A HYBRID ATTENTION MODEL LEARNING UNIFIED TIME-AWARE PATCH REPRESENTATION FOR IRREGULAR MULTIVARIATE TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Time series foundation models (TSFMs) have recently delivered impressive zero-shot performance across diverse forecasting tasks. However, real-world decision-making frequently relies on *irregular multivariate time series* (IMTS), where inconsistent inter-observation intervals and asynchronous sampling across variables coexist with informative missingness. In this paper, we propose a hybrid attention model that learns a unified time-aware patch representation for IMTS forecasting. We first design a *time-aware patch encoder* that maps a variable number of intra-patch timestamps into a fixed-size embedding, producing a uniform format for irregular patches without resorting to imputation. We then introduce a *time bias attention* mechanism that calibrates inter-patch temporal misalignment and asynchronous cross-channel dependencies as an auxiliary attention offset. Finally, on top of a decoder-only Transformer backbone, we adopt a *hybrid causal mask* that preserves a bidirectional full view over the historical context while keeping the forecast horizon strictly autoregressive. To support large-scale pretraining under irregular settings, we also curate VersaTSA, an expanded archive of over 231B observations that retains the native sampling sparsity of its sources. Experiments on three IMTS benchmarks and a standard regular-MTS benchmark show that our model achieves state-of-the-art zero-shot performance on IMTS and remains competitive when transferred to regular forecasting.

1 INTRODUCTION

With the advancement of foundational models (FMs) (Bommasani et al., 2022), time series analysis is undergoing a profound transformation. Numerous efforts have been made to establish a unified paradigm for Time Series Foundation models (TSFMs) that possess exceptional generalization capabilities for multivariate time series forecasting, such as Sundial (Liu et al., 2025), Chronos (Ansari et al., 2024), Moirai (Woo et al., 2024), and TimesFM (Das et al., 2024). However, existing large-scale TSFMs inherently struggle to handle irregularly sampled time series. They require data preprocessing methods such as interpolation to align the data to regular sampling intervals, thereby introducing additional noise (Che et al., 2018) and obscuring informative missingness (RUBIN, 1976). This approach leads to mismatches when applying these models to real-world scenarios characterized by inherent temporal inconsistencies. Thus, in this paper, we aim to explore how to build a universal model for irregular multivariate time series (IMTS).

By investigating irregular time series in various real-world domains, such as financial analysis (Sezer et al., 2020), cyber security (Sousa et al., 2020; Fawaz et al., 2019), and weather science (Ravuri et al., 2021), we can conclude that irregular multivariate time series exhibit two characteristics: (1) *inconsistent time intervals* between observation points and (2) *diverse sampling frequencies* among variables. Existing TSFMs rely on traditional positional encoding (e.g., Sinusoidal (Vaswani et al., 2023), RoPE (Su et al., 2024), T5 Bias (Press et al., 2021)), which can only encode the sequential order of indices. When temporal intervals are irregular, the positional encoding fails to describe the frequency changes and the natural differences among patches. Concurrently, some models for irregularly sampled time series address the issue by point-by-point processing (Rubanova et al., 2019; Luo et al., 2025; Zhang et al., 2024). While this approach offers flexibility in handling irregular data, it reduces the information density and increases computational complexity.

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

Based on the above observations, our idea starts from a natural understanding of irregular intervals. Empirically, continuous-time encoding transforms timestamps into a continuous space (Kazemi et al., 2019), enabling models to "perceive" temporal dimensions. Thus, we can build a universal temporal representation learning method based on TSFMs to achieve continuous modeling of time. To achieve this target, the following challenges need to be addressed. First, existing continuous-time encoding approaches predominantly embed time using timestamps as units. But multiple timestamps may exist within a single patch. Although the average time encoding provides a solution to represent the temporal information of patches, the capture of **intricate inner-patch temporal structure** remains underexplored. Secondly, regarding inter-patch position encoding, conventional methods encode the sequence order of the index. It neglects the **unevenness between patches**. Third, existing base models often employ Decoder-Only Transformers with causal masks to maintain the autoregressive manner. However, inconsistent sampling rates or missing values across different variables lead to overlapping patches during the chunking process, rendering **causal masking ineffective**.

To address the above challenges, we propose a hybrid attention model to learn the **UNI**fied **T**ime-aware patch embeddings for **I**rregular **M**ultivariate Time **S**eries forecasting (UNITIMS). We specifically decouple the input time series into two parts: the time series values and times. For time series values, we transform them into normalized patch tokens as most TSFMs do. For times, we design a patch-based temporal structure encoding approach, named *time-aware patch encoder*, to represent temporal structural information. It can represent inner-patch temporal information in a universal format. Then, we introduce *time bias attention* to calibrate attention biases arising from inter-patch temporal structural discrepancies. Finally, we introduce the *hybrid causal mask strategy*, full attention for the historical window (input) and causal attention for the autoregressively generated target (future window) based on the common decoder-only architecture used in foundation models. This preserves causal information extraction from irregular time series while preventing unintended exposure of irregular temporal inputs.

In this work, we conduct a systematic investigation of the pretraining behavior of time series foundation models on irregular multivariate time series (IMTS). By collecting and curating existing time series datasets, we construct a unified dataset collection termed the Versatile Time Series Archive (VersaTSA), which, after a subsequent expansion, includes over 231 billion observations from diverse domains and preserves the irregularity of the originally collected time series. To accommodate pretraining under irregular settings, we reconstruct the timestamps of each sequence and represent time series using sequences of triplets that store the observed time series values, timestamps, and corresponding channels. We summarize our contributions as:

- We propose a time-aware patch encoding method to represent the irregular temporal information for IMTS, allowing TSFMs to be capable of forecasting irregular time series data without decreasing information density.
- We present UNITIMS, which pioneers the use of large-scale decoder-only foundation models based on hybrid attention mechanisms to address irregular sequential data with a unified learning of time series values and continuous times.
- We extend and release VersaTSA, a large-scale time series dataset comprising over 231B observations, and use it for pre-training UNITIMS. Experimental results on various datasets show that UNITIMS achieves state-of-the-art performance, outperforming both full-shot time series models and zero-shot results from other TSFMs.

2 Related Work

Irregular multivariate time series (IMTS) involve asynchronous observations and non-uniform temporal intervals. Early studies mainly treated irregular observations as missing values and relied on kernel methods (Rehfeld et al., 2011), multiple imputation (White et al., 2011), and EM algorithms (García-Laencina et al., 2010). However, separating imputation from forecasting may overlook informative missing patterns (Wells et al., 2013). Recent end-to-end methods directly model temporal irregularity, including GRU-D (Che et al., 2018), Neural ODE-based approaches (Chen et al., 2018; Rubanova et al., 2019; Zhang et al., 2025b), and attention-based methods (Shukla & Marlin, 2021; Zhang et al., 2024; Luo et al., 2025; Sun et al., 2025) that exploit explicit temporal representations such as Time2Vec (Kazemi et al., 2019). Despite their effectiveness, these models are generally designed for specific tasks or data distributions, limiting their OOD generalization.

Recent Time Series Foundation Models (TSFMs) have achieved strong zero-shot and few-shot forecasting through large-scale pre-training. Patch-based architectures (Nie et al., 2023) have become prevalent, with models such as Timer (Liu et al., 2024b), TimesFM (Das et al., 2024), and AutoHFormer (Zhang et al., 2025a) adopting channel-independent modeling, while Chronos-2 (Ansari et al., 2025) further captures cross-variate context through synchronous Group Attention. Other TSFMs, including Moirai (Woo et al., 2024), TimeGPT (Garza et al., 2024), and Sundial (Liu et al., 2025), demonstrate broad cross-domain generalization. However, these models are predominantly developed under regular sampling grids: patch indices implicitly correspond to fixed temporal intervals, and synchronous cross-variate modeling assumes aligned observations. Such assumptions prevent them from explicitly capturing asynchronous dependencies and irregular temporal intervals in IMTS. Motivated by this gap, we aim to extend the foundation-model paradigm to irregular domains by natively modeling asynchronous multivariate contexts and irregular temporal intervals without imputation.

3 Problem Definition

Consider an irregular multivariate time series dataset $\mathcal{D} = \{(u_k, t_k, c_k)\}_{k=1}^{N}$, where N denotes the number of observed variates, $u_k \in \mathbb{R}$ is the observed value, $t_k \in \mathbb{R}^+$ is the timestamp, and $c_k \in \mathcal{C}$ is the channel identifier. Based on the channel partition $\mathcal{C} = \mathcal{C}_\mathcal{Y} \cup \mathcal{C}_\mathcal{Z}$, the dataset $\mathcal{D}$ is decomposed into target variables $\mathcal{Y} = \mathcal{D}\big|_{\mathcal{C}_\mathcal{Y}} = \{(u, t, c) \mid c \in \mathcal{C}_\mathcal{Y}\}$ and the covariate series $\mathcal{Z} = \mathcal{D}\big|_{\mathcal{C}_\mathcal{Z}}$. The goal of the IMTS forecasting task is to predict the values for a specific query set $Q = \{(t, c) \mid t \in [t_L, t_{L+H}], c \in \mathcal{C}_\mathcal{Y}\}$, where $[t_L, t_{L+H}]$ defines the future forecast horizon. Formally, the IMTS forecasting problem can be formulated as follows:

$$\mathcal{F}(\mathcal{Y}_{t_1:t_L}, \mathcal{Z}_{t_1:t_{L+H}}, Q) \to \hat{\mathcal{Y}}_Q, \tag{1}$$

where $\mathcal{Y}_{t_1:t_L} = \{(u_k, t_k, c_k) \mid c_k \in \mathcal{C}_\mathcal{Y}, t_k \in [t_1, t_L]\}$ is the lookback window in the history interval $[t_1, t_L]$, and $\hat{\mathcal{Y}}_Q = \{(u_k, t_k, c_k) \mid (t, c) \in Q\}$ is the set of predicted values for the given queries.

4 Method

As shown in Figure 1, UniTIMS models irregular multivariate time series with a patch-based, decoder-only architecture. The input is split into non-overlapping patches (Nie et al., 2023) of length l_p. For patch m, we form $p_m^{(c)} = \{(u_k, t_k, c) \mid (m-1)l_p + 1 \le k \le ml_p\}$ and flatten patches into a single sequence (Woo et al., 2024). For each channel, values $\{u_{(m-1)l_p}, \ldots, u_{ml_p}\}$ are encoded into $\mathbf{x}_m$, and timestamps $\{t_{(m-1)l_p}, \ldots, t_{ml_p}\}$ into $\mathbf{p}_m$ via time-aware patch encoding (Sec. 4.1) to capture intra-patch time. These embeddings are fed to a Transformer to autoregressively generate outputs. We replace standard attention with time bias attention (Sec. 4.2) to mitigate inter-patch bias, and use a hybrid causal mask (Sec. 4.3) during generation to model complex dependencies.

4.1 Time-Aware Patch Encoder

Irregular multivariate time series inputs typically consist of two parts: the observed time series values and their corresponding timestamp. For the time series values, we follow the standard paradigm of mainstream TSFMs, transforming them into continuous patch tokens via patching and normalization. Regarding the timestamp, we construct a continuous temporal structure representation that unifies the format for extracting temporal features from irregularly sampled patch-based data.

