TEXT-GUIDED IN-CONTEXT FORECASTING FOR MULTIMODAL TIME SERIES PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world time series forecasting often contains both numerical histories and textual context, such as news, reports, or event descriptions. Existing multimodal forecasting methods typically incorporate text through LLM-based prompting or representation-level fusion, which presents a modelling challenge when textual signals are heterogeneous, noisy and unevenly informative. We propose TAG-ICF, a text-attention guided in-context learning (ICL) forecasting framework that uses a frozen, pretrained ICL backbone for numerical prediction while using text to modulate attention over historical examples. Textual inputs are analyzed using LLMs and embedded as lagged features to compute query-context relevance scores, serving as an attention gate that prioritizes contextually relevant historical patterns. TAG-ICF consistently outperforms all multimodal baselines across twelve datasets, yielding superior average ranks of 1.58 and 1.58 in MSE and MAE, respectively. Our framework surpasses numerical-only ICL on nearly all datasets and delivers up to 15.57% MSE improvement relative to the strongest competing methods in diverse settings.

1 INTRODUCTION

Time series forecasting is a cornerstone of decision-making in financial markets, energy systems, and macroeconomics (Liu et al., 2024; Williams et al., 2025). In these domains, predictive signals are inherently multimodal: numerical data captures endogenous historical dynamics, while unstructured text, such as news reports, policy briefings, or event descriptions, provides crucial context on exogenous drivers and regime shifts that are unobservable from numerical histories alone.

Existing multimodal forecasting approaches incorporate textual information via LLM-based prompting (Xie et al., 2023; Yu et al., 2023; Wang et al., 2024) or embedding fusion (Jin et al., 2024; Sun et al., 2024; Jia et al., 2024; Cheng et al., 2025; Liu et al., 2025; Tang et al., 2025; Wu et al., 2026). Prompting-based methods leverage the zero-shot reasoning of pretrained LLMs over temporal sequences; however, LLMs often remain unreliable for precise numerical reasoning (Tan et al., 2024). Fusion-based methods, on the other hand, integrate textual representations with numerical features through embedding fusion or alignment. While effective in many multimodal settings, these approaches often require the model to learn not only the numerical forecasting structure, but also when and how textual context should influence the prediction. This challenge is pronounced in realistic settings where the textual context is heterogeneous and unevenly informative (Zhang et al., 2025). However, texts can also be useful in identifying historical situations, such as specific market shocks or events that are relevant to the current period, in contrast to numerical histories alone, which may fail to fully capture underlying temporal shifts. This insight aligns naturally with the In-Context Learning (ICL) framework, which leverages historical examples as context to perform prediction during inference. Such alignment enables the use of text as contextual guidance to modulate relevant historical patterns, bypassing the complexities of traditional representation-level fusion.

Recent ICL-based tabular foundation models, such as TabDPT (Ma et al., 2025) and TabPFN (Hollmann et al., 2023; 2025; Grinsztajn et al., 2025; 2026), have demonstrated strong performance on structured prediction tasks under this paradigm. Recent work has further extended ICL to time series forecasting, with approaches such as TabPFN-TS (Hoo et al., 2025) showing that structured in-context learning can be effective for sequential prediction. Separately, methods such as ContextTab (Spinaci et al., 2025) explore incorporating textual information using representation-level fusion

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.

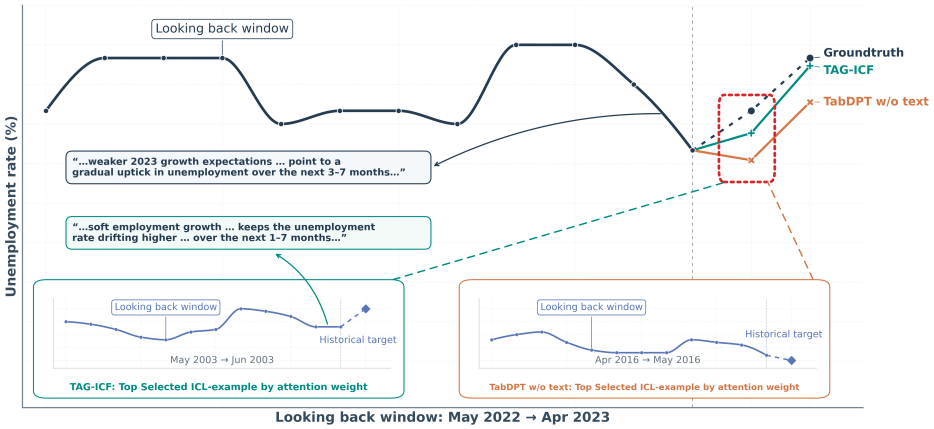


Figure 1: A motivating example for TAG-ICF. With TabDPT as the ICL backbone, textual guidance shifts attention from a numerically similar historical ICL example (**orange box**) to one whose textual signal is more closely aligned (**green box**), which exhibits a textual signal consistent with the current textual features—suggesting a slight increase in unemployment over the following months. After reweighting the attention scores toward this textually aligned ICL example, TAG-ICF shifts the forecast from the downward trend produced without text to an upward trend that is correctly aligned with the ground-truth direction and reduces the forecasting error.

into ICL-based models for tabular prediction. However, multimodal time series forecasting with textual context under an ICL framework remains largely underexplored.

We propose TAG-ICF, a Text-Attention Guided In-Context Forecasting framework for multimodal numerical time-series prediction. Our framework uses a frozen, pretrained ICL forecasting backbone to model temporal numerical structure, while textual inputs are analyzed using LLMs, embedded, and organized into lagged representations that determine the relevance of candidate historical examples to the current query. These relevance scores modulate the backbone's attention over the in-context examples, enabling text-informed forecasting without requiring representation-level fusion between textual and numerical features, as illustrated in Figure 1. Our core novelty lies in using text to guide the selection and weighting of numerical context, rather than treating it as an additional feature, thereby preserving the tabular foundation model as the primary numerical reasoner.

We evaluate TAG-ICF on diverse multimodal benchmarks over 12 datasets spanning diverse domains. Our method outperforms the numerical-only ICL by 3.44% in MAE and 7.21% in MSE on average, consistently surpassing multimodal baselines with top average ranks of 1.58 (MSE) and 1.58 (MAE). TAG-ICF demonstrates the most stable and significant gains from textual integration, improving MSE by up to 15.57% relative to the strongest baseline while maintaining robustness across noisy, heterogeneous datasets.

2 Related Work and Background

2.1 Multimodal Time Series Forecasting with Text

Existing approaches to time series forecasting with textual context can be broadly grouped into prompting-based methods and fusion-based methods. Prompting-based methods serialize numerical time series and textual context into natural language prompts and rely on pretrained language models to generate forecasts. Representative works include Xie et al. (2023); Yu et al. (2023); Wang et al. (2024), as well as PromptCast (Xue & Salim, 2023), ChatTime (Wang et al., 2025a), and LLMTime (Gruver et al., 2023), which also explore applying LLMs to time series forecasting by converting numerical inputs into textual form. More recently, Context is Key (Williams et al., 2025) highlights the importance of textual context for LLM forecasters when it contains predictive signals. However, such prompting-based methods rely on language-token reasoning over numerical sequences, which can be brittle for precise temporal extrapolation and structured numerical prediction (Tan et al., 2024).

Conventional fusion-based methods instead learn modality-specific representations and combine or align them before prediction. One group of methods focuses on aligning time-series representations with language-model or textual embedding spaces, including Time-LLM (Jin et al., 2024), TEST (Sun et al., 2024), InstructTime (Cheng et al., 2025), TimeCMA (Liu et al., 2025), and LLM-PS (Tang et al., 2025). Another group designs dedicated multimodal architectures that fuse temporal and textual features through encoders, cross-attention, decomposition, or other learned interaction modules, such as GPT4MTS (Jia et al., 2024), MM-TSFlib (Liu et al., 2024), TGForecaster (Xu et al., 2024), MoAT (Lee et al., 2024), CTPD (Wang et al., 2025b), TimeCAP (Lee et al., 2025), ContextFormer (Chattopadhyay et al., 2025), Time-VLM (Zhong et al., 2025), TimeXL (Jiang et al., 2026), Aurora (Wu et al., 2026), and TaTS (Li et al., 2026). These methods can capture richer cross-modal interactions than direct prompting, but typically rely on supervised training of task-specific multimodal architectures and carefully designed alignment or fusion modules. This can be challenging in realistic settings where aligned text–time series datasets are noisy, or domain-specific.

Overall, prior work primarily asks either how to prompt language models with time-series information, or how to fuse textual and temporal representations in supervised multimodal models. In contrast, our work studies how textual context can be incorporated into structured in-context forecasting backbones, enabling text to guide temporal reasoning without relying on LLM prompting or direct mapping of joint representations to numerical signals.

2.2 Structured in-context learning

In-context learning (ICL) has recently emerged as a promising paradigm for structured prediction, where models adapt to new tasks by conditioning on examples provided at inference time rather than relying on task-specific retraining. In the tabular domain, foundation models such as TabPFN (Hollmann et al., 2023; 2025; Grinsztajn et al., 2025; 2026) and its real-data variant (Garg et al., 2025), TabDPT (Ma et al., 2025), and TabICL (Qu et al., 2025; 2026) have demonstrated strong performance by learning to infer predictive patterns from contextual examples. This makes structured ICL particularly attractive for low-data settings, where training a task-specific model can be unreliable.

Recent work has begun extending this paradigm beyond standard tabular prediction. TabPFN-TS (Hoo et al., 2025) adapts TabPFN-style in-context learning to time series forecasting, showing that structured ICL can be effective for sequential prediction. In parallel, ContextTab (Spinaci et al., 2025) extends tabular ICL to multimodal tabular settings by incorporating textual information alongside tabular features. The versatility of tabular foundation models is further demonstrated by their ability to perform diverse tasks such as causal prediction (Balazadeh et al., 2025; Robertson et al., 2025), generation (Ma et al., 2023; van Breugel et al., 2024), and distillation (Ma et al., 2024; Feuer et al., 2024). Together, these works suggest that ICL-based structured prediction can support both temporal forecasting and text-enhanced prediction.

3 Methodology

We propose Text-Guided In-Context Forecasting (TAG-ICF), a trainable framework for multimodal time-series prediction in which textual context modulates the relevance of historical examples rather than serving directly as predictive features. As illustrated in Figure 2, numerical histories are processed by a pre-trained, frozen, transformer-based in-context learning (ICL) backbone, while temporally aligned textual histories are used to estimate query–context relevance scores. Our core technical contribution is a text-derived attention guidance mechanism that uses textual information to reweight historical numerical contexts rather than directly fusing text with numerical inputs. The resulting relevance scores define a text-conditioned attention gate that directs the model toward contextually relevant historical patterns while preserving the numerical inductive biases of the pre-trained forecasting backbone.

We first establish the formal framework for multimodal horizon forecasting in Section 3.1. In Section 3.2, we describe the transformation of the multimodal data and the reformulation of the horizon forecasting task to better align with the ICL paradigm. Finally, Section 3.3 introduces the text-guided attention gate, which modulates the backbone's query–context attention using text-derived relevance scores.

