TASK-AWARE TEMPORAL COARSENING UNDER FINITE BUDGETS: EXACT RISK, TRACTABLE SURROGATES, AND DOWNSTREAM TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Compressing an H-step history into $K \ll H$ contiguous measurements requires deciding where to allocate temporal resolution. Standard heuristics allocate it uniformly or logarithmically, ignoring both the history covariance Σ and the downstream task geometry M, while the exact task-weighted compression risk $J(\mathcal{P}; \Sigma, M)$ is not block-additive and cannot be minimized exactly by the classical contiguous-partition dynamic program. We study task-aware temporal coarsening with Σ and M estimated on development data only: a dynamic program globally minimizes a block-additive surrogate $\widetilde{J}$, and its minimizer is refined under exact J. This separates three gaps that are usually read as a single gap: approximation, optimization, and transfer. Over $20{,}000$ audit partitions per dataset–horizon cell at $K = 16$, surrogate rank fidelity ranges from 0.953 on Weather to -0.081 on Electricity. Exact J orders six controlled constructions against a held-out linear reader at mean rank agreement 0.980, against 0.362 for the surrogate, yet in a post-hoc validation audit, lower-J partitions reduce neural-reader error in only six of thirteen cases. Under the linear reader, task-aware allocation beats both uniform and logarithmic spacing in all 36 cells; under a Transformer reader, it beats uniform in all twelve cells, negligibly in one, but does not dominate a fixed logarithmic grid. Evaluating coarsened representations, therefore, requires auditing approximation, search, and transfer separately rather than attributing downstream performance to a single objective.

1 INTRODUCTION

When compressing an H-step observation history into a budget of $K \ll H$ contiguous temporal measurements, a model must determine how to allocate resolution across past lags. Standard heuristics set this resolution without reference to the downstream task: uniform grids distribute measurements evenly, while logarithmic grids allocate higher resolution to recent observations under an assumption of temporal decay. However, neither heuristic identifies which temporal distinctions in the history of the target prediction task actually require. That question becomes answerable only when the history covariance Σ and the downstream task geometry M are specified.

Existing approaches typically set temporal resolution using fixed architectural scales (Nie et al., 2023), dynamic routing and token merging (Chen et al., 2024; Ankireddy et al., 2026; Wang et al., 2025; Götz et al., 2025; Pagnoni et al., 2025), or end-to-end optimization within a specific forecasting architecture (Zucchi et al., 2026). In parallel, compressive memory and latent-bottleneck models constrain the stored internal state rather than the input resolution (Rae et al., 2020; Jaegle et al., 2021; Dai et al., 2020; Gu et al., 2020). In contrast, we examine whether a fixed temporal partition can be derived prior to training a downstream predictor, computed from an explicit task-weighted compression criterion, and held constant across different reader architectures. This formulation specializes classical problems in linear model aggregation (Fisher, 1962; Chipman & Winker, 1992), goal-oriented optimal experimental design (Attia et al., 2018), and task-based quantization (Shlezinger et al., 2019; Shlezinger & Eldar, 2021) to the case of contiguous temporal groupings under a fixed measurement budget. While those frameworks optimize continuous acquisition filters or vector quantizers, our formulation restricts representations to contiguous temporal aggregations fixed before downstream training. In related time-series literature, Fechete et al. (2026) reweights error

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.

contributions across forecast horizons while keeping input resolution constant, and Ding et al. (2026) adjusts patch granularity dynamically inside a forecasting model (see Appendix B). This work provides a finite-budget characterization of when the tractable surrogate objective, the exact-risk solution, and empirical downstream performance align, and when they diverge, using a development-only estimation framework.

This coarsening pipeline involves three distinct steps, each with an independent source of error: approximating the non-additive exact risk J with a block-additive surrogate $\widetilde{J}$ to permit dynamic programming; searching for a partition $\widehat{\mathcal{P}}$ under finite computational budgets; and transferring the resulting partition to downstream readers that were not involved in partition selection. The experiments in §5 evaluate these three transitions in sequence.

