Empowering Time Series Analysis with Large-Scale Multimodal Pretraining

Peng Chen[1], **Siyuan Wang**[1], **Shiyan Hu**[1], **Xingjian Wu**[1], **Yang Shu**[1], **Zhongwen Rao**[2], **Meng Wang**[2], **Yijie Li**[2], **Bin Yang**[1], **Chenjuan Guo**[1*]
[1]East China Normal University, [2]Huawei Technologies Co., Ltd.
{pchen,sywang,syhu,xjwu}@stu.ecnu.edu.cn,
{yshu,cjguo,byang}@dase.ecnu.edu.cn,
{raozhongwen,wangmeng71,liyijie5}@huawei.com

Abstract

While existing time series foundation models primarily rely on large-scale unimodal pretraining, they lack complementary modalities to enhance time series understanding. Building multimodal foundation models is a natural next step, but it faces key challenges: 1) lack of a unified multimodal pretraining paradigm and large-scale multimodal corpora for time series analysis; 2) how to effectively integrate heterogeneous modalities and enhance model generalization. To address these challenges, we take an early step toward multimodal foundation models for time series analysis. We first propose a multimodal pretraining paradigm that leverages time series with endogenous modalities (derived images and text) and exogenous knowledge (real-world news), providing a comprehensive multi-view perspective for time series analysis. To support this, we develop an automated data construction pipeline to curate MM-TS, the large-scale multimodal time series dataset spanning six domains, with up to one billion points. Then we propose HORAI, a frequency-enhanced multimodal foundation model. It integrates two core components: the Frequency-enhanced Cross-Modality Encoder and the Time-Frequency Decoder, designed to effectively fuse multimodal features and enhance model generalization across modalities and domains. After pretraining on MM-TS, HORAI achieves state-of-the-art zero-shot performance on time series forecasting and anomaly detection tasks, demonstrating strong generalization. We provide our code at `https://github.com/decisionintelligence/HORAI`.

1 Introduction

Time series analysis is widely applied across diverse domains, including energy management, environment monitoring, and financial forecasting [Tian et al., 2024, 2026, Mei et al., 2025]. Existing time series analysis approaches, ranging from time-series-specific models [Nie et al., 2023, Qiu et al., 2025a, Liu et al., 2024a, Wu et al., 2025a, Chen et al., 2024a] to recent time series foundation models [Woo et al., 2024, Shi et al., 2025, Wu et al., 2025b], primarily rely on time series numerical modality to capture temporal patterns and uncover underlying regularities. While these methods have achieved competitive performance, this single-modality paradigm remains limited in its ability to capture the complex and multifaceted nature of real-world temporal dynamics [Xu et al., 2024a].

At the same time, foundation models in NLP and multimodal learning [Brown et al., 2020, Bai et al., 2023, Wu et al., 2024, Chen et al., 2024b] have shown that large-scale pretraining on massive datasets with complementary modalities can enhance generalization and adaptability across tasks. Inspired by these, we propose developing multimodal foundation models for time series analysis.

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

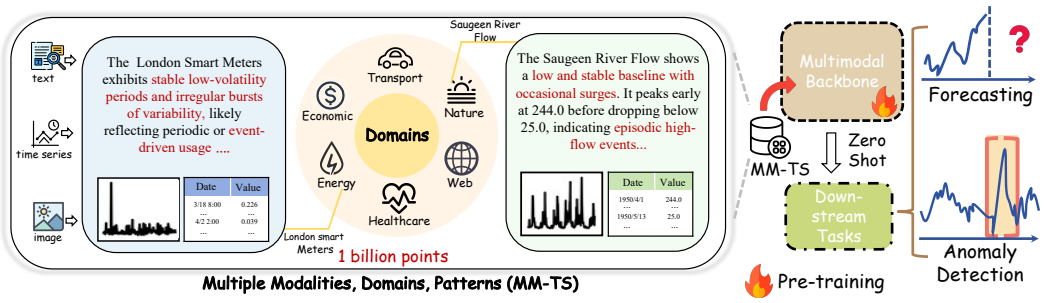


Figure 1: Left: The large-scale multimodal time series dataset MM-TS is characterized by its coverage of various modalities, heterogeneous domains, and diverse temporal patterns. Right: The foundation model HORAI is pre-trained on the MM-TS and evaluated on downstream scenarios and tasks.

By incorporating additional modalities for pretraining, such as text and images, the model leverages textual semantics and visual information to better capture complex temporal dynamics and strengthen time series understanding.

However, the development of multimodal foundation models faces several significant challenges. First, the field lacks a multimodal pretraining paradigm and large-scale multimodal corpora for time series analysis. Existing multimodal methods are tailored to specific datasets, lacking a unified pretraining paradigm for broad cross-modal analysis Jia et al. [2024], Li et al. [2025]. Moreover, existing large-scale datasets contain only numerical time series and lack complementary modalities such as text or images, which limits their suitability for multimodal pretraining Liu et al. [2024b], Woo et al. [2024]. Second, the architectural design for integrating different modalities in time series analysis remains underexplored. Each modality exhibits unique characteristics: text provides rich semantic information and offers a holistic, global description of events, whereas images capture localized details and spatial structures [Zhong et al., 2025]. Directly fusing time series with textual or visual modalities [Phuong and Lampert, 2019, Kim and Rush, 2016] may result in suboptimal alignment and ineffective representation learning. Therefore, it is critical to design fusion mechanisms that explicitly leverage the unique characteristics of each modality. Third, time series data from different domains exhibit diverse patterns, and the incorporation of multiple modalities further amplifies this diversity. Effectively modeling the heterogeneous patterns across modalities and domains, while enhancing the generalization ability of pretrained models, remains a challenge. Consequently, advancing multimodal foundation models for time series analysis requires further research and exploration.

In this paper, we take an early step toward developing multimodal foundation models for time series analysis. *On the pretraining paradigm and dataset side*, we propose a multimodal pretraining paradigm that leverages time series with endogenous modalities (derived images and text) and exogenous knowledge (real-world news), providing a comprehensive multi-view perspective for time series analysis. Endogenous modalities capture intrinsic time series characteristics through semantic and visual structures, while exogenous news provides supplementary background context. To support this, we develop an *automated data construction pipeline* that maps raw sequences to visual and textual modalities, aligns them with external news sources, and incorporates multiple LLMs for rigorous quality verification. This process curates MM-TS, the large-scale multimodal time series dataset. As illustrated in Figure 1, MM-TS integrates three modalities, spanning six diverse domains and a wide range of temporal patterns, with up to one billion points. The three modalities exhibit strong correlations and complementary characteristics, making MM-TS well-suited for multimodal pretraining to learn generalized representations. This dataset provides a solid foundation for studying multimodal foundation models.

On the modeling side, we propose **HORAI**, a frequency-enhanced multimodal time series foundation model built on an autoregressive architecture, which consists of two core components: Frequency-enhanced Cross-Modality Encoder and Time-Frequency Decoder. *In the Frequency-enhanced Cross-Modality Encoder*, we leverage the correspondence between modality-specific information and different frequency components of time series to align multiple modalities and enhance time series understanding. Specifically, time series are decomposed into multiple frequency bands, where low-frequency components capture long-term dynamics and align with the global semantics embedded in text, while mid- and high-frequency components encode rapid variations that tend to correspond to the localized patterns present in visual inputs. *In the Time-Frequency Decoder*, we design a

Time-Frequency MoE-FFN to learn generalized multimodal representations from multi-domain data. We introduce a time-frequency router that dynamically assigns each token to the suitable expert based on both temporal and frequency features. By incorporating frequency-domain features, the router gains additional cues to better distinguish similar patterns and group them coherently, which enhances feature consistency and improves generalization. Specifically, our contributions are as follows:

- **New Paradigm and Pretraining Dataset:** We propose a multimodal pretraining paradigm that leverages time series along with endogenous modalities (derived images and text) and exogenous knowledge (news), enhancing time series analysis from a comprehensive multi-view perspective. To support this, we develop an automated data construction pipeline to curate MM-TS, the large-scale multimodal time series dataset.
- **New Model:** We propose HORAI, a frequency-enhanced multimodal foundation model for Time series analysis, which incorporates two core components, the frequency-enhanced cross-modality encoder and the time-frequency decoder, designed to effectively fuse multi-modal features and enhance model generalization across modalities and domains.
- **Comprehensive Evaluation:** After pre-training on large-scale multimodal time series data, HORAI achieves state-of-the-art performance in time series forecasting and anomaly detection across zero-shot inference and few-shot learning situations, which demonstrates strong task versatility and generalization ability.

2 Related work

2.1 Time Series Analysis

Time series analysis spans a wide range of tasks, including forecasting and anomaly detection [Qiu et al., 2024, Chen et al., 2026, Darban et al., 2025, Qiu et al., 2025b, Paparrizos et al., 2022a]. Existing approaches can be broadly divided into unimodal and multimodal methods. **Unimodal methods** focus on time series data and employ diverse architectures to model temporal dynamics and channel correlations. These include MLP-based models [Zeng et al., 2023, Xu et al., 2024b], RNN-based models [Flunkert et al., 2017], CNN-based models [Wu et al., 2023, Cheng et al., 2026a], GNN-based models [Wu et al., 2021], as well as Transformer-based architectures for capturing long-range dependencies [Zhang and Yan, 2023, Nie et al., 2023]. In contrast, **multimodal methods** integrate additional modalities or external knowledge to enhance time series analysis [Jin et al., 2024, Chen et al., 2025, Pan et al., 2024, Hu et al., 2026]. Time-LLM [Jin et al., 2024] incorporates textual prompts when adapting pretrained language models to forecasting, while Time-VLM [Zhong et al., 2025] integrates temporal features with generated visual and textual representations through pretrained vision-language models. External text provides another source of information: GPT4MTS [Jia et al., 2024] combines numerical sequences with news-derived textual context, and TaTS [Li et al., 2025] models paired texts as auxiliary time series variables. Time-MMD [Liu et al., 2024c] further supports multimodal evaluation with aligned numerical and textual data across multiple domains.

Building on these studies, we construct MM-TS for large-scale multimodal pretraining. It combines images and descriptions derived from historical observations with external news. This construction distinguishes complementary views of the observed sequence from external contextual information and supports studying their joint contribution to transferable temporal representations.

2.2 Time Series Foundation Models

Time series foundation models learn transferable temporal representations through pretraining on diverse datasets [Ansari et al., 2025, Das et al., 2024, Goswami et al., 2024, Ekambaram et al., 2024, Podest et al., 2026, Shi et al., 2025, Cheng et al., 2026b,c]. Among models pretrained on numerical sequences, Timer [Liu et al., 2024b] uses an autoregressive next-token prediction objective, while MOIRAI Woo et al. [2024] handles variations in sampling frequencies, dimensionality, and predictive distributions. ROSE [Wang et al., 2025a] combines decomposed frequency learning with time series registers to capture generalized and domain-specific representations. Multimodal foundation models further incorporate information beyond numerical observations. ChatTime [Wang et al., 2025b] unifies numerical and textual processing within a language-model framework and supports zero-shot

forecasting. Aurora [Wu et al., 2025c] supports multimodal zero-shot forecasting through modality-guided temporal modeling and prototype-guided flow matching. These developments establish multimodal pretraining as an existing direction for time series foundation models.

HORAI investigates frequency information as a guide for multimodal integration and expert routing. Its encoder aligns low-frequency temporal components with text and mid-to-high-frequency components with images, while its decoder incorporates frequency features into expert selection. Together with MM-TS, this design combines sequence-derived views and external news within an autoregressive pretraining framework, evaluated on forecasting and anomaly detection.

3 Methodology

3.1 Large-Scale Multimodal Time Series Dataset

Large-scale datasets are the cornerstone of pre-training foundation models, enabling them to acquire transferable knowledge and improve generalization across diverse downstream scenarios. However, existing large-scale time series corpora are mostly confined to unimodal time series, which limits the potential of multimodal learning. To address this, we conduct MM-TS, a large-scale multimodal time series dataset for pre-training. As shown in Figure 1, MM-TS integrates three modalities: time series, text, and image, covering six diverse domains, including Energy, Healthcare, Web, Nature, Transport, and Economics. In total, MM-TS contains over 1 billion time points.

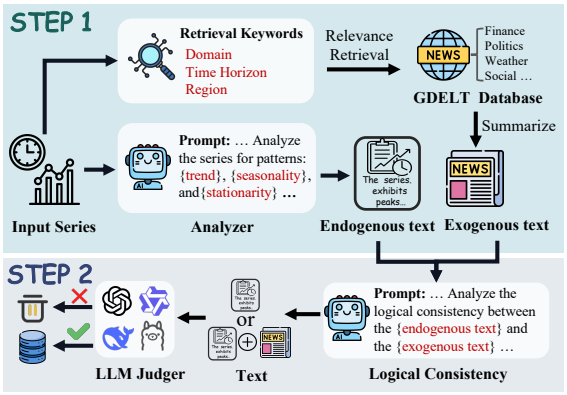


Figure 2: The data construction pipeline for multimodal text. It comprises two main stages: 1) Contextual Synthesis, involving endogenous pattern analysis and exogenous news retrieval; and 2) Quality Alignment, which ensures consistency between the synthesized texts and filters low-quality data via an LLM judger.

For the time series modality, MM-TS spans multiple temporal granularities, including seconds, minutes, hours, and months, and captures diverse characteristics such as periodicity, trends, and non-stationarity (see Appendix A.1 for details). For the visual modality, we construct line-plot images directly from time series, offering an intuitive view of temporal fluctuations and structural patterns.

