MECHANISM-EQUIVALENCE ROUTING FOR TIME-SERIES FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse mixture-of-experts models scale capacity by conditionally reusing a small subset of parameters, yet the relation that should govern this reuse remains unspecified in time-series foundation models. Routing by trajectories or learned embeddings can merge similar observations with different predictive responses and split a common response across different operating conditions. We formulate *mechanism equivalence* as a support-conditioned operational reuse criterion: two target-window predictors may share response parameters when their baseline-subtracted multi-horizon responses agree on declared domains under admissible affine coordinate changes. Under canonical squared-error risk and a well-specified equivariant expert family, this relation admits a shared population-optimal response component with a unit-specific horizon baseline. MER operationalizes the relation with finite response probes, a canonical fingerprint, and prototype routing. A controlled CausalTime audit tests routing alignment under designed response families. In frozen target-stage evaluation, MER gives lower mean MSE than pattern routing in all ten directions and than the protocol-matched sparse baseline in all 40 dataset–horizon cells. These end-to-end results support frozen target-stage reuse under the evaluated same-backbone protocol.

1 INTRODUCTION

Time-series foundation models forecast across domains, sampling frequencies, dimensions, and horizons with shared parameters (Woo et al., 2024; Das et al., 2024; Goswami et al., 2024). Sparse mixture-of-experts (MoE) layers expand capacity without activating every parameter, but shift a central decision to the router: which target windows should reuse an expert? Moirai-MoE demonstrates the practical value of sparse specialization for time-series foundation models (Liu et al., 2025), while leaving the criterion for reuse implicit.

Input similarity is insufficient. Matched trajectories, spectra, autocorrelations, or embeddings can yield different conditional responses; conversely, the same response can appear under different offsets, units, noise levels, exogenous marginals, and operating points. Task and representation embeddings summarize a dataset or network state, but do not specify which target-window responses should share conditional parameters. A pattern-oriented router can therefore merge dynamics requiring different computation and split dynamics for which parameters remain reusable. This raises our central question: *what equivalence relation should determine conditional parameter reuse in a time-series foundation model?*

We propose *mechanism equivalence*: a support-conditioned, predictor-level reuse relation rather than an identified structural mechanism. Two target windows are reuse-equivalent if their normalized multi-horizon response contrasts agree on declared parent-lag supports and domains after admissible affine coordinate changes. Horizon baselines remain unit-specific; the relation retains sign, lag, active support, interaction, relative cross-horizon response amplitude, and non-affine changes. Under a well-specified canonical squared-error risk, equivalent responses share a population-optimal response component.

Mechanism-Equivalence Routing (MER) approximates this relation using a canonical function-response fingerprint (Figure 1). Its central contribution is the support-conditioned response-equivalence criterion for expert reuse; response encoding, clustering, and prototype routing operationalize that criterion rather than constitute separate novelty claims. A frozen source-trained

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

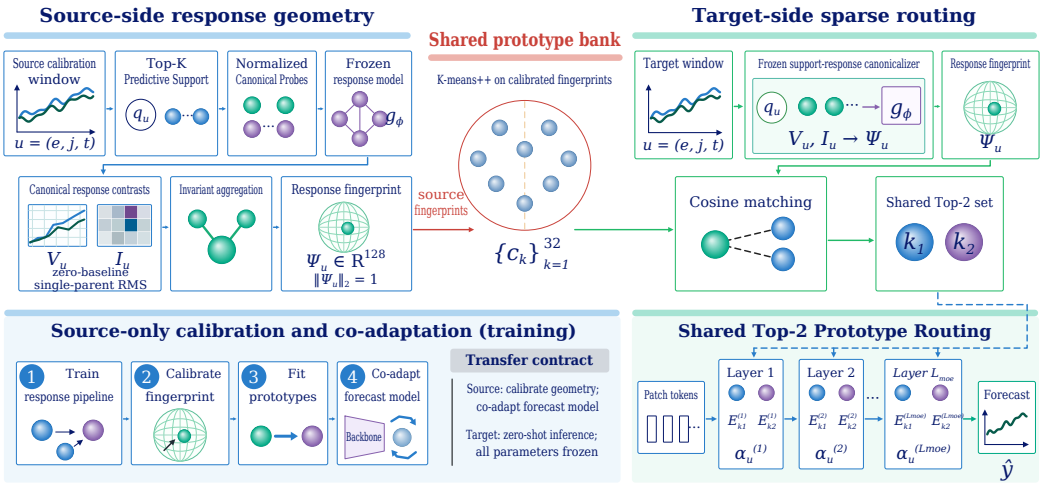


Figure 1: Overview of MER. Source calibration produces canonical, baseline-subtracted response fingerprints and a shared prototype bank. At target time, a frozen canonicalizer maps each window to a fingerprint; cosine matching selects shared Top-2 prototype indices, while layer-specific experts and mixture weights produce the forecast. All parameters remain frozen during target inference.

model is queried on normalized parent values and pairwise combinations, and a permutation-invariant encoder summarizes multi-horizon responses and lags. A fixed prototype bank gives expert indices common reuse semantics across layers.

Forecast accuracy alone cannot establish whether the intended reuse relation was learned. We therefore introduce the Pattern–Mechanism Decoupling Audit (PMDA), contrasting matched-pattern/different-response pairs with different-pattern/same-response pairs. Controlled CausalTime constructions expose declared reuse classes; held-out compositions test constituent recombination; and frozen cross-dataset experiments test reuse without target-stage fitting. We separate pretraining exposure, MER source-stage data, source-bank formation, and complete routing-pipeline choice. Together, these evaluations connect response-aligned routing in designed audits to forecasting and transfer on real series.

Our contributions are threefold:

- We formulate support-conditioned response equivalence as a parameter-reuse criterion and show its sufficiency, under stated assumptions, for a shared population-optimal response component.
- We operationalize this criterion with finite multi-horizon response probes, a canonical fingerprint, and shared prototypes that give expert indices consistent cross-layer semantics.
- We introduce PMDA and an evidence chain spanning routing alignment, forecasting, held-out composition, and frozen reuse, while contrasting source-bank formation with complete routing-pipeline choice.

2 Related Work

2.1 Time-series foundation models and sparse experts

Forecasting architectures exploit decomposition, frequency structure, patches, variate tokens, and multiscale mixing (Wu et al., 2021; Zhou et al., 2022; Nie et al., 2023; Liu et al., 2024; Wang et al., 2024); foundation models extend this line with shared patched decoders, masked encoders, representations, or probabilistic tokens (Das et al., 2024; Woo et al., 2024; Goswami et al., 2024; Ansari et al., 2024). Sparse MoEs activate a small expert subset per input (Shazeer et al., 2017; Lepikhin et al., 2021; Fedus et al., 2022). For time series, MoLE routes by periodicity, Time-MoE

scales sparse decoder-only forecasting, Moirai-MoE initializes routes from representation clusters, and FAME learns cost-aware suitability from forecastability fingerprints (Ni et al., 2024; Shi et al., 2025; Liu et al., 2025; Li et al., 2026). MER instead defines when target-window responses are eligible to reuse parameters and audits whether routes respect that relation.

2.2 Pattern and structure as routing signals

Temporal shape, spectra, embeddings, and input-dependent pathways are useful routing signals (Nie et al., 2023; Wu et al., 2023; Chen et al., 2024), but similar patterns can support different responses and shared responses can occur under different marginals. Directed support and lagged dependence add structure (Zheng et al., 2018; Tank et al., 2022; Zhou et al., 2024), yet equal support does not fix sign, saturation, threshold, or interaction. AME-TS routes by interpretable temporal structure, ACT couples causal attention with sparse experts, and Fuzzy-MoE uses learned regime memberships (Wang et al., 2026; Zhang et al., 2026; Guo et al., 2026). MER compares the normalized response itself; pattern, correlation, support, graph, and local-response descriptors enter the same-backbone ladder as substantive alternatives.

2.3 Functional similarity and reusable modules

Invariant prediction studies conditionals stable across environments (Peters et al., 2016). Independent Causal Mechanisms learns transferable inverse-mechanism experts, while Recurrent Independent Mechanisms and routing networks pursue sparsely interacting or dynamically selected modules (Parascandolo et al., 2018; Goyal et al., 2021; Rosenbaum et al., 2018). TRACE models atomic mechanisms and mixture trajectories; Causal Delta Embeddings encode sparse, context-invariant intervention changes (Fan et al., 2026; Alimisis & Diou, 2026). Together, these works motivate modular mechanisms; MER builds on this foundation to define when target windows should reuse parameters.

Model-reuse alternatives compare domains, tasks, or representations: Expert Gate uses reconstruction, Task2Vec a Fisher signature, and CKA representation geometry (Aljundi et al., 2017; Achille et al., 2019; Kornblith et al., 2019). MER instead compares target-window responses modulo admissible positive-affine transformations and assigns expert indices shared semantics across layers. Separate invariance, separation, and matched-routing audits test this window-level reuse relation. A standard Deep Sets encoder (Zaheer et al., 2017) and clustering instantiate the geometry; the contribution lies in the reuse criterion, shared index semantics, and its audit.

3 Problem Formulation

Let $\boldsymbol{X}^{e}_{1:T} \in \mathbb{R}^{T\times C_e}$ be a multivariate time series in environment e. At origin t, a model maps $\boldsymbol{X}^{e}_{t-L+1:t}$ to an H-step forecast for target channel j. Its L_{moe} sparse layers contain K experts; for target-window unit $u=(e,j,t)$, layer ℓ returns

$$\boldsymbol{\alpha}_u^{(\ell)} = R^{(\ell)}(u) \in \Delta^{K-1}, \qquad \|\boldsymbol{\alpha}_u^{(\ell)}\|_0 \le k \ll K. \tag{1}$$

Expert architecture and active parameter budget are fixed. We seek the relation under which two units should reuse expert parameters.

Let $\mathcal{H}_r \subseteq \{1,\dots,H\}$ be a pre-specified response-horizon set. For unit u, $\mathcal{S}_u$ is its finite active parent-lag role set, $\mathcal{D}_u \subset \mathbb{R}^{|\mathcal{S}_u|}$ is a compact response domain, and $f_u : \mathcal{D}_u \to \mathbb{R}^{|\mathcal{H}_r|}$ is the conditional response over $\mathcal{H}_r$. Roles are defined by lag and functional contribution rather than environment-specific channel name. A coordinate is active only if f_u is nonconstant in it. The mechanism is $M_u = (\mathcal{S}_u, \mathcal{D}_u, f_u)$; on observational data it is a predictive-model object, not necessarily an identified interventional response.

For equal-size supports, let $d = |\mathcal{S}_u| = |\mathcal{S}_v|$ and let $\pi : \mathcal{S}_u \to \mathcal{S}_v$ be lag-preserving, with permutation matrix $\boldsymbol{P}_\pi$. An admissible input map

$$T_{uv}(\boldsymbol{p}) = \boldsymbol{P}_\pi(\boldsymbol{A}\boldsymbol{p} + \boldsymbol{b}), \quad \boldsymbol{A} = \mathrm{diag}(a_1,\dots,a_d),\ a_m > 0, \tag{2}$$

maps $\mathcal{D}_u$ bijectively to $\mathcal{D}_v$; these admissible input maps form $\mathcal{G}_{\mathrm{aff}}$. Outputs separately admit $\boldsymbol{y} \mapsto a_y\boldsymbol{y} + \boldsymbol{b}_y$, where $a_y > 0$ is shared across response horizons and $\boldsymbol{b}_y \in \mathbb{R}^{|\mathcal{H}_r|}$ is a horizon baseline. The relation quotients out baseline level and common output gain while retaining the relative cross-horizon response to parent perturbations.