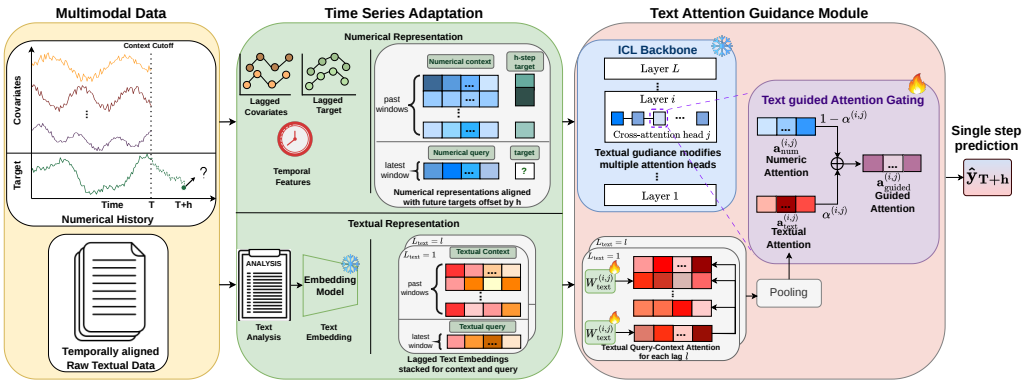


Figure 2: **Architectural overview of the TAG-ICF framework.** Multimodal inputs are processed through separate numerical and textual pipelines: the numerical representation $\mathbf{c}_T$ combines lagged numerical covariates ($\mathbf{x}_T$) and targets (y_T) with temporal features ($\boldsymbol{\tau}_{T+h}$) associated with the forecast timestamp $T+h$, while the raw textual input $\mathbf{s}_T^{\text{raw}}$ is synthesized by an LLM and encoded into a lagged embedding matrix S_T. The TAG Module projects embeddings via $W_{\text{text}}^{(i,j)}$ to compute a query–context relevance vector $\mathbf{a}_{\text{text}}$, averaged across L_{text} lookback windows. This is fused with the frozen backbone's numerical attention $\mathbf{a}_{\text{num}}^{(i,j)}$ via a trainable attention gating coefficient $\alpha^{(i,j)}$, creating guided attention that prioritizes contextually relevant historical windows. The resulting representation generates stepwise forecasts over H steps. Trainable parameters ($W_{\text{text}}^{(i,j)}$, $\alpha^{(i,j)}$) and frozen components (LLM, embedding model, ICL backbone) are marked with flame and snowflake symbols, respectively.

We instantiate this framework and conduct experiments primarily using TabDPT (Ma et al., 2025), chosen for its efficient row-based tokenization, fast inference, and strong empirical performance. We further conduct ablation studies using TabPFNv3 (Grinsztajn et al., 2026), a state-of-the-art tabular foundation model, as an alternative backbone to demonstrate the flexibility of our framework. Our mechanism operates at the level of query–context attention and can therefore be applied to transformer-based ICL backbones whose internal attention mechanisms are accessible and amenable to modification.

3.1 Problem Formulation

Given a target time series $\mathbf{y}_{1:T} = [y_1, y_2, \ldots, y_T]^\top$, where $y_t \in \mathbb{R}$, our objective is to predict the target values for a prediction horizon H, denoted as $\mathbf{y}_{T+1:T+H} = [y_{T+1}, y_{T+2}, \ldots, y_{T+H}]^\top$. Each time step t is associated with multimodal historical context $\{t, \mathbf{x}_t, \mathbf{s}_t^{\text{raw}}, y_t\}$, where $\mathbf{x}_t \in \mathbb{R}^{d_n}$ represents numerical covariates and $\mathbf{s}_t^{\text{raw}}$ represents textual covariates in natural language vocabulary space.

Following the In-Context Learning (ICL) paradigm, we formulate forecasting as a conditional generation problem. A key constraint in our setting is that for the prediction horizon $t \in \{T+1, \ldots, T+H\}$, the external numerical and textual covariates $\{\mathbf{x}_t, \mathbf{s}_t^{\text{raw}}\}$ are unavailable. However, we assume that temporal features derived from step t, $\boldsymbol{\tau}_t \in \mathbb{R}^{d_\tau}$ (e.g., hour of day, day of week, and seasonal indices inferred from raw timestamp data) are available for all time steps.

Given this history and the future timestamp sequence, the model generates the predicted sequence:

$$\hat{\mathbf{y}}_{T+1:T+H} = \text{TAG-ICF}(\{\boldsymbol{\tau}_t\}_{t=T+1}^{T+H} \mid \{\mathbf{x}_t, \mathbf{s}_t^{\text{raw}}, y_t\}_{t=1}^{T}) \tag{1}$$

Performance is assessed over an evaluation period $\mathbf{y}_{T+1:T_{\text{eval}}}$, using a rolling-window approach. At each step $k \in [0, \ldots T_{\text{eval}-H}]$, the model uses the sequence up to $T+k$ as context to predict the subsequent horizon $\hat{\mathbf{y}}_{T+k+1:T+k+H}$. This iterative formulation ensures that the model is evaluated on its ability to sustain predictive accuracy across the entire test duration while strictly adhering to the constraint that only historical data $\{\mathbf{x}_t, \mathbf{s}_t^{\text{raw}}, y_t\}_{t=1}^{T}$ and future timestamp features $\{\boldsymbol{\tau}_t\}_{t=T+1}^{T+H}$ are accessible at each step.

3.2 TIME SERIES ADAPTATIONS FOR ICL

To align multimodal observations with the ICL input paradigm, we first apply modality-specific transformations to numerical covariates, timestamp features, and textual data while preserving temporal dependencies. We then reformulate the forecasting task into horizon-specific context examples, where each example combines lagged numerical and textual representations with the corresponding future target values. This structured representation enables the observed history to be used directly within an ICL backbone for time-series forecasting.

Numerical representation: Given the numerical covariates $\mathbf{x}_t$ and target y_t, we construct a lagged feature vector. At time step t, we concatenate the covariates and target over lookback windows of lengths L_{num} and L_y, respectively, resulting in an initial vector in $\mathbb{R}^{L_{\text{num}} \cdot d_n + L_y}$. Furthermore, following Hoo et al. (2025), we augment the feature set with timestamp features to capture both linear and periodic temporal dynamics. When predicting the future target at time $t + h$, we use the corresponding future timestamp to construct timestamp features $\boldsymbol{\tau}_{t+h} \in \mathbb{R}^{d_\tau}$. These features include a running index representing monotonic time progression and calendar attributes (e.g., hour, day, month, etc.) encoded using sine and cosine transformations to preserve their cyclical nature. We also include m automated seasonal features extracted via Fourier Transforms to capture the top m domain-specific periodicities of the target variable. Together, these features enable the model to represent multi-scale temporal patterns within a tabular framework, resulting in the final numerical representation denoted by $\mathbf{c}_t \in \mathbb{R}^{L_{\text{num}} \cdot d_n + L_y + d_\tau}$.

Textual representation: Raw textual data $\mathbf{s}_t^{\text{raw}}$ is often highly verbose and contains noise irrelevant to the time series. We first utilize an LLM to perform an analytical synthesis of each textual input, extracting high-level insights relevant to the temporal dynamics. These insights are subsequently encoded into an embedding of dimension d_t using a pre-trained text embedding model. By applying a consistent lookback window L_{text}, we construct the textual context matrix $S_t \in \mathbb{R}^{L_{\text{text}} \times d_t}$ for each time step t. The specific LLM model, embedding model, and prompts are detailed in Appendix. B.3 and Appendix A respectively.

Horizon-specific ICL tasks: To apply the ICL backbone to multimodal forecasting, we reformulate the horizon-H prediction problem into a series of structured in-context examples.

We decompose the global forecasting task into H horizon-specific conditional prediction sub-tasks. Specifically, to predict a future value at step $T + h$ (where $h \in \{1, \ldots, H\}$), we align the historical context vectors with their corresponding future targets using a temporal offset of h. For a given horizon h, each training example in the context window consists of the triplet $(\mathbf{c}_t, S_t, y_{t+h})$. Thus, the context for ICL is defined as the set $\{(\mathbf{c}_t, S_t, y_{t+h})\}_{t=1}^{T-h}$. To obtain a prediction for h steps ahead at $T + h$, the ICL query consists of the pair $(\mathbf{c}_T, S_T)$. This allows the ICL model to learn the specific mapping between current multimodal observations and the target value h steps ahead. By iterating this process for all $h \in \{1, \ldots, H\}$, the model independently generates each point in the prediction horizon to construct the full sequence $\hat{\mathbf{y}}_{T+1:T+H}$.

We formulate the forecasting problem as an iterative process rather than a more common approach that uses a fixed context set $\{(\mathbf{c}_t, S_t, y_{t+h})\}_{t=1}^{T-H}$ and query set $\{(\mathbf{c}_t, S_t)\}_{t=T-H+1}^{T}$ to construct the entire sequence $\hat{\mathbf{y}}_{T+1:T+H}$. While the common method is computationally cheaper, our iterative formulation is designed to maximize context utilization, ensuring the ICL model leverages the maximal historical observations for every step in the prediction horizon.

Target representation: Tabular foundation models may struggle with extrapolation, particularly when time-series targets fall outside the range observed in the context and thus constitute out-of-distribution targets (Hoo et al., 2025; Loza et al., 2026). To mitigate this issue, we optionally apply target differencing, a common transformation used to reduce non-stationarity induced by strong trends in time series. Specifically, rather than directly predicting the target (y_{t+h}), we predict its first-order difference:

$$y'_{t+h} = y_{t+h} - y_{t+h-1} \tag{2}$$

Under this formulation, the context set is given by $(\mathbf{c}_t, S_t, y'_{t+h})_{t=1}^{T-h}$, while the query pair is $(\mathbf{c}_T, S_T)$. Following the procedure described in the previous section, the model predicts the full sequence of

first-order target differences $\hat{\mathbf{y}'}_{T+1:T+H}$, from which the final forecast is reconstructed recursively as

$$\hat{y}_{T+h} = y_T + \sum_{k=1}^{h} \hat{y}'_{T+k}, \qquad h = 1, \ldots, H. \tag{3}$$

Whether target differencing is applied is determined automatically based on validation-set performance. For brevity, we continue to denote the final forecast from our framework as $\hat{\mathbf{y}}_{T+1:T+H}$, regardless of whether target differencing is used.

3.3 Text Attention Guidance Module

In this section, we introduce the Text Attention Guidance (TAG) Module, the primary trainable component of our framework. TAG modulates the query–context attention of the frozen ICL backbone by integrating textual relevance scores, allowing semantic context to guide the numerical forecasting process.

Transformer-based ICL models predict targets by attending from queries to context examples. In a numerical-only forecasting backbone, the attention weight vector $\mathbf{a}_{\text{num}}^{(i,j)} \in \mathbb{R}^{1\times(T-h)}$ for the transformer layer i and the cross-attention head j is:

$$\mathbf{a}_{\text{num}}^{(i,j)} = \text{Softmax}\left(\frac{\mathbf{q}_{\text{num}}^{(i,j)}(K_{\text{num}}^{(i,j)})^\top}{\sqrt{d_{\text{num}}}}\right) \tag{4}$$

where $\mathbf{q}_{\text{num}}^{(i,j)}$ is the transformation of query at step T and $K_{\text{num}}^{(i,j)}$ is the transformation of historical context matrix (up to $T-h$) used to predict y_{T+h}. Here, d_{num} denotes the dimensionality of the feature space after projection. The resulting weights aggregate the value matrix $V_{\text{num}}^{(i,j)}$ into the attended representation $\mathbf{a}_{\text{num}}^{(i,j)}V_{\text{num}}^{(i,j)}$.