For the textual modality, to address the scarcity of high-quality, semantically aligned time series and textual pairs, we devise an automated data construction pipeline that uniquely integrates endogenous text with exogenous news. Endogenous text captures the semantic features of time series patterns, and exogenous news provides supplementary background context. Figure 2 shows that the pipeline includes two main stages: Contextual Synthesis and Quality Alignment. In the Contextual Synthesis stage, we adopt a dual-source generation strategy. Given a time series sample, an LLM analyzer GPT-4o first analyzes patterns, such as trends and seasonalities, and converts them into structured semantic descriptions as Endogenous text. In parallel, we retrieve and summarize semantically related news events based on keywords from the GDELT Database to serve as Exogenous text. Subsequently, the Quality Alignment stage ensures logical coherence and generation fidelity via two substeps: 1) Logical Consistency Filtering, where GPT-4o discards exogenous news conflicting with endogenous patterns to prevent spurious correlations; and 2) Ensemble Quality Evaluation, where the unified text is scored by multiple LLM judges on plausibility and relevance, retaining only high-confidence text. Further details on the text generation are in the Appendix A.2.

For each forecasting example, we first split the sequence into a historical context $\mathbf{X}_{1:L}$ and a future target $\mathbf{X}_{L+1:L+F}$. Endogenous descriptions and line-plot images are generated exclusively from $\mathbf{X}_{1:L}$, while exogenous news is restricted to information available before the forecast origin. The same temporal boundary is maintained during training and evaluation.

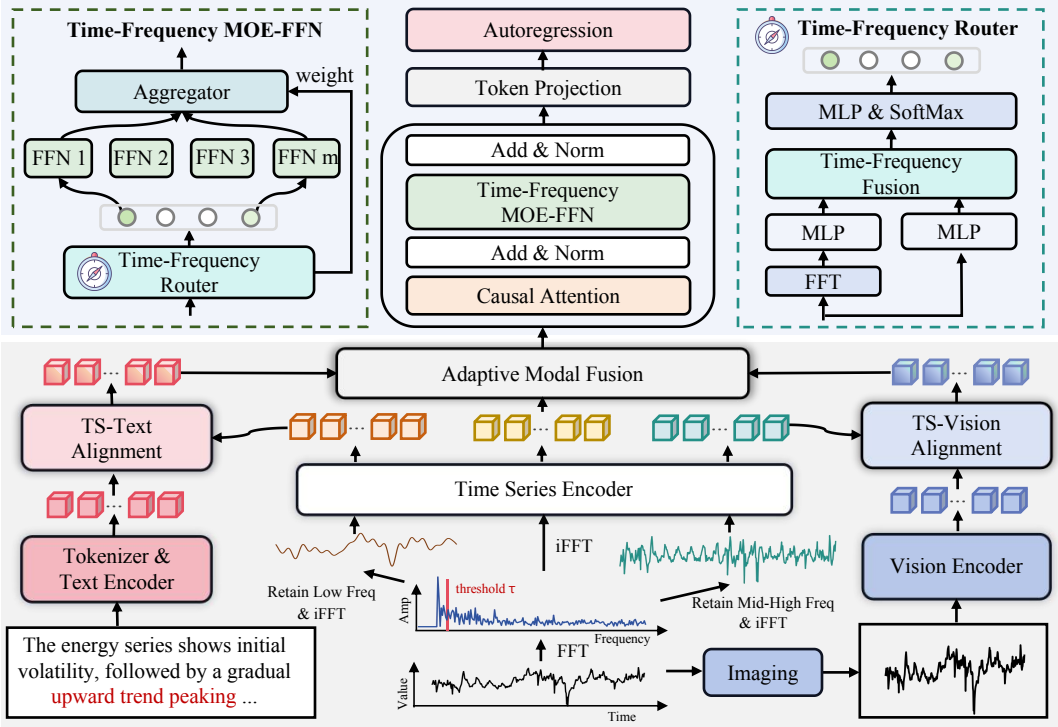


Figure 3: The framework of the proposed HORAI consists of a Frequency-Enhanced Cross-Modality Encoder (gray region) and a Time-Frequency Decoder (blue region).

3.2 HORAI

To better leverage cross-modal and cross-domain features for time series understanding, we propose HORAI, a frequency-enhanced multimodal foundation model for time series analysis. It consists of two core components: the Frequency-guided Cross-Modality Encoder and the Time-Frequency Decoder. As illustrated in Figure 3, in the Cross-Modality Encoder, the input time series is first decomposed into low-frequency and mid-to-high-frequency components, which are aligned with textual and visual features, respectively. Then, an adaptive modality fusion module combines these aligned representations to produce unified multimodal representations. In the Time-Frequency Decoder, the multimodal representations are first passed into a Time-Frequency MoE-FFN, which is designed to capture diverse patterns across multiple domains. To guide the routing of tokens to appropriate experts, both temporal-domain and frequency-domain features are incorporated. The inclusion of frequency information provides additional cues that help distinguish similar patterns and group them coherently, enhancing the cross-modality and cross-domain generalization. Finally, the representations are projected through a token projection layer for autoregressive pre-training.

3.2.1 Frequency-Enhanced Cross-Modality Encoder

Multimodal Embedding. For notational simplicity, we describe the method using a univariate time series, which can be easily extended to the multivariate case by treating each channel independently. Given a historical input $\mathbf{X}_{ts} = X_{1:L} \in \mathbb{R}^T$, our goal is to forecast the subsequent target sequence $\mathbf{Y} = X_{L+1:L+F} \in \mathbb{R}^H$, where L and F denote the input length and forecast horizon, respectively. We first apply instance normalization [Kim et al., 2021] to the historical input to mitigate distribution shift, resulting in $\mathbf{X}_{\text{norm}} \in \mathbb{R}^L$.

Since different frequency components capture different aspects of temporal dynamics, with low-frequency components reflecting global trends and mid-to-high-frequency components capturing local variations, we transform the normalized sequence into the frequency domain by the Fast Fourier Transform (FFT), obtaining $\mathbf{X}_{\text{freq}} \in \mathbb{R}^{L/2+1}$. To separate different frequency bands, we set a ratio parameter α to define a cutoff threshold $\tau = \alpha \cdot (L/2 + 1)$. Based on this threshold,

we construct two binary masks: $\mathbf{M}_{\text{low}} \in \{0,1\}^{L/2+1}$ to retain the low-frequency components and $\mathbf{M}_{\text{mh}} \in \{0,1\}^{L/2+1}$ to retain the mid-to-high-frequency components. Applying these masks to $\mathbf{X}_{\text{freq}}$ by element-wise multiplication, which are then transformed back into the time domain using the inverse FFT (iFFT). This process produces the low-frequency sequence $\mathbf{X}_{\text{low}} \in \mathbb{R}^L$ and the mid-to-high-frequency sequence $\mathbf{X}_{\text{mh}} \in \mathbb{R}^L$.

$$\mathbf{X}_{\text{low}} = \text{iFFT}(\mathbf{X}_{\text{freq}} \odot \mathbf{M}_{\text{low}}), \qquad \mathbf{X}_{\text{mh}} = \text{iFFT}(\mathbf{X}_{\text{freq}} \odot \mathbf{M}_{\text{mh}}). \tag{1}$$

Subsequently, we employ a patching strategy to divide $\mathbf{X}_{\text{low}}$, $\mathbf{X}_{\text{mh}}$, and $\mathbf{X}_{\text{norm}}$ into N_{ts} patches with patch size S. These patches are projected and fed into the time-series encoder [Nie et al., 2023], producing corresponding time series representations: $\mathbf{E}_{\text{low}}$, $\mathbf{E}_{\text{mh}}$, and $\mathbf{E}_{\text{ts}} \in \mathbb{R}^{N_{ts} \times D_{ts}}$.

For the textual input $\mathbf{X}_{\text{text}} \in \mathbb{R}^{L_{\text{text}}}$, we employ a text tokenizer followed by a pre-trained text encoder Qwen-0.5B Bai et al. [2023] to extract semantic features, yielding $\mathbf{E}_{\text{text}} \in \mathbb{R}^{L_{\text{text}} \times D_{\text{text}}}$. For the visual input $\mathbf{X}_{\text{img}}$, we apply a patching strategy and a pre-trained vision encoder ViT-Base Dosovitskiy et al. [2021] to obtain image representations $\mathbf{E}_{\text{img}} \in \mathbb{R}^{N_{\text{img}} \times D_{\text{img}}}$.

Frequency-enhanced Cross-Modality Alignment. Time series often exhibit rich frequency-dependent patterns, where low-frequency components capture global trends and mid-to-high-frequency components reflect local variations. Meanwhile, different modalities contribute differently to these patterns: textual information tends to describe global trends, aligning with low-frequency time series components, whereas visual information focuses more on short-term variation, corresponding to mid-to-high-frequency components [Zhong et al., 2025]. Motivated by this, we propose a frequency-enhanced cross-modal fusion that explicitly leverages the characteristic correspondence between modalities and frequency components. Additionally, given the large number of tokens in text and image modalities, we further integrate a Flow-Attention-based alignment mechanism to efficiently model cross-modal interactions.

In the TS-Text Alignment module, the low-frequency time series embeddings and textual embeddings are projected by MLPs into a shared representation space D_{model}. Cross-modal fusion is then performed efficiently using the Flow-Attention mechanism Wu et al. [2022]. The core idea is to treat attention as a flow of information and leverage the flow conservation principle to optimize the transmission and aggregation of features across modalities. Specifically, the low-frequency time series embeddings $\mathbf{E}_{\text{low}}$ are mapped to serve as the Query $\mathbf{Q}$, while the textual embeddings $\mathbf{E'}_{\text{text}}$ are mapped to serve as the Key $\mathbf{K}$ and Value $\mathbf{V}$. The information flow is computed as:

$$\begin{aligned} \mathbf{I}_i = \phi(\mathbf{Q}_i) \sum_{j=1}^{N_{\text{text}}} \phi(\mathbf{K}_j)^T, \quad \mathbf{O}_j = \phi(\mathbf{K}_j) \sum_{i=1}^{N_{\text{ts}}} \phi(\mathbf{Q}_i)^T, \quad \hat{\mathbf{O}} = \phi(\mathbf{K}) \sum_{i=1}^{N_{\text{ts}}} \frac{\phi(\mathbf{Q}_i)^T}{\mathbf{I}_i}, \\ \mathbf{E'}_{\text{text}} = \frac{\phi(\mathbf{Q})}{\mathbf{I}} (\phi(\mathbf{K})^T (\text{Softmax}(\hat{\mathbf{O}}) \odot \mathbf{V})), \end{aligned} \tag{2}$$

$\phi(\cdot)$ denotes the non-linear projection to the flow space, $\mathbf{I}_i$ and $\mathbf{O}_j$ represent the total outgoing and incoming flows for each token. The output $\mathbf{E'}_{\text{text}}$ is a flow-attention enhanced textual embedding, which has been adaptively aligned with the low-frequency time-series features.

Similar to the low-frequency time-series and text fusion, the TS-Vision Alignment module also leverages the Flow-Attention mechanism to integrate mid-to-high-frequency time-series embeddings $\mathbf{E}_{\text{mh}}$ with image embeddings $\mathbf{E}_{\text{img}}$, yielding aligned image representations $\mathbf{E'}_{\text{img}} \in \mathbb{R}^{N_{\text{ts}} \times D_{model}}$ for subsequent multimodal fusion.

Adaptive Modal Fusion. Considering that the contributions of image and text representations vary across different time series patterns, we adaptively fuse the aligned image and text embeddings. The aligned image embeddings $\mathbf{E'}_{\text{img}}$ and text embeddings $\mathbf{E'}_{\text{text}}$ are concatenated along the feature dimension and then passed through a linear projection followed by a sigmoid function σ to perform gated weighting $\mathbf{G}$, producing the multimodal representation $\mathbf{E}_{\text{mm}}$. This representation is subsequently added to the time series embeddings $\mathbf{E}_{\text{ts}}$ to obtain the fused representation $\mathbf{E}_{\text{fused}} \in \mathbb{R}^{N_{ts} \times D_{model}}$.

$$\mathbf{G} = \sigma(W_g[\mathbf{E'}_{\text{img}}, \mathbf{E'}_{\text{text}}] + b_g), \quad \mathbf{E}_{\text{fused}} = \mathbf{G} \odot \mathbf{E'}_{\text{img}} + (1 - \mathbf{G}) \odot \mathbf{E'}_{\text{text}} + \mathbf{E}_{\text{ts}}. \tag{3}$$

3.2.2 Time-Frequency Decoder

Large-scale time series data inevitably involves diverse domains, which gives rise to a wide variety of temporal patterns [Wang et al., 2025a, Woo et al., 2024]. The incorporation of textual and visual modalities further amplifies the diversity. To address this challenge, we propose a Time-Frequency Decoder designed to capture and adapt to different patterns, enhancing the generalization ability of pre-trained models. As illustrated in Figure 3, the Time-Frequency Decoder consists of key components including Causal Attention, Normalization layers, and a Time-Frequency MoE-FFN.

Time-Frequency MoE-FFN. Different expert networks can capture distinct patterns from large-scale data, so effectively routing multimodal features to the appropriate experts is crucial. However, relying only on temporal-domain features may lead to entangled representations across different patterns, which makes pattern discrimination less straightforward. By incorporating frequency-domain features, similar patterns can be represented more compactly, offering additional cues for more accurate expert routing. Motivated by this, we propose the Time-Frequency Router, which integrates both temporal and frequency information to enhance the routing process.

