TSCOMP: CROSS-SCALE COMPOSITIONAL PRE-TRAINING FOR TIME SERIES FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Time series foundation models (TSFMs) aim to learn transferable temporal representations across diverse downstream tasks. However, time series are inherently multi-scale: the observed signal is jointly composed of distinct temporal structures at different scales, each capturing different temporal dynamics. This poses two challenges for transferable pre-training: scale-specific structures should be explicitly distinguished and their cross-scale correspondence must also be preserved because their roles depend on how they interact across scales. Conventional time-domain masked reconstruction does not explicitly enforce either property, leaving both scale-specific structures and cross-scale correspondence implicitly supervised. To address these limitations, we propose TSComp, a cross-scale compositional pre-training framework for TSFMs. TSComp uses scale-balanced tokenization to construct aligned scale-specific representations, scale coefficient attention to model interactions among corresponding structures across scales, and multi-scale directional masking to infer missing components from complementary cross-scale context. Its pre-training objectives jointly supervise scale energy distribution calibration, signal recomposition, and coefficient reconstruction, making scale-specific recovery and cross-scale composition explicit learning targets. The resulting scale-specific representations are adapted to downstream tasks through a scale-aware adaptation method. Experiments across forecasting, imputation, classification, and anomaly detection demonstrate effective transfer across diverse downstream settings and competitive performance against both multi-task foundation models and task-specific architectures. Our code is available at `https://anonymous.4open.science/r/TSComp-66B0`.

1 INTRODUCTION

Time series foundation models (TSFMs) aim to learn general-purpose temporal representations that can be transferred across diverse downstream tasks (Goswami et al., 2024). While recent TSFMs have been developed for specific objectives such as forecasting, classification, and anomaly detection (Shentu et al., 2025; Liu et al., 2025; Xie et al., 2026), their broader goal is to serve as unified backbones that can be efficiently adapted to diverse downstream tasks. Masked pre-training has emerged as an effective paradigm towards this goal, encouraging models to recover missing observations from temporal context and thereby capture intrinsic dependencies and dynamics (Seo & Lim, 2025; Ekambaram et al., 2026).

However, time series are inherently multi-scale: long-term trends and low-frequency variations coexist with short transients and local fluctuations, jointly forming the observed signal across different temporal resolutions (Wang et al., 2024). This presents two challenges for transferable pre-training. First, modeling temporal patterns from different resolutions without explicitly separating them by scale can obscure their distinct structural characteristics, making it difficult to learn scale-specific dynamics. Second, separating scales alone is insufficient, as scale-specific structures are not independent; they jointly describe the underlying temporal dynamics, and their meaning depends on how they interact across resolutions. Therefore, transferable time series representations should explicitly distinguish structures at different temporal scales while preserving the cross-scale compositional relationships through which they coherently form the observed signal.

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.

Table 1: Time-domain and scale-wise reconstruction quality and cross-scale correspondence analysis averaged across eight datasets (Wu et al., 2021). For scale-wise reconstruction, we measure the Pearson correlation between corresponding wavelet coefficients of the original time series and each model's time-domain reconstruction. For cross-scale correspondence, we construct scale-misaligned inputs and train an MLP probe on frozen TSFM representations to distinguish aligned from misaligned inputs. Higher B. Acc., AUROC, and F1 indicate more readily decodable cross-scale correspondence.

Model	Scale-wise Reconstruction					Cross-scale Correspondence Probe		
	Time	A_3	D_3	D_2	D_1	B. Acc.	AUROC	F1
MOMENT (Goswami et al., 2024)	0.043	0.380	-0.071	0.010	-0.007	<u>0.877</u>	<u>0.949</u>	<u>0.886</u>
GTM (He et al., 2026)	0.904	0.925	0.848	<u>0.786</u>	<u>0.652</u>	0.815	0.885	0.823
TSPulse (Ekambaram et al., 2026)	<u>0.940</u>	<u>0.989</u>	<u>0.899</u>	0.725	0.376	0.658	0.702	0.669
TSComp	**0.999**	**1.000**	**1.000**	**0.999**	**0.999**	**0.999**	**1.000**	**0.999**

However, conventional time-domain masked reconstruction does not explicitly address either requirement. By supervising only the reconstructed signal as a whole, it aggregates errors across temporal scales, with each scale contributing in proportion to its signal energy (see Theorem 3). Moreover, time-domain masking removes all scales within the same region, preventing direct supervision of cross-scale inference from the observed scales. Consequently, existing TSFMs can exhibit uneven preservation of scale-specific temporal structures across resolutions. Our analysis in Table 1 reveals that high time-domain reconstruction fidelity does not necessarily imply consistent preservation across temporal scales. The cross-scale correspondence probe further reveals a mismatch between reconstruction fidelity and correspondence discriminability: models with similar or even higher reconstruction fidelity can differ substantially in how clearly cross-scale correspondence is encoded in their representations. These observations indicate that conventional time-domain reconstruction does not consistently guarantee either scale-specific preservation or cross-scale correspondence, motivating pre-training that explicitly targets both properties (see Section 2 for details).

To address these limitations, we propose TSComp, a cross-scale compositional pre-training framework for time series foundation models. Rather than relying on a shared time-domain representation in which structures at different scales remain implicit, TSComp uses scale-balanced tokenization to decompose the time series into scale-specific components and transform them into aligned and balanced token representations. This explicitly separates structures by scale while preserving their temporal correspondence. Building on this representation, the scale coefficient attention (SCA) backbone models interactions across scale-specific coefficients, allowing each scale to retain its distinct structure while incorporating information from corresponding structures at other scales. To make these properties part of the pre-training objective, multi-scale directional masking (MDM) imposes directional masking across aligned scales, allowing the model to infer missing components from complementary cross-scale context. The pre-training objectives jointly supervise scale energy distribution calibration, signal recomposition, and coefficient reconstruction, making scale-specific recovery and cross-scale composition explicit pre-training targets. Finally, the learned scale-specific representations are effectively adapted to downstream tasks, allowing each task and dataset to exploit the temporal scales most relevant to its characteristics. Our main contributions are as follows:

- We identify a limitation of conventional pre-training with time-domain masked reconstruction: scale-specific structures and cross-scale correspondence remain implicitly supervised, leading to uneven scale-wise preservation and leaving cross-scale relations implicit.
- We propose TSComp, a cross-scale compositional pre-training framework that explicitly learns scale-specific components and their cross-scale correspondence through scale-balanced tokenization, scale coefficient attention, multi-scale directional masking, and scale-aware pre-training objectives.
- We develop a scale-aware downstream adaptation method that exploits the learned scale-specific representations according to task- and dataset-specific temporal characteristics. Extensive experiments across forecasting, imputation, anomaly detection, and classification demonstrate that the learned representations transfer effectively across diverse downstream settings, achieving competitive performance against both multi-task foundation models and task-specific architectures.

2 LIMITATIONS OF TIME-DOMAIN RECONSTRUCTION FOR MULTI-SCALE PRE-TRAINING

Scale-wise Reconstruction Fidelity. To examine whether TSFMs pre-trained by time-domain reconstruction preserve structures across temporal resolutions, we apply a three-level discrete wavelet transform (DWT) with the Symlet-4 wavelet and periodization boundary extension to the input series and its reconstruction. This produces $\{A_3, D_3, D_2, D_1\}$, ranging from the coarse approximation to progressively finer detail streams. We measure reconstruction fidelity at each scale using Pearson correlation, which captures agreement in temporal structure independently of dataset- and variable-specific magnitudes.

Cross-scale Correspondence. We further examine whether cross-scale temporal correspondence is encoded in TSFM representations. Using the same three-level DWT, we construct a subtly misaligned series by retaining A_3, shifting one randomly selected detail component from D_3, D_2, D_1 by a single coefficient position, and applying inverse DWT while leaving the remaining components unchanged. We then extract representations of the aligned and misaligned inputs from each frozen TSFM and train an MLP probe to distinguish between them. Higher classification performance suggests that cross-scale temporal correspondence is more explicitly represented in the TSFM embeddings.

Results. Table 1 reports the analysis results averaged over three random seeds. The results show that high time-domain fidelity does not guarantee consistent preservation of scale-specific structures across scales: existing TSFMs generally deteriorate from the coarse approximation A_3 toward the finest detail D_1, whereas TSComp maintains consistently high fidelity across scales. MOMENT produces smooth reconstructions, but its low time-domain and near-zero detail-scale correlations indicate poor preservation of the input's temporal and fine-scale patterns (see Figure 3(a)). More importantly, high reconstruction fidelity does not necessarily imply that cross-scale correspondence is well represented in the learned embeddings: TSPulse shows the strongest reconstruction but the weakest probe performance, while MOMENT shows the reverse. TSComp maintains near-perfect fidelity across all scales while also making cross-scale misalignment distinguishable from its representations. **These results indicate that time-domain reconstruction alone does not reliably induce representations that preserve scale-specific structures and encode cross-scale correspondence.**

3 RELATED WORK

Time-domain Masked Reconstruction in TSFMs. Time-domain masked reconstruction has become a central pre-training objective for learning transferable time series representations. MOMENT masks input patches and reconstructs the missing observations in the time domain (Goswami et al., 2024), while TimesBERT complements masked patch modeling with functional token prediction to capture temporal and inter-variate patterns (Zhang et al., 2025). Subsequent approaches extend this paradigm under the same basic time-domain reconstruction framework: LPTM learns dataset-adaptive segmentation rather than relying on a fixed patch length (Kamarthi & Prakash, 2024), and SymTime aligns masked time series representations with symbolic descriptions (Wang et al., 2025). However, because time-domain reconstruction supervises the recovery of the observed signal in which multiple temporal scales are jointly composed, it does not explicitly separate or supervise the structures associated with individual scales. Consequently, scale-specific reconstruction errors can be obscured within the aggregate time-domain objective. To explicitly preserve scale-specific dynamics, TSComp uses scale-balanced tokenization to decompose the time series into scale-specific components and transform them into scale-separated yet temporally aligned token representations.

Multi-scale Modeling for Time Series Representation Learning. Temporal decomposition and frequency-domain modeling have been widely used to expose structures that are difficult to distinguish in the original signal. FreTS transforms time series into the Fourier domain and applies complex-valued MLPs to learn intra- and inter-series dependencies (Yi et al., 2023). TimeMixer constructs downsampled multi-scale sequences, decomposes them into seasonal and trend components, and mixes information in fine-to-coarse and coarse-to-fine directions for forecasting (Wang et al., 2024). SimpleTM instead uses stationary wavelet transforms to construct multi-resolution tokens and models their interactions through attention-based operations (Chen et al., 2025). Recent TSFMs extend the time-domain reconstruction paradigm by incorporating decomposed or spectral views during pre-training. TSPulse jointly supervises temporal, spectral, and semantic representations (Ekambaram

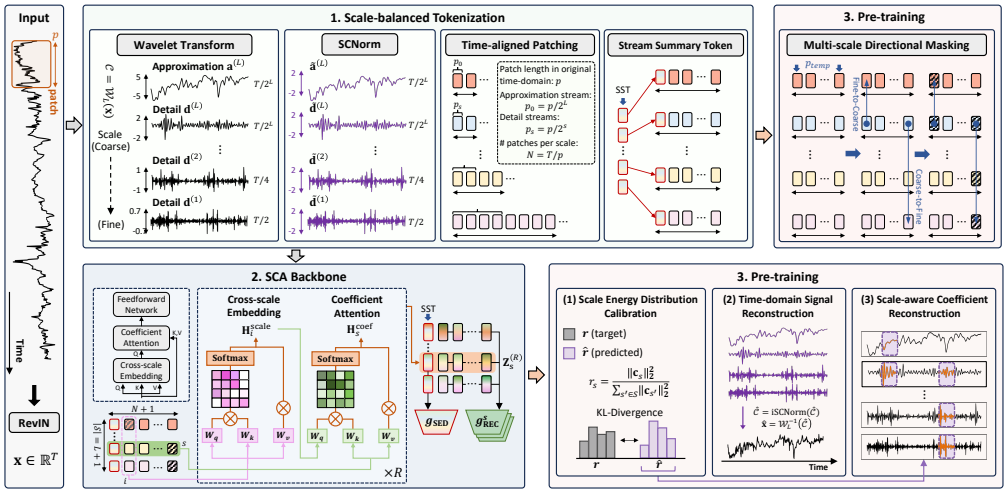


Figure 1: Overview of TSComp. Orange arrows indicate operations used only during pre-training.

et al., 2026), while GTM combines temporal and frequency-domain attention to capture time-granularity-aware features (He et al., 2026). However, these approaches do not explicitly formulate the relations among temporally corresponding scale components as a pre-training target, leaving cross-scale correspondence to be learned only indirectly. In contrast, TSComp introduces cross-scale compositional pre-training that explicitly models and supervises cross-scale correspondence among aligned scale-specific representations.

4 TSComp Architecture

In this section, we present the architecture and pre-training design of TSComp. As shown in Figure 1, TSComp consists of three stages. First, *Scale-balanced Tokenization* decomposes the input time series into scale-specific wavelet coefficient streams and transforms them into balanced, temporally aligned token representations. Second, *Scale Coefficient Attention* (SCA) encodes these representations by modeling cross-scale interactions among temporally corresponding tokens together with temporal dependencies within each coefficient stream. Third, *Multi-scale Directional Masking* (MDM) imposes directional masking across aligned scales, requiring the encoder to infer missing scale-specific components from complementary cross-scale context. The pre-training objectives jointly supervise scale energy distribution calibration, time-domain signal reconstruction, and scale-aware coefficient reconstruction, making scale-specific recovery and cross-scale composition explicit learning targets. The learned representations are transferred to downstream tasks through scale-aware adaptation strategies.

4.1 Scale-balanced Tokenization

Wavelet Transform. Given a raw univariate input window $\mathbf{x}^{\text{raw}} \in \mathbb{R}^T$, TSComp first applies Reversible Instance Normalization (RevIN) (Kim et al., 2022) to mitigate distribution shifts, yielding $\mathbf{x} \in \mathbb{R}^T$. To support scale-specific structure learning and cross-scale composition, the decomposition should provide temporal localization, an explicit scale hierarchy, perfect recomposition, and energy preservation. We adopt an orthonormal wavelet transform using Symlet-4, which satisfies these structural properties; further details are provided in Appendix A. It applies an L-level DWT $\mathcal{W}_L$ and obtains $\mathcal{C} = \mathcal{W}_L(\mathbf{x}) = \{\mathbf{a}^{(L)}, \mathbf{d}^{(L)}, \ldots, \mathbf{d}^{(1)}\}$, where $\mathbf{a}^{(L)}$ is the coarsest approximation stream and $\mathbf{d}^{(\ell)}$ is the detail stream at level ℓ. For unified indexing, we define the stream index set $\mathcal{S} = \{0, 1, \ldots, L\}$ and denote the coefficient stream indexed by $s \in \mathcal{S}$ as $\mathbf{c}_s$, with $\mathbf{c}_0 = \mathbf{a}^{(L)}$ and $\mathbf{c}_\ell = \mathbf{d}^{(\ell)}$ for $\ell = 1, \ldots, L$. Additional background on the wavelet transform is provided in Appendix B.

Scale-wise Coefficient Normalization. Although RevIN normalizes the time-domain input, wavelet streams can have substantially different magnitudes across scales. Directly processing raw coefficients may bias cross-scale operations toward high-energy coarse streams and suppress informative fine-scale variations. Thus, we introduce *Scale-wise Coefficient Normalization* (SCNorm), which normalizes

each stream before tokenization. The approximation stream is standardized as $\tilde{\mathbf{c}}_0 = (\mathbf{c}_0 - \mu_0)/(\sigma_0 + \epsilon)$, where μ_0 and σ_0 are the mean and standard deviation of $\mathbf{c}_0$, and $\epsilon > 0$ is a small constant. Each detail stream is RMS-normalized as $\tilde{\mathbf{c}}_s = \mathbf{c}_s/(\mathrm{RMS}(\mathbf{c}_s) + \epsilon)$ for $s = 1, \ldots, L$. SCNorm provides a scale-balanced representation space for learning cross-scale structural relations, and the original coefficient scales are restored via inverse SCNorm (iSCNorm) before inverse DWT (iDWT).

Time-aligned Patching. Let $\tilde{\mathcal{C}} = \{\tilde{\mathbf{c}}_s : s \in \mathcal{S}\}$ denote the normalized streams. To preserve temporal correspondence across resolutions, each stream is partitioned so that the i-th patch corresponds to the same temporal region of the original input across all scales. We define p as the patch length in the original time domain. For each detail stream $s = 1, \ldots, L$, whose temporal resolution is $1/2^s$ of the input, we set $p_s = p/2^s$. For the approximation stream ($s = 0$, corresponding to DWT level L), we set $p_0 = p/2^L$. This gives all streams the same number of patches, $N = T/p$. We denote the i-th patch of normalized stream $\tilde{\mathbf{c}}_s$ by $\mathbf{p}_{s,i} \in \mathbb{R}^{p_s}$, for $s \in \mathcal{S}$ and $i = 1, \ldots, N$. Each patch $\mathbf{p}_{s,i}$ is then projected by a stream-specific embedding layer E_s, producing $\mathbf{z}_{s,i} = E_s(\mathbf{p}_{s,i}) \in \mathbb{R}^D$.

Stream Summary Token. We prepend a learnable *stream summary token* (SST) to each stream, denoted by $\mathbf{z}_{s,0} = \mathbf{z}_s^{\mathrm{SST}} \in \mathbb{R}^D$. Similar to the `[CLS]` token in Vision Transformers (Dosovitskiy et al., 2021), the SST aggregates stream-level information at each scale. The input sequence for stream s is $\mathbf{Z}_s^{(0)} = [\mathbf{z}_{s,0}; \mathbf{z}_{s,1}; \ldots; \mathbf{z}_{s,N}] \in \mathbb{R}^{(N+1)\times D}$. These SSTs provide compact scale-specific summaries and are later used to estimate the window-specific scale energy distribution.

4.2 Scale Coefficient Attention

Understanding time series requires modeling both temporal dynamics within each resolution and compositional interactions across resolutions. After the wavelet transform, each coefficient stream captures a single scale-specific component of the signal, while the complete temporal pattern emerges from their joint composition. Thus, we introduce *scale coefficient attention* (SCA), a factorized attention block with two stages: (i) *cross-scale embedding*, which exchanges information among resolutions over the same temporal support, and (ii) *coefficient attention*, which models temporal dependencies within each coefficient stream conditioned on the cross-scale context.

At the r-th SCA block, let $\mathbf{Z}_s^{(r-1)} = [\mathbf{z}_{s,0}^{(r-1)}; \mathbf{z}_{s,1}^{(r-1)}; \ldots; \mathbf{z}_{s,N}^{(r-1)}] \in \mathbb{R}^{(N+1)\times D}$ denote the token sequence of stream s. For each aligned token index $i \in \{0, 1, \ldots, N\}$, where $i = 0$ corresponds to the SSTs and $i \geq 1$ corresponds to patch tokens, we collect tokens from all streams as $\mathbf{Z}_i^{(r-1)} = [\mathbf{z}_{s,i}^{(r-1)}]_{s\in\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|\times D}$. First, cross-scale embedding applies self-attention over $\mathbf{Z}_i^{(r-1)}$:

$$\mathbf{H}_i^{\mathrm{scale}} = \mathrm{MHA}\Big(\mathbf{Z}_i^{(r-1)}, \mathbf{Z}_i^{(r-1)}, \mathbf{Z}_i^{(r-1)}\Big) \in \mathbb{R}^{|\mathcal{S}|\times D}. \tag{1}$$

This stage allows each token to aggregate information from other resolutions that describe the same temporal region. The outputs $\{\mathbf{H}_i^{\mathrm{scale}}\}_{i=0}^N$ are then rearranged into stream-wise sequences $\mathbf{H}_s^{\mathrm{scale}} \in \mathbb{R}^{(N+1)\times D}$. Then, coefficient attention uses the cross-scale contextualized sequence as queries and the original stream sequence as keys and values:

$$\mathbf{H}_s^{\mathrm{coef}} = \mathrm{MHA}\Big(\mathbf{H}_s^{\mathrm{scale}}, \mathbf{Z}_s^{(r-1)}, \mathbf{Z}_s^{(r-1)}\Big) \in \mathbb{R}^{(N+1)\times D}. \tag{2}$$

This stage integrates cross-scale context with within-stream temporal dependencies. Compared with applying attention over all tokens jointly, SCA imposes a wavelet-aligned inductive bias by modeling (i) cross-scale interactions among temporally aligned coefficients and (ii) temporal dependencies within each coefficient stream. After R stacked SCA blocks, the final representations are denoted by $\mathbf{Z}^{(R)} = \{\mathbf{Z}_s^{(R)}\}_{s\in\mathcal{S}}$, with $\mathbf{Z}_s^{(R)} \in \mathbb{R}^{(N+1)\times D}$.

4.3 Pre-training

Multi-scale Directional Masking. Because time-aligned patching ensures that the i-th patch corresponds to the same temporal region across streams, we define the aligned patch group $\mathcal{A}_i = \{\mathbf{p}_{s,i} : s \in \mathcal{S}\}$. To encourage learning of cross-scale composition, we propose *Multi-scale Directional Masking* (MDM). Given a temporal selection ratio p_{temp}, we first sample a subset of temporal indices $\mathcal{I}$ with $|\mathcal{I}| \approx p_{\mathrm{temp}} N$. For each $i \in \mathcal{I}$, we choose a reference stream patch in $\mathcal{A}_i$ and

a masking direction, either coarse-to-fine or fine-to-coarse, masking the reference patch together with its aligned finer- or coarser-scale patches, respectively. This forces the model to reconstruct missing components from complementary scales within the same temporal support. The resulting masked normalized coefficient streams are denoted by $\tilde{\mathcal{C}}_\Omega = \{\tilde{\mathbf{c}}_{s,\Omega} : s \in \mathcal{S}\}$, with mask indices $\Omega = \{\Omega_s : s \in \mathcal{S}\}$. Under uniform sampling of the reference stream and direction, the expected masking ratio is $p_{\text{temp}}(|\mathcal{S}| + 1)/(2|\mathcal{S}|)$. In MDM, masking coefficient patches after DWT avoids the scale-dependent boundary deformation caused by pre-DWT time-domain masking (see Appendix G.1).

Scale Energy Distribution Calibration. SCNorm removes explicit scale-dependent magnitude differences, enabling the encoder to learn cross-scale structural relations without relying on raw coefficient amplitudes. However, relative amplitudes remain essential for faithful cross-scale composition, as different streams contribute unequally to the original time series. To retain this compositional information, we introduce *Scale Energy Distribution Calibration*. For each unnormalized stream $s \in \mathcal{S}$, we define the target scale energy distribution $\mathbf{r} = (r_s)_{s\in\mathcal{S}}$, where $r_s = \|\mathbf{c}_s\|_2^2 / \sum_{s'\in\mathcal{S}} \|\mathbf{c}_{s'}\|_2^2$. Under an orthonormal DWT with energy-preserving boundary handling, Parseval's identity makes r_s the fraction of the input-window energy assigned to stream s (Mallat, 1999). We predict this distribution from the final SST representations $\mathbf{Z}_0^{(R)} = [\mathbf{z}_{s,0}^{(R)}]_{s\in\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|\times D}$ as $\hat{\mathbf{r}} = \mathrm{Softmax}\big(g_{\text{SED}}(\mathbf{Z}_0^{(R)})\big)$, and optimize $\mathcal{L}_{\text{SED}} = D_{\text{KL}}(\mathbf{r} \,\|\, \hat{\mathbf{r}})$. This objective complements normalized coefficient modeling by encouraging the SSTs to encode window-specific scale importance. As a result, TSComp learns not only structural dependencies across scales but also the relative contribution of each multi-scale component to the observed signal.

Time-domain Signal Reconstruction. Given the masked streams $\tilde{\mathcal{C}}_\Omega$ from MDM, TSComp reconstructs normalized coefficient patches with scale-specific heads g_{REC}^s. The reconstructed patches are unraveled, rescaled by iSCNorm, and recomposed through iDWT to get $\hat{\mathbf{x}} = \mathcal{W}_L^{-1}(\hat{\mathcal{C}})$. We define the time-domain reconstruction loss as $\mathcal{L}_{\text{TIME}} = \|\hat{\mathbf{x}} - \mathbf{x}\|_2^2$, which encourages cross-scale composition by supervising whether scale-specific components recover the original signal. However, under an orthonormal DWT, $\mathcal{L}_{\text{TIME}}$ is equivalent to an energy-weighted sum of relative scale-wise reconstruction errors, allowing low-energy scales to contribute weakly even when poorly reconstructed (Theorem 3). Thus, we introduce scale-separated supervision in the wavelet domain.

Scale-aware Coefficient Reconstruction. Let $\mathbf{p}_{s,i}$ denote the normalized target patch from stream s at index i, and let $\hat{\mathbf{p}}_{s,i}$ denote its reconstruction from the stream-specific head g_{REC}^s. We define the coefficient reconstruction loss as $\mathcal{L}_{\text{COEF}} = \sum_{s\in\mathcal{S}} \hat{r}_s \sum_{i=1}^{N} m_{s,i} \|\hat{\mathbf{p}}_{s,i} - \mathbf{p}_{s,i}\|_2^2$, where $m_{s,i} = 1$ if patch (s, i) is masked and $m_{s,i} = 0$ otherwise. Weighting each stream by the predicted scale energy $\hat{r}_s$ aligns patch-level reconstruction with window-specific scale importance, encouraging the model to allocate greater emphasis on streams that contribute more strongly to the input composition. The overall pre-training objective is $\mathcal{L} = \mathcal{L}_{\text{SED}} + \mathcal{L}_{\text{TIME}} + \mathcal{L}_{\text{COEF}}$.

4.4 Fine-tuning on Downstream Tasks

For downstream transfer, TSComp introduces task-specific, scale-aware pipelines that adapt stream-specific representations from the pre-trained encoder rather than aggregating all scales into a single embedding. **(a) Coarse-to-detail Forecasting:** for forecasting, TSComp predicts future coefficients at each scale through scale-specific heads and reconstructs the forecast by iSCNorm and iDWT. **(b) Locality-aware Multi-scale Imputation:** for imputation, TSComp reconstructs masked coefficient streams through the pre-trained reconstruction heads before recomposition in the time domain. **(c) Scale-discriminative Classification:** TSComp computes stream-wise logits with the scale-specific heads and aggregates them using the predicted scale energy distribution. **(d) Scale-residual Anomaly Scoring:** TSComp computes scale-specific reconstruction residuals through the pre-trained reconstruction heads and converts them back to the original time axis. These task-specific adaptations are aligned with the pre-training architecture and effectively leverage scale-specific information for each task and dataset. Details are provided in Appendix C.4.1.

5 Experiments

We evaluate TSComp on four representative time series tasks: long-term forecasting, imputation, classification, and anomaly detection. We pre-train two variants, TSComp-small and TSComp-base,

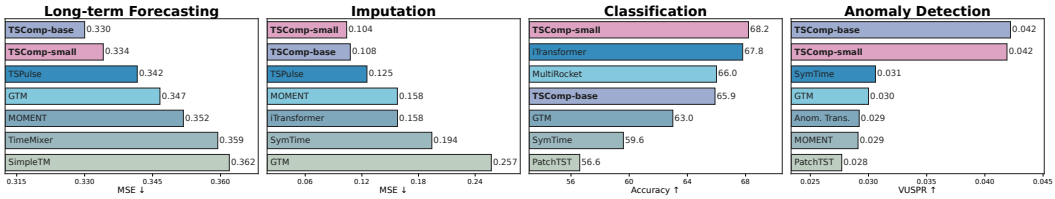


Figure 2: Overall downstream performance across four tasks, showing the top seven models for each task ranked by average performance across benchmark datasets.