Our goal is to use textual context to guide which historical examples should be emphasized, without directly mixing textual embedding with numerical feature representation. We compute a text-conditioned relevance vector $\mathbf{z}_{\text{text}}^{(i,j)}$ from lagged textual representations. Using a shared projection $W_{\text{text}}^{(i,j)}$, we project each lagged text embedding into a similarity space and average query–context similarities:

$$\mathbf{z}_{\text{text}}^{(i,j)} = \frac{1}{L_{\text{text}}}\sum_{l=1}^{L_{\text{text}}} \frac{(\mathbf{s}_T^{(l)}W_{\text{text}}^{(i,j)})(S_{1:T-h}^{(l)}W_{\text{text}}^{(i,j)})^\top}{\sqrt{d_{\text{text}}}} \tag{5}$$

where $\mathbf{s}_T^{(l)}$ is the query vector and $S_{1:T-h}^{(l)}$, the context matrix for lag l. To maintain parameter efficiency, $W_{\text{text}}^{(i,j)}$ is shared across transformer layers i. We define the textual attention as $\mathbf{a}_{\text{text}}^{(i,j)} = \text{Softmax}(\mathbf{z}_{\text{text}}^{(i,j)})$. To integrate this into the ICL backbone, we compute a gated sum of textual and numerical attention for layer i and cross attention head j:

$$\mathbf{a}_{\text{guided}}^{(i,j)} = (1-\alpha^{(i,j)})\mathbf{a}_{\text{num}}^{(i,j)} + \alpha^{(i,j)}\mathbf{a}_{\text{text}}^{(i,j)} \tag{6}$$

The trainable parameter $\alpha^{(i,j)} \in [0,1]$ controls the strength of textual guidance, with $\mathbf{a}_{\text{guided}}^{(i,j)}$ replacing the vanilla attention $\mathbf{a}_{\text{num}}^{(i,j)}$ during value aggregation. The set of guided transformer layers i is treated as a hyperparameter. We use a scalar convex gate $\alpha^{(i,j)}$ as a deliberate design choice, as it provides a simple and interpretable interpolation between numerical and textual attention while preserving the normalization of the resulting attention distribution. This lightweight formulation introduces minimal additional model capacity, allowing us to isolate the effect of textual guidance and more clearly attribute performance gains to the proposed cross-modal mechanism rather than to increased model expressiveness. More expressive modulation mechanisms may offer further improvements, and we leave their exploration to future work.

4 Experiment

4.1 Experimental setup

Datasets: To evaluate our framework, we utilize nine real-world datasets from TimeMMD (Liu et al., 2024) spanning 2000–2023 across diverse domains, along with three additional news-associated

datasets, Bitcoin price, electricity, and the AUD/USD exchange rate, from Wang et al. (2024). For TimeMMD, we specifically employ web search results rather than scheduled reports, as their temporal density aligns with numerical data to provide contextual signals. For details on dataset construction and numerical covariates, see Liu et al. (2024) and Appendix B.1.

Evaluation Details: Datasets are divided into train, validation, and test subsets using an 80/10/10% split. We evaluate models across prediction horizons $H \in \{4, 6, 8\}$ using a rolling-window approach (stride 1), normalizing all data based on training set statistics. Performance is measured via Mean Squared Error (MSE) and Mean Absolute Error (MAE) in the normalized target space, providing a complementary assessment of large-error sensitivity and overall error magnitude.

Detailed information on the experimental setup, including dataset statistics (Appendix B.1), baseline descriptions (Appendix B.2), implementation specifics (Appendix B.3), and hyperparameter configurations (Appendix B.4) is provided in the Appendix.

Table 1: Forecasting performance of TAG-ICF compared with several baselines on several datasets for a forecast horizon of 6. Best results are shown in bold and second-best results are underlined. Average ranks are computed using all methods Table 8 in the appendix.

Benchmark	**Datasets**	GPT4MTS Jia et al. (2024)		MM-TSFlib Liu et al. (2024)		ConTextTab Spinaci et al. (2025)		TaTS Li et al. (2026)		TimeVLM Zhong et al. (2025)		Chronos-2 Ansari et al. (2025)		TAG-ICF (ours)	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Time-MMD: Multimodal Time Series Liu et al. (2024)	Agriculture	0.797	0.781	1.000	0.809	1.150	0.821	1.186	0.909	0.723	0.687	0.883	0.769	**0.621**	**0.600**
	Climate	0.215	0.351	0.187	0.334	0.234	0.355	0.207	0.353	0.212	0.351	0.210	0.340	**0.185**	**0.323**
	Economy	1.026	0.812	0.895	0.756	1.026	0.801	1.013	0.824	0.766	0.683	0.591	0.585	**0.537**	**0.578**
	Energy	0.103	0.221	0.095	0.211	0.163	0.288	0.115	0.245	0.112	0.230	0.098	0.215	**0.091**	**0.206**
	Environment	0.435	0.464	0.538	0.568	0.390	0.451	0.318	0.402	0.439	0.469	**0.313**	**0.382**	0.326	0.391
	Health	0.642	0.503	0.633	0.504	0.504	0.415	0.716	0.547	0.594	0.495	0.501	**0.389**	**0.423**	**0.389**
	Security	20.920	3.706	12.562	2.717	8.403	2.457	15.123	3.165	9.012	2.560	7.862	**2.120**	**7.681**	2.343
	Social Good	0.238	0.363	0.511	0.621	0.598	0.550	1.403	1.060	0.510	0.548	**0.064**	**0.171**	0.108	0.242
	Traffic	0.376	0.470	1.507	1.078	0.730	0.594	0.375	0.473	0.322	0.432	**0.129**	**0.296**	0.250	0.402
FNF: Forecast with News Wang et al. (2024)	Bitcoin Price	3.138	1.385	3.735	1.575	6.527	1.984	3.680	1.501	3.354	1.375	3.504	1.456	**3.127**	**1.371**
	Electricity	0.782	0.629	0.756	0.684	0.782	0.677	0.720	0.614	0.702	0.623	0.663	**0.565**	**0.646**	0.602
	Exchange Rate	0.020	0.113	**0.018**	**0.106**	0.026	0.131	0.031	0.139	0.027	0.132	**0.018**	0.107	0.019	**0.106**
Average Rank		5.42	5.00	5.50	5.83	6.75	6.42	5.83	6.42	4.42	4.33	2.33	2.00	**1.58**	**1.58**

Table 2: Impact of the Text Module on Forecast Accuracy (MSE) for a forecast horizon of 6. The full Table 9 is provided in the appendix.

Datasets	GPT4MTS			MM-TSFlib			ConTextTab			TaTS			TimeVLM			TAG-ICF (Ours)		
	w/o Text	w/ Text	Gain	w/o Text	w/ Text	Gain	w/o Text	w/ Text	Gain	w/o Text	w/ Text	Gain	w/o Text	w/ Text	Gain	w/o Text	w/ Text	Gain
Agriculture	0.914	0.797	12.80%	1.006	1.000	0.60%	1.133	1.150	-1.50%	1.177	1.186	-0.76%	0.717	0.723	-0.84%	0.694	0.621	10.52%
Climate	0.209	0.215	-2.87%	0.201	0.187	6.97%	0.230	0.234	-1.74%	0.202	0.207	-2.48%	0.213	0.212	0.47%	0.180	0.185	-2.78%
Economy	0.988	1.026	-3.85%	0.823	0.895	-8.75%	0.990	1.026	-3.64%	1.007	1.013	-0.60%	0.766	0.766	0.00%	0.576	0.537	6.77%
Energy	0.098	0.103	-5.10%	0.094	0.095	-1.06%	0.152	0.163	-7.24%	0.108	0.115	-6.48%	0.112	0.112	0.00%	0.094	0.091	3.19%
Environment	0.387	0.435	-12.40%	0.545	0.538	1.28%	0.392	0.390	0.51%	0.362	0.318	12.15%	0.439	0.439	0.00%	0.337	0.326	3.26%
Health	0.641	0.642	-0.16%	0.632	0.633	-0.16%	0.509	0.504	0.98%	0.673	0.716	-6.39%	0.597	0.594	0.50%	0.445	0.423	4.94%
Security	12.891	20.920	-62.28%	6.670	12.562	-88.34%	8.293	8.403	-1.33%	16.577	15.123	8.77%	9.025	9.012	0.14%	8.491	7.681	9.54%
Social Good	0.253	0.238	5.93%	0.743	0.511	31.22%	0.697	0.598	14.20%	1.403	1.403	0.00%	0.236	0.510	-116.10%	0.131	0.108	17.56%
Traffic	0.374	0.376	-0.53%	1.454	1.507	-3.65%	0.705	0.730	-3.55%	0.318	0.375	-17.92%	0.321	0.322	-0.31%	0.293	0.250	14.68%
Bitcoin Price	3.234	3.138	2.97%	3.482	3.735	-7.27%	6.732	6.527	3.05%	3.441	3.680	-6.95%	3.357	3.354	0.09%	3.293	3.127	5.04%
Electricity	0.739	0.782	-5.82%	0.631	0.756	-19.81%	0.782	0.782	0.00%	0.735	0.720	2.04%	0.702	0.702	0.00%	0.708	0.646	8.76%
Exchange Rate	0.021	0.020	4.76%	0.017	0.018	-5.88%	0.025	0.026	-4.00%	0.033	0.031	6.06%	0.027	0.027	0.00%	0.020	0.019	5.00%

Table 3: Impact of the Text Module on Forecast Accuracy (MAE) for a forecast horizon of 6. The full Table 10 is provided in the appendix.

Datasets	GPT4MTS			MM-TSFlib			ConTextTab			TaTS			TimeVLM			TAG-ICF (Ours)		
	w/o Text	w/ Text	Gain	w/o Text	w/ Text	Gain	w/o Text	w/ Text	Gain	w/o Text	w/ Text	Gain	w/o Text	w/ Text	Gain	w/o Text	w/ Text	Gain
Agriculture	0.813	0.781	3.94%	0.830	0.809	2.53%	0.816	0.821	-0.61%	0.909	0.909	0.00%	0.686	0.687	-0.15%	0.639	0.600	6.10%
Climate	0.345	0.351	-1.74%	0.330	0.334	-1.21%	0.350	0.355	-1.43%	0.344	0.353	-2.62%	0.352	0.351	0.28%	0.316	0.323	-2.22%
Economy	0.798	0.812	-1.75%	0.784	0.756	3.57%	0.783	0.801	-2.30%	0.820	0.824	-0.49%	0.683	0.683	0.00%	0.604	0.578	4.30%
Energy	0.217	0.221	-1.84%	0.208	0.211	-1.44%	0.257	0.288	-12.06%	0.239	0.245	-2.51%	0.230	0.230	0.00%	0.211	0.206	2.37%
Environment	0.427	0.464	-8.67%	0.547	0.568	-3.84%	0.451	0.451	0.00%	0.433	0.402	7.16%	0.469	0.469	0.00%	0.400	0.391	2.25%
Health	0.503	0.503	0.00%	0.501	0.504	-0.60%	0.413	0.415	-0.48%	0.524	0.547	-4.39%	0.493	0.495	-0.41%	0.395	0.389	1.52%
Security	2.900	3.706	-27.79%	2.137	2.717	-27.14%	2.443	2.457	-0.57%	3.414	3.165	7.29%	2.562	2.560	0.08%	2.517	2.343	6.91%
Social Good	0.375	0.363	3.20%	0.766	0.621	18.93%	0.615	0.550	10.57%	1.060	1.060	0.00%	0.410	0.548	-33.66%	0.262	0.242	7.63%
Traffic	0.442	0.470	-6.33%	1.049	1.078	-2.76%	0.584	0.594	-1.71%	0.443	0.473	-6.77%	0.431	0.432	-0.23%	0.420	0.402	4.29%
Bitcoin Price	1.373	1.385	-0.87%	1.501	1.575	-4.93%	2.043	1.984	2.89%	1.466	1.501	-2.39%	1.375	1.375	0.00%	1.384	1.371	0.94%
Electricity	0.606	0.629	-3.80%	0.606	0.684	-12.87%	0.675	0.677	-0.30%	0.620	0.614	0.97%	0.623	0.623	0.00%	0.630	0.602	4.44%
Exchange Rate	0.117	0.113	3.42%	0.104	0.106	-1.92%	0.127	0.131	-3.15%	0.142	0.139	2.11%	0.132	0.132	0.00%	0.109	0.106	2.75%

4.2 Main Results

Performance Gains over Multimodal Baselines. Table 1 summarizes the forecasting performance of TAG-ICF relative to state-of-the-art multimodal baselines on the Time-MMD benchmark and three additional datasets (forecasting horizon $H = 6$). For clarity, Table 1 reports only the strongest baseline according to average rank, while the complete results for all baselines are provided in Appendix C. We also provide extended results for varying forecasting horizons in Appendix D.1.

TAG-ICF demonstrates superior consistency across all benchmarks, achieving the leading average rank (1.58 MSE; 1.58 MAE) and outperforming all competing methods on eight datasets in MSE and seven in MAE.

