HIRA: GRADIENT-FREE RESIDUAL RETRIEVAL FOR LONG-HORIZON TIME-SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Can a frozen time-series foundation model adapt to long-horizon forecasting without additional gradient training? We introduce Hidden-state Residual Adaptation (HIRA), which turns the backbone's own context representations into an index of historical forecast errors. Suffix-pooled final tokens retrieve observed residuals, and horizon-wise configurations convert them into corrections without training an encoder, retriever, or fusion network. Under chronological selection with memory and configurations frozen before test, HIRA yields positive mean gains on four backbones at both 336 and 720 steps. On Chronos-2, its agreement-gated variant improves all seven datasets at the frozen-protocol 720-step horizon, reducing MSE by 13.15% on average, compared with 12.61% for the tested matched SARAF-style control. The ungated recipe achieves long-horizon mean reductions of 4.54%, 7.00%, and 4.74% on TimesFM, Moirai 2.0, and Sundial, respectively. Mechanism controls and a recoverable-value analysis connect these gains to query-accessible historical errors and clarify transfer limits. These results establish residual memory as a competitive route to long-horizon adaptation without additional gradient training.

1 INTRODUCTION

Foundation forecasters such as Chronos, TimesFM, and Moirai learn reusable temporal structure from large corpora and can forecast a new series without task-specific training (Ansari et al., 2024; Das et al., 2024; Woo et al., 2024). Yet a strong pretrained forecast can still leave systematic local errors: recurring phases, level transitions, or volatility states can reveal target-specific error patterns in the observed history. These completed historical forecasts provide an additional source of adaptation information, beyond the patterns encoded in the pretrained model. Figure 3 illustrates this opportunity: historical residual retrieval can refine peak amplitudes and local waveforms while preserving the backbone's broad temporal pattern, without updating its parameters. This motivates a gradient-free approach to long-horizon adaptation that reuses the frozen forecaster's own representations and observed errors.

Retrieval-augmented forecasting finds historical analogs and incorporates their continuations into a forecaster (Tire et al., 2024; Ning et al., 2025). However, similar contexts need not imply compatible forecast errors, especially under nonstationarity (Zhou et al., 2026). Learned retrieval and fusion can introduce additional trainable components; HIRA instead reuses representations already computed by the frozen forecaster. The key distinction is therefore not simply which historical future resembles the current task, but which historical errors can usefully correct the same frozen model. We ask: *can a frozen forecaster's own context representations retrieve historical errors that improve long-horizon predictions without additional gradient training?*

We introduce HIRA (Hidden-state Residual Adaptation), a gradient-free adapter that turns a frozen forecaster's own context representations into a *model-specific error index*. Suffix-pooled final context tokens serve as retrieval keys for fully observed residuals from completed historical forecasts. Rather than retrieving a replacement future, HIRA retrieves how the same frozen model erred in related contexts and aggregates those errors into an additive correction. Because useful historical corrections may differ across forecast ranges, HIRA selects retrieval views and correction strengths separately for different horizon blocks. These configurations are selected using purged chronological validation and frozen before testing; retrieval uses only previously revealed outcomes, with no additional

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.

gradient-trained encoder, retriever, or fusion network. We evaluate this interface across TimesFM, Chronos-2, Moirai 2.0, and Sundial.

Our contributions are:

- **Model-specific residual retrieval from frozen context tokens.** We introduce a gradient-free adaptation interface that pools the last r tokens of a backbone's final-layer context states to index its own historical forecast errors. Rather than retrieving future trajectories for learned fusion, HiRA retrieves residuals and configures their corrections separately across forecast-horizon blocks, without training an additional encoder, retriever, or fusion network.
- **Cross-backbone long-horizon adaptation without additional gradient training.** Under a frozen chronological protocol, HiRA achieves positive mean gains across seven datasets on each of four backbones at both 336 and 720 steps (Tables 1, 2, and 16). Its agreement-gated Chronos-2 variant improves all seven datasets at the frozen-protocol 720-step horizon, reducing MSE by 13.15% on average. The same ungated residual-retrieval recipe also yields positive long-horizon mean gains on TimesFM, Moirai 2.0, and Sundial.
- **Mechanistic evidence and a theory of recoverable correction.** Ablations examine the roles of frozen representations, suffix-token spans, and horizon-wise configuration, alongside comparisons with fitted adapters. A recoverable-value decomposition distinguishes hindsight-oracle headroom from correction gains accessible through the query, while a two-world lower bound characterizes limits of historical-error transfer.

2 Related Work

Foundation models and adaptation. TimesFM, Chronos, and Moirai establish pretrained forecasting across domains (Das et al., 2024; Ansari et al., 2024; Woo et al., 2024); Chronos-2 extends this paradigm to multivariate and covariate-informed forecasting with in-context information sharing (Ansari et al., 2025). Moirai 2.0 uses quantile-based forecasting, while Sundial models continuous futures through flow matching (Liu et al., 2025a;b). Tiny Time Mixers offer a compact pretrained alternative (Ekambaram et al., 2024). In-context fine-tuning explicitly trains a backbone to exploit related series as inference-time examples (Faw et al., 2025), while LoRA learns low-rank weight updates (Hu et al., 2022). For deployment adaptation, TAFAS uses partially observed targets and gated calibration (Kim et al., 2025); ELF learns a lightweight online forecaster and combines it with foundation-model predictions (Lee et al., 2025); ORCA models contextual errors through black-box online adaptation (Dai et al., 2026). TimeRouter instead routes among multiple foundation models (Ning et al., 2026). HiRA explores a complementary adaptation paradigm: its frozen internal representations access historical error memory without gradient-training an encoder, retriever, or fusion network. Our evaluation freezes this memory and its configuration before test, distinguishing it from feedback-driven online adaptation.

Retrieval-augmented and training-free forecasting. Retrieval in pretrained representation spaces has precedents in nearest-neighbor language models (Khandelwal et al., 2020). In forecasting, RAF retrieves historical contexts and continuations (Tire et al., 2024); TS-RAG combines a pretrained retrieval encoder with a learned projector and adaptive retrieval mixer (Ning et al., 2025); TimeRAF combines a learned retriever with channel prompting (Zhang et al., 2024). Reference-token augmentation provides another integration route (Xiao et al., 2026). SARAF balances temporal relevance and diversity using stationarity-aware retrieval and aggregation (Zhou et al., 2026). Without gradient-based fitting, KReF uses handcrafted or frozen random-feature embeddings to retrieve historical futures, forming an empirical predictive distribution and a weighted-mean forecast (Zhang & Su, 2026). Relative to these continuation-based approaches, HiRA shifts the retrieval target to model-specific forecast errors: the backbone's unmodified hidden states index residual memory, supplying output-level corrections rather than a replacement future trajectory or learned representation-level fusion. Unlike prompt-token alignment for forecasting in TimeCMA (Liu et al., 2024), HiRA uses native frozen forecast-context tokens to retrieve model-specific historical errors for horizon-wise correction.

Multi-step error correction and nonparametric calibration. Forecast combination and stacking combine predictive information (Bates & Granger, 1969; Wolpert, 1992), while Rectify and sub-

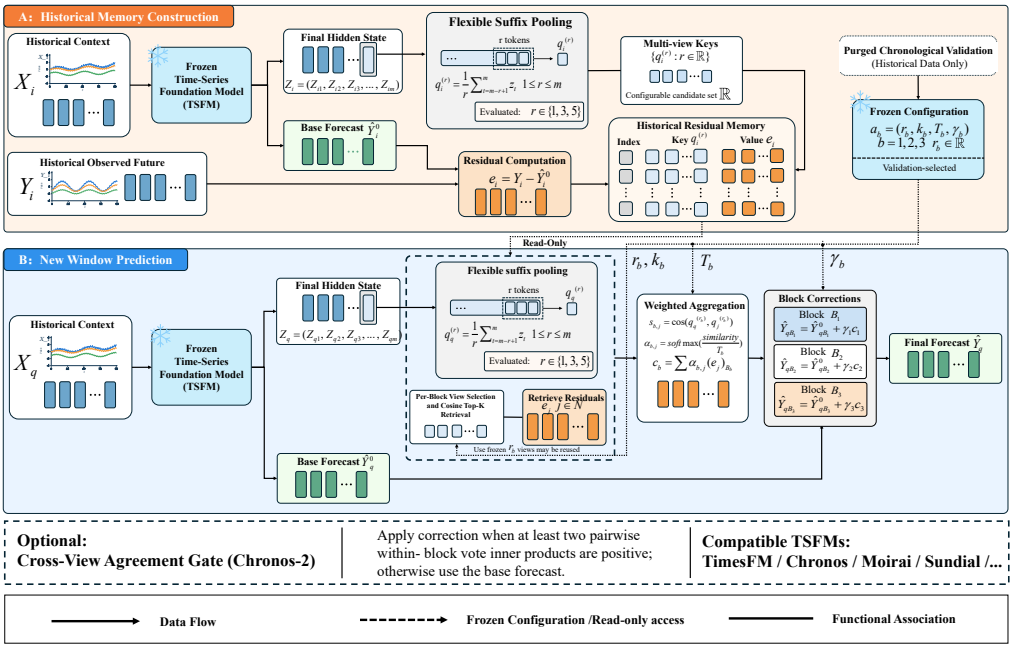


Figure 1: **Overview of HIRA: adapting frozen forecasts without gradient training.** HIRA uses a frozen backbone both to forecast and to encode the query. Historical residuals are indexed by suffix-pooled final context states; cosine-similarity softmax-weighted residuals correct three contiguous equal-length horizon blocks. Configurations are validation-selected and frozen before test; test targets never enter memory or configuration selection. In the diagram, c_b abbreviates $c_{a_b}(q)_{B_b}$. Input curves are schematic; the optional agreement gate is used only in the Chronos-2 study.

sequent multi-step boosting adjust base forecasts with horizon-specific error models (Ben Taieb & Hyndman, 2012; 2014). Our similarity-weighted residual vote connects to local kernel regression (Nadaraya, 1964), with neighborhoods defined by frozen forecasting representations rather than an additional fitted encoder. HIRA couples this hidden-state error memory with horizon-wise correction, allowing different forecast ranges to use distinct retrieval configurations. Its recoverable-value analysis further distinguishes available correction headroom from gains identifiable through the query, connecting the adaptation mechanism to its transfer limits.