Table 2: Results across four downstream tasks, averaged over three random seeds. Long-term forecasting results are further averaged over $\{96, 192, 336, 720\}$ horizons, and imputation results over four mask ratios. The best and second-best results are highlighted in bold and underlined, respectively. Full evaluation results are provided in Appendix H.

(a) Long-term Forecasting

Dataset	Proposed				Multi-task TSFMs								Wavelet-based				Frequency-based				Multi-scale			
	TSComp-small		TSComp-base		GTM		TSPulse		SymTime		MOMENT		SimpleTM		AdaWaveNet		FreTS		FEDformer		TimeMixer		TimesNet	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	0.426	0.441	**0.421**	**0.439**	0.452	0.462	0.448	0.457	0.548	0.515	0.447	0.455	0.492	0.483	0.474	0.475	0.463	0.469	0.461	0.473	0.506	0.493	0.502	0.493
ETTh2	0.360	0.397	**0.348**	**0.388**	0.383	0.415	0.360	0.403	0.422	0.442	0.361	0.405	0.374	0.413	0.402	0.430	0.529	0.497	0.410	0.460	0.388	0.425	0.441	0.459
ETTm1	0.361	0.388	0.357	0.385	**0.356**	0.387	0.362	**0.384**	0.440	0.443	0.369	0.392	0.370	0.396	0.385	0.402	0.375	0.396	0.423	0.449	0.376	0.399	0.392	0.413
ETTm2	0.263	0.320	0.260	**0.316**	0.274	0.329	**0.256**	0.318	0.277	0.337	0.263	0.324	0.264	0.325	0.274	0.330	0.291	0.338	0.332	0.377	0.269	0.328	0.301	0.349
Exchange	0.449	0.450	**0.440**	**0.447**	0.499	0.493	0.495	0.478	0.872	0.634	0.502	0.483	0.503	0.498	0.585	0.526	0.744	0.594	0.947	0.741	0.461	**0.447**	0.746	0.629
Weather	0.234	0.271	0.232	0.270	0.239	0.277	0.229	**0.267**	0.246	0.283	0.238	0.278	0.237	0.278	0.242	0.279	**0.226**	0.276	0.320	0.366	0.232	0.275	0.253	0.290
Electricity	0.166	0.259	0.168	0.265	0.166	**0.258**	**0.162**	0.260	0.176	0.274	0.183	0.283	0.187	0.287	0.169	0.267	0.177	0.279	0.234	0.341	0.187	0.289	0.212	0.315
Traffic	0.414	0.284	0.411	0.285	**0.404**	**0.275**	0.422	0.305	0.405	0.292	0.451	0.325	0.468	0.338	0.412	0.295	0.472	0.339	0.642	0.387	0.456	0.329	0.642	0.348
Top-2 Count	9		11		5		7		1		0		0		0		1		0		1		0	

(b) Imputation

Dataset	Proposed				Multi-task TSFMs								Task-specific Baselines			
	TSComp-small		TSComp-base		GTM		TSPulse		SymTime		MOMENT		PatchTST		iTransformer	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	**0.181**	**0.273**	0.196	0.285	0.475	0.466	0.279	0.349	0.321	0.392	0.279	0.343	0.671	0.548	0.271	0.350
ETTh2	**0.092**	**0.192**	0.100	0.200	0.241	0.342	0.115	0.229	0.270	0.359	0.112	0.218	0.269	0.354	0.177	0.287
ETTm1	**0.085**	**0.183**	0.092	0.192	0.130	0.228	0.099	0.201	0.145	0.260	0.117	0.206	0.662	0.532	0.130	0.242
ETTm2	**0.054**	**0.140**	0.057	0.144	0.173	0.279	0.057	0.148	0.136	0.244	0.059	0.142	0.168	0.279	0.091	0.199
Exchange	**0.013**	**0.071**	**0.013**	0.072	0.068	0.188	0.019	0.092	0.157	0.277	0.027	0.111	0.127	0.254	0.090	0.197
Weather	0.061	0.085	0.064	0.093	0.079	0.117	**0.058**	0.081	0.066	0.112	0.059	**0.073**	0.191	0.260	0.071	0.117
Electricity	0.072	0.177	0.074	0.181	0.132	0.244	**0.067**	**0.166**	0.106	0.225	0.127	0.241	0.785	0.720	0.087	0.194
Traffic	0.273	0.217	**0.266**	**0.210**	0.759	0.409	0.309	0.231	0.352	0.285	0.481	0.333	1.332	0.793	0.345	0.270
Top-2 Count	14		11		0		6		0		3		0		0	

(c) Classification

Model	Avg. Acc.
Proposed	
TSComp-small	**68.2**
TSComp-base	65.9
Multi-task TSFMs	
GTM	63.0
TSPulse	46.2
SymTime	59.6
MOMENT	52.5
Task-specific	
PatchTST	56.6
iTransformer	67.8
MultiRocket	66.0

(d) Anomaly Detection

Metric	Proposed		Multi-task TSFMs				Task-specific Baselines			
	TSComp-small	TSComp-base	GTM	TSPulse	SymTime	MOMENT	PatchTST	TSLANet	Anom. Trans.	Nonstat. Trans.
rAUCROC	0.654	0.641	**0.679**	0.615	0.646	0.632	0.628	0.616	0.629	0.629
rAUCPR	**0.042**	**0.042**	0.029	0.025	0.030	0.029	0.027	0.020	0.029	0.025
VUSROC	0.644	0.632	**0.673**	0.609	0.640	0.628	0.622	0.610	0.622	0.620
VUSPR	**0.042**	**0.042**	0.030	0.025	0.031	0.029	0.028	0.020	0.029	0.025

Anom. Trans. denotes Anomaly Transformer; Nonstat. Trans. denotes Non-stationary Transformer.

where both variants are pre-trained on the same 27.43 GB corpus with 4.12B observations. We compare TSComp with recent multi-task TSFMs, including GTM (He et al., 2026), TSPulse (Ekambaram et al., 2026), SymTime (Wang et al., 2025), and MOMENT (Goswami et al., 2024) as well as strong task-specific supervised baselines selected for each downstream task. Details of the pre-training setup and downstream evaluation are provided in Appendix C. Figure 2 summarizes the overall results, showing that TSComp consistently ranks among the top-performing models across all four tasks. We further evaluate TSComp on a real-world industrial regression task, demonstrating practical generalization to industrial settings (Appendix D).

Long-term Forecasting. As shown in Table 2(a), TSComp achieves strong overall performance across eight benchmark datasets (Zhou et al., 2021), with TSComp-base achieving the highest Top-2 count. Unlike multi-task TSFMs that adapt shared representations, TSComp preserves scale-specific representations during forecasting by predicting future wavelet coefficients and explicitly recomposing them in the time domain. This differs from frequency-domain forecasters that emphasize global spectral structure and multi-scale forecasters such as TimeMixer that combine predictions across

downsampled resolutions. These results are consistent with the benefit of preserving scale-specific structure and cross-scale relations throughout forecasting.

Imputation. For imputation, we follow the patch-masking protocol in prior TSFM evaluation (Goswami et al., 2024), where contiguous temporal patches are masked. As shown in Table 2(b), TSComp achieves superior performance across benchmark datasets and outperforms TSFM baselines on most datasets. Since patch-level missing regions can affect both local variations and broader temporal structures, recovery benefits from complementary information across wavelet scales. As missing patches may disrupt structures at multiple temporal resolutions, TSComp reconstructs scale-specific coefficients using cross-scale context before recomposing them into the time domain.

Classification. We evaluate TSComp on 27 UEA benchmark datasets (Bagnall et al., 2018), as shown in Table 2(c). TSComp-small achieves the best mean accuracy while TSComp-base achieves modest performance. Compared with multi-task TSFM baselines, TSComp obtains higher average accuracy, demonstrating that the representations learned by TSComp transfer effectively to classification. Compared with task-specific baselines, TSComp-small remains highly competitive, despite not being pre-trained specifically for classification.

Anomaly Detection. We evaluate TSComp on 245 UCR Anomaly Archive datasets (Wu & Keogh, 2023) using four range-aware metrics: rAUCROC, rAUCPR, VUSROC, and VUSPR. As shown in Table 2(d), TSComp performs on par with or better than state-of-the-art TSFM baselines across all four metrics, with particularly competitive results on rAUCPR and VUSPR. While GTM obtains the best rAUCROC and VUSROC, both TSComp variants remain among the top-performing methods overall. This pattern is meaningful because anomaly detection often depends on localized deviations that may be weak in aggregate time-domain errors but clearer at specific temporal resolutions. By measuring residuals across scale-specific streams, TSComp can capture anomalous evidence from both coarse structural changes and fine-grained irregularities. The strong precision-recall-oriented results suggest that these residuals provide useful signals for identifying anomalous regions rather than only separating normal and abnormal windows globally.

6 Ablation Studies

We analyze three aspects of TSComp: (a) the benefit of explicitly modeling relations among aligned scales, (b) the effect of scale-aware downstream pipelines, and (c) the effects of its pre-training objectives across downstream tasks.

Effect of Cross-scale Relation Modeling. Table 3(a) examines two distinct components of SCA. The cross-scale embedding exchanges information among temporally aligned resolutions, while coefficient attention models within-scale temporal dependencies conditioned on the resulting cross-scale context. Removing the cross-scale embedding degrades performance across all four downstream tasks, indicating the broad utility of information exchange across resolutions. Removing coefficient attention also consistently reduces performance across all tasks, although the magnitude of degradation varies by task, with larger effects on Weather imputation and classification and smaller effects on forecasting and anomaly detection. The full model performs best across all four tasks, highlighting the importance of both cross-scale and within-scale temporal modeling.

Effect of Scale-aware Downstream Adaptation. Table 3(b) evaluates the proposed scale-aware downstream pipelines against a w/o Scale variant that averages representations across scales and applies a single downstream head. TSComp consistently improves performance across imputation, forecasting, anomaly detection, and classification, although the gain varies by task. These results indicate that preserving scale-specific information throughout downstream adaptation is more effective than collapsing the scale representations. Further evidence is provided in Appendix F.3.

Effect of Pre-training Objectives. Table 3(c) shows that the objectives impose distinct pre-training biases, not uniformly improving each downstream task. $\mathcal{L}_{\mathrm{COEF}}$ alone achieves the lowest MSE in imputation, whereas forecasting performs best when $\mathcal{L}_{\mathrm{COEF}}$ is removed. Removing $\mathcal{L}_{\mathrm{SED}}$ yields small improvements on both ETTm1 tasks, indicating that explicitly calibrating the scale energy distribution does not directly translate into lower task-specific reconstruction error in this setting. Unlike the reconstruction losses, $\mathcal{L}_{\mathrm{SED}}$ does not directly supervise coefficient values; instead, it preserves the relative contribution of each scale and indirectly influences coefficient reconstruction

Table 3: Ablation studies of TSComp on (a) cross-scale relation modeling, (b) scale-aware downstream adaptation, and (c) pre-training objectives.

(a) Cross-scale Relation Modeling

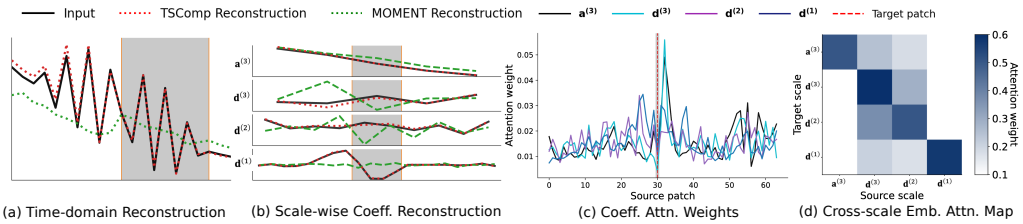
Variant	Weather Imp. MSE↓	ETTh1 Forecast MSE↓	Anomaly VUSPR↑	Classification Avg. Acc.↑
w/o Cross-scale Emb.	0.0643	0.4389	0.0297	67.62
w/o Coeff. Attn.	0.1503	0.4274	0.0418	66.64
TSComp	**0.0605**	**0.4257**	**0.0421**	**68.18**

(b) Scale-aware Downstream Adaptation

Variant	ETTh1 Imp. MSE↓	ETTh1 Forecast MSE↓	Anomaly VUSPR↑	Classification Avg. Acc.↑
w/o Scale	0.1885	0.4288	0.0343	67.00
TSComp	**0.1813**	**0.4257**	**0.0421**	**68.18**

(c) Pre-training Objectives

	ETTm1	
Variant	Imp. MSE↓	Forecast MSE↓
$\mathcal{L}_{\text{COEF}}$	**0.0824**	0.3783
$\mathcal{L}_{\text{TIME}}$	0.0992	0.3939
w/o $\mathcal{L}_{\text{TIME}}$	0.0909	0.3562
w/o $\mathcal{L}_{\text{COEF}}$	0.0887	**0.3556**
w/o $\mathcal{L}_{\text{SED}}$	0.0843	0.3575
TSComp	0.0849	0.3609



(a) Time-domain Reconstruction (b) Scale-wise Coeff. Reconstruction (c) Coeff. Attn. Weights (d) Cross-scale Emb. Attn. Map

Figure 3: Qualitative analysis. (a) Time-domain reconstruction in the selected region, comparing TSComp and MOMENT. (b) Scale-wise coefficient reconstruction. (c) Coefficient attention weights of the target patch. (d) Cross-scale embedding attention map of the selected target patch.

through scale-aware weighting. The full combination provides the most balanced trade-off across the evaluated tasks, even though individual removal can improve specific settings.

7 Qualitative Analysis

We qualitatively examine whether TSComp learns scale-specific structures and the cross-scale interactions through which they compose the observed signal. We generate synthetic time series and examine a selected target patch with its aligned event window, marked by the gray shaded region (see Appendix F.2). As shown in Figure 3(a–b), TSComp closely reconstructs both the time-domain signal and its scale-specific structures, whereas MOMENT smooths local variations and exhibits larger deviations in the detail scales. Figure 3(c) shows that attention patterns over source patches differ across scales, consistent with scale-specific temporal dependencies. Figure 3(d) further shows that each target scale assigns different attention weights to the aligned source scales, revealing scale-dependent cross-scale interactions. These analyses suggest how TSComp captures scale-specific structures while modeling cross-scale relations required for their coherent composition.

8 Conclusion

In this paper, we propose TSComp, a cross-scale compositional pre-training framework for multi-task TSFMs. Motivated by the limitations of conventional time-domain masked reconstruction in explicitly capturing scale-specific structures and their cross-scale correspondence, TSComp learns temporal structures at different resolutions together with the relations through which they compose the observed signal. Its scale-aware design separates temporal structures by scale, preserves their correspondence, and explicitly models both within-scale dynamics and cross-scale interactions during pre-training. Experiments across four downstream tasks demonstrate strong and consistent performance. Ablation and qualitative analyses further support the benefits of cross-scale modeling, scale-aware adaptation, and explicit multi-scale representation learning. These results establish cross-scale compositional learning as a promising direction for transferable time series foundation models. Future work will explore adaptive decomposition and task-adaptive cross-scale modeling to further improve transfer across diverse time series settings.

REFERENCES

Alexander Alexandrov, Konstantinos Benidis, Michael Bohlke-Schneider, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Danielle C Maddix, Syama Rangapuram, David Salinas, Jasper Schulz, et al. Gluonts: Probabilistic and neural time series modeling in python. *Journal of Machine Learning Research*, 21(116):1–6, 2020.

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

Christoph Bergmeir, Quang Bui, Frits de Nijs, and Peter Stuckey. Residential power and battery data, August 2023. URL `https://doi.org/10.5281/zenodo.8219786`.

Hui Chen, Viet Luong, Lopamudra Mukherjee, and Vikas Singh. Simpletm: A simple baseline for multivariate time series forecasting. In *International Conference on Learning Representations*, volume 2025, pp. 15431–15456, 2025.

Hojin Cho and Chiehyeon Lim. Multistage net: Learning continuous multistage manufacturing processes of liquid products without intermediate output and lead-time labels. *IEEE Transactions on Industrial Informatics*, 20(12):13765–13775, 2024. doi: 10.1109/TII.2024.3432134.

Ingrid Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *The Ninth International Conference on Learning Representations*, 2021.

Vijay Ekambaram, Subodh Kumar, Arindam Jati, Sumanta Mukherjee, Tomoya Sakai, Pankaj Dayama, Wesley Gifford, and Jayant Kalagnanam. Tspulse: Tiny pre-trained models with disentangled representations for rapid time-series analysis. In *International Conference on Learning Representations*, volume 2026, pp. 62090–62133, 2026.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, and Xiaoli Li. Tslanet: Rethinking transformers for time series representation learning. In *International Conference on Machine Learning*, pp. 12409–12428. PMLR, 2024.

Patrick Emami, Abhijeet Sahu, and Peter Graf. Buildingsbench: A large-scale dataset of 900k buildings and benchmark for short-term load forecasting. *Advances in Neural Information Processing Systems*, 36:19823–19857, 2023.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. In *Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: a family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 16115–16152, 2024.

Cheng He, Xu Huang, Gangwei Jiang, Zhaoyi Li, Defu Lian, Hong Xie, Enhong Chen, Patrick PC Lee, et al. Gtm: A general time-series model for enhanced representation learning of time-series data. In *International Conference on Learning Representations*, volume 2026, pp. 7280–7309, 2026.

Jiawei Jiang, Chengkai Han, Wenjun Jiang, Wayne Xin Zhao, and Jingyuan Wang. Libcity: A unified library towards efficient and comprehensive urban spatial-temporal prediction. *arXiv preprint arXiv:2304.14343*, 2023.

Harshavardhan Kamarthi and B. Aditya Prakash. Large pre-trained time series models for cross-domain time series analysis tasks. In *Advances in Neural Information Processing Systems*, volume 37, pp. 56190–56214, 2024.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International conference on learning representations*, 2022.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. *Advances in neural information processing systems*, 35: 9881–9893, 2022.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, volume 2024, 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, 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. In *International Conference on Machine Learning*, pp. 39295–39317. PMLR, 2025.

S.G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989. doi: 10.1109/34.192463.

Stephane Mallat. *A wavelet tour of signal processing*. Academic Press, 1999.

Soukayna Mouatadid, Paulo Orenstein, Genevieve Flaspohler, Miruna Oprescu, Judah Cohen, Franklyn Wang, Sean Knight, Maria Geogdzhayeva, Sam Levang, Ernest Fraenkel, et al. Subseasonalclimateusa: A dataset for subseasonal forecasting and benchmarking. *Advances in Neural Information Processing Systems*, 36:7960–7992, 2023.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023.

Hyunwoo Seo and Chiehyeon Lim. St-mtm: Masked time series modeling with seasonal-trend decomposition for time series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 1209–1220, 2025.

Qichao Shentu, Beibu Li, Kai Zhao, Yang Shu, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. Towards a general time series anomaly detector with adaptive bottlenecks and dual adversarial decoders. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=aKcd7ImG5e.

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

Chang Wei Tan, Angus Dempster, Christoph Bergmeir, and Geoffrey I Webb. Multirocket: multiple pooling operators and transformations for fast and effective time series classification. *Data Mining and Knowledge Discovery*, 36(5):1623–1646, 2022.

Martin Vetterli and Jelena Kovacevic. *Wavelets and subband coding*. Prentice Hall PTR Englewood Cliffs, NJ, 1995.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y Zhang, and JUN ZHOU. Timemixer: Decomposable multiscale mixing for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024.

Wenxuan Wang, Kai Wu, Yujian Betterest Li, Dan Wang, and Xiaoyu Zhang. Synthetic series-symbol data generation for time series foundation models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Zhixian Wang, Qingsong Wen, Chaoli Zhang, Liang Sun, Leandro Von Krannichfeldt, Shirui Pan, and Yi Wang. Benchmarks and custom package for energy forecasting. *arXiv preprint arXiv:2307.07191*, 2023.

Gerald Woo, Chenghao Liu, Akshat Kumar, and Doyen Sahoo. Pushing the limits of pre-training for time series forecasting in the cloudops domain. *arXiv preprint arXiv:2310.05063*, 2023.

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, 2024.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

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

Renjie Wu and Eamonn J Keogh. Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. *IEEE transactions on knowledge and data engineering*, 35(3): 2421–2429, 2023.

Shifeng Xie, Vasilii Feofanov, Jianfeng Zhang, Themis Palpanas, and Ievgen Redko. Cauker: Classification time series foundation models can be pretrained on synthetic data. In *International Conference on Learning Representations*, 2026.

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

Kun Yi, Qi Zhang, Wei Fan, Shoujin Wang, Pengyang Wang, Hui He, Ning An, Defu Lian, Longbing Cao, and Zhendong Niu. Frequency-domain mlps are more effective learners in time series forecasting. *Advances in neural information processing systems*, 36:76656–76679, 2023.

Han Yu, Peikun Guo, and Akane Sano. Adawavenet: Adaptive wavelet network for time series analysis. *Transactions on machine learning research*, 2024:2798, 2024.

Haoran Zhang, Yong Liu, Yunzhong Qiu, Haixuan Liu, Zhongyi Pei, Jianmin Wang, and Mingsheng Long. Timesbert: A bert-style foundation model for time series understanding. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10975–10983, 2025.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11106–11115, 2021.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*, pp. 27268–27286. PMLR, 2022.

APPENDIX CONTENTS

A Structural Requirements for a Scale-decomposed Space **14**

B Wavelet Transform **15**

C Experiment Settings **16**

C.1 Pre-training Datasets 16

C.2 Pre-training Settings 17

C.3 Fine-tuning Settings and Baseline Implementations 20

C.4 Task-specific Baseline Models and Implementations 22

D Real-world Industrial Time Series Regression **25**

E Additional Ablation Studies **26**

E.1 Effect of Decomposition Strategy 26

E.2 Effect of Decomposition Level 27

E.3 Effect of Wavelet Basis 27

E.4 Effect of Stream Summary Tokens 28

E.5 Effect of Multi-scale Directional Masking 28

E.6 Effect of Scale-wise Coefficient Normalization 29

F Qualitative Analysis **29**

F.1 TSComp Reconstruction Visualization 29

F.2 Qualitative Analysis of Multi-scale Composition 29

F.3 Case Studies on Downstream Tasks 31

G Theoretical Properties **33**

G.1 Boundary Effects and the Motivation for Coefficient-Domain Masking 34

G.2 Stability of the Scale Energy Distribution under Coefficient Perturbation 37

G.3 Population-level Targets of the Pre-training Objectives 37

G.4 Energy-weighted Decomposition of Time-domain Reconstruction 39

H Full Results on Downstream Tasks **40**

H.1 Full Results on Long-term Forecasting 40

H.2 Full Results on Imputation 40

H.3 Full Results on Anomaly Detection 40

H.4 Full Results on Classification 40

I Limitations **41**

J AI Use Statement **41**

K Ethics Statement and Reproducibility Statement **41**

A STRUCTURAL REQUIREMENTS FOR A SCALE-DECOMPOSED SPACE

Table 4: Comparison of structural properties among different decomposition strategies

Decomposition	Temporal localization	Scale hierarchy	Perfect recomposition	Energy preservation
Fourier transform	✗	✗	✓	✓
Average pooling (Wang et al., 2024)	✓	✓	✗	✗
Residual pyramid	✓	✓	✓	✗
Wavelet transform (biorthogonal basis)	✓	✓	✓	✗
Wavelet transform (orthonormal basis)	✓	✓	✓	✓

To construct representations that explicitly capture temporal structures at different scales, and to make scale-specific recovery and cross-scale composition explicit pre-training targets, TSComp requires an appropriate decomposition space. We define four structural requirements for such a decomposition: temporal localization, an explicit scale hierarchy, perfect recomposition, and energy preservation. Temporal localization requires each decomposed component to retain information about where a temporal pattern occurs, while an explicit scale hierarchy organizes the components according to distinct temporal resolutions, from coarse structures to progressively finer variations. Perfect recomposition requires that the original signal can be exactly recovered from the complete set of decomposed components. Energy preservation requires that the energy of the original signal equals the sum of the energies of its decomposed components.

These properties directly support the cross-scale compositional pre-training design of TSComp. Temporal localization and the scale hierarchy allow components at different resolutions to be aligned to the same temporal regions, enabling time-aligned patching, cross-scale embedding, and multi-scale directional masking. Perfect recomposition allows the reconstructed scale-specific components to be recombined exactly into the time-domain signal, providing the basis for time-domain signal reconstruction. Energy preservation gives each stream's energy ratio a direct interpretation as the fraction of input-window energy carried by that scale, thereby quantifying its relative contribution to the observed signal composition. This makes the scale energy distribution a meaningful calibration target, while the predicted energy ratios are subsequently used to weight scale-aware coefficient reconstruction, assigning greater reconstruction emphasis to scales predicted to contribute more strongly to the input composition. These requirements define a decomposition space in which scale-specific structures can be modeled separately while their temporal correspondence and composition remain explicitly accessible to the pre-training objectives.

Table 4 compares representative transformation and decomposition strategies with respect to the four structural requirements. The Fourier transform represents a signal using sinusoidal components at different frequencies, each spanning the entire input window. Its components capture global frequency patterns rather than temporally localized structures, and although they are ordered by frequency, they do not form an explicit multi-resolution scale hierarchy. Nevertheless, the Fourier transform supports perfect reconstruction through its invertible basis, while Parseval's identity guarantees energy preservation across the frequency coefficients (Mallat, 1999). Average-pooling-based downsampling constructs a multi-scale hierarchy by repeatedly averaging neighboring observations to produce progressively coarser temporal resolutions (Wang et al., 2024). While the resulting representations retain temporal localization and form an explicit hierarchy across resolutions, average pooling is a many-to-one operation that discards fine-scale information and therefore prevents perfect recomposition. A residual pyramid addresses this limitation by preserving the discarded information as residuals between each input and its upsampled coarse representation, enabling exact recomposition from the final coarse component and the residual hierarchy. However, these components do not generally form an orthogonal decomposition, so their squared norms do not provide an additive partition of the input-window energy and energy preservation is not guaranteed.

Among the candidate strategies, the wavelet transform is structurally well aligned with the four requirements defined above. Its localized multi-resolution representation provides both temporal localization and an explicit scale hierarchy, while a perfect-reconstruction wavelet transform allows the original signal to be exactly recovered from the decomposed components. With an orthonormal basis and energy-preserving boundary handling, Parseval's identity further gives an additive decomposition of the input-signal energy across coefficient streams. We therefore adopt an orthonormal

wavelet transform as the underlying decomposition space for TSComp. Detailed formulation and interpretation of the wavelet representation are provided in Appendix B.

Importantly, our choice of the wavelet transform is based on its structural compatibility with the proposed pre-training framework rather than on an assumption that the wavelet transform is universally optimal. Our decomposition ablations show that the best-performing strategy varies across downstream tasks (Appendix E.1). Accordingly, the contribution of this work is not the use of wavelet transform itself, but the formulation of cross-scale compositional pre-training that explicitly targets scale-specific recovery and cross-scale composition. The wavelet transform serves as the underlying decomposition space that satisfies the structural requirements needed to realize these learning objectives.

B Wavelet Transform

The wavelet transform provides a multi-resolution representation of a signal by decomposing it into one coarse approximation component and a sequence of detail components at progressively finer temporal scales (Mallat, 1989). In our setting, an L-level discrete wavelet transform (DWT) maps an input window $\mathbf{x} \in \mathbb{R}^T$ to an approximation coefficient stream $\mathbf{a}^{(L)}$ and detail coefficient streams $\{\mathbf{d}^{(\ell)}\}_{\ell=1}^{L}$, as illustrated in Figure 4. The approximation stream $\mathbf{a}^{(L)}$ represents the low-frequency trend at the coarsest resolution, while each detail stream $\mathbf{d}^{(\ell)}$ captures scale-specific fluctuations relative to a coarser representation.