The compression objective depends on two distinct matrices: the history covariance Σ, which characterizes how historical observations co-vary, and the task geometry matrix M, which defines how the downstream objective weights reconstruction errors. A finite measurement budget requires discarding information, and this selection cannot, in general, be determined from input geometry alone. If two distinct prediction tasks share the same history covariance Σ but have different optimal partitions, any allocation rule operating solely on input statistics, $\mathcal{P} = f(\Sigma)$, assigns an identical partition to both tasks. Such a rule cannot simultaneously recover the task-optimal allocation for both targets. An interval of history may exhibit substantial variance yet contribute little predictive utility to the task, or, conversely, remain relatively stable while carrying an essential signal.

For an observation history $q \in \mathbb{R}^H$ with covariance Σ and task geometry $M \succeq 0$, compressing q into K contiguous bands via an aggregation matrix $P_{\mathcal{P}}$ yields a residual task risk $J(\mathcal{P}; \Sigma, M)$, formulated as in goal-oriented A-optimal design and linear model aggregation (Attia et al., 2018; Fisher, 1962). Formally, J measures the residual mean squared error when reconstructing a linear task signal Cq from the compressed measurements $P_{\mathcal{P}}q$ via an optimal linear estimator, with task metric $M = C^\top C$. We take this criterion as a starting point, not as a contribution.

The primary computational challenge arises from the mathematical structure of J. Exact contiguous partitioning via dynamic programming (Fisher, 1958; Bellman, 1961) requires an objective that decomposes additively across intervals. When Σ is diagonal, J is strictly block-additive, making global optimization via dynamic programming exact. When Σ contains general cross-block covariances, J lacks block additivity. Under these conditions, dynamic programming instead optimizes a block-additive surrogate $\widetilde{J}$, which corresponds to evaluating J on a block-truncated covariance matrix (Proposition 1). Special correlated classes make the two coincide exactly (Theorem 1), but in general nothing guarantees that $\widetilde{J}$ orders partitions as J does. On development-estimated covariances, this ordering is nearly preserved on one benchmark and absent on another (§5.1), so discarding cross-block covariance can materially change which partition is selected.

To evaluate this formulation on empirical data without test-set leakage, we construct an estimator using only development data:

$$\text{TRAIN} \longrightarrow (\widehat{\Sigma}, \widehat{C}) \xrightarrow{\text{VAL selects } \lambda} \widehat{M} \longrightarrow \widetilde{J}\text{-DP} \longrightarrow J\text{-refine} \longrightarrow \mathcal{P}.$$

The sample covariance $\widehat{\Sigma}$ and pilot task mapping $\widehat{C}$ are estimated on the training split, with the ridge regularization parameter λ chosen by validation error to define $\widehat{M} = \widehat{C}^\top \widehat{C}$. An initial partition is generated by globally minimizing the surrogate $\widetilde{J}$ via dynamic programming, followed by local greedy refinement under the exact risk J. The resulting partition is fixed before test evaluation and supplied to downstream neural and linear models. Evaluating this development-only pipeline on the ETTm1, Weather, and Electricity benchmarks (Zhou et al., 2021; Wu et al., 2021) yields four contributions:

1. *Approximation.* Although dynamic programming globally minimizes $\widetilde{J}$, the surrogate can severely misorder partitions: an objective-independent audit of $20{,}000$ uniformly sampled partitions per (dataset, horizon) cell at $K = 16$ gives rank agreement from 0.953 on Weather to -0.081 on Electricity, and the surrogate minimizer has a median excess exact risk of 27.5%.
2. *Optimization.* Local exact-risk refinement recovers a median 89.7% of that gap, but the refined partitions remain a median 3.0% above the lowest exact risk found, so search limitations persist at finite budgets.

3. *Transfer.* Against a held-out linear reader at $K = 16$, exact J orders six controlled constructions at mean rank agreement 0.980, against 0.362 for the surrogate. In a post-hoc validation audit under a fixed neural reader, lower-J substitutions reduce error in six of thirteen cases and increase it in seven. Under a Transformer reader, task-aware allocation has a lower mean TEST MSE than uniform in all twelve cells, negligibly in one, but does not uniformly beat logarithmic spacing, and the ETTm1 criterion fixed before TEST evaluation is not met.
4. *Theory.* The high-resolution law does not explain the strength of logarithmic spacing: the condition $A(x) \propto x^{-3}$ under which $\rho^\star \propto A^{1/3}$ makes a log grid optimal fails on all twelve estimated task geometries.

