WILD: Widely Linear Conditioning for Time Series Forecasting

Binli Luo[*]
Central South University
lblhandsome@gmail.com

Wanrong Ma[*]
National University of Defense Technology
mawanrong@nudt.edu.cn

Ning Gui[†]
Central South University
ninggui@gmail.com

Xianhan Tan[†]
Central South University
txhts1606@gmail.com

Abstract

Transformer-based time series forecasting models often suffer from *block-wise* attention collapse, where temporal smoothness and cross-variable correlations concentrate attention locally and restrict global modeling capacity. We argue that this collapse is partly a pre-attention representation-conditioning problem: attention logits inherit the spectral concentration and low effective rank of the input representation. We propose **W**idely L**i**near Frequency-**D**omain (**WILD**) conditioning, a lightweight, model-agnostic module inserted before self-attention. WILD maps inputs to the frequency domain, applies data-dependent spectral preconditioning to redistribute over-dominant frequency energy, and then uses asymmetric real/imaginary projections to realize a widely linear transform beyond fixed time-domain channel mixing. The preconditioning step is essential: it constructs the usable normalized frequency basis on which the widely linear branch can operate. Thus, WILD uses frequency-domain modeling to condition an existing attention layer, not to replace it with a new forecasting architecture. Extensive experiments on eight benchmark datasets show positive gains on 105 of 112 model-dataset-metric configurations across seven backbones. Attention diagnostics show that WILD substantially increases effective rank over the baseline, and a $W_r = W_i$ ablation identifies spectral preconditioning as the primary mechanism, with widely linear mixing providing a secondary, dataset-dependent gain.

1 Introduction

Transformer models are strong long-horizon forecasters, but their attention maps on smooth multivariate time series often collapse into local blocks: adjacent queries attend to nearly identical keys, weakening global dependency modeling. Prior work mainly addresses this by redesigning the attention architecture [Nie et al., 2023, Chen et al., 2024, Wu et al., 2021, Liu et al., 2022a]. We argue that the bottleneck is more basic: the representation fed to attention is often low effective-rank and poorly conditioned. Since

$$\mathbf{Q}\mathbf{K}^\top = \mathbf{X}\mathbf{W}_Q\mathbf{W}_K^\top\mathbf{X}^\top,$$

attention logits inherit the column-space and spectral concentration of the input representation. Within the common fixed channel-mixing view, a time-domain right-multiplication $X \mapsto XM$ can only rotate or reweight within $\mathrm{Col}(X)$; it does not directly rebalance the temporal spectrum once smoothness and cross-variable correlation have already concentrated the singular values.

[*]Equal contribution.
[†]Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

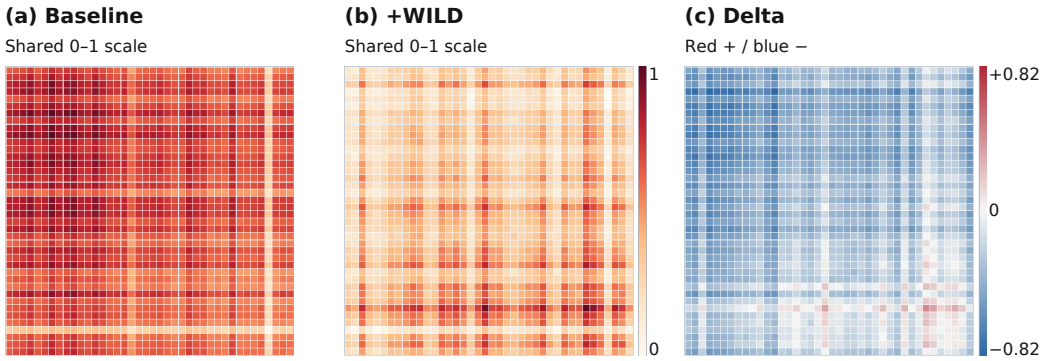


Figure 1: WILD attention case study on ETTm2. Panels (a) and (b) use the same min–max scale; panel (c) shows $\Delta_{ij} = A_{ij}^{\mathrm{WILD}} - A_{ij}^{\mathrm{base}}$. Red means WILD increases relative attention at a query-key location, blue means it decreases it, and white means little change. The delta view shows WILD suppressing broad block-wise concentration and reallocating attention to sharper selective structures.

Figure 1 illustrates the attention patterns motivating **W**idely L**i**near Frequency-**D**omain conditioning (**WILD**), a lightweight plug-and-play module inserted before attention. WILD first applies data-dependent spectral preconditioning in the Fourier domain, balancing over-dominant frequency energy and improving stable-rank conditions. This preconditioned spectrum is the substrate on which the next step can operate: asymmetric real/imaginary channel projections $W_r \neq W_i$ implement a widely linear transform containing a conjugate term and are not reducible to any fixed time-domain right-multiplication. After inverse FFT, residual fusion with the original input conditions the representation without changing the attention kernel.

Our contribution and novelty are threefold. First, we cast block-wise attention collapse as a representation-conditioning failure caused by spectral concentration, rather than only an attention-architecture limit. Second, WILD provides a pre-attention spectral conditioner: data-dependent preconditioning creates a usable frequency basis, and asymmetric real/imaginary projections realize a widely linear transform unavailable to fixed time-domain mixing. Third, we connect mechanism to evidence across eight datasets and seven Transformer backbones: WILD yields positive gains on 105 of 112 model-dataset-metric configurations, improves attention-rank diagnostics, and remains effective as a model-agnostic plugin. Additional attention visualizations, full horizon-level tables, and implementation details are provided in the Appendix.

2 Related Work

Time-series Transformers. Transformer forecasters improve long-horizon modeling by redesigning temporal and cross-variable dependency modeling. Autoformer introduces decomposition and auto-correlation [Wu et al., 2021]; Pyraformer and Pathformer use hierarchical or multi-scale pathways [Liu et al., 2022a, Chen et al., 2024]; Crossformer and iTransformer emphasize cross-dimension or inverted-token structure [Zhang and Yan, 2023, Liu et al., 2024]; and Nonstationary Transformer addresses distribution shift [Liu et al., 2022b]. These methods mainly alter the backbone architecture or attention pattern. WILD is complementary: it leaves the attention kernel and backbone intact, and instead improves the representation supplied to attention.

Frequency-domain forecasting. Spectral operations have been used for token mixing, efficient filtering, decomposition, and forecasting directly in the frequency domain. FNet replaces attention-like token mixing with Fourier transforms [Lee-Thorp et al., 2021]; AFNO uses adaptive Fourier operators [Guibas et al., 2022]; FEDformer combines decomposition with frequency-enhanced attention [Zhou et al., 2022]; FITS, Fredformer, TIMEMIXER++, and FreDF use frequency-domain parameterization, debiasing, or pattern mixing for forecasting [Xu et al., 2024, Piao et al., 2024, Wang et al., 2024, 2025]. WILD differs in objective, placement, and mechanism: it uses the frequency domain to condition the representation supplied to an existing attention layer, not to replace attention, filter tokens, or build a new forecasting architecture. Data-dependent spectral preconditioning first balances the usable frequency basis, and asymmetric real/imaginary projections then realize a widely linear transform that standard time-domain channel mixing cannot express. A detailed comparison with representative frequency-domain methods is given in Appendix F (Table 22). Recent parameter-efficient linear and frequency-domain forecasters are not direct substitutes for

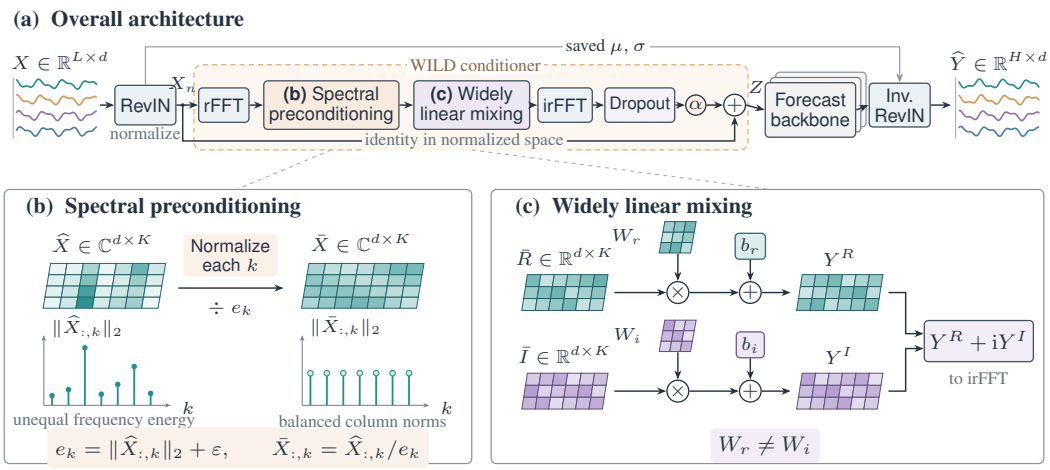


Figure 2: Overview of WILD. (a) The conditioner precedes the forecasting backbone; saved RevIN statistics are reused for final denormalization. (b) Spectral preconditioning balances per-frequency channel norms. (c) Separate affine real/imaginary projections implement widely linear mixing. After irFFT, dropout and the learnable gain α act on the transformed branch; the residual remains in normalized space. Tensor and spectrum sketches are schematic.

a pre-attention conditioner; Appendix L reports paired add-on tests, which are mixed rather than uniformly positive.

3 Problem Analysis and Method

3.1 Block-wise Attention as Representation Collapse

Let $X \in \mathbb{R}^{L\times d}$ denote a multivariate input window. In temporal self-attention, the unnormalized score matrix satisfies

$$G = QK^\top = XW_QW_K^\top X^\top, \qquad \operatorname{rank}(G) \le \min(\operatorname{rank}(X), d_k). \tag{1}$$

Smooth temporal dynamics and correlated channels render nearby rows of X highly similar, concentrating its singular spectrum. By Eq. (1), rows of the resulting attention matrix become nearly duplicated, inducing block-wise local attention and low effective rank—which we quantify via the entropy-based metric $\operatorname{erank}(M) = \exp(-\sum_i p_i \log p_i)$ with $p_i = \sigma_i(M)/\sum_j \sigma_j(M)$ (see Appendix G for visual diagnostics). This bottleneck directly formalizes the motivation from Section 1: because any fixed time-domain channel projection $X \mapsto XM$ remains confined to $\mathrm{Col}(X)$, escaping representation collapse requires intervening before attention. WILD achieves this by reconditioning the singular spectrum and applying a structurally distinct widely linear mixing step while preserving the original signal via a residual pathway.

3.2 WILD: Widely Linear Frequency-Domain Conditioning

As illustrated in Figure 2, WILD is inserted directly before a Transformer backbone. After window-level stationarization, we compute a unitary one-sided real FFT along the temporal dimension to separate real and imaginary components:

$$\widehat{X} = \mathrm{rFFT}(X^\top) \in \mathbb{C}^{d\times K}, \qquad R = \mathrm{Re}(\widehat{X}), \quad I = \mathrm{Im}(\widehat{X}). \tag{2}$$

For each frequency bin $k \in \{1, \dots, K\}$, we normalize spectral energy across channels:

$$e_k = \sqrt{\sum_{j=1}^{d} \left(R_{j,k}^2 + I_{j,k}^2\right)} + \varepsilon, \qquad \bar{R}_{j,k} = \frac{R_{j,k}}{e_k}, \quad \bar{I}_{j,k} = \frac{I_{j,k}}{e_k}. \tag{3}$$

Because e_k in Eq. (3) depends on the current window, normalization is data-dependent and nonlinear. WILD applies asymmetric channel projections, reconstructs $\widetilde{X} = \mathrm{irFFT}((Y^R + \mathrm{i}Y^I)^\top)$, and forms the residual output:

$$Y^R_{:,k} = W_r \bar{R}_{:,k} + b_r, \qquad Y^I_{:,k} = W_i \bar{I}_{:,k} + b_i, \qquad Z = X + \alpha \cdot \mathrm{Dropout}(\widetilde{X}, p). \tag{4}$$

The learnable parameters are $\{W_r, W_i, b_r, b_i, \alpha\}$. Throughout our experiments, Eq. (4) employs shape-preserving $d \to d$ projections, enabling WILD to be seamlessly integrated into existing backbones without modifying the attention mechanism. Base and WILD use the same RevIN protocol: the residual fusion stays in the normalized representation, and the final forecast is denormalized with the saved input-window statistics.

3.3 Why WILD Improves Conditioning

To justify Eqs. (2)–(4), we analyze conditioning via stable rank, $\mathrm{sr}(A) = \|A\|_F^2/\|A\|_2^2$, and the expressivity of asymmetric complex projections.

Lemma 3.1 (Spectral preconditioning improves stable rank under incoherence)**.** *Let $\widehat{X} = [a_1, \ldots, a_K] \in \mathbb{C}^{d\times K}$ be the rFFT representation in Eq.* (2)*, with active frequency norms $s_k = \|a_k\|_2 > 0$, $s_{\max} = \max_k s_k$, and normalized columns $u_k = a_k/s_k$ corresponding to Eq.* (3) *as $\varepsilon \to 0$. Let $U = [u_1, \ldots, u_K]$. If $\lambda_{\max}(U^*U) \le 1 + \delta$, then*

$$\mathrm{sr}(U) \ge \frac{K}{1+\delta}, \qquad \mathrm{sr}(\widehat{X}) \le \frac{\sum_{k=1}^K s_k^2}{s_{\max}^2}. \tag{5}$$

Consequently, spectral preconditioning strictly improves stable rank whenever $K/(1+\delta) > \sum_k s_k^2/s_{\max}^2$.

