ONE PATCH, THREE ROLES: WHAT IS ACTUALLY COUPLED IN AUTOREGRESSIVE TIME-SERIES FORECASTING?

Anonymous authors
Paper under double-blind review

ABSTRACT

Patch-based autoregressive time-series forecasting often ties input representation, learned transition span, and per-call output length to one patch length. We investigate whether these roles can be adjusted separately. For encoding, we explore building patches from short temporal atoms. We find that grouping preferences vary across data sets and model widths, but performance is more sensitive to width than to the number of atoms per patch on the evaluated grid. For decoding, we ask whether models can predict more subsequent patches per call. Lightweight parallel exits fit a parent's recursive trajectory substantially more easily than the observed future, motivating Autoregressive Trajectory Distillation (ATD). ATD compiles the frozen parent's rollout into selectable multi-patch execution. ATD-8 achieves a 5.54× end-to-end speedup with stable quality across widths in a paired four-data-set comparison. However, imitation can preserve the parent's accumulated forecast error. We therefore ask whether this error can be mitigated independently of execution width. We find a correctable residual projection along a train-selected periodic history direction, motivating Spectrum Tangent. This correction adds no neural parameters or Transformer calls. At horizon 720, it reduces mean squared error (MSE) and mean absolute error (MAE) by 2.54% and 2.33% across seven data sets and two execution widths, retaining a 3.24× speedup over recursive inference. Experiments on three public autoregressive parents and four data sets demonstrate both methods' transferability, with Tangent providing further long-horizon improvements.

1 INTRODUCTION

Autoregressive (AR) forecasting feeds predicted blocks back as context. Patch-based Transformers encode consecutive observations as tokens for next-patch supervision and flexible forecast horizons (Liu et al., 2024d; 2025b; 2024c; Das et al., 2024). A general forecaster must accommodate different temporal scales while supporting efficient long-horizon execution.

A patch plays a broader role here than in a standard Vision Transformer for image classification (Dosovitskiy et al., 2021). It sets input token granularity and, in a common next-patch AR design, also the supervised target and recursive writeback span. For example, forecasting 96 points with 24-point outputs takes four recursive calls; 48-point outputs take two. Larger outputs reduce feedback steps, but also change the learning problem, so fewer calls alone do not guarantee lower error. Patch length thus links representation, training, execution cost, and the path along which forecast errors accumulate.

We ask: *which roles assigned to an AR patch are actually coupled?* Let p_0 denote the representation atom, P the learned transition span (training target and recursive writeback), and K the transitions committed per call, giving a KP-point output. Conventional one-patch execution imposes:

$$\underbrace{p_0 = P}_{\text{representation = transition}} \quad , \qquad \underbrace{K = 1}_{\text{transition = commit}} \quad . \tag{1}$$

We relax these constraints one at a time, testing whether fixed atoms can encode a transition patch and whether one call can emit several learned transitions.

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

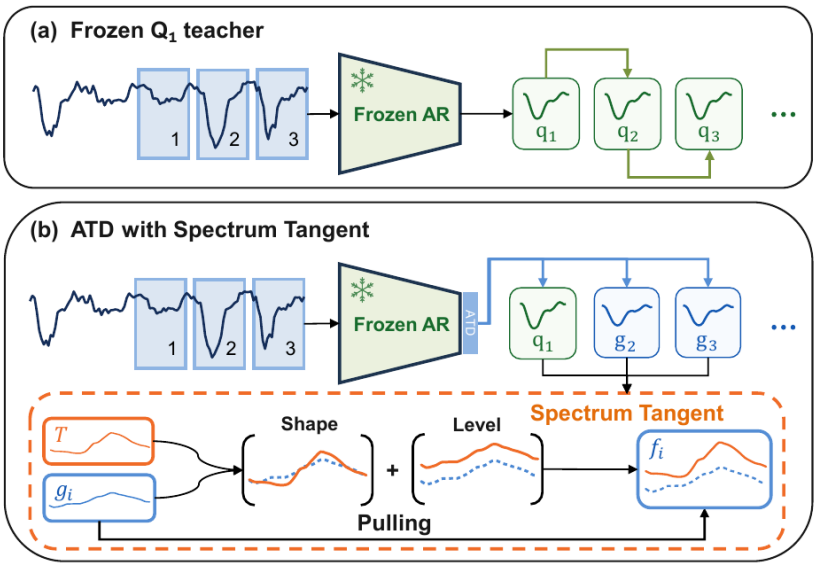

Figure 1: **The execution-decoupling branch and its residual correction.** **(a)** The frozen one-patch autoregressive (AR) parent (snowflake) recursively produces teacher patches q_i. **(b)** Autoregressive Trajectory Distillation (ATD) emits $G = (q_1, g_2, \ldots, g_K)$ in one call, training lightweight exits g_i to imitate q_i. Spectrum Tangent moves later proposals toward a train-selected history template T before writeback. Level and mean-free shape are diagnostic views of this single displacement. Dashed blue and solid orange curves show the proposal and corrected forecast; inference uses no future labels. The supporting atomic-encoding study appears in section 4.2 and appendix B.

We first examine the representation–transition constraint $p_0 = P$. Adaptive patching and frequency-specific projections highlight the importance of temporal scale for heterogeneous data (Chen et al., 2024; Woo et al., 2024). As a supporting study, we encode each patch with a shared projection of 12-point atoms, concatenate their representations chronologically, and project the result to model width D, retaining the original P-point training target. With an appropriate width, this atomic construction maintains accuracy comparable to direct whole-patch projection, with greater sensitivity to model width than to atom grouping on the evaluated grid.

We next examine the transition–execution constraint $K = 1$: can a model predict more subsequent patches per call while retaining its learned transition span? We find that lightweight parallel exits fit a parent's recursive trajectory substantially more easily than the observed future. This finding motivates Autoregressive Trajectory Distillation (ATD): exits imitate the frozen parent's later patches while preserving its first-patch output. One compiled model supports ATD-1/2/4/8, with ATD-1 exactly recovering the parent. ATD-8 achieves a 5.54× end-to-end speedup with little quality variation across widths in the paired four-data-set comparison.

Fewer calls, however, need not remove errors already present in the parent trajectory. In the 96-point example, ATD-2 can approximate the four-step parent rollout in two calls, including its errors. Parent-trajectory supervision improves fidelity in all 21 matched seed runs but forecast accuracy in only 15 against less stable clean-future exits. Efficient execution and residual correction are thus distinct objectives.

We therefore ask whether accumulated forecast error can be mitigated without changing patch length or execution width. Inspired by seasonal forecasting, we find a correctable residual projection toward a train-selected periodic history template. This motivates Spectrum Tangent, which fits one nonnegative coefficient along that direction without adding neural parameters or Transformer calls. Level and mean-free shape provide diagnostic views whose signs can disagree. Correction gains strengthen at longer horizons.

Our contributions are:

- We show that several choices commonly tied to patch length can be adjusted separately. A supporting atomic-encoding study shows that smaller atoms composed at an appropriate model width can maintain comparable forecast performance. A parent's readily fitted recursive trajectory supports committing multiple patches per call without relearning the transition span. Periodic history also provides a direction for correcting accumulated forecast error without changing commit size.
- We develop ATD to turn trajectory compressibility into selectable multi-patch execution from a frozen parent. ATD-1 exactly recovers the parent; wider modes accelerate inference while maintaining near-constant forecast quality in the matched comparison.
- We develop Spectrum Tangent to exploit the correctable residual projection using a train-selected periodic template and coefficient, without adding neural parameters or Transformer calls. Experiments on seven local data sets evaluate the two methods, and tests on three public AR parents across four data sets demonstrate their transferability.

2 Related work

Long-horizon forecasting. Long-horizon models use sparse attention, decomposition, patching, multiscale mixing, or cross-variable structure (Zhou et al., 2021; Wu et al., 2021; Zhou et al., 2022; Nie et al., 2023; Liu et al., 2024b; Wang et al., 2024); basis-expansion and MLP models offer alternatives (Oreshkin et al., 2020; Challu et al., 2023; Das et al., 2023). These methods optimize clean-future forecasts; our question concerns the roles inside a frozen AR patch interface. Simple baselines can also expose architectural assumptions that benchmark progress has left entangled, as DLinear did for long-horizon Transformer comparisons (Zeng et al., 2023). Our question is complementary to backbone design: which roles assigned to an AR patch are actually inseparable?

Autoregressive and foundation forecasters. DeepAR uses probabilistic autoregressive forecasting (Salinas et al., 2020); Timer/Timer-XL, AutoTimes, and TimesFM extend next-patch generation to modern and pretrained forecasters (Liu et al., 2024d; 2025b; 2024c; Das et al., 2024). Universal and zero-shot models broaden this direction (Ansari et al., 2024; Woo et al., 2024; Shi et al., 2025; Liu et al., 2025c; Auer et al., 2025). Moirai 2.0 predicts multiple tokens (Liu et al., 2025a), whereas Timer-S1 scales serial prediction (Liu et al., 2026b); SE-LLM is a recent 672-point-context/96-point-output reference (Liu et al., 2026a). These systems establish AR and multi-token forecasting as useful interfaces. Our question is how far representation and execution can vary while retaining a selected parent's learned transition.

Parallel autoregressive decoding. Blockwise decoding, multi-token prediction, and auxiliary heads produce multiple future tokens per call (Stern et al., 2018; Gloeckle et al., 2024; Cai et al., 2024). Speculative decoding uses a draft with online target verification (Leviathan et al., 2023); STRIDE adapts this design to continuous time-series patches (Subbaraman et al., 2025). ATD instead makes ATD-1 an exact bypass and removes the online parent pass for wider commits. We report rollout fidelity and forecast accuracy separately; the protocol and operating-point comparison is in tables 11 and 12.

Jacobi Forcing post-trains on its own parallel Jacobi decoding trajectories (Hu et al., 2026); ATD distills numeric patches together with their recursive writeback.

Exposure and rollout supervision. Classical multi-step forecasting distinguishes recursive, direct, DirRec, MIMO, and DIRMO strategies (Ben Taieb et al., 2012). ATD revisits this taxonomy at deployment: it retains the frozen recursive transition but emits selectable blocks. Unlike direct or MIMO forecasting, later outputs target the parent's rollout rather than clean futures. Boosting, scheduled sampling, and dataset aggregation instead address error or exposure mismatch (Ben Taieb & Hyndman, 2014; Bengio et al., 2015; Ross et al., 2011); rollout-aware refinement also appears in neural simulation (Lippe et al., 2023). Time-series distillation transfers teacher representations or predictions to a separate forecasting model (Guo et al., 2026; Fu et al., 2026). ATD instead freezes a validation-selected AR parent, compiles its deployed patch path into selectable widths, and preserves the original one-transition path as ATD-1.

Patch roles and spectral structure. PatchTST demonstrates the value of patch tokens (Nie et al., 2023). Pathformer adapts pathways across patch scales (Chen et al., 2024), and Moirai uses multiple patch-size projections to accommodate different frequencies (Woo et al., 2024). These approaches make temporal grouping an explicit adaptation choice. Our supporting atomic-encoding study complements them by comparing sensitivity to grouping and model width. TimesFM already allows different input and output patch lengths (Das et al., 2024); ATD examines widening execution while fitting a frozen parent's recursive path. Classical seasonal methods expose macroscopic level and seasonal structure (Winters, 1960; Cleveland et al., 1990), while frequency-aware models learn it inside the backbone (Wu et al., 2021; Zhou et al., 2022; Wu et al., 2023). Spectrum Tangent instead applies one external direction from an immutable history template and a period selected on training trajectories. Level and mean-free shape are diagnostic coordinates of this direction, rather than separate learned forecasting components.

3 Decoupling patch roles

3.1 Coupled baseline

Let $h_t \in \mathbb{R}^{C\times W}$ be the current history with C variables and W observed points, P the number of points predicted by the original one-transition forecaster, and K the number of successive transitions emitted per model call. Let $U(h, p)$ append patch p and retain the latest W points. The frozen parent f_θ includes instance normalization and its inverse, and returns $\hat{p}_{t,1} \in \mathbb{R}^{C\times P}$ in the history's data coordinates:

$$\hat{p}_{t,1} = f_\theta(h_t), \qquad h_{t+1} = U(h_t, \hat{p}_{t,1}). \tag{2}$$

Local parents and exits are channel-independent with shared weights across variables; no cross-variable mixing is used. We train f_θ from scratch using next-patch MSE and next-patch validation; ATD then freezes every parent parameter. "Timer-style" names this transition/writeback contract, not the exact Timer architecture.

