FROM WINDOWS TO REGIONS: REGIONAL GRADIENTS FOR CONTROLLING FORECASTING TRAJECTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Time-series forecasters train on overlapping sliding windows, where localized temporal segments repeatedly contribute to parameter updates. We reveal that while recurring high-loss regions provide constructive immediate gradients, their directional alignment with the remaining data collapses over training, causing multi-step momentum accumulation to exert severe trajectory drag. Physical data withholding confirms that this phenomenon is strictly anchored to temporal location, yet physically discarding observations risks forfeiting essential structural dynamics. To resolve this dilemma, we introduce *Regional First-Moment Memory Control ($\mathcal{H}$-short)*, shifting the intervention from data withholding to optimizer memory. Operating with complete data retention, $\mathcal{H}$-short dampens the accumulated momentum staleness of targeted temporal regions within Adam, absorbing current learning signals without inheriting outdated directional bias. Across diverse forecasters and benchmarks, $\mathcal{H}$-short consistently alleviates trajectory drag in expressive deep architectures while safeguarding critical physical transitions where data withholding fails, establishing source-aware memory control as a safe, non-destructive paradigm for temporal optimization.

1 INTRODUCTION

Time-series forecasting aims to learn temporal dynamics that reliably generalize beyond the observed history. Modern forecasters—ranging from linear models to Transformers and frequency-based architectures (Zeng et al., 2023; Liu et al., 2024; Nie et al., 2023; Yi et al., 2024)—are standardly trained via empirical risk minimization (ERM) over overlapping history–forecast windows. Because neighboring windows typically advance by a single stride, a local temporal segment redundantly appears in up to $L + H$ training pairs (Li et al., 2026). Rather than sampling independent instances, training repeatedly visits these contiguous temporal regions, turning each segment into a persistent gradient source that directly steers the optimization trajectory.

Regulating which training errors drive learning can improve generalization (Paul et al., 2021; Cho et al., 2022; Fu et al., 2025). However, existing interventions typically view high-loss samples statically—assuming that observations beneficial to the next update also benefit continued training. By tracking fixed *forecast origins* across epochs, we find that recurring high-loss regions are learnable (Section C.7) and improve all 27 one-step validation probes on a MixLinear/ETTh1 diagnostic (Section E.1). Yet, as training proceeds, their gradient alignment with the remaining data declines (Figure 1c, Section 3), and their multi-step accumulation through optimizer momentum degrades validation performance. Thus, a region's immediate learning signal can be useful even when its continued participation across subsequent updates is not.

To evaluate the effect of regional participation, we introduce *Robust Adaptive Masking with Past losses* (RAMP). RAMP fixes high-loss support $\mathcal{H}$ after one initial epoch using median absolute deviation (MAD) and withholds its subsequent gradients on the unnormalized empirical scale ($1/N$). Across five forecasters, eight datasets and four horizons, RAMP improves test MAE in 122/160 cells without per-cell tuning (Section 5.2). Controlled displacement tests show that shifting the support by sixteen origins retains 93% of its mean loss yet gives up 2.74 percentage points of relative MSE reduction across six configurations. Withholding therefore suppresses unfavorable later momentum by removing observations that may still provide useful current supervision.

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

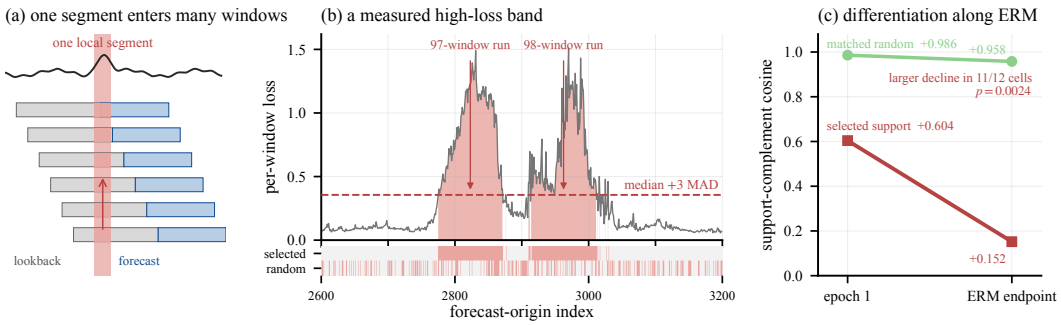


Figure 1: Temporal regions and their changing gradient contribution. **(a)** Overlapping windows repeatedly contain the same local segment. **(b)** First-epoch losses identify contiguous high-loss regions (iTransformer/ETTh2, $H = 96$, seed 1), versus equal-count random origins. **(c)** Across 12 paired measurements ($H = 96$), the selected group's alignment with remaining windows collapses at the ERM endpoint, while random groups remain aligned.

This dilemma motivates moving from data participation to **optimizer memory control**: rather than discarding informative observations, we seek to retain every window while regulating its accumulated influence. In momentum-based optimizers such as Adam, a region's persistent pull on the parameter trajectory is fundamentally carried by its exponentially smoothed first-moment history. To retain current learning signals without inheriting stale momentum, we introduce **Regional First-Moment Memory Control** ($\mathcal{H}$**-short**). Operating with complete data participation, $\mathcal{H}$-short decomposes Adam's first moment into separate regional accumulators for the selected support $\mathcal{H}$ and its complement $\mathcal{C}$. For region $\mathcal{H}$, we replace its accumulated historical momentum with its current gradient contribution scaled by the exact same cumulative coefficient, while keeping the complement's momentum, the full-data second-moment preconditioner, and the step norm unchanged. By resetting regional memory staleness at each step, the optimizer fully absorbs what difficult regions offer at the current parameter state, while preventing their past gradients from exerting outdated directional bias on subsequent updates. Our main contributions are summarized as follows:

- **Decoupling Immediate Value from Continued Participation**: We reveal that in sliding-window forecasting, recurring high-loss regions are learnable and provide gradients improving immediate one-step updates. However, as their gradients diverge from the remaining data over training, their continued multi-step participation degrades validation performance via momentum accumulation.
- **Controlled Regional Participation Analysis via RAMP**: We introduce RAMP as an effective diagnostic and training framework that withholds regional gradients on an unrenormalized empirical scale. Through matched temporal displacement controls, RAMP allows us to confirm that optimization benefits are tied to physical temporal location, while improving test MSE in 97/160 cells and MAE in 122/160 cells across a 5-model, 8-dataset benchmark.
- **Regional Memory Control with Full Participation** ($\mathcal{H}$**-short**): We propose $\mathcal{H}$-short, a principled optimizer intervention that eliminates harmful historical momentum while preserving 100% data participation. $\mathcal{H}$-short improves MSE over standard Adam in all 18/18 paired runs (by 1.05%–5.33%), outperforms equal-count random memory truncation by 2.85 percentage points (17/18 wins), and continuously scales with a tunable memory coefficient λ.

2 Related Work

Time-series training interventions. Deep forecasters capture temporal dependencies through decomposition, frequency transforms, or attention (Zeng et al., 2023; Nie et al., 2023; Liu et al., 2024; Yi et al., 2024). Because these models are standardly trained on sliding windows, recent methods regulate training by mitigating rolling-window redundancy (Li et al., 2026), bounding prediction errors (Cho et al., 2022), or masking target residuals within windows (Fu et al., 2025). However, these interventions operate primarily *intra-window* or at the instantaneous loss level, treating updates as isolated steps. In contrast, we track *fixed forecast origins* across training epochs, investigating how recurring temporal regions accumulate gradient momentum over long-term parameter trajectories.

Example selection and data pruning. Curriculum learning and robust training prioritize samples by difficulty or label reliability (Bengio et al., 2009; Kumar et al., 2010; Shen & Sanghavi, 2019), while pruning methods such as *Data Diet* (Paul et al., 2021) and forgetting dynamics (Toneva et al., 2019)

discard instances early based on loss or gradient norms. These approaches rely on an I.I.D. assumption and adopt an *all-or-nothing withholding* paradigm. In continuous time series, however, high-loss regions often contain critical structural patterns rather than mere noise. Completely discarding them suppresses harmful late-stage momentum accumulation, but also sacrifices informative supervision that benefits the current update.

Gradient control and optimizer dynamics. Gradient surgery methods, such as PCGrad (Yu et al., 2020) and CAGrad (Liu et al., 2021), project conflicting task gradients in multi-task optimization, while temporal attribution studies analyze instance influence across training stages (Wang et al., 2025). Yet, gradient projection addresses only instantaneous, step-local conflicts rather than cross-step momentum accumulation in adaptive optimizers like Adam (Kingma & Ba, 2015), and attribution methods remain post-hoc diagnostics. We bridge this gap by moving beyond data withholding: using RAMP as a diagnostic baseline for participation, we propose $\mathcal{H}$-short to directly reset regional first-moment memory, achieving full data participation while eliminating stale momentum.

3 Empirical Motivation: The Single-Step vs. Multi-Step Paradox

Time-series forecasters train via empirical risk minimization (ERM) on overlapping rolling windows, where a localized temporal segment redundantly enters up to $L + H$ training pairs (Li et al., 2026). By tracking fixed forecast origins across epochs, we uncover a central training paradox: *recurring high-loss regions provide constructive immediate gradients, yet their multi-step accumulation through optimizer momentum exerts severe trajectory drag as training proceeds.*

Continuous bands and alignment collapse. Prediction errors naturally cluster into contiguous temporal bands (Figure 1b) that persist across epochs ($p \approx 0.96$; Section B.1). These high-loss regions $\mathcal{H}$ capture learnable physical dynamics rather than label noise, with their training loss declining faster than the complement in the majority of current diagnostic settings (Section C.7). Early in training, $\mathcal{H}$'s gradient load aligns constructively with the remaining complement $\mathcal{C}$ ($\cos \approx +0.55$). However, this cooperative alignment collapses to $+0.15$ by the validation endpoint (Figure 1c): as shared lower frequencies are fitted, the component of $G_{\mathcal{H}}$ parallel to $G_{\mathcal{C}}$ contracts by nearly 40%, whereas its orthogonal component persists undiminished, causing regional updates to increasingly conflict with the collective trajectory.

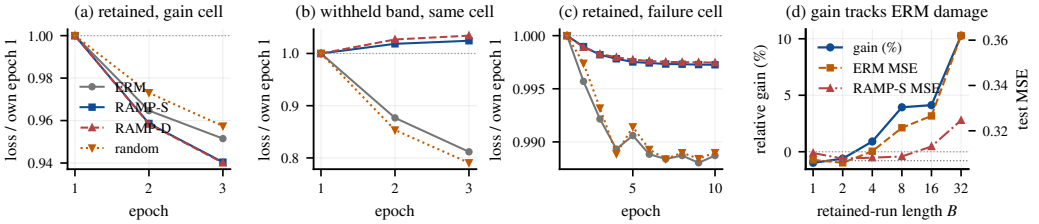


Figure 2: Empirical evidence of regional trajectory drag ($H = 96$, seed 1). (a) On ETTh2, standard ERM optimization is heavily dragged: withholding $\mathcal{H}$ accelerates the fitting of regular windows. **(b)** ERM expends massive capacity memorizing $\mathcal{H}$, whereas withholding halts this. **(c)** On smooth periodic series (Weather), physical withholding fails as shared dynamics are disrupted. **(d)** Gains scale with contiguous run length B, confirming drag originates from window overlap.

Trajectory drag and the continuation paradox. As visualized in Figure 2a–b, this alignment collapse causes severe trajectory drag: standard ERM expends massive capacity memorizing $\mathcal{H}$, actively slowing the convergence of regular windows (Figure 2a). Yet, controlled continuation probes on MixLinear/ETTh1 demonstrate that $\mathcal{H}$ is not inherently toxic (Section E.1):

- *Immediate utility*: In single-step evaluations, retaining $\mathcal{H}$ improves validation MSE across all 27 evaluations, confirming it supplies constructive immediate supervision.
- *Multi-step turnover*: Over 32-step continuations, however, continuing to train on $\mathcal{H}$ degrades validation error, decisively reversing from beneficial to detrimental by step 32 (Section E.5).
- *Momentum as the driver*: Because Adam exponentially smooths past gradients, outdated directional signals from $\mathcal{H}$ persist into subsequent steps after parameters move, dragging the trajectory.

When momentum smoothing is removed ($\beta_1 = 0$), this degradation is substantially attenuated (Table 21).

Mechanistic origin: compound window and momentum amplification. Why does this trajectory drag emerge uniquely in sliding-window forecasting? The root cause lies in the compound interaction between dataset construction and optimizer memory:

- *Sliding-window redundancy*: Because rolling windows advance by unit stride, an episode at timestamp $t^\star$ redundantly enters up to $L + H$ consecutive training pairs. Unlike i.i.d. training where difficult samples appear as isolated, zero-mean gradient perturbations, contiguous temporal regions repeatedly inject co-directional gradient pulses across adjacent mini-batches.
- **Compound momentum persistence:** In momentum optimizers like Adam, the first moment smooths gradients over an effective horizon of $\tau_{\text{eff}} \approx \frac{1}{1-\beta_1}$ steps. Although gradient weights normalize to unit sum, the high recurrence of localized segments compounds the temporal persistence of regional directions along the optimization trajectory.

Early in training, this compound amplification accelerates convergence toward shared low-frequency dynamics. However, once shared patterns are fitted and the region's gradient demands diverge orthogonally ($G_{\mathcal{H}} \perp G_{\mathcal{C}}$), the accumulated first moment continues driving parameters along obsolete historical directions, creating the structural trajectory drag observed in Figure 2d.

The withholding dilemma. The simplest response to trajectory drag is *physical data withholding*—excising $\mathcal{H}$ from the training set. While this halts momentum drag, it is a blunt instrument that discards valuable immediate supervision alongside stale history. On datasets where high-loss episodes capture critical transitions (or smooth periodic series like Weather, Figure 2c), physical omission forfeits indispensable temporal dynamics. This dilemma motivates resolving trajectory drag directly in optimizer memory rather than discarding data (Section 4).

4 Proposed Method: From Participation to Memory Control

4.1 From Rolling Windows to Regional Gradient Sources

For a multivariate series $\mathbf{x} \in \mathbb{R}^{T\times C}$ with lookback L and horizon H, stride-one rolling windows yield $N = T - L - H + 1$ training pairs indexed by forecast origins $o_i = L + i$. Standard ERM minimizes

$$R_{\text{ERM}}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell_i(\theta), \tag{1}$$

with gradients $\mathbf{g}_i(\theta) = \nabla_\theta \ell_i(\theta)$. Due to rolling overlap, any localized observation at $t^\star$ redundantly enters up to $L + H$ consecutive windows $\mathcal{B}(t^\star) = \{i : t^\star - H < o_i \le t^\star + L\}$. Partitioning all origins into a targeted region $\mathcal{H}$ and complement $\mathcal{C} = [N] \setminus \mathcal{H}$, the full ERM gradient decomposes additively into regional loads:

$$\nabla_\theta R_{\text{ERM}}(\theta) = G_{\mathcal{C}}(\theta) + G_{\mathcal{H}}(\theta), \qquad \text{where } G_{\mathcal{H}}(\theta) = \frac{1}{N}\sum_{i\in\mathcal{H}} \mathbf{g}_i(\theta). \tag{2}$$

Intervening on regional dynamics means regulating $G_{\mathcal{H}}(\theta)$'s influence on parameter trajectories.

4.2 Participation Baseline: Unrenormalized Masking (RAMP)

The simplest response to trajectory drag is to excise $G_{\mathcal{H}}$ from the objective. In epoch 1, RAMP identifies $\mathcal{H}$ by thresholding window losses $\mathcal{S}^{(1)} = \{\ell_i^{(1)}\}_{i=1}^N$ via MAD:

$$\tau = \text{Med}(\mathcal{S}^{(1)}) + k\ \text{MAD}(\mathcal{S}^{(1)}), \qquad \text{MAD}(\mathcal{S}) = \text{Med}(|\mathcal{S} - \text{Med}(\mathcal{S})|), \tag{3}$$

For the MSE benchmark, the per-window score is

$$\ell_i^{(r)} = \frac{1}{HC}\|\hat{\mathbf{y}}_i^{(r)} - \mathbf{y}_i\|_2^2, \tag{4}$$

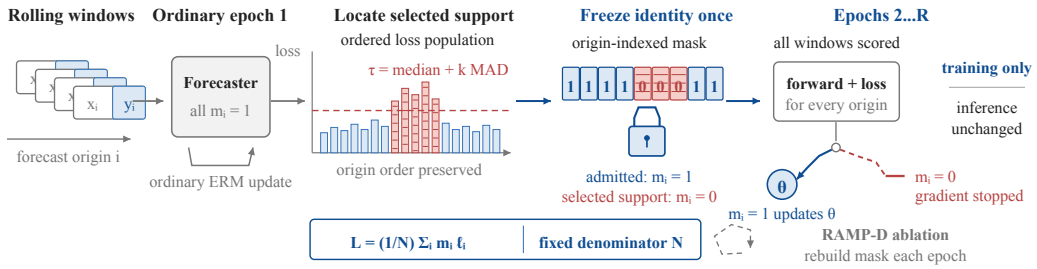


Figure 3: Overview of regional identification and participation control (RAMP). Epoch 1 records online window losses and locates support $\mathcal{H}$ via MAD. Subsequent epochs withhold gradients from $\mathcal{H}$ on the unnormalized $1/N$ scale ($m_i = 0$). $\mathcal{H}$'s temporal coordinates are shared with $\mathcal{H}$-short, which resets momentum staleness instead of discarding data.

which defines a scale-invariant selection criterion via Proposition 1. Subsequent epochs apply fixed binary indicators:

$$m_i = \mathbb{1}[\ell_i^{(1)} \leq \tau], \tag{5}$$

defining the masked empirical risk on the original unnormalized $1/N$ scale:

$$R_{\text{RAMP}}(\theta) = \frac{1}{N}\sum_{i=1}^{N} m_i \ell_i(\theta) \implies \nabla_\theta R_{\text{ERM}}(\theta) - \nabla_\theta R_{\text{RAMP}}(\theta) \equiv G_{\mathcal{H}}(\theta). \tag{6}$$

By preserving the $1/N$ denominator, RAMP subtracts the exact gradient load $G_{\mathcal{H}}$ without reweighting the complement $\mathcal{C}$ (Figure 3). We denote this fixed schedule as RAMP-S, and also compare an epoch-recomputed variant (RAMP-D) in our experimental ablations.

The withholding dilemma. While RAMP confirms that participation benefits are anchored to physical temporal support (Figure 4), physically zeroing out $\mathcal{H}$ treats an optimizer accumulation problem as a data quality problem. It terminates stale momentum drag only by discarding informative immediate supervision, degrading generalization on series with vital physical transitions (e.g., Solar).

4.3 Optimizer Memory Control: $\mathcal{H}$-short

To eliminate trajectory drag while preserving 100% data participation, $\mathcal{H}$-short intervenes directly on Adam's first-moment accumulation (Kingma & Ba, 2015). For mini-batch B_j after a standard ERM warmup epoch, let $g_{A,j} = |B_j|^{-1} \sum_{i \in B_j \cap A} \nabla_\theta \ell_i(\theta_j)$ denote regional batch gradients for $A \in \{\mathcal{H}, \mathcal{C}\}$. Unrolling Adam's first moment reveals that regional histories accumulate linearly:

$$m_j = \beta_1^j m_0 + q_{\mathcal{H},j} + q_{\mathcal{C},j}, \qquad \text{where } q_{A,j} = \beta_1 q_{A,j-1} + (1-\beta_1) g_{A,j}. \tag{7}$$

While $\mathcal{H}$'s immediate gradient $g_{\mathcal{H},j}$ provides constructive supervision at θ_j, its accumulated history $q_{\mathcal{H},j}$ carries outdated directional bias. We quantify this lag as the *momentum staleness vector*:

$$d_j = q_{\mathcal{H},j} - a_j g_{\mathcal{H},j} = (1-\beta_1) \sum_{k=1}^{j} \beta_1^{j-k} (g_{\mathcal{H},k} - g_{\mathcal{H},j}), \tag{8}$$

where $a_j = 1 - \beta_1^j$. Vector d_j captures the drift between accumulated history and current gradients, reflecting both parameter evolution and mini-batch sampling variance. Rather than withholding observations, **$\mathcal{H}$-short** retains all data but subtracts d_j from Adam's first-moment numerator:

$$\widetilde{m}_j(\lambda) = m_j - \lambda d_j = \beta_1^j m_0 + \left[(1-\lambda) q_{\mathcal{H},j} + \lambda a_j g_{\mathcal{H},j}\right] + q_{\mathcal{C},j}, \tag{9}$$

with $\lambda \in$ controlling attenuation strength ($\lambda = 1$ for full $\mathcal{H}$-short). Equation (9) decouples optimization dynamics: complement $\mathcal{C}$ retains standard momentum smoothing $q_{\mathcal{C},j}$, while $\mathcal{H}$'s decaying history is replaced by its immediate gradient $a_j g_{\mathcal{H},j}$. The second-moment preconditioner (computed over full data) and warmup m_0 remain untouched, and the update is scaled to match standard Adam's L_2 step norm (Section E.3). In this manner, $\mathcal{H}$-short absorbs immediate learning signals from difficult regions without inheriting their historical trajectory drag.

4.4 Theoretical Properties of Regional Memory Control

$\mathcal{H}$-short satisfies three core properties (proofs in Appendices A.2 and E.4): scale invariance, bounded staleness, and conditional improvement over Adam.

Scale invariance of temporal support identification. A primary requirement for hyperparameter stability is that identifying $\mathcal{H}$ should not depend on dataset units or arbitrary loss scaling.

Proposition 1 (Positive Scale Invariance)**.** *Let $\mathcal{S}^{(1)} = \{\ell_i^{(1)}\}_{i=1}^N$ denote the first-epoch window losses. For any uniform positive scaling $\ell_i' = a\,\ell_i^{(1)}$ ($a > 0$), the adaptive threshold rescales as $\tau'^{(1)} = a\,\tau^{(1)}$, and the resulting temporal support $\mathcal{H}' = \{i \in [N] : \ell_i' > \tau'^{(1)}\}$ is identical to $\mathcal{H}$.*

Because the median and median absolute deviation are positively homogeneous ($\mathrm{Med}(a\mathcal{S}) = a\,\mathrm{Med}(\mathcal{S})$ and $\mathrm{MAD}(a\mathcal{S}) = a\,\mathrm{MAD}(\mathcal{S})$), Proposition 1 ensures that region selection is non-parametric with respect to scale. This explains why a single global cutoff ($k = 3$) applies across diverse benchmarks without per-dataset threshold tuning.

Upper bound on accumulated momentum staleness. We next characterize the magnitude of the staleness vector $d_j = q_{\mathcal{H},j} - a_j g_{\mathcal{H},j}$ eliminated by $\mathcal{H}$-short.

Proposition 2 (Bound on Regional Staleness)**.** *Let $\beta_1 \in [0, 1)$ be the momentum decay factor. If consecutive regional gradients along the visited parameter trajectory satisfy $\|g_{\mathcal{H},k} - g_{\mathcal{H},k-1}\| \le \delta$ for all $k \le j$, then the accumulated staleness vector d_j in* (8) *satisfies:*

$$\|d_j\| \le \frac{\beta_1}{1-\beta_1}\,\delta. \tag{10}$$

Proposition 2 demonstrates that momentum staleness is scaled by the effective memory horizon $\frac{\beta_1}{1-\beta_1}$. Under Adam ($\beta_1 = 0.9$), local directional variations δ are magnified up to 9-fold in the first moment. As regional gradients differentiate orthogonally from the complement (Section 3), this lag exerts persistent trajectory drag. Setting $\lambda = 1$ in $\mathcal{H}$-short eliminates this accumulated error.

Conditional local improvement relative to Adam. Finally, we examine the local optimization effect of substituting the standard Adam proposal u_j with the $\mathcal{H}$-short proposal $\widetilde{u}_j$, evaluated under the shared step norm $\rho_j = \|u_j\|$.

Proposition 3 (Sufficient Local Improvement Condition)**.** *Let $e_j = u_j/\|u_j\|$ and $\widetilde{e}_j = \widetilde{u}_j/\|\widetilde{u}_j\|$ denote the unit directions of standard Adam and $\mathcal{H}$-short at state θ_j. Assuming the objective F has M-Lipschitz continuous gradients on the local update ball, a sufficient condition for $\mathcal{H}$-short to yield a strictly larger local reduction than standard Adam, $F(\theta_j + \rho_j\widetilde{e}_j) < F(\theta_j + \rho_j e_j)$, is:*

$$\nabla F(\theta_j)^\top(\widetilde{e}_j - e_j) < -M\rho_j. \tag{11}$$

Proposition 3 connects the elimination of staleness d_j directly to objective reduction: whenever outdated momentum pulls native Adam away from the true local descent vector $\nabla F(\theta_j)$, the directional correction $\widetilde{e}_j - e_j$ aligns negatively with the gradient. When this directional correction aligns sufficiently with the negative gradient to surpass the curvature penalty $M\rho_j$, $\mathcal{H}$-short achieves a strictly larger local reduction than standard Adam.

5 Experiments

Having established regional gradient dynamics in Section 3, we investigate three sequential questions: (i) *How does regional memory control perform across architectures and benchmarks?* (Section 5.2); (ii) *Why does $\mathcal{H}$-short transcend physical data withholding?* (Section 5.3); and (iii) *How does it scale to long contexts with minimal computational overhead?* (Section 5.4).

5.1 Experimental Setup

Datasets and forecasting benchmarks. We evaluate eight standard multivariate benchmarks: ETTh1, ETTh2, ETTm1, ETTm2 (Zhou et al., 2021), ECL (Godahewa et al., 2020a), Weather (Wu

et al., 2021), Solar (Godahewa et al., 2020b), and Traffic (Godahewa et al., 2020c), using standard chronological splits ($12/4/4$ months for ETT; $70/10/20\%$ for others). Input lookback is fixed to $L = 96$, and forecasting horizons span $H \in \{96, 192, 336, 720\}$. Details are in Appendix Table 19.