Sketch. Since $\|u_k\|_2 = 1$, we have $\|U\|_F^2 = K$ and $\|U\|_2^2 = \lambda_{\max}(U^*U) \le 1 + \delta$. For $\widehat{X}$, $\|\widehat{X}\|_F^2 = \sum_k s_k^2$ and $\|\widehat{X}\|_2^2 \ge s_{\max}^2$, yielding Eq. (5). Full derivations, including the ε-regularized case in Eq. (3), are provided in Appendix M. □

Lemma 3.2 (Widely linear transforms exceed fixed right-multiplications)**.** *Let $L \ge 3$ and $X_f = R + \mathrm{i}I \in \mathbb{C}^{d\times K}$. If $W_r, W_i \in \mathbb{R}^{d\times d}$ with $W_r \ne W_i$, then the frequency-domain projection in Eq.* (4)*, $X \mapsto \mathrm{irFFT}((W_r R + \mathrm{i}W_i I)^\top)$, cannot be reduced for all real inputs to a fixed time-domain channel mixing $X \mapsto XM$ for any $M \in \mathbb{R}^{d\times d}$.*

Sketch. We can decompose the asymmetric projection into linear and conjugate-linear branches:

$$W_r R + \mathrm{i}W_i I = \frac{W_r + W_i}{2} X_f + \frac{W_r - W_i}{2} \overline{X_f}. \tag{6}$$

At any non-DC, non-Nyquist frequency, the rFFT coefficient of a real signal spans all $z \in \mathbb{C}^d$. Equivalence to a fixed right-multiplication $X \mapsto XM$ would require $Az + B\bar{z} = Cz$ for all $z \in \mathbb{C}^d$ in Eq. (6), forcing the conjugate branch $B = (W_r - W_i)/2 = 0$ and thus $W_r = W_i$, a contradiction. See Appendix M for full details. □

Together, Lemmas 3.1 and 3.2 clarify the synergy within WILD. Spectral preconditioning in Eq. (3) acts as the enabling step: by preventing dominant frequency bins from monopolizing the spectrum (Eq. (5)), it provides the subsequent projections in Eq. (4) with a balanced frequency basis where the conjugate branch in Eq. (6) can effectively operate. This widely linear mixing unlocks a degree of freedom unavailable to standard time-domain channel projections. Rather than artificially inflating the algebraic rank beyond $\min(L, d)$, WILD substantially elevates the effective conditioning of the representation supplied to self-attention in Eq. (1).

4 Experiments

We empirically evaluate WILD through a sequence of claim-driven tests: (1) **Forecasting accuracy**—does WILD improve performance over strong baselines? (Section 4.1); (2) **Generality**—does WILD benefit diverse backbone architectures? (Section 4.2); (3) **Mechanism**—does WILD improve

effective rank and attention diversity as predicted by the theory? (Section 4.3 and Section 4.4). A time complexity analysis is provided in Appendix E.

Datasets: We evaluate on eight widely used multivariate time series forecasting datasets: ETTh1, ETTh2, ETTm1, and ETTm2 [Zhou et al., 2021]; Weather, Exchange, ECL (Electricity), and Traffic [Lai et al., 2018]. We follow standard preprocessing and evaluation protocols established in prior work [Zhou et al., 2021, Zeng et al., 2023, Hebrail and Berard, 2012]. Detailed dataset statistics and preprocessing procedures are provided in Appendix A.

Baselines: We integrate WILD into seven representative Transformer architectures: iTransformer [Liu et al., 2024], PatchTST [Nie et al., 2023], Crossformer [Zhang and Yan, 2023], Pyraformer [Liu et al., 2022a], Autoformer [Wu et al., 2021], Informer [Zhou et al., 2021], and Reformer [Kitaev et al., 2020]. All models use Reversible Instance Normalization (RevIN) [Kim et al., 2021] for fairness. Unless specified otherwise, iTransformer serves as our default backbone. Further implementation details are in Appendix A.

4.1 Main Results

Table 1 summarizes forecasting results, with the best and second-best scores marked in **bold** and underlined, respectively. Lower MSE and MAE indicate better accuracy. The WILD column is iTransformer equipped with WILD, averaged over horizons {96, 192, 336, 720}, matching the plug-and-play study. With iTransformer as the backbone, WILD ranks first on all 16 metrics across eight datasets, with the largest gains on ECL and Exchange. On ECL it reduces MSE to 0.167 and MAE to 0.256 (6.2%/5.2% over iTransformer); on Exchange it improves MSE by 10.0% and MAE by 3.2%. A matched five-seed rerun corroborates this on ECL and Weather: ECL MSE improves at all horizons (5/5 seeds; $p \leq 1.2 \times 10^{-2}$) and Weather-H96 MSE by 4.9% ($p = 9.4 \times 10^{-4}$), while ECL MAE gains only at H=96 and H=720 (Appendix J).

These results establish the accuracy side of the claim; the mechanism is tested directly in the ablations and diagnostics below. Table 10 in Appendix C.1 reports the full horizon-level results.

Table 1: Horizon-averaged forecasting performance. Bold and underline mark the lowest and second-lowest distinct values per dataset and metric; ties receive the same mark. WILD with iTransformer achieves the best MSE and MAE on all eight datasets. # Best counts first-place results across datasets.

Models	**WILD Ours**		iTransformer ICLR 2024		PatchTST ICLR 2023		Crossformer ICLR 2023		Pyraformer ICLR 2023		Autoformer Neurips 2022		Informer AAAI 2021		Reformer ICLR 2020	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	**0.444**	**0.431**	0.454	0.447	0.469	0.454	0.529	0.522	0.865	0.731	0.518	0.500	1.078	0.813	0.961	0.757
ETTh2	**0.367**	**0.395**	0.383	0.407	0.387	0.407	0.942	0.684	3.755	1.551	0.432	0.451	3.490	1.532	3.574	1.525
ETTm1	**0.380**	**0.384**	0.407	0.410	0.387	0.400	0.513	0.496	0.750	0.615	0.583	0.513	0.948	0.717	0.928	0.688
ETTm2	**0.277**	**0.318**	0.288	0.332	0.281	0.326	0.757	0.610	1.509	0.845	0.332	0.370	1.489	0.867	1.415	0.862
Weather	**0.244**	**0.265**	0.258	0.278	0.259	0.281	0.259	0.315	0.278	0.342	0.317	0.359	0.723	0.605	0.485	0.500
ECL	**0.167**	**0.256**	0.178	0.270	0.205	0.290	0.244	0.334	0.298	0.389	0.230	0.339	0.377	0.449	0.302	0.392
Exchange	**0.324**	**0.390**	0.360	0.403	0.367	0.404	0.940	0.707	1.308	0.945	0.493	0.493	1.411	0.968	1.000	0.837
Traffic	**0.423**	**0.266**	0.428	0.282	0.481	0.304	0.550	0.304	1.185	0.553	0.761	0.479	0.868	0.472	0.648	0.347
# Best	**8**	**8**	0	0	0	0	0	0	0	0	0	0	0	0	0	0

4.2 Plug-and-Play Integration Experiments

To evaluate whether WILD is genuinely plug-and-play rather than tailored to one backbone, we integrate it into seven Transformer architectures and measure before/after forecasting accuracy under the same unified setting. Figure 3 visualizes the relative gain map: across 8 datasets and 7 backbones (56 model-dataset pairs, 112 metrics), WILD yields positive gains on 105 of 112 metric cells (93.8%). The iTransformer gains are computed from the same runs summarized as WILD in Table 1.

The gain map preserves the key finding while making the integration pattern easier to inspect: WILD improves 105 of 112 metric cells across heterogeneous backbones. Large gains appear for weaker or more collapse-prone settings such as Pyraformer/Reformer on ETTh1 and Informer/Reformer on ETTh2, while the few degradations are concentrated in settings with overlapping spectral structure or sparse-key selection. This pattern supports the intended use of WILD as a representation conditioner: it is most helpful when the backbone's attention map becomes overly local or low-rank, and less

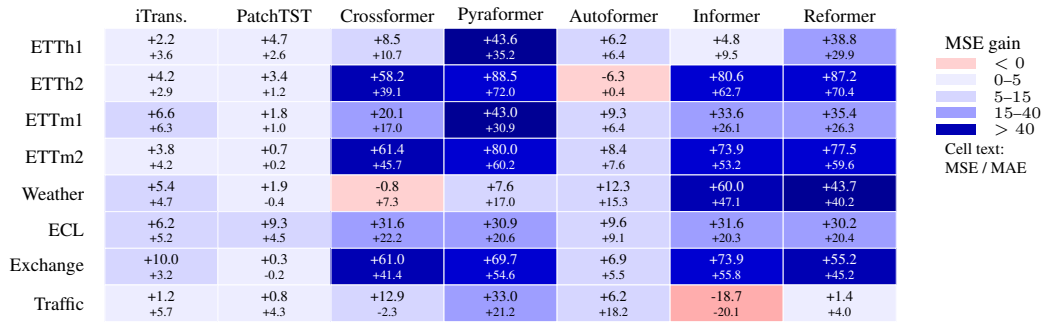


Figure 3: Plug-and-play gain map from adding WILD. Cell color encodes MSE gain; cell text reports MSE/MAE gain (%). Positive values mean lower error after adding WILD. Full Original/+WILD values are provided in Appendix Table 11.

helpful when the backbone already imposes a strong sparse or decomposition prior that partially overlaps with the spectral correction. Detailed, dataset-specific results are in Appendix C.2.

4.3 Ablation Studies

Beyond accuracy, we add targeted mechanism experiments: when the conjugate branch becomes active, how much asymmetric $W_r \neq W_i$ helps beyond symmetric complex-linear mixing, and how much FFT-domain conditioning and residual fusion contribute to the final score.

Why is the widely linear effect dataset-dependent? To understand when the asymmetric projection matters, we decompose the frequency-domain operation into its linear and conjugate components. The key algebraic identity is:

$$W_r\bar{R} + \mathrm{i}\, W_i\bar{I} = \underbrace{\frac{W_r + W_i}{2}}_{W_{\text{avg}}} X_f^{\text{pre}} \; + \; \underbrace{\frac{W_r - W_i}{2}}_{W_{\text{diff}}} \overline{X_f^{\text{pre}}},$$

where $X_f^{\text{pre}} = \bar{R} + \mathrm{i}\bar{I}$ is the preconditioned spectrum. The first term is a standard complex-linear transform; the second term couples the signal with its complex conjugate and is unique to the widely linear formulation. Table 2 reports the norms of W_{avg} and W_{diff} and the resulting conjugate energy fraction on real data. On ETT datasets ($d = 7$), the conjugate term contributes $< 0.2\%$ of output energy; on Weather ($d = 21$) and ECL ($d = 321$), it contributes 5–7%, showing that the widely linear branch is an activated secondary mechanism rather than the sole source of WILD's gains.

Table 2: Linear (W_{avg}) and conjugate (W_{diff}) decomposition. Conjugate energy is measured on test data; d is the channel dimension.

Dataset	**ETTh1**	**ETTh2**	**ETTm1**	**ETTm2**	**Weather**	**ECL**
d	7	7	7	7	21	321
$\|W_{\text{diff}}\|/\|W_{\text{avg}}\|$	0.016	0.016	0.044	0.043	0.199	0.330
Conjugate energy (%)	0.03	0.03	0.19	0.18	7.44	4.69

Widely linear ablation. To test whether the asymmetric projection ($W_r \neq W_i$) contributes beyond standard complex-linear processing, we compare WILD against **WILD$_{\text{sym}}$**, where $W_r = W_i$. In WILD$_{\text{sym}}$, the module still contains data-dependent spectral preconditioning, but the explicit conjugate term vanishes. Table 3 therefore isolates the incremental value of the widely linear branch.

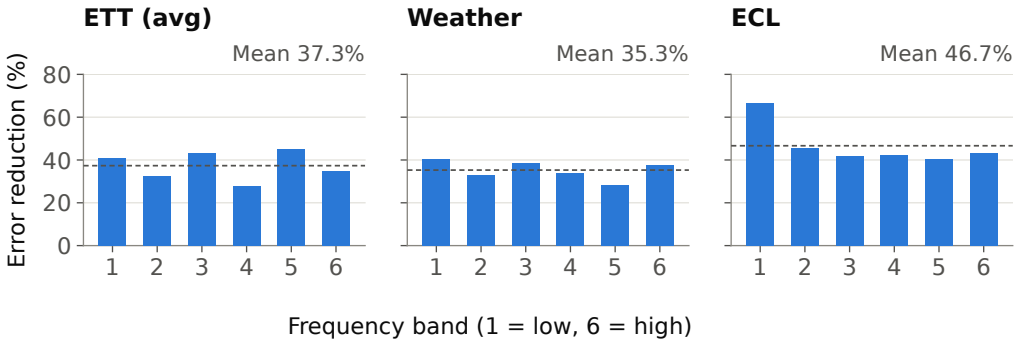


Figure 4: Error reduction over the baseline across six frequency bands, ordered from low (1) to high (6), at prediction horizon 96. ETT averages the four ETT datasets; dashed lines mark within-panel means. Gains span the spectrum rather than concentrating in one band.

Table 3: Widely linear ablation: WILD vs WILD_{sym} ($W_r = W_i$, symmetric). MSE at four prediction horizons (96, 192, 336, 720). Δ is computed from the horizon-averaged MSE values as (WILD_{sym} − WILD) / WILD. Positive Δ indicates WILD is better.

Dataset	ETTh1		ETTm1		ECL		Weather	
	WILD	WILD_{sym}	WILD	WILD_{sym}	WILD	WILD_{sym}	WILD	WILD_{sym}
96	0.386	0.391	0.348	0.339	0.139	0.139	0.169	0.170
192	0.442	0.447	0.384	0.395	0.160	0.160	0.220	0.223
336	0.488	0.486	0.438	0.442	0.171	0.175	0.277	0.280
720	0.497	0.527	0.503	0.510	0.206	0.202	0.356	0.355
Δ (avg)	**+2.0%**		**+1.0%**		**0.0%**		**+0.4%**	

Reading the result. Table 3 is a useful negative control: WILD wins in 10 of 16 displayed settings, but the average gains are modest. This matches the decomposition in Table 2. When the conjugate branch has little energy, as on low-dimensional ETT datasets, WILD behaves mostly as a spectral conditioner; when cross-variable structure is richer, the asymmetric branch becomes an additional degree of freedom. Thus the ablation supports a hierarchy rather than a single-cause story: spectral preconditioning is necessary, while widely linear mixing is activated in a dataset-dependent way. A matched five-seed H96 rerun confirms that asymmetry is not a uniform gain: WILDsym slightly lowers MSE (0.1396 vs. 0.1404; $p = 0.059$) and significantly lowers MAE ($p = 0.017$).

