PERCEIVE, ROUTE, MODULATE: DYNAMIC PATTERN RECALIBRATION FOR TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Local temporal patterns in real-world time series continuously shift across regimes. While existing forecasting backbones express input-dependent responses through shared transformations, they lack an explicit mechanism to separate shared representation learning from local feature-gain adaptation. We introduce **Dynamic Pattern Recalibration** (DPR), a backbone-agnostic mechanism that realizes this separation via token-level modulation. Through a lightweight "Perceive-Route-Modulate" pipeline, DPR computes a soft-routing distribution over a compact learned response basis, generating a context-dependent modulation vector that recalibrates hidden states via a residual Hadamard product. Across ten forecasting backbones, DPR matches or improves the raw backbone's MSE in 87 of 95 backbone-dataset pairs (91.6%), with larger gains associated with greater local volatility variation. Its minimalist standalone instantiation, DPRNet, achieves strong forecasting performance across eight benchmarks with low measured computational cost. Our code is available at `https://anonymous.4open.science/r/DPRNet`.

1 INTRODUCTION

Time series forecasting is central to decision-making in systems with shifting dynamics, ranging from financial markets and energy grids to climate, epidemiology, and system diagnostics (Idrees et al., 2019; Karevan & Suykens, 2020; Deb et al., 2017; Zheng & Huang, 2020; Ruan et al., 2025; Zhang et al., 2024; Liu et al., 2025; Zou et al., 2025). Such systems frequently alternate among stable trends, periodic oscillations, and sudden anomalies (Hamilton, 1990; Liu et al., 2023a), where local dynamics can shift rapidly within a single sequence (Figure 1(a)). Forecasting therefore requires models that adapt to changing local regimes (Liu et al., 2022; 2023b). Yet standard forecasting backbones rely on globally shared transformations across temporal tokens, lacking an explicit mechanism to recalibrate feature responses across shifting local dynamics.

Recent advances span deep forecasting architectures (Nie et al., 2023; Liu et al., 2024), linear and minimalist models, and large or sparse models (Das et al., 2024; Ansari et al., 2024; Shi et al., 2025; Ma et al., 2025). These architectures typically apply feature transformations using parameters Θ shared across temporal tokens. While nonlinearities, attention, and normalization already make their effective responses input-dependent, a complementary design opportunity is to explicitly separate shared representation learning from local, context-conditioned feature gains. Such a path provides a direct parameterization for adapting the response of individual tokens as local dynamics shift. Sequence-level preprocessing such as RevIN (Kim et al., 2022) adjusts input scale, whereas this adaptation acts directly on hidden representations within the backbone.

Adaptive feature transformations have explored diverse conditioning schemes: SE-Net (Hu et al., 2018) excites channels via globally pooled gates, FiLM (Perez et al., 2018) applies feature-wise affine conditioning, and CondConv (Yang et al., 2019) mixes dynamic convolution kernels. In temporal modeling, Time-MoE (Shi et al., 2025) scales conditional capacity through sparse expert routing, while TimeExpert (Ma et al., 2025) selects temporal context experts and FreqMoE (Liu, 2025) combines frequency-band experts. Sparse expert dispatch introduces routing and load-balancing costs (Fedus et al., 2022). For lightweight forecasting, we instead focus on fine-grained local conditioning combined with a compact, continuous mixture of feature-gain vectors (Figure 1(b–d)).

We study whether an explicit token-level feature-gain path can improve forecasting while preserving the shared backbone architecture. **Dynamic Pattern Recalibration** (DPR), a three-stage Perceive–

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

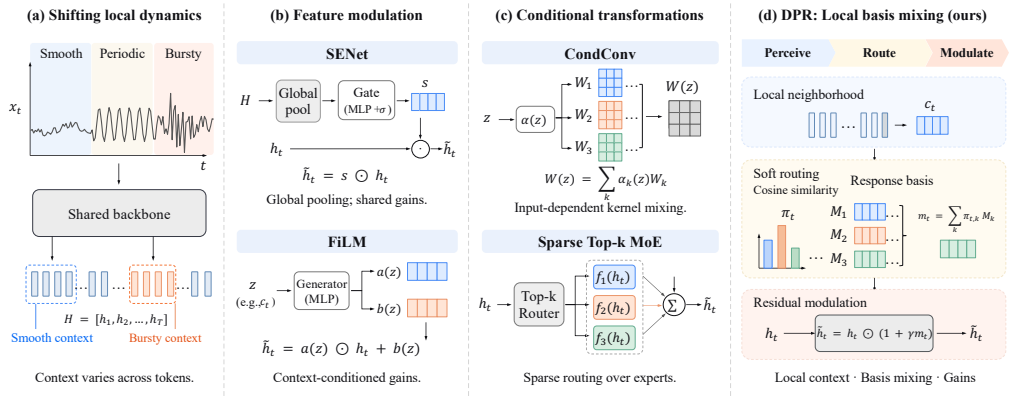


Figure 1: **From shifting local dynamics to dynamic pattern recalibration.** **(a)** Local dynamics vary within a sequence, giving tokens different local contexts. **(b)** Feature-modulation approaches recalibrate via globally pooled or conditioning-dependent responses. **(c)** Conditional-transformation approaches adapt full operators by mixing kernels or routing to experts. **(d)** DPR perceives each token's local context, composes a shared response basis, and applies a feature-wise residual gain.

Route–Modulate adapter, realizes this idea through local context perception, continuous routing over a compact shared response basis, and residual feature-wise modulation (Figure 1(d)), separating shared representation learning from lightweight local response adaptation. Perceive extracts multi-scale local context with depthwise convolutions (or patch-local features) and compresses each token into a routing query. Route applies a scaled cosine softmax over K learned routing prototypes; the resulting probabilities compose a continuous modulation vector from an orthogonally regularized response basis. Modulate then applies this vector via a residual Hadamard product to recalibrate hidden states (e.g., immediately prior to the prediction head in the plug-in setting). The resulting design combines local token-level conditioning, continuous basis mixing, and feature-wise residual modulation while preserving the host backbone's temporal aggregation. The adapter scales linearly in token count and removes the need for token dispatch and load balancing. We evaluate DPR across ten mainstream backbones and introduce a standalone minimalist instantiation, DPRNet. Our contributions are:

- **Dynamic Pattern Recalibration.** We formulate local adaptation as a structured response-basis parameterization: each token selects a continuous combination of shared feature-response directions, separating shared representation learning from local gain control.
- **Cross-Backbone Validation and Inductive-Bias Characterization.** Across ten diverse forecasting backbones (attention, convolution, MLP, GNN), DPR matches or improves the raw backbone's MSE in 87 of 95 evaluations (91.6%), with minimal parameter overhead. Gains correlate with local volatility variation, revealing when response recalibration is most useful.
- **A Minimalist Instantiation (DPRNet).** We instantiate DPR within a minimalist patch-based MLP, achieving strong accuracy across eight benchmarks with low measured computational cost compared to leading Transformer and mixing architectures.

2 Preliminary

2.1 Problem Formulation

In multivariate time series forecasting, we aim to predict a future horizon $\mathcal{Y} \in \mathbb{R}^{S\times C}$ given a historical lookback window $\mathcal{X} \in \mathbb{R}^{T\times C}$, where T, S, and C denote lookback length, forecast horizon, and the number of variates, respectively. A forecaster typically encodes $\mathcal{X}$ into hidden representations $\mathbf{H} \in \mathbb{R}^{B\times L\times d}$, with batch size B, L temporal tokens, and hidden dimension d.

For a sequence token at position $\ell \in \{1, \dots, L\}$ and layer j, let $\mathcal{F}_\Theta$ denote a shared transformation (e.g., FFN, linear projection, or attention) parameterized by Θ. Standard architectures apply shared mappings where token responses can already implicitly depend on inputs through nonlinearities:

$$\mathbf{h}_\ell^{(j)} = \mathcal{F}_\Theta\big(\tilde{\mathbf{h}}_\ell^{(j-1)}\big), \qquad \tilde{\mathbf{h}}_\ell^{(j)} = \mathbf{h}_\ell^{(j)} \odot \big(\mathbf{1} + \mathcal{M}(\mathbf{c}_\ell)\big), \tag{1}$$

where $\tilde{\mathbf{h}}_\ell^{(0)} = \mathbf{h}_\ell^{(0)}$ denotes the initial embedding. The second expression introduces an explicit, token-specific residual modulation via $\mathcal{M}(\mathbf{c}_\ell) \in \mathbb{R}^d$, where $\mathbf{c}_\ell$ summarizes the local hidden-state context surrounding position ℓ and $\odot$ denotes the element-wise Hadamard product. In this formulation, $\mathcal{M}(\mathbf{c}_\ell)$ dynamically amplifies or attenuates feature responses per token, while Θ governs shared representation learning. DPR parameterizes $\mathcal{M}(\mathbf{c}_\ell)$ as a continuous convex combination over a compact, learned response basis. When $\mathcal{M}(\mathbf{c}_\ell) \to \mathbf{0}$, Eq. (1) smoothly reduces to the unmodulated base model. For late-stage plug-in integration across diverse hosts, this recalibration is applied once to the final hidden representation immediately prior to the prediction head.

2.2 Related Work

Deep Time Series Forecasting Architectures. Recent advances in deep forecasting generally fall into two categories: architectural priors and capacity scaling. Architectural innovations introduce specialized structural inductive biases, spanning temporal/channel attention (Vaswani et al., 2017; Zhou et al., 2021; Wu et al., 2021; Zhang & Yan, 2023; Nie et al., 2023; Liu et al., 2024), linear operators (Zeng et al., 2023), 2D convolutions (Wu et al., 2023), multi-scale mixing (Wang et al., 2024; Murad et al., 2025), and graph filtration (Hu et al., 2025). Capacity-scaling methods expand model expressiveness via dense foundation models (Das et al., 2024; Ansari et al., 2024; Woo et al., 2024) or sparse Mixture-of-Experts (MoE) routing (Shazeer et al., 2017; Fedus et al., 2022; Shi et al., 2025). Forecasting-specific MoE variants also select temporal context experts (Ma et al., 2025) or combine frequency-band experts (Liu, 2025). Soft MoE further explores differentiable assignment over full tokens (Puigcerver et al., 2024). Unlike sparse MoE architectures that route tokens through distinct sub-networks with dispatch and load-balancing overhead, DPR routes over a compact feature-response basis on a shared input representation. Operating with strictly linear token complexity, this intra-token gain modulation complements the host backbone's temporal aggregation.

Conditional Feature Recalibration. Adaptive recalibration has been extensively studied in computer vision: SE-Net (Hu et al., 2018) excites channels via global pooling, CBAM (Woo et al., 2018) adds spatial attention, FiLM (Perez et al., 2018) modulates features via task-conditioned affine parameters, and CondConv (Yang et al., 2019) dynamically generates convolutional weights. For sequential tasks, Temporal FiLM (Birnbaum et al., 2019) explores recurrent modulation in audio processing. DPR constrains local gains to a shared response set and adapts by mixing within this set, a parameterization distinct from the direct affine generation of FiLM-style conditioning. In time series forecasting, methods addressing distribution shifts predominantly operate at the sequence or temporal operator level: RevIN (Kim et al., 2022) applies sequence-level affine normalization to counter non-stationarity, while Non-stationary Transformers (Liu et al., 2022) and Koopa (Liu et al., 2023b) deconstruct non-stationary dynamics via modified attention and Koopman operators. DPR introduces a distinct latent-space parameterization: rather than normalizing input statistics in the time domain, it performs self-conditioned, token-level feature-gain recalibration directly in the latent representation space. By composing gains from a compact, orthogonally regularized basis, DPR explicitly separates shared representation learning from local pattern adaptation.

3 Methodology

We instantiate the dynamic mapping formulated in Eq. (1) as **Dynamic Pattern Recalibration** (DPR) (Figure 2). In the plug-in setting, DPR recalibrates hidden states immediately prior to the prediction head across diverse backbones. In the standalone setting, we introduce DPRNet, a minimalist patch-based MLP. DPR operates via a three-stage Perceive–Route–Modulate pipeline that computes token-level feature modulation conditioned on localized temporal dynamics.

3.1 Dynamic Pattern Recalibration (DPR)

The DPR mechanism extracts time-aware modulation vectors from a compact learned basis of adaptive response patterns. For a hidden representation $\mathbf{h}_{b,\ell} \in \mathbb{R}^d$ at temporal position $\ell \in \{1, \ldots, L\}$, DPR modulates feature responses through three sequential stages:

Perceive: Multi-Scale Context. To dynamically capture localized temporal shifts (e.g., sudden spikes or meso-scale regime transitions), DPR extracts context features via parallel depthwise 1D convolutions. Permuting the input to a channel-first layout $\mathbf{H}_{\text{cf}} \in \mathbb{R}^{B \times d \times L}$, we apply two parallel

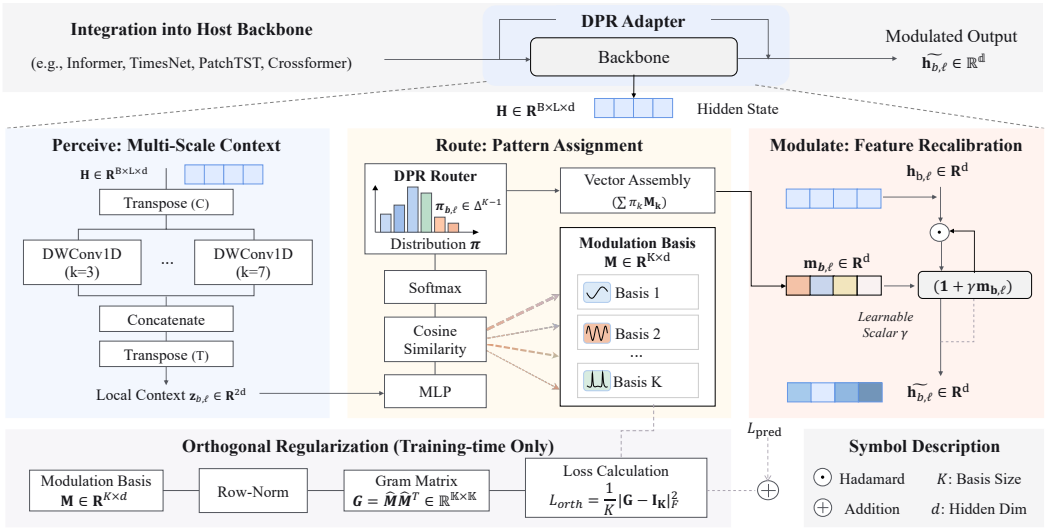


Figure 2: Overview of DPR. The plug-in path recalibrates hidden states before the prediction head. Local queries are compared with routing keys $\mathbf{E} \in \mathbb{R}^{K\times d_c}$; the resulting probabilities mix feature-gain values $\mathbf{M} \in \mathbb{R}^{K\times d}$ for a residual Hadamard gain.

depthwise convolutions with complementary receptive fields ($k_1 = 3, k_2 = 7$) using same-padding, and concatenate them along the feature dimension:

$$\mathbf{Z}_{\text{cf}} = [\text{DWConv}_{k_1}(\mathbf{H}_{\text{cf}})\,;\,\text{DWConv}_{k_2}(\mathbf{H}_{\text{cf}})] \in \mathbb{R}^{B\times 2d\times L}, \tag{2}$$

which is permuted back to yield the perceived context $\mathbf{Z} \in \mathbb{R}^{B\times L\times 2d}$. DPR consumes only the hidden states available to the host at inference. These states may represent historical patches or forecast-position latent tokens (Appendix C). Depthwise convolutions process hidden-feature channels independently. For hosts that process variables independently, reshaping to $(B \cdot C) \times L_p \times d$ and applying DPR separately to each variable strictly preserves Channel Independence (CI) (Nie et al., 2023), where L_p is the patch-token count. For patch-based architectures where each individual token already aggregates temporal dependencies across a localized window of length P, a pointwise convolution ($k = 1$) is employed to extract pattern queries directly from the patch-level channel representations. In general, let d_{ctx} denote the perceived context dimension ($d_{\text{ctx}} = 2d$ for multi-scale depthwise convolutions, or d for pointwise convolutions).

Route: Pattern Assignment. The perceived context $\mathbf{z}_{b,\ell} \in \mathbb{R}^{d_{\text{ctx}}}$ is projected via a shallow MLP into a compact context query $\mathbf{c}_{b,\ell} = \text{GELU}(\mathbf{W}_1\mathbf{z}_{b,\ell} + \mathbf{b}_1) \in \mathbb{R}^{d_c}$, where $\mathbf{W}_1 \in \mathbb{R}^{d_c\times d_{\text{ctx}}}$ and $d_c = \max(16, \lfloor d/4 \rfloor)$. We introduce K learnable routing prototypes $\mathbf{E} \in \mathbb{R}^{K\times d_c}$ acting as routing keys that anchor different regions of the local-context space. The soft-routing probability vector $\boldsymbol{\pi}_{b,\ell} \in \Delta^{K-1}$ is determined via temperature-scaled cosine similarity over the centroids:

$$\boldsymbol{\pi}_{b,\ell} = \text{softmax}\left(\tau\,\hat{\mathbf{E}}\hat{\mathbf{c}}_{b,\ell}\right), \quad \text{where} \quad \pi_{b,\ell,k} = \frac{\exp(\tau\langle\hat{\mathbf{c}}_{b,\ell}, \hat{\mathbf{E}}_k\rangle)}{\sum_{k'=1}^{K}\exp(\tau\langle\hat{\mathbf{c}}_{b,\ell}, \hat{\mathbf{E}}_{k'}\rangle)}, \tag{3}$$

where $\hat{\mathbf{c}}$ and $\hat{\mathbf{E}}$ denote L_2-normalized vectors, and τ is a learnable logit scale initialized to 2.0. Normalization prevents magnitude from dominating assignment, while τ controls routing entropy.

