Unlocking the Value of Text: Event-Driven Reasoning and Multi-Level Alignment for Time Series Forecasting

Siyuan Wang, Peng Chen, Yihang Wang, Wanghui Qiu, Chenjuan Guo, Bin Yang, Yang Shu
{sywang,pchen,yhwang,onehui}@stu.ecnu.edu.cn, {cjguo,byang,yshu}@dase.ecnu.edu.cn
East China Normal University

EAST CHINA NORMAL UNIVERSITY 華東師範大學

You can subscribe to our Lab:
DecisionIntelligence @ ECNU

ICLR

Introduction

Time series forecasting is a fundamental task in numerous domains. Deep learning-based forecasting methods have achieved competitive performance, yet they solely rely on numerical time series data and fail to capture complex event-driven patterns. As shown in Figure \ref{motivation}, regular temporal trends are predictable, but external event-driven abrupt changes cannot be captured by numerical data alone, which highlights the complementary value of textual information.

Despite the complementary value of textual information, existing multimodal forecasting schemes still suffer from two critical defects:

The first defect is **insufficient text utilization**: relevant text information either overlaps with numerical features and ignores external drivers, or only achieves shallow representation fusion without mining deep semantic value. The second defect is **ineffective cross-modal alignment**: the huge modality gap prevents integrating the complementary strengths of text and numerical time series thoroughly.

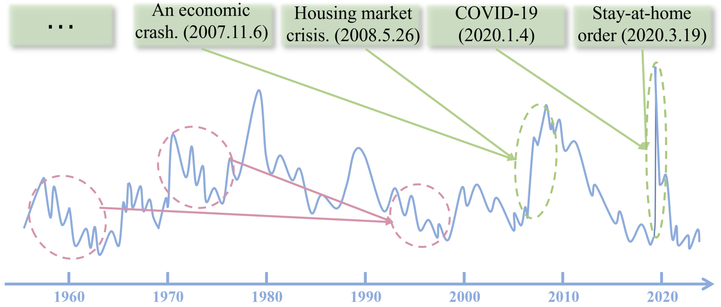


Figure 1 Unemployment rate time series. While certain patterns (pink) exhibit predictable temporal regularities, abrupt changes (green) driven by external events necessitate the integration of textual information to complement numerical forecasting.

Contributions

- We introduce an event-driven reasoning method to extract the forecasting-related information from exogenous text and obtain numerical predictions. It is enhanced by Historical In-Context Learning (HIC), which retrieves historical reasoning examples as prompts to provide error-informed guidance for reasoning. The method improves the reasoning ability of LLMs and unlocks the value of text.
- We propose a multi-level alignment approach. Specifically, we introduce the Endogenous Text Alignment (ETA) for representation-level alignment and the Adaptive Frequency Fusion (AFF) for prediction-level alignment. Through comprehensive alignment, we achieve complementary advantages across both modalities.
- We conduct extensive experiments on 10 real-world datasets from different domains and achieve state-of-the-art prediction accuracy. Moreover, we conduct thorough ablation and analysis experiments to demonstrate our effective utilization of text.

VoT Architecture

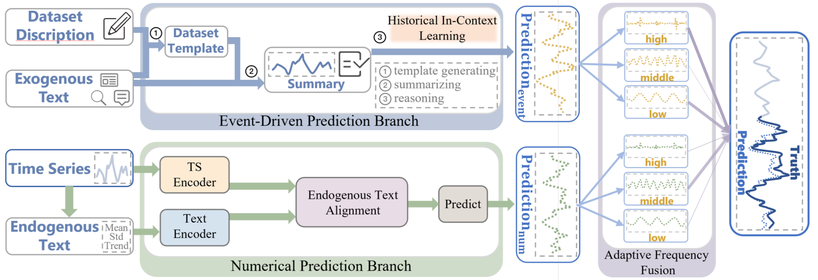


Figure 2 The overview of Aurora. n the Aurora Encoder, the multimodal information is extracted, distilled, and fused. Modality-Guided Multi-head Self-Attention is introduced to inject the domainspecific knowledge into temporal modeling. In the Aurora Decoder, the Prototype-Guided Flow Matching is introduced to support generative probabilistic forecasting.