4.1.1 Continuous Time Mapping

We firstly project the original continuous timestamp t into a high-dimensional feature vector to obtain initial time embedding $\mathbf{\Phi}(t) \in \mathbb{R}^d$:

$$\mathbf{\Phi}(t)[i] = \begin{cases} w[i] \cdot t + a[i], & \text{if } i = 0 \\ a[i] \cdot \sin(w[i] \cdot t), & \text{if } i \text{ is odd} \\ a[i] \cdot \cos(w[i] \cdot t), & \text{otherwise} \end{cases}, \tag{2}$$

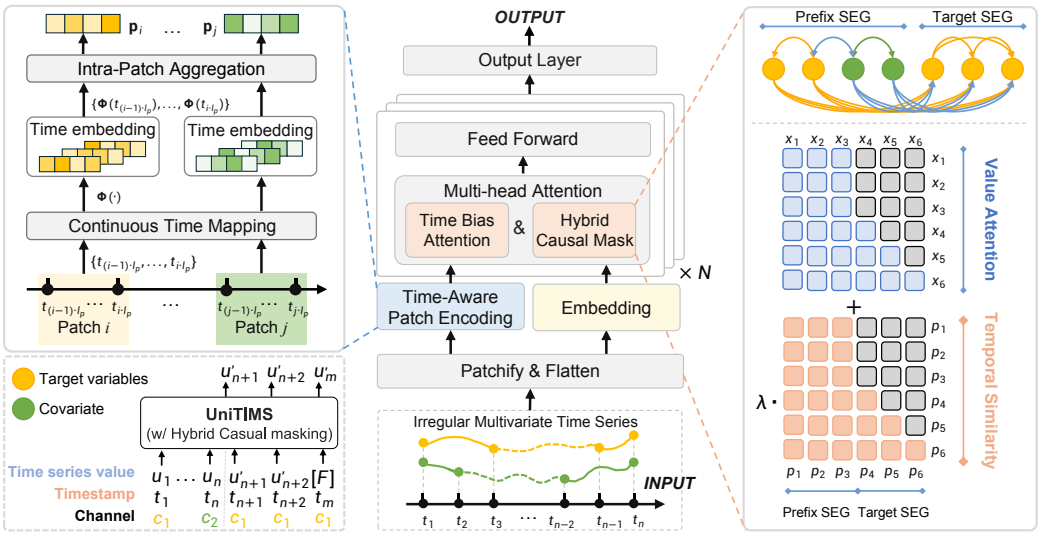

Figure 1: Overview of UNITIMS. The input sequence is split into patches of size 96. Timestamp and value patches are embedded separately and fed into Transformer blocks with Time Bias Attention and a Hybrid Causal Mask, followed by an output layer for value prediction. For simplicity, we illustrate a single target and one observable covariate; multiple variables are flattened and processed identically. During step-by-step inference, predictions at the special token [F] are fed back, and [F] advances one step forward.

where $\mathbf{\Phi}(t)[i]$ is the i-th element of $\mathbf{\Phi}(t)$, d denotes the dimension of vector, $w \in \mathbb{R}^d$ and $a \in \mathbb{R}^d$ are learnable frequency and amplitude parameters. The linear term captures the global temporal evolution trend, while the periodic components are extracted using Fourier terms at different frequencies.

4.1.2 INTRA-PATCH AGGREGATION

To align temporal information with patch representation, given the timestamp set $\mathcal{T}_m = \{t_{k_1}, \cdots, t_{k_{n_m}}\}$ of the m-th patch (where $n_m \leq l_p$ is the number of *actually observed* timestamps in this patch and may vary across patches and channels under IMTS), we first encode each timestamp via $\mathbf{\Phi}(\cdot)$ to obtain a sequence of n_m time embeddings. We then aggregate these variable-length embeddings into a single d-dimensional patch embedding $\mathbf{p}_m$ by an attention-based pooling that is invariant to the number of valid observations:

$$\mathbf{p}_m = \sum_{i=1}^{n_m} \alpha_i \, \mathbf{\Phi}(t_{k_i}), \quad \alpha_i = \frac{\exp\left(\mathbf{q}^\top \mathbf{\Phi}(t_{k_i})\right)}{\sum_{j=1}^{n_m} \exp\left(\mathbf{q}^\top \mathbf{\Phi}(t_{k_j})\right)}, \tag{3}$$

where $\mathbf{q} \in \mathbb{R}^d$ is a learnable query vector shared across patches. For implementation, we materialize each patch as a tensor of length l_p with a binary observation mask $\mathbf{m}_m \in \{0,1\}^{l_p}$ marking the n_m valid positions; unobserved entries are excluded from the softmax in eq. (3) (mask value $-\infty$ on logits) and contribute neither to the gradient nor to $\mathbf{p}_m$. This design differs fundamentally from imputation- or zero-padding-based encoders: the resulting $\mathbf{p}_m$ depends only on the observed timestamps and their continuous-time embeddings, so patches with different sampling densities map to the *same* representation space without injecting spurious values. This masked attention pooling effectively handles inconsistencies in sampling frequency and time intervals inherent to irregular multivariate time series, while keeping the output dimension fixed regardless of n_m.

4.2 TIME BIAS ATTENTION

In the Transformer backbone, given input value embeddings $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^d$ from patches i and j, the initial attention weight matrix $\mathbf{A}_{i,j}^{(l)}$ is calculated as the scaled dot product between the query and key projections, inevitably introducing temporal misalignment bias. To address this, we design a

time bias attention mechanism that calibrates attention scores by integrating the temporal similarity matrix across patches with both inter-variable and intra-variable correlation matrices. Specifically, we first measure the correlation of temporal structures between patches via a bilinear mapping, computing the inter-patch similarity as follows:

$$\mathbf{G}_{i,j} = \mathbf{p}_i^\top \mathbf{W}_E \mathbf{p}_j, \tag{4}$$

where $\mathbf{W}_E \in \mathbb{R}^{d\times d}$ is a learnable parameter matrix, and $\mathbf{p}_i$ and $\mathbf{p}_j$ are the time-aware patch embedding corresponding to patches i and j.

Then, to model multivariable correlations, we introduce the any-variate bias (Woo et al., 2024) by constructing the intra-variate matrix $v_1^{(l)}\mathbb{I}_{\{c_i=c_j\}}$ and the inter-variate matrix $v_2^{(l)}\mathbb{I}_{\{c_i\neq c_j\}}$, which enhance the attention mechanism's capacity to represent multivariate time series. Here, c_i and c_j denote the channel indices of the i-th and j-th patches, respectively. $\mathbb{I}_{\{\cdot\}}$ is the indicator function used to differentiate intra-channel and inter-channel dependencies. The learnable scalars $v_1^{(l)}$ and $v_2^{(l)}$ allow the model to adaptively balance between channel-independent and channel-mixing modes.

Finally, we incorporate the aforementioned temporal structure prior and multivariate bias as additive inductive biases directly into the initial attention weights. The updated attention matrix is $\hat{\mathbf{A}}_{i,j}^{(l)}$ computed as follows:

$$\hat{\mathbf{A}}_{i,j}^{(l)} = \mathbf{A}_{i,j}^{(l)} + \lambda^{(l)} \cdot \mathbf{G}_{i,j} + v_1^{(l)}\mathbb{I}_{\{c_i=c_j\}} + v_2^{(l)}\mathbb{I}_{\{c_i\neq c_j\}}, \tag{5}$$

where $\lambda^{(l)}$ is a learnable scaling factor specific to the l-th layer, which dynamically adjusts the influence of the temporal bias $\mathbf{G}$. Overall, we remove positional encodings from the Transformer architecture and instead rely on time bias attention to enable the UniTIMS to perceive inter-patch correlations in irregular multivariate time series modeling.

4.3 Hybrid Causal Mask

Based on the updated attention matrix $\hat{\mathbf{A}}$, we design a hybrid causal mask (HCM) strategy to maintain autoregressive generation while maximizing the model's capacity for global modeling of irregular contexts. Specifically, given a prefix length k, let $\mathbf{x}_i$ and $\mathbf{x}_j$ denote the query and key tokens derived from the value embeddings of their respective patches. The element of the mask matrix $\mathbf{M}_{i,j}$ is defined as:

$$\mathbf{M}_{i,j} = \begin{cases} 0, & \text{if } j \leq \max(i,k) \\ -\infty, & \text{otherwise} \end{cases}, \tag{6}$$

Subsequently, incorporating this mask matrix, the final masked attention matrix at the l-th layer is computed as:

$$\mathbf{E}_{i,j}^{(l)} = \mathrm{softmax}(\hat{\mathbf{A}}_{i,j}^{(l)} + \mathbf{M}_{i,j}). \tag{7}$$

The HCM strategy enables the model to perform bidirectional full attention within the prefix segment ($i \leq k$), thereby fully capturing global dependencies in IMTS. In contrast, for the target segment ($i > k$), the model is strictly constrained by unidirectional causal masking to ensure autoregressive consistency in time series forecasting.

4.4 VersaTSA Dataset

4.4.1 Dataset Description

To train foundation models tailored for IMTS, we collected and constructed the Versatile Time Series Archive dataset (VersaTSA), designed to support IMTS pretraining by preserving the original sampling sparsity inherent in the data. VersaTSA comprises the following datasets: 1) Asynchronously sampled torque data from offshore wind power and wave height measurements, with a maximum sampling frequency of 50 Hz, collected over one week. 2) The LOTSA dataset (Woo et al., 2024), from which padding values have been removed to restore the original irregular sampling distribution. 3) The irregular multivariate time series portion of the Time-IMM dataset (Chang et al., 2025), where only the time series data are retained, excluding unstructured textual information. To facilitate efficient integration with deep learning pipelines, VersaTSA employs Apache Arrow (Richardson et al.,

2025) for unified serialized storage. Following its subsequent expansion, the archive contains over 231 billion observations. Detailed dataset statistics and preprocessing metadata are provided in the project repository [1].

4.4.2 HOLD-OUT PROTOCOL

To prevent any data leakage in the downstream zero-shot evaluation, the three IMTS evaluation datasets used in Section 5.3.1 (PhysioNet, Human Activity, and USHCN) and the six regular MTS evaluation datasets used in Section 5.3.2 are *strictly excluded* from VersaTSA. We further filter out any sub-corpus from LOTSA whose source overlaps with these benchmarks. Pretraining and downstream evaluation therefore operate on disjoint data sources.

5 EXPERIMENTS

Extensive experiments are conducted to evaluate the performance of UNITIMS, which mainly focus on 1) zero-shot forecasting capability on typical IMTS datasets and general MTS datasets, 2) the ablation study of each module, and 3) the further analysis of attention mechanisms and scalability. Supplementary material, including further discussion of fev-bench results, inference-time comparisons, varying context and horizon settings, training dynamics, and qualitative showcases, is provided in Appendix D and Appendix E.

5.1 EXPERIMENTAL SETTING

To comprehensively evaluate the performance of UNITIMS, we consider both Irregular Multivariate Time Series (IMTS) forecasting and regular Multivariate Time Series (MTS) forecasting.

IMTS Forecasting We follow prior work (Luo et al., 2025) and utilize the following three datasets to experiment with models' performance on the IMTS forecasting task: PhysioNet (Silva et al., 2012), Human Activity (Vidulin et al., 2010), and USHCN (Menne et al., 2016), which span healthcare, biomechanics, and climate science domains. We compare UNITIMS against 15 representative baselines, which are divided into three categories: 1) Regular MTS forecasting methods: DLinear (Zeng et al., 2022), PatchTST (Nie et al., 2023), iTransformer (Liu et al., 2024a), TimesNet (Wu et al., 2023), and TimeMixer (Wang et al., 2025). 2) IMTS forecasting methods: GRU-D (Che et al., 2018), CRU (Schirmer et al., 2022), mTAND (Shukla & Marlin, 2021), NeuralFlow (Biloš et al., 2021), Latent-ODE (Rubanova et al., 2019), T-PatchGNN (Zhang et al., 2024) and Hi-Patch (Luo et al., 2025). 3) TSFM methods: Chronos (Ansari et al., 2024), Chronos-2 (Ansari et al., 2025), Moirai (Woo et al., 2024), and Sundial (Liu et al., 2025).

