SPAR: SPARSE PHASE-AWARE ADAPTATION OF REPRESENTATIONS FOR TIME SERIES FORECASTING WITH EXOGENOUS VARIABLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Exogenous variables provide complementary information for time series forecasting, but redundant or unstable cross-variable dependencies can impair predictions. Existing methods primarily model or select these dependencies, with limited ability to directly regulate their predictive effects in learned representations. Using sparse autoencoders (SAEs), we find that representation shifts induced by incorporating exogenous variables concentrate in a small set of sparse features, and that intervening on these features can substantially alter forecasting behavior. Motivated by this observation, we propose Sparse Phase-aware Adaptation of Representations (SPAR), a plug-in that leverages sparse adaptation shifts as an intervention interface to selectively regulate the predictive effects of exogenous variables. SPAR learns phase-aware corrections from forecasting error, enabling beneficial cross-variable effects to be preserved while mitigating harmful ones. Experiments on four backbones demonstrate consistent improvements, with average relative MSE reductions of 5.78%, 8.02%, 6.84%, and 5.74%, respectively, across seven datasets, highlighting SPAR's strong generalizability across diverse backbones and datasets. Code is available at: https://anonymous.4open.science/r/spar-Wzcc-Xzcz.

1 INTRODUCTION

Time series forecasting supports industrial equipment maintenance (Zhou et al., 2025b; Yang et al., 2025; Huo et al., 2025), and networked control systems (Zhao et al., 2009; Wang et al., 2010), where target variables often depend on other variables. Incorporating exogenous variables is an important route to better predictions. However, more variables do not necessarily improve forecasts: neighboring road flows can inform traffic prediction, while fluctuations on unrelated roads may introduce interference. When dependencies vary over time, information from the same variable can shift between complementary and redundant. The challenge is therefore to assess and regulate the predictive effects of cross-variable dependencies.

Existing methods exploit exogenous variables by explicitly modeling cross-variable dependencies, as in iTransformer (Liu et al., 2024) and TimeXer (Wang et al., 2024), or adaptively selecting dependencies, as in DUET (Qiu et al., 2025b) and CCM (Chen et al., 2024). Yet learned dependencies can not directly indicate predictive utility: a variable's influence may propagate through multiple internal features, combining complementary information with components that induce negative transfer. Variable-level selection offers limited control over these mixed effects. Here, we ask: **can we locate cross-variable adaptation effects in learned representations and selectively intervene on them?**

We use Sparse AutoEncoder (SAE) to investigate this question, which can decompose dense representations into sparse features, and leverage SAE to compare an endogenous-only (one-to-one, O2O) model with an endogenous-exogenous (many-to-one, M2O) model. As shown in figure 1, we observe three interesting phenomena. **(a) Shifts concentrate in few features:** compared with raw representation coordinates, a small fraction of SAE features captures most of the representation shift induced by incorporating exogenous variables. **(b) Shift magnitude tracks prediction changes:** the magnitude of shifts in these sparse SAE features is strongly associated with the magnitude of forecasting-error changes. **(c) Few features enable effective intervention:** reverting only

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.

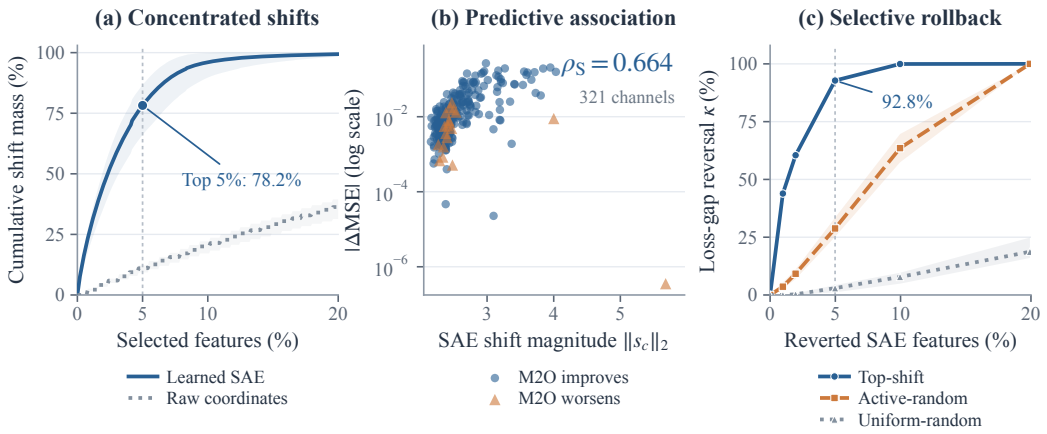


Figure 1: **Sparse features reveal cross-variable adaptation effects** (iTransformer on Electricity). (a) The top 5% of SAE features capture 78.2% of shift mass, versus 11.0% for raw coordinates. (b) The horizontal axis is the L2 norm of the SAE feature shift vector; the vertical axis is the absolute O2O–M2O MSE difference $|\Delta\text{MSE}| = |\text{MSE}_{\text{M2O}} - \text{MSE}_{\text{O2O}}|$, where Spearman correlation is 0.664. (c) After incorporating exogenous variables, reverting the top 5% of SAE features with the largest shifts to their O2O values reverses 92.8% of the O2O–M2O loss gap, versus 28.8% for active-random selection (from features with nonzero shifts) and 3.0% for uniform-random selection (from all SAE features), at the same feature budget.

the most shifted SAE features toward their O2O values recovers most of the O2O–M2O loss gap, substantially outperforming random feature reversion under the same budget. These findings support sparse features as coordinates for locating and intervening on adaptation effects. However, an association with absolute error changes does not identify beneficial or harmful effects; correction directions must still be learned from the forecasting objective.

We therefore propose **SPAR** (**S**parse **P**hase-aware **A**daptation of **R**epresentations). SPAR selects high-shift candidate features in the shared SAE space and learns phase-aware signed coefficients from forecasting error. These coefficients allow feature adaptation shifts to be retained, attenuated, reversed, or amplified, with different correction rules at different temporal positions. Corrections are added to representations as residual increments along SAE decoder directions, mapped to forecast corrections through the original M2O prediction head, and combined with auxiliary output calibration. O2O, M2O, and the SAE remain frozen during plug-in training; only additional parameters are fitted. SPAR thus connects sparse feature analysis of cross-variable effects to forecast optimization through a common representation intervention interface for different backbones.

Our main contributions are:

- We reveal that cross-variable adaptation effects exhibit a sparse structure in learned representations, and show through shift concentration, predictive association, and selective rollback that a small subset of SAE features captures most of these effects.
- We propose SPAR, a novel plug-in that leverages sparse adaptation shifts to selectively regulate the predictive effects of exogenous variables through forecast-supervised, phase-aware representation corrections.
- We conduct extensive experiments across different backbones and datasets. Specifically, with SPAR, four backbones achieve average relative MSE reductions of 5.78%, 8.02%, 6.84%, and 5.74%, across seven datasets, demonstrating consistent improvements and strong generalizability.

2 Related Work

Temporal modeling and channel strategies. Forecasting architectures span hybrid ARIMA–neural models (Zhang, 2003), NeuralProphet (Triebe et al., 2021), long- and short-term neural

modeling (Lai et al., 2018), efficient attention (Zhou et al., 2021), and convolution (Luo & Wang, 2024). FMamba and THCVformer explore gated and heterogeneous temporal modeling (Ma et al., 2026a;b). Channel strategies offer another perspective (Qiu et al., 2025a): DLinear (Zeng et al., 2023) and PatchTST (Nie et al., 2023) use channel independence, while iTransformer (Liu et al., 2024), TimesNet (Wu et al., 2023), and Crossformer (Zhang & Yan, 2023) model multiple channels jointly. C3RL (Ma et al., 2026c) combines independence and mixing through representation learning. DUET (Qiu et al., 2025b), CCM (Chen et al., 2024), channel masks (Lee et al., 2024), and TimeFilter (Hu et al., 2025) offer selective dependence modeling. Leading indicators (Zhao & Shen, 2024) inform cross-channel selection.

Exogenous information and adaptive correction. TimeXer (Wang et al., 2024), CrossLinear (Zhou et al., 2025a), XLinear (Chen et al., 2026), GCGNet (Li et al., 2026b), and DAG (Qiu et al., 2026) incorporate exogenous inputs into target forecasting. xCPD (Li et al., 2026a) routes channel-patch dependencies through graph spectral decomposition. More broadly, dynamic expert routing (Huang et al., 2024) adapts computation to tasks, while temporal-slice normalization (Liu et al., 2023) addresses non-stationarity. SPAR complements these approaches with forecast-supervised edits to cross-variable adaptation effects in frozen representations.

Sparse representations for forecasting. OLinear (Yue et al., 2026) forecasts in an orthogonally transformed domain. Sparse autoencoders (Ng et al., 2011; Makhzani & Frey, 2013) provide a learned sparse dictionary. TimeSAE (Oublal et al., 2026) and SAE analysis of Chronos (Mishra, 2026) investigate explanations and feature structure in forecasting models. SPAR uses shared SAE coordinates to locate adaptation shifts, then learns their correction strengths and directions from forecasting error.

3 PRELIMINARIES

3.1 TIME SERIES FORECASTING WITH EXOGENOUS VARIABLES

For a C-variable series $\{\mathbf{x}_t\}_{t=1}^{T}$, with $\mathbf{x}_t \in \mathbb{R}^C$, forecasting uses L historical steps to predict the next H steps. For target channel c and forecast origin t, the endogenous history is $\mathbf{x}^{(c)}_{t-L:t-1} \in \mathbb{R}^L$, the remaining channels form the exogenous history $\mathbf{X}^{(-c)}_{t-L:t-1} \in \mathbb{R}^{L\times(C-1)}$, and the target is $\mathbf{y}_c = \mathbf{x}^{(c)}_{t:t+H-1} \in \mathbb{R}^H$. O2O uses endogenous history alone, $\widehat{\mathbf{y}}^O_c = f_{\vartheta_O}(\mathbf{x}^{(c)}_{t-L:t-1})$, whereas M2O also uses exogenous history, $\widehat{\mathbf{y}}^M_c = f_{\vartheta_M}(\mathbf{x}^{(c)}_{t-L:t-1}, \mathbf{X}^{(-c)}_{t-L:t-1})$. These settings specify each target's input access. $\mathbb{E}_{\mathcal{D}}[\cdot]$ denotes the empirical average over windows in a set $\mathcal{D}$. Over training windows $\mathcal{D}_b$, models minimize $\mathcal{L}_{\text{pred}}(\vartheta) = (CH)^{-1}\mathbb{E}_{\mathcal{D}_b}[\sum_{c=1}^{C} \|\mathbf{y}_c - \widehat{\mathbf{y}}_c(\vartheta)\|_2^2]$.

3.2 SPARSE AUTOENCODERS

A TopK SAE represents a hidden state $\mathbf{h} \in \mathbb{R}^d$ using at most k active features in a dictionary of width $m > d$. With encoder preactivations $\mathbf{u}(\mathbf{h}) = \mathbf{W}_E(\mathbf{h} - \mathbf{b}_D) + \mathbf{b}_E$, its sparse code and reconstruction are

$$\mathbf{a}(\mathbf{h}) = \mathrm{ReLU}(\mathrm{TopK}_k(\mathbf{u}(\mathbf{h}))), \qquad \widehat{\mathbf{h}} = \mathbf{D}\mathbf{a}(\mathbf{h}) + \mathbf{b}_D, \tag{1}$$

where $\mathbf{W}_E \in \mathbb{R}^{m\times d}$ and $\mathbf{b}_E \in \mathbb{R}^m$ are encoder parameters; $\mathbf{D} \in \mathbb{R}^{d\times m}$ and $\mathbf{b}_D \in \mathbb{R}^d$ define the decoder. TopK_k retains the k largest entries by value and zeros the rest; ReLU ensures $\|\mathbf{a}(\mathbf{h})\|_0 \le k$. Column $\mathbf{D}_j$ is feature j's decoder direction.

The SAE is fitted to O2O training states $\mathcal{H}_O$ by minimizing

$$\min_{\mathbf{W}_E,\mathbf{b}_E,\mathbf{D},\mathbf{b}_D} \frac{1}{|\mathcal{H}_O|\cdot d} \sum_{\mathbf{h}\in\mathcal{H}_O} \|\mathbf{h} - \widehat{\mathbf{h}}\|_2^2 \quad \text{subject to} \quad \|\mathbf{D}_j\|_2 = 1 \quad \forall j. \tag{2}$$

TopK imposes sparsity, while unit-norm decoder directions give activations a consistent scale. With residual $\boldsymbol{\epsilon}(\mathbf{h}) = \mathbf{h} - \widehat{\mathbf{h}}$, the decomposition $\mathbf{h} = \mathbf{D}\mathbf{a}(\mathbf{h}) + \mathbf{b}_D + \boldsymbol{\epsilon}(\mathbf{h})$ supports coefficient edits that preserve the unreconstructed component.

4 METHOD

SPAR turns sparse representation shifts induced by cross-variable adaptation into learnable forecast corrections. As shown in Figure 2, we locate candidate features in a shared SAE space, apply phase-aware signed interventions along their decoder directions, and connect representation corrections directly to forecasting error through the frozen original prediction head.

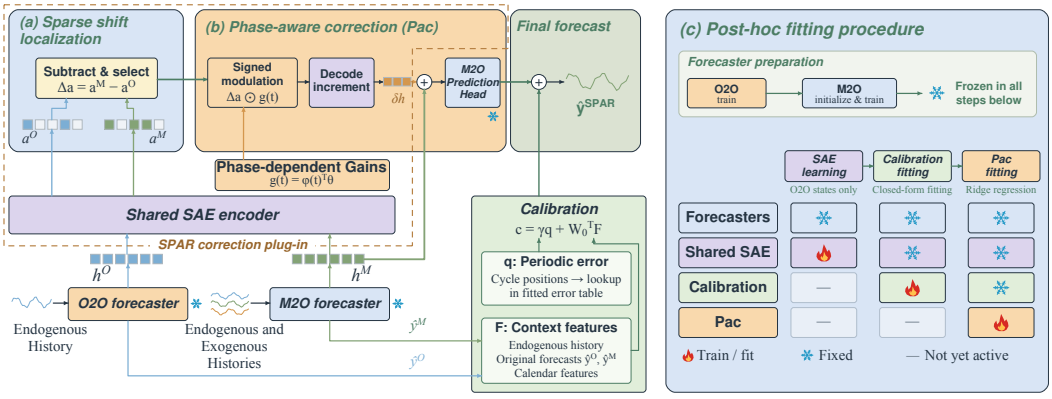


Figure 2: **Overview and post-hoc fitting of SPAR.** (a) Identify candidate features in a shared SAE space based on O2O–M2O representation shifts. (b) Phase-dependent gains produce an additive representation-level correction $\delta\mathbf{h}$, which is added before the frozen M2O prediction head; Calibration adjusts the output. (c) After forecaster preparation, the SAE, Calibration, and Phase-aware correction are fitted sequentially. Flames denote fitting and snowflakes denote fixed parameters.

