# MATCHED ENDPOINTS, DIFFERENT FUTURES: FORECASTING UNDER ORDER-DEPENDENT NON-STATIONARITY

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

## ABSTRACT

Forecasting models can adapt to distribution shift and encode long histories, but these capabilities do not isolate whether temporal order remains predictive once the observable endpoint and the experienced transformations are fixed. We formalize this phenomenon as *order-dependent predictive non-stationarity* (ODN): matched-endpoint histories containing the same transformation instances in different orders induce different conditional futures. Under squared loss and balanced paired orders, collapsing a reversed-order pair incurs minimum excess risk equal to one quarter of the per-target squared separation between its conditional means. We introduce ORBIT (Ordered Residual Bias for Interacting Transitions), a compact forecasting adapter that converts local predictive changes into bounded low-rank corrections and composes them in temporal order. We analyze when this interaction separates reversed paths and how approximation error propagates through the composition. On the primary context-held-out $AB \to BA$ benchmark, ORBIT reduces unseen-order NMSE from $.158$ for Mamba to $.124$, a $21.5\%$ reduction under matched training budgets and similar reported parameter scales. Removing or perturbing ordered composition increases paired error. An independent rotation diagnostic shows larger gain point estimates across increasing future-gap bins, and a no-auxiliary-loss variant retains a positive paired gain over Mamba. Across five semi-synthetic datasets, unseen-order error decreases by $12.25\%$. Together, these results support ordered composition as a useful inductive bias for forecasting under path-dependent shift.

## 1 INTRODUCTION

Modern forecasting models adapt to distribution shift and encode long histories. Yet these capabilities leave a basic question unresolved: if two histories end at the same observable endpoint and contain the same transformations, can their futures still depend on the order in which those transformations occurred?

Existing methods address changing levels, scales, periodicities, and dynamics through reversible normalization (Kim et al., 2022), explicit de-stationarization (Liu et al., 2022b), learned distribution alignment (Fan et al., 2023), local dynamics models (Liu et al., 2023), and flexible sequence backbones (Nie et al., 2023; Wang et al., 2024). Recurrent networks, attention, controlled differential equations, and signature representations also encode ordered paths (Hochreiter & Schmidhuber, 1997; Vaswani et al., 2017; Kidger et al., 2020; Morrill et al., 2021; Lou et al., 2024). These capabilities motivate a more specific test: whether the path carries predictive information beyond the endpoint and the collection of transformations experienced along it.

Consider two histories formed from the same transformation instances $A_\alpha$ and $B_\beta$. One history applies $A_\alpha$ followed by $B_\beta$; the other applies them in reverse. Suppose their predefined observable endpoints are identical, or matched within a declared tolerance. Different conditional futures then constitute *order-dependent predictive non-stationarity* (ODN). ODN is a controlled test of observable-endpoint sufficiency, not a new definition of statistical stationarity: it fixes the endpoint, transformation identities, and instance parameters while varying order, and asks whether the conditional forecasting law changes. Figure 1 illustrates this reversed-order construction.



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.

Exact-endpoint constructions isolate order at an identical observed endpoint; summary matching bounds the measured endpoint discrepancy. Commuting controls test the zero-interaction case. The endpoint is an observable representation, not a complete latent Markov state.

We first establish a method-independent consequence of this setting. Under mean squared error and balanced reversed-order cases, an order-blind representation pays excess risk $\frac{1}{4}\left\|\boldsymbol{\mu}_{AB} - \boldsymbol{\mu}_{BA}\right\|_{\mathrm{mse}}^2$, where the two matrices are the conditional future means. We then instantiate a minimal order-sensitive bias with ORBIT (Ordered Residual Bias for Interacting Transitions; Figure 1). ORBIT estimates local descriptors of predictive dynamics, converts descriptor changes into bounded low-rank residual corrections, and composes those corrections in temporal order. The ordered product retains interactions among the context-conditioned corrections while adapting the forecasting backbone.

We make three contributions. First, we formalize ODN as a controlled test of endpoint sufficiency and derive the exact excess MSE incurred by collapsing reversed-order histories under balanced paired orders. Second, we introduce ORBIT, combining change-only input, identity at zero, bounded low-rank corrections, and temporal composition, and characterize its order-separation mechanism and finite-path perturbation behavior. Third, we evaluate this bias through held-out-order comparisons, structural ablations, separate diagnostic benchmarks, and semi-synthetic transfer. The primary comparison reduces unseen-order NMSE from .158 for Mamba to .124. Independent diagnostics relate analytic future gaps to gain point estimates, while paired tests with two backbones and descriptive four-horizon results on three standard datasets broaden the empirical scope.

## 2 Related Work

### 2.1 Forecasting under distribution and dynamics shift

Work on concept drift motivates adaptation to changing predictive relations (Gama et al., 2014; Lu et al., 2019). In forecasting, normalization and de-stationarization adjust the observed series (Kim et al., 2022; Liu et al., 2022b), while distribution alignment and context-dependent dynamics adapt representations or transitions (Fan et al., 2023; Liu et al., 2023). TimeBridge distinguishes short- and long-range roles of non-stationarity (Liu et al., 2025). Flexible backbones capture temporal structure through recurrence and latent dynamics (Salinas et al., 2020; Rangapuram et al., 2018; Lai et al., 2018), attention and decomposition (Lim et al., 2021; Oreshkin et al., 2020; Zhou et al., 2021; Wu et al., 2021; Liu et al., 2022a), or patching and multiscale mixing (Nie et al., 2023; Liu et al., 2024; Wang et al., 2024; Ekambaram et al., 2023). Linear, frequency-domain, and temporal-variation models offer complementary predictive biases (Zeng et al., 2023; Zhou et al., 2022; Wu et al., 2023). These approaches address adaptation and representation; ODN instead isolates whether permuting fixed transition instances changes the conditional future at a matched observable endpoint.

### 2.2 History dependence, partial observation, and path representations

History can resolve predictive ambiguity under partial observation and delay (Schlaginhaufen et al., 2021), and recurrent and attention models already encode ordered observations (Hochreiter & Schmidhuber, 1997; Vaswani et al., 2017). Path-based models make this structure explicit through controlled differential equations, signatures, and rough differential equations (Kidger et al., 2020; Bonnier et al., 2019; Morrill et al., 2021). Path Development uses noncommutative multiplication (Lou et al., 2024); signature-patched Transformers connect path representations to continuous-time attention (Moreno-Pino et al., 2024). Our distinction is the evaluation question and adapter restriction: ODN controls the endpoint and transformation multiset, while ORBIT composes bounded corrections driven only by local changes. The controlled comparisons evaluate this configuration alongside order-sensitive alternatives.

### 2.3 Input-dependent operators and systematic generalization

Selective state-space models make transition parameters functions of the input (Gu & Dao, 2024), while fast-weight programmers dynamically construct rapidly changing weights (Irie et al., 2021). Structured state-space layers, neural ODEs, and related sequence operators provide dynamic parameterizations (Gu et al., 2022; Smith et al., 2023; Chen et al., 2018). ORBIT constrains input-dependent operations through change-only inputs, identity-at-zero updates, bounded low-rank corrections,

and ordered composition. Its Mamba comparison tests this joint inductive bias against general input-dependent sequence processing using complete model configurations.

Systematic generalization tests transfer to unseen compositions (Lake & Baroni, 2018; Keysers et al., 2020); ordinal templates reuse order structure across sequences (Zou et al., 2023). ODN adds endpoint and transformation-multiset controls to this question. The contribution lies in the joint adapter restriction and evaluation protocol, rather than in low rank, residual updates, fast weights, or multiplication individually. Ablations and coverage sweeps test that combined design and its limits.

## 3 Problem Formulation

Let $\mathbf{X}_{1:T} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ be a multivariate history and $\mathbf{Y}_{T+1:T+F} \in \mathbb{R}^{F\times p}$ its forecast target. To keep theory and reported mean squared error in the same units, define

$$\langle \mathbf{U}, \mathbf{V}\rangle_{\mathrm{mse}} = \frac{1}{Fp}\,\mathrm{tr}(\mathbf{U}^\top \mathbf{V}), \qquad \|\mathbf{U}\|^2_{\mathrm{mse}} = \frac{1}{Fp}\|\mathbf{U}\|_F^2. \tag{1}$$

Thus every squared norm used for a forecast matrix is per-target MSE, rather than a matrix spectral norm.

We write $H \in \mathcal{H}$ for the complete observed history and $e(H) \in \mathcal{E}$ for a predefined observable endpoint representation. The history is generated by a possibly latent predictive mechanism $\mathbf{\Theta}_t \in \mathcal{D}$. A transformation instance is a map on this mechanism, such as $A_\alpha : \mathcal{D} \to \mathcal{D}$; the formulation does not require $\mathbf{\Theta}_t$ or the maps to be observed by the forecaster. A structural key $S = s$ fixes every evaluator-controlled attribute other than order and history noise: the initial condition, transformation families and instance parameters, application times, endpoint target, horizon, and target variables. A coupling variable $U = u$ fixes the common random numbers used to materialize a paired history, but never a future innovation. We write $Q = (S, U)$ for the complete evaluator-only pairing context. It induces two histories $H_{AB}(q)$ and $H_{BA}(q)$ containing the same instances $A_\alpha, B_\beta \in \mathcal{T}$ in reversed order. The pairing key, mechanism, transformation labels, and order label are evaluator-only; a forecasting model receives $\mathbf{X}_{1:T}$.

The pair is *endpoint matched* under either the exact or approximate rule

$$e(H_{AB}(q)) = e(H_{BA}(q)), \qquad d_{\mathcal{E}}(e(H_{AB}(q)), e(H_{BA}(q))) \le \epsilon,$$

respectively, using a metric and tolerance fixed before evaluation. The endpoint is an observed summary, not a claim about a complete latent Markov state; the terminal latent predictive mechanism need not match. Let $\mathbf{Y}(o; q)$ be the potential forecast target under an evaluator intervention that changes only the order $o$ while fixing $q$. Expectation below integrates future innovations after that intervention. Define

$$\boldsymbol{\mu}_o(q) = \mathbb{E}[\mathbf{Y}(o; q)], \qquad D_{\mathrm{future}}(q) = \|\boldsymbol{\mu}_{AB}(q) - \boldsymbol{\mu}_{BA}(q)\|_{\mathrm{mse}}. \tag{2}$$

**Definition 1** (Order-dependent predictive non-stationarity)**.** A matched pair exhibits ODN at separation $\gamma > 0$ if it uses the same transformation instances in different temporal order and $D_{\mathrm{future}}(q) \ge \gamma$.

For a representation $\phi$, an *order-blind collapse* at $q$ means $\phi(H_{AB}(q)) = \phi(H_{BA}(q))$; any deterministic head on $\phi$ must then issue the same prediction for both members of the pair.

For a multiset of $K$ transformation instances, let $\mathfrak{S}_K$ be its distinct temporal permutations and $\mathfrak{S}_{\mathrm{train}} \subseteq \mathfrak{S}_K$ those allowed in training. Here $K$ counts generator-level instances, not observed samples or model update steps. Permutation coverage is

$$\kappa_{\mathrm{perm}} = \frac{|\mathfrak{S}_{\mathrm{train}}|}{|\mathfrak{S}_K|}. \tag{3}$$

For $0 < \kappa_{\mathrm{perm}} < 1$, the permutation generalization gap is

$$\mathrm{PGG} = \mathcal{L}_{\mathrm{unseen}} - \mathcal{L}_{\mathrm{seen}}, \tag{4}$$

where both losses use new contexts from the same held-out distribution. At $\kappa_{\mathrm{perm}} = 0$, no composed permutation is seen, while at $\kappa_{\mathrm{perm}} = 1$ no permutation remains unseen; PGG is undefined at both endpoints. Let $\mathcal{K}_{<1}$ be the tested coverage levels below one. We define $\kappa^\star_{5\%} = \min\{\kappa \in$

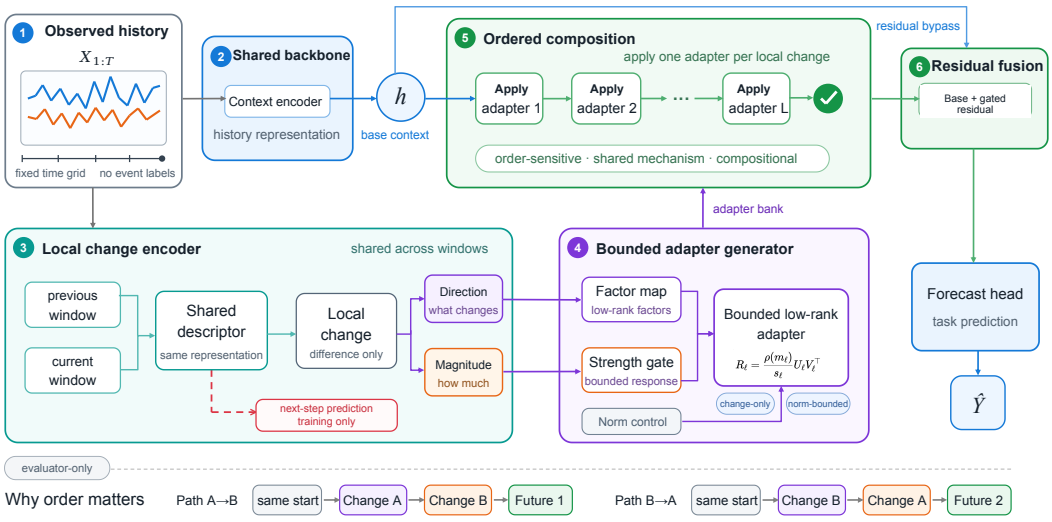



Figure 1: ORBIT architecture. A shared backbone encodes history; shared local descriptor changes generate bounded low-rank adapters, applied in temporal order and fused residually before forecasting. The evaluator-only reversed-order comparison uses histories with matched predefined observable endpoints.

