PROJECTION-CALIBRATED RESIDUAL ADAPTER FOR FROZEN TIME-SERIES FORECASTERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Time-series forecasters can leave systematic prediction errors that admit correction without modifying the frozen backbone. Effective correction requires estimating both the temporal shape of the residual and the strength with which it should be corrected. We introduce the *Projection-Calibrated Residual Adapter (PCRA)*, which supervises these factors separately. PCRA learns orthogonal temporal directions to represent complementary error patterns and uses analytical projection targets to supervise their signed strengths. For a fixed unit direction, the residual projection minimizes reconstruction error; orthogonality extends this result to decoupled coefficient targets. Implemented with lightweight MLPs, PCRA also supports test-time adaptation by updating the entire strength gate using fully observed historical outcomes. A conditional risk analysis characterizes when predictable residual energy captured by the directions exceeds coefficient prediction error. Static experiments on six supervised forecasters show a mean MSE reduction of 4.52% on five datasets; additional results demonstrate applicability to four frozen foundation models.

1 INTRODUCTION

Time-series forecasting informs critical decision making across energy, transportation, and finance, advancing from specialized deep architectures (Nie et al., 2023; Liu et al., 2024b) to pre-trained time-series foundation models (TSFMs) (Zhou et al., 2023; Goswami et al., 2024; Ekambaram et al., 2024). Despite their strong predictive capabilities, forecasters inevitably produce prediction errors that are not mere noise but often encode learnable temporal patterns and systematic biases (Liu et al., 2025b; Nguyen et al., 2026) (Figure 1a). Rather than retraining the entire model, a plug-in adapter can learn to compensate for these residuals and refine predictions while keeping the base forecaster frozen. Moreover, in practical deployment, learnable error patterns can be affected by distribution shifts over time (Kim et al., 2025), making adaptive residual correction essential.

Existing residual adapters typically learn the multi-step correction through a shared reconstruction objective (Liu et al., 2025b; Liang et al., 2026; Chen et al., 2024; Kim et al., 2025). Effective correction nevertheless requires distinguishing the error's temporal pattern from its magnitude: similar error shapes may demand different correction strengths, while a large residual does not justify a strong correction along a poorly aligned direction. Joint residual regression leaves these two aspects implicit. While the principle of vector decomposition is well established in weight normalization (Salimans & Kingma, 2016) and adaptation (Liu et al., 2024a), existing residual adapters still treat forecast errors as an entangled target. We instead introduce separate supervision to help the adapter learn temporal error patterns and calibrate their contribution to the forecast.

However, realizing this decoupled formulation presents several key challenges. First, decomposing prediction residuals into direction and magnitude inherently suffers from scaling ambiguity, requiring a principled mechanism to derive well-defined supervision targets. Second, forecast errors typically exhibit diverse temporal patterns; extending correction to multiple directions risks mutual interference unless the individual components remain strictly decoupled through an orthogonal basis. Finally, real-world temporal dynamics are non-stationary, meaning that test-time distribution shifts alter error behaviors and degrade static correction rules.

We propose the *Projection-Calibrated Residual Adapter (PCRA)*, which calibrates multi-step residual corrections by factorizing temporal error directions from their correction scale. Specifically,

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.

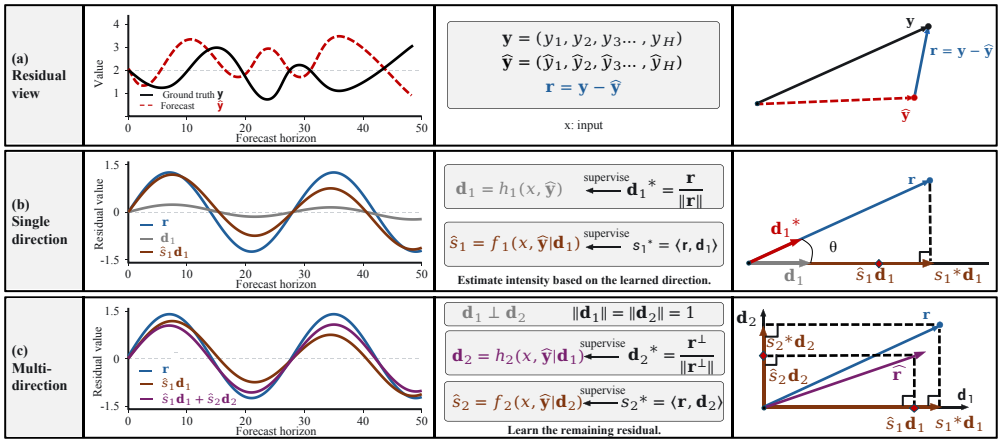


Figure 1: Direction–strength decomposition for residual correction. Forecast errors are represented as temporal residual vectors **(top)**. Unit-norm directions separate temporal shape from correction scale, enabling analytical projection targets for strength supervision **(middle)**. Orthogonal extension captures complementary residual structure while preserving decoupled coefficient targets **(bottom)**. The temporal and geometric views distinguish predicted corrections from their analytical targets.

PCRA identifies the optimal correction strength through residual projection, which inherently couples error magnitude with directional alignment. A strength predictor is then trained to estimate these projection targets on the learned directions (Figure 1b). To capture complementary error modes, PCRA expands this construction into an input-dependent orthogonal basis, in which a primary direction head captures the dominant error pattern and subsequent heads fit the remaining unexplained residuals along the orthogonal complement of preceding directions (Figure 1c). As a lightweight plug-in adapter for frozen forecasters, PCRA naturally supports test-time adaptation under distribution shifts: updating the entire strength gate accommodates non-stationary dynamics while preserving directional stability. We summarize our key contributions as follows:

- We formulate multi-step residual correction as decoupled direction and strength estimation, and propose PCRA to realize this principle by sequentially extracting an orthogonal temporal basis and calibrating correction strengths through analytical residual projections.
- We provide a prediction risk analysis that characterizes the tradeoff introduced by basis expansion: a correction improves forecasting risk when the predictable residual energy captured by the basis exceeds the error in predicting its strengths.
- We demonstrate that PCRA serves as an effective, lightweight plug-in adapter across both specialized deep forecasters and time-series foundation models (TSFMs) without backbone retraining. Furthermore, PCRA enables test-time adaptation under distribution shifts, which benefits long-term time-series forecasting in non-stationary environments.

2 Related Work

2.1 Post-forecast Correction

While normalization methods (Kim et al., 2022; Fan et al., 2023; Liu et al., 2023) and forecasting architectures (Liu et al., 2022; Lin et al., 2024; Li et al., 2026) mitigate distribution shifts and temporal structure internally within the model, post-forecast correction instead operates on the generated outputs to directly rectify residual errors. Among existing approaches, PIR (Liu et al., 2025b) retrieves historical patterns to rectify prediction failures, GLAFF (Wang et al., 2024) incorporates timestamp-derived global guidance, and UEC-STD (Nguyen et al., 2026) estimates structured errors in autoregressive forecasts to counteract progressive error accumulation.

The output-side Ada-Y variant of δ-Adapter (Liang et al., 2026) learns an additive residual correction for a frozen forecaster, relating error reduction to directional alignment and step size. Unlike classical signal pursuit and boosting where residuals are directly projected onto directional bases (Mallat & Zhang, 1993; Friedman, 2001), forecast correction must predict these projection coef-

ficients because future targets are unobserved at prediction time. PCRA advances this geometric perspective by decoupling the correction into normalized temporal directions and analytical projection targets. Orthogonality further ensures that multi-directional coefficients remain mutually decoupled, structuring supervision rather than relying on unconstrained residual regression.

2.2 Test-time Adaptation for Forecasting

Test-time adaptation aims to mitigate distribution shifts during inference, typically through unsupervised objectives such as self-supervision and entropy minimization (Sun et al., 2020; Wang et al., 2021). In time-series forecasting, this process is strictly constrained by temporal causality: future targets are unavailable at prediction time. Existing methods thus navigate this constraint under distinct feedback assumptions (Wang et al., 2026). SOLID (Chen et al., 2024) retrieves previously labeled historical contexts, while other frameworks adapt to unlabeled dynamics or propagate manifold errors (Dang et al., 2026; Liu et al., 2026). Alternatively, protocols in TAFAS (Kim et al., 2025) and PETSA (Medeiros et al., 2025) exploit partially observed target prefixes. The timing and latency of ground-truth feedback are therefore central to time-series adaptation.

Among these approaches, COSA (Im & Kwon, 2026) is a closely related output-space adapter that combines a linear residual correction with a learned gate and a context vector summarizing previously observed target values. PCRA similarly controls correction strength, but explicitly defines that strength relative to normalized temporal directions and learns it from analytical projection targets. While offline training uses training data, online adaptation updates the entire strength gate from fully observed historical horizons while keeping the direction predictors fixed.

3 Method

Given a historical context $\mathbf{X} \in \mathbb{R}^{L\times D}$, a frozen forecaster F (e.g., PatchTST or MOMENT) generates a base forecast $\widehat{\mathbf{Y}} = F(\mathbf{X}) \in \mathbb{R}^{H\times D}$, leaving an unobserved ground-truth residual $\mathbf{R} = \mathbf{Y} - \widehat{\mathbf{Y}}$. Without modifying F, PCRA learns a lightweight adapter A_θ to estimate this residual, producing the calibrated prediction:

$$\widetilde{\mathbf{Y}} = \widehat{\mathbf{Y}} + \widehat{\mathbf{R}}, \qquad \text{where} \quad \widehat{\mathbf{R}} = A_\theta(\mathbf{X}, \widehat{\mathbf{Y}}) \approx \mathbf{R}. \tag{1}$$

The adapter is fitted on historical training windows and uses only the pair $(\mathbf{X}, \widehat{\mathbf{Y}})$ available when a forecast is issued. However, direct residual regression entangles the temporal direction of the error with its correction strength. PCRA supervises these two aspects separately, with parameters shared across channels, as depicted in Figure 2.

3.1 Direction–Strength Decomposition

For a given channel with context $\mathbf{x} \in \mathbb{R}^L$, base forecast $\mathbf{b} \in \mathbb{R}^H$, and true residual $\mathbf{r} \in \mathbb{R}^H$, we first illustrate our approach along a single temporal direction: $\hat{\mathbf{r}} = \hat{s}\mathbf{d}$, with direction vector $\mathbf{d} \in \mathbb{R}^H$ and scalar correction strength $\hat{s} \in \mathbb{R}$. To eliminate the scaling ambiguity from $(q\mathbf{d})(\hat{s}/q) = \hat{\mathbf{r}}$ for any scalar $q \neq 0$, we constrain the direction to unit Euclidean norm (i.e., $\|\mathbf{d}\|_2 = 1$), allowing $\hat{s}$ to uniquely quantify the signed correction strength along $\mathbf{d}$. Given an observed training residual $\mathbf{r}$, the reconstruction loss $\ell(s) = \|\mathbf{r} - s\mathbf{d}\|_2^2$ is a scalar quadratic in s whose unique minimizer is the residual projection $s^* = \langle \mathbf{r}, \mathbf{d}\rangle / \|\mathbf{d}\|_2^2$, which for unit $\mathbf{d}$ simplifies to $s^* = \langle \mathbf{r}, \mathbf{d}\rangle = \|\mathbf{r}\|_2 \cos(\mathbf{r}, \mathbf{d})$.

Thus, the optimal strength depends on both residual magnitude and directional alignment: the residual norm alone does not account for how much of the error lies along the chosen direction. Subtracting this optimal correction leaves an unexplained residual $\mathbf{r}^\perp := \mathbf{r} - s^*\mathbf{d}$, which is orthogonal to the direction ($\langle \mathbf{r}^\perp, \mathbf{d}\rangle = 0$). By the Pythagorean theorem, the base squared error partitions as $\|\mathbf{r}\|_2^2 = (s^*)^2 + \|\mathbf{r}^\perp\|_2^2$, leaving an irreducible residual norm $\|\mathbf{r}^\perp\|_2^2 = \|\mathbf{r}\|_2^2 - \langle \mathbf{r}, \mathbf{d}\rangle^2$. At inference, the true residual is unobserved, and the adapter predicts $\hat{s}$ from observed inputs. Writing the test error as $\mathbf{r} - \hat{s}\mathbf{d} = \mathbf{r}^\perp + (s^* - \hat{s})\mathbf{d}$, its squared norm expands as:

$$\begin{aligned} \|\mathbf{r} - \hat{s}\mathbf{d}\|_2^2 = \|\mathbf{r}^\perp + (s^* - \hat{s})\mathbf{d}\|_2^2 &= \|\mathbf{r}^\perp\|_2^2 + 2(s^* - \hat{s})\underbrace{\langle \mathbf{r}^\perp, \mathbf{d}\rangle}_{=0} + (s^* - \hat{s})^2 \underbrace{\|\mathbf{d}\|_2^2}_{=1} \\ &= \|\mathbf{r}^\perp\|_2^2 + (\hat{s} - s^*)^2 = \|\mathbf{r}\|_2^2 - \langle \mathbf{r}, \mathbf{d}\rangle^2 + (\hat{s} - s^*)^2. \end{aligned} \tag{2}$$

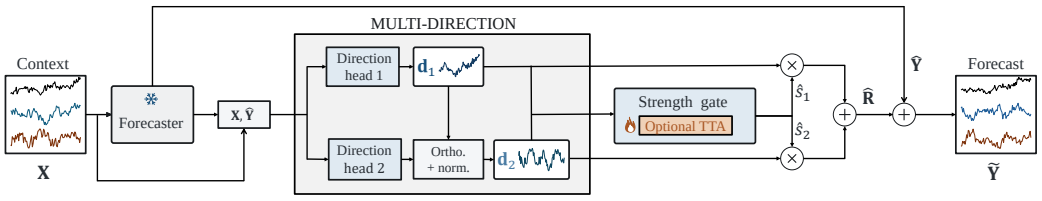


Figure 2: Overview of PCRA. Orthogonal direction experts and a strength gate produce an additive correction to the frozen forecaster. Optional TTA updates the entire strength gate from fully observed historical outcomes while keeping the direction experts fixed.

The decomposition separates residual energy orthogonal to the chosen direction, $\|\mathbf{r}^{\perp}\|_2^2$, from strength prediction error, $(\hat{s} - s^*)^2$. Observed training residuals provide the projection target s^*, which the strength gate learns to predict from inputs available at forecasting time.

3.2 Orthogonal Multi-Direction Correction

However, a single direction captures one residual mode, leaving complementary temporal structure unexplained. We represent the correction with K input-dependent directions. For channel c,

$$\widehat{\mathbf{r}}_c = \mathbf{D}_c \hat{\mathbf{s}}_c = \sum_{k=1}^{K} \hat{s}_{c,k} \mathbf{d}_{c,k}, \qquad \mathbf{D}_c = [\mathbf{d}_{c,1}, \ldots, \mathbf{d}_{c,K}] \in \mathbb{R}^{H \times K}. \tag{3}$$

For a fixed direction matrix, minimizing $\|\mathbf{r} - \mathbf{D}\mathbf{s}\|_2^2$ gives the normal equations $\mathbf{D}^\top \mathbf{D}\mathbf{s}^* = \mathbf{D}^\top \mathbf{r}$. The Gram matrix couples the coefficients when directions overlap. We therefore impose $\mathbf{D}^\top \mathbf{D} = I_K$, extending the scalar projection in Section 3.1 to decoupled coefficient targets.

Proposition (Orthogonal decoupling of strengths)**.** *For $\mathbf{D}^\top \mathbf{D} = I_K$, the unique least-squares coefficient vector is $\mathbf{s}^* = \mathbf{D}^\top \mathbf{r}$. For any predicted strengths $\hat{\mathbf{s}}$,*

$$\|\mathbf{r} - \mathbf{D}\hat{\mathbf{s}}\|_2^2 = \|(I - \mathbf{D}\mathbf{D}^\top)\mathbf{r}\|_2^2 + \|\hat{\mathbf{s}} - \mathbf{s}^*\|_2^2. \tag{4}$$

Under the orthogonality condition, the squared reconstruction error decomposes into two orthogonal contributions: residual energy in the complement of the direction span and squared error in predicting the projection coefficients. This is the multi-direction analogue of the single-direction decomposition in Eq. (2); the proof is given in Appendix F.4.

As shown in Figure 2, PCRA implements the direction head and strength gate with lightweight MLPs shared across channels. The direction head uses K separately parameterized predictors $f_{\psi_k} : \mathbb{R}^{L+H} \to \mathbb{R}^H$. Each receives $\mathbf{u}_c = [\mathcal{N}(\mathbf{x}_c); \mathcal{N}(\mathbf{b}_c)]$, where $\mathcal{N}(\mathbf{v}) = \mathbf{v}/\|\mathbf{v}\|_2$ for nonzero $\mathbf{v}$, and its output is orthogonalized and normalized:

$$\mathbf{d}_{c,k} = \mathcal{N}\left((I - P_{c,<k}) f_{\psi_k}(\mathbf{u}_c)\right), \qquad P_{c,<k} = \sum_{j<k} \mathbf{d}_{c,j}\mathbf{d}_{c,j}^\top, \quad P_{c,<1} = 0. \tag{5}$$

For linearly independent proposals with $K \leq H$, this Gram–Schmidt construction preserves their span and enforces orthogonality at both training and inference. The strength gate predicts $\hat{\mathbf{s}}_c = g_\phi(\mathbf{z}_c) \in \mathbb{R}^K$ from $\mathbf{z}_c = \Phi_c(\mathbf{x}_c, \mathbf{b}_c)$, which summarizes context and forecast statistics together with channel information. Its signed coefficients have no sum-to-one constraint and weight the directions in Eq. (3) to form the additive correction. We use $K = 2$ by default.

Projection defines the optimal coefficients for an observed residual, whereas forecasting requires them to be inferred before the future residual is available. We therefore characterize when the learned direction–strength predictor improves upon the frozen forecast. Let Z denote the information available at prediction time, with $\boldsymbol{\mu}(Z) = \mathbb{E}[\mathbf{r} \mid Z]$ and $\Sigma(Z) = \mathrm{Cov}(\mathbf{r} \mid Z)$. Assume finite second moments and that $\mathbf{D}_K$ and $\hat{\mathbf{s}}$ are functions of Z, with $\mathbf{D}_K^\top \mathbf{D}_K = I_K$. Applying the orthogonal decomposition to $\boldsymbol{\mu}(Z)$ yields the conditional risk

$$\mathcal{R}_K(Z) := \mathbb{E}[\|\mathbf{r} - \mathbf{D}_K\hat{\mathbf{s}}\|_2^2 \mid Z] = \mathrm{tr}\,\Sigma + \|(I - P_K)\boldsymbol{\mu}\|_2^2 + \|\hat{\mathbf{s}} - \mathbf{D}_K^\top \boldsymbol{\mu}\|_2^2, \tag{6}$$

where $P_K = \mathbf{D}_K \mathbf{D}_K^\top$. Subtracting this risk from the base risk $\mathcal{R}_0(Z) = \mathbb{E}[\|\mathbf{r}\|_2^2 \mid Z]$ cancels the shared conditional noise, giving

$$\mathbb{E}[\mathcal{R}_0(Z) - \mathcal{R}_K(Z)] = \underbrace{\mathbb{E}\|\mathbf{D}_K^\top \boldsymbol{\mu}\|_2^2}_{\text{captured predictable energy}} - \underbrace{\mathbb{E}\|\hat{\mathbf{s}} - \mathbf{D}_K^\top \boldsymbol{\mu}\|_2^2}_{\text{strength prediction error}}. \tag{7}$$

Eq. (7) characterizes the central tradeoff: a learned correction is beneficial when the predictable residual energy captured by the basis exceeds the error in predicting its strengths. In particular, append a unit direction $\mathbf{d}$ orthogonal to $\mathbf{D}_K$ while holding the existing directions and strengths fixed. Writing $a(Z) = \mathbf{d}^\top \boldsymbol{\mu}(Z)$ gives

$$\mathbb{E}[\mathcal{R}_{K+1}(Z) - \mathcal{R}_K(Z)] = \mathbb{E}[(\hat{s}_{\text{new}} - a)^2] - \mathbb{E}[a^2]. \tag{8}$$

The same tradeoff governs basis expansion: a new direction is useful when the predictable residual energy it captures exceeds the error in predicting its coefficient. This nested comparison does not determine the best K for independently retrained models. Full derivations and an analysis of projection-target noise are provided in Appendix F.