Table 1: Comprehensive tri-paradigm benchmark across eight forecasting datasets and five forecasters (4 horizons averaged, 3 seeds). We benchmark standard ERM/Adam, physical participation control (RAMP-S), and regional memory control ($\mathcal{H}$-short). **Bold** marks the lowest MSE across all three paradigms.

Dataset	Paradigm	MixLinear	PaiFilter	TimePro	iTransformer	PatchTST
ETTh1	ERM / Adam	0.4455	0.4904	0.4456	0.4591	0.4541
	RAMP-S	0.4463	**0.4740**	**0.4428**	**0.4560**	0.4511
	$\mathcal{H}$-short	**0.4453**	0.4873	0.4446	0.4570	**0.4446**
ETTh2	ERM / Adam	**0.3851**	0.3869	0.3852	0.3898	0.3856
	RAMP-S	0.3867	**0.3635**	0.3763	**0.3846**	0.3750
	$\mathcal{H}$-short	0.3866	0.3701	**0.3726**	0.3893	**0.3714**
ETTm1	ERM / Adam	**0.4226**	0.3850	0.3914	0.4176	0.3871
	RAMP-S	0.4227	**0.3820**	0.3898	**0.4059**	0.3857
	$\mathcal{H}$-short	0.4252	0.3829	**0.3892**	0.4120	**0.3852**
ETTm2	ERM / Adam	0.2941	0.2765	0.2813	0.2914	0.2888
	RAMP-S	0.2949	0.2751	0.2774	**0.2810**	**0.2788**
	$\mathcal{H}$-short	**0.2940**	**0.2731**	**0.2759**	0.2883	0.2812
ECL	ERM / Adam	0.2248	**0.1924**	0.1739	0.1769	0.2055
	RAMP-S	**0.2242**	0.1926	**0.1727**	**0.1737**	**0.2053**
	$\mathcal{H}$-short	0.2248	0.1930	0.1738	0.1754	0.2062
Weather	ERM / Adam	**0.2784**	**0.2477**	0.2522	0.2610	**0.2552**
	RAMP-S	0.2798	0.2480	0.2505	**0.2592**	0.2553
	$\mathcal{H}$-short	0.2792	0.2477	**0.2483**	0.2605	0.2554
Solar	ERM / Adam	0.4796	**0.2635**	0.2377	**0.2443**	**0.2533**
	RAMP-S	0.4802	0.2761	0.2426	0.2525	0.2650
	$\mathcal{H}$-short	**0.4795**	0.2659	**0.2349**	0.2460	0.2579
Traffic	ERM / Adam	0.6365	0.5109	**0.4406**	**0.4322**	**0.5006**
	RAMP-S	0.6371	**0.5105**	0.4459	0.4374	0.5013
	$\mathcal{H}$-short	**0.6363**	0.5107	**0.4403**	0.4333	0.5010

Forecaster architectures. Our testbed covers five forecasters with distinct inductive biases: linear (*MixLinear* (Ma et al., 2026)), frequency-domain (*PaiFilter* (Yi et al., 2024)), attention-based (*iTransformer* (Liu et al., 2024), *PatchTST* (Nie et al., 2023)), and representation-based (*TimePro* (Ma et al., 2025)), using their published configuration presets (Section D.1).

Compared methods and control arms. We benchmark: (i) **ERM**: standard training on all windows with Adam; (ii) **RAMP-S**: physical participation control withholding $\mathcal{H}$ on the unnormalized $1/N$ scale under a fixed first-epoch schedule; (iii) **$\mathcal{H}$-short (Ours)**: regional memory control retaining 100% data participation while replacing $\mathcal{H}$'s post-warmup first-moment history at strengths $\lambda \in \{0, 0.25, 0.5, 0.75, 1\}$; and (iv) **Optimizer controls**: *Random-short* (truncating momentum on an equal-count random support to isolate target specificity; additional controls such as *C-short* and *All-short* are evaluated in Section E.3).

Training and evaluation protocol. Models are trained with Adam (Kingma & Ba, 2015) ($(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$) and early-stopped on validation MSE. Test MSE and MAE are evaluated at validation-best checkpoints across five paired seeds for the 160-cell grid. Significance is assessed via two-sided paired Wilcoxon signed-rank tests.

5.2 Main Benchmark and Architectural Generality

We benchmark standard ERM (Adam), physical participation control (RAMP-S), and regional memory control ($\mathcal{H}$-short) across eight forecasting benchmarks and five forecasters (Table 1). H-short outperforms standard Adam in 99/160 cells (Table 26). It is particularly effective for expressive deep architectures (achieving win rates of 84.4% on TimePro and 68.8% on iTransformer), while showing neutral or mixed effects on compact linear models like MixLinear (37.5% win rate) where representational drift is minimal.

Table 2: Architectural generality of $\mathcal{H}$-short versus Adam across the 160-cell grid (5 forecasters $\times$ 8 datasets $\times$ 4 horizons). Gain is relative test-MSE reduction (%). Per-cell details appear in Appendix Tables 27–31.

Model Architecture	Inductive Bias	$\mathcal{H}$-short Wins	Mean Gain (%)
TimePro	Representation	27 / 32 (84.4%)	+1.12
iTransformer	Cross-variate Attention	22 / 32 (68.8%)	+0.41
PaiFilter	Frequency Filter	20 / 32 (62.5%)	+0.68
PatchTST	Patch Attention	18 / 32 (56.3%)	+0.78
MixLinear	Direct Linear	12 / 32 (37.5%)	−0.15
All Benchmark Configurations	Combined	99 / 160 (61.9%)	**+0.57**

Capacity-linked response closes the theoretical loop. Breaking down benchmark gains by architecture, we reveals that responsiveness to regional memory control scales with model capacity:

- *Expressive deep architectures*: High-capacity models with deep nonlinear representations achieve decisive win rates: TimePro leads with 84.4% (27/32, +1.12%), followed by iTransformer (68.8%), PaiFilter (62.5%), and PatchTST (56.3%). In these networks, recurring high-loss regions continuously induce representational drift and orthogonal gradient drag ($G_{\mathcal{H}} \perp G_{\mathcal{C}}$), making historical momentum reset particularly potent.
- *Compact linear models*: On MixLinear, $\mathcal{H}$-short yields a near-neutral win rate of 37.5% (12/32, mean gain −0.15%). This lack of response is theoretically expected rather than an empirical defect: shallow linear mappings lack multi-layer feature representations and experience negligible gradient orthogonality during rolling-window optimization (Section 3). Because linear models accumulate minimal multi-step trajectory drag, intervening on optimizer memory produces marginal effects.

5.3 Why $\mathcal{H}$-short Transcends Data Withholding: Empirical Insights

Comparing $\mathcal{H}$-short directly against physical data withholding (RAMP-S) clarifies the operational advantages of memory control over sample deletion across three key diagnostic perspectives:

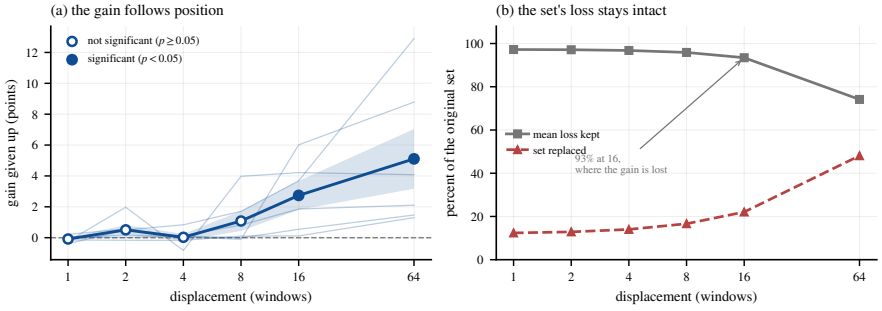


Figure 4: Sensitivity to temporal displacement (6 cells, 5 seeds). **(a)** Gain lost when shifting support by 1–64 origins. **(b)** Shifting 16 origins retains 93% loss but drops 2.74 points gain, confirming benefits follow temporal position rather than loss scale.

Physical location anchors both participation and memory. A vital question is whether $\mathcal{H}$-short merely acts as a generic momentum regularizer or targets genuine temporal coordinates. We test this through two rigorous matched controls:

- *Temporal displacement sweep*: In Figure 4, shifting the selected support by sixteen origins preserves 93% of its first-epoch loss magnitude, yet immediately gives up 2.74 percentage points of relative MSE reduction ($p = 0.033$). This proves that the underlying optimization gain follows temporal coordinates rather than sample loss magnitude.
- *Randomized memory truncation*: Applying first-moment truncation to an arbitrary, equal-count random temporal support (*Random-short*; Table 22) yields an insignificant average gain of only 0.63%. In contrast, anchoring truncation to the true temporal support $\mathcal{H}$ boosts the gain to 3.48%, winning in 17 of 18 paired runs by an average margin of 2.85 percentage points. Thus, memory intervention is effective only when targeted at the physically grounded sources of trajectory drag.

Complete retention preserves vital physical transitions. While physical withholding isolates localized noise, it causes structural loss whenever high-loss episodes contain indispensable dynamics:

- **Solar ramp boundaries:** High-loss episodes in Solar mark morning ramp-up and evening ramp-down boundaries calibrating the diurnal cycle. Excising them under RAMP-S strips away physical scales and degrades accuracy across architectures (Table **??**). While retaining full data under H-short still faces challenges on certain baselines (such as PaiFilter and PatchTST), it substantially narrows these withholding losses and yields gains on TimePro, offering a non-destructive optimization pathway that avoids physically discarding transition dynamics.
- *Traffic macro-synchronization*: In Traffic, dense spatial coupling keeps regional gradients aligned with the collective network (Section G.3). Because directional conflict is naturally absent, physical withholding merely discards valid sensor observations, whereas $\mathcal{H}$-short maintains complete coverage without destabilizing training.

Table 3: Continuous regional memory control across five replacement strengths λ (18 paired runs). Values are mean paired validation-MSE reductions relative to standard Adam.

λ	pairs	mean gain vs Adam	median
0	18	-0.01%	-0.00%
0.25	18	$+0.62\%$	$+0.62\%$
0.5	18	$+1.61\%$	$+1.43\%$
0.75	18	$+2.55\%$	$+2.24\%$
1	18	$+3.21\%$	$+3.10\%$

Smooth, continuous memory attenuation. Evaluating the memory attenuation strength $\lambda \in \{0, 0.25, 0.5, 0.75, 1\}$ (Table 3) demonstrates a graded effect: mean validation-MSE reductions are -0.01%, $+0.62\%$, $+1.61\%$, $+2.55\%$, and $+3.21\%$, respectively. The result supports continuous regional history control while the membership rule and training protocol remain fixed.

5.4 Scalability, Computational Overhead, and Practical Boundaries

Long-context scaling ($L = 512$). Extending context to $L = 512$ broadens rolling-window overlap, compounding momentum staleness. Across long-context benchmarks on iTransformer and PatchTST (Table 41), $\mathcal{H}$-short consistently improves generalization, confirming that regional memory regulation scales effectively and becomes increasingly critical as overlap intensifies.

Computational and memory overhead. Unlike gradient projection or sample re-scoring schemes that require auxiliary networks, H-short introduces minimal overhead. Regional gradients $g_{\mathcal{H},j}$ and $g_{\mathcal{C},j}$ share a single forward pass. In the optimizer, tracking $q_{\mathcal{H},j}$ requires only one additional parameter-sized momentum buffer (with zero extra trainable parameters), and the element-wise correction $-\lambda d_j$ is applied in-place within the standard Adam step with negligible runtime cost.

Practitioner's guide. These findings provide practical deployment guidance: in volatile series with localized shocks (e.g., ETT), H-short delivers substantial gains by curbing momentum drag; in series with critical physical transitions (e.g., Solar), it provides a non-destructive alternative that mitigates the severe information loss of data pruning; and in highly synchronized networks (e.g., Traffic), it closely matches standard Adam.

6 Conclusion

By tracking fixed forecast origins in sliding-window time-series forecasting, we uncover a divergence in training dynamics: contiguous high-loss regions provide constructive immediate gradients, yet degrade generalization across multi-step training as gradient alignment with the complement collapses. While participation control via RAMP demonstrates that this effect is strictly anchored to physical temporal location, physically withholding data inevitably discards informative current observations. To resolve this dilemma, we introduce $\mathcal{H}$-short to directly intervene on the optimizer's first-moment accumulation. By resetting regional momentum staleness, $\mathcal{H}$-short eliminates parameter trajectory drag while preserving complete data participation, establishing source-aware memory control as an effective paradigm for non-stationary temporal optimization.

AI USE STATEMENT

We used generative AI tools for manuscript language editing, experimental scripting, and analysis support. We did not use generative AI tools to generate data, select experimental results, or make final scientific decisions. All AI-assisted text, code, and analyses were reviewed and verified by the authors, and the authors take responsibility for the final content of this work, including all claims and artifacts.

REFERENCES

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In Andrea Pohoreckyj Danyluk, Léon Bottou, and Michael L. Littman (eds.), *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, pp. 41–48. ACM, 2009. doi: 10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

California Department of Transportation. Performance measurement system (PeMS) data source, n.d. URL https://dot.ca.gov/programs/traffic-operations/mpr/pems-source. Traffic detector data service; accessed August 25, 2026.

Youngin Cho, Daejin Kim, Dongmin Kim, Mohammad Azam Khan, and Jaegul Choo. Wavebound: Dynamic error bounds for stable time series forecasting. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022. doi: 10.52202/068431-1423. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/7b99e3c648898b9e4923dea0aeb4afa1-Abstract-Conference.html.

Yisong Fu, Zezhi Shao, Chengqing Yu, Yujie Li, Zhulin An, Qi Wang, Yongjun Xu, and Fei Wang. Selective learning for deep time series forecasting. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, 2025. doi: 10.52202/085713-3277. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/8cf54ff53f44835b9bdab2c546a1ca6d-Abstract-Conference.html.

Rakshitha Godahewa, Christoph Bergmeir, Geoff Webb, Rob Hyndman, and Pablo Montero-Manso. Electricity hourly dataset. Zenodo, 2020a. URL https://doi.org/10.5281/zenodo.4656140.

Rakshitha Godahewa, Christoph Bergmeir, Geoff Webb, Rob Hyndman, and Pablo Montero-Manso. Solar dataset (10 minutes observations). Zenodo, 2020b. URL https://doi.org/10.5281/zenodo.4656144.

Rakshitha Godahewa, Christoph Bergmeir, Geoff Webb, Rob Hyndman, and Pablo Montero-Manso. Traffic hourly dataset. Zenodo, 2020c. URL https://doi.org/10.5281/zenodo.4656132.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

M. Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In John D. Lafferty, Christopher K. I. Williams, John Shawe-Taylor, Richard S. Zemel, and Aron Culotta (eds.), *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada*, pp. 1189–1197. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/hash/e57c6b956a6521b28495f2886ca0977a-Abstract.html.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 95–104, 2018. doi: 10.1145/3209978.3210006. URL https://doi.org/10.1145/3209978.3210006.

Mingjie Li, Rui Liu, Guangsi Shi, Mingfei Han, Changlin Li, Lina Yao, Xiaojun Chang, and Ling Chen. Mitigating data redundancy to revitalize transformer-based long-term time series forecasting system. *ACM Transactions on Intelligent Systems and Technology*, 17(3):58:1–58:21, 2026. doi: 10.1145/3735651. URL https://doi.org/10.1145/3735651.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, 2021.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=JePfAI8fah.

Aitian Ma, Dongsheng Luo, and Mo Sha. Mixlinear: Extreme low resource multivariate time series forecasting with 0.1k parameters. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=QUj0KuCumD.

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

Max Planck Institute for Biogeochemistry. Weather station beutenberg campus, n.d. URL https://www.bgc-jena.mpg.de/en/servicegroups/fieldexperiements/locations/beutenberg. Meteorological data service; accessed August 25, 2026.

National Renewable Energy Laboratory. Solar power data for integration studies, n.d. URL https://www.nrel.gov/grid/solar-power-data.html. Dataset; accessed August 25, 2026.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021. arXiv:2107.07075.

Yanyao Shen and Sujay Sanghavi. Learning with bad training data via iterative trimmed loss minimization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5739–5748. PMLR, 2019. URL https://proceedings.mlr.press/v97/shen19e.html.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJlxm30cKm.

Artur Trindade. ElectricityLoadDiagrams20112014. UCI Machine Learning Repository, 2015. URL https://doi.org/10.24432/C58C86. Dataset.

Jiachen (Tianhao) Wang, Dawn Song, James Y. Zou, Prateek Mittal, and Ruoxi Jia. Capturing the temporal dependence of training data influence. In *International Conference on Learning Representations*, volume 2025, pp. 11236–11273, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/1e0d38c676d5855bcfab7f6d29d20ad9-Paper-Conference.pdf.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/bcc0d400288793e8bdcd7c19a8ac0c2b-Abstract.html.

Kun Yi, Jingru Fei, Qi Zhang, Hui He, Shufeng Hao, Defu Lian, and Wei Fan. Filternet: Harnessing frequency filters for time series forecasting. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. doi: 10.52202/079017-1749. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6323d96f79d5d49e0d3fc88835c082cd-Abstract-Conference.html.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, 2020.

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. doi: 10.1609/aaai.v37i9.26317. URL https://ojs.aaai.org/index.php/AAAI/article/view/26317.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 11106–11115. AAAI Press, 2021. doi: 10.1609/AAAI.V35I12.17325. URL https://doi.org/10.1609/aaai.v35i12.17325.

A FORMAL PROPERTIES OF REGIONAL GRADIENT PARTICIPATION

The appendix follows the distinction between identifying a source and deciding how it should influence training. Section A connects fixed forecast origins to the exact gradient term removed by RAMP. Section B establishes recurrence and cross-backbone agreement. Section C measures the same source across checkpoints and reports participation effects and the complete result grid; Section D specifies the protocol. Section E separates immediate usefulness from continued participation, tests all-data memory control, and isolates temporal location through matched interventions. Section F compares membership schedules, scores, and rate rules. Section G examines longer input context, channel-level resolution, and forecaster response heterogeneity.

A.1 FROM A FIXED TEMPORAL SUPPORT TO A CHANGED OBJECTIVE GRADIENT

Recurrence and consensus controls establish that independently recovered supports occupy similar origins (Section B). Holding membership fixed then attributes checkpoint differences to gradient evolution rather than changing selection (Section C.5). To intervene on that same source, the objective must preserve the complement's coefficients. The following identity shows that RAMP removes exactly the measured regional load at shared parameters.

Proposition 4 (Fixed-support gradient difference)**.** *Fix a support $\mathcal{H} \subseteq [N]$ from the first training epoch. For differentiable losses, the ERM and RAMP objectives at the same θ satisfy*

$$\nabla R_{\mathrm{ERM}}(\theta) - \nabla R_{\mathrm{RAMP}}(\theta) = \frac{1}{N} \sum_{i \in \mathcal{H}} \nabla \ell_i(\theta) = G_{\mathcal{H}}(\theta).$$

For full-batch gradient descent from the same start and step size η, the next-step difference is $\theta^{+}_{\mathrm{RAMP}} - \theta^{+}_{\mathrm{ERM}} = \eta G_{\mathcal{H}}(\theta)$.

Proof. Subtracting Equation (6) from Equation (1) cancels coefficients on the retained complement and leaves $1/N$ on $\mathcal{H}$. Differentiation and substitution into $\theta^{+} = \theta - \eta \nabla R(\theta)$ give the gradient and step identities. □

The gradient identity holds at each shared parameter state. In our Adam experiments, moment estimates also enter the parameter step, and the two arms subsequently visit different states. The matched interventions in Sections C.6 and E.9 measure those training consequences. The identity supplies the exact link between the measured regional load and the changed objective; the intervention results establish its effect on forecasting.

A.2 SCALE-INVARIANT REGION IDENTIFICATION

Section 4 identifies a candidate temporal region through a loss threshold standardized over the epoch population. Normalization choices and dataset units can rescale the loss by any positive factor, so membership should not depend on it. The proposition below proves that invariance.

Proposition 5 (Positive scale invariance)**.** *For any positive rescaling $\ell'_i = a\,\ell_i$ with $a > 0$ applied uniformly to all per-window losses, the RAMP mask* (5) *is exactly invariant: $m'_i = m_i$ for every i. Equivalently, the robust deviation* (3) *satisfies $d'_i = d_i$ in the limit $\epsilon \to 0$.*

Proof. Let $G = \{\ell_1, \ldots, \ell_N\}$ and $\ell'_i = a\ell_i$ with $a > 0$. The median is positively homogeneous: $\mathrm{Med}(G') = a\,\mathrm{Med}(G)$. For the median absolute deviation,

$$\mathrm{MAD}(G') \;=\; \mathrm{Med}\big(|G' - \mathrm{Med}(G')|\big) \;=\; \mathrm{Med}\big(a\,|G - \mathrm{Med}(G)|\big) \;=\; a\,\mathrm{MAD}(G),$$

since $a > 0$ factors out of the absolute value and the median. The selection threshold rescales identically:

$$\tau' \;=\; \mathrm{Med}(G') + k\,\mathrm{MAD}(G') \;=\; a\,\tau,$$

so (5) satisfies $\ell'_i \le \tau' \iff a\ell_i \le a\tau \iff \ell_i \le \tau$ for every i, and $m'_i = m_i$ exactly. For the analysis score in (3), $d'_i \to d_i$ as $\epsilon \to 0$; the training membership uses the raw threshold τ and remains exactly invariant. □

A.3 HISTORY-CONDITIONED INTERPRETATION UNDER MSE

Under the grid's MSE objective, fixing that membership acts on the region's contribution to the objective gradient, whose form is residual–Jacobian at each selected origin.

Let $\ell_i^{\mathrm{MSE}}(\theta) = (HC)^{-1}\|f_\theta(\mathbf{x}_i) - \mathbf{y}_i\|_2^2$ and define the half-squared residual $e_i = \frac{1}{2}\ell_i^{\mathrm{MSE}}$. Writing the prediction Jacobian as $\mathbf{J}_i = \nabla_\theta f_\theta(\mathbf{x}_i)$ and the residual as $\mathbf{r}_i = f_\theta(\mathbf{x}_i) - \mathbf{y}_i$, we have

$$\nabla_\theta \ell_i^{\mathrm{MSE}} \;=\; \frac{2}{HC}\,\mathbf{J}_i^\top \mathbf{r}_i \;=\; 2\,\nabla_\theta e_i,$$

and the masked gradient contribution is

$$m_i\,\nabla_\theta \ell_i^{\mathrm{MSE}}(\theta) \;=\; \frac{2\,m_i}{HC}\,\mathbf{J}_i^\top \mathbf{r}_i.$$

The gradient magnitude is $\|2\,\mathbf{J}_i^\top \mathbf{r}_i/(HC)\|_2$: a large residual yields a large parameter gradient only insofar as its projection through $\mathbf{J}_i^\top$ is large. Under RAMP's fixed schedule, the first-epoch mask gates every subsequent residual–Jacobian product at the same forecast origin. Thus epoch-one prediction errors determine later gradient participation, whereas later residuals and Jacobians determine the admitted gradient values. This derivation specifies the fixed gate under MSE; (5) applies to any differentiable per-window loss. The cross-architecture agreement in Section 3 addresses a separate question: whether different forecasters identify support at the same temporal origins.

B REGIONAL EVIDENCE AND IDENTIFICATION

B.1 EVIDENCE CHAIN: ADDITIONAL DETAIL

To complement the loss-field evidence in Section 3, we use an analytic head-gradient diagnostic to test two additional properties: whether score mass is concentrated and whether the selected membership recurs across epochs. Training remains standard ERM in this diagnostic; we retrospectively threshold each epoch's score field to form an epoch-indexed candidate mask $m^{(r)}$. These measurement masks are distinct from the fixed training mask used by RAMP. The diagnostic covers four iTransformer configurations at one seed and uses a different score, so it characterizes support geometry rather than the full-grid effect size. The analytic and loss scores agree in ranking at Spearman 0.89 but induce different thresholds at the same k.

Concentration of the analytic score. The top-q score-mass share uses the concentration statistic in Equation (16), with the analytic head-gradient score replacing the loss. It measures the fraction of total score mass concentrated in the top-$q\%$ of windows. On iTransformer/ETTh1 ($H = 96$, seed 1, scores measured at the end of each epoch) we observe $C_1 = 0.028$, $C_5 = 0.119$, $C_{10} = 0.202$: the top 5% of windows carry 11.9% of the score mass and the top 10% carry 20.2%. The analytic score mass is therefore concentrated in this run under standard ERM.

Temporal concentration. The unequal mass is not distributed independently over origins. Shared inputs and targets give adjacent windows a common temporal footprint; the measured score field determines whether high losses actually cluster within it. The epoch-wise candidate mask has a lag-1 autocorrelation of 0.94 (versus -0.001 for randomly permuted masks with matched retention), and masked windows form contiguous runs of average length 18 (maximum 143) along the window index. Upper-tail loss is therefore temporally clustered in this run. Cross-epoch measurements determine whether the same locations remain difficult as training proceeds.

Cross-epoch recurrence. Temporal concentration supplies a region, while cross-epoch recurrence determines whether that region remains relevant to later updates. The candidate supports of consecutive epochs overlap with Jaccard indices 0.86, 0.92, and 0.95 (increasing across epochs 1–4), and the one-step transition probability $P(m_i^{(r+1)} = 0 \mid m_i^{(r)} = 0)$ is 0.94–0.97: a window masked in one epoch remains masked in the next with high probability. The candidate set is therefore a temporally recurrent support in this run rather than transient score noise. This recurrence makes first-epoch membership informative for later epochs. The cross-backbone tests in Section B.2 determine whether that support is anchored to the same temporal coordinates across forecasters.