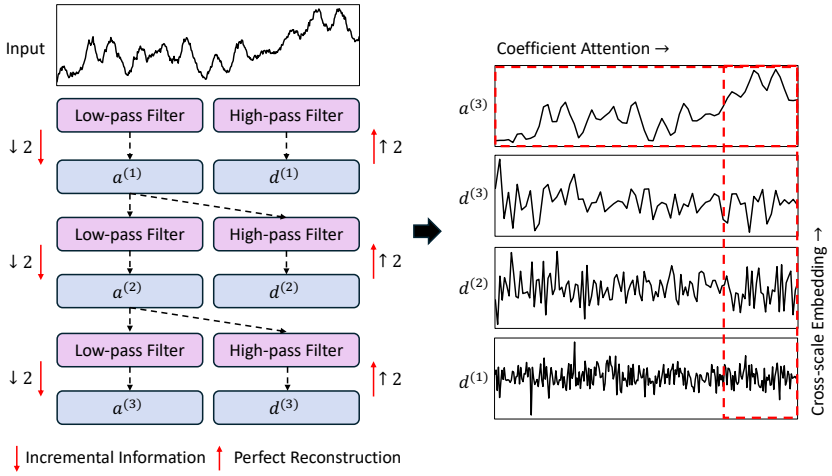


Figure 4: Depiction of wavelet decomposition and scale coefficient attention. Multi-level discrete wavelet transform decomposes the input into one approximation stream and multiple detail streams. Cross-scale embedding exchanges information among resolutions over the same temporal support. Coefficient attention models temporal dependencies within each coefficient stream.

Forward Transform $\mathcal{W}_L$. Let $h[\cdot]$ and $g[\cdot]$ denote the low-pass and high-pass analysis filters associated with a chosen wavelet family. We define the level-0 approximation coefficients as $\mathbf{a}^{(0)} = \mathbf{x}$. For each level $\ell = 1, \ldots, L$, the DWT applies filtering followed by dyadic downsampling to produce the next approximation stream and the corresponding detail stream:

$$a^{(\ell)}[k] = \sum_n h[n-2k]\, a^{(\ell-1)}[n], \qquad d^{(\ell)}[k] = \sum_n g[n-2k]\, a^{(\ell-1)}[n].$$

The resulting coefficient-stream collection is

$$\mathcal{C} = \mathcal{W}_L(\mathbf{x}) = \{\mathbf{a}^{(L)}, \mathbf{d}^{(L)}, \ldots, \mathbf{d}^{(1)}\}.$$

Choice of Wavelet Basis. In our implementation, we use Symlet-4 (`sym4`) as the default wavelet basis. Symlets are compactly supported wavelets designed to improve symmetry relative to Daubechies

wavelets of comparable support, which can reduce phase distortion in discrete-time representations while preserving efficient filter-bank computation. Compared with shorter wavelet bases such as `Haar`, `sym4` provides smoother approximation streams and less abrupt detail responses, while remaining computationally lightweight compared with higher-order wavelet bases. This gives a practical balance between temporal localization, frequency selectivity, and numerical efficiency.

Incremental Information across Levels. The multi-level construction of $\mathcal{W}_L$ can be interpreted as an incremental resolution process. At each level ℓ, the approximation stream $\mathbf{a}^{(\ell)}$ provides a coarser representation of $\mathbf{a}^{(\ell-1)}$, while the detail stream $\mathbf{d}^{(\ell)}$ represents the complementary high-pass subband required, together with $\mathbf{a}^{(\ell)}$, to reconstruct $\mathbf{a}^{(\ell-1)}$. Thus, $\mathbf{d}^{(\ell)}$ should not be viewed as arbitrary auxiliary information, but as the scale-specific residual information that is separated from the lower-resolution approximation. Collecting detail streams across levels therefore yields a hierarchy of residual components, from coarse to fine, that jointly describe temporal variation at multiple resolutions.

Coefficient Lengths and Scale Interpretation. Each decomposition level applies dyadic downsampling, so the native temporal resolution decreases as ℓ increases. Ignoring boundary-extension effects, the coefficient lengths scale approximately as

$$|\mathbf{a}^{(\ell)}| \approx \left\lceil \frac{T}{2^\ell} \right\rceil, \qquad |\mathbf{d}^{(\ell)}| \approx \left\lceil \frac{T}{2^\ell} \right\rceil.$$

In practical finite-window implementations, the exact coefficient-stream lengths depend on the extension mode and wavelet filter length. Up to these finite-window effects, dyadic downsampling gives a level-ℓ coefficient stream a nominal native resolution of approximately $T/2^\ell$. Thus, $\mathbf{a}^{(L)}$ represents the coarsest approximation, whereas $\mathbf{d}^{(1)}$ captures the finest-scale detail variations. This scale-dependent resolution motivates the patching rule used in TSComp. Let p denote the patch length in the original time-domain. For a detail stream at level ℓ, we set $p_\ell = p/2^\ell$, and for the approximation stream $\mathbf{a}^{(L)}$, we set $p_0 = p/2^L$. With p divisible by 2^L, each stream is mapped to the same number of patches, $N = T/p$, so the i-th patch remains temporally aligned across scales. This construction preserves cross-scale correspondence while allowing each stream to retain its scale-specific resolution and semantics.

Inverse Transform $\mathcal{W}_L^{-1}$ and Perfect Reconstruction. Let $\tilde{h}[\cdot]$ and $\tilde{g}[\cdot]$ denote the low-pass and high-pass synthesis filters. For $\ell = L, \ldots, 1$, the inverse recursion reconstructs $\mathbf{a}^{(\ell-1)}$ from $\mathbf{a}^{(\ell)}$ and $\mathbf{d}^{(\ell)}$ by dyadic upsampling and filtering:

$$a^{(\ell-1)}[n] = \sum_k \tilde{h}[n-2k]\; a^{(\ell)}[k] + \sum_k \tilde{g}[n-2k]\; d^{(\ell)}[k].$$

Applying this recursion from $\ell = L$ down to $\ell = 1$ yields the inverse mapping $\mathbf{x} = \mathcal{W}_L^{-1}(\mathcal{C})$. If $(h, g, \tilde{h}, \tilde{g})$ form a perfect reconstruction filter bank and the same extension rule is used in analysis and synthesis, then $\mathcal{W}_L^{-1}\big(\mathcal{W}_L(\mathbf{x})\big) = \mathbf{x}$ holds up to numerical precision.

Boundary Handling. Because filtering requires samples beyond the finite window, practical DWT implementations adopt a boundary extension rule (periodic or symmetric extension). In TSComp, $\mathcal{W}_L$ and $\mathcal{W}_L^{-1}$ use the same boundary handling so that reconstruction remains stable and coefficient streams align consistently across windows.

C Experiment Settings

C.1 Pre-training Datasets

TSComp is pre-trained on a diverse collection of public time series datasets spanning multiple domains, temporal frequencies, and task sources. The pre-training corpus includes selected subsets from widely used repositories and previously curated large-scale collections, including Monash (Godahewa et al., 2021), UTSD (Liu et al., 2024b), LOTSA (Woo et al., 2024), and Time Series Pile (Goswami et al., 2024), as well as datasets from CloudOps (Woo et al., 2023), BuildingsBench (Emami et al.,

2023), GluonTS (Alexandrov et al., 2020), and LibCity (Jiang et al., 2023). The pre-training corpus includes the Monash Electricity and Traffic datasets at hourly and weekly resolutions, which are based on the same underlying series as the corresponding downstream Electricity and Traffic benchmarks. Results on these datasets therefore include pre-training exposure to the same underlying series. Apart from these cases, all other downstream benchmark datasets are excluded from the pre-training corpus. Following prior work, we convert multivariate series into channel-wise univariate sequences under a channel-independent setting (Shi et al., 2025; Xie et al., 2026). As summarized in Tables 5 and 6, the resulting pre-training corpus contains 27.43 GB of data, 873,441 time series, and 4.12B observations in total. Tables 7–14 provide dataset-level statistics, including dataset name, sampling frequency, number of series, and number of observations, where frequency denotes the temporal resolution of each series.

Table 5: Summary statistics of the full pre-training corpus used in TSComp.

Total Size (GB)	Total # Time Series	Total # Observations (B)
27.43	873,441	4.12

Table 6: Key statistics of the TSComp pre-training corpus by domain.

	Energy	**Transport**	**Climate**	**CloudOps**	**Web**	**Sales**	**Nature**	**Econ/Fin**	**Healthcare**	**Other**	**Total**
# Datasets	28	20	5	3	2	3	2	5	3	6	77
# Series	8,273	121,452	37,076	361,267	192,738	148,868	2	419	1,034	2,312	873,441
# Obs.	1.197B	464.032M	135.776M	2.151B	157.105M	19.687M	97.665K	260.712K	128.125K	235.038K	4.126B
%	29.024	11.247	3.291	52.136	3.808	0.477	0.002	0.006	0.003	0.006	100.000

Table 7: Datasets and key properties from CloudOps (Woo et al., 2023).

Dataset	Frequency	# Time Series	# Obs.
Alibaba Cluster Trace 2018	5T	58409	190,385,060
Azure VM Traces 2017	5T	159472	885,522,908
Borg Cluster Data 2011	5T	143386	1,075,105,708

C.2 Pre-training Settings

We summarize the architectural and pre-training configurations of TSComp in Tables 15 and 16, respectively. We pre-train TSComp with three objectives: scale energy distribution calibration $\mathcal{L}_{\mathrm{SED}}$, time-domain signal reconstruction $\mathcal{L}_{\mathrm{TIME}}$, and scale-aware coefficient reconstruction $\mathcal{L}_{\mathrm{COEF}}$. The input time series length is set to 512, with a patch length of 8 and a patch stride of 8. We apply Reversible Instance Normalization (RevIN) (Kim et al., 2022) in the time domain and further enable Scale-wise Coefficient Normalization (SCNorm) to reduce scale-wise magnitude imbalance across decomposition levels. For the Scale-balanced Tokenization, TSComp uses a Symlet-4 wavelet basis with a decomposition level of $L = 3$ and periodization boundary extension. For masking, we adopt Multi-scale Directional Masking (MDM). We set the temporal selection ratio to $p_{\mathrm{temp}} = 0.8$, which selects a subset of temporally aligned patch groups before applying directional masking. Since the reference stream and masking direction are sampled uniformly, the resulting expected coefficient masking ratio is $p_{\mathrm{temp}}(|\mathcal{S}| + 1)/(2|\mathcal{S}|) = 0.5$ with $|\mathcal{S}| = 4$. The masking direction is randomly sampled for each sample, following either the coarse-to-fine or fine-to-coarse masking direction. We optimize TSComp using AdamW with a linear warmup cosine learning-rate schedule. The initial learning rate is 1×10^{-4}, the minimum learning rate is 1×10^{-5}, and the number of warmup steps is set to 6,000. We use a weight decay of 0.05, gradient clipping with a maximum norm of 5, and bfloat16 mixed-precision training with AMP. TSComp-small is pre-trained with a batch size of 2048 for 2 epochs, while TSComp-base is pre-trained with a batch size of 512 for 5 epochs. Pre-training takes approximately 20 hours for TSComp-small using 2 NVIDIA A100-80GB GPUs and approximately 98 hours for TSComp-base using 4 NVIDIA A100-80GB GPUs.

Table 8: Datasets and key properties from BuildingsBench (Emami et al., 2023).

Dataset	Frequency	# Time Series	# Obs.
BDG-2-Bear	H	91	1,482,312
BDG-2-Fox	H	135	2,324,568
BDG-2-Panther	H	105	919,800
BDG-2-Rat	H	280	4,728,288
Borealis	H	15	83,269
IDEAL	H	217	1,255,253
Sceaux	H	1	34,223
SMART	H	5	95,709

Table 9: Datasets and key properties from GluonTS (Alexandrov et al., 2020).

Dataset	Frequency	# Time Series	# Obs.
M5	D	30490	457,350
Taxi	30T	67984	54,999,056
Uber TLC Daily	D	262	47,087
Uber TLC Hourly	H	262	1,129,444
Wiki Rolling	D	47675	40,619,100

Table 10: Datasets and key properties from LibCity (Jiang et al., 2023).

Dataset	Frequency	# Time Series	# Obs.
Beijing Subway	30T	276	867,744
HZMETRO	15T	80	380,320
Loop Seattle	5T	323	33,953,760
Los-Loop	5T	207	7,094,304
PEMS03	5T	358	9,382,464
PEMS04	5T	307	15,649,632
PEMS07	5T	883	24,921,792
PEMS08	5T	170	9,106,560
PEMS Bay	5T	325	16,941,600
Q-Traffic	15T	45148	264,386,688
SHMetro	15T	288	5,073,984
SZ-Taxi	15T	156	464,256

Table 11: Datasets and key properties from Residential Power and Battery (Bergmeir et al., 2023).

Dataset	Frequency	# Time Series	# Obs.
Residential Load Power	T	271	437,983,677
Residential PV Power	T	233	376,016,850

Table 12: Datasets and key properties from SubseasonalClimateUSA (Mouatadid et al., 2023).

Dataset	Frequency	# Time Series	# Obs.
Subseasonal	D	862	56,788,560
Subseasonal Precipitation	D	862	9,760,426

Table 13: Datasets and key properties from ProEnFo (Wang et al., 2023).

Dataset	Frequency	# Time Series	# Obs.
BDG-2 Bull	H	41	719,304
BDG-2 Cockatoo	H	1	17,544
BDG-2 Hog	H	24	421,056
Covid19 Energy	H	1	31,912
ELF	H	1	21,792
Gfc12	H	20	788,280
Gfc14	H	1	17,520
Gfc17	H	8	140,352
PDF	H	1	17,520

Table 14: Datasets and key properties from Monash Time Series Forecasting Repository (Godahewa et al., 2021).

Dataset	Frequency	# Time Series	# Obs.
Australian Electricity Demand	30T	5	1,155,264
Bitcoin	D	18	75,364
Car Parts	M	2674	136,374
CIF 2016	M	72	7,108
COVID Deaths	D	266	56,392
Dominick	W	115704	19,092,987
Electricity Hourly	H	321	8,443,584
Electricity Weekly	W	321	50,076
FRED MD	M	107	77,896
Hospital	M	767	64,428
Extended Web Traffic	D	145063	116,485,589
KDD Cup 2018	H	270	2,942,364
London Smart Meters	30T	5560	166,528,896
M1 Monthly	M	617	55,998
M1 Quarterly	Q	203	9,944
M1 Yearly	Y	181	4,515
NN5 Daily	D	111	87,801
NN5 Weekly	W	111	12,543
Pedestrian Counts	H	66	3,132,346
Rideshare	H	2304	1,246,464
Saugeen	D	1	23,741
Solar	M	137	7,200,720
Solar Power	4S	1	7,397,222
Solar Weekly	W	137	7,124
Sunspot	D	1	73,924
Temperature Rain	D	32072	23,252,200
Tourism Monthly	M	366	109,280
Tourism Quarterly	Q	427	42,544
Tourism Yearly	Y	518	12,757
Traffic Hourly	H	862	15,122,928
Traffic Weekly	W	862	89,648
US Births	D	1	7,305
Vehicle Trips	D	329	42,382
Weather	D	3010	43,032,000
Wind Power	4S	1	7,397,147
Wind Farms	M	339	172,178,060

Table 15: Architectural configurations of TSComp.

	TSComp-small	**TSComp-base**
Model architecture		
Encoder layers	3	6
Hidden size (d_{model})	256	512
Attention heads	4	8
FFN size (d_{ff})	1024	2048
Dropout	0.1	0.1
Params	12.17M	47.93M
Scale-balanced Tokenization		
Wavelet basis	Symlet-4 (`sym4`)	
Decomposition level (L)	3	
Boundary extension mode	Periodization	

Table 16: Pre-training configurations.

Configuration	**TSComp-small**	**TSComp-base**
Pre-training masking and objectives		
MDM temporal selection ratio p_{temp}	0.8	
Expected coefficient masking ratio	0.5	
Masking direction	Random (coarse-to-fine / fine-to-coarse)	
Optimization		
Pre-training objectives	$\mathcal{L}_{\text{SED}}, \mathcal{L}_{\text{TIME}}, \mathcal{L}_{\text{COEF}}$	
Optimizer	AdamW	
LR scheduler	Linear warmup cosine decay	
Initial LR	1×10^{-4}	
Minimum LR	1×10^{-5}	
Warmup steps	6,000	
Weight decay	0.05	
Gradient clipping max norm	5	
Training precision	bfloat16 with AMP	
Training batch size	2048	512
Epochs	2	5

C.3 Fine-tuning Settings and Baseline Implementations

C.3.1 TSComp Implementation

For downstream adaptation, we use the same input length and patch configuration as in pre-training. All downstream experiments are conducted over three random seeds. Unless stated otherwise, we freeze the encoder parameters of TSComp and train only the task-specific scale-aware downstream pipelines for each downstream task. Detailed hyperparameters for each downstream task are provided in Table 17, and task-specific formulations are described in the corresponding subsections. Tables 18–20 summarize the benchmark datasets used for downstream evaluation, including forecasting, imputation, UEA classification, and UCR anomaly detection datasets.

Table 17: Downstream task fine-tuning hyperparameters for TSComp.

Task	LR	LR Scheduler	Loss	Batch Size	Epochs
Long-term Forecasting	5×10^{-5}	OneCycleLR	$\mathcal{L}_{\text{SED}}, \mathcal{L}_{\text{TIME}}, \mathcal{L}_{\text{COEF}}$	64	10
Imputation	1×10^{-3}		$\mathcal{L}_{\text{TIME}}, \mathcal{L}_{\text{COEF}}$	64	10
Classification	1×10^{-4}		Cross Entropy	16	60
Anomaly Detection	5×10^{-4}		$\mathcal{L}_{\text{TIME}}$	64	5

C.3.2 Multi-task TSFM Baseline Implementations

We compare TSComp with multi-task TSFM baselines across long-term forecasting, imputation, anomaly detection, and classification. **TSPulse** (Ekambaram et al., 2026) learns disentangled temporal, spectral, and semantic representations and selectively combines them using lightweight task-dependent fusers. **GTM** (He et al., 2026) combines frequency-domain attention with reconstruction and autoregressive objectives through a unified hybrid-masking pre-training framework. **SymTime** (Wang et al., 2025) pre-trains time series representations using synthetically generated series paired with their corresponding symbolic expressions. **MOMENT** (Goswami et al., 2024) pre-trains a Transformer encoder on a large multi-domain time-series corpus using masked time series reconstruction over patch-level tokens. In this work, MOMENT denotes MOMENT-base throughout. For all baselines, we load the released pre-trained checkpoints and follow the official downstream fine-tuning procedures as closely as possible, including official heads, optimization settings, and model-specific adaptation strategies when available. We unify core evaluation conditions such as input length, dataset split, metrics, and training epochs within each task. Specifically, we set the input length to 512 and use a patch length and stride of 8 for patch-based models, unless a model requires its original pre-training configuration. When official implementations do not directly support a downstream task, we adapt the released pre-trained backbone using the closest task-consistent protocol. Specifically, TSPulse does not provide a long-term forecasting head and its next-point prediction head is not intended for this setting, so we implement a minimal forecasting adaptation for TSPulse while preserving its released preprocessing, backbone, and decoder.

TSPulse Long-term Forecasting Implementation. We initialize TSPulse from the released pre-trained checkpoint and use the `TSPulseForReconstruction` backbone with its default configuration (Ekambaram et al., 2026). Given an input context $\mathbf{X}_{1:T} \in \mathbb{R}^{T \times C}$ with $T{=}512$, we feed the past window as `past_values` and use the target horizon $\mathbf{X}_{T+1:T+H}$ as `future_values`. The model applies its built-in scaling and partitions the input into non-overlapping length-8 patches. It then encodes them with the pre-trained backbone and decodes them through the official reconstruction decoder. We attach a linear forecasting head to the decoder representation to produce `forecast_output` $\in \mathbb{R}^{B \times H \times C}$, which is trained with a supervised forecasting loss against `future_values`. No additional frequency-domain or task-specific objective is introduced. Thus, this adaptation changes only the final task head while preserving the released TSPulse pipeline for a fair and reproducible long-term forecasting comparison.

Table 18: Dataset statistics used in long-term forecasting and imputation experiments (Wu et al., 2021; Zhou et al., 2021).

Dataset	Channels	Train	Val	Test
ETTh1	7	8,033	2,785	2,785
ETTh2	7	8,033	2,785	2,785
ETTm1	7	33,953	11,425	11,425
ETTm2	7	33,953	11,425	11,425
Exchange	8	4,704	665	1,422
Weather	21	31,010	10,445	10,444
Electricity	321	17,805	2,537	5,165
Traffic	862	11,673	1,661	3,413

Table 19: Dataset statistics for classification experiments (Bagnall et al., 2018).

Dataset	Dim	Num Classes	Train	Val	Test
ArticularyWordRecognition	9	25	275	300	300
AtrialFibrillation	2	3	15	15	15
BasicMotions	6	4	40	40	40
Cricket	6	12	108	72	72
EigenWorms	6	5	128	131	131
Epilepsy	3	4	137	138	138
ERing	4	6	30	270	270
EthanolConcentration	3	4	261	263	263
FaceDetection	144	2	5,890	3,524	3,524
FingerMovements	28	2	316	100	100
HandMovementDirection	10	4	160	74	74
Handwriting	3	26	150	850	850
Heartbeat	61	2	204	205	205
JapaneseVowels	12	9	270	370	370
Libras	2	15	180	180	180
LSST	6	14	2,459	2,466	2,466
MotorImagery	64	2	278	100	100
NATOPS	24	6	180	180	180
PenDigits	2	10	7,494	3,498	3,498
PhonemeSpectra	11	39	3,315	3,353	3,353
RacketSports	6	4	151	152	152
SelfRegulationSCP1	6	2	268	293	293
SelfRegulationSCP2	7	2	200	180	180
SpokenArabicDigits	13	10	6,599	2,199	2,199
StandWalkJump	4	3	12	15	15
UWaveGestureLibrary	3	8	120	320	320
CharacterTrajectories	3	20	1,422	1,436	1,436

Table 20: Dataset statistics for anomaly detection experiments (Wu & Keogh, 2023). Dataset names are compacted by removing the UCR archive index, the `UCR_Anomaly` prefix, perturbation prefixes such as `DISTORTED`/`NOISE`, and trailing anomaly-location indices. When multiple numbered variants belong to the same dataset family, only one representative variant is reported. Channels denotes the number of variables, and Train/Val/Test denote the number of sliding windows.

Dataset	Channels	Train	Val	Test	Dataset	Channels	Train	Val	Test
sddb40	1	24,489	9,489	3,969	park3m	1	3,775	16,632	2,489
BIDMC	1	1,274	204	7,988	qtdbSel1005V	1	2,346	632	2,690
CIMIS44AirTemperature	1	2,346	632	3,672	qtdbSel100MLII	1	2,346	632	2,089
ECG	1	6,631	2,347	19,488	resperation	1	6,632	2,347	9,489
GP711MarkerLFM5z	1	3,060	918	6,488	s20101mML	1	8,060	2,918	2,289
InternalBleeding	1	2,346	632	2,903	s20101m	1	6,631	2,347	2,489
Lab2Cmac011215EPG	1	3,060	918	24,489	sddb49	1	13,774	5,204	5,489
MesoplodonDensirostris	1	6,631	2,347	14,155	sel840mECG	1	11,631	4,347	3,789
PowerDemand	1	5,917	2,061	20,419	tiltAPB	1	6,632	2,347	2,490
TkeepFifthMARS	1	1,989	489	7,322	CHARISfive	1	6,631	2,347	2,624
TkeepFirstMARS	1	1,989	489	7,394	CHARISten	1	17,439	6,669	16,358
TkeepForthMARS	1	1,989	489	7,296	Fantasia	1	13,060	4,918	6,789
TkeepSecondMARS	1	1,989	489	7,376	Italianpowerdemand	1	2,070	9,810	7,859
TkeepThirdMARS	1	1,989	489	7,296	STAFFIIIDatabase	1	23,211	8,978	26,168
WalkingAceleration	1	560	0	4,672	mit14046longtermecg	1	3,498	15,525	15,027
apneaecg	1	6,631	2,347	2,340	mit14134longtermecg	1	7,511	2,698	4,297
gait	1	13,774	5,204	3,989	mit14157longtermecg	1	12,998	4,893	6,869
gaitHunt	1	12,703	4,775	4,039	taichidbS0715Master	1	13,064	4,919	70,252
insectEPG	1	1,631	347	6,489	tilt12744mtable	1	6,632	2,347	18,156
ltstdbs30791AI	1	12,028	4,505	3,346	tilt12754table	1	6,633	2,347	19,514
ltstdbs30791AS	1	15,917	6,061	2,689	tilt12755mtable	1	6,647	2,353	19,455
ltstdbs30791ES	1	13,774	5,204	2,989	weallwalk	1	917	61	5,920

C.4 Task-specific Baseline Models and Implementations

For task-specific model baselines, we reproduce results using officially released implementations whenever available. Most model implementations are adopted from Time-Series-Library (Wu et al., 2023); otherwise, we use the corresponding official GitHub repositories. For fair comparison, we fix the input length to 512 and set the patch length and stride to 8 for patch-based models, unless the original implementation requires a different configuration. We also match the number of training epochs used for the TSFM baselines within each downstream task to control the optimization budget.

SimpleTM (Chen et al., 2025) constructs multi-scale representations using a learnable stationary wavelet transform and models their interactions through attention and geometric-algebra-based operations. **AdaWaveNet** (Yu et al., 2024) performs multi-scale modeling using a learnable lifting-based wavelet transformation for adaptive decomposition and reconstruction. **TimeMixer** (Wang et al., 2024) constructs multiple temporal resolutions through downsampling and models dependencies across scales using Past-Decomposable-Mixing and Future-Multipredictor-Mixing modules. **iTransformer** (Liu et al., 2024a) represents each variate as a token and applies self-attention across variates to model multivariate dependencies. **TSLANet** (Eldele et al., 2024) combines an Adaptive Spectral Block with an Interactive Convolution Block to capture long- and short-term temporal patterns while suppressing high-frequency noise. **FreTS** (Yi et al., 2023) converts time series signals into Fourier representations and applies frequency-domain MLPs to learn inter-series and intra-series dependencies. **TimesNet** (Wu et al., 2023) transforms one-dimensional time series into multiple period-dependent 2D representations and models intra-period and inter-period variations using 2D convolutional kernels. **PatchTST** (Nie et al., 2023) tokenizes time series into subseries-level patches and applies a shared Transformer independently across channels under a channel-independent design. **FEDformer** (Zhou et al., 2022) combines seasonal-trend decomposition with frequency-enhanced Transformer blocks that operate on a sparse subset of frequency components for long-term forecasting. **Non-stationary Transformer** (Liu et al., 2022) combines Series Stationarization with De-stationary Attention to improve predictability while preserving non-stationary temporal dependencies. **Anomaly Transformer** (Xu et al., 2022) models the discrepancy between prior and series associations through Anomaly-Attention and uses the resulting Association Discrepancy as an unsupervised anomaly criterion. **MultiRocket** (Tan et al., 2022) extracts classification features by applying random convolutional kernels and multiple pooling operators to the original series and its first-order differences.