3.3 Training and Test-Time Adaptation

PCRA follows the direction–strength decomposition in its training procedure: direction predictors first learn complementary temporal structure, and the strength gate subsequently learns the associated projection coefficients. For the kth predictor, preceding directions are fixed and supervision is derived from the deflated residual $\mathbf{r}_c^{(k)} = (I - P_{c,<k})\mathbf{r}_c$. Writing $\bar{\mathbf{r}}_c^{(k)} = \mathcal{N}(\mathbf{r}_c^{(k)})$, the cosine objective and its population minimizer over admissible unit directions are

$$\mathcal{L}_{\text{dir},k} = \mathbb{E}\Big[1 - \mathbf{d}_{c,k}^\top \bar{\mathbf{r}}_c^{(k)}\Big], \qquad \mathbf{d}_k^*(\mathbf{u}_c) = \mathcal{N}\Big(\mathbb{E}[\bar{\mathbf{r}}_c^{(k)} \mid \mathbf{u}_c]\Big), \tag{9}$$

where the second expression applies when the conditional expectation is nonzero. For nonzero deflated residuals, the loss equals half the squared distance between unit vectors, giving each example equal weight irrespective of residual magnitude. Its population solution therefore identifies the conditional mean orientation of normalized residuals, separating temporal shape supervision from correction scale. The derivation is provided in Appendix G.2.

With the learned direction predictors fixed, observed training residuals determine the coefficient targets $\mathbf{s}_c^* = \mathbf{D}_c^\top \mathbf{r}_c$. The strength gate learns to infer these signed projections from features available at forecasting time, using an L1 objective on standardized targets. This stage calibrates the contribution of each temporal direction while preserving the basis that defines its supervision. The mapping g_ϕ includes the inverse standardization, so its outputs are expressed in residual units when forming the correction (Appendix G).

PCRA-T extends this formulation to delayed-feedback adaptation by updating the entire strength gate while retaining the frozen forecaster and direction predictors. Fully observed historical windows provide residuals $\mathbf{r}_m = \mathbf{y}_m - \mathbf{b}_m$ and renewed projection targets $\mathbf{s}_m^* = \mathbf{D}_m^\top \mathbf{r}_m$, yielding the online objective:

$$\mathcal{L}_{\text{TTA}}(\phi) = \frac{1}{M} \sum_{m=1}^{M} \|g_\phi(\mathbf{z}_m) - \mathbf{s}_m^*\|_2^2. \tag{10}$$

Orthogonality connects this coefficient objective directly to forecast accuracy: by Eq. (4), corrected-forecast squared error equals projection-coefficient squared error plus an orthogonal residual term independent of ϕ. Minimizing the online objective therefore minimizes the corresponding forecast error within the fixed direction span, adapting correction strengths without relearning temporal directions. The equivalence and feedback protocol are detailed in Appendix E.3.

4 Experiments

4.1 Experimental Setup

Datasets and backbones. We evaluate on seven real-world datasets: four ETT variants (ETTh1, ETTh2, ETTm1, ETTm2), Weather, Electricity, and Traffic, under standard chronological splits (Zhou et al., 2021; Wu et al., 2021). The backbones span six supervised forecasters and four zero-shot foundation models; the main text covers PatchTST (Nie et al., 2023), iTransformer (Liu

et al., 2024b), and CycleNet (Lin et al., 2024), alongside Chronos (Ansari et al., 2024), Moirai (Liu et al., 2025a), TimesFM (Das et al., 2024), and MOMENT-large (Goswami et al., 2024), with DLinear (Zeng et al., 2023), FreTS (Yi et al., 2023), and TQNet (Lin et al., 2025) deferred to Appendix B.

Baselines. We benchmark PCRA against both static and test-time adaptation (TTA) baselines: (i) *Static (non-TTA) methods*, which adjust predictions without test-time updates, including PIR (Liu et al., 2025b), Ada-Y (Liang et al., 2026), and UEC-STD (Nguyen et al., 2026); (ii) *Online TTA methods*, which adapt under streaming feedback, including TAFAS (Kim et al., 2025), PETSA (Medeiros et al., 2025), Ada-T (the test-time online variant of Ada-Y), and COSA (Im & Kwon, 2026), with DynaTTA (Grover & Etemad, 2025) and FAC (Wang et al., 2026) additionally evaluated on foundation models. All comparisons share identical frozen backbone forecasts and dataset splits.

PCRA implementation. PCRA is trained offline on the training data with the backbone completely frozen, computing normalization statistics strictly in-sample. By default, we employ two orthogonal residual directions, parameterizing direction predictors and the strength gate as lightweight MLPs with hidden widths of 64 and 32, respectively. For online adaptation, PCRA-T updates the strength gate under causally delayed feedback from fully observed historical windows, while direction predictors and the backbone remain fixed. Architecture specifications, training hyperparameters, and adaptation protocols are detailed in Appendix B.

4.2 Main Results

In this subsection, we evaluate PCRA in two complementary scenarios: static calibration and online adaptation under causally delayed feedback. We assess whether residual projection reliably enhances frozen forecasters on both supervised backbones and zero-shot foundation models (TSFMs). Performance is primarily measured by Mean Squared Error (MSE) and relative MSE reduction over frozen baselines, with Mean Absolute Error (MAE) additionally reported for TSFMs comparisons.

Table 1: Static adaptation of supervised forecasters (MSE). UEC: UEC-STD. Results average four horizons ($H \in \{96, 192, 336, 720\}$) on every dataset. For Traffic, PCRA uses the validation-selected correction-norm cap. Avg. is the unweighted mean over all seven datasets. **Bold**/underline mark the best/second-best values within each backbone at the displayed precision.

Dataset	PatchTST					iTransformer					CycleNet				
	Base	PIR	Ada-Y	UEC	PCRA	Base	PIR	Ada-Y	UEC	PCRA	Base	PIR	Ada-Y	UEC	PCRA
ETTh1	0.555	0.557	0.546	0.551	**0.528**	0.559	0.554	0.552	0.559	**0.530**	0.546	0.558	0.538	0.555	**0.523**
ETTh2	0.311	0.323	0.298	0.310	**0.292**	0.351	0.331	0.321	0.348	**0.312**	0.312	0.314	0.302	0.312	**0.291**
ETTm1	0.483	0.471	0.476	0.476	**0.462**	0.472	0.464	0.467	0.472	**0.458**	0.469	0.465	0.471	0.469	**0.461**
ETTm2	0.237	0.241	0.240	0.236	**0.233**	0.249	0.248	0.243	0.248	**0.235**	0.216	0.220	0.218	**0.214**	0.216
Weather	0.256	0.254	**0.240**	**0.240**	0.243	0.257	0.255	0.242	0.243	**0.241**	0.245	0.247	0.234	**0.231**	0.235
Electricity	0.209	**0.189**	0.192	0.200	0.195	0.189	**0.180**	0.184	0.190	0.182	0.168	0.176	0.163	0.184	**0.162**
Traffic	0.481	**0.451**	0.471	0.477	0.472	0.452	0.447	0.446	0.453	**0.442**	0.487	**0.452**	0.484	0.505	0.479
Avg.	0.362	0.355	0.352	0.356	**0.346**	0.361	0.354	0.351	0.359	**0.343**	0.349	0.347	0.344	0.353	**0.338**

Table 1 indicates that static PCRA consistently reduces forecasting error across all three supervised architectures, achieving the lowest MSE on 12 out of 15 ETT and Weather tasks. On ETTh1, for instance, PCRA lowers PatchTST error from 0.555 to 0.528, outperforming post-hoc baselines such as PIR and Ada-Y. The most prominent reduction occurs on ETTh2 under iTransformer, where MSE drops from 0.351 to 0.312. These improvements extend to high-dimensional Electricity, where PCRA achieves the best error of 0.162 on CycleNet. On Traffic, we report calibration with a validation-selected correction-norm cap: at $H = 336$, this capped setting reduces MSE from 0.483 to 0.473 on PatchTST, from 0.459 to 0.448 on iTransformer, and from 0.480 to 0.466 on CycleNet.

Beyond specialized supervised forecasters, we examine whether PCRA can calibrate zero-shot time-series foundation models (TSFMs) without altering their pretrained weights. As presented in Table 2, augmenting frozen backbones—Moirai (Liu et al., 2025a), TimesFM (Das et al., 2024), Chronos (Ansari et al., 2024), and MOMENT-large (Goswami et al., 2024)—yields an average relative MSE reduction of 25.90% (and 12.61% in MAE). The gains are particularly marked on non-stationary datasets such as ETTm1, where PCRA cuts average error by 48.06% (lowering Chronos

Table 2: **Zero-shot performance after applying our proposed PCRA to frozen time-series foundation models.** We report the mean MSE/MAE averaged over the four horizons ($H \in \{96, 192, 336, 720\}$). Avg. is the unweighted mean over the five displayed datasets. The best and second-best mean performances are highlighted in **bold** and underlined, respectively.

Dataset	Moirai		+PCRA		TimesFM		+PCRA		Chronos		+PCRA		MOMENT-large		+PCRA		Avg. Imp.	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	0.631	0.536	0.549	0.512	0.613	0.525	0.568	0.515	1.060	0.689	0.684	0.591	0.535	0.523	**0.510**	**0.508**	15.11%	5.85%
ETTh2	0.494	0.441	0.319	0.380	0.330	**0.375**	0.320	0.379	0.506	0.464	0.384	0.419	0.335	0.398	**0.315**	0.385	17.21%	6.40%
ETTm1	1.288	0.727	0.487	0.463	1.296	0.729	0.495	0.474	1.556	0.826	0.545	0.508	0.404	0.436	**0.391**	**0.421**	48.06%	28.39%
ETTm2	0.375	0.404	0.239	0.321	0.295	0.360	0.248	0.328	0.517	0.450	0.294	0.356	0.218	0.308	**0.211**	**0.300**	24.74%	13.21%
Weather	0.307	0.300	0.241	**0.265**	0.316	0.300	0.248	0.301	0.617	0.435	0.276	0.326	**0.229**	0.270	0.230	0.268	24.36%	9.18%
Avg.	0.619	0.482	0.367	0.388	0.570	0.458	0.376	0.399	0.851	0.573	0.437	0.440	0.344	0.387	**0.331**	**0.376**	25.90%	12.61%

MSE from 1.556 to 0.545). These results confirm that factorized residual projection effectively mitigates systematic bias in zero-shot forecasts.

While static calibration captures recurring error patterns, streaming time series often exhibit temporal distribution shifts that alter residual scales over time. We therefore evaluate online test-time adaptation (TTA) under causally delayed feedback, where PCRA-T dynamically updates its strength gate. Table 3 reports four-horizon averages on six datasets. On ETTh1, PCRA-T achieves the lowest error across all three backbones (e.g., 0.527 on PatchTST vs. 0.549 for TAFAS). Across these datasets, PCRA-T obtains mean MSEs of 0.324, 0.324, and 0.313 on PatchTST, iTransformer, and CycleNet, respectively—the lowest or tied-lowest mean on every backbone—although individual dataset results remain mixed.

Table 3: Online adaptation of supervised forecasters (MSE). Ada-T and PCRA-T use fully revealed horizons; other baselines retain native protocols (partially observed targets), so only Ada-T is feedback-matched. Each dataset averages $H \in \{96, 192, 336, 720\}$; Avg. is the unweighted six-dataset mean. **Bold**/underline mark the best/second-best per backbone at the displayed precision.

Dataset	PatchTST						iTransformer						CycleNet					
	Base	TAFAS	PETSA	Ada-T	COSA	PCRA-T	Base	TAFAS	PETSA	Ada-T	COSA	PCRA-T	Base	TAFAS	PETSA	Ada-T	COSA	PCRA-T
ETTh1	0.555	0.549	0.555	0.536	0.544	**0.527**	0.559	0.551	0.545	0.543	0.554	**0.527**	0.546	0.542	0.545	0.530	0.545	**0.523**
ETTh2	0.311	0.309	0.306	0.301	0.311	**0.292**	0.351	0.342	0.337	0.323	0.349	**0.311**	0.312	0.311	0.306	0.302	0.312	**0.292**
ETTm1	0.483	0.466	0.477	**0.452**	0.457	0.456	0.472	0.447	0.445	**0.441**	0.449	0.453	0.469	0.451	0.448	**0.441**	0.454	0.457
ETTm2	0.237	0.237	0.237	**0.229**	0.238	0.232	0.249	0.242	0.239	**0.233**	0.249	0.234	0.216	0.214	0.213	**0.208**	0.216	0.214
Weather	0.256	0.247	0.251	**0.240**	0.250	**0.240**	0.257	0.248	0.250	0.242	0.250	**0.238**	0.245	0.237	0.238	0.233	0.239	**0.231**
Electricity	0.209	0.203	0.209	**0.188**	0.209	0.194	0.189	0.189	0.189	0.183	0.189	**0.181**	0.168	0.168	0.167	**0.162**	0.168	**0.162**
Avg.	0.342	0.335	0.339	**0.324**	0.335	**0.324**	0.346	0.337	0.334	0.328	0.340	**0.324**	0.326	0.321	0.320	**0.313**	0.322	**0.313**

The performance of PCRA-T on supervised forecasters demonstrates the advantage of factorized adaptation under causally delayed feedback. By updating only the strength gate while freezing the directional basis, PCRA-T dynamically tracks non-stationary scale shifts without destabilizing the learned error geometry, avoiding the representation drift common in unrestricted online baselines. We next evaluate whether this lightweight calibration mechanism extends to frozen foundation models, where updating backbone parameters is computationally prohibitive.

Table 4: Online adaptation of foundation models (MSE). Each dataset averages $H \in \{96, 192, 336, 720\}$. Avg. is the unweighted mean over the six displayed datasets. TAFAS and PETSA use full-horizon feedback; other baselines retain native protocols. PCRA-T updates the full strength gate with projection MSE on fully revealed horizons. **Bold**/underline mark the best/second-best per model at the displayed precision.

Dataset	Chronos								TimesFM							
	Base	TAFAS	PETSA	DynaTTA	Ada-Y	COSA	FAC	PCRA-T	Base	TAFAS	PETSA	DynaTTA	Ada-Y	COSA	FAC	PCRA-T
ETTh1	1.060	0.970	1.008	1.002	1.055	0.972	1.043	**0.681**	0.613	0.625	0.634	0.612	0.611	0.623	0.611	**0.567**
ETTh2	0.506	0.512	0.502	0.501	0.504	0.516	0.506	**0.381**	0.330	0.343	0.337	0.327	0.331	0.354	0.330	**0.319**
ETTm1	1.556	1.037	1.134	1.140	1.545	1.034	1.397	**0.538**	1.296	0.845	0.857	0.878	1.286	0.910	1.131	**0.478**
ETTm2	0.517	0.478	0.454	0.482	0.513	0.515	0.502	**0.279**	0.295	0.290	0.284	0.284	0.293	0.344	0.290	**0.243**
Weather	0.617	0.441	0.429	0.442	0.579	0.507	0.587	**0.268**	0.316	0.286	0.295	0.285	0.305	0.355	0.307	**0.239**
Electricity	0.622	0.593	0.566	0.539	0.562	0.594	0.610	**0.285**	0.239	0.242	0.249	0.244	0.229	0.240	0.237	**0.202**
Avg.	0.813	0.672	0.682	0.684	0.793	0.690	0.774	**0.405**	0.515	0.438	0.443	0.438	0.509	0.471	0.484	**0.341**

Table 4 reports four-horizon averages on ETT, Weather, and Electricity. Across these six datasets, PCRA-T obtains mean MSEs of 0.405 on Chronos and 0.341 on TimesFM, compared with 0.672 and 0.438 for the strongest displayed baselines. Traffic is omitted because its $H = 720$ evaluations are incomplete; horizon-level results appear in Table 13.

Figure 3 examines how the benefit of residual calibration scales with the forecast horizon on frozen iTransformer. Static PCRA yields monotone gains, the relative MSE reduction rising from 3.3% at $H = 96$ to 9.1% at $H = 720$, indicating that cumulative multi-step drift leaves increasingly predictable residual structure at longer horizons. Online strength updates add a consistent increment at every horizon, reaching 9.7% at $H = 720$.

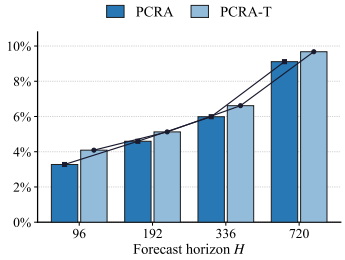


Figure 3: Relative MSE reduction (%) across horizons H.

In summary, experimental results demonstrate that factorized residual calibration consistently outperforms competing baselines across diverse real-world datasets. PCRA achieves consistent accuracy gains across supervised deep forecasters and zero-shot TSFMs, in both static calibration and online adaptation settings. We next analyze the underlying mechanisms of PCRA in Section 4.3, followed by ablation and sensitivity analyses in Section 4.4.

4.3 Multi-Direction Correction

To examine whether complementary directions yield forecast gains, we compare adapters with a single direction and two orthogonal directions (Section 3.2) via their *residual alignment*—the cosine similarity between the residual $\mathbf{r}$ and its projection $\mathbf{D}\mathbf{D}^\top\mathbf{r}$ onto the learned basis.

As shown in Figure 4a, two orthogonal directions consistently improve residual alignment across all five datasets. However, the resulting forecast gains (Figure 4b) are more modest and less uniform. This aligns with our conditional risk analysis in Eq. (7): expanding the basis captures more residual energy, but improves forecasts only when the corresponding projection strengths remain predictable from observable inputs (detailed in Appendix D.1).

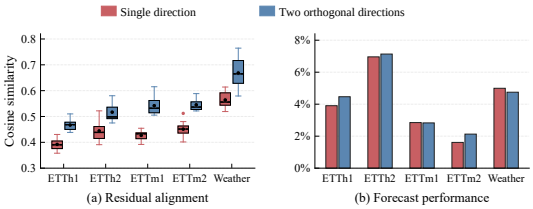
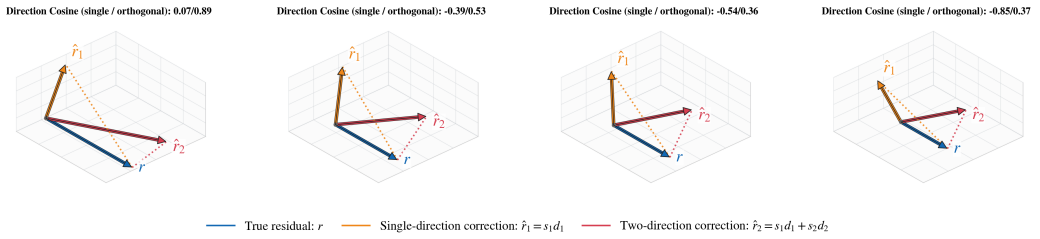


Figure 4: Comparison of residual alignment and forecast gains for single-/two-direction adapters.