All gaps are measured conditional on the development-estimated geometry $(\widehat{\Sigma}, \widehat{M})$; the gap between population and estimated geometry is not separately identified. Within each reader pipeline, architecture, capacity, and training protocol are fixed across arms, so temporal allocation is the manipulated variable.

2 Problem formulation: task-aware temporal compression

Let $q = (q_1, \dots, q_H)^\top \in \mathbb{R}^H$ denote an observation history vector of length H, indexed such that larger indices correspond to earlier historical observations (lags). Let $\Sigma = \mathrm{Cov}(q) \in \mathbb{R}^{H\times H}$ be the covariance matrix of q, assumed symmetric and strictly positive definite ($\Sigma \succ 0$). A measurement budget of $K \ll H$ contiguous intervals is defined by a partition $\mathcal{P} = \{I_1, \dots, I_K\}$ of the lag index set $\{1, \dots, H\}$ determined by integer boundaries $0 = h_0 < h_1 < \cdots < h_K = H$, where each interval is $I_k = \{h_{k-1} + 1, \dots, h_k\}$. The compression operation is represented by an aggregation matrix $P_\mathcal{P} \in \mathbb{R}^{K\times H}$. Throughout the coarsenings studied here, row k is uniform within interval I_k, so $p_{I_k} = \alpha_k \mathbf{1}_{I_k}^\top$ for some $\alpha_k \neq 0$, producing the compressed measurement vector $z = P_\mathcal{P} q \in \mathbb{R}^K$. Standard heuristics such as uniform binning, logarithmic spacing, and recent-heavy grids correspond to specific configurations of $\mathcal{P}$.

For a downstream prediction task governed by a symmetric positive semidefinite task geometry matrix $M \in \mathbb{R}^{H\times H}$ ($M \succeq 0$), the residual task risk after compression is defined as

$$J(\mathcal{P}; \Sigma, M) = \mathrm{tr}\big(M\, \mathcal{R}(\mathcal{P}; \Sigma)\big), \qquad \mathcal{R}(\mathcal{P}; \Sigma) = \Sigma - \Sigma P_\mathcal{P}^\top (P_\mathcal{P} \Sigma P_\mathcal{P}^\top)^{-1} P_\mathcal{P} \Sigma. \tag{1}$$

The matrix $\mathcal{R}(\mathcal{P}; \Sigma) \in \mathbb{R}^{H\times H}$ is the error covariance of the minimum-mean-squared-error linear reconstruction of history q from compressed measurements z. In forecasting where target $y = Cq$ with task mapping $C \in \mathbb{R}^{S\times H}$, expected reconstruction loss is $\mathrm{tr}(C\mathcal{R}(\mathcal{P}; \Sigma)C^\top) = \mathrm{tr}(M\mathcal{R}(\mathcal{P}; \Sigma))$ with task geometry $M = C^\top C$. Because invertible reparameterizations of the retained measurements cancel in the projection $P_\mathcal{P}^\top (P_\mathcal{P} \Sigma P_\mathcal{P}^\top)^{-1} P_\mathcal{P}$, J is invariant under them (Lemma 1); in particular, rescaling any row of $P_\mathcal{P}$ leaves J unchanged, so band sums and band means are equivalent for the objective.

The covariance Σ and task geometry M serve distinct structural functions: Σ characterizes second-order history dynamics, while M weights lag reconstruction errors according to target demands. In general, neither matrix alone determines the optimal allocation. The arms of §5 that supply Σ alone, M alone, or both are controls on which component is supplied; they do not causally isolate either component's contribution under every reader.

