RADIAN: Regime-Adaptive Dilated Inter-Attention Network for Non-Stationary Financial Time-Series Forecasting

Nabeel Ahmad Saidd *

Posted Date: 14 April 2026

doi: 10.20944/preprints202604.0884.v1

Keywords: financial time series forecasting; non-stationary sequence modeling; regime detection; mixture-of-experts decoder; dilated causal convolution

 doi:10.20944/preprints202604.0884.v1

Article

RADIAN: Regime-Adaptive Dilated Inter-Attention Network for Non-Stationary Financial Time-Series Forecasting

Nabeel Ahmad Saidd

Dr. A. P. J. Abdul Kalam Technical University; nabeelahmadsaidd@gmail.com

Abstract

Financial forecasting is challenged by non-stationarity, volatility clustering, and regime transitions. Many neural forecasting pipelines use expressive encoders but fixed decoder mappings, which can reduce robustness under distribution shift. This paper introduces RADIAN, a forecasting architecture that keeps representation learning regime-agnostic and applies deterministic, causal regime conditioning only in fusion and decoding modules. The conditioning signal is a four-dimensional regime vector computed from normalized one-step target returns within the input window. RADIAN is evaluated under a fixed protocol on nine hourly financial datasets spanning equities, foreign exchange, cryptocurrencies, indices, and commodities, against eight baseline models and three random seeds. Across datasets, RADIAN attains the lowest mean test MAE, with an average improvement of 0.6704% relative to the strongest per-dataset baseline (range: 0.2542%–2.1234%). Paired statistical testing yields three unadjusted significant comparisons at $p < 0.05$, of which two remain significant after Benjamini–Hochberg correction; paired effect sizes are large in magnitude (median $|d| = 1.9312$). Component ablations indicate that decoder-side mechanisms contribute materially to performance: removing regime features increases RMSE by 1.38% and decreases directional accuracy by 0.91 percentage points, while removing path decoding increases RMSE by 4.38%. These results support decoder-side regime conditioning as an effective mechanism for robustness in non-stationary financial forecasting.

Keywords: financial time series forecasting; non-stationary sequence modeling; regime detection; mixture-of-experts decoder; dilated causal convolution

1. Introduction

Financial forecasting is a distribution-shift setting in which heavy tails, volatility clustering, and regime transitions are empirically common [1–3]. The challenge is acute at hourly horizons, where forecasts are consumed at fixed decision boundaries and local dynamics can change rapidly.

Neural forecasting models have improved representation learning through recurrent, probabilistic, and attention-based architectures [4–7]. However, many pipelines still use decoder mappings that are fixed after training and shared across heterogeneous market conditions [6,8,9]. This creates a structural mismatch: representation modules are increasingly adaptive, while the prediction rule remains regime-agnostic.

The working hypothesis in this study is that robustness depends not only on representational capacity but also on where adaptation is introduced. Early regime injection can entangle shared representations with short-horizon fluctuations; by contrast, decode-time conditioning can preserve reusable encodings while allowing the prediction rule to adapt locally.

Formally, let $\mathbf{X}_{t-T+1:t} \in \mathbb{R}^{T \times F}$ denote the input window and $\mathbf{y}_{t+1:t+H}$ the forecasting horizon. Standard approaches learn a global mapping

$$\hat{\mathbf{y}}_{t+1:t+H} = f_\theta(\mathbf{X}_{t-T+1:t}), \tag{1}$$

 doi:10.20944/preprints202604.0884.v1

which implicitly assumes that a single decoder is sufficient across heterogeneous regimes. The proposed formulation instead conditions prediction on deterministic, causal regime descriptors:

$$\hat{\mathbf{y}}_{t+1:t+H} = g_\psi\big(\mathrm{Enc}_\phi(\mathbf{X}_{t-T+1:t}),\, y_t,\, \mathbf{r}_t\big), \tag{2}$$

where $\mathbf{r}_t \in \mathbb{R}^4$ is a compact regime vector computed from one-step normalized target returns within the input window. This formulation isolates representation learning from conditional synthesis and enables direct testing of decoder-centric adaptation.

Equation 1 defines the baseline decoder-agnostic mapping, whereas Equation 2 defines the decoder-conditioned formulation used in RADIAN.

The resulting architecture, RADIAN, applies late conditioning: regime statistics are injected only in the fusion and decoding modules, while temporal and cross-variable encoders remain regime-agnostic. The conditioning signal is deterministic and causal because it is computed only from in-window target dynamics.

Empirical evaluation on nine hourly financial datasets spanning equities, foreign exchange, cryptocurrencies, indices, and commodities, under a fixed protocol against eight baselines, shows that RADIAN attains the lowest mean test MAE across datasets, with an average improvement of 0.6704% relative to the strongest per-dataset baseline. Statistical testing reports multiple dataset-level wins under corrected thresholds, and mechanism-level ablations attribute the dominant gains to decoder-side components.

Contributions.

- **Decoder-centric forecasting paradigm.** The paper formalizes adaptation at the synthesis stage rather than within shared representations.
- **Deterministic causal regime conditioning.** The model introduces a compact regime descriptor derived solely from in-window target dynamics, ensuring causal validity and interpretability.
- **Mechanism-level empirical validation.** Ablations isolate the contribution of decoder-side components.
- **Rigorous statistical evaluation.** The evaluation includes significance testing with multiple-hypothesis correction.

The remainder of the paper is organized as follows. Section 2 reviews prior work in statistical, neural, and regime-aware forecasting. Section 3 presents the full model specification. Section 4 describes the datasets and evaluation protocol. Section 5 reports quantitative and qualitative results. Section 6 discusses implications and limitations, and Section 7 concludes.

2. Related Work

This section reviews prior work in six parts, progressing from statistical regime modeling and neural forecasting architectures to Transformer-based forecasting, non-stationarity-focused methods, conditional computation, and the gap addressed by this study.

2.1. Statistical Foundations and Regime Dynamics

Classical forecasting is grounded in autoregressive integrated moving average (ARIMA), vector autoregression (VAR), and state-space formulations [10–12]. In financial settings, ARCH/GARCH families provide parametric volatility adaptation [13,14], and regime-switching models capture structural transitions through latent-state dynamics [3].

The stylized-facts literature reports heavy tails, volatility clustering, and distributional instability in asset returns [1,2,15]. These findings motivate adaptation, but they do not specify where adaptation should be introduced in modern neural encoder–decoder pipelines.

 doi:10.20944/preprints202604.0884.v1

2.2. Neural Time-Series Forecasting

Neural forecasting broadened the modeling space from recurrent and attention-based models to convolutional sequence architectures. Representative developments include encoder–decoder attention models, hybrid long/short-horizon designs, and causal convolutional backbones [16–18].

Probabilistic global models further integrated cross-series learning with stochastic objectives [5]. Feedforward decomposition models and covariate-aware attention architectures improved long-horizon interpretability and multi-source fusion [6,19]. Across these lines, the emphasis remains on sequence encoding rather than decoder adaptation.

2.3. Transformer-Based Forecasting Architectures

Following Vaswani et al. [7], Transformer variants for forecasting diversified along tokenization, sparsification, and decomposition axes.

Point-wise attention models reduce quadratic cost through sparse attention or autocorrelation-style decomposition [20,21]. Frequency-enhanced decomposition variants extend this line to long-range components [22].

Locality-aware and representation-enhanced variants constrain attention neighborhoods or strengthen input representations [23].

Comparative analyses also show that performance conclusions are sensitive to benchmark design and inductive-bias choices [24].

Across these families, architectural changes are concentrated in the encoder and tokenization stack, while forecast heads are often lightweight projections with limited explicit regime conditioning.

2.4. Handling Non-Stationarity and Distribution Shift

Several methods target distribution shift directly. Reversible instance normalization (RevIN) mitigates scale and location shifts [25], and Non-stationary Transformers introduce de-stationary transformations for evolving temporal patterns [26].

These methods improve robustness but couple adaptation with representation changes, making it difficult to attribute gains cleanly to normalization, encoding, or decoding.

2.5. Conditional Computation and Adaptive Modeling

Conditional computation provides a principled mechanism for input-dependent adaptation. Jacobs [27] introduced adaptive mixtures of experts (MoE), and modern sparse MoE layers scale this routing principle to larger parameter sets [28,29].

For forecasting, this suggests conditioning prediction functions on local regime descriptors while keeping a shared encoder backbone.

2.6. Positioning and Gap

Prior work shows gains from stronger sequence encoders and non-stationarity-aware processing [8,9,25,26]. However, these gains are typically realized through encoder design, tokenization, or normalization, leaving decoder-side effects hard to isolate.

Accordingly, this paper evaluates the hypothesis that robustness to regime shift can be improved by conditioning the decoder while keeping the encoder regime-agnostic. Restricting adaptation to the synthesis stage with deterministic causal regime descriptors enables direct attribution of improvements to decode-time conditioning.

3. Method

This section first defines the forecasting problem and notation, then details input normalization and deterministic regime extraction, followed by the temporal and cross-variable encoders, decoder-side fusion and output generation, and finally the forward algorithm and complexity analysis.

 doi:10.20944/preprints202604.0884.v1

3.1. Problem Formulation

Given a batch of multivariate history windows $\mathbf{X} \in \mathbb{R}^{B\times T\times F}$, RADIAN learns a forecasting map

$$\mathcal{F}_\theta : \mathbf{X} \mapsto \hat{\mathbf{y}} \in \mathbb{R}^{B\times H}, \tag{3}$$

Equation 3 defines the end-to-end map from history windows to horizon forecasts. For each sample b, the target history is $\mathbf{y}_b = \mathbf{X}_{b,:,f^*} \in \mathbb{R}^T$, and the per-sample forecast is $\hat{\mathbf{y}}_b = [\hat{y}_{b,1}, \ldots, \hat{y}_{b,H}]^\top \in \mathbb{R}^H$. Here b, t, and f index batch, time, and feature dimensions, respectively; B is the batch size, T the input length, F the number of features, D the model width, and H the forecast horizon. In the released experiments, $T = 24$, $H = 4$, $F = 13$, and the target feature is `Close`. The default pipeline normalizes $\mathbf{X}$, extracts deterministic regime features from $\mathbf{y}_b^{(\text{norm})}$, decodes $\hat{\mathbf{y}}_b^{(\text{norm})}$, and maps the result back to the original scale. Table 1 summarizes the symbols used throughout the section and aligns them with the equations and implementation entities.

3.2. Overview and Design Principle

RADIAN maps $\mathbf{X} \in \mathbb{R}^{B\times T\times F}$ to $\hat{\mathbf{y}} \in \mathbb{R}^{B\times H}$. The guiding principle is an explicit separation of responsibilities: encoder branches are regime-agnostic and learn reusable representations, whereas adaptation is confined to decode-time modules (`GatedFusion` and `RegimeAwareDirectionalHead`). Figure 1 shows the full computation graph and the decoder-only regime-injection point.

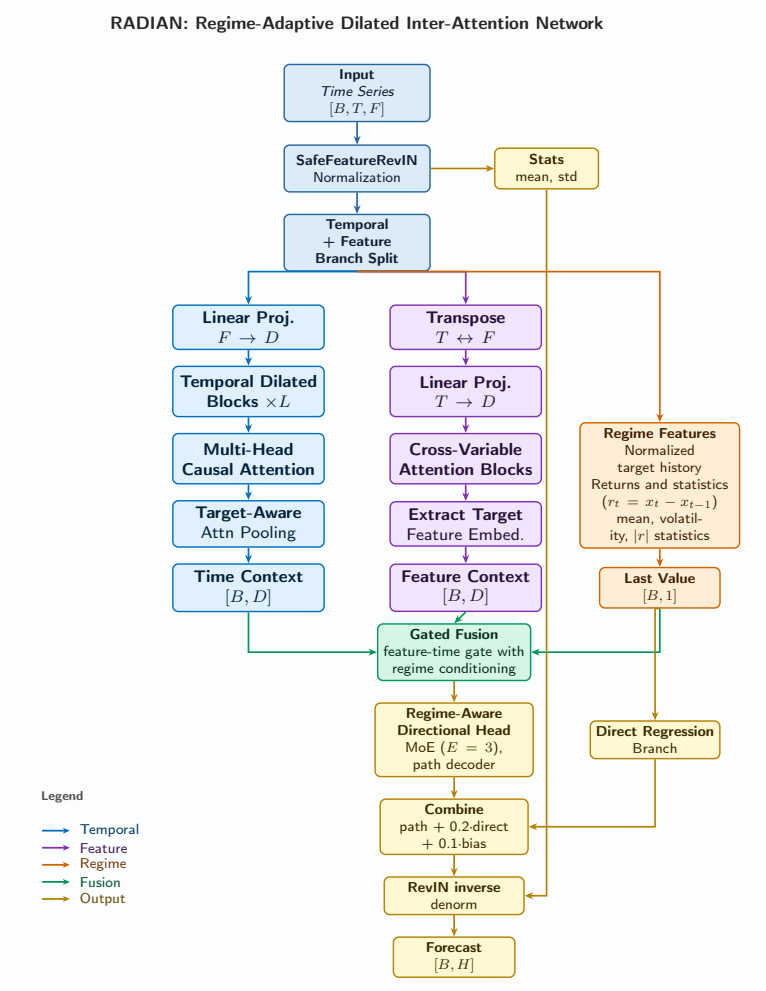


Figure 1. Architectural diagram of the proposed RADIAN model. The framework consists of three processing branches—temporal, feature, and regime—followed by gated fusion and a regime-aware directional head to produce multi-step forecasts.

 doi:10.20944/preprints202604.0884.v1

3.3. Input Normalization: SafeFeatureRevIN

For each sample b and feature f, SafeFeatureRevIN performs instance-wise normalization in the spirit of reversible normalization approaches for distribution shift handling [25]. With affine mode enabled, the learned feature-wise scale γ_f and shift β_f are applied after standardization:

$$\mu_{b,f} = \frac{1}{T}\sum_{t=1}^{T} X_{b,t,f}, \tag{4}$$

$$\sigma_{b,f} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(X_{b,t,f} - \mu_{b,f})^2 + \varepsilon}, \ \varepsilon = 10^{-5}, \tag{5}$$

$$\tilde{X}_{b,t,f} = \frac{X_{b,t,f} - \mu_{b,f}}{\sigma_{b,f}}, \tag{6}$$

$$X^{(\mathrm{norm})}_{b,t,f} = \gamma_f \tilde{X}_{b,t,f} + \beta_f. \tag{7}$$

The module stores $\mathbf{S}_b \in \mathbb{R}^{2\times F}$, where $\mathbf{S}_b[0,f] = \mu_{b,f}$ and $\mathbf{S}_b[1,f] = \sigma_{b,f}$.

This stage reduces sample-specific scale mismatch before the temporal and feature branches, allowing them to focus on shape and interaction patterns rather than raw magnitude. The stored statistics preserve exact per-sample inversion for the target channel.

3.4. Deterministic Regime Vector Extraction

Let $\mathbf{y}_b^{(\mathrm{norm})} = X^{(\mathrm{norm})}_{b,:,f^*} \in \mathbb{R}^T$. The causal one-step returns are

$$\rho_{b,t} = y^{(\mathrm{norm})}_{b,t} - y^{(\mathrm{norm})}_{b,t-1}, \quad t = 2,\ldots,T. \tag{8}$$

The return sequence underlies the regime statistics. The regime-feature extractor computes

$$\bar{\rho}_b = \frac{1}{T-1}\sum_{t=2}^{T}\rho_{b,t},$$

$$s_b = \sqrt{\frac{1}{T-1}\sum_{t=2}^{T}(\rho_{b,t} - \bar{\rho}_b)^2},$$

$$m_b = \frac{1}{T-1}\sum_{t=2}^{T}|\rho_{b,t}|,$$

$$M_b = \max_{t=2,\ldots,T}|\rho_{b,t}|,$$

and concatenates

$$\mathbf{r}_b = [\bar{\rho}_b, s_b, m_b, M_b]^\top \in \mathbb{R}^4. \tag{9}$$

This four-dimensional vector serves as the decoder conditioning signal; no latent regime labels, clustering, or future-data signals are used.

Figure 2 illustrates the extracted components across calm, transition, and volatile market windows.

 doi:10.20944/preprints202604.0884.v1

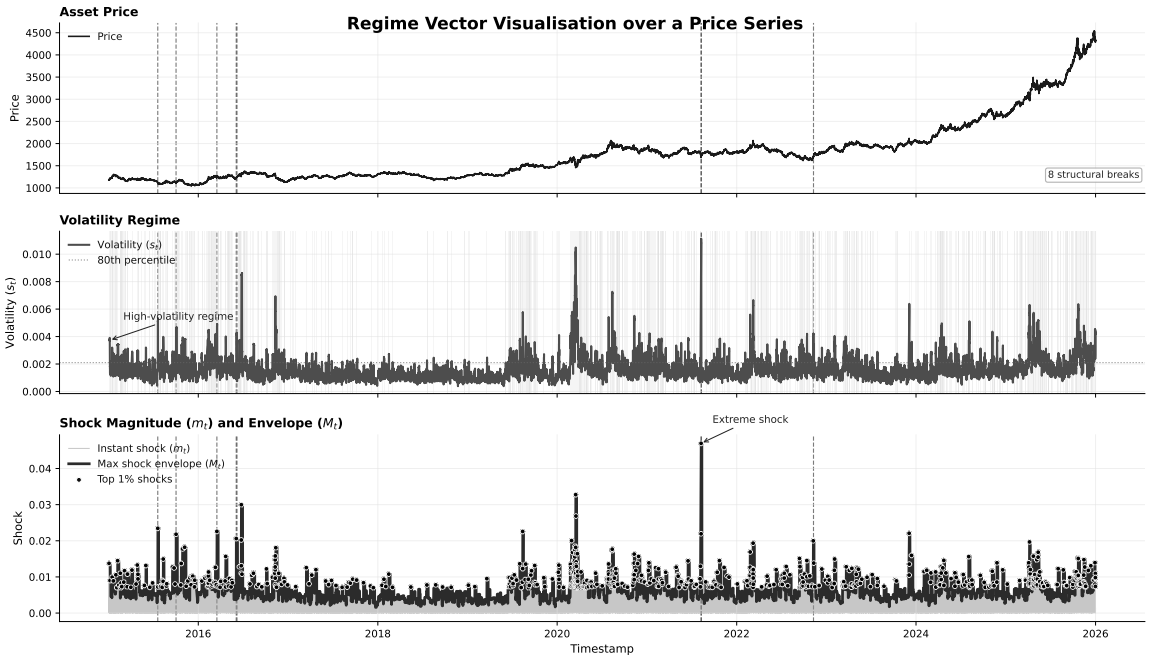


Figure 2. Regime vector decomposition of the price process. The plot illustrates the four components of the regime vector $\mathbf{r}_b = [\bar{\rho}_b, s_b, m_b, M_b]^\top$, with s_b (volatility) in the middle panel and m_b, M_b (shock magnitude and envelope) in the lower panel. The asset price evolution is shown above, and vertical markers denote structural breakpoints.

