CoSTAR: Common-State Adaptive Refinement for Multivariate Time Series Forecasting

Anonymous authors
Paper under double-blind review

Abstract

Multivariate time-series forecasting is central to applications such as energy management, traffic prediction, and weather analysis, where channels exhibit both shared and channel-specific dynamics. Existing forecasters capture these dependencies through their internal architectures; however, incorporating an additional refinement signal commonly requires altering the backbone. To address this limitation, we propose **CoSTAR** (Common-State Adaptive Refinement), an external module that operates only through a forecaster's input and output interfaces. CoSTAR factorizes the history into a common component and a residual input, summarizes the common trajectory with four window-level statistics, and decodes a gated residual forecast correction. We evaluate CoSTAR on eight long-term benchmarks and four PEMS traffic datasets. Across paired comparisons of five backbones on six datasets, CoSTAR reduces horizon-averaged MSE in 29 of 30 backbone–dataset pairs, with one tie at the reported precision; the largest reduction is 72.4% for SparseTSF on PEMS04. Ablations, five-seed analyses, and frozen-backbone experiments further characterize the contribution and scope of the design.

1 Introduction

Multivariate time series forecasting supports applications in energy, traffic, and weather, where multiple channels evolve together (Wen et al., 2023). Forecasting architectures capture this structure in different ways: linear models emphasize temporal decomposition (Zeng et al., 2023), patch-based Transformers model local temporal segments (Nie et al., 2023), and variate-token or convolutional designs model dependencies across channels and time (Liu et al., 2024; Luo & Wang, 2024). These choices improve representation learning, but incorporating a new mechanism can require changes to the backbone. A complementary direction is to refine an existing forecaster through an external module.

A useful refinement signal should capture shared variation without requiring access to the backbone's hidden representations. Channel-independent models do not explicitly mix channels, whereas channel-dependent models can exploit their interactions at additional modeling cost (Nie et al., 2023; Zhang & Yan, 2023). Reversible instance normalization changes and restores per-instance statistics at the input and output to address distribution shift (Kim et al., 2022); it does not construct a cross-channel common state. Existing approaches address the channel-interaction tradeoff through channel clustering, leading indicators, temporal retrieval, or parameter generation (Chen et al., 2024; Zhao & Shen, 2024; Cao et al., 2026; Savchenko & Kachan, 2026). Global–local factor models and series-core fusion also show the value of separating shared information from individual series (Wang et al., 2019; Han et al., 2024). **Research question.** *Can a shared representation condition a backbone's input and correct its output through a single external pathway?*

We explore this question through the *common state*, a compact summary of a learned common component. The selected PEMS08 channels in Figure 1(a) illustrate temporal variation shared across channels alongside distinct local details. CoSTAR uses factorized channel projections to represent such common variation, subtracts a learned fraction from the backbone's input, and summarizes the common trajectory for output refinement. This coupling allows the backbone and the correction pathway to use different views of the same history. Figure 1(b) previews its empirical effect: attaching CoSTAR to five backbones reduces horizon-averaged MSE on PEMS04 by 20.8%–72.4%.

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.

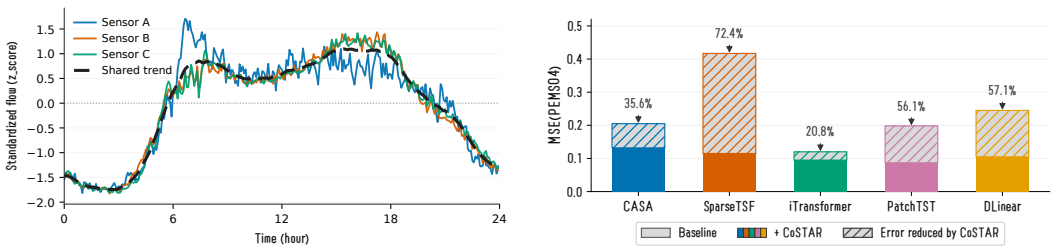

Figure 1: Motivation for common-state refinement. *(a)* The displayed PEMS08 channels exhibit a shared trend alongside channel-specific variation. *(b)* On PEMS04, attaching CoSTAR to five representative backbones lowers horizon-averaged MSE. The hatched region visualizes the reduction relative to each corresponding backbone, which ranges from 20.8% to 72.4% in this comparison.

CoSTAR comprises three stages. **Low-Rank State Factorization** constructs the common component and the residual input. **Statistical Context Encoding** summarizes the latent common trajectory through its mean, standard deviation, final state, and endpoint difference. **Adaptive Residual Correction** decodes this common state into a forecast correction and scales it with a learned gate. The factorized projection imposes a rank constraint when its latent width is smaller than the channel count. This is a representational bias toward shared structure, whose forecasting value we assess empirically.

The target setting of this work is therefore specific. CoSTAR is most useful when the input contains many channels whose dynamics exhibit measurable cross-channel regularity, while the underlying forecasting backbone has not already extracted that regularity as effectively as a dedicated channel model. In such settings, the common component provides a compact route for sharing information across channels, and the specific component preserves the local deviations required by the backbone. When a model already captures channel interactions particularly well, or when the channels are weakly coupled, the module is not expected to provide the same benefit. This scope is reflected in our experiments: we report results on both channel-rich traffic and electricity data and on smaller ETT datasets, rather than presenting CoSTAR as a universally dominant replacement for specialized architectures.

The distinction between these two branches is also important for interpreting the method. The common branch is not treated as a ground-truth latent factor, and the specific branch is not assumed to be statistically independent of it. They are two learned views of the same observed history. The usefulness of the decomposition is judged by the paired forecasting experiments, the component ablations, the cross-split correlation diagnostic, and the frozen-backbone study. This empirical framing keeps the architectural interpretation aligned with what the available results can support.

Our contributions are summarized as follows:

- **Common-state refinement.** We introduce an external module that couples input residualization, statistical common-state encoding, and gated output correction while preserving a forecaster's input and output shapes.
- **Compatibility across backbones.** Paired evaluations cover five backbones on six datasets. CoSTAR improves horizon-averaged MSE in 29 of 30 pairs, with one tie; frozen-backbone experiments also show that the module can be trained while the forecaster remains fixed.
- **Empirical characterization.** Long- and short-term comparisons, component ablations, lookback and hyperparameter analyses, five-seed results, and plug-in efficiency measurements characterize both the benefits and the limits of the design.

2 Related Work

Forecasting architectures and channel modeling. Transformer forecasters use sparse attention, temporal decomposition, patch tokens, or variate tokens to model time series (Zhou et al., 2021; Wu et al., 2021; Nie et al., 2023; Liu et al., 2024). Other designs use temporal 2D representations, convolution, or multiscale mixing (Wu et al., 2023; Luo & Wang, 2024; Wang et al., 2024). Cross-

former explicitly models cross-dimension dependencies, while temporal queries and local cross-variate modeling offer alternative interaction mechanisms (Zhang & Yan, 2023; Lin et al., 2025; Liu et al., 2026). CoSTAR complements these architectural choices through input and output refinement.

Shared representations for forecasting. Low-dimensional shared structure has a substantial history in forecasting. Temporal regularized matrix factorization models high-dimensional series through latent factors with temporal regularization (Yu et al., 2016). Deep Factors combines global latent structure with local models (Wang et al., 2019). More recently, SOFTS aggregates series representations into a global core and redistributes it through series-core fusion (Han et al., 2024). CoSTAR shares the motivation to expose information common to multiple series, but it neither redistributes a global core into per-series representations nor claims an identifiable latent factor. Instead, its common component conditions the residual input, and window-level statistics of that component drive an external forecast correction.

Forecasting modules and adaptation. CCM groups related channels, and LIFT uses leading indicators to refine multivariate forecasts (Chen et al., 2024; Zhao & Shen, 2024). GTR retrieves global temporal information, while HN-MVTS generates forecasting-layer parameters from learned channel embeddings (Cao et al., 2026; Savchenko & Kachan, 2026). CoRA targets correlation-aware adaptation of time series foundation models (Cheng et al., 2026). CoSTAR instead couples input residualization and output correction through a statistical common state. Its factorized channel map also differs from LoRA, which constrains weight updates for parameter-efficient adaptation (Hu et al., 2022). Our evaluation focuses on task-specific forecasters, including both joint training and frozen-backbone adaptation.

3 METHODOLOGY

3.1 PROBLEM STATEMENT AND OVERVIEW

Let $\mathbf{X} \in \mathbb{R}^{B\times L\times N}$ denote the input with batch size B, lookback length L, and N channels. The target is $\mathbf{Y} \in \mathbb{R}^{B\times H\times N}$, where H is the forecasting horizon. CoSTAR constructs a residual input $\mathbf{X}_{\text{specific}}$ for a backbone f_θ and a common state for a correction decoder. Its final prediction is

$$\hat{\mathbf{Y}} = f_\theta(\mathbf{X}_{\text{specific}}) + \alpha\, \Delta\hat{\mathbf{Y}}, \tag{1}$$

where $\Delta\hat{\mathbf{Y}}$ is an input-dependent correction and α is a learned scalar. Figure 2 shows the three stages; Algorithm 1 gives the forward pass.

3.2 LOW-RANK STATE FACTORIZATION

Two learned projections mix the channels independently at each time step:

$$\mathbf{X}_{\text{common}} = W_{\text{up}}\big(W_{\text{down}}(\mathbf{X})\big), \qquad W_{\text{down}} : \mathbb{R}^N \to \mathbb{R}^r, \quad W_{\text{up}} : \mathbb{R}^r \to \mathbb{R}^N. \tag{2}$$

Both maps share parameters across time. We form the backbone input as

$$\mathbf{X}_{\text{specific}} = \mathbf{X} - \sigma(\gamma_r)\, \mathbf{X}_{\text{common}}, \tag{3}$$

where σ is the sigmoid and γ_r controls the amount subtracted. The term *specific* denotes this residual input; the learned decomposition does not enforce orthogonality or statistical independence.

For $r < N$, the linear channel map has rank at most r, limiting the dimension of the common representation. For $r \geq N$, including the default $r = 16$ on seven-channel ETT data, it need not reduce channel rank. The factorization therefore provides a low-rank constraint on channel-rich datasets, rather than a universal guarantee of noise removal (Appendix B).

3.3 STATISTICAL CONTEXT ENCODING

We reuse the down-projection to obtain the latent common trajectory,

$$\mathbf{Z} = W_{\text{down}}(\mathbf{X}_{\text{common}}) \in \mathbb{R}^{B\times L\times r}. \tag{4}$$

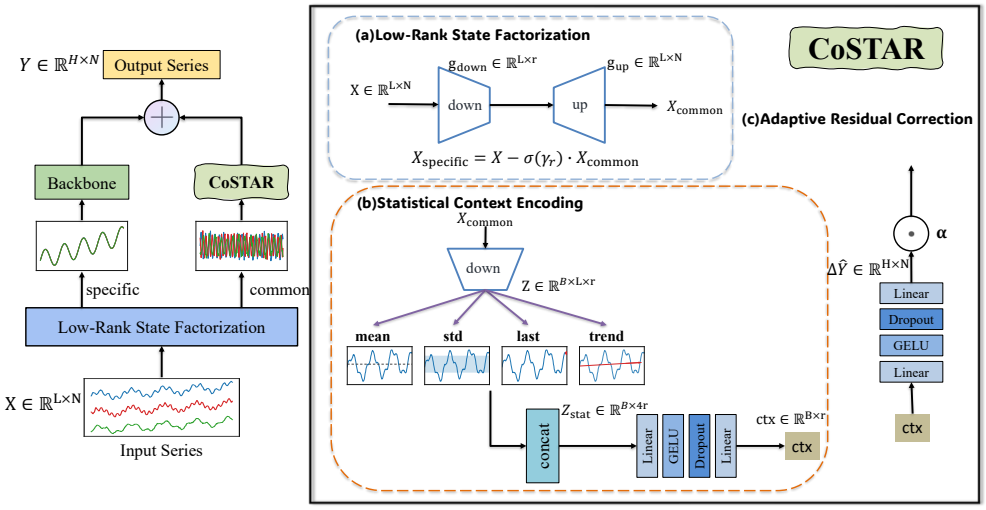

Figure 2: Overview of CoSTAR. *(a) Low-Rank State Factorization* decomposes the input $\mathbf{X}$ into a specific component $\mathbf{X}_{\text{specific}}$ sent to the backbone and a low-rank common component $\mathbf{X}_{\text{common}}$. *(b) Statistical Context Encoding* aggregates four temporal statistics of $\mathbf{X}_{\text{common}}$ into a context vector $\mathbf{ctx}$. *(c) Adaptive Residual Correction* maps $\mathbf{ctx}$ to a residual and adds the gated correction $\alpha\,\Delta\hat{\mathbf{Y}}$ to the backbone's preliminary prediction $\hat{\mathbf{Y}}_{\text{base}}$.

Four statistics summarize its level, variability, recent state, and net change:

$$\mathbf{Z}_{\text{stat}} = [\,\bar{\mathbf{z}}; \mathbf{s}_{\mathbf{z}}; \mathbf{z}_L; \mathbf{z}_L - \mathbf{z}_1\,] \in \mathbb{R}^{B\times 4r}, \tag{5}$$

where $\bar{\mathbf{z}} = L^{-1}\sum_t \mathbf{z}_t$ and $\mathbf{s}_{\mathbf{z}}$ is the temporal standard deviation. The endpoint difference measures net change over the window. A context MLP produces the *common state*,

$$\mathbf{ctx} = \text{MLP}_{\text{ctx}}(\mathbf{Z}_{\text{stat}}) \in \mathbb{R}^{B\times r}. \tag{6}$$

The common state has no explicit time axis, but its statistics retain temporal information. It is the interface between the observed history and the correction decoder (Figure 2). The representation avoids forwarding the full $L \times N$ tensor while retaining the level, variability, recent state, and net change of the shared trajectory. Table 4 tests the contribution of each statistic and pooling choice.

3.4 Adaptive Residual Correction

The third stage converts the common state into a correction for every horizon and channel (Figure 2). It is a residual correction, rather than an alternative forecast produced by a second backbone, and targets systematic errors left by the main forecaster:

$$\Delta\hat{\mathbf{Y}} = \text{reshape}\big(\text{MLP}_{\text{res}}(\mathbf{ctx}), H, N\big). \tag{7}$$

The final prediction follows Equation 1, with

$$\alpha = \sigma(\gamma_a)\,\sigma(\gamma_s). \tag{8}$$

Here γ_a and γ_s are learned global scalars; the correction is input-dependent through $\mathbf{ctx}$. With $\gamma_a = -3$ and $\sigma(\gamma_s) = 0.10$, the initial output scale is approximately 0.0047. This initialization starts with a small output correction while allowing the residual input to differ from the original input.

The three stages are intentionally coupled rather than independent add-on blocks. The first stage determines which part of the history is presented to the backbone. The second stage compresses the shared branch into a representation whose dimension does not grow with the lookback length. The third stage uses that representation to produce a correction in the same output space as the backbone. Consequently, the common state influences the prediction through two observable paths: it changes

the input seen by the backbone and it generates the residual added after the backbone. This design makes the source of the refinement explicit and allows each path to be removed or replaced in the ablation study.