Event-Driven Reasoning

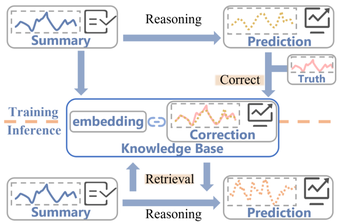


Figure 3: The processing procedure of the Historical In-Context Learning (HIC).

$$\hat{\mathbf{Y}}_i^{\text{event}}, \mathcal{R}_i = \text{Reasoner}(\mathcal{P}_{\text{reason}}, \mathcal{S}_i, \mathbf{X}_i)$$

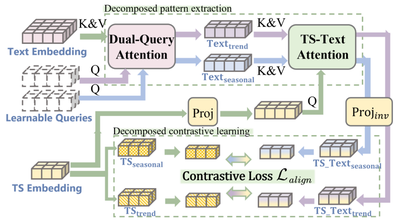
$$\mathcal{C}_i = \text{Reasoner}(\mathcal{P}_{\text{correct}}, \hat{\mathbf{Y}}_i^{\text{event}}, \mathcal{R}_i, \mathbf{Y}_i, \mathbf{X}_i)$$

$$\mathcal{K} = \{(\text{Embed}(\mathcal{S}_i), \mathcal{C}_i)\}_{i=1}^{M}$$

$$\tilde{i} = \arg \max_{(\mathcal{S}_i, \mathcal{C}_i) \in \mathcal{K}} \text{simi}(\text{Embed}(\mathcal{S}_j), \text{Embed}(\mathcal{S}_i))$$

$$\hat{\mathbf{Y}}_j^{\text{event}} = \text{Reasoner}(\mathcal{P}_{\text{ICL}}, \mathcal{C}_{\tilde{i}}, \mathcal{S}_j, \mathbf{X}_j)$$

Multi-Level Alignment



Figutr 4: The processing procedure of the Endogenous Text Alignment (ETA).

$$\mathbf{E}^{\text{tr}} = \text{Attention}(\mathbf{Q}^{\text{tr}}, \mathbf{H}^{\text{text}}, \mathbf{H}^{\text{text}})$$

$$\mathbf{E}^{\text{se}} = \text{Attention}(\mathbf{Q}^{\text{se}}, \mathbf{H}^{\text{text}}, \mathbf{H}^{\text{text}})$$

$$\mathbf{Z}^* = \text{Cross-Attention}(\text{Proj}(\mathbf{H}^{\text{ts}}), \mathbf{E}^*, \mathbf{E}^*)$$

$$\tilde{\mathbf{Z}}^* = \text{Proj}_{\text{inv}}(\mathbf{Z}^*)$$

$$\mathcal{L}_{\text{align}} = \frac{1}{2} \sum (-\log \frac{\exp(\text{sim}(\bar{\mathbf{H}}_i^*, \bar{\mathbf{Z}}_i^*))}{\sum_{j=1}^{B} \exp(\text{sim}(\bar{\mathbf{H}}_i^*, \bar{\mathbf{Z}}_j^*))} - \log \frac{\exp(\text{sim}(\bar{\mathbf{Z}}_i^*, \bar{\mathbf{H}}_i^*))}{\sum_{j=1}^{B} \exp(\text{sim}(\bar{\mathbf{Z}}_i^*, \bar{\mathbf{H}}_j^*))})$$

Experiments

Main Results

Table 1: Forecasting results of ts-only and text-enhanced methods and VoT.