Execution may commit K such transitions at once, spanning KP points. For forecast horizon H, the number of structural backbone calls is

$$C(H; P, K) = \left\lceil \frac{H}{KP} \right\rceil. \tag{3}$$

We separate representation from transition, then transition from execution, measuring both parent-trajectory fidelity and forecast accuracy.

3.2 An encoding-side observation

Against a monolithic baseline $R : \mathbb{R}^P \to \mathbb{R}^D$, we let $p_0 = 12$ and $m = P/p_0$. One shared linear map $\phi : \mathbb{R}^{12} \to \mathbb{R}^{16}$ lifts the m chronological atoms; we concatenate them without pooling and project the $16m$ coordinates to the data-set-selected width $D = d_{\text{model}}$. The atom and lift widths are global interface constants; each parent learns its own ϕ shared across atom positions and chooses P (hence m) and D.

If ϕ has full column rank, the concatenated features retain the ordered raw patch before the final projection, and every raw linear patch map remains representable. Section 4.2 compares matched training and the sensitivity along m and D; appendix B gives the proof and complete grid.

3.3 Recursive trajectories as compilation targets

From each training forecast origin h_0, we construct the frozen parent's deployed teacher trajectory:

$$h_0^Q = h_0, \quad q_1 = f_\theta(h_0^Q), \quad h_j^Q = U(h_{j-1}^Q, q_j), \quad q_{j+1} = f_\theta(h_j^Q), \qquad j = 1, \ldots, K-1. \tag{4}$$

For the local parent, each call recomputes per-series mean and standard deviation over the current W-point window, predicts in this normalization frame, maps the patch back for writeback, shifts, and renormalizes. This is additional to global benchmark scaling. Before one-shot regression, recursive teacher patches are re-expressed in the initial origin's normalization frame as $\bar{q}_j = (q_j - \mu_0)/\sigma_0$, channelwise. The local state $s_0 \in \mathbb{R}^{C\times D}$ is the last observed token after the final Transformer

block's post-normalization, before the output projection. Named-parent readouts are specified in appendix A.3. An independent residual multilayer perceptron (MLP) r_j predicts each later patch:

$$\bar{g}_1 = \bar{q}_1, \qquad \bar{g}_j = \bar{q}_1 + r_j(s_0) \in \mathbb{R}^{C\times P}, \quad j = 2, \dots, K. \tag{5}$$

Only the r_j are trained, using the origin-normalized rollout targets:

$$\mathcal{L}_{\mathrm{ATD}} = \frac{1}{CP(K-1)} \sum_{j=2}^{K} \|\bar{g}_j - \bar{q}_j\|_F^2 . \tag{6}$$

The data-space proposal $G = [g_1 \| \cdots \| g_K]$ uses $g_j = \mu_0 + \sigma_0 \bar{g}_j$. We call this compiled model ATD and denote execution that emits k transitions per call by ATD-k; ATD-1 bypasses every auxiliary exit and exactly recovers the frozen parent. Teacher states and recursive targets may be cached before fitting the later-patch exits.

Matched target controls. Frozen Direct uses the same frozen parent and independent residual exits, but replaces $\bar{q}_j$ in equation (6) by the identically normalized clean future patch $\bar{p}_j^\star$. Thus target is the only changed factor. Joint Direct also uses clean targets but jointly trains the parent-initialized backbone and K learned future placeholders to decode K clean-future patches in one causal pass. These controls isolate the target and distinguish ordinary joint optimization.

Selectable widths. An ATD-8 compiled model contains exits $r_2, \dots, r_8$. Width k evaluates only exits $r_2, \dots, r_k$ and writes back the resulting k patches. Local proposals therefore share exact prefixes at a common history, although complete $k > 1$ rollouts may diverge after different writebacks.

Proposition 1 (Exact ATD-1 fallback)**.** *The complete ATD-1 rollout is pointwise identical to the frozen parent rollout at every horizon.*

Proof. Both paths emit $f_\theta(h_0)$ and apply the same deterministic update U. Their next histories are identical; induction proves equality thereafter. □

3.4 A FIDELITY–ACCURACY MISMATCH

The parent trajectory already contains forecast error, including error propagated through recursive writeback. Fitting this trajectory can preserve that error even when execution uses fewer calls. Let $Q_\theta(h)$ denote the K-patch serial parent trajectory from history h, $G(h)$ the ATD proposal, and y the matching future. Define the rollout and forecast losses as $\mathcal{L}_{\mathrm{roll}}(G) = \mathbb{E}\|G - Q_\theta\|_2^2$ and $\mathcal{L}_{\mathrm{fore}}(G) = \mathbb{E}\|G - y\|_2^2$. With $e_{\mathrm{roll}} = G - Q_\theta$ and $e_Q = Q_\theta - y$, the identity

$$\mathcal{L}_{\mathrm{fore}}(G) = \mathcal{L}_{\mathrm{roll}}(G) + \mathcal{L}_{\mathrm{fore}}(Q_\theta) + 2\mathbb{E}\langle e_{\mathrm{roll}}, e_Q\rangle \tag{7}$$

shows why closer imitation need not improve accuracy: parent error is inherited and the cross term can have either sign. Section 4 tests this mismatch.

3.5 CORRECTING TRUTH ERROR OUTSIDE THE FROZEN PARENT

Spectrum Tangent asks whether the residual $e = y - G$ has a correctable projection along a simple history-derived direction while leaving the parent frozen. Inspired by periodic forecasting, it forms a phase-wise template T_τ from the normalized observed history at a train-selected, representation-aligned period τ. Its phase means are fixed at the forecast origin; modulo-τ lookup aligns each future point, and the current call's mean and scale map them to G's coordinates (appendix B). "Spectrum" refers to selecting τ over a period grid, while "Tangent" denotes the resulting one-dimensional correction of the frozen proposal. For absolute forecast endpoint ℓ, the complete correction direction is

$$d_\tau(G, \ell) = b(\ell)\big(T_\tau - G\big), \tag{8}$$

where $b(\ell)$ is a dimensionless scalar, constant within each patch, taking $0.25, 0.5, 1, 2, 3$ in successive bands separated by endpoints 96, 192, 336, and 672 (equation (14)). Here ℓ is measured from the forecast origin; every call's first slot has zero direction. Train-only origins determine

$$A_\tau = \mathbb{E}\|d_\tau\|_2^2, \quad B_\tau = \mathbb{E}\langle d_\tau, y - G\rangle, \quad V = \mathbb{E}\|y - G\|_2^2. \tag{9}$$

Among periods attaining the largest positive one-direction explained error $\max(B_\tau, 0)^2/(A_\tau V)$, the selector chooses the smallest and freezes

$$\alpha^* = \max(B_{\tau^*}/A_{\tau^*}, 0). \tag{10}$$

This is the exact nonnegative minimizer of a convex quadratic (derivation and selection protocol in appendix B). The frozen rule applies to slots 2–K:

$$F = G + \alpha^* d_{\tau^*}(G, \ell). \tag{11}$$

The minimizer is exact on the fixed proposal trajectory used to estimate its moments; closed-loop benefit remains empirical because corrected writeback changes later histories. The explicit wrapper uses only observed history and ATD proposals. It adds no trainable neural parameters or Transformer calls, although phase lookup and arithmetic add measurable latency. The inference rule uses the complete displacement in equation (8); section 4.5 later decomposes it into orthogonal level and mean-free shape coordinates only to diagnose where its alignment comes from.

4 Experiments

The experiments follow the empirical chain. We ask: (RQ1) how do atomic grouping and model width affect forecasting after factorizing the input patch; (RQ2) is a frozen transition's recursive trajectory sufficiently compressible to support wider output blocks with near-constant forecast quality; (RQ3) does improved rollout fidelity order forecast accuracy, and does the remaining residual align with a history-derived direction; and (RQ4) do trajectory compilation, the intervening mismatch, and the correction recur across named AR parents and data sets?

4.1 Protocol

Data, selection, and reporting. We use ETTh1, ETTh2, ETTm1, ETTm2 (Zhou et al., 2021), Weather, Electricity (ECL), and Traffic with lookback $W = 672$, training-set global standardization, and seeds 2021–2023. We report rolling horizons $H \in \{96, 192, 336, 720\}$ over every chronological test origin. We abbreviate these as H96, H192, H336, and H720. Parents use next-patch validation and multi-transition decoders use closed-loop H720 validation. Tangent uses one train-only period and coefficient per data-set–width configuration, shared across its three parent seeds; chronological selection and pooling controls are detailed in appendix B. All data-set-specific parent and tangent settings are deferred to table 8.

Metrics and systems. Let i, s, K, and b index data set, seed, commit width, and chronology block. We call (i, K) a configuration, (i, K, s) a seed run, and (i, K, s, b) a seed–block. Within the named split, time-ordered origins are partitioned into four contiguous, near-equal-count blocks. Training blocks serve the selector described above; test seed–blocks are reporting-only and never affect selection. Block counts are descriptive slices, not independent random trials. Mean squared error (MSE) and mean absolute error (MAE) use standardized benchmark coordinates. "Equal-configuration" is an unweighted mean over the stated (i, K) configurations after averaging all three seeds and test origins. Batch-one RTX5880-Ada-48Q timing includes rolling normalization, writeback, exits, and wrapper arithmetic; structural calls and milliseconds remain separate.

Table 1 first establishes the frozen transition as a reasonable local AR parent. The external block is contextual rather than a controlled comparison with the local rows. Following the compact convention of Liu et al. (2025b), it provides four-horizon context from Timer, UniTST, iTransformer, and PatchTST (Liu et al., 2024d;a;b; Nie et al., 2023); the appendix restores DLinear (Zeng et al., 2023) and the other reported baselines, together with every per-horizon MSE/MAE entry. The matched evidence below follows the two separations and the correction of inherited forecast error.

4.2 Atomic encoding: sensitivity to grouping and model width

The matched encoding controls retain the data-set-specific P, D configurations that performed best among our preceding controlled experiments. These settings are fixed before the raw-versus-atomic comparison, which keeps the original P-point training target and changes only the input parameterization.

Table 1: Parent qualification and local evidence chain under W672 rolling. MSE/MAE (↓) use each source's reported H96/H192/H336/H720 Avg. row for external methods and the corresponding unrounded mean for local runs. † methods are transcribed from Timer-XL Table 12 (Liu et al., 2025b); the local group shows three-seed means over all test origins, with seed standard deviations in table 7. External and local rows are grouped by provenance because their parent-patch protocols differ (table 8); the external block provides context, not a controlled ranking against the local rows. Bold and underline mark the best and second-best distinct metric at displayed precision within each group; displayed ties share a mark.

