Spectral Text Fusion: A Frequency-Aware Approach to Multimodal Time-Series Forecasting

Huu Hiep Nguyen **Minh Hoang Nguyen** **Dung Nguyen** **Hung Le**
Applied Artificial Intelligence Initiative
Deakin University
Geelong, Australia

Abstract

Multimodal time series forecasting is crucial in real-world applications, where decisions depend on both numerical data and contextual signals. The core challenge is to effectively combine temporal numerical patterns with the context embedded in other modalities, such as text. While most existing methods align textual features with time-series patterns one step at a time, they neglect the multiscale temporal influences of contextual information such as time-series cycles and dynamic shifts. This mismatch between local alignment and global textual context can be addressed by spectral decomposition, which separates time series into frequency components capturing both short-term changes and long-term trends. In this paper, we propose SpecTF, a simple yet effective framework that integrates the effect of textual data on time series in the frequency domain. Our method extracts textual embeddings, projects them into the frequency domain, and fuses them with the time series' spectral components using a lightweight cross-attention mechanism. This adaptively reweights frequency bands based on textual relevance before mapping the results back to the temporal domain for predictions. Experimental results demonstrate that SpecTF significantly outperforms state-of-the-art models across diverse multi-modal time series datasets while utilizing considerably fewer parameters. Code is available at `https://github.com/hiepnh137/SpecTF`.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

1 INTRODUCTION

Time series forecasting plays a critical role across numerous domains, including finance [Sezer et al., 2019] [Mondal et al., 2014], healthcare [Zhang et al., 2024a] [Kaushik et al., 2020], energy [Deb et al., 2017] [Koprinska et al., 2018], material science [Le et al., 2025], and economics [King, 1965] [Franses, 1998], where accurate predictions enable informed decision-making. While traditional forecasting approaches rely solely on historical numerical data, real-world time series are often influenced by external contextual information available as text-news articles, policy documents, social media, and domain-specific reports. This textual information frequently provides crucial insights into factors driving temporal dynamics that pure numerical analysis cannot capture [Liu et al., 2025].

Although recent efforts have been made to incorporate multi-modality into time series forecasting, existing approaches remain limited. MM-TSFlib [Liu et al., 2024a] employs a late-fusion architecture that views the whole text sequence as a vector by a separate pipeline with time series before combining outputs with a simple linear weighting mechanism, neglecting the cross-modal interaction. Meanwhile, TaTS [Li et al., 2025] relies on the text-periodicity assumption to align text with time series by treating time series-paired texts as auxiliary variables of the time series itself. We argue that they still cannot exploit the textual information well, which frequently encodes contextual signals that simultaneously influence multiple temporal scales. For instance, a central bank policy announcement may trigger immediate market volatility (high-frequency components) while also initiating a prolonged economic trend (low-frequency components). Similarly, news of supply chain disruptions might cause both short-term price spikes and longer-term inventory adjustments.

To overcome the above problems, we explore a novel direction to model the multiple temporal scale influ-

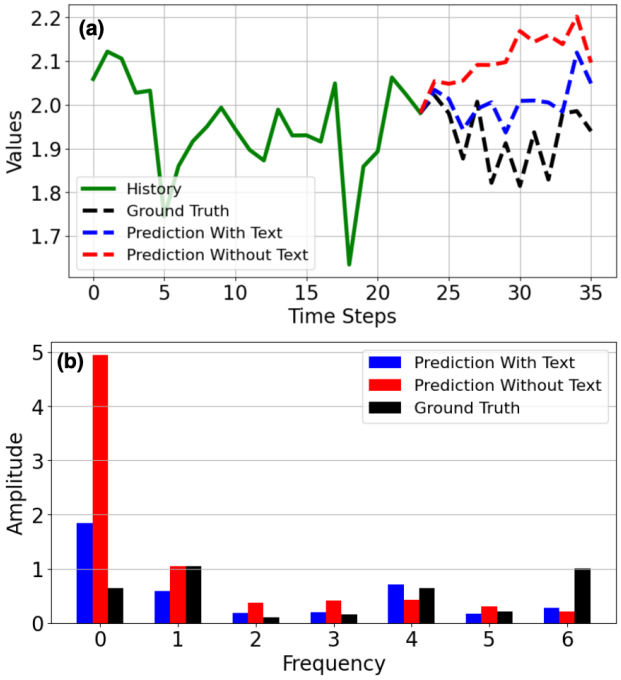


Figure 1: An example of SpecTF's predictions in Agriculture data. We provide both time-domain comparison (a) and frequency-domain comparison (b) in the predictions when incorporating text, compared to when it is not used.

ence through the spectral domain, where the influence at high-frequency components represents short-term volatility and the influence at low-frequency components captures prolonged trends. This spectral perspective enables a more nuanced integration of textual context with time series dynamics at their appropriate temporal scales while maintaining parameter efficiency through energy compaction [Yi et al., 2023], as real-world time series often concentrate energy in fewer frequency components.

Specifically, we propose Spectral Text Fusion (SpecTF), a novel framework that models the effect of textual data on time series in the frequency domain. SpecTF decomposes time series into spectral components and explicitly models how textual information modulates distinct frequency bands. As illustrated in Figure 1, our approach demonstrates superior forecasting performance when incorporating text, compared to when it is not used in the Agriculture dataset. To visualize in the spectral domain, we first normalize predictions and the ground truth by the historical mean and variance before using the Fourier Transform [Sundararajan, 2001] to convert to the frequency domain. The time-domain visualization (Figure 1a) shows that predictions incorporating textual information (blue dashed line) align more closely with the ground truth while the frequency-domain visualization (Figure 1b) reveals that these predictions more accurately match the ground truth across most frequency components, particularly in the 0th, 2nd, 3rd, 4th, 5th, and 6th frequency bands.

The proposed SpecTF framework consists of five components that process multimodal data through the frequency domain (see Figure 2). The Time Series Embedding module represents frequency components of the time series by complex vectors. Concurrently, the Text Embedding module processes textual data using a pretrained language model and projects the resulting embeddings into a complex frequency space. The heart of our approach lies in the Frequency Cross-Modality Fusion module (FreqCMF), which implements frequency-domain cross-modal fusion through a frequency attention mechanism that adaptively reweights spectral components based on the relevant textual information. Finally, the Forecaster maps the fused representations to prediction frequency representations before the Projection module converts results back to the temporal domain using the inverse Fourier Transform.

Our contributions can be summarized as follows:

- We pioneer a frequency-based approach for multimodal time series that explicitly models the multiscale temporal influence of text on time series by fusing textual and time series data in the spectral domain.
- We develop the Frequency Cross-Modality Fusion mechanism (FreqCMF) that uses attention to query relevant textual features from frequency components, enabling adaptive integration of text information based on frequency-specific relevance.
- We demonstrate through comprehensive experiments on the Time-MMD benchmark that our method not only surpasses existing multimodal forecasting approaches but also outperforms state-of-the-art time series models adapted for multimodal settings, across diverse domains, achieving state-of-the-art performance while using significantly fewer parameters.

2 RELATED WORKS

Unimodal time series forecasting. Traditional time series forecasting relies on statistical models to capture temporal patterns. ARIMA [Nandutu et al., 2022] [Mondal et al., 2014] (seasonal variant) models linear dependencies using autoregressive and moving average terms. Modern architectures leverage self-attention to model temporal dependencies at scale.

Temporal-based models like PatchTST [Nie et al., 2023] segment time series into patches for local feature extraction, while iTransformer [Liu et al., 2024b] inverts token-channel relationships to improve noise robustness. TimeMixer [Wang et al., 2024a] introduces a multiscale mixing architecture that decomposes time series into seasonal and trend components across multiple sampling scales, applying bottom-up seasonal mixing and top-down trend mixing to disentangle intricate temporal variations. Despite their impressive performance, these methods cannot exploit the external information from other modalities like text, which carries rich semantic information for time-series analysis [Liu et al., 2024a, Zhang et al., 2024b].

Frequency-domain approaches in time series. The application of frequency-domain techniques to time series forecasting has gained renewed interest due to their ability to isolate multiscale temporal patterns through spectral decomposition [Yi et al., 2025]. Auto-Former [Wu et al., 2021] proposes the auto-correlation mechanism using Fourier transforms, and FEDformer [Zhou et al., 2022] employs Fourier-enhanced attention for long-term forecasting. FreTS [Yi et al., 2023] further optimizes frequency-domain learning by redesigning MLPs to exploit global dependencies and energy concentration in spectral components. FITS [Xu et al., 2024] introduces a lightweight framework using complex-valued neural networks for frequency-domain interpolation. However, these approaches focus exclusively on unimodal numerical series, leaving unresolved the critical challenge of modeling the interaction between auxiliary modalities (like text), and time series data in spectral space. Our SpecTF addresses this gap as the first framework to enable frequency-aware multimodal time series forecasting through an attention-based cross-modality fusion mechanism to model the influence of textual data in different frequency components.

Multimodal time series forecasting with textual data. Multimodal approaches address the limitations of unimodal methods by integrating complementary data sources [Jiang et al., 2025]. Text2Freq [Lo et al., 2024] pioneers frequency-domain alignment but it utilizes textual data that essentially serves as a description or representation of the time series itself; the text is closely tied to the numerical data and does not provide additional or external contextual information. MM-TSFlib [Liu et al., 2024a] linearly combines the predictions from text and time series from separate models, neglecting the cross-modal interaction. TaTS [Li et al., 2025] projects text embeddings as auxiliary time series variables but conflates their influence across frequencies. TimeXer [Wang et al., 2024b] represents a significant advancement by empowering canonical Transformers, employing a global token mechanism to bridge information between exogenous and endogenous variables by using the attention mechanism at the temporal domain. While effective for general information transfer, this approach treats the influence of exogenous variables uniformly across all temporal dynamics. Compared to TaTS or TimeXer, which use time-domain fusion, our SpecTF, using cross-attention in the frequency domain, explicitly models the influence of textual data at multiple temporal scales through different frequency components.