Definition 1 (reuse equivalence). $M_u \sim_{\text{reuse}} M_v$ if there exist a lag-preserving π, $T_{uv} \in \mathcal{G}_{\text{aff}}$, $a_y > 0$, and $\boldsymbol{b}_y \in \mathbb{R}^{|\mathcal{H}_r|}$ such that

$$f_v(T_{uv}(\boldsymbol{p})) = a_y f_u(\boldsymbol{p}) + \boldsymbol{b}_y, \qquad \forall \boldsymbol{p} \in \mathcal{D}_u. \tag{3}$$

Proposition 1 (Well-defined reuse classes)**.** *On the declared mechanism space, $\sim_{\text{reuse}}$ is an equivalence relation.*

The exact relation is a reference object rather than a numerical router. Appendix A defines an operational quotient discrepancy D_{reuse} and uses $D_{\text{reuse}} \leq \epsilon$ as approximate reuse. MER estimates proximity by $\widehat{d}_\Psi = \|\boldsymbol{\Psi}_u - \boldsymbol{\Psi}_v\|_2$ and then quantizes the estimate through prototypes; reversing either implication requires response-approximation and prototype-margin conditions.

Let $\widetilde{\boldsymbol{P}}_u, \widetilde{\boldsymbol{Y}}_u$ denote parent and target variables after the canonical location-scale normalization used below. Let $e_\theta : \mathbb{R}^d \to \mathbb{R}^{|\mathcal{H}_r|}$ be a shared role-permutation-equivariant response expert. For a unit-specific horizon baseline $\boldsymbol{\beta}_u \in \mathbb{R}^{|\mathcal{H}_r|}$, define the population risk under the joint law of $(\widetilde{\boldsymbol{P}}_u, \widetilde{\boldsymbol{Y}}_u)$ associated with u:

$$\mathcal{R}_u(\boldsymbol{\beta}_u, \theta) = \mathbb{E}\left\|\widetilde{\boldsymbol{Y}}_u - \left(\boldsymbol{\beta}_u + e_\theta(\widetilde{\boldsymbol{P}}_u)\right)\right\|_2^2. \tag{4}$$

Proposition 2 (Response-component reuse sufficiency)**.** *Suppose $M_u \sim_{\text{reuse}} M_v$, the input and gain normalization statistics transform equivariantly, and the predictor family decomposes into a unit-specific horizon baseline plus a role-permutation-equivariant response expert. If that expert family contains the common normalized response contrast on the declared domains, then one response-expert parameterization, up to the role permutation, is population-optimal for both risks after optimizing their respective baselines.*

Here $\boldsymbol{\beta}_u$ is a proof-level decomposition, not an additional trainable MoE head. Equivalence is sufficient, not necessary, for reuse of the canonical response component; it does not require two units to share a horizon baseline and does not imply identical optima for arbitrary layerwise hidden representations or finite training. The relation removes coordinate nuisances while retaining sign, lag, support, interaction, relative cross-horizon response shape, and non-affine changes. We seek neither universal causal discovery nor arbitrary reparameterization invariance, and prototype identity is not proof of exact equivalence.

4 Mechanism-Equivalence Routing

4.1 Support and multi-horizon response model

For target-window $u = (e, j, t)$, MER first selects predictive parent-lag roles. For channel $i \in \{1, \ldots, C_e\}$ and lag $\tau \in \{1, \ldots, \tau_{\max}\}$, a shared window encoder produces a target query $\boldsymbol{q}_u$ and candidate representation $\boldsymbol{h}_{i,\tau}$; their support score is

$$s_{u,i,\tau} = \text{MLP}_s([\boldsymbol{q}_u, \boldsymbol{h}_{i,\tau}, \boldsymbol{q}_u \odot \boldsymbol{h}_{i,\tau}]). \tag{5}$$

We set $K_p = 6$ and $\tau_{\max} = 8$, and retain the top-K_p candidates as $\widehat{\mathcal{S}}_u$. This fixed-budget set approximates $\mathcal{S}_u$ and can include proxies or inactive candidates. Training executes deterministic hard top-K_p selection in the forward pass, while an untempered softmax relaxation over the same support logits passes prediction-loss gradients backward; no stochastic support sampling is used.

Let $\boldsymbol{p}_u \in \mathbb{R}^{K_p}$ collect selected values, and let $\boldsymbol{c}_u \in \mathbb{R}^{384}$ encode observed context ending before the largest candidate lag; canonical queries hold $\boldsymbol{c}_u$ fixed while varying only parent coordinates (architecture: Appendix Table 11). If selected role m uses channel i_m, define the observed histories $\mathcal{C}^p_{u,m} = \{X^e_{s,i_m} : s = t - L + 1, \ldots, t\}$ and $\mathcal{C}^y_u = \{X^e_{s,j} : s = t - L + 1, \ldots, t\}$. At every forecast origin, MER computes

$$\mu^p_{u,m} = \text{median}(\mathcal{C}^p_{u,m}), \qquad s^p_{u,m} = \begin{cases} \text{IQR}(\mathcal{C}^p_{u,m}), & \text{IQR}(\mathcal{C}^p_{u,m}) > 0, \\ 1, & \text{otherwise}, \end{cases}$$

$$\mu^y_u = \text{median}(\mathcal{C}^y_u), \qquad s^y_u = \begin{cases} \text{IQR}(\mathcal{C}^y_u), & \text{IQR}(\mathcal{C}^y_u) > 0, \\ 1, & \text{otherwise}. \end{cases} \tag{6}$$

MER uses $\widetilde{p}_{u,m} = (p_{u,m} - \mu^p_{u,m})/s^p_{u,m}$ and $\widetilde{y}_{u,t+h} = (X^e_{t+h,j} - \mu^y_u)/s^y_u$ for every h. Zero-IQR histories are flagged and mapped to zero after centering; the unit-scale fallback lies outside the affine-equivariant case of Proposition 4. Thus normalization uses only the observed window, shares one target location and scale across horizons, and never fits statistics on target training data. We use the relative grid

$$\mathcal{H}_r(H) = \{1, \lceil H/4 \rceil, \lceil H/2 \rceil, H\}, \tag{7}$$

with duplicate indices removed; the four entries represent immediate, quarter-range, half-range, and terminal response roles when tasks use different H. A two-layer, 128-hidden-unit GELU head predicts all selected horizons jointly:

$$\widehat{\widetilde{\boldsymbol{y}}}_{u,t+\mathcal{H}_r} = g_\phi(\boldsymbol{c}_u, \widetilde{\boldsymbol{p}}_u) \in \mathbb{R}^{|\mathcal{H}_r|}. \tag{8}$$

The encoder, scorer, context network, and response head are trained jointly through the straight-through support relaxation with the normalized multi-horizon forecasting loss in Appendix C, and then frozen. The support scorer therefore receives forecasting supervision rather than mechanism labels. The descriptor estimates predictive responses for the direct multi-horizon task.

Canonical queries may be weakly supported when parents are dependent. A source-only reservoir of $10{,}000$ normalized parent tuples defines squared-Euclidean 5-nearest-neighbor coverage at its 95th-percentile cutoff. The resulting mask restricts $\mathcal{L}_{\text{query}}$ to source-covered queries during calibration; neither the mask nor the underlying distance enters the fingerprint or target-time route.

4.2 Canonical function-response fingerprint

For anchors $\mathcal{A} = \{-1, -0.5, 0, 0.5, 1\}$, parent m, and horizon $h \in \mathcal{H}_r$, let $\boldsymbol{e}_m$ be the mth standard basis vector in $\mathbb{R}^{K_p}$ and define

$$\Delta r_{u,m,a,h} = g_{\phi,h}(\boldsymbol{c}_u, a\boldsymbol{e}_m) - g_{\phi,h}(\boldsymbol{c}_u, \boldsymbol{0}). \tag{9}$$

To retain relative response amplitude across horizons, we use one root-mean-square over the single-parent response tensor,

$$\sigma_{r,u} = \left(\frac{1}{K_p |\mathcal{A}| |\mathcal{H}_r|} \sum_{m=1}^{K_p} \sum_{a \in \mathcal{A}} \sum_{h \in \mathcal{H}_r} \Delta r^2_{u,m,a,h} \right)^{1/2}, \qquad \bar{\sigma}_{r,u} = \max(\sigma_{r,u}, \epsilon_{\text{flat}}), \tag{10}$$

where ϵ_{flat} is fixed on source validation. All reported experiments use this single-parent RMS; raw pairwise contrasts do not enter the scale. We then use

$$V_{u,m,a,h} = \Delta r_{u,m,a,h} / \bar{\sigma}_{r,u}. \tag{11}$$

Input normalization removes coordinate units, zero-query subtraction implements the horizon-baseline quotient in Definition 1, and, when the scale floor is inactive, the common scale removes positive output gain while preserving relative parent and cross-horizon perturbation responses. The baseline in Equation 4 is theoretical, not an extra trainable head. The floor is computed from the single-parent tensor rather than the full response: $g(p_1, p_2) = p_1 p_2$ has zero single-parent contrasts but nonzero pairwise contrasts. Such a response lies outside the nondegenerate-scale condition of Proposition 4; the floor prevents division by zero but does not bound normalized pairwise contrasts, so this case can be numerically ill-conditioned.

For the three highest-ranked parents $\mathcal{Q}_u$, pair $(m, n) \subset \mathcal{Q}_u$, and $(a, b) \in \{(-1, -1), (1, 1)\}$, the interaction contrast is

$$\begin{aligned} I_{u,mn,h}(a, b) = \bar{\sigma}^{-1}_{r,u} [& g_{\phi,h}(\boldsymbol{c}_u, a\boldsymbol{e}_m + b\boldsymbol{e}_n) - g_{\phi,h}(\boldsymbol{c}_u, a\boldsymbol{e}_m) \\ & - g_{\phi,h}(\boldsymbol{c}_u, b\boldsymbol{e}_n) + g_{\phi,h}(\boldsymbol{c}_u, \boldsymbol{0})]. \end{aligned} \tag{12}$$

Because $V_{u,m,0,h} = 0$ identically, the parent descriptor excludes it:

$$\boldsymbol{z}_{u,m} = [\tau_m / \tau_{\max}, \{V_{u,m,a,h} : a \in \mathcal{A} \setminus \{0\}, h \in \mathcal{H}_r\}]. \tag{13}$$

Two-layer, 256-hidden-unit GELU encoders map $\boldsymbol{z}_{u,m}$ and the symmetric endpoint interactions to 128-dimensional embeddings. A two-layer, 64-hidden-unit GELU map ρ aggregates their sums; LayerNorm and L_2 normalization produce the fingerprint

$$\boldsymbol{\Psi}_u = \text{norm}\left(\rho \left[\sum_m \boldsymbol{h}_{u,m}, \sum_{(m,n) \subset \mathcal{Q}_u} \boldsymbol{h}_{u,mn} \right] \right). \tag{14}$$

The pooling encoder is permutation-invariant for consistently permuted selected-role descriptors and interaction endpoints. Full-pipeline equivariance additionally requires corresponding support selection and Top-3 tie handling (Proposition 4). The finite grid approximates the response relation on the sampled queries.

4.3 Context calibration and shared prototypes

On controlled pairs known to share a mechanism across environments, context leakage is penalized by

$$\mathcal{L}_{\mathrm{ctx}} = \|\boldsymbol{\Psi}_u - \boldsymbol{\Psi}_v\|_2^2. \tag{15}$$

No such pair is inferred from unlabeled observational data. In real-data training, $\mathcal{L}_{\mathrm{ctx}}$ is disabled; controlled pairs and their mechanism labels never enter the real-data pathway. Let $\mathcal{Q}_u^\star$ be the declared fingerprint query family: the zero query, all one-parent anchor queries, and the pre-specified two-anchor queries for pairs in $\mathcal{Q}_u$. For each covered query, define the canonical response contrast

$$\widetilde{\boldsymbol{r}}_u(\boldsymbol{p}^\star) = \frac{g_\phi(\boldsymbol{c}_u, \boldsymbol{p}^\star) - g_\phi(\boldsymbol{c}_u, \mathbf{0})}{\bar{\sigma}_{r,u}}. \tag{16}$$

A temporary decoder receives the fingerprint and a canonical parent query and is trained to reconstruct this contrast rather than the discarded horizon baseline. Queries are sampled uniformly from the source-covered entries of the fixed query family $\mathcal{Q}_u^\star$:

$$\mathcal{L}_{\mathrm{query}} = \|d_\omega(\boldsymbol{\Psi}_u, \boldsymbol{p}^\star) - \widetilde{\boldsymbol{r}}_u(\boldsymbol{p}^\star)\|_2^2. \tag{17}$$

The temporary decoder receives six-dimensional queries in a fixed order and learns role correspondence implicitly, without an explicit coordinate-to-role map. The pooling invariance applies to the fingerprint; query-conditioned reconstruction uses the supplied coordinate order. After calibration, the fingerprint encoder is frozen and the decoder discarded.

