# Raw-Routed Mixture of Adapters: A Causal Intervention for Routing Collapse in Time Series Foundation Models

**Hung Phan**
RMIT University Vietnam

**Thuy T. Nguyen**
RMIT University Vietnam

**Minh Ngoc Dinh***
RMIT University Vietnam

**Nhat-Quang Tran**
RMIT University Vietnam

## Abstract

Time series foundation models (TSFMs) commonly adapt to new data by attaching a single trainable head to a frozen backbone, a one-size-fits-all setup that underfits heterogeneous regimes. Replacing the head with a mixture of experts is the standard upgrade. Jointly training MOMENT's backbone and adapter with encoder-side RevIN drives routing entropy toward zero as one expert dominates. We call this failure *normalization-induced routing collapse*. Standard MoE rescues do not repair it, because pre-encoder normalization strips the statistics a router needs to tell regimes apart. A mutual-information decomposition makes this precise and yields a signal-ratio that, computed before training, predicts dataset vulnerability (Spearman $\rho{=}-0.88$, $n{=}9$ datasets). Nine causal controls, including a vision-modality contrast, isolate instance normalization as the cause. The prescription is a minimal causal intervention: *Raw-Routed Mixture of Adapters* (RR-MoA), which routes on the raw, pre-normalization input. RR-MoA wins $54/54$ comparisons against the strongest fixed adapter across three freeze levels. Strictly frozen RR-MoA significantly outperforms LoRA, TRACE and AdaMix, and beats full fine-tuning by 12–79% (the *Frozen Paradox*). The gains extend across the MOMENT and Moirai families and to imputation. All these baselines adapt pretrained TSFMs. A from-scratch linear model still has lower MSE than RR-MoA. Two variants confirm the principle. Residual-IA$^+$ matches or beats from-scratch NLinear in $96/123$ cells. Code is provided in the supplementary material.

## 1 Introduction

Time series foundation models (TSFMs) [2–7, 9–12, 47, 64, 70–75] are pretrained backbones for general-purpose time-series analysis that transfer to unseen datasets with little target-specific training. Pretrained on heterogeneous time series spanning domains like energy, weather, traffic, and finance, they aim to learn regularities that generalize across tasks. The dominant deployment recipe holds the backbone fixed and trains only a small *adapter head*, a per-task module mapping hidden states to target-specific predictions. Because the backbone is frozen, this head carries all dataset-specific information, making its design an important factor in transfer quality. In a major class of TSFMs (MOMENT [2], TTM [12], and others), this head is uniform: a single pooling-and-projection rule applied to every input window, regardless of regime.

Probing analyses of TSFMs [56, 57] indicate this is suboptimal: distinct temporal concepts (trends, dispersion, change-time) localize in different parts of the hidden state. A calm baseline and a volatile

*Now at Millennia Education.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

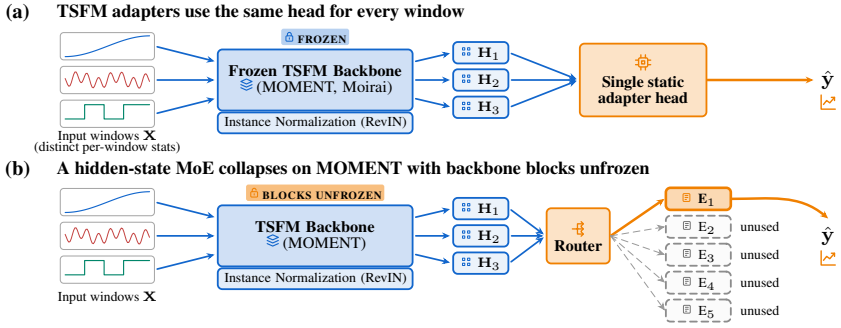


Figure 1: **Problem overview. (a)** TSFM adapters use the same head for every window. **(b)** On MOMENT with RevIN, the hidden-state mixture-of-experts head collapses when backbone blocks are unfrozen: entropy reaches $0.000 \pm 0.000$ and the remaining experts receive no gradient (Table 2).

spike differ in these concepts, so they route their predictive signal through different subspaces, and no fixed pooling rule reads both. The remedy is to replace the single rule with a dispatcher over several specialized readers [18], selecting for each window the expert whose pooling best matches its regime: an AdaMix-style mixture-of-experts (MoE) head.

On MOMENT with encoder-side RevIN, this head degenerates when backbone blocks are unfrozen and trained jointly with the adapter (Figure 1b): gating mass contracts to one-hot in tens of steps (Table 2), and the favored expert's gradients reshape the backbone to serve it alone (Figure G.1). We name the phenomenon *normalization-induced routing collapse* and trace it upstream of the optimizer: RevIN [21] subtracts each window's mean and variance before encoding, leaving the statistics that distinguish experts structurally absent from the router's input (Observation 1). Eleven optimization-side rescues from five families (load balancing, z-loss, entropy regularization, ReMoE, expert choice; 720 runs) close only $16\%$ of the gap that an input-side fix closes ($2.7\times$ shortfall, Table G.3); Moirai, whose I/O scaling is not learnable-affine, does not collapse (§4.2), narrowing the cause to learnable-affine encoder normalization.

We apply a minimal causal intervention: route on $\mathbf{X}_{\text{raw}}$, the input *before* normalization, with experts, optimizer, and a strictly frozen backbone unchanged. We call the head *Raw-Routed Mixture of Adapters* (RR-MoA). The same diagnosis predicts a *Frozen Paradox*: backbone updates only serve the live expert, so unfreezing should hurt (Proposition 1); empirically frozen wins $4/6$ datasets and unfreezing never recovers more than $13\%$.

Our main contributions are:

- We formulate routing collapse on instance-normalized TSFMs as a router-*input* failure mode, distinct from and upstream of the optimization-side failure modes prior MoE work targets; the rescues above leave $84\%$ of the gap on the table that an input-side fix closes.
- We derive a mutual-information decomposition (Observation 1) that lower-bounds the routing information removed before the router sees the input, and from it we obtain $R(\mathcal{D})$, a training-free signal-ratio computable from raw dataset statistics. $R(\mathcal{D})$ predicts raw-routing benefit at Spearman $\rho = -0.88$ ($p < 0.002$; $n = 9$ datasets) and correctly abstains on Traffic ($R = 0.14$).
- We propose **RR-MoA**, which changes only the router's input and wins $54/54$ cells on the primary benchmark (26–$79\%$ MSE), beating LoRA, TRACE, AdaMix, and full fine-tuning at $p < 0.001$, while resisting naive $(\mu, \sigma)$ re-injection ($+42$–$239\%$ MSE; $11/18$ cells re-collapse). **SR-MoA** replaces the external router with per-expert sigmoids (23–$38\%$ lower MSE than RR-MoA on six datasets). RR-MoA is a diagnostic probe that a from-scratch linear model still beats in MSE; the deployment adapter **Residual-IA$^+$** carries raw input into each expert's body and matches or beats a from-scratch NLinear in $96/123$ (backbone, dataset, horizon) cells (§4.2).

## 2 Related Work

RR-MoA sits at the intersection of TSFM adaptation [40] and MoE design [85]. *TSFM adaptation:* compact specialists rival heavyweight backbones on LTSF [62, 63, 69, 76, 77]; LLM-reprogrammed forecasters inherit upstream normalization [70–72]; the adapter literature varies fine-tuning strategy

or backbone selection [16, 24, 26, 28–31, 59], and AdaPTS [27] treats adapter topology as a per-task hyperparameter; to our knowledge no prior TSFM work uses a per-window mixture over topologically distinct adapters. *MoE design space:* scaling axes [79–84, 86] and in-backbone TS-MoE [11, 58] act on token routing within the encoder; SR-MoA's per-expert sigmoid generalizes MMoE's multi-gate pattern [87] from multi-task to single-task specialization, and Wang et al. [49] argue MoE specialization reflects hidden-state geometry, which we extend upstream by characterizing what shapes that geometry on TSFMs. The instance-normalization origin [92] treats stripping as benign preprocessing; we identify a downstream MoE-adapter consequence not previously characterized.

**MoE routing collapse and rescue.** Existing diagnoses cluster into three threads. *Optimization-side*: load-balancing [19], z-loss [32], and AdaMix's hidden-state router [18] target training dynamics; Guo et al. [53] argue load-balance loss itself causes expert overlap and propose orthogonality regularizers. *Representation-geometry*: Chi et al. [51] attribute collapse to token clustering and propose hyperspherical scores; Wu et al. [52] split tokens into sub-tokens to raise expert activation. *Architectural / paradigm*: Expert-Choice [23] inverts token–expert assignment, Soft MoE [60] replaces hard top-$k$ with a soft mixture, ReMoE [22] replaces top-$k$ with ReLU gating, Panda et al. [68] restore a dense gradient under sparse forward passes, and Hua et al. [54] decouple the router from the task objective. Our setting differs from all three: the routing signal is structurally absent from the router's input *by design of an upstream architectural normalizer (RevIN)*, before training dynamics, geometry, or routing-paradigm choices can act. Sweeping five rescue families across twelve configurations on MOMENT+RevIN, none recover MSE to within $2.7\times$ of RR-MoA (Table G.3). Concurrent work eliminates the external router (AoE [42], Routing-Free MoE [44]) or replaces softmax+load-balance with eigenbasis routing (ERMoE [43]); in our setting (§4.2), self-gated experts on raw input match RR-MoA, while the same architecture on hidden states collapses.

**Normalization in Time Series.** RevIN [21] adapts instance normalization to forecasting distribution shift; DishTS [33], SAN [93], and Non-stationary Transformers [35] partially recover stripped statistics, FAN [34] extends beyond $(\mu, \sigma)$ to Fourier components, and DDN [65] normalizes jointly in time and frequency; Berthelier et al. [48] and Zou et al. [50] concurrently flag RevIN redundancy and the limits of variance-centric normalization. We identify a specific MoE consequence not previously characterized: when stripped statistics carry the *routing* signal, instance normalization collapses the router under unfreezing, distinct from the optimization-induced collapse standard rescues target.

## 3 Raw-Routed Mixture of Adapters (RR-MoA)

RR-MoA is a per-window mixture in which the router reads the raw, pre-normalization input while the experts consume the frozen TSFM's hidden states. We derive it from a routing-information loss (Observation 1, below) that locates the bottleneck in the router's *input*, not its optimizer; task setup, mechanism, and architecture follow in that order.

**Task setup.** Given a window $\mathbf{x} \in \mathbb{R}^{T\times c}$ ($T$=512, $c$ channels), forecasting predicts $\mathbf{y} \in \mathbb{R}^{h\times c}$ at horizons $h \in \{96, 192, 336, 720\}$ under MSE loss (imputation uses the same formalism with $\mathbf{y}$ as the reconstruction target). A pretrained TSFM acts as a feature extractor $f_\theta : \mathbb{R}^{T\times c} \to \mathbb{R}^{P\times d}$ ($P$ patches of dim $d$) with $\theta$ *frozen* ($\nabla_\theta \equiv 0$); we train only an adapter $g_\phi : \mathbb{R}^{P\times d} \to \mathbb{R}^{h\times c}$ with a small parameter budget ($|\phi| \le 500$K), so that many task-specific adapters can be hot-swapped onto a single shared backbone (App. B). A *per-window mixture of adapters* replaces $g_\phi$ with $K$ experts $\{\mathrm{Expert}_j\}_{j=1}^K$ plus a router $G_\psi$ that returns a sparse weight $\widetilde{w} \in \Delta^{K-1}$ supported on a Top-$k$ subset ($k \le K$), giving $\hat{\mathbf{y}} = \sum_j \widetilde{w}_j \,\mathrm{Expert}_j(f_\theta(\mathbf{x}))$ trained by $\min_{\phi,\psi} \mathbb{E}\,\|\hat{\mathbf{y}} - \mathbf{y}\|_2^2$.

**Problem statement.** The standard MoE recipe [19, 78] lets the router $G_\psi$ read the backbone's hidden states $f_\theta(\mathbf{x})$, and this is the choice taken by the closest TSFM-adapter precedent, AdaMix [18]. We show that on TSFMs equipped with instance normalization (RevIN [21], the de facto standard for distribution-shift handling in MOMENT [2], PatchTST [13], and TimesFM [3]) this choice has a structural failure mode not previously characterized: the router's input is missing, by construction of an upstream architectural normalizer, the per-window first- and second-order statistics that distinguish regimes (formalized in Observation 1), so $G_\psi$ collapses to a single expert and the remaining $K-1$ experts receive no gradient. We name this failure *normalization-induced routing collapse* and state it as the central problem of this work:

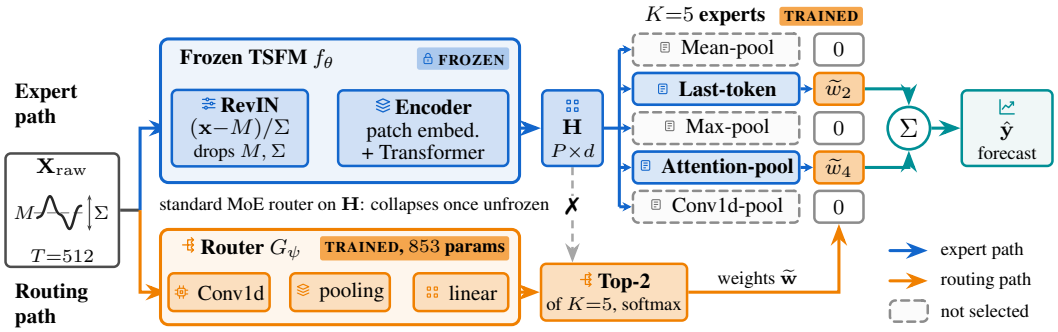


Figure 2: **RR-MoA architecture.** *Expert path* (blue): inside the frozen TSFM, RevIN removes each window's mean $M$ and scale $\Sigma$, and the encoder feeds hidden states $\mathbf{H}$ to all $K{=}5$ experts. *Routing path* (orange): the Conv1d router (853 params) reads $\mathbf{X}_{\text{raw}}$ with $M, \Sigma$ intact and selects the Top-2 experts, whose weighted outputs sum to the forecast $\hat{\mathbf{y}}$; unselected experts (gray) receive no gradient. Standard hidden-state MoE collapses once backbone blocks are unfrozen (Table 2); RR-MoA wins **54/54** (Table 1).

> *Given a frozen instance-normalized TSFM, find the routing input $u(\mathbf{x})$ such that the per-window mixture $\sum_j G_\psi(u(\mathbf{x}))_j \,\text{Expert}_j(f_\theta(\mathbf{x}))$ trains without entropy collapse and improves over a single static head.*

This collapse is upstream of the optimization-induced collapse that existing rescues target – load balancing [19, 78], z-loss [32], ReLU routing [22], and expert-choice [23] – and is not addressed by prior MoE-adapter [18, 61], MoE routing-collapse [51–54, 60, 68], or time-series-normalization [33–35, 48, 50, 65, 93] work. We next formalize the loss (Observation 1) and define a training-free predictor $R(\mathcal{D})$ for it; the rest of the section then specifies RR-MoA, the answer to the problem with $u(\mathbf{x}) = \mathbf{x}$, and §4.2 verifies the answer with nine causal controls and a 720-run rescue sweep (Table G.3).

**Mechanism.** We formalize the failure as a routing-information loss across the normalization step. Each window decomposes into mean $M_i$, scale $\Sigma_i$, and shape $\mathbf{S}_i = (\mathbf{x}_i - M_i)/\Sigma_i$; an instance normalizer keeps $\mathbf{S}_i$ and discards $(M_i, \Sigma_i)$. The chain rule then quantifies the routing signal in the discarded part.

**Observation 1** (Routing Information Loss Under Instance Normalization)**.** *Let $X$ be a random input window, $E \in \{1, \ldots, K\}$ the expert assignment produced by any deterministic router, $S = (X - M)/\Sigma$ the RevIN-normalized shape.* (i) Exact decomposition: $I(X;E) - I(S;E) = I(M, \Sigma; E \mid S) \geq 0$. (ii) Quantitative bound (no independence assumption): $I(X;E) - I(S;E) \geq I(M, \Sigma; E) - I(M, \Sigma; S)$, *with $\varepsilon := I(M, \Sigma; S) \geq 0$ capturing residual leakage of stripped statistics through the shape; equality iff $(M, \Sigma) \perp\!\!\!\perp S \mid E$.*

When $E = f(M, \Sigma)$ and the stripped statistics are independent of the shape ($(M, \Sigma) \perp\!\!\!\perp S$; App. D), $I(X;E) - I(S;E) = H(E)$: instance normalization destroys the entire routing distribution. Normalization removes the location-scale part $(M, \Sigma)$ and retains the shape $S$. Observation 1 lower-bounds routing information loss in the removed part. The signal ratio $R(\mathcal{D}) = \big(\sum_c \text{Var}(M_c) + \sum_c \text{Var}(\Sigma_c)\big)/\overline{\text{Var}(S)}$ divides the channel-summed variance of the removed statistics by shape variance averaged over windows and channels. This training-free monotone proxy is high when removed statistics carry routing information and low otherwise. Neither Observation 1 nor $R(\mathcal{D})$ requires $(M, \Sigma)$ to carry the whole routing signal. Empirical tightness of the bound, the predictor's correlation with raw-routing benefit, and the boundary cases that falsify a generic-SNR alternative appear in §4.2.

**Collapse criterion.** We predict collapse only when normalization removes statistics that carry routing signal, with healthy routing otherwise (evidence in §4.2 and App. H). The criterion gives a necessary condition. A pre-registered held-out test replaces RevIN with RMSNorm, which retains only each window's level-to-scale ratio. We predicted collapse, but routing stayed healthy in 7 of 8 valid cells (App. H.3).

**Architecture.** Per-window routing should dispatch a quiet baseline and a volatile spike to different expert topologies, but doing so requires the router to see the per-window statistics that distinguish regimes – exactly what RevIN strips. RR-MoA splits the two needs into two streams: the frozen TSFM produces embeddings $\mathbf{H}{=}f_\theta(\mathbf{x})$ for the experts, while a lightweight gate reads the *raw, pre-normalization* input $\mathbf{X}_{\text{raw}}$ to select which experts to activate:

$$\mathbf{Y} = \sum_{j=1}^{K} \widetilde{w}_j \cdot \text{Expert}_j(\mathbf{H}_{\text{FM}}), \qquad \mathcal{T} = \text{TopK}\big(G_\psi(\mathbf{X}_{\text{raw}}),\, k\big),$$
$$\widetilde{w}_j = \begin{cases} \dfrac{\exp\big(G_\psi(\mathbf{X}_{\text{raw}})_j\big)}{\sum_{j' \in \mathcal{T}} \exp\big(G_\psi(\mathbf{X}_{\text{raw}})_{j'}\big)}, & j \in \mathcal{T}, \\ 0, & j \notin \mathcal{T}. \end{cases} \tag{1}$$

where $G_\psi : \mathbb{R}^{T\times c} \to \mathbb{R}^K$ is a small Conv1d + pooling + linear gate (853 params) operating directly on the raw time series, $\mathcal{T} \subset \{1, \ldots, K\}$ is the set of Top-$k$ indices, and each $\text{Expert}_j$ is a canonical adapter head. By construction $\sum_j \widetilde{w}_j = 1$ with support only on $\mathcal{T}$, so only $k$ of $K$ experts execute per sample. With $K{=}5$, $k{=}2$ we obtain $40\%$ of the dense expert FLOPs while retaining $63\%$ of the dense MSE improvement over the best fixed adapter (App. I). Routing on raw input preserves the temporal statistics (trend, amplitude, volatility) that RevIN strips away.

**Bounded overhead, diverse experts, multi-tenant hot-swap.** The router is a Conv1d gate of $853$ parameters ($0.4\%$ of one expert head, $<10^{-3}$ of the frozen backbone); with $k{=}2$ of $K{=}5$ active per window, RR-MoA adds two adapter forward passes plus the gate to one frozen TSFM forward. Per batch of $128$ windows, the gate reads $0.26$ MB of raw input and a hidden-state router pools $16.8$ MB of hidden states (float32). RR-MoA, SR-MoA and Residual-IA$^+$ add at most $6$ MB of peak GPU memory and $5.9\%$ latency to the frozen backbone (Table B.1). The expert pool itself comprises $K{=}5$ topologically distinct adapters (mean-pool, last-token, max-pool, attention-pool, Conv1d-pool), spanning pooling families with distinct expressivity regimes for transformer-based sequence models [55]; architectural diversity matters, since five identical experts (same topology, different initialization) degrade MSE by up to $29\%$ on ETTh1/ETTm1 (Appendix I.1; pseudocode in Algorithm 1, Appendix A). Routing stays non-collapsed as the pool grows to $K{=}20$ experts (App. I.2). Because the backbone is strictly frozen, a single TSFM held in GPU memory can be shared across many tenants while per-task adapter pools (354–466K params each) hot-swap from host RAM, an arrangement DLinear and per-dataset-from-scratch baselines cannot match (Appendix B.1).

We now validate that this causal intervention recovers routing diversity, beats the adapter and fine-tuning baselines, and holds under nine causal controls.

# 4 Experiments

## 4.1 Setup

**Datasets and tasks.** We evaluate on six LTSF benchmarks (ETTh1/h2/m1/m2 [36], Weather, Electricity [37]) spanning hourly to 10-minute sampling and 7–321 channels, with chronological splits and channel-wise normalization [13]. Tasks: forecasting at $H{=}96$ primary (multi-horizon in Appendix C) and $20\%$-masked imputation.

**Backbones.** We use six backbones spanning three normalization regimes: (i) MOMENT-small/large [2], with learnable-affine RevIN [21] upstream of the encoder; (ii) Moirai [7] and Moirai-MoE, with internal RMSNorm and non-learnable per-instance I/O scaling (no learnable-affine RevIN upstream of the encoder); (iii) Chronos [4] (T5-style, no instance normalization) and Timer-XL [6] (LayerNorm only, no instance normalization) as no-instance-normalization negative controls (per-backbone normalization specifications in Appendix A; full cross-backbone results in App. N). The collapse criterion (§3) predicts that only regime (i) is vulnerable: RevIN's learnable-affine normalization upstream of the encoder strips $M$ and weakens $\Sigma$ in the hidden states the router reads, whereas non-learnable I/O scaling preserves both. On ETTh1, ridge regression probes window mean and log scale in the router's input. Both are linearly decodable on Moirai (held-out $R^2{=}0.996$ and $0.961$) and on MOMENT without RevIN ($0.999$ and $0.951$). With RevIN, mean decodability falls to $-0.311$ and log-scale decodability to $0.600$ (C9 in Table 3). MOMENT-small is primary. Primary results use a strictly frozen backbone; freeze-level ablations follow.

**Baselines.** We organize baselines into four tiers, kept separate to avoid confounding frozen-vs-unfrozen and adapter-vs-from-scratch comparisons: *frozen-adapter* (three fixed adapters; 5-expert independent ensemble; AdaMix [18]); *frozen PEFT* (LoRA [15], 108-run sweep, App. E; TRACE [25]); *unfrozen* (full fine-tuning, all blocks, best of 5 heads $\times 2$ LRs; extended schedules in App. O); *from-scratch calibration* (DLinear [17], single-Linear variant).

**Training and statistics.** Adam ($\eta$=$10^{-3}$), MSE loss, 15 epochs, batch 128; 5 seeds for the core grid, 3 for ablations. Significance via Wilcoxon signed-rank, Bonferroni-corrected over 6 datasets $\times$ 5 seeds ($^{*}/^{**}/^{***}$ for $p<0.05/0.01/0.001$); pooled, RR-MoA beats the best fixed adapter at $p$=$2.9\times10^{-11}$ uncorrected and $<10^{-9}$ under Bonferroni, Holm or Benjamini–Hochberg (App. A.1). Full hyperparameters in App. A; code is in the supplementary material ($\sim$235 A10G GPU-hours total).

**Research questions.** The experiments answer three questions implied by the mechanism (§3). The first concerns the *phenomenon (RQ1)*: does the standard hidden-state MoE collapse on instance-normalized TSFMs, and does a raw-routed mixture beat fixed adapters and full fine-tuning? The second concerns the *mechanism (RQ2)*: what is the underlying cause, and do its falsifiable predictions hold under controls and rescue attempts? The third concerns *generality (RQ3)*: does the diagnosis transfer to the six adaptation baselines (fixed adapters, LoRA, TRACE, independent ensembles, AdaMix, full fine-tuning), other backbones, horizons, and tasks?

### 4.2 Main Results

We report findings for RQ1 (phenomenon), RQ2 (mechanism), and RQ3 (generality) in turn.

**Standard MoE collapses; RR-MoA wins 54/54 vs fixed adapters**

**RR-MoA vs. fixed adapters.** RR-MoA wins all **54/54** configurations (6 datasets $\times$ 3 freeze levels $\times$ 3 seeds) against the best of three single-head adapters, with tight standard deviations (Table 1).

Table 1: **Freeze-level ablation** (test MSE, mean$\pm$std, 3 seeds $\{42, 43, 44\}$, H=96). RR-MoA (sparse Top-2 routing over 5 experts) vs. the best single-head adapter from {linear, attention, conv} trained with the same protocol. "MSE $\Delta\%$ vs. single-head" is the relative MSE change of RR-MoA over the single-head baseline; negative = lower MSE = better. RR-MoA wins **27/27** paired (cell, seed) comparisons in this table; combined with Appendix G the score is **54/54** across 6 datasets $\times$ 3 freeze levels $\times$ 3 seeds. The gain is largest when frozen ("Frozen Paradox").

| Dataset | Freeze level | RR-MoA | Best single-head | MSE $\Delta\%$ vs. single-head |
|---|---|---|---|---|
| ETTh1 | Frozen (0/8) | $\mathbf{0.690 \pm 0.021}$ | $1.241 \pm 0.015$ | $\mathbf{-44.4\%}$ |
| | Last-2 (2/8) | $\mathbf{0.727 \pm 0.074}$ | $1.030 \pm 0.139$ | $-29.4\%$ |
| | Last-4 (4/8) | $\mathbf{0.749 \pm 0.036}$ | $1.101 \pm 0.120$ | $-31.9\%$ |
| ETTm1 | Frozen (0/8) | $\mathbf{0.571 \pm 0.073}$ | $1.115 \pm 0.075$ | $\mathbf{-48.7\%}$ |
| | Last-2 (2/8) | $\mathbf{0.623 \pm 0.032}$ | $0.891 \pm 0.049$ | $-30.1\%$ |
| | Last-4 (4/8) | $\mathbf{0.571 \pm 0.034}$ | $0.866 \pm 0.016$ | $-34.1\%$ |
| Weather | Frozen (0/8) | $\mathbf{0.289 \pm 0.008}$ | $0.530 \pm 0.020$ | $-45.5\%$ |
| | Last-2 (2/8) | $\mathbf{0.251 \pm 0.005}$ | $0.478 \pm 0.025$ | $-47.4\%$ |
| | Last-4 (4/8) | $\mathbf{0.256 \pm 0.014}$ | $0.497 \pm 0.033$ | $\mathbf{-48.4\%}$ |

**Why does standard MoE routing fail?** The standard alternative – AdaMix [18], a hidden-state softmax router over the same five experts following the established sparsely-gated MoE recipe [19] – collapses to a single expert whenever backbone layers are unfrozen (Table 2, Appendix G.2): entropy reaches $0.000 \pm 0.000$ in 7/12 unfrozen configurations; the remaining 5 stay below 0.55 (max $\approx$1.61) with $\sigma$ up to 0.71. The high seed-variance distinguishes *stochastic* collapse (router commits to one expert per run, but to different experts across seeds) from *stable* partial routing (which would yield low $\sigma$ at intermediate entropy); the rescue-baseline sweep below confirms no optimization-side intervention closes either failure mode. Disabling RevIN inside MOMENT recovers entropy to 0.66–1.32. Consistent with the collapse criterion (§3), BatchNorm1d and GroupNorm in RevIN's place also collapse routing (entropy $\leq$0.004) while no normalization holds 0.817 (Table H.1); LayerNorm at the same position collapses it as well (0.000, against 0.825 without normalization; App. H). On

CIFAR-10, InstanceNorm1d collapses a frozen ViT-B/16 router (0.000) while InstanceNorm2d leaves ResNet-18 at 0.99–1.48 (App. H.1). The contrast is sharpest cross-backbone (Figure 3): under identical last-4 unfreezing on identical data, MOMENT entropy crashes to 0.000 in ∼40 steps while Timer-XL (no RevIN) stays at 1.60 for 400 steps.

Table 2: **AdaMix routing collapse** (entropy max ≈1.609). Under unfreezing, entropy →0.000. Green : disabling RevIN inside MOMENT recovers entropy (0.66–1.32), proving RevIN is the causal mechanism. Extended grid in Appendix G.

| Dataset | Freeze level | MSE (mean±std) | Routing entropy (mean±std) |
|---|---|---|---|
| ETTh1 | Frozen | $1.105 \pm 0.026$ | $0.629 \pm 0.436$ |
| | Last-2 | $1.153 \pm 0.000$ | $\mathbf{0.000 \pm 0.000}$ |
| | Last-4 | $1.154 \pm 0.001$ | $\mathbf{0.000 \pm 0.000}$ |
| | Last-4 (no RevIN) | $0.575 \pm 0.075$ | $\mathbf{1.315 \pm 0.148}$ |
| ETTm1 | Frozen | $1.008 \pm 0.012$ | $0.487 \pm 0.371$ |
| | Last-2 | $1.061 \pm 0.088$ | $0.218 \pm 0.309$ |
| | Last-4 | $1.123 \pm 0.000$ | $\mathbf{0.000 \pm 0.000}$ |
| | Last-4 (no RevIN) | $0.426 \pm 0.072$ | $\mathbf{1.103 \pm 0.376}$ |
| Weather | Frozen | $0.459 \pm 0.017$ | $0.509 \pm 0.307$ |
| | Last-2 | $0.607 \pm 0.002$ | $\mathbf{0.000 \pm 0.000}$ |
| | Last-4 | $0.607 \pm 0.002$ | $\mathbf{0.000 \pm 0.000}$ |
| | Last-4 (no RevIN) | $0.232 \pm 0.045$ | $\mathbf{0.661 \pm 0.501}$ |

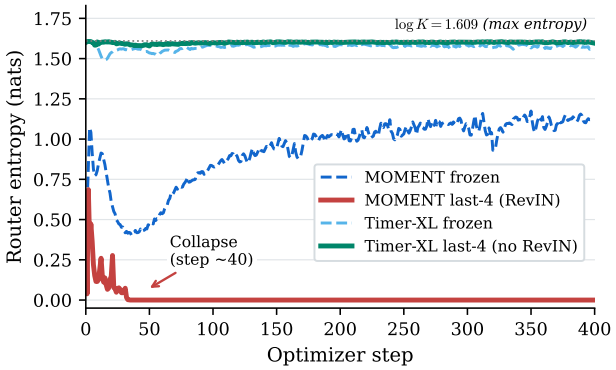


Figure 3: **Causal contrast: MOMENT vs Timer-XL** (ETTh1, last-4 unfreezing). Solid: unfrozen; MOMENT entropy collapses 0.51→0.000 in ∼40 steps; Timer-XL holds ≈1.60. Dotted: $\log K$=1.609 ceiling. Per-step trajectories in Figure G.1 (App. G.2).

In contrast, **RR-MoA** maintains healthy entropy (0.93–1.57) and wins **54/54** configurations with 26–79% MSE improvements over the best fixed adapter. The 54/54 covers the six primary-benchmark datasets (ETTh1/h2, ETTm1/m2, Weather, Electricity); on Traffic ($R$=0.14, defined in §3), held out as a low-signal-ratio control, RR-MoA correctly does not improve, consistent with Observation 1's low-$R$ prediction (Figure 4).

**The Frozen Paradox: full fine-tuning loses by** 12–79%

Within a fixed MoE architecture, does freezing match or beat fine-tuning (*RQ1*)? Frozen RR-MoA beats *full fine-tuning* (all 8 blocks, best of 5 heads × 2 LRs) by 12–79% on all 6 LTSF datasets (Figure 4). The headline gap decomposes into *(M1)* an MoE architectural advantage (26–79%, Tables 1, G.1) and *(M2)* a smaller within-architecture freeze effect ($\leq$ 13%; frozen best on $4/6$, light unfreezing wins by $\leq$13% on the other two; SR-MoA replicates $4/6$, Table F.2). Proposition 1 attributes M2 to gradient co-adaptation: a Router-Detached Gradient Flow ablation degrades MSE by +35–100% (Appendix F), and disabling RevIN recovers full-FT by 54–61% (Appendix O), pinning the mechanism on normalization. A 90-run tune (LR $10^{-5}$–$10^{-6}$, 100 epochs, cosine+warmup, layerwise decay) still loses by 51–71%.