4.1 SHARED SPARSE REPRESENTATIONS AND CANDIDATE FEATURES

We first train an O2O model, then initialize M2O from its parameters and continue training with exogenous information through the backbone's native cross-variable mechanism. Let $\mathbf{h}_c^O, \mathbf{h}_c^M \in \mathbb{R}^d$ denote the representations before the original prediction heads for target channel c. The SAE is fitted only to O2O training representations, then frozen together with both forecasting models. Appendix A.1 describes each backbone's native cross-variable mechanism and O2O implementation.

We encode both representations with the same SAE and select candidate features by their mean absolute activation shifts over training windows $\mathcal{D}_{\text{tr}}$:

$$S_c = \text{TopKIndices}_j(\mathbb{E}_{\mathcal{D}_{\text{tr}}}[|\Delta a_{c,j}|]), \tag{3}$$

where S_c is the set of K candidate feature indices for channel c, $\Delta a_{c,j} = a_j(\mathbf{h}_c^M) - a_j(\mathbf{h}_c^O)$ denotes the activation difference of SAE feature j between the M2O and O2O representations of channel c, and TopKIndices returns the indices of the K largest scores. The set is determined solely from training-input representations and remains fixed at inference. Shift magnitude identifies candidates without determining their predictive utility; correction directions and strengths are learned from forecasting supervision.

4.2 PHASE-AWARE CORRECTION

Cross-variable dependencies may have different predictive effects at different temporal positions. We therefore condition gains on the known forecast origin t. Let $P_1, \ldots, P_R$ be R specified cycle lengths in sampling steps (e.g., 24 and 168 for daily and weekly cycles in hourly data). Sine and cosine terms encode cycle position to define

$$\begin{aligned}\delta\mathbf{h}_c &= \alpha \sum_{j \in S_c} \mathbf{D}_j \Delta a_{c,j} g_{c,j}(t), \\ g_{c,j}(t) &= \boldsymbol{\phi}(t)^\top \boldsymbol{\theta}_{c,j}, \\ \boldsymbol{\phi}(t) &= \left[1, \sin\frac{2\pi t}{P_1}, \cos\frac{2\pi t}{P_1}, \ldots, \sin\frac{2\pi t}{P_R}, \cos\frac{2\pi t}{P_R}\right]^\top,\end{aligned} \tag{4}$$

where $\delta\mathbf{h}_c \in \mathbb{R}^d$ is the representation increment, $\mathbf{D}_j$ is feature j's decoder direction, and $g_{c,j}(t)$ is its signed gain. The phase vector $\boldsymbol{\phi}(t) \in \mathbb{R}^{1+2R}$ includes a constant for phase-independent correction, and $\boldsymbol{\theta}_{c,j} \in \mathbb{R}^{1+2R}$ contains learnable coefficients. Validation selects the overall shrinkage $\alpha \in [0, 1]$. The corrected representation is $\widetilde{\mathbf{h}}_c^M = \mathbf{h}_c^M + \delta\mathbf{h}_c$.

An effective coefficient $\alpha g = 0$ retains the adaptation shift; values in $(-1, 0)$ attenuate it; -1 restores the selected decoder coefficient to its O2O value; and positive values amplify the shift. Coefficients are not restricted to $[-1, 0]$, allowing reverse corrections beyond the O2O reference. By correcting the original representation along SAE decoder directions, SPAR preserves information not captured by SAE reconstruction.

4.3 Learning Interventions from Forecasting Error

Auxiliary calibration and supervision. Forecast error for an endogenous variable can reflect both its own temporal dynamics and the influence of exogenous variables. Before correcting cross-variable effects, we first account for predictable errors associated with the target's temporal context. SPAR fits auxiliary output calibration using target history, calendar features, and the original forecasts, then learns sparse representation corrections from the remaining error. We use $\mathbf{c}_c = \gamma\mathbf{q}_c + \mathbf{W}_0^\top \mathbf{F}_c \in \mathbb{R}^H$, where γ controls periodic calibration and $\mathbf{W}_0$ contains linear weights shared across channels. The periodic component is

$$\mathbf{q}_c = \left[T_{c,\, t \bmod P_{\text{cal}}}, \ldots, T_{c,\, (t+H-1) \bmod P_{\text{cal}}}\right]^\top, \tag{5}$$

where P_{cal} is the calibration cycle length in sampling steps and $T_{c,p}$ is the shrinkage mean of M2O training errors for channel c at cycle position p. The vector $\mathbf{q}_c \in \mathbb{R}^H$ retrieves these fitted estimates at known future cycle positions. The context vector is

$$\mathbf{F}_c = \left[\sigma_c \mathcal{N}\Big(\left[\overline{\mathbf{x}}_{t-L:t-1}^{(c)}; \overline{\widehat{\mathbf{y}}}_c^{O}; \overline{\widehat{\mathbf{y}}}_c^{M}; \boldsymbol{\psi}(t); \mu_c; \log\sigma_c\right]\Big); \sigma_c; 1\right], \tag{6}$$

where semicolons concatenate column vectors, μ_c and σ_c are the target history's mean and stabilized standard deviation, and bars denote normalization by these statistics. The vector $\boldsymbol{\psi}(t)$ contains fixed calendar Fourier features, and $\mathcal{N}$ standardizes context entries using calibration-training statistics and clips them to $[-10, 10]$. For $\mathbf{F}_c \in \mathbb{R}^{d_F}$, $\mathbf{W}_0 \in \mathbb{R}^{d_F \times H}$. Calendar features use their known timestamps. Appendix A.3 gives construction and fitting details. Freezing the calibrator yields $\widehat{\mathbf{y}}_c^{\text{cal}} = \widehat{\mathbf{y}}_c^M + \mathbf{c}_c$ and the supervised residual $\mathbf{r}_c = \mathbf{y}_c - \widehat{\mathbf{y}}_c^{\text{cal}}$.

Forecasting through the original head. The frozen M2O prediction head $\mathcal{P}_c$ maps the corrected representation to a forecast, to which we add auxiliary calibration:

$$\widehat{\mathbf{y}}_c^{\text{SPAR}} = \mathcal{P}_c(\mathbf{h}_c^M + \delta\mathbf{h}_c) + \mathbf{c}_c. \tag{7}$$

Learning coefficients. With fixed input statistics, an affine prediction mapping makes forecast increments linear in the phase coefficients (Appendix A.2). Setting $\alpha = 1$, we collect increments from candidate feature–phase pairs in $\mathbf{Z}_c \in \mathbb{R}^{H \times K(1+2R)}$. Dividing each column by its training root mean square gives $\widetilde{\mathbf{Z}}_c$, putting regularization on comparable forecast scales. For each channel, we solve

$$\boldsymbol{\beta}_c^* = \arg\min_{\boldsymbol{\beta}_c} \frac{1}{H} \mathbb{E}_{\mathcal{D}_{\text{tr}}}\left[\left\|\mathbf{r}_c - \widetilde{\mathbf{Z}}_c\boldsymbol{\beta}_c\right\|_2^2\right] + \lambda\|\boldsymbol{\beta}_c\|_2^2, \tag{8}$$

where $\boldsymbol{\beta}_c \in \mathbb{R}^{K(1+2R)}$ contains the rescaled design coefficients and $\lambda \geq 0$ controls ridge regularization. Undoing column scaling recovers $\boldsymbol{\theta}$. Validation selects λ and α, including $\alpha = 0$ for no intervention. Only phase coefficients are learned here; inference uses fixed parameters. Appendix A.4 details the design matrix, joint solve, and parameter recovery.

Table 1: Forecasting performance. MSE in units of 10^{-3} (↓).

System	Traffic	Electricity	ETTh2	ETTm2	ETTh1	ETTm1	Weather	Avg.Gain
iTransformer								
E2E M2O	161.20±0.95	365.80±1.40	199.68±0.95	128.65±0.73	75.48±0.60	53.10±0.53	1.88±0.15	
Adapted M2O	156.40±0.80	365.23±1.35	199.68±0.98	126.93±0.70	75.28±0.60	53.33±0.53	1.88±0.15	
+SPAR	**142.18**±0.85	**325.10**±1.15	**188.35**±0.95	**119.00**±0.38	**70.90**±0.58	**51.20**±0.50	1.90±0.13	**5.78%**
TimeXer								
E2E M2O	147.13±0.95	355.63±0.78	188.95±0.58	121.40±0.78	73.28±0.60	52.19±0.43	1.95±0.13	
Adapted M2O	152.20±0.98	355.68±0.88	189.25±0.60	120.30±0.73	73.30±0.58	52.28±0.43	1.95±0.13	
+SPAR	**134.43**±0.83	**290.43**±0.95	**178.08**±0.75	**113.30**±0.80	**69.90**±0.50	**49.85**±0.38	**1.85**±0.13	**8.02%**
CrossLinear								
E2E M2O	148.75±1.10	322.58±1.15	185.63±0.55	118.30±0.53	72.78±0.45	52.35±0.40	1.93±0.10	
Adapted M2O	154.78±0.98	323.08±1.18	186.18±0.58	118.38±0.78	72.43±0.45	52.33±0.40	1.93±0.10	
+SPAR	**130.88**±0.83	**289.20**±1.48	**173.43**±0.75	**111.35**±0.53	**69.38**±0.50	**50.43**±0.33	**1.90**±0.10	**6.84%**
DUET								
E2E M2O	144.80±0.75	318.08±1.05	186.70±0.85	118.30±0.65	73.48±0.53	51.93±0.45	1.70±0.13	
Adapted M2O	142.98±1.00	318.40±1.00	184.28±0.85	118.10±0.58	73.43±0.53	52.00±0.45	1.70±0.13	
+SPAR	**134.23**±0.78	**287.33**±1.03	**173.25**±0.83	**111.40**±0.45	**70.23**±0.58	**50.00**±0.40	**1.63**±0.13	**5.74%**

5 EXPERIMENTS

Benchmarking datasets. We evaluate SPAR on seven public datasets: ETTh1, ETTh2, ETTm1, ETTm2, Electricity, Traffic, and Weather, covering transformer temperature, electricity load, traffic, and weather forecasting. For each target variable, historical observations of the remaining variables serve as exogenous inputs. Forecast horizons are set to $H \in \{96, 192, 336, 720\}$, and forecasting errors are measured by MSE and MAE. Dataset details are provided in Appendix A.5.

Backbones and baselines. We evaluate iTransformer (Liu et al., 2024), TimeXer (Wang et al., 2024), CrossLinear (Zhou et al., 2025a), and DUET (Qiu et al., 2025b), spanning linear and Transformer-based architectures. In Table 1, E2E M2O is trained end to end, Adapted M2O is initialized from O2O and adapted with exogenous inputs, and SPAR adds auxiliary calibration and sparse correction to the frozen models.

Implementation details. A shared SAE is fitted to O2O training representations and frozen with both forecasting models before calibration and sparse correction are fitted sequentially. Training and calibration data determine candidate features and regression coefficients; validation selects regularization and correction strengths. Parameters are fixed at inference. Appendices A and C provide implementation and aggregation details, MAE, and per-horizon results.

5.1 OVERALL FORECASTING PERFORMANCE

Main Results. Table 1 compares E2E M2O, Adapted M2O, and full SPAR across four backbones and seven datasets. Relative to Adapted M2O, SPAR lowers horizon-averaged MSE in 27 of 28 combinations. The mean relative reductions are 5.78%, 8.02%, 6.84%, and 5.74% for iTransformer, TimeXer, CrossLinear, and DUET, respectively. Electricity and Traffic have the largest mean MSE gains, 12.39% and 10.58%. On Weather, TimeXer, CrossLinear, and DUET improve by 5.13%, 1.30%, and 4.41%, while iTransformer worsens by 1.33%, showing that gains depend on both dataset and backbone. MAE results appear in Table C.1.

Comparison with competitive plug-ins. Table 2 compares SPAR with two competitive plug-ins, CCM (Chen et al., 2024) and xCPD (Li et al., 2026a), which model cross-variable effects at the variable level, across iTransformer, TimeXer, CrossLinear, and DUET. Each metric is first averaged over $H \in \{96, 192, 336, 720\}$. Relative gains are computed against the corresponding backbone without a plug-in within each dataset, then averaged equally over seven datasets. W/T/L denotes

the numbers of wins/ties/losses against the corresponding backbone over 14 dataset–metric pairs. For each backbone and dataset, all plug-ins are applied to the same frozen M2O checkpoint and use identical training, validation, and test splits. Only the plug-in modules differ. This separate comparison uses the supplied dataset-level means in Appendix C.2.

Table 2: Mean relative forecasting gains of plug-ins over each backbone without a plug-in (↑).

Backbone	Method	Mean Rel. MSE Gain (%)	Mean Rel. MAE Gain (%)	W/T/L vs Base
iTransformer	CCM	0.09	0.03	8/5/1
	xCPD	2.58	1.23	10/3/1
	SPAR	**6.54**	**4.06**	**13/1/0**
TimeXer	CCM	0.87	0.75	11/3/0
	xCPD	2.30	1.46	12/2/0
	SPAR	**6.16**	**4.18**	**13/1/0**
CrossLinear	CCM	0.88	0.75	11/3/0
	xCPD	2.33	1.48	12/2/0
	SPAR	**6.09**	**3.96**	**13/1/0**
DUET	CCM	0.28	−0.91	8/2/4
	xCPD	3.12	−0.01	11/0/3
	SPAR	**6.04**	**3.94**	**14/0/0**

SPAR achieves the largest mean relative MSE and MAE gains on all four backbones. Its MSE reductions are 6.54%, 6.16%, 6.09%, and 6.04% for iTransformer, TimeXer, CrossLinear, and DUET, respectively; the corresponding MAE reductions are 4.06%, 4.18%, 3.96%, and 3.94%. Across all four backbones, SPAR improves 53 of the 56 metric–dataset pairs relative to the base models, with three ties and no losses. The ties are Weather MSE on iTransformer, TimeXer, and CrossLinear at the supplied precision. Detailed dataset-level comparisons appear in Tables C.8 and C.9.

5.2 Ablation of Core Mechanisms

To separate representation correction from differences in system training, we ablate the same adapted model in Table 3.

Table 3 uses the same adapted TimeXer on Electricity, Traffic, ETTh2, and ETTm2, averaging MSE over $H \in \{96, 192, 336, 720\}$. Let C denote output calibration and S sparse representation correction. Removal variants omit C, S, or both; replacement variants retain C and change one mechanism of S. Reductions below use horizon means, computed against the specified comparator within each dataset and then averaged equally. Full SPAR achieves the lowest horizon-averaged MSE on all four datasets. Full per-horizon results are in Appendix C.1.