Component ablation. We next ablate WILD's two key components: "w/o FFT" applies the enhancement in the time domain, while "w/o Residual" keeps spectral filtering but removes the shortcut that preserves the original signal. Table 4 reports their horizon-averaged effects. Conjugate energy in Table 2 measures only the $W_r \neq W_i$ branch; "w/o FFT" removes the entire frequency-domain conditioning path, so the two quantify different mechanisms.

Table 4: Ablation study on WILD. "w/o FFT" removes frequency-domain decomposition. "w/o Residual" removes the residual connection. The WILD row uses the same horizon-averaged iTransformer+WILD results as Table 1 and Figure 3.

Method	ETTh1		ETTh2		ECL		Exchange		Traffic	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
WILD w/o FFT	0.473	0.461	0.417	0.429	0.229	0.317	0.390	0.421	0.458	0.310
WILD w/o Residual	0.455	0.447	0.397	0.417	0.196	0.283	0.368	0.409	0.429	0.288
WILD	**0.444**	**0.431**	**0.367**	**0.395**	**0.167**	**0.256**	**0.324**	**0.390**	**0.423**	**0.266**

Component hierarchy. Removing FFT is the most damaging change, especially on ECL and Exchange, where MSE rises from 0.167 to 0.229 and from 0.324 to 0.390. This matches the frequency-band diagnostic in Figure 4: WILD reduces errors across low, middle, and high bands, so

replacing spectral conditioning with time-domain mixing loses more than a cosmetic implementation detail. Removing the residual is milder but still consistently worse, indicating that WILD should condition the representation without discarding the original temporal trend. Together with Table 3, the ablations show a clear ordering: FFT-domain spectral preconditioning carries the main effect, residual fusion stabilizes the conditioned signal, and asymmetric widely linear mixing gives a secondary refinement whose contribution is dataset-dependent.

Table 5: Relative degradation of ablated variants over full WILD. Values are computed from the horizon-averaged rows in Table 4; Avg. is an unweighted mean across listed datasets. The full horizon-level ablation is reported in Appendix Table 20.

Variant	Metric	ETTh1	ETTh2	ECL	Exchange	Traffic	Avg.
w/o FFT	MSE	+6.5%	+13.6%	+37.1%	+20.4%	+8.3%	+17.2%
	MAE	+7.0%	+8.6%	+23.8%	+7.9%	+16.5%	+12.8%
w/o Residual	MSE	+2.5%	+8.2%	+17.4%	+13.6%	+1.4%	+8.6%
	MAE	+3.7%	+5.6%	+10.5%	+4.9%	+8.3%	+6.6%

Ablation takeaway. Table 5 summarizes component removal without relying on a single dataset. FFT removal is worse than residual removal on every reported dataset and on both metrics. The gap is largest on ECL and Exchange for MSE, while Traffic shows a larger MAE penalty when FFT is removed, suggesting that spectral conditioning is especially useful when multivariate coupling, non-stationarity, or heavy local fluctuations make the input spectrum poorly balanced.

4.4 Mechanism Diagnostics

We further audit whether the observed accuracy gains match the theory. Attention heatmaps show that WILD reduces local block concentration; effective-rank and top-singular-value diagnostics show that WILD generally activates a broader singular spectrum, with architecture-specific exceptions discussed in Appendix G. The decomposition in Table 2 and band-error evidence in Figure 4 connect these attention changes to the two proposed mechanisms. A frozen ECL-H96 intervention that raises spectral concentration while preserving phase, total energy, and the attention consumer increases layer-0 row redundancy and lowers stable rank for both models; WILD attenuates the sensitivity (row-redundancy slope $+0.210$ vs. $+0.315$; stable-rank slope -0.0031 vs. -0.0045; 95% bootstrap CIs in Appendix K). This is a scoped geometry result, not a mediation proof for natural-distribution accuracy. Horizon-level tables and sensitivity studies appear in Appendices C.1 and D; additional heatmaps are in Appendix G.

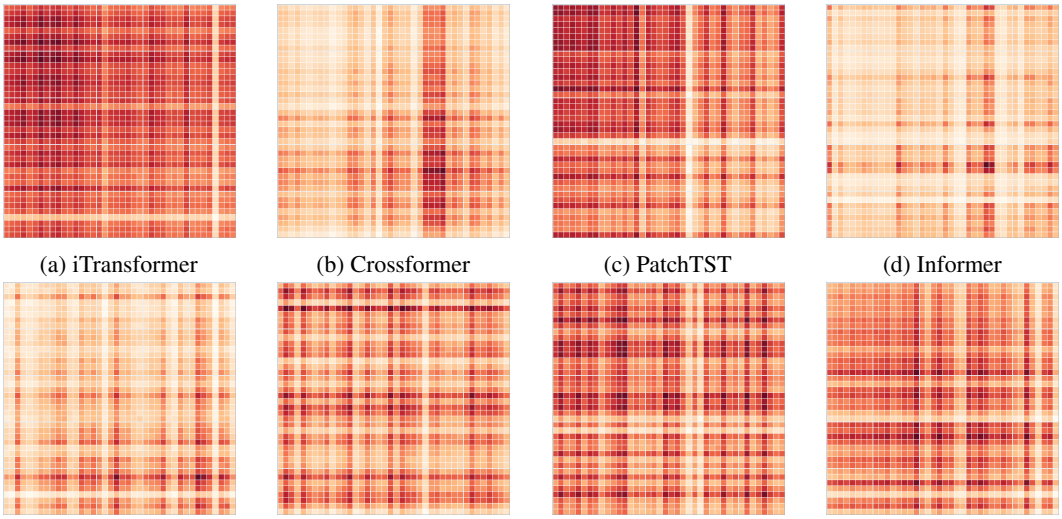
Attention structure. For iTransformer and PatchTST, the vanilla maps in Figure 5 contain repeated local blocks and WILD reduces diagonal concentration; Crossformer and Informer are shown for breadth, and we do not claim the same reduction for every architecture. Figure 6 and Table 6 quantify diagonal ratios and rank-spectrum diagnostics, respectively. A lower diagonal ratio alone does not establish better attention: the objective is to reduce redundant concentration while retaining selective dependencies. The accompanying increase in effective rank and decrease in top-5 spectral energy provide complementary evidence of a less concentrated attention spectrum.

Table 6: iTransformer attention diagnostics: effective rank (erank) and top-5 singular-value energy ratio (SVER); see Appendix G. Higher erank and lower SVER indicate less spectral concentration.

Dataset	erank Baseline	erank +WILD	Top-5 SVER Baseline	Top-5 SVER +WILD
Weather	1.87	6.42	0.9997	0.9755
ECL	2.78	45.71	0.9997	0.8966

4.5 Efficiency and Scope

WILD is designed as a conditioning layer rather than a replacement for attention. Its cost comes from rFFT/iFFT, per-frequency normalization, and two $d \times d$ channel projections, giving $\Theta(BdL \log L +$



(a) iTransformer (b) Crossformer (c) PatchTST (d) Informer

(e) iTrans.+WILD (f) Crossf.+WILD (g) PatchTST+WILD (h) Informer+WILD

Figure 5: Attention heatmaps on ETTm2, independently min–max normalized. Darker red indicates higher normalized attention; absolute values are not comparable across panels. Baselines show local blocks; WILD yields more distributed attention for iTransformer and PatchTST.

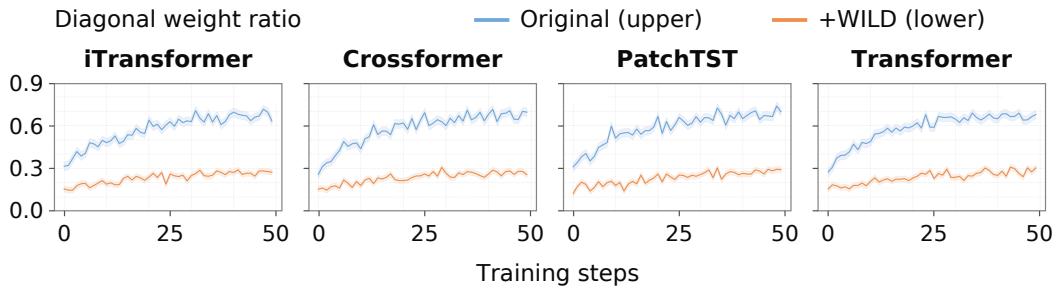


Figure 6: Diagonal attention ratio during training. WILD reduces diagonal concentration across all four backbones.

BLd^2) dominant FLOPs for batch size B, sequence length L, and channel dimension d. Table 7 shows measured forward time below 0.5ms at sequence length 96 on an RTX 4090, with negligible parameter overhead for ETT, Exchange, and Weather. For high-dimensional Traffic, a rank-32 low-rank variant reduces WILD parameters from 1.49M to 112K (≈1.7%), while rank-64 uses 222K; low rank lowers parameters and projection FLOPs but not measured end-to-end latency (Appendix E).

Qualitative forecast behavior. Appendix Figure 10 shows illustrative single-window forecasts: vanilla models flatten some peaks, valleys, and slope changes, while WILD retains more local variation. These examples are for illustration only and are not used as quantitative evidence.

Table 7: Measured efficiency of WILD (batch size 16, sequence length 96).

Dataset	d	Params	Time (ms)	Memory (MB)	Overhead
ETT (all)	7	113	0.48	8.7	$< 0.1\%$
Exchange	8	145	0.47	8.7	$< 0.1\%$
Weather	21	925	0.47	9.8	$< 0.1\%$
ECL	321	206,725	0.48	34.2	$\sim 3\%$
Traffic	862	1,487,813	0.47	80.4	$\sim 23\%$

Practical scope. WILD best benefits dense, low-rank, or block-wise attention with rich cross-channel coupling; when spectral correction overlaps sparse-key or decomposition priors, the residual scale α refines representations without disrupting them.

5 Conclusion

WILD mitigates block-wise attention collapse by conditioning representations before attention. Spectral preconditioning redistributes frequency energy, and asymmetric projections add a dataset-dependent widely linear refinement. Plug-and-play experiments show gains across diverse backbones; ablations identify spectral preconditioning as primary and residual fusion as stabilizing, while rank diagnostics and attention maps link accuracy gains to a broader singular spectrum. Qualitative examples are included for illustration only and are not used as quantitative evidence; overall, WILD upgrades existing forecasters without architectural changes.

Limitations Evaluation covers public benchmarks and Transformer backbones only. Matched five-seed checks expose a negative boundary on ETTh1-H96 (MSE 1/5 seeds, $p = 0.209$; MAE $p = 0.011$) and for a frequency-specialized FITS backbone. Gains are therefore conditional on a correctable redundancy pattern; efficiency and ablation analyses concern the studied configurations, and broader characterization remains future work.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers for their constructive feedback. This work is supported by the National Science Foundation of China (62476289).

References

Peng Chen, Yingying Zhang, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang, and Chenjuan Guo. Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024.

Robert M. Gray. Toeplitz and circulant matrices: A review. *Foundations and Trends in Communications and Information Theory*, 2(3):155–239, 2006. doi: 10.1561/0100000006. URL `https://www.nowpublishers.com/article/Details/CIT-006`.

John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Efficient token mixing for transformers via adaptive fourier neural operators. In *International Conference on Learning Representations*, 2022.

Georges Hebrail and Alice Berard. Individual Household Electric Power Consumption. UCI Machine Learning Repository, 2012.

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*, 2021.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

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

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. Fnet: Mixing tokens with fourier transforms. *arXiv preprint arXiv:2105.03824*, 2021.

Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X. Liu, and Schahram Dustdar. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International Conference on Learning Representations*, 2022a.

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

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

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Xihao Piao, Zheng Chen, Taichi Murayama, Yasuko Matsubara, and Yasushi Sakurai. Fredformer: Frequency debiased transformer for time series forecasting. *arXiv preprint arXiv:2406.09009*, 2024.

Hao Wang, Licheng Pan, Zhichao Chen, Degui Yang, Sen Zhang, Yifei Yang, Xinggao Liu, Haoxuan Li, and Dacheng Tao. Fredf: Learning to forecast in the frequency domain. *arXiv preprint arXiv:2402.02399*, 2025.

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. *arXiv preprint arXiv:2410.16032*, 2024.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 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.

Zhijian Xu, Ailing Zeng, and Qiang Xu. FITS: Modeling time series with 10k parameters. In *The Twelfth International Conference on Learning Representations*, 2024.

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, pages 11121–11128, 2023.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The Eleventh International Conference on Learning Representations*, 2023.

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, pages 11106–11115, 2021.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*, pages 27268–27286. PMLR, 2022.

A Detailed Experimental Settings

A.1 Datasets

Datasets. We conducted extensive experiments on eight widely-used benchmark datasets to evaluate the performance of baseline models and the proposed WILD framework. The ETT dataset [Zhou et al., 2021] contains seven features related to electricity transformer temperatures, sampled at different frequencies. It is divided into four subsets: ETTh1 and ETTh2 are recorded hourly, while ETTm1 and ETTm2 are recorded every 15 minutes, covering the period from July 2016 to July 2018. The Weather dataset [Lai et al., 2018] includes 21 meteorological variables collected every 10 minutes in 2020 from the Max Planck Institute for Biogeochemistry. The Exchange dataset [Lai et al., 2018] consists of daily exchange rates from eight countries spanning from 1990 to 2016. The ECL (Electricity) dataset [Lai et al., 2018] records hourly electricity consumption from 321 clients. Lastly, the Traffic dataset [Lai et al., 2018] captures hourly road occupancy rates from 862 sensors located on freeways in the San Francisco Bay Area, covering January 2015 to December 2016. These datasets span diverse domains such as energy systems, finance, meteorology, and transportation, posing a variety of forecasting challenges—particularly in long-horizon settings. Detailed statistics are summarized in Table 8.