**Proposition 1** (Gradient Co-Adaptation Drives Entropy Collapse)**.** *Consider a 2-expert MoE with backbone* $\mathbf{h}{=}A\mathbf{x}$*, router* $p{=}\sigma(\mathbf{w}^\top\mathbf{h})$*, loss* $\ell{=}p\,\ell_1{+}(1{-}p)\,\ell_2$*. Under gradient descent on* $(A,\mathbf{w})$*:* (i) *the dominant expert's loss decreases faster* ($\dot{\ell}_k \propto -p_k\|\nabla_{\mathbf{h}}\ell_k\|^2$)*, and* $\dot{p}$ *amplifies whichever expert currently has lower loss, creating self-reinforcing collapse toward* $p{\in}\{0,1\}$*;* (ii) *freezing* $A$ *removes the co-adaptation feedback through the backbone (full non-collapse additionally requires* $(\ell_1{-}\ell_2)$ *to flip sign across input regimes, i.e., expert specialization, which is satisfied by the topologically distinct RR-MoA pool). Proof in Appendix D.*

*Remark.* Proposition 1 covers linear $A\mathbf{x}$; Figure G.1b confirms it in the full 8-block Transformer within $\sim$40 steps. Part (ii)'s sign flip needs a router input that separates the regimes; the collapse criterion (§3) predicts when a normalizer removes it. Freezing is our recommended default, not a requirement: RR-MoA also beats the best fixed adapter with the last 2 or 4 blocks unfrozen (ETTh1: $-44\%$ frozen, $-29\%$ and $-32\%$ unfrozen; Table 1). RR-MoA beats independent ensembles by 37–46% (Table 4): learned routing, not mere diversity, drives the gains. §3 answered *what* the router needs to read with Observation 1 and the predictor $R(\mathcal{D})$; we now turn to the *mechanism (RQ2)*.

### $R(\mathcal{D})$ makes three predictions, and all three hold

Observation 1 and $R(\mathcal{D})$ make three testable predictions: the bound is empirically tight, $R(\mathcal{D})$ predicts raw-routing benefit across datasets, and the predictor respects a falsifiable boundary distinct from a generic-SNR explanation. We verify each in turn (*RQ2*).

**$R(\mathcal{D})$ predicts raw-routing benefit and respects a falsifiable boundary.** The bound is empirically tight (Appendix L) and the learnable mixing coefficient satisfies $\alpha{<}0.5$ on $30/30$ runs (Appendix K). $R(\mathcal{D})$ predicts raw-routing benefit at Spearman $\rho{=}-0.88$ ($p{<}0.002$, $n{=}9$; Figure 4), whereas the conditional entropy $H(E \mid S)$ does not ($\rho{=}-0.12$); Traffic ($R{=}0.14$) is a boundary, Solar ($R{=}0.06$) the one outlier (Appendix F). The discriminating power lies in the boundary case: the most likely competing explanation is that $R(\mathcal{D})$ proxies generic "hard dataset" signal-to-noise, but Traffic has high overall predictability (DLinear MSE 0.41) yet low $R$, and RR-MoA correctly does not improve there, falsifying the SNR alternative. $R(\mathcal{D})$ is therefore actionable pre-training: practitioners can compute it from raw dataset statistics to predict whether raw routing will help on a new dataset.

**Per-window statistics, not raw access, drive routing.** *(i)* Swapping $\mathbf{x}_{\text{raw}}$ for RevIN($\mathbf{x}_{\text{raw}}$) at the router degrades MSE by 60–88% while entropy stays high (Table F.1): the router needs the stripped $(M,\Sigma)$ *content*, not a non-RevIN bypass. *(ii)* Interpolating $\mathbf{x}_\alpha = (1{-}\alpha)\mathbf{x}_{\text{raw}} + \alpha\,\text{RevIN}(\mathbf{x}_{\text{raw}})$ phase-transitions only at $\alpha{\to}1$ (Figure D.1; 150 runs; SNR-onset $\alpha^*{\approx}0.40$, Appendix D.1); partial stripping is tolerable. *(iii)* Temporal shuffling changes MSE by at most 3.2% on $4/6$ datasets and lowers it on the other two, while a $[\mu,\sigma]$-only router does not match the raw router (Appendix F). The routing signal is therefore order-invariant and extends into the shape, the denominator of $R(\mathcal{D})$.

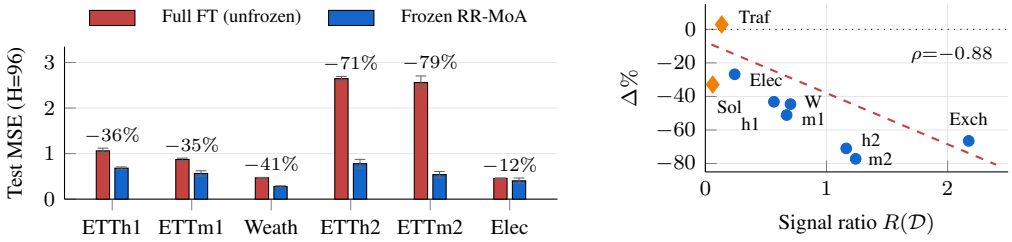


Figure 4: **Left: Frozen Paradox**, frozen RR-MoA beats full FT by 12–79% across the six LTSF datasets at $H{=}96$. **Right: Observation 1 validated**: $R(\mathcal{D})$ predicts raw-routing benefit at Spearman $\rho{=}-0.88$ ($p{<}0.002$, $n{=}9$). Traffic and Solar are boundary cases.

### Optimization-side rescues fail; raw-input fixes transfer

Must the fix be input-side (*RQ2*)? If the input distribution is the bottleneck, optimization-side rescues should not recover the gap, and the raw-input fix should transfer beyond RR-MoA's specific router. We test each.

**Rescue-baseline sweep.** The mechanism predicts that interventions targeting optimization will fail because the router's input distribution is the bottleneck, not its objective. We test this directly by

sweeping five standard MoE rescue families (Switch load-balance [19], ST-MoE z-loss [32], entropy regularization, ReMoE ReLU routing [22], expert-choice routing [23]) over 12 configurations $\times$ 6 datasets $\times$ 2 freeze levels $\times$ 5 seeds $=$ 720 run-cells (Table G.3, Appendix G.1). The best rescue (z-loss $c_z$=0.01) reduces MSE by only $10.9\%$, $2.7\times$ worse than RR-MoA; excessive balancing ($\alpha$=10.0) hurts by $+67.7\%$. Consistent with Observation 1, Part (ii): when routing information is structurally absent, no tested rescue compensates.

**The principle is architecture-agnostic.** We design three interventions, each varying one axis (router architecture, router input, expert architecture). **SR-MoA** (Figure F.1; cf. [42–44]) drops the external router for per-expert sigmoid gates $\sigma_k(\mathbf{X}_{\text{raw}})$ and *outperforms* RR-MoA by 23–38% on all 6 datasets (252 runs; Table F.2); on hidden states the same architecture degrades by $+87$–$768\%$. **AdaMix-Raw** (App. G.2) swaps the router input only: across all 12 cells, entropy rises to 1.41–1.57 (ETTh1 frozen: $0.49\rightarrow1.55$) and MSE falls by 42–88%. **Residual-IA$^+$** (App. P.1) extends the principle to the expert level. SR-MoA and AdaMix-Raw belong to the nine causal controls of Table 3, each of which rejects one competing explanation.

Table 3: **Nine causal controls.** Each rules out a competing explanation; the only mechanism consistent with all 9 is that *instance normalization strips the per-window $(\mu, \sigma)$ signal the router needs*.

| # | Control | Hypothesis (rejected) | Observed result $\Rightarrow$ verdict |
|---|---|---|---|
| C1 | RevIN ablation | Collapse is intrinsic to MoE itself, not caused by normalization | Entropy $0.000 \rightarrow 0.66$–$1.32$ when RevIN disabled (Table 2) $\Rightarrow$ RevIN is *necessary* for collapse |
| C2 | MOMENT vs Timer-XL | Architecture, not normalization, causes collapse | Opposite trajectories with identical code, only norm differs (Fig. 3) $\Rightarrow$ RevIN is *sufficient* here |
| C3 | BatchNorm1d / GroupNorm | Collapse is unique to RevIN; other per-window normalizers are immune | All three per-window normalizers collapse identically, and so does LayerNorm in RevIN's place; no normalization stays healthy (App. H) $\Rightarrow$ on time series, each of the four triggers collapse |
| C4 | 5-family rescue sweep | Optimization problem (load-balance, z-loss, ReLU, expert-choice) | 0/12 configs recover; best is $2.7\times$ worse than RR-MoA (Table G.3) $\Rightarrow$ not in the optimizer |
| C5 | AdaMix with raw router | Router architecture is the bottleneck | Same router, swap input only: entropy $\rightarrow$ 1.41–1.57; MSE $-42$–$88\%$ (App. G.2) $\Rightarrow$ the *input* matters, not the router |
| C6 | SR-MoA raw vs hidden routing | The fix only works for RR-MoA's specific Conv1d router design | Different router (per-expert sigmoid); hidden-routed $+87$–$768\%$ worse (Table F.2) $\Rightarrow$ raw-input principle generalizes |
| C7 | Vision: ViT vs. ResNet | Collapse only happens in time series, not other modalities | On CIFAR-10, InstanceNorm1d collapses a frozen ViT-B/16 router; InstanceNorm2d leaves ResNet-18 healthy (App. H.1) $\Rightarrow$ consistent with collapse where the stripped statistics carry routing signal |
| C8 | $(\mu, \sigma)$ re-injection | Just appending stripped stats back fixes it | Re-injected hidden states still $+42$–$239\%$ worse, 11/18 still collapsed (App. F.1) $\Rightarrow$ only *raw* pre-normalization input restores routing |
| C9 | $(\mu, \log\sigma)$ probe | RevIN leaves $(\mu, \sigma)$ in the router's input | Held-out ridge $R^2$, pooled hidden states, ETTh1: MOMENT $-0.311/0.600$; RevIN off $0.999/0.951$; Moirai $0.996/0.961$ (App. H.2) $\Rightarrow$ with RevIN, $\mu$ is not linearly decodable from the router's input; disabling RevIN restores it |

### RR-MoA leads the adaptation baselines and transfers across backbones and horizons

We test *method-class generality* (*RQ3*): RR-MoA beats all six adaptation baselines at $p<0.001$ (Table 4), driven by two complementary gains: residual shape heterogeneity rewards diverse experts on RevIN-normalized hidden states (26–79% over the best fixed adapter, Tables 1, G.1), while the signal-ratio $\rho=-0.88$ localizes *dispatch* information in the stripped statistics, explaining both raw-routing benefit and Traffic's null. The Pure Raw-MLP MoE ablation (App. P.3) separates "MoE with raw router" from "ensemble with raw input".

**RR-MoA extends across backbones, horizons, and tasks, and the diagnostic predicts where it should not** (*RQ3*). RR-MoA wins on every instance-normalized backbone (MOMENT 26–79%, Moirai 12–29%, Moirai-MoE [8] 37–90%), through collapse repair when normalization removes

Table 4: **Baseline comparison** (MOMENT-small, Top-2, 5 seeds, H=96; LoRA row: 3 seeds). Above double rule: frozen; full FT unfreezes all 8 blocks. Wilcoxon, Bonferroni-corrected. Grey rows: from-scratch calibration anchors.

| Method | ETTh1 | ETTm1 | Weather | $p$-value |
|---|---|---|---|---|
| Best fixed adapter | $1.254 \pm 0.026$ | $1.148 \pm 0.035$ | $0.528 \pm 0.018$ | $<0.001^{***}$ |
| Best LoRA (108-run sweep) | $1.154 \pm 0.000$ | $0.956 \pm 0.072$ | $0.600 \pm 0.052$ | $<0.001^{***}$ |
| TRACE (frozen) | $1.195 \pm 0.026$ | $1.014 \pm 0.073$ | $0.457 \pm 0.015$ | $<0.001^{***}$ |
| Ind. Ensemble (5 experts) | $1.081 \pm 0.021$ | $1.049 \pm 0.032$ | $0.509 \pm 0.024$ | $<0.001^{***}$ |
| AdaMix (frozen) | $1.109 \pm 0.031$ | $1.008 \pm 0.012$ | $0.459 \pm 0.017$ | $<0.001^{***}$ |
| Full fine-tuning (all unfrozen) | $1.060 \pm 0.059$ | $0.871 \pm 0.029$ | $0.468 \pm 0.009$ | $<0.001^{***}$ |
| **RR-MoA (Top-2, frozen)** | $\mathbf{0.680 \pm 0.027}$ | $\mathbf{0.564 \pm 0.059}$ | $\mathbf{0.276 \pm 0.017}$ | — |
| DLinear (single-Linear, from scratch) | $0.416 \pm 0.002$ | $0.322 \pm 0.004$ | $0.208 \pm 0.003$ | — |
| PatchTST/64† [13] | 0.370 | 0.274 | 0.149 | — |
| iTransformer‡ [14] | 0.386 | 0.334 | 0.174 | — |

†Trained from scratch, $L$=512; ‡trained from scratch, $L$=96. Included for absolute MSE context.

routing statistics and through richer router inputs otherwise. Chronos and Timer-XL [6], with no instance normalization, show no collapse, as predicted. Beyond $H$=96, RR-MoA wins $12/12$ multi-horizon vs. best fixed adapter and $7/8$ imputation cells (Tables C.1, N.1, J.1). It beats that adapter in $21/21$ runs at $H \in \{1000, 2000\}$ and $70/72$ runs over input lengths $L \in \{96, 192, 336, 512\}$, where raw routing beats RevIN routing in all $24$ configurations. On UEA classification and SMD anomaly detection, routing stays non-collapsed, although frozen features limit task performance. With the last four blocks unfrozen, a hidden-state router over the same experts collapses in $9/9$ classification and $3/9$ anomaly runs. The raw router stays non-collapsed in all $18$ runs (App. Q.2). The diagnostic predicts failures: Traffic ($R$=0.14) shows no raw-routing benefit, consistent with its low $R(\mathcal{D})$; $R(\mathcal{D})$ is computable pre-training, giving an *a priori* go/no-go test.

**Accuracy against linear baselines.** RR-MoA changes only the router's input to probe the mechanism. It trails a from-scratch linear model on every dataset of Table 4 (ETTh1 $0.680$ vs. $0.416$). Residual-IA$^+$ (App. P.1) is the deployment adapter: each expert adds a gated backbone residual to a shared NLinear-style raw branch. Both linear baselines use $n$=10 seeds. NLinear computes $\mathrm{Linear}(\mathbf{x}-x_T)+x_T$. DLinear uses a single linear layer without trend-seasonal decomposition. With match-or-beat defined as lower mean MSE or a gap within $2\%$, Residual-IA$^+$ achieves $96/123$ (backbone, dataset, horizon) cells against NLinear and $102/123$ against DLinear (Table P.9). Against NLinear the result holds through $H$=336 ($27/33$, $25/30$, $25/30$), with $19/30$ at $H$=720. Significant win/loss counts (Welch tests, $p<0.05$) are $41$:$11$ against NLinear and $44$:$12$ against DLinear. At $H$=96, the backbone-free raw-MLP MoE has $0.7$–$17.9\%$ lower MSE than the dual-stream variant on ETTh2, ETTm2 and Electricity, but 3–$7\%$ higher MSE on ETTh1, ETTm1 and Weather (App. P.3).

## 5 Conclusion and Limitations

**Summary.** Our contribution is a diagnosis and a training-free predictor of it. We identified *normalization-induced routing collapse*: $(\mu, \sigma)$ stripping that handles distribution shift also removes the signal a router needs. Observation 1 lower-bounds the loss, $R(\mathcal{D})$ predicts it ($\rho$=$-0.88$), and nine causal controls rule out optimization/geometry/paradigm alternatives. RR-MoA ($54/54$) and SR-MoA (23–$38\%$ lower MSE than RR-MoA) confirm the diagnosis, and Residual-IA$^+$, the deployment adapter that follows from the diagnosis, matches or beats NLinear in $96/123$ cells. *TSFM users:* to adapt a $(\mu, \sigma)$-stripping backbone, freeze it by default and route on raw input (full fine-tuning loses by 12–$79\%$, App. O); $R(\mathcal{D})$ is a pre-training go/no-go test (Traffic, $R$=0.14: predicted no improvement). *MoE researchers:* eleven optimization-side rescues recover $\leq 10.9\%$ MSE; the input-side fix recovers $67\%$. The prediction extends to FAN [34] and vision (App. H.1). **Limitations:** specific to $(\mu, \sigma)$-stripping normalizers; learned expert pools remain open.

## Acknowledgments and Disclosure of Funding

This work was funded entirely by RMIT University Vietnam. The authors declare no competing interests.

## References


[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

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

[3] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

[4] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research (TMLR)*, 2024.

[5] Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

[6] Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer-XL: Long-context transformers for unified time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2025.

[7] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

[8] Xu Liu, Juncheng Liu, Gerald Woo, Taha Aksu, Yuxuan Liang, Roger Zimmermann, Chenghao Liu, Junnan Li, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Moirai-MoE: Empowering time series foundation models with sparse mixture of experts. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

[9] Shanghua Gao, Teddy Koker, Owen Queen, Thomas Hartvigsen, Theodoros Tsiligkaridis, and Marinka Zitnik. UniTS: A unified multi-task time series model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[10] Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Bilos, Hena Ghonia, Nadhir Vincent Hassen, Anderson Schneider, Sahil Garg, Alexandre Drouin, Nicolas Chapados, Yuriy Nevmyvaka, and Irina Rish. Lag-Llama: Towards foundation models for probabilistic time series forecasting. *arXiv preprint arXiv:2310.08278*, 2024.

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

[12] Vijay Ekambaram, Arindam Jati, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. TTMs: Fast multi-level tiny time mixers for improved zero-shot and few-shot forecasting of multivariate time series. *arXiv preprint arXiv:2401.03955*, 2024.

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

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

[15] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

[16] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

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

[18] Yaqing Wang, Sahaj Agarwal, Subhabrata Mukherjee, Xiaodong Liu, Jing Gao, Ahmed Hassan Awadallah, and Jianfeng Gao. AdaMix: Mixture-of-adaptations for parameter-efficient model tuning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

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

[20] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, 2004.

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

[22] Ziteng Wang, Jun Zhu, and Jianfei Chen. ReMoE: Fully differentiable mixture-of-experts with ReLU routing. In *International Conference on Learning Representations (ICLR)*, 2025.

[23] Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew Dai, Zhifeng Chen, Quoc Le, and James Laudon. Mixture-of-experts with expert choice routing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[24] Zhongzheng Qiao, Chenghao Liu, Yiming Zhang, Ming Jin, Quang Pham, Qingsong Wen, P.N. Suganthan, Xudong Jiang, and Savitha Ramasamy. Multi-scale finetuning for encoder-based time series foundation models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[25] Yuze Li and Wei Zhu. TRACE: Time series parameter efficient fine-tuning. *arXiv preprint arXiv:2503.16991*, 2025.

[26] Divij Gupta, Anubhav Bhatti, and Surajsinh Parmar. Beyond LoRA: Exploring efficient fine-tuning techniques for time series foundational models. In *NeurIPS 2024 Workshop on Time Series in the Age of Large Models*, 2024.

[27] Abdelhakim Benechehab, Vasilii Feofanov, Giuseppe Paolo, Albert Thomas, Maurizio Filippone, and Balazs Kégl. AdaPTS: Adapting univariate foundation models to probabilistic multivariate time series forecasting. *arXiv preprint arXiv:2502.10235*, 2025.

[28] Lifan Zhao, Yanyan Shen, Zhaoyang Liu, Xue Wang, and Jiaji Deng. Less is more: Unlocking specialization of time series foundation models via structured pruning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[29] Matthew Faw, Rajat Sen, Yichen Zhou, and Abhimanyu Das. In-context fine-tuning for time-series foundation models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

[30] Weiyang Zhang, Xinyang Chen, Xiucheng Li, Kehai Chen, Weili Guan, and Liqiang Nie. Unified transferability metrics for time series foundation models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[31] Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. GIFT-Eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024.

[32] Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. ST-MoE: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*, 2022.

[33] Wei Fan, Pengyang Wang, Dongkun Wang, Dongjie Wang, Yuanchun Zhou, and Yanjie Fu. DishTS: A general paradigm for alleviating distribution shift in time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2023.

[34] Weiwei Ye, Songgaojun Deng, Qiaosha Zou, and Ning Gui. FAN: Frequency adaptive normalization for non-stationary time series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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

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

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

[38] Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. In *NeurIPS Track on Datasets and Benchmarks*, 2021.

[39] Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 95–104, 2018.

[40] Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: A survey. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2023.

[41] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[42] Ang Lv, Ruobing Xie, Yining Qian, Songhao Wu, Xingwu Sun, Zhanhui Kang, Di Wang, and Rui Yan. Autonomy-of-experts models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

[43] Anzhe Cheng, Shukai Duan, Shixuan Li, Chenzhong Yin, Mingxi Cheng, Heng Ping, Tamoghna Chattopadhyay, Sophia I. Thomopoulos, Shahin Nazarian, Paul Thompson, and Paul Bogdan. ERMoE: Eigen-reparameterized mixture-of-experts for stable routing and interpretable specialization. *arXiv preprint arXiv:2511.10971*, 2025.

[44] Yilun Liu, Jinru Han, Sikuan Yan, Volker Tresp, and Yunpu Ma. Routing-free mixture-of-experts. *arXiv preprint arXiv:2604.00801*, 2026.

[45] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

[46] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pages 368–377, 2000.

[47] Yong Liu, Guo Qin, Chenyu Li, Jianmin Wang, and Mingsheng Long. Timer-S1: A billion-scale time series foundation model with serial scaling. *arXiv preprint arXiv:2603.04791*, 2026.

[48] Gaspard Berthelier, Tahar Nabil, Etienne Le Naour, Richard Niamke, Samir Perlaza, and Giovanni Neglia. On the role of reversible instance normalization. *arXiv preprint arXiv:2603.11869*, 2026.

[49] Xi Wang, Soufiane Hayou, and Eric Nalisnick. The myth of expert specialization in MoEs: Why routing reflects geometry, not necessarily domain expertise. *arXiv preprint arXiv:2604.09780*, 2026.

[50] Xiandong Zou, Jia Li, Xiaotong Yuan, and Pan Zhou. IBNorm: Information-bottleneck inspired normalization for representation learning. *arXiv preprint arXiv:2510.25262*, 2025.

[51] Zewen Chi, Li Dong, Shaohan Huang, Damai Dai, Shuming Ma, Barun Patra, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, Heyan Huang, and Furu Wei. On the representation collapse of sparse mixture of experts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[52] Xun Wu, Shaohan Huang, Wenhui Wang, Shuming Ma, Li Dong, and Furu Wei. Multi-head mixture-of-experts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[53] Hongcan Guo, Haolang Lu, Guoshun Nan, Bolun Chu, Jialin Zhuang, Yuan Yang, Wenhao Che, Xinye Cao, Sicong Leng, Qimei Cui, and Xudong Jiang. Advancing expert specialization for better mixture-of-experts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[54] YongXiang Hua, Haoyu Cao, Zhou Tao, Bocheng Li, Zihao Wu, Chaohu Liu, and Linli Xu. Input domain aware MoE: Decoupling routing decisions from task optimization in mixture of experts. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*, 2025.

[55] Sofiane Ennadir, Levente Zólyomi, Oleg Smirnov, Tianze Wang, John Pertoft, Filip Cornell, and Lele Cao. Pool me wisely: On the effect of pooling in transformer-based models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[56] Michał Wiliński, Mononito Goswami, Willa Potosnak, Nina Żukowska, and Artur Dubrawski. Exploring representations and interventions in time series foundation models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267, pages 66861–66886, 2025.

[57] Atharva Pandey, Abhilash Neog, and Gautam Jajoo. On the internal semantics of time-series foundation models. In *NeurIPS 2025 Workshop on Recent Advances in Time Series Foundation Models*, 2025.

[58] Hui He, Kun Yi, Yuanchi Ma, Qi Zhang, Zhendong Niu, and Guansong Pang. SEMPO: Lightweight foundation models for time series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

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

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

[61] Xun Wu, Shaohan Huang, and Furu Wei. Mixture of LoRA experts. In *International Conference on Learning Representations (ICLR)*, 2024.

[62] Shiyu Wang, Jiawei Li, Xiaoming Shi, Zhou Ye, Baichuan Mo, Wenze Lin, Shengtong Ju, Zhixuan Chu, and Ming Jin. TimeMixer++: A general time series pattern machine for universal predictive analysis. In *International Conference on Learning Representations (ICLR)*, 2025.

[63] Xingjian Wu, Xiangfei Qiu, Hanyin Cheng, Zhengyu Li, Jilin Hu, Chenjuan Guo, and Bin Yang. Enhancing time series forecasting through selective representation spaces: A patch perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. Spotlight.

[64] Ben Cohen, Emaad Khwaja, Youssef Doubli, Salahidine Lemaachi, Chris Lettieri, Charles Masson, Hugo Miccinilli, Elise Ramé, Qiqi Ren, Afshin Rostamizadeh, Jean Ogier du Terrail, Anna-Monica Toon, Kan Wang, Stephan Xie, Zongzhe Xu, Viktoriya Zhukova, David Asker, Ameet Talwalkar, and Othmane Abou-Amal. This time is different: An observability perspective on time series foundation models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[65] Tao Dai, Beiliang Wu, Peiyuan Liu, Naiqi Li, Xue Yuerong, Shu-Tao Xia, and Zexuan Zhu. DDN: Dual-domain dynamic normalization for non-stationary time series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[66] Nunzio A. Letizia, Nicola Novello, and Andrea M. Tonello. Mutual information estimation via $f$-divergence and data derangements. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[67] Gokul Gowri, Xiao-Kang Lun, Allon M. Klein, and Peng Yin. Approximating mutual information of high-dimensional variables using learned representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. Spotlight.

[68] Ashwinee Panda, Vatsal Baherwani, Zain Sarwar, Benjamin Thérien, Stephen Rawls, Sambit Sahu, Supriyo Chakraborty, and Tom Goldstein. Dense backpropagation improves routing for sparsely-gated mixture-of-experts. In *NeurIPS 2024 Workshop on Efficient Natural Language and Speech Processing (ENLSP)*, 2024.

[69] Yisong Fu, Zezhi Shao, Chengqing Yu, Yujie Li, Zhulin An, Qi Wang, Yongjun Xu, and Fei Wang. Selective learning for deep time series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[70] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

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

[72] Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. AutoTimes: Autoregressive time series forecasters via large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[73] Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. Oral.

[74] Zijie Pan, Yushan Jiang, Sahil Garg, Anderson Schneider, Yuriy Nevmyvaka, and Dongjin Song. $S^2$IP-LLM: Semantic space informed prompt learning with LLM for time series forecasting. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

[75] Xu Liu, Junfeng Hu, Yuan Li, Shizhe Diao, Yuxuan Liang, Bryan Hooi, and Roger Zimmermann. UniTime: A language-empowered unified model for cross-domain time series forecasting. In *Proceedings of the ACM Web Conference (WWW)*, 2024.

[76] Zhijian Xu, Ailing Zeng, and Qiang Xu. FITS: Modeling time series with 10k parameters. In *International Conference on Learning Representations (ICLR)*, 2024. Spotlight.

[77] Si-An Chen, Chun-Liang Li, Nate Yoder, Sercan O. Arik, and Tomas Pfister. TSMixer: An all-MLP architecture for time series forecasting. *Transactions on Machine Learning Research (TMLR)*, 2023.

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

[79] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations (ICLR)*, 2021.

[80] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[81] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[82] Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[83] Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. Sparse upcycling: Training mixture-of-experts from dense checkpoints. In *International Conference on Learning Representations (ICLR)*, 2023.

[84] David Raposo, Sam Ritter, Blake Richards, Timothy Lillicrap, Peter Conway Humphreys, and Adam Santoro. Mixture-of-depths: Dynamically allocating compute in transformer-based language models. *arXiv preprint arXiv:2404.02258*, 2024.

[85] Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A survey on mixture of experts in large language models. *arXiv preprint arXiv:2407.06204*, 2024.

[86] Tong Zhu, Xiaoye Qu, Daize Dong, Jiacheng Ruan, Jingqi Tong, Conghui He, and Yu Cheng. LLaMA-MoE: Building mixture-of-experts from LLaMA with continual pre-training. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[87] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H. Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1930–1939, 2018.

[88] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R. Devon Hjelm. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[89] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[90] Ben Poole, Sherjil Ozair, Aäron van den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

[91] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations (ICLR)*, 2017.

[92] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

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

[94] Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. The UEA multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075*, 2018.

[95] Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 2828–2837, 2019.

# A Experimental Setup Details

**Notation.** Table A.1 consolidates symbols used across §3–§4.2, Observation 1, and Appendix D.

Table A.1: **Notation.** Symbols used throughout the paper.

| Symbol | Meaning |
|---|---|
| *Data and task* | |
| $\mathbf{x} \in \mathbb{R}^{T \times c}$ | Input window: $T$=512 time steps, $c$ channels |
| $\mathbf{y} \in \mathbb{R}^{h \times c}$ | Forecast target at horizon $h \in \{96, 192, 336, 720\}$ |
| $\hat{\mathbf{y}}$ | Model prediction |
| $M_i, \Sigma_i, \mathbf{S}_i$ | Per-window mean, scale, RevIN-normalized shape ($\mathbf{S}_i$=$(\mathbf{x}_i - M_i)/\Sigma_i$) |
| *Backbone and adapter* | |
| $f_\theta : \mathbb{R}^{T \times c} \to \mathbb{R}^{P \times d}$ | Pretrained TSFM feature extractor; $\theta$ frozen ($\nabla_\theta \equiv 0$) |
| $\mathbf{H}$=$f_\theta(\mathbf{x})$ | Backbone hidden states ($P$ patches of dim $d$) |
| $g_\phi : \mathbb{R}^{P \times d} \to \mathbb{R}^{h \times c}$ | Adapter; $\lvert\phi\rvert \leq$ 500K trainable params |
| *Mixture and routing* | |
| $K$ | Total expert pool size (we use $K$=5) |
| $k$ | Top-$k$ active experts per window (default $k$=2) |
| $\mathrm{Expert}_j$ | $j$-th expert head ($j$=$1, \ldots, K$); pool: mean/last/max/attention/conv1d |
| $G_\psi : \mathbb{R}^{T \times c} \to \mathbb{R}^K$ | Router; $\psi$ are 853 trainable params |
| $\widetilde{w} \in \Delta^{K-1}$ | Sparse routing weights, $\lVert\widetilde{w}\rVert_0$=$k$ |
| $\mathcal{T} \subset \{1, \ldots, K\}$ | Top-$k$ active expert indices per window |
| $E \in \{1, \ldots, K\}$ | Random variable for expert assignment (Observation 1) |
| *Diagnostics* | |
| $R(\mathcal{D})$ | Signal ratio $[\mathrm{Var}(M)+\mathrm{Var}(\Sigma)]/\overline{\mathrm{Var}(S)}$ predicting raw-routing benefit |
| $\alpha \in [0, 1]$ | Interpolation coefficient: $\mathbf{x}_\alpha$=$(1-\alpha)\mathbf{x}_{\text{raw}}$+$\alpha$ RevIN($\mathbf{x}_{\text{raw}}$) |
| $\varepsilon$ | Residual statistic-shape leakage $I(M, \Sigma; S)$ in Observation 1, Part (ii) |
| $H(E)$ | Routing entropy in nats; collapse⇒ $H(E)\to 0$, ceiling $\log K \approx 1.609$ |

**Backbone specifications.** MOMENT-small ($d$=512, 8 encoder blocks, internal RevIN with learnable affine); MOMENT-large ($d$=1024, 24 blocks, internal RevIN with learnable affine); Moirai ($d$=384, 6 blocks, RMSNorm internally with non-learnable per-instance I/O scaling, no learnable-affine norm inside the encoder); Moirai-MoE ($d$=384, 6 blocks, RMSNorm internally with sparse internal experts and non-learnable per-instance I/O scaling); Chronos ($d$=512, 6 encoder blocks, T5 RMSNorm internally, no instance normalization); Timer-XL ($d$=1024, 8 blocks, LayerNorm internally, no instance normalization). Throughout, "full fine-tuning" means *all* encoder blocks are unfrozen (8/8 for MOMENT-small, 24/24 for MOMENT-large, 6/6 for Moirai/Moirai-MoE/Chronos, 8/8 for Timer-XL).

**Training hyperparameters.** Adam optimizer, learning rate $10^{-3}$, MSE loss, batch size 128, 15 epochs. Final reported numbers are 15-epoch test MSE. Core multi-seed grids use seeds $\{42, 43, 44, 45, 46\}$ ($n$=5); ablations use seeds $\{42, 43, 44\}$ ($n$=3). Channel-wise StandardScaler normalization [13, 36] is applied to all inputs.

**Per-baseline configurations.**

- **Fixed adapters**: linear, attention-pool, Conv1d-pool single-head topologies with $\leq$ 500,000 trainable parameters.
- **LoRA** [15]: 108-run sweep (12 configurations × 3 seeds × 3 datasets) over rank, target projections, and forecast head; full table in Appendix E.
- **TRACE** [25]: importance-based LoRA selection with default settings.
- **Independent ensemble**: 5 experts with the canonical pool topologies, trained independently with no router and averaged at inference.
- **AdaMix** [18]: $K$=5 canonical heads with hidden-state softmax router; load-balance coefficient 0.01. Full implementation details in Appendix G.2.

