RAP: RECONSTRUCTION-AUGMENTED PREDICTORS FOR ROBUST TIME-SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term time-series forecasting (LTSF) supports applications in energy, traffic, and healthcare, yet state-of-the-art models remain brittle when inputs are corrupted by sensor or preprocessing noise. To resolve this problem, we introduce a *reconstruction-augmented predictor* (RAP) that improves robustness *without* data manipulation by coupling the standard forecasting loss with a self-reconstruction objective on the history window. A single encoder feeds two linear heads—reconstruction and prediction—so the shared representation is explicitly regularized to preserve fine-grained, multi-scale structure; analytically, the reconstruction head *implicitly* penalizes the predictor's input–output Jacobian (noise sensitivity). We derive that in a small-noise regime, regularizing the reconstruction term reduces noise sensitivity under mild spectral conditions. To instantiate RAP, we adopt a dual-axis Transformer that conditions on the history and alternates attention across temporal and channel dimensions, enabling full-horizon generation while capturing spatiotemporal interactions. Across representative LTSF benchmarks in high-dimensional datasets, RAP consistently outperforms competitive baselines in clean conditions and sustains its advantage as synthetic or real noise levels increase, with the largest gains on the most challenging settings. Ablations quantify the effect of the reconstruction head, and analysis highlights failure modes of augmentation- and filtering-based robustness strategies.

1 INTRODUCTION

Time-series data play a central role across power systems (Susto et al., 2018), finance (Sezer et al., 2020), and healthcare (Kaushik et al., 2020; Morid et al., 2023). Building on the success of deep learning in numerous areas such as image generation and classification, deep learning approaches have recently achieved strong performance on time-series tasks (Lim & Zohren, 2021; Torres et al., 2021; Lara-Benítez et al., 2021). Within this landscape, long-term time-series forecasting (LTSF)—predicting a horizon H steps ahead from a history window of length T—has emerged as a major focus due to its practical importance and modeling challenges such as nonstationarity (Liu et al., 2022c), multi-scale dynamics (Franceschi et al., 2019; Tonekaboni et al., 2021; Yang & Hong, 2022; Yue et al., 2022; Liu et al., 2022b), and covariate shift (Ulyanov et al., 2016; Kim et al., 2021).

Despite these advances, state-of-the-art forecasters are often brittle when observations are corrupted by noise. Notably, Figure 3 shows transformer-based SOTA models are easily deteriorated by small perturbation. Corruptions arise naturally in real sensors and pipelines (e.g., quantization, jitter, missingness, clock drift) and may also be adversarial or unmodeled (e.g., abrupt spikes, regime changes). Common robustness strategies include (i) augmenting the training set with gaussian noise (Tashiro et al., 2021; Rasul et al., 2021; Kollovieh et al., 2023) and (ii) suppressing high-frequency components via filtering (Zhou et al., 2022; Wu et al., 2023; Piao et al., 2024). However, both approaches risk discarding fine-grained, multi-scale structure which might be important information so that degrading accuracy on complex signals (cf. Section 2). Moreover, augmentation distributions rarely match the true, unknown corruption processes, and frequency-domain filtering can induce bias by over-smoothing informative transients.

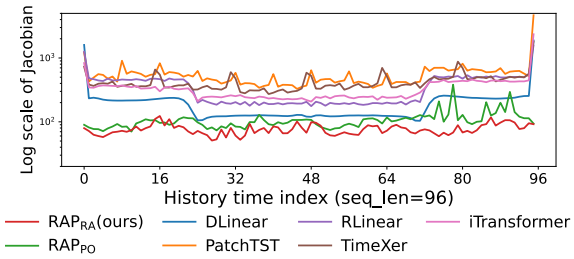

Figure 1: Log of the summed Jacobian energy w.r.t. each history step, averaged over prediction length 96, 192, 336, 720 on Traffic dataset. Lower curves indicate reduced input-noise sensitivity (greater robustness). We use Hutchinson estimator for calculating Jacobian by following Algorithm 3.

To address these issues, we propose a *reconstruction-augmented predictor* (RAP) that enhances robustness *without* augmenting the training data. Concretely, our model produces both a forecast and a self-reconstruction terms from the same shared representation(c.f. Equation 5). The reconstruction term acts as a structure-preserving regularizer that encourages the encoder to mitigate the tendency to overfit spurious or noise-specific artifacts by regularizing the Jacobian of predictor $\mathbb{E}_{\mathbf{x}_{1:T}} \|J_{\text{pred}}(\mathbf{x}_{1:T})\|_F^2$ for input sequence $\mathbf{x}_{1:T}$, named of *noise sensitivity* term. Unlike data-augmentation methods, RAP does not require the corruption of training data preserving detailed information while regularizing noise sensitivity, as Figure 1 shows.

Complementing this objective, we adopt a *dual-axis Transformer* (Vaswani et al., 2017) that alternates attention along temporal and feature (channel) axes to capture cross-time and cross-variable dependencies efficiently. The decoder generates the full target sequence conditioned on the history window, inspired by DiffWave-style sequence generation (Kong et al., 2020); this design naturally accommodates both prediction-only training ($\lambda = 0$) and reconstruction-augmented training ($\lambda = 1$) under a single architecture.

Empirically, on representative, complex benchmarks spanning energy, traffic, and electricity, RAP shows comparable results with SOTA baselines [1] in the clean setting and maintains its advantage as synthetic or real noise levels increase. We further analyze failure modes of augmentation-based robustness (Section 2), detail the model architecture, training and inference process (Section 5.1), provide theoretical justification for when reconstruction improves robustness (Section 4), and present extensive experiments and ablations (Section 7.2).

To summarize, our contributions are:

- We identify and theoretically characterize a failure mode of augmentation-based robustness for time-series forecasting, showing how mismatch between augmentation and true corruptions can erase detailed information (cf. Section 2).
- We introduce *RAP*, a reconstruction-augmented training objective with theoretical guarantees on structure preservation, together with a compatible dual-axis Transformer that supports both prediction-only and reconstruction-augmented methods under a unified design.
- Across diverse, complex datasets, our approach outperforms strong baselines and remains robust as the level and type of injected noise increase, with the largest gains on the most challenging settings, highlighting the effect of the reconstruction head on robustness.