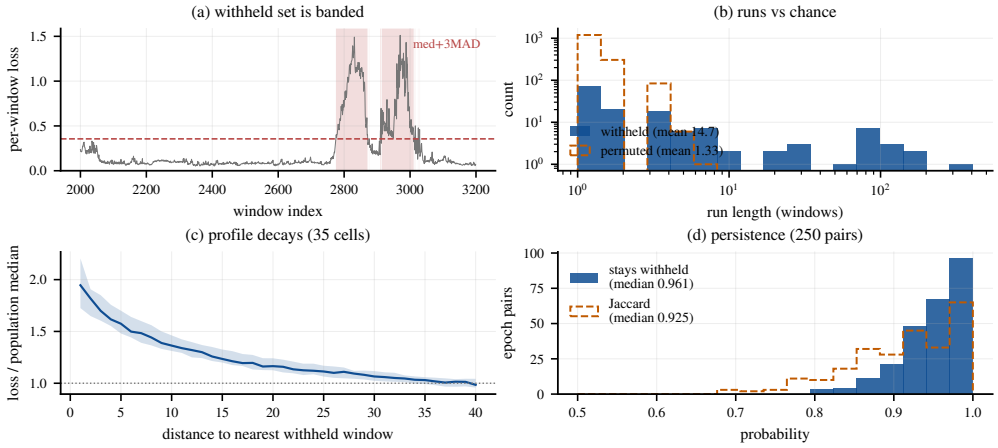


Figure 5: Additional temporal loss-field summaries. Selected windows form long contiguous runs against a matched permuted mask, loss decays with distance from the support, and membership persists across epochs.

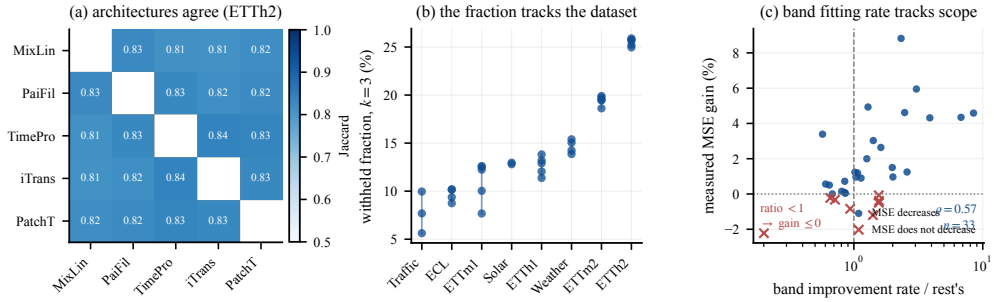


Figure 6: Additional summaries of regional support and scope. Independent forecasters agree on selected ETTh2 origins; the score-derived withholding rate varies across datasets; and, over the available $H = 96$ cells, the fitted-loss ratio is associated with the measured RAMP response.

The evidence chain across diagnostic configurations. Across these four diagnostic configurations, unequal mass, temporal concentration, and recurrence move together. This analysis uses one seed to expose the mechanism in detail; the declared grid supplies the separate multi-seed effect-size estimate. Table 4 reports the three evidence statistics across the same four iTransformer configurations. Within this diagnostic quartet, the ETTh2 $H = 192$ run has the highest measured values on every axis: $C_5 = 0.371$, mask autocorrelation 0.96, mean run length 32 (maximum 408), Jaccard 0.98, and $P(0 \to 0) = 0.998$. ETTh1 and ETTm1 show intermediate values, while the remaining configuration (ETTh1, $H = 336$) shows the weakest concentration ($C_5 = 0.091$). These diagnostics establish a candidate region through concentration and recurrence. The matched interventions below test whether its membership, rather than its size alone, changes optimization.

Single-run recurrence does not by itself establish that the support is anchored in the data. We therefore ask whether a consensus of several backbones predicts the support of one held out.

B.2 Cross-backbone consensus predicts a held-out support

Pairwise agreement can arise when two models share one bias, so the stronger test predicts an unseen backbone from the others. For each (dataset, horizon), we normalize every first-epoch loss field by its own median and median absolute deviation, hold one backbone out, and take the coordinatewise median of the remaining fields; the consensus support is the set of origins whose consensus score exceeds 3. We compare the held-out field and consensus by correlation, Jaccard overlap of the withheld sets, and score recovery: the mean consensus score over the held-out selected origins divided by the mean held-out score there, clipped to $[0, 1]$.

The reference is a circular-shift null. Rotating one field along the time index preserves marginal distribution, autocorrelation, and run structure but destroys cross-backbone alignment. Over 124

Table 4: Evidence-chain statistics across four iTransformer configurations (seed 1, $k = 3$). C_q: top-$q\%$ score-mass share; AC(1): mask lag-1 autocorrelation; run: mean (max) contiguous masked run length; J: consecutive-epoch candidate-support Jaccard; $P(0 \to 0)$: persistence of the masked state. This table reports score geometry for the four diagnostic configurations.

Config	C_5	AC(1)	run	J	$P(0 \to 0)$
ETTh2, H=192	0.371	0.963	32 (408)	0.982	0.998
ETTh1, H=96	0.119	0.939	18 (143)	0.950	0.969
ETTm1, H=192	0.143	0.941	19 (208)	0.891	0.932
ETTh1, H=336	0.091	0.902	11 (64)	0.815	0.911

Table 5: Masked controls at matched withholding count (iTransformer, H=96, k=3, 5-seed mean test MSE, validation-best checkpoints throughout). Relative change against ERM in parentheses; positive favors the mask. "random" withholds as many windows per epoch as the median–MAD rule but at uniformly random positions; "shuffled" permutes the RAMP-D mask over the window index, preserving the count. Both are recomputed each epoch. The separate fixed-set factorial in Table 34 compares rotation, which preserves run lengths, with permutation, which preserves count. Its cell and seed selection is specified in Section E.9.

Mask	ETTh1	ETTh2	ETTm1
ERM	0.38901	0.30682	0.34129
random (matched retention)	0.38911 (−0.03)	0.30593 (+0.28)	0.34105 (+0.07)
shuffled (matched count)	0.38874 (+0.07)	0.30520 (+0.51)	0.33965 (+0.47)
RAMP-D	0.38342 (+1.44)	0.29454 (+3.95)	0.33616 (+1.50)
RAMP-S	**0.38317** (**+1.50**)	**0.29346** (**+4.32**)	**0.33445** (**+2.00**)

held-out fits spanning 28 (dataset, horizon) combinations, correlation is 0.977 against 0.001, Jaccard is 0.821 against 0.036, and score recovery on the held-out support is 0.963 against 0.057. Every fit beats its null on all three quantities ($p = 4 \times 10^{-22}$). The held-out forecaster locates the same or nearby forecast origins as the other backbones, not merely a smooth field of the same shape.

Cross-backbone consensus establishes a reproducible temporal support. Matched-retention controls ask whether that choice changes the update beyond withholding as many windows.

Table 5 reports these matched-retention controls on three cells as five-seed means. Their separation from RAMP shows that the evidence chain depends on the retained set's membership rather than on withholding the same number of windows. The remaining ambiguity is whether that membership changes the update direction or merely its scale.

The 2×2 decomposition of the RAMP update. At a given epoch, let $\mathcal{C}$ be the retained set, $p = |\mathcal{C}|/N$ its retention rate, and $\bar{G}_{\mathcal{C}} = |\mathcal{C}|^{-1} \sum_{i \in \mathcal{C}} \nabla_\theta \ell_i$ its mean gradient. The masked objective passes $p\bar{G}_{\mathcal{C}}$ to the optimizer, separating a direction ($G_{\text{all}} \to \bar{G}_{\mathcal{C}}$) from a scalar gain ($p$); the Adam step itself also depends on the moment states. We attribute the gain by running all four combinations on iTransformer, H=96, five seeds, under one common score: ERM (both off); *gain-only*, which multiplies the unmasked batch loss by that batch's retention fraction $|\mathcal{C}_b|/N_b$ so that it sees RAMP's scale sequence with ERM's direction; *direction-only*, which masks but renormalizes to $1/|\mathcal{C}|$; and full RAMP. Table 6 reports the result using the RAMP-D membership schedule in all masked arms.

The four arms resolve that ambiguity in these cells. Gain-only is $+0.10\%$, $+0.17\%$ and $+0.19\%$ (none significant, $p \geq 0.14$), while direction-only recovers $+1.51\%$, $+3.99\%$ and $+1.53\%$ against RAMP's $+1.44\%$, $+3.95\%$ and $+1.50\%$. The seed-aligned paired difference between direction-only and RAMP is $+0.07\%$, -0.01% and $+0.02\%$ ($p = 0.32$, 0.89, 0.66). This matches Adam's normalized-moment update (Kingma & Ba, 2015): under a constant positive loss rescaling, the same factor enters $\eta\, \hat{m}/(\sqrt{\hat{v}} + \epsilon)$ in the numerator and denominator up to ϵ. We confirm this prediction directly: the epoch-global factor p applied to the loss reproduces ERM *bitwise* in four of five seeds on ETTh2. We separately scale the learning rate by p on top of direction-only: $+0.02\%$, -0.01% and $+0.12\%$ over direction-only ($p = 0.79$, 0.92, 0.23).

Widening the denominator comparison to 19 model–dataset–k combinations (five forecasters, eight datasets, $k \in \{1, 3, 5\}$) gives a median $1/|\mathcal{C}|$-versus-$1/N$ difference of $+0.02\%$, with a single significant advantage for renormalization anywhere ($+0.05\%$, PatchTST on ETTh1) against RAMP's own effects of $+0.4\%$ to $+5.5\%$ on the same cells. No denominator effect is resolved at the precision of these experiments. We report $1/N$ throughout, which preserves every retained origin's ERM

Table 6: 2×2 decomposition of the RAMP update (iTransformer, H=96, 5-seed mean test MSE, k=3, one common score across arms). "Direction" restricts the update to retained windows; "gain" scales it by the retention fraction. Relative improvement over ERM in parentheses; p from seed-aligned paired t-tests against ERM. The cell and protocol match Table 5; Section B.2 treats the denominator this decomposition varies.

Update	ETTh1	ETTh2	ETTm1
ERM	0.38901	0.30682	0.34129
Gain-only (per-batch p_b)	0.38862 (+0.10, p=0.22)	0.30627 (+0.17, p=0.37)	0.34064 (+0.19, p=0.14)
Direction-only ($1/\lvert\mathcal{C}\rvert$)	0.38315 (+1.51, p=0.003)	0.29456 (+3.99, p=0.045)	0.33608 (+1.53, p=0.018)
RAMP-D ($1/N$)	**0.38342** (+1.44, p=0.002)	**0.29454** (+3.95, p=0.046)	**0.33616** (+1.50, p=0.019)
Direction-only − RAMP-D	+0.07, p=0.32	−0.01, p=0.89	+0.02, p=0.66

Table 7: Boundary design ablation (iTransformer, ETTh1, H=96, 5-seed mean ± std). The fixed-q rows use rates different from the 12.8% derived by median–MAD on this cell. At equal retention, top-q and median–MAD select identical masks, so the comparison isolates adaptive extent selection. Positive gain means lower error than the comparison column; Table 38 extends the rate analysis to 17 cells.

Boundary	MSE (mean ± std)	gain vs. ERM	gain vs. RAMP
ERM	0.3890 ± 0.0019	*n/a*	-1.4%
mean $+2\sigma$	0.3890 ± 0.0010	0.0%	-1.4%
mean $+3\sigma$	0.3890 ± 0.0015	0.0%	-1.4%
fixed top-5%	0.3891 ± 0.0011	-0.0%	-1.5%
fixed top-10%	0.3900 ± 0.0017	-0.3%	-1.7%
median+kMAD (RAMP)	$\mathbf{0.3834 \pm 0.0012}$	$\mathbf{+1.4\%}$	*n/a*

coefficient. Together, the four-arm control locates the measured accuracy change in the composition of the retained set rather than in its aggregation weight.

B.3 Score boundary and regional extent

Reference statistic. We compare the median–MAD rule with four alternatives on iTransformer/ETTh1 (H=96, 5 seeds): mean-plus-standard-deviation boundaries (k=2, 3) and fixed top-q pruning (q=5%, 10%). A fixed top-q cutoff and a median–MAD cutoff select *bitwise identical* masks whenever they retain the same number of windows. This ablation therefore tests rate estimation rather than boundary geometry. The median–MAD rule derives a 12.8% rate from the score population and reaches 0.3834 MSE; the four alternatives use different rates and fall between 0.3890 and 0.3900, against ERM's 0.3890. The mean–std boundary derives a different rate because its scale is inflated by the upper tail. Table 7 reports the full ablation.

Threshold choice determines how many origins are selected. Their temporal arrangement requires a separate measurement of dependence along the origin axis.

B.4 Dependence induced by overlap

The support-control view meets a forecasting-specific structure: adjacent windows share large parts of their input and target, so their losses can remain correlated along the origin axis. At a frozen checkpoint, let $\rho(s)$ denote the empirical autocorrelation at lag s of the ordered per-window loss field. Let s_+ be the last lag in the initial positive sequence, so $\rho(j) > 0$ for $j = 1, \ldots, s_+$. We summarize this dependence by the truncated effective sample size

$$N_{\text{eff}} \;=\; \frac{N}{1 + 2\sum_{s=1}^{s_+} \left(1 - s/N\right) \rho(s)}. \tag{12}$$

We compute s_+ and N_{eff} from the loss field rather than substituting $L + H$: sharing timestamps need not make two losses identical. Over 32 cells at $H = 96$, the median loss field has $\rho(1) = 0.98$ and an initial positive sequence extending through lag 377. This is longer than the $L + H = 192$ span created by direct input–target-window overlap, so the dependence extends beyond direct overlap. Equation (12) gives a median N/N_{eff} of 166 (interquartile range 123 to 243).

This dependence explains regional extent, not gradient direction. A timestamp appears in up to H target windows and L input windows, so stride one expands a local segment into a run of up to $L + H$ training pairs. Section 3 verifies that upper-tail losses follow this geometry. Loss autocorrelation alone cannot establish whether the corresponding gradients align or

cancel. We therefore reserve the regional-load claim for the direct additive-gradient measurements in Section C.5. With discovery and gradient evidence separated, the next section locates the intervention within the hierarchy of training operators.

B.5 Training-stage hierarchy and the empirical measure

Three stages: within-window loss shaping, within-window coordinate selection, and across-window participation. The dependence analysis identifies complete windows as the unit whose repeated mass enters the empirical gradient. Equation (6) therefore acts at a different level from loss shaping or per-coordinate target selection. To make the distinction precise, at epoch r write the per-window residual as $\mathbf{r}_i^{(r)} = \hat{\mathbf{y}}_i^{(r)} - \mathbf{y}_i \in \mathbb{R}^{H\times C}$, the output gradient as $u_i^{(r)} = \nabla_{\hat{\mathbf{y}}_i^{(r)}} \ell_i$, and the prediction Jacobian as $J_i^{(r)} = \partial \hat{\mathbf{y}}_i^{(r)} / \partial \theta$. The corresponding parameter gradient is $\mathbf{g}_i^{(r)} = J_i^{(r)\top} u_i^{(r)}$. Three families of methods act on three different operators in the training pipeline:

- *Loss shaping* (MSE, Huber, Tukey, redescending M-estimators) modifies the map $\mathbf{r}_i^{(r)} \mapsto u_i^{(r)}$: it changes how a residual becomes an output gradient *within* a window, i.e., the local loss geometry.
- *Coordinate selection* (selective learning (Fu et al., 2025)) chooses a subset of target coordinates within each window, modifying the coordinate support of $u_i^{(r)}$. If $M_{i,h,c} \in \{0,1\}$ indicates whether horizon–channel coordinate (h,c) is retained and $|M_i| = \sum_{h,c} M_{i,h,c}$, its conditional risk $|M_i|^{-1} \sum_{h,c} M_{i,h,c} \ell_{i,h,c}$ renormalizes the surviving coordinatewise losses $\ell_{i,h,c}$.
- *Window participation* (RAMP) leaves the original within-window loss intact and modifies the *empirical measure* under which complete window gradients are aggregated, while deliberately *not* renormalizing away the removed mass.

Concretely, ordinary ERM aggregates window gradients under the uniform empirical measure $\mu_r = \frac{1}{N}\sum_i \delta_{(i,\mathbf{g}_i^{(r)})}$, where δ_z denotes a unit point mass at z; hence μ_r has total mass one. Let $\mathcal{C} = \{i : m_i = 1\}$ be RAMP's fixed retained set, $\mathcal{H} = [N] \setminus \mathcal{C}$ the selected support, and $p = |\mathcal{C}|/N$ the retained fraction. For every epoch $r \geq 2$, RAMP aggregates under the sub-probability measure

$$\widetilde{\mu}_r \;=\; \frac{1}{N}\sum_i m_i \, \delta_{(i,\mathbf{g}_i^{(r)})}. \tag{13}$$

Its total mass is $p \leq 1$. This expression shows both effects directly: the selected support's mass $1-p$ leaves the aggregation, while retained windows keep their original $1/N$ weight. RAMP therefore changes the *support* of the empirical measure over window gradients while leaving each retained term untouched. The denominator control above detects no difference from renormalizing the remaining mass at the precision of those comparisons. Write the retained- and selected-set mean gradients as $\bar{G}_{\mathcal{C}}^{(r)} = \frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}} \mathbf{g}_i^{(r)}$ and $\bar{G}_{\mathcal{H}}^{(r)} = \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \mathbf{g}_i^{(r)}$. For full-batch SGD with learning rate η, the RAMP update is therefore

$$\theta_{r+1} = \theta_r - \eta\, p\, \bar{G}_{\mathcal{C}}^{(r)}, \tag{14}$$

and (6) decomposes the corresponding ERM gradient as $G_{\mathrm{ERM}}^{(r)} = p\bar{G}_{\mathcal{C}}^{(r)} + (1-p)\bar{G}_{\mathcal{H}}^{(r)}$ and keeps only the first term.

The empirical-measure form leads directly to the two channels measured by the earlier 2×2 intervention: *support control* (the update direction uses only retained windows) and a *gain* channel (the scalar p multiplying that direction). On this panel, direction-only masking reproduces the main gain, while the scalar gain controls show small changes that are not detected as significant. The optimizer is mini-batch Adam, whose update is $\eta\,\hat{m}/(\sqrt{\hat{v}}+\epsilon)$ for bias-corrected first and second moments $\hat{m}$ and $\hat{v}$. Under a constant positive rescaling, the same factor enters the numerator and denominator up to ϵ; scaling the learning rate instead tests a direct step-size change. Section B.2 confirms both points: multiplying the ordinary loss by the global retained fraction p or multiplying each batch by its own retained fraction $p_b = |\mathcal{C}_b|/N_b$ remains statistically indistinguishable from ERM. Applying p to the learning rate on top of direction-only masking remains indistinguishable from direction-only masking, and replacing the $1/N$ denominator by $1/|\mathcal{C}|$ leaves it indistinguishable from RAMP across 19 model–dataset–k comparisons. We therefore retain $1/N$ and, on this panel, attribute the effect to the retained support rather than to the scalar gain. Accordingly (14) should be read as an exact account of the objective, not as a claim about the effective step size.

Table 8: MixLinear on the ETT family (5-seed mean test MSE), reported for the original forecaster and every RAMP-D threshold k. Best entry per row is bold. MixLinear occupies the low-ratio, low-response end of Table 40; Table 1 reports the complete benchmark grid.

Dataset	H	orig	k=1	k=2	k=3	k=4	k=5
ETTh1	96	0.3958	0.3944	**0.3937**	0.3939	0.3941	0.3946
ETTh1	192	0.4494	**0.4478**	**0.4478**	0.4496	0.4502	0.4506
ETTh1	336	0.4807	0.4818	0.4810	0.4807	0.4807	0.4807
ETTh1	720	**0.4615**	0.4624	0.4623	0.4616	0.4618	0.4615
ETTh2	96	**0.3060**	0.3083	0.3086	0.3096	0.3099	0.3094
ETTh2	192	**0.3884**	0.3932	0.3932	0.3920	0.3913	0.3910
ETTh2	336	**0.4236**	0.4287	0.4301	0.4297	0.4304	0.4315
ETTh2	720	**0.4220**	0.4229	0.4233	0.4234	0.4233	0.4238
ETTm1	96	0.3513	**0.3485**	0.3502	0.3507	0.3513	0.3503
ETTm1	192	0.4024	**0.3993**	0.4007	0.4014	0.4011	0.4015
ETTm1	336	0.4368	**0.4344**	*n/a*	0.4355	*n/a*	*n/a*

B.6 THRESHOLD-SPECIFIC EVIDENCE

With the participation mechanism fixed, MixLinear has the lowest median RAMP response in the full grid (Table 40). The ETT-family threshold sweep below tests whether that architecture-dependent response is specific to the global $k = 3$ choice.

The sweep leaves MixLinear's low-response ordering unchanged across thresholds, including the ETTh2 rows. With the cross-architecture sweep of Table 43, this stability identifies an architecture-dependent response rather than a special choice of $k = 3$. The next section tests the regional support directly, separating its existence from the value of withholding it.

B.7 PER-CELL EVIDENCE

The MixLinear sweep establishes this architecture-dependent response. We now return to the region itself and expand each link in the phenomenon claim from temporal geometry to cross-backbone agreement, model-conditioned location, and training response.

Contiguity. Withheld windows organize into runs far longer than chance. Their contiguous runs average 5.5 to 41.7 windows against a chance baseline of 1.05 to 1.37 under a permuted mask of matched retention, a median of 8.5 times longer and 4.8 to 38.2 times across cells. The mask indicator has lag-one autocorrelation between 0.79 and 0.97. On iTransformer/ETTh1 a training set of $8{,}449$ windows is covered by 39 runs; on iTransformer/ETTh2 at $H = 192$ a single run spans 408 windows, 5% of the training period, and is withheld at every epoch.

A smooth profile, not a level set of noise. The long runs are accompanied by a graded neighborhood rather than a binary jump. The main text reports the unsigned median profile across the diagnostic panel; the analysis below preserves the side of each boundary and averages over 34 pairs. Aligning every non-withheld window by its signed distance to the nearest withheld one, the per-window loss decays monotonically away from the selected support: 1.87 times the population median at distance one, then 1.80, 1.73, 1.63, 1.44, 1.32 and 1.24 at distances 2, 3, 5, 10, 15 and 20, reaching 1.03 only at distance 40. The profile is nearly symmetric (1.88 on the earlier side, 1.86 on the later side). A threshold therefore cuts the summit of a wide elevation whose flanks remain in training, which is the geometry that the ablations in Section 5 keep encountering.

Persistence. A smooth region could still drift too quickly to guide the next epoch. Instead, a withheld window is withheld again at the next epoch with probability 0.90 to 0.998, and consecutive-epoch masks overlap at Jaccard 0.68 to 0.98. The union of all masks approaches its final size quickly: on iTransformer/ETTh1 it grows $8.3 \to 9.1 \to 9.4 \to 9.5\%$ over four epochs, with increments of 0.8, 0.3 and 0.1 points.

Shared across architectures. Persistence within one model raises the stronger question of whether different forecasters locate the same region. On the same dataset, five forecasters spanning a linear

Table 9: Mask transfer across architectures (H=96, five seeds, k=3). A mask is computed from one epoch of the *donor* and frozen into the training of the *recipient*, which never scores its own windows. Gains are relative to the recipient's own ERM. The own-mask column uses the recipient's frozen mask, matching the donor arm's persistence. Ratios are reported when the recipient's own-mask gain exceeds 0.5%; the summary therefore uses 6 rows. On those rows the donor mask retains a median 87% of the own-mask gain (range 65–119%), and the two gains agree in sign in 9 of 9 rows. The recipient's own-mask column is the same quantity Table 1 reports per cell, and the donor sets agree with the recipient's at the Jaccard of Figure 6(a).

Recipient	Donor	ERM MSE	own frozen mask	donor mask	ratio
MixLinear/ETTh1	iTransformer	0.39577	+0.56%	+0.31%	—
MixLinear/ETTh2	iTransformer	0.30599	−0.39%	−0.33%	—
MixLinear/ETTm1	iTransformer	0.35128	+0.04%	+0.72%	—
PatchTST/ETTh1	MixLinear	0.38203	+1.19%	+1.42%	119%
PatchTST/ETTh2	MixLinear	0.30142	+5.95%	+5.09%	85%
PatchTST/ETTm1	MixLinear	0.32498	+0.72%	+0.64%	88%
iTransformer/ETTh1	MixLinear	0.38901	+1.50%	+1.49%	100%
iTransformer/ETTh2	MixLinear	0.30682	+4.32%	+3.32%	77%
iTransformer/ETTm1	MixLinear	0.34129	+2.00%	+1.29%	65%

Table 10: Series statistics against the withheld set on iTransformer/ETTh1. "Recovery" selects as many windows as RAMP withholds, by that statistic, and reports the fraction also selected by RAMP (12.1% at chance). The last column re-computes the rank correlation *within* the withheld set alone. This single-cell analysis tests scalar explanations for the support agreement in Section B.2.

Statistic	Spearman vs. loss	Jaccard	Recovery	Spearman inside the set
Mean channel std	+0.307	0.103	18.7%	−0.253
Spectral entropy	+0.202	0.063	11.8%	+0.219
Max $\lvert z\rvert$	+0.194	0.034	6.6%	−0.175
Spike ratio	+0.174	0.090	16.6%	−0.185
Roughness	+0.079	0.068	12.7%	−0.051
|Slope|	+0.022	0.072	13.5%	−0.083
|Level|	−0.018	0.072	13.4%	−0.258

model, a frequency-domain model, two attention models, and a representation-based model select overlapping sets: across the model pairs of the five reported architectures the mask Jaccard averages 0.721 (per dataset 0.481 to 0.790) against a chance level of 0.04 to 0.15, and the score rankings correlate at Spearman 0.891. The overlapping sets also transfer across models. Across 9 complete transfers, donor and recipient gains agree in sign in $9/9$ rows. To quantify retained benefit without dividing by a near-zero recipient effect, we report ratios when the recipient's own frozen-mask gain exceeds 0.5%. Over these 6 rows, the donor mask retains a median 87% of the recipient gain (range 65–119%).