Based on the fused multi-modal representation $\mathbf{E}_{\text{fused}}$, we obtain representation $\mathbf{H}$ through causal attention followed by normalization. In the router, each token of $\mathbf{H}$ is projected in parallel across both temporal and frequency domains: (i) an MLP produces temporal representations $\mathbf{H}_{\text{temp}}$, while (ii) an FFT followed by an MLP yields frequency representations $\mathbf{H}_{\text{freq}}$. These dual-domain signals are adaptively integrated via a learnable gating function G_{router}, resulting in router representation $\mathbf{H}_r \in \mathbb{R}^{N_{ts} \times D_{model}}$:

$$\mathbf{H}_{\text{r}}^{i} = \mathbf{G}_{\text{router}} \odot \text{MLP}(\mathbf{H}^{i}) + (1 - \mathbf{G}_{\text{router}}) \odot \text{MLP}(\text{FFT}(\mathbf{H}^{i})). \quad i = 1, \cdots, N_{ts}. \tag{4}$$

Given $\mathbf{H}_r$, the router applies an MLP-based routing function to compute routing weights $\mathbf{W} \in \mathbb{R}^{M}$, which determine expert assignment. Following a Top-K strategy, the router selects the K experts with the highest weights, denoting the set of their indexes as $\mathcal{K}$. Then their outputs are aggregated through weight-normalized fusion, producing the representation $\mathbf{H}_{\text{moe}} \in \mathbb{R}^{N_{ts} \times D_{model}}$:

$$\mathbf{H}_{\text{moe}}^{i} = \sum_{j \in \mathcal{K}} \frac{\exp(\mathbf{W}_j)}{\sum_{m \in \mathcal{K}} \exp(\mathbf{W}_m)} \text{FFN}_j(\mathbf{H}^{i}), \quad i = 1, \cdots, N_{ts}. \tag{5}$$

Autoregressive Training. Given the strong performance of the autoregressive paradigm in both NLP Bai et al. [2023], Brown et al. [2020] and time series domains [Liu et al., 2025, 2024b], we adopt a GPT-style training objective to predict the next token. This autoregressive formulation not only supports variable input and output lengths flexibly during inference but also excels at iterative, multi-step generation. Specifically, HORAI conditions on the historical input $\mathbf{X}_{ts}$ and its associated multimodal inputs to autoregressively generate future patches, yielding the forecast $\widehat{\mathbf{Y}} \in \mathbb{R}^{F}$. The forecasting loss is computed exclusively over the future target values:

$$\mathcal{L}_{\text{train}} = \frac{1}{F} \sum_{f=1}^{F} \left(\widehat{Y}_f - Y_f \right)^2, \qquad Y_f = X_{L+f}.$$

4 Experiments

Datasets. We perform pre-training of HORAI on our proposed MM-TS dataset and ensure no overlap between the pre-training and downstream evaluation datasets, except for the related-series overlap in Social Good. To assess HORAI's capability for time series analysis, we use the widely used evaluation datasets [Liu et al., 2024c] for forecasting and anomaly detection tasks. Specific dataset information is in Appendix A.

Baselines. *For the forecasting task*, we select eight SOTA foundation models as baselines: Aurora [Wu et al., 2025c], Chronos2 [Ansari et al., 2025], TimesFM [Das et al., 2024], ChatTime [Wang et al., 2025b], Tirex 2 [Podest et al., 2026], VisionTS [Chen et al., 2024c], ROSE [Wang et al., 2025a], MOIRAI [Woo et al., 2024], and four multimodal time-series-specific models: GPT4MTS [Jia et al., 2024], TATS [Li et al., 2025], GPT4TS [Zhou et al., 2023], TimeVLM [Zhong et al., 2025]. *For the anomaly detection task*, we select three foundation models: DADA [Shentu et al.,

2025], Timer [Liu et al., 2024b], UniTS [Gao et al., 2024], and nine time-series-specific models: GPT4TS, LLMMixer [Kowsher et al., 2024], TimesNet [Wu et al., 2023], DCdetector [Yang et al., 2023], Anomlay Transformer(A.T.) [Xu et al., 2022], PatchTST [Nie et al., 2023], HBOS [Goldstein and Dengel, 2012], IForest [Liu et al., 2008], and PCA [Shyu et al., 2003].

Settings. During pre-training, HORAI is optimized using the Adam optimizer with an initial learning rate of 0.0005 and trained for 20 epochs, employing an early stopping strategy with a patience of 5 epochs. For the forecasting task, all methods predict future values at four horizons to ensure a fair comparison. Additionally, *none of the models employ the drop-last strategy* [Qiu et al., 2024]. For the anomaly detection task, evaluation is conducted using three score-based metrics: AUC-ROC, VUS-ROC, and VUS-PR [Paparrizos et al., 2022b], which are threshold-independent. Notably, **time series foundation models perform zero-shot inference directly, whereas time-series-specific models are trained in a full-shot setting for comparison**.

4.1 Main Results

Forecasting Results. As shown in Table 1, HORAI achieves state-of-the-art forecasting performance compared to both unimodal foundation models and multimodal time-series-specific models, achieving top performance on 5 out of 9 cases. Specifically, relative to unimodal foundation models, HORAI outperforms ROSE with reductions of 32.4% in MSE. These results indicate that HORAI effectively leverages multimodal information to enhance time series understanding and improve predictive accuracy. Compared to multimodal time-series-specific models trained in a full-shot manner, HORAI achieves superior performance even in the zero-shot setting: exceeding GPT4MTS by 11.3% in MSE, and surpassing TimeVLM by 12.1% in MSE. This demonstrates that pre-training on the large-scale multimodal time series dataset equips HORAI with strong generalization ability.

Table 1: Time series forecasting results under zero-shot and full-shot settings, reported as the average across four prediction horizons. The best results are highlighted in **bold**, and the second-best results are underlined. Full results are presented in Tables 16 and 17. Comparisons with unimodal time-series-specific models and LLM-based reasoning models are presented in Tables 9 and 10.

Type	Time Series Foundation Models (Zero-Shot)									Time-Series-Specific Models (Full-Shot)			
Models	HORAI	Aurora	ChatTime	Chronos2	TimesFM	TiRex 2	ROSE	VisionTS	MOIRAI	GPT4MTS	TATS	GPT4TS	TimeVLM
Metrics	MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE
Agriculture	0.201	0.272	0.369	**0.197**	0.254	0.203	0.345	0.290	0.272	0.225	0.215	0.220	0.237
Climate	**0.857**	0.865	1.860	1.489	1.157	1.558	1.475	1.307	1.921	1.182	1.180	1.184	1.195
Energy	**0.208**	0.255	0.247	0.242	0.282	0.240	0.386	0.304	0.324	0.262	0.255	0.260	0.260
Environment	0.309	**0.276**	0.365	0.349	0.339	0.332	0.392	0.354	0.351	0.323	0.319	0.322	0.319
Social Good	0.819	0.838	1.069	1.128	**0.805**	0.886	1.141	1.126	1.430	0.920	0.918	0.917	0.868
Traffic	0.165	**0.161**	0.596	0.416	0.184	0.438	0.341	0.281	0.406	0.203	0.179	0.206	0.216
EWJ	**0.589**	0.605	0.887	0.619	0.738	0.760	0.706	0.645	0.937	0.626	0.612	0.607	0.609
KR	**0.551**	0.578	0.565	0.560	0.559	0.616	0.741	0.671	0.992	0.555	0.578	0.578	0.584
MDT	**0.376**	0.396	0.496	0.389	0.421	0.429	0.461	0.433	0.606	0.385	0.389	0.391	0.392
1st Count	**5**	2	0	1	1	0	0	0	0	0	0	0	0

Anomaly Detection Results. As illustrated in Table 2, HORAI achieves state-of-the-art anomaly detection performance compared to both unimodal foundation models and time-series-specific models, attaining top results on 12 out of 15 cases. Compared to DADA, a general time series anomaly detector, HORAI outperforms it by 13.4%, 19.5%, and 19.2% in AUC-ROC, VUS-ROC, and VUS-PR, respectively, under the zero-shot setting. This highlights that integrating multimodal data, such as text and images, enables the model to identify anomalous patterns better. Against time-series-specific models, HORAI outperforms GPT4TS by 12.2%, 20.2%, and 22.6% in AUC-ROC, VUS-ROC, and VUS-PR. These results demonstrate that pre-training on large-scale, multi-domain data equips HORAI with general detection capability, effectively distinguishing between diverse normal and anomalous patterns.

4.2 Ablation Study

To evaluate the effectiveness of each component in HORAI, we conduct ablation experiments. Figure 4 illustrates the unique impact of each module. Removing the image and text modalities (W/O Modality) leads to a drop in performance, demonstrating that HORAI effectively leverages textual semantics and visual spatial information to enhance time series modeling. In the Modality Exchange variant, mid- and high-frequency time series features are aligned with

Table 2: Time series anomaly detection results under zero-shot and full-shot settings. The best results are in **bold**, and the second-best results are underlined. More metric results are in Table 18.

Type		Time Series Foundation Models (Zero-Shot)				Time-Series-Specific Models (Full-Shot)								
Datasets	Metrics	HORAI	DADA	Timer	UniTS	GPT4TS	LLMMixer	TimesNet	DCdetector	A.T.	PatchTST	HBOS	IForest	PCA
EWJ	AUC_ROC	**86.95**	79.11	76.15	79.87	75.58	57.69	82.39	53.40	43.81	78.53	71.82	69.20	54.35
	VUS_ROC	**82.78**	71.79	67.72	73.91	67.95	52.79	75.76	47.10	31.75	71.96	62.07	59.24	45.26
	VUS_PR	**48.27**	43.36	33.17	39.32	35.63	15.13	43.15	15.37	10.85	36.08	41.19	37.81	19.38
MDT	AUC_ROC	**91.22**	79.04	75.65	73.19	74.79	60.30	86.67	53.82	56.44	84.55	60.26	63.92	54.51
	VUS_ROC	**86.82**	66.76	60.28	58.67	62.30	46.80	83.40	45.02	44.53	77.69	55.30	54.02	44.09
	VUS_PR	**56.88**	46.81	38.38	37.61	44.81	15.21	52.13	15.72	15.93	41.67	44.77	35.32	22.93
KR	AUC_ROC	**96.38**	79.53	66.72	80.95	78.30	65.77	85.88	52.97	51.25	82.15	75.16	74.45	63.58
	VUS_ROC	**93.54**	70.82	75.99	73.93	67.81	47.06	79.00	43.04	41.97	74.65	58.77	60.70	47.51
	VUS_PR	**60.76**	45.90	51.41	43.32	38.23	19.10	51.60	8.49	7.94	36.18	54.17	43.31	24.19
Energy	AUC_ROC	**68.44**	62.33	60.54	63.38	66.54	61.31	68.36	48.75	38.68	66.70	60.80	60.32	61.14
	VUS_ROC	**62.42**	54.37	46.03	51.15	53.10	53.04	59.47	45.93	31.56	58.31	51.50	53.61	53.07
	VUS_PR	35.24	34.18	29.46	31.04	31.68	30.35	38.61	22.57	19.69	34.41	42.57	**46.03**	44.30
Weather	AUC_ROC	81.49	66.37	80.86	81.22	74.47	79.60	81.10	47.90	47.11	**82.02**	64.47	67.81	67.71
	VUS_ROC	80.40	61.03	73.22	75.08	70.03	71.71	**81.91**	45.56	43.32	79.97	54.16	56.45	57.38
	VUS_PR	**50.76**	30.00	43.21	44.35	41.30	43.47	50.09	18.33	19.17	50.13	46.58	49.66	47.13
1st Count		**12**	0	0	0	0	0	1	0	0	0	0	1	0

texts, while low-frequency features are aligned with images. In contrast, HORAI aligns low-frequency features with text and mid- to high-frequency features with images, effectively exploiting the correspondence between modality-specific information and different frequency components of the time series, which improves modeling performance. This demonstrates that frequency-aware cross-modality alignment is crucial for capturing complementary patterns across modalities. Replacing the Time-Frequency MoE-FFN with a standard FFN (W/O MoE-FFN) shows that the MoE-FFN allows each expert to capture distinct patterns, thereby enhancing the model's generalization ability. Removing frequency information from the router (W/O Router) demonstrates that incorporating frequency information helps guide multimodal tokens to the most appropriate FFN experts, further improving performance.

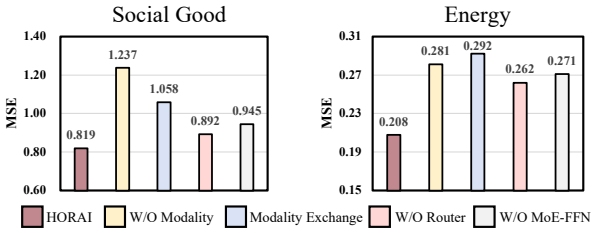


Figure 4: Ablation study on the Social Good dataset and the Energy dataset.

4.3 Sensitivity Analysis

We conduct sensitivity experiments on two key parameters: the frequency threshold α and the number of selected experts K. As shown in Table 3, while the absolute optimum fluctuates slightly between 0.025 and 0.1, setting α to 0.05 consistently emerges as the most robust global default. This value distinctly partitions low-frequency from mid-to-high-frequency features, facilitating optimal alignment with text and image modalities. A larger α retains more frequency components in the low-frequency branch, potentially incorporating local variations that interfere with global semantic alignment. Conversely, an overly small α removes useful low-frequency information, weakening the representation of long-term temporal patterns. As shown in Table 4, selecting the Top-2 or Top-3 experts yields superior performance. Activating all experts tends to introduce redundancy from irrelevant experts, thereby diluting the model's generalization. Whereas selecting only a single expert limits the representational capacity, preventing the model from modeling diverse patterns.

Table 3: Hyper-parameter sensitivity analysis about the frequency threshold α.