Modulate: Feature Recalibration. DPR maps the routing distribution to a learned basis of adaptive response patterns $\mathbf{M} \in \mathbb{R}^{K\times d}$. It composes a continuous modulation vector via convex combination and recalibrates the hidden representation through a residual Hadamard product:

$$\mathbf{m}_{b,\ell} = \sum_{k=1}^{K}\pi_{b,\ell,k}\mathbf{M}_{k,:}, \qquad \tilde{\mathbf{h}}_{b,\ell} = \mathbf{h}_{b,\ell} \odot \left(\mathbf{1} + \gamma\mathbf{m}_{b,\ell}\right), \tag{4}$$

where γ is a learnable scaling factor initialized to 0.1. The residual formulation supports a near-identity mapping when $\|\gamma\mathbf{m}_{b,\ell}\|_\infty \ll 1$, ensuring stable initialization without perturbing the host backbone's learned representations at the onset of training. While a scalar gate scales all channels

uniformly, this basis mixture supports fine-grained, token-dependent, feature-wise recalibration across response modes (see Appendix D for a mechanism-level analysis of this inductive bias).

Orthogonal Regularization. To eliminate basis redundancy and encourage geometrically diverse response directions, we impose an orthogonal penalty during training. Let $\hat{\mathbf{M}}$ denote the row-wise L_2-normalized basis and $\mathbf{G} = \hat{\mathbf{M}}\hat{\mathbf{M}}^\top \in \mathbb{R}^{K\times K}$ its Gram matrix:

$$\mathcal{L}_{\text{orth}} = \frac{1}{K}\left\|\mathbf{G} - \mathbf{I}_K\right\|_F^2, \qquad \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pred}} + \lambda_{\text{orth}}\mathcal{L}_{\text{orth}}. \tag{5}$$

Because the rows of $\hat{\mathbf{M}}$ are unit-normalized, the diagonal elements of $\mathbf{G}$ are identically 1. Hence, $\mathcal{L}_{\text{orth}}$ strictly penalizes off-diagonal elements, directly minimizing the sum of squared off-diagonal cosines $\sum_{i\neq j}\cos^2(\mathbf{M}_i, \mathbf{M}_j)$ across response directions without test-time computational cost.

Response-basis view. DPR parameterizes local adaptation as a continuous mixture over a compact set of shared response directions: each token selects a mixture of diagonal feature-response transformations on the same hidden representation. This yields a structured adaptation space distinct from direct gate generation and from routing through full expert networks (Appendix D).

3.2 DPRNet: A Minimalist Instantiation

DPRNet instantiates this mechanism within a minimalist, Transformer-free patch MLP architecture. Given input $\mathcal{X} \in \mathbb{R}^{T\times C}$, we adopt CI and RevIN (Kim et al., 2022) to normalize instance-level distribution shifts. The series is partitioned into L overlapping patches of length P and stride s, and linearly projected to dimension d, yielding initial tokens $\mathbf{H}^{(0)} \in \mathbb{R}^{B\times L\times d}$ (where B denotes the effective batch size $B \times C$ under CI). The network stacks N_{blocks} DPR Blocks. At layer $j \in \{1, \ldots, N_{\text{blocks}}\}$, a standard residual MLP first constructs the shared base representation:

$$\mathbf{U}^{(j)} = \mathbf{H}^{(j-1)} + \mathcal{F}_{\text{MLP}}\left(\text{LayerNorm}(\mathbf{H}^{(j-1)})\right). \tag{6}$$

We then apply DPR to the normalized hidden state:

$$\mathbf{H}^{(j)} = \text{DPR}\left(\text{LayerNorm}(\mathbf{U}^{(j)})\right). \tag{7}$$

Because the DPR operation natively incorporates an internal residual identity ($\tilde{\mathbf{h}} = \mathbf{h} \odot (\mathbf{1} + \gamma\mathbf{m})$), no outer residual is required around the DPR block. When $\gamma \to 0$, the block smoothly degenerates to an MLP block followed by normalization. Finally, the sequence $\mathbf{H}^{(N_{\text{blocks}})}$ is flattened and mapped to the forecast horizon $\widehat{\mathcal{Y}} \in \mathbb{R}^{S\times C}$ through a linear head, followed by RevIN denormalization.

Parameter and Computational Efficiency. DPR introduces a compact set of parameters: depthwise convolutions ($d \cdot k$ per kernel), a context projection ($d_{\text{ctx}} \cdot d_c$), K routing prototypes of size d_c, and a response basis of size $K \times d$. With standard settings ($K = 8, d = 256$), the basis contains merely 2,048 parameters, completely decoupled from sequence length or forecast horizon. Per sequence, context projection scales as $\mathcal{O}(Ld_{\text{ctx}}d_c)$, routing as $\mathcal{O}(LKd_c)$, basis assembly as $\mathcal{O}(LKd)$, and Hadamard modulation as $\mathcal{O}(Ld)$. Because soft routing avoids discrete token dispatch and load-balancing losses, all operations remain dense, parallel, and strictly linear in sequence length L.

4 Experiments

Our core evaluation covers 11 real-world datasets to systematically investigate four research questions:

- **RQ1 (Cross-Backbone Transfer)**: Can DPR enhance diverse backbones? (Section 4.1)
- **RQ2 (Feature Modulation & Volatility)**: How does DPR compare with parameter-matched modulation alternatives, and does its gain systematically correlate with local volatility? (Section 4.2)
- **RQ3 (Standalone Forecasting)**: Can DPR support an accurate and computationally efficient minimalist forecaster? (Section 4.3)
- **RQ4 (Design Principles & Scaling)**: How does DPR compare against static capacity scaling, sparse MoE routing, and structural ablations? (Section 4.4)

Datasets and metrics. Our evaluation centers on an 11-dataset matrix spanning energy, finance, climatology, epidemiology, healthcare, cloud telemetry, and solar physics, covering monthly to sub-hourly resolutions and diverse multivariate series. Figure 3 maps spectral entropy (H_s, frequency complexity) (Inouye et al., 1991) against volatility-of-volatility (VoV, a local volatility-variation diagnostic defined in Appendix A.2). We evaluate models using standard MSE and MAE; dataset descriptions, splits, and prediction horizons are detailed in Appendices A and G.

Backbones and protocol. The plug-in study focuses on seven backbones spanning diverse model families: Informer (Zhou et al., 2021), Crossformer (Zhang & Yan, 2023), TimesNet (Wu et al., 2023), PatchTST (Nie et al., 2023), TimeMixer (Wang et al., 2024), TimeFilter (Hu et al., 2025), and WPMixer (Murad et al., 2025). Adapter experiments retain all host hyperparameters and apply a unified late-stage insertion rule, tuning only DPR-specific parameters. Default DPR parameters are set to: basis size $K = 8$, routing bottleneck width d_c, logit scale τ initialized to 2.0, and residual gain γ initialized to 0.1. Models are optimized with Adam (Kingma & Ba, 2014); full implementation details are in Appendix B. Primary host results are reported in Table 1, with full horizon breakdowns in Appendix Table 12.

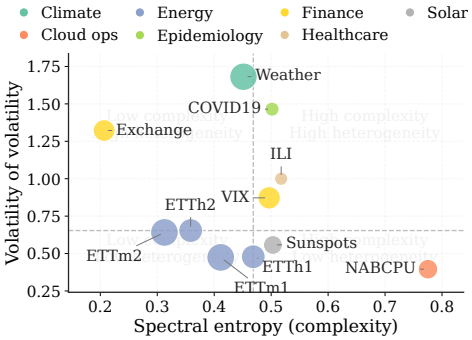


Figure 3: **Dataset diversity landscape.** Spectral entropy (H_s) versus volatility-of-volatility (VoV); bubble area denotes series length and color encodes domain.

4.1 DPR Transfers Across Forecasting Backbones (RQ1)

Transfer consistency. We evaluate DPR across forecasting backbones as a plug-in recalibrating hidden states prior to the prediction head. For multi-scale backbones, the adapter is shared across late-stage branches. Figure 4 summarizes the gain distributions and association with local volatility. The core matrix spans 77 backbone–dataset pairs (11 datasets $\times$ 7 hosts), plus 18 pairs from three hosts (OLinear, TimeMixer++, TimeBase) across six benchmarks in Appendix G.3. Across ten architectures and 95 evaluations, DPR matches or improves MSE in 87 pairs (91.6%). Improvements are typically modest on regular, low-heterogeneity benchmarks such as ETTh1 and ETTm1, consistent with the weaker benefit of local recalibration in more regular regimes.

Gain magnitude. Figure 4(a) summarizes relative MSE reductions across the seven core backbones in Table 1, averaged over nine datasets. Mean relative MSE reductions reach 1.7%, 3.3%, 2.9%, and 1.7% for PatchTST, TimeMixer, TimeFilter, and WPMixer, respectively, alongside a 25.7% reduction on Informer. Across all eleven datasets, modern hosts achieve 35 improvements, five ties, and four degradations across 44 raw comparisons.

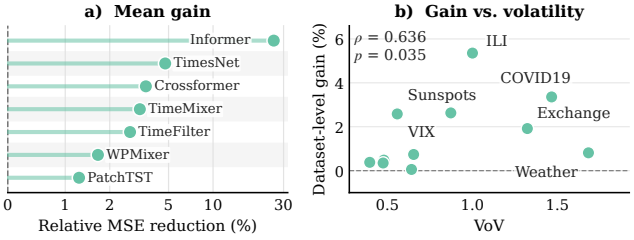


Figure 4: **DPR transfers across forecasting backbones.** **(a)** Mean relative MSE reduction over the nine datasets and seven backbones in Table 1 (symlog axis). **(b)** Dataset-level gain vs. VoV ($\rho = 0.636$, unadjusted $p = 0.035$, $n = 11$).

4.2 Modulation Controls and Volatility Association (RQ2)

Matched conditional-modulation controls. To isolate the efficacy of our shared-basis parameterization, we compare DPR against four competitive modulation mechanisms inserted at identical locations and strictly matched within 5% parameter budget. Global SE (Hu et al., 2018) pools temporal tokens into a global channel gate; Local SE utilizes DPR's receptive field to predict channel scales directly; Local FiLM (Perez et al., 2018) generates per-token affine parameters from local queries; and Gated Residual directly predicts element-wise gates without basis composition. Across 12 host–dataset configurations (Table 2), DPR outperforms the unadapted baseline in every case and achieves the best or tied-best MSE in 9 of the 12. These controlled results support the effectiveness of continuous shared-basis mixing relative to the evaluated direct-gating alternatives.

Table 1: **Backbone-agnostic DPR enhancement.** Four-horizon averages; gray = +DPR; ↑xx% = MSE reduction vs. Raw; last row = mean relative MSE/MAE reduction. Full results: Appendix Table 12.

Dataset	Metric	Informer (2021)		Crossformer (2023)		TimesNet (2023)		PatchTST (2023)		TimeMixer (2024)		TimeFilter (2025)		WPMixer (2025)	
		Raw	+DPR	Raw	+DPR	Raw	+DPR	Raw	+DPR	Raw	+DPR	Raw	+DPR	Raw	+DPR
ILI	MSE	7.094	**5.740** ↑19.1%	5.035	**4.699** ↑6.7%	6.021	**4.420** ↑26.6%	3.321	**3.114** ↑6.2%	3.182	**3.147** ↑1.1%	2.438	**2.294** ↑5.9%	3.096	**2.965** ↑4.2%
	MAE	1.896	**1.761** ↑7.1%	1.542	**1.466** ↑4.9%	1.309	**1.208** ↑7.7%	1.110	**1.095** ↑1.4%	1.148	**1.145** ↑0.3%	0.978	**0.951** ↑2.8%	1.069	**1.062** ↑0.7%
COVID19	MSE	1.917	**1.742** ↑9.1%	1.401	**1.326** ↑5.4%	1.067	**1.035** ↑3.0%	0.839	**0.814** ↑3.0%	0.930	**0.864** ↑7.1%	0.908	**0.797** ↑12.2%	0.839	**0.786** ↑6.3%
	MAE	**0.632**	0.634 -0.3%	**0.476**	**0.476** 0.0%	0.538	**0.531** ↑1.3%	0.362	**0.359** ↑0.8%	0.401	**0.386** ↑3.7%	0.394	**0.374** ↑5.1%	0.363	**0.359** ↑1.1%
VIX	MSE	1.261	**0.939** ↑25.5%	1.139	**1.102** ↑3.2%	1.207	**1.134** ↑6.0%	1.144	**1.121** ↑2.0%	1.141	**1.117** ↑2.1%	1.115	**1.094** ↑1.9%	1.119	**1.105** ↑1.3%
	MAE	0.744	**0.669** ↑10.1%	0.685	**0.678** ↑1.0%	0.715	**0.690** ↑3.5%	0.692	**0.684** ↑1.2%	0.694	**0.683** ↑1.6%	0.677	**0.672** ↑0.7%	0.686	**0.679** ↑1.0%
Exchange	MSE	2.782	**1.723** ↑38.1%	0.894	**0.840** ↑6.0%	0.494	**0.476** ↑3.6%	0.453	**0.448** ↑1.1%	**0.440**	**0.440** 0.0%	0.446	**0.443** ↑0.7%	0.445	**0.442** ↑0.7%
	MAE	1.235	**0.948** ↑23.2%	0.646	**0.625** ↑3.3%	0.488	**0.480** ↑1.6%	0.454	**0.451** ↑0.7%	**0.446**	**0.446** 0.0%	0.452	**0.449** ↑0.7%	0.449	**0.447** ↑0.4%
NABCPU	MSE	2.689	**1.926** ↑28.4%	1.257	**1.236** ↑1.7%	1.220	**1.217** ↑0.2%	1.195	**1.194** ↑0.1%	1.196	**1.194** ↑0.2%	1.197	**1.195** ↑0.2%	1.202	**1.190** ↑1.0%
	MAE	1.035	**0.777** ↑24.9%	0.389	**0.372** ↑4.4%	0.354	**0.352** ↑0.6%	0.318	**0.316** ↑0.6%	0.305	**0.300** ↑1.6%	0.307	**0.306** ↑0.3%	0.334	**0.312** ↑6.6%
Sunspots	MSE	2.562	**1.514** ↑40.9%	0.376	**0.357** ↑5.1%	0.702	**0.696** ↑0.9%	0.684	**0.680** ↑0.6%	0.722	**0.689** ↑4.6%	0.744	**0.705** ↑5.2%	0.679	**0.672** ↑1.0%
	MAE	1.296	**1.001** ↑22.8%	0.441	**0.424** ↑3.9%	0.578	**0.575** ↑0.5%	0.576	**0.575** ↑0.2%	0.586	**0.576** ↑1.7%	0.590	**0.578** ↑2.0%	0.576	**0.573** ↑0.5%
Weather	MSE	0.484	**0.386** ↑20.2%	0.241	**0.239** ↑0.8%	0.289	**0.285** ↑1.4%	0.251	**0.248** ↑1.2%	0.282	**0.243** ↑13.8%	0.246	**0.244** ↑0.8%	**0.243**	**0.243** 0.0%
	MAE	0.432	**0.381** ↑11.8%	0.268	**0.267** ↑0.4%	0.305	**0.303** ↑0.7%	0.270	**0.269** ↑0.4%	0.301	**0.263** ↑12.6%	0.267	**0.265** ↑0.7%	**0.264**	**0.264** 0.0%
ETTh1	MSE	1.324	**1.000** ↑24.5%	0.449	**0.441** ↑1.8%	0.534	**0.531** ↑0.6%	0.459	**0.457** ↑0.4%	0.458	**0.456** ↑0.4%	**0.451**	0.454 -0.7%	0.454	**0.453** ↑0.2%
	MAE	0.816	**0.735** ↑9.9%	0.445	**0.440** ↑1.1%	**0.492**	0.493 -0.2%	0.432	**0.415** ↑3.9%	0.429	**0.428** ↑0.2%	**0.429**	0.430 -0.2%	**0.431**	**0.431** 0.0%
ETTm1	MSE	1.480	**1.097** ↑25.9%	0.413	**0.400** ↑3.1%	0.519	**0.517** ↑0.4%	0.396	**0.395** ↑0.3%	0.393	**0.392** ↑0.3%	0.395	**0.394** ↑0.3%	0.390	**0.389** ↑0.3%
	MAE	0.855	**0.766** ↑10.4%	0.404	**0.397** ↑1.7%	0.474	**0.472** ↑0.4%	**0.387**	**0.387** 0.0%	**0.385**	**0.385** 0.0%	0.387	**0.386** ↑0.3%	**0.384**	**0.384** 0.0%
Avg. Impv.		**25.7% / 13.3%**		**3.8% / 2.3%**		**4.7% / 1.8%**		**1.7% / 1.0%**		**3.3% / 2.4%**		**2.9% / 1.4%**		**1.7% / 1.1%**	

Table 2: **Matched conditional-modulation controls.** MSE / MAE at the shared late-stage point. Gray = DPR; bold/underline = best/second-best per metric; adapters matched within 5% of parameters.