2 Motivation : Failure modes of previous robustness techniques

There have been two types of machine learning methods for robustness: insert gaussian on input data (Tashiro et al., 2021; Rasul et al., 2021; Kollovieh et al., 2023) and remove high-frequency noise, e.g. fourier transform (Zhou et al., 2022; Wu et al., 2023; Piao et al., 2024). Also in time-series domain, these method can be applied by i) injecting gaussian perturbation in history sequence and ii) removing high-frequency noise from viewed window. In this section, we theoretically investigate on the failure mode of these methods by proving the following conjecture:

Gaussian noise injection training and High-pass filter for robust time-series prediction deteriorate original history data.

Setup. We study multi-step univariate time-series forecasting under an unknown distribution μ over history–future pairs (X, Y). Let $X \in \mathbb{R}^d$ with density p_X and Bayes regressor is $m(x) := \mathbb{E}[Y \mid X = x]$. Assume p_X, m in d-dimensional Schwartz space $\mathcal{S}(\mathbb{R}^d)$ with $p_X > 0$ everywhere.[2] We evaluate clean-test risk $R_0(f) := \mathbb{E}[(f(X) - Y)^2]$ (expectation w.r.t. p_X), and write the score $s(x) := \nabla \log p_X(x)$. Denote by $\varphi_\sigma(x) = (2\pi\sigma^2)^{-d/2} \exp(-\|x\|^2/(2\sigma^2))$ and by

$$(e^{t\Delta} g)(x) = \big(g * \phi_t\big)(x), \qquad \phi_t(x) = (4\pi t)^{-d/2} \exp\big(-\|x\|^2/(4t)\big), \quad t > 0.$$

2.0.1 Gaussian-noise training on $\mathbb{R}^d$

We train with corrupted inputs $U := X + Z$, $Z \sim \mathcal{N}(0, \sigma^2 I_d)$ independent of (X, Y), by minimizing $\mathcal{L}_\sigma(f) = \mathbb{E}[(f(U) - Y)^2]$ over a rich function class.

[1] Our baselines accommodate representative baselines in LTSF domain from the following github repository : https://github.com/thuml/Time-Series-Library

[2] Gaussian p_X and m with polynomially bounded derivatives also suffice.

Proposition 1 (Small-noise expansion on $\mathbb{R}^d$)**.** *The population minimizer is the conditional expectation*

$$f_\sigma(u) = \mathbb{E}[Y \mid U = u] = \mathbb{E}[m(X) \mid U = u] = \frac{\big((m\,p_X) * \varphi_\sigma\big)(u)}{\big(p_X * \varphi_\sigma\big)(u)}.$$

Moreover, as $\sigma \downarrow 0$,

$$f_\sigma(x) = m(x) + \frac{\sigma^2}{2}\Big(\Delta m(x) + 2\,\nabla m(x)\cdot s(x)\Big) + O(\sigma^4), \tag{1}$$

with the $O(\sigma^4)$ remainder uniform on $\mathbb{R}^d$ (hence also in $L^2(p_X)$). Consequently,

$$\Delta R_0^{\text{noise}}(\sigma) := R_0(f_\sigma) - R_0(m) = \frac{\sigma^4}{4}\,\mathbb{E}\Big[\big(\Delta m + 2\,\nabla m\cdot s\big)^2\Big] + O(\sigma^6). \tag{2}$$

Remark 2. By Plancherel on $\mathbb{R}^d$, $\|\Delta m\|^2_{L^2(dx)} = \int \|\xi\|^4 |\widehat{m}(\xi)|^2\, d\xi$, so the leading term in equation 2 accentuates high frequencies. If the score s is bounded (e.g., Gaussian p_X), then

$$\mathbb{E}\big[(\Delta m + 2\nabla m \cdot s)^2\big] \;\geq\; \tfrac{1}{2}\,\mathbb{E}[(\Delta m)^2] - 2\,\|s\|^2_\infty\,\mathbb{E}[\|\nabla m\|^2],$$

so HF-dominant m makes $\Delta R_0^{\text{noise}}(\sigma)$ large even for small σ.

2.0.2 Low-pass conditioning on $\mathbb{R}^d$

Let $P_{\leq\omega_c}$ be the Fourier low-pass projector (Fourier multiplier by $\mathbf{1}_{\{\|\xi\|\leq\omega_c\}}$) acting on arguments: the model only receives $U := P_{\leq\omega_c}X$. The population optimal predictor is $f^{\text{LPF}}(x) := \mathbb{E}[m(X) \mid U = P_{\leq\omega_c}x]$.

Proposition 3 (Decomposition and general lower bound)**.** *For any p_X,*

$$R_0(f^{\text{LPF}}) - R_0(m) = \mathbb{E}[\mathit{Var}(m(X) \mid P_{\leq\omega_c}X)] = \inf_h\; \mathbb{E}\Big[\big(m(X) - h(P_{\leq\omega_c}X)\big)^2\Big],$$

where the infimum is over all measurable $h : \mathbb{R}^d \to \mathbb{R}$. Moreover, for any $R > 0$,

$$R_0(f^{\text{LPF}}) - R_0(m) \;\geq\; \Big(\inf_{\|x\|\leq R} p_X(x)\Big) \int_{\|x\|\leq R} \big(m(x) - \Pi_{\mathcal{H}_L} m(x)\big)^2\, dx.$$

Remark 4. In images, mild manipulation of high frequencies often preserves perceptual quality; in LTSF, however, fine, phase-sensitive structure at high frequencies is often *predictive* and the above propositions formalize this. Thus, in high-frequency regimes these tactics erase predictive microstructure rather than harmless detail, making them poor generic robustness fixes for time series and motivating our reconstruction-augmented approach.

3 Preliminaries

Inspired by failure modes of previous robustness techniques, we propose a *reconstruction-augmented predictor* (RAP) that guarantees robust prediction without data manipulation. In this section, we argue problem statement and describe theoretical derivation of our method.

3.1 Problem Statement

Setting and notation. We study multi-step univariate time-series forecasting under an unknown distribution μ over history–future pairs $(X, Y) \in \mathbb{R}^T \times \mathbb{R}^H$.[3] Here X collects the most recent T samples in a lookback window and Y denotes the next H samples to be predicted. We write $\mathbb{E}[\cdot]$ for expectation with respect to the relevant variables (explicitly indicated when needed), $\|\cdot\|$ for the Euclidean norm, $\|\cdot\|_{\text{op}}$ for the spectral (operator) norm, and $\|\cdot\|_F$ for the Frobenius norm.

