WHEN EVENTS MOVE THE CURVE: AN EVENT-INFORMED REASONING FRAMEWORK FOR MULTIMODAL TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Strong time series forecasting models can effectively capture endogenous patterns from numerical histories, but they cannot directly access the external events described in text. Existing multimodal methods either compress text into latent features, underutilizing event semantics, or rely on LLMs to directly generate numerical forecasts, risking imprecision due to the language–number modality gap. Consequently, they struggle to retain the forecasting capability of numerical models while fully exploiting textual event semantics and accurately translating them into future numerical changes. We propose AFTER (**A**djustment **F**ramework through **T**extual **E**vent **R**easoning), an event-informed framework that retains a trained numerical backbone and uses textual events to adjust its predictions without additional parameter optimization. AFTER first employs History-Aware Event Reasoning to identify forecast-relevant events from the textual context. It then derives forecast offsets through contextual event reasoning, case-based analogical reasoning over an Event-Response Memory, and direct numerical transfer of retrieved continuations. These complementary offset candidates are robustly integrated and added to the backbone forecast. Experiments across nine real-world multimodal time series datasets and multiple forecasting horizons demonstrate the effectiveness and broad applicability of AFTER against strong numerical and multimodal baselines. Our code is available at `https://anonymous.4open.science/r/AFTER-ICLR--0638/`

1 INTRODUCTION

Time series forecasting is a fundamental task that supports planning and decision-making in a wide range of real-world domains, including energy, transportation, economics, weather, and public health (Lim & Zohren, 2020; Li et al., 2018; Kamarthi et al., 2021). In recent years, substantial progress has been made in numerical time series forecasting, with a series of effective models achieving strong performance across diverse forecasting benchmarks (Wu et al., 2021; Nie et al., 2023; Wu et al., 2023; Liu et al., 2024b). While these models primarily learn predictive patterns from historical numerical observations, many real-world time series are also accompanied by textual information, such as news reports, policy announcements, and domain-specific reports, which may provide additional information beyond the observed values. This has motivated the emerging study of multimodal time series forecasting, where numerical sequences are modeled together with relevant texts (Liu et al., 2024a). Despite its potential, how to effectively extract and incorporate forecast-relevant information from noisy and heterogeneous texts remains a central challenge.

This naturally raises two questions: how does textual information influence future forecasts, and can the strong time series forecasting models developed in prior work be leveraged? As shown in Figure 1, following the 2011 Libya oil shock and the 2020 intervention in the oil price war, the observed gasoline-price series continues to rise, whereas iTransformer predicts either a decline or a nearly flat trajectory. Text can inform future forecasts by revealing external events that may alter subsequent temporal dynamics. Strong time series forecasting models can effectively learn endogenous patterns from numerical history, but they cannot directly obtain the external event information contained in text, which can cause their forecasts to miss the resulting changes. Consequently, if the influence of these events can be inferred from text and used to adjust the numerical forecast, we can jointly

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

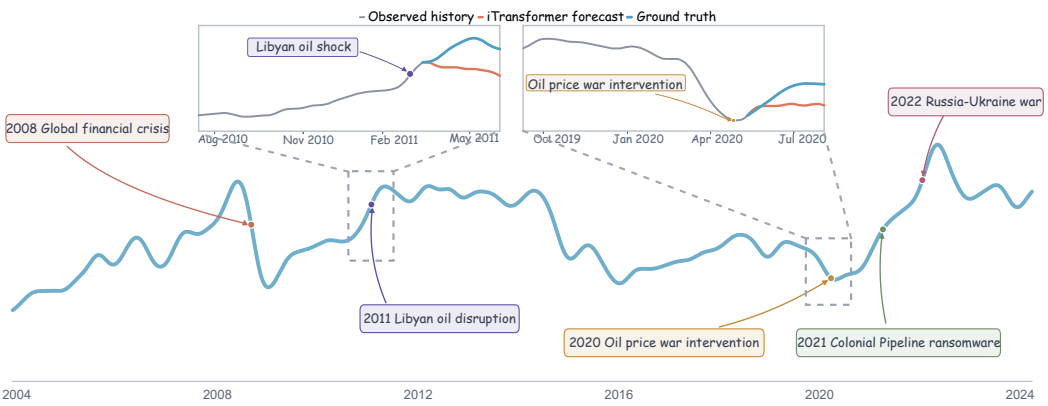


Figure 1: By analyzing the gasoline price series and its corresponding texts in the energy domain of Time-MMD (Liu et al., 2024a), we find that external events described in the texts can influence subsequent numerical dynamics. Meanwhile, strong numerical forecasting models such as iTransformer (Liu et al., 2024b) cannot access such event information from historical numerical values alone and may consequently mispredict the subsequent trajectory.

leverage the external event evidence provided by text and the predictive capability of strong time series models to produce more accurate future forecasts.

Prior studies have extensively explored how to use textual information to improve time series forecasting. One line of work encodes textual context into representations and incorporates them into forecasting models (Liu et al., 2024a; Li et al., 2026), while another employs pretrained language models to jointly process textual and numerical inputs (Jia et al., 2024; Wang et al., 2025a). Nevertheless, the deeper semantic information contained in text, particularly information about external events and their potential effects on future dynamics, remains underutilized. More recently, reasoning-oriented approaches have sought to leverage the reasoning capabilities of LLMs for event-aware and step-by-step forecasting (Guan et al., 2026; Zhou et al., 2026). Yet this paradigm must bridge a fundamental modality gap between high-level linguistic reasoning and fine-grained continuous values, which can compromise numerical precision when LLMs directly generate complete forecasts (Sun et al., 2024).

Based on these observations, we propose AFTER (**A**djustment **F**ramework through **T**extual **E**vent **R**easoning), a forecasting framework that reasons over textual events to adjust the predictions of a trained numerical backbone without additional parameter optimization. AFTER comprises two coordinated branches: a numerical forecasting branch and an event-informed adjustment branch. The numerical branch employs a trained backbone to produce a base forecast from the historical observations. The adjustment branch begins with History-Aware Event Reasoning (HAER), which uses the LLM to interpret the available textual events in light of the historical numerical dynamics and textual context, producing an event state grounded in the observed history. Conditioned on this event state and anchored to the base forecast, Complementary Offset Reasoning (COR) then estimates forecast offsets through three complementary routes: contextual event reasoning, case-based analogical reasoning, and direct numerical transfer of retrieved continuations, where the first two are performed by the LLM. Finally, Robust Offset Integration (ROI) consolidates these offset candidates and adds the resulting offset to the base forecast to produce the final prediction. In this way, AFTER combines the endogenous pattern modeling ability of strong numerical forecasting models with semantic reasoning over external events, while requiring no additional parameter optimization and retaining a transparent evidence trail for each forecast adjustment. Our main contributions are summarized as follows:

- We introduce an event-informed adjustment method for multimodal time series forecasting. Rather than using text to generate complete forecasts, we leverage textual events to infer future changes that are difficult to anticipate from numerical histories alone and use them to complement strong numerical forecasting models.

- We propose AFTER, a framework that adjusts the predictions of a trained forecasting backbone without additional parameter optimization. Through History-Aware Event Reasoning (HAER), Complementary Offset Reasoning (COR), and Robust Offset Integration (ROI), AFTER grounds event reasoning in the observed history, derives offsets from complementary semantic and numerical evidence, and robustly incorporates them into the backbone forecast.
- Extensive experiments across nine real-world multimodal time series datasets and multiple forecasting horizons demonstrate the effectiveness and generalizability of AFTER. Detailed case studies further provide traceable insights into how textual events contribute to individual forecast adjustments.

2 RELATED WORK

Time Series Forecasting. A wide range of methods have been developed for time series forecasting. Linear and MLP-based methods employ lightweight feed-forward structures to learn temporal patterns efficiently (Zeng et al., 2023; Chen et al., 2023). Transformer-based methods use attention mechanisms to capture long-range temporal dependencies and relationships among variables (Zhou et al., 2021; Wu et al., 2021; Nie et al., 2023; Liu et al., 2024b). Multi-scale methods model temporal patterns at different resolutions through adaptive decomposition and cross-scale information interaction (Hu et al., 2025; Wang et al., 2023; 2024). Frequency-aware approaches further separate and model heterogeneous frequency components (Zhou et al., 2022b; Huang et al., 2025). More recently, time series foundation models have been pretrained on large and diverse time series corpora to support zero-shot forecasting across datasets and domains (Ansari et al., 2024; Woo et al., 2024). Despite their strong performance, these methods primarily focus on unimodal numerical time series.

Multimodal Time Series Forecasting. A growing body of research has explored the use of textual and other contextual information for time series forecasting. Framework-based approaches provide general mechanisms for incorporating paired texts into existing time series models (Liu et al., 2024a; Li et al., 2026). LLM-based adaptation methods repurpose pretrained language models for forecasting through temporal-language alignment or prompt-based modeling (Zhou et al., 2023; Jia et al., 2024; Jin et al., 2024; Liu et al., 2025). Beyond text, Time-VLM jointly explores temporal, textual, and visual modalities using pretrained vision-language models (Zhong et al., 2025). Recent foundation models, such as ChatTime and Aurora, further support unified multimodal inputs and zero-shot forecasting (Wang et al., 2025a; Wu et al., 2026). Across these directions, text is primarily used as an auxiliary representation or conditioning signal, leaving its deeper semantic content underutilized. More recently, reasoning-oriented methods such as TimeOmni-1 and Time-R1 have employed LLMs for event-aware and step-by-step forecasting (Guan et al., 2026; Zhou et al., 2026). However, the modality gap between language and time series makes it difficult to translate qualitative event reasoning into quantitatively calibrated forecasts (Sun et al., 2024).

3 METHODOLOGY

3.1 PROBLEM FORMULATION

Let $\mathbf{X} \in \mathbb{R}^{L \times C}$ denote the historical observations of a C-variate time series over a look-back window of length L, and let $\mathcal{T}$ denote the textual context available at the forecast origin. Given $\mathbf{X}$ and $\mathcal{T}$, multimodal time series forecasting aims to predict the future sequence $\mathbf{Y} \in \mathbb{R}^{H \times C}$ over a forecasting horizon of length H:

$$\widehat{\mathbf{Y}} = \mathcal{F}(\mathbf{X}, \mathcal{T}) \tag{1}$$

where $\mathcal{F}$ denotes the forecasting model, and $\widehat{\mathbf{Y}}$ is expected to approximate the ground-truth future sequence $\mathbf{Y}$.

3.2 FRAMEWORK OVERVIEW

As illustrated in Figure 2, AFTER comprises two coordinated branches: a numerical forecasting branch and an event-informed adjustment branch. Given the historical time series $\mathbf{X}$, the numerical

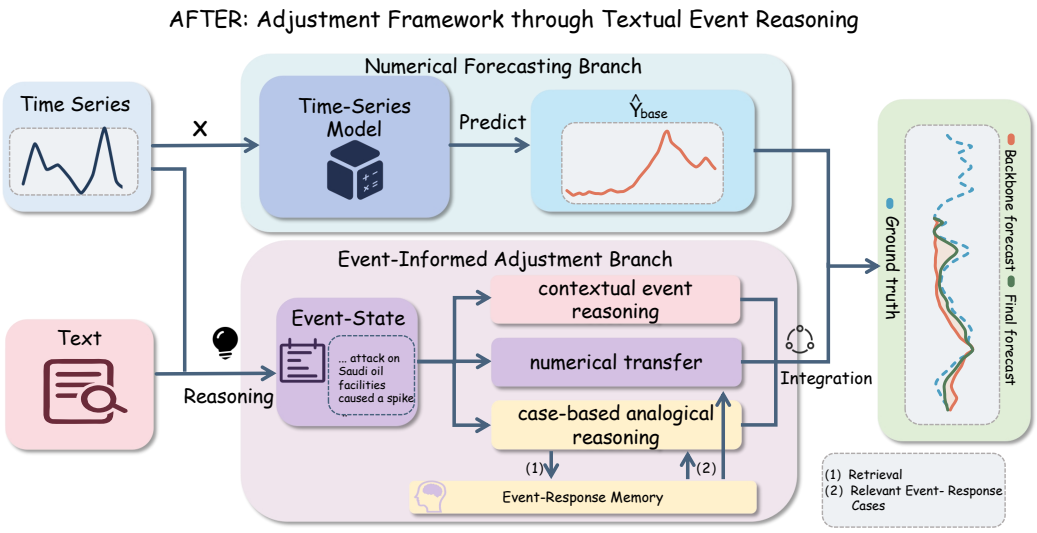


Figure 2: Overall architecture of AFTER. A numerical forecasting branch produces the base forecast, while an event-informed adjustment branch uses textual events and retrieved historical cases to adjust it.

branch employs a trained time series model f_θ to produce a base forecast:

$$\widehat{\mathbf{Y}}_{\text{base}} = f_\theta(\mathbf{X}) \tag{2}$$

The event-informed adjustment branch jointly uses the historical time series and textual context. History-Aware Event Reasoning (HAER) first uses the LLM to interpret the textual events in relation to the observed numerical dynamics, producing a history-aware event state. Conditioned on this state and anchored to the base forecast, Complementary Offset Reasoning (COR) estimates forecast offsets through three complementary routes: contextual event reasoning, case-based analogical reasoning, and direct numerical transfer of retrieved continuations. To support the latter two routes, an Event-Response Memory is constructed from the training set, where each record consists of a historical time series, its corresponding event state, and the observed future. Relevant historical cases are then retrieved for the current sample. Finally, Robust Offset Integration (ROI) consolidates the complementary offset candidates into an event-informed offset $\widehat{\mathbf{\Delta}}$. Rather than replacing the numerical forecast, AFTER applies this offset to the base forecast:

$$\widehat{\mathbf{Y}} = \widehat{\mathbf{Y}}_{\text{base}} + \widehat{\mathbf{\Delta}} \tag{3}$$

This formulation retains the forecasting capability of the numerical backbone while using textual events to account for changes that cannot be inferred from the numerical history alone, without additional parameter optimization.

3.3 History-Aware Event Reasoning

Textual context may contain noise, repeated reports, and events whose effects are already reflected in the observed series. Treating all texts as predictive evidence may therefore introduce irrelevant or stale influences. History-Aware Event Reasoning (HAER) addresses this issue by jointly analyzing the historical series $\mathbf{X}$ and textual context $\mathcal{T}$ to identify which events may still affect future numerical dynamics.

To make the time series accessible to the LLM, we serialize $\mathbf{X}$ as a timestamped table (Wang et al., 2025b; Hoo et al., 2025). The first observation is specified as a reference, and each table entry records its difference from this reference:

$$\widetilde{x}_{\tau,c} = x_{\tau,c} - x_{1,c} \tag{4}$$

where $x_{\tau,c}$ denotes variable c at time step τ. This relative tabular representation explicitly presents the direction and magnitude of temporal changes, facilitating the identification of trends, fluctuations, and turning points.

Using this representation, HAER employs an LLM to relate the texts to the observed temporal evolution, filtering irrelevant or repeated information and identifying the events that remain relevant to future dynamics. HAER represents the resulting event state as

$$\mathcal{S} = (s, c) = \mathrm{LLM}_{\mathrm{HAER}}(\mathbf{X}, \mathcal{T}) \tag{5}$$

where s denotes the *context status*, which categorizes the influence of the textual evidence as *none*, *weak*, or *active*. Specifically, *none* indicates no relevant textual influence, *weak* indicates a plausible but uncertain influence, and *active* indicates concrete evidence that may affect the forecasting horizon. The variable c denotes the *context summary*, which summarizes the relevant events and their likely influence on future dynamics.

3.4 Complementary Offset Reasoning

After HAER identifies the events that may affect future dynamics, the next challenge is to translate the resulting qualitative event state into a quantitative offset to the base forecast. A natural first step is to let the LLM reason from the current event state and base forecast, yielding a context-dependent estimate. However, reasoning from the current context alone lacks empirical evidence about how similar events have affected time series in the past. We therefore introduce an Event-Response Memory and retrieve relevant historical cases to support case-based analogical reasoning. Since both estimates still rely on the numerical judgments of the LLM, we further construct a deterministic reference by directly transferring the observed continuations of the retrieved cases. This progression leads to three complementary routes: contextual event reasoning, case-based analogical reasoning, and direct numerical transfer. Together, they provide contextual flexibility, historical grounding, and numerical stability, respectively.