Backbone	Dataset	None	Matched adapters				DPR
			Global SE	Local SE	Local FiLM	Gated	
PatchTST (2023)	ILI (24→24)	3.633 / 1.079	3.352 / 1.049	3.403 / 1.132	3.245 / 1.053	3.398 / 1.131	**3.106 / 1.043**
	COVID19 (36→7)	0.335 / 0.216	0.329 / 0.218	0.329 / 0.219	0.330 / 0.216	**0.324 / 0.215**	0.327 / 0.217
	VIX (96→96)	0.942 / 0.539	0.950 / 0.548	0.948 / 0.551	0.973 / 0.551	0.950 / 0.551	**0.940 / 0.538**
	ETTh1 (96→96)	0.394 / **0.392**	**0.392** / 0.393	0.400 / 0.396	0.396 / 0.395	0.400 / 0.396	**0.392 / 0.392**
Crossformer (2023)	ILI (24→24)	4.736 / 1.480	4.690 / 1.462	4.764 / 1.473	4.719 / 1.461	4.718 / 1.461	**4.593 / 1.428**
	COVID19 (36→7)	0.609 / 0.295	0.638 / 0.309	0.682 / 0.307	0.610 / 0.293	0.689 / 0.314	**0.587 / 0.269**
	VIX (96→96)	1.023 / 0.571	**0.932 / 0.541**	1.004 / 0.566	0.999 / 0.565	1.004 / 0.568	0.970 / 0.561
	ETTh1 (96→96)	0.394 / 0.404	0.400 / 0.412	0.402 / 0.417	0.387 / 0.400	0.401 / 0.416	**0.382 / 0.397**
WPMixer (2025)	ILI (24→24)	3.173 / **1.022**	3.150 / 1.039	3.169 / 1.043	2.976 / 1.075	3.177 / 1.046	**2.796** / 1.046
	COVID19 (36→7)	0.343 / 0.218	0.351 / 0.223	0.321 / **0.214**	0.341 / 0.218	0.321 / **0.214**	**0.318** / 0.218
	VIX (96→96)	0.957 / 0.547	0.991 / 0.567	0.949 / 0.548	1.005 / 0.577	0.949 / 0.548	**0.938 / 0.538**
	ETTh1 (96→96)	0.382 / 0.388	0.385 / 0.388	0.381 / 0.387	**0.379 / 0.386**	0.383 / 0.389	0.381 / 0.387

Association with changing local volatility. Figure 4(b) investigates the relationship between forecast gains and VoV across the 7×11 matrix. For each dataset and horizon, we compute the median gain across the seven primary hosts and average the four horizon medians. We observe a positive rank correlation ($\rho = 0.636$, $n = 11$, two-sided unadjusted $p = 0.035$), indicating that DPR's gains are closely associated with the presence of shifting local volatility.

4.3 Standalone Accuracy and Cost (RQ3)

Standalone accuracy. Table 3 summarizes standalone performance across eight benchmarks under four-horizon averaging (full results in Appendix Table 11). DPRNet achieves the lowest average MSE on COVID19, VIX, and NABCPU, ties for best on ETTh1, Exchange, and Weather, and secures best or tied-best results in six of eight benchmarks, with second-lowest MSE on ILI. Compared with recent linear and mixing models (Appendix Table 15), DPRNet outperforms unadapted backbones on ILI, COVID19, VIX, and Exchange, while OLinear excels on regular ETT benchmarks. A five-model, three-seed comparison (Appendix Table 17) shows DPRNet secures the lowest mean MSE overall.

Table 4: **Standalone efficiency comparison.** Four-horizon MAE; costs at $(96 \to 96)$. Gray = DPRNet; bold/underline denote best/second.

Model	MAE	Params	GMAC	Train s/epoch	Infer ms/batch	Train/Infer GB
OLinear	0.435	4.519M	0.032	4.971 ± 0.398	2.574 ± 0.060	**0.148/0.122**
TimeMixer++	0.456	**0.326M**	0.789	22.212 ± 0.220	47.702 ± 0.008	1.917/0.640
PatchTST	0.432	1.089M	0.069	4.329 ± 0.182	2.234 ± 0.004	0.207/0.133
DPRNet	**0.429**	0.602M	**0.028**	$\mathbf{4.149 \pm 0.135}$	$\mathbf{2.052 \pm 0.005}$	0.172/0.124

Table 3: **Standalone forecasting across eight benchmarks.** Four-horizon MSE / MAE; bold/underline = best/second MSE. Per-horizon results are in Table 11.

Dataset	Informer (2021)	TimesNet (2023)	PatchTST (2023)	iTransformer (2024)	TimeMixer (2024)	TimeFilter (2025)	WPMixer (2025)	DPRNet (Ours)
COVID19	1.917 / 0.632	1.067 / 0.538	0.839 / 0.362	0.902 / 0.379	0.930 / 0.401	0.908 / 0.394	0.839 / 0.363	**0.804 / 0.366**
ILI	7.094 / 1.896	6.021 / 1.309	3.321 / 1.110	3.163 / 1.069	3.182 / 1.148	**2.438 / 0.978**	3.096 / 1.069	2.888 / 1.071
Weather	0.484 / 0.432	0.289 / 0.305	0.251 / 0.270	0.268 / 0.280	0.282 / 0.301	0.246 / 0.267	**0.243 / 0.264**	**0.243 / 0.263**
VIX	1.261 / 0.744	1.207 / 0.715	1.144 / 0.692	1.153 / 0.699	1.141 / 0.694	1.115 / 0.677	1.119 / 0.686	**1.108 / 0.682**
NABCPU	2.689 / 1.035	1.220 / 0.354	1.195 / 0.318	1.200 / 0.338	1.196 / 0.305	1.197 / 0.307	1.202 / 0.334	**1.192 / 0.314**
Exchange	2.782 / 1.235	0.494 / 0.488	0.453 / 0.454	0.446 / 0.452	**0.440 / 0.446**	0.446 / 0.452	0.445 / 0.449	**0.440 / 0.446**
ETTh1	1.324 / 0.816	0.534 / 0.492	0.459 / 0.432	**0.448 / 0.431**	0.458 / 0.429	0.451 / 0.429	0.454 / 0.431	**0.448 / 0.429**
ETTm1	1.480 / 0.855	0.519 / 0.474	0.396 / 0.387	0.399 / 0.390	0.393 / 0.385	0.395 / 0.387	**0.390 / 0.384**	0.396 / 0.385

Accuracy and measured cost. Table 4 reports four-horizon average MAE alongside benchmarked hardware costs on an isolated GPU for ETTh1 (efficiency tradeoffs on COVID19 in Appendix Figure 9). Among all profiled models, DPRNet attains the lowest MAE, GMAC, training latency, and inference time. Operating with only 0.602M parameters and 0.028 GMAC per sample, it trains in 4.15 ± 0.14 s per epoch and executes inference in 2.05 ± 0.01 ms per batch.

4.4 Design Evidence Beyond Static Capacity (RQ4)

Static capacity scaling. We examine whether DPR's performance gains stem from added parameter capacity. We test three hosts (PatchTST, TimesNet, and TimeFilter) across ETTh1, ILI, and Exchange under six configurations: base, 2× depth, 2× width, 2× both, +DPR, and a parameter-matched static baseline (54 total settings). Figure 5 shows that uniform static scaling often leads to over-parameterization and accuracy drops. DPR improves or preserves accuracy across nine host–dataset settings (Appendix Table 13), outperforming the parameter-matched static baseline in all nine settings.

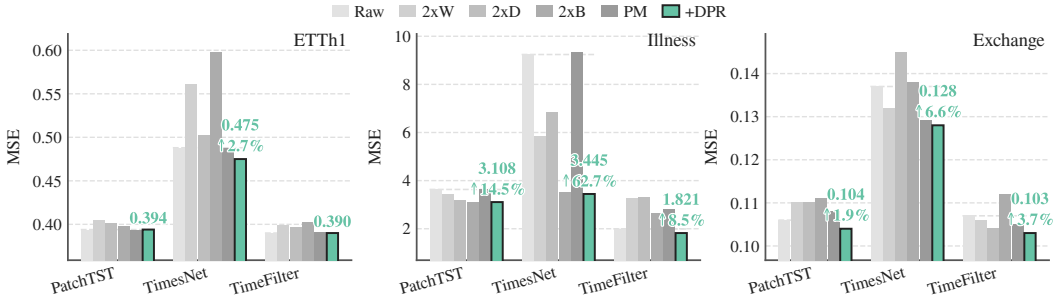


Figure 5: **Static parameter scaling versus DPR.** MSE for width, depth, combined, parameter-matched, and DPR variants across three hosts. Uniform scaling yields erratic effects across datasets, whereas token-level recalibration provides targeted adaptation. Full results are in Appendix Table 13.

Component ablations and basis diversity. Table 5 reports 18 Crossformer ablations on ETTh1, ILI, and Exchange. Full DPR attains the lowest MSE in all settings, improving 6.0% on average over the unadapted backbone. Randomizing gain initialization or disabling multi-scale perception hurts performance, most on volatile ILI (-8.3%). A PatchTST geometry study confirms orthogonal regularization: $\mathcal{L}_{\text{orth}}$ suppresses the mean off-diagonal basis cosine from 0.943 to 0.079 on ILI and 0.898 to 0.003 on ETTh1 while maintaining accuracy, eliminating basis redundancy. Across three seeds (Appendix Table 16), DPR lowers mean MSE on all twelve volatile-benchmark settings for Informer, PatchTST, WPMixer, and TimeFilter; ETTh1 changes stay small for modern hosts.

Controlled routing comparison. Table 6 compares DPR against sparse Top-k expert routing within the DPRNet architecture across ten settings on ILI ($24 \rightarrow 24$) and ETTh1 ($96 \rightarrow 96$). While sparse MoE scales to 818K–1.1M parameters,

Table 6: **DPR vs. MoE routing.** MoE = sparse Top-k ($k \in \{1, 2, 4\}$). Gray = DPR; bold/underline = best/second; Time = train / inf. (s).

Variant	Params	ILI MSE / MAE	ILI Time	ETTh1 MSE / MAE	ETTh1 Time
w/o DPR	287K–563K	3.347 / 1.091	0.81 / 0.71	0.405 / 0.398	15.84 / 5.07
MoE Top-1	818K–1.1M	3.715 / 1.156	1.06 / 0.77	0.411 / 0.403	62.62 / 18.68
MoE Top-2	818K–1.1M	3.755 / 1.097	1.18 / 0.90	0.419 / 0.405	89.12 / 16.05
MoE Top-4	818K–1.1M	4.146 / 1.122	1.02 / 0.77	0.417 / 0.404	45.89 / 13.43
DPRNet	325K–602K	**2.970 / 1.071**	0.85 / 0.73	**0.383 / 0.389**	29.90 / 8.60

Table 5: **Component ablations and basis geometry.** Single-run Crossformer ablations vs. no adapter; gray = DPR, bold/underline = best/second-best, AVG = mean Impv. over the three datasets.

Variant	Configuration				ETTh1 (96→96)		ILI (24→24)		Exchange (96→96)		AVG
	M-Scale	Ortho	Init	Route	MSE / MAE	Impv.	MSE / MAE	Impv.	MSE / MAE	Impv.	
No adapter	–	–	–	–	0.394 / 0.404	–	4.736 / 1.480	–	0.269 / 0.349	–	–
w/o multi-scale	× ($k = 1$)	✓	✓	Soft	0.397 / 0.406	↓0.8%	5.129 / 1.546	↓8.3%	0.266 / 0.346	↑1.1%	-2.7%
w/o orthogonal reg.	✓	×	✓	Soft	0.388 / 0.401	↑1.5%	4.661 / 1.448	↑1.6%	0.242 / 0.338	↑10.0%	↑4.4%
random gain init.	✓	✓	×	Soft	0.401 / 0.410	↓1.8%	5.062 / 1.525	↓6.9%	0.279 / 0.367	↓3.7%	-4.1%
Discrete routing	✓	✓	✓	Hard	0.385 / 0.399	↑2.3%	4.721 / 1.458	↑0.3%	0.238 / 0.335	↑11.5%	↑4.7%
DPR	✓	✓	✓	Soft	**0.382 / 0.397**	↑3.0%	**4.593 / 1.428**	↑3.0%	**0.237 / 0.335**	↑11.9%	↑6.0%

PatchTST mean off-diagonal basis cosine (w/o → w/ orthogonal reg.): ETTh1 $0.898 \rightarrow 0.003$; ILI $0.943 \rightarrow 0.079$.

DPR attains lower MSE with 325K–602K parameters. Avoiding discrete dispatch and gradient bottlenecks, DPRNet offers a better accuracy-parameter Pareto frontier.

Sensitivity. Appendix Figure 8 examines 18 hyperparameter configurations spanning 72 horizon evaluations. The default $K = 8$ consistently provides the optimal trade-off between expressiveness and basis compactness; MSE is remarkably stable across varying λ_{orth} values, and preferred perception kernels naturally align with host tokenization granularity.

4.5 Interpretability: Pattern-Conditioned Response

Qualitative response behavior. Figure 6 illustrates the qualitative impact of DPR recalibration. Panels (a–b) demonstrate visibly tighter forecast tracking and reduced prediction error during sudden trend shifts and volatility spikes. Panel (c) overlays the raw time series with local volatility, while panel (d) reveals that routing probabilities $\boldsymbol{\pi}_{b,\ell}$ dynamically diverge from the uniform prior $1/K = 0.125$ synchronously with regime transitions. These qualitative observations are consistent with context-dependent routing around local regime transitions.

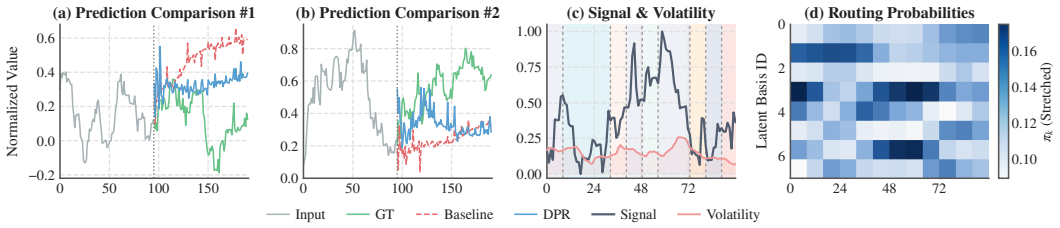


Figure 6: **Qualitative case study. (a–b)** Ground truth and forecasts from the base host and the same host with DPR. **(c)** Signal and local volatility, and **(d)** routing probabilities, over the highlighted interval. The uniform-routing reference is $1/K = 0.125$.

5 Conclusion

We introduced **Dynamic Pattern Recalibration** (DPR), a lightweight mechanism that parameterizes local adaptation as a structured mixture over a shared response basis, decoupling shared representation learning from token-level feature-gain control. Central to DPR is its three-stage "Perceive–Route–Modulate" pipeline, which routes temporal context over a compact, orthogonally regularized response basis and recalibrates hidden representations via a residual Hadamard product. Across ten backbones and 95 evaluations, DPR matches or improves MSE in 87 pairs (91.6%), with gains correlating with local volatility. As a standalone instantiation, DPRNet achieves strong accuracy across eight benchmarks with low measured computational cost. These results support continuous response-basis modulation as a parameter-efficient paradigm for non-stationary forecasting tasks.

Limitations and Future Work. DPR provides targeted gains in dynamic environments with shifting regimes, and diminishes on near-stationary, low-heterogeneity series where local pattern recalibration is rarely warranted. Our plug-in study focuses on late-stage recalibration; future work includes multi-layer integration, exogenous covariates, and scaling localized modulation to foundation models.

AI USE STATEMENT

Generative AI tools assisted with manuscript review, language editing, bibliography verification, and LaTeX formatting. The authors reviewed the AI-assisted output and take responsibility for the final methods, results, and claims.

REPRODUCIBILITY STATEMENT

We release the code, data preparation and split-generation scripts, and configurations at `https://anonymous.4open.science/r/DPRNet`. Models are implemented in PyTorch and trained on four NVIDIA A800 (80GB) GPUs with the Adam optimizer and early stopping (patience 10). DPRNet uses $d = 256$; in plug-in experiments, DPR inherits each host backbone's native hidden width. The default mechanism uses a shared response basis of size $K = 8$, routing logit scale initialized to 2.0, and gain $\gamma = 0.1$; default hyperparameters are in Appendix B (Table 9). The plug-in study preserves each host's learning rate, dropout, batch size, and layer count, and applies one late-stage insertion rule. Multi-seed stability results use seeds 2023, 2024, and 2025, and adapter reruns share the same preprocessing and BasicTS evaluation code across all datasets.

