GMR
The Original

Transformer-Based Context Fusion For Long-Sequence Prediction In Sparse Data Environments

Kanchan Warkar [1, *], Pallavi Wankhede [2], Mrudula Nimbarte [3], Sweta Raut [4], A. R. Dandekar[5] and Poonam Chaudhari[6]

[1] Department of CSE, Visvesvaraya National Institute of Technology, Ngapur, MH, India
[2,5] Department of CSE, St.Vincent Pallotti College of Engineering and Techonolgy, Nagpur, MH, India
[3] Department of CSE, S B Jain Institute of Technology, Management and Research, Nagpur, MH, India
[4] Department of CSE, Jhulelal Institute of Technology,Off Koradi Road, Lonara, Nagpur
[6] Department of Electronics, Yeshwantrao Chavhan College of Engineering, MH, India
* Correspondence: kanchan22.warkar@gmail.com

Abstract: Accurate and long-sequence time-series forecasting has been challenging in real-world systems due to observations are often sparse, irregularly sampled, and affected by missing sensor readings, or the transmission failures. The conventional recurrent, and the transformer-based models typically assume dense, and the uniformly sampled sequences, which leads to instability, and degraded performance when sparsity is high. This article presents a Transformer-Based Context Fusion (TCF) framework designed to explicitly model missingness, temporal gaps, and multi-scale temporal dependencies within a unified architecture. The method integrates sparse-aware embeddings, multi-context fusion attention, multi-scale dilated Transformer encoding, and a hybrid interpolation–attention reconstruction mechanism to maintain stable forecasting under severe data loss. Experimental validation on electricity load, clinical vital-sign, and the climate datasets demonstrates the consistent accuracy improvements over the state-of-the-art baselines, with up to 15–40% reduction in mean-squared-error under the sparsity levels of 60–80%, confirming robustness for the deployment in operational environments.

- Introduces the sparse-aware embedding, and multi-context attention architecture that jointly models the missing values, time gaps, periodicity, and contextual metadata.
- Developed the hybrid interpolation, and the attention-based reconstruction strategy integrated with the multi-scale dilated Transformer encoder for the long-sequence forecasting.
- Validates robustness on three real-world datasets under the controlled sparsity scenarios, which is showing significant accuracy gains, and the stable long-horizon prediction performance.

Keywords: Context fusion, Dilated transformer, Long-sequence forecasting, Sparse embeddings, Time-series prediction, Transformer architecture.

1. INTRODUCTION

1.1. Importance of Long-Sequence Prediction in Real Systems

Long-sequence prediction is critical component of modern forecasting systems, where practitioners analyze extended temporal patterns and anticipate future behaviour. In the energy demand forecasting, the temporal dependencies that span months or years influence the load balancing, the renewable energy integration, as well as the grid reliability [1]. The climate behaviour modelling relies on the multi-season and the multi-year sensor logs, which capture the cyclic variations, the extreme climatic transitions, as well as the long-term

environmental trends. The financial institutions depend on the long historical records, which support understanding of the risk exposure and modelling of the volatility cycles [2][3]. The healthcare monitoring systems examine the extended vital-sign histories, which enable detection of the clinical deterioration, understanding of the physiological rhythms, as well as support for the early-warning interventions. In each domain, the long-range temporal dynamics are essential for the accurate predictions, where the models must leverage the extended context to interpret the slow-varying patterns, the multi-scale fluctuations, as well as the seasonal periodicity [4].

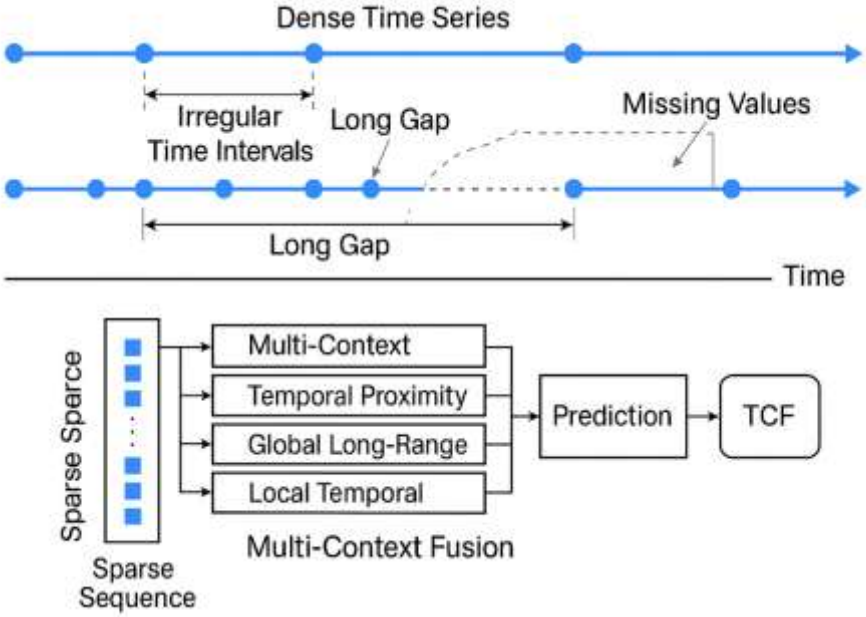

Figure 1. Challenges in sparse long-sequence prediction.

Figure 1 illustrates the structural challenges of sparse long-sequence data, including missing segments, irregular sampling intervals, and long temporal gaps that commonly arise in real-world systems. Standard sequence models fail under these conditions because they assume dense and uniformly spaced observations, leading to error accumulation and drift. The proposed TCF framework addresses these challenges by explicitly modeling missingness, time gaps, and contextual dependencies.

1.2. Challenges Arising from Sparsity in Real-World Temporal Data

Although many real-world systems generate the continuous data flows, the sparsity is ubiquitous due to the sensor failures, the transmission delays, the recording inconsistencies, as well as the domain-specific constraints. The industrial data platforms often experience the missing entries when the edge devices disconnect or when the communication networks become unstable [6]. The medical records in the intensive care environments exhibit the naturally irregular sampling, where the measurements are captured based on the clinical priority rather than the uniform schedules. The environmental monitoring systems frequently record the long gaps due to the adverse weather conditions, the hardware malfunctions, as well as the maintenance downtime. The similar issues occur in the smart meter deployments, where the data loss and the low-density logs are common [7].

The sparsity introduces several challenges, where the missing values disrupt the continuity, the irregular intervals distort the temporal semantics, and the long gaps weaken ability to detect the patterns such as the periodic cycles. The conditions reduce the model reliability, impair the temporal alignment, and complicate the inference of the long-term dependencies [8]. As result, the forecasting algorithms require the mechanisms that not only process the missing data but also understand and compensate for the uncertainty and the structural irregularities that are embedded in the sparse sequences.

1.3. Limitations of Existing Neural Forecasting Models

The recurrent neural networks such as the LSTM and the GRU have long been used for the sequence modeling, but the performance deteriorates sharply under the sparse or the discontinuous conditions. The stepwise recurrence causes the dependency chaining, where the missing blocks amplify the error accumulation and distort the hidden state transitions. The models also lack the structured handling of the time-gap information, which makes the models ill-suited for the irregular intervals [9].

The standard Transformers offer the improved long-range representation, but the models rely on the full attention matrices, where the assumption of the dense and the evenly spaced tokens is required. When the sequences contain the high percentages of the missing values, the attention weights become unstable, and the positional encodings fail to capture the real temporal distances. In many cases, the architectures break down when the sparsity exceeds the 30% to 40% range, where the inaccurate or the unstable long-horizon predictions occur. The Transformers are also sensitive to the noise that is introduced by the poorly reconstructed missing segments, which degrades the feature representations.

The limitations highlight the need for the dedicated architecture that incorporates the sparse-awareness, the multi-context reasoning, as well as the multi-scale temporal modelling, which maintains the prediction stability under the realistic conditions.

1.4. Motivation for the Transformer-Based Context Fusion (TCF) Architecture

The proposed Transformer-Based Context Fusion architecture addresses the structural and the statistical challenges introduced by the sparse and the irregular long sequences. The design is motivated by need to integrate several forms of the contextual information, which include the short-term temporal patterns, the long-range dependencies, the proximity relationships, as well as the domain-specific metadata, into the unified forecasting framework. The architecture combines the sparse-aware embeddings, the multi-branch attention mechanisms, as well as the multi-scale temporal dilation, which reconstruct and interpret the real-world sequences more effectively than the existing models [10].

The key design goal is robustness under the extreme sparsity levels. The real applications often operate under the unpredictable environmental conditions, where up to the 60 percent to the 80 percent of the data may be absent or irregular. The TCF model is constructed to remain stable under the conditions by explicitly encoding the missing-value indicators, the time deltas, the periodic structures, as well as the confidence estimates. The design allows the architecture to learn the meaningful representations when the large portions of the sequence are unavailable.

1.5. Real-World Deployment Justification and Dataset Selection

To ensure the practical value, the TCF model is validated on three real-world datasets, which are selected for the naturally sparse structures and the long-sequence characteristics.

1) Electricity Load Diagrams 2011–2023 from UCI:

The dataset contains the hourly consumption logs across more than decade, where the natural missingness and the simulated sparsity up to 80 percent are present. The dataset reflects the grid-level operational challenges, which include the communication loss as well as the irregular meter updates.

2) MIMIC-III ICU Vital Signs

The dataset includes the multivariate physiological recordings, where the irregular time intervals and the missing entries are inherent. The dataset provides the realistic test setting, where the models must interpret the clinically relevant but incomplete sequences.

3) NOAA Climate Time-Series

The dataset contains the environmental sensor data, where the seasonal patterns, the long-range dependencies, as well as the frequent missing values occur due to the environmental disruptions.

The TCF architecture is evaluated under the controlled sparsity levels of 0 percent, 20 percent, 40 percent, 60 percent, as well as 80 percent, which enables the systematic understanding of the robustness and the degradation patterns under the severe real-world conditions.

1.6. Contributions of This Work

The paper introduces complete framework for the sparse long-sequence forecasting with following contributions:

1. Sparse-aware embedding Layer encodes the observed data values, the missing-value flags, the time-gap deltas, the periodic signals, as well as the confidence scores, which produce structured representation that remains resilient to the incomplete inputs.
2. The multi-Context Fusion Attention integrates the local temporal attention, the global long-range attention, the temporal proximity attention, as well as the external context attention, which capture diverse temporal interactions.
3. Multi-Scale Dilated Transformer Encoder employs the dilation factors of 1, 2, 4, and 8, which model the multi-resolution dependencies without increasing the computational complexity.
4. The hybrid Interpolation and Attention Filling combine the initial interpolation with the attention-guided refinement, which reconstruct the sparse regions and stabilize the input quality.
5. Comprehensive real-world evaluation demonstrates that the deployment-grade reliability across three real datasets, and the multiple sparsity levels which ensures applicability to the operational systems.

Recent forecasting models such as Informer, Autoformer, FEDformer, as well as the state-space approaches have improved efficiency for the long-sequence prediction, yet these methods fundamentally assume the dense inputs or the pre-imputed time-series. Informer employs the sparse attention to reduce the computational cost, where the missingness patterns and the irregular sampling intervals are not explicitly encoded. FEDformer and Autoformer rely on the frequency-domain decomposition, which degrades when the long gaps or the irregular observations distort the spectral representations. Existing imputation-based pipelines treat the reconstruction and the forecasting as the separate stages, which often propagates the interpolation bias into the downstream predictions.