Contextual event reasoning. The first route conditions the LLM on the historical series, event state, and base forecast. To support their joint interpretation, the history and base forecast are represented as one continuous trajectory using the same reference value. The LLM then reasons about how the events summarized in $\mathcal{S}$ may influence future numerical dynamics and expresses the inferred influence as an offset to the base forecast:

$$\widehat{\boldsymbol{\Delta}}_{\mathrm{ctx}} = \mathrm{LLM}_{\mathrm{ctx}}\left(\mathbf{X}, \mathcal{S}, \widehat{\mathbf{Y}}_{\mathrm{base}}\right) \tag{6}$$

When the context status is *none*, HAER identifies no forecast-relevant textual influence, and the contextual offset is set to zero, while memory retrieval and numerical transfer are skipped; consequently, the final offset is zero and the backbone forecast remains unchanged.

Event-response memory. To provide historical grounding for offset reasoning, we construct an Event-Response Memory using only the training set, as illustrated in Figure 3. Each memory record consists of a historical series, its event state, and the corresponding observed future:

$$\mathcal{M} = \{(\mathbf{X}_i, \mathcal{S}_i, \mathbf{Y}_i) \,|\, i \in \mathcal{I}_{\mathrm{train}},\ s_i \in \{\textit{weak}, \textit{active}\}\} \tag{7}$$

For a query sample, we retrieve relevant records according to both event semantics and historical dynamics. Specifically, semantic retrieval computes the cosine similarity between the embeddings of the current context summary c and a memory summary c_i. In parallel, numerical retrieval compares their normalized historical change patterns. The two scores are combined using their geometric mean:

$$r_i^{\mathrm{sem}} = \max\{0, \cos(e(c), e(c_i))\}, \quad r_i^{\mathrm{shape}} = \exp[-\,\mathrm{RMSE}\,(\phi(\mathbf{X}), \phi(\mathbf{X}_i))], \quad r_i = \sqrt{r_i^{\mathrm{sem}} r_i^{\mathrm{shape}}} \tag{8}$$

Here, $e(\cdot)$ denotes the textual embedding function, and $\phi(\cdot)$ extracts the normalized first-difference pattern of a time series. Memory records are ranked according to r_i, and the top-K records form the retrieved case set $\mathcal{R}$.

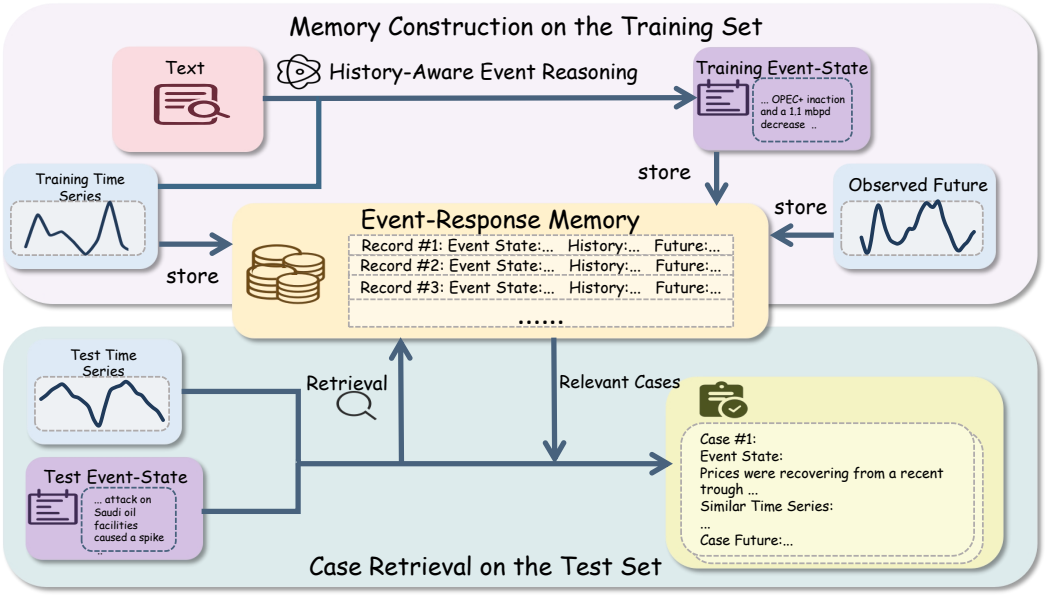


Figure 3: Construction and retrieval of the Event-Response Memory. The memory stores historical series, event states, and observed futures from selected training samples and retrieves relevant cases for each test query.

Case-based analogical reasoning. Reasoning from the current context alone lacks empirical evidence about how similar events have affected time series in the past. To address this limitation, we provide the retrieved case set $\mathcal{R}$, together with the current history, event state, and base forecast, to the LLM. The LLM reasons by analogy with these historical cases and estimates a memory-conditioned offset:

$$\widehat{\boldsymbol{\Delta}}_{\text{mem}} = \text{LLM}_{\text{mem}}\left(\mathbf{X}, \mathcal{S}, \widehat{\mathbf{Y}}_{\text{base}}, \mathcal{R}\right) \tag{9}$$

Direct numerical transfer. Both preceding routes rely on the numerical judgments of the LLM. To provide a stable numerical reference, we directly transfer the observed continuations of the retrieved cases. For each case $i \in \mathcal{R}$, its continuation $\mathbf{Y}_i - \mathbf{x}_{i,L}$ is translated to the current endpoint $\mathbf{x}_L$ and compared with the base forecast. The resulting numerical-analog offset is:

$$\widehat{\boldsymbol{\Delta}}_{\text{ana}} = \text{Median}_{i\in\mathcal{R}}\left[\mathbf{x}_L + (\mathbf{Y}_i - \mathbf{x}_{i,L}) - \widehat{\mathbf{Y}}_{\text{base}}\right] \tag{10}$$

where $\mathbf{x}_L$ and $\mathbf{x}_{i,L}$ denote the final observations of the current series and retrieved case i, respectively.

3.5 Robust Offset Integration

The three routes rely on different evidence and inference mechanisms, and their offset estimates may therefore disagree. Robust Offset Integration (ROI) combines them using an element-wise median:

$$\widehat{\Delta} = \text{Median}\left(\widehat{\Delta}_{\text{ctx}}, \widehat{\Delta}_{\text{mem}}, \widehat{\Delta}_{\text{ana}}\right) \tag{11}$$

The median prevents any single abnormal estimate from dominating the final adjustment without introducing learned fusion parameters. The final forecast is then obtained according to Equation 3.

4 Experiment

Datasets. We evaluate AFTER on nine real-world multimodal time series datasets from Time-MMD (Liu et al., 2024a), covering Agriculture, Climate, Economy, Energy, Environment, Health,

Security, Social Good, and Traffic. For each domain, we use the numerical time series together with the timestamp-aligned factual texts available within the observation window. Since the original Economy series is arranged in reverse chronological order, we reorder it by ascending timestamp before constructing the data splits. Each series is then chronologically divided into training, validation, and test sets with a ratio of 7:1:2. Detailed dataset information is provided in Appendix A.

Baselines. We compare AFTER with three categories of baselines. The numerical forecasting models include iTransformer (Liu et al., 2024b), PatchTST (Nie et al., 2023), and TimeKAN (Huang et al., 2025), which are also adopted as the backbones of AFTER. The backbone-compatible multimodal frameworks include MM-TSFLib (Liu et al., 2024a) and TaTS (Li et al., 2026). The standalone multimodal forecasting methods include GPT4TS (Zhou et al., 2023), GPT4MTS (Jia et al., 2024), Time-LLM (Jin et al., 2024), ChatTime (Wang et al., 2025a), Time-VLM (Zhong et al., 2025), and Aurora (Wu et al., 2026). For the framework-level comparison, AFTER, MM-TSFLib (Liu et al., 2024a), and TaTS (Li et al., 2026) are instantiated with the same numerical backbone and evaluated under identical experimental settings, together with the corresponding backbone forecast. The three AFTER variants are further compared with the standalone multimodal methods to evaluate their overall predictive performance. We use the official implementations released by the authors for all baseline methods and follow their recommended configurations.

Evaluation. We evaluate forecasting performance using Mean Squared Error (MSE) and Mean Absolute Error (MAE) (Hyndman & Koehler, 2006), where lower values indicate better performance. All reported results are averaged over multiple random seeds. All compared methods use the same data splits, prediction horizons, and random seeds within each experiment. Detailed implementation settings are provided in Appendix D.

4.1 MAIN RESULTS

Table 1: Framework-level comparison on representative datasets and overall improvements over the numerical backbones. Each pair reports MSE/MAE averaged over four forecasting horizons.

(a) Framework comparison

Dataset	Backbone	AFTER	Base	MM-TSFLib	TaTS
Economy	iTransformer	**0.1930 / 0.3402**	0.2006 / 0.3483	0.1953 / 0.3415	0.1947 / 0.3426
	PatchTST	0.1984 / 0.3478	0.2031 / 0.3520	**0.1949 / 0.3439**	0.1988 / 0.3487
	TimeKAN	**0.2008 / 0.3500**	0.2065 / 0.3551	0.2107 / 0.3603	0.2092 / 0.3577
Social Good	iTransformer	**1.0931 / 0.4555**	1.2078 / 0.4940	1.1436 / 0.4821	1.1956 / 0.4967
	PatchTST	**1.0359 / 0.4725**	1.1216 / 0.5057	1.1069 / 0.4966	1.1397 / 0.5075
	TimeKAN	**1.0301 / 0.4931**	1.1615 / 0.5379	1.1513 / 0.5357	1.1819 / 0.5440
Traffic	iTransformer	**0.1957 / 0.2264**	0.2130 / 0.2346	0.2114 / 0.2374	0.2140 / 0.2377
	PatchTST	**0.1679 / 0.2152**	0.1756 / 0.2185	0.1768 / 0.2190	0.1751 / 0.2186
	TimeKAN	**0.1872 / 0.2580**	0.1992 / 0.2643	0.1990 / 0.2633	0.2006 / 0.2718

(b) Overall improvement

Backbone	Metric	Improved	Reduction
iTransformer	MSE	31/36 (86.1%)	+4.42%
	MAE	27/36 (75.0%)	+2.35%
PatchTST	MSE	32/36 (88.9%)	+2.43%
	MAE	32/36 (88.9%)	+1.75%
TimeKAN	MSE	32/36 (88.9%)	+3.16%
	MAE	30/36 (83.3%)	+2.00%
Overall	**MSE**	**95/108 (88.0%)**	**+3.34%**
	MAE	**89/108 (82.4%)**	**+2.03%**

Table 1(a) shows that AFTER remains effective with different numerical backbones and compares favorably with existing multimodal frameworks on the representative datasets. Moreover, the consistent improvements over the corresponding backbones in Table 1(b) demonstrate that inferring event-driven changes from text and using them to adjust numerical forecasts can effectively improve forecasting performance. On Environment, however, the adjusted predictions remain close to the backbone forecasts. Further analysis shows that the texts in this dataset are highly repetitive and rarely contain concrete evidence of future numerical changes. Consequently, HAER assigns most queries the *none* status and produces zero adjustments. This result reflects the conservative behavior of AFTER when reliable textual evidence is unavailable. Detailed framework-level results and further analysis of Environment are provided in Appendix M.1.

Table 2 compares the three AFTER variants with existing multimodal forecasting methods using results averaged over four forecasting horizons. Across diverse datasets, the AFTER variants consistently achieve leading performance and favorable average ranks for both MSE and MAE, demonstrating strong overall competitiveness under different backbone instantiations. Detailed results for individual forecasting horizons are provided in Appendix M.2.

Table 2: Comparison with multimodal forecasting methods across nine datasets. Each entry reports MSE/MAE averaged over four forecasting horizons. Lower values are better. The best results are shown in bold and the second-best results are underlined.

Dataset	AFTER (iTransformer)	AFTER (PatchTST)	AFTER (TimeKAN)	GPT4TS	GPT4MTS	ChatTime	Aurora	TimeLLM	TimeVLM
Agriculture	**0.1118/0.2350**	0.1139/0.2376	0.1234/0.2510	0.1407/0.2800	0.1344/0.2685	0.2105/0.2850	0.1879/0.3045	0.1537/0.2956	0.1536/0.2915
Climate	1.1464/0.8627	1.1724/0.8777	**1.0587/0.8367**	1.1378/0.8531	1.0987/0.8466	1.5377/1.0010	1.1984/0.8573	1.1698/0.8648	1.2845/0.9087
Economy	**0.1930/0.3402**	0.1984/0.3478	0.2008/0.3500	0.2451/0.4007	0.2353/0.3842	0.3171/0.4450	0.2987/0.4271	0.2815/0.4303	0.2607/0.4119
Energy	**0.0957/0.2065**	0.1017/0.2110	0.1002/0.2115	0.1104/0.2271	0.1337/0.2587	0.1430/0.2555	0.1758/0.3108	0.1085/0.2186	0.1095/0.2239
Environment	**0.2736/0.3766**	0.2894/0.3817	0.2881/0.3799	0.2881/0.3780	0.2906/0.3834	0.3503/0.4042	0.3768/0.4345	0.2901/0.3818	0.2901/0.3798
Health	1.1181/**0.6132**	1.1030/0.6552	**1.0762**/0.6528	1.1829/0.7091	1.1873/0.7020	1.6123/0.7734	2.0031/1.0172	1.1683/0.6825	1.2723/0.7335
Security	113.9105/5.4374	112.6219/5.3735	**107.2957**/4.9344	108.4874/**4.8636**	109.4047/4.9907	119.0982/5.7052	114.9462/4.9531	113.7648/5.3997	120.2729/5.6737
Social Good	1.0931/**0.4555**	1.0359/0.4725	1.0301/0.4931	**0.9523**/0.5106	1.5360/0.6313	1.2993/0.6687	1.4654/0.8104	1.0663/0.5445	1.0485/0.5739
Traffic	0.1957/0.2264	**0.1679/0.2152**	0.1872/0.2580	0.2123/0.2794	0.1895/0.2441	0.3088/0.3724	0.2152/0.3461	0.2517/0.3386	0.2562/0.3478
Average Rank	3.11/**2.69**	3.22/3.33	**2.39/2.69**	3.75/4.14	4.75/4.69	8.25/7.89	7.61/7.36	5.56/5.86	6.28/6.33

4.2 ABLATION STUDY

To examine the complementary roles of COR's three routes and the staged, base-anchored design, we conduct the comparisons in Table 3. Panel (a) starts from the numerical backbone and progressively adds contextual reasoning, case-based reasoning, and direct numerical transfer. Contextual reasoning helps on Climate and Security but not Economy; adding cases restores the improvement on Economy, and the complete three-route model achieves the strongest overall performance. A complete comparison of all single-route and two-route combinations is provided in Appendix E. Panel (b) compares AFTER with Direct Offset Reasoning and Direct Value Forecasting. Direct Offset Reasoning performs worse, indicating that deriving a history-aware event state before estimating offsets separates qualitative event interpretation from numerical judgment and yields more reliable adjustments. Direct Value Forecasting degrades further, showing that the LLM is better used to adjust a numerical forecast than to generate the complete future trajectory.

Table 3: Ablation results of AFTER. In (a), Base is the numerical backbone, w/o Memory uses only the contextual offset, and w/o Transfer combines contextual and case-based offsets without direct numerical transfer. In (b), iTransformer is used as the backbone: Direct Offset Reasoning estimates an offset relative to its forecast without HAER, whereas Direct Value Forecasting predicts the complete future sequence without a base forecast.

(a) Component ablation

Dataset	Backbone	AFTER	Base	w/o Memory	w/o Transfer
Climate	iTransformer	**1.1464 / 0.8627**	1.1553 / 0.8650	1.1498 / 0.8650	1.1531 / 0.8645
	PatchTST	**1.1724 / 0.8777**	1.1839 / 0.8811	1.1729 / 0.8786	1.1767 / 0.8789
	TimeKAN	1.0587 / 0.8367	1.0719 / 0.8411	**1.0567 / 0.8361**	1.0610 / 0.8374
Economy	iTransformer	**0.1930 / 0.3402**	0.2006 / 0.3483	0.2017 / 0.3488	0.1995 / 0.3466
	PatchTST	**0.1984 / 0.3478**	0.2031 / 0.3520	0.2064 / 0.3550	0.2026 / 0.3513
	TimeKAN	**0.2008 / 0.3500**	0.2065 / 0.3551	0.2095 / 0.3578	0.2061 / 0.3544
Security	iTransformer	**113.911 / 5.437**	115.479 / 5.595	114.926 / 5.525	114.483 / 5.504
	PatchTST	**112.622 / 5.374**	114.825 / 5.554	114.190 / 5.507	113.430 / 5.435
	TimeKAN	**107.296 / 4.934**	109.645 / 5.073	108.251 / 5.024	107.962 / 4.961

(b) Reasoning strategies

Dataset	Strategy	MSE/MAE
Climate	AFTER	**1.1464 / 0.8627**
	Direct Offset Reasoning	1.1812 / 0.8758
	Direct Value Forecasting	1.9391 / 1.1039
Economy	AFTER	**0.1930 / 0.3402**
	Direct Offset Reasoning	0.1998 / 0.3467
	Direct Value Forecasting	0.3686 / 0.4818
Security	AFTER	**113.911 / 5.437**
	Direct Offset Reasoning	115.854 / 5.638
	Direct Value Forecasting	119.075 / 5.560

4.3 ROLE OF TEXTUAL EVENT SEMANTICS

To test whether the gains of AFTER come from event information rather than the mere presence of text, we replace the original event texts with either "No text available." or static descriptions derived from the historical series. As shown in Table 4, event texts achieve the lowest MSE and MAE on all four datasets. With no text, the model lacks event evidence and falls back to the backbone forecast. Static descriptions mostly restate patterns already present in the numerical history and provide little evidence about external drivers of future changes. Their results are therefore close to the no-text condition. This shows that the improvements depend on the event semantics in the texts, not simply on supplying additional words.

