# ONE-CORE: EVIDENCE-GATED COORDINATE SEARCH FOR TIME-SERIES INDUCTIVE BIASES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Architecture comparisons bundle observation, temporal, channel, capacity, and objective-side inductive biases, making whole-model winners difficult to interpret or transfer. We introduce ONE-CORE, a factorized testbed that audits one mechanism-linked bias coordinate at a time. Matched interventions measure its response, direct training-data descriptors propose where that response may change, and nested leave-one-dataset-out evaluation admits conditioning only when it improves on a fixed prior. Unsupported updates abstain locally; joint updates require a measured endpoint.

Forecasting and remaining useful life (RUL) exhibit different compression preferences within the audited settings: a six-dataset forecasting audit retains a hierarchical prior, whereas C-MAPSS favors a compact single bottleneck before conditioning capacity. For MSE forecasting, a normalization gate reduces mean regret from 10.15% with fixed RevIN to 2.11% across twelve nested-LODO folds, with selective improvements on two folds. An in-family RUL capacity gate reaches 0.00% regret versus 2.66% for fixed dim16 across four C-MAPSS subsets; external bearing audits provide only directional corroboration. Anomaly detection (AD) instead audits readout, evidence source, and calibration. Its six-source diagnostic envelope reaches 0.438 macro oracle Best-F1 versus 0.412 for the best fixed source, but five datasets do not support a routing gate. These outcomes provide local design evidence with explicit scope and uncertainty, rather than a global architecture oracle.

## 1 INTRODUCTION

The hidden variable in many time-series comparisons is not only scale or training recipe, but the inductive biases bundled inside an architecture. Forecasting extrapolates temporal structure, RUL compresses a degradation trajectory into a health state, and AD converts deviations into a ranked score and an operating decision. Success may therefore come from normalization, locality, bottleneck capacity, channel treatment, residual formation, aggregation, or calibration. Comparing only complete models obscures which choice helped, whether that choice transfers, and what should change on a new dataset. AD adds a further identification problem: an apparent preference for one residual family can be induced by differences in reconstruction fidelity within the other. ONE-Core therefore factorizes AD into readout geometry, score geometry, normality-evidence source, and calibration rather than treating reconstruction versus forecasting as a single model-family decision. This factorization also lets us compare point residual, residual induction, local profile, Q2 profile, context profile, and profile fusion under the same backbone, training checkpoints, and evaluation interface.

ONE-Core treats design as local, derivative-free coordinate search over a declared candidate space. Matched interventions estimate one coordinate's response; descriptors propose conditional updates; validation admits or rejects them. The experiments here audit individual axes around task-specific common references and then measure selected joint endpoints. They do not execute a fully automatic sequence of coordinate-descent updates. Repeated use may be path dependent; alternative audit orders and convergence are not evaluated.

We call this problem setting *evidence-gated inductive-bias experimentation*. Unlike whole-architecture search or learned model recommendation, its unit of inquiry is one mechanism-linked bias coordinate and its output is an evidence record: retain a fixed prior, condition the coordinate



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.

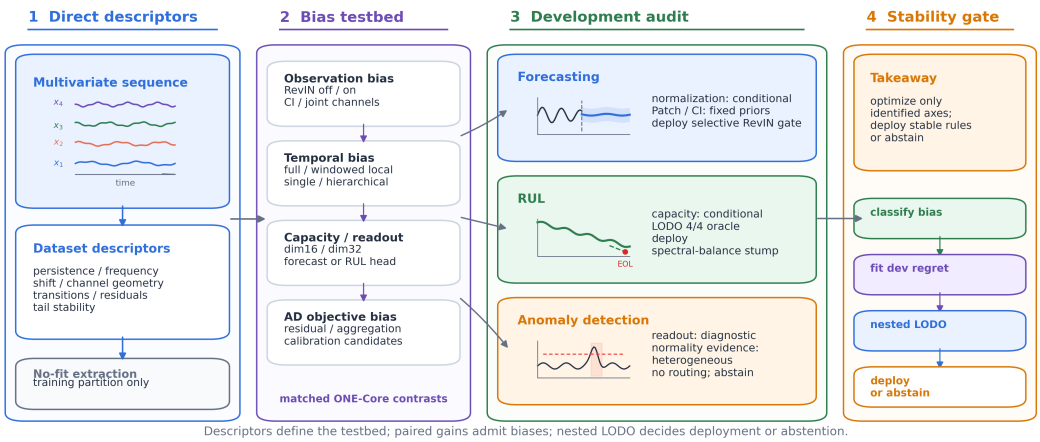


Figure 1: **Bias-specific experimental testbed.** Direct training-data descriptors motivate separate bias hypotheses. ONE-Core measures each axis with a matched intervention; paired responses fix, reject, or admit that axis to a restricted regret probe. Nested LODO either validates its conditional gate or inherits the fixed task prior. In AD, score geometry remains a controlled diagnostic; normality evidence is heterogeneous, but routing abstains.

on direct dataset evidence, or reject the proposed override. Domain knowledge generates testable descriptor–bias hypotheses, but held-out response evidence, rather than the hypothesis alone, determines admission.

At H96, dim16, and $K{=}2$, the five-seed compression audit covers four ETT variants, Weather, and Exchange. Hierarchy wins on ETTm1 and Weather; four comparisons remain unresolved. Refitting a context-memory descriptor probe within six-dataset LODO retains hierarchy on every fold, supporting the current prior within this matched scope. RUL separately favors compact summarization within the audited C-MAPSS configurations.

ONE-Core is not a universal architecture or a claim that one global descriptor distance solves architecture selection. It controls one task-relevant bias in one implementation, measures its paired response, and tests only that axis for dataset conditioning. The question is therefore not "which complete model wins?" or "have all biases been discovered?" but "does the next supported coordinate update improve on the current task prior out of dataset?"

Raw sequences fit the predictor; direct descriptors expose the dataset-level variable being tested for a coordinate override. Persistence motivates memory/span, spectral balance motivates capacity, window-scale dispersion motivates normalization, and covariance rank and lagged coupling motivate channel treatment. Transition, innovation, and tail statistics motivate AD scoring and calibration. Computed only from training partitions, these mechanism-linked summaries neither learn a corpus representation nor replace held-out response validation.

The method has three inferential layers. A *mechanistic hypothesis* maps descriptor block $\phi_j$ to the information-processing burden of bias $j$. A *controlled identification* layer changes only $j$ and measures its paired multi-seed response. An *empirical admission* layer permits dataset conditioning only when that response is heterogeneous and a restricted gate improves over the best fixed prior under nested dataset-level validation. Otherwise the coordinate inherits that prior. A descriptor is therefore a surrogate direction to test, not a universal law or a guarantee of the winning action. This makes the unit of inference a bias response, not dataset identity.

This outcome space makes negative experiments first-class evidence. A resolved no-gain intervention rejects the proposed update, a harmful joint endpoint rejects that composition, and an unstable mapping rejects the conditional rule. These outcomes reduce the admissible design space rather than count as failed model runs. By contrast, an underpowered or protocol-confounded comparison is only unresolved evidence and triggers local abstention; it is not converted into a negative claim.

Our contributions follow this inferential order. **(i) Coordinate-local formulation.** We formulate evidence-gated inductive-bias experimentation as local search over declared updates, with one-factor audits and separate endpoint checks in this study. **(ii) Mechanism-linked evidence.** Each conditional probe is restricted to direct descriptors that measure the burden of its bias coordinate, reducing hidden search freedom. **(iii) Empirical admission.** Nested LODO, fold-wise non-degradation, descriptor-row permutation, and local abstention test a restricted gate against the best fixed prior. **(iv) Cross-task evidence.** Multi-seed audits recover forecasting and RUL priors and conditional gates, reject a harmful composition, and show in AD how fidelity control can invalidate an apparent routing rule while a normality-evidence envelope remains diagnostic rather than deployable. Together, the accepted and rejected interventions establish the value of a common factorized testbed without turning every observed winner into a policy.

## 2 Related Work

**Time-series forecasting.** Informer (Zhou et al., 2021), Autoformer (Wu et al., 2021), FEDformer (Zhou et al., 2022), DLinear (Zeng et al., 2023), PatchTST (Nie et al., 2023), TimesNet (Wu et al., 2023), iTransformer (Liu et al., 2024a), and TimePerceiver (Lee & Lee, 2025) encode distinct representation and aggregation choices. NHITS (Challu et al., 2023) uses hierarchical interpolation and multirate sampling; TimeMixer (Wang et al., 2024) mixes decomposed multiscale series, while ModernTCN (Luo & Wang, 2024) revisits convolution. Our contribution is a matched abstraction audit within one implementation, rather than a new multiscale primitive. RevIN (Kim et al., 2022b) motivates the separate observation-normalization coordinate.

**RUL prediction.** RUL estimation differs from forecasting because the target is a compact health-state summary rather than a future sequence. CNN, recurrent, Transformer, and graph-based prognostics models have improved accuracy on CMAPSS-like benchmarks (Saxena et al., 2008; Lei et al., 2018; Babu et al., 2016; Li et al., 2018; Zheng et al., 2017; Li et al., 2024), and recent work has studied foundation-model transfer for engine prognostics (Dinten & Zorrilla, 2025). ONE-Core makes the contrast between RUL summarization and forecasting extrapolation a controlled configuration difference: compact single bottlenecks for RUL versus hierarchical aggregation for forecasting.

**Anomaly detection.** Multivariate time-series anomaly detection includes reconstruction, forecasting residual, association, and graph-based scoring families (Su et al., 2019; Audibert et al., 2020; Zhao et al., 2020; Deng & Hooi, 2021; Tuli et al., 2022; Xu et al., 2022; Lai et al., 2023; Yang et al., 2023). Benchmark studies emphasize that reported performance is highly sensitive to score aggregation, thresholding, point-adjustment, and calibration (Tatbul et al., 2018; Hundman et al., 2018; Abdulaal et al., 2021; Lai et al., 2021; Kim et al., 2022a). We therefore factorize anomaly detection into readout geometry, residual-score geometry, normality-evidence source, and calibration rather than treating an AD method as a single indivisible family.

**Unified and foundation time-series models.** Recent unified and pretrained models pursue broad coverage through scale, pretraining, or shared interfaces (Das et al., 2024; Gao et al., 2024; Goswami et al., 2024; Liu et al., 2024b; Woo et al., 2024). Their strength is coverage; their weakness for our purpose is diagnostic opacity. ONE-Core instead sacrifices scale to make bias choices visible.

**NAS, meta-selection, and inductive-bias experimentation.** Feature-based forecasting methods such as FFORMA learn to map time-series features to model weights or model combinations (Montero-Manso et al., 2020), and NAS methods search forecasting modules directly (Liang & Sun, 2024). Recent meta-learning jointly embeds datasets and candidate architectures to predict architecture and hyperparameter performance over accumulated search histories (Moeini et al., 2026). Such methods target efficient recommendation over a broad search space. ONE-Core poses a different problem: it neither searches a monolithic architecture nor learns a surrogate that must name a winner. It uses no-fit, mechanism-linked descriptors to motivate matched interventions inside one factorized testbed, and returns reusable bias-response evidence together with a Fix, Condition, or Abstain decision. Its stability diagnostic may therefore retain a task prior rather than force a recommendation. This also differs from selective prediction, which abstains on individual outputs to trade coverage for predictive risk (Geifman & El-Yaniv, 2017); ONE-Core always produces the task output and abstains only from an unsupported design-coordinate override. Foundation-model

analysis also characterizes how patching, embedding, and training loss induce temporal, geometric, and regression-to-the-mean biases in forecasting-oriented TSFMs (Yu et al., 2026). ONE-Core asks a complementary question: given a known task and direct dataset evidence, which observation-, structure-, and objective-side biases survive controlled intervention, and is there enough cross-dataset evidence to condition them further? This distinction makes the burden of evidence explicit: conditioning must be evaluated against task-only, universal, random, and oracle controls.

## 3 Bias-Wise Intervention and Evidence-Gated Admission

**Problem formulation.** Given a task $\tau \in \{\text{forecasting}, \text{RUL}, \text{AD}\}$ and its training partition $\mathcal{D}$, ONE-Core is an experimental testbed that audits a predefined lightweight candidate space rather than optimizing over unrestricted architectures. Forecasting produces a future sequence, RUL compresses a degradation trajectory to a scalar health estimate, and AD forms a score sequence from reconstruction or short-horizon prediction residuals; across all three, the design question is which declared bias coordinates should be fixed, conditioned, or left at the task prior. Matched responses $\Delta_j$ are measured across development datasets only. For an unseen dataset, a validated gate receives only the training-partition descriptor $\phi_j(\mathcal{D})$; fixed and abstained axes inherit their task prior. Consequently, $\hat{a}(\mathcal{D}, \tau)$ denotes an evidence-admitted configuration within the declared candidate space, not a globally optimal architecture. The catalogue and evidence corpus may expand over repeated audits; at each iteration the method estimates only local coordinate responses around the current controlled reference. This is coordinate search under an explicit experimental budget, not exhaustive enumeration of time-series structure. Its output is therefore both an admitted configuration and an evidence record over accepted, rejected, and unresolved bias hypotheses.

**Direct descriptor design.** For each dataset $\mathcal{D}$, we compute a direct training-partition descriptor $\phi = [\phi_{\text{persist}}, \phi_{\text{freq}}, \phi_{\text{shift}}, \phi_{\text{channel}}, \phi_{\text{transition}}, \phi_{\text{tail}}]$ organized by data phenomenon rather than benchmark identity. The full vector audits corpus coverage only. Each gate sees a declared mechanism-linked block $\phi_j \subset \phi$; unrelated coordinates cannot enter, and no multivariate dataset-to-winner classifier is fitted. These statistics instantiate properties implicated by forecasting normalization and interfaces, RUL compression, and AD scoring (Kim et al., 2022b; Nie et al., 2023; Lei et al., 2018; Deng & Hooi, 2021; Lai et al., 2021). Descriptor families arise from mechanism-driven exploratory analysis; once a coordinate is submitted for admission, its descriptor block and rule class are frozen before nested validation (Appendix A).

Two examples are central. RUL capacity uses a run-wise log fast/slow spectral balance $B_{\text{fast}}$, computed before any learned modality adapter, to measure the observation bandwidth that a bottleneck must preserve. Forecast normalization uses relative input-window scale drift $D_\sigma^L$, which directly measures how strongly the standard deviation removed by RevIN varies across model contexts. Context-to-future shift is retained as a complementary diagnostic but does not fit the admitted gate. All values are extracted from training partitions; exact equations and aggregation rules appear in Appendix A. Table 1 declares the only descriptor-to-bias links eligible for admission.

Table 1: **Mechanism-linked descriptor-to-bias catalogue.** This table declares intervention hypotheses, not validated or deployed gates. Each direct descriptor block is assigned to a matched bias coordinate; neither global vector distance nor nearest-dataset identity labels a winning architecture.

| Descriptor family | Bias coordinate | Controlled candidate |
|---|---|---|
| Context spectra / variation | temporal interface / query access | Patch/Global; $K{=}2/4$ |
| Spectral balance / complexity | capacity | dim16/32/64 |
| Context-scale memory | abstraction form | Single-8 / 32→8 |
| Memory / period–patch alignment | hierarchy range | 32→8 / 64→16 |
| Input-window scale drift | normalization | RevIN off/on |
| Channel rank / lagged coupling | channel treatment | CI/joint |
| Transition / innovation geometry | residual basis | reconstruction/forecast |
| Transition / innovation / channel | normality-evidence source | point residual/residual induction/local, Q2, or context profile/profile fusion |
| Residual concentration | score aggregation | mean/top-$k$ |
| Tail variation / drift | calibration | frozen percentile candidates |

We call an axis with a declared descriptor block an *admission coordinate*. A matched axis introduced only to isolate a confound is instead a *controlled diagnostic coordinate*: it may establish consequen-

tial heterogeneity but cannot produce a gate. Compact versus token-wise reconstruction has this latter role and is therefore absent from the descriptor catalogue.

The catalogue declares which measurements may explain a response; it does not assume the sign of that response or require every coordinate to be represented by one scalar. A block may encode competing benefit and information-loss terms, but remains deliberately low-dimensional. ETT admits CI by paired gains, the 12-dataset forecasting matrix admits a selective MSE normalization override, and CMAPSS capacity passes conditioning; remaining mappings are hypotheses subject to abstention.

