SWiT: Spectral Wiener Correction for Test-Time Adaptation in Time Series Forecasting

Anonymous authors
Paper under double-blind review

Abstract

Time-series forecasting models degrade after deployment because the statistical properties of the data shift over time in real-world environments. Test-time adaptation (TTA) keeps the deployed backbone frozen and updates a small adapter from the ground truth that arrives after each forecast. Because that ground truth arrives one step at a time, the freshest windows are only partially observed when adaptation runs, so an adapter must learn from incomplete labels without admitting any unobserved value. Existing methods answer this with modules updated by stochastic gradient descent, so incomplete labels can be used only through a gradient step; the correction depends on the learning rate and the number of steps taken, and the cost grows quadratically with the horizon. We propose Spectral Wiener correction for Test-time adaptation (simply, SWiT), which corrects a frozen backbone's forecast in the frequency domain and obtains that correction in closed form rather than by gradient descent. The prediction and its input history are transformed by the Fast Fourier Transform; every bin receives an affine correction (i.e., a scale, an input term, and a rotational bias phase-locked to absolute time), and the result is transformed back. By Parseval's theorem, the time-domain mean squared error (MSE) separates into independent per-bin quadratics, so the per-batch optimum is a closed-form Wiener-type solution over running sums, with no optimizer state or learning rate, at $\mathcal{O}(H \log H)$ cost in the forecasting horizon H. A causal protocol splices only observed steps, the prefix of each in-progress window, into those same running sums, so the freshest evidence enters the closed-form solve itself, and an evidence gate scales each bin's correction by the benefit it has delivered on past windows. Across 216 settings spanning six backbones, nine benchmarks, and four horizons, SWiT attains the lowest average MSE, 10.3% below the frozen backbone and $5.3 \sim 7.9\%$ below the state-of-the-art TTA methods, with $10 \sim 257\times$ fewer parameters and the shortest adaptation time. The code is available at `https://anonymous.4open.science/r/SWiT-808C`.

1 Introduction

Time-series forecasting is central to domains as diverse as energy (Hong et al., 2016), finance (Lim et al., 2021), and climate science (Bi et al., 2023). However, time series are inherently non-stationary. Their mean, variance, and temporal patterns change continuously, so deployed forecasters suffer accuracy drops as the test distribution drifts away from the training distribution (Liu et al., 2022).

Prior remedies, architectural (Liu et al., 2022), normalization-based (Kim et al., 2021; Fan et al., 2023; Dai et al., 2024; Liu et al., 2023), and online or continual learning (Zhang et al., 2024; Zhao & Shen, 2025; Chen et al., 2025; Kirkpatrick et al., 2017; Pan et al., 2024), either bake the adaptation strategy into training or update a substantial part of the parameters at test time, which suits neither already-deployed models nor compute-constrained edge environments.

Test-time adaptation (TTA) takes a more practical and recent stance. The backbone remains frozen in its deployed state, and only an attachable, lightweight adapter is updated online using the ground truth (GT) observed after each prediction. Recent TTA methods for time-series forecasting (TS-TTA) (Kim et al., 2025; Im & Kwon, 2026) follow this template, but it comes with two difficulties. The first is the update itself. Dense linear modules trained by stochastic gradient descent (SGD) incur $\mathcal{O}(H^2)$ FLOPs in the forecasting horizon H, and their correction depends on the learning rate and on

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

the number of steps taken. The second, common to every TTA method for forecasting, is that the GT of an H-step forecast arrives one step at a time. Therefore, when adaptation runs, the windows that carry the freshest evidence are exactly those whose labels are still incomplete. An adapter must draw on that evidence without using any value that has not yet been observed.

In this work, we propose **S**pectral **Wi**ener correction for **T**est-time adaptation (simply, SWiT), a TTA module that corrects the backbone forecast in the frequency domain. SWiT transforms the prediction and its input history by the Fast Fourier Transform (FFT), applies an affine correction to every frequency bin, consisting of a scale on the prediction, an input term, and a rotational bias phase-locked to absolute time, and reconstructs the corrected prediction by an inverse real FFT (irFFT). Because Parseval's theorem separates the time-domain MSE into independent per-bin quadratics, the per-batch optimum over all correction coefficients admits a single closed-form solution, Wiener-filter in form, computed from accumulated running sums. Adaptation therefore requires no iterative computation, no backward pass, and no optimizer state. Under a causal protocol that separates fully matured from partially observed ground truth (POGT) (Kim et al., 2025), SWiT introduces a splicing rule that folds only the observed prefix of in-progress windows into those running sums, and an evidence gate that scales each bin's correction by the benefit that correction has actually delivered on past windows.

The contributions are summarized as follows.

- **Model-agnostic spectral correction.** SWiT corrects the backbone's forecast in the frequency domain with a bin-wise affine map consisting of a scale, an input term, and a rotational bias. SWiT uses only the backbone's prediction and its input history, so it attaches to any forecaster. Across six backbones from three different architecture families, SWiT lowers MSE relative to the frozen backbone on every dataset, with per-backbone average reductions from 8.02% to 14.46%.
- **Closed-form adaptation.** By Parseval's theorem, the adaptation objective splits into independent per-bin quadratics. The per-batch optimum is therefore computed exactly from running sums in $\mathcal{O}(H \log H)$, with no SGD, no optimizer state, and no learning rate. Correcting in the frequency domain is also what keeps the adapter small: its bin-wise map is diagonal and carries only $\mathcal{O}(H)$ coefficients, yet it matches a dense time-domain corrector with $\mathcal{O}(H^2)$ parameters. The basis carries much of the accuracy as well, since one gradient step on the same per-bin objective already matches the strongest prior method and the closed form adds the rest (Sec. 4.3). SWiT is 10–$257\times$ smaller and 3.1–$11.4\times$ faster to adapt.
- **Leak-free adaptation under delayed supervision.** Under a causal protocol that separates matured GT from POGT, SWiT introduces a splicing rule that admits only the observed prefix of in-progress windows into the closed-form statistics, and an evidence gate that scales each bin's current correction by how well that bin's past corrections agreed with the GT. The splicing rule improves average accuracy by 3.1% and the evidence gate by 2.1%, and the two together by 4.9% relative to the adapter with neither.
- **State-of-the-art accuracy.** Over 216 settings (six backbones, nine benchmarks, four horizons), SWiT achieves the largest gain among TTA methods, lowering forecast error by 10.3% against the frozen backbone and by 5.3–7.9% against existing methods.

2 Related Work

2.1 Frequency-Domain Forecasting

FEDformer (Zhou et al., 2022b) pairs seasonal-trend decomposition with attention computed on a few Fourier modes, FiLM (Zhou et al., 2022a) filters a Legendre-memory representation of the history in frequency to suppress noise, and FreTS (Yi et al., 2023) reports that the same MLPs learn more effectively on the spectrum than on the time axis. FITS (Xu et al., 2024) forecasts by interpolating in the complex frequency domain and discarding high-frequency components; with about 10K parameters, it matches models orders of magnitude larger and is small enough for edge deployment. In these models, spectral parameterization buys accuracy at a fraction of the capacity, which SWiT exploits. These models leave drift after deployment, because SGD fixes their spectral structure at training time. SWiT instead estimates a per-bin affine correction of the backbone's residual error at test time, for any already trained backbone.

2.2 Non-Stationarity in Time-Series Forecasting

Non-stationarity, the drift in statistical properties of time series over time, was first handled inside the model by the shift-aware attention of Non-stationary Transformers (Liu et al., 2022), and later by normalization. RevIN (Kim et al., 2021) normalizes each instance, Dish-TS (Fan et al., 2023) models the input and output distributions separately, SAN (Liu et al., 2023) normalizes each slice of the input, and DDN (Dai et al., 2024) normalizes dynamically with sliding-window statistics in both the time and frequency domains.

A complementary line adapts the parameters themselves. Online learning methods such as FS-Net (Pham et al., 2022), PROCEED (Zhao & Shen, 2025), and drift-detecting or meta-learned adapters (Zhang et al., 2024; Chen et al., 2025) update the model as observations arrive, and continual learning methods (Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017; Buzzega et al., 2020; Pan et al., 2024) accommodate new distributions without catastrophic forgetting. Both require updating a substantial portion of the backbone after deployment, whereas the setting studied here keeps it frozen.

2.3 Test-Time Adaptation

TTA originated in vision (Sun et al., 2020; Wang et al., 2020), where no label follows the prediction, so adaptation relies on unsupervised proxies. Extending TTA to time-series forecasting adds three requirements: 1) The temporal order of inputs must be preserved during adaptation, 2) distribution shift manifests as gradual statistical drift rather than covariate shift, and 3) supervision is delayed, so the most recent windows are always only partially observed. Recent TS-TTA methods adapt with direct MSE or MAE losses on the GT that arrives after each batch. TAFAS (Kim et al., 2025) attaches lightweight calibration modules at the backbone input and output, updates them via SGD under a periodicity-aware schedule, and introduces POGT by attaching the loss to the observed prefix of in-progress windows, a route a closed-form adapter cannot take. PETSA (Medeiros et al., 2025) and DynaTTA (Grover & Etemad, 2025) extend this design toward parameter efficiency and shift-awareness, respectively. COSA (Im & Kwon, 2026) regresses a linear residual at the backbone output from a sliding context buffer, also via SGD. TAFAS and COSA incur $\mathcal{O}(H^2)$ adaptation FLOPs per horizon, and because their modules take one gradient step per batch, the correction they apply depends on the learning rate and the number of steps taken.

SWiT differs from these methods in three ways: 1) it corrects in the frequency domain rather than the time domain, and at the output of the frozen backbone alone, 2) it updates in closed form from running sums instead of via an SGD step, removing the optimizer state, learning rate, and iterative updates, and 3) it folds POGT into those statistics and scales each bin's correction by how well that bin's past corrections agreed with the GT.

3 Proposed Method

3.1 Notation and Problem Formulation

Setting. SWiT operates on univariate forecasting. For multivariate series, it treats each channel as an independent univariate forecasting problem. The backbone f_θ is frozen and maps a lookback window $\mathbf{x}$ of length W to a forecast $\hat{\mathbf{y}}^{(0)} = f_\theta(\mathbf{x}) \in \mathbb{R}^H$ of length H. Windows are indexed by their forecast origin; window j forecasts time indices $j, \dots, j+H-1$. Its GT is written $\mathbf{y}_j$, and the τ-th entry of $\mathbf{y}_j$ is $y_j[\tau]$ ($\tau = 0, \dots, H-1$). Besides the backbone output $\hat{\mathbf{y}}^{(0)}$, SWiT uses the input history $\mathbf{h}_j \in \mathbb{R}^H$, the H steps immediately preceding the origin of window j, accumulated from the past inputs of length W and zero-filled where fewer than H past steps have been observed. The only quantity updated at test time is the parameter ϕ_t of the adapter, and the corrected forecast is written $\hat{\mathbf{y}}_j = \phi_t(\hat{\mathbf{y}}_j^{(0)}, \mathbf{h}_j)$. The goal is to minimize $\mathcal{L}(\phi_t)$ under the test distribution. All notations are presented in Table 4, Appendix A.

$$\mathcal{L}(\phi_t) = \mathbb{E}_{(\mathbf{x},\mathbf{h},\mathbf{y})}\left[\|\phi_t(f_\theta(\mathbf{x}), \mathbf{h}) - \mathbf{y}\|_2^2\right].$$

Two Kinds of Ground Truth. The GT $\mathbf{y}_j$ of a forecast is not observed at once but one step at a time over H steps. If $m_j = t - j$ denotes the time elapsed for window j at time index t, the number of observed ground-truth steps is $\min(m_j, H)$. **Matured GT** is the GT of a window whose labels

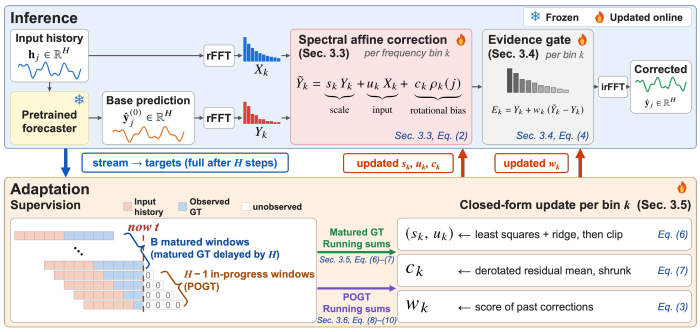


Figure 1: Overview of SWiT.

have all arrived ($m_j \geq H$, the set $\mathcal{M}_t$): complete, but slow to arrive. **POGT** (Kim et al., 2025) is the first m_j steps of an in-progress window ($1 \leq m_j < H$, the set $\mathcal{A}_t$): incomplete, but fresh. At every time step, one window moves from the POGT set $\mathcal{A}_t$ to the matured set $\mathcal{M}_t$, and the POGT of every remaining window grows by one step. Using both kinds of GT is the central design problem of TTA under delayed supervision.

Causal Protocol. Adaptation at time t uses only matured GT, which arrives H steps late, and POGT. Each batch runs two stages in sequence. (i) **Inference** emits the corrected forecasts with ϕ_{t-1}, estimated from the information available so far, and holds each window until its labels arrive. (ii) **Adaptation** folds the newly matured pairs and the POGT of the in-progress windows into the statistics, scores the corrections whose GT has just arrived, and obtains ϕ_t.

3.2 Overview

Figure 1 shows the overall structure of SWiT, which separates inference from adaptation.

Inference (top). For every window, the frozen backbone produces its forecast. The forecast and the preceding input history are transformed into the frequency domain via real FFT (rFFT). Each frequency component is corrected by an affine map comprising a scaled forecast, a term drawn from the history, and a bias locked to absolute time (Sec. 3.3). The evidence gate decides, component by component, how much of that correction to trust (Sec. 3.4). The inverse transform (irFFT) then returns the corrected forecast.

Adaptation (bottom). After each batch, the windows whose labels have fully arrived, one horizon after the forecast, are added once to the running sums. Because the error separates into independent per-component terms, the correction coefficients are then read off these sums in closed form, with no gradient descent (Sec. 3.5). Windows whose labels are still incomplete contribute their observed prefix (i.e., POGT) through a splicing rule that folds it into temporary sums without touching an unobserved value (Sec. 3.6), and the gate weights are updated from how much past corrections actually helped (Sec. 3.4). The updated coefficients and weights are used from the next batch on. Appendix B gives the complete procedure.

3.3 Spectral Correction

A frozen backbone errs in a frequency bin in one of three ways: 1) the component it produces there is mis-scaled, 2) a component it should produce is missing, or 3) a bias locked to absolute time shifts with the window. SWiT corrects each bin independently, with one affine term for each case. For window j, the same rFFT is applied to the backbone forecast $\hat{\mathbf{y}}_j^{(0)}$ and to the input history $\mathbf{h}_j \in \mathbb{R}^H$, yielding $K = \lfloor H/2 \rfloor + 1$ complex coefficients $\mathbf{Y}_j = [Y_1, \ldots, Y_K]^\top$ and $\mathbf{X}_j = [X_1, \ldots, X_K]^\top$ ($k = 1, \ldots, K$).

$$Y_k = \sum_{\tau=0}^{H-1} \hat{y}_j^{(0)}[\tau]\, e^{-\mathrm{i}\, 2\pi(k-1)\tau/H}, \qquad X_k = \sum_{\tau=0}^{H-1} h_j[\tau]\, e^{-\mathrm{i}\, 2\pi(k-1)\tau/H}. \tag{1}$$

The correction of window j is a frequency-bin-wise affine transformation.

$$\tilde{Y}_k = s_k\, Y_k + u_k\, X_k + c_k\, \rho_k(j), \qquad \tilde{\mathbf{Y}}_j = [\tilde{Y}_1, \ldots, \tilde{Y}_K]^\top, \qquad \tilde{\mathbf{y}}_j = \mathrm{irFFT}(\tilde{\mathbf{Y}}_j). \tag{2}$$

Because both transforms have length H, X_k and Y_k refer to the same frequency $(k-1)/H$, and G_k, obtained by applying the same transform to the GT $\mathbf{y}_j$, is the target the correction must match. The correction coefficients are $s_k, u_k \in \mathbb{R}$ and $c_k \in \mathbb{C}$, and $\rho_k(j) = e^{\mathrm{i}\,2\pi(k-1)j/H}$ is the phase rotation factor for the forecast origin of window j. 1) The scale term $s_k Y_k$ corrects the magnitude of the components the forecast already contains. 2) The input term $u_k X_k$ brings in from the input history the components the backbone has missed, that is, the bins where $Y_k \approx 0$ and a scale alone cannot act. 3) The rotational bias $c_k \rho_k(j)$ accounts for a bias whose phase is locked to absolute time: since the rFFT of each window measures phase relative to that window's start, a component fixed to the absolute time index is observed in window j rotated by $\rho_k(j)$, so c_k is a single window-independent constant that is multiplied by $\rho_k(j)$ when used in window j. The parameters are four real values per bin and channel (s_k, u_k, and the real and imaginary parts of c_k). Per window, the transforms cost $\mathcal{O}(H \log H)$ and every per-bin step costs $\mathcal{O}(H)$ in total, so adaptation is dominated by the rFFT, against $\mathcal{O}(H^2)$ for a dense $H \times H$ correction when updated by SGD (Appendix B). The final output forecast combines this correction with the backbone forecast through Eq. (4) (Sec. 3.4).

3.4 Evidence Gate

The evidence gate is a per-bin shrinkage coefficient that decides how much of the correction $\tilde{Y}_k$ of Eq. (2) to apply. Its coefficients are estimated in closed form from accumulated statistics, so the correction is fitted to those statistics. However, that fit fails to carry over to the current window when its distribution departs from the one over which the statistics were gathered, and in bins where the backbone error has no spectral structure to exploit because it is unstructured or falls between frequency bins. In such cases, the correction acts as noise and harms the forecast. The evidence gate therefore scores how the same corrector actually performed on past windows and applies the correction only to the extent that score warrants.

Scoring. When the corrected output of window j is emitted, the correction in force at that moment is stored as $C_k^{(j)} = \tilde{Y}_k^{(j)}$. When the window matures, the stored $C_k^{(j)}$ is scored against the matured GT. Here, $\Delta_k^{(j)} = C_k^{(j)} - Y_k^{(j)}$ is the displacement the proposed correction made, and $D_k^{(j)} = G_k^{(j)} - Y_k^{(j)}$ is the displacement required to reach the GT.

Gating. The state the gate holds is two accumulators, updated whenever a window matures.

$$p_k = \sum_{j \in \mathcal{M}_t} \mathrm{Re}\big(\overline{\Delta_k^{(j)}}\, D_k^{(j)}\big), \qquad d_k = \sum_{j \in \mathcal{M}_t} \big|\Delta_k^{(j)}\big|^2. \tag{3}$$

p_k is the real inner product of the applied and required displacements, that is, how well the correction moved in the direction the GT wanted, and d_k is how far it moved. Two real numbers per bin and channel hold the whole state.

The coefficient $E_k^{(j)}$ of the corrected output is a convex combination of the backbone forecast and the correction, and the mixing weight w_k is the ratio of the two accumulators.

$$w_k = \mathrm{clip}\left(\frac{p_k}{d_k},\, 0,\, 1\right), \qquad E_k^{(j)} = Y_k^{(j)} + w_k \Delta_k^{(j)}. \tag{4}$$

p_k/d_k is the value of w that would have minimized the squared error of $Y_k + w\Delta_k$ on the past windows (Appendix C). The gate clips it to $[0, 1]$, and sets $w_k = 0$ in a bin whose correction has never moved, $d_k = 0$. With $w_k = 1$, the correction is emitted as is; with $w_k = 0$, the backbone forecast is emitted as is.

3.5 Adaptation

Closed-form Optimum Estimation. Because the correction of SWiT has the linear structure of a bin-wise affine transformation, the error accumulated up to the current batch has an exact closed-form minimizer, with no iterative optimization. At time index t, the available observations are the matured

GT of $\mathcal{M}_t$ and the POGT of $\mathcal{A}_t$; this section fits on the matured GT alone, and Sec. 3.6 adds the POGT terms to the same objective.

Since the coefficients differ from window to window, those of window j are written $Y_k^{(j)}, X_k^{(j)}, G_k^{(j)}$. By Parseval's theorem, the time-domain ℓ_2 error decomposes into a weighted sum of per-bin errors where $\gamma_k = 1$ at DC and Nyquist; 2 otherwise.

$$\left\|\tilde{\mathbf{y}}_j - \mathbf{y}_j\right\|_2^2 = \frac{1}{H}\sum_{k=1}^{K} \gamma_k \left| s_k Y_k^{(j)} + u_k X_k^{(j)} + c_k \rho_k(j) - G_k^{(j)} \right|^2, \tag{5}$$

Since neither $\frac{1}{H}$ nor γ_k depends on the coefficients, they do not affect the per-bin minimizer and are omitted below.