Models	Author-reported†										Local, three seeds									
	Timer-XL (2025b)		Timer (2024d)		UniTST (2024a)		iTrans. (2024b)		PatchTST (2023)		Parent (Recursive)		Joint-4 (Clean)		Joint-8 (Clean)		ATD-8 (Rollout)		+Tangent-8 (Corrected)	
Dataset	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	**0.409**	**0.430**	0.418	0.436	0.429	0.447	0.421	0.445	0.412	0.435	0.362	0.405	0.409	0.427	0.430	0.439	0.361	0.403	**0.355**	**0.398**
ETTh2	**0.352**	0.402	0.382	0.418	0.384	0.428	0.389	0.421	0.359	**0.400**	0.307	0.371	0.357	0.403	0.348	0.408	0.316	0.376	**0.305**	**0.370**
ETTm1	0.359	**0.382**	0.352	0.383	0.352	0.388	0.376	0.403	**0.349**	0.385	0.346	0.381	0.350	0.391	0.363	0.401	0.342	0.379	**0.339**	**0.375**
ETTm2	0.271	0.322	0.275	0.327	0.265	**0.306**	0.290	0.340	**0.261**	0.318	0.282	0.330	0.273	0.328	0.281	0.332	0.277	0.328	**0.269**	**0.321**
ECL	**0.155**	**0.246**	0.161	0.251	0.163	0.257	0.164	0.258	0.169	0.268	0.153	0.248	0.162	0.262	0.169	0.271	0.153	0.248	**0.152**	**0.247**
Traffic	**0.374**	**0.255**	0.384	0.259	0.385	0.265	0.384	0.274	0.391	0.275	0.384	0.263	0.384	0.263	0.385	0.264	0.386	0.265	**0.382**	**0.261**
Weather	0.240	0.273	0.232	0.270	0.231	0.272	0.266	0.291	**0.226**	**0.268**	0.235	0.272	0.250	0.299	0.281	0.329	**0.234**	0.272	**0.234**	**0.271**
Avg.	**0.309**	**0.330**	0.315	0.335	0.316	0.338	0.327	0.347	0.310	0.336	0.295	0.324	0.312	0.339	0.323	0.349	0.296	0.324	**0.291**	**0.320**

A shared full-rank lift of 12-point atoms to 16 coordinates, followed by chronological concatenation, preserves every raw linear patch projection (proposition 2); all 21 selected local checkpoints satisfy the rank condition. Across five matched data sets, this atomic encoding wins 9/15 seed runs and changes pooled H720 MSE by -0.51%. The mixed signs in table 15 show small, mixed changes after independent training.

The more distinctive observation concerns sensitivity. In the four-data-set $m \times D$ grid, the mean within-group normalized MSE/MAE range from changing $m = P/12$ at fixed D is 2.85%/2.03%, versus 5.08%/4.16% when changing D at fixed m. The larger variation lies along width, whose preferred value differs by data set: ETTm2 and Weather attain their lowest MSE at $D = 64$, ECL and Traffic at $D = 256$. Width therefore warrants attention alongside patch grouping when adapting representations to heterogeneous data. Here m also changes parent span P and D changes backbone capacity; the full grid and interpretation are in appendix B.

4.3 Recursive trajectories can be compiled into wider outputs

The parent trajectory is a substantially easier matched compilation target. On 512 train-only origins per seed, its mean squared magnitude is 0.150× that of the clean future; more importantly, after normalizing each fit by its own target magnitude, the unexplained fraction is 0.109 versus 0.845 for clean futures, lower in all 42/42 seed runs and 168/168 training seed–blocks. An independent TimesFM validation analysis repeats the ordering in 12/12 runs and 48/48 blocks. Together they identify the deployed trajectory as the simpler matched target for the lightweight exits (full evidence in table 9).

One compiled model spans ATD-1 through ATD-8 with only 0.00056 H720 forecast-MSE range in the paired comparison; all prefixes are exact at a common origin and complete ATD-1 rollouts equal the frozen parent in 12 seed runs. ATD-4 and ATD-8 reach 3.22× and 5.54× end-to-end speedup. ATD-1's 0.96× also separates exact fallback from measured runtime. Thus self-trajectory prediction widens the per-call output with nearly constant forecasting performance in this comparison, while the frozen parent transition remains unchanged and exactly available through ATD-1.

4.4 Trajectory fidelity does not order forecast accuracy

The matched targets separate trajectory imitation from forecast accuracy. Table 3 changes only later-patch supervision: ATD improves parent-trajectory fidelity in 21/21 seed runs but forecast MSE in only 15/21. ETTm2 and Traffic are direct counterexamples—every seed becomes more faithful while forecast MSE worsens. Closer imitation can retain the parent's existing forecast errors even with wider outputs: rollout fidelity does not order forecast accuracy. The data-set means, block counts, raw values, and the ETTh2 outlier are retained in table 10.

Table 2: Two-stage deployment accounting. (a) The paired ETTh1, Weather, ECL, and Traffic prefix comparison (three seeds) tests whether one compiled model preserves H720 forecast MSE as commit width changes; ATD-1 bypasses later-patch exits. (b) The seven-data-set × two-width row records the cost of the later macro correction, analyzed in section 4.5. Rollout MSE measures error against the serial parent; speedup is relative to the parent, and retained speedup is relative to ATD without correction.

(a) Selectable prefixes

Mode	Calls	Forecast MSE	Rollout MSE	Latency (ms)	Speedup
ATD-1	15.25	0.3402	0.0000	23.19	0.96×
ATD-2	7.75	0.3405	0.0069	12.89	1.73×
ATD-4	4.00	0.3400	0.0089	6.93	3.22×
ATD-8	2.00	0.3401	0.0112	4.03	**5.54**×

(b) Tangent wrapper timing

Wrapper	Δ latency	Speedup retained	Speedup vs parent
+Tangent	+15.08%	86.89%	3.24×

Table 3: ATD-8 matched-target control at H720. ATD and Frozen Direct share the parent, exit topology, parameter budget, and selector; only supervision changes. Deltas are relative changes from Frozen Direct to ATD, so negative is better. Rollout error is measured against the serial parent. Bold rows improve fidelity but worsen forecast accuracy. Raw MSE values are in table 10.

	Forecast error		Rollout error	
Dataset	ΔMSE (%)	Seed wins	ΔMSE (%)	Seed wins
ETTh1	-8.1	3/3	-85.1	3/3
ETTh2	-81.0	3/3	-99.3	3/3
ETTm1	-1.8	3/3	-70.2	3/3
ETTm2	**+3.5**	0/3	-45.8	3/3
Weather	-2.3	3/3	-69.4	3/3
ECL	-2.3	3/3	-79.1	3/3
Traffic	**+0.9**	0/3	-69.2	3/3

Joint Direct is the shared-backbone clean-future control in table 1. On the four-horizon average, widening it from Joint Direct-4 to Joint Direct-8 improves only 3/21 seeds and 1/7 data-set means, while MSE changes from 0.3121 to 0.3225 (+0.010378). Its mean seed standard deviation grows from 0.0069 to 0.0124, versus 0.0016/0.0017 for ATD-4/8. ATD-8 wins 19/21 seeds and 6/7 means. Thus clean-future widening is less stable on this matched grid (full statistics in table 7).

4.5 Long-horizon residual aligns with a macroscopic history direction

Spectrum Tangent addresses the forecast residual inherited from the parent trajectory. It tests whether a periodic, history-derived direction can reduce this error outside the frozen parent. At H720, MSE falls from 0.371391 to 0.361963 (-2.54%) and MAE from 0.377865 to 0.369045 (-2.33%) across all 14 data-set–width means and 42/42 seeds. Yet only 143/168 chronology blocks improve; the ETTm2 ATD-4 worst case is $+0.025758$.

We use Exchange as a weak-seasonality stress test rather than an eighth positive benchmark cell. The unchanged train-only correction worsens H720 MSE/MAE by +2.82%/+1.22% for Tangent-4 and +4.26%/+1.81% for Tangent-8 (table 21).

Gains strengthen with horizon: H96 has six wins, two losses, and six exact ties among 14 configurations, versus 14/14 improvements at H720. This is consistent with the accumulated-error interpretation (table 4; full counts in table 22).

Table 4: Horizon and cross-parent evidence. (a) Tangent on the seven-data-set grid; W/L counts data-set–width improvements/regressions. (b) Under each parent, subcolumns report ATD H720 forecast-MSE seed wins against the parent and +Tan. ΔMSE (%) against ATD. Negative +Tan. values improve; all are 3/3 seed wins. MSE is not pooled across parents.

(a) Horizon correction

Horizon	Config. W/L	Seed wins	Mean ΔMSE
96	6/2	18/42	-0.0011
192	12/2	36/42	-0.0034
336	13/1	40/42	-0.0061
720	14/0	42/42	**-0.0094**

(b) Cross-parent transfer at H720

	AutoTimes		Timer		TimesFM	
Dataset	ATD Seed wins	+Tangent ΔMSE (%)	ATD Seed wins	+Tangent ΔMSE (%)	ATD Seed wins	+Tangent ΔMSE (%)
ETTh1	0/3	-1.56	3/3	-1.93	3/3	-2.04
ETTh2	2/3	-6.14	0/3	-4.19	3/3	-6.84
ETTm1	3/3	-0.84	3/3	-15.66	3/3	-5.06
Weather	3/3	-3.07	3/3	-5.83	3/3	-12.32

Diagnostic level–shape view. ETTh2 and ECL/Traffic exhibit opposing level–shape signs. The complete direction is positively aligned in 42/42 runs, while neither coordinate alone describes all data sets; full results are in tables 17, 18 and 20.

The wrapper adds no neural parameters or Transformer calls but incurs 15.08% latency overhead; it remains 3.24× faster than recursive parent inference.

4.6 The execution-side phenomena recur across named AR parents

Across AutoTimes–GPT2, Timer-base-84m, and TimesFM-2.5-200M, all 36/36 ATD-1 rollouts recover the parent. Panel (b) of table 4 separates the two ATD-8 stages: ATD improves on the recursive parent in 29/36 seeds; Tangent then improves the matched ATD in 36/36 (-0.84% to -15.66%). Across ATD-2/4/8, Tangent improves 108/108 runs and 377/432 blocks; the ATD-8 target control gives 36/36 fidelity but only 27/36 forecast wins (full results in table 13).

5 Scope and limitations

The scope is patch-based AR point forecasting. Atomic encoding is evaluated with separately trained parents; it does not establish a shared universal encoder. ATD requires teacher trajectories; only ATD-1 is exact. Tangent assumes recurrence: H96 retains regressions, 143/168 H720 blocks improve, and Exchange degrades. Horizon trends do not isolate recursive error from other forecast errors. Timing is hardware-specific and measures batch-one latency; greater GPU utilization at larger batches may reduce the speedup. Offline caching and exit fitting must be amortized over repeated forecasts (appendix A); named-parent evidence establishes compatibility, not throughput.

6 Conclusion

An AR patch often couples representation, training transition, and execution. Atomic encoding shows greater sensitivity to model width than to grouping on the evaluated grid, while parent trajectories support wider ATD execution. Fewer calls can retain the parent's existing errors: closer imitation does not order truth accuracy. Spectrum Tangent corrects an aligned component of the long-horizon residual using periodic history; level and shape are diagnostic views. These findings clarify which patch roles can be separated, while showing that more efficient execution and forecast-error correction remain distinct objectives.

REPRODUCIBILITY STATEMENT

All local configurations use seeds 2021–2023. Parent and decoder checkpoints are validation-selected. Tangent selects its period and coefficient from training origins alone: blocks 1–3 fit and block 4 confirms the frozen choice; test labels never affect selection. We report all test origins in four reporting-only blocks. The appendices document checkpoint lineage, normalization, parameters, calls, timing hardware, selectors, hashes, and negative cases.

STATEMENT ON THE USE OF AI TOOLS

Generative AI tools assisted code inspection, experiment orchestration, plotting, and language editing. The authors designed the method and experiments and verified the implementation, numerical evidence, citations, and claims.

REFERENCES

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

Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. TiRex: Zero-shot forecasting across long and short horizons with enhanced in-context learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/5356603f9c47399adfd372f77a677057-Abstract-Conference.html.

Souhaib Ben Taieb and Rob J. Hyndman. Boosting multi-step autoregressive forecasts. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 109–117, 2014. URL https://proceedings.mlr.press/v32/taieb14.html.

Souhaib Ben Taieb, Gianluca Bontempi, Amir F. Atiya, and Antti Sorjamaa. A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. *Expert Systems with Applications*, 39(8):7067–7083, 2012. doi: 10.1016/j.eswa.2012.01.039.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/e995f98d56967d946471af29d7bf99f1-Abstract.html.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D. Lee, Deming Chen, and Tri Dao. Medusa: Simple LLM inference acceleration framework with multiple decoding heads. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5209–5235, 2024. URL https://proceedings.mlr.press/v235/cai24b.html.

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

Peng Chen, Yingying Zhang, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang, and Chenjuan Guo. Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/2be6705de7412adf107900add727a795-Abstract-Conference.html.

Robert B. Cleveland, William S. Cleveland, Jean E. McRae, and Irma Terpenning. STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1):3–73, 1990. URL https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/stl-a-seasonal-trend-decomposition-procedure-based-on-loess.pdf.