	α = **0.01**	α= **0.025**	α= **0.05**	α= **0.1**	α = **0.25**	α = **0.5**
Metrics	MSE	MSE	MSE	MSE	MSE	MSE
Agriculture	0.210	0.206	**0.201**	0.207	0.225	0.247
Climate	0.868	0.864	0.857	**0.855**	1.054	1.200
Energy	0.228	0.208	**0.208**	0.215	0.284	0.266
Environment	0.315	**0.306**	0.309	0.314	0.332	0.333

Table 4: Hyper-parameter sensitivity analysis about the number of selected experts K.

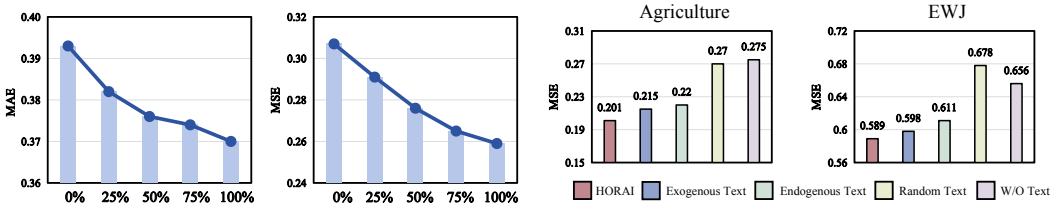
	K=1	**K=2**	**K=3**	**K=4**
Metrics	MSE	MSE	MSE	MSE
Agriculture	0.227	**0.201**	0.206	0.220
Climate	1.048	**0.857**	0.880	0.891
Energy	0.217	0.208	**0.200**	0.215
Environment	0.321	**0.309**	0.316	0.326

4.4 Model Analysis

Due to the limitation of space, we provide the analysis of specific modalities and alignment strategies, and experiments on replacing different text and vision encoders in Appendices F.4 and F.5.

Fine-tune with downstream data. To examine how the amount of fine-tuning data affects downstream performance, we evaluate HORAI by progressively enlarging the training portion of the Environment dataset. As shown in Figure 5 (a), the forecasting accuracy steadily improves as more data is used, reaching its best with the full dataset. Specifically, the MAE decreases from 0.393 to 0.370, and the MSE decreases from 0.309 to 0.259. These results highlight HORAI's strong adaptability to downstream data availability.

Text Ablation. To examine whether HORAI genuinely leverages semantic information from two types of text, we conduct a series of text ablation experiments with four variants: randomly generated unrelated text (Random Text), exogenous text only (Exogenous Text), endogenous text only (Endogenous Text), and removing the entire text modality (W/O Text). As shown in Figure 5(b), the full HORAI model consistently achieves the lowest MSE across both datasets, demonstrating the synergistic benefit of integrating endogenous descriptions with exogenous real-world news. Notably, utilizing either Exogenous Text or Endogenous Text independently yields better performance than W/O Text, confirming that both distinct sources provide valuable information. Introducing random text leads to a substantial performance drop, even worse than removing the text modality on the EWJ dataset, indicating that HORAI does not simply rely on the presence of text but actually understands and exploits its semantic content.



(a) Fine-tune HORAI with differentpercentages

(b) Text Replacement

Figure 5: (a) Fine-tuning HORAI with different data percentages on the Environment dataset. (b) Text replacement experiments on the Agriculture dataset and the EWJ dataset.

5 Conclusion

In this paper, we first propose a multimodal pretraining paradigm that leverages time series with endogenous images and text and exogenous news, providing a comprehensive multi-view perspective for time series analysis. To support this, we develop an automated data construction pipeline to curate MM-TS, the large-scale multimodal time series dataset spanning six domains. Then we propose HORAI, a frequency-enhanced multimodal foundation model. It integrates two core components: the Frequency-guided Cross-Modality Encoder and the Time-Frequency Decoder, effectively fusing different multimodal features and enhancing model generalization. After pre-training on MM-TS, HORAI achieves SOTA performance in time series forecasting and anomaly detection tasks, which demonstrates strong task versatility and generalization ability.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (62406112, 62372179).

References

Jindong Tian, Yuxuan Liang, Ronghui Xu, Peng Chen, Chenjuan Guo, Aoying Zhou, Lujia Pan, Zhongwen Rao, and Bin Yang. Air quality prediction with physics-informed dual neural odes in open systems. *CoRR*, abs/2410.19892, 2024.

Jindong Tian, Yifei Ding, Ronghui Xu, Hao Miao, Chenjuan Guo, and Bin Yang. Arrow: An adaptive rollout and routing method for global weather forecasting. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Jieyuan Mei, Jindong Tian, Ronghui Xu, Hanyue Wei, Chenjuan Guo, and Bin Yang. Find3: A dual 3d state space model with dynamic hypergraph for financial stock prediction. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 2084–2094, 2025.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *ICLR*, 2023.

Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. DUET: Dual clustering enhanced multivariate time series forecasting. In *SIGKDD*, pages 1185–1196, 2025a.

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 *ICLR*, 2024a.

Xingjian Wu, Xiangfei Qiu, Zhengyu Li, Yihang Wang, Jilin Hu, Chenjuan Guo, Hui Xiong, and Bin Yang. CATCH: Channel-aware multivariate time series anomaly detection via frequency patching. In *ICLR*, 2025a.

Peng Chen, Yingying Zhang, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang, and Chenjuan Guo. Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting. In *ICLR*, 2024a.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *ICML*, 2024.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-moe: Billion-scale time series foundation models with mixture of experts. In *ICLR*, 2025.

Xingjian Wu, Hanyin Cheng, Xiangfei Qiu, Zhengyu Li, Jilin Hu, Chenjuan Guo, and Bin Yang. Flame: Flow enhanced legendre memory models for general time series forecasting. *arXiv preprint arXiv:2512.14253*, 2025b.

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal LLM. In *ICML*, 2024.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024b.

Furong Jia, Kevin Wang, Yixiang Zheng, Defu Cao, and Yan Liu. Gpt4mts: Prompt-based large language model for multimodal time-series forecasting. In *AAAI*, volume 38, pages 23343–23351, 2024.

Zihao Li, Xiao Lin, Zhining Liu, Jiaru Zou, Ziwei Wu, Lecheng Zheng, Dongqi Fu, Yada Zhu, Hendrik Hamann, Hanghang Tong, et al. Language in the flow of time: Time-series-paired texts weaved into a unified temporal narrative. *arXiv preprint arXiv:2502.08942*, 2025.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. In *ICML*, 2024b.

Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qingsong Wen, and Yuxuan Liang. Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting. In *ICML*, 2025.

Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *ICML*, volume 97, pages 5142–5151, 2019.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. *CoRR*, abs/1606.07947, 2016.

Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, Zhenli Sheng, and Bin Yang. TFB: towards comprehensive and fair benchmarking of time series forecasting methods. *Proc. VLDB Endow.*, 17(9):2363–2377, 2024.

Peng Chen, Zhihao Zhuang, Hongzhou Chen, Junhao Huang, Aiping Yang, Mengsen Wu, Yiding Liu, Xilin Dai, and Zewei Dong. Beyond numerical time series: A unified benchmark for multimodal forecasting with heterogeneous context. *CoRR*, abs/2609.15087, 2026.

Zahra Zamanzadeh Darban, Geoffrey I. Webb, Shirui Pan, Charu Aggarwal, and Mahsa Salehi. Deep learning for time series anomaly detection: A survey. *ACM Comput. Surv.*, 57(1):15:1–15:42, 2025.

Xiangfei Qiu, Zhe Li, Wanghui Qiu, Shiyan Hu, Lekui Zhou, Xingjian Wu, Zhengyu Li, Chenjuan Guo, Aoying Zhou, Zhenli Sheng, Jilin Hu, Christian S. Jensen, and Bin Yang. TAB: Unified benchmarking of time series anomaly detection methods. In *Proc. VLDB Endow.*, pages 2775–2789, 2025b.

John Paparrizos, Yuhao Kang, Paul Boniol, Ruey S Tsay, Themis Palpanas, and Michael J Franklin. Tsb-uad: an end-to-end benchmark suite for univariate time-series anomaly detection. *VLDB*, 15 (8):1697–1711, 2022a.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *AAAI*, pages 11121–11128, 2023.

Zhijian Xu, Ailing Zeng, and Qiang Xu. FITS: modeling time series with 10k parameters. In *ICLR*, 2024b.

Valentin Flunkert, David Salinas, and Jan Gasthaus. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *CoRR*, abs/1704.04110, 2017.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *ICLR*, 2023.

Hanyin Cheng, Xingjian Wu, Xiangfei Qiu, Yang Shu, Bin Yang, and Chenjuan Guo. CCD: Capturing cross-correlations with deformable convolutional networks for multivariate time series forecasting. In *KDD*, 2026a.

Xinle Wu, Dalin Zhang, Chenjuan Guo, Chaoyang He, Bin Yang, and Christian S. Jensen. Autocts: Automated correlated time series forecasting. *VLDB*, 15(4):971–983, 2021.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *ICLR*, 2023.

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 *ICLR*, 2024.

Peng Chen, Yihang Wang, Yang Shu, Yunyao Cheng, Kai Zhao, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. Cc-time: Cross-model and cross-modality time series forecasting. *arXiv preprint arXiv:2508.12235*, 2025.

Zijie Pan, Yushan Jiang, Sahil Garg, Anderson Schneider, Yuriy Nevmyvaka, and Dongjin Song. s^2 ip-llm: Semantic space informed prompt learning with llm for time series forecasting. *arXiv preprint arXiv:2403.05798*, 2024.

Shiyan Hu, Jianxin Jin, Yang Shu, Peng Chen, Bin Yang, and Chenjuan Guo. Towards multimodal time series anomaly detection with semantic alignment and condensed interaction. In *ICLR*, 2026.

Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Prabhakar Kamarthi, Aditya Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, et al. Time-mmd: Multi-domain multimodal dataset for time series analysis. *NeurIPS*, 37:77888–77933, 2024c.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *ICML*, 2024.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *ICML*, 2024.

Vijay Ekambaram, Arindam Jati, Pankaj Dayama, Sumanta Mukherjee, Nam Nguyen, Wesley M. Gifford, Chandra Reddy, and Jayant Kalagnanam. Tiny time mixers (ttms): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series. In *NeurIPS*, 2024.

Patrick Podest, Marco Pichler, Elias Bürger, Levente Zólyomi, Bernhard Voggenberger, Wilhelm Berghammer, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. Tirex-2: Generalizing tirex to multivariate data and streaming. *CoRR*, abs/2607.01204, 2026.

Hanyin Cheng, Xingjian Wu, Yang Shu, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. CoRA: Boosting time series foundation models for multivariate forecasting through correlation-aware adapter. In *ICLR*, 2026b.

Hanyin Cheng, Ruitong Zhang, Yuning Lu, Peng Chen, Meng Wang, Yang Shu, Bin Yang, and Chenjuan Guo. STAR: Boosting time series foundation models for anomaly detection through state-aware adapter. In *NeurIPS*, 2026c.

Yihang Wang, Yuying Qiu, Peng Chen, Kai Zhao, Yang Shu, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. Towards a general time series forecasting model with unified representation and adaptive transfer. In *ICML*, 2025a.

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*, pages 12694–12702, 2025b.

Xingjian Wu, Jianxin Jin, Wanghui Qiu, Peng Chen, Yang Shu, Bin Yang, and Chenjuan Guo. Aurora: Towards universal generative multimodal time series forecasting. *CoRR*, abs/2509.22295, 2025c.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *ICLR*, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Haixu Wu, Jialong Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Flowformer: Linearizing transformers with conservation flows. *arXiv preprint arXiv:2202.06258*, 2022.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. *CoRR*, abs/2502.00816, 2025.

Mouxiang Chen, Lefei Shen, Zhuo Li, Xiaoyun Joy Wang, Jianling Sun, and Chenghao Liu. Visionts: Visual masked autoencoders are free-lunch zero-shot time series forecasters. *arXiv preprint arXiv:2408.17253*, 2024c.

Tian Zhou, Peisong Niu, Xue Wang, Liang Sun, and Rong Jin. One fits all: Power general time series analysis by pretrained LM. In *NeurIPS*, 2023.

Qichao Shentu, Beibu Li, Kai Zhao, Yang Shu, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. Towards a general time series anomaly detector with adaptive bottlenecks and dual adversarial decoders. In *ICLR*, 2025.

Shanghua Gao, Teddy Koker, Owen Queen, Tom Hartvigsen, Theodoros Tsiligkaridis, and Marinka Zitnik. Units: A unified multi-task time series model. In *NeurIPS*, 2024.

Md Kowsher, Md Shohanur Islam Sobuj, Nusrat Jahan Prottasha, E Alejandro Alanis, Ozlem Ozmen Garibay, and Niloofar Yousefi. Llm-mixer: Multiscale mixing in llms for time series forecasting. *arXiv preprint arXiv:2410.11674*, 2024.

Yiyuan Yang, Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun. Dcdetector: Dual attention contrastive representation learning for time series anomaly detection. In *SIGKDD*, pages 3033–3045, 2023.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. In *ICLR*, 2022.

Markus Goldstein and Andreas Dengel. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *KI-2012: poster and demo track*, 1:59–63, 2012.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *ICDM*, pages 413–422, 2008.

Mei-Ling Shyu, Shu-Ching Chen, Kanoksri Sarinnapakorn, and LiWu Chang. A novel anomaly detection scheme based on principal component classifier. 2003.

John Paparrizos, Paul Boniol, Themis Palpanas, Ruey S Tsay, Aaron Elmore, and Michael J Franklin. Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection. *Proceedings of the VLDB Endowment*, 15(11):2774–2787, 2022b.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I Webb, Rob J Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. *arXiv preprint arXiv:2105.06643*, 2021.

