Time-IMM: A Dataset and Benchmark for Irregular Multimodal Multivariate Time Series

UCLA | NATIONAL YANG MING CHIAO TUNG UNIVERSITY | NEURAL INFORMATION PROCESSING SYSTEMS

Ching Chang[1,2*] Jeehyun Hwang[1] Yidan Shi[1] Haixin Wang[1] Wen-Chih Peng[2] Tien-Fu Chen[2] Wei Wang[1]

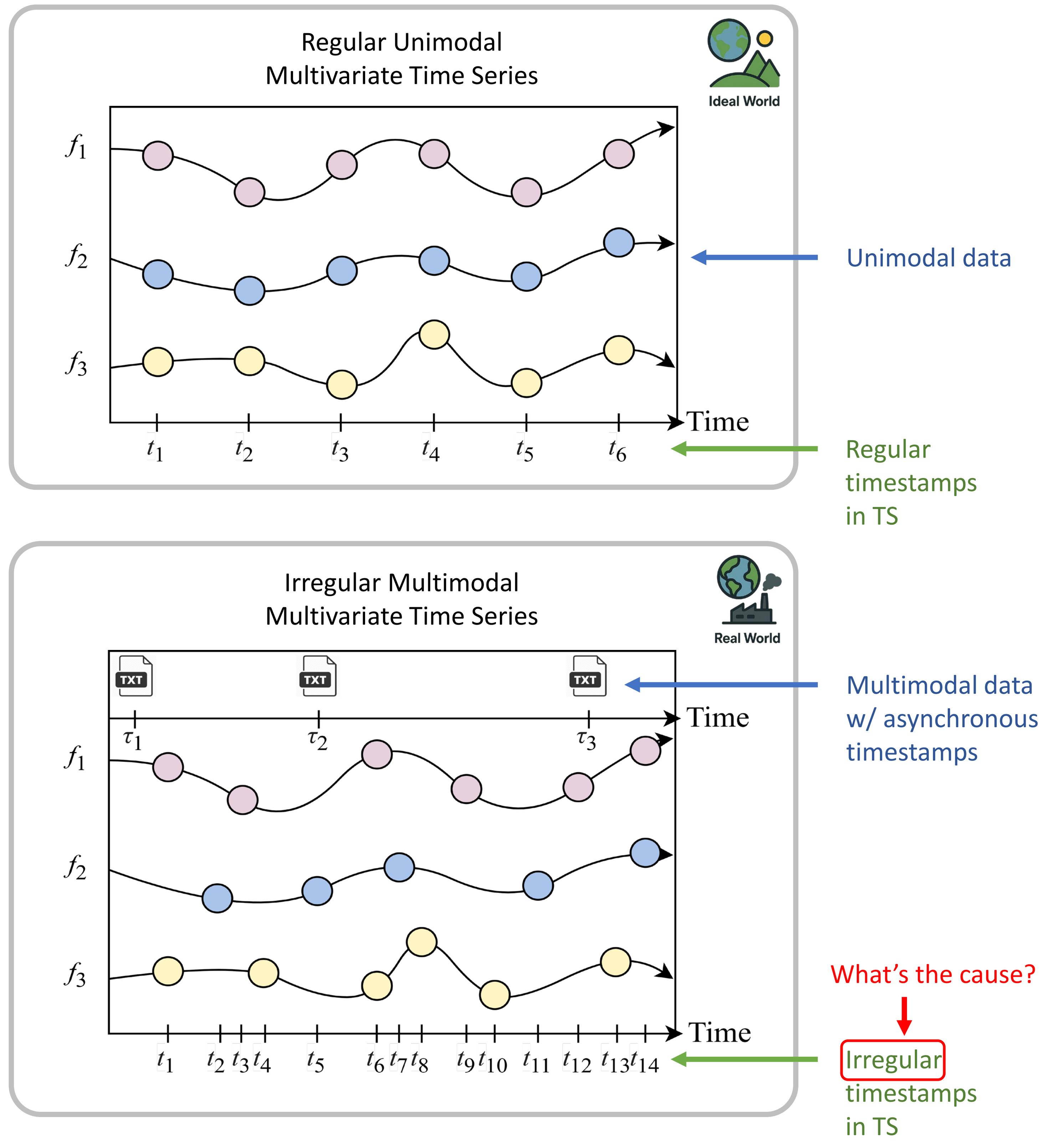
[1]University of California, Los Angeles [2]National Yang Ming Chiao Tung University