Representation correction provides the main incremental gains. Relative to Adapted M2O, S only reduces MSE by 7.70%, versus 0.59% for C only. Full SPAR further reduces MSE by 9.91% over C only and by 3.03% over S only. Sparse correction thus improves forecasts without auxiliary calibration and complements output calibration.

Channel-specific adaptation differences guide candidate selection. *Active-random support* draws K candidates from active SAE features instead of ranking adaptation shifts. *Global support* shares candidates across channels but keeps channel-specific coefficients. *M2O activations* uses a^M instead of $\Delta a = a^M - a^O$ for correction. *Output-difference gate* applies a phase-aware scalar per channel to $\hat{\mathbf{y}}_c^M - \hat{\mathbf{y}}_c^O$, replacing feature-level edits. Full SPAR reduces MSE by 6.63%, 4.33%, and 5.57% relative to active-random support, global support, and M2O activations a^M, respectively. These comparisons support selecting high-shift features, retaining channel specificity, and using O2O as the reference for adaptation differences. Its 2.44% MSE reduction over output-difference gating further supports feature-level intervention beyond the compared output-level correction.

Phase and signed coefficients enable flexible corrections. *Static coefficients* removes the sinusoidal phase terms, leaving one time-independent gain per selected feature. *Mismatched phase*

Table 3: Ablation study of SPAR on TimeXer. Horizon-averaged MSE in units of 10^{-3} (↓).

Variants		Electricity	Traffic	ETTh2	ETTm2	Avg.
Full SPAR $(C+S)$		**290.4**±1.0	**134.4**±0.8	**178.1**±0.8	**113.3**±0.8	**179.1**±0.8
Remove	w/o C (S only)	307.7±1.0	139.0±0.9	181.0±0.7	115.2±0.8	185.7±0.8
	w/o S (C only)	353.8±0.9	151.3±1.0	188.1±0.6	119.6±0.7	203.2±0.8
	w/o C, S (adapted M2O)	355.7±0.9	152.2±1.0	189.3±0.6	120.3±0.7	204.4±0.8
Replace	Output-difference gate	304.0±1.0	138.2±0.9	180.4±0.7	114.8±0.8	184.3±0.8
	Active-random support	329.8±0.9	145.1±0.9	184.8±0.7	117.5±0.8	194.3±0.8
	Global support	315.4±1.0	141.1±0.9	182.4±0.7	116.0±0.8	188.7±0.8
	M2O activations (a^M)	322.9±0.9	143.3±0.9	183.7±0.7	116.8±0.8	191.7±0.8
	Static coefficients	309.7±0.9	139.8±0.9	181.4±0.7	115.4±0.7	186.6±0.8
	Mismatched phase	342.7±0.9	148.7±1.0	187.0±0.6	118.9±0.7	199.3±0.8
	Attenuation only	319.4±0.9	142.2±0.9	183.0±0.7	116.4±0.8	190.3±0.8

breaks the alignment between phase features and forecast origins while retaining the same coefficient count. *Attenuation only* restricts the effective edit coefficient $\alpha g_{c,j}(t)$ to $[-1, 0]$, permitting only retention or attenuation of each selected adaptation shift toward its O2O value. Full SPAR reduces MSE by 3.42% over static coefficients and 8.57% over mismatched phase. The latter performs close to C only, supporting the importance of correct phase alignment. Full SPAR also reduces MSE by 4.98% over attenuation-only coefficients in $[-1, 0]$, supporting corrections beyond attenuation.

5.3 Model Analysis

Analysis of the role of SAE. Figure 3(a,b) compares corrections in SAE, raw representation, and PCA coordinates on Electricity and Weather. Here K is the number of candidate correction directions selected for each channel: SAE decoder directions, raw representation coordinates, or PCA components, respectively. We vary this correction budget over $K \in \{8, 16, 32, 64, 128\}$. Test MSE reductions are measured against the same calibrated reference C and averaged over four forecast horizons.

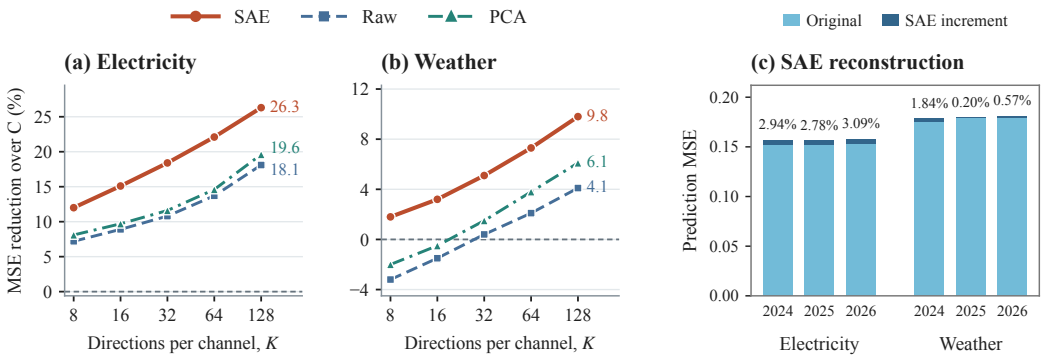


Figure 3: **SAE correction performance and forecasting fidelity.** (a,b) Varying the budget for different corrections. MSE reductions over C are averaged across $H \in \{96, 192, 336, 720\}$. The experiments are conducted on iTransformer. (c) iTransformer M2O forecasting MSE after O2O-trained SAE reconstruction. Light blue shows original MSE and dark blue shows its increase. The total height gives reconstructed MSE. Labels show percentage increases over original MSE.

SAE yields larger MSE reductions than Raw and PCA at all five budgets on both datasets. At $K = 128$, SAE reduces MSE by 26.3% on Electricity, exceeding Raw and PCA by 8.2 and 6.7 percentage points, respectively. On Weather, the corresponding reduction is 9.8%, exceeding Raw and PCA by 5.7 and 3.7 percentage points. This advantage persists as the direction budget increases. Raw and PCA also need larger budgets to match SAE's gain at $K = 8$: approximately $K = 64$ for

both on Electricity, and $K = 64$ for Raw and $K = 32$ for PCA on Weather. Together, these results show that SAE coordinates support stronger corrections at a fixed budget and use fewer directions to reach comparable gains, supporting their role in representation correction.

SAE reconstruction under adaptation. The SAE is trained only on O2O representations but applied to adapted M2O representations, raising a distribution-shift concern. To test whether the shift undermines forecasting, we replace each adapted representation with its SAE reconstruction and evaluate it using the frozen M2O head. At $H = 96$, the mean relative increase in test MSE across three seeds is only $2.94\% \pm 0.15\%$ on Electricity and $0.87\% \pm 0.86\%$ on Weather (Figure 3(c)). Thus, even full-representation replacement preserves nearly all aggregate M2O forecast accuracy. Together with the selective rollback in Figure 1(c), this supports the O2O-trained SAE as a forecast-relevant space for sparse interventions after adaptation.

Inference efficiency. As a plug-in, SPAR primarily incurs additional computation at inference time. Figure 4 compares the inference efficiency of different plug-ins on Electricity with iTransformer. As shown in Figure 4 (a), SPAR achieves an MSE of 0.1336 with 2.938ms latency and 20.85MiB peak inference memory. In contrast, CCM and xCPD provide little to no MSE improvement over the M2O backbone in this setting. The standalone M2O backbone remains faster and more memory-efficient, reflecting the additional inference cost required for SPAR's accuracy gains.

Figure 4 (b) decomposes SPAR's inference latency into the O2O branch (0.817ms), M2O branch (0.564ms), and SAE+C+S modules (1.557ms). The two forecasting branches account for 47.0% of the total profiled latency, while calibration and representation correction account for 53.0%. Although the SAE+C+S computation is not the sole source of overhead, SPAR additionally retains the O2O branch to obtain the O2O–M2O representation difference required by its correction mechanism. This explains why SPAR remains more expensive than a single-backbone plug-in such as xCPD, while still being substantially lighter than CCM in overall inference cost.

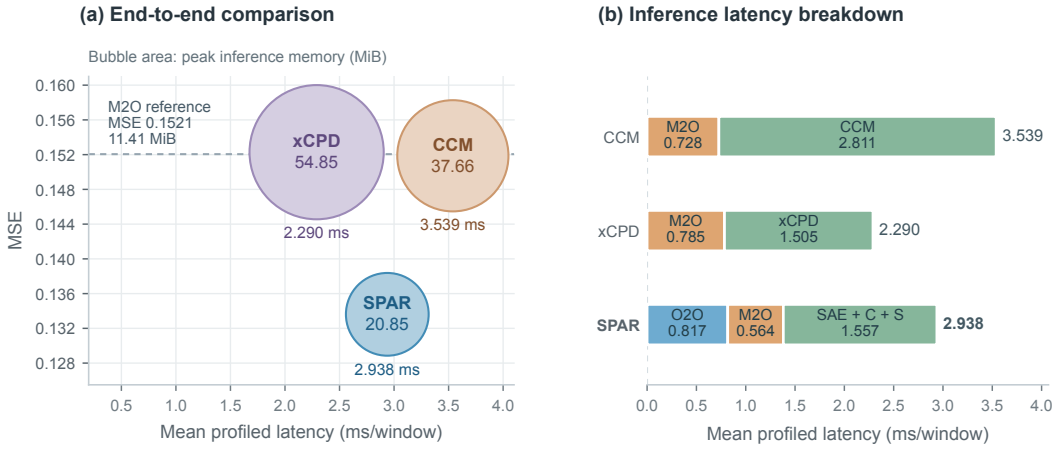


Figure 4: **End-to-end efficiency and inference cost decomposition.** iTransformer on Electricity. Results are averaged across three seeds. (a) MSE versus median inference latency; bubble area denotes peak GPU allocated memory. (b) C and S denote Calibration and SPAR. For CCM and xCPD, M2O denotes the encoder, with its replacement head included in CCM.

6 Conclusion

SPAR regulates exogenous information through forecast-supervised, phase-aware signed edits in a shared SAE space. Experiments across seven datasets and four backbones support its effectiveness; ablations and budget comparisons support gains from sparse correction beyond output calibration. This connects the analysis of cross-variable adaptation to forecast optimization, offering a complementary route to dependency modeling in time series forecasting.

AI USE STATEMENT

We used generative AI tools to brainstorm research ideas and assist with experimental design and execution; the authors reviewed all AI-assisted work and take responsibility for the final paper.

REFERENCES

Jialin Chen, Jan Eric Lenssen, Aosong Feng, Weihua Hu, Matthias Fey, Leandros Tassiulas, Jure Leskovec, and Rex Ying. From similarity to superiority: Channel clustering for time series forecasting. In *Advances in Neural Information Processing Systems*, volume 37, pp. 130635–130663, 2024.

Xinyang Chen, Huidong Jin, Yu Huang, and Zaiwen Feng. Xlinear: a lightweight and accurate mlp-based model for long-term time series forecasting with exogenous inputs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20325–20335, 2026.

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. *CoRR*, abs/2501.13041, January 2025.

Quzhe Huang, Zhenwei An, Nan Zhuang, Mingxu Tao, Chen Zhang, Yang Jin, Kun Xu, Liwei Chen, Songfang Huang, and Yansong Feng. Harder task needs more experts: Dynamic routing in moe models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12883–12895, 2024.

Zixuan Huo, Hang Wu, Ruifang Xu, Mingxing Yuan, and Xuebo Zhang. Toward efast autonomous robotic ultrasound imaging: System integrations and experimental studies. *Sci. China Inf. Sci.*, 68(3), 2025.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 95–104, 2018.

Seunghan Lee, Taeyoung Park, and Kibok Lee. Partial channel dependence with channel masks for time series foundation models. 2024.

Dongyuan Li, Shun Zheng, Chang Xu, Jiang Bian, and Renhe Jiang. Routing channel-patch dependencies in time series forecasting with graph spectral decomposition. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Zhengyu Li, Xiangfei Qiu, Yuhan Zhu, Xingjian Wu, Jilin Hu, Guo Guo, and Bin Yang. Gcgnet: Graph-consistent generative network for time series forecasting with exogenous variables. In *International Conference on Learning Representations*, volume 2026, pp. 27904–27926, 2026b.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 11116–11140, 2024.

Zhiding Liu, Mingyue Cheng, Zhi Li, Zhenya Huang, Qi Liu, Yanhu Xie, and Enhong Chen. Adaptive normalization for non-stationary time series forecasting: A temporal slice perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Donghao Luo and Xue Wang. Moderntcn: A modern pure convolution structure for general time series analysis. In *International conference on learning representations*, volume 2024, pp. 31728–31770, 2024.

Shusen Ma, Yun-Bo Zhao, and Yu Kang. Fmamba: Bi-mamba based on fast gated-attention for multivariate time-series forecasting. *IEEE Transactions on Artificial Intelligence*, 2026a.

Shusen Ma, Yun-Bo Zhao, and Yu Kang. Thcvformer: Modeling temporal-dependencies, heterogeneity, and correlations for enhanced time-series forecasting. *Knowledge-Based Systems*, pp. 115964, 2026b.

Shusen Ma, Yunbo Zhao, and Yu Kang. C3RL: Rethinking the Combination of Channel-independence and Channel-mixing from Representation Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24281–24289, 2026c.

Alireza Makhzani and Brendan Frey. K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.

Anurag Mishra. Dissecting chronos: Sparse autoencoders reveal causal feature hierarchies in time series foundation models. *arXiv preprint arXiv:2603.10071*, 2026.

Andrew Ng et al. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19, 2011.

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.

Khalid Oublal, Quentin Bouniot, Qi Gan, Stephan Clémençon, and Zeynep Akata. TimeSAE: Causal sparse decoding for faithful explanations of black-box time series models. In *Forty-third International Conference on Machine Learning*, 2026.

Xiangfei Qiu, Hanyin Cheng, Xingjian Wu, Jilin Hu, and Chenjuan Guo. A comprehensive survey of deep learning for multivariate time series forecasting: A channel strategy perspective. *CoRR*, abs/2502.10721, February 2025a.

Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. Duet: Dual clustering enhanced multivariate time series forecasting. In *Proceedings of the 31st ACM SIGKDD conference on knowledge discovery and data mining v. 1*, pp. 1185–1196, 2025b.

Xiangfei Qiu, Yuhan Zhu, Zhengyu Li, Xingjian Wu, Bin Yang, and Jilin Hu. Dag: A dual correlation network for time series forecasting with exogenous variables. In *ICML*, 2026.

Oskar Triebe, Hansika Hewamalage, Polina Pilyugina, Nikolay Laptev, Christoph Bergmeir, and Ram Rajagopal. Neuralprophet: Explainable forecasting at scale. *arXiv preprint arXiv:2111.15397*, 2021.

Rui Wang, Guo-Ping Liu, Wei Wang, David Rees, and Yun-Bo Zhao. h_∞ control for networked predictive control systems based on the switched lyapunov function method. *IEEE Transactions on Industrial Electronics*, 57(10):3565–3571, 2010.