Table 4: Effect of textual content on AFTER with iTransformer. Event Text uses the original aligned texts; No Text supplies "No text available."; Static Descriptive Text describes the historical numerical series without external events.

Dataset	Event Text	No Text	Static Descriptive Text
Security	**113.911 / 5.437**	115.479 / 5.595	115.481 / 5.593
Climate	**1.1464 / 0.8627**	1.1553 / 0.8650	1.1554 / 0.8651
Economy	**0.1930 / 0.3402**	0.2006 / 0.3483	0.2000 / 0.3482
Traffic	**0.1957 / 0.2264**	0.2130 / 0.2346	0.2129 / 0.2348

4.4 CASE STUDY OF TRACEABLE FORECAST ADJUSTMENT

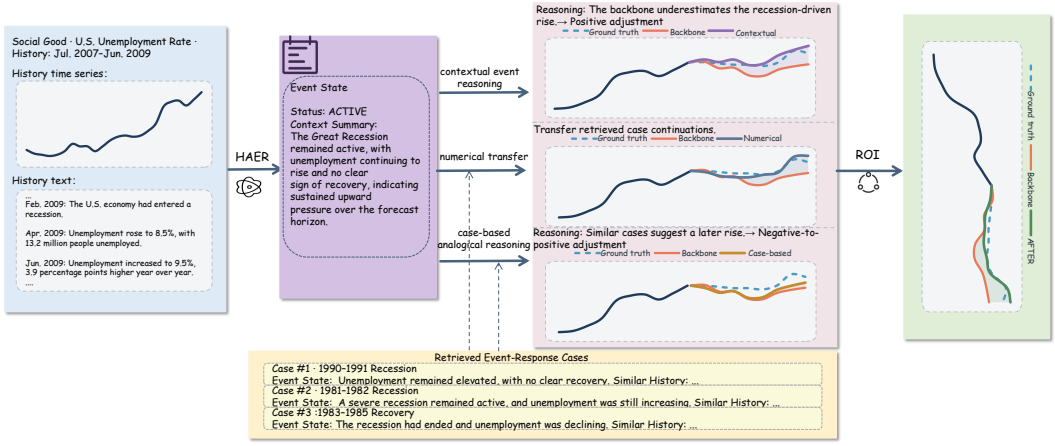


Figure 4: An illustrative example of the traceable forecast adjustment process of AFTER. The figure exposes the intermediate evidence and results at every stage, including the input history and texts, the event state inferred by HAER, the retrieved event-response cases, the reasoning and adjustment produced by each complementary route, and the final forecast obtained through robust integration.

Figure 4 presents an end-to-end evidence trail for an unemployment-rate forecast from the Social Good dataset. Although the backbone predicts an early decline, HAER identifies that the Great Recession remains active with no clear sign of recovery. Contextual reasoning therefore supports an upward correction, while the retrieved recession cases provide historical evidence for case-based reasoning and numerical transfer. ROI robustly integrates the three estimates, producing a trajectory closer to the ground truth and reducing the MSE from 0.195 to 0.024. By retaining the event state, retrieved cases, route-specific reasoning, and intermediate forecasts, the final adjustment can be directly traced.

5 CONCLUSION

In this paper, we introduced AFTER, an event-informed framework that complements trained numerical forecasting models with textual event reasoning without additional parameter optimization. Starting from a backbone forecast, AFTER grounds relevant events in the observed history, estimates forecast offsets through contextual reasoning, historical cases, and direct numerical transfer, and robustly integrates these complementary estimates. Experiments across nine datasets, multiple forecasting horizons, and three numerical backbones show that AFTER improves numerical forecasts in most settings and performs favorably against strong multimodal baselines, while retaining a traceable evidence trail for each adjustment. By assigning endogenous pattern modeling to numerical backbones and external-event interpretation to LLMs, AFTER establishes a practical and interpretable paradigm for combining numerical prediction with textual reasoning in multimodal time series forecasting.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing, translation, and code generation. We did not use generative AI tools for other research tasks requiring disclosure. All AI-assisted text was reviewed by the authors, and all AI-generated code was manually inspected and tested for correctness. We take full responsibility for the final content of this work, including any text or code produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only publicly available benchmark datasets containing aggregate time series and public texts. It does not involve human subjects or private individual-level information.

REPRODUCIBILITY STATEMENT

We provide an anonymized implementation of AFTER at `https://anonymous.4open.science/r/AFTER-ICLR--0638/`. The appendix details the datasets, preprocessing procedures, experimental settings, prompt templates, Event-Response Memory construction, and complete experimental results. Together, these materials support the reproduction and verification of all reported results.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Türkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda-Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Bernie Wang. Chronos: Learning the language of time series. *Trans. Mach. Learn. Res.*, 2024, 2024.

Si-An Chen, Chun-Liang Li, Nate Yoder, Sercan Ö. Arik, and Tomas Pfister. Tsmixer: An all-mlp architecture for time series forecasting. *CoRR*, abs/2303.06053, 2023. doi: 10.48550/ARXIV.2303.06053.

Tong Guan, Zijie Meng, Dianqi Li, Shiyu Wang, Chao-Han Huck Yang, Qingsong Wen, Zuozhu Liu, Sabato Siniscalchi, Ming Jin, and Shirui Pan. Timeomni-1: Incentivizing complex reasoning with time series in large language models. In *International Conference on Learning Representations*, volume 2026, pp. 152139–152170, 2026.

Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. The tabular foundation model tabpfn outperforms specialized time series forecasting models based on simple features. *CoRR*, abs/2501.02945, 2025. doi: 10.48550/ARXIV.2501.02945.

Yifan Hu, Peiyuan Liu, Peng Zhu, Dawei Cheng, and Tao Dai. Adaptive multi-scale decomposition framework for time series forecasting. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 17359–17367. AAAI Press, 2025. doi: 10.1609/AAAI.V39I16.33908.

Songtao Huang, Zhen Zhao, Can Li, and Lei Bai. Timekan: Kan-based frequency decomposition learning architecture for long-term time series forecasting. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2006.03.001.

Furong Jia, Kevin Wang, Yixiang Zheng, Defu Cao, and Yan Liu. GPT4MTS: prompt-based large language model for multimodal time-series forecasting. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI*

2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada, pp. 23343–23351. AAAI Press, 2024. doi: 10.1609/AAAI.V38I21.30383.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-llm: Time series forecasting by reprogramming large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Harshavardhan Kamarthi, Lingkai Kong, Alexander Rodríguez, Chao Zhang, and B. Aditya Prakash. When in doubt: Neural non-parametric uncertainty quantification for epidemic forecasting. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 19796–19807, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Zihao Li, Xiao Lin, Zhining Liu, Jiaru Zou, Ziwei Wu, Lecheng Zheng, Dongqi Fu, Yada Zhu, Hendrik Hamann, Hanghang Tong, et al. Language in the flow of time: Time-series-paired texts weaved into a unified temporal narrative. In *International Conference on Learning Representations*, volume 2026, pp. 24437–24484, 2026.

Bryan Lim and Stefan Zohren. Time series forecasting with deep learning: A survey. *CoRR*, abs/2004.13408, 2020.

Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Kamarthi, Aditya B. Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, and B. Aditya Prakash. Time-mmd: Multi-domain multimodal dataset for time series analysis. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 77888–77933. Curran Associates, Inc., 2024a. doi: 10.52202/079017-2476.

Peiyuan Liu, Hang Guo, Tao Dai, Naiqi Li, Jigang Bao, Xudong Ren, Yong Jiang, and Shu-Tao Xia. CALF: aligning llms for time series forecasting via cross-modal fine-tuning. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 18915–18923. AAAI Press, 2025. doi: 10.1609/AAAI.V39I18.34082.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 11116–11140, 2024b.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

Chenxi Sun, Hongyan Li, Yaliang Li, and Shenda Hong. TEST: text prototype aligned embedding to activate llm's ability for time series. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Chengsen Wang, Qi Qi, Jingyu Wang, Haifeng Sun, Zirui Zhuang, Jinming Wu, Lei Zhang, and Jianxin Liao. Chattime: A unified multimodal time series foundation model bridging numerical

and textual data. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 12694–12702. AAAI Press, 2025a. doi: 10.1609/AAAI.V39I12.33384.

Huiqiang Wang, Jian Peng, Feihu Huang, Jince Wang, Junhui Chen, and Yifei Xiao. MICN: multi-scale local and global context modeling for long-term series forecasting. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

Jiahao Wang, Mingyue Cheng, Qingyang Mao, Yitong Zhou, Daoyu Wang, Qi Liu, Feiyang Xu, and Xin Li. Tabletime: Reformulating time series classification as training-free table understanding with large language models, 2025b.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53140–53164. PMLR / OpenReview.net, 2024.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

Xingjian Wu, Jianxin Jin, Wanghui Qiu, Peng Chen, Yang Shu, Bin Yang, and Guo Guo. Aurora: Towards universal generative multimodal time series forecasting. In *International Conference on Learning Representations*, volume 2026, pp. 103239–103277, 2026.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 11121–11128. AAAI Press, 2023. doi: 10.1609/AAAI.V37I9.26317.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *CoRR*, abs/2506.05176, 2025. doi: 10.48550/ARXIV.2506.05176.

Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qingsong Wen, and Yuxuan Liang. Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative*

Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pp. 11106–11115. AAAI Press, 2021. doi: 10.1609/AAAI.V35I12.17325.

Tian Zhou, Ziqing Ma, Xue Wang, Qingsong Wen, Liang Sun, Tao Yao, Wotao Yin, and Rong Jin. Film: Frequency improved legendre memory model for long-term time series forecasting. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022a.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27268–27286. PMLR, 2022b.

Tian Zhou, Peisong Niu, Xue Wang, Liang Sun, and Rong Jin. One fits all: Power general time series analysis by pretrained LM. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Yitong Zhou, Yucong Luo, Mingyue Cheng, Qi Liu, Jiahao Wang, Daoyu Wang, and Enhong Chen. Time series forecasting via reasoning: A slow-thinking approach with reinforcement fine-tuned llms, 2026.

A Dataset Details

A.1 Dataset Description

We evaluate AFTER on nine multimodal time series datasets from Time-MMD (Liu et al., 2024a), covering diverse domains and temporal resolutions.

- **Agriculture:** Monthly U.S. retail broiler composite prices and aligned market reports from January 1983 to April 2024.
- **Climate:** Monthly precipitation measurements and aligned NOAA climate reports for the contiguous United States from January 1983 to April 2024.
- **Economy:** The monthly U.S. international trade balance from January 1989 to March 2024. The original reverse-chronological records are reordered by ascending timestamp.
- **Energy:** Weekly U.S. retail gasoline prices and aligned energy-market reports from January 1996 to April 2024.
- **Environment:** Daily Air Quality Index measurements and aligned environmental reports for the New York metropolitan area from January 1982 to September 2023.
- **Health:** Weekly U.S. influenza-like illness statistics and aligned public-health reports from September 1997 to May 2024.
- **Security:** Monthly FEMA disaster and emergency grant amounts and aligned disaster reports from September 1999 to May 2024.
- **Social Good:** The monthly U.S. unemployment rate and aligned labor-market reports from January 1950 to April 2024.
- **Traffic:** Monthly U.S. vehicle miles traveled and aligned traffic reports from January 1980 to March 2024.

A.2 Dataset Statistics and Forecasting Settings

Table 5 summarizes the temporal resolution, number of observations, time span, and forecasting configuration of each dataset. Since the datasets are sampled at different frequencies, using identical sequence and prediction lengths would correspond to substantially different real-world durations. We therefore adopt frequency-aware forecasting settings. For the six monthly datasets, we use a look-back length of 24 and prediction horizons of $\{6, 8, 10, 12\}$ months. For the weekly Energy and Health datasets, we use horizons of $\{4, 8, 12, 24\}$ weeks to cover forecasting periods from approximately one to six months. For the daily Environment dataset, we increase the look-back length to 48 and use horizons of $\{7, 14, 30, 48\}$ days to cover weekly and monthly air-quality dynamics.

Table 5: Dataset statistics and forecasting configurations. L, L_{label}, and H denote the look-back length, label length, and prediction horizons, respectively.

Dataset	Frequency	Time points	Time span	L	L_{label}	H
Agriculture	Monthly	496	1983–2024	24	12	$\{6, 8, 10, 12\}$
Climate	Monthly	496	1983–2024	24	12	$\{6, 8, 10, 12\}$
Economy	Monthly	423	1989–2024	24	12	$\{6, 8, 10, 12\}$
Energy	Weekly	1,479	1996–2024	24	12	$\{4, 8, 12, 24\}$
Environment	Daily	15,248	1982–2023	48	24	$\{7, 14, 30, 48\}$
Health	Weekly	1,389	1997–2024	24	12	$\{4, 8, 12, 24\}$
Security	Monthly	297	1999–2024	24	12	$\{6, 8, 10, 12\}$
Social Good	Monthly	892	1950–2024	24	12	$\{6, 8, 10, 12\}$
Traffic	Monthly	531	1980–2024	24	12	$\{6, 8, 10, 12\}$

A.3 TRAINING-ONLY MEMORY CONSTRUCTION.

To avoid the risk of information leakage, the Event-Response Memory for each experimental setting is constructed exclusively from the corresponding training split. Each memory record contains a training history, its event state inferred by HAER, and the subsequently observed training continuation. Validation and test samples are never inserted into or used to update the memory. During test inference, the memory remains fixed and is only queried to retrieve relevant training cases.

A.4 DATASET VISUALIZATION

Figure 5 presents the complete temporal evolution of the OT variable in each dataset. The series exhibit diverse characteristics, including long-term trends, recurring seasonal patterns, different levels of volatility, and abrupt changes. This diversity enables a comprehensive evaluation of forecasting methods across domains with substantially different temporal dynamics.



Figure 5: Overview of the OT variables in the nine datasets.

B BASELINES

Table 6 lists the official code repositories used to implement the baseline methods. We use the implementations released by the original authors. The data preprocessing pipeline and experimental protocol in the official MM-TSFLib implementation are used as a common reference (Liu et al., 2024a), with the dataset-specific modifications described in Appendix A.2. For MM-TSFLib (Liu et al., 2024a) and TaTS (Li et al., 2026), we use the same numerical backbones as AFTER to ensure a consistent framework-level comparison.

C METRICS

Mean Squared Error. Given N pairs of predicted and ground-truth values, MSE is defined as

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\widehat{y}_i - y_i)^2\,, \tag{12}$$

where $\widehat{y}_i$ and y_i denote the predicted and ground-truth values, respectively.

Table 6: Official code repositories for the baseline methods.

Category	Method	Official repository
Numerical forecasting	iTransformer	https://github.com/thuml/iTransformer
	PatchTST	https://github.com/yuqinie98/patchtst
	TimeKAN	https://github.com/huangst21/timekan
Backbone-compatible multimodal frameworks	MM-TSFLib	https://github.com/AdityaLab/MM-TSFlib
	TaTS	https://github.com/iDEA-iSAIL-Lab-UIUC/TaTS
Standalone multimodal forecasting	GPT4TS	https://github.com/DAMO-DI-ML/NeurIPS2023-One-Fits-All
	GPT4MTS	https://github.com/Flora-jia-jfr/GPT4MTS-Prompt-based-Large-Language-Model-for-Multimodal-Time-series-Forecasting
	Time-LLM	https://github.com/KimMeen/Time-LLM
	ChatTime	https://github.com/forestsking/chattime
	Time-VLM	https://github.com/CityMind-Lab/ICML25-TimeVLM
	Aurora	https://github.com/decisionintelligence/Aurora
	TimeOmni-1	https://github.com/AntonGuan/TimeOmni-1

Mean Absolute Error. MAE is defined as

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\widehat{y}_i - y_i| \,. \tag{13}$$

Cosine Similarity. Given two vectors $\mathbf{a}$ and $\mathbf{b}$, their cosine similarity is defined as

$$\mathrm{CosSim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\|_2 \, \|\mathbf{b}\|_2}. \tag{14}$$

Root Mean Squared Error. Given two D-dimensional vectors $\mathbf{u}$ and $\mathbf{v}$, RMSE is defined as

$$\mathrm{RMSE}(\mathbf{u}, \mathbf{v}) = \sqrt{\frac{1}{D}\sum_{j=1}^{D} (u_j - v_j)^2}. \tag{15}$$

D Experimental settings

The numerical backbones are optimized using Adam (Kingma & Ba, 2015) with a learning rate of 10^{-4} and a batch size of 32. Training is conducted for at most 10 epochs with an early-stopping patience of 5. DeepSeek-V4-Flash is used for textual event reasoning, while Qwen3-Embedding-0.6B is used to encode event-state summaries for memory retrieval (Zhang et al., 2025). The number of retrieved cases is set to $K = 3$. All reported results are averaged over multiple random seeds. Further implementation details and complete experimental configurations are provided in the accompanying anonymous code repository.