Calibration is separated by evidence source: controlled runs use $\mathcal{L}_{\mathrm{query}} + \lambda_{\mathrm{ctx}}\mathcal{L}_{\mathrm{ctx}}$, whereas real-data runs use $\mathcal{L}_{\mathrm{query}}$ alone. The nonnegative λ_{ctx} is selected on controlled source validation and never applied to observational runs, preventing synthetic equivalence labels from supervising the real-data router.

K-means++ with a fixed seed clusters $100{,}000$ source fingerprints into one unit-normalized bank $\{\boldsymbol{c}_k\}_{k=1}^{32}$. The bank size equals the 32 expert slots of the pretrained architecture. These prototypes quantize fingerprint geometry: they are neither estimates of the number of mechanisms nor assumed to correspond one-to-one with reuse-equivalence classes. Routing uses

$$z_{u,k}^{(\ell)} = \langle \boldsymbol{\Psi}_u, \boldsymbol{c}_k \rangle / T_\ell, \qquad \boldsymbol{\alpha}_u^{(\ell)} = \mathrm{Top2Softmax}(\boldsymbol{z}_u^{(\ell)}), \tag{18}$$

where $T_\ell = \mathrm{softplus}(t_\ell) > 0$ is initialized at 0.13 and is the only trainable routing quantity after geometry is fixed. The route is shared by the target window's patch tokens and fixed for a direct H-step forecast. Because the prototype bank is shared and T_ℓ is scalar, MER intentionally selects the same Top-2 prototype indices in every sparse layer. Layer-specific experts $E_k^{(\ell)}$ implement different computations for that quantized region, while T_ℓ controls only mixture sharpness. This design trades layer-specific routing freedom for shared cross-layer prototype semantics.

4.4 Training order and cost

MER freezes response geometry before forecast co-adaptation. Controlled models train from scratch; named-source ETT/PEMS models initialize from Moirai-MoE and execute Appendices C–D, whereas ILI/Weather load a declared source-mixture checkpoint with inherited stage provenance. Final co-adaptation updates the backbone, experts, and temperatures under fixed response, fingerprint, and prototype geometry using forecasting loss plus $0.01\mathcal{L}_{\mathrm{lb}}$. Comparisons fix the source artifact, backbone and active-expert budget, co-adaptation optimizer and updates, checkpoint rule, splits, test windows, horizons, and MSE aggregation. The comparison therefore evaluates the full source-trained pipeline; Table 1 reports inference overhead.

The fingerprint uses 30 logical single-parent and six pairwise evaluations per target channel; caching the zero query leaves 31 unique queries, each returning all $|\mathcal{H}_r|$ horizons. For C targets the logical scale is $O(CK_p|\mathcal{A}|)$, with $C \times 31$ queries batched into one or a few response-head calls. We report per-target-window FLOPs and latency; origin-level cost also depends on the batched target-channel count.

5 ANALYSIS

We analyze two questions: why common observational descriptors may fail to identify reuse classes, and when the canonical fingerprint preserves reuse equivalence. For the first, consider the scalar $h = 1$ component on the compact domain $[-1, 1]$, with $X_t \overset{\text{iid}}{\sim} \text{Unif}[-1, 1]$, $G(x) = x^3 - \frac{3}{5}x$, and $Y_{t+1} = f_Z(X_t) + \varepsilon_{t+1}$, where $f_Z(x) = ZaG(x)$, $Z \in \{-1, +1\}$ is uniform, $a > 0$, and ε_t is iid, independent of X, and identically distributed across classes.

Proposition 3 (Ambiguity under second-order statistics)**.** *The classes have identical input and output marginals, parent support, and population auto- and cross-covariances, although f_{+1} and f_{-1} are not reuse-equivalent.*

This counterexample establishes a general insufficiency of these summaries. For a routing descriptor T with the same class-conditional law here, it also gives a forecasting consequence.

Corollary 1 (Descriptor-limited routing)**.** *If $P(T \mid Z = +1) = P(T \mid Z = -1)$, every classifier based on T has error at least $1/2$. If a wrong expert incurs conditional excess loss at least $\Delta_{\min} > 0$, expected excess loss is at least $\Delta_{\min}/2$.*

More generally, descriptor error $\mathcal{E}_T$ implies

$$\mathcal{R}_T - \mathcal{R}_{\text{oracle}} \geq \Delta_{\min}\mathcal{E}_T. \tag{19}$$

The bound characterizes population excess risk under the stated descriptor and loss assumptions.

We next address the second question under exact response evaluation.

Proposition 4 (Exact canonical invariance)**.** *Suppose $M_u \sim_{\text{reuse}} M_v$, their context-derived normalization statistics transform equivariantly, and the response model is exact on the canonical query domain. Assume the selected support, interaction Top-3 set, and any tie decisions transform consistently under the role bijection. If both single-parent response scales in Equation 10 exceed ϵ_{flat}, the corresponding normalized single-parent and pairwise contrasts coincide, and a common permutation-invariant encoder yields $\mathbf{\Psi}_u = \mathbf{\Psi}_v$.*

On a finite grid, non-equivalent responses can coincide, so fingerprint agreement is approximate rather than evidence of exact or causal identity. Under bounded loss, Appendix B separates mixture-weight perturbation with a stable Top-2 set from support changes, while an unrestricted predictor also incurs fixed-bank approximation error. The needed Lipschitz constant, margin distribution, and real-data rates of nondegenerate scale, support correspondence, stable Top-3 sets, and tie consistency are unmeasured; the result is therefore conditional, not a numerical generalization or empirical-coverage claim.

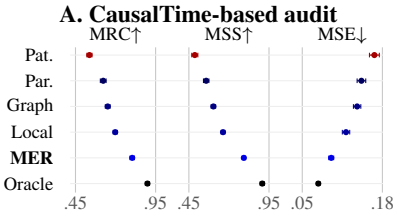
6 EXPERIMENTS

We test whether MER separates pattern from declared response class, improves matched end-to-end forecasting, recombines held-out constituents, and supports frozen cross-dataset reuse. Figure 2 and Tables 1–3 give the main evidence; the appendix supplies full results and controls.

6.1 EXPERIMENTAL SETUP

Controlled evaluations. The CausalTime-based audit (Cheng et al., 2024) uses five MER-defined instances across four response families and five declared reuse classes. These are designed diagnostics, not original benchmark labels or a representative sample of real-world mechanisms (Appendix Tables 4 and 5). Composition uses $g_1(x) = \tanh x$, $g_2(x) = \frac{1}{2}\tanh x$, $g_3(x) = x^2$, and $g_4(x, z) = xz$; training on $\{1, 2\}$, $\{2, 3\}$, and $\{1, 4\}$ holds out $\{3, 4\}$. Because g_1, g_2 are affine-equivalent, this tests three reuse classes (Appendix Table 6).

Real datasets and exposure. Forecasting uses ETT (Zhou et al., 2021), PEMS (Song et al., 2020), CDC ILI, and MPI-BGC Weather (Appendix Tables 12 and 15). ETT/PEMS use distinct source–target datasets, while ILI/Weather use a frozen source-mixture checkpoint. *Zero-shot target adaptation* denotes no target-stage parameter updates; pretraining and source-stage provenance are summarized in Appendix Table 15.



B. Relative MSE reduction (%)

	ETTh1	ETTh2	ETTm1	ETTm2	P03	P04	P07	P08	ILI	Weather
H_1	9.8	10.1	10.5	8.9	7.9	12.3	14.3	16.1	9.4	9.6
H_2	10.0	10.1	10.6	9.0	7.9	11.7	13.4	15.0	9.5	9.3
H_3	9.8	10.2	10.6	9.0	8.6	10.8	12.3	15.4	9.3	9.4
H_4	10.7	11.1	9.2	11.5	4.2	12.1	15.5	22.3	9.5	9.4

Figure 2: Controlled routing fidelity and real-data MSE reductions. **A**: mean ± one standard deviation over five CausalTime-based audit runs ($n_{\mathrm{MRC}} = 120$, $n_{\mathrm{MSS}} = 160$ retained test pairs). **B**: relative MSE reduction of MER over matched Moirai-MoE, positive across all 40 dataset–horizon cells; H_1–H_4 denote each dataset's four reported horizons, and P03–P08 abbreviate the PEMS datasets. Exact values are reported in Appendix Tables 18 and 20.

Table 1: Matched forecasting/co-adaptation and per-window inference cost. MER adds support-response training, fingerprint calibration, and prototype formation (executed for named-source transfers; inherited for ILI/Weather). Parentheses: macro-MSE reductions; lower in 40/40 cells. Mean±SD: MSE over five seed-level family macros; FLOPs across dataset–horizon settings; latency over 1,000 synchronized passes on one checkpoint. Cells: Appendix Table 20.

Method	ETT MSE ↓	PEMS MSE ↓	ILI+Weather MSE ↓	FLOPs ↓	Latency ↓	Aux. params
Moirai-MoE	.342±.013	.124±.005	.308±.013	3.3±.1G	17.5±.6ms	–
MER	**.308±.012**	**.108±.004**	**.279±.014**			
(lower: 40/40)	(−9.9%)	(−12.9%)	(−9.4%)	3.5±.1G	20.4±.7ms	2.1M

Architectures and controls. The controlled and same-backbone routing studies use a common six-layer, 32-expert Moirai-MoE Small backbone. Five runs match splits, optimizer, co-adaptation updates, expert architecture, and active parameters. MER additionally performs source-side response training, fingerprint calibration, and prototype formation on the named ETT/PEMS source or inherits them with the ILI/Weather source-mixture checkpoint. Real-data models initialize from `Salesforce/moirai-moe-1.0-R-small` (Liu et al., 2025). Native Moirai-MoE defines forecasting reductions; our Pattern control defines MRA. Parent, Graph, and Local-response controls replace complete descriptor pipelines, so comparisons are system-level (Appendix C).

The real-data path uses no controlled pairs and freezes every component at zero shot (Appendix Table 8). At 1/5/10% target data, only T_ℓ and $\widehat{y}_j^{\mathrm{cal}} = a_j \widehat{y}_j + b_j$ update. The 4×4 factorial substitutes complete routing pipelines, not one router module.

6.2 Metrics and Reporting

PMDA measures mechanism reuse consistency (MRC) and mechanism switch sensitivity (MSS) using fixed trajectory, spectral, autocorrelation, and frozen MOMENT distances. These fixed distance definitions are used throughout the PMDA evaluation. With validation CDFs $\widehat{F}_k^{\mathrm{val}}$, let $D_P = \frac{1}{4}\sum_{k=1}^{4} \widehat{F}_k^{\mathrm{val}}(d_k)$ and freeze its 95th/5th percentiles $q_{0.95}, q_{0.05}$ before test filtering. Let d_R be mean layerwise Jensen–Shannon routing divergence divided by $\log 2$. Then

$$\begin{aligned} \mathrm{MRC}_{\mathrm{PMDA}} &= 1 - \mathbb{E}[d_R(u,v) \mid M_u \sim_{\mathrm{reuse}} M_v,\ D_P(u,v) \geq q_{0.95}], \\ \mathrm{MSS}_{\mathrm{PMDA}} &= \mathbb{E}[d_R(u,v) \mid M_u \not\sim_{\mathrm{reuse}} M_v,\ D_P(u,v) \leq q_{0.05}]. \end{aligned} \tag{20}$$

Appendix Table 16 and Appendix C give full baselines, metrics, and scope.

