LEARNING LATENT ACTIONS FROM JOINT HISTORIES FOR MULTIVARIATE TIME-SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

In multivariate time series forecasting, the same target history can warrant different futures as surrounding observations change. Forecast accuracy alone leaves the context-induced revision unlocalized. ACT defines a forecasting state–action interface: it retains a representation encoded from the target history alone, uses joint histories to select a replaceable action, and reads the updated state with a shared forecast head. A fixed-state intervention protocol replaces only that action, making its forecast effect directly measurable and exactly replayable. LAB is one realization of this interface, using structured feature and variable interactions for exact action extraction and low-rank implementation; direct scaling provides a second realization under the same interface. Across 20 models on four datasets, actions selected from the observed joint history beat zero, equal-magnitude noise, and shuffled actions; direction and target matching matter beyond magnitude. Across five Solar sites and 2,527 active site–time pairs, the observed context lowers six-hour mean squared forecast error by 28.0% relative to substitutions matched on target history and daily phase; replay reconstructs the forecast difference to numerical precision. ACT is competitive with 12 forecasters across eight benchmarks. Relative to dense LAB, low-rank LAB cuts training-and-validation time per epoch by about 42% on Electricity and 56% on Traffic.

1 INTRODUCTION

Multivariate forecasting uses *context* to determine how a target's recent trajectory should continue. A decline in solar power may follow the fleet's daily progression or mark a shared departure, so identical target histories can warrant different conditional forecasts (Figure 1).

Temporal mappings and patch encoders capture each series' past, while variable-token attention and mixers combine histories (Zeng et al., 2023; Li et al., 2023; Nie et al., 2023; Liu et al., 2024; Ekambaram et al., 2023). TimeXer separates targets from external information; TimeFilter, TQNet, and FACT adapt dependencies through graph filtration, temporal queries, and across-variable convolutions (Wang et al., 2024b; Hu et al., 2025; Lin et al., 2025; Wang et al., 2026). These designs improve how histories inform prediction; ACT makes the context-dependent revision an explicit object by keeping the target-history representation and forecast head fixed while varying only the update selected by surrounding observations.

ACT makes this fixed-reference test explicit. An encoder retains each target's history state, joint histories select a *latent action*, and a shared head reads the updated state. Conditional modulation names the feature operation; ACT contributes an intervenable forecasting state–action interface and a fixed-state intervention protocol: only the action is replaced, and its forecast revision is replayed against the same head's no-action forecast.

The Latent Action Block (LAB) is one realization of this interface: its feature, variable, and residual routes reuse inputs multiplicatively, permitting exact recovery of the effective action even at zero-valued retained features. Low-rank factors control capacity and cost. Direct feature scaling is also competitive on Weather and Traffic under the same interface, demonstrating that the state–action interface supports multiple generator designs.

Across all 20 validation-selected checkpoints, native context-selected updates outperform zero, norm-matched noise, and globally shuffled updates in both MSE and MAE at the same target state and fixed head. Zeroing tests whether the update contributes, norm-matched noise holds its magnitude while changing its direction, and shuffling preserves the update distribution while breaking

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.

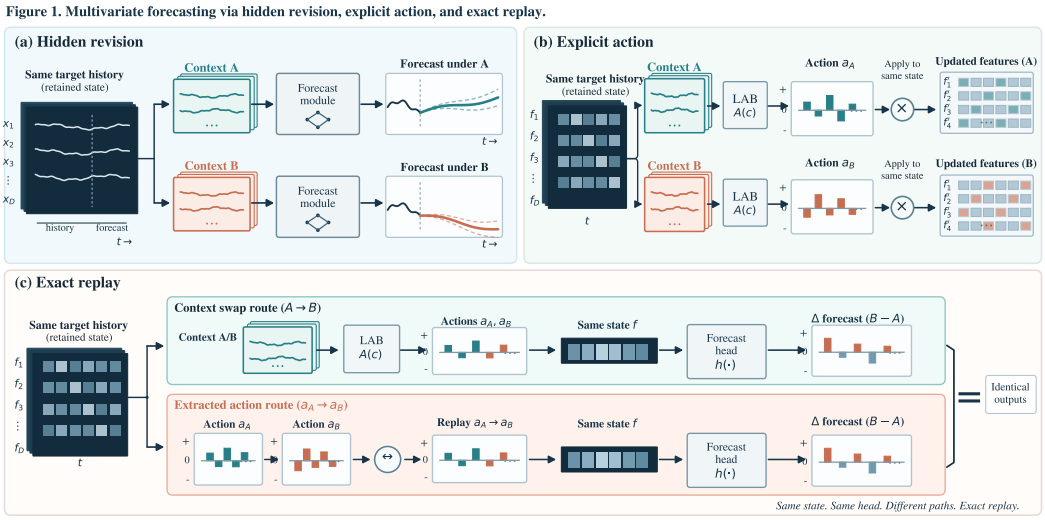


Figure 1: From a hidden contextual revision to an intervenable forecasting state–action interface. (a) With target history fixed, alternative surrounding histories can change the forecast, while the final outputs leave the revision unlocalized. (b) ACT retains a target-history state and represents the context-selected feature update as a latent action; LAB is one realization that analytically exposes the effective action of its composite routes. (c) At a fixed target state and prediction head, replacing context and replaying the extracted action change yield the same forecast difference. Feature strips and signs are schematic, not measured trajectories.

recipient pairing. Native updates win in every model, showing that predictive value depends on both the direction of the revision and its compatibility with the state it updates.

Context-label readouts ask whether information about surrounding observations is recoverable from the action; context substitution tests whether it improves the forecast. On Solar, adding other sites' histories improves six-hour mean squared error over target-history and calendar controls by 8.12% (shuffled context: −0.89%). For the fleet anomaly—the fleet-wide deviation from its daily profile—a 16-coordinate readout improves the coefficient of determination from 0.659 for the retained state to 0.706 for the action. Replacing context while fixing the target history across five Solar sites shows 28.0% lower six-hour mean squared error with the observed context than with matched substitutions, and action replay recovers the forecast difference. Separately, combining two context groups yields predictive effects beyond the sum of their separate contributions on Electricity, Traffic, and Weather. Across eight benchmarks, ACT is competitive with twelve forecasters; low-rank LAB speeds per-epoch training and validation by 1.72× on Electricity and 2.30× on Traffic relative to dense LAB.

2 Related Work

Temporal modeling and multivariate context. Forecasting architectures organize temporal structure and multivariate context. Linear and patch-based models encode histories (Zeng et al., 2023; Li et al., 2023; Ma et al., 2026; Nie et al., 2023), while mixers, multiscale models, and frequency transforms combine representations (Ekambaram et al., 2023; Wang et al., 2024a; Yi et al., 2024b;a; Wu et al., 2023). Cross-variable attention and target–external separation organize context in iTransformer and TimeXer (Liu et al., 2024; Wang et al., 2024b). TimeFilter, TQNet, FACT, and TimePro adapt graph, query, convolutional, or variable–time representations (Hu et al., 2025; Lin et al., 2025; Wang et al., 2026; Ma et al., 2025). Deep state-space models and neural relational inference model hidden dynamics and interactions (Rangapuram et al., 2018; Kipf et al., 2018). ACT isolates how surrounding histories revise a retained target-history state for conditional forecasting.

Latent actions and their operational domain. Latent actions serve distinct operations: LAPO learns transition codes for policies, while Genie uses them to generate videos in interactive environments (Schmidt & Jiang, 2024; Bruce et al., 2024). In embodied control, UniPi decodes controls from video plans, LAPA grounds pretrained transition codes in robot actions, and villa-X learns codes through visual and proprioceptive prediction (Du et al., 2023; Ye et al., 2025; Chen et al.,

2026). ACT defines a forecasting latent action as a replaceable model-internal update to a retained forecasting state, evaluated by fixed-state forecast intervention (Table 1). Its operational semantics are forecasting-state revision: joint histories select the update, the shared head reads it, and the protocol replays its effect at a fixed reference state.

Conditional feature updates. Feature-wise linear modulation (FiLM) conditions features through learned scales and offsets, and temporal FiLM extends this mechanism to sequences (Perez et al., 2017; Birnbaum et al., 2019). HyperNetworks generate weights, whereas cross networks multiply learned transformations by the original input (Ha et al., 2016; Wang et al., 2017; 2021). ACT organizes conditional feature updates into a forecasting interface: surrounding histories select coefficients that update an independently encoded target-history state, the shared head maps the updated state to a prediction, and fixed-state replacement replays the corresponding forecast correction. Direct scaling and LAB are alternative generators for this interface; LAB adds analytic extraction from axis-structured routes and low-rank implementation (Section 4.4).

Table 1: Connections to related models. iTrans. denotes iTransformer; DCN-V2 denotes Deep & Cross Network v2. Checks describe the cited constructions, with axis split for temporal/variable operators and low rank for factorized maps. Products are elementwise; base reuse anchors route updates. Transition targets are transition-derived training codes; control decoders yield environment actions; ACT defines a replaceable forecast-state update evaluated by fixed-state intervention, and LAB is its structured realization with exact extraction.

(a) Update construction					(b) Action learning and use		
Method	Axis split	Products	Base reuse	Low rank	Method	Transition targets	Control decoder
iTrans.	✓	×	×	×	LAPO	✓	✓
TimeXer	✓	×	×	×	UniPi	×	✓
FiLM	×	✓	×	×	LAPA	✓	✓
DCN-V2	×	✓	✓	✓	villa-X	✓	✓
ACT / LAB	✓	✓	✓	✓	**ACT**	×	×

3 ACT: LEARNING LATENT ACTIONS FROM JOINT HISTORIES

Given L past observations of C variables, $X \in \mathbb{R}^{L\times C}$, we predict H future observations $Y \in \mathbb{R}^{H\times C}$. For target c, let $x_c = X_{:,c}$ be its target history and X_{-c} its surrounding context; their union is the joint history X. ACT defines a forecasting state–action interface: it independently encodes x_c as retained target-history state $z_c \in \mathbb{R}^d$, where d is the feature width, while the joint history generates action coefficient $a_c \in \mathbb{R}^d$ (Figure 2). Across variables, these states and coefficients form matrices Z and A. We call a_c a latent action because it specifies the model-internal operation $\mathcal{T}_{a_c}(z_c) = z_c + z_c \odot a_c$: $d_c = z_c \odot a_c$ is the applied state update, and $u_c = z_c + d_c$ is the updated state read by the forecast head. The head maps u_c to $\widehat{y}_c$; the forecast correction $\delta_c = \widehat{y}_c - \widehat{y}_{0,c}$ compares this forecast with the same head's zero-action forecast from z_c. With target history and forecast head fixed, surrounding context selects the forecast revision. We use a_c/A, d_c/D, u_c/U, and δ_c for the latent action, applied state update, updated state, and forecast correction, respectively; LAB and direct scaling instantiate the interface.

3.1 RETAINING THE HISTORY, LEARNING ITS UPDATE

A shared temporal encoder independently processes each normalized history $\widetilde{x}_c \in \mathbb{R}^L$:

$$z_c = \mathrm{Dropout}_{0.1}[\mathrm{GELU}(W_{\mathrm{in}}\widetilde{x}_c + b_{\mathrm{in}})], \qquad Z = [z_1, \ldots, z_C]^\top \in \mathbb{R}^{C\times d}. \tag{1}$$

Here $W_{\mathrm{in}} \in \mathbb{R}^{d\times L}$ and $b_{\mathrm{in}} \in \mathbb{R}^d$ are shared across variables. GELU, $\mathrm{GELU}(v) = v\Phi(v)$, uses the standard Gaussian cumulative distribution function Φ (Hendrycks & Gimpel, 2016); dropout masks and rescales features during training. When enabled, reversible instance normalization (RevIN) uses each variable's history statistics and reverses the transform after prediction (Kim et al., 2022). Thus, variables share a temporal map, but each z_c depends only on its own history.

From the joint state, the generator produces $A = \mathrm{Act}(Z) \in \mathbb{R}^{C\times d}$, with row c equal to a_c. The matrices $U = Z + Z \odot A$ and $D = U - Z = Z \odot A$ collect the updated states and state updates, respectively. A shared linear head maps each updated state to H future values:

$$\widehat{y}_c = \mathcal{N}_c^{-1}(W_{\mathrm{out}}[z_c + z_c \odot a_c] + b_{\mathrm{out}}), \qquad W_{\mathrm{out}} \in \mathbb{R}^{H\times d}. \tag{2}$$

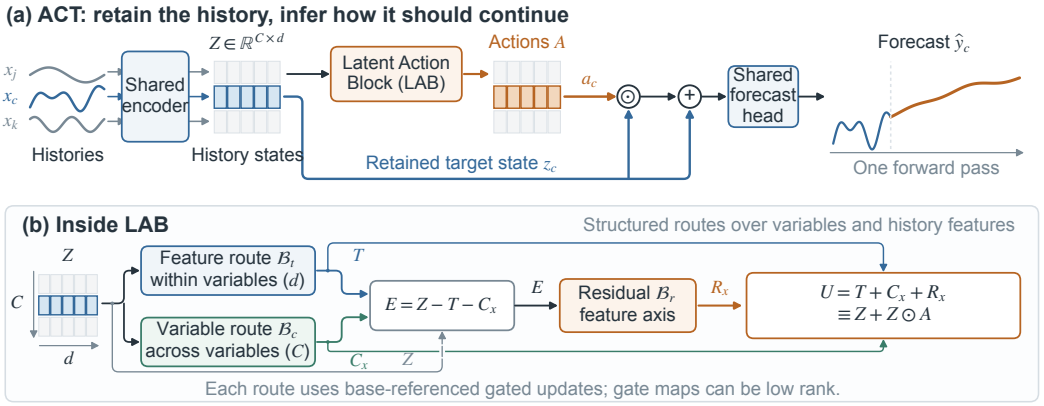


Figure 2: ACT independently encodes histories into retained states Z and generates action coefficients A from their joint state. LAB composes history-feature, cross-variable, and residual routes into $U = Z + Z \odot A$, with A analytically recoverable from route scales. Blue marks the state, amber the action, and normalization is omitted.

Here, $\mathcal{N}_c^{-1}$ reverses the target's normalization and $b_{\text{out}} \in \mathbb{R}^H$ is the bias. Because each state uses only its own history, changing X_{-c} leaves z_c fixed and changes its forecast through a_c. A target receiving a transferred or replacement action is its *recipient*.

The encoder, generator, and head jointly minimize mean squared error (MSE), $\mathcal{L} = \mathbb{E}[\|\widehat{Y} - Y\|_F^2/(HC)]$, where $\|\cdot\|_F^2$ sums squared entries. Actions use past observations only; future targets supervise the forecast loss but enter neither the generator nor action labels.

3.2 Axis-structured action construction

As the axis-structured composite implementation of the ACT interface, LAB builds the updated state from joint history Z through three routes: a *history-feature route* mixes the d features within each variable, a *cross-variable route* mixes the C variables at each feature, and a *residual route* transforms what remains after the first two outputs:

$$\begin{aligned} T &= \mathcal{B}_t(Z), & C_x &= \mathcal{B}_c(Z^\top)^\top, \\ E &= Z - T - C_x, & R_x &= \mathcal{B}_r(E), \qquad U = T + C_x + R_x. \end{aligned} \tag{3}$$

All five matrices have shape $C \times d$; C_x is the cross-variable output, not the variable count. The maps $\mathcal{B}_t$ and $\mathcal{B}_r$ act along features, while $\mathcal{B}_c$ acts along variables. Since E is the remainder after T and C_x, passing it unchanged gives $U = Z$.

Following cross networks (Wang et al., 2017; 2021), each route anchors its repeated updates at input u_0 while its gate depends on the evolving state:

$$u_{\ell+1} = u_\ell + \mathcal{D}_\ell\left[\text{GELU}(u_0 \odot g_\ell(u_\ell))\right], \qquad \ell = 0, \ldots, k-1. \tag{4}$$

The *gate* g_ℓ produces input-dependent coefficients through four affine maps: a nonlinear feature path and a path controlled by sigmoid $\sigma(v) = 1/(1+e^{-v})$ (Appendix A.1). $\mathcal{D}_\ell$ applies training dropout with probability 0.5; layers and routes have separate parameters.

A width-m gate map is either dense, $W \in \mathbb{R}^{m \times m}$, or factorized as $W = PQ$, with $P \in \mathbb{R}^{m \times r}$ and $Q \in \mathbb{R}^{r \times m}$. Rank r controls each map's interaction capacity and depth k the number of updates. At $k = 0$, every route is the identity, so $u_k = u_0$ and $u_k - u_0 = 0$. These choices vary capacity and cost while preserving the route-update interface (Section 4.4, Appendix E).

3.3 Exact action extraction and forecast effect

At evaluation, Equation (4) and $\text{GELU}(v) = v\Phi(v)$ imply $u_k = u_0 \odot s_k$, where s_k accumulates the route's scales. Let $S_t, S_c, S_r \in \mathbb{R}^{C \times d}$ denote the history-feature, cross-variable, and residual scales, respectively, transposing the cross-variable scale to match Z. Because the residual route scales the remainder $E = Z - T - C_x$, summing the three outputs yields the base-relative coefficient specified below. The proposition characterizes LAB's realization of the interface; direct scaling provides a same-interface generator whose coefficient is explicit by construction.

Proposition 1 (Exact factorization)**.** *For evaluation-mode LAB, with* $\mathbf{1}$ *denoting the all-ones matrix, the exact action coefficient on* Z *is*

$$A = (\mathbf{1} - S_t - S_c) \odot (S_r - \mathbf{1}) \tag{5}$$

The resulting updated state is $U = Z \odot (\mathbf{1} + A) = Z + Z \odot A$*, including coordinates where* $Z = 0$*.*

Appendix B gives the recurrence and proof. The replacement targets are distinct: replacing D transfers the update itself, whereas replacing A forms $Z \odot A'$ on the recipient's retained state. Before inverse normalization, the linear head maps the updated state to a forecast:

$$\widehat{y}_c^{\text{norm}} = W_{\text{out}} \operatorname{diag}(\mathbf{1} + a_c) z_c + b_{\text{out}}, \tag{6}$$

where $\operatorname{diag}(v)$ forms a diagonal matrix. For original and substituted-context actions a_c and a'_c, changing the surrounding histories while holding z_c fixed gives

$$\Delta\widehat{y}_c^{\text{norm}} = W_{\text{out}}(z_c \odot [a'_c - a_c]). \tag{7}$$

At horizon position h, coordinate j contributes $W_{\text{out},hj} z_{cj} a_{cj}$ before inverse normalization. Equation (7) replays a context replacement through its induced action change. With the state fixed, action increments add in forecast space; context edits are evaluated separately through the generator.