3 METHOD

3.1 FROZEN BACKBONE AND RESIDUAL MEMORY

Let $X_i \in \mathbb{R}^L$ be a context, $Y_i \in \mathbb{R}^H$ its future, and F a frozen foundation forecaster. Its point forecast is $\widehat{Y}_i^0 = F(X_i)$. The final Transformer layer produces patch states $Z_i = (z_{i1}, \ldots, z_{im}) \in \mathbb{R}^{m \times d}$. For the evaluated candidate set, we define

$$q_i^{(r)} = \frac{1}{r} \sum_{t=m-r+1}^{m} z_{it}, \qquad r \in \{1, 3, 5\}. \tag{1}$$

The same pooling operation is defined for any integer $1 \le r \le m$; all experiments restrict the candidate set to $\{1, 3, 5\}$. With $L = 160$, TimesFM-2.5's native patch length 32 gives $m = 5$ and $d = 1280$. These tensors are captured from the first original, non-sign-flipped input context prefill. For $H > 128$, later autoregressive states and the flip-invariance pass are deliberately excluded.

After a historical target is observed, memory stores $(q_j^{(r)}, e_j)$ where $e_j = Y_j - \widehat{Y}_j^0$. For a new query q, exact cosine search returns indices $\mathcal{N}_k(q)$. Attention weights and the retrieved correction are

$$s_j = \frac{\langle q, q_j \rangle}{\|q\|\|q_j\|}, \qquad \alpha_j = \frac{\exp(s_j/T)}{\sum_{u \in \mathcal{N}_k(q)} \exp(s_u/T)}, \tag{2}$$

$$c(q) = \sum_{j \in \mathcal{N}_k(q)} \alpha_j e_j, \qquad \widehat{Y}(q) = \widehat{Y}^0(q) + \gamma c(q), \tag{3}$$

where $k \in \{4, 16, 64\}$, $T \in \{0.03, 0.1, 0.3\}$, and $\gamma \in \{0, 0.25, 0.5, 1\}$. Including $\gamma = 0$ makes the unchanged backbone forecast an exact candidate.

3.2 Horizon-wise correction

Let B_1, B_2, B_3 partition $\{1, \ldots, H\}$ into contiguous equal blocks. Each block chooses one configuration $a_b = (r_b, k_b, T_b, \gamma_b)$, yielding

$$\widehat{Y}_{B_b} = \widehat{Y}^0_{B_b} + \gamma_b c_{a_b}(q)_{B_b}. \tag{4}$$

This permits near-term and far-term errors to use different temporal summaries and memories without training a horizon router. A global control constrains $a_1 = a_2 = a_3$.

3.3 Chronological selection and label isolation

Validation is divided into thirds. The middle and final thirds are two selection folds; for each fold, memory contains training episodes and only validation episodes at least H start positions earlier. A configuration is eligible only if its relative MSE gain is nonnegative in both folds, after which the largest mean-gain configuration is selected. If none improves, the zero-shrinkage candidate is used. The final test memory contains training and validation only; no test residual is added to this memory. For context starts s_j and s_i, a historical episode is available at the query origin only if $s_j + L + H - 1 < s_i + L$, equivalently $s_j \le s_i - H$. This constrains label availability; it does not require nonoverlapping rolling contexts. The prediction API takes hidden queries and historical residuals but has no test-target argument.

3.4 Label-free cross-view agreement

The primary TimesFM results use ungated HiRA. The following extension, agreement-gated HiRA, is evaluated in the Chronos-2 transfer study. Validation stability is an aggregate criterion and may not detect an unreliable individual query. For each selected block configuration, we therefore recompute the residual vote using last-1, last-3, and last-5 queries while keeping its neighbor count and temperature fixed. Let these three votes be $c^{(1)}_{B_b}, c^{(3)}_{B_b}, c^{(5)}_{B_b}$, restricted to block B_b. We apply the selected correction only when at least two of the three within-block pairwise cosines are positive (zero-norm votes do not count):

$$g_b(q) = \mathbf{1}\left\{ \sum_{u<v} \mathbf{1}[\langle c^{(u)}_{B_b}, c^{(v)}_{B_b} \rangle > 0] \ge 2 \right\}, \qquad \widehat{Y}_{B_b} = \widehat{Y}^0_{B_b} + g_b(q)\gamma_b c_{a_b}(q)_{B_b}. \tag{5}$$

The gate observes only the query and residuals of already revealed historical episodes. It has no learned parameters, target input, or test-selected threshold; disagreement returns the exact backbone forecast on that block.

3.5 What retrieval can identify

For a representation $Q = \phi(X)$ and base residual $E = Y - F(X)$, squared loss is minimized by the conditional correction $c^*(Q) = \mathbb{E}[E \mid Q]$. Thus hidden-state retrieval is useful only when the representation preserves information about the backbone's future error.

Let $R(c) = \mathbb{E}\|F(X) + c(Q) - Y\|^2 = \mathbb{E}\|c(Q) - E\|^2$ and let $c_{\text{or}} = E$ be the sample-wise hindsight oracle. Define oracle headroom $V_{\text{or}} = R(0) - R(c_{\text{or}})$ and query-identifiable headroom $V_{\text{id}} = R(0) - \inf_{c \in \mathcal{C}_Q} R(c)$, where $\mathcal{C}_Q$ contains all square-integrable corrections $c(Q, U)$ and U is independent algorithmic randomness (deterministic rules are included).

Proposition 1 (Recoverable-value decomposition). *Assume $\mathbb{E}\|E\|^2 < \infty$ and write $m(Q) = \mathbb{E}[E \mid Q]$. Then the Bayes correction is $c^*(Q) = m(Q)$ and*

$$V_{\mathrm{id}} = \mathbb{E}\|m(Q)\|^2, \tag{6}$$

$$V_{\mathrm{or}} - V_{\mathrm{id}} = \mathbb{E}\|E - m(Q)\|^2 = \mathbb{E}\ \mathrm{tr}\big(\mathrm{Cov}(E \mid Q)\big). \tag{7}$$

Moreover every $c \in \mathcal{C}_Q$ satisfies

$$R(c) - R(c^*) = \mathbb{E}\|c(Q,U) - m(Q)\|^2. \tag{8}$$

Thus oracle headroom is recoverable exactly to the extent that the query resolves the conditional mean residual; conditional residual variance is an irreducible oracle gap.

The result follows by conditioning on Q and applying the Pythagorean identity in L_2; independent router randomization cannot reduce the second term. A complete derivation is given in Appendix A.

Theorem 1 (Query aliasing). *For every $v > 0$, there is a squared-loss forecasting problem with oracle headroom v in which every possibly randomized correction measurable from Q has zero identifiable gain and regret at least v to the oracle.*

The proof is given in Appendix A.1. The negative-transfer cells in our results do not prove complete aliasing, but they demonstrate why validation fallback is not a formal safety guarantee.

The preceding construction averages ambiguity within one distribution. The following result gives the stronger two-process statement: even the entire revealed history can be identical while the optimal future correction changes sign.

Theorem 2 (Indistinguishable deployment worlds). *For every $v > 0$, there exist two deployment processes P_+ and P_- that induce the same law for all information D available before prediction, including the revealed training and validation history and the current query, but whose optimal corrections have opposite signs. For any possibly randomized decision rule $C = A(D, U)$, where U is independent algorithmic randomness,*

$$\max_{s \in \{-1,+1\}} \mathbb{E}_{P_s}\big[(C - s\sqrt{v})^2\big] \geq v. \tag{9}$$

The world-aware oracle has zero loss in both processes, so every rule incurs oracle regret at least v in at least one observationally indistinguishable world.

The proof is given in Appendix A.2.

4 Experimental Protocol

Backbone and data. The primary study uses the official TimesFM-2.5-200M PyTorch checkpoint. We evaluate the univariate target of ETTh1, ETTh2, ETTm1, ETTm2, Weather, Electricity, and Traffic with their chronological TSLib boundaries and $H \in \{96, 192, 336, 720\}$. Every valid within-split rolling origin is evaluated at stride one; training memory retains the latest 8,192 origins. Weather's 50 missing-value sentinels (-9999) are handled by causal forward filling, with a training-only median fallback for leading missing values.

Development and confirmation. H=96 is explicitly a development horizon because its test output was inspected during construction. We froze the representation set, search grids, three blocks, two-fold stability rule, and zero fallback before evaluating H=192/336/720. Those 21 cells are frozen-protocol cross-horizon evaluations. Here, confirmation means that the horizon-specific results were not used to select the frozen recipe. These evaluations reuse previously inspected datasets, with potentially overlapping target timestamps across horizons; confirmation therefore refers to the frozen protocol rather than independent datasets.

After completing the TimesFM study, we run a cross-backbone check with the official 120M-parameter Chronos-2 checkpoint (Ansari et al., 2025). The initial $H \in \{96, 336\}$ matrices are development evidence used to diagnose negative transfer and define the agreement gate. We then freeze that gate and its execution protocol before opening $H \in \{192, 720\}$; these 14 cells form the frozen-protocol cross-horizon evaluation set. We retain the same 160-point context, candidate

views, retrieval grid, three-block rule, validation folds, and exact-zero fallback. Chronos-2 has 16-point patches. We pool only its final-encoder observed-context patches; the regression token and all masked forecast-query patches are excluded. All four horizons are reported, with their development/confirmation status kept explicit.

We additionally test channel breadth at $H = 96$. Before loading values or targets, we predeclare every numeric channel when a dataset has at most 32 channels (all 49 channels across ETTh1/2, ETTm1/2, and Weather). For Electricity and Traffic, we rank names by SHA-256 of dataset and column and take the first 12 from each, giving 73 series total. Each series uses the same frozen method and matched memory/evaluation sizes: 8,192 training-memory origins and 2,048 validation and test origins. Channel identity is therefore independent of forecast quality; no test target enters selection or memory.

Additional backbone transfer. We apply the unchanged ungated recipe to Moirai 2.0 (all four horizons) and Sundial ($H = 336, 720$), keeping all seven target series, $L = 160$, and chronological origins. Their native 16-point patches yield last-1/3/5 views over 16/48/80 input positions; hidden states can encode earlier context. Moirai uses its official median forecast, and Sundial uses a fixed 20-sample mean. Within each backbone, all adapters share identical cached base forecasts; configurations use train/validation data, and predictions are frozen before separate scoring. These experiments assess cross-backbone transfer on the existing benchmarks, including previously inspected Moirai results on ETTh1, ETTh2, and Weather. Full protocols and intervals are in Appendix J.