6.3 Does MER Track Reuse Equivalence?

MER yields MRC/MSS 0.804/0.793 and MSE 0.097 in the designed audit (Figure 2A; Appendix Table 18). Cell-wise mean MSE reductions span 4.2–22.3% across 40 correlated cells (Figure 2B; Appendix Table 20); matched per-window latency rises 16.6% (Table 1). MER's pipeline has the lowest ETT macro MSE for every fixed bank (Table 2B; Appendix Table 19). These results quantify end-to-end gains under the matched forecasting and co-adaptation protocol.

Table 2: Δ = MER − Pattern; pp: percentage points. A: run-wise environment averages; CompGap: relative held-out-vs.-seen MSE increase; NA: no sparse slots. B: ETT macro MSE by bank and full routing pipeline; SDs: Appendix Table 19. Pat./Par.: Pattern/Parent; blue/red: lower/higher MSE.

A. Held-out function composition

Method	Held MSE ↓	CompGap ↓	Top-2 ↑
Dense	.124±.005	61.0%	NA
Pattern	.130±.006	51.2%	44.2±1.5%
MER	**.079±.004**	**29.5%**	**81.5±2.65%**
Δ vs. Pat.	−.051	−21.7 pp	+37.3 pp

B. ETT macro: bank × routing pipeline

Bank	Pat.	Par.	Graph	MER
Pattern	.33876	.34512	.34183	**.33241**
Parent	.33125	.32510	.32743	**.31987**
Graph	.32217	.31974	.31641	**.31128**
MER	.31452	.31183	.30946	**.30775**

Table 3: Graph/MER MSE; MRA: Pattern-relative mechanism reuse advantage. All rows: zero target-stage updates. "Src.": target excluded from MER source-stage data; "Frozen": source-mixture checkpoint. Provenance: Appendix Table 15; full results: Appendix Table 16.

Direction	Graph	MER	MRA ↑	Scope
ETTm1→ETTm2	.31382	**.30425**	9.1%	Src.
ETTm2→ETTm1	.33217	**.32425**	9.0%	Src.
ETTh1→ETTh2	.30418	**.29575**	9.1%	Src.
ETTh2→ETTh1	.31546	**.30675**	9.4%	Src.
ILI target	.28574	**.27850**	9.2%	Frozen

Direction	Graph	MER	MRA ↑	Scope
PEMS03→PEMS04	.11138	**.10850**	9.5%	Src.
PEMS04→PEMS03	.11024	**.10750**	9.1%	Src.
PEMS07→PEMS08	.10971	**.10700**	8.9%	Src.
PEMS08→PEMS07	.11205	**.10900**	9.3%	Src.
Weather target	.28715	**.27950**	9.2%	Frozen

Bank formation and routing-pipeline choice. The factorial in Table 2B compares these two design choices separately. With the Pattern pipeline fixed, the MER-formed bank lowers ETT macro MSE from .33876 to .31452; with the Pattern bank fixed, the MER pipeline lowers it to .33241. Combining both yields .30775. The MER bank has the lowest mean in every routing column, and MER routing has the lowest mean in every bank row. This two-way ordering connects the gains to both response-aligned bank formation and routing-pipeline choice.

Composition and frozen reuse. MER improves CompGap (51.2% to 29.5%) and Top-2 recovery (44.2% to 81.5%) (Table 2A; Appendix Table 14). The held-out case combines the even response x^2 with the interaction xz: each constituent appears in training, while their combination is reserved for testing. Improved held-out MSE together with higher constituent Top-2 recovery therefore connects forecasting gains to reuse of known response components in a new composition. This task-level check complements PMDA's pairwise routing diagnostic. Across ten frozen directions, MER has lower MSE than Graph and 8.9–9.5% Pattern-relative MRA (Table 3; Appendix Table 17). Appendix Table 17 additionally reports the remaining gap to full adaptation.

Adapting with fixed response geometry. On ETTm1→ETTm2, MER lowers MSE from .30425 at zero shot to .27800 with 10% target data (Appendix Figure 3). It has the lowest mean among the four routing pipelines at each tested budget, and its MSE advantage over Graph widens from .00957 to .01300. Only routing temperatures and the channelwise output affine map update; the response model, fingerprint, prototypes, backbone, and experts remain frozen. This result illustrates how the source-built reuse representation can support both frozen forecasting and adaptation through a small set of calibration parameters.

7 Discussion and Conclusion

MER defines sparse parameter reuse through support-conditioned predictive response equivalence. Controlled audits, held-out composition, and frozen cross-dataset evaluation favor its full pipeline under the stated same-backbone protocols. The factorial associates both MER-formed banks and MER routing with lower mean MSE. Together, these results support response equivalence as an organizing principle for sparse parameter reuse in time-series forecasting.

AI USE STATEMENT

Generative AI tools were used to aid and polish the writing of this manuscript and to assist with literature discovery. All scientific and technical content, methods, results, figures, tables, and implementation choices are the authors' own work. The authors have reviewed and take full responsibility for the entire manuscript.

REPRODUCIBILITY STATEMENT

The main paper specifies the forecasting interface, reuse relation, routing construction, and comparison protocol. Appendices C and D provide stage-wise optimization settings, frozen components, model dimensions, prototype construction, data splits, transfer protocols, exposure labels, hardware, and software versions. Appendix Table 20 reports the per-dataset–horizon results underlying Table 1 and Figure 2B. Controlled and same-backbone forecasting metrics are summarized over five training runs. An anonymous implementation and data package containing the audit registry, class labels, configurations, run-level outputs, and table-generation scripts is available at `https://anonymous.4open.science/r/MER-8B61/`.

REFERENCES

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Stefano Soatto, and Pietro Perona. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Panagiotis Alimisis and Christos Diou. Learning robust intervention representations with delta embeddings. In *International Conference on Learning Representations*, 2026.

Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3366–3375, 2017.

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

Peng Chen, Yingying Zhang, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang, and Chenjuan Guo. Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting. In *International Conference on Learning Representations*, 2024.

Yuxiao Cheng, Ziqian Wang, Tingxiong Xiao, Qin Zhong, Jinli Suo, and Kunlun He. CausalTime: Realistically generated time-series for benchmarking of causal discovery. In *International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=iad1yyyGme`.

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

Shicheng Fan, Kun Zhang, and Lu Cheng. TRACE: Trajectory recovery for continuous mechanism evolution in causal representation learning. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

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.

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

Anirudh Goyal, Alex Lamb, Jordan Hoffmann, Shagun Sodhani, Sergey Levine, Yoshua Bengio, and Bernhard Schölkopf. Recurrent independent mechanisms. In *International Conference on Learning Representations*, 2021.

Lan Guo, Jie Xiao, Zhao Su, Jun Shen, Haoran Li, Weixia Ma, Qingguo Zhou, and Binbin Yong. Fuzzy-MoE: Interpretable regime-conditioned expert routing for non-stationary multivariate time series forecasting. *arXiv preprint arXiv:2608.20761*, 2026.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, 2019.

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*, 2021.

Qianyang Li, Xingjun Zhang, Shaoxun Wang, Tao Peng, and Jia Wei. FAME: Forecastability-aware mixture of experts for heterogeneous time series forecasting. *arXiv preprint arXiv:2606.08896*, 2026. URL https://arxiv.org/abs/2606.08896.

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*, volume 267 of *Proceedings of Machine Learning Research*, pp. 38940–38962, 2025.

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*, 2024.

Ronghao Ni, Zinan Lin, Shuaiqi Wang, and Giulia Fanti. Mixture-of-linear-experts for long-term time series forecasting. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pp. 4672–4680. PMLR, 2024.

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

Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4036–4044, 2018.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

Clemens Rosenbaum, Tim Klinger, and Matthew Riemer. Routing networks: Adaptive selection of non-linear functions for multi-task learning. In *International Conference on Learning Representations*, 2018.

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

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-MoE: Billion-scale time series foundation models with mixture of experts. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=e1wDDFmlVu.

Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 914–921, 2020.

Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B. Fox. Neural Granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4267–4279, 2022.

Rui Wang, Renhao Xue, Ray Razi, Huan Song, and Hannah R. Marlowe. AME-TS: Anchored mixture-of-experts for time series forecasting. *arXiv preprint arXiv:2605.25166*, 2026.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations*, 2024.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53140–53164, 2024.

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

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

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Han Zhang, Yufeng Chen, and Xinyi Niu. ACT: Anomaly-aware causal transformer with mixture-of-experts for time-series forecasting. *Knowledge-Based Systems*, 341:115774, 2026.

Xun Zheng, Bryon Aragam, Pradeep K. Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

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

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27268–27286, 2022.

Wanqi Zhou, Shuanghao Bai, Shujian Yu, Qibin Zhao, and Badong Chen. Jacobian regularizer-based neural Granger causality. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61763–61782, 2024.

A Operational Scope of Mechanism Equivalence

The reuse relation in Definition 1 concerns the baseline-subtracted predictive response represented by the model on a declared support and domain. It is neither global function equality outside that domain nor a causal statement without assumptions connecting selected lags, observational responses, and interventions. We reserve causal language for controlled generators with known structural equations. On observational data, the fingerprint is a model-based functional response descriptor. Horizon baselines remain outside the reuse class: MER asks whether perturbation-response parameters can be shared, not whether two units have the same unconditional forecast level.

To distinguish exact from approximate reuse, let $\mathcal{C}$ apply the declared canonicalization and let ν have full support on the canonical compact domain. For compatible lag-role sets, define

$$D_{\mathrm{reuse}}(M_u, M_v) = \inf_{\pi, T \in \mathcal{G}_{\mathrm{aff}}} \|\mathcal{C} f_u - \mathcal{C}(f_v \circ T)\|_{L^2(\nu)}, \tag{21}$$

where T respects the lag-preserving role bijection π; set the discrepancy to $+\infty$ for incompatible active supports. Definition 1 implies zero discrepancy for non-flat reuse-equivalent mechanisms. We use $D_{\mathrm{reuse}} \leq \epsilon$ only as an approximate operational relation; finite fingerprint distance and shared-prototype identity need not be equivalent to it. Without additional compatibility conditions on the canonical domains and reference measures, we do not claim symmetry or the triangle inequality for D_{reuse}, and the router does not use it as a metric.

Limitations. MER defines an operational criterion for predictive parameter reuse on declared supports and domains. Learned support selection, finite response probes, and a fixed prototype bank approximate the ideal relation in Section 3; its guarantees apply under the stated response and correspondence assumptions. The evaluation focuses on same-backbone sparse routing and target-channel-specific forecasts. Additional backbones, explicit role alignment, and joint multivariate prediction are directions for extending this framework.

B Proofs

B.1 Proof of Proposition 1

Reflexivity uses identity role, input, and output maps. For symmetry, every admissible input map is invertible because its permutation matrix is orthogonal and every diagonal scale is positive. The output inverse has positive scalar gain a_y^{-1} and horizon baseline $-a_y^{-1}\boldsymbol{b}_y$. For transitivity, compose the lag-preserving role bijections and positive input- and output-affine maps. A scalar gain followed by a horizon-baseline translation remains in the same output family, and the composed input map remains admissible and maps the first compact domain bijectively to the third. Hence $\sim_{\mathrm{reuse}}$ is reflexive, symmetric, and transitive.