Computational cost. Axis-wise low-rank factorization gives a variable-linear alternative to dense joint-history conditioning. Ignoring biases and constant factors, dense routes cost $O(k[2Cd^2 + dC^2])$ per window, compared with $O(kCdr)$ for rank-r routes. The encoder and forecast head add $O(Cd[L + H])$, so for fixed d, k, r, L, H, total computation scales linearly with the number of variables. Separately, for an individual evaluation-mode route of width m and depth k, the Jacobian with respect to u_0 decomposes into a diagonal part and an interaction term of rank at most $\min(m, 2kr)$ (Proposition 2, Appendix B.3). This bound applies to one route, not the full action.

4 Experiments

4.1 Forecasting performance

Across eight chronological benchmarks—the four Electricity Transformer Temperature (ETT) datasets (ETTh1, ETTh2, ETTm1, ETTm2), Electricity (ECL), Solar, Traffic, and Weather—we evaluate lookback $L = 96$ and horizons $H \in \{96, 192, 336, 720\}$. Mean squared error (MSE) and mean absolute error (MAE) are measured in the dataset-standardized scale; lower is better. For each dataset–horizon cell, three-seed mean validation MSE selects one of 20 rank/depth settings; two further seeds use that setting, and we report all five test scores (Appendix A).

Against twelve forecasters, ACT has best or tied-best horizon-averaged MSE on four datasets and MAE on four (Table 2). For each dataset and metric, the table gives each of five model families' lowest model-level four-horizon average (Tables 11–13). This establishes endpoint accuracy; the fixed-state controls below test how the learned revision contributes.

Throughout, percentage MSE gain is $100(M_{\text{ref}} - M_{\text{method}})/M_{\text{ref}}$ for the stated reference: positive values mean improvement. Absolute MSE gains retain the original MSE units.

Table 2: Horizon-averaged MSE↓/MAE↓ against five baseline families. Trans., Freq./Period, and Dep.-aware abbreviate Transformer, frequency/period, and dependency-aware models. Bold and first-place counts include ties at three decimals. Appendix C reports all twelve models and four horizons.

Dataset	ACT	Linear	Mixer	Trans.	Freq./Period	Dep.-aware
ETTh1	0.444/0.434	0.444/**0.431**	0.456/0.446	**0.442**/0.440	0.451/0.438	0.445/0.435
ETTh2	0.373/**0.398**	0.384/0.403	**0.368/0.398**	0.379/0.405	0.380/0.405	0.378/0.403
ETTm1	**0.377/0.392**	0.404/0.400	0.401/0.406	0.404/0.406	0.393/0.402	0.384/0.399
ETTm2	**0.272/0.318**	0.287/0.329	0.290/0.333	0.281/0.328	0.279/0.323	0.280/0.326
ECL	0.179/0.272	0.210/0.293	0.179/0.278	0.175/0.270	0.183/0.276	**0.170/0.264**
Solar	**0.207**/0.266	0.327/0.350	0.257/0.292	0.239/0.279	0.248/0.287	0.232/**0.263**
Traffic	0.470/0.294	0.624/0.375	0.506/0.335	0.462/0.287	0.472/0.312	**0.448/0.283**
Weather	**0.242/0.273**	0.266/0.289	0.249/0.276	0.248/0.275	0.247/0.275	0.249/0.276
1st count↑	4/4	0/1	1/1	1/0	0/0	2/3

4.2 PREDICTIVE VALUE OF THE LEARNED UPDATE

We first test whether the *total learned action* is useful, with the recipient state, model weights, and normalization frozen. The native applied update D is selected from the observed joint history; each panel uses five validation-selected $H = 96$ models on Weather, Solar, ECL, and Traffic. Replacements are zero, norm-matched Gaussian noise, or unscaled D permuted across non-overlapping test histories (Table 3). This panel establishes the predictive value of the complete action; the context panels below identify when surrounding histories supply a useful revision.

Table 3: Total learned action under fixed-state replacement. Full-test MSE↓ at $L = H = 96$, averaged over five validation-selected seeds. Noise is injected into the applied update D with matched latent or forecast norm. This table establishes the value of the total learned action; context interventions identify the sources of its forecast revision. Full means and variances appear in Table 14.

Applied update	Weather	Solar	ECL	Traffic
Native D	**0.1547**	**0.1911**	**0.1463**	**0.4475**
Zero	0.1811	0.2561	0.2081	0.7245
Noise, latent-norm matched	0.1829	0.2629	0.2349	0.7776
Noise, forecast-norm matched	0.2158	0.3666	0.2915	0.9820
Shuffled D	0.2360	0.3634	0.2986	1.0205

Across all 20 checkpoints, native D beats every replacement in MSE and MAE. ECL MSE is 0.1463 with native D, 0.2081 with zero, and 0.2915 with correction-norm-matched noise; Traffic gives 0.4475, 0.7245, and 0.9820. Permutation raises MSE to 0.2986 and 1.0205 while preserving update distribution and cross-variable structure. Norm matching tests direction, permutation tests recipient pairing, and the later context panels test whether surrounding histories supply the useful revision.

The correction improves on the zero-action forecast when its alignment with zero-action error exceeds its energy. Let $e = Y - \widehat{Y}_0$ and $\delta = \widehat{Y} - \widehat{Y}_0$:

$$\mathrm{MSE}_0 - \mathrm{MSE} = 2\langle e, \delta\rangle - \langle \delta, \delta\rangle, \tag{8}$$

Inner products average over test windows, forecast positions, and variables; MSE improves when alignment $2\langle e, \delta\rangle$ exceeds correction energy $\langle \delta, \delta\rangle$ (Table 4, same models).

Table 4: Error alignment explains the learned update's absolute MSE gain. Full-test, five-model means from Table 3's checkpoints. Native and forecast-norm-matched noise share correction energy; gain is alignment minus energy (Equation (8)).

	Error alignment		Matched energy	Absolute MSE gain	
Dataset	Native	Noise	Both	Native	Noise
Weather	0.06104	−0.00005	0.03463	0.02641	−0.03468
Solar	0.17551	0.00004	0.11057	0.06494	−0.11054
ECL	0.14514	0.00001	0.08343	0.06171	−0.08343
Traffic	0.53452	0.00004	0.25755	0.27698	−0.25751

On all four datasets, native corrections have greater alignment than energy. On Traffic, 0.53452 alignment minus 0.25755 energy gives a 0.27698 MSE gain; forecast-norm-matched noise instead has near-zero alignment and adds 0.25751 MSE, nearly its 0.25755 energy (Appendix D.1).

To test action-coefficient (A) compatibility separately from replacing the applied update D, we transfer donor coefficients with the recipient state and forecast head fixed. Nearest-own-history matching among 256 training donors beats random transfer in all 25 dataset–seed comparisons; Traffic MSE is 0.482 versus 0.851 for random and 0.389 for native actions (Table 16). The panel uses four channels and non-overlapping windows.

Three further fixed-state probes characterize how context changes the action. Interpolating surrounding encoded histories toward another window raises endpoint MSE to 2.89× the native value on ECL and 2.58× on Traffic, while ETT resolutions remain nearly unchanged (Table 9). On Traffic, a same-window action from another variable gives MSE 0.286 versus 0.455 for a permuted-window action. Splitting context groups also leaves substantial residual interaction beyond additive reconstruction. These probes show that the action is a joint-context-dependent forecast revision, not a magnitude-only feature modulation; Appendix B.6 gives the full protocol.

With the recipient state and forecast head fixed, we compare the full-context action coefficient a_{12} with the target-only coefficient a_0. Coefficients a_1 and a_2 each use the target plus one disjoint context group; their additive reconstruction is $a_{\text{add}} = a_1 + a_2 - a_0$, subtracting the shared target-only contribution once (Appendix B.6). The difference $a_{12} - a_{\text{add}}$ captures context-group interaction. Relative to a_0, native actions reduce target-channel mean MSE by 53.5% on ECL, 45.7% on Traffic, and 32.8% on Weather, and beat a_{add} in all 30 model–partition comparisons per dataset (Table 5).

Table 5: Predictive value of joint context at a fixed target state. Five-seed mean target-channel MSE at $H = 96$; a_0 retains only the target's encoded history and $a_{\text{add}} = a_1 + a_2 - a_0$ reconstructs the full action from two context groups. R_a and R_y compare the full-minus-additive interaction with the full-minus-target-only context effect in action and forecast space, respectively; Appendix B.6 gives the norm ratios. Six partitions are averaged within each seed.

	Target-channel MSE↓			Interaction	
Dataset	Native a_{12}	Target-only a_0	Additive a_{add}	R_a	R_y
ECL	0.2422	0.5205	0.3673	0.743	0.562
Traffic	0.1426	0.2628	0.2337	0.948	0.943
Weather	0.0788	0.1172	0.0872	0.567	0.431
ETTh1	0.0541	0.0541	0.0541	0.034	0.044
ETTm1	0.0297	0.0298	0.0297	0.291	0.233

4.3 Context Information and Its Forecast Effect

Solar and Weather then identify when surrounding histories add readable, forecast-relevant information with the target state and forecast head fixed. Together, these panels distinguish the value of the complete action from the dataset- and context-dependent sources of its revision.

A context-label readout uses ridge regression on a principal component analysis (PCA) projection fitted on training data, with identical target-history and calendar controls and validation-selected regularization. Test coefficient of determination (R^2) measures conditional linear decodability, not predictive utility. Superscript $[q]$ gives the coordinate budget: $x_c^{[q]}, z_c^{[q]}, a_c^{[q]}$ are target history, state, and action; $X^{[q]}, Z^{[q]}$ are flattened joint histories and states.

Solar: departures from the daily fleet profile. For five fixed Solar targets, other sites provide context: power is scaled by each site's training 99th percentile and centered by its training median within-day profile; the cross-site median residual is the *fleet anomaly*. Adding fleet-level and anomaly features to a target-history/calendar predictor reduces six-hour MSE by 8.12%, while shuffled context gives −0.89% (Table 29), establishing signal beyond target history and daily phase.

On active histories (past-hour normalized fleet power > 0.05), the five-target Solar fleet-anomaly readout gives $R^2 = 0.659$ from $z_c^{[16]}$ and 0.706 from $a_c^{[16]}$, versus 0.626 from shuffled actions (Table 6). Full joint-state readouts are reported in Table 32; Figure 3 visualizes the action projections.

Table 6: Context readout on active histories: test $R^2 \uparrow$, five-seed mean ± sample standard deviation (SD). All probes use 16 dimensions and identical own-history/calendar controls. Solar averages five sites within each seed.

Context quantity	$z_c^{[16]}$	$a_c^{[16]}$	Shuffled $a_c^{[16]}$
Solar: fleet anomaly	0.659 ± 0.002	0.706 ± 0.011	0.626 ± 0.001
Solar: anomaly change	0.260 ± 0.007	0.302 ± 0.027	0.185 ± 0.001
Weather: relative radiation	0.658 ± 0.004	0.699 ± 0.025	0.629 ± 0.001

Predictive utility is tested by replacing only non-target context with a training donor from another site, matched on target history and within-day phase and retained by a validation-distance threshold. At fixed target state z_c, the generator recomputes a_c and the unchanged head forecasts. Across 2,527 active site–time pairs, native context lowers six-hour MSE by 28.00 ± 3.52%, with positive gains in all 25 site–model pairs (Figure 4). Replaying the native–donor action difference through Equation (7) reproduces the forecast difference within 10^{-6}.

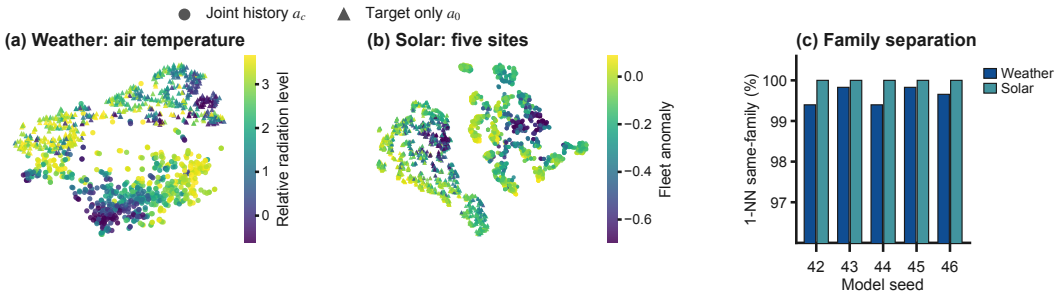


Figure 3: Actions at the readout checkpoints. (a,b) Seed-42 t-SNE projections: joint-history a_c (circles), target-only a_0 (triangles); color denotes Weather radiation or Solar fleet anomaly. Seed-42 readout R^2 (a_c vs. a_0) is 0.699 vs. 0.644 on Weather and 0.704 vs. 0.658 on Solar. (c) Same-action-type nearest-neighbour rates on 50-dimensional PCA of raw actions across seeds. R^2 measures decodability; t-SNE is a projection.

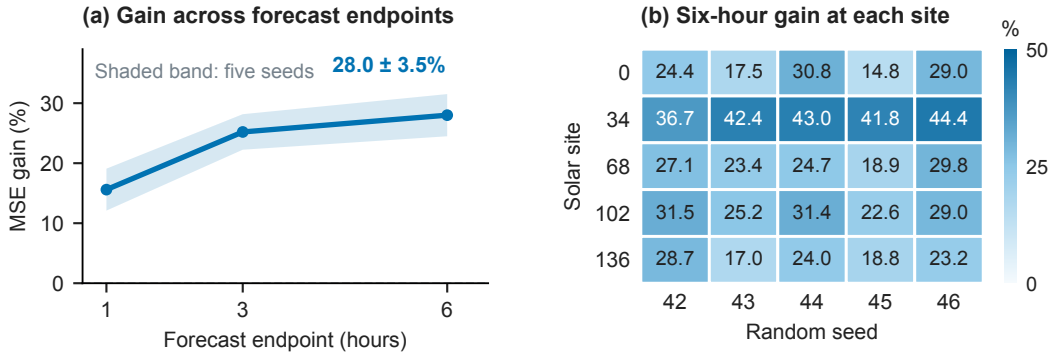


Figure 4: Native versus matched Solar context at fixed target history. (a) Endpoint MSE reductions; (b) six-hour gains across 25 site–model pairs. Values are mean ± sample SD across five models.

Weather: radiation and moisture. For air-temperature forecasting, relative radiation centered by its input-history mean and scaled by its standard deviation has probe $R^2 = 0.699$ from a_c versus 0.658 from z_c. On histories with past-hour radiation ≥ 10 W/m^2, native radiation and moisture context improve six-hour MSE by 4.59 ± 4.22% and 4.89 ± 0.71% over matched substitutions. Appendix H gives the protocol and five-seed results by context family and endpoint.

Future-summary probes. For level change (horizon mean minus last observation) and slope (a linear fit over the horizon), Table 17 compares $z_c^{[16]}$, $a_c^{[16]}$, $d_c^{[16]}$, and $z_c^{[8]} \oplus a_c^{[8]}$, where $\oplus$ concatenates eight coordinates from each source. The joint readout lowers slope probe MSE versus $z_c^{[16]}$ on Weather, ECL, and Traffic, including Traffic 0.544 to 0.373; $z_c^{[16]}$ is better for level change on ETTh1 and ECL. These probes reveal complementary information (Appendix D).

4.4 LAB CAPACITY AND IMPLEMENTATION

Generator controls isolate the state–action interface from LAB; the remaining sweeps identify the structural choices used by this realization (Appendix E).

At fixed $k = 2$, direct scaling and direct affine realize the same interface as LAB; their Weather MSEs span 0.15563–0.15644, while Traffic favors rank-16 LAB in MSE (0.44493) and direct scaling in MAE (0.27892). The complete five-seed generator comparison is reported in Appendix E; a matched 40-run panel gives complementary dataset-wise results (Table 18). LAB adds axis-wise route composition and exact recovery from route scales.

Table 7: Action generators at $L = H = 96$, $d = 512$, $k = 2$ (mean ± SD, five seeds).

Dataset	Generator	Params (M)↓	MSE↓	MAE↓
Weather	Dense LAB	4.305	**0.15563** ± 0.00090	0.20246 ± 0.00128
	Direct scaling	4.305	0.15620 ± 0.00209	**0.20195** ± 0.00271
	Direct affine	4.568	0.15589 ± 0.00088	0.20242 ± 0.00096
	Rank-16 LAB	0.375	0.15644 ± 0.00201	0.20334 ± 0.00255
Traffic	Dense LAB	10.254	0.45630 ± 0.00443	0.28701 ± 0.00151
	Direct scaling	10.254	0.44790 ± 0.00485	**0.27892** ± 0.00668
	Direct affine	10.517	0.46619 ± 0.00462	0.30118 ± 0.00140
	Rank-16 LAB	0.599	**0.44493** ± 0.00430	0.28558 ± 0.00206

At fixed rank $r = 4$, $k = 3$ lowers test MSE versus $k = 0$ by 7.1% on ECL and 4.5% on Traffic, while ETTm2 is lowest at $k = 2$. At fixed $k = 2$, the corresponding rank sweep is lowest at rank 8 on ECL and rank 16 on Traffic (Tables 19 and 20; validation-selected checkpoints, seeds 42–44). The full structural and efficiency profile is provided in Appendix G.

On ECL, post-encoder placement gives MSE 0.14755 (0.14691 before the encoder; 0.14799 for identity), while a current-state factor gives 0.15062 (Table 22); we therefore use post-encoder placement for fixed-state replacement (Section 4.2).

4.5 Route geometry and contribution

Route geometry describes representations; contribution requires testing the computation. Using validation-selected $L = H = 96$ checkpoints across eight datasets, residual-route CKA with its input ranges from 0.81 to 0.996, yet replacing the residual output by its input increases MSE on every dataset (Appendix F). On ECL, the native route gives 29.1% gain over identity replacement and 74.4% over zero replacement (Table 23). Effective-rank and route-label readouts provide complementary geometry diagnostics (Table 24). Thus similarity or route readability alone does not establish forecast contribution; fixed-state replacement identifies whether the learned transformation remains forecast-relevant.

These diagnostics assign distinct roles to the same intervention interface. The readout asks whether context labels remain accessible in the action; a context swap asks whether changing those labels changes the forecast while the target state and head remain fixed; route replacement asks whether the implemented transformation contributes beyond passing its input through unchanged. Their agreement connects action readability to a controlled forecast revision. This is the operational meaning of ACT's state–action interface: a replaceable forecasting computation, not a claim about environmental or embodied action.

5 Conclusion

ACT provides a fixed-state forecasting interface: a replaceable action updates a target-history state read by a fixed head. Replacement controls make the forecast revision measurable while holding the recipient state and head fixed; norm-matched and shuffled replacements separate directional alignment from recipient pairing. Context swaps then change only surrounding observations, identifying when they induce a useful revision.

The evidence separates the interface from its realization: joint-history actions improve over zero, matched-magnitude noise, and shuffled updates, while Solar and Weather interventions show that useful context depends on the target and context family. LAB supplies exact extraction and low-rank gates; direct scaling realizes the same interface under another generator. ACT therefore provides a reusable object for intervention and comparison, with LAB as an efficient structured realization.