C.4.1 Fine-tuning Pipelines for Downstream Tasks

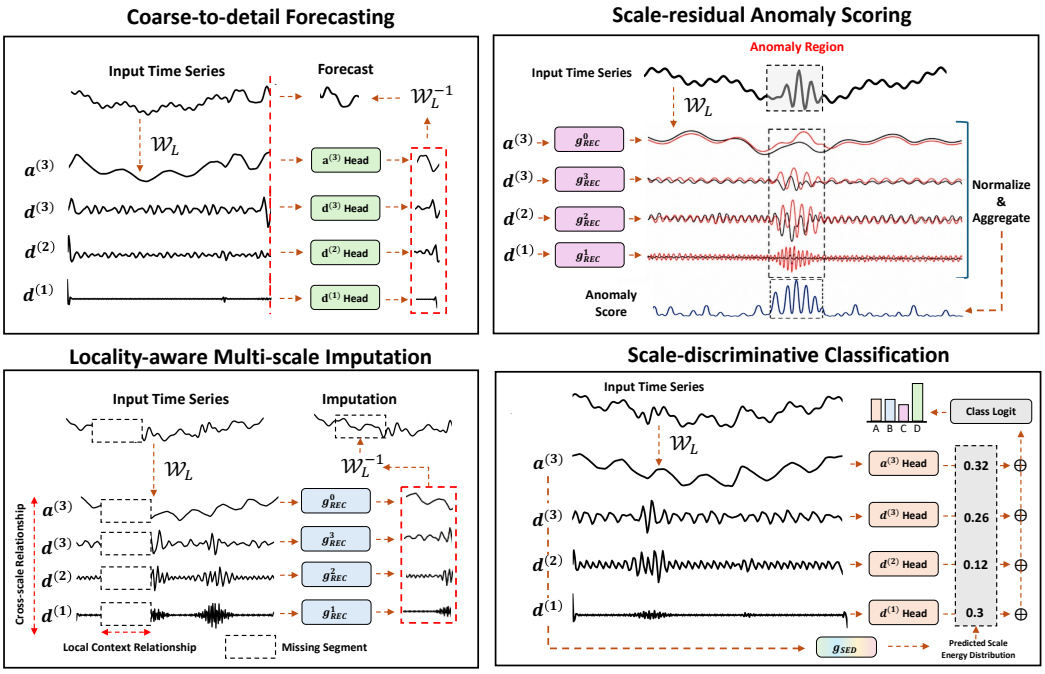


Figure 5: Overview of the downstream task pipelines of TSComp. For each task, the scale-specific representations are adapted through a task-specific scale-aware downstream pipeline.

Figure 5 summarizes how the scale-specific representations of TSComp are adapted to each downstream task. Rather than collapsing all information into a single time-domain representation, TSComp transfers scale-specific representations through task-specific scale-aware pipelines for forecasting, imputation, classification, and anomaly detection. This design keeps the downstream adaptation aligned with the pre-training objective by preserving scale-specific temporal structures and their cross-scale composition.

C.4.2 COARSE-TO-DETAIL FORECASTING

For long-term forecasting, TSComp uses a coarse-to-detail pipeline that transfers the scale-specific representations into scale-specific future prediction. Given the encoder outputs $\{\mathbf{Z}_s^R\}_{s\in\mathcal{S}}$, the pipeline applies a scale-specific forecasting head to each stream to estimate its future coefficient sequence. The predicted future coefficient streams are then recomposed through inverse SCNorm (iSCNorm) and inverse DWT (iDWT) to obtain the final time-domain forecast. Unlike direct time-domain forecasting from a single scale-agnostic representation, this pipeline preserves the compositional structure of the forecast. It models coarse future dynamics and fine-scale temporal variations separately, combining them only at the final recomposition stage.

Training is performed by jointly optimizing three objectives. First, the coefficient forecasting loss supervises scale-specific future coefficients, $\mathcal{L}_{\mathrm{COEF}} = \sum_{s\in\mathcal{S}} \ell\big(\mathbf{c}_s^{\mathrm{fut}}, \hat{\mathbf{c}}_s^{\mathrm{fut}}\big)$, where $\mathbf{c}_s^{\mathrm{fut}}$ is the ground-truth future coefficient stream at scale s. Second, the time-domain forecasting loss enforces accurate recomposed forecasts after iDWT, $\mathcal{L}_{\mathrm{TIME}} = \ell\Big(\mathbf{Y}, \hat{\mathbf{Y}}\Big)$. Third, the scale energy distribution loss encourages consistency in the scale-wise allocation of energy in the future window, $\mathcal{L}_{\mathrm{SED}} = D_{\mathrm{KL}}\big(\mathbf{r}^{\,\mathrm{fut}}, \hat{\mathbf{r}}^{\mathrm{fut}}\big)$, where $\mathbf{r}^{\mathrm{fut}}$ and $\hat{\mathbf{r}}^{\,\mathrm{fut}}$ denote the ground-truth and predicted scale energy distribution, respectively. The overall forecasting objective is $\mathcal{L}_{\mathrm{forecast}} = \mathcal{L}_{\mathrm{SED}} + \mathcal{L}_{\mathrm{TIME}} + \mathcal{L}_{\mathrm{COEF}}$.

This design makes the coarse-to-detail forecasting pipeline consistent with the pre-training objective. Specifically, the model is trained not only to forecast future values, but also to preserve how future dynamics are distributed across scales and how those scale-specific components are recomposed into the future time series.

C.4.3 LOCALITY-AWARE MULTI-SCALE IMPUTATION

For imputation, TSComp uses a locality-aware multi-scale imputation pipeline that preserves the pre-trained scale-specific representations throughout masked-region recovery and recomposition. Given the encoder outputs $\{\mathbf{Z}_s^R\}_{s\in\mathcal{S}}$, the pipeline applies a scale-specific reconstruction head to each stream to recover the masked coefficient sequence. The reconstructed coefficient streams are then recomposed through inverse SCNorm followed by inverse DWT (iDWT) to obtain the final imputed time series. Unlike direct time-domain recovery from a single scale-agnostic representation, this pipeline preserves the compositional structure of imputation. It reconstructs coarse contextual structure and fine local variations separately, combining them only at the final recomposition stage.

Training is performed by jointly optimizing two objectives. First, the scale-aware coefficient reconstruction supervises masked-region recovery at each scale, $\mathcal{L}_{\mathrm{COEF}} = \sum_{s\in\mathcal{S}} \ell\big(\mathbf{c}_s^{\mathrm{imp}}, \hat{\mathbf{c}}_s^{\mathrm{imp}}\big)$, where $\mathbf{c}_s^{\mathrm{imp}}$ denotes the ground-truth masked coefficient stream at scale s. Second, the time-domain reconstruction loss enforces consistency after recomposition, $\mathcal{L}_{\mathrm{TIME}} = \ell\Big(\mathbf{X}, \hat{\mathbf{X}}\Big)$. The overall imputation objective is $\mathcal{L}_{\mathrm{imp}} = \mathcal{L}_{\mathrm{COEF}} + \mathcal{L}_{\mathrm{TIME}}$. This design keeps the locality-aware multi-scale imputation pipeline aligned with the pre-training objective. The model is required not only to recover masked values, but also to preserve how the missing region is represented across scales and how those recovered components are recomposed into the observed time series. For downstream adaptation, we perform full end-to-end fine-tuning of TSComp, updating both the pre-trained encoder and the reconstruction head. The imputation head is architecturally identical to and initialized from the reconstruction head used during pre-training.

C.4.4 SCALE-DISCRIMINATIVE CLASSIFICATION

For classification, TSComp uses a scale-discriminative classification pipeline that transfers the scale-specific representations to class prediction while preserving scale-specific temporal evidence. Given the encoder outputs $\{\mathbf{Z}_s^R\}_{s\in\mathcal{S}}$, the pipeline applies an independent scale-specific classifier g_s to each stream, producing stream-wise logits $\mathbf{o}_s = g_s(\mathbf{Z}_s^R)$. These per-scale outputs are then aggregated using the predicted scale energy distribution learned during pre-training, so that the final class prediction is obtained as $\hat{\mathbf{y}} = \sum_{s\in\mathcal{S}} \hat{r}_s\, \mathbf{o}_s$, where $\hat{r}_s$ denotes the predicted scale energy distribution at scale s. In this way, the pipeline can emphasize coarse-scale evidence for global temporal morphology and fine-scale evidence for local discriminative variations according to the input time series.

Training is performed with $\mathcal{L}_{\mathrm{CLS}} = \mathrm{CE}(y, \hat{\mathbf{y}})$, where CE denotes the cross-entropy loss. During downstream adaptation, the pre-trained encoder is frozen, and only the classification modules within

the scale-discriminative pipeline are trained. This design keeps the scale-discriminative classification pipeline aligned with the pre-training objective. Rather than collapsing all scales into a single shared representation, TSComp forms the final decision by selectively aggregating scale-specific class evidence according to the predicted contribution of each scale to the input time series.

C.4.5 SCALE-RESIDUAL ANOMALY SCORING

For anomaly detection, TSComp uses a scale-residual anomaly scoring pipeline that transfers the pre-trained scale-specific representations to anomaly scoring through scale-wise reconstruction residuals. Given the encoder outputs $\{\mathbf{Z}_s^R\}_{s\in\mathcal{S}}$, the pipeline applies a scale-specific reconstruction head to each stream to reconstruct the coefficient sequence, yielding $\hat{\mathcal{C}} = \{\hat{\mathbf{c}}_s\}_{s\in\mathcal{S}}$. The reconstructed coefficient streams are then recomposed through inverse SCNorm followed by inverse DWT (iDWT) to obtain the time-domain reconstruction $\hat{\mathbf{X}}$. During downstream adaptation, the pre-trained encoder is frozen, and only the reconstruction modules within the anomaly-scoring pipeline are trained using the time-domain reconstruction loss $\mathcal{L}_{\text{TIME}} = \ell(\mathbf{X}, \hat{\mathbf{X}})$.

Anomaly evidence is not taken directly from a single global time-domain reconstruction error. Instead, for each scale $s \in \mathcal{S}$, the pipeline computes a coefficient-level residual $e_s = \ell(\tilde{\mathbf{c}}_s, \hat{\tilde{\mathbf{c}}}_s)$ between the normalized ground-truth and reconstructed coefficients. Each residual sequence is then robustly normalized within the input window as $\tilde{e}_s = \max\big((e_s - \text{median}(e_s))/(\text{MAD}(e_s) + \epsilon), 0\big)$, where $\text{MAD}(e_s) = \text{median}(|e_s - \text{median}(e_s)|)$. The normalized coefficient residuals are projected back to the original time axis according to their temporal support, producing a time-aligned score map $\mathbf{a}_s \in \mathbb{R}^T$ for each scale. The final anomaly score is obtained by uniform aggregation across scales, $\mathbf{a} = \sum_{s\in\mathcal{S}} w_s \mathbf{a}_s$ with $w_s = 1/|\mathcal{S}|$. This design keeps the scale-residual anomaly scoring pipeline aligned with the pre-training objective while making anomaly detection sensitive to residual evidence that appears at different temporal resolutions. Rather than relying on a single scale-agnostic reconstruction error, TSComp detects anomalies from scale-specific residual structure aligned back to the original time axis. The anomaly head is initialized from, and kept architecturally identical to, the reconstruction head used during pre-training.

D REAL-WORLD INDUSTRIAL TIME SERIES REGRESSION

Beyond the four benchmark tasks in the main experiments, we further evaluate TSComp on a real-world industrial regression task using a dataset collected from an industrial manufacturing process. This experiment assesses whether the multi-scale representations learned during pre-training transfer effectively to industrial regression tasks, where temporal dynamics are heterogeneous across scales and standard benchmark evaluations may not fully capture practical generalization.

Dataset Description. We utilize a lube base oil production dataset collected from the industrial manufacturing process described in Cho & Lim (2024), which involves four sequential stages that refine unconverted oil into high-quality base oil products. The four stages consist of two chemical reaction towers followed by two distillation towers. Each stage is equipped with multiple sensors monitoring control parameters including temperature, pressure, flow rate, and additive input. The dataset contains 18,331 observations gathered at 10-minute intervals from 22 channels (sensors and raw material variables), covering March 2021 to August 2021. The regression target is Pour Point, a key quality indicator that characterizes the flowability of the final base oil product. Accurate prediction of Pour Point enables operators to monitor product quality in real time and adjust process conditions accordingly, which is essential for maintaining consistent product quality throughout the production process. The dataset is split chronologically into training (60%), validation (20%), and test (20%) partitions, and the model predicts the Pour Point at the end of each input window.

Implementation. Following the channel-independent convention adopted in TSComp pre-training, each channel is independently encoded, and the resulting per-channel representations are concatenated to form a unified representation before applying a regression head. We evaluate two fine-tuning strategies: linear probing, which freezes the pre-trained encoder and trains only the regression head, and end-to-end fine-tuning, which updates all model parameters. All models are optimized using the MSE loss. Each experiment is repeated with three random seeds, and the mean and standard

deviation are reported. All other implementation details follow the long-term forecasting setting described in Appendix C.3.

Table 21: Pour Point regression results averaged over three random seeds. MSE and MAE are reported as mean ± standard deviation for each adaptation strategy. The best and second-best results within each strategy are highlighted in bold and underlined, respectively.

	Linear probing		End-to-end	
Model	MSE	MAE	MSE	MAE
TSComp-small	0.231 ±0.021	0.372 ±0.017	**0.230** ±0.002	<u>0.183</u> ±0.007
TSComp-base	**0.076** ±0.003	**0.230** ±0.006	**0.203** ±0.007	**0.163** ±0.008
GTM	0.160 ±0.045	0.279 ±0.040	0.281 ±0.046	0.260 ±0.041
TSPulse	<u>0.103</u> ±0.012	<u>0.258</u> ±0.014	<u>0.270</u> ±0.034	0.279 ±0.025
SymTime	0.187 ±0.025	0.319 ±0.024	0.345 ±0.071	0.317 ±0.062
MOMENT	0.391 ±0.026	0.484 ±0.013	0.470 ±0.028	0.478 ±0.022

Results. Table 21 reports the Pour Point regression results averaged over three random seeds under two fine-tuning strategies. Under linear probing, TSComp-base achieves the best performance in both MSE and MAE, followed by TSPulse. TSComp-base also shows relatively low standard deviations across random seeds, indicating stable performance under the frozen-encoder setting. Under end-to-end fine-tuning, TSComp-small and TSComp-base achieve the lowest MSE, while TSComp-base obtains the lowest MAE. The two TSComp variants show similar MSE, whereas TSComp-base performs better in terms of MAE. TSComp-small also exhibits a low standard deviation in MSE across the three runs.

The relative performance of the two fine-tuning strategies varies across models. End-to-end fine-tuning improves TSComp-small, while linear probing yields a lower MSE for TSComp-base. A similar trend is observed for several pre-trained baselines, for which linear probing performs better than end-to-end fine-tuning. This suggests that updating the entire encoder does not consistently improve performance on this industrial regression dataset. Overall, TSComp achieves the best results under both fine-tuning strategies. These results show that the multi-scale representations learned during pre-training can be transferred to industrial regression tasks. In this experiment, the representations from individual channels are simply concatenated before the regression head, without using a task-specific scale-aware regression pipeline. A regression-specific design that more directly utilizes the multi-scale representations may provide further improvements.

E Additional Ablation Studies

All ablation studies are conducted using TSComp-small under the same experimental settings as the main experiments, varying only the component under investigation. We provide additional analyses of the key design choices below.

E.1 Effect of Decomposition Strategy

To examine the effect of decomposition strategies within TSComp, we replace the orthonormal DWT with Fourier decomposition, a learnable wavelet transform, and a residual pyramid while keeping the remaining pre-training and the scale-aware pipeline unchanged. The residual pyramid recursively downsamples the signal using average pooling and computes detail residuals by subtracting the upsampled coarse representation at each level, yielding a final coarse stream and a hierarchy of residual streams.

As shown in Table 22, no single decomposition strategy is uniformly optimal across all tasks. The orthonormal DWT achieves the best forecasting and imputation performance on ETTh1 and remains competitive in classification, while the learnable wavelet transform performs slightly better for classification. Beyond task-specific performance, the orthonormal DWT is the only evaluated strategy that jointly satisfies the four structural requirements for a scale-decomposed space: temporal localization, an explicit scale hierarchy, perfect recomposition, and energy preservation (Table 4). These properties directly support the scale-balanced tokenization, cross-scale modeling, time-domain

Table 22: Effect of decomposition strategies on ETTh1 forecasting and imputation, together with average classification accuracy across 27 UEA datasets. The best and second-best results are highlighted in bold and underlined, respectively.

	Decomposition Strategy			
Task	Fourier	Learnable	Residual Pyramid	Orthonormal DWT (TSComp)
Forecasting (ETTh1, MSE)	0.4388	0.4288	<u>0.4267</u>	**0.4257**
Imputation (ETTh1, MSE)	0.2548	<u>0.1819</u>	0.1842	**0.1813**
Classification (Avg. Acc.)	53.13	**68.32**	66.10	<u>68.18</u>

reconstruction, and energy-based objectives of TSComp. Thus, although our choice for TSComp is not universally optimal for every downstream task, its consistently strong performance and structural compatibility make it a balanced and well-matched choice for the proposed pre-training framework.

E.2 Effect of Decomposition Level

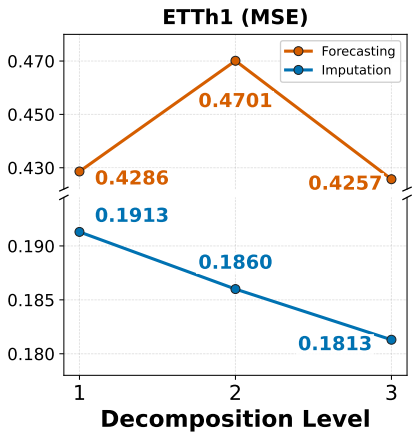


Figure 6: Effect of wavelet decomposition level on downstream performance.

We investigate the effect of wavelet decomposition depth on downstream performance. This ablation shows that the decomposition level shapes the quality of multi-scale representations. As shown in Figure 6, if the decomposition level is too low, the resulting wavelet streams fail to disentangle coarse trends and fine-grained variations, making it difficult for the model to capture scale-specific temporal patterns and their composition. In particular, the decomposition depth determines how temporal variation is distributed between approximation and detail streams, which affects how clearly coarse and fine-scale dynamics can be separated and transferred. A suitable decomposition level therefore provides a more balanced representation space in which long-range structure and local variation can be modeled jointly for downstream adaptation. These results further support that benefit of TSComp comes not only from applying wavelet decomposition, but from organizing the time series into multiple scales where scale-specific dynamics and cross-scale composition can be learned.

E.3 Effect of Wavelet Basis

Table 23: Effect of wavelet basis selection across four downstream tasks. Forecasting and imputation report ETTh1 MSE averaged over prediction horizons and mask ratios, respectively. Anomaly detection reports average VUSPR across 245 datasets, and classification reports average accuracy across 27 UEA datasets. The best and second-best results are highlighted in bold and underlined, respectively.

Wavelet Basis	ETTh1 Forecasting MSE↓	ETTh1 Imputation MSE↓	Anomaly Detection VUSPR↑	Classification Avg. Acc.↑
Symlet-4 (TSComp)	**0.4257**	<u>0.1813</u>	<u>0.0421</u>	68.18
Bior 4.4	0.4286	**0.1782**	0.0369	69.26
Db4	0.4325	0.1943	0.0412	**70.50**
Haar	<u>0.4284</u>	0.1876	**0.0562**	<u>69.89</u>

We evaluate the effect of wavelet basis selection by comparing Symlet-4, Bior 4.4, Db4, and Haar. Table 23 shows that the best-performing basis varies across the evaluated settings: Symlet-4 achieves the lowest forecasting MSE on ETTh1, Bior 4.4 yields the lowest imputation MSE on ETTh1, Db4 attains the highest average classification accuracy, and Haar achieves the highest average anomaly detection VUSPR. These results indicate that basis selection affects downstream performance, with

no single basis consistently achieving the best results across these settings. We therefore use Symlet-4 as the default basis for its combination of orthogonality, compact support, near symmetry, and four vanishing moments. This choice provides a balance of mathematical properties for multi-scale decomposition, without implying universal downstream optimality. These findings motivate future research on selective pre-training strategies that tailor the wavelet basis to the characteristics of specific datasets.

E.4 Effect of Stream Summary Tokens

Table 24: Effect of Stream Summary Tokens (SST) across four downstream tasks. Forecasting and imputation report ETTh1, anomaly detection reports average VUSPR across 245 datasets, and classification reports average accuracy across 27 UEA datasets.

Task	TSComp	w/o SST
Forecasting (ETTh1, MSE↓)	0.4257	**0.4253**
Imputation (ETTh1, MSE↓)	**0.1813**	0.1836
Anomaly Detection (Avg. VUSPR↑)	**0.0421**	0.0373
Classification (Avg. Acc.↑)	**68.18**	65.57

SSTs aggregate information within each coefficient stream to predict the input window's scale energy distribution during pre-training. Alongside patch-level coefficient reconstruction, $\mathcal{L}_{\mathrm{SED}}$ directly supervises SST-based prediction of the window-level scale energy distribution, making this aggregate property an explicit target of representation learning. To evaluate this design, we pre-train a variant without SST and the $\mathcal{L}_{\mathrm{SED}}$ objective. As shown in Table 24, the full model achieves lower imputation MSE, higher anomaly detection VUSPR, and higher classification accuracy, while yielding slightly higher forecasting MSE. In classification, SST-based energy estimates directly guide the aggregation of scale-specific class evidence. The gains in imputation and anomaly detection extend beyond this aggregation mechanism, as both tasks reconstruct coefficients from patch representations. However, supervision of the observed window's energy composition does not necessarily improve prediction of future values. Overall, these results suggest that SSTs trained with $\mathcal{L}_{\mathrm{SED}}$ offer benefits beyond scale energy distribution calibration, supporting classification, imputation, and anomaly detection in the evaluated settings, although improvements are not observed across all evaluated tasks.

E.5 Effect of Multi-scale Directional Masking

Table 25: Effect of Multi-scale Directional Masking (MDM) across four downstream tasks. Forecasting and imputation report ETTh1 MSE, anomaly detection reports average VUSPR across 245 datasets, and classification reports average accuracy on 27 UEA datasets.

Task	TSComp	w/o MDM
Forecasting (ETTh1, MSE↓)	**0.4257**	0.4275
Imputation (ETTh1, MSE↓)	**0.1813**	0.1950
Anomaly Detection (Avg. VUSPR↑)	0.0421	**0.0473**
Classification (Avg. Acc.↑)	**68.18**	67.48

We compare Multi-scale Directional Masking (MDM) with random masking, denoted as w/o MDM, to examine the contribution of the pre-training masking strategy to downstream performance. Random masking selects coefficient patches without imposing a directional relationship across scales. MDM instead masks temporally aligned coefficient patches along coarse-to-fine or fine-to-coarse directions, encouraging reconstruction from complementary scale information in both directions. Both strategies use the same expected coefficient-patch masking ratio, with the remaining pre-training and downstream adaptation settings held fixed. Table 25 reports forecasting and imputation MSE on ETTh1, average classification accuracy, and average anomaly-detection VUSPR. MDM yields modest improvements in forecasting and classification and a more pronounced improvement in imputation, whereas random masking achieves higher average anomaly-detection VUSPR. The imputation result

suggests that structuring missing coefficient patches through MDM provides useful pre-training conditions for recovering missing observations. These results support the utility of MDM particularly for imputation on ETTh1, although it does not improve performance across all evaluated downstream settings.

E.6 Effect of Scale-wise Coefficient Normalization

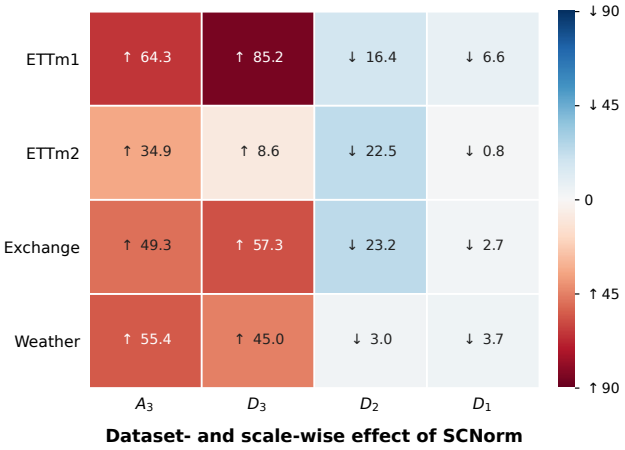


Figure 7: Effect of SCNorm on coefficient reconstruction across four datasets. Each cell reports the percentage decrease or increase in NRMSE with SCNorm, relative to TSComp without SCNorm.

SCNorm reduces magnitude differences among wavelet coefficient streams before tokenization, aiming to reduce the influence of raw coefficient amplitudes on cross-scale representation learning. Here, scale balancing refers to comparable input magnitudes rather than equal importance across scales; relative scale contributions are supervised through scale energy distribution calibration, and original coefficient magnitudes are restored through inverse SCNorm before recomposition. Figure 7 examines the effect of this design on coefficient reconstruction. Although the magnitude of the effect varies across datasets, a recurring scale-dependent pattern emerges: SCNorm generally improves reconstruction of the finest detail stream, D_1, while increasing errors for the coarse streams, A_3 and D_3, and improving D2 by a dataset-dependent margin. These results reveal a trade-off between coarse- and fine-scale reconstruction fidelity associated with SCNorm. This pattern is consistent with its intended role of reducing the dominance of high-magnitude coarse streams and redistributing representational emphasis toward finer-scale variations.

F Qualitative Analysis

F.1 TSComp Reconstruction Visualization

Figure 8 provides qualitative evidence that TSComp faithfully reconstructs observed time series by learning the scale-specific components that compose each signal. Across time series with smooth trend-dominant patterns, TSComp closely follows the global trajectory while preserving gradual turning points. For signals with denser local oscillations and more irregular short-term fluctuations, the reconstruction remains well aligned with the input even in rapidly changing regions. It also captures both slowly varying structure and localized deviations, while preserving irregular and non-periodic movements without oversmoothing. These visualizations show that TSComp does not merely reconstruct signals, but learns how different temporal patterns are organized across scales and how those scale-specific elements are jointly composed into the observed time series. As a result, TSComp achieves faithful reconstruction across heterogeneous temporal patterns.

F.2 Qualitative Analysis of Multi-scale Composition

Figure 9 provides qualitative evidence that the reconstruction behavior of TSComp is governed by input-dependent scale-specific temporal dependency learning and adaptive cross-scale composition.

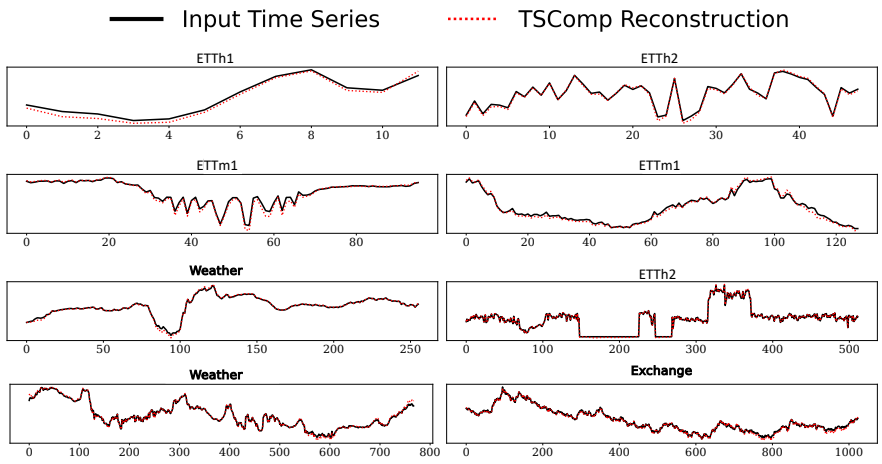


Figure 8: Reconstruction visualization. We visualize eight reconstruction examples drawn from five datasets (ETTh1, ETTh2, ETTm1, Exchange, Weather). The input time series is shown in black, and the model reconstruction is shown in red. By learning both scale-specific temporal structures and their cross-scale composition in the wavelet domain, TSComp faithfully reconstructs the original signal across diverse temporal patterns.