**Definition 1: matched bias response.** The catalogue is a hypothesis map: if a data property should matter, ONE-Core must expose the corresponding choice as a matched contrast. Let $a_{-j}$ denote all nuisance coordinates held fixed. For binary bias axis $\mathcal{A}_j = \{a_j^0, a_j^1\}$, we measure the signed paired relative response

$$\Delta_j(\mathcal{D}; a_{-j}) = \frac{s_j \left[m(\mathcal{D}, a_j^1, a_{-j}) - m(\mathcal{D}, a_j^0, a_{-j})\right]}{\max(|m(\mathcal{D}, a_j^0, a_{-j})|, \epsilon)},$$

where $s_j = 1$ for maximized and $-1$ for minimized metrics. Candidate and reference share split, seeds, backbone family, aggregation, and metric. Their roles are predeclared: $a_j^0$ is the reference and $a_j^1$ the candidate, so $\Delta_j$ is a reference-relative intervention response rather than a symmetric architecture distance. For paired seeds $r = 1, \ldots, n$, we evaluate the same expression as $\Delta_j^{(r)}$, then report $\bar{\Delta}_j$ and the paired Student-$t$ interval $\bar{\Delta}_j \pm t_{0.975,n-1} s_{\Delta_j}/\sqrt{n}$, where $s_{\Delta_j}$ is the sample standard deviation of the paired responses. These intervals are descriptive screens, not multiplicity-corrected confirmatory tests; conditional admission is evaluated separately by nested LODO. We call a response non-tied only when this interval excludes zero and $|\bar{\Delta}_j| > \delta_{\mathrm{PE}} = 0.005$. This predeclared engineering tolerance prevents stable but operationally negligible gains from creating additional routing complexity. Appendix Table 5 verifies that the final admission decisions are unchanged for 0.25–1.0% margins. Uniform resolved responses define a fixed prior or rejection; resolved sign changes make the axis eligible for conditioning; unresolved responses inherit the predeclared fixed prior rather than creating a conditional override. A contrast that changes multiple coordinates is diagnostic and cannot fit an axis gate. Thus $\Delta_j$ is an intervention response, not a label for the dataset's globally best architecture. Definition 1 governs binary admission coordinates. Multiway coordinates, such as the six-source AD audit, are evaluated only as matched diagnostics and do not enter COMPASS gate fitting or admission.

**Necessary condition for conditioning.** Let $r_j(\mathcal{D}, a)$ be axis-wise regret and define candidate gain $g_j(\mathcal{D}) = r_j(\mathcal{D}, a_j^0) - r_j(\mathcal{D}, a_j^1)$. Under the metric normalizations above and below, $g_j(\mathcal{D}) = c_j(\mathcal{D})\Delta_j(\mathcal{D})$ with $c_j(\mathcal{D}) = \max(|m_0|, \epsilon)/\max(|m^*|, \epsilon) > 0$; hence their pointwise signs agree, where $m_0 = m(\mathcal{D}, a_j^0, a_{-j})$ and $m^*$ is the measured axis-wise oracle metric, both with the same nuisance setting $a_{-j}$ (equivalently $m_{\mathcal{D}}(a_j^0)$ and $m_{\mathcal{D}}^*$ below). For the binary matched axis considered here, if $\nu_j(\phi) = \mathbb{E}[g_j(\mathcal{D}) \mid \phi_j(\mathcal{D}) = \phi]$ has the same nonzero sign almost surely over the descriptor support, the same action minimizes conditional expected regret everywhere. Opposite signs on descriptor regions with nonzero probability are therefore necessary, but not sufficient, for a descriptor-conditioned gate to outperform the best fixed action. Finite data add a third regime: unresolved responses inherit the predeclared fixed prior rather than manufacturing heterogeneity from noise.

**ONE-Core factorization.** ONE-Core instantiates this catalogue as a small family of modular models:

$$X \rightarrow \text{local evidence encoder} \rightarrow \text{global temporal abstraction} \rightarrow \text{task readout / score}.$$

A local encoder preserves short-range evidence. A global aggregator compresses temporal information either through a single compact bottleneck or a hierarchical sequence of bottlenecks. A task query forms the primary prediction. Optional auxiliary components are diagnostic branches. The purpose is not to make every task identical, but to change one declared bias while retaining a common implementation and evaluation contract. Accordingly, Single-8 and the 32→8 hierarchy are controlled bias coordinates rather than claims of a new attention primitive.

Single-8 updates the shared task queries once from the last eight encoder tokens; 32→8 updates the same queries first from a wide 32-token context and then from the recent eight. Input stem, local encoder, query count, width, auxiliary branch, and readout remain fixed. Exact tensor shapes and parameter differences appear in Appendix A.

**Coordinate-local admission and abstention.** After the paired-response audit, a restricted candidate rule probes each eligible axis independently. For bias $j$, only its declared descriptor block is visible:

$$\widetilde{S}_j : (\tau, \phi_j(\mathcal{D})) \to \mathcal{A}_j = \{a_j^0, a_j^1\}.$$

Validation then converts the candidate into an admitted coordinate decision,

$$S_j(\tau, \phi_j(\mathcal{D})) = \begin{cases} \widetilde{S}_j(\tau, \phi_j(\mathcal{D})), & j \in \mathcal{J}_{\text{valid}}, \\ \perp, & \text{otherwise.} \end{cases}$$

Only a jointly checked subset $\mathcal{J}_{\text{comp}} \subseteq \mathcal{J}_{\text{valid}}$ may override the fixed configuration:

$$\hat{a}(\mathcal{D}, \tau) = a_\tau^{\text{fixed}} \oplus_{j \in \mathcal{J}_{\text{comp}}} S_j(\tau, \phi_j(\mathcal{D})).$$

The subset's combined endpoint must pass the measured comparison with the retained configuration; individual admission alone does not qualify the union. Here $a_\tau^{\text{fixed}}$ contains admitted task-level coordinates. A $\perp$ on one axis means "no conditional override": it inherits that axis's fixed prior, with $a_j \oplus \perp = a_j$, and leaves all other decisions intact. It neither suppresses a sample prediction nor reduces task coverage. We call this family of interpretable gates COMPASS. It is neither a learned task classifier nor unrestricted architecture search: matched responses act as finite differences, while descriptor blocks propose restricted surrogate directions whose utility must still be measured. Exploratory multi-coordinate probes are limited to depth two over one mechanism-linked descriptor block and one matched response matrix. The admitted forecast normalization and in-family RUL capacity gates are stricter one-dimensional single-split stumps over $D_\sigma^{96}$ and $B_{\text{fast}}$, respectively. Crucially, ONE-Core does not choose whichever descriptor axis happens to minimize training regret inside a fold; that would introduce a second, hidden model-selection layer.

A nested leave-one-dataset-out (LODO) audit refits every threshold and leaf without the held-out dataset. Admission requires lower mean regret than the best fixed prior and no worse performance on at least two thirds of folds. Descriptor-row permutation is reported separately as a specificity diagnostic; we use exact enumeration where feasible and fixed-seed Monte Carlo for normalization, not as a calibrated significance test or hard admission cutoff. Composition is permitted only after the combined endpoint has been measured; individually favorable one-axis effects are not assumed additive. These are finite-sample empirical guardrails, not an asymptotic confidence guarantee. Permutation breaks the descriptor–response pairing and is a specificity stress test; with four C-MAPSS subsets, 8/24 permutations attain zero regret (Appendix A). Table 2 summarizes the resulting decisions; exact threshold generation, leaf enumeration, corpus roles, and permutation scopes appear in Appendix A.

**Protocol.** Each bias uses its own matched matrix and declared training-only descriptor block. Forecasting and RUL gates use nested LODO over datasets and C-MAPSS subsets, respectively; the held-out metric never enters its own fit. AD uses development and secondary falsification diagnostics but fits no readout, evidence-source, or granularity gate. The broader 44-dataset corpus checks descriptor coverage only; complete dataset roles and permutation scopes appear in Appendix A.

**Algorithm 1: bias-wise evidence-gated coordinate search.** Given tasks $\tau$, development datasets $\mathcal{D}_{\text{dev}}$, direct extractor $\phi$, specified bias catalogue $\mathcal{B} = \{(j, \phi_j, a_j^0, a_j^1)\}$, and module map $M$: (1) compute $\phi(\mathcal{D})$ from training partitions; (2) for every $j$, instantiate the paired intervention $M(a_j^0, a_{-j})/M(a_j^1, a_{-j})$; (3) estimate $\Delta_j$ with paired seeds and classify the axis as fixed, rejected, tied, conditional, or confounded; (4) for each conditional $j$, fit the candidate $\widetilde{S}_j$ using only $\phi_j$ and that axis's response matrix; (5) rerun Step 4 inside every LODO fold and report descriptor-row permutation separately; (6) set $S_j = \widetilde{S}_j$ only when it lowers mean LODO regret and is no worse than the best single fixed action on at least $\lceil 2n/3 \rceil$ of $n$ folds (regret tolerance $10^{-12}$), otherwise set $S_j = \perp$; (7) measure the endpoint before composing multiple validated axes; and (8) evaluate wrong-choice and calibration controls. Fixed task biases are already produced by Steps 1–3, so abstention is local

Table 2: **Bias-specific stability decisions.** Forecast evidence is candidate/reference/tie under paired 95% intervals and a 0.5% equivalence margin. Fixed regret uses the best single action, including RUL dim16 (2.66%); its fold-fitted constant is a separate diagnostic (5.34%). The abstraction-form probe selects hierarchy on all six folds. The RevIN row uses the homogeneous 12-dataset, five-seed, MSE-trained matrix; MAE is excluded from admission and reported only for benchmark comparison.

| Task / bias | Evidence | Gate regret | Fixed regret | Decision |
|---|---|---|---|---|
| Forecast / RevIN | 7 on / 5 off (MSE) | **2.11%** | 10.15% | admit low-$D_\sigma$ override |
| Forecast / CI vs joint | 4/0/0 | – | – | fixed CI |
| Forecast / Global vs Patch | 0/2/6 | 0.24% | 0.24% | fix Patch |
| Forecast / 32→8 vs Single-8 | 2/0/4 | 0.49% | 0.49% | fix 32→8 |
| Forecast / $K$=2 vs 4 | 0/0/6 | 0.29% | 0.29% | fix $K$=4 |
| Forecast / dim16 vs 32 | 0/3/3 | 0.00% | 0.00% | fix dim32 |
| Forecast / 64→16 vs 32→8 | 0/3/3 | 0.03% | 0.03% | fix 32→8 |
| RUL / dim32 vs 16 | 2/2/0 | **0.00%** | 2.66% | admit in-family gate |
| AD / readout geometry | heterogeneous | – | – | diagnostic; no routing |
| AD / normality evidence | heterogeneous | – | – | diagnostic envelope; abstain |

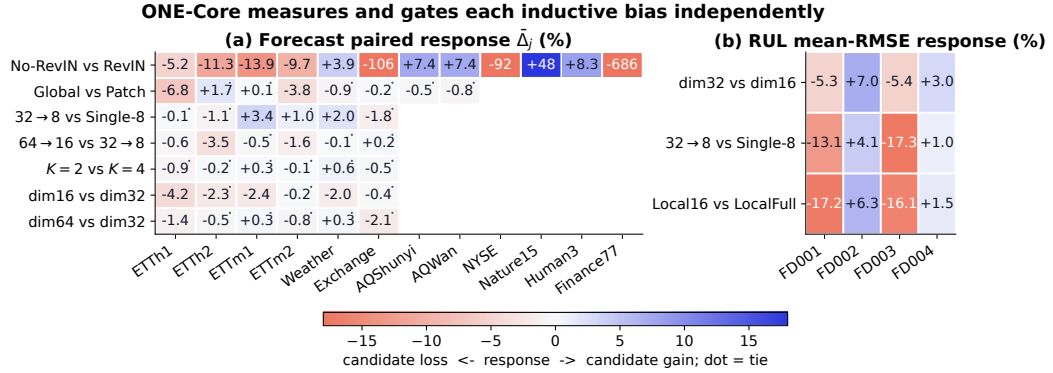


Figure 2: **Matched bias-response landscape.** The 12-dataset No-RevIN row is the homogeneous MSE normalization matrix; positive response favors no-RevIN. Colors are clipped at $\pm 18\%$ for cross-axis readability, while annotations show unclipped responses rounded for display. Abstraction form and hierarchy range are separate axes; form is compared on four ETT variants, Weather, and Exchange at H96, dim16, $K$=2. A small dot marks paired uncertainty/equivalence. Forecast values average paired seed responses; RUL values are relative changes of archived mean RMSE. RUL one-axis responses separate FD001/FD003 from FD002/FD004 and motivate, but do not substitute for, a measured composition endpoint. Table 2 maps these responses to fixed, conditioned, or abstained decisions.

to one remaining coordinate. A subsequent audit starts from the retained configuration and may introduce a new mechanism-linked coordinate or descriptor block. The procedure therefore accumulates locally validated updates; it neither enumerates all possible time-series properties nor claims convergence to a global architecture oracle.

**Admission risk and empirical guardrails.** For axis $j$, let $\mathcal{A}_j = \{a_j^0, a_j^1\}$ and let $m_{\mathcal{D}}(a)$ denote its evaluation metric. Normalized regret relative to the dataset-wise oracle is

$$r_j(\mathcal{D}, a) = \begin{cases} \dfrac{m_{\mathcal{D}}(a) - m^*_{\mathcal{D}}}{\max(|m^*_{\mathcal{D}}|, \epsilon)}, & m^*_{\mathcal{D}} = \min_{a' \in \mathcal{A}_j} m_{\mathcal{D}}(a') \quad \text{(lower is better)}, \\ \dfrac{m^*_{\mathcal{D}} - m_{\mathcal{D}}(a)}{\max(|m^*_{\mathcal{D}}|, \epsilon)}, & m^*_{\mathcal{D}} = \max_{a' \in \mathcal{A}_j} m_{\mathcal{D}}(a') \quad \text{(higher is better)}. \end{cases}$$

Thus the oracle has zero regret and every inferior action has positive regret; $\epsilon > 0$ stabilizes near-zero denominators. This piecewise definition performs the same metric-direction alignment as $s_j$ in the matched response, but changes the baseline from the predeclared reference to the axis-wise

dataset oracle. A descriptor-conditioned candidate is a restricted function $\widetilde{S}_j(\tau, \phi_j(\mathcal{D})) \in \mathcal{A}_j$ fit by

$$\min_{\widetilde{S}_j} \sum_{\mathcal{D} \in \mathcal{D}_{\text{dev}}} r_j(\mathcal{D}, \widetilde{S}_j(\tau, \phi_j(\mathcal{D}))).$$

This objective prevents one gate from compensating for a bias by silently selecting another architecture coordinate. Bias admission precedes it: a uniformly useful bias is fixed, a confounded contrast is excluded, and an unstable conditional rule inherits the task prior.

More explicitly, let $a_j^{\text{fix}}$ be the fixed action with lowest mean axis regret and let $R_j(\widetilde{S}_j) = \mathbb{E}_{\mathcal{D}}[r_j(\mathcal{D}, \widetilde{S}_j(\phi_j(\mathcal{D})))]$. The population-level admission principle is $R_j(\widetilde{S}_j) < R_j(a_j^{\text{fix}})$. ONE-Core does not claim to estimate a tight confidence bound on this quantity from these small dataset corpora; Algorithm 1 operationalizes the principle with nested LODO and the stated guardrails.

**Composition: one-axis admission does not imply composition.** For two binary axes, let $m_{uv}(\mathcal{D})$ denote their measured endpoint and define $I_{jk} = m_{11} - m_{10} - m_{01} + m_{00}$. Zero denotes additivity on this metric scale. Dividing all four endpoints by the same positive reference/oracle denominator rescales $I_{jk}$ and preserves zero; separate normalizers or nonlinear transforms need not do so. The RUL experiment checks a three-axis endpoint against the retained capacity gate. Its failure rejects that proposed update, but does not estimate pairwise $I_{jk}$ or isolate which interaction caused the loss: intermediate two-axis endpoints were not measured.

# 4 Experiments

We ask RQ1: which matched biases help; RQ2: does conditioning transfer; RQ3: what is the wrong-choice cost; and RQ4: when should the prior be retained? Complete matrices and negative controls appear in Appendix C. Each task uses its standard split, validation-selected checkpoints, and five matched seeds. Forecasting uses input length 96, MSE, and horizons 96/192/336/720; RUL uses NASA CMAPSS splits; AD uses normal-only training without point adjustment.

## 4.1 Task-Level Testbed Results

**Forecasting.** The structural reference is Patch+CI, dim32, $K{=}4$, and 32→8. The separate Local-Full, dim16, $K{=}2$ abstraction audit adds Weather ($+1.98\%$ hierarchy gain, 95% CI $[1.36, 2.60]\%$) and Exchange (unresolved). Its six-dataset descriptor probe matches fixed hierarchy at 0.49% regret versus 1.09% for fixed Single-8, retaining the hierarchy prior (Table 19). Other structural gates likewise do not reduce nested-LODO regret; CI removal costs 3.34–23.13% MSE. At H96, RevIN wins 7/12 datasets. The scale-drift gate reduces mean regret from 10.15% to 2.11% (family-LODO: 2.92%): two folds improve, nine tie, and one worsens. Both median regrets are zero; Nature15 contributes 95.54% of total regret reduction (fold-distribution details in Appendix D). High-dispersion AQShunyi, AQWan, and Weather still favor no-RevIN, so dispersion alone does not determine the preferred action. Final ETT H96 MSEs are 0.3850/0.3301/0.3369/0.1836; Table 9 provides literature comparisons with published forecasting baselines.

**RUL.** RUL favors Single-8 summarization and then conditions capacity. Nested LODO uses training-only $B_{\text{fast}}$ to select dim16 for FD001/FD003 and dim32 for FD002/FD004, reducing regret from 2.66% for a single fixed dim16 to 0.00%. The fold-fitted constant baseline has 5.34% regret. ONE-Core averages 13.73 RMSE with 17K–53K parameters; composition fails, while XJTU-SY (Wang et al., 2020) and PRONOSTIA (Nectoux et al., 2012) only directionally corroborate dim32 (Tables 28 and 29).