Alexander Alexandrov, Konstantinos Benidis, Michael Bohlke-Schneider, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Danielle C. Maddix, Syama Sundar Rangapuram, David Salinas, Jasper Schulz, Lorenzo Stella, Ali Caner Türkmen, and Yuyang Wang. Gluonts: Probabilistic and neural time series modeling in python. *J. Mach. Learn. Res.*, 21:116:1–116:6, 2020.

Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. The ucr time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305, 2019.

Shuyi Zhang, Bin Guo, Anlan Dong, Jing He, Ziping Xu, and Song Xi Chen. Cautionary tales on air-quality improvement in beijing. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2205):20170457, 2017.

Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. The uea multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075*, 2018.

Yihe Wang, Yu Han, Haishuai Wang, and Xiang Zhang. Contrast everything: A hierarchical contrastive framework for medical time-series. *NeurIPS*, 36, 2024.

Minhao Liu, Ailing Zeng, Muxi Chen, Zhijian Xu, Qiuxia Lai, Lingna Ma, and Qiang Xu. Scinet: Time series modeling and forecasting with sample convolution and interaction. *NeurIPS*, 35: 5816–5828, 2022.

Jingyuan Wang, Jiawei Jiang, Wenjun Jiang, Chengkai Han, and Wayne Xin Zhao. Towards efficient and comprehensive urban spatial-temporal prediction: A unified library and performance benchmark. *CoRR*, abs/2304.14343, 2023.

Michael W McCracken and Serena Ng. Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589, 2016.

Souhaib Ben Taieb, Gianluca Bontempi, Amir F Atiya, and Antti Sorjamaa. A review and comparison of strategies for multi-step ahead time series forecasting based on the nn5 forecasting competition. *Expert systems with applications*, 39(8):7067–7083, 2012.

Zihan Dong, Xinyu Fan, and Zhiyuan Peng. FNSPID: A comprehensive financial news dataset in time series. In Ricardo Baeza-Yates and Francesco Bonchi, editors, *SIGKDD*, pages 4918–4927, 2024.

Zhe Xie, Zeyan Li, Xiao He, Longlong Xu, Xidao Wen, Tieying Zhang, Jianjun Chen, Rui Shi, and Dan Pei. Chatts: Aligning time series with llms via synthetic data for enhanced understanding and reasoning. *Proc. VLDB Endow.*, 18(8):2385–2398, 2025.

A Datasets

A.1 Pre-train Dataset MM-TS of Time Series Modality

For time series modality, we assemble a large and diverse set of publicly available time series datasets covering domains such as energy, nature, transportation, web, health, and economics. The corpus contains around 1 billion time points, with a strict separation from all target evaluation datasets. The datasets vary widely in their sampling frequencies, from millisecond-level measurements to monthly observations, reflecting both the heterogeneity of real-world scenarios and the complexity of temporal dynamics.

Pretraining–Evaluation Overlap. We examine potential overlaps between MM-TS and the downstream forecasting datasets. We identify one related-series overlap: the Social Good benchmark uses UNRATENSA, while MM-TS contains UNRATE. Although these series are not numerically identical, they are seasonally unadjusted and adjusted versions of the same underlying unemployment measure, with overlapping temporal coverage. For the remaining eight forecasting datasets, our audit of series identifiers and temporal coverage reveals no analogous overlap.

Table 5: List of pretraining datasets of time series modality.

Domain	Dataset	Frequency	Time Pionts	Source
Energy	Aus. Electricity Demand	Half Hourly	1155264	Monash [Godahewa et al., 2021]
	Wind	4 Seconds	7397147	Monash [Godahewa et al., 2021]
	Wind Farms	Minutely	172178060	Monash [Godahewa et al., 2021]
	Solar Power	4 Seconds	7397222	Monash [Godahewa et al., 2021]
	London Smart Meters	Half Hourly	166527216	Monash [Godahewa et al., 2021]
	BDG-2 Rat	Hourly	4596080	[Alexandrov et al., 2020]
	BDG-2 Panther	Hourly	893840	[Alexandrov et al., 2020]
	BDG-2 Fox	Hourly	2285288	[Alexandrov et al., 2020]
Nature	Phoneme	-	2160640	UCRDau et al. [2019]
	PRSA	Hourly	4628448	[Zhang et al., 2017]
	Temperature Rain	Daily	23252200	Monash [Godahewa et al., 2021]
	StarLightCurves	-	9457664	UCR [Dau et al., 2019]
	Worms	0.033 Seconds	232200	UCR [Dau et al., 2019]
	Saugeen River Flow	Daily	23741	Monash [Godahewa et al., 2021]
	Sunspot	Daily	73924	Monash [Godahewa et al., 2021]
	Weather	Daily	43032000	Monash [Godahewa et al., 2021]
	KDD Cup 2018	Daily	2942364	MonashGodahewa et al. [2021]
	US Births	Daily	7305	Monash [Godahewa et al., 2021]
Healthcare	MotorImagery	0.001 Seconds	72576000	UEA [Bagnall et al., 2018]
	AtrialFibrillation	0.008 Seconds	38400	UEA [Bagnall et al., 2018]
	PigArtPressure	-	624000	UCR [Dau et al., 2019]
	TDbrain	0.002 Seconds	79232703	[Wang et al., 2024]
Transport	Pems03	5 Minute	9382464	[Liu et al., 2022]
	Pems08	5 Minute	3035520	[Liu et al., 2022]
	Pems-bay	5 Minute	16937700	[Liu et al., 2022]
	Pedestrian_Counts	Hourly	3132346	Monash [Godahewa et al., 2021]
	SZ-Taxi	15 Minute	464256	[Wang et al., 2023]
	Taxi	Half Hourly	40584636	[Alexandrov et al., 2020]
	Uber TLC	Hourly	510284	[Alexandrov et al., 2020]
Web	Web Traffic	Daily	116485589	Monash [Godahewa et al., 2021]
Economic	FRED_MD	Monthly	77896	[McCracken and Ng, 2016]
	Bitcoin	Daily	75364	Monash [Godahewa et al., 2021]
	NN5	Daily	87801	[Taieb et al., 2012]

A.2 Pre-train Dataset MM-TS of Text Modality

To address the scarcity of high-quality, semantically aligned time series and text pairs, we design a two-stage automated data construction pipeline. The pipeline uniquely integrates endogenous text (time

series patterns description) with exogenous real-world news events through a "Generate-then-Filter" paradigm.

Stage 1: Contextual Synthesis Stage. We adopt a parallel generation strategy to synthesize the textual modality:

- Endogenous Text Generation: An LLM Analyzer GPT-4o is employed to interpret the numerical time series. As shown in Figure 6, guided by specific prompts targeting statistical properties (e.g., trend, seasonality, and stationarity), the analyzer converts raw temporal dynamics into structured Endogenous Text descriptions.
- Exogenous News Retrieval: Simultaneously, we extract metadata constraints—specifically Domain, Time Horizon, and Region—to formulate search queries. We utilize a Relevance Retrieval to fetch the relevant news events from the GDELT Database. These raw articles are then summarized by GPT-4o to form concise Exogenous News, providing the necessary background context for the time series. Specific exogenous news summarization prompt is in Figure 7.

Stage 2: Quality Alignment Stage. To ensure the reliability and logical consistency of the synthesized data, the raw generations undergo a rigorous quality control process:

- Logical Consistency Check: A dedicated LLM GPT-4o evaluates the alignment between the Endogenous Text and Exogenous News. This step filters out hallucinations or irrelevant pairings by verifying, for example, that a significant market dip is explained by a corresponding negative news event rather than a positive one. Specific logical consistency filtering prompt is in Figure 8.
- Ensemble Quality Evaluation: To mitigate the bias inherent in single-model evaluation, the unified text is scrutinized by a heterogeneous ensemble of LLM judges. We employ a consensus-based filtering mechanism, where only samples achieving high aggregate scores on factual plausibility and semantic clarity are incorporated into the final dataset.

Endogenous Text Generation Prompt

Role：
You are a domain expert in time series analysis, specializing in interpreting numerical {**energy domain**} data and translating it into structured, semantically meaningful textual descriptions.

Task：
Analyze the provided univariate time series and generate a concise yet informative endogenous information summary that describes the intrinsic temporal patterns of the data.

Guidelines：
1.Pattern Identification: Focus exclusively on statistical characteristics of the time series, including overall trends, seasonal or cyclic behaviors, and general fluctuations.
2.Objective Tone: Use precise, professional, and neutral language, avoiding external assumptions or speculative interpretations.
3.Output Constraint: Produce a single paragraph that summarizes the temporal dynamics. Your description should begin with: **"The {London Smart Meters} series exhibits …"**

Dataset Description：
The dataset is sourced from the {**London Smart Meters**} project within the {**energy domain**}, with data collected at {**30-minute intervals**}.

Input Data：
[time series data]

Figure 6: The prompt is designed for generating endogenous descriptions of the London Smart Meters dataset.

A.3 Evaluation Dataset

To evaluate HORAI in a multi-task setting, we employ widely used benchmark datasets for both forecasting and anomaly detection. 1) Forecasting: As shown in Table 6, experiments are conducted on

Exogenous News Summarization Prompt

Role:
You are an expert Data Scientist specializing in time series analysis and macroeconomic forecasting. Your goal is to convert fragmented news headlines into a structured "exogenous information summary" .

Task:
Analyze the provided list of news titles and synthesize them into a single, cohesive summary of the external environment.

Guidelines:
1. **Denoising**: Ignore trivial, entertainment, or irrelevant news. Focus on events that could impact economic, social, or environmental variables (e.g., policy changes, extreme weather, strikes, energy crises, or major holidays).
2. **Synthesis**: If multiple titles refer to the same event, merge them into a single comprehensive description.
3. **Objective Tone**: Use professional and neutral language. Avoid emotional bias.
4. **Output Constraint**: Provide ONLY the "Exogenous Information Summary".

Input Data:
[Paste your list of news titles here]

Figure 7: The prompt is designed for summarizing the relevant exogenous news.

Logical Consistency Filtering Prompt

Role:
You are an expert Time Series Analyst and Logic Auditor. Your task is to rigorously evaluate the causal plausibility and logical consistency between endogenous time series patterns description and exogenous news summary.

Task:
You will be provided with two text inputs: Endogenous Description: A textual description of the time series‘ data patterns (e.g., trends, seasonality). Exogenous News Summary: A summary of real-world events occurring during the same time of the domain.
Your Goal: Determine if the Exogenous News logically explains or aligns with the Endogenous Description.

Guidelines:
1.Directional Alignment Check: Ensure the news sentiment aligns with the data trajectory: associate rises with positive factors (growth/holidays), declines with negative shocks (crises/outages), and stability with the absence of disruptive events.
2.Magnitude Consistency: A minor data fluctuation should not be paired with a catastrophic global event, and a massive anomaly should not be explained by trivial news.
3.Reject Contradictions: If the news explicitly contradicts the physical reality of the data (e.g., "Electricity usage dropped to zero" vs. "Heatwave causes record power demand"), mark as Inconsistent.

Input Data:
Endogenous Description: [endogenous_description]
Exogenous News Summary: [exogenous_summary]

Output Format:
"consistency_score": An integer from 1 (Complete Contradiction) to 5 (Perfect Alignment).
"is_consistent": Boolean (true if score >= 3, else false).

Figure 8: The prompt is designed for judgment logical consistency for the endogenous description and exogenous news.

TimeMMD [Liu et al., 2024c] and additional datasets, covering diverse domains such as Agriculture, Climate, Energy, Environment, Social Good, Traffic, EWJ, KR, and MDT. 2) Anomaly Detection: We evaluate HORAI on five datasets—Weather, Energy, KR, EWJ, and MDT—with anomaly ratios ranging from 5.81% to 17.23%. Detailed statistics are provided in Table 7.

Table 6: The statistics of evaluation datasets for the forecasting task.

Tasks	Dataset	Variate	Frequency	Dataset Size	Timespan
Forecasting	Agriculture	1	Monthly	496	1983-2024
	Climate	5	Monthly	496	1983-2024
	Energy	9	Weekly	1479	1996-2024
	Environment	4	Daily	11102	1982-2023
	Social Good	1	Monthly	900	1950-2024
	Traffic	1	Monthly	531	1980-2024
	EWJ	1	Daily	2658	2009-2020
	KR	1	Daily	2655	2009-2020
	MDT	1	Daily	2732	2009-2020

Table 7: The statistics of evaluation datasets for the anomaly detection task.

Tasks	Dataset	Anomaly Ratio	Frequency	Dataset Description
Detection	Weather	17.10%	Monthly	Temperature and humidity information collected from government websites.
	Energy	17.23%	Weekly	The dataset records weekly U.S. gasoline prices (dollars per gallon).
	KR	6.21%	Daily	The dataset is collected from Yahoo, NASDAQ finance websites.
	MDT	11.17%	Monthly	The dataset is collected from Yahoo, NASDAQ finance websites.
	EWJ	9.96%	Daily	The dataset is collected from Yahoo, NASDAQ finance websites.

B Baselines

We categorize the baselines into three groups: *Time Series Foundation Models*, *Multimodal Time-Series-Specific Models*, and *Unimodal Time-Series-Specific Models*. Time Series Foundation Models are pre-trained on large-scale, cross-domain unimodal time series data, enabling direct inference on downstream tasks and demonstrating generalization capabilities. In contrast, Time-Series-Specific Models require training on each downstream dataset and can be further divided by input type. Multimodal Time-Series-Specific Models leverage additional modalities, such as text or images, or reuse LLM representations to enhance time series understanding. Unimodal Time-Series-Specific Models, on the other hand, design tailored modules to exploit the inherent characteristics of time series data.