3 BACKGROUNDS

3.1 Multimodal Time Series Forecasting

In the context of multimodal time series forecasting, the primary goal is to predict future values of a time series while utilizing additional modalities that are temporally paired with the time series. This paper focuses specifically on text-enhanced forecasting. In particular, given a collection of multivariate time series samples and textual data with a lookback window L, the number of variables N, we denote the input as pairs $(x_1, s_1), \ldots, (x_L, s_L)$, where $X^h = [x_1, \ldots, x_L] \in \mathbb{R}^{N \times L}$ is the numerical data and $S^h = [s_1, \ldots, s_L]$ denotes the associated textual data. The objective of multimodal time series forecasting is to predict future numerical values $Y = [x_{L+1}, \ldots, x_{L+H}] \in \mathbb{R}^{N \times H}$, where H is the forecast horizon.

3.2 Frequency Domain in Time Series

Frequency domain analysis of time series data provides a powerful framework for decomposing temporal signals into constituent periodic components. This approach offers significant advantages, particularly for capturing multi-scale patterns and global dependencies. The Discrete Fourier Transform (DFT) [Sundararajan, 2001] serves as the fundamental mathematical tool for this transformation, mapping a discrete time series $X = \{x_t\}_{t=0}^{n-1}$ into frequency domain $\hat{X} = \{\mathcal{X}_k\}_{k=0}^{n-1}$:

$$\begin{aligned} \hat{X} &= \{\mathcal{X}_k\}_{k=0}^{n-1} = \mathrm{DFT}(X), \\ \mathcal{X}_k &= \sum_{t=0}^{n-1} x_t \exp -2\pi j \frac{kt}{n} \end{aligned} \tag{1}$$

where $\mathcal{X}_k$ represents the complex amplitude and phase of the frequency component at $\omega_k = \frac{2\pi k}{n}$. The inverse Discrete Fourier Transform (iDFT) enables the reconstruction of the original time series from its frequency representation:

$$X = \{x_t\}_{t=0}^{n-1} = \text{iDFT}(\hat{X}),$$
$$x_t = \frac{1}{n}\sum_{k=0}^{n-1} \mathcal{X}_k \exp 2\pi j \frac{kt}{n}. \tag{2}$$

The bidirectional mapping formed by DFT and iDFT provides a framework for manipulating specific frequency bands, enabling precise control over different temporal scales, from high-frequency components (representing short-term fluctuations) to low-frequency components (capturing long-term trends).

The Fast Fourier Transform (FFT) [Brigham and Morrow, 1967] is an efficient algorithm for computing the DFT, reducing computational complexity from $\mathbf{O}(n^2)$ to $\mathbf{O}(n \log n)$, which is crucial for large-scale or real-time signal processing. The Real Fast Fourier Transform (rFFT) [Brigham and Morrow, 1967] is a specialized variant of the FFT designed to exploit this redundancy by calculating only the first $n/2 + 1$ frequency components, corresponding to the non-redundant positive frequencies. The inverse operation, known as the inverse rFFT (irFFT), reconstructs the original real-valued time series from its reduced set of frequency components.

$$\hat{X} = \{\mathcal{X}_k\}_{k=0}^{n/2} = \text{rFFT}(X),$$
$$X = \{x_t\}_{t=0}^{n-1} = \text{irFFT}(\hat{X}). \tag{3}$$

3.3 Frequency-domain MLP

Frequency-domain MLPs exploit these advantages through architectures redesigned for complex-valued inputs. Unlike time-domain MLPs that suffer from point-wise mapping constraints and information bottlenecks [Yi et al., 2023], frequency-domain variants operate on compact spectral features. Yi et al. [2023] and Xu et al. [2024] use FreqMLP, a variant of MLPs with complex parameters, to learn in the frequency domain. Specifically, the FreqMLP can be formulated as follows: $(\hat{Z}\hat{W} + \hat{b})$,

$$Y = \text{FreqMLP}(\hat{Z}) = \sigma(\hat{Z}\hat{W} + \hat{b}), \tag{4}$$

where $\hat{Z} \in \mathbb{C}^{L\times d}$ is the frequency spectrum, $\hat{W} = (W_r + jW_i) \in \mathbb{C}^{d\times d'}$ is the complex number weight matrix with $W_r \in \mathbb{R}^{d\times d'}$ and $W_i \in \mathbb{R}^{d\times d'}$, and $\hat{b} = (b_r + jb_i) \in \mathbb{C}^{d'}$ are the complex number biases with $b_r \in \mathbb{R}^{d'}$ and $b_i \in \mathbb{R}^{d'}$. $Y \in \mathbb{C}^{L\times d'}$ is the output of frequency-domain MLP.

4 METHODS

4.1 SpecTF Pipeline

Our proposed SpecTF framework, as illustrated in Figure 2, consists of five key components designed to effectively model the influence of textual information on time series across multiple frequency bands. We employ the channel-independent strategy [Nie et al., 2023] [Zeng et al., 2023] for our approach, where each variate in the input time series is fed independently into SpecTF.

Time Series Embedding. Given a historical univariate time series $X^h \in \mathbb{R}^{1\times L}$ of length L and associated textual information S^h, we first transform X^h into the frequency domain using rFFT, resulting in complex-valued spectral components. Then, we use a frequency-domain MLP (FreqMLP) as mentioned in Section 3.3 as an embedding layer to map the spectrum to a higher-dimensional space. Additionally, we incorporate the positional encoding [Vaswani et al., 2017] to assign each frequency band (from low to high) a unique high-dimensional signature with particular patterns across embedding dimensions:

$$\hat{X}^h = \text{rFFT}(X^h),$$
$$\hat{X}^h_{emb} = \text{FreqMLP}(\hat{X}^h) + \text{PoE}(\hat{X}^h) \tag{5}$$

where $\hat{X}^h_{emb} \in \mathbb{C}^{(L/2+1)\times d}$, d is the dimension of frequency vectors, $\text{PoE}(\cdot)$ is the positional encoding of a sequence.

Text Embedding. Concurrently, the textual data S^h is processed through our Text Embedding block, which projects the input text into a complex embedding space $\hat{S}^h = \text{TextEmbedding}(S^h)$ with $\hat{S}^h \in \mathbb{C}^{L\times d}$. Here, we employ a pre-trained language model to extract contextual embeddings, followed by an MLP layer that maps these embeddings into the complex embedding space (see details in Section 4.2).

FreqCMF. Our framework's key innovation lies in the fusion approach implemented via Frequency Cross-Modality Fusion (FreqCMF) block. It incorporates textual context into historical frequency components $\hat{X}^h_{fuse} = \text{FreqCMF}(\hat{X}^h_{emb}, \hat{S}^h)$, where $\hat{X}^h_{fuse} \in \mathbb{C}^{(L/2+1)\times d}$, The inner operations of FreqCMF will be given in Section 4.3. This fusion block enables our model to leverage textual information for understanding historical patterns.

Forecaster. Then, we employ a Forecaster model that maps these historical frequency components to future ones $\hat{X}^p_{emb} \in \mathbb{C}^{(H/2+1)\times d}$. Here, the Forecaster is implemented as a FreqMLP layer:

$$\hat{X}^p_{emb} = \text{FreqMLP}\big((\hat{X}^h_{fuse})^\top\big)^\top.$$

Projection. After that, we use a Projection layer to map the embedding of prediction frequency to the original dimension of the spectrum before converting

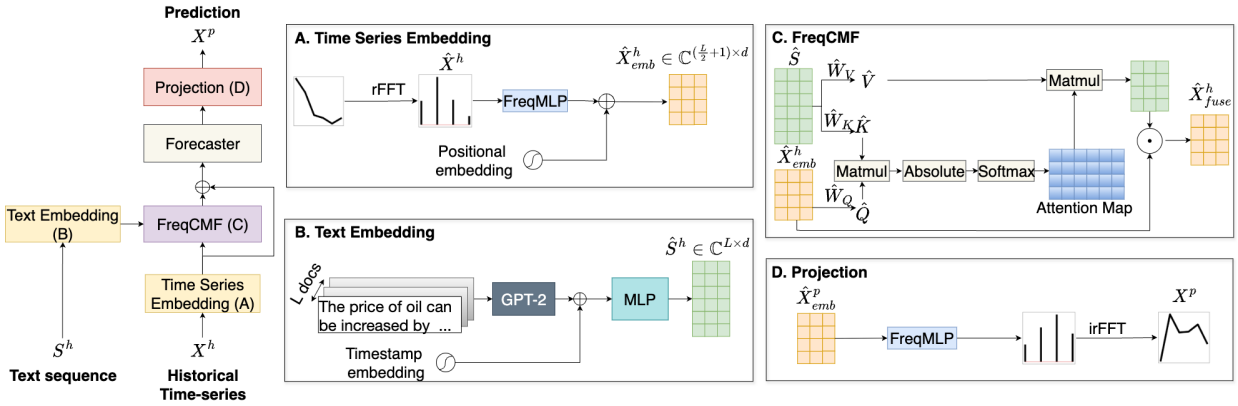


Figure 2: **An overview of the proposed SpecTF framework.** The time series input is passed through **Time Series Embedding** to be encoded in the frequency domain, and concurrently, **Text Embedding** represents each document in the frequency domain. **Frequency Cross-Modality Fusion (FreqCMF)** integrates textual information into time series frequencies through attention mechanisms and multiplication fusion based on the complex multiplication operation $\odot$. **Forecaster** that maps historical frequency representations to prediction representations in the frequency domain. **Projection** projects the representations to the original dimension of the spectrum and converts them back to the temporal domain.

it back to the temporal domain X^p through the inverse Real Fast Fourier Transform (irFFT):

$$\begin{aligned} \hat{X}^p &= \text{FreqMLP}(\hat{X}^p_{emb}), \\ X^p &= \text{irFFT}(\hat{X}^p) \end{aligned} \tag{6}$$

where $X^p \in \mathbb{R}^{1\times H}$ is the prediction of SpecTF based on the input X^h. By operating primarily in the frequency domain and leveraging complex-valued operations, SpecTF achieves remarkable parameter efficiency while explicitly modeling the multi-scale influence of textual information on time series dynamics.

