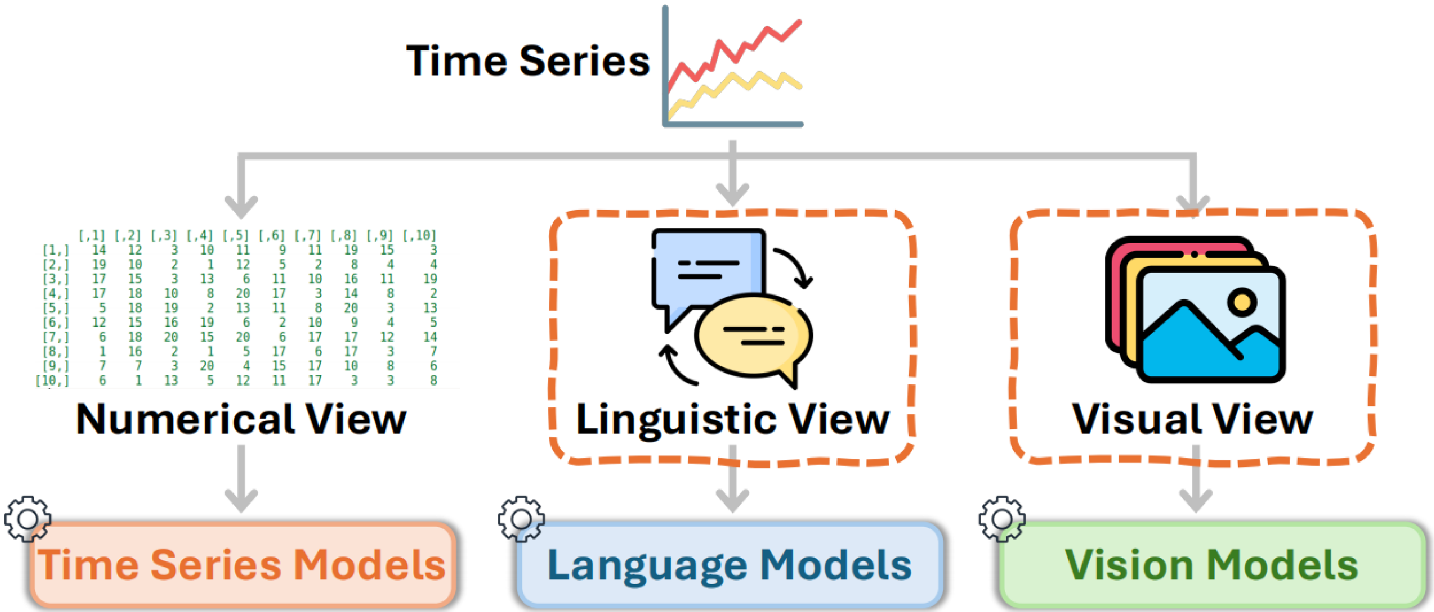
3) Real World Problems, e.g. AQI

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

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

- Knowledge transfer across multi-domain is a promising direction.



Thank you!

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Siru Zhong

- siruzhong@outlook.com
- https://siruzhong.github.io/

Yuxuan Liang

- yuxliang@outlook.com
- https://yuxuanliang.com/