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. In *Advances in neural information processing systems*, volume 37, pp. 469–498, 2024.

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

Le Yang, Zelin Yang, Fan Li, and CL Philip Chen. Single-layer feedforward neural networks with dynamic width for domain adaptation. *Science China Information Sciences*, 68(8):182205, 2025.

Wenzhen Yue, Yong Liu, Hao Wang, Haoxuan Li, Xianghua Ying, Ruohao Guo, Bowei Xing, and Ji Shi. Olinear: A linear model for time series forecasting in orthogonally transformed domain. In *Advances in Neural Information Processing Systems*, volume 38, pp. 60573–60632, 2026.

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.

G Peter Zhang. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175, 2003.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The eleventh international conference on learning representations*, 2023.

Lifan Zhao and Yanyan Shen. Rethinking channel dependence for multivariate time series forecasting: Learning from leading indicators. In *International Conference on Learning Representations*, volume 2024, pp. 41433–41451, 2024.

Yun-Bo Zhao, Guo-Ping Liu, and David Rees. Design of a packet-based control framework for networked control systems. *IEEE Transactions on Control Systems Technology*, 17(4):859–865, 2009.

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

Pengfei Zhou, Yunlong Liu, Junli Liang, Qi Song, and Xiangyang Li. Crosslinear: Plug-and-play cross-correlation embedding for time series forecasting with exogenous variables. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 4120–4131, 2025a.

Qianli Zhou, Li Zhu, and Yong Deng. Measure-based uncertainty with dempster-shafer structure. *Science China Information Sciences*, 68(10):202204, 2025b.

A Implementation Details and Key Properties of SPAR

A.1 Native Model Mechanisms and Representation Extraction

O2O restricts each target to its own history by masking cross-channel access. Adapted M2O starts from the selected O2O checkpoint and restores access to the other channels' histories, retaining the same parameter shapes. A shared SAE trained on O2O states encodes both models at the pre-head interfaces below. E2E M2O in Table 1 is trained independently.

iTransformer (Liu et al., 2024). Each variable's history is a token. O2O applies diagonal masks to every variable-attention layer; M2O restores cross-variable attention. We extract the final target token immediately before the linear prediction head.

TimeXer (Wang et al., 2024). The target is represented by patch tokens and a global token, which queries variable tokens through cross-attention. O2O allows each target's global token to attend only to its own variable token; within-target temporal attention is unchanged. M2O restores the other variable tokens. We retain the complete final token array and its original ordering into the flattening/linear head.

Diagonal attention restriction. For iTransformer, TimeXer, and DUET, the O2O cross-variable attention is

$$\mathbf{A}^O = \mathrm{softmax}_{\mathrm{row}}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} + \mathbf{B}^O\right), \qquad B^O_{c,j} = \begin{cases} 0, & j = c, \\ -\infty, & j \neq c. \end{cases} \tag{A1}$$

Here $\mathbf{Q}, \mathbf{K}$ are queries and keys, and d_k is the key dimension. The $-\infty$ mask is applied before softmax, making forbidden weights exactly zero. A zero logit or a finite negative penalty does not guarantee isolation.

DUET (Qiu et al., 2025b). DUET first routes each channel's history through temporal experts, then fuses the resulting features by channel attention. In M2O, a learned Mahalanobis metric on Fourier amplitudes and a Gumbel–Softmax sampler generate an input-dependent channel mask. O2O bypasses this graph generator, supplies a fixed identity mask to every channel-attention layer, and enforces equation A1. Temporal experts, routing, and normalization operate independently on each channel in both modes. Thus no exogenous history enters through either the temporal branch or the mask itself. M2O restores the native learned-mask attention. We extract the fused target state before the final linear head and retain target-wise RevIN de-normalization. DUET's internal channel clustering module is distinct from the CCM plug-in in Table 2.

CrossLinear (Zhou et al., 2025a). CrossLinear combines each normalized target history with a convolutional cross-channel embedding. For convolution weights $W_{c,j,r}$, where c is the output channel, j the input channel, and r the temporal offset, O2O uses

$$W^O_{c,j,r} = W_{c,j,r}\,\mathbf{1}[j = c], \qquad z^O_{c,t} = b_c + \sum_r W_{c,c,r}\,\overline{x}^{(c)}_{t+r}. \tag{A2}$$

Bars denote normalization within each channel's input window. The diagonal mask is broadcast over all temporal offsets: it removes cross-channel mixing while preserving each channel's temporal convolution, bias, and input shape. M2O uses the full convolution weights. Both modes normalize patch and hidden features separately within each channel, retaining channel-specific affine parameters; normalization across channels would otherwise bypass the mask. This is a normalization-modified CrossLinear variant. We extract the state after the final LayerNorm/ReLU and before the last prediction Linear layer. The remaining head and target de-normalization are affine for fixed input statistics, as required by Appendix A.4.

Interface checks. With parameters and random draws fixed, changing other channels' histories must not change an O2O target forecast. Reinserting an unedited pre-head state must reproduce the native forecast, including token reshaping and target de-normalization. These checks concern input access and the prediction interface; adaptation also changes model parameters.

A.2 Properties Connecting Feature Interventions to Prediction

A.2.1 Sparse coefficient edits preserve the original prediction mapping

Property. Fix an input window and its normalization statistics, and suppose that the full original prediction mapping $\mathcal{P}$ is affine. Let $\boldsymbol{\delta}$ be a coefficient increment supported only on the selected SAE features. The edit $\widetilde{\mathbf{h}} = \mathbf{h} + \mathbf{D}\boldsymbol{\delta}$ preserves the SAE reconstruction residual, and its forecast increment follows from applying the original mapping to these decoder directions.

Proof. Writing $\mathbf{h} = \mathbf{D}\mathbf{a} + \mathbf{b}_D + \boldsymbol{\epsilon}$ gives

$$\widetilde{\mathbf{h}} = \mathbf{D}(\mathbf{a} + \boldsymbol{\delta}) + \mathbf{b}_D + \boldsymbol{\epsilon}. \tag{A3}$$

Thus, the increment leaves the original reconstruction residual in place. Define $\mathcal{L}(\mathbf{u}) = \mathcal{P}(\mathbf{h} + \mathbf{u}) - \mathcal{P}(\mathbf{h})$. Since $\mathcal{P}$ is affine, $\mathcal{L}$ is linear, and hence

$$\mathcal{P}(\widetilde{\mathbf{h}}) - \mathcal{P}(\mathbf{h}) = \sum_j \delta_j \mathcal{L}(\mathbf{D}_j). \tag{A4}$$

Setting $\delta_j = -\Delta a_j$ restores the corresponding O2O coefficient in the decoder expansion, recovering the restoration intervention in Section 1 as a special case. Restoring decoder coefficients does not require all re-encoded activations to remain unchanged. □

For models using RevIN, the original mapping includes the full DE-REVIN operation with the same input statistics as the original forecast. Taking prediction differences cancels its translation term. Applying the full DE-REVIN operation directly to an increment would incorrectly add this translation again.

A.2.2 Shift magnitude does not determine whether an edit helps

Property. Let $\mathbf{r} = \mathbf{y} - \widehat{\mathbf{y}}^{\mathrm{cal}}$. Adding a forecast increment $\mathbf{v}$ changes squared error by

$$\|\mathbf{r} - \mathbf{v}\|_2^2 - \|\mathbf{r}\|_2^2 = -2\langle \mathbf{r}, \mathbf{v} \rangle + \|\mathbf{v}\|_2^2. \tag{A5}$$

Proof. Expanding the squared norm and canceling $\|\mathbf{r}\|_2^2$ gives the identity. The error decreases exactly when $2\langle \mathbf{r}, \mathbf{v} \rangle > \|\mathbf{v}\|_2^2$; the magnitude of the increment or feature shift alone cannot determine whether this holds. □

This explains why concentrated shifts and restoration effects identify candidate interventions, whereas the sign and strength of each edit must be learned against the forecasting objective. It guarantees neither improvements on unseen data nor a benefit from phase conditioning.

A.3 Construction of Base Calibration

Base calibration consists of a periodic error table and shared linear calibration, which are fitted, selected using validation, and then frozen. To define the table in equation 5, let $e_{c,s}(t)$ be the original M2O error $y_{c,s}(t) - \widehat{y}^{M}_{c,s}(t)$ at forecast offset $s \in \{0, \ldots, H-1\}$ for the training window with origin t. Here we explicitly show the window argument only to specify cycle grouping. For a cycle of P_{cal} sampling steps, define

$$\begin{aligned} \mathcal{I}_p &= \{(t, s) : t \in \mathcal{D}_{\mathrm{cal}},\ 0 \le s < H,\ (t+s) \bmod P_{\mathrm{cal}} = p\}, \\ T_{c,p} &= \frac{\sum_{(t,s)\in\mathcal{I}_p} e_{c,s}(t)}{|\mathcal{I}_p| + \tau}, \qquad p = 0, \ldots, P_{\mathrm{cal}} - 1. \end{aligned} \tag{A6}$$

Here $\mathcal{D}_{\mathrm{cal}}$ contains calibration-training forecast origins, $\mathcal{I}_p$ groups forecast steps by cycle position, and $\tau > 0$ shrinks sparsely observed entries toward zero. For a nonempty group, this shrinkage mean equals the ordinary mean error multiplied by $|\mathcal{I}_p|/(|\mathcal{I}_p| + \tau)$. Thus positions with fewer training errors receive stronger shrinkage, while an empty group yields $T_{c,p} = 0$. Each window–offset pair contributes one error, including when forecast windows overlap. At prediction time, entries at known future cycle positions form $\mathbf{q}_c$; the table is fixed and no future observations are required.

To construct the ordinary context, we normalize target history and both original forecasts using the target window's mean and a stabilized standard deviation. Specifically, $\mu_c = L^{-1} \sum_{\ell=t-L}^{t-1} x_\ell^{(c)}$ and

$\sigma_c = (L^{-1}\sum_{\ell=t-L}^{t-1}(x_\ell^{(c)} - \mu_c)^2 + \epsilon_x)^{1/2}$, with $\epsilon_x > 0$ a numerical stabilizer. The bars in equation 6 denote $(\mathbf{v} - \mu_c \mathbf{1})/\sigma_c$ for each history or forecast vector $\mathbf{v}$. The calendar vector $\boldsymbol{\psi}(t)$ concatenates daily and weekly sine/cosine features, with the first four harmonics of each cycle; the periods are expressed in sampling steps. Unlike the phase gains in equation 4, these features are inputs to the shared output calibrator. They are concatenated with normalized history, both normalized forecasts, the window mean, and log scale. The operator $\mathcal{N}$ standardizes each context coordinate using its mean and stabilized standard deviation over calibration-training windows and channels, then clips to $[-10, 10]$. Multiplication by σ_c restores the current window scale. Appending σ_c and a constant gives $\mathbf{F}_c \in \mathbb{R}^{d_F}$, where $d_F = L + 2H + \dim(\boldsymbol{\psi}) + 4$. This construction uses the three observed inputs in Figure 2 and their known timestamps, without SAE features.

For each candidate periodic coefficient γ, the shared weights $\mathbf{W}_0$ fit $\mathbf{y} - \widehat{\mathbf{y}}^M - \gamma\mathbf{q}$ using multi-output ridge regression:

$$\min_{\mathbf{W}_0} \quad \frac{1}{C}\mathbb{E}_{\mathcal{D}_{\text{cal}}}\left[\sum_{c=1}^{C} \left\|\mathbf{y}_c - \widehat{\mathbf{y}}_c^M - \gamma\mathbf{q}_c - \mathbf{W}_0^\top \mathbf{F}_c\right\|_2^2\right] + \lambda_0 \|\mathbf{W}_0\|_F^2. \tag{A7}$$

The objective averages over samples and channels, while each squared vector norm sums over forecast steps; this specifies the regularization scale used by the solver. Validation selects γ and λ_0, including a zero-calibration candidate. Base calibration and sparse correction may use training windows with different coverage; the temporal split boundaries are part of the experimental protocol.

A.4 Solving for Sparse Corrections

Let $B = 1 + 2R$ be the number of phase terms, so that each channel has $J = KB$ parameters. Use the combined index $\ell = (j, f)$, where $j \in S_c$ and $f \in \{1, \ldots, B\}$ indexes a phase term. The original prediction mapping from Appendix A.2.1 gives

$$\begin{aligned} \mathbf{e}_{c,j} &= \mathcal{P}_c(\mathbf{h}_c^M + \mathbf{D}_j \Delta a_{c,j}) - \mathcal{P}_c(\mathbf{h}_c^M), \\ \mathbf{Z}_{c,:,\ell} &= \phi_f(t)\mathbf{e}_{c,j}, \\ \omega_{c,\ell} &= \max\left\{\sqrt{\frac{1}{H}\mathbb{E}_{\mathcal{D}_{\text{tr}}}[\|\mathbf{Z}_{c,:,\ell}\|_2^2]}, \varepsilon\right\}, \\ \widetilde{\mathbf{Z}}_{c,:,\ell} &= \mathbf{Z}_{c,:,\ell}/\omega_{c,\ell}. \end{aligned} \tag{A8}$$

Here, $\mathbf{e}_{c,j} \in \mathbb{R}^H$ is the forecast increment along candidate j, $\phi_f(t)$ is the fth phase-vector entry, and $\mathbf{Z}_{c,:,\ell}$ is column ℓ of the design matrix. The scale $\omega_{c,\ell}$ is the column's root mean square forecast increment; $\varepsilon > 0$ is its numerical floor. The training objective is equation 8. Its linear system and the recovery of the actual modulation coefficients are

$$\begin{aligned} \left(\frac{1}{H}\mathbb{E}_{\mathcal{D}_{\text{tr}}}[\widetilde{\mathbf{Z}}_c^\top \widetilde{\mathbf{Z}}_c] + \lambda\mathbf{I}\right)\boldsymbol{\beta}_c &= \frac{1}{H}\mathbb{E}_{\mathcal{D}_{\text{tr}}}[\widetilde{\mathbf{Z}}_c^\top \mathbf{r}_c], \\ \theta_{c,j,f} &= \beta_{c,(j,f)}/\omega_{c,(j,f)}. \end{aligned} \tag{A9}$$

The solve retains all cross-column terms and jointly handles correlated decoder directions. The implementation accumulates sufficient statistics using the factorized structure, without rerunning the model for every feature or storing the complete design tensor.

For each λ, validation computes the unshrunk increment $\mathbf{v} = \widetilde{\mathbf{Z}}\boldsymbol{\beta}$ and selects

$$\alpha = \text{clip}_{[0,1]} \frac{\mathbb{E}_{\mathcal{D}_{\text{val}}}[\sum_{c=1}^{C}\langle \mathbf{r}_c, \mathbf{v}_c\rangle]}{\mathbb{E}_{\mathcal{D}_{\text{val}}}[\sum_{c=1}^{C}\|\mathbf{v}_c\|_2^2]}. \tag{A10}$$

Here, $\mathcal{D}_{\text{val}}$ indexes validation windows. If the denominator is zero, we set $\alpha = 0$. The validation error after shrinkage selects a shared λ, including the option of no intervention.