4.2 Text Embedding

In this block, textual data are presented in a complex vector space. This process involves three stages: (1) Language Model Encoding, (2) Temporal Alignment and (3) Complex Projection.

Language Model Encoding. We extract contextual embeddings from text using a pre-trained language model such as BERT [Devlin et al., 2019] or GPT-2 [Radford et al., 2019]: $e_t = \text{LM}(s_t)$ where $e_t \in \mathbb{R}^{L\times d_{LM}}$ is the contextual embedding of the text instance corresponding to the time series at step t, d_{LM} is the dimension of the embedding.

Temporal Alignment. To align with time series data, we incorporate temporal information by adding timestamp embeddings [Zhou et al., 2021]: $e_t^{aligned} = e_t + e_t^{temporal}$, where $e_t^{temporal}$ encodes the timestamp at step t.

Complex Projection. We map the aligned embeddings to real and imaginary components via separate MLPs:

$$\begin{aligned} Re(\hat{s}_t) &= \text{MLP}_{Re}(e_t^{aligned}), \\ Im(\hat{s}_t) &= \text{MLP}_{Im}(e_t^{aligned}) \end{aligned} \tag{7}$$

where $\hat{s}_t = Re(\hat{s}_t) + jIm(\hat{s}_t) \in \mathbb{C}^{L\times d}$, $Re(\cdot)$ and $Im(\cdot)$ represent the real and imaginary parts of a complex vector. As a result, $\hat{S}^h = [\hat{s}_1, \ldots, \hat{s}_L]$ is the output of the Text Embedding module, which is resided in the same complex frequency space as the time series, enabling direct modulation of the spectral component.

4.3 Frequency Cross-Modality Fusion (FreqCMF)

This novel component, the cornerstone of our approach, enables the direct integration of textual information with time series spectral components in the frequency domain. Unlike existing methods that operate in the time domain and struggle to capture text's multi-scale influence, our fusion block explicitly models these interactions through operations in the frequency domain. The fusion process involves Frequency Cross Attention (FCA) and Multiplication Fusion (MF) components.

Frequency Cross Attention. Given frequency representations of time series $\hat{X}^h_{emb}$ and the textual complex representation $\hat{S}^h$, we propose to compute the query, key, and value using FreqMLPs:

$$\begin{aligned} \hat{Q} &= \text{FreqMLP}_q(\hat{X}^h_{emb}), \\ \hat{K} &= \text{FreqMLP}_k(\hat{S}^h), \\ \hat{V} &= \text{FreqMLP}_v(\hat{S}^h) \end{aligned} \tag{8}$$

Table 1: Forecasting results on Time-MMD benchmark [Liu et al., 2024a]. The results are averaged across three different seeds and all prediction lengths. The best numbers in each row are shown in **bold**, and the second-best numbers are underlined. The last column denotes the percentage reduction in MSE and MAE achieved by SpecTF compared to the best-performing baseline. Full results are in Appendix C.1

Models	SpecTF (ours)		TaTS		MM-TSF		TimeXer		FreTS		TimeLLM		ChatTime		Promotion	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Agriculture	**0.103**	**0.218**	0.112	0.237	0.123	0.263	0.132	0.263	0.120	0.249	0.157	0.301	0.305	0.345	8.35%	8.02%
Climate	**0.938**	**0.762**	1.002	0.790	1.086	0.832	1.053	0.805	1.289	0.888	1.175	0.868	1.528	1.002	6.38%	3.54%
Economy	0.0085	0.0780	**0.0083**	**0.0770**	0.0121	0.0883	0.0096	0.0817	0.0152	0.0986	0.0319	0.1438	0.0547	0.1786	-2.41%	-1.30%
Energy	**0.246**	**0.359**	0.264	0.374	0.270	0.393	0.288	0.400	0.299	0.401	0.264	0.369	0.270	0.393	6.82%	4.01%
Environment	**0.259**	**0.367**	0.266	0.370	0.391	0.494	0.272	0.371	0.372	0.482	0.342	0.441	0.391	0.494	2.63%	0.81%
Health	**1.276**	**0.733**	1.340	0.763	1.508	0.897	1.394	0.973	1.542	0.814	1.699	0.871	1.508	0.897	4.78%	3.93%
Security	**108.411**	**4.745**	110.600	5.069	116.003	5.572	109.472	4.870	128.873	6.260	108.812	4.836	136.273	5.985	0.97%	2.56%
Social Good	**0.962**	**0.443**	1.012	0.449	1.183	0.497	1.002	0.478	1.086	0.520	1.084	0.573	1.183	0.497	4.94%	1.34%
Traffic	**0.171**	**0.210**	0.192	0.218	0.197	0.233	0.176	0.210	0.205	0.241	0.171	0.388	0.197	0.233	2.85%	0%

Table 2: Forecasting results on TTC benchmark [Kim et al., 2024]. The results are averaged across three different seeds and all prediction lengths. The best numbers in each row are shown in **bold**, and the second-best numbers are underlined. The last column denotes the percentage reduction in MSE and MAE achieved by SpecTF compared to the best-performing baseline. Full results are in Appendix C.2

Models	SpecTF (ours)		TaTS		MM-TSF		TimeXer		FreTS		TimeLLM		ChatTime		Promotion	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Climate	**0.693**	**0.605**	0.790	0.636	0.802	0.662	0.760	0.631	0.878	0.682	0.699	0.611	1.327	0.836	8.81%	4.12%
Medical	**1.298**	**0.828**	1.313	0.953	1.436	1.072	1.311	0.880	1.511	1.073	1.572	1.104	1.610	1.117	1.00%	5.90%

where $\hat{Q} \in \mathbb{C}^{(L/2+1)\times d_k}$, $\hat{K} \in \mathbb{C}^{L\times d_k}$ and $\hat{V} \in \mathbb{C}^{L\times d}$. Then we compute the attention output:

$$\hat{O} = \text{softmax}\left(\frac{|\hat{Q}\hat{K}^\top|}{\sqrt{d_k}}\right) \cdot \hat{V} \tag{9}$$

where d_k is the dimension of query and key, $|\cdot|$ extracts the absolute value of the complex products, ensuring that attention weights reflect the strength of alignment between time series queries and textual keys. This attention mechanism identifies relevant textual information for each frequency component.

Multiplication Fusion. Following attention computation, we apply a complex multiplication operation to incorporate textual influence:

$$\begin{aligned}\hat{X}_{fuse} &= \hat{X}_{emb} \odot \hat{O} \\ &= \left(Re(\hat{X}_{emb})Re(\hat{O}) - Im(\hat{X}_{emb})Im(\hat{O})\right) + \\ &\quad j\left(Re(\hat{X}_{emb})Im(\hat{O}) + Im(\hat{X}_{emb})Re(\hat{O})\right).\end{aligned} \tag{10}$$

This operation in the frequency domain simultaneously modifies both amplitude and phase information, allowing textual context to selectively emphasize or attenuate specific frequency bands while also adjusting their temporal alignment. The multiplication in the frequency domain corresponds to convolution in the time domain [Yi et al., 2023], enabling a more nuanced integration of textual influence across multiple temporal scales. We provide the theoretical support in Appendix A.

5 EXPERIMENTS

Datasets. We evaluate SpecTF on two comprehensive multimodal time series benchmarks: the Time-MMD dataset [Liu et al., 2024a] and the TimeText Corpus (TTC) [Kim et al., 2024]. Further dataset details are provided in Appendix B.

Baselines. We compare our SpecTF with the representative and state-of-the-art models for both unimodal and multimodal time series forecasting. For *unimodal* forecasting, we compare SpecTF against FreTS [Yi et al., 2023], a frequency-based method. For *multimodal* time series forecasting, we include MM-TSF [Liu et al., 2024a] and TaTS [Li et al., 2025] with their best variant using iTransformer [Liu et al., 2024b] backbone for comparison. In addition, we also include TimeXer [Wang et al., 2024b] with text sequence as an exogenous variable by using our Text Embedding block rather than the exogenous embedding layer. For *LLM-based models*, we compare SpecTF against TimeLLM [Jin et al., 2024] and ChatTime [Wang et al., 2025].