B.2 Proof of Proposition 2

Equivariant input and gain normalization cancels the corresponding admissible maps in Definition 1. Subtracting the value at a reference input removes the horizon baseline, so u and v have the same normalized response contrast after applying the role permutation. Let r denote this common contrast, let $\boldsymbol{\beta}_u$ and $\boldsymbol{\beta}_v$ be unit-specific horizon baselines, and write $m(\widetilde{\boldsymbol{P}}) = \mathbb{E}[\widetilde{\boldsymbol{Y}} \mid \widetilde{\boldsymbol{P}}]$. For squared error, conditional bias–variance decomposition gives

$$\mathbb{E}[\|\widetilde{\boldsymbol{Y}} - (\boldsymbol{\beta} + e_\theta(\widetilde{\boldsymbol{P}}))\|_2^2 \mid \widetilde{\boldsymbol{P}}] = \mathbb{E}[\|\widetilde{\boldsymbol{Y}} - m(\widetilde{\boldsymbol{P}})\|_2^2 \mid \widetilde{\boldsymbol{P}}] + \|m(\widetilde{\boldsymbol{P}}) - (\boldsymbol{\beta} + e_\theta(\widetilde{\boldsymbol{P}}))\|_2^2. \tag{22}$$

The first term is irreducible. By well specification, some $\theta^\star$ represents r on the declared domain; the respective $\boldsymbol{\beta}_u$ and $\boldsymbol{\beta}_v$ absorb the two horizon baselines, and role-permutation equivariance supplies the corresponding representation after π. Thus the same response-expert parameterization, up to role indexing, minimizes both population risks after their baselines are optimized. This establishes existence for the stated decomposed predictor family. It does not imply that an approximate finite optimizer will recover $\theta^\star$, nor that arbitrary layerwise hidden representations share an identical full-network optimum.

B.3 Proof of Proposition 3

The uniform distribution on $[-1, 1]$ is symmetric and G is odd, so $G(X_t)$ and $-G(X_t)$ have the same distribution. Consequently, after adding identically distributed independent noise, the output marginals under $Z = +1$ and $Z = -1$ agree. Both classes also use the same input and parent support. Moreover,

$$\mathbb{E}[X_t G(X_t)] = \mathbb{E}[X_t^4] - \tfrac{3}{5}\mathbb{E}[X_t^2] = \tfrac{1}{5} - \tfrac{3}{5} \cdot \tfrac{1}{3} = 0. \tag{23}$$

Independence across time makes all nonzero-lag input and output autocovariances equal in the two classes, while the displayed identity makes the relevant input–output cross-covariance zero in both. On the declared domain, the only positive-affine bijection from $[-1, 1]$ to itself is the identity. The responses therefore differ by a negative output gain, which Definition 1 excludes, so they are not reuse-equivalent. The result concerns the listed population summaries; it does not assert equality of the full paired distribution of (X_t, Y_{t+1}).

B.4 Proof of Corollary 1

Equal class-conditional descriptor laws imply that the posterior class probability remains the uniform prior almost surely. Hence the Bayes error based on T is $1/2$, and no classifier measurable with respect to T can do better. Multiplying the wrong-route probability by the assumed conditional excess loss $\Delta_{\min}$ gives the stated forecasting lower bound.

B.5 Proof of Proposition 4

Let $\boldsymbol{\mu}_u^p$ and $\boldsymbol{s}_u^p$ be the context-derived coordinatewise parent medians and interquartile ranges in Equation 6, and let s_u^y be the corresponding target scale. Equivariance gives $\boldsymbol{\mu}_v^p = \boldsymbol{P}_\pi(\boldsymbol{A}\boldsymbol{\mu}_u^p + \boldsymbol{b})$, $\boldsymbol{s}_v^p = \boldsymbol{P}_\pi \boldsymbol{A}\boldsymbol{s}_u^p$, and $s_v^y = a_y s_u^y$. By the selection-equivariance assumption, the selected roles and interaction pairs correspond under π, including any ties. Input normalization cancels the positive input-affine map, and zero-query subtraction removes the horizon baseline. The common positive gain cancels under the single-parent RMS in Equation 10; the assumed scale bounds ensure that the scale floor is inactive. Single-parent and pairwise normalized contrasts therefore agree after π. The descriptors differ only by role permutation, so the common pooling encoder returns the same fingerprint. This conditional statement does not establish equivariance of the learned support scorer or its tie handling.

Proposition 5 (Top-2 stability)**.** *Let $\boldsymbol{\Psi}$ and $\widehat{\boldsymbol{\Psi}}$ be unit-norm oracle and estimated fingerprints, let all prototypes be unit-norm, and let $\gamma = s_{(2)} - s_{(3)}$ be the oracle cosine-score margin. If $\|\widehat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}\|_2 < \gamma/2$, both fingerprints select the same top-2 prototype set.*

B.6 Proof of Proposition 5

For any unit prototype $\boldsymbol{c}_k$, Cauchy–Schwarz gives

$$\left|\langle \widehat{\boldsymbol{\Psi}}, \boldsymbol{c}_k\rangle - \langle \boldsymbol{\Psi}, \boldsymbol{c}_k\rangle\right| \le \|\widehat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}\|_2. \tag{24}$$

Let this error be $\varepsilon < \gamma/2$. The estimated score of either true top-2 prototype can decrease by at most ε, while that of any remaining prototype can increase by at most ε. Their separation therefore remains at least $\gamma - 2\varepsilon > 0$, preserving the top-2 set.

Risk consequence. For unit u, write $\varepsilon_u = \|\widehat{\boldsymbol{\Psi}}_u - \boldsymbol{\Psi}_u\|_2$ and let γ_u be the oracle second-to-third cosine-score margin. Let $\boldsymbol{\alpha}_u^{\star(\ell)}$ and $\widehat{\boldsymbol{\alpha}}_u^{(\ell)}$ be the Top2Softmax weights obtained from the oracle and estimated fingerprints with the same prototypes and temperature T_ℓ. Assume the conditional loss ℓ_u lies in $[0, B]$ and, whenever the Top-2 sets agree, is L_{mix}-Lipschitz in $\sum_\ell \|\widehat{\boldsymbol{\alpha}}_u^{(\ell)} - \boldsymbol{\alpha}_u^{\star(\ell)}\|_1$. Then

$$\mathcal{R}(\widehat{\boldsymbol{\alpha}}) - \mathcal{R}(\boldsymbol{\alpha}^\star) \le L_{\mathrm{mix}} \left(\sum_{\ell=1}^{L_{\mathrm{moe}}} T_\ell^{-1}\right) \mathbb{E}[\varepsilon_u \mathbf{1}\{\varepsilon_u < \gamma_u/2\}] + B \Pr[\varepsilon_u \ge \gamma_u/2]. \tag{25}$$

On the event $\varepsilon_u < \gamma_u/2$, Proposition 5 fixes the selected set. Each cosine score changes by at most ε_u, so their difference changes by at most $2\varepsilon_u$. The logistic representation of the two-way softmax has derivative at most $1/4$, giving $\|\widehat{\boldsymbol{\alpha}}_u^{(\ell)} - \boldsymbol{\alpha}_u^{\star(\ell)}\|_1 \le \varepsilon_u/T_\ell$. Lipschitz continuity bounds the resulting within-set loss change. On the complementary event, boundedness contributes at most B. Thus Equation 25 includes both changed weights on a stable set and changed support; it does not equate set agreement with identical routing. The oracle-routed source bank may still differ from an unrestricted optimal predictor, leaving a separate bank-approximation term.

C Experimental Protocol

Data and preprocessing. Each benchmark retains its timestamp order and channel semantics and uses the deterministic chronological boundaries declared in Appendix D. Context normalization is recomputed from each observed window by Equation 6; no corpus-level statistic is fitted on a target training split during zero-shot evaluation. ILI uses the CDC weekly series from 2002–2021. Weather uses the MPI-BGC local meteorological series sampled at 10-minute resolution (June 2020 release).

The controlled generators use disjoint training, validation, and test seeds and disjoint parameter ranges. Validation data alone determine PMDA empirical CDFs, tail thresholds, query-coverage cutoffs, and probe hyperparameters.

The CausalTime-based controlled audit uses the five MER-defined instances in Table 4. The instance identifiers, response families, and reuse classes belong to our evaluation protocol rather than the original CausalTime benchmark. Each label is fixed from the declared response before training or evaluation. Positive-affine variants within a reuse class form equivalent pairs; a cross-instance pair is non-equivalent only when its responses cannot be related by Definition 1. A pair without a predeclared label is omitted rather than classified post hoc from MER output.

Table 4: MER-defined registry for the CausalTime-based controlled audit. Counts are target-window units $u = (e, j, t)$, not independent sequences. The five instances span four response families and five reuse-equivalence classes.

MER audit instance	Response family	Reuse-class ID	Equivalence basis	Train/val/test
`hermite_sign_pos`	Hermite sign	`hermite_pos`	Positive output scaling within the $+H_3$ class, $a_y > 0$.	5000/1000/1000
`hermite_sign_neg`	Hermite sign	`hermite_neg`	No admissible map to `hermite_pos`; that map would require $a_y < 0$.	5000/1000/1000
`saturation_tanh`	Saturation	`saturation_affine_eq`	Positive-affine maps within the $\tanh$ response class; centered sigmoid appears only in the composition study.	4000/800/800
`even_response_sq`	Even response	`even_sq`	Positive-affine coordinates with the induced response reparameterization and positive output gain.	3000/600/600
`interaction_xz`	Interaction	`interaction_xz`	Positive-affine coordinates with the induced response reparameterization and positive output gain.	3000/600/600

Table 4 contains 20,000, 4,000, and 4,000 training, validation, and test target-window units; Tables 5 and 6 specify nuisance splits and response families. The CausalTime-based audit contains only the $\tanh$ saturation instance; centered sigmoid and $\tanh$ form a same-class control only in the separately evaluated composition study.

Table 5: Disjoint nuisance-parameter supports for the CausalTime-based audit. Half-open intervals remove shared endpoints. The operating-range split deliberately evaluates extrapolation outside the central training range.

Nuisance parameter	Training	Validation	Test
Input scale	$[0.5, 1.5)$	$[1.5, 2.5)$	$[2.5, 3.5]$
Input offset	$[-1.0, 0.0)$	$[0.0, 1.0)$	$[1.0, 2.0]$
Output scale	$[0.8, 1.2)$	$[1.2, 1.6)$	$[1.6, 2.0]$
Output offset	$[-0.5, 0.0)$	$[0.0, 0.5)$	$[0.5, 1.0]$
Noise scale	$[0.01, 0.05)$	$[0.05, 0.10)$	$[0.10, 0.20]$
Operating range	$[-2, 2)$	$[-4, -2)$	$[2, 4]$

Table 6: Response families across the controlled evaluations. Nuisance variables vary independently of the structural response; Table 5 gives the CausalTime-based audit ranges.

Family	Declared response	Audit role
Hermite sign	$f_{\pm}(x) = \pm a(x^3 - 3x)$, $a > 0$	Matched second-order statistics, different response sign.
Saturation	$\tanh(ax)$ and $\sigma(2ax) - \frac{1}{2} = \frac{1}{2}\tanh(ax)$	Affine-equivalent positive control; never a separation pair.
Even response	x^2 under shifted/scaled marginals	Operating-point and marginal variation.
Interaction	xz with signed and lagged alternatives	Pairwise contribution and interaction separation.
Composition	$g_1 = \tanh x$, $g_2 = \sigma(2x) - .5$, $g_3 = x^2$, $g_4 = xz$	Three reuse classes; train on $\{1,2\}, \{2,3\}, \{1,4\}$ and hold out $\{3,4\}$.