Abhimanyu Das, Weihao Kong, Andrew Leach, Shaan Mathur, Rajat Sen, and Rose Yu. Long-term forecasting with TiDE: Time-series dense encoder. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=pCbC3aQB5W.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10148–10167, 2024. URL https://proceedings.mlr.press/v235/das24c.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.11929.

Da Fu, Yuan Zuo, Yuan Shi, Hao Lin, Yansong Liu, and Junjie Wu. REDNet: Two-stage knowledge distillation for robust edge time series forecasting. *Pattern Recognition Letters*, 208:258–264, 2026. doi: 10.1016/j.patrec.2026.08.011.

Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Roziere, David Lopez-Paz, and Gabriel Synnaeve. Better & faster large language models via multi-token prediction. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15706–15734, 2024. URL https://proceedings.mlr.press/v235/gloeckle24a.html.

Suhan Guo, Bingxu Wang, Shaodan Zhang, and Furao Shen. T-LLM: Teaching large language models to forecast time series via temporal distillation. *arXiv preprint arXiv:2602.01937*, 2026. URL https://arxiv.org/abs/2602.01937.

Lanxiang Hu, Siqi Kou, Yichao Fu, Samyam Rajbhandari, Tajana Rosing, Yuxiong He, Zhijie Deng, and Hao Zhang. Fast and accurate causal parallel decoding using jacobi forcing. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://openreview.net/forum?id=ORy2hyPd4G.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19274–19286, 2023. URL https://proceedings.mlr.press/v202/leviathan23a.html.

Phillip Lippe, Bastiaan S. Veeling, Paris Perdikaris, Richard E. Turner, and Johannes Brandstetter. PDE-refiner: Achieving accurate long rollouts with neural PDE solvers. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d529b943af3dba734f8a7d49efcb6d09-Abstract-Conference.html.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. Moirai 2.0: When less is more for time series forecasting. *arXiv preprint arXiv:2511.11698*, 2025a. URL https://arxiv.org/abs/2511.11698.

Hao Liu, Xiaoxing Zhang, Chun Yang, and Xiaobin Zhu. Semantic-enhanced time-series forecasting via large language models. In *International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a5e2853be2ff6218a8d87a5ce4b65062-Abstract-Conference.html.

Juncheng Liu, Chenghao Liu, Gerald Woo, Yiwei Wang, Bryan Hooi, Caiming Xiong, and Doyen Sahoo. UniTST: Effectively modeling inter-series and intra-series dependencies for multivariate time series forecasting. *arXiv preprint arXiv:2406.04975*, 2024a. URL https://arxiv.org/abs/2406.04975.

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, 2022. URL https://papers.neurips.cc/paper_files/paper/2022/hash/4054556fcaa934b0bf76da52cf4f92cb-Abstract-Conference.html.

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*, 2024b. URL https://arxiv.org/abs/2310.06625.

Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. AutoTimes: Autoregressive time series forecasters via large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024c. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/dcf88cbc8d01ce7309b83d0ebaeb9d29-Abstract-Conference.html.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32369–32399, 2024d. URL https://proceedings.mlr.press/v235/liu24cb.html.

Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer-XL: Long-context transformers for unified time series forecasting. In *International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=KMCJXjlDDr.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39295–39317, 2025c. URL https://proceedings.mlr.press/v267/liu25be.html.

Yong Liu, Xingjian Su, Shiyu Wang, Haoran Zhang, Haixuan Liu, Yuxuan Wang, Zhou Ye, Yang Xiang, Jianmin Wang, and Mingsheng Long. Timer-S1: A billion-scale time series foundation model with serial scaling. *arXiv preprint arXiv:2603.04791*, 2026b. URL https://arxiv.org/abs/2603.04791.

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.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

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. URL https://doi.org/10.1016/j.ijforecast.2019.07.001.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-MoE: Billion-scale time series foundation models with mixture of experts. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2409.16040.

Mitchell Stern, Noam Shazeer, and Jakob Uszkoreit. Blockwise parallel decoding for deep autoregressive models. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/c4127b9194fe8562c64dc0f5bf2c93bc-Abstract.html.

Pranav Subbaraman, Fang Sun, Jinxi Yu, Yue Yao, Huacong Tang, Xiao Luo, and Yizhou Sun. Accelerating time series foundation models with speculative decoding. *arXiv preprint arXiv:2511.18191*, 2025. URL https://arxiv.org/abs/2511.18191v2.

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://proceedings.iclr.cc/paper_files/paper/2024/hash/a7ac8a21e5a27e7ab31a5f42a0117bdb-Abstract-Conference.html.

Peter R. Winters. Forecasting sales by exponentially weighted moving averages. *Management Science*, 6(3):324–342, 1960. doi: 10.1287/mnsc.6.3.324.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53140–53164, 2024. URL https://proceedings.mlr.press/v235/woo24a.html.

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. URL https://proceedings.neurips.cc/paper/2021/hash/bcc0d400288793e8bdcd7c19a8ac0c2b-Abstract.html.

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://arxiv.org/abs/2210.02186.

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. URL https://arxiv.org/abs/2205.13504.

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. URL https://arxiv.org/abs/2012.07436.

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.

A Experimental details and complete result tables

A.1 Complete four-horizon forecasting results

The reported entries below are transcribed entry-for-entry from Timer-XL Table 12 (Liu et al., 2025b). It covers Timer, UniTST, iTransformer, DLinear, PatchTST, TimesNet, Non-stationary Transformer, and Autoformer (Liu et al., 2024d;a;b; Zeng et al., 2023; Nie et al., 2023; Wu et al., 2023; Liu et al., 2022; Wu et al., 2021). These are author-reported comparisons, not local reruns. We therefore keep their P96-output protocol separate from our data-set-specific-P local group even though both use W672 rolling forecasts to H96/H192/H336/H720.

Joint Direct appends learned future placeholders to the observed token sequence and jointly updates them with the parent-initialized backbone against clean future patches. It is Moirai-2.0-style in its multi-token objective (Liu et al., 2025a), but is a point-MSE control rather than a reproduction of Moirai 2.0's probabilistic architecture. On the four-horizon average, the wider Joint Direct-8 control has larger average seed dispersion than Joint Direct-4 and improves only 3/21 seed pairs, whereas ATD's frozen backbone keeps dispersion substantially smaller.

Table 5: Complete reported MSE/MAE entries transcribed from Timer-XL Table 12. One source model per data set is trained at W672/P96 and rolled to four horizons; Avg. reproduces the source table's printed Avg. row rather than re-averaging its rounded entries. Bold/underline mark the best/second-best distinct reported metric at displayed precision in each row; ties share a mark.

Models	Timer-XL (2025b)		Timer (2024d)		UniTST (2024a)		iTrans. (2024b)		DLinear (2023)		PatchTST (2023)		TimesNet (2023)		Nonstat. (2022)		Autoformer (2021)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1																		
96	**0.364**	**0.397**	0.371	0.404	0.379	0.415	0.387	0.418	0.369	0.400	0.373	0.403	0.452	0.463	0.452	0.478	0.467	0.499
192	**0.405**	0.424	0.407	0.429	0.415	0.438	0.416	0.437	**0.405**	**0.422**	**0.405**	0.425	0.474	0.477	0.484	0.510	0.492	0.523
336	0.427	**0.439**	0.434	0.445	0.440	0.454	0.434	0.450	0.435	0.445	**0.423**	0.440	0.493	0.489	0.511	0.522	0.519	0.531
720	**0.439**	**0.459**	0.461	0.466	0.482	0.482	0.447	0.473	0.493	0.508	0.445	0.471	0.560	0.534	0.571	0.543	0.589	0.560
Avg.	**0.409**	**0.430**	0.418	0.436	0.429	0.447	0.421	0.445	0.426	0.444	0.412	0.435	0.495	0.491	0.505	0.513	0.517	0.528
ETTh2																		
96	**0.277**	**0.343**	0.285	0.344	0.343	0.398	0.304	0.362	0.305	0.371	0.289	0.347	0.340	0.374	0.348	0.403	0.358	0.397
192	**0.348**	**0.391**	0.365	0.400	0.376	0.420	0.372	0.407	0.412	0.439	0.360	0.393	0.402	0.414	0.408	0.448	0.435	0.451
336	**0.375**	**0.418**	0.412	0.440	0.399	0.435	0.418	0.440	0.527	0.508	0.389	0.420	0.452	0.452	0.424	0.457	0.454	0.475
720	0.409	0.458	0.468	0.487	0.419	0.457	0.463	0.476	0.830	0.653	**0.398**	**0.440**	0.462	0.468	0.448	0.476	0.479	0.492
Avg.	**0.352**	0.402	0.382	0.418	0.384	0.428	0.389	0.421	0.518	0.493	0.359	**0.400**	0.414	0.427	0.407	0.446	0.431	0.454
ETTm1																		
96	0.290	0.341	**0.281**	**0.338**	0.289	0.348	0.311	0.365	0.307	0.350	0.285	0.346	0.338	0.375	0.414	0.414	0.466	0.466
192	0.337	0.369	0.330	**0.368**	0.332	0.375	0.353	0.390	0.337	**0.368**	**0.329**	0.372	0.371	0.387	0.524	0.482	0.504	0.496
336	0.374	0.392	0.367	0.393	0.365	0.397	0.387	0.411	0.366	**0.387**	**0.363**	0.394	0.410	0.411	0.541	0.497	0.574	0.530
720	0.437	0.428	0.432	0.433	0.421	0.431	0.452	0.445	**0.419**	**0.419**	0.421	0.426	0.478	0.450	0.578	0.509	0.596	0.558
Avg.	0.359	0.382	0.352	0.383	0.352	0.388	0.376	0.403	0.357	**0.381**	**0.349**	0.385	0.399	0.406	0.514	0.475	0.535	0.512
ETTm2																		
96	0.175	**0.257**	0.175	**0.257**	0.171	0.260	0.183	0.272	**0.167**	0.263	0.172	0.259	0.187	0.267	0.237	0.306	0.255	0.339
192	0.242	0.301	0.239	0.301	**0.228**	**0.230**	0.250	0.315	0.230	0.311	0.233	0.299	0.249	0.309	0.330	0.387	0.279	0.335
336	0.293	0.337	0.293	0.342	0.282	0.336	0.311	0.356	0.298	0.361	**0.280**	**0.331**	0.321	0.351	0.404	0.424	0.331	0.374
720	0.376	0.390	0.392	0.407	0.380	0.398	0.417	0.419	0.432	0.446	**0.357**	**0.382**	0.497	0.403	0.525	0.486	0.413	0.450
Avg.	0.271	0.322	0.275	0.327	0.265	**0.306**	0.290	0.340	0.282	0.345	**0.261**	0.318	0.314	0.333	0.374	0.401	0.320	0.374
ECL																		
96	**0.127**	**0.219**	0.129	0.221	0.130	0.225	0.133	0.229	0.138	0.238	0.132	0.232	0.184	0.288	0.185	0.287	0.256	0.357
192	**0.145**	**0.236**	0.148	0.239	0.150	0.244	0.158	0.258	0.152	0.251	0.151	0.250	0.192	0.295	0.282	0.368	0.291	0.376
336	**0.159**	**0.252**	0.164	0.256	0.166	0.262	0.168	0.262	0.167	0.268	0.171	0.272	0.200	0.303	0.289	0.377	0.290	0.379
720	**0.187**	**0.277**	0.201	0.289	0.206	0.297	0.205	0.294	0.203	0.302	0.222	0.318	0.228	0.325	0.305	0.399	0.320	0.403
Avg.	**0.155**	**0.246**	0.161	0.251	0.163	0.257	0.164	0.258	0.165	0.265	0.169	0.268	0.201	0.303	0.265	0.358	0.289	0.379
Traffic																		
96	**0.340**	**0.238**	0.348	0.240	0.359	0.250	0.353	0.259	0.399	0.285	0.359	0.255	0.593	0.315	0.610	0.322	0.675	0.412
192	**0.360**	**0.247**	0.369	0.250	0.373	0.257	0.373	0.267	0.409	0.290	0.377	0.265	0.596	0.317	0.626	0.346	0.679	0.423
336	**0.377**	**0.256**	0.388	0.260	0.386	0.265	0.386	0.275	0.422	0.297	0.393	0.276	0.600	0.319	0.633	0.352	0.688	0.440
720	**0.418**	**0.279**	0.431	0.285	0.421	0.286	0.425	0.296	0.461	0.319	0.436	0.305	0.619	0.335	0.651	0.366	0.693	0.457
Avg.	**0.374**	**0.255**	0.384	0.259	0.385	0.265	0.384	0.274	0.423	0.298	0.391	0.275	0.602	0.322	0.630	0.347	0.684	0.433
Weather																		
96	0.157	0.205	0.151	**0.202**	0.152	0.206	0.174	0.225	0.169	0.229	**0.149**	**0.202**	0.169	0.228	0.185	0.241	0.355	0.409
192	0.206	0.250	0.196	**0.245**	0.198	0.249	0.227	0.268	0.211	0.268	**0.194**	**0.245**	0.222	0.269	0.286	0.325	0.421	0.450
336	0.259	0.291	0.249	0.288	0.251	0.291	0.290	0.309	0.258	0.306	**0.244**	**0.285**	0.290	0.310	0.323	0.347	0.452	0.465
720	0.337	0.344	0.330	0.344	0.322	0.340	0.374	0.360	0.320	0.362	**0.317**	**0.338**	0.376	0.364	0.436	0.401	0.513	0.496
Avg.	0.240	0.273	0.232	0.270	0.231	0.272	0.266	0.291	0.239	0.291	**0.226**	**0.268**	0.264	0.293	0.308	0.329	0.435	0.455