Controls and statistics. We compare unchanged TimesFM; one fixed token span at a time; validation selection among last-1/3/5; z-normalized raw-context retrieval with the same search grid; a query-independent historical mean-residual correction; and a single global correction. At all horizons, stronger controls include exact Cholesky-solved, standardized multi-output ridge regression from genuine hidden states to future residuals (Hoerl & Kennard, 1970), and blockwise Hedge over delayed-feedback rolling residual experts (Freund & Schapire, 1997). Their hyperparameters use the same purged validation folds. Ridge is frozen before test. Hedge receives a strictly more permissive causal protocol: a test outcome updates later predictions only after its full horizon is observable, never before. Relative gain is $(\mathrm{MSE}_0 - \mathrm{MSE})/\mathrm{MSE}_0$. Confidence intervals use a paired circular moving-block bootstrap with block length H and 2,000 draws. We additionally test block lengths $H/2$ and $2H$, and treat the dataset—not the horizon cell—as the independent unit in a second bootstrap and an exact sign-flip test.

5 RESULTS

We test whether frozen residual memory supports competitive long-horizon adaptation across backbones, then examine its design choices and learned controls. Development/confirmation splits remain unchanged. The controlled recoverability experiment is reported in Appendix K.

5.1 CROSS-HORIZON ACCURACY

Table 1 reports the horizon-wise adapter. It wins 20/28 cells, with 5.75% mean and 3.15% median improvement. Fifteen cells have a strictly positive 95% interval. Every horizon is positive on average, including all three confirmation horizons; those horizons contain 15 wins in 21 cells.

Qualitative forecast refinement. Figure 3 illustrates peak and waveform refinement without updating the backbone. The complete seven-dataset visualization and its selection protocol are provided in Appendix H.

5.2 LONG-HORIZON ADAPTATION ON CHRONOS-2

Long-horizon adaptation. Agreement-gated HIRA improves Chronos-2 on all seven datasets at both $H = 336$ and $H = 720$ (Table 2). At the frozen-protocol $H = 720$ horizon, it achieves a 13.15% mean relative MSE reduction, with strictly positive moving-block 95% intervals on five datasets and a maximum reduction of 24.34% on ETTm1. These results highlight the potential

Table 1: Relative MSE gain (%) of horizon-wise hidden residual retrieval over the identical TimesFM forecast. H=96 is development evidence; H=192/336/720 are frozen-rule confirmation evidence. * marks a strictly positive moving-block 95% confidence interval.

Dataset	96	192	336	720	Mean
ETTh1	-4.999	-1.338	-5.967	0.552	-2.938
ETTh2	0.727	-2.610	-1.531	11.639*	2.056
ETTm1	5.842*	2.765*	0.810	0.473	2.473
ETTm2	33.648*	26.505*	19.862*	13.949*	23.491
Weather	-2.810	0.867	-0.831	-1.206	-0.995
Electricity	12.214*	9.069*	8.485*	3.535*	8.326
Traffic	9.844*	7.708*	8.168*	5.559*	7.820
Macro mean	7.781	6.138	4.142	4.929	5.747
Wins	5/7	5/7	4/7	6/7	20/28

Table 2: Agreement-gated relative MSE gain (%) over Chronos-2. H=96/336 are development diagnostics; H=192/720 are frozen-protocol confirmation cells. * marks a strictly positive moving-block 95% interval; † marks a strictly negative one.

Dataset	96	192	336	720	Mean
ETTh1	-23.719†	-9.632	5.137	10.240	-4.494
ETTh2	4.214*	1.698	6.797	11.352*	6.015
ETTm1	9.569*	10.735*	14.478*	24.335*	14.779
ETTm2	16.462*	12.065*	8.428	15.582*	13.134
Weather	7.330*	5.505*	4.852*	3.270	5.239
Electricity	28.968*	22.230*	20.779*	20.625*	23.150
Traffic	5.643*	4.443*	4.866*	6.654*	5.401
Macro mean	6.924	6.721	9.334	13.151	9.032
Wins	6/7	6/7	7/7	7/7	26/28

of historical residual memory for long-horizon adaptation without updating backbone parameters; $H = 336$ remains development evidence.

Table 2 applies agreement-gated HIRA to Chronos-2. On the frozen-protocol $H = 192, 720$ cells it wins 13/14, with 9.94% mean and 10.49% median gain. The complete matrix wins 26/28 and averages 9.03%; all four horizon means are positive. Twenty cells have a strictly positive moving-block interval. These results extend the evidence beyond a single backbone. Performance remains dataset-dependent: ETTh1 has a significant -23.72% gain at $H = 96$, while improving at $H = 336, 720$. Over the full four-horizon matrix, exact hidden ridge and HIRA have nearly identical mean gains (9.04% versus 9.03%), with 25 and 26 positive cells, respectively; raw retrieval and constant calibration reach 5.34% and 1.94%.

5.3 LONG-HORIZON TRANSFER BEYOND TWO BACKBONES

Table 16 tests whether the same suffix-token error memory remains useful beyond TimesFM and Chronos-2. Without adding the agreement gate or changing its candidate grid, HIRA improves all 14 Moirai 2.0 long-horizon settings: mean MSE reduction rises from 5.73% at $H = 336$ to 8.27% at $H = 720$. Its 7.00% long-horizon mean exceeds constant calibration (1.42%) and hidden ridge (4.76%), winning 14/14 and 11/14 paired settings, respectively. Six individual gain intervals are strictly positive. This supports portability of the error-memory interface, not only the Chronos-2 agreement-gated variant.

On Sundial, HIRA improves 10/14 settings and attains the highest long-horizon macro mean among these four adapters: 4.74%, versus 3.01% for constant calibration, 1.94% for ridge, and 2.99% for the learned Future Retrieval Mixer (FRM). The evidence is complementary: Moirai supplies consistent positive long-horizon transfer, while Sundial supplies competitive mean performance against fitted controls. The comparison with fitted fusion is backbone-dependent: HiRA achieves a higher long-horizon mean gain on Sundial (4.74% versus 2.99% for FRM), whereas FRM achieves

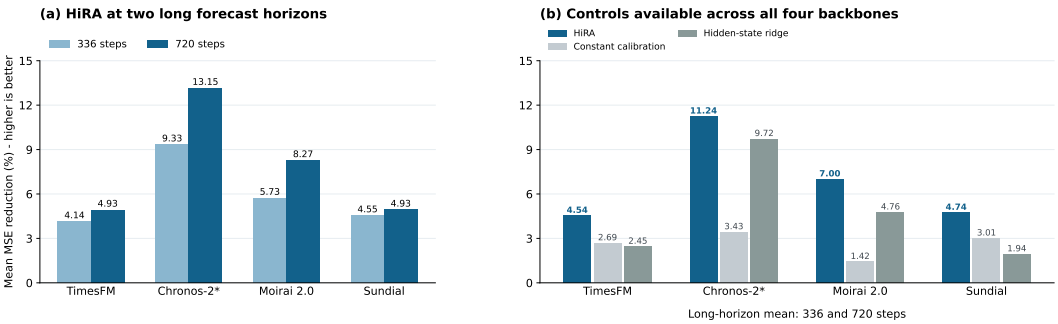


Figure 2: **Gradient-free long-horizon adaptation across four frozen backbones.** (a) Seven-dataset macro-mean MSE reductions at $H = 336$ and 720. (b) Mean gains averaged equally over both horizons, comparing HiRA with constant residual calibration and hidden-state ridge, the controls available across all four backbones. Gains are relative to each adapter's own identical frozen backbone; higher is better. Raw-input retrieval remains in the ablation results (Table 3). *Chronos-2 uses agreement-gated HiRA and ungated controls (336: development; 720: frozen-protocol confirmation); other HiRA variants are ungated. These are descriptive mean comparisons, not tests of statistical superiority or gate-matched ablations.

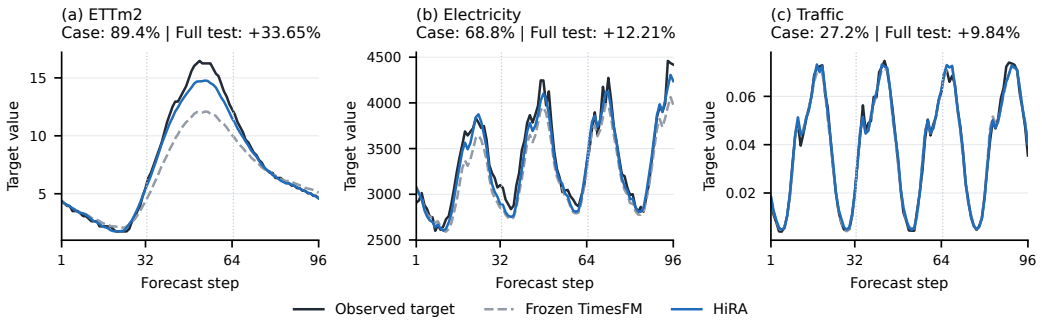


Figure 3: **HiRA in Action: Trend, Peak, and Waveform Refinement.** On ETTm2, residual retrieval reduces peak underestimation; on Electricity, it improves the amplitudes of recurring peaks; on Traffic, it refines an already strong periodic forecast. Black, dashed gray, and blue denote observations, frozen TimesFM, and HiRA. Each panel shows the full 96-step horizon in original units; percentages report window-level and full-test MSE reductions separately. Examples illustrate these improvements; selection details appear in Appendix H.

a higher mean on Moirai (11.00% versus 7.00% for HiRA). Thus the results support a transferable gradient-free residual interface rather than uniform superiority over learned future fusion. Mean rankings are descriptive; Appendix J retains every cell and protocol difference.

Figure 2 summarizes HiRA's long-horizon adaptation across four frozen backbones. HiRA yields positive mean gains at both horizons and exceeds constant residual calibration and hidden-state ridge in long-horizon mean gain on each backbone.

5.4 Representation and Horizon-wise Correction

Horizon-wise correction wins 20 cells with a 5.75% mean gain, compared with 19 cells and 5.38% for a single global configuration. The benefit varies across horizons: it is largest at $H = 720$, while the global configuration is ahead by 1.11 percentage points at $H = 336$.

Hidden retrieval outperforms raw-input retrieval in 20/28 paired cells, with a mean gain difference of 0.80 percentage points. Span selection adds 0.21 percentage points over last-1, while last-3 has the highest median gain among fixed spans. Constant residual calibration achieves a 1.54% mean

Table 3: Full 28-cell ablation. All retrieval variants use identical memories, grids, chronological selection, and horizon blocks.

Representation or control	Wins	Mean gain	Median gain
Selected last-1/3/5	20	5.747	3.150
Last-1 only	19	5.535	2.513
Last-3 only	20	4.324	3.544
Last-5 only	19	3.802	2.346
Raw-input RAG	19	4.945	3.847
Constant residual calibration	20	1.535	0.285

gain, compared with 5.75% for hidden retrieval. This comparison supports query-dependent residual adaptation over the tested query-independent calibration, while the representation and span controls assess specific design choices.