A single encoder $\Phi_\theta : \mathbb{R}^T \to \mathbb{R}^D$ produces a shared representation $z = \Phi_\theta(x)$, on which we place two linear heads:

$$f_{\text{rec}}(x) = W_c\,\Phi_\theta(x) \in \mathbb{R}^T, \qquad f_{\text{pred}}(x) = W_p\,\Phi_\theta(x) \in \mathbb{R}^H, \tag{3}$$

[3]We point out that the univariate arguments can be extend directly to the multivariate setting: replace derivatives by Jacobians and scalar inner products by trace forms. We refer to Section 4.

where $W_c \in \mathbb{R}^{T\times D}$ and $W_p \in \mathbb{R}^{H\times D}$. The *reconstruction* head f_{rec} reconstructs the input history X and serves as a structure-preserving self-supervised signal, while the *prediction* head f_{pred} forecasts Y using the same representation. For later use, we denote the Jacobians with respect to the input:

$$J_\Phi(x) := \nabla_x \Phi_\theta(x) \in \mathbb{R}^{D\times T}, \quad J_{\text{rec}}(x) := \nabla_x f_{\text{rec}}(x) = W_c\, J_\Phi(x), \quad J_{\text{pred}}(x) := \nabla_x f_{\text{pred}}(x) = W_p\, J_\Phi(x).$$

Also, we operate under mild spectral constraints (standard in practice):

$$\sigma_{\min}(W_c) \;\geq\; \beta > 0, \qquad \|W_p\|_{\text{op}} \;\leq\; B_p. \tag{4}$$

Intuitively, the first condition prevents the reconstruction head from collapsing informative directions (so the "conditioning" signal is well-posed), and the second bounds the Lipschitz constant of the prediction head, curbing gradient explosion and improving stability.

Training objectives. Conventional forecasters ignore f_{rec} and optimize only the prediction loss

$$\mathcal{L}_{\text{pred}}(\theta, W_p) := \mathbb{E}_{(X,Y)\sim\mu}\, \|f_{\text{pred}}(X) - Y\|^2.$$

In our reconstruction-augmented formulation, we add a reconstruction term and minimize the composite objective

$$\mathcal{L}_{\text{RAP}}(\theta, W_c, W_p) := \mathcal{L}_{\text{pred}}(\theta, W_p) + \lambda \mathcal{L}_{\text{rec}}(\theta, W_c), \quad \mathcal{L}_{\text{rec}}(\theta, W_c) := \mathbb{E}_X\, \|f_{\text{rec}}(X) - X\|^2, \tag{5}$$

where $\lambda \in \{0, 1\}$. This regularizer encourages the encoder to retain fine-grained, multi-scale information in X that is predictive for Y, without requiring data augmentation. Specifically, we refer to the two settings as (i) $\lambda = 0$: prediction-only (RAP_{PO}) and (ii) $\lambda = 1$: reconstruction-augmented (RAP_{RA}).

For convenience, we instantiate a shared latent of size $D = T + H$ and choose $W_c = [\mathbf{I}_d, \mathbf{0}]$ and $W_p = [\mathbf{0}, \mathbf{I}_k]$ so that the model can generate the concatenated sequence $[f_{\text{rec}}(x),\, f_{\text{pred}}(x)] \approx [X, Y]$. Therefore, it is straightforward that the $\mathcal{L}_{\text{RAP}}$ is represented by

$$\mathcal{L}_{\text{RAP}}(\theta) = \mathbb{E}_{X,Y}[\, \underbrace{\|\hat{Y} - Y\|}_{\text{prediction}}^2 + \lambda \underbrace{\|\hat{X} - X\|}_{\text{reconstruction}}^2],$$

where $\hat{X}, \hat{Y}$ are history, prediction part of $\Phi_\theta(X)$. This choice simplifies implementation but is *not* required by our analysis; at inference we report forecasts from f_{pred}.

3.2 Robust prediction under noisy inputs

To capture robustness to sensor, quantization, or preprocessing noise at deployment, we evaluate the *noisy risk*

$$\mathcal{R}_\sigma(f_{\text{pred}}) := \mathbb{E}_{X,\varepsilon}\, \big\|f_{\text{pred}}(X + \sigma\varepsilon) - Y\big\|^2, \qquad \varepsilon \sim \mathcal{N}(0, I_d) \;\text{ independent of } (X, Y), \tag{6}$$

and compare it to the clean risk $\mathcal{R}_0(f_{\text{pred}}) = L_{\text{pred}}$. Here σ controls the noise level and I_d denotes the $d \times d$ identity.

Assuming differentiability and square-integrability of the Jacobian, a first-order Taylor expansion around X yields, for small $\sigma > 0$,

$$\mathbb{E}_{X,\varepsilon}\, \|f_{\text{pred}}(X + \sigma\varepsilon) - Y\|^2 \;\simeq\; \mathbb{E}_{(X,Y)}\, \|f_{\text{pred}}(X) - Y\|^2 \;+\; \sigma^2\, \mathbb{E}_X\, \|J_{\text{pred}}(X)\|_F^2.$$

Following standard small-noise analysis, we define the *noise sensitivity*

$$S \;:=\; \mathbb{E}_X\, \|J_{\text{pred}}(X)\|_F^2, \tag{7}$$

which serves as a first-order proxy for the increase $\mathcal{R}_\sigma - \mathcal{R}_0$ when σ is small. Intuitively, S measures how strongly the predictor amplifies infinitesimal, isotropic perturbations of the input; lower S indicates a more stable (robust) mapping. In Section 4, we show how the reconstruction term in L_{RAP} reduces S while maintaining low clean risk.

4 Small-noise robustness via Jacobian energy

In this section, we analyze the effect of small additive input noise on the prediction risk and begin with a standard smoothness assumption.

Assumption 5. The map f_{pred} is twice continuously differentiable in x, with bounded second and third derivatives on the support of μ. Equivalently, for each coordinate $g_\ell(x) := [f_{\text{pred}}(x) - Y]_\ell$ we have $g_\ell \in C^3$ with $\sup_x \big\|\nabla^2 g_\ell(x)\big\|_{\text{op}} \leq H$ and $\sup_x \big\|\nabla^3 g_\ell(x)\big\| \leq M$ for some $H, M < \infty$.