AI USE STATEMENT

In this work, we used generative AI tools only for language editing and proofreading of the manuscript. We have not used generative AI tools for research framing, method design, code, data generation, experiments, data analysis, literature search, or scientific decisions. The authors reviewed all AI-assisted edits and take responsibility for the final content of this work, including its text, claims, and artifacts.

REFERENCES

Sawyer Birnbaum, Volodymyr Kuleshov, S. Zayd Enam, Pang Wei Koh, and Stefano Ermon. Temporal FiLM: Capturing long-range sequence dependencies with feature-wise modulations. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/2afc4dfb14e55c6face649a1d0c1025b-Abstract.html.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments, 2024. URL https://arxiv.org/abs/2402.15391.

Xiaoyu Chen, Hangxing Wei, Pushi Zhang, Chuheng Zhang, Kaixin Wang, Yanjiang Guo, Rushuai Yang, Yucen Wang, Xinquan Xiao, Li Zhao, Jianyu Chen, and Jiang Bian. villa-X: Enhancing latent action modeling in vision-language-action models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=y5CaJb17Fn.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.07289.

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 9156–9172. Curran Associates, Inc., 2023. doi: 10.52202/075280-0403. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1d5b9233ad716a43be5c0d3023cb82d0-Paper-Conference.pdf.

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.

David Ha, Andrew Dai, and Quoc V. Le. Hypernetworks, 2016. URL https://arxiv.org/abs/1609.09106.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Yifan Hu, Guibin Zhang, Peiyuan Liu, Disen Lan, Naiqi Li, Dawei Cheng, Tao Dai, Shu-Tao Xia, and Shirui Pan. TimeFilter: Patch-specific spatial-temporal graph filtration for time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 24893–24911, 2025. URL https://proceedings.mlr.press/v267/hu25ac.html.

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.

Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems, 2018. URL https://arxiv.org/abs/1802.04687.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Zhe Li, Shiyi Qi, Yiduo Li, and Zenglin Xu. Revisiting long-term time series forecasting: An investigation on linear mapping. *arXiv preprint arXiv:2305.10721*, 2023.

Shengsheng Lin, Haojun Chen, Haijie Wu, Chunyun Qiu, and Weiwei Lin. Temporal query network for efficient multivariate time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37797–37814, 2025. URL https://proceedings.mlr.press/v267/lin25e.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 *The Twelfth International Conference on Learning Representations*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Aitian Ma, Dongsheng Luo, and Mo Sha. Mixlinear: Extreme low resource multivariate time series forecasting with 0.1k parameters, 2026. URL https://arxiv.org/abs/2410.02081.

Xiaowen Ma, Zhen-Liang Ni, Shuai Xiao, and Xinghao Chen. TimePro: Efficient multivariate long-term time series forecasting with variable- and time-aware hyper-state. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 42096–42111, 2025. URL https://proceedings.mlr.press/v267/ma25p.html.

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

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer, 2017. URL https://arxiv.org/abs/1709.07871.

Syama Sundar Rangapuram, Matthias W Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski. Deep state space models for time series forecasting. *Advances in neural information processing systems*, 31, 2018.

Dominik Schmidt and Minqi Jiang. Learning to act without actions, 2024. URL https://arxiv.org/abs/2312.10812.

Huiqiang Wang, Jieming Shi, and Qing Li. FACT: Fine-grained across-variable convolution for multivariate time-series forecasting. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=j3gNYqrHtl.

Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. Deep & cross network for ad click predictions. In *Proceedings of the ADKDD'17*, ADKDD'17, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450351942.

Ruoxi Wang, Rakesh Shivanna, Derek Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed Chi. Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the Web Conference 2021*, pp. 1785–1797, 2021.

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 *The Twelfth International Conference on Learning Representations*, 2024a.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. TimeXer: Empowering transformers for time series forecasting with exogenous variables. *arXiv preprint arXiv:2402.19072*, 2024b. doi: 10.48550/ARXIV.2402.19072.

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.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos, 2025. URL https://arxiv.org/abs/2410.11758.

Kun Yi, Jingru Fei, Qi Zhang, Hui He, Shufeng Hao, Defu Lian, and Wei Fan. Filternet: Harnessing frequency filters for time series forecasting, 2024a. URL https://arxiv.org/abs/2411.01623.

Kun Yi, Qi Zhang, Wei Fan, Shoujin Wang, Pengyang Wang, Hui He, Ning An, Defu Lian, Longbing Cao, and Zhendong Niu. Frequency-domain mlps are more effective learners in time series forecasting. *Advances in Neural Information Processing Systems*, 36, 2024b.

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.

A Implementation and experimental protocol

This appendix records ACT/LAB implementation and evaluation protocols, derives the LAB factorization and sensitivity bound, and reports extended forecasting and intervention evidence.

A.1 Gate and complete forward computation

Each route multiplies its fixed input by a gate computed from the current representation. For a route of width m, the gate in layer ℓ is

$$\begin{aligned} v &= \mathrm{GELU}(W_1 u + b_1), \\ w &= v + W_2(u + v) + b_2, \\ q &= \mathrm{ELU}(w) + \mathbf{1}, \\ g_\ell(u) &= (W_v q + b_v) \odot \mathrm{sigmoid}(W_g q + b_g). \end{aligned} \tag{9}$$

All four maps preserve width m. Low-rank routes use $W_j = P_j Q_j$ with an r-dimensional intermediate representation. The down projection has no bias; the up projection carries the bias. GELU is defined in Section 3. The exponential linear unit (ELU) equals w for $w \geq 0$ and $e^w - 1$ otherwise (Hendrycks & Gimpel, 2016; Clevert et al., 2016). Each layer and route has its own gate parameters.

The shared input map projects from L to d, followed by GELU and dropout 0.1; a shared output map projects from d to H. Each of the three routes follows Equation (4); during training, dropout 0.5 is applied to the GELU-transformed update before it is added to the running state. Training minimizes forecasting MSE averaged over batch elements, future positions, and variables. All dropout is disabled for evaluation and action extraction.

A.2 Data, normalization, and metrics

The benchmark comprises the four ETT datasets (seven variables each), Weather (21 variables), ECL (321 variables), Traffic (862 variables), and Solar (137 variables). ETTh1/ETTh2, ECL, and Traffic use hourly observations; ETTm1/ETTm2 use 15-minute observations; and Weather and Solar use 10-minute observations. ETT follows the conventional 12/4/4-month train/validation/test split, while the other datasets use chronological 70%/10%/20% splits. At split boundaries, an input history window may include observations preceding the boundary, while its forecast targets remain within the designated split.

Data loading and preprocessing follow the Time-Series-Library pipeline. StandardScaler is fitted on the training segment and applied to the validation and test segments. Models with RevIN additionally normalize each variable using statistics from its input history and learned affine parameters. Inverse RevIN maps the output back to the dataset-standardized scale for training-loss computation and benchmark MSE/MAE evaluation. Solar ACT uses dataset standardization without RevIN.

Algorithm 1 ACT evaluation and exact latent-action extraction

Require: History X, trained parameters, lookback L, horizon H

$\widetilde{X}, \nu \leftarrow \mathrm{Normalize}(X)$ ▷ Own-history statistics ν
$Z \leftarrow \mathrm{Enc}(\widetilde{X})$
$(T, S_t) \leftarrow \mathrm{RouteWithScale}(Z, \mathcal{B}_t)$
$(C_x^\top, S_c^\top) \leftarrow \mathrm{RouteWithScale}(Z^\top, \mathcal{B}_c)$
$C_x \leftarrow (C_x^\top)^\top$
$S_c \leftarrow (S_c^\top)^\top$ ▷ Restore the shape of Z
$E \leftarrow Z - T - C_x$
$(R_x, S_r) \leftarrow \mathrm{RouteWithScale}(E, \mathcal{B}_r)$
$A \leftarrow (\mathbf{1} - S_t - S_c) \odot (S_r - \mathbf{1})$
$D \leftarrow Z \odot A$
$U \leftarrow Z + D$
$\widehat{Y} \leftarrow \mathrm{Denormalize}(\mathrm{Head}(U), \nu)$
return $\widehat{Y}, Z, A$
function $\mathrm{RouteWithScale}(u_0, \mathcal{B})$
  $u \leftarrow u_0$; $s \leftarrow \mathbf{1}$
  for $\ell = 0, \ldots, k-1$ **do**
    $g \leftarrow g_\ell(u)$; $v \leftarrow u_0 \odot g$
    $s \leftarrow s + g \odot \Phi(v)$; $u \leftarrow u + \mathrm{GELU}(v)$
  end for
  return u, s
end function

All benchmark forecasts use a lookback length of $L = 96$. For each dataset–horizon–model combination, let $e \in \mathbb{R}^{N \times H \times C}$ denote the test-error tensor; MSE is $(NHC)^{-1} \sum e^2$ and MAE is $(NHC)^{-1} \sum |e|$. For each metric, the five test scores q_s correspond to the same dataset, forecast horizon, and model. We report their mean $\bar{q} = 5^{-1} \sum_s q_s$ and sample SD $[4^{-1} \sum_s (q_s - \bar{q})^2]^{1/2}$. Horizon averages are computed within each seed before summarizing across seeds, so their SD measures model-seed variation. SD values, including 0.000, are rounded.

A.3 Configuration selection and optimization

For each dataset–horizon cell, we evaluate 20 candidate configurations by crossing four action-route depths, $k \in \{0, 1, 2, 3\}$, with five gate-map choices: dense or factor rank $r \in \{2, 4, 8, 16\}$. At $k = 0$, the routes contain no update layers and produce a zero action. The shared encoder maps each variable's history to a 512-dimensional history state ($d = 512$) in every configuration; the batch size is 32. The learning rate is 10^{-4} for the four ETT datasets and 10^{-3} for Weather, ECL, Traffic, and Solar. We use AdamW (Loshchilov & Hutter, 2019) with zero weight decay and a constant learning rate. Non-Solar ACT runs allow 50 epochs and stop after ten epochs without validation improvement. Solar runs disable RevIN, allow 30 epochs, and stop after three epochs without validation improvement.

Validation MSE selects both checkpoints and configurations. For each search run, we retain the checkpoint with the lowest validation MSE. Within each dataset–horizon cell, we average these minima across seeds 42–44 for each configuration and select the configuration with the lowest mean. Seeds 45–46 use the selected configuration and choose their own checkpoints by validation MSE; test scores enter neither decision. The search comprises $8 \times 4 \times 20 \times 3 = 1{,}920$ runs: eight datasets, four prediction horizons, 20 configurations per cell, and three search seeds, followed by evaluation with seeds 45–46. Table 8 lists all selected settings.

Table 8: Validation-selected ACT settings. All cells use lookback $L = 96$, state width $d = 512$, and batch size 32. Select one setting by mean validation MSE over seeds 42–44, then freeze it for seeds 45–46. Horizons with identical settings are grouped; every dataset–horizon cell is included.

Dataset	H	LAB	r	k	Learning rate	RevIN
ETTh1	96, 192, 336, 720	Dense	–	3	10^{-4}	Yes

Continued on next page

Table 8 (continued)

Dataset	H	LAB	r	k	Learning rate	RevIN
ETTh2	96	Dense	–	3	10^{-4}	Yes
ETTh2	192, 336	Dense	–	2	10^{-4}	Yes
ETTh2	720	Dense	–	0	10^{-4}	Yes
ETTm1	96, 192	Low-rank	16	3	10^{-4}	Yes
ETTm1	336	Low-rank	16	2	10^{-4}	Yes
ETTm1	720	Dense	–	1	10^{-4}	Yes
ETTm2	96	Low-rank	8	3	10^{-4}	Yes
ETTm2	192, 336	Low-rank	16	3	10^{-4}	Yes
ETTm2	720	Low-rank	4	3	10^{-4}	Yes
Weather	96	Low-rank	4	3	10^{-3}	Yes
Weather	192	Low-rank	2	2	10^{-3}	Yes
Weather	336	Low-rank	8	2	10^{-3}	Yes
Weather	720	Low-rank	8	1	10^{-3}	Yes
ECL	96, 192	Dense	–	3	10^{-3}	Yes
ECL	336	Low-rank	2	3	10^{-3}	Yes
ECL	720	Low-rank	2	1	10^{-3}	Yes
Traffic	96	Dense	–	3	10^{-3}	Yes
Traffic	192, 336, 720	Dense	–	2	10^{-3}	Yes
Solar	96	Low-rank	8	3	10^{-3}	No
Solar	192, 336, 720	Dense	–	1	10^{-3}	No

Reproduction commands. The experiment entrypoint accepts the complete configuration explicitly. The following loop runs Weather-96 with each seed from 42 through 46:

```
for seed in 42 43 44 45 46; do
  python run_tim_single.py --dataset Weather --variant lowrank \
    --rank 4 --e_layers 3 --d_model 512 --seq_len 96 \
    --pred_len 96 --lr 0.001 --batch_size 32 \
    --dropout 0.1 --seed "$seed" \
    --max_epochs 50 --early_stopping_rounds 10 --norm_type revin \
    --out_dir ./act_results --ckpt_root ./act_checkpoints
done
```

Each backslash-continued argument block is one shell command. Solar-96 uses rank 8, the same depth/width/learning rate, `--norm_type no`, at most 30 epochs, and patience three. Dense settings use `--variant full`; Table 8 gives the remaining settings.

B Derivations and properties of the action interface

We first establish why surrounding observations can change the squared-error-optimal conditional forecast, then derive LAB's exact route factorization and low-rank sensitivity bound. Fixed-state interventions test the resulting predictive revision with the recipient state and head held fixed.

B.1 Context changes the conditional continuation

Under squared error, let $m_c = \mathbb{E}[y_c \mid x_c]$ and $m_J = \mathbb{E}[y_c \mid X]$ denote the optimal target-history and joint-history predictors. They satisfy

$$\mathbb{E}\|y_c - m_c\|^2 - \mathbb{E}\|y_c - m_J\|^2 = \mathbb{E}\|m_J - m_c\|^2. \tag{10}$$

Equation (10) is the squared-error projection identity. Write $y_c - m_c = (y_c - m_J) + (m_J - m_c)$ and expand the squared norm. The cross term has expectation zero because $\mathbb{E}[y_c - m_J \mid X] = 0$ and $m_J - m_c$ is measurable with respect to X. Thus, under squared error, context's predictive value is a change in the conditional continuation. The identity establishes the population-level value of context, while m_c serves as the Bayes-optimal target-history predictor and ACT's zero-action forecast serves as its model-specific reference. To identify how context produces a prediction change, retain the target-history state and head and vary only the action; the next subsection proves LAB's exact factorization.

B.2 Proof of the state–action factorization

Base anchoring expresses each route output as a coordinatewise scaling of its original input, including at zero coordinates. Let Φ denote the standard Gaussian cumulative distribution function. In evaluation mode, the route scale is

$$s_0 = \mathbf{1}, \qquad s_{\ell+1} = s_\ell + g_\ell(u_\ell) \odot \Phi(u_0 \odot g_\ell(u_\ell)) \,. \tag{11}$$

For a single route, assume inductively that $u_\ell = u_0 \odot s_\ell$, with $s_0 = \mathbf{1}$. At evaluation, dropout is the identity. Using $\mathrm{GELU}(v) = v \odot \Phi(v)$ gives

$$\begin{aligned} u_{\ell+1} &= u_\ell + \mathrm{GELU}(u_0 \odot g_\ell(u_\ell)) \\ &= u_0 \odot [s_\ell + g_\ell(u_\ell) \odot \Phi(u_0 \odot g_\ell(u_\ell))] = u_0 \odot s_{\ell+1}. \end{aligned} \tag{12}$$

Thus, the factorization holds at every depth. Applying it to LAB gives the temporal and context outputs, together with the residual route's input and output:

$$\begin{aligned} T &= Z \odot S_t, \qquad C_x = Z \odot S_c, \\ E &= Z \odot (\mathbf{1} - S_t - S_c), \\ R_x &= Z \odot (\mathbf{1} - S_t - S_c) \odot S_r. \end{aligned} \tag{13}$$

The updated state sums the two direct outputs and the residual-route output, $U = T + C_x + R_x$; E is the residual-route input, not an additional term in this sum. Substituting the route outputs gives

$$\begin{aligned} U &= Z \odot [S_t + S_c + (\mathbf{1} - S_t - S_c) \odot S_r] \\ &= Z \odot [\mathbf{1} + (\mathbf{1} - S_t - S_c) \odot (S_r - \mathbf{1})] = Z \odot (\mathbf{1} + A), \end{aligned} \tag{14}$$

proving Proposition 1. Extracting the action uses only products and sums, with no division by Z. At a zero coordinate of Z, the action defined by this parameterization remains finite whenever the network computation is finite, and its contribution to the updated state is zero. This decomposition is canonical for the specified parameterization, although the same predictor can admit different state–action representations.

B.3 Proof of the low-rank local sensitivity bound

Factor-rank control bounds how much each route update can revise coordinatewise propagation. In each gate, the first two rank-r maps form the shared intermediate $q(u)$; its Jacobian right-multiplies the derivative of the final product and cannot increase its rank. The product has a value-transform branch and a sigmoid-gate branch. Its derivative therefore yields two terms, each of rank at most r. Thus each update contributes at most $2r$ rank, while direct differentiation through the base-state path contributes a diagonal term, as made explicit by the Jacobian below.

Proposition 2 (Local sensitivity of a low-rank route)**.** *For a depth-k evaluation-mode LAB route of width m, whose four gate maps per layer have factor rank at most r, the Jacobian with respect to the route input is $J_{\mathcal{B}} = \Lambda + K$, where Λ is diagonal and* $\mathrm{rank}(K) \le \min(m, 2kr)$.

Write the last line of the gate as $g(u) = v(q(u)) \odot s(q(u))$, with $v(q) = W_v q + b_v$ and $s(q) = \mathrm{sigmoid}(W_g q + b_g)$. Its Jacobian is

$$J_g = \left[\mathrm{diag}(s) W_v + \mathrm{diag}(v)\,\mathrm{diag}(s \odot [\mathbf{1} - s]) W_g\right] J_q. \tag{15}$$

Each summand inside brackets has rank at most r, so $\mathrm{rank}(J_g) \le 2r$, regardless of the intermediate derivative J_q. Biases do not change this bound.

Let $J_\ell = \partial u_\ell / \partial u_0$ and $v_\ell = u_0 \odot g_\ell(u_\ell)$. Differentiating Equation (4) in evaluation mode yields

$$J_{\ell+1} = J_\ell + \mathrm{diag}(\mathrm{GELU}'(v_\ell) \odot g_\ell(u_\ell)) + \mathrm{diag}(\mathrm{GELU}'(v_\ell) \odot u_0) J_{g_\ell} J_\ell. \tag{16}$$

The middle term is diagonal, and since $\mathrm{rank}(J_{g_\ell}) \le 2r$, each last term has rank at most $2r$. Starting from $J_0 = I$, define the cumulative diagonal contribution $D_k = I + \sum_{\ell=0}^{k-1} \mathrm{diag}(\mathrm{GELU}'(v_\ell) \odot g_\ell(u_\ell))$ and the residual $K = \sum_{\ell=0}^{k-1} \mathrm{diag}(\mathrm{GELU}'(v_\ell) \odot u_0) J_{g_\ell} J_\ell$, so that $J_k = D_k + K$. Subadditivity gives $\mathrm{rank}(K) \le \min(m, 2kr)$, proving Proposition 2. Thus, the bound applies to the residual in a single route's local prediction-state Jacobian.