A.5 Datasets and Data Use

We use the seven public long-term forecasting datasets described by Wang et al. (2024). Table A.1 summarizes their variables and sampling frequencies. The ETT collection comprises two hourly subsets (ETTh1/ETTh2) and two 15-minute subsets (ETTm1/ETTm2), each containing oil temperature and six power-load measurements. Electricity records hourly consumption from 321 clients. Traffic contains hourly road-occupancy rates from 862 freeway sensors in the San Francisco Bay Area. Weather records 21 meteorological variables at the Max Planck Institute for Biogeochemistry weather station in 2020, sampled every ten minutes.

Table A.1: Benchmark datasets. C counts all observed variables; each target has $C - 1$ historical exogenous variables.

Dataset	C	Sampling interval	Measurements
ETTh1, ETTh2	7	1 hour	Oil temperature and power load
ETTm1, ETTm2	7	15 minutes	Oil temperature and power load
Electricity	321	1 hour	Client electricity consumption
Traffic	862	1 hour	Road occupancy
Weather	21	10 minutes	Meteorological factors

For each target channel, O2O uses its own history and M2O additionally uses all other channels' histories. No future observations of either are inputs. The forecasting benchmark uses $H \in \{96, 192, 336, 720\}$ sampling steps and reports MSE and MAE. The same step count therefore spans different physical durations across datasets; calendar periods are expressed in sampling steps.

O2O training precedes M2O adaptation. The SAE is fitted on O2O training states; SPAR candidate selection and regression fitting use training data, with calibration fitted before sparse correction. Validation selects regularization and shrinkage, and all fitted components remain fixed at inference. Periodic lookups use known timestamps and training errors. Diagnostic split boundaries and feature-selection populations are specified in Appendix B.

B ADDITIONAL SAE ANALYSES

B.1 CROSS-MODEL EVIDENCE FOR THE THREE OBSERVATIONS

Figures 5–7 examine shift concentration, association with forecasting-error changes, and selective feature restoration on Electricity (321 channels, $L = H = 96$). All three backbones use the O2O masks in Appendix A.1 and a shared O2O-trained SAE with 1,024 features and TopK activation budget 51.

For each channel, $s_c(j)$ is the mean absolute O2O–M2O activation difference. Panel (a) reports the mass in its largest coordinates; panel (b) correlates $\|s_c\|_2$ with the absolute native O2O–M2O MSE difference. In panel (c), selected M2O coefficients are restored to their O2O values. Writing $d_c = L^{\text{dec}}_{O,c} - L^{\text{dec}}_{M,c}$, the reported statistic is

$$\kappa = \text{median}_{c\in\mathcal{E}} \frac{L^{\text{dec}}_{\text{edit},c} - L^{\text{dec}}_{M,c}}{d_c}, \qquad \mathcal{E} = \left\{c : d_c > \max\left(10^{-4},\, 0.1\,\text{median}_{j:d_j>0}\, d_j\right)\right\}.$$

All decoded losses use the same frozen M2O head and input statistics. The eligible channels are fixed across budgets; ratios are not clipped. This measures conditional decoded loss-gap reversal. Active-random samples features with $s_c(j) > 10^{-12}$; uniform-random samples the full dictionary at the same budget.

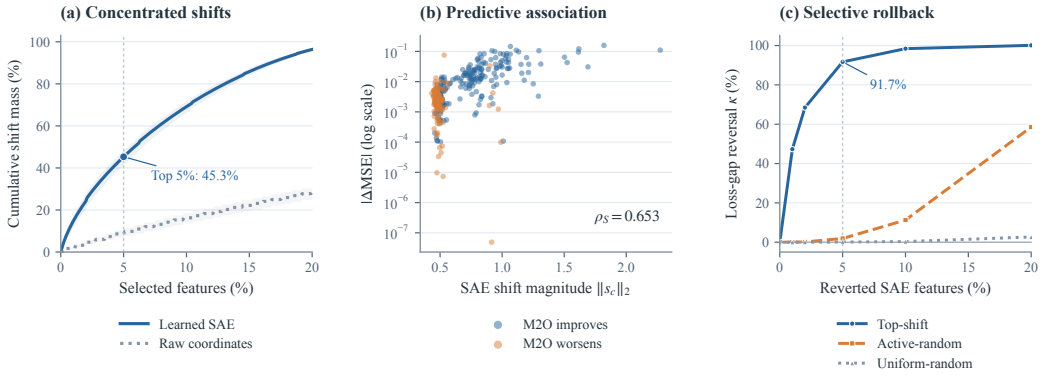


Figure 5: **Three SAE observations in TimeXer on Electricity** . (a) Top-5% features capture 45.3% of shift mass, versus 8.6% for raw coordinates. (b) $\rho_S = 0.653$. (c) Restoring 51 of 1,024 features gives 91.7% decoded loss-gap reversal, versus 1.90% for active-random and 0.07% for uniform-random.

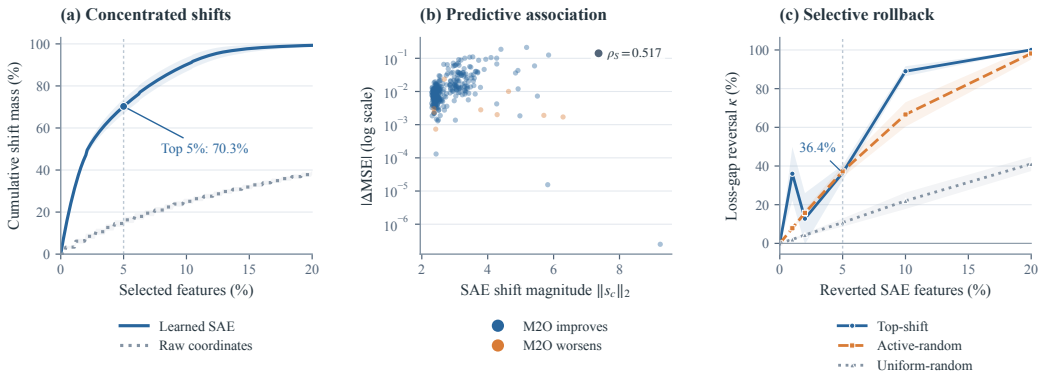


Figure 6: **Three SAE observations in DUET on Electricity.** (a) Top-5% features capture 70.3% of shift mass, versus 14.6% for raw. (b) $\rho_S = 0.517$. (c) Restoring 51 features gives 36.4% decoded loss-gap reversal, versus 37.2% for active-random and 10.8% for uniform-random.

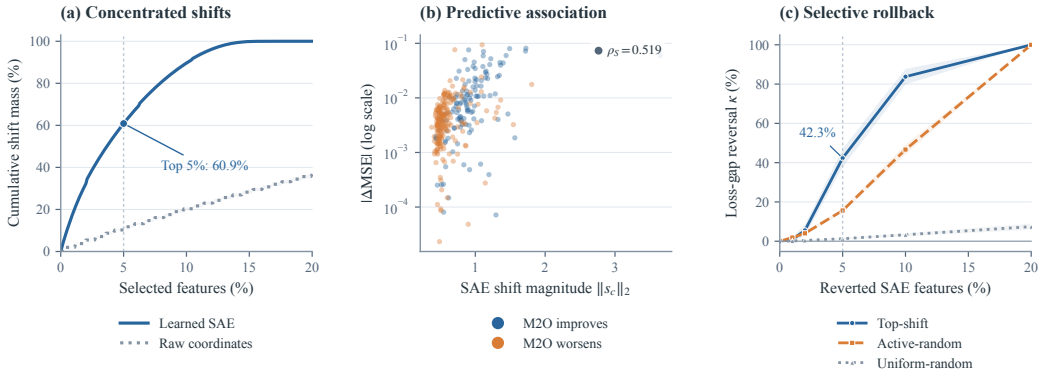


Figure 7: **Three SAE observations in CrossLinear on Electricity.** (a) Top-5% features capture 60.9% of shift mass, versus 10.2% for raw. (b) $\rho_S = 0.519$. (c) Restoring 51 features gives 42.3% decoded loss-gap reversal, versus 15.6% for active-random and 1.3% for uniform-random.

C Full Forecasting Results

Table C.1 complements the MSE results in Table 1 with horizon-averaged MAE. Tables C.2–C.5 report MSE and MAE for seven datasets, four backbones, three systems, and all four horizons. E2E M2O is trained independently; Adapted M2O starts from O2O; SPAR adds calibration and sparse correction to the frozen models.

Aggregation and units. For metric $Q \in \{\mathrm{MSE}, \mathrm{MAE}\}$, dataset d, backbone m, and system v, we average the four horizon means before rounding:

$$\overline{\mu}^{v}_{d,m,Q} = \frac{1}{4} \sum_{H \in \{96,192,336,720\}} \mu^{v}_{d,m,H,Q}. \tag{C1}$$

For each seed, we first average the metric over $H \in \{96, 192, 336, 720\}$. We then report the mean and sample standard deviation of these horizon-averaged values across seeds. MAE Avg. standard deviations retain the supplied summaries. Summary tables use units of 10^{-3}; per-horizon tables use original units. Bold marks the lowest displayed summary mean within each backbone and dataset; displayed ties are unbolded.

Relative gains. The SPAR Avg.Gain column compares with Adapted M2O within each dataset and then averages the seven relative reductions equally:

$$\mathrm{Avg.Gain}_{m,Q} = \frac{100}{7} \sum_{d=1}^{7} \frac{\overline{\mu}^{\text{Adapted M2O}}_{d,m,Q} - \overline{\mu}^{\text{SPAR}}_{d,m,Q}}{\overline{\mu}^{\text{Adapted M2O}}_{d,m,Q}} \%. \tag{C2}$$

This uses unrounded horizon means. The separate plug-in comparison uses its own Base results and supplied precision (Appendix C.2).

Table C.1: Forecasting performance. MAE in units of 10^{-3} (↓).

System	Traffic	Electricity	ETTh2	ETTm2	ETTh1	ETTm1	Weather	Avg.Gain
iTransformer								
E2E M2O	283.48±0.6	420.08±0.8	324.85±0.6	261.70±0.5	205.85±0.5	175.15±0.4	32.73±0.1	
Adapted M2O	292.95±0.4	430.65±1.0	333.15±0.8	269.58±0.6	212.40±0.5	180.38±0.4	33.90±0.2	
SPAR	**265.05**±0.8	**393.78**±0.8	**313.95**±0.6	**252.53**±0.4	**198.63**±0.3	**170.25**±0.3	**31.88**±0.2	**6.89%**
TimeXer								
E2E M2O	280.20±0.5	399.48±0.8	317.63±0.5	255.40±0.4	203.88±0.4	173.73±0.4	33.60±0.1	
Adapted M2O	287.85±0.9	415.40±1.0	326.98±0.7	264.45±0.6	211.23±0.4	179.30±0.4	34.45±0.2	
SPAR	**261.08**±0.7	**374.95**±0.7	**306.03**±0.4	**246.90**±0.4	**197.85**±0.4	**168.60**±0.3	**32.08**±0.1	**7.33%**
CrossLinear								
E2E M2O	277.45±0.8	397.73±0.9	316.25±0.6	253.75±0.5	203.25±0.4	174.40±0.4	33.55±0.1	
Adapted M2O	287.03±0.6	412.13±1.0	326.53±0.7	261.55±0.7	210.20±0.6	180.10±0.5	34.78±0.2	
SPAR	**260.18**±0.7	**374.23**±0.8	**303.95**±0.5	**244.95**±0.5	**197.80**±0.5	**169.18**±0.3	**31.80**±0.1	**7.48%**
DUET								
E2E M2O	273.75±0.5	396.60±0.8	315.73±0.6	254.48±0.5	205.23±0.4	174.50±0.3	31.55±0.1	
Adapted M2O	281.83±0.8	408.58±1.2	328.23±0.7	263.63±0.5	211.70±0.5	180.88±0.4	32.50±0.2	
SPAR	**256.58**±0.7	**374.65**±0.9	**304.68**±0.6	**246.18**±0.4	**199.35**±0.4	**169.70**±0.4	**30.50**±0.1	**7.03%**

Notes. MAE means are averaged over $H \in \{96, 192, 336, 720\}$; ± retains the supplied MAE Avg. standard deviations. Multiply displayed means and standard deviations by 10^{-3} to recover original values. Avg.Gain is SPAR's relative MAE reduction from Adapted M2O, averaged equally over seven datasets using unrounded means. Bold marks the lowest displayed MAE per backbone; ties are unbolded.

Table C.2: Full forecasting results for iTransformer. Per-horizon entries retain the supplied mean ± standard deviation. MSE Avg. rows show the mean of horizon means ± the standard deviations. MAE Avg. rows retain the supplied summaries.