Two structural properties govern the optimization of equation 1. First, under the uniform aggregation already assumed in this work, when history covariance is diagonal, $\Sigma = D = \mathrm{diag}(d_1, \dots, d_H)$, cross-block covariances vanish. Residual covariance $\mathcal{R}(\mathcal{P}; D)$ decomposes into independent block-diagonal components. The risk is then strictly additive across intervals:

$$J(\mathcal{P}; D, M) = \sum_{k=1}^{K} C_D(I_k), \qquad C_D(I) = \mathrm{tr}(M_{II} D_I) - \frac{d_I^\top M_{II} d_I}{\sum_{j\in I} d_j}, \tag{2}$$

where M_{II} and D_I are principal submatrices restricted to interval I, and $d_I = (d_j)_{j\in I}$ is the diagonal variance vector. Because this cost is additive across intervals, the globally optimal partition is computed by an $O(KH^2)$ Bellman recurrence once interval costs are tabulated (Fisher, 1958; Bellman, 1961).

Second, when Σ is a general non-diagonal covariance matrix, the matrix inverse $(P_\mathcal{P} \Sigma P_\mathcal{P}^\top)^{-1}$ in equation 1 couples all partition intervals together, preventing an exact additive decomposition. To

maintain computational tractability, dynamic programming can instead optimize the block-additive surrogate

$$\widetilde{J}(\mathcal{P};\Sigma,M)=\sum_{k=1}^{K}\operatorname{tr}(M_{I_kI_k}R_{I_k}), \tag{3}$$

where $R_I=\Sigma_{II}-\Sigma_{II}p_I^\top(p_I\Sigma_{II}p_I^\top)^{-1}p_I\Sigma_{II}\in\mathbb{R}^{|I|\times|I|}$ denotes the isolated within-block residual covariance for interval I with aggregation vector p_I, and Σ_{II} is the principal submatrix of Σ restricted to I. The surrogate $\widetilde{J}$ replaces the global coupling in Σ with independent block evaluations, enabling polynomial-time dynamic programming for general covariance structures.

3 Tractable reductions at finite budget and high resolution

This section addresses two theoretical questions: under finite measurement budgets, when the block-additive surrogate $\widetilde{J}$ preserves the ranking of the exact risk J; and in the high-resolution limit $K\to\infty$, which density profile characterizes optimal temporal coarsening.

To formalize the surrogate, let $B_{\mathcal{P}}(\Sigma)\in\mathbb{R}^{H\times H}$ denote the block-diagonal truncation of covariance Σ under partition $\mathcal{P}$, formed by setting all cross-block entries $(\Sigma_{ij})_{i\in I_k,j\in I_l}$ with $k\neq l$ to zero.

Proposition 1 (Surrogate equivalence under block truncation)**.** *For any task metric $M\succeq 0$ and any aggregation matrix $P_{\mathcal{P}}\in\mathbb{R}^{K\times H}$ with one non-zero row supported within each interval I_k such that $P_{\mathcal{P}}B_{\mathcal{P}}(\Sigma)P_{\mathcal{P}}^\top$ is invertible,*

$$\widetilde{J}(\mathcal{P};\Sigma,M)=J\big(\mathcal{P};B_{\mathcal{P}}(\Sigma),M\big). \tag{4}$$

Proposition 1 establishes that dynamic programming optimizes the exact risk under a block-diagonal covariance model. Because the truncation $B_{\mathcal{P}}(\Sigma)$ depends on $\mathcal{P}$, each allocation is evaluated under its own induced block structure. Compound symmetry makes the surrogate exact (Theorem 1, Appendix A) even though the discarded cross-block mass can be almost the whole matrix, so neither the number nor the magnitude of discarded entries controls fidelity, and a bound of the form $|\Delta_{\mathcal{P}}|\le C\|\Sigma-B_{\mathcal{P}}(\Sigma)\|$ is blind to the mechanism. For general correlated covariances, $\Delta_{\mathcal{P}}=J(\mathcal{P};\Sigma,M)-\widetilde{J}(\mathcal{P};\Sigma,M)$ varies across partitions and can take either sign (Proposition 2, Appendix A). Because partition ranking depends on relative differences $\Delta_{\mathcal{P}}-\Delta_{\mathcal{Q}}$ rather than absolute errors, $\widetilde{J}$ carries no rank-preservation guarantee. Theorem 3 (Appendix A.3) identifies one regime in which correlation is harmless: on a fixed $\mathrm{AR}(1)$ lattice, correlation of vanishing physical length rescales the isotropic objective by a constant.