The use of temporal statistics is a deliberate compromise between retaining the entire common trajectory and reducing the correction pathway to a single instantaneous value. The mean captures the overall level, the standard deviation captures temporal variability, the final state describes the most recent shared condition, and the endpoint difference provides a simple trend descriptor. These statistics are computed after the channel projection, so they summarize shared latent directions rather than independently aggregating the raw channels. The context MLP can combine them nonlinearly before the residual decoder maps the result to the future horizon.

The gate is initialized to make the added correction small, but it does not force the final model to remain close to the backbone throughout training. The add gate and strength gate are learned jointly with the forecasting objective. Thus, a useful correction can grow when the common state explains a systematic forecasting error, while an unhelpful correction can remain suppressed. This initialization is an optimization choice and should not be interpreted as a guarantee that the module will improve every dataset or every horizon.

3.5 Compatibility and Complexity

CoSTAR preserves input and output shapes and requires no access to hidden representations. It can wrap a forecaster that accepts an $L \times N$ input and returns an $H \times N$ prediction. The backbone and module can be trained jointly, or only CoSTAR can be trained while the backbone is frozen.

Per input window, the projection and decoding costs are $O(LNr + rHN + r^2)$. The implementation uses a residual MLP with hidden width $2r$, whose final weight matrix has $2rHN$ parameters. Thus, the added cost scales linearly with N and H for fixed r, but its size relative to the backbone depends on the setting. Appendix E reports paired latency and peak-memory measurements.

4 Experiments

4.1 Experimental Setup

Datasets and metrics. We evaluate eight long-term benchmarks: ETTh1, ETTh2, ETTm1, ETTm2, Electricity, Traffic, Weather, and Solar (Zhou et al., 2021; Wu et al., 2021). Four PEMS datasets (PEMS03/04/07/08) provide additional traffic forecasting tasks (Liu et al., 2024). We report MSE and MAE, for which lower values are better. Unless varied explicitly, the lookback is $L = 96$; horizons are $\{96, 192, 336, 720\}$ for long-term benchmarks and $\{12, 24, 48, 96\}$ for PEMS. Dataset details appear in Appendix A.1.

Baselines. Long-term comparisons include VPNet, FreqCycle, SimpleTM, CASA, WPMixer, ModernTCN, iTransformer, PatchTST, TimesNet, and DLinear (Liu et al., 2026; Zhang et al., 2026; Chen et al., 2025; Lee et al., 2025; Murad et al., 2025; Luo & Wang, 2024; Liu et al., 2024; Nie et al., 2023; Wu et al., 2023; Zeng et al., 2023). The paired backbone study additionally includes SparseTSF (Lin et al., 2024). Short-term comparisons include RAFT, xPatch, and GTR (Han et al., 2025; Stitsyuk & Choi, 2025; Cao et al., 2026). Appendix A.2 distinguishes standalone baselines from the backbones used with CoSTAR.

Protocol and reporting. For each paired backbone comparison, the backbone-only and CoSTAR-equipped models use identical preprocessing, chronological splits, lookback length, horizons, optimization protocol, and evaluation code. We retain the backbone-specific hyperparameter settings recommended by the original implementation. Appendix A.4 specifies the complete training, checkpoint-selection, and hardware protocol.

4.2 Forecasting Accuracy

4.2.1 Long-Term Forecasting

Table 1 summarizes horizon-averaged results on eight long-term benchmarks. In this table, CoSTAR is attached to VPNet and uses the same VPNet configuration. It attains the best value in six of the

16 dataset–metric entries, comprising one MSE and five MAE entries. Relative to the reported iTransformer results, its MSE is lower on Electricity (0.159 versus 0.176) and Weather (0.239 versus 0.261), reductions of 9.7% and 8.4%, respectively. Performance remains dataset dependent: ModernTCN leads on ETTh2, and VPNet leads on ETTm1. These results show competitive overall accuracy, especially in MAE, rather than uniform superiority over specialized forecasters. Table 10 gives the horizon-wise results.

Table 1: Long-term forecasting results. The lookback length is fixed at 96, and results are averaged across horizons $\{96, 192, 336, 720\}$. Best and second-best results are shown in **bold red** and blue underlined, respectively. The 1st Count row tallies first-place entries. See Table 10 for horizon-wise results.

Dataset	CoSTAR (Ours)		VPNet 2026		FreqCycle 2026		SimpleTM 2025		CASA 2025		WPMixer 2025		ModernTCN 2024		iTransformer 2024		PatchTST 2023		TimesNet 2023		DLinear 2023	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	0.431	**0.424**	0.435	0.428	0.428	0.427	**0.426**	0.432	0.442	0.440	0.447	0.434	0.435	0.428	0.455	0.448	0.451	0.442	0.461	0.455	0.458	0.451
ETTh2	0.370	0.393	0.375	0.394	0.371	0.398	0.355	0.392	0.381	0.406	0.385	0.408	**0.353**	**0.388**	0.383	0.407	0.366	0.395	0.400	0.417	0.498	0.479
ETTm1	0.381	**0.380**	0.376	0.382	**0.373**	0.391	0.382	0.396	0.389	0.400	0.400	0.405	0.388	0.402	0.408	0.413	0.383	0.396	0.412	0.418	0.405	0.409
ETTm2	0.271	**0.312**	0.280	0.317	**0.265**	**0.312**	0.276	0.323	0.278	0.323	0.277	0.322	0.281	0.324	0.291	0.335	0.285	0.328	0.294	0.331	0.310	0.367
Electricity	**0.159**	**0.248**	0.164	0.252	0.168	0.260	0.174	0.268	0.176	0.268	0.180	0.269	0.212	0.290	0.176	0.267	0.190	0.275	0.199	0.300	0.225	0.318
Traffic	0.451	0.292	0.462	0.295	0.452	**0.266**	0.448	0.283	0.428	0.277	0.525	0.339	0.625	0.376	**0.422**	0.282	0.470	0.294	0.626	0.333	0.625	0.384
Solar	0.197	0.224	0.203	**0.212**	0.254	0.288	**0.190**	0.252	0.226	0.263	0.263	0.286	0.486	0.516	0.235	0.262	0.261	0.296	0.276	0.291	0.326	0.397
Weather	0.239	**0.261**	**0.238**	0.263	0.242	0.270	0.240	0.270	0.246	0.274	0.245	0.272	0.239	0.272	0.261	0.281	0.258	0.280	0.258	0.285	0.266	0.317
1st Count	1	5	1	1	2	2	2	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0

4.2.2 Short-Term Forecasting

Table 2 compares methods on the four PEMS datasets. CoSTAR and GTR both use iTransformer. CoSTAR lowers the backbone's horizon-averaged MSE on all four datasets; on PEMS03, it decreases from 0.152 to 0.128. Relative to GTR, the average MSE reductions are 12.3%, 37.1%, 29.8%, and 38.0% on PEMS03, PEMS04, PEMS07, and PEMS08, respectively. CoSTAR leads in five of the eight averaged dataset–metric entries, while RAFT leads in three. The complete results appear in Table 11.

Table 2: Short-term forecasting results on PEMS. The lookback length is fixed at 96, and results are averaged across horizons $\{12, 24, 48, 96\}$. Best and second-best results are shown in **bold red** and blue underlined, respectively. The 1st Count row tallies first-place entries. See Table 11 for horizon-wise results.

Dataset	CoSTAR (Ours)		VPNet 2026		GTR (iTransformer) 2026		RAFT 2025		xPatch 2025		CASA 2025		SparseTSF 2024		iTransformer 2024		PatchTST 2023		DLinear 2023	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
PEMS03	**0.128**	**0.233**	0.149	0.260	0.146	0.257	0.144	0.235	0.188	0.279	0.174	0.274	0.398	0.424	0.152	0.258	0.176	0.294	0.245	0.342
PEMS04	**0.095**	**0.206**	0.148	0.259	0.151	0.266	0.102	0.209	0.203	0.290	0.205	0.301	0.417	0.444	0.120	0.232	0.198	0.317	0.246	0.350
PEMS07	0.099	**0.193**	0.140	0.250	0.141	0.239	**0.095**	0.200	0.185	0.271	0.174	0.267	0.396	0.433	0.176	0.267	0.216	0.331	0.277	0.371
PEMS08	0.163	0.236	0.138	0.247	0.263	0.333	**0.124**	**0.226**	0.181	0.267	0.241	0.307	0.362	0.413	0.196	0.264	0.188	0.314	0.238	0.345
1st Count	2	3	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0

4.2.3 Backbone Compatibility

Table 3 compares four representative backbones on three datasets; Table 12 extends the comparison to five backbones and six datasets. Across the 30 backbone–dataset pairs, CoSTAR lowers horizon-averaged MSE in 29, with one tie at the reported precision (CASA on Solar). It lowers average MAE in 28 pairs; PatchTST on Electricity and Solar are the two exceptions. On PEMS04, average MSE reductions range from 20.8% for iTransformer to 72.4% for SparseTSF. This paired evidence supports compatibility across architectures while distinguishing improvement in MSE from improvement in both metrics.

4.3 Component and Sensitivity Analyses

Component ablations. Table 4 compares the full DLinear variant with alternatives that remove input decomposition or output correction, or replace parts of common-state encoding with pooling. The full model achieves the lowest horizon-averaged MSE and MAE on all four reported datasets. The broader horizon-wise results in Table 13 show that individual alternatives can perform better in

Table 3: Paired backbone results on three benchmark datasets. The lookback length is fixed at 96. Original and + CoSTAR denote the backbone alone and the backbone equipped with CoSTAR, respectively; Improve is the relative error reduction.

Model		CASA						SparseTSF						iTransformer						DLinear					
Setup		Original		+ CoSTAR		Improve ↑		Original		+ CoSTAR		Improve ↑		Original		+ CoSTAR		Improve ↑		Original		+ CoSTAR		Improve ↑	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Electricity	96	0.139	0.234	**0.136**	**0.234**	2.2%	0.0%	0.196	0.269	**0.158**	**0.262**	19.4%	2.6%	0.148	0.240	**0.137**	**0.235**	7.4%	2.1%	0.210	0.301	**0.155**	**0.254**	26.2%	15.6%
	192	0.162	0.256	**0.156**	**0.250**	3.7%	2.3%	0.195	0.272	**0.165**	**0.268**	15.4%	1.5%	0.166	0.257	**0.158**	**0.252**	4.8%	1.9%	0.210	0.304	**0.170**	**0.268**	19.0%	11.8%
	336	0.178	0.271	**0.173**	**0.268**	2.8%	1.1%	0.208	0.286	**0.176**	**0.280**	15.4%	2.1%	0.179	0.272	**0.171**	**0.267**	4.5%	1.8%	0.223	0.319	**0.180**	**0.282**	19.3%	11.6%
	720	0.220	0.308	**0.208**	**0.302**	5.5%	1.9%	0.251	0.321	**0.209**	**0.307**	16.7%	4.4%	0.210	0.299	**0.197**	**0.292**	6.2%	2.3%	0.257	0.349	**0.224**	**0.321**	12.8%	8.0%
	Avg	0.175	0.267	**0.168**	**0.264**	4.0%	1.1%	0.212	0.287	**0.177**	**0.279**	16.5%	2.8%	0.176	0.267	**0.166**	**0.262**	5.7%	1.9%	0.225	0.318	**0.182**	**0.281**	19.1%	11.6%
PEMS04	12	0.091	0.199	**0.075**	**0.181**	17.6%	9.0%	0.154	0.273	**0.086**	**0.203**	44.2%	25.6%	0.082	0.191	**0.074**	**0.179**	9.8%	6.3%	0.111	0.227	**0.070**	**0.175**	36.9%	22.9%
	24	0.124	0.237	**0.095**	**0.208**	23.4%	12.2%	0.270	0.376	**0.097**	**0.217**	64.1%	42.3%	0.101	0.213	**0.084**	**0.193**	16.8%	9.4%	0.183	0.301	**0.086**	**0.199**	53.0%	33.9%
	48	0.221	0.325	**0.136**	**0.255**	38.5%	21.5%	0.468	0.489	**0.145**	**0.277**	69.0%	43.4%	0.133	0.247	**0.100**	**0.213**	24.8%	13.8%	0.303	0.400	**0.116**	**0.238**	61.7%	40.5%
	96	0.385	0.443	**0.220**	**0.329**	42.9%	25.7%	0.774	0.636	**0.133**	**0.260**	82.8%	59.1%	0.165	0.276	**0.123**	**0.238**	25.5%	13.8%	0.385	0.471	**0.150**	**0.276**	61.0%	41.4%
	Avg	0.205	0.301	**0.132**	**0.243**	35.6%	19.3%	0.417	0.444	**0.115**	**0.239**	72.4%	46.2%	0.120	0.232	**0.095**	**0.206**	20.8%	11.2%	0.245	0.350	**0.105**	**0.222**	57.1%	36.6%
PEMS07	12	0.069	0.173	**0.060**	**0.156**	13.0%	9.8%	0.139	0.261	**0.076**	**0.190**	45.3%	27.2%	0.066	0.164	**0.058**	**0.152**	12.1%	7.3%	0.100	0.218	**0.058**	**0.161**	42.0%	26.1%
	24	0.102	0.211	**0.082**	**0.183**	19.6%	13.3%	0.259	0.369	**0.120**	**0.251**	53.7%	32.0%	0.087	0.190	**0.071**	**0.170**	18.4%	10.5%	0.186	0.303	**0.076**	**0.188**	59.1%	38.0%
	48	0.183	0.286	**0.123**	**0.226**	32.8%	21.0%	0.471	0.489	**0.124**	**0.251**	73.7%	48.7%	0.229	0.328	**0.088**	**0.189**	61.6%	42.4%	0.354	0.443	**0.107**	**0.230**	69.8%	48.1%
	96	0.343	0.399	**0.188**	**0.288**	45.2%	27.8%	0.713	0.611	**0.267**	**0.387**	62.6%	36.7%	0.322	0.384	**0.179**	**0.261**	44.4%	32.0%	0.466	0.521	**0.146**	**0.272**	68.7%	47.8%
	Avg	0.174	0.267	**0.113**	**0.213**	35.1%	20.2%	0.395	0.432	**0.147**	**0.270**	62.8%	37.5%	0.176	0.267	**0.099**	**0.193**	43.7%	27.7%	0.277	0.371	**0.097**	**0.213**	65.0%	42.6%

some settings. Thus, the aggregate evidence supports the combined design without implying that every component is necessary at every horizon.

Table 4: Ablation study of CoSTAR on DLinear. The lookback length is fixed at 96, and results are averaged across forecasting horizons. Best results are in **bold**.

Variant	Electricity		PEMS03		PEMS07		Solar	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
DLinear + CoSTAR	**0.182**	**0.281**	**0.117**	**0.235**	**0.097**	**0.213**	**0.273**	**0.345**
w/o Input Decomposition	0.190	0.284	0.126	0.243	0.105	0.222	0.277	0.348
w/o Output Correction	0.201	0.296	0.157	0.276	0.132	0.253	0.299	0.380
Common → Mean Pooling	0.206	0.292	0.164	0.277	0.160	0.276	0.283	0.356
Statistics → Mean Pooling	0.198	0.294	0.142	0.262	0.118	0.238	0.282	0.360
Statistics → Attention Pooling	0.198	0.295	0.155	0.276	0.135	0.256	0.294	0.376

Lookback length. Figure 3 varies $L \in \{48, 96, 192, 336\}$ at $H = 96$. Across the four datasets, the average relative MSE reduction is 40.8% for DLinear and 20.7% for iTransformer at $L = 48$, compared with 16.9% and 3.2% at $L = 336$. On PEMS03/04, DLinear with CoSTAR at $L = 48$ achieves MSE 0.215/0.172, below the unmodified DLinear at $L = 96$ (0.386/0.385). Within this sweep, the larger gains at shorter lookbacks suggest that the refinement pathway can be especially useful when the backbone receives less historical context.