The paired cells in Table 9 turn cross-architecture agreement into an intervention: where donor and recipient select overlapping regions, the donor's region retains most of the recipient's own frozen-mask effect. Shared location therefore transfers across backbones. The next comparison asks how much of that location can be recovered from scalar series statistics without a forecaster's loss field.

Shared support reflects a data–model interaction. The transfer result establishes coordinates shared across backbones, while the construction argument attributes their temporal dependence to the series and overlapping windows. We next test whether scalar series statistics alone recover the selected support. Those statistics track the per-window loss outside its upper tail, but their relation changes inside the support selected by RAMP. For a standardized input window, mean channel standard deviation averages the temporal standard deviation over channels; spectral entropy averages the entropy of each channel's normalized Fourier power; max $|z|$ and spike ratio are the largest absolute within-channel z-score and the fraction exceeding three; roughness is one minus the mean lag-one channel autocorrelation; and slope and level are, respectively, the mean per-channel linear trend and the window mean. We rank by the absolute slope and level. All comparisons below are computed from the released per-window table for iTransformer/ETTh1 (`data/spl_samples.csv`: 8,449 windows, 1,026 withheld, giving withheld fraction $q = 0.121$ and chance Jaccard $q/(2-q) = 0.065$); Table 10 collects the comparisons.

The comparison in Table 10 makes mean within-window channel variation the strongest scalar proxy. Its rank correlation with the loss is $+0.218$ over the lower half of the score population,

$+0.029$ over the upper half, and -0.253 inside the withheld set: among the windows RAMP removes, larger mean channel standard deviation goes with *lower* loss. Ranking by this statistic consequently recovers only 18.7% of RAMP's selected origins. Combined with the cross-architecture agreement of Figure 6a and the mask transfer of Table 9, the selected support reflects an interaction between data and model: its location is stable across architectures, while scalar series statistics alone do not determine membership.

Offset-kernel audit. Scalar statistics do not locate the selected support, so we test a richer offset-based proxy: ridge regression from squared one-step changes over the lookback–forecast span to online window loss. Across 24 fits, its median correlation with the loss field is 0.371, but half of its coefficient mass requires 91 offsets. Local variation therefore tracks part of the forecasting difficulty recorded by the loss field, over a wide range of offsets rather than one scale.

Many selected regions are fitted faster than their complement. Scalar series statistics recover only a small fraction of the selected support and the offset kernel correlates only partially with the loss field, so we return to the model-derived partition and measure its relative fitting rate. Under ERM, the selected loss falls *faster* than the complement's in the majority of the current diagnostic configurations: their improvement-rate ratio ranges from 0.20 to 8.48, reaching 3.88 on iTransformer/ETTh2 and 1.99 on iTransformer/ETTh1, but falling to 0.20 on MixLinear/Weather. A faster relative decline is a different quantity from the endpoint's absolute loss ranking; Section G.4 reports the ratio's forecaster-level ordering.

Regional extent varies chiefly by dataset in the $H = 96$ panel. The fitted-loss ratio summarizes forecaster-level response ordering, while regional extent records how much of each loss population belongs to the selected support. Across the 135 cells with recorded support size, the single $k = 3$ rule derives rates from 3.3% to 26.0%, with median 13.6%. At $H = 96$, dataset-level means span 7.8% to 25.5%, whereas the median cross-backbone spread within a dataset is only 1.5 percentage points. When the candidate mask is retrospectively recomputed for diagnosis, its fraction moves by only 0.1 to 3.7 points across epochs. This scale separation places the dominant variation between datasets in this panel, motivating deriving the rate from each score population rather than fixing it globally.

B.8 Implementation and empirical score conventions

The preceding evidence determines what the implementation must preserve: origin-indexed membership, scores standardized over the full loss population, and the original aggregation weight. Per-window scores are the per-window losses already produced by the training forward pass, recorded by sample index at no additional cost, with no extra forward or backward pass. Masked losses are zeroed per sample and then averaged over the full batch size, retaining the $1/N$ denominator. The one-time mask computation after epoch 1 is $O(N)$. All runs use the official hyperparameters of each paper with unified $L = 96$. Implementation and configuration records accompany the release.

Run tables, indexed score fields, and checkpointed gradients generate all manuscript numbers; the resulting score field distinguishes region membership from aggregate influence.

RAMP reallocates participation, not loss concentration. Recording every origin also exposes how the intervention redistributes optimization. The share of total score mass held by the top decile of windows increases under RAMP ($54.2 \to 56.5\%$ on iTransformer/ETTh2 and $57.1 \to 60.8\%$ on PaiFilter/ETTh2), while it decreases under ERM ($51.0 \to 48.3\%$ and $53.7 \to 49.1\%$). Withheld windows stop supplying direct gradients but remain affected by updates from retained windows. Their losses fall more slowly and occupy a larger share of the recorded loss mass. The intervention changes gradient participation rather than homogenizing the loss field.

Selection protocol for k. The recorded score population also determines the selected extent. Once that population and its spread are defined, k controls the extent selected from that population. The headline protocol fixes $k = 3$ globally, without per-cell selection. A complete RAMP-D sweep over $k \in \{1, \ldots, 5\}$ measures threshold sensitivity on every core (model, dataset, horizon) cell, and a separate validation-selected table quantifies the the corresponding tuning range. These analyses leave the fixed-k headline unchanged. The per-epoch measurements and complete threshold tables specify both protocols for reproducibility (Tables 8, 39 and 43).

C GRADIENT MECHANISM, RESULT SURFACE, AND TRAJECTORY

C.1 FROM SCORE MASS TO GRADIENT LOAD

The implementation turns the formal results above into two measurable properties of the score field: membership depends on relative scale, and potential influence depends on mass concentration. Positive scale invariance applies to a uniform rescaling of the scalar losses: only each window's position within the score population determines membership. It does not identify changes caused by learning rate, initialization, backbone, or a different loss, none of which generally rescales every window uniformly. Those choices are evaluated empirically in Section B.3.

The MSE interpretation in Section A.3 then relates fixed membership to the underlying gradient. Under the $1/(HC)$-normalized squared error, the score (4) reduces to the mean squared residual. Writing the residual as $\mathbf{r}_i = f_\theta(\mathbf{x}_i) - \mathbf{y}_i$ and the prediction Jacobian as $\mathbf{J}_i = \nabla_\theta f_\theta(\mathbf{x}_i)$, the effective per-window gradient of (6) satisfies

$$\nabla_\theta\big(m_i\,\ell_i^{\mathrm{MSE}}\big) \;=\; \frac{2}{HC}\,m_i\,\mathbf{J}_i^\top\mathbf{r}_i, \tag{15}$$

The fixed indicator m_i, formed from epoch-one loss, therefore determines whether the later residual–Jacobian product at origin i enters the update. This interpretation is specific to MSE; RAMP itself requires only the per-window score and fixed membership in (4)–(5).

Relative membership alone does not determine regional gradient load. We therefore measure the *loss-mass concentration* of the top-$q\%$ of windows,

$$C_q^{(r)} \;=\; \frac{\sum_{i\in\text{Top-}q\%}\ell_i^{(r)}}{\sum_i \ell_i^{(r)}}, \tag{16}$$

i.e., the fraction of total loss mass carried by a small subset of windows. This scalar measures magnitude concentration, not vector influence: large losses can still produce gradients that cancel. Directional influence is measured separately in Section F.3: the per-window loss is strongly correlated with head-gradient magnitude (Spearman 0.85) and unrelated to its direction (Spearman -0.07). Ranking by anti-consensus instead selects different origins and reduces accuracy in the tested comparisons. Loss therefore provides a magnitude-related selection signal, not a test of opposition. A persistent C_q identifies a candidate source. Its net magnitude and relation to the complement require the additive-gradient measurements in Section C.5.

Mass concentration identifies a candidate region but does not establish that its temporal organization matters. The next intervention asks whether its effect depends on overlap-induced arrangement rather than training-set size alone.

C.2 THE WINDOW-ARRANGEMENT INTERVENTION: CONFOUNDS AND CONTROLS

Concentration shows that a small set of origins carries substantial loss mass; the arrangement intervention asks whether its effect depends on overlap-induced temporal organization rather than training-set size alone. Table 11 gives five-seed means for every arm on iTransformer/ETTh2, together with an ETTh1 cell that provides a contrasting response. All arms below the first train on $1{,}057$ windows with epochs and patience multiplied by eight. The ERM column is that arm's own baseline, and gains are computed within an arm.

Run length, at a fixed one eighth of windows: RAMP-S gains -0.98, -0.61, $+0.91$, $+3.94$, $+4.12$ and $+10.28\%$ for runs of $1, 2, 4, 8, 16$ and 32, against arm ERMs of 0.30717, 0.30587, 0.31097, 0.32133, 0.32657 and 0.36199. The relative gain and the arm's ERM correlate at Spearman 0.94, and RAMP-S's absolute error over the same sequence is 0.310, 0.308, 0.308, 0.309, 0.313, 0.325, never better than full-data ERM. The monotone relative series therefore reflects how much sparsification costs ERM rather than a benefit that grows with clustering.

This intervention varies clustering rather than adjacent overlap. At a spacing of eight with $L = H = 96$, adjacent retained windows still share 88 of 96 input timestamps, no adjacent pair is disjoint, and the mean adjacent overlap is identical ($88.0/96$) in both geometries, since seven gaps of one and one gap of fifty-seven also average eight. What differs is the *clustering* of the retained forecast origins:

Table 11: Window-arrangement intervention, all arms, $H = 96$, five seeds. "phase" is the number of distinct start-hour residues the retained origins occupy. The block and jitter arms match the uniform arm on window count and on mean adjacent overlap ($88.0/96$); they differ in clustering and, for the uniform arm, in phase coverage. The final-eighth arm has the best ERM of all settings; its -0.28% response separates recency from mask geometry. Table 13 carries the two principal geometries across 6 cells and Table 12 the spacing sweep on one; this table is their union, arm by arm.

Cell	retained windows	phase	ERM	RAMP-D	RAMP-S
iTr./ETTh2	all 8,449	24	0.30682	$+3.95\%$	$+4.32\%$
	every 8th (uniform)	3	0.30717	-1.79%	-0.98%
	jittered (dispersed)	24	0.30273	-1.34%	-0.55%
	runs of 8 (clustered)	24	0.32133	$+2.68\%$	$+3.94\%$
	final eighth	24	0.29481	*n/a*	-0.28%
	every 32nd	3	0.32808	-0.82%	-0.92%
iTr./ETTh1	all 8,449	24	0.38901	$+1.44\%$	$+1.50\%$
	every 8th (uniform)	3	0.43286	$+0.44\%$	$+0.82\%$
	runs of 8 (clustered)	24	0.41640	$+0.09\%$	$+0.42\%$

Table 12: Window-arrangement intervention in detail on one cell, iTransformer/ETTh2 at $H = 96$, with arms this table alone carries (spacings of 1, 8, 32 and 96, and a contiguous tail). Table 13 is the across-cell version. Five seeds. Rows two through four train on the same 1,057 windows with the same number of gradient steps and differ only in *which* windows. All three have heavily overlapping neighbors ($88/96$ shared input timestamps on average in the first two); what differs is whether the retained origins are uniformly spaced, clustered into runs of eight, or confined to the final eighth of the series. Gains are relative to that row's own ERM, which is not constant across rows. The uniform and clustered rows also differ in start-hour coverage ($3/24$ against $24/24$); see the phase-balanced control in the appendix.

Training windows	span	origins	RAMP-D	RAMP-S
all 8,449	full	every window	$+3.95\%$	$+4.32\%$
1,057, every 8th	full	uniform	-1.79%	-0.98%
1,057, runs of 8	full	clustered	$\mathbf{+2.68\%}$	$\mathbf{+3.94\%}$
1,057, final eighth	$1/8$	contiguous	*n/a*	-0.28%
264, every 32nd	full	uniform	-0.82%	-0.92%

only the contiguous geometry contains runs of near-duplicate windows, which is the structure our account requires. The two geometries also differ in a way that has nothing to do with clustering: on hourly data a spacing of eight occupies only three of the twenty-four start-hour residues, whereas contiguous runs cover all of them. The jittered, phase-balanced arm in Table 11 removes this seasonal-phase confound while keeping the retained origins dispersed. The protocol also fixes each geometry across the five optimizer seeds, so seed replication does not vary the geometry. Moreover, per-epoch halving of the learning rate means the sparsified arms decay eight times more often than the stride-one arm, so only comparisons among sparsified arms are schedule-matched; and the ERM baseline is not constant across arms (0.30717, 0.32133 and 0.29481 for the three $1{,}057$-window geometries), so every gain is computed within its own arm. The third geometry keeps the final eighth of the series contiguously and yields -0.28% ($p = 0.006$). Its ERM is the best of all four settings, a difference consistent with a recency effect rather than with the mask geometry shared by the other sparse arms.

These constraints bound the arrangement result to comparisons among schedule-matched sparsified arms. Within that boundary, clustering restores relative gains when uniform spacing does not. The across-cell intervention below tests the same contrast on its predeclared panel.

Under clustered retention RAMP-S improves 6 of 6 cells; under uniform spacing 2 cells are negative. The difference-in-differences also tracks the cost each geometry imposes on its own ERM ($r = 0.81$, $p = .049$), so the result supports a relative arrangement effect rather than an absolute advantage over dense ERM. Having bounded the spatial claim, the next section asks whether RAMP-D's epoch-to-epoch revisions preserve the same regional membership.

C.3 Composition of the per-epoch mask revisions

The arrangement intervention isolates membership across space; RAMP-D varies membership across epochs. Its revisions are composed as follows: 70–95% of the windows entering the mask at any epoch are within two indices of a window already in it, and only 42–65% of them are still withheld

Table 13: Retained-window arrangement over 6 predeclared cells, five seeds each. Both geometries retain one eighth of the windows with matched training budgets; each RAMP-S result is compared with that geometry's own ERM. $\Delta\Delta$ is clustered gain minus uniform gain.

	every eighth		runs of eight		
Cell	ERM	RAMP-S	ERM	RAMP-S	$\Delta\Delta$
iTransformer/ETTh2/H=96	0.3072	-0.98	0.3213	$+3.94$	$+4.92$
iTransformer/ETTh1/H=96	0.4329	$+0.82$	0.4164	$+0.42$	-0.41
PatchTST/ETTh2/H=96	0.3149	$+6.04$	0.3147	$+6.02$	-0.02
PaiFilter/ETTh2/H=96	0.2967	-2.82	0.3186	$+5.19$	$+8.01$
TimePro/ETTh2/H=96	0.3020	$+1.85$	0.3223	$+5.67$	$+3.82$
iTransformer/ETTm2/H=96	0.1860	$+1.73$	0.1993	$+4.58$	$+2.85$

one epoch later. The revisions therefore remain near stable boundaries but are short-lived. RAMP-D repeatedly changes the boundary membership of an otherwise stable support. Persistent membership, rather than repeated region discovery, is the remaining difference from RAMP-S.

C.4 Mechanism measurement table

The revision statistics describe membership, not its optimization effect. Table 14 measures what each masking rule does to the retained and withheld groups and places that response beside the accuracy interventions used in the main text.

Table 14: Mechanism measurement, comparing changes on retained windows with the accuracy changes in Table 34. Improvement of the retained windows under each intervention, expressed as the difference in percentage points from the same windows' improvement under ERM (3 seeds; a positive value means the retained windows are fitted better than they would have been). "random" withholds the same number of windows per epoch at uniformly random positions. Right-hand column: improvement of the withheld set itself, where a negative value means its loss increased.

		retained windows (Δpp vs. ERM)			withheld band
Pair	MSE gain	random	RAMP-D	RAMP-S	ERM $\rightarrow$ RAMP-D
PaiFilter/ETTh2	$+3.04\%$	$+1.31$	$+3.84$	$\mathbf{+4.90}$	$+24.2 \rightarrow -5.0$
iTransformer/ETTh2	$+3.95\%$	-0.05	$\mathbf{+1.34}$	$+1.28$	$+19.5 \rightarrow -3.1$
iTransformer/Weather	$+0.54\%$	-0.74	$+1.25$	$\mathbf{+1.26}$	$+8.7 \rightarrow -2.2$
MixLinear/Weather	-2.22%	$\mathbf{+0.00}$	-0.98	-0.96	$+0.2 \rightarrow -0.5$

On the three positive cells, targeted withholding accelerates the retained group by 1.25–3.84 points while the withheld loss stops decreasing; on MixLinear/Weather, accuracy degrades and the advantage reverses. The optimization response thus tracks accuracy in both directions.

The benchmark in Table 1 covers the complete 5×8 cross of 40 model–dataset pairs, four horizons, and five seeds. ERM, RAMP-S, and RAMP-D share the fixed-$k = 3$ protocol across the grid. The threshold sweep evaluates RAMP-D at $k \in \{1, \ldots, 5\}$ on the 32 model–dataset pairs retained in the current benchmark.

The declared grid establishes breadth, while Section B.2 establishes that different backbones select the same or nearby origins. Decomposing the objective gradient of that shared support and of its complement is what reveals the size and direction of its contribution; Table 15 provides those quantities cell by cell.

C.5 The region inside the objective, cell by cell

Section 3 reports the change in alignment between the two ERM checkpoints; Table 15 gives the regional diagnostic medians at mask formation. Table 15 gives the regional measurements per cell, at the checkpoint where the mask is formed and at the validation-selected ERM checkpoint.

Table 15: Contribution ratio R, support–complement cosine c and displacement $D = 1 - \cos(G_{\mathcal{C}}, G_{\text{full}})$ per cell, at the first-epoch checkpoint and at the validation-selected ERM checkpoint, ordered by the gain the frozen rule achieves on that cell. † marks a cell whose best validation epoch was the first, so that its two checkpoints are the same weights and its two readings coincide by construction. Such cells are excluded from the paired mask-forming-against-endpoint tests.

Model	Dataset	gain	epoch 1			ERM endpoint		
		(%)	R	c	D	R	c	D
PatchTST	ETTh2	$+5.95$	1.78	$+0.655$	0.151	1.64	$+0.768$	0.093
iTransformer	ETTm2	$+4.62$	0.71	-0.219	0.227	0.89	-0.304	0.347
PaiFilter	ETTh2	$+4.35$	5.54	$+0.084$	0.744	1.22	-0.590	0.727
iTransformer†	ETTh2	$+4.32$	2.91	-0.019	0.692	2.91	-0.019	0.692
PaiFilter	ETTm2	$+3.03$	0.73	$+0.659$	0.062	2.12	$+0.066$	0.526
iTransformer	ETTm1	$+2.00$	0.22	$+0.554$	0.013	0.30	$+0.543$	0.022
iTransformer	ETTh1	$+1.50$	0.38	-0.015	0.066	0.34	-0.076	0.056
PaiFilter	ETTm1	$+1.23$	0.17	$+0.968$	0.001	0.55	$+0.871$	0.016
iTransformer	ECL	$+0.96$	0.15	$+0.874$	0.002	0.18	$+0.377$	0.012
iTransformer	Weather	$+0.51$	0.72	$+0.291$	0.131	0.75	$+0.238$	0.149
MixLinear	ECL	$+0.00$	0.27	$+0.812$	0.008	0.20	-0.472	0.018
PatchTST	Weather	-0.22	0.22	$+0.339$	0.019	0.70	-0.369	0.247
PatchTST	Traffic	-0.32	0.09	$+0.808$	0.001	0.09	$+0.530$	0.003

The two regional gradients are accumulated as $1/N$ sums so that they add to the full gradient coordinatewise, and the full gradient is accumulated a third time independently, over all windows, purely to check that they do. The residual $\|G_{\mathcal{H}} + G_{\mathcal{C}} - G_{\text{full}}\|/\|G_{\text{full}}\|$ stays at 3×10^{-5} or below on every cell reported, and TimePro/ETTm1 exceeds this tolerance by five orders of magnitude in its fused scan kernels and is reported separately rather than pooled. The accumulation and every inner product use double precision because the squared norms overflow in single precision on the widest dataset.

These additivity and precision checks establish the quantities in Table 15. The reported values are read at a fixed checkpoint with the model in evaluation mode, so they describe the gradient field at one point on the trajectory rather than an average along it. Both regional sums run over the complete training split, so a checkpoint reading carries no mini-batch sampling variance by construction. The following experiment measures the variability introduced by estimating the same quantities from sampled origins.

Sampling variance of the same reading. We calibrate this variability at the validation-selected ERM endpoints of PaiFilter on ETTh2 and ETTm2, both at $H = 96$ and seed 1. With the same fixed masks and $1/N$ accumulation, independent draws of $256, 1024, 4096$ training origins give standard deviations of 0.248–0.328, 0.238–0.292, and 0.093–0.154 in the displacement D over 20 draws each, against full-split values of 0.5256–0.7272. Even samples of 4096 origins leave substantial variability in D. Full-split readings hold both membership and the evaluated origins fixed, so their checkpoint differences measure changes in the gradient field as parameters move. This calibration concerns fixed-checkpoint geometry; the history difference in Section E.4 also contains changes in mini-batch composition along a training trajectory. The separate Traffic extension in Section G.3 changes the support resolution and therefore reports its shared-prediction-head decomposition separately.

The same reading under an adaptive preconditioner. The fixed-checkpoint reading is expressed first in raw gradient geometry, whereas Adam preconditions the direction in which parameters move. We therefore repeat the displacement under a diagonal preconditioner $\text{diag}(1/(\sqrt{\hat{v}} + \epsilon))$, with $\hat{v}$ estimated from the second moment of batch gradients at the same checkpoint. The estimate is formed at the same checkpoint from the batch second moments. The preconditioned displacement ranks the cells at $+0.758$ ($p = 0.003$) against $+0.714$ raw, and loses its signal under the same control for regional extent ($+0.297$, $p = 0.325$). Adam's rescaling compresses the displacement without changing its ordering across cells. The measured regional ordering therefore persists under this adaptive preconditioner.

Regional differentiation over optimization. Preconditioning preserves the ordering, so we next follow the raw, unpreconditioned components over training. The support–complement cosine com-

Table 16: Regional differentiation cell by cell at $H = 96$, seed 1. The first four numeric columns hold the first-epoch region membership fixed and read its support–complement cosine at the mask-forming checkpoint and at the validation-selected ERM endpoint, against an extent-matched random region. The last two columns compare the selected-support contribution ratio at the validation-selected ERM and RAMP checkpoints; under RAMP the withheld gradient is measured but not applied.

		selected cosine		random cosine		selected ratio R	
Model	Dataset	epoch 1	ERM	epoch 1	ERM	ERM	RAMP
MixLinear	ECL	$+0.812$	-0.472	$+0.998$	$+0.995$	$+0.200$	$+0.550$
PaiFilter	ETTh2	$+0.084$	-0.590	$+0.767$	$+0.868$	$+1.221$	$+22.980$
PaiFilter	ETTm1	$+0.968$	$+0.871$	$+1.000$	$+0.995$	$+0.553$	$+0.339$
PaiFilter	ETTm2	$+0.659$	$+0.066$	$+0.991$	$+0.962$	$+2.122$	$+3.188$
iTransformer	ETTh1	-0.015	-0.076	$+0.853$	$+0.827$	$+0.342$	$+1.128$
iTransformer	ETTh2	-0.019	-0.019	$+0.943$	$+0.903$	$+2.913$	$+16.394$
iTransformer	ETTm1	$+0.554$	$+0.543$	$+0.981$	$+0.974$	$+0.298$	$+0.366$
iTransformer	ETTm2	-0.219	-0.304	$+0.916$	$+0.927$	$+0.889$	$+2.297$
iTransformer	ECL	$+0.874$	$+0.377$	$+0.992$	$+0.943$	$+0.182$	$+0.362$
iTransformer	Weather	$+0.291$	$+0.238$	$+0.940$	$+0.977$	$+0.751$	$+3.399$
PatchTST	ETTh2	$+0.655$	$+0.768$	$+0.919$	$+0.956$	$+1.644$	$+7.491$
PatchTST	Weather	$+0.339$	-0.369	$+0.997$	$+0.911$	$+0.698$	$+1.131$
PatchTST	Traffic	$+0.808$	$+0.530$	$+0.991$	$+0.961$	$+0.091$	$+0.098$

pares each region with the retained direction at the same checkpoint, so it does not inherit the rotation of a fixed global reference. Fixing the first-epoch region membership, the cosine falls from $+0.604$ to $+0.152$ at the validation-selected ERM endpoint ($p = 0.007$, down in 11 of 12 cells). An extent-matched random region remains nearly aligned with its complement, moving from $+0.986$ to $+0.958$ ($p = 0.68$). The decline is larger for the selected support in 11 of 12 paired cells ($p = 0.0024$).

The cosine decline establishes differentiation; the component split identifies its starting asymmetry. At the mask-forming checkpoint, the ratio of the complement-orthogonal component $\|G_{\mathcal{H}}\|\sqrt{1-c^2}$ to the complement-parallel component $\|G_{\mathcal{H}}\||c|$ is 1.50 for the selected support and 0.20 for the matched random region. The selected-support ratio is larger in all 13 of 13 cells, by a median factor of 10.4 ($p = 0.0002$). From that checkpoint to the validation-selected ERM endpoint, the parallel component contracts to 0.62 of its initial magnitude ($p = 0.047$), while the orthogonal component remains at 1.04 ($p = 0.94$). The random split shows the same component-wise evolution, 0.66 and 1.05 ($p = 0.001$ and $p = 0.52$). The parallel component contracts more strongly; the selected support starts with a larger orthogonal-to-parallel ratio.