Table 6: Complete local MSE/MAE entries. Every entry averages all test origins and seeds 2021–2023; Avg. is the four-horizon mean, and bold/underline mark the best/second-best distinct local metric at displayed precision in each row; ties share a mark. Joint Direct uses the same parent initialization and closed-loop H720 selector; ATD and Tangent use the frozen selections reported in table 8.

Models	Parent		Joint-4		ATD-4		+Tangent-4		Joint-8		ATD-8		+Tangent-8	
Horizon	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1														
96	0.326	0.375	0.349	0.386	0.325	0.374	**0.321**	**0.370**	0.360	0.393	0.325	0.374	0.322	0.371
192	0.350	0.393	0.385	0.409	0.348	0.392	**0.342**	**0.386**	0.400	0.417	0.350	0.392	0.345	0.387
336	0.363	0.408	0.414	0.430	0.361	0.406	**0.352**	**0.397**	0.436	0.443	0.362	0.405	0.354	0.398
720	0.408	0.443	0.487	0.483	0.405	0.440	**0.397**	**0.433**	0.522	0.502	0.405	0.439	0.398	0.434
Avg.	0.362	0.405	0.409	0.427	0.360	0.403	**0.353**	**0.396**	0.430	0.439	0.361	0.403	0.355	0.398
ETTh2														
96	0.235	0.312	0.296	0.364	0.235	0.312	**0.233**	**0.311**	0.296	0.372	0.235	0.312	**0.233**	0.312
192	0.279	0.348	0.321	0.382	0.282	0.351	**0.276**	**0.347**	0.321	0.390	0.282	0.350	0.277	0.348
336	0.313	0.380	0.348	0.402	0.322	0.386	**0.308**	**0.376**	0.343	0.407	0.326	0.388	0.314	0.380
720	0.402	0.445	0.463	0.465	0.419	0.455	**0.392**	**0.435**	0.433	0.461	0.421	0.455	0.397	0.439
Avg.	0.307	0.371	0.357	0.403	0.314	0.376	**0.302**	**0.367**	0.348	0.408	0.316	0.376	0.305	0.370
ETTm1														
96	**0.272**	**0.339**	0.297	0.360	**0.272**	**0.339**	**0.272**	**0.339**	0.319	0.376	**0.272**	**0.339**	**0.272**	**0.339**
192	0.321	0.368	0.326	0.378	0.320	0.367	0.317	**0.364**	0.342	0.391	0.319	0.366	**0.316**	**0.364**
336	0.363	0.391	0.360	0.396	0.360	0.390	**0.355**	**0.384**	0.371	0.406	0.359	0.389	**0.355**	**0.384**
720	0.426	0.426	0.417	0.428	0.421	0.424	0.415	0.415	0.422	0.432	0.419	0.422	**0.412**	**0.413**
Avg.	0.346	0.381	0.350	0.391	0.343	0.380	0.340	**0.375**	0.363	0.401	0.342	0.379	**0.339**	**0.375**
ETTm2														
96	0.182	**0.263**	**0.181**	0.265	0.184	0.266	**0.181**	0.264	0.190	0.270	0.184	0.266	**0.181**	0.264
192	0.245	0.306	0.242	0.306	0.244	0.306	**0.237**	**0.302**	0.252	0.311	0.243	0.305	**0.237**	**0.302**
336	0.300	0.343	0.293	0.342	0.297	0.342	0.287	0.334	0.303	0.346	0.296	0.340	**0.286**	**0.333**
720	0.399	0.409	0.374	0.398	0.388	0.401	0.376	0.387	0.381	0.400	0.387	0.399	**0.373**	**0.386**
Avg.	0.282	0.330	0.273	0.328	0.278	0.329	0.270	0.322	0.281	0.332	0.277	0.328	**0.269**	**0.321**
ECL														
96	**0.116**	**0.216**	0.129	0.234	**0.116**	**0.216**	**0.116**	**0.216**	0.137	0.245	**0.116**	**0.216**	**0.116**	**0.216**
192	**0.136**	0.234	0.146	0.248	0.137	0.234	**0.136**	**0.233**	0.153	0.258	0.137	0.234	**0.136**	**0.233**
336	0.157	0.253	0.165	0.265	0.157	0.253	**0.155**	**0.252**	0.172	0.274	0.157	0.253	0.156	**0.252**
720	0.201	0.289	0.209	0.301	0.202	0.290	**0.199**	**0.287**	0.213	0.305	0.202	0.290	**0.199**	**0.287**
Avg.	0.153	0.248	0.162	0.262	0.153	0.248	**0.152**	**0.247**	0.169	0.271	0.153	0.248	**0.152**	**0.247**
Traffic														
96	**0.348**	**0.242**	0.353	0.247	**0.348**	**0.242**	**0.348**	**0.242**	0.355	0.249	**0.348**	**0.242**	**0.348**	**0.242**
192	**0.366**	**0.252**	0.367	0.254	0.369	0.254	**0.366**	**0.252**	0.369	0.255	0.369	0.255	**0.366**	**0.252**
336	0.386	0.265	0.385	0.263	0.388	0.267	**0.383**	**0.262**	0.385	0.265	0.388	0.267	0.384	0.263
720	0.436	0.293	0.431	0.289	0.438	0.295	**0.430**	**0.286**	0.431	0.288	0.437	0.295	0.431	0.287
Avg.	0.384	0.263	0.384	0.263	0.386	0.265	**0.382**	**0.260**	0.385	0.264	0.386	0.265	**0.382**	0.261
Weather														
96	**0.154**	**0.204**	0.173	0.238	0.156	0.206	0.156	0.206	0.208	0.277	0.156	0.206	0.156	0.206
192	0.201	0.250	0.215	0.276	**0.200**	**0.249**	0.201	0.250	0.246	0.307	0.201	0.250	0.201	0.250
336	0.254	0.290	0.269	0.316	**0.252**	0.290	**0.252**	**0.289**	0.292	0.339	**0.252**	0.290	**0.252**	0.290
720	0.330	0.342	0.344	0.364	0.328	0.342	**0.324**	**0.339**	0.380	0.393	0.328	0.342	0.325	**0.339**
Avg.	0.235	0.272	0.250	0.299	0.234	0.272	**0.233**	**0.271**	0.281	0.329	0.234	0.272	0.234	**0.271**

Table 7: Three-seed four-horizon stability of the clean-future joint predictor and ATD. For each seed, MSE/MAE is first averaged over H96/H192/H336/H720; entries then report mean±sample standard deviation over seeds 2021–2023. The Avg. column is the equal-data-set mean and, after ±, the mean of the seven data-set-wise standard deviations; it is not a standard deviation across heterogeneous data sets.

Method	Metric	ETTh1	ETTh2	ETTm1	ETTm2
Joint Direct-4	MSE	0.4090 ± 0.0162	0.3571 ± 0.0108	0.3499 ± 0.0002	0.2728 ± 0.0082
	MAE	0.4270 ± 0.0097	0.4032 ± 0.0015	0.3906 ± 0.0019	0.3277 ± 0.0045
Joint Direct-8	MSE	0.4295 ± 0.0291	0.3483 ± 0.0124	0.3635 ± 0.0015	0.2812 ± 0.0079
	MAE	0.4389 ± 0.0123	0.4076 ± 0.0117	0.4010 ± 0.0015	0.3319 ± 0.0048
ATD-4	MSE	0.3598 ± 0.0013	0.3144 ± 0.0056	0.3432 ± 0.0001	0.2785 ± 0.0006
	MAE	0.4029 ± 0.0008	0.3760 ± 0.0036	0.3799 ± 0.0003	0.3288 ± 0.0022
ATD-8	MSE	0.3606 ± 0.0016	0.3161 ± 0.0048	0.3421 ± 0.0002	0.2773 ± 0.0012
	MAE	0.4028 ± 0.0013	0.3765 ± 0.0028	0.3790 ± 0.0005	0.3277 ± 0.0026

Method	Metric	ECL	Traffic	Weather	Avg.
Joint Direct-4	MSE	0.1622 ± 0.0021	0.3840 ± 0.0050	0.2501 ± 0.0062	0.3121 ± 0.0069
	MAE	0.2621 ± 0.0010	0.2635 ± 0.0011	0.2987 ± 0.0025	0.3390 ± 0.0032
Joint Direct-8	MSE	0.1687 ± 0.0010	0.3849 ± 0.0073	0.2814 ± 0.0277	0.3225 ± 0.0124
	MAE	0.2706 ± 0.0012	0.2643 ± 0.0026	0.3289 ± 0.0270	0.3490 ± 0.0087
ATD-4	MSE	0.1530 ± 0.0002	0.3859 ± 0.0014	0.2340 ± 0.0023	0.2955 ± 0.0016
	MAE	0.2483 ± 0.0002	0.2645 ± 0.0009	0.2719 ± 0.0023	0.3246 ± 0.0015
ATD-8	MSE	0.1529 ± 0.0003	0.3856 ± 0.0013	0.2342 ± 0.0023	0.2955 ± 0.0017
	MAE	0.2482 ± 0.0004	0.2646 ± 0.0008	0.2719 ± 0.0023	0.3244 ± 0.0015

Table 8: Data-set-specific parent and tangent hyperparameters. P, D, and L are parent patch length, model width, and Transformer depth. For each K, τ^* and α_K^* are fit after pooling seeds 2021–2023 on training blocks 1–3 and are shared by their three runs. The separately selected local ATD-4/8 periods coincide, so panel (a) shows τ^* once. Panel (b) supplies all named-parent ATD-2/4/8 selections used in tables 13 and 19, including α_2^*. No local ATD-2+Tangent result is reported.