B.4 Fixed-state context interventions

Because the history encoder operates independently on each variable, replacing histories other than c while holding x_c fixed leaves z_c unchanged. With recipient RevIN statistics fixed, subtracting the forecasts in Equation (6) gives Equation (7). In dataset-standardized output space, the difference is additionally multiplied by the recipient inverse-RevIN scale; with RevIN disabled, no additional transform applies.

Numerical replay validates the pathway in the trained predictor. Across 25 checkpoints and 166,160 full-test window–seed evaluations, native forward replay reproduces the original computation; factorized replay matches the native predictions numerically, with a maximum discrepancy of 7.63×10^{-6}. Across 1,600 context interventions, the target-state change is zero and the maximum context-through-action replay discrepancy is 7.15×10^{-7}. The separate Solar/Weather physical panel gives a maximum replay error below 10^{-6}. The replacement and transfer experiments below assess predictive utility beyond this pathway check.

B.5 Interpretation of a latent action

We call a_c a latent action because it specifies the model-internal operation $\mathcal{T}_{a_c}(z_c) = z_c + z_c \odot a_c$: joint history supplies the coefficient that updates the independently encoded target-history state, and forecast error trains the generator with the rest of the predictor. The shared head maps the updated state to a conditional forecast; δ_c denotes its difference from the zero-action forecast. This operational forecasting definition gives the action a precise semantic role: it is a replaceable operator on a retained state whose forecast effect can be measured and replayed. The explicit operation can be replaced and replayed while the reference state and head remain fixed.

For fixed coefficients, the action family has identity $\mathcal{T}_0(z) = z$ and the following closed composition rule:

$$\mathcal{T}_b(\mathcal{T}_a(z)) = \mathcal{T}_{a+b+a\odot b}(z). \tag{17}$$

This follows by expanding $z \odot (\mathbf{1} + a) \odot (\mathbf{1} + b)$. The rule describes sequential operations on the same prediction state with their coefficients held fixed; it does not combine two context groups and recompute a coefficient from the combined context.

The interventions vary distinct parts of the predictor: setting $a_c = 0$ tests the applied update's contribution; replacing X_{-c} while retaining x_c tests context-to-action-to-forecast dependence; context readouts measure which labels are linearly decodable from a_c; and state-compatible transfer compares native action coefficients with nearest-state, random-donor, and mean-action replacements. Equation (7) gives the forecast difference for coefficient replacement with the state and forecast head fixed.

With the state and normalization fixed, define $F_{z_c}(a) = \mathcal{N}_c^{-1}(W_{\text{out}}[z_c + z_c \odot a] + b_{\text{out}})$. Its increment is linear: $F_{z_c}(a + \Delta a) - F_{z_c}(a) = M_c \Delta a$, where M_c combines $W_{\text{out}}\,\text{diag}(z_c)$ with the inverse-normalization scale. Consequently, specified action increments have additive effects on the forecast under this fixed-state map. This additivity describes the forecast response; context edits are evaluated separately through their generator-induced action increments.

B.6 Operational content of the action

Three fixed-state diagnostics use 25 frozen, validation-selected $L = H = 96$ checkpoints from ETTh1, ETTm1, Weather, ECL, and Traffic, with seeds 42–46; test windows are sampled every sixth window. The recipient is Weather air temperature `T (degC)` (zero-based channel 1), ETT oil temperature `OT` (channel 6), ECL channel 320, or Traffic channel 861. The corresponding test-window counts are 1,741 for Weather, 465 for ETTh1, 1,905 for ETTm1, 861 for ECL, and 569 for Traffic.

Using each recipient's original normalization statistics, we replay its native action coefficient at the fixed target state through the same forecast head and compare the resulting forecast with the checkpoint's native forward forecast. The maximum absolute discrepancy across all 25 runs is 3.34×10^{-6}. The retained target-history state changes by zero. These target-channel populations differ from the all-variable benchmark and update-replacement panels.

For these prediction probes, the non-target channels define the context, with donor contexts supplied by encoded histories from sampled test windows. Holding the target's encoded history fixed, we define the target-only baseline by setting all non-target rows to zero.

Graded context replacement. This test measures how the forecast at a fixed target state responds to replacing the encoded context. Let Z_{-c} collect the encoded histories of all non-target variables. Interpolation sets $Z_{-c}(\lambda) = (1-\lambda)Z_{-c} + \lambda Z'_{-c}$ for $\lambda \in \{0, 1/4, 1/2, 3/4, 1\}$, where Z'_{-c} comes from a random permutation of sampled test windows. The same permutation is used at all five levels and may contain fixed points. The target state z_c remains fixed while the generator recomputes $a_c(\lambda)$, yielding the update $d_c(\lambda) = z_c \odot a_c(\lambda)$ and updated state $u_c(\lambda) = z_c + d_c(\lambda)$; the head maps this updated state to a forecast. Table 9 reports the seed-averaged MSE(λ)/MSE(0); its monotone column counts non-decreasing five-point curves. Mean endpoint ratios are 2.892 on ECL, 2.575 on Traffic, and 1.484 on Weather, with non-decreasing curves in three, four, and five of five seeds, respectively. On ETTh1 and ETTm1, the endpoint ratios are 1.000 and 1.003, with non-decreasing curves in two and zero of five seeds. These ratios quantify forecast sensitivity to encoded-context replacement at a fixed target state, not an environmental or physical causal effect.

Cross-variable transfer. For each recipient, we replace its action coefficient with another variable's coefficient while holding z_c, the forecast head, and the recipient's normalization statistics fixed. The donor coefficient comes either from the recipient's own window or from a random permutation of sampled test windows. We sample eight distinct non-target donor channels, or all six on ETT, without performance selection, and average donor-variable errors within each model. Table 9 reports the five-seed mean same-window MSE divided by the five-seed mean permuted-window MSE: a ratio below one means same-window coefficients yield lower recipient error, indicating greater predictive compatibility with shared contemporaneous context under the fixed head; a ratio above one favors permuted-window coefficients. This window-paired replacement measures predictive compatibility; the same-variable training-donor experiment in Appendix D tests history matching. Partition and donor-channel sampling use the model seed, while window permutations use a separate random stream seeded by the model seed plus 10,000.

Table 9: Context response and cross-variable transfer on frozen $H = 96$ checkpoints, seeds 42–46. Dose entries average seed-level MSE(λ)/MSE(0), interpolating every non-target encoded history toward a permuted test window. Mono. counts non-decreasing five-point curves. Transfer applies a different variable's action: ratio divides the mean same-window MSE by mean permuted-window MSE, and MSE reports the same-window value. Each row evaluates the target channel specified in Appendix B.6.

	Dose response MSE(λ)/MSE(0)					Transfer	
Dataset	1/4	1/2	3/4	1	Mono.	Ratio↓	MSE
ECL	1.012	1.565	2.442	2.892	3/5	0.83	0.421
Traffic	1.236	2.139	2.496	2.575	4/5	0.63	0.286
Weather	1.065	1.231	1.376	1.484	5/5	0.85	0.117
ETTh1	1.000	1.000	1.000	1.000	2/5	1.01	0.059
ETTm1	0.999	1.000	1.001	1.003	0/5	0.99	0.036

Context-group interaction. Partition the non-target variables into disjoint groups G_1, G_2. Let $Z^{(0)}$ retain only z_c, with all other rows set to zero; let $Z^{(1)}$ and $Z^{(2)}$ additionally retain the native rows in G_1 and G_2, respectively. Define $a_0 = \mathrm{Act}(Z^{(0)})_c$, $a_1 = \mathrm{Act}(Z^{(1)})_c$, $a_2 = \mathrm{Act}(Z^{(2)})_c$, and $a_{12} = \mathrm{Act}(Z)_c$. The additive reconstruction and its interaction residual are

$$a_{\mathrm{add}} = a_1 + a_2 - a_0, \qquad I_c = a_{12} - a_{\mathrm{add}}. \tag{18}$$

Subtracting a_0 counts the shared target-only action once. For a generator of the additive form $b(z_c) + f_1(z_c, Z_{G_1}) + f_2(z_c, Z_{G_2})$, $I_c = 0$. We use contiguous halves, interleaved halves, and four seeded random partitions. The relative action interaction is $R_a = \mathbb{E}\|I_c\|_2 / \mathbb{E}\|a_{12} - a_0\|_2$, with expectations taken over sampled windows within each model and partition. The corresponding forecast interaction is $R_y = \mathbb{E}\|M_c I_c\|_2 / \mathbb{E}\|M_c(a_{12} - a_0)\|_2$. These are ratios of expected vector norms. We average the six partitions within each seed and then summarize across five seeds. A denominator no larger than $\max(10^{-10}, 10^{-6}q)$ is flagged as degenerate, where q is the mean native-action norm in the corresponding space; no denominator is flagged in these runs. Forecast MSE under $a_{12}, a_0, 0$, and a_{add} tests whether the interaction changes prediction.

On ECL, Traffic, and Weather, the native action lowers MSE relative to a_0 in all five models and relative to a_{add} in all six partitions of every model (Table 10). Relative to target-only conditioning, seed-mean MSE decreases by 53.5%, 45.7%, and 32.8%, respectively; replacing the native action

Table 10: Context-group intervention results at fixed target state. Five-seed mean ± sample SD, with six partitions averaged within each model. (a) MSE uses the target channels in Appendix B.6; zero action removes the entire state update, whereas a_0 retains the action generated by target-only conditioning. (b) R_a and R_y quantify interaction relative to the context effect in action and forecast space. Native/additive counts compare native MSE with each partition reconstruction across the five models; partitions are not independent model replicates.

(a) Target-channel MSE↓

Dataset	Native a_{12}	Target-only a_0	Zero action	Additive a_{add}
ECL	0.242238 ± 0.009335	0.520474 ± 0.063983	0.454247 ± 0.008361	0.367273 ± 0.061482
Traffic	0.142646 ± 0.005455	0.262770 ± 0.019021	0.362569 ± 0.007102	0.233699 ± 0.044168
Weather	0.078763 ± 0.003290	0.117245 ± 0.015903	0.103366 ± 0.006063	0.087218 ± 0.007284
ETTh1	0.054075 ± 0.000287	0.054076 ± 0.000287	0.062121 ± 0.000257	0.054074 ± 0.000287
ETTm1	0.029666 ± 0.001029	0.029751 ± 0.000989	0.033288 ± 0.000400	0.029655 ± 0.001037

(b) Relative interaction and forecast comparison

Dataset	R_a	R_y	Native better than additive
ECL	0.743 ± 0.044	0.562 ± 0.059	30/30
Traffic	0.948 ± 0.051	0.943 ± 0.211	30/30
Weather	0.567 ± 0.064	0.431 ± 0.121	30/30
ETTh1	0.034 ± 0.015	0.044 ± 0.017	6/30
ETTm1	0.291 ± 0.186	0.233 ± 0.115	10/30

with a_{add} raises mean MSE by 51.6%, 63.8%, and 10.7%. Thus, on these datasets, the native action improves forecasts beyond the additive action. On ETT, context interventions yield only small forecast differences, although the nonzero native action still improves over zero action.

C Complete forecasting results

Tables 11–13 extend Section 4.1 to all four horizons and twelve reference models. ACT entries are five-run means at the validation-selected settings in Table 8; baseline entries are reference point estimates.

Table 11: Complete reference comparison: linear and mixer models. Lower MSE/MAE is better. ACT gives validation-selected five-seed means; baselines are reference point estimates. Avg. averages four horizons; bold marks panel-local minima, including ties.

		Ours		Linear						Mixer			
Dataset	H	ACT		MixLinear		DLinear		RLinear		TSMixer		TimeMixer	
		MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓
ETTh1	96	**0.368**	**0.390**	0.404	0.408	0.386	0.399	0.385	0.393	0.384	0.403	0.408	0.413
	192	**0.428**	**0.422**	0.438	0.423	0.434	0.428	0.436	**0.422**	0.444	0.435	0.457	0.442
	336	**0.476**	0.445	0.480	0.445	0.482	0.460	**0.476**	**0.442**	0.491	0.460	0.505	0.467
	720	0.502	0.480	**0.464**	**0.457**	0.504	0.502	0.478	0.467	0.505	0.485	0.492	0.478
	Avg.	**0.444**	0.434	0.447	0.433	0.452	0.447	**0.444**	**0.431**	0.456	0.446	0.466	0.450
ETTh2	96	**0.284**	**0.336**	0.304	0.347	0.329	0.384	0.290	0.340	0.304	0.353	0.293	0.342
	192	**0.368**	**0.389**	0.385	0.396	0.435	0.448	0.378	0.395	0.402	0.409	0.375	0.394
	336	0.414	0.425	0.426	0.430	0.563	0.526	0.430	0.439	0.444	0.445	**0.398**	**0.424**
	720	0.426	0.442	0.420	0.437	0.775	0.634	0.442	0.453	0.436	0.450	**0.406**	**0.432**
	Avg.	0.373	**0.398**	0.384	0.403	0.526	0.498	0.385	0.407	0.397	0.414	**0.368**	**0.398**
ETTm1	96	**0.311**	**0.351**	0.391	0.397	0.345	0.371	0.350	0.368	0.321	0.361	0.333	0.368
	192	**0.357**	**0.377**	0.415	0.406	0.383	0.394	0.388	0.386	0.370	0.388	0.376	0.393
	336	**0.388**	**0.399**	0.444	0.426	0.414	0.414	0.419	0.406	0.415	0.414	0.408	0.418
	720	**0.451**	**0.439**	0.496	0.454	0.474	0.453	0.480	0.440	0.497	0.461	0.493	0.464
	Avg.	**0.377**	**0.392**	0.437	0.421	0.404	0.408	0.409	0.400	0.401	0.406	0.403	0.411

Continued on next page

Dataset	H	Ours ACT MSE↓	Ours ACT MAE↓	Linear MixLinear MSE↓	Linear MixLinear MAE↓	Linear DLinear MSE↓	Linear DLinear MAE↓	Linear RLinear MSE↓	Linear RLinear MAE↓	Mixer TSMixer MSE↓	Mixer TSMixer MAE↓	Mixer TimeMixer MSE↓	Mixer TimeMixer MAE↓
ETTm2	96	**0.173**	**0.254**	0.199	0.281	0.186	0.282	0.182	0.265	0.183	0.267	0.183	0.267
	192	**0.234**	**0.294**	0.259	0.315	0.270	0.347	0.247	0.306	0.249	0.309	0.260	0.318
	336	**0.292**	**0.332**	0.317	0.350	0.362	0.414	0.309	0.344	0.310	0.347	0.309	0.346
	720	**0.390**	**0.392**	0.414	0.403	0.527	0.507	0.408	0.400	0.417	0.407	0.438	0.420
	Avg.	**0.272**	**0.318**	0.297	0.337	0.336	0.388	0.287	0.329	0.290	0.333	0.298	0.338
ECL	96	**0.147**	**0.243**	0.220	0.293	0.195	0.277	0.198	0.274	0.156	0.258	0.153	0.253
	192	**0.164**	**0.259**	0.206	0.281	0.194	0.280	0.198	0.277	0.174	0.274	0.169	0.270
	336	**0.180**	**0.274**	0.219	0.295	0.208	0.297	0.212	0.293	0.187	0.288	0.184	0.285
	720	0.226	0.311	0.260	0.330	0.243	0.330	0.254	0.326	0.216	0.309	**0.209**	**0.305**
	Avg.	**0.179**	**0.272**	0.226	0.300	0.210	0.296	0.216	0.293	0.183	0.282	**0.179**	0.278
Solar	96	**0.191**	**0.250**	0.515	0.446	0.285	0.372	0.305	0.329	0.214	0.264	0.234	0.279
	192	**0.208**	**0.267**	0.543	0.458	0.316	0.393	0.344	0.348	0.257	0.292	0.277	0.306
	336	**0.216**	**0.274**	0.556	0.456	0.352	0.413	0.386	0.364	0.280	0.307	0.284	0.307
	720	**0.214**	**0.272**	0.524	0.432	0.355	0.411	0.389	0.358	0.278	0.304	0.278	0.300
	Avg.	**0.207**	**0.266**	0.535	0.448	0.327	0.397	0.356	0.350	0.257	0.292	0.268	0.298
Traffic	96	**0.452**	**0.286**	0.663	0.394	0.650	0.398	0.646	0.386	0.487	0.338	0.472	0.316
	192	**0.458**	**0.288**	0.611	0.367	0.599	0.371	0.599	0.362	0.496	0.338	0.494	0.328
	336	**0.470**	**0.292**	0.617	0.369	0.607	0.375	0.607	0.366	0.514	0.349	0.518	0.347
	720	**0.502**	**0.310**	0.655	0.389	0.648	0.398	0.645	0.385	0.541	0.368	0.540	0.350
	Avg.	**0.470**	**0.294**	0.637	0.380	0.626	0.386	0.624	0.375	0.510	0.348	0.506	0.335
Weather	96	**0.155**	**0.202**	0.208	0.250	0.196	0.255	0.193	0.232	0.159	0.208	0.160	0.207
	192	**0.206**	**0.250**	0.252	0.283	0.238	0.297	0.236	0.268	0.214	0.254	0.226	0.265
	336	**0.264**	**0.292**	0.302	0.317	0.283	0.333	0.288	0.304	0.273	0.294	0.286	0.307
	720	**0.344**	**0.347**	0.372	0.361	0.348	0.385	0.359	0.350	0.349	0.348	0.372	0.358
	Avg.	**0.242**	**0.273**	0.284	0.303	0.266	0.318	0.269	0.289	0.249	0.276	0.261	0.284

The twelve references span linear (MixLinear, DLinear, RLinear), mixer (TSMixer, TimeMixer), Transformer (PatchTST, iTransformer), frequency/period (FilterNet, FreTS), and dependency-aware (TimeFilter, FACT, TimePro) families. For dataset d, family $\mathcal{F}$, and metric q, Table 2 reports $\min_{m\in\mathcal{F}} \frac{1}{4}\sum_{H\in\{96,192,336,720\}} q_{dHm}$. Each model is averaged over horizons before the family minimum is taken, so the MSE and MAE minima may come from different models. Baseline averages use per-horizon point estimates. For ACT, five seed-level test values are averaged at each horizon, then over horizons. Values are rounded to three decimals after aggregation; displayed ties count as first places, and bold entries in the complete tables mark panel-local minima.

Table 12: Complete reference comparison: Transformer and frequency/period models. Metrics, aggregation, and highlighting follow Table 11.