An illustrative retrieval trace shows how historical residuals yield a correction aligned with peak underestimation, refining forecast amplitude without updating the backbone (Appendix H.1).

5.5 Learned and delayed-online baselines

Exact multi-output hidden ridge improves 22/28 cells by 3.82%; HiRA wins their paired comparison in 21/28 and leads by 1.92 points (16/21 and 2.02 points on confirmation horizons). Delayed blockwise Hedge, allowed to update only after each test target is fully observable, improves 2/28 cells by 0.18%. Table 17 gives the full horizon summary.

5.6 Competitiveness without gradient training

Agreement-gated HiRA approaches the tested SARAF-style control on Chronos-2 at $H = 336$ (9.33% versus 9.59%) and slightly exceeds it at $H = 720$ (13.15% versus 12.61%) without additional gradient training. These are matched mechanism controls, not official checkpoint reproductions; evaluation origins and label availability are shared, while fitting and context augmentation differ. On TimesFM, the matched learned-fusion control achieves a higher full-matrix mean gain (9.71% versus 5.75% for HiRA), indicating that relative performance depends on the backbone and adaptation mechanism. Tables 18 and 8 retain all horizons and training distinctions. Expanded multi-series and uncertainty audits are reported in Appendices B and C.

6 Discussion and Limitations

Frozen forecasting representations provide a practical interface for gradient-free adaptation, with positive long-horizon mean gains across four backbones. Performance varies across datasets and comparators; aggregate rankings are descriptive. Chronos-2 uses agreement-gated HiRA, so comparisons with ungated controls assess the complete adapter rather than retrieval alone. The appendices report detailed comparisons and uncertainty analyses. Future work will examine when historical residuals remain transferable under distribution shift and when corrections should be suppressed.

7 Conclusion

HiRA shows how frozen forecasting representations can support long-horizon adaptation through model-specific historical residual memory. Suffix-pooled context tokens and horizon-wise corrections yield positive mean gains across four backbones at both long horizons without training an additional retriever or fusion network. At 720 steps, agreement-gated Chronos-2 improves all seven datasets by 13.15% on average and slightly exceeds the tested matched SARAF-style retrieval control in mean gain. Together with the mechanism controls and recoverable-value analysis, these findings establish a practical role for frozen context tokens beyond forecast generation: retrieving historical errors to refine future predictions.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Turkmen, et al. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

John M. Bates and Clive W. J. Granger. The combination of forecasts. *Journal of the Operational Research Society*, 20(4):451–468, 1969.

Souhaib Ben Taieb and Rob J. Hyndman. Recursive and direct multi-step forecasting: The best of both worlds. Working paper, Monash University, 2012. URL https://robjhyndman.com/papers/rectify.pdf.

Souhaib Ben Taieb and Rob J. Hyndman. Boosting multi-step autoregressive forecasts. In *International Conference on Machine Learning*, pp. 109–117, 2014. URL https://proceedings.mlr.press/v32/taieb14.html.

Xilin Dai, Yiding Liu, Hongjie Xia, Yifan Hu, Zewei Dong, Jiang-Ming Yang, and Qiang Xu. Learning the context of errors: Black-box online adaptation of time series foundation models. *arXiv preprint arXiv:2606.14222*, 2026.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *International Conference on Machine Learning*, 2024.

Vijay Ekambaram, Arindam Jati, Pankaj Dayama, Sumanta Mukherjee, Nam H. Nguyen, Wesley M. Gifford, Chandra 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*, 2024. URL https://arxiv.org/abs/2401.03955.

Matthew Faw, Rajat Sen, Yichen Zhou, and Abhimanyu Das. In-context fine-tuning for time-series foundation models. In *International Conference on Machine Learning*, pp. 16355–16374, 2025. URL https://proceedings.mlr.press/v267/faw25b.html.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1911.00172.

HyunGi Kim, Siwon Kim, Jisoo Mok, and Sungroh Yoon. Battling the non-stationarity in time series forecasting via test-time adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(17):17868–17876, 2025. doi: 10.1609/aaai.v39i17.33965.

Thomas L. Lee, William Toner, Rajkarn Singh, Artjom Joosen, and Martin Asenov. Lightweight online adaption for time series foundation model forecasts. In *International Conference on Machine Learning*, pp. 33736–33764, 2025. URL https://proceedings.mlr.press/v267/lee25ag.html.

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. URL https://arxiv.org/abs/2511.11698.

Chenxi Liu, Qianxiong Xu, Hao Miao, Sun Yang, Lingzheng Zhang, Cheng Long, Ziyue Li, and Rui Zhao. Timecma: Towards llm-empowered multivariate time series forecasting via cross-modality alignment. *arXiv preprint arXiv:2406.01638*, 2024. URL https://arxiv.org/abs/2406.01638. Accepted at AAAI 2025.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. *arXiv preprint arXiv:2502.00816*, 2025b. URL https://arxiv.org/abs/2502.00816.

E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964. doi: 10.1137/1109020.

Kanghui Ning, Zijie Pan, Yu Liu, Yushan Jiang, James Y. Zhang, Kashif Rasul, Anderson Schneider, Lintao Ma, Yuriy Nevmyvaka, and Dongjin Song. TS-RAG: Retrieval-augmented generation based time series foundation models are stronger zero-shot forecaster. In *Advances in Neural Information Processing Systems*, 2025.

Kanghui Ning, Yushan Jiang, Kashif Rasul, Anderson Schneider, Yuriy Nevmyvaka, and Dongjin Song. Timerouter: Efficient and adaptive routing of time-series foundation models. *arXiv preprint arXiv:2606.11625*, 2026.

Kutay Tire, Ege Onur Taga, Muhammed Emrullah Ildiz, and Samet Oymak. Retrieval augmented time series forecasting. *arXiv preprint arXiv:2411.08249*, 2024.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *International Conference on Machine Learning*, 2024.

Yixiong Xiao, Congxi Xiao, and Jingbo Zhou. TS-RAG: Retrieval augmented generation for time series forecasting. *arXiv preprint arXiv:2608.06223*, 2026.

Huanyu Zhang, Chang Xu, Yi-Fan Zhang, Zhang Zhang, Liang Wang, Jiang Bian, and Tieniu Tan. TimeRAF: Retrieval-augmented foundation model for zero-shot time series forecasting. *arXiv preprint arXiv:2412.20810*, 2024. URL https://arxiv.org/abs/2412.20810.

Yang Zhang and Rui Su. KReF: Training-free retrieval for long-term time-series forecasting and predictive uncertainty. *arXiv preprint arXiv:2608.06748*, 2026. URL https://arxiv.org/abs/2608.06748.

Shiqiao Zhou, Holger Schöner, Zipeng Wu, Edouard Fouché, IAG Wilson, and Shuo Wang. Stationarity-aware retrieval-augmented time series forecasting. *arXiv preprint arXiv:2606.04135*, 2026.

AI USE STATEMENT

No generative AI tools were used in conducting this research or preparing the manuscript and supplementary materials.

APPENDIX

A PROOF OF THE RECOVERABLE-VALUE DECOMPOSITION

Let $M = \mathbb{E}[E \mid Q]$. For a deterministic query-measurable correction, conditioning on Q gives

$$\mathbb{E}[\|c - E\|^2 \mid Q] = \|c - M\|^2 + \mathbb{E}[\|E - M\|^2 \mid Q], \tag{10}$$

because the cross term contains $\mathbb{E}[E - M \mid Q] = 0$. The same identity holds for a randomized rule after also conditioning on its internal randomness, assumed independent of E given Q. Therefore $c^* = M$ minimizes risk and

$$\inf_{c \in \mathcal{C}_Q} R(c) = \mathbb{E}\|E - M\|^2. \tag{11}$$

Applying the same orthogonal decomposition to $E = M + (E - M)$ yields $R(0) = \mathbb{E}\|M\|^2 + \mathbb{E}\|E - M\|^2$. The sample-wise oracle has zero risk, hence $V_{\text{or}} = R(0)$, $V_{\text{id}} = \mathbb{E}\|M\|^2$, and their difference is the expected trace of the conditional covariance. Subtracting $R(c^*)$ from the first identity gives the excess-risk statement.

A.1 PROOF OF QUERY ALIASING

Proof. Let Q be constant and let the scalar future residual be $-\sqrt{v}$ or $+\sqrt{v}$ with equal probability. Then $m(Q) = 0$, so the proposition gives $V_{\text{id}} = 0$, $V_{\text{or}} = v$, and $R(c) \geq v$ for every possibly randomized query-measurable correction. Its regret to the zero-risk hindsight oracle is therefore at least v. □

A.2 PROOF OF INDISTINGUISHABLE DEPLOYMENT WORLDS

Proof. Choose identical deterministic D under both processes and let the deployment residual be $E = s\sqrt{v}$ under P_s. Because D has the same law and U is independent, C has the same distribution in both worlds. Averaging their risks gives

$$\frac{1}{2}\mathbb{E}[(C - \sqrt{v})^2 + (C + \sqrt{v})^2] = \mathbb{E}[C^2] + v \geq v. \tag{12}$$

At least one risk is therefore at least v, while choosing $C = s\sqrt{v}$ after observing the world has zero loss. □

B EXPANDED MULTI-SERIES CHECK

The target-channel matrix could overstate breadth if its seven conventional targets happened to favor adaptation. Table 4 reports the predeclared expanded check. HIRA improves 53/73 series, has 5.21% series-macro mean and 3.41% median gain, and improves six of seven dataset means. Twenty-five series have a strictly positive 95% interval, four have a strictly negative interval, and five invoke the exact zero correction.

Table 4: Matched $H = 96$ multi-series check. All channels are used for datasets with at most 32; 12 channels are hash-selected from each larger dataset.

Dataset	Mean gain	Median gain	Wins
ETTh1	2.752	0.292	4/7
ETTh2	-0.839	0.000	1/7
ETTm1	2.744	2.016	5/7
ETTm2	3.339	4.220	5/7
Weather	3.601	3.683	16/21
Electricity	16.938	12.424	12/12
Traffic	3.803	3.450	10/12
Series macro	5.212	3.405	53/73

Weather improves on 16/21 channels despite losing slightly on the conventional target, while ETTh2 remains difficult. Electricity improves on all 12 sampled channels. The aggregate therefore extends beyond seven target columns, but the four negative intervals keep series-level negative transfer visible.

C Robustness, Evidence Status, and Efficiency Audit

For dataset-level inference on TimesFM, gains across the four forecast horizons are averaged within each dataset before resampling the seven datasets. Twenty thousand bootstrap draws give $[0.43, 12.51]\%$, while an exact sign-flip randomization gives $p = 0.125$. For $H = 96$, all seven clean frozen results are exactly reconstructed from cache. ETTm1, ETTm2, Electricity, and Traffic have strictly positive lower endpoints for block lengths 48, 96, and 192; ETTh1, ETTh2, and Weather do not.