B.1 Time Series Foundation Models

- Chronos2 [Ansari et al., 2025] is a universal time series forecasting model that extends the original Chronos from univariate to multivariate prediction through unified tokenization.
- TimesFM [Das et al., 2024] is a decoder-only time series foundation model pre-trained on large-scale real-world and synthetic time series data, enabling zero-shot forecasting performance across diverse domains, horizons, and temporal granularities.
- ChatTime [Wang et al., 2025b] proposes a unified multimodal time series foundation model that tokenizes both numerical sequences and text prompts into a shared vocabulary, enabling forecasting and question answering tasks with a single generative language-model backbone.
- Sundial [Liu et al., 2025] proposes a TimeFlow Loss that predicts the distribution of the next patch, enabling Transformer training without discrete tokenization and supporting probabilistic forecasting.
- VisionTS [Chen et al., 2024c] converts time series data into image form and uses visual mask autoencoders for unsupervised feature learning.
- ROSE [Wang et al., 2025a] combines frequency decomposition with time-series registers to jointly learn both domain-invariant and domain-specific representations, facilitating knowledge transfer to downstream tasks.

- Timer [Liu et al., 2024b] adopts a decoder-only architecture employing autoregressive modeling for generative pre-training.
- MOIRAI [Woo et al., 2024] introduces multi-scale patch projections to model diverse patterns and an any-variate attention mechanism that allows flexible handling of time series with arbitrary dimensionality.
- DADA [Shentu et al., 2025] leverages adaptive bottleneck and dual-adversarial decoding to enable robust zero-shot anomaly detection across diverse domains.
- UniTS [Gao et al., 2024] proposes a novel unified network backbone for classification, forecasting, and anomaly detection.

B.2 Multimodal Time-Series-Specific Models

- GPT4MTS [Jia et al., 2024] propose a prompt tuning-based LLM for time series forecasting with multimodal input.
- TATS [Li et al., 2025] propose a plug-and-play multimodal time series forecasting framework, which transforms text representations into auxiliary variables.
- GPT4TS [Zhou et al., 2023] fine-tunes the limited parameters of LLM, demonstrating competitive performance by transferring knowledge from large-scale pre-training text data.
- LLMMixer [Kowsher et al., 2024] adapts LLMs for time series forecasting by breaking down the data into different time scales.
- TimeVLM [Zhong et al., 2025] leverages pre-trained VLMs to enhance time series forecasting by unifying temporal, visual, and textual information.

B.3 Unimodal Time-Series-Sepcific Models

- TimesNet [Wu et al., 2023] transforms the 1D time series into a set of 2D tensors based on multiple periods to handle the multi-periodicity of the time series.
- DCdetector [Yang et al., 2023] leverages dual-attention contrastive representation learning, extracting normal feature representations through self-supervised learning and dual-attention mechanisms.
- Anomaly Transformer [Xu et al., 2022] leverages a self-attention mechanism to capture both short- and long-term dependencies in time series, and detects anomalies by analyzing differences in association matrices.
- PatchTST [Nie et al., 2023] segments time series into subseries-level patches that serve as input tokens to the Transformer and applies the channel-independence strategy for training on multivariate time series.
- HBOS [Goldstein and Dengel, 2012] is a fast unsupervised anomaly detection method based on histogram density estimation.
- IForest [Liu et al., 2008] detects anomalies by recursively partitioning data to isolate outliers, rather than modeling normal behavior.
- PCA [Shyu et al., 2003] detects anomalies by measuring deviations in the principal component space, assuming outliers lie far from the normal distribution.

C Experiment Setting

During pre-training, HORAI is optimized using the Adam optimizer with an initial learning rate of 0.0005 and trained for 20 epochs, with early stopping applied using a patience of 5 epochs. The batch size is set to 2048, the input time series length to 576, and the patch size to 48. The Time-Frequency Decoder is configured with 6 layers, the model dimension D_{model} is set to 768, and the ratio parameter α for high- and low-frequency decomposition is fixed at 0.05. All experiments are implemented in PyTorch, and pre-training is conducted on four NVIDIA Tesla A800 80GB GPUs.

For forecasting, To ensure fairness, we remove the drop-last strategy for HORAI and all baselines, since using it would result in inconsistent numbers of test samples across different batch sizes Qiu et al. [2024]. For each dataset, we evaluate four prediction horizons for both HORAI and the baselines.

Specifically, Agriculture, Climate, Social Good, Traffic, EWJ, KR, and MDT are evaluated with horizons $\{6, 8, 10, 12\}$, Environment with $\{48, 96, 192, 336\}$, and Energy with $\{12, 24, 36, 48\}$.

D Discussion with ChatTime

We further provide a clarified discussion comparing ChatTime and HORAI, including the following aspects: (1) **Model perspective**: HORAI is **specifically architected as a multimodal foundation model** integrating time series, images, and text. It leverages modality-specific encoders to extract distinct features and employs a novel frequency-enhanced alignment to fuse these representations from multiple perspectives explicitly. In contrast, ChatTime **adapts general-purpose LLMs** for time series analysis. While leveraging LLMs' inherent reasoning abilities for time series analysis offers generalization, discretizing continuous numerical values into textual tokens results in precision loss, making it challenging to capture time series patterns. (2) **Data perspective**: HORAI is pretrained on a large-scale multimodal dataset incorporating aligned text and images. These modalities capture diverse characterizations of temporal dynamics from multiple perspectives and simultaneously introduce some external context, providing relevant supervision that improves generalization. However, ChatTime relies only on simple prompts such as "Please predict the following sequence," which offer limited text regarding the specific time series characteristics.

E Performance on FNSPID

To further validate HORAI, we select five stock datasets from FNSPID [Dong et al., 2024] for additional comparison, including FCX, AVGO, FCAU, CSX, and GILD. The Table 8 demonstrates that HORAI consistently achieves SOTA forecasting performance on the RMSE and MAE metrics.

Table 8: Model performance comparison on the FNSPID benchmark.

	HORAI		**ChatTime**		**VisionTS**		**MOIRAI**		**TimesFM**		**Chronos2**		**TATS**		**GPT4MTS**		**GPT4TS**		**TimeVLM**	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
FCX	**0.062**	0.051	0.065	0.051	0.074	0.057	0.072	0.065	0.065	0.057	0.062	0.054	0.065	**0.048**	0.068	0.050	0.073	0.057	0.064	0.049
AVGO	**0.222**	0.159	0.253	0.174	0.254	0.164	0.243	0.168	0.228	0.167	0.235	0.162	0.225	0.154	0.224	**0.150**	0.244	0.158	0.235	0.154
FCAU	0.182	**0.123**	0.192	0.127	0.210	0.138	**0.176**	0.156	0.189	0.130	0.184	**0.123**	0.180	0.124	0.179	0.125	0.178	0.133	0.180	0.138
CSX	0.440	**0.282**	0.455	0.324	0.485	0.320	0.396	0.349	0.442	0.300	**0.423**	0.291	0.435	0.284	0.436	0.285	0.477	0.310	0.431	0.289
GILD	**0.080**	**0.061**	0.106	0.075	0.083	0.064	0.084	0.073	0.080	0.069	0.085	0.068	0.085	0.061	0.084	0.062	0.085	0.062	0.088	0.064

F Model Analysis

F.1 Compare with Unimodal Time Series Specific Models

We compare HORAI with four strong unimodal task-specific models: DLinear, ModernTCN, PatchTST, and TimeMixer. These baselines are trained on the full downstream training splits, whereas HORAI is evaluated zero-shot. As shown in Table 9, HORAI achieves the lowest MSE on eight of the nine datasets. On these eight datasets, its relative MSE reduction over the strongest task-specific baseline ranges from 1.6% to 18.4%. DLinear performs slightly better on EWJ, with an MSE of 0.582 compared with HORAI's 0.589. These results support HORAI's transfer capability without downstream fine-tuning, including when compared with competitive numerical-only forecasting models.

F.2 Comparison with multimodal foundation models and LLM-based reasoning models

Table 10 compares HORAI with Aurora [Wu et al., 2025c], ChatTS-8B [Xie et al., 2025], and Gemma-4-26B-A4B-it under zero-shot evaluation. HORAI achieves lower MSE than ChatTS-8B and Gemma on all nine datasets and outperforms Aurora on most datasets, although Aurora performs better on Environment. These comparisons extend the evaluation to both pretrained multimodal forecasting models and a general-purpose multimodal LLM. We also include an equal-weight average of the numerical forecasts from Chronos-2 and Gemma to examine whether a simple prediction-level combination is competitive with HORAI's learned multimodal fusion.

Table 9: Comparison with time-series-specific forecasting models. The best results are highlighted in **bold**, and the second-best results are underlined.

Dataset	HORAI (zero-shot)	DLinear	ModernTCN	PatchTST	TimeMixer
Metric	MSE	MSE	MSE	MSE	MSE
Agriculture	**0.201**	0.515	0.218	0.228	0.230
Climate	**0.857**	1.050	1.275	1.184	1.159
Energy	**0.208**	0.247	0.235	0.250	0.240
Environment	**0.309**	0.520	0.322	0.317	0.325
Social Good	**0.819**	0.855	0.957	0.944	1.093
Traffic	**0.165**	0.237	0.211	0.176	0.212
EWJ	0.589	**0.582**	0.607	0.603	0.604
KR	**0.551**	0.562	0.584	0.575	0.578
MDT	**0.376**	0.382	0.393	0.387	0.386

Table 10: Comparison with multimodal foundation models and LLM-based reasoning models.

Dataset	HORAI (Zero-shot)	Aurora (Zero-shot)	ChatTS-8B (Zero-shot)	Gemma-4-26B-A4B-it (Zero-shot)	Chronos-2+Gemma-4-26B-A4B-it (Average)
Metric	**MSE**	**MSE**	**MSE**	**MSE**	**MSE**
Agriculture	0.201	0.272	0.210	0.249	**0.193**
Climate	**0.857**	0.865	2.058	1.629	1.335
Energy	**0.208**	0.255	0.238	0.474	0.293
Environment	0.309	**0.276**	0.560	0.485	0.428
Social Good	**0.819**	0.838	1.165	0.973	0.845
Traffic	**0.165**	0.161	0.442	0.885	0.623
EWJ	**0.589**	0.605	0.812	0.936	0.755
KR	**0.551**	0.578	0.858	0.799	0.682
MDT	**0.376**	0.396	0.572	0.659	0.458

F.3 Continued Pretraining on MM-TS and Full-Shot TSFM Comparisons

To examine whether a strong numerical-only foundation model can benefit similarly from MM-TS, we continue pretraining Chronos-2 on its numerical time-series portion and evaluate the resulting checkpoint both zero-shot and after full-shot downstream fine-tuning. Table 11 shows that HORAI achieves lower zero-shot MSE on all nine datasets, with a mean relative reduction of 20.4%. Downstream fine-tuning improves Chronos-2 on every dataset, but HORAI remains better on eight of nine, with Agriculture as the exception. HORAI also outperforms full-shot UniTS, ROSE, Timer, and MOIRAI on all nine datasets. Thus, HORAI remains competitive without downstream adaptation even against these fine-tuned models. The Chronos-2 comparison controls for access to the numerical portion of MM-TS, but not for multimodal supervision or the models' original pretraining histories.

Table 11: Forecasting results of HORAI and baseline methods. Lower MSE is better, and the best result in each row is highlighted in **bold**.

Dataset	HORAI (MM-TS + Zero-Shot)	Chronos2 (MM-TS + Zero-Shot)	Chronos2 (MM-TS + Full-Shot)	UniTS (Full-Shot)	ROSE (Full-Shot)	Timer (Full-Shot)	MOIRAI (Full-Shot)
Metric	MSE	MSE	MSE	MSE	MSE	MSE	MSE
Agriculture	0.201	0.210	**0.196**	0.311	0.246	0.222	0.303
Climate	**0.857**	1.009	0.992	1.242	1.183	1.201	1.028
Energy	**0.208**	0.268	0.240	0.306	0.246	0.274	0.256
Environment	**0.309**	0.347	0.326	0.329	0.337	0.349	0.342
Social Good	**0.819**	1.044	0.955	1.012	0.984	1.166	0.815
Traffic	**0.165**	0.372	0.217	0.281	0.184	0.199	0.187
EWJ	**0.589**	0.677	0.609	0.612	0.609	0.609	0.598
KR	**0.551**	0.752	0.581	0.591	0.556	0.566	0.566
MDT	**0.376**	0.438	0.386	0.397	0.388	0.387	0.384

F.4 Ablation Analysis on Specific Modalities and Alignment Strategies

We perform ablation studies to evaluate the contributions of individual modalities (text, image) and the efficacy of our frequency-based alignment strategy. Specifically, we analyze four settings: 1) only text and time series; 2) only image and time series ; 3) text, image, and time series without

frequency-based alignment (w/o Freq-Align); and 4) swapping modalities by fusing low-frequency time series with images and mid-to-high frequency time series with text (Modality Exchange). As shown in Table 12, both visual and textual modalities contribute to performance gains, though their relative impact varies depending on the dataset characteristics. For datasets exhibiting clear long-term trends, such as Agriculture and Energy, the text modality contributes more significantly. Conversely, for datasets dominated by local fluctuations, such as Climate, the image modality proves more critical. Crucially, the significant performance drop observed when removing frequency-based alignment and modality exchange underscores the validity of our design: it confirms that aligning images with mid-to-high frequency components and text with low-frequency components is the effective strategy.

Table 12: Ablation analysis about each modality.