3.5. Temporal Branch: Temporal Dilated Block (TDB) + Causal Attention

The temporal branch first projects each time step from $\mathbb{R}^F$ to $\mathbb{R}^D$. It combines causal dilated convolutions [18] with pre-norm LayerNorm [30] and masked self-attention [7]:

$$\mathbf{H}^{(0)}_{b,t,:} = \mathbf{W}_{\text{time}} \mathbf{X}^{(\text{norm})}_{b,t,:} + \mathbf{b}_{\text{time}}. \tag{10}$$

Each `TemporalDilatedBlock` applies pre-norm, depthwise dilated causal convolution, a gated pointwise transformation, and a residual update:

$$\begin{aligned}
\mathbf{Z}^{(l)} &= \text{LayerNorm}(\mathbf{H}^{(l)}), \\
\mathbf{C}^{(l)} &= \text{DWConv1D}_{\text{causal}}(\mathbf{Z}^{(l)}), \\
[\mathbf{U}^{(l)}, \mathbf{G}^{(l)}] &= \text{Linear}_{D \to 2D}(\mathbf{C}^{(l)}), \\
\mathbf{F}^{(l)} &= \mathbf{U}^{(l)} \odot \sigma(\mathbf{G}^{(l)}), \\
\mathbf{H}^{(l+1)} &= \mathbf{H}^{(l)} + \text{Dropout}(\text{Linear}_{D \to D}(\mathbf{F}^{(l)})).
\end{aligned}$$

Left-only padding enforces causality, and masked self-attention is applied after the L temporal blocks:

$$\mathbf{A} = \text{MHA}(\text{LN}(\mathbf{H}^{(L)}), \text{LN}(\mathbf{H}^{(L)}), \text{LN}(\mathbf{H}^{(L)}); \mathbf{M}_{\text{causal}}), \tag{11}$$

$$\mathbf{H}^{\text{time}} = \mathbf{H}^{(L)} + \mathbf{A} \in \mathbb{R}^{B \times T \times D}. \tag{12}$$

This branch captures local multi-scale dynamics through causal convolutions and extends the receptive field with masked attention without future leakage. Pre-norm LayerNorm and residual updates improve optimization stability under non-stationary inputs. Equations 10–12 define the temporal encoder output passed to pooling.

 doi:10.20944/preprints202604.0884.v1

3.6. Cross-Variable Branch: Cross-Variable Attention Block (CVAB)

To form feature tokens, the normalized tensor is transposed over its last two axes to $\mathbf{X}^{(\mathrm{norm})\top} \in \mathbb{R}^{B\times F\times T}$ and then projected:

$$\mathbf{Z}^{(0)}_{b,f,:} = \mathbf{W}_{\mathrm{feat}}\mathbf{X}^{(\mathrm{norm})}_{b,:,f} + \mathbf{b}_{\mathrm{feat}}. \tag{13}$$

Each CVAB applies pre-norm LayerNorm [30], feature-token self-attention [7], and a residual MLP refinement,

$$\begin{aligned}
\tilde{\mathbf{Z}}^{(m)} &= \mathrm{LN}(\mathbf{Z}^{(m)}),\\
\mathbf{A}^{(m)} &= \mathrm{MHA}(\tilde{\mathbf{Z}}^{(m)}, \tilde{\mathbf{Z}}^{(m)}, \tilde{\mathbf{Z}}^{(m)}),\\
\mathbf{Z}'^{(m)} &= \mathbf{Z}^{(m)} + \mathbf{A}^{(m)},\\
\mathbf{Z}^{(m+1)} &= \mathbf{Z}'^{(m)} + \mathrm{FFN}(\mathbf{Z}'^{(m)}),
\end{aligned}$$

where $\mathrm{FFN}(\cdot)$ denotes the position-wise LayerNorm–GELU–dropout MLP sublayer. The final feature context is given by

$$\mathbf{c}_{\mathrm{feat}} = \mathbf{Z}^{(M)}_{:,f^*,:} \in \mathbb{R}^{B\times D}. \tag{14}$$

In the released configuration, $L = 3$, so $M = \max(1, \lfloor L/2 \rfloor) = 1$.

The resulting feature context summarizes cross-variable interactions around the target channel rather than collapsing variables too early. This complementary signal helps the fusion stage correct purely temporal forecasts when inter-series dependencies matter. Equations 13–14 define the cross-variable branch.

3.7. Target-Aware Pooling and Gated Fusion (GF)

Target-aware temporal pooling constructs a query conditioned on the most recent normalized target and uses it to form an attention-weighted summary of the temporal encoder states, following the scaled dot-product mechanism of Vaswani et al. [7]. Formally, for each sample b, a query vector is obtained as

$$\mathbf{q} * b = \mathbf{W} * q\mathbf{y}^{(\mathrm{norm})} * b + \mathbf{b} * q, \qquad \alpha * b, t = \mathrm{softmax} * t!\left(\frac{\langle \mathbf{H}^{\mathrm{time}} * b, t, :, \mathbf{q} * b\rangle}{\sqrt{D}}\right), \tag{15}$$

where $\alpha * b, t$ denotes the normalized attention weight over the temporal index t. The resulting context vector is computed as a convex combination of temporal features:

$$\mathbf{c} * \mathrm{time}, b = \sum * t = 1^T \alpha * b, t\mathbf{H}^{\mathrm{time}}_{b,t,:}. \tag{16}$$

To integrate complementary information sources, we form a joint representation by concatenating the temporal context $\mathbf{c} * \mathrm{time}, b$, the feature-wise context $\mathbf{c} * \mathrm{feat}, b$, the most recent normalized target value $y^{(\mathrm{norm})} * b, T$, and the regime descriptor $\mathbf{r} * b$:

$$\mathbf{z} * b = [\mathbf{c} * \mathrm{time}, b; \mathbf{c} * \mathrm{feat}, b; y^{(\mathrm{norm})} * b, T; \mathbf{r}_b] \in \mathbb{R}^{2D+5}. \tag{17}$$

Gated Fusion (GF) then adaptively combines these signals by computing a sigmoid gate and a candidate projection:

$$\mathbf{g}_b = \sigma(\mathrm{MLP}_g(\mathbf{z}_b)), \quad \mathbf{u}_b = \mathrm{MLP}_p(\mathbf{z}_b), \tag{18}$$

and producing the final fused representation via element-wise interpolation:

$$\mathbf{c}_b = \mathbf{g}_b \odot \mathbf{u}_b + (1 - \mathbf{g} * b) \odot \mathbf{c} * \mathrm{time}, b. \tag{19}$$

Equations 15–19 collectively specify the full target-aware aggregation pipeline: the construction of a target-conditioned query, the derivation of attention weights and pooled temporal context, the

 doi:10.20944/preprints202604.0884.v1

formation of the composite fusion input, and the subsequent gated interpolation that yields the final context representation.

Operationally, `GatedFusion` performs context-dependent soft selection between the transformed joint context and the temporal anchor, conditioned on $y_{b,T}^{(\text{norm})}$ and regime statistics. This preserves a stable temporal reference while still allowing the decoder to amplify informative joint-context features.

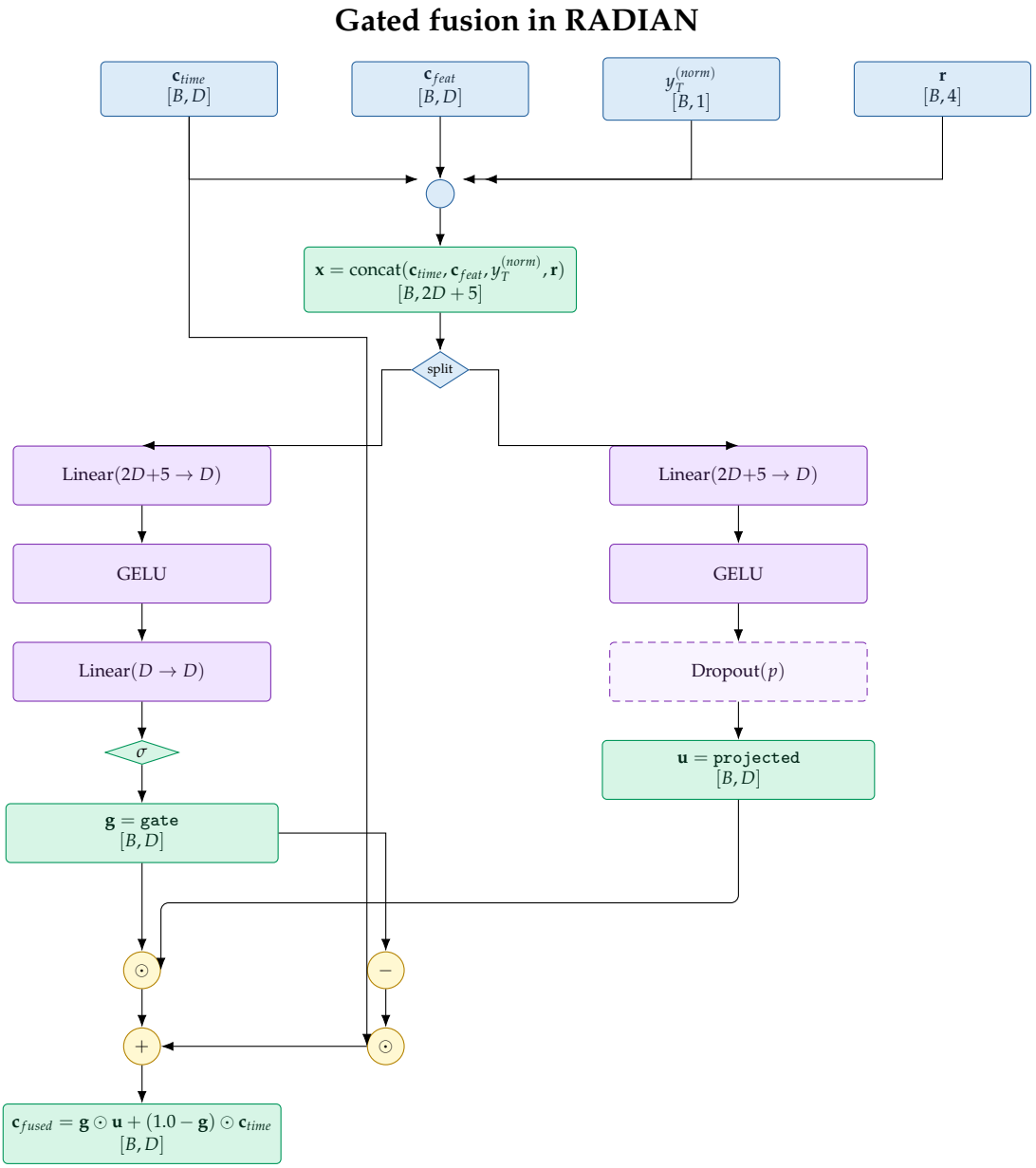


Figure 3. Illustration of the gated residual fusion mechanism. The figure shows how RADIAN fuses multiple conditioning signals: the gate branch learns which parts of the projected representation $\mathbf{u}$ to retain, and the residual connection preserves the original temporal context $\mathbf{c}_{\text{time}}$ where the gate is close to zero. This design allows dynamic, regime-dependent feature selection.

3.8. Decoder: Regime-Aware Directional Head (RADH)

The decoder operates on a regime-conditioned representation by concatenating the fused context, the most recent normalized target value, and the regime descriptor into a single input vector. Formally, the decoder state is defined in Equation 20—the decoder input definition—as

$$\mathbf{h}_b = [\mathbf{c}_b; y_{b,T}^{(\text{norm})}; \mathbf{r}_b] \in \mathbb{R}^{D+5}. \tag{20}$$

 doi:10.20944/preprints202604.0884.v1

Conditional computation is realized through a mixture-of-experts (MoE) mechanism, wherein a gating network produces a probability simplex over E experts. The routing weights are computed according to Equation 21—the MoE gating distribution—

$$\boldsymbol{\pi}_b = \mathrm{softmax}(\mathrm{MLP}_{\mathrm{gate}}(\mathbf{h}_b)) \in \mathbb{R}^E, \tag{21}$$

which determines the contribution of each expert to the aggregated output. The expert outputs are then combined as a convex mixture, as specified in Equation 22—the MoE aggregation rule—

$$\mathbf{o}_b = \sum_{e=1}^{E} \pi_{b,e} \mathbf{o}_{e,b}, \quad \mathbf{o}_{e,b} \in \mathbb{R}^{3H}. \tag{22}$$

The aggregated vector $\mathbf{o}_b$ is partitioned into three components, $(\mathbf{s}_b^{\mathrm{raw}}, \mathbf{a}_b^{\mathrm{raw}}, \boldsymbol{\ell}_b^{\mathrm{lvl}})$, each in $\mathbb{R}^H$, which parameterize the directional forecasting mechanism. A non-negative scaling factor is obtained via $\mathbf{s}_b^{\mathrm{scale}} = \mathrm{softplus}(\mathbf{a}_b^{\mathrm{raw}})$, ensuring stable magnitude modulation. Directional information is incorporated by forming $\mathbf{s}_b^{\mathrm{dir}} = \tanh(\mathbf{s}_b^{\mathrm{raw}}) \odot \mathbf{s}_b^{\mathrm{scale}}$, which constrains the signal while preserving sign structure. The final increment signal is then defined as a convex combination $\mathbf{s}_b^{\mathrm{blend}} = (1-\lambda)\mathbf{s}_b^{\mathrm{raw}} + \lambda \mathbf{s}_b^{\mathrm{dir}}$, where $\lambda = \texttt{directional\_mix} = 0.35$ controls the strength of directional regularization.

Forecast generation proceeds through a path-based decoding process that constructs the output trajectory cumulatively. Starting from the last observed normalized target, the trajectory evolves according to Equation 23—the path decoding recursion—

$$p_{b,0} = y_{b,T}^{(\mathrm{norm})}, \qquad p_{b,h} = p_{b,h-1} + s_{b,h}^{\mathrm{blend}}, \quad h = 1, \ldots, H, \tag{23}$$

thereby producing a sequence of intermediate states. The predicted trajectory is collected as $\mathbf{p}_b = [p_{b,1}, \ldots, p_{b,H}]^\top \in \mathbb{R}^H$.

To account for local deviations not captured by the structured trajectory, a residual correction branch is introduced. This branch computes a direct adjustment term as defined in Equation 24—the direct correction mapping—

$$\mathbf{d}_b = \mathbf{W}_{\mathrm{direct}}[\mathbf{c}_b; y_{b,T}^{(\mathrm{norm})}] + \mathbf{b}_{\mathrm{direct}}, \tag{24}$$

which is subsequently integrated with the trajectory and a learned level component to produce the final normalized forecast. The complete output is given by Equation 25—the normalized prediction rule—

$$\hat{\mathbf{y}}_b^{(\mathrm{norm})} = \mathbf{p}_b + 0.2\mathbf{d}_b + 0.1\boldsymbol{\ell}_b^{\mathrm{lvl}}. \tag{25}$$

In the default configuration, the decoder employs $E = 3$ experts. Collectively, Equations 20 through 25 define a unified decoding framework that integrates regime-aware conditioning, expert specialization via MoE routing, directionally constrained increment modeling, and path-based trajectory construction. The inclusion of a residual correction term further enhances flexibility by allowing localized adjustments without compromising the global structure of the predicted path.

 doi:10.20944/preprints202604.0884.v1

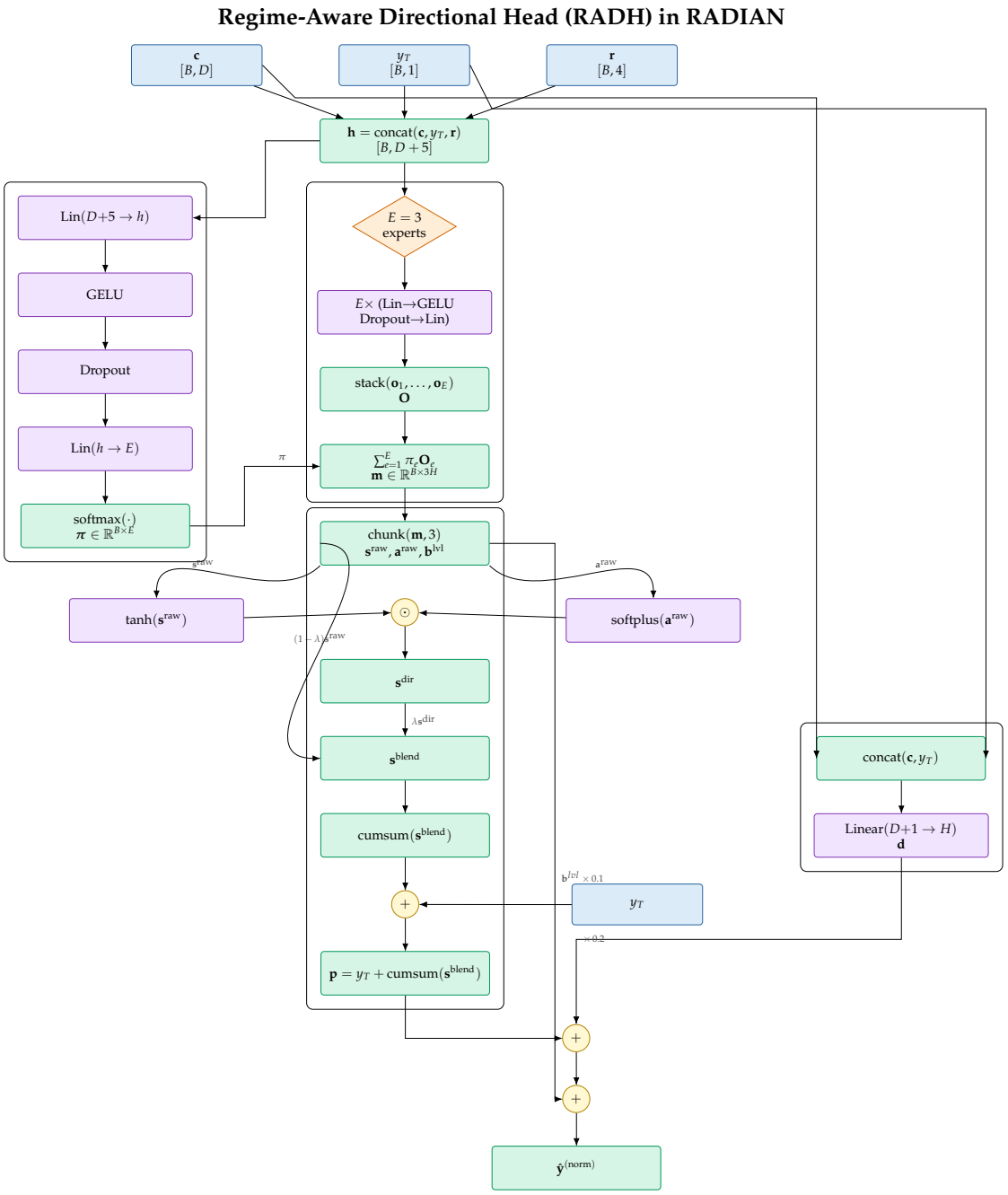


Figure 4. Regime-Aware Directional Head (RADH) architecture illustrating the integration of contextual representations, the most recent observation, and regime descriptors through a gated mixture-of-experts mechanism to produce normalized multi-step forecasts.