[*]Correspondence to: Ching Chang <chingchang0730@ucla.edu>

Project Website | Personal Website

Introduction

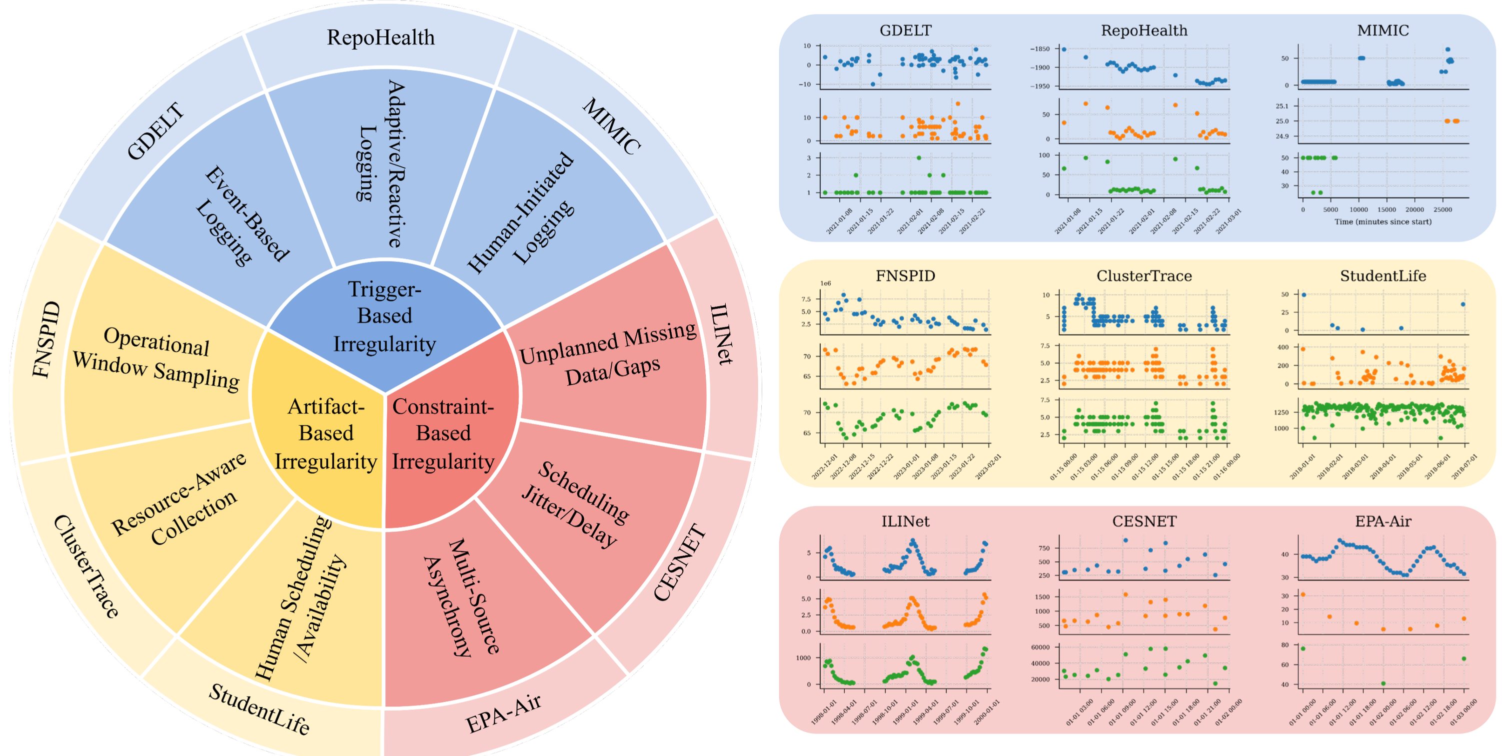
- **3 main challenges in current time series benchmarks**
  - Regular-only assumptions → *unrealistic in practice*
  - Multimodal integration with synchronous timestamps → *ignores asynchrony*
  - No understanding of irregularity causes → *limits interpretability*