Dataset	H	MSE ↓			MAE ↓		
		E2E M2O	Adapted M2O	SPAR	E2E M2O	Adapted M2O	SPAR
Traffic	96	0.1562 ± 0.0009	0.1510 ± 0.0006	0.1405 ± 0.0010	0.2781 ± 0.0009	0.2905 ± 0.0006	0.2625 ± 0.0004
	192	0.1569 ± 0.0006	0.1523 ± 0.0005	0.1366 ± 0.0009	0.2794 ± 0.0009	0.2866 ± 0.0006	0.2592 ± 0.0009
	336	0.1547 ± 0.0014	0.1503 ± 0.0013	0.1360 ± 0.0007	0.2781 ± 0.0005	0.2859 ± 0.0006	0.2602 ± 0.0012
	720	0.1770 ± 0.0009	0.1720 ± 0.0008	0.1556 ± 0.0008	0.2983 ± 0.0007	0.3088 ± 0.0006	0.2783 ± 0.0013
	Avg.	0.16120 ± 0.00095	0.15640 ± 0.00080	0.14218 ± 0.00085	0.2835 ± 0.0006	0.2930 ± 0.0004	0.2651 ± 0.0008
Electricity	96	0.2992 ± 0.0012	0.2993 ± 0.0012	0.2650 ± 0.0015	0.3803 ± 0.0008	0.3885 ± 0.0008	0.3575 ± 0.0007
	192	0.3205 ± 0.0009	0.3212 ± 0.0007	0.2849 ± 0.0007	0.3947 ± 0.0014	0.4068 ± 0.0008	0.3684 ± 0.0013
	336	0.3812 ± 0.0022	0.3790 ± 0.0022	0.3353 ± 0.0014	0.4298 ± 0.0014	0.4409 ± 0.0016	0.4014 ± 0.0008
	720	0.4623 ± 0.0013	0.4614 ± 0.0013	0.4152 ± 0.0010	0.4755 ± 0.0009	0.4864 ± 0.0017	0.4478 ± 0.0015
	Avg.	0.36580 ± 0.00140	0.36523 ± 0.00135	0.32510 ± 0.00115	0.4201 ± 0.0008	0.4307 ± 0.0010	0.3938 ± 0.0008
ETTh2	96	0.1371 ± 0.0007	0.1371 ± 0.0007	0.1306 ± 0.0005	0.2698 ± 0.0006	0.2805 ± 0.0010	0.2617 ± 0.0004
	192	0.1865 ± 0.0010	0.1870 ± 0.0011	0.1761 ± 0.0008	0.3158 ± 0.0007	0.3228 ± 0.0007	0.3054 ± 0.0009
	336	0.2210 ± 0.0014	0.2210 ± 0.0014	0.2090 ± 0.0015	0.3442 ± 0.0008	0.3522 ± 0.0013	0.3317 ± 0.0011
	720	0.2541 ± 0.0007	0.2536 ± 0.0007	0.2377 ± 0.0010	0.3696 ± 0.0009	0.3771 ± 0.0008	0.3570 ± 0.0008
	Avg.	0.19968 ± 0.00095	0.19968 ± 0.00098	0.18835 ± 0.00095	0.3249 ± 0.0006	0.3332 ± 0.0008	0.3140 ± 0.0006
ETTm2	96	0.0733 ± 0.0007	0.0713 ± 0.0008	0.0672 ± 0.0003	0.1983 ± 0.0008	0.2039 ± 0.0003	0.1919 ± 0.0004
	192	0.1122 ± 0.0009	0.1082 ± 0.0009	0.1027 ± 0.0004	0.2448 ± 0.0007	0.2514 ± 0.0008	0.2383 ± 0.0004
	336	0.1401 ± 0.0006	0.1400 ± 0.0004	0.1306 ± 0.0003	0.2792 ± 0.0004	0.2893 ± 0.0010	0.2688 ± 0.0003
	720	0.1890 ± 0.0007	0.1882 ± 0.0007	0.1755 ± 0.0005	0.3245 ± 0.0004	0.3337 ± 0.0009	0.3111 ± 0.0009
	Avg.	0.12865 ± 0.00073	0.12693 ± 0.00070	0.11900 ± 0.00038	0.2617 ± 0.0005	0.2696 ± 0.0006	0.2525 ± 0.0004
ETTh1	96	0.0572 ± 0.0004	0.0572 ± 0.0004	0.0549 ± 0.0003	0.1793 ± 0.0006	0.1840 ± 0.0007	0.1741 ± 0.0003
	192	0.0742 ± 0.0006	0.0748 ± 0.0006	0.0700 ± 0.0007	0.2055 ± 0.0003	0.2143 ± 0.0005	0.1986 ± 0.0004
	336	0.0849 ± 0.0006	0.0846 ± 0.0006	0.0796 ± 0.0005	0.2191 ± 0.0008	0.2266 ± 0.0007	0.2113 ± 0.0005
	720	0.0856 ± 0.0008	0.0845 ± 0.0008	0.0791 ± 0.0008	0.2195 ± 0.0008	0.2247 ± 0.0008	0.2105 ± 0.0006
	Avg.	0.07548 ± 0.00060	0.07528 ± 0.00060	0.07090 ± 0.00058	0.2059 ± 0.0005	0.2124 ± 0.0005	0.1986 ± 0.0003
ETTm1	96	0.0288 ± 0.0005	0.0290 ± 0.0005	0.0278 ± 0.0005	0.1307 ± 0.0005	0.1337 ± 0.0004	0.1269 ± 0.0002
	192	0.0449 ± 0.0006	0.0450 ± 0.0006	0.0437 ± 0.0005	0.1632 ± 0.0006	0.1665 ± 0.0004	0.1583 ± 0.0006
	336	0.0599 ± 0.0006	0.0603 ± 0.0006	0.0579 ± 0.0005	0.1894 ± 0.0004	0.1972 ± 0.0006	0.1843 ± 0.0004
	720	0.0788 ± 0.0004	0.0790 ± 0.0004	0.0754 ± 0.0005	0.2173 ± 0.0006	0.2241 ± 0.0006	0.2115 ± 0.0004
	Avg.	0.05310 ± 0.00053	0.05333 ± 0.00053	0.05120 ± 0.00050	0.1752 ± 0.0004	0.1804 ± 0.0004	0.1703 ± 0.0003
Weather	96	0.0010 ± 0.0001	0.0010 ± 0.0001	0.0011 ± 0.0001	0.0241 ± 0.0001	0.0249 ± 0.0001	0.0235 ± 0.0001
	192	0.0022 ± 0.0001	0.0022 ± 0.0001	0.0022 ± 0.0002	0.0358 ± 0.0001	0.0369 ± 0.0002	0.0349 ± 0.0001
	336	0.0020 ± 0.0002	0.0020 ± 0.0002	0.0020 ± 0.0001	0.0342 ± 0.0002	0.0358 ± 0.0002	0.0332 ± 0.0003
	720	0.0023 ± 0.0002	0.0023 ± 0.0002	0.0023 ± 0.0001	0.0368 ± 0.0002	0.0380 ± 0.0003	0.0359 ± 0.0003
	Avg.	0.00188 ± 0.00015	0.00188 ± 0.00015	0.00190 ± 0.00013	0.0327 ± 0.0001	0.0339 ± 0.0002	0.0319 ± 0.0002

Table C.3: Full forecasting results for TimeXer. Per-horizon entries retain the supplied mean ± standard deviation. MSE Avg. rows show the mean of horizon means ± the standard deviations. MAE Avg. rows retain the supplied summaries.

Dataset	H	MSE ↓			MAE ↓		
		E2E M2O	Adapted M2O	SPAR	E2E M2O	Adapted M2O	SPAR
Traffic	96	0.1423 ± 0.0005	0.1497 ± 0.0007	0.1310 ± 0.0009	0.2744 ± 0.0006	0.2805 ± 0.0011	0.2556 ± 0.0009
	192	0.1438 ± 0.0008	0.1496 ± 0.0008	0.1338 ± 0.0005	0.2762 ± 0.0008	0.2873 ± 0.0009	0.2575 ± 0.0006
	336	0.1510 ± 0.0012	0.1485 ± 0.0014	0.1313 ± 0.0012	0.2751 ± 0.0005	0.2820 ± 0.0014	0.2547 ± 0.0009
	720	0.1514 ± 0.0013	0.1610 ± 0.0010	0.1416 ± 0.0007	0.2951 ± 0.0007	0.3016 ± 0.0012	0.2765 ± 0.0008
	Avg.	0.14713 ± 0.00095	0.15220 ± 0.00098	0.13443 ± 0.00083	0.2802 ± 0.0005	0.2879 ± 0.0009	0.2611 ± 0.0007
Electricity	96	0.2608 ± 0.0003	0.2611 ± 0.0007	0.2353 ± 0.0004	0.3561 ± 0.0006	0.3728 ± 0.0009	0.3370 ± 0.0007
	192	0.3249 ± 0.0007	0.3212 ± 0.0007	0.2804 ± 0.0005	0.3929 ± 0.0012	0.4057 ± 0.0014	0.3693 ± 0.0013
	336	0.3758 ± 0.0012	0.3790 ± 0.0012	0.3250 ± 0.0016	0.4245 ± 0.0008	0.4438 ± 0.0008	0.3972 ± 0.0008
	720	0.4610 ± 0.0009	0.4614 ± 0.0009	0.3210 ± 0.0013	0.4244 ± 0.0014	0.4393 ± 0.0016	0.3963 ± 0.0009
	Avg.	0.35563 ± 0.00078	0.35568 ± 0.00088	0.29043 ± 0.00095	0.3995 ± 0.0008	0.4154 ± 0.0010	0.3750 ± 0.0007
ETTh2	96	0.1311 ± 0.0006	0.1320 ± 0.0006	0.1247 ± 0.0008	0.2656 ± 0.0006	0.2746 ± 0.0004	0.2559 ± 0.0006
	192	0.1808 ± 0.0009	0.1813 ± 0.0007	0.1710 ± 0.0009	0.3120 ± 0.0007	0.3206 ± 0.0011	0.3014 ± 0.0003
	336	0.2226 ± 0.0004	0.2230 ± 0.0004	0.2110 ± 0.0007	0.3469 ± 0.0006	0.3559 ± 0.0008	0.3354 ± 0.0005
	720	0.2213 ± 0.0004	0.2207 ± 0.0007	0.2056 ± 0.0006	0.3460 ± 0.0008	0.3568 ± 0.0011	0.3314 ± 0.0006
	Avg.	0.18895 ± 0.00058	0.18925 ± 0.00060	0.17808 ± 0.00075	0.3176 ± 0.0005	0.3270 ± 0.0007	0.3060 ± 0.0004
ETTm2	96	0.0673 ± 0.0003	0.0673 ± 0.0003	0.0635 ± 0.0005	0.1933 ± 0.0003	0.2011 ± 0.0007	0.1869 ± 0.0003
	192	0.1012 ± 0.0004	0.1012 ± 0.0004	0.0966 ± 0.0006	0.2376 ± 0.0008	0.2457 ± 0.0008	0.2307 ± 0.0005
	336	0.1320 ± 0.0011	0.1302 ± 0.0009	0.1241 ± 0.0004	0.2701 ± 0.0009	0.2762 ± 0.0009	0.2626 ± 0.0008
	720	0.1851 ± 0.0013	0.1825 ± 0.0013	0.1690 ± 0.0017	0.3206 ± 0.0004	0.3348 ± 0.0008	0.3074 ± 0.0006
	Avg.	0.12140 ± 0.00078	0.12030 ± 0.00073	0.11330 ± 0.00080	0.2554 ± 0.0004	0.2645 ± 0.0006	0.2469 ± 0.0004
ETTh1	96	0.0577 ± 0.0006	0.0576 ± 0.0007	0.0543 ± 0.0003	0.1805 ± 0.0007	0.1877 ± 0.0005	0.1744 ± 0.0005
	192	0.0709 ± 0.0004	0.0712 ± 0.0004	0.0686 ± 0.0004	0.2012 ± 0.0006	0.2100 ± 0.0006	0.1952 ± 0.0006
	336	0.0805 ± 0.0009	0.0802 ± 0.0007	0.0760 ± 0.0004	0.2140 ± 0.0005	0.2197 ± 0.0006	0.2080 ± 0.0004
	720	0.0840 ± 0.0005	0.0842 ± 0.0005	0.0807 ± 0.0009	0.2198 ± 0.0004	0.2275 ± 0.0004	0.2138 ± 0.0007
	Avg.	0.07328 ± 0.00060	0.07330 ± 0.00058	0.06990 ± 0.00050	0.2039 ± 0.0004	0.2112 ± 0.0004	0.1979 ± 0.0004
ETTm1	96	0.0284 ± 0.0003	0.0282 ± 0.0003	0.0270 ± 0.0002	0.1293 ± 0.0006	0.1338 ± 0.0005	0.1256 ± 0.0002
	192	0.0434 ± 0.0003	0.0432 ± 0.0003	0.0412 ± 0.0004	0.1604 ± 0.0003	0.1674 ± 0.0003	0.1554 ± 0.0006
	336	0.05796 ± 0.0007	0.0582 ± 0.0007	0.0554 ± 0.0005	0.1866 ± 0.0005	0.1905 ± 0.0007	0.1811 ± 0.0004
	720	0.0790 ± 0.0004	0.0795 ± 0.0004	0.0758 ± 0.0004	0.2186 ± 0.0008	0.2255 ± 0.0007	0.2123 ± 0.0006
	Avg.	0.05219 ± 0.00043	0.05228 ± 0.00043	0.04985 ± 0.00038	0.1737 ± 0.0004	0.1793 ± 0.0004	0.1686 ± 0.0003
Weather	96	0.0012 ± 0.0001	0.0012 ± 0.0001	0.0012 ± 0.0001	0.0265 ± 0.0001	0.0272 ± 0.0003	0.0257 ± 0.0001
	192	0.0023 ± 0.0002	0.0023 ± 0.0002	0.0020 ± 0.0001	0.0367 ± 0.0001	0.0375 ± 0.0003	0.0341 ± 0.0001
	336	0.0020 ± 0.0001	0.0020 ± 0.0001	0.0021 ± 0.0002	0.0343 ± 0.0003	0.0352 ± 0.0002	0.0333 ± 0.0002
	720	0.0023 ± 0.0001	0.0023 ± 0.0001	0.0021 ± 0.0001	0.0369 ± 0.0001	0.0379 ± 0.0003	0.0352 ± 0.0002
	Avg.	0.00195 ± 0.00013	0.00195 ± 0.00013	0.00185 ± 0.00013	0.0336 ± 0.0001	0.0345 ± 0.0002	0.0321 ± 0.0001

Table C.4: Full forecasting results for CrossLinear. Per-horizon entries retain the supplied mean ± standard deviation. MSE Avg. rows show the mean of horizon means ± the standard deviations. MAE Avg. rows retain the supplied summaries.