3.9. Target Denormalization

For the target feature $f^* =$ `target_idx`:

$$\hat{y}_{b,h} = \left(\frac{\hat{y}^{(\text{norm})}_{b,h} - \beta_{f^*}}{\gamma_{f^*}} \right) \sigma_{b,f^*} + \mu_{b,f^*}. \tag{26}$$

The implementation checks the output shape (B, H) and finite values, catching numerical issues early and preserving the forecast scale expected by downstream evaluation.

All reported error metrics are computed after this inverse transform in the original target scale.

 doi:10.20944/preprints202604.0884.v1

3.10. Notation Summary

Table 1. Core notation used in the RADIAN formulation. Symbols are aligned with implementation-level entities so that equations and complexity terms remain consistent.

Symbol	Meaning
B, T, F, D, H	batch size, window length, #features, model width, horizon
f^*	target feature index (`target_idx`)
$\mathbf{X}, \mathbf{X}^{(\text{norm})}$	raw and normalized inputs
$\mathbf{S}_b$	stored RevIN statistics (mean/std)
$\mathbf{r}_b \in \mathbb{R}^4$	deterministic regime vector
$\mathbf{H}^{\text{time}}$	temporal-branch sequence representation
$\mathbf{c}_{\text{time},b}, \mathbf{c}_{\text{feat},b}, \mathbf{u}_b, \mathbf{c}_b$	pooled temporal context, feature context, projected fusion context, fused context
$\mathbf{p}_b$	path-decoded trajectory vector before direct correction
$\hat{\mathbf{y}}_b^{(\text{norm})}, \hat{\mathbf{y}}_b$	normalized and denormalized forecasts

3.11. Forward Algorithm

Algorithm 1 summarizes the modular data flow of the RADIAN model, with each component operating in sequence from normalization to denormalization. This makes the implementation transparent enough to map directly onto the equations above.

Algorithm 1 RADIAN forward pass

Require: $\mathbf{X} \in \mathbb{R}^{B\times T\times F}$, target index f^*

$(\mathbf{X}^{(\text{norm})}, \mathbf{S}) \leftarrow \texttt{SafeFeatureRevIN}(\mathbf{X})$
$\mathbf{y}^{(\text{norm})} \leftarrow \mathbf{X}^{(\text{norm})}_{:,:,f^*}$
$\mathbf{r} \leftarrow \texttt{\_extract\_regime\_features}(\mathbf{y}^{(\text{norm})})$
$\mathbf{H}^{\text{time}} \leftarrow \texttt{TemporalBranch}(\mathbf{X}^{(\text{norm})})$
$\mathbf{Z}^{\text{feat}} \leftarrow \texttt{FeatureBranch}(\mathbf{X}^{(\text{norm})\top})$
$\mathbf{c}_{\text{time}} \leftarrow \texttt{TargetAwarePool}(\mathbf{H}^{\text{time}}, \mathbf{y}^{(\text{norm})})$
$\mathbf{c}_{\text{feat}} \leftarrow \mathbf{Z}^{\text{feat}}_{:,f^*,:}$
$\mathbf{y}_{\text{last}}^{(\text{norm})} \leftarrow \mathbf{y}^{(\text{norm})}_{:,-1}$
$\mathbf{c} \leftarrow \texttt{GatedFusion}(\mathbf{c}_{\text{time}}, \mathbf{c}_{\text{feat}}, \mathbf{y}_{\text{last}}^{(\text{norm})}, \mathbf{r})$
$\hat{\mathbf{y}}^{(\text{norm})} \leftarrow \texttt{RegimeAwareDirectionalHead}(\mathbf{c}, \mathbf{y}_{\text{last}}^{(\text{norm})}, \mathbf{r})$
$\hat{\mathbf{y}} \leftarrow \texttt{denorm\_target}(\hat{\mathbf{y}}^{(\text{norm})}, \mathbf{S}, f^*)$
return $\hat{\mathbf{y}}$

3.12. Complexity Analysis

Let L and M denote the numbers of temporal and feature blocks, K the kernel size, E the number of experts, and h the decoder hidden size.

Dominant FLOPs.

$$\mathcal{O}(BTFD) + \mathcal{O}\big(BLT(KD + D^2)\big) + \mathcal{O}(BT^2D) + \mathcal{O}(BFTD) + \mathcal{O}(BMF^2D) + \mathcal{O}\big(B(E(D{+}5)h + Eh(3H))\big). \quad (27)$$

Decoder parameter scaling.

$$\#\theta_{\text{dec}} \approx (D+5)h + hE + E\big((D+5)h + 3Hh\big) + (D+1)H, \quad (28)$$

which is linear in expert count E. With released defaults (D=128, E=3, H=4), the decoder contributes a limited additional parameter cost relative to the full model (Table 10). Equations 27 and 28 summarize the dominant FLOPs and decoder parameter scaling, respectively.

 doi:10.20944/preprints202604.0884.v1

4. Experimental Setup

This section documents the datasets and split policy, shared training protocol, baseline set, evaluation metrics, statistical testing pipeline, robustness diagnostics, and reproducibility constraints.

4.1. Datasets, Features, and Splits

Evaluation is conducted on nine hourly financial datasets: AAPL, AUDUSD, BTCUSD, DAX30, ETHUSD, FTSE100, US500, USDJPY, and XAUUSD. Each dataset follows the raw schema `Datetime, Open, High, Low, Close, Volume`.

To ensure reproducibility and to prevent data leakage, a fixed chronological split is applied: 80% training, 10% validation, and 10% test. This split is consistent across all models, thereby preserving temporal ordering and preventing contamination between training and test sets.

Input sequences use a window length of $T = 24$ and a forecast horizon of $H = 4$. The feature set comprises 13 channels:

- Raw prices: `Open, High, Low, Close, Volume`
- Technical indicators: `macd, macd_signal, macd_hist, rsi_14, ema_9, bb_middle, bb_upper, bb_lower`

The target variable is `Close` (`target_idx=3`).

Additional details on data scope, provenance, and raw file statistics are provided in Appendix A (Table A1 and Table A2).

4.2. Training Protocol

All models are trained using Adam [31] with an MSE objective, a learning rate of 5×10^{-4}, a batch size of 64, and 50 epochs. The seed set is fixed at {417, 153, 999}. Deterministic execution is enforced through global seeding, deterministic CuDNN settings, and deterministic algorithm selection when available.

Table 2 summarizes the key hyperparameters required to reproduce the reported results in the main pipeline. All core experimental controls are shared and fixed across models, thereby ensuring that differences reported in Section 5 are attributable to architectural variation rather than disparities in training budget.

Table 2. Primary RADIAN hyperparameters used in all reported experiments. The table lists fixed settings shared across datasets.

Setting	Value
Input window T	24
Forecast horizon H	4
Target feature	`close`
Loss	MSE
Optimizer	Adam
Learning rate	5×10^{-4}
Batch size	64
Epochs	50
Seeds	{417, 153, 999}
RADIAN width D	128
Attention heads	8
Temporal layers L	3
Experts E	3
Directional mix λ	0.35

Checkpointing stores the model corresponding to the best validation RMSE. The released training procedure does not employ patience-based early stopping; each run executes the full epoch budget, after which the best-performing checkpoint is restored.

 doi:10.20944/preprints202604.0884.v1

4.3. Baselines

RADIAN is compared against eight implemented baselines: DLinear [24], LSTM [4], iTransformer [9], PatchTST [8], TimesNet [32], N-HiTS [33], TimeXer [34], and ModernTCN.

All models use identical data splits, seeds, and optimizer family; model-specific hyperparameters are listed in Appendix B.

The baseline set spans linear projection, recurrent memory, and multiple transformer tokenization strategies [4,8,9,24], enabling an assessment of whether decoder-centric adaptation provides consistent benefits across architectural families.

4.4. Evaluation Metrics

Primary metrics are test MSE, RMSE, MAE, and directional accuracy (DA) [12], where $i = 1, \ldots, N$ indexes test windows and N denotes the number of test samples:

$$\text{MSE} = \frac{1}{NH} \sum_{i=1}^{N} \sum_{h=1}^{H} (\hat{y}_{i,h} - y_{i,h})^2, \tag{29}$$

$$\text{RMSE} = \sqrt{\text{MSE}}, \tag{30}$$

$$\text{MAE} = \frac{1}{NH} \sum_{i=1}^{N} \sum_{h=1}^{H} |\hat{y}_{i,h} - y_{i,h}|, \tag{31}$$

$$\text{DA} = \frac{1}{NH} \sum_{i=1}^{N} \sum_{h=1}^{H} \mathbf{1}[\text{sign}(\hat{y}_{i,h} - y_{i,h-1}) = \text{sign}(y_{i,h} - y_{i,h-1})], \tag{32}$$

Equations 29, 30, 31, and 32 define the reported metrics. Here, $y_{i,0}$ denotes the last observed target value in the input window, and for $h > 1$, $y_{i,h-1}$ denotes the prior realized value in the forecast horizon. MSE and RMSE emphasize larger deviations, MAE measures central absolute error, and DA measures sign consistency of predicted and realized increments.

4.5. Statistical Testing and Multiple Comparisons

For each dataset, the released main results tables provide paired p-values for RADIAN versus the best baseline. Both raw p-values and Benjamini–Hochberg (BH) adjusted p-values over 9 tests are reported. Effect size is computed using paired Cohen's d based on seed-level MAE differences in the detailed experiment records, where $\overline{\Delta}$ is the mean paired MAE difference across seeds and s_Δ is the corresponding sample standard deviation:

$$d = \frac{\overline{\Delta}}{s_\Delta}, \qquad \Delta_s = \text{MAE}_{\text{RADIAN},s} - \text{MAE}_{\text{baseline},s}. \tag{33}$$

Equation 33 defines the effect-size calculation. The BH adjustment controls the false discovery rate under multiple comparisons [35]. Paired effect sizes complement p-values by quantifying the magnitude of differences [36]. Full per-dataset statistics, including raw and adjusted p-values and Cohen's d, are reported in Table A18.

4.6. Robustness and Ablation Diagnostics

The robustness analysis uses released robustness artifacts together with ablation logs (A0–A9) from the published ablation summary. In ablation experiments, each switch corresponds exactly to the released configuration; no architectural modifications beyond these controlled flags are introduced.

This one-switch-at-a-time protocol isolates the effect of each architectural component.

4.7. Reproducibility Constraints

Method descriptions in this manuscript are restricted to the released model modules: `SafeFeatureRevIN`, `TemporalDilatedBlock`, `CrossVariableAttentionBlock`, `GatedFusion`, and `RegimeAwareDirectionalHead`.

 doi:10.20944/preprints202604.0884.v1

Appendix Sections A–F provide dataset inventories, hyperparameter tables, extended result logs, and the reproducibility checklist used to maintain alignment with released artifacts.

5. Results

This section reports benchmark accuracy, regime-stratified robustness, ablation outcomes, directional accuracy, and efficiency trade-offs. The presentation follows the order of Tables 3 and **??**, moving from accuracy to robustness, ablation, and computational cost.

5.1. Main Accuracy and Statistical Strength

As quantified in Table 3, RADIAN achieves the lowest root mean square error (RMSE) across all nine evaluated datasets—explicitly outperforming the strongest baseline, ModernTCN—while simultaneously securing the highest directional accuracy (DA) on six heterogeneous asset classes. The benchmark spans multiple asset classes with diverse volatility profiles to rigorously validate this performance edge.

Table 3. Dataset-level RMSE (↓) and directional accuracy (DA, %, ↑) for six models across nine financial datasets. Bold values indicate the best result per row/metric; underlined values indicate the second-best result.

	RADIAN		ModernTCN		PatchTST		TimeXer		iTransformer		N-HiTS	
Dataset	RMSE	DA	RMSE	DA	RMSE	DA	RMSE	DA	RMSE	DA	RMSE	DA
AAPL	**2.43452**	**52.0399**	2.47554	50.4395	2.54399	50.6782	2.68699	51.0796	2.68113	50.6944	2.47771	51.0796
AUDUSD	**0.0012**	51.2706	0.0012	51.0947	0.0012	50.6322	0.0012	51.1427	0.0013	**51.3198**	0.0013	51.2079
BTCUSD	**661.786**	51.5112	665.049	**51.5498**	669.78	51.2684	674.672	51.3298	680.603	51.3044	664.178	51.4578
DAX30	**85.5863**	**50.6762**	85.7886	50.238	85.7011	50.4135	86.2227	50.4368	86.5245	50.4643	91.9907	49.9202
ETHUSD	**37.9566**	**51.7844**	38.1581	51.5597	38.406	51.2828	38.813	51.4085	38.8142	50.6233	38.2602	51.5597
FTSE100	**23.4105**	**52.2703**	23.5147	51.7304	23.4354	51.7304	23.5771	51.6581	23.8746	51.4663	25.2417	51.6153
US500	**22.3209**	50.3227	22.4534	50.0683	22.4146	50.1003	22.6872	**50.4303**	22.6837	50.0829	23.266	50.1846
USDJPY	**0.287882**	**50.7606**	0.288621	50.5977	0.289673	50.5977	0.291194	50.5977	0.293137	50.5658	0.288265	50.5475
XAUUSD	**13.1594**	**50.8788**	13.2079	50.7655	13.2321	50.7848	13.3127	50.4696	13.2972	50.6073	13.8924	49.7311

As aggregated by the critical-difference diagram in Figure 5, RADIAN attains the most optimal mean rank across the evaluated datasets ($\alpha = 0.05$), establishing a statistically significant performance advantage over legacy baselines such as LSTM while strictly superseding top-tier architectures like PatchTST.

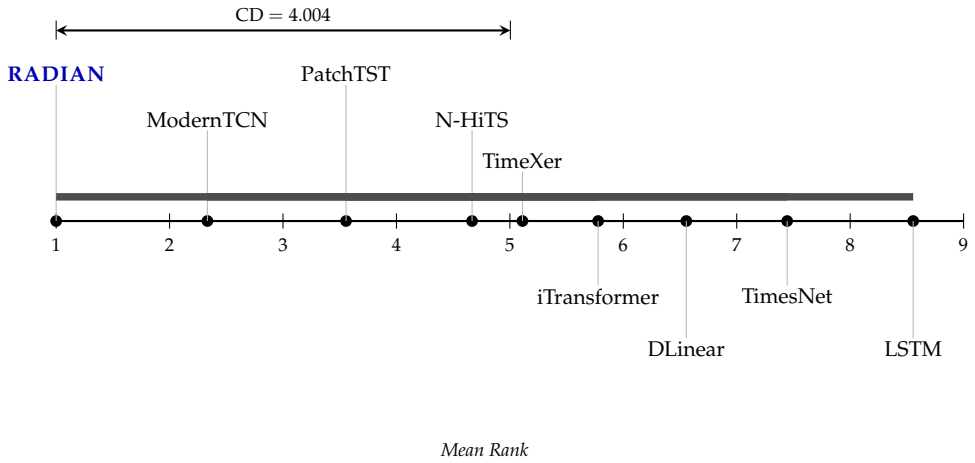


Figure 5. Comparative performance of deep learning forecasting models expressed as mean ranks. Non-significant differences ($\alpha = 0.05$) are indicated by horizontal brackets. RADIAN demonstrates the most favorable rank, followed by ModernTCN and PatchTST, whereas LSTM and TimesNet show the weakest relative performance.