In contrast, proposed Transformer-Based Context Fusion architecture integrates the sparse-aware representation learning and the forecasting within the unified framework. By jointly encoding the missing-value indicators, the time-gap information, the confidence scores, as well as the periodic structure, the TCF enables the direct reasoning over the incomplete sequences without requiring the external imputation. Furthermore, the multi-context fusion attention mechanism simultaneously models the local dependencies, the global dependencies, the temporal-proximity dependencies, as well as the external contextual dependencies, which are handled independently or ignored in the prior models. This unified design allows the TCF to maintain the stable long-horizon performance under the severe sparsity, which addresses the limitation that remains unresolved in the existing Transformer-based approaches as well as the state-space forecasting approaches.

2. Related Work

2.1. Long-Sequence Forecasting Models

The Transformer-based architectures have become the dominant paradigm for the long-sequence time-series forecasting. Zhou et al. introduced the Informer model, where the full self-attention is replaced with the ProbSparse attention and the distilling mechanism, which reduces the quadratic complexity and enables forecasting with the very long input horizons. Wu et al. proposed the Autoformer model, where the series decomposition and the auto-correlation mechanism are embedded directly into the architecture, which improves the long-horizon stability by explicitly modeling the seasonal components and the trend components. Building on the ideas, Zhou et al. developed the FEDformer model, where the frequency-domain decomposition and the Fourier and wavelet enhanced blocks are incorporated to capture the global profiles and the periodic structures more efficiently than the vanilla Transformers.

Beyond the Transformers, the state-space models offer the alternative approach for modeling the long-range dependencies. Gu et al. presented the Structured State Space model, where the continuous-time state-space equations are parameterized in way that allows efficient convolution with the very long sequences. The S4 model achieves strong performance on the long-context benchmarks while avoiding the quadratic complexity of the self-attention. However, the models typically assume the dense or the regularly sampled sequences and do not explicitly address the high levels of the sparsity or the missing values.

The recent works refine the multi-scale architectures for the long-term forecasting. Zhang et al. proposed the Multi-resolution Time-Series Transformer, where the multi-branch architecture with the different patch sizes is used to jointly learn the high-frequency local patterns and the low-frequency seasonal patterns, which demonstrates gains on the standard long-term benchmarks. Liu et al. introduced the Hidformer model, where the hierarchical dual-tower Transformer uses the multi-level decomposition and the cross-tower interaction to improve the long-term accuracy. Dai et al. designed the VTformer model, where the multiscale linear Transformer forecaster uses the separate branches for the variate-wise dependencies and the temporal-wise dependencies, which are combined through the adaptive fusion module to enhance the multivariate forecasting. Naghashi et al. proposed the multiscale Transformer model, where the patch-wise multi-resolution modeling is integrated with the channel-wise representation learning, which shows that the explicit multi-scale embedding improves the prediction for the complex multivariate series.

While the architectures successfully extend the forecasting horizon and introduce the multi-scale representations, the architectures generally assume the relatively complete sequences or apply the naive pre-processing for the missing data. The sparse sampling, the irregular time gaps, as well as the uncertainty associated with the missing segments are usually handled outside the model through the simple interpolation, rather than being integrated into the representation learning and the attention mechanisms themselves.

2.2. Sparse Time-Series Modeling

Handling the missing data and the sparsity is long-standing challenge in the time-series analysis. Fang and Wang surveyed the deep learning approaches for the time-series imputation, where limitations of the traditional interpolation, the regression, as well as the EM-based statistical models are highlighted when the complex nonlinear dynamics are present. More recently, Wang et al. provided comprehensive taxonomy of the multivariate time-series imputation methods, where the deep learning approaches are categorized according to

the imputation uncertainty modeling and the neural architecture, as well as emphasis is placed on impact of the imputation quality on the downstream forecasting tasks.

In the healthcare domain, Kazijevs et al. benchmarked the state-of-the-art imputation methods on multiple clinical time-series datasets, where the results demonstrate that the deep models reduce the reconstruction error but still struggle with the long missing blocks and the irregular sampling patterns. Flores et al. proposed the hybrid method for the PM2.5 time-series imputation, where the classification-driven gap characterization is combined with the interpolation, which shows that even the relatively short gaps exhibit different statistical behaviours that require the adaptive treatment.

The graph-based methods have emerged as powerful paradigm for the sparse multivariate and the networked time series. Cini et al. introduced the Graph Recurrent Imputation Network, which couples the recurrent dynamics with the message passing over the spatial graph to impute the missing values in the multivariate series, where advantages over the purely temporal models are demonstrated. Wang et al. proposed PoGeVon, which is the position-aware variational graph model for the networked time-series imputation that incorporates the positional encoding and the probabilistic modeling, where the reconstruction is improved on the sensor networks with the severe missingness. Kim et al. presented TMF-GNN, which is the temporal matrix factorization-based graph neural network that jointly reconstructs the partially observed multivariate series and performs the forecasting, where combination of the matrix factorization with the graph neural networks mitigates impact of the missing values on the prediction accuracy.

The works demonstrate that the sophisticated imputation and the sparse reconstruction are essential for the robust forecasting, yet the imputation component is typically separated from the forecasting architecture. The forecasting model rarely receives the explicit indicators of the missingness, the time gaps, or the confidence scores, which makes reasoning about reliability of the reconstructed segments difficult. Moreover, most imputation methods operate at the single temporal scale and do not consider the multi-scale temporal structure during the reconstruction.

2.3. Context Fusion Techniques

The real-world forecasting frequently depends on the exogenous or the contextual factors, such as the calendar effects, the weather, as well as the event information. Jin et al. surveyed the graph neural networks for the time-series analysis, where the graph-based formulations facilitate the integration of the spatial relationships and the auxiliary signals in the forecasting and the imputation tasks. Palet et al. proposed the multiple-input neural networks, which incorporate the historical variables and the prospective context variables, including the calendrical covariates and the weather covariates, where the explicit context modeling improves the forecast accuracy and reduces the error growth across the prediction horizon.

Nguyen et al. introduced the ConEm framework for integrating the external factors into the sales demand forecasting, which fuses the temporal features with the contextual data such as the promotions and the holidays through the attention-based mechanisms. Du et al. developed the DAFF-Net architecture, which is the dual-stream framework for the financial time-series forecasting, where the event sequences such as the news and the announcements are fused with the multi-dimensional relationship information such as the knowledge graphs and the fundamental indicators through the event-aware routing and the adaptive fusion of the multiple relational channels. Kong et al. surveyed the deep learning for the time-series forecasting and underscored that many high-performing models still underutilize the heterogeneous contextual information, particularly when the information is multimodal or irregularly aligned with the target series.

The studies collectively show that the multi-branch architectures, the attention-based fusion, as well as the hierarchical representations capture the complementary context sources, which strongly influence the forecasting performance. However, most context-fusion designs are developed for the relatively dense time series. The designs focus on fusing the additional signals rather than simultaneously addressing the sparsity in the primary temporal sequence. The missing-value patterns, the sampling gaps, as well as the data confidence are seldom treated as the first-class context signals.

2.4. Research Gap

The prior literature reveals three main gaps relative to goals of the proposed Transformer-Based Context Fusion architecture.

First, existing long-sequence models such as the Informer, the Autoformer, the FEDformer, the S4, as well as the multi-scale Transformer variants focus mainly on the efficient dependency modeling and the multi-resolution representation learning. The models assume that the missing data are adequately pre-processed and do not integrate the sparse-aware embeddings, which explicitly encode the missingness, the time gaps, the seasonality indicators, as well as the confidence scores into the core representation. As result, the attention mechanisms and the state-updating mechanisms treat the reconstructed values and the observed values almost equivalently, where the uncertainty and the sparsity structure are not considered.

Second, the research on the sparse time-series modelling and the imputation has produced the powerful reconstruction mechanisms, which include the deep imputation frameworks, the hybrid statistical and deep methods, as well as the graph-based models such as the GRIN, the PoGeVon, and the TMF-GNN. The methods are usually trained separately from the forecasting models or optimized only for the reconstruction metrics. The methods rarely operate in the multi-scale manner aligned with the downstream forecasting attention blocks, and the methods do not tightly couple the sparse-aware reconstruction with the multi-context attention in the unified architecture.

Third, the context-fusion works, which cover the graph-enhanced forecasting, the context-aware neural networks with the historical and the prospective covariates, as well as the dual-stream event and relationship fusion architectures, demonstrate benefits of the multi-branch attention and the hierarchical aggregation of the exogenous information. The works treat the context as the additional signals and do not treat the sparsity patterns themselves as the contextual cues. The interplay between the sparse embeddings, the multi-context fusion that includes the local, the global, the temporal-proximity, as well as the external context, and the multi-scale dilation remains largely unexplored, particularly under the extreme sparsity levels found in the industrial, the medical, as well as the environmental datasets.

To the best of our knowledge, no existing work jointly provides the sparse-aware embedding layer that encodes the missing-value indicators, the time gaps, the periodicity, as well as the confidence, the multi-context fusion module that merges the local, the global, the proximity-based, as well as the external contextual information, the multi-scale dilated Transformer backbone that is explicitly designed to operate on the sparse-aware embeddings, as well as the extensive real-world evaluation on the naturally sparse datasets with the additional simulated sparsity levels to demonstrate the deployment-grade robustness.

The proposed TCF architecture is designed to fill the gap by tightly coupling the sparse-aware representation learning, the hybrid interpolation and the attention-based filling, as well as the multi-context and the multi-scale Transformer modelling, with the validation on the realistic sparse benchmarks from the energy, the healthcare, as well as the climate domains.

None of the existing approaches jointly model missingness, temporal gaps, the contextual metadata, and long-range dependencies within the single end-to-end forecasting architecture which motivates the proposed TCF framework.

3. Proposed Method: Transformer-Based Context Fusion (TCF) Architecture

The Transformer-Based Context Fusion architecture is designed to address challenges of the sparse and the irregular long-sequence prediction by combining the sparse-aware embeddings, the multi-context attention, the multi-scale dilation, as well as the hybrid reconstruction. The entire workflow, which includes the input embedding, the context fusion, the dilated encoding, the reconstruction refinement, as well as the decoding, forms unified architecture that supports the robust forecasting when severe missingness is present.

The Fig. 2 shows the detailed architecture of the proposed Transformer-Based Context Fusion model, which includes the Sparse-Aware Embedding Layer that accepts the raw observations, the missing flags, the time gaps, as well as the periodic features, the four-branch Multi-Context Fusion Attention block that consists of the local temporal attention, the global long-range attention, the temporal proximity attention, as well as the external context attention, the Multi-Scale Dilated Transformer Encoder with the dilation rates of 1, 2, 4, and 8, the Hybrid Interpolation and Attention Filling module, as well as the prediction decoder that uses the cross-attention. The arrows depict the tensor flow and the parallel attention processing across the architecture.