E Complete Route-Combination Ablation

To comprehensively examine the complementary roles of the three routes in COR, we evaluate all possible route combinations on Climate, Economy, and Security using iTransformer, PatchTST, and TimeKAN as numerical backbones. We denote contextual event reasoning, case-based analogical reasoning, and direct numerical transfer by C, M, and N, respectively. A single-route variant retains only the corresponding offset, while a two-route variant combines the two available offsets element-wise.

Table 7: Average ranks of different route combinations on Climate, Economy, and Security. C, M, and N denote contextual event reasoning, case-based analogical reasoning, and direct numerical transfer, respectively.

Routes	iTransformer	PatchTST	TimeKAN	Overall
Base	4.92 / 5.08	5.08 / 5.08	5.17 / 5.25	5.06 / 5.14
C	4.42 / 4.67	4.58 / 4.83	4.00 / 4.42	4.33 / 4.64
M	5.00 / 4.67	4.25 / 3.92	4.50 / 3.92	4.58 / 4.17
N	6.00 / 6.00	6.17 / 6.00	6.42 / 6.00	6.19 / 6.00
C+M	3.58 / 3.42	3.67 / 3.83	3.83 / 3.92	3.69 / 3.72
C+N	5.17 / 5.25	5.17 / 5.17	5.08 / 5.08	5.14 / 5.17
M+N	4.83 / 4.75	4.67 / 4.83	4.75 / 4.92	4.75 / 4.83
C+M+N	**2.08 / 2.17**	**2.42 / 2.33**	**2.25 / 2.50**	**2.25 / 2.33**

As shown in Table 7, the complete three-route model achieves the best average MSE and MAE ranks under all three numerical backbones. Contextual and case-based reasoning provide useful but incomplete evidence, while their combination yields more stable rankings. Direct numerical transfer is less reliable when used alone because retrieved continuations may not be fully transferable, but it provides complementary numerical evidence when robustly integrated with the two reasoning routes. The consistently leading ranks of C+M+N support both the complementary route design and the use of ROI for robust integration.

F Inference-Time Analysis

Table 8: Average per-sample inference time of the two branches in AFTER, with a breakdown of the Event-Informed Adjustment Branch.

Component	Time (s/sample)
Numerical Forecasting Branch	0.0018
Event-Informed Adjustment Branch	
History-Aware Event Reasoning	2.7098
Contextual Event Reasoning	2.3163
Case-Based Analogical Reasoning	2.5929
Retrieval, numerical transfer, and ROI	0.1002
Total	7.7192

We compare the average per-sample inference time of the numerical forecasting branch and the event-informed adjustment branch in AFTER. As shown in Table 8, the Event-Informed Adjustment Branch requires approximately 7.72 seconds per sample, with 98.7% of its latency arising from the three LLM reasoning stages. In contrast, memory retrieval, direct numerical transfer, and ROI together require only about 0.10 seconds. The numerical branch is substantially faster, completing a backbone forward pass in approximately 1.8 milliseconds.

This additional latency is an inherent cost of using LLMs to interpret external events, compare historical responses, and retain a traceable reasoning process. However, AFTER does not require any additional parameter optimization: it operates directly on a trained numerical backbone and introduces event reasoning only during inference. Moreover, the datasets considered in this work are sampled at daily, weekly, or monthly intervals. Relative to these sampling frequencies, a delay of several seconds remains small and does not prevent the forecast from being produced before the next observation becomes available. For applications with strict real-time constraints, the trained numerical backbone can still be deployed independently, since AFTER adjusts its predictions without modifying the backbone itself.

G Analysis of Robust Offset Integration

Robust Offset Integration (ROI) combines the offsets produced by contextual event reasoning, case-based analogical reasoning, and direct numerical transfer. To examine the choice of the element-wise median, we compare it with Mean Integration and Conservative Integration.

Integration rules. Mean Integration calculates the element-wise arithmetic mean. Conservative Integration selects the estimate with the smallest absolute magnitude at each forecast step and variable:

Overall comparison. Table 9 compares the three integration rules over all nine datasets and forecasting horizons.

Table 9: Overall comparison of offset integration rules across nine datasets.

Backbone	Integration	MSE	MAE
iTransformer	Mean	13.7598	0.9810
	Conservative	13.3182	0.9954
	Median	**13.1264**	**0.9726**
PatchTST	Mean	13.6063	0.9857
	Conservative	13.2069	0.9972
	Median	**12.9783**	**0.9747**
TimeKAN	Mean	12.9445	0.9544
	Conservative	12.5943	0.9488
	Median	**12.3734**	**0.9297**

The median achieves the best overall MSE and MAE under all three backbones. Mean integration can be affected by an abnormal route-specific estimate, whereas conservative integration may suppress useful adjustments by consistently selecting the smallest offset. The median balances these two behaviors and provides consistent performance across different numerical forecasting architectures.

Table 10: Comparison of offset integration rules on representative datasets.

Dataset	Backbone	Mean	Conservative	Median
Climate	iTransformer	1.1885 / 0.8845	1.1527 / 0.8641	**1.1464 / 0.8627**
	PatchTST	1.2149 / 0.8966	1.1813 / 0.8804	**1.1724 / 0.8777**
	TimeKAN	1.1310 / 0.8672	1.0682 / 0.8401	**1.0587 / 0.8367**
Energy	iTransformer	0.0969 / 0.2070	0.0959 / 0.2070	**0.0957 / 0.2065**
	PatchTST	0.1024 / 0.2116	0.1019 / 0.2116	**0.1017 / 0.2110**
	TimeKAN	0.1014 / 0.2121	0.1003 / 0.2117	**0.1002 / 0.2115**
Security	iTransformer	119.6634 / 5.5394	115.4201 / 5.5805	**113.9105 / 5.4374**
	PatchTST	118.2805 / 5.4936	114.5698 / 5.5302	**112.6219 / 5.3735**
	TimeKAN	112.4333 / 5.1867	109.1409 / 5.0492	**107.2957 / 4.9344**

Representative datasets. To further examine the robustness of different integration rules under diverse event-response patterns, we report detailed results on Climate, Energy, and Security in Table 10. The median consistently achieves the best results across all three backbones. Its advantage is more evident on Climate and Security, while the smaller but consistent improvements on Energy show that it remains effective when the differences among the three offset estimates are limited.

H Analysis of Dual-Similarity Retrieval

To examine the roles of event semantics and historical dynamics in case retrieval, we compare the complete dual-similarity retrieval with two variants that remove textual similarity or historical-

shape similarity, respectively. Memory retrieval is applied only to samples whose HAER status is *weak* or *active*, which we refer to as Memory-eligible samples; samples with a *none* status remain unchanged across retrieval variants. We therefore report results on both the full test set and the Memory-eligible subset: the former measures the effect on overall forecasting performance, while the latter isolates samples for which retrieval is actually performed and more directly evaluates the retrieval mechanism.

Table 11: Ablation of the dual-similarity retrieval mechanism using iTransformer as the numerical backbone. The complete method combines textual and historical-shape similarities, while each variant removes one of them.

Evaluation Set	Dataset	Dual Similarity	w/o Text Similarity	w/o Shape Similarity
All Test Samples	Climate	**1.1464 / 0.8627**	1.1485 / 0.8638	1.1497 / **0.8627**
	Economy	**0.1930 / 0.3402**	0.1936 / 0.3411	0.1944 / 0.3429
	Traffic	**0.1957 / 0.2264**	0.1959 / 0.2266	0.1970 / 0.2282
Memory-Eligible	Climate	**1.1473 / 0.8638**	1.1493 / 0.8649	1.1508 / 0.8639
	Economy	**0.2925 / 0.4290**	0.2936 / 0.4309	0.2957 / 0.4347
	Traffic	**0.2698 / 0.3544**	0.2705 / 0.3560	0.2757 / 0.3618

As shown in Table 11, dual-similarity retrieval generally achieves the best performance under both evaluation scopes, with clearer differences on Memory-eligible samples. Removing textual similarity allows numerically similar but semantically mismatched cases to enter the retrieval results, whereas removing shape similarity may retrieve events occurring under different historical dynamics. The larger degradation after removing shape similarity on Economy and Traffic further shows that similar event semantics alone do not guarantee transferable numerical responses. Together, the two signals provide complementary evidence for retrieving relevant historical cases.

I Sensitivity to the Number of Retrieved Cases

To examine the effect of the number of retrieved cases, we vary K among $\{1, 3, 5, 10\}$ while keeping the remaining settings unchanged. Top-3 is the default setting used in the main experiments. We report results on the full test set to measure the overall forecasting effect and on the Memory-eligible subset to isolate the samples for which retrieval is actually performed.

Table 12: Sensitivity to the number of retrieved cases using iTransformer as the numerical backbone. Top-3 is the default setting.

Evaluation Set	Dataset	Top-1	**Top-3**	Top-5	Top-10
All Test Samples	Agriculture	0.1121 / 0.2356	**0.1118** / 0.2350	0.1120 / 0.2353	0.1120 / **0.2346**
	Economy	0.1964 / 0.3441	**0.1930 / 0.3402**	0.1949 / 0.3425	0.1953 / 0.3432
Memory-Eligible	Agriculture	0.1161 / 0.2416	**0.1158** / 0.2410	0.1160 / 0.2413	0.1160 / **0.2405**
	Economy	0.2998 / 0.4373	**0.2925 / 0.4290**	0.2966 / 0.4338	0.2975 / 0.4353

As shown in Table 12, increasing K from one to three generally improves forecasting performance, showing that multiple cases provide more reliable historical evidence than a single example. Further increasing K often brings no additional improvement; when gains do occur, they are generally small, while lower-ranked and less relevant cases may instead degrade performance. This pattern becomes clearer on the Memory-eligible subset, where retrieval directly affects every sample. Overall, Top-3 provides the most stable balance between evidence coverage and retrieval noise and is therefore adopted as the default setting.

J Analysis of Time-Series Representation

To examine how the time series should be presented to the LLM, we compare three representations. The proposed representation provides the first observation as a reference and expresses the subse-

quent values as relative changes in a timestamped table. The other two representations present the sequence values directly, either in a timestamped table or as an ordered value list.

Table 13: Comparison of time-series representations using iTransformer as the numerical backbone. Reference + Relative Table is the proposed representation; Value Table presents the sequence values directly with timestamps; Value List presents the same values as an ordered list.

Dataset	Reference + Relative Table	Value Table	Value List
Economy	**0.1930 / 0.3402**	0.1973 / 0.3453	0.1964 / 0.3439
Social Good	**1.0931 / 0.4555**	1.1210 / 0.4594	1.1809 / 0.4725

As shown in Table 13, the two direct-value representations show no consistent advantage over each other, indicating that changing the serialization format alone is insufficient. Introducing a reference value and explicitly expressing relative changes, however, consistently reduces forecasting errors. This representation exposes the direction and magnitude of temporal changes, helping the LLM better understand the time series and relate textual events to its observed evolution.

K Robustness to Different LLM Backends

To examine whether AFTER depends on a specific LLM, we replace the reasoning backend with Qwen3.8-Flash and GLM-5.3-Flash while keeping the remaining settings unchanged. All methods use iTransformer as the numerical backbone.

Table 14: Comparison of different LLM reasoning backends with iTransformer as the numerical backbone. Base denotes the numerical forecast without event-informed adjustment.

Dataset	Base	DeepSeek-v4-Flash	Qwen3.8-Flash	GLM-5.3-Flash
Agriculture	0.1124 / 0.2345	0.1118 / 0.2350	0.0994 / 0.2228	0.1071 / 0.2314
Economy	0.2006 / 0.3483	0.1930 / 0.3402	0.1831 / 0.3329	0.1787 / 0.3252
Traffic	0.2130 / 0.2346	0.1957 / 0.2264	0.1917 / 0.2323	0.1608 / 0.2105

As shown in Table 14, all three LLM backends improve forecasting performance over the numerical backbone across the selected datasets. Although the magnitude of the gains varies across reasoning models, the event-informed adjustment remains effective with different LLM APIs. This indicates that AFTER is not tied to a particular LLM and can accommodate different reasoning backends.

L Verification against Information Leakage

To verify that the improvements of AFTER are not caused by information leakage, we examine three possible channels: future information contained in the textual inputs, test-set information introduced through the Event-Response Memory, and identifying cues in the prompts—including absolute dates, geographic and institutional names, and named historical events—that could allow the LLM to identify a specific historical episode and recall its subsequent outcome.

No future information in textual inputs. Time-MMD (Liu et al., 2024a) explicitly separates factual content from predictions during textual data construction and stores them in different fields. We exclusively use the extracted factual texts and exclude all prediction fields. Each text is also associated with verified start and end timestamps. Following the protocol of MM-TSFLib, the latest end time of the input texts is constrained not to exceed the end of the observed input window. Therefore, neither future texts nor future numerical observations are available when forecasting a test sample.

Training-only Event-Response Memory. The Event-Response Memory is constructed exclusively from the training split. Each record contains a training history, its event state inferred by HAER, and the corresponding continuation observed within the training set. Validation and test

samples are never inserted into or used to update the memory. During test inference, the memory remains fixed and only retrieves cases from the training set. Consequently, the retrieved continuations cannot reveal the future values of the current test sample.

Verification under de-identified inputs. A remaining concern is that the LLM may identify well-known historical episodes from absolute dates and entity names and recall their subsequent outcomes. To examine this possibility, we conduct a de-identified evaluation on Economy and Social Good using iTransformer as the numerical backbone. All absolute timestamps are replaced with relative positions, such as $t = -23, \ldots, 0, 1, \ldots, H$. Dataset, country, institution, and named-event identifiers are replaced with typed placeholders, while event mechanisms and temporal order are retained. The same transformation is applied during both training- memory construction and test inference, and HAER and all subsequent reasoning stages are rerun using the de-identified inputs.

Table 15: Forecasting performance with de-identified textual inputs using iTransformer as the numerical backbone.

Dataset	Base	Original AFTER	De-identified AFTER
Economy	0.2006 / 0.3483	0.1930 / 0.3402	**0.1925 / 0.3399**
Social Good	1.2078 / 0.4940	1.0931 / 0.4555	**1.0834 / 0.4517**

As shown in Table 15, removing absolute timestamps and identifying information does not reduce forecasting performance. De-identified AFTER reduces MSE and MAE over the numerical backbone by 4.05% and 2.41% on Economy, and by 10.30% and 8.55% on Social Good, respectively. It even slightly outperforms the original version, possibly because removing identity- specific cues helps the LLM focus on transferable event mechanisms. These results indicate that the improvements arise from reasoning over the provided event mechanisms and historical dynamics, rather than relying on memorized outcomes associated with identifiable historical episodes.

M DETAILED MAIN EXPERIMENT RESULTS

This section provides the detailed results of the two comparisons in our main experiments: the framework-level comparison under shared numerical backbones and the overall comparison with standalone multimodal forecasting methods.

M.1 FRAMEWORK-LEVEL COMPARISON

The framework-level comparison examines the effectiveness and generalizability of AFTER across diverse datasets and numerical forecasting backbones. We instantiate AFTER, MM-TSFLib, and TaTS with the same backbone and compare them with the corresponding numerical backbone under identical experimental settings. The comparison is conducted separately using iTransformer, PatchTST, and TimeKAN as the numerical backbone.