Within a bin, we solve for the scale and input coefficients in two stages, then fit the rotational bias c_k to the residual. The two minimizations of Eq. (6) are Stage 1, the unpenalized least-squares pair $(\hat{s}_k, \tilde{u}_k)$, and then Stage 2, which re-solves for u with $\hat{s}_k$ fixed under a ridge penalty, the ℓ_2 term $\lambda_u \Sigma_k^{XX} u^2$; Eq. (7) then fits the bias, where $\lambda_c\, n\, |c|^2$ plays the same role ($\lambda_u, \lambda_c \geq 0$). Stage 2 is not a second solve: by Proposition 1 its minimizer is $\tilde{u}_k$ rescaled.

$$(\hat{s}_k, \tilde{u}_k) \;=\; \underset{s,u\in\mathbb{R}}{\arg\min}\; \mathcal{L}_k(s,u), \qquad \hat{u}_k \;=\; \underset{u\in\mathbb{R}}{\arg\min}\; \mathcal{L}_k(\hat{s}_k, u) \;+\; \lambda_u\, \Sigma_k^{XX}\, u^2, \tag{6}$$

$$\hat{c}_k \;=\; \underset{c\in\mathbb{C}}{\arg\min} \sum_{j\in\mathcal{M}_t} \left| c\, \rho_k(j) - R_k^{(j)} \right|^2 \;+\; \lambda_c\, n\, |c|^2. \tag{7}$$

Here, $\mathcal{L}_k(s,u) = \sum_{j\in\mathcal{M}_t} \left| s\, Y_k^{(j)} + u\, X_k^{(j)} - G_k^{(j)} \right|^2$ is the squared error of bin k over the matured windows, $\Sigma_k^{XX} = \sum_{j\in\mathcal{M}_t} |X_k^{(j)}|^2$ is the energy of the input coefficients, and $n = |\mathcal{M}_t|$. To limit over-correction, both coefficients are clipped before use: $s_k = \mathrm{clip}_{[0.5,\,1.5]}(\hat{s}_k)$ and $u_k = \mathrm{clip}_{[-0.3,\,0.3]}(\hat{u}_k)$, and $R_k^{(j)} = G_k^{(j)} - s_k Y_k^{(j)} - u_k X_k^{(j)}$ is the residual they leave. Both problems are convex quadratics in (s,u) and in c, respectively, so their solutions are available in closed form.

Proposition 1 (Closed-form optimum). *Fix a bin k and $\lambda_u, \lambda_c \geq 0$, and suppose the 2×2 normal matrix of Stage 1 of Eq.* (6) *is positive definite. Then, Stage 1 has the unique minimizer $(\hat{s}_k, \tilde{u}_k)$ given by that linear system, and Stage 2 has the unique minimizer $\hat{u}_k = (1-\alpha_u)\,\tilde{u}_k$ with $\alpha_u = \lambda_u/(1+\lambda_u)$. Since $|\rho_k(j)| = 1$, Eq.* (7) *has the unique minimizer $\hat{c}_k = (1-\alpha_c)\, S_k/n$ with $\alpha_c = \lambda_c/(1+\lambda_c)$, where $S_k = \sum_{j\in\mathcal{M}_t} \overline{\rho_k(j)}\, R_k^{(j)}$ is the derotated residual sum. All three are functions of running sums over $\mathcal{M}_t$ alone, and the same statements hold with the POGT terms of Sec. 3.6 added (Appendix C).*

Accumulation via Running Sums. Solving Eqs. (6) and (7) requires statistics of Y_k, X_k, G_k over $\mathcal{M}_t$. SWiT keeps them as running sums. A window's contribution is added once, at the moment it matures, and the window itself is never revisited. The running sums accumulate over the entire stream, with no forgetting rule such as decay or a sliding window.

Energy-proportional Ridge. A penalty of fixed strength λu^2 acts differently from bin to bin. The data term scales with the bin's energy Σ_k^{XX}, while the penalty is uniform. It is therefore swamped where $\Sigma_k^{XX} \gg \lambda$, typically the low-frequency bins, and excessive where $\Sigma_k^{XX} \ll \lambda$. The same holds along the stream, since Σ_k^{XX} grows with every matured window, so a fixed λ would fade as evidence accumulates. Eqs. (6) and (7) therefore weight the penalty by the energy of the term the coefficient multiplies. This is Σ_k^{XX} for u, and the sample count n for c, since $|\rho_k| = 1$. With $\hat{s}$ fixed, the penalized solution then satisfies $\hat{u}_k = (1-\alpha_u)\,\tilde{u}_k$ with $\alpha_u = \lambda_u/(1+\lambda_u)$, where $\tilde{u}_k$ is the unpenalized solution, so the shrinkage is the same fraction in every bin and at every point of the stream. Likewise $\hat{c}_k$ is $(1-\alpha_c)$ times the mean of the derotated residuals, with $\alpha_c = \lambda_c/(1+\lambda_c)$. The scale s_k rescales the backbone output itself and carries no penalty. The values used throughout are $\lambda_u = \lambda_c = 1$, that is, $\alpha_u = \alpha_c = 1/2$.

Relation to the Wiener Filter. Under a wide-sense-stationary reading of the samples, the $\lambda_u = 0$ solution of Eq. (6) is the two-channel Wiener filter for G_k (Wiener, 1949; Kolmogorov, 1941). The computation itself is plain least squares (Appendix C).

3.6 POGT SPLICING

In gradient-based TTA, using POGT is straightforward. The loss is simply restricted to the observed steps (Kim et al., 2025). SWiT does not push gradients through a loss. It solves for the coefficients directly from accumulated statistics. Because those statistics are built from FFT coefficients of fixed length H, a POGT of length m_j cannot be used as-is, so we need a separate rule to add it to the statistics.

Splicing. For an in-progress window $j \in \mathcal{A}_t$, we build a provisional target of length H, denoted with $\breve{\cdot}$. The observed first m_j steps keep the GT, and the rest is filled with 0. Applying Eq. (1) yields the provisional target coefficients $\breve{G}^{(j)}$, and we multiply the corresponding input coefficients by the observed fraction m_j/H.

$$\breve{g}_j[\tau] = \begin{cases} y_j[\tau], & \tau < m_j \\ 0, & \text{otherwise} \end{cases} \qquad \breve{G}^{(j)} = \mathrm{rFFT}(\breve{\mathbf{g}}_j), \qquad \breve{X}_k^{(j)} = \tfrac{m_j}{H} X_k^{(j)}. \tag{8}$$

m_j/H is the mean of the zero-padding mask, so in expectation it is the factor relating the coefficients of the provisional target to those the fully observed window would produce. The input coefficients are scaled by the same factor so the two sides match. The spectral leakage caused by zero-padding has a different phase in every window and therefore survives only as residual noise in the sum over many windows.

The zero-frequency (DC) bin is the exception. For the zero-padded target, its coefficient is the sum of the POGT, which, once rescaled by H/m_j, has a variance that grows as the POGT shortens (Appendix D). For the DC bin only, the unobserved part is therefore filled with the backbone forecast $\hat{\mathbf{y}}_j^{(0)}$ and the input coefficient is left unweighted, $\breve{X}_1^{(j)} = X_1^{(j)}$ (Appendix C).

Row-separation Rule. The in-progress windows add their POGT terms to both minimizations of Eq. (6) and to the bias fit of Eq. (7).

$$(\hat{s}_k, \tilde{u}_k) \;=\; \operatorname*{arg\,min}_{s,u\in\mathbb{R}} \; \breve{\mathcal{L}}_k(s,u), \qquad \hat{u}_k \;=\; \operatorname*{arg\,min}_{u\in\mathbb{R}} \; \breve{\mathcal{L}}_k(\hat{s}_k, u) \;+\; \lambda_u\big(\Sigma_k^{XX} + \breve{\Sigma}_k^{XX}\big)u^2, \tag{9}$$

$$\hat{c}_k \;=\; \operatorname*{arg\,min}_{c\in\mathbb{C}} \; \sum_{j\in\mathcal{M}_t} \big|c\,\rho_k(j) - R_k^{(j)}\big|^2 + \sum_{j\in\mathcal{A}_t} \big|c\,\rho_k(j) - \breve{R}_k^{(j)}\big|^2 + \lambda_c\big(n + |\mathcal{A}_t|\big)|c|^2. \tag{10}$$

Here, $\breve{\mathcal{L}}_k(s,u) = \mathcal{L}_k(s,u) + \sum_{j\in\mathcal{A}_t}\big|u\,\breve{X}_k^{(j)} - \breve{G}_k^{(j)}\big|^2$, $\breve{\Sigma}_k^{XX} = \sum_{j\in\mathcal{A}_t}|\breve{X}_k^{(j)}|^2$, and $\breve{R}_k^{(j)} = \breve{G}_k^{(j)} - \frac{m_j}{H}\big(s_k Y_k^{(j)} + u_k X_k^{(j)}\big)$ is the residual of the provisional target that the bias of Eq. (10) fits, with the correction scaled by the observed fraction of Eq. (8).

The penalties stay proportional to the energy of the term each coefficient multiplies ($\Sigma_k^{XX} + \breve{\Sigma}_k^{XX}$ for u, $n + |\mathcal{A}_t|$ for c), so the shrinkage rates α_u and α_c are unchanged by POGT.

In Eq. (9), the term added to Stage 1 contains u but no $s_k Y_k$. As a result, the $\mathcal{A}_t$ sums enter only the u row of the 2×2 system, leaving the s row to $\mathcal{M}_t$ alone. Stage 2 of the same equation has no such separation and takes the $\mathcal{A}_t$ term as is. This is because s_k decides how far the forecast is trusted over the whole horizon, a judgment the first m_j steps cannot support because the error grows with lead time, whereas u_k measures how well the input segment explains only the steps POGT covers. Appendix E.2 tests the two alternatives, adding the $\mathcal{A}_t$ term to the s row as well and removing the $s_k Y_k$ share from the u-row target with $\hat{s}_k$ fixed; neither improves on the rule.

The POGT terms are kept apart from the running sums and used only for the solve of the current step. With a running discrete Fourier transform (DFT), each adaptation batch of B windows updates them using only its B new in-progress windows and its B newly matured windows (Algorithm 1), keeping adaptation at $\Theta(H \log H)$ per window (Appendix B). Since there are never more than $H-1$ in-progress windows while the matured samples grow with the stream, the relative weight of POGT, $\mathcal{O}(H/n)$, decays as evidence accumulates, without any schedule.

4 EVALUATION

Comparison Methods. None denotes the frozen-backbone forecast without TTA. We compare the existing TTA methods for time-series forecasting, TAFAS (TFS) (Kim et al., 2025),

Table 1: Per-backbone MSE at each forecasting horizon, averaged over nine datasets (per-dataset results in Appendix G.1).

	iTransformer						DLinear						FreTS					
H	None	TFS	PTS	COS	SWGD	SWiT	None	TFS	PTS	COS	SWGD	SWiT	None	TFS	PTS	COS	SWGD	SWiT
96	.2691	.2610	.2609	.2612	.2546	**.2496**	.3020	.2958	.2952	.2967	.2847	**.2811**	.2706	.2660	.2652	.2677	.2605	**.2577**
192	.3114	.3039	.3038	.3030	.2966	**.2865**	.3397	.3322	.3317	.3335	.3209	**.3079**	.3129	.3073	.3067	.3091	.3013	**.2927**
336	.3658	.3553	.3545	.3521	.3486	**.3348**	.3907	.3815	.3804	.3812	.3624	**.3445**	.3653	.3587	.3576	.3585	.3486	**.3351**
720	.4813	.4675	.4630	.4316	.4640	**.4253**	.5084	.4971	.4924	.4631	.4851	**.4453**	.4838	.4751	.4704	.4399	.4693	**.4323**
Avg.	.3569	.3469	.3456	.3370	.3409	**.3240**	.3852	.3766	.3749	.3686	.3633	**.3447**	.3582	.3518	.3500	.3438	.3449	**.3294**
	PatchTST						OLS						MICN					
H	None	TFS	PTS	COS	SWGD	SWiT	None	TFS	PTS	COS	SWGD	SWiT	None	TFS	PTS	COS	SWGD	SWiT
96	.2641	.2607	.2759	.2586	.2527	**.2497**	.3037	.2969	.2960	.2980	.2867	**.2820**	.2951	.2832	.2858	.2813	.2642	**.2547**
192	.3083	.3032	.3122	.3018	.2961	**.2876**	.3407	.3330	.3320	.3340	.3222	**.3089**	.3475	.3319	.3356	.3316	.3161	**.2994**
336	.3622	.3540	.3607	.3525	.3476	**.3340**	.3907	.3815	.3793	.3810	.3664	**.3469**	.4117	.3875	.3902	.3869	.3738	**.3544**
720	.4901	.4783	.4816	.4422	.4772	**.4318**	.5065	.4959	.4910	.4608	.4840	**.4437**	.5844	.5545	.5500	.5086	.5491	**.4932**
Avg.	.3562	.3490	.3576	.3388	.3434	**.3258**	.3854	.3768	.3746	.3685	.3648	**.3454**	.4097	.3893	.3904	.3771	.3758	**.3504**

Table 2: Computational overhead summary.

Method	Avg. MSE	# Params	Peak GPU (MB)	Adapt. (ms/batch)	Infer. (ms/batch)	Adapt. (MFLOPs)	Infer. (MFLOPs)	E2E (s)
TFS	.3651	151,666	3,065	98.21	9.11	115.60	17.20	14.67
PTS	.3655	6,805	2,864	78.66	10.86	72.10	22.90	16.94
COS	.3556	173,041	2,641	293.67	9.88	151.50	16.60	39.00
SWGD	.3555	**674**	2,568	29.01	0.36	1.81	**1.22**	9.16
SWiT	**.3366**	**674**	**2,565**	**25.67**	**0.32**	**1.80**	**1.22**	**8.75**

PETSA (PTS) (Medeiros et al., 2025), and COSA (COS) (Im & Kwon, 2026), and SWiT-GD (SWGD), which keeps every part of SWiT but replaces the closed-form solve of (s_k, u_k) with one Adam step per batch on the same objective (Appendix F.5). Appendix F describes the evaluation settings. We do not compare DynaTTA (Grover & Etemad, 2025) for the reason given in Appendix F. In the tables of this section, **bold** marks the best value and underline the second best.

4.1 Main Results

Table 1 summarizes the results per backbone, Table 17 per forecasting horizon, and Table 16 in Appendix G.1 per dataset. SWiT is best or second best in almost every scenario. Appendix G.2 reports every setting with its spread over five backbone seeds, and Appendix G.3 the same comparison in MAE, where the ordering holds. Two patterns stand out: 1) SWiT lowers the MSE of every backbone by 8.02–14.46%. 2) SWiT improves accuracy at every forecasting horizon and attains the lowest overall average MSE. This is a 10.29% improvement over None, the largest among the compared methods, against 2.71% for TAFAS, 2.60% for PETSA, and 5.23% for COSA. The gains are not a re-fitting of the instance normalization of Sec. 2.2, which the backbones already apply to every input window: no term of the correction, the level-correcting bias included, carries the reduction on its own (Appendix E.1). The one exception is Exchange, where SWiT is second to COSA. The series is close to a random walk, which a per-time-step correction follows more closely than a per-bin one.

4.2 Computational Overhead

Table 2 summarizes the computational overhead of the existing TTA methods and of SWiT. Adaptation never backpropagates through the backbone, and the closed form removes even the optimizer step that SWiT-GD still takes, so SWiT has the shortest adaptation time (25.67 ms per batch; the cost is derived in Appendix B) and, when correcting in the frequency domain, the shortest inference time (0.32 ms), while reaching the highest accuracy with 674 parameters per channel.

Table 3: Ablation of the two components that use delayed supervision, MSE over 9 datasets × 3 backbones, seed 0. The per-horizon breakdown is in Table 14, Appendix E.5.

H	None	SWiT	w/o POGT	w/o Gate	w/o Both
96	.2809	**.2632**	<u>.2647</u>	.2703	.2741
720	.4950	**.4376**	.4657	<u>.4395</u>	.4740
Avg.	.3690	**.3341**	.3440	<u>.3404</u>	.3514

4.3 Ablation Study

We ablate the composition of the affine terms, the row-separation rule, the closed-form update, the adaptation batch size, the POGT mechanism, and the evidence gate on one backbone per architecture family (DLinear, iTransformer, FreTS) over the same 9 datasets × 4 horizons, timing each cell alone on an idle GPU. Appendix H shows sensitivity to the hyperparameters.

POGT and the Evidence Gate. Table 3 compares SWiT with and without the POGT mechanism and the gate. Disabling POGT worsens the average MSE by 2.95%, disabling the gate by 1.87%, and disabling both by 5.16%. Removing both costs more than the two losses combined; Appendix E.5 traces each loss against the horizon, and Appendix I follows the coefficients and the gate weight over a stream, bin by bin.

Affine Terms. Appendix E.1 ablates the three terms of Eq. (2). The scale alone reaches 5.6% of the 9.5% that the full map reaches, and the two remaining terms each contribute more when the other is present.

Closed Form vs. Gradient-based Update. SWiT-GD, which runs one gradient step on the quadratic objective held in the running sums, matches the best existing TTA method (.3507 vs. .3522 for COSA) at a fraction of the cost, so the per-bin spectral structure remains accurate even under gradient descent (Table 13, Appendix E.4). SWiT is nonetheless more accurate at every batch size, since it reaches the per-bin optimum rather than stepping toward it.

Frequency versus Time Domain. Appendix E.3 compares the same affine corrector in both domains, the FFT being only a change of basis: the frequency-domain diagonal corrector beats its time-domain counterpart on 51 of 72 cells and matches a dense one with $\mathcal{O}(H^2)$ parameters.

Adaptation Batch Size. The batch B sets how often the coefficients are re-solved. Adapting more often helps only slightly while the wall-clock time grows much faster, so we use 48 (Table 13).

Online Learning Methods. Appendix J compares SWiT with FSNet (Pham et al., 2022) and PROCEED (Zhao & Shen, 2025), which update the backbone at test time. PROCEED is best when the backbone is trained on a fifth of the series, SWiT under the protocol of Sec. 4.

Row Separation. Appendix E.2 tests where the POGT statistics enter. Restricting them to the u equation, as Sec. 3.6 does, lowers MSE by 9.5% over None against 8.8% when they enter both, and the variant that removes the $s_k Y_k$ share from the u-row target wins only 31 of 108 cells.

5 Conclusion

We presented SWiT, which corrects a frozen backbone's forecast per frequency bin and solves each adaptation batch in closed form: Parseval's theorem separates the objective into per-bin quadratics whose sufficient statistics running sums carry, so the optimum is read off without gradient descent. The causal protocol, splicing rule for POGT, and evidence gate let the adapter use fresh but incomplete labels without admitting any unobserved value. On 216 settings SWiT attains the lowest average MSE among TTA methods, 10.3% below the frozen backbone, with 674 parameters per channel and the shortest adaptation time.

SWiT processes channels independently, targets point forecasts under squared error, keeps its hyperparameters fixed rather than scheduled, and is compared only against TTA methods that keep the backbone frozen. Its gains are smallest at short horizons, where a good backbone leaves little spectral structure; a cross-channel solve and other objectives are natural next steps.

Ethics Statement

This work studies test-time adaptation for time-series forecasting on nine publicly available benchmark datasets that contain no personal or sensitive information. It involves no human subjects, data collection, or deployment. We do not foresee negative societal impacts specific to the method beyond those shared by forecasting systems in general.

Reproducibility Statement

We list all SWiT hyperparameters and those of the compared methods, the seed policy, the data splits, and the hardware in Appendix F. Appendix B presents the complete inference and adaptation loop as Algorithm 1, and Appendix C derives the closed-form solutions. An anonymized implementation, including the baselines and a quick-start script that reproduces one setting end to end, is available at `https://anonymous.4open.science/r/SWiT-808C`.

Statement on the Use of Large Language Models

We used large language models for three purposes: translating draft sections of the manuscript into English, editing the text for readability, and assisting with writing the experiment code that implements the method, the baselines, and the evaluation pipeline. The authors reviewed and tested all generated code, and every reported number comes from runs the authors executed. They were not used to generate data, to develop the theoretical framework or to formulate or prove the mathematical claims, to propose hypotheses, to design the methodology or the experiments, to clean or reformat datasets, or to interpret the results. All technical content is the authors' own.

References

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020.

Weiqi Chen, Zhaoyang Zhu, Yifan Zhang, Lefei Shen, Linxiao Yang, Qingsong Wen, and Liang Sun. Learning to extrapolate and adjust: Two-stage meta-learning for concept drift in online time series forecasting. In *IJCAI*, pp. 4869–4877, 2025.

Tao Dai, Beiliang Wu, Peiyuan Liu, Naiqi Li, Xue Yuerong, Shu-Tao Xia, and Zexuan Zhu. Ddn: Dual-domain dynamic normalization for non-stationary time series forecasting. *Advances in Neural Information Processing Systems*, 37:108490–108517, 2024.

Wei Fan, Pengyang Wang, Dongkun Wang, Dongjie Wang, Yuanchun Zhou, and Yanjie Fu. Dish-ts: a general paradigm for alleviating distribution shift in time series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 7522–7529, 2023.

Shivam Grover and Ali Etemad. Shift-aware test time adaptation and benchmarking for time-series forecasting. In *Second Workshop on Test-Time Adaptation: Putting Updates to the Test! at ICML 2025*, 2025.

Tao Hong, Pierre Pinson, Shu Fan, Hamidreza Zareipour, Alberto Troccoli, and Rob J Hyndman. Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond, 2016.