Dataset	H	MSE ↓			MAE ↓		
		E2E M2O	Adapted M2O	SPAR	E2E M2O	Adapted M2O	SPAR
Traffic	96	0.1392 ± 0.0010	0.1405 ± 0.0010	0.1261 ± 0.0009	0.2741 ± 0.0011	0.2812 ± 0.0004	0.2559 ± 0.0006
	192	0.1407 ± 0.0013	0.1469 ± 0.0006	0.1242 ± 0.0006	0.2748 ± 0.0004	0.2831 ± 0.0007	0.2575 ± 0.0009
	336	0.1496 ± 0.0011	0.1547 ± 0.0014	0.1244 ± 0.0007	0.2744 ± 0.0013	0.2840 ± 0.0012	0.2574 ± 0.0011
	720	0.1655 ± 0.0010	0.1770 ± 0.0009	0.1488 ± 0.0011	0.2865 ± 0.0013	0.2998 ± 0.0008	0.2699 ± 0.0008
	Avg.	0.14875 ± 0.00110	0.15478 ± 0.00098	0.13088 ± 0.00083	0.2775 ± 0.0008	0.2870 ± 0.0006	0.2602 ± 0.0007
Electricity	96	0.2510 ± 0.0010	0.2508 ± 0.0011	0.2278 ± 0.0024	0.3505 ± 0.0009	0.3626 ± 0.0007	0.3325 ± 0.0011
	192	0.2940 ± 0.0013	0.2944 ± 0.0013	0.2647 ± 0.0012	0.3805 ± 0.0007	0.3967 ± 0.0008	0.3603 ± 0.0011
	336	0.3444 ± 0.0016	0.3439 ± 0.0016	0.3036 ± 0.0012	0.4123 ± 0.0015	0.4254 ± 0.0017	0.3842 ± 0.0009
	720	0.4009 ± 0.0007	0.4032 ± 0.0007	0.3607 ± 0.0011	0.4476 ± 0.0010	0.4638 ± 0.0015	0.4199 ± 0.0010
	Avg.	0.32258 ± 0.00115	0.32308 ± 0.00118	0.28920 ± 0.00148	0.3977 ± 0.0009	0.4121 ± 0.0010	0.3742 ± 0.0008
ETTh2	96	0.1308 ± 0.0004	0.1314 ± 0.0004	0.1230 ± 0.0006	0.2659 ± 0.0005	0.2768 ± 0.0006	0.2566 ± 0.0004
	192	0.1766 ± 0.0003	0.1784 ± 0.0003	0.1674 ± 0.0008	0.3108 ± 0.0007	0.3174 ± 0.0007	0.3004 ± 0.0009
	336	0.2141 ± 0.0006	0.2141 ± 0.0009	0.1982 ± 0.0009	0.3411 ± 0.0005	0.3540 ± 0.0014	0.3265 ± 0.0007
	720	0.2210 ± 0.0009	0.2208 ± 0.0007	0.2051 ± 0.0007	0.3472 ± 0.0011	0.3579 ± 0.0013	0.3323 ± 0.0005
	Avg.	0.18563 ± 0.00055	0.18618 ± 0.00058	0.17343 ± 0.00075	0.3163 ± 0.0006	0.3265 ± 0.0007	0.3040 ± 0.0005
ETTm2	96	0.0641 ± 0.0008	0.0640 ± 0.0008	0.0612 ± 0.0005	0.1891 ± 0.0006	0.1957 ± 0.0006	0.1837 ± 0.0006
	192	0.0978 ± 0.00011	0.0983 ± 0.0011	0.0926 ± 0.0003	0.2349 ± 0.0007	0.2407 ± 0.0010	0.2261 ± 0.0006
	336	0.1308 ± 0.0008	0.1282 ± 0.0008	0.1211 ± 0.0005	0.2689 ± 0.0009	0.2787 ± 0.0011	0.2599 ± 0.0007
	720	0.1805 ± 0.0004	0.1830 ± 0.0004	0.1705 ± 0.0008	0.3221 ± 0.0004	0.3311 ± 0.0008	0.3101 ± 0.0003
	Avg.	0.11830 ± 0.00053	0.11838 ± 0.00078	0.11135 ± 0.00053	0.2538 ± 0.0005	0.2616 ± 0.0007	0.2450 ± 0.0005
ETTh1	96	0.0552 ± 0.0003	0.0550 ± 0.0003	0.0532 ± 0.0004	0.1770 ± 0.0003	0.1854 ± 0.0007	0.1728 ± 0.0004
	192	0.0729 ± 0.0007	0.0726 ± 0.0007	0.0697 ± 0.0005	0.2038 ± 0.0008	0.2095 ± 0.0006	0.1987 ± 0.0006
	336	0.0819 ± 0.0003	0.0820 ± 0.0003	0.0779 ± 0.0007	0.2171 ± 0.0004	0.2227 ± 0.0007	0.2100 ± 0.0005
	720	0.0811 ± 0.0005	0.0801 ± 0.0005	0.0767 ± 0.0004	0.2151 ± 0.0006	0.2232 ± 0.0010	0.2097 ± 0.0008
	Avg.	0.07278 ± 0.00045	0.07243 ± 0.00045	0.06938 ± 0.00050	0.2033 ± 0.0004	0.2102 ± 0.0006	0.1978 ± 0.0005
ETTm1	96	0.0284 ± 0.0005	0.0280 ± 0.0005	0.0273 ± 0.0004	0.1293 ± 0.0003	0.1345 ± 0.0007	0.1257 ± 0.0004
	192	0.0444 ± 0.0003	0.0442 ± 0.0003	0.0426 ± 0.0003	0.1628 ± 0.0007	0.1674 ± 0.0003	0.1581 ± 0.0004
	336	0.0567 ± 0.0004	0.0570 ± 0.0004	0.0558 ± 0.0003	0.1853 ± 0.0007	0.1912 ± 0.0008	0.1801 ± 0.0005
	720	0.0799 ± 0.0004	0.0801 ± 0.0004	0.0760 ± 0.0003	0.2202 ± 0.0003	0.2273 ± 0.0009	0.2128 ± 0.0005
	Avg.	0.05235 ± 0.00040	0.05233 ± 0.00040	0.05043 ± 0.00033	0.1744 ± 0.0004	0.1801 ± 0.0005	0.1692 ± 0.0003
Weather	96	0.0013 ± 0.0001	0.0013 ± 0.0001	0.0011 ± 0.0001	0.0276 ± 0.0002	0.0283 ± 0.0002	0.0252 ± 0.0001
	192	0.0021 ± 0.0001	0.0021 ± 0.0001	0.0022 ± 0.0001	0.0352 ± 0.0001	0.0369 ± 0.0002	0.0343 ± 0.0002
	336	0.0020 ± 0.0001	0.0020 ± 0.0001	0.0023 ± 0.0001	0.0344 ± 0.0001	0.0354 ± 0.0003	0.0334 ± 0.0003
	720	0.0023 ± 0.0001	0.0023 ± 0.0001	0.0020 ± 0.0001	0.0370 ± 0.0003	0.0385 ± 0.0003	0.0343 ± 0.0001
	Avg.	0.00193 ± 0.00010	0.00193 ± 0.00010	0.00190 ± 0.00010	0.0336 ± 0.0001	0.0348 ± 0.0002	0.0318 ± 0.0001

Table C.5: Full forecasting results for DUET. Per-horizon entries retain the supplied mean ± standard deviation. MSE Avg. rows show the mean of horizon means ± the standard deviations. MAE Avg. rows retain the supplied summaries.

Dataset	H	MSE ↓			MAE ↓		
		E2E M2O	Adapted M2O	SPAR	E2E M2O	Adapted M2O	SPAR
Traffic	96	0.1510 ± 0.0006	0.1497 ± 0.0007	0.1423 ± 0.0005	0.2653 ± 0.0004	0.2747 ± 0.0006	0.2503 ± 0.0005
	192	0.1366 ± 0.0009	0.1338 ± 0.0008	0.1323 ± 0.0005	0.2674 ± 0.0005	0.2753 ± 0.0010	0.2494 ± 0.0007
	336	0.1360 ± 0.0007	0.1370 ± 0.0012	0.1303 ± 0.0013	0.2708 ± 0.0005	0.2774 ± 0.0014	0.2505 ± 0.0010
	720	0.1556 ± 0.0008	0.1514 ± 0.0013	0.1320 ± 0.0008	0.2915 ± 0.0009	0.2999 ± 0.0013	0.2761 ± 0.0013
	Avg.	0.14480 ± 0.00075	0.14298 ± 0.00100	0.13423 ± 0.00078	0.2738 ± 0.0005	0.2818 ± 0.0008	0.2566 ± 0.0007
Electricity	96	0.2557 ± 0.0005	0.2562 ± 0.0005	0.2267 ± 0.0006	0.3554 ± 0.0010	0.3655 ± 0.0011	0.3315 ± 0.0011
	192	0.2919 ± 0.0011	0.2922 ± 0.0011	0.2677 ± 0.0008	0.3805 ± 0.0007	0.3927 ± 0.0014	0.3616 ± 0.0009
	336	0.3392 ± 0.0009	0.3394 ± 0.0009	0.3077 ± 0.0015	0.4111 ± 0.0008	0.4222 ± 0.0017	0.3902 ± 0.0011
	720	0.3855 ± 0.0017	0.3858 ± 0.0015	0.3472 ± 0.0012	0.4394 ± 0.0013	0.4539 ± 0.0018	0.4153 ± 0.0012
	Avg.	0.31808 ± 0.00105	0.31840 ± 0.00100	0.28733 ± 0.00103	0.3966 ± 0.0008	0.4086 ± 0.0012	0.3747 ± 0.0009
ETTh2	96	0.1332 ± 0.0008	0.1303 ± 0.0008	0.1246 ± 0.0009	0.2657 ± 0.0006	0.2772 ± 0.0004	0.2585 ± 0.0005
	192	0.1813 ± 0.0005	0.1803 ± 0.0005	0.1673 ± 0.0008	0.3133 ± 0.0005	0.3280 ± 0.0010	0.3011 ± 0.0007
	336	0.2090 ± 0.0012	0.2091 ± 0.0012	0.1976 ± 0.0007	0.3382 ± 0.0012	0.3498 ± 0.0006	0.3264 ± 0.0005
	720	0.2233 ± 0.0009	0.2174 ± 0.0009	0.2035 ± 0.0009	0.3457 ± 0.0007	0.3579 ± 0.0012	0.3327 ± 0.0010
	Avg.	0.18670 ± 0.00085	0.18428 ± 0.00085	0.17325 ± 0.00083	0.3157 ± 0.0006	0.3282 ± 0.0007	0.3047 ± 0.0006
ETTm2	96	0.0644 ± 0.0009	0.0642 ± 0.0006	0.0611 ± 0.0005	0.1900 ± 0.0008	0.1971 ± 0.0008	0.1839 ± 0.0005
	192	0.0992 ± 0.0007	0.0988 ± 0.0007	0.0929 ± 0.0003	0.2363 ± 0.0005	0.2463 ± 0.0006	0.2289 ± 0.0005
	336	0.1291 ± 0.0006	0.1295 ± 0.0006	0.1221 ± 0.0006	0.2712 ± 0.0007	0.2778 ± 0.0004	0.2629 ± 0.0004
	720	0.1805 ± 0.0004	0.1799 ± 0.0004	0.1695 ± 0.0004	0.3204 ± 0.0008	0.3333 ± 0.0004	0.3090 ± 0.0008
	Avg.	0.11830 ± 0.00065	0.11810 ± 0.00058	0.11140 ± 0.00045	0.2545 ± 0.0005	0.2636 ± 0.0005	0.2462 ± 0.0004
ETTh1	96	0.0550 ± 0.0003	0.0554 ± 0.0003	0.0530 ± 0.0006	0.1782 ± 0.0004	0.1821 ± 0.0006	0.1735 ± 0.0005
	192	0.0715 ± 0.0006	0.0711 ± 0.0006	0.0686 ± 0.0007	0.2023 ± 0.0004	0.2090 ± 0.0008	0.1968 ± 0.0003
	336	0.0841 ± 0.0005	0.0840 ± 0.0005	0.0800 ± 0.0004	0.2205 ± 0.0008	0.2260 ± 0.0005	0.2137 ± 0.0007
	720	0.0833 ± 0.0007	0.0832 ± 0.0007	0.0793 ± 0.0006	0.2199 ± 0.0004	0.2297 ± 0.0005	0.2134 ± 0.0009
	Avg.	0.07348 ± 0.00053	0.07343 ± 0.00053	0.07023 ± 0.00058	0.2052 ± 0.0004	0.2117 ± 0.0005	0.1994 ± 0.0004
ETTm1	96	0.0277 ± 0.0003	0.0282 ± 0.0003	0.0271 ± 0.0002	0.1302 ± 0.0003	0.1340 ± 0.0005	0.1264 ± 0.0003
	192	0.0434 ± 0.0003	0.0434 ± 0.0003	0.0418 ± 0.0003	0.1618 ± 0.0004	0.1667 ± 0.0005	0.1576 ± 0.0003
	336	0.0574 ± 0.0007	0.0574 ± 0.0007	0.0555 ± 0.0005	0.1866 ± 0.0004	0.1933 ± 0.0007	0.1817 ± 0.0007
	720	0.0792 ± 0.0005	0.0790 ± 0.0005	0.0756 ± 0.0006	0.2194 ± 0.0005	0.2295 ± 0.0004	0.2131 ± 0.0006
	Avg.	0.05193 ± 0.00045	0.05200 ± 0.00045	0.05000 ± 0.00040	0.1745 ± 0.0003	0.1809 ± 0.0004	0.1697 ± 0.0004
Weather	96	0.0012 ± 0.0001	0.0012 ± 0.0001	0.0011 ± 0.0001	0.0266 ± 0.0001	0.0272 ± 0.0002	0.0254 ± 0.0001
	192	0.0014 ± 0.0001	0.0014 ± 0.0001	0.0013 ± 0.0001	0.0288 ± 0.0002	0.0296 ± 0.0002	0.0277 ± 0.0001
	336	0.0022 ± 0.0001	0.0022 ± 0.0001	0.0021 ± 0.0001	0.0362 ± 0.0002	0.0373 ± 0.0003	0.0352 ± 0.0002
	720	0.0020 ± 0.0002	0.0020 ± 0.0002	0.0020 ± 0.0002	0.0346 ± 0.0002	0.0359 ± 0.0003	0.0337 ± 0.0003
	Avg.	0.00170 ± 0.00013	0.00170 ± 0.00013	0.00163 ± 0.00013	0.0316 ± 0.0001	0.0325 ± 0.0002	0.0305 ± 0.0001

C.1 Full TimeXer Ablation Results

Tables C.6 and C.7 expand Table 3 to all four horizons on Electricity, Traffic, ETTh2, and ETTm2. The 11 variants comprise full SPAR, three component-removal variants, and seven mechanism replacements, all based on the same adapted TimeXer within each setting.

The final column further averages over datasets. These $\pm$ values summarize per-setting variability, not uncertainty of the aggregate estimator. Entries retain original units and four decimal places; the main table uses units of 10^{-3}. Percentage reductions in Section 5.2 use unrounded horizon means, compute the relative reduction within each dataset, and average equally over the four datasets.

Table C.6: Full TimeXer ablation results: Full SPAR and removal variants. MSE (↓) in original units.