$\mathcal{K}_{<1} : \mathcal{L}_{\text{unseen}}(\kappa) \le 1.05\mathcal{L}_{\text{all}}(1)\}$, where $\mathcal{L}_{\text{all}}(1)$ is the 100% all-order test NMSE. If no tested level qualifies, we report $> 80\%$; the 100% value is not called unseen-order error.

Context keys are partitioned before either order is materialized, so no $q$ crosses data splits. In the primary $AB \to BA$ protocol, training contains $AB$ only for $q \in \mathcal{Q}_{\text{train}}$; test contains paired $AB$ and $BA$ histories for new $q \in \mathcal{Q}_{\text{test}}$. The test pair reuses identical instances internally. The protocol tests unseen order on held-out contexts; context-disjoint splits do not require the transformation instance pairs themselves to be disjoint across splits.

This interventional notation applies to the controlled generator; no causal estimand is asserted for the natural observational audits.

## 4 Ordered Residual Bias for Interacting Transitions

ORBIT is a forecasting adapter that makes the effect of local predictive changes depend on their order of occurrence. Let a backbone encode the observed history as $\mathbf{h} = E_\theta(\mathbf{X}_{1:T}) \in \mathbb{R}^d$ and let $G_\omega$ map an adapted representation to an $F$-step forecast. The adapter operates on a fixed descriptor grid $\tau_0 < \tau_1 < \cdots < \tau_L = T$; thus it does not require transition labels or a learned event detector at inference time. The $L$ grid corrections need not correspond one-to-one to the $K$ generator-level transformation instances. The grid, rank, and norm bound are selected hyperparameters; all neural components and the residual gate are learned jointly from observed histories. Figure 1 summarizes the complete data flow from the shared backbone and local change encoder to ordered adaptation and forecasting.

### 4.1 Local Predictive Changes

For a window of width $w$ ending at $\tau_\ell$, the descriptor network gives

$$\mathbf{d}_\ell = D_\psi(\mathbf{X}_{\tau_\ell - w + 1:\tau_\ell}) \in \mathbb{R}^q, \qquad \Delta\mathbf{d}_\ell = \mathbf{d}_\ell - \mathbf{d}_{\ell-1}, \quad \bar{\mathbf{d}}_\ell = \frac{\Delta\mathbf{d}_\ell}{\|\Delta\mathbf{d}_\ell\|_2 + \varepsilon_d}. \tag{5}$$

Here $\varepsilon_d > 0$ is fixed for numerical stability. The inputs are normalized using statistics fitted on the training split. The descriptor is learned for forecasting utility. The auxiliary objective predicts the next $s$ observations from $\mathbf{d}_\ell$, using only targets inside a training history; its weight $\lambda_{\text{aux}} \ge 0$ is a hyperparameter, with zero disabling the term. The scalar $m_\ell = \|\Delta\mathbf{d}_\ell\|_2$ measures the magnitude

of an estimated local change, while its direction is retained by $\bar{\mathbf{d}}_\ell$. This separates the factor head's directional input from the explicit magnitude control below.

### 4.2 Bounded low-rank corrections

A shared correction network $C_\chi$ maps $\bar{\mathbf{d}}_\ell$ to factors $\mathbf{U}_\ell, \mathbf{V}_\ell \in \mathbb{R}^{d\times r}$. We use

$$\mathbf{R}_\ell = \rho(m_\ell)\,\frac{\mathbf{U}_\ell \mathbf{V}_\ell^\top}{s_\ell}, \quad s_\ell = \max(1, \|\mathbf{U}_\ell\|_F \|\mathbf{V}_\ell\|_F), \quad \rho(m) = \rho_{\max}\frac{m}{m+c}, \tag{6}$$

where $c > 0$, $0 < \rho_{\max} < 1$, and $r \ll d$. Because $\|\mathbf{U}\mathbf{V}^\top\|_2 \le \|\mathbf{U}\|_F \|\mathbf{V}\|_F$, the normalization is computed directly from the factors and guarantees $\|\mathbf{R}_\ell\|_2 \le \rho_{\max}$. The gate also imposes the identity limit $\rho(0) = 0$: if no descriptor change is estimated, ORBIT applies no correction. Both the gate and normalization are needed. The former couples correction strength to the estimated change; the latter prevents an arbitrarily large factorization from bypassing that coupling.

The factor head receives neither the absolute descriptor $\mathbf{d}_\ell$, the backbone state $\mathbf{h}$, the grid index, nor a transformation or order label. This *change-only* restriction shares a map across normalized predictive changes while leaving absolute context to the backbone. Since $\Delta\mathbf{d}_\ell$ can depend on the surrounding history, $\mathbf{R}_\ell$ is a context-conditioned local correction, not an identified physical transformation.

### 4.3 Path-ordered adaptation

As shown by the ordered-composition stage in Figure 1, corrections act on the backbone representation in temporal order,

$$\mathbf{P}_{1:L} = \prod_{\leftarrow \ell=1}^{L} (\mathbf{I} + \mathbf{R}_\ell) = (\mathbf{I} + \mathbf{R}_L)\cdots(\mathbf{I} + \mathbf{R}_1). \tag{7}$$

To isolate the product's order sensitivity, consider two fixed reference corrections $\mathbf{R}_A, \mathbf{R}_B$. Reversing these two factors produces

$$\mathbf{P}_{AB} - \mathbf{P}_{BA} = \mathbf{R}_B\mathbf{R}_A - \mathbf{R}_A\mathbf{R}_B. \tag{8}$$

This identity exposes the intended inductive bias: first-order effects are shared across the two orders, while their interaction preserves order information. A commutative sum of the same two corrections discards this term, although a general recurrent or attention aggregator need not be order blind. Here $A, B$ index an idealized two-factor comparison; reversing physical events need not simply permute the grid corrections. Section 5 treats the resulting alignment and context dependence explicitly.

The adapted representation and forecast are

$$\widetilde{\mathbf{h}} = \mathbf{h} + \eta(\mathbf{P}_{1:L} - \mathbf{I})\mathbf{h}, \qquad \widehat{\mathbf{Y}} = G_\omega(\widetilde{\mathbf{h}}), \tag{9}$$

where the gate is parameterized as $\eta = \mathrm{sigmoid}(a_\eta)$ for a trainable scalar $a_\eta$. Equivalently, $\widetilde{\mathbf{h}} = (1-\eta)\mathbf{h} + \eta\mathbf{P}_{1:L}\mathbf{h}$. The limits $\eta \to 0$ and $\eta \to 1$ recover the backbone and complete ordered transport, respectively; the no-correction ablation clamps $\eta = 0$. Controlled comparisons use common encoder/head architectures and descriptor grids at similar rounded parameter scales, but compare full configurations, not aggregation alone (Appendices C, E).

**Inductive-bias contract.** Change-only input prevents direct copying of absolute state or grid position; the magnitude gate makes zero change the identity; normalization bounds every update; and low rank limits capacity and cost. Multiplication retains successive interactions. Recurrent, attention, and path models can encode order, whereas an additive adapter discards the explicit interaction in Eq. (8). ORBIT therefore provides a targeted inductive bias for held-out permutation generalization, complementing general sequence architectures that can also represent temporal order. The shared-encoder comparison tests the complete contract; the additive, gate, normalization, and rank ablations provide supporting sensitivity checks.

### 4.4 Learning, computation, and stability

The product is never materialized. Starting at $\mathbf{z}_0 = \mathbf{h}$, ORBIT performs the recurrence

$$\mathbf{z}_\ell = \mathbf{z}_{\ell-1} + \frac{\rho(m_\ell)}{s_\ell}\mathbf{U}_\ell(\mathbf{V}_\ell^\top \mathbf{z}_{\ell-1}), \qquad \ell = 1, \ldots, L, \tag{10}$$

and substitutes $\mathbf{z}_L$ for $\mathbf{P}_{1:L}\mathbf{h}$ in equation 9. All components are trained end to end with

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{forecast}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}}, \tag{11}$$

where the main loss is MSE and $\lambda_{\text{aux}} = 0$ when the descriptor auxiliary task is disabled. No transformation identity, order label, composition loss, or commuting label is supplied to ORBIT.

Appendix C gives the descriptor and factor-head architectures and the search specification.

Each transport step costs $O(dr)$, so an $L$-step path costs $O(Ldr)$ beyond descriptor and factor generation. Streaming inference stores the hidden vector and current factors, requiring $O(d + dr)$ working memory; backpropagation through all factors requires $O(Ldr)$ activation memory before checkpointing. With linear factor heads, $C_\chi$ contributes $2dr(q + 1)$ parameters; the descriptor and forecast heads contribute to the total counts in the controlled comparison. Finite-path norm and perturbation controls appear in Section 5; they are worst-case bounds rather than contraction claims.

## 5 Prediction Cost and Composition Error

The risk analysis quantifies the cost of discarding order, while the composition analysis characterizes approximation error in ORBIT's ordered parameterization. Architectural comparisons are evaluated empirically in Section 6. For fixed paired context $Q = q$, define

$$\begin{aligned}\mathcal{R}_{\text{blind}}(q) &= \inf_{\mathbf{a}} \frac{1}{2} \sum_{o\in\{AB,BA\}} \mathbb{E}[\|\mathbf{Y}(o;q) - \mathbf{a}\|^2_{\text{mse}}],\\ \mathcal{R}_{\text{order}}(q) &= \frac{1}{2} \sum_{o\in\{AB,BA\}} \inf_{\mathbf{a}_o} \mathbb{E}[\|\mathbf{Y}(o;q) - \mathbf{a}_o\|^2_{\text{mse}}].\end{aligned} \tag{12}$$

If $\phi(H_{AB}(q)) = \phi(H_{BA}(q))$, every deterministic head on $\phi$ must issue one prediction for the pair and is subject to the blind restriction.

### 5.1 Exact and approximate order collapse

**Theorem 1** (Exact order-blind excess risk)**.** *Under balanced orders at fixed $Q = q$,*

$$\mathcal{R}_{\text{blind}}(q) - \mathcal{R}_{\text{order}}(q) = \frac{1}{4}\,\|\boldsymbol{\mu}_{AB}(q) - \boldsymbol{\mu}_{BA}(q)\|^2_{\text{mse}}\,. \tag{13}$$

*Consequently, the same lower bound applies to any predictor whose representation collapses the matched pair.*

This is per-target excess MSE, not total risk: within-order observation noise appears in both terms and cancels. For a distribution of matched contexts, the population penalty is the expectation of equation 13 over $Q$. With order probability $p_{AB}(q)$, the factor $1/4$ becomes $p_{AB}(q)(1 - p_{AB}(q))$.

Natural pairs are usually matched only in a declared endpoint metric. The next result limits how much predictive separation such mismatch alone can explain.

**Theorem 2** (Approximate endpoint matching)**.** *Fix $Q = q$ with $d_{\mathcal{E}}(e(H_{AB}), e(H_{BA})) \le \epsilon$. If an endpoint-only predictor $g : \mathcal{E} \to \mathbb{R}^{F\times p}$ satisfies $\|g(e) - g(e')\|_{\text{mse}} \le L_g d_{\mathcal{E}}(e, e')$, its average squared bias relative to the two order-aware means is at least*

$$\frac{1}{4}\left(\|\boldsymbol{\mu}_{AB}(q) - \boldsymbol{\mu}_{BA}(q)\|_{\text{mse}} - L_g\epsilon\right)^2_+\,. \tag{14}$$

The result applies to endpoint-only prediction, not to an arbitrary history encoder. It motivates reporting endpoint and future gaps separately and does not treat approximate matching as exact equality.

### 5.2 Order separation and perturbation control

For a common base vector $\mathbf{h}$ and linear structural readout $\mathbf{W}$, context-invariant reference corrections give

$$\widehat{\mathbf{y}}^{\text{str}}_{AB} - \widehat{\mathbf{y}}^{\text{str}}_{BA} = \mathbf{W}(\mathbf{R}_B\mathbf{R}_A - \mathbf{R}_A\mathbf{R}_B)\mathbf{h}. \tag{15}$$

This identity isolates the transport branch with $\eta = 1$, a common base vector, and a linear readout. The full predictor also depends on its history-specific base state and gate. Commuting reference corrections have identical pairwise structural contributions, whereas a commutative sum of the same corrections cannot retain this interaction. General recurrent or attention aggregators are not covered by that statement.