Jeonghwan Im and Hyuk-Yoon Kwon. Cosa: Context-aware output-space adapter for test-time adaptation in time series forecasting. In *The Fourteenth International Conference on Learning Representations*, 2026.

HyunGi Kim, Siwon Kim, Jisoo Mok, and Sungroh Yoon. Battling the non-stationarity in time series forecasting via test-time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17868–17876, 2025.

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.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.

Andrey Kolmogorov. Interpolation and extrapolation of stationary random sequences. *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya*, 5:3, 1941.

Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International journal of forecasting*, 37(4): 1748–1764, 2021.

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, 2022.

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

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

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

Heitor R Medeiros, Hossein Sharifi-Noghabi, Gabriel L Oliveira, and Saghar Irandoust. Accurate parameter-efficient test-time adaptation for time series forecasting. *arXiv preprint arXiv:2506.23424*, 2025.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Zijie Pan, Yushan Jiang, Dongjin Song, Sahil Garg, Kashif Rasul, Anderson Schneider, and Yuriy Nevmyvaka. Structural knowledge informed continual multivariate time series forecasting, 2024.

Quang Pham, Chenghao Liu, Doyen Sahoo, and Steven CH Hoi. Learning fast and slow for online time series forecasting. 2022.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.

William Toner and Luke Darlow. An analysis of linear time series forecasting models. *arXiv preprint arXiv:2403.14587*, 2024.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

Huiqiang Wang, Jian Peng, Feihu Huang, Jince Wang, Junhui Chen, and Yifei Xiao. Micn: Multi-scale local and global context modeling for long-term series forecasting. In *The eleventh international conference on learning representations*, 2023.

Norbert Wiener. *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*. The MIT press, 1949.

Zhijian Xu, Ailing Zeng, and Qiang Xu. Fits: Modeling time series with $10k$ parameters. In *International Conference on Learning Representations*, volume 2024, pp. 26295–26318, 2024.

Kun Yi, Qi Zhang, Wei Fan, Shoujin Wang, Pengyang Wang, Hui He, Ning An, Defu Lian, Longbing Cao, and Zhendong Niu. Frequency-domain mlps are more effective learners in time series forecasting. volume 36, pp. 76656–76679, 2023.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

YiFan Zhang, Weiqi Chen, Zhaoyang Zhu, Dalin Qin, Liang Sun, Xue Wang, Qingsong Wen, Zhang Zhang, Liang Wang, and Rong Jin. Addressing concept shift in online time series forecasting: Detect-then-adapt. *arXiv preprint arXiv:2403.14949*, 2024.

Lifan Zhao and Yanyan Shen. Proactive model adaptation against concept drift for online time series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 2020–2031, 2025.

Tian Zhou, Ziqing Ma, Qingsong Wen, Liang Sun, Tao Yao, Wotao Yin, Rong Jin, et al. Film: Frequency improved legendre memory model for long-term time series forecasting. *Advances in neural information processing systems*, 35:12677–12690, 2022a.

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*, pp. 27268–27286. PMLR, 2022b.

A NOTATIONS

Table 4 lists the symbols used in Sec. 3 and in the appendices. Channels are processed independently, so every symbol refers to one channel. The bin index k and the window index (j) are omitted when clear from context.

Table 4: Notation.

Symbol	Meaning	Shape
Problem setting		
W, H, B	Lookback length, forecasting horizon, adaptation batch size	scalars
j, t, $m_j = t - j$	Forecast origin of a window, current time index, steps elapsed since j	scalars
j_0	Forecast origin of the first window in the current batch (Algorithm 1)	scalar
f_θ, ϕ_t	Frozen backbone, adapter with parameters at time t	–
$\mathbf{x}_j$	Lookback window of window j	$\mathbb{R}^W$
$\hat{\mathbf{y}}_j^{(0)}$, $\mathbf{y}_j$, $\hat{\mathbf{y}}_j$	Base prediction $f_\theta(\mathbf{x}_j)$, ground truth, corrected prediction, $y_j[\tau]$ is the τ-th entry, $\tau = 0, \dots, H-1$	$\mathbb{R}^H$
$\mathbf{h}_j$	Input history: the H steps preceding origin j, in time order, $h_j[0]$ at $j-H$ and $h_j[H-1]$ at $j-1$	$\mathbb{R}^H$
$\mathcal{M}_t$, $n = \lvert\mathcal{M}_t\rvert$	Matured windows ($m_j \geq H$) and their count	set, scalar
$\mathcal{L}(\phi_t)$	Expected squared error under the test distribution	scalar
$\mathcal{A}_t$, $\tilde{n} = \lvert\mathcal{A}_t\rvert$	In-progress windows ($1 \leq m_j < H$) and their count	set, scalar
$\mathcal{B}$	Buffer of emitted windows whose targets have not matured (Algorithm 1)	set
Spectral correction		
K, k	Number of rFFT bins $\lfloor H/2 \rfloor + 1$, bin index ($k = 1$ is DC)	scalars
$Y_k^{(j)}$, $X_k^{(j)}$, $G_k^{(j)}$	k-th rFFT coefficient of $\hat{\mathbf{y}}_j^{(0)}$, $\mathbf{h}_j$, $\mathbf{y}_j$, stacked as $\mathbf{Y}_j, \mathbf{X}_j, \mathbf{G}_j \in \mathbb{C}^K$	$\mathbb{C}$
s_k, u_k	Scaling factor and input factor of bin k	$\mathbb{R}$
c_k	Rotational bias of bin k	$\mathbb{C}$
$\rho_k(j)$	Phase rotation factor $e^{\mathrm{i}\,2\pi(k-1)j/H}$, unit modulus	$\mathbb{C}$
$\tilde{Y}_k^{(j)}$, $\tilde{\mathbf{y}}_j$	Corrected coefficient $s_k Y_k^{(j)} + u_k X_k^{(j)} + c_k \rho_k(j)$, its irFFT	$\mathbb{C}$, $\mathbb{R}^H$
δ_s, δ_u	Clipping ranges, $s_k \in [1-\delta_s, 1+\delta_s]$, $u_k \in [-\delta_u, \delta_u]$	scalars
Adaptation		
$\hat{s}_k, \hat{u}_k, \hat{c}_k$	Minimizers of Eqs. (6) and (7); clipping them gives s_k, u_k	$\mathbb{R}, \mathbb{R}, \mathbb{C}$
$\tilde{u}_k$	Unpenalized input factor with $\hat{s}_k$ fixed ($\hat{u}_k = (1-\alpha_u)\tilde{u}_k$)	$\mathbb{R}$
$R_k^{(j)}$	Stage-1 residual $G_k^{(j)} - s_k Y_k^{(j)} - u_k X_k^{(j)}$	$\mathbb{C}$
Σ_k^{AB}	Running sum $\sum_{j \in \mathcal{M}_t} \mathrm{Re}(\overline{A_k^{(j)}} B_k^{(j)})$, $AB \in \{YY, XY, XX, YG, XG\}$	$\mathbb{R}$
P_k, Q_k, T_k	Derotated running sums $\sum_{j \in \mathcal{M}_t} \overline{\rho_k(j)}\,(Y_k^{(j)}, X_k^{(j)}, G_k^{(j)})$	$\mathbb{C}$
$\lambda_u, \lambda_c, \alpha_u, \alpha_c$	Energy-proportional ridge strengths, shrinkage rates $\lambda/(1+\lambda)$	scalars
$\mathbf{A}$, $\mathbf{b}$	Normal-equation matrix and right-hand side of stage 1	$\mathbb{R}^{2\times 2}$, $\mathbb{R}^2$
S_k	Derotated residual sum $T_k - s_k P_k - u_k Q_k + \check{T}_k$	$\mathbb{C}$
POGT splicing		
$\check{\mathbf{g}}_j$, $\check{G}_k^{(j)}$	Provisional target (observed prefix, zero-filled) and its rFFT coefficient	$\mathbb{R}^H$, $\mathbb{C}$
$\tilde{G}_k^{(j)}$	DC-bin provisional target with the unobserved part filled by $\hat{\mathbf{y}}_j^{(0)}$	$\mathbb{C}$
$\check{X}_k^{(j)}$	Ratio-weighted input coefficient $(m_j/H)\,X_k^{(j)}$	$\mathbb{C}$
$\check{\Sigma}_k^{XX}$, $\check{\Sigma}_k^{XG}$	POGT sums $\sum_{j \in \mathcal{A}_t} \lvert\check{X}_k^{(j)}\rvert^2$, $\sum_{j \in \mathcal{A}_t} \mathrm{Re}(\overline{\check{X}_k^{(j)}} \check{G}_k^{(j)})$	$\mathbb{R}$
$\check{R}_k^{(j)}$, $\check{T}_k$	POGT residual coefficient and its derotated sum $\sum_{j \in \mathcal{A}_t} \overline{\rho_k(j)}\,\check{R}_k^{(j)}$	$\mathbb{C}$
Evidence gate		
$C_k^{(j)}$	Correction $\tilde{Y}_k^{(j)}$ frozen at emission	$\mathbb{C}$
$\Delta_k^{(j)}$, $D_k^{(j)}$	Applied displacement $C_k^{(j)} - Y_k^{(j)}$, target displacement $G_k^{(j)} - Y_k^{(j)}$	$\mathbb{C}$
p_k, d_k	Gate accumulators $\sum_{j \in \mathcal{M}_t} \mathrm{Re}(\overline{\Delta_k^{(j)}} D_k^{(j)})$, $\sum_{j \in \mathcal{M}_t} \lvert\Delta_k^{(j)}\rvert^2$, $\check{p}_k, \check{d}_k$ their POGT counterparts at cold start	$\mathbb{R}$
w_k	Gate weight $\mathrm{clip}(p_k/d_k, 0, 1)$	$[0, 1]$
$E_k^{(j)}$	Emitted coefficient $Y_k^{(j)} + w_k \Delta_k^{(j)}$	$\mathbb{C}$
Operators		
$\mathrm{rFFT}(\cdot)$, $\mathrm{irFFT}(\cdot)$	Real FFT of a length-H sequence and its inverse	$\mathbb{R}^H \to \mathbb{C}^K$, $\mathbb{C}^K \to \mathbb{R}^H$
$\overline{z}$, $\mathrm{Re}(z)$, $\lvert z\rvert$	Complex conjugate, real part, modulus of z	$\mathbb{C} \to \mathbb{C}$, $\mathbb{R}$, $\mathbb{R}_{\geq 0}$
$\mathrm{clip}_{[a,b]}(x)$	$\min(\max(x, a), b)$, also written $\mathrm{clip}(x, a, b)$	$\mathbb{R} \to [a, b]$
$\lVert\mathbf{v}\rVert_2$	Euclidean norm of a vector	$\mathbb{R}^H \to \mathbb{R}_{\geq 0}$

B ALGORITHM OF SWIT INFERENCE AND ADAPTATION

Algorithm 1 is the complete online loop of SWiT, which corrects the backbone forecast with a per-bin affine map under delayed supervision. A correction is applied only to the extent that out-of-sample evidence warrants it. Each batch runs two stages, evaluation and adaptation. The corrected outputs

are first emitted with the coefficients left by the previous batch, and the coefficients are then re-solved from the GT that has already arrived. The backbone stays frozen; the only quantities updated are the per-bin coefficients (s, u, c) and the gate weight w. Every update is closed-form, so there is no learning rate, step count, or optimizer state, and the clipping ranges and shrinkage rates are fixed constants shared by all evaluation settings. There are two kinds of state. The running sums, the intercept sums, and the gate ledger, all filled by matured windows, persist for the whole stream. The POGT statistics built from in-progress windows are rebuilt at every batch, used for that batch's solve, and discarded.

Cold Start. At the start of the stream, before any window has matured, s_k is fixed to 1 and u_k, c_k are estimated from POGT splicing alone. Since no window has matured, the gate accumulators p_k, d_k are empty, so the evidence gate opens only the DC bin and applies only level correction during the cold period. Once matured windows accumulate, the gate corrects every bin through its scoring.

The pre-gate correction is stored because mixing in the gate weight would make the gate score its own output, a circularity. We store the correction in force at emission time, rather than the latest coefficients at maturity, because the latter have already learned from the steps overlapping that window, and scoring with them would be in-sample evaluation.

Notation follows Sec. 3. The buffer $\mathcal{B}$ is the set of windows that have been emitted but have not yet matured. Its subset with $1 \le m_j < H$ is the set of in-progress windows $\mathcal{A}$.

Time Complexity. We count the cost of Algorithm 1 per channel and per adaptation batch of B windows, with $K = \lfloor H/2 \rfloor + 1 = \Theta(H)$ bins. Table 5 lists the cost of each stage.

Table 5: Cost of Algorithm 1 per channel and per batch of B windows. Line numbers refer to Algorithm 1, $\breve{n} = |\mathcal{A}_t| \le H - 1$.

Stage	Lines	Cost per batch
Evaluation: three transforms and the suffix sums, per window	5–7	$\mathcal{O}(B\,H \log H)$
Running DFT and moment update, per newly observed step	9–12	$\mathcal{O}(B\,K)$
Matured windows: one rFFT and the running-sum updates, per window	13–19	$\mathcal{O}(B\,H \log H)$
POGT sums from the moments, and their DC-bin overrides	20–25	$\mathcal{O}(K + \breve{n})$
Closed-form solve, shrinkage, and clipping	26–27	$\mathcal{O}(K)$
Intercept POGT sum: the vector $\mathbf{g}$ and one rFFT	28–30	$\mathcal{O}(H \log H + \breve{n})$
Intercept and gate	31–39	$\mathcal{O}(K + \breve{n})$
Total		$\mathcal{O}\big((B+1)\,H \log H\big)$

Per batch. The per-window stages cost $\mathcal{O}(H \log H)$ each and are executed B times; the running DFT costs $\mathcal{O}(K)$ per newly observed step and there are B such steps per batch, and every stage that runs once per batch costs at most $\mathcal{O}(H \log H)$, since K, $\breve{n}$, and H are all $\mathcal{O}(H)$. The total is therefore $\mathcal{O}\big((B+1)\,H \log H\big)$ per batch, that is, $\Theta(H \log H)$ per window for any $B \ge 1$, and the constant does not depend on B. A batch of B windows performs $3B + 1$ transforms of length H and B more on the matured side. The key is that no stage iterates over the in-progress set $\mathcal{A}_t$ with a transform inside the loop. The zero-padded prefix transforms $\breve{G}^{(j)}$ of Eq. (8) are never materialized. Line 21 expresses each of them as a difference of two states of the running DFT Φ, and lines 22–23 and 29 obtain the three POGT sums that the solve needs, $\breve{\Sigma}^{XX}$, $\breve{\Sigma}^{XG}$, and $\breve{T}$, from the moments A_q, U_q, W_q, $\breve{P}_q$, $\breve{Q}_q$, which the procedure MOMENTS updates in $\mathcal{O}(K)$ when a window enters or leaves $\mathcal{A}_t$. The only loops over $\mathcal{A}_t$ that remain, lines $24 \sim 25$, 30, and 35, touch the DC bin alone and cost $\mathcal{O}(\breve{n}) = \mathcal{O}(H)$.

Memory. The buffer $\mathcal{B}$ holds every window whose ground truth is incomplete, at most $H - 1 + B$ windows, each with three coefficient vectors of length K and the suffix sums Ψ of length H, so the working set is $\Theta(HK) = \Theta(H^2)$ per channel; the running sums and moments add $\mathcal{O}(K)$. This buffer is what a method that adapts on partially observed ground truth must keep in any case, and it is the source of the peak-memory figure in Table 2.

The direct evaluation. Evaluating the POGT sums directly costs $\mathcal{O}(H^2 \log H)$ per batch through two paths. First, the $\breve{G}^{(j)}$ of Eq. (8) is an rFFT of a zero-padded prefix for each of the $\breve{n} = \mathcal{O}(H)$

Algorithm 1 Inference and adaptation loop of SWiT (per channel, bin index k omitted, $k = 1$ is the DC bin, where $\rho \equiv 1$). Costs in braces are per batch, except in the evaluation stage and the matured loop, where they are per window; every line is $\mathcal{O}(K)$ unless noted. The total is $\Theta(H \log H)$ per window.

Require: Frozen backbone f_θ, input history $\mathbf{h}_j$, horizon H, batch size B, clip ranges $s \in [0.5, 1.5]$, $|u| \le 0.3$, shrinkage $\alpha_u = \alpha_c = 1/2$ $(\lambda_u = \alpha_u/(1-\alpha_u) = 1)$
Ensure: Corrected predictions $\{\hat{\mathbf{y}}_j\}$

$\Sigma^{YY}, \Sigma^{XX}, \Sigma^{XY}, \Sigma^{YG}, \Sigma^{XG}, P, Q, T, n, p, d \leftarrow 0, \quad \Phi \leftarrow 0; \quad \mathcal{B} \leftarrow \emptyset, \quad (s, u, c, w) \leftarrow (1, 0, 0, 0)$
$\{A_q\}_{q=0}^{2}, \{U_q, W_q, \breve{P}_q, \breve{Q}_q\}_{q=0}^{1}, \breve{n} \leftarrow 0$ {q-th order moments in j' over $\mathcal{A}$, and $\breve{n} = |\mathcal{A}|$, all maintained by MOMENTS}
for each batch of B windows $j = j_0, \dots, j_0 + B - 1$ in the test stream **do**
  (i) Inference {emit with the state left by the previous batch, $\mathcal{O}(H \log H)$}
  $Y \leftarrow \mathrm{rFFT}\big(f_\theta(\mathbf{x}_j)\big), \quad X \leftarrow \mathrm{rFFT}(\mathbf{h}_j), \quad C \leftarrow sY + uX + c\,\rho(j)$
  $\Psi[m] \leftarrow \sum_{\tau \ge m} \hat{y}_j^{(0)}[\tau], \quad 0 \le m < H$ {suffix sums of the backbone prediction, for the DC bin, $\mathcal{O}(H)$ once}
  $\hat{\mathbf{y}}_j \leftarrow \mathrm{irFFT}\big(Y + w\,(C - Y)\big), \quad \mathcal{B} \leftarrow \mathcal{B} \cup \{(j, Y, X, C, \Psi)\}$ {freeze C, hold the window until it matures}
  (ii) Adaptation {GT up to $j_0 - 1$: window j' has $m_{j'} = j_0 - j'$ elapsed steps, $\min(m_{j'}, H)$ of them observed}
  for each newly observed step τ **do**
    $\Phi_\tau \leftarrow \Phi, \quad \Phi \mathrel{+}= \overline{\rho(\tau)}\, y[\tau]$ {running DFT, Φ_τ kept for the current batch only, with $\Phi_{j_0} \equiv \Phi$ after the loop, Φ_1 grows linearly in t, so accumulate it in double precision}
    MOMENTS$(\tau, +1, \Phi_\tau), \quad$ retain $\Phi_{\tau,1}$ {window τ enters $\mathcal{A}$, only the DC scalar of Φ_τ is kept, and only while $\tau \in \mathcal{A}$}
  end for
  for $(j', Y', X', C', \Psi') \in \mathcal{B}$ with $m_{j'} \ge H$, then drop it (and $\Phi_{j',1}$) from $\mathcal{B}$ **do**
    $G' \leftarrow \mathrm{rFFT}(\mathbf{y}_{j'})$ {matured GT running sum update, $\mathcal{O}(H \log H)$}
    $(\Sigma^{YY}, \Sigma^{XX}) \mathrel{+}= (|Y'|^2, |X'|^2), \quad (\Sigma^{XY}, \Sigma^{YG}, \Sigma^{XG}) \mathrel{+}= \mathrm{Re}\big(\overline{Y'}X',\ \overline{Y'}G',\ \overline{X'}G'\big)$
    $(P, Q, T) \mathrel{+}= \overline{\rho(j')}\,(Y', X', G'), \quad n \mathrel{+}= 1$ {intercept sums}
    $d \mathrel{+}= |C' - Y'|^2, \quad p \mathrel{+}= \mathrm{Re}\big(\overline{(C' - Y')}\,(G' - Y')\big)$ {gate ledger}
    MOMENTS$\big(j', -1, \Phi_{j'+H} - \overline{\rho(j')}\,G'\big)$ {j' leaves $\mathcal{A}$; $\Phi_{j'} = \Phi_{j'+H} - \overline{\rho(j')}\,G'$}
  end for
  $t \leftarrow j_0$ {$\mathcal{A} = \{j' \in \mathcal{B} : 1 \le m_{j'} < H\}, \quad \breve{n} = |\mathcal{A}|; \quad m_{j'} = t - j'$ for every $j' \in \mathcal{A}$}
  $\breve{G}' \equiv \mathrm{rFFT}\big(\mathbf{y}_{j'}[0{:}m_{j'}] \,\|\, \mathbf{0}\big) = \rho(j')\,(\Phi - \Phi_{j'}), \quad \breve{X}' \equiv \frac{m_{j'}}{H} X'$ {POGT terms for $k \ge 2$, never materialized, the DC bin is line 24}
  $\breve{\Sigma}^{XX} \leftarrow \big(t^2 A_0 - 2tA_1 + A_2\big)/H^2$ {$= \sum_{j' \in \mathcal{A}} |\breve{X}'|^2$}
  $\breve{\Sigma}^{XG} \leftarrow \big[\mathrm{Re}\big(\Phi\,(tU_0 - U_1)\big) - (tW_0 - W_1)\big]/H$ {$= \sum_{j' \in \mathcal{A}} \mathrm{Re}(\overline{\breve{X}'}\,\breve{G}')$}
  $\breve{G}'_1 \leftarrow (\Phi_1 - \Phi_{j',1}) + \Psi'[m_{j'}], \quad \breve{X}'_1 \leftarrow X'_1 \quad$ for $j' \in \mathcal{A}$ {DC bin overrides line 21: GT prefix from the running DFT ($\rho \equiv 1$) plus backbone-fill tail from Ψ', and no $\frac{m_{j'}}{H}$ weight since no mass is lost, $\mathcal{O}(1)$ each}
  $\breve{\Sigma}^{XX}_1 \leftarrow [A_0]_1, \quad \breve{\Sigma}^{XG}_1 \leftarrow \sum_{j' \in \mathcal{A}} \mathrm{Re}\big(\overline{X'_1}\,\breve{G}'_1\big)$ {DC bin of the two sums above, $[A_0]_1 = \sum_{j' \in \mathcal{A}} |X'_1|^2$ already, $\mathcal{O}(\breve{n})$}
  $(\hat{s}, \tilde{u}) \leftarrow$ solution of $\begin{pmatrix} \Sigma^{YY} & \Sigma^{XY} \\ \Sigma^{XY} & \Sigma^{XX} + \breve{\Sigma}^{XX} \end{pmatrix} \begin{pmatrix} s \\ u \end{pmatrix} = \begin{pmatrix} \Sigma^{YG} \\ \Sigma^{XG} + \breve{\Sigma}^{XG} \end{pmatrix}$ {POGT in the u row only}
  $\hat{u} \leftarrow (1 - \alpha_u)\,\tilde{u}, \quad s \leftarrow \mathrm{clip}(\hat{s}), \quad u \leftarrow \mathrm{clip}(\hat{u})$ {ridge penalty on u with $\hat{s}$ fixed, then clip}
  $g[\tau \bmod H] \leftarrow (\tau - t + \breve{n} + 1)\,y[\tau]$ for $t - \breve{n} \le \tau < t$ {# windows of $\mathcal{A}$ containing τ, $\mathcal{O}(H)$}
  $\breve{T} \leftarrow \mathrm{rFFT}(\mathbf{g}) - \frac{s}{H}\big(t\breve{P}_0 - \breve{P}_1\big) - \frac{u}{H}\big(t\breve{Q}_0 - \breve{Q}_1\big)$ {$= \sum_{j' \in \mathcal{A}} \overline{\rho(j')}\,\breve{R}', \quad \breve{R}' \equiv \breve{G}' - \frac{m_{j'}}{H}(sY' + uX')$, $\mathcal{O}(H \log H)$}
  $\breve{T}_1 \leftarrow \sum_{j' \in \mathcal{A}} \breve{G}'_1 - s\,[\breve{P}_0]_1 - u\,[\breve{Q}_0]_1$ {DC bin overrides line 29: $\breve{R}'_1 = \breve{G}'_1 - (sY'_1 + u\breve{X}'_1)$ with line 24, unweighted, $\mathcal{O}(\breve{n})$}
  $c \leftarrow (1 - \alpha_c)\big(T - sP - uQ + \breve{T}\big)/(n + \breve{n})$
  if $n \ge 2$ **then**
    $w \leftarrow \mathrm{clip}\big(p/\max(d,\ 10^{-30}),\ 0,\ 1\big)$ {a bin whose correction never moved has $p = d = 0$, so $w = 0$ and the backbone passes}
  else if $\breve{n} \ge 1$ **then**
    $\breve{d}_1 \leftarrow \sum_{j' \in \mathcal{A}} |C'_1 - Y'_1|^2, \quad \breve{p}_1 \leftarrow \sum_{j' \in \mathcal{A}} \mathrm{Re}\big(\overline{(C'_1 - Y'_1)}\,(\breve{G}'_1 - Y'_1)\big)$ {POGT gate ledger, DC bin only, $\mathcal{O}(\breve{n})$}
    $w_1 \leftarrow \mathrm{clip}\big((p + \breve{p})_1/\max\big((d + \breve{d})_1,\ 10^{-30}\big),\ 0,\ 1\big), \quad w_{k \ge 2} \leftarrow 0$ {cold start: DC bin only}
  else
    $w \leftarrow 0$ {no evidence: the backbone passes through}
  end if
end for