(a) Local parents

Dataset	P	D	L	τ^*	α_4^*	α_8^*
ETTh1	24	64	1	24	0.4815	0.3997
ETTh2	48	64	1	24	0.2828	0.2141
ETTm1	96	64	1	96	0.1827	0.1592
ETTm2	24	64	2	96	0.3977	0.3459
Weather	48	64	2	12	0.0851	0.0856
ECL	96	256	2	168	0.1133	0.0907
Traffic	96	256	2	168	0.1235	0.1114

(b) Named-parent tangent selections

Parent	Dataset	τ_2^*	α_2^*	τ_4^*	α_4^*	τ_8^*	α_8^*
AutoTimes	ETTh1	24	0.1799	24	0.1152	36	0.0424
AutoTimes	ETTh2	24	0.5033	24	0.3760	24	0.2347
AutoTimes	ETTm1	96	0.2341	96	0.2051	96	0.1903
AutoTimes	Weather	144	0.3473	144	0.3104	144	0.2777
Timer	ETTh1	24	0.1626	24	0.1502	24	0.1596
Timer	ETTh2	24	0.2413	24	0.2423	24	0.2599
Timer	ETTm1	96	0.3327	96	0.3050	96	0.3081
Timer	Weather	144	0.3037	144	0.2799	144	0.2839
TimesFM	ETTh1	24	0.1917	24	0.1818	24	0.1935
TimesFM	ETTh2	24	0.2389	24	0.2236	24	0.2302
TimesFM	ETTm1	96	0.2931	96	0.3018	96	0.3123
TimesFM	Weather	144	0.2993	144	0.3015	144	0.3107

A.2 Data-varying configurations

All local runs use $W = 672$, training-set global standardization, seeds 2021–2023, eight attention heads, and feed-forward width $4D$. Parent checkpoints use next-patch validation; ATD-4/8 checkpoints use closed-loop H720 validation. Table 8 records every data-varying parent configuration and the frozen tangent selection shared by all three seeds in each data-set–width configuration. Each 12-point atom is lifted to 16 coordinates; α^* is the effective coefficient applied in equation (11). The candidate period grid is the common set of 12-sample multiples through 216, the largest multiple represented at least three times in $W = 672$. The global $b(\ell) = (0.25, 0.5, 1, 2, 3)$ ramp at boundaries $(96, 192, 336, 672)$ is fixed before moment fitting and is unchanged across data sets, seeds, and K. The grid and ramp are design constants, not data-varying selected hyperparameters.

Local ATD uses channel-independent, zero-initialized two-layer MLP exits while the parent remains in evaluation mode. Local instance normalization uses $\mu_c = W^{-1}\sum_t h_{c,t}$ and $\sigma_c = \sqrt{W^{-1}\sum_t (h_{c,t} - \mu_c)^2 + 0.001}$; a bar denotes this frame, distinct from training-set global standardization. Across all 42 ATD/Direct seed runs, complete ATD-1 differences are zero and full-origin H720 MSE matches the frozen evaluation records within 1.46×10^{-10}.

The 42 local ATD-4/8 post-training jobs took 0.17–5.38 minutes each in the campaign logs, including process startup, teacher-state/target caching, exit fitting, and validation, but excluding parent training and final testing. These are logged job wall times, not isolated GPU-time measurements. Deployment savings amortize this one-time cost only after roughly $N > C_{\text{post}}/(t_{\text{parent}} - t_{\text{ATD}})$ forecasts, when the denominator is positive and all times refer to a matched hardware/batch configuration; we do not infer a break-even count by mixing unmatched timing records.

A.3 Named-parent state extraction

For each named parent, $s_0 \in \mathbb{R}^{C \times D_{\text{parent}}}$ contains one final-token hidden vector per channel at the current ATD call origin. The implementation folds channels into the batch axis; exit weights are shared across channels, without cross-variable mixing.

AutoTimes-GPT2 ($D_{\text{parent}} = 768$). We select `last_hidden_state[:, -1]` from GPT2: the last observed 96-point token after the final Transformer block and GPT2's final `ln_f`, before the AutoTimes MLP decoder.

Timer-base-84m ($D_{\text{parent}} = 1024$). We select `hidden_states[-1][:, -1]` from Timer's prediction output: the last consumed 96-point token after the final decoder block and the stack's final LayerNorm, before `lm_head`.

TimesFM-2.5-200M ($D_{\text{parent}} = 1280$). We select `output_embeddings[:, -1]` from the native forward pass: the last consumed 32-point input token after the final Transformer block, before the point and quantile output projections. There is no additional stack-final normalization. This hidden vector feeds exits predicting 128-point output blocks; it is not the forecast or a pooled output block.

All three readouts enter the residual exit MLPs directly, without additional state normalization, projection, or temporal/layer pooling. The frozen parent remains in evaluation mode. AutoTimes recomputes the rolling-window encoding at each call; Timer and TimesFM retain their native KV caches and refresh the readout after consuming the committed patches. Thus s_0 is call-local, not a vector held fixed throughout the forecast. KV caches and normalization statistics remain part of parent execution, not extra exit inputs.

A.4 Additional analyses

Table 9: Matched-target fitting audits. The local parent uses train-only origins over seven data sets and ATD-4/8; TimesFM uses a shared validation cache over four data sets at ATD-8. Each has three seeds. Energy ratio is parent trajectory over the clean future; unexplained fraction is fitted MSE divided by each target's own energy. Runs/seed–blocks count a lower fraction for the parent trajectory.

	Energy	Unexplained		Parent easier	
Parent / split	Parent/future	Parent	Future	Runs	Seed–blocks
Local parent / train	0.150×	**0.109**	0.845	42/42	168/168
TimesFM / val.	0.139×	**0.126**	0.861	12/12	48/48

Table 10: Raw values underlying table 3. Entries average three seeds; ATD and Frozen Direct differ only in the later-patch target.

	Forecast MSE		Parent-traj. MSE		Wins	
Data	Direct	ATD	Direct	ATD	Forecast	Fidel.
ETTh1	0.4410	0.4052	0.0661	0.0098	3/3	3/3
ETTh2	2.3903	0.4208	1.7784	0.0104	3/3	3/3
ETTm1	0.4261	0.4186	0.0488	0.0145	3/3	3/3
ETTm2	0.3736	0.3865	0.0613	0.0332	0/3	3/3
Weather	0.3356	0.3278	0.0385	0.0116	3/3	3/3
ECL	0.2063	0.2017	0.0187	0.0038	3/3	3/3
Traffic	0.4328	0.4368	0.0305	0.0094	0/3	3/3

The absence of a STRIDE row in the matched result tables is a protocol boundary rather than an omitted nearest baseline. Its Timer-XL operating point couples a P96/context-1536 target, a 0.125× draft, an acceptance temperature, and acceptance-dependent target verification. Our local system instead times one W672 compiled model with data-specific P and fixed commit widths. A controlled systems comparison would need to hold fixed the accepted parent checkpoint, draft construction, context and patch sizes, batch size, hardware, accuracy gate, and all draft, verification, fallback, and wrapper work.

Table 11: Nearby parallel-generation designs. "Runtime target" means invocation of the accepted autoregressive model after proposal. STRIDE keeps the target frozen but runs a separate draft–verify path online.

Method	Parallel signal	Post-train	Runtime target
Moirai 2.0	clean future patches	joint	no
STRIDE	separate-draft proposals	no	verify
Jacobi Forcing	Jacobi generation trajectory	yes	refine
ATD	serial deployed parent trajectory	exits	no

The four data sets are the common frozen compatibility grid across all three named-parent campaigns: ETTh1, ETTh2, ETTm1, and Weather. This study changes the parent architecture rather than repeating the seven-data-set local benchmark. ETTm2, ECL, and Traffic have no named-parent results in this grid; the transfer claim is limited to the four evaluated data sets.

Table 12: Closest acceleration reference without cross-protocol ranking. STRIDE† entries are author-reported Timer-XL H720 points from arXiv v2, Table 1 (Subbaraman et al., 2025): batch one, K3, acceptance temperature 0.25, and GPU-compute timing on one H200. Local speed uses batch-one RTX5880-Ada-48Q timing on the four-data-set paired timing subset; local accuracy averages seven data sets and three seeds. Thus rows expose the operating contracts but are not a controlled speed leaderboard.

Method	Reported/local scope	Target online?	Accuracy report	Speedup
STRIDE†	Timer-XL, ECL, H720, B1, width 3	verify	MSE 0.200→0.240	1.82×
STRIDE†	Timer-XL, Weather, H720, B1, width 3	verify	MSE 0.320→0.362	1.42×
ATD-8	local H720, B1	no	7-data MSE 0.3720→0.3711	**5.54**×
+Tangent	local H720, B1	no	7-data MSE 0.3711→0.3620	3.24×

Table 13: Transfer across three named AR parents and four data sets. (a) ATD-8 H720 counts are seeds out of three; Tangent delta is relative to the matched ATD model, and Frozen Direct changes only the later-patch target. ATD-1 exact denotes bit-exact complete-rollout recovery (36/36 overall). (b) Test-block wins across ATD-2/4/8; each parent–width entry contains 48 blocks, and each width total contains 144 blocks. MSE is never pooled across parents.

(a) ATD-8 H720 seed counts

			ATD/parent	Tangent		ATD/Direct	
Parent	Data	ATD-1 exact	Forecast wins	Δ (%)	Wins	Forecast wins	Fidel. wins
AutoTimes	ETTh1	3/3	0/3	-1.56	3/3	3/3	3/3
	ETTh2	3/3	2/3	-6.14	3/3	3/3	3/3
	ETTm1	3/3	3/3	-0.84	3/3	3/3	3/3
	Weather	3/3	3/3	-3.07	3/3	3/3	3/3
Timer	ETTh1	3/3	3/3	-1.93	3/3	3/3	3/3
	ETTh2	3/3	0/3	-4.19	3/3	3/3	3/3
	ETTm1	3/3	3/3	-15.66	3/3	0/3	3/3
	Weather	3/3	3/3	-5.83	3/3	0/3	3/3
TimesFM	ETTh1	3/3	3/3	-2.04	3/3	3/3	3/3
	ETTh2	3/3	3/3	-6.84	3/3	3/3	3/3
	ETTm1	3/3	3/3	-5.06	3/3	3/3	3/3
	Weather	3/3	3/3	-12.32	3/3	0/3	3/3

(b) Multiwidth test-block wins

ATD width	AutoTimes	Timer	TimesFM	**Total**
ATD-2	38/48	33/48	48/48	**119/144**
ATD-4	40/48	33/48	48/48	**121/144**
ATD-8	44/48	45/48	48/48	**137/144**

With ATD-8, compilation improves over the recursive parent forecast MSE in 29/36 parent–data–seed runs; the matched target control favors ATD in 36/36 fidelity but only 27/36 forecast comparisons. Tangent then improves ATD in 36/36 runs; complete values are in table 13.

B ATOMIC ENCODING: CONSTRUCTION AND SENSITIVITY

We fix $p_0 = 12$ because it is a common divisor of every studied parent span $P \in \{24, 48, 96\}$; on hourly data it also corresponds to half a day, but it has no universal physical-period meaning across the mixed sampling rates. The atom width 16 is likewise a fixed overcomplete lift that permits full column rank, not a data-set-specific semantic choice or a separately tuned module. In the tables, P12–16 denotes this 12-point atom and 16-coordinate lift.

For $m = P/12$ chronological atoms $u_q \in \mathbb{R}^{12}$, each parent learns one map ϕ shared across atom positions and uses

$$z_P = \psi_P[\phi(u_1)\| \cdots \|\phi(u_m)], \qquad \phi : \mathbb{R}^{12} \to \mathbb{R}^{16}, \quad \psi_P : \mathbb{R}^{16m} \to \mathbb{R}^{d_{model}}. \tag{12}$$

The following proposition records the class-preservation property used by the encoding-side control.

Proposition 2 (Information-preserving atomic factorization)**.** *If ϕ has full column rank, the concatenated atomic encoding is injective. Moreover, for every raw linear patch projection $R : \mathbb{R}^{12m} \to \mathbb{R}^{d_{model}}$ there exists a ψ_P such that equation* (12) *equals R for every input.*

Proof. Full column rank guarantees a left inverse $L \in \mathbb{R}^{12\times 16}$ satisfying $L\phi = I_{12}$. Let $L_m = \mathrm{diag}(L, \ldots, L)$ and choose $\psi_P = RL_m$. Then $\psi_P \,\mathrm{diag}(\phi, \ldots, \phi) = R$. □

Thus each ordered raw patch can be recovered from its atomic encoding, and chronological concatenation preserves the raw linear function class. This establishes class preservation; optimization equivalence is evaluated empirically rather than guaranteed by the construction.

This injectivity statement extends algebraically to multivariate patches: applying the lift independently to C variables gives a block-diagonal map with left inverse $I_C \otimes L_m$. Thus any linear cross-variable patch projection can also be represented with a suitable subsequent joint map. Our implemented parents instead share a per-variable projection and have no cross-variable mixing. The sensitivity results therefore do not establish the same empirical behavior for cross-variable models, nor does a later width-reducing projection necessarily preserve all input information.