Implementation Details. We conduct all experiments on a single NVIDIA Tesla V100. We take MSE (Mean Squared Error) as the loss function and

Table 3: Ablation study on Agriculture, Health, Environment datasets. **w/o multiplication fusion** removes multiplication fusion. **w/o frequency cross attention** replaces cross attention by summation operation. **w/o textual data** excludes textual data. **w/o time series data** removes time series data. **w/ real-imaginary MLP** replace the FreqMLPs by applying MLPs independently to the real and imaginary parts. The best numbers in each column are shown in **bold**, and the second-best numbers are underlined.

Models	Agriculture		Health		Environment	
	MSE	MAE	MSE	MAE	MSE	MAE
SpecTF (ours)	**0.103**	**0.218**	**1.276**	**0.733**	**0.259**	**0.367**
w/o multiplication fusion	0.104	0.222	1.289	0.752	0.263	0.369
w/o frequency cross attention	0.115	0.234	1.291	0.751	0.272	0.370
w/ real–imaginary MLP	0.113	0.230	1.297	0.750	0.330	0.423
w/o textual data	0.132	0.262	1.342	0.776	0.355	0.368
w/o time series data	0.160	0.311	1.668	0.895	0.379	0.491

MAE (Mean Absolute Error) as the evaluation metric. GPT2 [Radford et al., 2019] is used to embed the text. For additional implementation details, please refer to Appendix B.

5.1 Benchmarking Results

Table 1 presents the comparative results of SpecTF against all baselines across the nine domains from Time-MMD benchmark. For each dataset, we report the results averaged across three different seeds and all prediction lengths. Overall, SpecTF outperforms both unimodal and multimodal baselines in terms of MSE and MAE metrics on 8 out of 9 domains. When compared to the unimodal baseline FreTS and LLM-based approaches (TimeLLM, ChatTime), SpecTF demonstrates consistent improvements across all domains. This indicates that incorporating textual information in the frequency domain provides valuable additional context for forecasting. In comparison to the runner-up among multimodal approaches (TimeXer, MM-TSF, and TaTS), SpecTF still achieves superior performance in most domains, with the average improvement across 9 datasets of 3.82% and 2.25% in MSE and MAE, respectively. The most significant improvements over multimodal baselines are observed in the Energy, Environment, Health, and Security domains, where textual information often contains signals about regime shifts or policy changes that manifest at different temporal scales. However, SpecTF underperforms on the Economy dataset, where TaTS achieves marginally better results because the textual data contains more noise and shorter-term economic indicators that may be better suited to the time-domain alignment mechanism of TaTS.

Table 2 shows SpecTF's performance on TTC benchmark, averaged over prediction lengths. Due to computational constraints, we limit our comparisons to the best-performing baselines from earlier experiments. SpecTF achieves superior results on both datasets: 8.81% improvements in MSE on Climate, and 5.90% improvements in MAE on Medical, demonstrating robustness across diverse multimodal datasets.

5.2 Model Analysis

Ablation study. We perform an ablation study spanning three domains (Table 3), comparing SpecTF with four alternative model variants: (1) w/o multiplication fusion, which removes the multiplication fusion; (2) w/o frequency cross attention, substituting cross-attention with direct embedding addition; (3) w/ real–imaginary MLP, replacing FreqMLP with separate MLPs applied independently to real and imaginary components; (4) w/o textual data, excluding all textual inputs to assess modality dependence; and (5) w/o time series data, removing numerical sequences to test text-only forecasting viability. The results demonstrate the multiplication fusion mechanism's critical role, with its removal causing performance degradation, specifically by about 2.6% in MAE in the Health dataset. The cross-attention module proves vital for text-frequency alignment when its substitution with summation reduces accuracy, particularly in Agriculture, with a drop in MSE of 11.7%. The real–imaginary variant also shows consistent degradation, increasing MAE by 2.3% in the Environment dataset, confirming the benefit of modeling spectral interactions jointly of FreqMLPs.

When textual data is excluded, performance deteriorates significantly across all domains, with particularly pronounced effects in the Environment dataset. Figure 1 provides a clear case study of this phenomenon, where predictions without text deviate substantially

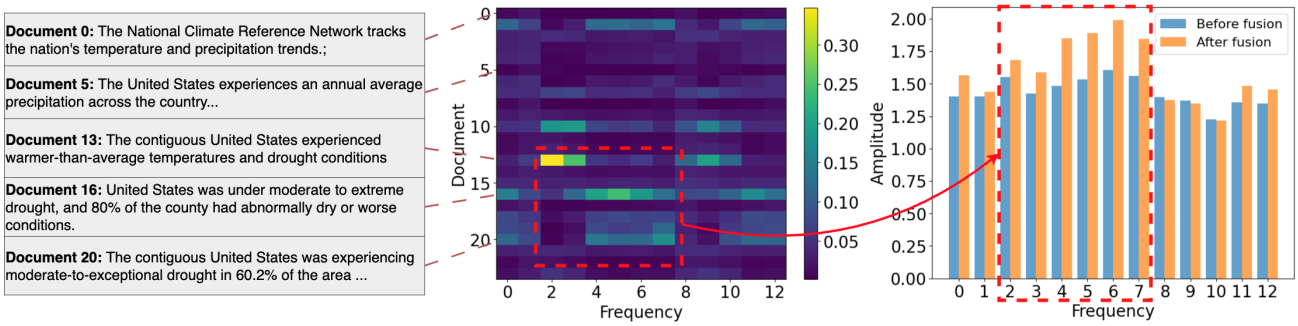


Figure 3: **Visualization of frequency-text interactions in the Climate dataset.** The attention map (center) reveals document-frequency relationships, with brightness indicating attention strength. Documents 13, 16, and 20 (left panel) show strong attention weights in middle-frequency bands (2-7), highlighted by the red rectangle. The frequency amplitude comparison (right) confirms this relationship: amplitudes after fusion (orange) show significant increases in these middle frequencies compared to the values before fusion (blue), as indicated by the corresponding red rectangle.

from ground truth in the time domain (Figure 1). In contrast, predictions with text track the ground truth more closely. The frequency-domain comparison in Figure 1b further illustrates this difference, showing how text-enhanced predictions better match the amplitude distribution across frequency bands. Conversely, removing time series data results in the most dramatic performance drops, most significantly in the Health domain with 30.7% reduction in MSE.

Visualization of Cross-modality Fusion. Figure 3 illustrates how FreqCMF modulates time-series frequency components in the Climate dataset based on textual context. The right figure displays frequency amplitude changes in one dimension before and after the FreqCMF block. We observe that the middle-frequency bands (2-7) show more substantial amplitude increases compared to higher frequencies, suggesting that textual information in this domain primarily influences medium-term dynamics. The centre attention heatmap reveals which documents influence specific frequency bands. It shows that documents 13, 16, and 20, containing specific information about "warm-than-average temperatures", "extreme drought conditions", "abnormally dry conditions", and "moderate-to-exceptional drought", make a strong impact on the middle-frequency band. In contrast, documents 0 and 5, which do not contain valuable information, contribute less to the frequency components. This demonstrates SpecTF's ability to extract the relevant texts with different frequency bands.

Model Efficiency. Table 4 shows that our SpecTF uses significantly fewer trainable parameters than TimeXer and far fewer than MM-TSF and TaTS. This leads to much lower computational costs, as shown by Multiply-Accumulate Operations (MACs) metric that counts the total number of multiplication and addition operations in a neural network. Besides, SpecTF achieves superior inference speed, outperforming all baselines. These results demonstrate that SpecTF improves efficiency without compromising performance, making it suitable for both resource-limited and large-scale applications.

Table 4: Number of trainable parameters, MACs and inference time of models under forecast horizon of 48 in Energy dataset. The best results are shown in **bold**.

Models	Parameters	MACs	Inference time
SpecTF (ours)	**397.1K**	**1.8M**	**9.1ms**
TimeXer	1.72M	11.1M	88.4ms
MM-TSF	6.41M	8.23M	14.1ms
TaTS	6.41M	84.23M	41.3ms

6 CONCLUSION

In conclusion, SpecTF demonstrates that modeling textual influences through frequency-domain decomposition enables more effective modeling of multiscale temporal patterns than traditional time-domain approaches. By fusing text embeddings with spectral components via cross-attention, our framework captures how language data modulates both short-term fluctuations and long-term trends, achieving superior performance across diverse datasets with considerably fewer parameters than leading baselines. SpecTF delivers the greatest benefit when the accompanying text exhibits multiscale temporal influence and offers targeted fusion by reweighting and phase-aligning only those frequency bands modulated by the text.

ACKNOWLEDGMENT

This research was funded (partially or fully) by the Australian Government through the Australian Research Council. Dr Hung Le is the recipient of an Australian Research Council Discovery Early Career Researcher Award (project number DE250100355) funded by the Australian Government.

References

E. O. Brigham and R. E. Morrow. The fast fourier transform. *IEEE Spectrum*, 4(12):63–70, 1967. doi: 10.1109/MSPEC.1967.5217220.

Chirag Deb, Fan Zhang, Junjing Yang, Siew Eang Lee, and Kwok Wei Shah. A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews*, 74:902–924, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Philip Hans Franses. *Time series models for business and economic forecasting*. Cambridge university press, 1998.

Yushan Jiang, Kanghui Ning, Zijie Pan, Xuyang Shen, Jingchao Ni, Wenchao Yu, Anderson Schneider, Haifeng Chen, Yuriy Nevmyvaka, and Dongjin Song. Multi-modal time series analysis: A tutorial and survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 6043–6053, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714542. doi: 10.1145/3711896.3736567. URL https://doi.org/10.1145/3711896.3736567.