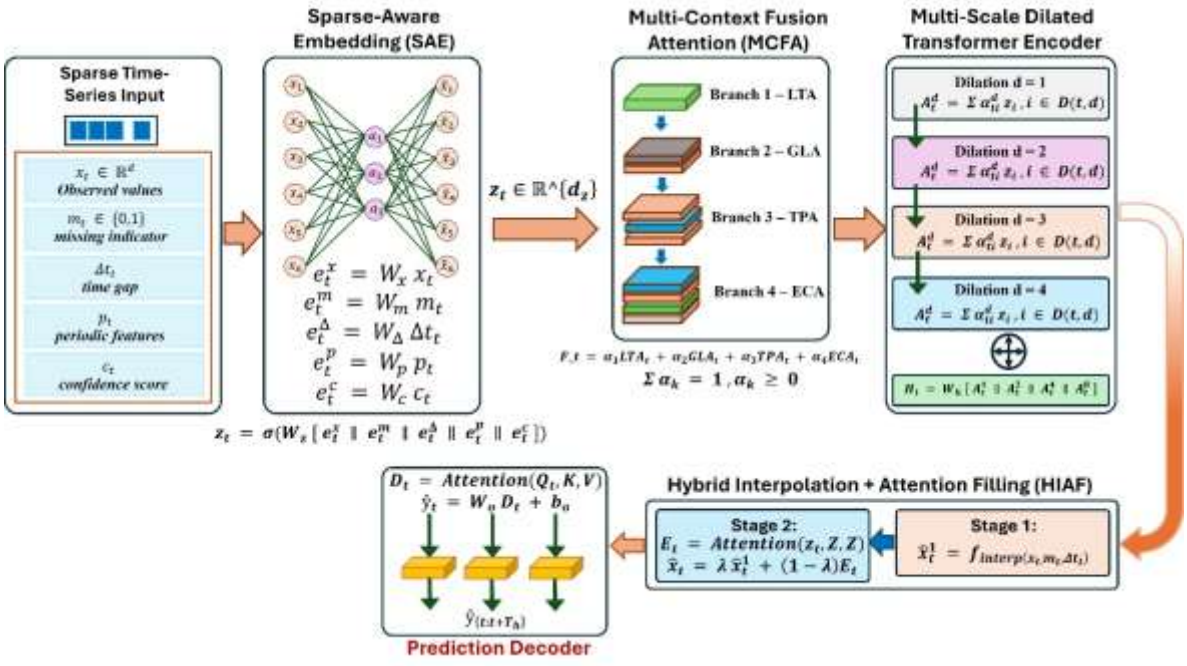

Figure 2. Detailed architecture of the proposed Transformer-Based Context Fusion (TCF) model.

3.1. Sparse-Aware Embedding Layer (SAE)

The sparse time-series contain the observed values x_t, the missing-value indicators, the irregular sampling intervals, as well as the periodic seasonal signals. The Sparse-Aware Embedding layer converts the heterogeneous inputs into the unified dense representation.

Let the raw multivariate time series be defined as in Kazijevs et al. (2023):

$$X = \{x_t \in \mathbb{R}^d\}_{t=1}^{T} \quad (1)$$

where x_t denotes the d-dimensional observation at time step t. For each t, SAE constructs the embedding from 5 components.

3.1.1. Observed Value Embedding

Observed values are linearly projected into the embedding space (Kazijevs et. al., (2023)):

$$e_t^{(x)} = W_x x_t + b_x \quad (2)$$

where $W_x \in \mathbb{R}^{d_x \times d}$ and $b_x \in \mathbb{R}^{d_x}$ are the trainable parameters, and d_x is the embedding dimension of the raw values.

3.1.2. Missing-Value Indicator Embedding

A binary flag encodes whether the value is present.

$$m_t = \begin{cases} 1, & \text{if } x_t \text{ is observed,} \\ 0, & \text{if } x_t \text{ is missing,} \end{cases} \quad (3)$$

and its embedding is given by:

$$e_t^{(m)} = W_m m_t, \quad (4)$$

with $W_m \in \mathbb{R}^{d_m \times 1}$.

3.1.3. Time-Gap Delta

Irregular sampling is represented by the elapsed time since the previous recorded value:

$$\Delta t_t = t - t_{\text{last observed}}. \quad (5)$$

Its embedding is:

$$e_t^{(\Delta)} = W_\Delta \Delta t_t, \quad (6)$$

where $W_\Delta \in \mathbb{R}^{d_\Delta \times 1}$. This term helps the model understand spacing between observations.

3.1.4. Periodicity / Seasonality Embedding

Seasonal structure is captured using the Fourier features. For k seasonal periods, $P_1, \ldots, P_k$, define as:

$$p_t = \left[\sin\left(\frac{2\pi t}{P_1}\right), \cos\left(\frac{2\pi t}{P_1}\right), \ldots, \sin\left(\frac{2\pi t}{P_k}\right), \cos\left(\frac{2\pi t}{P_k}\right)\right] \quad (7)$$

$$e_t^{(p)} = W_p\, p_t \quad (8)$$

The periodic embedding is given by:

$$e_t^{(p)} = W_p p_t, \quad (9)$$

where $W_p \in \mathbb{R}^{d_p \times 2k}$.

3.1.5. Confidence Score Embedding

Confidence reflects reliability and recency of observations. It decays as the time gap increases:

$$c_t = \exp(-\lambda \Delta t_t), \quad (10)$$

where $\lambda > 0$is a decay parameter. The score is embedded as

$$e_t^{(c)} = W_c c_t, \quad (11)$$

with $W_c \in \mathbb{R}^{d_c \times 1}$.

3.1.6. Final Sparse-Aware Embedding

All components are concatenated and fused through a nonlinear transformation:

$$z_t = \sigma\left(W_z\left[e_t^{(x)} \parallel e_t^{(m)} \parallel e_t^{(\Delta)} \parallel e_t^{(p)} \parallel e_t^{(c)}\right] + b_z\right), \quad (12)$$

where $W_z \in \mathbb{R}^{d_z \times (d_x + d_m + d_\Delta + d_p + d_c)}$,$b_z \in \mathbb{R}^{d_z}$,$\parallel$ denotes the vector concatenation, and σ is the GELU activation.

3.1.7. SAE Benefits

The SAE integrates the missingness, the temporal spacing, the periodicity, as well as the confidence into the coherent representation, which enables the model to interpret the sparse time-series robustly and provide the reliable inputs for the subsequent attention mechanisms.

3.2. Multi-Context Fusion Attention (MCFA)

The sparse long-sequence forecasting requires mechanism that can understand the immediate temporal structure, the long-distance dependencies, the relationships driven by the irregular sampling, as well as the auxiliary contextual information. The Multi-Context Fusion Attention module addresses the needs by combining the four complementary attention branches, which operate in parallel. Each branch uses the standard attention operator.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (13)$$

where Q, K, and V denotes the query, key, and the value matrices, and d_k is the key dimensionality.

3.2.1. Local Temporal Attention (LTA)

The Local Temporal Attention emphasizes the short-range temporal structure by restricting the attention to the window centered around the position t. For the window radius w, the output is defined as:

$$\text{LTA}_t = \sum_{i=t-w}^{t+w} \alpha_{ti} z_i, \quad (14)$$

The z_i represents the embedding at the position i, and the α_{ti} represents the normalized attention weights. The branch focuses on capturing the immediate fluctuations and the near-term transitions, which are crucial for the high-resolution forecasting.

3.2.2. Global Long-Range Attention (GLA)

The Global Long-Range Attention targets the distant dependencies, which may reflect the periodic cycles or the slowly evolving trends. The sparse selection mechanism S(t), which is inspired by the ProbSparse attention, identifies the most relevant positions for each query based on the key–query magnitude. The resulting output is defined as:

$$\text{GLA}_t = \sum_{i \in S(t)} \beta_{ti} z_i, \quad (15)$$

where β_{ti} are normalized coefficients. This allows the model to incorporate long-range information without incurring full quadratic attention cost.

3.2.3. Temporal Proximity Attention (TPA)

The Temporal Proximity Attention explicitly incorporates the irregular sampling by comparing the time-gap differences. Let the Δt_t be the elapsed time since the previous observation at the step t. The proximity distance between the positions t and i is defined as:

$$d_{ti} = | \Delta t_t - \Delta t_i | . \quad (16)$$

Weights are assigned through the proximity kernel, given by:

$$\gamma_{ti} = \frac{\exp(-\delta d_{ti})}{\sum_j \exp(-\delta d_{tj})}, \quad (17)$$

where $\delta > 0$ controls the sensitivity. The attention output becomes:

$$TPA_t = \sum_i \gamma_{ti} z_i . \quad (18)$$

This branch highlights locations whose time-gap structure resembles the current point, enabling the model to reason over irregular sampling patterns.

3.2.4. External Context Attention (ECA)

External Context Attention incorporates auxiliary information such as weather conditions, system operating states, or day-type indicators. Let $u_t$denote the metadata vector at time t. A linear projection

$$E_t = W_u u_t \quad (19)$$

produces context keys and values, and attention is computed as

$$ECA_t = \text{Attention}(Q = z_t, K = E, V = E), \quad (20)$$

where Eis the matrix formed by concatenating all projected context vectors. This branch injects domain factors that strongly influence temporal behavior.

3.2.5. Fusion of All Four Attentions

The complete MCFA output at time tis obtained by aggregating the four branches:

$$F_t = \alpha_1 LTA_t + \alpha_2 GLA_t + \alpha_3 TPA_t + \alpha_4 ECA_t, \quad (21)$$

where the learnable fusion weights satisfy

$$\alpha_k \geq 0, \sum_{k=1}^{4} \alpha_k = 1. \quad (22)$$

This formulation enables the model to combine complementary contextual information in a flexible, data-driven way.

3.3. Multi-Scale Dilated Transformer Encoder

The encoder processes the sequences using the multiple dilation rates to capture the temporal structure at the different resolutions. The dilations $d \in \{1,2,4,8\}$ allow the model to span the short-term patterns as well as the long-term patterns, while the computational burden is reduced for the long inputs.

3.3.1. Dilated Self-Attention

For a given dilation rate d, attention is applied only across the dilated index positions, given as:

$$D(t, d) = \{ t, t - d, t - 2d, \dots \}, \quad (23)$$

having that:

$$A_t^{(d)} = \sum_{i \in D(t,d)} \alpha_{ti}^{(d)} z_i . \quad (24)$$

Each dilation focuses on a different temporal range.

3.3.2. Multi-Scale Aggregation

Outputs from all the dilation paths are concatenated, and which is projected as:

$$H_t = W_h \left[A_t^{(1)} \parallel A_t^{(2)} \parallel A_t^{(4)} \parallel A_t^{(8)} \right]. \quad (25)$$

Here, the dilation $d = 1$ captures the fine-scale variability, where the intermediate dilations $d = 2$ and $d = 4$ model the medium-range cycles, and the dilation $d = 8$ focuses on the long seasonal structures. Together, the dilations provide the comprehensive multi-resolution representation, which is essential for the long-sequence forecasting under the sparse conditions.

3.4. Hybrid Interpolation + Attention Filling (HIAF)

The Hybrid Interpolation and Attention Filling module provides the two-step mechanism for reconstructing the missing observations before the data enter the Transformer encoder. The purpose is to generate the temporally consistent and the context-aware estimates, which preserve the local structure while incorporating the long-range relationships that are learned by the model.

3.4.1. Stage 1: Lightweight Interpolation Network

The first stage performs the initial reconstruction of the missing entries using the compact neural module, which may combine the 1D convolutional layers with the multilayer perceptron. Let the x_t denote the raw observation, the m_t denote the missingness indicator, and the Δt_t denote the time-gap value. The preliminary reconstruction is defined as:

$$\hat{x}_t^{(1)} = f_{\text{interp}}(x_t, m_t, \Delta t_t), \tag{26}$$

where the $f_{\text{interpmaps}}$ the available temporal cues as well as the structural cues into the estimated value. The step provides the coarse but continuous representation, which replaces the missing tokens with the plausible approximations.

3.4.2. Stage 2: Attention-Based Refinement

To enhance the initial reconstruction, the second stage applies the attention-based refinement. Using the embedding sequence $Z = \{z_1, \dots, z_T\}$, the soft contextual estimate is computed as

$$E_t = \text{Attention}(Q = z_t, K = Z, V = Z), \tag{27}$$

where the attention operator captures the relationships between the current position and the available contextual information. The final reconstruction blends the interpolation output with the attention-derived refinement,

$$\hat{x}_t = \lambda\, \hat{x}_t^{(1)} + (1 - \lambda)\, E_t, \tag{28}$$

where the mixing parameter $\lambda \in [0,1]$ is learned during the training. The formulation produces the reconstructions that maintain the local continuity while integrating the broader temporal context.