Under Assumption 5, the difference between noisy and clean risks admits a small-noise expansion.

Theorem 6. *Let $\varepsilon \sim \mathcal{N}(0, I_d)$ be independent of (X, Y) and define $\mathcal{R}_\sigma(f_{\text{pred}}) := \mathbb{E}_{X,\varepsilon} \left\| f_{\text{pred}}(X + \sigma\varepsilon) - Y \right\|^2$ and $\mathcal{R}_0(f_{\text{pred}}) := L_{\text{pred}}$. Under Assumption 5*

$$\mathcal{R}_\sigma(f_{\text{pred}}) = \mathcal{R}_0(f_{\text{pred}}) \;+\; \sigma^2\, \mathbb{E}\, \|J_{\text{pred}}(X)\|_{\text{F}}^2 \;+\; \frac{\sigma^2}{2}\, \mathbb{E}\, \langle f_{\text{pred}}(X) - Y,\, \Delta f_{\text{pred}}(X)\rangle \;+\; O(\sigma^4), \tag{8}$$

where Δ is the componentwise Laplacian. In particular, there exist explicit constants $c_{d,k}, C$ depending only on (d, k) and the derivative bounds such that

$$\mathcal{R}_\sigma(f_{\text{pred}}) - \mathcal{R}_0(f_{\text{pred}}) \;\le\; \sigma^2\, \mathbb{E}\, \|J_{\text{pred}}(X)\|_{\text{F}}^2 \;+\; c_{d,k}\, \sigma^2\, H \sqrt{\mathbb{E}\, \|f_{\text{pred}}(X) - Y\|^2} \;+\; C\, \sigma^4. \tag{9}$$

By Theorem 6, when the model is well-trained, $L_{\text{pred}} = \mathbb{E}\, \|f_{\text{pred}}(X) - Y\|^2$ became small, the Laplacian inner-product term in equation 8 is negligible compared to the Jacobian energy. Thus the *noise sensitivity* $S = \mathbb{E}\, \|J_{\text{pred}}(X)\|_{\text{F}}^2$ governs the dominant σ^2-order increase in risk, and reducing S directly improves small-noise robustness.

Next, we prove that the sensitivity S, which is the problematic term in equation 9, is controllable via the reconstruction objective $\mathcal{L}_{\text{rec}}(\theta, W_c) := \mathbb{E}\, \|f_{\text{rec}}(X) - X\|^2$. We assume a classical C^2 hypothesis and a local Hessian bound to obtain an ε–δ type inequality.

Assumption 7. There exists a bounded convex domain $\Omega \subset \mathbb{R}^d$ with diameter D such that $\text{supp}(\mu) \subset \Omega$. Let $u(x) := f_{\text{rec}}(x) - x \in C^2(\Omega; \mathbb{R}^d)$ and assume

$$H \;:=\; \sup_{x \in \Omega} \left\| \nabla^2 u(x) \right\|_{\text{op}} \;<\; \infty.$$

Theorem 8. *Under Assumption 7, there exists a constant $C_d > 0$ (depending only on d) such that*

$$\mathbb{E}\, \|J_{\text{rec}}(X) - I\|_{\text{F}}^2 \;\le\; \frac{C_d}{D^2}\, \mathbb{E}\, \|f_{\text{rec}}(X) - X\|^2 \;+\; C_d\, H^2\, D^2. \tag{10}$$

Equivalently, for any $\varepsilon > 0$, if

$$\mathbb{E}\, \|f_{\text{rec}}(X) - X\|^2 \;\le\; \frac{\varepsilon^2}{2C_d}\, D^2 \quad \textit{and} \quad H\, D \;\le\; \frac{\varepsilon}{\sqrt{2C_d}},$$

then $\mathbb{E}\, \|J_{\text{rec}}(X) - I\|_{\text{F}} \;\le\; \varepsilon$.

For any $t \in (0, D/2]$ and $u := f_{\text{rec}} - \text{id}$, Theorem 8 also implies

$$\mathbb{E}\, \|\nabla u(X)\|_{\text{F}}^2 \;\le\; \frac{c_d}{t^2}\, \mathbb{E}\, \|u(X)\|^2 \;+\; C_d\, H^2\, t^2.$$

Choosing $t_\star = \left(\frac{c_d\, \mathbb{E}\|u\|^2}{C_d\, H^2}\right)^{1/4}$ (whenever $t_\star \le D/2$, as is typical for a well-trained model) yields

$$\mathbb{E}\, \|J_{\text{rec}}(X) - I\|_{\text{F}} \;\le\; \sqrt{2}\, (c_d C_d)^{1/4} \sqrt{H} \left(\mathbb{E}\, \|f_{\text{rec}}(X) - X\|^2\right)^{1/4}.$$

Thus, with bounded curvature $H < \infty$, driving the reconstruction error $\mathbb{E}\|f_{\text{rec}}(X) - X\|^2 \to 0$ ensures $\mathbb{E}\|J_{\text{rec}}(X) - I\|_{\text{F}} \to 0$. If instead $t_\star > D/2$, taking $t \simeq D$ recovers equation 10; in that regime, the condition $HD \lesssim \varepsilon$ is a convenient sufficient criterion. Practically, one can (i) rescale inputs to reduce D, (ii) regularize curvature (e.g., spectral/Jacobian/Hessian penalties) to control H, or (iii) localize the analysis to smaller patches (smaller effective D).

Scope and multivariate extension. For clarity, several theoretical statements in Section 4 are presented for the univariate case (d=k=1). The arguments extend directly to the multivariate setting: replace derivatives by Jacobians and scalar inner products by trace forms. Under channel-correlated Gaussian input noise $\varepsilon \sim \mathcal{N}(0, \Sigma)$ with $\Sigma \succeq 0$, the small-noise expansion holds with the sensitivity term weighted by Σ, i.e., $\mathbb{E}\, \text{tr}\big(J_{\text{pred}}(X)\, \Sigma\, J_{\text{pred}}(X)^\top\big)$, which reduces to equation 7 when $\Sigma = I_d$. All bounds carry over with constants scaling at most by $\|\Sigma\|_{\text{op}}$ and the spectral constraints in equation 4; equivalently, one may argue in whitened coordinates via $\Sigma^{-\frac{1}{2}}$ without changing the structure of the proofs.