On Economy, TAG-ICF reduces MSE by 29.90% and MAE by 15.37% relative to TimeVLM, the strongest competing multimodal baseline on this dataset. More importantly, when compared with the best-performing competing baseline on each individual dataset, TAG-ICF still delivers notable gains, including MSE reductions of 15.57% on Health and 14.11% on Agriculture, as well as MAE reductions of 12.66% on Agriculture and 3.29% on Climate.

Generalization Across domains. A notable finding is the substantial performance variability exhibited by existing multimodal forecasting methods across domains. For example, MM-TSFlib remains highly competitive on Climate and Energy, ranking second in both MSE and MAE, but drops substantially in relative performance on Traffic and Security. GPT4MTS and Time-VLM similarly exhibit considerable variation in their rankings across datasets, highlighting the challenge of maintaining consistent performance across domains. While TAG-ICF does not achieve the best result on every individual metric, it attains the strongest aggregate average rank (1.58 for both MSE and MAE), demonstrating consistently competitive performance across diverse datasets and domains.

Impact of Multimodal Integration. Table 2 and Table 3 quantify the value of textual modalities by comparing TAG-ICF against its "no-text" numerical baseline (forecasting horizon $H = 6$). All numerical preprocessing steps, including target differencing, are applied identically to both the TabDPT no-text baseline and TAG-ICF. This controlled comparison ensures that the gains reported in Table 2 and Table 3 can be attributed to the proposed text-attention module rather than differences in numerical preprocessing.

Overall, TAG-ICF exhibits the most consistent and stable gains from incorporating text compared with the other multimodal forcasting baselines. In particular, on Social Good, TAG-ICF improves MSE by 17.56% and MAE by 7.63%; on Traffic, it improves MSE by 14.68% and MAE by 4.29%; and on Agriculture, it improves MSE by 10.52% and MAE by 6.10%. These results indicate that the proposed framework effectively extracts complementary semantic information from the textual context rather than relying solely on numerical forecasting capacity.

4.3 Evaluating the Significance of Textual Semantics

To investigate whether the performance gains of TAG-ICF stem from meaningful multimodal learning or simply from increased model capacity (parameter count), we conduct two targeted ablation studies inspired by Li et al. (2026): a *text shuffling* test and a *random text dropping* test.

We first evaluate the model's sensitivity to textual content by randomly shuffling the textual contexts across the dataset, thereby decoupling the semantic information from the corresponding numerical time series while keeping the parameter count constant. As shown in Table 4, TAG-ICF outperforms the shuffled variants across the five evaluated datasets. Notably, the performance of the most variants is very close to the "No Text" baseline in terms of MAE, although the differences vary across datasets. This confirms that the model does not achieve gains through parameter scaling alone; rather, it actively utilizes the specific semantic content of the text to modulate attention toward relevant historical examples.

To quantify the model's dependence on textual availability, we systematically drop a percentage of the textual contexts during inference. As reported in Table 4, the MAE and MSE increases by 4.83% and 6.56%, respectively when 30% of the text is removed, suggesting a degree of robustness to sparse signals. Performance also steadily declines at a 60% drop rate, where TAG-ICF maintains a 4.42% (MAE) and 5.50% (MSE), respectively. Although the degradation is not strictly monotonic with the proportion of text removed, both dropping variants consistently underperform the full-text model across all five datasets. These results suggest that TAG-ICF benefits from the available textual guidance while remaining reasonably robust to partial text removal.

4.4 Evaluating an Alternative Tabular Foundation Model Backbone

To demonstrate the flexibility of the TAG module, we evaluate our framework with an additional tabular foundation model backbone, TabPFNv3 (Grinsztajn et al., 2026), which is among the top-

Table 4: Impact of Shuffling and Dropping Text on Forecast Accuracy.

Datasets	TAG-ICF		Shuffled		Drop 30%		Drop 60%		No Text	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Economy	0.537	0.578	0.549	0.582	0.551	0.587	0.550	0.586	0.576	0.604
Environment	0.326	0.391	0.335	0.398	0.338	0.399	0.339	0.399	0.337	0.400
Security	7.681	2.343	8.230	2.415	7.893	2.367	7.997	2.375	8.491	2.517
Social Good	0.108	0.242	0.131	0.283	0.130	0.284	0.123	0.279	0.131	0.262
Electricity	0.646	0.602	0.676	0.621	0.668	0.615	0.666	0.614	0.708	0.630

performing tabular foundation models on widely used public benchmarks such as TabArena (Erickson et al., 2026) and BeyondArena (Purucker et al., 2026). Beyond standard tabular prediction, TabPFNv3 has also demonstrated strong performance across a broader range of problem settings, highlighting its versatility beyond conventional classification and regression tasks.

We evaluate TAG-ICF with TabPFNv3 on five datasets, comparing performance with and without the textual module, as shown in Table 5. Incorporating textual information yields consistent improvements on four out of five datasets, with average gains of 3.27% in MAE and 3.79% in MSE. These results demonstrate that the benefits of TAG-ICF generalize across different tabular foundation model backbones without extensive TabPFNv3-specific hyperparameter tuning or backbone-specific optimization. Notably, TabPFNv3 is already a strong forecaster without textual information: it outperforms all baselines in Table 1 on Social Good for both MAE and MSE, as well as on Economy for MSE. Despite this strong baseline performance, TAG-ICF achieves further improvements by effectively incorporating textual context.

For further analysis, see the Appendix for: varying forecast horizons (D.1), lookback window L_{text} (D.2), embedding impact (D.3), and LLM-based raw text utility (D.4).

Table 5: Impact of TAG-ICF with TabPFN v3 Performance (Horizon 6). Bold values indicate the better error for each metric.

Dataset	MSE			MAE		
	TabPFNv3 w/o Text	TAG-ICF (TabPFNv3)	Gain	TabPFNv3 w/o Text	TAG-ICF (TabPFNv3)	Gain
Economy	0.570	**0.554**	2.81%	0.591	**0.580**	1.86%
Environment	0.329	**0.327**	0.61%	0.392	**0.383**	2.30%
Security	8.457	**7.903**	6.55%	2.439	**2.300**	5.70%
Social Good	0.057	**0.051**	10.53%	0.166	**0.158**	4.82%
Electricity	**0.770**	0.782	-1.56%	0.662	**0.651**	1.66%

5 Conclusion and Limitations

We propose TAG-ICF, an in-context forecasting framework that leverages temporally aligned textual context to modulate attention within pretrained backbones. By selectively prioritizing relevant historical patterns, TAG-ICF achieved superior average ranks (1.58 MSE, 1.58 MAE) spanning diverse domains, demonstrating that attention-level guidance offers a robust alternative to utilize textual context.

Despite these promising results, the current framework faces certain limitations. First, because the model is inherently text-dependent, it may not function optimally if the textual granularity does not match that of the numerical data. Second, the system's efficacy remains fundamentally tethered to the robustness of the underlying ICL backbone. While our module provides strategic guidance, it cannot entirely compensate for scenarios where the base temporal representations are fundamentally weak.

Several avenues remain for future work, beginning with the transition from globally learned convex gating parameters to either a dynamic, instance-aware mechanism or more expressive combination functions such as MLP-based gating. Additionally, we aim to extend our module to a broader range of ICL architectures and multimodal tabular learning settings. Ultimately, we hope this work catalyzes further research into attention-guided multimodal reasoning and adaptive context selection for more resilient in-context learning systems.

AI USE STATEMENT

In this work, we used generative AI tools to clean and reformat the textual dataset, specifically to summarize raw textual features and filter for useful and relevant textual information, as described in the paper. We did not use generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, design or provide feedback on research methodology or experiments, implement methods, support qualitative or thematic data analysis, or interpret results. Generating synthetic datasets, assisting in the writing of proofs, proposing or refining hypotheses, and assisting with translation were not applicable to this work.

Additionally, we used generative AI tools to create or modify scientific figures or images, create or edit software code, summarize or analyze existing literature, source and search for information, edit the paper to improve readability, and identify relevant literature. We reviewed all AI-assisted work.; AI-assisted code and figures were reviewed before inclusion; literature and useful information identified with AI assistance was checked and verified against the original sources; and all AI-assisted writing edits were reviewed by at least two authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We acknowledge and adhere to the conference's Code of Conduct. There are no ethical concerns associated with this work.

REPRODUCIBILITY STATEMENT

We have taken several steps to support reproducibility. Appendix A and B provide detailed experimental information, including the LLM prompts, dataset preprocessing procedures and statistics, computational resources, and hyperparameter settings. Appendix D.5 further reports results across different random seeds, together with their standard deviations, to assess variability and reproducibility deviations across runs. We also provided the source code as supplementary material. Upon acceptance, we will make the GitHub repository public, together with a README containing instructions for reproducing the experiments and running the models.

REFERENCES

Abdul Fatir Ansari et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025. URL https://arxiv.org/abs/2510.15821.

Vahid Balazadeh, Hamidreza Kamkari, Valentin Thomas, Junwei Ma, Bingru Li, Jesse C Cresswell, and Rahul Krishnan. CausalPFN: Amortized causal effect estimation via in-context learning. In *Advances in Neural Information Processing Systems*, 2025.

Sameep Chattopadhyay, Pulkit Paliwal, Sai Shankar Narasimhan, Shubhankar Agarwal, and Sandeep P. Chinchali. Context matters: Leveraging contextual features for time series forecasting, 2025. URL https://openreview.net/forum?id=xW4J2QlqRx.

Mingyue Cheng, Yiheng Chen, Qi Liu, Zhiding Liu, Yucong Luo, and Enhong Chen. Instructime: Advancing time series classification with multimodal language modeling. WSDM '25. Association for Computing Machinery, 2025. ISBN 9798400713293. doi: 10.1145/3701551.3703499.

Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Desai, David Salinas, and Frank Hutter. Tabarena: A living benchmark for machine learning on tabular data. *Advances in Neural Information Processing Systems*, 38, 2026.

Benjamin Feuer, Robin T Schirrmeister, Valeriia Cherepanova, Chinmay Hegde, Frank Hutter, Micah Goldblum, Niv Cohen, and Colin White. TuneTables: Context optimization for scalable prior-data fitted networks. In *Advances in Neural Information Processing Systems*, 2024.

Anurag Garg, Muhammad Ali, Noah Hollmann, Lennart Purucker, Samuel Müller, and Frank Hutter. Real-TabPFN: Improving tabular foundation models via continued pre-training with real-world data. In *1st ICML Workshop on Foundation Models for Structured Data*, 2025.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Benjamin Jäger, Dominik Safaric, Simone Alessi, Adrian Hayler, et al. TabPFN-2.5: Advancing the state of the art in tabular foundation models. *arXiv preprint arXiv:2511.08667*, 2025.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, et al. Tabpfn-3: Technical report. *arXiv preprint arXiv:2605.13986*, 2026.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. Large language models are zero-shot time series forecasters. *Advances in neural information processing systems*, 36:19622–19635, 2023.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations 2023*, 2023.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. From tables to time: Extending tabpfn-v2 to time series forecasting. *arXiv preprint arXiv:2501.02945*, 2025.

Furong Jia, Kevin Wang, Yixiang Zheng, Defu Cao, and Yan Liu. Gpt4mts: Prompt-based large language model for multimodal time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 23343–23351, 2024.

Yushan Jiang, Wenchao Yu, Geon Lee, Dongjin Song, Kijung Shin, Wei Cheng, Yanchi Liu, and Haifeng Chen. TimeXL: Explainable multi-modal time series prediction with LLM-in-the-loop. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=WRwr2YZ4zt.

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 *International Conference on Learning Representations (ICLR)*, 2024.

Geon Lee, Wenchao Yu, Wei Cheng, and Haifeng Chen. MoAT: Multi-modal augmented time series forecasting, 2024. URL https://openreview.net/forum?id=uRXxnoqDHH.