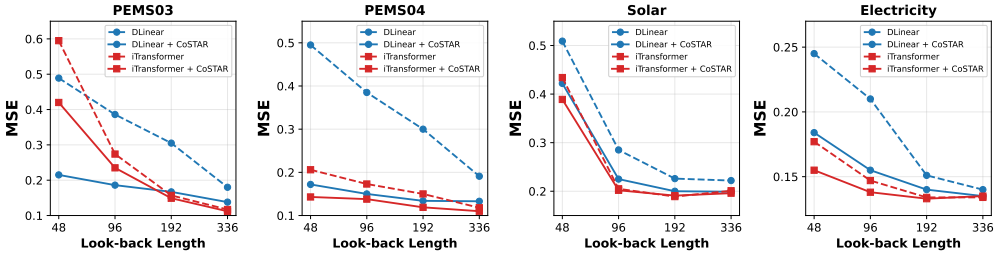

Figure 3: Influence of lookback length. The forecasting horizon is fixed at $H = 96$ and the lookback length varies over $\{48, 96, 192, 336\}$. Dashed lines denote the original backbones, and solid lines denote their CoSTAR-enhanced counterparts. CoSTAR consistently improves both backbones across all window lengths, with the largest relative gains at the shortest window.

4.4 Efficiency and Common-State Analysis

Cross-split correlation. Figure 4 compares training- and test-split cross-channel correlations on PEMS04 using PatchTST. The mean absolute correlation change is 0.025 for the raw input and 0.0115 for the common component. Electricity shows the same pattern (0.1204 versus 0.0128;

Appendix C.6). The common representation therefore exhibits smaller correlation changes across these splits. This diagnostic characterizes second-order structure; it does not establish distributional invariance.

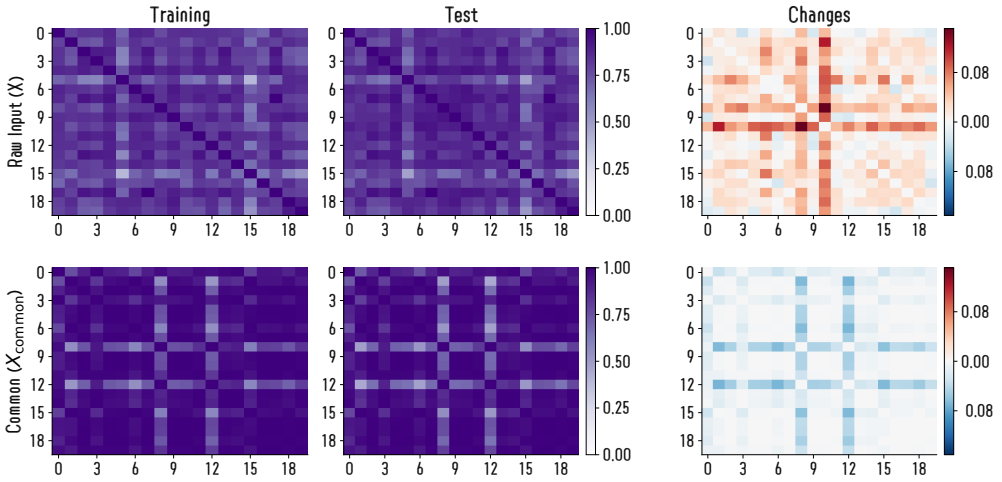

Figure 4: Cross-split stability of the learned common state on PEMS04 (PatchTST backbone, look-back 96, prediction length 96). *Top:* cross-channel Pearson correlation of the raw input on the training split (left), the test split (middle), and their difference (right). *Bottom:* the same three panels for the learned common component. The common component has a smaller measured train–test correlation change (0.0115) than the raw input (0.025). This descriptive diagnostic does not establish distributional invariance or identify a unique latent factor.

Hyperparameters. Table 5 varies the projection width from 4 to 64. Average MSE decreases from 0.121 to 0.097 on PEMS04 and from 0.192 to 0.176 on Electricity. The larger relative change on PEMS04 indicates a meaningful capacity–accuracy tradeoff. Appendix D reports the complete sweeps for projection width, correction strength, and gate initialization.

Plug-in efficiency. Figure 5 gives an implementation-specific comparison of CoSTAR, CCM, and LIFT on PEMS07 with $L = H = 96$ (Chen et al., 2024; Zhao & Shen, 2024). In this configuration, CoSTAR reports MSE 0.133 with 0.645 ms/iter latency and 100 MB peak memory. The corresponding MSE/latency/memory measurements are 1.247/157/2818 for CCM and 0.245/2.1/261 for LIFT. Because cross-implementation measurements depend on the implementation and evaluation environment, the figure is diagnostic. Appendix E reports the complete paired base-versus-CoSTAR efficiency analysis under one shared protocol.

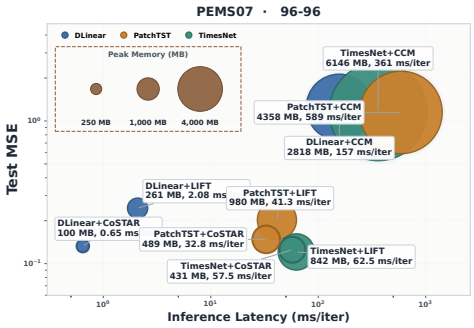

Figure 5: Plug-in efficiency on PEMS07 (96-96). The horizontal axis is inference latency (ms/iter), the vertical axis is test MSE, and bubble size represents peak memory (MB). The figure summarizes measurements for DLinear, PatchTST, and TimesNet under the plotted evaluation configuration.

4.5 Plug-in Comparison and Frozen-Backbone Adaptation

Table 6 compares CoSTAR with GTR and HN-MVTS on DLinear and PEMS04 (Cao et al., 2026; Savchenko & Kachan, 2026). CoSTAR has the lowest MSE and MAE at every reported horizon. At $H = 96$, its MSE of 0.150 is 21.5% below GTR and 85.0% below HN-MVTS. Tables 15 and 16 extend the comparison to PatchTST and PEMS07.

Table 5: Sensitivity to the common-state rank r at lookback 96.

Dataset	Metric	Rank r 4	8	16	32	64
PEMS04	MSE	0.121	0.119	0.108	0.107	**0.097**
	MAE	0.235	0.239	0.225	0.221	**0.210**
Electricity	MSE	0.192	0.192	0.182	0.178	**0.176**
	MAE	0.289	0.291	0.282	0.280	**0.279**

Table 6: Comparison with GTR and HN-MVTS on DLinear and PEMS04 (lookback 96).

Horizon	CoSTAR (Ours) MSE	MAE	GTR MSE	MAE	HN-MVTS MSE	MAE
12	**0.070**	**0.175**	0.083	0.191	0.120	0.232
24	**0.086**	**0.199**	0.111	0.222	0.224	0.323
48	**0.116**	**0.238**	0.154	0.262	0.506	0.510
96	**0.150**	**0.276**	0.191	0.294	1.002	0.754

Table 7: Frozen-backbone adaptation at lookback 96.

Backbone	Dataset	Backbone	Joint	Frozen
DLinear	Electricity	0.225	**0.182**	0.185
DLinear	PEMS03	0.245	**0.117**	0.134
iTransformer	PEMS04	0.120	**0.095**	0.102
iTransformer	PEMS03	0.152	0.128	**0.126**

Frozen-backbone adaptation. We freeze a pretrained backbone and train only CoSTAR. Table 7 shows improved horizon-averaged MSE for two representative settings. Appendix 17 provides the full horizon-wise MSE/MAE comparison.

The five-seed stability results provide a complementary view of reliability. A small standard deviation indicates that the reported configuration is not dominated by one favorable initialization, but it is not by itself a statistical significance test against every baseline. We therefore use the stability table together with the paired comparisons and ablations. The combination shows whether the gain is repeated across seeds, whether it persists after removing a stage, and whether it transfers to another backbone or dataset.

5 Limitations

The benefit of CoSTAR varies with the backbone, dataset, horizon, and metric. Its factorized channel map and window-level statistics limit the information available to the correction pathway, and its decoder grows with the horizon and channel count. The evaluation covers deterministic forecasting with joint and frozen-backbone training, including lookback sensitivity. It does not establish zero-shot transfer of a trained module across backbones or datasets, nor performance under missing data, streaming updates, or probabilistic forecasting.

6 Conclusion

CoSTAR couples input residualization, statistical common-state encoding, and gated forecast correction through an external backbone interface. Across the reported paired comparisons, it lowers horizon-averaged MSE in 29 of 30 backbone–dataset pairs, with one tie at the reported precision. Ablations and frozen-backbone experiments indicate that the module can improve a fixed forecaster after task-specific training. The evidence supports CoSTAR as a refinement module for multivariate settings with exploitable cross-channel regularity, while its gains remain dependent on the backbone, dataset, horizon, and metric.

Reproducibility Statement

The main text formalizes the CoSTAR architecture with equations and gives its forward pass in Algorithm 1. Appendix A provides the dataset descriptions, metrics, model configuration, training protocol, and checkpoint-selection rule. The remaining appendices report the complete horizon-wise results, ablations, five-seed summaries, frozen-backbone protocol, hyperparameter sweeps, and paired efficiency measurements.

AI USE STATEMENT

Generative AI tools assisted with language editing, manuscript organization, limited literature discovery and reference formatting, and experimental instrumentation code for recording efficiency measurements. The authors determined the research question, method, experimental design, data processing, and scientific interpretation. All AI-assisted text and code were reviewed and verified by the authors, who take full responsibility for the final manuscript, results, references, and claims.

REFERENCES

Fanpu Cao, Lu Dai, Jindong Han, and Hui Xiong. Enhancing multivariate time series forecasting with global temporal retrieval. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7a3284ce53f620d244fca0c79d94c478-Abstract-Conference.html.

Hui Chen, Viet Luong, Lopamudra Mukherjee, and Vikas Singh. SimpleTM: A simple baseline for multivariate time series forecasting. In *International Conference on Learning Representations*, 2025.

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*, 2024.

Hanyin Cheng, Xingjian Wu, Yang Shu, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. CoRA: Boosting time series foundation models for multivariate forecasting through correlation-aware adapter. In *International Conference on Learning Representations*, 2026.

Lu Han, Xu-Yang Chen, Han-Jia Ye, and De-Chuan Zhan. SOFTS: Efficient multivariate time series forecasting with series-core fusion. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/754612bde73a8b65ad8743f1f6d8ddf6-Abstract-Conference.html.

Sungwon Han, Seungeon Lee, Meeyoung Cha, Sercan O. Arik, and Jinsung Yoon. Retrieval augmented time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

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

Minhyuk Lee, HyeKyung Yoon, and MyungJoo Kang. CASA: CNN autoencoder-based score attention for efficient multivariate long-term time-series forecasting. In *International Joint Conference on Artificial Intelligence*, 2025.

Dongyuan Li, Shun Zheng, Chang Xu, Jiang Bian, and Renhe Jiang. Routing channel-patch dependencies in time series forecasting with graph spectral decomposition. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2603.13702.

Shengsheng Lin, Weiwei Lin, Wentai Wu, Haojun Chen, and Junjie Yang. SparseTSF: Modeling long-term time series forecasting with 1k parameters. In *International Conference on Machine Learning*, 2024.

Shengsheng Lin, Haojun Chen, Haijie Wu, Chunyun Qiu, and Weiwei Lin. Temporal query network for efficient multivariate time series forecasting. In *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.

Kun Liu, Renjun Jia, Ruifeng Yang, Xirui Zeng, Yuqi Liang, and Cen Chen. Are global dependencies necessary? scalable time series forecasting via local cross-variate modeling. In *International Conference on Learning Representations*, 2026.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024.

Donghao Luo and Xue Wang. ModernTCN: A modern pure convolution structure for general time series analysis. In *International Conference on Learning Representations*, 2024.

Md Mahmuddun Nabi Murad, Mehmet Aktukmak, and Yasin Yilmaz. WPMixer: Efficient multi-resolution mixing for long-term time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

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

Andrey Savchenko and Oleg Kachan. HN-MVTS: HyperNetwork-based multivariate time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 25200–25208, 2026. URL https://ojs.aaai.org/index.php/AAAI/article/view/39711.

Artyom Stitsyuk and Jaesik Choi. xPatch: Dual-stream time series forecasting with exponential seasonal-trend decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2025.

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

Shiyu Wang, Jiawei Li, Xiaoming Shi, Zhou Ye, Baichuan Mo, Wenze Lin, Shengtong Ju, Zhixuan Chu, and Ming Jin. Timemixer++: A general time series pattern machine for universal predictive analysis. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.16032.

Yuyang Wang, Alex Smola, Danielle Maddix, Jan Gasthaus, Dean Foster, and Tim Januschowski. Deep factors for forecasting. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6607–6617, 2019. URL https://proceedings.mlr.press/v97/wang19k.html.

Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: A survey. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 6778–6786, 2023.

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

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

Hsiang-Fu Yu, Nikhil Rao, and Inderjit S. Dhillon. Temporal regularized matrix factorization for high-dimensional time series prediction. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/hash/85422afb467e9456013a2a51d4dff702-Abstract.html.

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.

Boya Zhang, Shuaijie Yin, Huiwen Zhu, and Xing He. FreqCycle: A multi-scale time-frequency analysis method for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 28159–28167, 2026.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *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*, 2024.

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.

A Implementation Details

A.1 Datasets

We evaluate CoSTAR on twelve public multivariate time-series benchmarks spanning energy, meteorology, and traffic. The Electricity Transformer Temperature (ETT) dataset contains seven channels and is split into ETTh1/ETTh2 (hourly) and ETTm1/ETTm2 (15-minute) (Zhou et al., 2021). Weather contains 21 meteorological channels (Wu et al., 2021). Electricity records hourly electricity consumption of 321 clients (Wu et al., 2021). Traffic records hourly road occupancy from 862 sensors (Wu et al., 2021). Solar records 10-minute solar power from 137 plants (Wu et al., 2021). PEMS03/04/07/08 are 5-minute traffic-flow datasets (Liu et al., 2024). We follow the standard chronological train/validation/test split without leakage. For the eight long-term benchmarks the horizons are $\{96, 192, 336, 720\}$, and for PEMS03/04/07/08 the horizons are $\{12, 24, 48, 96\}$; the lookback length is fixed at 96. Table 8 summarizes the dataset statistics.

Table 8: Dataset statistics. Sizes are reported as (train, validation, test).