ETHICS STATEMENT

This work develops a general-purpose numerical forecasting adapter and evaluates it on publicly available time-series benchmarks; it involves no human participants and no personally identifiable information. In high-stakes domains such as energy, healthcare, and finance, forecasts should be paired with uncertainty estimation and expert oversight. We will comply with the licenses of all source datasets, and all authors have read and will follow the ICLR Code of Ethics.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Bernie Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. URL `https://mlanthology.org/tmlr/2024/ansari2024tmlr-chronos/`.

Sawyer Birnbaum, Volodymyr Kuleshov, Zayd Enam, Pang Wei Koh, and Stefano Ermon. Temporal FiLM: Capturing long-range sequence dependencies with feature-wise modulations. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL `https://proceedings.neurips.cc/paper/2019/hash/2afc4dfb14e55c6face649a1d0c1025b-Abstract.html`.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10148–10167. PMLR, 2024.

Chirag Deb, Fan Zhang, Junjing Yang, Siew Eang Lee, and Kwok Wei Shah. A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews*, 74:902–924, 2017.

Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 20(5):533–534, 2020. doi: 10.1016/S1473-3099(20)30120-1. URL `https://pubmed.ncbi.nlm.nih.gov/32087114/`.

William Fedus, Barret Zoph, and Noam Shazeer. Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

James D. Hamilton. Analysis of time series subject to changes in regime. *Journal of Econometrics*, 45(1-2):39–70, 1990.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

Yifan Hu, Guibin Zhang, Peiyuan Liu, Disen Lan, Naiqi Li, Dawei Cheng, Tao Dai, Shu-Tao Xia, and Shirui Pan. TimeFilter: Patch-specific spatial-temporal graph filtration for time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 24893–24911. PMLR, 2025. URL https://proceedings.mlr.press/v267/hu25ac.html.

Qihe Huang, Zhengyang Zhou, Kuo Yang, Zhongchao Yi, Xu Wang, and Yang Wang. TimeBase: The power of minimalism in efficient long-term time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 26227–26246. PMLR, 2025. URL https://proceedings.mlr.press/v267/huang25az.html.

Sheikh Mohammad Idrees, M Afshar Alam, and Parul Agarwal. A prediction approach for stock market volatility based on time series data. *IEEE Access*, 7:17287–17298, 2019.

T. Inouye, K. Shinosaki, H. Sakamoto, S. Toi, S. Ukai, A. Iyama, Y. Katsuda, and M. Hirano. Quantification of EEG irregularity by use of the entropy of the power spectrum. *Electroencephalography and Clinical Neurophysiology*, 79(3):204–210, 1991.

Zahra Karevan and Johan AK Suykens. Transductive LSTM for time-series prediction: An application to weather forecasting. *Neural Networks*, 125:1–9, 2020.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations (ICLR)*, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jiayi Liu, Donghua Yang, Kaiqi Zhang, Hong Gao, and Jianzhong Li. Anomaly and change point detection for time series with concept drift. *World Wide Web*, 26(5):3229–3252, 2023a.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 9881–9893, 2022.

Yong Liu, Chenyu Li, Jianmin Wang, and Mingsheng Long. Koopa: Learning non-stationary time series dynamics with koopman predictors. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023b. doi: 10.52202/075280-0538.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2024. Spotlight.

Yongzheng Liu, Siru Zhong, Gefeng Luo, Weilin Ruan, and Yuxuan Liang. Towards multi-scenario forecasting of building electricity loads with multimodal data. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2188–2196, 2025. doi: 10.1145/3746027.3755775. URL https://doi.org/10.1145/3746027.3755775.

Ziqi Liu. FreqMoE: Enhancing time series forecasting through frequency decomposition mixture of experts. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258, pp. 3430–3438. PMLR, 2025.

Xiaowen Ma, Shuning Ge, Fan Yang, Xiangyu Li, Yun Chen, Mengting Ma, Wei Zhang, and Zhipeng Liu. Timeexpert: Boosting long time series forecasting with temporal mix of experts. *arXiv preprint arXiv:2509.23145*, 2025.

Md Mahmuddun Nabi Murad, Mehmet Aktukmak, and Yasin Yilmaz. WPMixer: Efficient multi-resolution mixing for long-term time series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18):19581–19588, April 2025. ISSN 2159-5399. doi: 10.1609/aaai.v39i18.34156. URL http://dx.doi.org/10.1609/aaai.v39i18.34156.

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

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. In *International Conference on Learning Representations (ICLR)*, 2024.

Weilin Ruan, Wenzhuo Wang, Siru Zhong, Wei Chen, Li Liu, and Yuxuan Liang. Cross space and time: A spatio-temporal unitized model for traffic flow forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 26(11):21296–21308, 2025. doi: 10.1109/TITS.2025.3601630. URL https://doi.org/10.1109/TITS.2025.3601630.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 5998–6008, 2017.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/a7ac8a21e5a27e7ab31a5f42a0117bdb-Abstract-Conference.html.

Shiyu Wang, Jiawei Li, Xiaoming Shi, Zhou Ye, Baichuan Mo, Wenze Lin, Shengtong Ju, Zhixuan Chu, and Ming Jin. TimeMixer++: A general time series pattern machine for universal predictive analysis. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2b187165e28fdfdc0ffb34d1bfff2b0c-Abstract-Conference.html.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53140–53164. PMLR, 2024. URL https://proceedings.mlr.press/v235/woo24a.html.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, 2018.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 22419–22430, 2021.

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

Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. Condconv: Conditionally parameterized convolutions for efficient inference. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Wenzhen Yue, Yong Liu, Hao Wang, Haoxuan Li, Xianghua Ying, Ruohao Guo, Bowei Xing, and Ji Shi. OLinear: A linear model for time series forecasting in orthogonally transformed domain. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 67394–67453, 2025. doi: 10.52202/085713-2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/571c7e164fb1ffbcf2f84a63784451ec-Abstract-Conference.html.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):11121–11128, jun 2023. ISSN 2159-5399. doi: 10.1609/aaai.v37i9.26317. URL http://dx.doi.org/10.1609/aaai.v37i9.26317.

Huaiwu Zhang, Yutong Xia, Siru Zhong, Kun Wang, Zekun Tong, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. Predicting carpark availability in singapore with cross-domain data: A new dataset and a data-driven approach. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 7554–7562, 2024. doi: 10.24963/ijcai.2024/836. URL https://doi.org/10.24963/ijcai.2024/836.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2023.

Jianhu Zheng and Mingfang Huang. Traffic flow forecast through time series analysis based on deep learning. *IEEE Access*, 8:82562–82570, 2020.

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

Xingchen Zou, Weilin Ruan, Siru Zhong, Yuehong Hu, and Yuxuan Liang. Fine-grained urban heat island effect forecasting: A context-aware thermodynamic modeling framework. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 4226–4237, 2025. doi: 10.1145/3711896.3736962. URL https://doi.org/10.1145/3711896.3736962.

Appendix

Table of Contents

A Dataset Statistics and Local Volatility Diagnostics 14
A.1 Dataset Descriptions 14
A.2 Quantifying Local Volatility Variation 15
A.3 Observations and Discussion 16

B Implementation Details and Hyperparameters 17
B.1 DPR Default Configuration 17
B.2 DPR Vectorized Pseudocode 17

C Integration Details for Diverse Architectures. 18
C.1 Patch/Channel-Aware Topologies 18
C.2 WPMixer Branch-Level Integration 19
C.3 Kernel Configuration 19

D Inductive Bias Analysis of DPR 19
D.1 Shared Transformations and Explicit Gain Control 19
D.2 Design Principles: Continuity and Basis Diversity 19
D.3 Mechanism-Level Jacobian and Conditional Stability 20

E Hyperparameter Sensitivity 20

F Complete Experimental Results 20
F.1 Full Main Results 20
F.2 Full Adapter Results 20
F.3 Parameter Scaling versus DPR 20

G Controls, Robustness, and Evaluation Details 21
G.1 Control Protocol Details 22
G.2 Matched Conditional-Modulation Controls 23
G.3 Extended Architecture Evaluation 23
G.4 Multi-seed Stability and Efficiency 23
G.5 Additional Standalone Benchmarks 24

A Dataset Statistics and Local Volatility Diagnostics

The evaluation spans 17 unique datasets across energy, finance, climatology, epidemiology, healthcare, cloud operations, solar physics, and transportation. They range from monthly to sub-hourly resolution, contain 966 to 57,600 timesteps, and cover univariate to highly multivariate series. Table 7 gives detailed statistics for the eleven datasets represented in the complete per-horizon backbone matrices; configurations for ECL, Traffic, and PEMS03/04/07/08 are specified in Appendix G.

A.1 Dataset Descriptions

- **ETT (Electricity Transformer Temperature)**: Oil temperature and six loads from two Chinese counties (2016/7/1–2018/6/26). Four subsets: *ETTh1/h2* (hourly, 14,400 steps) and *ETTm1/m2* (15-min, 57,600 steps). Variables: HUFL, HULL, MUFL, MULL, LUFL, LULL, OT.
- **Weather**: 21 meteorological indicators (temperature, humidity, pressure, wind, precipitation, etc.) sampled every 10 minutes in 2020 (52,696 steps). Variables range from slowly drifting (temperature) to highly volatile (rain, wind gust).
- **Illness (ILI)**: Weekly ILI patient ratios from the U.S. CDC (2002–2020, 966 steps). Seven variables: weighted/unweighted ILI ratios, age-specific counts, and total provider numbers. Small sample size and long-term dependencies with sharp seasonal outbreaks.
- **ExchangeRate**: Daily exchange rates for eight countries (Australia, Britain, Canada, Switzerland, China, Japan, New Zealand, Singapore) from 1990 to 2016 (7,588 steps). Highly volatile, chaotic fluctuations with minimal periodicity challenge models that rely on regular cyclic patterns.

- **COVID19**: Confirmed COVID-19 cases from JHU CSSE (Dong et al., 2020) (2020/1/22–2023/3/9, 1,143 steps). The benchmark contains daily new-case counts for the top 7 countries plus a global aggregate (8 channels total). Asymmetric pattern shifts: long plateaus punctuated by exponential variant waves.
- **VIX (Volatility Index)**: Measures 30-day S&P 500 implied volatility (1990/1/2–2026/4/15, 9,165 trading days, 1 channel). Canonical volatility clustering: narrow bands in calm markets, vertical spikes at crises (2008, 2020, 2022), then decay.
- **NABCPU**: AWS CloudWatch traces from the Numenta Anomaly Benchmark (2014/2/14–2014/2/28, 4,031 steps at 5-minute resolution, 3 channels). Sharp step-change bursts (scheduled jobs, memory leaks, traffic spikes) sit on top of irregular idle baselines, testing how well models recover from abrupt transitions.
- **Sunspots**: Monthly mean total sunspot number from WDC-SILSO, Royal Observatory of Belgium (1749/1–2026/3, 3,327 months, 1 channel). The $\sim$11-year cycle alternates between sharp spikes and smooth plateaus, producing very high VoV.

Train, validation, and test splits follow Table 7: 6:2:2 for the ETT datasets and 7:1:2 for all others.

Table 7: Statistics of the eleven datasets represented in the full per-horizon backbone results. For ETT, we adopt the truncated 20-month protocol introduced by Informer (Zhou et al., 2021); all other datasets use the full publicly available horizons. The datasets are grouped by their respective domains.

Dataset	**Vars** (C)	**Timesteps** (T)	**Split**	**Frequency**	**Look-back** (L)	**Pred. Len.** (H)	**Raw coverage**	**Domain**
ETTh1	7	14,400	6:2:2	Hourly	96	{96, 192, 336, 720}	2016/7–2018/6	Energy
ETTh2	7	14,400	6:2:2	Hourly	96	{96, 192, 336, 720}	2016/7–2018/6	
ETTm1	7	57,600	6:2:2	15 Min	96	{96, 192, 336, 720}	2016/7–2018/6	
ETTm2	7	57,600	6:2:2	15 Min	96	{96, 192, 336, 720}	2016/7–2018/6	
Weather	21	52,696	7:1:2	10 Min	96	{96, 192, 336, 720}	2020/1–2021/1	Climatology
Illness (ILI)	7	966	7:1:2	Weekly	24	{24, 36, 48, 60}	2002–2020	Healthcare
COVID19	8	1,143	7:1:2	Daily	36	{7, 14, 28, 60}	2020/1–2023/3	Epidemiology
ExchangeRate	8	7,588	7:1:2	Daily	96	{96, 192, 336, 720}	1990–2016	Finance
VIX	1	9,165	7:1:2	Daily	96	{96, 192, 336, 720}	1990–2026	
NABCPU	3	4,031	7:1:2	5 Min	96	{24, 48, 96, 192}	2014/2	Cloud Ops
Sunspots	1	3,327	7:1:2	Monthly	36	{12, 24, 48, 96}	1749–2026	Solar Phys.

A.2 Quantifying Local Volatility Variation

The qualitative characterization in Table 7 is complemented by three quantitative diagnostics that directly probe the *local* pattern structure motivating our method. All diagnostics are computed on the full series (training, validation, and test splits concatenated) with each channel z-normalized, so they are scale-invariant. Spectral entropy and VoV are averaged over all C channels, whereas the ADF test is restricted to the first $\min(C, 20)$ channels and at most $2{,}000$ decimated points to bound runtime.

- **ADF p-value.** We test the regression $\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p} \delta_i \Delta y_{t-i} + \varepsilon_t$ ($H_0 : \gamma = 0$, unit root). Larger p-values indicate weaker evidence against a unit root.
- **Spectral entropy** $H_s \in [0, 1]$**.** $H_s = -\frac{1}{\log N_f} \sum_f P(f) \log P(f)$ where $P(f)$ is the normalized power spectrum (DC excluded). Values near 0 signify concentrated, quasi-periodic dynamics; values near 1 indicate broadband stochastic dynamics in which no single frequency dominates.
- **Volatility-of-Volatility (VoV).** VoV $= \frac{\mathrm{std}(\sigma_t)}{\mathrm{mean}(\sigma_t)}$, where $\sigma_t = \mathrm{std}(x_{t-w+1:t})$ is a trailing rolling standard deviation (population convention, $\mathrm{ddof} = 0$); channels with $\mathrm{mean}(\sigma_t) = 0$ are discarded. The window is $w = \mathrm{clip}(\mathrm{round}(1440/\Delta),\ 12,\ 256)$ for a sampling interval of Δ minutes. For sub-daily series, w targets approximately one day; for daily and coarser series, we use the minimum window of 12 observations for estimator stability. Concretely, $w = 12$ for the daily, weekly, and monthly series (VIX, ExchangeRate, COVID19, Illness, Sunspots), $w = 24$ for hourly (ETTh1/h2), $w = 96$ for 15-minute (ETTm1/m2), $w = 144$ for 10-minute (Weather), and $w = 256$ for 5-minute (NABCPU). VoV describes variation in local sample dispersion. Finite-window variation can occur even under a stationary process, so VoV serves as a descriptive measure of local volatility heterogeneity.

The score in Table 8 is computed as $rank(H_s) + rank(\text{VoV})$, where higher values indicate more local frequency complexity and volatility variation. It summarizes these diagnostics within the eleven-dataset reference set.

A.3 Observations and Discussion

The diagnostics summarize complementary aspects of the datasets:

1. **COVID19** and **Illness** have VoV values of 1.46 and 1.00, respectively, indicating substantial variation in local sample dispersion alongside their temporal structure.
2. **Weather** has the highest channel-averaged VoV (1.68), reflecting temporal variation in within-channel local dispersion across its meteorological variables.
3. **VIX** (Score 14) shows volatility clustering: narrow low-volatility bands during calm periods with vertical spikes at crises (2008, 2020, 2022).
4. **NABCPU** (Score 12) has the highest spectral entropy (0.78), indicating nearly broadband dynamics from overlapping periodicities: daily cycles, weekly patterns, and intermittent computational bursts. **Sunspots** (Score 13) shows the $\sim$11-year solar cycle with strong amplitude variation.
5. **ExchangeRate** (Score 10) combines a high mean ADF p-value (0.55), concentrated low-frequency power ($H_s = 0.21$), and high VoV (1.32). These describe unit-root evidence, spectral concentration, and local dispersion, respectively. VoV captures local dispersion, which does not by itself imply a change in the predictable conditional mean.
6. **ETTh1, ETTh2, ETTm1, ETTm2** have comparatively regular periodic structure (eleven-benchmark scores 6–9). They serve as regular-periodic controls for measuring DPR's effect when unadapted backbones are already strong.