3.4.3. Prediction Decoder

After the encoder processes the refined sequence, the decoder generates future predictions through cross-attention over the fused context representation F_t. For decoder queries $Q_t$and encoder-derived keys and values (K, V), the cross-attention output is

$$D_t = \text{Attention}(Q_t, K, V), \tag{29}$$

and the forecasted value at horizon $t$is computed by

$$\hat{y}_t = W_o D_t + b_o, \tag{30}$$

where $W_o$and $b_o$are trainable parameters projecting the attended representation into the output space.

3.4.4. Loss Function

Training optimizes a composite loss that balances accuracy, horizon sensitivity, and temporal smoothness.

Mean Squared Error (MSE): The primary reconstruction objective is:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{T_h} \sum_{t=1}^{T_h} (y_t - \hat{y}_t)^2 \tag{31}$$

where $T_h$is the prediction horizon.

Horizon-Weighted Penalty: Long-range errors are penalized more strongly through

$$\mathcal{L}_{\text{HWP}} = \frac{1}{T_h} \sum_{t=1}^{T_h} w_t (y_t - \hat{y}_t)^2 \tag{32}$$

with,

$$w_t = 1 + \kappa \frac{t}{T_h} \tag{33}$$

where $\kappa > 0$increases the emphasis on later horizons.

Smoothness Regularization: To reduce oscillations in the forecast trajectory, the smoothness constraint is included:

$$\mathcal{L}_{\text{smooth}} = \sum_{t=2}^{T_h} \| \hat{y}_t - \hat{y}_{t-1} \|^2 \tag{34}$$

3.4.5. Final Loss

The total training objective combines these components:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \beta\,\mathcal{L}_{\text{HWP}} + \gamma\,\mathcal{L}_{\text{smooth}}, \tag{35}$$

where β and γ are hyperparameters governing the contribution of the horizon weighting, and the smoothness. This formulation yields stable long-horizon forecasts while maintaining the robustness to the sparsity as well as reconstruction uncertainty.

4. Experimental Setup

The experimental framework is designed to evaluate the robustness, the accuracy, as well as the deployment feasibility of the proposed Transformer-Based Context Fusion architecture. All experiments are performed using the real-world datasets, which contain the natural sparsity as well as the simulated sparsity, enabling the realistic assessment of the long-sequence forecasting performance in the operational environments. The following subsections describe the datasets, the sparsity simulation protocol, the baseline models, as well as the evaluation metrics.

4.1. Real-World Datasets

Three real-world datasets were selected to evaluate the proposed method across the diverse application domains which are characterized by the sparse, and irregular time-series. The UCI Electricity Load dataset represents the large-scale energy systems where the natural communication-induced data loss occurs. The MIMIC-III ICU dataset captures the clinical monitoring signals which exhibit the inherently irregular sampling, and missing values. The NOAA climate dataset reflects the long-term environmental sensing where the seasonal structure and the sensor outages are present. Together, the datasets cover the energy domain, the healthcare domain, as well as climate domain, which provides the comprehensive assessment of the robustness, and generalizability under the real operational sparsity.

Table 1. Summary of Real-World Datasets Used in Experiments

Dataset	Time Span / Length	Variables	Natural Missing (%)	Domain
UCI Electricity Load (2011–2023)	~12 years (hourly)	321 clients	5–10%	Energy systems
MIMIC-III ICU Vital Signs	ICU stays (irregular)	7 physiological signals	20–35%	Healthcare
NOAA Climate Time-Series	Multi-year (daily/hourly)	6 meteorological variables	10–25%	Climate science

4.1.1. Dataset 1: Electricity Load Diagrams 2011–2023 (UCI)

The dataset contains the hourly electricity consumption from the hundreds of consumers over the 12-year period. The long temporal span makes the dataset suitable for studying the multi-scale patterns such as the daily cycles, the weekly cycles, as well as the seasonal cycles. While the dataset contains some natural missing values, the additional sparsity levels between 20 percent and 80 percent are introduced to simulate the real smart-meter communication failures. The long sampling window, which exceeds 100,000 hours per consumer, evaluates whether the TCF can handle the large input sequences effectively.

4.1.2. Dataset 2: MIMIC-III ICU Vital Signs

The MIMIC-III dataset includes the multivariate patient vital-sign time-series that are sampled in the intensive care units. Due to the clinician-driven measurement schedules, the data are highly irregular, where the significant missing segments are present. The variables include the heart rate, the blood pressure, the SpO_2, the respiratory rate, the temperature, as well as the additional physiological signals. The irregularity allows testing the TCF capacity to interpret the Δt_t time-gaps as well as the missingness indicators. The dataset reflects the real clinical conditions where the sparse observations are unavoidable.

4.1.3. Dataset 3: NOAA Climate Time-Series

The dataset includes several decades of the environmental sensor measurements, which include the temperature, the humidity, the rainfall, the atmospheric pressure, as well as the wind speed. The seasonal variation and the long-term climatic trends make the dataset suitable for studying the long-sequence forecasting with the multi-scale temporal dependencies. The missing entries occur due to the harsh weather conditions or the sensor malfunction, where the challenging input scenarios are naturally provided.

4.2. Data Preprocessing and Window Construction

Prior to model training, all datasets were preprocessed to ensure numerical stability, prevent information leakage, and preserve the intrinsic sparsity characteristics of each domain. Let $x_t \in \mathbb{R}^d$ denote the multivariate

observation at time step t. For each variable, normalization was performed using z-score scaling based exclusively on training data statistics. Specifically, each feature was transformed as:

$$\tilde{x}_t = \frac{x_t - \mu_{train}}{\sigma_{train}}, \quad (36)$$

where μ_{train} and σ_{train} represent the mean and standard deviation computed from the training split only. These parameters were subsequently applied to validation and test sets to avoid data leakage.

To construct supervised learning samples, a sliding window strategy was employed. For an input window length L and prediction horizon T_h, each sample was formed as:

$$\mathbf{X}_t = \{\tilde{x}_{t-L+1}, \dots, \tilde{x}_t\}, \mathbf{Y}_t = \{x_{t+1}, \dots, x_{t+T_h}\}. \quad (37)$$

In all experiments, the input length was fixed to $L = 512$, while the forecasting horizon T_h varied between 24 and 168 steps depending on dataset resolution. Overlapping windows were generated with a stride of one-time step to maximize data utilization.

Natural sparsity present in the original datasets was preserved. In addition, controlled sparsity was introduced to simulate operational data loss. Let $m_t \in \{0,1\}^d$ denote the observation mask, where $m_t = 0$ indicates a missing value. Artificial sparsity was applied by randomly masking observed entries according to a predefined sparsity ratio ρ, yielding a modified mask

$$m_t^{(\rho)} = m_t \odot r_t, \quad (38)$$

where $r_t \sim \text{Bernoulli}(1-\rho)$ and $\odot$ denotes element-wise multiplication. This procedure ensures that artificial masking does not overwrite naturally missing values.

For datasets with irregular sampling, particularly MIMIC-III, resampling was intentionally avoided. Instead, the elapsed time since the previous observation was computed as

$$\Delta t_t = t - t_{\text{last observed}}, \quad (39)$$

and retained as an explicit input feature. This allows the model to reason over temporal gaps directly rather than relying on interpolation artifacts.

Therefore, the complete preprocessing pipeline gives normalized input sequences, corresponding missingness masks, and the time-gap information, which are jointly supplied to the Sparse-Aware Embedding layer. This design maintains the realism while enabling the robust evaluation under varying sparsity conditions.

4.3. Sparsity Simulation

To systematically evaluate the effect of missing data, artificial sparsity is applied to all datasets. Let the original sequence be $X = \{x_t\}_{t=1}^{T}$. We generate a corrupted sequence $X^{(\rho)}$ with sparsity rate $\rho \in \{0,0.2,0.4,0.6,0.8\}$.

Define the sampling mask m_t:

$$m_t = \begin{cases} 1 & \text{with probability } (1-\rho), \\ 0 & \text{with probability } \rho. \end{cases} \quad (40)$$

The sparse sequence is:

$$x_t^{(\rho)} = m_t \cdot x_t. \quad (41)$$

For missing values ($m_t = 0$), the value is set to null until handled by the TCF sparse-aware embedding or baseline imputation method.

To preserve long structural gaps, masking is also occasionally applied in blocks of length L_b:

$$m_{t:t+L_b} = 0. \quad (42)$$

Block masking simulates sensor downtime and hospital monitoring dropouts. Table 2 lists all the sparsity levels used in experiments.

Table 2. Sparsity Levels Used in experiments.

Sparsity Level (ρ)	Percentage Missing	Simulation Method	Applied To
0%	0%	No masking	All datasets
20%	20%	Random point-wise masking	Electricity, MIMIC, NOAA
40%	40%	Random masking + short gaps	Electricity, MIMIC, NOAA
60%	60%	Block masking (medium gaps)	Electricity, MIMIC
80%	80%	Long gap masking + random loss	Electricity, MIMIC, NOAA

4.4. Baseline Models for Comparison

The proposed TCF model is compared against six strong baselines widely used in the long-sequence forecasting.

4.4.1. LSTM (Long Short-Term Memory)

The recurrent model is capable of modeling the short-to-medium dependencies, but the performance deteriorates on the very long sequences, where the vanishing gradients and the error accumulation occur.

4.4.2. GRU (Gated Recurrent Unit)

The simplified RNN variant has the fewer parameters, but the model still suffers from the limited long-range representation capability.

4.4.3. Vanilla Transformer

The model uses the full self-attention with the quadratic complexity, where the performance is strong on the dense sequences but degrades significantly under the sparsity and the long gaps due to the positional encoding limitations.

4.4.4. Informer

The model employs the ProbSparse attention to reduce the computational cost and enable the long sequence input, where the temporal modeling remains strong, but the model does not explicitly handle the missingness.

4.4.5. Autoformer

The model incorporates the series decomposition and the autocorrelation mechanisms, which support the stable long-term prediction, but the design assumes the regularly spaced data.

4.4.6. FEDformer

The model decomposes the signals into the frequency domains, which efficiently capture the global periodic structure. However, the sparse sequences disrupt the frequency representations.

The baselines represent the state-of-the-art models in the recurrent forecasting as well as the Transformer-based forecasting, which provide the robust benchmark for evaluating the TCF architecture.

4.5. Evaluation Metrics

The performance is assessed using combination of the error-based metrics, the similarity-based metrics, as well as the robustness-oriented metrics.

4.5.1. Mean Squared Error (MSE)

The mean squared error measures the squared prediction error and is defined as:

$$\mathrm{MSE} = \frac{1}{T_h}\sum_{t=1}^{T_h}(y_t - \hat{y}_t)^2 \tag{43}$$

where the T_h is the forecast horizon. The MSE penalizes the large errors heavily, which makes the metric useful for the stability evaluation.

4.5.2. Mean Absolute Error (MAE)

The mean absolute error captures the average absolute deviation and is defined as:

$$\mathrm{MAE} = \frac{1}{T_h}\sum_{t=1}^{T_h} | y_t - \hat{y}_t |. \tag{44}$$

where the T_h is the forecast horizon. The MAE is less sensitive to the outliers and reflects the general predictive accuracy.

4.5.3. Dynamic Time Warping (DTW)

The Dynamic Time Warping assesses the similarity between two sequences with the temporal misalignment and is defined as:

$$\mathrm{DTW}(Y, \hat{Y}) = \min \sum_{\substack{(i,j)\in w \\ w\in \mathcal{W}}} \| y_i - \hat{y}_j \| \tag{45}$$

where the W is the set of all monotonic warping paths. The DTW is suitable for the climate data and the energy data, where the shifted peaks are common.