For an idealized pair with aligned correction identities $A, B$, write $\mathbf{R}_A^{AB}, \mathbf{R}_B^{AB}$ for the two corrections on an $AB$ path and $\mathbf{R}_A^{BA}, \mathbf{R}_B^{BA}$ for those on a $BA$ path. Let $\mathbf{R}_A^\star, \mathbf{R}_B^\star$ be context-independent references and $\varepsilon_i^o = \|\mathbf{R}_i^o - \mathbf{R}_i^\star\|_2$.

**Proposition 1** (Context-aware order decomposition)**.** *If every actual and reference correction has spectral norm at most $\rho_{\max}$, then*

$$\mathbf{P}_{AB} - \mathbf{P}_{BA} = \mathbf{R}_B^\star \mathbf{R}_A^\star - \mathbf{R}_A^\star \mathbf{R}_B^\star + \mathbf{E}_{\mathrm{ctx}}, \tag{16}$$

$$\|\mathbf{E}_{\mathrm{ctx}}\|_2 \le (1 + \rho_{\max}) \sum_{o \in \{AB, BA\}} (\varepsilon_A^o + \varepsilon_B^o). \tag{17}$$

Let $\mathbf{C}^\star = \mathbf{R}_B^\star \mathbf{R}_A^\star - \mathbf{R}_A^\star \mathbf{R}_B^\star$ and define $\|\mathbf{W}\|_{2\to\mathrm{mse}} = \sup_{\mathbf{v}\neq\mathbf{0}} \|\mathbf{W}\mathbf{v}\|_{\mathrm{mse}} / \|\mathbf{v}\|_2$.

**Corollary 1** (Separation under context drift)**.** *Under Proposition 1, a common base vector* $\mathbf{h}$*, and the linear structural readout* $\mathbf{W}$*, the two structural forecasts remain distinct whenever*

$$\|\mathbf{W}\mathbf{C}^\star\mathbf{h}\|_{\mathrm{mse}} > \|\mathbf{W}\|_{2\to\mathrm{mse}} \|\mathbf{h}\|_2 (1 + \rho_{\max}) \sum_{o \in \{AB, BA\}} (\varepsilon_A^o + \varepsilon_B^o). \tag{18}$$

The reverse triangle inequality lower-bounds the output separation by the left side of equation 18 minus its right side. This is a sufficient separation condition for the aligned structural branch, not a claim that training attains the required margin or that separation alone lowers risk.

The decomposition separates a reference interaction from the effect of context-dependent corrections. Its two-factor interpretation requires an alignment; the general grid-level bound below does not require primitive identities.

Consider two positionally aligned sequences of $N \ge 1$ bounded corrections, $\mathbf{R}_j^c$ and $\mathbf{R}_j^{c,\star}$, for a history context $c$. Let $\mathbf{P}^c$ and $\mathbf{P}^{c,\star}$ be their temporal products. A telescoping argument gives

$$\|\mathbf{P}^c - \mathbf{P}^{c,\star}\|_2 \le (1 + \rho_{\max})^{N-1} \sum_{j=1}^{N} \|\mathbf{R}_j^c - \mathbf{R}_j^{c,\star}\|_2. \tag{19}$$

For ORBIT's descriptor grid, $N = L$: references are compared at the same grid positions, without identifying either sequence with the $K$ physical changes. A primitive-permutation interpretation additionally requires one bounded correction per instance in a common representation space, so $N = K$; the descriptor grid alone supplies no such correspondence.

With a Lipschitz forecast head, the product error, base-state error, and reference approximation error yield the prediction bound and uniform corollary in Appendix A. These are deterministic sufficient conditions; they do not establish that training makes the errors small. The factor $(1 + \rho_{\max})^{N-1}$ also permits growth with path length, rather than guaranteeing contraction.

**Theory–evidence map.** Theorem 1 quantifies information loss but does not rank encoders; Theorem 2 motivates the endpoint audit; Eq. (8) motivates the ordered ablations; and Eq. (19) concerns an aligned reference. Together, these results characterize the ordered parameterization under explicit alignment and approximation assumptions; they neither imply convergence nor identify learned grid corrections with physical primitives. Held-out-order comparisons and ablations provide the empirical test of this inductive bias.

# 6 Experiments

## 6.1 Experimental setup

**Data and splits.** Primary-benchmark context keys are split 12,000/2,000/2,000 before generation. Training and validation contain $AB$ only on disjoint keys; paired $AB/BA$ test histories use new

keys. Thus $BA$ enters neither optimization nor model selection. Section 3 and Appendix C define its finite pairing, split, and selection rules; Appendix B diagnostics apply only to named regimes. Coverage uses 28,800/5,760/28,800 histories with admitted validation orders; semi-synthetic studies use 10,000/2,000/2,000 split-local base windows and AB-only validation. Ordinary forecasting uses $T = 96$, four horizons, chronological splits, and training-only normalization.

**Comparisons, metrics, and uncertainty.** The primary benchmark fixes TCN/head architectures, descriptor grid, 12,000 training examples, 5,000 updates, 30 trials, and a rounded 1.18–1.23M scale. Seed 2027 selects by AB-only validation NMSE (ORBIT ties favor smaller $r$); the update-5,000 checkpoint is frozen, rerun for seeds 2027–2031, and tested once per seed. NMSE/NMAE use validation-target variance/standard deviation; $\mathrm{PGG} = \mathrm{NMSE}_{\mathrm{unseen}} - \mathrm{NMSE}_{\mathrm{seen}}$. Unless stated otherwise, controlled and semi-synthetic 95% confidence intervals use 2,000 paired hierarchical seed–context bootstrap replicates with percentile endpoints; full resampling details appear in Appendix C.2.

Table 1: Fixed paired context-held-out $AB \to BA$ comparison under common budgets and similar rounded parameter scales. The paired best-comparator Mamba–ORBIT difference is .034 [.012,.056].

| Aggregator | Params | Seen NMSE | Unseen NMSE [95% CI] | Unseen NMAE | PGG |
|---|---|---|---|---|---|
| GRU | 1.21M | .104 | .197 [.177,.215] | .277 | .094 |
| Transformer | 1.23M | .095 | .185 [.165,.207] | .267 | .091 |
| Mamba | 1.22M | .097 | .158 [.139,.177] | .261 | .061 |
| Signature | 1.18M | .108 | .207 [.185,.227] | .289 | .098 |
| Path Development | 1.22M | .112 | .215 [.197,.235] | .297 | .104 |
| **ORBIT** | 1.21M | .100 | **.124 [.104,.144]** | **.242** | **.024** |

*Core paired effects (variant minus full ORBIT)*

| Comparison | ΔNMSE | Paired 95% CI |
|---|---|---|
| Additive only − ordered product | .051 | [.028,.074] |
| Shuffled residual order − full ORBIT | .111 | [.088,.134] |
| No correction − full ORBIT | .093 | [.072,.114] |

### 6.2 Held-out-order comparison and structural evidence

**Primary benchmark.** ORBIT achieves .124 unseen-order NMSE compared with .158 for Mamba, a 21.5% reduction, while reducing PGG from .061 to .024 (Table 1). It generalizes more effectively to the held-out $BA$ order under this shared evaluation protocol; separate controlled diagnostics examine matched-endpoint effects. Against Mamba, ORBIT uses 6.1% more FLOPs, 4.4% higher latency, and 9.1% more memory; versus Transformer, the respective increases are 15.6%, 13.7%, and 14.3% (Appendix Table 8).

**Restriction tests.** Each structural perturbation increases paired error: additive composition raises NMSE by .051, shuffled residual order by .111, and removing the correction by .093 (Table 1). Additional variants also perform worse: absolute+change reaches .142 [.117,.167], no factor-norm normalization .157 [.130,.187], and no magnitude gate .146 [.120,.173], versus full ORBIT's .124 [.104,.144] (marginal model CIs; Appendix Table 16). These ablations support the full design rather than attributing its gain to a single component.

Table 2: Descriptive four-horizon mean NMSE. The macro .3032 is 5.3% below .3203; these point estimates provide complementary evidence beyond the controlled benchmarks.

| Dataset | DLinear | TimesNet | FEDformer | PatchTST | iTransformer | TimeMixer | **PatchTST+ORBIT** |
|---|---|---|---|---|---|---|---|
| ETTh1 | .467 | .435 | .452 | .4090 | .3984 | .3973 | **.3825** |
| Weather | .267 | .247 | .255 | .21925 | .2090 | .2143 | **.1936** |
| ETTm1 | .385 | .357 | .370 | .3634 | .3536 | .3500 | **.3336** |
| Macro | .373 | .346 | .359 | .3306 | .3203 | .3205 | **.3032** |

**Independent diagnostic benchmark.** On the rotation benchmark (10k $AB$ training contexts/seed), ORBIT reaches .126 $BA$ NMSE versus Mamba's .161 (paired gain .035 [.013,.057]). Gain point estimates increase across validation-fixed future-gap bins, without a between-bin significance claim (Figure 2a). No-aux ORBIT retains a .028 [.009,.047] gain, so auxiliary supervision alone does not explain the advantage. Fusion paths differ: these are complete-configuration comparisons (Table 20c).

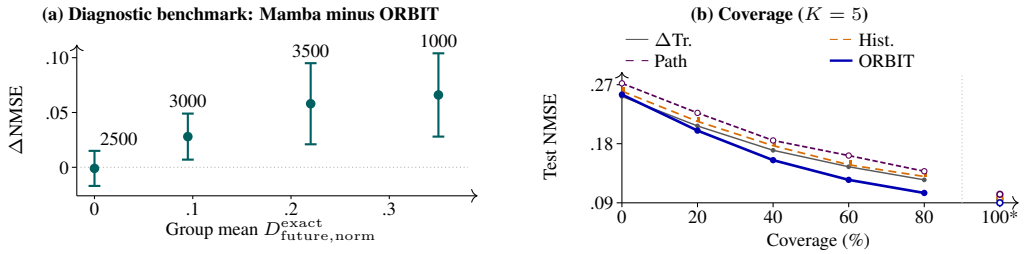
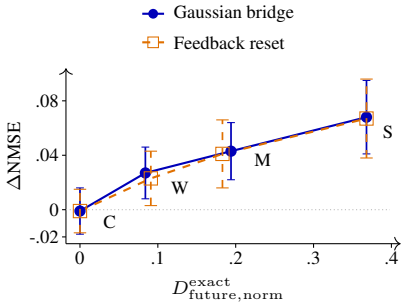



Figure 2: (a) $BA$ Mamba–ORBIT gain by validation-fixed $D$ bin (paired 95% CIs; pooled counts; no between-bin test). (b) $K = 5$ coverage; open points are the all-order reference. ORBIT is not best at 0%; AUPC gain versus Delta Transformer is .013 [.003,.023].

### 6.3 Rotation-angle regimes and transfer

Table 3: Mechanism response, transfer, and portability; positive ΔNMSE favors ORBIT.




**(b) Noncommuting semi-synthetic transfer**

| Data | Best | Comp. | ORBIT | Δ [95% CI] | Rel. |
|---|---|---|---|---|---|
| ETTh1 | P | .460 | .407 | .053 [.021,.091] | 11.5% |
| ETTm1 | iT | .432 | .374 | .058 [.020,.090] | 13.4% |
| Weather | TM | .330 | .280 | .050 [.022,.078] | 15.2% |
| Electricity | P | .260 | .225 | .035 [.014,.060] | 13.5% |
| Traffic | iT | .502 | .455 | .047 [.008,.080] | 9.4% |
| **Macro** | D-best | **.397** | **.348** | **.049 [.021,.077]** | **12.25%** |

**(c) Cross-backbone controlled portability**

| Backbone | None | +Hist. | +Delta | +ORBIT | Δ [95% CI] |
|---|---|---|---|---|---|
| PatchTST | .415 | .407 | .397 | .377 | .038 [.015,.055] |
| TimeMixer | .427 | .418 | .403 | .385 | .042 [.018,.058] |

(a) Transformer minus ORBIT, paired 95% CIs; C/W/M/S denote commuting/weak/medium/strong regimes. (b) P/iT/TM: PatchTST/iTransformer/TimeMixer; D-best is dataset-wise best. (c) Full metrics: Table 15.

**Regimes and transfer.** Across both analytic constructions, ORBIT's gain increases from weak to strong noncommuting regimes; the commuting controls have zero future gap and a $-.001$ gain with CIs spanning zero. This pattern is consistent with the intended order-sensitive mechanism. All five semi-synthetic datasets favor ORBIT, with relative reductions of 9.4–15.2% and a 12.25% macro reduction (.049 [.021,.077]). The commuting macro is $-.002$ [$-.011$,.006], and all commuting ΔPGG intervals include zero. Selected pairs satisfy $\epsilon_{\text{match}} = .15$ (maximum .1498). These studies test rotation-based transfer against the reported comparators; cross-backbone portability is evaluated in the controlled setting.