Table 14: Atomic input interfaces and a direct checkpoint audit of the sufficient condition in proposition 2. The final column counts selected seed checkpoints whose learned 16×12 atom lift has full column rank. This is a structural audit, not an accuracy comparison or an exhaustive search over atom sizes.

Dataset	P12–16 interface			Parent	Full rank
	P	$m = P/12$	concat. $16m$	D	$\mathrm{rank}(\phi) = 12$
ETTh1	24	2	32	64	3/3
ETTh2	48	4	64	64	3/3
ETTm1	96	8	128	64	3/3
ETTm2	24	2	32	64	3/3
Weather	48	4	64	64	3/3
ECL	96	8	128	256	3/3
Traffic	96	8	128	256	3/3

Table 15: Matched H720 representation control. Raw and P12–16 entries are three-seed mean $\pm$ sample standard deviation of MSE; the final column averages seed-paired relative changes (negative favors P12–16). ETTm2* uses the stricter same-draw, exact-transplant initialization; both arms are then trained independently. Other rows retain matched architectures with independent draws.

Dataset	Architecture		H720 MSE		Paired Δ
	P	D	Raw	P12–16	(%)
ETTh1	24	64	$\mathbf{0.4079}\pm0.0037$	0.4084 ± 0.0012	+0.15
ETTh2	48	64	0.4174 ± 0.0138	$\mathbf{0.4025}\pm0.0087$	-3.46
ETTm1	96	64	0.4292 ± 0.0012	$\mathbf{0.4264}\pm0.0011$	-0.64
ETTm2*	24	64	$\mathbf{0.3858}\pm0.0029$	0.3905 ± 0.0031	+1.23
Weather	48	64	$\mathbf{0.3278}\pm0.0050$	0.3301 ± 0.0050	+0.72

The matched ablation in table 15 changes only the raw patch projection versus the chronological atomic parameterization while retaining the same parent patch length, backbone width and depth, feed-forward width, training protocol, full test origins, and seeds. It covers every clean matched control available in the frozen evaluation records: five data sets and three seeds.

The atomic encoding wins 9 of 15 matched seed–data-set runs, with an equally weighted pooled H720 change of -0.51%. The mixed per-data-set signs show modest changes after independent training despite expressive equivalence. They motivate examining grouping and width separately in the following grid.

Table 16: Atomic grouping and model-width sensitivity on four representative data sets. Each entry is the three-seed mean ± sample standard deviation of MSE or MAE averaged over $H \in \{96, 192, 336, 720\}$; lower is better and the best and second-best distinct means in each data-set/metric group are bold and underlined, respectively. All 36 settings use $W = 672$, full test origins, fixed P12–16 atoms, and depth two.

Dataset	Atoms	Metric	Model width D		
	m		64	128	256
ETTm2	2	MSE	0.2816±0.0014	0.2769±0.0058	0.2999±0.0429
		MAE	0.3302±0.0020	0.3250±0.0015	0.3367±0.0221
	4	MSE	**0.2764**±0.0039	<u>0.2765</u>±0.0031	0.2854±0.0026
		MAE	**0.3231**±0.0029	<u>0.3242</u>±0.0021	0.3290±0.0014
	8	MSE	0.2835±0.0034	0.2844±0.0031	0.2862±0.0065
		MAE	0.3257±0.0012	0.3276±0.0025	0.3295±0.0046
Weather	2	MSE	**0.2333**±0.0007	0.2508±0.0248	0.2408±0.0062
		MAE	<u>0.2708</u>±0.0006	0.2842±0.0219	0.2750±0.0079
	4	MSE	0.2348±0.0032	0.2373±0.0043	0.2497±0.0127
		MAE	0.2718±0.0029	**0.2698**±0.0022	0.2806±0.0092
	8	MSE	<u>0.2343</u>±0.0003	0.2401±0.0042	0.2365±0.0021
		MAE	0.2723±0.0005	0.2749±0.0019	0.2739±0.0028
ECL	2	MSE	0.1609±0.0003	0.1556±0.0006	0.1572±0.0007
		MAE	0.2569±0.0003	0.2504±0.0007	0.2499±0.0004
	4	MSE	0.1632±0.0006	0.1568±0.0005	<u>0.1538</u>±0.0001
		MAE	0.2593±0.0004	0.2509±0.0003	**0.2470**±0.0002
	8	MSE	0.1631±0.0005	0.1570±0.0002	**0.1526**±0.0002
		MAE	0.2596±0.0003	0.2525±0.0000	<u>0.2476</u>±0.0003
Traffic	2	MSE	0.4006±0.0012	0.3938±0.0006	0.3921±0.0016
		MAE	0.2756±0.0016	0.2678±0.0009	0.2655±0.0006
	4	MSE	0.4113±0.0011	0.3919±0.0014	**0.3828**±0.0009
		MAE	0.2858±0.0006	0.2703±0.0005	**0.2628**±0.0012
	8	MSE	0.4138±0.0010	0.3971±0.0009	<u>0.3841</u>±0.0013
		MAE	0.2879±0.0006	0.2739±0.0006	<u>0.2629</u>±0.0009

Across the 12 data-set/D groups, changing m at fixed D gives a mean within-group normalized range $(\max - \min)/\operatorname{mean}$ of 2.85% MSE and 2.03% MAE. Across the 12 data-set/m groups, changing D at fixed m gives 5.08% and 4.16%, respectively. The observed range is smaller along m and larger along D, but D also changes backbone capacity. Its effect is data-set-specific: ETTm2 and Weather attain their lowest MSE at $D = 64$, while ECL and Traffic do so at $D = 256$; width is not monotonically beneficial.

Because $m = P/12$, changing m also changes the inherited parent span P; this is a parent-scale/atom-count sensitivity audit, not an isolated atom-count intervention or a test-set model-selection rule. In the starred ETTm2 P24/$D = 64$ row of table 15, exact Raw-to-P12–16 transplantation gives zero initial discrepancy, but independent training changes H720 by $+1.226\%$ (4/12 test seed–block wins). Preserving the raw projection class therefore does not guarantee identical results after training.

Frozen tangent drift geometry. For diagnostic analysis only, let $\Pi_{\text{lev}} = \mathbf{1}\mathbf{1}^\top/n$, where $\mathbf{1} \in \mathbb{R}^n$ is the all-ones vector on the n analyzed time coordinates, and decompose each channel's later-patch displacement into $d_{\text{lev}} = \Pi_{\text{lev}}(T_\tau - G)$ and $d_{\text{shape}} = (I - \Pi_{\text{lev}})(T_\tau - G)$. These coordinates are orthogonal, sum to $T_\tau - G$, and are never fitted or deployed separately.

We first audit the paper's main local ATD structure. Every train-selected τ^* is reused without reselection, and only the held-forward fourth training block is measured. At H720, table 17 reports later-patch coordinates, excluding the protected first patch of each commit. Level and shape cosines are computed within their respective orthogonal residual coordinates; "full–shuffle" is $\cos(T_{\tau^*} - G, y - G)$ minus the mean cosine of eight deterministic phase-permuted templates.

The complete direction is positively aligned in 42/42 seed runs, whereas the level and shape components are individually positive in 31/42 and 33/42 runs. ECL/Traffic contain level-negative runs, while ETTh2 is shape-negative in all six width–seed runs. This heterogeneity is why the decomposition is used as a view rather than as two independently sufficient mechanisms.

We then repeat the geometry audit across the three named transfer parents. For ATD-2, table 19 compares native recursion depths 1 and 3 on later-patch coordinates under the same held-forward protocol.

The level fraction rises in 33/36 paired summaries, and the phase-specific cosine margin rises in 34/36. Across all cumulative, band, and depth summaries over later-patch coordinates, the mean-free shape direction has

Table 17: Frozen held-forward training-block-four H720 geometry for the main local ATD structure, averaged over three seeds. Component signs vary by data set; these coordinates are a diagnostic view of the shared direction, not separately deployed correction mechanisms.

Data/width	Level frac.	Level cos.	Shape cos.	Full cos.	Full–shuffle
ETTh1 ATD-4	0.265	0.407	0.167	0.243	0.246
ETTh1 ATD-8	0.254	0.317	0.171	0.214	0.224
ETTh2 ATD-4	0.502	0.128	-0.027	0.056	0.013
ETTh2 ATD-8	0.506	0.137	-0.031	0.068	0.013
ETTm1 ATD-4	0.277	0.318	0.140	0.193	0.158
ETTm1 ATD-8	0.273	0.258	0.166	0.195	0.173
ETTm2 ATD-4	0.533	0.241	0.001	0.124	0.020
ETTm2 ATD-8	0.524	0.196	0.009	0.107	0.019
Weather ATD-4	0.270	0.314	0.099	0.118	0.000
Weather ATD-8	0.267	0.311	0.094	0.118	0.000
ECL ATD-4	0.287	-0.051	0.130	0.089	0.037
ECL ATD-8	0.294	-0.021	0.127	0.081	0.030
Traffic ATD-4	0.063	-0.030	0.109	0.102	0.051
Traffic ATD-8	0.064	-0.071	0.101	0.093	0.047
Macro ATD-4	0.314	0.189	0.088	0.132	0.075
Macro ATD-8	0.312	0.161	0.091	0.125	0.072

Table 18: Positive parent–data–seed runs out of 36, with mean explained error in parentheses, on the three-parent held-forward grid. The combined direction is consistently stronger than either diagnostic coordinate.

Direction	ATD-2, depth 1	ATD-2, depth 3	ATD-8, depth 1
Level	33/36 (0.026)	36/36 (0.053)	33/36 (0.036)
Shape	36/36 (0.052)	36/36 (0.070)	36/36 (0.058)
Combined	**36/36 (0.070)**	**36/36 (0.118)**	**36/36 (0.089)**

Table 19: Frozen held-forward training-block-four ATD-2 transfer-parent drift geometry, averaged over four data sets and three seeds per parent. Native depths span 192 points for AutoTimes and Timer and 256 for TimesFM.

Parent/depth	Level frac.	Level cos.	Shape cos.	Full cos.	Full–shuffle
AutoTimes, depth 1	0.428	0.061	0.263	0.169	0.091
AutoTimes, depth 3	0.571	0.193	0.274	0.242	0.117
Timer, depth 1	0.418	0.208	0.360	0.290	0.248
Timer, depth 3	0.542	0.261	0.400	0.351	0.294
TimesFM, depth 1	0.445	0.339	0.221	0.275	0.159
TimesFM, depth 3	0.634	0.307	0.317	0.334	0.240
Macro, depth 1	0.430	0.203	0.281	0.245	0.166
Macro, depth 3	0.582	0.254	0.330	0.309	0.217

positive residual alignment in 996/996 rows. At ATD-8 H720, the true direction exceeds the largest shuffled control in 35/36 parent–data–seed runs; the exception is AutoTimes–ETTh1 seed 2023. Raw cosine itself rises from depth 1 to 3 in only 28/36 paired summaries, so this audit supports repeatable drift geometry but does not support a universal monotone-horizon claim. These are overlapping descriptive summaries on held-forward training block 4, not independent significance trials.

Complete tangent selection. The representation-aligned candidate grid is $\tau \in \{12, 24, \ldots, 216\}$: multiples of the fixed 12-sample atom, capped at the largest multiple with at least three occurrences in $W = 672$. The initial history has times $t = -W, \ldots, -1$ and normalized values $u_{c,t} = (h_{0,c,t} - \mu_{0,c})/\sigma_{0,c}$. For phase $r \in \{0, \ldots, \tau - 1\}$, let $I_r = \{t \in \{-W, \ldots, -1\} : t \bmod \tau = r\}$ and $S_r = I_r \setminus \{\min I_r\}$. The stored template is $v_{\tau,c,r} = |S_r|^{-1} \sum_{t \in S_r} u_{c,t}$; it excludes the oldest occurrence of each phase and is never updated. If a patches have already been committed, local slot $j \in \{1, \ldots, K\}$ and offset $s \in \{0, \ldots, P - 1\}$ correspond to $t = (a + j - 1)P + s$ and endpoint $\ell = (a + j)P$. Using the current rolling history's mean and scale $\mu_{a,c}, \sigma_{a,c}$, the aligned data-space template in equation (8) is

$$(T_\tau)_{c,j,s} = \mu_{a,c} + \sigma_{a,c}\, v_{\tau,c,t \bmod \tau}. \tag{13}$$

Thus T_τ and G both have shape $C \times KP$ after concatenating slots. The stored phase values stay fixed; only their data-space mapping follows the current call. Equivalently, the code subtracts normalized proposals from v_τ, then multiplies the displacement by σ_a before computing data-space moments or writing back the corrected patch. The fixed endpoint ramp is the dimensionless scalar

$$b(\ell) = \begin{cases} 0.25, & 0 \le \ell < 96, \\ 0.5, & 96 \le \ell < 192, \\ 1, & 192 \le \ell < 336, \\ 2, & 336 \le \ell < 672, \\ 3, & 672 \le \ell. \end{cases} \tag{14}$$

It applies uniformly to all P points and all channels of that patch; at a boundary the higher band applies. Local slot $j = 1$ is hard-zeroed in every call, not only at the start of the complete forecast. These boundaries follow the reporting horizons and lookback rather than a data-set search. The implementation's fixed factor $\beta = 0.0625$ is absorbed into the reported $\alpha^* = \beta\gamma$; the tables report the effective coefficient, not γ.