Visualizing corresponding sequential predictions in Figure 6, RADIAN maintains demonstrably tighter structural alignment with the ground truth trajectory during volatile and transition regimes, systematically suppressing the absolute error magnitudes relative to standard convolutional benchmarks.

 doi:10.20944/preprints202604.0884.v1

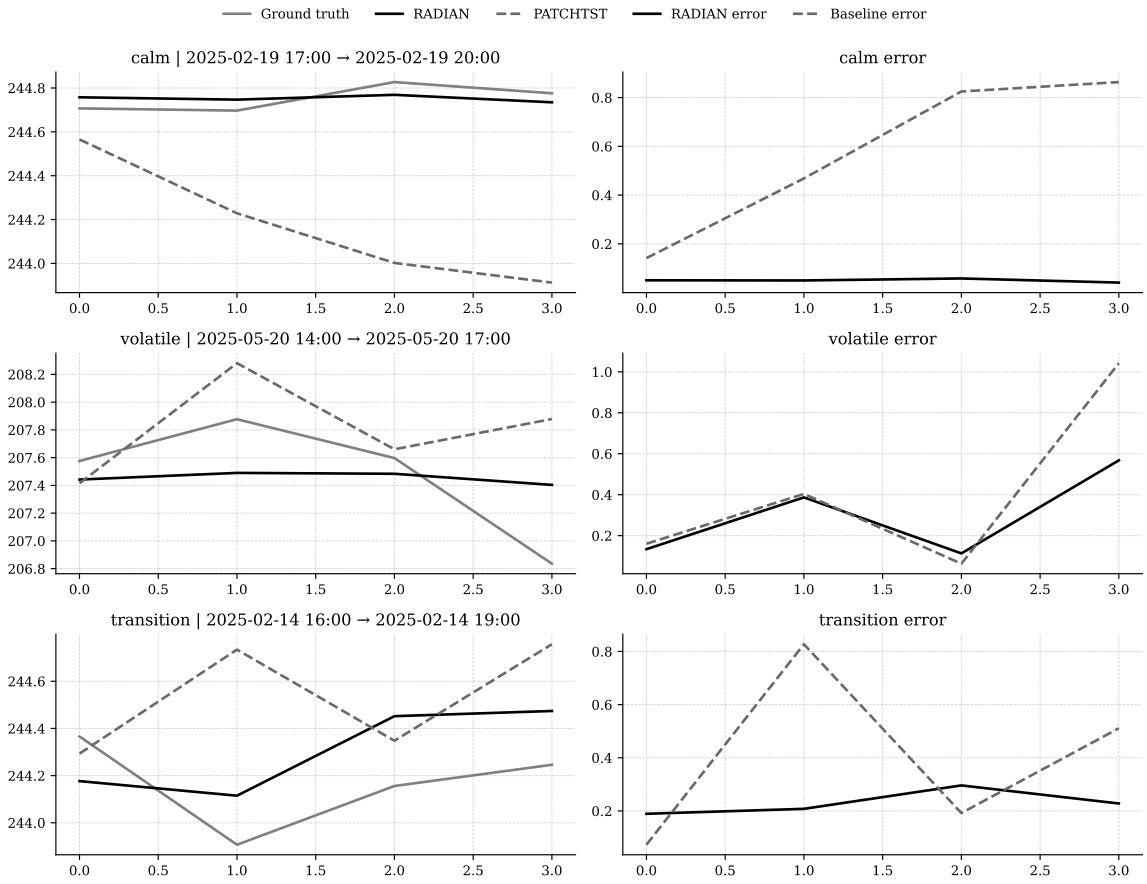


Figure 6. Comparative forecast performance across market regimes. For each regime (calm, volatile, transition), the left column shows ground truth versus predictions from RADIAN and the selected baseline; the right column displays absolute prediction errors. Window timestamps are indicated in each subplot title.

For completeness, expanded dataset/model breakdowns are reported in Appendix Section C (Tables A6–A8 and A6).

5.2. Robustness Under Non-Stationarity

Stratifying evaluation by market conditions in Tables 4, 5, and 6, RADIAN systematically bounds error inflation during high-volatility events, and as corroborated by the structural metrics in Table 7, it yields the flattest MAE–volatility response slope alongside an unmatched worst-decile mean squared error.

Table 4. Regime-stratified comparison for calm windows (lowest-volatility tertile): RMSE (↓) and directional accuracy (DA, %, ↑). Bold values indicate the best result per row/metric; underlined values indicate the second-best result.

	RADIAN		ModernTCN		PatchTST		TimeXer		iTransformer		N-HiTS	
Dataset	**RMSE**	**DA**	**RMSE**	**DA**	**RMSE**	**DA**	**RMSE**	**DA**	**RMSE**	**DA**	**RMSE**	**DA**
AAPL	**1.17473**	**52.636**	1.20872	51.2179	1.26594	51.3013	1.31382	52.4525	1.29356	51.4681	1.20079	51.4681
AUDUSD	**0.00072**	**51.4299**	0.00072	50.9983	0.00072	50.7956	0.00072	51.0433	0.00072	51.2497	0.00073	51.0996
BTCUSD	**352.208**	**51.7604**	352.71	51.6143	355.171	51.6143	356.243	51.6143	360.051	50.9086	359.585	51.5413
DAX30	**44.3731**	**50.2517**	44.5435	49.5143	44.8666	49.766	44.8154	49.7969	45.1181	49.7969	49.2693	49.3687
ETHUSD	**17.5892**	51.2257	17.5988	**51.9627**	17.8298	50.9976	18.004	50.8796	18.0882	49.8941	17.853	51.4537
FTSE100	**12.8137**	**51.8568**	12.8365	51.7851	13.0164	51.7851	12.97	51.7851	13.1161	51.3635	14.0828	51.7851
US500	**9.00642**	**50.3714**	9.08592	49.8541	9.10911	49.5181	9.17819	50.305	9.21651	49.4518	9.13131	49.7966
USDJPY	**0.173258**	**50.4363**	0.174598	50.1939	0.173849	50.3133	0.174331	50.3133	0.175922	50.3133	0.173637	50.2685
XAUUSD	**6.38012**	50.4186	6.39978	**50.5086**	6.41445	50.2973	6.44058	50.4812	6.46872	50.2074	6.93114	50.3208

 doi:10.20944/preprints202604.0884.v1

Table 5. Regime-stratified comparison for transition windows (middle-volatility tertile): RMSE (↓) and directional accuracy (DA, %, ↑). Bold values indicate the best result per row/metric; underlined values indicate the second-best result.

	RADIAN		ModernTCN		PatchTST		TimeXer		iTransformer		N-HiTS	
Dataset	RMSE	DA	RMSE	DA	RMSE	DA	RMSE	DA	RMSE	DA	RMSE	DA
AAPL	**1.83344**	**50.2488**	1.88874	49.334	1.95201	49.9117	2.02056	49.5105	2.00147	49.3821	1.87434	49.5105
AUDUSD	**0.0010**	**50.6978**	0.0010	50.4572	0.0010	49.4947	0.0010	50.4572	0.0010	50.3905	0.0010	50.5423
BTCUSD	**578.607**	**51.7295**	581.206	50.7613	587.679	50.6207	588.969	50.5404	591.834	50.8979	583.478	50.9341
DAX30	**59.5844**	**50.9066**	59.91	50.6838	59.9799	50.8236	60.3654	50.2731	60.219	50.71	64.0141	49.5259
ETHUSD	**32.0134**	**51.7055**	32.2486	51.2948	32.3061	51.0128	32.7433	51.2948	32.648	50.5578	32.5292	51.1619
FTSE100	**16.9644**	**51.7447**	17.2029	51.5447	17.2282	51.5713	17.1642	51.558	17.3029	51.5447	18.1041	50.9535
US500	**15.5185**	**50.3307**	15.5425	50.0985	15.6328	50.2212	15.7309	50.2562	15.7362	50.1248	15.6127	50.2343
USDJPY	**0.240022**	**50.758**	0.241052	50.4555	0.241524	50.4555	0.242232	50.4555	0.243828	50.4555	0.24051	50.3154
XAUUSD	**10.1542**	**51.8514**	10.2179	51.8359	10.2582	51.5606	10.3688	51.2194	10.332	51.4094	11.3654	49.7926

Table 6. Regime-stratified comparison for volatile windows (highest-volatility tertile): RMSE (↓) and directional accuracy (DA, %, ↑). Bold values indicate the best result per row/metric; underlined values indicate the second-best result.

	RADIAN		ModernTCN		PatchTST		TimeXer		iTransformer		N-HiTS	
Dataset	RMSE	DA	RMSE	DA	RMSE	DA	RMSE	DA	RMSE	DA	RMSE	DA
AAPL	**3.5935**	**52.2727**	3.63695	50.8059	3.7242	50.8382	3.96303	51.6763	3.96679	50.951	3.65134	51.225
AUDUSD	**0.0017**	**52.2977**	0.0017	51.8056	0.0017	51.5885	0.0018	51.5342	0.0018	52.0517	0.0018	51.9612
BTCUSD	**917.765**	52.0882	923.788	51.8208	928.748	51.4472	937.946	51.6793	947.414	51.8208	918.272	**52.2416**
DAX30	**127.548**	50.8342	127.582	50.5048	127.582	50.6417	128.108	50.6588	128.671	**51.0481**	136.789	50.6588
ETHUSD	**54.2712**	**52.7121**	54.548	52.3336	54.9536	51.8251	55.489	51.9473	55.5167	51.3955	54.519	52.1444
FTSE100	**34.1298**	52.0311	34.3318	**52.4052**	34.2826	51.8528	34.4648	51.5049	34.9445	51.4005	37.037	51.8528
US500	**33.987**	**50.7973**	34.2125	50.2401	34.0909	50.5401	34.5578	50.4758	34.5388	50.6559	35.7383	50.5058
USDJPY	**0.398577**	**51.3059**	0.398976	50.9885	0.401254	50.57	0.403874	50.671	0.406406	50.4185	0.398945	50.9885
XAUUSD	**19.2459**	**50.4854**	19.3046	50.3034	19.3278	50.3148	19.4245	49.7345	19.4034	50.2124	19.9109	49.0937

Table 7. Stress-test summary across volatility escalation, distribution-shift windows, and extreme-volatility tails. Lower values indicate lower error or reduced sensitivity to error growth.

Stress scenario	Metric	RADIAN	TimesNet	ModernTCN	PatchTST	TimeXer
Volatility escalation (percentile sweep)	MAE–volatility slope (↓)	**1.4069**	1.4149	1.4170	1.4186	1.4663
Distribution-shift window	Worst-decile MSE (↓)	**93916.2351**	106277.6696	96250.9116	96515.6212	98774.4408
Extreme-volatility tail	Tail RMSE (↓)	**594.9410**	614.6372	602.3759	597.7476	604.1600

Aligning absolute prediction errors to structural market shifts in Figure 7, RADIAN demonstrates a substantially narrower 95% confidence interval and a faster error recovery post-transition, structurally outperforming the best baseline exactly at the transition boundary ($t = 0$).

Contrasting cross-variable dependency mappings in Figure 8, the model transitions from a uniform, diffuse attention distribution during calm periods to a highly sparse, localized feature routing structure distinctly adapted to volatile regimes.

Parametrizing prediction accuracy against volatility percentiles in Figure 9, mean absolute error strictly increases for all structures; crucially, RADIAN maintains a much shallower error-growth slope than TimesNet inside the 95% confidence interval.

 doi:10.20944/preprints202604.0884.v1

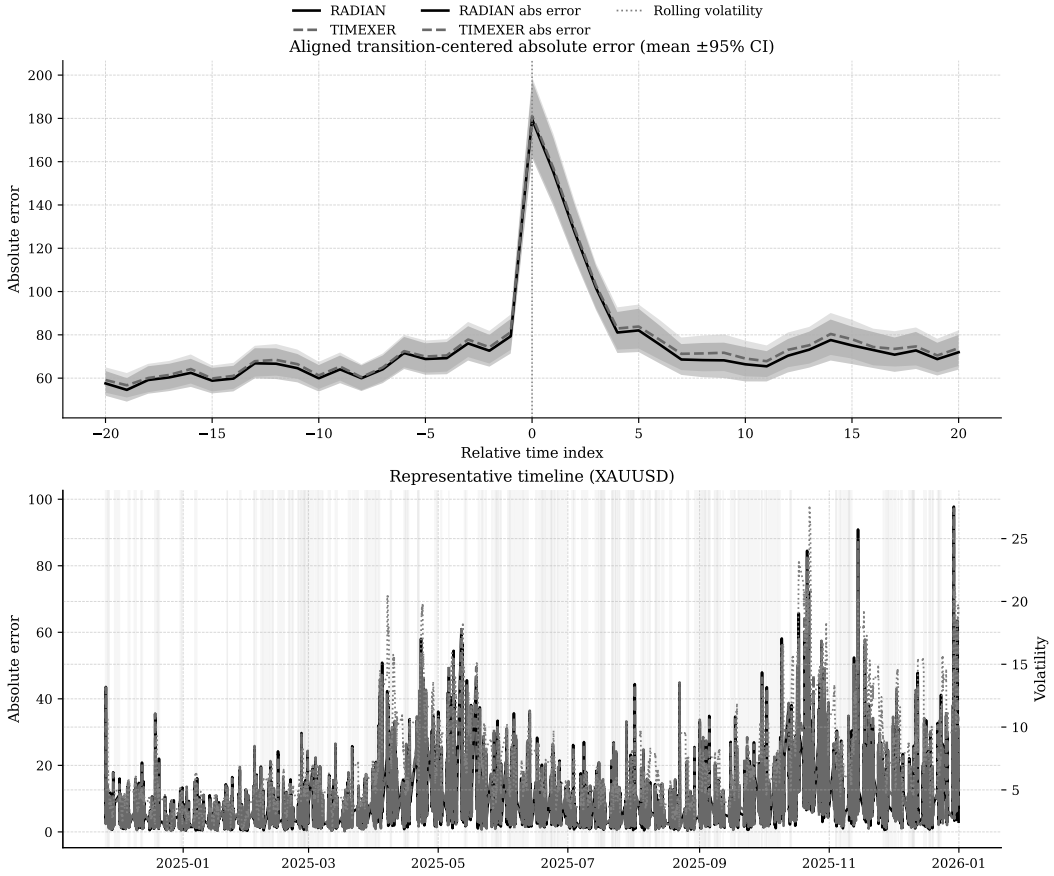


Figure 7. Transition-centered absolute prediction error across all datasets. The upper panel displays mean absolute error (±95% CI) aligned at transition points ($t = 0$) for RADIAN and the best-performing baseline. The lower panel shows a representative timeline of absolute errors for both models with overlaid rolling volatility; shaded regions indicate identified transition periods.

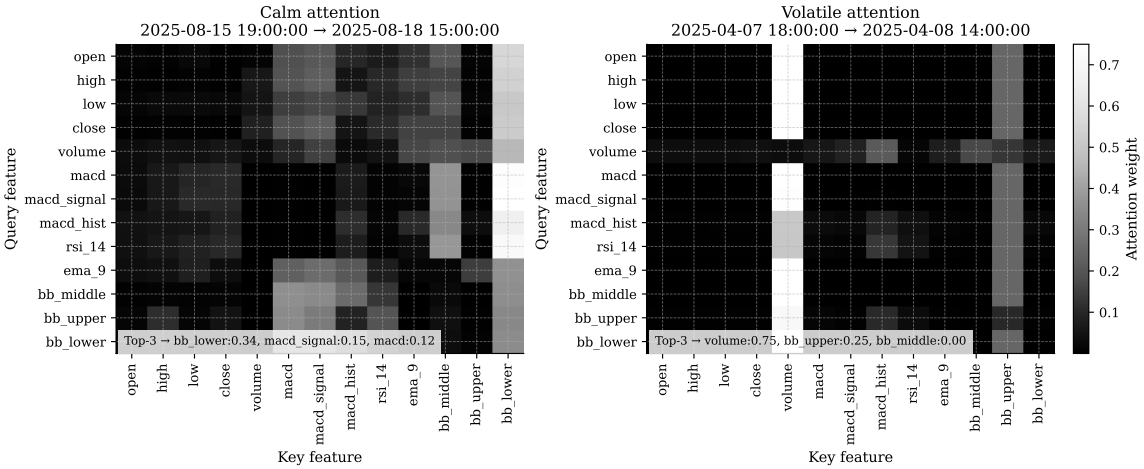
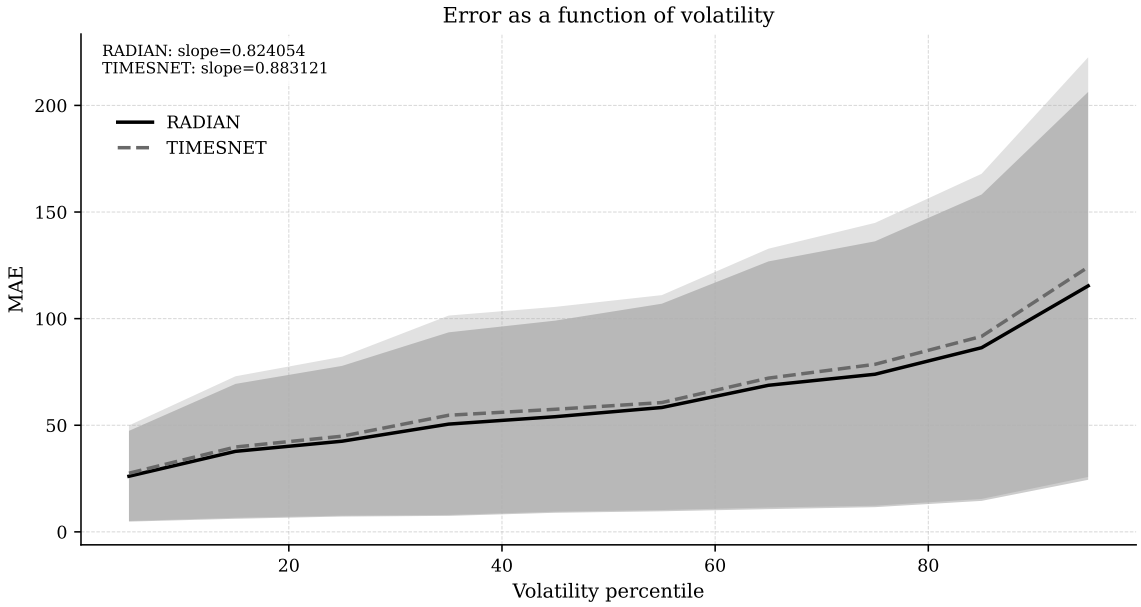


Figure 8. Cross-variable attention weights from the RADIAN model under calm (left) and volatile (right) market regimes. Each heatmap shows the average attention across heads, with rows representing query features and columns key features.

 doi:10.20944/preprints202604.0884.v1



MAE increases with volatility; RADIAN maintains lower error growth under high-volatility regimes.