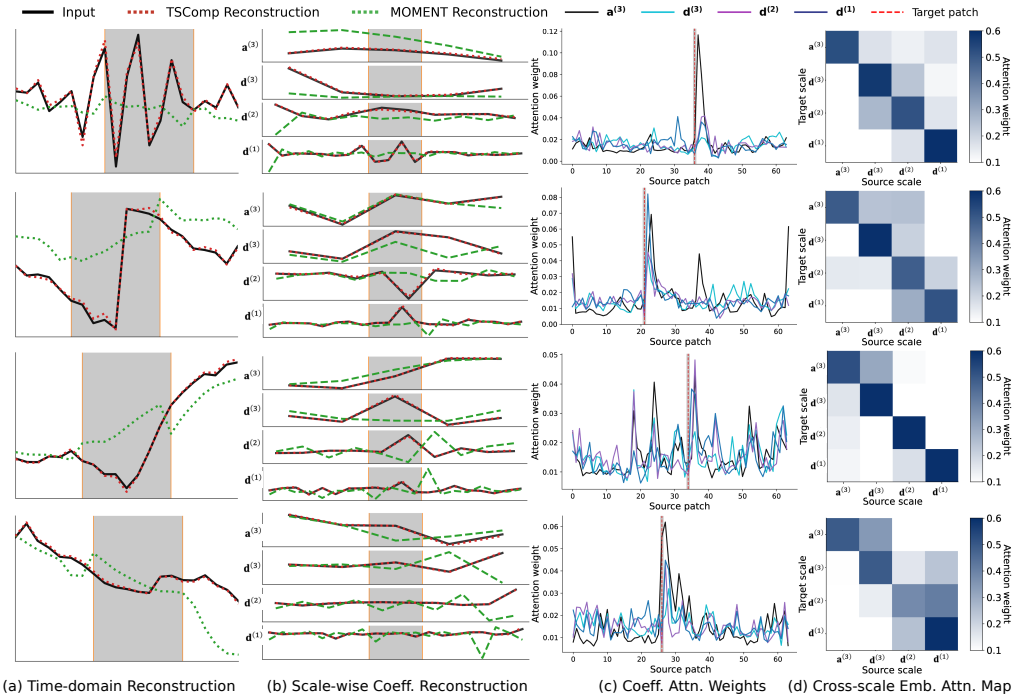


Figure 9: Qualitative examples of multi-scale composition. Each row shows a different synthetic event pattern. From left to right, we compare time-domain reconstruction, scale-wise coefficient reconstruction in the selected event window, coefficient attention weights over source patches, and the corresponding cross-scale embedding attention map.

For the analysis, we generate synthetic time series through the additive compositional process so that each signal contains overlapping coarse structure, local transient events, and fine-scale irregularities:

$$\mathbf{x}^{\text{syn}} = \text{Norm}\left(\sum_{m=1}^{M} \alpha_m \frac{\mathbf{f}_m}{\text{Std}(\mathbf{f}_m)} + \beta\, \mathbf{e} + \boldsymbol{\epsilon}^{(g)} + \boldsymbol{\epsilon}^{(i)} \right), \tag{3}$$

where $M \in \{2, \ldots, 6\}$ is the number of randomly selected primitive functions, $\mathbf{f}_m$ denotes the m-th primitive component, α_m is its random coefficient, $\mathbf{e}$ is a localized event pattern with random type and support, β is a random event-amplitude coefficient controlling the strength of $\mathbf{e}$, $\boldsymbol{\epsilon}^{(g)}$ is low-amplitude Gaussian noise, $\boldsymbol{\epsilon}^{(i)}$ is sparse impulse noise, and $\mathrm{Norm}(\cdot)$ denotes final sequence-level standardization. The primitive pool includes sinusoidal waves, polynomial trends, chirp-like frequency-varying patterns, piecewise trends, step changes, local bursts, sparse spikes, localized oscillations, frequency-jump patterns, and amplitude-modulated oscillations. This construction yields controlled signals with overlapping coarse and fine temporal structures, making them suitable for analyzing whether TSComp learns scale-specific components and their cross-scale composition.

In general, both the coefficient attention and the cross-scale embedding reveal that the multi-scale dependency structure varies with the input time series pattern. In the first example, which exhibits a high-frequency oscillatory event, faithful reconstruction is accompanied by strong local alignment in the detail streams, especially $\mathbf{d}^{(1)}$ and $\mathbf{d}^{(2)}$, together with concentrated coefficient attention around the target patch, indicating that sharp local dynamics are recovered from tightly localized multi-scale context, whereas MOMENT mainly follows the overall trend without faithfully reconstructing the fine-grained dynamics. In the second example, which contains a valley-to-jump transition, the approximation stream $\mathbf{a}^{(3)}$ and the coarser detail streams show stronger cross-scale interaction, suggesting that reconstruction depends more on structural transition information rather than on fine local variation alone. In the third example, where the event follows a smoother valley-to-rise pattern, the coefficient attention becomes more distributed across nearby source patches, indicating that the model integrates broader temporal context when the event evolves over a longer support. In the fourth example, which shows a gradual decay, the cross-scale embedding map exhibits more broadly distributed off-diagonal interactions, suggesting that faithful reconstruction requires more active information exchange between coarse trend structure and lower-scale corrective details. Taken together, these examples show that the relations among scales are not fixed, but are adaptively organized according to the input time series. This further supports the central claim of TSComp: faithful reconstruction is achieved not by a single shared scale-agnostic representation, but by learning input-dependent scale-specific temporal dependencies and their cross-scale composition.

F.3 Case Studies on Downstream Tasks

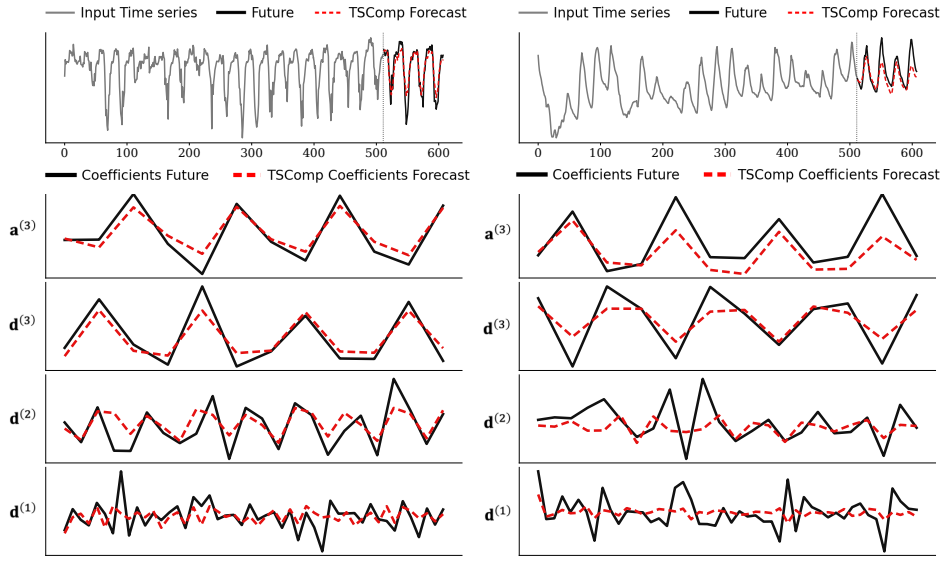


Figure 10: Case studies on long-term forecasting for ETTh1 and ETTh2 with input length 512 and horizon 96. The top panels show the input time series, the ground-truth future window, and the forecast of TSComp in the time domain, while the bottom panels compare the corresponding future wavelet coefficient streams at each scale.

Long-term Forecasting. Figure 10 provides qualitative evidence of how TSComp performs long-term forecasting through Coarse-to-detail Forecasting. In both ETTh1 and ETTh2, the top panels show that TSComp accurately follows the future trajectory in the time domain, preserving not only the overall trend but also local oscillatory patterns and turning points. The bottom panels further show that this forecasting quality arises from faithful forecasts of the future wavelet coefficient streams at multiple scales. In particular, the approximation stream captures the coarse future structure, while the detail streams preserve finer temporal variations, indicating that TSComp forecasts the future by modeling scale-specific dynamics rather than by directly regressing a single time-domain sequence. These case studies further support that long-term forecasting benefits from explicitly forecasting scale-specific temporal patterns and recomposing them into the future horizon.

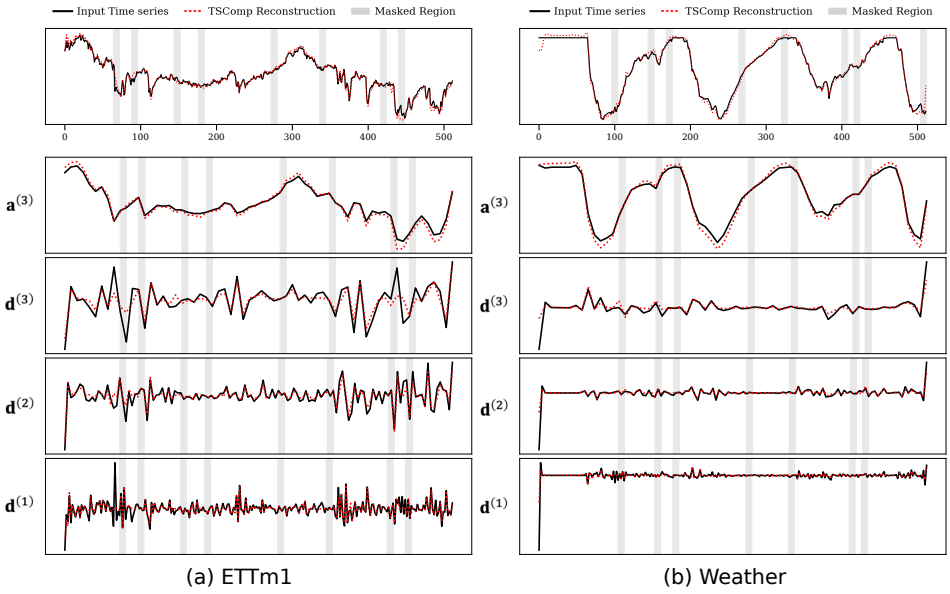


Figure 11: Case studies on imputation for ETTm1 and Weather under a mask ratio of 0.125. The top panels compare the masked input time series and the reconstruction of TSComp in the time domain, while the bottom panels show the corresponding wavelet coefficient streams and their reconstructions at each scale.

Imputation. Figure 11 shows case studies illustrating how TSComp reconstructs masked regions through scale-specific reconstruction. In the top panels, TSComp recovers the missing time-domain segments while preserving their continuity with the surrounding observed context. The bottom panels further show that this imputation quality arises from faithful reconstruction of the corresponding wavelet coefficient streams across multiple scales. In particular, the approximation stream preserves coarse structural continuity, while the detail streams recover finer local variations within the masked region. Since the masked patches are distributed across the sequence, these examples also suggest that TSComp performs imputation by exploiting complementary multi-scale context beyond local interpolation alone. Moreover, the two datasets show different reconstruction patterns across scales, indicating that the relative importance of coarse and fine-scale recovery can vary with the temporal structure of the data. These examples illustrate that TSComp performs imputation not by relying only on direct time-domain recovery, but by reconstructing complementary multi-scale components and recomposing them into the observed time series.

Classification. Figure 12 provides qualitative evidence that the temporal scales relevant for classification vary substantially across datasets and classes. In some datasets, discriminative evidence is concentrated in the approximation stream, indicating that coarse global structure is most informative, whereas in others the detail streams are more strongly activated, suggesting that class distinction depends more on localized or transient variations. For example, in AtrialFibrillation, the strongest activations are concentrated in the approximation stream, implying that the coarse pattern plays the

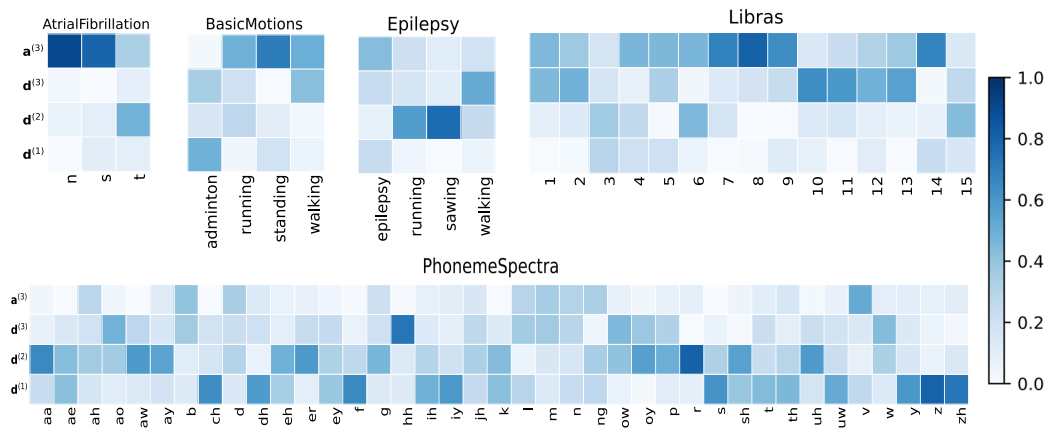


Figure 12: Visualization of class-wise scale importance on representative classification datasets. Rows correspond to wavelet scales and columns correspond to classes. Different datasets and classes exhibit distinct scale-activation patterns, showing that class-discriminative temporal evidence is not uniformly distributed across scales. The color bar indicates the normalized relative contribution of each scale to each class, where darker colors denote higher scale importance.

dominant role in distinguishing classes. In contrast, PhonemeSpectra shows much stronger activation in the detail streams, indicating that finer local variations are more discriminative than the global shape for the specific dataset. Epilepsy exhibits a more class-dependent pattern, where different classes emphasize different scales, suggesting that no single temporal resolution is sufficient to capture all discriminative evidence. Similarly, BasicMotions and Libras show that coarse and mid-scale structures are activated differently across classes, indicating that class-relevant temporal information is distributed across multiple scales. These examples show that class-relevant temporal evidence is not uniformly distributed across scales. These results support our motivation for scale-discriminative classification. Rather than collapsing all temporal information into a single representation, TSComp aggregates scale-specific representations according to their discriminative relevance. This enables more effective classification across heterogeneous time series patterns.

Anomaly Detection. As shown in Figure 13, anomaly-detection case studies provide qualitative evidence of how TSComp identifies anomalous regions through scale-specific residual structure. In the top panels, the time-domain reconstruction remains visually close to the ground-truth signal, yet the anomaly score becomes highly activated only around the true anomalous segments. The coefficient-stream panels explain this behavior more clearly: anomalies are not expressed uniformly across scales, but instead emerge as reconstruction discrepancies across different scales. The approximation stream captures broader structural deviations, while the detail streams highlight more localized and sharp irregularities. Moreover, the two examples show different score distributions across scales, indicating that the temporal resolution at which anomaly evidence appears depends on the anomaly type. In Figure 13(a), the anomalous segment is mainly detected through the approximation stream $\mathbf{a}^{(3)}$ and the detail stream $\mathbf{d}^{(3)}$, whereas in Figure 13(b), the anomaly evidence is concentrated more strongly in the $\mathbf{d}^{(2)}$ and $\mathbf{d}^{(1)}$ streams. This difference shows that anomalous patterns are not expressed at a fixed temporal scale, but can emerge through different combinations of coarse structural deviation and fine local irregularity depending on the anomaly. These results suggest that TSComp detects anomalies not from a single aggregated reconstruction error, but by preserving and integrating complementary residual evidence across multiple scales, enabling more effective range-aware anomaly detection.

G Theoretical Properties

This section analyzes four properties of the pre-training formulation in Section 4: the boundary effect avoided by coefficient-domain masking, the stability of the scale energy distribution, the statistical targets of the reconstruction and calibration losses, and the inherent bias that the time-

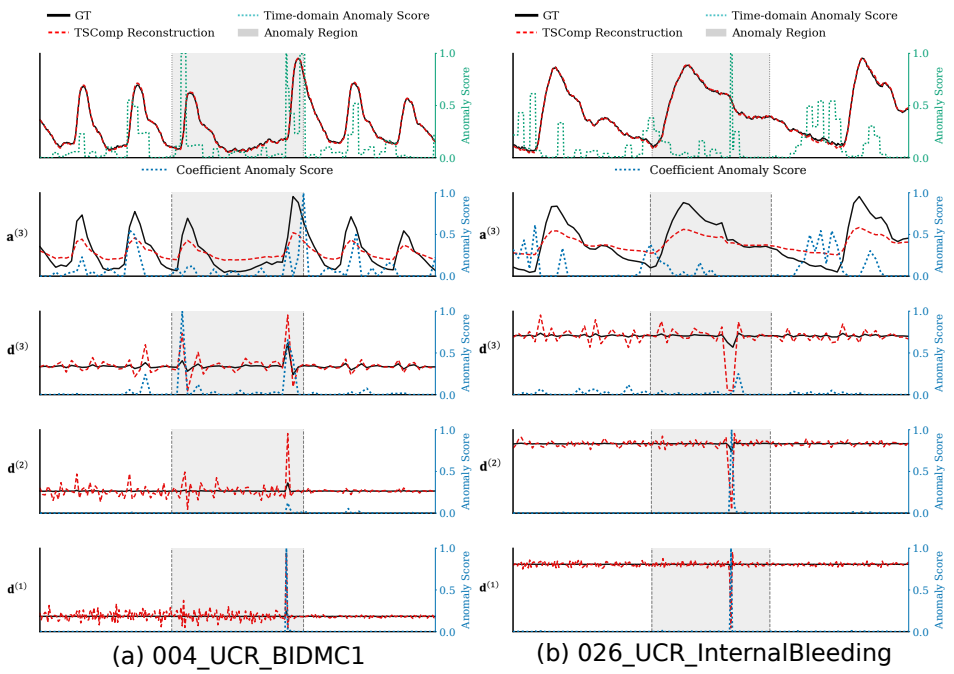


Figure 13: Case studies on anomaly detection for two UCR datasets. The top panels compare the ground-truth signal, the reconstruction of TSComp, and the time-domain anomaly score, while the bottom panels show the corresponding wavelet coefficient streams and coefficient anomaly scores across scales.

domain reconstruction loss imposes on scale-wise reconstruction. We use the stream index set $\mathcal{S} = \{0, 1, \ldots, L\}$, with $\mathbf{c}_0 = \mathbf{a}^{(L)}$ and $\mathbf{c}_s = \mathbf{d}^{(s)}$ for $s = 1, \ldots, L$.

G.1 Boundary Effects and the Motivation for Coefficient-Domain Masking

TSComp applies the DWT and SCNorm before masking selected normalized coefficient patches. We compare this design with a counterfactual *mask-then-decompose* pipeline, in which a time-domain interval is masked before the DWT. Since wavelet coefficients depend on local neighborhoods of input samples, time-domain masking can also change coefficients whose nominal locations lie outside the masked interval (Daubechies, 1992; Vetterli & Kovacevic, 1995).

Let $\mathbf{x} \in \mathbb{R}^T$ be the RevIN-normalized input, and let $B \subseteq \{1, \ldots, T\}$ be a time-domain masked set. Define $M_B\mathbf{x}$ by setting x_t to zero for $t \in B$ and leaving the remaining samples unchanged. For stream s, let A_s be the finite-dimensional DWT analysis operator, including the chosen linear boundary handling. Write

$$\mathbf{c}_s = A_s\mathbf{x}, \qquad \mathbf{c}_s^B = A_s M_B\mathbf{x}, \qquad c_{s,k} = \langle \mathbf{x}, \eta_{s,k} \rangle, \tag{4}$$

where $\eta_{s,k} \in \mathbb{R}^T$ is the analysis vector for coefficient (s, k). Its effective support is $I_{s,k} = \{t : \eta_{s,k}(t) \neq 0\}$. Define the observed and structural deformation sets as

$$\mathcal{O}_s(\mathbf{x}, B) = \{k : c_{s,k}^B \neq c_{s,k}\}, \qquad \mathcal{D}_s(B) = \{k : I_{s,k} \cap B \neq \emptyset\}. \tag{5}$$

The observed set depends on the signal values, whereas the structural set identifies all coefficients that can be affected by the mask. Its cardinality $L_s(B) = |\mathcal{D}_s(B)|$ counts the potentially affected coefficients.

Proposition 1 (Structural deformation under time-domain masking)**.** *For every stream s and coefficient index k,*

$$c_{s,k}^B - c_{s,k} = -\sum_{t \in I_{s,k} \cap B} x_t \eta_{s,k}(t). \tag{6}$$

Consequently, $\mathcal{O}_s(\mathbf{x}, B) \subseteq \mathcal{D}_s(B)$, and

$$k \notin \mathcal{D}_s(B) \quad \Longleftrightarrow \quad c^B_{s,k} = c_{s,k} \quad \textit{for every } \mathbf{x} \in \mathbb{R}^T. \tag{7}$$

If $I_{s,k} \subseteq B$, then $c^B_{s,k} = 0$.

Proof. Linearity gives $c^B_{s,k} - c_{s,k} = \langle (M_B - I)\mathbf{x}, \eta_{s,k} \rangle$, which yields Eq. 6. If $I_{s,k} \cap B = \emptyset$, this difference vanishes for every input. Conversely, if the intersection is nonempty, choosing $x_t = \eta_{s,k}(t)$ on B and $x_t = 0$ elsewhere gives a difference of $-\sum_{t \in B} \eta_{s,k}(t)^2 < 0$. Thus, $\mathcal{D}_s(B)$ is exactly the set of potentially affected coefficients. The final statement follows because every contributing sample is masked. □

Scale-dependent support expansion. Let m_ψ be an upper bound on the lengths of the low-pass and high-pass analysis filters, and define $q(0) = L$ and $q(s) = s$ for $s \geq 1$. Before finite-window boundary handling, a level-j analysis filter is supported within an interval of at most

$$S_j = 1 + (m_\psi - 1)(2^j - 1) \tag{8}$$

input samples. Indeed, cascading the filters gives the support-span bound $S_1 = m_\psi$ and $S_j = S_{j-1} + (m_\psi - 1)2^{j-1}$ (Vetterli & Kovacevic, 1995).

For interior coefficients, choose a nominal center $\xi_{s,k}$ in this supporting interval and set $\tau_s = S_{q(s)} - 1$. Under periodization with $2^L \mid T$, the same construction applies on the periodic time axis, using circular distance $\mathrm{dist}_T(t, u) = \min_{n \in \mathbb{Z}} |t - u + nT|$. In the interior case, all coefficient-index sets in the following construction and inclusion are restricted to interior coefficients; under periodization, they include all coefficients. Use the corresponding distance and let $\mathrm{dist}(\xi, B) = \inf_{t \in B} \mathrm{dist}(\xi, t)$, with distance to the empty set defined as $+\infty$. Then

$$I_{s,k} \subseteq \{t : \mathrm{dist}(t, \xi_{s,k}) \leq \tau_s\}, \qquad \mathcal{U}_s(B) = \{k : \mathrm{dist}(\xi_{s,k}, B) \leq \tau_s\}. \tag{9}$$

If $I_{s,k}$ intersects B, any point in the intersection verifies that $k \in \mathcal{U}_s(B)$. Hence,

$$\mathcal{O}_s(\mathbf{x}, B) \subseteq \mathcal{D}_s(B) \subseteq \mathcal{U}_s(B). \tag{10}$$

For other boundary rules, $\mathcal{D}_s(B)$ remains exact, while the upper set requires a support bound for the actual finite-dimensional operator. An interior Euclidean-distance bound cannot be applied unchanged to supports that wrap around the window.

For Symlet-4, $m_\psi = 8$, so

$$S_{q(s)} = 1 + 7(2^{q(s)} - 1), \qquad \tau_s = 7(2^{q(s)} - 1). \tag{11}$$

In these settings, the uniform radius $\tau = \max_s \tau_s = 7(2^L - 1)$ bounds the support radii of all streams. The support bound grows with wavelet level in original-time coordinates. This does not imply that the number of affected coefficient indices increases with level, because coarser streams are also more strongly downsampled. Figure 14 illustrates the distinction between observed deformation, structural support overlap, and a conservative support bound.

Preservation under coefficient-domain masking. In the actual pipeline,

$$\mathcal{C} = \mathcal{W}_L(\mathbf{x}), \qquad \tilde{\mathcal{C}} = \mathrm{SCNorm}(\mathcal{C}), \qquad \tilde{\mathcal{C}}_\Omega = \mathcal{M}_\Omega(\tilde{\mathcal{C}}), \tag{12}$$

where Ω_s contains the masked patch indices in stream s. Let P_{Ω_s} replace only these patches. For every coefficient k in an unmasked patch,

$$(P_{\Omega_s} \tilde{\mathbf{c}}_s)_k = \tilde{c}_{s,k}. \tag{13}$$

By construction, with the DWT and SCNorm statistics fixed before masking, MDM introduces no additional coefficient deformation outside its selected patches. This is a preservation statement in coefficient space; it does not assert that visible wavelet coefficients depend only on unmasked time-domain samples. Similarly, time-aligned patching aligns nominal temporal intervals, not identical analysis supports across scales. The argument applies to coefficient-domain masking generally and does not establish an advantage of the directional sampling rule over other coefficient masks.