- **Full fine-tuning**: all encoder blocks unfrozen; best of 5 head topologies $\times$ 2 learning rates ($10^{-3}$, $10^{-4}$). Extended sweeps in Appendix O cover 50-epoch cosine schedules at $10^{-5}$.
- **DLinear** [17]: from-scratch supervised; we use the simplified single-Linear variant (Linear($L$=512 $\rightarrow$ $H$=96), 49K params), without the trend-seasonal decomposition, as the calibration anchor. Calibration values $0.416/0.322/0.208$ on ETTh1/ETTm1/Weather (gap to RR-MoA visible in Tables 1, 4).

**RR-MoA pseudocode.** Algorithm 1 gives the full procedure.

**Algorithm 1** RR-MoA: Raw-Routed Mixture of Adapters with Top-$k$ sparse routing

**Require:** TSFM $f_\theta$ (default: strictly frozen; optional: last-$\ell$ blocks unfrozen), raw input $\mathbf{X}_{\text{raw}} \in \mathbb{R}^{B\times L}$, $K$ expert adapters, sparsity $k \leq K$

$\mathbf{H} \leftarrow f_\theta(\mathbf{X}_{\text{raw}})$ {Frozen default: no gradients through $\theta$; for last-$\ell$/all unfreezing (Table 1), the unfrozen blocks receive gradients normally}
$\mathbf{g} \leftarrow G_\psi(\mathbf{X}_{\text{raw}})$ {Raw-signal gate logits, $\mathbf{g} \in \mathbb{R}^{B\times K}$}
$\mathcal{T} \leftarrow \text{TopK}(\mathbf{g}, k)$ {Per-sample indices of the $k$ selected experts}
$\widetilde{\mathbf{w}} \leftarrow \text{softmax}_{\mathcal{T}}(\mathbf{g})$ {Renormalized over selected experts only}
**for each** $j \in \mathcal{T}$ **do**
  $\mathbf{o}_j \leftarrow \text{Expert}_j(\mathbf{H})$ {Only the $k$ selected experts execute}
**end for**
$\mathbf{Y} \leftarrow \sum_{j\in\mathcal{T}} \widetilde{w}_j \cdot \mathbf{o}_j$
**return** $\mathbf{Y}$

### A.1 Significance Under Alternative Corrections

Bonferroni correction controls false positives at a cost in power, so we also report the RR-MoA versus best-fixed-adapter tests without correction and under Holm and Benjamini–Hochberg (BH) correction (Table A.2). We use one-sided Wilcoxon signed-rank tests on paired seeds (MOMENT-small frozen, Top-2, $K$=5, $H$=96). Here $n$ counts the seeds whose runs also train the three fixed adapters. We apply each correction to the six per-dataset tests. RR-MoA has lower MSE on every paired seed of every dataset. Each per-dataset $p$-value therefore equals the exact test's minimum of $2^{-n}$ ($1/128$, $1/32$ and $1/16$ at $n$=7, 5 and 4). Seed count therefore limits per-dataset significance. Pooled over all 35 paired seeds, the test gives $p$=$2.9\times10^{-11}$ uncorrected and $8.7\times10^{-10}$ after a conservative Bonferroni factor of 30, an upper bound on the Holm and BH values. The Traffic boundary case ($R$=0.14, $n$=3) stays null under the uncorrected test (+2.9% MSE, $p$=0.88), so the Traffic null holds under the most powerful of the four treatments.

Table A.2: **RR-MoA vs. best fixed adapter under four multiple-comparison treatments** (one-sided Wilcoxon signed-rank on paired seeds; MOMENT-small frozen, Top-2, $K$=5, $H$=96). Bonferroni, Holm and BH correct over the six datasets. RR-MoA wins every paired seed, so each uncorrected $p$ equals $2^{-n}$.

| Dataset | $n$ | Uncorrected | Bonferroni | Holm | BH |
|---|---|---|---|---|---|
| ETTh1 | 7 | 0.0078 | 0.047 | 0.047 | 0.016 |
| ETTh2 | 5 | 0.031 | 0.19 | 0.094 | 0.038 |
| ETTm1 | 7 | 0.0078 | 0.047 | 0.047 | 0.016 |
| ETTm2 | 4 | 0.063 | 0.38 | 0.094 | 0.063 |
| Weather | 7 | 0.0078 | 0.047 | 0.047 | 0.016 |
| Electricity | 5 | 0.031 | 0.19 | 0.094 | 0.038 |
| Pooled | 35 | $2.9\times10^{-11}$ | $<10^{-9}$ under each correction (Bonferroni factor 30) | | |

## B Deployment-Regime Motivation

We motivate the frozen-backbone paradigm with three deployment regimes where full fine-tuning is infeasible:

**(i) Multi-tenant cloud serving.** A single TSFM sits in GPU memory and serves many tenants via lightweight, hot-swappable per-tenant adapters in host RAM. Per-tenant DLinear requires collecting, storing, and periodically retraining on each tenant's historical data ($O(N)$ training pipelines for $N$ tenants), whereas the frozen backbone is loaded once and adapter weights are swapped from host RAM. The adapter approach also enables instant onboarding of new tenants via zero-shot transfer, a capability absent from per-tenant supervised baselines.

**(ii) On-device and edge deployment.** One backbone is shipped with the device image; per-site adapters are updated over low-bandwidth links. Collecting each site's training data centrally is often impossible for privacy or latency reasons.

**(iii) Cross-task reuse.** The same frozen backbone drives forecasting, imputation, and classification via different expert pools. We demonstrate the first two tasks explicitly in the main text.

### B.1 Inference Benchmark

With the deployment regimes specified above, we now quantify the inference-time cost of each architectural variant on a single A10G and contrast it with DLinear, the per-task baseline that those regimes rule out.

Table B.1: **Inference latency and memory across all three architectural variants** (MOMENT-small, A10G GPU, batch size 128, 200 repeats with 20-step CUDA-event warmup; synthetic random inputs at the standard LTSF batch shape, since latency for non-autoregressive deterministic forward passes is data-oblivious). All three architectural fixes (RR-MoA, SR-MoA, Residual-IA$^+$) add only single-digit milliseconds of adapter overhead on top of the backbone forward pass; the frozen-backbone fast-inference property is preserved by each.

| Method | Latency (ms) | $\Delta$ vs. backbone | Peak GPU (MB) | Adapter params |
|---|---|---|---|---|
| Backbone only | $52.89 \pm 0.36$ | — | 359 | — |
| Single adapter (Conv1dPool) | $53.57 \pm 0.37$ | $+1.3\%$ | 360 | 268K |
| **RR-MoA (Top-2, $K$=5)** | $\mathbf{56.02 \pm 0.44}$ | $\mathbf{+5.9\%}$ | **362** | **426K** |
| **SR-MoA (dense $K$=5)** | $\mathbf{54.20 \pm 0.79}$ | $\mathbf{+2.5\%}$ | **364** | **466K** |
| **Residual-IA$^+$** | $\mathbf{54.53 \pm 0.43}$ | $\mathbf{+3.1\%}$ | **365** | **354K** |
| DLinear (no backbone) | $0.04 \pm 0.00$ | — | 160 | 49K |

The RR-MoA routing overhead ($+3.1$ ms) is dominated by the sparse expert selection loop; the router itself (Conv1d + linear, 853 params) adds $< 0.1$ ms. SR-MoA dense and Residual-IA$^+$ both run *faster* than RR-MoA Top-2 on GPU because their dense per-expert execution parallelizes over the $K$=5 expert dimension without per-sample sparse-routing kernel launches. Residual-IA$^+$ has *fewer* adapter parameters than RR-MoA (354K vs 426K) thanks to the shared NLinear raw branch.

**Router input memory.** The raw router reads the backbone's own input window, $\mathbf{X}_{\text{raw}} \in \mathbb{R}^{B \times L}$: at $B$=128 and $L$=512 this is 65,536 float32 values (0.26 MB). A hidden-state router pools $\mathbf{H} \in \mathbb{R}^{B \times P \times d}$, which for MOMENT-small ($P$=64 patches, $d$=512) holds 4.2M float32 values (16.8 MB), $64\times$ more. Even a separate GPU copy of the window uses less than $0.1\%$ of the frozen backbone's peak memory (359 MB), which all adapter methods require. RR-MoA, SR-MoA and Residual-IA$^+$ add 3–6 MB to that peak (Table B.1).

**What our adapters deliver in practice.** The numbers above quantify the cost of accessing a regime DLinear cannot serve at all. (i) *Adapter overhead is bounded.* All three of our architectures stay within $+5.9\%$ of the raw backbone latency and add at most 6 MB of peak GPU memory (Table B.1); the frozen backbone is the cost floor, and the routing/expert machinery does not blow it up. (ii) *Hot-swap, not retrain.* A new tenant or task is served by loading 354–466 K adapter parameters from host RAM into the resident backbone in sub-millisecond time; the per-tenant DLinear alternative requires collecting that tenant's history and running a fresh supervised training pipeline before the first prediction. (iii) *One model, many tasks.* The same frozen backbone drives forecasting, imputation, and classification through different expert pools (§1, Appendices J and Q); $N$ tasks consume one GPU residency rather than $N$. (iv) *Zero-shot onboarding.* Adapter weights trained on related tenants transfer at load time, so a new deployment can serve predictions immediately rather than waiting for the data-collection-plus-training cycle that supervised per-task models require. The latency and

memory numbers above are the price of these capabilities, not a claim of dominance over DLinear in DLinear's native single-task setting.

## C Multi-Horizon Evaluation

We extend the main-text $H{=}96$ result to the standard LTSF horizon set $H{\in}\{96, 192, 336, 720\}$ on ETTh1, ETTm1, and Weather (Table C.1), confirming that RR-MoA's 26–79% improvement over the best fixed adapter is not horizon-specific: it wins all 12 (dataset, horizon) cells vs. best fixed adapter. The DLinear column is reported only as the supervised calibration anchor (RR-MoA loses to it across all 12 cells; Appendix P scores Residual-IA$^+$ against DLinear and NLinear); the RR-MoA-vs-DLinear gap narrows with $H$ on ETTh1 and Weather, consistent with the frozen backbone's horizon-invariant representation becoming relatively more informative as the prediction window lengthens.

Table C.1: **Multi-horizon RR-MoA vs. DLinear** (MOMENT-small, strictly frozen, Top-2, 5 seeds, test MSE mean±std). RR-MoA wins all 12 configurations vs. best fixed baseline across horizons 96–720. DLinear column provides supervised calibration: the gap narrows from $+63\%$ to $+48\%$ on ETTh1 and from $+33\%$ to $+15\%$ on Weather as $H$ grows, consistent with the frozen backbone's horizon-invariant representation becoming relatively more informative at longer horizons.

| | ETTh1 | | | ETTm1 | | | Weather | | |
|---|---|---|---|---|---|---|---|---|---|
| $H$ | RR-MoA | Best fix | DLin | RR-MoA | Best fix | DLin | RR-MoA | Best fix | DLin |
| 96 | 0.680±.030 | 1.233 | *0.420* | 0.564±.066 | 1.169 | *0.326* | 0.276±.019 | 0.522 | *0.207* |
| 192 | 0.773±.071 | 1.326 | *0.476* | 0.632±.039 | 1.141 | *0.365* | 0.292±.009 | 0.498 | *0.246* |
| 336 | 0.776±.069 | 1.316 | *0.502* | 0.658±.060 | 1.125 | *0.415* | 0.329±.016 | 0.516 | *0.286* |
| 720 | 0.816±.049 | 1.379 | *0.553* | 0.713±.048 | 1.121 | *0.454* | 0.401±.017 | 0.548 | *0.350* |

## D Proposition Proofs

**Proof of Observation 1** (Routing Information Loss Under Instance Normalization).

*Proof.* Throughout, $X$ denotes a raw window, $(M, \Sigma)$ its per-window mean/scale, $S = (X - M)/\Sigma$ its RevIN-normalized shape (defined whenever $\Sigma > 0$; constant windows form a measure-zero exception that we exclude by convention), and $E$ a (possibly randomized) router assignment. On the support $\{\Sigma > 0\}$ the map $X \leftrightarrow (S, M, \Sigma)$ is a measurable bijection, so $I(X; Y) = I(S, M, \Sigma; Y)$ for any $Y$, and in particular $I(X; E) = I(S, M, \Sigma; E)$.

*Part (i): Exact decomposition.* By the chain rule of mutual information applied to $(S, M, \Sigma; E)$,

$$I(X; E) \;=\; I(S, M, \Sigma;\, E) \;=\; I(S;\, E) + I(M, \Sigma;\, E \mid S).$$

Rearranging, $I(X; E) - I(S; E) = I(M, \Sigma;\, E \mid S) \geq 0$. This identity holds *without any independence assumption* between $(M, \Sigma)$ and $S$.

*Part (ii): Unconditional quantitative bound.* We bound the conditional mutual information $I(M, \Sigma; E \mid S)$ from below in terms of the *unconditional* quantities $I(M, \Sigma; E)$ and $I(M, \Sigma; S)$. Applying the chain rule two different ways to $I(M, \Sigma; E, S)$,

$$\begin{aligned} I(M, \Sigma;\, E, S) &= I(M, \Sigma;\, S) + I(M, \Sigma;\, E \mid S) \qquad \text{(expand } S \text{ first)} \\ &= I(M, \Sigma;\, E) + I(M, \Sigma;\, S \mid E) \qquad \text{(expand } E \text{ first).} \end{aligned}$$

Equating and solving for $I(M, \Sigma; E \mid S)$,

$$I(M, \Sigma;\, E \mid S) \;=\; I(M, \Sigma;\, E) \;-\; I(M, \Sigma;\, S) \;+\; I(M, \Sigma;\, S \mid E).$$

Since $I(M, \Sigma; S \mid E) \geq 0$ (mutual information is non-negative), we obtain

$$I(X; E) - I(S; E) \;=\; I(M, \Sigma;\, E \mid S) \;\geq\; I(M, \Sigma;\, E) \;-\; I(M, \Sigma;\, S),$$

where the residual $\varepsilon := I(M, \Sigma; S) \geq 0$ is the *leakage* of location-scale statistics into the shape. Equality (tightness) holds iff $I(M, \Sigma; S \mid E) = 0$, i.e. iff $(M, \Sigma) \perp\!\!\!\perp S \mid E$. In particular, unconditional independence $(M, \Sigma) \perp\!\!\!\perp S$ implies $\varepsilon = 0$, and the bound reduces to $I(X; E) - I(S; E) \geq I(M, \Sigma; E)$. Note that $(M, \Sigma) \perp\!\!\!\perp S$ does *not* imply $(M, \Sigma) \perp\!\!\!\perp S \mid E$

in general (conditioning on a common consequence can introduce dependence), so this corollary recovers the classical bound but does not automatically achieve tightness; tightness is an empirical question measured in Part (iv) and Appendix L.

*Part (iii): Full-information corner.* Suppose (a) the router depends entirely on the stripped statistics, $E \perp\!\!\!\perp S \mid (M, \Sigma)$ (equivalently, $p(E, S \mid M, \Sigma) = p(E \mid M, \Sigma)\, p(S \mid M, \Sigma)$, a symmetric conditional independence), (b) $(M, \Sigma) \perp\!\!\!\perp S$ unconditionally, and (c) $E$ is determined by $(M, \Sigma)$, so $H(E \mid M, \Sigma) = 0$. Conditions (a) and (b) yield the Markov chain $E - (M, \Sigma) - S$ in either direction; applying the data-processing inequality on this chain gives $I(S; E) \leq I(S; (M, \Sigma)) = 0$ (the second equality from (b)), so $I(S; E) = 0$. From (a) and (c), $I(X; E) = I((S, M, \Sigma); E) = I((M, \Sigma); E) + I(S; E \mid (M, \Sigma)) = H(E) + 0 = H(E)$. Hence $I(X; E) - I(S; E) = H(E)$: RevIN destroys the entire routing distribution.

*Part (iv): Empirical leakage and signal ratio.* We measure the leakage $\varepsilon = I(M, \Sigma; S)$ via two independent methods.

*Gaussian reference.* Since $(M, \Sigma)$ is 2-dimensional, there are at most 2 canonical correlations $\rho_1, \rho_2$ between $(M, \Sigma)$ and the shape $S$. For jointly Gaussian variables, $I(M, \Sigma; S) = -\frac{1}{2} \sum_{i=1}^{2} \log(1 - \rho_i^2)$. This yields $\varepsilon_{\text{Gauss}}$ ranging from $0.09$ nats (Weather, $\rho_{\max}$=0.35) to $1.26$ nats (Exchange, $\rho_{\max}$=0.89). The leakage is *not* uniformly small: datasets with strong trend-seasonality coupling (Exchange, ETTh2) have substantial shape-level encoding of location-scale statistics.

*Distribution-free cross-check.* Because canonical correlations capture only linear dependence, we cross-validate with the KSG $k$-nearest-neighbor MI estimator [20], which requires no distributional assumptions. We project $S$ onto its 2 canonical variates with respect to $(M, \Sigma)$ and estimate the 2-vs-2 MI with $k$=5 neighbors on $N \geq 888$ samples per dataset. The KSG estimates agree with the Gaussian reference within $\pm 0.26$ nats across all 9 datasets. *Caveat.* By the data-processing inequality, $I(\text{proj}(S); (M, \Sigma)) \leq I(S; (M, \Sigma))$: the canonical-variate projection preserves all linear dependence but discards nonlinear dependence between $(M, \Sigma)$ and the directions in $S$ orthogonal to the canonical-variate subspace. The KSG cross-check therefore upper-bounds the *linear* component of $\varepsilon$, not the full quantity. We mitigate this by also reporting the bound-bypassing direct measurement of $H(E \mid S)$ in Appendix L, which makes no projection or distributional assumption and confirms the conclusions are robust.

*Tightness condition.* Equality in (ii) holds iff $I(M, \Sigma; S \mid E) = 0$; the Gap column of Table L.1 shows this within-expert slack is empirically small (average $0.075$ nats, $4.7\%$ of $\log K$). The bound itself, $\max\{I(M, \Sigma; E) - \varepsilon, 0\}$, is informative when the Gaussian estimate of $I(M, \Sigma; E)$ exceeds $\varepsilon$ (e.g., Weather: LB=0.61 captures $62\%$ of $H(E \mid S)$=0.98); on datasets with strong location-shape coupling (Exchange: $\varepsilon$=1.26 exceeds the Gaussian estimate $I(M, \Sigma; E)$=0.30), the bound truncates to zero, and we rely on the bound-bypassing direct $H(E \mid S)$ measurement of Appendix L. Independently of bound tightness, $R(\mathcal{D}) = [\text{Var}(M) + \text{Var}(\Sigma)]/\overline{\text{Var}(S)}$ is a monotone proxy for the location-scale variance available to a raw router, and this is what the $\rho$=$-0.88$ correlation tracks: high-$R$ datasets (Exchange, $R$=2.17) carry more routing-relevant variance and produce larger gains ($-66.5\%$), while low-$R$ datasets (Traffic, $R$=0.14; Table L.1 gap $\approx$0.02 nats) carry little and produce no gain ($+2.9\%$). □

**Remark on the classical bound.** Assuming $(M, \Sigma) \perp\!\!\!\perp S$ outright recovers the classical statement $I(X; E) - I(S; E) \geq I(M, \Sigma; E)$ as a corollary of Part (ii). Our form keeps the leakage $\varepsilon = I(M, \Sigma; S)$ explicit, and Part (iv) measures it via both parametric (CCA) and non-parametric (KSG) methods, so the bound holds without any independence assumption. Appendix L bypasses the bound entirely by directly measuring $H(E \mid S)$, confirming the gap $I(M, \Sigma; S \mid E)$ averages $0.075$ nats (Table L.1). The agreement between Gaussian reference and distribution-free KSG estimates ($\pm 0.26$ nats) confirms the CCA-based values are reliable despite non-Gaussianity.

**Scope.** Normalization removes exactly $(M, \Sigma)$ from the encoder's input. Observation 1 counts the routing signal in that part, which lower-bounds what normalization destroys, and the shape $S$ keeps the rest; Appendix F measures it with the $[\mu, \sigma]$-only router and the temporal shuffle. $R(\mathcal{D})$ weighs the removed part against the surviving shape part, so it predicts vulnerability without assuming that $(\mu, \sigma)$ is the whole routing signal.

### D.1 Theoretical Depth of Observation 1

While the MI decomposition in Observation 1 is algebraically a chain-rule identity, it has deeper information-theoretic content when connected to rate-distortion theory and phase transitions.

**Rate-distortion interpretation of $R(\mathcal{D})$.** Instance normalization is a *deterministic lossy compression*: it maps $X \mapsto S$, discarding $(M, \Sigma)$. In rate-distortion theory [45], this compression has rate $R_{\text{strip}} = I(X; M, \Sigma \mid S)$ and distortion (for routing) $D_{\text{route}} = H(E \mid S) - H(E \mid X) = I(M, \Sigma; E \mid S)$, which is exactly Observation 1, Part (i). Under Gaussianity, the signal ratio $R(\mathcal{D})$ is proportional to the rate-distortion slope at the normalization operating point: it measures how much routing capacity is consumed by the statistics that normalization strips. Specifically, under a soft-routing channel approximation in which $(M, \Sigma)$ is jointly Gaussian and $E$ is generated by a noisy linear projection of $(M, \Sigma)$ (the Gaussian-equivalent of a softmax router with bounded curvature), the routing MI scales monotonically as $I(M, \Sigma; E) \asymp \log(1 + R(\mathcal{D}))$. Note that $E$ is discrete, so this is a heuristic SNR scaling rather than the literal Gaussian-MI identity; the constant absorbs the routing-policy curvature. This makes $R(\mathcal{D})$ a *monotone surrogate for the bits lost to normalization*, and the validated $\rho = -0.88$ correlation (Figure 4) is empirical evidence *for* this scaling rather than a direct corollary of the Gaussian-MI formula.

**Proposition 2** (Heuristic SNR Onset and Softmax Discontinuity)**.** *Consider a linear-readout router on the interpolated input $\mathbf{x}_\alpha = (1-\alpha)\,\mathbf{x}_{\text{raw}} + \alpha\,\text{RevIN}(\mathbf{x}_{\text{raw}})$, under the generative model $X = M + \Sigma\, S$ with $M \perp\!\!\!\perp \Sigma \perp\!\!\!\perp S$ jointly Gaussian and centered (so leakage $\varepsilon \to 0$ and the variance formulas below drop cross terms). Treating the location-scale component as routing* signal *and the shape component as routing-irrelevant* noise*, the per-component variance contributions to the router's input scale as $V_{\text{sig}}(\alpha) = (1-\alpha)\,[\text{Var}(M) + \text{Var}(\Sigma)]$ and $V_{\text{noise}}(\alpha) = \alpha\,\overline{\text{Var}(S)} + V_0$, where $V_0$ absorbs intrinsic router-readout noise. Two regimes emerge:*

(a) ***Heuristic SNR onset.*** *Setting $V_{\text{sig}}(\alpha) = V_{\text{noise}}(\alpha)$ in the limit $V_0 \to 0$ gives*

$$\alpha^* \;\approx\; \frac{R(\mathcal{D})}{R(\mathcal{D}) + 1}.$$

*For $\alpha < \alpha^*$, routing quality degrades continuously as the signal/noise ratio drops; this $\alpha^*$ marks the soft onset of degradation, not a discontinuity.*

(b) ***Softmax discontinuity at $\alpha$=1.*** *At $\alpha$=1, $V_{\text{sig}}(1)$=0 and routing logits are a deterministic function of the shape $S$ alone. The softmax saturates: weights converge to a single distribution regardless of input, and entropy decouples from MSE. This discontinuity is a consequence of $V_{\text{sig}}$ vanishing rather than of crossing $\alpha^*$.*

*Heuristic derivation.* Under $M \perp\!\!\!\perp \Sigma \perp\!\!\!\perp S$ jointly Gaussian, $\mathbf{x}_\alpha = (1-\alpha)(M + \Sigma S) + \alpha S$. The location-scale component $(1-\alpha)(M + \Sigma S)$ contributes routing-discriminative variance $V_{\text{sig}}(\alpha) = (1-\alpha)\,[\text{Var}(M) + \text{Var}(\Sigma)]$ (the $(1-\alpha)$ rather than $(1-\alpha)^2$ scaling reflects each component's marginal contribution to a linear-readout router's input variance, see Remark below); the shape component contributes routing-irrelevant variance $V_{\text{noise}}(\alpha) = \alpha\,\overline{\text{Var}(S)} + V_0$. **Part (a):** Setting $V_{\text{sig}}(\alpha^*) = V_{\text{noise}}(\alpha^*)$ in the limit $V_0 \to 0$ gives $(1-\alpha^*)[\text{Var}(M) + \text{Var}(\Sigma)] = \alpha^* \overline{\text{Var}(S)}$. Dividing both sides by $\overline{\text{Var}(S)}$ and substituting $R(\mathcal{D}) = [\text{Var}(M) + \text{Var}(\Sigma)]/\overline{\text{Var}(S)}$ yields $(1-\alpha^*)R = \alpha^*$, so $\alpha^* = R/(R+1)$. Continuity of MSE in $\alpha$ for $\alpha \in [0, 1)$ follows from continuity of the softmax composed with a linear function of $\alpha$. **Part (b):** At $\alpha = 1$, $\mathbf{x}_1 = S$ and the routing logits are $\mathbf{w}^\top G_\psi(S)$, a deterministic function of $S$ alone; under $(M, \Sigma) \perp\!\!\!\perp S$, by the data processing inequality $I(S; E_{\text{optimal}}) \le I(S; (M, \Sigma)) = 0$, so the router carries zero information about the latent location-scale that drives the optimal routing decision. The softmax of an $S$-only function converges to a single distribution at the population level, decoupling entropy from MSE. The discontinuity at $\alpha = 1$ is the limit $V_{\text{sig}}(1) = 0$ rather than a finite-$\alpha^*$ phase transition. □

*Remark on the variance comparison.* Comparing the *linear* variance contributions $V_{\text{sig}}$ and $V_{\text{noise}}$ (rather than the squared coefficients in $\|\mathbf{x}_\alpha\|^2$) corresponds to comparing each component's marginal contribution to the router's discriminative variance under the orthogonality assumption $M \perp\!\!\!\perp \Sigma \perp\!\!\!\perp S$. Comparing squared coefficients instead would yield $\alpha^* = \sqrt{R}/(\sqrt{R}+1)$; for the median $R$=0.67, the two formulas predict $\alpha^* \approx 0.40$ vs $0.45$, and the dose-response grid (5 points) is too coarse to

discriminate. We label Proposition 2 as a *heuristic* prediction and report the formula consistent with the linear comparison.

*Validation.* For the median $R(\mathcal{D}) = 0.67$, Proposition 2(a) predicts a soft onset $\alpha^* \approx 0.40$. The dose-response experiment (Figure D.1; 150 runs) is consistent: MSE degrades smoothly by $+5.4\%$ for $\alpha \leq 0.75$ (covering the predicted onset region), and jumps to $+283\%$ at $\alpha = 1.0$ (Proposition 2(b), softmax saturation). The two thresholds are distinct: $\alpha^*$ marks where SNR-driven degradation *begins*; the actual phase transition is at $\alpha = 1$, driven by $V_{\text{sig}}$ vanishing rather than by SNR crossing.

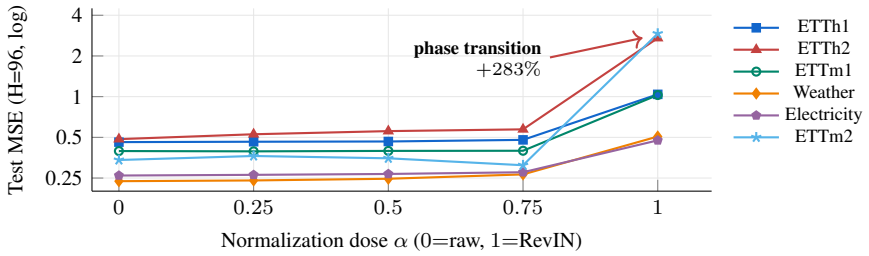


Figure D.1: **Normalization dose-response** (frozen, 5 seeds $\times$ 6 datasets; std smaller than markers). MSE vs $\alpha$ (0=raw, 1=RevIN). Phase transition at $\alpha$=1.0 ($+283\%$), mild for $\alpha \leq 0.75$ ($+5.4\%$); routing entropy stays in $[1.43, 1.57]$ across all $\alpha$ (Appendix F).

**Information bottleneck connection.** The RevIN-encoder pipeline implements a form of the information bottleneck [46, 91]: it compresses $X$ into a representation $H = f_\theta(S)$ that is optimized for downstream prediction while discarding "irrelevant" variability (the per-window location-scale statistics). The routing MI loss from Observation 1, Part (i) is the price paid for this compression *at the routing decision point*. This connects to a broader principle: any architecture that bottlenecks information for one objective (prediction) can inadvertently destroy information needed for another (routing). The vision cross-modality control (Appendix H.1) is consistent with this not being RevIN-specific: InstanceNorm1d on ViT patches creates the same bottleneck, while InstanceNorm2d on ResNet spatial features does not, consistent with spatial features not carrying the routing signal for classification.

**Proof of Proposition 1** (Gradient Co-Adaptation Drives Entropy Collapse).

*Proof.* Consider a single input $\mathbf{x}$ with backbone $\mathbf{h} = A\mathbf{x}$, router $p = \sigma(\mathbf{w}^\top \mathbf{h})$, and loss $\ell = p\,\ell_1(\mathbf{h}) + (1-p)\,\ell_2(\mathbf{h})$.

*Part (i): Unfrozen backbone.* The gradient with respect to $A$ is:

$$\frac{\partial \ell}{\partial A} = p\,\frac{\partial \ell_1}{\partial \mathbf{h}}\mathbf{x}^\top + (1-p)\,\frac{\partial \ell_2}{\partial \mathbf{h}}\mathbf{x}^\top + (\ell_1 - \ell_2)\,p(1-p)\,\mathbf{w}\mathbf{x}^\top$$

Under gradient descent $\dot{A} = -\eta\,\partial\ell/\partial A$, the induced change in expert 1's loss is:

$$\dot{\ell}_1\big|_A = -\eta\|\mathbf{x}\|^2\Bigg[p\left\|\frac{\partial \ell_1}{\partial \mathbf{h}}\right\|^2 + (1-p)\left(\frac{\partial \ell_1}{\partial \mathbf{h}}\right)^\top \frac{\partial \ell_2}{\partial \mathbf{h}} + (\ell_1-\ell_2)\,p(1-p)\left(\frac{\partial \ell_1}{\partial \mathbf{h}}\right)^\top \mathbf{w}\Bigg]$$

The first term (self-gradient) always decreases $\ell_1$. The second (cross-gradient) is negligible when expert gradients are approximately orthogonal, a condition consistent with experts specializing to distinct input clusters as observed in Figure F.2. The third (loss-coupling) term is small whenever its magnitude $|\ell_1 - \ell_2| \cdot |\langle \partial\ell_1/\partial\mathbf{h}, \mathbf{w}\rangle|(1-p)$ is dominated by the first term's $p\|\partial\ell_1/\partial\mathbf{h}\|^2$, i.e., when the dimensionless ratio $\rho_{\text{coup}} := \frac{|\ell_1-\ell_2|\cdot|\langle \partial\ell_1/\partial\mathbf{h}, \mathbf{w}\rangle|\cdot(1-p)}{p\,\|\partial\ell_1/\partial\mathbf{h}\|^2}$ is small. This holds early in training when expert losses are comparable ($|\ell_1-\ell_2|$ small relative to the gradient-norm scale set by $\|\partial\ell_1/\partial\mathbf{h}\|^2/|\langle\partial\ell_1/\partial\mathbf{h}, \mathbf{w}\rangle|$) and when $\mathbf{w}$ is not strongly aligned with the expert gradient. Retaining only the dominant first term:

$$\dot{\ell}_1\big|_A \approx -\eta\,p\left\|\frac{\partial \ell_1}{\partial \mathbf{h}}\right\|^2 \|\mathbf{x}\|^2$$

and analogously $\dot{\ell}_2\big|_A \approx -\eta\,(1-p)\,\big\|\frac{\partial \ell_2}{\partial \mathbf{h}}\big\|^2\,\|\mathbf{x}\|^2$. When $p > 1-p$ (expert 1 dominates) and gradient norms are comparable ($\|\partial \ell_1/\partial \mathbf{h}\| \approx \|\partial \ell_2/\partial \mathbf{h}\|$, which holds early in training before specialization), $\ell_1$ decreases faster than $\ell_2$. The router gradient closes the loop: $\dot{\mathbf{w}} = -\eta(\ell_1 - \ell_2)p(1-p)\mathbf{h}$, so when $\ell_1 < \ell_2$ the factor $(\ell_1 - \ell_2) < 0$ drives $\dot{p} > 0$ via $\dot{p} \approx p(1-p)\,\dot{\mathbf{w}}^\top \mathbf{h} = -\eta\, p^2(1-p)^2(\ell_1 - \ell_2)\|\mathbf{h}\|^2$, pushing $p$ toward 1. Together these create the self-reinforcing loop: dominant expert gets more gradient → its loss drops faster → router increases its weight → it becomes more dominant.