**Anomaly detection.** Under one encoder, R0/R1/Forecast rankings depend on score geometry and calibration. On SWaT, R1-MAE exceeds Forecast-MAE by $+0.0310$ oracle Best-F1 [0.0204, 0.0416], but remains worse under frozen-threshold F1 (0.2591 vs. 0.3267; Table 36). This diagnostic reversal alone does not establish deployment preference. Six normality-evidence paths yield fixed-path macro Best-F1 values of 0.297–0.412 versus a 0.438 dataset-wise envelope. Only five heterogeneous datasets support this six-way audit, so routing abstains; no evidence-source gate is fitted (Tables 34–35).

## 5 Discussion, Limitations, and Conclusion

**Supported claims and boundaries.** ONE-Core provides local evidence for forecasting structural priors, corpus-dependent forecasting normalization, and in-family RUL capacity; AD exposes heterogeneous evidence sources without admitting a router. The testbed supports matched coordinate comparisons. Published benchmark results already include recent strong forecasting baselines, notably TimePerceiver (2025) and TimeMixer (2024), alongside RUL and AD specialists (Tables 9–11). These comparisons locate ONE-Core's performance; external methods have not been jointly rerun with matched preprocessing, seeds, tuning budgets, and evaluation.

**Empirical resolution.** RUL's four C-MAPSS subsets support a feasibility result, not field readiness. The dim16/dim32 split coincides with the single-/multiple-operating-condition split, so spectral balance is not disentangled from regime count. Exactly 8/24 descriptor permutations also attain zero regret; 33.3% is weak specificity evidence, not statistical significance. Bearing intervals remain broad. $D_\sigma^L$ measures scale dispersion, not nonstationarity itself: periodic amplitude variation can also raise it. Its amplitude-preservation rationale remains a hypothesis. The binary in-family RUL probe and six-way cross-domain AD audit differ in both corpus and rule complexity.

**Practical value and next steps.** Accumulated audits could support *cumulative model-design intelligence*: a reusable record of validated priors, rejected updates, and unresolved hypotheses that informs subsequent design decisions across deployments. The expanded abstraction audit illustrates prior retention after refitting; future descriptors may support an update if they pass the same validation. Broader claims require larger independent corpora, matched specialist reruns, order-sensitive search experiments, and intermediate composition ablations. Cross-deployment transfer and a complete sequential search remain unevaluated. Within this scope, matched bias-wise interventions and local abstention provide reusable, falsifiable model-design evidence from heterogeneous architecture outcomes.

### AI use statement

We developed the core model-design ideas and research hypotheses. Generative AI tools were used to assist with planning and operating experiments based on those ideas, including experimental controls, ablations, seed configurations, evaluation procedures, and run orchestration; limited portions of code implementation and debugging; organizing and checking experimental results and tables; conceptual comparison with related work; and editing the manuscript for clarity, consistency, and formatting; mathematical formulation and proof checking; and translation for manuscript review. All experimental decisions, code changes, numerical results, citations, and claims were independently inspected and verified by us. AI-generated outputs were not treated as empirical evidence; we take full responsibility for the final content.

### Reproducibility statement

Dataset splits, seeds, candidate configurations, descriptor extraction rules, nested-LODO fitting, and evaluation procedures are documented in Appendices A and C. The anonymous supplementary package contains the model implementation, data modules, experiment manifests, gate evaluator, and machine-readable result summaries used to verify the paper's decisions.

## References

Ahmed Abdulaal, Zhuanghua Liu, and Tomer Lancewicki. Practical approach to asynchronous multivariate time series anomaly detection and localization. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2485–2494, 2021. doi: 10.1145/3447548.3467174.

Julien Audibert et al. USAD: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2020.

Giduthuri Sateesh Babu, Peilin Zhao, and Xiao-Li Li. Deep convolutional neural network based regression approach for estimation of remaining useful life. In *Database Systems for Advanced Applications*, pp. 214–228. Springer, 2016. doi: 10.1007/978-3-319-32025-0_14.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Cristian Challu, Kin G. Olivares, Boris N. Oreshkin, Federico Garza Ramirez, Max Mergenthaler Canseco, and Artur Dubrawski. NHITS: Neural hierarchical interpolation for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 6989–6997, 2023. doi: 10.1609/aaai.v37i6.25854.

Abhimanyu Das et al. A decoder-only foundation model for time-series forecasting. In *International Conference on Machine Learning*, 2024.

Ailin Deng and Bryan Hooi. Graph neural network-based anomaly detection in multivariate time series. In *AAAI Conference on Artificial Intelligence*, 2021.

Raul Dinten and Marta Zorrilla. Using time series foundation models for few-shot remaining useful life prediction of aircraft engines. *Computer Modeling in Engineering & Sciences*, 2025.

Xingjian Gao et al. UniTS: Building a unified time series model. In *Advances in Neural Information Processing Systems*, 2024.

Astha Garg, Wenyu Zhang, Jules Samaran, Savitha Ramasamy, and Chuan-Sheng Foo. An evaluation of anomaly detection and diagnosis in multivariate time series. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2508–2517, 2022.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Mononito Goswami et al. MOMENT: A family of open time-series foundation models. In *International Conference on Machine Learning*, 2024.

Juntao Hua, Yupeng Zhang, Dingcheng Zhang, Jiayuan He, Jie Wang, and Xia Fang. Dynamic feature-aware graph convolutional network with multisensor for remaining useful life prediction of turbofan engines. *IEEE Sensors Journal*, 24(18):29414–29428, 2024. doi: 10.1109/JSEN.2024.3435071.

Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2018.

Siwon Kim, Kukjin Choi, Hyun-Soo Choi, Byunghan Lee, and Sungroh Yoon. Towards a rigorous evaluation of time-series anomaly detection. In *AAAI Conference on Artificial Intelligence*, 2022a.

Taesung Kim et al. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2022b.

Chih-Yu Lai, Fan-Keng Sun, Zhengqi Gao, Jeffrey H. Lang, and Duane Boning. Nominality score conditioned time series anomaly detection by point/sequential reconstruction. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-3348.

Kwei-Herng Lai et al. Revisiting time series outlier detection: Definitions and benchmarks. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

J. Lee and H. Lee. TimePerceiver: An encoder-decoder framework for generalized time-series forecasting. In *Advances in Neural Information Processing Systems*, 2025.

Yaguo Lei et al. Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mechanical Systems and Signal Processing*, 2018.

Junhuai Li, Kan Wang, Xiangwang Hou, Dapeng Lan, Yunwen Wu, Huaijun Wang, Lei Liu, and Shahid Mumtaz. A dual-scale transformer-based remaining useful life prediction model in industrial internet of things. *IEEE Internet of Things Journal*, 11(16):26656–26667, 2024. doi: 10.1109/JIOT.2024.3376706.

Xiang Li, Qian Ding, and Jian-Qiao Sun. Remaining useful life estimation in prognostics using deep convolution neural networks. *Reliability Engineering & System Safety*, 2018.

Zixuan Liang and Yanan Sun. Evolutionary neural architecture search for multivariate time series forecasting. In *Proceedings of the 15th Asian Conference on Machine Learning*, pp. 771–786, 2024.

Yong Liu et al. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024a.

Yong Liu et al. Timer: Generative pre-trained transformers are large time series models. In *International Conference on Machine Learning*, 2024b.

Donghao Luo and Xue Wang. ModernTCN: A modern pure convolution structure for general time series analysis. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vpJMJerXHU.

Erfan Moeini, Christopher Vox, Marie Anastacio, Wadie Skaf, Mitra Baratchi, and Holger H. Hoos. Neural architecture and hyperparameter selection through meta-learning on time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24405–24413, 2026. doi: 10.1609/aaai.v40i29.39622.

Pablo Montero-Manso, George Athanasopoulos, Rob J. Hyndman, and Thiyanga S. Talagala. FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting*, 36 (1):86–92, 2020.

Patrick Nectoux, Rafael Gouriveau, Kamal Medjaher, Emmanuel Ramasso, Brigitte Chebel-Morello, Noureddine Zerhouni, and Christophe Varnier. PRONOSTIA: An experimental platform for bearings accelerated degradation tests. In *IEEE International Conference on Prognostics and Health Management*, pp. 1–8, 2012.

Yuqi Nie et al. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, Zhenli Sheng, and Bin Yang. TFB: Towards comprehensive and fair benchmarking of time series forecasting methods. *Proceedings of the VLDB Endowment*, 17 (9):2363–2377, 2024.

Abhinav Saxena et al. Damage propagation modeling for aircraft engine run-to-failure simulation. In *International Conference on Prognostics and Health Management*, 2008.

Ya Su et al. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2019.

Nesime Tatbul, Tae Jun Lee, Stan Zdonik, Mejbah Alam, and Justin Gottschlich. Precision and recall for time series. In *Advances in Neural Information Processing Systems*, 2018.

Shreshth Tuli, Giuliano Casale, and Nicholas R. Jennings. TranAD: Deep transformer networks for anomaly detection in multivariate time series data. *Proceedings of the VLDB Endowment*, 2022.

Biao Wang, Yaguo Lei, Naipeng Li, and Ningbo Li. A hybrid prognostics approach for estimating remaining useful life of rolling element bearings. *IEEE Transactions on Reliability*, 69(1):401–412, 2020. doi: 10.1109/TR.2018.2882682.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7oLshfEIC2.

Gerald Woo et al. Unified training of universal time series forecasting transformers. In *International Conference on Machine Learning*, 2024.

Haixu Wu et al. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, 2021.

Haixu Wu et al. TimesNet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

Jiehui Xu et al. Anomaly transformer: Time series anomaly detection with association discrepancy. In *International Conference on Learning Representations*, 2022.

Yiyuan Yang, Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun. DCdetector: Dual attention contrastive representation learning for time series anomaly detection. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023. doi: 10.1145/3580305.3599295.

Annan Yu, Danielle C. Maddix, Boran Han, Xiyuan Zhang, Abdul Fatir Ansari, Oleksandr Shchur, Christos Faloutsos, Andrew Gordon Wilson, Michael W. Mahoney, and Yuyang Wang. Understanding the implicit biases of design choices for time series foundation models. In *International Conference on Learning Representations*, 2026.

Ailing Zeng et al. Are transformers effective for time series forecasting? In *AAAI Conference on Artificial Intelligence*, 2023.

Hang Zhao, Yujing Wang, Juanyong Duan, Congrui Huang, Defu Cao, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, and Qi Zhang. Multivariate time-series anomaly detection via graph attention network. In *2020 IEEE International Conference on Data Mining*, pp. 841–850, 2020. doi: 10.1109/ICDM50108.2020.00093.

Shuai Zheng, Kosta Ristovski, Ahmed Farahat, and Chetan Gupta. Long short-term memory network for remaining useful life estimation. In *2017 IEEE International Conference on Prognostics and Health Management*, pp. 88–95, 2017. doi: 10.1109/ICPHM.2017.7998311.

Haoyi Zhou et al. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *AAAI Conference on Artificial Intelligence*, 2021.

Tian Zhou et al. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*, 2022.

## A Descriptor Definitions and Gate Implementation

**Direct descriptor vocabulary.** Persistence includes autocorrelation decay and the Hurst exponent; frequency geometry includes spectral entropy, low/high-frequency ratios, dominant-component mass, and context-relative variation; shift includes trend and block mean/scale drift; channel geometry includes covariance effective rank and lagged coupling; and AD statistics include transition density, innovation scale, residual concentration, and block-tail variation. Descriptor roles are coordinate-local. The concatenated vector audits corpus coverage and collinearity, but a gate consumes only its declared block.

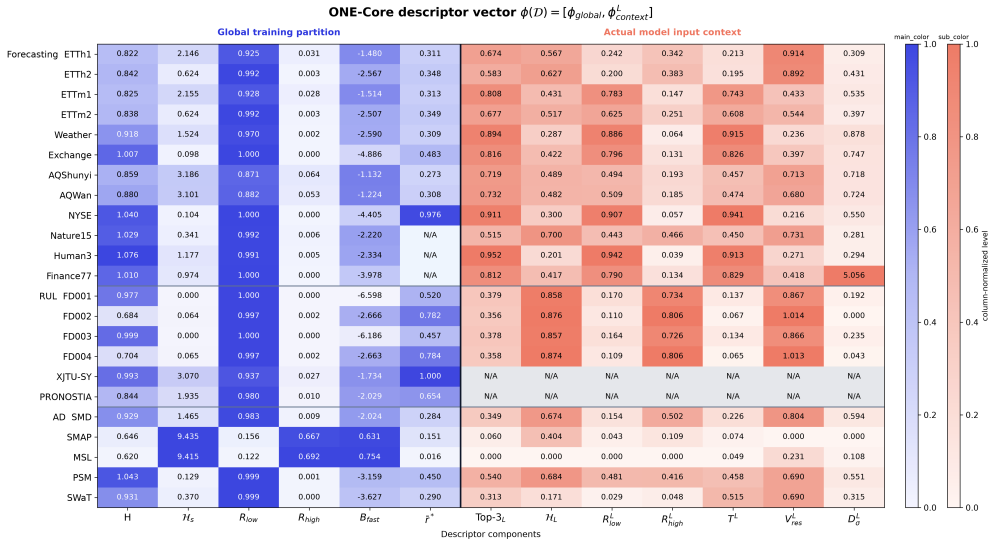


Figure 3: **Descriptor-block coverage audit.** Each row displays global training-partition statistics and input-context statistics computed at the model's actual receptive field ($L$=96/50/100 for Forecast/RUL/AD); context bands are relative to $L$. The figure's global/context grouping refers to measurement scale; each group draws from the phenomenon-based blocks $\phi_{\text{persist}}, \ldots, \phi_{\text{tail}}$ in Section 3, rather than defining a different descriptor vector. $D^L_\sigma$ is the relative input-window scale drift used by the MSE forecasting-normalization gate; $B_{fast}$ is the run-level RUL capacity descriptor. $D^L_\sigma$ is displayed for every directly comparable task input but is consumed only by the forecasting gate. N/A denotes an undefined channel statistic for univariate data or a non-comparable input representation; XJTU and PRONOSTIA context cells are omitted because their candidate-dependent raw-vibration adapters change width. Evaluation roles are reported in the protocol tables; this coverage map does not encode development/holdout status, and each gate consumes only its declared block.

The direct statistics used here come from the standing training-only catalogue, rather than from held-out performance. Descriptor-to-bias mappings and low-dimensional aggregations may be refined during exploratory mechanism analysis. Once an admission mapping is declared, its extractor and rule class remain fixed within nested and family-grouped LODO. These resampling tests therefore evaluate the declared mapping conditional on the catalogue; they do not claim to correct unrestricted descriptor discovery.

**RUL fast/slow spectral balance.** For each training run $e$, let $x_{e,c}(t)$ be the cycle-wise sensor trajectory for C-MAPSS and the per-snapshot RMS trajectory for bearings. XJTU-SY reduces each 32,768-sample, two-channel snapshot to two RMS values (one snapshot per minute); PRONOSTIA similarly reduces each 2,560-sample, two-channel snapshot (one per 10 seconds). Thus bearing spectra describe slow RMS evolution, not the within-snapshot vibration spectrum; the predictor separately uses raw vibration through a learned adapter.

For a trajectory of length $n_e$, the released FFT extractor retains exactly $\mathcal{K}_e = \{0, \ldots, \lfloor n_e/2 \rfloor - 1\}$, corresponding to `fft(x)[:n//2]`. Write $f_k = k/n_e$ and $f_N = 1/2$ in observation-grid units and normalize $p_{e,c,k} = |\mathcal{F}x_{e,c}[k]|^2 / \sum_{\ell \in \mathcal{K}_e} |\mathcal{F}x_{e,c}[\ell]|^2$ (all ratios are zero when retained power is zero). This half-open convention excludes Nyquist for even $n_e$ and also omits the highest positive-frequency bin for odd $n_e$; it is not an inclusive real FFT. Let $R^{(e)}_{\text{low}} = \text{median}_c \sum_{k \in \mathcal{K}_e : f_k < 0.1 f_N} p_{e,c,k}$ and $R^{(e)}_{\text{high}} = \text{median}_c \sum_{k \in \mathcal{K}_e : f_k \geq 0.3 f_N} p_{e,c,k}$. Then

$$B_{\text{fast}}(D) = \text{median}_{e \in D} \log_{10} \left[ \frac{R^{(e)}_{\text{high}} + \epsilon}{R^{(e)}_{\text{low}} + \epsilon} \right], \qquad \epsilon = 10^{-8}. \tag{1}$$

Trajectories are neither mean-centered nor joined across run boundaries. Retaining DC makes this a level-relative statistic. Extraction precedes any learned modality adapter and cannot depend on

candidate width. Bands refer to the observation grid, not shared physical Hz across modalities. Operating-regime counts are contextual information, not gate inputs.