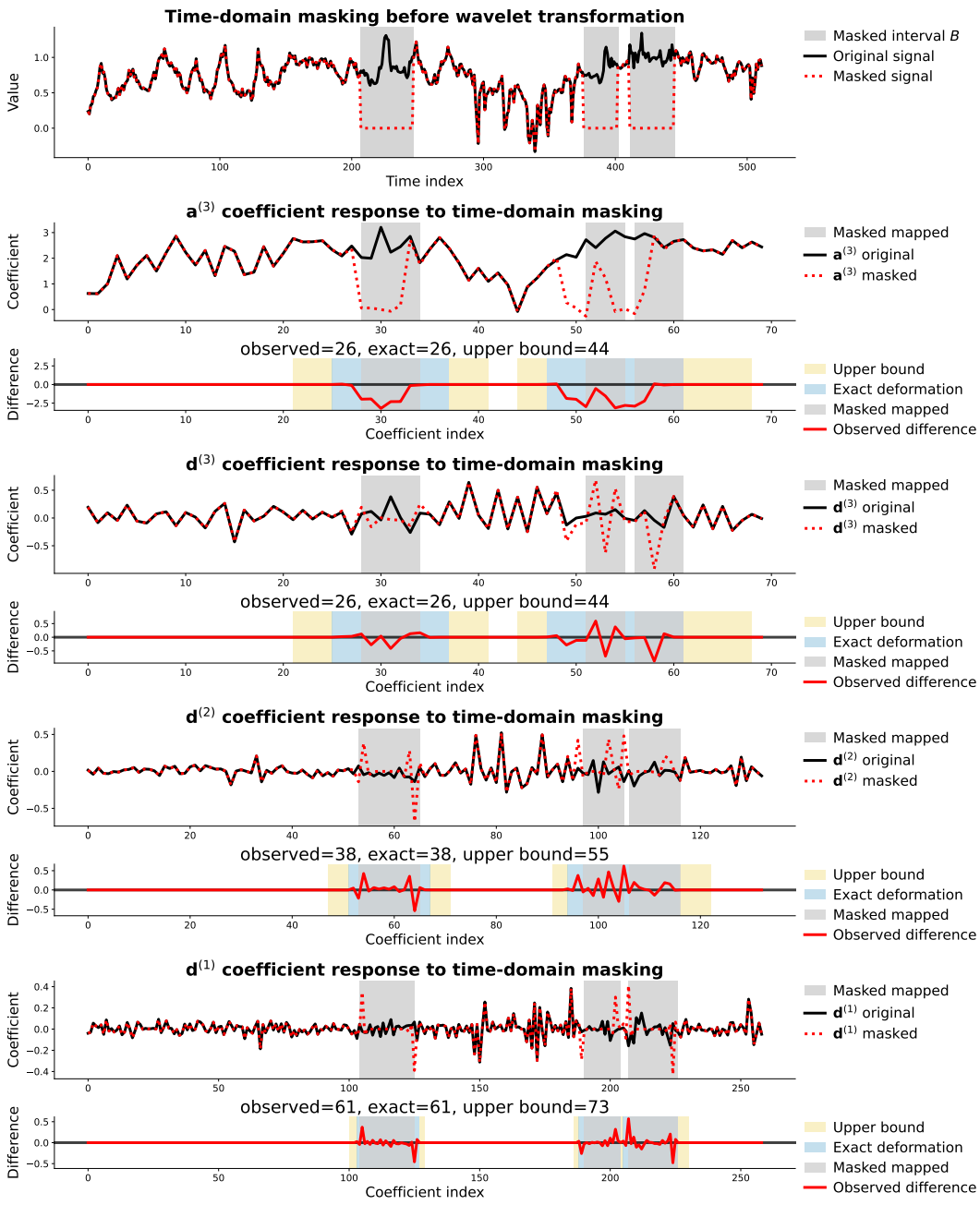


Figure 14: Boundary effect induced by the counterfactual mask-then-decompose pipeline. The top panel shows a time-domain mask B, and the coefficient panels compare $\mathcal{W}_L(\mathbf{x})$ with $\mathcal{W}_L(M_B\mathbf{x})$. The difference panels distinguish the observed deformation set $\mathcal{O}_s(\mathbf{x}, B)$, the structural deformation set $\mathcal{D}_s(B)$, and the compact-support upper set $\mathcal{U}_s(B)$. Support overlap can change coefficients outside the nominal masked interval, motivating masking directly in coefficient space.

G.2 Stability of the Scale Energy Distribution under Coefficient Perturbation

For a nonzero coefficient collection $\mathcal{C}$, define

$$E = \sum_{s\in\mathcal{S}} \|\mathbf{c}_s\|_2^2, \qquad r_s(\mathcal{C}) = \frac{\|\mathbf{c}_s\|_2^2}{E}. \tag{14}$$

The resulting scale energy distribution $\mathbf{r}(\mathcal{C})$ belongs to the probability simplex $\Delta^L = \{\mathbf{q} \in \mathbb{R}^{L+1} : q_s \geq 0,\ \sum_s q_s = 1\}$. For an orthonormal finite-dimensional DWT, Parseval's identity gives $E = \|\mathbf{x}\|_2^2$, so r_s is the fraction of input-window energy assigned to stream s (Mallat, 1999; Daubechies, 1992). This identity requires orthonormality of the finite transform, not merely an orthogonal wavelet family. A periodized orthonormal DWT with $2^L \mid T$ satisfies this condition. Zero-energy windows require a separately specified target convention and are excluded from the analysis below.

Let $\hat{\mathcal{C}}$ denote reconstructed coefficients after iSCNorm. We write $\mathbf{r}^{\mathrm{rec}} = \mathbf{r}(\hat{\mathcal{C}})$ for their induced SED, distinguishing it from the separately predicted SED-head output $\hat{\mathbf{r}}$.

Theorem 1 (Stability of the coefficient-induced scale energy distribution)**.** *Let $\mathcal{C} = \{\mathbf{c}_s\}_{s\in\mathcal{S}}$ and $\hat{\mathcal{C}} = \{\hat{\mathbf{c}}_s\}_{s\in\mathcal{S}}$ have matching stream dimensions, and define*

$$\varepsilon = \left(\sum_{s\in\mathcal{S}} \|\hat{\mathbf{c}}_s - \mathbf{c}_s\|_2^2\right)^{1/2}. \tag{15}$$

If $E > 0$ and $\varepsilon < \sqrt{E}$, then $\hat{\mathcal{C}} \neq 0$ and

$$\|\mathbf{r}^{\mathrm{rec}} - \mathbf{r}\|_1 \leq \frac{2\varepsilon}{\sqrt{E}}. \tag{16}$$

Proof. Stack the streams into vectors $\mathbf{c}$ and $\hat{\mathbf{c}}$. The reverse triangle inequality gives $\|\hat{\mathbf{c}}\|_2 \geq \sqrt{E} - \varepsilon > 0$. Set $\mathbf{u} = \mathbf{c}/\sqrt{E}$ and $\mathbf{v} = \hat{\mathbf{c}}/\|\hat{\mathbf{c}}\|_2$, with corresponding stream blocks $\mathbf{u}_s$ and $\mathbf{v}_s$. Writing $\alpha = \langle \mathbf{u}, \mathbf{v}\rangle$, Cauchy–Schwarz gives

$$\begin{aligned}\|\mathbf{r}^{\mathrm{rec}} - \mathbf{r}\|_1 &= \sum_s |\langle \mathbf{v}_s - \mathbf{u}_s, \mathbf{v}_s + \mathbf{u}_s\rangle| \\ &\leq \|\mathbf{v} - \mathbf{u}\|_2 \, \|\mathbf{v} + \mathbf{u}\|_2 = 2\sqrt{1-\alpha^2}.\end{aligned} \tag{17}$$

Since $\hat{\mathbf{c}}$ lies in the span of $\mathbf{v}$, orthogonal projection onto that span yields

$$E(1-\alpha^2) = \min_{\beta\in\mathbb{R}} \|\mathbf{c} - \beta\mathbf{v}\|_2^2 \leq \|\mathbf{c} - \hat{\mathbf{c}}\|_2^2 = \varepsilon^2. \tag{18}$$

Combining the two inequalities proves the result. □

The bound depends on relative coefficient error and does not require positive energy in every stream. It holds for any partition into coefficient streams; orthonormality additionally identifies ε with $\|\hat{\mathbf{x}} - \mathbf{x}\|_2$, where $\hat{\mathbf{x}} = \mathcal{W}_L^{-1}(\hat{\mathcal{C}})$. Thus, small original-scale reconstruction error relative to signal energy controls the induced SED error. The perturbation must be measured after iSCNorm, since rescaling can amplify normalized reconstruction errors. This result establishes stability of the coefficient-derived energy target, not an error bound for the separate prediction $\hat{\mathbf{r}}$.

G.3 Population-level Targets of the Pre-training Objectives

We finally examine how SCNorm and the predicted SED weights affect the reconstruction and calibration targets. The analysis uses the losses in Section 4, including the dependence of $\mathcal{L}_{\mathrm{COEF}}$ on $\hat{\mathbf{r}}$. We consider unrestricted measurable predictors of $U = (\tilde{\mathcal{C}}_\Omega, \Omega)$, abstracting the capacity constraints of the patch embeddings and shared encoder. The stored SCNorm statistics are used for iSCNorm but are not added to the input U of the learned predictors.

Write the SCNorm inverse for stream s as

$$\begin{aligned}\mathbf{c}_s &= a_s\tilde{\mathbf{c}}_s + b_s\mathbf{1}, \\ (a_0, b_0) &= (\sigma_0 + \epsilon, \mu_0), \qquad (a_s, b_s) = (\mathrm{RMS}(\mathbf{c}_s) + \epsilon, 0) \quad (s \geq 1),\end{aligned} \tag{19}$$

where $\epsilon > 0$ is the SCNorm stabilizer and $\mathbf{1}$ is the all-ones vector of the appropriate stream length. The statistics $\Gamma = \{a_s, b_s : s \in \mathcal{S}\}$ are computed before masking and reused to rescale the predictions. They may contain information not determined by U.

Assume the finite-dimensional DWT is orthonormal and the patches partition every stream, with $2^L \mid p$ and $p \mid T$ in the dyadic setting of the methodology. Let $\hat{\mathbf{p}}_{s,i}(U)$ predict the normalized target patch $\mathbf{p}_{s,i}$, and let $m_{s,i}$ indicate whether that patch is masked. Because the same SCNorm statistics rescale the prediction and target, their offsets cancel. Parseval's identity then gives the exact loss representation

$$\mathcal{L}_{\mathrm{TIME}} = \sum_{s\in\mathcal{S}} a_s^2 \sum_{i=1}^{N} \|\hat{\mathbf{p}}_{s,i}(U) - \mathbf{p}_{s,i}\|_2^2,$$

$$\mathcal{L}_{\mathrm{TIME}} + \mathcal{L}_{\mathrm{COEF}} = \sum_{s\in\mathcal{S}} \sum_{i=1}^{N} \big(a_s^2 + \hat{r}_s(U) m_{s,i}\big) \|\hat{\mathbf{p}}_{s,i}(U) - \mathbf{p}_{s,i}\|_2^2. \tag{20}$$

Thus, time-domain reconstruction provides full-stream supervision in original coefficient units, while coefficient reconstruction adds masked-patch supervision in normalized units. In the orthonormal setting, the time-domain squared loss does not introduce cross-stream error terms through iDWT.

Theorem 2 (Conditional targets of reconstruction and SED calibration)**.** *Assume $0 < \|\mathbf{x}\|_2$ almost surely and $\mathbb{E}\|\mathbf{x}\|_2^2 < \infty$. Under the normalization, orthonormality, and patching assumptions above, consider the population risk $\mathbb{E}[\mathcal{L}_{\mathrm{TIME}} + \mathcal{L}_{\mathrm{COEF}} + \mathcal{L}_{\mathrm{SED}}]$ over measurable patch predictors satisfying $\mathbb{E}[\mathcal{L}_{\mathrm{TIME}}] < \infty$ and measurable SED predictors taking values in Δ^L. The following statements characterize conditional minimizers with the other output held fixed, almost surely.*

For a fixed SED predictor $\hat{\mathbf{r}}(U) \in \Delta^L$, the unique minimizer of the conditional reconstruction risk is

$$\hat{\mathbf{p}}^{\star}_{s,i}(U; \hat{\mathbf{r}}) = \frac{\mathbb{E}[a_s^2 \mathbf{p}_{s,i} \mid U] + \hat{r}_s(U) m_{s,i}\, \mathbb{E}[\mathbf{p}_{s,i} \mid U]}{\mathbb{E}[a_s^2 \mid U] + \hat{r}_s(U) m_{s,i}}. \tag{21}$$

For a fixed reconstruction predictor, define

$$\eta_s(U) = \mathbb{E}[r_s(\mathcal{C}) \mid U], \qquad e_s(U) = \mathbb{E}\left[\sum_{i=1}^{N} m_{s,i} \|\hat{\mathbf{p}}_{s,i}(U) - \mathbf{p}_{s,i}\|_2^2 \,\middle|\, U\right]. \tag{22}$$

The SED calibration loss alone is minimized by $\hat{\mathbf{r}}(U) = \boldsymbol{\eta}(U)$. Under the coupled objective, the SED minimizers instead solve

$$\hat{\mathbf{r}}^{\star}(U; \hat{\mathbf{p}}) \in \operatorname*{arg\,min}_{\mathbf{q}\in\Delta^L} \left\{ D_{\mathrm{KL}}(\boldsymbol{\eta}(U) \,\|\, \mathbf{q}) + \sum_{s\in\mathcal{S}} e_s(U) q_s \right\}. \tag{23}$$

Here $\hat{\mathbf{p}}$ denotes the collection of patch predictors. If $\eta_s(U) > 0$ for every stream, this minimizer is unique and satisfies

$$\hat{r}^{\star}_s(U; \hat{\mathbf{p}}) = \frac{\eta_s(U)}{e_s(U) + \nu(U)}, \qquad \sum_{s\in\mathcal{S}} \frac{\eta_s(U)}{e_s(U) + \nu(U)} = 1, \tag{24}$$

where $\nu(U) > -\min_s e_s(U)$ is the unique scalar satisfying the normalization equation. We use the usual extended-value KL convention: a positive target mass assigned zero predicted probability has infinite loss, and terms with zero target mass contribute zero.

Proof. For fixed $\hat{\mathbf{r}}$, condition Eq. 20 on U. The mask indicators and predicted weights are determined by U, whereas a_s need not be. Each patch contributes a quadratic whose coefficient is $\mathbb{E}[a_s^2 \mid U] + \hat{r}_s(U) m_{s,i} \geq \epsilon^2 > 0$. Setting its derivative to zero gives Eq. 21. The required moments are finite because SCNorm bounds the normalized stream norms, and the finite second moment of $\mathbf{x}$ implies $\mathbb{E}[a_s^2] < \infty$. The minimizing patch predictions are also bounded and have finite reconstruction risk. Moreover, since $a_s \geq \epsilon$, the assumed finite time-domain risk gives

$$\mathbb{E}\left[\sum_{s\in\mathcal{S}} \sum_{i=1}^{N} \|\hat{\mathbf{p}}_{s,i}(U) - \mathbf{p}_{s,i}\|_2^2\right] \leq \epsilon^{-2} \mathbb{E}[\mathcal{L}_{\mathrm{TIME}}] < \infty.$$

Thus every $e_s(U)$ is finite almost surely.

For a fixed reconstruction predictor, nonnegativity implies that $r_s = 0$ conditionally almost surely whenever $\eta_s(U) = 0$. With the stated zero-mass convention, the forward KL identity gives

$$\mathbb{E}[D_{\mathrm{KL}}(\mathbf{r} \,\|\, \hat{\mathbf{r}}) \,|\, U] = \mathbb{E}[D_{\mathrm{KL}}(\mathbf{r} \,\|\, \boldsymbol{\eta}) \,|\, U] + D_{\mathrm{KL}}(\boldsymbol{\eta} \,\|\, \hat{\mathbf{r}}). \tag{25}$$

The first term is independent of $\hat{\mathbf{r}}$, so the calibration loss alone is minimized by $\boldsymbol{\eta}$, consistent with the logarithmic scoring rule (Gneiting & Raftery, 2007). The time-domain loss is also constant with respect to this fixed-output variation, while the conditional coefficient loss is $\sum_s e_s(U)\hat{r}_s(U)$. This proves Eq. 23. The objective is lower semicontinuous on the compact simplex and finite at the uniform distribution, so a minimizer exists.

If all η_s are positive, the objective in Eq. 23 is strictly convex and diverges on the simplex boundary. Its interior first-order conditions are $-\eta_s/\hat{r}_s + e_s + \nu = 0$, yielding Eq. 24. The sum in that equation decreases continuously from $+\infty$ to zero as ν increases from $-\min_s e_s$ to $+\infty$, establishing existence and uniqueness. □

Equation 21 reduces to $\mathbb{E}[\mathbf{p}_{s,i} \mid U]$ when the normalization scales are determined by U; in general, original-scale reconstruction induces a weighted conditional mean. Likewise, the KL term supervises the conditional energy distribution, but using $\hat{\mathbf{r}}$ as non-detached reconstruction weights also makes its conditional optimum depend on reconstruction error. With all $\eta_s > 0$, the coupled SED optimum equals $\boldsymbol{\eta}$ exactly when the e_s are equal across streams. These are conditional minimizers with the other output held fixed, not a claim that the shared neural parameterization attains them jointly. When some $\eta_s = 0$, Eq. 23 remains valid on the closed simplex. Such zero entries do not by themselves force a boundary optimizer. If a minimizer lies on the simplex boundary, its objective value can be approached by interior probability vectors, but the minimizer cannot be represented by finite Softmax logits.

G.4 Energy-weighted Decomposition of Time-domain Reconstruction

We characterize how a time-domain reconstruction objective distributes reconstruction error across temporal scales under an orthonormal wavelet decomposition.

Theorem 3 (Energy-weighted decomposition of time-domain reconstruction)**.** *Let $\mathbf{x}, \hat{\mathbf{x}} \in \mathbb{R}^T$ denote the original and reconstructed time series, respectively, and let $\mathcal{W}_L$ be an L-level orthonormal DWT with energy-preserving boundary handling. Denote their wavelet coefficients by*

$$\mathcal{C} = \mathcal{W}_L(\mathbf{x}) = \{\mathbf{c}_s\}_{s\in\mathcal{S}}, \qquad \hat{\mathcal{C}} = \mathcal{W}_L(\hat{\mathbf{x}}) = \{\hat{\mathbf{c}}_s\}_{s\in\mathcal{S}},$$

where $\mathcal{S}$ contains the approximation and detail scales.

For each scale s with nonzero target energy, define $E_s = \|\mathbf{c}_s\|_2^2$, and the target-energy-normalized coefficients

$$\bar{\mathbf{c}}_s = \frac{\mathbf{c}_s}{\sqrt{E_s}}, \qquad \bar{\hat{\mathbf{c}}}_s = \frac{\hat{\mathbf{c}}_s}{\sqrt{E_s}}.$$

Then the time-domain reconstruction loss $\mathcal{L}_{\mathrm{TIME}} = \|\hat{\mathbf{x}} - \mathbf{x}\|_2^2$ can be exactly decomposed as

$$\boxed{\mathcal{L}_{\mathrm{TIME}} = \sum_{s\in\mathcal{S}} E_s \left\|\bar{\hat{\mathbf{c}}}_s - \bar{\mathbf{c}}_s\right\|_2^2}.$$

Therefore, for the same target-energy-normalized reconstruction error, the contribution of each scale to the time-domain objective is proportional to the signal energy carried by that scale.

Proof. Since $\mathcal{W}_L$ is orthonormal, it preserves the Euclidean norm. By Parseval's identity, $\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 = \|\mathcal{W}_L(\hat{\mathbf{x}} - \mathbf{x})\|_2^2$. Using the linearity of the DWT, $\mathcal{W}_L(\hat{\mathbf{x}} - \mathbf{x}) = \mathcal{W}_L(\hat{\mathbf{x}}) - \mathcal{W}_L(\mathbf{x}) = \hat{\mathcal{C}} - \mathcal{C}$. Because the complete wavelet coefficient vector is partitioned into the scale-specific streams,

$$\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 = \sum_{s\in\mathcal{S}} \|\hat{\mathbf{c}}_s - \mathbf{c}_s\|_2^2.$$

From the definition of the target-energy-normalized coefficients,

$$\hat{\mathbf{c}}_s = \sqrt{E_s}\,\bar{\hat{\mathbf{c}}}_s, \qquad \mathbf{c}_s = \sqrt{E_s}\,\bar{\mathbf{c}}_s.$$

Hence,

$$\begin{aligned}\|\hat{\mathbf{c}}_s - \mathbf{c}_s\|_2^2 &= \left\| \sqrt{E_s} \left(\bar{\hat{\mathbf{c}}}_s - \bar{\mathbf{c}}_s \right) \right\|_2^2 \\ &= E_s \left\| \bar{\hat{\mathbf{c}}}_s - \bar{\mathbf{c}}_s \right\|_2^2 .\end{aligned}$$

Substituting this expression into the scale-wise decomposition gives

$$\mathcal{L}_{\mathrm{TIME}} = \sum_{s \in \mathcal{S}} E_s \left\| \bar{\hat{\mathbf{c}}}_s - \bar{\mathbf{c}}_s \right\|_2^2 ,$$

which proves the result. □

Implication. Theorem 3 shows that time-domain reconstruction does not impose equal reconstruction pressure across scales. The same normalized scale-wise error incurs a larger penalty at high-energy scales and a smaller penalty at low-energy scales. Consequently, low-energy scale-specific structures can be weakly penalized by the overall time-domain objective even when their relative reconstruction error is substantial.

H Full Results on Downstream Tasks

H.1 Full Results on Long-term Forecasting

Table 26 reports the full long-term forecasting results across all datasets and horizons. TSComp performs strongly overall, with TSComp-base achieving the best average MSE and MAE on ETTh1, ETTh2, and Exchange and the highest overall Top-2 count. Both variants also frequently rank among the top-performing models across individual horizons. On Weather, Electricity, and Traffic, where other baselines achieve the best results, TSComp remains competitive.

H.2 Full Results on Imputation

Table 27 reports the full imputation results across all datasets and mask ratios. TSComp performs consistently well across varying mask ratios, with particularly strong results on ETTh1, ETTh2, ETTm1, ETTm2, Exchange, and Traffic. TSComp-small achieves the highest overall Top-2 count and often matches or outperforms TSComp-base, indicating that increased model capacity does not consistently improve imputation. On Weather and Electricity, TSPulse and MOMENT remain competitive or superior, while TSComp retains strong overall performance.

H.3 Full Results on Anomaly Detection

Table 28 reports the full anomaly detection results across 245 UCR anomaly datasets using four range-aware metrics. Overall, TSComp variants remain consistently competitive across a wide variety of anomaly types and frequently achieve the best or second-best results in rAUCROC, VUSROC, rAUCPR, and VUSPR, indicating that the proposed scale-residual anomaly scoring provides informative residual-based anomaly evidence. While the strongest baseline varies across datasets, the full results consistently support that explicitly modeling reconstruction discrepancies across multiple temporal scales improves range-aware anomaly detection by capturing both fine-scale irregularities and their contribution to the observed time series.

H.4 Full Results on Classification

Table 29 reports the full classification results on 27 UEA benchmark datasets. TSComp-small achieves the best mean accuracy, while TSComp-base remains competitive across the diverse datasets. Although the most discriminative temporal patterns vary substantially across datasets, TSComp preserves scale-specific representations and aggregates class evidence across scales during downstream adaptation. The strong average performance is consistent with the benefit of exploiting multi-scale representations for classification, as further supported by the ablation studies.

I LIMITATIONS

Although TSComp provides a general framework for cross-scale compositional pre-training, several limitations remain. First, the current implementation uses a fixed wavelet decomposition and scale hierarchy. Our decomposition and basis studies suggest that the preferred representation can vary across tasks, motivating adaptive or learnable decomposition strategies for cross-scale compositional pre-training. Second, TSComp currently adopts a single balanced pre-training configuration across downstream tasks. While this configuration provides consistent performance across all four tasks, the relative benefits of individual objectives and components vary by task, suggesting that task-adaptive weighting or configuration could further improve transfer.

J AI USE STATEMENT

Generative AI tools were used for manuscript editing, literature and source checking, and consistency checks across the paper. All AI-assisted outputs were reviewed by the authors, and relevant references and factual claims were verified against their original sources. The authors take full responsibility for the final content of this work, including all text, claims, and other content developed with the assistance of generative AI.

K ETHICS STATEMENT AND REPRODUCIBILITY STATEMENT

This work uses publicly available benchmark time series datasets together with an industrial process dataset previously described in prior work. The study does not involve human-subject experiments or personally identifiable information.

We provide the information required to reproduce our experiments throughout the paper and appendices. Appendix C specifies the pre-training corpus, model architecture, optimization settings, downstream datasets, fine-tuning protocols, and baseline implementations. Complete downstream results are reported in Appendix H. An anonymized implementation is available at the code repository linked in the main paper.

Table 26: Long-term forecasting results. For each dataset, MSE and MAE are reported as mean±standard deviation over three random seeds for each horizon. The Avg row reports the average across horizons. The best and second-best results for each row and metric are highlighted in bold and underlined, respectively. The Top-2 count row counts best or second-best results across all horizon and Avg rows for both MSE and MAE, including ties.