Variant	H	Electricity	Traffic	ETTh2	ETTm2	Avg.
Full SPAR ($C + S$)	96	0.2353±0.0004	0.1310±0.0009	0.1247±0.0008	0.0635±0.0005	0.1386±0.0007
	192	0.2804±0.0005	0.1338±0.0005	0.1710±0.0009	0.0966±0.0006	0.1705±0.0006
	336	0.3250±0.0016	0.1313±0.0012	0.2110±0.0007	0.1241±0.0004	0.1979±0.0010
	720	0.3210±0.0013	0.1416±0.0007	0.2056±0.0006	0.1690±0.0017	0.2093±0.0011
	Avg.	0.2904±0.0010	0.1344±0.0008	0.1781±0.0008	0.1133±0.0008	0.1791±0.0008
w/o C (S only)	96	0.2422±0.0005	0.1358±0.0009	0.1266±0.0007	0.0645±0.0004	0.1423±0.0006
	192	0.2908±0.0006	0.1379±0.0006	0.1738±0.0008	0.0978±0.0005	0.1751±0.0006
	336	0.3389±0.0015	0.1359±0.0013	0.2140±0.0006	0.1257±0.0005	0.2036±0.0010
	720	0.3587±0.0012	0.1465±0.0008	0.2095±0.0006	0.1726±0.0016	0.2218±0.0011
	Avg.	0.3077±0.0010	0.1390±0.0009	0.1810±0.0007	0.1152±0.0008	0.1857±0.0008
w/o S (C only)	96	0.2597±0.0007	0.1489±0.0007	0.1313±0.0006	0.0669±0.0003	0.1517±0.0006
	192	0.3195±0.0007	0.1488±0.0008	0.1802±0.0007	0.1006±0.0004	0.1873±0.0007
	336	0.3770±0.0012	0.1476±0.0014	0.2216±0.0005	0.1294±0.0008	0.2189±0.0010
	720	0.4588±0.0009	0.1600±0.0010	0.2193±0.0007	0.1813±0.0014	0.2549±0.0010
	Avg.	0.3538±0.0009	0.1513±0.0010	0.1881±0.0006	0.1196±0.0007	0.2032±0.0008
w/o C, S (adapted M2O)	96	0.2611±0.0007	0.1497±0.0007	0.1320±0.0006	0.0673±0.0003	0.1525±0.0006
	192	0.3212±0.0007	0.1496±0.0008	0.1813±0.0007	0.1012±0.0004	0.1883±0.0007
	336	0.3790±0.0012	0.1485±0.0014	0.2230±0.0004	0.1302±0.0009	0.2202±0.0010
	720	0.4614±0.0009	0.1610±0.0010	0.2207±0.0007	0.1825±0.0013	0.2564±0.0010
	Avg.	0.3557±0.0009	0.1522±0.0010	0.1893±0.0006	0.1203±0.0007	0.2044±0.0008

Notes. Per-horizon dataset cells show the supplied mean $\pm$ standard deviation. Avg. rows average over horizons; the final column averages over datasets. The final entry of each Avg. row averages all 16 dataset–horizon entries.

Table C.7: Full TimeXer ablation results: Replacement variants. MSE (↓) in original units.

Variant	H	Electricity	Traffic	ETTh2	ETTm2	Avg.
Output-difference gate	96	0.2406±0.0005	0.1351±0.0009	0.1262±0.0007	0.0643±0.0004	0.1416±0.0006
	192	0.2893±0.0006	0.1370±0.0006	0.1731±0.0008	0.0976±0.0006	0.1743±0.0007
	336	0.3360±0.0016	0.1349±0.0012	0.2136±0.0006	0.1253±0.0005	0.2025±0.0010
	720	0.3501±0.0012	0.1458±0.0007	0.2087±0.0006	0.1718±0.0017	0.2191±0.0011
	Avg.	0.3040±0.0010	0.1382±0.0009	0.1804±0.0007	0.1148±0.0008	0.1843±0.0008
Active-random support	96	0.2510±0.0006	0.1422±0.0007	0.1290±0.0007	0.0658±0.0004	0.1470±0.0006
	192	0.3049±0.0006	0.1432±0.0007	0.1773±0.0008	0.0994±0.0005	0.1812±0.0007
	336	0.3570±0.0013	0.1418±0.0013	0.2182±0.0006	0.1277±0.0007	0.2112±0.0010
	720	0.4064±0.0010	0.1532±0.0009	0.2145±0.0007	0.1772±0.0014	0.2378±0.0010
	Avg.	0.3298±0.0009	0.1451±0.0009	0.1848±0.0007	0.1175±0.0008	0.1943±0.0008
Global support	96	0.2453±0.0005	0.1379±0.0009	0.1275±0.0007	0.0650±0.0004	0.1439±0.0006
	192	0.2956±0.0006	0.1398±0.0007	0.1750±0.0008	0.0983±0.0005	0.1772±0.0007
	336	0.3457±0.0015	0.1379±0.0013	0.2155±0.0006	0.1264±0.0006	0.2064±0.0010
	720	0.3751±0.0012	0.1488±0.0008	0.2114±0.0006	0.1742±0.0016	0.2274±0.0011
	Avg.	0.3154±0.0010	0.1411±0.0009	0.1824±0.0007	0.1160±0.0008	0.1887±0.0008
M2O activations (a^M)	96	0.2481±0.0006	0.1403±0.0008	0.1284±0.0007	0.0654±0.0004	0.1456±0.0006
	192	0.3006±0.0006	0.1418±0.0006	0.1761±0.0008	0.0989±0.0005	0.1794±0.0006
	336	0.3525±0.0014	0.1398±0.0013	0.2170±0.0006	0.1272±0.0007	0.2091±0.0010
	720	0.3904±0.0011	0.1512±0.0008	0.2133±0.0007	0.1757±0.0015	0.2327±0.0010
	Avg.	0.3229±0.0009	0.1433±0.0009	0.1837±0.0007	0.1168±0.0008	0.1917±0.0008
Static coefficients	96	0.2428±0.0005	0.1367±0.0008	0.1269±0.0008	0.0646±0.0004	0.1428±0.0006
	192	0.2929±0.0005	0.1386±0.0006	0.1740±0.0009	0.0980±0.0005	0.1759±0.0006
	336	0.3412±0.0014	0.1363±0.0013	0.2147±0.0006	0.1259±0.0005	0.2045±0.0010
	720	0.3620±0.0012	0.1476±0.0008	0.2101±0.0006	0.1729±0.0015	0.2232±0.0010
	Avg.	0.3097±0.0009	0.1398±0.0009	0.1814±0.0007	0.1154±0.0007	0.1866±0.0008
Mismatched phase	96	0.2560±0.0007	0.1461±0.0008	0.1305±0.0006	0.0665±0.0003	0.1498±0.0006
	192	0.3133±0.0007	0.1463±0.0008	0.1793±0.0007	0.1003±0.0004	0.1848±0.0007
	336	0.3677±0.0013	0.1451±0.0014	0.2207±0.0004	0.1289±0.0008	0.2156±0.0010
	720	0.4337±0.0010	0.1572±0.0009	0.2175±0.0007	0.1798±0.0014	0.2471±0.0010
	Avg.	0.3427±0.0009	0.1487±0.0010	0.1870±0.0006	0.1189±0.0007	0.1993±0.0008
Attenuation only	96	0.2469±0.0006	0.1392±0.0008	0.1279±0.0007	0.0652±0.0005	0.1448±0.0007
	192	0.2982±0.0006	0.1407±0.0006	0.1756±0.0008	0.0986±0.0005	0.1783±0.0006
	336	0.3485±0.0014	0.1390±0.0013	0.2162±0.0006	0.1268±0.0006	0.2076±0.0010
	720	0.3840±0.0011	0.1500±0.0008	0.2122±0.0007	0.1751±0.0015	0.2303±0.0010
	Avg.	0.3194±0.0009	0.1422±0.0009	0.1830±0.0007	0.1164±0.0008	0.1903±0.0008

Notes. Per-horizon dataset cells show the supplied mean ± standard deviation. Avg. rows average over horizons; the final column averages over datasets. The final entry of each Avg. row averages all 16 dataset–horizon entries.

C.2 Full Plug-in Comparison Results

Tables C.8 and C.9 provide the dataset-level results behind Table 2: both group results by backbone and report the supplied means and standard deviations in units of 10^{-3}. Each error already averages over $H \in \{96, 192, 336, 720\}$.

For each backbone and metric Q, the mean relative gain of plug-in v is $\frac{100}{7}\sum_{d=1}^{7}(e_{d,Q}^{\text{Base}} - e_{d,Q}^{v})/e_{d,Q}^{\text{Base}}$ percent, using the supplied means before rounding the final result to two decimal places. The denominator is that comparison's Base backbone, not Adapted M2O in Table 1. W/T/L counts lower/equal/higher errors across seven datasets and two metrics (14 comparisons), with ties defined at the supplied precision. Standard deviations do not determine these counts.

Table C.8: TimeXer and CrossLinear plug-in results. MSE and MAE in units of 10^{-3} (↓).

TimeXer

Dataset	Base		+CCM		+xCPD		+SPAR	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Traffic	156.00±0.90	235.00±1.00	154.00±1.03	232.00±1.10	151.00±0.60	230.00±0.70	**137.00**±0.60	**217.00**±0.50
Electricity	327.00±1.05	408.00±0.80	322.00±0.40	403.00±0.60	317.00±0.50	399.00±0.50	**290.00**±1.05	**381.00**±0.60
ETTh2	189.00±0.45	342.00±0.70	187.00±0.80	339.00±0.50	184.00±0.60	336.00±0.80	**178.00**±0.40	**330.00**±0.70
ETTm2	120.00±0.33	258.00±0.90	119.00±1.13	256.00±0.85	117.00±1.42	254.00±1.55	**113.00**±0.18	**249.00**±0.53
ETTh1	73.00±0.30	209.00±1.10	72.00±1.33	208.00±0.67	71.00±0.60	206.00±0.30	**70.00**±1.10	**204.00**±1.05
ETTm1	52.00±0.40	171.00±0.30	52.00±0.60	170.00±1.27	51.00±0.65	169.00±1.67	**50.00**±0.60	**167.00**±1.50
Weather	2.00±0.10	31.00±0.10	2.00±0.10	31.00±0.10	2.00±0.10	31.00±0.10	2.00±0.10	**30.00**±0.10

CrossLinear

Dataset	Base		+CCM		+xCPD		+SPAR	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Traffic	152.00±0.55	231.00±0.30	150.00±0.30	228.00±1.55	147.00±1.33	226.00±1.60	**135.00**±0.80	**216.00**±0.90
Electricity	323.00±1.20	405.00±0.60	318.00±0.70	400.00±1.67	313.00±0.50	396.00±1.30	**289.00**±0.90	**380.00**±1.10
ETTh2	186.00±0.40	339.00±0.60	184.00±0.40	336.00±0.90	181.00±0.40	333.00±1.10	**173.00**±0.40	**327.00**±0.80
ETTm2	118.00±0.45	254.00±1.20	117.00±0.50	252.00±0.70	115.00±0.40	250.00±0.80	**111.00**±0.40	**246.00**±1.20
ETTh1	72.00±0.30	209.00±0.80	71.00±0.60	208.00±0.80	70.00±0.50	206.00±1.20	**69.00**±0.40	**203.00**±0.80
ETTm1	52.00±0.30	170.00±0.80	52.00±0.20	169.00±0.80	51.00±0.60	168.00±1.10	**50.00**±0.33	**166.00**±1.10
Weather	2.00±0.10	32.00±0.10	2.00±0.10	32.00±0.10	2.00±0.10	32.00±0.10	2.00±0.10	**31.00**±0.10

Notes. Entries retain the supplied mean ± standard deviation for results averaged over $H \in \{96, 192, 336, 720\}$. Multiply both components by 10^{-3} to recover original units. Table 2 uses the means to compute relative gains and W/T/L. Bold marks the lowest mean for each backbone, dataset, and metric; tied minima are bold unless all methods tie.

Table C.9: iTransformer and DUET plug-in results. MSE and MAE in units of 10^{-3} (↓).

iTransformer

Dataset	Base		+CCM		+xCPD		+SPAR	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Traffic	161.20±1.00	283.50±0.60	161.00±1.40	283.00±2.60	150.40±0.70	282.00±0.60	**142.20**±0.90	**265.10**±0.80
Electricity	365.80±1.40	420.10±0.80	365.00±0.90	419.50±0.90	344.40±1.30	408.90±1.40	**325.10**±1.15	**393.80**±0.80
ETTh2	199.70±0.95	324.90±0.60	199.70±1.20	324.90±1.50	199.70±1.20	324.90±1.50	**188.35**±0.95	**314.00**±0.60
ETTm2	128.70±0.73	261.70±0.50	128.60±1.20	261.50±1.70	126.60±1.20	261.50±1.70	**119.00**±0.38	**252.50**±0.40
ETTh1	75.50±0.60	205.90±0.50	75.50±0.80	205.60±0.70	72.70±2.20	206.60±2.30	**70.90**±0.58	**198.60**±0.30
ETTm1	53.10±0.53	175.20±0.40	53.00±0.50	175.80±1.50	53.00±1.10	174.40±1.80	**51.20**±0.50	**170.30**±0.30
Weather	1.90±0.15	32.70±0.10	1.90±0.20	32.70±0.10	1.90±0.20	**31.00**±0.10	1.90±0.13	31.90±0.20

DUET

Dataset	Base		+CCM		+xCPD		+SPAR	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Traffic	144.80±0.80	273.80±0.50	143.30±0.40	273.50±0.50	143.00±2.10	272.90±1.10	**134.23**±0.78	**256.60**±0.70
Electricity	318.08±1.05	396.60±0.80	316.90±1.30	396.20±1.40	316.20±0.90	395.50±1.00	**287.33**±1.03	**374.70**±0.90
ETTh2	186.70±0.85	315.70±0.60	185.90±0.90	332.00±1.10	185.40±1.40	331.60±2.20	**173.25**±0.83	**304.70**±0.60
ETTm2	118.30±0.65	254.50±0.50	118.50±0.70	254.00±0.90	118.50±0.80	253.60±1.30	**111.40**±0.45	**246.20**±0.40
ETTh1	73.48±0.53	205.20±0.40	72.90±1.50	204.10±0.60	72.20±1.30	203.60±0.80	**70.23**±0.58	**199.40**±0.40
ETTm1	51.93±0.45	174.50±0.30	52.20±0.60	178.20±1.30	51.90±0.40	175.00±1.70	**50.00**±0.40	**169.70**±0.40
Weather	1.70±0.13	31.60±0.10	1.70±0.10	31.60±0.10	**1.40**±0.10	**30.50**±0.10	1.63±0.13	**30.50**±0.10

Notes. Entries retain the supplied horizon-averaged mean ± standard deviation. Multiply both components by 10^{-3} to recover original units; two decimals here preserve up to five decimals in the original values. Table 2 uses the means to compute relative gains and W/T/L, while the supplied standard deviations are retained here. Bold marks the lowest mean for each backbone, dataset, and metric; tied minima are bold unless all methods tie.

D EXPERIMENTAL HYPERPARAMETERS

The experiment uses an NVIDIA RTX 4090 D with float32 precision, with seeds set as 2024, 2025, 2026.

Backbone training. Both O2O training and M2O adaptation use Adam with learning rate 10^{-3} and batch size 32. O2O is trained for at most 30 epochs; M2O is initialized from the selected O2O checkpoint and adapted for at most 10 epochs. Checkpoints are selected by validation MSE, with early-stopping patience of five epochs.

SAE and SPAR. The TopK SAE has 1024 dictionary features and activation sparsity $k = 51$. It is fitted on O2O representations using Adam with learning rate 4×10^{-4}, batch size 4096, and 30–60 epochs, with early-stopping patience of five epochs after the minimum training period. SPAR fixes the correction budget at $K = 51$ per channel. Validation selects calibration parameters $\lambda_0 = 1$ and $\gamma = 0.25$, and sparse-correction parameters $\lambda = 0.1$ and $\alpha = 0.095$. All components are frozen at inference.