Overall framework comparison. Tables 17, 18, and 19 show that AFTER achieves the best average rank for both MSE and MAE under all three numerical backbones. With iTransformer, AFTER obtains average ranks of 1.61 and 1.75 for MSE and MAE, respectively. The corresponding ranks are 1.81 and 1.81 with PatchTST, and 1.47 and 1.58 with TimeKAN. In comparison, MM-TSFLib is consistently the second-ranked framework on average. The leading ranks achieved by AFTER across different backbone architectures demonstrate that its effectiveness is not limited to a particular numerical forecasting model. Beyond forecasting accuracy, AFTER provides a more transparent adjustment process than methods that primarily incorporate textual information through latent representations. For each forecast, AFTER retains the inferred event state, retrieved historical cases, route-specific offset estimates, and the final integrated adjustment. This evidence trail makes it possible to trace how textual events contribute to an individual numerical forecast.

Improvement over numerical backbones. Tables 20, 21, and 22 further evaluate the adjustments made by AFTER to the corresponding backbone forecasts. With iTransformer, AFTER improves MSE in 31 of the 36 settings and MAE in 27 settings. With PatchTST, both MSE and MAE are

improved in 32 settings. With TimeKAN, AFTER improves MSE in 32 settings and MAE in 30 settings. Across all three backbones, AFTER achieves lower MSE in 95 of the 108 settings and lower MAE in 89 of them, corresponding to improvement rates of 88.0% and 82.4%, respectively. The average relative reductions in MSE and MAE are 3.34% and 2.03%.

Dataset-level observations. The benefits of AFTER are particularly consistent on Economy, Security, Social Good, and Traffic, where it improves both MSE and MAE over all three backbones and four forecasting horizons. On Social Good, the average MSE reduction ranges from 7.75% to 11.43% across the three backbones, while the MAE reduction ranges from 6.66% to 8.44%. On Traffic, the corresponding reductions are 4.47%–8.16% for MSE and 1.56%–3.46% for MAE. Consistent improvements are also observed on Economy and Security. Environment exhibits a different

Table 16: Text redundancy and HAER context-state distribution on Environment. Text statistics are calculated over the test interval and its look-back window.

Text statistics

Metric	Value
Timestamped texts	3,097
Unique texts	441
Unique-text rate	14.24%
Duplicate rate	85.76%

HAER context-state distribution

H	Queries	*none*	*weak*	*active*
7	3,040	94.18%	5.26%	0.56%
14	3,008	94.35%	5.05%	0.60%
30	3,008	93.95%	5.45%	0.60%
48	2,976	94.49%	4.97%	0.54%
Overall	**12,032**	**94.24%**	**5.19%**	**0.57%**

pattern, where the adjusted forecasts remain close to the original backbone forecasts. As shown in Table 16, 85.76% of the timestamped texts in the test interval are exact duplicates, and HAER assigns the *none* status to 94.24% of the test queries. Only 5.19% and 0.57% of the queries are categorized as *weak* and *active*, respectively. The high text redundancy and the dominance of the *none* status indicate that the textual context rarely provides concrete evidence for future air-quality changes. Consequently, AFTER largely preserves the backbone forecasts rather than introducing unsupported adjustments.

Dependence on backbone quality. A small number of settings do not achieve the best performance despite improving the corresponding backbone forecast. This is partly attributable to the variability of the numerical backbone across random seeds. In several cases, the backbone produces substantially larger errors for particular seeds than for the remaining runs. Because AFTER estimates a residual adjustment anchored to the backbone forecast rather than replacing it entirely, the inferred offset can reduce such errors but may not fully compensate for an unusually inaccurate base forecast. This dependence on backbone quality represents an inherent limitation of the adjustment-based formulation.

M.2 Overall Comparison with Multimodal Forecasting Methods

While the preceding framework-level comparison controls the numerical backbone, we further compare the overall forecasting performance of AFTER with standalone multimodal forecasting methods. We report three instantiations of AFTER, using iTransformer, PatchTST, and TimeKAN as the numerical backbone, respectively, and compare them with GPT4TS, GPT4MTS, ChatTime, Aurora, TimeLLM, and TimeVLM. Table 25 reports the MSE and MAE results over all datasets and forecasting horizons. Each result is averaged over the evaluated random seeds, while the *Avg.* row reports the mean over four horizons.

The three AFTER variants occupy the top three average positions for both metrics. AFTER(TimeKAN) achieves the best average MSE rank of 2.39, followed by AFTER(iTransformer) with 3.11. For MAE, AFTER(iTransformer) and AFTER(TimeKAN) share the best average rank of 2.69, while AFTER(PatchTST) ranks second with 3.33. Together, these results demonstrate the strong overall competitiveness of AFTER against existing multimodal forecasting methods across different backbone choices.

Table 17: Framework-level comparison using iTransformer as the numerical backbone. Each entry reports MSE/MAE averaged over multiple random seeds. The *Avg.* row reports the mean over four forecasting horizons. Lower values are better. The best results are shown in bold and the second-best results are underlined.

Dataset	H	AFTER	iTransformer	MM-TSFLib	TaTS
Agriculture	6	0.0651 / 0.1779	**0.0644 / 0.1762**	0.0683 / 0.1815	0.0677 / 0.1789
	8	**0.0895** / 0.2079	0.0897 / **0.2068**	0.0964 / 0.2184	0.0934 / 0.2111
	10	0.1296 / 0.2588	0.1301 / 0.2586	0.1281 / **0.2545**	**0.1272** / 0.2602
	12	0.1628 / 0.2955	0.1652 / 0.2964	0.1631 / 0.2926	**0.1601 / 0.2905**
	Avg.	**0.1118** / 0.2350	0.1124 / **0.2345**	0.1140 / 0.2368	0.1121 / 0.2352
Climate	6	1.1122 / 0.8461	1.1260 / 0.8468	**1.0997 / 0.8347**	1.1663 / 0.8669
	8	1.1435 / 0.8602	1.1520 / 0.8638	**1.1168 / 0.8459**	1.1863 / 0.8722
	10	1.1536 / 0.8645	1.1598 / 0.8672	**1.1375 / 0.8594**	1.2055 / 0.8867
	12	1.1763 / 0.8800	1.1835 / 0.8821	**1.1166 / 0.8520**	1.1830 / 0.8830
	Avg.	1.1464 / 0.8627	1.1553 / 0.8650	**1.1176 / 0.8480**	1.1853 / 0.8772
Economy	6	**0.1647 / 0.3142**	0.1743 / 0.3251	0.1667 / 0.3201	0.1658 / 0.3208
	8	0.1851 / 0.3350	0.1930 / 0.3432	0.1868 / 0.3338	**0.1837 / 0.3325**
	10	**0.2016 / 0.3469**	0.2078 / 0.3537	0.2045 / 0.3486	0.2108 / 0.3553
	12	0.2204 / 0.3646	0.2274 / 0.3714	0.2230 / 0.3633	**0.2185 / 0.3617**
	Avg.	**0.1930 / 0.3402**	0.2006 / 0.3483	0.1953 / 0.3415	0.1947 / 0.3426
Energy	4	**0.0247 / 0.1059**	0.0257 / 0.1079	0.0322 / 0.1274	0.0279 / 0.1144
	8	**0.0624 / 0.1734**	0.0647 / 0.1773	0.0653 / 0.1839	0.0654 / 0.1788
	12	**0.0973 / 0.2187**	0.0981 / **0.2187**	0.1055 / 0.2289	0.1055 / 0.2261
	24	0.1986 / 0.3281	0.1970 / 0.3274	**0.1902 / 0.3256**	0.2164 / 0.3406
	Avg.	**0.0957 / 0.2065**	0.0964 / 0.2078	0.0983 / 0.2164	0.1038 / 0.2150
Environment	7	0.2698 / 0.3717	**0.2690** / 0.3712	0.2700 / 0.3719	0.2735 / **0.3711**
	14	0.2761 / 0.3770	0.2762 / 0.3770	**0.2702 / 0.3720**	0.2770 / 0.3754
	30	0.2713 / 0.3757	0.2710 / 0.3755	**0.2679 / 0.3748**	0.2764 / 0.3777
	48	0.2770 / 0.3820	0.2769 / 0.3819	**0.2721 / 0.3786**	0.2844 / 0.3866
	Avg.	0.2736 / 0.3766	0.2733 / 0.3764	**0.2701 / 0.3743**	0.2778 / 0.3777
Health	4	0.4131 / 0.3851	0.4855 / 0.4057	**0.4042 / 0.3790**	0.5113 / 0.4101
	8	**0.7474 / 0.5249**	0.9734 / 0.5567	0.8832 / 0.5477	0.9308 / 0.5581
	12	1.2391 / 0.6734	1.4774 / 0.7007	**1.1797 / 0.6697**	1.2918 / 0.7041
	24	2.0728 / 0.8694	2.1124 / 0.8770	**1.9286 / 0.8484**	1.9802 / 0.8777
	Avg.	1.1181 / 0.6132	1.2622 / 0.6350	**1.0989 / 0.6112**	1.1785 / 0.6375
Security	6	**113.2056 / 5.3958**	114.8335 / 5.5974	114.7177 / 5.4952	114.7815 / 5.6168
	8	**114.3334 / 5.4868**	116.0226 / 5.6560	116.6433 / 5.6536	116.4235 / 5.6203
	10	**113.2471 / 5.3701**	115.1472 / 5.5425	114.7239 / 5.4686	116.1852 / 5.6143
	12	**114.8561 / 5.4971**	115.9119 / 5.5846	116.9027 / 5.6198	118.7691 / 5.7790
	Avg.	**113.9105 / 5.4374**	115.4788 / 5.5951	115.7469 / 5.5593	116.5398 / 5.6576
Social Good	6	**0.9607 / 0.3931**	1.0531 / 0.4257	1.0178 / 0.4181	1.0627 / 0.4300
	8	**1.0430 / 0.4347**	1.1684 / 0.4762	1.1215 / 0.4681	1.1785 / 0.4737
	10	**1.1678 / 0.4893**	1.3000 / 0.5313	1.2196 / 0.5114	1.2691 / 0.5308
	12	**1.2010 / 0.5051**	1.3098 / 0.5426	1.2153 / 0.5310	1.2721 / 0.5522
	Avg.	**1.0931 / 0.4555**	1.2078 / 0.4940	1.1436 / 0.4821	1.1956 / 0.4967
Traffic	6	**0.1821 / 0.2174**	0.2007 / 0.2257	0.2001 / 0.2294	0.2038 / 0.2297
	8	**0.1901 / 0.2285**	0.2123 / 0.2381	0.2060 / 0.2380	0.2122 / 0.2421
	10	**0.1997 / 0.2404**	0.2175 / 0.2493	0.2196 / 0.2555	0.2179 / 0.2512
	12	**0.2109 / 0.2193**	0.2214 / 0.2251	0.2199 / 0.2268	0.2219 / 0.2277
	Avg.	**0.1957 / 0.2264**	0.2130 / 0.2346	0.2114 / 0.2374	0.2140 / 0.2377
Average Rank (MSE/MAE)		**1.61 / 1.75**	2.97 / 2.72	2.19 / 2.14	3.19 / 3.28

Table 18: Framework-level comparison using PatchTST as the numerical backbone. Each entry reports MSE/MAE averaged over multiple random seeds. The *Avg.* row reports the mean over four forecasting horizons. Lower values are better. The best results are shown in bold and the second-best results are underlined.

Dataset	H	AFTER	PatchTST	MM-TSFLib	TaTS
Agriculture	6	0.0695 / 0.1847	**0.0692** / 0.1836	0.0718 / 0.1928	0.0695 / **0.1813**
	8	**0.0952** / **0.2180**	0.0959 / **0.2180**	0.1017 / 0.2293	0.0984 / 0.2214
	10	0.1304 / 0.2580	0.1329 / 0.2604	0.1270 / 0.2566	**0.1255** / **0.2524**
	12	0.1605 / 0.2898	0.1631 / 0.2911	0.1641 / 0.2911	**0.1601** / **0.2856**
	Avg.	0.1139 / 0.2376	0.1153 / 0.2383	0.1162 / 0.2425	**0.1134** / **0.2352**
Climate	6	1.1801 / 0.8793	1.1930 / 0.8815	**1.1341** / **0.8493**	1.1613 / 0.8614
	8	1.1829 / 0.8808	1.1930 / 0.8848	**1.1362** / **0.8618**	1.1836 / 0.8842
	10	1.1638 / 0.8767	1.1746 / 0.8804	**1.1307** / **0.8601**	1.1808 / 0.8792
	12	1.1629 / 0.8740	1.1751 / 0.8777	**1.1307** / **0.8642**	1.1716 / 0.8795
	Avg.	1.1724 / 0.8777	1.1839 / 0.8811	**1.1330** / **0.8588**	1.1743 / 0.8761
Economy	6	0.1717 / 0.3199	0.1738 / 0.3223	0.1647 / **0.3129**	**0.1634** / 0.3131
	8	0.1889 / 0.3409	0.1917 / 0.3431	**0.1823** / **0.3327**	0.1923 / 0.3434
	10	0.2080 / 0.3583	0.2150 / 0.3652	**0.2071** / **0.3568**	0.2198 / 0.3707
	12	0.2250 / 0.3721	0.2320 / 0.3775	0.2256 / 0.3732	**0.2199** / **0.3677**
	Avg.	0.1984 / 0.3478	0.2031 / 0.3520	**0.1949** / **0.3439**	0.1988 / 0.3487
Energy	4	**0.0240** / **0.1054**	0.0246 / 0.1073	0.0321 / 0.1276	0.0277 / 0.1135
	8	**0.0622** / 0.1719	0.0631 / 0.1740	0.0650 / 0.1815	0.0624 / **0.1702**
	12	**0.1066** / 0.2286	0.1078 / 0.2293	0.1084 / 0.2353	0.1080 / **0.2283**
	24	0.2141 / **0.3382**	0.2136 / 0.3386	**0.2125** / 0.3432	0.2148 / 0.3391
	Avg.	**0.1017** / **0.2110**	0.1023 / 0.2123	0.1045 / 0.2219	0.1032 / 0.2128
Environment	7	0.2709 / **0.3694**	0.2700 / 0.3708	**0.2656** / 0.3706	0.2713 / 0.3700
	14	0.2832 / 0.3795	0.2814 / 0.3786	**0.2755** / **0.3749**	0.2811 / 0.3792
	30	0.2906 / 0.3834	0.2909 / 0.3841	**0.2845** / **0.3794**	0.2927 / 0.3866
	48	0.3127 / 0.3946	0.3129 / 0.3943	**0.2981** / **0.3899**	0.3125 / 0.3970
	Avg.	0.2894 / 0.3817	0.2888 / 0.3819	**0.2809** / **0.3787**	0.2894 / 0.3832
Health	4	0.4816 / 0.4398	0.4949 / 0.4492	**0.4543** / **0.4261**	0.4962 / 0.4427
	8	0.8320 / 0.5711	0.8593 / 0.5806	**0.7842** / **0.5611**	0.8815 / 0.5872
	12	1.2080 / 0.7147	1.2476 / 0.7269	**1.1437** / **0.6911**	1.2360 / 0.7139
	24	1.8905 / 0.8953	1.9151 / 0.9040	**1.7137** / **0.8682**	1.8385 / 0.8902
	Avg.	1.1030 / 0.6552	1.1292 / 0.6652	**1.0240** / **0.6366**	1.1131 / 0.6585
Security	6	**108.8765** / **5.2163**	111.6659 / 5.3930	111.1097 / 5.3233	111.7296 / 5.3855
	8	**112.4414** / **5.3524**	114.5890 / 5.5678	113.7669 / 5.5408	113.1674 / 5.4522
	10	**113.1932** / **5.3716**	115.4210 / 5.5642	114.7072 / 5.4771	116.4950 / 5.6072
	12	**115.9766** / **5.5537**	117.6240 / 5.6922	118.8750 / 5.7686	117.5928 / 5.6296
	Avg.	**112.6219** / **5.3735**	114.8250 / 5.5543	114.6147 / 5.5274	114.7462 / 5.5186
Social Good	6	**0.8686** / **0.3968**	0.9557 / 0.4310	0.9572 / 0.4276	0.9959 / 0.4381
	8	**0.9758** / **0.4464**	1.0721 / 0.4817	1.0612 / 0.4686	1.0783 / 0.4796
	10	**1.0814** / **0.4992**	1.1537 / 0.5292	1.1474 / 0.5194	1.2015 / 0.5363
	12	**1.2180** / **0.5478**	1.3048 / 0.5809	1.2620 / 0.5708	1.2832 / 0.5759
	Avg.	**1.0359** / **0.4725**	1.1216 / 0.5057	1.1069 / 0.4966	1.1397 / 0.5075
Traffic	6	**0.1545** / **0.2109**	0.1607 / 0.2134	0.1618 / 0.2136	0.1611 / 0.2143
	8	**0.1532** / **0.2126**	0.1631 / 0.2150	0.1672 / 0.2191	0.1644 / 0.2152
	10	**0.1651** / **0.2201**	0.1743 / 0.2255	0.1740 / 0.2248	0.1712 / 0.2261
	12	**0.1988** / **0.2170**	0.2043 / 0.2204	0.2041 / 0.2186	0.2035 / 0.2188
	Avg.	**0.1679** / **0.2152**	0.1756 / 0.2185	0.1768 / 0.2190	0.1751 / 0.2186
Average Rank (MSE/MAE)		**1.81** / **1.81**	3.08 / 3.17	2.19 / 2.17	2.86 / 2.75

Table 19: Framework-level comparison using TimeKAN as the numerical backbone. Each entry reports MSE/MAE averaged over multiple random seeds. The *Avg.* row reports the mean over four forecasting horizons. Lower values are better. The best results are shown in bold and the second-best results are underlined.