Dataset	Horizon	TSComp-small		TSComp-base		GTM		TSPulse		SynTime		MOMENT		SimpleTM		AdaWaveNet		FreTS		FedFormer		TimeMixer		TimesNet	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	0.383 ± 0.001	0.407 ± 0.002	**0.378** ± 0.001	**0.403** ± 0.001	0.390 ± 0.001	0.416 ± 0.001	0.409 ± 0.000	0.429 ± 0.000	0.468 ± 0.001	0.466 ± 0.002	0.395 ± 0.001	0.419 ± 0.000	0.436 ± 0.000	0.445 ± 0.002	0.412 ± 0.003	0.434 ± 0.002	0.397 ± 0.001	0.420 ± 0.002	0.398 ± 0.002	0.442 ± 0.002	0.408 ± 0.001	0.435 ± 0.002	0.433 ± 0.001	0.450 ± 0.001
	192	0.416 ± 0.001	0.428 ± 0.001	**0.410** ± 0.000	**0.425** ± 0.001	0.451 ± 0.022	0.457 ± 0.014	0.438 ± 0.001	0.447 ± 0.000	0.514 ± 0.007	0.494 ± 0.004	0.425 ± 0.000	0.438 ± 0.000	0.481 ± 0.001	0.471 ± 0.002	0.458 ± 0.021	0.464 ± 0.011	0.435 ± 0.001	0.447 ± 0.002	0.441 ± 0.001	0.464 ± 0.000	0.449 ± 0.014	0.462 ± 0.011	0.501 ± 0.011	0.493 ± 0.010
	336	0.430 ± 0.002	0.444 ± 0.003	**0.429** ± 0.001	**0.443** ± 0.001	0.473 ± 0.001	0.475 ± 0.001	0.455 ± 0.001	0.460 ± 0.000	0.567 ± 0.004	0.523 ± 0.002	0.454 ± 0.004	0.462 ± 0.003	0.503 ± 0.001	0.488 ± 0.000	0.487 ± 0.013	0.483 ± 0.009	0.475 ± 0.007	0.476 ± 0.007	0.477 ± 0.021	0.503 ± 0.040	0.493 ± 0.021	0.499 ± 0.007	0.493 ± 0.004	0.493 ± 0.001
	720	0.472 ± 0.005	**0.485** ± 0.002	**0.468** ± 0.000	0.486 ± 0.000	0.496 ± 0.021	0.499 ± 0.011	0.491 ± 0.001	0.495 ± 0.002	0.644 ± 0.016	0.578 ± 0.013	0.515 ± 0.001	0.502 ± 0.000	0.548 ± 0.002	0.527 ± 0.006	0.539 ± 0.028	0.521 ± 0.018	0.545 ± 0.004	0.535 ± 0.002	0.529 ± 0.004	0.511 ± 0.009	0.666 ± 0.131	0.581 ± 0.053	0.574 ± 0.007	0.535 ± 0.007
	Avg	0.426 ± 0.001	0.441 ± 0.001	**0.421** ± 0.000	**0.439** ± 0.001	0.452 ± 0.005	0.462 ± 0.004	0.448 ± 0.000	0.457 ± 0.000	0.548 ± 0.006	0.515 ± 0.001	0.447 ± 0.001	0.455 ± 0.001	0.492 ± 0.001	0.483 ± 0.002	0.474 ± 0.006	0.475 ± 0.002	0.463 ± 0.002	0.469 ± 0.002	0.461 ± 0.011	0.473 ± 0.004	0.506 ± 0.027	0.493 ± 0.010	0.502 ± 0.007	0.493 ± 0.002
ETTh2	96	0.299 ± 0.004	0.349 ± 0.003	**0.284** ± 0.000	**0.341** ± 0.001	0.328 ± 0.004	0.373 ± 0.003	0.298 ± 0.002	0.358 ± 0.000	0.334 ± 0.006	0.380 ± 0.001	0.317 ± 0.000	0.369 ± 0.000	0.302 ± 0.002	0.364 ± 0.001	0.333 ± 0.007	0.382 ± 0.001	0.324 ± 0.001	0.383 ± 0.001	0.417 ± 0.031	0.472 ± 0.014	0.385 ± 0.010	0.353 ± 0.008	0.396 ± 0.011	0.427 ± 0.010
	192	0.365 ± 0.004	0.391 ± 0.002	**0.348** ± 0.003	**0.380** ± 0.002	0.383 ± 0.003	0.406 ± 0.002	0.354 ± 0.000	0.393 ± 0.001	0.414 ± 0.002	0.432 ± 0.003	0.361 ± 0.001	0.398 ± 0.001	0.372 ± 0.002	0.405 ± 0.001	0.405 ± 0.007	0.428 ± 0.001	0.429 ± 0.026	0.446 ± 0.013	0.380 ± 0.013	0.434 ± 0.012	0.358 ± 0.011	0.400 ± 0.004	0.411 ± 0.002	0.439 ± 0.001
	336	0.375 ± 0.005	0.408 ± 0.001	0.366 ± 0.002	**0.400** ± 0.001	0.395 ± 0.002	0.422 ± 0.001	0.375 ± 0.000	0.414 ± 0.000	0.442 ± 0.007	0.458 ± 0.007	**0.363** ± 0.000	0.408 ± 0.000	0.400 ± 0.002	0.429 ± 0.001	0.429 ± 0.034	0.448 ± 0.009	0.524 ± 0.054	0.504 ± 0.037	0.389 ± 0.010	0.442 ± 0.011	0.466 ± 0.004	0.477 ± 0.004	0.459 ± 0.023	0.470 ± 0.012
	720	0.402 ± 0.008	0.438 ± 0.004	**0.397** ± 0.001	**0.432** ± 0.001	0.454 ± 0.007	0.457 ± 0.003	0.412 ± 0.001	0.447 ± 0.000	0.499 ± 0.026	0.497 ± 0.013	0.408 ± 0.002	0.444 ± 0.002	0.422 ± 0.001	0.453 ± 0.001	0.443 ± 0.004	0.463 ± 0.001	0.840 ± 0.057	0.656 ± 0.021	0.456 ± 0.007	0.493 ± 0.009	0.443 ± 0.016	0.468 ± 0.022	0.496 ± 0.022	0.498 ± 0.012
	Avg	0.360 ± 0.001	0.397 ± 0.001	**0.348** ± 0.001	**0.388** ± 0.001	0.383 ± 0.002	0.415 ± 0.001	0.360 ± 0.000	0.403 ± 0.000	0.422 ± 0.004	0.442 ± 0.003	0.361 ± 0.001	0.405 ± 0.001	0.374 ± 0.001	0.413 ± 0.001	0.402 ± 0.003	0.430 ± 0.002	0.529 ± 0.031	0.497 ± 0.009	0.410 ± 0.014	0.460 ± 0.012	0.388 ± 0.024	0.425 ± 0.019	0.441 ± 0.015	0.459 ± 0.008
ETTm1	96	0.308 ± 0.000	0.359 ± 0.002	**0.304** ± 0.002	0.358 ± 0.001	0.311 ± 0.001	0.359 ± 0.001	0.309 ± 0.000	**0.354** ± 0.000	0.374 ± 0.001	0.403 ± 0.001	0.316 ± 0.000	0.363 ± 0.000	0.317 ± 0.000	0.363 ± 0.001	0.324 ± 0.007	0.367 ± 0.001	0.317 ± 0.002	0.361 ± 0.002	0.371 ± 0.004	0.421 ± 0.005	0.317 ± 0.015	0.369 ± 0.013	0.321 ± 0.005	0.372 ± 0.004
	192	0.340 ± 0.001	0.376 ± 0.001	**0.334** ± 0.002	**0.374** ± 0.001	0.343 ± 0.001	0.378 ± 0.001	0.344 ± 0.001	**0.374** ± 0.001	0.438 ± 0.013	0.441 ± 0.009	0.350 ± 0.000	0.381 ± 0.000	0.349 ± 0.001	0.384 ± 0.000	0.356 ± 0.001	0.385 ± 0.002	0.356 ± 0.002	0.384 ± 0.001	0.412 ± 0.002	0.445 ± 0.002	0.357 ± 0.003	0.390 ± 0.003	0.385 ± 0.004	0.409 ± 0.008
	336	0.372 ± 0.001	0.394 ± 0.000	**0.370** ± 0.002	0.391 ± 0.000	**0.370** ± 0.001	0.394 ± 0.001	0.371 ± 0.000	**0.389** ± 0.000	0.448 ± 0.004	0.448 ± 0.002	0.380 ± 0.001	0.398 ± 0.001	0.380 ± 0.000	0.403 ± 0.000	0.399 ± 0.004	0.413 ± 0.001	0.383 ± 0.001	0.400 ± 0.001	0.432 ± 0.004	0.453 ± 0.001	0.386 ± 0.001	0.403 ± 0.001	0.397 ± 0.000	0.416 ± 0.007
	720	0.424 ± 0.010	0.422 ± 0.001	0.422 ± 0.000	0.418 ± 0.000	**0.400** ± 0.002	**0.416** ± 0.001	0.424 ± 0.000	0.418 ± 0.001	0.501 ± 0.003	0.479 ± 0.003	0.429 ± 0.001	0.425 ± 0.000	0.432 ± 0.001	0.434 ± 0.000	0.459 ± 0.031	0.443 ± 0.013	0.443 ± 0.008	0.440 ± 0.006	0.477 ± 0.007	0.477 ± 0.005	0.442 ± 0.001	0.434 ± 0.001	0.467 ± 0.010	0.457 ± 0.010
	Avg	0.361 ± 0.001	0.388 ± 0.001	0.357 ± 0.001	0.385 ± 0.000	**0.356** ± 0.000	0.387 ± 0.000	0.362 ± 0.001	**0.384** ± 0.000	0.440 ± 0.004	0.443 ± 0.003	0.369 ± 0.000	0.392 ± 0.000	0.370 ± 0.000	0.396 ± 0.000	0.385 ± 0.002	0.402 ± 0.002	0.375 ± 0.002	0.396 ± 0.002	0.423 ± 0.003	0.449 ± 0.002	0.376 ± 0.005	0.399 ± 0.004	0.392 ± 0.005	0.413 ± 0.004
ETTm2	96	0.170 ± 0.001	0.260 ± 0.001	0.170 ± 0.000	**0.259** ± 0.000	0.197 ± 0.002	0.276 ± 0.001	**0.169** ± 0.000	0.261 ± 0.000	0.180 ± 0.000	0.274 ± 0.001	0.180 ± 0.000	0.270 ± 0.000	0.177 ± 0.000	0.270 ± 0.000	0.183 ± 0.001	0.272 ± 0.001	0.183 ± 0.001	0.265 ± 0.000	0.272 ± 0.000	0.343 ± 0.010	0.171 ± 0.005	0.260 ± 0.004	0.195 ± 0.004	0.282 ± 0.003
	192	0.230 ± 0.001	0.299 ± 0.001	0.225 ± 0.001	**0.294** ± 0.001	0.250 ± 0.011	0.311 ± 0.003	**0.223** ± 0.000	0.297 ± 0.000	0.239 ± 0.002	0.314 ± 0.002	0.230 ± 0.000	0.303 ± 0.000	0.233 ± 0.001	0.307 ± 0.000	0.232 ± 0.001	0.304 ± 0.001	0.260 ± 0.002	0.313 ± 0.001	0.305 ± 0.002	0.361 ± 0.001	0.244 ± 0.019	0.316 ± 0.020	0.254 ± 0.009	0.325 ± 0.007
	336	0.285 ± 0.001	0.336 ± 0.001	0.280 ± 0.002	**0.329** ± 0.002	0.289 ± 0.005	0.340 ± 0.003	**0.274** ± 0.000	0.330 ± 0.000	0.298 ± 0.009	0.353 ± 0.006	0.278 ± 0.000	0.335 ± 0.000	0.281 ± 0.001	0.337 ± 0.000	0.287 ± 0.007	0.339 ± 0.001	0.316 ± 0.010	0.356 ± 0.005	0.344 ± 0.003	0.383 ± 0.001	0.284 ± 0.011	0.357 ± 0.007	0.311 ± 0.009	0.359 ± 0.008
	720	0.367 ± 0.008	0.387 ± 0.003	0.365 ± 0.002	**0.383** ± 0.001	**0.359** ± 0.002	0.388 ± 0.001	**0.369** ± 0.000	**0.385** ± 0.000	0.390 ± 0.005	0.408 ± 0.003	0.363 ± 0.001	0.389 ± 0.000	0.366 ± 0.002	0.389 ± 0.001	0.393 ± 0.012	0.407 ± 0.008	0.404 ± 0.015	0.419 ± 0.010	0.408 ± 0.000	0.422 ± 0.000	0.377 ± 0.006	0.397 ± 0.005	0.443 ± 0.003	0.429 ± 0.005
	Avg	0.263 ± 0.001	0.320 ± 0.001	0.260 ± 0.000	**0.316** ± 0.000	0.274 ± 0.004	0.329 ± 0.002	**0.256** ± 0.000	0.318 ± 0.000	0.277 ± 0.003	0.337 ± 0.003	0.263 ± 0.000	0.324 ± 0.000	0.264 ± 0.000	0.325 ± 0.000	0.274 ± 0.003	0.330 ± 0.003	0.291 ± 0.006	0.338 ± 0.004	0.332 ± 0.000	0.377 ± 0.001	0.269 ± 0.004	0.328 ± 0.004	0.301 ± 0.004	0.349 ± 0.003
Exchange	96	0.100 ± 0.001	0.225 ± 0.001	0.101 ± 0.001	0.225 ± 0.001	0.145 ± 0.004	0.279 ± 0.007	0.133 ± 0.001	0.262 ± 0.000	0.354 ± 0.212	0.351 ± 0.062	0.142 ± 0.002	0.272 ± 0.001	0.170 ± 0.002	0.300 ± 0.001	0.183 ± 0.057	0.315 ± 0.061	0.184 ± 0.021	0.317 ± 0.020	0.589 ± 0.173	0.584 ± 0.100	**0.095** ± 0.002	**0.216** ± 0.003	0.243 ± 0.004	0.377 ± 0.019
	192	0.207 ± 0.006	0.325 ± 0.001	0.211 ± 0.002	0.327 ± 0.002	0.274 ± 0.014	0.384 ± 0.012	0.238 ± 0.002	0.353 ± 0.001	0.460 ± 0.034	0.499 ± 0.014	0.250 ± 0.009	0.363 ± 0.007	0.283 ± 0.001	0.392 ± 0.000	0.274 ± 0.037	0.388 ± 0.031	0.346 ± 0.075	0.438 ± 0.047	0.795 ± 0.351	0.693 ± 0.127	**0.195** ± 0.004	**0.315** ± 0.004	0.518 ± 0.094	0.543 ± 0.045
	336	**0.382** ± 0.098	**0.447** ± 0.054	0.385 ± 0.001	0.450 ± 0.001	0.498 ± 0.025	0.521 ± 0.014	0.409 ± 0.000	0.468 ± 0.001	1.091 ± 0.190	0.752 ± 0.030	0.425 ± 0.005	0.476 ± 0.001	0.456 ± 0.002	0.503 ± 0.002	0.426 ± 0.017	0.486 ± 0.011	0.782 ± 0.132	0.662 ± 0.049	0.939 ± 0.025	0.762 ± 0.013	0.405 ± 0.045	0.455 ± 0.021	0.784 ± 0.161	0.685 ± 0.024
	720	1.108 ± 0.032	0.803 ± 0.094	**1.061** ± 0.001	**0.785** ± 0.000	1.080 ± 0.049	0.787 ± 0.012	1.198 ± 0.001	0.830 ± 0.001	1.584 ± 0.777	0.934 ± 0.332	1.190 ± 0.021	0.822 ± 0.009	1.103 ± 0.004	0.796 ± 0.012	1.458 ± 0.034	0.917 ± 0.037	1.667 ± 0.131	0.959 ± 0.025	1.464 ± 0.287	0.927 ± 0.096	1.149 ± 0.141	0.804 ± 0.047	1.438 ± 0.057	0.913 ± 0.026
	Avg	0.449 ± 0.000	0.450 ± 0.002	**0.440** ± 0.001	**0.447** ± 0.001	0.499 ± 0.022	0.493 ± 0.010	0.495 ± 0.001	0.478 ± 0.001	0.872 ± 0.228	0.634 ± 0.057	0.502 ± 0.006	0.483 ± 0.003	0.503 ± 0.089	0.498 ± 0.000	0.585 ± 0.014	0.526 ± 0.005	0.744 ± 0.031	0.594 ± 0.007	0.947 ± 0.089	0.741 ± 0.040	0.461 ± 0.039	**0.447** ± 0.017	0.746 ± 0.069	0.629 ± 0.006
Weather	96	0.162 ± 0.001	0.214 ± 0.001	0.161 ± 0.000	0.215 ± 0.000	0.168 ± 0.001	0.222 ± 0.002	0.154 ± 0.000	**0.206** ± 0.001	0.169 ± 0.006	0.225 ± 0.003	0.166 ± 0.000	0.222 ± 0.000	0.158 ± 0.000	0.213 ± 0.000	0.170 ± 0.003	0.223 ± 0.002	**0.152** ± 0.002	0.211 ± 0.003	0.281 ± 0.001	0.347 ± 0.001	0.154 ± 0.009	0.212 ± 0.013	0.167 ± 0.002	0.222 ± 0.001
	192	0.205 ± 0.001	0.251 ± 0.001	0.204 ± 0.000	0.250 ± 0.000	0.208 ± 0.000	0.257 ± 0.000	**0.197** ± 0.000	**0.246** ± 0.000	0.213 ± 0.002	0.259 ± 0.002	0.210 ± 0.000	0.258 ± 0.000	0.206 ± 0.000	0.255 ± 0.000	0.213 ± 0.000	0.259 ± 0.002	**0.197** ± 0.004	0.256 ± 0.007	0.302 ± 0.003	0.354 ± 0.000	**0.197** ± 0.002	0.248 ± 0.002	0.221 ± 0.005	0.269 ± 0.001
	336	0.252 ± 0.000	0.286 ± 0.000	0.250 ± 0.000	0.285 ± 0.000	0.259 ± 0.001	0.295 ± 0.001	0.246 ± 0.000	**0.283** ± 0.000	0.270 ± 0.009	0.305 ± 0.006	0.256 ± 0.000	0.293 ± 0.000	0.257 ± 0.002	0.296 ± 0.001	0.262 ± 0.002	0.296 ± 0.002	**0.244** ± 0.000	0.292 ± 0.000	0.322 ± 0.001	0.364 ± 0.001	0.252 ± 0.006	0.294 ± 0.009	0.275 ± 0.003	0.309 ± 0.001
	720	0.317 ± 0.001	0.333 ± 0.001	0.315 ± 0.000	**0.332** ± 0.000	0.322 ± 0.001	0.334 ± 0.002	0.317 ± 0.000	0.334 ± 0.000	0.346 ± 0.003	0.342 ± 0.000	0.322 ± 0.000	0.339 ± 0.000	0.328 ± 0.001	0.350 ± 0.001	0.324 ± 0.003	0.340 ± 0.001	**0.311** ± 0.008	0.342 ± 0.000	0.374 ± 0.002	0.397 ± 0.001	0.324 ± 0.002	0.346 ± 0.004	0.349 ± 0.022	0.360 ± 0.013
	Avg	0.234 ± 0.001	0.271 ± 0.001	0.232 ± 0.000	0.270 ± 0.000	0.239 ± 0.000	0.277 ± 0.000	0.229 ± 0.000	**0.267** ± 0.000	0.246 ± 0.005	0.283 ± 0.002	0.238 ± 0.000	0.278 ± 0.000	0.237 ± 0.000	0.278 ± 0.000	0.242 ± 0.001	0.279 ± 0.002	**0.226** ± 0.000	0.276 ± 0.002	0.320 ± 0.001	0.366 ± 0.000	0.232 ± 0.004	0.275 ± 0.007	0.253 ± 0.003	0.290 ± 0.003
Electricity	96	0.139 ± 0.000	0.235 ± 0.000	0.141 ± 0.000	0.241 ± 0.000	0.139 ± 0.000	**0.233** ± 0.000	**0.136** ± 0.000	0.236 ± 0.001	0.144 ± 0.006	0.248 ± 0.006	0.157 ± 0.000	0.261 ± 0.000	0.153 ± 0.000	0.256 ± 0.000	0.140 ± 0.001	0.240 ± 0.001	0.146 ± 0.008	0.251 ± 0.000	0.217 ± 0.001	0.328 ± 0.006	0.149 ± 0.002	0.256 ± 0.002	0.198 ± 0.001	0.304 ± 0.002
	192	**0.152** ± 0.000	**0.246** ± 0.001	0.154 ± 0.000	0.252 ± 0.000	0.154 ± 0.000	0.248 ± 0.000	0.153 ± 0.000	0.252 ± 0.001	0.159 ± 0.001	0.261 ± 0.002	0.170 ± 0.000	0.272 ± 0.000	0.173 ± 0.000	0.275 ± 0.000	0.159 ± 0.000	0.258 ± 0.001	0.160 ± 0.002	0.262 ± 0.001	0.224 ± 0.001	0.332 ± 0.001	0.165 ± 0.001	0.270 ± 0.001	0.206 ± 0.000	0.311 ± 0.007
	336	0.167 ± 0.000	**0.262** ± 0.000	0.170 ± 0.000	0.268 ± 0.000	0.169 ± 0.000	0.263 ± 0.000	**0.166** ± 0.000	0.267 ± 0.001	0.179 ± 0.007	0.280 ± 0.007	0.185 ± 0.000	0.285 ± 0.000	0.191 ± 0.000	0.292 ± 0.000	0.173 ± 0.002	0.274 ± 0.002	0.176 ± 0.000	0.279 ± 0.000	0.235 ± 0.001	0.345 ± 0.004	0.197 ± 0.016	0.299 ± 0.016	0.213 ± 0.007	0.317 ± 0.006
	720	0.206 ± 0.000	0.294 ± 0.000	0.210 ± 0.000	0.300 ± 0.000	0.201 ± 0.000	0.290 ± 0.001	**0.207** ± 0.001	0.221 ± 0.010	0.312 ± 0.006	0.221 ± 0.000	0.314 ± 0.000	0.233 ± 0.000	0.326 ± 0.004	0.204 ± 0.000	0.298 ± 0.001	0.224 ± 0.008	0.325 ± 0.000	0.259 ± 0.002	0.357 ± 0.000	0.238 ± 0.014	0.330 ± 0.019	0.231 ± 0.009	0.330 ± 0.007
	Avg	0.166 ± 0.000	0.259 ± 0.000	0.168 ± 0.000	0.265 ± 0.000	0.166 ± 0.000	**0.258** ± 0.000	**0.162** ± 0.000	0.260 ± 0.000	0.176 ± 0.007	0.274 ± 0.007	0.183 ± 0.000	0.283 ± 0.000	0.187 ± 0.000	0.287 ± 0.000	0.169 ± 0.000	0.267 ± 0.000	0.177 ± 0.000	0.279 ± 0.001	0.234 ± 0.003	0.341 ± 0.005	0.187 ± 0.004	0.289 ± 0.010	0.212 ± 0.004	0.315 ± 0.003
Traffic	96	0.393 ± 0.001	0.276 ± 0.000	0.389 ± 0.000	0.274 ± 0.000	0.383 ± 0.000	**0.265** ± 0.000	0.395 ± 0.002	0.290 ± 0.002	**0.376** ± 0.002	0.278 ± 0.001	0.430 ± 0.001	0.316 ± 0.001	0.424 ± 0.000	0.313 ± 0.000	0.381 ± 0.000	0.279 ± 0.001	0.416 ± 0.000	0.310 ± 0.001	0.621 ± 0.010	0.377 ± 0.005	0.430 ± 0.003	0.314 ± 0.002	0.652 ± 0.001	0.342 ± 0.003
	192	0.403 ± 0.001	0.278 ± 0.000	0.400 ± 0.000	0.278 ± 0.000	**0.395** ± 0.000	**0.270** ± 0.000	0.408 ± 0.001	0.299 ± 0.001	**0.395** ± 0.004	0.286 ± 0.001	0.441 ± 0.001	0.320 ± 0.001	0.458 ± 0.000	0.333 ± 0.000	0.405 ± 0.001	0.291 ± 0.002	0.448 ± 0.002	0.324 ± 0.002	0.629 ± 0.009	0.378 ± 0.009	0.450 ± 0.011	0.325 ± 0.006	0.655 ± 0.012	0.348 ± 0.011
	336	0.411 ± 0.001	0.282 ± 0.000	0.410 ± 0.000	0.283 ± 0.000	**0.403** ± 0.000	**0.275** ± 0.000	0.423 ± 0.001	0.308 ± 0.002	0.406 ± 0.002	0.292 ± 0.002	0.450 ± 0.001	0.324 ± 0.001	0.475 ± 0.001	0.343 ± 0.000	0.410 ± 0.001	0.294 ± 0.001	0.469 ± 0.002	0.332 ± 0.002	0.651 ± 0.014	0.394 ± 0.006	0.458 ± 0.007	0.333 ± 0.005	0.639 ± 0.012	0.342 ± 0.003
	720	0.447 ± 0.001	0.301 ± 0.000	0.446 ± 0.000	0.303 ± 0.000	**0.434** ± 0.000	**0.292** ± 0.000	0.460 ± 0.002	0.325 ± 0.000	0.444 ± 0.001	0.310 ± 0.002	0.482 ± 0.000	0.339 ± 0.001	0.514 ± 0.001	0.361 ± 0.000	0.453 ± 0.000	0.316 ± 0.001	0.556 ± 0.000	0.388 ± 0.001	0.667 ± 0.013	0.399 ± 0.011	0.486 ± 0.001	0.345 ± 0.000	0.670 ± 0.004	0.358 ± 0.001
	Avg	0.414 ± 0.001	0.284 ± 0.000	0.411 ± 0.000	0.285 ± 0.000	**0.404** ± 0.000	**0.275** ± 0.000	0.422 ± 0.001	0.305 ± 0.000	0.405 ± 0.001	0.292 ± 0.001	0.451 ± 0.001	0.325 ± 0.001	0.468 ± 0.001	0.338 ± 0.000	0.412 ± 0.002	0.295 ± 0.001	0.472 ± 0.002	0.339 ± 0.002	0.642 ± 0.006	0.387 ± 0.007	0.456 ± 0.003	0.329 ± 0.003	0.642 ± 0.007	0.348 ± 0.002
Top-2 Count		43		53		24		31		5		3		0		1		6		0		10		0	

Table 27: Imputation results across different mask ratios on benchmark datasets. For each dataset, MSE and MAE are reported as mean±standard deviation over three random seeds under different mask ratios, and the Avg row reports the average across mask ratios. The best and second-best results for each row and metric are highlighted in bold and underlined, respectively. The Top-2 count row counts best or second-best results across all mask-ratio and Avg rows for both MSE and MAE, including ties.

Dataset	Mask ratio	TSComp-small		TSComp-base		GTM		TSPulse		SynTime		MOMENT		PatchTST		iTransformer	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	0.125	**0.150** ± 0.001	**0.249** ± 0.001	0.160 ± 0.006	0.256 ± 0.009	0.657 ± 0.008	0.553 ± 0.012	0.240 ± 0.002	0.325 ± 0.001	0.281 ± 0.003	0.368 ± 0.004	0.237 ± 0.003	0.319 ± 0.001	0.668 ± 0.006	0.548 ± 0.002	0.234 ± 0.001	0.321 ± 0.000
	0.25	**0.168** ± 0.003	**0.262** ± 0.002	0.178 ± 0.005	0.274 ± 0.004	0.451 ± 0.068	0.460 ± 0.038	0.265 ± 0.001	0.340 ± 0.001	0.314 ± 0.008	0.387 ± 0.006	0.260 ± 0.002	0.332 ± 0.001	0.670 ± 0.002	0.547 ± 0.001	0.254 ± 0.001	0.339 ± 0.000
	0.375	**0.190** ± 0.000	**0.280** ± 0.002	0.214 ± 0.019	0.297 ± 0.013	0.401 ± 0.035	0.430 ± 0.023	0.294 ± 0.001	0.359 ± 0.001	0.329 ± 0.007	0.396 ± 0.005	0.289 ± 0.001	0.348 ± 0.001	0.671 ± 0.002	0.548 ± 0.001	0.282 ± 0.001	0.360 ± 0.000
	0.5	**0.217** ± 0.000	**0.301** ± 0.000	0.233 ± 0.006	0.313 ± 0.004	0.392 ± 0.013	0.422 ± 0.008	0.315 ± 0.002	0.371 ± 0.001	0.360 ± 0.001	0.417 ± 0.002	0.331 ± 0.000	0.371 ± 0.000	0.673 ± 0.002	0.549 ± 0.000	0.313 ± 0.001	0.379 ± 0.001
	Avg	**0.181** ± 0.001	**0.273** ± 0.001	0.196 ± 0.005	0.285 ± 0.005	0.475 ± 0.030	0.466 ± 0.016	0.279 ± 0.001	0.349 ± 0.001	0.321 ± 0.003	0.392 ± 0.003	0.279 ± 0.001	0.343 ± 0.001	0.671 ± 0.001	0.548 ± 0.001	0.271 ± 0.001	0.350 ± 0.001
ETTh2	0.125	**0.079** ± 0.001	**0.177** ± 0.002	0.083 ± 0.002	0.181 ± 0.004	0.247 ± 0.009	0.347 ± 0.007	0.101 ± 0.000	0.215 ± 0.001	0.157 ± 0.002	0.283 ± 0.002	0.097 ± 0.000	0.203 ± 0.000	0.271 ± 0.012	0.355 ± 0.008	0.106 ± 0.001	0.224 ± 0.001
	0.25	**0.084** ± 0.001	**0.183** ± 0.001	0.093 ± 0.004	0.194 ± 0.004	0.236 ± 0.006	0.339 ± 0.006	0.109 ± 0.001	0.223 ± 0.001	0.194 ± 0.005	0.313 ± 0.004	0.105 ± 0.001	0.211 ± 0.000	0.267 ± 0.017	0.352 ± 0.012	0.135 ± 0.002	0.256 ± 0.002
	0.375	**0.095** ± 0.001	**0.196** ± 0.001	0.103 ± 0.003	0.203 ± 0.004	0.235 ± 0.007	0.338 ± 0.007	0.119 ± 0.000	0.233 ± 0.001	0.291 ± 0.021	0.380 ± 0.014	0.116 ± 0.000	0.222 ± 0.000	0.268 ± 0.016	0.353 ± 0.011	0.182 ± 0.002	0.301 ± 0.001
	0.5	**0.111** ± 0.000	**0.213** ± 0.000	0.121 ± 0.005	0.223 ± 0.005	0.244 ± 0.010	0.345 ± 0.007	0.131 ± 0.001	0.244 ± 0.000	0.438 ± 0.024	0.461 ± 0.012	0.132 ± 0.000	0.237 ± 0.000	0.271 ± 0.013	0.355 ± 0.009	0.282 ± 0.004	0.368 ± 0.002
	Avg	**0.092** ± 0.001	**0.192** ± 0.001	0.100 ± 0.001	0.200 ± 0.001	0.241 ± 0.005	0.342 ± 0.004	0.115 ± 0.000	0.229 ± 0.001	0.270 ± 0.005	0.359 ± 0.003	0.112 ± 0.000	0.218 ± 0.000	0.269 ± 0.015	0.354 ± 0.010	0.177 ± 0.000	0.287 ± 0.000
ETTm1	0.125	**0.061** ± 0.000	**0.158** ± 0.000	0.070 ± 0.002	0.171 ± 0.004	0.102 ± 0.001	0.207 ± 0.002	0.075 ± 0.000	0.177 ± 0.001	0.104 ± 0.003	0.219 ± 0.002	0.079 ± 0.001	0.176 ± 0.001	0.659 ± 0.008	0.531

Table 28: Per-dataset anomaly detection results on UCR anomaly datasets. Results are averaged over three random seeds and within each grouped dataset, and are reported using four range-aware metrics: rAUCROC, rAUCPR, VUSROC, and VUSPR. Higher values indicate better performance. The best and second-best results for each metric are highlighted in bold and underlined, respectively. The Top-2 count row counts best or second-best results across the four metrics for all datasets, including ties.