## 7 Limitations

Our study centers on controlled and semi-synthetic interventions, with analytic constructions enabling matched-endpoint tests of ODN. Extending these diagnostics beyond rotational interventions to naturally occurring shifts is an important direction for future work. The theory characterizes the specified setting and ordered parameterization, while descriptive ordinary-forecasting results complement the controlled studies with performance comparisons on standard benchmarks.

## 8 Conclusion

We introduced ODN to test whether temporal order remains predictive at matched observable endpoints for the same transformation instances. We derived the exact excess risk of collapsing reversed-order pairs and proposed ORBIT, a bounded low-rank adapter that composes local predictive changes in temporal order. ORBIT improves held-out-order prediction on the primary benchmark and in all five noncommuting semi-synthetic comparisons, with structural ablations and rotation diagnostics supporting ordered composition. These findings motivate path-aware adaptation under distribution shift: the route to an observable endpoint can matter for forecasting.

## AI USE STATEMENT

In this work, we used generative AI tools to improve readability of the manuscript and to identify and retrieve related literature. We have not used generative AI tools for generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, translation, cleaning or reformatting datasets, qualitative or thematic data analysis, or interpreting results, and these tasks are not applicable to this work. We have reviewed all AI-assisted work; experiments and numerical measurements were executed by the authors, and we take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

## REPRODUCIBILITY STATEMENT

The main text specifies the forecasting task, context-disjoint splits, evaluation metrics, comparison budgets, and uncertainty conventions. Section 3 defines pairing and held-out-order evaluation. Appendix A provides complete proofs; Appendix B specifies the analytic controlled and semi-synthetic reference constructions; Appendix C documents the primary benchmark's partition and pairing protocol, model architectures, hyperparameter search, training and selection procedures, seeds, leakage controls, and bootstrap resampling; Appendix D reports detailed experimental results; and Appendix E specifies the independent diagnostic benchmark and its comparison protocol. An anonymous code repository for ORBIT is available at `https://anonymous.4open.science/r/ORBIT-D6F8/`.

## REFERENCES

Patric Bonnier, Patrick Kidger, Imanol Perez Arribas, Cristopher Salvi, and Terry Lyons. Deep signature transforms. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Vijay Ekambaram, Arindam Jati, Nam Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. TSMixer: Lightweight MLP-mixer model for multivariate time series forecasting. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 459–469, 2023. doi: 10.1145/3580305.3599533.

Wei Fan, Pengyang Wang, Dongkun Wang, Dongjie Wang, Yuanchun Zhou, and Yanjie Fu. Dish-TS: A general paradigm for alleviating distribution shift in time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7522–7529, 2023. doi: 10.1609/aaai.v37i6.25914.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):44:1–44:37, 2014. doi: 10.1145/2523813.

Albert Gu and Tri Dao. Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024. URL `https://openreview.net/forum?id=tEYskw1VY2`.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=uYLFoz1vlAC`.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Kazuki Irie, Imanol Schlag, Róbert Csordás, and Jürgen Schmidhuber. Going beyond linear transformers with recurrent fast weight programmers. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SygcCnNKwr.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=cGDAkQo1C0p.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 95–104, 2018. doi: 10.1145/3209978.3210006.

Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2873–2882, 2018. URL https://proceedings.mlr.press/v80/lake18a.html.

Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4): 1748–1764, 2021. doi: 10.1016/j.ijforecast.2021.03.012.

Peiyuan Liu, Beiliang Wu, Yifan Hu, Naiqi Li, Tao Dai, Jigang Bao, and Shu-Tao Xia. TimeBridge: Non-stationarity matters for long-term time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39815–39840, 2025. URL https://proceedings.mlr.press/v267/liu25cb.html.

Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X. Liu, and Schahram Dustdar. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=0EXmFzUn5I.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9881–9893, 2022b. doi: 10.52202/068431-0718.

Yong Liu, Chenyu Li, Jianmin Wang, and Mingsheng Long. Koopa: Learning non-stationary time series dynamics with koopman predictors. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0538.

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. URL https://openreview.net/forum?id=JePfAI8fah.

Hang Lou, Siran Li, and Hao Ni. Path development network with finite-dimensional lie group. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=YoWBLu74TL.

Jie Lu, Anjin Liu, Fan Dong, Feng Gu, João Gama, and Guangquan Zhang. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12):2346–2363, 2019. doi: 10.1109/TKDE.2018.2876857.

Fernando Moreno-Pino, Álvaro Arroyo, Harrison Waldon, Xiaowen Dong, and Álvaro Cartea. Rough transformers: Lightweight and continuous time series modelling through signature patching. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3371.

James Morrill, Cristopher Salvi, Patrick Kidger, and James Foster. Neural rough differential equations for long time series. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7829–7838, 2021. URL https://proceedings.mlr.press/v139/morrill21b.html.

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. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1ecqn4YwB.

Syama Sundar Rangapuram, Matthias W. Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski. Deep state space models for time series forecasting. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3): 1181–1191, 2020. doi: 10.1016/j.ijforecast.2019.07.001.

Andreas Schlaginhaufen, Philippe Wenk, Andreas Krause, and Florian Dörfler. Learning stable deep dynamics models for partially observed or delayed dynamical systems. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Jimmy T. H. Smith, Andrew Warrington, and Scott W. Linderman. Simplified state space layers for sequence modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Ai8Hw3AXqks.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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

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, 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. URL https://openreview.net/forum?id=ju_Uqw384Oq.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317.

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. doi: 10.1609/aaai.v35i12.17325.

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. URL https://proceedings.mlr.press/v162/zhou22g.html.

Xiaolong Zou, Zhikun Chu, Qinghai Guo, Jie Cheng, Bo Ho, Si Wu, and Yuanyuan Mi. Learning and processing the ordinal information of temporal sequences in recurrent neural circuits. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1473.

# A PROOFS

## A.1 PROOF OF THEOREM 1

For a common prediction $\mathbf{a}$, conditional variance decomposition in the inner product of equation 1 gives

$$\frac{1}{2}\sum_{o}\mathbb{E}[\|\mathbf{Y}(o;q)-\mathbf{a}\|_{\mathrm{mse}}^2]=\frac{1}{2}\sum_{o}\mathbb{E}[\|\mathbf{Y}(o;q)-\boldsymbol{\mu}_o(q)\|_{\mathrm{mse}}^2]+\frac{1}{2}\sum_{o}\|\boldsymbol{\mu}_o(q)-\mathbf{a}\|_{\mathrm{mse}}^2\,. \tag{20}$$

The first line is $\mathcal{R}_{\mathrm{order}}(q)$. The second is minimized at the midpoint $\mathbf{a}=(\boldsymbol{\mu}_{AB}+\boldsymbol{\mu}_{BA})/2$ and equals one quarter of their squared separation. For order probability $p_{AB}(q)$, the weighted midpoint yields $p_{AB}(q)(1-p_{AB}(q))\,\|\boldsymbol{\mu}_{AB}-\boldsymbol{\mu}_{BA}\|_{\mathrm{mse}}^2$.

## A.2 PROOF OF THEOREM 2

Write $\widehat{\mathbf{y}}_o=g(e(H_o))$. Lipschitzness gives $\|\widehat{\mathbf{y}}_{AB}-\widehat{\mathbf{y}}_{BA}\|_{\mathrm{mse}}\le L_g\epsilon$. The triangle inequality then implies

$$\begin{aligned}&\|\boldsymbol{\mu}_{AB}-\widehat{\mathbf{y}}_{AB}\|_{\mathrm{mse}}+\|\boldsymbol{\mu}_{BA}-\widehat{\mathbf{y}}_{BA}\|_{\mathrm{mse}}\\&\quad\ge(\|\boldsymbol{\mu}_{AB}-\boldsymbol{\mu}_{BA}\|_{\mathrm{mse}}-L_g\epsilon)_+.\end{aligned} \tag{21}$$

Applying $(a^2+b^2)/2\ge(a+b)^2/4$ proves equation 14.

## A.3 PROOF OF PROPOSITION 1

Define the reference products $\mathbf{P}^\star_{AB}=(\mathbf{I}+\mathbf{R}^\star_B)(\mathbf{I}+\mathbf{R}^\star_A)$ and $\mathbf{P}^\star_{BA}=(\mathbf{I}+\mathbf{R}^\star_A)(\mathbf{I}+\mathbf{R}^\star_B)$. Direct expansion gives $\mathbf{P}^\star_{AB}-\mathbf{P}^\star_{BA}=\mathbf{R}^\star_B\mathbf{R}^\star_A-\mathbf{R}^\star_A\mathbf{R}^\star_B$. For either order, the two-factor telescoping identity and submultiplicativity give

$$\|\mathbf{P}_o-\mathbf{P}^\star_o\|_2\le(1+\rho_{\max})(\varepsilon^o_A+\varepsilon^o_B).$$

Set $\mathbf{E}_{\mathrm{ctx}}=(\mathbf{P}_{AB}-\mathbf{P}^\star_{AB})-(\mathbf{P}_{BA}-\mathbf{P}^\star_{BA})$ and apply the triangle inequality.

## A.4 PERTURBATION OF ALIGNED CORRECTION SEQUENCES

Fix a sequence length $N\ge1$ and a history context $c$, including its order when applicable. For each position $j$, compare a learned correction $\mathbf{R}^c_j$ with a reference $\mathbf{R}^{c,\star}_j$ in the same representation space. The references may depend on $c$; no cross-context primitive identity is needed. Define

$$\mathbf{P}^{c,\star}=\prod_{\leftarrow j=1}^{N}(\mathbf{I}+\mathbf{R}^{c,\star}_j),\qquad\mathbf{P}^{c}=\prod_{\leftarrow j=1}^{N}(\mathbf{I}+\mathbf{R}^{c}_j).$$

**Theorem 3** (Prediction error under aligned corrections)**.** *Assume, for all $j$, $\|\mathbf{R}^{c,\star}_j\|_2,\|\mathbf{R}^c_j\|_2\le\rho_{\max}$, $\|\mathbf{R}^c_j-\mathbf{R}^{c,\star}_j\|_2\le\varepsilon^c_j$, and $\|\mathbf{h}^c-\mathbf{h}^c_\star\|_2\le\varepsilon^c_h$. Fix $\eta\in[0,1]$. Let $G_\omega$ be $K_G$-Lipschitz from vector $\ell_2$ distance to $\|\cdot\|_{\mathrm{mse}}$, define $\widehat{\mathbf{Y}}^c=G_\omega((1-\eta)\mathbf{h}^c+\eta\mathbf{P}^c\mathbf{h}^c)$, and let the reference predictor's approximation error relative to the conditional target mean $\boldsymbol{\mu}(c)$ be*

$$\xi^c=\|G_\omega((1-\eta)\mathbf{h}^c_\star+\eta\mathbf{P}^{c,\star}\mathbf{h}^c_\star)-\boldsymbol{\mu}(c)\|_{\mathrm{mse}}\,.$$

*Then*

$$\left\|\widehat{\mathbf{Y}}^c-\boldsymbol{\mu}(c)\right\|_{\mathrm{mse}}\le\xi^c+K_G\big[\{1-\eta+\eta(1+\rho_{\max})^N\}\varepsilon^c_h+\eta(1+\rho_{\max})^{N-1}\|\mathbf{h}^c_\star\|_2\sum_{j=1}^{N}\varepsilon^c_j\big]. \tag{22}$$

**Proof.** For fixed $c$, let $\mathbf{A}_j = \mathbf{I} + \mathbf{R}_j^c$ and $\mathbf{A}_j^\star = \mathbf{I} + \mathbf{R}_j^{c,\star}$. A telescoping expansion gives

$$\mathbf{P}^c - \mathbf{P}^{c,\star} = \sum_{j=1}^{N} \mathbf{A}_N \cdots \mathbf{A}_{j+1}(\mathbf{R}_j^c - \mathbf{R}_j^{c,\star})\mathbf{A}_{j-1}^\star \cdots \mathbf{A}_1^\star.$$

Submultiplicativity and $\|\mathbf{A}_j\|_2, \|\mathbf{A}_j^\star\|_2 \le 1 + \rho_{\max}$ yield

$$\|\mathbf{P}^c - \mathbf{P}^{c,\star}\|_2 \le (1+\rho_{\max})^{N-1} \sum_{j=1}^{N} \varepsilon_j^c, \quad \|\mathbf{P}^c\|_2 \le (1+\rho_{\max})^N. \tag{23}$$

Subtract the reference adapted state from the learned state and add and subtract $\eta \mathbf{P}^c \mathbf{h}_\star^c$. Applying equation 23 gives the two bracketed terms in equation 22. Lipschitzness of $G_\omega$ and a final triangle inequality with the reference approximation error $\xi^c$ complete the proof.