5 RAP: RECONSTRUCTION-AUGMENTED PREDICTORS

In Section 4, we proved that a simple reconstruction loss reduces the dominant small-noise term (the Jacobian energy) of the predictor. Here, we detail how we instantiate the model and how we train and evaluate it for both accuracy and robustness.

5.1 MODEL ARCHITECTURE

Our backbone follows DiffWave (Kong et al., 2020) in spirit but differs in three key aspects. Because DiffWave-style generators map inputs to outputs of the *same shape* to learn score function, we feed the model the concatenated sequence $[\mathbf{x}_{1:T}, \mathbf{0}]$ (zero-padding the future) and train it to produce $[\hat{\mathbf{x}}_{1:T}, \hat{\mathbf{y}}_{1:H}]$. This single-pass, same-shape design naturally supports both the forecasting and reconstruction objectives, enabling compare the effect of reconstruction term within single architecture.

First, we *remove the diffusion-step (t) embedding and the entire diffusion schedule*, training a one-shot forecaster with a plain MSE objective. Second, each residual block mixes information via *dual-axis self-attention* (along time and along features) instead of dilated convolutions, explicitly capturing long-range temporal patterns and cross-variable interactions. Third, we employ *LayerNorm* and *Dropout* within blocks. For forecasting, we optionally apply *Reversible Instance Normalization* (RevIN). The decoder generates the full H-step target in one shot conditioned on the encoded history. Further architectural description appears in Appendix B.

5.2 TRAINING AND INFERENCE

Given a history $\mathbf{x}_{1:T}$ and target $\mathbf{y}_{1:H}$, RAP produces a forecast $\hat{\mathbf{y}}_{1:H}$ and a self-reconstruction $\hat{\mathbf{x}}_{1:T}$ from a shared representation. We train the model by using the Equation 5. Again, we refer to the two settings as (i) $\lambda = 0$: prediction-only ($\mathrm{RAP}_{\mathrm{PO}}$) and (ii) $\lambda = 1$: reconstruction-augmented ($\mathrm{RAP}_{\mathrm{RA}}$). Note that both share the *same* architecture and generate the full sequence; they differ only in whether the reconstruction term is optimized. Losses are computed in the normalized space. By default we use MSE for both terms and optimize with Adam and a cosine learning-rate scheduler.

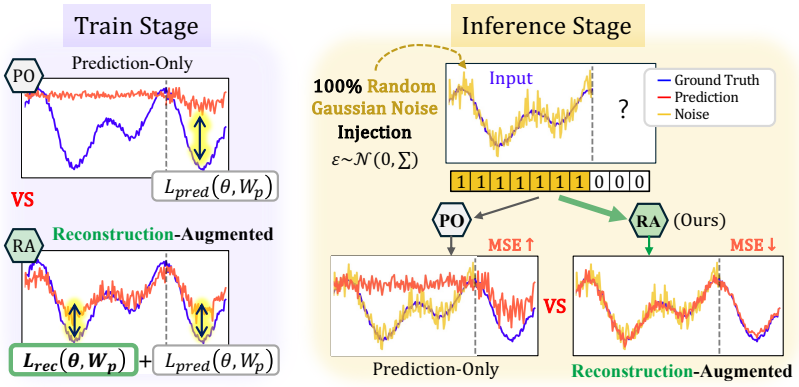

Figure 2: Overall visualization of the training and inference process

At inference time, we evaluate both standard LTSF accuracy and robustness under input noise. For robustness, we inject i.i.d. Gaussian noise into the history window, $\mathbf{x}_{1:T} \mapsto \mathbf{x}_{1:T} + \sigma\varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I)$, and report the MSE/MAE as σ increases. We compare $\mathrm{RAP}_{\mathrm{RA}}$, $\mathrm{RAP}_{\mathrm{PO}}$, and representative SOTA baselines under identical settings in Section 7.2. This protocol directly probes the small-noise sensitivity analyzed in Section 4 while holding the forecasting setup fixed. We describe detailed training and inference process in Algorithm 1, 2 in Appendix C.

6 RELATED WORK

Research on long-term time-series forecasting (LTSF) has advanced along several complementary axes, including temporal decomposition, spectral modeling, segmentation/patching, cross-dimension dependency modeling, and normalization. Below we summarize representative approaches that we partly include as baselines.

Decomposition-centric models. Autoformer (Wu et al., 2021) introduces a progressive decomposition block and an auto-correlation mechanism that replaces point-wise self-attention, improving efficiency while explicitly disentangling trend and seasonal components. This decomposition perspective motivated follow-ups such as FEDformer (Zhou et al., 2022), DLinear (Zeng et al., 2023), and xPatch (Stitsyuk & Choi, 2025), which further explore how separating low- and high-frequency behavior can stabilize long-horizon prediction.

Spectral and multi-period modeling. TimesNet (Wu et al., 2023) leverages the Fourier transform to decompose signals into frequency components with varying periods, strengthening modeling of multi-scale temporal variation. Beyond forecasting, TimesNet is proposed as a task-general backbone for imputation, anomaly detection, and classification, and its official codebase offers a standardized evaluation protocol and unified baselines.

Segmentation and patch tokens. Crossformer (Zhang & Yan, 2023) and PatchTST (Nie et al., 2023) segment long sequences into subseries-level patches that act as input tokens for Transformers. PatchTST focuses on local temporal patterns via patching, whereas Crossformer emphasizes explicit cross-dimension dependencies to capture interactions among variables.

Simplicity and normalization. DLinear (Zeng et al., 2023) shows that a simple linear architecture, coupled with series decomposition and careful preprocessing, can be surprisingly competitive with Transformer-based models. RLinear (Li et al., 2023) further highlights the role of normalization, combining linear forecasting with Reversible Instance Normalization (RevIN) (Kim et al., 2021) to stabilize training and improve generalization across datasets.

Cross-variate attention at scale. iTransformer (Liu et al., 2024) takes an "inverted" view that embeds time points into variate tokens, shifting attention from temporal correlations to cross-variate correlations. By treating time implicitly and focusing attention across variables, iTransformer is particularly effective for high-dimensional multivariate settings where variable interactions dominate.

Utilizing exogenous signals. TimeXer (Wang et al., 2024) targets scenarios in which exogenous series provide crucial context for forecasting endogenous targets. It augments the Transformer with exogenous embeddings, patch-wise self-attention, variate-wise cross-attention, and global endogenous tokens, enabling scalable fusion of external and internal signals.