Category			Time series-only						Text-enhanced					
Models	VoT		PatchTST (2023)		iTransformer (2024b)		RaFT (2025)		PatchTST*		iTransformer*		RaFT*	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Agriculture	**0.209**	**0.302**	0.228	0.303	0.220	0.308	0.226	0.322	0.232	0.316	0.229	0.310	0.246	0.333
Climate	**1.078**	**0.840**	1.184	0.888	1.135	0.865	1.289	0.926	1.178	0.887	1.117	0.858	1.342	0.944
Economy	**0.201**	**0.353**	0.210	0.363	0.222	0.378	0.265	0.411	0.219	0.370	0.213	0.367	0.275	0.420
Energy	**0.222**	**0.343**	0.250	0.363	0.269	0.382	0.254	0.367	0.253	0.365	0.265	0.383	0.246	0.360
Environment	**0.268**	**0.380**	0.317	0.395	0.276	0.386	0.339	0.423	0.318	0.397	0.278	0.390	0.337	0.422
Health	**1.205**	**0.714**	1.432	0.804	1.519	0.833	1.833	0.975	1.360	0.768	1.713	0.915	1.788	0.963
Security	**70.117**	**3.937**	72.027	4.062	75.042	4.217	77.204	4.473	72.721	4.177	74.032	4.154	76.587	4.448
Social Good	**0.804**	**0.389**	0.944	0.475	0.961	0.463	0.968	0.484	0.909	0.427	1.027	0.515	0.970	0.477
Traffic	**0.169**	**0.232**	0.176	0.234	0.184	0.238	0.288	0.382	0.174	0.239	0.184	0.237	0.300	0.394
Weather	**0.968**	**0.706**	1.145	0.751	1.231	0.803	1.099	0.746	1.036	0.707	1.004	0.709	1.096	0.745
1st counts	**20**		0		0		0		0		0		0	

Table 2: Forecasting results of multimodal methods and VoT.

Models	VoT		GPT4TS (2025)		GPT4MTS (2024)		TaTS (2026)		Time-VLM (2025)		CALF (2025a)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Agriculture	**0.209**	0.302	0.220	**0.294**	0.225	0.298	0.215	0.301	0.238	0.303	0.250	0.315
Climate	**1.078**	**0.840**	1.184	0.891	1.182	0.889	1.180	0.887	1.195	0.899	1.286	0.922
Economy	**0.201**	**0.353**	0.217	0.371	0.208	0.363	0.215	0.368	0.229	0.384	0.207	0.357
Energy	**0.222**	**0.343**	0.260	0.376	0.262	0.380	0.255	0.368	0.260	0.374	0.244	0.365
Environment	**0.268**	**0.380**	0.322	0.393	0.323	0.400	0.319	0.396	0.320	0.398	0.325	0.387
Health	**1.205**	**0.714**	1.341	0.777	1.464	0.799	1.356	0.767	1.490	0.835	1.491	0.775
Security	**70.117**	**3.937**	71.165	4.047	71.487	4.068	72.406	4.097	73.731	4.182	76.376	4.300
Social Good	**0.804**	**0.389**	0.917	0.476	0.920	0.450	0.918	0.428	0.869	0.444	0.906	0.401
Traffic	**0.169**	**0.232**	0.206	0.266	0.203	0.261	0.179	0.238	0.217	0.320	0.222	0.293
Weather	**0.968**	**0.706**	1.048	0.708	0.986	0.711	1.037	**0.706**	1.061	0.717	1.098	0.714
1st counts	**19**		1		0		1		0		0	

Ablation Study

		TS-only	w/o ETA	w/o HIC	w/o Event	VoT
Energy	MSE	0.250	0.241	0.238	0.243	**0.222**
	MAE	0.363	0.355	0.363	0.360	**0.343**
Social Good	MSE	0.944	0.840	0.845	0.876	**0.804**
	MAE	0.475	0.424	0.410	0.436	**0.389**

Table 3: Ablation study results on Energy and Social Good datasets

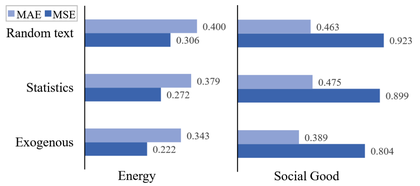


Figure 4: Ablation study on different text sources for event-driven reasoning.

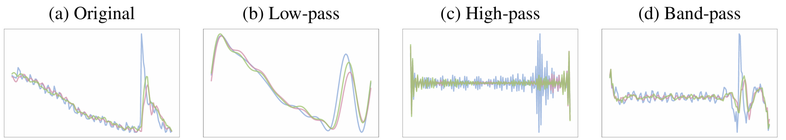


Figure 6: Frequency domain analysis of time series predictions (Social Good). (a) Original signals without frequency decomposition. (b)-(d) Frequency-filtered components: (b) Low-pass filtered signals, (c) High-pass filtered signals, and (d) Band-pass filtered signals. Ground Truth (blue), Time series-only prediction (pink), and event-driven prediction (green)