Dataset Split. We follow the data processing procedures and chronological data split strategy adopted in TimesNet [Wu et al., 2023]. For all datasets, the training, validation, and test sets are partitioned sequentially without shuffling to preserve temporal dependencies. Specifically, we apply a (0.7, 0.1, 0.2) split ratio for Weather, Traffic, ECL, and Exchange datasets, and a (0.6, 0.2, 0.2) split ratio for the ETT subsets (ETTh1, ETTh2, ETTm1, and ETTm2), consistent with prior literature [Zhou et al., 2021, Zeng et al., 2023]. This setup ensures a fair and standardized evaluation protocol, especially for long-term forecasting benchmarks.

Forecasting setting. We standardize the experimental configurations across all models to ensure a fair comparison within a unified framework. Specifically, the input sequence length is fixed to 96, while the prediction horizons vary among {96, 192, 336, 720}.

Table 8: Dataset summary.

Dataset	Dim	Time Points	Frequency	Domain
ETTh1	7	17,420	1-hour	Electricity
ETTh2	7	17,420	1-hour	Electricity
ETTm1	7	69,680	15-min	Electricity
ETTm2	7	69,680	15-min	Electricity
ECL	321	26,304	1-hour	Electricity
Traffic	862	17,544	1-hour	Transportation
Weather	21	52,696	10-min	Weather
Exchange	8	7,588	1-day	Financial

A.2 Experiment Setting

We standardized the parameters across all models to ensure a fair comparison within a unified platform (Time-Series-Library). Specifically, in main experiment, the batch size (`batch_size`) was set to 16. To prevent overfitting, a dropout rate of 0.1 was applied. The number of training epochs was fixed at 30. For temporal encoding, we adopted the `timeF` embedding method. Additional parameter details are provided in Table 9. Experiments are implemented in PyTorch [Paszke et al., 2019] and executed on a single NVIDIA RTX 4090 GPU (24GB). In the plug-and-play experiment, we adopt the experimental settings as defined in the original papers of the seven baseline models.

Table 9: Experiment configuration of WILD.

Dataset	ETTh1	ETTh2	ETTm1	ETTm2	Weather	ECL	Exchange	Traffic
seq_len	96	96	96	96	96	96	96	96
label_len	48	48	48	48	96	48	96	48
e_layers	2	2	2	2	2	2	2	2
d_layers	1	1	1	1	1	1	1	1
factor	3	3	1	1	3	3	3	3
enc_in	7	7	7	7	21	321	8	862
dec_in	7	7	7	7	21	321	8	862
c_out	7	7	7	7	21	321	8	862
d_model	512	512	512	512	32	512	64	512
d_ff	–	16	–	–	32	–	64	–
n_heads	8	8	8	8	8	8	8	8
lr	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4	1e-3

B Algorithm of WILD

Algorithm 1 WILD: Widely Linear Frequency-Domain Conditioning Module

Input: Input window $\mathbf{X} \in \mathbb{R}^{L\times d}$, dropout rate $p \in [0, 0.2]$
Output: Enhanced representation $\mathbf{Z} \in \mathbb{R}^{L\times d}$
Parameters: $\mathbf{W}_r, \mathbf{W}_i \in \mathbb{R}^{d\times d}$, $\mathbf{b}_r, \mathbf{b}_i \in \mathbb{R}^d$, $\alpha \in \mathbb{R}$

// Pre-normalization (stationarization)
De-mean and z-normalize each feature of $\mathbf{X}$ within the current window

// Step 1: Unitary real FFT along time axis
$\widehat{\mathbf{X}} \leftarrow \mathrm{rFFT}(\mathbf{X}^\top)$ ▷ $\widehat{\mathbf{X}} \in \mathbb{C}^{d\times K},\ K = \lfloor L/2 \rfloor + 1$
$\mathbf{R} \leftarrow \mathrm{Re}(\widehat{\mathbf{X}}), \mathbf{I} \leftarrow \mathrm{Im}(\widehat{\mathbf{X}})$ ▷ $\mathbf{R}, \mathbf{I} \in \mathbb{R}^{d\times K}$

// Step 2: Spectral preconditioning (per-frequency across channels)
for $k = 1$ **to** K **do**
  $e_k \leftarrow \sqrt{\sum_{j=1}^{d} (\mathbf{R}_{j,k}^2 + \mathbf{I}_{j,k}^2)} + \varepsilon$
  $\bar{\mathbf{R}}_{:,k} \leftarrow \mathbf{R}_{:,k} \,/\, e_k$
  $\bar{\mathbf{I}}_{:,k} \leftarrow \mathbf{I}_{:,k} \,/\, e_k$
end for

// Step 3: Asymmetric channel mixing ($\mathbf{W}_r \neq \mathbf{W}_i$, widely linear)
$\mathbf{Y}^R \leftarrow \mathbf{W}_r \cdot \bar{\mathbf{R}} + \mathbf{b}_r$ ▷ $\mathbf{Y}^R \in \mathbb{R}^{d\times K}$
$\mathbf{Y}^I \leftarrow \mathbf{W}_i \cdot \bar{\mathbf{I}} + \mathbf{b}_i$ ▷ $\mathbf{Y}^I \in \mathbb{R}^{d\times K}$

// Step 4: Inverse FFT
$\widetilde{\mathbf{X}} \leftarrow \mathrm{irFFT}\big((\mathbf{Y}^R + \mathrm{i} \cdot \mathbf{Y}^I)^\top\big)$ ▷ $\widetilde{\mathbf{X}} \in \mathbb{R}^{L\times d}$

// Step 5: Residual fusion with dropout
$\mathbf{Z} \leftarrow \mathbf{X} + \alpha \cdot \mathrm{Dropout}(\widetilde{\mathbf{X}},\ p)$
return Z

C FULL RESULTS

C.1 Main Results

We evaluate the forecasting performance of our proposed method across eight widely used multivariate time series datasets: ETTh1, ETTh2, ETTm1, ETTm2, Electricity, Exchange, Traffic, and Weather. The evaluation is conducted under four different prediction horizons (96, 192, 336, and 720), using Mean Squared Error (MSE) and Mean Absolute Error (MAE) as performance metrics. To assess the effectiveness of the proposed WILD module, we adopt iTransformer as the default backbone and

integrate WILD into it. The resulting model is then compared with a range of evaluated Transformer baselines.

Table 10: Forecasting performance (MSE and MAE) on eight datasets under different prediction horizons. The average row corresponds to the WILD column in Table 1 and to the iTransformer+WILD averages used for Figure 3.

Dataset	**ETTh1**		**ETTh2**		**ETTm1**		**ETTm2**		**ECL**		**Exchange**		**Traffic**		**Weather**	
metric	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**
96	0.383	0.393	0.286	0.338	0.310	0.341	0.174	0.252	0.139	0.228	0.084	0.204	0.390	0.251	0.158	0.195
192	0.434	0.423	0.360	0.386	0.363	0.370	0.239	0.295	0.157	0.245	0.170	0.297	0.416	0.259	0.208	0.241
336	0.471	0.441	0.401	0.420	0.395	0.392	0.299	0.334	0.170	0.259	0.299	0.403	0.425	0.269	0.267	0.286
720	0.487	0.466	0.420	0.438	0.453	0.431	0.395	0.390	0.202	0.289	0.744	0.656	0.460	0.286	0.345	0.337
Avg.	0.444	0.431	0.367	0.395	0.380	0.384	0.277	0.318	0.167	0.256	0.324	0.390	0.423	0.266	0.244	0.265

C.2 Plug-And-Play

We apply WILD to Transformer variants such as Reformer [Kitaev et al., 2020], Informer [Zhou et al., 2021], and iTransformer [Liu et al., 2024]. Table 11 summarizes the results; Tables 12–19 provide the full horizon-level comparisons.

Table 11: Mean plug-and-play comparison before and after integrating WILD. This table reports the Original, +WILD, and relative gain values summarized in Figure 3; metrics are horizon-averaged MSE/MAE.

	Models / **Ours**	iTransformer ICLR 2024		PatchTST ICLR 2023		Crossformer ICLR 2023		Pyraformer ICLR 2023		Autoformer NeurIPS 2022		Informer AAAI 2021		Reformer ICLR 2020	
	Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	Original	0.454	0.447	0.469	0.454	0.529	0.522	0.865	0.731	0.518	0.500	1.078	0.813	0.961	0.757
	+WILD	0.444	0.431	0.447	0.443	0.484	0.466	0.488	0.474	0.486	0.468	1.026	0.736	0.588	0.530
	Gain	+2.2%	+3.6%	+4.7%	+2.6%	+8.5%	+10.7%	+43.6%	+35.2%	+6.2%	+6.4%	+4.8%	+9.5%	+38.8%	+29.9%
ETTh2	Original	0.383	0.407	0.387	0.407	0.942	0.684	3.755	1.551	0.432	0.451	3.490	1.532	3.574	1.525
	+WILD	0.367	0.395	0.374	0.402	0.394	0.416	0.433	0.434	0.459	0.449	0.677	0.571	0.459	0.451
	Gain	+4.2%	+2.9%	+3.4%	+1.2%	+58.2%	+39.1%	+88.5%	+72.0%	-6.3%	+0.4%	+80.6%	+62.7%	+87.2%	+70.4%
ETTm1	Original	0.407	0.410	0.387	0.400	0.513	0.496	0.750	0.615	0.583	0.513	0.948	0.717	0.928	0.688
	+WILD	0.380	0.384	0.380	0.396	0.410	0.411	0.427	0.425	0.529	0.480	0.629	0.530	0.599	0.507
	Gain	+6.6%	+6.3%	+1.8%	+1.0%	+20.1%	+17.0%	+43.0%	+30.9%	+9.3%	+6.4%	+33.6%	+26.1%	+35.4%	+26.3%
ETTm2	Original	0.288	0.332	0.281	0.326	0.757	0.610	1.509	0.845	0.332	0.370	1.489	0.867	1.415	0.862
	+WILD	0.277	0.318	0.279	0.326	0.292	0.332	0.302	0.336	0.304	0.342	0.389	0.406	0.318	0.348
	Gain	+3.8%	+4.2%	+0.7%	+0.2%	+61.4%	+45.7%	+80.0%	+60.2%	+8.4%	+7.6%	+73.9%	+53.2%	+77.5%	+59.6%
Weather	Original	0.258	0.278	0.259	0.281	0.259	0.315	0.278	0.342	0.317	0.359	0.723	0.605	0.485	0.500
	+WILD	0.244	0.265	0.254	0.281	0.261	0.292	0.257	0.284	0.278	0.304	0.289	0.320	0.273	0.299
	Gain	+5.4%	+4.7%	+1.9%	-0.4%	-0.8%	+7.3%	+7.6%	+17.0%	+12.3%	+15.3%	+60.0%	+47.1%	+43.7%	+40.2%
ECL	Original	0.178	0.270	0.205	0.290	0.244	0.334	0.298	0.389	0.230	0.339	0.377	0.449	0.302	0.392
	+WILD	0.167	0.256	0.186	0.277	0.167	0.260	0.206	0.309	0.208	0.308	0.258	0.358	0.210	0.312
	Gain	+6.2%	+5.2%	+9.3%	+4.5%	+31.6%	+22.2%	+30.9%	+20.6%	+9.6%	+9.1%	+31.6%	+20.3%	+30.2%	+20.4%
Exchange	Original	0.360	0.403	0.367	0.404	0.940	0.707	1.308	0.945	0.493	0.493	1.411	0.968	1.000	0.837
	+WILD	0.324	0.390	0.365	0.405	0.367	0.414	0.396	0.429	0.459	0.466	0.368	0.428	0.448	0.459
	Gain	+10.0%	+3.2%	+0.3%	-0.2%	+61.0%	+41.4%	+69.7%	+54.6%	+6.9%	+5.5%	+73.9%	+55.8%	+55.2%	+45.2%
Traffic	Original	0.428	0.282	0.481	0.304	0.550	0.304	1.185	0.553	0.761	0.479	0.868	0.472	0.648	0.347
	+WILD	0.423	0.266	0.477	0.291	0.479	0.311	0.794	0.436	0.713	0.392	1.030	0.567	0.638	0.333
	Gain	+1.2%	+5.7%	+0.8%	+4.3%	+12.9%	-2.3%	+33.0%	+21.2%	+6.2%	+18.2%	-18.7%	-20.1%	+1.4%	+4.0%

Detailed horizon-level results are split by dataset in Tables 12–19. Averaging the four horizons recovers the corresponding Original/+WILD entries in Table 11; the iTransformer+WILD averages also recover the WILD column in Table 1.

Table 12: Detailed plug-and-play results on ETTh1. Bold and underline mark the lowest and second-lowest distinct displayed errors across all fourteen variants at each horizon; ties share the same mark. Reported gain values are listed by horizon.