The methods above advance LTSF via stronger inductive biases (decomposition, spectral filters), scalable tokenization (patching), and training stability (normalization). Our reconstruction-augmented predictor (RAP) is orthogonal to these directions: instead of modifying the data distribution (e.g., via augmentation or filtering), RAP regularizes the *noise sensitivity* by jointly forecasting the future and reconstructing the history for robust prediction. This encourages the encoder to preserve fine-grained, multi-scale structure that remains predictive at long horizons, and it can be paired with strong LTSF backbones.

7 Experiments

To evaluate the proposed RAP model, we conducted two experiments. The first experiment is a general Long-term Time Series Forecasting (LTSF) experiment that evaluates the predictive ability of the model. In contrast, the second experiment is a robustness experiment that evaluates the model's resilience to noise. We also attach ablation study about model architecture in Section D.1.

7.1 Experimental Setup

Following previous works, we use Mean Squared Error (MSE) and Mean Absolute Error (MAE) metrics to assess the performance of the LTSF models. All the experiments are implemented in PyTorch (Paszke et al., 2019), and conducted on a single Quadro RTX 6000 GPU. For the general LTSF experiment, we follow the forecasting protocol proposed by TimesNet (Wu et al., 2023) with a lookback length $L = 96$ and prediction length $T = \{96, 192, 336, 720\}$. For the robustness experiment, we explore the performance of models on the data corrupted with random Gaussian noise $\varepsilon = \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$.

Table 1: Detailed dataset descriptions.

Dataset	Dim	Dataset Size	Frequency
Solar	137	(36792,5271,10540)	10 min
Elec.	321	(18317,2633,5261)	Hourly
Traffic	862	(12185,1757,3509)	Hourly
PEMS03	358	(15617,5135,5135)	5 min
PEMS04	307	(10172,3375,3375)	5 min
PEMS07	883	(16911,5622,5622)	5 min

7.1.1 Datasets and Baselines

We conduct extensive experiments on six real-world multivariate time series datasets, including Solar-energy (Lai et al., 2018), Electricity (ECL), Traffic, (Wu et al., 2021), and Performance Measurement System (PEMS) 3, 4, 7 (Chen et al., 2001) introduced in (Liu et al., 2022a). We focus on datasets that have a high number of channels and possess more complex temporal features for evaluating the robustness of the proposed approach. Table 4 summarizes details of statistics of datasets.

- Solar-energy[4]: contains the solar power production records from 137 PV plants in 2006.
- Electricity (ECL)[5]: describes the hourly electricity consumption data of 321 clients from 2012 to 2014.

[4] https://www.nrel.gov/grid/solar-power-data.html
[5] https://archive.ics.uci.edu/dataset/321/electricityloaddiagrams20112014

Table 2: Full results of the long-term multivariate forecasting task (MSE / MAE). Ranking notation: lower is better; 1st=**bold**, 2nd=underline, 3rd=*emph*.

Dataset	Horizon	**RAP$_{\text{RA}}$(Ours)**		TimeXer		iTrans.		RLinear		PatchTST		Cross.		TimesNet		DLinear		Auto.	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ECL	96	*0.149*	*0.247*	**0.140**	0.242	0.148	**0.240**	0.201	0.281	0.195	0.285	0.219	0.314	0.168	0.272	0.197	0.282	0.201	0.317
	192	*0.173*	*0.274*	**0.157**	0.256	0.162	**0.253**	0.201	0.283	0.199	0.289	0.231	0.322	0.184	0.289	0.196	0.285	0.222	0.334
	336	*0.191*	*0.290*	**0.176**	0.275	0.178	**0.269**	0.215	0.298	0.215	0.305	0.246	0.337	0.198	0.300	0.209	0.301	0.231	0.338
	720	0.227	0.321	**0.211**	**0.306**	*0.225*	0.317	0.257	0.331	0.256	0.337	0.280	0.363	0.220	*0.320*	0.245	0.333	0.254	0.361
	Avg	*0.185*	*0.283*	**0.171**	**0.270**	0.178	0.270	0.219	0.298	0.216	0.304	0.244	0.334	0.192	0.295	0.212	0.300	0.227	0.338
Traffic	96	**0.312**	0.340	*0.428*	0.271	0.395	**0.268**	0.649	0.389	0.462	*0.295*	0.522	0.290	0.593	0.321	0.650	0.396	0.613	0.388
	192	**0.334**	0.345	*0.448*	0.282	0.417	**0.276**	0.601	0.366	0.466	*0.296*	0.530	0.293	0.617	0.336	0.598	0.370	0.616	0.382
	336	**0.385**	0.379	*0.473*	0.289	0.433	**0.283**	0.609	0.369	0.482	*0.304*	0.558	0.305	0.629	0.336	0.605	0.373	0.622	0.337
	720	**0.421**	0.393	0.516	0.307	0.467	**0.302**	0.647	0.387	*0.514*	*0.322*	0.589	0.328	0.640	0.350	0.645	0.394	0.660	0.408
	Avg	**0.363**	0.364	*0.466*	0.287	0.428	**0.282**	0.626	0.378	0.481	*0.304*	0.550	*0.304*	0.620	0.336	0.625	0.383	0.628	0.379
Solar	96	**0.145**	**0.168**	*0.207*	*0.259*	0.203	0.237	0.322	0.339	0.234	0.286	0.310	0.331	0.250	0.292	0.290	0.378	0.884	0.711
	192	**0.153**	**0.180**	*0.237*	*0.292*	0.233	0.261	0.359	0.356	0.267	0.310	0.734	0.725	0.296	0.318	0.320	0.398	0.834	0.692
	336	**0.210**	**0.217**	0.248	*0.293*	0.248	0.273	0.397	*0.369*	*0.290*	0.315	0.750	0.735	0.319	0.330	0.353	0.415	0.941	0.723
	720	0.348	0.292	0.258	*0.298*	**0.249**	**0.275**	0.397	0.356	*0.289*	0.317	0.769	0.765	0.338	0.337	0.356	0.413	0.882	0.717
	Avg	**0.214**	**0.214**	*0.238*	*0.286*	0.233	0.262	0.369	0.356	0.270	0.307	0.641	0.639	0.301	0.319	0.330	0.401	0.885	0.711
PEMS03	12	**0.067**	**0.170**	0.067	*0.178*	*0.070*	0.176	0.114	0.233	0.078	0.185	0.090	0.203	0.085	0.192	0.105	0.220	0.272	0.385
	24	**0.079**	**0.186**	0.088	0.204	*0.097*	*0.209*	0.193	0.309	0.121	0.230	0.121	0.240	0.118	0.223	0.183	0.296	0.334	0.440
	48	**0.122**	**0.222**	0.140	0.253	0.164	0.275	0.329	0.421	0.213	0.306	0.202	0.317	*0.155*	*0.260*	0.318	0.407	1.032	0.782
	96	**0.131**	**0.238**	0.451	0.501	0.264	*0.360*	0.455	0.513	0.364	0.410	*0.262*	0.367	0.228	0.317	0.451	0.507	1.031	0.796
	Avg	**0.100**	**0.204**	0.186	0.284	*0.149*	*0.255*	0.273	0.369	0.194	0.283	0.169	0.282	0.146	0.248	0.264	0.358	0.667	0.601
PEMS04	12	**0.074**	**0.180**	0.074	0.185	0.089	*0.194*	0.135	0.256	0.099	0.206	0.098	0.218	*0.087*	0.195	0.115	0.228	0.424	0.491
	24	**0.084**	**0.193**	0.088	0.202	*0.124*	*0.232*	0.212	0.325	0.165	0.271	0.131	0.256	0.103	0.215	0.189	0.299	0.459	0.509
	48	**0.097**	**0.210**	0.114	0.231	0.195	0.296	0.342	0.424	0.310	0.383	0.205	0.326	0.136	0.250	0.322	0.405	0.646	0.610
	96	**0.110**	**0.225**	0.162	0.283	0.479	0.497	0.443	0.496	0.527	0.515	0.402	0.457	*0.190*	*0.303*	0.428	0.483	0.912	0.748
	Avg	**0.091**	**0.202**	0.109	0.225	0.222	0.305	0.283	0.375	0.275	0.344	0.209	0.314	*0.129*	*0.241*	0.264	0.354	0.610	0.590
PEMS07	12	**0.059**	**0.156**	0.059	0.163	*0.068*	*0.168*	0.108	0.231	0.080	0.189	0.094	0.200	0.082	*0.181*	0.100	0.215	0.199	0.336
	24	**0.071**	**0.171**	0.077	0.182	*0.096*	*0.200*	0.202	0.318	0.131	0.238	0.139	0.247	0.101	0.204	0.189	0.302	0.323	0.420
	48	**0.088**	**0.189**	0.117	0.227	0.185	0.289	0.389	0.448	0.259	0.339	0.311	0.369	*0.134*	*0.238*	0.377	0.439	0.390	0.470
	96	**0.106**	**0.202**	0.159	0.266	0.214	0.311	0.584	0.547	0.429	0.445	0.396	0.442	*0.181*	*0.279*	0.580	0.539	0.554	0.578
	Avg	**0.081**	**0.179**	0.103	0.209	0.141	0.242	0.321	0.386	0.225	0.303	0.235	0.315	*0.124*	*0.226*	0.311	0.374	0.367	0.451