The two ERM checkpoints establish the component change before intervention. Reading the same quantities at the RAMP endpoint then characterizes the geometry reached after withholding. At the validation-selected checkpoint, the support–complement cosine is $+0.263$ under RAMP and $+0.066$ under ERM ($p = 0.84$). The regional-to-complement contribution ratio rises from 0.70 to 1.13 ($p = 0.0017$, higher in 12 of 13 cells). Under RAMP, $G_{\mathcal{H}}$ is evaluated for this diagnostic but excluded from the update; its larger relative contribution is therefore an effect of withholding, not the force that produces the accuracy gain.

These paired checkpoint measurements establish differentiation under ERM at fixed data-time support; cross-epoch recovery of similar membership appears in Section B.1.

The next test changes participation after the ERM endpoint to determine whether the same support remains consequential at a later training stage.

C.6 Activating the frozen region after a validation-selected ERM endpoint

Both arms restart from identical validation-selected ERM weights and fresh Adam states, take one matched unmasked epoch, and then differ only in whether the first-epoch frozen region participates in subsequent updates. Holding the pre-ERM membership fixed makes the contrast about participation at the new starting point. Because restarting changes the optimizer state, the late-start gain is measured against the matched continuation arm rather than the original grid ERM value. Both arms otherwise use the standard early-stopping and test protocol.

Table 17: Post-ERM activation by model–dataset pair at $H = 96$. Means and standard deviations use the same 3 seeds for the late-start and early-start estimands; positive values favor withholding the frozen first-epoch region.

Model	Dataset	late start		early start	
		mean	s.d.	mean	s.d.
MixLinear	ETTh1	$+0.95$	1.36	$+0.43$	0.89
MixLinear	ETTh2	-0.18	0.82	-0.39	0.64
MixLinear	ETTm1	$+0.06$	0.20	$+0.08$	0.12
MixLinear	ETTm2	-0.55	0.17	-0.82	0.07
MixLinear	ECL	$+0.04$	0.04	-0.02	0.09
MixLinear	Weather	-1.90	0.82	-2.06	1.06
MixLinear	Traffic	-0.04	0.04	-0.12	0.06
PaiFilter	ETTh1	$+2.53$	1.53	$+2.96$	0.82
PaiFilter	ETTh2	$+4.73$	1.62	$+3.88$	0.52
PaiFilter	ETTm1	$+0.70$	0.93	$+1.61$	1.59
PaiFilter	ETTm2	$+4.25$	1.43	$+2.87$	0.74
TimePro	ETTh1	$+1.03$	0.61	$+0.85$	0.17
TimePro	ETTh2	$+3.60$	0.66	$+4.06$	0.34
TimePro	ETTm1	$+1.03$	0.29	$+0.93$	0.26
TimePro	ETTm2	$+2.49$	0.54	$+0.99$	1.36
TimePro	ECL	$+7.67$	13.76	$+0.86$	0.56
TimePro	Weather	$+0.77$	3.41	$+1.33$	1.02
TimePro	Solar	$+4.26$	3.27	-1.47	0.55
iTransformer	ETTh1	$+2.23$	0.96	$+1.80$	0.48
iTransformer	ETTh2	$+3.73$	0.03	$+3.60$	1.55
iTransformer	ETTm1	$+1.46$	1.91	$+2.19$	1.07
iTransformer	ETTm2	$+7.67$	1.10	$+4.60$	0.84
iTransformer	ECL	$+1.29$	0.29	$+0.93$	0.03
iTransformer	Weather	$+1.37$	0.24	$+0.42$	1.04
iTransformer	Traffic	-0.41	0.18	-1.22	0.09
PatchTST	ETTh1	$+3.41$	2.17	$+1.45$	0.89
PatchTST	ETTh2	$+6.41$	2.69	$+5.28$	1.96
PatchTST	ETTm1	$+0.70$	0.47	$+0.75$	0.29
PatchTST	ETTm2	$+2.89$	0.74	$+0.83$	0.39
PatchTST	ECL	$+0.17$	0.09	$+0.14$	0.05
PatchTST	Weather	$+0.03$	0.50	-0.14	0.30
PatchTST	Traffic	-0.33	0.03	-0.30	0.02

Late activation improves 27 of 32 cells, with median gain $+1.29\%$ and $p = 9 \times 10^{-6}$; agreement with the early-start effect shows that the region remains actionable after ERM rather than acting only through warm-up.

C.7 Regional extent and model response

Regional differentiation identifies a fixed region; regional extent asks the complementary question of how much response to expect across cells. The relation between regional extent and gain, the per-cell RAMP-S test-MSE change tabulated in Table 1, is fitted over the 128 cells formed by the 32 registered model–dataset pairs at the four horizons, with additive effects for architecture, dataset and horizon, so that extent is asked to explain what remains after those are absorbed. With one slope shared by every architecture the coefficient is $+10.27$ percentage points of relative MSE reduction per unit of extent ($p = 0.087$) and the fit explains 0.467 of the variance. Allowing one slope per architecture raises that to 0.563, and the joint test that the slopes are equal is rejected at $F(5, 110) = 4.80$, $p = 0.001$. Table 18 lists the slopes.

The architecture-specific slopes establish that extent and response are not coupled uniformly. The same fit then separates variation across datasets from variation across forecasters. Ranking cells within one architecture and horizon, extent orders the gain across datasets ($+0.344$, $p = 7 \times 10^{-5}$ over 128 within-group comparisons). Ranking within one dataset and horizon, it does not order the gain across architectures ($+0.069$, $p = 0.45$ over 124). Extent orders which series a given forecaster

Table 18: Response of the measured gain to regional extent, one slope per architecture, from the fit with additive architecture, dataset and horizon effects over 128 cells. Units are points of mean squared error improvement per unit of extent, so a cell whose region covers ten per cent more of the origins carries one tenth of the tabulated coefficient.

Architecture	λ_m	p	cells
PatchTST	$+17.2$	0.013	28
iTransformer	$+15.3$	0.047	28
PaiFilter	$+6.7$	0.438	16
TimePro	$+5.6$	0.456	28
MixLinear	-5.1	0.447	28

stands to gain on. Across architectures within a fixed series the rank correlation is not detected, and Table 18 reports the per-architecture slopes that describe response heterogeneity on a given series.

D Experimental protocol and implementation

D.1 Protocol details

The extent analysis draws on a heterogeneous grid, so its interpretation depends on a common training and selection protocol. We evaluate eight standard long-term forecasting datasets: ETTh1, ETTh2, ETTm1, ETTm2, ECL, Weather, Solar, and Traffic. The four ETT variants contain six transformer-load variables and oil temperature at two sites and two resolutions (Zhou et al., 2021). ECL is the 321-client hourly benchmark derived from the 370-client, 15-minute UCI archive (Trindade, 2015); we use its archived hourly forecasting release (Godahewa et al., 2020a; Lai et al., 2018). Weather records 21 pressure, temperature, humidity, wind, rain, and radiation channels from the Max Planck Institute for Biogeochemistry station, following the 10-minute Autoformer benchmark (Max Planck Institute for Biogeochemistry, n.d.; Wu et al., 2021). Solar is the 137-site Alabama subset of NREL's synthetic 2006 PV integration data (National Renewable Energy Laboratory, n.d.; Godahewa et al., 2020b), while Traffic is the 862-sensor hourly subset of Caltrans PeMS road occupancy data (California Department of Transportation, n.d.; Godahewa et al., 2020c). Both processed releases follow the LSTNet benchmark lineage (Lai et al., 2018).

Table 19: Datasets used in the experiment grid. Timesteps and channels describe the exact files consumed by our loader; channels exclude the timestamp. The preceding text records both the originating archive and the processed forecasting release. All splits are chronological.

Dataset	Measured system	Timesteps	Channels	Cadence	Split
ETTh1, ETTh2	Transformer load and oil temperature, two sites	17,420 each	7	1 hour	12/4/4 mo.
ETTm1, ETTm2	The same two sites at finer resolution	69,680 each	7	15 min	12/4/4 mo.
ECL	Electricity consumption by benchmark-retained clients	26,304	321	1 hour	70/10/20%
Weather	Meteorological station measurements	52,696	21	10 min	70/10/20%
Solar	Simulated PV power at Alabama sites	52,560	137	10 min	70/10/20%
Traffic	San Francisco Bay Area freeway occupancy	17,544	862	1 hour	70/10/20%

Across those datasets, the grid covers five architectures with different inductive biases: a linear forecaster (MixLinear (Ma et al., 2026)), the frequency-domain PaiFilter variant of FilterNet (Yi et al., 2024), two attention-based forecasters (iTransformer (Liu et al., 2024), PatchTST (Nie et al., 2023)), and a representation-based forecaster (TimePro (Ma et al., 2025)). We preserve each architecture's published configuration family. For model–dataset pairs covered by a public pair-specific preset, we use that preset; otherwise, we transfer a published preset by a rule fixed before execution: the MixLinear, iTransformer, and PatchTST Solar extensions; the dataset-invariant ETT preset for PaiFilter; and the high-dimensional ECL preset for TimePro/Traffic. No validation or test result selects among presets. Every paired ERM–RAMP comparison then changes only gradient participation. All runs use a unified look-back $L = 96$ and $H \in \{96, 192, 336, 720\}$. Fixing L isolates the declared forecaster–dataset–horizon grid. A separate controlled $L = 512$ study tests whether the effect

Algorithm 1 RAMP training with fixed first-epoch membership

1: Train epoch 1 on all windows and record the online loss field $\mathcal{S}^{(1)}$.
2: $\tau \leftarrow \mathrm{Med}(\mathcal{S}^{(1)}) + k\,\mathrm{MAD}(\mathcal{S}^{(1)})$.
3: $m_i \leftarrow \mathbb{1}[\ell_i^{(1)} \le \tau]$ for every origin i.
4: **for** $r = 2, 3, \ldots$ **do**
5: Train epoch r on (6) using the same fixed m.
6: **end for**

survives longer input context (Section G.2). We restrict the grid to comparisons that share the MSE objective and complete five-seed protocol.

The main benchmark compares ERM, RAMP-S, and RAMP-D at a global $k = 3$ across all 40 model–dataset pairs, four horizons, and five seeds. A separate threshold sweep evaluates RAMP-D at $k \in \{1, 2, 3, 4, 5\}$ on 32 pairs under the same four-horizon, five-seed protocol. The fixed-k benchmark measures performance without per-cell threshold selection; the sweep measures sensitivity to k. Each horizon-level test MSE and MAE is averaged over five seeds; the pair-level pivot in Table 1 then averages the four horizons. The MAE companion uses the same checkpoints and masks. Per-seed values, hyperparameters, and training logs accompany the paper. The Traffic origin–channel study is a separate PatchTST/H=96 extension and does not enter the 160-cell denominator. Its four arms use five seeds, validation-selected checkpoints, the same first standard ERM epoch, and the original $1/(NHC)$ loss denominator.

In the benchmark, model selection is identical across arms. Every run, baseline and masked arm alike, trains with early stopping on validation loss at its model's own patience and is then evaluated once at the checkpoint with the lowest validation loss. Test MSE is recorded at that checkpoint. Epoch selection uses validation loss, the headline threshold is fixed at $k = 3$, and all five seeds are reported. The responsive panels used for subsequent mechanism studies have their own configuration-selection rules, stated in Sections E.3 and E.9; the descriptive rate oracle is defined in Section F.5. The same rule governs the intervention arms: in the displacement factorial, the contiguity split and the displacement sweep, each arm selects its checkpoint on its own validation curve rather than inheriting the baseline's stopping point. Where a comparison selects among candidates, as in the five-candidate threshold protocol of Section F.7, the criterion is validation loss and is stated at the point of use.

Table 19 gives the exact file-level dimensions and splits. Each channel is standardized using training-partition statistics only. Validation and test segments reuse the preceding $L = 96$ observations as input context, but their prediction targets remain inside their respective partitions. All reported models use the squared loss. Benchmark runs used eight NVIDIA A100 GPUs. The accompanying materials include run configurations, per-seed MSE and MAE, training logs, and the per-epoch window scores used in the regional analyses.

D.2 Training loop

The protocol fixes the evaluation units and checkpoint rule; Algorithm 1 specifies how RAMP's fixed schedule changes participation inside each run.

RAMP-D changes only the update of the mask: after every epoch it recomputes the threshold and the next mask from that epoch's recorded loss field. Thus RAMP-S and RAMP-D share the first standard ERM pass, its first-epoch score and threshold, denominator, and training budget, and differ only in whether the origin-indexed membership persists. The same fixed identity carries the all-data memory intervention, which keeps every window and changes only how the selected support's first moment accumulates.

E Matched interventions on regional support

E.1 One-step value and short-trajectory value

The comparisons below connect three questions: whether a region helps the next update, how that judgment changes over a continuation, and which optimizer history contributes to the difference. The

intermediate-step readout in Section E.5 complements the endpoint comparisons, while Section E.2 holds preconditioning and update norms fixed to distinguish first-moment treatments.

The same regional information can be evaluated over one update or a continuation from the same state. We use the original MixLinear implementation and ETTh1 preset at $L = H = 96$, with seeds 1, 2, and 3. Each trajectory runs for the preset maximum of 30 epochs without early stopping, saving complete model, Adam, and random states after epochs 1, 15, and 30. The standard split contains 8,449 training origins and 2,785 validation origins. First-epoch online losses select a fixed median-plus-3-MAD support containing 1,112, 1,097, and 936 origins, respectively. Each support is held fixed across the three checkpoints. No test split is used.

At each saved state, four arms retain all origins, withhold $\mathcal{H}$, withhold a circularly shifted support, or withhold an equal-count random support. They share the initial optimizer state, random state, and 32 predetermined batches of size 256. Masking retains the original full-batch denominator. The common diagnostic learning rate is 0.003, one tenth of the preset initial rate, and no scheduler advances during the probe. This rate is distinct from the native collection schedule, whose late-stage updates approach FP32 resolution. All reported stage comparisons below use the common rate.

Define the regional retention value after s steps as

$$V_s = 100 \frac{L_{\mathrm{val}}(\theta_s^{\mathrm{masked}}) - L_{\mathrm{val}}(\theta_s^{\mathrm{full}})}{L_{\mathrm{val}}(\theta_s^{\mathrm{full}})}. \tag{17}$$

Positive V_s favors retaining the region; negative V_s favors withholding it. Every readout covers the complete validation split. Each seed supplies one trajectory per stage, not 32 independent repetitions.

Table 20: The value of retaining a fixed region depends on the number of subsequent updates. Entries are three-seed mean V_s (%) on MixLinear/ETTh1. One-step values use the first shared batch; independent one-step probes on the first three batches favor retention in all 27 seed–stage–batch comparisons. Parentheses in the 32-step selected-support column count seeds favoring withholding.

Checkpoint	One step, $\mathcal{H}$	32 steps, $\mathcal{H}$	32 steps, shifted	32 steps, random
Epoch 1	$+0.01396$	$+0.12734$ (1/3)	$+0.04113$	$+0.09917$
Epoch 15	$+0.00238$	-0.10304 (3/3)	-0.04282	$+0.00484$
Epoch 30	$+0.00243$	-0.12003 (3/3)	-0.03686	$+0.02370$

The 32-step selected-support values for seeds 1, 2, and 3 are $(-0.09872, +0.11349, +0.36723)\%$ at epoch 1, $(-0.03455, -0.21370, -0.06086)\%$ at epoch 15, and $(-0.05129, -0.16685, -0.14194)\%$ at epoch 30. Thus middle and late checkpoints agree across seeds, while the early response is mixed. Keeping the second moment fixed gives mean middle/late values $-0.10175/-0.11835\%$; fixing the entire preconditioner gives $-0.09842/-0.11711\%$, with withholding favored in all three seeds in both cases. Continuing second-moment adaptation is therefore not necessary for this short-trajectory effect.

E.2 Inherited momentum and continuing accumulation

Using the same nine checkpoints and fixed regional support, we next hold the checkpoint-derived preconditioner common across arms and compare inherited first-moment history, zero initial first moment with continuing $\beta_1 = 0.9$, and no continuing first-moment smoothing ($\beta_1 = 0$). The second-moment state and optimizer age are retained; the learning rate stays 0.003. For each step, both the full and masked arms match an actual parameter-displacement norm from either a frozen full-data reference trajectory or a frozen masked reference trajectory. These two reference sequences are fixed without validation selection. They differ from the branch-local norm reference used by the all-data training intervention below.

Clearing inherited momentum increases the mean middle/late withholding advantage, whereas removing continuing smoothing reduces it under both norm references. A smaller full-versus-masked gap is not itself a better all-data optimizer: under the full reference, no smoothing raises full-data MSE by 0.768×10^{-3} and 0.346×10^{-3} at the middle and late stages. Zero initial momentum raises it by 0.902×10^{-3} and 0.206×10^{-3}. These results motivate separating a particular source's continuing accumulation from both its present gradient and the inherited optimizer state.

An additional order control retains the same batch multiset and within-batch ordering, comparing the original sequence, its reverse, and a circular shift by 16 batches. With inherited momentum and natural

Table 21: First-moment controls at matched actual step norms. Entries are mean masked-minus-full validation MSE after 32 updates, in units of 10^{-3}, over three seeds. Negative values favor withholding. The fixed preconditioner is common; full and masked references specify the common norm sequence.

First-moment treatment	Norm reference	Epoch 1	Epoch 15	Epoch 30
Inherited, continuing	Full	$+0.411$	-0.507	-0.499
Inherited, continuing	Masked	$+0.421$	-0.529	-0.776
Zero initial, continuing	Full	$+0.140$	-0.864	-0.927
Zero initial, continuing	Masked	$+0.231$	-0.876	-1.049
No first-moment smoothing	Full	-0.250	-0.074	-0.246
No first-moment smoothing	Masked	-0.157	-0.126	-0.292

norms, mean middle/late masked-minus-full differences are $(-0.746, -0.885)$, $(-0.856, -0.942)$, and $(-0.547, -0.625)$ in units of 10^{-3}. Original and reversed orders favor withholding in all three seeds at both stages; the shifted order does so in two. These finite order perturbations retain the average effect while changing individual trajectories. Across the norm, memory, and order comparisons there are 162 arms and 81 paired curves; the 32 updates within a curve are dependent observations.

E.3 Regional first-moment memory with complete data participation

The all-data experiment tests the design implied by the continuation controls. Adam supplies the unmodified reference; Random-short tests equal-count regional targeting; C-short and All-short change the scope of history replacement; H-masked changes participation instead. The secondary readouts and strength sweep then test whether the benefit extends across training and varies with the amount of history replaced.

Table 22: Regional-memory and participation controls across six configurations.

Model	Dataset/H	Adam	H-short	Random-short	C-short	All-short	RAMP-S$_{\text{warmup}}$	H-short gain (%)	RAMP-S gain (%)
PaiFilter	ETTh2/720	0.4569	0.4324	0.4511	0.4552	0.4401	**0.4055**	5.33	11.23
PaiFilter	ETTh2/336	0.4242	0.4049	0.4234	0.4264	0.4152	**0.3992**	4.54	5.89
iTransformer	ETTh2/96	0.3142	0.2999	0.3103	0.3126	0.2998	**0.2915**	4.50	7.21
PaiFilter	ETTh2/96	0.2927	0.2846	0.2926	0.2951	0.2880	**0.2825**	2.75	3.45
PaiFilter	ETTm2/96	0.1771	0.1723	0.1759	0.1760	0.1731	**0.1720**	2.69	2.87
PaiFilter	ETTh1/96	0.3819	0.3779	0.3802	0.3781	0.3794	**0.3724**	1.05	2.49

Selection and paired readout. The all-data memory intervention keeps every window in training and changes only how the selected support's first moment accumulates, isolating the history term from the act of withholding. It uses configurations where fixed-$k = 3$ RAMP-S improves test MSE in all five original seeds. The six configurations are PaiFilter on ETTh1, ETTh2, and ETTm2 at $H = 96$, PaiFilter/ETTh2 at $H = 336, 720$, and iTransformer/ETTh2 at $H = 96$. Configuration selection uses the benchmark's paired RAMP-S results, independently of memory outcomes. The memory comparisons use seeds 1, 2, and 3 from the same five-seed pool. Table 22 reports all six arms on test at each arm's own validation-selected epoch. Table 23 reports final-epoch validation MSE for Adam, H-short, C-short, All-short, and H-masked.

Table 23: Validation readout of the same regional-memory continuations: final-epoch complete-validation MSE (three-seed means; lowest in bold). Short-memory arms retain all windows. Gain is $\mathcal{H}$ short's mean paired reduction (%) relative to Adam, with improved-seed counts. The main text reports the test readout of these runs in Table 22.

Model	Dataset/H	Adam	$\mathcal{H}$ short	$\mathcal{C}$ short	All short	$\mathcal{H}$ masked	Gain (wins)
PaiFilter	ETTh2/96	0.222923	**0.216063**	0.222762	0.216902	0.218097	3.08 (3/3)
PaiFilter	ETTm2/96	0.128035	0.122774	0.127144	0.124037	**0.120105**	4.11 (3/3)
PaiFilter	ETTh1/96	0.693816	0.679661	0.686052	0.684533	**0.676584**	2.04 (3/3)
PaiFilter	ETTh2/336	0.378537	**0.370232**	0.378495	0.372857	0.377824	2.19 (3/3)
PaiFilter	ETTh2/720	0.645046	**0.628567**	0.636207	0.629338	0.655115	2.56 (3/3)
iTransformer	ETTh2/96	0.240174	0.227519	0.236105	0.227245	**0.216086**	5.26 (3/3)

Every cell–seed pair shares one complete native-Adam warmup epoch. Online pre-update losses select $\mathcal{H}$ by the median-plus-3-MAD rule, with $\mathcal{C}$ its complement. All six arms branch from the

same weights, first and second moments, optimizer age, model buffers, and random state. They share the subsequent epoch permutations. Each epoch visits every origin, including the final partial mini-batch; only the participation arm zeros the selected losses. Within each all-data arm, a single stochastic forward pass computes per-window losses, preserving identical dropout realizations across both regions. To obtain the unrenormalized regional gradients $g_{\mathcal{H},j}$ and $g_{\mathcal{C},j}$ defined in Equation (**??**), the loss tensors corresponding to sample indices in $B_j \cap \mathcal{H}$ and $B_j \cap \mathcal{C}$ are aggregated separately and backpropagated (either sequentially with `retain_graph=True` sharing the cached activations, or via vectorized per-sample gradient routines). While evaluating two distinct gradient vectors introduces an additional backward projection pass compared to standard single-objective ERM, it operates entirely on the existing forward graph and introduces zero auxiliary neural networks or iterative inner-loop projections. Random-short draws one fixed set of $|\mathcal{H}|$ origins per cell and seed and applies the H-short formula to that set.

Separating current and accumulated contributions. Let $j \geq 1$ index updates after warmup, s_0 be the warmup update count, and m_0, v_0 its Adam moments. For the current mini-batch B_j, define

$$g_{A,j} = \frac{1}{|B_j|} \sum_{i \in B_j \cap A} \nabla_\theta \ell_i(\theta_j), \quad A \in \{\mathcal{H}, \mathcal{C}\}, \qquad g_{F,j} = g_{\mathcal{H},j} + g_{\mathcal{C},j}. \tag{18}$$

The original mini-batch denominator preserves each window's coefficient. Within each all-data arm, native moments and post-warmup regional histories follow

$$\begin{aligned} m_j &= \beta_1 m_{j-1} + (1-\beta_1) g_{F,j}, \qquad v_j = \beta_2 v_{j-1} + (1-\beta_2) g_{F,j} \odot g_{F,j}, \\ q_{A,j} &= \beta_1 q_{A,j-1} + (1-\beta_1) g_{A,j}, \quad q_{A,0} = 0. \end{aligned} \tag{19}$$

Thus $m_j = \beta_1^j m_0 + q_{\mathcal{H},j} + q_{\mathcal{C},j}$. With $a_j = 1 - \beta_1^j$, replace the chosen history by its current contribution at the same accumulated coefficient:

$$\begin{aligned} \widetilde{m}_j^{\mathcal{H}\text{-short}} &= m_j - q_{\mathcal{H},j} + a_j g_{\mathcal{H},j}, \\ \widetilde{m}_j^{\mathcal{C}\text{-short}} &= m_j - q_{\mathcal{C},j} + a_j g_{\mathcal{C},j}, \\ \widetilde{m}_j^{\text{all-short}} &= \beta_1^j m_0 + a_j g_{F,j}. \end{aligned} \tag{20}$$

The inherited warmup moment remains in every arm. These replacements change the update numerator while preserving the native moment recurrences. At $j = 1$ they equal ordinary Adam, so all five all-data arms take the same first update.

Adaptive scaling and controls. All-data arms use the full-gradient second moment, including the cross terms between $g_{\mathcal{H},j}$ and $g_{\mathcal{C},j}$. Let

$$P_j = \sqrt{v_j / (1 - \beta_2^{s_0+j})} + \epsilon, \qquad u_j = -\eta_j \frac{\widetilde{m}_j / (1 - \beta_1^{s_0+j})}{P_j}. \tag{21}$$

All operations are coordinatewise; complex parameters use separate real and imaginary coordinates. Each modified proposal is scaled by a positive scalar to match the actual parameter-update L2 norm of native Adam at that arm's current state. The numerical tolerance is $\max(10^{-7}, 10^{-4}\|\Delta\theta_j^{\text{Adam}}\|)$. The reference is evaluated at each arm's current state; different arms can therefore have different parameters and second moments. Moment histories and optimizer age retain their native recurrences. C-short changes which set receives the memory intervention; Random-short in Table 22 matches the selected count. The $\mathcal{H}$-masked arm, named RAMP-S$_{\text{warmup}}$ in the main table, feeds $g_{\mathcal{C},j}$ to native Adam with its own first and second moments and natural step norm. The subscript identifies the shared complete warmup state and fixed training budget used by this panel.