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*, 2024. URL https://openreview.net/forum?id=Unb5CVPtae.

Shruti Kaushik, Abhinav Choudhury, Pankaj Kumar Sheron, Nataraj Dasgupta, Sayee Natarajan, Larry A Pickett, and Varun Dutt. Ai in healthcare: time-series forecasting using statistical, neural, and ensemble architectures. *Frontiers in big data*, 3:4, 2020.

Kai Kim, Howard Tsai, Rajat Sen, Abhimanyu Das, Zihao Zhou, Abhishek Tanpure, Mathew Luo, and Rose Yu. Multi-modal forecaster: Jointly predicting time series and textual data, 2024. URL https://arxiv.org/abs/2411.06735.

Benjamin F. King. Market and Industry Factors in Stock Price Behavior. *The Journal of Business*, 39: 139–139, 1965. doi: 10.1086/294847.

Irena Koprinska, Dengsong Wu, and Zheng Wang. Convolutional neural networks for energy time series forecasting. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2018. doi: 10.1109/IJCNN.2018.8489399.

Hung Le, Sherif Abbas, Minh Hoang Nguyen, Van Dai Do, Huu Hiep Nguyen, and Dung Nguyen. Accelerating long-term molecular dynamics with physics-informed time-series forecasting. In *IEEE International Conference on Data Mining*, 2025.

Zihao Li, Xiao Lin, Zhining Liu, Jiaru Zou, Ziwei Wu, Lecheng Zheng, Dongqi Fu, Yada Zhu, Hendrik F. Hamann, Hanghang Tong, and Jingrui He. Language in the flow of time: Time-series-paired texts weaved into a unified temporal narrative. *CoRR*, abs/2502.08942, 2025. doi: 10.48550/ARXIV.2502.08942. URL https://doi.org/10.48550/arXiv.2502.08942.

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*, 2024a. URL https://openreview.net/forum?id=fuD0h4R1IL.

Haoxin Liu, Harshavardhan Kamarthi, Zhiyuan Zhao, Shangqing Xu, Shiyu Wang, Qingsong Wen, Tom Hartvigsen, Fei Wang, and B Aditya Prakash. How can time series analysis benefit from multiple modalities? a survey and outlook. *arXiv preprint arXiv:2503.11835*, 2025.

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 *The Twelfth International Conference on Learning Representations*,

2024b. URL https://openreview.net/forum?id=JePfAI8fah.

Ming-Chih Lo, Ching Chang, and Wen-Chih Peng. Text2freq: Learning series patterns from text via frequency domain. In *NeurIPS Workshop on Time Series in the Age of Large Models*, 2024. URL https://openreview.net/forum?id=Pi6sA1MSSr.

Prapanna Mondal, Labani Shit, and Saptarsi Goswami. Study of effectiveness of time series modeling (arima) in forecasting stock prices. *International Journal of Computer Science, Engineering and Applications*, 4(2):13, 2014.

Irene Nandutu, Marcellin Atemkeng, Nokubonga Mgqatsa, Sakayo Toadoum Sari, Patrice Okouma, Rockefeller Rockefeller, Theophilus Ansah-Narh, Jean Louis Ebongue Kedieng Fendji, and Franklin Tchakounte. Error correction based deep neural networks for modeling and predicting south african wildlife–vehicle collision data. *Mathematics*, 10(21), 2022. ISSN 2227-7390. doi: 10.3390/math10213988.

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*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Omer Berat Sezer, M. Ugur Gudelek, and Ahmet Murat Özbayoglu. Financial time series forecasting with deep learning : A systematic literature review: 2005-2019. *CoRR*, abs/1911.13288, 2019. URL http://arxiv.org/abs/1911.13288.

D. Sundararajan. *The Discrete Fourier Transform: Theory, Algorithms and Applications*. G - Reference,Information and Interdisciplinary Subjects Series. World Scientific, 2001. ISBN 9789810245214.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

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 *AAAI Conference on Artificial Intelligence*, 2025.

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

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=INAeUQO4lT.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: decomposition transformers with auto-correlation for long-term series forecasting. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Zhijian Xu, Ailing Zeng, and Qiang Xu. FITS: Modeling time series with $10k$ parameters. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=bWcnvZ3qMb.

Kun Yi, Qi Zhang, Wei Fan, Shoujin Wang, Pengyang Wang, Hui He, Defu Lian, Ning An, Longbing Cao, and Zhendong Niu. Frequency-domain mlps are more effective learners in time series forecasting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 76656–76679, 2023.

Kun Yi, Qi Zhang, Wei Fan, Longbing Cao, Shoujin Wang, Hui He, Guodong Long, Liang Hu, Qingsong Wen, and Hui Xiong. A survey on deep learning based time series analysis with frequency transformation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 6206–6215, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714542. doi: 10.1145/3711896.3736571. URL https://doi.org/10.1145/3711896.3736571.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):11121–11128, Jun. 2023. doi: 10.1609/aaai.v37i9.26317.

Xi Nicole Zhang, Yuan Pu, Yuki Kawamura, Andrew Loza, Yoshua Bengio, Dennis Shung, and Alexander Tong. Trajectory flow matching with applications to clinical time series modelling. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=fNakQltI1N.

Xiyuan Zhang, Ranak Roy Chowdhury, Rajesh K. Gupta, and Jingbo Shang. Large language models for time series: A survey. In Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8335–8343. International Joint Conferences on Artificial Intelligence Organization, 8 2024b. doi: 10.24963/ijcai.2024/921. URL `https://doi.org/10.24963/ijcai.2024/921`. Survey Track.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12): 11106–11115, May 2021. doi: 10.1609/aaai.v35i12.17325.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proc. 39th International Conference on Machine Learning (ICML 2022)*, 2022.

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes, please refer to Appendix A]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes, the analysis of complexity is described in the Section 5]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes, we provide source code and guidelines in the supplementary.]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes, please refer to Appendix A]
   (b) Complete proofs of all theoretical results. [Yes, please refer to Appendix A]
   (c) Clear explanations of any assumptions. [Yes, please refer to Appendix A]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes, we provide source code, data and guidelines in the supplementary.]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes, please refer to Appendix B]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes, please refer to Appendix C]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes, we provide details the infrastructure used in Appendix B]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable, we do not introduce any new assets]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [No]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials for
Spectral Text Fusion: A Frequency-Aware Approach to Multimodal Time-Series Forecasting

A THEORETICAL ANALYSIS

This section provides formal theoretical justification for why frequency-domain fusion offers principled advantages over time-domain approaches in multimodal time series forecasting.

A.1 Complex Multiplication Properties

When two complex numbers are multiplied, the operation simultaneously affects both amplitude and phase characteristics. Given two complex numbers $X_1 = A_1 e^{j\phi_1}$ and $X_2 = A_2 e^{j\phi_2}$, their multiplication yields:

$$X = X_1 \cdot X_2 = |A_1 A_2| e^{j(\phi_1 + \phi_2)} \tag{11}$$

where $|A_1 A_2|$ represents the combined amplitude modulation and $(\phi_1 + \phi_2)$ represents the combined phase shift.

In our SpecTF framework, this property enables **dual-modal spectral modulation** because textual information can simultaneously influence both the energy distribution across frequencies (amplitude characteristics) and the temporal alignment of frequency components (phase characteristics) of time series data. This is mathematically impossible to achieve with equivalent efficiency in the time domain, where amplitude and phase modifications require separate operations.

A.2 Convolution Theorem and Global Influence

The theoretical superiority of frequency-domain fusion is further established by the **convolution theorem**, which states that convolution in the time domain equals multiplication in the frequency domain:

$$\mathcal{F}\{x(t) * h(t)\} = X(\omega) \cdot H(\omega) \tag{12}$$

where $\mathcal{F}$ denotes the Fourier transform, $*$ represents convolution, and $\cdot$ represents point-wise multiplication.

This theorem reveals why our frequency-domain text embedding can modulate every timestep of the time series in one unified operation, rather than relying on local, step-wise interactions typical in time-domain approaches. The mathematical equivalence ensures that our multiplication fusion $\hat{X}_{fuse} = \hat{X}_{emb} \odot \hat{O}$ corresponds to a global convolution operation across all temporal positions simultaneously.

A.3 Energy Conservation Guarantees

Parseval's Energy Conservation Theorem provides mathematical guarantees for information preservation in our approach:

$$\int_{-\infty}^{\infty} |x(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} |X(f)|^2 df \tag{13}$$

This theorem ensures that our frequency-domain operations preserve signal energy, providing a mathematical guarantee that no information is lost during domain transformation. The energy conservation property is crucial for maintaining the integrity of both temporal patterns and textual influences throughout the fusion process.

A.4 Multi-Scale Influence Modeling

Time-domain fusion approaches face fundamental limitations in modeling multi-scale temporal influences. Consider a textual signal $s(t)$ that needs to influence a time series $x(t)$ at multiple temporal scales. In the time domain, this requires:

$$y(t) = \sum_k \alpha_k \cdot s(t) * h_k(t) * x(t) \tag{14}$$

where $h_k(t)$ represents different temporal scale filters and α_k are scale-specific weights. This formulation requires multiple convolution operations and explicit scale decomposition.

In contrast, our frequency-domain approach naturally decomposes signals into scale-specific components through the Fourier transform, where each frequency bin k corresponds to a specific temporal scale $\frac{2\pi}{k}$. The fusion operation:

$$\hat{Y}(k) = \hat{X}(k) \cdot \hat{S}(k) \tag{15}$$

simultaneously addresses all temporal scales with a single multiplication operation per frequency component.