Task	Dataset	Dim	Horizons	Size (train, val, test)	Freq.	Domain
Long-term Forecasting	ETTm1	7	$\{96, 192, 336, 720\}$	$(34465, 11521, 11521)$	15min	Temperature
	ETTm2	7	$\{96, 192, 336, 720\}$	$(34465, 11521, 11521)$	15min	Temperature
	ETTh1	7	$\{96, 192, 336, 720\}$	$(8545, 2881, 2881)$	Hourly	Temperature
	ETTh2	7	$\{96, 192, 336, 720\}$	$(8545, 2881, 2881)$	Hourly	Temperature
	Electricity	321	$\{96, 192, 336, 720\}$	$(18317, 2633, 5261)$	Hourly	Electricity
	Traffic	862	$\{96, 192, 336, 720\}$	$(12185, 1757, 3509)$	Hourly	Transportation
	Weather	21	$\{96, 192, 336, 720\}$	$(36792, 5271, 10540)$	10min	Weather
	Solar	137	$\{96, 192, 336, 720\}$	$(36601, 5161, 10417)$	10min	Electricity
Short-term Forecasting	PEMS03	358	$\{12, 24, 48, 96\}$	$(15617, 5135, 5135)$	5min	Transportation
	PEMS04	307	$\{12, 24, 48, 96\}$	$(10172, 3375, 3375)$	5min	Transportation
	PEMS07	883	$\{12, 24, 48, 96\}$	$(16911, 5622, 5622)$	5min	Transportation
	PEMS08	170	$\{12, 24, 48, 96\}$	$(10690, 3548, 3548)$	5min	Transportation

A.2 Backbone Models

The standalone long-term comparison uses ten forecasters: VPNet, FreqCycle, SimpleTM, CASA, WPMixer, ModernTCN, iTransformer, PatchTST, TimesNet, and DLinear. The paired backbone study attaches CoSTAR to CASA, SparseTSF, iTransformer, PatchTST, and DLinear, covering Transformer, convolutional, and linear/MLP designs. The short-term comparison additionally includes RAFT, xPatch, and iTransformer with GTR. We use official or widely used open-source implementations and the model-specific configurations described by their papers.

A.3 Metrics

We report mean squared error (MSE) and mean absolute error (MAE). For predictions $\hat{\mathbf{Y}} \in \mathbb{R}^{H \times N}$ and ground truth $\mathbf{Y} \in \mathbb{R}^{H \times N}$, they are defined as

$$\mathrm{MSE} = \frac{1}{HN} \sum_{h,n} \big(\hat{Y}_{h,n} - Y_{h,n}\big)^2, \qquad \mathrm{MAE} = \frac{1}{HN} \sum_{h,n} \big|\hat{Y}_{h,n} - Y_{h,n}\big|. \tag{9}$$

A.4 Training and CoSTAR Configuration

Unless a sensitivity analysis states otherwise, all runs use Adam, an initial learning rate of 1×10^{-4}, and the ℓ_2 forecasting loss. Experiments use Python 3.10.16 and PyTorch 2.6.0+cu124 on an Ubuntu server with seven Tesla P100-PCIE GPUs (16 GB) and two NVIDIA GeForce RTX 4090 GPUs (24 GB). For each baseline and plug-in comparison, we use the hyperparameter settings recommended by the original authors, including the original configuration of the backbone receiving

CoSTAR. Within every paired comparison, the baseline and CoSTAR-equipped variants share preprocessing, chronological train/validation/test splits, lookback length, prediction horizons, optimizer family, initial learning rate, loss, and evaluation code. The backbone and CoSTAR are trained jointly unless stated otherwise. Checkpoint selection uses the lowest validation loss under the relevant training protocol, after which the selected checkpoint is evaluated once on the test set. The epoch-wise test curves in Appendix F are diagnostic only and are never used for checkpoint selection or hyperparameter tuning. This protocol follows the evaluation practice of recent multivariate forecasting studies such as TimeMixer++ (Wang et al., 2025) and xCPD (Li et al., 2026). CoSTAR uses rank $r = 16$ by default, with $\gamma_a = -3.0$, $\gamma_r = -2.0$, and strength 0.10 (i.e., $\sigma(\gamma_s) = 0.10$). Table 9 lists the default configuration, and Algorithm 1 gives the forward pass.

Table 9: Default CoSTAR configuration.

Hyperparameter	Value
Rank r	16
Strength	0.10
γ_a init	-3.0
γ_r init	-2.0
Dropout	0.0
Context projection	Linear$(4r, r)$→GELU→Dropout→Linear(r, r)
Residual MLP	Linear$(r, 2r)$→GELU→Dropout→Linear$(2r, HN)$
Optimizer	Adam
Learning rate	1×10^{-4}
Loss	MSE

Algorithm 1 CoSTAR forward pass attached to a backbone f_θ.

Require: input $\mathbf{X} \in \mathbb{R}^{B \times L \times N}$, backbone f_θ, rank r
Ensure: refined prediction $\hat{\mathbf{Y}} \in \mathbb{R}^{B \times H \times N}$
// 1. low-rank state factorization
$\mathbf{X}_{\text{common}} \leftarrow W_{\text{up}}(W_{\text{down}}(\mathbf{X}))$
$\mathbf{X}_{\text{specific}} \leftarrow \mathbf{X} - \sigma(\gamma_r)\, \mathbf{X}_{\text{common}}$
// 2. backbone preliminary prediction
$\hat{\mathbf{Y}}_{\text{base}} \leftarrow f_\theta(\mathbf{X}_{\text{specific}})$
// 3. statistical context encoding
$\mathbf{Z} \leftarrow W_{\text{down}}(\mathbf{X}_{\text{common}})$
$\mathbf{ctx} \leftarrow \text{MLP}_{\text{ctx}}([\bar{\mathbf{z}}; \mathbf{s}_{\mathbf{z}}; \mathbf{z}_L; \mathbf{z}_L - \mathbf{z}_1])$
// 4. adaptive residual correction
$\Delta\hat{\mathbf{Y}} \leftarrow \text{reshape}(\text{MLP}_{\text{res}}(\mathbf{ctx}), H, N)$
$\alpha \leftarrow \sigma(\gamma_a)\, \sigma(\gamma_s)$
$\hat{\mathbf{Y}} \leftarrow \hat{\mathbf{Y}}_{\text{base}} + \alpha\, \Delta\hat{\mathbf{Y}}$
return $\hat{\mathbf{Y}}$

B Representation Constraints and Interpretation

Factorized representations are established tools for modeling shared structure in multivariate series (Yu et al., 2016; Wang et al., 2019). CoSTAR uses the same principle for the channel map in its external refinement module. Using a column-vector convention, the two projections define $W_{\text{down}} \in \mathbb{R}^{r \times N}$ and $W_{\text{up}} \in \mathbb{R}^{N \times r}$, and their composition is the channel map $M = W_{\text{up}} W_{\text{down}} \in \mathbb{R}^{N \times N}$. It satisfies

$$\text{rank}(M) \leq \min(N, r). \tag{10}$$

Thus, when $r < N$, the common branch can transmit at most r linearly independent channel directions. This is a capacity constraint and an inductive bias toward shared structure; it does not establish that the retained directions are the true shared factors. When $r \geq N$, the construction does not impose rank compression. Unlike LoRA's low-rank update of model weights (Hu et al., 2022), CoSTAR constrains the channel representation used for input conditioning and output correction.

The forecasting objective does not enforce factor identifiability, orthogonality, statistical independence, or noise removal. Accordingly, X_{common} and X_{specific} should be understood as learned functional representations, rather than identifiable latent factors or a theoretically proven denoising decomposition. The low-rank channel map also does not imply that a decoded $H \times N$ correction matrix is low rank, because the decoder can apply nonlinear transformations after the common state is formed.

Our empirical analyses are intended to test utility rather than prove these stronger properties. Paired forecasting results and ablations assess whether the representation improves accuracy; the cross-split correlation analysis is a descriptive diagnostic of the measured correlation structure. The frozen-backbone experiment tests whether CoSTAR can be adapted while the forecaster parameters remain fixed for the target task. None of these experiments proves distributional invariance, zero-shot transfer, factor recovery, or a generalization theorem. Together, they provide empirical evidence for the practical value of the representation constraint within the evaluated forecasting settings.

C Full Results

C.1 Long-Term Forecasting

Table 10 reports horizon-wise MSE and MAE on the eight long-term benchmarks. It complements the averaged comparison in Table 1 by showing the variation across horizons and metrics, including settings in which competing forecasters achieve lower error.

Table 10: Horizon-wise long-term forecasting results. The lookback length is fixed at 96. Best and second-best results are shown in **bold red** and blue underlined, respectively. Avg reports the average over the four forecasting horizons, and 1st Count tallies first-place entries.

Model		CoSTAR (Ours)		VPNet 2026		FreqCycle 2026		SimpleTM 2025		CASA 2025		WPMixer 2025		ModernTCN 2024		iTransformer 2024		PatchTST 2023		TimesNet 2023		DLinear 2023	
Year																							
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	0.373	0.389	0.375	**0.387**	0.371	0.391	**0.366**	0.393	0.382	0.404	0.386	0.399	0.384	0.396	0.386	0.405	0.394	0.408	0.389	0.412	0.382	0.395
	192	0.426	**0.417**	0.428	0.418	**0.420**	0.419	0.421	0.421	0.427	0.428	0.428	0.424	0.439	0.425	0.442	0.436	0.445	0.434	0.442	0.443	0.432	0.425
	336	0.466	0.437	0.470	0.444	0.457	0.438	**0.438**	0.439	0.474	0.451	0.483	0.448	0.442	**0.426**	0.486	0.458	0.484	0.453	0.495	0.471	0.491	0.467
	720	**0.459**	**0.454**	0.467	0.463	0.465	0.459	0.479	0.474	0.484	0.475	0.490	0.467	0.474	0.464	0.506	0.491	0.480	0.471	0.519	0.494	0.528	0.518
	Avg	0.431	**0.424**	0.435	0.428	0.428	0.427	**0.426**	0.432	0.442	0.440	0.447	0.434	0.435	0.428	0.455	0.448	0.451	0.442	0.461	0.455	0.458	0.451
ETTh2	96	0.285	0.332	0.286	**0.331**	0.282	0.335	0.282	0.338	0.296	0.347	0.298	0.347	**0.281**	0.336	0.299	0.350	0.295	0.344	0.335	0.370	0.329	0.380
	192	0.364	0.384	0.372	**0.381**	0.362	0.388	**0.357**	0.388	0.375	0.397	0.373	0.396	0.361	0.387	0.381	0.399	0.377	0.393	0.403	0.413	0.431	0.443
	336	0.412	0.420	0.444	0.438	0.415	0.427	0.366	0.401	0.422	0.433	0.424	0.434	**0.349**	**0.388**	0.422	0.432	0.382	0.410	0.435	0.440	0.458	0.461
	720	0.418	0.436	**0.399**	**0.425**	0.425	0.442	0.417	0.439	0.430	0.446	0.446	0.456	0.421	0.440	0.432	0.448	0.411	0.433	0.428	0.444	0.773	0.631
	Avg	0.370	0.393	0.375	0.394	0.371	0.398	0.355	0.392	0.381	0.406	0.385	0.408	**0.353**	**0.388**	0.384	0.407	0.366	0.395	0.400	0.417	0.498	0.479
ETTm1	96	**0.304**	**0.336**	0.313	0.342	0.312	0.355	0.321	0.360	0.321	0.360	0.335	0.366	0.322	0.364	0.340	0.375	0.323	0.360	0.333	0.375	0.345	0.373
	192	0.359	**0.364**	0.361	0.372	**0.350**	0.374	0.362	0.381	0.372	0.388	0.387	0.398	0.368	0.389	0.382	0.396	0.364	0.384	0.408	0.414	0.382	0.392
	336	0.393	**0.388**	**0.385**	0.389	**0.385**	0.400	0.390	0.403	0.403	0.407	0.396	0.406	0.398	0.410	0.419	0.419	0.391	0.402	0.415	0.422	0.416	0.417
	720	0.469	0.432	**0.443**	**0.424**	0.445	0.434	0.454	0.438	0.461	0.445	0.483	0.451	0.462	0.444	0.490	0.462	0.454	0.439	0.492	0.462	0.479	0.456
	Avg	0.381	**0.380**	0.376	0.382	**0.373**	0.391	0.382	0.396	0.389	0.400	0.400	0.405	0.388	0.402	0.408	0.413	0.383	0.396	0.412	0.418	0.405	0.409
ETTm2	96	0.167	0.245	0.176	0.248	**0.162**	**0.244**	0.175	0.258	0.176	0.258	0.174	0.257	0.174	0.256	0.185	0.271	0.178	0.260	0.185	0.262	0.186	0.281
	192	0.233	0.290	0.236	0.291	**0.226**	**0.287**	0.238	0.298	0.243	0.303	0.239	0.298	0.238	0.298	0.252	0.312	0.249	0.307	0.251	0.306	0.272	0.348
	336	0.293	0.328	0.304	0.334	**0.282**	**0.325**	0.298	0.339	0.296	0.336	0.302	0.340	0.312	0.346	0.316	0.352	0.309	0.345	0.319	0.348	0.343	0.394
	720	**0.390**	**0.387**	0.404	0.395	0.391	0.392	0.394	0.396	0.399	0.396	0.393	0.393	0.399	0.394	0.413	0.406	0.404	0.400	0.420	0.407	0.438	0.443
	Avg	0.271	**0.312**	0.280	0.317	**0.265**	**0.312**	0.276	0.323	0.278	0.323	0.277	0.322	0.281	0.324	0.291	0.335	0.285	0.328	0.294	0.331	0.310	0.366
Electricity	96	**0.130**	**0.219**	0.135	0.225	0.138	0.232	0.150	0.245	0.140	0.236	0.150	0.241	0.189	0.270	0.148	0.240	0.166	0.252	0.167	0.271	0.210	0.301
	192	**0.149**	**0.237**	0.151	0.239	0.154	0.247	0.151	0.248	0.164	0.256	0.162	0.252	0.196	0.276	0.166	0.257	0.174	0.260	0.187	0.289	0.210	0.304
	336	**0.166**	**0.257**	0.168	**0.257**	0.171	0.264	0.189	0.284	0.179	0.273	0.180	0.270	0.210	0.290	0.179	0.272	0.190	0.277	0.203	0.304	0.223	0.319
	720	**0.192**	**0.278**	0.202	0.288	0.210	0.297	0.206	0.296	0.219	0.307	0.230	0.313	0.253	0.324	0.210	0.299	0.229	0.310	0.238	0.334	0.257	0.349
	Avg	**0.159**	**0.248**	0.164	0.252	0.168	0.260	0.174	0.268	0.175	0.268	0.180	0.269	0.212	0.290	0.176	0.267	0.190	0.275	0.199	0.299	0.225	0.318
Traffic	96	0.399	**0.236**	0.419	0.277	0.444	0.263	0.420	0.270	0.396	0.262	0.521	0.344	0.652	0.390	**0.394**	0.269	0.450	0.286	0.589	0.315	0.649	0.396
	192	0.445	0.299	0.437	0.282	0.437	**0.262**	0.435	0.277	0.416	0.272	0.507	0.332	0.599	0.363	**0.413**	0.277	0.456	0.287	0.614	0.323	0.599	0.371
	336	0.462	0.307	0.480	0.309	0.448	**0.265**	0.450	0.284	0.433	0.278	0.530	0.336	0.606	0.366	**0.423**	0.283	0.471	0.293	0.638	0.341	0.606	0.374
	720	0.496	0.326	0.513	0.314	0.478	**0.274**	0.488	0.303	0.467	0.297	0.541	0.344	0.644	0.385	**0.459**	0.301	0.503	0.312	0.664	0.353	0.646	0.395
	Avg	0.451	0.292	0.462	0.295	0.452	**0.266**	0.448	0.283	0.428	0.277	0.525	0.339	0.625	0.376	**0.422**	0.282	0.470	0.294	0.626	0.333	0.625	0.384
Solar	96	**0.175**	0.201	0.179	**0.193**	0.211	0.260	0.180	0.244	0.194	0.234	0.230	0.269	0.452	0.497	0.203	0.235	0.230	0.277	0.241	0.268	0.285	0.372
	192	0.189	0.220	0.193	**0.210**	0.244	0.282	**0.182**	0.248	0.227	0.269	0.273	0.293	0.478	0.510	0.238	0.262	0.260	0.295	0.268	0.294	0.315	0.393
	336	0.216	0.229	0.217	**0.222**	0.278	0.303	**0.198**	0.262	0.240	0.273	0.276	0.292	0.513	0.532	0.250	0.275	0.277	0.306	0.288	0.297	0.349	0.413
	720	0.209	0.245	0.225	**0.223**	0.284	0.305	**0.200**	0.254	0.241	0.275	0.272	0.291	0.499	0.523	0.250	0.276	0.276	0.305	0.307	0.306	0.354	0.410
	Avg	0.197	0.224	0.204	**0.212**	0.254	0.288	**0.190**	0.252	0.226	0.263	0.263	0.286	0.485	0.516	0.235	0.262	0.261	0.296	0.276	0.291	0.326	0.397
Weather	96	**0.152**	**0.190**	0.155	0.193	0.158	0.202	0.156	0.201	0.161	0.206	0.163	0.208	0.157	0.205	0.176	0.215	0.178	0.219	0.168	0.219	0.196	0.255
	192	0.203	**0.239**	**0.202**	0.240	0.205	0.246	0.204	0.247	0.208	0.251	0.209	0.249	0.205	0.250	0.226	0.259	0.225	0.259	0.224	0.264	0.239	0.297
	336	0.260	**0.281**	0.260	0.283	0.262	0.289	0.263	0.289	0.267	0.293	0.266	0.289	**0.254**	0.291	0.283	0.300	0.278	0.298	0.283	0.305	0.283	0.332
	720	0.342	**0.335**	**0.337**	0.337	0.344	0.343	0.339	0.341	0.349	0.346	0.341	0.342	0.340	0.343	0.359	0.351	0.350	0.346	0.355	0.352	0.348	0.384
	Avg	**0.239**	**0.261**	**0.239**	0.263	0.242	0.270	0.240	0.270	0.246	0.274	0.245	0.272	**0.239**	0.272	0.261	0.281	0.258	0.280	0.258	0.285	0.266	0.317
1st Count		11	20	6	11	8	8	8	0	0	0	0	0	5	3	5	0	0	0	0	0	0	0