Geon Lee, Wenchao Yu, Kijung Shin, Wei Cheng, and Haifeng Chen. Timecap: Learning to contextualize, augment, and predict time series events with large language model agents, 2025. URL https://arxiv.org/abs/2502.11418.

Zihao Li, Xiao Lin, Zhining Liu, Jiaru Zou, Ziwei Wu, Lecheng Zheng, Dongqi Fu, Yada Zhu, Hendrik Hamann, Hanghang Tong, and Jingrui He. Language in the flow of time: Time-series-paired texts weaved into a unified temporal narrative, 2026. URL https://arxiv.org/abs/2502.08942.

Chenxi Liu, Qianxiong Xu, Hao Miao, Sun Yang, Lingzheng Zhang, Cheng Long, Ziyue Li, and Rui Zhao. Timecma: Towards llm-empowered multivariate time series forecasting via cross-modality alignment, 2025. URL https://arxiv.org/abs/2406.01638.

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 *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=fuD0h4R1IL.

Malena Loza, David Chushig-Muzo, Eva Milara, Luis Bote-Curiel, Luis Estrada-Petrocelli, and Felipe Grijalva. Empirical evaluation of out-of-distribution performance of tabular foundation models. *arXiv preprint arXiv:2607.26000*, 2026.

Junwei Ma, Apoorv Dankar, George Stein, Guangwei Yu, and Anthony Caterini. TabPFGen–tabular data generation with TabPFN. In *NeurIPS 2023 Second Table Representation Learning Workshop*, 2023.

Junwei Ma, Valentin Thomas, Guangwei Yu, and Anthony Caterini. In-context data distillation with TabPFN. *arXiv preprint arXiv:2402.06971*, 2024.

Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Alex Labach, Hamidreza Kamkari, Jesse C. Cresswell, Keyvan Golestan, Guangwei Yu, Anthony L. Caterini, and Maksims Volkovs. TabDPT: Scaling tabular foundation models on real data. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=pIZxEOZCId.

Lennart Purucker, Andrej Tschalzev, Nick Erickson, Gioia Blayer, David HolzmÃ1/4ller, Alan Arazi, Alexander Pfefferle, Mustafa Tajjar, GaÃ Varoquaux, Frank Hutter, et al. Beyond iid: How general are tabular foundation models, really? *arXiv preprint arXiv:2606.30410*, 2026.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learning on large data. In *International Conference on Machine Learning*, 2025.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. *arXiv preprint arXiv:2602.11139*, 2026.

Jake Robertson, Arik Reuter, Siyuan Guo, Noah Hollmann, Frank Hutter, and Bernhard Schölkopf. Do-PFN: In-context learning for causal effect estimation. In *Advances in Neural Information Processing Systems*, 2025.

Marco Spinaci, Marek Polewczyk, Maximilian Schambach, and Sam Thelin. Contexttab: A semantics-aware tabular in-context learner. *arXiv preprint arXiv:2506.10707*, 2025.

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*, 2024. URL https://openreview.net/forum?id=Tuh4nZVb0g.

Mingtian Tan, Mike A Merrill, Vinayak Gupta, Tim Althoff, and Thomas Hartvigsen. Are language models actually useful for time series forecasting? *Advances in Neural Information Processing Systems*, 37:60162–60191, 2024.

Jialiang Tang, Shuo Chen, Chen Gong, Jing Zhang, and Dacheng Tao. Llm-ps: Empowering large language models for time series forecasting with temporal patterns and semantics. *arXiv preprint arXiv:2503.09656*, 2025.

Boris van Breugel, Jonathan Crabbé, Rob Davis, and Mihaela van der Schaar. LaTable: Towards large tabular models. *arXiv preprint arXiv:2406.17673*, 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 12694–12702, 2025a.

Fuying Wang, Feng Wu, Yihan Tang, and Lequan Yu. Ctpd: Cross-modal temporal pattern discovery for enhanced multimodal electronic health records analysis, 2025b. URL https://arxiv.org/abs/2411.00696.

Xinlei Wang, Maike Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. From news to forecast: Integrating event analysis in llm-based time series forecasting with reflection. *Advances in Neural Information Processing Systems*, 37:58118–58153, 2024.

Andrew Robert Williams, Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Jithendaraa Subramanian, Roland Riachi, James Requeima, Alexandre Lacoste, Irina Rish, Nicolas Chapados, and Alexandre Drouin. Context is key: A benchmark for forecasting with essential textual information. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=ih2WuBT1Fn.

Xingjian Wu, Jianxin Jin, Wanghui Qiu, Peng Chen, Yang Shu, Bin Yang, and Chenjuan Guo. Aurora: Towards universal generative multimodal time series forecasting. In *ICLR*, 2026.

Qianqian Xie, Weiguang Han, Yanzhao Lai, Min Peng, and Jimin Huang. The wall street neophyte: A zero-shot analysis of chatgpt over multimodal stock movement prediction challenges. *arXiv preprint arXiv:2304.05351*, 2023.

Zhijian Xu, Hao Wang, and Qiang Xu. Intervention-aware forecasting: Breaking historical limits from a system perspective. *arXiv preprint arXiv:2405.13522*, 2024.

Hao Xue and Flora D Salim. Promptcast: A new prompt-based learning paradigm for time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6851–6864, 2023.

Xinli Yu, Zheng Chen, and Yanbin Lu. Harnessing LLMs for temporal data - a study on explainable financial time series forecasting. In Mingxuan Wang and Imed Zitouni (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-industry.69.

Xiyuan Zhang, Boran Han, Haoyang Fang, Abdul Fatir Ansari, Shuai Zhang, Danielle C Maddix, Cuixiong Hu, Andrew Gordon Wilson, Michael W Mahoney, Hao Wang, et al. When does multimodality lead to better time series forecasting? *arXiv preprint arXiv:2506.21611*, 2025.

Siru Zhong, Weilin Ruan, Ming Jin, Li Huan, Qingsong Wen, and Yuxuan Liang. Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

In this Appendix, we provide supplementary details to support the main text. We have also additionally provided further empirical results and analysis. The sections are organized as follows:

- LLM Prompt for Text Summarization(Appendix A)
- Experimental Details (Appendix B)
  - Dataset Statistics and Details (Appendix B.1)
  - Baselines (Appendix B.2)
  - Implementation Details (Appendix B.3)
  - Hyperparameters (Appendix B.4)
  - Code and Reproducibility (Appendix B.5)
  - Hardware and Compute Environment (Appendix B.6)
- Full Experiment Results (Appendix C)
- Ablations (Appendix D)
  - Effect of Different Forecast Horizons (Appendix D.1)
  - Impact of Text Lookback Window (Appendix D.2)
  - Effect of Text Embedder (Appendix D.3)
  - Effect of LLM Analyzed Textual Data (Appendix D.4)
  - Effect of Random Seed (Appendix D.5)

A LLM-Based Text Summarization

A.1 Prompt

```
You are a {domain} expert specializing in {target}.

Given a set of related news and articles from the same {week/month},
analyze the information to infer the likely impact on {target} over
the next 1 to 7 {weeks/months}.  Focus on multi-{week/month} drivers.

Instructions:

     • Provide a high-level analysis or insight based strictly on the
       provided articles, including a directional {target} forecast if
       the data supports a confident outlook.
     • Limit your response to a maximum of two sentences.
     • If specific numbers are necessary for the analysis, generalize
       highly specific, granular numbers, e.g., use "roughly 3%"
       instead of "3.12%", to maintain focus on broader trends.
     • STRICTLY AVOID extraneous introductory phrases, e.g., "The
       articles state," "Based on the news".  Start your explanation
       directly.

Articles:  {articles}
```

A.2 Example of Text Before and After LLM Summarization

Raw text:

The unemployment rate in the US is currently at 3.4%, with a record low in April; ... The unemployment rate has been decreasing, with 20 states reporting an unemployment rate under 3% in April 2023, and the overall unemployment rate being lower in 2022 than before COVID-19; the unemployment rate in the U.S. transportation sector was 3.6% (not seasonally adjusted) in May 2023, according to Bureau of Labor Statistics (BLS) data; ... the unemployment rate among Black workers is the lowest on record, with a 22-year high employment rate among Black women; ... the unemployment rate for young people is the lowest in 70 years; ... the unemployment rate for Black Americans reached historic lows in Q1 2023, but employment equality has not been achieved; ...

Summarized text:

> Unemployment is likely to remain very low and could edge slightly lower over the next 1–7 months, as broad-based labor-market tightness—near-record overall unemployment, many states below roughly 3%, and historically low joblessness among Black workers and young people—suggests sustained hiring demand across sectors such as transportation. ... implying a broadly stable rate around the mid-3% range rather than a sharp increase.

B Experiment Details

This section details the experimental setup used to evaluate the TAG-ICF framework. We first describe the nine datasets from the Time-MMD benchmark (Liu et al., 2024) and three additional datasets from Wang et al. (2024) in Appendix B.1, followed by an overview of the prompting and fusion-based baselines in Appendix B.2. Technical implementation details, including the ICL backbone and textual processing pipeline, are provided in Appendix B.3, with hyperparameter configurations listed in Appendix B.4.

B.1 Dataset Statistics and Details

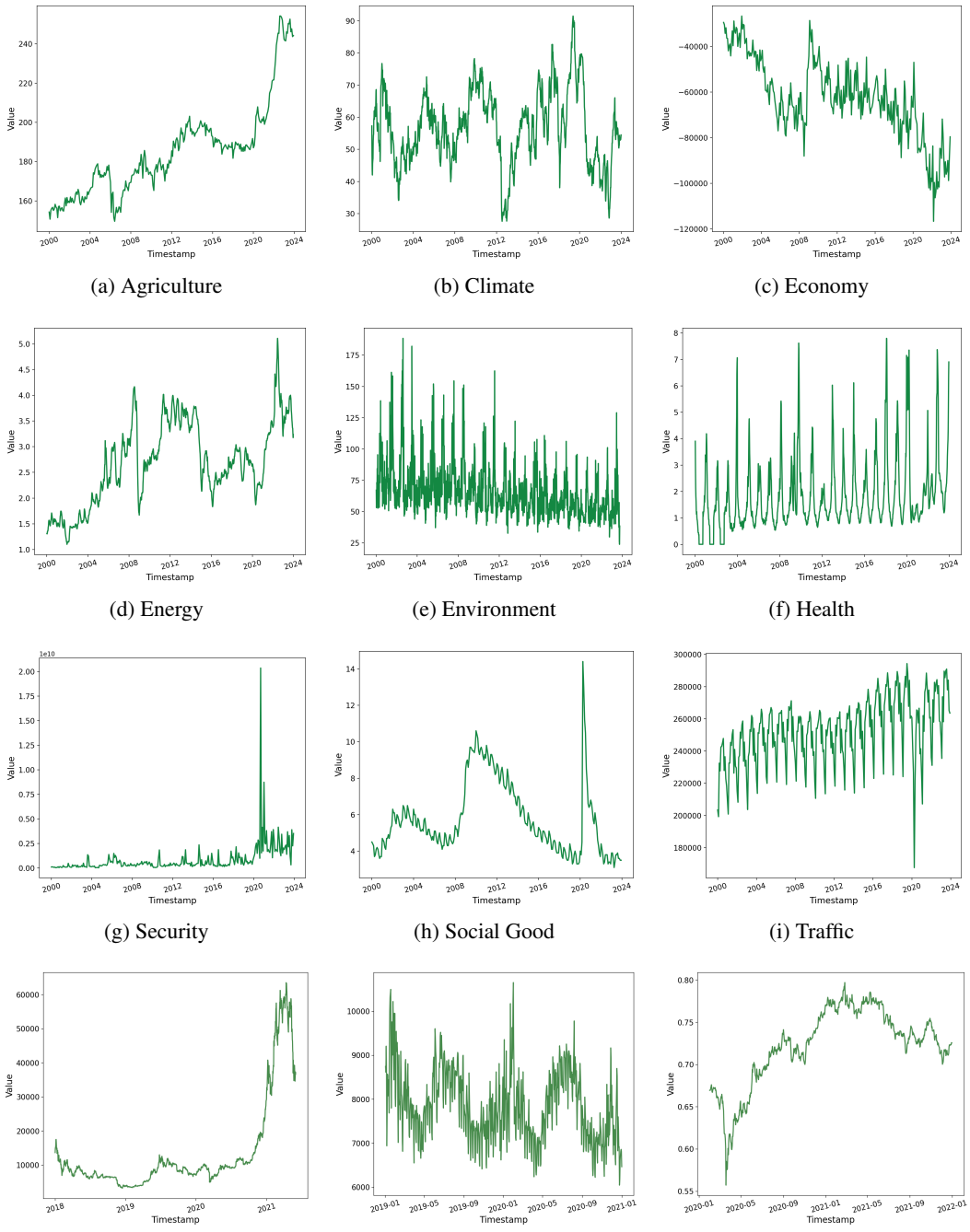
We conduct our experiments on twelve datasets: nine domains from the Time-MMD benchmark (Liu et al., 2024) and three news-associated datasets, Bitcoin price, electricity, and the AUD/USD exchange rate, from Wang et al. (2024). The Time-MMD datasets cover domains including agriculture, climate, economy, energy, environment, health, security, social good, and traffic. They are filtered to the period 2000–2023 to ensure temporal consistency across numerical and textual modalities. The environment dataset was originally available at a daily frequency; we aggregated it to a weekly frequency to match the temporal granularity of the textual data. The three additional datasets retain the daily frequency and date ranges of the CSV files used in this study. For Bitcoin, we directly join the daily price series with its associated news by date. For electricity demand, we average the 48 half-hourly readings each day and select NSW, which has the greatest coverage of region-specific news among the five regions. For exchange rates, we screen the Australian news using keywords followed by LLM filtering to identify potentially relevant economic, trade, and financial articles, select the most recent period with sufficient relevant-news coverage, and pair the retained text with daily AUD/USD observations by timestamp.