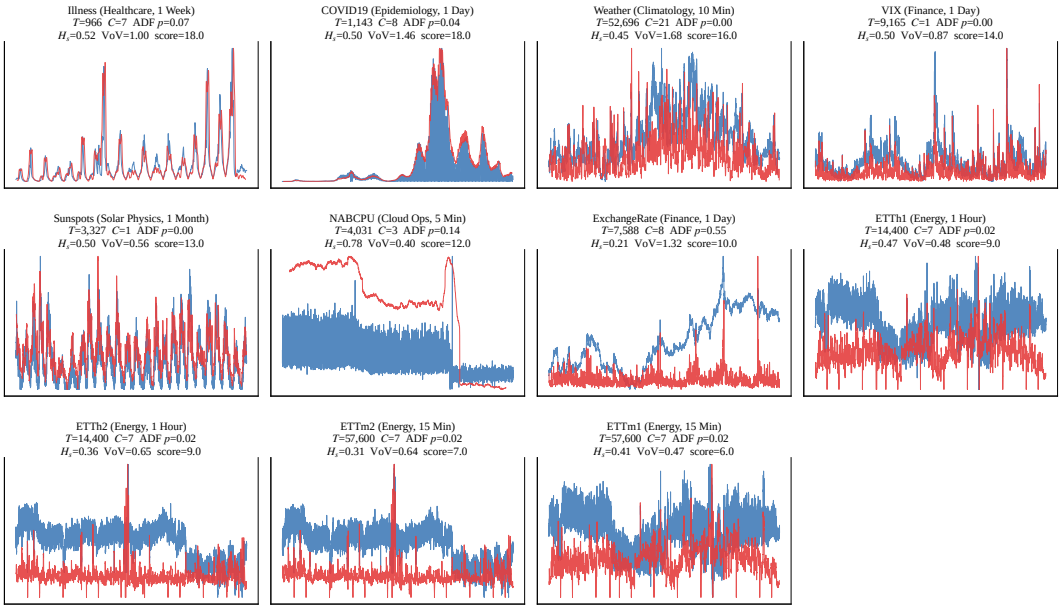


Figure 7: Local volatility variation across the eleven datasets represented in the full per-horizon backbone results. For each dataset, a representative z-normalised channel (cyan) is overlaid with its rolling standard deviation over its dataset-specific rolling window w defined in Appendix A.2 (orange, right axis). Flat and varying orange curves indicate relatively homogeneous and heterogeneous local sample dispersion, respectively. Panel titles report length T, channels C, ADF p-value, spectral entropy H_s, and volatility-of-volatility VoV.

Table 8 and Figure 7 illustrate why a single global diagnostic is insufficient. The exploratory gain correlation in Appendix G relates DPR's measured benefit to local volatility variation.

Table 8: **Dataset quantitative diagnostics across eleven core benchmarks.** The composite **Score** is computed as *rank*(H_s) + *rank*(VoV). Larger values indicate higher frequency complexity and local volatility variation. Bold and underlined entries identify the first- and second-ranked values per metric within this dataset set.

Metric	Illness	COVID19	Weather	VIX	NABCPU	Sunspots	Exchange	ETTh1	ETTh2	ETTm2	ETTm1
ADF p-val ↑	0.0722	0.0438	0.0033	0.0000	<u>0.1374</u>	0.0002	**0.5499**	0.0165	0.0249	0.0249	0.0165
Spectral Entropy (H_s) ↑	<u>0.5176</u>	0.5016	0.4514	0.4965	**0.7754**	0.5030	0.2067	0.4686	0.3586	0.3126	0.4114
Volatility-of-Volatility (VoV) ↑	0.9995	<u>1.4648</u>	**1.6813**	0.8722	0.3955	0.5573	1.3221	0.4786	0.6536	0.6411	0.4743
rank(H_s)	<u>10</u>	8	5	7	**11**	9	1	6	3	2	4
rank(VoV)	8	<u>10</u>	**11**	7	1	4	9	3	6	5	2
Composite Score	**18**	**18**	<u>16</u>	14	12	13	10	9	9	7	6

B Implementation Details and Hyperparameters

B.1 DPR Default Configuration

All models are implemented in PyTorch and trained on four NVIDIA A800 (80GB) GPUs. We use the Adam optimizer (Kingma & Ba, 2014) across all experiments. Across the 77 backbone–dataset pairs with complete per-horizon adapter results, injecting DPR preserves the host backbone's original learning rate, dropout, batch size, and layer count. Early stopping is applied with patience 10 epochs based on validation loss. Table 9 reports the complete default hyperparameters for our minimalist DPRNet backbone. Unless stated otherwise, adapter experiments adopt only the DPR-mechanism hyperparameters (bottom half of the table) while preserving the host backbone's native configuration.

Table 9: Default hyperparameters for DPRNet and the DPR mechanism.

Hyperparameter	Default
Backbone	
Patch length P / stride s	16 / 8
Hidden dimension d	256
MLP: layers / expansion / dropout	2 / 2.0× / 0.1
DPR Mechanism	
Number of patterns K	8
Response basis init $\mathbf{M}$	$\mathcal{N}(0, 0.02^2)$
Projection d_c	$\max(16, \lfloor d/4 \rfloor)$
Orthogonal penalty λ_{orth}	10^{-4}
Logit scale τ	2.0 (learned)
Gain init γ	0.1
Multi-scale kernels	$k_1 = 3, k_2 = 7$
Training	
Optimizer / learning rate	Adam / 10^{-3}
Batch size	32
Early-stopping patience	10 epochs

See Table 7 in Appendix A for look-back windows and horizons.

Near-identity initialization. The response basis rows are sampled from $\mathcal{N}(0, 0.02^2)$ and the residual gain is set to $\gamma = 0.1$. The small initial basis magnitude ensures $\|\gamma\, \mathbf{m}_{b,\ell}\|_\infty \ll 1$ at the onset of training, so the DPR operation begins as a near-identity map that leaves the host backbone's learned representation essentially unperturbed. The "random gain init." row in Table 5 instead samples $\gamma \sim \mathcal{N}(0, 0.01^2)$ rather than fixing $\gamma = 0.1$, while leaving the basis initialization unchanged.

B.2 DPR Vectorized Pseudocode

Algorithm 1 provides the vectorized forward pass of DPR, executing dense parallel matrix multiplications without token-level dispatch or loop overhead.

Algorithm 1 Dynamic Pattern Recalibration (DPR) – Vectorized Forward Pass

Require: Hidden states $\mathbf{H} \in \mathbb{R}^{B\times L\times d}$; response basis $\mathbf{M} \in \mathbb{R}^{K\times d}$; routing prototypes $\mathbf{E} \in \mathbb{R}^{K\times d_c}$; context dimension d_{ctx} ($2d$ for multi-scale depthwise convolutions, d for pointwise); projection parameters $\mathbf{W}_1 \in \mathbb{R}^{d_c\times d_{\text{ctx}}}$, $\mathbf{b}_1 \in \mathbb{R}^{d_c}$; logit scale τ; residual gain γ; kernel sizes k_1, k_2.
Ensure: Recalibrated hidden representations $\tilde{\mathbf{H}} \in \mathbb{R}^{B\times L\times d}$.

% Stage 1: Perceive multi-scale local context
$\mathbf{H}_{\text{cf}} \leftarrow \text{Permute}(\mathbf{H}, (0, 2, 1))$ ▷ To channel-first layout: $\mathbb{R}^{B\times d\times L}$
$\mathbf{Z}_{\text{cf}} \leftarrow [\text{DWConv}_{k_1}(\mathbf{H}_{\text{cf}})\,;\, \text{DWConv}_{k_2}(\mathbf{H}_{\text{cf}})]$ ▷ Concatenate along feature axis: $\mathbb{R}^{B\times d_{\text{ctx}}\times L}$
$\mathbf{Z} \leftarrow \text{Permute}(\mathbf{Z}_{\text{cf}}, (0, 2, 1))$ ▷ Restore sequence layout: $\mathbb{R}^{B\times L\times d_{\text{ctx}}}$
% Stage 2: Route tokens over routing prototypes (Parallel batch projection)
$\mathbf{C} \leftarrow \text{GELU}(\mathbf{Z}\mathbf{W}_1^\top + \mathbf{b}_1^\top)$ ▷ Project context query: $\mathbb{R}^{B\times L\times d_c}$
$\hat{\mathbf{C}} \leftarrow \mathbf{C}/\|\mathbf{C}\|_2, \quad \hat{\mathbf{E}} \leftarrow \mathbf{E}/\|\mathbf{E}\|_2$ ▷ L_2-normalization along feature dimensions
$\mathbf{S} \leftarrow \tau \cdot (\hat{\mathbf{C}}\,\hat{\mathbf{E}}^\top)$ ▷ Scaled cosine logits: $\mathbb{R}^{B\times L\times K}$
$\mathbf{\Pi} \leftarrow \text{softmax}(\mathbf{S}, \text{dim} = -1)$ ▷ Continuous routing distributions: $\mathbb{R}^{B\times L\times K}$
% Stage 3: Modulate hidden states via residual Hadamard product
$\mathbf{M}_{\text{dyn}} \leftarrow \mathbf{\Pi}\,\mathbf{M}$ ▷ Soft mixture of basis patterns: $\mathbb{R}^{B\times L\times d}$
$\tilde{\mathbf{H}} \leftarrow \mathbf{H} \odot (\mathbf{1} + \gamma\,\mathbf{M}_{\text{dyn}})$ ▷ Element-wise recalibration: $\mathbb{R}^{B\times L\times d}$
return $\tilde{\mathbf{H}}$

C Integration Details for Diverse Architectures

DPR is backbone-agnostic. We use one structural insertion rule across hosts: apply a shared DPR adapter to the late-stage hidden representation immediately before the final prediction head. This location exposes contextualized features while preserving the host's feature-extractor architecture and keeps the comparison independent of backbone-specific layer tuning. For backbones without standard token-form hidden states, we apply lightweight reshaping before and after DPR. Table 10 gives implementation-aligned mappings for representative hidden-state topologies; OLinear, TimeMixer++, and TimeBase follow the same late-stage rule.

Table 10: Implementation-aligned DPR integration across representative backbone topologies. C denotes variates, L_p denotes the number of patch tokens, and H denotes prediction horizon.

Backbone	Hidden Shape	Insert Position	Reshape?	Kernel
Informer	$B \times L \times d$	After decoder, before head	No	$(3, 7)$
Crossformer	$(B \cdot C) \times P_{\text{out}} \times d$	Decoder: final scale branch	No	$(3, 7)$
TimesNet	$B \times (L{+}H) \times d$	After backbone, before head	No	$(3, 7)$
PatchTST	$B \times C \times L_p \times d$	After encoder, before head	Yes	1
TimeMixer	$B \times H \times d$ (per scale)	Per-scale: after temporal projection	No	$(3, 7)$
TimeFilter	$B \times C \times L_p \times d$	After graph backbone, before head	Yes	1
WPMixer	$B \times C \times L_p \times d$	Per-resolution: after mixer, before head	Yes	1

The local neighborhood follows the host token axis: it can represent historical patches, forecast-position latent states, or a combined history–forecast representation. Thus, local context refers to neighboring hidden states rather than uniformly to observed historical timesteps. For sequential token topologies (Informer, TimesNet, and the per-scale path of TimeMixer), DPR consumes hidden states in $\mathbb{R}^{B\times L\times d}$ (or forecast-length variants) and applies depthwise convolutions without reshaping. Each backbone uses one shared DPR adapter at late-stage representations, with reuse across scale or branch paths where needed. The evaluation focuses on this late-stage configuration; earlier-layer and multi-layer placements remain open design choices.

C.1 Patch/Channel-Aware Topologies

For PatchTST and TimeFilter, hidden states are organized as $\mathbb{R}^{B\times C\times L_p\times d}$. We reshape to

$$\mathbf{H}_{\text{in}} \in \mathbb{R}^{B\times C\times L_p\times d} \xrightarrow{\text{reshape}} \hat{\mathbf{H}}_{\text{in}} \in \mathbb{R}^{(B\cdot C)\times L_p\times d}, \tag{8}$$

apply DPR independently on each variable-token sequence, then restore the original layout.

C.2 WPMixer Branch-Level Integration

In WPMixer, DPR is applied inside each resolution branch. The branch feature map is reshaped from $\mathbb{R}^{B\times C\times L_p\times d}$ to $\mathbb{R}^{B\times(C\cdot L_p)\times d}$, modulated, and reshaped back before the branch prediction head.

C.3 Kernel Configuration

The DPR adapter defaults to multi-scale kernels $(3, 7)$. In the baseline scripts, we set $k = 1$ for patch-like backbones (PatchTST, TimeFilter, WPMixer) because each token already summarizes a local span, while retaining $(3, 7)$ for sequential token topologies. We fix this choice before evaluation and hold it constant across test windows.

D Inductive Bias Analysis of DPR

We analyze the explicit token-level gain parameterization introduced by DPR.

D.1 Shared Transformations and Explicit Gain Control

For a linear transformation $\mathcal{F}_\Theta(\mathbf{h}) = \mathbf{W}\mathbf{h}$, the Jacobian is the constant matrix $\mathbf{W}$. This illustrates the linear case. A shared nonlinear FFN, $\mathcal{F}_\Theta(\mathbf{h}) = \mathbf{W}_2\phi(\mathbf{W}_1\mathbf{h} + \mathbf{b}_1) + \mathbf{b}_2$, already has an input-dependent Jacobian:

$$\mathbf{J}_{\mathcal{F}}(\mathbf{h}) = \mathbf{W}_2\,\mathrm{diag}\big(\phi'(\mathbf{W}_1\mathbf{h} + \mathbf{b}_1)\big)\mathbf{W}_1.$$

Parameter sharing is therefore compatible with input-dependent responses. DPR adds a structured, explicit path for local feature-gain adaptation.

DPR uses $\tilde{\mathbf{h}} = \mathbf{h} \odot (\mathbf{1} + \gamma\,\mathbf{m})$, where $\mathbf{m}$ is a context-weighted combination of learned basis vectors. Holding the modulation vector fixed gives the conditional derivative:

$$\left.\frac{\partial\tilde{\mathbf{h}}}{\partial\mathbf{h}}\right|_{\mathbf{m}\ \text{fixed}} = \mathrm{diag}(\mathbf{1} + \gamma\,\mathbf{m}). \tag{9}$$

Each token receives context-conditioned feature gains. Since $\sum_k \pi_k = 1$, the same transformation can be written as $\tilde{\mathbf{h}} = \sum_k \pi_k[(\mathbf{1} + \gamma\mathbf{M}_{k,:}) \odot \mathbf{h}]$, a mixture of diagonal transformations on a shared input. Its compact basis parameterization differs fundamentally from evaluating full expert networks.

D.2 Design Principles: Continuity and Basis Diversity

Continuity bias via soft routing. With soft routing (softmax over pattern logits), the convex combination $\mathbf{m}_{b,\ell} = \sum_k \pi_{b,\ell,k}\,\mathbf{M}_{k,:}$ varies smoothly with local context $\mathbf{z}_{b,\ell}$, so small context changes induce small modulation changes. This gives the gain path a continuous response to local context.

Discrete routing remains differentiable. The discrete-routing ablation in Table 5 replaces the full softmax with a hard top-2 selection: the router keeps the two largest logits, applies a softmax over only these selected logits, and scatters the resulting weights back onto the basis. The modulation is therefore still a differentiable convex combination of the selected basis rows, trained end-to-end without straight-through or Gumbel–Softmax estimators.

Orthogonal basis as a compact coordinate system. The regularizer $\mathcal{L}_{\mathrm{orth}}$ penalizes Gram-matrix deviation from identity, encouraging near-orthogonal pattern vectors. Because the basis rows are L_2-normalized, the Gram diagonal is identically one, so the penalty acts solely on the off-diagonal entries and directly minimizes the sum of squared off-diagonal cosines $\sum_{i\neq j}\cos^2(\mathbf{M}_i, \mathbf{M}_j)$ across response directions. This discourages redundant modes and promotes geometric diversity among the learned response directions without any inference-time cost.

Distinct response directions. Correlated response directions can make routing assignments redundant. The PatchTST geometry results in the note to Table 5 show that the Gram penalty reduces mean off-diagonal basis cosine, while forecast accuracy is nearly unchanged.

D.3 Mechanism-Level Jacobian and Conditional Stability

Equation 9 holds the modulation fixed. For the full derivative, include its dependence on the input through perception and routing. Let $\mathbf{s} = \mathbf{1} + \gamma\,\mathbf{m}$ denote the scaling vector. The (i, j)-th element of the total derivative is

$$\frac{\partial \tilde{h}_i}{\partial h_j} = s_i\,\delta_{ij} + h_i\,\frac{\partial s_i}{\partial h_j}, \tag{10}$$

where δ_{ij} is the Kronecker delta. In matrix form, $\partial\tilde{\mathbf{h}}/\partial\mathbf{h} = \mathrm{diag}(\mathbf{s}) + \mathrm{diag}(\mathbf{h})\,(\partial\mathbf{s}/\partial\mathbf{h})$. Since $\mathbf{s} = \mathbf{1} + \gamma\,\mathbf{m}$, we have $\partial\mathbf{s}/\partial\mathbf{h} = \gamma\,(\partial\mathbf{m}/\partial\mathbf{h})$, so the off-diagonal term is proportional to γ. Near-identity feature scaling requires $\|\gamma\mathbf{m}\|_\infty \ll 1$; the magnitude of the second term also depends on the perception and routing derivatives. Thus $\mathrm{diag}(\mathbf{s})$ describes the direct gain path, while the full Jacobian includes the context-dependent term.

Conditional norm bound. DPR's multiplicative modulation admits a direct bound on forward-representation amplification. Since $\tilde{\mathbf{h}} = \mathbf{h} \odot (\mathbf{1} + \gamma\,\mathbf{m})$ and each component of $\mathbf{m}$ is a convex combination of basis vectors with coefficients in $[0, 1]$, the following conditional bound holds for any finite learned gain:

$$\|\tilde{\mathbf{h}}\|_2 \;\le\; \max_i |1 + \gamma\,m_i| \cdot \|\mathbf{h}\|_2 \;\le\; (1 + |\gamma|\,\max_k \|\mathbf{M}_{k,:}\|_\infty) \cdot \|\mathbf{h}\|_2. \tag{11}$$

The bound is close to one when $|\gamma|\max_k \|\mathbf{M}_{k,:}\|_\infty \ll 1$. It characterizes forward amplification at the current parameter values through the learned gain and basis magnitudes.