**Forecast context and normalization descriptors.** Length-$L$ training windows are mean-centered before positive Fourier power is normalized. We aggregate entropy, top-$k$ mass, architecture-relative low/high bands, and variation; low periods are at least $2P$ and high periods at most $P$ for patch length $P$. Primary experiments use $L$=96, $P$=16, and top-3 mass. A centered moving average of span $2\lfloor L/8 \rfloor + 1$ yields trend strength $T^L = 1 - \mathrm{Var}(x - \mathrm{MA}(x))/\mathrm{Var}(x)$, residual volatility $V^L_{\mathrm{res}} = \mathrm{MAD}(x - \mathrm{MA}(x))/\mathrm{MAD}(x)$, and normalized drift $D^L = |\operatorname{med}(x_{3L/4:L}) - \operatorname{med}(x_{0:L/4})|/\mathrm{IQR}(x)$.

RevIN directly divides each input window by its channel-wise standard deviation. The primary normalization descriptor therefore measures how much that acted-on statistic varies over training contexts. For channel $c$, let $\sigma_{w,c}$ be the standard deviation of length-$L$ input window $w$. We define relative window-scale drift

$$D^L_\sigma(D) = \operatorname{median}_c \frac{Q_{0.75,w}(\sigma_{w,c}) - Q_{0.25,w}(\sigma_{w,c})}{\operatorname{median}_w(\sigma_{w,c}) + 10^{-12}}. \tag{2}$$

Standard deviations use population normalization. Only training windows enter; starts are spaced by $\lfloor L/4 \rfloor$ and, if over budget, uniformly subsampled by index. The primary normalization extractor caps channels at 32 (uniformly spaced indices) and windows at 512. The general context-coverage extractor instead defaults to 16 channels and 256 windows, with each eligible entity allocated $\lceil 256/n_{\mathrm{entities}} \rceil$ windows. Within each entity, nonfinite or sentinel values ($\leq -9000$) are replaced by channel medians (all-missing medians become zero), and channels with training standard deviation $\leq 10^{-10}$ are removed; if all are constant, the first is retained. For multiple entities the final median pools their channel-wise ratios without crossing boundaries. Thus Eq. (2)'s stabilizer is distinct from Eq. (1)'s $10^{-8}$, and no learned descriptor weight is used. An intermittent channel may have zero median window scale despite nonzero training variation; it remains and uses the same $10^{-12}$ denominator stabilizer. For the primary $L$=96 MSE matrix, the full-corpus split lies near $D^{96}_\sigma = 0.301$: lower values trigger no-RevIN and higher values retain RevIN. Every LODO fold regenerates the midpoint threshold and both leaf actions without its held-out response. Low scale dispersion motivates the hypothesis that preserving absolute amplitude can help; the fitted gate uses RevIN as its high-dispersion fallback. Dispersion alone neither proves nuisance variation nor distinguishes stochastic nonstationarity from periodic scale changes. This is an empirically admitted selective override, not a complete law: high-drift AQShunyi, AQWan, and Weather remain measured no-RevIN exceptions.

For diagnostic completeness, we also compute context-to-future shift burden $D^{L,H}_{\mathrm{shift}}$ from paired training windows. It does not lower nested regret as reliably as $D^L_\sigma$ and therefore does not enter the admitted rule. The earlier scalar $P_{\mathrm{scale}}$ analysis is treated only as the exploratory hypothesis that generated the Weather/AQ closure.

**Abstraction-form context-memory probe.** For the six-dataset abstraction audit, the descriptor block comprises four native-context autocorrelation summaries. Using training windows of length 96, stride 24, at most 256 evenly sampled windows and 16 evenly sampled channels, we demean each window/channel and compute zero-padded FFT autocorrelation normalized by lag-zero energy (floor $10^{-12}$). We average its absolute value over lags 1–16 (short), 17–48 (long), and 1–48 (integral), and compute long/(short+$10^{-12}$). Each descriptor is the median over window/channel pairs. Preprocessing follows the general context extractor above; the channel cap applies to descriptors, not to model inputs. This memory block was specified during exploratory analysis of the expanded matrix, then frozen for the completed five-seed LODO audit, consistent with the catalogue-conditional validation scope above.

The probe uses the existing regression-tree class: depth at most two, minimum leaf size two, squared-error splitting, and random seed zero. Training targets are paired mean responses, with unresolved responses set to zero. Every fold refits splits and leaf predictions using the other five datasets; predicted responses above $+0.005$ select hierarchy, below $-0.005$ select Single-8, and otherwise inherit the training-fold best fixed action. The absence of a resolved sign change already favors prior retention; we also run this restricted LODO probe to check for added conditional value. It selects hierarchy on all six folds. The full-corpus tree splits at long ACF 0.185507, but both leaves

select hierarchy, so this threshold creates no conditional action change. All $6! = 720$ descriptor permutations match its regret.

**Abstraction implementation.** The encoder produces $T_E$ tokens in $\mathbb{R}^d$ and both abstraction paths return the same $K$ learned task-query tokens. Single-8 performs one cross-attention/FFN update from the last $\min(T_E, 8)$ encoder tokens; "8" is the context length, not the number of output latents. The 32→8 path applies two distinct residual cross-attention/FFN updates to the same queries, first from the last $\min(T_E, 32)$ tokens and then from the last $\min(T_E, 8)$ tokens. Primary forecasting uses $K$=4, $d$=32; the extra stage adds 12,768 parameters (131,767 vs. 118,999 total). At the RUL setting $K$=2, $d$=16, it adds 3,312 parameters. For forecasting, $L$=96, patch length 16, stride 8, and right replication padding 8 yield $T_E = \lfloor (96 + 8 - 16)/8 \rfloor + 1 = 12$ tokens. Consequently, 32→8 accesses 12 then 8 tokens, whereas 64→16 accesses 12 then 12. These are encoder-token access spans, not raw-input receptive fields; encoder attention can mix information before abstraction. Both the additional stage and its parameter cost remain real.

**Threshold and leaf fitting.** For the admitted one-dimensional stumps, unique training-fold descriptor values are sorted and their consecutive midpoints form candidate thresholds. Both assignments of $a_j^0/a_j^1$ to the two leaves are evaluated exhaustively by summed axis regret, together with the best fixed action as a constant-rule candidate. The held-out dataset contributes neither its descriptor-response pair nor its metric to fitting. This procedure is rerun independently for every held-out dataset.

Two constant-action comparators must be distinguished. The RUL fold-fitted baseline selects dim16 or dim32 on each training fold, giving 5.34% held-out regret. A single action fixed over all four C-MAPSS subsets gives 2.66% for dim16 and 2.68% for dim32. The gate has 0.00% regret against either comparator: it strictly improves all four folds over the fold-fitted baseline, but two folds over either single fixed action, tying the other two. These values use the same released capacity matrix, not additional training. Exactly 8 of the 24 descriptor-row permutations reproduce zero nested-LODO regret. For the observed balanced two-versus-two response split, the number of full-data orderings separable by a stump with either orientation is $2 \times 2! \times 2! = 8$. The released nested procedure attains the same count for this particular descriptor geometry; this is not a universal theorem about four-subset LODO. The tied rank 33.3% provides no strong specificity evidence.

Forecast interface responses use ETTh1/2, ETTm1/2, Weather, Exchange, and two AQ datasets; abstraction form uses the four ETT variants, Weather, and Exchange; query, capacity, and hierarchy range use the six fully factorial datasets. RUL uses FD001–FD004 for in-family LODO and full XJTU-SY and PRONOSTIA as external bearing-modality audits. Forecast normalization uses the homogeneous 12-dataset H96 matrix and five matched seeds; its $D_\sigma^{96}$ gate additionally groups all ETT and AQ rows in a family-LODO robustness audit. Descriptor-row permutation is repeated per axis ($8!$, $6!$, or $4!$; normalization uses 10,000 fixed-seed random permutations) as a specificity diagnostic, not a calibrated test. SMD/SMAP/PSM are AD development diagnostics and MSL/SWaT are secondary falsification diagnostics. The five datasets span different entity, transition, rank, and coupling regimes, so neither an AD readout gate nor an evidence-source gate is fitted or deployed.

**Prospective organizational use.** Repeated audits could accumulate into what we call *cumulative model-design intelligence*: a reusable body of model-design evidence in which validated priors transfer across deployments and unsupported coordinates are isolated for targeted reevaluation. In industrial settings with heterogeneous operating regimes and sensing configurations, such accumulated intervention evidence may eventually become more reusable than any particular backbone. This prospective implication is not evaluated here.

Table 3: **Matched bias-response audit.** Forecast counts are candidate/reference/tie under paired 95% intervals and a 0.5% equivalence margin. RUL directions come from the matched five-seed one-axis matrix; only capacity is admitted within the CMAPSS family. AD contains a first-stage three-way readout diagnostic and a second-stage six-way normality-evidence diagnostic.

| Task | Matched contrast | Response | Descriptor block | Decision |
|---|---|---|---|---|
| Forecast | RevIN on/off | 7 RevIN / 5 no-RevIN (MSE) | $D_\sigma^{96}$ | admit low-drift override |
| Forecast | CI/joint | 4/0/0; +3.34–23.13% | channel rank/coupling | fixed CI |
| Forecast | Global/Patch | 0/2/6 | context spectrum/variation | fix Patch |
| Forecast | 32→8 / Single-8 | 2/0/4 | context-scale memory | fix 32→8 |
| Forecast | $K$=2/4 | 0/0/6 | modes/memory variation | fix $K$=4 |
| Forecast | dim16/32 | 0/3/3 | innovation/temporal rank | fix dim32 |
| Forecast | 64→16 / 32→8 | 0/3/3 | memory/period alignment | fix 32→8 |
| RUL | dim32/16 | 2/2/0 | log fast/slow balance | admit in-family gate |
| RUL | hierarchy/Single-8 | 2/2/0 | persistence/complexity | endpoint audit |
| RUL | local16/full | 2/2/0 | persistence/complexity | endpoint audit |
| RUL | composed/reference endpoint | 1/2/1 | matched endpoint | retain capacity gate |
| AD | R0/R1/Forecast | heterogeneous | controlled diagnostic | no routing |
| AD | normality-evidence source | heterogeneous; envelope 0.438 | transition/channel profile | diagnostic; no gate |

## B Additional Trajectory Figures

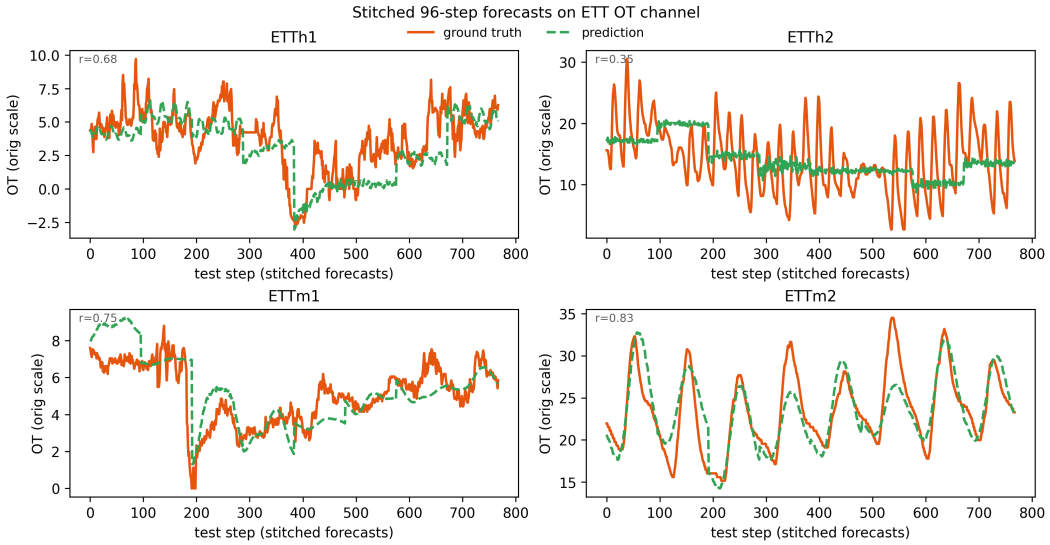


Figure 4: **Qualitative ETT forecasts.** Stitched 96-step forecasts on the official LTSF test split using the final post-LODO Patch+RevIN+CI ETT policy. For every dataset, a model-independent rule selects the aligned 768-step ground-truth span whose variance is closest to the median across eligible test spans. Because this rule is independent of prediction quality, the display includes both well-tracked and difficult regions rather than a best-looking sample. Each 96-step block receives a new, fully observed past context; stride-96 blocks are stitched across 768 steps, not recursively fed back as a 768-step autoregressive forecast. The plotted channel is OT and seed is 42. Quantitative comparisons are reported in Table 4.

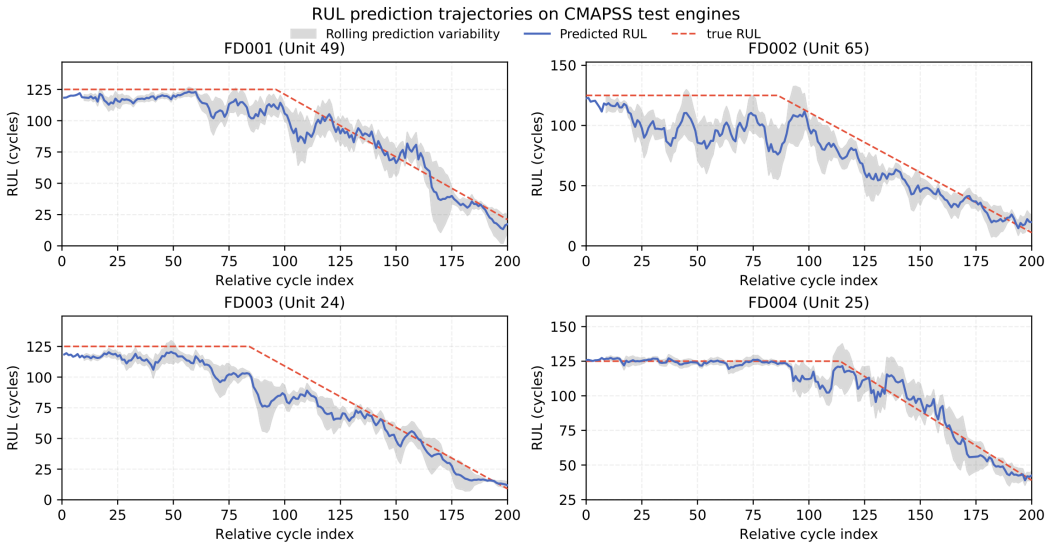


Figure 5: **CMAPSS RUL trajectories.** Representative trajectories show compact summarization behavior across the four nested-LODO subsets. Gray bands begin after the first complete 10-point window. At time $t$ they show $\hat{y}_t \pm 2\ \mathrm{std}(\hat{y}_{t-9:t})$ with population standard deviation and the lower bound clipped at zero. Each panel is one run, using seed 42 for FD001/FD003 and 240224 for FD002/FD004; the bands are temporal prediction variability, not seed uncertainty or coverage intervals.

## C Reproducibility Notes

**Artifacts.** This appendix specifies data splits, seeds, aggregation policies, candidate configurations, and nested evaluation; the repository contains scripts, manifests, and machine-readable summaries.

**Splits and aggregation.** The optional conditioning diagnostic is evaluated by nested dataset-level LODO: each fold refits the threshold and leaves without its held-out dataset. Forecasting, RUL, and AD multi-seed results use seeds $\{0, 1, 42, 101925, 240224\}$. Tables report arithmetic means and population standard deviations over seeds; paired confidence intervals use the sample standard error, and the ETT uncertainty audit pairs identical seeds. The primary bias-specific interface audit uses eight datasets; query, capacity, and hierarchy audits use the six-dataset factorial matrix. The older six-row Patch-star union is retained only as a confounded whole-configuration baseline.

Forecasting follows the standard ETT long-term forecasting setup with input length 96 and prediction horizons 96 / 192 / 336 / 720. Models optimize MSE and select checkpoints by validation RMSE, its monotone equivalent; MSE therefore defines matched responses, regret, and admission, while MAE is comparison-only. RUL uses NASA CMAPSS train/test splits with validation-selected checkpoints and reports test RMSE, MAE, and PHM score. AD trains on normal data and reports AUROC, AUPRC, oracle Best-F1, and strict F1 under a frozen validation-score percentile; point adjustment is excluded from primary comparisons. Unless otherwise noted, all task results average seeds $\{0, 1, 42, 101925, 240224\}$. External specialized models contextualize task difficulty only; the central matched comparisons hold the ONE-Core implementation family fixed and change one bias choice.