Figure 9. Mean absolute error (MAE) as a function of volatility percentile for RADIAN and TimesNet. Solid lines denote the mean MAE across datasets and seeds; shaded regions represent 95% confidence intervals. The figure demonstrates that prediction error increases monotonically with volatility, with RADIAN exhibiting consistently lower error and a shallower slope than the baseline.

5.3. Ablation Evidence for Decoder-Centric Gains

Isolating the prognostic value of discrete neural components in Table 8, the structural ablations (A0–A9) collectively verify that omitting path decoding outright (A9) triggers maximum RMSE instability, while displacing gated fusion (A6) systematically suppresses directional accuracy across the decoder-coupled variants.

Table 8. Ablation results for selected datasets (BTCUSD, US500, USDJPY, XAUUSD) using RMSE (↓) and directional accuracy (DA, %, ↑). Bold values indicate the best result per row/metric; underlined values indicate the second-best result.

Ablation	BTCUSD		US500		USDJPY		XAUUSD	
	RMSE (↓)	DA (↑)	RMSE (↓)	DA (↑)	RMSE (↓)	DA (↑)	RMSE (↓)	DA (↑)
Full RADIAN	**660.14**	**51.76**	**22.21**	**50.46**	**0.29**	**51.15**	**13.06**	**51.53**
Remove RevIN	661.31	51.19	22.34	49.44	0.29	50.67	13.38	49.74
Remove temporal backbone	662.08	51.08	22.38	50.19	0.29	50.70	13.13	51.45
Remove cross-variable branch	663.63	51.44	22.34	50.07	0.29	50.72	13.15	51.07
Replace target-aware with mean pooling	662.93	51.29	22.25	50.45	0.29	51.00	13.13	51.29
Remove regime features	662.92	51.19	22.30	49.98	0.29	50.67	13.14	51.28
Replace gated fusion with simple addition	664.89	50.69	22.45	49.69	0.29	50.70	13.17	50.93
Single expert (remove MoE)	663.86	51.43	22.27	50.00	0.29	50.68	13.14	51.29
Remove directional component	664.54	51.06	22.31	50.24	0.29	50.44	13.14	51.30
Remove path decoding	667.42	51.74	22.44	50.13	0.29	50.61	13.27	51.05

Comprehensive ablation outputs are systematically collated in Appendix C.2.

Demonstrating the explicit prognostic utility of context vectors across a representative volatile segment in Figure 10, the fully assembled RADIAN model (A0) strictly bounds mean absolute error growth, whereas striking regime features (A5) systemically inflates deviation from true market targets.

 doi:10.20944/preprints202604.0884.v1

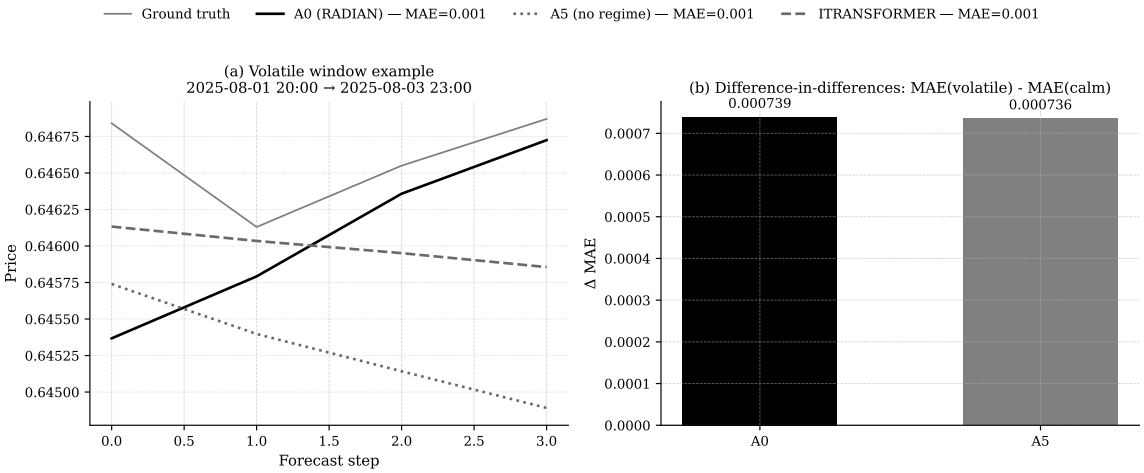


Figure 10. Impact of removing regime conditioning on forecasting accuracy. The left panel presents a volatile-window forecast comparison between the complete RADIAN model (A0), its variant lacking regime conditioning (A5), a baseline, and actual values. The right panel quantifies the regime-specific performance gap, showing that A0 achieves a smaller increase in MAE from calm to volatile conditions than A5.

Decomposing the underlying mixture-of-experts structural assignments via Figure 11, the gate probability sequences visibly adapt and re-weight specific functional layers across calm, shifting, and volatile boundaries, validating parameter reconfiguration matched directly to market variance.

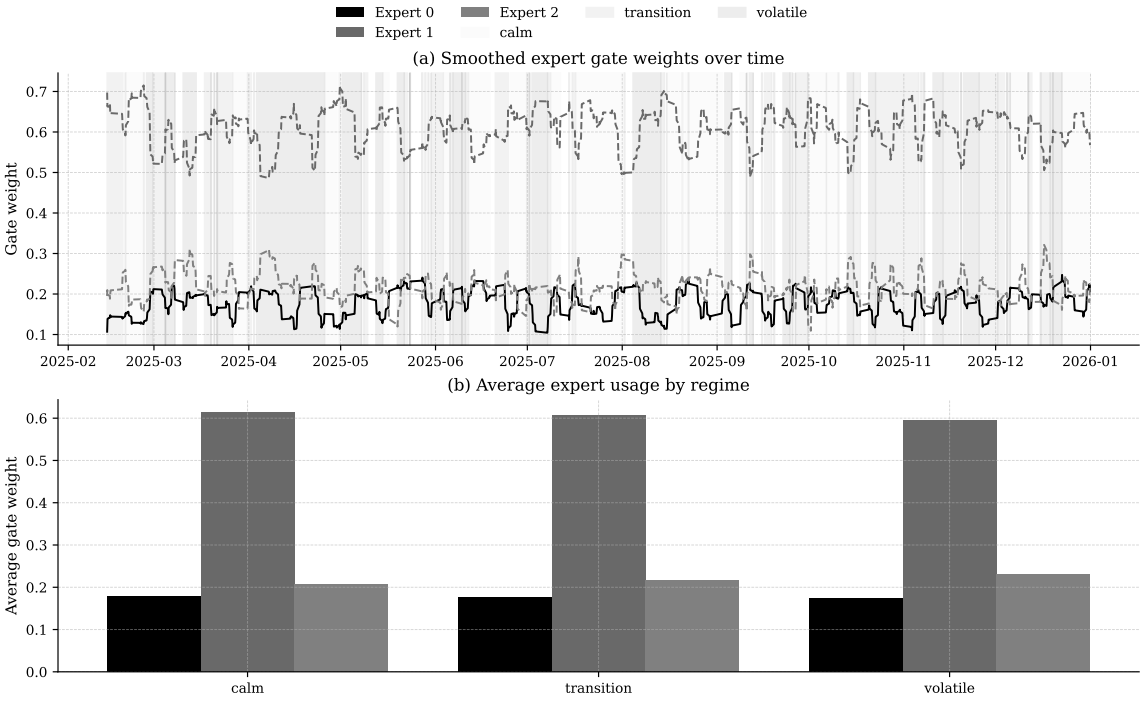


Figure 11. Mixture-of-experts (MoE) gate weight dynamics across market regimes. Panel (a) shows temporally smoothed expert assignment probabilities over the test period, with background shading indicating calm, transition, and volatile regimes. Panel (b) reports the average gate weight per expert for each regime, revealing systematic variation in expert utilisation with volatility conditions.

Contrasting the native autoregressive path integration strategy (A0) against single-shot direct generation (A9) in Figure 12, the fully recursive trajectories structurally mitigate random-walk step variance, compressing broad sign inconsistency rates while systematically smoothing predictive outputs.

 doi:10.20944/preprints202604.0884.v1

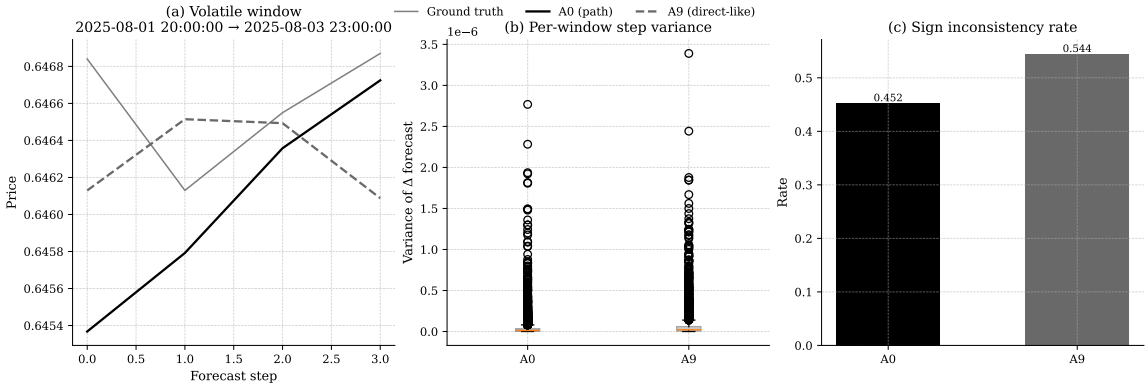


Figure 12. Comparison of path-based (A0) and direct-like (A9) decoding strategies. Panel (a) shows forecasts for a representative volatile window; A0 more closely follows the ground truth trajectory. Panel (b) reports per-window step variance of predictions, and panel (c) presents the sign inconsistency rate, both indicating smoother and more consistent forecasts from A0.

For AUDUSD at seed 417, removing regime features (A5) increases RMSE by 1.38% and reduces DA by 0.91 percentage points relative to A0. Larger RMSE increases are observed for A7 (single expert, +3.32%) and A9 (no path decoding, +4.38%).

5.4. Directional Performance and Stability

Synthesizing comparative classification performance broadly throughout Table 9, RADIAN records a robust +0.31 percentage-point directional accuracy edge on average compared to sector-optimal legacy pipelines, confirming that its enhanced sequence trajectory logic functionally maps securely across diverse market indices rather than over-fitting specific localized windows.

Table 9. Directional accuracy (DA, %; mean ± standard deviation across three seeds) on nine financial datasets. Bold values indicate the best result per row; underlined values indicate the second-best result.

Dataset	RADIAN	ModernTCN	PatchTST	TimeXer	iTransformer	N-HiTS
AAPL	**52.0399 ± 0.273483**	50.4395 ± 0.333956	50.6782 ± 0.535957	51.0796 ± 0.522948	50.6944 ± 0.315744	51.0796 ± 0.522948
AUDUSD	51.2706 ± 0.0765479	51.0947 ± 0.146877	50.6322 ± 0.0899592	51.1427 ± 0.177825	**51.3198 ± 0.130428**	51.2079 ± 0.0928639
BTCUSD	51.5112 ± 0.14627	**51.5498 ± 0.27139**	51.2684 ± 0.517188	51.3298 ± 0.699599	51.3044 ± 0.288453	51.4578 ± 0.10604
DAX30	**50.6762 ± 0.421722**	50.238 ± 0.126865	50.4135 ± 0.17044	50.4368 ± 0.102676	50.4643 ± 0.337482	49.9202 ± 0.411971
ETHUSD	**51.7844 ± 0.249453**	51.5597 ± 0.264979	51.2828 ± 0.440908	51.4085 ± 0.361445	50.6233 ± 0.140526	51.5597 ± 0.264979
FTSE100	**52.2703 ± 0.151442**	51.7304 ± 0.0565018	51.7304 ± 0.0565018	51.6581 ± 0.240057	51.4663 ± 0.118778	51.6153 ± 0.258394
US500	50.3227 ± 0.0265298	50.0683 ± 0.0838191	50.1003 ± 0.0795895	**50.4303 ± 0.397956**	50.0829 ± 0.182141	50.1846 ± 0.632555
USDJPY	**50.7606 ± 0.153839**	50.5977 ± 0.21293	50.5977 ± 0.21293	50.5977 ± 0.21293	50.5658 ± 0.0194426	50.5475 ± 0.126804
XAUUSD	**50.8788 ± 0.268592**	50.7655 ± 0.475718	50.7848 ± 0.14283	50.4696 ± 0.220528	50.6073 ± 0.397548	49.7311 ± 0.491006

Visualizing the dispersion bounds across datasets via Figure 13, RADIAN strictly limits worst-case standard deviation spread across unique seed instances, structurally lowering RMSE variance while sustaining tight win margins against dataset-specific legacy baselines.

 doi:10.20944/preprints202604.0884.v1

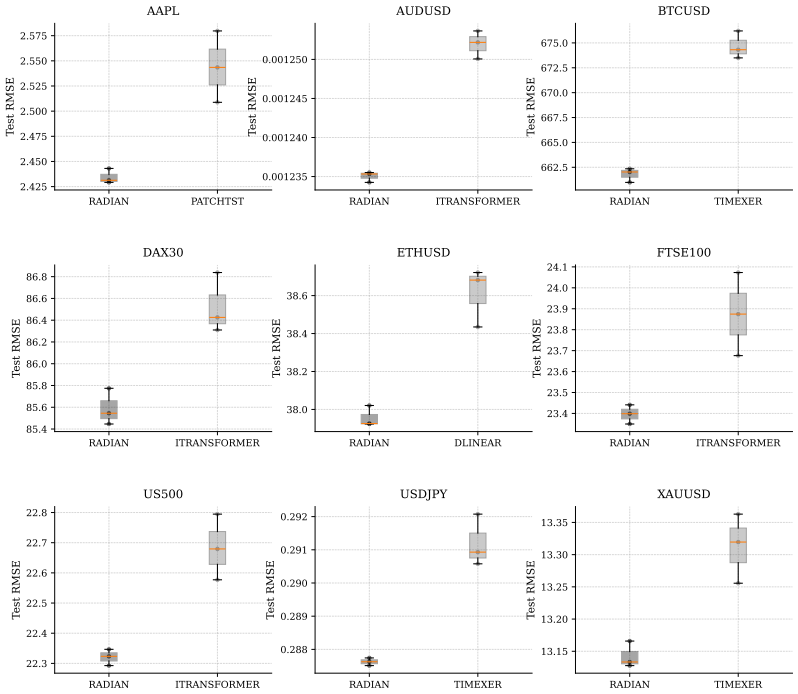


Figure 13. Multi-dataset robustness comparison of RADIAN against dataset-specific baselines. For each dataset, the left boxplot (RADIAN) and right boxplot (baseline) summarise test RMSE variability across seeds, with scatter points representing individual runs.

5.5. Efficiency and Accuracy–Cost Trade-Off

Benchmarking hardware limits relative to predictive boundaries in Table 10, RADIAN establishes the global minimum test RMSE (94.10) whereas deploying an extremely sparse layout of 516,425 parameters, returning a fully processed batch vector in 4.628 ms on standard CPU resources.

Table 10. Accuracy–efficiency summary across models. The table reports aggregated RMSE (↓), parameter count, and median CPU inference latency (ms) at $B = 1$ and $B = 64$. Bold values indicate column-wise minima.

Model	Agg. test RMSE	Params	Latency ms (B=1)	Latency ms (B=64)
RADIAN	**94.1048**	516,425	4.62815	24.7172
ModernTCN	94.5486	608,862	6.0382	83.7451
PatchTST	95.0893	400,929	4.89115	27.3331
TimeXer	95.8071	601,692	4.90595	12.5903
iTransformer	96.5303	268,932	2.90395	12.2155
N-HiTS	95.5107	455,326	1.59575	3.32515
DLinear	98.0036	**7,432**	**0.6078**	**1.351**
TimesNet	98.226	873,556	16.8479	145.299
LSTM	1,984.61	333,188	1.8045	9.36025

Quantifying computational execution overhead directly relative to accuracy, the latency plotting in Figure 14 places RADIAN firmly on the optimal Pareto frontier boundary, asserting strict predictive superiority over deep learning networks such as PatchTST with fractional processor utilization.

 doi:10.20944/preprints202604.0884.v1

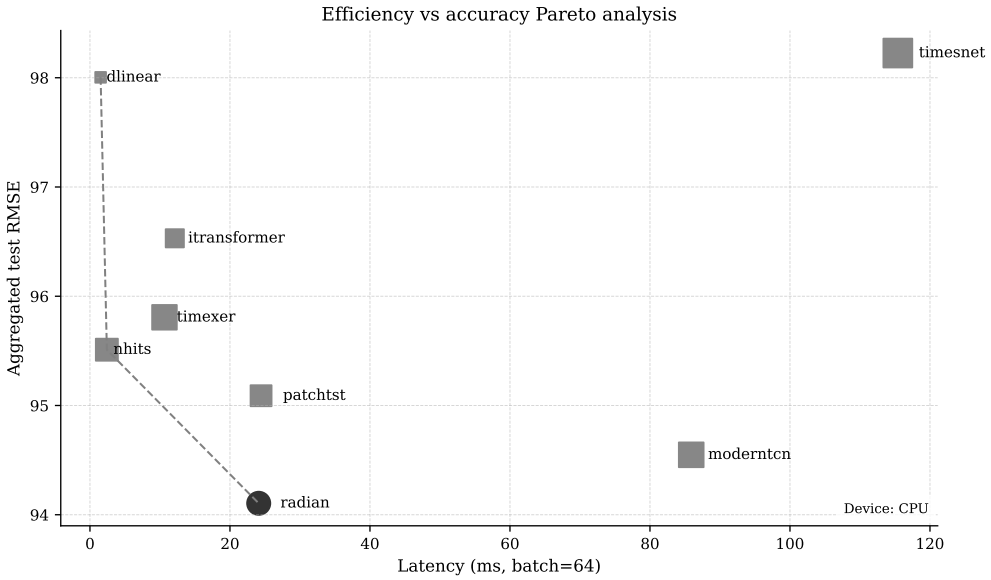


Figure 14. Trade-off between computational cost and prediction error. Each point represents a model; the x-axis shows inference latency (ms) for a batch of 64 samples, and the y-axis shows mean test RMSE aggregated across datasets. RADIAN occupies a position on the Pareto boundary, indicating competitive accuracy with moderate latency.