Regular MTS Forecasting We evaluate the performance of UNITIMS on regular multivariate time series forecasting using the point forecasting benchmark (Liu et al., 2025), which comprises the 6 datasets, including ETTm1, ETTm2, ETTh1, ETTh2, ECL, Weather, covering 4 different prediction horizons. We compare UNITIMS against SOTA TSFM baselines: Sundial (Liu et al., 2025), MOIRAI (Woo et al., 2024), Chronos (Ansari et al., 2024) and TimesFM (Das et al., 2024).

To ensure a thorough and fair comparison, we include all available model sizes for each TSFM baseline that provides multiple variants. For comparability and fairness considerations, three categories of baselines are evaluated under deliberately different protocols. Mean Squared Error (MSE) and Mean Absolute Error (MAE) are used as primary forecasting metrics, computed only over the observed target values and averaged across variables; the full definitions for regular MTS and IMTS are given in Appendix C.1.

5.2 IMPLEMENT DETAILS

We implement UNITIMS in three configurations: Small (60M), Base (403M), and Large (1.3B); the detailed layer and dimension settings are reported in Appendix B.3. All models are trained with the AdamW optimizer under a linear-warmup cosine-decay schedule on NVIDIA H100-NVL GPUs, with full optimization, sampling, and inference details provided in Appendix B.

[1] https://anonymous.4open.science/r/UniTIMS-27A4

Table 1: Overall performance is evaluated by MSE and MAE (mean ± std). The best and second-best results are highlighted in **red bold** and blue underline, respectively. Ties are determined at the reported precision. The TSFM model is evaluated under the zero-shot setting. The IMTS forecasting methods use the full-shot setting, and their results are obtained from Luo et al. (2025).

	PhysioNet		USHCN		Human Activity	
	MSE ($\times 10^{-3}$)	MAE ($\times 10^{-2}$)	MSE ($\times 10^{-1}$)	MAE ($\times 10^{-1}$)	MSE ($\times 10^{-3}$)	MAE ($\times 10^{-2}$)
DLinear	41.86 ± 0.05	15.52 ± 0.03	6.21 ± 0.00	3.88 ± 0.02	4.03 ± 0.01	4.21 ± 0.01
PatchTST	5.21 ± 0.33	5.10 ± 0.20	5.88 ± 0.10	3.66 ± 0.13	25.56 ± 4.42	10.90 ± 1.08
iTransformer	3.97 ± 0.10	4.30 ± 0.08	6.17 ± 0.07	4.18 ± 0.55	53.55 ± 19.59	16.87 ± 4.51
TimesNet	3.79 ± 0.05	4.28 ± 0.04	5.62 ± 0.12	3.56 ± 0.12	9.30 ± 0.70	5.50 ± 0.34
TimeMixer	4.97 ± 0.31	5.02 ± 0.16	5.88 ± 0.10	3.59 ± 0.07	13.98 ± 0.31	6.88 ± 0.09
GRU-D	5.76 ± 0.34	4.53 ± 0.15	5.17 ± 0.06	3.21 ± 0.05	3.94 ± 0.29	4.37 ± 0.21
CRU	3.03 ± 0.04	3.60 ± 0.04	5.15 ± 0.50	3.18 ± 0.03	6.43 ± 0.62	4.51 ± 0.16
mTAND	3.14 ± 0.09	3.71 ± 0.06	5.03 ± 0.05	3.00 ± 0.06	6.18 ± 0.31	4.44 ± 0.19
NeuralFlow	4.29 ± 0.63	4.61 ± 0.43	5.41 ± 0.05	3.35 ± 0.06	7.68 ± 0.37	4.84 ± 0.19
Latent-ODE	3.32 ± 0.10	3.91 ± 0.08	5.16 ± 0.04	3.21 ± 0.07	6.85 ± 0.28	4.77 ± 0.17
T-PatchGNN	2.79 ± 0.09	3.24 ± 0.06	5.00 ± 0.03	3.07 ± 0.05	5.06 ± 0.10	3.75 ± 0.07
Hi-Patch	2.57 ± 0.02	3.11 ± 0.03	4.94 ± 0.05	2.96 ± 0.04	4.86 ± 0.03	3.62 ± 0.07
Chronos-2	3.94 ± 0.02	4.12 ± 0.04	4.71 ± 0.06	2.83 ± 0.03	4.65 ± 0.10	3.54 ± 0.07
Moirai-Base	4.31 ± 0.30	4.71 ± 0.43	6.04 ± 0.10	4.11 ± 0.14	5.72 ± 0.06	4.63 ± 0.10
Moirai-Large	4.19 ± 0.26	4.27 ± 0.04	6.00 ± 0.11	4.09 ± 0.12	5.68 ± 0.13	5.59 ± 0.10
Sundial-Base	4.79 ± 0.02	4.91 ± 0.14	5.78 ± 0.06	3.83 ± 0.05	5.79 ± 0.12	5.67 ± 0.33
Sundial-Large	4.76 ± 0.04	4.82 ± 0.13	5.65 ± 0.04	3.76 ± 0.04	5.76 ± 0.13	5.57 ± 0.31
UniTIMS-Small	3.41 ± 0.04	3.89 ± 0.05	4.92 ± 0.03	2.91 ± 0.04	4.63 ± 0.12	3.44 ± 0.06
UniTIMS-Base	**2.37 ± 0.03**	**2.74 ± 0.03**	3.62 ± 0.03	**2.29 ± 0.03**	3.41 ± 0.10	2.07 ± 0.07
UniTIMS-Large	2.52 ± 0.03	3.07 ± 0.03	**3.53 ± 0.02**	2.32 ± 0.02	**2.44 ± 0.11**	**1.61 ± 0.05**

Table 2: Point forecasting results of regular MTS. Results are averaged across prediction lengths 96, 192, 336, 720. Best results are highlighted in **red bold**, and second-best results are highlighted in blue underline. Ties are determined at the reported precision. Baseline results are obtained from Liu et al. (2025).

	ETTm1		ETTm2		ETTh1		ETTh2		ECL		Weather	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
TimesFM	0.433	0.418	0.328	0.346	0.473	0.443	0.392	0.406	-	-	-	-
Chronos-Base	0.645	0.500	0.310	0.350	0.591	0.468	0.405	0.410	0.214	0.278	0.292	0.315
Chronos-Large	0.555	0.465	0.295	0.338	0.588	0.466	0.455	0.427	0.204	0.273	0.279	0.306
Moirai-Base	0.406	0.385	0.311	0.337	0.417	0.419	0.362	0.382	0.187	0.274	0.287	0.281
Moirai-Large	0.422	0.391	0.329	0.343	0.480	0.439	0.367	0.377	0.186	0.270	0.264	0.273
Sundial-Base	0.336	0.377	0.258	0.320	0.411	0.434	**0.333**	0.387	0.169	0.265	**0.234**	**0.270**
Sundial-Large	**0.331**	**0.369**	0.254	**0.315**	**0.395**	0.420	0.334	0.387	0.166	0.262	0.238	0.275
UniTIMS-Small	0.400	0.381	0.283	0.331	0.431	0.436	0.354	0.378	0.169	0.269	0.261	0.283
UniTIMS-Base	0.384	0.374	0.271	0.318	0.420	0.422	0.351	**0.374**	0.167	0.260	0.254	0.279
UniTIMS-Large	0.362	0.371	**0.253**	**0.315**	0.403	**0.418**	0.352	0.376	**0.162**	**0.257**	0.241	0.274

5.3 Overall Performance

5.3.1 IMTS Forecasting

Table 1 reports the zero-shot forecasting performance on three IMTS datasets. The experimental results show that UniTIMS-Base and UniTIMS-Large achieve superior performance, consistently ranking best or second-best across all evaluation metrics, with several findings: 1) The UniTIMS-Small using fewer parameters achieves the averaged MSE reduction of 3.14% and the averaged MAE reduction of 1.86% compared to the state-of-the-art TSFM baseline, Chronos-2. This improvement is mainly attributed to the widespread use of positional encodings in existing TSFMs, which limits their ability to effectively capture the inherent irregularities and inter-variable dependencies in IMTS. 2) The zero-shot forecast results of UniTIMS-Base and UniTIMS-Large surpass those of specialized full-shot models extensively trained on the target datasets, demonstrating that leveraging high-quality, large-scale data such as VersaTSA can effectively activate model parameters and enhance the generalization performance of TSFMs.

To further evaluate robustness under different observation-to-forecast ratios, we evaluate UniTIMS on Human Activity and PhysioNet under both balanced and short-history, long-horizon configura-

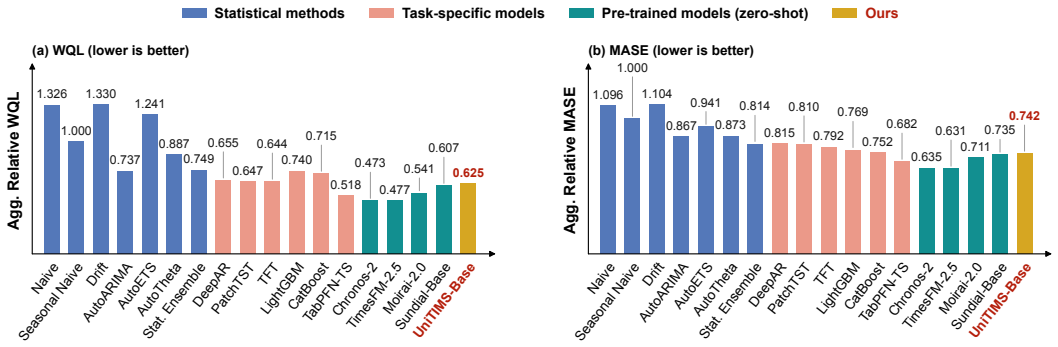

Figure 2: Fev-bench results with (a) aggregate relative WQL and (b) aggregate relative MASE. Colors distinguish statistical methods, task-specific models, and pre-trained models evaluated zero-shot. UniTIMS-Base is shown in gold, with its name and scores highlighted in red bold. Lower values indicate better performance.

tions, with full results reported in Appendix D.2 (Tables 10 and 11). The proposed variants maintain strong performance across all settings, indicating stable adaptability to shifting temporal windows.

5.3.2 Regular MTS Forecasting

Table 2 reports regular MTS forecasting performance averaged over the four prediction horizons in the point forecasting benchmark, with the full per-horizon results deferred to Appendix D.1. Remarkably, UniTIMS achieves state-of-the-art average performance on the ECL dataset and competitive performance on ETTh2. It is important to note that our model was pre-trained on the irregular VersaTSA dataset without the aid of synthetic generation tools like KernelSynth (Ansari et al., 2024), demonstrating that large-scale irregular real-world data can also support competitive transfer to regular forecasting. Meanwhile, UniTIMS-Large achieves the lowest reported average MSE on ETTm2 and ties with Sundial-Large for the lowest average MAE at the reported precision.

On the open leaderboard fev-bench (100 forecasting tasks, none seen during training), UniTIMS demonstrates robust and trustworthy performance despite being designed for irregular time series, and all pre-trained models under zero-shot inference show clear superiority over task-specific and statistical baselines. Figure 2 reports the aggregate relative WQL and MASE results, with further discussion in Appendix D.3. We further compare the wall-clock inference time of representative TSFMs under the same hardware configuration in Appendix D.4: UniTIMS achieves competitive inference speed relative to foundation models of comparable scale, indicating that the time-aware hybrid attention design does not introduce prohibitive computational overhead.