Dataset	H	AFTER	TimeKAN	MM-TSFLib	TaTS
Agriculture	6	**0.0775** / 0.1906	0.0781 / **0.1905**	0.0832 / 0.2067	0.0811 / 0.1971
	8	**0.1050** / **0.2319**	0.1067 / 0.2322	0.1113 / 0.2432	0.1163 / 0.2503
	10	**0.1433** / **0.2789**	0.1437 / 0.2793	0.1476 / 0.2859	0.1460 / 0.2815
	12	**0.1679** / **0.3026**	0.1710 / 0.3053	0.1896 / 0.3385	0.1846 / 0.3264
	Avg.	**0.1234** / **0.2510**	0.1249 / 0.2518	0.1329 / 0.2686	0.1320 / 0.2638
Climate	6	1.0357 / 0.8275	1.0573 / 0.8322	**1.0103** / **0.8114**	1.0311 / 0.8165
	8	1.0500 / 0.8324	1.0624 / 0.8371	**1.0383** / **0.8257**	1.0579 / 0.8300
	10	**1.0791** / **0.8448**	1.0909 / 0.8504	1.0795 / **0.8448**	1.0931 / 0.8485
	12	1.0700 / 0.8423	1.0771 / 0.8448	**1.0457** / **0.8303**	1.0796 / 0.8451
	Avg.	1.0587 / 0.8367	1.0719 / 0.8411	**1.0434** / **0.8281**	1.0654 / 0.8350
Economy	6	0.1703 / 0.3212	0.1736 / 0.3248	**0.1648** / **0.3178**	0.1714 / 0.3230
	8	**0.1821** / **0.3318**	0.1874 / 0.3357	0.1956 / 0.3463	0.1937 / 0.3452
	10	**0.2380** / **0.3845**	0.2447 / 0.3904	0.2489 / 0.3981	0.2470 / 0.3936
	12	**0.2126** / **0.3623**	0.2205 / 0.3694	0.2334 / 0.3790	0.2249 / 0.3692
	Avg.	**0.2008** / **0.3500**	0.2065 / 0.3551	0.2107 / 0.3603	0.2092 / 0.3577
Energy	4	0.0246 / 0.1072	0.0252 / 0.1087	0.0321 / 0.1296	**0.0243** / **0.1070**
	8	0.0619 / **0.1728**	0.0633 / 0.1744	0.0675 / 0.1879	**0.0614** / 0.1735
	12	0.1012 / 0.2245	**0.1011** / **0.2238**	0.1045 / 0.2340	0.1014 / 0.2261
	24	**0.2131** / **0.3415**	0.2135 / 0.3429	0.2200 / 0.3544	0.2174 / 0.3496
	Avg.	**0.1002** / **0.2115**	0.1008 / 0.2124	0.1060 / 0.2265	0.1011 / 0.2140
Environment	7	0.2710 / 0.3690	0.2710 / **0.3684**	**0.2671** / 0.3702	0.2710 / 0.3685
	14	0.2805 / 0.3764	0.2806 / 0.3764	**0.2760** / **0.3750**	0.2815 / 0.3753
	30	0.2935 / 0.3817	0.2932 / 0.3814	**0.2837** / **0.3729**	0.2954 / 0.3902
	48	0.3075 / 0.3924	0.3075 / **0.3922**	**0.2996** / 0.3924	0.3140 / 0.4015
	Avg.	0.2881 / 0.3799	0.2881 / 0.3796	**0.2816** / **0.3776**	0.2905 / 0.3839
Health	4	0.4921 / 0.4247	0.5033 / 0.4339	**0.4706** / **0.4098**	0.5037 / 0.4309
	8	0.8763 / 0.5969	0.9090 / 0.6084	**0.8497** / **0.5966**	0.9208 / 0.6049
	12	1.2270 / 0.7243	1.2659 / 0.7372	**1.1647** / **0.7158**	1.2589 / 0.7426
	24	1.7093 / 0.8654	1.7277 / 0.8728	**1.6261** / **0.8432**	1.7240 / 0.8717
	Avg.	1.0762 / 0.6528	1.1015 / 0.6631	**1.0278** / **0.6413**	1.1018 / 0.6626
Security	6	**103.2083** / **4.6928**	105.9332 / 4.8266	107.7798 / 4.8811	107.9218 / 5.0072
	8	**106.8915** / **4.8352**	108.2990 / 4.9425	110.7981 / 5.1760	108.3112 / 4.9792
	10	**108.0075** / **5.0000**	111.0473 / 5.1780	110.5550 / 5.0759	111.1096 / 5.2016
	12	**111.0755** / 5.2095	113.2996 / 5.3443	112.2720 / 5.2190	111.7893 / **5.2013**
	Avg.	**107.2957** / **4.9344**	109.6448 / 5.0729	110.3513 / 5.0880	109.7830 / 5.0973
Social Good	6	**0.8846** / **0.4201**	1.0179 / 0.4649	0.9883 / 0.4622	1.0688 / 0.4807
	8	**0.9971** / **0.4812**	1.1471 / 0.5332	1.1531 / 0.5345	1.1708 / 0.5360
	10	**1.0616** / **0.5095**	1.1823 / 0.5525	1.1844 / 0.5397	1.1804 / 0.5529
	12	**1.1773** / **0.5614**	1.2988 / 0.6009	1.2795 / 0.6063	1.3075 / 0.6062
	Avg.	**1.0301** / **0.4931**	1.1615 / 0.5379	1.1513 / 0.5357	1.1819 / 0.5440
Traffic	6	**0.1718** / **0.2476**	0.1830 / 0.2521	0.1841 / 0.2541	0.1864 / 0.2615
	8	**0.1699** / **0.2478**	0.1853 / 0.2558	0.1819 / 0.2496	0.1873 / 0.2651
	10	**0.1810** / **0.2587**	0.1913 / 0.2653	0.1896 / 0.2588	0.1905 / 0.2692
	12	**0.2262** / **0.2777**	0.2373 / 0.2840	0.2403 / 0.2906	0.2382 / 0.2912
	Avg.	**0.1872** / **0.2580**	0.1992 / 0.2643	0.1990 / 0.2633	0.2006 / 0.2718
Average Rank (MSE/MAE)		**1.47** / **1.58**	2.69 / 2.61	2.53 / 2.58	3.17 / 3.06

Table 20: Improvement of AFTER over the iTransformer backbone. "Improved horizons" reports the number and percentage of forecasting horizons for which AFTER achieves a lower error than iTransformer. "Avg. reduction" reports the mean relative error reduction over the four horizons. Positive values indicate improvement.

Dataset	MSE		MAE	
	Improved horizons	Avg. reduction	Improved horizons	Avg. reduction
Agriculture	3/4 (75.0%)	+0.24%	1/4 (25.0%)	-0.32%
Climate	4/4 (100.0%)	+0.78%	4/4 (100.0%)	+0.26%
Economy	4/4 (100.0%)	+3.92%	4/4 (100.0%)	+2.37%
Energy	3/4 (75.0%)	+1.86%	2/4 (50.0%)	+0.96%
Environment	1/4 (25.0%)	-0.10%	0/4 (0.0%)	-0.05%
Health	4/4 (100.0%)	+14.03%	4/4 (100.0%)	+3.89%
Security	4/4 (100.0%)	+1.36%	4/4 (100.0%)	+2.82%
Social Good	4/4 (100.0%)	+9.50%	4/4 (100.0%)	+7.80%
Traffic	4/4 (100.0%)	+8.16%	4/4 (100.0%)	+3.46%
Overall	**31/36 (86.1%)**	**+4.42%**	**27/36 (75.0%)**	**+2.35%**

Table 21: Improvement of AFTER over the PatchTST backbone. "Improved horizons" reports the number and percentage of forecasting horizons for which AFTER achieves a lower error than PatchTST. "Avg. reduction" reports the mean relative error reduction over the four horizons. Positive values indicate improvement.

Dataset	MSE		MAE	
	Improved horizons	Avg. reduction	Improved horizons	Avg. reduction
Agriculture	3/4 (75.0%)	+0.94%	2/4 (50.0%)	+0.19%
Climate	4/4 (100.0%)	+0.97%	4/4 (100.0%)	+0.39%
Economy	4/4 (100.0%)	+2.24%	4/4 (100.0%)	+1.18%
Energy	3/4 (75.0%)	+1.19%	4/4 (100.0%)	+0.85%
Environment	2/4 (50.0%)	-0.20%	2/4 (50.0%)	+0.06%
Health	4/4 (100.0%)	+2.58%	4/4 (100.0%)	+1.59%
Security	4/4 (100.0%)	+1.93%	4/4 (100.0%)	+3.26%
Social Good	4/4 (100.0%)	+7.75%	4/4 (100.0%)	+6.66%
Traffic	4/4 (100.0%)	+4.47%	4/4 (100.0%)	+1.56%
Overall	**32/36 (88.9%)**	**+2.43%**	**32/36 (88.9%)**	**+1.75%**

Table 22: Improvement of AFTER over the TimeKAN backbone. "Improved horizons" reports the number and percentage of forecasting horizons for which AFTER achieves a lower error than TimeKAN. "Avg. reduction" reports the mean relative error reduction over the four horizons. Positive values indicate improvement.

Dataset	MSE		MAE	
	Improved horizons	Avg. reduction	Improved horizons	Avg. reduction
Agriculture	4/4 (100.0%)	+1.11%	3/4 (75.0%)	+0.28%
Climate	4/4 (100.0%)	+1.24%	4/4 (100.0%)	+0.52%
Economy	4/4 (100.0%)	+2.76%	4/4 (100.0%)	+1.43%
Energy	3/4 (75.0%)	+1.17%	3/4 (75.0%)	+0.60%
Environment	1/4 (25.0%)	-0.02%	0/4 (0.0%)	-0.07%
Health	4/4 (100.0%)	+2.49%	4/4 (100.0%)	+1.65%
Security	4/4 (100.0%)	+2.14%	4/4 (100.0%)	+2.73%
Social Good	4/4 (100.0%)	+11.43%	4/4 (100.0%)	+8.44%
Traffic	4/4 (100.0%)	+6.12%	4/4 (100.0%)	+2.40%
Overall	**32/36 (88.9%)**	**+3.16%**	**30/36 (83.3%)**	**+2.00%**

N GENERALIZATION TO ADDITIONAL NUMERICAL BACKBONES

To examine whether AFTER generalizes beyond the numerical backbones used in the main experiments, we further instantiate it with DLinear (Zeng et al., 2023), Autoformer (Wu et al., 2021), Informer (Zhou et al., 2021), and FiLM (Zhou et al., 2022a) on Security, Social Good, and Traffic.

Table 23: Performance of AFTER with four additional numerical backbones. Each pair in (a) reports MSE/MAE. Panel (b) reports the relative error reductions on each dataset and their macro average. Positive values indicate improvement.

(a) Forecasting performance

Dataset	Backbone	AFTER	Base
Social Good	DLinear	**1.2277 / 0.7491**	1.3663 / 0.7935
	Autoformer	**1.1728 / 0.6418**	1.2972 / 0.6911
	Informer	**0.8445 / 0.4737**	0.8936 / 0.4984
	FiLM	**1.1825 / 0.6815**	1.3480 / 0.7368
Security	DLinear	**105.883 / 4.650**	108.139 / 4.730
	Autoformer	**111.836 / 5.054**	113.935 / 5.190
	Informer	**128.438 / 6.423**	134.706 / 6.840
	FiLM	**112.751 / 5.229**	115.465 / 5.396
Traffic	DLinear	**0.2774 / 0.4119**	0.2844 / 0.4163
	Autoformer	**0.1909 / 0.2663**	0.2006 / 0.2717
	Informer	**0.2251 / 0.3857**	0.2300 / 0.3901
	FiLM	**0.2170 / 0.3333**	0.2245 / 0.3385

(b) Relative error reduction

Backbone	Dataset	MSE	MAE
DLinear	Social Good	+10.15%	+5.60%
	Security	+2.09%	+1.69%
	Traffic	+2.43%	+1.06%
	Overall	**+4.89%**	**+2.78%**
Autoformer	Social Good	+9.59%	+7.12%
	Security	+1.84%	+2.62%
	Traffic	+4.85%	+1.98%
	Overall	**+5.43%**	**+3.91%**
Informer	Social Good	+5.49%	+4.96%
	Security	+4.65%	+6.10%
	Traffic	+2.16%	+1.13%
	Overall	**+4.10%**	**+4.07%**
FiLM	Social Good	+12.28%	+7.50%
	Security	+2.35%	+3.10%
	Traffic	+3.33%	+1.56%
	Overall	**+5.99%**	**+4.05%**

As shown in Table 23, AFTER consistently improves the forecasts produced by all four numerical backbones across the evaluated datasets. The overall MSE reductions range from 4.10% to 5.99%, while the MAE reductions range from 2.78% to 4.07%. These results show that the event-informed adjustment is not restricted to a particular backbone architecture and can be applied to diverse numerical forecasting models.

O COMPARISON WITH REASONING-BASED FORECASTING

To complement the preceding comparisons with numerical and multimodal forecasting methods, we further compare AFTER with TimeOmni-1 (Guan et al., 2026), a unified time-series reasoning model that includes event-aware forecasting as one of its core tasks. We use its official implementation under the same data splits, input lengths, and forecasting horizons as in our main experiments.

Table 24: Comparison with the reasoning-based model TimeOmni-1.

Dataset	AFTER (iTransformer)	AFTER (PatchTST)	AFTER (TimeKAN)	TimeOmni-1
Agriculture	**0.1118 / 0.2350**	0.1139 / 0.2376	0.1234 / 0.2510	0.1487 / 0.2529
Economy	**0.1930 / 0.3402**	0.1984 / 0.3478	0.2008 / 0.3500	1.2649 / 0.7109
Social Good	1.0931 / **0.4555**	1.0359 / 0.4725	**1.0301** / 0.4931	1.0962 / 0.5213
Traffic	0.1957 / 0.2264	**0.1679 / 0.2152**	0.1872 / 0.2580	1.9542 / 0.7427

As shown in Table 24, all three AFTER variants outperform TimeOmni-1 in both MSE and MAE across the evaluated datasets. Although TimeOmni-1 is trained for multiple time-series reasoning tasks, translating such reasoning directly into accurate numerical forecasts remains challenging. In contrast, AFTER retains the forecasting capability of a trained numerical backbone and uses LLM reasoning to estimate event-informed offsets. These results show that anchoring textual reasoning to a numerical forecast enables more reliable quantitative predictions without additional parameter optimization.

P PROMPT TEMPLATES AND CASE INSTANTIATION

This section presents the prompts used by the LLM-based components of AFTER. For each reasoning stage, we provide the general prompt template and its structured output template, followed by the generated result for the Social Good case shown in Figure 4. Direct numerical transfer and ROI are deterministic operations and therefore do not require additional LLM prompts.

P.1 HISTORY-AWARE EVENT REASONING

History-Aware Event Reasoning (HAER) jointly examines the historical time series and its textual context to determine whether any textual event remains relevant to the forecasting horizon. At this stage, the LLM only produces a structured event state and does not estimate a numerical offset.