C.2 Short-Term Forecasting

Table 11 reports the full horizon-wise short-term forecasting results on the four PEMS benchmarks with the lookback length fixed at 96. CoSTAR and GTR are both attached to the iTransformer backbone for a fair plug-in comparison.

Table 11: Horizon-wise short-term forecasting results on PEMS. The lookback length is fixed at 96. Best and second-best results are shown in **bold red** and blue underlined, respectively. Avg reports the average over the four forecasting horizons.

Dataset		CoSTAR (Ours)		VPNet 2026		GTR (iTransformer) 2026		RAFT 2025		xPatch 2025		CASA 2025		SparseTSF 2024		iTransformer 2024		PatchTST 2023		DLinear 2023	
Horizon		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
PEMS03	12	**0.062**	**0.166**	0.077	0.185	0.088	0.200	0.076	0.181	0.084	0.191	0.073	0.182	0.144	0.261	0.069	0.175	0.097	0.216	0.108	0.220
	24	**0.083**	**0.191**	0.111	0.224	0.111	0.224	0.104	0.207	0.126	0.233	0.110	0.223	0.245	0.352	0.098	0.209	0.143	0.269	0.183	0.291
	48	**0.133**	**0.244**	0.175	0.290	0.173	0.282	0.158	0.248	0.219	0.310	0.188	0.297	0.443	0.469	0.166	0.277	0.220	0.337	0.302	0.392
	96	0.235	0.331	0.232	0.341	**0.212**	0.321	0.239	**0.303**	0.321	0.381	0.323	0.395	0.760	0.615	0.274	0.370	0.244	0.355	0.386	0.466
	Avg	**0.128**	**0.233**	0.149	0.260	0.146	0.257	0.144	0.235	0.188	0.279	0.174	0.274	0.398	0.424	0.152	0.258	0.176	0.294	0.245	0.342
PEMS04	12	**0.074**	**0.179**	0.085	0.193	0.089	0.204	0.078	0.181	0.092	0.197	0.091	0.199	0.154	0.273	0.082	0.191	0.110	0.236	0.111	0.227
	24	**0.084**	**0.193**	0.118	0.231	0.115	0.234	0.091	0.195	0.135	0.240	0.124	0.237	0.270	0.376	0.101	0.213	0.152	0.275	0.183	0.301
	48	**0.100**	**0.213**	0.173	0.285	0.185	0.296	0.108	0.216	0.235	0.320	0.221	0.325	0.468	0.489	0.133	0.247	0.222	0.343	0.303	0.400
	96	**0.123**	**0.238**	0.215	0.327	0.215	0.330	0.130	0.242	0.351	0.401	0.385	0.443	0.774	0.636	0.165	0.276	0.307	0.414	0.385	0.471
	Avg	**0.095**	**0.206**	0.148	0.259	0.151	0.266	0.102	0.209	0.203	0.290	0.205	0.301	0.417	0.444	0.120	0.232	0.198	0.317	0.246	0.350
PEMS07	12	**0.058**	**0.152**	0.072	0.178	0.079	0.177	0.065	0.164	0.077	0.180	0.069	0.173	0.139	0.261	0.066	0.164	0.108	0.234	0.100	0.218
	24	**0.071**	**0.170**	0.105	0.217	0.108	0.210	0.080	0.184	0.123	0.227	0.102	0.211	0.259	0.369	0.087	0.190	0.147	0.280	0.186	0.303
	48	**0.088**	**0.189**	0.162	0.276	0.165	0.266	0.104	0.211	0.214	0.301	0.183	0.286	0.471	0.489	0.229	0.328	0.285	0.385	0.354	0.443
	96	0.179	0.261	0.221	0.328	0.212	0.303	**0.131**	**0.241**	0.325	0.376	0.343	0.399	0.713	0.611	0.322	0.384	0.324	0.425	0.466	0.521
	Avg	0.099	**0.193**	0.140	0.250	0.141	0.239	**0.095**	0.200	0.185	0.271	0.174	0.267	0.396	0.433	0.176	0.267	0.216	0.331	0.277	0.371
PEMS08	12	0.082	0.187	0.077	0.182	0.092	0.200	**0.071**	**0.177**	0.082	0.184	0.088	0.195	0.136	0.254	0.088	0.193	0.109	0.230	0.102	0.217
	24	0.124	0.232	0.111	0.220	0.138	0.247	**0.090**	**0.200**	0.123	0.224	0.137	0.243	0.234	0.352	0.136	0.242	0.150	0.280	0.169	0.287
	48	0.204	0.244	0.161	0.273	0.263	0.355	**0.122**	**0.235**	0.202	0.292	0.258	0.340	0.415	0.462	0.240	0.279	0.213	0.343	0.295	0.399
	96	0.242	**0.280**	**0.204**	0.313	0.560	0.531	0.212	0.292	0.315	0.368	0.481	0.449	0.661	0.584	0.320	0.343	0.279	0.401	0.387	0.477
	Avg	0.163	0.236	0.138	0.247	0.263	0.333	**0.124**	**0.226**	0.181	0.267	0.241	0.307	0.362	0.413	0.196	0.264	0.188	0.314	0.238	0.345

C.3 Backbone Compatibility

Table 12 reports paired results for five backbones and six datasets. At the reported precision, horizon-averaged MSE improves in 29 of 30 pairs, with one tie; MAE improves in 28 pairs. PatchTST on Electricity and Solar illustrates that a lower MSE need not imply a lower MAE. The comparison evaluates separately trained backbone–module combinations, rather than transfer of one trained module across architectures.

Table 12: Paired backbone results across five backbones and six datasets. The lookback length is fixed at 96. Original and + CoSTAR denote the backbone alone and the backbone equipped with CoSTAR, respectively; Improve is the relative error reduction.

Model		CASA						SparseTSF						iTransformer						PatchTST						DLinear			
Setup		Original		+ CoSTAR		Improve ↑		Original		+ CoSTAR		Improve ↑		Original		+ CoSTAR		Improve ↑		Original		+ CoSTAR		Improve ↑		Original		+ CoSTAR		Improve ↑	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Electricity	96	0.139	0.234	**0.136**	**0.234**	2.2%	0.0%	0.196	0.269	**0.158**	**0.262**	19.4%	2.6%	0.148	0.240	**0.137**	**0.235**	7.4%	2.1%	0.192	0.277	**0.182**	**0.283**	5.2%	-2.2%	0.210	0.301	**0.155**	**0.254**	26.2%	15.6%
	192	0.162	0.256	**0.156**	**0.250**	3.7%	2.3%	0.195	0.272	**0.165**	**0.268**	15.4%	1.5%	0.166	0.257	**0.158**	**0.252**	4.8%	1.9%	0.194	0.281	**0.187**	**0.285**	3.6%	-1.4%	0.210	0.304	**0.170**	**0.268**	19.0%	11.8%
	336	0.178	0.271	**0.173**	**0.268**	2.8%	1.1%	0.208	0.286	**0.176**	**0.280**	15.4%	2.1%	0.179	0.272	**0.171**	**0.267**	4.5%	1.8%	0.212	0.299	**0.201**	**0.297**	5.2%	0.7%	0.223	0.319	**0.180**	**0.282**	19.3%	11.6%
	720	0.220	0.308	**0.208**	**0.302**	5.5%	1.9%	0.251	0.321	**0.209**	**0.307**	16.7%	4.4%	0.210	0.299	**0.197**	**0.292**	6.2%	2.3%	0.250	0.326	**0.243**	**0.330**	2.8%	-1.2%	0.257	0.349	**0.224**	**0.321**	12.8%	8.0%
	Avg	0.175	0.267	**0.168**	**0.264**	4.0%	1.1%	0.212	0.287	**0.177**	**0.279**	16.5%	2.8%	0.176	0.267	**0.166**	**0.262**	5.7%	1.9%	0.212	0.296	**0.203**	**0.299**	4.2%	-1.0%	0.225	0.318	**0.182**	**0.281**	19.1%	11.6%
PEMS03	12	0.073	0.182	**0.066**	**0.172**	9.6%	5.5%	0.144	0.261	**0.080**	**0.194**	44.4%	25.7%	0.069	0.175	**0.062**	**0.166**	10.1%	5.1%	0.097	0.216	**0.079**	**0.190**	18.6%	12.0%	0.108	0.220	**0.066**	**0.174**	38.9%	20.9%
	24	0.110	0.223	**0.092**	**0.202**	16.4%	9.4%	0.245	0.352	**0.097**	**0.215**	60.4%	38.9%	0.098	0.209	**0.083**	**0.191**	15.3%	8.6%	0.143	0.269	**0.100**	**0.216**	30.1%	19.7%	0.183	0.291	**0.090**	**0.206**	50.8%	29.2%
	48	0.188	0.297	**0.145**	**0.256**	22.9%	13.8%	0.443	0.469	**0.140**	**0.256**	68.4%	45.4%	0.166	0.277	**0.133**	**0.244**	19.9%	11.9%	0.220	0.337	**0.111**	**0.229**	49.5%	32.0%	0.302	0.392	**0.128**	**0.253**	57.6%	35.5%
	96	0.323	0.395	**0.247**	**0.343**	23.5%	13.2%	0.760	0.615	**0.187**	**0.308**	75.4%	49.9%	0.274	0.370	**0.235**	**0.331**	14.2%	10.5%	0.244	0.355	**0.129**	**0.249**	47.1%	29.9%	0.386	0.466	**0.186**	**0.307**	51.8%	34.1%
	Avg	0.173	0.274	**0.138**	**0.243**	20.2%	11.3%	0.398	0.424	**0.126**	**0.243**	68.3%	42.7%	0.152	0.258	**0.128**	**0.233**	15.8%	9.7%	0.176	0.294	**0.105**	**0.221**	40.3%	24.8%	0.245	0.342	**0.117**	**0.235**	52.2%	31.3%
PEMS04	12	0.091	0.199	**0.075**	**0.181**	17.6%	9.0%	0.154	0.273	**0.086**	**0.203**	44.2%	25.6%	0.082	0.191	**0.074**	**0.179**	9.8%	6.3%	0.110	0.236	**0.073**	**0.181**	33.6%	23.3%	0.111	0.227	**0.070**	**0.175**	36.9%	22.9%
	24	0.124	0.237	**0.095**	**0.208**	23.4%	12.2%	0.270	0.376	**0.097**	**0.217**	64.1%	42.3%	0.101	0.213	**0.084**	**0.193**	16.8%	9.4%	0.152	0.275	**0.082**	**0.192**	46.1%	30.2%	0.183	0.301	**0.086**	**0.199**	53.0%	33.9%
	48	0.221	0.325	**0.136**	**0.255**	38.5%	21.5%	0.468	0.489	**0.145**	**0.277**	69.0%	43.4%	0.133	0.247	**0.100**	**0.213**	24.8%	13.8%	0.222	0.343	**0.093**	**0.211**	58.1%	38.5%	0.303	0.400	**0.116**	**0.238**	61.7%	40.5%
	96	0.385	0.443	**0.220**	**0.329**	42.9%	25.7%	0.774	0.636	**0.133**	**0.260**	82.8%	59.1%	0.165	0.276	**0.123**	**0.238**	25.5%	13.8%	0.307	0.414	**0.102**	**0.223**	66.8%	46.1%	0.385	0.471	**0.150**	**0.276**	61.0%	41.4%
	Avg	0.205	0.301	**0.132**	**0.243**	35.6%	19.3%	0.417	0.444	**0.115**	**0.239**	72.4%	46.2%	0.120	0.232	**0.095**	**0.206**	20.8%	11.2%	0.198	0.317	**0.087**	**0.202**	56.1%	36.3%	0.245	0.350	**0.105**	**0.222**	57.1%	36.6%
PEMS07	12	0.069	0.173	**0.060**	**0.156**	13.0%	9.8%	0.139	0.261	**0.076**	**0.190**	45.3%	27.2%	0.066	0.164	**0.058**	**0.152**	12.1%	7.3%	0.108	0.234	**0.081**	**0.201**	25.0%	14.1%	0.100	0.218	**0.058**	**0.161**	42.0%	26.1%
	24	0.102	0.211	**0.082**	**0.183**	19.6%	13.3%	0.259	0.369	**0.120**	**0.251**	53.7%	32.0%	0.087	0.190	**0.071**	**0.170**	18.4%	10.5%	0.147	0.280	**0.078**	**0.192**	46.9%	31.4%	0.186	0.303	**0.076**	**0.188**	59.1%	38.0%
	48	0.183	0.286	**0.123**	**0.226**	32.8%	21.0%	0.471	0.489	**0.124**	**0.251**	73.7%	48.7%	0.229	0.328	**0.088**	**0.189**	61.6%	42.4%	0.285	0.385	**0.099**	**0.220**	65.3%	42.9%	0.354	0.443	**0.107**	**0.230**	69.8%	48.1%
	96	0.343	0.399	**0.188**	**0.288**	45.2%	27.8%	0.713	0.611	**0.267**	**0.387**	62.6%	36.7%	0.322	0.384	**0.179**	**0.261**	44.4%	32.0%	0.324	0.425	**0.120**	**0.242**	63.0%	43.1%	0.466	0.521	**0.146**	**0.272**	68.7%	47.8%
	Avg	0.174	0.267	**0.113**	**0.213**	35.1%	20.2%	0.395	0.432	**0.147**	**0.270**	62.8%	37.5%	0.176	0.267	**0.099**	**0.193**	43.7%	27.7%	0.216	0.331	**0.095**	**0.214**	56.0%	35.3%	0.277	0.371	**0.097**	**0.213**	65.0%	42.6%
PEMS08	12	0.088	0.195	**0.081**	**0.183**	8.0%	6.2%	0.136	0.254	**0.089**	**0.208**	34.6%	18.1%	0.088	0.193	**0.082**	**0.187**	6.8%	3.1%	0.109	0.230	**0.096**	**0.212**	11.9%	7.8%	0.102	0.217	**0.069**	**0.176**	32.4%	18.9%
	24	0.137	0.243	**0.118**	**0.222**	13.9%	8.6%	0.234	0.352	**0.107**	**0.231**	54.3%	34.4%	0.136	0.242	**0.124**	**0.232**	8.8%	4.1%	0.150	0.280	**0.141**	**0.269**	6.0%	3.9%	0.169	0.287	**0.090**	**0.206**	46.7%	28.2%
	48	0.258	0.340	**0.202**	**0.297**	21.7%	12.6%	0.415	0.462	**0.128**	**0.252**	69.2%	45.5%	0.240	0.279	**0.204**	**0.244**	15.0%	12.5%	0.213	0.343	**0.108**	**0.230**	49.3%	32.9%	0.295	0.399	**0.143**	**0.271**	51.5%	32.1%
	96	0.481	0.449	**0.353**	**0.384**	26.6%	14.5%	0.661	0.584	**0.191**	**0.317**	71.1%	45.7%	0.320	0.343	**0.242**	**0.280**	24.4%	18.4%	0.279	0.401	**0.140**	**0.267**	49.8%	33.4%	0.387	0.477	**0.217**	**0.345**	43.9%	27.7%
	Avg	0.241	0.307	**0.189**	**0.272**	21.6%	11.4%	0.361	0.413	**0.129**	**0.252**	64.3%	39.0%	0.196	0.264	**0.163**	**0.236**	16.8%	10.6%	0.188	0.314	**0.121**	**0.244**	35.6%	22.3%	0.238	0.345	**0.130**	**0.249**	45.4%	27.8%
Solar	96	0.197	0.238	**0.198**	**0.238**	-0.5%	0.0%	0.253	0.316	**0.178**	**0.249**	29.6%	21.2%	0.203	0.235	**0.200**	**0.234**	1.5%	0.4%	0.230	0.277	**0.203**	**0.294**	11.7%	-6.1%	0.285	0.372	**0.225**	**0.312**	21.1%	16.1%
	192	0.227	0.269	**0.228**	**0.260**	-0.4%	3.3%	0.255	0.305	**0.203**	**0.254**	20.4%	16.7%	0.249	0.281	**0.246**	**0.275**	1.2%	2.1%	0.260	0.295	**0.239**	**0.323**	8.1%	-9.5%	0.315	0.393	**0.258**	**0.338**	18.1%	14.0%
	336	0.245	0.276	**0.244**	**0.273**	0.4%	1.1%	0.265	0.308	**0.209**	**0.278**	21.1%	9.7%	0.249	0.275	**0.246**	**0.272**	1.2%	1.1%	0.277	0.306	**0.242**	**0.323**	12.6%	-5.6%	0.349	0.413	**0.299**	**0.363**	14.3%	12.1%
	720	0.245	0.277	**0.246**	**0.277**	-0.4%	0.0%	0.262	0.303	**0.215**	**0.281**	17.9%	7.3%	0.250	0.276	**0.250**	**0.275**	0.0%	0.4%	0.276	0.305	**0.268**	**0.331**	2.9%	-8.5%	0.354	0.410	**0.309**	**0.367**	12.7%	10.5%
	Avg	0.229	0.265	**0.229**	**0.262**	0.0%	1.1%	0.259	0.308	**0.201**	**0.266**	22.4%	13.6%	0.238	0.267	**0.235**	**0.264**	1.3%	1.1%	0.261	0.296	**0.238**	**0.318**	8.8%	-7.4%	0.326	0.397	**0.273**	**0.345**	16.3%	13.1%