B IMPLEMENTATION DETAILS

B.1 Dataset

Our experimental evaluation utilizes two comprehensive multimodal time series benchmarks with distinct characteristics and evaluation protocols.

Time-MMD Benchmark. We employ the Time-MMD dataset comprising 9 multimodal datasets across daily, weekly, and monthly frequencies, with standardized temporal parameters adapted from established forecasting protocols. Following the setup in [Li et al., 2025], we adopt a fixed 24-step lookback window across all domains to enable consistent cross-domain comparisons. Prediction horizons follow frequency-specific configurations: daily datasets use horizons 48, 96, 192, 336 steps, weekly datasets employ 12, 24, 36, 48 steps, and monthly datasets utilize 6, 8, 10, 12 steps. Table 6 describes the statistics of each and the corresponding lookback window and forecasting horizon.

TimeText Corpus (TTC) Benchmark. For the TTC evaluation, we conduct experiments on the Medical and Climate datasets following the standardized protocol established in [Kim et al., 2024]. We maintain consistency by adopting a 24-step lookback window for both datasets, with prediction horizons configured as 12, 24, 36, 48 steps across all scenarios.

B.2 Code and Reproducibility

The experimental implementation leverages NVIDIA V100 GPUs for training, with all models undergoing three independent runs using random seeds to ensure statistical reliability through averaged metrics. Hyperparameter configurations follow the specifications detailed in Table 5. The codebase, implemented in PyTorch 1.13 with CUDA 11.7 dependencies, along with comprehensive installation instructions and environment setup guidelines, is provided in the supplementary materials to ensure reproducibility.

B.3 Metrics

Our implementation utilizes Mean Squared Error (MSE) as the primary optimization objective and Mean Absolute Error (MAE) as the key evaluation metric, aligning with established practices for regression tasks in machine learning literature

The MSE loss function calculates the average squared difference between predicted values ($\hat{y}_i$) and ground truth

Table 5: Default parameters of SpecTF.

Hyperparameter	Description	Choices
batch_size	The batch size for training	32
seq_len	Lookback window length	24
prior_weight	Weight for combination	{0.1 0.2 0.4 0.5}
train_epochs	Number of training epochs	50
patience	Early stopping patience	20
pool_type	Pooling type for text embedding	avg
dropout	Dropout	0.1

Table 6: Overview of numerical data in Time-MMD and the lookback window, forecasting horizon in our implementation.

Dataset	Frequency	Number of samples	Lookback window	Forecasting horizon
Agriculture	Monthly	496	24	{6, 8, 10, 12}
Climate	Monthly	496	24	{6, 8, 10, 12}
Economy	Monthly	423	24	{6, 8, 10, 12}
Energy	Weekly	1479	24	{12, 24, 36, 48}
Environment	Daily	11102	24	{48, 96, 192, 336}
Health	Weekly	1389	24	{12, 24, 36, 48}
Security	Monthly	297	24	{6, 8, 10, 12}
Social Good	Monthly	900	24	{6, 8, 10, 12}
Traffic	Monthly	531	24	{6, 8, 10, 12}

labels (y_i) across N samples:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{16}$$

This quadratic penalty strongly penalizes large prediction errors, making it particularly effective for gradient-based optimization landscapes. For final model evaluation, MAE provides an interpretable measure of average absolute deviation:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|^2 \tag{17}$$

This MAE metric's linear scaling makes it robust to outliers compared to MSE, providing complementary insights into model performance.

Specifically, in our problem, MSE and MAE are computed as follows:

$$\begin{aligned} \text{MSE} &= \frac{1}{H}\sum_{i=1}^{H}(y_i - x_i^p)^2 \\ \text{MAE} &= \frac{1}{H}\sum_{i=1}^{H}|y_i - x_i^p|^2 \end{aligned} \tag{18}$$

where $Y = [y_1, \dots y_H]$ is the ground truth, $X^p = [x_1^p, \dots x_H^p]$ is the prediction, H is the forecast horizon.

C FULL EXPERIMENT RESULTS

C.1 Forecasting Results on Time-MMD benchmark

We provide full forecasting results in Table 7 at multiple forecast horizons as described in Appendix B. Overall, SpecTF outperforms both unimodal and multimodal baselines across 8 in 9 datasets at different prediction length.

C.2 Forecasting Results on TTC benchmark

We have conducted additional experiments (Table 8) on two diverse datasets, Medical and Climate, from the TimeText Corpus (TTC) [Kim et al., 2024]. These datasets originate from MIMIC-III clinical notes and National Weather Service forecasts, respectively, and introduce different characteristics such as higher noise levels (in medical text) and finer temporal granularity (in weather data).

We follow the same experimental setup with a lookback window of 24 and compare our method against TaTS, the strongest baseline in our main experiments. These results confirm that our frequency-domain fusion generalises well to very different data regimes, reinforcing that the gains reported on Time-MMD are not benchmark-specific.

C.3 Hyperparameter Study

Figure 4 provides a comprehensive analysis of the method's robustness to learning rate choices. Performances remain stable within a broad learning rate range (1e-5 to 5e-4). However, the Security domain shows significant sensitivity, where high rates (> 0.0001) cause MSE spikes (> 140), necessitating precise tuning. Long-horizon tasks (e.g., Health 48-step forecasts) also require lower rates to maintain stability, highlighting domain-specific optimization needs despite general adaptability. These edge cases highlight the framework's default stability while underscoring domain-specific tuning needs for volatile or long-horizon tasks.

Figure 5 reveals several key insights about how model dimensionality affects forecasting performance. The plots demonstrate that increasing model dimensions generally improves accuracy (reduces MSE) but with domain-specific patterns and clear diminishing returns. While complex domains like Energy show substantial performance gains when scaling from small to medium dimensions ($8 \rightarrow 32$), many other domains exhibit performance plateaus beyond certain thresholds, particularly for simpler or more periodic data patterns. Longer forecast horizons typically benefit more from increased dimensions than shorter ones, suggesting higher capacity requirements for extended predictions. The Security domain notably shows minimal dimension sensitivity, indicating that some forecasting challenges remain difficult regardless of model capacity. These patterns highlight the importance of balanced dimensionality selection—increasing capacity where complexity demands it, while avoiding unnecessary computation where simpler models suffice, enabling efficient resource allocation in practical deployments.

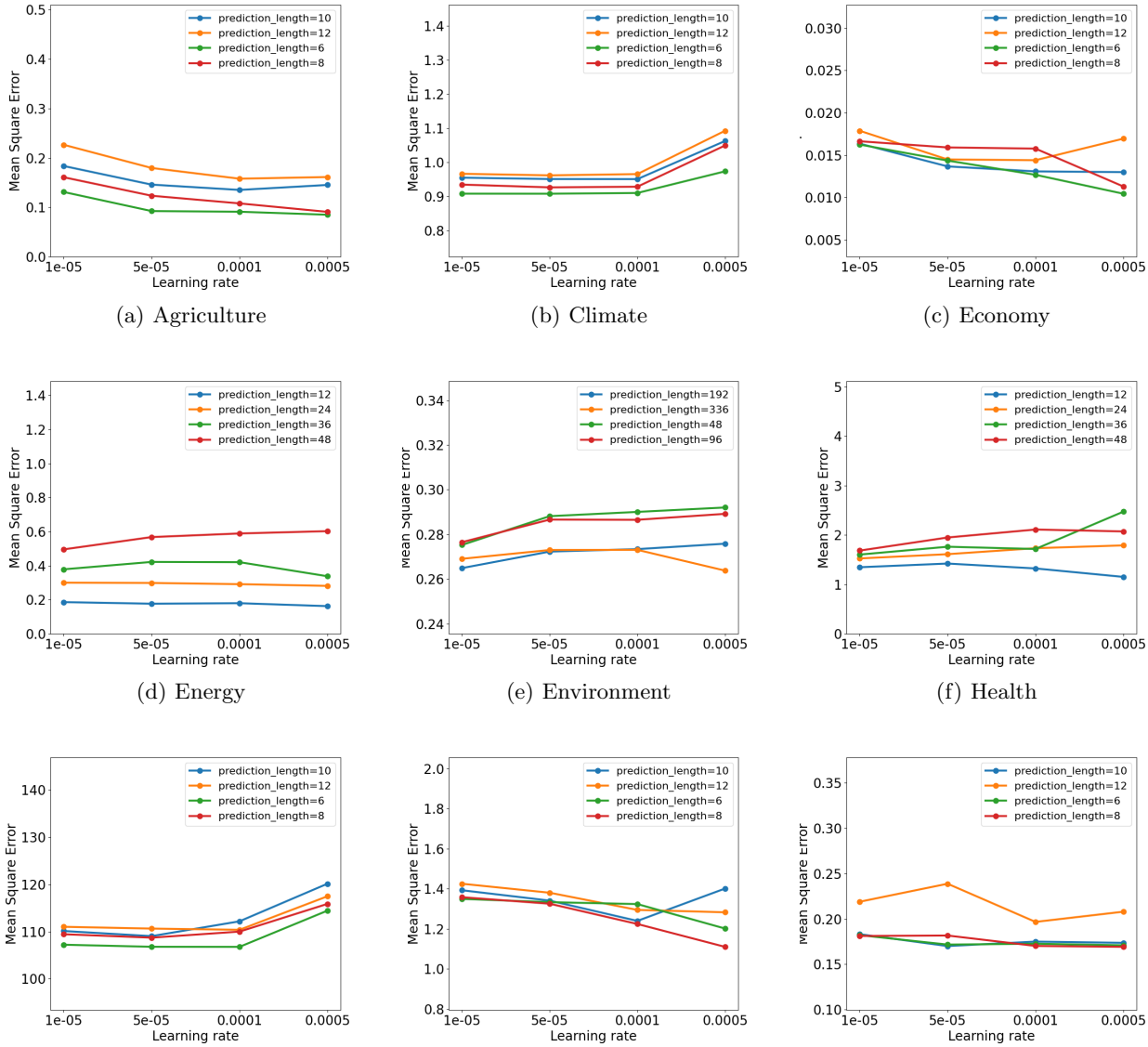
Figure 6 demonstrates that increasing model scale consistently improves forecasting accuracy, with larger LMs (e.g., GPT-XL) achieving significantly lower MSE compared to smaller counterparts (e.g., BERT) in complex domains like Security (Figure A3g) and long-horizon predictions (e.g., 48-step Energy forecasts). This aligns with scaling laws in LLM research, where model capacity enhances pattern recognition in volatile or multi-scale data. However, smaller models (e.g., BERT-large) remain competitive in stable domains like Agriculture (Figure 6a), achieving comparable accuracy with reduced computational costs. These results highlight a critical trade-off: practitioners should prioritize larger LLMs for high-stakes, irregular tasks while leveraging compact architectures for efficiency in predictable scenarios.