- Traffic[6], records hourly road occupancy rates measured by 862 sensors on San Francisco Bay area freeways from January 2015 to December 2016.
- Performance Measurement System (PEMS)[7] contains the public traffic network data in California collected by 5-minute windows.

For baselines, we select the last state-of-the-art LTSF models[8], namely Autoformer (Wu et al., 2021), DLinear (Zeng et al., 2023), TimesNet (Wu et al., 2023), Crossformer (Zhang & Yan, 2023), PatchTST (Nie et al., 2023), RLinear (Li et al., 2023), iTransformer (Liu et al., 2024), and TimeXer (Wang et al., 2024). We describe the details in Section 6.

7.2 Experimental Results

7.2.1 General LTSF Experiment

In the general LTSF experiment, we evaluate the regular forecasting abilities of the prediction models. For a fair comparison, we use unified settings for all models: a lookback length $L = 96$, prediction lengths $T = \{12, 24, 48, 96\}$ for the PEMS dataset, and $L = 96$, $T = \{96, 192, 336, 720\}$ for all other datasets. To handle data distribution shift, we apply reversible instance normalization (Kim et al., 2021).

The results of this experiment demonstrate that the proposed RAP model is comparable with all baseline methods in the LTSF task. Notably, RAP achieves comparable results with other SOTA models, scoring 3rd best performance on ECL and 1st on the most of other cases. This confirms that RAP effectively captures long-term temporal dependencies and achieves superior performance in the general LTSF setting.

[6] https://pems.dot.ca.gov
[7] https://pems.dot.ca.gov
[8] Our baseline selection stems from the following github repository : https://github.com/thuml/Time-Series-Library

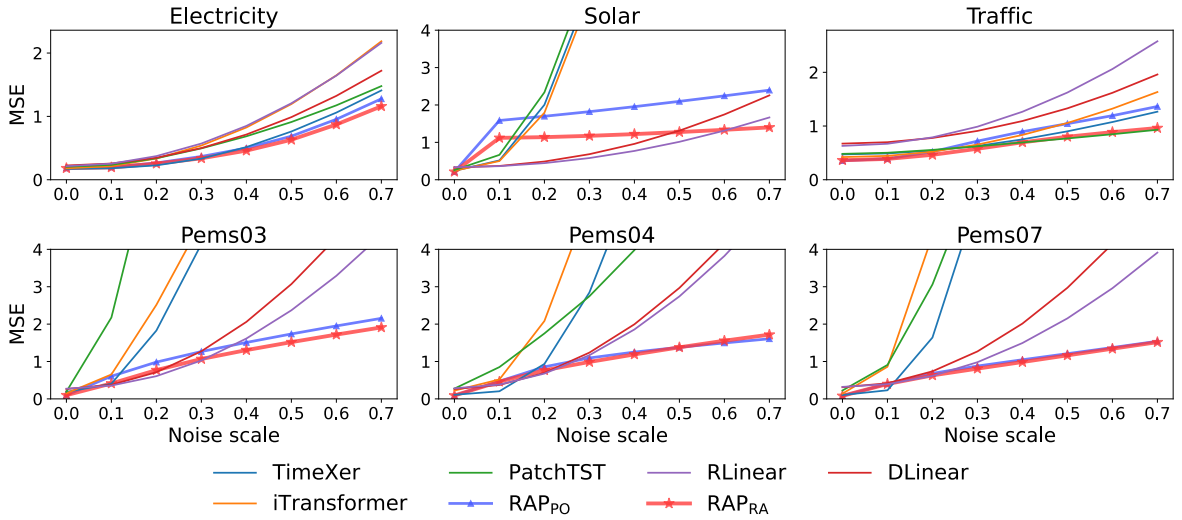

Figure 3: Average MSE vs. noise scale for Electricity, Traffic, and Solar datasets. Results are averaged over prediction lengths $H = \{96, 192, 336, 720\}$.