- **Time-IMM solves the absence of realistic, cause-driven irregular multimodal time series benchmarks**
  - 9 multimodal (numerical + text) real datasets capturing distinct causes of irregularity
  - A unified multimodal forecasting library (IMM-TSF)
  - Modular fusion strategies for asynchronous numerical–text data
  - Empirical proof that modeling multimodality under irregularity yields robust forecasting gains.

Time-IMM: Dataset for Irregular Multimodal Multivariate Time Series

- **Real-world irregularities arise from three fundamental causes, each with unique modeling challenges.**
  - **Trigger-Based**: Observations occur only when external events or internal triggers happen.
  - **Artifact-Based**: Irregularity caused by system faults, delays, or multi-source asynchrony.
  - **Constraint-Based**: Sampling limited by operational schedules, resource availability, or human timing.

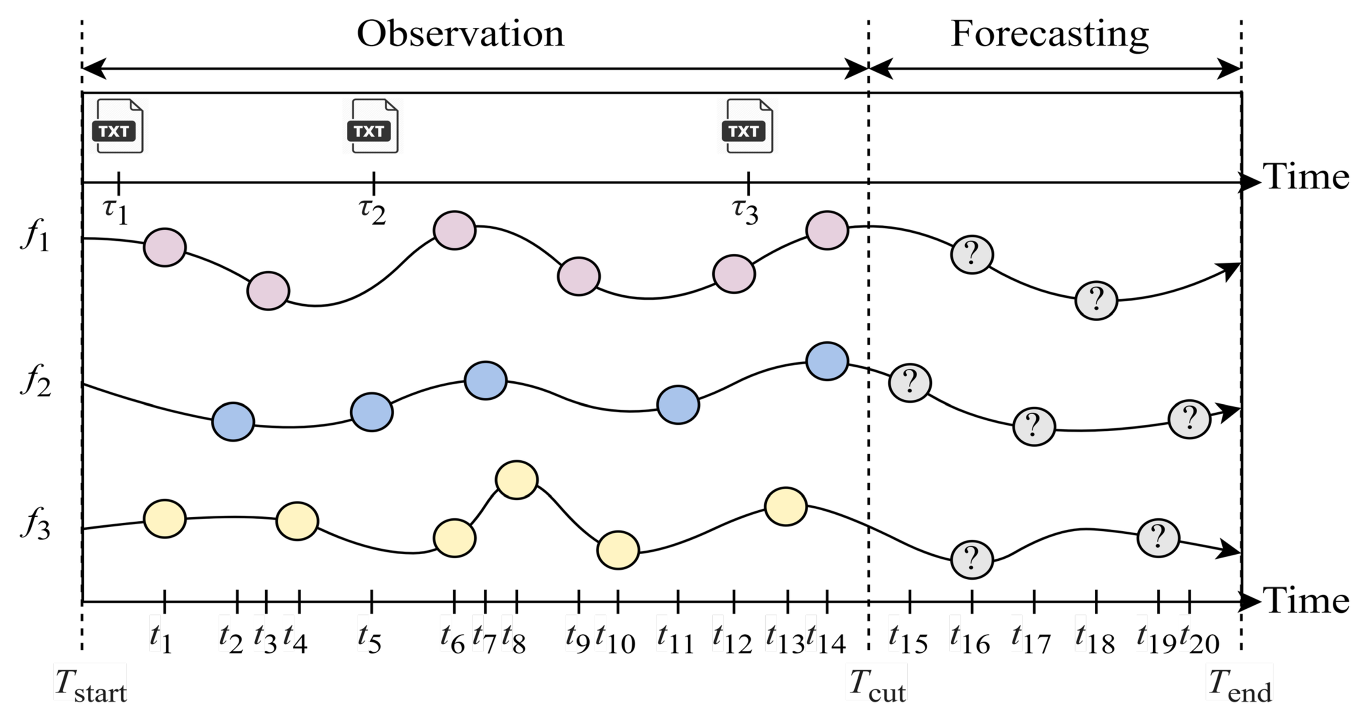


- **Dataset Construction Pipeline**
  1. *Numerical Data*
     - Real-world time series for each irregularity type
     - Preserve native timestamps (no resampling)
  2. *Textual Data*
     - Collect relevant reports, logs, or notes linked to each dataset
     - Filter & summarize using GPT-4.1 Nano
     - Retain original timestamps for text entries
  3. *Multimodal Integration*
     - Combine numerical and textual data while preserving asynchronous timestamps

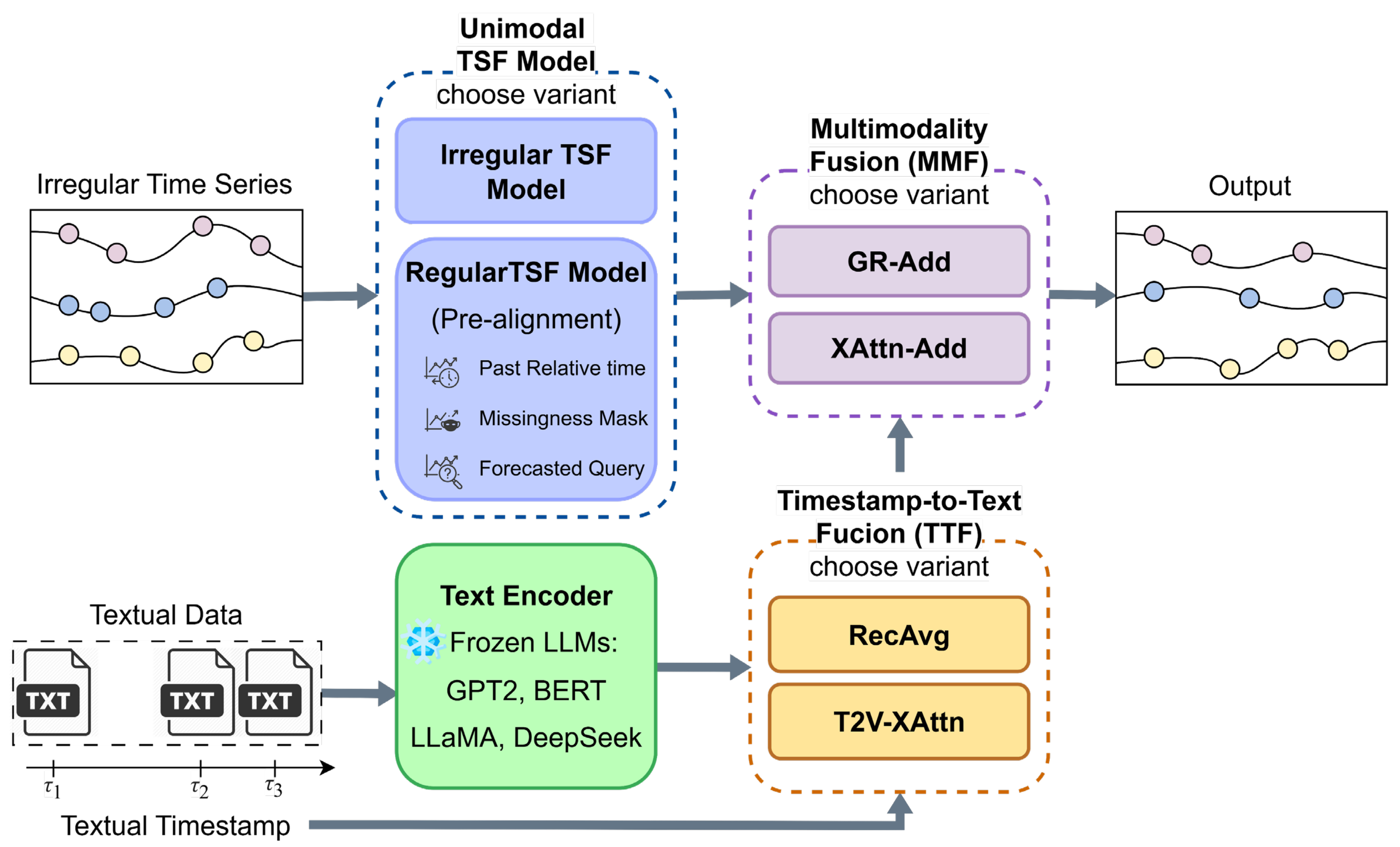
Dataset	# Entities	# Features	# Unique Timestamps	# Observations	Feature Observability Entropy	Temporal Observability Entropy	Mean Inter-Obseration Interval	# Text Entries†	Textual Temporal Observability Entropy†
GDELT	8	5	34317	193205	1	0.9964	7.2364 hours	14357	0.9896
RepoHealth	4	10	6783	67830	1	0.9658	1.8217 days	12310	0.9821
MIMIC	20	30	91098	219949	0.8461	0.6556	14.6157 minutes	1593	0.6758
FNSPID	10	6	3659	209688	1	0.9969	1.4507 days	20826	0.9488
ClusterTrace	3	11	12615	69001	0.893	0.9753	18.1425 minutes	688	0.9971
StudentLife	20	9	1743	153610	0.92	0.9775	1.0191 days	6623	0.9761
ILINet	1	11	4918	4918	0.9267	1	6.989 days	650	1
CESNET	30	10	51107	512760	1	1	1.17 hours	224	0.9869
EPA-Air	8	4	6587	49552	0.3777	0.9576	1.0242 hours	1244	0.9956