6. Discussion

6.1. Evidence-Grounded Summary

The empirical picture is directional consistency with moderate variance rather than a uniformly dominant margin. Table 3 show the lowest RMSE on every dataset and the best directional accuracy (DA) on six of nine datasets; Figure 5 places RADIAN at the top mean rank under the critical-difference analysis. The same pattern persists under regime stratification: Tables 4, 5, and 6 show that RADIAN remains the lowest-error model in calm, transition, and volatile windows, while Table 7 reports the smallest slope of MAE against volatility (1.41), worst-decile MSE (9.39×10^4), and tail RMSE (5.95×10^2) among the compared models. Figures 7 and 9 further indicate that error growth with volatility is slower for RADIAN than for the baselines. Directional gains are smaller and dataset-dependent, so the principal effect is reduced error sensitivity rather than uniform improvement across all metrics. The paired tests in Appendix Table A18 are directionally consistent with this pattern, but the three-seed design and Benjamini–Hochberg correction limit inferential strength.

6.2. Mechanistic Interpretation

The ablations indicate that decoder-side components exhibit the largest ablation sensitivity. In the representative AUDUSD seed-level ablation, removing regime features is associated with an RMSE increase of 1.4%, collapsing mixture-of-experts routing to a single expert with a 3.3% increase, and removing path decoding with a 4.4% increase; the cross-dataset summary in Table 8 preserves the same ordering on BTCUSD, US500, USDJPY, and XAUUSD, and Appendix Table A9 shows the same qualitative pattern across the full ablation set. In that table, replacing gated fusion with simple addition and removing path decoding produce the largest RMSE degradations, whereas removing the regime vector alone yields a smaller but still systematic loss. Encoder removals also affect performance, but the decoder-side variants are the most sensitive. Figure 10 shows that the effect of regime conditioning is most visible in volatile windows, Figure 11 shows regime-dependent routing variation, and Figure 12 shows lower step variance and sign inconsistency under path decoding. These observations are mechanism-consistent, but they do not establish causal identification for individual decoder submodules; the ablations also alter capacity and optimization geometry.

 doi:10.20944/preprints202604.0884.v1

6.3. Design Principle Extraction

The evidence supports a modular design principle: keep the shared encoder stack regime-agnostic and concentrate adaptation at the prediction interface, where the model can condition on $\mathbf{r}_t$ without entangling representation learning with short-horizon noise. This principle is most defensible when the regime descriptor is causally computable from the observed window, as in RADIAN, and when the shift is local enough that a compact conditional policy can react without retraining the full network. In that setting, decoder-side conditioning provides an interpretable adaptation point and preserves a reusable representation backbone. Transfer of the same separation to other non-stationary sequence domains is plausible, but remains a hypothesis outside the reported financial benchmarks.

6.4. Comparison to Alternative Non-Stationarity Approaches

Online learning methods adapt by updating parameters as data arrive [37]; RADIAN instead keeps parameters fixed at test time and moves adaptation into conditional inference at the decoder. Adaptive normalization methods such as RevIN [25] adapt preprocessing, so the adapted object is the input scale and location rather than the forecast rule itself. Non-stationary and decomposition-based Transformer variants such as the Non-stationary Transformer, FEDformer, PatchTST, and iTransformer [8,9,22,26] shift adaptation into the encoder or tokenization stack, where the model representation, frequency decomposition, or attention geometry is modified before prediction. By contrast, RADIAN keeps the representation backbone regime-agnostic and adapts the decision rule itself through decoder-side conditioning. Time-varying parameter and regime-switching formulations [3,10,11] make latent dynamics explicit, but they impose stronger parametric assumptions about state evolution than the compact conditional policy used here.

6.5. Broader Applicability

Beyond finance, the same separation between regime sensing and decoder-side adaptation may be relevant in climate and environmental forecasting [38], network telemetry [39], and biomedical time-series analysis [40]. In climate settings, the regime signal would be a causal summary of recent circulation, anomaly, or extreme-event conditions, while decoder-side adaptation would condition short-horizon forecasts on those summaries instead of altering the shared encoder. In telemetry, the analogue would be a rolling congestion or anomaly descriptor derived from recent traffic statistics, with the decoder adjusting load or alert predictions while the encoder remains fixed. In biomedical signals, the regime descriptor would be a real-time patient-state summary from observed vitals or sensor dynamics, and decoder-side adaptation would modulate risk or trajectory outputs without retraining the feature extractor. This design principle is most likely to transfer when causal regime descriptors are available in real time, distribution shift is local enough for a compact conditional policy to track, and the learned routing remains stable across windows; it is less likely to hold when regimes are latent or weakly observable, when structural drift unfolds over long horizons, or when the shift changes the representation space itself rather than only the prediction rule.

6.6. Limitations

- **Lagged regime response.** The regime vector $\mathbf{r}_t$ is computed from the window ending at time t, so abrupt events at $t+1$ can only be anticipated indirectly. The transition-window error dynamics in Figure 7 expose this limitation most clearly.
- **Partial observability.** The four-statistic regime summary captures only return-based local shape. It cannot represent cross-asset dependencies, macroeconomic shocks, or other exogenous drivers of market structure.
- **Inference power and ablation confounding.** Appendix Table A18 relies on three seeds, and only two comparisons remain significant after Benjamini-Hochberg correction. In addition, the architectural ablations change capacity and optimization geometry together with the targeted mechanism, so the decoder-centric interpretation is supported by sensitivity evidence but not fully isolated causally.

 doi:10.20944/preprints202604.0884.v1

- **Incomplete diagnostics.** The released artifacts do not include per-regime aggregate summaries, directional confusion matrices, or FLOPs measurements. The aggregated RMSE reported in Table 10 should therefore be read as a coarse comparative indicator rather than a fully scale-normalized cross-asset score.

6.7. Future Directions

- **Streaming regime updates.** Recompute or smooth $\mathbf{r}_t$ online with rolling, exponentially weighted, or change-point-triggered updates so that abrupt shifts are reflected more quickly.
- **Richer regime observability.** Augment the four return statistics with cross-asset, macroeconomic, or latent-volatility features, or learn a sparse regime embedding under causal constraints, to reduce omission from the current descriptor.
- **Mechanism isolation.** Run matched-parameter ablations, expert-swapping tests, and representation-similarity probes across multiple seeds to separate decoder mechanism from capacity and optimization effects.
- **Operational diagnostics.** Expose routing entropy, expert load, regime-drift percentiles, transition-window error, and FLOPs in the evaluation pipeline; if drift persists, restrict adaptation to decoder-side low-rank adapters or periodic refreshes on recent windows.

6.8. Practical Implications

In settings where regime observability is adequate and routing behavior remains stable across recent windows, plausible deployment signals are regime drift, expert-utilization concentration, and transition-window error. Because $\mathbf{r}_t$ is computed causally, large shifts in its percentile rank relative to training data provide an indicator of distribution change; sustained concentration of MoE weights in Figure 11 can flag reduced conditional diversity; and spikes in Figure 7 identify regimes where adaptation may be insufficient. When these signals move outside their reference bands and the underlying assumptions hold, a plausible intervention is to recalibrate normalization statistics, refresh the decoder modules, or retrain on recent windows rather than rebuilding the full encoder stack. The model remains on a favorable accuracy-latency frontier in Table 10 and Figure 14, so these diagnostics can be monitored without materially changing inference cost in the reported benchmark.

7. Conclusion

Financial time series under shifting dynamics are difficult because the mapping from recent history to near term movement changes as volatility, dependence structure, and local market regime evolve. The central problem is therefore not simply to encode more context, but to determine where adaptation should enter so that the forecast rule can respond to changing conditions without destabilizing the shared representation.

The proposed model addresses this problem by separating representation learning from prediction synthesis. The encoder stack remains regime independent, while a compact causal summary of the observed window conditions the decoding pathway at the point where forecasts are formed. This design localizes adaptation at the prediction interface rather than dispersing it throughout the network, preserving reusable structure while still allowing the model to react to short term market state.

The empirical evidence supports this design choice in a qualified but consistent way. Across heterogeneous financial benchmarks, the model reduces forecast error relative to strong alternatives, and the advantage is clearest when the input distribution is unstable. The results further indicate that the improvement is not merely cosmetic: error growth is dampened in volatile and transition periods, suggesting greater robustness to regime change. Sensitivity analyses reinforce this interpretation by showing that weakening the regime conditioned decoding path degrades performance, whereas the gains in directional accuracy are smaller and less uniform than the gains in error reduction. Accordingly, the strongest claim warranted by the study is not universal superiority on every metric, but improved resilience of the forecast rule under changing conditions.

 doi:10.20944/preprints202604.0884.v1

These conclusions remain bounded by the study design. The regime descriptor is intentionally compact and derived only from the observed window, so it cannot represent all structural drivers of market behavior and may respond imperfectly to abrupt shifts. Likewise, the evidence supports the proposed modular design principle within financial forecasting, but it does not establish that the same conditioning strategy will transfer unchanged to every sequence problem with changing dynamics. Even so, the broader implication is that observably causal regime information can be used to adapt the decoder while keeping the representation backbone stable, and future work should examine more responsive regime updates, richer conditional descriptors, and stricter tests that separate conditional routing effects from capacity and optimization effects.

Conflict of Interest

The author declares that there are no conflicts of interest, financial or otherwise, that could have influenced the research, authorship, or publication of this work. The study was conducted independently, and no external funding, affiliations, or personal relationships exist that could be perceived as affecting the objectivity, integrity, or interpretation of the results presented in this article.

Data and Code Availability

The implementation code, along with all relevant configuration files required to reproduce the experiments presented in this manuscript, is publicly available at the following GitHub repository:
https://github.com/NabeelAhmad9/Radian

Processed data artifacts generated and analyzed during this study are not publicly available due to proprietary or ethical restrictions. They are available from the corresponding author upon reasonable request, subject to a signed data use agreement.

This appendix is organized as follows: dataset details and split statistics; hyperparameters and implementation settings; extended results; statistical tests; additional figures; and reproducibility. It documents the released data splits, model settings, evaluation summaries, diagnostic figures, and reproducibility materials used in the experiments.

Appendix A. Dataset Details and Splits

This section summarizes the raw OHLCV schema, engineered inputs, chronological split statistics, and raw-file inventory used in the experiments.

Appendix A.1. Raw Schema and Scope

All nine datasets are hourly open-high-low-close-volume (OHLCV) time series with the schema `datetime, open, high, low, close, volume`. The datasets cover equities (AAPL), foreign exchange (FX; AUDUSD, USDJPY), cryptocurrencies (BTCUSD, ETHUSD), indices (DAX30, FTSE100, US500), and commodities (XAUUSD).

Appendix A.2. Feature Pipeline

The model input uses 13 channels: the five base OHLCV features and eight engineered indicators—`macd, macd_signal, macd_hist, rsi_14, ema_9, bb_middle, bb_upper, bb_lower`. The target is `close`.

Appendix A.3. Chronological Split Statistics

Table A1 reports sample counts and chronological boundaries for each asset. The differences in observation span across assets help contextualize the cross-dataset performance variation discussed in Sections 5 and 6.

 doi:10.20944/preprints202604.0884.v1

Table A1. Dataset inventory for the nine financial series used in this study, listing asset class, date range, and chronological split counts. N denotes the total number of samples, while N_{tr}, N_{val}, and N_{te} denote the counts in the training, validation, and test splits, respectively.

Symbol	Class	Start	End	N	N_{tr}	N_{val}	N_{te}
AAPL	equity	2017-02-01	2025-12-31	15,706	12,564	1,570	1,572
AUDUSD	FX	2015-01-01	2025-12-31	68,568	54,854	6,856	6,858
BTCUSD	crypto	2018-01-01	2025-12-31	62,826	50,260	6,282	6,284
DAX30	index	2015-01-01	2025-12-31	57,816	46,252	5,781	5,783
ETHUSD	crypto	2018-01-01	2025-12-31	62,816	50,252	6,281	6,283
FTSE100	index	2015-01-01	2025-12-31	56,800	45,440	5,680	5,680
US500	index	2015-01-02	2025-12-31	57,698	46,158	5,769	5,771
USDJPY	FX	2015-01-01	2025-12-31	68,569	54,855	6,856	6,858
XAUUSD	commodity	2015-01-01	2025-12-31	65,126	52,100	6,512	6,514

Appendix A.4. Raw CSV Inventory

Table A2 lists the raw files, confirms the six-column OHLCV schema, and reports row counts. It serves as a data-ingestion check for reproducibility.

Table A2. Inventory of raw CSV artifacts used in the study, reporting row and column counts and the first and last feature columns for each asset dataset.

CSV	Rows	Columns	First column	Last column
AAPL	15,706	6	datetime	volume
AUDUSD	68,568	6	datetime	volume
BTCUSD	62,826	6	datetime	volume
DAX30	57,816	6	datetime	volume
ETHUSD	62,816	6	datetime	volume
FTSE100	56,800	6	datetime	volume
US500	57,698	6	datetime	volume
USDJPY	68,569	6	datetime	volume
XAUUSD	65,126	6	datetime	volume

Appendix B. Hyperparameters

This section summarizes the shared training controls, model-family-specific settings, and implementation-specific values used for reproduction.

Table A3 summarizes experiment-level controls shared across all models. Table A4 lists the model-family-specific differences in depth, width, heads, and dropout that account for the parameter and latency variation reported in Table 10. Table A5 records architecture-specific settings needed for faithful implementation.

Table A3. Global training configuration and fixed data parameters used across all models and datasets.

Parameter	Value
Loss	MSE
Optimizer	Adam
Learning Rate	5×10^{-4}
Epochs	50
Batch Size	64
Window Size	24
Forecast Horizon	4

The random seeds are fixed at {417, 153, 999}. Global settings include $T = 24$, $H = 4$, a batch size of 64, a learning rate of 5×10^{-4}, and 50 epochs.

 doi:10.20944/preprints202604.0884.v1

Table A4. Model architecture comparison across primary hyperparameters; - indicates parameters that do not apply to a given model.

Hyperparameter	RADIAN	PatchTST	iTrans.	ModernTCN	DLinear	LSTM	N-HiTS	TimesNet	TimeXer
d_{model}	128	128	128	–	–	–	–	64	164
Layers	3	3	2	–	–	2	–	2	2
Heads	8	4	8	–	–	–	–	–	8
FFN Dim	–	256	256	–	–	–	–	96	256
Dropout	0.0	0.0	0.0	0.1	–	0.05	0.05	0.05	0.05
RevIN	–	True	–	True	–	–	–	–	–
Activation	–	–	GELU	–	–	–	ReLU	–	GELU
Hidden Size	–	–	–	–	–	96	256	–	–
Kernel Size	5	–	–	25	25	–	–	–	–

Table A5. Model-specific architecture parameters that are not directly comparable across models.

Model	Specific Details
RADIAN	Dilations: [1, 2, 4]; Regime Hidden: 128
PatchTST	Patch Length: 16; Stride: 8
ModernTCN	Patch/Stride: 16/8; Dims: [32, 64, 96]; Blocks: [1, 1, 1]
DLinear	Projection Size: 128
LSTM	MLP Hidden: 128; Bidirectional: True
N-HiTS	Stacks/Blocks/Layers: 3/2/2; Pooling Kernel: [2, 2, 2]; Freq Downsample: [4, 2, 1]
TimesNet	Top-k: 5
TimeXer	Patch Length: 8

Appendix C. Extended Results

This section collects the full dataset-level comparison tables and the complete ablation results referenced in Section 5.

Appendix C.1. Main Results

Tables A6–A8 report the full dataset-level comparison across all models and metrics (MSE, MAE, RMSE, DA).

Table A6. Dataset-level results for RADIAN, ModernTCN, and PatchTST across MSE (↓), MAE (↓), RMSE (↓), and DA (↑) on nine datasets. Bold values denote best results per row/metric; underlined values denote second-best results.