E Hyperparameter Sensitivity

Figure 8 evaluates 18 hyperparameter configurations and reports 72 horizon-level MSE points. The default $K = 8$ provides the best accuracy–compactness trade-off, with larger bases yielding similar MSE. Orthogonal regularization remains a training-only geometry constraint, with accuracy stable across the tested λ_{orth} values. Kernel choice follows the host token type. PatchTST prefers $k = 1$ because each patch already summarizes a local span, whereas Crossformer prefers the multi-scale $(3, 7)$ pair used for sequential tokens.

F Complete Experimental Results

This section reports the full per-horizon results and adapter integration gains that complement the averaged tables in the main paper.

F.1 Full Main Results

Table 3 reports four-horizon averages for DPRNet and seven baselines on eight representative benchmarks. Table 11 reports the same eight benchmarks at every prediction horizon. DPRNet's average advantage emerges from horizon-specific outcomes: it ties the lowest reported MSE at the 7-day COVID19 horizon and has the lowest MSE at 60 days, while other models lead at 14 and 28 days.

F.2 Full Adapter Results

Table 12 reports per-dataset, per-horizon gains when DPR is integrated into forecasting backbones. Improvements are most pronounced on COVID19, VIX, and Illness. On the regular ETT controls, averages for local or mixing backbones remain close to the baseline. The improvement magnitude is positively associated with the composite diagnostic scores from Table 8, as quantified by the rank correlations in Table 14. This correlation captures a dataset-level tendency across hosts and horizons.

F.3 Parameter Scaling versus DPR

Table 13 reports the MSE, MAE, parameter, and FLOP values behind Figure 5. Costs are profiled on ETTh1 ($96 \to 96$). DPR improves or preserves accuracy in all nine host–dataset settings and

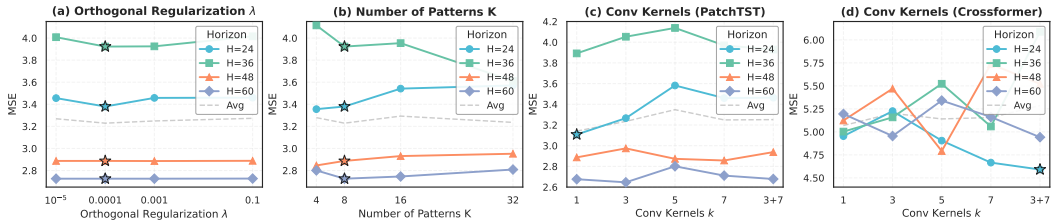


Figure 8: **Sensitivity analysis. (a)** Orthogonal penalty λ_{orth}; **(b)** basis size K; **(c–d)** kernel configurations for PatchTST and Crossformer. Solid lines show individual horizons, dashed lines show horizon averages, and stars mark the preferred settings.

Table 11: **Full per-horizon forecasting results** on eight benchmarks. MSE / MAE; bold/underline = best/second MSE per row (ties allowed after rounding).

Dataset	Len.	Informer	TimesNet	PatchTST	iTransformer	TimeMixer	TimeFilter	WPMixer	DPRNet
		Base	Base	Base	Base	Base	Base	Base	(Ours)
ILI	24	7.005 / 1.868	9.241 / 1.389	3.633 / 1.079	3.507 / 1.071	3.124 / 1.136	**1.991 / 0.873**	3.173 / 1.022	2.970 / 1.071
	36	7.201 / 1.898	7.371 / 1.438	4.019 / 1.192	3.974 / 1.152	3.538 / 1.214	**2.481 / 0.976**	3.720 / 1.147	3.349 / 1.149
	48	7.030 / 1.903	4.175 / 1.237	2.939 / 1.099	2.513 / 1.005	3.055 / 1.130	**2.468 / 0.990**	2.770 / 1.061	2.541 / 1.012
	60	7.140 / 1.916	3.298 / 1.172	2.695 / 1.071	**2.657 / 1.049**	3.010 / 1.114	2.812 / 1.072	2.722 / 1.044	2.692 / 1.052
	Avg	7.094 / 1.896	6.021 / 1.309	3.321 / 1.110	3.163 / 1.069	3.182 / 1.148	**2.438 / 0.978**	3.096 / 1.069	2.888 / 1.071
COVID19	7	0.824 / 0.391	0.490 / 0.390	0.335 / 0.216	**0.327 / 0.217**	0.361 / 0.238	0.331 / 0.220	0.343 / 0.218	**0.327 / 0.224**
	14	0.849 / 0.433	0.690 / 0.484	0.440 / 0.264	0.439 / 0.269	0.464 / 0.288	**0.434 / 0.275**	0.473 / 0.267	0.440 / 0.275
	28	3.027 / 0.813	0.927 / 0.550	0.837 / 0.387	**0.787 / 0.385**	0.855 / 0.410	0.866 / 0.439	0.872 / 0.397	0.813 / 0.396
	60	2.970 / 0.893	2.161 / 0.728	1.743 / 0.582	2.055 / 0.643	2.041 / 0.668	2.002 / 0.643	1.667 / 0.569	**1.637 / 0.569**
	Avg	1.917 / 0.632	1.067 / 0.538	0.839 / 0.362	0.902 / 0.379	0.930 / 0.401	0.908 / 0.394	0.839 / 0.363	**0.804 / 0.366**
Weather	96	1.094 / 0.753	0.181 / 0.227	0.173 / 0.208	0.185 / 0.217	0.180 / 0.226	**0.161 / 0.200**	0.163 / 0.201	0.165 / 0.202
	192	0.294 / 0.328	0.246 / 0.283	0.215 / 0.248	0.235 / 0.258	0.246 / 0.282	0.211 / 0.245	**0.208 / 0.242**	**0.208 / 0.241**
	336	**0.260 / 0.312**	0.311 / 0.324	0.269 / 0.288	0.291 / 0.299	0.309 / 0.323	0.268 / 0.286	0.262 / 0.281	0.261 / 0.280
	720	**0.289 / 0.336**	0.417 / 0.387	0.345 / 0.337	0.362 / 0.345	0.395 / 0.375	0.346 / 0.337	0.340 / 0.332	0.337 / 0.329
	Avg	0.484 / 0.432	0.289 / 0.305	0.251 / 0.270	0.268 / 0.280	0.282 / 0.301	0.246 / 0.267	**0.243 / 0.264**	**0.243 / 0.263**
VIX	96	1.077 / 0.668	1.005 / 0.562	0.942 / 0.539	0.990 / 0.560	0.967 / 0.550	0.959 / 0.545	0.957 / 0.547	**0.935 / 0.542**
	192	2.059 / 0.975	1.275 / 0.693	1.268 / 0.692	1.208 / 0.674	1.203 / 0.670	**1.181 / 0.662**	1.193 / 0.665	1.193 / 0.668
	336	**1.011 / 0.700**	1.342 / 0.802	1.224 / 0.762	1.176 / 0.751	1.236 / 0.768	1.172 / 0.739	1.170 / 0.749	1.174 / 0.746
	720	**0.896 / 0.634**	1.205 / 0.801	1.141 / 0.773	1.238 / 0.809	1.160 / 0.786	1.147 / 0.764	1.156 / 0.782	1.130 / 0.774
	Avg	1.261 / 0.744	1.207 / 0.715	1.144 / 0.692	1.153 / 0.699	1.141 / 0.694	1.115 / 0.677	1.119 / 0.686	**1.108 / 0.682**
NABCPU	24	2.949 / 1.213	1.176 / 0.367	**1.128 / 0.306**	1.133 / 0.327	1.140 / 0.305	1.134 / 0.302	1.137 / 0.325	1.130 / 0.311
	48	3.218 / 1.070	1.170 / 0.358	1.151 / 0.320	1.155 / 0.335	1.147 / 0.298	1.153 / 0.301	1.159 / 0.328	**1.146 / 0.309**
	96	2.334 / 0.978	1.216 / 0.337	1.201 / 0.316	1.209 / 0.343	1.198 / 0.299	1.203 / 0.306	1.212 / 0.338	**1.197 / 0.304**
	192	2.254 / 0.880	1.316 / 0.355	1.300 / 0.329	1.304 / 0.348	1.301 / 0.318	1.298 / 0.320	1.300 / 0.343	**1.297 / 0.330**
	Avg	2.689 / 1.035	1.220 / 0.354	1.195 / 0.318	1.200 / 0.338	1.196 / 0.305	1.197 / 0.307	1.202 / 0.334	**1.192 / 0.314**
Exchange	96	2.295 / 1.093	0.137 / 0.266	0.106 / 0.229	0.106 / 0.231	0.102 / 0.224	0.107 / 0.230	0.102 / 0.224	**0.101 / 0.223**
	192	4.254 / 1.581	0.243 / 0.368	0.214 / 0.332	0.206 / 0.327	0.204 / 0.325	0.213 / 0.332	0.204 / 0.325	**0.203 / 0.325**
	336	1.419 / 0.893	0.429 / 0.489	0.414 / 0.467	0.394 / 0.459	**0.384 / 0.450**	0.396 / 0.456	0.393 / 0.454	0.389 / 0.451
	720	3.160 / 1.374	1.166 / 0.830	1.077 / 0.789	1.078 / 0.791	1.069 / 0.784	**1.068 / 0.788**	1.080 / 0.792	**1.068 / 0.783**
	Avg	2.782 / 1.235	0.494 / 0.488	0.453 / 0.454	0.446 / 0.452	**0.440 / 0.446**	0.446 / 0.452	0.445 / 0.449	**0.440 / 0.446**
ETTh1	96	1.569 / 0.902	0.488 / 0.475	0.394 / 0.392	0.384 / 0.391	0.401 / 0.395	0.390 / 0.390	**0.382 / 0.388**	0.383 / 0.389
	192	1.410 / 0.856	0.513 / 0.478	0.447 / 0.423	**0.438 / 0.422**	0.443 / 0.420	0.442 / 0.421	0.441 / 0.420	0.442 / 0.421
	336	1.221 / 0.762	0.581 / 0.505	0.490 / 0.444	0.487 / 0.446	0.492 / 0.441	0.487 / 0.442	0.492 / 0.448	**0.484 / 0.442**
	720	1.094 / 0.742	0.556 / 0.509	0.506 / 0.470	**0.481 / 0.466**	0.496 / 0.460	0.486 / 0.462	0.501 / 0.470	0.484 / 0.463
	Avg	1.324 / 0.816	0.534 / 0.492	0.459 / 0.432	**0.448 / 0.431**	0.458 / 0.429	0.451 / 0.429	0.454 / 0.431	**0.448 / 0.429**
ETTm1	96	1.502 / 0.870	0.455 / 0.439	0.327 / 0.348	0.330 / 0.351	0.318 / 0.342	0.323 / 0.347	**0.317 / 0.344**	0.327 / 0.348
	192	1.264 / 0.792	0.524 / 0.472	0.374 / 0.370	0.379 / 0.376	0.373 / 0.372	0.375 / 0.374	**0.371 / 0.369**	0.374 / 0.370
	336	1.453 / 0.827	0.543 / 0.489	0.406 / 0.395	0.416 / 0.399	0.407 / 0.394	0.409 / 0.396	**0.403 / 0.393**	0.407 / 0.392
	720	1.700 / 0.929	0.554 / 0.495	0.477 / 0.433	0.473 / 0.433	0.474 / 0.432	0.473 / 0.432	**0.469 / 0.430**	0.475 / 0.429
	Avg	1.480 / 0.855	0.519 / 0.474	0.396 / 0.387	0.399 / 0.390	0.393 / 0.385	0.395 / 0.387	**0.390 / 0.384**	0.396 / 0.385

outperforms the parameter-matched static baseline in every one, using parameters comparable to or below the scaled variants.

G Controls, Robustness, and Evaluation Details

This section reports matched-control protocols, the TimeBase contrast, the four-host seed study, isolated-GPU timing, and additional standalone-benchmark configurations. Unless stated otherwise, values are MSE/MAE, lower is better, and are averaged over the four predefined horizons. Adapter reruns use the same preprocessing, BasicTS evaluation code, and late-stage insertion rule.

Table 12: **Full forecasting results of the DPR enhancement.** Per-horizon MSE / MAE; gray = +DPR; bold = lower or equal MSE vs. Base at the reported precision.