C.4 Ablation

Table 13 reports horizon-wise ablations across four backbones. The full design performs best in the DLinear horizon-averaged comparison in the main text, while some alternatives improve individual settings. For example, removing input decomposition benefits PatchTST on Solar at $H = 192$. These differences show that the contribution of each stage depends on the backbone and forecasting task.

C.5 Stability

Table 14 reports mean and standard deviation over five seeds for DLinear and PatchTST with CoSTAR. Figure 6 visualizes the corresponding seed-level distributions. These results character-

Table 13: Horizon-wise ablation results. The lookback length is fixed at 96. Best results are in **bold**; Original + CoSTAR denotes the full model.

		Original + CoSTAR		w/o Input Decomposition		w/o Output Correction		Common → Mean Pooling		Statistics → Mean Pooling		Statistics → Attention Pooling	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
DLinear													
Electricity	96	**0.155**	**0.254**	0.159	**0.254**	0.179	0.276	0.192	0.275	0.175	0.271	0.171	0.268
	192	0.170	0.268	**0.167**	**0.262**	0.189	0.282	0.191	0.278	0.183	0.278	0.196	0.293
	336	**0.180**	0.282	0.182	**0.279**	0.201	0.297	0.203	0.293	0.197	0.294	0.193	0.293
	720	**0.224**	**0.321**	0.250	0.340	0.235	0.329	0.236	0.324	0.238	0.334	0.231	0.328
PEMS03	12	**0.066**	**0.174**	0.073	0.181	0.074	0.186	0.086	0.196	0.070	0.181	0.072	0.183
	24	**0.090**	**0.206**	0.098	0.213	0.105	0.227	0.129	0.243	0.099	0.221	0.106	0.229
	48	**0.128**	**0.253**	0.143	0.266	0.173	0.303	0.188	0.306	0.161	0.289	0.173	0.302
	96	**0.186**	**0.307**	0.190	0.311	0.276	0.387	0.251	0.363	0.238	0.355	0.268	0.388
PEMS07	12	**0.058**	**0.161**	0.063	0.169	0.062	0.169	0.076	0.186	0.061	0.166	0.062	0.168
	24	**0.076**	**0.188**	0.080	0.194	0.088	0.209	0.121	0.242	0.083	0.201	0.090	0.212
	48	**0.107**	**0.230**	0.117	0.241	0.163	0.292	0.176	0.301	0.140	0.268	0.178	0.308
	96	**0.146**	**0.272**	0.160	0.284	0.215	0.340	0.267	0.375	0.189	0.316	0.211	0.337
Solar	96	**0.225**	**0.312**	0.226	**0.312**	0.256	0.352	0.230	0.317	0.233	0.326	0.254	0.349
	192	**0.258**	0.338	0.261	**0.337**	0.287	0.375	0.263	0.341	0.267	0.353	0.284	0.372
	336	**0.299**	**0.363**	0.306	0.369	0.324	0.396	0.312	0.378	0.308	0.378	0.317	0.392
	720	**0.309**	**0.367**	0.316	0.374	0.331	0.396	0.325	0.386	0.318	0.382	0.320	0.392
PatchTST													
Electricity	96	**0.182**	0.283	0.195	0.299	0.208	0.304	0.189	**0.279**	0.183	0.280	0.199	0.297
	192	**0.187**	0.285	0.189	**0.279**	0.208	0.303	0.190	0.280	0.190	0.285	0.200	0.298
	336	**0.201**	0.297	0.204	**0.294**	0.224	0.316	0.207	0.295	0.213	0.309	0.217	0.312
	720	**0.243**	0.330	0.246	0.343	0.264	0.346	0.246	**0.326**	0.246	0.332	0.257	0.343
PEMS03	12	**0.079**	**0.190**	0.099	0.213	0.099	0.223	0.097	0.214	0.097	0.219	0.094	0.211
	24	**0.100**	**0.216**	0.103	0.223	0.152	0.276	0.122	0.243	0.132	0.259	0.133	0.255
	48	**0.111**	**0.229**	0.112	0.230	0.194	0.316	0.150	0.268	0.185	0.309	0.186	0.308
	96	**0.129**	**0.249**	0.174	0.294	0.218	0.339	0.173	0.295	0.214	0.334	0.210	0.335
PEMS07	12	0.081	0.201	0.081	0.200	0.082	0.201	0.084	0.203	0.083	0.206	**0.080**	**0.199**
	24	**0.078**	**0.192**	0.080	0.201	0.135	0.265	0.089	0.210	0.124	0.254	0.145	0.283
	48	**0.099**	**0.220**	0.103	0.233	0.190	0.325	0.121	0.248	0.170	0.302	0.185	0.320
	96	**0.120**	**0.242**	0.156	0.292	0.236	0.362	0.147	0.273	0.206	0.338	0.214	0.343
Solar	96	0.203	0.294	**0.194**	0.259	0.254	0.310	0.200	**0.258**	0.207	0.289	0.224	0.287
	192	0.239	0.323	**0.207**	**0.267**	0.299	0.351	0.218	0.276	0.237	0.312	0.258	0.310
	336	0.242	0.323	**0.236**	**0.284**	0.328	0.357	0.244	0.290	0.262	0.323	0.289	0.323
	720	0.268	0.331	**0.236**	**0.287**	0.343	0.378	0.246	0.293	0.265	0.330	0.282	0.324
CASA													
Electricity	96	**0.136**	0.234	0.142	0.239	0.139	0.238	0.137	**0.233**	0.139	0.238	0.148	0.251
	192	**0.156**	**0.250**	0.159	0.253	0.158	0.254	0.161	0.256	0.159	0.254	0.159	0.254
	336	0.173	0.268	0.180	0.273	**0.171**	**0.267**	0.175	0.269	0.173	0.269	0.176	0.272
	720	0.208	0.302	**0.203**	**0.297**	0.204	**0.297**	0.220	0.308	0.219	0.309	0.209	0.310
PEMS03	12	**0.066**	**0.172**	0.070	0.177	0.070	0.177	0.072	0.179	0.070	0.177	0.070	0.178
	24	**0.092**	**0.202**	0.099	0.211	0.100	0.214	0.103	0.216	0.101	0.214	0.100	0.214
	48	**0.145**	**0.256**	0.164	0.275	0.157	0.271	0.170	0.282	0.157	0.271	0.156	0.269
	96	**0.247**	**0.343**	0.265	0.359	0.255	0.351	0.288	0.376	0.256	0.354	0.256	0.354
PEMS07	12	**0.060**	**0.156**	0.062	0.162	0.063	0.165	0.068	0.172	0.064	0.166	0.063	0.164
	24	**0.082**	**0.183**	0.087	0.192	0.091	0.200	0.099	0.210	0.092	0.201	0.090	0.197
	48	**0.123**	**0.226**	0.135	0.245	0.155	0.267	0.168	0.278	0.151	0.262	0.143	0.256
	96	**0.188**	**0.288**	0.214	0.318	0.234	0.338	0.272	0.359	0.235	0.337	0.223	0.324
Solar	96	0.198	0.238	0.198	0.240	0.195	0.235	**0.193**	0.236	0.194	**0.234**	0.196	0.235
	192	0.228	0.260	0.231	0.264	0.229	0.260	0.227	0.263	**0.225**	**0.258**	0.228	0.259
	336	0.244	0.273	0.245	0.277	0.243	**0.272**	**0.238**	0.273	0.243	0.273	0.242	0.273
	720	0.246	0.277	0.242	0.277	0.246	**0.276**	**0.240**	**0.276**	0.247	**0.276**	0.248	**0.276**
iTransformer													
Electricity	96	**0.137**	**0.235**	0.140	0.236	0.141	0.237	0.146	0.239	0.141	0.237	0.140	0.237
	192	**0.158**	0.252	**0.158**	**0.251**	0.160	0.256	0.161	0.254	**0.158**	0.254	0.160	0.255
	336	**0.171**	**0.267**	0.172	**0.267**	0.175	0.272	0.179	0.273	0.177	0.273	0.172	0.269
	720	**0.197**	**0.292**	0.204	0.296	0.205	0.300	0.214	0.302	0.202	0.297	0.206	0.301
PEMS03	12	**0.062**	**0.166**	0.065	0.170	0.065	0.171	0.065	0.170	0.064	0.170	0.064	0.169
	24	**0.083**	**0.191**	0.088	0.199	0.087	0.199	0.091	0.203	0.088	0.200	0.088	0.199
	48	**0.133**	**0.244**	0.150	0.264	0.139	0.249	0.254	0.351	0.138	0.247	0.137	0.248
	96	0.235	0.331	0.419	0.473	**0.218**	**0.321**	0.696	0.608	0.419	0.471	0.224	0.328
PEMS07	12	**0.058**	**0.152**	0.059	0.153	0.062	0.159	0.061	0.156	0.061	0.159	0.061	0.158
	24	**0.071**	**0.170**	0.074	0.174	0.079	0.183	0.076	0.176	0.076	0.179	0.078	0.182
	48	**0.088**	**0.189**	0.201	0.284	0.098	0.208	0.249	0.326	0.094	0.200	0.095	0.201
	96	0.179	0.261	0.233	0.315	0.224	0.321	0.278	0.355	**0.116**	0.220	0.118	**0.219**
Solar	96	0.200	0.234	**0.198**	**0.231**	0.207	0.237	0.199	0.233	0.205	0.237	0.202	0.235
	192	0.246	**0.275**	0.245	0.277	0.246	0.277	**0.243**	0.277	0.244	0.276	0.246	0.278
	336	0.246	**0.272**	0.246	**0.272**	0.247	**0.272**	**0.245**	**0.272**	**0.245**	0.273	0.246	0.274
	720	0.250	0.275	0.246	**0.273**	0.250	0.275	**0.245**	**0.273**	0.252	0.276	0.249	0.276

ize variability within each evaluated configuration; they are not a paired statistical significance test against every baseline.

C.6 Cross-Split Stability of the Common State

Figure 7 extends the correlation diagnostic to Electricity using PatchTST with $L = H = 96$. The mean absolute training–test correlation change is 0.1204 for the raw input and 0.0128 for the common component, approximately 9.4 times smaller. Together with the PEMS04 result, this supports greater stability of the measured correlation structure on these two datasets.

Table 14: Five-seed stability analysis. The lookback length is fixed at 96, and entries report mean ± standard deviation over five runs.