	RADIAN				ModernTCN				PatchTST			
Dataset	MSE	MAE	RMSE	DA	MSE	MAE	RMSE	DA	MSE	MAE	RMSE	DA
AAPL	**5.92691**	**1.52104**	**2.43452**	**52.0399**	6.12855	1.57219	2.47554	50.4395	6.47271	1.62803	2.54399	50.6782
AUDUSD	**1.53e-06**	**0.00083**	**0.0012**	51.2706	**1.53e-06**	**0.00083**	**0.0012**	51.0947	1.54e-06	0.00084	**0.0012**	50.6322
BTCUSD	**437,961**	**439.905**	**661.786**	51.5112	442,291	440.942	665.049	**51.5498**	448,608	447.973	669.78	51.2684
DAX30	**7,325.02**	**50.6815**	**85.5863**	**50.6762**	7,359.68	50.8776	85.7886	50.238	7,344.75	51.1913	85.7011	50.4135
ETHUSD	**1,440.7**	**24.3112**	**37.9566**	**51.7844**	1,456.04	24.3824	38.1581	51.5597	1,475.03	24.7972	38.406	51.2828
FTSE100	**548.056**	**14.1489**	**23.4105**	**52.2703**	552.944	14.1735	23.5147	51.7304	549.224	14.2725	23.4354	51.7304
US500	**498.223**	**12.6922**	**22.3209**	50.3227	504.157	12.8259	22.4534	50.0683	502.413	12.9033	22.4146	50.1003
USDJPY	**0.0828764**	**0.194877**	**0.287882**	**50.7606**	0.0833023	0.195489	0.288621	50.5977	0.0839107	0.196941	0.289673	50.5977
XAUUSD	**173.17**	**8.63918**	**13.1594**	**50.8788**	174.45	8.66934	13.2079	50.7655	175.09	8.71226	13.2321	50.7848

 doi:10.20944/preprints202604.0884.v1

Table A7. Dataset-level results for TimeXer, iTransformer, and N-HiTS across MSE (↓), MAE (↓), RMSE (↓), and DA (↑) on nine datasets. Bold values denote best results per row/metric; underlined values denote second-best results.

	TimeXer				iTransformer				N-HiTS			
Dataset	**MSE**	**MAE**	**RMSE**	**DA**	**MSE**	**MAE**	**RMSE**	**DA**	**MSE**	**MAE**	**RMSE**	**DA**
AAPL	7.22221	1.71715	2.68699	51.0796	7.189	1.73288	2.68113	50.6944	6.13925	1.56205	2.47771	51.0796
AUDUSD	1.56e-06	0.00084	**0.0012**	51.1427	1.57e-06	0.00084	0.0013	**51.3198**	1.58e-06	0.00084	0.0013	51.2079
BTCUSD	455,183	450.541	674.672	51.3298	463,256	454.571	680.603	51.3044	441,135	442.273	664.178	51.4578
DAX30	7,434.38	51.5304	86.2227	50.4368	7,486.54	51.4502	86.5245	50.4643	8,502.55	55.4249	91.9907	49.9202
ETHUSD	1,506.57	25.0133	38.813	51.4085	1,506.55	25.243	38.8142	50.6233	1,463.97	24.5964	38.2602	51.5597
FTSE100	555.889	14.3705	23.5771	51.6581	570.021	14.4754	23.8746	51.4663	639.156	15.3499	25.2417	51.6153
US500	514.716	13.0455	22.6872	**50.4303**	514.559	13.0917	22.6837	50.0829	541.339	13.0211	23.266	50.1846
USDJPY	0.0847944	0.197836	0.291194	50.5977	0.0859317	0.200002	0.293137	50.5658	0.0830971	0.195236	0.288265	50.5475
XAUUSD	177.23	8.81781	13.3127	50.4696	176.816	8.78136	13.2972	50.6073	193.059	9.42913	13.8924	49.7311

Table A8. Dataset-level results for DLinear, TimesNet, and LSTM across MSE (↓), MAE (↓), RMSE (↓), and DA (↑) on nine datasets. Bold values denote best results per row/metric; underlined values denote second-best results.

	DLinear				TimesNet				LSTM			
Dataset	**MSE**	**MAE**	**RMSE**	**DA**	**MSE**	**MAE**	**RMSE**	**DA**	**MSE**	**MAE**	**RMSE**	**DA**
AAPL	6.02404	1.55404	2.45437	51.0362	8.58647	1.85958	2.92909	51.0796	6.48732	1.64317	2.54682	50.4015
AUDUSD	1.58e-06	0.00086	0.0013	50.5941	1.62e-06	0.00087	0.0013	50.7466	1.69e-06	0.00089	0.0013	50.6138
BTCUSD	477,236	479.532	690.752	50.8709	478,319	467.057	691.589	51.4578	1.63e+08	11,165.9	12,749.6	49.6439
DAX30	7,893.8	54.8744	88.838	49.8302	7,686.26	53.0993	87.6711	50.3018	1.07e+07	3,144.01	3,257.67	47.9758
ETHUSD	1,490.95	25.1391	38.6127	50.9831	1,593.56	26.2599	39.9192	51.0179	1,543.15	25.8084	39.2783	51.4246
FTSE100	620.196	16.1955	24.8691	50.8453	587.76	14.9406	24.243	51.5903	86,275.8	222.365	292.385	49.6592
US500	513.685	13.1832	22.6642	50.3111	566.267	13.7503	23.7956	50.3111	279,938	456.968	529.034	47.8672
USDJPY	0.0876014	0.205403	0.295914	50.6173	0.0898571	0.206436	0.299758	50.3233	0.364396	0.502048	0.603065	50.1066
XAUUSD	183.479	9.07443	13.5453	50.4529	184.592	8.98366	13.5864	50.422	984,358	867.561	990.381	47.7845

Appendix C.2. Ablation Results

Table A9 reports the complete ablation study, including validation and test performance across all four metrics (MSE, RMSE, MAE, DA) for each dataset.

Table A9. Ablation results on AAPL for MSE (↓), RMSE (↓), MAE (↓), and DA (↑).

Ablation	**MSE**	**RMSE**	**MAE**	**DA**
Full RADIAN	**5.9**	**2.4254**	**1.5195**	51.8221
Remove RevIN	6.0565	2.461	1.5524	50.651
Remove temporal backbone (dilated blocks + temporal attention)	5.9064	2.4303	1.522	50.8626
Remove cross-variable branch	5.97	2.4434	1.5217	51.6764
Replace target-aware pooling with mean pooling	5.9108	2.4312	1.5225	50.7487
Remove regime features	5.9134	2.4317	1.5211	51.1556
Replace gated fusion with simple addition	6.046	2.4589	1.545	50.6673
Single expert (remove MoE)	6.0246	2.4545	1.5395	50.9115
Remove directional component	5.9483	2.4389	1.5356	50.3255
Remove path decoding (no cumulative sum)	6.0385	2.4573	1.5536	**51.8947**

Table A10. Ablation results on AUDUSD for MSE (↓), RMSE (↓), MAE (↓), and DA (↑).

Ablation	**MSE**	**RMSE**	**MAE**	**DA**
Full RADIAN	**1.5197e-06**	**0.0012**	**8.2643e-04**	**51.5195**
Remove RevIN	1.5239e-06	0.0012	8.3176e-04	51.0959
Remove temporal backbone (dilated blocks + temporal attention)	1.5260e-06	0.0012	8.3154e-04	51.4428
Remove cross-variable branch	1.5260e-06	0.0012	8.3179e-04	51.3727
Replace target-aware pooling with mean pooling	1.5288e-06	0.0012	8.3378e-04	51.1993
Remove regime features	1.5300e-06	0.0012	8.3216e-04	51.155
Replace gated fusion with simple addition	1.5307e-06	0.0012	8.3328e-04	51.1845
Single expert (remove MoE)	1.5324e-06	0.0012	8.3347e-04	51.1513
Remove directional component	1.5294e-06	0.0012	8.3227e-04	51.2251
Remove path decoding (no cumulative sum)	1.5434e-06	0.0012	8.3499e-04	51.2915

 doi:10.20944/preprints202604.0884.v1

Table A11. Ablation results on BTCUSD for MSE (↓), RMSE (↓), MAE (↓), and DA (↑).

Ablation	MSE	RMSE	MAE	DA
Full RADIAN	**437097.0876**	**660.1415**	**437.89**	**51.7647**
Remove RevIN	437325.0303	661.3055	440.9926	51.1884
Remove temporal backbone (dilated blocks + temporal attention)	438352.2572	662.0818	439.0286	51.0843
Remove cross-variable branch	440403.2604	663.6289	439.3764	51.4405
Replace target-aware pooling with mean pooling	439475.7174	662.9296	438.8183	51.2884
Remove regime features	439460.7768	662.9184	440.0287	51.1884
Replace gated fusion with simple addition	442078.3505	664.8897	442.9135	50.6922
Single expert (remove MoE)	440705.3587	663.8564	440.3045	51.4325
Remove directional component	441609.2919	664.5369	440.9801	51.0603
Remove path decoding (no cumulative sum)	445454.8654	667.4241	447.8995	51.7366

Table A12. Ablation results on DAX30 for MSE (↓), RMSE (↓), MAE (↓), and DA (↑).

Ablation	MSE	RMSE	MAE	DA
Full RADIAN	**7273.7417**	**85.0944**	**50.4437**	**50.6129**
Remove RevIN	7357.0796	85.7734	50.631	50.5659
Remove temporal backbone (dilated blocks + temporal attention)	7306.3348	85.4771	50.6293	50.1001
Replace target-aware pooling with mean pooling	7300.4951	85.4429	50.5837	50.3483
Remove regime features	7307.6027	85.4845	50.5965	50.2438
Replace gated fusion with simple addition	7344.0208	85.6973	50.8668	50.4135
Single expert (remove MoE)	7300.3243	85.4419	50.6745	50.1654
Remove directional component	7315.2242	85.5291	50.7094	50.1437
Remove path decoding (no cumulative sum)	7460.4246	86.3737	51.239	50.2655

Table A13. Ablation results on ETHUSD for MSE (↓), RMSE (↓), MAE (↓), and DA (↑).

Ablation	MSE	RMSE	MAE	DA
Full RADIAN	**1431.4516**	**37.7494**	**24.1806**	**51.9076**
Remove RevIN	1435.6804	37.8904	24.2964	51.0674
Remove temporal backbone (dilated blocks + temporal attention)	1434.8956	37.88	24.2165	51.5048
Remove cross-variable branch	1441.7076	37.9698	24.3448	51.7897
Replace target-aware pooling with mean pooling	1436.7215	37.9041	24.2489	51.6573
Remove regime features	1439.0035	37.9342	24.2617	51.87
Replace gated fusion with simple addition	1443.9107	37.9988	24.3987	51.1035
Single expert (remove MoE)	1438.8224	37.9318	24.2651	51.6974
Remove directional component	1448.5209	38.0594	24.43	51.3283
Remove path decoding (no cumulative sum)	1492.9191	38.6383	25.4056	50.4535

Table A14. Ablation results on FTSE100 for MSE (↓), RMSE (↓), MAE (↓), and DA (↑).

Ablation	MSE	RMSE	MAE	DA
Full RADIAN	**539.8955**	**23.2182**	**14.0573**	**52.2908**
Remove RevIN	581.2706	24.1096	15.1931	50.9869
Remove cross-variable branch	540.1113	23.2403	14.0857	52.0225
Replace target-aware pooling with mean pooling	544.4953	23.3344	14.0922	52.1685
Remove regime features	547.6916	23.4028	14.1379	51.9738
Replace gated fusion with simple addition	552.5675	23.5068	14.1914	51.4383
Single expert (remove MoE)	544.2855	23.3299	14.0897	52.0933
Remove directional component	546.846	23.3847	14.1542	52.0358
Remove path decoding (no cumulative sum)	566.9445	23.8106	14.4379	51.5711

 doi:10.20944/preprints202604.0884.v1

Table A15. Ablation results on FTSE100 for MSE (↓), RMSE (↓), MAE (↓), and DA (↑).

Ablation	MSE	RMSE	MAE	DA
Full RADIAN	**494.6048**	**22.2064**	**12.6655**	**50.4575**
Remove RevIN	498.9179	22.3365	12.7199	49.4417
Remove temporal backbone (dilated blocks + temporal attention)	501.0854	22.3849	12.7587	50.1919
Remove cross-variable branch	499.287	22.3447	12.7378	50.0741
Replace target-aware pooling with mean pooling	494.9128	22.2466	12.7015	50.4536
Remove regime features	497.501	22.3047	12.7028	49.9826
Replace gated fusion with simple addition	504.0322	22.4507	12.7944	49.6947
Single expert (remove MoE)	495.8116	22.2668	12.7041	50
Remove directional component	497.7845	22.3111	12.7084	50.2442

Table A16. Ablation results on USDJPY for MSE (↓), RMSE (↓), MAE (↓), and DA (↑).

Ablation	MSE	RMSE	MAE	DA
Full RADIAN	**0.0822**	**0.2859**	**0.1935**	**51.1549**
Remove temporal backbone (dilated blocks + temporal attention)	0.083	0.2881	0.195	50.7018
Remove cross-variable branch	0.0834	0.2889	0.1954	50.7238
Replace target-aware pooling with mean pooling	0.0826	0.2875	0.1945	51.0031
Remove regime features	0.0827	0.2875	0.1945	50.6687
Replace gated fusion with simple addition	0.0834	0.2887	0.1956	50.7018
Single expert (remove MoE)	0.0827	0.2875	0.1946	50.6761
Remove directional component	0.0831	0.2883	0.1952	50.4409
Remove path decoding (no cumulative sum)	0.0861	0.2935	0.2011	50.6136

Table A17. Ablation results on XAUUSD for MSE (↓), RMSE (↓), MAE (↓), and DA (↑).

Ablation	MSE	RMSE	MAE	DA
Full RADIAN	**171.673**	**13.0631**	**8.5652**	**51.5298**
Remove RevIN	179.001	13.3791	8.9185	49.7414
Remove temporal backbone (dilated blocks + temporal attention)	172.3059	13.1265	8.6015	51.4513
Remove cross-variable branch	173.0049	13.1531	8.6372	51.0653
Replace target-aware pooling with mean pooling	172.5208	13.1347	8.6133	51.2853
Remove regime features	172.6089	13.1381	8.6082	51.2815
Replace gated fusion with simple addition	173.424	13.1691	8.6646	50.9302
Single expert (remove MoE)	172.7184	13.1422	8.6282	51.293
Remove directional component	172.599	13.1377	8.6148	51.3046
Remove path decoding (no cumulative sum)	175.9996	13.2665	8.7416	51.0499

Appendix D. Statistical Tests

This section presents the results of the paired statistical significance tests and corresponding effect size estimates employed in the main study.

Table A18 reports the raw paired p-values alongside their Benjamini–Hochberg-adjusted counterparts, accounting for multiple hypothesis testing across nine datasets. Effect sizes are quantified using paired Cohen's d, computed from seed-level MAE differences. Negative values of Cohen's d indicate superior performance of RADIAN relative to the baseline in terms of lower MAE. After correction for multiple comparisons, the number of statistically significant results is reduced, reflecting a more conservative assessment of significance.

 doi:10.20944/preprints202604.0884.v1

Table A18. Summary of paired statistical tests, including raw p-values, Benjamini–Hochberg (BH) adjusted p-values across nine comparisons, and Cohen's d effect sizes computed from seed-level MAE differences.

Dataset	p	p_{BH}	Cohen's d
AAPL	0.0789	0.1283	-1.9312
AUDUSD	0.0339	0.1017	-3.0552
BTCUSD	0.1071	0.1283	-1.6193
DAX30	0.0088	0.0409	-6.1299
ETHUSD	0.0091	0.0409	-6.0182
FTSE100	0.1114	0.1283	-1.5821
US500	0.0501	0.1128	-2.4802
USDJPY	0.1662	0.1662	-1.2334
XAUUSD	0.1141	0.1283	-1.5596

Appendix E. Additional Figures

This appendix presents supplementary visualizations that provide a more granular and diagnostic view of the empirical results discussed in the main text. The figures collectively examine model behavior from complementary perspectives, including cross-dataset performance consistency, sensitivity to random initialization, and qualitative forecast fidelity under varying market conditions. By consolidating ranking-based comparisons, variance diagnostics, and extended prediction plots, this section aims to substantiate the robustness and generalization characteristics of RADIAN relative to established baselines.