Table 7: Full forecasting results across nine datasets in TimeMMD benchmark [Liu et al., 2024a] at multiple forecast horizons.

Models		SpecTF (ours)		TaTS		MM-TSF		TimeXer		FreTS		TimeLLM		ChatTime	
Method		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Agriculture	6	0.064	0.170	0.075	0.196	0.084	0.216	0.088	0.211	0.082	0.201	0.092	0.218	0.193	0.283
	8	0.084	0.192	0.097	0.219	0.112	0.253	0.109	0.239	0.101	0.229	0.151	0.303	0.264	0.320
	10	0.110	0.236	0.127	0.254	0.133	0.276	0.144	0.2811	0.125	0.262	0.176	0.330	0.345	0.371
	12	0.152	0.275	0.167	0.291	0.165	0.305	0.187	0.322	0.172	0.303	0.210	0.354	0.418	0.407
	Avg	0.103	0.218	0.112	0.237	0.123	0.263	0.132	0.263	0.120	0.249	0.157	0.301	0.305	0.345
Climate	6	0.904	0.748	0.990	0.780	1.074	0.822	1.027	0.794	0.286	0.878	1.126	0.854	1.490	0.981
	8	0.927	0.756	0.998	0.786	1.067	0.823	1.059	0.804	0.283	0.883	1.165	0.866	1.530	1.006
	10	0.965	0.775	1.010	0.794	1.100	0.839	1.067	0.812	0.294	0.893	1.200	0.877	1.519	1.005
	12	0.960	0.770	1.012	0.798	1.101	0.842	1.058	0.810	0.293	0.897	1.210	0.875	1.570	1.019
	Avg	0.938	0.762	1.002	0.790	1.053	0.805	1.053	0.805	1.289	0.888	1.175	0868	1.528	1.002
Economy	6	0.0084	0.0770	0.0080	0.0759	0.0116	0.0856	0.0133	0.0935	0.0150	0.0975	0.0307	0.1414	0.478	0.169
	8	0.0086	0.0785	0.0083	0.0778	0.0117	0.0880	0.0140	0.0958	0.0149	0.0981	0.0310	0.1435	0.056	0.179
	10	0.0085	0.0779	0.0083	0.0773	0.0126	0.0906	0.0143	0.0952	0.0155	0.0992	0.0323	0.1444	0.056	0.181
	12	0.0086	0.0784	0.0084	0.0767	0.0125	0.0891	0.0148	0.0982	0.0154	0.0996	0.0336	0.1460	0.059	0.184
	Avg	0.0085	0.0780	0.0083	0.0770	0.0096	0.0817	0.0096	0.0817	0.0152	0.0986	0.0319	0.1438	0.0547	0.1786
Energy	12	0.096	0.219	0.103	0.227	0.117	0.256	0.137	0.278	0.127	0.251	0.105	0.225	0.135	0.282
	24	0.199	0.326	0.206	0.336	0.210	0.355	0.242	0.369	0.245	0.373	0.216	0.340	0.238	0.369
	36	0.282	0.396	0.313	0.420	0.307	0.431	0.324	0.429	0.337	0.433	0.307	0.408	0.328	0.421
	48	0.406	0.493	0.433	0.514	0.447	0.520	0.447	0.523	0.487	0.548	0.427	0.503	0.435	0.516
	Avg	0.246	0.359	0.264	0.374	0.270	0.393	0.288	0.400	0.299	0.401	0.264	0.369	0.284	0.397
Environment	48	0.265	0.370	0.267	0.374	0.288	0.391	0.270	0.374	0.350	0.464	0.314	0.393	0.342	0.429
	96	0.265	0.368	0.268	0.371	0.294	0.392	0.280	0.376	0.367	0.479	0.352	0.414	0.351	0.428
	192	0.255	0.365	0.268	0.367	0.292	0.398	0.271	0.367	0.385	0.492	0.354	0.410	0.345	0.423
	336	0.253	0.366	0.262	0.368	0.291	0.397	0.267	0.367	0.386	0.493	0.348	0.426	0.342	0.424
	Avg	0.259	0.367	0.266	0.370	0.391	0.494	0.272	0.371	0.372	0.482	0.342	0.411	0.345	0.426
Health	12	0.985	0.647	0.979	0.663	1.196	0.828	1.181	0.744	1.141	0.680	1.308	0.752	1.382	0.773
	24	1.256	0.718	1.297	0.741	1.440	0.865	1.402	0.786	1.546	0.785	1.724	0.868	1.758	0.888
	36	1.395	0.769	1.442	0.787	1.669	0.920	1.455	0.803	1.691	0.881	1.869	0.914	1.904	0.938
	48	1.470	0.797	1.519	0.820	1.755	0.973	1.538	0.841	1.790	0.910	1.895	0.949	1.930	0.973
	Avg	1.276	0.733	1.340	0.763	1.508	0.897	1.394	0.973	1.542	0.814	1.699	0.871	1.730	0.893
Security	6	106.364	4.507	108.460	4.929	113.759	5.418	106.859	4.754	126.132	6.265	106.535	4.694	113.701	5.835
	8	106.947	4.769	110.180	5.042	115.563	5.542	109.134	4.855	129.167	6.215	108.205	4.718	117.083	5.877
	10	109.767	4.616	111.352	5.104	116.793	5.611	110.484	4.915	129.848	6.237	109.566	4.905	123.805	6.009
	12	110.567	5.089	112.405	5.201	117.897	5.717	111.411	4.956	130.346	6.322	110.865	5.028	190.503	6.219
	Avg	108.411	4.745	110.600	5.069	116.003	5.572	109.472	4.870	128.873	6.260	108.812	4.836	136.273	5.985
Social Good	6	0.907	0.404	0.943	0.388	1.088	0.469	0.924	0.437	1.008	0.469	0.888	0.432	1.193	0.487
	8	0.939	0.421	1.012	0.426	1.183	0.499	0.978	0.459	1.017	0.500	0.916	0.473	1.117	0.493
	10	0.962	0.434	0.996	0.448	1.198	0.505	1.018	0.485	1.145	0.541	1.039	0.527	1.241	0.511
	12	1.040	0.514	1.058	0.473	1.264	0.516	1.089	0.532	1.174	0.571	1.492	0.859	1.189	0.513
	Avg	0.962	0.443	1.012	0.449	1.183	0.497	1.002	0.478	1.086	0.520	1.084	0.573	1.185	0.501
Traffic	6	0.169	0.202	0.182	0.208	0.186	0.225	0.166	0.205	0.200	0.232	0.271	0.383	0.255	0.344
	8	0.169	0.210	0.184	0.219	0.190	0.231	0.167	0.211	0.199	0.238	0.276	0.385	0.257	0.347
	10	0.173	0.218	0.190	0.230	0.195	0.241	0.176	0.219	0.205	0.241	0.279	0.385	0.259	0.347
	12	0.175	0.210	0.213	0.214	0.217	0.236	0.197	0.205	0.215	0.253	0.333	0.399	0.261	0.349
	Avg	0.171	0.210	0.192	0.218	0.197	0.233	0.176	0.210	0.205	0.241	0.289	0.388	0.258	0.347

Table 8: Full forecasting results across nine datasets in TTC benchmark at multiple forecast horizons.