Model and training configuration. The experiment uses the original PaiFilter and iTransformer implementations with lookback $L = 96$, seven input/output channels, and multivariate MSE. PaiFilter uses hidden size 256, batch size 16, initial learning rate 0.005, and 15 total epochs. iTransformer uses two encoder layers, eight heads, model and feed-forward dimensions 512, batch size 32, initial learning rate 10^{-4}, and 10 total epochs. Adam uses $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, and no weight decay. The original learning-rate schedule uses η_0 in epochs 1 and 2, then $\eta_0 2^{-(e-2)}$ in epoch $e \geq 3$. Memory and bias-correction age continue across epoch boundaries. Training runs in FP32 on

Table 24: All paired final-epoch regional-memory results. MSE is evaluated on the complete validation split at epoch 15 for PaiFilter and epoch 10 for iTransformer. Gain is $100(\mathrm{MSE}_{\mathrm{Adam}} - \mathrm{MSE}_{\mathcal{H}\text{-short}})/\mathrm{MSE}_{\mathrm{Adam}}$. No validation-best epoch or seed is selected.

Model	Dataset/H	Seed	Adam	$\mathcal{H}$ short	$\mathcal{C}$ short	All short	$\mathcal{H}$ masked	Gain
PaiFilter	ETTh2/96	1	0.222901	0.216015	0.223332	0.216583	0.218092	3.09
PaiFilter	ETTh2/96	2	0.222523	0.216621	0.222298	0.217795	0.217493	2.65
PaiFilter	ETTh2/96	3	0.223344	0.215552	0.222656	0.216329	0.218707	3.49
PaiFilter	ETTm2/96	1	0.128403	0.122813	0.127630	0.124001	0.119977	4.35
PaiFilter	ETTm2/96	2	0.128494	0.122798	0.127295	0.124004	0.120353	4.43
PaiFilter	ETTm2/96	3	0.127209	0.122711	0.126507	0.124107	0.119985	3.54
PaiFilter	ETTh1/96	1	0.693424	0.680044	0.685544	0.683841	0.674207	1.93
PaiFilter	ETTh1/96	2	0.697179	0.682160	0.688232	0.686498	0.678862	2.15
PaiFilter	ETTh1/96	3	0.690846	0.676780	0.684379	0.683261	0.676682	2.04
PaiFilter	ETTh2/336	1	0.379126	0.371723	0.378241	0.372881	0.377662	1.95
PaiFilter	ETTh2/336	2	0.374411	0.368879	0.374147	0.371313	0.377806	1.48
PaiFilter	ETTh2/336	3	0.382074	0.370094	0.383097	0.374377	0.378003	3.14
PaiFilter	ETTh2/720	1	0.643923	0.623889	0.631089	0.625664	0.651281	3.11
PaiFilter	ETTh2/720	2	0.645925	0.631579	0.639732	0.631244	0.658755	2.22
PaiFilter	ETTh2/720	3	0.645290	0.630233	0.637802	0.631106	0.655308	2.33
iTransformer	ETTh2/96	1	0.232615	0.220145	0.227352	0.220114	0.214787	5.36
iTransformer	ETTh2/96	2	0.240842	0.230225	0.236613	0.230124	0.217078	4.41
iTransformer	ETTh2/96	3	0.247064	0.232185	0.244351	0.231497	0.216394	6.02

Table 25: Prespecified secondary readouts for $\mathcal{H}$ short versus Adam. Entries are mean paired relative MSE reductions (%); parentheses give improved seeds. Each seed's MSE is first averaged over the indicated epochs before computing its relative reduction. E is the model's fixed final epoch.

Model	Dataset/H	Last five epochs	Epochs 2–E
PaiFilter	ETTh2/96	3.07 (3/3)	2.99 (3/3)
PaiFilter	ETTm2/96	4.11 (3/3)	3.78 (3/3)
PaiFilter	ETTh1/96	2.03 (3/3)	1.90 (3/3)
PaiFilter	ETTh2/336	2.18 (3/3)	2.46 (3/3)
PaiFilter	ETTh2/720	2.55 (3/3)	2.46 (3/3)
iTransformer	ETTh2/96	5.27 (3/3)	4.77 (3/3)

CPU with two threads for the complete epoch budget and evaluates every validation origin. For the main table, each arm's minimum-validation-MSE epoch determines its test readout. The tables below instead use the fixed final epoch or the stated averages of validation errors along the continuation.

H-short improves on Adam in all 18 paired runs under each of these validation summaries. The benefit persists across the final epoch, the last five epochs, and the continuation as a whole. On the final-epoch validation readout, H-short improves on masking by 0.93%, 2.01%, and 4.05% on PaiFilter/ETTh2 at $H = 96, 336, 720$, with agreement in all nine paired runs. At $H = 336$, it also exceeds All-short by 0.70% in all three seeds. Across the panel, H-short improves mean final-epoch validation MSE over All-short in $5/6$ configurations ($14/18$ seed-level wins). Masking has the lowest final-epoch validation error on the other three configurations. These comparisons describe training-stage behavior; Table 22 reports test performance after validation-based checkpoint selection.

Across forecasters and datasets. The shared-state six-arm panel isolates the effect of regional history. We also report the complete H-short readout over all 160 model–dataset–horizon cells in the $5 \times 8 \times 4$ grid. For each arm, seeds 1–3 contribute test MSE at that arm's validation-selected epoch; the detailed tables give each arm's mean $\pm$ sample standard deviation and the paired H-short gain relative to ERM/Adam. The main-grid Adam runs and H-short continuations follow their respective batch orders, so this readout measures benchmark coverage rather than the shared-state contrast above. H-short has lower mean test MSE in 99/160 cells, including 27/32 TimePro, 22/32 iTransformer, 20/32 PaiFilter, 18/32 PatchTST, and 12/32 MixLinear cells. Table 26 gives the model-level summary; Tables 27 to 31 give every cell.

Table 26: H-short versus the ERM/Adam baseline on the complete 160 model–dataset–horizon cells. A win means lower three-seed mean test MSE. Gain is $100(\overline{\mathrm{MSE}}_{\mathrm{Adam}} - \overline{\mathrm{MSE}}_{\mathrm{H-short}})/\overline{\mathrm{MSE}}_{\mathrm{Adam}}$, summarized across cells as mean $\pm$ sample standard deviation (%).

Model	Cells	Wins	Gain (%)
MixLinear	32	12	-0.15 $\pm$ 0.40
PaiFilter	32	20	+0.68 $\pm$ 1.65
TimePro	32	27	+1.12 $\pm$ 1.23
iTransformer	32	22	+0.41 $\pm$ 1.01
PatchTST	32	18	+0.78 $\pm$ 1.86
All	160	99	+0.57 $\pm$ 1.38

Table 27: Validation-selected test MSE on MixLinear across all 32 cells (8 datasets $\times$ 4 horizons). We compare standard ERM/Adam, physical participation control (RAMP-S), and regional memory control ($\mathcal{H}$-short). Bold marks the lowest test MSE across all three paradigms. Gains (Δ) are relative to standard Adam (%).

Dataset	H	ERM/Adam	RAMP-S	$\mathcal{H}$-short (Ours)	Δ_{RAMP} (%)	$\Delta_{\mathcal{H}\text{-short}}$ (%)
ETTh1	96	0.3920	**0.3898**	0.3902	+0.56	+0.46
	192	**0.4503**	**0.4503**	0.4506	0.00	-0.06
	336	0.4779	**0.4777**	0.4783	+0.05	-0.09
	720	**0.4616**	0.4618	0.4620	-0.04	-0.09
ETTh2	96	**0.3057**	0.3069	0.3059	-0.39	-0.08
	192	**0.3890**	0.3904	0.3895	-0.36	-0.14
	336	**0.4235**	0.4270	0.4273	-0.83	-0.91
	720	**0.4222**	0.4230	0.4236	-0.18	-0.34
ETTm1	96	0.3494	**0.3493**	0.3545	+0.04	-1.47
	192	0.4027	**0.4023**	0.4027	+0.10	0.00
	336	**0.4366**	0.4370	0.4379	-0.10	-0.30
	720	0.5017	**0.5003**	0.5056	+0.27	-0.79
ETTm2	96	0.1899	0.1915	**0.1898**	-0.84	+0.10
	192	0.2548	0.2555	**0.2547**	-0.27	+0.06
	336	**0.3153**	0.3155	0.3154	-0.07	-0.00
	720	0.4162	**0.4161**	0.4163	+0.03	-0.02
ECL	96	**0.2147**	**0.2147**	**0.2147**	0.00	-0.00
	192	0.2055	0.2055	**0.2055**	+0.02	+0.02
	336	0.2186	**0.2184**	0.2185	+0.09	+0.04
	720	0.2602	**0.2601**	0.2604	+0.02	-0.08
Weather	96	**0.1980**	0.2024	0.2004	-2.22	-1.21
	192	0.2463	0.2473	**0.2459**	-0.41	+0.16
	336	**0.2980**	**0.2979**	0.2985	+0.02	-0.17
	720	**0.3714**	**0.3711**	0.3718	+0.07	-0.13
Solar	96	0.4190	0.4229	**0.4190**	-0.93	+0.01
	192	0.4688	**0.4680**	0.4681	+0.18	+0.16
	336	0.5128	**0.5122**	0.5131	+0.12	-0.06
	720	0.5176	**0.5172**	0.5178	+0.08	-0.02
Traffic	96	**0.6626**	0.6631	0.6631	-0.07	-0.07
	192	0.6113	0.6115	**0.6109**	-0.03	+0.07
	336	0.6171	0.6178	**0.6167**	-0.11	+0.05
	720	0.6549	0.6555	**0.6545**	-0.09	+0.06

Table 28: Validation-selected test MSE on PaiFilter across all 32 cells (8 datasets $\times$ 4 horizons). We compare standard ERM/Adam, physical participation control (RAMP-S), and regional memory control ($\mathcal{H}$-short). Bold marks the lowest test MSE across all three paradigms. Gains (Δ) are relative to standard Adam (%).

Dataset	H	ERM/Adam	RAMP-S	$\mathcal{H}$-short (Ours)	Δ_{RAMP} (%)	$\Delta_{\mathcal{H}\text{-short}}$ (%)
ETTh1	96	0.3849	**0.3747**	0.3785	+2.64	+1.66
	192	0.4465	**0.4338**	0.4473	+2.85	-0.16
	336	0.5239	**0.5067**	0.5180	+3.29	+1.14
	720	0.6062	**0.5824**	0.6054	+3.92	+0.13
ETTh2	96	0.2935	**0.2807**	0.2898	+4.35	+1.25
	192	0.3698	0.3627	**0.3559**	+1.91	+3.76
	336	0.4297	**0.3973**	0.4049	+7.54	+5.77
	720	0.4544	**0.4092**	0.4299	+9.95	+5.40
ETTm1	96	0.3219	0.3179	**0.3191**	+1.23	+0.90
	192	0.3634	0.3628	**0.3619**	+0.17	+0.39
	336	0.3945	**0.3944**	**0.3942**	+0.03	+0.08
	720	0.4601	0.4551	**0.4563**	+1.08	+0.81
ETTm2	96	0.1770	**0.1716**	0.1726	+3.03	+2.49
	192	0.2389	0.2371	**0.2357**	+0.75	+1.35
	336	0.2964	0.2967	**0.2925**	-0.09	+1.32
	720	0.3937	0.3936	**0.3918**	+0.03	+0.48
ECL	96	0.1665	0.1667	0.1675	-0.14	-0.61
	192	**0.1767**	**0.1766**	0.1772	+0.03	-0.29
	336	**0.1931**	**0.1930**	0.1936	+0.04	-0.27
	720	0.2332	0.2337	0.2334	-0.23	-0.12
Weather	96	0.1662	**0.1661**	0.1660	+0.07	+0.12
	192	**0.2102**	0.2116	0.2107	-0.66	-0.22
	336	0.2679	**0.2675**	0.2678	+0.15	+0.03
	720	**0.3463**	0.3467	0.3465	-0.12	-0.05
Solar	96	**0.2254**	0.2419	0.2280	-7.30	-1.15
	192	**0.2591**	0.2716	0.2626	-4.81	-1.35
	336	**0.2859**	0.2974	0.2874	-4.02	-0.55
	720	**0.2836**	0.2937	0.2857	-3.56	-0.77
Traffic	96	0.4962	**0.4958**	0.4961	+0.09	+0.02
	192	0.4967	0.4970	**0.4955**	-0.06	+0.25
	336	0.5088	0.5089	**0.5084**	-0.02	+0.09
	720	**0.5417**	0.5408	0.5429	+0.16	-0.23

Table 29: Validation-selected test MSE on TimePro across all 32 cells (8 datasets $\times$ 4 horizons). We compare standard ERM/Adam, physical participation control (RAMP-S), and regional memory control ($\mathcal{H}$-short). Bold marks the lowest test MSE across all three paradigms. Gains (Δ) are relative to standard Adam (%).

Dataset	H	ERM/Adam	RAMP-S	$\mathcal{H}$-short (Ours)	Δ_{RAMP} (%)	$\Delta_{\mathcal{H}\text{-short}}$ (%)
ETTh1	96	0.3790	**0.3754**	0.3798	+0.96	-0.23
	192	0.4346	**0.4296**	0.4333	+1.16	+0.30
	336	0.4782	**0.4737**	0.4765	+0.93	+0.36
	720	0.4904	0.4915	**0.4888**	-0.23	+0.34
ETTh2	96	0.3011	**0.2873**	0.2890	+4.59	+4.02
	192	0.3729	0.3656	**0.3601**	+1.95	+3.44
	336	0.4242	0.4194	**0.4182**	+1.14	+1.43
	720	0.4424	0.4365	**0.4231**	+1.33	+4.36
ETTm1	96	0.3251	0.3248	**0.3242**	+0.08	+0.27
	192	0.3674	0.3647	**0.3651**	+0.73	+0.62
	336	0.4009	**0.3969**	0.4001	+0.99	+0.22
	720	0.4722	**0.4667**	0.4673	+1.17	+1.04
ETTm2	96	0.1754	**0.1732**	**0.1730**	+1.24	+1.37
	192	0.2427	0.2403	**0.2386**	+0.98	+1.71
	336	0.3052	0.2987	**0.2982**	+2.13	+2.27
	720	0.4020	0.3973	**0.3937**	+1.17	+2.08
ECL	96	0.1430	**0.1360**	0.1417	+4.93	+0.90
	192	0.1597	0.1582	**0.1584**	+0.91	+0.80
	336	0.1772	0.1802	**0.1772**	-1.70	-0.16
	720	0.2159	**0.2114**	0.2175	+2.08	-0.75
Weather	96	0.1656	**0.1600**	0.1632	+3.39	+1.45
	192	0.2142	0.2139	**0.2099**	+0.16	+1.99
	336	0.2717	0.2741	**0.2697**	-0.89	+0.75
	720	0.3574	0.3539	**0.3502**	+0.99	+2.01
Solar	96	0.2063	0.2073	**0.2013**	-0.50	+2.47
	192	0.2357	0.2408	**0.2327**	-2.15	+1.25
	336	0.2527	0.2623	**0.2504**	-3.78	+0.92
	720	0.2563	0.2605	**0.2551**	-1.62	+0.45
Traffic	96	0.4049	0.4085	**0.4031**	-0.89	+0.45
	192	0.4307	0.4342	**0.4307**	-0.82	-0.15
	336	0.4481	0.4520	**0.4466**	-0.87	+0.33
	720	**0.4787**	0.4869	0.4802	-1.72	-0.32

Table 30: Validation-selected test MSE on iTransformer across all 32 cells (8 datasets $\times$ 4 horizons). We compare standard ERM/Adam, physical participation control (RAMP-S), and regional memory control ($\mathcal{H}$-short). Bold marks the lowest test MSE across all three paradigms. Gains (Δ) are relative to standard Adam (%).

Dataset	H	ERM/Adam	RAMP-S	$\mathcal{H}$-short (Ours)	Δ_{RAMP} (%)	$\Delta_{\mathcal{H}\text{-short}}$ (%)
ETTh1	96	0.3901	**0.3842**	0.3874	+1.50	+0.68
	192	0.4458	**0.4393**	0.4407	+1.45	+1.15
	336	0.4888	0.4866	**0.4862**	+0.44	+0.53
	720	**0.5116**	**0.5116**	0.5137	0.00	-0.41
ETTh2	96	0.3040	**0.2909**	0.2989	+4.32	+1.69
	192	0.3792	**0.3711**	0.3753	+2.14	+1.02
	336	0.4339	0.4339	**0.4325**	0.00	+0.33
	720	**0.4418**	**0.4418**	0.4503	0.00	-1.92
ETTm1	96	0.3438	0.3369	**0.3377**	+2.00	+1.78
	192	0.3940	**0.3815**	0.3885	+3.16	+1.38
	336	0.4309	**0.4185**	0.4213	+2.87	+2.23
	720	0.5019	**0.4890**	0.5006	+2.56	+0.26
ETTm2	96	0.1845	**0.1760**	0.1810	+4.62	+1.89
	192	0.2514	**0.2434**	0.2491	+3.20	+0.91
	336	0.3153	**0.3032**	0.3118	+3.83	+1.09
	720	0.4145	**0.4007**	0.4111	+3.32	+0.81
ECL	96	0.1481	**0.1467**	0.1477	+0.96	+0.30
	192	0.1652	0.1603	**0.1626**	+2.96	+1.58
	336	0.1786	**0.1747**	0.1772	+2.20	+0.81
	720	0.2157	0.2173	**0.2140**	-0.75	+0.79
Weather	96	0.1738	**0.1729**	0.1758	+0.51	-1.18
	192	0.2246	**0.2236**	0.2242	+0.44	+0.19
	336	0.2851	**0.2817**	0.2825	+1.20	+0.94
	720	0.3604	0.3601	**0.3593**	+0.09	+0.29
Solar	96	**0.2092**	0.2138	0.2125	-2.21	-1.58
	192	0.2419	0.2490	**0.2413**	-2.93	+0.27
	336	**0.2614**	0.2726	0.2627	-4.28	-0.47
	720	**0.2648**	0.2755	0.2676	-4.04	-1.06
Traffic	96	**0.4006**	0.4054	0.4026	-1.20	-0.50
	192	**0.4221**	0.4274	0.4228	-1.26	-0.16
	336	**0.4367**	0.4429	0.4379	-1.41	-0.27
	720	**0.4694**	0.4745	0.4702	-1.09	-0.16

Table 31: Validation-selected test MSE on PatchTST across all 32 cells (8 datasets × 4 horizons). We compare standard ERM/Adam, physical participation control (RAMP-S), and regional memory control ($\mathcal{H}$-short). Bold marks the lowest test MSE across all three paradigms. Gains (Δ) are relative to standard Adam (%).

Dataset	H	ERM/Adam	RAMP-S	$\mathcal{H}$-short (Ours)	Δ_{RAMP} (%)	$\Delta_{\mathcal{H}\text{-short}}$ (%)
ETTh1	96	0.3829	**0.3783**	0.3804	+1.19	+0.65
	192	0.4338	0.4293	**0.4280**	+1.03	+1.35
	336	0.4713	**0.4650**	0.4673	+1.34	+0.86
	720	0.5283	0.5286	**0.5028**	-0.06	+4.84
ETTh2	96	0.2996	**0.2818**	0.2869	+5.95	+4.25
	192	0.3755	0.3658	**0.3619**	+2.59	+3.61
	336	0.4243	0.4156	**0.4093**	+2.05	+3.54
	720	0.4428	0.4323	**0.4275**	+2.37	+3.45
ETTm1	96	0.3258	**0.3235**	0.3234	+0.72	+0.75
	192	0.3657	0.3667	**0.3653**	-0.28	+0.12
	336	0.3991	**0.3974**	0.3962	+0.43	+0.74
	720	0.4576	0.4576	**0.4561**	+0.01	+0.34
ETTm2	96	0.1770	0.1754	**0.1753**	+0.90	+0.93
	192	0.2500	**0.2394**	0.2416	+4.25	+3.35
	336	0.3100	**0.3008**	0.3026	+2.96	+2.40
	720	0.4182	**0.4010**	0.4053	+4.11	+3.10
ECL	96	0.1819	**0.1816**	0.1825	+0.15	-0.33
	192	0.1887	**0.1882**	0.1897	+0.25	-0.53
	336	0.2047	**0.2042**	0.2058	+0.22	-0.53
	720	**0.2467**	0.2473	0.2468	-0.25	-0.05
Weather	96	**0.1730**	0.1734	0.1733	-0.22	-0.16
	192	**0.2200**	0.2201	0.2201	+0.12	-0.08
	336	0.2754	**0.2753**	0.2758	+0.05	-0.13
	720	0.3524	0.3525	**0.3523**	-0.04	+0.01
Solar	96	**0.2221**	0.2341	0.2258	-5.39	-1.70
	192	**0.2517**	0.2606	0.2580	-3.55	-2.52
	336	**0.2700**	0.2826	0.2725	-4.68	-0.92
	720	**0.2693**	0.2807	0.2754	-4.24	-2.25
Traffic	96	**0.4809**	0.4824	0.4817	-0.32	-0.17
	192	0.4859	0.4864	**0.4849**	-0.11	+0.21
	336	**0.5007**	0.5016	0.5017	-0.19	-0.19
	720	**0.5348**	0.5363	0.5356	-0.28	-0.14

Adjustable regional history replacement. Equation (9) changes only the strength of the H-short numerator replacement. We evaluate $\lambda \in \{0, 0.25, 0.5, 0.75, 1\}$ from the same warmup states, keeping support, batch order, learning-rate schedule, epoch budget, and norm matching fixed. This gives 90 continuations over six configurations and three seeds. The readout is complete-validation MSE at the final epoch. At $\lambda = 1$, all $18/18$ runs reproduce the original H-short metrics exactly. At $\lambda = 0$, the native numerator is used through the regional implementation and its actual-update norm matching; the standard Adam arm follows its natural-norm implementation. The mean relative difference is -0.01%, as reported in Table 3. The five measured mean gains increase with replacement strength. This coefficient controls optimizer history; the soft participation weights in Section F.4 act on window losses instead.

E.4 What the first-moment intervention removes

Unrolling the zero-initialized selected history gives $q_{\mathcal{H},j} = (1-\beta_1)\sum_{k=1}^{j}\beta_1^{j-k} g_{\mathcal{H},k}$. The weights sum to $a_j = 1-\beta_1^j$, so the numerator discrepancy is exactly

$$d_j = q_{\mathcal{H},j} - a_j g_{\mathcal{H},j} = (1-\beta_1)\sum_{k=1}^{j}\beta_1^{j-k}(g_{\mathcal{H},k} - g_{\mathcal{H},j}). \tag{22}$$

All terms belong to one arm's history, so d_j measures how its accumulated regional contribution differs from the current one. Both parameter movement and mini-batch composition contribute.

Checkpoint cosines describe regional geometry at two states; evaluating d_j additionally requires the intervening gradient history.

The identity connects the modification to gradient evolution. If consecutive regional contributions satisfy $\|g_{\mathcal{H},k+1} - g_{\mathcal{H},k}\| \le \delta$ along a given history, the triangle inequality yields

$$\|d_j\| \le (1-\beta_1)\delta \sum_{h=0}^{j-1} h\beta_1^h \le \frac{\beta_1}{1-\beta_1}\delta. \tag{23}$$

Constant regional gradients give $d_j = 0$ and no intervention; variation is needed for the two directions to differ. Variation alone does not imply that removing d_j improves the objective.

To state the improvement condition, let u_j be the native same-state Adam proposal and $\widetilde{u}_j$ the proposal with numerator $m_j - d_j$, before norm matching. With P_j from (21),

$$\widetilde{u}_j - u_j = \frac{\eta_j}{1-\beta_1^{s_0+j}} \frac{d_j}{P_j}. \tag{24}$$

For nonzero proposals define $e_j = u_j/\|u_j\|$, $\widetilde{e}_j = \widetilde{u}_j/\|\widetilde{u}_j\|$, and $\rho_j = \|u_j\|$. Suppose a target loss F has M-Lipschitz gradient on the two step segments. Applying the first-order remainder bound to both steps gives

$$F(\theta_j + \rho_j\widetilde{e}_j) - F(\theta_j + \rho_j e_j) \le \rho_j \nabla F(\theta_j)^\top(\widetilde{e}_j - e_j) + M\rho_j^2. \tag{25}$$

Thus a sufficient local improvement condition is $\nabla F(\theta_j)^\top(\widetilde{e}_j - e_j) < -M\rho_j$. This is a same-state, equal-norm directional condition. The implemented intervention uses training gradients alone. The paired training results test whether its directional change benefits the full continuation.

E.5 Where the value of continued participation turns over

We read intermediate steps from the original-order controls in Section E.2, using the fixed checkpoint-derived preconditioner, inherited first-moment history, and natural update norms at the same nine starting checkpoints. Table 32 follows full-validation MSE at six shared update counts within these continuations. Mean masked-minus-full MSE is positive through the measured 16-step point and negative at 32, locating an aggregate sign change between those points. Individual trajectories vary: one favors withholding at four steps and one at eight, none at sixteen, and seven at thirty-two. The nine trajectories are the three seeds at each of three starting checkpoints.

Table 32: Continuation-step ladder on MixLinear/ETTh1. Mean masked-minus-full validation MSE is in units of 10^{-3}; negative favors withholding. Counts use the same nine stage–seed trajectories at every step.