**Corollary 2** (Uniform deterministic control)**.** *If, over a context set $\mathcal{C}_N$ of aligned length-$N$ sequence pairs, $\xi^c \le \bar{\xi}$, $\varepsilon_h^c \le \bar{\varepsilon}_h$, $\|\mathbf{h}_\star^c\|_2 \le H_\star$, and $\varepsilon_j^c \le \bar{\varepsilon}_R$ for all positions $j$, then*

$$\sup_{c \in \mathcal{C}_N} \left\| \widehat{\mathbf{Y}}^c - \boldsymbol{\mu}(c) \right\|_{\text{mse}} \le \bar{\xi} + K_G\big[\{1 - \eta + \eta(1+\rho_{\max})^N\}\bar{\varepsilon}_h + \eta(1+\rho_{\max})^{N-1} H_\star N \bar{\varepsilon}_R\big]. \tag{24}$$

Taking the supremum of Theorem 3 and using $\sum_{j=1}^N \varepsilon_j^c \le N\bar{\varepsilon}_R$ proves the result.

**Grid application and primitive-permutation specialization.** For the implemented adapter, set $N = L$ and compare learned and reference corrections position by position on the descriptor grid. References can vary with context, so this application imposes no primitive-reuse assumption. For a stronger primitive-level interpretation, assume instead an explicit one-to-one correspondence between the $K$ transformation instances and $K$ bounded corrections in the same representation space. Set $N = K$ and, for $\pi \in \mathfrak{S}_K$, identify $\mathbf{R}_j^c = \mathbf{R}_{\pi_j}^c$ and $\mathbf{R}_j^{c,\star} = \mathbf{R}_{\pi_j}^\star$, with references shared across contexts. The theorem then applies to each such permutation, and the uniform corollary applies if its error assumptions hold uniformly across them. This is an additional alignment assumption, not a consequence of the fixed grid: one event may affect several descriptor differences, and one difference may combine several events. In particular, grouping grid factors does not automatically preserve the per-correction norm bound. The results neither construct an alignment nor establish small reference errors after training.

# B Reference Construction and Analytic Properties

## B.1 Exact-endpoint reference generator

The following construction defines an analytic matched-endpoint reference. All quantities in this subsection are properties of the declared generator; they are not estimated from simulated future innovations.

The latent predictive parameter is $\boldsymbol{\theta} \in \mathbb{R}^{p^2}$ with $p = 8$, and $\mathbf{x}_t \in \mathbb{R}^p$. Here $\mathrm{mat} : \mathbb{R}^{p^2} \to \mathbb{R}^{p \times p}$ is the fixed row-major reshape, $[\mathrm{mat}(\boldsymbol{\theta})]_{ij} = \theta_{(i-1)p+j}$; it is not a reshape of an eight-dimensional vector. It defines a stable VAR(1) matrix

$$\mathbf{M}(\boldsymbol{\theta}) = 0.95 \, \frac{\mathrm{mat}(\boldsymbol{\theta})}{\max(1, \|\mathrm{mat}(\boldsymbol{\theta})\|_2)}, \qquad \mathbf{x}_{t+1} = \mathbf{M}(\boldsymbol{\theta}_t)\mathbf{x}_t + \boldsymbol{\varepsilon}_t,$$

where $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ and $\sigma \in \{0.01, 0.05, 0.1\}$, with $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. A primitive is a Givens rotation in parameter space: $A_\alpha(\boldsymbol{\theta}) = \mathbf{Q}_{ij}(\alpha)\boldsymbol{\theta}$. Commuting controls use two rotations in the same coordinate plane; interacting cases use planes $(1,2)$ and $(2,3)$ with $\alpha, \beta \in \{\pi/36, \pi/12, \pi/6\}$. In addition to the parameter-space discrepancy, the analytic predictive interaction is

$$C_M^\star(q) = \left\| \mathbf{M}(B_\beta(A_\alpha(\boldsymbol{\theta}_0))) - \mathbf{M}(A_\alpha(B_\beta(\boldsymbol{\theta}_0))) \right\|_F. \tag{25}$$

This quantity is defined after the stability map that produces the forecasting dynamics. Weak, medium, and strong denote prescribed rotation-angle regimes, rather than estimates of $C_M^\star$.

Histories have $T = 96$. A paired context samples two application times $\tau_1 \in \{24, \dots, 40\}$ and $\tau_2 \in \{56, \dots, 72\}$; $AB$ applies $A_\alpha$ at $\tau_1$ and $B_\beta$ at $\tau_2$, whereas $BA$ swaps them. The same times are used inside each pair. Splits separate complete context keys, not individual time pairs. Application times vary rather than tying every change to a fixed descriptor boundary.

**Conditional Gaussian bridge.** For order $o$, the time-varying linear system induces a joint Gaussian law for $\mathbf{Z}_o = (\mathbf{x}_1, \dots, \mathbf{x}_{T-1})$ and $\mathbf{x}_T$. Given a context target $\mathbf{x}_\star$, sample the history from the exact conditional law

$$\mathbf{Z}_o \mid \mathbf{x}_T = \mathbf{x}_\star \sim \mathcal{N}\big(\mathbf{m}_{Z,o} + \mathbf{\Sigma}_{ZT,o}\mathbf{\Sigma}_{TT,o}^{-1}(\mathbf{x}_\star - \mathbf{m}_{T,o}), \\ \mathbf{\Sigma}_{ZZ,o} - \mathbf{\Sigma}_{ZT,o}\mathbf{\Sigma}_{TT,o}^{-1}\mathbf{\Sigma}_{TZ,o}\big). \tag{26}$$

The moments are obtained by the standard linear-Gaussian forward recursion. A shared standard-normal vector is mapped through each order-specific conditional Cholesky factor, providing common-random-number coupling without violating either dynamics. Every point in both paths therefore follows the declared conditional dynamical model, and the exact endpoint is $e(H) = \mathbf{x}_T = \mathbf{x}_\star$. A reference measure of the conditioning effect is

$$B_o(q) = \frac{\|\mathbf{m}_{Z|T,o}(q) - \mathbf{m}_{Z,o}\|_2}{\sqrt{\mathrm{tr}(\mathbf{\Sigma}_{ZZ,o})} + 10^{-8}}, \tag{27}$$

where $\mathbf{m}_{Z|T,o}$ is the conditional mean in equation 26. This quantity characterizes the conditioning shift; it is not used to define ODN.

**Feedback-reset control.** As a construction check, an unconstrained trajectory is generated through $T - 1$, followed by the common terminal-control rule

$$\mathbf{u}_{T-1}^o = \mathbf{x}_\star - \mathbf{M}(\boldsymbol{\theta}_o)\mathbf{x}_{T-1}^o, \qquad \mathbf{x}_T^o = \mathbf{M}(\boldsymbol{\theta}_o)\mathbf{x}_{T-1}^o + \mathbf{u}_{T-1}^o = \mathbf{x}_\star.$$

The policy and target are identical across orders; only the realized feedback action differs. This provides a second analytic route to a common endpoint.

For both constructions, conditional future means are analytically available as $[\mathbf{M}(\boldsymbol{\theta}_o)^k \mathbf{x}_\star]_{k=1}^F$. The reference uses $F = 24$, with $F \in \{12, 24, 48\}$ as alternative horizons. Analytic future gaps use the per-target norm in equation 1, so $D_{\mathrm{future}}^2/4$ and the order-blind excess risk have identical units.

Reference targets and conditional means use the same coordinate-wise validation standard deviations $s_j$ as the forecasting metrics. Writing $\widetilde{\mu}_{o,kj} = \mu_{o,kj}/s_j$, the normalized separation is

$$D_{\mathrm{future,norm}}^2(q) = \frac{1}{Fp}\sum_{k=1}^{F}\sum_{j=1}^{p}(\widetilde{\mu}_{AB,kj} - \widetilde{\mu}_{BA,kj})^2, \qquad \Delta\mathrm{NMSE}_{\mathrm{blind}} = \tfrac{1}{4}D_{\mathrm{future,norm}}^2(q).$$

Thus the theoretical per-target MSE identity and the experimental NMSE differ only by the declared validation-scale normalization. This common scaling also preserves the commuting zero-gap conclusion below.

**Analytic zero and nonzero gaps.** Same-plane rotations satisfy $\mathbf{Q}_{ij}(\beta)\mathbf{Q}_{ij}(\alpha) = \mathbf{Q}_{ij}(\alpha)\mathbf{Q}_{ij}(\beta)$, hence the final parameters and their stability-mapped matrices coincide. At the common endpoint,

$$\mathbf{M}_{AB} = \mathbf{M}_{BA} \quad \Longrightarrow \quad C_M^\star(q) = 0, \qquad D_{\mathrm{future}}(q) = 0. \tag{28}$$

These are exact statements for each reference context, not small positive noise floors. Conversely, noncommuting rotations admit a parameter vector on which their products differ. Scale it so both terminal vectors have $\ell_2$ norm below one; the stability map then scales both reshaped matrices by $0.95$. Choose an endpoint outside the kernel of their matrix difference. The first forecast-step means differ, yielding ODN for a positive separation threshold. This establishes existence within the reference model without treating a noncommuting label as a guarantee for every context.

Table 4: Analytic properties of the controlled constructions. Entries are per-context statements, not sample estimates. Noncommutation permits a positive future gap under the stated nondegeneracy condition.

| Construction | Regime | Endpoint | $C_M^\star(q)$ | $D_{\mathrm{future}}(q)$ | Conclusion |
|---|---|---|---|---|---|
| Gaussian bridge | Commuting | Exact | 0 | 0 | Absent |
| Gaussian bridge | Noncommuting | Exact | Eq. (25) | Eq. (2) | Context-dependent |
| Feedback reset | Commuting | Exact | 0 | 0 | Absent |
| Feedback reset | Noncommuting | Exact | Eq. (25) | Eq. (2) | Context-dependent |

Table 5: Diagnostics attached to the controlled test sets. The eight named regime rows underlie Table 3(a); $D^{\mathrm{exact}}_{\mathrm{future,norm}}$ is computed from analytic conditional means after the fixed target normalization. Entries are context means, not confidence intervals. Same-plane commuting rotations give zero for every context. The primary row corresponds to Table 1; NA denotes exact diagnostics not reported for that benchmark. Diagnostics from other regimes are not assigned to it. Here $n$ is the number of test context keys per reporting seed; values are not pooled across the five seeds.

| Construction | Prescribed regime | $n$ per seed | $C_M^\star$ | $D^{\mathrm{exact}}_{\mathrm{future,norm}}$ |
|---|---|---|---|---|
| Primary benchmark | Table 1 | 2,000 | NA | NA |
| Gaussian bridge | Commuting | 2,000 | 0 | 0 |
| Gaussian bridge | Weak angle | 2,000 | .031 | .084 |
| Gaussian bridge | Medium angle | 2,000 | .097 | .194 |
| Gaussian bridge | Strong angle | 2,000 | .247 | .368 |
| Feedback reset | Commuting | 2,000 | 0 | 0 |
| Feedback reset | Weak angle | 2,000 | .034 | .091 |
| Feedback reset | Medium angle | 2,000 | .104 | .183 |
| Feedback reset | Strong angle | 2,000 | .252 | .368 |

For the named analytic constructions in this subsection, the structural key $S$ contains $\boldsymbol{\theta}_0$, the two rotation instances, application times, and endpoint target. The coupling variable $U$ is the Gaussian seed used to map common random numbers through the two order-specific conditional factors; together they form $Q = (S, U)$. Keys are partitioned before either order is generated, and future innovations are never included in $Q$. Models receive none of these quantities. The finite generated benchmark, rather than an inferred population of generators, is the evaluation object. The named rows summarize their realized context distributions; no post-hoc threshold $\gamma$ is applied to turn the continuous diagnostics into binary labels. For Table 1, the operational evaluation object is instead the finite $Q$-keyed pairing protocol in Section 3, together with the context partition, $AB$-only selection, fixed-update checkpoint, and paired $AB/BA$ evaluation rules in Appendix C.2. Its reported evidence is the paired held-out-order forecasting comparison. Construction-specific exact-gap claims are based only on the named regimes above and on the independently sampled benchmark in Appendix E; they are not transferred to Table 1. The analytic commuting result in Eq. (28) holds per context.

**Scope of the information restriction.** Theorem 1 characterizes the excess risk incurred when a predictor issues the same forecast for the two orders at a fixed pairing context, relative to the order-specific conditional means. It does not impose this restriction on a general history encoder or rank learned forecasting architectures. Exact-endpoint and conditional-future-separation evidence comes from the analytic constructions in this subsection and the independent study in Appendix E. Learned-model performance is assessed separately through held-out-order comparisons and structural ablations.