Dataset	TSComp-small				TSComp-base				GTM				TSPulse				SymTime				MOMENT				PatchTST				TSLANet				Anom. Trans.				Nonstat. Trans.			
	rAUCROC	rAUCPR	VUSROC	VUSPR	rAUCROC	rAUCPR	VUSROC	VUSPR	rAUCROC	rAUCPR	VUSROC	VUSPR	rAUCROC	rAUCPR	VUSROC	VUSPR	rAUCROC	rAUCPR	VUSROC	VUSPR	rAUCROC	rAUCPR	VUSROC	VUSPR	rAUCROC	rAUCPR	VUSROC	VUSPR	rAUCROC	rAUCPR	VUSROC	VUSPR	rAUCROC	rAUCPR	VUSROC	VUSPR	rAUCROC	rAUCPR	VUSROC	VUSPR
epmaocg	**0.882**	0.014	**0.890**	0.015	0.815	0.015	0.826	0.015	0.865	0.010	0.870	0.011	0.848	0.015	0.847	0.016	0.995	0.012	0.993	0.013	0.573	0.008	0.575	0.008	0.810	0.016	0.841	0.016	0.523	0.018	0.518	0.019	0.958	**0.019**	0.957	**0.020**
BIDMC1	0.648	0.064	0.630	0.062	0.599	0.058	0.577	0.057	**0.814**	**0.100**	**0.796**	**0.096**	0.479	0.055	0.661	0.053	0.653	0.063	0.636	0.061	0.650	0.080	0.629	0.078	0.639	0.068	0.622	0.067	0.634	0.048	0.619	0.047	0.657	0.095	0.642	0.091	0.675	0.084	0.658	0.083
CHARISfive	0.581	0.004	0.563	0.004	**0.721**	0.013	**0.718**	0.012	0.597	0.004	0.556	0.004	0.589	0.005	0.567	0.006	0.636	0.006	0.618	0.006	0.590	0.007	0.572	0.007	0.610	0.007	0.582	0.007	0.628	0.006	0.612	0.006	**0.664**	**0.035**	0.629	**0.038**	0.607	0.006	0.578	0.006
CHARISfive	0.524	0.007	0.504	0.006	0.570	0.005	0.559	0.005	0.610	0.005	0.586	0.005	0.587	**0.011**	0.584	**0.009**	**0.654**	0.005	**0.634**	0.005	0.631	0.007	0.613	0.007	0.609	0.005	0.592	0.005	0.554	0.004	0.541	0.004	0.590	0.005	0.566	0.004	0.610	0.005	0.589	0.005
CIMIS44AirTemperature	0.699	0.033	0.684	0.035	0.697	**0.070**	0.687	**0.071**	**0.825**	0.046	**0.818**	0.046	0.649	0.030	0.643	0.030	0.696	0.036	0.686	0.036	0.659	0.033	0.654	0.033	0.671	0.034	0.658	0.033	0.619	0.019	0.624	0.018	0.663	0.032	0.660	0.032	0.654	0.021	0.629	0.021
DISTORTED1sddb40	0.639	0.030	0.624	0.029	0.660	0.029	0.643	0.029	0.652	0.029	0.635	0.029	0.694	**0.071**	0.680	**0.071**	0.706	0.049	0.693	0.048	0.722	0.056	0.725	0.056	0.705	0.039	0.693	0.039	**0.740**	0.058	0.728	0.059	0.652	0.035	0.636	0.034	0.725	0.060	0.711	0.059
DISTORTED1sddb40	0.444	0.010	0.413	0.010	0.376	0.009	0.375	0.009	0.505	0.010	0.490	0.010	0.444	0.011	0.437	0.011	0.489	0.010	0.468	0.010	**0.823**	**0.013**	0.824	**0.013**	0.437	0.009	0.414	0.009	0.578	0.014	0.564	0.013	**0.813**	**0.023**	**0.813**	**0.023**	0.553	0.011	0.542	0.011
DISTORTED1sddb40	0.620	0.014	0.620	0.014	0.638	**0.030**	0.623	**0.031**	0.673	0.015	0.653	0.015	0.670	0.020	0.678	0.020	0.668	0.019	0.681	0.018	**0.759**	0.021	**0.757**	0.021	0.691	0.017	0.687	0.017	0.652	0.015	0.640	0.015	0.636	0.015	0.627	0.015	0.633	0.015	0.624	0.015
DISTORTED1tiagesacq	0.629	0.013	0.611	0.013	0.662	0.019	0.644	0.019	**0.696**	0.012	**0.702**	0.012	0.611	0.014	0.617	0.014	0.581	0.009	0.574	0.009	0.584	0.009	0.589	0.009	0.597	**0.020**	0.577	0.020	0.603	0.009	0.594	0.009	0.532	0.010	0.518	0.019	0.610	**0.020**	0.602	**0.021**
DISTORTED1EDACC	0.652	0.071	0.624	0.067	0.630	0.060	0.613	0.058	**0.881**	0.044	**0.783**	0.089	0.791	0.064	0.683	0.061	0.671	0.078	0.652	0.075	0.641	0.084	0.618	0.081	0.644	0.068	0.630	0.068	0.646	0.055	0.633	0.055	0.669	0.095	0.655	0.095	0.661	0.090	0.640	0.092
DISTORTED1IMS44AirTemperature	0.244	0.006	0.308	0.008	0.258	0.006	0.303	0.007	0.254	0.006	0.301	0.007	0.249	0.006	0.299	0.007	0.225	0.005	0.268	0.006	0.422	0.010	0.467	0.010	0.427	0.009	0.472	0.010	0.422	0.008	0.458	0.009	0.503	0.009	0.541	0.009	0.456	0.010	0.497	0.010
DISTORTED1ECG	0.592	0.023	0.581	0.023	0.606	0.022	0.596	0.021	0.674	0.028	**0.666**	0.028	0.547	0.017	0.531	0.016	**0.679**	**0.030**	0.664	**0.029**	0.608	0.029	0.592	0.027	0.618	0.029	0.603	0.028	0.596	0.011	0.587	0.011	0.604	0.022	0.599	0.022	0.642	0.016	0.641	0.016
DISTORTED1gait1	0.722	0.022	**0.703**	0.022	0.595	0.016	0.586	0.016	0.691	0.023	0.674	0.022	0.587	0.020	0.618	0.019	0.692	0.025	0.685	0.025	0.602	0.024	0.583	0.024	0.685	0.027	0.667	0.027	0.593	0.025	0.589	0.025	0.623	0.016	0.614	0.016	0.609	0.019	0.603	0.019
DISTORTED1gait1MarkerLFM5z3	0.680	0.026	0.596	0.028	0.643	0.029	0.618	0.029	0.781	0.037	0.751	0.037	0.587	0.029	0.564	0.029	0.696	0.025	0.694	0.025	0.600	0.024	0.593	0.024	0.621	0.027	0.607	0.027	0.593	0.025	0.589	0.025	**0.670**	**0.034**	**0.657**	**0.035**	0.617	0.031	0.608	0.032
DISTORTED1insectEPG	0.576	0.011	0.557	0.011	0.558	0.011	0.541	0.011	0.725	0.015	0.705	0.015	0.576	0.011	0.565	0.011	0.725	0.015	0.707	0.015	0.566	0.010	0.564	0.010	0.566	0.010	0.556	0.010	0.553	0.011	0.553	0.011	0.614	0.015	0.596	0.015	0.572	0.014	0.564	0.014
DISTORTED1InternalBleeding	0.691	0.063	0.654	0.088	**0.738**	**0.120**	**0.732**	**0.122**	0.653	0.051	0.652	0.053	0.591	0.045	0.583	0.045	0.622	0.052	0.613	0.053	0.607	0.045	0.604	0.046	0.626	0.048	0.615	0.047	0.607	0.045	0.602	0.046	0.624	0.044	0.616	0.045	0.639	0.048	0.631	0.048
DISTORTED1sddb220sub3112USPG	0.677	0.050	0.652	0.050	0.666	0.047	0.643	0.048	0.745	0.017	0.719	0.017	0.639	0.013	0.668	0.013	0.668	0.029	0.654	0.027	0.655	0.016	0.666	0.016	0.645	0.019	0.639	0.018	0.623	0.018	0.620	0.018	0.683	0.035	0.670	0.035	0.662	0.035	0.639	0.035
DISTORTED1sddb30791A1	**0.963**	**0.088**	**0.858**	**0.083**	0.709	0.013	0.695	0.013	0.653	0.012	0.649	0.013	0.707	0.047	0.704	0.050	0.695	0.017	0.677	0.023	0.705	0.049	0.716	0.055	0.651	0.046	0.667	0.049	0.624	0.012	0.618	0.012	0.702	0.050	0.720	0.056	0.736	0.050	0.726	0.055
DISTORTED1sddb30791A2	**0.931**	0.042	**0.862**	0.054	0.764	0.019	0.764	0.018	0.723	0.018	0.658	0.018	0.665	0.043	0.732	0.028	0.722	0.002	0.737	0.023	0.709	0.018	0.716	0.023	0.655	0.008	0.693	0.008	0.690	0.007	0.732	0.008	0.754	0.008	0.754	0.008	0.687	0.015	0.682	0.129
DISTORTED1MesoplodonDensirostris	0.615	0.024	0.660	0.026	0.693	0.024	0.712	0.027	0.658	0.023	0.658	0.026	0.455	0.019	0.463	0.020	0.714	**0.063**	0.716	**0.068**	0.691	0.059	0.701	0.063	0.690	0.054	0.696	0.058	0.717	0.050	0.714	0.053	0.656	0.050	0.663	0.060	**0.745**	**0.062**	**0.734**	0.064
DISTORTED1sddb40	**0.824**	**0.075**	**0.818**	**0.074**	0.648	0.027	0.646	0.027	0.762	0.022	0.765	0.023	0.654	0.047	0.671	0.045	0.611	0.024	0.700	0.033	0.678	0.031	0.661	0.031	0.693	0.042	0.676	0.041	0.665	0.037	0.640	0.038	0.674	0.037	0.665	0.038	0.665	0.039	0.649	0.039
DISTORTED1PowerDemand	0.608	0.024	0.595	0.024	0.646	0.041	0.614	0.040	**0.719**	0.033	**0.715**	0.033	0.522	0.009	0.513	0.009	0.663	0.030	0.653	0.029	0.617	0.035	0.608	0.035	0.599	0.020	0.572	0.019	0.564	0.019	0.554	0.020	0.572	0.019	0.564	0.019	0.654	0.032	0.644	0.032
DISTORTED1sddb30791A3	0.533	0.028	0.528	0.024	0.669	0.031	0.653	0.033	0.788	0.023	0.774	0.024	0.559	0.019	0.545	0.019	0.750	0.052	0.744	0.052	0.708	0.035	0.706	0.035	0.714	0.037	0.678	0.038	0.654	0.026	0.640	0.026	0.578	0.028	0.566	0.026	0.554	0.026	0.546	0.026
DISTORTED1sddb21005V	**0.677**	**0.031**	**0.676**	**0.031**	0.570	0.022	0.582	0.023	0.519	0.018	0.518	0.019	0.672	**0.031**	0.652	0.030	0.571	0.021	0.556	0.021	0.573	0.021	0.564	0.021	0.547	0.020	0.542	0.021	0.534	0.019	0.527	0.019	0.640	0.024	0.621	0.025	0.591	0.022	0.586	0.022
DISTORTED1tiagesstation	0.392	0.006	0.385	0.006	0.488	0.008	0.481	0.008	0.501	0.008	0.501	0.008	0.411	0.007	0.423	0.007	**0.856**	**0.013**	0.843	**0.013**	0.577	0.009	0.577	0.009	0.553	0.003	0.556	0.003	0.418	0.008	0.424	0.004	0.491	0.003	0.506	0.003	0.418	0.008	0.424	0.004
DISTORTED1ts2010160ML2	**0.979**	**0.262**	**0.972**	**0.231**	0.716	0.010	0.714	0.010	0.577	0.006	0.577	0.007	0.625	0.007	0.642	0.008	0.707	0.010	0.711	0.012	0.615	0.008	0.612	0.010	0.577	0.007	0.574	0.010	0.642	0.008	0.654	0.009	0.568	0.007	0.590	0.009	0.603	0.007	0.601	0.009
DISTORTED1sddb40	**0.782**	**0.024**	**0.761**	**0.024**	**0.816**	**0.031**	**0.803**	**0.032**	0.653	0.022	0.653	0.022	0.696	0.011	0.689	0.011	0.648	0.016	0.643	0.016	0.709	0.016	0.707	0.016	0.647	0.009	0.643	0.009	0.657	0.011	0.656	0.011	0.658	0.017	0.657	0.017	0.651	0.013	0.651	0.013
DISTORTED1ts2010160ECG	0.602	0.020	0.629	0.020	0.574	0.016	0.577	0.017	0.591	0.016	0.594	0.017	**0.656**	**0.022**	**0.649**	**0.022**	0.619	0.018	0.613	0.019	0.634	0.019	0.628	0.019	0.609	0.018	0.605	0.018	0.634	0.019	0.621	0.020	0.605	0.018	0.600	0.018	0.559	0.016	0.550	0.016
DISTORTED1tiwaterAPE	0.276	0.006	0.264	0.006	0.545	0.008	0.554	0.008	0.671	0.009	0.674	0.010	0.661	0.009	0.664	0.009	0.714	0.010	0.713	0.010	**0.727**	**0.015**	**0.730**	0.014	0.728	0.015	0.731	**0.021**	0.704	0.008	0.705	0.008	0.696	0.010	0.695	0.010	0.705	0.010	0.698	0.010
DISTORTED1TkeepFifthMARS	0.650	0.025	0.616	0.024	0.808	0.021	0.779	0.019	0.718	**0.042**	0.722	**0.041**	**0.748**	0.036	0.732	0.031	0.654	0.026	0.633	0.026	0.647	0.026	0.619	0.024	0.643	0.023	0.623	0.023	0.523	0.020	0.495	0.020	0.647	0.026	0.629	0.025
DISTORTED1TkeepFirstMARS	0.284	0.004	0.298	0.005	0.479	0.005	0.488	0.005	0.284	0.004	0.298	0.005	0.482	0.005	0.492	0.005	0.588	0.003	0.583	0.004	0.524	0.003	0.543	0.003	0.525	0.003	0.537	0.003	0.493	0.003	0.513	0.003	0.554	0.004	0.568	0.004	0.554	0.004	0.568	0.004
DISTORTED1TkeepFourthMARS	0.653	0.030	0.679	0.030	0.874	0.042	0.864	0.038	**0.798**	**0.038**	**0.779**	0.042	0.748	0.029	0.758	0.029	0.641	0.016	0.626	0.016	0.633	0.018	0.628	0.018	0.671	0.029	0.651	0.028	0.620	0.024	0.609	0.024	0.662	0.028	0.643	0.028	0.636	0.027	0.616	0.026
DISTORTED1TkeepThirdMARS	0.455	0.015	0.447	0.014	0.643	0.014	0.624	0.013	0.615	0.012	0.615	0.012	0.544	0.013	0.551	0.014	0.608	0.022	0.603	0.022	0.573	0.021	0.562	0.021	0.572	0.020	0.575	0.020	0.558	0.020	0.553	0.020	0.508	0.020	0.529	0.020	0.526	0.020	0.556	0.020
DISTORTED1WalkingAcceleration	0.637	0.096	0.633	0.097	0.787	0.087	0.767	0.085	0.906	0.071	0.906	**0.074**	**0.743**	0.066	0.740	0.070	0.692	0.089	0.665	0.090	0.696	0.088	0.688	0.089	0.663	0.069	0.655	0.069	0.682	0.077	0.674	0.076	0.722	**0.130**	0.710	0.131	0.647	0.072	0.644	0.073
HCG	0.601	0.023	0.583	0.022	0.626	0.020	0.613	0.027	0.628	0.029	0.617	0.028	0.543	0.027	**0.652**	**0.033**	**0.668**	**0.032**	0.668	0.029	0.594	0.028	0.642	0.032	0.627	**0.032**	0.606	0.012	0.598	0.011	0.583	0.025	0.700	0.024	0.629	0.019	0.626	0.019	0.626	0.019
Fantasia	**0.680**	0.009	**0.694**	0.010	0.660	0.009	0.686	0.010	0.585	0.007	0.582	0.007	0.561	**0.023**	0.572	**0.024**	0.591	0.008	0.579	0.008	0.627	0.008	0.641	0.009	0.601	0.008	0.612	0.008	0.637	0.010	0.633	0.010	0.666	0.010	0.677	0.010	0.657	0.009	0.652	0.011
gait	0.657	0.022	0.643	0.022	0.671	0.016	0.671	0.016	0.649	**0.048**	0.638	0.048	0.633	0.020	0.633	0.020	0.638	0.017	0.628	0.017	0.760	0.014	0.774	0.014	0.618	0.016	0.619	0.015	0.611	0.016	0.602	0.016	0.646	0.018	0.637	0.018	0.593	0.015	0.590	0.015
gaitHeat	0.604	0.024	0.605	0.024	**0.681**	0.027	0.674	0.027	0.593	0.027	0.577	0.027	0.594	0.024	0.574	0.024	0.600	0.024	0.581	0.024	0.579	0.026	0.576	0.025	0.635	0.026	0.628	0.026	0.608	0.024	0.594	0.025	0.643	0.038	0.630	0.039	0.629	0.025	0.595	0.023
GP711MarkerLFM5z	0.686	0.041	0.687	0.042	0.685	0.041	0.682	0.042	0.594	0.058	0.738	0.055	0.593	0.035	0.716	**0.082**	0.711	**0.078**	0.669	0.030	0.599	0.021	0.622	0.031	0.617	0.031	0.638	0.036	0.629	0.035	0.593	0.033	0.592	0.033	0.646	0.019	0.644	0.019	0.640	0.019
insectEPG	0.614	0.012	0.599	0.012	0.566	0.011	0.550	0.011	0.535	0.009	0.545	0.009	0.615	0.010	0.610	0.011	0.596	0.010	0.598	0.010	0.592	0.010	0.590	0.010	0.587	0.013	0.592	0.011	**0.629**	0.010	0.622	0.010	0.575	**0.019**	0.579	**0.019**	0.575	0.012	0.574	0.011
InternalBleeding	**0.702**	**0.120**	**0.697**	**0.120**	0.690	0.104	0.688	0.104	0.647	0.038	0.645	0.038	0.628	0.053	0.619	0.055	0.669	0.046	0.668	0.047	0.608	0.045	0.603	0.045	0.623	0.048	0.613	0.048	0.615	0.057	0.612	0.059	0.635	0.048	0.621	0.047
ItalianPowerDemand	0.498	0.000	0.489	0.000	0.687	**0.003**	**0.678**	**0.003**	0.683	0.003	0.671	0.003	**0.686**	0.003	0.667	0.003	0.614	0.003	0.601	0.003	0.672	0.003	0.661	0.003	0.671	0.003	0.653	0.003	0.616	0.003	0.597	0.002	0.618	0.003	0.608	0.003	0.604	0.005	0.585	0.002
LabCTAcq10132SPG	0.655	0.002	0.651	0.002	0.597	0.006	0.602	0.005	0.716	0.018	0.728	0.017	0.618	0.003	0.639	0.003	0.656	0.027	0.644	0.026	0.671	0.044	0.661	0.045	0.654	0.042	0.642	0.043	0.654	0.018	0.652	0.018	0.683	0.054	0.682	0.054	0.654	0.047	0.647	0.048
lsudb30791A1	**0.951**	**0.407**	**0.897**	**0.360**	0.715	0.013	0.703	0.018	0.652	0.013	0.651	0.013	0.657	0.038	0.660	0.039	0.701	0.027	0.686	0.028	0.696	0.049	0.705	0.055	0.649	0.050	0.664	0.052	0.653	0.013	0.644	0.014	0.784	0.093	0.774	0.098	0.679	0.046	0.661	0.049
lsudb30791A5	**0.853**	0.073	**0.817**	0.059	0.722	0.027	0.703	0.027	0.748	0.035	0.745	0.035	0.685	0.044	0.712	0.046	0.726	0.104	0.726	0.104	0.758	**0.133**	0.770	**0.139**	0.707	0.042	0.702	0.042	0.704	0.033	0.703	0.033	0.756	0.106	0.756	0.106	0.736	0.116	0.731	0.111
MesoplodonDensirostris	0.706	0.027	0.722	0.030	0.674	0.023	0.691	0.025	0.669	0.023	0.668	0.026	0.467	0.019	0.473	0.022	0.707	0.041	0.713	**0.067**	0.671	0.058	0.687	0.063	0.706	0.055	0.714	0.060	0.698	0.043	0.693	0.045	0.676	**0.064**	0.690	0.066	**0.717**	0.064	0.713	0.066
mit40446hypertrnsg	**0.754**	**0.048**	**0.748**	**0.048**	0.617	0.027	0.625	0.027	0.720	0.041	0.719	0.041	0.625	0.022	0.625	0.022	0.704	0.031	0.694	0.031	0.662	0.025	0.654	0.025	0.654	0.042	0.642	0.043	0.654	0.018	0.652	0.018	0.688	0.028	0.672	0.028	0.654	0.027	0.647	0.027
mit41436hypertrnsg	0.656	0.008	0.641	0.008	0.574	0.004	0.573	0.004	0.569	0.004	0.570	0.004	**0.692**	**0.011**	**0.676**	**0.010**	0.656	0.007	0.637	0.007	0.642	0.008	0.618	0.008	0.602	0.006	0.586	0.006	0.617	0.006	0.600	0.006	0.681	0.008	0.674	0.008	0.656	0.007	0.646	0.007
NOISEGaitMix1	0.615	0.034	0.597	0.034	0.628	0.033	0.611	0.033	0.627	0.026	0.618	0.026	0.571	0.025	0.572	0.025	0.715	0.035	0.707	0.035	0.702	0.031	0.699	0.031	0.724	0.036	0.715	0.036	0.705	0.031	0.701	0.031	0.705	0.037	0.703	0.037	0.713	0.032	0.709	0.032
NOISEGaitMix2	0.614	0.034	0.597	0.034	0.628	0.033	0.611	0.033	0.627	0.026	0.618	0.026	0.571	0.025	0.572	0.025	0.715	0.035	0.707	0.035	0.702	0.031	0.699	0.031	0.724	0.036	0.715	0.036	0.705	0.031	0.701	0.031	0.705	0.037	0.703	0.037	0.713	0.032	0.709	0.032
NOISEgait1	0.602	0.004	0.601	0.005	0.607	0.019	0.602	0.020	**0.825**	**0.105**	**0.833**	**0.107**	0.590	0.019	0.593	0.019	0.608	0.019	0.609	0.019	0.510	0.015	0.516	0.015	0.597	0.019	0.604	0.020	0.554	0.018	0.557	0.018	0.547	0.017	0.547	0.017	0.533	0.010	0.534	0.010
NOISEgait1bout2	0.787	0.027	0.765	0.027	0.781	0.027	0.767	0.027	0.615	0.017	0.615	0.017	0.614	0.014	0.587	0.012	0.605	0.028	0.588	0.028	0.577	0.027	0.559	0.027	0.624	0.030	0.624	0.031	0.629	0.000	0.628	0.000	0.667	0.032	0.644	0.032	0.594	0.029	0.585	0.030
NOISEGP711MarkerLFM5z3	0.610	0.012	0.609	0.012	0.680	0.016	0.700	0.018	**0.739**	0.042	0.717	**0.043**	0.697	0.024	0.718	0.026	**0.752**	0.022	0.726	0.024	0.708	0.018	**0.727**	0.020	0.683	0.015	0.675	0.015	0.632	0.014	0.636	0.016	0.678	0.015	0.678	0.015	0.646	0.014	0.648	0.015
NOISEinsectEPG1	0.530	0.018	0.524	0.018	**0.674**	0.035	**0.675**	0.034	0.561	0.016	0.551	0.016	0.656	**0.027**	0.641	**0.028**	0.650	0.032	0.618	0.031	0.661	0.031	0.636	0.031	**0.674**	0.024	0.651	0.023	0.611	0.021	0.598	0.021	0.638	0.032	0.621	0.032	0.653	0.023	0.645	0.023
NOISEInternalBleeding	0.691	0.030	0.678	0.032	0.724	0.041	0.705	0.042	0.618	0.041	0.622	0.041	0.460	0.028	0.471	0.030	0.586	0.032	0.583	0.033	0.523	0.026	0.539	0.040	0.665	0.035	0.656	0.036	0.618	0.031	0.614	0.031	0.752	0.036	0.746	0.038	0.687	0.032	0.662	0.031
NOISELabCTAcq10132SPG	0.689	0.050	0.673	0.052	0.677	0.052	0.673	0.053	0.768	0.024	0.754	0.024	0.697	0.032	0.694	0.033	0.697	0.032	0.684	0.033	0.706	0.041	0.691	0.041	0.676	0.030	0.664	0.030	0.674	0.029	0.675	0.030	0.684	0.029	0.674	0.030	0.685	0.038	0.678	0.051
NOISEMesoplodonDensirostris	0.675	0.025	0.682	0.027	0.691	0.025	0.696	0.027	0.658	0.023	0.648	0.025	0.447	0.019	0.452	0.021	**0.700**	**0.039**	**0.706**	**0.044**	0.671	0.035	0.686	0.040	0.686	0.054	0.693	0.059	0.693	0.038	0.691	0.042	0.623	0.053	0.634	0.055	0.692	0.027	0.689	0.030
NOISETkeepThirdMARS	**0.693**	0.026	**0.673**	0.022	0.651	**0.031**	0.625	**0.030**	0.655	0.024	0.628	0.024	0.578	0.025	0.568	0.025	0.687	0.027	0.677	0.027	0.693	0.024	0.693	0.024	0.664	0.023	0.658	0.023	0.594	0.023	0.586	0.023	0.666	0.028	0.656	0.028	0.682	0.025	0.659	0.024
pamLtm	**0.774**	0.039	**0.773**	0.041	0.691	0.030	0.683	0.030	0.583	0.026	0.566	0.023	0.595	0.042	0.586	0.043	0.602	0.023	0.588	0.023	0.609	0.024	0.601	0.025	0.686	0.033	0.674	0.033	0.667	0.030	0.657	0.030	0.680	0.026	0.672	0.025	0.630	0.024	0.629	0.025
PowerDemand	0.628	0.025	0.615	0.025	0.653	0.025	0.633	0.025	0.688	0.023	0.688	0.023	0.664	0.014	0.664	0.014	0.641	0.014	0.641	0.014	0.503	0.021	0.519	0.021	0.514	0.019	0.558	0.019	0.603	0.023	0.603	0.023	0.573	0.024	0.573	0.024	0.603	0.021	0.603	0.021
spdb5e1036MLH	0.499	0.022	0.508	0.024	0.502	0.026	0.512	0.028	0.547	0.022	0.552	0.022	0.599	0.022	0.598	0.022	0.610	0.026