procedure MOMENTS(j', σ, φ) {$\sigma = +1$: enter, -1: leave; (Y', X') of window j' from $\mathcal{B}$, $\varphi = \Phi_{j'}$}
  $\breve{n} \mathrel{+}= \sigma$ {$\breve{n} = |\mathcal{A}|$}
  $A_q \mathrel{+}= \sigma\, {j'}^q\, |X'|^2, \quad q \in \{0, 1, 2\}; \quad U_q \mathrel{+}= \sigma\, {j'}^q\, \overline{X'}\rho(j'), \quad W_q \mathrel{+}= \sigma\, {j'}^q\, \mathrm{Re}\big(\overline{X'}\rho(j')\,\varphi\big), \quad q \in \{0, 1\}$
  $\breve{P}_q \mathrel{+}= \sigma\, {j'}^q\, \overline{\rho(j')}\, Y', \quad \breve{Q}_q \mathrel{+}= \sigma\, {j'}^q\, \overline{\rho(j')}\, X', \quad q \in \{0, 1\}$

in-progress windows, and the prefix of every one of them grows at every batch, so all of them are recomputed, $\mathcal{O}(H)$ transforms per batch. Second, an earlier implementation formed the residual $\breve{R}^{(j)}$ that the intercept fits by transforming the correction $sY^{(j)} + uX^{(j)}$ to the time domain, masking the unobserved steps, and transforming back, which is one more pair of transforms per in-progress window. Algorithm 1 removes the first path with the running DFT and the second by defining the residual as $\breve{R}^{(j)} = \breve{G}^{(j)} - \frac{m_j}{H}\big(sY^{(j)} + uX^{(j)}\big)$: the correction is scaled by the observed fraction

instead of being masked, the same factor that Eq. (8) applies to the input coefficients $\breve{X}^{(j)}$, and it lets $\breve{T}$ be assembled from the moments in line 29. All results in Sec. 4 use this residual. The implementation evaluates the POGT sums by two routes that compute the same sums, the running DFT of the moments and a direct transform over the in-progress set, and takes the running DFT where the number of bins makes it the cheaper of the two; measured differences between the routes are at the 10^{-7} level in MSE.

C Closed-Form Solutions

We give the solutions of the two objectives of Sec. 3.5, Eqs. (6) and (7), with the POGT term of Sec. 3.6, Eq. (9), added, and of the gate weight of Sec. 3.4, Eq. (4). The bin subscript k is omitted.

Running Sums. The only statistics the solutions need are plain sums over past matured windows and the POGT statistics of the windows currently in progress. For (s, u), these are the five real sums:

$$\Sigma^{AB} = \sum_{i \in \mathcal{M}_t} \mathrm{Re}\big(\overline{A^{(i)}} B^{(i)}\big), \qquad (A, B) \in \{YY, XY, XX, YG, XG\}, \tag{11}$$

For c, the three complex sums as follows.

$$P = \sum_{i \in \mathcal{M}_t} \overline{\rho(i)}\, Y^{(i)}, \qquad Q = \sum_{i \in \mathcal{M}_t} \overline{\rho(i)}\, X^{(i)}, \qquad T = \sum_{i \in \mathcal{M}_t} \overline{\rho(i)}\, G^{(i)}, \tag{12}$$

Each sum is maintained by adding a window's contribution once, when it matures. The POGT sums over in-progress windows are not accumulated but rebuilt at every time step.

$$\breve{\Sigma}^{XX} = \sum_{j \in \mathcal{A}_t} |\breve{X}^{(j)}|^2, \qquad \breve{\Sigma}^{XG} = \sum_{j \in \mathcal{A}_t} \mathrm{Re}\big(\overline{\breve{X}^{(j)}} \breve{G}^{(j)}\big), \qquad \breve{T} = \sum_{j \in \mathcal{A}_t} \overline{\rho(j)}\, \breve{R}^{(j)}, \tag{13}$$

On the DC bin, $\breve{G}^{(j)}$ filled with the backbone forecast replaces $\breve{G}^{(j)}$, and $X^{(j)}$ replaces $\breve{X}^{(j)}$ (Sec. 3.6).

Scale and Input Coefficients. Adding the POGT term to Eq. (6) gives the two stages.

$$J_1(s, u) = \sum_{i \in \mathcal{M}_t} \big| s\, Y^{(i)} + u\, X^{(i)} - G^{(i)} \big|^2 + \sum_{j \in \mathcal{A}_t} \big| u\, \breve{X}^{(j)} - \breve{G}^{(j)} \big|^2, \tag{14}$$

$$J_2(u\,;\, \hat{s}) = J_1(\hat{s}, u) + \lambda_u \big(\Sigma^{XX} + \breve{\Sigma}^{XX}\big) u^2. \tag{15}$$

Expanding the squared moduli and summing yields a quadratic in (s, u) that is written in terms of the sums above alone.

$$\begin{aligned} J_1(s, u) &= \begin{pmatrix} s & u \end{pmatrix} \mathbf{A} \begin{pmatrix} s \\ u \end{pmatrix} - 2 \begin{pmatrix} s & u \end{pmatrix} \mathbf{b} + \mathrm{const}, \\ \mathbf{A} &= \begin{pmatrix} \Sigma^{YY} & \Sigma^{XY} \\ \Sigma^{XY} & \Sigma^{XX} + \breve{\Sigma}^{XX} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} \Sigma^{YG} \\ \Sigma^{XG} + \breve{\Sigma}^{XG} \end{pmatrix}. \end{aligned} \tag{16}$$

Because the POGT term contains no s, it enters only the $(2, 2)$ element of $\mathbf{A}$ and the second component of $\mathbf{b}$. This is the row separation. $\mathbf{A}$ is a sum of per-window positive semidefinite matrices, hence positive semidefinite, and positive definite on every bin whose $\{Y^{(j)}\}$ and $\{X^{(j)}\}$ are not collinear, so Stage 1 has the unique minimizer.

$$\mathbf{A} \begin{pmatrix} \hat{s} \\ \tilde{u} \end{pmatrix} = \mathbf{b}. \tag{17}$$

Stage 2 is scalar. Differentiating J_2 and using the second row of Eq. (17). $\big(\Sigma^{XX} + \breve{\Sigma}^{XX}\big)\tilde{u} = \Sigma^{XG} + \breve{\Sigma}^{XG} - \hat{s}\, \Sigma^{XY}$, gives

$$\hat{u} = \frac{\Sigma^{XG} + \breve{\Sigma}^{XG} - \hat{s}\, \Sigma^{XY}}{(1 + \lambda_u)\big(\Sigma^{XX} + \breve{\Sigma}^{XX}\big)} = (1 - \alpha_u)\, \tilde{u}, \qquad \alpha_u = \frac{\lambda_u}{1 + \lambda_u}, \tag{18}$$

So $\Sigma^{XX} + \breve{\Sigma}^{XX}$ cancels, every bin is shrunk by the same fraction, and $\hat{s}$ is left untouched by λ_u. As a 2×2 problem, Stage 1 is solved with a single determinant division. For bins whose determinant is close to 0, add a small constant to the diagonal and the determinant for stability (Algorithm 1).

Why Two Stages. We do not solve the three coefficients jointly because s_k, u_k are real while c_k is complex. A joint solve would require splitting c_k into real and imaginary parts and setting up a 4×4 linear system with four unknowns for every bin. Solved in stages, stage 1 is a 2×2 linear system with two unknowns, and stage 2 needs no matrix at all because $|\rho_k| = 1$. The answer is the mean of the derotated residuals.

Wiener Correspondence. Eq. (6) minimizes the squared error over the samples in $\mathcal{M}_t$ and places no distributional assumption on $\{Y_k^{(j)}, X_k^{(j)}, G_k^{(j)}\}$. If the samples are regarded as drawn from a wide-sense stationary process and the running sums are replaced by the population moments of that process, then at $\lambda_u = 0$ the pair $(\hat{s}_k, \hat{u}_k)$ is the Wiener filter that estimates G_k from the two channels Y_k and X_k in bin k (Wiener, 1949; Kolmogorov, 1941). Under a further joint-Gaussian assumption, this linear estimator is the minimum mean-squared-error estimator. We borrow the name because the estimator coincides with the Wiener filter under those assumptions. The computation itself is plain least squares on the observed samples.

Rotational Bias. Fix the applied (s, u) and write the residuals as $R^{(i)} = G^{(i)} - sY^{(i)} - uX^{(i)}$ for $i \in \mathcal{M}_t$ and $\breve{R}^{(j)}$ for $j \in \mathcal{A}_t$. Adding the POGT term to Eq. (7) gives the objective as follows.

$$J(c) = \sum_{i \in \mathcal{M}_t} \left|c\,\rho(i) - R^{(i)}\right|^2 + \sum_{j \in \mathcal{A}_t} \left|c\,\rho(j) - \breve{R}^{(j)}\right|^2 + \lambda_c\big(n + |\mathcal{A}_t|\big)|c|^2. \tag{19}$$

Because $|\rho| = 1$, this equals $(1 + \lambda_c)(n + |\mathcal{A}_t|)|c|^2 - 2\,\mathrm{Re}(\bar{c}S) + \mathrm{const}$, and because the residuals are linear in (s, u).

$$S = \sum_{i \in \mathcal{M}_t} \overline{\rho(i)}\, R^{(i)} + \breve{T} = T - s\,P - u\,Q + \breve{T} \tag{20}$$

This is expressible solely through the sums in Eq. (12). Since the coefficient of $|c|^2$ is positive, the unique minimizer is:

$$\hat{c} = \frac{S}{(1 + \lambda_c)\big(n + |\mathcal{A}_t|\big)} = (1 - \alpha_c)\,\frac{S}{n + |\mathcal{A}_t|}, \qquad \alpha_c = \frac{\lambda_c}{1 + \lambda_c}. \tag{21}$$

S multiplies each residual by $\overline{\rho}$ and thereby unwinds (derotates) the phase of its window, so $\hat{c}$ is the mean of the derotated residuals shrunk by the factor $(1 - \alpha_c)$. Algorithm 1 computes S with the (s, u) clipped as in Sec. 3.5.

Gate Weight. The gate weight of Eq. (4) is the same kind of problem to solve with one real unknown. Over the matured windows, the squared error of the mixture $Y^{(j)} + w\Delta^{(j)}$ expands as follows, where $\Delta^{(j)} = C^{(j)} - Y^{(j)}$ and $D^{(j)} = G^{(j)} - Y^{(j)}$.

$$\sum_{j \in \mathcal{M}_t} \left|w\Delta^{(j)} - D^{(j)}\right|^2 = w^2 \underbrace{\sum_{j \in \mathcal{M}_t} \left|\Delta^{(j)}\right|^2}_{d} - 2w \underbrace{\sum_{j \in \mathcal{M}_t} \mathrm{Re}\big(\overline{\Delta^{(j)}}\, D^{(j)}\big)}_{p} + \sum_{j \in \mathcal{M}_t} \left|D^{(j)}\right|^2, \tag{22}$$

a real quadratic in w whose minimizer is p/d whenever $d > 0$. The real part is what enters because the two cross terms, $w\Delta^{(j)}\overline{D^{(j)}}$ and $w\overline{\Delta^{(j)}}D^{(j)}$, are conjugates of each other, so their imaginary parts cancel and their sum is $2w\,\mathrm{Re}(\overline{\Delta^{(j)}}D^{(j)})$. Equivalently, p is the inner product of the applied and required displacements read as vectors in $\mathbb{R}^2$, which is the only part of the complex product a single real w can act on. The gate therefore holds p and d and nothing else. Clipping p/d to $[0, 1]$ keeps the emitted coefficient a convex combination of $Y^{(j)}$ and $C^{(j)}$, and because the quadratic is convex in the single real w, the clipped value is the minimizer over that interval. The gate weight is therefore an exact optimum of its own objective, which the pair (s, u) is not: their joint solution is shrunk and then clipped coordinate-wise. Finally, $d = 0$ marks a bin whose correction never moved, which forces $p = 0$ as well, so flooring the denominator as Algorithm 1 does leaves $w = 0$ and the backbone forecast in place.

D VARIANCE OF THE DC-BIN PROVISIONAL TARGET

Sec. 3.6 fills the unobserved part of the provisional target with the backbone forecast only in the DC bin, because the zero-filled DC coefficient, once rescaled by H/m_j, has a variance that grows as the POGT shortens. This section reports the measurement behind that statement.

Setup. For every test window of iTransformer at $H = 336$ on the nine datasets (seed 0) and every observed fraction m/H, we form the two provisional DC coefficients, the sum of the m observed steps rescaled by H/m (zero-fill) and the sum of the m observed steps plus the backbone forecast of the remaining $H - m$ steps (backbone-fill), and compare each with the DC coefficient of the completed window. All three are divided by H so that the error e is in window-mean units. Table 6 reports the variance of that error over all test windows and channels, $\mathrm{Var}(e) = \mathbb{E}[e^2] - (\mathbb{E}[e])^2$, at seven observed fractions, and Figure 2 plots the ratio of the two variances over the range of m/H. The variance is centered, so it excludes the bias of the backbone forecast. In the DC bin, that bias is a level offset, which the rotational bias c_1 absorbs (Sec. 3.5), and the variance is the part of the error that the fill convention itself controls.

Table 6: Variance of the DC-bin provisional-target error (window-mean units), iTransformer, $H = 336$, seed 0. zero (zero-fill rescaled by H/m), back (backbone-fill). The columns use $m = 3, 17, 34, 84, 168, 252, 302$, the nearest integers to the stated fractions of $H = 336$. **Bold** marks the lower variance of the two.

	$m/H = 1\%$		$m/H = 5\%$		$m/H = 10\%$		$m/H = 25\%$		$m/H = 50\%$		$m/H = 75\%$		$m/H = 90\%$	
Dataset	zero	back	zero	back	zero	back	zero	back	zero	back	zero	back	zero	back
ETTh1	0.7941	**0.1324**	0.2348	**0.1284**	0.1536	**0.1221**	**0.0792**	0.0970	**0.0308**	0.0518	**0.0086**	0.0157	**0.0019**	0.0031
ETTh2	0.2351	**0.1516**	0.1578	**0.1482**	**0.1275**	0.1413	**0.0772**	0.1116	**0.0291**	0.0557	**0.0073**	0.0158	**0.0012**	0.0029
ETTm1	0.8034	**0.1382**	0.6239	**0.1346**	0.4007	**0.1261**	**0.0888**	0.1001	**0.0373**	0.0540	**0.0099**	0.0160	0.0051	**0.0041**
ETTm2	0.1722	**0.1274**	0.1332	**0.1244**	**0.1028**	0.1185	**0.0529**	0.0973	**0.0221**	0.0545	**0.0059**	0.0165	**0.0013**	0.0032
Weather	0.1975	**0.1310**	0.1598	**0.1289**	0.1315	**0.1240**	**0.0788**	0.1030	**0.0322**	0.0573	**0.0086**	0.0172	**0.0016**	0.0033
Exchange	**0.1654**	0.1857	**0.1511**	0.1835	**0.1350**	0.1788	**0.0932**	0.1537	**0.0411**	0.0888	**0.0101**	0.0268	**0.0016**	0.0047
Solar	0.7040	**0.0479**	0.5912	**0.0459**	0.4284	**0.0417**	0.1359	**0.0326**	0.0202	**0.0186**	0.0147	**0.0074**	0.0053	**0.0027**
Electricity	0.7235	**0.0432**	0.1599	**0.0420**	0.0937	**0.0398**	0.0401	**0.0318**	**0.0112**	0.0178	**0.0046**	0.0054	0.0012	**0.0011**
Traffic	1.1578	**0.0302**	0.2161	**0.0289**	0.1134	**0.0272**	0.0428	**0.0217**	**0.0091**	0.0125	0.0049	**0.0041**	0.0015	**0.0010**

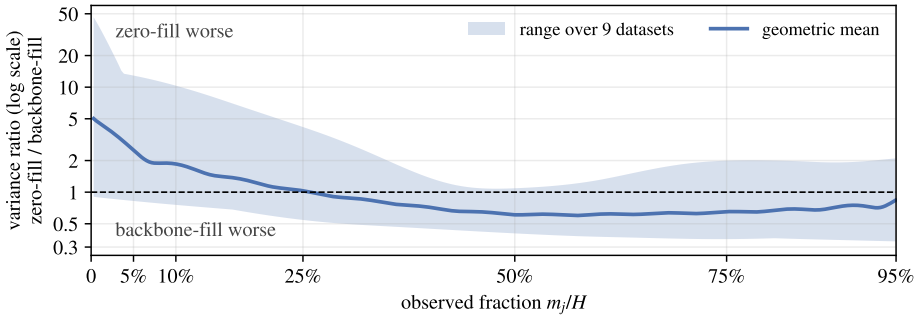


Figure 2: Ratio of the two variances of Table 6, $\mathrm{Var}(\text{zero-fill})/\mathrm{Var}(\text{backbone-fill})$, as a function of the observed fraction m/H: range and geometric mean over the nine datasets (log scale, above the dashed line, zero-fill has the larger variance). The range stops at $m/H = 95\%$, since both variances vanish as $m \to H$ and their ratio is undefined in the limit.