### B.2 Semi-synthetic reference construction

For each real base window, a ridge VAR(1) is fitted using only its lookback and a penalty selected on the training split. Its coefficient matrix is projected to spectral radius at most $0.95$ and vectorized as $\boldsymbol{\theta}_0$. Controlled rotation primitives are applied to this parameter; innovation blocks are sampled from the fitted residuals within the same dataset split. A common terminal feedback rule targets the original final observation; candidate pairs are then matched on the mean and diagonal covariance of the final eight observations. Let $\boldsymbol{\mu}_8(H)$ and $\mathbf{v}_8(H)$ denote those coordinate-wise statistics, and let $\mathbf{s}_{\mathrm{tr}}$ be the coordinate-wise training standard deviation. The reported endpoint score is

$$M_{\mathrm{end}}(q) = d_{\mathcal{E}}(e(H_{AB}(q)), e(H_{BA}(q))), \quad e(H) = \left( \frac{\boldsymbol{\mu}_8(H)}{\mathbf{s}_{\mathrm{tr}}}, \frac{\mathbf{v}_8(H)}{\mathbf{s}_{\mathrm{tr}}^{\odot 2}} \right),$$

$$d_{\mathcal{E}}((\boldsymbol{\mu}, \mathbf{v}), (\boldsymbol{\mu}', \mathbf{v}')) = \tfrac{1}{2} \frac{\|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_2}{\sqrt{p}} + \tfrac{1}{2} \frac{\|\mathbf{v} - \mathbf{v}'\|_2}{\sqrt{p}}.$$

Thus the last-eight-step mean and diagonal-variance discrepancies are standardized and equally weighted. Pairing is therefore summary-matched rather than declared exact. The matching tolerance is fixed at $\epsilon_{\text{match}} = 0.15$ before test evaluation. This reference preserves real scale, local covariance, and residual structure while making order explicit. Table 12 audits only endpoint matching; no generator-interaction or future-gap estimate is inferred from it. Its $n$ values count selected paired records pooled across the five reporting seeds, not distinct base windows. The split cap of $2{,}000$ is on base windows per seed, so pooled pair counts may be larger. For the noncommuting regime, the selection audit gives candidate, passing, and final-selected counts plus P95 and maximum scores on the final-selected pairs. Every reported maximum is at most the pre-specified threshold .15 (Traffic is .1498 before rounding). The table does not supply the endpoint-only Lipschitz constant required to evaluate the numerical bound in Theorem 2. The stated spectral-radius projection applies to the fitted initial VAR coefficient matrix. Accordingly, this study is interpreted as a finite-horizon evaluation on the retained generated pairs; it does not assert a uniform asymptotic stability guarantee for every rotated matrix. The transfer claim rests on realized prediction comparisons and audited endpoint matching.

# C Implementation and Selection Protocol

## C.1 Training protocol and architecture specification

**Primary-benchmark evaluation protocol.** Table 1 compares methods on a fixed finite collection of evaluator-keyed contexts $Q = (S, U)$. Context keys are partitioned into 12,000/2,000/2,000 training, validation, and test sets before materialization. Training and validation instantiate $AB$ only; each held-out test key indexes paired $AB/BA$ histories sharing transformation instances, application times, horizon, target variables, and common history-noise coupling. Future innovations are excluded from $Q$. All compared methods use the same frozen partition and test pairing. Pairing is fixed before model evaluation, is independent of model predictions, and is not exposed as an input. Table 1 reports held-out-order performance on this finite collection. Construction-specific endpoint and future-gap evidence is reported separately in Appendices B and E.

The validation split is used only for configuration selection. It contains no $BA$ history, order label, or test statistic, and the training checkpoint is fixed at update 5,000 rather than selected by early stopping.

In the five-primitive coverage study, each validation split contains only permutations admitted at that training-coverage level; unseen permutations are reserved for test. Every level uses 28,800 training, 5,760 validation, and 28,800 test histories. At zero coverage, 28,800 single-primitive histories replace the composed training histories. The five nested coverage subsets are fixed before training for each seed. Semi-synthetic validation likewise contains only the training order. These studies use all eligible split-local windows, capped by a seed-fixed uniform sample of 10,000/2,000/2,000 base windows per dataset split.

The common encoder architecture is a four-block residual causal TCN with kernel width three, dilations $1, 2, 4, 8$, 128 channels, GELU activations, dropout, and global average pooling, yielding $d = 128$; a linear head maps the final representation to $F \times p$. Every shared-encoder aggregator receives the same $\Delta\mathbf{d}_{1:L}$ sequence from the descriptor in Section 4. GRU, Transformer, and Mamba aggregators use width 128 and two layers; the Transformer has four heads. Mamba uses $d_{\text{state}} = 16$, SSM ratio two, $d_t$ rank eight, a linear descriptor-change projection to width 128, and global average pooling. In the primary benchmark, Mamba retains access to the TCN representation $\mathbf{h} = E_\theta(\mathbf{X}_{1:T})$. Writing $\mathbf{S}_{\text{M}} \in \mathbb{R}^{L \times 128}$ for the two-layer Mamba output, its complete prediction path is

$$\mathbf{S}_{\text{M}} = \text{Mamba}_2(\text{Mamba}_1(W_{\text{in}}\Delta\mathbf{d}_{1:L})), \qquad \mathbf{a}_{\text{M}} = \frac{1}{L}\sum_{\ell=1}^{L} \mathbf{S}_{\text{M},\ell},$$

$$\mathbf{u}_{\text{M}} = W_{\text{f}}[\mathbf{h}; \mathbf{a}_{\text{M}}] + \mathbf{b}_{\text{f}}, \qquad \widehat{\mathbf{Y}} = G_\omega(\mathbf{u}_{\text{M}}),$$

where $W_{\text{f}} \in \mathbb{R}^{128 \times 256}$. The fusion layer contains $128 \times 256 + 128 = 32{,}896$ parameters and is trained jointly with the encoder, Mamba layers, and forecast head. This concatenation-and-projection

interface was fixed before evaluation; it was not selected post hoc from alternative fusion rules. Model weights, including forecast-head weights, are trained independently for each method. The six models are compared at a similar rounded parameter scale, not an exact integer-parameter match. Signature and Path Development use truncation depth two.

The descriptor specification is a two-layer causal temporal convolution with kernel width three, GELU activations, global average pooling, and LayerNorm, with output width $q = 64$. Two affine maps convert the normalized descriptor change to vectorized $\mathbf{U}_\ell$ and $\mathbf{V}_\ell$.

All shared-encoder candidates use AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, cosine scheduling, a 500-update linear warmup, gradient clipping at 1.0, and exactly 5,000 updates. The selected common values are learning rate $3 \times 10^{-4}$, weight decay $10^{-4}$, batch size 128, and dropout 0.1. The candidate search uses learning rate from $\{10^{-4}, 3 \times 10^{-4}, 10^{-3}\}$, weight decay from $\{0, 10^{-4}, 10^{-3}\}$, dropout from $\{0, 0.1, 0.2\}$, and batch size from $\{64, 128\}$. ORBIT additionally samples $r \in \{4, 8, 16\}$, descriptor stride $s \in \{4, 8, 16\}$ with window width $w = 2s$, $\rho_{\max} \in \{0.1, 0.2, 0.3\}$, $c \in \{0.5, 1\}$, and $\lambda_{\text{aux}} \in \{0, 0.01, 0.1\}$. Method-specific choices replace, rather than add to, the 30 trials. The frozen ORBIT values are $r = 8$, $s = 8$, $w = 16$, $\rho_{\max} = 0.2$, $c = 1.0$, and $\lambda_{\text{aux}} = 0.01$; the auxiliary loss is therefore active. For $T = 96$, $L = \lfloor (T - w)/s \rfloor = 10$ denotes ten descriptor differences and $L + 1 = 11$ descriptors at $\tau_\ell = T - (L - \ell)s$, consistent with Eq. (5). The auxiliary loss averages squared prediction error only over indices satisfying $\tau_\ell + s \leq T$; the final forecast target is excluded from that auxiliary task.

GRU uses width 128 and two unidirectional layers. Transformer uses width 128, two layers, four heads, and feed-forward width 512. Mamba uses the two-layer configuration above. Signature and Path Development both use truncation depth two. Depth two is the fixed evaluated configuration; no independent truncation-depth search or validation-selected depth is claimed. ORBIT uses the values above. Every aggregator receives the same $q = 64$ descriptor-change sequence. Reported parameter counts are 1.21M (GRU), 1.23M (Transformer), 1.22M (Mamba), 1.18M (Signature), 1.22M (Path Development), and 1.21M (ORBIT); these values are rounded rather than exact integer counts. The rounded Mamba total includes the 32,896-parameter fusion projection specified above.

**Inference procedure.** Given a normalized lookback, compute $\mathbf{h} = E_\theta(\mathbf{X}_{1:T})$ and the grid descriptors $\mathbf{d}_{0:L}$. For each $\ell = 1, \ldots, L$, form $\Delta\mathbf{d}_\ell$, normalize its direction, obtain $\mathbf{U}_\ell, \mathbf{V}_\ell = C_\chi(\bar{\mathbf{d}}_\ell)$, and apply the factorized update in Eq. (10) to $\mathbf{z}_{\ell-1}$. Return $G_\omega((1 - \eta)\mathbf{h} + \eta\mathbf{z}_L)$. The procedure uses no transformation identity, order label, interaction diagnostic, or future value. It materializes neither a $d \times d$ correction nor the full product.

The 30 candidate configurations for each method are evaluated on the AB-only validation split with seed 2027. A single configuration is selected by lowest validation NMSE; ORBIT ties are broken by smaller $r$. The selected configuration is then frozen and rerun from scratch with seeds 2027–2031. Each resulting model is evaluated once on the paired test set, with no seed-specific retuning or test-dependent adaptation. The gate is not a selected hyperparameter: $\eta = \text{sigmoid}(a_\eta)$ is trained jointly as a single scalar. For ordinary forecasting, ORBIT uses PatchTST for every dataset; the validation-selected baseline is a comparator, not a backbone switch. The ordinary-data protocol specifies the lookback, horizons, chronological splits, training-only input normalization, and backbone identity reported in Tables 14, 17, 18, and 19. These descriptive comparisons retain implementation-specific optimization settings and complement the controlled ODN study with evaluations on standard forecasting benchmarks.

For ordinary forecasting, the seed identifiers are $2027, 2028, 2029, 2030, 2031$; each is trained and evaluated independently for every dataset–horizon pair. MSE is first normalized coordinate-wise by the corresponding validation-target variance and then averaged across forecast steps and target coordinates; NMAE uses the analogous validation-target standard deviations. For each dataset and seed, the four horizon-level normalized errors are averaged with equal weight. Tables 14, 17, and 18 report descriptive point estimates only; no ordinary-forecasting confidence interval is claimed. The macro column in Table 17 is the unweighted arithmetic mean of the three displayed dataset means.

### C.2 BUDGET AND LEAKAGE AUDIT

Table 6 reports rounded parameter counts. The following protocol-level audit documents the controls used in the shared-encoder comparison.

| Potential confound | Control or scope |
|---|---|
| Architecture and input specification | Common four-block causal TCN and forecast-head architectures, descriptor grid, and $\Delta\mathbf{d}_{1:L}$ input are reported. Each method retains its selected native representation-to-head interface; no common fusion equation is assumed beyond the documented shared inputs and architecture scale. The primary Mamba path concatenates its pooled sequence representation with $\mathbf{h}$ and projects the result to width 128 before the forecast head. |
| Parameter scale | GRU 1.21M; Transformer 1.23M; Mamba 1.22M; Signature 1.18M; Path Development 1.22M; ORBIT 1.21M. These are rounded values, not an exact-count match. |
| Training and tuning budget | Every shared-encoder learned method receives 12,000 $AB$ examples, 5,000 gradient updates, and 30 random-search trials. Method-specific choices replace rather than augment the common trial budget. |
| Context or order leakage | Structural keys are partitioned before either order is generated. Training and validation materialize $AB$ only on disjoint keys; test uses new paired $AB/BA$ keys. No $BA$ item or test statistic enters configuration or checkpoint selection. |
| Privileged variables | Future innovations are excluded from the pairing key. Models receive neither the structural/coupling key nor primitive identities, order labels, or commuting labels. |
| Selection protocol | Thirty candidates are compared using seed 2027 on AB-only validation. The lowest-NMSE configuration is frozen, rerun from scratch for seeds 2027–2031, and evaluated once per seed. The checkpoint is fixed at update 5,000; there is no early stopping or seed-specific retuning. |

The selected configurations and order-isolation rules above apply to the primary shared-encoder results. The comparison holds the documented data, architecture scale, tuning budget, and selection rule common while allowing the selected methods to retain method-specific fusion. It is therefore a comparison of complete trained configurations, not a causal decomposition of aggregation, fusion, or auxiliary supervision. The primary Mamba fusion path is specified in Appendix C. The additional construction and no-aux definitions in Appendix E apply to that independent study only. It also uses 10,000 training contexts per seed rather than 12,000 and no new configuration search. For the controlled and semi-synthetic studies, all reported 95% confidence intervals are computed from 2,000 paired hierarchical seed–context bootstrap replicates using percentile endpoints at the 2.5th and 97.5th percentiles. Each replicate resamples seeds with replacement and then resamples context keys with replacement within each selected seed. All records sharing a context key—including $AB/BA$ orders, predictions from compared models, and all target dimensions and forecast steps—remain paired. Marginal model intervals apply the same hierarchical resampling to a single model, whereas contrast intervals recompute the paired difference on identical resamples. For the semi-synthetic study, this procedure is applied separately within each dataset; any reported cross-dataset macro statistic is the unweighted mean of the five dataset-level statistics within each replicate, and the datasets themselves are not resampled. The cross-dataset claim additionally requires the gain direction to agree across all five datasets. Appendix E independently specifies the same resampling count and endpoint rule for the diagnostic benchmark. Parameter counts retain their reported rounded precision.