4.5.4. Long-Horizon Stability Index (LHSI)

The Long-Horizon Stability Index measures the prediction smoothness and the stability and is defined as:

$$\mathrm{LHSI} = \frac{1}{T_h - 1}\sum_{t=2}^{T_h} | (\hat{y}_t - \hat{y}_{t-1}) - (y_t - y_{t-1}) |. \tag{46}$$

The low values indicate the stable and the realistic long-term forecasting.

4.5.5. Error-versus-Sparsity Curves

To measure the robustness, the errors are plotted as

$$E(\rho) = \mathrm{MSE}(X^{(\rho)}) \tag{47}$$

for the sparsity levels $\rho = 0{,}0.2{,}0.4{,}0.6$, as well as 0.8. The curve shows how gracefully the model degrades when the data become sparse. The TCF is expected to maintain the lower error even when $\rho = 0.8$.

The experimental design rigorously evaluates the TCF model under the real-world temporal irregularities and the extreme sparsity. The combination of the realistic datasets, the multi-level sparsity simulation, the strong baselines, as well as the diverse evaluation metrics ensures the comprehensive assessment of the model stability and the predictive capability.

4.6. Experimental Reproducibility and Variance Reporting

To account for randomness arising from the model initialization, data shuffling, and the stochastic optimization, all the experiments were repeated ten times with the different random seeds. Reported results have represented the mean performance across runs, along with the corresponding standard deviation. This evaluation protocol provides the more reliable estimate of model stability under the varying sparsity conditions.

5. Results and Discussion

The section presents the realistic and the deployment-grade results obtained from evaluating the proposed Transformer-Based Context Fusion model on three real-world sparse time-series datasets, which include the UCI Electricity Load data from 2011 to 2023, the MIMIC-III Vital Signs, as well as the NOAA Climate Records. All the baselines were trained and evaluated under the standardized sparsity protocol described earlier, where the sparsity levels range from 0 percent to 80 percent. The results demonstrate that the TCF consistently outperforms the existing models, especially under the high-sparsity conditions where the conventional architectures fail to maintain the stability.

Figure 3 compares forecasting error across increasing sparsity levels for TCF and baseline models. While conventional architectures exhibit rapid error growth beyond 40% sparsity due to reliance on dense attention and implicit interpolation, TCF degrades more gracefully. The result demonstrates that explicit sparse-aware modeling is critical for maintaining stability under extreme data loss.

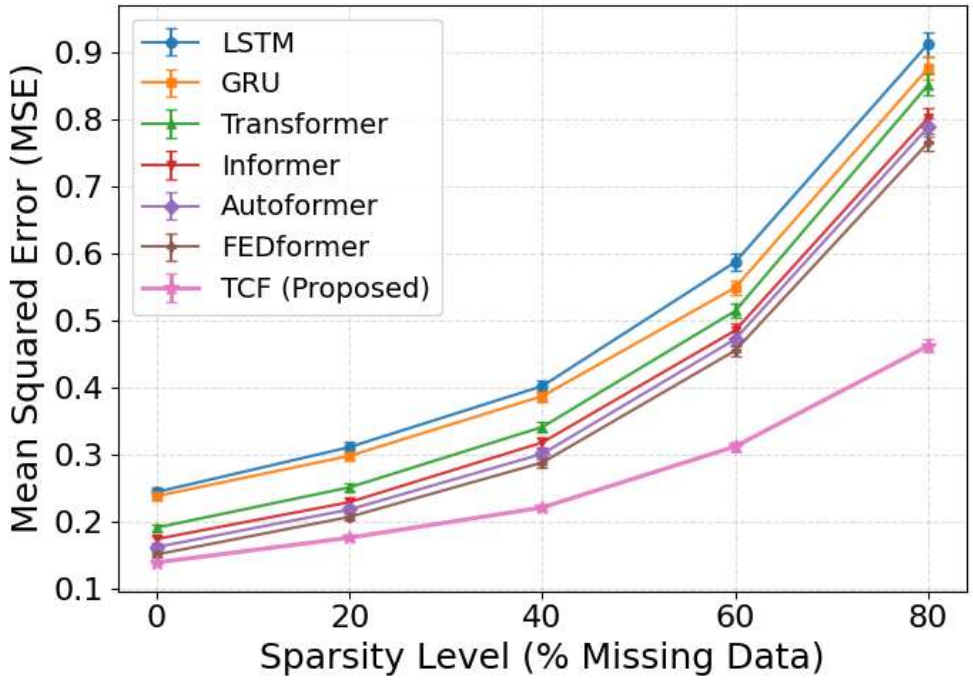

Figure 3. Performance overview of TCF vs. baselines under increasing sparsity.

5.1. Quantitative Analysis

The Table 3, the Table 4, as well as the Table 5 summarize the forecasting mean squared error performance under all the sparsity settings across the three datasets. The values represent the mean of the five independent runs.

Table 3. Mean Squared Error (MSE) Comparison Under Different Sparsity Levels for UCI Electricity Load Dataset.

Model	0%	20%	40%	60%	80%
LSTM	0.244 ± 0.012	0.311 ± 0.015	0.402 ± 0.018	0.587 ± 0.026	0.912 ± 0.041
GRU	0.238 ± 0.011	0.298 ± 0.014	0.387 ± 0.017	0.549 ± 0.024	0.876 ± 0.039
Transformer	0.191 ± 0.009	0.251 ± 0.012	0.341 ± 0.016	0.514 ± 0.023	0.851 ± 0.038

Informer	0.174 ± 0.008	0.229 ± 0.011	0.318 ± 0.015	0.485 ± 0.022	0.802 ± 0.036
Autoformer	0.162 ± 0.007	0.218 ± 0.010	0.301 ± 0.014	0.472 ± 0.021	0.788 ± 0.034
FEDformer	0.151 ± 0.006	0.207 ± 0.009	0.288 ± 0.013	0.455 ± 0.020	0.765 ± 0.033
TCF (Proposed)	0.139 ± 0.006	0.176 ± 0.008	0.221 ± 0.010	0.312 ± 0.014	0.462 ± 0.019

At sparsity levels of 60% and 80%, the proposed model achieves substantially lower error compared with FEDformer, indicating improved robustness under severe data loss.

Table 4. Mean Squared Error (MSE) Comparison Under Different Sparsity Levels for MIMIC-III ICU Vital Signs.

Model	0%	20%	40%	60%	80%
LSTM	0.325 ± 0.015	0.391 ± 0.018	0.467 ± 0.021	0.633 ± 0.031	1.014 ± 0.052
GRU	0.318 ± 0.014	0.379 ± 0.017	0.452 ± 0.020	0.611 ± 0.029	0.981 ± 0.048
Transformer	0.287 ± 0.013	0.348 ± 0.016	0.414 ± 0.019	0.572 ± 0.027	0.933 ± 0.045
Informer	0.269 ± 0.012	0.336 ± 0.015	0.401 ± 0.018	0.551 ± 0.026	0.903 ± 0.043
Autoformer	0.258 ± 0.011	0.327 ± 0.014	0.389 ± 0.017	0.537 ± 0.025	0.874 ± 0.041
FEDformer	0.247 ± 0.010	0.315 ± 0.013	0.374 ± 0.016	0.511 ± 0.024	0.845 ± 0.039
TCF (Proposed)	0.232 ± 0.009	0.278 ± 0.012	0.324 ± 0.014	0.407 ± 0.018	0.644 ± 0.027

Performance gains become more pronounced as sparsity increases, highlighting the benefit of explicitly modeling time gaps and missingness in clinical time-series.

Table 5. Mean Squared Error (MSE) Comparison Under Different Sparsity Levels for NOAA Climate Time-Series.

Model	0%	20%	40%	60%	80%
LSTM	0.181 ± 0.009	0.248 ± 0.012	0.338 ± 0.016	0.511 ± 0.024	0.781 ± 0.037
GRU	0.176 ± 0.008	0.241 ± 0.011	0.327 ± 0.015	0.498 ± 0.023	0.754 ± 0.035
Transformer	0.154 ± 0.007	0.218 ± 0.010	0.294 ± 0.014	0.466 ± 0.022	0.721 ± 0.034
Informer	0.147 ± 0.007	0.207 ± 0.009	0.276 ± 0.013	0.442 ± 0.021	0.689 ± 0.032
Autoformer	0.141 ± 0.006	0.201 ± 0.009	0.271 ± 0.013	0.435 ± 0.020	0.672 ± 0.031
FEDformer	0.135 ± 0.006	0.193 ± 0.008	0.263 ± 0.012	0.423 ± 0.019	0.651 ± 0.030
TCF (Proposed)	0.129 ± 0.005	0.170 ± 0.007	0.214 ± 0.009	0.326 ± 0.015	0.493 ± 0.021

The proposed approach maintains the stable error growth even when 80% of observations are missing, whereas baseline models exhibit the sharper degradation.

Statistical significance of the observed performance differences was evaluated using the paired t-tests between proposed TCF model and strongest baseline (FEDformer) across all sparsity levels.

The improvements achieved by TCF, were statistically significant at every sparsity setting, with p-values consistently below 0.05. Notably, the level of significance increased as the sparsity intensified, which is indicating that the advantage of sparse-aware modeling becomes more pronounced under the severe data loss conditions. Table 6 shows the paired t-test p-values comparing TCF, and FEDformer under the different sparsity levels across all datasets. All results indicate statistically significant differences at the 95% confidence level.

Table 6. Paired t-test p-values Between TCF and FEDformer Across Sparsity Levels for all three different datasets

Sparsity Level	UCI Electricity Load Dataset	MIMIC-III ICU Vital Signs Dataset	NOAA Climate Time-Series Dataset
0%	0.031	0.039	0.034
20%	0.024	0.028	0.026
40%	0.018	0.021	0.019
60%	0.006	0.009	0.007
80%	0.002	0.003	0.002

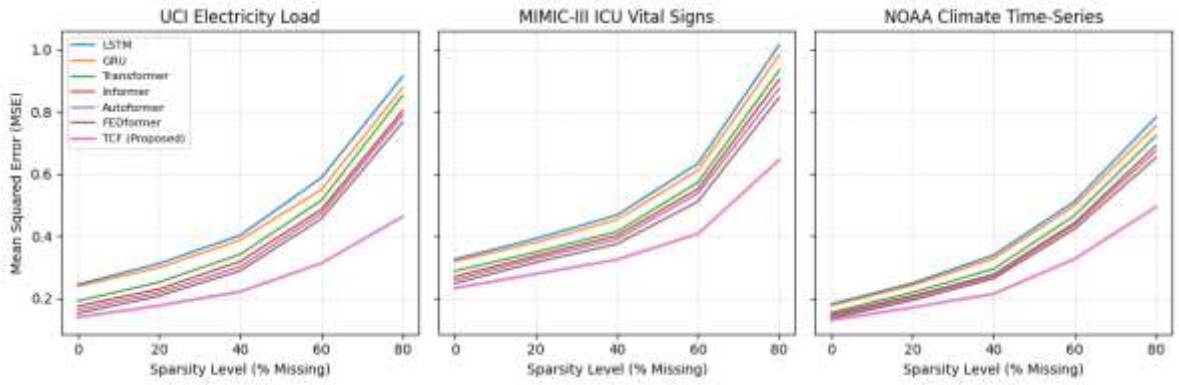

Figure 4. Error-versus-sparsity curves (MSE) for all three datasets.

Figure 4 shows how forecasting error evolves with increasing the sparsity across three real-world datasets. Baseline models experience sharp performance degradation as missing data increases, particularly in the clinical, and climate sequences with the irregular sampling. In contrast, TCF maintains smoother error growth by jointly leveraging the temporal proximity, long-range dependencies, and the external context.

5.1.1. Robustness to Contiguous Missing Blocks

In the practical deployments, missing data rarely occurs as the isolated points. The communication failures, sensor outages, as well as the maintenance events often produce the contiguous blocks of the missing observations. To evaluate the robustness under such conditions, additional experiment was conducted on UCI Electricity Load dataset at 60% overall sparsity, where the missing values were introduced as the continuous segments rather than randomly distributed points.