Dataset	H	Ours ACT MSE↓	Ours ACT MAE↓	Transformer PatchTST MSE↓	Transformer PatchTST MAE↓	Transformer iTransformer MSE↓	Transformer iTransformer MAE↓	Frequency/Period FilterNet MSE↓	Frequency/Period FilterNet MAE↓	Frequency/Period FreTS MSE↓	Frequency/Period FreTS MAE↓
ETTh1	96	**0.368**	**0.390**	0.383	0.402	0.384	0.403	0.384	0.399	0.400	0.409
	192	**0.428**	**0.422**	0.435	0.431	0.434	0.431	0.446	0.432	0.455	0.440
	336	0.476	**0.445**	**0.470**	0.452	0.482	0.457	0.487	0.449	0.496	0.460
	720	0.502	0.480	**0.479**	0.476	0.491	0.482	0.488	**0.471**	0.506	0.481
	Avg.	0.444	**0.434**	**0.442**	0.440	0.448	0.443	0.451	0.438	0.464	0.448
ETTh2	96	**0.284**	**0.336**	0.292	0.344	0.302	0.352	0.299	0.349	0.298	0.348
	192	**0.368**	**0.389**	0.373	0.395	0.379	0.399	0.375	0.398	0.382	0.399
	336	**0.414**	**0.425**	0.417	0.431	0.418	0.430	0.417	0.430	0.426	0.436
	720	**0.426**	**0.442**	0.434	0.450	0.427	0.447	0.428	0.444	0.448	0.457
	Avg.	**0.373**	**0.398**	0.379	0.405	0.382	0.407	0.380	0.405	0.389	0.410

Continued on next page

Dataset	H	Ours ACT		Transformer PatchTST		Transformer iTransformer		Frequency/Period FilterNet		Frequency/Period FreTS	
		MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓
ETTm1	96	**0.311**	**0.351**	0.377	0.424	0.337	0.371	0.330	0.367	0.358	0.394
	192	**0.357**	**0.377**	0.417	0.439	0.376	0.388	0.372	0.388	0.399	0.411
	336	**0.388**	**0.399**	0.465	0.466	0.423	0.414	0.402	0.408	0.433	0.439
	720	**0.451**	**0.439**	0.517	0.501	0.480	0.449	0.468	0.446	0.538	0.509
	Avg.	**0.377**	**0.392**	0.444	0.458	0.404	0.406	0.393	0.402	0.432	0.438
ETTm2	96	**0.173**	**0.254**	0.176	0.261	0.184	0.268	0.175	0.257	0.180	0.262
	192	**0.234**	**0.294**	0.242	0.305	0.252	0.312	0.242	0.301	0.247	0.306
	336	**0.292**	**0.332**	0.303	0.344	0.317	0.352	0.298	0.337	0.304	0.342
	720	**0.390**	**0.392**	0.402	0.402	0.411	0.405	0.399	0.398	0.406	0.402
	Avg.	**0.272**	**0.318**	0.281	0.328	0.291	0.334	0.279	0.323	0.284	0.328
ECL	96	0.147	**0.243**	0.206	0.309	**0.146**	0.244	0.152	0.248	0.184	0.271
	192	0.164	0.259	0.214	0.321	**0.162**	**0.256**	0.167	0.259	0.188	0.277
	336	**0.180**	**0.274**	0.218	0.325	**0.180**	**0.274**	0.184	0.278	0.203	0.294
	720	0.226	0.311	0.254	0.352	**0.213**	**0.305**	0.227	0.317	0.248	0.335
	Avg.	0.179	0.272	0.223	0.327	**0.175**	**0.270**	0.183	0.276	0.206	0.294
Solar	96	**0.191**	**0.250**	0.223	0.328	0.203	0.256	0.214	0.265	0.250	0.308
	192	**0.208**	**0.267**	0.246	0.353	0.233	0.271	0.249	0.287	0.268	0.328
	336	**0.216**	**0.274**	0.260	0.365	0.266	0.304	0.262	0.296	0.285	0.336
	720	**0.214**	**0.272**	0.246	0.350	0.254	0.286	0.265	0.298	0.269	0.315
	Avg.	**0.207**	**0.266**	0.244	0.349	0.239	0.279	0.248	0.287	0.268	0.322
Traffic	96	0.452	0.286	0.475	**0.277**	**0.427**	0.299	0.434	0.298	0.542	0.357
	192	0.458	0.288	0.489	**0.278**	**0.451**	0.302	0.459	0.305	0.537	0.358
	336	0.470	0.292	0.500	**0.291**	**0.464**	0.304	0.479	0.315	0.553	0.363
	720	**0.502**	0.310	0.535	**0.302**	0.506	0.324	0.514	0.331	0.590	0.380
	Avg.	0.470	0.294	0.500	**0.287**	**0.462**	0.307	0.472	0.312	0.556	0.365
Weather	96	**0.155**	**0.202**	0.165	0.211	0.168	0.211	0.164	0.210	0.167	0.213
	192	**0.206**	**0.250**	0.212	0.253	0.214	0.254	0.210	0.251	0.241	0.272
	336	**0.264**	**0.292**	0.268	**0.292**	0.273	0.296	0.268	0.293	0.269	0.295
	720	**0.344**	0.347	0.346	**0.344**	0.351	0.347	0.346	0.345	0.346	0.346
	Avg.	**0.242**	**0.273**	0.248	0.275	0.252	0.277	0.247	0.275	0.256	0.282

Table 13: Complete reference comparison: dependency-aware models. Metrics, aggregation, and highlighting follow Table 11.

Dataset	H	Ours ACT		TimeFilter		Dependency-aware FACT		TimePro	
		MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓
ETTh1	96	**0.368**	**0.390**	0.378	0.396	0.383	0.401	0.383	0.404
	192	**0.428**	**0.422**	0.445	0.428	0.437	0.431	0.434	0.431
	336	0.476	**0.445**	0.487	0.451	**0.474**	0.458	0.479	0.454
	720	0.502	0.480	**0.470**	**0.465**	0.525	0.496	0.486	0.481
	Avg.	**0.444**	**0.434**	0.445	0.435	0.455	0.447	0.446	0.443
ETTh2	96	**0.284**	**0.336**	0.293	0.341	0.292	0.342	0.300	0.349
	192	**0.368**	**0.389**	0.371	0.396	0.374	0.397	**0.368**	0.394
	336	**0.414**	**0.425**	0.420	0.429	0.434	0.437	0.421	0.433
	720	**0.426**	**0.442**	0.429	0.446	0.508	0.480	0.441	0.452
	Avg.	**0.373**	**0.398**	0.378	0.403	0.402	0.414	0.383	0.407
ETTm1	96	**0.311**	**0.351**	0.322	0.360	0.321	0.359	0.336	0.371
	192	**0.357**	**0.377**	0.367	0.386	0.369	0.384	0.376	0.391
	336	**0.388**	**0.399**	0.392	0.405	0.404	0.411	0.406	0.409
	720	**0.451**	**0.439**	0.456	0.446	0.478	0.454	0.471	0.447
	Avg.	**0.377**	**0.392**	0.384	0.399	0.393	0.402	0.397	0.405

Continued on next page

Dataset	H	Ours ACT		TimeFilter		Dependency-aware FACT		TimePro	
		MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓
ETTm2	96	0.173	0.254	0.178	0.263	**0.172**	**0.252**	0.178	0.260
	192	**0.234**	**0.294**	0.242	0.303	0.247	0.306	0.242	0.303
	336	**0.292**	**0.332**	0.304	0.343	0.302	0.340	0.300	0.339
	720	**0.390**	**0.392**	0.413	0.409	0.426	0.410	0.401	0.401
	Avg.	**0.272**	**0.318**	0.284	0.330	0.287	0.327	0.280	0.326
ECL	96	0.147	0.243	0.143	0.238	**0.142**	0.238	**0.142**	**0.237**
	192	0.164	0.259	0.162	0.255	0.160	0.253	**0.158**	**0.252**
	336	0.180	0.274	0.176	0.271	0.177	0.270	**0.173**	**0.268**
	720	0.226	0.311	0.224	0.313	0.213	0.305	**0.208**	**0.299**
	Avg.	0.179	0.272	0.176	0.269	0.173	0.267	**0.170**	**0.264**
Solar	96	**0.191**	0.250	0.202	0.240	0.203	0.250	0.193	**0.230**
	192	**0.208**	0.267	0.236	0.270	0.257	0.297	0.251	**0.263**
	336	**0.216**	**0.274**	0.244	0.281	0.254	0.284	0.250	0.275
	720	**0.214**	**0.272**	0.246	0.283	0.258	0.291	0.268	0.285
	Avg.	**0.207**	0.266	0.232	0.269	0.243	0.281	0.241	**0.263**
Traffic	96	0.452	0.286	**0.407**	**0.269**	0.448	0.287	0.435	0.286
	192	0.458	0.288	**0.431**	**0.273**	0.466	0.292	0.442	0.282
	336	0.470	0.292	0.468	0.290	0.490	0.299	**0.462**	**0.288**
	720	0.502	0.310	**0.484**	**0.298**	0.516	0.316	0.508	0.313
	Avg.	0.470	0.294	**0.448**	**0.283**	0.480	0.299	0.462	0.292
Weather	96	**0.155**	**0.202**	0.165	0.209	0.166	0.210	0.166	0.208
	192	**0.206**	**0.250**	0.215	0.255	0.210	0.252	0.214	0.254
	336	**0.264**	**0.292**	0.269	0.294	0.269	0.294	0.271	0.295
	720	**0.344**	0.347	0.351	0.347	0.351	0.348	0.349	**0.346**
	Avg.	**0.242**	**0.273**	0.250	0.276	0.249	0.276	0.250	0.276

D Update direction and recipient compatibility

This section distinguishes applied-update tests on D, coefficient-transfer tests on A at a fixed recipient state and head, and linear readouts of future summaries from z_c, a_c, and d_c.

D.1 Direction and pairing at matched update magnitude

Averaged over the five trained models on each dataset, native updates yield lower MSE and MAE than zero, norm-matched noise, or shuffled updates on all four datasets (Table 14). For each recipient, the applied update is $D = U - Z$ on its retained predictive state Z. Zeroing isolates the update's contribution, norm-matched noise tests its direction beyond magnitude, and shuffling tests whether pairing it with its own Z matters beyond preserving the empirical update distribution. The evaluation covers five end-to-end trained models per dataset (20 total), with seeds 42–46 and $L = H = 96$, $d = 512$, and $k = 3$. Weather uses rank 4, Solar rank 8 without RevIN, and ECL and Traffic dense maps. Each run retains its original training recipe and validation-selected checkpoint.

We evaluate every test window and variable with all parameters fixed, replacing the learned update D at prediction state Z with D'. The trained forecast head maps $Z + D'$ to a forecast, which the recipient's inverse normalization decodes. Setting $D' = 0$ removes the update within this trained predictor; it is distinct from training a predictor without an update.

For each recipient window i and variable c, draw $\epsilon_{ic} \sim \mathcal{N}(0, I_d)$. Latent-norm matching sets the replacement applied update to a Gaussian direction with the same norm as the native applied update:

$$d_{ic}^{\text{latent}} = \frac{\|d_{ic}\|_2}{\|\epsilon_{ic}\|_2}\epsilon_{ic}. \tag{19}$$

Let $F_{ic}(v)$ denote the forecast head followed by recipient inverse normalization applied to latent vector v, and define $\delta_{ic} = F_{ic}(z_{ic} + d_{ic}) - F_{ic}(z_{ic})$ as the native forecast correction. Forecast-norm matching rescales the same Gaussian direction to match the norm of this correction:

$$d_{ic}^{\text{forecast}} = \frac{\|\delta_{ic}\|_2}{\|F_{ic}(z_{ic} + \epsilon_{ic}) - F_{ic}(z_{ic})\|_2}\epsilon_{ic}. \tag{20}$$

Because the fixed-recipient decoder is affine, this replacement produces a forecast correction with the same norm as the native forecast correction. Both conditions replace the applied latent update rather than perturbing targets; a zero norm of the quantity being matched yields a zero replacement. Norms are taken over the 512 latent coordinates or 96 forecast positions, separately for each recipient and variable.

The shuffled condition assigns $D'_i = D_{\pi(i)}$ using one global permutation of complete multivariate updates, preserving their empirical latent distribution and within-donor cross-variable structure. Starting from a random permutation, the procedure visits initially too-close recipients in increasing index order and accepts a swap only when both resulting recipient–donor pairs meet the separation $|i - \pi(i)| \geq L + H = 192$. It rejects sequences with at most $4(L + H) = 768$ windows and allows at most 10,000 randomly sampled swap partners per recipient; if none yields a valid swap, it stops with an error rather than relaxing the constraint. Thus, paired input-plus-target spans share no raw time point. Permutation selection uses window indices, not target values or forecast outcomes. Because donors may occur later in the test sequence, this is an offline diagnostic of update–history pairing, not an intervention on the forecasting environment. The recipient state, head, and normalization remain fixed.

We use five fixed random repetitions, seeded 2026091100–2026091104 with dataset offsets 0, 100, 200, and 300 for Weather, Solar, ECL, and Traffic. Permutations use the corresponding seed plus 500000. Randomness is shared across model seeds within a dataset, and the two noise conditions share Gaussian directions. Repetitions are averaged within each model before computing the five-model mean and sample variance $s^2 = \sum_{s=1}^{5}(q_s - \bar{q})^2/4$. The original benchmark test splits are retained. Together with the alignment decomposition in Equation (8), these replacement results support the interpretation that predictive gains depend on pairing error-directed corrections with the recipient state.

Table 14: Native updates outperform all replacements in both metrics across all 20 models. Full-test evaluation at $L = H = 96$ reports means and sample variances over five models per dataset; variance columns display $10^6 s^2$. Random repetitions are averaged within each model first.

		MSE↓		MAE↓	
Dataset	Applied update	Mean	Variance	Mean	Variance
Weather	Native	**0.15471**	1.41	**0.20157**	0.91
	Zero	0.18112	11.08	0.22941	12.90
	Noise, latent-norm matched	0.18291	10.41	0.23149	12.48
	Noise, forecast-norm matched	0.21580	12.00	0.25979	14.23
	Shuffled	0.23597	286.18	0.26534	57.33
Solar	Native	**0.19113**	76.05	**0.25295**	27.49
	Zero	0.25606	22.93	0.31573	84.29
	Noise, latent-norm matched	0.26292	26.87	0.32386	85.51
	Noise, forecast-norm matched	0.36660	374.83	0.40630	136.47
	Shuffled	0.36340	450.65	0.40262	109.71
ECL	Native	**0.14635**	0.17	**0.24154**	0.38
	Zero	0.20805	5.85	0.28898	23.77
	Noise, latent-norm matched	0.23495	13.08	0.31044	31.85
	Noise, forecast-norm matched	0.29148	27.66	0.34840	60.57
	Shuffled	0.29861	12.14	0.34297	26.00
Traffic	Native	**0.44750**	4.91	**0.28579**	11.17
	Zero	0.72448	133.85	0.41227	7.29
	Noise, latent-norm matched	0.77761	155.51	0.47295	21.94
	Noise, forecast-norm matched	0.98199	409.50	0.61028	66.47
	Shuffled	1.02048	331.93	0.56858	36.01

D.2 Action-coefficient replacement and transfer

Transferring coefficients at a fixed recipient state. At a fixed recipient state Z, coefficient replacement substitutes the donor coefficient A' and recomputes the applied update as $Z \odot A'$. Update replacement instead retains Z as the base state but substitutes the donor's already-applied update for the recipient update D, rather than recomputing it from Z. Both interventions retain the encoder, forecast head, recipient Z, and normalization statistics, isolating the effect of transferred context on

the forecast at the same reference state. This panel uses the $H = 96$ settings in Table 8 for ETTh1, ETTm1, Weather, ECL, and Traffic, with seeds 42–46, separate from the update-replacement models above. Each run's validation MSE selects its checkpoint. Training-mean and donor actions come from a fixed bank of 256 training windows, preserving variable identity.

Full-test replacement. The native, zero, training-mean, and random-action conditions in Table 15 are evaluated over every test window and variable. Across five datasets and five seeds, there are three replacement comparisons for each of 25 dataset–seed pairs, or 75 in total. Native actions achieve lower full-test MSE in all 75. Setting the action to zero feeds Z to the trained forecast head, while random replacement samples an action from the training bank.

Table 15: Full-test action replacement at $L = H = 96$. MSE↓, five-seed mean ± sample standard deviation. The trained predictor and recipient state remain fixed while the action is replaced.

Dataset	Native A	$A = 0$	Training-mean A	Random training A
ETTh1	0.368 ± 0.003	0.510 ± 0.007	0.394 ± 0.002	0.414 ± 0.004
ETTm1	0.311 ± 0.010	0.362 ± 0.030	0.363 ± 0.009	0.463 ± 0.043
Weather	0.155 ± 0.001	0.180 ± 0.001	0.169 ± 0.002	0.181 ± 0.004
ECL	0.147 ± 0.001	0.207 ± 0.003	0.201 ± 0.001	0.297 ± 0.021
Traffic	0.452 ± 0.004	0.727 ± 0.005	0.689 ± 0.004	0.785 ± 0.011

State-compatible transfer. We evaluate state-compatible transfer on non-overlapping forecast windows, sampling four variable indices uniformly from each dataset's variable axis before evaluation. For each recipient variable, we select from the training bank the donor whose history state z_c is nearest to the recipient's, without using the recipient future. On this sampled population, nearest-state transfer yields lower MSE than random transfer for every dataset and seed (Table 16). On Traffic, nearest-state transfer gives an MSE of 0.482, compared with 0.851 for random transfer and 0.389 for native actions. Native, nearest-state, random, and mean-action conditions use this same population. The mean action provides a fixed-update reference, while nearest-state transfer tests whether matching the recipient state improves the predictive utility of a transferred action.

Table 16: Action transfer with the recipient state fixed. MSE↓, five-seed mean ± sample standard deviation, on four fixed channels and non-overlapping windows.

Dataset	Native	Nearest-state donor	Random donor	Mean action
ETTh1	0.388 ± 0.003	0.401 ± 0.005	0.417 ± 0.005	0.404 ± 0.006
ETTm1	0.443 ± 0.019	0.462 ± 0.014	0.673 ± 0.085	0.513 ± 0.019
Weather	0.137 ± 0.002	0.143 ± 0.002	0.165 ± 0.009	0.152 ± 0.003
ECL	0.323 ± 0.015	0.426 ± 0.012	0.578 ± 0.050	0.389 ± 0.005
Traffic	0.389 ± 0.006	0.482 ± 0.004	0.851 ± 0.023	0.669 ± 0.008

Complementary future-summary information. For these future-summary probes, the state representation and latent action carry complementary predictive information: joint features improve slope readout over the state alone on Weather, ECL, and Traffic, whereas state features perform better for level change on ETTh1 and ECL (Table 17). The probes use a 16-dimensional PCA representation of z_c, a_c, or d_c; the matched-budget joint probe concatenates eight coordinates from z_c and eight from a_c. Standardization and PCA are fitted on training, and validation selects ridge regularization. Targets use window-history scale followed by training standardization. The reported test MSE is for these probe targets. Near-constant histories enlarge the scaled targets, particularly on ETTm1, so comparisons are made within each dataset and target.