The negative-transfer warning averages five validation-only risk percentiles and freezes the validation upper quartile as threshold. Its post-hoc evaluation gives TP=1, FP=1, FN=1, TN=4 and AUROC 0.50. Because all benchmark outputs had already been seen, this is a development audit, not untouched confirmation.

Table 5 gives a descriptive matched-cell efficiency check. The common TimesFM forecast/cache generation is excluded; time includes adapter selection or fitting and frozen test application. The single run is descriptive, not a controlled hardware-normalized latency comparison. Selection/fitting and application are aggregated, so these numbers do not isolate online query latency.

Table 5: Adapter-side efficiency on ETTh1, $H = 96$ (RTX 4090D host).

Method	Time (s)	Trainable params	State (MiB)
HiRA	3.70	0	164.1
Exact hidden ridge	554.58	245,952	1.88
Matched TS-RAG-style	1.28	33	64.7

This documents the absence of additional gradient-trained parameters and shows a large adapter-workflow time advantage over hidden ridge in this run, but matched TS-RAG-style retrieval is faster and smaller in this cell. Evidence roles remain distinct: TimesFM $H = 96$ and the warning are development evidence; TimesFM $H = 192/336/720$ and Chronos-2 $H = 192/720$ are frozen-protocol cross-horizon evidence; multiseries, block-length, efficiency, and LoRA results are post-hoc robustness audits.

D Learned Future Equivalence and Reliability Memory

We tested two preregistered extensions of the positive matched Future Retrieval Mixer (FRM). First, a query-only encoder was trained on ETTh2, ETTm1, ETTm2, Electricity, and Traffic so that pairwise similarities among frozen TimesFM hidden states reproduce pairwise geometry among context-normalized historical futures. Futures act only as teachers for completed source episodes; inference receives hidden states alone. Epoch selection uses ETTh1/Weather validation, and the matched control differs only in its hidden-cosine retrieval metric.

Table 6: Frozen $H = 96$ development comparison. Positive values would favor the learned future-equivalence metric over matched FRM.

Dataset	Matched FRM MSE	Learned metric MSE	Relative gain (%)
ETTh1	4.308874	4.360861	-1.207
Weather	133.656262	134.620245	-0.721
Macro			-0.964

The gate required at least 1% macro gain and no dataset below -0.5%; neither condition holds. Second, a reliability-gated residual memory stacked after FRM changes FRM by -7.67% macro on ETTh1/ETTh2/Weather, winning one of three cells. Checkpoints and predictions for both audits were frozen before separate target attachment. The failures distinguish hindsight-compatible historical futures from future equivalence identifiable by a present query. They do not rule out all learned retrieval, but they prevent promoting either extension as the proposed method. A never-opened external pool is reserved for a future method that first passes a new development protocol.

E Reproducibility Contract

Targets and split boundaries. The primary target is the column named `OT` in each TSLib-format CSV. After timestamp sorting, the exclusive train/validation/test end indices are $(8640, 11520, 14400)$ for ETTh1/2, $(34560, 46080, 57600)$ for ETTm1/2, $(36886, 42157, 52696)$ for Weather, $(18412, 21044, 26304)$ for Electricity, and $(12185, 13942, 17451)$ for Traffic. Within a split $[a, b)$, eligible context starts are $s = a, \ldots, b - L - H$, giving $b - a - L - H + 1$ stride-one windows before any cap. The primary study uses $L = 160$ and at most the latest 8,192 training-memory origins. The multi-series audit additionally caps validation and test at 2,048 origins.

Cache contents and access contract. The cache for every dataset, split, and horizon records the TimesFM forecast, target, start index, three 1280-dimensional hidden views, context length, patch length, checkpoint identifier, missing-value policy, and a test-label policy marker. Archive contents alone do not enforce physical test-label isolation: test targets must not be passed to retrieval, configuration selection, or fitting. Cache generation and evaluation are separate programs. Unit tests check split boundaries, zero fallback, exact cosine neighbors, residual attention, long-horizon hook counts, missing-sentinel handling, and metric assembly. The Chronos-2 cache additionally records the number of excluded regression and forecast-query tokens. Foundation inference uses one RTX 4090D with 24GB memory.

F Public Retrieval Protocol Audit

Scope of the matched fusion control. The TS-RAG-style control is an inspired future-retrieval convex mixer, not an implementation of the official ARM architecture. Last-token cosine search retrieves ten historical context–future segments. Each complete historical segment is centered and scaled using its own mean and standard deviation; its future is then mapped to the observed query-context scale. For each of three horizon blocks, eleven nonnegative weights summing to one combine the base forecast with ten ranked retrieved futures (33 stored coefficients). Weights minimize validation MSE by projected gradient descent on the simplex, with at most 5,000 iterations and a maximum-coordinate change tolerance of 10^{-11}. Validation references come from training only; frozen test references come from training plus validation. This fit uses all validation origins, not HiRA's two-fold eligibility rule. Thus origins, label availability, and the underlying backbone are matched, but selection budgets and fusion architectures are not identical. RAF-style augmentation re-runs the backbone on an augmented context; it does not reuse the same point forecast as its final output. Neither replay establishes a ranking against the official released systems.

The nearest released systems answer related questions under materially different contracts. TS-RAG uses a learned Adaptive Retrieval Mixer with Chronos-Bolt/MOMENT, context 512, and a released $H = 64$ zero-shot script. RAF retrieves with Chronos-T5 embeddings and prepends a normalized historical context–future segment; its released ETTh configuration uses $H = 20$ and one roll. SARAF trains a multivariate TSLib forecaster with context 720 and three seeds. Consequently, their published absolute errors are not inserted into Tables 1–17. We instead port each usable retrieval mechanism under the same frozen forecast, chronological origins, memory, and test-label contract, and label such results "matched mechanism" rather than "official reproduction."

This separation prevents two common but invalid comparisons: treating a gain over a method-specific backbone as an absolute ranking against TimesFM, or crediting a retrieval method with future values that were unavailable at the forecast origin. Exact repository revisions, commands, and machine-readable outputs are included in the supplementary reproducibility package.

Long-horizon competitiveness without gradient training. Table 8 isolates the Chronos-2 comparison from Table 18. Agreement-gated HiRA performs no additional gradient training of the backbone, encoder, retriever, or fusion network; its retrieval and shrinkage configurations are selected on validation data and frozen before test. The matched SARAF-style control uses a stationarity-aware residual-retrieval mechanism, with results averaged over three seeds. Both adapters perform no additional gradient training; the three seeds control stochastic neighbor selection. At $H = 336$, HiRA approaches this control (9.33% versus 9.59% mean MSE reduction); at $H = 720$, it achieves a slightly higher mean (13.15% versus 12.61%, a 0.54 percentage-point difference). Thus historical residual memory can deliver competitive long-horizon adaptation without additional gradient training.

Table 7: Protocol-separated audit of the closest public retrieval systems. Native results retain each release's own backbone and horizon; matched results in Table 18 instead hold our evaluation contract fixed.

Method	Released contract	Reproduction status and paired result
TS-RAG	Chronos-Bolt, $L = 512$, $H = 64$	Native ETTh1 rerun: MSE $0.3615 \rightarrow 0.3556$ ($+1.63\%$); kept separate from our four-horizon TimesFM table.
RAF	Chronos-T5, $H = 20$, one roll	Released construction improves MASE by 34.82% over four ETTh rows, but becomes -7.19% after enforcing forecast-time availability; excluded from leakage-free evidence.
SARAF	Trained multivariate TSLib model, $L = 720$	Native training contract is not commensurate; its stationarity/diversity retrieval is evaluated as a three-seed matched replay in Table 18.
TimeRAF	TTM-Base, $L = 512$, external knowledge base	The released repository lacks the trained retriever and knowledge base. Our matched HiRA-TTM control is not called a native reproduction or a direct comparison with the published TimeRAF system.

SARAF-style has higher means at the other horizons and across the complete matrix. These are descriptive mean comparisons, not evidence of pairwise statistical superiority; the control is a matched mechanism rather than an official SARAF reproduction.

Table 8: Comparison of gradient-free retrieval adapters on Chronos-2. Entries are seven-dataset macro-mean relative MSE reductions (%) over the unchanged backbone. Additional gradient training refers to adaptation, not backbone pretraining. For the HiRA study, $H = 96/336$ are development horizons and $H = 192/720$ use the frozen protocol. SARAF-style averages three seeds.

Method	Additional gradient training	96	192	336	720
Agreement-gated HiRA	No	6.92	6.72	9.33	13.15
Matched SARAF-style	No	9.35	7.83	9.59	12.61

G Frozen LoRA Complementarity Ablation

We give HiRA one preregistered opportunity to complement weight adaptation. This post-hoc $H = 96$ audit selects ETTh1, ETTm2, and Weather to cover a negative, strongly positive, and weak frozen-HiRA cell. We use the official TimesFM-2.5-200M Transformers/PEFT path and fit one train-only LoRA adapter per dataset (rank 4, alpha 8, dropout 0.05, all linear modules, 5,000 seed-42 windows, learning rate 10^{-4}, batch 32, at most 10 epochs). Exactly 512 validation windows select the epoch; all three select epoch 1. HiRA retains its original grid and two-fold selection, but its memory and hidden tokens are recomputed from the adapted backbone. All four columns in Table 9 are regenerated using the Transformers/PEFT inference path rather than copied from the primary TimesFM cache. Its frozen ETTh1 HiRA gain is -5.176%, compared with -4.999% in Table 1. These are separate inference runs; complementarity is assessed within this four-way run, not by mixing entries across tables.

LoRA alone has macro relative gain 22.27%, driven by ETTm2, but it hurts ETTh1. More importantly, HiRA after LoRA is negative in all three cells. This shows no complementarity for the three audited settings. Residual absorption, representation changes, retrieval mismatch, and selection instability remain possible explanations; this experiment does not distinguish them. We stop the LoRA branch without trying another rank, learning rate, seed, or favorable subset. Prediction files were frozen without test truth and hashed before a separate evaluation process loaded the targets; trainable parameters, histories, checkpoint and artifact hashes are provided in the supplement.

Table 9: Frozen four-way LoRA ablation (MSE). "Increment" is the relative gain of LoRA+HiRA over LoRA, the test of complementarity. This is post-hoc evidence and no second LoRA configuration was attempted.