Result. The variance of zero-fill is unbounded as m shrinks. At $m/H = 1\%$ it ranges from .17 to 1.16 across the datasets, whereas backbone-fill stays between .03 and .19, the level set by the error of the backbone itself. Backbone-fill is lower on 8 of the 9 datasets at 1% (geometric mean of the ratio 4.6, up to 38 on Traffic), on 8 of 9 at 5% (2.5), and on 6 of 9 at 10% (1.9), with the largest gaps on the high-dimensional datasets, Traffic, Electricity, and Solar. The two conventions cross near $m/H = 25\%$ (geometric mean 1.03), and for longer POGTs zero-fill is lower, on 8 of 9 datasets at 50% (0.61), and it stays lower through 75% and 90%, where both variances have fallen below .03. By then both variances are small. The cost of keeping backbone-fill at 50% is at most .05 in window-mean units (Exchange) and .03 elsewhere, whereas the cost of zero-fill at 1% is up to 1.13 (Traffic). SWiT therefore keeps backbone-fill for every m. It bounds the variance where it is worst, on the freshest windows, at a small cost on the oldest. Exchange is the one dataset on which zero-fill is lower at every m/H, consistent with its random-walk behavior noted in Sec. 4.1. The observed step level better predicts the window level.

E ADDITIONAL ABLATIONS

This appendix holds the tables for the ablations of Sec. 4.3. The composition of the affine terms (Appendix E.1), the row-separation rule (Appendix E.2), the closed-form update with the adaptation batch size (Appendix E.4), and the POGT mechanism with the evidence gate are measured on 108 cells (DLinear, iTransformer, FreTS $\times$ 9 datasets $\times$ 4 horizons, seed 0), reported as per-dataset and per-horizon averages. We also report the basis ablation in a separate setting of 72 cells with POGT spliced off (Appendix E.3), with the cost side and the per-horizon breakdown of the component ablation collected in Appendix E.5.

E.1 AFFINE TERMS

Table 7 reports the per-dataset average MSE, and Table 8 reports the per-horizon average MSE. These tables also show whether the gains come from re-fitting the normalization. The backbones already normalize every input window internally (Appendix F), so a per-window level or scale drift is removed before SWiT acts, and no single term, the level-correcting bias included, accounts for the reduction on its own. All percentages below are MSE reductions over None. The full model with all three terms, SWiT, reaches 9.46%.

(i) **Contribution of sY.** The scale term sY alone reduces MSE by 5.6%, which is 59.3% of the full reduction of SWiT. Its reduction grows with the horizon, from 3.0% at $H{=}96$ to a maximum of 8.8% at $H{=}720$.

(ii) **Contribution of uX.** Adding the input term uX to sY raises the reduction from 5.6% to 6.9%, and adding it to $sY{+}c\rho$ raises it from 7.7% to 9.5%. Per horizon, the contribution is largest at $H{=}336$, where 4.8% becomes 7.4%. At $H{=}720$ adding uX to sY barely changes the reduction (8.8% to 8.9%), whereas adding it to $sY{+}c\rho$ raises 10.4% to 11.6%.

(iii) **Contribution of $c\rho$.** Adding the rotational bias $c\rho$ to sY raises the reduction from 5.6% to 7.7%, and adding it to $sY{+}uX$ raises it from 6.9% to 9.5%. It adds $1.6 \sim 2.7$ pp at every horizon, and at $H{=}96$ it lifts 3.0% to 5.2%, a larger contribution than uX under the same condition ($3.0\% \rightarrow 4.3\%$).

The 9.5% of the full model exceeds every partial configuration, and each of uX and $c\rho$ contributes more when the other term is present. Per dataset, the full configuration is lowest on eight of the nine datasets. On Exchange, sY alone is lowest and adding uX or $c\rho$ increases the error (Table 7).

Table 7: Affine-term ablation, per-dataset MSE (mean over 3 backbones $\times$ 4 horizons, seed 0). **bold** = best, underline = second.

Dataset	None	sY	$sY{+}uX$	$sY{+}c\rho$	SWiT ($sY{+}uX{+}c\rho$)
ETTh1	.5582	.5326	.5293	<u>.5286</u>	**.5243**
ETTh2	.3231	.3174	<u>.3047</u>	.3155	**.2896**
ETTm1	.4781	.4489	.4393	<u>.4361</u>	**.4250**
ETTm2	.2315	.2272	<u>.2192</u>	.2201	**.2102**
Weather	.2657	.2430	<u>.2405</u>	.2414	**.2405**
Exchange	.3643	<u>.3039</u>	.3363	**.2977**	.3232
Solar	.3184	.2987	.2896	<u>.2859</u>	**.2786**
Electricity	.2157	.2088	<u>.1994</u>	.2027	**.1961**
Traffic	.5660	.5543	<u>.5319</u>	.5382	**.5197**
Avg.	.3690	.3483	.3434	<u>.3407</u>	**.3341**

E.2 ROW SEPARATION

Table 9 reports the per-dataset MSE, and Table 10 reports the per-horizon average MSE. Adding the POGT statistics to the u equation only yields an MSE reduction of 9.5% over None; adding them to both equations yields 8.8%. The difference has the same sign at all four horizons and ranges from 0.44 pp at $H{=}96$ to 0.89 pp at $H{=}720$. Per dataset, the u-only variant is lower on eight of the nine datasets. On Exchange, adding POGT to both equations is lower.

Table 8: Affine-term ablation, per-horizon MSE (mean over 3 backbones × 9 datasets, seed 0).

H	None	sY	$sY{+}uX$	$sY{+}c\rho$	SWiT ($sY{+}uX{+}c\rho$)
96	.2809	.2725	.2688	<u>.2663</u>	**.2632**
192	.3236	.3104	.3049	<u>.3043</u>	**.2967**
336	.3765	.3586	.3487	<u>.3486</u>	**.3391**
720	.4951	.4517	.4511	<u>.4436</u>	**.4376**
Avg.	.3690	.3483	.3434	<u>.3407</u>	**.3341**

The rule leaves one asymmetry, Stage 1 reads the provisional target as $\breve{G}^{(j)} \approx u\,\breve{X}^{(j)}$, whereas the residual of Stage 2 reads it as $\frac{m_j}{H}(sY^{(j)} + uX^{(j)})$, so the share of $\breve{G}^{(j)}$ that $sY^{(j)}$ already explains is absorbed by u in Stage 1. The third column tests whether this matters. It keeps $\hat{s}$ at the value of the previous batch, which the matured windows determine, and fits u on $\breve{G}^{(j)} - \frac{m_j}{H} sY^{(j)}$, so that both stages assume the same model. Taking $\hat{s}$ from the current batch instead would make Stage 1 a two-pass solve. The variant matches the u-only rule on average (.3340 against .3341), but the tie comes from Exchange alone. Excluding it, the u-only rule is lower by .0027, and over the 108 cells the variant wins only 31 ($p = 2.3 \times 10^{-6}$). Its gain is confined to $H{=}720$, where the relative weight of the POGT terms, $\mathcal{O}(H/n)$, is largest, and it loses at the three shorter horizons. The asymmetry is therefore real but small, and correcting it overcorrects where the POGT weight is low, so we keep the rule as stated.

Table 9: Row-separation ablation, per-dataset MSE (mean over 3 backbones × 4 horizons, seed 0). **Bold** = lowest and <u>underline</u> = second lowest of the three.

Dataset	None	SWiT (POGT in u row)	POGT in u row, target $\breve{G}^{(j)} - \frac{m_j}{H} sY^{(j)}$	POGT in both rows
ETTh1	.5582	<u>.5243</u>	**.5231**	.5263
ETTh2	.3231	**.2896**	<u>.2941</u>	.3020
ETTm1	.4781	**.4250**	.4273	<u>.4269</u>
ETTm2	.2315	**.2102**	.2178	<u>.2139</u>
Weather	.2657	**.2405**	<u>.2421</u>	.2443
Exchange	.3643	.3232	**.3004**	<u>.3172</u>
Solar	.3184	**.2786**	.2796	<u>.2794</u>
Electricity	.2157	**.1961**	<u>.1985</u>	.1986
Traffic	.5660	**.5197**	.5230	<u>.5212</u>
Avg.	.3690	<u>.3341</u>	**.3340**	.3366

Table 10: Row-separation ablation, per-horizon MSE (mean over 3 backbones × 9 datasets, seed 0). The last two rows compare each variant with the adopted u-row rule over the 108 paired cells. The win count is the number of cells in which the variant attains the lower MSE, and p is a two-sided Wilcoxon signed-rank test.

H	None	SWiT (POGT in u row)	POGT in u row, target $\breve{G}^{(j)} - \frac{m_j}{H} sY^{(j)}$	POGT in both rows
96	.2809	**.2632**	<u>.2642</u>	.2645
192	.3236	**.2967**	.2991	<u>.2986</u>
336	.3765	**.3391**	<u>.3403</u>	.3415
720	.4951	<u>.4376</u>	**.4324**	.4420
Avg.	.3690	<u>.3341</u>	**.3340**	.3366
Wins vs. u row	–	–	31/108	17/108
Wilcoxon p	–	–	$2.3{\times}10^{-6}$	$1.2{\times}10^{-10}$

Table 11: Time-domain versus frequency-domain correction operators. FD diag is SWiT without POGT splicing. "vs FD diag" counts the cells in which an arm attains a lower MSE than FD diag. p is the two-sided Wilcoxon signed-rank test against FD diag, bold when $p < 0.05$.

		iTransformer					DLinear					FreTS				
Dataset	H	None	FD diag	FD dense	TD diag	TD dense	None	FD diag	FD dense	TD diag	TD dense	None	FD diag	FD dense	TD diag	TD dense
ETTh1	96	.4507	**.4229**	.4249	.4379	.4279	.4595	.4329	**.4317**	.4476	.4328	.4462	**.4266**	.4293	.4411	.4305
	192	.5078	.4808	.4897	.4960	**.4793**	.5125	.4828	.4876	.5052	**.4776**	.5022	.4811	.4877	.4998	**.4773**
	336	.5658	.5555	.5513	.5611	**.5495**	.5659	**.5452**	.5455	.5560	.5570	.5544	.5453	**.5422**	.5530	.5477
	720	.7038	.6876	.6883	**.6805**	.7038	.7117	**.6831**	.6902	.6958	.7117	.7182	**.6988**	.7001	.7041	.7182
ETTh2	96	.2577	.2486	.2463	.2501	**.2456**	.2323	.2292	**.2270**	.2312	.2272	.2384	**.2325**	.2337	.2379	.2328
	192	.3161	**.2841**	.2972	.2915	.3038	.2862	**.2731**	.2773	.2780	.2801	.2866	**.2733**	.2782	.2789	.2810
	336	.3545	**.3224**	.3355	.3250	.3403	.3252	.3092	.3109	.3137	**.3062**	.3317	**.3107**	.3144	.3164	.3151
	720	.4276	.4081	.4245	**.4038**	.4276	.4087	**.3855**	.4102	.3865	.4087	.4119	.3870	.4097	**.3857**	.4119
ETTm1	96	.3823	.3311	.3204	.3317	**.3192**	.3715	**.3264**	.3333	.3313	.3325	.3675	**.3289**	.3326	.3308	.3330
	192	.4423	.3839	**.3808**	.3971	.3834	.4438	**.3864**	.3892	.4039	.3908	.4325	**.3868**	.3902	.3953	.3945
	336	.5093	.4625	**.4396**	.4844	.4461	.5183	.4680	.4448	.4956	**.4441**	.5005	.4630	**.4480**	.4827	.4488
	720	.6065	.5345	**.5095**	.5603	.5117	.5929	.5305	.5034	.5684	**.5022**	.5704	.5231	.5047	.5487	**.4954**
ETTm2	96	.1647	**.1539**	.1557	.1599	.1559	.1598	**.1503**	.1553	.1546	.1579	.1581	**.1486**	.1541	.1542	.1543
	192	.2209	.1911	.1981	.1932	**.1873**	.1930	**.1790**	.1881	.1810	.1804	.1923	**.1788**	.1885	.1800	.1807
	336	.2727	.2474	.2378	.2578	**.2299**	.2324	.2210	.2216	.2325	**.2190**	.2320	.2256	.2226	.2311	**.2182**
	720	.3451	**.2981**	.3003	.3211	.3053	.3062	.2845	**.2833**	.2958	.2868	.3012	**.2770**	.2797	.2920	.2843
Exchange	96	.0882	**.0869**	.0915	.0880	.0876	.0913	.0948	.0910	.0942	**.0908**	.0828	.0832	.0846	**.0828**	.0832
	192	.1811	.1798	.2024	**.1753**	.1999	.1827	.1804	.2101	**.1803**	.2133	.1734	.1716	.1985	**.1693**	.1984
	336	.3428	.3389	.3434	**.3302**	.3428	.3277	.3265	.3301	**.3213**	.3277	.3240	.3222	.3272	**.3164**	.3240
	720	.8540	.8540	.8540	.8540	.8540	.8873	.8873	.8873	.8873	.8873	.8368	.8368	.8368	.8368	.8368
Weather	96	.1755	.1622	.1534	.1612	**.1532**	.1954	.1805	.1580	.1759	**.1566**	.1856	.1745	.1575	.1713	**.1566**
	192	.2232	.2078	**.1974**	.2088	.1980	.2403	.2186	**.1985**	.2203	.2001	.2310	.2133	**.1986**	.2149	.2015
	336	.2800	.2591	.2680	**.2588**	.2770	.2918	**.2679**	.2703	.2694	.2828	.2843	**.2633**	.2715	.2644	.2807
	720	.3571	.3279	.3364	**.3275**	.3421	.3643	.3330	.3413	**.3303**	.3453	.3599	.3300	.3391	**.3277**	.3409
Avg.		.3762	**.3512**	.3519	.3565	.3529	.3709	**.3490**	.3494	.3565	.3508	.3634	**.3451**	.3471	.3506	.3477
vs FD diag		–	–	9/24	7/24	11/24	–	–	8/24	5/24	10/24	–	–	6/24	6/24	7/24
p (Wilcoxon)		–	–	6.5e-01	**1.5e-02**	5.0e-01	–	–	2.7e-01	**1.9e-03**	4.3e-01	–	–	1.5e-01	**4.7e-03**	1.8e-01

Table 12: Per-horizon and per-dataset means over the 72 cells of Table 11. The two comparison rows are computed over all 72 cells.

Horizon	None	SWiT (FD diag)	FD dense	TD diag	TD dense
$H = 96$	.2504	.2341	.2322	.2379	**.2321**
$H = 192$	.3093	**.2863**	.2921	.2927	.2904
$H = 336$	.3785	.3585	**.3569**	.3650	.3587
$H = 720$	.5424	**.5148**	.5166	.5226	.5208
Avg.	.3702	**.3484**	.3495	.3545	.3505
vs FD diag	–	–	23/72	18/72	28/72
p (Wilcoxon)	–	–	1.2e-01	**9.6e-07**	1.3e-01

Dataset	None	SWiT (FD diag)	FD dense	TD diag	TD dense
ETTh1	.5582	**.5369**	.5391	.5482	.5428
ETTh2	.3231	**.3053**	.3137	.3082	.3150
ETTm1	.4781	.4271	**.4164**	.4442	.4168
ETTm2	.2315	**.2129**	.2154	.2211	.2133
Exchange	.3643	.3635	.3714	**.3613**	.3705
Weather	.2657	.2448	**.2408**	.2442	.2446

E.3 Time-Domain Counterparts

The FFT is a change of basis, so the question arises whether the frequency domain contributes anything beyond the affine structure itself. We therefore compare four correctors that share the two-stage fit (rotational bias in the second stage) and the causal protocol of SWiT and differ only in the basis and in the degrees of freedom (TD, time domain, FD, frequency domain).

- **TD diagonal**: $\tilde{y}_\tau = s_\tau\, \hat{y}_\tau + u_\tau\, x_\tau + b_{(j+\tau) \bmod H}$, one scale and one input factor per lead time τ and a bias indexed by absolute time ($3H$ parameters).
- **TD dense**: $\tilde{\mathbf{y}} = \mathbf{W}_y \hat{\mathbf{y}} + \mathbf{W}_x \mathbf{x} + \mathbf{b}$ with $\mathbf{W}_y, \mathbf{W}_x \in \mathbb{R}^{H \times H}$, every lead time influencing every other ($2H^2 + H$ parameters).

Table 13: Avg. MSE and wall-clock execution time across different batch sizes. Numbers in parentheses denote the batch size. TFS, TAFAS (Kim et al., 2025); PTS, PETSA (Medeiros et al., 2025); COS, COSA (Im & Kwon, 2026); SWGD, SWiT-GD.

Metric	None	TFS	COS	PTS	SWGD	SWiT (8)	SWiT (16)	SWiT (24)	SWiT (48)	SWiT (96)
MSE	.3690	.3607	.3522	.3591	.3507	.3344	.3343	<u>.3342</u>	**.3341**	.3345
Time (s)	–	12.70	37.74	15.04	8.24	12.90	9.76	8.69	<u>7.74</u>	**7.73**

- **FD diagonal**: SWiT's correction $\tilde{Y}_k = s_k Y_k + u_k X_k$ plus the rotational bias ($2K + H$ parameters, $K = \lfloor H/2 \rfloor + 1$).
- **FD dense**: $\tilde{Y}_k = \sum_{k'} W^y_{kk'} Y_{k'} + \sum_{k'} W^x_{kk'} X_{k'}$ with complex $\mathbf{W}^y, \mathbf{W}^x \in \mathbb{C}^{K \times K}$, mixing frequency components ($4K^2 + H$ parameters).

The diagonal arms use clipping and the energy-proportional ridge of SWiT. The dense arms, for which clipping has no analog, use a ridge equal to 1% of the diagonal energy. We apply the evidence gate in the frequency basis for all four arms because a time-domain gate degrades accuracy, and we disable POGT splicing throughout so only the corrector changes. The comparison covers three backbones (DLinear, iTransformer, FreTS), six datasets (ETTh1, ETTh2, ETTm1, ETTm2, Weather, Exchange), and four horizons, 72 cells in all, with seed 0 (Tables 11 and 12).

Three patterns emerge. 1) With the same $\mathcal{O}(H)$ degrees of freedom, the frequency-domain corrector is more accurate than its time-domain counterpart on 51 of the 72 cells (18 cells favor TD diagonal, 3 are ties), and the difference is significant under a two-sided Wilcoxon signed-rank test ($p = 9.6 \times 10^{-7}$). The reversals concentrate on Exchange and Weather. 2) Against the dense time-domain corrector, which has $\mathcal{O}(H^2)$ degrees of freedom, FD diagonal is lower on 41 of 72 cells and lower on average (.3484 vs. .3505), and the test finds no difference ($p = 0.13$). The two are statistically indistinguishable, while FD diagonal uses fewer parameters. 3) Making the frequency-domain corrector dense does not help either. FD dense is lower than FD diagonal on 23 of 72 cells ($p = 0.12$) and averages .3495, so once the correction may mix all components, the basis no longer matters, as the change-of-basis argument predicts. The spectral basis therefore pays off exactly where SWiT exploits it. It makes a diagonal, per-component correction competitive with a dense one, and its diagonal structure admits the per-bin closed form and the per-bin gate.

E.4 Closed Form versus Gradient Update and Batch Size

Table 13 compares the average MSE and the wall-clock time of the SGD-based TTA methods, SWiT-GD, and SWiT at five adaptation batch sizes. SWiT-GD reaches .3507 against .3522 for COSA in 8.24 s against 37.74 s, so one gradient step on the accumulated quadratic already matches the best existing method at a fraction of its cost, and the closed form lowers the MSE to .3341 in 7.74 s. Across the five batch sizes the MSE spans .0004, from .3341 at $B = 48$ to .3345 at $B = 96$, while the time falls from 12.90 s at $B = 8$ to 7.73 s at $B = 96$. $B = 48$ is therefore where the accuracy is highest and the time is already near its floor.

E.5 Component Ablations

Table 14 collects the two component ablations of Sec. 4.3. Panel (a) adds the time, memory, and parameter cost of the four correctors in Appendix E.3, with accuracy in Tables 11 and 12. Panel (b) is the per-horizon breakdown of Table 3: in w/o POGT every path that uses POGT during adaptation is disabled, and in w/o Gate $w_k = 1$ in Eq. (4), so the correction is emitted as is.

The two losses move in opposite directions as the horizon increases: without POGT the loss grows from 0.54% at H=96 to 6.43% at H=720, while without the gate it shrinks from 2.70% to 0.45%. This first trend is what Sec. 3.6 anticipates: a longer horizon delays the first matured GT, leaves $H - 1$ windows incomplete at every time step, and raises the relative weight $\mathcal{O}(H/n)$ of the POGT terms. The second trend follows from the gate's purpose. Since w_k is the mixing ratio that would have reduced the error most on past windows, it converges to 1 wherever the corrections are mostly right, as at H=720, so fixing $w_k = 1$ changes little there. At H=96 the backbone is already accurate,

Table 14: Component ablations. (a) Frequency versus time basis for the same closed form, six datasets, POGT splicing off. (b) Avg. MSE with and without POGT splicing and the evidence gate.