5.4 Ablation Study

We present a series of ablation results in Figure 3 on PhysioNet, Human Activity, and USHCN datasets, starting from the default UniTIMS-Small. Here, **w/ PE** adds sinusoidal positional encoding (PE), while **w/ [MASK]** uses a special `[MASK]` token to supply query information at the corresponding positions. First, we ablate the Patch-Aware Time Embedding component (**w/o Patch-Aware**). Instead of concatenating the vector representations of multiple timestamps within a patch and projecting them via a linear layer, we employed the average of these vector representations to represent the temporal features of the patch. We observed a significant deterioration in normalized MAE when removing the Patch-aware time embedding. Subsequently, we conducted ablation on the Time Bias Attention module (**w/o Time Bias Attention**), again noting an increase in normalized MAE, indicating diminished model performance. Next, we removed the learnable layer-specific scaling factor λ of the temporal bias within Time Bias Attention (**w/o** λ), observing a slight shift in MAE. Finally, we conducted ablation experiments on the Hybrid Causal Mask (**w/o Hybrid Causal Mask**). Upon replacing it with the Full Attention Mask, experimental results demonstrated a significant decline in the model's prediction accuracy. From the ablation study, we can conclude that each component in our framework contributes to the final improvements for irregular multivariate time series forecasting tasks.

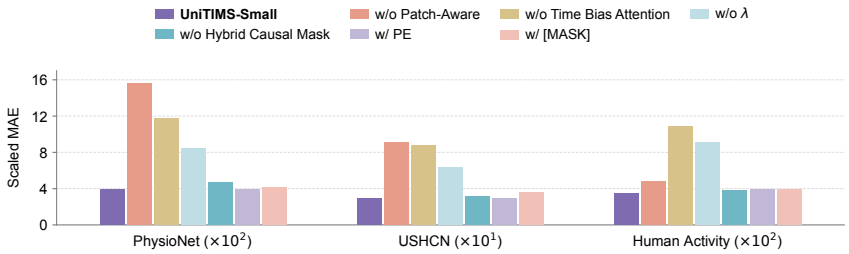

Figure 3: Ablation results of UniTIMS-Small on three IMTS datasets. MAE is scaled as indicated on the horizontal axis; lower is better. PE denotes sinusoidal positional encoding; `[MASK]` is a special token used to supply query information at corresponding positions.

We observed that introducing additional positional information via PE did not yield significant performance gains for the model. Furthermore, the **w/ [MASK]** variant demonstrated that this backfilling approach exhibited a suppressing effect on the model's capabilities.

5.5 Hybrid Causal Mask Analysis

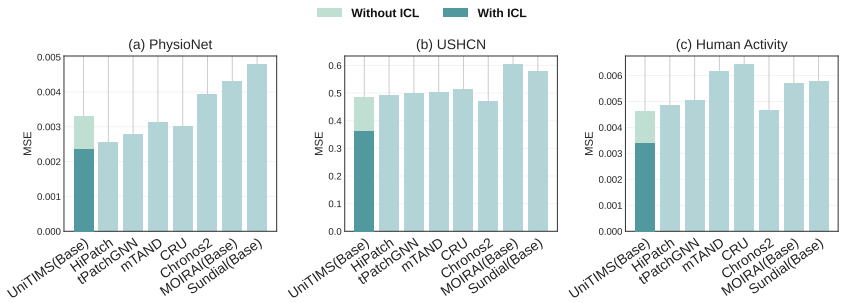

Figure 4: The analysis of Hybrid Causal Mask on three datasets of IMTS. Evaluated by MSE, where lower values indicate better performance.

We design a comparative study to examine the contextual modeling capability of the proposed Hybrid Causal Mask. Specifically, we compare two settings for IMTS forecasting: (i) *without ICL* by replacing the Hybrid Causal Mask with a standard Causal Attention Mask, where multiple variate time series are provided in a channel-independent (CI) manner; and (ii) *with ICL* by adopting the Hybrid Causal Mask and concatenating multiple multivariate time series into a single sequence as contextual input. Figure 4 reports the results evaluated by MSE. Across three irregularly sampled time-series datasets, the Hybrid Causal Mask consistently achieves better performance, indicating a stronger ability to leverage contextual information.

Qualitative zero-shot forecasting visualizations of UniTIMS-Large on both IMTS and regular MTS datasets are provided in Appendix E, where the predicted curves preserve the overall evolution of the ground truth under irregular intervals, asynchronous observations, and varying frequencies.

6 Conclusion

We introduce UniTIMS, a time-aware decoder-only Transformer with a hybrid causal mask for irregular multivariate time series forecasting (IMTS), with a particular focus on informative missingness and complex contextual dependencies issues in IMTS. We also introduced VersaTSA, a large collection of IMTS data for pretraining. Experiments on several datasets and benchmarks demonstrate UniTIMS achieves superior zero-shot performance on IMTS and remains competitive on regular multivariate time series. Despite its strong out-of-the-box performance on IMTS, the inference efficiency of UniTIMS remains an important practical consideration. Incorporating inference acceleration techniques like KV cache may significantly improve its efficiency and scalability.

Reproducibility statement

The model architecture and training objective are described in Sections 4.1–4.3 and Section 4.4. Dataset and baseline choices, evaluation protocols, metrics, model configurations, optimization hyperparameters, and inference details are reported in Section 5.1 and Section 5.2. The anonymous project repository referenced in Section 4.4 provides additional material for reproducing the data-processing pipeline.

AI use statement

Generative AI tools were used to assist with migrating the manuscript source to the ICLR 2027 LaTeX format. No scientific claims, experimental results, or conclusions were generated as part of this formatting step. The authors remain responsible for reviewing the final manuscript and should update this statement if generative AI tools were used elsewhere in the research or writing process.

References

Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Bernie Wang. Chronos: Learning the Language of Time Series. *Transactions on Machine Learning Research*, May 2024. ISSN 2835-8856.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From Univariate to Universal Forecasting, October 2025.

Marin Biloš, Johanna Sommer, Syama Sundar Rangapuram, Tim Januschowski, and Stephan Günnemann. Neural Flows: Efficient Alternative to Neural ODEs, October 2021.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models, July 2022.

Ching Chang, Jeehyun Hwang, Yidan Shi, Haixin Wang, Wen-Chih Peng, Tien-Fu Chen, and Wei Wang. Time-IMM: A Dataset and Benchmark for Irregular Multimodal Multivariate Time Series, October 2025.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports*, 8(1): 6085, April 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-24271-9.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural Ordinary Differential Equations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

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

Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: A review. *Data Mining and Knowledge Discovery*, 33(4):917–963, July 2019. ISSN 1384-5810, 1573-756X. doi: 10.1007/s10618-019-00619-1.

Pedro J. García-Laencina, José-Luis Sancho-Gómez, and Aníbal R. Figueiras-Vidal. Pattern classification with missing data: A review. *Neural Computing and Applications*, 19(2):263–282, March 2010. ISSN 1433-3058. doi: 10.1007/s00521-009-0295-6.

Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. TimeGPT-1, May 2024.

Robert W. Johnson. Applications of the beta distribution part 1: Transformation group approach, 2016. URL https://arxiv.org/abs/1307.6437.

Seyed Mehran Kazemi, Rishab Goel, Sepehr Eghbali, Janahan Ramanan, Jaspreet Sahota, Sanjay Thakur, Stella Wu, Cathal Smyth, Pascal Poupart, and Marcus Brubaker. Time2Vec: Learning a Vector Representation of Time, July 2019.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting, March 2024a.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative Pre-trained Transformers Are Large Time Series Models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 32369–32399. PMLR, July 2024b.

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, 2025.

Yicheng Luo, Bowen Zhang, Zhen Liu, and Qianli Ma. Hi-Patch: Hierarchical Patch GNN for Irregular Multivariate Time Series. In *Forty-Second International Conference on Machine Learning*, June 2025.

M. J. Menne, Jr C. N. Williams, and R. S. Vose. Long-Term Daily and Monthly Climate Records from Stations Across the Contiguous United States (U.S. Historical Climatology Network). 2016.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers, March 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. URL https://arxiv.org/abs/1912.01703.

Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.

Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, Rachel Prudden, Amol Mandhane, Aidan Clark, Andrew Brock, Karen Simonyan, Raia Hadsell, Niall Robinson, Ellen Clancy, Alberto Arribas, and Shakir Mohamed. Skillful Precipitation Nowcasting using Deep Generative Models of Radar. *Nature*, 597(7878):672–677, September 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-03854-z.

K. Rehfeld, N. Marwan, J. Heitzig, and J. Kurths. Comparison of correlation analysis techniques for irregularly sampled time series. *Nonlinear Processes in Geophysics*, 18(3):389–404, June 2011. ISSN 1023-5809. doi: 10.5194/npg-18-389-2011.

Neal Richardson, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Bryce Mecum, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. *Arrow: Integration to 'Apache' 'Arrow'*, 2025.

Yulia Rubanova, Ricky T. Q. Chen, and David Duvenaud. Latent ODEs for Irregularly-Sampled Time Series, July 2019.

DONALD B. RUBIN. Inference and missing data. *Biometrika*, 63(3):581–592, December 1976. ISSN 0006-3444. doi: 10.1093/biomet/63.3.581.

Mona Schirmer, Mazin Eltayeb, Stefan Lessmann, and Maja Rudolph. Modeling Irregular Time Series with Continuous Recurrent Units, July 2022.

Omer Berat Sezer, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. Financial time series forecasting with deep learning : A systematic literature review: 2005–2019. *Applied Soft Computing*, 90:106181, May 2020. ISSN 1568-4946. doi: 10.1016/j.asoc.2020.106181.

Satya Narayan Shukla and Benjamin M. Marlin. Multi-Time Attention Networks for Irregularly Sampled Time Series, June 2021.

Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. Predicting In-Hospital Mortality of ICU Patients: The PhysioNet/Computing in Cardiology Challenge 2012. *Computing in cardiology*, 39:245–248, 2012. ISSN 2325-8861.

Rafael T. Sousa, Lucas A. Pereira, and Anderson S. Soares. Improving Irregularly Sampled Time Series Learning with Dense Descriptors of Time, March 2020.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, February 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.127063.

Yu Sun, Xinyu Yang, Shaoxu Song, Ying Zhang, and Xiaojie Yuan. Collaborative Imputation for Multivariate Time Series with Convergence Guarantee . In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, pp. 1111–1124, Los Alamitos, CA, USA, May 2025. IEEE Computer Society. doi: 10.1109/ICDE65448.2025.00088. URL https://doi.ieeecomputersociety.org/10.1109/ICDE65448.2025.00088.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023.

Vedrana Vidulin, Mitja Lustrek, Bostjan Kaluza, Rok Piltaver, and Jana Krivec. Localization Data for Person Activity. UCI Machine Learning Repository, 2010.

Shiyu Wang, Jiawei Li, Xiaoming Shi, Zhou Ye, Baichuan Mo, Wenze Lin, Shengtong Ju, Zhixuan Chu, and Ming Jin. TimeMixer++: A General Time Series Pattern Machine for Universal Predictive Analysis, March 2025.

Brian J. Wells, Kevin M. Chagin, Amy S. Nowacki, and Michael W. Kattan. Strategies for Handling Missing Data in Electronic Health Record Derived Data. *eGEMs*, 1(3):1035, December 2013. ISSN 2327-9214. doi: 10.13063/2327-9214.1035.