To examine whether asymptotic theory explains heuristic spacings, we consider the high-resolution limit $K\to\infty$. On compact interval $[a,c]$ with $0<a<c$, let $f_1,\dots,f_R\in C^2([a,c])$ denote target task profiles with weights $\pi_r>0$, $\sum_{r=1}^R\pi_r=1$. Under isotropic covariance, the continuous counterpart of equation 1 is population-weighted L^2 error:

$$J_{\mathrm{part}}(\mathcal{P})=\sum_{r=1}^{R}\pi_r\sum_{I\in\mathcal{P}}\int_I\big(f_r(x)-\bar{f}_{r,I}\big)^2\,dx,\qquad A(x)=\sum_{r=1}^{R}\pi_r f_r'(x)^2, \tag{5}$$

where $\bar{f}_{r,I}=|I|^{-1}\int_I f_r(u)\,du$ and $A(x)>0$ is local task roughness. For partitions generated by quantiles of a boundary density ρ, the asymptotic error is minimized by the cube-root law $\rho^\star(x)\propto A(x)^{1/3}$ (Theorem 2, Appendix A). This law is classical: it is the order-one case of variable-knot spline asymptotics and the Panter–Dite point density (Barrow & Smith, 1978; Panter & Dite, 1951; Bennett, 1948); what is specific here is that A arises from task geometry rather than from a source distribution. Comparing this optimal density to logarithmic spacing yields a direct diagnostic:

Corollary 1 (Asymptotic optimality condition for logarithmic spacing)**.** *On $[a,c]$ with $a>0$, the normalized logarithmic boundary density $\rho_{\log}(x)=1/(x\log(c/a))$ coincides with the optimal density $\rho^\star$ almost everywhere if and only if $A(x)=Cx^{-3}$ for a constant $C>0$. When this condition fails, the asymptotic inefficiency ratio satisfies*

$$R_{\log}=\frac{\int_a^c A(x)\rho_{\log}(x)^{-2}\,dx}{\big(\int_a^c A(x)^{1/3}\,dx\big)^3}\ge 1, \tag{6}$$

with equality if and only if $A(x)=Cx^{-3}$.

Algorithm 1 Task-Aware Temporal Coarsening (TATC)

Require: History length H, budget K, training observations (q, y), validation split.
Ensure: Fixed partition $\widehat{\mathcal{P}} = \{I_1, \dots, I_K\}$ with boundaries $0 = h_0 < h_1 < \cdots < h_K = H$.
1: Compute $\widehat{\Sigma} = \mathrm{Cov}_{\mathrm{TRAIN}}(q)$ and fit $\widehat{C} = \arg\min_C \|y - Cq\|^2_{\mathrm{TRAIN}} + \lambda\|C\|^2_F$ (λ tuned on VAL); set $\widehat{M} = \widehat{C}^\top \widehat{C}$.
Stage 1: Surrogate Dynamic Programming Initialization
2: Tabulate interval costs $\mathrm{tr}(\widehat{M}_{II} R_I)$ ($O(H^4)$ arithmetic, $O(H^2)$ storage); solve the Bellman DP $\mathcal{P}_0 \leftarrow \arg\min_{\mathcal{P}} \widetilde{J}(\mathcal{P}; \widehat{\Sigma}, \widehat{M})$ in $O(KH^2)$.
Stage 2: Deterministic Local Exact-Risk Refinement
3: Initialize partition $\mathcal{P} \leftarrow \mathcal{P}_0$.
4: **repeat**
5: For each internal boundary $k \in \{1, \dots, K-1\}$ and perturbation $\delta \in \{-1, +1\}$ with $|I'_k| \geq 1, |I'_{k+1}| \geq 1$:
6: Evaluate candidate exact risk $J(\mathcal{P}'; \widehat{\Sigma}, \widehat{M})$ in $O(K^3)$ via 2D integral-image prefix sums.
7: Adopt shift with the largest strict decrease in J; break ties lexicographically.
8: **until** no single-boundary perturbation strictly decreases J
9: **return** $\widehat{\mathcal{P}} \leftarrow \mathcal{P}$.