Table 6 summarizes the datasets. Dimension counts numerical variables, excluding dates and text; the number of samples counts dated observations in the corresponding dataset. The datasets vary in temporal granularity, coverage, sequence length, and the availability of associated text. This variation presents forecasting challenges involving trends, seasonality, abrupt changes, and domain-specific noise.

Table 6: Summary of the nine Time-MMD datasets and three news-associated datasets. Dimension is the number of numerical variables.

Benchmark	Dataset/Domain	Dimension	Frequency	Samples	Coverage
Time-MMD: Multimodal Time Series Liu et al. (2024)	Agriculture	1	Monthly	288	2000–2023
	Climate	5	Weekly	1252	2000–2023
	Economy	3	Monthly	288	2000–2023
	Energy	9	Weekly	1252	2000–2023
	Environment	4	Weekly	1239	2000–2023
	Health	11	Weekly	1251	2000–2023
	Security	1	Monthly	288	2000–2023
	Social Good	1	Monthly	288	2000–2023
	Traffic	1	Monthly	288	2000–2023
From News to Forecast Wang et al. (2024)	Bitcoin price	6	Daily	1248	2018–2021
	Electricity	1	Daily	731	2019–2021
	Exchange rate	7	Daily	700	2020–2021

Representative time series are shown in Figure 3. Agriculture shows an upward trend, whereas traffic exhibits regular seasonal variation around a gradually increasing level. Climate and electricity demand



(a) Agriculture (b) Climate (c) Economy

(d) Energy (e) Environment (f) Health

(g) Security (h) Social Good (i) Traffic

(j) Bitcoin Price (k) Electricity (l) Exchange Rate

Figure 3: Representative numerical time series from the nine Time-MMD domains after preprocessing and filtering to 2000–2023, together with Bitcoin closing price, daily electricity demand, and the AUD/USD exchange rate over their respective coverage periods.

display recurring fluctuations, while health has periodic peaks of varying magnitude. Environment and security contain sharp spikes, and economy, energy, and social good show pronounced changes in level or direction. Among the additional datasets, Bitcoin price rises sharply near the end of its coverage period, electricity demand fluctuates substantially, and AUD/USD declines and subsequently recovers. Together, these series span trend-dominated, seasonal, and highly variable forecasting regimes.

Time-MMD provides separate numerical and textual datasets for each domain, where the two modalities may exhibit different temporal granularities and coverage periods. To integrate these modalities, each numerical forecasting window is aligned with temporally relevant textual information using an interval-overlap strategy. Specifically, a textual record is associated with a numerical time window if its temporal intervals overlap:

$$\text{text.start_date} \leq \text{ts.end_date} \quad \wedge \quad \text{text.end_date} \geq \text{ts.start_date}.$$

This alignment strategy enables the incorporation of temporally relevant contextual information while preserving chronological consistency and preventing future information leakage.

B.2 Baselines

We compare our method against representatives from two main families of multimodal time series forecasting approaches as discussed in Section 2: prompting-based approaches and traditional fusion-based approaches. For prompting-based forecasting, we include ChatTime (Wang et al., 2025a), which adapts pretrained language models to time-series inputs through numerical tokenization and vocabulary expansion. For traditional fusion-based forecasting, we include MM-TSFLib (Liu et al., 2024), GPT4MTS (Jia et al., 2024), ConTextTab (Spinaci et al., 2025), TaTS (Li et al., 2026), Time-VLM (Zhong et al., 2025) and Aurora (Wu et al., 2026) to cover a diverse range of fusion strategies. MM-TSFLib independently encodes numerical and textual modalities and combines their respective predictions through a linear weighting mechanism; GPT4MTS adopts an encoder-decoder framework where inputs are first processed by modality-specific encoders before being passed to a GPT-2 decoder for multimodal fusion and final forecasting; and ConTextTab incorporates textual embeddings into a structured ICL model, providing a text-aware ICL comparison; Time-VLM transforms time series into visual representations and encodes textual context using a pretrained vision-language model, subsequently fusing the resulting multimodal embeddings with temporal features for forecasting; TaTS treats time-series-paired text as auxiliary variables and integrates them directly into numerical time-series models without modifying the underlying forecasting architecture; and Aurora extracts domain knowledge from textual and visual modalities and injects it into temporal representations through modality-guided multi-head self-attention for generative forecasting. To ensure a fair comparison, all baselines were fitted on the training data, with hyperparameter optimization conducted via random search across a comprehensive range of model-specific parameters.

Note that ChatTime, a prompting-based forecasting model, typically aggregates textual metadata for an entire series rather than individual timestamps. To adapt it to our setting, we strategically selected textual context from the most recent historical windows; this ensures that the model's forecasting is guided by information temporally closest to the target horizon. While this may be suboptimal as ChatTime was not natively trained in this specific configuration, it nonetheless provides a more consistent evaluation framework and ensures the model is provided with the most relevant contextual signals for the task at hand.

B.3 Implementation details

In this section, we provide a comprehensive overview of the technical configuration of our framework.

ICL Backbone and Architectural Configuration. We primarily utilize TabDPT (Ma et al., 2025) as the core ICL backbone. TabDPT is a Tabular Foundation Model designed to generalize across diverse, unseen datasets without task-specific fine-tuning, having been pre-trained on a broad collection of 123 real-world datasets encompassing 32M rows and 2B cells. The model architecture employs a Transformer encoder that treats each table row as a single token, a design choice that ensures row-order invariance and high computational efficiency. Our specific configuration consists of $B = 16$ layers with a 768-dimensional embedding space; each layer includes multi-head self-attention with 8 heads and a feedforward network with an expansion factor of 2. A critical feature of this architecture is the cross-attention split, which permits query points to attend to all context points while restricting context points to attend only to one another. We apply our textual guidance module exclusively to these cross-attention heads to intervene on the interaction between query and context, leaving the self-attention of context tokens intact. The pipeline concludes with a frozen MLP regression output head to generate our final predictions.

Textual Processing and Embedding Pipeline. Our textual components are processed through a multi-stage pipeline. We first utilize GPT-5.2 to analyse the raw text data, extracting high-level

insights grounded in the facts provided by the textual input. For vector representation, we employ Qwen3-Embedding-8B , a model trained with Matryoshka Representation Learning . This allows us to balance semantic richness with computational efficiency by extracting the first $d_t = 1024$ dimensions from the original 4096-dimensional embeddings.

Attention Gating Mechanism and Optimization Strategy. Before computing the text-guided attention scores, the text embeddings are first linearly projected through $W_{text}^{i,j}$ in eq.(5) into head-specific subspaces. Per-head LayerNorm is then applied separately to the projected query and key representations, which helps stabilize their scale within each head during training. To integrate textual guidance while maintaining numerical stability, we reparameterize the attention gating parameter α_{ij} as $\sigma(o_{ij})$, where $\sigma(\cdot)$ is the sigmoid function and o_{ij} is a randomly initialized latent parameter in the logit space. To prevent the gate from becoming saturated at the boundaries $\{0, 1\}$, we enforce a clamping constraint on o_{ij} within the range $[-5, 5]$. We employ a decoupled learning rate strategy, using a specific rate η_{gate} for the attention gating parameters to coordinate their convergence with the pre-trained weights of the backbone. Both the global learning rate and η_{gate} are treated as hyperparameters.

B.4 Hyperparameters

We list all the hyperparameters used by our pipeline in Table 7. We perform random search over the defined ranges for hyperparameter optimization.

Table 7: Sumamary of used Hyperparameter settings

Hyperparameter	Description	Value or Choices
`criterion`	The criterion for calculating loss	MAE
`optimizer`	The optimizer	AdamWScheduleFree
`tune_batch_size`	Batch sizes used for fine-tuning	{16, 32, 64, 128, 256}
`train_epochs`	Maximal number of training epochs	50
`patience`	Early stopping patience	5
`text embedding lookback window` (L_{text})	Textual data lookback window	3
`text_attn_layers` (i)	Text attention layer sets	{[1, 2], [13, 14, 15, 16], [15, 16], [16]}
`gate_lr` (η_{gate})	Learning rate for gate parameters	{0.01, 0.03, 0.1, 0.3}
`text_attn_lr`	Learning rate for text attention layers	{0.0003, 0.001, 0.003, 0.01}
`max_context`	Maximum context length	{64, 128, 256, 512, dataset length}
`dropout`	Dropout rate	0.1
`gate_logit_clamp`	Clamp value for gate logits	5.0
`text_attention_logit_l2`	L2 regularization on text attention logits	0.001
`use_flash`	Whether to use FlashAttention	True

B.5 Code and Reproducibility

To support reproducibility, the complete codebase and datasets will be publicly released upon acceptance.

B.6 Hardware and Compute Environment.

All computational experiments were executed on a system equipped with two NVIDIA RTX 6000 Ada Generation GPUs with 48 GB of VRAM.

C Full Experiment Results

Tables 8 and 9 provide the full experimental results, including all evaluated baselines and the TAG-ICF performance.

D Ablations

This section presents extended ablation studies and sensitivity analyses to rigorously evaluate the robustness and individual components of the TAG-ICF framework. We first assess the model's stability across varying temporal scales by evaluating performance at forecast horizons of 4 and 8 in

Table 8: Forecasting performance of TAG-ICF compared with baselines for a forecast horizon of 6. Best results are shown in bold and second-best results are underlined. Average ranks are computed across all methods.