Figure 5 illustrates the residual correction geometry on representative test windows from PatchTST, iTransformer, and CycleNet. Single-direction corrections frequently fail to align with the true residual $\mathbf{r}$, remaining near-orthogonal or even anti-aligned (with cosines dropping as low as -0.85). In contrast, two orthogonal directions consistently steer the correction toward the residual, establishing positive alignment in all cases (with cosines rising to 0.36–0.89). Each window is isometrically projected onto the 3D subspace spanned by the residual and correction vectors (Appendix D.4).



(a) PatchTST / ETTh2 / $H = 192$ (b) iTransformer / ETTh2 / $H = 336$ (c) CycleNet / ETTh1 / $H = 336$ (d) iTransformer / ETTh1 / $H = 192$

Figure 5: Geometry of residual corrections w.r.t. diverse backbones.

4.4 Ablation and Sensitivity Analysis

Key module ablations. To isolate each module's contribution, we evaluate four variants: *(i)* w/o factorization directly regresses residuals via MLP (Eq. 3); *(ii)* w/o input dependence fixes a static PCA basis (Eq. 5); and *(iii)* w/o cosine loss uses unnormalized squared projection (Eq. 9).

Table 5: Ablation results (% error reduction; best in **bold**).

Setting	ETTh1	ETTh2	ETTm1	ETTm2	Weather
PCRA	4.47%	**7.13%**	2.83%	**2.13%**	4.75%
w/o factorization	4.70%	1.74%	2.69%	-9.12%	**6.32%**
w/o input dependence	3.34%	4.45%	2.08%	1.27%	5.75%
w/o cosine loss	-1.36%	-9.82%	2.58%	0.75%	4.31%
w/o decoupled supervision	**6.58%**	1.95%	**3.18%**	-4.48%	6.22%

We evaluate all variants across ETT and Weather. To ensure a fair comparison, the direct MLP in *(i)* is parameter-matched to PCRA ($< 1\%$ difference) under identical inputs. As shown in Table 5, structural factorization provides crucial regularization: direct regression remains competitive on ETTh1 and Weather, but severely degrades on ETTm2 (-9.12%). Next, dynamic directions provide vital flexibility over a static PCA basis, enabling PCRA to lead on ETTh2, ETTm1, and ETTm2. Normalized alignment also proves consistently beneficial: PCRA outperforms squared-projection supervision across every dataset average, validating the decoupling of error shape from residual scale. Finally, staged supervision trades peak gains for reliability: joint reconstruction training attains the largest reductions on ETTh1 and ETTm1, yet forfeits most of the ETTh2 gain and degrades sharply on ETTm2 (-4.48%), so that its average reduction falls short of PCRA (2.69% vs. 4.26%), which remains positive on every dataset. Complete absolute MSE comparisons are detailed in Appendix C.

Direction count. We select $K=2$ by the highest mean validation performance on the complete four-horizon grid (Appendix D.1). Test gains from larger K are backbone-dependent: iTransformer peaks at $K=4$, while PatchTST and CycleNet improve more gradually (Figure 6a). Thus, the compact validation-selected default is not universally test-optimal.

Online parameter sensitivity. We examine the sensitivity of online strength adaptation to the learning rate, buffer size, and update interval (Figure 6b–d). Among these factors, the learning rate dominates, with overly aggressive updates nearly eliminating the adaptation gains, most markedly for CycleNet. Frequent updates are consistently beneficial, whereas the buffer size is largely inconsequential, indicating that timely incorporation of feedback outweighs its accumulation. Full experimental protocols appear in Appendix E.4, separate from the main PCRA-T evaluation.

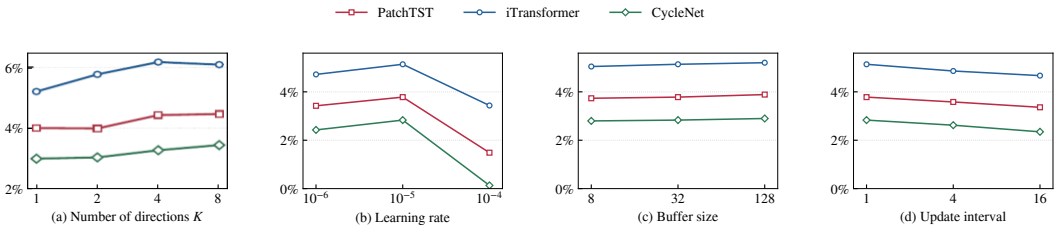


Figure 6: Sensitivity across backbones. (a) Test improvement versus K, averaged over ETT/Weather and four horizons. (b–d) Improvement of the full-gate projection-MSE protocol at $H=192$. Higher is better.

5 Conclusion and Discussion

We propose PCRA, a plug-in adapter that factorizes time-series residual correction into normalized temporal directions and dynamic projection strengths. By projecting prediction residuals onto unit directions, PCRA derives analytical strength targets that eliminate scale ambiguities, while orthogonal bases capture complementary error modes with decoupled supervision. Theoretically, our risk analysis formalizes the fundamental trade-off governing residual calibration, proving that net forecast risk is reduced whenever the predictable residual energy captured by the orthogonal basis exceeds strength estimation error.

Experimental results on diverse real-world datasets show that PCRA consistently outperforms competing baselines on both specialized deep forecasters and TSFMs, while effectively supporting online test-time adaptation under delayed feedback. Together, these findings establish structured residual projection as a principled, lightweight paradigm for time-series forecasting, demonstrating that frozen deep and foundation models can be reliably calibrated without the computational burden or representation drift of parameter retraining.

Limitations and future work. Despite its demonstrated effectiveness, PCRA remains constrained by the residual subspace spanned by the basis directions, where expanding the basis rank is beneficial only when captured predictable energy exceeds strength estimation noise. Furthermore, holding direction predictors fixed during test-time adaptation favors numerical stability, yet limits rapid adaptation to abrupt distribution shifts that introduce entirely new temporal error geometries. Developing online adapters that dynamically evolve directional bases while maintaining reliable, noise-resilient strength estimation constitutes a natural and promising future direction.

AI USE STATEMENT

Generative AI tools assisted with language editing, mathematical exposition and derivation checks, figure preparation, LaTeX formatting, and code or compilation checks. The authors are responsible for verifying all AI-assisted content and for the final manuscript.

REPRODUCIBILITY STATEMENT

The appendix documents the derivations, architecture, evaluation scope, and feedback protocols, and reports the complete static and online PCRA results. A provenance manifest and generation scripts reproduce every table and figure from full-precision records; checkpoint identities, split boundaries, and random seeds are logged with each result.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, 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 Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

Mouxiang Chen, Lefei Shen, Han Fu, Zhuo Li, Jianling Sun, and Chenghao Liu. Calibration of time-series forecasting: Detecting and adapting context-driven distribution shift. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024.

Hanyin Cheng, Xingjian Wu, Yang Shu, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. CoRA: Boosting time series foundation models for multivariate forecasting through correlation-aware adapter. In *The Fourteenth International Conference on Learning Representations*, 2026.

Ting Dang, Soumyajit Chatterjee, Hong Jia, Yu Wu, Flora Salim, and Fahim Kawsar. AdaNODEs: Test time adaptation for time series forecasting using neural ODEs. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2026.

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, 2024.

Vijay Ekambaram, Arindam Jati, Pankaj Dayama, Sumanta Mukherjee, Nam H. Nguyen, Wesley M. Gifford, Chandra K. Reddy, and Jayant Kalagnanam. Tiny time mixers (TTMs): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Wei Fan, Pengyang Wang, Dongkun Wang, Dongjie Wang, Yuanchun Zhou, and Yanjie Fu. Dish-TS: A general paradigm for alleviating distribution shift in time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7522–7529, 2023.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16115–16152, 2024.

Shivam Grover and Ali Etemad. Shift-aware test time adaptation and benchmarking for time-series forecasting. In *ICML 2025 Workshop on Test-Time Adaptation*, 2025.

Pengxin Guo, Pengrong Jin, Ziyue Li, Lei Bai, and Yu Zhang. Online test-time adaptation of spatial-temporal traffic flow forecasting. *arXiv preprint arXiv:2401.04148*, 2024.

Lu Han, Han-Jia Ye, and De-Chuan Zhan. SIN: Selective and interpretable normalization for long-term time series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024.

Jeonghwan Im and Hyuk-Yoon Kwon. COSA: Context-aware output-space adapter for test-time adaptation in time series forecasting. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=L7Z5wBMPrW.

HyunGi Kim, Siwon Kim, Jisoo Mok, and Sungroh Yoon. Battling the non-stationarity in time series forecasting via test-time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17868–17876, 2025.

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

Indar Kumar, Akanksha Tiwari, Sai Krishna Jasti, and Ankit Hemant Lade. RG-TTA: Regime-guided meta-control for test-time adaptation in streaming time series. *arXiv preprint arXiv:2603.27814*, 2026.

Thomas L. Lee, William Toner, Rajkarn Singh, Artjom Joosen, and Martin Asenov. Lightweight online adaption for time series foundation model forecasts. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 33736–33764, 2025.

Yujie Li, Zezhi Shao, Chengqing Yu, Yisong Fu, Tao Sun, Yongjun Xu, and Fei Wang. APT: Affine prototype-timestamp for time series forecasting under distribution shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15180–15188, 2026.

Daojun Liang, Qi Li, Yinglong Wang, Jing Chen, Hu Zhang, Xiaoxiao Cui, Qizheng Wang, and Shuo Li. The forecast after the forecast: A post-processing shift in time series. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=syfWdclGE1.

Shengsheng Lin, Weiwei Lin, Xinyi Hu, Wentai Wu, Ruichao Mo, and Haocheng Zhong. CycleNet: Enhancing time series forecasting through modeling periodic patterns. In *Advances in Neural Information Processing Systems*, volume 37, pp. 106315–106345, 2024.

Shengsheng Lin, Haojun Chen, Haijie Wu, Chunyun Qiu, and Weiwei Lin. Temporal query network for efficient multivariate time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37797–37814, 2025.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. Moirai 2.0: When less is more for time series forecasting. *arXiv preprint arXiv:2511.11698*, 2025a.

Jiaqi Liu, Yifan Ouyang, Zhifei Song, Sim Kuan Goh, and Ashwaq Qasem. STEPS: A temporal smooth error propagation solver on the manifolds for test-time adaptation in time series forecasting. *arXiv preprint arXiv:2605.08005*, 2026.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32100–32121. PMLR, 2024a. URL https://proceedings.mlr.press/v235/liu24bn.html.

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*, volume 35, 2022.

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

Zhiding Liu, Mingyue Cheng, Zhi Li, Zhenya Huang, Qi Liu, Yanhu Xie, and Enhong Chen. Adaptive normalization for non-stationary time series forecasting: A temporal slice perspective. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Zhiding Liu, Mingyue Cheng, Guanhao Zhao, Jiqian Yang, Qi Liu, and Enhong Chen. Improving time series forecasting via instance-aware post-hoc revision. In *Advances in Neural Information Processing Systems*, 2025b.

Sisuo Lyu, Siru Zhong, Tiegang Chen, Weilin Ruan, Qingxiang Liu, Taiqiang Lv, Qingsong Wen, Raymond Chi-Wing Wong, and Yuxuan Liang. TS-Memory: Plug-and-play memory for time series foundation models. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026. doi: 10.1145/3770855.3817614.

Stéphane G. Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993. doi: 10.1109/78.258082.

Heitor Rapela Medeiros, Hossein Sharifi-Noghabi, Gabriel L. Oliveira, and Saghar Irandoust. Accurate parameter-efficient test-time adaptation for time series forecasting. In *Second Workshop on Test-Time Adaptation: Putting Updates to the Test! at ICML 2025*, 2025. URL https://openreview.net/forum?id=uFj4EL4GTB.

Minh Hoang Nguyen, Van Dai Do, Huu Hiep Nguyen, Dung Nguyen, Kien Do, and Hung Le. Reviving error correction in modern deep time-series forecasting. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

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

Tim Salimans and Diederik P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/hash/ed265bc903a5a097f61d3ec064d96d2e-Abstract.html.

Giuseppe Soriano, Nicola Tonellotto, and Alberto Gotta. Black-Mamba: Biologically-inspired leaky accumulation for conceptual knowledge under distribution drift. *arXiv preprint arXiv:2607.18899*, 2026.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9229–9248, 2020.

Chengsen Wang, Qi Qi, Jingyu Wang, Haifeng Sun, Zirui Zhuang, Jinming Wu, and Jianxin Liao. Rethinking the power of timestamps for robust time series forecasting: A global-local fusion perspective. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Haochun Wang, Ruichen Xu, Georgios Kementzidis, Karen Cho, Sebastian Ramirez Villarreal, and Yuefan Deng. Towards principled test-time adaptation for time series forecasting. *arXiv preprint arXiv:2605.17250*, 2026.

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*, volume 34, pp. 22419–22430, 2021.

Yurui Wu, Qingying Deng, Wonou Chung, and Mairui Li. Test-time adaptation for non-stationary time series: From synthetic regime shifts to financial markets. *arXiv preprint arXiv:2602.00073*, 2026.

Weiwei Ye, Songgaojun Deng, Qiaosha Zou, and Ning Gui. Frequency adaptive normalization for non-stationary time series forecasting. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Kun Yi, Qi Zhang, Wei Fan, Shoujin Wang, Pengyang Wang, Hui He, Ning An, Defu Lian, Longbing Cao, and Zhendong Niu. Frequency-domain MLPs are more effective learners in time series forecasting. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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

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

Tian Zhou, Peisong Niu, Xue Wang, Liang Sun, and Rong Jin. One fits all: Power general time series analysis by pretrained LM. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Dataset	Variates	Time steps	Frequency	Domain
ETTh1	7	17,420	1 hour	Electricity
ETTh2	7	17,420	1 hour	Electricity
ETTm1	7	69,680	15 min	Electricity
ETTm2	7	69,680	15 min	Electricity
Weather	21	52,696	10 min	Weather
Electricity	321	26,304	1 hour	Electricity
Traffic	862	17,544	1 hour	Transportation

Table 6: Characteristics of the benchmark datasets used in our experiments.

A Experimental Setup and Reproducibility

A.1 Datasets and Splits

We evaluate PCRA on seven public multivariate long-term forecasting benchmarks following standard chronological protocols (Zhou et al., 2021; Wu et al., 2021). The ETT datasets record electricity-transformer measurements: ETTh1 and ETTh2 are sampled hourly, whereas ETTm1 and ETTm2 are sampled every 15 minutes. Weather contains meteorological measurements from the Max Planck Biogeochemistry Institute sampled every 10 minutes. Electricity and Traffic represent larger-scale, high-dimensional benchmarks: Electricity collects hourly electricity consumption of 321 customers, and Traffic records hourly road occupancy rates from 862 sensors. Table 6 summarizes their basic characteristics.

A.2 Backbones and Checkpoints

We evaluate PCRA across ten diverse frozen forecasters, encompassing modern supervised deep models and zero-shot foundation models:

Supervised Forecasters.

- **PatchTST** (Nie et al., 2023): A representative channel-independent Transformer that segments time series into subseries-level patches, capturing local semantic patterns while reducing quadratic attention complexity.
- **iTransformer** (Liu et al., 2024b): An inverted Transformer architecture that embeds the entire lookback sequence of each variable into an individual variate token to explicitly capture multivariate cross-channel correlations.
- **CycleNet** (Lin et al., 2024): A recurrent neural network designed with explicit cycle-trend modeling to accommodate periodic dynamics and non-stationarity in long-sequence forecasting.
- **DLinear** (Zeng et al., 2023): A simple yet competitive linear baseline that decomposes sequences into trend and seasonal components using moving average filters followed by one-layer linear projections.
- **FreTS** (Yi et al., 2023): A frequency-domain MLP-based architecture that transforms time series into the frequency domain to learn both inter-frequency and intra-frequency dependencies.
- **TQNet** (Lin et al., 2025): A state-of-the-art forecaster leveraging temporal-quantization mechanisms to model multiscale temporal representations.

Time-Series Foundation Models (TSFMs).

- **Chronos** (Ansari et al., 2024): A language-model-based time-series foundation model pretrained on massive synthetic and real-world time-series datasets using quantized tokenization and cross-entropy loss.
- **TimesFM** (Das et al., 2024): A decoder-only foundation model trained on diverse real-world time-series corpora, supporting zero-shot long-horizon multi-variate forecasting.
- **Moirai** (Liu et al., 2025a): A decoder-only time-series foundation model that adopts quantile forecasting and multi-token prediction for efficient zero-shot inference.
- **MOMENT-large** (Goswami et al., 2024): An open time-series foundation model pretrained via masked sequence modeling on a large public time-series corpus.

A.3 BASELINES

We compare PCRA against two distinct classes of forecasting adaptation methods:

Static (Non-TTA) Baselines. These methods calibrate frozen backbone forecasts without requiring test-time streaming labels:

- **PIR** (Liu et al., 2025b): Post-Inference Residual adapter that trains a lightweight revision network on source domains to directly predict residual adjustments for frozen base forecasts.
- **Ada-Y** (δ-Adapter) (Liang et al., 2026): A parameter-efficient adaptation paradigm that learns delta residual representations to calibrate target predictions.
- **UEC-STD** (Nguyen et al., 2026): Uncertainty-guided Error Correction with standardized calibration, updating predictions based on estimated residual uncertainty.

Online Test-Time Adaptation (TTA) Baselines. These methods dynamically update model parameters or adapter modules during inference using streaming feedback:

- **TAFAS** (Kim et al., 2025): A target-aware feedback adaptation system that dynamically adjusts forecasting representations under distribution shifts.
- **PETSA** (Medeiros et al., 2025): A parameter-efficient test-time adaptation framework that updates lightweight adapter modules online while keeping the backbone frozen.
- **Ada-T**: The online streaming test-time adaptation variant of Ada-Y, updating calibration parameters under delayed historical observations.
- **COSA** (Im & Kwon, 2026): Continual online selective adaptation that updates parameters under causal historical labels to prevent catastrophic forgetting.
- **DynaTTA** (Grover & Etemad, 2025): Dynamic test-time adaptation designed for time-series foundation models, leveraging frequency and statistical cues to adapt to non-stationary environments.
- **FAC** (Wang et al., 2026): Foundation-model adaptation under causal feedback constraints for streaming deployment.

A.4 TRAINING HYPERPARAMETERS AND MODEL SELECTION

Unless a table caption states otherwise, the direction predictors use hidden width 64 and one hidden layer, while the strength gate uses hidden width 32 and one hidden layer; both use dropout 0.1. The initial direction and strength stages use learning rate 10^{-3}, weight decay 10^{-4}, batch size 4,096, and at most 40 epochs. Checkpoints are selected using the validation objective of the corresponding stage, and all standardizers are fitted on the source training split only. The final reconstruction refinement starts from the selected stage-wise checkpoint with joint residual-MSE optimization at learning rate 10^{-5}. This offline refinement is separate from online adaptation.

A.5 VALIDATION AND MODEL SELECTION

Model and hyperparameter selection uses validation windows only; test windows are evaluated once after selection.

A.6 CORRECTION-NORM CAP

For Traffic, the correction-norm cap is a validation-selected safety control and is reported explicitly in the relevant captions. The cap is applied to the complete correction vector,

$$\Delta_{\text{cap}} = \Delta \min\left(1, \frac{\tau}{\|\Delta\|_2 + \epsilon}\right),$$

with τ selected on the validation split and then held fixed on test.

A.7 HARDWARE

We record the checkpoint identity, split boundaries, normalization statistics, selected epoch, device, and random seed with each result.

Offline jobs run on NVIDIA L20 and A800 GPUs. Hardware and timing boundaries for online measurements are specified in Appendix E.5.

B FULL EXPERIMENTAL RESULTS

The following tables expose the horizon-level records behind the aggregate comparisons. They retain the three main backbones and the additional supervised backbones reported in the appendix. A dash marks a horizon-level result that is not present in the verified archive; no partial horizon group is averaged or imputed. Electricity and Traffic are reported at all four standard horizons. For the three main backbones, the static Traffic results use the validation-selected correction-norm cap, consistently with Table 1. Their online Traffic results use the same validation-selected cap protocol as Table 3.