(a) Frequency versus time domain

	None	FD diag	FD dense	TD diag	TD dense
Avg. MSE	.3702	**.3484**	.3495	.3545	.3505
Time (s)	–	4.83	7.05	**4.76**	6.58
Peak GPU (MB)	–	90	193	**88**	316
Params (per channel)	–	$2K$ $+H$	$4K^2$ $+H$	$2H$ $+H$	$2H^2$ $+H$

(b) POGT splicing and the evidence gate

H	None	SWiT	w/o POGT	w/o Gate	w/o Both
96	.2809	**.2632**	.2647	.2703	.2741
192	.3236	**.2967**	.3005	.3044	.3083
336	.3765	**.3391**	.3450	.3473	.3491
720	.4950	**.4376**	.4657	.4395	.4740
Avg.	.3690	**.3341**	.3440	.3404	.3514

and the residual error it leaves has little spectral structure, so the correction in many bins acts as noise and the gate holds them near the backbone output.

F Evaluation Settings

Table 2 counts # Params per channel, averaged over the nine datasets, six backbones, and four horizons; for PETSA, whose module is shared across channels, the shared count is divided by the number of channels of each dataset. Its FLOPs are counted for one adaptation or inference pass, per channel and averaged over the four horizons, with the backbone forward excluded. We ran all methods on the same backbone checkpoints, data splits, and test streams. TAFAS, PETSA, and COSA were run with their official implementations unchanged. Seeds apply to backbone training. For every setting, the backbone is trained with five seeds, TTA is run once from each checkpoint, and we report the mean of the five values. The compared methods use the same five checkpoints. Only OLS is run once, since its training is a deterministic closed-form refit. The SWiT adaptation is closed-form and therefore deterministic. The adaptation random seed of the compared methods is fixed to 0. Every batch runs in the order evaluation then adaptation (Sec. 3.1).

F.1 Datasets and Forecasting Setup

Table 15: Datasets, sampling frequency, and splits.

Dataset	Channels	Frequency	Timesteps	Train	Validation	Test	Split
ETTh1	7	1 hour	17,420	10,452	3,484	3,484	6:2:2
ETTh2	7	1 hour	17,420	10,452	3,484	3,484	6:2:2
ETTm1	7	15 min	69,680	41,808	13,936	13,936	6:2:2
ETTm2	7	15 min	69,680	41,808	13,936	13,936	6:2:2
Weather	21	10 min	52,696	36,887	5,270	10,539	7:1:2
Exchange	8	1 day	7,587	5,310	760	1,517	7:1:2
Solar	137	10 min	52,560	36,792	5,256	10,512	7:1:2
Electricity	321	1 hour	26,304	18,412	2,632	5,260	7:1:2
Traffic	862	1 hour	17,544	12,280	1,756	3,508	7:1:2

We evaluate SWiT on nine standard time-series forecasting benchmarks, namely ETTh1/2, ETTm1/2, Exchange Rate, Weather, Electricity, Solar, and Traffic; Table 15 lists their sampling frequency and splits. The lookback window is fixed at $W = 96$, and we use four prediction horizons $H \in \{96, 192, 336, 720\}$. All input time series are processed as a variable-wise univariate forecasting task. Every series is z-score normalized per channel with statistics fitted on the training split and kept fixed thereafter. RevIN is not used, but each backbone applies instance normalization to its own input window internally. All TTA methods operate in the fixed z-scored space outside that instance normalization. TAFAS and PETSA at the backbone input and output, COSA and SWiT at the output only, and we report the MSE in this normalized space. The train/validation/test split ratio is set to 6:2:2 for the ETT datasets and 7:1:2 for the remaining datasets.

F.2 Backbone Models

We use six representative backbone models spanning different architecture families, namely Transformer-based (iTransformer (Liu et al., 2024), PatchTST (Nie et al., 2022)), Linear-based (DLinear (Zeng et al., 2023), OLS (Toner & Darlow, 2024)), and MLP-based (FreTS (Yi et al., 2023), MICN (Wang et al., 2023)).

F.3 Hyperparameters

SWiT uses an adaptation batch of 48, clipping ranges $s_k \in [0.5, 1.5]$ and $u_k \in [-0.3, 0.3]$, and shrinkage $\alpha_u = \alpha_c = 1/2$ in every setting, with no tuning per dataset, backbone, or horizon. We selected the clipping ranges and shrinkage rates once on the validation splits of ETTh1 and ETTm2 (Appendix H.1), and report their sensitivity in Appendix H. SWiT-GD uses a batch of 48 and one step per batch. Periodicity-aware adaptation scheduling (PAAS) is the rule proposed by TAFAS that sets the adaptation batch to the dominant period length found in the input FFT. TAFAS and PETSA use the PAAS batch and one step per batch. COSA uses a batch of 48, three steps per batch, and a context length of 10, identical to its original configuration. All methods use Adam; the learning rate and scheduler follow each method's official implementation. DynaTTA (Grover & Etemad, 2025) is cited in Sec. 2 but not compared. Its public repository contains only the entry script, while the model, data, trainer, and configuration modules it imports are absent. The paper does not specify the sensitivity, smoothing, and warm-up coefficients of its adaptation-rate schedule, the MSE-buffer size, or the gating network, and the code's default values differ from those in the paper's appendix. We could therefore not fix a setting whose results we could stand behind.

F.4 Hardware

We ran all experiments on a server with an Intel Xeon 6515P CPU and an NVIDIA RTX PRO 6000 Blackwell 96GB GPU. Each setting was run sequentially on a single GPU.

F.5 SWiT-GD

SWiT-GD is a variant of SWiT that isolates the contribution of the closed-form update. It shares the spectral correction module, the running sums, POGT splicing, and the evidence gate with SWiT and differs only in how it obtains (s_k, u_k). SWiT-GD starts from the same stage-1 objective $\breve{\mathcal{L}}_k$ of Eq. (9) as SWiT, takes the (s_k, u_k) of the previous batch as the initial point, computes the gradient once, and moves a single Adam step (learning rate 0.01). Because $\breve{\mathcal{L}}_k(s, u)$ is quadratic in (s, u), its gradient is written in terms of the same sums that SWiT uses to form the normal equations.

$$\begin{aligned} \frac{\partial \breve{\mathcal{L}}_k}{\partial s} &= 2\big(s\,\Sigma^{YY} + u\,\Sigma^{XY} - \Sigma^{YG}\big), \\ \frac{\partial \breve{\mathcal{L}}_k}{\partial u} &= 2\Big(s\,\Sigma^{XY} + u\,\big(\Sigma^{XX} + \breve{\Sigma}^{XX}\big) - \big(\Sigma^{XG} + \breve{\Sigma}^{XG}\big)\Big), \end{aligned} \tag{23}$$

$\Sigma^{AB} = \sum_{j \in \mathcal{M}_t} \mathrm{Re}(\overline{A^{(j)}} B^{(j)})$ are the running sums over matured windows and $\breve{\Sigma}^{XX}$, $\breve{\Sigma}^{XG} = \sum_{j \in \mathcal{A}_t} \mathrm{Re}(\overline{\breve{X}^{(j)}} \breve{G}^{(j)})$ are the POGT sums over in-progress windows. SWiT uses the solution of the 2×2 system obtained by setting this gradient to 0. SWiT-GD takes one step along the same gradient. The Adam state persists throughout the stream, and the (s_k, u_k) after the step receive the same shrinkage and clipping as in SWiT. The rotational bias c_k is solved in closed form on the resulting residual exactly as in SWiT, and the evidence gate passes the coefficients through under the same rule. The two methods therefore head for the same minimizer of the same objective. SWiT reaches it at every batch, whereas SWiT-GD moves toward it in one gradient step, so the accuracy difference reflects the gain of the closed-form update.

G Full Experimental Results

This appendix holds the complete result tables: the per-dataset and per-horizon means in MSE, every one of the 216 settings with its spread over five backbone seeds, and the same comparison in MAE.

G.1 Per-Dataset and Per-Horizon Means

Table 16 reports the MSE for each method, backbone, and dataset, averaged over the four horizons.

Table 16: Per-dataset MSE averaged over four horizons. TFS, TAFAS (Kim et al., 2025); PTS, PETSA (Medeiros et al., 2025); COS, COSA (Im & Kwon, 2026); SWGD, SWiT-GD.

	Transformer-based: iTransformer						Linear-based: DLinear						MLP-based: FreTS					
Dataset	None	TFS	PTS	COS	SWGD	SWiT	None	TFS	PTS	COS	SWGD	SWiT	None	TFS	PTS	COS	SWGD	SWiT
ETTh1	.574	.555	.549	.544	.563	**.531**	.559	.550	.544	.544	.550	**.522**	.555	.546	.540	.539	.552	**.525**
ETTh2	.344	.334	.332	.333	.332	**.307**	.312	.306	.304	.306	.308	**.282**	.317	.310	.308	.309	.310	**.289**
ETTm1	.497	.462	.461	.456	.444	**.428**	.481	.451	.450	.453	.437	**.426**	.465	.453	.449	.454	.433	**.421**
ETTm2	.246	.241	.237	.245	.230	**.219**	.223	.220	.219	.223	.214	**.206**	.221	.219	.218	.221	.211	**.205**
Exchange	.386	.379	.381	**.310**	.379	.335	.359	.355	.355	**.286**	.354	.321	.356	.351	.351	**.281**	.347	.314
Weather	.257	.248	.248	.245	.240	**.233**	.273	.261	.260	.259	.253	**.246**	.265	.254	.255	.253	.247	**.241**
Solar	.268	.264	.261	.262	.251	**.247**	.371	.359	.356	.362	.322	**.309**	.283	.273	.271	.281	.264	**.258**
Electricity	.188	.186	.187	.185	.183	**.177**	.229	.228	.227	.225	.212	**.204**	.211	.209	.209	.207	.204	**.196**
Traffic	.453	.452	.454	.452	.447	**.439**	.659	.659	.659	.659	.619	**.587**	.550	.550	.550	.550	.536	**.516**
Average	.357	.347	.346	.337	.341	**.324**	.385	.377	.375	.369	.363	**.345**	.358	.352	.350	.344	.345	**.329**

	Transformer-based: PatchTST						Linear-based: OLS						MLP-based: MICN					
Dataset	None	TFS	PTS	COS	SWGD	SWiT	None	TFS	PTS	COS	SWGD	SWiT	None	TFS	PTS	COS	SWGD	SWiT
ETTh1	.571	.551	.558	.544	.567	**.532**	.583	.576	.568	.568	.568	**.536**	.767	.720	.711	.687	.678	**.605**
ETTh2	.318	.310	.312	.308	.312	**.290**	.325	.319	.317	.319	.318	**.289**	.355	.347	.345	.344	.343	**.319**
ETTm1	.484	.468	.483	.454	.437	**.425**	.485	.455	.454	.456	.439	**.428**	.545	.511	.509	.497	.465	**.446**
ETTm2	.234	.234	.235	.234	.227	**.212**	.224	.222	.221	.224	.215	**.206**	.242	.238	.236	.243	.227	**.217**
Exchange	.405	.401	.405	**.332**	.397	.342	.360	.356	.357	**.287**	.355	.321	.537	.473	.475	**.416**	.513	.441
Weather	.256	.248	.253	.246	.241	**.235**	.274	.264	.262	.260	.254	**.247**	.258	.254	.265	.248	.245	**.240**
Solar	.255	.252	.273	.252	.243	**.240**	.379	.369	.362	.368	.326	**.313**	.266	.266	.271	.263	.250	**.245**
Electricity	.201	.198	.209	.198	.195	**.191**	.215	.211	.213	.211	.208	**.200**	.196	.190	.192	.192	.179	**.173**
Traffic	.482	.480	.490	.481	.472	**.464**	.623	.619	.616	.622	.601	**.568**	.520	.504	.511	.505	.482	**.469**
Average	.356	.349	.358	.339	.343	**.326**	.385	.377	.375	.368	.365	**.345**	.410	.389	.390	.377	.376	**.350**

G.2 Per-Setting Results over Five Seeds

Tables 18 and 19 report every setting separately, with each cell showing the mean±std over five backbone seeds.

G.3 Results in MAE

Table 20 repeats the main MAE comparison on the same 216 settings. Every cell is the mean over the same five backbone seeds as Table 16, and bold and underline mark the best and second-best TTA method. SWiT has the lowest average MAE on all six backbones and at all four horizons, .3610 overall against .3739 for None (3.45% lower) and .3668 for COSA, the strongest baseline. Tables 20a and 20b give the per-dataset and horizon-wise means. The margins are smaller than in MSE, as expected for a method that minimizes a squared objective, and on 8 of the 54 backbone–dataset cells, concentrated on Solar and Weather, SWiT is above None in MAE while below it in MSE.

H Hyperparameter Sensitivity

SWiT has four hyperparameters, the clipping ranges δ_s, δ_u and the shrinkage rates α_u, α_c. δ_s and δ_u clip the scale and input coefficients solved in closed form at every batch to $s_k \in [1-\delta_s, 1+\delta_s]$ and $u_k \in [-\delta_u, \delta_u]$ before they are used in the output, and α_u and α_c shrink the least-squares solutions of the input coefficient and the rotational bias by the factor $(1-\alpha)$. This section examines sensitivity to each. We vary one value at a time and report results on 108 cells (DLinear, iTransformer, FreTS $\times$ 9 datasets $\times$ 4 horizons, seed 0). The tables give the per-horizon average MSE (bold = lowest within the block, underline = second), and the figures give the per-dataset MSE reduction (mean over 12 cells).

Table 17: Horizon-wise average MSE. TFS, TAFAS (Kim et al., 2025); PTS, PETSA (Medeiros et al., 2025); COS, COSA (Im & Kwon, 2026); SWGD, SWiT-GD.

H	None	TFS	PTS	COS	SWGD	SWiT
96	.2841	.2773	.2799	.2773	.2672	**.2625**
192	.3268	.3186	.3203	.3188	.3089	**.2972**
336	.3811	.3697	.3704	.3687	.3579	**.3416**
720	.5091	.4947	.4914	.4577	.4881	**.4453**
Avg.	.3753	.3651	.3655	.3556	.3555	**.3366**

H.1 Selection

We fixed the adopted values once, before the main experiments, on the validation splits of ETTh1 and ETTm2 with FreTS at the four horizons (8 cells). Every combination of the grid below, $\alpha_u, \alpha_c \in \{0, 0.1, 0.3, 0.5, 0.7, 0.9, 1\}$, $\delta_u \in \{0.1, 0.2, 0.3, 0.5, \infty\}$, $\delta_s \in \{1/4, 1/2, 1, \infty\}$, 980 combinations in all, was ranked by validation MSE within each cell. The combination with the lowest mean rank over the 8 cells was adopted. The mean rank rather than the mean MSE was used so that cells with large errors do not dominate the choice. Table 21 lists the five best combinations. The adopted one, $\alpha_u = \alpha_c = 1/2$, $\delta_u = 0.3$, $\delta_s = 1/2$, ranks first, and all five share $\delta_s = 1/2$. The adaptation batch $B = 48$ was not part of the grid. We set it to the COSA batch size so the two methods refresh their statistics equally often, and Sec. 4.3 reports its effect.

H.2 Clipping

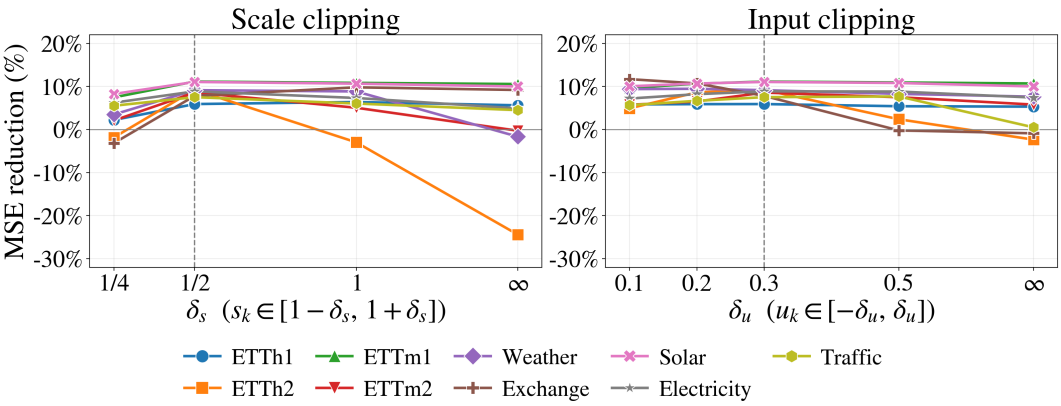


Figure 3: Clipping sensitivity per dataset (mean over 12 cells, dashed = adopted).

Because SWiT solves for its coefficients in closed form, a window that contains outliers or noise enters the solution directly, and the unclipped least-squares solution flips sign or becomes excessively large in some bins. Clipping is a safeguard that trims only these coefficients, so as long as the range is wide enough to trim them while leaving every other coefficient unchanged, their exact value barely affects the result (Table 22). For $\delta_u \in [0.1, 0.5]$ the overall MSE reduction is $7.3 \sim 9.5\%$, and removing the bound drops it to 5.1%. For δ_s the reduction is 9.5% at $1/2$, 7.9% at 1, and 3.8% at ∞. Narrowing it to $1/4$ trims needed corrections and gives 3.8%. The loss from removing a safeguard concentrates on particular datasets, ETTh2 at -25% for $\delta_s = \infty$ and Traffic at -29% for $\delta_u = \infty$, while the remaining datasets barely move (Figure 3).

H.3 Shrinkage

With $\alpha_u = 0$ the input coefficient is the plain least-squares solution and with $\alpha_u = 1$ the input term is switched off (Table 23, Figure 4). At $\alpha_u = 0$ the overall MSE reduction is 9.2%, 0.3 pp from the adopted value, and at $\alpha_u = 1$ it is 7.7%, with the largest losses on ETTh2 and ETTm2. In between, on $[0, 0.7]$, the reduction stays at 9.2–9.5% and the curve is flat on every dataset. With $\alpha_c = 0$, the

Table 18: Full results (MSE, mean±std over 5 seeds) TFS, TAFAS (Kim et al., 2025); PTS, PETSA (Medeiros et al., 2025); COS, COSA (Im & Kwon, 2026); SWGD, SWiT-GD.