		DLinear + CoSTAR		PatchTST + CoSTAR	
		MSE	MAE	MSE	MAE
Electricity	96	0.156±0.001	0.256±0.001	0.182±0.004	0.284±0.005
	192	0.168±0.001	0.267±0.001	0.191±0.001	0.287±0.001
	336	0.181±0.001	0.284±0.001	0.204±0.002	0.302±0.002
	720	0.232±0.004	0.331±0.004	0.244±0.002	0.332±0.003
Solar	96	0.224±0.001	0.311±0.001	0.186±0.003	0.263±0.005
	192	0.258±0.001	0.339±0.002	0.209±0.004	0.281±0.007
	336	0.298±0.001	0.362±0.001	0.228±0.003	0.290±0.005
	720	0.310±0.001	0.368±0.001	0.233±0.004	0.299±0.007
PEMS07	12	0.059±0.000	0.162±0.000	0.082±0.003	0.204±0.005
	24	0.075±0.000	0.188±0.001	0.121±0.003	0.252±0.004
	48	0.109±0.001	0.232±0.000	0.109±0.009	0.232±0.009
	96	0.147±0.003	0.274±0.003	0.139±0.026	0.270±0.028

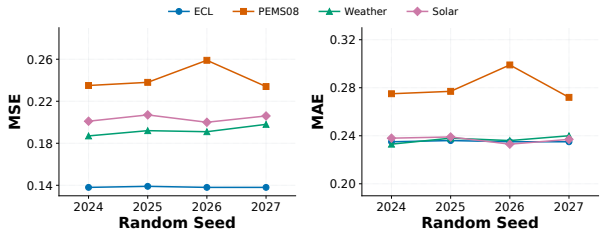

Figure 6: Seed analysis: distribution of test MSE across five random seeds for representative backbones equipped with CoSTAR.

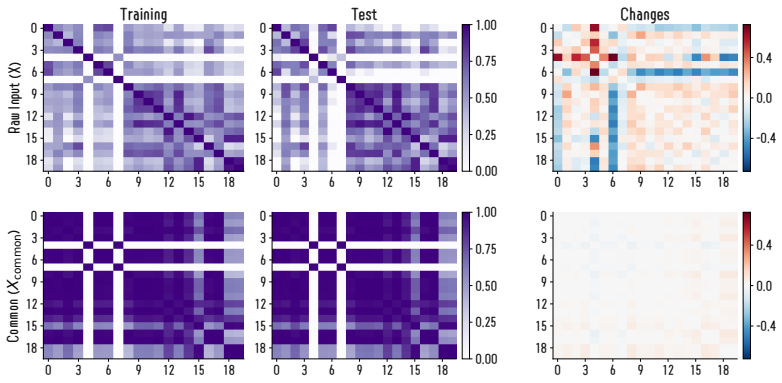

Figure 7: Cross-split stability of the learned common state on Electricity (PatchTST backbone, lookback 96, prediction length 96). *Top:* cross-channel Pearson correlation of the raw input on the training split (left), the test split (middle), and their difference (right). *Bottom:* the same three panels for the learned common component. The common component changes much less between splits (mean absolute change 0.0128) than the raw input (0.1204).

C.7 Plug-in Comparisons

This section reports the full plug-in comparison behind Table 6 in the main text. Tables 15 and 16 benchmark CoSTAR against GTR and HN-MVTS, respectively, under the same two backbones (DLinear and PatchTST) and two datasets (PEMS04 and PEMS07); green cells mark settings where

CoSTAR is better and red cells where it is worse. CoSTAR wins in nearly all settings, with only a few short-horizon PatchTST exceptions.

Table 15: Comparison with GTR at lookback 96.

Model	Horizon	PEMS04				PEMS07			
		Ours		GTR		Ours		GTR	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
DLinear	12	**0.070**	**0.175**	0.083	0.191	**0.058**	**0.161**	0.068	0.173
	24	**0.086**	**0.199**	0.111	0.222	**0.076**	**0.188**	0.103	0.212
	48	**0.116**	**0.238**	0.154	0.262	**0.107**	**0.230**	0.159	0.257
	96	**0.150**	**0.276**	0.191	0.294	**0.146**	**0.272**	0.209	0.297
PatchTST	12	**0.073**	**0.181**	0.101	0.217	0.081	0.201	**0.077**	**0.188**
	24	**0.082**	**0.192**	0.143	0.262	**0.078**	**0.192**	0.100	0.215
	48	**0.093**	**0.211**	0.229	0.340	**0.099**	**0.220**	0.177	0.287
	96	**0.102**	**0.223**	0.261	0.369	**0.120**	**0.242**	0.243	0.341

Table 16: Comparison with HN-MVTS at lookback 96.

Model	Horizon	PEMS04				PEMS07			
		Ours		HN-MVTS		Ours		HN-MVTS	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
DLinear	12	**0.070**	**0.175**	0.120	0.232	**0.058**	**0.161**	0.103	0.214
	24	**0.086**	**0.199**	0.224	0.323	**0.076**	**0.188**	0.210	0.309
	48	**0.116**	**0.238**	0.506	0.510	**0.107**	**0.230**	0.497	0.496
	96	**0.150**	**0.276**	1.002	0.754	**0.146**	**0.272**	0.996	0.737
PatchTST	12	**0.073**	**0.181**	0.099	0.215	0.081	0.201	**0.077**	**0.186**
	24	**0.082**	**0.192**	0.150	0.267	**0.078**	**0.192**	0.126	0.241
	48	**0.093**	**0.211**	0.265	0.361	**0.099**	**0.220**	0.230	0.327
	96	**0.102**	**0.223**	0.393	0.449	**0.120**	**0.242**	0.400	0.428

C.8 Frozen-Backbone Adaptation

To assess adaptation with fixed backbone parameters, we first pretrain a forecaster without CoSTAR, freeze it, and then train only CoSTAR on the target dataset. Table 17 compares this protocol with the backbone alone and joint training for DLinear and iTransformer on Electricity, PEMS04, and PEMS03. The lookback is fixed at 96. This experiment tests whether CoSTAR can improve a fixed forecaster through its input and output interface.

Every completed setting improves both MSE and MAE over the backbone alone. On Electricity, DLinear improves from average MSE 0.225 to 0.185 with frozen-backbone adaptation, compared with 0.182 under joint training. On PEMS04, the corresponding iTransformer values are 0.120, 0.102, and 0.095. On PEMS03, DLinear improves from 0.245 to 0.134, compared with 0.117 jointly. Thus, frozen-backbone adaptation recovers a substantial fraction of the joint-training improvement in these examples. The completed iTransformer–PEMS03 frozen runs average 0.126 MSE and 0.235 MAE, compared with $0.152/0.258$ for the backbone and $0.128/0.233$ for joint training. Frozen adaptation is strongest at $H = 96$, where it reaches 0.208 MSE and 0.316 MAE, improving on both the backbone ($0.274/0.370$) and joint training ($0.235/0.331$).

Table 17: Frozen-backbone adaptation at lookback 96.

Model	Dataset	Horizon	Backbone		Joint		Frozen	
			MSE	MAE	MSE	MAE	MSE	MAE
iTransformer	Electricity	96	0.148	0.240	**0.137**	**0.235**	0.140	0.236
		192	0.166	0.257	**0.158**	**0.252**	0.159	0.254
		336	0.179	0.272	**0.171**	**0.267**	0.172	0.268
		720	0.210	0.299	**0.197**	**0.292**	0.200	0.293
	PEMS04	12	0.082	0.191	0.074	**0.179**	**0.073**	**0.179**
		24	0.101	0.213	**0.084**	**0.193**	0.086	0.196
		48	0.133	0.247	**0.100**	**0.213**	0.110	0.222
		96	0.165	0.276	**0.123**	**0.238**	0.137	0.249
	PEMS03	12	0.069	0.175	**0.062**	**0.166**	0.066	0.172
		24	0.098	0.209	**0.083**	**0.191**	0.089	0.200
		48	0.166	0.277	**0.133**	**0.244**	0.141	0.253
		96	0.274	0.370	0.235	0.331	**0.208**	**0.316**
DLinear	Electricity	96	0.210	0.301	**0.155**	**0.254**	**0.155**	0.255
		192	0.210	0.304	**0.170**	**0.268**	0.171	0.269
		336	0.223	0.319	**0.180**	**0.282**	0.183	0.283
		720	0.257	0.349	**0.224**	**0.321**	0.229	0.324
	PEMS04	12	0.111	0.227	**0.070**	**0.175**	0.080	0.187
		24	0.183	0.301	**0.086**	**0.199**	0.100	0.213
		48	0.303	0.400	**0.116**	**0.238**	0.131	0.251
		96	0.385	0.471	**0.150**	**0.276**	0.186	0.301
	PEMS03	12	0.108	0.220	**0.066**	**0.174**	0.071	0.180
		24	0.183	0.291	**0.090**	**0.206**	0.096	0.212
		48	0.302	0.392	**0.128**	**0.253**	0.139	0.262
		96	0.386	0.466	**0.186**	**0.307**	0.229	0.340

D HYPERPARAMETER ANALYSIS

We vary the projection width r, initial correction strength $\sigma(\gamma_s)$, output-gate initialization γ_a, and input-removal initialization γ_r. Each sweep changes one quantity from the reference setting $r = 32$, $\sigma(\gamma_s) = 0.10$, $\gamma_a = -3.0$, and $\gamma_r = -2.0$. This sensitivity-study reference uses $r = 32$, whereas the default configuration in Table 9 uses $r = 16$. Tables 18–21 report horizon-wise results on PEMS04 and Electricity at $L = 96$. Increasing the projection width improves the averaged errors in the reported range; the remaining sweeps characterize the dependence on gate initialization.

Table 18: Sensitivity to the common-state rank r at lookback 96.

Dataset	Horizon	$r=4$		$r=8$		$r=16$		$r=32$		$r=64$	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
PEMS04	12	0.079	0.190	0.072	0.180	0.069	0.174	0.068	0.172	**0.067**	**0.170**
	24	0.105	0.224	0.090	0.205	0.087	0.200	0.081	0.192	**0.080**	**0.189**
	48	0.137	0.265	0.133	0.259	0.112	0.233	0.107	0.226	**0.097**	**0.213**
	96	0.164	0.262	0.181	0.312	0.162	0.292	0.173	0.294	**0.145**	**0.266**
Electricity	96	0.164	0.262	0.161	0.260	0.155	0.254	0.149	0.250	**0.146**	**0.248**
	192	0.175	0.272	0.171	0.269	0.166	0.265	**0.161**	**0.262**	0.161	0.262
	336	0.194	0.291	0.195	0.293	0.180	0.284	0.179	0.284	**0.175**	**0.282**
	720	0.235	0.330	0.242	0.341	0.227	0.326	0.222	0.322	**0.221**	**0.322**

Table 19: Sensitivity to the initial correction strength $\sigma(\gamma_s)$ at lookback 96.

Dataset	Horizon	$\sigma(\gamma_s)=0.01$		0.05		0.10		0.20		0.50	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
PEMS04	12	0.069	0.173	**0.068**	0.172	0.068	0.172	0.068	0.172	0.068	**0.171**
	24	0.084	0.197	0.082	0.193	**0.081**	0.192	0.081	**0.191**	0.081	0.192
	48	0.117	0.240	0.109	0.228	0.107	0.226	0.103	0.221	**0.101**	**0.218**
	96	0.152	0.277	0.140	0.265	0.145	0.271	**0.135**	**0.259**	0.152	0.271
Electricity	96	0.154	0.256	0.150	0.251	0.149	0.250	0.149	0.249	**0.148**	**0.248**
	192	0.167	0.268	0.162	0.263	0.161	0.262	**0.160**	0.262	0.162	0.266
	336	0.180	0.287	**0.178**	**0.283**	0.179	0.284	0.178	0.283	0.180	0.286
	720	0.223	0.324	0.225	0.325	0.222	0.322	**0.219**	**0.319**	0.223	0.322

Table 20: Sensitivity to the add-gate initialization γ_a at lookback 96.

Dataset	Horizon	$\gamma_a=-1$		$\gamma_a=-3$		$\gamma_a=-5$	
		MSE	MAE	MSE	MAE	MSE	MAE
PEMS04	12	**0.068**	**0.171**	0.068	0.172	0.069	0.173
	24	**0.081**	**0.192**	0.081	0.192	0.084	0.196
	48	**0.100**	**0.217**	0.107	0.226	0.114	0.236
	96	0.153	0.270	**0.145**	**0.271**	0.163	0.285
Electricity	96	**0.147**	**0.248**	0.149	0.250	0.153	0.255
	192	0.162	0.266	**0.161**	**0.262**	0.165	0.266
	336	**0.179**	0.286	0.179	**0.284**	0.180	0.286
	720	**0.216**	**0.316**	0.222	0.322	0.228	0.328

E INFERENCE EFFICIENCY

Tables 22 and 23 report paired inference measurements before and after adding CoSTAR. The overhead columns report Δ = CoSTAR − Base, so positive values indicate additional measured cost. Memory denotes peak allocated GPU memory, and latency is reported as mean ± standard deviation

Table 21: Sensitivity to the remove-gate initialization γ_r at lookback 96.

Dataset	Horizon	$\gamma_r = -1$ MSE	$\gamma_r = -1$ MAE	$\gamma_r = -2$ MSE	$\gamma_r = -2$ MAE	$\gamma_r = -3$ MSE	$\gamma_r = -3$ MAE
PEMS04	12	**0.068**	**0.172**	0.068	0.172	0.068	0.172
	24	**0.081**	**0.192**	0.081	0.192	0.083	0.194
	48	0.109	0.229	0.107	0.226	**0.106**	**0.225**
	96	**0.144**	**0.265**	0.145	0.271	0.153	0.271
Electricity	96	0.152	0.253	**0.149**	0.250	0.149	**0.248**
	192	0.162	0.264	**0.161**	0.262	0.161	**0.261**
	336	**0.179**	0.286	0.179	**0.284**	0.180	0.286
	720	0.226	0.326	0.222	0.322	**0.220**	**0.321**

in ms/iter over 100 forward passes after 10 warm-up passes. The table reports the exact horizon used in each efficiency configuration. For compact PEMS test sets, the loader is cycled during timing; forecasting metrics are still evaluated once on the test set.

Peak-memory differences remain small across the recorded settings. The largest observed increase is 2.18 MB for PatchTST and the largest absolute change is 1.66 MB for DLinear. On Electricity, PatchTST adds 6.41–70.95 ms/iter across the four horizons, while DLinear adds 17.31–35.28 ms/iter. These measurements are hardware-, batch-size-, and implementation-dependent, and the reported standard deviations quantify within-run latency variation.

Table 22: Paired inference overhead of CoSTAR on PatchTST. Δ is the CoSTAR-equipped value minus the backbone value. Positive values denote increased measured cost; $\uparrow$ marks the overhead columns. Latency is reported as mean $\pm$ standard deviation over 100 forward passes after 10 warm-up passes.