Pair construction is fixed before model evaluation. From the five predeclared audit classes in the test split, we deterministically sample without replacement 2,400 unordered same-class target-window pairs (480 per reuse class) and 3,200 unordered different-class pairs (320 per unordered pair of classes); a target-window unit may occur in multiple pairs. An additional 500 candidate pairs were drawn from auxiliary CausalTime generator variants outside the five predeclared audit classes in Table 4. Because their reuse relation was not certified a priori under Definition 1, they were excluded before model evaluation and PMDA tail filtering rather than assigned labels from MER outputs; they do not enter either eligible denominator. Validation pairs alone fit the four pattern-distance empirical CDFs and determine the frozen $q_{0.95}$ and $q_{0.05}$ thresholds. Applying these thresholds to the test candidate pool retains 120 same-class/high-pattern-distance pairs for MRC and 160 different-class/low-pattern-distance pairs for MSS, observed retention rates of 5.0% in both pools. No test-set quantile is used to enforce these rates. The anonymous release at `https://anonymous.4open.science/r/MER-8B61/` is available and includes the registry, pair construction, and table-generation files.

Terminology and exposure labels. FM exposure denotes possible inclusion of a target in foundation-model pretraining. MER source-stage data denotes the data used in this study to train the support-response pipeline, calibrate fingerprints, form prototypes, and co-adapt the sparse backbone and experts. Target-stage adaptation denotes optimization after a dataset is designated as the evaluation target. Unless stated otherwise, *zero-shot target adaptation* means zero target-stage parameter updates; it does not imply that the foundation model or MER source stage excluded the target. ETT FM exposure is unresolved and marked *Unknown*; PEMS, ILI, and Weather are conservatively marked *Yes*. ETT and PEMS directions name a distinct source dataset and therefore mark target-in-source as *No*. ILI and Weather use a declared source-mixture checkpoint. Available checkpoint provenance does not permit verified target-series exclusion; we therefore conservatively mark target-in-source as *Unknown*. No stronger exclusion claim is made.

Baseline identity and comparison matching. Pattern denotes our implemented pattern-based routing control on the Moirai-MoE backbone, rather than a separately published model. The backbone is Moirai-MoE (Liu et al., 2025), using the official Uni2TS implementation (`https://github.com/SalesforceAIResearch/uni2ts`) and the `Salesforce/moirai-moe-1.0-R-small` checkpoint. Pattern and the native Moirai-MoE baseline in Tables 1 and 20 are distinct routing configurations. Their names are not interchangeable: the native baseline supplies the denominator for forecasting reductions, and Pattern supplies the denominator for MRA.

Moirai-MoE and MER in Tables 1 and 20 use the same chronological splits in Table 12, identical test windows and forecasting horizons, and the same MSE computation and aggregation; they do not use separate original-paper default splits. The comparison also fixes the named source dataset or declared source artifact, backbone and active-expert budget, forecasting co-adaptation optimizer and update budget, and checkpoint-selection rule. Throughout the paper, *protocol-matched* refers to these evaluation and forecasting co-adaptation conditions. The MER pipeline additionally includes source-side response, fingerprint, and prototype calibration, executed for named-source transfers and inherited from the declared source-mixture checkpoint for ILI/Weather; matched forecasting updates therefore do not imply equal total training cost. Table 1 reports inference cost, not the cost of those additional source-side stages.

Evaluation scale and cell aggregation. All reported real-data MSEs use the observed-window normalized target scale of Equation 6. Model outputs are converted to this common scale before comparison with normalized targets; the reported MSE is not computed on the original data scale. Within each dataset–horizon cell, the shared evaluator averages squared errors equally over all test windows, forecast steps, and evaluated target channels. Moirai-MoE and MER use identical target-channel sets, with no single-channel or representative-channel selection, and observed context length $L = 512$.

Same-backbone routing controls. Pattern uses trajectory, Fourier-magnitude, and autocorrelation patterns; Parent represents selected parent support; Graph uses a graph embedding; and Local-response uses local response derivatives. These are alternative descriptor families, not interchangeable names for the native Moirai-MoE router. Each enters a complete source-trained routing pipeline. The

controls share the backbone, expert architecture, Top-2 active-expert budget, source data or declared source artifact, Stage-4 optimizer and update budget, load-balancing coefficient, and checkpoint-selection rule. In the factorial, source-derived index alignment and temperature accompany each pipeline, as specified below; zero-shot target evaluation freezes all components. This comparison does not equate the cost of constructing the different descriptors.

Controlled and real-data training pathways. The two pathways are independent. Controlled synthetic pairs are used only in controlled experiments and supply no labels or examples to real-data fingerprint calibration, prototype formation, or expert training. Tables 7 and 8 separate their optimization order and frozen components. For ETT/PEMS, S is the named source dataset in Table 15. For ILI/Weather, the source-stage entries describe the provenance of the loaded source-mixture checkpoint rather than a locally rerun Stage 1–4 pipeline on an identified dataset; the retained checkpoint record does not expose its constituent source mixture.

The support-response stage minimizes

$$\mathcal{L}_{\mathrm{resp}} = \mathbb{E}_u |\mathcal{H}_r|^{-1} \| g_\phi(\boldsymbol{c}_u, \widetilde{\boldsymbol{p}}_u) - \widetilde{\boldsymbol{y}}_{u,t+\mathcal{H}_r} \|_2^2. \tag{26}$$

The straight-through support scorer receives gradients only through this forecasting objective: the forward pass executes deterministic hard top-K_p selection, while the backward pass differentiates through an untempered softmax (temperature 1) over the same support logits, without stochastic support sampling. Fingerprint calibration uses $\mathcal{L}_{\mathrm{query}} + \lambda_{\mathrm{ctx}} \mathcal{L}_{\mathrm{ctx}}$ on controlled pairs and $\mathcal{L}_{\mathrm{query}}$ alone on observational data; λ_{ctx} is fixed from controlled source validation.

Table 7: Controlled-suite training pathway.

Stage	Component	Operation
0	Backbone	Random initialization.
1	Support response	Optimize Equation 26 on the controlled training split; then freeze.
2	Fingerprint	Optimize $\mathcal{L}_{\mathrm{query}} + \lambda_{\mathrm{ctx}} \mathcal{L}_{\mathrm{ctx}}$ on controlled training pairs; then freeze.
3	Prototypes	K-means++ on 100k fingerprints sampled with replacement from controlled training units; then freeze.
4	Forecast model	Co-adapt backbone, layer experts, and T_ℓ on the controlled training split under fixed fingerprint geometry.
–	Evaluation	Evaluate once on the held-out controlled split; validation alone fixes audit thresholds.

Table 8: Real-data source and target pathway. No controlled pair or mechanism label enters this pathway. Source stages are executed on the named source S for ETT/PEMS; for ILI/Weather, the source-stage column records inherited checkpoint provenance rather than locally repeated operations.

Stage	Component	MER source stage	Target evaluation
0	Backbone	Initialize from Moirai-MoE.	No operation.
1	Support response	Train on S with $\mathcal{L}_{\mathrm{resp}}$; freeze.	Frozen.
2	Fingerprint	Train on S with $\mathcal{L}_{\mathrm{query}}$; $\mathcal{L}_{\mathrm{ctx}} = 0$; freeze.	Frozen.
3	Prototypes	K-means++ on 100k fingerprints from S; freeze.	Frozen.
4	Forecast model	Co-adapt backbone, experts, and T_ℓ on S under fixed routing geometry.	Zero-shot: all frozen.
5	Optional few-shot	No operation.	Update only T_ℓ and the channelwise output affine map.

In both pathways, final co-adaptation minimizes the forecast loss plus an identical sparse-expert load-balancing regularizer for every protocol-matched router. The response pipeline, fingerprint encoder, and prototypes receive no gradients at this stage. In the real-data pathway, $\mathcal{L}_{\mathrm{query}}$ reconstructs only canonical contrasts of the frozen source-trained response model, so the real-data router receives no supervision from synthetic mechanism-equivalence labels.

Frozen and adapted transfer. The zero-shot protocol loads a source checkpoint and evaluates the target without optimizing any parameter or recomputing support, response, fingerprint, prototype, or audit threshold from target training data. Its only normalization is the deterministic observed-window transform in Equation 6. Few-shot protocols use the first 1%, 5%, or 10% of chronologically available target training windows and update only routing temperatures and the output map $\widehat{y}_j^{\mathrm{cal}} = a_j \widehat{y}_j + b_j$; the response pipeline, fingerprint encoder, prototypes, backbone, experts, attention blocks, and forecast head remain frozen. Each fraction runs exactly 4,800 updates and reports the final iterate; neither source nor target validation selects a Stage 5 checkpoint. Because the number of windows varies across fractions, this fixed update count does not equalize effective epochs or revisit counts, so the few-shot curves are descriptive. The fully adapted reference is a separate target-training protocol initialized from the source checkpoint at the end of Stage 4. It uses the Stage 4 architecture, optimizer, and hyperparameters in Table 9, with all model parameters trainable on 100% of the target training split, without subsampling. Checkpoints are selected by target-validation forecasting MSE using the Stage 4 stopping rule; this selection applies only to the fully adapted reference, not zero-shot or few-shot evaluation. The result is an empirical full-adaptation reference, not a guaranteed performance bound.

Coverage isolation. The calibration reservoir contains normalized parent tuples from source training data only. The cutoff is the 95th percentile of source-validation k-NN distances to the frozen source-training reservoir. A source-fixed coverage mask restricts the queries sampled by $\mathcal{L}_{\mathrm{query}}$ during fingerprint calibration. It never becomes a fingerprint feature, changes a target-time route, or uses target training data; no off-support query is interpreted as an intervention. The neighbor cutoff, ϵ_{flat}, and all PMDA thresholds are fixed before target designation.

Cost protocol. FLOPs and latency are normalized to one target-window unit $u = (e, j, t)$ and include support selection, all canonical response queries, fingerprint encoding, prototype matching, sparse expert execution, and the complete direct forecast horizon. Multichannel origin-level execution batches these target units and therefore depends on the number of targets. FLOPs report the mean $\pm$ standard deviation across dataset–horizon configurations. Protocol-matched latency is measured on one NVIDIA A100 80GB at batch size one and BF16, using 100 warm-ups and 1,000 synchronized passes. The displayed means rise from 17.5 ms to 20.4 ms (16.6%) under this matched hardware protocol. Latency reports mean $\pm$ standard deviation across these passes on one fixed checkpoint, whereas forecasting metrics aggregate five training runs.

Constituent Top-2 evaluation. For each test window, success requires the classes associated with the two highest-probability predicted slots to include both true constituent classes. Each window contributes one binary success indicator, so the Top-2 percentage is 100 times the number of successful test windows divided by the number of test windows. Selecting two slots assigned to the same class does not recover two distinct constituent classes. The slot-to-class assignment is fixed using source training data before test evaluation; multiple slots may map to one class. This class assignment is distinct from the router-to-index lookup in the bank-by-router factorial.

Aggregation and statistical scope. Dataset-family macro-averages are formed by first averaging horizons equally within each dataset and then datasets equally within each family. Appendix Table 20 exposes the per-dataset–horizon means underlying Table 1; Table 16 reports every available transfer direction. We use

$$\mathrm{CompGap} = 100\,\frac{\mathrm{MSE}_{\mathrm{held}} - \mathrm{MSE}_{\mathrm{seen}}}{\mathrm{MSE}_{\mathrm{seen}}},$$
$$\mathrm{FETG}_r = 100\,\frac{\mathrm{MSE}_r - \mathrm{MSE}_{\mathrm{adapt}}}{\mathrm{MSE}_{\mathrm{adapt}}}, \qquad \mathrm{MRA} = 100\,\frac{\mathrm{MSE}_{\mathrm{Pattern}} - \mathrm{MSE}_{\mathrm{MER}}}{\mathrm{MSE}_{\mathrm{Pattern}}}. \tag{27}$$

Controlled and same-backbone forecasting entries report mean $\pm$ standard deviation over five training repetitions under fixed splits. For Table 2A, each run first averages the three environments equally before aggregation across runs; the MER Top-2 standard deviation is 2.65 percentage points across these run-level averages. The runs are unpaired and the dataset–horizon cells share datasets, windows, and trained models, so 40/40 summarizes coverage rather than an inferential test. CompGap, FETG, and MRA are recomputed from the displayed aggregate means. In the factorial, each row fixes the bank, backbone, and forecast head, while each column substitutes the complete source-trained

routing pipeline—descriptor, source-derived index alignment, and temperature. This contrasts bank formation with complete pipeline choice. Run-level outputs and table-generation scripts are available in the anonymous release.