	HORAI		**Text + Time Series**		**Image + Time Series**		**W/O Freq-Align**		**Modality Exchange**	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Agriculture	**0.201**	**0.309**	0.248	0.327	0.275	0.340	0.234	0.324	0.292	0.352
Climate	**0.857**	**0.734**	1.102	0.828	0.982	0.797	0.928	0.786	1.321	0.856
Energy	**0.208**	**0.325**	0.255	0.372	0.276	0.395	0.248	0.355	0.292	0.402
Environment	**0.309**	**0.393**	0.344	0.412	0.320	0.398	0.325	0.396	0.360	0.426

F.5 Ablation Analysis about Text Encoder and Vision Encoder

To evaluate the model's performance with different encoders, we conduct additional experiments by replacing both text and visual encoders. By default, HORAI employs Qwen-0.5B and ViT-Base for text and visual modalities, respectively. Considering time and computational constraints, we select encoders with relatively small parameter sizes. Specifically, the text encoders include GPT2-small, LLaMA3-1B, and Qwen2.5-1.5B, while the visual encoder comparison uses Swin Transformer. As shown in the Table 13, for a given text encoder, models with larger parameter sizes tend to perform slightly better, and employing more advanced architectures (e.g., Qwen and LLaMA) generally yields further improvements. In the comparison of visual encoders, ViT and Swin Transformer achieve similar overall forecasting performance.

Table 13: Ablation analysis of different text encoders and image encoders.

	HORAI		**GPT2**		**Llama3-1B**		**Qwen-1.5B**		**Swin Trans-Base**	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Agriculture	0.201	0.309	0.238	0.325	0.210	0.315	0.204	0.314	**0.198**	**0.302**
Climate	0.857	0.734	0.913	0.840	0.851	0.730	**0.832**	**0.715**	0.863	0.758
Energy	0.208	0.325	0.229	0.348	0.204	0.322	**0.202**	**0.318**	0.212	0.336
Environment	0.309	0.393	0.325	0.398	0.310	0.395	0.302	0.390	**0.300**	**0.388**

F.6 Efficiency Analysis

We compare HORAI with time series foundation models using two common efficiency metrics: parameter count and inference time. All experiments are conducted on the Environment dataset with a batch size of 1. As shown in Table 14, despite incorporating multimodal information, HORAI maintains comparable efficiency to unimodal foundation models. Furthermore, compared to the multimodal foundation model ChatTime, HORAI demonstrates significant advantages in both parameter size and inference speed.

F.7 Few-Shot Learning for Forecasting

To evaluate the rapid adaptation capability of HORAI on downstream tasks, we further conduct 10% Few-Shot learning experiments for forecasting and compare them with the Zero-Shot setting. As shown in Table 15, HORAI can further improve forecasting performance on most datasets with only a small amount of downstream data for fine-tuning. These results demonstrate that HORAI not only possesses strong Zero-Shot generalization ability but can also adapt to target domains with very limited supervision, showing excellent data efficiency and cross-domain transfer capability.

Table 14: Efficiency analysis on the environment dataset.

	Parameters(M)	**Inference Time(s)**
ChatTime	6820	0.5497
Sundial	128	0.0813
VisionTS	112	0.0073
TimesFM	203	0.0370
MOIRAI	310	0.0510
TimeVLM	152	0.0576
TimeMOE	453	0.0314
HORAI	635	0.0756

Table 15: Forecasting Performance of HORAI: Zero Shot vs 10% Few Shot. Few Shot results represent the mean across five independent runs with different random seeds, with observed variations bounded within ± 0.005.

	Zero Shot		**10% Few Shot**	
Metrics	MSE	MAE	MSE	MAE
Agriculture	0.201	0.309	0.194	0.295
Climate	0.857	0.734	0.850	0.732
Energy	0.208	0.325	0.201	0.322
Environment	0.309	0.393	0.270	0.371
Social Good	0.819	0.497	0.798	0.484
Traffic	0.165	0.244	0.156	0.245
EWJ	0.589	0.543	0.582	0.538
KR	0.551	0.448	0.548	0.442
MDT	0.376	0.434	0.368	0.430

G Limitations and Future Work

While HORAI incorporates textual and visual modalities, the visual modality in MM-TS is currently constructed from rendered time series plots, following prior time-series imaging studies that have demonstrated the effectiveness of visual representations for temporal modeling. Nevertheless, such rendered images remain a relatively constrained form of visual modality compared to naturally occurring visual observations. Future work could further explore the collection and integration of external visual information. In addition, the introduction of textual and visual modalities inevitably increases model complexity and computational overhead during both pre-training and inference. We plan to explore techniques like model compression to further improve computational efficiency, potentially enabling more resource-efficient deployments.

H Broader Impacts

The paper first proposes a multimodal pretraining paradigm that leverages time series with endogenous images and text and exogenous news, providing a comprehensive multi-view perspective for time series analysis. To support this, the paper develops an automated data construction pipeline to curate MM-TS, the large-scale multimodal time series dataset spanning six domains. Then propose HORAI, a frequency-enhanced multimodal foundation model. We believe it can be a valuable addition to multimodal time series research.

The paper focuses on the data generation pipeline construction and algorithm design. Using all the codes and datasets strictly follows the corresponding licenses. There is no potential ethical risk or negative social impact.

Table 16: Full time series forecasting results measured by MSE for HORAI, time series foundation models, and time-series-specific models.

Type		Time Series Foundation Models (Zero-Shot)								Time-Series-Specific Models (Full-Shot)					
Models		HORAI	Aurora	ChatTime	Chronos 2	TimesFM	TiRex 2	Sundial	VisionTS	ROSE	MOIRAI	GPT4MTS	TATS	GPT4TS	TimeVLM
Metric		MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE
Agriculture	6	0.131	0.184	0.243	**0.113**	0.151	0.128	0.218	0.210	0.219	0.187	0.161	0.140	0.135	0.143
	8	0.171	0.242	0.349	**0.161**	0.215	0.175	0.319	0.266	0.278	0.245	0.207	0.187	0.198	0.215
	10	**0.219**	0.297	0.390	**0.219**	0.284	0.225	0.425	0.307	0.408	0.297	0.230	0.244	0.258	0.271
	12	**0.286**	0.365	0.497	0.291	0.364	**0.286**	0.530	0.376	0.474	0.357	0.301	0.290	0.291	0.322
	avg	0.201	0.272	0.369	**0.197**	0.254	0.203	0.373	0.290	0.345	0.272	0.225	0.215	0.220	0.237
Climate	6	**0.854**	0.859	1.884	1.494	1.186	1.619	1.180	1.316	1.488	1.624	1.199	1.194	1.207	1.218
	8	**0.852**	0.858	1.843	1.496	1.158	1.592	1.159	1.312	1.598	2.148	1.205	1.178	1.191	1.181
	10	**0.859**	0.868	1.806	1.482	1.148	1.534	1.141	1.302	1.401	1.983	1.173	1.170	1.169	1.179
	12	**0.865**	0.875	1.909	1.483	1.134	1.488	1.134	1.297	1.414	1.929	1.152	1.179	1.171	1.203
	avg	**0.857**	0.865	1.860	1.489	1.157	1.558	1.154	1.307	1.475	1.921	1.182	1.180	1.184	1.195
Energy	12	0.105	0.117	0.104	0.122	**0.095**	0.113	0.125	0.173	0.268	0.183	0.111	0.105	0.111	0.114
	24	**0.189**	0.226	0.203	0.214	0.205	0.215	0.234	0.264	0.363	0.290	0.232	0.216	0.223	0.227
	36	**0.248**	0.292	0.292	0.275	0.340	0.285	0.318	0.346	0.413	0.367	0.308	0.309	0.314	0.309
	48	**0.292**	0.383	0.389	0.357	0.487	0.347	0.410	0.434	0.501	0.457	0.398	0.391	0.393	0.390
	avg	**0.208**	0.255	0.247	0.242	0.282	0.240	0.272	0.304	0.386	0.324	0.262	0.255	0.260	0.260
Environment	48	0.306	**0.281**	0.343	0.338	0.342	0.331	0.330	0.345	0.402	0.352	0.315	0.307	0.320	0.304
	96	0.326	**0.284**	0.369	0.342	0.355	0.349	0.353	0.370	0.409	0.370	0.340	0.334	0.340	0.327
	192	0.306	**0.270**	0.377	0.348	0.337	0.335	0.343	0.360	0.389	0.350	0.336	0.332	0.330	0.328
	336	0.298	**0.269**	0.372	0.368	0.320	0.315	0.317	0.340	0.369	0.332	0.299	0.302	0.300	0.320
	avg	0.309	**0.276**	0.365	0.349	0.339	0.332	0.336	0.354	0.392	0.351	0.323	0.319	0.322	0.319
Social Good	6	**0.682**	0.701	0.988	0.977	0.724	0.751	0.861	0.957	0.939	0.966	0.718	0.753	0.717	0.732
	8	**0.779**	0.804	1.044	1.153	0.782	0.847	0.994	1.106	1.168	1.532	0.942	0.875	0.855	0.822
	10	0.869	0.886	1.098	1.186	**0.834**	0.934	1.100	1.164	1.187	1.551	0.929	0.991	0.930	0.916
	12	0.948	0.960	1.149	1.194	**0.878**	1.012	1.187	1.278	1.272	1.671	1.093	1.053	1.167	1.005
	avg	0.819	0.838	1.069	1.128	**0.805**	0.886	1.036	1.126	1.141	1.430	0.920	0.918	0.917	0.868
Traffic	6	0.163	**0.154**	0.609	0.384	0.176	0.443	0.273	0.275	0.331	0.349	0.192	0.164	0.199	0.210
	8	0.166	**0.158**	0.626	0.410	0.181	0.455	0.275	0.282	0.365	0.461	0.195	0.178	0.204	0.212
	10	0.164	**0.163**	0.572	0.424	0.186	0.435	0.270	0.286	0.326	0.414	0.204	0.185	0.210	0.222
	12	**0.165**	0.168	0.579	0.445	0.189	0.417	0.268	0.282	0.342	0.400	0.218	0.189	0.211	0.222
	avg	0.165	**0.161**	0.596	0.416	0.184	0.438	0.271	0.281	0.341	0.406	0.203	0.179	0.206	0.216
EWJ	6	**0.548**	0.555	0.808	0.564	0.690	0.695	0.571	0.583	0.634	0.751	0.579	0.550	0.550	0.552
	8	**0.581**	0.592	0.880	0.605	0.725	0.746	0.611	0.629	0.729	1.017	0.608	0.611	0.597	0.599
	10	**0.604**	0.630	0.920	0.634	0.755	0.783	0.650	0.665	0.716	0.982	0.644	0.627	0.632	0.629
	12	**0.623**	0.642	0.940	0.673	0.780	0.815	0.681	0.701	0.746	0.997	0.673	0.661	0.649	0.657
	avg	**0.589**	0.605	0.887	0.619	0.738	0.760	0.628	0.645	0.706	0.937	0.626	0.612	0.607	0.609
KR	6	0.533	0.550	**0.528**	0.543	0.538	0.582	0.549	0.628	0.687	0.793	**0.528**	0.542	0.539	0.550
	8	**0.549**	0.581	0.564	0.560	0.554	0.614	0.572	0.674	0.798	1.077	0.564	0.569	0.573	0.580
	10	**0.561**	0.588	0.570	0.562	0.570	0.632	0.586	0.685	0.727	1.063	0.566	0.600	0.594	0.601
	12	**0.562**	0.592	0.598	0.574	0.574	0.636	0.590	0.698	0.750	1.038	**0.562**	0.602	0.604	0.606
	avg	**0.551**	0.578	0.565	0.560	0.559	0.616	0.574	0.671	0.741	0.992	0.555	0.578	0.578	0.584
MDT	6	**0.361**	0.370	0.466	0.376	0.401	0.415	0.372	0.412	0.426	0.494	0.369	0.365	0.373	0.369
	8	**0.372**	0.387	0.474	0.382	0.420	0.424	0.387	0.431	0.483	0.668	0.377	0.383	0.386	0.385
	10	**0.382**	0.405	0.526	0.399	0.426	0.434	0.407	0.437	0.456	0.630	0.389	0.397	0.395	0.400
	12	**0.390**	0.420	0.518	0.400	0.435	0.445	0.420	0.453	0.477	0.632	0.405	0.411	0.411	0.414
	avg	**0.376**	0.396	0.496	0.389	0.421	0.429	0.397	0.433	0.461	0.606	0.385	0.389	0.391	0.392
1st Count		**28**	9	1	4	4	1	0	0	0	0	2	0	0	0

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the contribution and scope of the paper: we first propose a multimodal pretraining paradigm that leverages time series with endogenous images and text and exogenous news, providing a comprehensive multi-view perspective for time series analysis. To support this, we develop an automated data construction pipeline to curate MM-TS, the first large-scale multimodal time series dataset spanning six domains. Then we propose HORAI, a frequency-enhanced multimodal foundation model.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Table 17: Full time series forecasting results measured by MAE for HORAI, time series foundation models, and time-series-specific models.