Table 4: **Official ETT forecasting under the final post-LODO refit policy.** Values are 5-seed mean±population-std test MSE under the corrected LTSF/TimePerceiver split. Every row lies on the RevIN side of the fitted scale-drift gate and uses the same Patch+RevIN+CI policy.

| Dataset | 96 | 192 | 336 | 720 |
|---|---|---|---|---|
| ETTh1 | 0.3850±0.0017 | 0.4420±0.0028 | 0.4918±0.0073 | 0.5029±0.0079 |
| ETTh2 | 0.3301±0.0046 | 0.4124±0.0080 | 0.4458±0.0040 | 0.4573±0.0057 |
| ETTm1 | 0.3369±0.0033 | 0.3826±0.0025 | 0.4080±0.0047 | 0.4666±0.0018 |
| ETTm2 | 0.1836±0.0010 | 0.2538±0.0017 | 0.3233±0.0043 | 0.4285±0.0068 |

Table 5: **Practical-equivalence margin sensitivity.** Existing paired responses are reclassified without retraining. The final FIX/CONDITION/ABSTAIN decisions are invariant. At 1.0%, one forecast hierarchy-range response changes from reference to tie ($0/3/3 \rightarrow 0/2/4$), but the fixed 32→8 policy is unchanged.

| Margin | Forecasting | RUL capacity | AD | Cell-level change |
|---|---|---|---|---|
| 0.25% | RevIN cond.; rest fixed | condition | abstain | none |
| 0.50% | RevIN cond.; rest fixed | condition | abstain | baseline |
| 1.00% | RevIN cond.; rest fixed | condition | abstain | one hierarchy reference → tie |

Table 6: **Bias-specific nested-LODO stability.** Every fold refits one gate using only its matched response matrix and descriptor block. No rule may choose between different bias axes inside a fold. The Fixed prior column uses one action across the entire corpus, including RUL dim16. Its separate fold-fitted constant gives 5.34% regret and 4/4 better folds; this is not the primary fixed-action comparator. RUL permutation rank is 8/24, not strong specificity evidence.

| Bias | Fixed prior | Cond. gate | No-worse | Better | Perm. rank | Decision |
|---|---|---|---|---|---|---|
| Forecast RevIN ($D_\sigma^{96}$) | 10.15% | **2.11%** | 11/12 | 2/12 | 4.64% | admit gate within audited corpus |
| Forecast interface | **0.24%** | 0.24% | 8/8 | 0/8 | 100.0% | fix Patch |
| Forecast abstraction form | **0.49%** | 0.49% | 6/6 | 0/6 | 100.0% | fix 32→8 |
| Forecast query $K$ | **0.29%** | 0.29% | 6/6 | 0/6 | 100.0% | fix $K$=4 |
| Forecast dim16/32 | **0.00%** | 0.00% | 6/6 | 0/6 | 100.0% | fix dim32 |
| Forecast hierarchy range | **0.03%** | 0.03% | 6/6 | 0/6 | 100.0% | fix 32→8 |
| RUL capacity | 2.66% | **0.00%** | 4/4 | 2/4 | 33.3% | admit in-family gate |

For the normalization robustness audit, ETTh1/2/ETTm1/2 and AQShunyi/AQWan are also withheld as two complete families. The scale-drift gate retains 2.92% regret versus 10.15% for fixed RevIN and is no worse on 10/12 datasets. This grouped result is secondary to ordinary dataset-level LODO but verifies that the gain is not produced only by sibling ETT or AQ rows entering the fit.

Table 7: **Matched three-way AD readout audit under MSE scoring.** R0 is compact whole-window reconstruction, R1 is token-wise reconstruction, and F is H=1 forecasting. The encoder, data protocol, channel treatment, loss, aggregation, and point-level Best-F1 evaluation are shared. Pairwise gaps are R1 minus the named comparator. All datasets use five matched seeds. The observation column reports the matched response only; it is not a deployed readout policy. The MAE reversal occurs only for oracle Best-F1; the matched MSE/MAE and frozen-threshold results are in Table 36.

| Dataset | R0 | R1 | F | R1–R0 [95% CI] | R1–F [95% CI] | Observation |
|---|---|---|---|---|---|---|
| SMD | 0.2174 | 0.1772 | 0.1816 | -0.0402 $[-0.0476, -0.0346]$ | -0.0045 $[-0.0112, +0.0012]$ | compact |
| SMAP | 0.2393 | 0.3170 | 0.2720 | +0.0777 $[+0.0716, +0.0837]$ | +0.0450 $[+0.0370, +0.0522]$ | token-wise |
| MSL | 0.2190 | 0.2121 | 0.2110 | -0.0070 $[-0.0182, +0.0045]$ | +0.0011 $[-0.0088, +0.0120]$ | unresolved |
| PSM | 0.4561 | 0.4669 | 0.4390 | +0.0108 $[-0.0107, +0.0376]$ | +0.0279 $[+0.0058, +0.0556]$ | R1 over F; R0/R1 unresolved |
| SWaT | 0.6178 | 0.5819 | 0.6734 | -0.0359 $[-0.0656, -0.0063]$ | -0.0915 $[-0.1065, -0.0769]$ | F first under MSE only |

Table 8: **Dataset roles.** Forecasting and RUL conditioning diagnostics use nested LODO. AD uses development and secondary falsification diagnostics only; neither subset validates a deployment gate.

| Role | Forecasting | RUL | Anomaly detection |
|---|---|---|---|
| Nested LODO matrix | Normalization: 12 homogeneous rows; interface: six factorial rows + AQShunyi/AQWan; other axes: ETTh1/2, ETTm1/2, Weather, Exchange | FD001–FD004 | – |
| Coverage / modality diagnostic | 44-dataset corpus | XJTU-SY, PRONOSTIA | SMD, SMAP, PSM |
| Falsification / secondary diagnostic | – | – | MSL, SWaT |

Table 9: **Published long-term forecasting comparison (lower is better).** Baselines are ordered by publication year (newest first). All rows use input length $L$=96. ETT covers prediction lengths $H \in \{96, 192, 336, 720\}$, with Avg denoting their unweighted mean; Weather uses the same four-horizon protocol for an exploratory path, while Exchange provides single-horizon coverage at $H$=96. TimePerceiver, DLinear, PatchTST, and iTransformer results on ETT and Weather reproduce TimePerceiver Appendix Table 8 (Lee & Lee, 2025). TimeMixer results are the $L$=96 three-run means from its original Table 13 (Wang et al., 2024); Exchange baselines come from PatchTST and iTransformer (Nie et al., 2023; Liu et al., 2024a), with TimeMixer unreported. ONE-Core entries are five-seed means. Input lengths, horizons, and metrics align, while training and tuning budgets are source-specific. ETT uses the final Patch+RevIN+CI fallback at all horizons; Weather reports its exploratory Patch+$K$=2+no-RevIN path at all four horizons. The Weather row labels distinguish ONE-Core's exploratory and admitted paths. The separate admitted-policy row reports the H96 action. The actual Weather policy is $K$=4+RevIN, with H96 MSE $0.1721 \pm 0.0011$; its higher horizons were not measured here. Weather's exploratory closure appears in Table 16; Exchange and the actual Weather policy use the homogeneous $K$=4 RevIN reference in Table 18. Red bold and blue underline mark the best and second-best reported values; ties at the reported three-decimal precision share the same rank.

| Dataset | $H$ | TimePerceiver (2025) | | TimeMixer (2024) | | iTransformer (2024) | | PatchTST (2023) | | DLinear (2023) | | ONE-Core (Ours) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE |
| ETTh1 | 96 | **0.372** | **0.394** | 0.375 | 0.400 | 0.386 | 0.405 | 0.414 | 0.419 | 0.386 | 0.400 | 0.385 | 0.409 |
| | 192 | **0.422** | 0.425 | 0.429 | **0.421** | 0.441 | 0.436 | 0.460 | 0.445 | 0.437 | 0.432 | 0.442 | 0.440 |
| | 336 | **0.444** | **0.434** | 0.484 | 0.458 | 0.487 | 0.458 | 0.501 | 0.466 | 0.481 | 0.459 | 0.492 | 0.466 |
| | 720 | **0.451** | **0.453** | 0.498 | 0.482 | 0.503 | 0.491 | 0.500 | 0.488 | 0.519 | 0.516 | 0.503 | 0.489 |
| | Avg | **0.422** | **0.427** | 0.447 | 0.440 | 0.454 | 0.447 | 0.469 | 0.454 | 0.456 | 0.452 | 0.455 | 0.451 |
| ETTh2 | 96 | **0.285** | **0.338** | 0.289 | 0.341 | 0.297 | 0.349 | 0.302 | 0.348 | 0.333 | 0.387 | 0.330 | 0.376 |
| | 192 | **0.365** | **0.388** | 0.372 | 0.392 | 0.380 | 0.400 | 0.388 | 0.400 | 0.477 | 0.476 | 0.412 | 0.422 |
| | 336 | **0.366** | **0.414** | 0.386 | **0.414** | 0.428 | 0.432 | 0.426 | 0.433 | 0.594 | 0.541 | 0.446 | 0.451 |
| | 720 | **0.404** | 0.440 | 0.412 | **0.434** | 0.427 | 0.445 | 0.431 | 0.446 | 0.831 | 0.657 | 0.457 | 0.467 |
| | Avg | **0.355** | **0.395** | 0.364 | **0.395** | 0.383 | 0.407 | 0.387 | 0.407 | 0.559 | 0.515 | 0.411 | 0.429 |
| ETTm1 | 96 | **0.306** | **0.347** | 0.320 | 0.357 | 0.334 | 0.368 | 0.329 | 0.367 | 0.345 | 0.372 | 0.337 | 0.375 |
| | 192 | **0.343** | **0.369** | 0.361 | 0.381 | 0.377 | 0.391 | 0.367 | 0.385 | 0.380 | 0.389 | 0.383 | 0.401 |
| | 336 | **0.371** | **0.391** | 0.390 | 0.404 | 0.426 | 0.420 | 0.399 | 0.410 | 0.413 | 0.413 | 0.408 | 0.419 |
| | 720 | **0.428** | **0.426** | 0.454 | 0.441 | 0.491 | 0.459 | 0.454 | 0.439 | 0.474 | 0.453 | 0.467 | 0.452 |
| | Avg | **0.362** | **0.383** | 0.381 | 0.395 | 0.407 | 0.410 | 0.387 | 0.400 | 0.403 | 0.407 | 0.399 | 0.412 |
| ETTm2 | 96 | **0.170** | **0.249** | 0.175 | 0.258 | 0.180 | 0.264 | 0.175 | 0.259 | 0.193 | 0.292 | 0.184 | 0.269 |
| | 192 | **0.236** | **0.293** | 0.237 | 0.299 | 0.250 | 0.309 | 0.241 | 0.302 | 0.284 | 0.362 | 0.254 | 0.316 |
| | 336 | **0.298** | **0.330** | **0.298** | 0.340 | 0.311 | 0.348 | 0.305 | 0.343 | 0.369 | 0.427 | 0.323 | 0.360 |
| | 720 | 0.399 | **0.392** | **0.391** | 0.396 | 0.412 | 0.407 | 0.402 | 0.400 | 0.554 | 0.522 | 0.429 | 0.422 |
| | Avg | 0.276 | **0.316** | **0.275** | 0.323 | 0.288 | 0.332 | 0.281 | 0.326 | 0.350 | 0.401 | 0.297 | 0.342 |
| Weather (exploratory) | 96 | **0.153** | **0.202** | 0.163 | 0.209 | 0.174 | 0.214 | 0.177 | 0.218 | 0.196 | 0.255 | 0.163 | 0.223 |
| | 192 | **0.202** | **0.247** | 0.208 | 0.250 | 0.221 | 0.254 | 0.225 | 0.259 | 0.237 | 0.296 | 0.206 | 0.267 |
| | 336 | 0.258 | 0.288 | **0.251** | **0.287** | 0.278 | 0.296 | 0.278 | 0.297 | 0.283 | 0.335 | 0.258 | 0.309 |
| | 720 | 0.338 | **0.340** | 0.339 | 0.341 | 0.358 | 0.347 | 0.354 | 0.348 | 0.345 | 0.381 | **0.329** | 0.359 |
| | Avg | **0.238** | **0.269** | 0.240 | 0.271 | 0.258 | 0.278 | 0.259 | 0.281 | 0.265 | 0.317 | 0.239 | 0.289 |
| Weather (admitted policy) | 96 | – | – | – | – | – | – | – | – | – | – | 0.172 | 0.214 |
| Exchange | 96 | – | – | – | – | 0.086 | 0.206 | 0.088 | **0.205** | – | – | **0.084** | 0.206 |

Table 10: **Published RUL comparison (lower is better).** Baselines are ordered by publication year (newest first). Columns report RMSE and PHM Score. The PHM Score is CMAPSS-specific, so it is shown as – when unreported on C-MAPSS and N/A when inapplicable to bearings. Baselines are reported results and protocols differ across papers (Babu et al., 2016; Zheng et al., 2017; Li et al., 2024; Dinten & Zorrilla, 2025). DFAGCN (Hua et al., 2024) values are reproduced from the literature comparison reported by Dinten & Zorrilla (2025); the remaining baseline columns use their cited source reports. XJTU-SY and PRONOSTIA use the full-dataset protocols audited in this paper; external methods do not report protocol-matched values. Red bold and blue underline mark the best and second-best reported values for each metric; – denotes an unreported result.

| Dataset | DFAGCN (2024) | | DSFormer (2024) | | LSTM (2017) | | CNN (2016) | | ONE-Core (Ours) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | RMSE | Score | RMSE | Score | RMSE | Score | RMSE | Score | RMSE | Score |
| FD001 | 11.21 | **175** | **10.77** | 199.82 | 16.14 | 338 | 18.45 | 1286.7 | 13.31 | 291.96 |
| FD002 | 13.05 | **683** | 14.82 | 916.72 | 24.49 | 4450 | 30.29 | 13570 | **12.87** | 684.63 |
| FD003 | 11.23 | 188 | **10.12** | **147.38** | 16.18 | 852 | 19.82 | 1596.2 | 13.92 | 504.12 |
| FD004 | **14.01** | **855** | 18.17 | 1570.20 | 28.17 | 5550 | 29.16 | 7886.4 | 14.80 | 1181.60 |
| XJTU-SY | – | – | – | – | – | – | – | – | 3.08 | N/A |
| PRONOSTIA | – | – | – | – | – | – | – | – | 18.42 | N/A |

Table 11: **Published AD comparison under raw oracle Best-F1 (higher is better).** Learned baselines are ordered by publication year (newest first); Heuristic is a non-learned reference. Baseline values are the unadjusted $F_1^*$ results tabulated by Lai et al. (2023); NPSR reports its best tabulated variant per dataset. UAE denotes the Univariate Fully-Connected Auto-Encoder of Garg et al. (2022), with one model per channel. ONE-Core reports both a single fixed residual-induction path and a dataset-wise diagnostic envelope over six predefined normality-evidence paths after the readout–score audit. No evidence-source, reconstruction-granularity, or forecast-routing policy is fitted or deployed. The adjacent Path column identifies the evidence source attaining the reported mean. 'Profile fusion' averages robustly normalized local, Q2, and context scores; 'residual induction' propagates point residual evidence through a train-calibrated nominality gate. The envelope is oracle not only in threshold selection but also in evidence-source selection, and is reported only as a diagnostic upper bound. All ONE-Core Best-F1 entries are five-seed means. These oracle-threshold scores are not deployable calibration results (see Tables 36 and 37); preprocessing and training protocols differ across sources; the table provides scale context rather than a controlled superiority claim. Macro means weight the five datasets equally: ONE-Core uses unrounded seed results, while external means use the published values shown here. Rows are datasets and models are columns; red bold marks the best and blue underline the second-best value in each dataset row.

| Dataset | NPSR (2023) | TranAD (2022) | UAE (2022) | GDN (2021) | MTAD-GAT (2020) | Heuristic (ref.) | ONE-Core (Ours) | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Fixed induction | Envelope path | Envelope |
| SMD | **0.535** | 0.310 | 0.435 | 0.529 | 0.400 | 0.494 | 0.221 | Profile fusion | 0.252 |
| SMAP | **0.511** | 0.247 | 0.390 | 0.252 | 0.296 | 0.229 | 0.247 | Q2 profile | 0.323 |
| MSL | **0.551** | 0.251 | 0.451 | 0.217 | 0.275 | 0.239 | 0.207 | Profile fusion | 0.227 |
| PSM | 0.648 | **0.649** | 0.427 | 0.552 | 0.571 | 0.509 | 0.614 | Residual induction | 0.614 |
| SWaT | **0.839** | 0.669 | 0.453 | 0.810 | 0.784 | 0.789 | 0.773 | Residual induction | 0.773 |
| Macro (5 datasets) | 0.617 | 0.425 | 0.431 | 0.472 | 0.465 | 0.452 | 0.412 | dataset-wise max | 0.438 |

Table 12: **Admission diagnostic.** Nested LODO permits forecasting normalization and RUL capacity conditioning while retaining the remaining forecasting priors. Matched controls remain within ONE-Core families. For lower-is-better metrics, aligned gain is $(m_{\text{mismatch}} - m_{\text{aligned}})/m_{\text{mismatch}}$; these candidate contrasts do not replace the admitted ETT policy in Table 4. RUL abstraction controls report RMSE and share LocalFull, dim16, $K$=2, and length 50.

| Control | Dataset / task | Aligned choice | Mismatched control | Outcome / aligned gain |
|---|---|---|---|---|
| Nested LODO | Forecast RevIN | Conditional regret 2.11% | Fixed RevIN 10.15% | admit low-drift override |
| Nested LODO | RUL capacity | Conditional regret 0.00% | Fixed dim16 2.66% | admit in-family gate |
| Axis LODO | Forecast interface | Fixed/gated regret 0.24% | 0/8 better folds | fix Patch |
| Axis LODO | Forecast capacity | Fixed/gated regret 0.00% | 0/6 better folds | fix dim32 |
| Aligned choice | FD001 RUL | Single-8: 13.31 | 32→8: 15.06 | +11.6% |
| Aligned choice | FD003 RUL | Single-8: 13.92 | 32→8: 16.33 | +14.8% |
| Matched normalization | ETTh2 forecast | RevIN 0.3301 MSE | no-RevIN 0.3672 MSE | +10.1% |
| Matched normalization | Nature15 forecast | no-RevIN 0.3465 MSE | RevIN 0.6658 MSE | +48.0% |

**Task protocols.** ETT experiments use input length 96 and prediction horizons 96 / 192 / 336 / 720. Forecasting reports test MSE and MAE. CMAPSS experiments use the standard FD001–FD004 train/test subsets, a 50-cycle input window, RUL targets capped at 125 cycles for both training and testing, and validation-selected checkpoints. At cycle $t$, training labels are $y_{e,t} = \min(125, T_e - t)$; test labels are $y_{e,t} = \min(125, R_e^{\text{end}} + T_e^{\text{obs}} - t)$, where $R_e^{\text{end}}$ is the supplied last-observation RUL. This cycle scale differs from PRONOSTIA's per-bearing 0–100 uncapped time-to-EOL labels; XJTU uses its separate 0–100 degradation/plateau convention. Test metrics are RMSE, MAE, and PHM score. AD experiments train on normal data and evaluate anomaly scores on test windows. The SWaT three-way readout audit uses matched seeds with train stride 50 and test stride 1. R0/R1/Forecast comparisons use MSE; MAE/MSE score geometry is then evaluated test-only on the resulting fixed checkpoints. Q1 always forms the primary task readout. Q2 is trained as an auxiliary bootstrap/trajectory branch with an internal projection-moment regularizer (the code alias is `sigreg_loss`), evaluated separately from direct metric gains. Inspired by LeJEPA's Sketched Isotropic Gaussian Regularization (Balestriero & LeCun, 2025), this implementation is a simplified moment penalty, not the original distribution-matching objective. It batch-standardizes $z$ using sample standard deviation plus $10^{-6}$, projects it through a trainable Gaussian-initialized matrix, and penalizes $P^{-1}\sum_{p=1}^{P}\{\mu_p^2 + (v_p - 1)^2\}$, where $v_p$ uses population variance. Defaults are $P$=64 projections and weight 0.1.