Method	Metric	Time-MMD									FNF			Avg. Rank
		Agric.	Climate	Economy	Energy	Environ.	Health	Security	Social Good	Traffic	Bitcoin	Electricity	Exchange	
GPT4MTS	MSE	0.797	0.215	1.026	0.103	0.435	0.642	20.920	0.238	0.376	3.138	0.782	0.020	5.42
	MAE	0.781	0.351	0.812	0.221	0.464	0.503	3.706	0.363	0.470	1.385	0.629	0.113	5.00
MM-TSFlib	MSE	1.000	0.187	0.895	0.095	0.538	0.633	12.562	0.511	1.507	3.735	0.756	**0.018**	5.50
	MAE	0.809	0.334	0.756	0.211	0.568	0.504	2.717	0.621	1.078	1.575	0.684	**0.106**	5.83
ChatTime	MSE	1.650	0.298	2.440	0.365	0.497	0.949	27.815	0.142	0.604	4.218	0.728	0.035	8.25
	MAE	1.118	0.418	1.330	0.406	0.502	0.586	4.626	0.274	0.620	1.607	0.638	0.144	8.33
ConTextTab	MSE	1.150	0.234	1.026	0.163	0.390	0.504	8.403	0.598	0.730	6.527	0.782	0.026	6.75
	MAE	0.821	0.355	0.801	0.288	0.451	0.415	2.457	0.550	0.594	1.984	0.677	0.131	6.42
DLinear	MSE	9.773	0.252	1.245	0.146	0.329	0.826	8.615	2.682	0.396	4.924	1.127	0.034	7.92
	MAE	3.032	0.394	0.872	0.274	0.410	0.590	2.538	1.546	0.535	1.897	0.841	0.152	8.25
TaTS	MSE	1.186	0.207	1.013	0.115	0.318	0.716	15.123	1.403	0.375	3.680	0.720	0.031	5.83
	MAE	0.909	0.353	0.824	0.245	0.402	0.547	3.165	1.060	0.473	1.501	0.614	0.139	6.42
Aurora	MSE	0.911	0.286	0.997	0.617	0.509	0.808	7.772	0.082	0.621	5.124	0.757	0.027	6.67
	MAE	0.802	0.385	0.768	0.499	0.514	0.545	2.333	0.219	0.584	1.697	0.694	0.132	6.50
TimeVLM	MSE	0.723	0.212	0.766	0.112	0.439	0.594	9.012	0.510	0.322	3.354	0.702	0.027	4.42
	MAE	0.687	0.351	0.683	0.230	0.469	0.495	2.560	0.548	0.432	1.375	0.623	0.132	4.33
Chronos-2	MSE	0.883	0.210	0.591	0.098	**0.313**	0.501	7.862	**0.064**	**0.129**	3.504	0.663	**0.018**	2.33
	MAE	0.769	0.340	0.585	0.215	**0.382**	**0.389**	**2.120**	**0.171**	**0.296**	1.456	**0.565**	0.107	2.00
TAG-ICF (Ours)	MSE	**0.621**	**0.185**	**0.537**	**0.091**	0.326	**0.423**	**7.681**	0.108	0.250	**3.127**	**0.646**	0.019	**1.58**
	MAE	**0.600**	**0.323**	**0.578**	**0.206**	0.391	**0.389**	2.343	0.242	0.402	**1.371**	0.602	**0.106**	**1.58**

Table 9: Impact of the Text Module on Forecast Accuracy (MSE) for a Forecast Horizon of 6. Gains are computed from the displayed MSE values; positive gain indicates an improvement from text.

Method	Metric	Time-MMD									FNF		
		Agric.	Climate	Economy	Energy	Environ.	Health	Security	Social Good	Traffic	Bitcoin	Electricity	Exchange
GPT4MTS	w/o Text	0.914	0.209	0.988	0.098	0.387	0.641	12.891	0.253	0.374	3.234	0.739	0.021
	w/ Text	0.797	0.215	1.026	0.103	0.435	0.642	20.920	0.238	0.376	3.138	0.782	0.020
	Gain	12.80%	-2.87%	-3.85%	-5.10%	-12.40%	-0.16%	-62.28%	5.93%	-0.53%	2.97%	-5.82%	4.76%
MM-TSFlib	w/o Text	1.006	0.201	0.823	0.094	0.545	0.632	6.670	0.743	1.454	3.482	0.631	0.017
	w/ Text	1.000	0.187	0.895	0.095	0.538	0.633	12.562	0.511	1.507	3.735	0.756	0.018
	Gain	0.60%	6.97%	-8.75%	-1.06%	1.28%	-0.16%	-88.34%	31.22%	-3.65%	-7.27%	-19.81%	-5.88%
ChatTime	w/o Text	0.842	0.212	0.785	0.104	0.572	0.667	12.180	0.106	0.536	43.406	0.620	0.200
	w/ Text	1.650	0.298	2.440	0.365	0.497	0.949	27.815	0.142	0.604	4.218	0.728	0.035
	Gain	-95.96%	-40.57%	-210.83%	-250.96%	13.11%	-42.28%	-128.37%	-33.96%	-12.69%	90.28%	-17.42%	82.50%
ConTextTab	w/o Text	1.133	0.230	0.990	0.152	0.392	0.509	8.293	0.697	0.705	6.732	0.782	0.025
	w/ Text	1.150	0.234	1.026	0.163	0.390	0.504	8.403	0.598	0.730	6.527	0.782	0.026
	Gain	-1.50%	-1.74%	-3.64%	-7.24%	0.51%	0.98%	-1.33%	14.20%	-3.55%	3.05%	0.00%	-4.00%
TaTS	w/o Text	1.177	0.202	1.007	0.108	0.362	0.673	16.577	1.403	0.318	3.441	0.735	0.033
	w/ Text	1.186	0.207	1.013	0.115	0.318	0.716	15.123	1.403	0.375	3.680	0.720	0.031
	Gain	-0.76%	-2.48%	-0.60%	-6.48%	12.15%	-6.39%	8.77%	0.00%	-17.92%	-6.95%	2.04%	6.06%
Aurora	w/o Text	0.920	0.290	1.006	0.647	0.510	0.811	7.721	0.082	0.618	5.077	0.754	0.027
	w/ Text	0.911	0.286	0.997	0.617	0.509	0.808	7.772	0.082	0.621	5.124	0.757	0.027
	Gain	0.98%	1.38%	0.89%	4.64%	0.20%	0.37%	-0.66%	0.00%	-0.49%	-0.93%	-0.40%	0.00%
TimeVLM	w/o Text	0.717	0.213	0.766	0.112	0.439	0.597	9.025	0.236	0.321	3.357	0.702	0.027
	w/ Text	0.723	0.212	0.766	0.112	0.439	0.594	9.012	0.510	0.322	3.354	0.702	0.027
	Gain	-0.84%	0.47%	0.00%	0.00%	0.00%	0.50%	0.14%	-116.10%	-0.31%	0.09%	0.00%	0.00%
TAG-ICF (Ours)	w/o Text	0.694	0.180	0.576	0.094	0.337	0.445	8.491	0.131	0.293	3.293	0.708	0.020
	w/ Text	0.621	0.185	0.537	0.091	0.326	0.423	7.681	0.108	0.250	3.127	0.646	0.019
	Gain	10.52%	-2.78%	6.77%	3.19%	3.26%	4.94%	9.54%	17.56%	14.68%	5.04%	8.76%	5.00%

Appendix D.1. We also investigate the model's sensitivity to technical configurations in Appendix D.2 and Appendix D.3, specifically analyzing the impact of the text lookback window and the choice of embedding models. In addition, Appendix D.4 examines the effectiveness of using LLMs to analyze and compress raw textual data prior to forecasting. Finally, we evaluate performance across multiple random seeds to assess the stability and reproducibility of the model in Appendix D.5.

D.1 Effect of Different Forecast Horizons

Table 11 and Table 12 report the forecasting performance of TAG-ICF against several recent multimodal forecasting baselines on the Time-MMD benchmark for forecasting horizons of 4 and 8, respectively. Consistent with the results observed for horizon 6, TAG-ICF achieves the strongest and most stable performance across datasets, obtaining the best average rank in both MSE and MAE.

D.2 Impact of Text Lookback Window

Table 13 evaluates the performance of TAG-ICF on different text lookback windows (L_{text}). Overall, our model demonstrates stable performance across different text lookback window, indicating that performance improvements are achievable in general cases. For security, performance improves gradually as the L_{text} increases from 1 to 5, with MSE decreasing from 8.489 to 7.544 and MAE decreasing from 2.490 to 2.320.

For environment, social good and electricity, the best result is obtained at $L_{\text{text}} = 3$, while $L_{\text{text}} = 5$ only slightly degrades performance compared with $L_{\text{text}} = 3$. For economy, $L_{\text{text}} = 3$ gives the lowest MSE while $L_{\text{text}} = 5$ gives a slightly lower MAE. This suggests that although the optimal lookback window may vary across domains, our model still remains robust to different L_{text} choices. In particular, the performance differences between $L_{\text{text}} = 3$ and $L_{\text{text}} = 5$ are relatively small across the evaluated datasets, showing that the model can effectively leverage historical textual signals without being limited to a specific L_{text} value.

D.3 Effect of Text Embedder

Table 14 shows how different text embedding models influence the forecasting accuracy, but the performance of TAG-ICF does not collapse when replacing Qwen3-8B with the smaller Qwen3-0.6B model. Qwen3-0.6B achieves lower MAE on economy and social goods and lower MSE on Economy and Social Goods, suggesting that swapping to a smaller model in the same model family does not degrade the quality of the embedded textual signals used by the forecasting model. In the meantime, changing the embedding model to Nemotron-8B, a similar-size embedding model from a different model family, TAG-ICF still maintain similar performance on economy and social goods. Nemotron-8B improves both metrics on security. On social goods, it improves MSE but worsens MAE. On economy, MSE does not change and MAE increases slightly. Environment and Electricity get slightly worse performance on both MSE and MAE. This shows that our proposed framework has the generality and flexibility for different embedding models sizes and families. Overall, these results suggest that TAG-ICF is reasonably robust to embedding model selection.

D.4 Effect of LLM Analyzed Textual Data

Table 15 evaluates whether TAG-ICF depends on carefully processed textual inputs or can also benefit from direct embedding from noisy real-world text. LLM-analyzed text are cleaned and processed by GPT-5.2 using the prompt in Appendix A.1. Using LLM-Analyzed text improves both MAE and MSE across all five datasets, suggesting that text cleaning and semantic compression can help the model leverage more forecasting-relevant signals. Furthermore, the performance of our model using raw-text remains competitive, indicating that the performance benefit of the text module does not disappear when using less curated news-style inputs. Overall, these results suggest that TAG-ICF is robust to text quality, while LLM-based analysis provides a modest but consistent advantage on performance.

Table 10: Impact of the Text Module on Forecast Accuracy (MAE) for a Forecast Horizon of 6. Gains are computed from the displayed MAE values; positive gain indicates an improvement from text.