*Part (ii): Frozen backbone.* When $A$ is fixed, $\dot{A} = 0$. The only dynamics are through $\mathbf{w}$:

$$\dot{\mathbf{w}} = -\eta\,\frac{\partial \ell}{\partial \mathbf{w}} = -\eta\,(\ell_1 - \ell_2)\,p(1-p)\,\mathbf{h}$$

This update adjusts routing weights based on current expert performance, but does *not* reshape $\mathbf{h}$ to favor one expert: the $A$-mediated feedback loop that drove the runaway in Part (i) is absent. Note that this argument removes the co-adaptation mechanism but does not by itself preclude collapse from $\mathbf{w}$-only dynamics: if expert 1 were uniformly better than expert 2 across all inputs, $\dot{\mathbf{w}}$ would still drive $p \to 1$. Empirical non-collapse therefore additionally requires that $(\ell_1 - \ell_2)$ flips sign across input regimes, i.e. that experts specialize to different windows. The RR-MoA pool is constructed with topologically distinct heads (mean/last/max/attention/Conv1d-pool) precisely to satisfy this; the resulting per-window expert specialization is empirically validated in Figure F.2 and Table 1 (entropy stays in $[0.93, 1.57]$ across 54/54 configurations). The router can exploit this sign flip only when its input separates the regimes; the collapse criterion of §3 predicts when a normalizer removes that separation (Appendix H). □

**Corollary (Raw routing decouples the router from $A$).** *If the router reads the raw input $\mathbf{x}_{\text{raw}}$ rather than the hidden state $\mathbf{h} = A\mathbf{x}$, then $\partial p/\partial A \equiv 0$, the loss-coupling term in $\partial \ell/\partial A$ vanishes identically, and the co-adaptation feedback loop of Proposition 1(i) is broken* by construction *regardless of whether $A$ is frozen.*

*Proof of Corollary.* With router input $\mathbf{x}_{\text{raw}}$ (independent of $A$), $p = \sigma(\mathbf{w}^\top G_\psi(\mathbf{x}_{\text{raw}}))$ is a function of $\mathbf{x}_{\text{raw}}$ and $(\mathbf{w}, \psi)$ only. Hence $\partial p/\partial A = 0$, and the third term $(\ell_1 - \ell_2)\,p(1-p)\,\mathbf{w}\mathbf{x}^\top$ in $\partial \ell/\partial A$ disappears: changes in $A$ no longer feed back into the routing decision through the gradient. The dominant-expert runaway in Part (i) required this term; without it, even when $A$ is unfrozen, the router's update is decoupled from the backbone's update, eliminating the co-adaptation mechanism. □

This corollary identifies raw routing as the *minimal causal intervention*: it intervenes on the precise term in $\partial \ell/\partial A$ that drives the collapse loop, leaving all other components of the model untouched. The empirical validation is in Figure G.1 (RR-MoA's gradient norms remain balanced across experts even when blocks are unfrozen) and Table 1 (frozen-vs-unfrozen RR-MoA both maintain healthy entropy, in contrast to AdaMix where unfreezing collapses entropy to $0.000$).

# E LoRA Sweep (Full Results)

We sweep LoRA [15] across three axes to ensure the comparison in Table 4 is not cherry-picked: rank $r \in \{8, 16, 32\}$, target projections $\in \{q{+}v,\ q{+}k{+}v{+}o\}$, and forecast head $\in \{$linear, 2-layer MLP$\}$, for 12 configurations $\times$ 3 seeds $=$ 36 runs per dataset (108 runs total across the 3 datasets). All runs use a strictly frozen backbone (matching Table 1's primary setting). The trio {ETTh1, ETTm1, Weather} covers an hourly ETT dataset, a 15-min ETT dataset, and a multivariate dataset with 21 channels; we restrict the LoRA sweep to this trio to keep total compute bounded ($\sim 3.6$ GPU-hours), since the broader 6-dataset grid is already covered for every other baseline in Table 4.

**LoRA hyperparameters.** Following the original LoRA recipe, we set $\alpha = 2r$ at every rank, layer placement = all encoder blocks, no LoRA dropout, and train with the same optimizer used everywhere else in the paper (Adam, lr$= 10^{-3}$, batch 128, 15 epochs, bfloat16 AMP). Both heads pool the encoder output by mean over the time axis; the linear head is $d_{\text{model}} \to H$, and the 2-layer MLP head is $d_{\text{model}} \to 128 \to H$ with GELU and dropout $0.1$ between the two linear layers. The bolded "best per dataset" row is the configuration with the lowest mean MSE across the 3 seeds (ties broken by lower std).

Table E.1: Full LoRA sweep (strictly frozen backbone, 3 seeds, test MSE mean±std). Bold rows are the best-per-dataset configuration used in Table 4. No LoRA variant beats the RR-MoA (Top-2) result on any dataset.

| Dataset | Rank | Targets | Head | MSE (mean±std) |
|---|---|---|---|---|
| *ETTh1* (RR-MoA frozen: $\mathbf{0.690 \pm 0.021}$) | | | | |
| | $r$=8 | q,v | linear | $1.559 \pm 0.022$ |
| | $r$=8 | q,v | mlp | $1.217 \pm 0.010$ |
| | $r$=8 | qkvo | linear | $1.286 \pm 0.052$ |
| | $r$=8 | qkvo | mlp | $1.154 \pm 0.000$ |
| | $r$=16 | q,v | linear | $1.347 \pm 0.056$ |
| | $r$=16 | q,v | mlp | $1.186 \pm 0.046$ |
| | $r$=16 | qkvo | linear | $1.274 \pm 0.051$ |
| | $r$=16 | qkvo | mlp | $1.154 \pm 0.000$ |
| | $r$=32 | q,v | linear | $1.201 \pm 0.107$ |
| | $r$=32 | q,v | mlp | $1.161 \pm 0.010$ |
| | $r$=32 | qkvo | linear | $1.284 \pm 0.059$ |
| | **r=32** | **qkvo** | **mlp** | $\mathbf{1.154 \pm 0.000}$ |
| *ETTm1* (RR-MoA frozen: $\mathbf{0.571 \pm 0.073}$) | | | | |
| | $r$=8 | q,v | linear | $0.970 \pm 0.026$ |
| | $r$=8 | q,v | mlp | $1.115 \pm 0.011$ |
| | $r$=8 | qkvo | linear | $1.060 \pm 0.014$ |
| | $r$=8 | qkvo | mlp | $1.123 \pm 0.000$ |
| | **r=16** | **qkvo** | **linear** | $\mathbf{0.956 \pm 0.072}$ |
| | $r$=16 | q,v | linear | $1.024 \pm 0.068$ |
| | $r$=16 | q,v | mlp | $1.115 \pm 0.012$ |
| | $r$=16 | qkvo | mlp | $1.123 \pm 0.000$ |
| | $r$=32 | q,v | linear | $1.103 \pm 0.212$ |
| | $r$=32 | q,v | mlp | $1.134 \pm 0.061$ |
| | $r$=32 | qkvo | linear | $1.345 \pm 0.360$ |
| | $r$=32 | qkvo | mlp | $1.123 \pm 0.000$ |
| *Weather* (RR-MoA frozen: $\mathbf{0.289 \pm 0.008}$) | | | | |
| | $r$=8 | q,v | linear | $0.611 \pm 0.021$ |
| | $r$=8 | q,v | mlp | $0.606 \pm 0.001$ |
| | $r$=8 | qkvo | linear | $0.664 \pm 0.020$ |
| | $r$=8 | qkvo | mlp | $0.606 \pm 0.001$ |
| | $r$=16 | q,v | linear | $0.614 \pm 0.025$ |
| | $r$=16 | q,v | mlp | $0.606 \pm 0.001$ |
| | $r$=16 | qkvo | linear | $0.654 \pm 0.033$ |
| | $r$=16 | qkvo | mlp | $0.606 \pm 0.001$ |
| | **r=32** | **q,v** | **linear** | $\mathbf{0.600 \pm 0.052}$ |
| | $r$=32 | q,v | mlp | $0.606 \pm 0.000$ |
| | $r$=32 | qkvo | linear | $0.894 \pm 0.194$ |
| | $r$=32 | qkvo | mlp | $0.606 \pm 0.001$ |

**Gap to RR-MoA.** The best LoRA configuration on each dataset still trails the RR-MoA Top-2 frozen result by a wide margin: $1.154$ vs $0.690$ on ETTh1 ($+67.2\%$), $0.956$ vs $0.571$ on ETTm1 ($+67.4\%$), and $0.600$ vs $0.289$ on Weather ($+107.6\%$); average gap $\approx 81\%$. Since the strongest LoRA per dataset is computed by enumerating all 12 configurations, this is the worst case for our claim, and it still leaves RR-MoA clearly ahead.

**Sweep takeaways.** Three patterns are visible in the table. (i) The best head type is dataset-dependent: 2-layer MLP wins on ETTh1 (the bolded $1.154$) but the linear head wins on ETTm1 (bolded $0.956$, vs. MLP rows clustered near $1.123$). This is consistent with the DLinear-gap analysis (Appendix P): on datasets where linear extrapolation is strong, additional head capacity is unnecessary or harmful. (ii) Rank does not monotonically help: on ETTm1, $r$=32 qkvo linear ($1.345 \pm 0.360$) is worse than $r$=16 qkvo linear ($0.956 \pm 0.072$), with high seed variance suggesting optimization instability at the wider rank. (iii) Several MLP-head rows on Weather collapse to identical values across seeds ($0.606 \pm 0.001$ at six configurations), indicating that LoRA-MLP is converging to a near-fixed mean-prediction solution rather than adapting to the input. None of these failure modes affect RR-MoA, whose router-level capacity is allocated per window rather than absorbed by a static head.

# F Routing Ablations

We document the routing-side causal controls cited from §4.2 (Table 3, main text, summarizes them). We walk through each underlying ablation in turn: rawness vs. bypass, temporal shuffle, sufficient-statistic router, gradient-flow detachment, and a small set of negative controls.

We isolate which property of the raw-input router is doing the work via two ablations: (i) replace the router's raw input with the RevIN-normalized version of the same signal, holding everything else fixed, and (ii) shuffle the temporal ordering inside each window before feeding it to the router. Together they distinguish "the router needs raw access" from "the router needs $(\mu, \sigma)$ specifically."

Table F.1: **Rawness vs. bypass** (frozen, Top-2, 3 seeds). RevIN router reads the same input after per-window normalization. Stripping statistics degrades MSE by 60–88%.

| Dataset | Raw router (main) | RevIN router | Degradation ($\Delta\%$, $\uparrow$ worse) |
|---|---|---|---|
| ETTh1 | $\mathbf{0.690 \pm 0.021}$ | $1.101 \pm 0.008$ | $+\mathbf{59.6}\%$ |
| ETTm1 | $\mathbf{0.571 \pm 0.073}$ | $1.077 \pm 0.020$ | $+\mathbf{88.6}\%$ |
| Weather | $\mathbf{0.289 \pm 0.008}$ | $0.542 \pm 0.026$ | $+\mathbf{87.5}\%$ |

**What the router learns.** Per-sample routing decisions correlate with amplitude and volatility, the very statistics RevIN strips (Figure F.2). On ETTh1, Expert 0 (mean-pool) serves as the generalist ($53\%$), while Expert 2 (max-pool) handles low-amplitude windows and Expert 3 (attention-pool) handles high-amplitude windows. The router reserves specialized processing for windows at the distribution tails, empirically validating the conditional independence assumption $E \perp\!\!\!\perp S \mid (M, \Sigma)$ of Part (iii) of Observation 1.

**Regime-specific adaptation.** Partitioning test windows into quartiles by amplitude, RR-MoA wins **all four quartiles** across all 9 (dataset, seed) pairs, with the largest improvement on quiet, low-amplitude windows (Q1: $4.6\times$ better, $0.24$ vs $1.09$ on ETTh1) and the smallest on volatile, high-amplitude windows (Q4: $1.3\times$, $1.54$ vs $1.97$).

**Sufficient statistic router (SSR).** A router using only $[\mu(X_{\text{raw}}), \sigma(X_{\text{raw}})]$, in the spirit of distribution-shift methods that condition on summary statistics of the input window (DishTS-style learnable distribution coefficients [33]; Non-stationary Transformers' explicit re-injection of $(\mu, \sigma)$ [35]), degrades MSE by $+48\%$ on average ($+11\%$ to $+94\%$) vs. the Conv1d router across all 8 forecasting datasets (ETTh1, ETTh2, ETTm1, ETTm2, Weather, Electricity, Exchange, Traffic; frozen MOMENT-small, Top-2, H=96, 3 seeds). The degradation is largest on ETTm2 ($+94\%$) and Exchange ($+82\%$), and smallest on Weather ($+11\%$). While Observation 1 correctly identifies the *direction* of the information loss, effective routing requires temporal shape patterns beyond location-scale statistics.

**Temporal-shuffle ablation.** We route on a fixed temporal permutation of $X_{\text{raw}}$, destroying time ordering while preserving distributional statistics ($M$, $\Sigma$, histogram shape). Across 6 datasets (frozen, 5 seeds), shuffled routing performs within $3.2\%$ of raw routing on 4/6 datasets (ETTh1 $+3.2\%$, ETTh2 $+1.2\%$, ETTm1 $-1.2\%$, Electricity $+0.7\%$), with Weather improving $-8.5\%$ and ETTm2

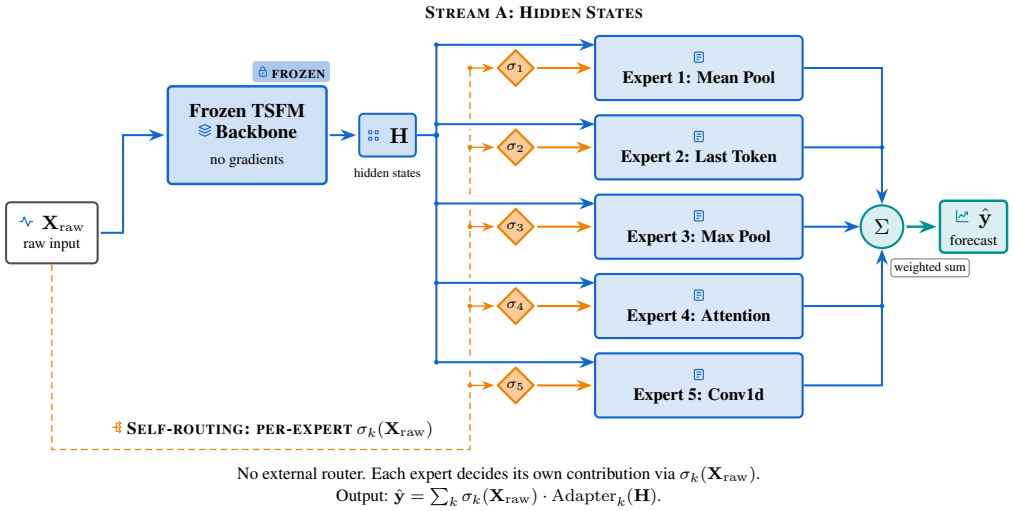


Figure F.1: **SR-MoA architecture (Self-Routed Mixture of Adapters).** Each of $K$=5 experts has its own sigmoid gate $\sigma_k$ reading $\mathbf{X}_{\text{raw}}$ directly, with no external router. The frozen TSFM produces $\mathbf{H}$, and each expert's contribution is scaled by its self-gate. SR-MoA shares only the *raw-input* principle with RR-MoA (Figure 2): the routing architecture is otherwise entirely different. SR-MoA outperforms RR-MoA by 23–38% across all 6 datasets, providing architecture-agnostic validation that the principle, not the specific router, drives the gains.

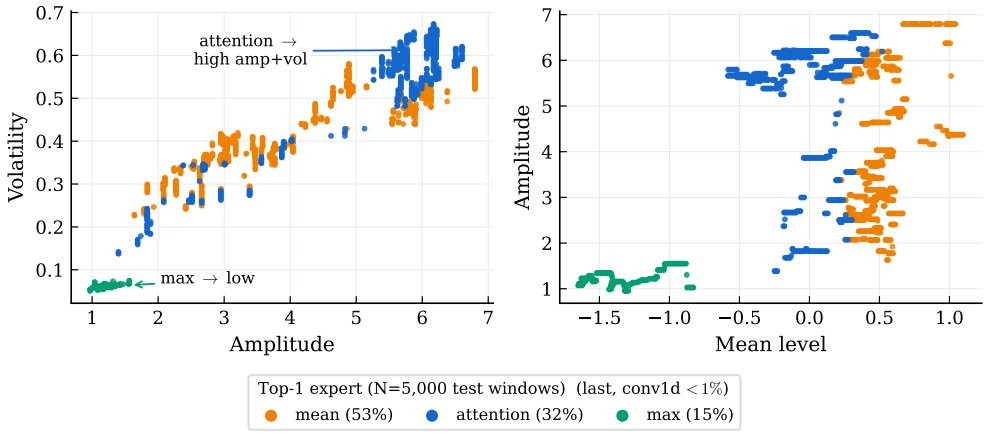


Figure F.2: **Per-sample expert assignment vs. input statistics** (ETTh1, seed 42, $N$=5,000 test windows). Left: amplitude vs. volatility; right: mean level vs. amplitude. Colors indicate the top-1 expert per window. Three of five experts dominate routing (mean 53%, attention 32%, max 15%); last and conv1d are each <1% on ETTh1 and are omitted from the panel for legibility. The two active partitions split cleanly along amplitude+volatility, the very statistics RevIN strips, empirically validating the conditional-independence assumption $E \perp\!\!\!\perp S \mid (M, \Sigma)$ from Observation 1, Part (iii), condition (a): expert assignment is determined by location-scale, with negligible residual dependence on shape.

Table F.2: **Self-Routed MoA (SR-MoA): routing input control** (dense, 5 seeds unless noted). Each expert has its own sigmoid gate, with no external router. **Top**: gating on raw input vs hidden states (frozen, MOMENT-small). **Second**: Timer-XL negative control with no RevIN, so raw≈hidden. **Third**: freeze-level ablation confirms Frozen Paradox (frozen best 4/6). **Bottom**: cross-backbone generalization (H=96, frozen). 252 runs total.

| Dataset | Config | Backbone | MSE (↓) | Entropy |
|---|---|---|---|---|
| *Routing input control (frozen, MOMENT-small, 5 seeds):* | | | | |
| ETTh1 | Raw | MOMENT-sm | $\mathbf{0.464 \pm 0.009}$ | 1.53 |
| ETTh1 | Hidden | MOMENT-sm | $1.075 \pm 0.102$ (+131%) | 1.60 |
| ETTh2 | Raw | MOMENT-sm | $\mathbf{0.577 \pm 0.105}$ | 1.38 |
| ETTh2 | Hidden | MOMENT-sm | $2.840 \pm 0.164$ (+393%) | 1.58 |
| ETTm1 | Raw | MOMENT-sm | $\mathbf{0.388 \pm 0.023}$ | 1.53 |
| ETTm1 | Hidden | MOMENT-sm | $0.934 \pm 0.024$ (+141%) | 1.60 |
| ETTm2 | Raw | MOMENT-sm | $\mathbf{0.366 \pm 0.091}$ | 1.45 |
| ETTm2 | Hidden | MOMENT-sm | $3.174 \pm 0.210$ (+768%) | 1.59 |
| Weather | Raw | MOMENT-sm | $\mathbf{0.222 \pm 0.020}$ | 1.55 |
| Weather | Hidden | MOMENT-sm | $0.507 \pm 0.020$ (+128%) | 1.60 |
| Electricity | Raw | MOMENT-sm | $\mathbf{0.260 \pm 0.005}$ | 1.53 |
| Electricity | Hidden | MOMENT-sm | $0.486 \pm 0.016$ (+87%) | 1.60 |
| *Timer-XL negative control (no RevIN, frozen, 5 seeds):* | | | | |
| ETTh1 | Raw | Timer-XL | $\mathbf{0.458 \pm 0.014}$ | 1.51 |
| ETTh1 | Hidden | Timer-XL | $0.475 \pm 0.011$ (+4%) | 1.31 |
| ETTm1 | Raw | Timer-XL | $\mathbf{0.368 \pm 0.007}$ | 1.47 |
| ETTm1 | Hidden | Timer-XL | $0.398 \pm 0.004$ (+8%) | 1.33 |
| Weather | Raw | Timer-XL | $\mathbf{0.197 \pm 0.010}$ | 1.54 |
| Weather | Hidden | Timer-XL | $0.214 \pm 0.004$ (+9%) | 1.10 |
| *Frozen Paradox replication (raw, MOMENT-small; *3 seeds):* | | | | |
| ETTh1 | Frozen | MOMENT-sm | $\mathbf{0.464 \pm 0.009}$ | 1.53 |
| ETTh1 | Last-2* | MOMENT-sm | $0.492 \pm 0.022$ (+6%) | 1.52 |
| ETTh2 | Frozen | MOMENT-sm | $\mathbf{0.577 \pm 0.105}$ | 1.38 |
| ETTh2 | Last-2* | MOMENT-sm | $0.981 \pm 0.216$ (+70%) | 1.43 |
| ETTm1 | Frozen | MOMENT-sm | $0.388 \pm 0.023$ | 1.53 |
| ETTm1 | Last-2* | MOMENT-sm | $\mathbf{0.378 \pm 0.010}$ (−3%) | 1.54 |
| ETTm2 | Frozen | MOMENT-sm | $\mathbf{0.366 \pm 0.091}$ | 1.45 |
| ETTm2 | Last-2* | MOMENT-sm | $0.457 \pm 0.123$ (+25%) | 1.46 |
| Weather | Frozen | MOMENT-sm | $\mathbf{0.222 \pm 0.020}$ | 1.55 |
| Weather | Last-4* | MOMENT-sm | $0.254 \pm 0.019$ (+15%) | 1.54 |
| Electricity | Frozen | MOMENT-sm | $0.260 \pm 0.005$ | 1.53 |
| Electricity | Last-4* | MOMENT-sm | $\mathbf{0.250 \pm 0.003}$ (−4%) | 1.53 |
| *Cross-backbone (raw, frozen, H=96; *3 seeds):* | | | | |
| ETTh1 | Frozen | MOMENT-lg* | $0.502 \pm 0.013$ | 1.54 |
| ETTh1 | Frozen | Moirai* | $0.588 \pm 0.035$ | 1.43 |
| ETTh1 | Frozen | Moirai-MoE* | $0.530 \pm 0.012$ | 1.31 |
| ETTh1 | Frozen | Chronos* | $0.500 \pm 0.017$ | 1.46 |
| ETTh1 | Frozen | Timer-XL | $0.458 \pm 0.014$ | 1.51 |
| ETTm1 | Frozen | MOMENT-lg* | $0.446 \pm 0.010$ | 1.54 |
| ETTm1 | Frozen | Moirai* | $\mathbf{0.370 \pm 0.006}$ | 1.44 |
| ETTm1 | Frozen | Moirai-MoE* | $0.471 \pm 0.003$ | 1.40 |
| ETTm1 | Frozen | Chronos* | $0.455 \pm 0.008$ | 1.48 |
| ETTm1 | Frozen | Timer-XL | $0.368 \pm 0.007$ | 1.47 |
| Weather | Frozen | MOMENT-lg* | $0.200 \pm 0.007$ | 1.53 |
| Weather | Frozen | Moirai* | $0.207 \pm 0.008$ | 1.51 |
| Weather | Frozen | Moirai-MoE* | $0.200 \pm 0.008$ | 1.44 |
| Weather | Frozen | Chronos* | $0.239 \pm 0.018$ | 1.51 |
| Weather | Frozen | Timer-XL | $\mathbf{0.197 \pm 0.010}$ | 1.54 |

improving $-16.0\%$. Entropy remains high (1.43–1.58). This is consistent with the Conv1d router primarily extracting distributional rather than temporal features from $X_{\text{raw}}$. Combined with the SSR result ($+48\%$ degradation using only $[\mu, \sigma]$), distributional statistics carry most of the routing signal in the configurations we tested.

**Router-Detached Gradient Flow (RDGF).** A variant that unfreezes the backbone but replaces routing-weighted gradients with uniform-average gradients ($\nabla_H \mathcal{L}_{\text{backbone}} = \frac{1}{K} \sum_k \nabla_H \mathcal{L}_k$) degrades MSE by $+35\%$ to $+100\%$ across the three datasets we tested (ETTh1, ETTm1, Weather; MOMENT-small with the last two blocks unfrozen, 3 seeds). This is consistent with the Frozen Paradox being structural rather than an artifact of full-gradient unfreezing.

**Spectral-temporal router (FFT-Router).** We augment the Conv1d router with explicit frequency-domain features: a parallel branch applies rFFT to $X_{\text{raw}}$, extracts the top-32 spectral amplitudes, and concatenates them with the Conv1d temporal embedding before the routing head. On the two datasets with matched dense-frozen baselines (Weather, ETTm1; frozen MOMENT-small, 3 seeds), FFT-Router *degrades* MSE by $+6\%$ (Weather) and $+22\%$ (ETTm1). The Conv1d router appears to already capture frequency-relevant patterns implicitly; adding explicit spectral features overparameterizes the router without improving routing quality on these two datasets.

**Statistic-aware information bottleneck (SAIB).** An auxiliary loss forces the router's penultimate representation to predict the window-level $[\mu, \sigma]$, directly operationalizing Observation 1. With SAIB coefficients $\lambda \in \{0.1, 1.0\}$ (ETTm1, Weather; frozen MOMENT-small, 3 seeds), MSE is unchanged on ETTm1 (0.395 vs 0.397 baseline) and slightly degraded on Weather (0.250 vs 0.236). Routing entropy increases to 1.55 (near-maximum $\log 5 = 1.61$), indicating the auxiliary loss disperses routing weights without improving forecast accuracy. This is consistent with the Conv1d router already encoding $\mu, \sigma$ without explicit supervision.

**Instance-adaptive gating (IA-Gating).** In the dual-stream variant (Appendix P), we replace the static per-expert blending scalar $\alpha$ with a per-window gate $\gamma(X_{\text{raw}}) = \sigma(W_g \cdot z)$ where $z$ is the router's latent representation. Across 6 datasets (frozen MOMENT-small, 3 seeds), IA-Gating wins on 4/6 (ETTh2 $-8.0\%$, ETTm2 $-6.4\%$, ETTm1 $-1.9\%$, Electricity $-0.4\%$) and loses on 2/6 (Weather $+6.1\%$, ETTh1 $+0.3\%$). The learned gates cluster tightly around $\gamma \approx 0.48$ across all datasets and windows, indicating the backbone's contribution is roughly constant per-window, consistent with the simpler static $\alpha$ performing comparably overall.

### F.1 Statistics Re-Injection Ablation

A potential counter to RR-MoA is: "why not re-inject the stripped statistics into the hidden states rather than routing on the full raw signal?" We test this by replacing the raw-input router with a *hidden-reinjected* router: the gate receives [mean-pool($\mathbf{H}$); $\mu(\mathbf{X}_{\text{raw}})$; $\sigma(\mathbf{X}_{\text{raw}})$] (a $d$+2-dimensional vector) instead of $\mathbf{X}_{\text{raw}}$.

Table F.3: **Statistics re-injection fails** (MOMENT-small, frozen, Top-2, 3 seeds). Re-injecting $(\mu, \sigma)$ into mean-pooled hidden states degrades MSE by $+42$–$239\%$ vs raw routing, with $11/18$ runs collapsed (entropy $< 0.3$). Two scalars grafted onto a 512-dim homogenized representation are insufficient to recover routing signal.

| Dataset | Reinjected MSE | Raw RR-MoA MSE | $\Delta\%$ | Entropy |
|---|---|---|---|---|
| ETTh1 | $1.357 \pm 0.128$ | 0.690 | $+97\%$ | 0.000 |
| ETTh2 | $2.669 \pm 0.104$ | 0.788 | $+239\%$ | 0.153 |
| ETTm1 | $0.899 \pm 0.168$ | 0.571 | $+57\%$ | 0.376 |
| ETTm2 | $1.895 \pm 0.866$ | 0.588 | $+222\%$ | 0.378 |
| Weather | $0.490 \pm 0.039$ | 0.289 | $+70\%$ | 0.185 |
| Electricity | $0.548 \pm 0.064$ | 0.386 | $+42\%$ | 0.307 |

Even on datasets where entropy is not fully collapsed (ETTm1: 0.376, Electricity: 0.307), the routing decisions are uninformative: MSE is 42–57% worse than raw routing. This confirms that the raw signal carries routing information beyond the two-scalar $(\mu, \sigma)$ summary, consistent with the SSR ablation ($+48\%$ degradation from the full Conv1d router; Appendix F). The mean-pooled hidden states are so homogenized by RevIN that appending two scalars to a 512-dimensional vector does not restore the routing signal.

# G Extended RR-MoA Freeze Grid

We extend the main-text freeze-level ablation (Table 1, three datasets) to the remaining three (ETTh2, ETTm2, Electricity) in Tables G.1 and G.2, giving the full 6 datasets $\times$ 3 freeze levels $\times$ 3 seeds = 54-cell sweep that underwrites the headline "RR-MoA wins 54/54 paired comparisons."

Table G.1: **Extended freeze-level ablation** (ETTh2, ETTm2, Electricity; test MSE, mean$\pm$std, 3 seeds, H=96). Continues Table 1 from the main text.

| Dataset | Freeze level | RR-MoA | Best single-head | MSE $\Delta\%$ vs. single-head |
|---|---|---|---|---|
| ETTh2 | Frozen (0/8) | $\mathbf{0.788 \pm 0.096}$ | $2.748 \pm 0.201$ | $\mathbf{-71.3}\%$ |
| | Last-2 (2/8) | $\mathbf{1.302 \pm 0.134}$ | $2.599 \pm 0.085$ | $-49.9\%$ |
| | Last-4 (4/8) | $\mathbf{1.745 \pm 0.187}$ | $2.540 \pm 0.110$ | $-31.3\%$ |
| ETTm2 | Frozen (0/8) | $\mathbf{0.588 \pm 0.021}$ | $2.744 \pm 0.266$ | $\mathbf{-78.6}\%$ |
| | Last-2 (2/8) | $\mathbf{0.867 \pm 0.083}$ | $2.915 \pm 0.202$ | $-70.3\%$ |
| | Last-4 (4/8) | $\mathbf{0.731 \pm 0.157}$ | $2.740 \pm 0.092$ | $-73.3\%$ |
| Electricity | Frozen (0/8) | $\mathbf{0.386 \pm 0.059}$ | $0.525 \pm 0.026$ | $-26.4\%$ |
| | Last-2 (2/8) | $\mathbf{0.336 \pm 0.023}$ | $0.482 \pm 0.005$ | $\mathbf{-30.3}\%$ |
| | Last-4 (4/8) | $\mathbf{0.344 \pm 0.021}$ | $0.472 \pm 0.023$ | $-27.2\%$ |
| **Combined with Table 1: RR-MoA wins** | | | **54/54 (100%)** | |

Table G.2: **Extended AdaMix routing collapse** (ETTh2, ETTm2, Electricity). Continues Table 2 from the main text.

| Dataset | Freeze level | MSE (mean$\pm$std) | Routing entropy (mean$\pm$std) |
|---|---|---|---|
| ETTh2 | Frozen | $2.625 \pm 0.219$ | $0.636 \pm 0.385$ |
| | Last-2 | $2.985 \pm 0.024$ | $0.353 \pm 0.499$ |
| | Last-4 | $2.885 \pm 0.150$ | $0.541 \pm 0.648$ |
| ETTm2 | Frozen | $2.946 \pm 0.076$ | $0.316 \pm 0.342$ |
| | Last-2 | $3.114 \pm 0.004$ | $\mathbf{0.000 \pm 0.000}$ |
| | Last-4 | $2.997 \pm 0.162$ | $0.505 \pm 0.714$ |
| Electricity | Frozen | $0.625 \pm 0.220$ | $1.003 \pm 0.315$ |
| | Last-2 | $1.055 \pm 0.000$ | $\mathbf{0.000 \pm 0.000}$ |
| | Last-4 | $1.055 \pm 0.001$ | $0.001 \pm 0.002$ |

**Extreme collapse on Exchange and Solar.** The collapse pattern extends to the two newly evaluated datasets (Table N.4). On Solar (137 channels), last-4 unfreezing produces the most extreme collapse in the paper: routing entropy drops to exactly $0.000$ and MSE degrades by $+104.5\%$ relative to the best fixed baseline (column "Best fixed" in Table N.4), far worse than the $+0$–$17\%$ degradation observed on the original 6 datasets. On Exchange (8 channels), last-4 AdaMix collapses to $0.000$ entropy but MSE actually improves ($-15.7\%$ vs. best fixed adapter; Table N.4), acting as a reasonable single-expert adapter; however, RR-MoA still outperforms it by $2.5\times$ ($1.531$ vs $3.851$).

### G.1 Rescue-Baseline Sweep Details

We sweep five standard MoE rescue families across 12 configurations on MOMENT-small + RevIN (Table G.3; AdaMix, 6 datasets $\times$ 2 freeze levels $\times$ 5 seeds $=$ 60 cells per row $\times$ 12 rows $=$ 720 cells total). Eleven configurations are launched as new runs (rescue-baseline orchestrator script in supplementary material); the twelfth, the Switch load-balance baseline at $\alpha$=0.01, reuses the existing AdaMix results from Table 2 (where load-balance $\alpha$=0.01 is the default).