**Additional RUL modality result.** XJTU-SY is included as supporting evidence that the compact RUL summarization prior and capacity sensitivity are not limited to CMAPSS-style turbofan data. Because it uses raw bearing-vibration inputs and an XJTU-specific signal representation block, we report it outside the in-family CMAPSS LODO claim. The block is a modality-specific observation interface for high-rate vibration; automatic selection from sampling geometry remains outside the matched catalogue and is future work.

Table 13: **Additional RUL modality result outside the in-family LODO claim.**

| Dataset | Configuration | Result | Role |
|---|---|---|---|
| XJTU-SY | LocalFull + Single-8 + dim32 + signal representation block | RMSE 3.08±0.36, MAE 2.79 | Supporting bearing-vibration RUL result |

Table 14: **Frozen configurations used in the main tables.** Source files and per-seed manifests are provided in the supplementary repository.

| Result | Frozen configuration |
|---|---|
| Official ETT forecasting | Patch + h32→8 + dim32 + $K$=4 + CI + RevIN at horizons 96/192/336/720 |
| Homogeneous normalization audit | Patch + h32→8 + dim32 + $K$=4 + CI at H96; RevIN only intervention |
| Weather policy / exploratory | policy: Patch + h32→8 + dim32 + $K$=4 + CI + RevIN (H96); exploratory: $K$=2, no RevIN (all horizons) |
| Exchange H96 benchmark | Patch + h32→8 + dim32 + $K$=4 + CI + RevIN |
| FD003 RUL | LocalFull + Single-8 + dim16 |
| FD004 RUL | LocalFull + Single-8 + dim32 |
| PSM AD | three-way readout audit; calibration evaluated separately |
| MSL AD | unresolved three-way readout diagnostic |
| SWaT AD | MSE/MAE readout–score interaction; no deployed calibration policy |
| AD normality-evidence audit | fixed nominality reference: dim32, h32→8, $K$=4, full local attention, global window 8, no CI, MLP context reconstruction + point bottleneck 10; MSE + 0.5 MAE reconstruction + Q2 SSL; only the post-hoc evidence source changes |

**Full dataset coverage.** The main paper emphasizes task-level bias interventions and resulting model configurations. For transparency, the appendix reports the candidate matrices and secondary conditioning diagnostics underlying every LODO fold.

Table 15: **Forecasting coverage under the official split.** Values are 5-seed mean±population-std test MSE for prediction length 96. The four ETT rows use the scale-drift gate's Patch+RevIN action. Weather reports both the admitted $K$=4 RevIN policy and its separate exploratory $K$=2 no-RevIN closure; Exchange is a normalization-LODO row whose reported H96 benchmark is the homogeneous RevIN reference.

| Dataset | Role | Reported task-level path | MSE |
|---|---|---|---|
| ETTh1 | LODO matrix | Patch P16/S8 + h32→8 + dim32 + $K$=4 + CI + RevIN | 0.3850±0.0017 |
| ETTh2 | LODO matrix | Patch P16/S8 + h32→8 + dim32 + $K$=4 + CI + RevIN | 0.3301±0.0046 |
| ETTm1 | LODO matrix | Patch P16/S8 + h32→8 + dim32 + $K$=4 + CI + RevIN | 0.3369±0.0033 |
| ETTm2 | LODO matrix | Patch P16/S8 + h32→8 + dim32 + $K$=4 + CI + RevIN | 0.1836±0.0010 |
| Weather | admitted policy | Patch P16/S8 + h32→8 + dim32 + $K$=4 + CI + RevIN | 0.1721±0.0011 |
| Weather | exploratory closure | Patch P16/S8 + h32→8 + dim32 + $K$=2 + CI, no RevIN | 0.1634±0.0013 |
| Exchange | LODO matrix | Patch P16/S8 + h32→8 + dim32 + $K$=4 + CI + RevIN | 0.0836±0.0010 |

Table 16: **Weather H96 interface–normalization closure.** Five paired seeds hold hierarchy, dim32, $K$=2, CI, and training protocol fixed. Lower is better. Withholding RevIN improves the primary MSE but worsens MAE; the MSE-trained model and Compass are nevertheless selected only by MSE. MAE is shown for standard forecasting comparison and does not alter the decision.

| Interface | RevIN MSE | RevIN MAE | No-RevIN MSE | No-RevIN MAE |
|---|---|---|---|---|
| Patch P16/S8 | 0.1711±0.0006 | **0.2126±0.0009** | **0.1634±0.0013** | 0.2227±0.0017 |
| LocalFull | 0.1741±0.0005 | **0.2153±0.0006** | **0.1670±0.0013** | 0.2264±0.0038 |

**Forecasting candidate diagnostics.** The original pred-96 matrix shows that RevIN is useful on its five audited rows and that the exact configuration winner varies across ETT variants. The Weather 2×2 closure refutes the stronger claim that normalization is uniformly beneficial under primary MSE, while confirming Patch independently of normalization. It identifies an exploratory $K$=2 no-RevIN path; that path is not the admitted $K$=4 RevIN fallback for Weather. The later homogeneous 12-dataset matrix, not this mixed-interface screen, fits and validates the selective scale-drift gate. Most other downstream differences remain ties and no structure fold is improved by conditioning.

Table 17: **External TFB interface diagnostics at horizon 96.** Both paths use the official TFB split (Qiu et al., 2024), all channels, full stride-1 test origins, RevIN, and five paired seeds; only Global versus Patch is changed. These two rows extend the interface LODO matrix but are not presented as a standalone forecasting benchmark. Negative change favors Patch.

| Dataset | Global MSE | Patch MSE | Patch change | Evidence |
|---|---|---|---|---|
| AQShunyi | 0.74148±0.00184 | **0.73758±0.00226** | −0.53% | practical tie |
| AQWan | 0.83318±0.00344 | **0.82642±0.00514** | −0.82% | practical tie |

These datasets were the two nearest *eligible external TFB candidates* to ETTh2 under the initial mixed whole-vector ranking; they were not the nearest rows to ETTh2 after pooling the controlled ETT datasets with the external corpus. Figure 3 makes the distinction visible: most global spectral coordinates and the context top-3/IQR coordinates of AQShunyi and AQWan align more closely with ETTh1/ETTm1, while their context low-band mass is intermediate. Their Patch-direction ties therefore falsify whole-vector proximity as an interface rule rather than provide positive ETTh2-like transfer. We retain them as matched interface responses in LODO and use mechanism-linked context descriptors, not their original aggregate distance, for interpretation.

The final normalization audit supersedes that earlier $K{=}2$ directional check. It fixes Patch P16/S8, hierarchy 32→8, dim32, $K{=}4$, CI, H96, split, and five seeds across all 12 datasets, changing only RevIN. All models optimize MSE and use validation RMSE for checkpoint selection; Table 18 therefore reports the MSE response used by Compass. MAE remains in the external comparison tables but does not define a winner or threshold here.

Table 18: **Homogeneous forecasting-normalization matrix at horizon 96.** Values are five-seed mean±population-std test MSE; lower is better. Every model uses the same MSE-trained Patch P16/S8, hierarchy 32→8, dim32, $K{=}4$, and CI implementation. $D_\sigma^{96}$ is training-only relative input-window scale drift. Direction is stable in at least four of five seeds for every dataset.

| Dataset | $D_\sigma^{96}$ | RevIN MSE | No-RevIN MSE | Winner | Gain vs. wrong | Seed wins |
|---|---|---|---|---|---|---|
| AQShunyi | 0.718 | 0.7376±0.0023 | **0.6832±0.0026** | no-RevIN | 7.37% | 5/5 |
| AQWan | 0.724 | 0.8264±0.0051 | **0.7649±0.0016** | no-RevIN | 7.44% | 5/5 |
| ETTh1 | 0.309 | **0.3850±0.0017** | 0.4049±0.0068 | RevIN | 4.92% | 5/5 |
| ETTh2 | 0.431 | **0.3301±0.0046** | 0.3672±0.0146 | RevIN | 10.10% | 5/5 |
| ETTm1 | 0.535 | **0.3369±0.0033** | 0.3837±0.0047 | RevIN | 12.21% | 5/5 |
| ETTm2 | 0.397 | **0.1836±0.0010** | 0.2014±0.0041 | RevIN | 8.87% | 5/5 |
| Exchange | 0.747 | **0.0836±0.0010** | 0.1726±0.0154 | RevIN | 51.56% | 5/5 |
| Finance77 | 5.056 | **0.0466±0.0007** | 0.3657±0.0281 | RevIN | 87.26% | 5/5 |
| Human3 | 0.294 | 0.1229±0.0084 | **0.1123±0.0012** | no-RevIN | 8.65% | 4/5 |
| NYSE | 0.550 | **2.8296±0.0116** | 5.4433±0.2451 | RevIN | 48.02% | 5/5 |
| Nature15 | 0.281 | 0.6658±0.0075 | **0.3465±0.0025** | no-RevIN | 47.96% | 5/5 |
| Weather | 0.878 | 0.1721±0.0011 | **0.1653±0.0010** | no-RevIN | 3.95% | 5/5 |

The full-corpus stump splits near $D_\sigma^{96} = 0.301$, assigning no-RevIN below the threshold and RevIN otherwise. Nested LODO refits this midpoint and both leaves inside every fold. Human3 and Nature15 independently reproduce the low-drift override; ETTh1 is the only harmful ordinary-LODO override. The high drift AQ and Weather winners are not captured, so the gate is deliberately reported as selective rather than oracle-complete.

**Expanded forecasting structure matrix.** For this orthogonal structure audit, we hold Patch P16/S8, the RevIN-on reference, and channel independence fixed, then specify a one-factor star over capacity (dim16/32/64), task-token count ($K{=}2/4$), and hierarchy range (32→8 versus 64→16). Six datasets, five candidates, and five seeds produce 150 primary cells. A matched Global × $K$ diagnostic adds 120 entries, of which 60 overlap with the Patch view, for 210 unique measured h96 cells. Figure 6 reports all unique cells and, separately, the matched four-horizon Patch/Global comparison available on ETTh2 and ETTm2. Color is normalized within each dataset so cross-dataset MSE scale does not obscure the structural preference.

Table 19: **Six-dataset forecast abstraction-form audit.** Single-8 and hierarchical 32→8 hold LocalFull, dim16, $K=2$, RevIN, CI, input/horizon 96/96, and five paired seeds fixed under each dataset's standard split. Positive gain favors hierarchy. Fixed hierarchy and the descriptor LODO probe both have 0.49% mean regret, versus 1.09% for fixed Single-8. Tie denotes unresolved evidence under the paired interval and 0.5% margin, not demonstrated equivalence.

| Dataset | Single-8 MSE | 32→8 MSE | Hierarchy gain [95% CI] | Evidence |
|---|---|---|---|---|
| ETTh1 | 0.4919 | 0.4924 | −0.14% [−5.93, +5.65] | tie |
| ETTh2 | 0.3236 | 0.3270 | −1.07% [−2.49, +0.35] | tie |
| ETTm1 | 0.3688 | 0.3562 | +3.42% [+2.07, +4.76] | hierarchy |
| ETTm2 | 0.1914 | 0.1895 | +0.97% [−2.28, +4.22] | tie |
| Weather | 0.1843 | 0.1807 | +1.98% [+1.36, +2.60] | hierarchy |
| Exchange | 0.0815 | 0.0829 | −1.78% [−3.95, +0.39] | tie |

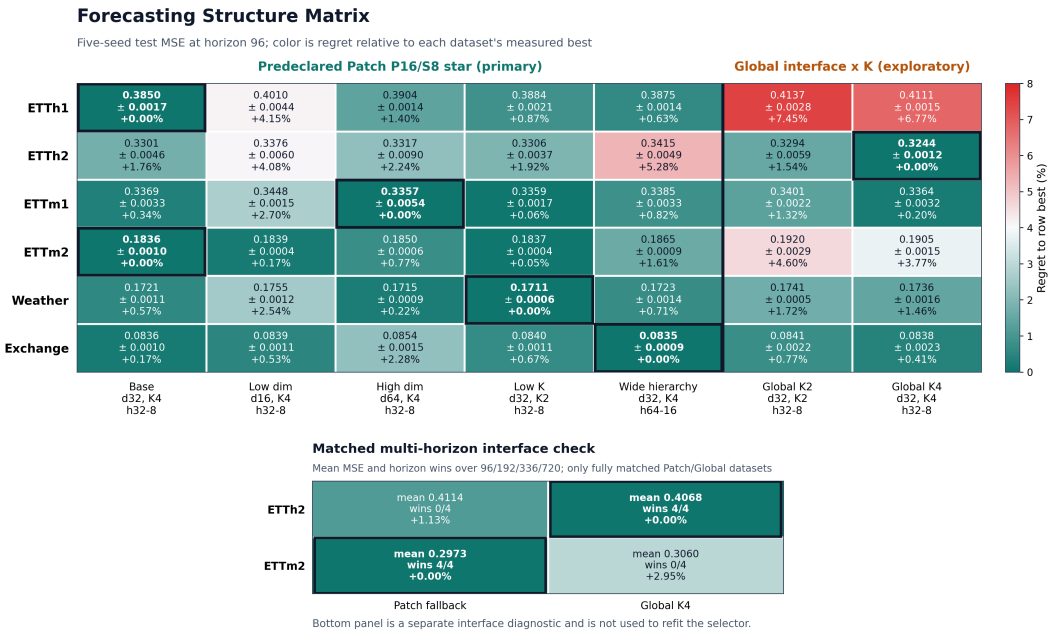


Figure 6: **Forecasting structure matrix.** Each cell reports five-seed test MSE mean, standard deviation, and regret relative to the best measured candidate for that dataset at horizon 96. The first five columns are the primary Patch star; the two Global columns are a matched interface × $K$ view. The lower panel is a separate matched-interface check: cells average MSE over horizons 96/192/336/720 and report horizon wins only for ETTh2/ETTm2, where both paths were measured throughout. Borders mark row winners. All cells retain the RevIN-on reference and do not refit the separate normalization gate; the lower panel is not used to refit the selector.

The matched horizon panel shows that interface heterogeneity is not an h96 artifact: Global wins all four ETTh2 horizons, whereas Patch wins all four ETTm2 horizons. Their mean-MSE gaps are 1.13% and 2.95%, respectively. This strengthens the evidence that interface choice is consequential. It also motivates evaluating frequency geometry inside the 96-step model context rather than relying on a global summary; that analysis is reported below. The stability gate still retains the low-regret Patch fallback.

**Accuracy–efficiency audit.** Exact-MSE selection can overstate the value of small gains. We therefore audit the cached forecasting matrix with strict Pareto dominance over test MSE and trainable parameter count. For a declared tolerance $\epsilon$, the $\epsilon$-knee is the smallest candidate whose MSE is within $\epsilon$ of the measured dataset oracle; this is an objective audit, not a refit of the primary selector. Within the primary Patch star at $\epsilon=1\%$, $K=2$ is selected on ETTh1, ETTh2, ETTm1, and Weather,

while dim16 is selected on ETTm2 and Exchange. The largest examples are ETTm1, where 71K parameters are within 0.06% of the 331K MSE winner, and ETTm2 and Exchange, where 29K-parameter candidates are within 0.17% and 0.53% of their respective winners.