B.1 SUPERVISED STATIC RESULTS

Table 7: Horizon-resolved static MSE on all seven datasets (ETT, Weather, Electricity, and Traffic). Traffic PCRA uses the validation-selected correction-norm cap, consistently with Table 1. **Bold**/underline mark the best/second-best available displayed values within each backbone; – indicates an unavailable result.

Dataset	H	PatchTST					iTransformer					CycleNet				
		Base	PIR	Ada-Y	UEC-STD	PCRA	Base	PIR	Ada-Y	UEC-STD	PCRA	Base	PIR	Ada-Y	UEC-STD	PCRA
ETTh1	96	0.432	0.431	0.429	0.429	**0.422**	0.443	0.436	0.443	0.443	**0.431**	0.438	0.433	0.439	0.434	**0.428**
	192	0.494	0.489	0.491	0.494	**0.476**	0.508	0.506	0.507	0.508	**0.480**	0.497	0.518	0.492	0.505	**0.478**
	336	0.549	0.552	0.545	0.548	**0.528**	0.574	0.569	0.567	0.574	**0.540**	0.548	0.558	0.543	0.555	**0.527**
	720	0.745	0.757	0.718	0.735	**0.686**	0.712	0.704	0.691	0.710	**0.668**	0.700	0.723	0.678	0.724	**0.658**
ETTh2	96	0.235	0.236	**0.230**	0.234	0.232	0.258	0.251	**0.244**	0.253	0.245	0.236	0.239	0.237	0.238	**0.229**
	192	0.283	0.295	0.284	0.282	**0.274**	0.308	0.302	0.299	0.307	**0.288**	0.282	0.285	0.285	0.288	**0.273**
	336	0.315	0.332	0.315	0.315	**0.300**	0.372	0.344	0.338	0.370	**0.327**	0.332	0.326	0.320	0.323	**0.304**
	720	0.411	0.427	0.365	0.410	**0.361**	0.464	0.426	0.403	0.462	**0.388**	0.397	0.405	0.366	0.399	**0.358**
ETTm1	96	0.401	0.391	0.400	0.393	**0.379**	0.374	**0.364**	0.375	0.374	0.366	0.388	0.388	0.396	0.394	**0.385**
	192	0.443	0.439	0.445	0.443	**0.431**	0.426	0.421	0.428	0.426	**0.418**	0.436	0.431	0.445	0.433	**0.430**
	336	0.501	0.497	0.494	0.499	**0.489**	0.508	0.499	0.500	0.509	**0.496**	0.493	0.489	0.491	0.495	**0.485**
	720	0.586	0.557	0.566	0.570	**0.551**	0.580	0.571	0.564	0.580	**0.551**	0.557	0.554	0.551	0.555	**0.544**
ETTm2	96	0.158	0.161	0.160	0.158	**0.156**	0.165	**0.160**	0.164	0.165	0.161	0.147	0.151	0.149	0.148	**0.145**
	192	0.213	0.219	0.216	0.213	**0.208**	0.218	0.218	0.220	0.218	**0.212**	0.185	0.188	0.191	**0.183**	0.186
	336	0.248	0.257	0.256	**0.246**	0.250	0.260	0.263	0.256	0.258	**0.253**	0.229	0.228	0.232	**0.227**	0.233
	720	0.331	0.327	0.328	0.326	**0.319**	0.354	0.353	0.330	0.351	**0.315**	0.303	0.312	**0.299**	**0.299**	**0.299**
Weather	96	0.174	0.172	**0.162**	0.163	0.169	0.172	0.170	**0.161**	0.165	0.165	0.161	0.160	**0.154**	**0.154**	0.156
	192	0.221	0.219	**0.207**	0.212	0.211	0.221	0.219	**0.207**	0.209	**0.207**	0.209	0.209	**0.200**	**0.200**	0.202
	336	0.275	0.275	**0.260**	**0.260**	**0.260**	0.277	0.276	0.260	0.259	**0.258**	0.266	0.269	0.255	**0.251**	0.254
	720	0.353	0.352	0.331	**0.324**	0.333	0.358	0.356	0.339	0.339	**0.333**	0.344	0.351	0.325	**0.320**	0.329
Electricity	96	0.192	**0.164**	0.171	0.178	0.175	0.162	**0.155**	**0.155**	0.162	0.156	0.136	0.147	0.134	0.154	**0.133**
	192	0.186	**0.173**	0.176	0.182	0.179	0.176	**0.168**	0.172	0.176	0.170	0.153	0.161	**0.150**	0.169	**0.150**
	336	0.215	**0.189**	0.194	0.201	0.196	0.192	**0.183**	0.188	0.192	0.185	0.171	0.177	0.167	0.187	**0.166**
	720	0.244	0.231	**0.224**	0.238	0.231	0.228	**0.215**	0.222	0.228	0.216	0.213	0.221	0.200	0.226	**0.199**
Traffic	96	0.458	**0.424**	0.442	0.453	0.450	0.420	0.413	**0.410**	0.420	0.412	0.456	**0.429**	0.454	0.486	0.451
	192	0.467	**0.440**	0.458	0.465	0.459	0.442	0.436	0.436	0.442	**0.432**	0.467	**0.441**	0.464	0.490	0.462
	336	0.483	**0.454**	0.474	0.478	0.473	0.459	0.454	0.453	0.459	**0.448**	0.480	**0.455**	0.473	0.505	0.466
	720	0.516	**0.485**	0.510	0.512	0.506	0.489	0.483	0.486	0.489	**0.478**	0.544	**0.482**	0.544	0.539	0.538

Table 8: Additional-backbone, horizon-resolved static MSE on all seven datasets (ETT, Weather, Electricity, and Traffic). **Bold**/underline mark the best/second-best available displayed values within each backbone; – indicates an unavailable result.

Dataset	H	DLinear					FreTS					TQNet				
		Base	PIR	Ada-Y	UEC-STD	PCRA	Base	PIR	Ada-Y	UEC-STD	PCRA	Base	PIR	Ada-Y	UEC-STD	PCRA
ETTh1	96	0.459	0.440	0.456	0.448	**0.434**	0.444	0.439	0.443	0.438	**0.433**	0.433	0.433	0.438	0.434	**0.425**
	192	0.512	0.495	0.508	0.504	**0.481**	0.500	0.498	0.495	0.494	**0.482**	0.491	0.489	0.491	0.492	**0.478**
	336	0.558	0.552	0.552	0.548	**0.525**	0.564	0.558	0.556	0.564	**0.529**	0.546	0.581	0.540	0.568	**0.519**
	720	0.705	0.689	0.681	0.687	**0.677**	0.719	0.713	0.698	0.714	**0.667**	0.705	0.720	0.686	0.710	**0.660**
ETTh2	96	0.230	0.227	0.235	0.230	**0.226**	0.239	0.239	0.240	0.237	**0.234**	0.237	0.244	**0.229**	0.240	0.236
	192	0.284	0.282	0.289	0.284	**0.277**	0.289	0.292	0.291	0.288	**0.282**	0.283	0.289	0.278	0.280	**0.272**
	336	0.324	0.322	0.320	0.324	**0.313**	0.333	0.331	0.321	0.328	**0.306**	0.309	0.324	0.300	0.311	**0.290**
	720	0.410	0.406	0.379	0.389	**0.351**	0.411	0.406	0.373	0.410	**0.370**	0.408	0.394	0.375	0.383	**0.370**
ETTm1	96	0.371	0.378	0.362	0.374	**0.348**	0.367	0.375	0.369	0.369	**0.361**	0.371	0.385	0.381	0.381	**0.366**
	192	0.444	0.434	0.428	0.433	**0.417**	0.431	0.432	0.430	0.430	**0.423**	0.427	0.432	0.428	0.431	**0.421**
	336	0.518	0.493	0.497	0.504	**0.481**	0.497	0.492	0.485	0.495	**0.477**	0.492	0.494	0.493	0.502	**0.485**
	720	0.592	0.564	0.563	0.558	**0.538**	0.565	0.559	0.550	0.560	**0.536**	0.558	0.570	0.548	0.572	**0.543**
ETTm2	96	0.160	0.158	0.164	0.156	**0.154**	0.158	0.160	0.161	**0.155**	0.157	0.157	0.158	0.159	0.159	**0.155**
	192	0.193	0.197	0.205	0.191	**0.190**	0.193	0.197	0.201	**0.192**	0.193	0.192	0.194	0.201	**0.191**	0.192
	336	0.232	0.238	0.243	**0.229**	0.233	0.233	0.236	0.245	**0.231**	0.237	0.241	0.243	0.247	**0.233**	0.243
	720	0.306	0.321	0.313	**0.300**	0.310	0.302	0.316	0.305	**0.301**	0.305	0.302	0.314	0.303	0.301	**0.297**
Weather	96	0.195	**0.171**	0.179	0.172	0.172	0.183	0.171	0.169	**0.165**	0.169	0.158	0.158	**0.153**	**0.153**	0.155
	192	0.240	**0.217**	0.222	0.220	0.218	0.229	0.216	0.214	0.213	**0.211**	0.206	0.205	**0.198**	**0.198**	**0.198**
	336	0.292	0.285	0.274	**0.266**	0.267	0.283	0.281	0.267	**0.258**	0.261	0.263	0.263	0.252	**0.246**	0.250
	720	0.364	0.359	0.347	**0.329**	0.341	0.359	0.358	0.342	**0.325**	0.336	0.342	0.344	0.328	0.328	**0.325**
Electricity	96	0.208	**0.171**	0.219	0.205	0.179	0.181	**0.164**	0.169	0.182	0.176	0.138	0.149	**0.136**	0.151	0.135
	192	0.208	**0.178**	0.216	0.205	0.184	0.187	**0.174**	0.180	0.187	0.182	0.157	0.165	**0.154**	0.170	0.153
	336	0.223	**0.192**	0.213	0.220	0.201	0.203	**0.190**	0.195	0.204	0.197	0.172	0.189	**0.169**	0.202	0.168
	720	0.264	**0.235**	0.268	0.262	0.237	0.246	0.234	**0.233**	0.245	0.234	0.204	0.232	**0.202**	0.241	0.196
Traffic	96	0.647	**0.423**	0.652	0.616	0.517	0.485	**0.430**	0.468	0.484	0.479	0.422	0.421	0.419	0.432	**0.418**
	192	0.600	**0.440**	0.603	0.590	0.511	0.488	**0.443**	0.479	0.486	0.490	0.443	**0.440**	**0.440**	0.455	0.436
	336	0.607	**0.457**	0.577	0.598	0.529	0.502	**0.457**	0.493	0.499	0.509	0.458	0.459	**0.455**	0.471	0.461
	720	0.646	**0.485**	0.650	0.635	0.574	0.536	**0.486**	0.530	0.532	0.581	0.495	**0.487**	0.494	0.505	0.508

B.2 Foundation-Model Static Results

The following table reports the complete static transfer scope for the foundation-model checkpoints. It preserves the two-horizon records for the smaller models and the four-horizon records for Moirai2-R-small and MOMENT-large, together with the TS-Memory (Lyu et al., 2026) and CoRA (Cheng et al., 2026) cached-forecast controls.

Table 10 expands the main foundation-model comparison in Table 2 to the individual horizons. It reports the matched Base and PCRA MSE/MAE values for all four foundation models on the five datasets used in the main table; the aggregate improvements are recomputed independently for each horizon.

B.3 Supervised Online Results

Table 11: Horizon-resolved online MSE on all seven datasets (ETT, Weather, Electricity, and Traffic). All four horizons ($H \in \{96, 192, 336, 720\}$) are reported separately. Ada-T and PCRA-T use fully revealed horizons. For Traffic, PCRA-T uses the validation-selected correction-norm cap, consistently with Table 3. Native TTA entries use validation-only update/no-update selection; a no-update selection equals Base. **Bold**/underline mark the best/second-best available displayed values within each backbone; – indicates an unavailable result.

Dataset	H	PatchTST						iTransformer						CycleNet					
		Base	TAFAS	PETSA	Ada-T	COSA	PCRA-T	Base	TAFAS	PETSA	Ada-T	COSA	PCRA-T	Base	TAFAS	PETSA	Ada-T	COSA	PCRA-T
ETTh1	96	0.432	0.425	0.432	0.426	0.426	**0.421**	0.443	0.435	0.443	0.440	0.435	**0.429**	0.438	0.433	0.438	0.436	0.438	**0.428**
	192	0.494	0.485	0.494	0.483	0.487	**0.475**	0.508	0.494	0.508	0.490	0.498	**0.477**	0.497	0.492	0.497	0.480	0.494	**0.477**
	336	0.549	0.541	0.549	0.533	0.550	**0.527**	0.574	0.564	0.564	0.553	0.569	**0.535**	0.548	0.545	0.544	0.528	0.550	**0.526**
	720	0.745	0.745	0.745	0.701	0.712	**0.686**	0.712	0.712	**0.665**	0.691	0.712	0.668	0.700	0.700	0.700	0.678	0.700	**0.660**

Table 11 (continued)

Dataset	H	PatchTST						iTransformer						CycleNet					
		Base	TAFAS	PETSA	Ada-T	COSA	PCRA-T	Base	TAFAS	PETSA	Ada-T	COSA	PCRA-T	Base	TAFAS	PETSA	Ada-T	COSA	PCRA-T
ETTh2	96	0.235	0.233	0.235	**0.228**	0.235	0.232	0.258	0.253	0.253	**0.243**	0.258	0.244	0.236	0.236	0.236	0.232	0.236	**0.229**
	192	0.283	0.283	0.283	0.283	0.283	**0.274**	0.308	0.308	0.299	0.299	0.301	**0.288**	0.282	0.278	0.282	0.286	0.282	**0.273**
	336	0.315	0.311	0.315	0.319	0.315	**0.302**	0.372	0.344	0.372	0.347	0.372	**0.327**	0.332	0.332	0.332	0.324	0.332	**0.305**
	720	0.411	0.411	0.393	0.373	0.411	**0.359**	0.464	0.464	0.422	0.405	0.464	**0.387**	0.397	0.397	0.376	0.367	0.397	**0.360**
ETTm1	96	0.401	0.389	0.401	**0.368**	0.374	0.376	0.374	0.352	0.355	**0.347**	0.356	0.365	0.388	0.366	0.369	**0.365**	0.371	0.381
	192	0.443	0.433	0.443	**0.421**	0.423	0.429	0.426	0.410	0.412	**0.405**	0.410	0.415	0.436	0.419	0.420	**0.414**	0.419	0.428
	336	0.501	0.487	0.501	**0.475**	0.482	0.486	0.508	0.479	0.476	**0.475**	0.482	0.490	0.493	0.475	0.470	**0.464**	0.478	0.483
	720	0.586	0.556	0.562	0.542	0.548	**0.533**	0.580	0.545	0.538	**0.537**	0.549	0.544	0.557	0.542	0.532	**0.521**	0.547	0.538
ETTm2	96	0.158	0.158	0.158	**0.153**	0.158	0.156	0.165	0.162	0.163	**0.156**	0.165	0.159	0.147	0.146	0.146	**0.143**	0.147	0.145
	192	0.213	0.213	0.213	**0.207**	0.214	**0.207**	0.218	0.214	0.215	**0.207**	0.219	0.211	0.185	0.184	0.184	**0.182**	0.185	0.184
	336	0.248	0.248	0.248	**0.242**	0.248	0.248	0.260	0.252	**0.245**	0.246	0.258	0.253	0.229	0.228	0.227	**0.223**	0.229	0.230
	720	0.331	0.331	0.331	**0.313**	0.331	0.318	0.354	0.339	0.333	0.322	0.354	**0.314**	0.303	0.299	0.295	**0.286**	0.303	0.296
Weather	96	0.174	0.172	0.174	**0.161**	0.170	0.166	0.172	0.167	0.168	**0.161**	0.168	0.162	0.161	0.157	0.158	**0.154**	0.158	**0.154**
	192	0.221	0.216	0.220	**0.207**	0.215	**0.207**	0.221	0.214	0.214	0.207	0.215	**0.205**	0.209	0.202	0.203	0.200	0.204	**0.199**
	336	0.275	0.265	0.270	0.260	0.266	**0.257**	0.277	0.266	0.265	0.260	0.267	**0.255**	0.266	0.256	0.255	0.253	0.258	**0.250**
	720	0.353	0.337	0.342	0.332	0.348	**0.328**	0.358	0.346	0.353	0.339	0.350	**0.329**	0.344	0.332	0.336	0.326	0.338	**0.321**
Electricity	96	0.192	0.185	0.192	**0.169**	0.191	0.175	0.162	0.161	0.162	**0.155**	0.161	**0.155**	0.136	0.135	0.136	0.135	0.136	**0.133**
	192	0.186	0.183	0.186	**0.174**	0.186	0.178	0.176	0.175	0.176	0.170	0.176	**0.169**	0.153	0.152	0.152	**0.149**	0.153	0.150
	336	0.215	0.206	0.215	**0.190**	0.215	0.195	0.192	0.191	0.192	0.185	0.192	**0.184**	0.171	0.171	0.170	**0.163**	0.171	0.165
	720	0.244	0.238	0.245	**0.221**	0.245	0.230	0.228	0.228	0.228	0.221	0.228	**0.215**	0.213	0.213	0.213	0.200	0.214	**0.199**
Traffic	96	0.458	0.458	0.458	**0.442**	0.458	0.451	0.420	0.420	0.420	**0.410**	0.420	0.413	0.456	**0.449**	0.455	0.454	0.452	0.451
	192	0.467	0.467	0.467	**0.460**	0.467	**0.460**	0.442	0.442	0.442	0.436	0.442	**0.433**	0.467	**0.460**	0.465	0.464	0.463	0.462
	336	0.483	0.483	0.483	0.474	0.483	**0.473**	0.459	0.459	0.459	0.453	0.459	**0.449**	0.480	0.480	0.480	0.473	0.480	**0.466**
	720	0.516	0.517	0.516	0.510	0.516	**0.505**	0.489	0.489	0.489	0.486	0.489	**0.478**	0.544	0.544	0.544	0.544	0.544	**0.538**

Table 12: Additional-backbone, horizon-resolved online MSE on all seven datasets (ETT, Weather, Electricity, and Traffic). Native TTA entries use validation-only update/no-update selection; a no-update selection equals Base. **Bold**/underline mark the best/second-best available displayed values within each backbone; – indicates an unavailable result.