IMM-TSF: Benchmark Library for Irregular Multimodal Multivariate Time Series Forecasting

- **Problem Formulation: Irregular Multimodal Multivariate Time Series Forecasting**
  - Predict future time series values using irregularly sampled numerical data and asynchronous textual context.

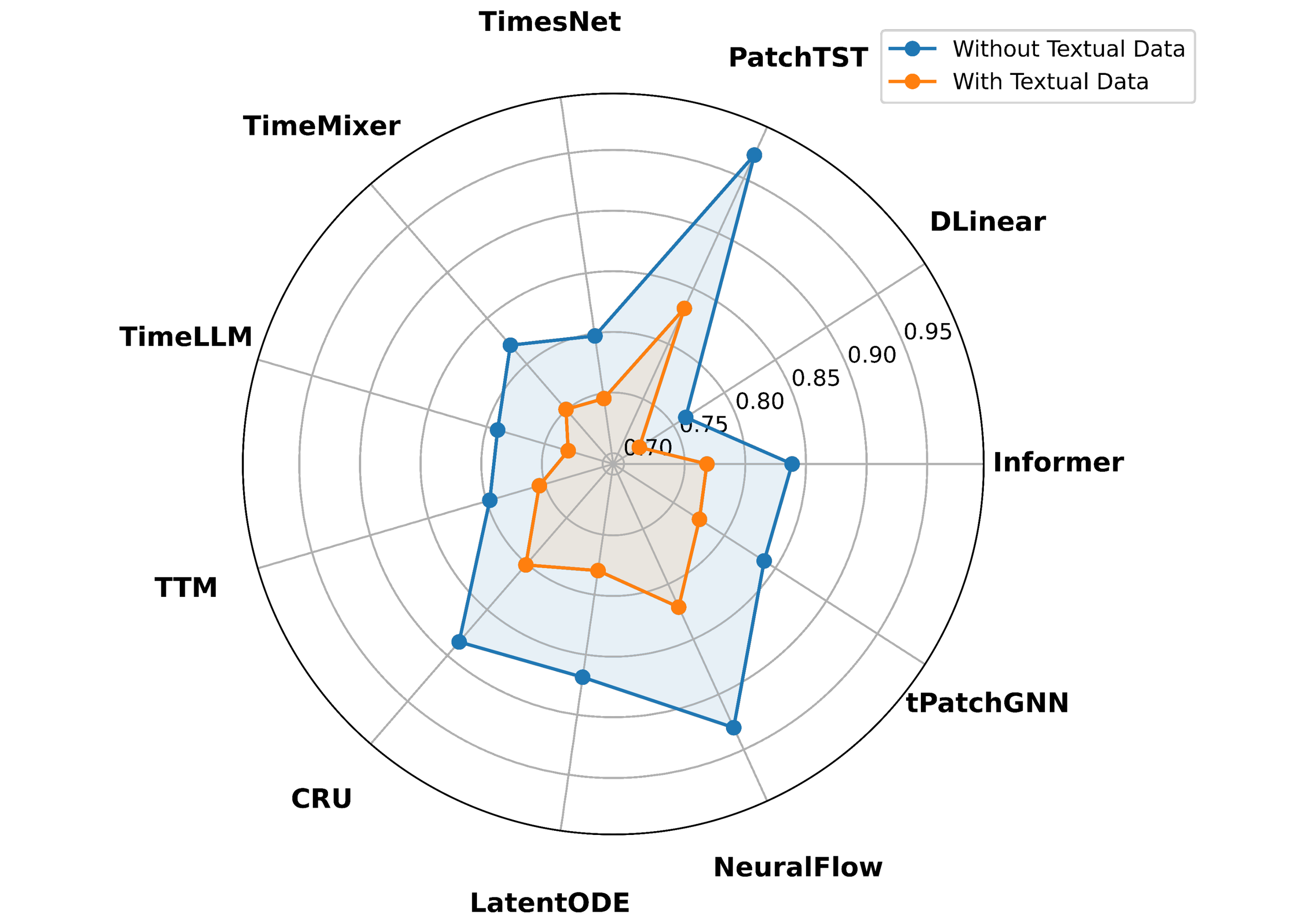


- **Multimodal TSF Library: A plug-and-play framework for forecasting on irregular multimodal time series.**
  - **Timestamp-to-Text Fusion (TTF)**
    - *RecAvg*: recency-weighted aggregation of past text embeddings
    - *T2V-XAttn*: Time2Vec-augmented cross-attention for temporal relevance
  - **Multimodality Fusion (MMF)**
    - *GR-Add*: GRU-gated residual addition for adaptive text influence
    - *XAttn-Add*: cross-attention addition between numerical and textual features



Experimental Results

- **Effectiveness of Multimodality**
  - Across all nine Time-IMM datasets, incorporating textual information consistently improves forecasting accuracy compared to unimodal (numerical-only) models.
    - Average MSE reduction: **6.7%**
    - Maximum improvement: **38.4%** in datasets with highly informative text



- **Multimodal Forecasting Analysis**
  a) Gains Across Datasets
     - Multimodal models outperform unimodal baselines on all datasets, with larger gains when text provides strong contextual signals (e.g., ClusterTrace).
  b) Fusion Strategies
     - *GR-Add* gives the most stable and accurate results; both *RecAvg* and *T2V-XAttn* perform similarly.
  c) Frozen LLM Backbones
     - Text encoder choice has limited effect — forecasting depends more on temporal alignment than on large-scale language understanding.