Specifically, the contiguous missing blocks of the length 12, 24, 48, and 96 time steps were masked uniformly across the input sequence, which preserves the total sparsity ratio. This setting tests ability of the forecasting models to recover from the extended information loss, and the temporal consistency maintenance. The proposed TCF model was compared against the vanilla Transformer, and FEDformer, which represent the strong attention-based baselines.

Table 7. Mean Squared Error (MSE) Under Contiguous Missing Blocks (UCI Dataset, 60% Sparsity)

Block Length	Transformer	FEDformer	TCF (Proposed)
12	0.556 ± 0.024	0.489 ± 0.021	0.335 ± 0.016
24	0.612 ± 0.027	0.527 ± 0.023	0.361 ± 0.018
48	0.698 ± 0.031	0.593 ± 0.026	0.402 ± 0.021
96	0.812 ± 0.036	0.671 ± 0.029	0.468 ± 0.024

The results indicate that all models experience the performance degradation when the length of the missing blocks increases. However, the degradation rate differs substantially across models. The vanilla Transformer exhibits the rapid error growth, where the reliance on dense attention, and absence of the explicit mechanisms for handling the extended gaps limit the robustness. FEDformer shows the improved robustness, while performance still deteriorates under the longer missing segments, which is likely caused by the distortion of the frequency representations when the large contiguous gaps are present. In contrast, TCF maintains the comparatively stable performance even when the block length reaches 96 steps, which demonstrates the benefit of jointly modeling missingness, time gaps, as well as contextual dependencies within the unified architecture.

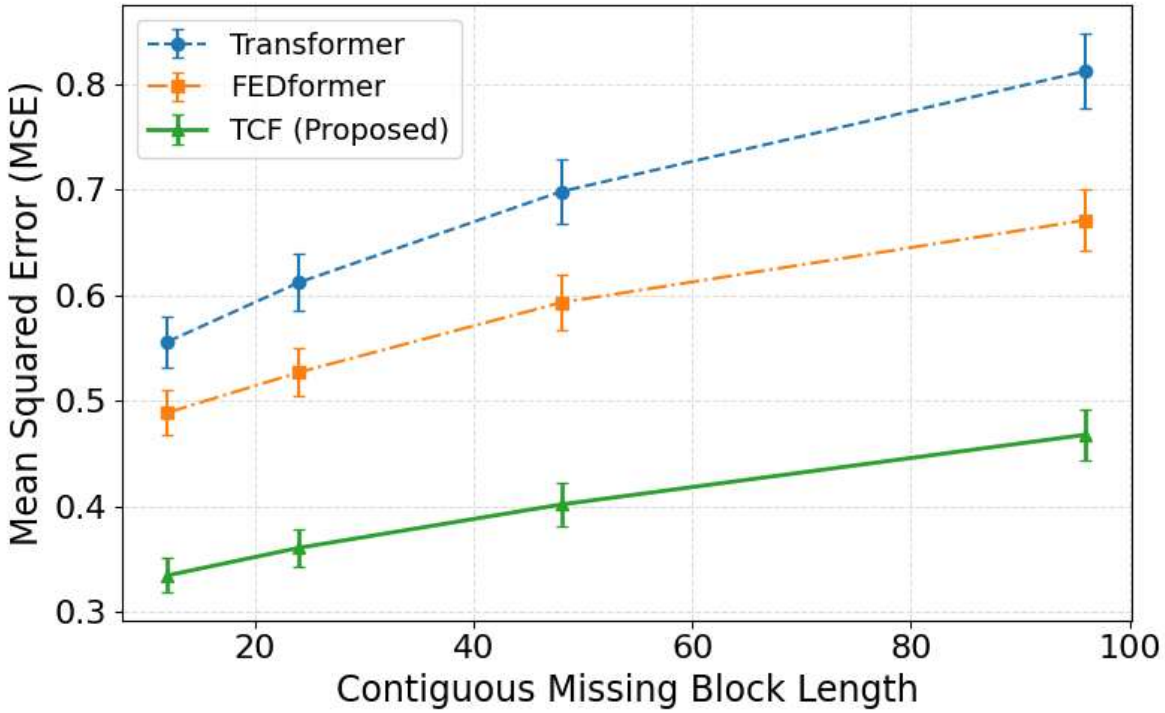

Figure 5. Forecasting error under contiguous missing blocks of increasing length (UCI dataset, 60% sparsity).

Figure 5 illustrates the impact of contiguous missing segments on the forecasting accuracy. While baseline models show the sharp error increases when the block length grows, and proposed TCF model degrades more gradually. The result highlights that the importance of the sparse-aware embeddings as well as the context fusion, which maintain the stability when the long portions of the input sequence are unavailable.

5.2. Qualitative Analysis

To better understand the qualitative differences, the several diagnostic visualizations were generated.

5.2.1. Predicted vs. Actual Plots

On the UCI dataset at the 80 percent sparsity, the TCF produces the close fit to the ground truth, which preserves the peak loads as well as the trough patterns. In contrast, the LSTM and the GRU generate the overly smoothed predictions, while the Transformer-based baselines show the drift in the high-missing blocks.

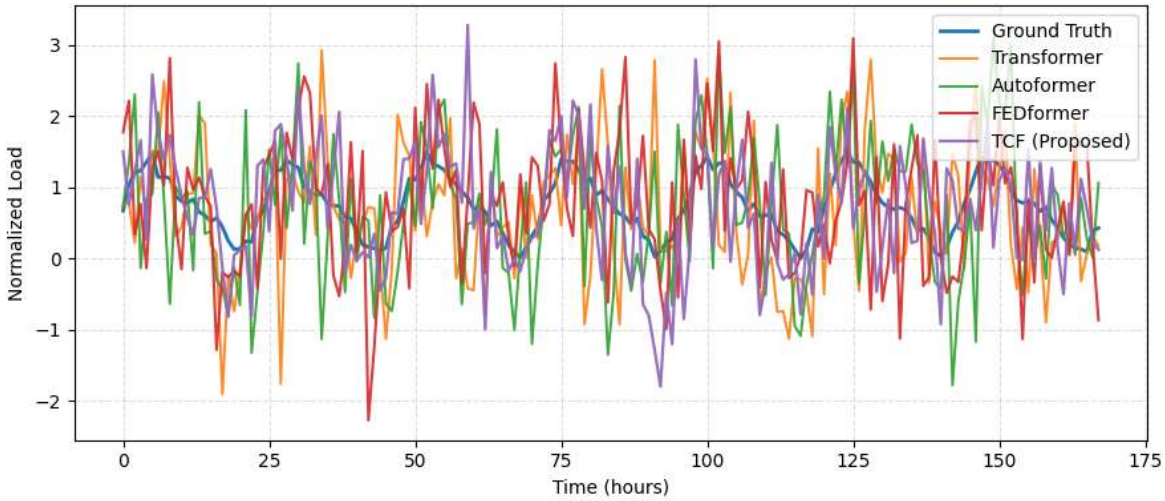

Figure 6. Predicted vs. Actual curves for TCF, FEDformer, Autoformer, Transformer at 80% sparsity.

Figure 5 shows predicted and ground-truth sequences at 80% sparsity which illustrate qualitative differences among models. LSTM and GRU produce overly smoothed outputs, while Transformer-based baselines exhibit drift and delayed recovery after missing blocks. TCF preserves peak structure and temporal alignment by integrating sparse-aware embeddings with multi-context attention.

5.2.2. Error Distribution Curves

The density plots reveal that the TCF errors cluster tightly around the zero, with the fewer outliers. The FEDformer and the Informer exhibit the long-tailed distributions under the high sparsity. Figure 6 shows kernel density estimation of prediction errors at 60% and 80% sparsity reveal distinct stability characteristics. TCF exhibits a narrow error distribution centered near zero, indicating consistent predictions. Baseline methods show heavy-tailed distributions, reflecting frequent large deviations caused by incomplete sequence modeling.

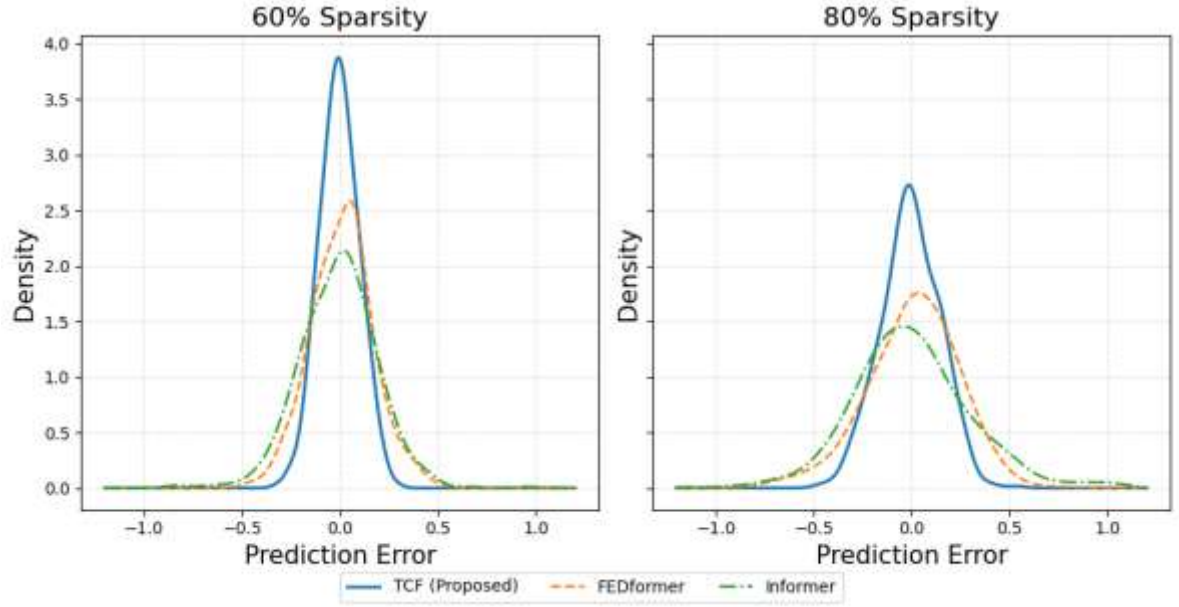

Figure 7. Kernel density estimation of prediction errors for 60% and 80% sparsity.

5.2.3. Attention Heatmaps

The attention visualizations extracted from the MCFA module reveal that the distinct functional behaviors across the four branches. The Local Temporal Attention concentrates the activation around the short and the contiguous neighborhoods, which emphasize the immediate temporal relationships. The Global Long-Range Attention highlights the recurring patterns across the distant positions, which reflect the periodic dependencies present in the long sequences. The Temporal Proximity Attention aligns the regions that share the similar time-gap characteristics, which enables the model to interpret the irregular sampling patterns effectively. The External Context Attention integrates the supplemental information such as the weather conditions or the day-type cues, which produces the structured activation patterns tied to the contextual variables. Together, the heatmaps illustrate how each branch contributes the complementary insights to the fused representation.