Method	Metric	Time-MMD									FNF		
		Agric.	Climate	Economy	Energy	Environ.	Health	Security	Social Good	Traffic	Bitcoin	Electricity	Exchange
GPT4MTS	w/o Text	0.813	0.345	0.798	0.217	0.427	0.503	2.900	0.375	0.442	1.373	0.606	0.117
	w/ Text	0.781	0.351	0.812	0.221	0.464	0.503	3.706	0.363	0.470	1.385	0.629	0.113
	Gain	3.94%	-1.74%	-1.75%	-1.84%	-8.67%	0.00%	-27.79%	3.20%	-6.33%	-0.87%	-3.80%	3.42%
MM-TSFlib	w/o Text	0.830	0.330	0.784	0.208	0.547	0.501	2.137	0.766	1.049	1.501	0.606	0.104
	w/ Text	0.809	0.334	0.756	0.211	0.568	0.504	2.717	0.621	1.078	1.575	0.684	0.106
	Gain	2.53%	-1.21%	3.57%	-1.44%	-3.84%	-0.60%	-27.14%	18.93%	-2.76%	-4.93%	-12.87%	-1.92%
ChatTime	w/o Text	0.724	0.346	0.708	0.224	0.528	0.505	2.698	0.242	0.558	1.416	0.602	0.112
	w/ Text	1.118	0.418	1.330	0.406	0.502	0.586	4.626	0.274	0.620	1.607	0.638	0.144
	Gain	-54.42%	-20.81%	-87.85%	-81.25%	4.92%	-16.04%	-71.46%	-13.22%	-11.11%	-13.49%	-5.98%	-28.57%
ConTextTab	w/o Text	0.816	0.350	0.783	0.257	0.451	0.413	2.443	0.615	0.584	2.043	0.675	0.127
	w/ Text	0.821	0.355	0.801	0.288	0.451	0.415	2.457	0.550	0.594	1.984	0.677	0.131
	Gain	-0.61%	-1.43%	-2.30%	-12.06%	0.00%	-0.48%	-0.57%	10.57%	-1.71%	2.89%	-0.30%	-3.15%
TaTS	w/o Text	0.909	0.344	0.820	0.239	0.433	0.524	3.414	1.060	0.443	1.466	0.620	0.142
	w/ Text	0.909	0.353	0.824	0.245	0.402	0.547	3.165	1.060	0.473	1.501	0.614	0.139
	Gain	0.00%	-2.62%	-0.49%	-2.51%	7.16%	-4.39%	7.29%	0.00%	-6.77%	-2.39%	0.97%	2.11%
Aurora	w/o Text	0.809	0.392	0.772	0.509	0.515	0.546	2.328	0.219	0.581	1.677	0.692	0.130
	w/ Text	0.802	0.385	0.768	0.499	0.514	0.545	2.333	0.219	0.584	1.697	0.694	0.132
	Gain	0.87%	1.79%	0.52%	1.96%	0.19%	0.18%	-0.21%	0.00%	-0.52%	-1.19%	-0.29%	-1.54%
TimeVLM	w/o Text	0.686	0.352	0.683	0.230	0.469	0.493	2.562	0.410	0.431	1.375	0.623	0.132
	w/ Text	0.687	0.351	0.683	0.230	0.469	0.495	2.560	0.548	0.432	1.375	0.623	0.132
	Gain	-0.15%	0.28%	0.00%	0.00%	0.00%	-0.41%	0.08%	-33.66%	-0.23%	0.00%	0.00%	0.00%
TAG-ICF (Ours)	w/o Text	0.639	0.316	0.604	0.211	0.400	0.395	2.517	0.262	0.420	1.384	0.630	0.109
	w/ Text	0.600	0.323	0.578	0.206	0.391	0.389	2.343	0.242	0.402	1.371	0.602	0.106
	Gain	6.10%	-2.22%	4.30%	2.37%	2.25%	1.52%	6.91%	7.63%	4.29%	0.94%	4.44%	2.75%

Table 11: Forecasting performance of TAG-ICF compared with baselines for a forecast horizon of 4. Best results are shown in bold and second-best results are underlined.

Method	Metric	Time-MMD									FNF			Avg. Rank
		Agric.	Climate	Economy	Energy	Environ.	Health US	Security	Social Good	Traffic	Bitcoin	Electricity	Exchange	
GPT4MTS	MSE	0.545	0.130	0.846	0.055	0.410	0.435	20.964	0.186	0.528	2.121	0.685	0.013	5.667
	MAE	0.620	0.267	0.736	0.163	0.452	0.414	3.828	0.316	0.573	1.132	0.578	0.089	6.222
MM-TSFlib	MSE	0.712	0.129	0.712	0.055	0.469	0.453	11.894	0.764	1.260	2.563	0.678	**0.013**	5.750
	MAE	0.670	0.258	0.670	**0.158**	0.519	0.433	2.626	0.763	0.979	1.309	0.648	0.088	6.000
ChatTime	MSE	1.103	0.197	1.916	0.215	0.486	0.712	27.553	0.158	0.599	3.537	0.621	0.025	8.889
	MAE	0.922	0.338	1.153	0.318	0.499	0.482	4.582	0.283	0.620	1.359	0.590	0.123	8.667
ConTextTab	MSE	0.709	0.133	0.785	0.108	0.369	0.322	8.377	0.348	0.640	5.442	0.685	0.020	6.000
	MAE	0.684	0.265	0.699	0.221	0.443	0.321	2.486	0.374	0.526	1.793	0.645	0.113	5.833
DLinear	MSE	9.282	0.182	0.907	0.105	0.330	0.616	8.601	0.736	0.366	3.581	1.059	0.025	7.083
	MAE	2.946	0.335	0.728	0.231	0.411	0.507	2.513	0.738	0.510	1.637	0.818	0.129	7.500
TaTS	MSE	0.514	0.141	0.915	0.072	0.336	0.448	8.434	0.619	0.261	2.812	0.674	0.023	5.333
	MAE	0.613	0.285	0.782	0.191	0.411	0.402	2.483	0.528	0.408	1.307	0.586	0.124	5.083
Aurora	MSE	0.595	0.197	0.804	0.489	0.481	0.643	11.770	0.124	0.495	3.794	0.651	0.021	6.833
	MAE	0.674	0.314	0.685	0.435	0.501	0.486	2.743	0.276	0.594	1.491	0.639	0.114	7.083
TimeVLM	MSE	0.519	0.147	0.646	0.063	0.393	0.381	8.247	0.110	0.346	**2.035**	0.682	0.017	4.167
	MAE	0.615	0.295	0.630	0.173	0.439	0.389	2.430	0.270	0.454	1.111	0.582	0.099	3.917
Chronos-2	MSE	0.486	0.135	0.505	0.058	**0.316**	0.320	7.968	**0.040**	**0.121**	2.331	0.597	**0.013**	2.167
	MAE	0.582	0.266	**0.533**	0.163	**0.384**	**0.306**	**2.154**	**0.136**	**0.279**	1.197	**0.539**	**0.087**	1.667
TAG-ICF (Ours)	MSE	**0.305**	**0.118**	**0.462**	**0.053**	0.336	**0.287**	**7.597**	0.064	0.272	2.066	**0.587**	**0.013**	**1.500**
	MAE	**0.427**	**0.248**	0.537	**0.158**	0.396	**0.306**	2.234	0.190	0.413	**1.106**	0.569	**0.087**	**1.583**

Table 12: Forecasting performance of TAG-ICF compared with baselines for a forecast horizon of 8. Best results are shown in bold and second-best results are underlined.

Method	Metric	Time-MMD									FNF			Avg. Rank
		Agric.	Climate	Economy	Energy	Environ.	Health US	Security	Social Good	Traffic	Bitcoin	Electricity	Exchange	
GPT4MTS	MSE	1.309	0.303	1.187	0.148	0.503	0.913	53.772	0.211	0.552	4.305	0.764	0.028	6.11
	MAE	0.951	0.421	0.897	0.267	0.499	0.618	6.128	0.359	0.537	1.573	0.652	0.137	6.22
MM-TSFlib	MSE	1.411	0.269	0.985	**0.138**	0.596	0.828	11.549	0.722	0.914	5.188	0.837	0.024	5.08
	MAE	0.972	**0.385**	0.789	0.259	0.614	0.581	2.569	0.779	0.816	1.854	0.724	0.131	5.25
ChatTime	MSE	2.361	0.414	2.729	0.524	0.496	1.156	27.464	0.227	0.589	7.829	0.917	0.060	8.44
	MAE	1.340	0.495	1.444	0.483	0.496	0.682	4.592	0.338	0.604	2.258	0.694	0.192	8.33
ConTextTab	MSE	1.780	0.375	1.277	0.226	0.402	0.692	8.726	0.850	0.714	7.712	0.878	0.032	6.67
	MAE	0.968	0.444	0.899	0.349	0.456	0.513	2.498	0.712	0.613	2.177	0.711	0.146	5.92
DLinear	MSE	10.232	0.321	1.601	0.186	0.327	1.018	16.388	2.864	0.391	6.389	1.251	0.091	7.08
	MAE	3.109	0.445	1.009	0.311	0.407	0.653	3.319	1.605	0.486	2.137	0.876	0.256	7.33
TaTS	MSE	1.268	0.330	1.058	0.151	0.356	0.863	18.317	1.457	0.400	6.082	0.788	0.042	5.67
	MAE	0.913	0.449	0.848	0.277	0.412	0.605	3.571	1.080	0.457	1.972	0.642	0.164	5.67
Aurora	MSE	1.245	0.366	1.113	0.187	0.496	0.919	7.582	0.115	0.670	8.086	0.848	0.032	5.67
	MAE	0.919	0.450	0.824	0.319	0.505	0.582	2.313	0.253	0.616	2.233	0.733	0.146	6.00
TimeVLM	MSE	**0.997**	0.311	0.864	0.158	0.458	0.825	10.057	0.770	0.327	4.408	0.763	0.036	4.17
	MAE	**0.790**	0.427	0.734	0.272	0.472	0.585	2.691	0.690	0.440	1.594	0.650	0.153	4.25
Chronos-2	MSE	1.355	0.282	0.687	0.140	**0.314**	0.668	7.951	**0.082**	**0.139**	4.756	**0.718**	**0.023**	2.27
	MAE	0.972	0.402	**0.625**	0.260	**0.381**	0.465	**2.110**	**0.197**	**0.314**	1.694	**0.588**	0.126	2.18
TAG-ICF (Ours)	MSE	1.250	**0.261**	**0.625**	0.139	0.330	**0.630**	**6.949**	0.182	0.269	**4.288**	0.748	0.024	**1.83**
	MAE	0.833	**0.385**	0.639	**0.256**	0.392	**0.507**	2.230	0.321	0.426	**1.559**	0.651	**0.124**	**1.83**

Table 13: Impact of Text Lookback Window on Forecast Accuracy

Datasets	$L_{\text{text}} = 1$		$L_{\text{text}} = 3$		$L_{\text{text}} = 5$	
	MSE	MAE	MSE	MAE	MSE	MAE
Economy	0.561	0.589	0.537	0.578	0.540	0.577
Environment	0.335	0.399	0.326	0.391	0.334	0.398
Security	8.489	2.490	7.681	2.343	7.544	2.320
Social Good	0.143	0.292	0.108	0.242	0.127	0.271
Electricity	0.654	0.612	0.646	0.602	0.673	0.618

Table 14: Impact of Text Embedder Selection on Forecast Accuracy.

Datasets	Qwen-8B		Qwen-0.6B		Nemotron-8B	
	MSE	MAE	MSE	MAE	MSE	MAE
Economy	0.537	0.578	0.534	0.576	0.537	0.580
Environment	0.326	0.391	0.330	0.394	0.346	0.406
Security	7.681	2.343	7.851	2.406	7.376	2.298
Social Good	0.108	0.242	0.100	0.239	0.106	0.250
Electricity	0.646	0.602	0.672	0.616	0.660	0.607

D.5 Effect of Random Seed

Table 16 examines the sensitivity of TAG-ICF to random seeds across five datasets. The three-seed mean remains close to the reported TAG-ICF result for both MSE and MAE. The largest MSE standard deviation is 0.069 on Security, relative to a mean of 7.641; the largest MAE standard deviation is 0.009 on Social Good, relative to a mean of 0.246. These small variations suggest that TAG-ICF's forecasting performance is stable across the evaluated seeds and that the reported results are not driven by a single favorable run.

Table 15: Impact of LLM based Analysis of Textual Data

Datasets	Raw Text		LLM Analyzed Text	
	MSE	MAE	MSE	MAE
Economy	0.547	0.581	0.537	0.578
Environment	0.333	0.398	0.326	0.391
Security	8.635	2.447	7.681	2.343
Social Good	0.129	0.277	0.108	0.242
Electricity	0.669	0.613	0.646	0.602

Table 16: Comparison of Fixed TAG-ICF References and Seed Results.

Datasets	TAG-ICF		3 Seed Average	
	MSE	MAE	MSE (Mean $\pm$ STD)	MAE (Mean $\pm$ STD)
Economy	0.537	0.578	0.541 ± 0.007	0.580 ± 0.003
Environment	0.326	0.391	0.328 ± 0.004	0.393 ± 0.004
Security	7.681	2.343	7.641 ± 0.069	2.339 ± 0.007
Social Good	0.108	0.242	0.109 ± 0.005	0.246 ± 0.009
Electricity	0.646	0.602	0.650 ± 0.003	0.603 ± 0.001