Dataset	Horizon	Peak memory (MB)			Time (ms/iter)		
		Base	+CoSTAR	**Δ** $\uparrow$	Base	+CoSTAR	**Δ** $\uparrow$
Electricity	96	202.45	202.71	+0.26	125.41 ± 35.94	131.81 ± 31.17	+6.40
Electricity	192	221.80	220.86	-0.94	67.83 ± 17.34	121.78 ± 28.70	+53.95
Electricity	336	247.42	249.10	+1.68	74.43 ± 18.14	128.08 ± 20.91	+53.65
Electricity	720	319.91	319.41	-0.50	79.43 ± 18.15	150.38 ± 30.12	+70.95
PEMS04	96	320.22	321.60	+1.38	164.12 ± 27.65	155.60 ± 40.22	-8.52
PEMS04	192	343.16	343.88	+0.72	169.85 ± 36.51	178.72 ± 31.13	+8.87
PEMS04	336	377.84	380.02	+2.18	107.23 ± 25.77	77.45 ± 14.22	-29.78
PEMS04	720	468.09	470.07	+1.98	131.62 ± 24.76	83.42 ± 23.64	-48.20
PEMS08	12	186.53	187.02	+0.49	129.86 ± 31.39	124.33 ± 30.72	-5.53
PEMS08	24	188.90	189.39	+0.49	123.71 ± 31.92	125.87 ± 33.52	+2.16
PEMS08	48	193.47	193.97	+0.50	127.67 ± 26.49	132.38 ± 34.99	+4.71
PEMS08	96	202.09	200.73	-1.36	118.61 ± 32.18	112.83 ± 30.46	-5.78

F VISUALIZATION RESULTS

Figures 8–10 plot test MSE over training epochs on Solar, PEMS03, PEMS04, PEMS07, PEMS08, and Electricity at $L = 96$. These curves describe the training trajectories of each backbone and its CoSTAR counterpart.

Table 23: Paired inference overhead of CoSTAR on DLinear. Δ is the CoSTAR-equipped value minus the backbone value. Positive values denote increased measured cost; $\uparrow$ marks the overhead columns. Latency is reported as mean $\pm$ standard deviation over 100 forward passes after 10 warm-up passes.

Dataset	Horizon	Peak memory (MB)			Time (ms/iter)		
		Base	+CoSTAR	$\Delta \uparrow$	Base	+CoSTAR	$\Delta \uparrow$
Electricity	96	55.17	55.17	+0.00	32.85 ± 10.42	50.16 ± 34.84	+17.31
Electricity	192	80.34	80.34	+0.00	20.14 ± 5.77	49.71 ± 11.93	+29.57
Electricity	336	120.05	120.05	+0.00	22.90 ± 6.69	58.17 ± 11.74	+35.27
Electricity	720	234.19	234.19	+0.00	17.57 ± 1.59	35.45 ± 7.33	+17.88
PEMS04	12	36.99	36.99	+0.00	45.16 ± 15.26	49.52 ± 12.51	+4.36
PEMS04	24	40.81	40.81	+0.00	48.72 ± 49.86	55.82 ± 14.47	+7.10
PEMS04	48	50.28	50.28	+0.00	39.44 ± 14.88	58.62 ± 33.35	+19.18
PEMS04	96	67.38	65.72	-1.66	49.68 ± 14.74	56.87 ± 13.38	+7.19
PEMS08	12	27.92	27.92	+0.00	49.04 ± 11.64	59.75 ± 30.29	+10.71
PEMS08	24	30.45	30.45	+0.00	48.33 ± 14.53	61.69 ± 13.80	+13.36
PEMS08	48	35.58	35.58	+0.00	46.20 ± 15.98	49.02 ± 14.00	+2.82
PEMS08	96	47.86	48.86	+1.00	30.16 ± 12.65	61.35 ± 15.90	+31.19

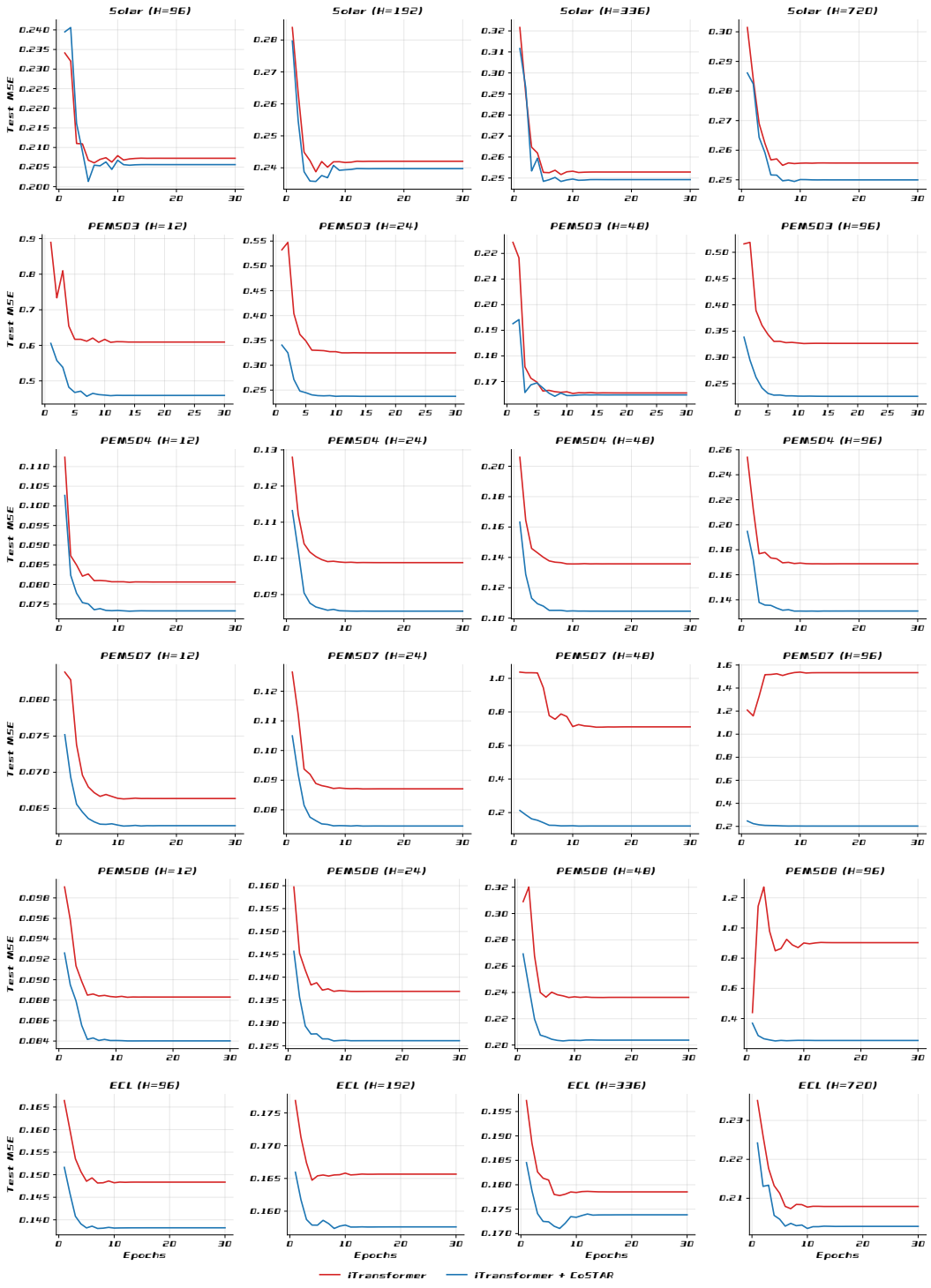

Figure 8: Training curves of MSE on a test set for the iTransformer base model and its CoSTAR-enhanced counterpart.

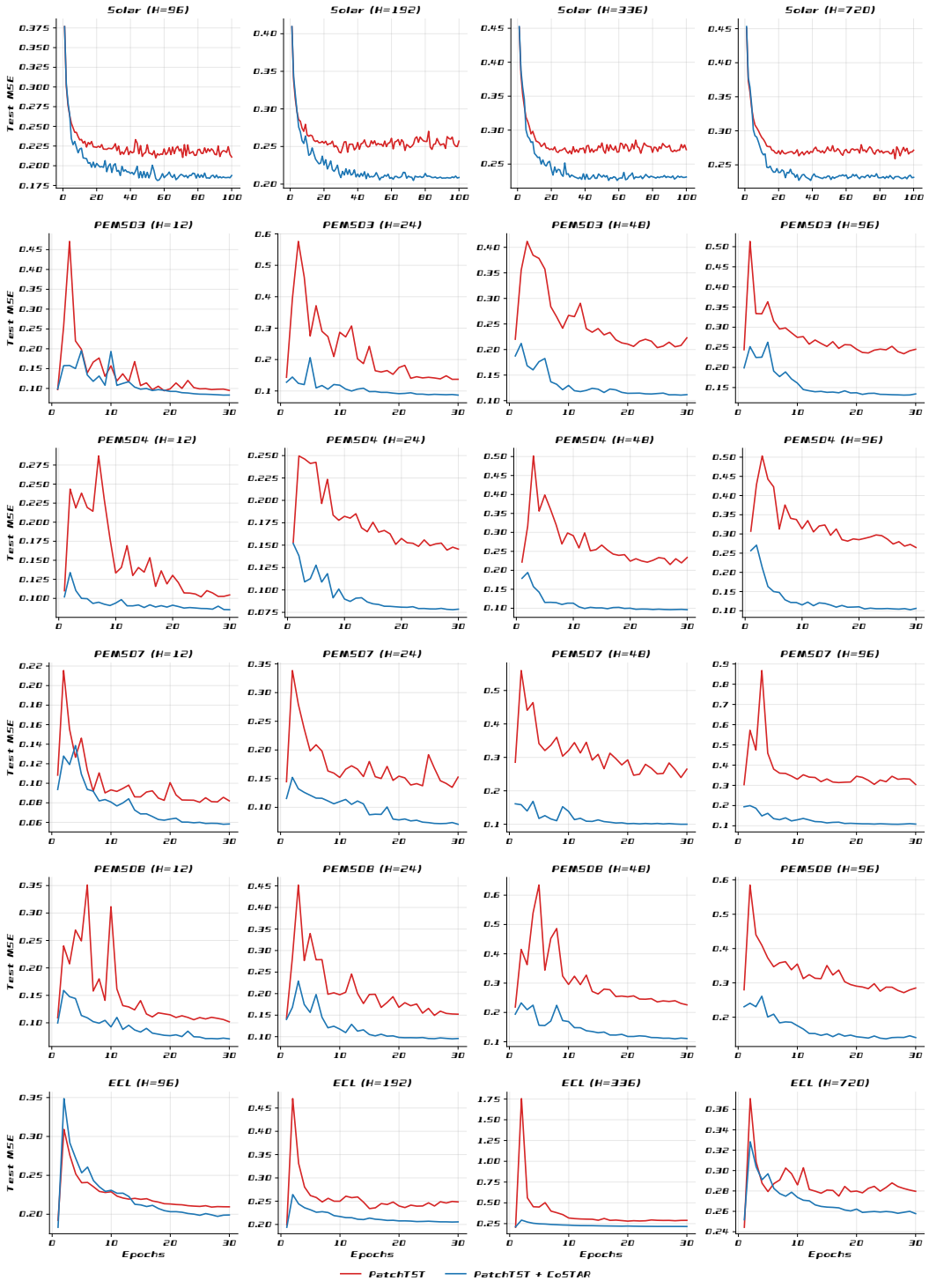

Figure 9: Training curves of MSE on a test set for the PatchTST base model and its CoSTAR-enhanced counterpart.

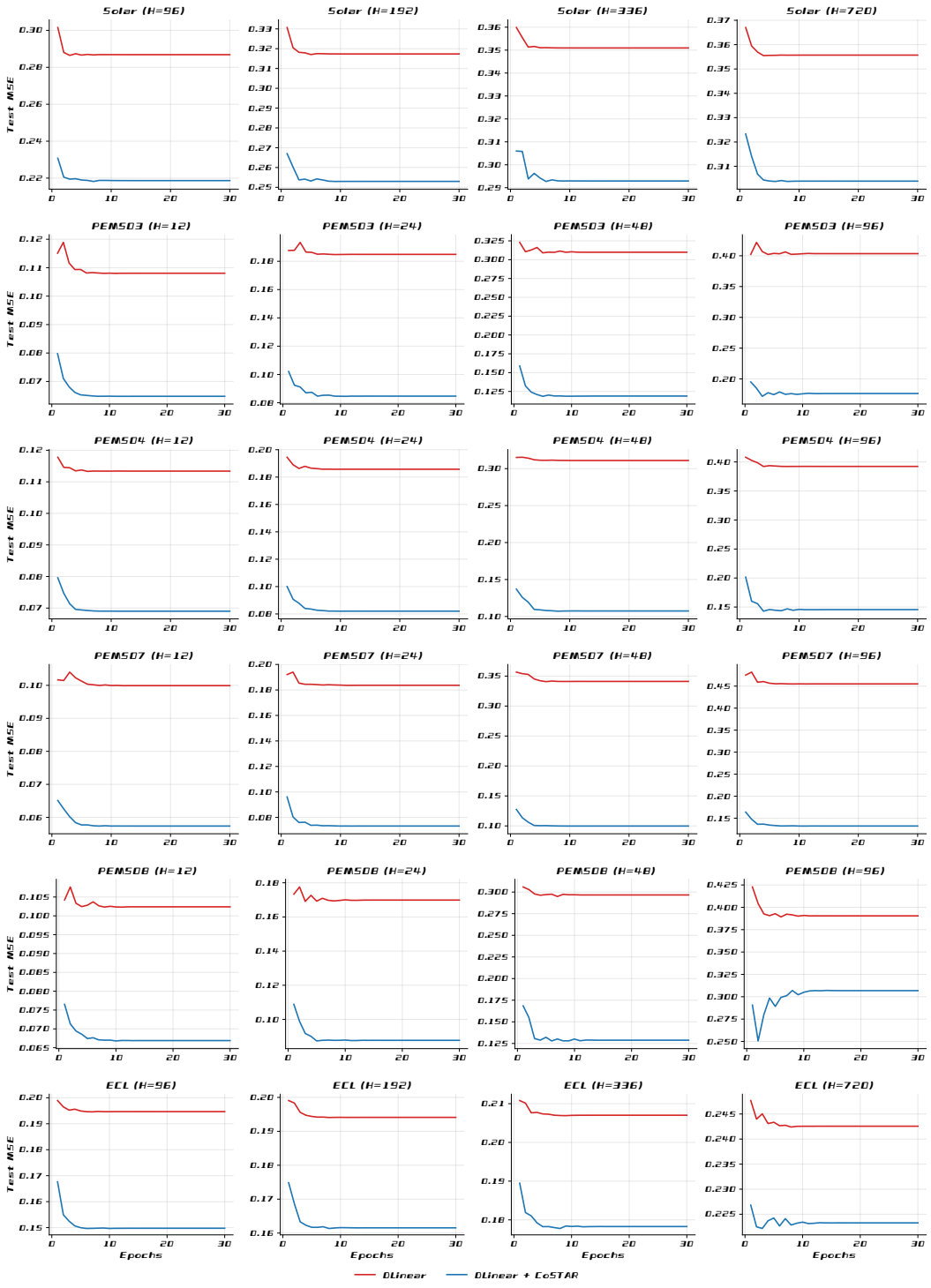

Figure 10: Training curves of MSE on a test set for the DLinear base model and its CoSTAR-enhanced counterpart.