Backbone	**Horizon 96**				**Horizon 192**			
	Original		+WILD		Original		+WILD	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
iTransformer	<u>0.386</u>	0.405	**0.383**	**0.393**	0.441	0.436	**0.434**	**0.423**
PatchTST	0.414	0.419	0.389	<u>0.404</u>	0.460	0.445	<u>0.439</u>	<u>0.435</u>
Crossformer	0.423	0.448	0.416	0.426	0.471	0.474	0.478	0.461
Pyraformer	0.790	0.696	0.424	0.432	0.738	0.662	0.478	0.464
Autoformer	0.485	0.468	0.418	0.433	0.554	0.517	0.526	0.478
Informer	1.006	0.780	0.848	0.641	0.927	0.721	0.977	0.718
Reformer	0.826	0.687	0.554	0.505	0.906	0.732	0.555	0.517
Reported gain (%)	—		16.97%	14.35%	—		11.48%	10.74%
Backbone	**Horizon 336**				**Horizon 720**			
	Original		+WILD		Original		+WILD	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
iTransformer	0.487	0.458	**0.471**	**0.441**	0.503	0.491	<u>0.487</u>	**0.466**
PatchTST	0.501	0.466	<u>0.474</u>	<u>0.453</u>	0.500	0.488	**0.484**	<u>0.480</u>
Crossformer	0.570	0.546	0.499	0.469	0.653	0.621	0.543	0.510
Pyraformer	0.922	0.766	0.509	0.485	1.011	0.799	0.539	0.514
Autoformer	0.503	0.491	0.491	0.469	0.531	0.523	0.509	0.491
Informer	1.197	0.878	1.058	0.765	1.182	0.873	1.220	0.819
Reformer	1.022	0.786	0.600	0.530	1.091	0.821	0.642	0.570
Reported gain (%)	—		17.46%	15.25%	—		16.34%	14.59%

Table 13: Detailed plug-and-play results on ETTh2. Bold and underline mark the lowest and second-lowest distinct displayed errors across all fourteen variants at each horizon; ties share the same mark. Reported gain values are listed by horizon.

Backbone	**Horizon 96**				**Horizon 192**			
	Original		+WILD		Original		+WILD	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
iTransformer	0.297	0.349	**0.286**	**0.338**	0.380	0.400	<u>0.360</u>	**0.386**
PatchTST	0.302	0.348	<u>0.288</u>	<u>0.341</u>	0.388	0.400	**0.357**	<u>0.388</u>
Crossformer	0.745	0.584	0.320	0.361	0.877	0.656	0.377	0.402
Pyraformer	1.324	0.870	0.328	0.372	4.264	1.608	0.422	0.420
Autoformer	0.350	0.397	0.365	0.389	0.430	0.440	0.447	0.440
Informer	2.535	1.272	0.440	0.458	3.557	1.522	0.787	0.608
Reformer	2.410	1.234	0.379	0.398	4.812	1.930	0.472	0.449
Reported gain (%)	—		43.18%	33.44%	—		45.90%	36.41%
Backbone	**Horizon 336**				**Horizon 720**			
	Original		+WILD		Original		+WILD	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
iTransformer	0.428	<u>0.432</u>	**0.401**	**0.420**	0.427	<u>0.445</u>	**0.420**	**0.438**
PatchTST	0.426	0.433	<u>0.425</u>	0.434	0.431	0.446	<u>0.426</u>	0.446
Crossformer	1.043	0.731	0.438	0.440	1.104	0.763	0.439	0.461
Pyraformer	5.218	1.944	0.486	0.462	4.213	1.780	0.498	0.483
Autoformer	0.463	0.471	0.497	0.475	0.484	0.497	0.528	0.492
Informer	4.388	1.725	0.734	0.604	3.478	1.609	0.747	0.614
Reformer	4.245	1.638	0.484	0.467	2.828	1.296	0.499	0.488
Reported gain (%)	—		45.28%	36.15%	—		43.45%	34.14%

Table 14: Detailed plug-and-play results on ETTm1. Bold and underline mark the lowest and second-lowest distinct displayed errors across all fourteen variants at each horizon; ties share the same mark. Reported gain values are listed by horizon.

Backbone	**Horizon 96**				**Horizon 192**			
	Original		+WILD		Original		+WILD	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
iTransformer	0.334	0.368	**0.310**	**0.341**	0.377	0.391	0.373	**0.363**
PatchTST	<u>0.329</u>	0.367	<u>0.329</u>	<u>0.366</u>	<u>0.370</u>	0.385	**0.356**	<u>0.380</u>
Crossformer	0.404	0.426	0.334	0.373	0.450	0.451	0.376	0.390
Pyraformer	0.609	0.540	0.371	0.391	0.606	0.537	0.422	0.418
Autoformer	0.519	0.490	0.475	0.444	0.567	0.503	0.545	0.492
Informer	0.765	0.655	0.427	0.427	0.696	0.586	0.554	0.493
Reformer	0.749	0.575	0.479	0.450	0.879	0.668	0.551	0.481
Reported gain (%)	—		20.82%	15.29%	—		16.07%	12.15%
Backbone	**Horizon 336**				**Horizon 720**			
	Original		+WILD		Original		+WILD	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
iTransformer	0.426	0.420	<u>0.395</u>	**0.392**	0.491	0.459	<u>0.453</u>	**0.431**
PatchTST	0.399	0.410	**0.390**	<u>0.403</u>	0.454	0.439	**0.444**	<u>0.435</u>
Crossformer	0.532	0.515	0.442	0.424	0.666	0.589	0.487	0.455
Pyraformer	0.864	0.666	0.428	0.430	0.919	0.718	0.486	0.463
Autoformer	0.558	0.507	0.539	0.496	0.689	0.551	0.557	0.488
Informer	1.098	0.799	0.620	0.525	1.234	0.827	0.914	0.675
Reformer	1.017	0.736	0.640	0.528	1.065	0.774	0.728	0.569
Reported gain (%)	—		22.62%	17.55%	—		22.34%	16.93%

Table 15: Detailed plug-and-play results on ETTm2. Bold and underline mark the lowest and second-lowest distinct displayed errors across all fourteen variants at each horizon; ties share the same mark. Reported gain values are listed by horizon.

Backbone	**Horizon 96**				**Horizon 192**			
	Original		+WILD		Original		+WILD	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
iTransformer	0.180	0.264	**0.174**	**0.252**	0.250	0.309	**0.240**	**0.295**
PatchTST	<u>0.175</u>	<u>0.259</u>	0.176	0.260	<u>0.241</u>	<u>0.302</u>	0.248	0.307
Crossformer	0.287	0.366	0.176	0.261	0.414	0.492	0.261	0.315
Pyraformer	0.415	0.480	0.194	0.272	0.647	0.609	0.260	0.312
Autoformer	0.234	0.313	0.198	0.279	0.292	0.349	0.263	0.317
Informer	0.487	0.551	0.204	0.295	0.568	0.592	0.314	0.376
Reformer	0.553	0.564	0.199	0.276	1.215	0.833	0.284	0.327
Reported gain (%)	—		33.17%	26.10%	—		32.73%	27.49%
Backbone	**Horizon 336**				**Horizon 720**			
	Original		+WILD		Original		+WILD	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
iTransformer	0.311	0.348	**0.299**	**0.334**	0.412	0.407	<u>0.395</u>	**0.390**
PatchTST	0.305	0.343	<u>0.300</u>	<u>0.340</u>	0.402	0.400	**0.394**	<u>0.395</u>
Crossformer	0.597	0.542	0.306	0.345	1.730	1.042	0.424	0.406
Pyraformer	1.297	0.881	0.319	0.348	3.676	1.411	0.435	0.413
Autoformer	0.337	0.375	0.333	0.360	0.464	0.443	0.423	0.412
Informer	1.567	0.970	0.400	0.427	3.333	1.353	0.638	0.527
Reformer	1.221	0.843	0.334	0.365	2.670	1.206	0.454	0.423
Reported gain (%)	—		39.34%	30.83%	—		48.58%	38.28%

Table 16: Detailed plug-and-play results on Weather. Bold and underline mark the lowest and second-lowest distinct displayed errors across all fourteen variants at each horizon; ties share the same mark. Reported gain values are listed by horizon.

Backbone	**Horizon 96**				**Horizon 192**			
	Original		+WILD		Original		+WILD	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
iTransformer	0.174	<u>0.214</u>	**0.158**	**0.195**	0.221	<u>0.254</u>	<u>0.208</u>	**0.241**
PatchTST	0.177	0.218	0.185	0.231	0.225	0.259	0.213	0.255
Crossformer	**0.158**	0.230	<u>0.172</u>	0.221	**0.206**	0.277	0.231	0.275
Pyraformer	0.210	0.294	<u>0.172</u>	0.222	0.240	0.314	0.223	0.263
Autoformer	0.235	0.304	0.195	0.243	0.301	0.354	0.253	0.293
Informer	0.333	0.393	0.204	0.258	0.858	0.652	0.267	0.306
Reformer	0.354	0.424	0.194	0.244	0.314	0.384	0.242	0.280
Reported gain (%)	—		16.25%	17.31%	—		15.89%	16.44%
Backbone	**Horizon 336**				**Horizon 720**			
	Original		+WILD		Original		+WILD	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
iTransformer	0.278	0.296	**0.267**	**0.286**	0.358	0.347	**0.349**	**0.337**
PatchTST	0.278	0.297	<u>0.269</u>	<u>0.295</u>	<u>0.354</u>	0.348	**0.349**	<u>0.345</u>
Crossformer	0.272	0.335	0.283	0.312	0.398	0.418	0.358	0.359
Pyraformer	0.295	0.356	0.278	0.301	0.367	0.404	0.355	0.351
Autoformer	0.329	0.367	0.299	0.321	0.403	0.410	0.364	0.359
Informer	0.664	0.583	0.317	0.340	1.035	0.793	0.370	0.374
Reformer	0.590	0.566	0.293	0.314	0.681	0.627	0.364	0.358
Reported gain (%)	—		17.13%	17.53%	—		19.52%	19.39%

Table 17: Detailed plug-and-play results on ECL. Bold and underline mark the lowest and second-lowest distinct displayed errors across all fourteen variants at each horizon; ties share the same mark. Reported gain values are listed by horizon.

Backbone	**Horizon 96**				**Horizon 192**			
	Original		+WILD		Original		+WILD	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
iTransformer	0.148	0.240	<u>0.139</u>	**0.228**	0.162	0.253	<u>0.157</u>	**0.245**
PatchTST	0.181	0.270	0.157	0.248	0.188	0.274	0.169	0.262
Crossformer	0.219	0.314	**0.136**	<u>0.230</u>	0.231	0.322	**0.156**	<u>0.248</u>
Pyraformer	0.285	0.377	0.178	0.285	0.289	0.384	0.197	0.301
Autoformer	0.196	0.313	0.173	0.279	0.220	0.330	0.211	0.308
Informer	0.340	0.420	0.219	0.325	0.378	0.452	0.241	0.346
Reformer	0.274	0.372	0.190	0.294	0.293	0.389	0.200	0.303
Reported gain (%)	—		24.78%	16.61%	—		21.19%	14.46%
Backbone	**Horizon 336**				**Horizon 720**			
	Original		+WILD		Original		+WILD	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
iTransformer	0.178	0.269	**0.170**	**0.259**	0.225	0.317	**0.202**	**0.289**
PatchTST	0.204	0.293	0.188	0.281	0.246	0.324	0.232	0.316
Crossformer	0.246	0.337	<u>0.172</u>	<u>0.266</u>	0.280	0.363	<u>0.203</u>	<u>0.295</u>
Pyraformer	0.308	0.400	0.212	0.315	0.309	0.396	0.239	0.334
Autoformer	0.246	0.350	0.210	0.312	0.259	0.362	0.238	0.333
Informer	0.387	0.460	0.271	0.372	0.403	0.464	0.301	0.390
Reformer	0.321	0.408	0.217	0.318	0.318	0.400	0.235	0.333
Reported gain (%)	—		21.19%	14.12%	—		16.67%	11.44%

Table 18: Detailed plug-and-play results on Exchange. Bold and underline mark the lowest and second-lowest distinct displayed errors across all fourteen variants at each horizon; ties share the same mark. Reported gain values are listed by horizon.

Backbone	**Horizon 96**				**Horizon 192**			
	Original		+WILD		Original		+WILD	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
iTransformer	<u>0.086</u>	0.206	**0.084**	**0.204**	0.177	0.299	**0.170**	**0.297**
PatchTST	0.088	<u>0.205</u>	<u>0.086</u>	0.206	<u>0.176</u>	0.299	<u>0.176</u>	<u>0.298</u>
Crossformer	0.256	0.367	0.088	0.211	0.470	0.509	0.184	0.307
Pyraformer	0.898	0.798	0.106	0.234	1.065	0.870	0.205	0.326
Autoformer	0.158	0.288	0.122	0.249	0.268	0.379	0.234	0.353
Informer	0.974	0.794	0.122	0.250	1.314	0.942	0.215	0.339
Reformer	0.820	0.758	0.121	0.247	0.900	0.782	0.237	0.349
Reported gain (%)	—		50.06%	37.38%	—		44.35%	32.45%
Backbone	**Horizon 336**				**Horizon 720**			
	Original		+WILD		Original		+WILD	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
iTransformer	0.331	0.417	**0.299**	<u>0.403</u>	0.847	0.691	**0.744**	**0.656**
PatchTST	<u>0.301</u>	**0.397**	0.331	0.417	0.901	0.714	0.868	0.701
Crossformer	1.268	0.883	0.367	0.443	1.767	1.068	0.829	0.694
Pyraformer	1.287	0.957	0.353	0.431	1.980	1.155	0.922	0.726
Autoformer	0.434	0.487	0.416	0.473	1.111	0.818	1.063	0.787
Informer	1.403	0.988	0.368	0.450	1.954	1.146	<u>0.768</u>	<u>0.675</u>
Reformer	1.080	0.877	0.403	0.464	1.199	0.929	1.029	0.775
Reported gain (%)	—		39.48%	29.20%	—		27.75%	19.54%

Table 19: Detailed plug-and-play results on Traffic. Bold and underline mark the lowest and second-lowest distinct displayed errors across all fourteen variants at each horizon; ties share the same mark. Reported gain values are listed by horizon.