Dataset	H	DLinear						FreTS						TQNet					
		Base	TAFAS	PETSA	Ada-T	COSA	PCRA-T	Base	TAFAS	PETSA	Ada-T	COSA	PCRA-T	Base	TAFAS	PETSA	Ada-T	COSA	PCRA-T
ETTh1	96	0.459	0.453	0.449	0.452	0.451	**0.433**	0.444	0.439	0.444	0.440	0.444	**0.432**	0.433	0.433	0.433	0.434	0.433	**0.424**
	192	0.512	0.504	0.504	0.492	0.505	**0.479**	0.500	0.495	0.500	**0.481**	0.496	**0.481**	0.491	0.482	0.491	0.480	0.484	**0.474**
	336	0.558	0.553	0.551	0.539	0.558	**0.522**	0.564	0.558	0.557	0.542	0.563	**0.525**	0.546	0.540	0.540	0.527	0.546	**0.516**
	720	0.705	0.705	**0.667**	0.670	0.705	0.678	0.719	0.719	0.678	0.683	0.719	**0.666**	0.705	0.705	0.663	0.675	0.705	**0.659**
ETTh2	96	0.230	0.230	0.230	0.229	0.230	**0.227**	0.239	0.239	0.239	0.236	0.239	**0.234**	0.237	0.234	0.237	**0.229**	0.237	0.236
	192	0.284	0.284	0.284	0.287	0.284	**0.276**	0.289	0.286	0.286	0.288	0.289	**0.282**	0.283	0.277	0.283	0.281	0.283	**0.272**
	336	0.324	0.324	0.324	0.326	0.324	**0.313**	0.333	0.333	0.333	0.328	0.333	**0.307**	0.309	0.309	0.309	0.305	0.309	**0.291**
	720	0.410	0.410	0.410	0.380	0.410	**0.352**	0.411	0.411	0.386	0.376	0.411	**0.370**	0.408	0.408	0.385	0.377	0.408	**0.372**
ETTm1	96	0.371	0.349	0.352	**0.346**	0.348	**0.346**	0.367	0.357	0.358	**0.348**	0.358	0.359	0.371	**0.350**	0.371	0.351	0.359	0.365
	192	0.444	0.417	0.418	**0.415**	0.417	**0.415**	0.431	0.421	0.419	**0.410**	0.421	0.421	0.427	0.413	0.417	**0.407**	0.414	0.420
	336	0.518	0.483	0.481	0.482	0.483	**0.477**	0.497	0.483	0.477	**0.467**	0.497	0.475	0.492	0.469	**0.468**	0.471	0.471	0.481
	720	0.592	0.556	0.549	0.543	0.559	**0.532**	0.565	0.552	0.541	**0.526**	0.554	0.530	0.558	0.536	0.530	**0.523**	0.540	0.540
ETTm2	96	0.160	0.159	0.159	**0.150**	0.160	0.154	0.158	0.157	0.157	**0.152**	0.158	0.156	0.157	0.157	0.157	**0.152**	0.157	0.154
	192	0.193	0.192	0.192	**0.184**	0.193	0.189	0.193	0.192	0.192	**0.189**	0.193	0.192	0.192	0.190	0.193	**0.189**	0.192	0.191
	336	0.232	0.230	0.230	**0.225**	0.232	0.231	0.233	0.231	0.231	**0.227**	0.233	0.234	0.241	0.238	0.238	**0.231**	0.241	0.242
	720	0.306	0.299	0.296	**0.291**	0.306	0.300	0.302	0.297	0.294	**0.288**	0.302	0.299	0.302	0.295	0.294	**0.287**	0.302	0.295
Weather	96	0.195	0.188	0.183	0.177	0.190	**0.170**	0.183	0.176	0.176	0.169	0.179	**0.167**	0.158	0.154	0.158	**0.153**	0.155	**0.153**
	192	0.240	0.229	0.229	0.222	0.231	**0.214**	0.229	0.218	0.218	0.214	0.222	**0.207**	0.206	0.199	0.206	0.198	0.201	**0.195**
	336	0.292	0.277	0.276	0.274	0.279	**0.262**	0.283	0.269	0.269	0.266	0.272	**0.258**	0.263	0.253	0.252	0.250	0.255	**0.246**
	720	0.364	0.350	0.350	0.343	0.357	**0.331**	0.359	0.347	0.350	0.339	0.353	**0.327**	0.342	0.329	0.338	0.329	0.336	**0.320**
Electricity	96	0.208	0.206	0.208	**0.187**	0.207	0.184	0.181	0.178	0.178	**0.168**	0.180	0.175	0.138	0.137	0.138	**0.136**	0.138	0.135
	192	0.208	0.207	0.205	**0.193**	0.208	0.187	0.187	0.185	0.184	**0.176**	0.187	0.182	0.157	0.157	0.157	**0.154**	0.157	0.153
	336	0.223	0.222	0.220	**0.206**	0.223	0.201	0.203	0.202	0.201	**0.189**	0.203	0.197	0.172	0.172	0.172	**0.169**	0.172	0.167
	720	0.264	0.264	0.265	**0.243**	0.266	0.237	0.246	0.245	0.246	**0.227**	0.247	0.234	0.204	0.204	0.204	**0.200**	0.206	0.196

Table 12 (continued)

Dataset	H	DLinear						FreTS						TQNet					
		Base	TAFAS	PETSA	Ada-T	COSA	PCRA-T	Base	TAFAS	PETSA	Ada-T	COSA	PCRA-T	Base	TAFAS	PETSA	Ada-T	COSA	PCRA-T
Traffic	96	0.647	0.647	0.647	**0.597**	0.647	0.517	0.485	0.485	0.485	**0.468**	0.485	0.477	0.422	**0.411**	0.422	0.419	0.422	0.418
	192	0.600	0.600	0.600	**0.566**	0.600	0.511	0.488	0.488	0.488	**0.480**	0.488	0.487	0.443	**0.437**	0.443	0.440	0.443	0.436
	336	0.607	0.607	0.607	**0.575**	0.607	0.527	0.502	0.502	0.502	**0.493**	0.502	0.505	0.458	0.458	0.458	0.455	**0.454**	0.458
	720	0.646	0.646	0.646	0.616	0.646	**0.571**	0.536	0.536	0.536	**0.530**	0.536	0.575	0.495	0.495	0.495	**0.494**	0.495	0.505

B.4 Foundation-Model Online Results

Table 13 provides the individual horizons behind the foundation-model online comparison. Traffic is retained here but excluded from the six-dataset aggregate in Table 4 because its $H = 720$ coverage is incomplete. Online adaptation is not uniformly beneficial: over $H \in \{96, 192, 336\}$, TimesFM/Traffic has mean MSE 0.632 with PCRA-T, compared with 0.589 for the frozen backbone and 0.573 for Ada-Y.

Table 13: Horizon-resolved online adaptation of foundation models (MSE). The six complete datasets underlie Table 4; available Traffic results are retained here, with – denoting missing results. Ada-Y retains its original online protocol; PCRA-T updates the full strength gate using projection MSE and fully revealed horizons. **Bold**/underline mark the best/second-best available values within each model at the displayed precision.

Dataset	H	Chronos								TimesFM							
		Base	TAFAS	PETSA	DynaTTA	Ada-Y	COSA	FAC	PCRA-T	Base	TAFAS	PETSA	DynaTTA	Ada-Y	COSA	FAC	PCRA-T
ETTh1	96	0.834	0.750	0.755	0.783	0.829	0.746	0.817	**0.522**	0.500	0.491	0.481	0.497	0.499	0.489	0.499	**0.458**
	192	0.982	0.880	0.899	0.921	0.978	0.886	0.957	**0.602**	0.573	0.567	0.571	0.572	0.572	0.571	0.571	**0.531**
	336	1.073	0.958	1.018	1.033	1.069	1.014	1.055	**0.723**	0.612	0.607	0.680	0.624	0.610	0.640	0.610	**0.561**
	720	1.349	1.293	1.360	1.271	1.344	1.244	1.344	**0.878**	0.766	0.837	0.803	0.755	0.762	0.793	0.765	**0.717**
ETTh2	96	0.381	0.378	0.375	0.380	0.380	0.368	0.381	**0.284**	0.251	0.249	0.251	0.249	0.251	0.250	0.251	**0.248**
	192	0.470	0.462	0.464	0.467	0.468	0.452	0.470	**0.345**	0.309	0.313	0.312	0.310	0.309	0.315	0.309	**0.301**
	336	0.528	0.527	0.522	0.525	0.525	0.542	0.527	**0.368**	0.343	0.349	0.352	0.342	0.344	0.378	0.343	**0.323**
	720	0.646	0.681	0.646	0.632	0.643	0.700	0.645	**0.528**	0.419	0.460	0.432	0.408	0.420	0.471	0.419	**0.403**
ETTm1	96	1.447	0.921	0.970	1.012	1.436	0.854	1.279	**0.450**	1.241	0.759	0.761	0.787	1.231	0.735	1.062	**0.383**
	192	1.511	0.979	1.056	1.096	1.499	0.954	1.357	**0.501**	1.311	0.812	0.827	0.852	1.301	0.842	1.144	**0.446**
	336	1.573	1.053	1.189	1.175	1.562	1.075	1.402	**0.559**	1.282	0.848	0.875	0.894	1.273	0.972	1.117	**0.512**
	720	1.694	1.196	1.323	1.279	1.684	1.254	1.549	**0.643**	1.348	0.963	0.966	0.977	1.340	1.091	1.202	**0.571**
ETTm2	96	0.351	0.328	0.309	0.335	0.348	0.313	0.342	**0.199**	0.225	0.218	0.215	0.218	0.223	0.221	0.222	**0.165**
	192	0.450	0.417	0.399	0.429	0.446	0.430	0.437	**0.254**	0.264	0.257	0.253	0.254	0.263	0.278	0.260	**0.216**
	336	0.534	0.493	0.476	0.499	0.531	0.568	0.517	**0.291**	0.304	0.303	0.295	0.292	0.302	0.367	0.298	**0.256**
	720	0.733	0.673	0.634	0.665	0.728	0.748	0.711	**0.373**	0.385	0.383	0.372	0.373	0.383	0.508	0.377	**0.335**
Weather	96	0.388	0.275	0.279	0.306	0.361	0.275	0.376	**0.193**	0.207	0.191	0.200	0.192	0.198	0.190	0.204	**0.164**
	192	0.510	0.363	0.351	0.386	0.479	0.417	0.487	**0.233**	0.264	0.243	0.246	0.240	0.256	0.270	0.257	**0.204**
	336	0.640	0.478	0.454	0.464	0.599	0.551	0.603	**0.283**	0.337	0.304	0.307	0.299	0.327	0.425	0.323	**0.256**
	720	0.929	0.648	0.633	0.613	0.875	0.784	0.883	**0.362**	0.454	0.408	0.427	0.407	0.437	0.533	0.442	**0.331**
Electricity	96	0.475	0.434	0.429	0.415	0.430	0.438	0.465	**0.241**	0.223	0.215	0.214	0.211	0.215	0.216	0.220	**0.184**
	192	0.542	0.515	0.493	0.474	0.489	0.517	0.530	**0.256**	0.220	0.221	0.215	0.220	0.212	0.218	0.218	**0.187**
	336	0.654	0.626	0.577	0.557	0.590	0.628	0.640	**0.289**	0.236	0.240	0.236	0.241	0.226	0.239	0.233	**0.200**
	720	0.816	0.798	0.767	0.712	0.739	0.794	0.806	**0.352**	0.278	0.291	0.329	0.306	0.263	0.285	0.276	**0.237**
Traffic	96	1.735	1.517	1.508	1.451	1.637	1.584	1.663	**0.687**	0.602	0.600	0.598	0.595	0.582	0.599	0.601	**0.539**
	192	1.761	1.586	1.575	1.465	1.659	1.634	1.698	**0.699**	0.575	0.573	0.572	0.568	0.560	0.573	0.574	**0.527**
	336	1.886	1.749	1.719	1.566	1.773	1.794	1.828	**0.858**	0.592	0.591	0.592	0.586	**0.577**	0.591	0.591	0.829
	720	2.044	1.960	1.915	**1.746**	1.920	1.979	2.017	–	–	–	–	–	–	–	–	–

C Controlled Ablations

C.1 Structural Controls

The direct residual MLP receives the union of PCRA's available inputs: normalized context, normalized base forecast, and 22 statistical/channel features. Its width matches the adapter parameter count within 1%, without test-based capacity selection. We evaluate the unshrunk correction and select checkpoints on validation loss. All 60 controlled tasks share the same frozen predictions and data splits. Table 14 reports the full grid, including results unfavorable to PCRA.

Figure 7 reports the same matched tasks as Table 5, but on the absolute MSE scale. The controls replace one structural choice at a time: cosine direction supervision, input-dependent directions, or the direction–strength factorization. Relative ablation values divide each task's MSE by the corresponding PCRA MSE before averaging, matching the performance definition in the main text. The in-sample target control changes the mean relative MSE by only 0.04% and is therefore retained as a training-order check rather than a separate component of the method.

		TTM-R2		Bolt-tiny		Bolt-mini		MOMENT-small		Moirai2-R-small				MOMENT-large			
Dataset	Metric	Base	PCRA	Base	PCRA	Base	PCRA	Base	PCRA	Base	TS-Memory	CoRA	PCRA	Base	TS-Memory	CoRA	PCRA
ETTh1	MSE	0.483	0.465	0.544	0.495	0.550	0.499	0.526	0.485	0.631	0.597	0.598	**0.549**	0.535	**0.508**	0.525	0.510
	MAE	0.488	0.478	0.490	0.483	0.488	0.482	0.516	0.487	0.536	0.539	0.531	**0.512**	0.523	0.511	0.517	**0.509**
ETTh2	MSE	0.286	0.285	0.308	0.281	0.307	0.276	0.291	0.272	0.494	0.348	0.388	**0.319**	0.335	0.331	0.331	**0.315**
	MAE	0.358	0.365	0.363	0.353	0.361	0.351	0.366	0.353	0.441	0.394	0.407	**0.380**	0.398	0.396	0.394	**0.385**
ETTm1	MSE	0.410	0.372	1.324	0.457	1.343	0.459	0.459	0.436	1.288	0.711	0.707	**0.487**	0.404	0.397	0.392	**0.391**
	MAE	0.419	0.400	0.674	0.441	0.682	0.441	0.457	0.429	0.727	0.573	0.551	**0.463**	0.436	0.430	0.425	**0.421**
ETTm2	MSE	0.195	0.186	0.255	0.208	0.258	0.206	0.223	0.206	0.375	0.252	0.266	**0.239**	0.218	0.211	**0.210**	0.211
	MAE	0.290	0.281	0.339	0.299	0.340	0.298	0.316	0.298	0.404	0.335	0.341	**0.321**	0.308	0.303	0.303	**0.300**
Weather	MSE	0.203	0.205	0.338	0.214	0.347	0.213	0.225	0.210	0.307	0.260	0.281	**0.241**	0.229	0.230	**0.225**	0.230
	MAE	0.241	0.242	0.300	0.246	0.301	0.245	0.273	0.247	0.300	0.285	0.294	**0.265**	0.270	0.269	**0.263**	0.268

Table 9: **Static transfer to frozen foundation models.** MSE/MAE are averaged over $H \in \{96, 336\}$ for TTM-R2, Bolt-tiny, Bolt-mini, and MOMENT-small, and over $H \in \{96, 192, 336, 720\}$ for Moirai2-R-small and MOMENT-large. Bolt, MOMENT, and Moirai use fixed $Q = 16$; all other choices are selected on validation. PCRA is trained on source labels and the foundation-model checkpoints remain frozen. For both Moirai2-R-small and MOMENT-large, TS-Memory and CoRA use their official implementations adapted to the same cached forecasts as Base and PCRA. Bold marks the lowest result within each backbone and metric.

Table 10: **Horizon-level zero-shot performance after applying our proposed PCRA to frozen time-series foundation models.** Each row reports one forecasting horizon. Avg. Imp. is the unweighted mean relative reduction over the four backbones at that horizon. The best and second-best horizon-level performances are highlighted in **bold** and underlined, respectively.

Dataset	H	**Moirai**		**+PCRA**		**TimesFM**		**+PCRA**		**Chronos**		**+PCRA**		**MOMENT-large**		**+PCRA**		**Avg. Imp.**	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	0.492	0.454	0.449	**0.443**	0.500	0.456	0.459	0.443	0.834	0.595	0.523	0.500	0.445	0.464	**0.430**	0.451	14.39%	5.98%
	192	0.571	0.499	0.504	**0.483**	0.573	0.498	0.534	0.490	0.982	0.661	0.605	0.554	0.494	0.496	**0.477**	0.484	15.13%	5.81%
	336	0.628	0.538	0.551	0.519	0.612	0.526	0.562	0.519	1.073	0.702	0.728	0.621	0.539	0.528	**0.517**	**0.514**	14.17%	4.79%
	720	0.834	0.652	0.691	0.603	0.766	0.620	0.716	0.609	1.349	0.797	0.879	0.687	0.661	0.606	**0.618**	**0.585**	16.26%	6.67%
ETTh2	96	0.265	0.334	0.251	0.333	0.251	0.326	0.249	0.329	0.381	0.404	0.285	0.355	0.239	0.335	**0.223**	**0.320**	9.41%	3.97%
	192	0.353	0.385	0.318	0.377	0.309	**0.362**	0.304	0.366	0.470	0.449	0.349	0.399	0.311	0.387	**0.291**	0.369	10.98%	4.20%
	336	0.515	0.457	**0.323**	**0.384**	0.343	0.384	0.324	0.384	0.528	0.476	0.373	0.419	0.344	0.406	0.329	0.400	19.14%	7.38%
	720	0.844	0.586	**0.384**	**0.426**	0.419	0.429	0.401	0.438	0.646	0.526	0.532	0.502	0.445	0.465	0.415	0.453	20.79%	8.10%
ETTm1	96	1.054	0.618	0.396	0.403	1.241	0.696	0.411	0.421	1.447	0.783	0.459	0.456	0.320	0.380	**0.312**	**0.367**	49.98%	29.92%
	192	1.165	0.676	0.455	0.442	1.311	0.726	0.460	0.453	1.511	0.814	0.507	0.487	0.386	0.423	**0.375**	**0.407**	48.78%	29.09%
	336	1.313	0.750	0.514	0.480	1.282	0.731	0.525	0.490	1.573	0.835	0.565	0.518	0.426	0.450	**0.418**	**0.438**	46.47%	27.35%
	720	1.619	0.866	0.584	0.525	1.348	0.765	0.583	0.532	1.694	0.873	0.650	0.570	0.485	0.492	**0.457**	**0.472**	46.99%	27.16%
ETTm2	96	0.204	0.300	0.163	0.259	0.225	0.314	0.166	0.263	0.351	0.377	0.203	0.294	0.154	0.258	**0.151**	**0.252**	22.62%	13.49%
	192	0.259	0.344	0.208	0.301	0.264	0.343	0.221	0.311	0.450	0.422	0.263	0.329	0.194	0.288	**0.183**	**0.279**	20.80%	11.73%
	336	0.388	0.421	0.266	0.347	0.304	0.369	0.261	0.342	0.534	0.466	0.309	0.373	0.227	0.317	**0.224**	**0.312**	22.17%	11.63%
	720	0.651	0.551	0.318	0.379	0.385	0.413	0.343	0.395	0.733	0.536	0.399	0.429	0.299	0.369	**0.285**	**0.356**	28.06%	14.77%
Weather	96	0.204	0.227	0.168	0.203	0.207	0.229	0.171	0.230	0.388	0.321	0.197	0.259	0.156	0.211	**0.151**	**0.197**	21.91%	9.01%
	192	0.256	0.272	0.211	**0.243**	0.264	0.277	0.211	0.272	0.510	0.395	0.240	0.300	**0.197**	0.246	0.205	0.250	21.69%	8.76%
	336	0.332	0.321	0.259	**0.282**	0.337	0.318	0.263	0.315	0.640	0.463	0.292	0.342	**0.246**	0.284	0.250	0.286	24.13%	9.59%
	720	0.437	0.380	0.327	**0.333**	0.454	0.375	0.350	0.388	0.929	0.562	0.375	0.403	0.317	0.337	**0.315**	0.339	27.13%	9.13%

Table 14: Matched ablation MSE over 60 tasks. w/o cosine uses squared-projection supervision; w/o input dep. uses a fixed PCA basis; w/o fact. uses direct residual regression; In-sample is the target-construction control. **Bold**/underline mark the best/second-best displayed values.