Ian R. White, Patrick Royston, and Angela M. Wood. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399, 2011. ISSN 1097-0258. doi: 10.1002/sim.4067.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified Training of Universal Time Series Forecasting Transformers. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 53140–53164. PMLR, July 2024.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. http://arxiv.org/abs/2210.02186, April 2023.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are Transformers Effective for Time Series Forecasting?, August 2022.

Qianru Zhang, Honggang Wen, Ming Li, Dong Huang, Siu-Ming Yiu, Christian S. Jensen, and Pietro Liò. Autohformer: Efficient hierarchical autoregressive transformer for time series prediction, 2025a. URL https://arxiv.org/abs/2506.16001.

Weijia Zhang, Chenlong Yin, Hao Liu, Xiaofang Zhou, and Hui Xiong. Irregular Multivariate Time Series Forecasting: A Transformable Patching Graph Neural Networks Approach. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 60179–60196. PMLR, July 2024.

Yudong Zhang, Xu Wang, Xuan Yu, Zhengyang Zhou, Xing Xu, Lei Bai, and Yang Wang. DIFFODE: Neural ODE with Differentiable Hidden State for Irregular Time Series Analysis . In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, pp. 1–14, Los Alamitos, CA, USA, May 2025b. IEEE Computer Society. doi: 10.1109/ICDE65448.2025.00160. URL https://doi.ieeecomputersociety.org/10.1109/ICDE65448.2025.00160.

Ying Zheng, Yangfan Jiang, and Kian-Lee Tan. Causalpre: Scalable and effective data preprocessing for causal fairness, 2026. URL https://arxiv.org/abs/2509.15199.

A Versatile Time Series Archive

Versa collects multiple datasets of both regular MTS and IMTS, and stores them in the form of three lists (U, T, C) to facilitate pretraining of TSFMs in IMTS. For regular MTS datasets, we filter out missing values and reconstruct the timestamps of each sequence using the provided start field. We further details on key attributes of each dataset, including the domain, the number of time series, and the number of observations. And the frequency is reported for regular MTS.

LOTSA Large-scale Open Time Series Archive (Woo et al., 2024) provides a large-scale collection of datasets spanning nine domains, ranging from Climate to Transport, covering a total of 231B observations. To enable the pretraining of Time Series Foundation Models (TSFMs), Woo et al. (2024) constructed and opened this extensive data archive, representing a pioneering effort in this direction. We reconstruct the timestamps using the provided start and freq (frequency) fields, and organize the data into sequences of triplets (u, τ, c) according to the variable associated with each time point. Triplets with missing values in u are subsequently removed from the sequences.

Table 4: Datasets and key properties from Time-IMM.

Dataset	Domain	# Time Series	# Obs.
GDELT	Web	8	193,205
RepoHealth	Health	4	67,830
FNSPID	Econ/Fin	10	209,688
ClusterTrace	CloudOps	3	69,001
StudentLife	Human Activity	20	1,743
ILINet	Network	1	4,918
CESNET	Network	30	51,107
EPA-Air	Environment	8	49,552

Time-IMM Time-IMM (Chang et al., 2025) provides a dataset specifically designed to capture cause-driven irregularity in multimodal multivariate time series. And it ranges from a varying taxonomy of irregularity in time series. (a) **Trigger-Based Irregularities**: GDELT and RepoHealth; (b) **Constraint-Based Irregularities**: FNSPID, ClusterTrace(different year from CloudOps LOTSA), and StudentLife; (c) **Artifact-Based Irregularities**: ILINet, CESNET, and EPA-Air.

Table 3: Datasets and key properties from LOTSA.

Dataset	**Domain**	**Frequency**	**# Time Series**	**# Obs.**
BDG-2 Panther	Energy	H	105	919,800
BDG-2 Fox	Energy	H	135	2,324,568
BDG-2 Rat	Energy	H	280	4,728,288
BDG-2 Bear	Energy	H	91	1,482,312
Low Carbon London	Energy	H	0	9,543,348
SMART	Energy	H	5	95,709
IDEAL	Energy	H	219	1,265,672
Sceaux	Energy	H	1	34,223
Borealis	Energy	H	15	83,269
Buildings900K	Energy	H	1,792,328	15,702,590,000
Azure VM Traces 2017	CloudOps	5T	159,472	885,522,908
Borg Cluster Data 2011	CloudOps	5T	143,386	537,552,854
Alibaba Cluster Trace 2018	CloudOps	5T	58,409	95,192,530
Taxi	Transport	30T	67,984	54,999,060
Uber TLC Daily	Transport	D	262	47,087
Uber TLC Hourly	Transport	H	262	1,129,444
Wiki-Rolling	Web	D	47,675	40,619,100
M5	Sales	D	30,490	58,327,370
LargeST	Transport	5T	42,333	4,452,510,528
PEMS03	Transport	5T	358	9,382,464
PEMS04	Transport	5T	307	5,216,544
PEMS07	Transport	5T	883	24,921,792
PEMS08	Transport	5T	170	3,035,520
PEMS Bay	Transport	5T	325	16,937,700
Los-Loop	Transport	5T	207	7,094,304
Loop Seattle	Transport	5T	323	33,953,760
SZ-Taxi	Transport	15T	156	464,256
Beijing Subway	Transport	30T	276	248,400
SHMetro	Transport	15T	288	1,934,208
HZMetro	Transport	15T	80	146,000
Rotterdam	Transport	2T	208	4,813,536
Q-Traffic	Transport	15T	45,148	264,386,688

Table 5: Datasets and key properties from OWP.

Dataset	Domain	# Time Series	# Obs.
OWP	Energy	10	98,106,540

OWP Offshore Wind Power and Wave Height measurements is a dataset released by us, consisting of 98,106,540 observations across 10 variables (Details in 6). It includes physical measurements related to offshore wind power, specifically a set of strain-gauge load components obtained through linear transformations, as well as wave height information derived from processed radar data, which captures both the magnitude and direction of ocean waves.

Table 3: Cont. Datasets and key properties from LOTSA.

... Continued from previous page				
Dataset	**Domain**	**Frequency**	**# Time Series**	**# Obs.**
KDD Cup 2022	Energy	10T	134	4,727,519
GoDaddy	Econ/Fin	M	3,135	128,535
Favorita Sales	Sales	D	111,840	139,179,538
Favorita Transactions	Sales	D	54	84,408
Restaurant	Sales	D	216	76,573
Hierarchical Sales	Sales	D	118	212,164
China Air Quality	Nature	H	437	5,739,234
Beijing Air Quality	Nature	H	12	420,768
Residential Load Power	Energy	T	271	145,994,559
Residential PV Power	Energy	T	233	125,338,950
CDC Fluview ILINet	Healthcare	W	75	63,903
CDC Fluview WHO NREVSS	Healthcare	W	74	41,760
Project Tycho	Healthcare	W	1,258	1,377,707
London Smart Meters	Energy	30T	5,520	166,238,880
Wind Farms	Energy	T	337	172,165,370
Wind Power	Energy	4S	1	7,397,147
Solar Power	Energy	4S	1	7,397,222
Oikolab Weather	Climate	H	8	800,456
Elecdemand	Energy	30T	1	17,520
Covid Mobility	Transport	D	362	148,602
Kaggle Web Traffic Weekly	Web	W	145,063	16,537,182
Extended Web Traffic	Web	D	145,063	370,926,091
M1 Yearly	Econ/Fin	Y	106	3,136
M1 Quarterly	Econ/Fin	Q	198	9,854
M1 Monthly	Econ/Fin	M	617	44,892
M3 Yearly	Econ/Fin	Y	645	18,319
M3 Quarterly	Econ/Fin	Q	756	37,004
M3 Monthly	Econ/Fin	M	1,428	141,858
M3 Other	Econ/Fin	Q	174	11,933
M4 Yearly	Econ/Fin	Y	22,739	840,644
M4 Quarterly	Econ/Fin	Q	24,000	2,214,108
M4 Monthly	Econ/Fin	M	48,000	10,382,411
M4 Weekly	Econ/Fin	W	359	366,912
M4 Hourly	Econ/Fin	H	414	353,500
M4 Daily	Econ/Fin	D	4,227	9,964,658
NN5 Daily	Econ/Fin	D	111	81,585
NN5 Weekly	Econ/Fin	W	111	11,655
Tourism Yearly	Econ/Fin	Y	419	11,198
Tourism Quarterly	Econ/Fin	Q	427	39,128
Tourism Monthly	Econ/Fin	M	366	100,496
CIF 2016	Econ/Fin	M	72	6,334
Traffic Weekly	Transport	W	862	82,752
Traffic Hourly	Transport	H	862	14,978,112
Australian Electricity Demand	Energy	30T	5	1,153,584
Rideshare	Transport	H	2,304	859,392
Saugeen	Nature	D	1	23,711
Sunspot	Nature	D	1	73,894
Temperature Rain	Nature	D	32,072	22,290,040
Vehicle Trips	Transport	D	329	32,512

Table 3: Cont. Datasets and key properties from LOTSA.

... Continued from previous page				
Dataset	**Domain**	**Frequency**	**# Time Series**	**# Obs.**
Weather	Climate	D	3,010	42,941,700
Car Parts	Sales	M	2,674	104,286
FRED MD	Econ/Fin	M	107	76,612
Pedestrian Counts	Transport	H	66	3,130,762
Hospital	Healthcare	M	767	55,224
COVID Deaths	Healthcare	D	266	48,412
KDD Cup 2018	Energy	H	270	2,897,004
Bitcoin	Econ/Fin	D	18	74,824
US Births	Healthcare	D	1	7,275
CMIP6	Climate	6H	1,351,680	1,973,453,000
ERA5	Climate	H	245,760	2,146,959,000
Covid19 Energy	Energy	H	1	31,912
GEF12	Energy	H	20	788,280
GEF14	Energy	H	1	17,520
GEF17	Energy	H	8	140,352
PDB	Energy	H	1	17,520
Spanish	Energy	H	1	35,064
BDG-2 Hog	Energy	H	24	421,056
BDG-2 Bull	Energy	H	41	719,304
BDG-2 Cockatoo	Energy	H	1	17,544
ELF	Energy	H	1	21,792
Subseasonal	Climate	D	862	14,097,148
Subseasonal Precipitation	Climate	D	862	9,760,426

Table 6: Description of Variables in the OWP Dataset

Variable	Unit	Description	#Example Value
Mtbx	kNm	Derived from tower-bottom x-axis sensor measurements	5186.6229
Mtby	kNm	Derived from tower-bottom y-axis sensor measurements	-60491.49
Mtmx	kNm	Derived from tower-middle x-axis sensor measurements	6149.6357
Mtmy	kNm	Derived from tower-middle x-axis sensor measurements	-33853.17
Mttx	kNm	Derived from tower-top x-axis sensor measurements	-2281.088
Mtty	kNm	Derived from tower-top y-axis sensor measurements	18087.884
Mttz	kNm	Derived from tower-top z-axis sensor measurements	7147.9924
Hm0_M	cm	Wave Height	158.24474
Th0	deg	azimuth angle	5.263157894736842

B Experimental Details

B.1 UniTIMS Training Strategy

The pre-training process is conducted on the VersaTSA archive to equip the model with universal representations under irregular temporal fluctuations, aiming to: 1) capture complex dependencies in irregular samples, and 2) enable robust zero-shot forecasting via large-scale autoregressive training. We implement a training strategy based on dynamic window partitioning. Specifically, for each training sample extracted from the long-term sequences in VersaTSA, we randomly select a partition point followed by the Beta distribution (Johnson, 2016), splitting the patch sequence $\{p_1, p_2, \ldots, p_{m+n}\}$ into a historical prefix segment $\mathcal{P}$ and a target segment $\mathcal{T}$. For example, given a partition point m sampled from a Beta distribution, $\mathcal{P}$ denotes the index set corresponding to the combined historical observations and covariates, i.e., $\mathcal{Y}_{t_1:t_m} \cup \mathcal{Z}_{t_1:t_{m+n}}$. The target segment $\mathcal{T}$

includes the indices of the query set $Q = \{(t, c) \mid t \in (t_m, t_{m+n}], c \in \mathcal{C}_\mathcal{Y}\}$ and the special forecasting token `[F]`. During autoregressive generation, the prefix serves as the historical context where bidirectional information flow is permitted, while the target is forecasted causally, step-by-step.