Updates	1	2	4	8	16	32
Mean difference	$+0.048$	$+0.115$	$+0.262$	$+0.622$	$+0.774$	-0.100
Withholding wins	$0/9$	$0/9$	$1/9$	$1/9$	$0/9$	$7/9$

E.6 Random-position control

It is the positions, not the count. Withholding as many windows per epoch at uniformly random positions yields $+1.31$, -0.05 and -0.74 points on the same pairs, against 3.84, 1.34 and 1.25 for the selected regions in the dynamic arm (Table 14), so the excess attributable to choosing the right positions is 2.53, 1.39 and 1.99 points. Random withholding has mixed effects: on PaiFilter/ETTh2, removing any twenty-six percent already lets the remainder fall faster, while on the other two pairs it slightly slows them, which is why the excess over the matched control, not the absolute acceleration, tracks test accuracy. On MixLinear/Weather it inverts ($+0.00$ against -0.98), so there the cost comes from removing the selected regions rather than from removing windows as such.

E.7 Batch-local trimming at a matched rate

The random-position control shows that count alone is insufficient. We next isolate ranking population and persistent origin membership by holding the retention rate close to RAMP's own rate per cell. The batch-local member drops the largest q fraction of *each batch*, ranked at the current step with

Table 33: Participation controls. Differences are paired percentage points of relative MSE reduction. The rows vary the withholding rate, the set over which losses are ranked, or cross-epoch identity.

Contrast	cells	median diff.	p
How the withheld count is obtained			
RAMP-D − best single fixed rate	17	-0.02	0.96
RAMP-D − rate swept on validation	17	-0.00	0.28
The population the losses are ranked in			
batch-local, frozen − RAMP-S	132	$+0.03$	3×10^{-4}
batch-local, per step − RAMP-D	17	-0.08	0.225
Whether the withheld set is held fixed			
RAMP-S − RAMP-D	160	$+0.09$	4×10^{-10}
batch-local, frozen − batch-local, per step	17	$+0.61$	8×10^{-4}
batch-local, frozen − RAMP-D	132	$+0.15$	5×10^{-10}
Population and persistence together			
RAMP-S − per-batch trimming at RAMP's rate	132	-0.028	0.86

no cross-epoch memory, and keeps the $1/N$ denominator. The rate is read from the static arm's first-epoch mask log, and each cell is matched to the nearest rate we swept; the residual mismatch is a median 0.42 and at most 2.52 percentage points. This supplied-rate arm measures how much aggregate accuracy batch-local ranking captures at matched count. A batch-local median–MAD cutoff can also derive a similar rate (Section F.1); the supplied version fixes the count for this comparison.

Section E.9 carries an exact-rate version of this arm over twelve cells, where RAMP-S is ahead in 7 of 12 ($+0.9$ points, $p = 0.774$). Across the 128 five-seed cells, per-batch trimming reaches a median $+0.63\%$ against ERM (94 positive, 38 negative), while RAMP-S reaches $+0.44\%$ (92/37). The paired comparison assigns no aggregate difference: RAMP-S is ahead in 63 cells by a median -0.028 points ($p = 0.86$). Matching the retention rate under the same score-based ranking therefore recovers the aggregate gain, and this arm does not separate rate from score selection on its own. The per-epoch variant sets RAMP-D against the same trimmer (43 of 128, median -0.174 points, $p < 10^{-4}$). Because the per-batch logs are not indexed by global origin, this arm tests aggregate accuracy rather than positional equivalence; the displacement intervention supplies the positional test. The comparison in which a trimmer derives its own count appears in Sections E.7 and F.5.

The aggregate paired result separates into two regimes. Ordering cells by RAMP-S's validation gain, a quantity fixed before any test set is read, the difference grows with it (Spearman $+0.38$, $p = 7 \times 10^{-6}$), RAMP-S leading by a median $+0.42$ points in the top third ($p = 2 \times 10^{-4}$) and trailing by -0.24 in the bottom third ($p = 0.040$). Two other candidate moderators do not separate the arms: the difference is uncorrelated with the cell's support-acceleration ratio (Pearson $+0.05$, $p = 0.78$, over the 32 cells that carry both) and with the retention rate itself (Spearman $+0.09$, $p = 0.28$). Figure 7 shows the cell-level agreement at Pearson $+0.81$; the displacement controls in Section E.9 separately establish that translating a fixed origin-space set eliminates the gain.

The control intentionally supplies RAMP's rate and begins selection in the first epoch, forming a strong matched-rate bound for batch-local selection. On iTransformer/ETTh2, the cell where we sweep q, trimming reaches $+2.81\%$ at RAMP's 25.9% rate and $+4.01\%$ at $q = 0.10$, against RAMP-S's $+4.32\%$. The first-epoch analysis in Section F.8 separately measures the role of RAMP's standard first epoch.

Together, these controls separate rate adaptation from origin-indexed membership. The withheld fraction varies from 3.3% to 26.0% across cells. Median–MAD derives a cell-adaptive rate from either the global or a batch-local score population (Section F.1), avoiding a validation sweep. Among arms that recompute membership every epoch, the derived rate gains $+1.80\%$, the best single fixed rate gains $+1.82\%$ (-0.02 points, $p = 0.96$), and a per-seed validation sweep over 6 candidates gains $+2.06\%$ (median -0.00, $p = 0.28$). Section F.5 expresses the same comparison as regret against a per-cell oracle. Persistence supplies the second component: freezing the first-epoch set instead of recomputing it is worth $+0.32$ points over the 160 cells (128 ahead, $p = 4 \times 10^{-10}$). The fixed set provides the temporal coordinate used by the displacement and delayed-activation interventions. On the 3 cell–rate pairs where a trimming run and a sweep run happen to sit at an identical q, per-batch trimming and a global per-epoch cutoff differ by -1.36 to $+0.24$ points, in both directions.

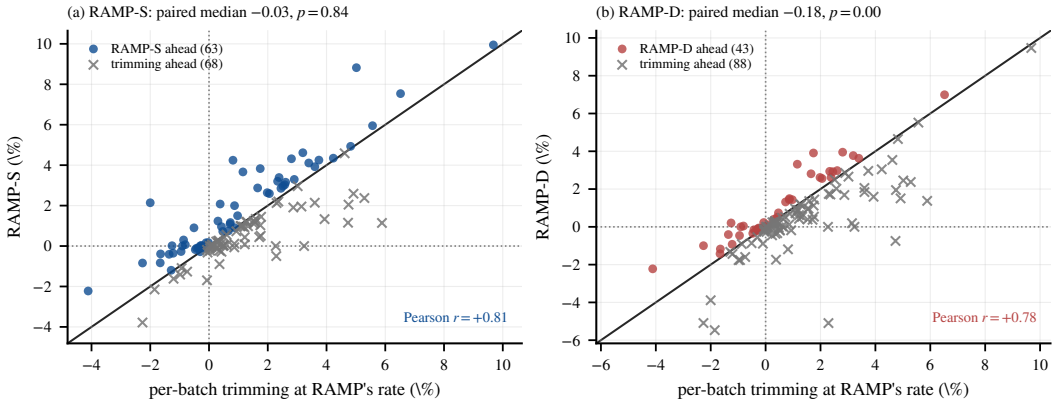


Figure 7: RAMP against per-batch trimming at RAMP-S's retention rate over the 128 five-seed cells; the last row of Table 33 gives the paired summary. Each point is one cell and the diagonal is equality. **(a)** The fixed schedule isolates aggregate accuracy at matched count. **(b)** Recomputing RAMP membership each epoch separates the schedule comparison from the fixed form.

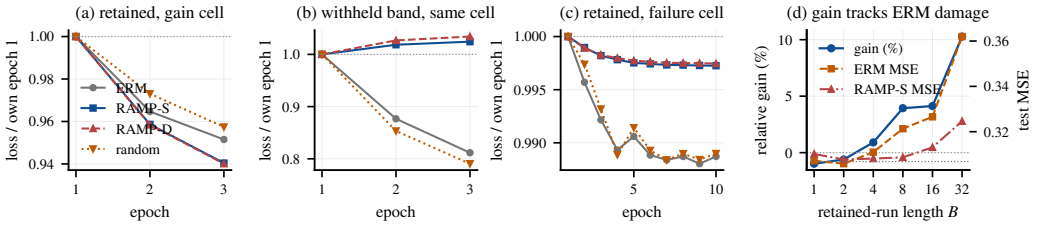


Figure 8: Loss trajectories of the retained and the withheld windows under each intervention, at H=96 and seed 1. The two groups are fixed once, by thresholding the first epoch of the ERM run, so every arm is measured on the same partition; each arm is truncated to the epochs all arms share and divided by its own first-epoch group mean, so the plotted quantity is that arm's improvement rate. "Random" withholds as many windows per epoch at uniformly random positions. **(a)** iTransformer/ETTh2, where RAMP improves accuracy: the retained windows fall faster under RAMP than under ERM or random withholding. **(b)** The withheld set on the same cell does the opposite: ERM and random withholding drive it down, under RAMP it rises. **(c)** MixLinear/Weather, where RAMP costs accuracy: the ordering inverts, and RAMP's retained windows improve *less* than ERM's. **(d)** Retaining one eighth of the windows in contiguous runs of B windows, sweeping B from 1 to 32 with the number of windows and the gradient-step budget held fixed: the relative gain rises with B but tracks how much each arm's own ERM degrades, and RAMP-S's absolute error never beats ERM trained on all windows (dotted).

E.8 Loss trajectories of the two groups

The matched-rate comparison removes selection count as the source of the aggregate difference. To expose what fixed membership makes observable, we compare the two trajectories on the same window groups. Scores are recorded for every window in every epoch even though training uses only the retained ones, so the quantities remain comparable after partitioning windows once by the first epoch of the ERM run (Figure 8).

Targeted withholding accelerates the rest. On the three pairs where RAMP improves accuracy, the windows remaining in training improve by 3.84, 1.34 and 1.25 percentage points more than under ERM in RAMP-D, with the same signs under RAMP-S (Figure 8a and Table 14; three seeds). The withheld set moves the other way: its loss falls 9 to 24% under ERM over the same epochs and *rises* 2 to 5% under RAMP. On MixLinear/Weather, where RAMP costs accuracy, the retained windows improve 0.98 points less than under ERM in RAMP-D (Figure 8c).

The arrangement of retained windows. The retained-group acceleration is location dependent: the effect also follows how the retained windows are spread over the series. Holding the number of training windows and the gradient-step budget fixed and keeping the retained windows in contiguous runs of eight, RAMP-S improves accuracy in 6 of 6 cells. Spacing the same number of windows uniformly instead, it degrades accuracy in 2 of them. Sparsification does not cost the two geometries

equally, so we report each geometry against its own ERM arm and give the absolute levels, together with the one confound this comparison carries, in Section C.2.

Dynamic versus fixed membership. Within the RAMP family, location dependence leaves cross-epoch membership as the remaining schedule choice. Over the complete grid RAMP-S is ahead of RAMP-D in 128 of the 160 cells, by a median of $+0.09$ points (two-sided paired Wilcoxon $p = 4 \times 10^{-10}$). Freezing the map is therefore the reference RAMP schedule. Two controls separate membership from count. The refreshed arm in Section E.9 reselects origins at the frozen count, whereas the rate controls in Section F.5 vary how a dynamic schedule obtains that count. Together they isolate revision of *which* origins participate from selection of *how many*.

E.9 The displacement factorial, cell by cell

The trajectory analysis connects accuracy to the retained group's response. The displacement factorial turns that association into a controlled location test. Figure 9 shows every cell behind the group-level summary of Table 34 lists them with the two further arms. Both directions of the effect are visible per cell: where RAMP-S gains, the two displaced arms do not, and where RAMP-S costs accuracy, displacing the same set returns the run to its baseline.

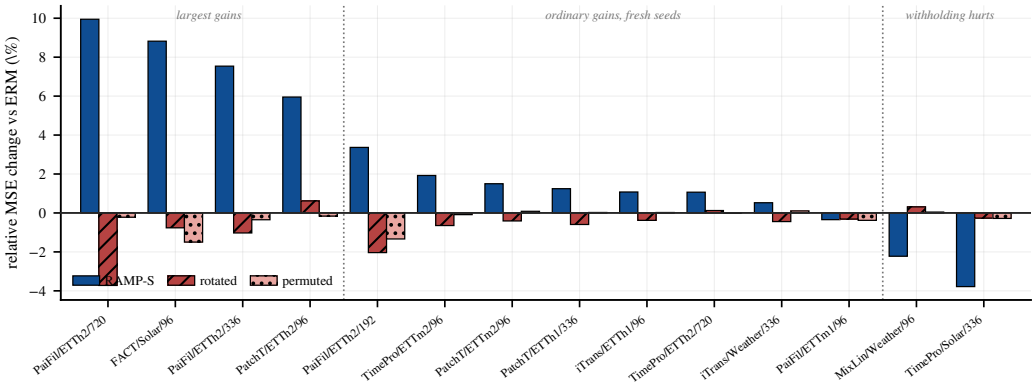


Figure 9: The displacement factorial cell by cell, five seeds each, against each cell's own ERM. *Rotated* shifts RAMP-S's own withheld set along the window index at an offset that splits no run, so the count and the multiset of run lengths are preserved exactly; *permuted* keeps the count alone. Dotted lines separate three predeclared groups with different roles. The largest-gain group was selected on the same outcome the contrast then measures. The ordinary-gain group comes from the middle of the gain distribution and uses seeds not used when the cells were predeclared. The remaining group contains cells where withholding costs accuracy.

The refreshed arm complements spatial displacement by changing membership over time. It recomputes membership from the current score field at the frozen count, preserving count and score-based selection while removing cross-epoch persistence. Its location and contiguity are therefore outcomes, not properties preserved by construction. It retains most of the gain: RAMP-S is ahead of it in 9 of 12 cells by a mean of $+0.8$ points ($p = 0.146$). The same frozen-versus-recomputed axis is measured with greater resolution over the 160 cells of the grid ($+0.32$ points, 128 ahead, $p = 4 \times 10^{-10}$; Table 33, Section E.8). The two arms are not identical, since refreshed recomputes membership at the frozen count while RAMP-D recomputes the count as well, but both re-select, and the grid-scale contrast is the one the paper relies on.

E.10 Constructions of the displacement arms

The arms. The factorial's interpretation depends on which property each arm preserves. Every arm below derives its withheld *count* from the same run's own first-epoch RAMP-S mask, so the number of windows removed is identical by construction rather than matched through an external rate. *Rotated* shifts that mask along the window index by an offset chosen so that no maximal run is split at the wrap point, which leaves the multiset of run lengths exactly unchanged and moves only where the runs sit. *Permuted* shuffles the mask over the window index, keeping the count and destroying contiguity. *Refreshed* recomputes the mask from the current epoch's scores at the frozen count,

Table 34: Every cell of the mask-displacement factorial of Section E.9. The table gives the complete cell-level comparison. Relative test-MSE change against each cell's own ERM, five seeds. The first four cells were selected as the largest RAMP-S gains of the grid and use seeds 1–5; the next eight were drawn from the ordinary part of the distribution and use seeds 6–10 with their own ERM arm. The final two are predeclared boundary cells where RAMP-S trails ERM and use seeds 1–5.

Cell	RAMP-S	rotated	permuted	refreshed	trimmed
Largest gains of the grid, seeds 1–5					
PaiFilter/ETTh2/H=720	+9.95	−3.72	−0.22	+9.94	+5.26
PaiFilter/ETTh2/H=336	+7.54	−1.03	−0.35	+7.01	+5.53
PatchTST/ETTh2/H=96	+5.95	+0.62	−0.16	+5.50	+4.96
Ordinary gains, seeds 6–10					
iTransformer/ETTh1/H=96	+1.08	−0.38	+0.02	+1.37	+1.27
PatchTST/ETTh1/H=336	+1.25	−0.59	+0.02	+0.68	+0.80
TimePro/ETTh2/H=720	+1.07	+0.12	+0.01	+1.56	+1.26
TimePro/ETTm2/H=96	+1.92	−0.65	−0.09	+1.18	+1.90
PaiFilter/ETTm1/H=96	−0.34	−0.31	−0.38	−0.57	+0.44
iTransformer/Weather/H=336	+0.52	−0.44	+0.10	+0.57	+0.82
PaiFilter/ETTh2/H=192	+3.37	−2.04	−1.34	+1.45	+3.89
PatchTST/ETTm2/H=96	+1.50	−0.41	+0.08	+0.03	+0.66
Cells where withholding costs accuracy, seeds 1–5					
TimePro/Solar/H=336	−3.78	−0.27	−0.28	−3.81	−0.81
MixLinear/Weather/H=96	−2.22	+0.31	+0.05	−2.22	−2.22

preserving count and score-based selection while removing cross-epoch persistence. *Trimmed* drops the largest fraction of *each mini-batch* at this run's own first-epoch rate, with the same first-epoch pass RAMP gives and an exact per-epoch quota. It keeps the rate without global ranking or persistent origin membership.

E.11 How far the withheld set may move

Figure 10 plots the sweep and Table 35 reports it step by step. Two channels have to be read against each other here: what a displacement costs in gain, and what it does to the withheld set itself. They come apart. Over the whole range in which the gain disappears the set's mean loss barely moves, so the level of the withheld windows' loss cannot be what the effect follows.

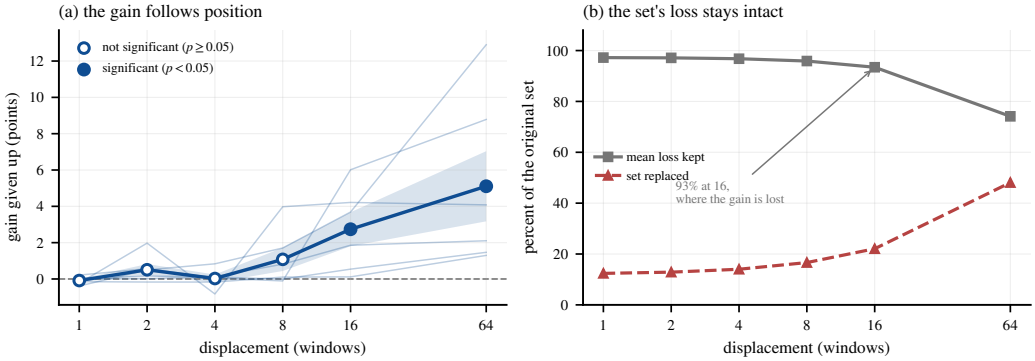


Figure 10: Displacement sweep over the 6 cells of Section E.11, five seeds each. **(a)** Points of gain given up against displacement; faint lines are individual cells, the heavy line their mean, and the shaded band one standard error. Filled circles mark displacements at which the paired loss reaches $p < 0.05$ and open circles those at which it does not. **(b)** The same displacements described through the set they produce: the mean loss of the withheld set as a percentage of its undisplaced value, and the percentage of the set's members that have been replaced.

E.12 Changing the region's width instead of its position

The displacement arms move the withheld region while holding its size fixed. Two further arms do the converse, changing how wide the region is while leaving where it sits, and they bound the width the same way the sweep bounds the position.

Table 35: Displacement of the frozen set along the window index, over 6 cells and five seeds each, continuing the rotated arm of the factorial comparison, which moves the set by a third of the series, down to displacements of a few windows, where Table 36 instead varies the set's shape at a matched score. "Score kept" is the displaced set's mean first-epoch loss as a fraction of the original set's; "paired loss" is the decrease in relative MSE-improvement points against undisplaced RAMP-S.

Displacement	score kept	set replaced	paired loss	cells losing	p
1	97%	12%	−0.09	2/6	.370
2	97%	13%	+0.51	5/6	.167
4	97%	14%	+0.03	4/6	.913
8	96%	17%	+1.08	4/6	.154
16	93%	22%	+2.74	6/6	.033
64	74%	48%	+5.11	6/6	.046

Cumulative keeps the first-epoch selection withheld and adds, at every later epoch, the origins that epoch's score field selects, so the withheld set is the running union of the first-epoch set and every later selection and grows monotonically instead of staying at the first epoch's count. Over 15 cells at $H = 96$ with five seeds each, RAMP-S is ahead of it in 13, by a mean of $+1.47$ points (sign test $p = 0.0074$, Wilcoxon $p = 0.0054$). Enlarging the withheld set along the later epochs' own rankings therefore carries a cost, and the first epoch's count is not a lower bound that any larger set would improve upon. The growth is substantial on the cells where it costs most: on TimePro/ECL the withheld set runs from 1,790 windows to 2,644, and on TimePro/Solar from 4,995 to 7,049, each an increase of between about 41 and 48% over the frozen count.

Because cumulative withholding changes both extent and membership over time, the *dilated* arm isolates extent at a fixed location. It takes the frozen first-epoch mask and extends every withheld run outward by r windows on each side, merging runs that meet. This widens the region by a known amount at an unchanged location. On iTransformer at $H = 96$ over ETTh1, ETTh2 and ETTm1 with five seeds each, $r = 2$ enlarges the withheld set by about a sixth and $r = 5$ by about a quarter, and RAMP-S is ahead of $r = 2$ in 3 of 3 cells (mean $+0.39$ points, $p = 0.2500$) and of $r = 5$ in 2 of 3 (mean -0.18, $p = 1.0000$). At this sample size neither dilation is distinguishable from the rule's own boundary.

The two arms enlarge the withheld set in different ways (Figure 11). Their size ranges overlap: cumulative withholding ends between 1.13 and 1.67 times the frozen count, dilation by five windows between 1.23 and 1.76 times it. Dilation extends the frozen region outward at an unchanged location; cumulative withholding admits whichever windows score highest at each later epoch, so it reselects members as well as adding them. The two comparisons are not matched: dilation covers 3 cells and is indistinguishable from RAMP-S at that sample size, cumulative withholding covers 15 and RAMP-S leads it on 13. The two arms give a descriptive comparison of two ways of enlarging the set: added windows at a fixed location, and added windows chosen afresh each epoch. Position is separated from size by the matched displacement control in Section E.9.

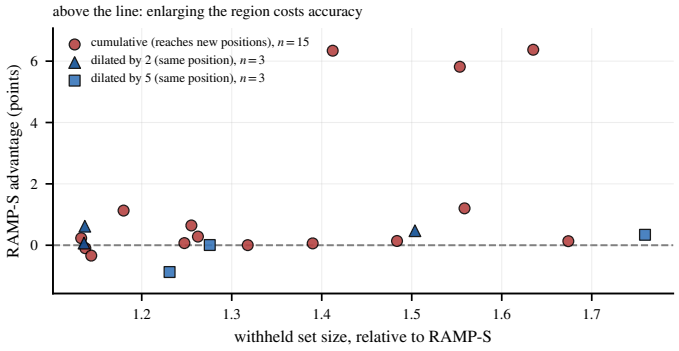


Figure 11: Enlarging the withheld region instead of moving it. Each point is one cell, five seeds; the horizontal axis is the arm's withheld-set size as a multiple of the count RAMP-S freezes, and the vertical axis is how many points RAMP-S is ahead of that arm. Dilation widens the same region symmetrically and holds its position; cumulative withholding admits newly high-scoring windows each epoch and so reaches new positions. The two families overlap in size but not in cost.

Table 36: Halving RAMP-S's withheld set three ways, varying its shape and its membership where Table 35 varied its position; the scattered column is the membership control Section E.9 reads against displacement. Relative test-MSE change against each cell's own ERM, five seeds. The first four cells use seeds 1–5, the rest seeds 6–10 with their own ERM arm.

Cell	RAMP-S	scattered	in runs	top half
PaiFilter/ETTh2/H=720	$+9.95$	$+3.75$	$+1.24$	$+10.19$
PaiFilter/ETTh2/H=336	$+7.54$	$+4.42$	$+1.09$	$+3.02$
PatchTST/ETTh2/H=96	$+5.95$	$+3.32$	-0.40	$+4.98$
iTransformer/ETTh1/H=96	$+1.08$	$+0.98$	$+0.55$	$+1.09$
PatchTST/ETTh1/H=336	$+1.25$	$+0.84$	$+0.71$	$+0.65$
TimePro/ETTh2/H=720	$+1.07$	$+0.38$	-0.11	$+1.80$
TimePro/ETTm2/H=96	$+1.92$	$+1.76$	-0.15	$+2.31$
PaiFilter/ETTm1/H=96	-0.34	$+0.22$	$+0.40$	$+0.49$
iTransformer/Weather/H=336	$+0.52$	$+0.39$	-0.01	$+0.07$
PaiFilter/ETTh2/H=192	$+3.37$	$+1.71$	-1.27	$+2.36$
PatchTST/ETTm2/H=96	$+1.50$	$+1.39$	$+0.68$	$+1.83$

E.13 Contiguity, at a matched score distribution

Section E.9 establishes that displacement reduces the benefit of withholding. Given the original location, does the *contiguity* of the set contribute anything of its own?

This construction cannot vary top rank and contiguity independently, because the top-ranked windows are themselves contiguous. Forcing a scattered set of high-loss windows by imposing a minimum spacing walks far enough down the ranking that the mean withheld score falls from 1.32 to 0.57 on iTransformer/ETTh2 against a population mean of 0.46, so such an arm confounds contiguity with score.

We therefore split RAMP-S's own withheld set. Each arm below removes about half of it, from inside the same supports, and differs only in which half: *scattered* takes every other window, giving runs of one and a mean withheld score that matches RAMP-S to within 0.005; *in runs* takes every other whole maximal run after sorting runs by mean score, leaving run lengths intact; *top half* takes the highest-scoring half.

Scattering wins (Table 36). Over 12 cells the scattered half beats the in-runs half in 11 of them, by a mean of $+1.6$ points, and the top-scoring half beats the scattered half in 9. The ordering is top half, then scattered, then in runs. Each half is defined by which windows it keeps, and the ordering follows that composition: taking whole runs retains their low-scoring edges while discarding half the cores, and the scattered half matches the mean withheld score of RAMP-S to within 0.005. What the ordering separates is a role for the observed region and a role for the membership chosen inside it. The runs locate the region, while subsampling within that support avoids retaining low-score run edges at a matched count. Region describes the spatial structure of the measured load; high-score membership determines which origins the tested intervention withholds.