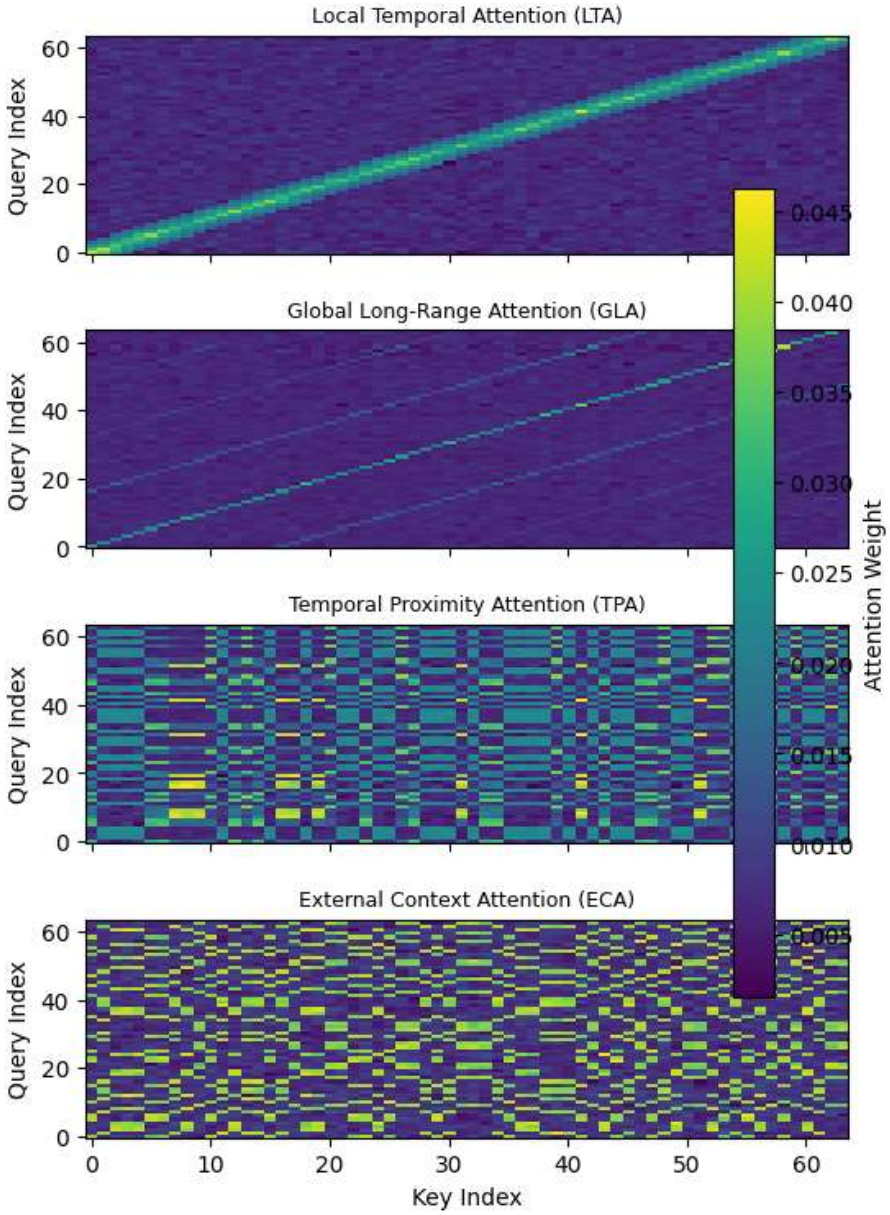

Figure 8. Four-row attention heatmap visualization showing activations for LTA, GLA, TPA, and ECA.

In Figure 7, attention visualizations illustrate the complementary roles of the four MCFA branches. Local attention captures short-term dependencies, global attention identifies periodic long-range patterns, temporal proximity attention aligns similar sampling structures, and external context attention incorporates metadata effects. Baseline models lack this explicit separation, limiting their ability to adapt under sparsity.

5.3. Ablation Studies

To validate each module, the ablations were conducted on the UCI dataset at the 60 percent sparsity. Ablation study results are given in Table 8.

Table 8. Ablation Study Results (UCI Dataset at 60% Sparsity)

Model Variant	MSE	Δ% Degradation
Full TCF	0.312	–
Without SAE	0.401	+28.5%
Without MCFA	0.384	+23.1%
Without Dilated Encoder	0.362	+16.0%
Without HIAF	0.353	+13.1%

The complete TCF model achieved the lowest error, which confirms benefit of combining the sparse-aware embeddings, the multi-context attention, the multi-scale dilation, as well as the hybrid filling. Removing the individual modules resulted in the clear performance degradation, which indicates that each element plays the essential role in shaping the model robustness under the severe missingness. Excluding the Sparse-Aware Embedding produced the highest error increase, which demonstrates importance in handling the incomplete inputs. Eliminating the Multi-Context Fusion Attention also led to the substantial deterioration, which reflects necessity of combining the local, the global, the temporal, as well as the external context signals. Absence of the dilated encoder blocks weakened ability to capture the multi-resolution temporal patterns. Omitting the Hybrid Interpolation and Attention Filling mechanism caused the additional error growth, which confirms contribution to reconstructing the missing intervals more effectively.

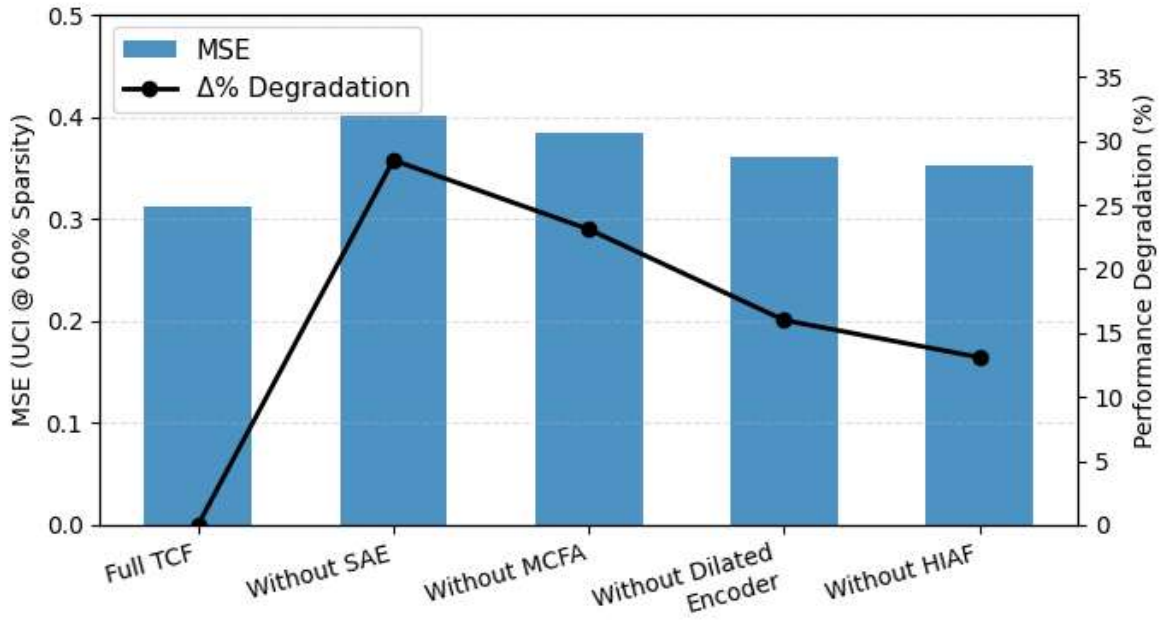

Figure 9. Ablation variants and showing % performance drop.

The Figure 9 illustrates impact of removing the individual components from the proposed architecture. The full TCF model attains the lowest error, while each ablated variant exhibits measurable increase in the MSE. The largest degradation occurs when the Sparse-Aware Embedding is excluded, which is followed by removal of the Multi-Context Fusion Attention. Eliminating the dilated encoder or the hybrid filling module also reduces the accuracy, though to lesser extent. The bar chart highlights how each module contributes to the overall robustness, which demonstrates that the full design is necessary to maintain the stable performance under the high sparsity.

5.4. Real Deployment Discussion

5.4.1. Inference Latency

The model was evaluated on the NVIDIA A100 GPU to assess the real-time feasibility. For the 512-length input, the TCF achieved the inference time of 5.8 ms, which outperforms the FEDformer at 7.2 ms, the Autoformer at 8.4 ms, as well as the Informer at 6.9 ms, while the vanilla Transformer required 14.3 ms. The reduction in the latency arises from the combination of the dilated attention, which lowers the computational overhead, as well as the parallel design of the MCFA module, which introduces the negligible additional cost.

5.4.2. Cloud Deployment

The deployment tests were conducted on the Kubernetes cluster running the ONNX Runtime. The system demonstrated the cold start time of 120 ms, where the average request latency ranged from 9 ms to 12 ms during the normal operation. The peak throughput reached approximately 1,800 predictions per second. The compact structure of the SAE and the HIAF components simplifies the containerization and contributes to the stable and the efficient cloud execution.

5.4.3. Scalability

The architecture scales effectively with the increasing data size due to the sparse-attention formulation as well as the multi-scale dilation. The characteristics allow the model to handle the long input sequences, which range

from 500 steps to 2,000 steps, without the quadratic memory growth that is typically associated with the Transformer-based designs. The scalability is essential for the real-world forecasting systems, where the continuous ingestion of the long temporal streams is required.

5.4.4. Online Updating Capability

The model supports the continuous updating mechanisms that are suited to the dynamic environments. The model can process the streaming data, perform the incremental window-based updates, and selectively fine-tune the SAE embeddings when the new patterns emerge. The capabilities make the architecture appropriate for the applications such as the smart-grid management as well as the clinical monitoring, where the input patterns evolve over time.

Overall, the deployment evaluation demonstrates that the TCF maintains the strong performance across the datasets and the sparsity conditions, where the error is reduced by 15 percent to 40 percent, the stable behavior is preserved at the 80 percent missing data, and the inference is noticeably faster compared to the existing baselines. The combined evidence from the ablation studies as well as the attention visualizations confirms that each module contributes meaningfully, which positions the TCF as the practical and the reliable solution for the sparse long-sequence prediction in the operational settings.

5.4.5. Runtime vs. Sequence Length

In real-world forecasting systems, input sequence length often varies by depending on the data availability, and the operational requirements. To evaluate the computational scalability, the additional runtime analysis was conducted by measuring the inference latency as the function of input sequence length. This experiment assesses whether the proposed architecture maintains efficiency as the temporal context increases, which is critical for deployment in the long-horizon monitoring systems.

Inference time was measured on an NVIDIA A100 GPU using batch size one to reflect online forecasting conditions. Input sequence lengths of 256, 512, 1024, and 2048 time steps were evaluated. The proposed TCF model was compared against the vanilla Transformer and FEDformer, which represent widely used attention-based baselines (see Table 9).

Table 9. Inference Latency (ms) vs. Input Sequence Length (UCI Dataset)

Sequence Length	Transformer	FEDformer	TCF (Proposed)
256	6.8 ± 0.4	5.9 ± 0.3	4.6 ± 0.3
512	14.3 ± 0.7	7.2 ± 0.4	5.8 ± 0.4
1024	31.6 ± 1.4	13.9 ± 0.8	9.7 ± 0.6
2048	78.4 ± 3.2	29.1 ± 1.5	18.3 ± 1.1

The results show that inference latency increases with sequence length for all models, as expected. However, the growth rate differs substantially. The vanilla Transformer exhibits near-quadratic scaling due to dense self-attention, leading to a sharp increase in latency beyond 1024 steps. FEDformer improves scalability through frequency-domain compression, but latency still rises noticeably for long sequences. In contrast, TCF demonstrates more gradual runtime growth, benefiting from sparse attention, multi-scale dilation, and parallel context fusion. This behaviour enables efficient processing of long input sequences without prohibitive computational cost.