Backbone	**Horizon 96**				**Horizon 192**			
	Original		+WILD		Original		+WILD	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
iTransformer	<u>0.395</u>	<u>0.268</u>	**0.390**	**0.251**	<u>0.417</u>	<u>0.276</u>	**0.416**	**0.259**
PatchTST	0.462	0.295	0.457	0.277	0.466	0.296	0.466	0.278
Crossformer	0.522	0.290	0.488	0.337	0.530	0.293	0.445	0.278
Pyraformer	0.839	0.466	0.844	0.471	1.086	0.602	0.701	0.384
Autoformer	0.613	0.385	0.680	0.372	0.953	0.612	0.700	0.396
Informer	0.877	0.475	0.973	0.517	0.791	0.431	0.924	0.516
Reformer	0.641	0.344	0.617	0.321	0.638	0.344	0.633	0.340
Reported gain (%)	—		-1.63%	-1.37%	—		7.93%	8.49%
Backbone	**Horizon 336**				**Horizon 720**			
	Original		+WILD		Original		+WILD	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
iTransformer	<u>0.433</u>	<u>0.283</u>	**0.425**	**0.269**	<u>0.467</u>	<u>0.302</u>	**0.460**	**0.286**
PatchTST	0.482	0.304	0.482	0.293	0.514	0.322	0.501	0.314
Crossformer	0.558	0.305	0.469	0.305	0.589	0.328	0.515	0.325
Pyraformer	0.850	0.470	0.843	0.460	1.244	0.675	0.789	0.430
Autoformer	0.691	0.432	0.699	0.379	0.785	0.488	0.773	0.422
Informer	0.921	0.495	1.112	0.622	0.884	0.487	1.112	0.612
Reformer	0.636	0.340	0.636	0.326	0.675	0.359	0.665	0.343
Reported gain (%)	—		-1.13%	-0.65%	—		3.70%	4.67%

C.3 Ablation Study

To validate the effectiveness of the WILD module, we conduct a comprehensive ablation study. Specifically, we examine two reduced variants: (1) WILD w/o FFT, which removes the frequency-domain decomposition and performs rank enhancement directly in the time domain; and (2) WILD w/o Residual, which retains frequency-domain operations but excludes the residual connection.

As shown in Table 20, both ablated versions consistently underperform the full WILD model across all datasets. This suggests that both frequency-domain processing and the residual pathway contribute meaningfully to the final performance. In particular, removing the FFT module leads to more significant degradation, highlighting the benefit of decorrelating temporal patterns via spectral decomposition.

Table 20: Ablation study on WILD without FFT and without Residual connection across five datasets. Metrics: MSE and MAE.

Dataset	Horizon	WILD_w/o_FFT		WILD_w/o_Residual		WILD	
		MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	0.406	0.422	0.389	0.408	**0.383**	**0.393**
	192	0.465	0.455	0.438	0.435	**0.434**	**0.423**
	336	0.493	0.466	0.483	0.455	**0.471**	**0.441**
	720	0.530	0.501	0.510	0.491	**0.487**	**0.466**
ETTh2	96	0.314	0.361	0.341	0.378	**0.286**	**0.338**
	192	0.401	0.413	0.414	0.421	**0.360**	**0.386**
	336	0.437	0.442	0.465	0.457	**0.401**	**0.420**
	720	0.435	0.451	0.448	0.460	**0.420**	**0.438**
ECL	96	0.169	0.259	0.204	0.296	**0.139**	**0.228**
	192	0.180	0.269	0.209	0.301	**0.157**	**0.245**
	336	0.197	0.286	0.230	0.320	**0.170**	**0.259**
	720	0.237	0.319	0.272	0.352	**0.202**	**0.289**
Exchange	96	0.090	0.212	0.091	0.214	**0.084**	**0.204**
	192	0.181	0.304	0.185	0.307	**0.170**	**0.297**
	336	0.330	0.417	0.339	0.422	**0.299**	**0.403**
	720	0.871	0.705	0.945	0.741	**0.744**	**0.656**
Traffic	96	0.400	0.274	0.428	0.299	**0.390**	**0.251**
	192	0.421	0.283	0.444	0.302	**0.416**	**0.259**
	336	0.427	0.287	0.462	0.310	**0.425**	**0.269**
	720	0.469	0.307	0.496	0.329	**0.460**	**0.286**

D Hyperparameter Sensitivity

To evaluate the robustness and adaptability of the proposed WILD framework under different hyperparameter configurations, we conduct a comprehensive sensitivity analysis in Figure 7. Specifically, we vary four key hyperparameters—learning rate (lr), batch size, number of encoder layers (e_layers), and model dimension (d_model)—across four widely-used multivariate time series datasets: ETTh1, ETTm1, ECL, and Exchange. The Mean Squared Error (MSE) is adopted as the evaluation metric.

E Time Complexity

WILD is designed as a lightweight plugin. Its computational cost is dominated by the FFT/iFFT and the channel mixing linear layers. The total overhead is modest relative to the self-attention mechanism it precedes.

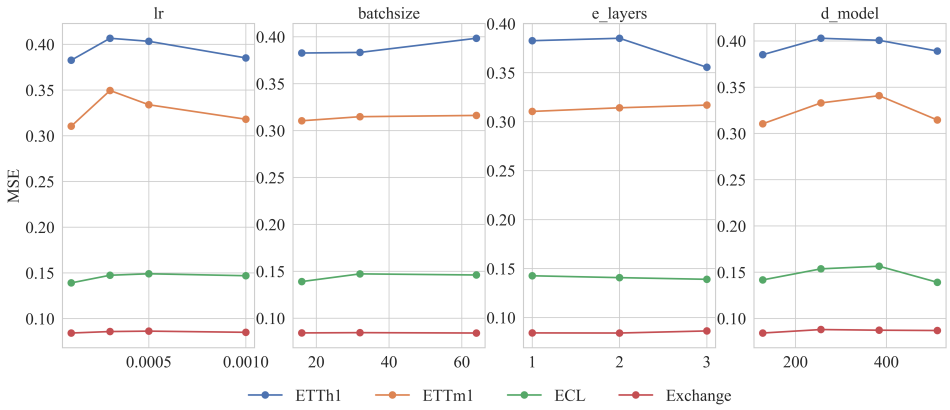


Figure 7: Hyperparameter sensitivity analysis of the WILD framework on four benchmark datasets (ETTh1, ETTm1, ECL, Exchange). We vary the learning rate, batch size, number of encoder layers, and model dimension, and report the corresponding MSE. Results demonstrate the robustness and stable performance of WILD across different configurations.

Complexity (per mini-batch of size B). We summarize time and parameter costs:

Operation	FLOPs (dominant terms)	Params
rFFT	$\Theta(B\,d\,L\log L)$	0
Spectral Preconditioning	$\Theta(B\,d\,L)$	0
Real/Imag Linear (channel mixing)	$\Theta(B\,L\,d^2)$	$2d^2+2d$
irFFT	$\Theta(B\,d\,L\log L)$	0
Residual & Dropout	$\Theta(B\,L\,d)$	1 (α)

Space: $\mathcal{O}(B\,L\,d)$ activations + $\mathcal{O}(d^2)$ weights. When d is large (e.g., Traffic with $d=862$), a low-rank factorization $W_\bullet = U_\bullet V_\bullet^\top$ with rank $r \ll d$ reduces each projection to $\Theta(B\,L\,dr)$ FLOPs. Because W_r and W_i are factorized independently, the pair costs $4dr+2d+1$ parameters (two factors, two biases, and the residual scale), not $2dr$. Low rank therefore reduces parameters and analytic projection FLOPs while leaving the dense backbone unchanged.

Measured efficiency. Table 7 reports the measured parameter count, per-forward-pass wall-clock time, and peak GPU memory of the WILD module across dataset configurations on an NVIDIA RTX 4090. The overhead is negligible for small-to-medium d and remains modest even at $d=862$.

Overhead is computed as WILD parameters divided by typical iTransformer backbone parameters ($\approx$6.6M for $d_{\text{model}}=512$). For Traffic, rank-32 reduces WILD parameters from 1.49M to 112K ($\approx$1.7%); rank-64 uses 222K. Low rank reduces parameters and analytic projection FLOPs but not end-to-end latency: measured forward time changes by $+10.9\%/+10.1\%$ on ECL (ranks 32/64) and $+2.2\%/+2.9\%$ on Traffic relative to full rank (Table 21). Wall-clock time is dominated by the dense backbone and kernel-launch overhead. Peak memory scales modestly with d and remains well within the 24GB budget of a single RTX 4090.

Table 21: Low-rank WILD profiler on an NVIDIA RTX 4090 (iTransformer, $L=96$, $H=96$). Parameters count the WILD block only; latency is full-model forward time and Δ is relative to the full-rank row.

Dataset	Rank	WILD params	Projection FLOPs	Forward (ms)	Δ latency
ECL	Full	206,725	323.1M	4.225	–
ECL	32	41,731	64.4M	4.687	+10.9%
ECL	64	82,819	128.9M	4.652	+10.1%
Traffic	Full	1,487,813	2.330G	18.719	–
Traffic	32	112,061	173.0M	19.128	+2.2%
Traffic	64	222,397	346.0M	19.254	+2.9%

F Comparison with Frequency-Domain Methods

Table 22 situates WILD relative to representative frequency-domain forecasting methods. The key distinguishing factor is *how* each method uses the frequency domain. Prior methods predominantly operate in the frequency domain as a computational or filtering tool—replacing attention, applying learnable filters, or performing task projections. WILD, in contrast, uses the frequency domain to enable a specific structural operation: asymmetric real-imaginary mixing that implements a widely linear transform not reducible to a time-domain right-multiplication. This column-space argument is unique to WILD and is orthogonal to the design choices of prior work—a model could, in principle, combine WILD-style conditioning with any of the backbone architectures listed below.

Table 22: Comparison of WILD with representative frequency-domain forecasting methods.

Method	**Freq. Operation**	**Modifies Attention?**	**Complex-Linear?**	**Not $X \mapsto XM$?**	**Plug-in?**
FNet [Lee-Thorp et al., 2021]	Fourier mixing	Replaces	Yes	No	No
AFNO [Guibas et al., 2022]	Freq. MLP	Replaces	Yes	No	No
FEDformer [Zhou et al., 2022]	Freq. attention	Replaces	Yes	No	No
FITS [Xu et al., 2024]	Learnable filter	Replaces	Yes	No	No
Fredformer [Piao et al., 2024]	Learnable filter	Replaces	Yes	No	No
FreDF [Wang et al., 2025]	Freq. projection	Replaces	Yes	No	No
WILD (Ours)	**Widely linear**	**No**	**No ($W_r \neq W_i$)**	**Yes**	**Yes**

Column descriptions: *Complex-Linear?*—does the method use tied/symmetric operations ($W_r = W_i$) in the frequency domain, which are equivalent to a standard complex-linear transform? *Not $X \mapsto XM$?*—is the frequency-domain operation provably not reducible to a fixed time-domain right-multiplication? *Plug-in?*—can the method be inserted into existing Transformer backbones without architectural modification?

The core distinction is structural. WILD is not an alternative to FEDformer or FITS—it is an orthogonal contribution. Where prior methods replace or redesign the attention mechanism in the frequency domain, WILD conditions the input representation before attention, leaving the backbone unchanged. The widely linear mechanism ($W_r \neq W_i$) is what makes this conditioning structurally non-trivial: it ensures the module can produce representations not achievable by standard time-domain preprocessing. This property does not depend on the specific choice of backbone and applies to any Transformer architecture that applies attention to time series inputs.

G Analysis of Attention Structure

To further investigate the effectiveness of the proposed WILD module, we visualize the attention distribution before and after incorporating WILD through heatmaps on the ETTh2 dataset. As shown in Figure 8, the baseline model exhibits block-like attention patterns that suggest limited expressiveness and an inability to differentiate between important and irrelevant temporal dependencies. These dense and repetitive structures indicate that the model tends to treat different time steps and features uniformly, which may lead to suboptimal forecasting. In contrast, after incorporating WILD, the attention maps display significantly enhanced granularity and discriminative patterns, with sharper focus on salient regions. This improvement demonstrates WILD's capacity to enrich the model's representational power by promoting more diverse and information-aware attention patterns.

Moreover, we further analyze the representational properties of the attention maps by evaluating two key metrics: attention entropy and the Top-5 Singular Value Energy Ratio (SVER), as shown in Figure 9. These metrics offer complementary perspectives on the diversity and expressive capacity of the learned attention patterns.