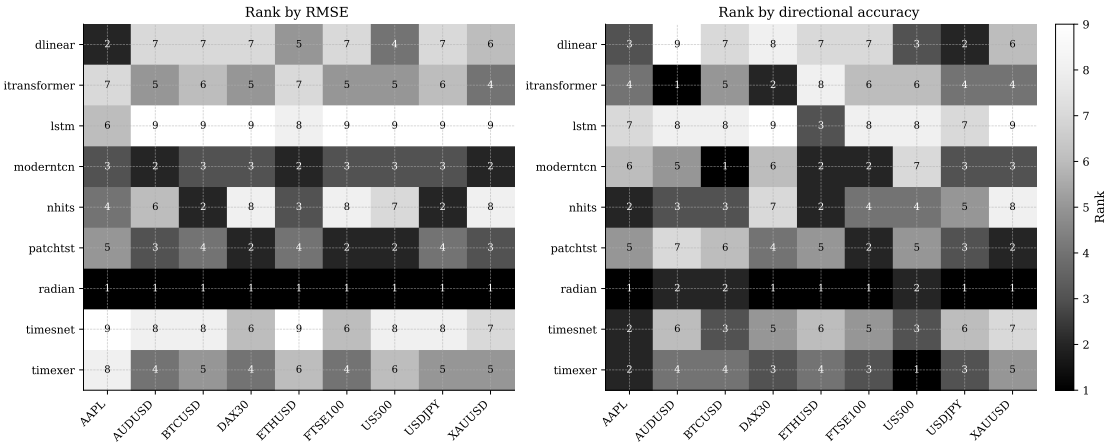


Figure A1. Comparative ranking of forecasting models across all datasets. For each dataset, models receive a rank (1 = best) according to RMSE (left) and directional accuracy (right). Grayscale intensity reflects rank order; annotated cell values provide exact ranks. The figure enables rapid identification of which models dominate particular datasets.

 doi:10.20944/preprints202604.0884.v1

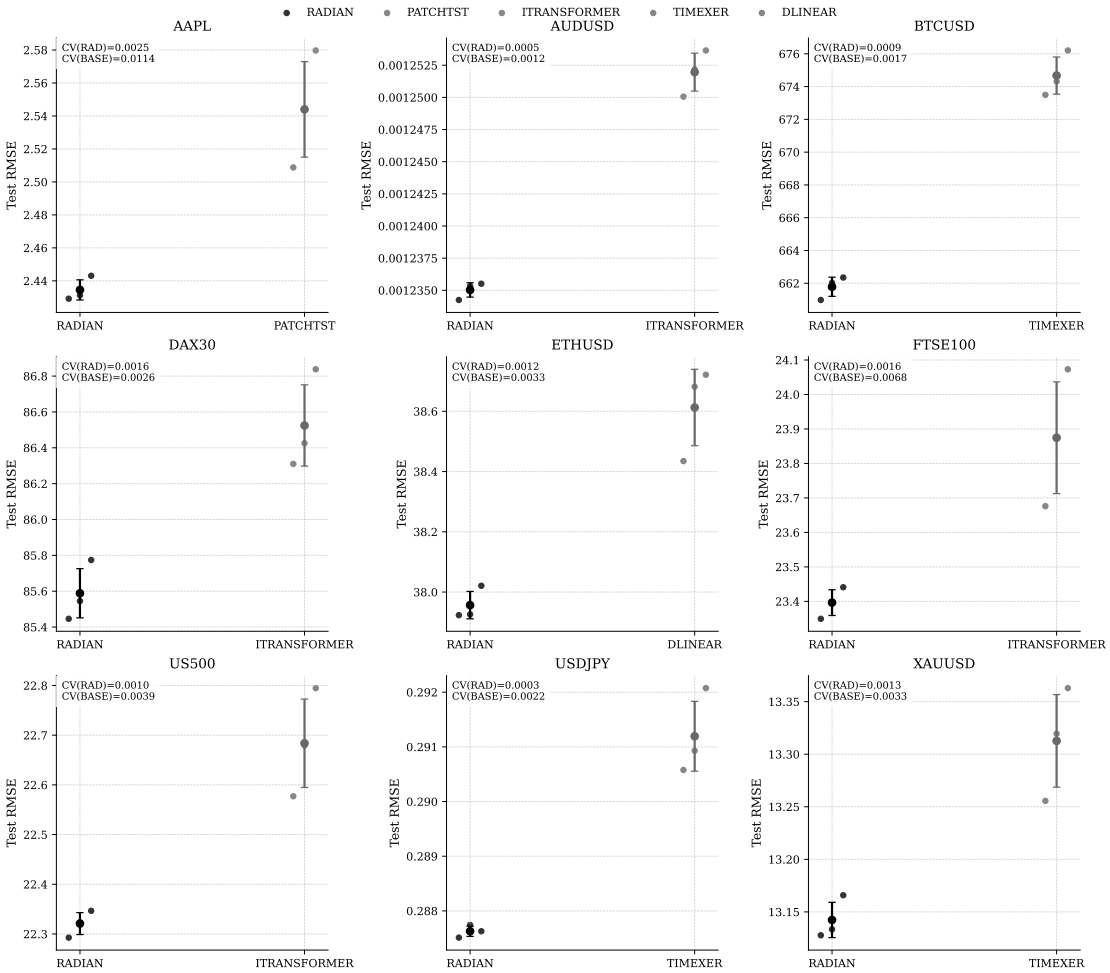


Figure A2. Diagnostic of model stability across training seeds. Each subplot corresponds to a dataset; RADIAN exhibits lower spread and coefficient of variation than the baseline in most cases, as shown by the vertical jitter plots and accompanying statistics.

 doi:10.20944/preprints202604.0884.v1

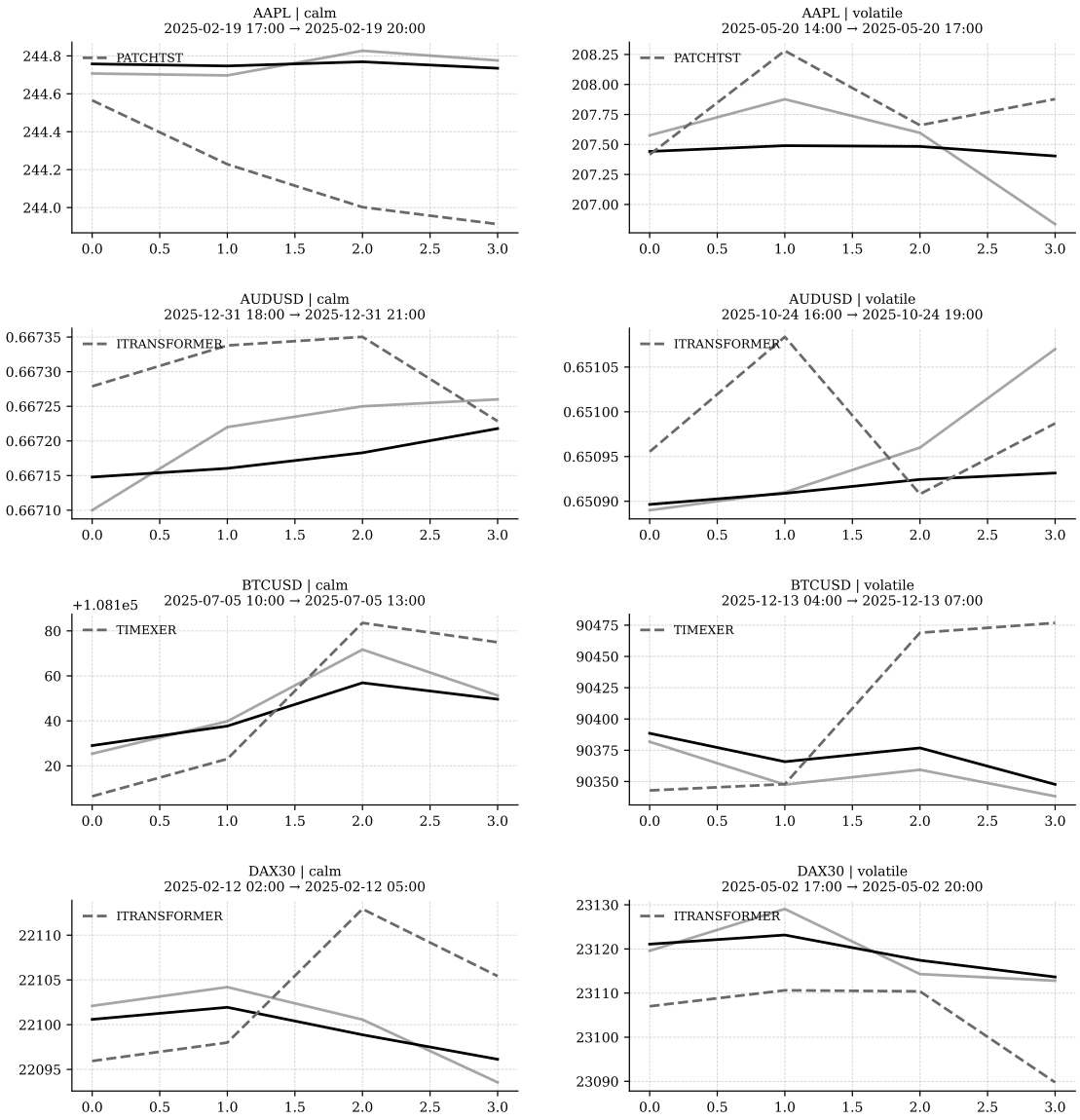


Figure A3. Forecast comparison across datasets and volatility regimes. Each panel shows the ground truth price series alongside RADIAN and baseline predictions over a short forecast horizon.

 doi:10.20944/preprints202604.0884.v1

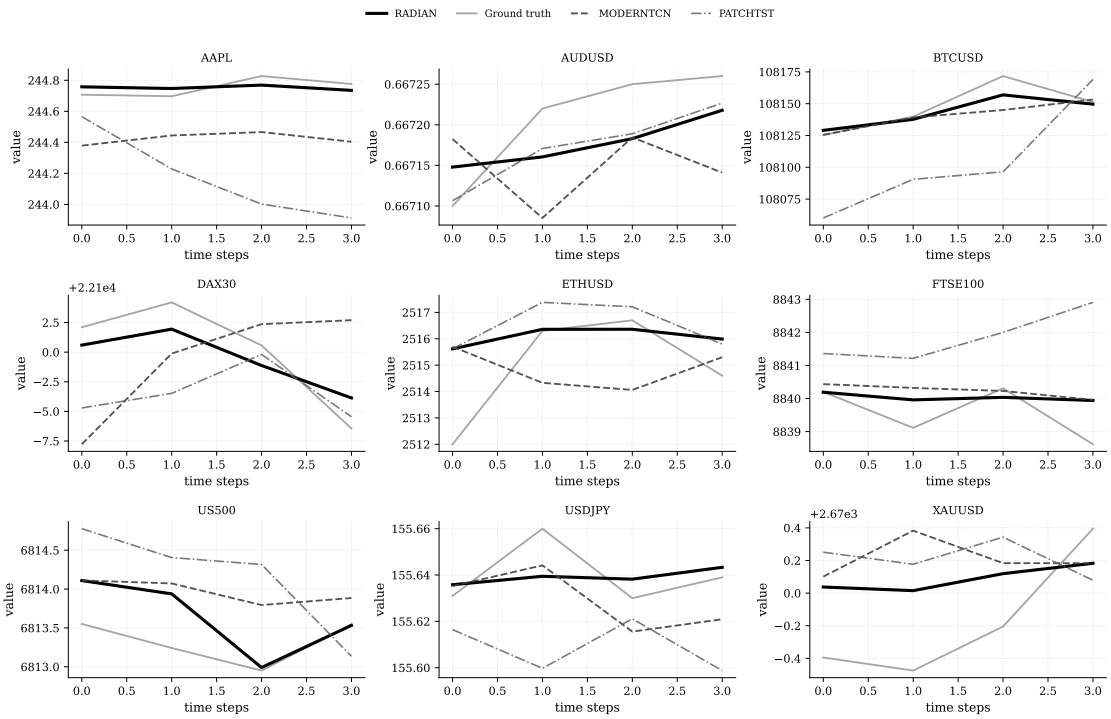


Figure A4. Comparative forecasts across multiple assets. Each panel displays the ground truth price series, RADIAN predictions (black solid line), and forecasts from ModernTCN and PatchTST baselines (distinguished by dashed and dash-dot styles).

Appendix F. Reproducibility

This section summarizes the artifacts and settings required to reproduce preprocessing, training, evaluation, and ablations.

Random seeds and determinism:
Each experiment uses three fixed seeds ({417, 153, 999}). Deterministic backend flags are enabled for GPU/CPU operations (for example, `torch.backends.cudnn.deterministic = True`).

Data splits:
A chronological split is used: 80% training, 10% validation, 10% test. The same split indices are applied consistently across all models and ablations.

Evaluation metrics:
Evaluation reports mean absolute error (MAE), root mean squared error (RMSE), mean squared error (MSE), and directional accuracy (DA). All metrics are computed by a unified evaluation script.

Hyperparameters and ablations:
Hyperparameters are documented in the configuration file included in the supplementary material. Ablation variants (A0–A9) are listed in Table A9 (Appendix C.2) and implemented through configuration flags.

Refer to the primary Data and Code Availability statement following the Conclusion for access to implementation artifacts and processed data.

References

1. Mandelbrot, B. Variation of Certain Speculative Prices. *The Journal of Business* **1963**, *36*, 394–419. https://doi.org/10.1086/294632.
2. Cont, R. Empirical Properties of Asset Returns. *Quantitative Finance* **2001**. https://doi.org/10.1080/713665670.
3. Hamilton, J.D. Economic Analysis of Nonstationary Time Series. *Econometrica* **1989**. https://doi.org/10.2307/1912559.

 doi:10.20944/preprints202604.0884.v1

4. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Computation* **1997**, *9*, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.
5. Salinas, D.; Flunkert, V.; Gasthaus, J.; Januschowski, T. DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks. *International Journal of Forecasting* **2020**, *36*, 1181–1191. https://doi.org/10.1016/j.ijforecast.2019.07.001.
6. Lim, B.; Arik, S.O.; Loeff, N.; Pfister, T. Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting. *International Journal of Forecasting* **2021**, *37*, 1748–1764. https://doi.org/10.1016/j.ijforecast.2021.03.012.
7. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems, 2017. https://doi.org/10.48550/arXiv.1706.03762.
8. Nie, Y.; Nguyen, N.H.; Sinthong, P.; Kalagnanam, J. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In Proceedings of the International Conference on Learning Representations, 2023. https://doi.org/10.48550/arXiv.2211.14730.
9. Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; Long, M. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. In Proceedings of the International Conference on Learning Representations, 2024. https://doi.org/10.48550/arXiv.2310.06625.
10. Box, G.E.P.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. *Time Series Analysis: Forecasting and Control*, 5 ed.; Wiley, 2015. https://doi.org/10.1002/9781118675021.
11. Brockwell, P.J.; Davis, R.A. *Introduction to Time Series and Forecasting*, 3 ed.; Springer, 2016. https://doi.org/10.1007/978-3-319-29854-2.
12. Hyndman, R.J.; Athanasopoulos, G. *Forecasting: Principles and Practice*; OTexts, 2021.
13. Engle, R.F. ARCH. *Econometrica* **1982**. https://doi.org/10.2307/1912773.
14. Bollerslev, T. Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics* **1986**, *31*, 307–327. https://doi.org/10.1016/0304-4076(86)90063-1.
15. Tsay, R.S. *Analysis of Financial Time Series*, 3 ed.; Wiley, 2010. https://doi.org/10.1002/9780470644560.
16. Qin, Y.; Song, D.; Chen, H.; Cheng, W.; Jiang, G.; Cottrell, G. A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction. In Proceedings of the Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017. https://doi.org/10.24963/ijcai.2017/366.
17. Lai, G.; Chang, W.C.; Yang, Y.; Liu, H. Modeling Long- and Short-Term Temporal Patterns with Deep Neural Networks. In Proceedings of the The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018. https://doi.org/10.1145/3209978.3210006.
18. Bai, S.; Kolter, J.Z.; Koltun, V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. In Proceedings of the International Conference on Learning Representations, 2018. https://doi.org/10.48550/arXiv.1803.01271.
19. Oreshkin, B.N.; Carpov, D.; Chapados, N.; Bengio, Y. N-BEATS: Neural Basis Expansion Analysis for Interpretable Time Series Forecasting. In Proceedings of the International Conference on Learning Representations, 2020. https://doi.org/10.48550/arXiv.1905.10437.
20. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence* **2021**, *35*, 11106–11115. https://doi.org/10.48550/arXiv.2012.07436.
21. Wu, H.; Xu, J.; Wang, J.; Long, M. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. *Advances in Neural Information Processing Systems* **2021**, *34*, 22419–22430. https://doi.org/10.48550/arXiv.2106.13008.
22. Zhou, T.; Ma, Z.; Wen, Q.; Wang, X.; Sun, L.; Jin, R.; et al. FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting. In Proceedings of the Proceedings of the 39th International Conference on Machine Learning, 2022. https://doi.org/10.48550/arXiv.2201.12740.
23. Li, S.; Jin, X.; Xuan, Y.; Zhou, X.; Chen, W.; Wang, Y.X.; Yan, X. Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting. In Proceedings of the Advances in Neural Information Processing Systems, 2019. https://doi.org/10.48550/arXiv.1907.00235.
24. Zeng, A.; Chen, M.; Zhang, L.; Xu, Q. Are Transformers Effective for Time Series Forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence* **2022**, *37*, 11121–11128. https://doi.org/10.48550/arXiv.2205.13504.

 doi:10.20944/preprints202604.0884.v1

25. Kim, T.; Kim, J.; Tae, Y.; Park, C.; Choi, J.H.; Choo, J. Reversible Instance Normalization for Accurate Time-Series Forecasting against Distribution Shift. In Proceedings of the International Conference on Learning Representations, 2022.
26. Liu, Y.; Wu, H.; Wang, J.; Long, M. Non-stationary Transformers: Exploring the Stationarity in Time Series Forecasting. In Proceedings of the Advances in Neural Information Processing Systems, 2022. https://doi.org/10.48550/arXiv.2205.14415.
27. Jacobs, R.A. Adaptive Mixtures of Local Experts. *Neural Computation* **1991**, *3*, 79–87. https://doi.org/10.1162/neco.1991.3.1.79.
28. Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; Dean, J. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. *arXiv preprint arXiv:1701.06538* **2017**. https://doi.org/10.48550/arXiv.1701.06538.
29. Fedus, W.; et al. Switch Transformers. *JMLR* **2022**.
30. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer Normalization. *arXiv preprint arXiv:1607.06450* **2016**. https://doi.org/10.48550/arXiv.1607.06450.
31. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations, 2015. https://doi.org/10.48550/arXiv.1412.6980.
32. Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; Long, M. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. *arXiv preprint arXiv:2210.02186* **2022**. https://doi.org/10.48550/arXiv.2210.02186.
33. Challu, C.; et al. N-HiTS: Neural Hierarchical Interpolation for Time Series Forecasting. In Proceedings of the AAAI, 2023. https://doi.org/10.48550/arXiv.2201.12886.
34. Wang, Y.; Wu, H.; Dong, J.; Qin, G.; Zhang, H.; Liu, Y.; Qiu, Y.; Wang, J.; Long, M. TimeXer: Empowering Transformers for Time Series Forecasting with Exogenous Variables. *arXiv preprint arXiv:2402.19072* **2024**. https://doi.org/10.48550/arXiv.2402.19072.
35. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B* **1995**, *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
36. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*; Lawrence Erlbaum, 1988. https://doi.org/10.4324/9780203771587.
37. Hazan, E. Introduction to Online Convex Optimization. *Foundations and Trends in Optimization* **2016**, *2*, 157–325. https://doi.org/10.1561/2400000013.
38. Reichstein, M. Deep Learning for Earth System Science. *Nature* **2019**, *566*, 195–204. https://doi.org/10.1038/s41586-019-0912-1.
39. Benson, T.; Anand, A.; Akella, A.; Zhang, M. Network Traffic Characteristics of Data Centers in the Wild. In Proceedings of the IMC, 2010, pp. 267–280. https://doi.org/10.1145/1879141.1879175.
40. Clifford, G.D.; et al. AF Classification from a Short Single Lead ECG Recording. *Computing in Cardiology* **2017**. https://doi.org/10.22489/CinC.2017.065-469.