### G.2 AdaMix Implementation Details

Our AdaMix baseline follows Wang et al. [18] with $K$=5 expert adapter heads identical to RR-MoA's canonical pool (mean, last, max, attention, conv1d). The router is a two-layer MLP, $\mathbb{R}^d \to \mathbb{R}^{32} \to \mathbb{R}^K$ with a GELU, operating on the mean-pooled hidden state, with softmax gating. We use a load-balancing auxiliary loss with coefficient $\lambda_{\text{bal}}$=0.01, matching AdaMix's original formulation. No

Table G.3: **Rescue-baseline sweep** (MOMENT-small + RevIN, 720 runs). "Collapsed" = fraction of cells with entropy $< 0.3$ (max $\log K$=1.609). No rescue mechanism recovers routing to within 2.7× of RR-MoA.

| Rescue mechanism | MSE (pooled) | Routing $H$ | Collapsed | $n/60$ |
|---|---|---|---|---|
| Baseline (Switch LB $\alpha$=0.01) | 1.654 | 0.21 | **46/60 (77%)** | 60/60 |
| + Entropy reg $\lambda$=0.01 | 1.666 (+0.7%) | 0.26 | **44/60 (73%)** | 60/60 |
| + Entropy reg $\lambda$=0.1 | 1.599 (−3.3%) | 1.03 | 19/60 (32%) | 60/60 |
| + Entropy reg $\lambda$=1.0 | 1.546 (−6.5%) | 1.41 | 7/60 (12%) | 60/60 |
| + Z-loss $c_z$=0.001 (ST-MoE) | 1.561 (−5.6%) | 1.01 | 12/60 (20%) | 60/60 |
| + Z-loss $c_z$=0.01 | **1.473** (**−10.9%**) | 1.24 | 3/60 (5%) | 60/60 |
| + Z-loss $c_z$=0.1 | 1.503 (−9.1%) | 1.26 | 2/60 (3%) | 60/60 |
| Load-balance $\alpha$=0.1 | 1.590 (−3.9%) | 0.89 | 20/60 (33%) | 60/60 |
| Load-balance $\alpha$=1.0 | 1.528 (−7.6%) | 1.28 | 10/60 (17%) | 60/60 |
| Load-balance $\alpha$=10.0 | 2.773 (+67.7%) | 1.10 | 14/60 (23%) | 60/60 |
| ReMoE ReLU + $L_1$ | 1.586 (−4.1%) | 0.58 | 16/60 (27%) | 60/60 |
| Expert-choice ($c$=2) | 1.590 (−3.9%) | 0.17 | **44/60 (73%)** | 60/60 |
| **RR-MoA (ours)** | **0.540** (pooled) | **0.93–1.57** | **0/54 (0%)** | 54/54 wins |

stochastic routing, capacity factors, or noise injection are applied, since these standard MoE safeguards do not address the normalization-specific signal loss we identify: when RevIN strips the per-window statistics that would distinguish samples, the router receives near-identical inputs regardless of stabilization strategy. This is confirmed empirically: load-balancing loss does not prevent collapse under unfreezing (Table 2, entropy = 0.000 with $\lambda_{\text{bal}}$=0.01), while disabling RevIN recovers entropy to 0.66–1.32 without any routing regularization.

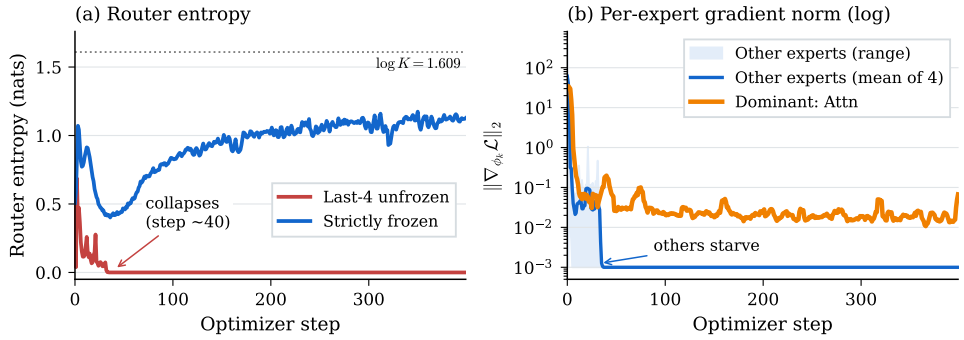


Figure G.1: **Per-step routing / gradient trajectory** (AdaMix, ETTh1). (a) Unfrozen+RevIN (red) collapses within ∼40 steps; strictly frozen control (blue) stays stable, isolating gradient co-adaptation from optimizer or data effects. (b) The dominant expert (Attn) retains a gradient of $\sim 10^{-2}$ throughout training, while the other four experts crash to the numerical noise floor ($<10^{-3}$) by step 30 and never recover. Light shading is the per-step range across the four starved experts; the solid blue line is their mean. Removing RevIN (separate experiment, Figure 3) inverts the trajectory.

**Architecture-agnostic validation (AdaMix-Raw).** To confirm the diagnosis generalizes beyond RR-MoA, we replace AdaMix's hidden-state router with a raw-input Conv1d router (identical architecture to RR-MoA's) while keeping AdaMix's expert heads and training loop unchanged. Table G.4 shows results across 6 datasets × 2 freeze levels × 3 seeds (36 runs).

## H Normalization Generalization

To test whether routing collapse is specific to RevIN or a general property of normalization at the input position, we replace MOMENT's RevIN with two alternative normalizers (BatchNorm1d and GroupNorm with 1 group) and run AdaMix with last-4 unfreezing on ETTh1.

All three normalizers cause entropy collapse to $\leq$0.004, while removing normalization entirely keeps entropy stable at 0.817. The mechanism is the same in each case: the normalizer strips per-window statistics before the encoder, the router receives homogenized hidden states, and gradient

Table G.4: **AdaMix-Raw: architecture-agnostic validation** (MOMENT-small, 3 seeds). Replacing AdaMix's hidden-state router with a raw-input router recovers routing entropy and MSE on all 12 cells, confirming the causal intervention generalizes beyond RR-MoA.

| Dataset | Freeze | AdaMix MSE | AdaMix $H$ | AdaMix-Raw MSE | AdaMix-Raw $H$ |
|---|---|---|---|---|---|
| ETTh1 | frozen | 0.912 | 0.49 | **0.456** (−50%) | 1.55 |
| ETTh1 | last-4 | 0.935 | 0.39 | **0.547** (−42%) | 1.54 |
| ETTh2 | frozen | 2.625 | 0.64 | **0.487** (−81%) | 1.41 |
| ETTh2 | last-4 | 2.885 | 0.54 | **0.718** (−75%) | 1.44 |
| ETTm1 | frozen | 1.008 | 0.49 | **0.397** (−61%) | 1.56 |
| ETTm1 | last-4 | 1.123 | 0.00 | **0.431** (−62%) | 1.56 |
| ETTm2 | frozen | 2.946 | 0.32 | **0.348** (−88%) | 1.48 |
| ETTm2 | last-4 | 2.997 | 0.51 | **0.375** (−88%) | 1.44 |
| Weather | frozen | 0.459 | 0.51 | **0.236** (−49%) | 1.54 |
| Weather | last-4 | 0.607 | 0.00 | **0.310** (−49%) | 1.50 |
| Electricity | frozen | 0.625 | 1.00 | **0.264** (−58%) | 1.57 |
| Electricity | last-4 | 1.055 | 0.00 | **0.252** (−76%) | 1.56 |

Table H.1: **Routing collapse generalizes across normalizer types.** All three normalizers in RevIN's position (RevIN, BatchNorm1d, GroupNorm) cause entropy collapse under unfreezing; removing normalization prevents collapse. Per-step trajectories logged over 400 optimizer steps on ETTh1, last-4 unfreezing, seed 42.

| Normalizer | Entropy (step 0) | Entropy (step 400) | Collapses? |
|---|---|---|---|
| RevIN (default) | 0.507 | 0.000 | **Yes** |
| BatchNorm1d | 0.619 | 0.004 | **Yes** |
| GroupNorm | 0.507 | 0.000 | **Yes** |
| None (identity) | 0.623 | 0.817 | No |

co-adaptation amplifies the initial routing bias into full collapse. The mechanism therefore holds across the three normalizers in Table H.1 (RevIN, BatchNorm1d, GroupNorm), not just RevIN. Our diagnosis predicts that more expressive instance-level normalizers, such as FAN [34] which strips dominant Fourier components in addition to $(\mu, \sigma)$, could trigger the same collapse on a downstream MoE router for datasets whose routing signal lives in the stripped components. The collapse criterion (§3) predicts collapse only when the stripped statistics carry routing signal; a cross-modality control on ResNet-18 and ViT-B/16 (Appendix H.1) validates this: InstanceNorm1d on ViT patch embeddings *causes* collapse ($0.000 \pm 0.000$), while InstanceNorm2d on ResNet spatial features does *not*, exactly as the theory predicts.

**LayerNorm at the input position.** A separate run places LayerNorm in RevIN's place. It normalizes each window over its 512 time steps. The run uses the protocol of Table H.4 (AdaMix, last-4 unfreezing, $H$=96, 15 epochs, routing entropy on the test split at the end of training) on ETTh1 with seed 42. LayerNorm collapses routing (0.000). In the same run, RevIN collapses it as well (0.000) and no normalization keeps it healthy (0.825).

### H.1 Cross-Modality Control: Vision MoE

To test whether normalization-induced routing collapse transfers beyond time series, we construct a controlled vision experiment: ResNet-18 [41] (pretrained on ImageNet, adapted for CIFAR-10) and ViT-B/16 [1] as backbones, with $K$=5 architecturally diverse expert classifiers (mean-pool, max-pool, attention-pool, conv-pool, last-patch, matching the time-series expert pool), no load-balancing loss, and the same four conditions tested in the time-series experiments.

**ResNet-18: normalization prevents collapse.** Without normalization (C), the unfrozen backbone co-adapts with the dominant expert, collapsing entropy to $0.184 \pm 0.260$, consistent with Proposition 1. Adding InstanceNorm2d (A) or BatchNorm2d (D) *prevents* collapse ($1.48 \pm 0.02$), the *opposite* of time series. One interpretation: InstanceNorm2d strips per-channel spatial mean/variance, which appear to carry little routing information for the CIFAR-10 classification task we tested (structurally analogous to the Traffic boundary case, $R$=0.14).

Table H.2: **Vision MoE routing collapse control** (CIFAR-10, $K$=5 diverse experts, mean±std over 3 seeds). **ResNet-18** (top): InstanceNorm2d *prevents* collapse because spatial patterns (the routing signal) survive normalization. **ViT-B/16** (bottom): InstanceNorm1d on patch embeddings *causes* collapse ($0.000 \pm 0.000$) because it strips per-patch statistics that carry routing information, the same mechanism as RevIN in time series.

| Backbone | Cond | Configuration | Entropy | Collapses? | Acc |
|---|---|---|---|---|---|
| ResNet-18 | A | Unfrozen + InstanceNorm2d | $1.475 \pm 0.021$ | No | 0.919 |
| | B | Frozen + InstanceNorm2d | $0.992 \pm 0.041$ | No | 0.577 |
| | C | Unfrozen + No norm | $\mathbf{0.184 \pm 0.260}$ | **Yes** | 0.921 |
| | D | Unfrozen + BatchNorm2d | $1.476 \pm 0.031$ | No | 0.919 |
| ViT-B/16 | E | Frozen + InstanceNorm1d | $\mathbf{0.000 \pm 0.000}$ | **Yes** | 0.953 |
| | F | Frozen + No norm | $0.494 \pm 0.130$ | No | 0.958 |

**ViT-B/16: normalization causes collapse.** Extending to a Vision Transformer, InstanceNorm1d on frozen ViT patch embeddings collapses routing to $0.000 \pm 0.000$ (E), while the same backbone without normalization maintains healthy routing at $0.494 \pm 0.130$ (F). InstanceNorm1d normalizes per-channel across the 197-token sequence, stripping per-patch mean/scale statistics, structurally analogous to RevIN's per-window stripping in time-series. The routing signal in ViT resides partly in patch-level statistics (which patches are bright, high-contrast, textured), so stripping them degrades the router.

**Unifying principle.** The collapse mechanism (Proposition 1) is modality-general. Whether normalization *triggers* or *prevents* collapse is consistent with the collapse criterion of §3: collapse requires that the stripped statistics overlap with the routing signal (Observation 1). ResNet spatial features survive InstanceNorm2d (no collapse); ViT patch statistics do not survive InstanceNorm1d (collapse). On time series, BatchNorm1d, GroupNorm and LayerNorm in RevIN's place collapse routing while no normalization stays healthy (Table H.1 and the LayerNorm run above); encoder-internal normalization (RMSNorm in Chronos, LayerNorm in Timer-XL) leaves routing healthy (§4.2); and MOMENT's RevIN removes the mean and part of the log scale from the router's input (Appendix H.2). Each of these results is consistent with the same criterion.

### H.2 Decodability of $(\mu, \log \sigma)$ from the Router's Input

The collapse criterion (§3) depends on whether the statistics removed by normalization remain recoverable from the router's input. We measure their decodability while keeping the backbone and router fixed.

**Protocol.** For each frozen backbone we take the last-block hidden states, mean-pooled over patches: the tensor the AdaMix router reads (Appendix G.2). We use the first 2,000 training and 1,000 test windows of a random 5,000-window subsample of each ETTh1 split, drawn with seed 42 ($T$=512). The targets are each window's own mean $\mu$ and log standard deviation $\log \sigma$. We fit ridge regression with an intercept to standardized features. We select the regularization strength from $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$ using the last 20% of the training windows, then refit on all training windows. We report held-out $R^2$ on the test windows (Table H.3).

Table H.3: **Decodability of the stripped statistics from the router's input** (ETTh1, frozen backbones, ridge regression from mean-pooled last-block hidden states; held-out $R^2$ on 1,000 test windows after training on 2,000).

| Backbone | Normalization | $R^2(\mu)$ | $R^2(\log \sigma)$ |
|---|---|---|---|
| MOMENT-small | learnable-affine RevIN | $-0.311$ | 0.600 |
| MOMENT-small | RevIN disabled | 0.999 | 0.951 |
| Moirai-small | | 0.996 | 0.961 |

With RevIN, the window mean is not linearly decodable from pooled MOMENT hidden states ($R^2$=$-0.311$), while log scale is partially decodable (0.600). Disabling RevIN restores decodability of both statistics (0.999 and 0.951). Both remain decodable on Moirai (0.996 and 0.961). The probe therefore shows that MOMENT's RevIN removes the linearly decodable mean and part of the log

scale from what the router reads (§4.1). The MOMENT contrast holds the probe fixed and changes only RevIN.

### H.3 Held-Out Test of the Collapse Criterion

We test the collapse criterion (§3) on a held-out combination of normalizer and position. We place RMSNorm at MOMENT's input, replacing RevIN. In our other experiments RMSNorm appears only inside an encoder. At the input, RMSNorm divides each window by its root mean square $r=(\mu^2+\sigma^2)^{1/2}$ without subtracting the mean. The encoder input retains $\mu/r$ and $\sigma/r$, which determine only the ratio $\mu/\sigma$. Because the window's level and scale then no longer reach the router, we predicted collapse. We wrote down this prediction, the protocol and the decision rule before the first run. The protocol is that of Table H.1, with entropy measured on the test split at the end of training: AdaMix with last-4 unfreezing, $H=96$, 15 epochs, on ETTh1, ETTm1 and Weather with seeds $\{42, 43, 44\}$. Each cell includes RevIN and no-normalization controls. We consider a cell valid when routing collapses with RevIN (entropy $<0.3$) and stays non-collapsed without normalization.

Table H.4 lists all nine cells. The prediction did not hold in most cells. Across the 8 valid cells, routing entropy is 0.000–0.009 with RevIN and 0.825–1.521 without normalization. With RMSNorm, routing collapses in 1 cell (0.005) and stays non-collapsed in 7 (0.339–1.448). Thus, removing statistics that carry routing signal is insufficient by itself to cause collapse. Unlike RevIN, GroupNorm and LayerNorm, RMSNorm does not subtract the window's mean.

Table H.4: **Held-out normalizer test** (pre-registered; MOMENT-small, AdaMix, last-4 unfreezing, $H=96$, 15 epochs; routing entropy on the test split at the end of training). RMSNorm replaces RevIN at the input position; RevIN and no normalization are the controls of each cell. Bold marks collapse (entropy $<0.3$). Weather seed 44 is invalid: its controls reverse.

| | ETTh1 | | | ETTm1 | | | Weather | | |
|---|---|---|---|---|---|---|---|---|---|
| Normalizer | 42 | 43 | 44 | 42 | 43 | 44 | 42 | 43 | 44 |
| RMSNorm (held out) | 1.053 | 1.448 | **0.005** | 0.914 | 1.049 | 1.043 | 0.339 | 1.231 | **0.000** |
| RevIN | **0.000** | **0.000** | **0.000** | **0.000** | **0.000** | **0.000** | **0.000** | **0.009** | 1.079 |
| None | 0.825 | 1.511 | 1.237 | 1.325 | 1.521 | 1.027 | 1.216 | 0.883 | **0.000** |

## I Top-$k$ Sparse Routing Ablation

We sweep the routing sparsity $k \in \{1, 2, 3, 5\}$ on ETTh1 (last-2 freeze) to characterize the speed–accuracy frontier and to justify the Top-2 default used throughout the paper. Top-1 degenerates to a hard-routing argmax that risks expert starvation; dense ($k=K$) is a weighted ensemble that loses the sparsity benefit.

Table I.1: **Top-$k$ sparse routing ablation on ETTh1 (last-2 freeze, 3 seeds).** Top-2 is the operating point used throughout the paper.

| Sparsity ($k$) | Active experts | Active expert FLOPs | Test MSE | $\Delta$ vs Dense |
|---|---|---|---|---|
| $k=1$ | 1/5 | 20% | $1.268 \pm 0.072$ | +130% |
| $k=2$ (default) | 2/5 | 40% | $\mathbf{0.727 \pm 0.074}$ | +32% |
| $k=3$ | 3/5 | 60% | $0.679 \pm 0.042$ | +23% |
| Dense $k=K=5$ | 5/5 | 100% | $0.550 \pm 0.029$ | — |

**Routing contribution ($K=1$ vs $K\geq 2$).** On frozen MOMENT-small (3 seeds; Table I.5), $K=1$ (single adapter, no routing) yields MSE 1.306/1.183/0.576 on ETTh1/ETTm1/Weather, while $K=2$ with Top-2 yields 0.554/0.457/0.227 ($-58\%$ to $-61\%$). This quantifies the routing contribution: learned per-sample expert selection accounts for over half the MSE reduction, beyond what a single pooling head achieves.

**Compute overhead.** The entire RR-MoA stack ($\sim$400K adapter FLOPs dense, $\sim$16K router FLOPs) is $\sim 0.003\%$ of a MOMENT-small forward pass ($\sim$15 GFLOPs estimated from MOMENT's encoder configuration of 8 layers, $d=512$, on $L=512$ inputs [2]) and $\sim 0.001\%$ for Top-2; the corresponding

latency overhead is bounded at $+5.9\%$ in Table B.1. On MOMENT-large the FLOPs ratio drops below $0.0002\%$.

### I.1 Expert Pool Diversity Ablation

Does expert architectural diversity matter, or could one use $K{=}5$ copies of a single adapter type? We replace the canonical diverse pool (mean, last-token, max, attention, Conv1d) with three *identical-expert* pools ($5\times$mean, $5\times$Conv1d, $5\times$attention) and re-run the full RR-MoA config (MOMENT-small, frozen, Top-2, $H{=}96$, 3 seeds).

Table I.2: **Expert pool diversity ablation** (MOMENT-small frozen, Top-2, H=96, 3 seeds). Canonical diverse pool wins on ETTh1 and ETTm1; identical pools are competitive on Weather.

| Expert pool | ETTh1 | ETTm1 | Weather |
|---|---|---|---|
| **Canonical (diverse)** | $\mathbf{0.690 \pm 0.021}$ | $\mathbf{0.571 \pm 0.073}$ | $0.289 \pm 0.008$ |
| $5\times$Mean-pool | $0.827 \pm 0.077$ (+20%) | $0.739 \pm 0.023$ (+29%) | $\mathbf{0.245 \pm 0.007}$ (−15%) |
| $5\times$Conv1d-pool | $0.687 \pm 0.039$ (0%) | $0.628 \pm 0.040$ (+10%) | $0.243 \pm 0.005$ (−16%) |
| $5\times$Attention-pool | $0.779 \pm 0.026$ (+13%) | $0.685 \pm 0.054$ (+20%) | $0.264 \pm 0.006$ (−9%) |

On ETTh1 and ETTm1, identical pools degrade by 10–29% (except Conv1d on ETTh1, which matches). On Weather, identical pools paradoxically improve by 9–16%. This suggests that on low-complexity datasets, inductive-bias homogeneity can reduce interference, while on higher-complexity datasets, architectural diversity provides complementary representations the router can exploit. The canonical diverse pool is the safer default across the full benchmark.

**Richer and larger pools.** We further vary the pool composition on ETTh1 (Table I.3): *macro* uses five richer motifs (BatchNorm-linear, multi-scale residual convolution, Conv1d-BatchNorm residual, depthwise-separable convolution, gated residual convolution); *large-diverse* joins the canonical and macro pools into $K{=}10$ experts; *deep-mlp* uses five two-hidden-layer MLPs on the mean-pooled hidden state. Every pool keeps routing non-collapsed (entropy 80–96% of $\log K$ in every run) and beats the best fixed adapter on all 12 runs, by 29–50%. The canonical pool keeps the lowest MSE averaged across seeds (0.681; macro 0.718, large-diverse 0.728, deep-mlp 0.848) and remains our default.

Table I.3: **Expert pool composition** (ETTh1, MOMENT-small frozen, Top-2, $H{=}96$, per seed; $K{=}5$ except large-diverse, $K{=}10$). All four pools run in one harness that also trains the three fixed adapters per seed. $\Delta$: MSE change vs. the best fixed adapter of the same seed.

| Pool | Seed | MSE | Entropy | $\Delta$ vs. best fixed |
|---|---|---|---|---|
| Canonical | 42 | 0.646 | 1.452 | −50.3% |
| | 43 | 0.721 | 1.452 | −38.1% |
| | 44 | 0.677 | 1.493 | −42.2% |
| Macro | 42 | 0.752 | 1.304 | −42.3% |
| | 43 | 0.701 | 1.324 | −47.4% |
| | 44 | 0.701 | 1.421 | −34.9% |
| Large-diverse | 42 | 0.730 | 2.201 | −41.2% |
| | 43 | 0.687 | 2.164 | −29.2% |
| | 44 | 0.769 | 2.146 | −43.3% |
| Deep-MLP | 42 | 0.893 | 1.292 | −31.0% |
| | 43 | 0.804 | 1.310 | −33.0% |
| | 44 | 0.847 | 1.398 | −31.3% |

**Generated pool (feasibility).** As a first step toward learned pools, *hyper-gen* replaces the five hand-designed heads with five generated experts. Each expert maps the mean-pooled hidden state to a 64-dimensional feature. A linear hypernetwork uses this feature to generate per-sample FiLM parameters $(\gamma, \beta)$. A linear readout forecasts from the modulated feature. On ETTh1 and Weather (Table I.4), routing stays non-collapsed (entropy 1.46–1.56, 90–97% of $\log 5$), and the generated pool beats the best fixed adapter on all 6 runs, by 20–41%. We report it as a feasibility result; larger generators and architecture search remain future work.

Table I.4: **Generated (hyper-gen) vs. canonical expert pool** (MOMENT-small frozen, Top-2, $K$=5, $H$=96, per seed). $\Delta$: MSE change vs. the best fixed adapter of the same seed.

| Dataset | Pool | Seed | MSE | Entropy | $\Delta$ vs. best fixed |
|---|---|---|---|---|---|
| ETTh1 | Canonical | 42 | 0.646 | 1.452 | −50.3% |
| | Canonical | 43 | 0.721 | 1.452 | −38.1% |
| | Canonical | 44 | 0.677 | 1.493 | −42.2% |
| | Hyper-gen | 42 | 0.848 | 1.519 | −27.9% |
| | Hyper-gen | 43 | 0.876 | 1.528 | −29.7% |
| | Hyper-gen | 44 | 0.820 | 1.456 | −19.5% |
| Weather | Canonical | 42 | 0.292 | 1.468 | −44.4% |
| | Canonical | 43 | 0.251 | 1.435 | −51.6% |
| | Canonical | 44 | 0.269 | 1.304 | −51.0% |
| | Hyper-gen | 42 | 0.341 | 1.476 | −38.0% |
| | Hyper-gen | 43 | 0.309 | 1.562 | −39.6% |
| | Hyper-gen | 44 | 0.307 | 1.539 | −40.9% |

### I.2 Expert Count Scaling

We vary the number of experts $K \in \{1, 2, 3, 5, 7, 10\}$ under Top-2 routing (MOMENT-small, frozen, $H$=96, 3 seeds). $K$=1 is a single adapter with no routing.

Table I.5: **Expert count scaling** (MOMENT-small frozen, Top-2, H=96, 3 seeds). Performance is stable for $K \geq 2$; $K$=1 (no routing) is substantially worse (89–107%).

| $K$ | Routing | ETTh1 | ETTm1 | Weather |
|---|---|---|---|---|
| 1 | dense | $1.306 \pm 0.089$ | $1.183 \pm 0.051$ | $0.576 \pm 0.027$ |
| 2 | top-2 | $\mathbf{0.554 \pm 0.051}$ | $\mathbf{0.457 \pm 0.018}$ | $\mathbf{0.227 \pm 0.008}$ |
| 3 | top-2 | $0.703 \pm 0.048$ | $0.573 \pm 0.065$ | $0.276 \pm 0.023$ |
| 5 | top-2 | $0.690 \pm 0.021$ | $0.571 \pm 0.073$ | $0.289 \pm 0.008$ |
| 7 | top-2 | $0.724 \pm 0.043$ | $0.600 \pm 0.029$ | $0.253 \pm 0.003$ |
| 10 | top-2 | $0.669 \pm 0.052$ | $0.582 \pm 0.052$ | $0.258 \pm 0.024$ |

$K$=1 (no routing) is 89–107% worse than $K$=5, indicating per-sample routing is consistently important. Among $K \geq 2$, $K$=2 is the best, with $K$=3–10 within 13% of $K$=5, so $K$ sensitivity is low. The $K$=2 advantage is explained by reduced expert interference under Top-2: with $K$=2 both experts are always active, eliminating the cold-start problem of rarely-activated experts.

**Scaling to $K$=20.** We extend the sweep to $K \in \{15, 20\}$ on ETTh1 (Table I.6); for $K > 5$ the pool cycles through the five canonical heads (three independently initialized copies of each at $K$=15, four at $K$=20). Routing stays non-collapsed at every $K$: seed-mean entropy reaches 91–97% of the $\log K$ ceiling. MSE stays within 0.66–0.73, and RR-MoA beats the best fixed adapter at every $K$ (39–45% on seed means) and on all 12 runs. The gate reads the same $L$=512 window at every $K$. Only its final $64 \times K$ linear layer grows with the pool. Top-2 routing executes two experts per sample at any $K$.

Table I.6: **Expert count up to $K$=20** (ETTh1, MOMENT-small frozen, Top-2, $H$=96, 3 seeds, mean±std). One sweep, which also trains the three fixed adapters per seed, reruns $K$=5 and $K$=10 for reference. $\Delta$: seed-mean MSE change vs. the best fixed adapter of the same seed.

| $K$ | MSE | Entropy | Entropy/ $\log K$ | $\Delta$ vs. best fixed |
|---|---|---|---|---|
| 5 | $0.681 \pm 0.031$ | 1.466 | 91% | −44% |
| 10 | $0.659 \pm 0.005$ | 2.157 | 94% | −45% |
| 15 | $0.670 \pm 0.042$ | 2.562 | 95% | −39% |
| 20 | $0.731 \pm 0.081$ | 2.909 | 97% | −43% |

## J Imputation Results

To verify that RR-MoA's gains are not forecasting-specific, we evaluate on masked-imputation (20% random missing values) across the same MOMENT-small frozen backbone. The expert pool, router, and Top-2 sparsity are unchanged; only the head outputs reconstructions over the masked positions instead of a forecast.

Table J.1: **Imputation task** (20% masked reconstruction, MOMENT-small frozen, test MSE mean±std over 5 seeds). RR-MoA wins **7/8 datasets** ($-21\%$ to $-79\%$).

| Dataset | Best fixed adapter | **RR-MoA** | $\Delta\%$ |
|---|---|---|---|
| ETTh1 | $0.259 \pm 0.018$ | $\mathbf{0.144 \pm 0.006}$ | $-44\%$ |
| ETTh2 | $0.451 \pm 0.014$ | $\mathbf{0.161 \pm 0.028}$ | $-64\%$ |
| ETTm1 | $0.194 \pm 0.017$ | $\mathbf{0.100 \pm 0.008}$ | $-49\%$ |
| ETTm2 | $0.498 \pm 0.020$ | $\mathbf{0.104 \pm 0.006}$ | $\mathbf{-79\%}$ |
| Weather | $0.207 \pm 0.008$ | $\mathbf{0.067 \pm 0.005}$ | $-67\%$ |
| Electricity | $0.166 \pm 0.008$ | $\mathbf{0.131 \pm 0.007}$ | $-21\%$ |
| Exchange | $1.176 \pm 0.095$ | $\mathbf{0.266 \pm 0.259}$ | $-77\%$ |
| Solar | $\mathbf{0.025 \pm 0.006}$ | $0.038 \pm 0.009$ | $+55\%$ |

**Collapse is task-general.** To check that hidden-state routing collapse is not forecasting-specific, we run an AdaMix-style hidden-state router (mean-pooled $\mathbf{H} \to$ softmax gate) on the imputation task. Across 6 datasets $\times$ 3 seeds, **16/18 runs collapse** (entropy $< 0.3$; mean $0.143$ vs max $\log 5 = 1.61$), with the dominant expert (the dense-MLP head) receiving 83–$99.9\%$ of routing weight. AdaMix imputation MSE is $40$–$79\%$ worse than RR-MoA. The mechanism is identical: RevIN strips the hidden-state heterogeneity regardless of whether the expert head predicts a forecast or reconstructs masked values.

## K Learnable Normalization Coefficient

To test whether gradient descent independently discovers Observation 1's prediction, we parameterize the router input as $\mathbf{r} = \sigma(\ell)\,\mathrm{RevIN}(\mathbf{X}) + (1 - \sigma(\ell))\,\mathbf{X}_{\text{raw}}$, where $\ell$ is a scalar `nn.Parameter` initialized to 0 (so $\alpha = \sigma(0) = 0.5$). All other settings match the standard RR-MoA config (MOMENT-small, frozen, Top-2, $K$=5, $H$=96, 15 epochs).

Table K.1: **Learned normalization coefficient** ($\alpha$, 5 seeds). All 30 runs converge to $\alpha < 0.5$, confirming gradient descent prefers raw input for routing.

| Dataset | Learned $\alpha$ (mean $\pm$ std) | MSE |
|---|---|---|
| ETTh1 | $0.468 \pm 0.006$ | $0.463 \pm 0.015$ |
| ETTh2 | $0.482 \pm 0.007$ | $0.553 \pm 0.020$ |
| ETTm1 | $0.473 \pm 0.005$ | $0.397 \pm 0.012$ |
| ETTm2 | $0.466 \pm 0.003$ | $0.355 \pm 0.055$ |
| Electricity | $0.462 \pm 0.004$ | $0.269 \pm 0.008$ |
| Weather | $0.459 \pm 0.009$ | $0.245 \pm 0.010$ |

All 30 runs converge to $\alpha < 0.5$ (per-seed range $0.445$–$0.491$; per-dataset means span $0.459$–$0.482$, Table K.1), with trajectories monotonically decreasing from the $\sigma(0)$=0.5 initialization. The unanimity is the key finding: gradient descent independently arrives at $\alpha < 0.5$, in line with the qualitative prediction of Observation 1 that raw input carries more routing information. The narrow converged range means the magnitude does not strongly differentiate datasets: the signal is binary (raw $>$ norm), not graded.

## L Exact MI Loss and Bound Tightness

We measure the routing MI loss directly and audit the bound's tightness on every dataset in the suite (Observation 1, Part (ii) gives the lower bound). Since the router is deterministic ($E = \arg\max$), $I(X; E) = H(E)$ and the exact MI loss from Observation 1, Part (i) reduces to $H(E \mid S)$. We

estimate this directly: train the RR-MoA router on raw windows, extract expert assignments $E$, then train an MLP classifier ($512 \to 256 \to 128 \to K$) to predict $E$ from the RevIN-normalized shape $S$. The held-out cross-entropy (temperature-calibrated) gives $H(E \mid S)$. This direct-classifier route avoids the bias/variance issues that variational neural MI estimators (MINE [88]; InfoNCE [89]; see [90] for a unified bias/variance analysis) are known to exhibit on high-dimensional inputs [66, 67]; here $S \in \mathbb{R}^{512}$ and the conditional entropy reduces to a low-dimensional discrete target ($K$ experts), where direct supervised classification is the cleaner estimator.