Models		SpecTF (ours)		TaTS		MM-TSF		TimeXer		FreTS		TimeLLM		ChatTime	
Method		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Climate	12	0.655	0.576	0.759	0.620	0.767	0.647	0.741	0.615	0.857	0.665	0.656	0.583	0.868	0.607
	24	0.690	0.603	0.785	0.641	0.802	0.658	0.759	0.627	0.865	0.679	0.698	0.605	0.906	0.629
	36	0.701	0.610	0.797	0.653	0.811	0.660	0.769	0.629	0.889	0.683	0.700	0.616	0.943	0.650
	48	0.727	0.631	0.819	0.668	0.831	0.684	0.771	0.653	0.905	0.701	0.743	0.639	0.987	0.671
	Avg	0.693	0.605	0.790	0.636	0.802	0.662	0.760	0.631	0.878	0.682	0.699	0.611	0.926	0.639
Medical	12	0.927	0.666	1.265	0.768	1.422	1.056	1.293	0.861	1.493	1.051	0.911	0.662	1.572	1.104
	24	1.271	0.821	1.297	0.998	1.427	1.064	1.309	0.873	1.507	1.067	1.275	0.822	1.597	1.108
	36	1.441	0.888	1.332	1.017	1.443	1.077	1.317	0.885	1.519	1.075	1.483	0.895	1.622	1.126
	48	1.554	0.936	1.357	1.032	1.451	1.092	1.325	0.901	1.525	1.099	1.637	0.964	1.649	1.131
	Avg	1.298	0.828	1.313	0.953	1.436	1.072	1.311	0.880	1.511	1.073	1.327	0.836	1.610	1.117

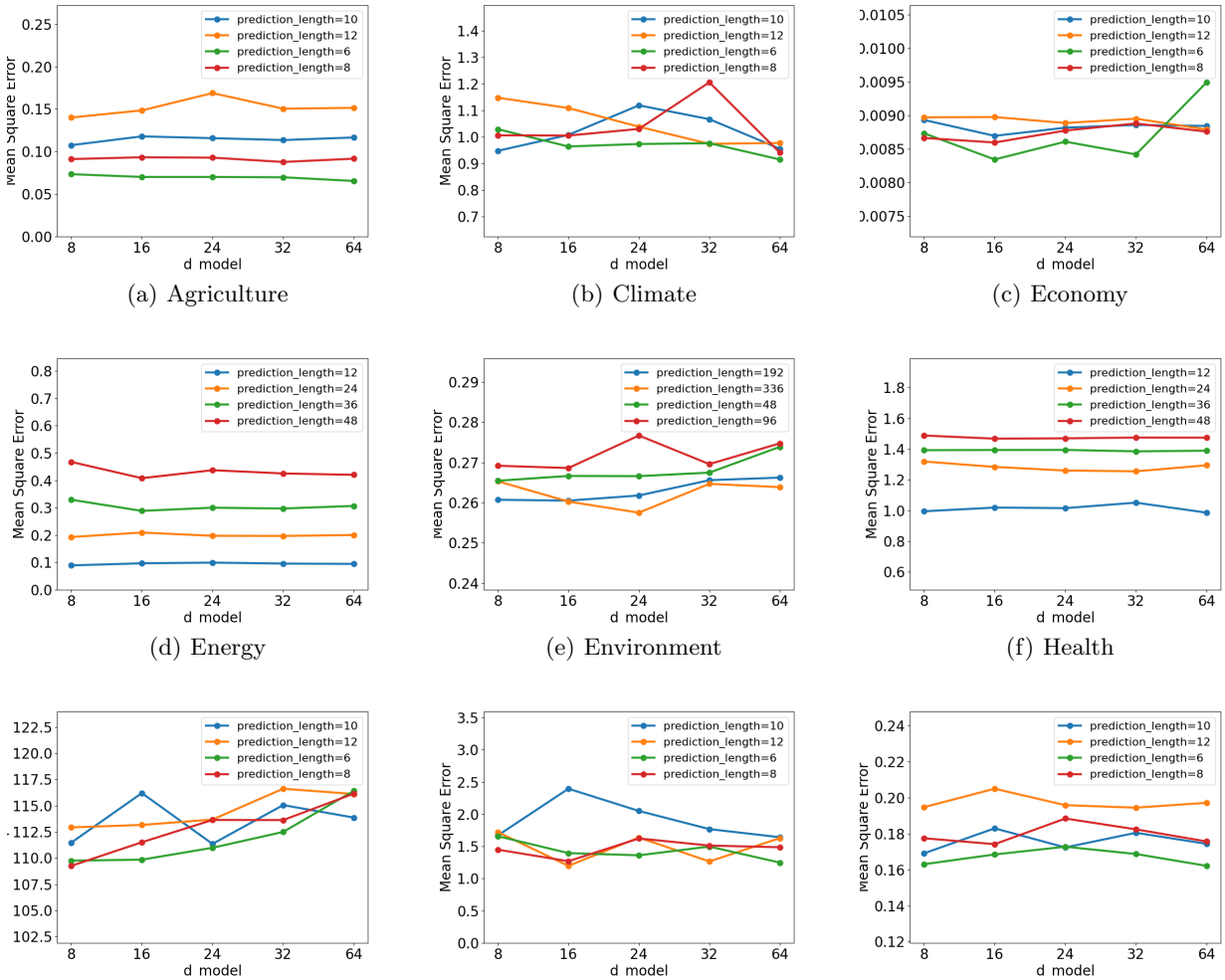


(a) Agriculture (b) Climate (c) Economy

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

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

Figure 4: Parameter study on learning rate.

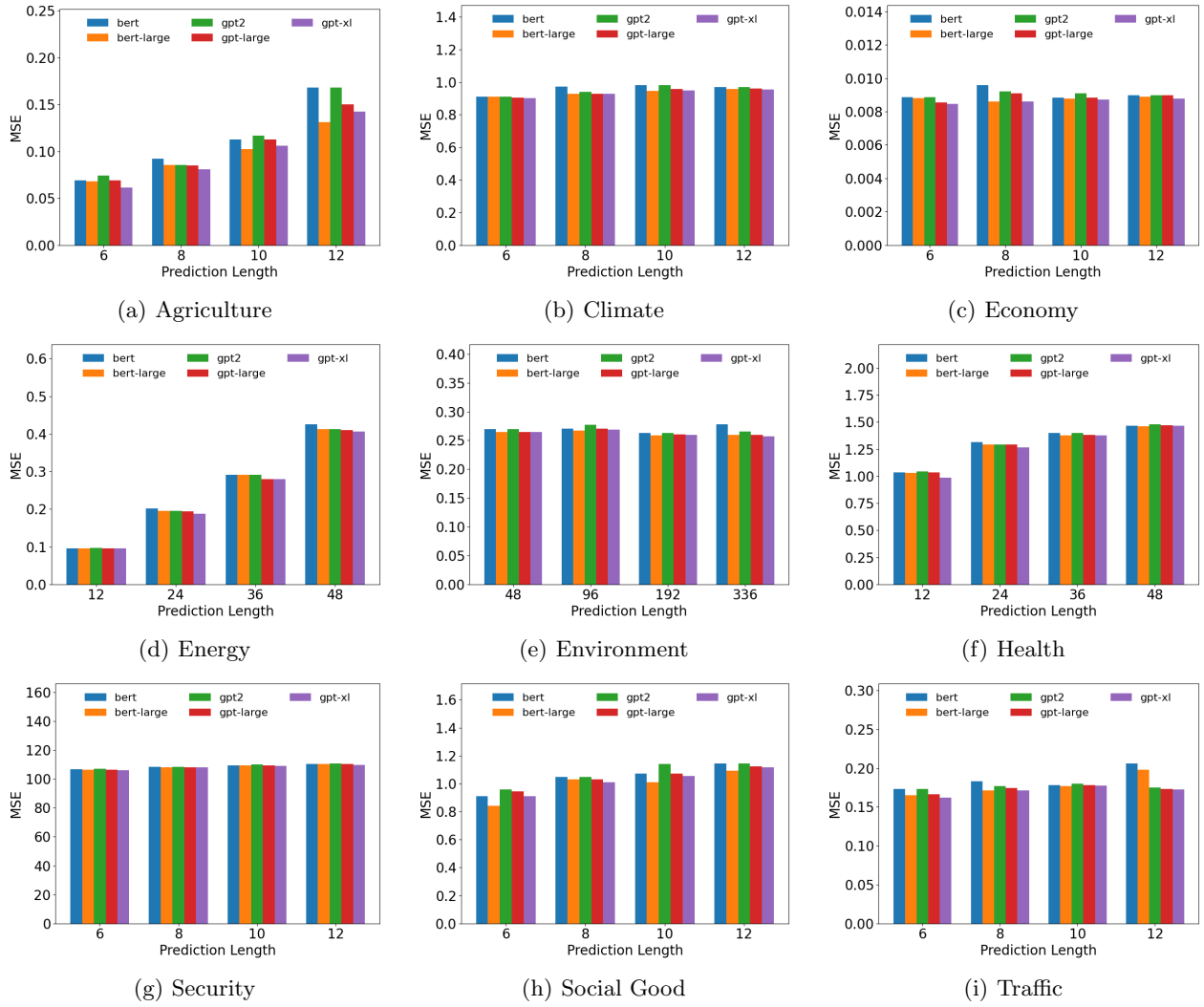


(a) Agriculture (b) Climate (c) Economy

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

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

Figure 5: Parameter study on model dimension d.



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

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

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

Figure 6: Performance comparison of different language models.

D BROADER IMPACTS

SpecTF framework advances multimodal forecasting accuracy while reducing computational costs, potentially benefiting applications like climate modeling and healthcare, where text-time series correlations drive critical decisions. However, ethical considerations arise regarding data privacy (e.g., sensitive textual inputs in medical records) and algorithmic bias propagation through spectral-text alignment, necessitating robust anonymization and fairness audits. The method's efficiency (via RFFT) lowers energy consumption compared to dense time-domain models, aligning with sustainable AI practices. Yet, deployment in high-stakes domains like finance or security risks misuse for market manipulation or surveillance if not governed by transparency protocols. Future work should integrate explainability tools to audit text-frequency interactions and ensure equitable outcomes across socioeconomic groups.