Dataset	H	**PatchTST**					**iTransformer**					**CycleNet**				
		PCRA	w/o fact.	w/o cosine	w/o input dep.	In-sample	PCRA	w/o fact.	w/o cosine	w/o input dep.	In-sample	PCRA	w/o fact.	w/o cosine	w/o input dep.	In-sample
ETTh1	96	**0.4221**	0.4272	0.4290	0.4249	0.4233	0.4307	0.4297	0.4457	0.4319	**0.4285**	0.4281	**0.4218**	0.4344	0.4353	0.4274
	192	0.4763	0.4768	0.5065	0.4981	**0.4704**	**0.4805**	0.4836	0.5205	0.4844	0.4809	0.4781	**0.4742**	0.5128	0.4754	0.4762
	336	0.5278	0.5360	0.5337	0.5334	**0.5215**	0.5398	0.5380	0.5607	0.5425	**0.5338**	0.5269	**0.5237**	0.5290	0.5300	0.5272
	720	0.6857	0.6908	0.7946	0.6999	**0.6846**	0.6684	0.6611	0.7523	0.6866	**0.6593**	0.6581	**0.6404**	0.7355	0.6570	0.6461
ETTh2	96	**0.2316**	0.2337	0.2323	0.2398	0.2327	**0.2452**	0.2636	0.2526	0.2592	0.2474	0.2292	0.2415	0.2349	0.2412	**0.2276**
	192	0.2741	0.2783	0.2918	**0.2737**	0.2738	0.2882	0.2951	0.3238	0.2937	**0.2832**	0.2728	0.2925	0.3010	**0.2715**	0.2825
	336	0.2999	0.3035	0.3170	**0.2932**	0.3021	0.3274	**0.3181**	0.3694	0.3262	0.3234	**0.3036**	0.3479	0.3309	0.3052	0.3108
	720	0.3607	0.3897	0.5519	0.3593	**0.3517**	**0.3881**	0.4240	0.6441	0.3957	0.3947	**0.3580**	0.4137	0.5305	0.4279	0.3616

Table 14 (continued)

Dataset	H	PatchTST					iTransformer					CycleNet				
		PCRA	w/o fact.	w/o cosine	w/o input dep.	In-sample	PCRA	w/o fact.	w/o cosine	w/o input dep.	In-sample	PCRA	w/o fact.	w/o cosine	w/o input dep.	In-sample
ETTm1	96	0.3789	0.3971	0.3894	0.4026	**0.3786**	0.3665	0.3700	0.3695	0.3649	**0.3641**	0.3848	0.3935	**0.3820**	0.3864	0.3882
	192	0.4305	0.4431	0.4327	0.4383	**0.4301**	0.4183	0.4172	**0.4167**	0.4169	0.4177	0.4302	0.4310	0.4326	0.4317	**0.4284**
	336	0.4892	0.4904	0.4923	0.4899	**0.4859**	0.4962	**0.4867**	0.4931	0.4901	0.4943	0.4849	0.4764	0.4760	0.4843	**0.4727**
	720	0.5507	**0.5416**	0.5588	0.5610	0.5421	0.5512	**0.5461**	0.5569	0.5565	0.5609	0.5438	**0.5269**	0.5383	0.5427	0.5292
ETTm2	96	**0.1562**	0.1648	0.1570	0.1568	0.1579	**0.1606**	0.1677	0.1636	0.1616	0.1614	0.1451	0.1534	**0.1450**	0.1465	0.1461
	192	0.2079	0.2218	0.2206	**0.2062**	0.2072	**0.2119**	0.2236	0.2150	0.2133	0.2123	0.1862	0.2041	**0.1856**	0.1882	0.1891
	336	0.2502	0.3001	0.2462	**0.2440**	0.2519	0.2530	0.2667	0.2562	0.2578	**0.2527**	0.2331	0.2801	**0.2289**	0.2349	0.2363
	720	**0.3186**	0.3834	0.3267	0.3268	0.3211	**0.3149**	0.3586	0.3455	0.3238	0.3191	0.2995	0.3602	**0.2944**	0.3066	0.3035
Weather	96	0.1690	**0.1642**	0.1686	0.1667	0.1677	0.1648	**0.1589**	0.1648	0.1629	0.1647	0.1564	**0.1536**	0.1554	0.1549	0.1568
	192	0.2115	0.2077	0.2123	**0.2073**	0.2092	0.2073	**0.2041**	0.2105	0.2072	0.2095	0.2023	**0.1973**	0.2029	0.1980	0.2034
	336	0.2602	0.2582	0.2618	**0.2581**	0.2614	0.2576	**0.2549**	0.2597	0.2566	0.2592	0.2544	0.2513	0.2538	**0.2497**	0.2547
	720	0.3333	**0.3316**	0.3381	0.3330	0.3334	0.3333	**0.3293**	0.3348	0.3321	0.3364	0.3289	0.3259	0.3328	**0.3243**	0.3306

The compact summary in the main text and the task-level table below use the same frozen predictions, splits, and default $K = 2$ configuration. The summary is averaged over tasks for readability; the detailed table is retained to expose the per-task variation. In particular, "w/o proj. strength" keeps the factorized architecture but replaces projection coefficients with residual norm targets, whereas "w/o staged train." keeps the same architecture and trains it end-to-end from random initialization with residual reconstruction MSE. These controls change one training choice at a time and are not used to select the main model.

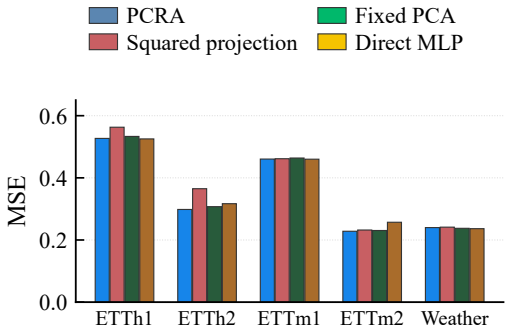


Figure 7: Absolute ablation MSE averaged over the controlled ETT/Weather tasks.

D Multi-Direction Analysis

D.1 Choice of K

The direction-count comparison evaluates independently trained adapters on three backbones, five ETT/Weather datasets, and four horizons, weighting all 60 tasks equally. Mean validation performance is highest for two directions among the evaluated counts, supporting the default used in the main results. This aggregate criterion differs from selecting a separate rank for each task.

Figures 8 and 9 distinguish representational capacity from realized benefit. Residual-derived projection coefficients yield increasing reconstruction gains as the span expands, while gains with predicted strengths approach a plateau. The gap contains both coefficient estimation error and future noise, whose contribution is formalized in Appendix F.7; it does not establish that estimation alone accounts for the observed saturation.

Figure 10 complements the four-horizon comparison in Figure 8 with repeated training on the 15 backbone–dataset tasks at $H = 192$. Both splits improve from one to two directions, but their mean gains peak at different larger bases: $K = 4$ on validation and $K = 8$ on test. These results support treating $K = 2$ as a compact default rather than a uniformly optimal direction count. The error bars

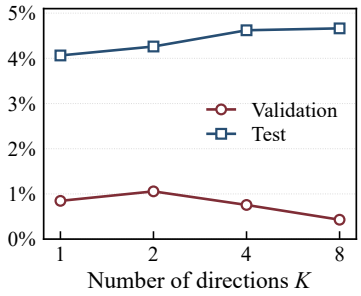


Figure 8: Validation and test performance on the complete four-horizon task grid.

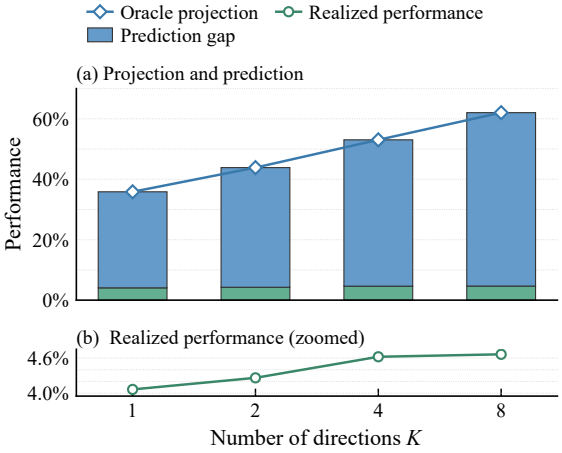


Figure 9: Residual representation and realized performance on the same test tasks. (a) Projection using observed residuals and its gap to predicted strengths. (b) Realized performance on an expanded vertical scale.

measure run-to-run variability on this fixed horizon, not uncertainty over the complete four-horizon grid.

Figure 6(a) in the main text examines whether the gain continues as K exceeds two: aggregate improvements grow more slowly after the first few directions, while individual backbone curves remain non-monotone. The direction-count study covers all three supervised backbones, five ETT/Weather datasets, and four horizons. Each $K \in \{1, 2, 4, 8\}$ is trained independently; $K = 2$ reuses the static checkpoints evaluated in the main tables. Figure 11 shows the 60 individual curves behind the aggregate result. No task is selected by the size or sign of its gain.

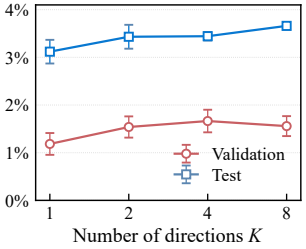


Figure 10: Repeated-run direction-count comparison at $H = 192$. Points are means across run-level task averages; error bars show one sample standard deviation across those averages.

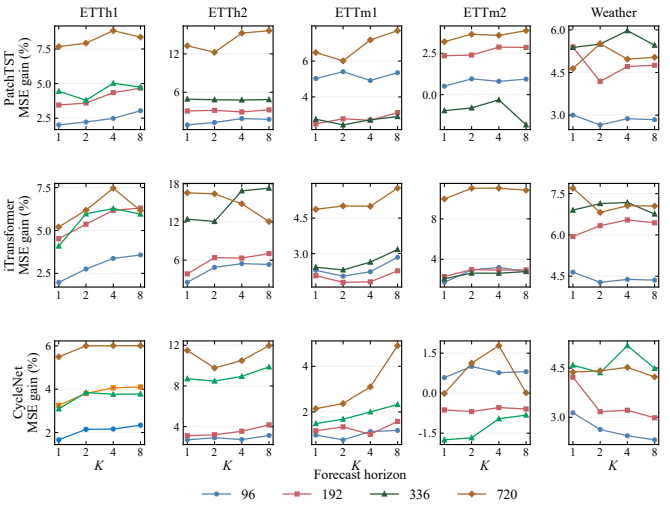


Figure 11: Direction-count curves by backbone and dataset; colors denote horizons.

D.2 Validation–Test Consistency

The validation and test curves need not agree. Across the full grid, validation gains for $K = 1, 2, 4, 8$ average 0.85%, 1.06%, 0.76%, and 0.43%, respectively. However, validation selects these ranks on 7, 12, 18, and 23 tasks, so the highest aggregate validation gain does not imply the most frequent task-wise selection. Validation and test select the same rank on 28/60 tasks. Selecting K per task using validation loss yields a mean test reduction of 4.49%. This retrospective comparison describes selection stability; the main results retain their fixed $K = 2$ setting.

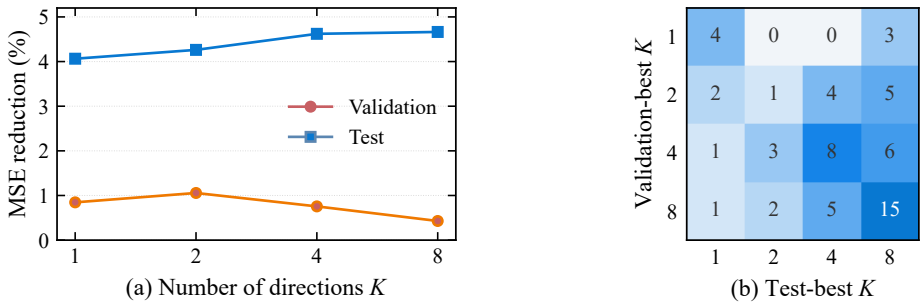


Figure 12: Validation/test comparison. Left: mean relative gains. Right: task counts for each pair of selected ranks.

D.3 Residual Energy and Coefficient Error

For the energy diagnostics, we replay all 240 saved checkpoints and verify test MSE against the recorded results within 3×10^{-6}. For each task, the observed projection-energy fraction is $\sum \|\mathbf{D}^\top \mathbf{r}\|^2 / \sum \|\mathbf{r}\|^2$, and the normalized coefficient error is $\sum \|\hat{\mathbf{s}} - \mathbf{D}^\top \mathbf{r}\|^2 / \sum \|\mathbf{r}\|^2$. Both sums cover the same test sample–channel pairs. Their difference is the relative MSE reduction by Eq. (4). We check this identity numerically before averaging across tasks.

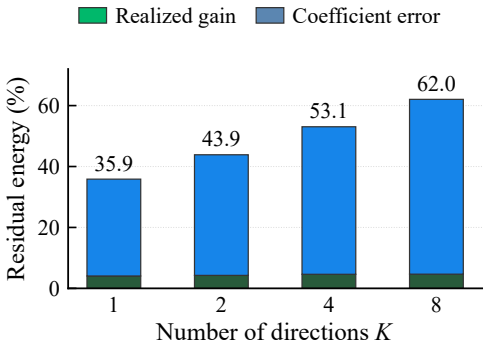


Figure 13: Observed projection-energy fraction, coefficient error, and realized performance across K.

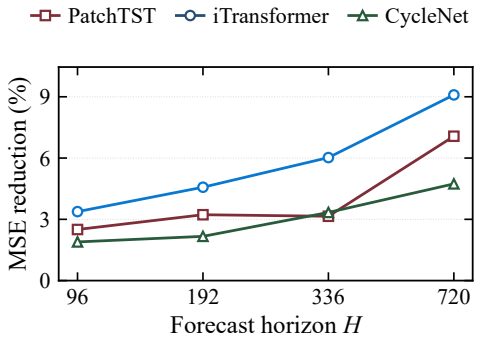


Figure 14: Static performance by forecast horizon on the five ETT/Weather datasets.

Figure 13 shows the empirical counterpart of the risk decomposition: increasing K exposes more residual energy, but the coefficient prediction term grows as well. The realized gain is their difference on the same task-level normalization.

To connect it to the risk analysis, fix the learned model and condition on the available information Z, as in Section 3.2. For orthonormal $\mathbf{D}$ measurable with respect to Z,

$$\mathbb{E}[\|\mathbf{D}^\top \mathbf{r}\|^2 \mid Z] = \|\mathbf{D}^\top \boldsymbol{\mu}\|^2 + \mathrm{tr}(\mathbf{D}^\top \Sigma \mathbf{D}), \tag{11}$$

$$\mathbb{E}[\|\hat{\mathbf{s}} - \mathbf{D}^\top \mathbf{r}\|^2 \mid Z] = \|\hat{\mathbf{s}} - \mathbf{D}^\top \boldsymbol{\mu}\|^2 + \mathrm{tr}(\mathbf{D}^\top \Sigma \mathbf{D}). \tag{12}$$

The shared projected-noise term cancels on subtraction, recovering Eq. (7) after averaging over Z. Consequently, a higher observed projection-energy fraction can reflect predictable residual structure, projected noise, or both. The diagnostic cannot identify their separate contributions. It illustrates why increased residual coverage need not produce a comparable gain in forecasting accuracy; it does not establish that larger independently retrained models are worse.

Figure 14 disaggregates the four-horizon means in the main tables. Every point uses the same five datasets, so the curves isolate the effect of horizon within the fixed correction protocol.

D.4 Qualitative Examples

Figures 5 and 15 examine four test windows spanning all three backbones on ETTh1 and ETTh2: (a) PatchTST/ETTh2, $H{=}192$, zero-based window 2568, channel 6; (b) iTransformer/ETTh2, $H{=}336$, window 1438, channel 6; (c) CycleNet/ETTh1, $H{=}336$, window 2383, channel 1; and (d) iTransformer/ETTh1, $H{=}192$, window 2499, channel 6. The cases are selected post hoc among windows whose cosine gain exceeds 0.1, whose two-direction alignment exceeds 0.3, whose relative MSE reduction exceeds 10%, and whose residual energy reaches the task median; they illustrate the correction geometry rather than its frequency. These are qualitative mechanism illustrations, not aggregate evidence. Across the repeated-run tasks, the mean MSE reduction of the two-direction adapter over the single-direction one is 0.599% on ETTh1 (8/12 tasks), 0.585% on ETTh2 (9/12), -0.105% on ETTm1 (3/12), and 0.353% on ETTm2 (7/11). All displayed curves are unmodified predictions and observations.

For each window, let $\mathbf{A} = [\mathbf{r}, \hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2] \in \mathbb{R}^{H\times 3}$, where the subscripts identify separately trained single- and two-direction adapters. An orthonormal basis $\mathbf{Q}$ of its column span gives coordinates $\mathbf{C} = \mathbf{Q}^\top \mathbf{A}$ satisfying $\mathbf{C}^\top \mathbf{C} = \mathbf{A}^\top \mathbf{A}$. This coordinate transformation preserves vector norms and pairwise inner products. For display, however, each resulting vector is normalized to unit length, and a common orthographic view is compressed vertically. The displayed lengths and screen angles therefore do not preserve residual magnitudes or high-dimensional angles. Dotted connectors compare the normalized endpoints rather than quantify forecast-error norms. Each case uses its own basis; the printed cosines are measured in the original H-dimensional space before display transformations.

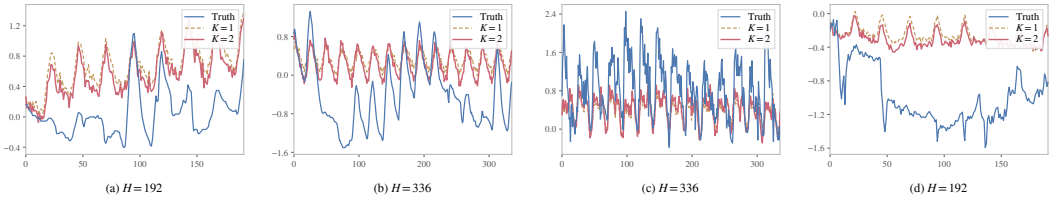


Figure 15: Temporal forecasts corresponding to Figure 5. Single- and two-direction adapters are evaluated on four selected windows: (a) PatchTST/ETTh2, $H{=}192$; (b) iTransformer/ETTh2, $H{=}336$; (c) CycleNet/ETTh1, $H{=}336$; (d) iTransformer/ETTh1, $H{=}192$.

E Test-Time Adaptation Protocol and Efficiency

E.1 Causal Feedback Timeline

For a forecast issued at index i, a historical window beginning at j is eligible for adaptation only after its full target is observed, namely $j + H \le i$. The prediction at i is emitted before any update from that window. After maturity, the cached features, frozen directions, and residual produce the delayed projection target; the resulting pair enters the replay buffer and updates the strength gate used by the next forecast. The supervised Ada-T comparisons on the three primary backbones use the same full-horizon eligibility constraint; their warmup and optimizer rules differ.

E.2 Matched and Native Protocols

PCRA-T and the primary supervised Ada-T comparison require fully observed historical horizons. This matches label eligibility, not the entire adaptation algorithm: Ada-T retains its 96-window warmup, learning rate 10^{-6}, gradient clipping, and persistent optimizer. Its update/no-update choice is selected on validation data, and the test stream restarts from the static checkpoint. TAFAS and PETSA use partially observed targets in the supervised study and full-horizon feedback in the foundation-model study. Foundation-model baselines can incur additional window or batch delays. Additional-backbone Ada-T and foundation-model Ada-Y records retain their native protocols. Consequently, the online tables do not constitute a fully matched cross-method replay.

For clarity, we use the following protocol names throughout the appendix. The *formal PCRA-T protocol* updates the complete strength gate with physical projection MSE, uses $\mathbf{s}^* = \mathbf{D}^\top \mathbf{r}$, one Adam step at 10^{-5} on the latest 32 matured windows, update interval one, and fixed offline standardizers.