Dataset	Frozen	+HiRA	LoRA	LoRA+HiRA	Increment
ETTh1	5.2633	5.5357	5.5807	5.6107	-0.54%
ETTm2	24.0542	15.9657	8.9280	9.1795	-2.82%
Weather	162.3686	166.9235	146.1818	147.9143	-1.19%
Macro incremental gain					-1.51%

H Qualitative Forecast Refinement Across Seven Datasets

Figure 4 displays one illustrative window per dataset from the frozen TimesFM $H = 96$ predictions. All 96 forecast steps are shown, without smoothing or changes to predictions. These are post-hoc qualitative examples, not a random sample or additional confirmation evidence.

Reproducible case selection. For each dataset, eligible windows have more than 10% relative MSE reduction, lower first-difference MSE than the backbone, and target standard deviation above the median across that dataset's test windows. Among eligible windows, we maximize $\mathrm{corr}(Y, \widehat{Y}) - 0.15\ \mathrm{MSE}(Y, \widehat{Y})/\ \mathrm{Var}(Y)$, using the first index in case of a tie. First-difference MSE compares successive changes in the forecast and target. Test outcomes are used only for this explicitly post-hoc visual selection, not to refit the predictor or change its frozen settings. Zero-based context starts in the timestamp-sorted series are 13647 (ETTh1), 13833 (ETTh2), 49194 (ETTm1), 56007 (ETTm2), 46985 (Weather), 23714 (Electricity), and 14405 (Traffic); each forecast starts 160 steps later. Window-level improvements do not imply dataset-wide gains: ETTh1 and Weather retain negative full-test gains, as reported in Table 1.

H.1 Retrieval trace: from historical errors to forecast refinement

Figure 5 opens the residual-memory mechanism behind the ETTm2 example in Figure 3. It follows the same previously selected window, rather than selecting a new case for this analysis. CPU replay uses the stored TimesFM context states, historical residuals, and frozen block configurations; the reconstructed forecast matches the saved prediction within 10^{-5} in original units. The context starts at zero-based index 56007, with forecast origin 56167 ($L = 160$, $H = 96$).

Reading the residual panel. The middle panel plots adjustments, not forecasts. Positive values indicate that the backbone needs to be raised; negative values indicate that it needs to be lowered. The black dashed curve is the current target minus the backbone forecast and is revealed only for evaluation. The green curve is the actual prediction-time adjustment, computed without current targets:

$$\delta_{q,B_b} = \gamma_b \sum_{j \in \mathcal{N}_b} \alpha_{b,j} (e_j)_{B_b}.$$

Here the frozen configurations use last-1 queries, $k_b = (16, 64, 64)$, $T_b = 0.03$, and $\gamma_b = 0.5$ for the three blocks. The faint curves show only the top three neighbors in each block, whereas the green curve uses *all* selected neighbors and includes shrinkage; it is not the simple average of the three displayed curves. Neighbor sets are block-specific, so faint segments across a boundary need not represent the same episode. All retrieved historical futures are fully observed before the query origin.

What the trace reveals about HiRA. The backbone captures the broad waveform but underestimates its central peak. Around steps 30–65, the historical residual vote supplies a positive, time-varying adjustment aligned with the required correction; outside this region the adjustment is much smaller. Thus the gain is not merely a uniform vertical offset. The adjustment raises the forecast toward the observed peak without replacing the backbone trajectory, although some peak underestimation remains. The trace makes the proposed interface concrete: frozen forecasting representations locate historical episodes, their *model-specific errors* provide correction values, and block-wise aggregation refines the output without gradient-training another encoder, retriever, or

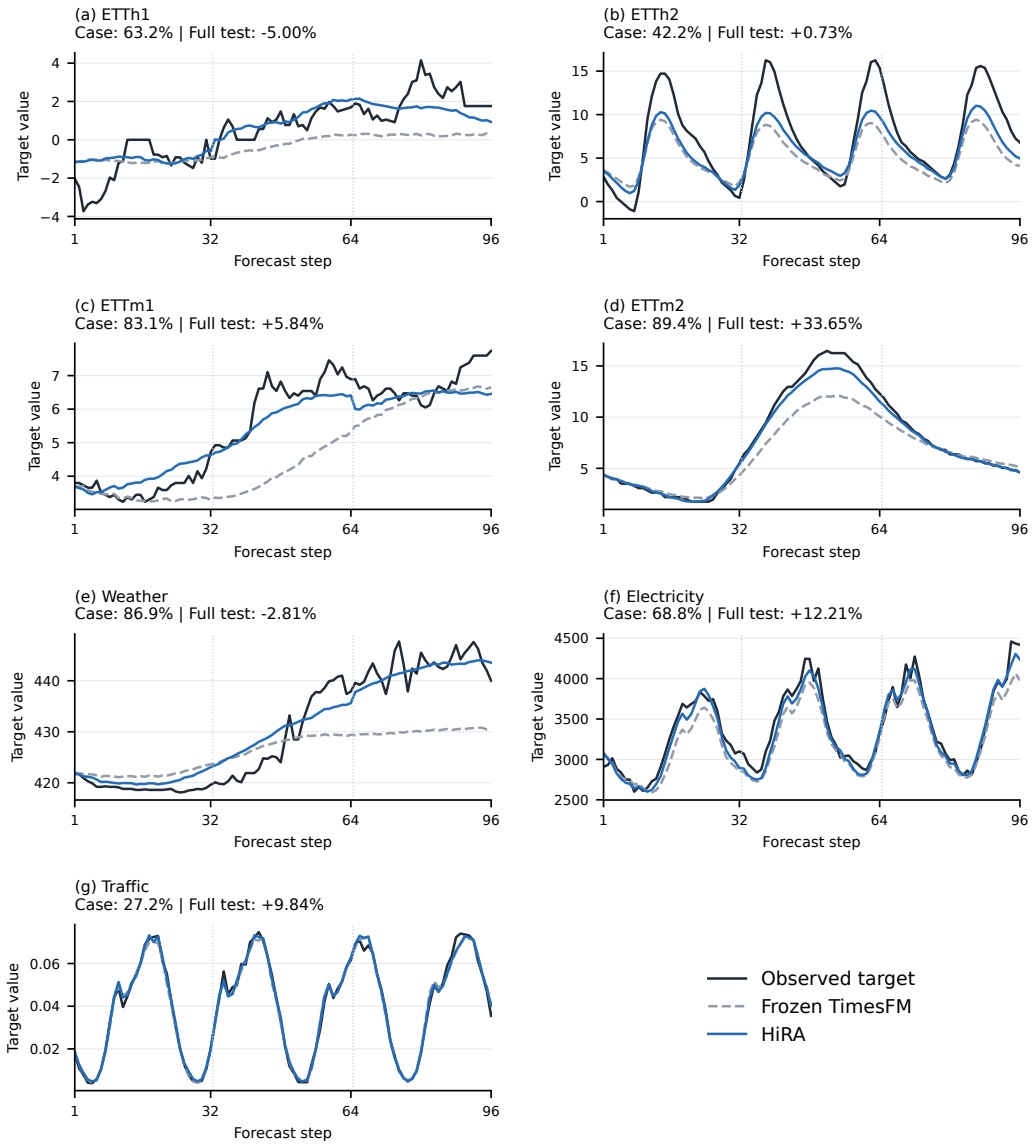


Figure 4: **Forecast refinement across all seven datasets.** The same display and selection rule is used throughout. "Case" and "Full test" denote relative MSE reductions for the illustrated window and the entire corresponding $H = 96$ test split, respectively. Positive values favor HiRA. Vertical dotted lines mark the three correction blocks. These examples highlight attainable improvements rather than typical performance.

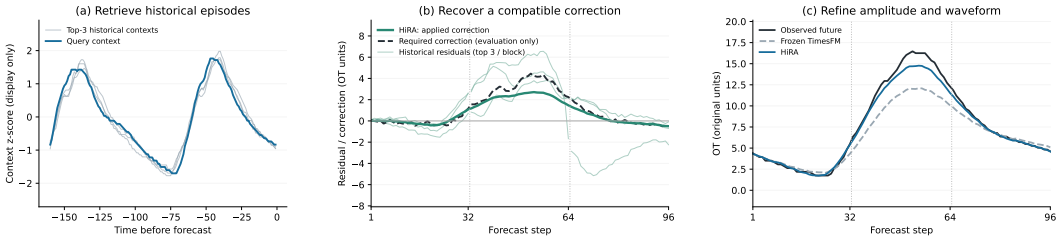


Figure 5: **Historical-error retrieval refines an underestimated forecast.** (a) Query context and the top three historical contexts retrieved for the first block; z-scoring is for display only. (b) The applied correction (green) and required correction, $Y_q - \widehat{Y}_q^0$ (black dashed, evaluation only), with three historical residuals per block shown faintly. (c) Adding the correction reduces peak underestimation while preserving the forecast waveform. The illustrated window has 89.43% MSE reduction; the full ETTm2 $H = 96$ test split has 33.65% reduction. This is a post-hoc mechanism illustration of the existing case, not independent confirmation or typical-window performance. Vertical dotted lines separate the three correction blocks.

fusion network. Together with the aggregate controls, this offers an interpretable example of how residual memory can complement a frozen forecaster. The example illustrates the mechanism; the ablations, rather than this single trace, assess the value of individual design choices. The selection rule remains the one stated above, and neither the displayed target nor its error is used in retrieval or prediction.

I Real-Data Correction Headroom and Transfer Diagnostics

Figure 6 audits existing ungated TimesFM predictions without refitting or changing configurations. For $H = 96$, the hindsight on/off oracle chooses, separately for each test window and each 32-step block, the lower squared error of the frozen HiRA prediction and the unchanged backbone. It uses test targets only for post-hoc scoring, is not deployable, and is neither a full candidate-grid oracle nor an estimate of query-identifiable value. CPU replay matched recorded relative gains within 10^{-5}.

The right panel includes all 28 dataset–horizon cells. Its horizontal coordinate averages the six recorded gains from three selected blocks and two validation folds; it is not the pooled validation MSE reduction. Its vertical coordinate is full-test MSE reduction. Eight cells have positive validation summaries but negative test gains. Validation was used for configuration selection, so the discrepancy does not isolate distribution shift or prove non-identifiability. The plots are descriptive audits, not independent confirmation.