F Participation schedule, score, and rate controls

F.1 Batch-local variants: derived rates and fixed membership

The matched per-batch trimming control uses the rate derived by $\mathcal{H}$-short. We next test whether a batch-local rule can derive a comparable rate directly. Applying the same median $+\ k\,$MAD cutoff *within each mini-batch*, at the current step, with no supplied rate and nothing remembered across batches or epochs, withholds a median 16.6% of windows where the global rule withholds 13.3% on the same cells. The batch-local rate follows the global one across cells, and the batch-local estimate is biased upward by at most 5.7 points. The batch estimate is computed from 32 windows at the current step.

On the 17 cells that carry every arm at five seeds, the batch-local rule gains a median $+1.01\%$ against ERM (16 cells positive, 1 negative). It is indistinguishable from $\mathcal{H}$-short (-0.08 points, $p = 0.225$) and from per-batch trimming at RAMP's own rate (-0.02, $p = 0.43$), and behind RAMP-S by a median $- - 0.38$ points (3 of 17, $p = 0.002$). Deriving the rate therefore does not require the global window population.

Table 37: Measured per-epoch wall-clock and test MSE, one GPU, arms interleaved, 2 repeats, epoch 1 excluded. Percentages are against that cell's own ERM; "kept" is the fraction of windows retained by the frozen mask, which bounds what dropping them can save. The MSE columns show that dropping them does not change the outcome.

Cell	kept	ERM s/epoch	$\mathcal{H}$-short	+skip	$\mathcal{H}$-short MSE	+skip MSE
PatchTST/Weather	86.2%	44.5 s	+10.6%	−3.2%	0.17381	0.17335
iTransformer/ECL	89.3%	39.2 s	+5.9%	−6.3%	0.14694	0.14711
iTransformer/ETTh1	86.6%	4.1 s	+19.8%	+9.2%	0.38151	0.38135

One batch-local first pass achieves comparable aggregate accuracy using scores from the forward pass already performed; its wall-clock cost is measured in Section F.2. Across both ranking populations, freezing first-pass membership improves on recomputation and gives the set stable membership on the time index. Matched-rate trimming evaluates the batch-local score selection under a supplied rate, while globally ranked fixed membership preserves the exact temporal coordinates manipulated in Section E.9.

F.2 Training cost per epoch

The frozen batch-local rule uses first-pass scores without a global ranking pass. Table 37 measures per-epoch cost on one GPU with one process at a time. The three arms are interleaved across 2 repeats per cell to balance host-load drift. Timing excludes the first, unmasked warmup epoch.

Once the set is frozen, dropping withheld windows from the loader changes median epoch cost by -3% against ERM (-6 to $+9\%$ across 3 cells), with absolute test-MSE differences no larger than 4.6×10^{-4}. The measurement implementation retains every forward pass for trajectory analysis and costs a median $+11\%$ more per epoch ($+6$ to $+20\%$); its overhead comes from per-window loss extraction and index bookkeeping. A reselected rule must score every window each epoch, so loader exclusion is specific to the frozen schedule. On the smallest cell, where an epoch takes about four seconds, the residual scatter of the host dominates the measurement; the reported ranges retain this variability.

F.3 Why the score is the loss

RAMP ranks windows by their own loss, which reports current prediction error and not the direction of the gradient. A natural alternative is to measure interference directly: the part of a window's gradient that opposes the direction the rest of the data agrees on. Direct interference scoring produces lower accuracy in every tested form.

The quantity is affordable in closed form. For a linear prediction head, $g_W = r_\ell z^\top$ and $g_b = \sum r_\ell$ with $r_\ell = (\hat{y} - y)\,\sigma$, so one batched pass yields every window's head gradient without a backward pass through the encoder. To check that this 49k-dimensional head-gradient representation preserves the direction of the whole model's gradient, we computed the exact per-sample gradient over all 3.26M parameters for every one of the $8{,}449$ training windows of iTransformer/ETTh2 and correlated the two cosines against the full-batch ERM direction: Spearman 0.980. Ranking by anti-consensus *mass* $\|g\| \max(0, -\cos)$, the head-level and the exact field select the same windows at Jaccard 0.842.

Two facts about that field explain the outcome. Loss and gradient magnitude are near-redundant (Spearman 0.846), while loss and gradient *direction* are unrelated (Spearman -0.067), so the alternative genuinely scores something else, selecting a set that overlaps RAMP's at Jaccard 0.277, with 17.5% of its windows below the median loss. The field is also banded, at 162 runs of mean length 13.1 against 1.33 under permutation, so region structure is not what is lost. What the two fields disagree about is which regions to remove, and every way of acting on the disagreement costs accuracy. On PatchTST/ETTh2 at a matched 25% withholding, withholding by conflict gives -1.08% against ERM where withholding by loss gives $+5.47\%$. Exempting the 26% of RAMP's withheld set that also carries anti-consensus mass and refilling from the next hardest windows gives $+1.98\%$, exempting without refilling $+1.50\%$, and withholding *only* the overlap $+1.85\%$, against RAMP-S's $+5.95\%$. In these comparisons the loss-selected support remains the more useful one to withhold; the outcome is consistent with the anti-consensus component being a correction the model is applying rather than damage it is absorbing.

Table 38: Regret against a per-cell oracle rate, over 17 cells. Each fixed rate is the same for every cell; the validation-selected row picks among the six per cell and seed at 6 times the training cost; RAMP-S reads its rate off each cell's own score population at one run per seed. Lower is better; the oracle is zero by construction. Table 33 states the same two rate rows as paired differences in gain, and Table 7 shows that at equal count the two cutoff shapes select the same mask, so these rows isolate the count and not the boundary.

Rate	mean regret	median regret	worst cell
fixed $q = 5\%$	1.188	1.065	2.37
fixed $q = 10\%$	0.672	0.469	1.95
fixed $q = 15\%$	0.630	0.304	2.16
fixed $q = 20\%$	0.585	0.432	1.69
fixed $q = 25\%$	0.633	0.319	2.58
fixed $q = 30\%$	0.725	0.348	3.27
q chosen per seed on validation loss	0.347	0.251	2.14
RAMP-D, rate read from the score population	0.607	0.318	1.99
RAMP-S, read from the score population	**0.104**	**0.079**	1.49

F.4 Soft participation weights around selected regions

This experiment varies window participation weights; Section E.3 instead varies optimizer history with all windows retained. Loss identifies the relevant regions, while their smooth spatial profile suggests a softer way to act on them. The elevation around a selected region decays over the window index with a 6–9 window half-life, which suggests replacing the $0/1$ mask with a graded weight: selected region at $w = 0$, neighbors ramping back to 1 over a scale ℓ, with the total withheld mass pinned to what median–MAD withholds so the hard core shrinks to pay for the neighborhood. Against a mass-matched hard mask on PatchTST/ETTh2 the graded form gains $+0.43$, $+0.60$, $+0.73$, $+0.81$ and $+0.91$ points at $\ell = 4, 8, 16, 32, 64$; beyond 64, the fixed-mass construction becomes infeasible. Selection over seven candidates identifies $\ell = 64$ on the same five seeds. A transfer test on three new cells then places the hard mask ahead on all three (-0.13, -0.08, and -0.24 for graded minus hard; pooled -0.15, $p = 0.24$), making the hard mask the fixed-support rule used in the main comparison.

F.5 How close the automatic rate comes to a per-cell oracle

The hard-mask transfer result fixes how membership is expressed. The remaining threshold question is how many windows enter that mask. We measure how well it does that against an oracle that is not available in practice: for each cell, the best of six fixed global rates, chosen after seeing that cell's test outcome. *Regret* is how many percentage points of gain an arm gives up against that oracle on the same cell, so zero is the oracle and larger is worse. The comparison runs on the 17 cells for which we swept all six rates ($q \in \{5, 10, 15, 20, 25, 30\}\%$, five seeds each), and note that a fixed top-q cutoff and the median–MAD rule select bitwise identical masks whenever their counts agree (Section B.3), so this isolates the choice of count alone. Every fixed-rate arm recomputes its mask each epoch, so RAMP-D is the matched row. RAMP denotes the fixed-support form, called RAMP-S in schedule ablations, and additionally freezes the first epoch's set. The following comparisons separate this persistence effect from rate.

The oracle distributes its selected extent across all six candidates (Figure 12 and Table 38), confirming that the useful rate is cell-dependent. The best single global choice, $q = 20\%$, has mean regret 0.59, and every fixed rate exceeds 1.6 points on at least one cell. Selected rates span 5 to 30%, with seed-level disagreement in 11 of the 17 cells. Per-seed validation selection has regret 0.35 at 6 training runs per seed; score-derived RAMP-D has regret 0.61 in one run, and freezing the resulting first-pass set reduces regret to 0.10. Section E.8 reports the same protocols as gains and separates adaptive extent from persistence.

F.6 What the automatic rate presumes about the score population

Deriving the rate avoids a cross-cell constant, but it introduces a distributional requirement. A median and a median absolute deviation are defined for any score population, so the cutoff is always computable. The threshold moves with k only when the median absolute deviation is positive; where it is zero the cutoff sits at the median for every k, and the rule can no longer be retuned through k.

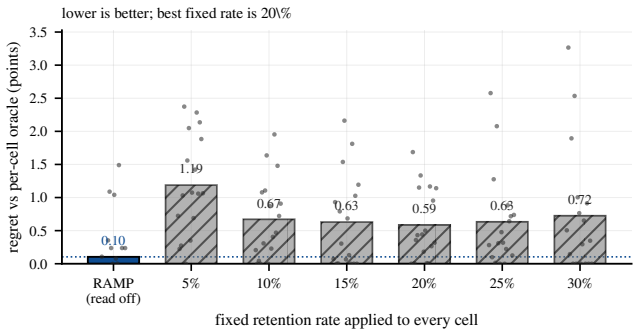


Figure 12: Regret against a per-cell oracle over the 17 cells for which all six rates were swept, five seeds each; Table 38 is the same comparison as a table. Bars are means and dots are individual cells. RAMP, which reads the rate off the score population and freezes the first epoch's set, gives up 0.10 points; the best single fixed rate, $q = 20\%$ chosen after seeing every cell's outcome, gives up 0.59, and RAMP-D, which derives the rate but recomputes the mask as those arms do, gives up 0.61, so the gap between the fixed rates and the fixed-support form is persistence, not rate selection.

Table 39: RAMP-D with k selected per seed on validation MSE, over the 128 core cells. Entries are (ERM − RAMP-D)/ERM in percent, horizon-averaged over $H \in \{96, 192, 336, 720\}$ and five seeds; bold exceeds $+2\%$.

Model	ETTh1	ETTh2	ETTm1	ETTm2	ECL	Weather	Solar	Traffic
MixLinear	-0.0	-0.9	$+0.5$	-0.2	$+0.0$	-0.7	*n/a*	-0.0
PaiFilter	$\mathbf{+6.2}$	$\mathbf{+5.2}$	$+0.7$	$+0.3$	*n/a*	*n/a*	*n/a*	*n/a*
TimePro	$+1.5$	$\mathbf{+2.2}$	$+1.1$	$+1.2$	$+1.8$	$+1.3$	-1.9	*n/a*
iTransformer	$+1.6$	$+1.6$	$\mathbf{+2.5}$	$\mathbf{+3.0}$	$\mathbf{+2.2}$	$+0.7$	*n/a*	-0.7
PatchTST	$+1.7$	$\mathbf{+3.7}$	$+0.6$	$\mathbf{+2.1}$	$+0.2$	$+0.0$	*n/a*	-0.1

The boundary appears on one fixed checkpoint. On iTransformer/ETTh2 the anti-consensus component of the head gradient (Section F.3) is *exactly* zero on every window whose gradient makes an acute angle with the full-batch ERM direction, which holds for 57.7% of the $8{,}449$ windows; the remaining 42.3% are the windows whose gradient opposes that direction. The median and the median absolute deviation of this field are both zero, so $\text{med} + k\,\text{MAD} = 0$ for every k and the rule withholds every window with a nonzero head score. The automatic rate is therefore a property of the score distribution, not of the median–MAD form in isolation. With more than half the scores at zero, k cannot change the withheld fraction, although the rule still produces its fixed 42.3%; retuning the extent then requires another rule or an externally supplied count, at which point Section E.7 applies. The experiments use the per-window loss (Section F.3).

F.7 The same grid with the threshold tuned per seed

On RAMP-D we select $k \in \{1, 2, 3, 4, 5\}$ independently for each cell and seed by validation loss, then read test MSE only after selection. Per-seed tuning improves 96 of the 128 core cells with median $+0.65\%$ against ERM. The main comparison retains the single global $k = 3$ schedule because it tests the regional intervention without per-cell search. Since this arm selects among five candidates, its per-cell p-values condition on that selection and are not directly comparable with a single fixed configuration. The fixed and tuned matrices preserve the same architecture ordering, with the purely linear forecaster at the low-response end in both.

F.8 Cells where validation selects the observational epoch

Threshold tuning changes which mask would be selected, but any mask still enters after the first unmasked ERM epoch. On a cell whose validation loss is lowest at the first epoch and does not improve afterwards, early stopping returns the epoch-1 checkpoint, which predates every mask update, so the mask cannot reach the weights that get evaluated. The masking rule still forms a nonempty set on schedule, but that mask never reaches the evaluated checkpoint, producing a reported 0.00%.

Across all 160 cells, the recorded test MSE of RAMP-S equals ERM on no seed in 142 cells, on one seed in 10, on two in 2, on three in 2, on four in 1, and on all five in three cells: iTransformer at ETTh1/H=720, ETTh2/H=336 and ETTh2/H=720. Equality of the recorded metrics does not establish identical checkpoints. These three iTransformer cells are the configurations where the epoch-2 mask withholds 652, 1,460 and 1,960 windows respectively and validation never improves again. All three all-seed-equal cells contribute exactly 0.00% and therefore fall into neither the positive nor the negative count.

This timing difference explains the trimming comparison in Section E.7: per-batch trimming acts from the first optimizer step, whereas RAMP's first epoch is standard ERM.

F.9 Dynamic versus fixed membership, in full

Dynamic versus fixed membership. The spatial interventions identify which origins matter. The full-grid comparison now isolates the advantage of retaining that membership within the RAMP family. On the complete 160-cell matrix at the same global threshold $k = 3$, RAMP-S is ahead of RAMP-D in 128 cells, by a mean $+0.32$ and median $+0.09$ percentage points ($p = 4 \times 10^{-10}$). Table 1 exposes the benchmark comparison. The two arms share the score, threshold, first unmasked ERM epoch, optimiser, and training budget. They differ in whether the first-epoch membership is fixed or recomputed.

Fixing the first-epoch mask at a threshold fixes its count along with its membership, whereas RAMP-D lets both change as the score distribution shifts across epochs. The full-grid comparison in Table 1 contrasts those two participation schedules and does not by itself separate member turnover from rate drift. Rate is a separate axis: Section F.5 compares two rules that both reselect every epoch, one deriving the count from a dynamic threshold and one holding a supplied count fixed, so the count rule varies while the schedule stays dynamic. Section C.3 reports which windows the revisions add and drop.

G Scope, boundaries, and extensions

G.1 Scope, by the phenomenon's own diagnostic

The previous subsections compare fixed with reselected membership; scope is a different question, asking where RAMP improves on its own ERM baseline. Huber loss supplies a complementary loss-shaping condition because it attenuates large residuals before origin-level participation is applied. On iTransformer/ETTh1 with five seeds, adding RAMP to Huber changes performance by -0.44% relative to Huber ERM ($p = 0.025$), so the intervention does not help under that loss. Under the squared-loss grid, the improvement-rate ratio supplies the corresponding empirical scope diagnostic. The scope summary below reports both forecaster medians and individual cells. Across forecasters, the ratio tracks the observed response ordering; within a forecaster, the per-cell cloud spans both sides of the threshold, so the ratio does not account for all cell-level variation. Section C.7 reports the extent and architecture-conditioned relations.

Table 40: Forecaster-level scope summary. For each forecaster, the table pairs its median improvement-rate ratio with its median RAMP gain over $H = 96$ cells; masking is disabled when the ratio is measured. Table 8 reports the linear forecaster at every threshold.

Forecaster	cells	ratio < 1	median ratio	median gain
PatchTST	7	4	0.85	$+0.72\%$
MixLinear	7	5	0.86	-0.07%
TimePro	7	2	1.29	$+1.24\%$
PaiFilter	4	0	1.52	$+2.83\%$
iTransformer	7	1	1.99	$+1.50\%$

G.2 Longer input context

The main grid fixes $L = 96$ to isolate regional participation across architectures, datasets, and horizons. The controlled extension sets $L = 512$ for iTransformer and PatchTST on ETTh2, ETTm2, ECL, and Weather at $H = 96$. Each ERM/$\mathcal{H}$-short cell uses seeds 1–3 and retains the

main-grid optimizer, early-stopping rule, loss, and global $k = 3$. $\mathcal{H}$-short improves $6/8$ long-context cells, with median $+1.48\%$ and mean $+3.13\%$ MSE reduction ($p = 4 \times 10^{-2}$). The unchanged runs improve MAE on $8/8$ cells, with median $+1.80\%$ and mean $+2.66\%$ reduction ($p = 8 \times 10^{-3}$). The largest paired reductions occur on ETTm2 for both architectures. The absolute MSE and MAE results are reported in Table 41.

Table 41: Long-context MSE/MAE at $L = 512$, $H = 96$ (three-seed mean $\pm$ standard deviation). All other settings match the main protocol. Lower values are bold.

(a) MSE

Model	Dataset	ERM	$\mathcal{H}$-short
iTransformer	ETTh2	0.3309 ± 0.0029	$\mathbf{0.3255 \pm 0.0049}$
iTransformer	ETTm2	0.1950 ± 0.0030	$\mathbf{0.1782 \pm 0.0020}$
iTransformer	ECL	0.1317 ± 0.0010	$\mathbf{0.1299 \pm 0.0012}$
iTransformer	Weather	$\mathbf{0.1690 \pm 0.0039}$	0.1690 ± 0.0040
PatchTST	ETTh2	0.3047 ± 0.0109	$\mathbf{0.2857 \pm 0.0049}$
PatchTST	ETTm2	0.1770 ± 0.0044	$\mathbf{0.1644 \pm 0.0021}$
PatchTST	ECL	$\mathbf{0.1363 \pm 0.0000}$	0.1364 ± 0.0002
PatchTST	Weather	0.1568 ± 0.0012	$\mathbf{0.1562 \pm 0.0018}$

(b) MAE

Model	Dataset	ERM	$\mathcal{H}$-short
iTransformer	ETTh2	0.3749 ± 0.0015	$\mathbf{0.3686 \pm 0.0054}$
iTransformer	ETTm2	0.2788 ± 0.0022	$\mathbf{0.2627 \pm 0.0019}$
iTransformer	ECL	0.2274 ± 0.0007	$\mathbf{0.2247 \pm 0.0004}$
iTransformer	Weather	0.2207 ± 0.0036	$\mathbf{0.2200 \pm 0.0026}$
PatchTST	ETTh2	0.3631 ± 0.0063	$\mathbf{0.3447 \pm 0.0022}$
PatchTST	ETTm2	0.2657 ± 0.0041	$\mathbf{0.2521 \pm 0.0006}$
PatchTST	ECL	0.2400 ± 0.0001	$\mathbf{0.2395 \pm 0.0002}$
PatchTST	Weather	0.2118 ± 0.0017	$\mathbf{0.2077 \pm 0.0034}$

G.3 Traffic on joint origin–channel support

Traffic illustrates a channel-aggregation boundary: its 862 channels share one scalar origin score and one RAMP masking decision. We preserve origin-indexed membership but resolve the first-epoch score by channel. Let N, H, and C denote the numbers of training origins, forecast steps, and channels. For origin i, channel c, and forecast step h, define

$$s_{i,c}^{(1)} = \frac{1}{H} \sum_{h=1}^{H} (\hat{y}_{i,h,c}^{(1)} - y_{i,h,c})^2, \qquad m_{i,c} = \mathbb{1}\left[s_{i,c}^{(1)} \le \mathrm{Med}_i(s_{\cdot,c}^{(1)}) + 3\,\mathrm{MAD}_i(s_{\cdot,c}^{(1)}) \right].$$

The first standard ERM epoch fixes $m_{i,c}$ for all later updates.

Channel-resolved load. At the mask-forming checkpoint, we partition each sensor's gradient to PatchTST's shared linear prediction head into withheld and complementary components. Let ϕ denote the parameters of this head and write the two components as $G^{\phi}_{\mathcal{H},c}$ and $G^{\phi}_{\mathcal{C},c}$. The first two rows of Table 42 aggregate these components in the same prediction-head parameter space and are therefore directly comparable. They remain separate from the full-parameter diagnostic in Table 15.

Table 42: Traffic at joint origin–channel resolution on PatchTST/H=96. The upper panel reports the seed-1 shared-prediction-head diagnostic at mask formation. The lower panel reports validation-selected test MSE over five seeds; positive Δ favours the named arm over ERM. Sensor wins are descriptive counts based on mean per-sensor change across seeds, not independent statistical replicates. The random arm preserves every channel's withholding count exactly.

Support in the shared prediction head	Withheld (%)	R	c
Origin-only aggregate	8.30	0.098	$+0.903$
Origin–channel aggregate	11.77	0.159	$+0.957$
Per-channel median [q25, q75]	–	$0.198\,[0.140, 0.304]$	$+0.710\,[+0.556, +0.825]$

Method	Test MSE	Δ (%)	Seed wins	Sensor wins
ERM	$\mathbf{0.480135 \pm 0.001329}$	–	–	–
RAMP-S	0.481687 ± 0.001190	-0.32	$0/5$	$126/862$
RAMP-C	0.481601 ± 0.001060	-0.31	$0/5$	$134/862$
Matched random	0.480416 ± 0.001297	-0.06	$0/5$	$98/862$

The support is structured before intervention. It withholds 11.77% of origin–channel entries. Runs are measured along the origin index within each channel: the median run length is 3.07 and the lag-1 Jaccard is 0.509, which exceeds a matched-random expectation of about 0.056 by 0.453. The top 5% of channels carry 14.9% of the withheld gradient norm. These are descriptive channel measurements from one seed; the performance unit remains the seed. In the common prediction-head space, the

aggregate support–complement cosine rises from $+0.903$ under the origin-only support to $+0.957$ under the origin–channel support. Resolving the support therefore exposes channel variation without producing an aggregate conflicting component for joint withholding to remove.

Matched intervention. RAMP-C applies the fixed support without renormalizing the retained entries,

$$\mathcal{L}_{\mathrm{RAMP-C}}(\theta) = \frac{1}{NHC} \sum_{i=1}^{N} \sum_{c=1}^{C} \sum_{h=1}^{H} m_{i,c} (\hat{y}_{i,h,c} - y_{i,h,c})^2.$$

Its matched-random control independently permutes origins within every channel, preserving all 862 withholding counts. RAMP-C changes MSE by -0.31% relative to ERM ($0/5$ seed wins); relative to matched random, its paired change is -0.25 percentage points ($0/5$ wins). The result shows that channel-resolved losses expose structure while leaving the aggregate gradient aligned. It therefore points to gradient direction, rather than finer loss resolution alone, as the next control variable and motivates direction-aware origin–channel routing that preserves aligned gradients.

G.4 The improvement-rate ratio

The improvement-rate ratio orders the observed median response across forecasters. Its cell-level resolution is measured over the 32 cells at $H{=}96$ for which both quantities are available: 12 fall on opposite sides of the ratio-one and zero-gain boundaries: 8 have a ratio below one and positive gain, while 4 have a ratio above one and negative gain. The rank correlation between ratio and gain is Spearman 0.57 ($p = .0005$).

The ratio associates forecaster-level response with the phenomenon in this benchmark, whereas cell-level response additionally depends on regional extent and architecture-specific gradient geometry. We analyze those variables in Section C.7.

G.5 Threshold families, in full

The forecaster-level scope diagnostic summarizes where the mechanism is expressed. The threshold family tests whether the operating pattern depends on the fixed $k = 3$ row. Table 43 reports RAMP-D across 128 forecaster–dataset–horizon cells, five seeds, and $k \in \{1, 2, 3, 4, 5\}$, with each threshold reported separately. The fixed global $k = 3$ RAMP arm remains the main comparison because it isolates cross-epoch membership without cell-wise threshold selection.

Table 43: Complete dynamic-mask threshold sweep. Each row aggregates the same 128 core forecaster–dataset–horizon cells and five paired seeds. Entries after the sign counts are relative test-MSE improvements over ERM in percent. This table is descriptive; Table 39 separately selects k by validation MSE per seed.

k	cells	+	−	0	median	mean	q3	max
1	132	64	68	0	−0.02	−0.15	+1.86	+12.78
2	132	73	59	0	+0.19	+0.33	+1.61	+9.83
3	132	79	50	3	+0.23	+0.64	+1.50	+9.47
4	132	87	42	3	+0.29	+0.82	+1.46	+9.13
5	132	86	43	3	+0.26	+0.77	+1.22	+7.21

The sweep also clarifies what changes across threshold families. Median–MAD, mean–standard-deviation, and fixed top-$q\%$ all threshold the same scalar score monotonically. Each therefore admits an upper set of the score ordering, and for any k there exists a q producing a bitwise identical mask. At matched count the selected windows are identical; the families differ in the count they derive. Section 5 verifies this directly: plotted against the withheld fraction, all three families fall on a single curve, and a tuned top-q matches the best median–MAD setting.

What median–MAD provides is therefore a *selector* for that fraction, read off the score population instead of fixed in advance. This matters because the useful fraction is not a constant of the problem: at one fixed k the recorded masks withhold between 3.3% and 26.0% depending on the cell, and the best rate varies across cells, so no single q attains the scan's best rate on every cell. A mean–standard-deviation reference instead has its own tail inflate both moments, landing it on the low-intervention end of the same curve (Section B.3).