Dataset	Len.	Informer		Crossformer		TimesNet		PatchTST		TimeMixer		TimeFilter		WPMixer	
		Base	+DPR	Base	+DPR	Base	+DPR	Base	+DPR	Base	+DPR	Base	+DPR	Base	+DPR
ILI	24	7.005 / 1.868	**5.666 / 1.764**	4.736 / 1.480	**4.593 / 1.428**	9.241 / 1.389	**3.445 / 1.054**	3.633 / 1.079	**3.108 / 1.042**	3.124 / 1.136	**3.123 / 1.142**	1.991 / 0.873	**1.821 / 0.848**	3.173 / 1.022	**2.796 / 1.046**
	36	7.201 / 1.898	**5.529 / 1.725**	5.153 / 1.561	**4.734 / 1.468**	7.371 / 1.438	**7.147 / 1.441**	4.019 / 1.192	**3.891 / 1.207**	3.538 / 1.214	**3.428 / 1.194**	2.481 / 0.976	**2.460 / 0.974**	3.720 / 1.147	**3.310 / 1.052**
	48	7.030 / 1.903	**4.887 / 1.614**	5.244 / 1.576	**4.669 / 1.466**	4.175 / 1.237	**3.835 / 1.194**	2.939 / 1.099	**2.857 / 1.095**	3.055 / 1.130	**3.047 / 1.126**	2.468 / 0.990	**2.438 / 0.982**	2.770 / 1.061	2.932 / 1.090
	60	7.140 / 1.916	**6.879 / 1.942**	5.006 / 1.550	**4.801 / 1.500**	3.298 / 1.172	**3.251 / 1.141**	2.695 / 1.071	**2.599 / 1.038**	3.010 / 1.114	**2.989 / 1.118**	2.812 / 1.072	**2.457 / 0.998**	2.722 / 1.044	2.822 / 1.060
	Avg	7.094 / 1.896	**5.740 / 1.761**	5.035 / 1.542	**4.699 / 1.466**	6.021 / 1.309	**4.420 / 1.208**	3.321 / 1.110	**3.114 / 1.095**	3.182 / 1.148	**3.147 / 1.145**	2.438 / 0.978	**2.294 / 0.951**	3.096 / 1.069	**2.965 / 1.062**
COVID19	7	0.824 / 0.391	0.941 / 0.437	0.609 / 0.295	**0.587 / 0.269**	0.490 / 0.390	**0.480 / 0.388**	0.335 / 0.216	**0.327 / 0.217**	0.361 / 0.238	0.363 / 0.238	0.331 / 0.220	**0.323 / 0.219**	0.343 / 0.218	**0.318 / 0.218**
	14	0.849 / 0.433	**0.676 / 0.369**	0.853 / 0.356	**0.823 / 0.355**	0.690 / 0.484	0.699 / 0.488	0.440 / 0.264	**0.435 / 0.262**	0.464 / 0.288	**0.455 / 0.285**	0.434 / 0.275	**0.425 / 0.272**	0.473 / 0.267	**0.434 / 0.265**
	28	3.027 / 0.813	**2.442 / 0.855**	1.444 / 0.531	**1.287 / 0.513**	0.927 / 0.550	**0.910 / 0.545**	0.837 / 0.387	**0.806 / 0.383**	0.855 / 0.410	**0.842 / 0.405**	0.866 / 0.439	**0.820 / 0.429**	0.872 / 0.397	**0.784 / 0.389**
	60	2.970 / 0.893	**2.907 / 0.876**	2.700 / 0.724	**2.608 / 0.768**	2.161 / 0.728	**2.052 / 0.705**	1.743 / 0.582	**1.687 / 0.573**	2.041 / 0.668	**1.796 / 0.614**	2.002 / 0.643	**1.622 / 0.577**	1.667 / 0.569	**1.606 / 0.562**
	Avg	1.917 / 0.632	**1.742 / 0.634**	1.401 / 0.476	**1.326 / 0.476**	1.067 / 0.538	**1.035 / 0.531**	0.839 / 0.362	**0.814 / 0.359**	0.930 / 0.401	**0.864 / 0.386**	0.908 / 0.394	**0.797 / 0.374**	0.839 / 0.363	**0.786 / 0.359**
Weather	96	1.094 / 0.753	**0.767 / 0.607**	0.166 / 0.204	0.169 / 0.208	0.181 / 0.227	0.183 / 0.228	0.173 / 0.208	**0.167 / 0.204**	0.180 / 0.226	**0.164 / 0.200**	0.161 / 0.200	**0.158 / 0.198**	0.163 / 0.201	**0.163 / 0.200**
	192	0.294 / 0.328	**0.238 / 0.288**	0.208 / 0.244	**0.207 / 0.243**	0.246 / 0.283	**0.245 / 0.281**	0.215 / 0.248	**0.214 / 0.248**	0.246 / 0.282	**0.208 / 0.241**	0.211 / 0.245	**0.206 / 0.241**	0.208 / 0.242	**0.207 / 0.241**
	336	0.260 / 0.312	**0.260 / 0.305**	0.260 / 0.286	**0.254 / 0.279**	0.311 / 0.324	0.312 / 0.325	0.269 / 0.288	**0.269 / 0.288**	0.309 / 0.323	**0.262 / 0.280**	0.268 / 0.286	**0.268 / 0.286**	0.262 / 0.281	**0.262 / 0.281**
	720	0.289 / 0.336	**0.280 / 0.325**	0.331 / 0.339	**0.328 / 0.337**	0.417 / 0.387	**0.400 / 0.378**	0.345 / 0.337	**0.344 / 0.336**	0.395 / 0.375	**0.339 / 0.332**	0.346 / 0.337	**0.345 / 0.336**	0.340 / 0.332	**0.340 / 0.332**
	Avg	0.484 / 0.432	**0.386 / 0.381**	0.241 / 0.268	**0.239 / 0.267**	0.289 / 0.305	**0.285 / 0.303**	0.251 / 0.270	**0.248 / 0.269**	0.282 / 0.301	**0.243 / 0.263**	0.246 / 0.267	**0.244 / 0.265**	0.243 / 0.264	**0.243 / 0.264**
VIX	96	1.077 / 0.668	**0.938 / 0.665**	1.023 / 0.571	**0.970 / 0.561**	1.005 / 0.562	**0.972 / 0.544**	0.942 / 0.539	**0.940 / 0.538**	0.967 / 0.550	**0.954 / 0.541**	0.959 / 0.545	**0.947 / 0.543**	0.957 / 0.547	**0.938 / 0.538**
	192	2.059 / 0.975	**1.115 / 0.734**	1.245 / 0.683	**1.181 / 0.655**	1.275 / 0.693	**1.198 / 0.676**	1.268 / 0.692	**1.222 / 0.679**	1.203 / 0.670	**1.188 / 0.665**	1.181 / 0.662	1.194 / 0.663	1.193 / 0.665	**1.187 / 0.663**
	336	1.011 / 0.700	**0.928 / 0.638**	1.240 / 0.769	**1.207 / 0.750**	1.342 / 0.802	**1.201 / 0.759**	1.224 / 0.762	**1.202 / 0.753**	1.236 / 0.768	**1.185 / 0.752**	1.172 / 0.739	**1.147 / 0.727**	1.170 / 0.749	**1.164 / 0.743**
	720	0.896 / 0.634	**0.774 / 0.638**	1.046 / 0.717	1.048 / 0.747	1.205 / 0.801	**1.166 / 0.781**	1.141 / 0.773	**1.120 / 0.765**	1.160 / 0.786	**1.140 / 0.775**	1.147 / 0.764	**1.089 / 0.755**	1.156 / 0.782	**1.130 / 0.774**
	Avg	1.261 / 0.744	**0.939 / 0.669**	1.139 / 0.685	**1.102 / 0.678**	1.207 / 0.715	**1.134 / 0.690**	1.144 / 0.692	**1.121 / 0.684**	1.141 / 0.694	**1.117 / 0.683**	1.115 / 0.677	**1.094 / 0.672**	1.119 / 0.686	**1.105 / 0.679**
NABCPU	24	2.949 / 1.213	**1.898 / 0.711**	1.189 / 0.366	**1.162 / 0.344**	1.176 / 0.367	**1.154 / 0.357**	1.128 / 0.306	**1.126 / 0.301**	1.140 / 0.305	**1.138 / 0.300**	1.134 / 0.302	1.135 / 0.301	1.137 / 0.325	**1.127 / 0.304**
	48	3.218 / 1.070	**1.754 / 0.758**	1.194 / 0.371	1.196 / 0.366	1.170 / 0.358	1.182 / 0.356	1.151 / 0.320	**1.150 / 0.314**	1.147 / 0.298	**1.146 / 0.294**	1.153 / 0.301	**1.148 / 0.300**	1.159 / 0.328	**1.142 / 0.308**
	96	2.334 / 0.978	**2.195 / 0.930**	1.257 / 0.400	**1.240 / 0.380**	1.216 / 0.337	1.222 / 0.338	1.201 / 0.316	**1.201 / 0.320**	1.198 / 0.299	**1.198 / 0.295**	1.203 / 0.306	**1.201 / 0.306**	1.212 / 0.338	**1.198 / 0.312**
	192	2.254 / 0.880	**1.857 / 0.710**	1.389 / 0.420	**1.346 / 0.397**	1.316 / 0.355	**1.311 / 0.359**	1.300 / 0.329	**1.297 / 0.329**	1.301 / 0.318	**1.296 / 0.313**	1.298 / 0.320	**1.296 / 0.317**	1.300 / 0.343	**1.291 / 0.325**
	Avg	2.689 / 1.035	**1.926 / 0.777**	1.257 / 0.389	**1.236 / 0.372**	1.220 / 0.354	**1.217 / 0.352**	1.195 / 0.318	**1.194 / 0.316**	1.196 / 0.305	**1.194 / 0.300**	1.197 / 0.307	**1.195 / 0.306**	1.202 / 0.334	**1.190 / 0.312**
Sunspots	12	3.702 / 1.545	**1.238 / 0.916**	0.200 / 0.332	**0.199 / 0.324**	0.249 / 0.365	0.250 / 0.366	0.251 / 0.365	**0.249 / 0.365**	0.254 / 0.368	**0.252 / 0.367**	0.245 / 0.363	0.252 / 0.366	0.251 / 0.369	**0.246 / 0.365**
	24	2.018 / 1.247	**1.706 / 1.090**	0.279 / 0.378	**0.267 / 0.369**	0.461 / 0.488	**0.439 / 0.477**	0.441 / 0.475	**0.437 / 0.475**	0.470 / 0.487	**0.444 / 0.475**	0.436 / 0.470	**0.431 / 0.465**	0.444 / 0.479	**0.431 / 0.470**
	48	1.375 / 0.970	1.563 / 0.998	0.462 / 0.509	**0.417 / 0.470**	1.006 / 0.709	1.013 / 0.709	0.954 / 0.702	**0.951 / 0.700**	1.034 / 0.717	**1.005 / 0.710**	1.081 / 0.737	**0.966 / 0.700**	0.974 / 0.709	**0.954 / 0.702**
	96	3.155 / 1.424	**1.549 / 1.001**	0.564 / 0.545	**0.546 / 0.532**	1.091 / 0.750	**1.082 / 0.748**	1.090 / 0.763	**1.084 / 0.759**	1.130 / 0.773	**1.055 / 0.751**	1.214 / 0.789	**1.171 / 0.780**	1.048 / 0.746	1.059 / 0.755
	Avg	2.562 / 1.296	**1.514 / 1.001**	0.376 / 0.441	**0.357 / 0.424**	0.702 / 0.578	**0.696 / 0.575**	0.684 / 0.576	**0.680 / 0.575**	0.722 / 0.586	**0.689 / 0.576**	0.744 / 0.590	**0.705 / 0.578**	0.679 / 0.576	**0.672 / 0.573**
Exchange	96	2.295 / 1.093	**1.936 / 0.975**	0.269 / 0.349	**0.237 / 0.335**	0.137 / 0.266	**0.128 / 0.260**	0.106 / 0.229	**0.104 / 0.226**	0.102 / 0.224	**0.101 / 0.224**	0.107 / 0.230	**0.103 / 0.226**	0.102 / 0.224	**0.102 / 0.223**
	192	4.254 / 1.581	**1.447 / 0.871**	0.570 / 0.529	**0.559 / 0.522**	0.243 / 0.368	**0.234 / 0.361**	0.214 / 0.332	**0.213 / 0.331**	0.204 / 0.325	0.207 / 0.327	0.213 / 0.332	**0.208 / 0.328**	0.204 / 0.325	0.207 / 0.326
	336	1.419 / 0.893	**1.245 / 0.803**	1.046 / 0.723	**0.873 / 0.663**	0.429 / 0.489	**0.422 / 0.485**	0.414 / 0.467	**0.400 / 0.459**	0.384 / 0.450	**0.384 / 0.450**	0.396 / 0.456	0.398 / 0.459	0.393 / 0.454	**0.392 / 0.454**
	720	3.160 / 1.374	**2.262 / 1.142**	1.692 / 0.983	**1.691 / 0.980**	1.166 / 0.830	**1.118 / 0.813**	1.077 / 0.789	**1.073 / 0.786**	1.069 / 0.784	**1.069 / 0.784**	1.068 / 0.788	**1.064 / 0.782**	1.080 / 0.792	**1.068 / 0.785**
	Avg	2.782 / 1.235	**1.723 / 0.948**	0.894 / 0.646	**0.840 / 0.625**	0.494 / 0.488	**0.476 / 0.480**	0.453 / 0.454	**0.448 / 0.451**	0.440 / 0.446	**0.440 / 0.446**	0.446 / 0.452	**0.443 / 0.449**	0.445 / 0.449	**0.442 / 0.447**
ETTh1	96	1.569 / 0.902	**1.062 / 0.770**	0.394 / 0.404	**0.382 / 0.397**	0.488 / 0.475	**0.475 / 0.466**	0.394 / 0.392	**0.392 / 0.392**	0.401 / 0.395	**0.397 / 0.393**	0.390 / 0.390	**0.390 / 0.390**	0.382 / 0.388	**0.381 / 0.387**
	192	1.410 / 0.856	**1.099 / 0.783**	0.436 / 0.431	**0.427 / 0.421**	0.513 / 0.478	0.520 / 0.482	0.447 / 0.423	**0.442 / 0.422**	0.443 / 0.420	0.446 / 0.420	0.442 / 0.421	0.443 / 0.420	0.441 / 0.420	**0.436 / 0.419**
	336	1.221 / 0.762	**0.919 / 0.675**	0.471 / 0.453	**0.456 / 0.442**	0.581 / 0.505	**0.554 / 0.496**	0.490 / 0.444	**0.488 / 0.444**	0.492 / 0.441	0.493 / 0.441	0.487 / 0.442	0.489 / 0.441	0.492 / 0.448	**0.491 / 0.447**
	720	1.094 / 0.742	**0.920 / 0.711**	0.495 / 0.493	0.498 / 0.500	0.556 / 0.509	0.576 / 0.529	0.506 / 0.470	0.507 / 0.401	0.496 / 0.460	**0.487 / 0.460**	0.486 / 0.462	0.493 / 0.468	0.501 / 0.470	0.504 / 0.473
	Avg	1.324 / 0.816	**1.000 / 0.735**	0.449 / 0.445	**0.441 / 0.440**	0.534 / 0.492	**0.531 / 0.493**	0.459 / 0.432	**0.457 / 0.415**	0.458 / 0.429	**0.456 / 0.428**	0.451 / 0.429	0.454 / 0.430	0.454 / 0.431	**0.453 / 0.431**
ETTh2	96	5.257 / 1.710	**3.206 / 1.359**	0.339 / 0.395	**0.311 / 0.358**	0.448 / 0.428	**0.435 / 0.421**	0.293 / 0.338	**0.293 / 0.336**	0.297 / 0.338	0.299 / 0.339	0.295 / 0.340	**0.291 / 0.340**	0.291 / 0.336	0.293 / 0.336
	192	2.512 / 1.150	**1.738 / 0.941**	0.438 / 0.440	**0.396 / 0.420**	0.520 / 0.470	**0.518 / 0.470**	0.376 / 0.391	**0.376 / 0.392**	0.375 / 0.390	**0.374 / 0.389**	0.374 / 0.391	0.377 / 0.393	0.366 / 0.385	0.370 / 0.388
	336	0.997 / 0.722	**0.702 / 0.610**	0.491 / 0.488	0.588 / 0.552	0.497 / 0.475	0.510 / 0.480	0.425 / 0.431	**0.418 / 0.430**	0.435 / 0.434	0.436 / 0.435	0.421 / 0.431	**0.417 / 0.426**	0.421 / 0.427	0.423 / 0.427
	720	0.969 / 0.713	0.999 / 0.723	1.913 / 1.072	**1.384 / 0.884**	0.456 / 0.465	**0.444 / 0.459**	0.441 / 0.450	**0.431 / 0.446**	0.437 / 0.447	**0.436 / 0.447**	0.430 / 0.444	0.445 / 0.450	0.439 / 0.446	**0.431 / 0.444**
	Avg	2.434 / 1.074	**1.661 / 0.908**	0.795 / 0.599	**0.670 / 0.553**	0.480 / 0.460	**0.477 / 0.458**	0.384 / 0.403	**0.380 / 0.401**	0.386 / 0.402	**0.386 / 0.403**	0.380 / 0.402	0.383 / 0.402	0.379 / 0.399	**0.379 / 0.399**
ETTm2	96	8.137 / 2.167	**3.183 / 1.361**	0.181 / 0.264	**0.180 / 0.264**	0.207 / 0.282	0.226 / 0.294	0.175 / 0.252	0.176 / 0.253	0.173 / 0.250	**0.172 / 0.250**	0.174 / 0.254	**0.173 / 0.253**	0.172 / 0.250	**0.172 / 0.250**
	192	4.557 / 1.490	**0.358 / 0.406**	0.237 / 0.306	0.256 / 0.328	0.289 / 0.332	0.290 / 0.333	0.240 / 0.296	**0.240 / 0.296**	0.239 / 0.295	0.240 / 0.295	0.238 / 0.296	**0.237 / 0.295**	0.236 / 0.293	**0.236 / 0.293**
	336	0.351 / 0.429	**0.332 / 0.409**	0.313 / 0.367	0.365 / 0.399	0.362 / 0.378	0.366 / 0.378	0.301 / 0.335	0.302 / 0.335	0.300 / 0.333	**0.300 / 0.334**	0.303 / 0.337	**0.301 / 0.336**	0.294 / 0.332	0.295 / 0.331
	720	0.384 / 0.425	0.419 / 0.458	0.784 / 0.642	**0.774 / 0.637**	0.534 / 0.463	**0.520 / 0.456**	0.402 / 0.394	**0.402 / 0.394**	0.398 / 0.391	**0.398 / 0.392**	0.407 / 0.395	**0.406 / 0.396**	0.394 / 0.390	0.396 / 0.392
	Avg	3.357 / 1.128	**1.073 / 0.658**	0.379 / 0.395	0.394 / 0.407	0.348 / 0.364	0.350 / 0.365	0.279 / 0.319	0.280 / 0.320	0.277 / 0.317	**0.277 / 0.318**	0.280 / 0.321	**0.279 / 0.320**	0.274 / 0.316	0.275 / 0.317
ETTm1	96	1.502 / 0.870	**1.069 / 0.776**	0.345 / 0.362	**0.326 / 0.351**	0.455 / 0.439	**0.454 / 0.439**	0.327 / 0.348	0.328 / 0.350	0.318 / 0.342	0.319 / 0.342	0.323 / 0.347	**0.322 / 0.346**	0.317 / 0.344	0.318 / 0.344
	192	1.264 / 0.792	**1.049 / 0.771**	0.378 / 0.381	**0.377 / 0.379**	0.524 / 0.472	0.525 / 0.472	0.374 / 0.370	**0.372 / 0.370**	0.373 / 0.372	0.374 / 0.371	0.375 / 0.374	**0.375 / 0.373**	0.371 / 0.369	**0.369 / 0.368**
	336	1.453 / 0.827	**1.229 / 0.756**	0.430 / 0.411	0.432 / 0.425	0.543 / 0.489	**0.535 / 0.487**	0.406 / 0.395	**0.406 / 0.394**	0.407 / 0.394	**0.404 / 0.394**	0.409 / 0.396	**0.407 / 0.394**	0.403 / 0.393	**0.403 / 0.393**
	720	1.700 / 0.929	**1.039 / 0.762**	0.498 / 0.463	**0.463 / 0.434**	0.554 / 0.495	**0.553 / 0.491**	0.477 / 0.433	**0.475 / 0.433**	0.474 / 0.432	**0.472 / 0.432**	0.473 / 0.432	**0.472 / 0.432**	0.469 / 0.430	**0.467 / 0.429**
	Avg	1.480 / 0.855	**1.097 / 0.766**	0.413 / 0.404	**0.400 / 0.397**	0.519 / 0.474	**0.517 / 0.472**	0.396 / 0.387	**0.395 / 0.387**	0.393 / 0.385	**0.392 / 0.385**	0.395 / 0.387	**0.394 / 0.386**	0.390 / 0.384	**0.389 / 0.384**

Table 13: **Parameter scaling vs. DPR. 2xW/2xD/2xB** = width/depth/both $\times 2$. Gray = +DPR; bold/underline = best/second MSE; $\uparrow/\downarrow$ = MSE change vs. Raw. Costs on ETTh1 ($96 \rightarrow 96$).