(a) iTransformer / PatchTST

	H	iTransformer						PatchTST					
		None	TFS	PTS	COS	SWGD	SWiT	None	TFS	PTS	COS	SWGD	SWiT
ETTh1	96	.4656±.0131	.4454±.0063	.4456±.0056	.4480±.0068	.4397±.0074	**.4248**±.0038	.4470±.0103	.4401±.0095	.4528±.0093	.4419±.0097	.4423±.0110	**.4276**±.0076
	192	.5130±.0086	.4965±.0051	.5001±.0051	.4985±.0048	.4986±.0052	**.4799**±.0040	.5145±.0156	.5014±.0129	.5105±.0137	.5030±.0131	.5067±.0151	**.4844**±.0108
	336	.5878±.0171	.5772±.0135	.5781±.0150	.5689±.0121	.5805±.0145	**.5594**±.0082	.5881±.0253	.5691±.0180	.5746±.0201	.5683±.0196	.5789±.0217	**.5549**±.0132
	720	.7278±.0214	.7021±.0120	.6719±.0083	.6597±.0123	.7325±.0143	**.6590**±.0076	.7327±.0177	.6931±.0071	.6950±.0071	.6625±.0096	.7402±.0090	**.6621**±.0057
ETTh2	96	.2781±.0247	.2754±.0237	.2747±.0216	.2766±.0244	.2715±.0251	**.2665**±.0144	.2456±.0059	.2442±.0053	.2476±.0043	.2445±.0051	**.2426**±.0053	.2446±.0059
	192	.3177±.0051	.3102±.0073	.3086±.0063	.3140±.0055	.3088±.0063	**.2899**±.0068	.2892±.0039	.2857±.0042	.2872±.0038	.2896±.0034	.2897±.0037	**.2748**±.0049
	336	.3563±.0159	.3453±.0145	.3435±.0129	.3484±.0135	.3433±.0145	**.3165**±.0138	.3246±.0067	.3154±.0040	.3172±.0043	.3213±.0034	.3209±.0047	**.2961**±.0063
	720	.4222±.0134	.4049±.0109	.4014±.0111	.3936±.0072	.4035±.0104	**.3550**±.0055	.4113±.0177	.3958±.0101	.3943±.0111	.3778±.0063	.3938±.0080	**.3435**±.0030
ETTm1	96	.3966±.0141	.3706±.0130	.3751±.0119	.3594±.0086	.3463±.0074	**.3351**±.0050	.4025±.0069	.3915±.0055	.4183±.0040	.3694±.0020	.3476±.0024	**.3381**±.0013
	192	.4496±.0122	.4258±.0077	.4269±.0066	.4175±.0039	.4047±.0051	**.3866**±.0026	.4595±.0065	.4456±.0054	.4622±.0054	.4293±.0041	.4097±.0040	**.3930**±.0015
	336	.5281±.0157	.4849±.0079	.4843±.0065	.4849±.0058	.4793±.0081	**.4634**±.0057	.5037±.0056	.4880±.0035	.4994±.0039	.4807±.0036	.4688±.0048	**.4577**±.0030
	720	.6131±.0059	.5649±.0038	.5566±.0025	.5642±.0009	.5450±.0036	**.5268**±.0028	.5693±.0052	.5457±.0031	.5523±.0027	.5373±.0008	.5210±.0015	**.5101**±.0023
ETTm2	96	.1715±.0049	.1691±.0046	.1674±.0024	.1713±.0051	.1610±.0037	**.1560**±.0025	.1594±.0021	.1588±.0018	.1620±.0019	.1587±.0017	.1508±.0015	**.1492**±.0009
	192	.2192±.0052	.2162±.0052	.2137±.0040	.2193±.0060	.2010±.0046	**.1926**±.0026	.2039±.0066	.2035±.0060	.2050±.0064	.2055±.0054	.1913±.0037	**.1846**±.0026
	336	.2610±.0082	.2555±.0069	.2505±.0070	.2607±.0054	.2450±.0032	**.2400**±.0034	.2483±.0038	.2483±.0032	.2487±.0035	.2514±.0032	.2455±.0020	**.2323**±.0021
	720	.3329±.0098	.3249±.0080	.3174±.0076	.3285±.0067	.3129±.0068	**.2863**±.0061	.3264±.0064	.3265±.0055	.3256±.0045	.3191±.0034	.3212±.0026	**.2830**±.0023
Weather	96	.1732±.0025	.1680±.0014	.1671±.0015	.1666±.0022	.1620±.0022	**.1604**±.0018	.1739±.0009	.1720±.0008	.1765±.0007	.1692±.0009	.1657±.0007	**.1647**±.0007
	192	.2214±.0014	.2136±.0016	.2127±.0013	.2131±.0016	.2069±.0017	**.2022**±.0015	.2207±.0014	.2159±.0012	.2196±.0012	.2137±.0011	.2090±.0011	**.2042**±.0009
	336	.2790±.0014	.2667±.0012	.2652±.0015	.2699±.0015	.2592±.0010	**.2533**±.0009	.2765±.0014	.2660±.0014	.2710±.0014	.2695±.0013	.2596±.0013	**.2544**±.0011
	720	.3551±.0022	.3439±.0021	.3479±.0070	.3292±.0018	.3308±.0013	**.3175**±.0011	.3546±.0015	.3376±.0012	.3436±.0016	.3309±.0012	.3309±.0011	**.3167**±.0007
Exchange Rate	96	.0999±.0069	.0972±.0059	.0976±.0057	.0975±.0064	.0971±.0062	**.0949**±.0056	.0877±.0050	.0858±.0045	.0896±.0045	.0849±.0043	.0846±.0039	**.0835**±.0038
	192	.2014±.0137	.1956±.0139	.1977±.0132	.1882±.0113	.1918±.0110	**.1832**±.0083	.1828±.0049	.1785±.0042	.1828±.0044	**.1714**±.0034	.1738±.0027	.1723±.0033
	336	.3578±.0138	.3479±.0132	.3512±.0139	**.3206**±.0124	.3492±.0126	.3220±.0080	.3639±.0273	.3531±.0256	.3619±.0269	.3308±.0237	.3487±.0266	**.3289**±.0180
	720	.8842±.1041	.8754±.1016	.8788±.1023	**.6349**±.0744	.8792±.0990	.7382±.0609	.9859±.0924	.9850±.0924	.9865±.0927	**.7419**±.0914	.9804±.0924	.7851±.0452
Solar	96	.2594±.0474	.2471±.0367	**.2412**±.0324	.2545±.0435	.2454±.0360	.2456±.0351	.2267±.0090	.2236±.0071	.2702±.0092	.2264±.0087	**.2209**±.0075	.2214±.0071
	192	.2657±.0157	.2631±.0150	.2588±.0144	.2622±.0149	**.2506**±.0128	**.2506**±.0119	.2517±.0034	.2493±.0033	.2719±.0030	.2517±.0034	**.2436**±.0031	.2450±.0030
	336	.2767±.0099	.2750±.0095	.2724±.0085	.2721±.0089	.2608±.0074	**.2579**±.0067	.2685±.0025	.2654±.0020	.2785±.0022	.2673±.0024	.2593±.0019	**.2583**±.0016
	720	.2715±.0048	.2710±.0047	.2698±.0044	.2604±.0039	.2458±.0021	**.2359**±.0016	.2713±.0024	.2686±.0018	.2730±.0022	.2627±.0023	.2464±.0013	**.2368**±.0011
Electricity	96	.1584±.0014	.1575±.0013	.1588±.0016	.1583±.0014	.1545±.0013	**.1532**±.0012	.1771±.0006	.1750±.0003	.1938±.0003	.1770±.0006	**.1717**±.0003	.1724±.0006
	192	.1724±.0009	.1719±.0008	.1724±.0009	.1721±.0009	.1696±.0008	**.1650**±.0008	.1842±.0010	.1828±.0006	.1941±.0008	.1839±.0010	.1801±.0008	**.1763**±.0008
	336	.1871±.0012	.1867±.0012	.1867±.0013	.1851±.0012	.1768±.0007	**.1702**±.0004	.2015±.0015	.1980±.0009	.2062±.0013	.1992±.0014	.1858±.0002	**.1781**±.0013
	720	.2340±.0088	.2295±.0109	.2314±.0110	.2238±.0086	.2308±.0087	**.2211**±.0097	.2423±.0010	.2368±.0005	.2437±.0003	**.2319**±.0010	.2423±.0006	.2364±.0009
Traffic	96	.4194±.0059	.4190±.0059	.4210±.0061	.4190±.0059	.4135±.0054	**.4097**±.0053	.4571±.0021	.4552±.0025	.4724±.0024	.4557±.0019	.4479±.0022	**.4459**±.0018
	192	.4422±.0050	.4419±.0051	.4435±.0052	.4419±.0051	.4370±.0047	**.4287**±.0043	.4683±.0015	.4658±.0011	.4766±.0011	.4679±.0018	.4612±.0013	**.4538**±.0015
	336	.4582±.0061	.4581±.0061	.4589±.0062	.4580±.0061	.4431±.0047	**.4302**±.0037	.4844±.0009	.4826±.0011	.4888±.0010	.4836±.0009	.4610±.0006	**.4451**±.0003
	720	.4910±.0073	.4909±.0073	.4916±.0074	.4901±.0074	.4955±.0069	**.4880**±.0063	.5173±.0009	.5157±.0011	.5203±.0012	.5156±.0010	.5190±.0010	**.5123**±.0007
Avg.		.3569±.0029	.3469±.0024	.3456±.0023	.3370±.0018	.3409±.0027	**.3240**±.0017	.3562±.0036	.3490±.0032	.3576±.0031	.3388±.0032	.3434±.0036	**.3258**±.0023

(b) DLinear / OLS

	H	DLinear						OLS					
		None	TFS	PTS	COS	SWGD	SWiT	None	TFS	PTS	COS	SWGD	SWiT
ETTh1	96	.4595±.0000	.4527±.0000	.4498±.0003	.4531±.0000	.4396±.0000	**.4308**±.0000	.4868	.4808	.4777	.4807	.4597	**.4454**
	192	.5124±.0001	.5045±.0001	.5034±.0003	.5062±.0000	.4961±.0000	**.4772**±.0000	.5379	.5309	.5296	.5316	.5143	**.4906**
	336	.5596±.0035	.5550±.0034	.5529±.0036	.5522±.0035	.5514±.0017	**.5347**±.0015	.5814	.5781	.5737	.5741	.5683	**.5478**
	720	.7065±.0029	.6894±.0031	.6686±.0037	.6661±.0031	.7113±.0017	**.6463**±.0014	.7271	.7128	.6910	.6868	.7291	**.6610**
ETTh2	96	.2307±.0009	.2287±.0009	.2292±.0007	.2304±.0009	**.2267**±.0003	.2309±.0001	.2456	.2438	.2442	.2451	.2371	**.2370**
	192	.2847±.0008	.2817±.0009	.2817±.0007	.2859±.0009	.2854±.0004	**.2632**±.0002	.2986	.2962	.2961	.2995	.2959	**.2703**
	336	.3246±.0004	.3185±.0003	.3181±.0003	.3236±.0005	.3228±.0001	**.2900**±.0001	.3374	.3320	.3307	.3367	.3326	**.2971**
	720	.4098±.0006	.3933±.0007	.3861±.0002	.3839±.0001	.3971±.0002	**.3429**±.0002	.4201	.4057	.3983	.3937	.4077	**.3513**
ETTm1	96	.3709±.0004	.3494±.0002	.3523±.0003	.3475±.0002	.3347±.0001	**.3291**±.0001	.3757	.3545	.3572	.3520	.3371	**.3312**
	192	.4436±.0004	.4175±.0001	.4188±.0003	.4175±.0001	.3993±.0001	**.3861**±.0001	.4476	.4218	.4237	.4212	.4013	**.3878**
	336	.5180±.0002	.4826±.0001	.4806±.0004	.4873±.0001	.4745±.0002	**.4664**±.0002	.5213	.4862	.4844	.4902	.4770	**.4689**
	720	.5926±.0002	.5562±.0000	.5496±.0001	.5581±.0001	.5381±.0001	**.5238**±.0001	.5945	.5592	.5520	.5596	.5396	**.5253**
ETTm2	96	.1599±.0002	.1588±.0001	.1589±.0001	.1596±.0001	.1500±.0001	**.1499**±.0000	.1617	.1606	.1606	.1615	.1510	**.1508**
	192	.1931±.0002	.1921±.0002	.1922±.0003	.1944±.0003	.1814±.0001	**.1785**±.0000	.1948	.1936	.1939	.1956	.1824	**.1794**
	336	.2325±.0002	.2300±.0002	.2304±.0004	.2349±.0003	.2284±.0002	**.2242**±.0001	.2342	.2317	.2322	.2365	.2296	**.2253**
	720	.3062±.0001	.2995±.0002	.2964±.0002	.3046±.0001	.2959±.0004	**.2695**±.0001	.3072	.3015	.2986	.3038	.2964	**.2703**
Weather	96	.1953±.0001	.1878±.0001	.1839±.0009	.1869±.0001	.1821±.0001	**.1810**±.0001	.1972	.1903	.1872	.1889	.1835	**.1824**
	192	.2402±.0001	.2291±.0001	.2280±.0007	.2290±.0001	.2224±.0001	**.2163**±.0001	.2420	.2318	.2314	.2306	.2238	**.2176**
	336	.2918±.0001	.2770±.0001	.2768±.0004	.2808±.0001	.2698±.0001	**.2639**±.0001	.2932	.2796	.2786	.2821	.2710	**.2650**
	720	.3646±.0002	.3501±.0001	.3513±.0009	.3387±.0002	.3369±.0001	**.3218**±.0001	.3655	.3527	.3520	.3394	.3375	**.3226**
Exchange Rate	96	.0872±.0023	.0856±.0023	.0854±.0022	.0852±.0022	.0857±.0023	**.0842**±.0020	.0910	.0894	.0893	.0889	.0893	**.0872**
	192	.1778±.0027	.1742±.0027	.1750±.0027	**.1688**±.0025	.1757±.0028	.1725±.0023	.1811	.1777	.1787	**.1715**	.1790	.1751
	336	.3231±.0026	.3161±.0027	.3176±.0028	**.2926**±.0024	.3152±.0025	.2957±.0017	.3255	.3190	.3210	**.2950**	.3179	.2975
	720	.8484±.0217	.8458±.0219	.8425±.0238	**.5974**±.0129	.8402±.0195	.7309±.0236	.8421	.8398	.8374	**.5921**	.8346	.7240
Solar	96	.3238±.0378	.3110±.0310	.3101±.0292	.3177±.0353	.2968±.0328	**.2904**±.0284	.3330	.3202	.3180	.3237	.3016	**.2966**
	192	.3601±.0322	.3466±.0252	.3429±.0237	.3546±.0288	.3215±.0268	**.3161**±.0258	.3697	.3573	.3501	.3614	.3256	**.3198**
	336	.3989±.0263	.3873±.0205	.3821±.0187	.3938±.0232	.3567±.0212	**.3489**±.0208	.4067	.3969	.3870	.4002	.3605	**.3523**
	720	.3998±.0208	.3927±.0173	.3892±.0140	.3828±.0186	.3146±.0090	**.2802**±.0053	.4052	.4001	.3947	.3872	.3171	**.2821**
Electricity	96	.2109±.0070	.2089±.0070	.2080±.0069	.2105±.0070	.1966±.0026	**.1937**±.0021	.1974	.1923	.1926	.1970	.1925	**.1898**
	192	.2113±.0072	.2102±.0072	.2085±.0072	.2109±.0072	.1974±.0030	**.1895**±.0022	.1972	.1937	.1935	.1967	.1924	**.1852**
	336	.2259±.0069	.2250±.0069	.2231±.0069	.2236±.0069	.1952±.0022	**.1833**±.0017	.2119	.2083	.2083	.2094	.1925	**.1803**
	720	.2673±.0066	.2666±.0067	.2674±.0067	.2566±.0066	.2598±.0032	**.2501**±.0028	.2536	.2504	.2570	**.2428**	.2535	.2438
Traffic	96	.6795±.0192	.6789±.0191	.6795±.0191	.6792±.0191	.6498±.0120	**.6400**±.0105	.6447	.6398	.6372	.6444	.6287	**.6173**
	192	.6343±.0204	.6340±.0204	.6344±.0203	.6341±.0204	.6091±.0139	**.5719**±.0098	.5978	.5939	.5908	.5975	.5851	**.5539**
	336	.6418±.0202	.6416±.0202	.6417±.0201	.6416±.0202	.5473±.0027	**.4931**±.0040	.6051	.6013	.5974	.6049	.5481	**.4878**
	720	.6807±.0202	.6806±.0202	.6806±.0200	.6798±.0202	.6717±.0166	**.6424**±.0148	.6429	.6405	.6379	.6420	.6404	**.6132**
Avg.		.3852±.0073	.3766±.0066	.3749±.0064	.3686±.0067	.3633±.0049	**.3447**±.0045	.3854	.3768	.3746	.3685	.3648	**.3454**

Table 19: Full results (MSE, mean±std over seeds, the Avg. row reports the std of the per-seed averages) on MLP-based backbones. **Bold** = best, underline = second among TTA methods. TFS, TAFAS (Kim et al., 2025); PTS, PETSA (Medeiros et al., 2025); COS, COSA (Im & Kwon, 2026); SWGD, SWiT-GD.

(a) FreTS / MICN

	H	FreTS						MICN					
		None	TFS	PTS	COS	SWGD	SWiT	None	TFS	PTS	COS	SWGD	SWiT
ETTh1	96	.4442±.0012	.4395±.0011	.4378±.0012	.4403±.0011	.4328±.0010	**.4234**±.0010	.6261±.0177	.5816±.0137	.5800±.0133	.5839±.0133	.5292±.0060	**.4778**±.0025
	192	.5007±.0008	.4950±.0007	.4944±.0003	.4968±.0006	.4933±.0004	**.4758**±.0002	.7044±.0377	.6473±.0297	.6534±.0271	.6495±.0308	.6206±.0184	**.5498**±.0112
	336	.5537±.0007	.5512±.0009	.5507±.0007	.5464±.0006	.5554±.0005	**.5405**±.0006	.7980±.0464	.7383±.0433	.7287±.0372	.7135±.0367	.6935±.0237	**.6227**±.0168
	720	.7195±.0015	.7000±.0014	.6780±.0050	.6726±.0016	.7278±.0013	**.6583**±.0009	.9405±.0138	.9119±.0193	.8803±.0158	.8020±.0169	.8684±.0084	**.7688**±.0084
ETTh2	96	.2403±.0016	.2385±.0015	.2382±.0015	.2396±.0017	**.2370**±.0018	.2401±.0013	.2554±.0050	.2499±.0043	**.2476**±.0040	.2526±.0045	.2493±.0034	.2506±.0028
	192	.2884±.0018	.2851±.0013	.2846±.0015	.2883±.0016	.2866±.0018	**.2722**±.0019	.3179±.0145	.3094±.0121	.3102±.0109	.3157±.0129	.3137±.0114	**.2923**±.0081
	336	.3261±.0037	.3192±.0030	.3182±.0033	.3237±.0029	.3213±.0024	**.2959**±.0012	.3792±.0069	.3688±.0100	.3679±.0100	.3723±.0101	.3646±.0077	**.3345**±.0069
	720	.4146±.0052	.3965±.0042	.3892±.0040	.3831±.0048	.3965±.0029	**.3459**±.0015	.4681±.0073	.4581±.0085	.4551±.0090	.4354±.0088	.4460±.0073	**.3985**±.0077
ETTm1	96	.3670±.0004	.3576±.0007	.3572±.0009	.3542±.0004	.3374±.0003	**.3296**±.0004	.4577±.0131	.4252±.0105	.4271±.0147	.3986±.0037	.3648±.0023	**.3485**±.0014
	192	.4312±.0008	.4208±.0008	.4188±.0008	.4197±.0009	.3989±.0006	**.3854**±.0004	.5154±.0116	.4817±.0133	.4828±.0148	.4642±.0104	.4238±.0036	**.4016**±.0025
	336	.4970±.0020	.4828±.0017	.4767±.0017	.4870±.0017	.4671±.0016	**.4575**±.0017	.5645±.0152	.5283±.0087	.5248±.0062	.5226±.0057	.5017±.0060	**.4832**±.0047
	720	.5657±.0026	.5522±.0022	.5413±.0021	.5537±.0016	.5270±.0020	**.5134**±.0018	.6440±.0144	.6096±.0138	.6012±.0136	.6009±.0102	.5680±.0086	**.5491**±.0083
ETTm2	96	.1577±.0003	.1570±.0002	.1569±.0004	.1577±.0003	.1489±.0001	**.1483**±.0003	.1660±.0084	.1636±.0058	.1658±.0054	.1649±.0084	.1538±.0054	**.1506**±.0035
	192	.1930±.0009	.1916±.0007	.1917±.0004	.1931±.0006	.1807±.0004	**.1783**±.0002	.2123±.0077	.2083±.0077	.2086±.0070	.2125±.0075	.1935±.0048	**.1888**±.0030
	336	.2324±.0004	.2304±.0004	.2303±.0002	.2353±.0008	.2254±.0005	**.2244**±.0002	.2516±.0077	.2450±.0095	.2447±.0083	.2489±.0099	.2365±.0073	**.2344**±.0044
	720	.3020±.0005	.2969±.0007	.2928±.0007	.2972±.0003	.2901±.0010	**.2688**±.0006	.3401±.0082	.3349±.0084	.3249±.0064	.3452±.0092	.3246±.0079	**.2950**±.0051
Weather	96	.1836±.0011	.1761±.0013	.1757±.0018	.1784±.0010	.1735±.0010	**.1720**±.0011	.1745±.0017	.1721±.0008	.1787±.0018	.1698±.0017	.1663±.0010	**.1656**±.0013
	192	.2296±.0008	.2190±.0009	.2193±.0007	.2220±.0009	.2155±.0007	**.2098**±.0006	.2236±.0020	.2205±.0027	.2287±.0028	.2169±.0020	.2132±.0017	**.2104**±.0017
	336	.2878±.0058	.2745±.0059	.2730±.0052	.2781±.0045	.2672±.0039	**.2617**±.0034	.2813±.0011	.2738±.0009	.2860±.0035	.2746±.0009	.2658±.0009	**.2623**±.0009
	720	.3592±.0005	.3480±.0006	.3537±.0058	.3346±.0004	.3338±.0003	**.3197**±.0002	.3521±.0014	.3507±.0039	.3649±.0034	.3292±.0014	.3365±.0014	**.3223**±.0014
Exchange Rate	96	.0833±.0005	.0813±.0005	.0812±.0003	.0808±.0004	.0803±.0005	**.0794**±.0004	.0961±.0037	.0922±.0039	.0921±.0045	.0925±.0028	.0921±.0024	**.0895**±.0021
	192	.1751±.0019	.1708±.0016	.1714±.0016	**.1646**±.0013	.1679±.0011	.1669±.0012	.2044±.0101	.1919±.0079	.1952±.0070	.1896±.0062	.1929±.0045	**.1884**±.0035
	336	.3247±.0006	.3162±.0005	.3185±.0004	**.2929**±.0006	.3080±.0001	.2969±.0004	.4249±.0551	**.3519**±.0429	.3672±.0450	.3648±.0420	.3900±.0439	.3660±.0390
	720	.8394±.0062	.8364±.0063	.8319±.0065	**.5848**±.0029	.8335±.0056	.7144±.0065	1.4223±.3268	1.2576±.2921	1.2454±.2795	**1.0177**±.2712	1.3757±.3031	1.1206±.2468
Solar	96	.2402±.0053	.2277±.0042	**.2247**±.0046	.2400±.0053	.2322±.0046	.2330±.0044	.2208±.0033	.2220±.0034	.2284±.0041	.2206±.0033	.2157±.0034	**.2152**±.0030
	192	.2777±.0038	.2652±.0035	.2627±.0052	.2778±.0040	**.2624**±.0034	.2628±.0031	.2602±.0020	.2601±.0019	.2650±.0009	.2600±.0022	**.2493**±.0018	**.2493**±.0019
	336	.3069±.0084	.2962±.0079	.2946±.0077	.3064±.0086	.2894±.0070	**.2864**±.0066	.2828±.0032	.2824±.0026	.2864±.0022	.2821±.0030	.2684±.0026	**.2655**±.0026
	720	.3084±.0093	.3015±.0087	.3005±.0087	.2999±.0093	.2703±.0051	**.2512**±.0034	.3002±.0079	.3005±.0074	.3023±.0096	.2908±.0089	.2655±.0064	**.2488**±.0048
Electricity	96	.1889±.0013	.1864±.0014	.1855±.0014	.1886±.0013	.1835±.0011	**.1811**±.0010	.1759±.0046	.1704±.0012	.1732±.0017	.1747±.0041	.1571±.0015	**.1523**±.0010
	192	.1930±.0011	.1917±.0011	.1904±.0010	.1926±.0011	.1881±.0009	**.1819**±.0007	.1897±.0035	.1820±.0006	.1835±.0009	.1875±.0030	.1712±.0012	**.1634**±.0011
	336	.2093±.0008	.2082±.0008	.2068±.0008	.2070±.0008	.1906±.0003	**.1796**±.0003	.1998±.0056	.1917±.0023	.1928±.0024	.1952±.0048	.1795±.0022	**.1727**±.0027
	720	.2525±.0011	.2516±.0011	.2531±.0012	**.2417**±.0011	.2526±.0009	.2433±.0008	.2200±.0028	.2150±.0017	.2190±.0024	.2088±.0022	.2102±.0017	**.2027**±.0009
Traffic	96	.5305±.0125	.5296±.0124	.5299±.0125	.5300±.0125	.5193±.0109	**.5121**±.0102	.4834±.0017	.4718±.0015	.4796±.0016	.4741±.0005	.4499±.0006	**.4419**±.0007
	192	.5272±.0115	.5269±.0116	.5271±.0116	.5270±.0115	.5183±.0105	**.5009**±.0089	.5000±.0026	.4857±.0010	.4932±.0017	.4885±.0013	.4669±.0011	**.4508**±.0005
	336	.5498±.0011	.5497±.0011	.5494±.0011	.5494±.0010	.5129±.0008	**.4734**±.0004	.5229±.0023	.5069±.0013	.5133±.0017	.5081±.0013	.4645±.0013	**.4484**±.0009
	720	.5931±.0017	.5930±.0017	.5928±.0017	.5917±.0016	.5925±.0016	**.5759**±.0015	.5722±.0012	.5524±.0012	.5565±.0013	.5476±.0005	.5466±.0007	**.5330**±.0004
Avg.		.3582±.0012	.3518±.0011	.3500±.0011	.3438±.0012	.3449±.0011	**.3295**±.0009	.4097±.0116	.3893±.0105	.3904±.0106	.3771±.0099	.3758±.0098	**.3504**±.0081