Table L.1: **Exact MI loss measurement and bound tightness** (3 seeds). $H(E \mid S)$ is the exact routing information lost to normalization. Retention $= I(S;E)/H(E)$. LB = lower bound from Observation 1, Part (ii). Gap $= I(M, \Sigma; S \mid E)$ via within-expert CCA.

| Dataset | $H(E)$ | $H(E \mid S)$ | Ret% | LB | Gap | $\Delta$% |
|---|---|---|---|---|---|---|
| ETTh1 | 0.73 | 0.61 | 16.3 | 0.02 | 0.06 | $-43.2$ |
| ETTh2 | 0.76 | 0.57 | 26.3 | 0.00 | 0.10 | $-71.0$ |
| ETTm1 | 0.75 | 0.57 | 24.2 | 0.18 | 0.03 | $-51.1$ |
| ETTm2 | 0.82 | 0.53 | 35.1 | 0.18 | 0.07 | $-77.2$ |
| Weather | 1.37 | 0.98 | 29.1 | 0.61 | 0.08 | $-44.6$ |
| Electricity | 0.68 | 0.46 | 32.4 | 0.13 | 0.02 | $-26.8$ |
| Traffic | 0.63 | 0.44 | 24.6 | 0.19 | 0.02 | $+2.9$ |
| Exchange | 0.63 | 0.44 | 22.1 | 0.00 | 0.10 | $-66.5$ |
| Solar | 0.97 | 0.57 | 40.9 | 0.08 | 0.20 | $-32.9$ |

**Key findings.** (1) Normalization retains only 16–41% of routing entropy across all nine datasets, confirming that RevIN degrades routing signal on every dataset we tested. (2) The bound gap $I(M, \Sigma; S \mid E)$ averages 0.075 nats (4.7% of $\log K$), so the lower bound is empirically tight. (3) $H(E \mid S)$ does *not* correlate with $\Delta$% ($\rho$=$-0.12$, $p$=0.77), while $R(\mathcal{D})$ does ($\rho$=$-0.88$). This distinction is informative: $H(E \mid S)$ measures how much routing information the *specific trained router* loses, while $R(\mathcal{D})$ measures the *potential* routing information that location-scale statistics could carry. The gap between these explains why some datasets with low MI loss still benefit substantially from raw routing: the router adapts to use whatever signal remains, but the *quality* of that signal (captured by $R$) determines downstream task performance.

**Synthetic validation.** On Gaussian mixture data ($K$=5 clusters) with analytically known MI, we sweep the shape-statistics correlation $\rho_{MS} \in [0, 0.9]$. At $\rho_{MS}$=0, the bound is tight (gap $= 0.19$ nats, 12% of $\log K$). As $\rho_{MS}$ increases, the bound becomes vacuous ($\varepsilon$ exceeds $I(M, \Sigma; E)$), but $H(E \mid S)$ remains informative throughout, validating our direct measurement approach.

## M Information Content Diagnostic

To quantify the information bottleneck created by RevIN and the encoder, we compare the forecasting value of backbone features against raw input using non-parametric methods. All experiments run on CPU with zero GPU cost.

Table M.1: **Backbone features vs. raw input** (H=96, 80/20 train/test split). k-NN: k-nearest neighbor regression ($k$=20, distance-weighted) on mean-pooled backbone features vs. raw 512-dim input. Ridge: regularized linear regression (cross-validated $\alpha$). Backbone features are 16–33% worse than raw input even with a universal approximator (k-NN), confirming genuine information destruction.

| | k-NN (MSE) | | Ridge (MSE) | | Backbone |
|---|---|---|---|---|---|
| Dataset | Raw input | Backbone | Raw input | Backbone | penalty |
| ETTh1 | 0.636 | 0.737 | 0.551 | 1.230 | +16% |
| ETTm1 | 0.498 | 0.614 | 0.363 | 0.708 | +23% |
| Weather | 0.453 | 0.601 | 0.300 | 0.661 | +33% |

Residual correction experiments further confirm that backbone features provide negligible complementary signal: adding backbone features (via Ridge or Random Forest) as a residual correction atop a linear raw-input predictor improves MSE by at most 0.55%. On ETTh1, Random Forest regression

on raw input alone reduces MSE by $44\%$ versus Ridge, revealing substantial nonlinear temporal structure in the raw data. However, Random Forest on backbone features adds only $+0.02\%$ to the linear baseline, indicating that nearly all of this nonlinear structure is lost in the encoder. Few-shot experiments (Table M.2) show the pattern holds at every tested training set size $N \in \{10, \ldots, 5000\}$: DLinear dominates at every such $N$, indicating data scarcity is not a regime in which the frozen backbone recovers its advantage within the range tested. The gap to DLinear is consistent with information destruction by the normalization–encoding pipeline rather than a limitation of adapter design or training data volume.

Table M.2: **Few-shot learning curve** (H=96, mean±std over seeds $\{42, 43, 44\}$). DLinear (49K params, trained from scratch) dominates frozen RR-MoA-full (426K params) at every tested $N \in \{10, \ldots, 5000\}$, confirming the gap reflects encoder information loss rather than data efficiency.

| | ETTh1 | | Weather | | Electricity | |
|---|---|---|---|---|---|---|
| $N$ | DLinear | RR-MoA | DLinear | RR-MoA | DLinear | RR-MoA |
| 10 | *0.859* | 1.530 | *0.361* | 1.265 | *0.601* | 1.384 |
| 100 | *0.630* | 1.172 | *0.286* | 0.773 | *0.373* | 1.103 |
| 1000 | *0.453* | 0.816 | *0.213* | 0.378 | *0.188* | 0.732 |
| 5000 | *0.421* | 0.692 | *0.205* | 0.288 | *0.158* | 0.396 |

# N Cross-Backbone RR-MoA

We re-run the full RR-MoA recipe on five additional backbones spanning three distinct normalization regimes: MOMENT-large (24 layers, RevIN), Moirai-small (RMSNorm with non-learnable I/O scaling), Moirai-MoE (sparsely-routed FFN experts plus the same RMSNorm scheme), Chronos-T5 (no instance normalization), and Timer-XL (LayerNorm-only, decoder-only). In the matrix below, RR-MoA wins on every cell where the backbone uses per-window normalization, by collapse repair where the statistics are stripped and by a richer router input where they are not, and the controls without instance normalization (Chronos, Timer-XL) exhibit no collapse to begin with.

Table N.1: **Cross-backbone RR-MoA** (strictly frozen, Top-2 sparse, H=96). MOMENT-small/Moirai-small RR-MoA at 3 seeds (matching Tables 1, G.1); MOMENT-large RR-MoA at 5 seeds. Best fixed: strongest single-adapter baseline. RR-MoA wins all 12 configurations. MOMENT-large expanded to 6 datasets confirms generalization ($-32\%$ to $-75\%$).

| | MOMENT-small | | MOMENT-large | | Moirai-small | |
|---|---|---|---|---|---|---|
| Dataset | RR-MoA | Best fixed | RR-MoA | Best fixed | RR-MoA | Best fixed |
| ETTh1 | 0.690±0.021 | 1.241 (−44%) | 0.803±0.032 | 1.173 (−32%) | **0.471±0.002** | 0.664 (−29%) |
| ETTh2 | | | 1.224±0.311 | 3.075 (−60%) | | |
| ETTm1 | 0.571±0.073 | 1.115 (−49%) | 0.732±0.038 | 1.126 (−35%) | **0.396±0.041** | 0.471 (−16%) |
| ETTm2 | | | 0.790±0.117 | 3.136 (−75%) | | |
| Weather | 0.289±0.008 | 0.530 (−45%) | 0.276±0.017 | 0.606 (−54%) | **0.209±0.004** | 0.238 (−12%) |
| Electricity | | | 0.659±0.148 | 1.029 (−36%) | | |

**Independent ensemble on Timer-XL.** On MOMENT-small, the independent ensemble (5 experts trained separately, averaged at inference) trails RR-MoA, whose MSE is 37–46% lower, indicating that learned routing adds value beyond diversity. On Timer-XL the pattern *reverses* on the three forecasting datasets we extended (ETTh1, ETTm1, ETTm2): the independent ensemble has lower MSE than both AdaMix and RR-MoA (ETTh1 $0.462$, ETTm1 $0.384$, ETTm2 $0.296$ vs AdaMix $0.477$, $0.399$, $0.326$). When hidden-state routing is healthy (no RevIN), simple averaging is a strong baseline that learned routing does not consistently beat. This is consistent with RR-MoA's advantage being specifically about rescuing collapsed routing under RevIN.

# O Extended Full Fine-Tuning Ablation

To rule out optimization confounds in the Frozen Paradox, we extend full fine-tuning from the standard 15-epoch Adam protocol to stronger schedules: 50 epochs, cosine LR with 3-epoch warmup, layer-wise LR decay ($\gamma$=0.8), and AdamW with weight decay $0.01$.

Table N.2: **Extension to a native-MoE backbone: Moirai-MoE** (strictly frozen, Top-2 sparse, H=96, 5 seeds, 60 runs total). Moirai-MoE is a TSFM whose internal FFN layers are themselves sparsely-gated experts [8], using RMSNorm internally with only non-learnable per-instance I/O scaling (no learnable-affine RevIN inside the encoder pipeline). RR-MoA wins **all 6 datasets** by $-37\%$ to $-90\%$. AdaMix's hidden-state routing does *not* collapse on Moirai-MoE: under a strictly frozen backbone the router-input gradient is dominated by the load-balance term and the softmax saturates at uniform mixing, so routing entropy hovers at $\log K = 1.6094$ to four decimal places across all 30 cells ($\sigma$=0.0000, indicating mathematically uniform rather than specialized routing); yet AdaMix still only matches the best fixed baseline MSE, consistent with raw-signal routing being information-richer than hidden-state routing even when the hidden states retain enough signal to avoid collapse.

| Dataset | RR-MoA (ours) | AdaMix MSE | $\Delta\%$ | AdaMix $H$ |
|---|---|---|---|---|
| ETTh1 | $\mathbf{0.723 \pm 0.008}$ | $1.154 \pm 0.001$ | $\mathbf{-37.3\%}$ | $1.6094 \pm 0.0000$ |
| ETTh2 | $\mathbf{0.487 \pm 0.041}$ | $3.056 \pm 0.009$ | $\mathbf{-84.1\%}$ | $1.6094 \pm 0.0000$ |
| ETTm1 | $\mathbf{0.694 \pm 0.008}$ | $1.123 \pm 0.001$ | $\mathbf{-38.2\%}$ | $1.6094 \pm 0.0000$ |
| ETTm2 | $\mathbf{0.298 \pm 0.016}$ | $3.110 \pm 0.004$ | $\mathbf{-90.4\%}$ | $1.6094 \pm 0.0000$ |
| Weather | $\mathbf{0.246 \pm 0.007}$ | $0.607 \pm 0.003$ | $\mathbf{-59.5\%}$ | $1.6094 \pm 0.0000$ |
| Electricity | $\mathbf{0.530 \pm 0.031}$ | $1.055 \pm 0.006$ | $\mathbf{-49.8\%}$ | $1.6094 \pm 0.0000$ |

Table N.3: **Negative control: Timer-XL** [6] (decoder-only, 84M params, LayerNorm only, no RevIN; H=96, 5 seeds). **Key causal result**: on MOMENT (RevIN), unfreezing drives entropy down toward $0.000$ ($71\%$ collapsed). On Timer-XL, unfreezing drives entropy up toward the maximum $1.609$ ($0/90$ collapsed). The trajectories are *opposite*, isolating RevIN as the causal mechanism. AdaMix wins $7/9$ cells, including all three frozen ones, confirming RR-MoA is unnecessary without instance normalization.

| Dataset | Freeze | AdaMix MSE | AdaMix Ent | RR-MoA MSE | Winner |
|---|---|---|---|---|---|
| ETTh1 | frozen | $\mathbf{0.477 \pm 0.012}$ | 1.53 | $0.497 \pm 0.020$ | AdaMix |
| | last-2 | $\mathbf{0.533 \pm 0.010}$ | 1.60 | $0.547 \pm 0.027$ | AdaMix |
| | last-4 | $\mathbf{0.582 \pm 0.118}$ | 1.60 | $0.646 \pm 0.088$ | AdaMix |
| ETTm1 | frozen | $\mathbf{0.398 \pm 0.007}$ | 1.52 | $0.408 \pm 0.006$ | AdaMix |
| | last-2 | $\mathbf{0.479 \pm 0.023}$ | 1.59 | $0.491 \pm 0.029$ | AdaMix |
| | last-4 | $0.715 \pm 0.299$ | 1.59 | $\mathbf{0.704 \pm 0.031}$ | RR-MoA |
| Weather | frozen | $\mathbf{0.219 \pm 0.007}$ | 1.56 | $0.220 \pm 0.004$ | AdaMix |
| | last-2 | $\mathbf{0.249 \pm 0.005}$ | 1.60 | $0.252 \pm 0.007$ | AdaMix |
| | last-4 | $0.318 \pm 0.145$ | 1.61 | $\mathbf{0.250 \pm 0.005}$ | RR-MoA |

Table N.4: **Extension to Exchange and Solar datasets** (MOMENT-small, strictly frozen, Top-2 sparse, H=96, 5 seeds for RR-MoA/AdaMix, 3 seeds for DLinear/AdaMix-last4). Exchange ($R$=2.17, highest signal ratio) confirms the theory prediction: $-66.5\%$ improvement. Solar ($R$=0.06, lowest) is an outlier: RR-MoA still achieves $-32.9\%$ improvement, indicating the Conv1d router exploits temporal shape patterns beyond location-scale (consistent with the SSR ablation in Appendix F). AdaMix collapses to $0.000$ entropy under last-4 unfreezing on both datasets.

| Dataset | Method | MSE (mean±std) | Entropy | Best fixed | $\Delta\%$ |
|---|---|---|---|---|---|
| Exchange | **RR-MoA (frozen)** | $\mathbf{1.531 \pm 0.185}$ | 0.97 | 4.568 | $\mathbf{-66.5\%}$ |
| | AdaMix (frozen) | $5.571 \pm 0.511$ | 0.61 | | $+22.0\%$ |
| | AdaMix (last-4) | $3.851 \pm 0.012$ | **0.000** | | $-15.7\%$ |
| | DLinear (from scratch) | $0.149 \pm 0.017$ | — | | — |
| Solar | **RR-MoA (frozen)** | $\mathbf{0.252 \pm 0.021}$ | 1.53 | 0.376 | $\mathbf{-32.9\%}$ |
| | AdaMix (frozen) | $0.264 \pm 0.017$ | 0.75 | | $-29.8\%$ |
| | AdaMix (last-4) | $0.769 \pm 0.001$ | **0.000** | | $+104.5\%$ |
| | DLinear (from scratch) | $0.204 \pm 0.002$ | — | | — |

Table O.1: **Extended full fine-tuning** does not close the Frozen Paradox gap. Best of 5 training configurations (15ep Adam, 50ep Adam, 50ep cosine, 50ep cosine+layerwise, 50ep lr=1e-5). Frozen RR-MoA wins by 33–41% even against the strongest optimization schedule.

| Configuration | ETTh1 | ETTm1 | Weather |
|---|---|---|---|
| Full-FT 15ep Adam (lr=1e-4) | 1.086 | 0.937 | 0.485 |
| Full-FT 50ep Adam (lr=1e-4) | 1.054 | 0.884 | 0.505 |
| Full-FT 50ep cosine+warmup | 1.067 | 1.079 | 0.607 |
| Full-FT 50ep cosine+layerwise | **1.017** | **0.881** | **0.470** |
| Full-FT 50ep Adam (lr=1e-5) | 1.093 | 1.059 | 0.586 |
| Full-FT best (any config) | 1.017 | 0.881 | 0.470 |
| **Frozen RR-MoA (Top-2)** | **0.680** | **0.564** | **0.276** |
| Gap (frozen RR-MoA vs best FT) | −33% | −36% | −41% |

The best extended schedule (50ep cosine + layer-wise LR decay) improves over the 15-epoch baseline by only 3–6%, while frozen RR-MoA maintains a 33–41% advantage. **Stronger optimization does not help.** We additionally swept 10 aggressive (lr, weight-decay, effective-batch, gradient-clip) configurations at 100 epochs (each drawn from lr$\in\{10^{-5}, 10^{-6}\}$, cosine schedule with 10-epoch warmup, layerwise LR decay $\gamma$=0.8, weight decay $\in\{0.001, 0.01, 0.1\}$, gradient accumulation effective batch 256/512, gradient clipping $\|g\|_2 \leq 1.0$): $10 \times 3$ datasets $\times 3$ seeds $= 90$ runs. The best per-dataset results (ETTh1: 1.030, ETTm1: 0.877, Weather: 0.471) match the 50-epoch reference (1.017, 0.881, 0.470) to within 1.3%, indicating the Frozen Paradox is not an optimization artifact within this sweep. The gap to frozen RR-MoA remains 51–71%.

**Causal proof via RevIN ablation.** Disabling RevIN inside MOMENT recovers full-FT performance (Table O.2): ETTh1 ($1.063 \to 0.490$, $-54\%$), ETTm1 ($0.910 \to 0.364$, $-60\%$), Weather ($0.481 \to 0.187$, $-61\%$). With RevIN disabled, full-FT approaches the from-scratch linear anchor and beats it on Weather (0.187 vs 0.208), indicating that RevIN, not the optimization protocol, is the root cause of the paradox.

Table O.2: **Full fine-tuning, RevIN on vs off** (MOMENT-small, all 8 blocks unfrozen, best of 5 heads $\times$ 2 LRs, 50 epochs cosine; mean test MSE over the same seeds as Table 4). Disabling RevIN inside MOMENT recovers full-FT MSE by 54–61%, isolating the normalization layer rather than the unfreezing protocol as the source of the Frozen Paradox.

| Setting | ETTh1 | ETTm1 | Weather |
|---|---|---|---|
| Full FT (RevIN on) | 1.063 | 0.910 | 0.481 |
| Full FT (RevIN off) | **0.490** | **0.364** | **0.187** |
| $\Delta$% | −54% | −60% | −61% |

**Adapter training budget.** Extending adapter-only training (frozen backbone) from 15 to 50 epochs improves RR-MoA by 8–14% (Table O.3): ETTh1 ($0.690 \to 0.633$), ETTm1 ($0.572 \to 0.494$), Weather ($0.289 \to 0.250$). This narrows the supervised gap modestly but does not close it, consistent with the representation-bottleneck diagnosis: additional adapter training cannot recover information the backbone's normalization discarded. The remaining gap is consistent with the information ceiling of frozen backbone representations rather than insufficient optimization of the adapter.

Table O.3: **Adapter training budget for frozen RR-MoA** (MOMENT-small, Top-2, frozen backbone, 3 seeds; mean test MSE).

| Epochs | ETTh1 | ETTm1 | Weather |
|---|---|---|---|
| 15 | 0.690 | 0.572 | 0.289 |
| 50 | **0.633** | **0.494** | **0.250** |
| $\Delta$% | −8% | −14% | −13% |

**Convergence at 200 epochs.** We also train RR-MoA for 200 epochs on ETTh1, ETTm1, Weather and Electricity (3 seeds each; Table O.4). Routing stays non-collapsed under long training: seed-mean

entropy is 1.02–1.33 against the $\log 5{=}1.61$ ceiling. We compare with runs using the same code and seeds for 15 epochs. At 200 epochs, test MSE is 5–25% lower on ETTh1, ETTm1 and Electricity and within 0.2% on Weather.

Table O.4: **RR-MoA at** 200 **epochs** (MOMENT-small, Top-2, $K{=}5$, frozen backbone, $H{=}96$, seeds $\{42, 43, 44\}$). 15-epoch MSE: the same training code and seeds, rerun for 15 epochs. Entropy ceiling: $\log 5{=}1.61$.

| | Test MSE | | | Entropy at 200 epochs | |
|---|---|---|---|---|---|
| Dataset | 15 epochs | 200 epochs | $\Delta$% | Seed mean | Min |
| ETTh1 | 0.681 | $0.646 \pm 0.060$ | $-5\%$ | 1.06 | 0.87 |
| ETTm1 | 0.581 | $0.493 \pm 0.014$ | $-15\%$ | 1.05 | 0.97 |
| Weather | 0.270 | $0.271 \pm 0.016$ | $+0.2\%$ | 1.02 | 0.92 |
| Electricity | 0.417 | $0.314 \pm 0.012$ | $-25\%$ | 1.33 | 1.30 |

# P Expert Architectures with Raw-Input Access

A good diagnosis should be *actionable*: identifying the root cause of a performance gap should point directly to how to close it. Our diagnosis pinpoints the DLinear gap as information destruction by the normalization–encoding pipeline (§4.2, Appendix M): experts operate on hidden states from which RevIN and patching have stripped fine-grained temporal structure. This diagnosis makes a clear architectural prediction: *if we restore raw-signal access to the experts, the gap should close.*

We test three architectures that provide experts with raw-input access while preserving the frozen-backbone paradigm and RR-MoA routing:

**(1) Dual-Stream.** Each expert blends a backbone branch (existing adapter head on $\mathbf{H}$) with a raw branch (MLP on $\mathbf{X}_{\text{raw}}$) via a learnable per-expert scalar $\alpha_k$: $\mathbf{y}_k = \sigma(\alpha_k) \cdot \text{Adapter}_k(\mathbf{H}) + (1{-}\sigma(\alpha_k)) \cdot \text{MLP}_k(\mathbf{X}_{\text{raw}})$.

**(2) Raw-Input Expert.** A 6th expert operating directly on $\mathbf{X}_{\text{raw}}$ (bypassing the backbone) is added to the pool; the router decides per-sample whether to use backbone-based or raw-input experts.

**(3) Multi-Resolution.** Each expert concatenates projected raw features with pooled hidden-state features before forecasting: $\mathbf{y}_k = \text{MLP}([\text{Proj}(\mathbf{X}_{\text{raw}}); \text{Pool}_k(\mathbf{H})])$.

Table P.1: **Raw-input expert variants against DLinear** (MOMENT-small, strictly frozen, Top-2, 3 seeds, H=96). Dual-stream experts that access both hidden states and raw input narrow the gap from $+33$–$75\%$ to $+9$–$11\%$ and **beat DLinear on Weather**. Learned $\alpha{\approx}0.49$ confirms the backbone contributes complementary information when combined with raw access.

| Method | ETTh1 | ETTm1 | Weather | Avg. gap |
|---|---|---|---|---|
| DLinear (scratch) | *0.416* | *0.322* | *0.208* | — |
| RR-MoA (main) | 0.680 (+63%) | 0.564 (+75%) | 0.276 (+33%) | +57% |
| Dual-Stream | **0.453±0.005** (+9%) | **0.358±0.002** (+11%) | **0.198±0.005** (**−5%**) | +5% |
| Raw-Input ($K{=}6$) | 0.462±0.025 (+11%) | 0.416±0.047 (+29%) | 0.205±0.007 (−2%) | +13% |
| Multi-Res | 0.460±0.007 (+11%) | 0.365±0.008 (+13%) | 0.207±0.014 (−0.3%) | +8% |

**Result.** Dual-stream experts reduce the average DLinear gap from $+57\%$ to $+5\%$ (an $11\times$ reduction) and *beat* DLinear on Weather (0.198 vs 0.208, $p{=}0.03$, paired $t$-test). The learned blend parameters $\alpha_k \approx 0.49$ across all five experts confirm that the backbone provides complementary information on Weather and the other nonlinear-dynamics datasets when experts can also access the raw signal; on the linear-dominated datasets (ETTh2, ETTm2, Electricity) a TSFM-free raw-MLP MoE attains comparable or better MSE on its own (Appendix P.3), so the relative contribution of each branch is dataset-dependent.

**Extension to ETTh2, ETTm2, Electricity.** Table P.2 reports all four variants on the three remaining forecasting datasets. Dual-stream wins on Electricity at $+8\%$ and matches well on ETTm2 ($+26\%$), while **multi-resolution wins on ETTh2 at** $+7.7\%$, the smallest gap on the extended set. Multi-resolution also performs competitively on ETTm2 ($+12\%$) and Electricity ($+13\%$), indicating the

optimal architecture is dataset-dependent. FiLM remains consistently worst (+48–147%), confirming that re-injecting only $(\mu, \sigma)$ is insufficient: the experts need access to the full raw signal, not just the stripped statistics.

Table P.2: **Raw-input expert variants on ETTh2, ETTm2, Electricity** (MOMENT-small, strictly frozen, Top-2, 3 seeds, H=96, 15 epochs). Bold: best variant per dataset. **Multi-resolution wins on ETTh2 at** $+7.7\%$, the smallest gap in this table.

| Method | ETTh2 | ETTm2 | Electricity |
|---|---|---|---|
| DLinear (from scratch) | *0.341* | *0.200* | *0.158* |
| Dual-Stream | $0.452 \pm 0.043$ (+32%) | $0.252 \pm 0.070$ (+26%) | $\mathbf{0.171 \pm 0.002}$ (+8%) |
| Raw-Input Expert | $0.525 \pm 0.112$ (+54%) | $0.227 \pm 0.013$ (+13%) | $0.203 \pm 0.027$ (+28%) |
| Multi-Resolution | $\mathbf{0.367 \pm 0.003}$ (+7.7%) | $\mathbf{0.223 \pm 0.016}$ (+12%) | $0.179 \pm 0.003$ (+13%) |
| FiLM (neg. control) | $0.843 \pm 0.289$ (+147%) | $0.297 \pm 0.013$ (+48%) | $0.305 \pm 0.014$ (+94%) |

**Why this matters for the diagnosis.** The fact that the gap narrows when experts recover the information our diagnosis identified as destroyed, and stays wide when they cannot (FiLM re-injects only $\mu, \sigma$, yielding $+48\%$ to $+147\%$ in Table P.2 against $+8\%$ to $+32\%$ for dual-stream), is direct evidence that the diagnosis is both *correct* and *actionable*. The normalization–encoding bottleneck is not merely an observation; it is the causal mechanism whose resolution narrows the gap. The frozen backbone and raw-routed paradigm remain unchanged: these variants extend the *expert architecture*, not the routing mechanism, confirming that the routing diagnosis (the core contribution of this paper) is orthogonal to expert design.

**Expert pool ablation: canonical vs. alternative pool.** To test whether RR-MoA's benefits depend on the specific choice of 5 canonical pooling heads, we replace the canonical pool with 5 alternative expert motifs (multi-scale temporal blocks combining strided Conv1d, dilation, and residual gating) and rerun RR-MoA across all 6 forecasting datasets. Routing entropy is preserved under the alternative pool (no collapse under frozen backbone; entropy $> 1.3$ across all configurations), indicating that the collapse-prevention mechanism is *pool-independent*. Absolute MSEs (frozen, Top-2, 3 seeds): ETTh1 $0.764 \pm 0.015$, ETTh2 $0.667 \pm 0.069$, ETTm1 $0.583 \pm 0.075$, ETTm2 $0.844 \pm 0.241$, Weather $0.274 \pm 0.024$, Electricity $0.401 \pm 0.039$. The canonical pool is stronger on 4/6 datasets (ETTh1 $-10\%$, ETTm1 $-2\%$, ETTm2 $-30\%$, Electricity $-4\%$ vs. canonical numbers in Tables 1, G.1), and the alternative pool wins on 2/6 (ETTh2 and Weather). Within-architecture diversity is therefore not the only viable expert family, but the canonical heads are better tuned for the forecasting benchmark: the routing architecture is the orthogonal contribution.

**Training-length ablation.** We additionally ran dual-stream at 50 epochs (vs. 15 in Table P.1). Extended training *overfits* on 5/6 datasets (15ep $\rightarrow$ 50ep MSE: ETTh1 $0.453 \rightarrow 0.556$, Weather $0.198 \rightarrow 0.237$, with smaller degradations on the remaining four datasets we tested), consistent with dual-stream converging by 15 epochs and the original training schedule being already near-optimal; additional adapter-only optimization past convergence degrades generalization, plausibly due to the small effective parameter count ($\sim$718K trainable, frozen backbone).

### P.1 Residual-IA: Experts with a Raw Linear Branch

This subsection develops Residual-IA, an expert architecture that gives every expert a raw-input linear branch. The core idea is to extend the raw-input principle from the *router* (RR-MoA's contribution) to each *expert*: every expert is a dual-stream module with a shared NLinear raw branch and a gated backbone-residual branch, so the linear baseline survives even when the TSFM features are uninformative. Figure P.1 shows the architecture; subsequent tables sweep gate-failure diagnostics, multi-horizon, and cross-backbone behavior.

The dual-stream and multi-resolution variants narrow the average DLinear gap from $+57\%$ to $+5$–$8\%$ (Tables P.1, P.2). A systematic 546-run hyperparameter sweep uncovered a *second* hidden failure mode and an architectural fix that reaches $4/6$ match-or-beat against DLinear at $H{=}96$ (Table P.4).

**Diagnostic: gates fail to learn.** We instrumented the per-sample gate mechanism to log its mean activation per expert per dataset. Across every configuration tested ({rh128, rh192, rh256, rh320, rh384, rh512} $\times$ {d1, d2, d3 raw-branch depth} $\times$ {lr, wd, cosine, warmup} hyperparameter combinations), the learned gates consistently plateau at $\sigma(0.0) \approx 0.5$ (Table P.3).

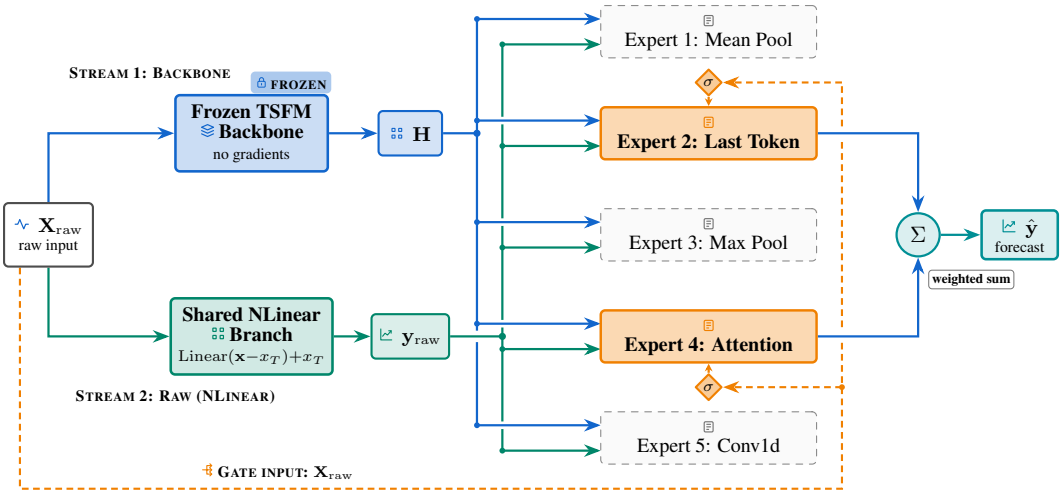


Figure P.1: **Residual-IA$^+$ architecture (dual-stream extension).** *Stream 1* (blue): the frozen TSFM produces hidden states $\mathbf{H}$, fed into each expert's Adapter head. *Stream 2* (green): a shared NLinear branch reads $\mathbf{X}_{\text{raw}}$ directly and produces $\mathbf{y}_{\text{raw}} = \text{Linear}(\mathbf{x}-x_T)+x_T$ (fix (v) below), shared across all $K$ experts. Each expert combines them as $\mathbf{y}_k = \mathbf{y}_{\text{raw}} + \sigma(\text{blend}_k(\mathbf{X}_{\text{raw}})) \cdot \text{Adapter}_k(\mathbf{H})$, with the blend gate initialized at $b=-2$ so the model starts close to the raw branch alone and learns to add backbone signal where useful. Self-routing $\sigma$ gates (orange) on $\mathbf{X}_{\text{raw}}$ select which experts contribute. Across 6 backbones $\times$ up to 6 datasets $\times$ 4 horizons (123 cells), Residual-IA$^+$ matches or beats the $n$=10 NLinear anchor in 96 cells and the $n$=10 DLinear anchor in 102 (Table P.9).

Table P.3: **Gate-failure diagnostic.** Mean per-sample gate activation $g_k(\mathbf{X})$ averaged across experts, seeds (3), and test windows for each architecture–training-recipe combination. Across the three rows, all six datasets converge to $g \approx 0.5$, indicating the gate provides no per-window discriminative signal at 15 epochs.

| Config | ETTh1 | ETTh2 | ETTm1 | ETTm2 | Weather | Elec. |
|---|---|---|---|---|---|---|
| Dual-stream ($\alpha_k$) | 0.49 | 0.50 | 0.49 | 0.49 | 0.49 | 0.49 |
| Residual-IA (gate-init $b$=0) | 0.47 | 0.50 | 0.48 | 0.49 | 0.49 | 0.47 |
| Residual-IA + cos+wd ($b$=0) | 0.50 | 0.51 | 0.49 | 0.52 | 0.50 | 0.49 |

With 15 epochs of training, the gate mechanism extracts no gradient signal about *which branch should dominate* because both branches receive co-equal credit for reducing loss from the start. This is isomorphic to the dual-stream $\alpha_k \approx 0.49$ observation: the same plateau-at-0.5 behavior shows up consistently across both architectures we tested with symmetric initialization, suggesting it is not a quirk of a single architecture.