7.2.2 Experiments about Robustness

In the robustness experiment, we evaluate how well the prediction models adapt to data corrupted with random Gaussian noise. Specifically, we inject random Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma I_d)$ with standard deviation σ into the input time series across the history part of all datasets, simulating real-world measurement errors and sensor disturbances. We report the MSE/MAE with ground truth as σ increases. All models are trained and tested under the same noise-injected conditions, with the lookback and prediction settings identical to those used in the general LTSF experiment for a fair comparison.

The Figure 3 shows that the proposed RAP model consistently achieves the best forecasting performance under noisy conditions. Compared to baseline models, RAP demonstrates a stronger ability to preserve temporal patterns and maintain stable predictions despite input perturbations. Although some baselines, such as DLinear and RLinear on Solar dataset, achieves robust prediction then other SOTA models on specific dataset but cannot generate good performance on the other datasets, $\mathrm{RAP}_{\mathrm{RA}}$ consistently generates relatively accurate prediction than any other baselines. This indicates that RAP is not only effective in clean scenarios but also robust to uncertainty and noise in practical forecasting environments. We also give Table 5 in Appendix D which possesses exact MSE/MAE results.

When comparing RAP variants, the performance difference between $\mathrm{RAP}_{\mathrm{PO}}$ and $\mathrm{RAP}_{\mathrm{RA}}$ is relatively small on simpler datasets, while on more complex datasets such as Solar or Traffic, the gap becomes noticeably larger. Nevertheless, even the simpler RAP prediction-only variant surpasses other state-of-the-art models under noisy conditions, suggesting that the architecture of RAP itself inherently provides robustness to perturbations.

8 Conclusion

We introduced *reconstruction-augmented predictors* (RAP) for robust long-term time-series forecasting. By coupling a standard forecasting loss with a self-reconstruction objective on the history window and using a dual-axis Transformer backbone, RAP reduces input-noise sensitivity while preserving fine-grained, predictive structure. Our small-noise analysis explains the observed robustness gains through an explicit control of the predictor's Jacobian energy, and extensive experiments across high-dimensional benchmarks show that RAP matches or exceeds strong baselines in clean conditions and maintains a clear advantage under synthetic and real perturbations.

Limitations and future work. As the first framework to regularize forecasting via self-reconstruction, RAP's *best-fit architectures* are still under-explored; our DiffWave-inspired dual-axis backbone is a single instantiation rather than the optimum. We also find certain models unusually robust on specific datasets, e.g. DLinear and RLinear on Solar; explaining which inductive biases (decomposition, spectral, patching, cross-variate attention, normalization) match which regimes could guide an "ideal" RAP design. Future work includes robustness-aware network architecture and broader real-world corruption tests.

REFERENCES

Chao Chen, Karl Petty, Alexander Skabardonis, Pravin Varaiya, and Zhanfeng Jia. Freeway performance measurement system: mining loop detector data. *Transportation research record*, 1748(1):96–102, 2001.

Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. Unsupervised scalable representation learning for multivariate time series. *Advances in neural information processing systems*, 32, 2019.

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

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*, 2021.

Marcel Kollovieh, Abdul Fatir Ansari, Michael Bohlke-Schneider, Jasper Zschiegner, Hao Wang, and Yuyang Bernie Wang. Predict, refine, synthesize: Self-guiding diffusion models for probabilistic time series forecasting. *Advances in Neural Information Processing Systems*, 36:28341–28364, 2023.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 95–104, 2018.

Pedro Lara-Benítez, Manuel Carranza-García, and José C Riquelme. An experimental review on deep learning architectures for time series forecasting. *International journal of neural systems*, 31(03):2130001, 2021.

Zhe Li, Shiyi Qi, Yiduo Li, and Zenglin Xu. Revisiting long-term time series forecasting: An investigation on linear mapping. *arXiv preprint arXiv:2305.10721*, 2023.

Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A*, 379(2194):20200209, 2021.

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

Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *# PLACE-HOLDER_PARENT_METADATA_VALUE#*, 2022b.

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, 2022c.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=JePfAI8fah

Mohammad Amin Morid, Olivia R Liu Sheng, and Joseph Dunbar. Time series prediction using deep learning methods in healthcare. *ACM Transactions on Management Information Systems*, 14(1):1–29, 2023.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Xihao Piao, Zheng Chen, Taichi Murayama, Yasuko Matsubara, and Yasushi Sakurai. Fredformer: Frequency debiased transformer for time series forecasting. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 2400–2410, 2024.

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International conference on machine learning*, pp. 8857–8868. PMLR, 2021.

Omer Berat Sezer, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied soft computing*, 90:106181, 2020.

Artyom Stitsyuk and Jaesik Choi. xpatch: Dual-stream time series forecasting with exponential seasonal-trend decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20601–20609, 2025.

Gian Antonio Susto, Angelo Cenedese, and Matteo Terzi. Time-series classification methods: Review and applications to power systems data. *Big data application in power systems*, pp. 179–220, 2018.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in neural information processing systems*, 34:24804–24816, 2021.

Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. *arXiv preprint arXiv:2106.00750*, 2021.

José F Torres, Dalil Hadjout, Abderrazak Sebaa, Francisco Martínez-Álvarez, and Alicia Troncoso. Deep learning for time series forecasting: a survey. *Big data*, 9(1):3–21, 2021.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. *Advances in Neural Information Processing Systems*, 37:469–498, 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.

Ling Yang and Shenda Hong. Unsupervised time-series representation learning with iterative bilinear temporal-spectral fusion. In *International conference on machine learning*, pp. 25038–25054. PMLR, 2022.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 8980–8987, 2022.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The eleventh international conference on learning representations*, 2023.

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.