J Additional Backbone Transfer: Full Results and Protocol

Models and observed-context extraction. We use Salesforce/moirai-2.0-R-small (Liu et al., 2025a), checkpoint revision `30f43ff08c8494f4943ae1521e9d4e94a0fbb389`, with uni2ts revision `cfd46d4510ed8896f263116f32928eede05b0a75`; and thuml/sundial-base-128m (Liu et al., 2025b), revision `3212e42564493f520593e5414af4367fc4b49226`. Both use float32, batch size 32, context 160, native patch length 16, and last-1/3/5 means of the first-forward final observed-context states. Future placeholders and recursively generated tokens are excluded. Moirai retains its official recursive quantile decoder and median point forecast; Sundial retains its 20-sample mean with seed 2027 and fixed dataset/split/batch seed derivation. Four-horizon smoke tests checked official forecast parity and observed/future-token isolation before expansion. Quantiles and sample counts were not changed in response to test MSE.

Scope and evidence status. The same seven univariate target columns, causal cleaning, chronological split boundaries, stride-one origins, and latest-at-most-8,192 training origins are used. We report all four Moirai horizons and both completed Sundial long horizons; Sundial $H = 96, 192$ are not part of this expansion comparison. Moirai ETTh1/ETTh2/Weather at all four horizons had been explored previously; their recomputation is replication, not new confirmation. The expansion grid

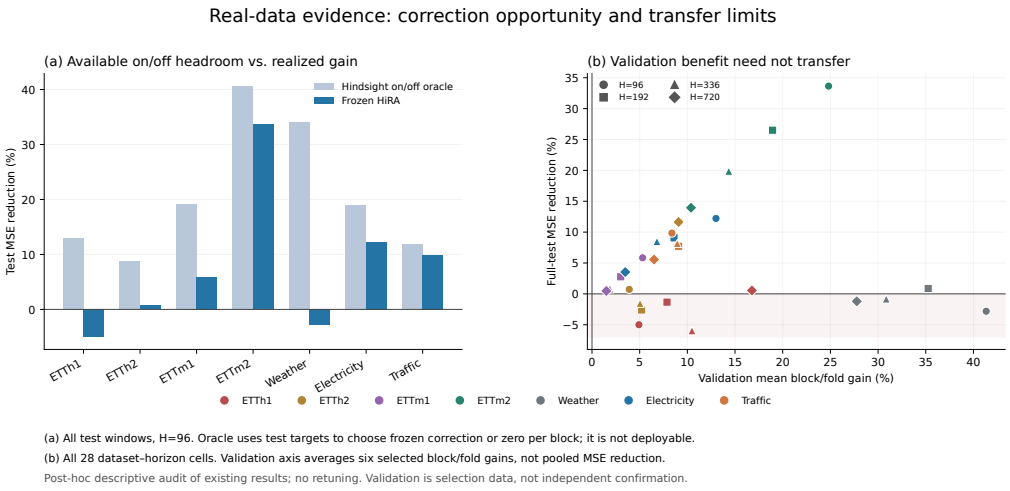


Figure 6: **Real-data correction opportunity and transfer limits on TimesFM.** Left: restricted hindsight on/off gains and realized HiRA gains on all $H = 96$ test windows. Right: validation summaries versus test gains over all four horizons. Negative outcomes are retained; the axes use different aggregation rules as detailed above.

and ungated HiRA recipe were fixed before evaluating the added settings. Sundial control fitting was specified after its HiRA scores were known, so this is a fixed-comparator audit, not untouched confirmation. All Moirai configuration selections and control fits were saved before test prediction. Separate processes select/fit, predict, and score; the query loader excludes truth, and prediction artifacts are hashed before target attachment. Synthetic label-poisoning tests verify that changing test targets does not change predictions. No test residual updates memory or refits an adapter.

Matched controls. Constant calibration uses the historical mean residual vector with block-wise shrinkage in $\{0, 0.25, 0.5, 1\}$. Hidden ridge standardizes last-1/3/5 features, uses multi-output Cholesky ridge with penalty $\{0.01, 1, 100\}$ and the same shrinkage set, and refits the selected maps on train plus validation. Both use HiRA's two purged chronological folds and nonnegative-both-folds eligibility rule. The standalone FRM uses last-token cosine top-10 retrieval and the historical segment normalization described in Appendix F. It fits 11 simplex coefficients per block (base plus ten future experts) on all validation origins, with training-only retrieval memory during fitting; test memory contains train plus validation. The fitting schedule and search budget therefore differ from HiRA, although the available labels, forecast origins, and cached backbone outputs match. FRM is an in-house mechanism control, not an official TS-RAG or SARAF implementation.

Metrics and interpretation. All entries below are relative MSE reductions against the identical cached backbone; zero denotes no net change, not missing evaluation. Brackets report individual paired circular moving-block 95% intervals with block length H, 2,000 draws, and seed 2027. Intervals are not adjusted for multiple comparisons; positivity against the backbone does not establish superiority to another adapter. Different horizons of a dataset are dependent. Long-horizon macro means weight the 14 dataset–horizon settings equally, not their window counts. The results compare adaptation gains within a backbone, not absolute accuracy across backbones or a normalized multivariate leaderboard. Moirai and Sundial also retain different point-forecast estimators (median versus sample mean).

The long-horizon gains are positive in 14/14 Moirai settings (six strictly positive intervals) and 10/14 Sundial settings (five strictly positive intervals). HiRA beats constant/ridge/FRM in 14/11/3 of the 14 Moirai settings and 8/9/7 of the 14 Sundial settings. Moirai's FRM advantage and Sundial's four negative HiRA cells remain visible. The original adapter and its grid are unchanged; no post-test choice between gated and ungated variants is made.

Table 10: Moirai 2.0 at $H = 96$: full dataset-level relative MSE gain (%) and individual 95% interval. All methods share the same backbone predictions.

Dataset	HIRA	Constant	Ridge	FRM
ETTh1	-8.403 [-20.86, 4.40]	-0.066 [-1.55, 1.02]	-12.359 [-22.20, -3.07]	1.136 [-6.10, 10.43]
ETTh2	1.514 [-10.08, 11.26]	-0.018 [-0.10, 0.06]	8.194 [4.56, 11.80]	12.843 [7.03, 18.88]
ETTm1	6.068 [1.75, 10.19]	0.000 [0.00, 0.00]	6.748 [1.71, 11.70]	9.897 [5.52, 14.30]
ETTm2	12.298 [3.22, 20.91]	0.000 [0.00, 0.00]	13.352 [5.58, 20.33]	16.391 [8.64, 23.72]
Weather	16.277 [7.16, 24.13]	-0.050 [-1.52, 1.63]	16.435 [6.81, 24.50]	19.799 [10.77, 27.92]
Electricity	5.570 [0.08, 10.07]	0.000 [0.00, 0.00]	6.567 [3.81, 9.10]	3.363 [2.28, 4.64]
Traffic	11.760 [5.86, 19.51]	0.105 [-0.02, 0.27]	7.718 [3.57, 12.47]	5.934 [0.60, 12.16]

Table 11: Moirai 2.0 at $H = 192$: full dataset-level relative MSE gain (%) and individual 95% interval. All methods share the same backbone predictions.

Dataset	HIRA	Constant	Ridge	FRM
ETTh1	-1.540 [-14.17, 11.24]	-0.133 [-1.13, 0.70]	-4.395 [-11.88, 3.15]	4.873 [-5.97, 17.97]
ETTh2	-2.254 [-13.17, 8.78]	0.000 [0.00, 0.00]	3.354 [-3.44, 9.43]	7.260 [0.64, 14.42]
ETTm1	3.562 [-0.11, 7.31]	0.000 [0.00, 0.00]	5.780 [2.48, 9.08]	10.384 [6.00, 14.87]
ETTm2	5.314 [-3.89, 14.00]	0.000 [0.00, 0.00]	4.609 [-4.54, 12.67]	10.932 [3.40, 18.23]
Weather	10.651 [3.48, 17.68]	0.631 [-1.50, 2.87]	10.581 [4.29, 16.67]	19.171 [13.14, 25.18]
Electricity	4.130 [-0.58, 7.91]	0.108 [-0.69, 0.75]	4.273 [2.92, 5.45]	3.565 [2.41, 5.14]
Traffic	8.628 [1.47, 17.71]	0.182 [-0.16, 0.53]	7.727 [3.56, 12.59]	5.613 [-0.53, 12.51]

Table 12: Moirai 2.0 at $H = 336$: full dataset-level relative MSE gain (%) and individual 95% interval. All methods share the same backbone predictions.

Dataset	HIRA	Constant	Ridge	FRM
ETTh1	7.434 [-10.08, 23.88]	0.000 [0.00, 0.00]	2.075 [-8.78, 11.43]	14.484 [-1.87, 28.92]
ETTh2	4.229 [-1.34, 9.30]	0.000 [0.00, 0.00]	3.289 [2.14, 4.37]	12.110 [8.37, 15.50]
ETTm1	2.113 [-4.40, 8.40]	0.759 [-0.94, 2.49]	6.042 [0.17, 11.36]	15.226 [7.10, 22.93]
ETTm2	1.505 [-8.25, 10.05]	0.000 [0.00, 0.00]	0.570 [-9.13, 8.79]	6.337 [-3.04, 14.38]
Weather	11.921 [4.41, 19.17]	2.786 [-1.32, 6.45]	10.826 [5.94, 15.53]	21.997 [18.11, 26.16]
Electricity	2.856 [0.68, 5.11]	-0.062 [-3.69, 2.72]	1.681 [0.75, 2.79]	2.890 [1.82, 4.32]
Traffic	10.018 [2.23, 19.75]	0.107 [-0.09, 0.33]	6.660 [1.67, 13.12]	5.146 [-1.24, 12.81]

Table 13: Moirai 2.0 at $H = 720$: full dataset-level relative MSE gain (%) and individual 95% interval. All methods share the same backbone predictions.

Dataset	HiRA	Constant	Ridge	FRM
ETTh1	12.928 [0.70, 22.89]	0.000 [0.00, 0.00]	5.362 [-0.10, 9.87]	14.089 [-8.50, 31.76]
ETTh2	20.026 [10.34, 29.02]	8.557 [5.46, 11.25]	11.278 [5.98, 15.51]	11.744 [8.10, 15.66]
ETTm1	3.472 [-12.63, 19.60]	2.161 [-2.28, 6.50]	5.455 [-9.01, 19.76]	18.111 [3.24, 32.33]
ETTm2	0.828 [-11.16, 11.74]	0.000 [0.00, 0.00]	-2.102 [-15.31, 9.79]	3.721 [-7.96, 14.09]
Weather	9.785 [-3.98, 20.82]	7.284 [-0.48, 12.68]	8.735 [-7.29, 20.22]	21.462 [15.68, 26.14]
Electricity	0.522 [-0.02, 1.14]	-1.925 [-7.84, 3.81]	0.640 [-1.57, 2.82]	2.150 [1.37, 2.93]
Traffic	10.325 [6.03, 15.58]	0.251 [-0.06, 0.63]	6.170 [2.37, 11.38]	4.576 [-1.68, 12.81]

Table 14: Sundial at $H = 336$: full dataset-level relative MSE gain (%) and individual 95% interval. All methods share the same backbone predictions.