We introduce row-averaged **attention entropy** $\mathcal{H}$ to quantify the uncertainty or dispersion in attention distributions. For a row-stochastic attention matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$, it is defined as:

$$\mathcal{H}(\mathbf{M}) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{M}_{ij} \log(\mathbf{M}_{ij} + \epsilon), \tag{7}$$

where ϵ is a small numerical constant. Lower entropy indicates more focused attention, while higher entropy indicates more diffuse attention. Attention entropy is related to, but distinct from, matrix effective rank, which is computed from the entropy of normalized singular values:

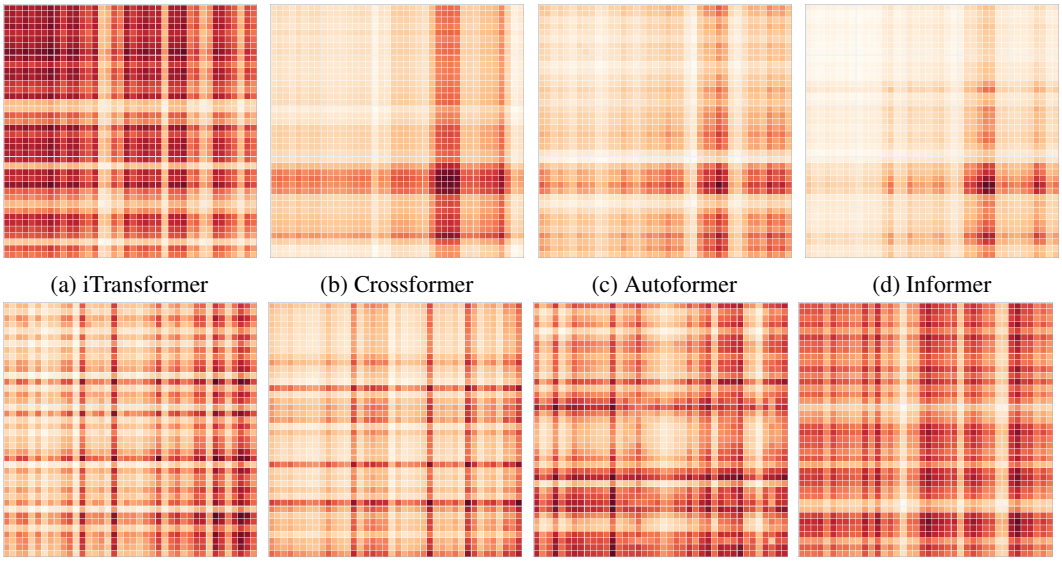
$$\operatorname{erank}(\mathbf{M}) = \exp\left(-\sum_i p_i \log p_i\right), \quad \text{where } p_i = \frac{\sigma_i}{\sum_j \sigma_j}, \tag{8}$$

with σ_i denoting the singular values of M. A skewed singular value distribution (i.e., a few dominant σ_i) results in lower entropy and thus a lower effective rank, indicating redundancy. Conversely, a higher effective rank suggests richer and more diverse representations. To further assess low-rankness, we compute the Top-5 Singular Value Energy Ratio (SVER):

$$\mathrm{SVER}_5(\mathbf{M}) = \frac{\sum_{i=1}^{5} \sigma_i^2}{\sum_{j=1}^{r} \sigma_j^2}, \qquad r = \operatorname{rank}(\mathbf{M}), \tag{9}$$

which measures the proportion of total matrix energy captured by the top-5 singular values. A higher SVER indicates that most of the information is concentrated in a few dominant components, reinforcing the interpretation of a low-rank structure.

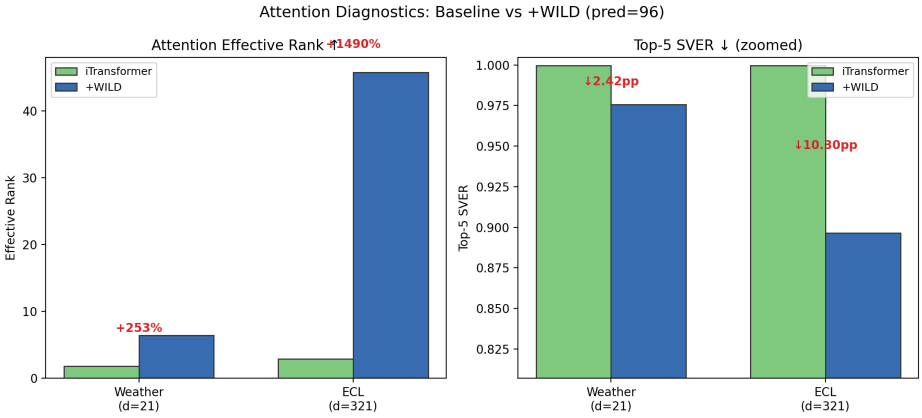
Discussion. The goal of introducing the WILD module is not simply to maximize row entropy, but to reduce redundant low-rank concentration while preserving selective attention. Figure 9 reports the effective rank and Top-5 SVER of attention matrices for the iTransformer backbone with and without WILD on the Weather (d=21) and ECL (d=321) datasets. WILD dramatically increases effective rank (Weather: 1.87 $\rightarrow$ 6.42, +243%; ECL: 2.78 $\rightarrow$ 45.71, +1544%) and reduces top-5 energy concentration (Weather SVER: 0.9997 $\rightarrow$ 0.9755; ECL: 0.9997 $\rightarrow$ 0.8966). The improvement is substantially larger on ECL, consistent with the conjugate decomposition analysis: larger channel dimension d provides more room for spectral preconditioning to balance the singular value spectrum.



(a) iTransformer (b) Crossformer (c) Autoformer (d) Informer

(e) iTransformer (WILD) (f) Crossformer (WILD) (g) Autoformer (WILD) (h) Informer (WILD)

Figure 8: Attention heatmaps of the original model (a–d) and the model enhanced with WILD (e–h) on the ETTh2 dataset. Each map is independently min–max normalized for structural comparison; darker red indicates higher normalized attention, and absolute values are not comparable across panels. The original model exhibits obvious block-wise attention patterns with low discriminability. After introducing WILD, the block structure is alleviated and the attention maps show improved focus and sharper distinctions, indicating better global awareness and feature separation.



(a) Effective rank (b) Top-5 SVER

Figure 9: (a) Attention effective rank of the iTransformer backbone with and without WILD on Weather (d=21) and ECL (d=321) datasets. WILD substantially increases effective rank, especially on large-channel datasets.(b) Top-5 singular value energy ratio (SVER). WILD reduces energy concentration, indicating more balanced attention spectra.

H Visualization of Prediction Results

Figures 10, 11, and 12 show illustrative single-window forecasts for three representative models—iTransformer, PatchTST, and Crossformer—in their vanilla and WILD-enhanced forms. These visualizations are qualitative examples only; they are not used to support the quantitative comparisons, which are reported in the main tables.

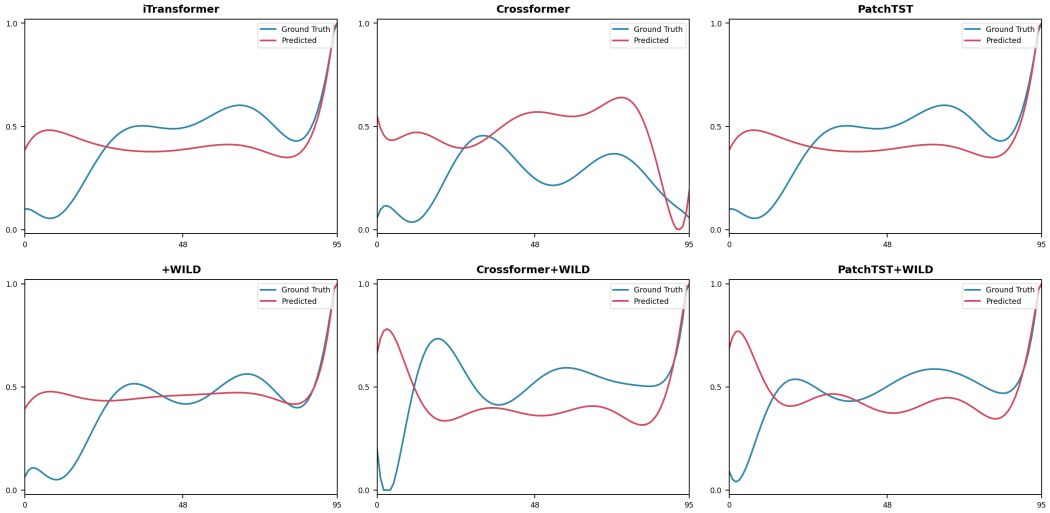


Figure 10: Illustrative prediction visualization on ETTh2 (input-96-predict-96), not used as quantitative evidence. Columns show iTransformer, Crossformer, and PatchTST, respectively; vanilla models are on top and their WILD-enhanced counterparts below. The horizontal coordinate is forecast step (0–95).

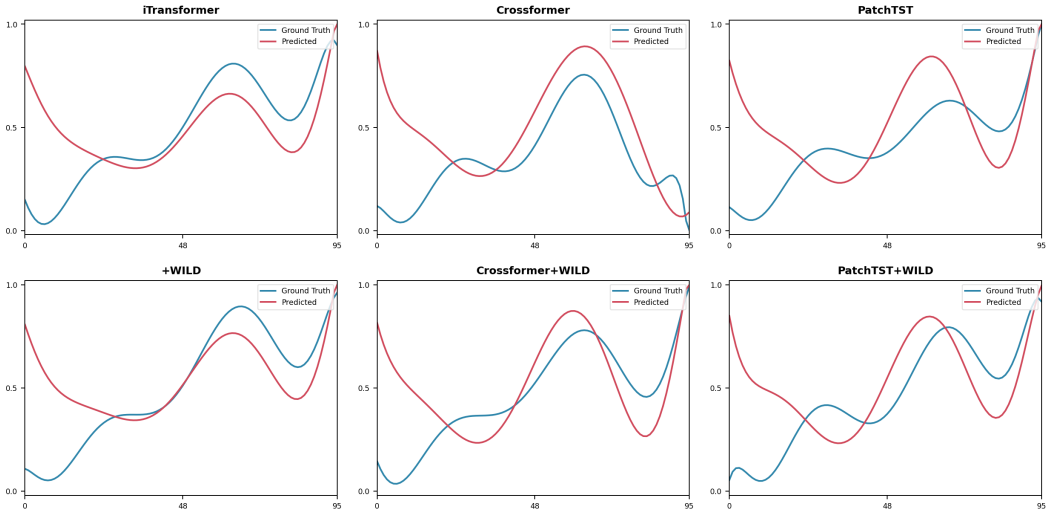


Figure 11: Prediction visualization on ETTm2 (input-96-predict-96). The horizontal coordinate is forecast step (0–95).

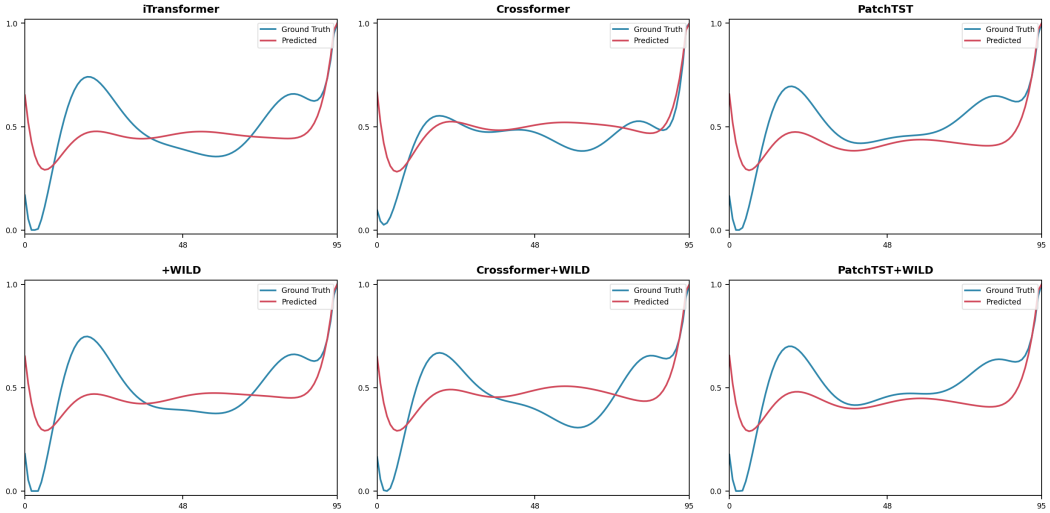


Figure 12: Prediction visualization on ECL (input-96-predict-96). The horizontal coordinate is forecast step (0–95).

I Block Attention Structure: Derived Patterns and Manifestations

In this section, we provide theoretical and empirical insights into several structural properties observed in attention matrices of transformer-based time series forecasting models. These patterns, especially block-wise structures, are further analyzed through mathematical formulations and explanations.

Diagonal Concentration and Band Patterns. The diagonally concentrated attention arises due to strong local temporal autocorrelation. Let x_t be a time series such that $\mathbb{E}[(x_t - \mu)(x_{t+\tau} - \mu)] = \gamma(\tau)$ is the autocovariance function. For many real-world time series, $\gamma(\tau)$ decays rapidly as $|\tau|$ increases.

Given that attention weights are computed via:

$$A_{ij} = \frac{\exp(q_i^\top k_j)}{\sum_{j'} \exp(q_i^\top k_{j'})}$$

where q_i, k_j depend on x_i, x_j, local autocorrelation implies $q_i^\top k_j$ is maximized when $|i-j|$ is small. Hence, attention concentrates along the diagonal.

Periodic Block Replication. For time series with period P, i.e., $x_t \approx x_{t+P}$, this periodicity induces a block structure in the attention maps, repeating every P steps. Formally, if $q_i = q_{i+P}$ and $k_j = k_{j+P}$ approximately hold, then:

$$A_{i,j} \approx A_{i+P,j+P}$$

This implies periodic blocks in the attention matrix. Fourier decomposition can be used to show that dominant frequencies in x_t lead to structural repetitions in learned attention patterns.

Asymmetric Block Structures. In causal attention, the attention mask enforces $A_{ij} = 0$ for $j > i$. This lower triangular structure leads to asymmetric blocks where attention is only allocated to the past:

$$A = \text{softmax}(QK^\top + M), \quad M_{ij} = \begin{cases} 0, & j \le i \\ -\infty, & j > i \end{cases}$$

The asymmetry is thus a direct consequence of this masking. Empirically, the asymmetry can be quantified via:

$$\text{Asym}(A) = \|A - A^\top\|_F$$

which is significantly nonzero in such cases.

Robustness to Noise and Overfitting. The block-wise structure in attention matrices can be characterized through a low-rank plus noise decomposition. Let $A = A^{\text{block}} + \epsilon$, where A^{block} is the structured component and ϵ denotes noise or overfitting-induced irregularity.

A bounded Frobenius norm of ϵ:

$$\|\epsilon\|_F^2 \ll \|A^{\text{block}}\|_F^2$$

implies model robustness. Furthermore, using a block mask or WILD-like structure restricts attention to semantically meaningful groups, effectively reducing variance and thus generalization error according to the bias-variance tradeoff.

J Protocol-Matched Five-Seed Results

This section reports the matched analysis used in the revision. All rows are five-seed means under a validation-selected protocol: configurations and checkpoints are selected on validation performance, and the corresponding test metrics are reported after selection. Paired tests are two-sided t-tests on per-seed deltas. Table 23 isolates the ECL-H96 arms, Table 24 reports the ECL horizons and Weather-H96, and Table 25 records the ETTh1-H96 negative boundary.