Within-variable action geometry. For each checkpoint and variable, we fit PCA separately to the observed actions. Averaged over checkpoints and variables within each dataset, the first 16 components explain 84.26%, 83.41%, 79.17%, 83.86%, and 60.90% of action variance on ETTh1, ETTm1, Weather, ECL, and Traffic, respectively. These means summarize empirical within-variable variation, not a global rank bound; Proposition 2 bounds a route's local input sensitivity.

Table 17: Linear readouts of future summaries: test MSE↓ on probe targets standardized using training statistics, five-seed mean ± SD (lower is better). PCA is fitted on training windows and ridge regularization is chosen on validation. Each representation has 16 PCA coordinates; $z_c^{[8]} \oplus a_c^{[8]}$ concatenates eight coordinates from each source. These probe errors are not forecasting MSE.

Dataset	Target	$z_c^{[16]}$	$a_c^{[16]}$	$d_c^{[16]}$	$z_c^{[8]} \oplus a_c^{[8]}$
ETTh1	level change	0.240 ± 0.012	0.303 ± 0.030	0.303 ± 0.016	0.312 ± 0.018
ETTh1	slope	0.791 ± 0.019	0.751 ± 0.021	0.840 ± 0.022	0.788 ± 0.019
ETTm1	level change	120.617 ± 0.197	122.233 ± 0.814	120.651 ± 0.073	121.938 ± 0.645
ETTm1	slope	9.139 ± 0.054	9.324 ± 0.135	9.111 ± 0.018	9.279 ± 0.067
Weather	level change	0.730 ± 0.021	0.853 ± 0.032	0.872 ± 0.157	0.713 ± 0.025
Weather	slope	0.751 ± 0.020	0.800 ± 0.068	0.896 ± 0.107	0.723 ± 0.020
ECL	level change	0.164 ± 0.012	0.395 ± 0.072	0.390 ± 0.038	0.235 ± 0.031
ECL	slope	0.318 ± 0.007	0.369 ± 0.035	0.320 ± 0.014	0.299 ± 0.010
Traffic	level change	0.182 ± 0.003	0.189 ± 0.025	0.179 ± 0.022	0.181 ± 0.021
Traffic	slope	0.544 ± 0.008	0.628 ± 0.089	0.619 ± 0.043	0.373 ± 0.028

E Capacity and action-generator controls

These comparisons separate interface transferability from LAB's structural contribution. Direct scaling tests the shared state–action interface; depth, rank, and placement study LAB's route structure and capacity.

E.1 Generators at the selected ACT configurations

Within each dataset, we compare LAB and direct scaling at that dataset's validation-selected ACT configuration, matching parameter count, linear multiply–accumulate count, encoder/head initialization, data, and training recipe. The comparison comprises 40 runs across Weather, Solar, ECL, and Traffic, with two generators and seeds 42–46. All runs use $L = H = 96$, $d = 512$, and $k = 3$; Weather uses rank 4, Solar rank 8, and ECL and Traffic dense maps. Training uses batch size 32 and AdamW with learning rate 10^{-3} and zero weight decay. Weather, ECL, and Traffic use RevIN, at most 50 epochs, and patience 10; Solar uses no RevIN, at most 30 epochs, and patience 3. Each run restores its minimum-validation-MSE checkpoint.

The direct generator reuses the initialized gate modules in sequential history-feature, variable, and history-feature stacks. Each stack applies $h_{\ell+1} = h_\ell + \mathcal{D}[\mathrm{GELU}(g_\ell(h_\ell))]$, and its final output A gives the action-updated forecasting state $U = Z \odot (\mathbf{1} + A)$, which the forecasting head maps to a forecast. Both arms use the same state–action interface, with different scale generators and no independent affine shift. Table 18 reports five-seed means and sample variances. LAB leads on ECL, direct scaling on Traffic, the results are close on Weather, and Solar splits by metric. Across these tested configurations, both generators implement the same interface.

Table 18: Matched action generators at validation-selected ACT configurations, $L = H = 96$, seeds 42–46. Sample variance is scaled by 10^6; lower mean MSE and MAE are better. Both generators match parameter count and linear multiply–accumulate cost within each dataset.

		MSE↓		MAE↓	
Dataset	Generator	Mean	$10^6 s^2$	Mean	$10^6 s^2$
Weather	LAB	0.15471	1.41	0.20157	0.91
	Direct scaling	0.15456	2.28	0.20148	2.51
Solar	LAB	0.19113	76.05	0.25295	27.49
	Direct scaling	0.18748	74.76	0.25634	86.12
ECL	LAB	0.14635	0.17	0.24154	0.38
	Direct scaling	0.14953	6.35	0.24584	5.98
Traffic	LAB	0.44750	4.91	0.28579	11.17
	Direct scaling	0.44339	40.82	0.28300	7.55

Table 19: Depth at $r = 4$. MSE↓ averages four horizons and three seeds; $k = 0$ gives no action.

Dataset	$k = 0$	$k = 1$	$k = 2$	$k = 3$
ETTh1	0.4490	0.4434	0.4410	**0.4391**
ETTh2	0.3761	0.3757	**0.3752**	**0.3752**
ETTm1	0.3829	**0.3782**	0.3791	0.3786
ETTm2	0.2791	0.2747	**0.2723**	0.2727
Weather	0.2482	0.2431	0.2423	**0.2416**
ECL	0.1914	0.1823	0.1817	**0.1778**
Traffic	0.4991	0.4855	0.4785	**0.4765**

Table 20: Rank at $k = 2$. MSE↓ averages four horizons and three seeds at unchanged settings.

Dataset	$r = 2$	$r = 4$	$r = 8$	$r = 16$
ETTh1	0.4434	0.4410	0.4399	**0.4366**
ETTh2	**0.3747**	0.3752	0.3755	0.3766
ETTm1	0.3793	0.3791	**0.3743**	0.3758
ETTm2	0.2750	**0.2723**	0.2729	0.2733
Weather	**0.2421**	0.2423	0.2428	0.2423
ECL	0.1835	0.1817	**0.1801**	0.1839
Traffic	0.4844	0.4785	0.4773	**0.4749**

E.2 Depth and rank as capacity choices

Relative to the separately trained $k = 0$ no-update baseline, whose action is identically zero, every tested positive depth yields lower mean MSE on all seven datasets, although the depth with the lowest mean varies by dataset (Table 19). The sweep fixes rank $r = 4$, width $d = 512$, and lookback $L = 96$, and evaluates depths $k \in \{0, 1, 2, 3\}$. Each dataset–depth entry averages MSE over four horizons and seeds 42–44, giving twelve runs per entry. Within each dataset, the learning rate, normalization, optimizer, and stopping rule are fixed; each run restores its minimum-validation-MSE checkpoint. All depths are reported without test-based selection.

The rank sweep fixes $k = 2$ and varies $r \in \{2, 4, 8, 16\}$, holding width, lookback, horizons, and seeds fixed (Table 20). All other training settings are fixed within each dataset. Among the tested ranks, $r = 8$ gives the lowest mean MSE on ECL, while $r = 16$ does so on Traffic. These results motivate validation-based rank selection; Appendix G reports training-plus-validation epoch time and parameter count for the profiled rank settings.

E.3 Direct scaling and affine modulation

This panel compares direct scaling, an added affine shift, and gate factorization at the fixed depth $k = 2$. It uses Weather and Traffic at $L = H = 96$ and $d = 512$, with seeds 42–46; rank $r = 16$ applies only to the low-rank LAB arm. All four arms use the same history-encoder and forecast-head architectures, RevIN, batch size 32, and AdamW with constant learning rate 10^{-3} and zero weight decay. Each arm is trained independently for at most 50 epochs; training stops after ten epochs without validation improvement, and we restore the minimum-validation-MSE checkpoint. These fixed-configuration generator ablations are separate from the selected-configuration panel above.

The *direct scaling* generator is a matched control that replaces LAB's base-anchored routes with three sequential gated residual stacks, applied over history features, variables, and history features, respectively. Each stack uses $h_{\ell+1} = h_\ell + \mathcal{D}[\mathrm{GELU}(g_\ell(h_\ell))]$, with the same gate modules as dense LAB. In order, the stacks implement $Z \to h_t \to h \to A$: writing their outputs as $h_t = F_t(Z)$, $h = F_c(h_t^\top)^\top$, and $A = F_r(h)$, the final state is $U = Z \odot (\mathbf{1} + A)$. This zero-shift, FiLM-style conditional scaling adapts the forecasting state. The control matches dense LAB in joint-history conditioning, multiplicative update, parameter count, and encoder/head initialization, changing only the scale generator.

The *direct affine* extension adds an independent shift, $\beta = B(h)$, predicted from the shared joint-history trunk by a width-d affine map. The resulting prediction state is $U = Z \odot (\mathbf{1} + A) + \beta$, which the shared forecasting head maps to a forecast. The shift map adds $d^2 + d = 262{,}656$ parameters; the zero-shift direct-scaling control uses the same gate modules as dense LAB and matches its parameter count, isolating the generator choice at equal budget. Rank-16 LAB preserves LAB's axis-structured routes while factorizing its gates at rank 16.

Table 21: Action generators at $L = H = 96$, $d = 512$, $k = 2$ (mean ± SD, five seeds).

Dataset	Generator	Params (M)↓	MSE↓	MAE↓
Weather	Dense LAB	4.305	**0.15563** ± 0.00090	0.20246 ± 0.00128
	Direct scaling	4.305	0.15620 ± 0.00209	**0.20195** ± 0.00271
	Direct affine	4.568	0.15589 ± 0.00088	0.20242 ± 0.00096
	Rank-16 LAB	0.375	0.15644 ± 0.00201	0.20334 ± 0.00255
Traffic	Dense LAB	10.254	0.45630 ± 0.00443	0.28701 ± 0.00151
	Direct scaling	10.254	0.44790 ± 0.00485	**0.27892** ± 0.00668
	Direct affine	10.517	0.46619 ± 0.00462	0.30118 ± 0.00140
	Rank-16 LAB	0.599	**0.44493** ± 0.00430	0.28558 ± 0.00206

In the five-seed comparison (seeds 42–46) at $L = H = 96$ on Weather and Traffic, the contextual forecasting interface accommodates direct scaling as well as LAB, with outcomes varying by dataset and metric (Table 4.4). Weather's four mean MSE values span 0.15563–0.15644. On Traffic, direct scaling lowers dense LAB's mean MSE by 0.00840 and mean MAE by 0.00809; it also gives the lowest mean MAE on both datasets. Rank-16 LAB gives the lowest mean MSE among the four Traffic generators, 0.44493, with 0.599M parameters—the fewest of the four (10.254M for dense LAB and direct scaling, and 10.517M for the affine arm). The affine arm has higher mean MSE and MAE than direct scaling on Traffic, so the independent shift adds no predictive gain in this comparison. Appendix G reports the computational cost of dense and factorized LAB.

E.4 Where the history-feature update is applied

Table 22: Update placement on ECL at $L = H = 96$, $d = 512$, $k = 2$, rank 8. Mean ± SD over three seeds. Only the history-feature route moves in the pre-encoder variant.

Variant	Parameters↓	MSE↓	MAE↓
ACT: history-state update	282,474	0.14755 ± 0.00131	0.24342 ± 0.00159
History-feature route = identity	212,842	0.14799 ± 0.00169	0.24409 ± 0.00173
History-feature route before encoder	225,898	0.14691 ± 0.00184	0.24290 ± 0.00181
Current-state factor in each update	282,474	0.15062 ± 0.00223	0.24561 ± 0.00081

The placement comparison uses ECL at $L = H = 96$, $d = 512$, $k = 2$, rank 8, batch size 32, learning rate 10^{-3}, and RevIN. All four variants use seeds 42–44 and validation-selected checkpoints, with at most 30 epochs and stopping after three epochs without validation improvement.

The default history-feature route updates the encoded history-state representation supplied to the forecasting head. The identity control replaces this route with $T = Z$, retaining the cross-variable and residual computations. The pre-encoder variant applies a width-L route to each raw normalized history before the shared encoder, then sets $T = Z$ after encoding; its cross-variable and residual routes remain unchanged. The current-state-factor control retains the default placement but replaces u_0 with u_ℓ in every route update, leaving the placement and number of parameters unchanged.

The pre-encoder and history-state placements both give accurate forecasts (Table 22). Applying the update to the encoded history state places the revision in the coordinates read by the forecast head, so the fixed-state tests in Section 4.2 isolate its predictive effect while holding the recipient state and forecast head fixed. In this setting, the original-state factor has slightly lower MSE than the evolving-state factor (0.14755 ± 0.00131 versus 0.15062 ± 0.00223); LAB's structured route parameterization gives exact action extraction.

F Similar representations can contribute differently

Similar representations can still make different predictive contributions. Residual input/output centered kernel alignment (CKA), a measure of representation similarity, ranges from 0.81 to 0.996 across the eight datasets, yet replacing the residual route's learned output with its input (the identity replacement) raises test-set MSE on every dataset. The five-seed mean of the per-checkpoint test-MSE increase, normalized by the identity-replacement MSE, ranges from 13.4% to 37.9% across datasets. We pair route geometry with these frozen-checkpoint replacement tests to distinguish representational similarity from predictive contribution.

The panel uses each dataset's validation-selected $L = H = 96$ configuration (Table 8) with seeds 42–46, yielding 40 frozen checkpoints, each with the complete width-512 LAB computation. Route-

geometry measurements characterize representation similarity, while frozen-checkpoint replacement tests assess the residual route's contribution to forecasting future targets.

The implementation names `feats_cross`, `times_cross`, and `res_cross` denote the history-feature route T, the cross-variable route C_x, and the residual route R_x, respectively. Routes T and R_x operate over the d history features, whereas C_x operates over the C variables after transposition. Diagnostic labels identify the computational axis of each route.

Representation similarity. For the history-feature route T, cross-variable route C_x, and residual route R_x, we collect 20 test batches, totaling 640 windows, and average each route's hidden outputs over variables. After centering each feature across windows, we compute CKA for each pair of routes. Let $V, W \in \mathbb{R}^{640\times d}$ denote the resulting matrices. Linear CKA is

$$\mathrm{CKA}(V, W) = \frac{\|V^\top W\|_F^2}{\|V^\top V\|_F \, \|W^\top W\|_F}. \tag{21}$$

CKA measures alignment between the sample-similarity matrices and is invariant to uniform scaling and orthogonal feature transformations (Kornblith et al., 2019). We also compute CKA between the residual-route output and its input. The replacement analysis measures how substituting the residual-route output affects prediction. Table 23 reports route geometry alongside these interventions.

Table 23: Route geometry and residual contribution at $L = H = 96$. Each dataset uses its validation-selected configuration (Table 8); entries are mean±sample SD over seeds 42–46. CKA uses 640 test windows per seed. MSE gains compare the native checkpoint with each replacement, using replacement MSE as the denominator.

	Linear CKA				Effective rank			MSE gain (%)↑	
Dataset	T, C_x	T, R_x	C_x, R_x	E, R_x	T	C_x	R_x	R_x=E	R_x=0
ETTh1	0.468±0.083	0.957±0.022	0.512±0.067	0.962±0.018	84.0±16.5	95.1±1.8	197.4±30.6	27.8±0.9	13.0±1.5
ETTh2	0.929±0.054	0.947±0.028	0.950±0.027	0.954±0.019	67.9±4.7	70.6±6.6	117.2±12.5	13.4±1.0	6.5±1.5
ETTm1	0.535±0.079	0.656±0.054	0.921±0.111	0.969±0.049	164.8±49.9	182.9±55.4	223.9±65.2	13.9±4.0	36.5±12.3
ETTm2	0.976±0.016	0.970±0.015	0.974±0.010	0.975±0.011	133.5±38.8	130.7±45.8	207.6±61.0	15.9±8.4	35.3±12.4
ECL	0.994±0.002	0.991±0.003	0.994±0.002	0.996±0.001	213.1±7.5	249.3±21.2	324.7±6.3	29.1±1.4	74.4±2.8
Solar	0.962±0.018	0.924±0.007	0.960±0.017	0.966±0.015	67.5±14.0	185.9±11.4	214.1±10.9	25.1±1.6	61.3±5.5
Traffic	0.916±0.016	0.645±0.031	0.778±0.026	0.809±0.020	169.6±4.2	318.7±4.3	341.4±3.1	37.9±0.8	35.0±1.2
Weather	0.829±0.051	0.850±0.040	0.947±0.016	0.964±0.010	207.2±5.6	305.6±6.0	307.3±6.6	13.8±1.0	22.0±5.1

Residual contribution. For each frozen checkpoint, we evaluate the complete test set with R_x replaced by E or by zero, without retraining. Identity replacement retains the signal entering the residual route, whereas zeroing removes that signal together with its learned transformation. The native-route MSE gain is $100(M_{\text{replacement}} - M_{\text{native}})/M_{\text{replacement}}$, with positive values favoring the native route. On ECL, the native route yields test-set MSE gains of 29.1% over identity replacement and 74.4% over zero replacement, showing that the learned residual transformation improves prediction relative to passing its input through unchanged.

Effective rank and route-label readout. Effective rank summarizes the dimensionality of a route's hidden representation, while the route-label readout measures how readily its outputs identify the route. For each route, we stack its hidden vectors across windows and variables into a matrix and compute its singular values σ_i. With $p_i = \sigma_i / \sum_j \sigma_j$, its effective rank is $\exp(-\sum_i p_i \log p_i)$. This equals the number of equally weighted singular directions when the nonzero singular values are equal; gate rank r, by contrast, specifies a parameter factorization. Table 24 also reports a multinomial logistic-regression classifier predicting route identity from variable-averaged hidden vectors collected over 400 test windows, for both a randomly initialized encoder and the trained checkpoint. The classifier is fit on 80% of the pooled vectors and evaluated on the remaining 20

Table 24: Route dimensionality and source-label readout at $L = H = 96$, at the same checkpoints as Table 23. Entries are mean±sample SD over seeds 42–46. Effective rank summarizes the spread of singular values; larger values indicate more representation directions. The probe assigns held-out hidden vectors to their source route, before and after model training.