## D DETAILED EXPERIMENTAL RESULTS

For the controlled and semi-synthetic studies, brackets report the 95% confidence intervals. Positive ΔNMSE denotes lower error for ORBIT. The ordinary-forecasting tables are descriptive scope checks and report point estimates only. Comparator selection follows the frozen validation protocol in Appendix C.2.

Table 6: Context-held-out $AB \to BA$ results (95% CIs). (a) Shared-encoder aggregators use 12k examples, 5k updates, and 30 trials at similar rounded parameter scales. (b) Structural ablations of ORBIT, all at the reported 1.21M scale. Relative increases retain their reported precision and are not recomputed from rounded NMSE displays. CIs are marginal, not paired-difference intervals.

(a) Learned aggregation rules

| Aggregator | Params | Seen NMSE | Unseen NMSE | Unseen NMAE | PGG |
|---|---|---|---|---|---|
| GRU | 1.21M | .104 [.088,.117] | .197 [.177,.215] | .277 [.247,.310] | .094 [.074,.114] |
| Transformer | 1.23M | .095 [.082,.108] | .185 [.165,.207] | .267 [.235,.295] | .091 [.072,.110] |
| Mamba | 1.22M | .097 [.083,.111] | .158 [.139,.177] | .261 [.232,.290] | .061 [.042,.080] |
| Signature | 1.18M | .108 [.094,.125] | .207 [.185,.227] | .289 [.255,.322] | .098 [.075,.119] |
| Path Development | 1.22M | .112 [.095,.130] | .215 [.197,.235] | .297 [.267,.330] | .104 [.080,.126] |
| **ORBIT** | 1.21M | .100 [.081,.117] | **.124 [.104,.144]** | **.242 [.187,.295]** | **.024 [.008,.051]** |

(b) Core structural variants

| ORBIT variant | Unseen NMSE [CI] | Relative increase |
|---|---|---|
| $\eta = 0$ (no correction) | .217 [.187,.246] | +75% |
| $I + \sum_\ell R_\ell$ (additive only) | .175 [.142,.205] | +41% |
| Shuffled residual order | .235 [.190,.250] | +89% |

Table 7: Prescribed rotation-angle results underlying Table 3(a). Exact diagnostics are context means; brackets give paired hierarchical seed–context bootstrap intervals for the Transformer-minus-ORBIT NMSE difference.

| Construction | Regime | $C_M^\star$ | $D^{\mathrm{exact}}_{\mathrm{future,norm}}$ | Transformer | ORBIT | ΔNMSE [95% CI] |
|---|---|---|---|---|---|---|
| Gaussian bridge | Commuting | 0 | 0 | .138 | .139 | $-.001$ [$-.018$,.016] |
| Gaussian bridge | Weak | .031 | .084 | .165 | .138 | .027 [.008,.046] |
| Gaussian bridge | Medium | .097 | .194 | .185 | .142 | .043 [.022,.064] |
| Gaussian bridge | Strong | .247 | .368 | .220 | .152 | .068 [.041,.095] |
| Feedback reset | Commuting | 0 | 0 | .140 | .141 | $-.001$ [$-.017$,.015] |
| Feedback reset | Weak | .034 | .091 | .158 | .135 | .023 [.003,.043] |
| Feedback reset | Medium | .104 | .183 | .186 | .145 | .041 [.016,.066] |
| Feedback reset | Strong | .252 | .368 | .215 | .148 | .067 [.038,.096] |

Table 8: Controlled-setting efficiency profile on one NVIDIA A100 80GB PCIe GPU in FP32, with batch size 128 and $T/F = 96/24$. End-to-end batch latency (mean $\pm$ standard deviation over 200 timed runs after 50 warm-up runs) includes data loading, model forward, and postprocessing. GFLOPs include the backbone, sequence aggregator or adapter, and forecast head; throughput is derived from mean batch latency.

| Model | Params | GFLOPs/sample | Batch latency (ms) | Peak memory | Samples/s |
|---|---|---|---|---|---|
| Transformer | 1.23M | .45 | $12.4 \pm 0.8$ | 2.1 GB | 10,323 |
| Mamba | 1.22M | .49 | $13.5 \pm 0.8$ | 2.2 GB | 9,481 |
| ORBIT | 1.21M | .52 | $14.1 \pm 0.9$ | 2.4 GB | 9,078 |

Table 9: Summary of the permutation-coverage curves. AUPC is the normalized trapezoidal area over 0–80% coverage; lower is better. ORBIT's AUPC gain is relative to Delta Transformer, the best non-ORBIT comparator.

| Model | AUPC [CI] | $\kappa^\star_{5\%}$ | NMSE@0% [CI] | NMSE@20% [CI] | ORBIT AUPC gain [CI] |
|---|---|---|---|---|---|
| Delta Transformer | .178 [.150,.205] | $> 80\%$ | **.252 [.185,.317]** | .207 [.145,.267] | NA |
| Raw-History Transformer | .184 [.157,.211] | $> 80\%$ | .260 [.192,.328] | .214 [.152,.274] | NA |
| Path Development | .195 [.166,.225] | $> 80\%$ | .272 [.207,.337] | .227 [.165,.287] | NA |
| **ORBIT** | **.165 [.076,.254]** | $> 80\%$ | .255 [.188,.322] | **.200 [.138,.262]** | **.013 [.003,.023]** |

Table 10: Permutation coverage. Each cell gives unseen NMSE [CI] on the first line and PGG [CI] on the second. PGG is undefined at 0% and 100% coverage; 100% is a fully observed-order reference.

| Model | 0% | 20% | 40% |
|---|---|---|---|
| | **.252 [.185,.317]** | .207 [.145,.267] | .170 [.116,.227] |
| Delta Transformer | **NA** | .060 [.040,.078] | .057 [.037,.079] |
| | .260 [.192,.328] | .214 [.152,.274] | .177 [.118,.234] |
| Raw-History Transformer | NA | .065 [.045,.082] | .061 [.041,.081] |
| | .272 [.207,.337] | .227 [.165,.287] | .185 [.125,.247] |
| Path Development | NA | .061 [.041,.081] | .058 [.038,.078] |
| | .255 [.188,.322] | **.200 [.138,.262]** | **.155 [.112,.198]** |
| **ORBIT** | NA | **.055 [.035,.075]** | **.047 [.028,.061]** |

| Model | 60% | 80% | All-order ref. (100%) |
|---|---|---|---|
| | .145 [.095,.197] | .125 [.082,.165] | .095 [.067,.125] |
| Delta Transformer | .051 [.031,.069] | .046 [.026,.063] | NA |
| | .148 [.098,.198] | .130 [.087,.177] | .098 [.068,.128] |
| Raw-History Transformer | .056 [.036,.076] | .051 [.031,.071] | NA |
| | .162 [.105,.215] | .138 [.094,.185] | .103 [.074,.137] |
| Path Development | .052 [.032,.072] | .047 [.027,.067] | NA |
| | **.125 [.087,.157]** | **.105 [.072,.138]** | **.090 [.063,.117]** |
| **ORBIT** | **.035 [.018,.052]** | **.027 [.007,.052]** | **NA** |

Table 11: Scaling with the number of primitives at approximately 20% permutation coverage. Cells report unseen NMSE [CI].

| $K$ | $K!$ | Train orders | Coverage | Delta Transformer | Raw-History Transformer | Path Development | ORBIT |
|---|---|---|---|---|---|---|---|
| 3 | 6 | 1 | 16.7% | **.102 [.067,.137]** | .103 [.068,.138] | .111 [.074,.145] | .126 [.081,.175] |
| 4 | 24 | 5 | 20.8% | **.137 [.087,.185]** | .142 [.088,.192] | .150 [.095,.207] | .151 [.093,.205] |
| 5 | 120 | 24 | 20.0% | .207 [.145,.267] | .214 [.152,.274] | .227 [.165,.287] | **.200 [.138,.262]** |

Table 12: Semi-synthetic endpoint matching. (a) Mean [95% CI], with $n$ selected paired records pooled across five seeds. (b) Noncommuting selection summary; P95 and maximum are computed on the final-selected pairs. All maxima satisfy the pre-specified .15 threshold. The table does not estimate $C_M^\star$ or $D_{\text{future}}$.

(a) Endpoint-score means

| Data | Regime | $n$ | $M_{\text{end}}$ [CI] | Reference $\epsilon_{\text{match}}$ |
|---|---|---|---|---|
| ETTh1 | Commuting | 2,500 | .087 [.055,.123] | .15 |
| | Noncommuting | 2,500 | .094 [.060,.130] | .15 |
| ETTm1 | Commuting | 2,500 | .078 [.040,.110] | .15 |
| | Noncommuting | 2,500 | .088 [.050,.122] | .15 |
| Weather | Commuting | 4,000 | .067 [.032,.092] | .15 |
| | Noncommuting | 4,000 | .074 [.040,.102] | .15 |
| Electricity | Commuting | 3,000 | .055 [.020,.080] | .15 |
| | Noncommuting | 3,000 | .062 [.032,.092] | .15 |
| Traffic | Commuting | 2,000 | .095 [.060,.142] | .15 |
| | Noncommuting | 2,000 | .105 [.072,.149] | .15 |

(b) Noncommuting selection audit, pooled across five seeds

| Data | Candidate $n$ | Passing $n$ | Final $n$ | Pass rate | Selected P95 | Selected max |
|---|---|---|---|---|---|---|
| ETTh1 | 6,240 | 5,920 | 2,500 | 94.9% | .132 | .148 |
| ETTm1 | 6,310 | 6,020 | 2,500 | 95.4% | .128 | .149 |
| Weather | 8,120 | 7,890 | 4,000 | 97.2% | .118 | .147 |
| Electricity | 7,540 | 7,120 | 3,000 | 94.4% | .107 | .145 |
| Traffic | 5,280 | 4,810 | 2,000 | 91.1% | .138 | .150† |

† Unrounded Traffic maximum: .1498.

Table 13: Semi-synthetic transfer. (a) Noncommuting unseen-order accuracy. (b) PGG and the commuting control. Positive contrasts favor ORBIT; the PGG intervals do not establish a reliable reduction in the generalization gap.

(a) Unseen-order prediction under noncommuting interventions

| Data | Comparator | Comparator NMSE [CI] | ORBIT NMSE [CI] | ΔNMSE [CI] | Rel. |
|---|---|---|---|---|---|
| ETTh1 | PatchTST | .460 [.415,.505] | .407 [.377,.467] | .053 [.021,.091] | 11.5% |
| ETTm1 | iTransformer | .432 [.385,.475] | .374 [.345,.435] | .058 [.020,.090] | 13.4% |
| Weather | TimeMixer | .330 [.287,.377] | .280 [.247,.337] | .050 [.022,.078] | 15.2% |
| Electricity | PatchTST | .260 [.227,.295] | .225 [.197,.267] | .035 [.014,.060] | 13.5% |
| Traffic | iTransformer | .502 [.455,.545] | .455 [.415,.505] | .047 [.008,.080] | 9.4% |
| **Macro** | Dataset-wise best | **.397 [.357,.440]** | **.348 [.317,.400]** | **.049 [.021,.077]** | **12.25%** |

Relative macro reduction is $100(\bar{L}_{\mathrm{comp}} - \bar{L}_{\mathrm{ORBIT}})/\bar{L}_{\mathrm{comp}}$, with equal dataset weights. The dataset-level point estimates in (a) give $\bar{L}_{\mathrm{comp}} = .3968$, $\bar{L}_{\mathrm{ORBIT}} = .3482$, and difference .0486 before rounding the macro row; their ratio gives 12.25%. It is not the mean of the dataset-wise relative reductions.