For each data-set–width configuration, 512 uniformly spaced training origins from each seed are divided chronologically into four contiguous 128-origin blocks. Moments from seeds 2021–2023 and blocks 1–3 select one period and coefficient shared by the three seed runs; block 4 confirms the frozen rule without reselection. The selector chooses the smallest period attaining the greatest positive explained error $\max(B_\tau, 0)^2/(A_\tau V)$ and then applies equation (10). Thus reported seed runs have distinct trained parents but are not independent selector refits.

For completeness, for a fixed proposal trajectory and $A_\tau > 0$,

$$R_\tau(\alpha) = \mathbb{E}\|G + \alpha d_\tau - y\|_2^2 = R_\tau(0) - 2\alpha B_\tau + \alpha^2 A_\tau. \tag{15}$$

This strictly convex quadratic has unconstrained minimizer B_τ/A_τ; projection onto $\alpha \ge 0$ gives equation (10). The result is exact for the fixed training proposal trajectory, while closed-loop correction is tested empirically because writeback changes later histories.

The ramp has one prescribed relative shape, not five fitted coefficients. Its global scale is analytically immaterial: for any $c > 0$, replacing d_τ by cd_τ gives $A'_\tau = c^2 A_\tau$, $B'_\tau = cB_\tau$, and $\alpha'^* = \alpha^*/c$. The explained-risk selector and applied correction $\alpha'^* c d_\tau = \alpha^* d_\tau$ are therefore unchanged; the five displayed ramp values do not constitute five fitted coefficients.

Table 20: Train-only Spectrum Tangent sensitivity over seven data sets, ATD-4/8, and three seeds. Training blocks 1–3 reselect and fit each listed variant; block 4 confirms the frozen rule. R^2 is equal-configuration held-forward explained error, and retained is relative to the fitted staged default. No validation or test example is read.

Factor	Choice	Cfg. wins	Seed wins	R^2 (%)	Retained (%)
Ramp	**staged (used)**	14/14	42/42	1.772	100.0
	flat	13/14	38/42	1.307	73.8
	linear	14/14	42/42	1.806	101.9
Period grid	**12:12:216 (used)**	14/14	42/42	1.772	100.0
	24:24:216	14/14	42/42	1.772	100.0
	12:12:168	14/14	42/42	1.772	100.0
Coefficient	0.50	14/14	42/42	1.298	73.2
	0.75	14/14	42/42	1.638	92.4
	1.00 (fit)	14/14	42/42	1.772	100.0
	1.25	14/14	42/42	1.701	96.0
	1.50	12/14	35/42	1.423	80.3

The linear ramp slightly exceeds the staged ramp on this diagnostic, while both retain 14/14 configuration and 42/42 seed wins. Thus the audit supports increasing long-horizon weight but does not identify the staged breakpoints as unique. The flat ramp drops to 13/14 configurations and 38/42 seeds. Both alternative period grids retain 14/14 and 42/42, as do coefficient multipliers from 0.50 through 1.25; at 1.50 the counts fall to 12/14 and 35/42. These are raw-trajectory held-forward training risks, not closed-loop test re-ranking.

Table 8 lists the pooled local ATD-4/8 and named-parent ATD-2/4/8 periods and effective coefficients. Within each data-set–width configuration, training blocks 1–3 are pooled across the three seeds to select one shared rule. A no-pooling control instead refits period and coefficient from each seed's blocks 1–3 and evaluates its block 4. It yields 42/42 positive checks; all 42 select the pooled τ^*, and coefficients are 0.887–1.124$\times$ the pooled values. A complementary leave-one-seed-out control fits on two seeds and evaluates the excluded seed's block 4. It likewise yields 42/42 positive checks and all 42 pooled periods; its coefficient is 0.941–1.051$\times$ the pooled value. Thus three-seed pooling is not required for held-forward selector stability; both controls are train-only and do not re-estimate test risk.

Table 21: Exchange-Rate stress test over all 798 test origins and three seeds. Entries are mean±sample standard deviation; positive Δ means the correction is worse. The route is selected only from training blocks and is not changed after observing this table.

		MSE			MAE		
Mode	H	ATD	+Tangent	Δ%	ATD	+Tangent	Δ%
ATD-4	96	0.1109 ± 0.0078	0.1109 ± 0.0078	+0.00	0.2269 ± 0.0078	0.2269 ± 0.0078	+0.00
ATD-4	192	0.2199 ± 0.0062	0.2282 ± 0.0069	+3.80	0.3285 ± 0.0071	0.3344 ± 0.0075	+1.80
ATD-4	336	0.4026 ± 0.0111	0.4205 ± 0.0124	+4.45	0.4573 ± 0.0096	0.4670 ± 0.0094	+2.12
ATD-4	720	1.0264 ± 0.0247	1.0554 ± 0.0266	+2.82	0.7610 ± 0.0101	0.7704 ± 0.0092	+1.22
ATD-8	96	0.1109 ± 0.0078	0.1109 ± 0.0078	+0.00	0.2269 ± 0.0078	0.2269 ± 0.0078	+0.00
ATD-8	192	0.2175 ± 0.0079	0.2258 ± 0.0088	+3.80	0.3273 ± 0.0082	0.3331 ± 0.0087	+1.76
ATD-8	336	0.3939 ± 0.0055	0.4121 ± 0.0075	+4.62	0.4527 ± 0.0069	0.4627 ± 0.0073	+2.20
ATD-8	720	0.9870 ± 0.0163	1.0291 ± 0.0144	+4.26	0.7475 ± 0.0032	0.7610 ± 0.0032	+1.81

Table 22: Detailed local H720 accounting. Every row averages all test origins and three seeds. Seed runs and test seed–blocks count strict improvements over the matched ATD model; negative Δ is better.

Dataset	Mode	ATD	+Tangent	Δ	Runs	Test blocks
ETTh1	ATD-4	0.4054	0.3970	-0.0084	3/3	11/12
ETTh1	ATD-8	0.4052	0.3980	-0.0072	3/3	11/12
ETTh2	ATD-4	0.4191	0.3917	-0.0274	3/3	11/12
ETTh2	ATD-8	0.4208	0.3972	-0.0237	3/3	11/12
ETTm1	ATD-4	0.4210	0.4151	-0.0060	3/3	9/12
ETTm1	ATD-8	0.4186	0.4120	-0.0066	3/3	9/12
ETTm2	ATD-4	0.3884	0.3760	-0.0124	3/3	9/12
ETTm2	ATD-8	0.3865	0.3726	-0.0140	3/3	9/12
Weather	ATD-4	0.3278	0.3245	-0.0033	3/3	11/12
Weather	ATD-8	0.3278	0.3246	-0.0032	3/3	12/12
ECL	ATD-4	0.2018	0.1990	-0.0028	3/3	9/12
ECL	ATD-8	0.2017	0.1994	-0.0023	3/3	8/12
Traffic	ATD-4	0.4384	0.4301	-0.0084	3/3	12/12
Traffic	ATD-8	0.4368	0.4305	-0.0063	3/3	11/12

Weak-seasonality stress test on Exchange-Rate. Exchange is deliberately kept outside the seven-dataset mean and used as a weak-seasonality stress test rather than another positive cell; we do not assign the selected period a known physical-seasonality interpretation. We reuse the same P96/D64/L1 parent interface and three seeds, train ATD-8 by the same validation-only selector as the other exits, and then apply the unchanged Spectrum Tangent protocol independently as Tangent-4 and Tangent-8. Both widths select $\tau = 12$, with $\alpha^* = 1.496$ and 1.518, and all 6/6 held-forward training checks are positive.

The preserved first proposal slot makes H96 identical, but the correction worsens H720 MSE/MAE by +2.82%/ +1.22% for Tangent-4 and +4.26%/ +1.81% for Tangent-8. It wins 0/6 H720 seed runs and only 6/24 reporting blocks. The gap between positive held-forward training geometry and negative test risk is therefore retained as a distribution-shift counterexample: the closed-form coefficient is exact for its fixed training trajectory, not a guarantee that a history template transports across regimes.

Horizon and H720 configuration detail.

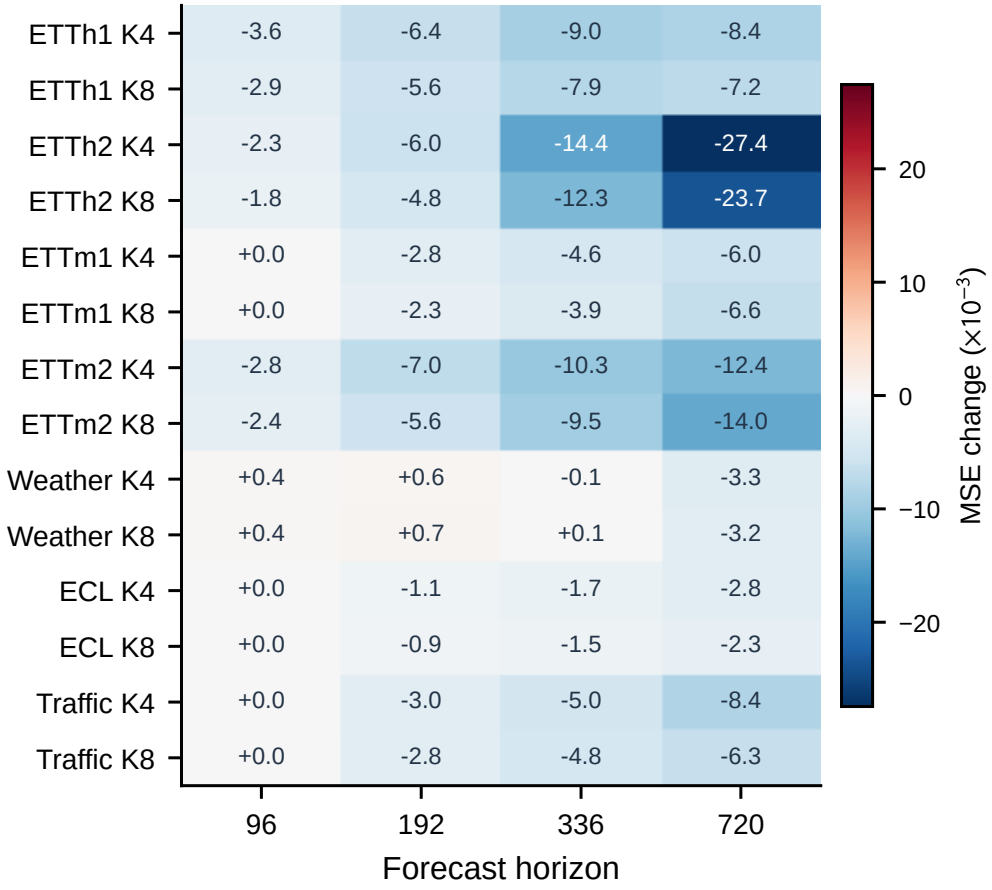

Figure 2: Per-configuration MSE change from ATD to Spectrum Tangent. The correction is strongest and most consistent at H720; short-horizon effects are smaller and mixed.