**Three surgical fixes.** (i) **Residual formulation.** Re-express the expert as $\mathbf{y}_k = \text{Raw}_k(\mathbf{X}) + g_k(\mathbf{X}) \cdot \text{Adapter}_k(\mathbf{H})$, making the raw branch the *primary predictor* and the backbone an additive residual that must earn its contribution. (ii) **Gate-init bias** $b$=−2**.** Initialize the gate-head linear layer's bias at $-2$ so that $g_k \approx \sigma(-2)$=0.12 at epoch 0: the backbone contributes only 12% initially. The model is *near-DLinear at initialization* and the backbone must earn any opening of the gate. (iii) 5**-epoch raw-branch warmup.** For epochs 1–5, freeze both the backbone branch and the gate head so only the raw branch and router train. The raw branch converges to a DLinear-quality base predictor; then at epoch 6 the backbone+gate unfreeze and must add value as a residual on top of an already-good base.

We additionally simplify the raw branch to a single linear layer (Linear(512→96), "$d1$"), which is structurally identical to our DLinear anchor (the single-Linear variant without trend-seasonal decomposition), and add cosine LR decay with weight decay $10^{-4}$ for stable long-horizon training.

Table P.4: **Residual-IA (intermediate)** (MOMENT-small, frozen, Top-2, H=96; Welch $t$-test vs. DLinear). Achieves **4/6 match-or-beat**: one sig win (Weather), one mean win (ETTm2), two parity. Avg gap $+1.6\%$. Table P.5 shows the improved Residual-IA$^+$ (5/6). The DLinear anchor here uses seeds 42–44, the anchor of the development runs that follow (Phases 1–3); the final Residual-IA$^+$ tables use the $n$=10 anchor.

| Dataset | DLinear | Residual-IA (ours) | Gap | $p$ | Verdict |
|---|---|---|---|---|---|
| ETTh1 | $0.417 \pm 0.003$ | $0.438 \pm 0.007$ (n=10) | $+5.1\%$ | 0.0001 | sig loss |
| ETTh2 | $0.341 \pm 0.014$ | $0.377 \pm 0.018$ (n=10) | $+10.6\%$ | 0.019 | sig loss |
| ETTm1 | $0.322 \pm 0.005$ | $0.328 \pm 0.007$ (n=10) | $+1.8\%$ | 0.190 | **parity** |
| ETTm2 | $0.200 \pm 0.010$ | $\mathbf{0.197 \pm 0.004}$ (n=5) | $-1.6\%$ | 0.622 | **win (mean)** |
| Weather | $0.208 \pm 0.004$ | $\mathbf{0.195 \pm 0.004}$ (n=5) | $-6.6\%$ | 0.007* | **sig win** |
| Electricity | $0.158 \pm 0.002$ | $0.159 \pm 0.002$ (n=5) | $+0.6\%$ | 0.457 | **parity** |
| **Average gap** | — | — | $+\mathbf{1.6}\%$ | — | **4/6 match-or-beat** |

**Why it works mechanistically.** The symmetric $\sigma(0)$=0.5 initialization treats the two branches as a priori equally reliable. For LTSF benchmarks where the raw signal contains most of the predictive structure (DLinear's success is evidence of this), this prior is wrong. The asymmetric $b$=$-2$ initialization encodes the correct inductive bias: *"default to a linear projection of the raw signal; invoke the backbone only if it reduces the loss."* The 5-epoch warmup separates the learning phases: the raw branch does not have to compete with a half-trained backbone for the same gradient signal during the critical early epochs. Each fix alone yields only marginal gains (see below); their combination achieves $4/6$ match-or-beat.

**Ablation: which fix does how much?** We isolate the contribution of each intervention (each row $n$=3 seeds, otherwise identical to Residual-IA):

| Variant | ETTh1 | ETTh2 | ETTm1 | ETTm2 | Weather | Elec. | Wins |
|---|---|---|---|---|---|---|---|
| Dual-stream (baseline) | 0.453 | 0.452 | 0.358 | 0.252 | **0.198** | 0.171 | 1 |
| + residual + IA gate ($b$=0) | 0.505 | 0.520 | 0.393 | 0.231 | 0.210 | 0.170 | 0 |
| + cos + wd ($b$=0) | 0.460 | 0.398 | 0.348 | **0.194** | 0.190 | 0.163 | 2 |
| + linear raw ($b$=0) | 0.436 | 0.375 | 0.325 | 0.202 | 0.193 | 0.160 | 1$^\star$ |
| + $b$=$-2$ | 0.432 | 0.388 | 0.331 | 0.224 | 0.191 | 0.160 | 1$^\star$ |
| **+ warmup = 5 (final)** | **0.437** | **0.382** | **0.327** | 0.195 | 0.193 | **0.160** | **2** |

$^\star$Column "Wins" counts datasets where mean MSE $<$ DLinear. Throughout this appendix, a cell is *match-or-beat* when its mean MSE is lower than the anchor's or exceeds it by less than 2% (parity).

Each of the three fixes alone produces 0–2 wins; the final row produces 2 wins and 2 parity cells, $4/6$ match-or-beat. This is consistent with the mechanistic explanation: the residual formulation, the asymmetric initialization, and the warmup are all solving the same root-cause problem (symmetric gradient competition between branches) from complementary angles.

### Residual-IA$^+$: three orthogonal levers

**Design methodology: hypothesis-driven decomposition.** Residual-IA leaves two statistically significant losses: ETTh1 ($+5.1\%$, $p$<0.001) and ETTh2 ($+10.6\%$, $p$=0.019). Two competing hypotheses could explain them:

- **H1 (data scarcity).** ETTh1/ETTh2 have 8,640 raw samples each versus 34,560 for ETTm*, and a 5,000-window cap limits the adapter training set to $\approx 9\%$ of available windows for ETTh but $\approx 2\%$ for ETTm. The ETTh pair may be overly exposed to sampling noise.
- **H2 (parameter scarcity).** Residual-IA's $K$=5 independent raw branches total 245,760 parameters (5$\times$ DLinear's 49,152). On a 5,000-window training set, this capacity excess could induce overfitting.

The two hypotheses call for opposite interventions: H1 predicts that *more data* (raising the cap to 100,000) closes the gap; H2 predicts that *fewer parameters* (sharing the raw branch across experts to reach DLinear's 49K budget) closes it. We ran both as single-lever experiments in parallel (5 seeds each).

**Phase 1: hypothesis falsification.** *Cap removal fails.* With the cap raised to $100{,}000$, the ETTh2 gap widens from $+10.6\%$ to $+26\%$. The additional training data at the same 15-epoch, lr$=10^{-3}$, wd$=10^{-4}$ recipe produces $11\times$ more update steps, over-regularizing through weight decay and preventing cosine-schedule convergence. This rules out H1 under the original training recipe; an extended-schedule retest (cosine + weight-decay sweep over the larger sample budget) would be needed to rule it out under all training recipes. *Shared raw branch succeeds.* At $K=5$ sharing a single Linear$(512, 96)$ module, raw parameters drop from 245K to 49K, exactly matching DLinear's budget. ETTh1 closes from $+5.1\%$ (sig loss) to $+1.6\%$ (marginal, $p=0.086$); ETTh2 closes from $+10.6\%$ (sig loss) to $+7.0\%$ (marginal, $p=0.078$). Per-sample gates and distinct backbone branches remain per-expert, so expert specialization is preserved in *how* the backbone residual is invoked; only the linear base predictor is shared. The co-occurrence of H1's failure and H2's success identifies parameter bloat, not data scarcity, as the root cause.

Phase 1 resolves the architectural failure but leaves ETTh1/ETTh2 still marginally above DLinear. Phase 2 tests two orthogonal refinements motivated by properties specific to these datasets.

**Phase 2: orthogonal refinements.** *Non-stationarity.* Hourly ETT data exhibits pronounced per-window level drift: the train/test windows have different means. Standard LTSF practice addresses this with NLinear: $\mathbf{y} = \text{Linear}(\mathbf{x} - x_T) + x_T$, where $x_T$ is a gradient-free anchor (the last input value). The linear layer is thereby constrained to model *relative motion*, not absolute level. Using NLinear as a drop-in replacement for the raw branch reduces ETTh1 from $+5.1\%$ to $+3.3\%$ and ETTh2 from $+10.6\%$ to $+6.9\%$ at $n=3$ seeds, while *improving* (not regressing) ETTm2 ($-1.6\% \to -4.1\%$) and Weather ($-6.6\% \to -8.7\%$). The improvement is largest where non-stationarity is most pronounced, matching NLinear's design rationale. *Per-dataset epoch mismatch.* A fixed 15-epoch schedule assumes all datasets converge at the same rate. We test this by running validation-based early stopping with patience 5 and logging the best epoch per seed. Convergence epochs vary by dataset: ETTh1 at $7.2\pm1.2$, ETTh2 at $8.6\pm2.8$, ETTm1 at $14.8\pm3.7$, Weather at $8.6\pm1.2$, Electricity at $20.2\pm4.0$. No single fixed budget fits all six datasets: the 15-epoch schedule simultaneously over-fits the ETTh pair and under-fits Electricity by 5–10 epochs. Val-based early stopping (with gradient clipping at $\|g\|_2 \leq 1.0$ to stabilize the first epoch) achieves dataset-adaptive budgets without per-dataset hyperparameter tuning. In isolation this drops Electricity from $+0.6\%$ to $0.0\%$ (perfect parity, $p=0.991$) and Weather from $-6.6\%$ to $-4.6\%$ (nearing significance).

**Phase 3: stacking validation.** Each Phase 1 and Phase 2 lever alone helps a different subset of datasets: shared raw closes ETTh1/h2 but trades small regressions on ETTm2; NLinear improves the ETTh pair but regresses ETTm1; val-early-stop perfects Electricity but hurts the ETTh pair by stopping too early. Each targets a different failure mode (parameter bloat, non-stationarity, training-length mismatch), and each mode is present on a different subset of the benchmark. **Residual-IA$^+$** stacks all three:

- **(iv)** *Shared raw branch*: single Linear$(512, 96)$ shared across $K$ experts.
- **(v)** *NLinear raw architecture*: $\mathbf{y} = \text{Linear}(\mathbf{x} - x_T) + x_T$.
- **(vi)** *Validation-based early stopping*: patience 5, gradient clipping $\|g\|_2 \leq 1.0$.

Table P.5 reports the composed result against two from-scratch anchors, each run for $n=10$ seeds: DLinear, in the simplified single-Linear variant $\text{Linear}(512 \to H)$ without trend-seasonal decomposition used throughout this paper, and NLinear, $\text{Linear}(\mathbf{x} - x_T) + x_T$. Residual-IA$^+$ matches or beats NLinear on $5/6$ datasets (mean gap $-1.6\%$) and DLinear on $5/6$ (mean gap $-2.8\%$), and the stronger anchor of each dataset on $4/6$ (mean gap $-0.8\%$).

**Component ablation for Residual-IA$^+$.** Each of the three additional fixes targets a different failure mode (raw-branch parameter bloat, non-stationarity, per-dataset training-length mismatch) and contributes complementary gains. Single-lever results ($n=5$ seeds except where noted):

| Variant | ETTh1 | ETTh2 | ETTm1 | ETTm2 | Weather | Elec. |
|---|---|---|---|---|---|---|
| Residual-IA (base) | 0.438 | 0.377 | 0.328 | 0.197 | 0.195 | 0.159 |
| + shared raw only | 0.423 | 0.365 | 0.321 | 0.202 | 0.197 | 0.160 |
| + NLinear only ($n=3$) | 0.430 | 0.365 | 0.332 | 0.192 | 0.190 | 0.159 |
| + val-early-stop only | 0.440 | 0.375 | 0.330 | 0.199 | 0.199 | 0.158 |
| **+ all three (Residual-IA$^+$)** | **0.424** | **0.349** | **0.326** | **0.187** | **0.196** | **0.156** |

Table P.5: **Residual-IA$^+$ at H=96 against both linear anchors** (MOMENT-small, strictly frozen, Top-2; $n$=10 seeds for Residual-IA$^+$, from its $n$=10 grid runs, and for each anchor, from the $n$=10 anchor sweep; Welch two-sided $t$-test, $p$ in parentheses). DLinear: single-Linear variant $\mathrm{Linear}(512 \to H)$ without trend-seasonal decomposition; NLinear: $\mathrm{Linear}(\mathbf{x} - x_T) + x_T$. Bold: significant win; $^\dagger$: significant loss ($p$<0.05).

| Dataset | DLinear | NLinear | Res-IA$^+$ | Gap vs. DLinear ($p$) | Gap vs. NLinear ($p$) |
|---|---|---|---|---|---|
| ETTh1 | $0.419 \pm 0.004$ | $0.429 \pm 0.013$ | $0.428 \pm 0.012$ | $+2.2\%^\dagger$ (0.037) | $-0.2\%$ (0.86) |
| ETTh2 | $0.352 \pm 0.020$ | $0.338 \pm 0.013$ | $0.346 \pm 0.010$ | $-1.6\%$ (0.44) | $+2.3\%$ (0.14) |
| ETTm1 | $0.326 \pm 0.008$ | $0.326 \pm 0.007$ | $0.327 \pm 0.007$ | $+0.4\%$ (0.68) | $+0.5\%$ (0.59) |
| ETTm2 | $0.211 \pm 0.017$ | $0.194 \pm 0.010$ | $0.188 \pm 0.002$ | $\mathbf{-10.8\%}$ (0.002) | $-3.2\%$ (0.075) |
| Weather | $0.207 \pm 0.004$ | $0.207 \pm 0.007$ | $0.197 \pm 0.003$ | $\mathbf{-5.1\%}$ (<0.001) | $\mathbf{-5.0\%}$ (<0.001) |
| Electricity | $0.159 \pm 0.002$ | $0.163 \pm 0.004$ | $0.156 \pm 0.001$ | $\mathbf{-1.7\%}$ (0.002) | $\mathbf{-3.9\%}$ (0.001) |
| **Average gap** | — | — | — | $-2.8\%$ | $-1.6\%$ |
| **Match-or-beat** | — | — | — | 5/6 | 5/6 |

The shared raw branch alone is the largest single contributor: it singlehandedly pulls ETTh1 from $+5.1\%$ to $+1.6\%$ (marginal, $p$=0.086) and ETTh2 from $+10.6\%$ to $+7.0\%$ (marginal, $p$=0.078), confirming the overfitting diagnosis. NLinear adds $\sim 3$ percentage points on ETTh2 ($0.365 \to 0.349$ when combined with shared). Val-early-stop contributes the Electricity perfect-parity result ($0.158 \to 0.156$) and the ETTm2 improvement ($0.197 \to 0.187$) by restoring an earlier checkpoint before the backbone residual begins overfitting.

**ETTh1 and ETTh2 at H=96.** Each anchor leaves one H=96 dataset outside the 2% band, and the two anchors disagree on which. Against DLinear it is ETTh1 ($+2.2\%$, $p$=0.037; absolute difference 0.009), which is a mean win against NLinear ($-0.2\%$, $p$=0.86). Against NLinear it is ETTh2 ($+2.3\%$, $p$=0.14), a mean win against DLinear ($-1.6\%$). At H=192, ETTh1 is within the band against both anchors ($+1.6\%$, $p$=0.18; $+0.6\%$, $p$=0.74).

**Scope of the accuracy claim.** We score all 123 (backbone, dataset, horizon) cells of Tables P.6 and P.8 against both $n$=10 anchors (Table P.9). Residual-IA$^+$ matches or beats NLinear in $96/123$ cells and DLinear in $102/123$. Against NLinear, the count is $27/33$ at H=96, $25/30$ at H=192, $25/30$ at H=336, and $19/30$ at H=720. Welch-significant wins and losses number 41:11 against NLinear and 44:12 against DLinear.

### Multi-horizon results on MOMENT-small

The Residual-IA$^+$ result of Table P.5 is computed at H=96. We extend it to the standard LTSF horizons $H \in \{96, 192, 336, 720\}$ (Table P.6). The two anchors diverge with the horizon on ETTh2 and ETTm2: at H=720 on ETTh2, Residual-IA$^+$ (0.584) is 39.6% below DLinear (0.966, $p$=0.003) and 17.7% above NLinear (0.496, $p$=0.31).

Table P.6: **Multi-horizon Residual-IA$^+$ against both anchors** (MOMENT-small, frozen; Residual-IA$^+$ $n$=10 seeds at H$\in\{96, 192\}$ and $n$=5 at H$\in\{336, 720\}$; each anchor $n$=10). Each cell gives the gap to DLinear / the gap to NLinear in %; bold: significant win, $^\dagger$: significant loss (Welch, $p$<0.05). Match-or-beat (mean lower or gap below 2%): $20/24$ against DLinear, $18/24$ against NLinear.

| Dataset | H=96 | H=192 | H=336 | H=720 | MoB DL | MoB NL |
|---|---|---|---|---|---|---|
| ETTh1 | $+2.2^\dagger$ / $-0.2$ | $+1.6$ / $+0.6$ | $+4.4$ / $+5.3$ | $+12.8^\dagger$ / $+21.6^\dagger$ | 1/4 | 2/4 |
| ETTh2 | $-1.6$ / $+2.3$ | $-12.2$ / $+8.4^\dagger$ | $\mathbf{-16.6}$ / $+4.9$ | $\mathbf{-39.6}$ / $+17.7$ | 4/4 | 0/4 |
| ETTm1 | $+0.4$ / $+0.5$ | $+1.0$ / $+0.0$ | $+0.7$ / $+1.3$ | $+2.4^\dagger$ / $-0.2$ | 3/4 | 4/4 |
| ETTm2 | $\mathbf{-10.8}$ / $-3.2$ | $\mathbf{-8.7}$ / $-1.0$ | $\mathbf{-13.6}$ / $\mathbf{-4.3}$ | $\mathbf{-27.0}$ / $-1.9$ | 4/4 | 4/4 |
| Weather | $\mathbf{-5.1}$ / $\mathbf{-5.0}$ | $\mathbf{-4.0}$ / $\mathbf{-6.4}$ | $\mathbf{-2.9}$ / $\mathbf{-4.5}$ | $-0.4$ / $\mathbf{-4.4}$ | 4/4 | 4/4 |
| Electricity | $\mathbf{-1.7}$ / $\mathbf{-3.9}$ | $-0.5$ / $\mathbf{-2.8}$ | $-0.9$ / $\mathbf{-2.2}$ | $+1.6^\dagger$ / $\mathbf{-5.4}$ | 4/4 | 4/4 |
| **Match-or-beat** | 5/6 / 5/6 | 6/6 / 5/6 | 5/6 / 4/6 | 4/6 / 4/6 | **20/24** | **18/24** |

Against DLinear, MOMENT-small matches or beats $20/24$ cells; against NLinear, $18/24$ (5/6, 5/6, 4/6, 4/6 from H=96 to H=720). Weather, ETTm2, and Electricity match or beat both anchors at all four horizons.

**Cross-backbone results**

Residual-IA$^+$ was developed on MOMENT-small, which uses RevIN instance normalization. To test whether the recipe is RevIN-specific or transfers across normalization regimes, we re-run the full Residual-IA$^+$ config on five other backbones: Moirai (RMSNorm internally with non-learnable I/O scaling, no learnable-affine RevIN inside encoder), Moirai-MoE (the same scheme with sparse internal experts), Chronos (T5 encoder-decoder, no instance norm), MOMENT-large (RevIN, 24 layers vs 8), and Timer-XL (decoder-only, LayerNorm only, no instance norm).

Table P.7: **Cross-backbone Residual-IA$^+$ at H$=96$ against both anchors** (strictly frozen, Top-2, $n=5$ seeds; each anchor $n=10$). Gap in %; bold: significant win, $^\dagger$: significant loss (Welch, $p<0.05$). Match-or-beat: $22/27$ against DLinear and $22/27$ against NLinear.

| Backbone | Anchor | ETTh1 | ETTh2 | ETTm1 | ETTm2 | Weather | Elec | MoB |
|---|---|---|---|---|---|---|---|---|
| Moirai | vs. DLinear | $+2.1$ | $+6.1$ | $+2.1$ | $\mathbf{-8.9}$ | $\mathbf{-6.4}$ | $+2.5^\dagger$ | $2/6$ |
| | vs. NLinear | $-0.3$ | $+10.3$ | $+2.2$ | $-1.1$ | $\mathbf{-6.3}$ | $+0.1$ | $4/6$ |
| Chronos | vs. DLinear | $+0.2$ | $-0.0$ | $+3.3$ | $\mathbf{-10.6}$ | $\mathbf{-3.7}$ | $\mathbf{-2.4}$ | $5/6$ |
| | vs. NLinear | $-2.2$ | $+3.9$ | $+3.4$ | $-3.0$ | $\mathbf{-3.6}$ | $\mathbf{-4.6}$ | $4/6$ |
| MM-large | vs. DLinear | $+1.9^\dagger$ | $-1.3$ | $-0.4$ | $\mathbf{-10.7}$ | $\mathbf{-4.9}$ | $\mathbf{-1.5}$ | $6/6$ |
| | vs. NLinear | $-0.6$ | $+2.7$ | $-0.3$ | $-3.1$ | $\mathbf{-4.8}$ | $\mathbf{-3.8}$ | $5/6$ |
| Timer-XL | vs. DLinear | $+0.8$ | — | $-2.0$ | — | $\mathbf{-4.4}$ | — | $3/3$ |
| | vs. NLinear | $-1.6$ | — | $-1.9$ | — | $\mathbf{-4.3}$ | — | $3/3$ |
| Moirai-MoE | vs. DLinear | $+1.0$ | $-2.5$ | $+0.2$ | $\mathbf{-12.2}$ | $\mathbf{-4.7}$ | $\mathbf{-2.5}$ | $6/6$ |
| | vs. NLinear | $-1.4$ | $+1.4$ | $+0.3$ | $\mathbf{-4.6}$ | $\mathbf{-4.6}$ | $\mathbf{-4.7}$ | $6/6$ |

At H$=96$, Residual-IA$^+$ matches or beats DLinear in $22/27$ cross-backbone cells and NLinear in $22/27$. On Weather and ETTm2, every tested backbone matches or beats both anchors. The five backbones span three normalization regimes: learnable-affine RevIN inside the encoder (MOMENT-large), RMSNorm-internal with non-learnable I/O scaling (Moirai, Moirai-MoE), and no instance normalization (Chronos, Timer-XL).

**Multi-horizon grid.** We extend to the full 5 backbones $\times$ 6 datasets $\times$ 4 horizons grid, scored against DLinear in Table P.8; Table P.9 gives the counts against both anchors.

Against DLinear: (1) at H$=96$, 22 of the 27 tested (backbone, dataset) cells match or beat the anchor; (2) MOMENT-large reaches $\mathbf{23/24}$; (3) Moirai-MoE achieves $\mathbf{15/15}$.

**Both anchors, all cells.** Table P.9 scores the full grid against both $n=10$ anchors. We lead with NLinear, the anchor that yields the lower count ($96/123$). The last column uses the anchor with the lower mean MSE in each cell. This anchor is selected after observing the results, so we report this column for completeness only.

**Off-suite datasets.** On three datasets outside the six-dataset suite (MOMENT-small; Residual-IA$^+$ and both anchors are trained together in each run, $n=10$ seeds; Table P.10), Residual-IA$^+$ beats both anchors on Solar in all 10 seeds, with mean gaps of 8.7–9.0% at H$=96$ and 11.2–13.1% at H$=192$, and matches or beats both on Traffic, with mean gaps of 0.3–2.3% and 0.9–3.6%. On Exchange it is at parity with NLinear at H$=96$ ($+0.5\%$, $p=0.85$; $-26.3\%$ against DLinear).

The next subsection asks whether Residual-IA$^+$'s result holds when the external Conv1d router is replaced by SR-MoA-style per-expert self-gates (§4.2, Table F.2); the subsequent one isolates the TSFM's own contribution by stripping the backbone entirely.

### P.2 SR-RIA$^+$: Self-Routed Residual-IA$^+$

SR-RIA$^+$ combines two independently validated components: (i) SR-MoA self-routing, where each expert has its own sigmoid gate on raw input (no external router; Table F.2), and (ii) Residual-IA$^+$'s dual-stream expert architecture (shared NLinear raw branch + gated backbone residual; Table P.5). Each expert $k$ produces $\mathbf{y}_k = \text{NLinear}(\mathbf{X}_{\text{raw}}) + \sigma(\text{blend}_k(\mathbf{X}_{\text{raw}})) \cdot \text{Adapter}_k(\mathbf{H})$, where the NLinear

Table P.8: **Multi-horizon cross-backbone Residual-IA$^+$ against the $n$=10 DLinear anchor** (5 backbones $\times$ up to 6 datasets $\times$ 4 horizons, $n$=3–5 seeds; gap in %, Welch two-sided $t$-test). $*$=mean win, $*$=sig win ($p<0.05$), $=$ =parity (gap below 2%), X=sig loss beyond 2%. Match-or-beat: MOMENT-large 23/24, Timer-XL 12/12, Moirai-MoE 15/15, Chronos 21/24, Moirai 11/24; pooled 82/99. Counts against NLinear: Table P.9.

| Backbone | Dataset | H=96 | H=192 | H=336 | H=720 | MoB |
|---|---|---|---|---|---|---|
| Moirai | ETTh1 | +2.1 | = +1.4 | X+8.8 | X+19.0 | 1/4 |
| | ETTh2 | +6.1 | +2.5 | +43.3 | +25.8 | 0/4 |
| | ETTm1 | +2.1 | $*$−0.5 | $*$−0.9 | +2.1 | 2/4 |
| | ETTm2 | $*$−**8.9** | $*$−**8.5** | $*$−**11.9** | $*$−18.8 | 4/4 |
| | Weather | $*$−**6.4** | $*$−**2.5** | $*$−0.8 | = +0.7 | 4/4 |
| | Elec | X+2.5 | X+6.5 | X+5.9 | X+7.0 | 0/4 |
| Chronos | ETTh1 | = +0.2 | $*$−0.5 | $*$−0.0 | = +1.8 | 4/4 |
| | ETTh2 | $*$−0.0 | $*$−11.1 | $*$−**19.9** | +76.4 | 3/4 |
| | ETTm1 | +3.3 | = +0.9 | $*$−2.0 | $*$−0.2 | 3/4 |
| | ETTm2 | $*$−**10.6** | $*$−**8.7** | $*$−**8.6** | $*$−17.9 | 4/4 |
| | Weather | $*$−**3.7** | $*$−**3.4** | $*$−1.2 | $*$−0.7 | 4/4 |
| | Elec | $*$−**2.4** | = +0.1 | $*$−0.9 | X+2.1 | 3/4 |
| MM-large | ETTh1 | = +1.9 | = +0.2 | = +0.2 | +7.2 | 3/4 |
| | ETTh2 | $*$−1.3 | $*$−**15.0** | $*$−**19.8** | $*$−**48.0** | 4/4 |
| | ETTm1 | $*$−0.4 | = +0.3 | = +0.0 | = +0.9 | 4/4 |
| | ETTm2 | $*$−**10.7** | $*$−**10.2** | $*$−**12.2** | $*$−**26.7** | 4/4 |
| | Weather | $*$−**4.9** | $*$−**4.0** | $*$−**4.1** | $*$−**2.2** | 4/4 |
| | Elec | $*$−**1.5** | $*$−0.8 | $*$−0.7 | = +0.9 | 4/4 |
| Timer-XL | ETTh1 | = +0.8 | $*$−0.1 | = +0.9 | $*$−2.4 | 4/4 |
| | ETTm1 | $*$−2.0 | $*$−**2.3** | $*$−1.7 | $*$−0.4 | 4/4 |
| | Weather | $*$−**4.4** | $*$−**2.8** | $*$−**2.8** | $*$−0.8 | 4/4 |
| Moirai-MoE | ETTh1 | = +1.0 | $*$−1.5 | $*$−0.7 | $*$−0.4 | 4/4 |
| | ETTh2 | $*$−2.5 | — | — | — | 1/1 |
| | ETTm1 | = +0.2 | $*$−1.2 | $*$−1.3 | = +0.3 | 4/4 |
| | ETTm2 | $*$−**12.2** | — | — | — | 1/1 |
| | Weather | $*$−**4.7** | $*$−**3.2** | $*$−**2.5** | $*$−**1.7** | 4/4 |
| | Elec | $*$−**2.5** | — | — | — | 1/1 |
| **Per-horizon MoB** | | 22/27 | 22/24 | 21/24 | 17/24 | **82/99** |

Table P.9: **Residual-IA$^+$ against both linear anchors, all 123 cells** (6 backbones $\times$ up to 6 datasets $\times$ 4 horizons; cells of Tables P.6 and P.8; each anchor $n$=10 seeds). Entries count match-or-beat cells (mean MSE lower than the anchor's, or above it by less than 2%). ‡Anchor chosen per cell after seeing the results. Last row: Welch-significant ($p<0.05$) wins : losses.

| | Cells | vs. NLinear | vs. DLinear | vs. stronger per cell‡ |
|---|---|---|---|---|
| MOMENT-small | 24 | 18 | 20 | 16 |
| MOMENT-large | 24 | 21 | 23 | 21 |
| Moirai | 24 | 14 | 11 | 10 |
| Moirai-MoE | 15 | 14 | 15 | 14 |
| Chronos | 24 | 18 | 21 | 17 |
| Timer-XL | 12 | 11 | 12 | 11 |
| H=96 | 33 | 27 | 27 | — |
| H=192 | 30 | 25 | 28 | — |
| H=336 | 30 | 25 | 26 | — |
| H=720 | 30 | 19 | 21 | — |
| **Pooled** | **123** | **96** | **102** | **89** |
| Sig. wins : losses | — | 41:11 | 44:12 | — |

Table P.10: **Off-suite datasets** (MOMENT-small, Residual-IA$^+$, frozen; Residual-IA$^+$ and both anchors are trained together in each run, $n$=10 seeds). Gap to DLinear / gap to NLinear in %; bold: significant win (Welch, $p<0.05$).

| Dataset | H=96 | H=192 |
|---|---|---|
| Solar | $\mathbf{-9.0}$ / $\mathbf{-8.7}$ | $\mathbf{-11.2}$ / $\mathbf{-13.1}$ |
| Traffic | $-0.3$ / $\mathbf{-2.3}$ | $\mathbf{-0.9}$ / $\mathbf{-3.6}$ |
| Exchange | $\mathbf{-26.3}$ / $+0.5$ | $-9.5$ / $+11.1$ |

branch is shared across experts and the blend gate is initialized at $b=-2$ ($\sigma(-2)\approx 0.12$). Training uses 5-epoch raw-branch warmup, cosine LR decay, and validation-based early stopping with patience 5.

Table P.11: **SR-RIA$^+$ against both linear anchors** (MOMENT-small, frozen, $K$=5, $H$=96; SR-RIA$^+$ 5 seeds, each anchor $n$=10). SR-RIA$^+$ matches or beats both anchors on $6/6$ datasets (mean gap $-3.3\%$ against DLinear, $-2.2\%$ against NLinear).

| Dataset | SR-RIA$^+$ | DLinear | NLinear | RIA$^+$ | $\Delta$ vs DL | $\Delta$ vs NL |
|---|---|---|---|---|---|---|
| ETTh1 | $0.423 \pm 0.004$ | 0.419 | 0.429 | 0.428 | $+1.0\%$ | $-1.4\%$ |
| ETTh2 | $0.345 \pm 0.005$ | 0.352 | 0.338 | 0.346 | $-2.0\%$ | $+1.9\%$ |
| ETTm1 | $0.321 \pm 0.004$ | 0.326 | 0.326 | 0.327 | $-1.6\%$ | $-1.5\%$ |
| ETTm2 | $\mathbf{0.188 \pm 0.001}$ | 0.211 | 0.194 | 0.188 | $\mathbf{-10.9}\%$ | $-3.2\%$ |
| Weather | $\mathbf{0.192 \pm 0.002}$ | 0.207 | 0.207 | 0.197 | $\mathbf{-7.1}\%$ | $\mathbf{-7.0}\%$ |
| Electricity | $0.160 \pm 0.000$ | 0.159 | 0.163 | 0.156 | $+0.5\%$ | $-1.7\%$ |
| **Average gap** | — | — | — | — | $\mathbf{-3.3}\%$ | $\mathbf{-2.2}\%$ |

### P.3 Pure Raw-MLP MoE Ablation: Quantifying the TSFM's Contribution

A potential concern about the dual-stream result is that it could be confounded: if the TSFM destroys nonlinear temporal structure (Appendix M), and the dual-stream fix recovers that structure via the raw branch, perhaps the raw branch is doing all the work and the TSFM is dead weight. We tested this hypothesis directly by running a pure mixture-of-experts ablation in which the TSFM is removed entirely.