	Effective rank			Route-label accuracy (%)↑	
Dataset	T	C_x	R_x	Untrained	Trained
ETTh1	84.0±16.5	95.1±1.8	197.4±30.6	100.0±0.0	99.9±0.2
ETTh2	67.9±4.7	70.6±6.6	117.2±12.5	99.7±0.5	99.0±0.6
ETTm1	164.8±49.9	182.9±55.4	223.9±65.2	100.0±0.0	100.0±0.0
ETTm2	133.5±38.8	130.7±45.8	207.6±61.0	98.6±1.3	99.8±0.4
ECL	213.1±7.5	249.3±21.2	324.7±6.3	100.0±0.0	100.0±0.0
Solar	67.5±14.0	185.9±11.4	214.1±10.9	100.0±0.0	100.0±0.0
Traffic	169.6±4.2	318.7±4.3	341.4±3.1	100.0±0.0	100.0±0.0
Weather	207.2±5.6	305.6±6.0	307.3±6.6	98.0±3.6	100.0±0.0

G TRAINING COST AND FORECASTING EFFICIENCY

The dense-versus-factorized LAB comparison measures a complete training-plus-validation epoch. We then report selected-model inference and training profiles, followed by configuration-search compute.

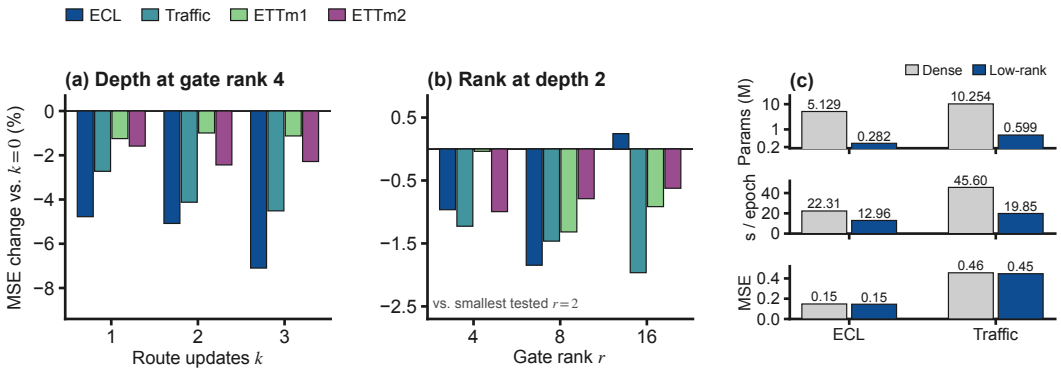


Figure 5: Complete LAB capacity and efficiency profile. (a,b) Structural depth and rank sweeps. (c) Dense versus low-rank LAB: parameters, training-plus-validation seconds per epoch, and complete H=96 test MSE at validation-selected ranks.

The selected Weather model uses 0.212 million parameters, processes 16,479 sequences/s, and occupies 37.7 MiB at batch size 32. The same validation-selected efficiency protocol gives 1.72× and 2.30× training-plus-validation speedups for rank-8 ECL and rank-16 Traffic, with 18.16× and 17.13× fewer parameters than dense LAB; Tables 25–26 give the complete profiles.

Profiles use an NVIDIA A100-SXM4-40GB (compute capability 8.0), Python 3.12.10, PyTorch 2.11.0 with CUDA 13.0, and Linux. ACT uses its selected $H = 96$ setting for each of ETTh1, ETTm1, Weather, ECL, and Traffic. DLinear and iTransformer are profiled in the same environment; iTransformer uses width 256 and two layers. The profiling seed is 42. Multi-seed accuracy measurements are specified separately below.

Static inference. The model and input tensors remain resident on the GPU. For each batch size (1 and 32), we exclude warmup and collect 200 timings, synchronizing CUDA immediately before and after every forward call. The p50 latency is the median of these calls, and throughput is batch size divided by mean latency. For each batch size, peak memory is the maximum CUDA-allocated tensor memory during that measurement, in MiB; it includes the resident model and input tensors, rather than reserved memory or incremental allocation above the resident baseline. Table 25 reports both operating points and both latency statistics.

Table 25: Static forward inference on one A100-SXM4-40GB, $L = H = 96$, seed 42. BS is batch size. Median and mean latency are milliseconds per batch from 200 synchronized post-warmup calls. Throughput is 1000 BS/mean latency in sequences/s; memory is total peak allocated MiB. Mean and median are both shown because occasional slow calls affect throughput.

Dataset	Model	BS	Median (ms)↓	Mean (ms)↓	Seq/s↑	Peak MiB↓
ETTh1	ACT	1	1.41	1.41	708	90.8
ETTh1	ACT	32	1.48	1.48	21,592	96.3
ETTh1	DLinear	1	0.31	0.32	3,157	16.6
ETTh1	DLinear	32	0.33	0.33	96,695	17.4
ETTh1	iTransformer	1	0.95	1.72	582	36.2
ETTh1	iTransformer	32	1.04	1.97	16,284	39.7
ETTm1	ACT	1	1.93	1.94	517	23.4
ETTm1	ACT	32	2.01	2.02	15,861	28.8
ETTm1	DLinear	1	0.31	0.31	3,194	16.6
ETTm1	DLinear	32	0.32	0.33	97,745	17.4
ETTm1	iTransformer	1	0.95	0.95	1,057	36.2
ETTm1	iTransformer	32	1.01	1.01	31,593	39.7
Weather	ACT	1	1.95	1.95	513	21.1
Weather	ACT	32	1.94	1.94	16,479	37.7
Weather	DLinear	1	0.69	0.67	1,488	17.4
Weather	DLinear	32	0.53	0.54	59,245	19.8
Weather	iTransformer	1	0.94	0.94	1,065	37.4
Weather	iTransformer	32	1.00	1.01	31,555	46.8
ECL	ACT	1	1.48	1.56	639	127.9
ECL	ACT	32	15.95	16.00	2,000	380.3
ECL	DLinear	1	0.31	0.31	3,197	32.8
ECL	DLinear	32	0.33	0.34	93,984	70.2
ECL	iTransformer	1	0.92	0.92	1,085	67.8
ECL	iTransformer	32	5.37	5.38	5,950	524.0
Traffic	ACT	1	2.28	2.28	438	278.4
Traffic	ACT	32	51.84	51.86	617	950.0
Traffic	DLinear	1	0.33	0.33	3,042	76.9
Traffic	DLinear	32	0.49	0.49	64,958	175.0
Traffic	iTransformer	1	1.15	1.25	797	188.9
Traffic	iTransformer	32	18.63	18.65	1,716	3181.8

Training and complete-test inference. Training timing spans from the end of model, data, and trainer setup through restoration of the best checkpoint. The reported total therefore includes training, validation, early stopping, and checkpoint restoration; mean seconds per completed epoch is this elapsed time divided by the number of completed epochs. Profiling runs allow up to 50 epochs and stop after ten epochs without validation improvement, so the selected checkpoint epoch need not be the final completed epoch. Complete-test timing includes loader iteration, device transfers, model evaluation, and output concatenation. Table 26 reports these complete-pipeline measurements alongside the training statistics.

Table 26: Training and complete-test inference costs on the same GPU. Training includes validation and restoration of the best checkpoint; seconds/epoch includes validation. All profiles allow 50 epochs with patience 10. Test time includes loader iteration, transfers, forwards, and output concatenation. Batch size 32, seed 42.

Dataset	Model	Params (M)↓	Epochs	Train (s)↓	s/epoch↓	Peak MiB↓	Test (s)↓
ETTh1	ACT	6.403	25	119.9	4.80	140.2	0.16
ETTh1	DLinear	0.019	15	10.6	0.70	18.2	0.06
ETTh1	iTransformer	1.630	13	37.2	2.87	79.1	0.17
ETTm1	ACT	0.507	14	412.8	29.49	75.6	0.86
ETTm1	DLinear	0.019	27	116.1	4.30	18.2	0.24
ETTm1	iTransformer	1.630	13	181.7	13.98	79.1	0.49

Continued on next page

Table 26 (continued)

Dataset	Model	Params (M)↓	Epochs	Train (s)↓	s/epoch↓	Peak MiB↓	Test (s)↓
Weather	ACT	0.212	29	891.2	30.73	145.0	0.80
Weather	DLinear	0.019	17	83.1	4.89	21.6	0.35
Weather	iTransformer	1.630	18	250.0	13.89	100.6	0.50
ECL	ACT	7.644	15	487.8	32.52	1950.4	3.00
ECL	DLinear	0.019	27	118.4	4.39	96.7	0.53
ECL	iTransformer	1.630	37	544.4	14.71	996.6	1.30
Traffic	ACT	15.331	14	916.2	65.45	5163.6	6.22
Traffic	DLinear	0.020	12	49.7	4.14	228.8	0.71
Traffic	iTransformer	1.631	50	1552.9	31.06	5300.7	2.76

Training cost of low-rank LAB. We profile the epoch-level training cost of otherwise matched dense- and factorized-LAB models, holding $L = H = 96$, $d = 512$, $k = 2$, learning rate 10^{-3}, batch size 32, seed 42, and the data pipeline fixed. The only model change is replacing dense LAB gates with rank-8 factors on ECL or rank-16 factors on Traffic; the encoder, forecast head, and state–action computation are retained. Each architecture receives a separate one-epoch warmup run. Each timed repeat constructs a fresh model and trainer at seed 42 and measures one training-plus-validation epoch, with one repeat per ECL configuration and three per Traffic configuration. Timing excludes model/data setup and test evaluation, includes validation and checkpoint callbacks, and ends after CUDA synchronization. With one repeat per ECL configuration, ECL timing does not estimate repeat-to-repeat variability; Traffic variability is SD across timing repeats, not training seeds. Peak memory is the largest allocation across repeats.

Table 27: Dense and low-rank LAB at $L = H = 96$, $d = 512$, $k = 2$, batch size 32, on an A100-SXM4-40GB. After a separate warmup run, each of n_t repeats starts a fresh seed-42 model and times one training-plus-validation epoch. Peak memory is the maximum allocated across repeats. MSE is mean ± sample SD from complete training at the same settings, seeds 42–44.

Dataset	LAB	Params (M)↓	n_t	s/epoch↓	Peak MiB↓	Speedup↑	MSE↓ (3 seeds)
ECL	Dense	5.129	1	22.31	1364.5	1.00×	0.14801 ± 0.00047
ECL	Rank 8	0.282	1	12.96	1318.0	1.72×	0.14554 ± 0.00066
Traffic	Dense	10.254	3	45.60 ± 0.25	3605.6	1.00×	0.45570 ± 0.00538
Traffic	Rank 16	0.599	3	19.85 ± 0.10	3528.8	2.30×	0.44677 ± 0.00730

Low-rank LAB reduces mean training-plus-validation time per epoch from 22.31 to 12.96 seconds on ECL and from 45.60 to 19.85 seconds on Traffic, corresponding to 1.72× and 2.30× speedups. Parameter counts fall by 18.16× and 17.13×, respectively. In this fixed-$k = 2$ comparison, factorizing the LAB gates reduces training cost while preserving the encoder, forecast head, and state–action interface. Table 26 profiles each selected forecasting configuration separately to early stopping. Hardware performance depends on matrix and batch size; the static inference table reports these operating points separately, including batch size one.

The complete configuration-grid runs show lower three-seed mean test MSE for low-rank LAB than for dense LAB on ECL (rank 8) and Traffic (rank 16) (Table 27). Checkpoints are selected by validation performance from runs with seeds 42–44, at most 50 epochs, and early stopping after ten epochs without validation improvement. These accuracy results come from complete training runs, distinct from the one-epoch timing runs and the selected main-table configurations.

Search accounting. Table 28 reports the compute cost of the 1,920-run configuration search, including the Solar no-RevIN selection grid. The search covers 20 configurations per dataset–horizon pair, three selection seeds, four horizons, and eight datasets; all 1,920 candidate–seed runs completed successfully. GPU-hours sum per-run durations rather than wall-clock time and exclude idle intervals. Averaged over the eight datasets, successful runs use 28.57 GPU-hours per dataset, or 7.14 minutes per successful run. The "With retries" totals include retry sessions; the per-run mean uses successful-run durations only.

Table 28: Configuration-search cost: 20 configurations per dataset/horizon, three selection seeds, four horizons. GPU-hours sum run durations and do not denote elapsed wall time under parallel execution. Solar configurations disable the per-window RevIN transform throughout this search.

Dataset	Runs	GPU-hours↓	With retries↓	Min/run↓
ETTh1	240	4.57	4.57	1.14
ETTh2	240	3.20	3.20	0.80
ETTm1	240	42.13	42.54	10.53
ETTm2	240	31.21	31.57	7.80
Weather	240	32.92	33.07	8.23
ECL	240	33.85	34.11	8.46
Traffic	240	66.52	68.20	16.63
Solar	240	14.12	14.12	3.53
Total	1920	228.52	231.38	–

H From observed context to a useful forecasting revision

For Solar and Weather, the evidence proceeds from context signal beyond target history and calendar, to linear decodability of context labels from the action, to predictive utility under matched context substitution at a fixed target history.

H.1 Data and physical quantities

Solar comprises 137 modeled photovoltaic-power series at 10-minute resolution, sourced from NREL and distributed with the forecasting benchmark. Each series is a site-level power profile; for a target site, the other sites' observed power trajectories provide the observation context for prediction. Weather consists of observations from the Jena weather station.[1]

Weather forecasts measured air temperature, distinct from the logger-temperature channel. Its contextual inputs comprise observed radiation (shortwave radiation, photosynthetically active radiation, and its maximum), moisture (relative humidity, dew point, and water-vapor-related variables), pressure, and wind (speed, maximum speed, and direction); the moisture group also includes derived measurements.

In the 137-channel Solar benchmark, the five fixed target channels are the zero-based indices $\{0, 34, 68, 102, 136\}$. For each target c, its context comprises the other 136 channels, including the other four fixed targets. Let $p_j(t)$ denote site j's power, s_j its training 99th-percentile power, and $m_j(\tau)$ its training median normalized power at within-day position $\tau \in \{0, \ldots, 143\}$. We define the site residual and leave-target-out fleet anomaly as

$$r_j(t) = p_j(t)/s_j - m_j(t \bmod 144), \qquad f_{-c}(t) = \text{median}_{j \neq c}\, r_j(t). \tag{22}$$

The normalized fleet power is the median of $p_j(t)/s_j$ over those same 136 sites, and $f_{-c}(t)$ is their median departure from their empirical daily profiles. The context probe uses the level and change of both quantities, where level is the past-hour mean and change is the difference between the past-hour and preceding-hour means.

For radiation, pressure, and moisture readouts, "level" is the past-hour mean and "change" is the difference from the preceding hour. "Relative level" subtracts the context variable's mean over the full input history and divides by its input-history SD; "relative change" divides its past-hour change by the same SD. Radiation readout labels use shortwave radiation; moisture labels use relative humidity. Wind labels are past-hour horizontal velocity components and their changes, derived from speed and meteorological direction. Absolute and relative labels are reported separately.

H.2 Sampling, support, and controls

We evaluate frozen $L = H = 96$ checkpoints from seeds 42–46: rank 4/depth 3 for Weather with RevIN, and rank 8/depth 3 for Solar without RevIN. History anchors are spaced one hour apart. The training bank contains at most 4,096 evenly selected anchors; validation and test retain hourly anchors. The 1-, 3-, and 6-hour endpoints use the same 96-step forecast. Because the history and forecast windows for adjacent hourly anchors overlap, the samples are temporally dependent.

[1] Benchmark data: https://github.com/laiguokun/multivariate-time-series-data. Weather variables: https://www.bgc-jena.mpg.de/wetter/Weatherstation.pdf.

Weather excludes histories containing nonfinite values, sentinel values below −900, or negative wind speeds. Small negative radiation is clipped for physical labels, while predictor inputs are retained.

Matching covariates comprise the target history's mean, standard deviation, last value, and slope; daily phase represented by its first two sine/cosine harmonics; and annual phase represented by sine and cosine. Weather phase uses timestamps, whereas Solar phase uses row position modulo the 144-step day. An anchor is active when the mean over its six 10-minute observations ending at that anchor is at least $10\,\mathrm{W\,m^{-2}}$ for Weather. For Solar, we divide each site's power by its training-period 99th-percentile power, take the leave-target-out median across sites at each observation, and require its six-observation mean to be at least 0.05. Both windows end at the anchor, and Solar's normalization is calibrated on training data; neither mask uses future values.

To vary external context while holding each recipient's history fixed, we match it to training donors with comparable target histories and within-day phases no more than one hour apart. Training-set statistics standardize own-history and control features. For the supported panel, we additionally require the nearest-donor distance to be no greater than the validation-set 90th percentile of nearest-donor distances. Replacing the recipient's external family or fleet histories with those of a matched donor preserves the recipient history and normalization, leaving its history state numerically unchanged. The resulting forecast difference quantifies the trained model's sensitivity to external-input replacement at the fixed recipient history.

Each model uses its validation-selected checkpoint under the prescribed early-stopping rule. Thresholds, sites, context families, and donor-matching criteria are fixed using training and validation information; the test split is used only to measure the resulting readouts and forecast gains. Five-seed SD measures model-seed variation.

H.3 Predictive information in the observed context

On Solar, fleet context carries predictive information beyond target history and calendar: the six-hour MSE gain is 8.12%, versus −0.89% for shuffled context (Table 29). We quantify this information with a ridge probe that predicts future mean target change from target-history PCA16 and common controls, then adds either a context family's four features or independently shuffled context of equal dimension. PCA and scaling use training data, and ridge strength is selected from $\{0.01, 1, 10, 100, 1000\}$ by mean standardized validation MSE across endpoints; gains are normalized by the target-history/calendar probe's MSE. These results show that Solar fleet context contains predictive signal beyond target history and calendar; the next subsection tests whether ACT's latent action makes that signal readable.

Table 29: Context information in the data, measured by diagnostic ridge forecasting. Test R^2 and MSE gain use the same panel. Own denotes target history plus calendar controls; percentage gains use its MSE as the denominator, so positive favors added context. Gains are computed per site before averaging. Solar includes five fixed sites.

Dataset/family	Hours	Own R^2 ↑	Context R^2 ↑	Shuf. R^2 ↑	Gain (%)↑	Shuf. (%)↑
Weather: radiation	1	0.718	0.720	0.717	0.71	−0.45
Weather: radiation	3	0.555	0.556	0.553	0.19	−0.38
Weather: radiation	6	0.454	0.453	0.451	−0.29	−0.58
Weather: pressure	1	0.718	0.717	0.716	−0.17	−0.59
Weather: pressure	3	0.555	0.549	0.553	−1.19	−0.35
Weather: pressure	6	0.454	0.447	0.453	−1.22	−0.26
Weather: moisture	1	0.718	0.711	0.717	−2.27	−0.46
Weather: moisture	3	0.555	0.545	0.555	−2.06	0.02
Weather: moisture	6	0.454	0.444	0.457	−1.84	0.45
Weather: wind	1	0.718	0.718	0.718	0.21	0.05
Weather: wind	3	0.555	0.559	0.555	0.93	0.16
Weather: wind	6	0.454	0.469	0.457	2.68	0.46
Solar: fleet	1	0.664	0.721	0.662	16.74	−0.60
Solar: fleet	3	0.731	0.788	0.729	20.63	−0.97
Solar: fleet	6	0.842	0.855	0.840	8.12	−0.89

H.4 READING CONTEXT FROM THE ACTION

On active histories, linear probes decode two selected context labels more accurately from $a_c^{[16]}$ than from $z_c^{[16]}$: Solar fleet anomaly (test R^2 0.706 versus 0.659; 0.626 with shuffled actions) and Weather relative radiation level (0.699 versus 0.658; 0.629 with shuffled actions). Tables 30 and 31 report complete readouts on active and all valid histories.