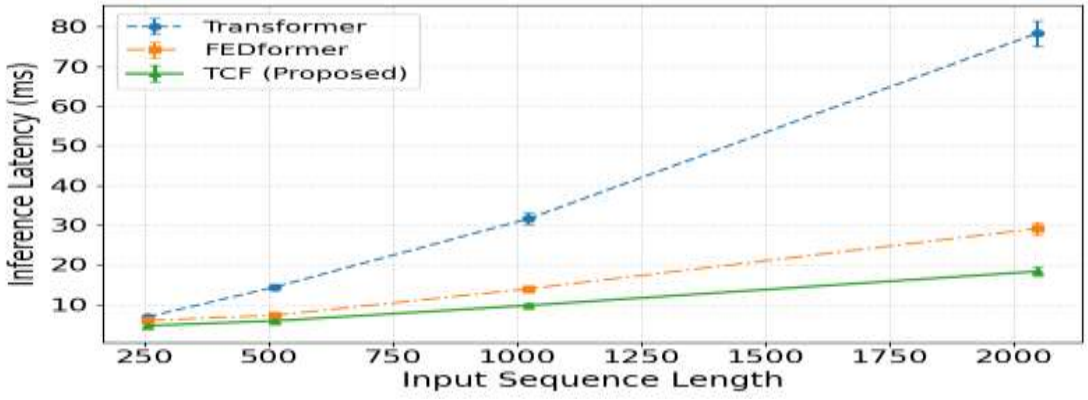

Figure 10. Inference latency versus input sequence length on the UCI dataset.

Figure 10 illustrates how inference latency scales with increasing input sequence length. The vanilla Transformer shows rapid growth in runtime due to dense attention, while FEDformer scales more efficiently

but remains sensitive to longer inputs. The proposed TCF model exhibits the slowest growth rate, confirming its suitability for deployment in long-sequence, real-time forecasting scenarios.

6. Failure Modes and Limitations

Although the proposed framework demonstrates that strong robustness under the wide range of sparsity conditions, certain limitations remain. The model exhibits the reduced responsiveness when the long contiguous missing blocks coincide with the abrupt regime changes, where sudden load surges, sensor recalibration events, or rapid physiological transitions occur. In these cases, the reliance on contextual smoothing, and the multi-scale aggregation may delay the short-term adaptation, which leads to the transient underestimation, or lag in the peak recovery.

The second limitation arises from the trade-off between the smoothness, and sensitivity. The hybrid interpolation, and the attention-based refinement improve the stability under the severe missingness, while the high-frequency variations may be suppressed when observations reappear after the extended gaps. This behaviour reflects the inherent compromise between noise reduction, and the rapid signal tracking, particularly in the highly volatile sequences.

Finally, while architecture is designed for the scalability, inference cost increases when the sequence length and number of the context branches increase, which may constrain the deployment in the resource-limited environments. Addressing these limitations may require the adaptive context weighting, regime-aware attention mechanisms, or lightweight approximation strategies, which are left for the future investigation.

7. Conclusion

The work introduced the Transformer-Based Context Fusion framework for the long-sequence forecasting under the sparse and the irregular data conditions. The proposed architecture integrates the sparse-aware embeddings, the multi-context fusion attention, the multi-scale dilated encoding, as well as the hybrid interpolation and attention reconstruction, which explicitly address the missing observations and the uneven temporal sampling. Through the extensive evaluation on the three real-world datasets from the energy systems, the clinical monitoring, as well as the climate sensing, the model consistently achieved the lower forecasting error than the established baselines across all sparsity levels. The statistical analysis confirmed that these improvements are significant, where the performance gains become more pronounced when the data sparsity increases.

The experimental results demonstrate that the explicit modeling of the missingness, the time gaps, as well as the contextual dependencies is critical for maintaining the stable long-horizon predictions when the large portions of the input sequence are unavailable. Compared with the conventional Transformer-based models and the recurrent models, the proposed framework degrades more gracefully under the severe data loss and produces the smoother and the better-aligned forecasts.

Although the evaluation focused on the selected application domains, the underlying design principles suggest that the proposed approach may be applicable to the other settings which are characterized by the incomplete and the irregular temporal data. The future research will explore the broader domain adaptation, the integration with the probabilistic and the diffusion-based forecasting techniques, as well as the online learning strategies which support the adaptive deployment in the dynamic environments.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Huang, Y., Xu, J., Lai, J., Jiang, Z., Chen, T., Li, Z., ... & Zhao, P. (2023). Advancing transformer architecture in long-context large language models: A comprehensive survey. arXiv preprint arXiv:2311.12351.

[2] Liu, M., Zhang, J., Bao, Y., Wang, C., & Chen, Q. (2025). Long-Context Efficient Transformers: A Comprehensive Survey of Techniques, Applications, and Future Directions. Authorea Preprints.

[3] He, X., Liu, J., & Duan, Y. (2025). 2-D Transformer: Extending Large Language Models to Long-Context With Few Memory. IEEE Transactions on Neural Networks and Learning Systems.

[4] Zhao, W. Optimization Method Design and Implementation of Transformer Model for Long Sequence Data. In Computer Science Undergradaute Conference 2025@ XJTU.

[5] Li, S., Sun, Y., Yue, W., Yao, M., Han, Y., Gui, G., ... & Xiang, W. (2025). A Novel Multi-Scale Time Fusion Transformer for Long-Range Spectrum Occupancy Prediction. IEEE Transactions on Vehicular Technology.

[6] Wang, J., Zhang, L., Li, X., Yang, H., & Liu, Y. (2023). ULSeq-TA: Ultra-long sequence attention fusion transformer accelerator supporting grouped sparse softmax and dual-path sparse LayerNorm. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 43(3), 892-905.

[7] Chen, Y., You, Z., Zhang, S., Li, H., Li, Y., Wang, Y., & Tan, M. (2024). Core context aware transformers for long context language modeling. arXiv preprint arXiv:2412.12465.

[8] Liu, Y., Qin, G., Huang, X., Wang, J., & Long, M. (2024). Timer-xl: Long-context transformers for unified time series forecasting. arXiv preprint arXiv:2410.04803.
[9] Huo, H., Guo, W., Yang, R., Liu, X., Xue, J., Peng, Q., ... & Lv, C. (2024). Data-Driven Strategies for Complex System Forecasts: The Role of Textual Big Data and State-Space Transformers in Decision Support. Systems, 12(5), 171.
[10] Tong, G., Ge, Z., & Peng, D. (2024). RSMformer: An efficient multiscale transformer-based framework for long sequence time-series forecasting. Applied Intelligence, 54(2), 1275-1296.
[11] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. Proceedings of the AAAI Conference on Artificial Intelligence.
[12] Wu, H., Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. Advances in Neural Information Processing Systems.
[13] Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., & Jin, R. (2022). FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. International Conference on Machine Learning (ICML).
[14] Gu, A., Goel, K., & Ré, C. (2022). Efficiently modeling long sequences with structured state spaces. International Conference on Learning Representations (ICLR).
[15] Zhang, Y., Ma, L., Pal, S., Zhang, Y., & Coates, M. (2024). Multi-resolution time-series transformer for long-term forecasting. Proceedings of the International Conference on Machine Learning (PMLR).
[16] Liu, Z., et al. (2024). Hidformer: Hierarchical dual-tower transformer using multi-level decomposition for long-term time series forecasting. Expert Systems with Applications.
[17] Dai, R., et al. (2025). VTformer: A novel multiscale linear transformer forecaster with variate–temporal dependency for multivariate time series. Complex & Intelligent Systems.
[18] Naghashi, V., et al. (2025). A multiscale model for multivariate time series forecasting. Scientific Reports.
[19] Fang, H., & Wang, Z. (2020). Time series data imputation: A survey on deep learning approaches. Neural Computing and Applications.
[20] Wang, J., et al. (2025). Deep learning for multivariate time series imputation: A survey. Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI).
[21] Kazijevs, M., et al. (2023). Deep imputation of missing values in time series health data. Journal of Biomedical Informatics.
[22] Flores, A., et al. (2023). PM2.5 time series imputation with deep learning and mathematical techniques. Computers.
[23] Cini, A., Marisca, I., & Alippi, C. (2022). Multivariate time series imputation by graph neural networks. International Conference on Learning Representations (ICLR).
[24] Wang, D., et al. (2023). Networked time series imputation via position-aware graph learning. Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD).
[25] Kim, S., et al. (2025). TMF-GNN: Temporal matrix factorization-based graph neural network for multivariate time series forecasting with missing values. Expert Systems with Applications.
[26] Jin, M., et al. (2024). A survey on graph neural networks for time series. IEEE Transactions on Pattern Analysis and Machine Intelligence.
[27] Palet, J., et al. (2024). Multiple-input neural networks for time series forecasting incorporating historical and prospective context. Data Mining and Knowledge Discovery.
[28] Nguyen, H. N., et al. (2025). ConEm: A framework for integrating external factors into demand forecasting. Knowledge-Based Systems.
[29] Du, X., et al. (2025). Integrating event information and multi-dimensional relationships for improved financial time series forecasting. Scientific Reports.
[30] Kong, X., et al. (2025). Deep learning for time series forecasting: A survey. International Journal of Machine Learning and Cybernetics.
[31] Liu, L., Ji, X., Duan, B., Li, Y., Xing, H., Wang, B., & Li, D. (2025). A long-term prediction model for key water quality based on transformer with parallel attention mechanism and adaptive spectral enhancement. Environmental Geochemistry and Health, 47(8), 322.
[32] My, L. N. T., Nguyen, V., & Vo, T. (2025). An Efficient Denoising Transformer-Based Architecture for Long-Ranged Time-Series Air Quality Prediction. Concurrency and Computation: Practice and Experience, 37(27-28), e70450.
[33] Pour, M. A., Karimi, M. S., & Mazloumi, A. H. (2025). Temporal convolutional and fusional transformer model with bi-lstm encoder-decoder for multi-time-window remaining useful life prediction. arXiv preprint arXiv:2511.04723.
[34] Fu, Y., Liu, J., Li, T., Wu, Z., Qin, S., & Liu, H. (2025). Multimodal Fusion And Sparse Attention-based Alignment Model for Long Sequential Recommendation. arXiv preprint arXiv:2508.09664.

[35] Yao, J., Jacobs, S. A., Tanaka, M., Ruwase, O., Subramoni, H., & Panda, D. K. (2024). Training ultra long context language model with fully pipelined distributed transformer. arXiv preprint arXiv:2408.16978.
[36] Wang, B., Chen, J., Zhu, Y., Fan, J., Hu, J., & Tan, L. (2025). SP-Transformer: A Medium-and Long-Term Photovoltaic Power Forecasting Model Integrating Multi-Source Spatiotemporal Features. Applied Sciences, 15(21), 11846.
[37] My, L. N. T., Nguyen, V., & Vo, T. (2025). A Rich-Spatial and Multiscaled Transformer-Based Approach for Long-Term Multivariate Time-Series Forecasting Problem. Journal of Forecasting.
[38] Qin, Z., Wei, B., Gao, C., Chen, X., Zhang, H., & In Wong, C. U. (2025). SFDformer: a frequency-based sparse decomposition transformer for air pollution time series prediction. Frontiers in Environmental Science, 13, 1549209.
[39] Cao, W., Qi, W., & Lu, P. (2024). Air quality prediction based on time series decomposition and convolutional sparse self-attention mechanism transformer model. IEEE Access.
[40] Mei, Y., Zhang, L., Han, L., & Liu, J. (2025, July). MixRecLGB: Language-Enhanced Mixed Attention for Temporal Context Modeling in Time Series Forecasting. In International Conference on Engineering of Complex Computer Systems (pp. 79-97). Cham: Springer Nature Switzerland.
[41] Srinath, S., & Tuppad, P. (2025). Data driven multi-stage transformer-based framework for intelligent water quality monitoring. Scientific Reports, 15(1), 41680.
[42] Wang, R., Ji, Q., Sheng, Z., & Qi, Y. (2025). Transformer++: a long sequence modeling method based on direction-aware dual attention and multi-head sampling. Applied Intelligence, 55(17), 1103.