Table 7: MTS and IMTS task settings with time-slicing and input configurations. $|p_m| = l_p$ denotes the same sequence length, while ΔT_i and Δt_i represent the time durations for MTS and IMTS within each patch, respectively. The inputs for target variables $\mathcal{X}_y$ are projected alongside covariates $\mathcal{X}_z$ into a unified d-dimensional representation space.

Task type	Time slicing	Inputs
MTS	$\lvert p_m\rvert = l_p, \Delta T_i = \Delta T_j$	$\begin{bmatrix} \underbrace{\begin{matrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_{M_y} \\ \boldsymbol{p}_1 & \cdots & \boldsymbol{p}_{M_y} \end{matrix}}_{\text{Target Variables } \mathcal{X}_y} & \underbrace{\begin{matrix} \boldsymbol{x}_{M_y+1} & \cdots & \boldsymbol{x}_M \\ \boldsymbol{p}_{M_y+1} & \cdots & \boldsymbol{p}_M \end{matrix}}_{\text{Covariates } \mathcal{X}_z} \end{bmatrix} \in \mathbb{R}^{2\times M\cdot d}$
IMTS	$\lvert p_m\rvert = l_p, \Delta t_i \neq \Delta t_j$	$\begin{bmatrix} \underbrace{\begin{matrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_{M_y} \\ \boldsymbol{p}_{\Delta t_1} & \cdots & \boldsymbol{p}_{\Delta t_{M_y}} \end{matrix}}_{\text{Target Variables } \mathcal{X}_y} & \underbrace{\begin{matrix} \boldsymbol{x}_{M_y+1} & \cdots & \boldsymbol{x}_M \\ \boldsymbol{p}_{\Delta t_{M_y+1}} & \cdots & \boldsymbol{p}_{\Delta t_M} \end{matrix}}_{\text{Covariates } \mathcal{X}_z} \end{bmatrix} \in \mathbb{R}^{2\times M\cdot d}$

Crucially, our model can fit both regular and irregular MTS forecasting tasks. We summarize the task settings and input representations of MTS and IMTS under the UNITIMS framework in Table 7. For both tasks, the input sequence is divided into non-overlapping patches where each patch contains the same number of sequence points ($|p_m| = l_p$). The difference is that MTS patches span uniform time durations ($\Delta T_i = \Delta T_j$), whereas IMTS patches span varying time durations ($\Delta t_i \neq \Delta t_j$) due to non-uniform sampling. Finally, inputs from both target variables $\mathcal{X}_y$ and covariates $\mathcal{X}_z$ are projected into a unified d-dimensional representation space.

B.2 Details of Evaluation Setups

We summarize the setup for three categories of baselines here to avoid misinterpretation of the results in Sec. 5.3.1.

- IMTS-specific methods (GRU-D, mTAND, T-PatchGNN, Hi-Patch, etc.) are evaluated under the *full-shot* setting, i.e., trained directly on each target dataset, and their numbers are taken from (Luo et al., 2025) using the identical splits and metric definitions reported there.
- TSFMs (Chronos-2, Moirai, Sundial) and UniTIMS are evaluated under the *zero-shot* setting on the same splits; for IMTS inputs, regular-grid TSFMs operate on the same observed-only timestamps via their public inference interfaces without any extra fine-tuning.
- UniTIMS is pretrained exclusively on VersaTSA, which strictly excludes the three IMTS evaluation datasets and the six regular MTS evaluation datasets (see the hold-out protocol in Sec. 4.4). Despite the larger pretraining scale of UniTIMS relative to full-shot baselines, all methods are evaluated on the *same* test windows and use *identical* MSE/MAE definitions over the observed targets, so the gains reported in Table 1 reflect the model's representational capability rather than benefiting from extra access to evaluation data.

All experiments are implemented using PyTorch (Paszke et al., 2019) and conducted on NVIDIA H100-NVL GPUs using TF32 precision. All models are trained with a global batch size of 1024 and with 8 attention heads. We utilize the AdamW optimizer with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.98$, weight decay of 0.1, and a peak learning rate of 1×10^{-3}. The learning rate schedule includes a linear warmup over the first 10,000 steps, followed by cosine decay. **Pretrain.** During the pretraining stage, we partition the dataset by time-series indices rather than constructing samples via sliding windows, which alleviates the burden on the sampler. We then randomly sample subsequences from each time series with a fixed window size of 3840, where the sampling positions are drawn from a Beta distribution with parameters $(1.6, 2.4)$. **Inference.** During inference, for any time segment

whose length is insufficient to form a complete patch, we pad it accordingly and ignore the padded values in the inference process.

B.3 Model Configurations

We implement UniTIMS in three configurations: Small (60M), Base (403M), and Large (1.3B). Table 8 details the number of layers, hidden dimensions, and parameter counts of each variant.

Table 8: Details of the model sizes for UniTIMS.

	Layers	d_{model}	d_{ff}	Params
UniTIMS-Small	3	512	2048	60M
UniTIMS-Base	24	1024	2048	403M
UniTIMS-Large	48	1280	5120	1.3B

C Forecasting Settings

C.1 Evaluation Metrics

Mean Squared Error (MSE) and MAE are commonly used to evaluate forecasting performance. However, their computation differs slightly between irregular multivariate time series IMTS and regular multivariate time series (MTS). In general, given a series x and its prediction $\hat{x}$, the MSE and MAE are defined as:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{x}_i - x_i)^2, \qquad \text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{x}_i - x_i\right|.$$

Due to asynchrony and missing values, different variables (including target variables and covariates) within the same IMTS may have different numbers of observed time points. As a result, for irregularly-sampled time series (IMTS) with C variates, let Ω_c denote the valid evaluated observations for variable c. The MSE and MAE are computed only over the observed target values and then averaged across variables:

$$\text{MSE} = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{|\Omega_c|}\sum_{i\in\Omega_c}(\hat{x}_i^{(c)} - x_i^{(c)})^2, \qquad \text{MAE} = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{|\Omega_c|}\sum_{i\in\Omega_c}\left|\hat{x}_i^{(c)} - x_i^{(c)}\right|.$$

C.2 Clarification on Autoregressive

In the following, we clarify how a hybrid mask can ensure autoregressive learning. Given a context sequence as the condition and a target sequence $S = (s_1, \ldots, s_n)$, any predictive model whose conditional distribution admits the factorization can be regarded as an autoregressive forecasting model, i.e.,

$$P(S \mid \text{context}) = \prod_{i=1}^{n} P(s_i \mid s_{<i}, \text{context}). \tag{8}$$

Given a context sequence as the condition and a target sequence $S = (s_1, \ldots, s_n)$, any predictive model whose conditional distribution admits the factorization: $P(S \mid \text{context}) = \prod_{i=1}^{n} P(s_i \mid s_{<i}, \text{context})$ is an autoregressive forecasting model.

An inference structure is autoregressive IFF, given a sequence of length m, the model predicts future values by unidirectional reasoning solely based on the observed historical information, without leveraging any information from future time steps, i.e., of the following form:

$$P(S) = \prod_{i=1}^{m+n} P(s_i|s_{<i}) \tag{9}$$

For the inference structure formulated, its formulation is already highly consistent with the standard autoregressive form given in Eq. (9), except for the presence of an additional conditioning term, $context$. To further examine the role of this condition, we note that time series inputs inherently follow a temporal order. Specifically, the context sequence of length m corresponds to a segment of observations that precede s_i in time (here we temporarily ignore the effect of covariates). As a result, a complete history can be constructed as:

$$y_{<i} = s_{<i} \cup context \tag{10}$$

Under this formulation, the target s_i can be regarded as a natural continuation of the historical sequence $y_{<i}$, i.e., $y_i = s_i$. Consequently, we obtain:

$$P(S \mid context) = \prod_{i=1}^{m+n} P(s_i|s_{<i}, context) = \prod_{i=1}^{m+n} P(y_i|y_{<i}) \tag{11}$$

D Supplementary Results

D.1 Full results of regular MTS Forecasting

Table 9 evaluates the zero-shot forecasting performance of various time series foundation models on regular Multivariate Time Series (MTS) tasks. Using six benchmark datasets (ETTm1, ETTm2, ETTh1, ETTh2, ECL, and Weather) across four forecast horizons (96, 192, 336, and 720), the study utilizes MSE and MAE metrics to quantify accuracy and assess the generalization capabilities of models such as Sundial, Moirai, Chronos, and TimesFM in handling unseen data distributions and long-term dependencies.

Table 9: Full results for Regular MTS point forecasting experiments. Best results are highlighted in **bold**, and second best results are underlined.