We measure the incremental linear readability of current context labels over shared controls by comparing predictions from controls alone with predictions from the same controls plus $x_c^{[16]}$, $z_c^{[16]}$, $a_c^{[16]}$, $z_c^{[8]} \oplus a_c^{[8]}$, or shuffled $a_c^{[16]}$. Each representation-based probe has a 16-dimensional PCA budget, and all probes use the same controls and train/validation split; the resulting R^2 scores quantify linearly accessible context information.

Table 30: Complete context readouts on active histories: five-seed mean test $R^2 \uparrow$ (higher is better). All methods include the same controls; Solar averages five sites within each seed. Negative R^2 means greater error than predicting the test-target mean. Absolute and relative labels are distinct targets. This includes the selected context labels shown in Table 6.

Dataset/family	Target	Control	$x_c^{[16]}$	$z_c^{[16]}$	$a_c^{[16]}$	$z_c^{[8]} \oplus a_c^{[8]}$	Shuf. a_c
Weather: radiation	level	−0.813	−0.447	−0.500	−0.555	−0.659	−0.816
Weather: radiation	change	0.473	0.524	0.372	0.411	0.388	0.470
Weather: radiation	relative level	0.627	0.624	0.658	0.699	0.672	0.629
Weather: radiation	relative change	0.474	0.519	0.564	0.570	0.568	0.475
Weather: pressure	level	−0.038	−0.051	−0.029	−0.036	−0.038	−0.038
Weather: pressure	change	0.139	0.181	0.125	0.173	0.145	0.138
Weather: pressure	relative level	0.113	0.128	0.094	0.179	0.145	0.112
Weather: pressure	relative change	0.147	0.213	0.181	0.184	0.193	0.148
Weather: moisture	level	0.210	0.199	0.138	0.206	0.196	0.189
Weather: moisture	change	0.382	0.652	0.383	0.391	0.382	0.390
Weather: moisture	relative level	0.361	0.419	0.436	0.385	0.474	0.364
Weather: moisture	relative change	0.149	0.427	0.452	0.211	0.439	0.157
Weather: wind	u level	−0.141	−0.067	−0.170	−0.116	−0.066	−0.166
Weather: wind	v level	0.121	0.056	0.182	0.090	0.153	0.120
Weather: wind	u change	0.006	0.010	−0.022	0.008	−0.019	−0.001
Weather: wind	v change	0.017	0.047	0.038	0.029	0.074	0.004
Solar: fleet	anomaly	0.626	0.637	0.659	0.706	0.691	0.626
Solar: fleet	anomaly change	0.188	0.254	0.260	0.302	0.308	0.185
Solar: fleet	fleet power	0.778	0.783	0.787	0.822	0.812	0.778
Solar: fleet	fleet power change	0.832	0.846	0.847	0.864	0.860	0.831

Table 31: Complete context readouts on all valid histories, including inactive periods. Entries are five-seed mean test $R^2 \uparrow$ (higher is better), with the same probes and aggregation as Table 30. Differences between these two tables reflect their evaluation populations.

Dataset/family	Target	Control	$x_c^{[16]}$	$z_c^{[16]}$	$a_c^{[16]}$	$z_c^{[8]} \oplus a_c^{[8]}$	Shuf. a_c
Weather: radiation	level	−0.602	−0.294	−0.327	−0.534	−0.563	−0.610
Weather: radiation	change	−0.047	−0.014	−0.084	−0.039	−0.095	−0.060
Weather: radiation	relative level	0.839	0.842	0.852	0.862	0.856	0.839
Weather: radiation	relative change	0.478	0.511	0.546	0.565	0.554	0.478
Weather: pressure	level	−0.067	−0.087	−0.063	−0.039	−0.051	−0.068
Weather: pressure	change	0.071	0.110	0.065	0.091	0.069	0.070
Weather: pressure	relative level	0.076	0.091	0.069	0.129	0.102	0.073
Weather: pressure	relative change	0.115	0.162	0.136	0.126	0.130	0.114
Weather: moisture	level	0.059	0.065	0.003	0.034	0.028	0.036
Weather: moisture	change	0.390	0.619	0.422	0.377	0.422	0.392
Weather: moisture	relative level	0.485	0.511	0.542	0.496	0.570	0.485
Weather: moisture	relative change	0.223	0.426	0.448	0.249	0.433	0.225

Continued on next page

Table 31 (continued)

Dataset/family	Target	Control	$x_c^{[16]}$	$z_c^{[16]}$	$a_c^{[16]}$	$z_c^{[8]} \oplus a_c^{[8]}$	Shuf. a_c
Weather: wind	u level	−0.407	−0.297	−0.419	−0.414	−0.391	−0.446
Weather: wind	v level	0.019	0.048	0.094	−0.003	0.055	0.013
Weather: wind	u change	−0.005	−0.010	−0.026	−0.014	−0.031	−0.013
Weather: wind	v change	0.005	0.021	0.007	−0.006	0.014	−0.002
Solar: fleet	anomaly	0.704	0.728	0.748	0.756	0.754	0.703
Solar: fleet	anomaly change	0.311	0.376	0.430	0.401	0.426	0.308
Solar: fleet	fleet power	0.903	0.910	0.911	0.919	0.918	0.902
Solar: fleet	fleet power change	0.722	0.762	0.765	0.779	0.789	0.721

H.5 Equal-budget joint-input readout controls

Table 32: Equal-budget context readouts on active histories. $X^{[16]}$ and $Z^{[16]}$ project all variables' histories and states; $a_c^{[16]}$ projects the target's action. $R^2 \uparrow$ averages relative-radiation level/change (Weather) and fleet-anomaly level/change (Solar, five sites). Levels are past-hour means; changes compare consecutive hours. $\Delta R^2 = R^2(a_c^{[16]}) - R^2(Z^{[16]})$ reports five-seed mean ± SD.

Dataset	$a_c^{[16]}$	$Z^{[16]}$	$X^{[16]}$	ΔR^2
Weather	0.63446	**0.90576**	0.54054	−0.27130 ± 0.02898
Solar	0.50385	**0.74485**	0.73030	−0.24100 ± 0.01873

To compare context-label readability at equal dimension, we project the flattened joint history X and the encoded state Z, which encodes the joint history, to 16 PCA coordinates and compare them with the target's $a_c^{[16]}$. Each probe combines one of these 16-dimensional representations with the same ten target-history and calendar controls. Standardization and PCA are fitted on training anchors, and validation selects ridge regularization. We evaluate active histories using the existing five $H = 96$ checkpoints per dataset. These readouts measure the linear accessibility of context labels; fixed-state replacement separately tests predictive utility.

Within each seed, the summary averages relative-radiation level and change for Weather and fleet-anomaly level and change for Solar; for Solar, it also averages across the five fixed sites before computing seed statistics. $Z^{[16]}$ yields higher context-label R^2 in all five seeds on both datasets (Table 32), indicating more accurate linear decoding of these labels.

H.6 Action and context interventions

We assess predictive utility with the target history fixed: native actions are compared with training-mean actions and $A = 0$, and matched external context is substituted with the action recomputed. Tables 33 and 34 report all families, endpoints, and comparators. Relative gain is $100(M_{\text{comparator}} - M_{\text{native}})/M_{\text{comparator}}$. We compute it per site and seed, average sites within each seed, and report the five-seed mean and sample SD.

At six hours, the supported-active panel contains 2,527 Solar site–anchors and 556 Weather anchors. For Solar, native-context forecasts have lower MSE than forecasts under matched-context substitution in all 25 site–seed comparisons, with a macro-averaged gain of 28.00 ± 3.52%. Relative to the mean action, native actions yield gains of 46.66 ± 3.16% on all 3,080 active Solar site–anchors and 29.93 ± 3.45% on the 581 active Weather anchors. Because the Weather target and mean-action control are shared across context families, the Weather result is one cross-family comparison.

Table 33: Native-action MSE gains over each intervention on active histories (%, ↑; five-seed mean ± sample SD). Positive values favor native actions. Anchors are counted once across seeds. Solar averages site-wise gains over five sites; Weather shares action controls across families.

Dataset/family	Hours	Anchors	Mean A	$A = 0$	Context swap
Weather: radiation	1	581	28.28 ± 10.41	38.52 ± 5.15	0.92 ± 1.97
Weather: radiation	3	581	33.03 ± 4.85	39.76 ± 5.79	2.96 ± 2.95
Weather: radiation	6	581	29.93 ± 3.45	34.85 ± 4.75	4.06 ± 3.72

Continued on next page

Table 33 (continued)

Dataset/family	Hours	Anchors	Mean A	$A = 0$	Context swap
Weather: pressure	1	581	28.28 ± 10.41	38.52 ± 5.15	0.40 ± 0.45
Weather: pressure	3	581	33.03 ± 4.85	39.76 ± 5.79	0.60 ± 0.86
Weather: pressure	6	581	29.93 ± 3.45	34.85 ± 4.75	0.96 ± 1.29
Weather: moisture	1	581	28.28 ± 10.41	38.52 ± 5.15	2.06 ± 1.67
Weather: moisture	3	581	33.03 ± 4.85	39.76 ± 5.79	4.26 ± 1.50
Weather: moisture	6	581	29.93 ± 3.45	34.85 ± 4.75	4.95 ± 0.62
Weather: wind	1	581	28.28 ± 10.41	38.52 ± 5.15	1.04 ± 1.68
Weather: wind	3	581	33.03 ± 4.85	39.76 ± 5.79	−0.45 ± 1.88
Weather: wind	6	581	29.93 ± 3.45	34.85 ± 4.75	0.17 ± 1.99
Solar: fleet	1	3080	19.89 ± 4.96	20.03 ± 6.16	15.48 ± 3.32
Solar: fleet	3	3080	38.57 ± 3.40	38.40 ± 5.01	25.47 ± 3.20
Solar: fleet	6	3080	46.66 ± 3.16	45.93 ± 4.85	28.75 ± 3.49

Table 34: Native-action MSE gains on supported active histories (%, ↑; five-seed mean ± sample SD). The validation-calibrated donor-distance threshold restricts the active panel and determines its anchor counts. Columns and aggregation follow Table 33.

Dataset/family	Hours	Anchors	Mean A	$A = 0$	Context swap
Weather: radiation	1	556	29.05 ± 11.44	40.01 ± 5.88	1.71 ± 2.23
Weather: radiation	3	556	32.25 ± 6.60	40.84 ± 7.19	3.61 ± 3.53
Weather: radiation	6	556	28.77 ± 3.87	34.40 ± 5.83	4.59 ± 4.22
Weather: pressure	1	556	29.05 ± 11.44	40.01 ± 5.88	0.44 ± 0.67
Weather: pressure	3	556	32.25 ± 6.60	40.84 ± 7.19	0.46 ± 0.94
Weather: pressure	6	556	28.77 ± 3.87	34.40 ± 5.83	0.75 ± 1.24
Weather: moisture	1	556	29.05 ± 11.44	40.01 ± 5.88	2.45 ± 1.49
Weather: moisture	3	556	32.25 ± 6.60	40.84 ± 7.19	4.51 ± 1.85
Weather: moisture	6	556	28.77 ± 3.87	34.40 ± 5.83	4.89 ± 0.71
Weather: wind	1	556	29.05 ± 11.44	40.01 ± 5.88	0.78 ± 1.35
Weather: wind	3	556	32.25 ± 6.60	40.84 ± 7.19	−0.71 ± 1.89
Weather: wind	6	556	28.77 ± 3.87	34.40 ± 5.83	−0.53 ± 1.77
Solar: fleet	1	2527	19.66 ± 5.62	20.00 ± 6.78	15.59 ± 3.50
Solar: fleet	3	2527	38.28 ± 4.19	38.59 ± 5.78	25.21 ± 2.96
Solar: fleet	6	2527	46.43 ± 4.14	46.54 ± 5.97	28.00 ± 3.52

With the target history fixed, Figure 4 summarizes Solar's native-context gains over matched context substitution across forecast endpoints and all 25 six-hour site–seed comparisons; Figure 6 breaks down the six-hour effects by Solar site and Weather context family.

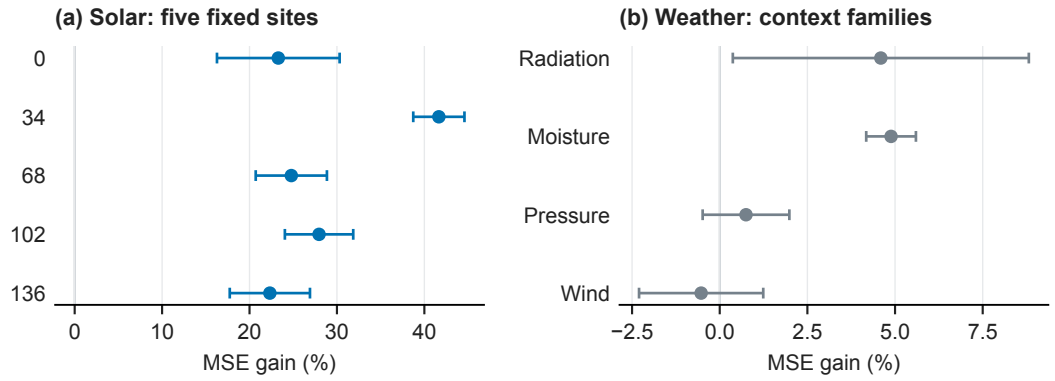


Figure 6: With the target history fixed, the context-dependent latent action updates the forecasting state used to form the conditional future forecast. Points show the six-hour MSE improvement from retaining native context rather than substituting matched training context; error bars are SD over five seeds. Solar includes all five fixed sites, and Weather all four context families. Donor support is calibrated on validation; activity masks use observed histories.

Large opposing trends. An additional active Solar subset isolates site–anchors where the target site's training-profile residual and the leave-target-out fleet anomaly change with opposite signs. Each change compares the most recent one-hour mean with the preceding one-hour mean; both absolute changes exceed their respective 75th-percentile thresholds, computed from active training anchors. Across these 47 site–anchors, native actions reduce six-hour MSE by 44.58 ± 11.31% relative to the mean action and by 30.77 ± 21.44% relative to matched context substitution. This subset illustrates useful context-dependent forecast revision at a fixed target history; the broader supported active panel provides the primary comparison.

Together, these results progress from context signal to decodability and fixed-history predictive utility. We next visualize the forecast correction associated with each coefficient at a common state.

H.7 Action geometry at a common forecasting state

We visualize native action coefficients and color each point by its forecast correction at one shared target-history state, using the same forecast head.

Action population and coordinates. We analyze the 516 supported active test anchors from Solar site 0, the first of the five fixed sites, using the seed-42 validation-selected checkpoint. The model has $L = H = 96$, latent dimension 512, rank 8, depth 3, and no RevIN; sampling and support follow the hourly-anchor protocol above. Each point represents the site's native action $a_i \in \mathbb{R}^{512}$ at one anchor. We center these coefficients and apply 50-component PCA, retaining 98.22% of their variance, followed by three-dimensional t-SNE with perplexity 30, PCA initialization, automatic learning rate, and random seed 0. Coordinates are fitted from coefficients alone; forecast effects enter only as colors. Ten-neighbor trustworthiness is 0.992 at perplexity 30, compared with 0.990 and 0.993 at perplexities 15 and 50. The t-SNE axes show a projection of local coefficient geometry, while color encodes the six-hour mean forecast correction relative to the zero-action forecast, evaluated at the same target-history state with the forecast head fixed and expressed in standardized units.

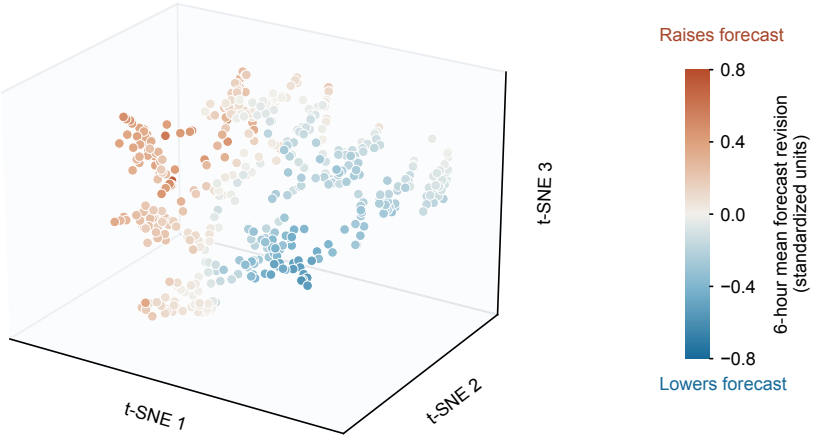


Figure 7: Action-coefficient geometry and conditional forecast revisions on Solar site 0. Each point is one of 516 native actions from the seed-42 validation-selected model; its position is the three-dimensional t-SNE projection fitted using action coefficients alone. For the color, each action updates the forecasting state at the same observed reference state, and the updated state is mapped to a forecast by the same trained head. Color shows the mean forecast correction over the first six hours, defined as the action-conditioned forecast minus the zero-action forecast. Blue indicates a lower mean forecast and terracotta a higher one; the symmetric color scale is in standardized target units.

A shared operational reference. Let z_i denote the native target state at anchor i. We select the observed state medoid $z_* = z_{j_*}$ by minimizing its mean Euclidean distance to the population, $j_* = \arg\min_j \frac{1}{516}\sum_i \|z_j - z_i\|_2$. The selected medoid is test window 8868, and selection uses states alone. To isolate each coefficient's contextual prediction revision, we hold the reference state z_* and trained head fixed, changing only the state supplied as the head's input. The head maps this updated state to a forecast; we define $\Delta\hat{y}_i$ as the difference from its zero-action forecast:

$$\Delta\hat{y}_i = h(z_* \odot (\mathbf{1} + a_i)) - h(z_*) = W_{\text{out}}(z_* \odot a_i), \qquad q_i = \frac{1}{36}\sum_{t=1}^{36} \Delta\hat{y}_{i,t}. \tag{23}$$

The zero-action forecast $h(z_*)$ is the common baseline, so q_i supplies the point color. The first 36 ten-minute steps span six hours. Solar has no RevIN inverse transformation, so these revisions are expressed directly in standardized target units. Positive and negative values of q_i indicate upward and downward mean forecast revisions, respectively. A value near zero indicates little net mean revision over this interval; temporal changes can still cancel in the average.

From coefficient geometry to forecast revision. The t-SNE projection shows coefficient neighborhoods with similar colors, linking local action geometry to forecast corrections at the shared state. Native-recipient comparisons above establish predictive utility on the histories that generated the actions.