Prompt template. The general prompt used by HAER is shown below.

```
``You are an expert assistant for history-aware contextual analysis in multimodal time-
    series forecasting.

Your job is to synthesize the recent historical series and its historical context texts
    into one forecast-relevant context state for forecast_time_range.

The context texts are noisy historical evidence, not a list of events that must be used.
    Understand them chronologically together with the target series. Distinguish concrete
     external drivers from target-value reports, generic background, local or mismatched
    evidence, repeated descriptions of the same event, and events whose effects are
    already visible and completed in the history. Produce one integrated assessment
    rather than selecting or separately reasoning over many texts.

Reason in this order:
1. Understand the exact target variable and the historical series: its level, trend,
    volatility, seasonal pattern, late turning points, and state at the end of
    input_time_range.
2. Relate the context texts to the observed historical evolution. Identify whether they
    explain an already-observed movement, indicate an active or delayed external
    influence, show weakening or recovery, contradict one another, or provide no target-
    aligned information. Separate an event's accumulated effect on the current level from
     evidence that it will cause further change after the forecast origin.
3. Decide whether any contextual influence remains relevant for forecast_time_range.
    Summarize its mechanism, current phase, likely next pressure on the target from the
    forecast origin, plausible timing, and scope. The next pressure may oppose the event'
    s original effect during recovery. Do not infer a numeric offset and do not compare
    against a base forecast, which is handled in Stage2.

The dataset information is as follows.
[Dataset Description]
{dataset_description}
[Dataset Guidance]
{context_guidance}
Use Dataset Guidance to understand domain mechanisms, target scope, and target-variable
    semantics, not as a keyword list.

The forecasting setup is as follows.
[Forecasting Setup]
target_variable: {target_variable}
input_length: {input_length}
prediction_length: {prediction_length}
frequency: {frequency}
input_time_range: {input_time_range}
forecast_time_range: {forecast_time_range}
variable_description: {variable_description}

The historical time series is as follows.
[History Time Series]
reference_values at {time_0}: {initial_values_text}
```

```
calculation: each table value is computed as the current value minus the reference value
     at {time_0}.
{time_series_dfloader_text}

The reference values and table values are scaled values, not raw real-world measurements.
     Interpret them only for relative temporal changes.

The historical context texts are as follows.
[Context Texts]
{context_texts_dfloader_text}

The output must be exactly one valid JSON object following the structured template below.
     Do not include Markdown code fences or any extra text outside the JSON.
[Output]
{template}

[Context Status]
- `none`: the texts provide no credible target-aligned influence that remains relevant
     beyond the observed history.
- `weak`: a relevant influence is plausible, or an event still explains the current level,
      but its additional effect after the forecast origin is uncertain.
- `active`: concrete target-aligned evidence supports an active, delayed, emerging, or
     recovery-phase influence that can plausibly alter the target path after the forecast
     origin.

[Notes]
- context_summary must synthesize the net contextual state after considering all texts; do
      not enumerate every candidate event.
- Treat later historical values and later texts as evidence about whether an earlier event
      persisted, weakened, reversed, or ended.
- Give the latest observed turning point substantial weight. Earlier repeated reports that
      explain the pre-turn movement are not evidence that the same movement will resume
     unless later concrete evidence supports it.
- Do not invent conditional escalation, reversal, or persistence merely because it is
     possible. Describe the net pressure supported at the forecast origin.
- A target can remain unusually high or low while recovering in the opposite direction.
     State both the remaining level effect and the likely next movement when they differ.
- A turning point visible only in the numerical history helps identify the current phase
     but is not itself external contextual evidence. Do not mark the context `active`
     unless the texts add credible information about the path after that observed point.
- An ongoing condition that merely explains the current abnormal level is `weak` when the
     texts do not show that it will further change the target or alter the ongoing
     recovery after the forecast origin.
- Reports of historical target values usually describe what is already reflected in the
     series and are not by themselves forward evidence.
- Respect the target's geographic, population, variable, and temporal scope. Local or
     related-domain evidence is weak unless it plausibly affects the stated aggregate
     target.
- Use `none` when the only defensible conclusion is the endogenous historical pattern
     already visible in the series.
- evidence_times should contain only the timestamps essential to the synthesized
     conclusion and may be empty.
``
```

Output template. The structured output template used by HAER is shown below.

```
{
  "context_status": "<none | weak | active>",
  "history_state_summary": "<brief summary of the observed historical evolution, late
     turning point, and state at the forecast origin>",
  "context_summary": "<brief integrated summary of what the texts explain in the observed
     history and what influence, if any, remains after the forecast origin, including
     mechanism, current phase, likely next pressure on the target, timing, and scope>",
  "evidence_times": ["<timestamp>"]
}
```

Example output. The HAER output for the Social Good case in Figure 4 is shown below.

```
{
  "context_status": "active",
  "history_state_summary": "The unemployment rate was relatively stable around 4.6--4.7\%
     from mid-2007 to early 2008, then began a sharp and sustained increase from mid-2008
     through June 2009, reaching a level about 2.94 points above the July 2007 reference.
     The series shows a clear upward trend with no sign of peaking by the forecast origin
     .",
  "context_summary": "The texts document the unfolding of the Great Recession, with rising
      unemployment from late 2007 through 2009. They report accelerating job losses and
     rising unemployment rates, with the rate reaching 9.4\% in May 2009 and 9.5\% in June
      2009. The recession was officially recognized as starting in December 2007. The
     evidence indicates that the unemployment rate is still climbing, with no indication
     of a peak or recovery by the forecast origin. The likely next pressure is continued
     upward movement in the unemployment rate in the near term, as the labor market
     deterioration persists.",
  "evidence_times": [
    "2009-06-08",
    "2009-06-15",
    "2009-06-22"
  ]
}
```

P.2 Contextual Event Reasoning

Contextual Event Reasoning estimates the influence of the current event state relative to the backbone forecast. It jointly considers the historical series, base forecast, and HAER output, but does not use retrieved historical cases.

Prompt template. The general prompt template used for Contextual Event Reasoning is shown below.

```
You are an expert assistant for contextual forecast deviation reasoning.

Your job is to estimate numeric context offsets by jointly interpreting the recent
     historical series, the base forecast from a time-series model, and the history-aware
     context summary.

Do not regenerate the forecast from scratch. The base forecast is the anchor. Estimate
     only the residual contextual deviation that should be added to it.

Reason in this order:
1. Understand the exact target variable and the historical series, including its level,
     pattern, recent movement, late turning points, and state at the forecast origin.
2. Understand what the base forecast predicts by interpreting it as the continuation of
     the historical series. Relate its complete path to the preceding historical evolution
     , including the transition from the end of history into the first forecast row and
     how the prediction then evolves over the actual forecast timestamps.
3. Interpret the history-aware context summary as evidence about the target, not as a
     command. Separate the event's effect already accumulated in the current target level
     from evidence about further movement after the forecast origin. Determine the
     remaining mechanism, current phase, likely next pressure, and plausible duration. Re-
     evaluate weak, stale, mismatched, speculative, or already-completed evidence.
4. Close the residual comparison. For each affected portion of the horizon, jointly
     compare the historical state, remaining contextual pressure, and the change already
     predicted by the base path. When the target is recovering from a shock, compare the
     recovery speed and shape supported by the combined history and context with the
     recovery speed and shape in the base; do not reapply the shock merely because the
     level remains abnormal. Decide explicitly whether the resulting forecast should be
     above, below, or close to the base forecast there. The offset sign must express this
     relation to the base, not the event's original direction or the target's distance
     from a normal level.
```

```
5. Set the offset magnitude and shape using observed normal steps, volatility, turning-
     point size, contextual strength, and plausible persistence. A large movement already
     realized in history cannot by itself justify another large residual offset. Then
     verify that final_forecast = base_forecast + context_offset is consistent with the
     stated relation to the base at every affected row.

The dataset information is as follows.
[Dataset Description]
{dataset_description}

The forecasting setup is as follows.
[Forecasting Setup]
target_variable: {target_variable}
input_length: {input_length}
prediction_length: {prediction_length}
frequency: {frequency}
input_time_range: {input_time_range}
forecast_time_range: {forecast_time_range}
variable_description: {variable_description}

The historical time series is as follows.
[History Time Series]
reference_values at {time_0}: {initial_values_text}
calculation: each table value is computed as the current historical value minus the
     historical initial value at {time_0}.
{time_series_dfloader_text}

The base forecast from the {base_backbone_name} backbone is shown as the continuation of
     the historical path, using the same historical reference as the history table.
[Base Forecast]
reference_values at {base_time_0}: {base_initial_values_text}
calculation: each table value is computed as the current base forecast value minus the
     historical initial value at {base_time_0}.
{base_forecast_dfloader_text}

The Stage1 history-aware context state is as follows.
[History-Aware Context State]
{history_aware_context_summary}

The output must be exactly one valid JSON object following the compact column-style
     template below. Do not include Markdown code fences or any extra text outside the
     JSON.

[Output]
{
  "context_offset_reasoning": "<briefly explain the continuous history-to-base forecast
     situation, the remaining contextual influence and phase, which rows should end above,
      below, or close to the base, and how the offset magnitude and shape implement that
     conclusion>",
  "context_offset_values": {
    "offsets_by_variable": {
      "<variable_name>": [0.0]
    }
  }
}

[Notes]
- Estimate only context offsets: final_forecast = base_forecast + context_offset.
- The History and Base Forecast tables share one historical reference and describe one
     continuous path. Compare their displayed values directly.
- A high or low base level does not by itself mean a contextual effect is already
     reflected. Judge whether the base path's direction, timing, and magnitude reflect the
      remaining influence.
- Likewise, a target remaining above or below its pre-event level does not justify another
      offset in the event's original direction. During weakening or recovery, the residual
      direction may reverse.
- A non-zero offset does not need one-to-one attribution to a single text. It may be
     supported by a reasoned joint inconsistency among the historical state, the context
```

```
    phase, and the base continuation, but do not arbitrarily replace the base forecast or
     infer a future event from its absence in historical texts.
- For `context_status = none`, use zero offsets. For `weak`, use zero or limited offsets
    unless the joint history-context-to-base comparison provides a strong residual case
    with a clear direction and affected interval. `active` still does not require an
    offset when the base already represents the influence.
- Do not default to tiny offsets when active evidence implies a material new deviation.
    Conversely, without evidence of a new shock or acceleration after the forecast origin
    , keep the residual magnitude within the scale of ordinary recent steps and use zero
    when only the current abnormal level is supported.
- Match the offset shape to onset, persistence, weakening, reversal, or recovery. Do not
    repeat one value across the horizon unless a sustained constant deviation is
    justified, and return to zero when the influence is no longer supported.
- In offsets_by_variable, each list follows the Base Forecast row order, has exactly
    prediction_length values, and does not repeat forecast_time.
- Write positive JSON numbers without a leading plus sign: use 0.015, not +0.015.
```

Example output. The output generated for the Social Good case in Figure 4 is shown below.

```
{
  "context_offset_reasoning": "The historical series shows a sharp, sustained rise in the
     unemployment rate from mid-2008 through June 2009, ending at about +2.94 relative to
     the July 2007 reference. The context summary confirms the Great Recession is ongoing,
      with unemployment still climbing and no peak or recovery by the forecast origin. The
      base forecast continues the upward trend but predicts a dip in August 2009 and a
     generally slower climb thereafter, which is inconsistent with the persistent upward
     pressure described in the context. The base already captures much of the rise, but
     the context suggests the unemployment rate should continue to increase more steadily
     and reach higher levels by early 2010. Therefore, the forecast should be above the
     base for most of the horizon, with the largest positive offsets in the later months
     as the labor market deterioration persists. The offset shape starts small in July,
     grows through the fall, and remains elevated through February 2010, reflecting
     continued upward pressure rather than a reversal.",
  "context_offset_values": {
    "offsets_by_variable": {
      "OT": [0.1, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]
    }
  }
}
```

P.3 Case-Based Analogical Reasoning

Case-Based Analogical Reasoning introduces historical evidence into offset estimation. In addition to the current history, base forecast, and event state, the LLM receives the event states and observed continuations of the retrieved historical cases. It determines which historical responses are transferable to the current sample and estimates a memory-conditioned offset relative to the base forecast.

Prompt template. The general prompt template used for Case-Based Analogical Reasoning is shown below. Its required output format is included directly in the template.

```
You are an expert assistant for example-guided contextual forecast deviation reasoning.

Your job is to estimate a memory-conditioned context offset from the current history,
     current base forecast, current history-aware context state, and retrieved historical
     cases. Do not forecast from scratch or copy an example's observed future.

Reason in this order:
1. Understand the current history-to-base path and what the base already predicts.
2. Compare the retrieved cases with the current sample in mechanism, phase, target scope,
     timing, and historical response. Decide what their observed outcomes imply relative
```

```
     to the current base forecast, separating transferable behavior from case-specific
     variation or later events.
3. Estimate the offset relative to the current base forecast, including its direction,
     magnitude, and evolution over the horizon. Use small or zero offsets where the cases
     lack a shared target-aligned mechanism, conflict with the current state, or describe
     effects already represented by the base.

[Dataset Description]
{dataset_description}

[Current Forecasting Setup]
target_variable: {target_variable}
input_length: {input_length}
prediction_length: {prediction_length}
frequency: {frequency}
input_time_range: {input_time_range}
forecast_time_range: {forecast_time_range}
variable_description: {variable_description}

[Current History Time Series]
reference_values at {time_0}: {initial_values_text}
calculation: each table value is computed as the current historical value minus the
     historical initial value at {time_0}.
{time_series_dfloader_text}

[Current Base Forecast]
reference_values at {base_time_0}: {base_initial_values_text}
calculation: each table value is computed as the current base forecast value minus the
     historical initial value at {base_time_0}.
{base_forecast_dfloader_text}

[Current History-Aware Context State]
{history_aware_context_summary}

[Retrieved Historical Examples]
{retrieved_memory_cases}

The output must be exactly one valid JSON object. Do not include Markdown fences or any
     text outside the JSON.

[Output]
{
  "context_offset_reasoning": "<briefly explain the current history-to-base situation,
     which case evidence is transferable, and how it determines the memory-conditioned
     offset path>",
  "context_offset_values": {
    "offsets_by_variable": {
      "<variable_name>": [0.0]
    }
  }
}

[Notes]
- Estimate offsets only: adjusted_forecast = current_base_forecast + context_offset.
- The observed future of a case is evidence about a possible response, not a command or a
     pure event effect.
- Judge early and late rows separately when a transferable influence starts, persists,
     weakens, or reverses.
- Every offset list must follow the Current Base Forecast row order and contain exactly
     prediction_length values.
- Write positive JSON numbers without a leading plus sign.
```

Retrieved event-response cases. For the Social Good sample in Figure 4, the retrieval module returns the following three historical cases. Each case contains a historical time series, its HAER event state, and the subsequently observed future. All numerical values use the corresponding case's initial historical value as their shared reference. Repeated monthly timestamps are represented by their time ranges for readability.

```
[Retrieved Historical Case 1]

input_time_range: 1989-07-01 to 1991-06-01
forecast_time_range: 1991-07-01 to 1992-02-01
reference_values at 1989-07-01: OT=-0.2202

[Memory History Time Series]
OT: [+0.0000, -0.1226, -0.1226, -0.1839, -0.0613, -0.1226, +0.4292, +0.3679, +0.1226,
     +0.0000, -0.0613, +0.0613, +0.1839, +0.1226, +0.1839, +0.1226, +0.3679, +0.4292,
     +1.1036, +1.2262, +1.1649, +0.7357, +0.8583, +1.0423]

[Memory History-Aware Context State]
{
  "context_status": "active",
  "history_state_summary": "The unemployment rate was relatively stable around the
     reference level from mid-1989 to mid-1990, then began a clear upward trend starting
     in late 1990, accelerating sharply in early 1991. By June 1991, the rate was about
     1.04 points above the July 1989 reference, indicating a significant rise.",
  "context_summary": "The historical texts indicate a U.S. recession beginning in mid
     -1990, with rising unemployment observed through 1991. Reports from late 1990 and
     early 1991 confirm the upward trend, and by May 1991 the unemployment rate reached
     6.9\%, the highest since 1986. The recession's effects are ongoing, with continued
     labor market weakness expected in the near term. The context supports an active
     influence on the unemployment rate during the forecast period, as the recession's
     impact is still unfolding and likely to keep unemployment elevated or rising further
     ."
}

[Memory Observed True Future]
OT: [+0.9196, +0.7970, +0.7357, +0.7357, +0.8583, +0.9809, +1.7167, +1.7780]
```

```
[Retrieved Historical Case 2]

input_time_range: 1980-07-01 to 1982-06-01
forecast_time_range: 1982-07-01 to 1983-02-01
reference_values at 1980-07-01: OT=1.3739

[Memory History Time Series]
OT: [+0.0000, -0.1839, -0.4292, -0.4905, -0.4905, -0.6131, +0.1839, +0.0613, -0.1226,
     -0.5518, -0.4905, -0.1226, -0.3679, -0.4292, -0.3679, -0.2452, +0.0000, +0.2452,
     +0.9196, +1.0423, +0.9809, +0.7970, +0.7357, +1.1649]

[Memory History-Aware Context State]
{
  "context_status": "active",
  "history_state_summary": "The OT series shows a decline from mid-1980 to early 1981,
     followed by a sharp rise starting in late 1981, with a notable acceleration in early
     1982. By June 1982, the level is significantly above the reference, indicating a
     strong upward trend in unemployment.",
  "context_summary": "The texts describe a severe recession beginning in 1981, with rising
      unemployment, high interest rates, and policy responses. The historical series
     confirms an upward trend in unemployment from late 1981. The context indicates that
     the recession is ongoing and unemployment is expected to continue rising, with
     mentions of the 1981--1982 recession pushing unemployment up and reports of high and
     rising unemployment in 1982. This supports an active influence that will likely
     continue to pressure the unemployment rate upward in the forecast period."
}

[Memory Observed True Future]
OT: [+1.1649, +1.0423, +1.1036, +1.2262, +1.5327, +1.5940, +2.1458, +2.0845]
```

```
[Retrieved Historical Case 3]
```

```
input_time_range: 1983-07-01 to 1985-06-01
forecast_time_range: 1985-07-01 to 1986-02-01
reference_values at 1983-07-01: OT=2.2935

[Memory History Time Series]
OT: [+0.0000, -0.1226, -0.3679, -0.6131, -0.7970, -0.8583, -0.3679, -0.6131, -0.7970,
     -1.1036, -1.3488, -1.2262, -1.1649, -1.2875, -1.4101, -1.4714, -1.5327, -1.4714,
     -0.8583, -0.9809, -1.1649, -1.4101, -1.4714, -1.1649]

[Memory History-Aware Context State]
{
  "context_status": "weak",
  "history_state_summary": "The unemployment rate declined sharply from mid-1983 through
     late 1984, reaching a low around -1.53 relative to the reference, then rose slightly
     in early 1985 before resuming a decline, ending at -1.16 in June 1985. The series
     shows a strong recovery from the early 1980s recession, with a recent dip and partial
      rebound.",
  "context_summary": "The texts document the strong economic recovery from the early 1980s
      recession, with unemployment falling from double digits to around 7\% by late 1984.
     Reports of continued declines and low unemployment in early 1985 suggest the recovery
      is ongoing but slowing. There is no strong evidence of a new external shock; the
     main influence is the tail end of the recovery, which may continue to exert mild
     downward pressure on unemployment, but the effect is weakening and uncertain."
}

[Memory Observed True Future]
OT: [-1.2262, -1.5327, -1.5327, -1.5940, -1.6554, -1.6554, -1.2875, -0.9809]
```