Table 23: ECL-H96 matched decomposition (five seeds). WILDsym sets $W_r = W_i$; the time-domain (TD) control applies two matched maps outside the FFT domain.

Arm	MSE	MAE
Base	0.1482	0.2398
Spectral only	0.1479	0.2395
Complex	0.1412	0.2400
WILD	0.1404	0.2381
WILDsym	0.1396	0.2368
Time-domain two-map	0.1553	0.2572

K Frozen Spectral-Concentration Intervention

We freeze ECL-H96 checkpoints and change only the spectral-energy concentration of the held-out input windows, preserving phase, cross-variable direction, total spectral energy, model parameters, and the downstream attention consumer. Table 26 reports layer-0 slopes over five checkpoint seeds, 30 contiguous test batches per seed, five perturbation strengths, and a circular block bootstrap with

Table 24: Matched five-seed accuracy. "Wins" counts seeds where WILD improves on Base; p is the paired t-test. ECL MAE is not significant at H=192 or H=336.

Setting	Base MSE	WILD MSE	ΔMSE	Wins	p (MSE)	Base MAE	WILD MAE	p (MAE)
ECL / 96	0.1482	0.1404	-5.30%	5/5	1.1×10^{-6}	0.2398	0.2381	7.0×10^{-4}
ECL / 192	0.1647	0.1606	-2.49%	5/5	3.7×10^{-4}	0.2559	0.2568	0.132
ECL / 336	0.1784	0.1743	-2.28%	5/5	1.2×10^{-2}	0.2712	0.2710	0.804
ECL / 720	0.2091	0.1996	-4.50%	5/5	2.7×10^{-3}	0.2989	0.2940	8.4×10^{-3}
Weather / 96	0.1753	0.1668	-4.86%	5/5	9.4×10^{-4}	0.2149	0.2106	4.6×10^{-3}

Table 25: ETTh1-H96 matched decomposition (five seeds), a negative boundary. WILD does not improve over Base on MSE (1/5 wins, $p = 0.209$) and degrades MAE (0/5 wins, $p = 0.011$).

Arm	Base	Spectral	Complex	WILD	WILDsym	TD
MSE	0.3897	0.3885	0.3935	0.3910	0.3937	0.3910
MAE	0.4059	0.4052	0.4088	0.4089	0.4088	0.4091

4,000 resamples. The intervention changes attention geometry in both models, and WILD attenuates the layer-0 sensitivity. This is a scoped geometry result on ECL-H96; it does not establish a full mediation chain from spectral rank to natural-distribution accuracy.

Table 26: ECL-H96 frozen intervention: layer-0 slopes as spectral concentration increases (mean [95% bootstrap CI]). Larger row redundancy and lower stable rank indicate a more concentrated attention geometry.

Layer-0 metric	Base slope	WILD slope
Row redundancy	$+0.315$ $[0.307, 0.323]$	$+0.210$ $[0.205, 0.216]$
Stable rank	-0.0045 $[-0.0047, -0.0044]$	-0.0031 $[-0.0033, -0.0030]$
Diagonal band mass	-2.380 $[-2.487, -2.277]$	-2.620 $[-2.729, -2.510]$

L Paired Add-On Tests on Linear and Frequency-Domain Backbones

Table 27 reports five-seed paired add-on tests under $L = 96$, $H = 96$. These results are mixed: DLinear and FreTS improve both metrics, SparseTSF and MixLinear improve MSE but worsen MAE, FEDformer is inconclusive, and FITS on ETTh1 is a negative boundary. Official results for methods with different native inputs or settings (e.g., CycleNet, SparseTSF-L720, FITS-L360, FAN) are not pooled as direct paired competitors.

Table 27: Paired add-on tests at $H = 96$ (five seeds). ΔMSE is relative to the reference backbone; the MAE column summarizes direction and paired significance.

Backbone	Setup	Ref. MSE	+WILD MSE	ΔMSE	MAE
DLinear	Traffic	0.6448	0.5873	-8.93%	improves ($p = 5.3 \times 10^{-5}$)
DLinear	Electricity	0.1975	0.1807	-8.53%	improves ($p = 8.6 \times 10^{-6}$)
FreTS	Electricity	0.1895	0.1747	-7.80%	improves ($p = 5.5 \times 10^{-7}$)
SparseTSF	Electricity	0.1970	0.1873	-4.92%	worsens ($p = 4.1 \times 10^{-4}$)
MixLinear	Electricity	0.2098	0.2070	-1.32%	worsens ($p = 3.8 \times 10^{-4}$)
FEDformer	Electricity	0.1924	0.1920	-0.24%	inconclusive ($p = 0.899$)
FITS	ETTh1	0.3849	0.3860	$+0.28\%$	not significant ($p = 0.411$)

M Proofs of Theoretical Results

M.1 Preliminaries

We denote by $\mathrm{Col}(A)$ the column space of matrix A, and by $\mathrm{rank}(A)$ its algebraic rank. The stable rank is defined as $\mathrm{sr}(A) = \|A\|_F^2/\|A\|_2^2$. For a wide-sense stationary (WSS) process with autocovariance $r(\tau)$ and spectral density $f(\omega)$, let $\mathbf{\Sigma}$ be its $L \times L$ Toeplitz covariance matrix.

M.2 Proof of Lemma 3.1: Spectral Preconditioning and Stable Rank

Proof. Write the rFFT representation as $\widehat{X} = [a_1, \ldots, a_K] \in \mathbb{C}^{d\times K}$, where a_k is the channel vector at frequency k. For active frequencies, let $s_k = \|a_k\|_2 > 0$ and $u_k = a_k/s_k$, and define $U = [u_1, \ldots, u_K]$. This is the ideal spectral-preconditioning step without the numerical ε term; the implemented normalization uses $s_k + \varepsilon$ and is a continuous perturbation of this map for active bins with $s_k \gg \varepsilon$.

Since each normalized column has unit norm,

$$\|U\|_F^2 = \sum_{k=1}^{K} \|u_k\|_2^2 = K.$$

By the stated incoherence assumption, $\lambda_{\max}(U^*U) \le 1 + \delta$. Because $\|U\|_2^2 = \lambda_{\max}(U^*U)$,

$$\mathrm{sr}(U) = \frac{\|U\|_F^2}{\|U\|_2^2} \ge \frac{K}{1+\delta}.$$

For the unnormalized matrix, $\|\widehat{X}\|_F^2 = \sum_{k=1}^{K} s_k^2$. The spectral norm dominates every column norm, so $\|\widehat{X}\|_2 \ge \max_k \|a_k\|_2 = s_{\max}$. Hence

$$\mathrm{sr}(\widehat{X}) = \frac{\|\widehat{X}\|_F^2}{\|\widehat{X}\|_2^2} \le \frac{\sum_{k=1}^{K} s_k^2}{s_{\max}^2}.$$

If $K/(1+\delta) > \sum_k s_k^2/s_{\max}^2$, the lower bound for $\mathrm{sr}(U)$ exceeds the upper bound for $\mathrm{sr}(\widehat{X})$, so $\mathrm{sr}(U) > \mathrm{sr}(\widehat{X})$. □

Connection to WSS intuition. For a WSS process, the DFT approximately diagonalizes the Toeplitz covariance matrix [Gray, 2006]: the columns of $\widehat{X}$ (indexed by frequency k) are approximately uncorrelated, with column k having expected squared norm proportional to the spectral density $f(\omega_k)$. When the time series exhibits temporal smoothness, $f(\omega)$ is concentrated at low frequencies. Spectral preconditioning normalizes each frequency column by $e_k = \sqrt{\sum_{j=1}^{d}(R_{j,k}^2 + I_{j,k}^2)} + \varepsilon$, which balances the contributions across frequency bands.

M.3 Proof of Lemma 3.2: Widely Linear Transforms Are Not Fixed Right-Multiplications

Lemma M.1 (Widely linear transform — restated)**.** *Let $X \in \mathbb{R}^{L\times d}$ with $L \ge 3$, and let $X_f = R + \mathrm{i}I \in \mathbb{C}^{d\times K}$ be its frequency-domain representation ($K = \lfloor L/2 \rfloor + 1$). Let $W_r, W_i \in \mathbb{R}^{d\times d}$ with $W_r \ne W_i$. Define the composite time-domain map:*

$$\mathcal{T}(X) = \mathrm{irFFT}\big((W_r R + \mathrm{i}\, W_i I)^\top\big) \in \mathbb{R}^{L\times d}.$$

Then $\mathcal{T}$ is not equivalent, as a map on all real inputs, to a right-multiplication $X \mapsto XM$ for any fixed $M \in \mathbb{R}^{d\times d}$.

Proof. Express the real and imaginary parts in terms of X_f and its complex conjugate:

$$R = \frac{X_f + \overline{X_f}}{2}, \qquad I = \frac{X_f - \overline{X_f}}{2\mathrm{i}}.$$

Then the transformed frequency-domain signal is:

$$W_r R + \mathrm{i}\, W_i I = W_r \frac{X_f + \overline{X_f}}{2} + \mathrm{i}\, W_i \frac{X_f - \overline{X_f}}{2\mathrm{i}} \tag{10}$$

$$= \frac{W_r + W_i}{2} X_f + \frac{W_r - W_i}{2} \overline{X_f}. \tag{11}$$

This is a *widely linear* transform: it involves both X_f and its complex conjugate $\overline{X_f}$. A standard complex linear transform would only involve X_f (i.e., the coefficient of $\overline{X_f}$ would be zero).

The key observation: if $W_r = W_i$, the $\overline{X_f}$ term vanishes and the operation reduces to $X_f \mapsto W_r X_f$, which is equivalent to the time-domain right-multiplication $X \mapsto X W_r^\top$ (since FFT is unitary). When $W_r \neq W_i$, the $\overline{X_f}$ term is nonzero, introducing a conjugate-linear component absent from any right-multiplication.

Formally, suppose for contradiction that $\mathcal{T}(X) = XM$ for some $M \in \mathbb{R}^{d \times d}$ and all real X. Then in the frequency domain, $\mathrm{rFFT}(\mathcal{T}(X)^\top) = \mathrm{rFFT}(M^\top X^\top) = M^\top X_f$. But our construction yields $\frac{W_r+W_i}{2} X_f + \frac{W_r-W_i}{2} \overline{X_f}$. For $L \geq 3$, there exists at least one frequency index k away from DC and Nyquist. At this frequency, we may choose any $z \in \mathbb{C}^d$ as the k-th column of X_f (since the real signal constraint imposes no restriction at non-endpoint frequencies). Equality for all such z requires the conjugate-linear coefficient to vanish, i.e., $W_r = W_i$, contradicting $W_r \neq W_i$.

Therefore, no fixed M can satisfy $\mathcal{T}(X) = XM$ for all X, and $\mathcal{T}$ is structurally distinct from any time-domain right-multiplication. □

Remark on column-space inclusion. The lemma establishes that $\mathcal{T}$ is not a right-multiplication, which means it is structurally richer than standard time-domain linear preprocessing. However, we note that "not a right-multiplication" does not by itself guarantee $\mathrm{Col}(\mathcal{T}(X)) \not\subseteq \mathrm{Col}(X)$ for every input X. The practical significance is that $\mathcal{T}$ provides a representational degree of freedom unavailable to any fixed M, and in combination with the data-dependent nonlinearity of spectral preconditioning, this permits the module to learn enriched representations that empirically improve forecasting accuracy.

M.4 Remark on Effective Rank vs. Algebraic Rank

The algebraic rank of $X \in \mathbb{R}^{L \times d}$ is bounded above by $\min(L, d)$. WILD does not claim to increase the algebraic rank beyond this trivial bound. Rather, it improves the *effective* conditioning: the spectral preconditioning makes the singular value distribution more uniform (higher stable rank), and the widely linear mixing can introduce representational directions that were not practically usable in the original low-effective-rank regime. The combination means that downstream attention layers receive input with a richer usable basis, which empirically translates to more diverse attention patterns and improved forecasting accuracy.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the contributions and scope, and Sections 4.1–4.4 support the empirical and mechanistic claims.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 discusses the empirical scope and the limits of the efficiency and ablation analyses, and identifies broader empirical and theoretical characterization as future work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The stable-rank lemma states explicit spectral-energy and incoherence assumptions and proves a sufficient condition for stable-rank improvement. The widely-linear lemma states the required $L \geq 3$ condition and proves non-equivalence to fixed right-multiplication in Appendix M.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Appendix A provides dataset statistics, preprocessing, training settings, and hyperparameters, and Appendix C.3 documents the ablation protocol.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The benchmark datasets are publicly available, but the source code is not released.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix A lists datasets, splits, batch size, dropout, training epochs, and hardware; the main text describes the evaluation protocol.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The main benchmark tables report MSE and MAE averaged over four prediction horizons. Appendix J reports protocol-matched five-seed results and two-sided paired t-tests on per-seed differences, including positive and negative results; Appendix L reports additional five-seed paired tests. Appendix K reports 95% confidence intervals from a circular block bootstrap with 4,000 resamples for the frozen intervention. These analyses cover the specified configurations, not every configuration in the main tables.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix A specifies a single NVIDIA RTX 4090 (24 GB), and Table 7 reports parameter counts, forward time, and peak memory.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses public benchmark datasets and involves no human subjects or sensitive data; it conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

Justification: The work is foundational research on a general forecasting module evaluated on public benchmarks; it is not tied to a specific deployment or application.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release models or data with a high risk of misuse; the proposed module is a general forecasting component.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: The datasets are cited in Appendix A, but their licenses and terms of use are not explicitly documented in the paper.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new datasets, models, or code; the proposed module is described in Section 3.2 and Appendix C.3.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work does not involve human subjects; no IRB approval was required.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method does not involve LLMs as an important, original, or non-standard component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.