Dataset	HiRA	Constant	Ridge	FRM
ETTh1	2.180 [1.13, 3.21]	4.679 [-7.76, 15.66]	0.000 [0.00, 0.00]	-0.391 [-3.55, 2.46]
ETTh2	5.786 [-0.31, 11.50]	0.000 [0.00, 0.00]	1.158 [-1.72, 3.81]	1.865 [-0.45, 4.65]
ETTm1	2.365 [-0.60, 4.91]	0.000 [0.00, 0.00]	-1.536 [-6.71, 2.69]	1.812 [0.69, 2.89]
ETTm2	4.812 [-3.79, 12.40]	0.151 [0.03, 0.27]	5.589 [0.88, 10.05]	8.363 [3.13, 13.57]
Weather	-0.421 [-7.40, 6.10]	0.000 [0.00, 0.00]	2.569 [-1.35, 6.04]	3.131 [1.54, 4.88]
Electricity	3.612 [2.44, 5.11]	-0.001 [-0.23, 0.17]	0.818 [0.26, 1.29]	3.821 [2.56, 5.51]
Traffic	13.495 [7.94, 20.92]	0.038 [0.01, 0.07]	8.200 [4.51, 13.23]	6.303 [2.14, 11.04]

Table 15: Sundial at $H = 720$: full dataset-level relative MSE gain (%) and individual 95% interval. All methods share the same backbone predictions.

Dataset	HiRA	Constant	Ridge	FRM
ETTh1	6.304 [-0.47, 11.88]	21.992 [3.46, 38.37]	0.184 [-4.20, 3.34]	0.070 [-2.47, 1.86]
ETTh2	20.629 [10.87, 27.30]	15.177 [9.70, 19.31]	3.592 [-0.46, 7.45]	4.034 [-0.86, 7.91]
ETTm1	-0.590 [-3.05, 2.22]	0.000 [0.00, 0.00]	0.505 [-0.87, 1.87]	0.813 [-0.29, 2.10]
ETTm2	-1.029 [-10.12, 7.92]	0.000 [0.00, 0.00]	0.508 [-3.18, 4.97]	4.265 [1.12, 7.59]
Weather	-3.626 [-13.02, 4.39]	-0.027 [-0.69, 0.65]	-0.385 [-5.54, 4.62]	2.326 [1.41, 3.60]
Electricity	1.605 [-1.07, 4.28]	0.009 [-0.01, 0.03]	-0.273 [-1.76, 1.20]	2.132 [1.40, 2.90]
Traffic	11.230 [7.21, 17.53]	0.080 [0.04, 0.15]	6.219 [3.94, 9.91]	3.328 [-0.25, 8.47]

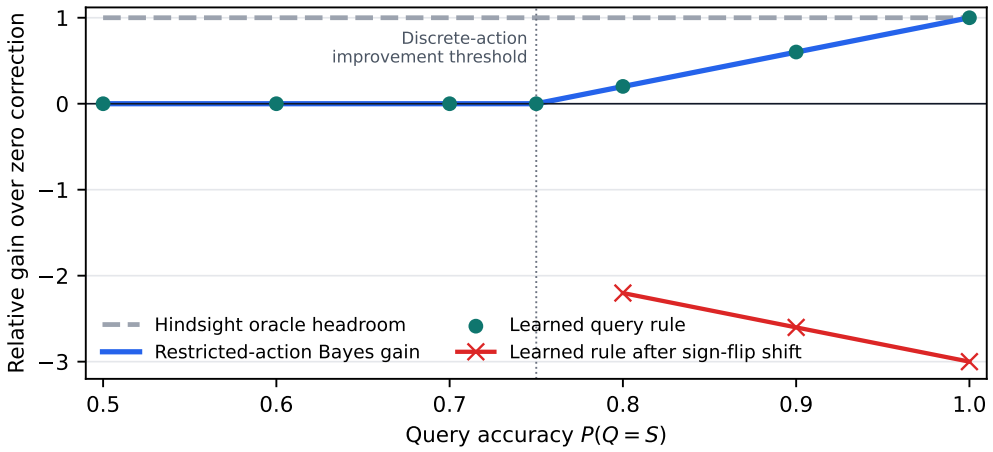


Figure 7: Oracle headroom versus value identifiable from the query. The stationary learned rule matches the restricted-action Bayes boundary; a sign-flip shift makes the same learned rule harmful. The $p = 0.75$ threshold applies only to the discrete action set $\{0, -1, +1\}$, not arbitrary real-valued corrections.

Table 16: Long-horizon backbone expansion. Entries are dataset-macro relative MSE reductions (%) over each method's identical cached backbone; Mean averages both horizons and Wins counts positive settings. Bold method names identify our proposed adapter. All HiRA rows are ungated. FRM is a matched future-retrieval convex mixer, not an official TS-RAG/SARAF reproduction; fitting schedules differ.

Backbone	Adapter	336	720	Mean	Wins
Moirai 2.0	**HiRA**	5.725	8.269	6.997	14/14
	Constant	0.513	2.333	1.423	7/14
	Hidden ridge	4.449	5.077	4.763	13/14
	FRM	11.170	10.836	11.003	14/14
Sundial	**HiRA**	4.547	4.932	4.740	10/14
	Constant	0.695	5.319	3.007	7/14
	Hidden ridge	2.400	1.479	1.939	10/14
	FRM	3.558	2.424	2.991	13/14

K Extended Experimental Evidence and Transfer Diagnostics

This section retains the controlled experiment, detailed comparison tables, and transfer diagnostics accompanying the main-text long-horizon synthesis.

K.1 Controlled recoverability and negative transfer

We instantiate the recoverability boundary and its failure under sign reversal with balanced $Y \in \{-1, +1\}$, zero base forecast, restricted correction actions $\{0, -1, +1\}$, and a symmetric noisy query sign that matches the future regime with probability $p \in [0.5, 1]$. The sample-wise oracle removes the entire unit loss. The restricted-action Bayes rule has risk $\min\{1, 4(1-p)\}$ and cannot improve until $p > 0.75$. This threshold is a consequence of the discrete action set, not a general identifiability threshold: with unrestricted real-valued corrections, the proposition gives $c^*(Q) = (2p-1)Q$, risk $4p(1-p)$, and positive gain for every $p > 0.5$. Figure 7 compares this exact curve with logistic regression trained on 20,000 samples and evaluated on 100,000 fresh samples, repeated independently 20 times; error bars are empirical 95% intervals. The learned rule tracks the identifiable value under stationarity. Flipping the query–regime relation only at test time turns increasingly informative training queries into increasingly harmful corrections, reaching -300% relative gain at $p = 1$. Thus both zero recoverability and negative transfer occur despite unchanged hindsight-oracle headroom.

Table 17: Stronger controls across horizons. Entries are mean relative MSE gain (%) over TimesFM; the last column counts positive cells. Hedge uses only causally revealed test outcomes, while HiRA and ridge are frozen.

Method	96	192	336	720	All / wins
HiRA	7.781	6.138	4.142	4.929	5.747 / 20
Exact hidden ridge	6.137	4.260	3.182	1.717	3.824 / 22
Delayed online Hedge	-0.744	0.140	1.318	0.000	0.179 / 2

Table 18: Matched-mechanism relative MSE gain (%). Methods within each backbone share evaluation origins and label-availability constraints, with fitting and context-augmentation differences detailed in Appendix F. These are not native checkpoint reproductions. SARAF-style values average three seeds.

Backbone	Method	96	192	336	720	All / wins
TimesFM	HiRA	7.78	6.14	4.14	4.93	5.75 / 20
TimesFM	TS-RAG-style	11.15	10.15	9.22	8.32	9.71 / 28
TimesFM	RAF-style	-42.33	-22.72	-10.90	-51.73	-31.92 / 9
Chronos-2	Gated HiRA	6.92	6.72	9.33	13.15	9.03 / 26
Chronos-2	SARAF-style	9.35	7.83	9.59	12.61	9.84 / 27

K.2 Transfer diagnostics and unsuccessful extensions

Appendix I complements the controlled experiment with real-data correction-headroom and validation-to-test diagnostics. This is positive frozen-backbone adaptation evidence, not a general SOTA claim: learned fusion is stronger on TimesFM and Moirai, whereas HiRA has the highest long-horizon mean among the tested Sundial adapters. Raw rolling MSE is not comparable to normalized multivariate tables, and only the predeclared horizons are confirmation evidence. ETTh1 and Weather show that similar hidden states need not imply compatible future errors. Indeed, a frozen validation-only warning using aliasing, agreement, shift, concentration, and fold stability attains AUROC 0.50, so we do not claim failure detection (Appendix C). Dataset-level uncertainty is assessed for the TimesFM four-horizon analysis in Appendix C. With only seven datasets, related ETT datasets may also be dependent, so dataset-level independence is an approximation rather than an established property. Finally, adding HiRA after frozen LoRA hurts all three audited cells (mean -1.51%), showing no observed complementarity in these settings (Appendix G). A stricter development audit also asks whether historical future geometry can supervise a query-only retrieval metric. Its source loss improves, but held-out neighbor quality peaks after one epoch and the frozen metric loses 0.96% to the matched future-retrieval mixer on ETTh1/Weather (0/2 wins). A subsequent reliability residual memory loses 7.67% to that same mixer (1/3 wins). We retain these failures rather than tuning against opened targets; details are in Appendix D.

K.3 Matched public retrieval mechanisms

These are inspired mechanism controls with matched evaluation origins and label availability, not native public checkpoints or identical training protocols (Appendix F). On Chronos-2, SARAF-style reaches 9.84%/27 wins versus 9.03%/26 for gated HiRA. Without additional gradient training, agreement-gated HiRA remains competitive with this matched SARAF-style retrieval control at long horizons: its mean gain approaches SARAF-style at $H = 336$ (9.33% versus 9.59%) and is slightly higher at $H = 720$ (13.15% versus 12.61%). Appendix F, Table 8, provides the full four-horizon comparison and adaptation-training status.

TimeRAF omits artifacts needed for an official rerun. A matched HiRA–TTM control improves 5/6 $H = 96$ cells by 6.12%, but is not labeled a native TimeRAF reproduction (Appendix F).

An expanded $H = 96$ check improves 53/73 predeclared series and six of seven dataset means, with four significant negative cases retained (Appendix B).

All four significant $H = 96$ cells remain significant for block lengths $H/2$, H, and $2H$ (Appendix C).