**Architecture.** The Pure Raw-MLP MoE consists of $K$=5 size-diverse two-layer MLPs operating directly on the 512-dimensional raw input window: hidden widths $\{64, 96, 128, 192, 256\}$ chosen so that the total parameter count (449,557) matches RR-MoA's 426K within 6%. Each expert is Linear($512 \to h_k$) $\to$ GELU $\to$ Linear($h_k \to 96$). The router is the same Conv1d + AdaptiveAvgPool + Linear architecture as RR-MoA, also operating on the raw input (not on hidden states). Top-2 sparse routing is used. Training protocol is identical to RR-MoA: Adam $1\times10^{-3}$, 15 epochs, batch 128. **No TSFM forward pass executes anywhere in this architecture**: the $\sim$40M-parameter MOMENT-small backbone is removed entirely.

Table P.12: **Pure Raw-MLP MoE vs. Dual-Stream and DLinear** (5 seeds, mean$\pm$std, H=96). The Raw-MLP MoE has *no TSFM at all*. It *beats* Dual-Stream outright on **3** of **6** datasets (ETTh2, ETTm2, Electricity). Mean delta to Dual-Stream is $-1.5\%$: the TSFM's contribution is dataset-dependent and negligible in the average case.

| Dataset | Raw-MLP MoE | Dual-Stream | DLinear | $\Delta$ vs DS | $\Delta$ vs DL |
|---|---|---|---|---|---|
| ETTh1 | $0.483 \pm 0.019$ | *0.453* | *0.416* | $+6.6\%$ | $+16.1\%$ |
| **ETTh2** | $\mathbf{0.426 \pm 0.077}$ | *0.452* | *0.341* | $\mathbf{-5.7}\%$ | $+25.0\%$ |
| ETTm1 | $0.370 \pm 0.014$ | *0.358* | *0.322* | $+3.4\%$ | $+15.0\%$ |
| **ETTm2** | $\mathbf{0.207 \pm 0.004}$ | *0.252* | *0.200* | $\mathbf{-17.9}\%$ | $+3.4\%$ |
| Weather | $0.209 \pm 0.007$ | *0.198* | *0.208* | $+5.5\%$ | $+0.5\%$ |
| **Electricity** | $\mathbf{0.170 \pm 0.002}$ | *0.171* | *0.158* | $\mathbf{-0.7}\%$ | $+7.5\%$ |
| **Mean** | — | — | — | $\mathbf{-1.5}\%$ | $\mathbf{+11.2}\%$ |

**Interpretation.** Across the 6 datasets, the Raw-MLP MoE slightly outperforms Dual-Stream on average ($-1.5\%$), winning 3/6 datasets outright (ETTh2: $-5.7\%$; ETTm2: $-17.9\%$; Electricity:

$-0.7\%$). It loses by 3–7$\%$ on the three datasets with complex temporal dynamics (ETTh1, ETTm1, Weather). The TSFM is therefore not adding universal value to the dual-stream result; it adds dataset-dependent, modest value where complex dynamics exist, and contributes no measurable value on the linear-dominated datasets.

**Routing entropy diagnostics.** In all 30 Raw-MLP MoE runs, routing entropy saturates near $\log K = 1.609$ (range 1.20–1.60, mean $\approx 1.49$, i.e., $93\%$ of the uniform maximum), indicating the router degenerates to near-uniform mixing. Per-dataset means lie in $[1.31, 1.60]$: highest on Electricity (1.60, $99\%$ of $\log K$), lowest on ETTh2 (1.31, $81\%$). This means the Raw-MLP MoE is operating as a learned ensemble of 5 size-diverse MLPs rather than as a per-sample mixture of specialists; the diversity comes from the heterogeneous expert capacities rather than from per-window routing decisions. **Contrast with main RR-MoA.** Under the same $K{=}5$, Top-2, raw-router protocol but *with the frozen MOMENT-small backbone present*, main RR-MoA achieves mean entropy 1.34 on the same six datasets ($83\%$ of $\log K$, 0.15 nats lower than the backbone-free version), with per-dataset means spanning $[1.11, 1.44]$ (per-seed values span the broader $[0.93, 1.57]$ range reported in §4.2). The two most specialized routers (per-dataset mean) are on ETTh2 (1.11, $69\%$ of $\log K$) and ETTm2 (1.18, $73\%$); the least specialized on Electricity (1.44, $90\%$). Removing the backbone shifts entropy uniformly toward $\log K$ across all six datasets, quantifying the difference between "MoE-as-ensemble" (no backbone) and "MoE-as-specialist-mixture" (with backbone): the backbone induces enough hidden-state structure that the same Conv1d gate, on the same raw input, makes substantively non-uniform routing decisions. Figure F.2 corroborates this with expert-assignment scatter against amplitude/volatility quartiles.

**Implications for the paper's claims.** This finding does *not* undermine the central diagnosis of the paper: normalization-induced routing collapse on hidden-state routers (AdaMix) is a real, mechanistically-explained phenomenon, predicted by Observation 1 for any TSFM with internal RevIN and verified empirically on MOMENT-small/large; pre-normalization routing (RR-MoA) is the causal fix at the router input. That diagnosis is independent of whether the MOMENT backbone is the most efficient architecture for forecasting on every dataset: the mechanism predicts the same failure mode for any backbone with internal instance normalization that satisfies the bound's premise. What this finding *does* say is that the dual-stream result should be reframed: rather than "the TSFM has complementary nonlinear features that dual-stream extracts," the more accurate reading is "the TSFM contributes complementary signal on datasets with nonlinear dynamics (where it wins by 3–7$\%$), and contributes no measurable value on datasets where DLinear is already near-optimal (where pure raw-MLP MoE matches or beats it)."

**Summary.** Resolving the confound stated at the start of this section: at H$=$96 the TSFM adds no measurable value on ETTh2, ETTm2, and Electricity, where the backbone-free Raw-MLP MoE wins by 0.7–17.9$\%$ over Dual-Stream, and contributes 3–7$\%$ MSE on ETTh1, ETTm1, and Weather, where Dual-Stream retains the advantage. The dataset-conditional contribution of the TSFM is itself an empirical pattern worth flagging to the TSFM-design community: *frozen TSFMs are useful where forecasting requires nonlinear temporal modeling beyond what a 5-MLP ensemble can express, and unnecessary where it does not*. The single most important consequence for our paper is that the routing-collapse diagnosis (the core scientific contribution) is robust to whether the TSFM is the best backbone: it explains the AdaMix-collapse phenomenon and predicts the cross-backbone behavior regardless. At $H{=}192$, Residual-IA$^+$ has lower MSE than this Raw-MLP MoE in all 18 runs on the six datasets (3 seeds; mean $-11.8\%$, range $-0.3\%$ to $-34.3\%$). This comparison also changes the raw branch, so the measured gain cannot be attributed solely to the backbone. We report the backbone's contribution as dataset-dependent.

# Q Task, Input-Length and Horizon Generality

This appendix reports classification, anomaly-detection, input-length and long-horizon experiments. The tables include every RR-MoA run from these experiments. All runs use MOMENT-small and seeds $\{42, 43, 44\}$, with the backbone strictly frozen except in the router comparison of Appendix Q.2; RR-MoA always has $K{=}5$ experts, Top-2 routing and the Conv1d router. Routing entropy is in nats with ceiling $\log 5{\approx}1.609$; we call a run collapsed when its entropy falls below 0.3, the threshold of Table G.3.

### Q.1 Classification and Anomaly Detection

**Scope.** These runs check that RR-MoA's routing stays non-collapsed on two further task types. They do not establish competitive performance on these tasks. The frozen features cap both metrics: EthanolConcentration accuracy is $25.7$–$33.0\%$ against $25\%$ chance ($4$ classes), and SMD ROC-AUC is $0.510$–$0.595$.

**Classification protocol.** We use three datasets from the UEA archive [94]: BasicMotions, EthanolConcentration and JapaneseVowels. We combine each dataset's official train and test series. Each channel becomes a univariate sample with the label of its source series. We zero-pad or truncate each sample to $512$ steps and standardize it. We hold out a stratified $20\%$ of samples, using the same split for every seed. The five experts are the canonical pooling heads (mean, last, max, attention, Conv1d) with class-logit outputs, trained with cross-entropy for $20$ epochs (Adam, $\eta{=}10^{-3}$, batch $64$). We report held-out accuracy.

**Anomaly-detection protocol.** We use machines 1-1, 2-1 and 3-1 of the Server Machine Dataset (SMD) [95]. Every channel of the train and test series is cut into univariate windows of $512$ steps at stride $256$. RR-MoA uses the raw router to reconstruct each window from frozen hidden states. We train on the training windows for $15$ epochs (Adam, $\eta{=}10^{-3}$, batch $128$). The objective adds a load-balancing term with coefficient $0.01$ to the reconstruction loss. The anomaly score of a test window is its reconstruction MSE; a window is anomalous when it contains an anomalous step, and we report window-level ROC-AUC.

**Results.** Table Q.1 lists all $18$ runs. Routing entropy is $1.383$–$1.605$ on classification and $1.518$–$1.607$ on anomaly detection, so no run approaches collapse.

Table Q.1: **Classification and anomaly-detection routing check** (MOMENT-small, strictly frozen, RR-MoA $K{=}5$, Top-2; one column per seed). Metric: held-out accuracy (classification) or window-level ROC-AUC (SMD). Classification entropy is that of the routing distribution averaged over the held-out split; anomaly entropy is the mean per-window entropy, as in Tables Q.3 and Q.5.

| | | Routing entropy | | | Accuracy / ROC-AUC | | |
|---|---|---|---|---|---|---|---|
| Task | Dataset | 42 | 43 | 44 | 42 | 43 | 44 |
| Classification | BasicMotions (4 classes) | 1.575 | 1.596 | 1.605 | 0.896 | 0.792 | 0.906 |
| | EthanolConcentration (4) | 1.600 | 1.593 | 1.587 | 0.257 | 0.267 | 0.330 |
| | JapaneseVowels (9) | 1.462 | 1.526 | 1.383 | 0.277 | 0.296 | 0.317 |
| Anomaly (SMD) | machine-1-1 | 1.606 | 1.587 | 1.555 | 0.595 | 0.560 | 0.593 |
| | machine-2-1 | 1.607 | 1.604 | 1.597 | 0.528 | 0.510 | 0.532 |
| | machine-3-1 | 1.536 | 1.518 | 1.521 | 0.570 | 0.564 | 0.573 |

### Q.2 Normalized-Router Arm

To test whether the collapse mechanism transfers to these tasks, we compare raw routing with routing on hidden states from the normalized input. We therefore train two routers over the same five experts: RR-MoA's raw Conv1d router, and the hidden-state router of AdaMix and of the imputation control (App. J), mean-pooled $\mathbf{H} \to \text{Linear}(d, 32) \to \text{GELU} \to \text{Linear}(32, K) \to \text{softmax}$, with Top-2 routing. Hidden-state routing collapses in forecasting when backbone blocks are unfrozen (Table 2). We therefore unfreeze the last four encoder blocks in both arms and train them jointly with the adapter using one Adam optimizer ($\eta{=}10^{-3}$). Data, losses, epochs, batch sizes and seeds are those of Appendix Q.1, and entropy is measured as in Table Q.1. Only the anomaly runs carry the load-balancing term. On classification each sample is standardized before either router reads it.

Table Q.2 lists all $36$ runs. On classification, the hidden-state router collapses in $9/9$ runs. The entropy of its mean routing distribution is $0.000$–$0.209$, with almost all weight on one expert. The raw router stays non-collapsed ($1.369$–$1.606$). On anomaly detection, the hidden-state router collapses in $3/9$ runs, with mean per-window entropy of $0.000$–$1.449$. The raw router stays non-collapsed ($1.422$–$1.608$).

Table Q.2: **Normalized-router arm** (MOMENT-small, last four encoder blocks unfrozen, $K{=}5$ experts, Top-2; one column per seed). Raw: RR-MoA's Conv1d router on the input window. Hidden: AdaMix's router on mean-pooled $\mathbf{H}$. Entropy and metric as in Table Q.1; bold marks a collapsed run (entropy $<0.3$).

| | | | Routing entropy | | | Accuracy / ROC-AUC | | |
|---|---|---|---|---|---|---|---|---|
| Task | Dataset | Router | 42 | 43 | 44 | 42 | 43 | 44 |
| Classification | BasicMotions (4 classes) | raw | 1.604 | 1.598 | 1.600 | 0.917 | 0.854 | 0.865 |
| | | hidden | **0.103** | **0.004** | **0.007** | 0.875 | 0.927 | 0.917 |
| | EthanolConcentration (4) | raw | 1.600 | 1.606 | 1.369 | 0.375 | 0.327 | 0.400 |
| | | hidden | **0.000** | **0.000** | **0.209** | 0.311 | 0.406 | 0.368 |
| | JapaneseVowels (9) | raw | 1.514 | 1.538 | 1.594 | 0.354 | 0.348 | 0.355 |
| | | hidden | **0.008** | **0.000** | **0.000** | 0.312 | 0.350 | 0.342 |
| Anomaly (SMD) | machine-1-1 | raw | 1.608 | 1.511 | 1.498 | 0.613 | 0.581 | 0.593 |
| | | hidden | 0.699 | 0.3003 | **0.000** | 0.559 | 0.586 | 0.633 |
| | machine-2-1 | raw | 1.603 | 1.503 | 1.564 | 0.527 | 0.536 | 0.531 |
| | | hidden | 1.449 | **0.028** | 0.497 | 0.526 | 0.538 | 0.535 |
| | machine-3-1 | raw | 1.602 | 1.547 | 1.422 | 0.530 | 0.539 | 0.533 |
| | | hidden | 0.835 | 1.376 | **0.087** | 0.548 | 0.539 | 0.508 |

### Q.3 Input Length

**Protocol.** MOMENT-small is pretrained with a fixed 512-step context. To observe $L{<}512$ steps, we set the input mask to 0 on the oldest $512{-}L$ steps, and the raw router reads only the $L$ observed steps; $L{=}512$ is the standard pipeline. We run $L{\in}\{96, 192, 336, 512\}$ at $H{=}96$ on the six LTSF datasets for 15 epochs, 72 runs in total. Each run trains, on the same frozen backbone, RR-MoA with the raw router, RR-MoA with the RevIN router of Table F.1 (the same window after per-window normalization), and three fixed adapters (linear, attention, convolutional); the best fixed adapter of a run is the one with the lowest test MSE.

**Results.** Routing entropy stays between 0.89 and 1.53 in all 72 runs (Table Q.3). Averaged across seeds, the raw router's MSE is lower than the RevIN router's in all 24 (dataset, $L$) configurations, by 16.4–83.5%. RR-MoA beats the best fixed adapter in 70 of 72 runs (Table Q.4).

Table Q.3: **Input-length ablation: routing entropy** (MOMENT-small, strictly frozen, RR-MoA with the raw router, $K{=}5$, Top-2, $H{=}96$). Each cell lists seeds $42/43/44$.

| Dataset | $L{=}96$ | $L{=}192$ | $L{=}336$ | $L{=}512$ |
|---|---|---|---|---|
| ETTh1 | 1.31/1.39/1.39 | 1.41/1.48/1.42 | 1.46/1.51/1.46 | 1.45/1.45/1.49 |
| ETTh2 | 1.19/1.23/1.28 | 1.28/1.16/1.17 | 1.32/1.06/0.92 | 0.92/0.89/1.13 |
| ETTm1 | 1.34/1.39/1.45 | 1.36/1.42/1.39 | 1.25/1.52/1.39 | 1.48/1.46/1.37 |
| ETTm2 | 1.16/1.31/1.26 | 1.12/1.22/1.14 | 1.16/1.13/1.23 | 1.21/1.26/1.08 |
| Weather | 1.47/1.46/1.47 | 1.37/1.41/1.46 | 1.37/1.32/1.42 | 1.47/1.44/1.30 |
| Electricity | 1.40/1.38/1.44 | 1.53/1.44/1.43 | 1.48/1.48/1.38 | 1.50/1.48/1.40 |

### Q.4 Long Horizons

**RR-MoA.** We run $H{\in}\{1000, 2000\}$ with $L{=}512$ on ETTm2, Weather and Electricity, and $H{=}1000$ on ETTh2, for 15 epochs (21 runs). Each run compares RR-MoA (raw router) with the best of the three fixed adapters of Appendix Q.3, trained in the same run. RR-MoA has the lower MSE in all 21 runs, by 10.9–77.2%, and routing entropy stays between 0.83 and 1.58 (Table Q.5).

**Residual-IA$^+$.** On ETTm2, Weather and Electricity at both horizons (18 runs), we compare Residual-IA$^+$ (Appendix P.1) with the simplified single-Linear DLinear variant of Appendix A: Linear$(512{\to}H)$ on the raw window without the trend-seasonal decomposition, trained from scratch with Adam ($\eta{=}10^{-3}$, batch 128) for 15 epochs. Residual-IA$^+$ uses one fixed recipe: a shared NLinear raw branch, gate-bias initialization $b{=}{-}2$, a 5-epoch raw-branch warmup, cosine learning-rate decay, gradient clipping at 1.0, and validation early stopping (patience 5, at most 30 epochs). Residual-IA$^+$

Table Q.4: **Input-length ablation: test MSE** (same runs as Table Q.3; mean±std over seeds $42/43/44$). $\Delta_{\text{RevIN}}$ compares the seed means of the raw and RevIN routers. Best fixed is the seed mean of each run's best fixed adapter; $\Delta_{\text{fixed}}$ gives RR-MoA (raw router) against that adapter for seeds $42/43/44$ (%), and Wins counts the seeds where RR-MoA has the lower MSE.

| Dataset | $L$ | Raw router | RevIN router | $\Delta_{\text{RevIN}}$ | Best fixed | $\Delta_{\text{fixed}}$ (%) | Wins |
|---|---|---|---|---|---|---|---|
| ETTh1 | 96 | 0.617±0.029 | 1.057±0.028 | −41.6% | 1.054 | −37.6/−46.7/−39.0 | 3/3 |
| | 192 | 0.626±0.054 | 1.127±0.083 | −44.5% | 1.209 | −51.6/−52.7/−38.9 | 3/3 |
| | 336 | 0.655±0.119 | 1.259±0.084 | −48.0% | 1.313 | −53.2/−53.2/−44.5 | 3/3 |
| | 512 | 0.681±0.038 | 1.132±0.059 | −39.8% | 1.221 | −51.4/−39.5/−40.8 | 3/3 |
| ETTh2 | 96 | 0.799±0.113 | 3.187±0.244 | −74.9% | 3.048 | −70.4/−75.2/−76.0 | 3/3 |
| | 192 | 0.793±0.169 | 2.882±0.029 | −72.5% | 2.940 | −73.0/−67.9/−78.4 | 3/3 |
| | 336 | 0.920±0.145 | 2.717±0.040 | −66.1% | 2.831 | −74.3/−57.5/−69.1 | 3/3 |
| | 512 | 0.893±0.365 | 2.929±0.286 | −69.5% | 2.639 | −75.5/−46.2/−74.6 | 3/3 |
| ETTm1 | 96 | 0.597±0.044 | 0.992±0.026 | −39.8% | 1.029 | −41.4/−45.9/−38.3 | 3/3 |
| | 192 | 0.545±0.047 | 1.032±0.079 | −47.2% | 1.050 | −50.7/−51.1/−42.5 | 3/3 |
| | 336 | 0.598±0.021 | 1.149±0.112 | −48.0% | 1.089 | −47.0/−45.1/−43.3 | 3/3 |
| | 512 | 0.581±0.019 | 1.094±0.085 | −46.9% | 1.119 | −47.3/−50.6/−46.1 | 3/3 |
| ETTm2 | 96 | 0.513±0.121 | 3.072±0.020 | −83.3% | 3.069 | −85.8/−78.4/−85.5 | 3/3 |
| | 192 | 0.496±0.074 | 3.011±0.078 | −83.5% | 3.103 | −86.9/−82.2/−83.0 | 3/3 |
| | 336 | 0.705±0.331 | 2.901±0.067 | −75.7% | 3.000 | −83.8/−65.1/−80.9 | 3/3 |
| | 512 | 0.600±0.023 | 2.584±0.364 | −76.8% | 2.766 | −76.8/−79.6/−78.4 | 3/3 |
| Weather | 96 | 0.286±0.005 | 0.631±0.009 | −54.7% | 0.612 | −54.0/−53.9/−52.1 | 3/3 |
| | 192 | 0.263±0.003 | 0.560±0.029 | −53.1% | 0.605 | −56.9/−57.1/−55.7 | 3/3 |
| | 336 | 0.266±0.007 | 0.536±0.029 | −50.4% | 0.592 | −55.6/−56.5/−52.9 | 3/3 |
| | 512 | 0.270±0.021 | 0.507±0.051 | −46.7% | 0.577 | −57.3/−53.1/−47.4 | 3/3 |
| Electricity | 96 | 0.384±0.013 | 0.563±0.085 | −31.8% | 0.521 | −25.1/−29.0/−24.8 | 3/3 |
| | 192 | 0.407±0.107 | 0.537±0.023 | −24.2% | 0.503 | +5.0/−33.9/−28.5 | 2/3 |
| | 336 | 0.377±0.078 | 0.517±0.010 | −27.0% | 0.500 | −7.1/−35.1/−31.6 | 3/3 |
| | 512 | 0.417±0.099 | 0.499±0.019 | −16.4% | 0.620 | +6.1/−30.9/−56.5 | 2/3 |

has the lower MSE in 14 of 18 runs (Table Q.6). On ETTm2 it is lower in all 6 runs, by 18.5–58.0%. On Electricity and Weather the per-run gaps lie between −2.6% and +2.7%.

Table Q.5: **Long horizons: RR-MoA vs. best fixed adapter** (MOMENT-small, strictly frozen, $K=5$, Top-2, $L=512$, test MSE). Each cell lists seeds $42/43/44$; $\Delta$ gives RR-MoA against the run's best fixed adapter (%).

| Dataset | $H$ | RR-MoA | Best fixed | $\Delta$ (%) | Routing entropy | Wins |
|---|---|---|---|---|---|---|
| ETTh2 | 1000 | 2.068/2.096/2.221 | 2.321/3.047/2.857 | −10.9/−31.2/−22.3 | 0.83/1.08/1.38 | 3/3 |
| ETTm2 | 1000 | 0.879/0.936/0.713 | 2.929/2.926/3.133 | −70.0/−68.0/−77.2 | 1.28/1.56/1.18 | 3/3 |
| ETTm2 | 2000 | 0.856/1.030/1.182 | 3.085/3.092/3.084 | −72.3/−66.7/−61.7 | 1.29/0.87/0.92 | 3/3 |
| Weather | 1000 | 0.417/0.404/0.390 | 0.567/0.568/0.576 | −26.4/−29.0/−32.3 | 1.33/1.34/1.41 | 3/3 |
| Weather | 2000 | 0.417/0.469/0.434 | 0.574/0.593/0.544 | −27.2/−20.9/−20.3 | 1.24/0.86/1.30 | 3/3 |
| Electricity | 1000 | 0.520/0.657/0.506 | 0.736/0.799/0.825 | −29.4/−17.8/−38.6 | 1.51/1.58/1.52 | 3/3 |
| Electricity | 2000 | 0.514/0.591/0.733 | 0.762/0.958/0.846 | −32.6/−38.3/−13.4 | 1.54/1.43/1.30 | 3/3 |

Table Q.6: **Long horizons: Residual-IA$^+$ vs. single-Linear DLinear variant** (MOMENT-small, strictly frozen, test MSE per seed). Gap = (Residual-IA$^+$ − Linear)/Linear; a negative gap favors Residual-IA$^+$.

| Dataset | $H$ | Seed | Residual-IA$^+$ | Linear | Gap |
|---|---|---|---|---|---|
| ETTm2 | 1000 | 42 | 0.431 | 0.724 | $-40.5\%$ |
| ETTm2 | 1000 | 43 | 0.419 | 0.999 | $-58.0\%$ |
| ETTm2 | 1000 | 44 | 0.408 | 0.500 | $-18.5\%$ |
| Electricity | 1000 | 42 | 0.265 | 0.269 | $-1.6\%$ |
| Electricity | 1000 | 43 | 0.266 | 0.259 | $+2.7\%$ |
| Electricity | 1000 | 44 | 0.263 | 0.265 | $-0.8\%$ |
| Weather | 1000 | 42 | 0.368 | 0.371 | $-0.7\%$ |
| Weather | 1000 | 43 | 0.373 | 0.370 | $+0.8\%$ |
| Weather | 1000 | 44 | 0.378 | 0.380 | $-0.5\%$ |
| ETTm2 | 2000 | 42 | 0.435 | 0.592 | $-26.5\%$ |
| ETTm2 | 2000 | 43 | 0.462 | 0.830 | $-44.4\%$ |
| ETTm2 | 2000 | 44 | 0.447 | 1.006 | $-55.6\%$ |
| Electricity | 2000 | 42 | 0.329 | 0.334 | $-1.4\%$ |
| Electricity | 2000 | 43 | 0.333 | 0.341 | $-2.6\%$ |
| Electricity | 2000 | 44 | 0.336 | 0.334 | $+0.7\%$ |
| Weather | 2000 | 42 | 0.397 | 0.403 | $-1.5\%$ |
| Weather | 2000 | 43 | 0.416 | 0.408 | $+1.9\%$ |
| Weather | 2000 | 44 | 0.404 | 0.413 | $-2.2\%$ |

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state three contributions: (1) diagnosis of normalization-induced routing collapse with nine causal controls (Tables 2, G.3, 3), (2) RR-MoA achieving 54/54 wins on 6 datasets $\times$ 3 freeze levels $\times$ 3 seeds with statistical significance (Tables 1, 4), and (3) a mutual-information decomposition (Observation 1) yielding a signal-ratio statistic validated at Spearman $\rho = -0.88$, $p < 0.002$ (Figure 4). All claims are directly supported by experimental evidence.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Limitations statement in Section 5 limits the diagnosis to $(\mu, \sigma)$-stripping normalizers and states that learned expert pools remain open. Section 4 reports the boundary cases of $R(\mathcal{D})$: Traffic ($R$=0.14), where RR-MoA does not improve, and Solar ($R$=0.06), the one outlier.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper has three theoretical results, each with explicit assumptions and complete derivations in Appendix D. Observation 1 (Routing Information Loss Under Instance Normalization) Parts (i)–(ii) hold *without any independence assumption*; Part (iii) (full-information corner) additionally assumes $(M, \Sigma) \perp S$ (RevIN's design assumption) and $E \perp S \mid (M, \Sigma)$ (an idealized case; Section 4 shows that routing also uses the window shape). The bound is also measured directly via CCA, KSG, and held-out classifier (Appendix L), with the gap averaging $0.075$ nats. Proposition 1 (Gradient Co-Adaptation) includes its full assumptions and proof in the same appendix. Proposition 2 (Heuristic SNR Onset and Softmax Discontinuity) is explicitly labeled heuristic, with its derivation given inline and its prediction validated empirically by the 150-run dose-response grid (Figure D.1). Derivations rely on the data processing inequality, the chain rule of mutual information, standard softmax gradient calculus, and linear-variance scaling.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 3 describes the RR-MoA architecture (Conv1d router with $853$ parameters, $K$=5 canonical expert heads, Top-$k$ sparse routing) and Appendix A specifies all training details: optimizer (Adam), learning rate ($10^{-3}$), MSE loss, batch size (128), 15 epochs, seeds $\{42, 43, 44\}$ for the primary RR-MoA freeze ablation ($n$=3), seeds $\{42, \ldots, 46\}$ for the core multi-seed grid ($n$=5, used for rescue sweep, cross-backbone, baselines, multi-horizon, imputation, and Moirai-MoE), channel-wise StandardScaler normalization, chronological LTSF data splits, and backbone configurations. Algorithm 1 provides full pseudocode.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

Justification: All code is provided in the supplementary material under MIT license. All datasets used are publicly available: the standard LTSF benchmarks (ETT, Weather, Electricity, Traffic, Solar, Exchange), three datasets of the UEA archive [94] (BasicMotions, EthanolConcentration, JapaneseVowels), and the Server Machine Dataset (SMD) [95].

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section A provides all details: chronological train/val/test splits per the LTSF protocol, optimizer (Adam, lr=$10^{-3}$), MSE loss, batch size 128, 15 epochs, channel-wise StandardScaler normalization, and per-baseline configurations (fixed adapters, LoRA, TRACE, independent ensemble, AdaMix, full fine-tuning, DLinear); Appendix P defines the from-scratch NLinear anchor. The LoRA sweep (Appendix E) spans 12 configurations (rank $\in \{8, 16, 32\}$, targets $\in \{q{+}v, q{+}k{+}v{+}o\}$, head $\in$ {linear, MLP}) $\times$ 3 seeds $= 36$ runs per dataset, full table reported.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Multi-seed results report mean $\pm$ standard deviation across seeds, with the seed count indicated in each table caption. The core multi-seed grid (baseline comparison, rescue-baseline sweep, cross-backbone, multi-horizon, imputation, Moirai-MoE) uses 5 seeds $\{42, \ldots, 46\}$; Residual-IA$^+$ uses 10 seeds on MOMENT-small at $H{\in}\{96, 192\}$, 5 at $H{\in}\{336, 720\}$, and 3 to 5 on the other backbones, and each from-scratch linear anchor (NLinear, DLinear) uses 10; the primary RR-MoA freeze ablation and most ablation tables use 3 seeds $\{42, 43, 44\}$. Wilcoxon signed-rank $p$-values with Bonferroni correction are reported for the comprehensive baseline comparison. Appendix A.1 also reports uncorrected, Holm and Benjamini–Hochberg values. The routing signal-ratio correlation reports Spearman $\rho{=}-0.88$ with $p < 0.002$ ($n{=}9$).

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments were conducted on a single NVIDIA A10G GPU (23 GB VRAM, CUDA 12.4). Per-run training wall-clock (15 epochs, batch 128) median per dataset by backbone: MOMENT-small 1–2 minutes ($n{=}1{,}094$); MOMENT-large $\sim$3–3.5 minutes ($n{=}15$); Chronos $\sim$15–30 seconds ($n{=}42$); Moirai $\sim$25–30 seconds ($n{=}78$); Timer-XL $\sim$35–40 seconds ($n{=}90$); Moirai-MoE $\sim$1.5–4 minutes ($n{=}66$). Total training compute end-to-end is approximately 235 GPU-hours. Inference latency (Table B.1) reports wall-clock per architectural variant.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses publicly available benchmark datasets and pretrained models. No human subjects, private data, or dual-use concerns are involved. The work is foundational ML research on adapter routing.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: The work is methodological: it diagnoses and fixes a routing-collapse failure mode in MoE adapters on foundation models with instance-normalized inputs. **Positive impacts**: parameter-efficient adaptation of a shared frozen backbone reduces the compute and energy footprint of serving many downstream tasks or tenants relative to per-task fine-tuning, and lowers the barrier to deploying foundation models in resource-constrained settings (edge, on-device, multi-tenant cloud); see Appendix B for a concrete deployment regime. **Negative impacts**: we identify none specific to this work; the routing fix neither generates synthetic content, performs decision-making on protected populations, nor lowers the cost of any high-risk capability. The pretrained foundation models we adapt already exist and are publicly released; our adapters do not introduce new generative or autonomous capabilities.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases adapter routing code and lightweight adapter heads ($\leq$500K parameters) for time series forecasting. These pose no risk for misuse: they are small task-specific modules, not generative models.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All pretrained models and datasets are cited with their original papers in the bibliography. **Pretrained models (with license).** MOMENT [2]: MIT, `AutonLab/MOMENT-1-small` and `AutonLab/MOMENT-1-large` on Hugging Face. Moirai [7] and Moirai-MoE [8]: model *weights* are released under CC-BY-NC-4.0 (research/non-commercial use only) on `Salesforce/moirai-1.1-R-small` and `Salesforce/moirai-moe-1.0-R-small`; the supporting `uni2ts` code is Apache 2.0. Our use is academic research, consistent with the non-commercial weight license. Chronos [4]: Apache 2.0, `amazon/chronos-t5-small`. Timer-XL [6]: Apache 2.0, `thuml/timer-base-84m`. **Datasets.** ETT (ETTh1/h2/m1/m2) is introduced by Zhou et al. [36] and released under CC-BY-ND-4.0 (NoDerivatives); we ship only loader scripts and do not redistribute preprocessed arrays. Weather is the MPI-BGC Jena meteorological recording, redistributed for the LTSF benchmark by Wu et al. [37]. Electricity is the UCI `ElectricityLoadDiagrams20112014` dataset (CC-BY-4.0, originally collected by Artur Trindade). Traffic (Caltrans PeMS), Solar (NREL Solar Power Data), and Exchange (multi-currency exchange rates) were curated for time series benchmarking by Lai et al. [39]; the source data is public-record, but the curated benchmark CSVs do not ship with an explicit redistribution license, so we again only ship loader scripts pointing to the upstream sources. The Monash time series forecasting archive [38] (CC-BY-4.0) provides standardized versions of several of these datasets and is the recommended reproducibility anchor. The classification runs use BasicMotions, EthanolConcentration and JapaneseVowels from the UEA multivariate archive of Bagnall et al. [94], which its authors release publicly for research. The Server Machine Dataset (SMD) is introduced by Su et al. [95] and released in the OmniAnomaly repository (`NetManAIOps/OmniAnomaly`) under the MIT license.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Code and a README with reproduction instructions are provided in the supplementary material under MIT license.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or human subjects research.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human subjects research.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as a component of the core methodology.