D REPRODUCIBILITY CONFIGURATION

This appendix records the fixed configuration used for the reported runs. Controlled and real-data values are abbreviated C and R. Split boundaries and all source-selected hyperparameters are deterministic and independent of the training seed.

D.1 STAGE-WISE OPTIMIZATION

Table 9 specifies the optimizer, schedule, and update budget for each training stage.

Table 9: Optimization configuration by stage. Learning rates, batch sizes, and maximum steps are reported as controlled/real-data values; Stage 5 uses the fixed 4,800-update schedule shown and applies only to real-data few-shot calibration.

Stage	Component	Optimizer (hyperparameters)	LR (C/R)	Schedule; warmup	WD	Batch (C/R)	Max steps (C/R)
0	Backbone initialization	–	–	–	–	–	–
1	Support–response	AdamW ($\beta_1 = .93, \beta_2 = .967$)	$2.7\times10^{-4}/4.3\times10^{-5}$	Cosine; .053	.047	37/53	87,300/62,100
2	Fingerprint encoder	AdamW ($\beta_1 = .91, \beta_2 = .973$)	$8.1\times10^{-4}/1.6\times10^{-4}$	Linear; .031	.013	37/53	52,400/38,100
3	Prototype formation	Offline K-means++	–	–	–	–	–
4	Forecast model	AdamW ($\beta_1 = .93, \beta_2 = .967$)	$7.3\times10^{-5}/9.7\times10^{-6}$	Cosine; .027	.052	37/53	124,700/89,300
5	Few-shot calibration	SGD (momentum 0)	3.1×10^{-3}	Fixed; –	0	16	4,800

Controlled Stage 0 uses random initialization; real-data Stage 0 loads `Salesforce/moirai-moe-1.0-R-small`. Stage 1 is evaluated every 873 steps and stops after seven evaluations without validation-MSE improvement. Stage 2 is evaluated every 654 steps; checkpoint selection uses $\mathcal{L}_{\text{query}} + \lambda_{\text{ctx}}\mathcal{L}_{\text{ctx}}$ for controlled data and $\mathcal{L}_{\text{query}}$ for real data. Stage 4 is evaluated every 1,247 steps and stops after five evaluations without validation-MSE improvement; its checkpoint criterion is validation forecasting MSE, not $\mathcal{L}_{\text{lb}}$. In Stages 1, 2, and 4, the best checkpoint is retained and training may stop before the stated maximum. Real-data Stage 5 instead runs the fixed 4,800-update schedule and evaluates its final iterate, without source- or target-validation checkpoint selection.

All stages use gradient clipping at norm 0.87. Dataset construction and chronological splits are fixed before training. Within each method, seeds $\{17, 203, 491, 768, 1355\}$ initialize five independent training repetitions. These repetitions were not constructed as matched randomized blocks across methods; their labels are therefore used for reproducibility, not paired inference. Eligible forward and loss operations run under CUDA BF16 autocast; model parameters and optimizer states remain FP32, and gradient scaling is disabled.

D.2 LOSS, COVERAGE, AND ROUTING CONSTANTS

Table 10 lists the fixed losses, coverage rules, and routing constants.

Table 10: Fixed loss, coverage, fingerprint, and routing configuration.

Item	Configuration
Context loss	$\lambda_{\mathrm{ctx}} = 0.1$ on controlled data, selected from $\{0.013, 0.047, 0.1, 0.18\}$ on controlled source validation; $\lambda_{\mathrm{ctx}} = 0$ on real data.
Load balancing	$\lambda_{\mathrm{lb}} = 0.01$ for every protocol-matched sparse router.
Flat-response floor	$\epsilon_{\mathrm{flat}} = 10^{-6}$, fixed globally after source-validation selection.
Coverage	$k_{\mathrm{cov}} = 5$, selected from $\{3, 5, 7\}$ on source validation; cutoff is the 95th percentile of source-validation k-NN distances to the frozen source-training reservoir.
Coverage reservoir	10,000 normalized parent tuples sampled uniformly from source training data; squared Euclidean distance.
Coverage use	Source-covered mask for $\mathcal{L}_{\mathrm{query}}$ sampling during calibration; never a fingerprint feature or routing input.
Fingerprint	LayerNorm followed by L_2 normalization; output lies on the unit sphere in $\mathbb{R}^{128}$.
Prototype temperature	$T_\ell = \mathrm{softplus}(t_\ell)$ with effective initialization $T_\ell = 0.13$ ($t_\ell \approx -1.97$); positive, trainable, and layer-specific.
Routing	Cosine similarity to unit-normalized prototypes followed by Top-2 softmax.
Support relaxation	Untempered softmax (temperature 1) in the backward pass; deterministic hard Top-K_p selection in the forward pass.

D.3 Auxiliary network architecture

Table 11 specifies the dimensions and nonlinearities of the MER-specific modules.

Table 11: Architecture of MER-specific modules. Output layers are linear unless stated otherwise.

Module	Layers	Hidden	Output	Activation	Input and structural details
Support scorer MLP_s	2	128	1	GELU	$[\boldsymbol{q}_u, \boldsymbol{h}_{i,\tau}, \boldsymbol{q}_u \odot \boldsymbol{h}_{i,\tau}] \in \mathbb{R}^{1152}$.
Context encoder	2	128	384	GELU	Four attention heads, pre-LN; produces $\boldsymbol{q}_u$ and $\boldsymbol{h}_{i,\tau}$.
Response head g_ϕ	2	128	4	GELU	Jointly predicts the four relative-grid horizons.
Single-parent encoder	2	256	128	GELU	17-dimensional input: one lag and 4×4 nonzero-anchor responses.
Interaction encoder	2	256	128	GELU	42-dimensional input: 17-d endpoint sums, 17-d products, and eight contrasts.
Aggregation map ρ	2	64	128	GELU	256-dimensional concatenation of the two summed 128-d embeddings.
Transient decoder d_ω	2	128	4	GELU	134-dimensional input: 128-d fingerprint and 6-d canonical parent query; predicts the baseline-subtracted canonical response contrast.
Final fingerprint	–	–	128	–	LayerNorm followed by L_2 unit normalization.
Prototype bank	–	–	$K = 32$	–	Centroids are L_2-normalized after clustering.

D.4 Prototype formation

We use `sklearn.cluster.KMeans` from scikit-learn 1.2.2 on 100,000 unit-normalized source fingerprints. For the controlled audit, this count comprises repeated sampling with replacement from the 20,000 training target-window units in Table 4; repetitions are retained and do not represent 100,000 independent windows or sequences. The configuration is $K = 32$, `init=k-means++`, `n_init=13`, `random_state=73`, `max_iter=300`, and `tol=`10^{-4}. K-means minimizes squared Euclidean distance; centroids are L_2-normalized after fitting, making subsequent cosine routing compatible with the unit-sphere fingerprint geometry.

D.5 Dataset configurations

Table 12 gives the forecasting horizons and chronological split boundaries.

Table 12: Dataset horizons and deterministic chronological splits. Counts are timesteps rather than sliding-window samples.

Dataset	Horizons H	Total	Train/val/test	Boundary or date protocol	Missing
ETTh1, ETTh2	$\{96, 192, 336, 720\}$	17,420	10,452/3,484/3,484	Train 0–10,451; val 10,452–13,935; test 13,936–17,419.	0
ETTm1, ETTm2	$\{96, 192, 336, 720\}$	69,680	41,808/13,936/13,936	Train 0–41,807; val 41,808–55,743; test 55,744–69,679.	0
PEMS03	$\{12, 24, 48, 96\}$	26,208	15,725/5,242/5,241	Chronological 6:2:2; 2018-09-01 to 2018-11-30.	None
PEMS04	$\{12, 24, 48, 96\}$	16,992	10,195/3,399/3,398	Chronological 6:2:2; 2018-01-01 to 2018-02-28.	None
PEMS07	$\{12, 24, 48, 96\}$	28,224	16,934/5,645/5,645	Chronological 6:2:2; 2017-05-01 to 2017-08-31.	None
PEMS08	$\{12, 24, 48, 96\}$	17,856	10,714/3,571/3,571	Chronological 6:2:2; 2016-07-01 to 2016-08-31.	None
ILI	$\{24, 36, 48, 60\}$	966	676/145/145	Approximately 70:15:15; 2002–2015/2016–2018/2019–2021, weekly.	None
Weather	$\{96, 192, 336, 720\}$	52,696	31,618/10,539/10,539	Fixed chronological boundaries; June 2020 release, 10-minute sampling.	0

Sliding windows are generated deterministically from these timestep boundaries with context length 512. Their sample counts depend on H and the boundary-context convention; exact per-(H, split) counts are available in the anonymous data manifest. ILI and Weather retain the conservative exposure labels in Table 15: FM exposure is marked *Yes*, both load the declared source-mixture checkpoint, and available checkpoint provenance does not permit verified target-series exclusion; target-in-source is therefore conservatively marked *Unknown*. They support frozen evaluation without target-stage fitting; target exclusion from MER source-stage data remains unverified.

D.6 Software and hardware

Table 13 records the software versions and hardware for the controlled and same-backbone routing studies.

Table 13: Execution environment.

Component	Version or setting
Python	3.9.16
PyTorch	2.0.1+cu118
CUDA / NVIDIA driver	11.8 / 525.105.17
scikit-learn / NumPy	1.2.2 / 1.24.3
Training hardware	4× NVIDIA A100 80GB (PCIe)
Inference hardware	One NVIDIA A100 80GB; batch size 1; BF16

Quantitative evidence comprises controlled audits and same-backbone routing comparisons under the evaluation and forecasting co-adaptation protocol defined in Appendix C; cross-architecture generality remains outside this evaluation.

E PRIMARY RESULT DETAILS

Table 14: Held-out even–interaction function composition in three designed nuisance calibrations referenced to aggregate marginal and noise scales. The ETT/PEMS/ILI–Weather labels do not imply dataset-level generative fidelity. The two saturation instances are affine-equivalent under Definition 1, so the study contains three reuse classes and does not test recovery of four distinct mechanisms. CompGap is computed from displayed means.

Environment	Method	Seen MSE ↓	Held-out MSE ↓	CompGap ↓	Constituent Top-2 ↑
ETT-calibrated	Dense	.075±.004	.120±.007	60.0%	NA
	Pattern	.084±.005	.128±.008	52.4%	43.1±1.8%
	MER	**.060±.003**	**.078±.005**	**30.0%**	**80.2±2.5%**
PEMS-calibrated	Dense	.078±.003	.126±.006	61.5%	NA
	Pattern	.087±.004	.132±.007	51.7%	44.8±1.6%
	MER	**.062±.003**	**.080±.004**	**29.0%**	**82.1±2.7%**
ILI/Weather-calibrated	Dense	.078±.004	.126±.007	61.5%	NA
	Pattern	.087±.005	.130±.008	49.4%	44.8±1.5%
	MER	**.061±.004**	**.079±.005**	**29.5%**	**82.1±2.8%**

In Table 14, MER limits the relative held-out degradation to 29.0–30.0% and recovers both true constituent classes in more than 80% of test windows under the Top-2 criterion in Appendix C. These summaries evaluate held-out compositions under distinct marginal and noise calibrations. Consistent with Definition 1, the protocol contains three reuse classes.

Table 15: Exposure audit for frozen evaluation. FM denotes possible foundation-model pretraining exposure. MER source is the data used for MER-specific source training. "Target in source?" concerns only that MER source stage.