Base Model	Variant	Params (M)	FLOPs (GMac)	ETTh1 ($96 \rightarrow 96$) MSE / MAE	Impv.	ILI ($24 \rightarrow 24$) MSE / MAE	Impv.	Exchange ($96 \rightarrow 96$) MSE / MAE	Impv.
PatchTST	Raw	1.08	0.069	0.394 / 0.392	-	3.633 / 1.079	-	0.106 / 0.229	-
	2xW	3.75	0.269	0.405 / 0.397	↓2.8%	3.415 / 1.073	↑6.0%	0.110 / 0.232	↓3.8%
	2xD	1.87	0.135	0.401 / 0.395	↓1.8%	3.199 / 1.041	↑11.9%	0.110 / 0.231	↓3.8%
	2xB	6.90	0.534	0.398 / 0.394	↓1.0%	3.114 / 1.024	↑14.3%	0.111 / 0.233	↓4.7%
	PM	1.08	0.069	0.394 / 0.392	-	3.633 / 1.079	-	0.108 / 0.231	↓1.9%
	+DPR	**1.10**	**0.070**	**0.392 / 0.392**	-	**3.108 / 1.042**	↑14.5%	**0.104 / 0.226**	↑1.9%
TimesNet	Raw	18.5	17.75	0.488 / 0.475	-	9.241 / 1.389	-	0.137 / 0.266	-
	2xW	73.7	70.47	0.561 / 0.509	↓15.0%	5.859 / 1.250	↑36.6%	0.132 / 0.263	↑3.6%
	2xD	36.8	36.55	0.503 / 0.481	↓3.1%	6.827 / 1.242	↑26.1%	0.145 / 0.277	↓5.8%
	2xB	147.0	142.4	0.598 / 0.518	↓22.5%	3.515 / 1.081	↑62.0%	0.138 / 0.270	↓0.7%
	PM	18.5	17.75	0.488 / 0.475	-	9.345 / 1.392	↓1.1%	0.129 / 0.262	↑5.8%
	+DPR	**18.5**	**17.84**	**0.475 / 0.466**	↑2.7%	**3.445 / 1.054**	↑62.7%	**0.128 / 0.260**	↑6.6%
TimeFilter	Raw	0.19	0.011	**0.390 / 0.390**	-	1.991 / 0.873	-	0.107 / 0.230	-
	2xW	2.99	0.214	0.399 / 0.397	↓2.3%	3.264 / 0.975	↓63.9%	0.106 / 0.229	↑0.9%
	2xD	1.50	0.107	0.397 / 0.395	↓1.8%	3.285 / 0.913	↓65.0%	0.104 / 0.229	↑2.8%
	2xB	5.39	0.424	0.403 / 0.398	↓3.3%	2.656 / 0.960	↓33.4%	0.112 / 0.237	↓4.7%
	PM	0.19	0.011	0.391 / 0.390	↓0.3%	2.808 / 0.986	↓41.0%	0.105 / 0.228	↑1.9%
	+DPR	**0.19**	**0.011**	**0.390 / 0.390**	-	**1.821 / 0.848**	↑8.5%	**0.103 / 0.226**	↑3.7%

G.1 Control Protocol Details

Parameter-matched static baseline and MoE control. The parameter-matched static baseline is a single-layer copy of the same host with DPR removed, its hidden width grid-searched over $\{32, 48, \ldots, 1024\}$ (feed-forward $4\times$ hidden) to match the base host plus DPR within 50%; it adds

parameters but no routing. The MoE control replaces DPR with 8 experts, each a two-layer MLP ($d \to d/2 \to d$, GELU, dropout 0.1), and a linear router ($d \to 8$) with noisy top-k gating ($k \in \{1, 2, 4\}$): train-time Gaussian noise is added to the logits, and a softmax over the selected top-k logits yields differentiable weights, with a CV^2 load-balancing term (coefficient 0.01) added to the loss. All variants use batch size 64, 100 epochs (early stopping, patience 10), and seed 42. Reported *Time* is wall-clock seconds for one training epoch and one full test-set pass on the same GPU, distinct from the isolated-latency protocol of Table 4.

G.2 Matched Conditional-Modulation Controls

We compare DPR against four conditional-modulation controls inserted at identical locations and matched within a 5% parameter budget: global SE, local SE, local FiLM, and a gated residual. These controls isolate the shared-basis parameterization: local FiLM generates per-token affine parameters directly from local queries, whereas DPR composes its gain from a shared, orthogonally regularized basis. Matched-control MSE/MAE are in main-text Table 2. The dataset-level correlation check in Table 14 first takes the median relative MSE reduction across the seven hosts with complete horizon records at each horizon, then averages the four horizon-wise medians. Among the individual diagnostics, volatility-of-volatility has the strongest association ($\rho = 0.636$, $p = 0.035$), whereas spectral entropy and ADF p-value do not.

Table 14: **Spearman association** between dataset-level DPR gain and dataset diagnostics ($n = 11$). Median host gain per horizon, then average over four horizons; p-values are unadjusted. Composite = $\mathrm{rank}(H_s) + \mathrm{rank}(\mathrm{VoV})$.

Diagnostic	ρ	two-sided p
ADF p-value	-0.009	0.979
Spectral entropy H_s	0.436	0.180
Volatility-of-volatility	0.636	0.035
Composite rank $\mathrm{rank}(H_s) + \mathrm{rank}(\mathrm{VoV})$	0.868	0.001

G.3 Extended Architecture Evaluation

Plug-in results for three additional hosts—OLinear (Yue et al., 2025), TimeMixer++ (Wang et al., 2025), and TimeBase (Huang et al., 2025)—are in Table 15. These contribute 18 backbone–dataset pairs across six shared benchmarks, supplementing the 7×11 core matrix to complete the 95 evaluations summarized in Section 4.1. TimeBase compresses historical segments along the time axis and decodes temporal components. DPR stores a feature-response basis $\mathbf{M} \in \mathbb{R}^{K \times d}$ whose local mixture produces a feature-wise gain for an already contextualized token, independent of look-back length, patch length, and forecast horizon.

G.4 Multi-seed Stability and Efficiency

We repeat Base/+DPR comparisons with three seeds for four hosts on ILI, COVID19, VIX, and ETTh1, yielding 16 host–dataset blocks and 48 paired seed comparisons. Table 16 reports all four hosts and paired relative MSE changes over seeds 2023, 2024, and 2025. Mean MSE decreases on all twelve volatile-benchmark settings for Informer, PatchTST, WPMixer, and TimeFilter. Nine of the 16 paired bootstrap intervals lie entirely above zero and seven include zero; with three seed pairs per setting, these intervals are descriptive. ETTh1 changes are small for the modern hosts, while Informer gains 28.3% (interval $[24.0, 32.3]$). Table 17 separately reports standalone model rankings.

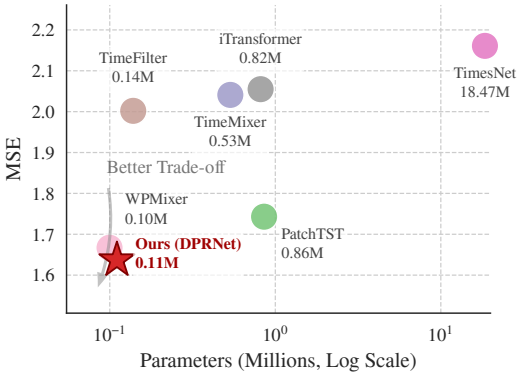


Figure 9: **Accuracy–parameter trade-off on COVID19.** Raw-model MSE at horizon 60 versus parameters in the corresponding configuration.

Isolated efficiency uses one exclusively occupied NVIDIA A800 80GB GPU on ETTh1 ($96 \to 96$), batch size 64, 133 batches per epoch, zero data-loader workers, 20 warm-up iterations, and 100 synchronized timed iterations repeated five times. Training values average ten epochs. Table 4 reports the resulting wall-clock, GMAC, and memory measurements; training time includes the orthogonal penalty, which is removed at inference.

Table 15: **DPR transfer and standalone DPRNet on six benchmarks.** Four-horizon MSE / MAE. Bold = lower or equal MSE within each Base/+DPR pair; gray = standalone DPRNet (bold MSE when best vs. all three Bases).

Backbone	Variant	ILI	COVID19	VIX	Exchange	ETTh1	ETTm1
OLinear	Base	3.344 / 1.058	0.815 / 0.368	1.134 / 0.695	0.489 / 0.468	0.444 / 0.435	0.383 / 0.384
	+DPR	**2.837 / 0.998**	**0.799** / 0.368	**1.118 / 0.694**	**0.451 / 0.453**	0.446 / 0.436	0.383 / 0.384
TimeMixer++	Base	3.444 / 1.172	1.248 / 0.485	1.187 / 0.734	0.478 / 0.471	0.458 / 0.456	0.398 / 0.409
	+DPR	**3.365 / 1.169**	**1.230 / 0.483**	**1.145 / 0.714**	**0.471 / 0.468**	**0.456 / 0.455**	**0.391 / 0.405**
TimeBase	Base	8.859 / 2.125	2.427 / 0.806	1.283 / 0.792	0.518 / 0.505	0.471 / 0.434	0.616 / 0.514
	+DPR	**8.833 / 2.119**	**2.356 / 0.771**	**1.276 / 0.791**	0.526 / 0.510	0.471 / **0.433**	**0.532 / 0.471**
DPRNet	Standalone	**2.888** / 1.071	**0.804** / 0.366	**1.108** / 0.682	**0.440** / 0.446	0.448 / 0.429	0.396 / 0.385

Table 16: **Paired three-seed stability.** Mean$_{\pm\mathrm{std}}$ (seeds 2023–2025); gray = +DPR; bold = lower mean MSE. Δ MSE (%) = paired relative reduction with bootstrap 95% interval ($n = 3$, descriptive).

Backbone	Variant	ILI (24→24)	COVID19 (36→7)	VIX (96→96)	ETTh1 (96→96)
Informer	Base	7.192 ± 0.163 / 1.906 ± 0.033	1.920 ± 0.950 / 0.688 ± 0.257	1.071 ± 0.008 / 0.681 ± 0.015	1.642 ± 0.076 / 0.927 ± 0.032
	+DPR	**6.106 ± 0.760 / 1.756 ± 0.130**	**1.718 ± 0.673 / 0.631 ± 0.168**	**0.957 ± 0.033 / 0.662 ± 0.014**	**1.177 ± 0.100 / 0.804 ± 0.030**
	Δ MSE (%)	+15.2 [4.4,22.0]	+5.0 [-14.2,15.5]	+10.6 [6.3,12.9]	+28.3 [24.0,32.3]
PatchTST	Base	3.326 ± 0.275 / 1.072 ± 0.008	0.330 ± 0.005 / 0.217 ± 0.002	0.950 ± 0.008 / 0.543 ± 0.003	0.395 ± 0.001 / 0.393 ± 0.001
	+DPR	**3.052 ± 0.048 / 1.048 ± 0.008**	**0.327 ± 0.001 / 0.217 ± 0.000**	**0.942 ± 0.006 / 0.541 ± 0.004**	**0.392 ± 0.001 / 0.393 ± 0.002**
	Δ MSE (%)	+7.9 [2.2,14.5]	+1.0 [-0.2,2.4]	+0.9 [0.2,1.5]	+0.7 [0.3,1.3]
WPMixer	Base	3.042 ± 0.211 / 1.037 ± 0.015	0.333 ± 0.015 / 0.218 ± 0.001	0.969 ± 0.034 / 0.558 ± 0.021	0.380 ± 0.002 / 0.387 ± 0.001
	+DPR	**2.840 ± 0.059 / 1.049 ± 0.004**	**0.328 ± 0.009 / 0.218 ± 0.000**	**0.962 ± 0.021 / 0.554 ± 0.014**	**0.380 ± 0.001 / 0.387 ± 0.000**
	Δ MSE (%)	+6.4 [-0.7,11.9]	+1.4 [-5.9,7.3]	+0.6 [-3.3,3.2]	+0.0 [-0.4,0.3]
TimeFilter	Base	2.341 ± 0.304 / 0.908 ± 0.031	0.333 ± 0.005 / 0.222 ± 0.004	0.955 ± 0.004 / 0.551 ± 0.005	**0.389 ± 0.001** / 0.389 ± 0.001
	+DPR	**2.205 ± 0.337 / 0.900 ± 0.046**	**0.323 ± 0.004 / 0.219 ± 0.000**	**0.947 ± 0.001 / 0.547 ± 0.005**	0.389 ± 0.001 / 0.390 ± 0.001
	Δ MSE (%)	+6.0 [3.6,8.5]	+3.0 [0.7,6.0]	+0.8 [0.5,1.3]	-0.0 [-0.7,0.4]

Table 17: **Direct three-seed standalone comparison.** Mean$_{\pm\mathrm{std}}$ of four-horizon averages (seeds 2023–2025). Gray = DPRNet; bold/underline = best/second MSE / MAE per dataset.

Model	**MSE / MAE**			
	ILI	COVID19	VIX	Exchange
OLinear	$\underline{3.224}_{\pm 0.144}$ / $\mathbf{1.030}_{\pm 0.015}$	$\underline{0.854}_{\pm 0.046}$ / $\underline{0.375}_{\pm 0.010}$	$1.139_{\pm 0.031}$ / $0.698_{\pm 0.006}$	$0.466_{\pm 0.013}$ / $0.461_{\pm 0.004}$
TimeMixer	$3.379_{\pm 0.077}$ / $1.224_{\pm 0.021}$	$0.931_{\pm 0.031}$ / $0.401_{\pm 0.006}$	$\underline{1.120}_{\pm 0.010}$ / $\underline{0.688}_{\pm 0.001}$	$\mathbf{0.440}_{\pm 0.002}$ / $\mathbf{0.447}_{\pm 0.001}$
TimeMixer++	$3.587_{\pm 0.261}$ / $1.184_{\pm 0.006}$	$1.191_{\pm 0.061}$ / $0.478_{\pm 0.010}$	$1.152_{\pm 0.015}$ / $0.721_{\pm 0.007}$	$0.478_{\pm 0.007}$ / $0.472_{\pm 0.003}$
TimeBase	$8.864_{\pm 0.016}$ / $2.122_{\pm 0.005}$	$2.423_{\pm 0.243}$ / $0.777_{\pm 0.053}$	$1.283_{\pm 0.004}$ / $0.793_{\pm 0.002}$	$0.520_{\pm 0.004}$ / $0.507_{\pm 0.003}$
DPRNet	$\mathbf{2.887}_{\pm 0.045}$ / $\underline{1.070}_{\pm 0.002}$	$\mathbf{0.807}_{\pm 0.009}$ / $\mathbf{0.365}_{\pm 0.002}$	$\mathbf{1.114}_{\pm 0.022}$ / $\mathbf{0.685}_{\pm 0.006}$	$\underline{0.441}_{\pm 0.003}$ / $\underline{0.448}_{\pm 0.002}$

G.5 Additional Standalone Benchmarks

ECL, Traffic, and Weather use $L = 96$, horizons $\{96, 192, 336, 720\}$, and a 7:1:2 split; PEMS03/04/07/08 use OLinear's 6:2:2 split with $L = 96$ and horizons $\{12, 24, 48, 96\}$. Table 18 reports these seven configurations. DPRNet entries use three-run estimates when available; Traffic is a single-run aggregate at $L = 96$. The $L = 336$ Traffic diagnostic is excluded because its look-back differs. Weather also appears in the eight-benchmark matrix (main-text Table 3); the column here is a three-seed comparison against TimeMixer++, DLinear, and PatchTST rather than the eight-baseline average. Dataset configurations use hidden width 128 with patch/stride 8/4 for ECL, Traffic, and Weather, and width 256 with patch/stride 16/8 for PEMS. ECL and Traffic use batch size 128 and learning rate 5×10^{-4}; Weather uses batch size 256 and learning rate 10^{-3} with MAE training; PEMS uses batch size 128 and learning rate 10^{-3}. The longest PEMS horizons use larger patches or adjusted weight decay as specified in the released configurations.

Table 18: **Standalone generalization extension.** ECL/Traffic/Weather/PEMS; DPRNet is three-run mean$_{\pm\mathrm{std}}$ (Traffic single-run). Gray = DPRNet; bold/underline = best/second MSE / MAE.

Model	PEMS03	PEMS04	PEMS07	PEMS08	Weather	Traffic	ECL
DPRNet	**0.144** / 0.254 ±0.023 / ±0.013	**0.136** / 0.249 ±0.004 / ±0.002	$\underline{0.159}$ / 0.249 ±0.026 / ±0.015	$\underline{0.206}$ / 0.269 ±0.005 / ±0.004	$\underline{0.243}$ / 0.263 ±0.033 / ±0.017	0.521 / 0.326	$\underline{0.188}$ / 0.277 ±0.003 / ±0.002
TimeMixer++	$\underline{0.165}$ / 0.263	**0.136** / 0.251	**0.152** / 0.258	**0.200** / 0.279	**0.226** / 0.262	**0.416** / 0.264	**0.165** / 0.253
DLinear	0.278 / 0.375	0.295 / 0.388	0.329 / 0.395	0.379 / 0.416	0.265 / 0.317	0.625 / 0.383	0.212 / 0.300
PatchTST	0.180 / 0.291	0.195 / 0.307	0.211 / 0.303	0.280 / 0.321	0.259 / 0.281	0.531 / 0.343	0.208 / 0.295

Table 18 contains 28 model–dataset aggregates (7 datasets × 4 reported models) against TimeMixer++, DLinear, and PatchTST. DPRNet has the lowest MSE on PEMS03, ties TimeMixer++ on PEMS04, and ranks second on the remaining five datasets.