		UniTIMS-Large		**UniTIMS-Base**		**UniTIMS-Small**		**Sundial-Large**		**Sundial-Base**		**Moirai-Large**		**Moirai-Base**		**Chronos-Large**		**Chronos-Base**		**TimesFM**	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	96	0.298	**0.327**	0.295	0.338	0.305	0.348	**0.273**	0.329	0.280	0.334	0.380	0.361	0.363	0.356	0.457	0.403	0.454	0.408	0.361	0.370
	192	0.337	**0.356**	0.330	0.365	0.347	0.370	**0.312**	0.357	0.321	0.366	0.412	0.383	0.388	0.375	0.530	0.450	0.567	0.477	0.414	0.405
	336	0.371	0.380	0.375	0.395	0.405	0.400	**0.343**	**0.378**	0.350	0.389	0.436	0.400	0.416	0.392	0.577	0.481	0.662	0.525	0.445	0.429
	720	0.442	0.421	0.536	**0.398**	0.543	0.406	0.397	0.413	**0.394**	0.418	0.462	0.420	0.460	0.418	0.660	0.526	0.900	0.591	0.512	0.471
ETTm2	96	**0.165**	**0.253**	0.182	0.262	0.172	0.259	0.172	0.255	0.170	0.256	0.211	0.274	0.205	0.273	0.197	0.271	0.199	0.274	0.202	0.270
	192	**0.217**	**0.291**	0.235	0.301	0.226	0.298	0.227	0.296	0.229	0.300	0.281	0.318	0.275	0.316	0.254	0.314	0.261	0.322	0.289	0.321
	336	**0.268**	**0.328**	0.284	0.335	0.285	0.334	0.275	0.331	0.281	0.337	0.341	0.355	0.329	0.350	0.313	0.353	0.326	0.366	0.360	0.366
	720	0.362	0.388	0.383	**0.374**	0.449	0.433	**0.343**	0.378	0.351	0.387	0.485	0.428	0.437	0.411	0.416	0.415	0.455	0.439	0.462	0.430
ETTh1	96	0.350	**0.382**	0.355	0.388	0.395	0.404	**0.346**	0.383	0.348	0.385	0.381	0.388	0.376	0.392	0.441	0.390	0.440	0.393	0.414	0.404
	192	**0.385**	**0.402**	0.400	0.412	0.418	0.425	0.386	0.410	0.393	0.418	0.434	0.415	0.412	0.413	0.502	0.524	0.492	0.426	0.465	0.434
	336	**0.410**	**0.424**	0.445	0.428	0.448	0.448	**0.410**	0.426	0.422	0.440	0.485	0.445	0.433	0.428	0.576	0.467	0.550	0.462	0.503	0.456
	720	0.467	0.464	0.480	0.460	0.463	0.467	**0.438**	0.459	0.481	0.493	0.611	0.510	0.447	**0.444**	0.835	0.583	0.882	0.591	0.511	0.481
ETTh2	96	0.285	**0.322**	**0.268**	0.331	0.300	0.340	0.269	0.330	0.271	0.333	0.296	0.330	0.294	0.330	0.320	0.345	0.308	0.343	0.315	0.349
	192	0.335	0.365	**0.324**	0.372	0.330	**0.360**	0.325	0.373	0.327	0.376	0.361	0.371	0.365	0.375	0.406	0.399	0.384	0.392	0.388	0.395
	336	0.365	0.392	0.356	**0.389**	0.370	0.390	**0.354**	0.400	**0.354**	0.402	0.390	0.390	0.376	0.390	0.492	0.453	0.429	0.430	0.422	0.427
	720	0.423	0.425	0.456	**0.404**	0.416	0.422	0.389	0.443	**0.381**	0.435	0.423	0.418	0.416	0.433	0.603	0.511	0.501	0.477	0.443	0.454
ECL	96	**0.127**	**0.223**	0.131	0.226	0.136	0.235	0.130	0.227	0.132	0.229	0.153	0.241	0.160	0.250	0.152	0.229	0.154	0.231	-	-
	192	**0.144**	**0.241**	0.149	0.244	0.156	0.255	0.150	0.247	0.152	0.250	0.169	0.255	0.175	0.263	0.172	0.250	0.179	0.254	-	-
	336	**0.164**	**0.261**	0.169	0.264	0.176	0.275	0.170	0.268	0.173	0.271	0.187	0.273	0.187	0.277	0.203	0.276	0.214	0.284	-	-
	720	0.213	**0.302**	0.219	0.306	**0.208**	0.311	0.214	0.307	0.218	0.311	0.237	0.313	0.228	0.309	0.289	0.337	0.311	0.346	-	-
Weather	96	0.162	0.208	0.170	0.215	0.160	0.210	**0.157**	0.208	**0.157**	**0.205**	0.199	0.211	0.220	0.217	0.194	0.235	0.203	0.238	-	-
	192	0.211	0.255	0.225	0.255	0.212	0.255	0.207	0.256	**0.205**	**0.251**	0.246	**0.251**	0.271	0.259	0.249	0.285	0.256	0.290	-	-
	336	0.262	0.293	0.270	0.295	0.270	0.295	0.259	0.295	**0.253**	**0.289**	0.274	0.291	0.286	0.297	0.302	0.327	0.314	0.336	-	-
	720	0.329	0.340	0.351	0.351	0.402	0.372	0.327	0.342	**0.320**	**0.336**	0.337	0.340	0.373	0.354	0.372	0.378	0.397	0.396	-	-
Count	1-st	8	13	2	4	1	1	9	1	7	4	0	1	0	0	0	0	0	0	0	0
	2-nd	2	6	3	5	2	3	9	12	7	0	0	4	1	2	0	0	0	0	0	0

Table 10: Performance of varying observation and forecast horizons on Human Activity dataset. Best results are highlighted in **bold**, and second best results are underlined.

	2000ms → 2000ms		1000ms → 3000ms	
	MSE ($\times 10^{-3}$)	MAE ($\times 10^{-2}$)	MSE ($\times 10^{-3}$)	MAE ($\times 10^{-2}$)
PatchTST	7.25±0.29	6.26±0.17	8.97±1.96	6.94±0.94
iTransformer	7.49±4.72	6.08±2.17	5.58±0.04	5.13±0.05
TimesNet	5.38±0.30	5.32±0.20	9.90±0.42	7.34±1.81
TimeMixer	5.39±0.54	5.05±0.38	5.96±0.19	5.36±0.10
GRU-D	5.93±0.10	5.66±0.66	6.14±0.76	5.75±0.49
CRU	4.12±0.08	4.43±0.06	4.85±0.09	4.86±0.07
mTAND	4.38±0.37	4.59±0.29	5.29±0.32	5.12±0.23
NeuralFlow	5.47±0.49	5.35±0.28	6.01±0.91	5.66±0.60
Latent-ODE	5.04±0.46	5.11±0.29	5.48±0.21	5.33±0.14
tPatchGNN	3.71±0.20	3.89±0.16	4.56±0.08	4.32±0.06
Hi-Patch	3.29±0.04	3.70±0.04	4.21±0.08	4.25±0.07
UniTIMS-Small	3.08±0.03	3.53±0.04	3.97±0.06	4.04±0.07
UniTIMS-Base	**2.14±0.02**	**1.89±0.03**	**2.83±0.02**	**2.32±0.03**
UniTIMS-Large	2.84±0.03	2.15±0.06	3.74±0.03	2.70±0.05

Table 11: Performance of varying observation and forecast horizons on PhysioNet dataset. Best results are highlighted in **bold**, and second best results are underlined.

	36h → 12h		12h → 36h	
	MSE ($\times 10^{-3}$)	MAE ($\times 10^{-2}$)	MSE ($\times 10^{-3}$)	MAE ($\times 10^{-2}$)
PatchTST	26.13±1.75	11.25±0.29	25.63±1.51	10.7±0.38
iTransformer	56.83±21.17	17.05±5.21	54.17±17.91	17.03±4.18
TimesNet	9.43±0.53	5.55±0.20	9.26±0.18	5.58±0.10
TimeMixer	12.52±0.45	6.47±0.21	19.86±0.17	8.38±0.18
GRU-D	6.85±0.37	4.88±0.18	7.80±0.22	5.13±0.13
CRU	6.74±0.21	4.82±0.11	7.66±0.14	4.97±0.05
mTAND	5.61±0.31	4.15±0.09	7.46±0.19	4.85±0.05
NeuralFlow	8.87±1.00	5.43±0.18	7.98±0.57	5.08±0.24
Latent-ODE	6.99±0.24	4.74±0.11	7.28±0.13	4.83±0.07
tPatchGNN	4.22±0.09	3.38±0.04	6.45±0.11	4.24±0.09
Hi-Patch	4.16±0.08	3.31±0.06	6.30±0.06	4.12±0.05
UniTIMS-Small	4.52±0.08	3.7±0.08	6.41±0.07	4.38±0.05
UniTIMS-Base	**3.93±0.04**	**2.41±0.03**	**6.06±0.04**	**4.04±0.03**
UniTIMS-Large	4.11±0.03	2.72±0.05	6.21±0.06	4.03±0.05

D.2 Varying Context and Horizon of IMTS Forecasting

On the Human Activity dataset (Table 10), we evaluate model performance under varying observation-to-forecast ratios, comparing general Transformers against specialized architectures. Across both the balanced (2000ms → 2000ms) and the challenging short-history (1000ms → 3000ms) configurations, the results quantify the superior adaptability of our proposed variants to shifting temporal windows.

Similarly, on the PhysioNet dataset (Table 11), the Base model achieves state-of-the-art results, significantly outperforming both general Transformers and specialized baselines like Hi-Patch. It demonstrates remarkable robustness when transitioning to the long-prediction (12h → 36h) scenario, effectively capturing irregularly sampled physiological features. Collectively, these findings validate the model's strong generalization capabilities and its efficacy in handling data-constrained environments across diverse domains.

D.3 Benchmark Results on fev-bench

We evaluate our model on the open leaderboard fev-bench, which consists of 100 forecasting tasks and offers the most comprehensive coverage of diverse real-world scenarios, including tasks with covariates. None of these datasets or tasks were seen during the training phase. Figure 2 shows the results on fev-bench with respect to the Agg. Relative WQL and Agg. Relative MASE metrics, which evaluate the point forecast accuracy (the precision of the model's mean or median (point) predictions under a scale-invariant lens) and the generalization over baselines (whether the foundation model has actually learned complex temporal patterns, like complex seasonality and trends, that give it a true edge over trivial, statistical baselines). From the results, we observe that all pre-trained models under zero-shot inference show significant superiority in scalability compared with task-specific models under supervised training and statistical models. Meanwhile, TabPFN-TS stood out because its native table-form feature extraction design can easily align with the covariance-rich characteristics of fev-bench. Despite being designed for irregular time series, our model still demonstrates robust and trustworthy performance on this benchmark.

D.4 Inference Time

To further evaluate the practical deployment efficiency of UniTIMS, we compare the inference time of representative TSFMs under the same hardware configuration. Figure 5 reports the wall-clock inference time across different models. UniTIMS achieves competitive inference speed relative to other foundation models of comparable scale, demonstrating that our time-aware hybrid attention design does not introduce prohibitive computational overhead despite modeling irregular temporal structures explicitly.

D.5 Training Dynamics

The λ value in Time Attention Bias reflects the extent to which temporal similarity modifies the model's attention. A higher λ value indicates a greater degree of adjustment to attention scores by temporal similarity. Figure 6 illustrates the variation in λ with increasing training steps during the pre-training process of the UniTIMS-Small model, with sampling conducted every 3,000 steps. It can be observed that as the model's training steps increase, the degree of correction applied by temporal similarity to the model grows, gradually trending towards a stable state in the later stages of training. The curve demonstrates that an optimal weight can be found as a bound of using temporal information in the universal representation of IMTS when the model converges.

We further investigate how the model's effectiveness evolves when the number of parameters scales up. Figure 7 presents the training curves for UniTIMSof different sizes. Compared to UniTIMS(Small), the large variant converges to a 24.4% lower training loss, indicating that increased model capacity leads to more effective optimization and stronger representational capability in generative forecasting. We also note that on PhysioNet, UniTIMS-Base marginally outperforms UniTIMS-Large in MSE/MAE (Table 1), which appears to invert the scaling trend observed in pretraining loss. We attribute this to the small evaluation size and high label sparsity of PhysioNet: under the same downstream evaluation setting, the larger variant exhibits higher generalization variance on small held-out IMTS targets, while still attaining the lowest pretraining loss. Across the larger Human Activity and USHCN benchmarks, UniTIMS-Large consistently delivers the best results, confirming that the scaling benefit holds in the regime where evaluation noise is comparatively small. We regard mitigating this small-dataset variance (e.g., via downstream-aware early stopping or temperature-controlled decoding) as a practical limitation worth further study.