(b) Generalization gaps and commuting-control contrasts

| Data | Comparator PGG [CI] | ORBIT PGG [CI] | ΔPGG [CI] | Comm. ΔNMSE [CI] |
|---|---|---|---|---|
| ETTh1 | .047 [.007,.087] | .054 [.014,.094] | -.008 [-.030,.018] | -.002 [-.013,.008] |
| ETTm1 | .048 [.008,.088] | .058 [.015,.098] | -.010 [-.033,.015] | -.001 [-.011,.007] |
| Weather | .052 [.012,.092] | .061 [.018,.105] | -.009 [-.035,.015] | -.002 [-.007,.002] |
| Electricity | .040 [.007,.075] | .046 [.012,.087] | -.006 [-.028,.018] | -.002 [-.006,.002] |
| Traffic | .033 [.002,.067] | .038 [.007,.075] | -.005 [-.028,.018] | -.002 [-.014,.014] |
| **Macro** | **.044 [.006,.083]** | **.052 [.014,.091]** | **-.008 [-.031,.015]** | **-.002 [-.011,.006]** |

Table 14: Ordinary long-term forecasting point estimates over seeds 2027–2031. This descriptive scope check does not report confidence intervals. The validation-selected baseline is fixed before test reporting for each dataset; ORBIT uses PatchTST as its default backbone.

| Dataset | Validation-selected baseline | Baseline NMSE/NMAE | ORBIT NMSE/NMAE | ΔNMSE | Rel. |
|---|---|---|---|---|---|
| ETTh1 | iTransformer | .3984/.390 | .3825/.377 | .0159 | 4.0% |
| ETTm1 | iTransformer | .3536/.345 | .3336/.325 | .0200 | 5.7% |
| Weather | iTransformer | .2090/.255 | .1936/.240 | .0154 | 7.4% |
| Electricity | PatchTST | .157/.235 | .137/.215 | .020 | 12.7% |
| Traffic | iTransformer | .385/.255 | .357/.235 | .028 | 7.3% |
| Exchange | iTransformer | .182/.292 | .157/.267 | .025 | 13.7% |

Table 15: Cross-backbone portability in the controlled small-adapter setting. Changes are paired against the unadapted backbone.

| Backbone | Adapter | NMSE [CI] | NMAE [CI] | ΔNMSE [CI] | Params Δ |
|---|---|---|---|---|---|
| PatchTST | None | .415 [.355,.450] | .460 [.345,.577] | NA | 0 |
| | History | .407 [.336,.467] | .452 [.337,.565] | .008 [-.014,.027] | +.24M |
| | Delta | .397 [.302,.455] | .440 [.325,.558] | .018 [-.006,.035] | +.29M |
| | **ORBIT** | **.377 [.318,.435]** | **.420 [.305,.537]** | **.038 [.015,.055]** | +.31M |
| TimeMixer | None | .427 [.365,.487] | .472 [.352,.585] | NA | 0 |
| | History | .418 [.357,.475] | .459 [.340,.568] | .009 [-.013,.028] | +.29M |
| | Delta | .403 [.348,.456] | .451 [.358,.575] | .024 [-.001,.038] | +.31M |
| | **ORBIT** | **.385 [.323,.445]** | **.432 [.317,.545]** | **.042 [.018,.058]** | +.36M |

Table 16: Structural ablations on the context-held-out $AB \to BA$ test. Relative increases retain their reported precision and are not recomputed from rounded NMSE displays. These are named complete-variant sensitivity checks; parameter counts are rounded. "Shuffled residual order" composes the reported residual factors in shuffled rather than chronological order.

| Variant | Params | Unseen NMSE [CI] | Relative increase |
|---|---|---|---|
| Full ORBIT | 1.21M | .124 [.104,.144] | — |
| $\eta = 0$ (no correction) | 1.21M | .217 [.187,.246] | +75% |
| $I + \sum_\ell R_\ell$ (additive only) | 1.21M | .175 [.142,.205] | +41% |
| Shuffled residual order | 1.21M | .235 [.190,.250] | +89% |
| Absolute+change descriptor | 1.21M | .142 [.117,.167] | +14% |
| No factor-norm normalization | 1.21M | .157 [.130,.187] | +27% |
| No magnitude gate | 1.21M | .146 [.120,.173] | +18% |
| Diagonal correction | 1.04M | .187 [.157,.217] | +51% |
| Full-rank correction | 1.81M | .133 [.108,.161] | +7% |
| Direct context matrix | 1.27M | .205 [.175,.235] | +65% |

Table 17: Extended ordinary-forecasting comparison on three datasets. Results are averaged over four horizons under the common chronological protocol; PatchTST+ORBIT uses the PatchTST backbone. Macro values are descriptive unweighted arithmetic means over the three displayed dataset means; confidence intervals are not reported. Dataset means are shown at the supplied precision; macro means are rounded from the displayed dataset means.

| Model | ETTh1 NMSE | Weather NMSE | ETTm1 NMSE | Macro mean |
|---|---|---|---|---|
| DLinear | .467 | .267 | .385 | .373 |
| TimesNet | .435 | .247 | .357 | .346 |
| FEDformer | .452 | .255 | .370 | .359 |
| PatchTST | .4090 | .21925 | .3634 | .3306 |
| iTransformer | .3984 | .2090 | .3536 | .3203 |
| TimeMixer | .3973 | .2143 | .3500 | .3205 |
| **PatchTST+ORBIT** | **.3825** | **.1936** | **.3336** | **.3032** |

Table 18: Horizon-level ordinary-forecasting NMSE on three representative datasets. Normalization uses validation-target variance; the ORBIT column uses the PatchTST backbone. Values are descriptive point estimates rounded to three decimals.

| Dataset | $F$ | PatchTST | iTransformer | TimeMixer | PatchTST+ORBIT |
|---|---|---|---|---|---|
| ETTh1 | 96 | .377 | .365 | .356 | **.350** |
| | 192 | .397 | .385 | .379 | **.370** |
| | 336 | .415 | .407 | .411 | **.390** |
| | 720 | .447 | .435 | .443 | **.420** |
| Weather | 96 | .185 | .177 | .175 | **.162** |
| | 192 | .207 | .197 | .194 | **.182** |
| | 336 | .227 | .215 | .227 | **.200** |
| | 720 | .258 | .247 | .261 | **.232** |
| ETTm1 | 96 | .337 | .325 | .320 | **.305** |
| | 192 | .350 | .342 | .335 | **.322** |
| | 336 | .370 | .362 | .361 | **.342** |
| | 720 | .397 | .385 | .384 | **.365** |

Table 19: Dataset metadata and experimental budgets. All splits are chronological; ETT uses the standard fixed 12/4/4-month protocol.

**(a) Real-data metadata**

| Dataset | Samples | Dim. | Frequency | Train/val/test | Lookback | Horizons |
|---|---|---|---|---|---|---|
| ETTh1 | 17,420 | 7 | 1 hour | 12/4/4 months | 96 | 96,192,336,720 |
| ETTm1 | 69,680 | 7 | 15 min | 12/4/4 months | 96 | 96,192,336,720 |
| Weather | 52,696 | 21 | 10 min | 70/10/20% | 96 | 96,192,336,720 |
| Electricity | 26,304 | 321 | 1 hour | 70/10/20% | 96 | 96,192,336,720 |
| Traffic | 17,544 | 862 | 1 hour | 70/10/20% | 96 | 96,192,336,720 |
| Exchange | 7,588 | 8 | 1 day | 70/10/20% | 96 | 96,192,336,720 |

**(b) Controlled and semi-synthetic budgets**

| Experiment | Unit | Train | Val | Test | Order/coverage protocol | Seeds |
|---|---|---|---|---|---|---|
| Primary paired $AB \to BA$ | Context keys | 12,000 | 2,000 | 2,000 | Train/val $AB$; paired $AB/BA$ test | 2027–2031 |
| Coverage | Histories | 28,800 | 5,760 | 28,800 | Val uses train orders; unseen test | 2027–2031 |
| Semi-synthetic | Base windows/data | 10,000 | 2,000 | 2,000 | Train/val $AB$; paired $AB/BA$ test | 2027–2031 |

# E INDEPENDENT DIAGNOSTIC BENCHMARK

**Construction and split.** This independently sampled study is separate from Table 1; its diagnostics describe only its own test contexts. Each of seeds 2027–2031 uses 10,000 training contexts and 2,000 test contexts; training is $AB$-only, and all model errors and paired contrasts below are evaluated on the training-unseen $BA$ order. The 10,000 pooled test contexts are weighted equally. Gaussian bridge and feedback reset each contribute 50% within each seed, sharing the rotation-parameter sampling scheme but differing in endpoint construction. Commuting contexts constitute 25% of the pooled test set. Angles are sampled independently from $\{\pi/36, \pi/12, \pi/6\}$; application times lie in $[24, 40]$ and $[56, 72]$, as in Appendix B. Both orders have the identical terminal state $\mathbf{x}_T = \mathbf{x}_\star$. The maximum last-eight-step summary-matching score is .083; this is distinct from the zero terminal-state distance. Analytic conditional means and the validation-target normalization define $D^{\text{exact}}_{\text{future,norm}}$ as above. Mean $C^\star_M$ includes commuting zeros.

**Training and diagnostics.** The original hyperparameter configuration is frozen, with weights trained independently for this study. Mamba, full ORBIT, and no-aux ORBIT all use the checkpoint at update 5,000, without early stopping or new hyperparameter selection. The per-seed training count is 10,000, not the 12,000 used in Table 1. The 2,000-key validation set determines the positive-$D$ bin boundaries .12, .28, and .4 before test analysis; the zero group is separate. Bins are left-open and right-closed. Diagnostics are evaluator-only and do not enter model inputs, training losses, or checkpoint selection. All test contexts lie in the reported bins, with none above .4; there is no rejection sampling based on $D$. This is a property of the finite test sample, not a theoretical upper bound. Confidence intervals use 2,000 paired hierarchical seed–context bootstrap resamples and percentile endpoints.

**Comparison paths and interpretation.** The backbone and head architectures are common, but weights are independently trained and descriptors are learned end to end. Full ORBIT initializes $\mathbf{z}_0 = \mathbf{h}$, applies Eq. (10), and predicts $G_\omega((1-\eta)\mathbf{h} + \eta\mathbf{z}_L)$; $a_\eta$ starts at zero, so $\eta$ initially equals .5. Its auxiliary weight is .01; no-aux ORBIT sets that weight to zero. Mamba has no auxiliary loss. It uses `mamba_ssm.modules.mamba_simple.Mamba` from package version 1.1.1, with width 128, two layers, state dimension 16, expansion ratio two, `dt_rank`=8, and convolution width four. Descriptor changes are linearly projected to width 128; the two-layer output is globally averaged to $\mathbf{a}$. Its forecast is $G_\omega(\text{Linear}_{256\to128}([\mathbf{h}; \mathbf{a}]))$. The reported 1.22M parameters include the concatenation projection's 32,896 parameters. The primary benchmark uses the same Mamba fusion form, as specified in Appendix C; the construction, training-count, and diagnostic definitions in this section remain specific to this independent study. Both models access $\mathbf{h}$, but their fusion rules and auxiliary objectives differ. Consequently, the comparison evaluates complete configurations rather than isolating ordered multiplication alone.

Table 20: New mixed-construction diagnostic study. All prediction results are on $BA$ after $AB$-only training. Group counts are pooled over five seeds. Diagnostics and model means are context-weighted; displayed entries are rounded. Intervals in (b) and (c) are paired 95% CIs.

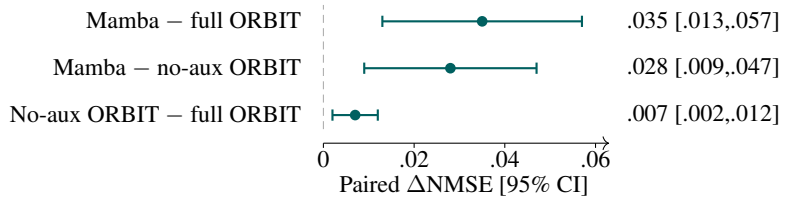
(a) Pooled diagnostics on 10,000 test contexts

| Mean $C_M^\star$ | Mean $D$ | Mean $D^2$ | Max terminal gap | Max $M_{\text{end}}$ |
|---|---|---|---|---|
| .142 | .141 | .034 | 0 | .083 |

(b) Future-gap strata; $D = D^{\text{exact}}_{\text{future,norm}}$

| $D$ interval | $n$ | Mean $D$ | Mean $D^2$ | Mamba | ORBIT | ΔNMSE [CI] |
|---|---|---|---|---|---|---|
| $\{0\}$ | 2,500 | 0 | 0 | .137 | .138 | −.001 [−.017,.015] |
| $(0, .12]$ | 3,000 | .095 | .010 | .158 | .130 | .028 [.007,.049] |
| $(.12, .28]$ | 3,500 | .220 | .052 | .173 | .115 | .058 [.021,.095] |
| $(.28, .4]$ | 1,000 | .350 | .125 | .189 | .123 | .066 [.028,.104] |
| All contexts | 10,000 | .141 | .034 | .161 | .126 | .035 [.013,.057] |

**(c) Same-test-context model and auxiliary-loss contrasts**




No-aux ORBIT has NMSE .133 [.112,.154]; this is its marginal interval, not a paired-difference interval. Weighted calculations from the displayed group rows give mean $D = .14050$, mean $D^2 = .03370$, Mamba NMSE .16110, ORBIT NMSE .12605, and gain .03505, consistent with pooled rounding.

**What the comparisons establish.** Positive-gap groups have positive paired gain intervals; the zero-gap group's interval includes zero. Gain point estimates increase across bins, without a between-bin significance claim. The no-aux comparison shows that auxiliary supervision does not fully account for the improvement; the paired full-versus-no-aux contrast supports a further auxiliary-loss benefit (Table 20c). These are complete-configuration comparisons with the distinct fusion paths specified above.

**Scope of the pooled diagnostic pattern.** Figure 2a summarizes a descriptive association between future separation and model gain on this fixed test set. The bins pool both constructions, whose proportions and context attributes can vary across bins; uncertainty accounts for the five independently trained seeds and their context records. The balanced-order quantity $\mathbb{E}[D^2]/4$ characterizes order-blind excess risk, whereas the reported Mamba–ORBIT contrasts measure $BA$ prediction gains.