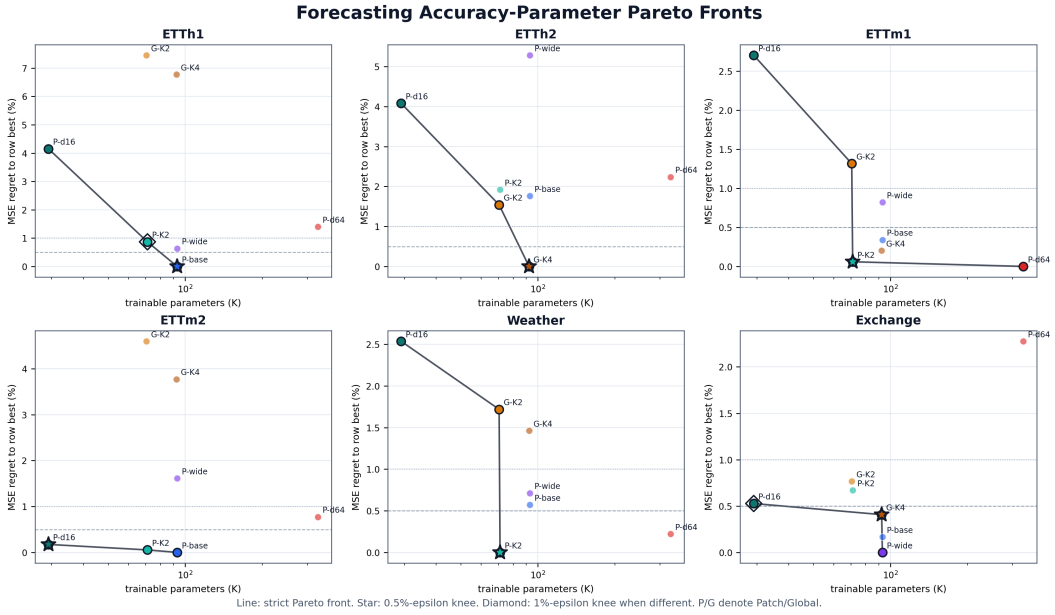


Figure 7: **Forecasting accuracy–parameter Pareto audit.** Points are cached five-seed h96 candidates; lines trace strict Pareto fronts. Stars and diamonds mark the smallest candidates within 0.5% and 1% of the measured MSE oracle. The audit exposes over-capacity exact winners but does not treat parameter count as latency.

Efficiency-aware routing remains unsupported. On the primary Patch star at $\epsilon$=0.5%, the restricted Compass point estimate reduces mean trainable size from 85.7K to 78.4K and composite loss from 0.655 to 0.618, but its bootstrap probability of beating the universal policy is only 33.2% and its exact-permutation rank is 74.4%. Allowing either leaf to select any Patch candidate does not support admission: at $\epsilon$=1% it selects a Pareto point on 6/6 folds and averages 57.1K parameters, yet increases mean quality regret from 0.30% to 0.63% and beats the universal policy in only 28.4% of bootstrap replicates. Thus Pareto-aware admission belongs in the testbed, while the unsupported efficiency coordinate still inherits its fixed task prior. Appendix D gives the exact piecewise objective.

Table 20: **Expanded forecasting conditioning diagnostic.** The primary Patch star is evaluated with nested six-dataset LODO and all 6!=720 exact descriptor-row permutations. The union is an exploratory stress test and does not alter the task-level configuration. Regret is normalized to the measured dataset-level oracle.

| Catalogue | Cells | Cond. rule | Universal | Fixed base | Random | No-worse | Top-1 | Outcome |
|---|---|---|---|---|---|---|---|---|
| Patch star (primary) | 150 | 0.50% | 0.32% | 0.18% | 0.92% | 4/6 | 2/6 | no override |
| Patch + Global union | 210 | 1.29% | 0.62% | 0.47% | 1.59% | 4/6 | 0/6 | no override |

This whole-configuration stress test is retained as a negative baseline for the bias-specific method. Its row winners are Patch base on ETTh1, ETTh2, and ETTm2; dim64 on ETTm1; $K$=2 on Weather; and the wider hierarchy on Exchange. Thus the matrix establishes configuration heterogeneity. The global-only conditional rule is strictly better than the fold-fitted universal candidate on 0/6 folds, the observed descriptor assignment is not an exact permutation optimum (89.4% of assignments attain equal or lower regret), and paired seed bootstrap gives only 37.6% probability that it beats the universal candidate. Because each leaf can differ in more than one bias, this table diagnoses

configuration heterogeneity but does not train the primary Compass gates. The axis-specific results in Table 6 replace this confounded selection objective.

**Exploratory whole-configuration descriptor baseline.** Global summaries alone can hide the geometry presented to a finite-context model. ETTh2 and ETTm2 have nearly identical global spectral entropy (0.6243/0.6241), complexity (0.0651/0.0655), and autocorrelation decay (50/50), yet Global $K{=}4$ improves over Patch by 1.73% on ETTh2 and degrades by 3.77% on ETTm2. The context-relative catalogue helps diagnose this mismatch. Within 96-step inputs, ETTm2 has top-3 spectral mass 0.677, almost identical to ETTh1 (0.674), while ETTh2 is lower at 0.583. The corresponding low/high band mass also separates ETTh2 (0.200/0.383) from ETTm2 (0.625/0.251). Thus the model-context view places ETTm2 in the Patch-compatible region rather than next to ETTh2.

Table 21: **Training-only context-spectrum descriptors.** Each row aggregates mean-centered length-96 inputs. Low contains periods $\geq$ 32 steps, high contains periods $\leq$ 16 steps, and IQR measures window/channel variation.

| Dataset | Top-3 mass | Entropy | Low mass | High mass | High IQR |
|---|---|---|---|---|---|
| ETTh1 | 0.674 | 0.567 | 0.242 | 0.342 | 0.239 |
| ETTh2 | 0.583 | 0.627 | 0.200 | 0.383 | 0.381 |
| ETTm1 | 0.808 | 0.431 | 0.783 | 0.147 | 0.150 |
| ETTm2 | 0.677 | 0.517 | 0.625 | 0.251 | 0.301 |
| Weather | 0.894 | 0.287 | 0.886 | 0.064 | 0.102 |
| Exchange | 0.816 | 0.422 | 0.796 | 0.131 | 0.109 |

Table 22: **Superseded whole-configuration routing baseline.** Every scalar coordinate uses the same six-dataset nested LODO over a candidate union that can vary both interface and $K$. It is an ablation showing why the primary method requires one matched response matrix per bias.

| Scope | Descriptor | LODO regret | No-worse | Top-1 | Perm. tied rank | $P$(beat universal) |
|---|---|---|---|---|---|---|
| Global | complexity | 1.27% | 3/6 | 0/6 | 55.0% | 6.4% |
| Global | spectral entropy | 2.28% | 3/6 | 0/6 | 94.0% | 10.5% |
| Context-96 | top-3 mass | **0.58%** | 6/6 | 2/6 | 9.4% | 37.3% |
| Context-96 | spectral entropy | **0.58%** | 6/6 | 2/6 | 9.4% | 35.3% |
| Context-96 | low-band mass | 1.56% | 5/6 | 2/6 | 65.8% | 1.7% |
| Context-96 | high-band mass | 1.56% | 5/6 | 2/6 | 65.8% | 5.4% |
| Context-96 | log low/high | 1.56% | 5/6 | 2/6 | 65.8% | 1.8% |
| Context-96 | high-band IQR | **0.58%** | 6/6 | 2/6 | 4.9% | 40.8% |

Context statistics improve this confounded baseline, but they cannot identify which of interface or query access caused a configuration-level winner. Under the primary matched analysis, Global/Patch is evaluated at fixed $K$, and $K{=}2/4$ at fixed Patch: the former yields two Patch wins and six ties, while the latter yields six ties. The correct decision is therefore to fix both priors, not to search for a more elaborate distance metric over the union.

Table 23: **Official ETT pred-96 measured-candidate matrix.** Values are 5-seed mean$\pm$population-std test MSE. Rows are measured whole configurations rather than a matched one-axis contrast; dedicated bias tables report the controlled contrasts used for admission.

| Mode | ETTh1 | ETTh2 | ETTm1 | ETTm2 |
|---|---|---|---|---|
| Patch + RevIN | **0.3850$\pm$0.0017** | 0.3301$\pm$0.0046 | 0.3369$\pm$0.0033 | **0.1836$\pm$0.0010** |
| Global + RevIN | 0.4111$\pm$0.0015 | 0.3244$\pm$0.0012 | **0.3364$\pm$0.0032** | 0.1905$\pm$0.0015 |
| Compact single-8 + RevIN | 0.4919$\pm$0.0382 | **0.3236$\pm$0.0015** | 0.3688$\pm$0.0049 | 0.1914$\pm$0.0018 |
| Global without RevIN | 0.4736$\pm$0.0293 | 0.3916$\pm$0.0374 | 0.4074$\pm$0.0106 | 0.2025$\pm$0.0027 |

Table 24: **Official ETT controlled appendix closure.** Entries are candidate-minus-baseline relative MSE; positive values are worse. CI-off and dim64 use five paired seeds, while Q2-off uses three paired seeds. "Resolved" means the 95% paired confidence interval excludes zero. Q2 is an auxiliary representation branch, so this column tests direct task-metric attribution rather than its representation utility.

| Dataset | Global CI off | Patch CI off | Q2 off | dim64, $K$=4 |
|---|---|---|---|---|
| ETTh1 | +23.13% (resolved) | +17.59% (resolved) | -0.41% (unresolved) | -3.15% (resolved) |
| ETTh2 | +9.32% (resolved) | – | +0.36% (unresolved) | -0.11% (unresolved) |
| ETTm1 | +16.23% (resolved) | – | -0.81% (unresolved) | – |
| ETTm2 | +3.34% (resolved) | +15.10% (resolved) | -0.70% (unresolved) | – |

Channel independence is therefore a fixed ETT observation prior in this testbed. Capacity sensitivity remains dataset-dependent beyond dim32: ETTh1 benefits from dim64 whereas ETTh2 is already saturated, but this evidence is insufficient to admit a capacity override. Q2's point-metric neutrality is consistent with the ONE-Core design, in which Q1 supplies the task readout and Q2 supplies trajectory-aware bootstrap, prediction, and anti-collapse regularization.

Table 25: **Exploratory one-factor Patch-star LODO fold audit.** Every fold refits the one-level descriptor rule on the other five datasets. Regret is relative to the best measured Patch-star candidate for the held-out dataset.

| Held out | Selected axis | LODO choice | Oracle | Cond. regret | Universal regret | Outcome |
|---|---|---|---|---|---|---|
| ETTh1 | capacity | high dim | base | 1.40% | 0.87% | miss |
| ETTh2 | capacity | base | base | 0.00% | 0.00% | top-1 |
| ETTm1 | tokens | base | high dim | 0.34% | 0.34% | miss |
| ETTm2 | capacity | base | base | 0.00% | 0.00% | top-1 |
| Weather | capacity | base | low $K$ | 0.57% | 0.57% | miss |
| Exchange | tokens | low $K$ | wide hierarchy | 0.67% | 0.17% | miss |

Table 26: **ETT interface diagnostic at all horizons.** Values are 5-seed mean test MSE. Global is a separately measured candidate, not the output of the withheld override; Patch is the retained fixed fallback.

| Dataset / path | 96 | 192 | 336 | 720 |
|---|---|---|---|---|
| ETTh2 global | 0.3244 | 0.4044 | 0.4453 | 0.4531 |
| ETTh2 patch fallback | 0.3301 | 0.4124 | 0.4458 | 0.4573 |
| ETTm2 global | 0.1905 | 0.2662 | 0.3331 | 0.4344 |
| ETTm2 patch fallback | 0.1836 | 0.2538 | 0.3233 | 0.4285 |

These diagnostics motivate the phrasing in the main text: official ETT is a fixed-bias admission case, not a failed selector. Context-relative spectra help interpret the ETTh2/ETTm2 mismatch, but the matched interface response contains no resolved Global win and six practical/statistical ties. The axis-specific gate therefore inherits the Patch prior, and the primary result uses Patch on all four ETT variants and horizons. On the two datasets with matched long-horizon Global runs, Patch is better throughout ETTm2 and remains within 2% on ETTh2. This supports Patch as a robust fallback rather than a per-dataset optimum; ETT's primary role is protocol-correct lightweight performance and sensitivity to observation-interface choices.

**Independent XJTU capacity audit.** The matched full-dataset audit retains the bearing split, signal representation, LocalFull encoder, Single-8 bottleneck, and 50-epoch training budget and changes only dim16 versus dim32. Its purpose is to test whether capacity response exists outside CMAPSS under the same full XJTU-SY protocol used for the support result.

dim32 wins RMSE on three of five seeds and lowers aggregate RMSE by 10.70%; the mean paired relative gain is 9.81% with a 95% interval of $[-8.32, 27.94]$%. Its two losses are near ties (1.48%

Table 27: **RUL coverage under the selected compact-bottleneck rule.** CMAPSS RMSE reports 5-seed mean±std; MAE/PHM are means. XJTU-SY reports RMSE and MAE for the selected side of the matched full-dataset capacity closure; the CMAPSS-specific PHM Score is not reported for bearing datasets.

| Dataset | Role | Configuration | RMSE | MAE / PHM |
|---|---|---|---|---|
| FD001 | LODO matrix | LocalFull + Single-8 + dim16 | 13.31±0.47 | 9.72 / 291.96 |
| FD002 | LODO matrix | LocalFull + Single-8 + dim32 | 12.87±0.24 | 9.47 / 684.63 |
| FD003 | LODO matrix | LocalFull + Single-8 + dim16 | 13.92±0.53 | 9.86 / 504.12 |
| FD004 | LODO matrix | LocalFull + Single-8 + dim32 | 14.80±0.27 | 10.68 / 1181.60 |
| XJTU-SY | external audit | LocalFull + Single-8 + dim32 + signal block | 3.08±0.36 | 2.79 / N/A |

Table 28: **Independent full-XJTU capacity closure.** Mean±std over five matched seeds. The aggregate favors dim32, while the paired interval keeps the claim at external capacity support rather than population significance.

| Capacity | RMSE | MAE |
|---|---|---|
| dim16 | 3.4516±0.3326 | 3.0815±0.2996 |
| dim32 | **3.0824±0.3560** | **2.7875±0.3486** |

and 0.43%), while two wins are substantially larger (15.76% and 32.83%). Slow RMS persistence alone would predict dim16, but the frozen full-XJTU training fast/slow balance places XJTU on the dim32 side ($B_{\text{fast}} = -1.734$, versus $-2.666/-2.663$ for FD002/FD004 and $-6.598/-6.186$ for FD001/FD003). The descriptor is computed from the training bearings and evaluated against the frozen CMAPSS boundary. This distinction shows why persistence is an abstraction-span descriptor rather than the primary capacity proxy.

**Independent PRONOSTIA capacity audit.** PRONOSTIA is a second two-channel raw-vibration bearing benchmark with three operating conditions (Nectoux et al., 2012). We use the canonical bearing-wise split (five train, three validation, and nine test bearings), fit condition-specific scaling only on the training bearings, and retain uncapped time-to-EOL labels normalized per bearing to 0–100. No FDP plateau or snapshot-level random split is introduced. The matched audit keeps the LocalFull encoder, Single-8 bottleneck, $K=2$, 30-snapshot context, and 50-epoch budget fixed and changes only dim16 versus dim32.

Per-bearing training RMS trajectories give $B_{\text{fast}} = -2.029$. Applying the frozen CMAPSS boundary ($-4.426$) therefore selects dim32. Table 29 reports the matched response.

Table 29: **Independent full-PRONOSTIA capacity closure.** RMSE uses the 0–100 uncapped RUL scale. The frozen descriptor selects dim32, which wins three seeds and has the lower mean and variance; the paired interval limits the result to directional external corroboration.

| Seed | dim16 | dim32 | dim32 gain | Winner |
|---|---|---|---|---|
| 0 | 20.2536 | 18.1949 | 10.16% | dim32 |
| 1 | 19.8836 | 17.7356 | 10.80% | dim32 |
| 42 | 17.3412 | 19.4565 | −12.20% | dim16 |
| 101925 | 18.6403 | 17.9767 | 3.56% | dim32 |
| 240224 | 17.7235 | 18.7335 | −5.70% | dim16 |
| Mean±std | 18.7684±1.2840 | **18.4194±0.6871** | 1.33% | 3/5 |

Across the same five seeds, selected dim32 has mean MAE 17.9469. The paired relative-gain interval is $[-11.17, 13.82]\%$. Thus PRONOSTIA reproduces the frozen gate's selected direction by majority vote, mean RMSE, and lower seed dispersion, while appropriately leaving the stronger claim of a modality-general effect unresolved.