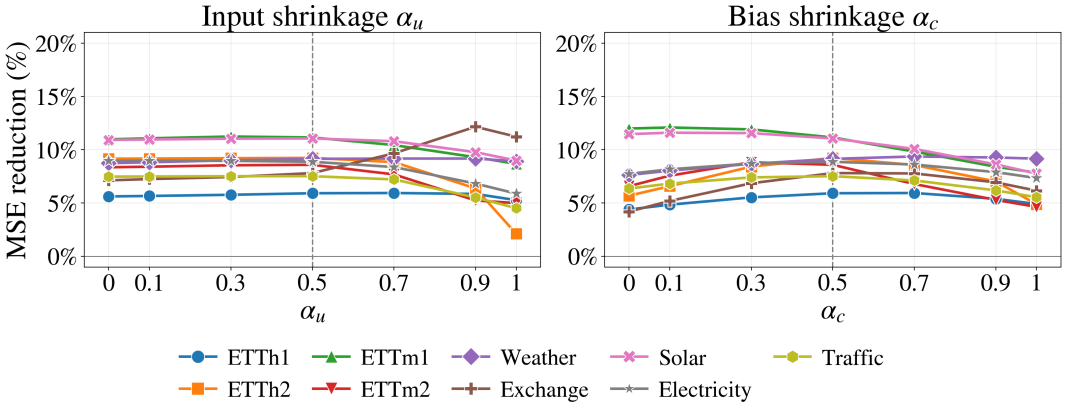


Figure 4: Shrinkage sensitivity per dataset (same convention as Figure 3).

rotational bias is the plain least-squares solution, and with $\alpha_c = 1$, the bias is switched off. Both ends are below the middle, 7.9% at $\alpha_c = 0$ and 6.9% at $\alpha_c = 1$, while $[0.1, 0.7]$ gives 8.5–9.5%. Because both shrinkage rates give nearly the same result anywhere inside these ranges, we set $\alpha_u = \alpha_c = 1/2$.

Table 20: Main experimental results in MAE. TFS, TAFAS (Kim et al., 2025); PTS, PETSA (Medeiros et al., 2025); COS, COSA (Im & Kwon, 2026); SWGD, SWiT-GD.

(a) Per-dataset MAE averaged over four horizons.

	Transformer-based: iTransformer						Linear-based: DLinear						MLP-based: FreTS					
Dataset	None	TFS	PTS	COS	SWGD	SWiT	None	TFS	PTS	COS	SWGD	SWiT	None	TFS	PTS	COS	SWGD	SWiT
ETTh1	.535	.528	.526	.523	.528	**.517**	.521	.518	.515	.515	.514	**.504**	.520	.517	.514	.513	.517	**.508**
ETTh2	.398	.391	.389	.392	.390	**.373**	.370	.366	.364	.368	.367	**.354**	.378	.373	.371	.375	.374	**.361**
ETTm1	.473	.457	.460	.454	.450	**.443**	.451	.435	.436	.436	**.430**	**.430**	.448	.443	.443	.444	.436	**.434**
ETTm2	.328	.325	.322	.325	.314	**.307**	.310	.308	.307	.308	.301	**.297**	.309	.308	.307	.308	.301	**.297**
Exchange	.423	.419	.420	**.396**	.420	.403	.403	.401	.401	**.377**	.402	.392	.399	.396	.396	**.373**	.396	.386
Weather	.278	.276	.278	**.275**	.279	.277	.291	**.288**	**.288**	**.288**	.292	.289	.284	**.280**	.282	.281	.286	.283
Solar	.289	**.287**	**.287**	.289	.305	.303	.358	**.356**	**.356**	.359	.363	.358	.302	.301	**.298**	.304	.315	.312
Electricity	.278	.276	.278	.276	.277	**.273**	.312	.312	.313	.311	.300	**.296**	.291	.290	.291	.289	.290	**.287**
Traffic	.303	.303	.305	.303	.303	**.301**	.402	.402	.402	.402	.388	**.379**	.357	.357	.357	.356	.353	**.346**
Average	.367	.362	.363	.359	.363	**.355**	.380	.376	.376	.374	.373	**.366**	.365	.363	.362	.360	.363	**.357**

	Transformer-based: PatchTST						Linear-based: OLS						MLP-based: MICN					
Dataset	None	TFS	PTS	COS	SWGD	SWiT	None	TFS	PTS	COS	SWGD	SWiT	None	TFS	PTS	COS	SWGD	SWiT
ETTh1	.537	.531	.533	.526	.532	**.520**	.537	.534	.530	.531	.526	**.513**	.607	.590	.587	.578	.573	**.547**
ETTh2	.379	.373	.374	.374	.375	**.362**	.381	.377	.375	.379	.375	**.360**	.405	.398	.396	.400	.396	**.382**
ETTm1	.464	.455	.462	.450	.443	**.440**	.454	.438	.440	.438	**.431**	**.431**	.499	.482	.482	.476	.463	**.453**
ETTm2	.317	.317	.318	.315	.310	**.301**	.311	.310	.309	.310	.302	**.298**	.323	.320	.318	.321	.311	**.305**
Exchange	.420	.417	.421	**.396**	.417	.400	.406	.404	.404	**.381**	.405	.394	.471	.443	.445	**.434**	.460	.438
Weather	.279	.282	.282	**.276**	.282	.279	.293	**.289**	.290	.290	.293	.290	.283	.285	.292	**.281**	.287	.284
Solar	.300	**.299**	.312	**.299**	.303	.306	.367	.365	**.364**	.367	.369	.365	.276	**.276**	.280	**.276**	.288	.284
Electricity	.287	.286	.295	.285	**.284**	.285	.292	.294	.296	.290	.292	**.289**	.297	.291	.293	.294	.284	**.276**
Traffic	.309	**.306**	.313	.309	.307	.310	.371	.371	.373	.371	.367	**.362**	.318	.317	.318	.317	.313	**.310**
Average	.366	.363	.368	.359	.361	**.356**	.379	.376	.376	.373	.373	**.367**	.386	.378	.379	.375	.375	**.364**

(b) Horizon-wise average MAE.

H	None	TFS	PTS	COS	SWGD	SWiT
96	.3135	.3102	.3121	.3103	.3077	**.3057**
192	.3465	.3427	.3442	.3430	.3409	**.3368**
336	.3827	.3779	.3790	.3782	.3764	**.3715**
720	.4530	.4475	.4472	.4357	.4474	**.4301**
Avg.	.3739	.3696	.3706	.3668	.3681	**.3610**

Table 21: Hyperparameter selection on the validation splits of ETTh1 and ETTm2. Mean rank is the average rank of a combination over the 8 cells. Mean MSE is the average validation MSE. None is the backbone without adaptation.

	Combination					Mean validation MSE		
	α_u	α_c	δ_u	δ_s	Mean rank	All	ETTh1	ETTm2
–	None				783.25	.3580	.4658	.2501
1	0.5	0.5	0.3	1/2	**198.38**	.3385	.4374	.2397
2	0.5	0.3	0.2	1/2	202.25	.3391	.4396	**.2387**
3	0.5	0.7	0.3	1/2	204.12	**.3376**	**.4340**	.2412
4	0.3	0.5	0.3	1/2	204.38	.3386	.4375	.2397
5	0.3	0.7	0.3	1/2	204.62	.3377	.4342	.2412

I Qualitative Case Study

I.1 Coefficient Trajectories

Figure 5 shows, for one channel of three streams, the coefficients s_k, u_k, $|c_k|$ and the gate weight w_k that SWiT produces at every adaptation batch, as heat maps over time $\times$ frequency bin (DLinear, seed 0).

Table 22: Clipping sensitivity, per-horizon MSE.

		δ_s				δ_u				
H	None	$1/4$	$\mathbf{1/2}$	1	∞	0.1	0.2	**0.3**	0.5	∞
96	.2809	.2694	.2632	**.2630**	.2732	.2644	.2634	**.2632**	.2640	.2648
192	.3236	.3112	**.2967**	.3014	.3145	.3018	.2993	**.2967**	.3004	.3032
336	.3765	.3605	**.3391**	.3475	.3655	.3446	.3394	**.3391**	.3473	.3670
720	.4951	.4793	**.4376**	.4481	.4665	.4354	**.4341**	.4376	.4567	.4658
Avg.	.3690	.3551	**.3341**	.3400	.3549	.3365	**.3341**	.3341	.3421	.3502

Table 23: Shrinkage sensitivity, per-horizon MSE.

		α_u							α_c						
H	None	0	0.1	0.3	**0.5**	0.7	0.9	1	0	0.1	0.3	**0.5**	0.7	0.9	1
96	.2809	.2640	.2638	.2633	**.2632**	.2641	.2655	.2663	.2631	.2626	**.2622**	.2632	.2652	.2676	.2688
192	.3236	.2973	.2971	.2968	**.2967**	.2981	.3026	.3043	.2994	.2979	**.2962**	.2967	.2997	.3033	.3049
336	.3765	.3401	.3399	.3395	**.3391**	.3392	.3439	.3486	.3433	.3416	.3393	**.3391**	.3407	.3454	.3487
720	.4951	.4387	.4386	.4385	.4376	.4337	**.4324**	.4436	.4533	.4493	.4417	**.4376**	.4399	.4446	.4511
Avg.	.3690	.3350	.3349	.3345	.3341	**.3338**	.3361	.3407	.3398	.3378	.3349	**.3341**	.3364	.3402	.3434

Behavior of s, u, and c. s splits to either side of 1 from bin to bin, a sizeable fraction of bins sit at the ends of the clipping range, $1/2$ or $3/2$, and on ETTm1 many bins stay near the lower bound for roughly the first 70 batches before crossing above 1. u starts with mixed signs across bins and, as samples accumulate, settles to a stable sign and a small magnitude per bin. $|\hat{c}|$ concentrates on the DC bin and the few bins above it and is more than an order of magnitude smaller elsewhere, consistent with the ablation in Appendix E that the rotational bias mainly performs level correction.

Opening and closing of the gate. The gate opens and closes per bin. During the first H steps, when no window has matured, and for a few batches thereafter, every bin is closed, and the backbone output passes through unchanged (the white vertical band in the figure). As scores accumulate, the bins open individually; of the bins that open, 59% close again at some later batch on ETTm1 channel 4, 65% on ETTh2 channel 5, and 45% on Exchange channel 7, and some reopen after closing. At the end of the stream, 94% of the bins are open on ETTm1, whereas only 37–39% are open on ETTh2 and Exchange, and the rest emit the backbone output.

J Comparison with Online Learning Methods

The main comparison is restricted to TTA methods that keep the backbone frozen (Sec. 2). This section compares SWiT with two online learners that update the backbone parameters at test time, FSNet (Pham et al., 2022) and PROCEED (Zhao & Shen, 2025), under both their protocol and ours. Following PROCEED, the backbone is a temporal convolutional network (TCN) with RevIN, the lookback is $W = 60$, the horizon is $H = 96$, and the five datasets are those for which PROCEED publishes per-dataset hyperparameters. We use PROCEED's official configuration throughout, including its choice to disable RevIN on Weather, and SWiT uses $B = 48$ as in the main experiments. GT arrives with an H-step delay for every method. We replace FSNet's original one-step delay with the same H-step delay, since a one-step delay reveals values that have not yet been observed. The two protocols differ in the split: online learning uses 20:5:75, so the backbone is trained on one fifth of the series and the remainder is adapted online, whereas the TS-TTA protocol of Sec. 4 uses 6:2:2 for ETT and 7:1:2 otherwise. For each seed, the frozen baseline, FSNet, PROCEED, and SWiT share the same trained backbone. Results are the mean and standard deviation over three seeds.

Table 24 shows that the two families win in different regimes. Under the 20:5:75 split, the backbone is trained on one fifth of the series, and its frozen error is large, 15.55 on average and 61.72 on Electricity. Here, updating the backbone matters, and PROCEED is best on every dataset (3.73 on average), ahead of FSNet (4.42) and SWiT (5.51). SWiT still lowers the frozen error on every dataset, from 61.72 to 16.62 on Electricity, but a corrector that leaves the backbone untouched cannot recover

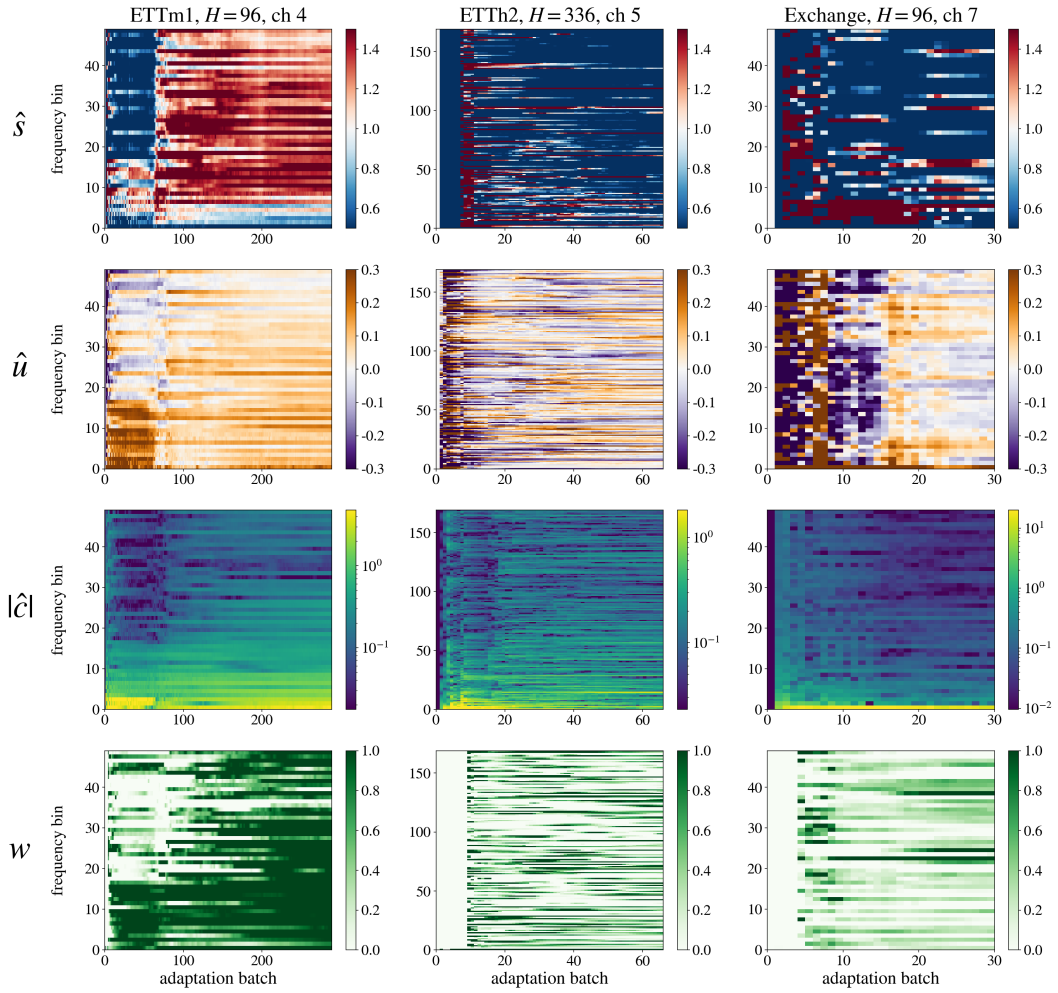


Figure 5: Coefficient trajectories of a single channel (adaptation batch × frequency bin, DLinear, seed 0). Rows: s, u, $|c|$ (log scale), gate weight w. Columns: ETTm1 H=96 (channel 4), ETTh2 H=336 (channel 5), Exchange H=96 (channel 7).

Table 24: MSE of online adaptation on a TCN backbone ($W = 60$, $H = 96$, mean±std over 3 seeds, same backbone per seed). Frozen is the backbone without adaptation (None in the main tables). FSNet (Pham et al., 2022) and PROCEED (Zhao & Shen, 2025) update the backbone online. **Bold** = best and underline = second among the three adaptive methods. Following PROCEED's official configuration, RevIN is disabled on Weather for every method. FSNet (Pham et al., 2022) and PROCEED (Zhao & Shen, 2025) update the backbone online.

Dataset	Split 20:5:75				Split 6:2:2 (ETT) / 7:1:2 (others)			
	Frozen	FSNet	PROCEED	SWiT	Frozen	FSNet	PROCEED	SWiT
ETTh2	6.2883±0.1237	6.0704±0.1610	**5.9450**±0.1257	6.2238±0.0202	0.3652±0.0152	0.4617±0.0125	0.3327±0.0170	**0.3267**±0.0083
ETTm1	1.7710±0.1764	1.1896±0.1869	**0.7819**±0.0020	0.9134±0.0485	0.4632±0.0283	0.6703±0.1238	0.3424±0.0068	**0.3209**±0.0063
Weather	7.2804±1.5075	1.6401±0.1119	**1.3293**±0.0348	3.2299±0.2874	0.4450±0.0103	0.2566±0.0667	**0.1586**±0.0017	0.2258±0.0013
Electricity	61.7214±0.2853	12.6430±0.5411	**10.0701**±0.3064	16.6184±0.2097	0.1671±0.0015	0.1841±0.0008	0.1677±0.0004	**0.1483**±0.0010
Traffic	0.6873±0.0106	0.5620±0.0021	**0.5257**±0.0232	0.5504±0.0051	0.6258±0.0063	0.5347±0.0015	0.5129±0.0128	**0.4743**±0.0099
Avg.	15.5497	4.4210	**3.7304**	5.5072	0.4132	0.4215	0.3029	**0.2992**

a backbone that is this far off. Under the TS-TTA split, the backbone is trained normally, and the picture reverses. SWiT has the lowest error on four of the five datasets and, on average, (.2992 against .3029 for PROCEED and .4132 for frozen), while FSNet is above the frozen backbone on average (.4215). The exception is Weather, where RevIN is disabled by PROCEED's configuration

and the frozen TCN is weak (.4450). PROCEED reaches .1586 and SWiT .2258. The comparison therefore places the two families side by side rather than ranking them: when the deployed backbone is barely trained, updating its parameters is necessary; when it is trained normally, correcting its output without touching the parameters is at least as accurate.