E Forecast Visualizations

Figure 8 and Figure 9 present qualitative zero-shot forecasting results of UniTIMS-Large on IMTS and regular MTS datasets, respectively. In Figure 8, four representative cases from the Human Activity dataset exhibit diverse temporal behaviors, including stable variations, gradual trends, and periodic fluctuations. Under irregular sampling, the predicted curves generally preserve the overall evolution of the ground truth and capture the major temporal patterns. Although abrupt outliers or sharp local changes are not fully recovered, the model maintains stable predictions without obvious

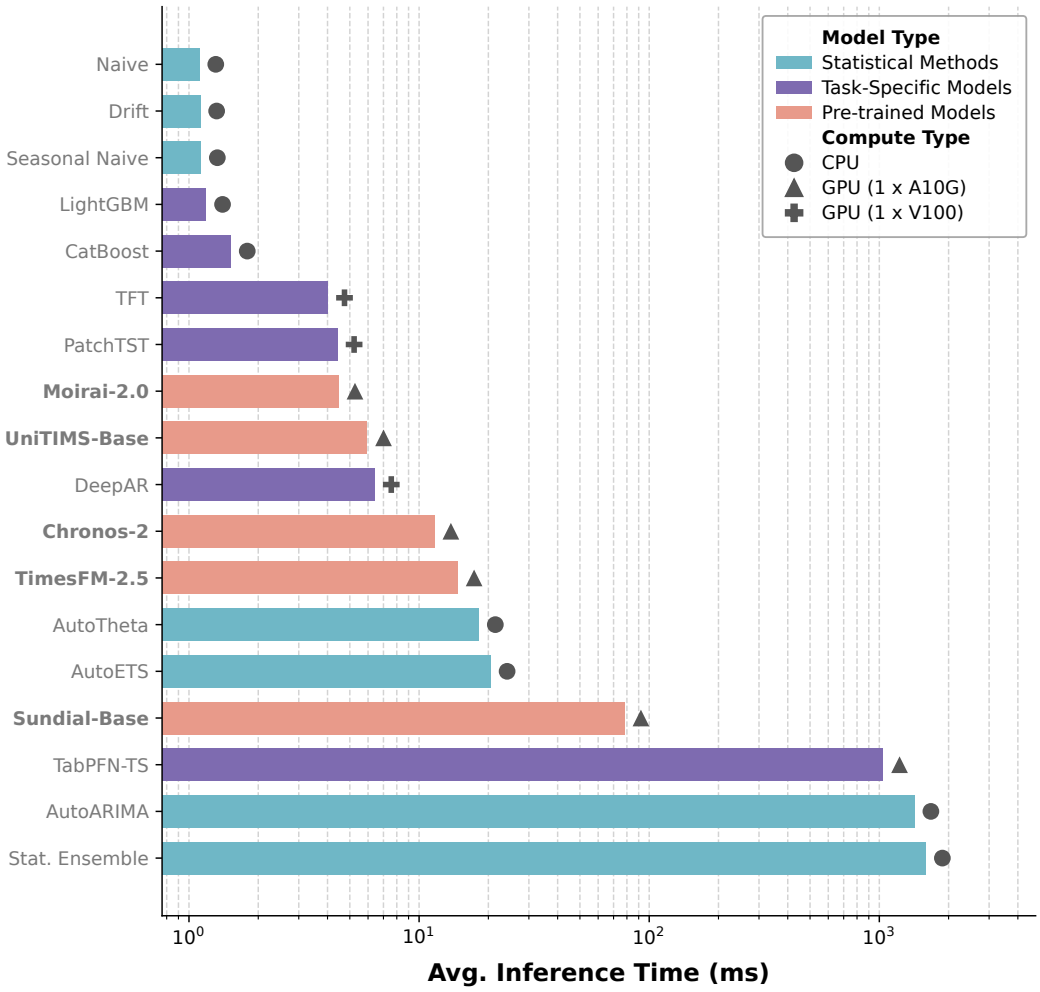

Figure 5: Inference time comparison across representative TSFMs. All models are evaluated under the same hardware setup.

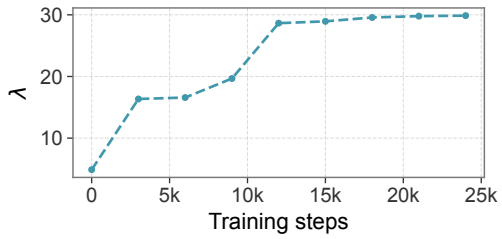

Figure 6: Evolution of λ during training.

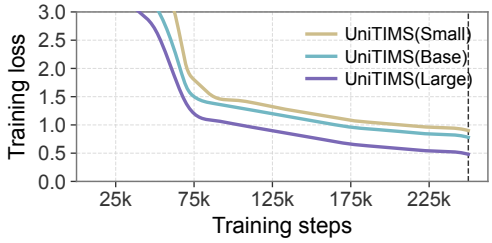

Figure 7: Training loss across model sizes.

trend deviation. Figure 9 further evaluates the transferability of UniTIMS-Large on regular MTS datasets, including ETTh1 and ETTm1. The predictions are smoothly connected with the historical context and remain well aligned with the future ground truth, covering both slowly varying trends and relatively high-frequency local fluctuations. Overall, these visualizations provide intuitive evidence that UniTIMS-Large achieves robust zero-shot forecasting across heterogeneous time-series scenarios, including irregular intervals, asynchronous observations, varying frequencies, trends, and periodic patterns.

E.1 IRREGULAR MULTIVARIATE TIME SERIES (IMTS)

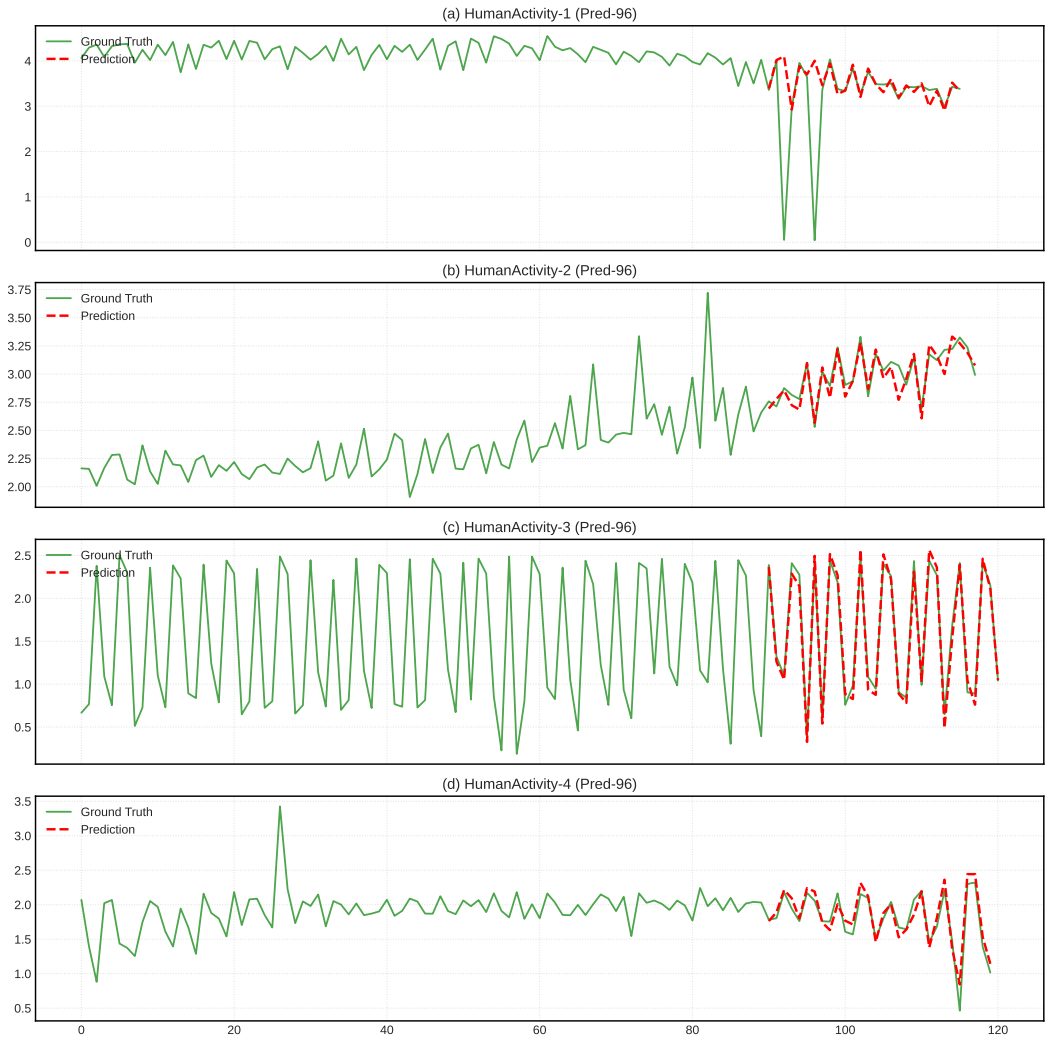

Figure 8: Visualizations of zero-shot forecasts from UniTIMS-Large on the start segments of the Human Activity dataset.

E.2 MULTIVARIATE TIME SERIES (MTS)

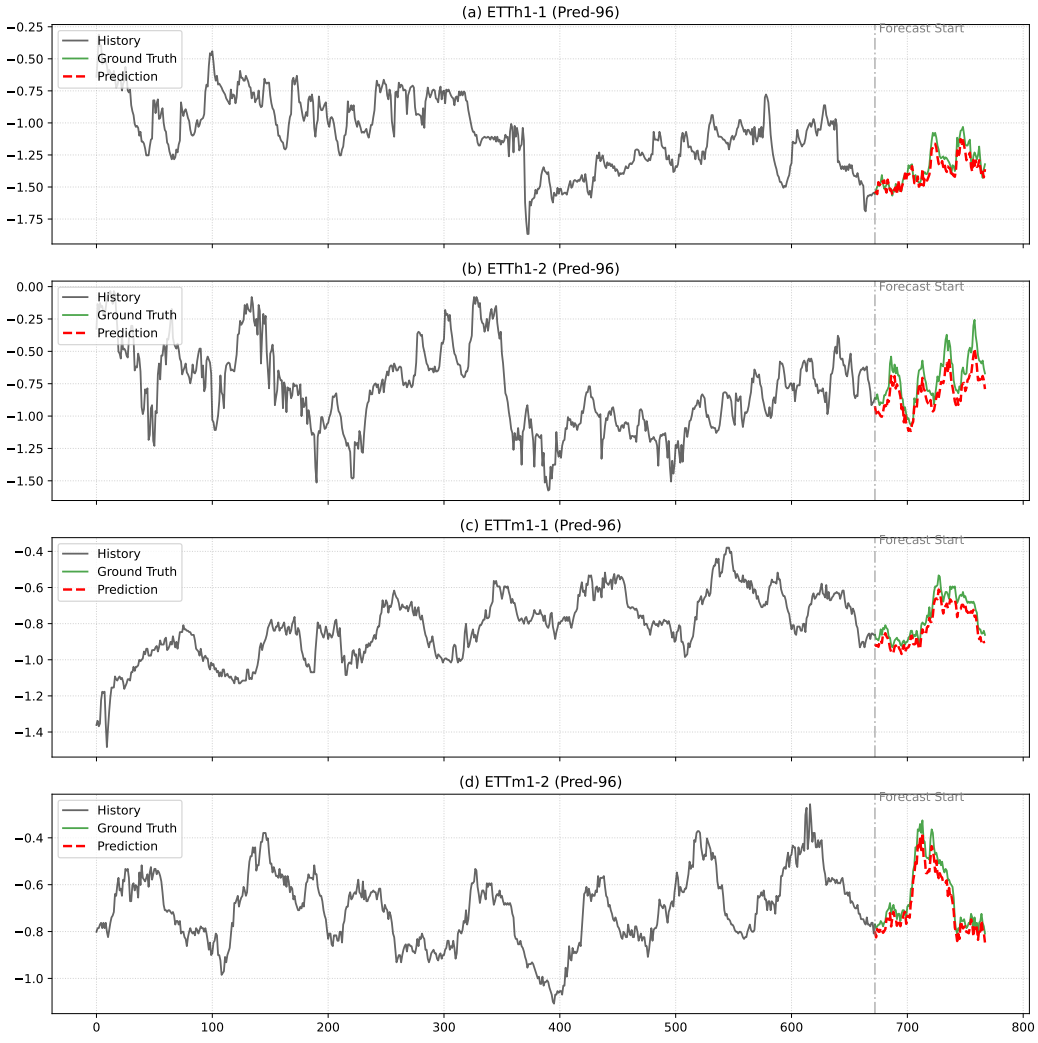

Figure 9: Visualizations of zero-shot forecasts from UniTIMS-Large on regular MTS datasets (ETTh1 and ETTm1).