Family	Representative methods	Deployment signal	Relation to PCRA
Static distribution handling	RevIN, Non-stationary Transformer, Dish-TS, SAN, SIN, FAN, APT Kim et al. (2022); Liu et al. (2022); Fan et al. (2023); Liu et al. (2023); Han et al. (2024); Ye et al. (2024); Li et al. (2026)	Fixed forward pass; methods generally alter normalization or the training pipeline	Complementary rather than strict frozen-output baselines
Static post-forecast revision	PIR, GLAFF, UEC-STD, δ-Adapter Liu et al. (2025b); Wang et al. (2024); Nguyen et al. (2026); Liang et al. (2026)	Fixed adapter at deployment	PIR and δ-Adapter are principal static comparisons; PCRA differs in direction-conditioned projection supervision
Retrieved or matured feedback	SOLID, COSA, FAC, ELF Chen et al. (2024); Im & Kwon (2026); Wang et al. (2026); Lee et al. (2025)	Historical labels, retrieved source samples, or online feedback	Comparable only after aligning label maturity and prediction–update order
Partial-prefix/current-window feedback	TAFAS, PETSA, DynaTTA, STEPS Kim et al. (2025); Medeiros et al. (2025); Grover & Etemad (2025); Liu et al. (2026)	Partially observed current target or revealed prefix	Native protocol is stronger than PCRA's matured-only boundary; report separately or rerun
Unlabeled or specialized TTA	AdaNODEs, normalization-only TTA, RG-TTA, ADCSD, Black-Mamba Dang et al. (2026); Wu et al. (2026); Kumar et al. (2026); Guo et al. (2024); Soriano et al. (2026)	Unlabeled windows, regime controllers, traffic structure, or adaptive memory	Establishes broader TTA boundary; not a direct primary baseline

Table 15: Protocol taxonomy of the static and dynamic forecasting-adaptation literature reviewed for this work. Classification is based on deployment-time information and updated components, not merely whether the output is sample-dependent.

The *matched protocol* shares the same eligibility rule and measurement boundary across methods while preserving each method's declared update rule. A *native protocol* is the authors' released implementation and may use a different warm-up, objective, update scope, or feedback schedule. The historical output-layer coefficient-MAE runs are reported only as controls; they are not substituted for formal PCRA-T.

E.3 Update Rule

Let ϕ denote all strength-gate parameters. Once a historical horizon is fully observed, its cached features, frozen directions, and residual yield a delayed projection target. PCRA-T updates the complete gate by minimizing physical projection MSE:

$$\mathcal{L}_{\mathrm{TTA}}(\phi) = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{H}\left\|g_\phi(\mathbf{z}_m) - \mathbf{s}_m^*\right\|_2^2, \tag{13}$$

where $\mathbf{s}_m^* = \mathbf{D}_m^\top(\mathbf{y}_m - \mathbf{b}_m)$ and both $g_\phi(\mathbf{z}_m)$ and $\mathbf{s}_m^*$ are in residual units. Here M counts the sample–channel pairs in the buffer. The implementation divides the objective in Eq. (10) by H to match the per-time-step reconstruction-MSE scale; it does not divide by K. If the network output q_ϕ is standardized, the same loss is implemented as $\frac{1}{MH}\sum_{m,k}\sigma_k^2(q_{\phi,k}(\mathbf{z}_m) - s_{m,k}^{*,\mathrm{std}})^2$. These scale factors apply only in standardized coordinates and must not be applied again after inverse standardization. For frozen orthonormal directions, reconstruction MSE differs from this loss only by the orthogonal residual energy, which is independent of ϕ. The update uses a buffer of fully observed historical windows. For the one-hidden-layer strength gate with input width 22, hidden width $h = 32$, and $K = 2$ outputs, all adaptable parameters number

$$P = (22+1)h + (h+1)K = 802. \tag{14}$$

The forecaster, direction heads, and offline standardizers remain fixed; the optimizer is reinitialized for each eligible update. A historical window whose final target is indexed by $j+H$ is eligible when $j + H \le t$. The prediction at index t is emitted before this update; the updated gate is first used for the next prediction. Each eligible update takes one Adam step at learning rate 10^{-5} on the latest 32 fully observed windows, with dropout disabled and offline standardizers fixed.

The offline stage uses the same projection target but a separate training schedule: direction predictors are fitted first, strength predictors are then trained on standardized projection coefficients, and the optional joint reconstruction refinement starts from those selected checkpoints. This stage-wise initialization is part of the formal offline PCRA definition; the online replay does not refit directions or restart the offline stages.

E.4 Hyperparameter Sensitivity

Figure 6(b–d) uses the current full-gate projection-MSE protocol on the 15 three-backbone ETT/Weather tasks at $H = 192$. Relative to the reference learning rate 10^{-5}, buffer 32, and update interval 1, the sweeps vary learning rate over $\{10^{-6}, 10^{-5}, 10^{-4}\}$, buffer size over $\{8, 32, 128\}$, and

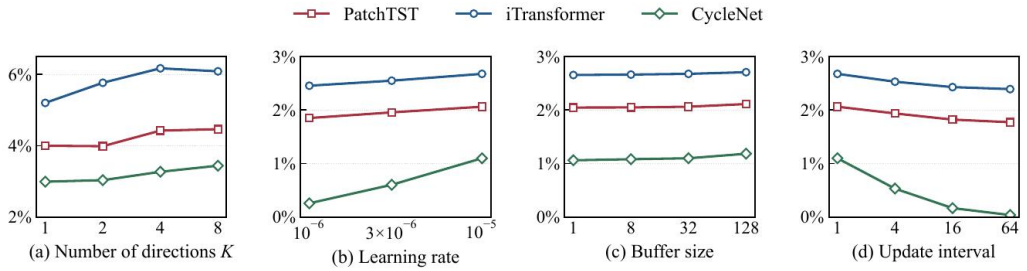


Figure 16: Historical output-layer coefficient-MAE sensitivity control. This figure is retained for reference and is separate from the unified full-gate projection-MSE sensitivity in the main text.

update interval over $\{1, 4, 16\}$. Only one setting changes at a time. These test-stream replays are distinct from the validation control below.

E.4.1 Output-Layer MAE Control

All TTA variants are separated by update scope and objective. The formal PCRA-T protocol used in the main results updates the complete strength gate with physical projection MSE. The output-layer projection-MSE variant is a parameter-count control, while the historical coefficient-MAE variant is retained only as an archival sensitivity control. Neither controlled variant defines the formal PCRA-T result.

The following figure preserves the historical parameter sweeps for the output-layer update with coefficient MAE. These runs use a different update scope and objective from the full-gate, projection-MSE PCRA-T in the main tables. Each sweep changes one setting while fixing the others at one gradient step, learning rate 10^{-5}, buffer size 32, and an update at every eligible window. The learning rates are $\{10^{-6}, 3 \times 10^{-6}, 10^{-5}\}$; buffer sizes are $\{1, 8, 32, 128\}$; update intervals are $\{1, 4, 16, 64\}$ windows. An interval of q triggers one update every q eligible windows. These historical sweeps use validation data from PatchTST, iTransformer, and CycleNet on the five ETT/Weather datasets at $H = 192$. Performance is averaged over datasets for each backbone. The corresponding full-gate projection-MSE sensitivity is now reported in Figure 6 in the main text.

E.5 Runtime and Memory

Table 16: Archived cost of the earlier online variants on 12 ETT tasks at $H = 192$. Time is seconds per 1,000 windows, excluding backbone inference.

Method	Updated params	MSE ↓	Time ↓
Ada-Y	1.64M	0.346	1.67
PCRA (output-layer ctrl.)	66	0.345	2.14

Only test-time execution is measured here; offline training costs are not included. The study covers 15 backbone–dataset pairs at $H = 192$ on RTX 4090 D GPUs. Cached-adapter measurements run on one physical GPU; native TTA measurements run on another of the same model. Each active case has three replays, with same-device metric/update parity and GPU-process isolation checks. The repetitions measure runtime variability, not training variability. Historical cross-device numerical differences are retained in the result archive rather than interpreted as changes in the forecasting results.

The cached replay table uses the historical output-layer control available in the runtime archive, whereas the main results use the complete strength gate. This control isolates replay overhead; it is not used to claim a runtime ranking for the full-gate method. Timing begins with cached forecasts and deserialized CPU weights and includes adapter initialization, feature and direction computation, online replay, and metric reduction. Backbone inference, disk I/O, and offline training are excluded. Peak memory is incremental allocated GPU memory. The timer for TAFAS, PETSA, and COSA includes native backbone forwarding, causal updates, metric reduction, and native reporting; initialization is recorded separately, and memory is total allocated peak. We do not subtract an independently batched backbone measurement to infer adaptation overhead.

Each task's median stream time over three replays is divided by its number of forecast windows and scaled to 1,000 windows. Table 16 averages these normalized times over the same 12 active ETT workloads for Ada-Y and the historical PCRA output-layer control; it also averages MSE from those replays. Figure 17 groups the same times by dataset. These costs are not the latency of a batch of 1,000 simultaneous requests. Active counts differ: 15 for the historical PCRA control, 12 for Ada-Y, 12 for TAFAS, 8 for PETSA, and 13 for COSA. Pooling their costs would not give a matched speed ranking. Ada-Y selects no online update on Weather but still incurs static adapter inference cost. Native no-update selections (3 TAFAS, 7 PETSA, 2 COSA) have separately measured backbone references; these are not presented as zero-cost adaptation.

The number of triggered updates also differs from the number of optimizer steps. The archived output-layer control takes three steps at each trigger; for example, CycleNet/ETTh1 has 3,101 triggers over 3,293 windows, or 9,303 optimizer steps. Update frequency, replay, preprocessing, and implementation overhead all contribute to runtime despite the 66-parameter update layer.

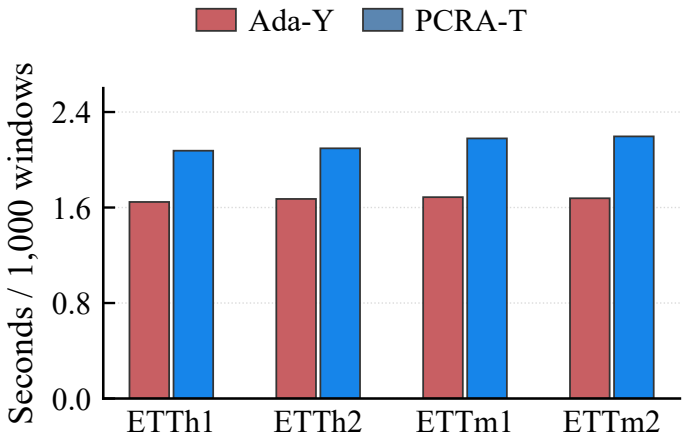


Figure 17: Archived cached-replay cost on matched ETT workloads. The PCRA series uses the 66-parameter output-layer control with three steps per trigger, not the formal one-step PCRA-T. Lower is better.

Figure 17 groups the same normalized times as Table 16. It reports adaptation overhead rather than end-to-end backbone latency, allowing the update mechanism to be compared on a common cached-forecast workload.

F Theoretical Results and Proofs

The following reconstruction analysis applies to one channel and one correction direction; subscripts are omitted for clarity. We provide complete derivations of the results stated in the main paper.

F.1 Identifiability of Direction–Strength Factorization

Proposition 1 (Non-identifiability of the unconstrained factorization, and its resolution by normalization). Let $\Phi(\boldsymbol{\epsilon}, s) = s\boldsymbol{\epsilon}$ for $\boldsymbol{\epsilon} \in \mathbb{R}^H \setminus \{\mathbf{0}\}$ and $s \in \mathbb{R}$. (i) Φ is constant on each orbit $O(\boldsymbol{\epsilon}, s) = \{(c\boldsymbol{\epsilon}, s/c) : c \neq 0\}$, so any loss of the form $\ell(\mathbf{r}, \Phi(\boldsymbol{\epsilon}, s))$ is constant on $O(\boldsymbol{\epsilon}, s)$ and the strength s is not identified. (ii) Restricted to $\|\boldsymbol{\epsilon}\|_2 = 1$, the fibre of Φ over any nonzero correction $\mathbf{v}$ is exactly $\{(\mathbf{v}/\|\mathbf{v}\|_2, \|\mathbf{v}\|_2), (-\mathbf{v}/\|\mathbf{v}\|_2, -\|\mathbf{v}\|_2)\}$, so $(\boldsymbol{\epsilon}, s)$ is identified up to a joint sign flip. (iii) For $\mathbf{r} \neq \mathbf{0}$ the direction objective of Eq. (9) is minimized uniquely at $\mathbf{e} = \mathbf{r}/\|\mathbf{r}\|_2$, which removes the residual sign ambiguity in (ii).

Proof. (i) $\Phi(c\boldsymbol{\epsilon}, s/c) = (s/c)(c\boldsymbol{\epsilon}) = s\boldsymbol{\epsilon} = \Phi(\boldsymbol{\epsilon}, s)$ for every $c \neq 0$; the loss depends on $(\boldsymbol{\epsilon}, s)$ only through Φ. (ii) If $\|\boldsymbol{\epsilon}\|_2 = 1$ and $s\boldsymbol{\epsilon} = \mathbf{v} \neq \mathbf{0}$, then $|s| = \|\mathbf{v}\|_2$ and $\boldsymbol{\epsilon} = \mathbf{v}/s$, leaving the two stated solutions. (iii) On the unit sphere, $1 - \mathbf{e}^\top \bar{\mathbf{r}}$ with $\bar{\mathbf{r}} = \mathbf{r}/\|\mathbf{r}\|_2$ is minimized by maximizing $\mathbf{e}^\top \bar{\mathbf{r}}$, and Cauchy–Schwarz gives $\mathbf{e}^\top \bar{\mathbf{r}} \leq 1$ with equality iff $\mathbf{e} = \bar{\mathbf{r}}$. □

Part (i) identifies the scale ambiguity of unconstrained factorization. Parts (ii)–(iii) motivate normalization and separate direction supervision, but do not prove that staged optimization is necessary or superior to joint training. The remaining propositions characterize the strength target once the direction has been realized.

F.2 Projection-Optimal Strength

Proposition 2 (MSE-optimal scalar coefficient on a fixed non-unit basis). For any nonzero $\mathbf{d}$, define $s^* = \langle \mathbf{r}, \mathbf{d} \rangle / \|\mathbf{d}\|_2^2$ and $\mathbf{r}^\perp = \mathbf{r} - s^*\mathbf{d}$. Then $\langle \mathbf{r}^\perp, \mathbf{d} \rangle = 0$ and

$$\|\mathbf{r} - \hat{s}\mathbf{d}\|_2^2 = \|\mathbf{r}^\perp\|_2^2 + \|\mathbf{d}\|_2^2(\hat{s} - s^*)^2. \tag{15}$$

Proof. Because $\mathbf{r} = \mathbf{r}^\perp + s^*\mathbf{d}$ and $\langle \mathbf{r}^\perp, \mathbf{d} \rangle = 0$,

$$\begin{aligned}\|\mathbf{r} - \hat{s}\mathbf{d}\|_2^2 &= \|\mathbf{r}^\perp + (s^* - \hat{s})\mathbf{d}\|_2^2 \\ &= \|\mathbf{r}^\perp\|_2^2 + \|\mathbf{d}\|_2^2(\hat{s} - s^*)^2.\end{aligned}$$

The same minimizer follows from the derivative $\ell'(\hat{s}) = 2(\hat{s}\|\mathbf{d}\|_2^2 - \langle \mathbf{r}, \mathbf{d} \rangle)$ of the scalar reconstruction loss; the second term is nonnegative and vanishes only at $\hat{s} = s^*$ because $\mathbf{d} \neq \mathbf{0}$. □

This is the reconstruction basis for the stage-wise direction-then-coefficient supervision within the fixed-direction scalar correction class. Applied column-wise to an orthogonal direction matrix, it yields the multi-direction decomposition in Eq. (4). This classical identity is used as a supervision principle rather than claimed as a new standalone theorem.

F.3 Norm–Projection Mismatch

Proposition 3 (Norm–projection mismatch under an imperfect fixed direction). Let $\mathbf{e} = \mathbf{d}/\|\mathbf{d}\|_2$, $m = \|\mathbf{r}\|_2$, and $\rho = \cos(\mathbf{r}, \mathbf{e})$ for $m > 0$. Applying the full residual norm along the normalized direction gives

$$\begin{aligned}\mathcal{L}_{\text{norm}} &= \|\mathbf{r} - m\mathbf{e}\|_2^2 = 2m^2(1 - \rho), \\ \mathcal{L}_{\text{proj}} &= \|\mathbf{r} - \langle \mathbf{r}, \mathbf{e} \rangle \mathbf{e}\|_2^2 = m^2(1 - \rho^2), \\ \mathcal{L}_{\text{norm}} - \mathcal{L}_{\text{proj}} &= m^2(1 - \rho)^2 \geq 0.\end{aligned} \tag{16}$$

Proof. Using $\langle \mathbf{r}, \mathbf{e} \rangle = m\rho$ and $\|\mathbf{e}\|_2 = 1$ gives the first two equalities by direct expansion; their difference is $2m^2(1 - \rho) - m^2(1 - \rho^2) = m^2(1 - \rho)^2$. □ Thus residual norm is not generally the MSE-optimal scalar strength within the fixed-direction correction class: even a perfect norm predictor incurs excess MSE whenever the proposed direction is imperfect. For $0 < \rho < 1$, it over-corrects; for $\rho < 0$, it also applies the correction with the wrong sign. This is a projection-mismatch analysis, not a claim of a new standalone mathematical theorem.

F.4 Projection Targets and Orthogonal Decoupling

For a nonzero direction $\mathbf{d}$, the unique least-squares coefficient and its residual are

$$s^* = \frac{\langle \mathbf{r}, \mathbf{d} \rangle}{\|\mathbf{d}\|_2^2}, \qquad \mathbf{r}^\perp = \mathbf{r} - s^*\mathbf{d}, \qquad \langle \mathbf{r}^\perp, \mathbf{d} \rangle = 0. \tag{17}$$

Pythagoras gives $\|\mathbf{r} - \hat{s}\mathbf{d}\|_2^2 = \|\mathbf{r}^\perp\|_2^2 + \|\mathbf{d}\|_2^2(\hat{s} - s^*)^2$, yielding Eq. (2) when $\|\mathbf{d}\|_2 = 1$.

For multiple unit directions, expand the squared error as

$$\Big\|\mathbf{r} - \sum_k s_k\mathbf{d}_k\Big\|_2^2 = \|\mathbf{r}\|_2^2 - 2\sum_k s_k\langle \mathbf{r}, \mathbf{d}_k \rangle + \sum_k s_k^2 + \sum_{k \neq k'} s_k s_{k'} \mathbf{d}_k^\top \mathbf{d}_{k'}. \tag{18}$$

With orthonormal directions, the cross terms vanish, and each remaining scalar quadratic is minimized uniquely by $s_k^* = \langle \mathbf{r}, \mathbf{d}_k \rangle$. Completing the squares gives

$$\|\mathbf{r} - \mathbf{D}\hat{\mathbf{s}}\|_2^2 = \|\mathbf{r}\|_2^2 - \sum_k (s_k^*)^2 + \sum_k (\hat{s}_k - s_k^*)^2.$$

Substituting $\|\mathbf{r}\|_2^2 - \sum_k (s_k^*)^2 = \|(I - \mathbf{D}\mathbf{D}^\top)\mathbf{r}\|_2^2$ proves Eq. (4).

For linearly independent but nonorthogonal directions, the optimum instead solves $\mathbf{D}^\top \mathbf{D}\mathbf{s} = \mathbf{D}^\top \mathbf{r}$. If the proposal matrix $\mathbf{V}$ has full column rank, Gram–Schmidt gives $\mathbf{V} = \mathbf{D}\mathbf{T}$ for an invertible triangular $\mathbf{T}$. Hence $\mathrm{span}(\mathbf{V}) = \mathrm{span}(\mathbf{D})$: orthogonalization preserves the correction space while changing its coordinates.