Table 30: **RUL nested LODO capacity selection.** Every fold refits the training-only fast/slow spectral-balance threshold and leaves without the held-out subset. The exact tied permutation rank is 8/24 (33.3%); this is weak specificity evidence.

| Held out | $B_{\text{fast}}$ | Selected | Oracle | Regret |
|---|---|---|---|---|
| FD001 | -6.598 | dim16 | dim16 | 0.00% |
| FD002 | -2.666 | dim32 | dim32 | 0.00% |
| FD003 | -6.186 | dim16 | dim16 | 0.00% |
| FD004 | -2.663 | dim32 | dim32 | 0.00% |

Table 31: **Measured RUL multi-axis composition closure.** The reference is LocalFull+Single-8+dim16; the composed candidate is Local16+32→8+dim32. Gains and 95% intervals are paired over five seeds. "Capacity gate" retains LocalFull+Single-8 and selects dim16/dim32. A binary descriptor choice between the reference and the full three-axis candidate has 0.47% mean regret relative to that two-endpoint oracle. This is a different candidate space from the capacity gate's dim16/dim32 oracle (0.00%). Composition is not admitted because it does not improve the retained capacity endpoint.

| Dataset | Reference | Composed | Composed gain [95% CI] | Capacity gate | Decision |
|---|---|---|---|---|---|
| FD001 | 13.3073 | 17.1414 | $-28.90\%$ $[-42.59, -15.21]$ | 13.3073 | reference |
| FD002 | 13.8445 | 12.8568 | $+7.15\%$ $[+6.64, +7.66]$ | 12.8741 | capacity gate; tie |
| FD003 | 13.9261 | 17.1460 | $-23.27\%$ $[-38.10, -8.45]$ | 13.9261 | reference |
| FD004 | 15.2620 | 15.5460 | $-1.94\%$ $[-9.22, +5.33]$ | 14.7984 | capacity gate |

Local16 retains all 50 input cycles and masks each encoder attention layer to at most 17 positions (eight neighbors on each side plus self); it does not truncate the input to 16 cycles. Encoder layers can propagate information beyond one such neighborhood. The endpoint combines locality, hierarchy, and capacity, so its loss cannot be assigned to input deletion or to a particular pairwise interaction. Intermediate two-axis closures were not measured.

Table 32: **RUL Q2 auxiliary-branch ablation.** Values are five-seed RMSE means; the paired delta is No-Q2 minus Q2. Every 95% paired interval contains zero, separating Q2's trajectory-aware auxiliary objective from claims about direct RMSE gains.

| Dataset | Q2 | No Q2 | Paired delta | 95% paired CI |
|---|---|---|---|---|
| FD001 | 13.3073 | 13.0665 | -0.2408 | [-0.9885, +0.5069] |
| FD002 | 12.8741 | 12.7757 | -0.0984 | [-0.4659, +0.2691] |
| FD003 | 13.9261 | 14.1089 | +0.1828 | [-1.1927, +1.5582] |
| FD004 | 14.7984 | 15.0848 | +0.2864 | [-0.1583, +0.7311] |

Table 33: **Final dataset-wise AD bias-envelope summary.** The first-stage R0/R1/Forecast and score-geometry audits remain controlled evidence in Table 36. The second stage compares predefined normality-evidence sources with the same checkpoints, normal-training calibration, and per-entity evaluation interface. 'Gain' is the selected path minus point Best-F1. These rows form a diagnostic testbed envelope, not a fitted or deployed router; SMAP's Q2 and point means are effectively unresolved given Q2 variance.

| Dataset | Diagnostic evidence path | Best-F1 | Gain vs point | Descriptor-linked interpretation |
|---|---|---|---|---|
| SMD | Profile fusion | 0.2521 | +0.0827 | 28-entity, high-change local regime |
| SMAP | Q2 profile (point residual tied) | 0.3232 | +0.0037 | 54-entity, low-change high-entropy regime |
| MSL | Profile fusion | 0.2271 | +0.0187 | high-rank, weakly coupled context |
| PSM | Residual induction | 0.6140 | +0.0870 | high-change, strongly correlated process |
| SWaT | Residual induction | 0.7732 | +0.0038 | low-rank, strongly coupled plant dynamics |

Table 34: **AD normality-evidence source audit.** Global oracle Best-F1, mean±population standard deviation over five matched seeds. Point residual uses the point head; residual induction applies a soft train-calibrated nominality gate and bidirectional temporal propagation; local profile and Q2 profile are diagonal normal-profile distances; context profile is the context-reconstruction residual; profile fusion averages robustly normalized local, Q2, and context components. All six paths reuse the same trained nominality-reference checkpoint per dataset and seed: dim32, h32→8, $K=4$, full local attention, global window 8, no CI, an MLP context decoder and point bottleneck 10, trained with MSE $+ 0.5$ MAE reconstruction and Q2 SSL. Only the post-hoc evidence source changes. Bold marks the largest mean in each row, not a deployable selection rule.

| Dataset | Point residual | Residual induction | Local profile | Q2 profile | Context profile | Profile fusion |
|---|---|---|---|---|---|---|
| SMD | 0.1694±0.0015 | 0.2209±0.0045 | 0.2508±0.0105 | 0.1646±0.0195 | 0.1973±0.0088 | **0.2521±0.0173** |
| SMAP | 0.3195±0.0070 | 0.2472±0.0058 | 0.3071±0.0215 | **0.3232±0.0451** | 0.2526±0.0035 | 0.2878±0.0125 |
| MSL | 0.2084±0.0043 | 0.2071±0.0088 | 0.1978±0.0051 | 0.2133±0.0099 | 0.2198±0.0057 | **0.2271±0.0037** |
| PSM | 0.5270±0.0144 | **0.6140±0.0091** | 0.4518±0.0106 | 0.4577±0.0150 | 0.5781±0.0085 | 0.5197±0.0243 |
| SWaT | 0.7694±0.0026 | **0.7732±0.0019** | 0.7465±0.0070 | 0.3266±0.0650 | 0.7615±0.0052 | 0.7482±0.0055 |
| Fixed macro mean | 0.3988 | **0.4125** | 0.3908 | 0.2971 | 0.4019 | 0.4070 |
| Diagnostic envelope | | | 0.4379 | | | |

Table 35: **Training-only AD descriptor regimes and observed responses.** These direct descriptors explain which mechanism each controlled path probes; they are not a fitted dataset-to-winner classifier. Effective rank and correlation are channel-geometry summaries, while change rate, innovation, and spectral entropy summarize temporal geometry. With only five nonexchangeable regimes, the table supports mechanistic diagnosis but not nested-LODO gate admission.

| Dataset | Entities | Change rate | Innovation | Spectral entropy | Effective rank | Corr. p90 | Observed response |
|---|---|---|---|---|---|---|---|
| SMD | 28 | 0.785 | 0.176 | 0.426 | 0.118 | 0.894 | local profile/profile fusion favored |
| SMAP | 54 | 0.038 | 4.979 | 0.924 | 0.540 | 0.441 | point residual/Q2 profile unresolved |
| MSL | 27 | 0.092 | 2.257 | 0.942 | 0.972 | 0.041 | context profile/profile fusion favored |
| PSM | 1 | 0.916 | 0.130 | 0.309 | 0.220 | 0.936 | residual induction favored |
| SWaT | 1 | 0.189 | 0.083 | 0.283 | 0.080 | 0.820 | residual induction favored |

Table 36: **AD readout–score interaction closure.** Mean±std over five seeds. SMD compares score geometry on compact R0 checkpoints. SWaT reports Forecast H=1 and token-wise R1 under MSE and MAE; the readout ordering reverses in ranking quality across score geometry, while the frozen operating-point F1 does not improve with R1. Global oracle Best-F1 and macro frozen F1 use different threshold scopes and should not be interpreted as ordered bounds. This interaction table is diagnostic and does not deploy a readout or score router.

| Dataset | Readout | Score | AUROC | AUPRC | Global oracle Best-F1 | Macro frozen F1 |
|---|---|---|---|---|---|---|
| SMD | Compact R0 | MSE | 0.8186±0.0019 | 0.2956±0.0028 | 0.2174±0.0026 | 0.2687±0.0019 |
| SMD | Compact R0 | MAE | **0.8419±0.0012** | **0.3479±0.0019** | **0.2269±0.0023** | **0.2837±0.0021** |
| SWaT | Forecast H=1 | MSE | 0.8304±0.0190 | 0.5409±0.0219 | 0.6734±0.0226 | **0.3288±0.0355** |
| SWaT | Forecast H=1 | MAE | 0.8453±0.0152 | 0.6292±0.0221 | 0.7223±0.0094 | 0.3267±0.0321 |
| SWaT | Token-wise R1 | MSE | 0.8183±0.0094 | 0.4616±0.0404 | 0.5819±0.0121 | 0.2608±0.0046 |
| SWaT | Token-wise R1 | MAE | **0.8454±0.0065** | **0.7132±0.0062** | **0.7533±0.0012** | 0.2591±0.0044 |

Paired MAE-minus-MSE 95% intervals are $[0.0507, 0.0539]$ (SMD) and $[0.0719, 0.1047]$ (SWaT Forecast) for AUPRC, and $[0.0081, 0.0109]$ and $[0.0251, 0.0726]$ for Best-F1. On SWaT, the paired R1-MAE minus Forecast-MAE Best-F1 gain is $+0.0310$ [0.0204, 0.0416], with R1 winning all five seeds. Top-$k$ pooling and unconditional reconstruction/forecast residual fusion are negative controls, so neither is promoted.

**AD metric interpretation.** AUROC, AUPRC, and Best-F1 measure score ranking quality, with AUPRC emphasizing the imbalanced positive class. Strict F1 measures the fixed point-wise operating decision. Score geometry is orthogonal to these evaluation metrics: MSE emphasizes rare large residuals, whereas MAE reduces their leverage. A test-only adapter audit therefore compares MSE and MAE on fixed checkpoints without retraining. This audit follows the decoder intervention: otherwise a score change can mask whether improvement came from residual geometry or from

reconstruction fidelity. Top-$k$ channel pooling and unconditional reconstruction/forecast fusion remain negative controls.

**Calibration interpretation.** This distinction matters on PSM: oracle ranking metrics remain high, but strict F1 under the frozen point-level p95 threshold collapses because only a tiny fraction of each anomalous range is detected. SWaT supplies a second decomposition: token-wise reconstruction and MAE improve ranking, but neither determines a deployable unlabeled threshold. Thus decoder fidelity, score geometry, and calibration cannot be collapsed into one AD family label.

**PSM point-level percentile transfer (appendix diagnostic).** For the same compact R0 CI-reconstruction checkpoints, a frozen validation percentile of p95 yields near-zero strict F1 ($0.021 \pm 0.004$), whereas p20 recovers oracle Best-F1 ($0.456 \pm 0.001$). AUROC is constant across the sweep, isolating threshold scale from ranking. PSM's train-tail validation split has no anomaly labels, so this F1-optimal percentile is not selectable.

Table 37: **PSM val-calibrated percentile → test transfer (point-level scores).** Absolute thresholds are fit on unlabeled validation scores and applied to test. Values are 5-seed mean±std. Bold marks the best grid operating point by test F1 (oracle among percentiles, not val-selectable).

| Val percentile | Strict F1 | Precision | Recall |
|---|---|---|---|
| p10 | 0.452±0.001 | 0.314±0.002 | 0.803±0.004 |
| p20 | **0.456±0.001** | 0.353±0.002 | 0.643±0.006 |
| p40 | 0.435±0.006 | 0.502±0.017 | 0.384±0.003 |
| p60 | 0.125±0.015 | 0.698±0.005 | 0.069±0.009 |
| p80 | 0.043±0.006 | 0.759±0.008 | 0.022±0.003 |
| p95 | 0.021±0.004 | 0.754±0.004 | 0.010±0.002 |
| p99 | 0.012±0.002 | 0.741±0.003 | 0.006±0.001 |

Table 38: **Diagnostic controls supporting the procedure.** Stability controls test whether dataset conditioning is useful over the cached candidate space; wrong-choice controls report the gain from selecting the aligned bias over a deliberately mismatched choice within the same small implementation family.

| Control | Evidence | Decision |
|---|---|---|
| Forecast RevIN LODO | Twelve nested-LODO folds: conditional regret 2.11% versus fixed RevIN 10.15%; 11/12 no-worse | admit low-drift override |
| Forecast RevIN family LODO | ETT and AQ withheld as families: conditional regret 2.92%; 10/12 no-worse | robustness passed |
| RUL LODO | Four C-MAPSS subsets: conditional regret 0.00% versus fixed dim16 2.66% (fold-fitted constant: 5.34%) | admit in-family gate |
| Forecast interface LODO | Eight datasets: gated and fixed-prior regret both 0.24%; 0/8 better folds | fix Patch |
| Forecast capacity LODO | Six datasets: gated and fixed-prior regret both 0.00%; 0/6 better folds | fix dim32 |
| AD factorized audit | R0/R1/F rankings are heterogeneous; six evidence paths yield fixed macro 0.297–0.412 and diagnostic envelope 0.438 | reject universal readout/profile fusion; retain diagnostic envelope and abstain |
| RUL aligned choice | FD001 / FD003: +11.6% / +14.8% RMSE gain | compact prior matters |

## D Additional Robustness and Sensitivity Checks

**Normalization fold distribution.** Across the twelve nested-LODO folds, the gate improves two, ties nine, and worsens one relative to fixed RevIN (comparison tolerance $10^{-12}$). Both methods have median regret zero, as does the median paired regret reduction. Nature15 alone contributes 7.681

percentage points to the 10.147% fixed-RevIN mean regret, and 95.54% of the total gate reduction. Across the original twelve folds, the regret IQR is 7.98 percentage points for fixed RevIN and 4.38 for the gate; maximum regret is 92.17% (Nature15) and 8.04% (AQWan), respectively. Quartiles use linear interpolation of dataset-level fold regrets. These summaries characterize selective improvements within the audited matrix rather than uniform gains across datasets.

**Post-hoc corpus-composition diagnostic.** Prompted by feedback on gain concentration, this additional post-hoc check reuses the released five-seed means without retraining predictors. Fitting all twelve datasets or the eleven remaining after removing Nature15 gives the same threshold, $D_\sigma^{96} = 0.301218$, with no-RevIN below it and RevIN otherwise. Table 39 separately repeats LODO after this deletion, using ten fitting datasets per fold instead of eleven. In the Human3 fold, both low-dispersion no-RevIN examples, Nature15 and Human3, are then absent from fitting; the selected rule becomes constant RevIN. Human3 is the only held-out action that changes, while the existing harmful ETTh1 override remains. This comparison diagnoses altered development coverage and does not replace the original LODO evaluation.

Table 39: **Additional post-hoc corpus-composition check.** Dataset-balanced mean regret (%; lower is better). The fitting algorithm, descriptor, and held-out regret definition are unchanged.

| Evaluation | Fit datasets/fold | Fixed RevIN | Gate |
|---|---|---|---|
| Original 12-dataset LODO | 11 | 10.1467 | 2.1072 |
| LODO after removing Nature15 | 10 | 2.6898 | 3.1601 |

**Exact accuracy–efficiency objective.** For candidate pool $\mathcal{A}$, write $p(a)$ for trainable parameter count and $q_D(a) = \max\{0, \mathrm{MSE}_D(a)/\min_{b\in\mathcal{A}} \mathrm{MSE}_D(b) - 1\}$. Let $F_D = \{a : q_D(a) \le \varepsilon + 10^{-12}\}$, $p_{F,D} = \min_{a\in F_D} p(a)$, $p_{\min} = \min_{a\in\mathcal{A}} p(a)$, and $p_{\max} = \max_{a\in\mathcal{A}} p(a)$. The released evaluator uses

$$\ell_D(a) = \begin{cases} \dfrac{\log[p(a)/p_{F,D}]}{\log[p_{\max}/p_{\min}]}, & a \in F_D, \\ 1 + \dfrac{q_D(a) - \varepsilon}{\max\{\max_{b\in\mathcal{A}}(q_D(b) - \varepsilon), 10^{-12}\}}, & a \notin F_D. \end{cases}$$

The reported composite is the unweighted dataset mean of $\ell_D$, not a weighted sum of MSE and parameters. All audited pools have $p_{\max} > p_{\min}$. The 0.655-to-0.618 comparison uses $\varepsilon = 0.005$ and the five-candidate Patch star (base, dim16, dim64, $K$=2, and wide hierarchy); $\varepsilon = 0.01$ denotes the separate 1% audit.

**Published comparisons and future extensions.** Tables 9, 10, and 11 compare ONE-Core with published baseline results. Forecasting includes TimePerceiver (2025) and TimeMixer (2024) at $L$=96. The tables identify source-reported results and ONE-Core runs; external methods have not been jointly retrained with controlled splits, preprocessing, tuning budgets, checkpoint selection, and score calibration. A direct rerun would need the same forecasting borders, input/horizon/channel settings and MSE objective; RUL window length, target cap and bearing splits; and AD point adjustment, aggregation and frozen-threshold protocol, with the same seed set and a declared validation-only tuning budget. Such joint reruns and intermediate two-axis RUL closures remain future extensions. The completed Weather/Exchange abstraction audit extends the H96 evidence; broader corpora and horizons remain opportunities to test further updates.