Target/direction	FM exposure	MER source-stage data	Target in source?
ETTm1→ETTm2	Unknown	ETTm1	No
ETTm2→ETTm1	Unknown	ETTm2	No
ETTh1→ETTh2	Unknown	ETTh1	No
ETTh2→ETTh1	Unknown	ETTh2	No
PEMS03→PEMS04	Yes	PEMS03	No
PEMS04→PEMS03	Yes	PEMS04	No
PEMS07→PEMS08	Yes	PEMS07	No
PEMS08→PEMS07	Yes	PEMS08	No
ILI target	Yes	Source-mixture checkpoint	Unknown
Weather target	Yes	Source-mixture checkpoint	Unknown

Table 16: Frozen target evaluation by direction. Entries are MSE ± standard deviation; zero-shot denotes no target-stage fitting. ILI/Weather target exclusion from MER source-stage data is unverified.

Target/direction	Pattern	Parent	Graph	MER	Fully adapted
ETTm1→ETTm2	.33472±.01381	.32145±.01274	.31382±.01163	**.30425±.01058**	.27385±.00924
ETTm2→ETTm1	.35641±.02017	.34128±.01832	.33217±.01695	**.32425±.01573**	.29174±.01402
ETTh1→ETTh2	.32538±.01592	.31274±.01456	.30418±.01321	**.29575±.01208**	.26583±.01075
ETTh2→ETTh1	.33852±.02413	.32491±.02207	.31546±.01984	**.30675±.01823**	.27519±.01654
PEMS03→PEMS04	.11983±.00945	.11472±.00883	.11138±.00812	**.10850±.00745**	.07915±.00562
PEMS04→PEMS03	.11827±.01012	.11358±.00941	.11024±.00873	**.10750±.00798**	.07108±.00513
PEMS07→PEMS08	.11745±.00876	.11293±.00805	.10971±.00742	**.10700±.00681**	.08752±.00601
PEMS08→PEMS07	.12018±.01054	.11527±.00968	.11205±.00897	**.10900±.00826**	.08073±.00584
ILI target	.30685±.01732	.29418±.01605	.28574±.01483	**.27850±.01362**	.25137±.01218
Weather target	.30792±.01847	.29543±.01712	.28715±.01576	**.27950±.01435**	.25189±.01345

Table 17: Direction-level mechanism reuse advantage (MRA) and fixed-expert transfer gap (FETG), recomputed from the displayed means in Table 16. MRA is the reduction relative to Pattern; FETG_0 is the remaining zero-shot gap relative to the fully adapted reference. "Src." denotes target exclusion from MER source-stage data; "Frozen" denotes zero target-stage updates with target exclusion from MER source-stage data unverified.

Target/direction	MRA ↑	FETG_0 ↓	Scope
ETTm1→ETTm2	9.1%	11.1%	Src.
ETTm2→ETTm1	9.0%	11.1%	Src.
ETTh1→ETTh2	9.1%	11.3%	Src.
ETTh2→ETTh1	9.4%	11.5%	Src.
PEMS03→PEMS04	9.5%	37.1%	Src.
PEMS04→PEMS03	9.1%	51.2%	Src.
PEMS07→PEMS08	8.9%	22.3%	Src.
PEMS08→PEMS07	9.3%	35.0%	Src.
ILI target	9.2%	10.8%	Frozen
Weather target	9.2%	11.0%	Frozen

Table 17 shows positive MRA in every direction, ranging from 8.9% to 9.5%. FETG quantifies the remaining gap to the fully adapted reference alongside MRA. Exposure and source-stage provenance follow Table 15. These direction-level results complement the per-horizon results in Table 20.

F Additional Results and Ablations

Table 18: CausalTime-based controlled-audit results underlying Figure 2. MER-defined reuse labels are declared from responses under Definition 1; pairs without a predeclared label are excluded. Values are five-run means ± standard deviations.

Routing signal	MRC ↑	MSS ↑	MSE ↓
Pattern	.537±.016	.487±.019	.167±.008
Parent support	.623±.018	.558±.017	.146±.007
Graph	.651±.015	.603±.014	.139±.006
Local response	.698±.013	.663±.012	.121±.006
MER	**.804±.014**	**.793±.011**	**.097±.004**
Oracle	.899±.013	.908±.013	.076±.003

The Oracle in Table 18 has access to true reuse classes and changes only routing decisions; it does not update the expert bank or other model parameters. The 32 prototype slots need not map one-to-one to reuse classes. Using the class assignment fixed on the source training split, the Oracle selects the highest-predicted-probability slot among those assigned to the required true class when multiple slots are eligible. The $\log 2$ normalization in Equation 20 applies to the Jensen–Shannon routing distance used by MRC/MSS; it does not define the constituent Top-2 assignment percentage in Table 2A.

Table 19: Bank-by-router factorial on frozen-target ETT macro MSE. Each row fixes the source-formed expert bank, backbone, and forecast head; columns substitute the complete source-trained routing pipeline. Blue intensity denotes lower MSE and red intensity higher MSE; entries are MSE ± standard deviation, with the best router per row in bold.

Source bank	Pattern router	Parent router	Graph router	MER router
Pattern	.33876±.01412	.34512±.01485	.34183±.01452	**.33241±.01378**
Parent	.33125±.01387	.32510±.01345	.32743±.01372	**.31987±.01305**
Graph	.32217±.01352	.31974±.01328	.31641±.01295	**.31128±.01263**
MER	.31452±.01287	.31183±.01254	.30946±.01228	**.30775±.01203**

In Table 19, the MER routing pipeline has the lowest ETT macro MSE for every fixed source bank, and the MER-formed bank has the lowest mean in every routing column. With the Pattern pipeline fixed, replacing the Pattern bank by the MER bank changes .33876 to .31452 ($-.02424$); with the Pattern bank fixed, replacing the Pattern pipeline by MER changes .33876 to .33241 ($-.00635$). This displayed slice associates a larger change with bank formation, but interactions and the system-level pipeline substitution preclude causal allocation to one module; MER bank and routing jointly attain .30775.

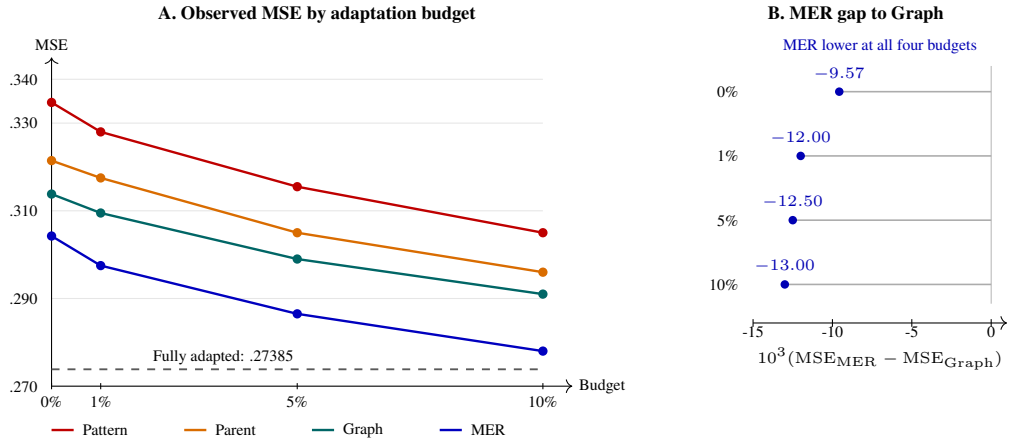


Figure 3: Single-direction few-shot illustration on ETTm1→ETTm2. MER has the lowest mean MSE among the four routers at 0%, 1%, 5%, and 10% target data; its advantage over Graph widens from .00957 to .01300. Panel A connects the four observed means as a visual guide and shows the fully adapted reference (.27385) as a horizontal dashed line. Panel B reports $\mathrm{MSE}_{\mathrm{MER}} - \mathrm{MSE}_{\mathrm{Graph}}$. At nonzero budgets, only routing temperatures and the output calibration map update; full target retraining is a separate reference.

Figure 3 is a single-direction illustration of the zero-shot result. MER has the lowest mean MSE among the four routers at all observed target-data budgets (0%, 1%, 5%, and 10%). Its MSE advantage over Graph increases from .00957 at zero-shot to .01300 at 10%, consistent with greater benefit from the permitted temperature and output-calibration updates in this direction. MER reaches .27800 MSE at 10%, remaining above the fully adapted reference of .27385.

G PER-DATASET AND PER-HORIZON FORECASTING RESULTS

Table 20: Per-dataset–horizon MSE underlying the family macros in Table 1. Each cell is the mean over five independently initialized runs with seeds $\{17, 203, 491, 768, 1355\}$. Family centers are obtained by first averaging horizons equally within each dataset and then datasets equally within each family. The standard deviations in Table 1 are computed over seed-level family macros and cannot be reconstructed from the displayed cell means alone; the run-level values and aggregation script are available in the anonymous release.

Family	Dataset	Horizon	Moirai-MoE	MER	Rel. ↓
ETT	ETTh1	96	.287	**.259**	9.8%
ETT	ETTh1	192	.320	**.288**	10.0%
ETT	ETTh1	336	.357	**.322**	9.8%
ETT	ETTh1	720	.401	**.358**	10.7%
ETT	ETTh2	96	.276	**.248**	10.1%
ETT	ETTh2	192	.306	**.275**	10.1%
ETT	ETTh2	336	.342	**.307**	10.2%
ETT	ETTh2	720	.397	**.353**	11.1%
ETT	ETTm1	96	.305	**.273**	10.5%
ETT	ETTm1	192	.340	**.304**	10.6%
ETT	ETTm1	336	.376	**.336**	10.6%
ETT	ETTm1	720	.423	**.384**	9.2%
ETT	ETTm2	96	.281	**.256**	8.9%
ETT	ETTm2	192	.312	**.284**	9.0%
ETT	ETTm2	336	.346	**.315**	9.0%
ETT	ETTm2	720	.409	**.362**	11.5%
PEMS	PEMS03	12	.076	**.070**	7.9%
PEMS	PEMS03	24	.089	**.082**	7.9%
PEMS	PEMS03	48	.105	**.096**	8.6%
PEMS	PEMS03	96	.190	**.182**	4.2%
PEMS	PEMS04	12	.081	**.071**	12.3%
PEMS	PEMS04	24	.094	**.083**	11.7%
PEMS	PEMS04	48	.111	**.099**	10.8%
PEMS	PEMS04	96	.206	**.181**	12.1%
PEMS	PEMS07	12	.084	**.072**	14.3%
PEMS	PEMS07	24	.097	**.084**	13.4%
PEMS	PEMS07	48	.114	**.100**	12.3%
PEMS	PEMS07	96	.213	**.180**	15.5%
PEMS	PEMS08	12	.087	**.073**	16.1%
PEMS	PEMS08	24	.100	**.085**	15.0%
PEMS	PEMS08	48	.117	**.099**	15.4%
PEMS	PEMS08	96	.220	**.171**	22.3%
ILI+Weather	ILI	24	.235	**.213**	9.4%
ILI+Weather	ILI	36	.263	**.238**	9.5%
ILI+Weather	ILI	48	.290	**.263**	9.3%
ILI+Weather	ILI	60	.442	**.400**	9.5%
ILI+Weather	Weather	96	.251	**.227**	9.6%
ILI+Weather	Weather	192	.280	**.254**	9.3%
ILI+Weather	Weather	336	.310	**.281**	9.4%
ILI+Weather	Weather	720	.393	**.356**	9.4%

The displayed means reproduce the three Table 1 centers after rounding: $.342/.308$ for ETT, $.124/.108$ for PEMS, and $.308/.279$ for ILI/Weather. The family summaries in Table 1 and cell-wise percentages in Figure 2B are computed from these means. MER is lower in every displayed cell; the cells are correlated configurations, so this is descriptive coverage rather than paired or 40-trial significance evidence.