Type		Time Series Foundation Models (Zero-Shot)										Time-Series-Specific Models (Full-Shot)			
Models		HORAI	Aurora	ChatTime	Chronos 2	TimesFM	TiRex 2	Sundial	VisionTS	ROSE	MOIRAI	GPT4MTS	TATS	GPT4TS	TimeVLM
Metric		MAE	MAE	MAE	MAE	MAE	MAE	MAE	MAE	MAE	MAE	MAE	MAE	MAE	MAE
Agriculture	6	0.260	0.295	0.340	**0.222**	0.266	0.233	0.304	0.289	0.299	0.342	0.257	0.251	0.242	0.245
	8	0.294	0.335	0.399	**0.260**	0.316	0.267	0.364	0.323	0.339	0.391	0.288	0.282	0.284	0.287
	10	0.323	0.365	0.418	**0.297**	0.359	0.298	0.423	0.348	0.406	0.423	0.305	0.320	0.313	0.320
	12	0.362	0.398	0.483	0.339	0.402	**0.334**	0.477	0.386	0.443	0.455	0.342	0.350	0.338	0.359
	avg	0.309	0.348	0.410	**0.280**	0.336	0.283	0.392	0.336	0.372	0.403	0.298	0.301	0.294	0.302
Climate	6	**0.736**	0.747	1.118	1.026	0.887	1.015	0.891	0.932	0.993	1.016	0.895	0.897	0.901	0.907
	8	**0.733**	0.746	1.100	1.017	0.876	1.014	0.885	0.935	1.031	1.152	0.899	0.886	0.892	0.914
	10	**0.734**	0.748	1.090	1.000	0.867	0.999	0.876	0.928	0.967	1.112	0.885	0.881	0.886	0.880
	12	**0.736**	0.753	1.117	0.988	0.858	0.986	0.870	0.925	0.957	1.101	0.876	0.885	0.883	0.896
	avg	**0.734**	0.749	1.106	1.007	0.873	1.003	0.881	0.930	0.987	1.095	0.889	0.887	0.891	0.899
Energy	12	0.226	0.245	0.222	0.230	**0.218**	0.226	0.242	0.313	0.401	0.309	0.244	0.232	0.243	0.253
	24	**0.314**	0.354	0.321	0.325	0.325	0.330	0.345	0.395	0.469	0.396	0.362	0.344	0.355	0.359
	36	**0.362**	0.409	0.396	0.379	0.406	0.389	0.409	0.454	0.497	0.449	0.418	0.418	0.423	0.410
	48	**0.399**	0.472	0.470	0.438	0.487	0.438	0.473	0.516	0.549	0.515	0.496	0.480	0.484	0.475
	avg	**0.325**	0.370	0.352	0.343	0.359	0.346	0.367	0.420	0.479	0.417	0.380	0.368	0.376	0.374
Environment	48	**0.379**	0.380	0.406	0.390	0.393	0.389	0.410	0.426	0.459	0.404	0.400	0.389	0.396	0.387
	96	0.396	**0.382**	0.465	0.402	0.401	0.399	0.423	0.441	0.465	0.415	0.401	0.402	0.401	0.405
	192	0.395	**0.375**	0.474	0.401	0.396	0.388	0.419	0.442	0.452	0.402	0.411	0.401	0.391	0.403
	336	0.403	**0.377**	0.478	0.412	0.395	0.379	0.411	0.436	0.447	0.390	0.390	0.391	0.383	0.395
	avg	0.393	**0.379**	0.456	0.400	0.397	0.389	0.416	0.436	0.456	0.403	0.400	0.396	0.393	0.397
Social Good	6	0.418	0.442	0.451	0.431	0.430	0.421	0.487	0.543	0.499	0.522	0.382	**0.370**	0.374	0.379
	8	0.474	0.493	0.488	0.482	0.478	0.463	0.549	0.605	0.588	0.653	0.505	**0.409**	0.459	0.427
	10	0.528	0.543	0.519	0.525	0.492	0.503	0.604	0.636	0.595	0.691	**0.446**	0.459	0.463	0.465
	12	0.569	0.587	0.554	0.556	0.560	0.538	0.651	0.688	0.642	0.736	**0.470**	0.474	0.608	0.505
	avg	0.497	0.516	0.503	0.498	0.490	0.481	0.573	0.618	0.581	0.651	0.451	**0.428**	0.476	0.444
Traffic	6	0.240	0.285	0.623	0.469	0.246	0.515	0.410	0.411	0.449	0.448	0.264	**0.226**	0.278	0.316
	8	0.245	0.286	0.636	0.483	0.252	0.514	0.408	0.410	0.455	0.499	0.256	**0.242**	0.262	0.313
	10	0.246	0.289	0.592	0.519	0.259	0.493	0.403	0.406	0.443	0.466	0.257	**0.243**	0.264	0.328
	12	0.247	0.294	0.592	0.507	0.264	0.477	0.401	0.402	0.458	0.458	0.268	**0.242**	0.260	0.322
	avg	0.244	0.289	0.610	0.495	0.256	0.500	0.405	0.407	0.451	0.468	0.261	**0.238**	0.266	0.319
EWJ	6	0.528	0.540	0.612	0.538	0.561	0.592	0.535	0.560	0.581	0.623	0.531	0.525	0.523	**0.521**
	8	**0.537**	0.576	0.641	0.545	0.576	0.607	0.551	0.580	0.626	0.714	0.540	0.544	0.538	0.541
	10	**0.550**	0.584	0.652	0.557	0.589	0.619	0.567	0.591	0.599	0.705	0.559	0.551	0.551	0.554
	12	**0.558**	0.597	0.659	0.573	0.599	0.627	0.580	0.607	0.613	0.709	0.566	0.563	0.560	0.562
	avg	**0.543**	0.574	0.641	0.553	0.581	0.611	0.559	0.584	0.605	0.688	0.549	0.546	**0.543**	0.544
KR	6	0.435	0.454	0.436	0.445	0.445	0.470	0.450	0.503	0.521	0.567	0.442	**0.426**	0.435	0.437
	8	0.446	0.463	0.452	0.450	0.448	0.488	0.462	0.524	0.572	0.650	0.452	0.446	**0.444**	0.451
	10	0.454	0.474	0.459	0.460	0.457	0.497	0.469	0.526	0.530	0.649	0.455	0.462	**0.452**	0.463
	12	0.458	0.480	0.473	0.464	0.463	0.502	0.475	0.535	0.547	0.649	**0.453**	0.461	0.459	0.467
	avg	**0.448**	0.468	0.455	0.455	0.452	0.489	0.464	0.522	0.542	0.629	0.450	0.449	**0.448**	0.454
MDT	6	0.426	0.433	0.455	0.445	0.442	0.462	0.433	0.471	0.476	0.521	0.436	0.423	**0.422**	0.423
	8	0.433	0.445	0.473	0.450	0.448	0.470	0.443	0.486	0.514	0.591	0.439	0.433	**0.432**	0.434
	10	**0.438**	0.451	0.494	0.483	0.452	0.477	0.453	0.487	0.486	0.580	0.444	0.440	0.448	0.443
	12	**0.442**	0.463	0.494	0.485	0.456	0.483	0.460	0.495	0.499	0.582	0.450	0.447	0.452	0.448
	avg	**0.434**	0.448	0.479	0.466	0.450	0.473	0.447	0.485	0.493	0.569	0.442	0.436	0.438	0.437
1st Count		**18**	4	0	4	1	1	0	0	0	0	3	9	6	1

Justification: The paper discusses the limitations of HORAI in Appendix G.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

Table 18: Time series anomaly detection results under zero-shot and full-shot settings with multiple metrics. The best results are in **bold**, and the second-best results are underlined.

Type	Time Series Foundation Models (Zero-Shot)					Time-Series-Specific-Models (Full-Shot)								
Datasets	Metric	HORAI	DADA	Timer	UniTS	GPT4TS	LLMMixer	TimesNet	DCdetector	A.T.	PatchTST	HBOS	IForest	PCA
EWJ	Aff-F1	**82.54**	81.26	78.06	77.61	76.65	66.86	81.82	48.10	59.03	75.82	71.03	67.55	51.06
	F1	**56.28**	49.33	41.21	39.18	48.33	18.95	49.37	17.09	14.39	45.39	44.80	41.67	18.68
	Range-AUC-ROC	**79.84**	68.44	64.71	67.15	55.57	41.75	74.22	45.69	27.50	69.26	61.02	57.94	43.78
	Range-AUC-PR	43.33	41.61	30.67	44.31	32.83	12.36	41.84	12.52	9.01	33.37	**46.37**	34.77	16.49
	AUC-PR	51.97	**55.24**	44.01	50.33	46.75	18.81	54.99	10.88	8.97	47.91	25.24	24.16	10.99
	AUC-ROC	**86.95**	79.11	76.15	79.87	75.58	57.69	82.39	53.40	43.81	78.53	71.82	69.20	54.35
	VUS-ROC	**82.78**	71.79	67.72	73.91	67.95	52.79	75.76	47.10	31.75	71.96	62.07	59.24	45.26
	VUS-PR	**48.27**	43.36	33.17	39.32	35.63	15.13	43.15	15.37	10.85	36.08	41.19	37.81	19.38
MDT	Aff-F1	80.66	77.99	78.51	75.57	**80.81**	67.65	80.08	47.33	66.12	79.47	52.33	53.74	54.66
	F1	**59.36**	53.70	48.39	51.70	49.14	27.71	54.88	19.54	25.46	49.40	43.84	38.10	20.75
	Range-AUC-ROC	**86.59**	63.94	58.98	58.78	59.00	42.44	77.01	43.65	41.41	75.61	54.86	53.22	41.90
	Range-AUC-PR	**51.11**	44.63	37.54	36.14	42.48	15.30	48.60	13.30	13.20	13.11	43.16	33.63	19.53
	AUC-PR	61.98	63.03	55.86	53.44	60.40	19.86	**65.57**	11.59	15.29	54.11	28.66	22.41	12.29
	AUC-ROC	**91.22**	79.04	75.65	73.19	74.79	60.30	86.67	53.82	56.44	84.55	60.26	63.92	54.51
	VUS-ROC	**86.82**	66.76	60.28	58.67	62.30	46.80	83.40	45.02	44.53	77.69	55.30	54.02	44.09
	VUS-PR	**56.88**	46.81	38.38	37.61	44.81	15.21	52.13	15.72	15.93	41.67	44.77	35.32	22.93
KR	Aff-F1	85.44	84.22	**89.55**	82.24	79.56	71.80	85.47	61.94	70.99	79.52	64.78	69.38	58.11
	F1	71.89	49.48	58.04	30.23	**74.01**	20.25	58.14	11.98	11.10	36.64	60.71	53.97	22.76
	Range-AUC-ROC	**86.16**	69.91	74.61	71.29	65.15	49.01	78.29	41.75	40.18	72.72	61.80	61.10	51.01
	Range-AUC-PR	**59.64**	46.95	51.59	40.75	37.53	13.25	52.83	6.04	7.44	35.22	51.69	43.07	18.99
	AUC-PR	**72.91**	63.55	66.72	55.39	56.78	28.19	67.47	8.10	7.01	53.60	41.09	32.21	10.18
	AUC-ROC	**96.38**	79.53	66.72	80.95	78.30	65.77	85.88	52.97	51.25	82.15	75.16	74.45	63.58
	VUS-ROC	**93.54**	70.82	75.99	73.93	67.81	47.06	79.00	43.04	41.97	74.65	58.77	60.70	47.51
	VUS-PR	**60.76**	45.90	51.41	43.32	38.23	19.10	51.60	8.49	7.94	36.18	54.17	43.31	24.19
Energy	Aff-F1	**71.37**	64.38	60.20	63.84	66.37	65.85	66.00	47.07	43.39	66.85	55.85	62.03	57.65
	F1	**37.71**	31.54	31.71	31.66	33.22	33.08	33.95	12.63	12.05	34.81	34.83	34.39	35.12
	Range-AUC-ROC	**62.93**	55.78	46.82	52.12	53.54	55.25	61.56	45.39	31.52	61.39	51.06	52.64	52.64
	Range-AUC-PR	33.24	33.47	28.81	30.70	31.10	30.59	38.17	21.77	19.24	35.25	42.14	**45.19**	43.89
	AUC-PR	39.82	37.81	38.05	27.51	33.75	32.85	**42.05**	17.69	14.02	34.25	21.55	21.17	21.69
	AUC-ROC	**68.44**	62.33	60.54	63.38	66.54	61.31	68.36	48.75	38.68	66.70	60.80	60.32	61.14
	VUS-ROC	**62.42**	54.37	46.03	51.15	53.10	53.04	59.47	45.93	31.56	58.31	51.50	53.61	53.07
	VUS-PR	35.24	34.18	29.46	31.04	31.68	30.35	38.61	22.57	19.69	34.41	42.57	**46.03**	44.30
Weather	Aff-F1	**80.84**	69.01	75.46	76.17	72.56	73.68	80.58	42.80	49.22	77.17	47.70	54.06	64.91
	F1	47.44	35.29	46.42	50.00	40.16	43.13	**51.58**	11.14	15.59	49.60	42.94	49.21	40.41
	Range-AUC-ROC	80.61	61.95	73.37	75.55	71.43	72.54	**83.11**	45.41	43.11	80.47	54.12	56.69	57.80
	Range-AUC-PR	**50.88**	29.86	43.20	44.31	41.37	43.40	50.58	18.06	18.85	49.81	46.37	49.65	47.47
	AUC-PR	49.16	29.80	48.87	49.91	44.12	49.71	47.56	17.08	16.71	**53.39**	31.16	35.44	25.02
	AUC-ROC	81.49	66.37	80.86	81.22	74.47	79.60	81.10	47.90	47.11	**82.02**	64.47	67.81	67.71
	VUS-ROC	80.40	61.03	73.22	75.08	70.03	71.71	**81.91**	45.56	43.32	79.97	54.16	56.45	57.38
	VUS-PR	**50.76**	30.00	43.21	44.35	41.30	43.47	50.09	18.33	19.17	50.13	46.58	49.66	47.13
	1st Count	**26**	1	1	0	2	0	5	0	0	2	1	2	0

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not provide theoretical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper fully discloses all the information needed to reproduce the main experimental results of the paper.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide our code at https://github.com/decisionintelligence/HORAI.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies all the training and test details in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Please refer to the Appendix F.7.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: The paper provides sufficient information on the computer resources needed to reproduce the experiments in Appendix F.6.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: Our research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses both potential positive societal impacts and negative societal impacts of the work performed in Appendix H.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks because all the experiments are conducted on publicly available

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We conduct experiments on the open-source datasets, and the creators or original owners of assets used in the paper are properly credited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper uses the LLMs to generate the text of MM-TS in Appendix A.2, and uses the LLMs as the text encoder of HORAI.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.