F.5 Conditional Risk and Forecasting Gain

Consider a forecast after fitting the model. Let Z denote the information available at that time, including the fitted parameters and any eligible historical feedback. Both $\mathbf{D}_K(Z)$ and $\hat{\mathbf{s}}(Z)$ must be measurable with respect to this information, without using the current future residual. Assume finite second moments, and define

$$\boldsymbol{\mu}(Z) = \mathbb{E}[\mathbf{r} \mid Z], \qquad \boldsymbol{\varepsilon} = \mathbf{r} - \boldsymbol{\mu}(Z), \qquad \Sigma(Z) = \mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top \mid Z].$$

The term "irreducible" refers to the information in Z, not to an absolute limit under additional observations. No independence across horizon positions or Gaussian assumption is needed.

Conditioning on Z, the mean-zero noise eliminates the cross term:

$$\begin{aligned}\mathcal{R}_K(Z) &= \mathbb{E}[\|\boldsymbol{\mu} - \mathbf{D}_K\hat{\mathbf{s}} + \boldsymbol{\varepsilon}\|_2^2 \mid Z] \\ &= \|\boldsymbol{\mu} - \mathbf{D}_K\hat{\mathbf{s}}\|_2^2 + \operatorname{tr}\Sigma.\end{aligned}$$

Write $P_K = \mathbf{D}_K\mathbf{D}_K^\top$ and $\boldsymbol{\alpha}_K = \mathbf{D}_K^\top\boldsymbol{\mu}$. Then

$$\boldsymbol{\mu} - \mathbf{D}_K\hat{\mathbf{s}} = (I - P_K)\boldsymbol{\mu} + \mathbf{D}_K(\boldsymbol{\alpha}_K - \hat{\mathbf{s}}).$$

The two terms are orthogonal, and $\|\mathbf{D}_K(\boldsymbol{\alpha}_K - \hat{\mathbf{s}})\|_2^2 = \|\boldsymbol{\alpha}_K - \hat{\mathbf{s}}\|_2^2$. Substitution proves Eq. (6). The coefficient term includes approximation, estimation, and optimization error relative to the conditional mean coefficients.

For the uncorrected forecaster, the same conditional mean decomposition gives $\mathcal{R}_0(Z) = \|\boldsymbol{\mu}\|_2^2 + \operatorname{tr}\Sigma$. Pythagoras yields $\|\boldsymbol{\mu}\|_2^2 = \|(I - P_K)\boldsymbol{\mu}\|_2^2 + \|\boldsymbol{\alpha}_K\|_2^2$. Subtracting the corrected risk therefore gives

$$\mathcal{R}_0(Z) - \mathcal{R}_K(Z) = \|\boldsymbol{\alpha}_K\|_2^2 - \|\hat{\mathbf{s}} - \boldsymbol{\alpha}_K\|_2^2.$$

Taking expectations gives Eq. (7). Within the stated information and orthogonality assumptions, strict improvement is equivalent to $\mathbb{E}\|\hat{\mathbf{s}} - \boldsymbol{\alpha}_K\|_2^2 < \mathbb{E}\|\boldsymbol{\alpha}_K\|_2^2$. The identity characterizes the condition under which a correction improves forecasting risk.

Proposition 4 (Optimal deployable coefficient predictor). Let D denote a non-unit direction, $W = \|D\|_2^2$, $T = \langle R, D\rangle$, and $s^* = T/W$. Let Z contain only inference-visible projection-strength features. Assume $\mathbb{E}[W \mid Z] > 0$. For a coefficient predictor $g(Z)$,

$$\mathcal{R}(g) = \mathbb{E}\|R - g(Z)D\|_2^2. \tag{19}$$

The population-risk minimizer is the weighted conditional projection

$$g^*(Z) = \frac{\mathbb{E}[T \mid Z]}{\mathbb{E}[W \mid Z]}. \tag{20}$$

When W is measurable with respect to Z—for example, when Z retains the direction norm—this simplifies to $g^*(Z) = \mathbb{E}[T \mid Z]/W$. Indeed, conditioning the expanded squared loss on Z yields a quadratic in $g(Z)$ with positive coefficient $\mathbb{E}[W \mid Z]$; differentiating it gives Eq. (20). Substituting the minimizer and completing the square gives Eq. (21). Writing $\bar{s}(Z) = g^*(Z), \omega(Z) = \mathbb{E}[W \mid Z]$, and $R^\perp = R - s^*D$, the risk decomposes as

$$\begin{aligned}\mathcal{R}(g) = \mathbb{E}\|R^\perp\|_2^2 &+ \mathbb{E}\left[W(s^* - \bar{s}(Z))^2\right] \\ &+ \mathbb{E}\left[\omega(Z)\big(g(Z) - \bar{s}(Z)\big)^2\right].\end{aligned} \tag{21}$$

The three terms quantify direction-limited error, irreducible weighted coefficient uncertainty, and learnable weighted coefficient error. In particular,

$$\mathcal{R}(g^*) = \mathbb{E}\|R\|_2^2 - \mathbb{E}\left[\frac{\mathbb{E}[T \mid Z]^2}{\mathbb{E}[W \mid Z]}\right] \le \mathbb{E}\|R\|_2^2. \tag{22}$$

The guarantee is population-level and within the fixed-direction correction class; a finite-sample learned controller still needs regularization and safety controls. It characterizes the ideal reconstruction-risk solution. The implemented standardized L1 objective is a robust surrogate for heavy-tailed oracle coefficients and is not claimed to attain this conditional MSE optimum exactly.

F.6 The Marginal Value of an Additional Direction

Consider a nested extension $\mathbf{D}_{K+1} = [\mathbf{D}_K, \mathbf{d}]$ with $\mathbf{D}_K^\top \mathbf{d} = 0$ and $\|\mathbf{d}\|_2 = 1$. Hold the first K predicted coefficients fixed and append $\hat{s}_{\text{new}}(Z)$. For $s^*_{\text{new}}(Z) = \mathbf{d}^\top \boldsymbol{\mu}(Z)$,

$$\|(I - P_{K+1})\boldsymbol{\mu}\|_2^2 = \|(I - P_K)\boldsymbol{\mu}\|_2^2 - (s^*_{\text{new}})^2,$$

while the coefficient-error term increases by $(\hat{s}_{\text{new}} - s^*_{\text{new}})^2$. The irreducible-noise term is unchanged. Subtracting the two risks and taking expectations proves Eq. (8). Thus the extra direction strictly helps if and only if $\mathbb{E}[(\hat{s}_{\text{new}} - s^*_{\text{new}})^2] < \mathbb{E}[(s^*_{\text{new}})^2]$.

If the earlier coefficient predictions also change, the difference in their coefficient errors must be included. If the earlier directions are retrained, the change in uncovered predictable residual must likewise be evaluated for the new span. The conditional risk decomposition applies to both models, but the single-direction increment is no longer their full difference. The result therefore explains a tradeoff rather than asserting a universal optimum at $K = 2$. With unrestricted coefficient prediction and nested spans, the population optimum cannot worsen, since setting the new coefficient to zero recovers the previous correction.

F.7 Projected Noise, Span Alignment, and Oracle Gains

An *oracle coefficient* uses the realized future residual: $\mathbf{s}^* = \mathbf{D}_K^\top \mathbf{r}$. For a fixed learned direction span, it minimizes reconstruction error, but it is unavailable when a forecast is issued. The associated gain is therefore a diagnostic bound within that span, not an attainable forecasting guarantee or a bound across all predictors.

The exact projection target contains both predictable and unpredictable components:

$$\mathbf{s}^* = \boldsymbol{\alpha}_K + \boldsymbol{\eta}_K, \qquad \boldsymbol{\eta}_K = \mathbf{D}_K^\top \boldsymbol{\varepsilon}, \qquad \operatorname{Cov}(\boldsymbol{\eta}_K \mid Z) = \mathbf{D}_K^\top \Sigma \mathbf{D}_K. \tag{23}$$

Orthogonal directions eliminate geometric cross terms but need not make label noise statistically independent. Their projected noise energy is $\operatorname{tr}(\mathbf{D}_K^\top \Sigma \mathbf{D}_K)$, equal to $K\sigma^2$ when $\Sigma = \sigma^2 I$. Enlarging a fixed span does not create future noise; it exposes additional noise coordinates in the supervised coefficients. With finite training data, fitting these coordinates can incur estimation error.

For a nonzero residual, span alignment is

$$A_K = \frac{\|\mathbf{D}_K^\top \mathbf{r}\|_2}{\|\mathbf{r}\|_2}, \qquad A_1 = \frac{|\mathbf{d}_1^\top \mathbf{r}|}{\|\mathbf{r}\|_2}, \qquad A_2 = \frac{\sqrt{(\mathbf{d}_1^\top \mathbf{r})^2 + (\mathbf{d}_2^\top \mathbf{r})^2}}{\|\mathbf{r}\|_2}. \tag{24}$$

Its square is the captured energy fraction for that residual. It measures alignment with a span, not the cosine of the gate's predicted correction. The unweighted mean of A_K does not determine the energy-weighted oracle gain.

To relate the analysis to MSE reductions, let $R_0 = \mathbb{E}\|\mathbf{r}\|_2^2 > 0$ and define the oracle gain O_K and the predicted correction's gain G_K:

$$O_K = \frac{\mathbb{E}\|\mathbf{D}_K^\top \mathbf{r}\|_2^2}{R_0} = \frac{\mathbb{E}\|\boldsymbol{\alpha}_K\|_2^2 + \mathbb{E}\operatorname{tr}(\mathbf{D}_K^\top \Sigma \mathbf{D}_K)}{R_0}, \tag{25}$$

$$G_K = 1 - \frac{\mathbb{E}\|\mathbf{r} - \mathbf{D}_K \hat{\mathbf{s}}\|_2^2}{R_0} = \frac{\mathbb{E}\|\boldsymbol{\alpha}_K\|_2^2 - \mathbb{E}\|\hat{\mathbf{s}} - \boldsymbol{\alpha}_K\|_2^2}{R_0}. \tag{26}$$

The same expectation and normalization must be used for both quantities. Their difference is

$$O_K - G_K = \frac{\mathbb{E}\|\hat{\mathbf{s}} - \boldsymbol{\alpha}_K\|_2^2 + \mathbb{E}\operatorname{tr}(\mathbf{D}_K^\top \Sigma \mathbf{D}_K)}{R_0} = \frac{\mathbb{E}\|\hat{\mathbf{s}} - \mathbf{s}^*\|_2^2}{R_0}. \tag{27}$$

Consequently, a growing oracle gain may partly reflect additional captured noise, and the oracle–actual gap contains both coefficient prediction error and projected future noise. The gap alone does not identify their separate magnitudes. The conditional moments $\boldsymbol{\mu}$ and Σ are theoretical quantities; empirical oracle diagnostics do not directly measure them. These identities require coefficients in residual units and the exact projection, rather than standardized labels or ridge-shrunk coefficients.

G Implementation Details

This section gives the numerical operations behind the direction–strength adapter. The analysis uses exact unit directions; implementations stabilize normalization as $\mathcal{N}_\epsilon(\mathbf{v}) = \mathbf{v}/(\|\mathbf{v}\|_2 + \epsilon)$. The projection identities assume nondegenerate, orthonormal directions, while numerical checks use the realized Gram matrix.

G.1 Coefficient Labels and Standardization

For a learned direction matrix $\mathbf{D}_c$, the coefficient target is the projection

$$\mathbf{s}_c^* = \mathbf{D}_c^\top \mathbf{r}_c, \qquad \mathbf{D}_c^\top \mathbf{D}_c = I. \tag{28}$$

The formal online replay computes $\mathbf{D}_c^\top \mathbf{r}_c$ directly from matured residuals without a ridge term. A least-squares target for a nonorthogonal control instead requires $\mathbf{D}_c^\dagger \mathbf{r}_c$; it is not interchangeable with the orthonormal projection target.

Let μ_k and σ_k denote training-split statistics of $s_{c,k}^*$. The gate is fitted to standardized labels with an L1 objective,

$$s_{c,k}^{*,\mathrm{std}} = \frac{s_{c,k}^* - \mu_k}{\sigma_k}, \qquad \mathcal{L}_{\mathrm{str}} = \frac{1}{MK} \sum_{c \in \mathcal{M}, k} \left| \hat{s}_{c,k}^{\mathrm{std}} - s_{c,k}^{*,\mathrm{std}} \right|, \tag{29}$$

where $\mathcal{M}$ indexes the M sample–channel pairs in a minibatch. Coefficients are restored to residual units through $\hat{s}_{c,k} = \mu_k + \sigma_k \hat{s}_{c,k}^{\mathrm{std}}$. The mapping g_ϕ in the main text includes this fixed inverse transformation. L1 is used as a robust coefficient-learning objective; the exact projection's least-squares optimality does not imply that this learned predictor attains minimum population MSE.

The feature map Φ contains context and base-forecast statistics, discrepancy summaries, channel metadata, and summaries of the predicted directions, all available before the target is revealed. Input and label standardizers are fitted on training data and held fixed for validation, static inference, and online adaptation. Direction checkpoints are selected before strength training, and the final reconstruction refinement starts from this stage-wise solution.

G.2 Cosine Direction Supervision

For $\mathbf{r} \neq \mathbf{0}$, let $\bar{\mathbf{r}} = \mathbf{r}/\|\mathbf{r}\|_2$ and consider the unit-direction loss $\ell(\mathbf{d}, \mathbf{r}) = 1 - \mathbf{d}^\top \bar{\mathbf{r}}$. If $\mathbb{E}[\bar{\mathbf{r}} \mid \mathbf{z}] \neq \mathbf{0}$, its conditional population minimizer is

$$\mathbf{d}^*(\mathbf{z}) = \frac{\mathbb{E}[\bar{\mathbf{r}} \mid \mathbf{z}]}{\|\mathbb{E}[\bar{\mathbf{r}} \mid \mathbf{z}]\|_2}. \tag{30}$$

Indeed, conditioning on $\mathbf{z}$ reduces the risk to a constant minus $\mathbf{d}^\top \mathbb{E}[\bar{\mathbf{r}} \mid \mathbf{z}]$; Cauchy–Schwarz attains its maximum at Eq. (30). Thus cosine supervision learns conditional residual orientation, with each nonzero sample contributing after normalization rather than in proportion to its residual norm. The same argument applies to the residual left after projecting previously learned directions.

G.3 Training and Online Adaptation Algorithm

Algorithm 1 PCRA training, forecasting, and optional test-time adaptation

Require: Frozen forecaster F; source data $\mathcal{S}$; horizon H; K directions; buffer size B; TTA rate η
Ensure: Corrected forecasts $\widetilde{\mathbf{y}}_t$; direction parameters ψ; strength parameters ϕ

$\mathbf{b} \leftarrow F(\mathbf{x})$, $\mathbf{r} \leftarrow \mathbf{y} - \mathbf{b}$, $\mathbf{u} \leftarrow [\mathcal{N}(\mathbf{x}); \mathcal{N}(\mathbf{b})]$ for $(\mathbf{x}, \mathbf{y}) \in \mathcal{S}$
Offline direction learning
for $k = 1, \ldots, K$ **do**
  $P_{<k} \leftarrow \sum_{j<k} \mathbf{d}_j \mathbf{d}_j^\top$, $\quad \mathbf{r}^{(k)} \leftarrow (I - P_{<k})\mathbf{r}$
  for each training minibatch $\mathcal{M} \subset \mathcal{S}$ **do**
    $\mathbf{d}_k \leftarrow \mathcal{N}\big((I - P_{<k}) f_{\psi_k}(\mathbf{u})\big)$
    $\mathcal{L}_{\mathrm{dir},k} \leftarrow \mathrm{mean}_{\mathcal{M}}[1 - \mathbf{d}_k^\top \mathcal{N}(\mathbf{r}^{(k)})]$
    $\psi_k \leftarrow \mathrm{AdamWStep}(\psi_k, \nabla_{\psi_k} \mathcal{L}_{\mathrm{dir},k})$
  end for
  Select ψ_k on validation loss; freeze ψ_k
end for
Offline strength learning
$\mathbf{D} \leftarrow [\mathbf{d}_1, \ldots, \mathbf{d}_K]$, $\quad \mathbf{s}^* \leftarrow \mathbf{D}^\top \mathbf{r}$, $\quad \mathbf{z} \leftarrow \Phi(\mathbf{x}, \mathbf{b})$
$(\boldsymbol{\mu}_s, \boldsymbol{\sigma}_s) \leftarrow \mathrm{MeanStd}_{\mathcal{S}}(\mathbf{s}^*)$, $\quad \mathbf{s}^{*,\mathrm{std}} \leftarrow (\mathbf{s}^* - \boldsymbol{\mu}_s) \oslash \boldsymbol{\sigma}_s$
for each training minibatch $\mathcal{M} \subset \mathcal{S}$ **do**
  $\hat{\mathbf{s}}^{\mathrm{std}} \leftarrow q_\phi(\mathbf{z})$, $\quad \mathcal{L}_{\mathrm{str}} \leftarrow \mathrm{mean}_{\mathcal{M},k} \, |\hat{s}_k^{\mathrm{std}} - s_k^{*,\mathrm{std}}|$
  $\phi \leftarrow \mathrm{AdamWStep}(\phi, \nabla_\phi \mathcal{L}_{\mathrm{str}})$
end for
Select ϕ on validation loss; define $g_\phi(\mathbf{z}) = \boldsymbol{\mu}_s + \boldsymbol{\sigma}_s \odot q_\phi(\mathbf{z})$
Forecasting and optional PCRA-T
$\mathcal{C} \leftarrow \varnothing$, $\quad \mathcal{B} \leftarrow \varnothing$ ▷ Pending windows and replay buffer
for forecast time $t = 1, 2, \ldots$ **do**
  $\mathbf{b}_t \leftarrow F(\mathbf{x}_t)$; compute $\mathbf{D}_t$ by Eq. (5); $\mathbf{z}_t \leftarrow \Phi(\mathbf{x}_t, \mathbf{b}_t)$
  $\widetilde{\mathbf{y}}_t \leftarrow \mathbf{b}_t + \mathbf{D}_t g_\phi(\mathbf{z}_t)$; emit $\widetilde{\mathbf{y}}_t$
  if PCRA-T is enabled **then**
    $\mathcal{C}[t] \leftarrow (\mathbf{z}_t, \mathbf{b}_t, \mathbf{D}_t)$
    $\mathcal{A}_t \leftarrow \{j \in \mathrm{keys}(\mathcal{C}) : j + H \le t\}$ ▷ Fully observed historical horizons
    if $\mathcal{A}_t \neq \varnothing$ **then**
      $\mathcal{B} \leftarrow \mathrm{Latest}_B(\mathcal{B} \cup \{(\mathbf{z}_j, \mathbf{b}_j, \mathbf{D}_j, \mathbf{y}_j) : j \in \mathcal{A}_t\})$
      Remove keys $\mathcal{A}_t$ from $\mathcal{C}$
      $\mathbf{s}_j^* \leftarrow \mathbf{D}_j^\top (\mathbf{y}_j - \mathbf{b}_j)$ for each sample–channel pair in $\mathcal{B}$
      $\mathcal{L}_{\mathrm{TTA}} \leftarrow H^{-1} \, \mathrm{mean}_{\mathcal{B}} \, \|g_\phi(\mathbf{z}_j) - \mathbf{s}_j^*\|_2^2$
      Reset optimizer; $\quad \phi \leftarrow \mathrm{AdamStep}_\eta(\phi, \nabla_\phi \mathcal{L}_{\mathrm{TTA}})$ ▷ One full-gate step
    end if
  end if
end for