Example output. The Case-Based Analogical Reasoning output for the Social Good case is shown below.

```
{
  "context_offset_reasoning": "The current base forecast already reflects a high and still
     -rising unemployment rate after the sharp 2008--2009 increase, with a slight decline
     in the first forecast month and then a mild upward drift. The retrieved cases are all
      recession/recovery episodes with active or weak context states, but their mechanisms
      and phases differ. Case 1 (1991) is a recession with unemployment still rising, but
     its observed future shows a near-term plateau and then a renewed rise, implying the
     current base may be slightly too high early and too low late. Case 2 (1982) is a
     severe recession with unemployment still rising, and its observed future shows a
     continued upward trend, supporting a positive offset later in the horizon. Case 3
     (1985) is a recovery phase with weak context and declining unemployment, which
     conflicts with the current recessionary state and should not be transferred. Overall,
      the transferable evidence suggests a small negative offset in the first month, near-
     zero in the next few months, and a modest positive offset toward the end of the
     horizon as recessionary pressures persist.",
  "context_offset_values": {
    "offsets_by_variable": {
      "OT": [-0.15, -0.05, 0.0, 0.05, 0.1, 0.15, 0.2, 0.25]
    }
  }
}
```

Table 25: Overall comparison with multimodal forecasting methods. Each entry reports MSE/MAE averaged over the evaluated random seeds. The *Avg.* row reports the mean over four forecasting horizons. Lower values are better. The best results are shown in bold and the second-best results are underlined.

Dataset	H	AFTER (iTransformer)	AFTER (PatchTST)	AFTER (TimeKAN)	GPT4TS	GPT4MTS	ChatTime	Aurora	TimeLLM	TimeVLM
Agriculture	6	**0.0651 / 0.1779**	0.0695 / 0.1847	0.0775 / 0.1906	0.1002 / 0.2360	0.1015 / 0.2372	0.1480 / 0.2371	0.1328 / 0.2602	0.1045 / 0.2407	0.1090 / 0.2454
	8	**0.0895 / 0.2079**	0.0952 / 0.2180	0.1050 / 0.2319	0.1202 / 0.2577	0.1146 / 0.2463	0.1889 / 0.2691	0.1661 / 0.2897	0.1330 / 0.2731	0.1325 / 0.2713
	10	**0.1296** / 0.2588	0.1304 / **0.2580**	0.1433 / 0.2789	0.1552 / 0.3000	0.1377 / 0.2692	0.2317 / 0.3021	0.2030 / 0.3189	0.1694 / 0.3171	0.1694 / 0.3105
	12	0.1628 / 0.2955	**0.1605 / 0.2898**	0.1679 / 0.3026	0.1873 / 0.3264	0.1837 / 0.3211	0.2733 / 0.3316	0.2495 / 0.3492	0.2079 / 0.3513	0.2036 / 0.3389
	Avg.	**0.1118 / 0.2350**	0.1139 / 0.2376	0.1234 / 0.2510	0.1407 / 0.2800	0.1344 / 0.2685	0.2105 / 0.2850	0.1879 / 0.3045	0.1537 / 0.2956	0.1536 / 0.2915
Climate	6	1.1122 / 0.8461	1.1801 / 0.8793	**1.0357 / 0.8275**	1.1031 / 0.8441	1.1002 / 0.8485	1.5223 / 0.9981	1.1447 / 0.8353	1.1658 / 0.8671	1.2236 / 0.8909
	8	1.1435 / 0.8602	1.1829 / 0.8808	**1.0500 / 0.8324**	1.1404 / 0.8547	1.0820 / 0.8380	1.5788 / 1.0136	1.1831 / 0.8508	1.1119 / 0.8432	1.3016 / 0.9174
	10	1.1536 / 0.8645	1.1638 / 0.8767	**1.0791 / 0.8448**	1.1563 / 0.8584	1.1062 / 0.8489	1.5099 / 0.9910	1.2226 / 0.8669	1.1870 / 0.8704	1.3057 / 0.9137
	12	1.1763 / 0.8800	1.1629 / 0.8740	**1.0700 / 0.8423**	1.1513 / 0.8551	1.1066 / 0.8511	1.5397 / 1.0013	1.2431 / 0.8761	1.2146 / 0.8784	1.3071 / 0.9127
	Avg.	1.1464 / 0.8627	1.1724 / 0.8777	**1.0587 / 0.8367**	1.1378 / 0.8531	1.0987 / 0.8466	1.5377 / 1.0010	1.1984 / 0.8573	1.1698 / 0.8648	1.2845 / 0.9087
Economy	6	**0.1647 / 0.3142**	0.1717 / 0.3199	0.1703 / 0.3212	0.2505 / 0.4048	0.2098 / 0.3618	0.2846 / 0.4217	0.2574 / 0.3961	0.2818 / 0.4312	0.2436 / 0.3977
	8	0.1851 / 0.3350	0.1889 / 0.3409	**0.1821 / 0.3318**	0.2341 / 0.3917	0.2217 / 0.3725	0.3165 / 0.4433	0.2874 / 0.4205	0.2625 / 0.4139	0.2521 / 0.4048
	10	**0.2016 / 0.3469**	0.2080 / 0.3583	0.2380 / 0.3845	0.2519 / 0.4072	0.2408 / 0.3886	0.3263 / 0.4520	0.3141 / 0.4390	0.2926 / 0.4406	0.2760 / 0.4246
	12	0.2204 / 0.3646	0.2250 / 0.3721	**0.2126 / 0.3623**	0.2437 / 0.3990	0.2688 / 0.4138	0.3411 / 0.4631	0.3357 / 0.4526	0.2892 / 0.4355	0.2711 / 0.4204
	Avg.	**0.1930 / 0.3402**	0.1984 / 0.3478	0.2008 / 0.3500	0.2451 / 0.4007	0.2353 / 0.3842	0.3171 / 0.4450	0.2987 / 0.4271	0.2815 / 0.4303	0.2607 / 0.4119
Energy	4	0.0247 / 0.1059	**0.0240 / 0.1054**	0.0246 / 0.1072	0.0373 / 0.1355	0.0461 / 0.1505	0.0441 / 0.1414	0.0966 / 0.2359	0.0289 / 0.1169	0.0363 / 0.1329
	8	0.0624 / 0.1734	0.0622 / **0.1719**	**0.0619** / 0.1728	0.0720 / 0.1898	0.1048 / 0.2376	0.1101 / 0.2301	0.1406 / 0.2840	0.0724 / 0.1853	0.0746 / 0.1903
	12	**0.0973 / 0.2187**	0.1066 / 0.2286	0.1012 / 0.2245	0.1114 / 0.2368	0.1312 / 0.2658	0.1680 / 0.2850	0.1862 / 0.3247	0.1107 / 0.2280	0.1117 / 0.2348
	24	**0.1986 / 0.3281**	0.2141 / 0.3382	0.2131 / 0.3415	0.2208 / 0.3462	0.2526 / 0.3809	0.2498 / 0.3657	0.2797 / 0.3986	0.2221 / 0.3443	0.2154 / 0.3374
	Avg.	**0.0957 / 0.2065**	0.1017 / 0.2110	0.1002 / 0.2115	0.1104 / 0.2271	0.1337 / 0.2587	0.1430 / 0.2555	0.1758 / 0.3108	0.1085 / 0.2186	0.1095 / 0.2239
Environment	7	0.2698 / 0.3717	0.2709 / 0.3694	0.2710 / 0.3690	**0.2690 / 0.3666**	0.2717 / 0.3727	0.3346 / 0.3935	0.3530 / 0.4186	0.2722 / 0.3706	0.2699 / 0.3681
	14	**0.2761** / 0.3770	0.2832 / 0.3795	0.2805 / 0.3764	0.2795 / 0.3751	0.2799 / **0.3744**	0.3471 / 0.4010	0.3647 / 0.4271	0.2814 / 0.3776	0.2801 / 0.3746
	30	**0.2713 / 0.3757**	0.2906 / 0.3834	0.2935 / 0.3817	0.2915 / 0.3801	0.2949 / 0.3864	0.3540 / 0.4070	0.3857 / 0.4411	0.2947 / 0.3843	0.2950 / 0.3826
	48	**0.2770 / 0.3820**	0.3127 / 0.3946	0.3075 / 0.3924	0.3124 / 0.3901	0.3160 / 0.4002	0.3653 / 0.4153	0.4039 / 0.4514	0.3120 / 0.3949	0.3154 / 0.3939
	Avg.	**0.2736 / 0.3766**	0.2894 / 0.3817	0.2881 / 0.3799	0.2881 / 0.3780	0.2906 / 0.3834	0.3503 / 0.4042	0.3768 / 0.4345	0.2901 / 0.3818	0.2901 / 0.3798
Health	4	**0.4131 / 0.3851**	0.4816 / 0.4398	0.4921 / 0.4247	0.5714 / 0.4925	0.5822 / 0.4944	0.6793 / 0.4731	1.2307 / 0.7971	0.5578 / 0.4617	0.6155 / 0.5125
	8	**0.7474 / 0.5249**	0.8320 / 0.5711	0.8763 / 0.5969	0.9889 / 0.6616	1.0188 / 0.6631	1.3902 / 0.7195	1.8008 / 0.9716	1.0071 / 0.6301	1.1015 / 0.6956
	12	1.2391 / **0.6734**	**1.2080** / 0.7147	1.2270 / 0.7243	1.3637 / 0.7873	1.3911 / 0.7770	1.8638 / 0.8646	2.2344 / 1.0894	1.3376 / 0.7546	1.4466 / 0.8034
	24	2.0728 / 0.8694	1.8905 / 0.8953	**1.7093 / 0.8654**	1.8077 / 0.8951	1.7571 / 0.8736	2.5160 / 1.0364	2.7464 / 1.2108	1.7708 / 0.8834	1.9258 / 0.9223
	Avg.	1.1181 / **0.6132**	1.1030 / 0.6552	**1.0762** / 0.6528	1.1829 / 0.7091	1.1873 / 0.7020	1.6123 / 0.7734	2.0031 / 1.0172	1.1683 / 0.6825	1.2723 / 0.7335
Security	6	113.2056 / 5.3958	108.8765 / 5.2163	**103.2083** / 4.6928	106.7255 / 4.7517	106.0658 / 4.7476	116.2920 / 5.5332	107.5873 / **4.6161**	114.3026 / 5.5387	115.2021 / 5.3290
	8	114.3334 / 5.4868	112.4414 / 5.3524	**106.8915** / 4.8352	108.0057 / **4.8215**	111.0346 / 5.1354	118.3159 / 5.6387	112.5078 / 4.8495	115.4524 / 5.4789	120.6756 / 5.6626
	10	113.2471 / 5.3701	113.1932 / 5.3716	**108.0075** / 5.0000	108.9359 / **4.8679**	109.8788 / 4.9893	120.4965 / 5.8005	118.2118 / 5.0817	112.9196 / 5.3176	121.3145 / 5.8006
	12	114.8561 / 5.4971	115.9766 / 5.5537	111.0755 / 5.2095	**110.2826 / 5.0132**	110.6397 / 5.0905	121.2885 / 5.8485	121.4779 / 5.2650	112.3846 / 5.2636	123.8993 / 5.9028
	Avg.	113.9105 / 5.4374	112.6219 / 5.3735	**107.2957** / 4.9344	108.4874 / **4.8636**	109.4047 / 4.9907	119.0982 / 5.7052	114.9462 / 4.9531	113.7648 / 5.3997	120.2729 / 5.6737
Social Good	6	0.9607 / **0.3931**	0.8686 / 0.3968	0.8846 / 0.4201	0.8503 / 0.4462	1.3808 / 0.5656	1.2040 / 0.6008	1.3262 / 0.7484	0.9405 / 0.4678	**0.8484** / 0.4693
	8	1.0430 / **0.4347**	0.9758 / 0.4464	0.9971 / 0.4812	**0.9132** / 0.4889	1.4138 / 0.5946	1.2748 / 0.6519	1.4345 / 0.7949	1.0255 / 0.5110	1.1299 / 0.5918
	10	1.1678 / **0.4893**	1.0814 / 0.4992	1.0616 / 0.5095	**1.0110** / 0.5578	1.7277 / 0.6839	1.3268 / 0.6907	1.5178 / 0.8323	1.1525 / 0.6046	1.0382 / 0.5746
	12	1.2010 / **0.5051**	1.2180 / 0.5478	1.1773 / 0.5614	**1.0348** / 0.5497	1.6216 / 0.6809	1.3918 / 0.7315	1.5833 / 0.8660	1.1467 / 0.5947	1.1774 / 0.6598
	Avg.	1.0931 / **0.4555**	1.0359 / 0.4725	1.0301 / 0.4931	**0.9523** / 0.5106	1.5360 / 0.6313	1.2993 / 0.6687	1.4654 / 0.8104	1.0663 / 0.5445	1.0485 / 0.5739
Traffic	6	0.1821 / 0.2174	**0.1545 / 0.2109**	0.1718 / 0.2476	0.1912 / 0.2685	0.1743 / 0.2361	0.2945 / 0.3682	0.2113 / 0.3478	0.2249 / 0.3111	0.2320 / 0.3336
	8	0.1901 / 0.2285	**0.1532 / 0.2126**	0.1699 / 0.2478	0.1975 / 0.2727	0.1811 / 0.2448	0.2973 / 0.3718	0.2126 / 0.3467	0.2400 / 0.3344	0.2384 / 0.3388
	10	0.1997 / 0.2404	**0.1651 / 0.2201**	0.1810 / 0.2587	0.2011 / 0.2744	0.1718 / 0.2337	0.3026 / 0.3738	0.2131 / 0.3478	0.2588 / 0.3648	0.2476 / 0.3443
	12	0.2109 / 0.2193	**0.1988 / 0.2170**	0.2262 / 0.2777	0.2592 / 0.3020	0.2308 / 0.2619	0.3409 / 0.3760	0.2236 / 0.3423	0.2832 / 0.3441	0.3067 / 0.3744
	Avg.	0.1957 / 0.2264	**0.1679 / 0.2152**	0.1872 / 0.2580	0.2123 / 0.2794	0.1895 / 0.2441	0.3088 / 0.3724	0.2152 / 0.3461	0.2517 / 0.3386	0.2562 / 0.3478
Average Rank (MSE/MAE)		3.11 / **2.69**	3.22 / 3.33	**2.39 / 2.69**	3.75 / 4.14	4.75 / 4.69	8.25 / 7.89	7.61 / 7.36	5.56 / 5.86	6.28 / 6.33