A logarithmic grid is therefore asymptotically optimal only when task roughness decays as x^{-3}; it encodes a specific assumption about the task, not a generic consequence of multiscale structure. Theorem 2 and Corollary 1 are $K \to \infty$ statements about limiting boundary densities, whereas our experiments use finite budgets; §5.4 tests whether the condition holds on the estimated task geometries, and we do not identify a finite budget behavior with the limit.

4 Task-aware temporal coarsening (TATC): estimation and optimization

Task-aware temporal coarsening (TATC) derives a static temporal partition $\mathcal{P} = \{I_1, \dots, I_K\}$ of H observation lags into $K \ll H$ contiguous intervals prior to downstream training, without conditioning on any specific forecasting backbone. While Sections 2 and 3 assume exact population matrices (Σ, M), practical forecasting estimates both operators on development data alone. For standardized histories $q \in \mathbb{R}^H$ and targets $y \in \mathbb{R}^S$ over horizon S, sample covariance $\widehat{\Sigma} \in \mathbb{R}^{H \times H}$ and pilot mapping $\widehat{C} \in \mathbb{R}^{S \times H}$ are fit on the training split:

$$\widehat{\Sigma} = \mathrm{Cov}_{\mathrm{TRAIN}}(q), \qquad \widehat{C} = \arg\min_{C \in \mathbb{R}^{S \times H}} \|y - Cq\|^2_{\mathrm{TRAIN}} + \lambda\|C\|^2_F, \qquad \widehat{M} = \widehat{C}^\top \widehat{C} \in \mathbb{R}^{H \times H}, \tag{7}$$

where the ridge penalty $\lambda > 0$ is selected by validation grid search identically across all arms. No test data enters estimation or hyperparameter selection (splits detailed in §5 and Table 11). The pilot matrix $\widehat{M}$ provides an exact second-order metric for the linear task signal: for any aggregation $P_{\mathcal{P}} \in \mathbb{R}^{K \times H}$, optimal linear reconstruction satisfies $\min_{W \in \mathbb{R}^{S \times K}} \mathbb{E}[\|\widehat{C}q - W P_{\mathcal{P}} q\|_2^2] = J(\mathcal{P}; \widehat{\Sigma}, \widehat{M})$ with decoder W. This identity is exact for reconstructing the fitted pilot signal $\widehat{C}q$, not the true target y unless y is itself a linear function of q; $\widehat{M}$ is therefore a linear proxy for the forecasting task, and every claim about J inherits that substitution (§5.3). All quantities are accumulated from single-pass streamed sufficient statistics, so they are exact rather than subsampled.

Partition optimization operates in two deterministic stages (Algorithm 1). A dynamic program returns the global optimum $\mathcal{P}_0$ of the block-additive surrogate $\widetilde{J}$ in $O(KH^2)$ operations after interval-cost tabulation; deterministic unit-step descent on exact J, evaluated in $O(K^3)$ by integral-image rectangle sums over $\widehat{\Sigma}$ and $\widehat{\Sigma}\widehat{M}\widehat{\Sigma}$, then refines it until no single-boundary move strictly decreases J. There are no random restarts, no tuned search budget, and no downstream error in the loop, so the partition is fixed before any reader is trained. $\mathcal{P}_0$ is the exact global optimum of $\widetilde{J}$; $\widehat{\mathcal{P}}$ is best-found under J and is never called an optimum of it. We know no tractable globally exact optimizer for J under correlated Σ; exhaustive enumeration, which gate G2 uses at small H, is infeasible at $H = 512$.

5 EXPERIMENTAL DESIGN AND RESULTS

The empirical evaluation tests three standard benchmarks: ETTm1, Weather, and Electricity. We use standard chronological splits with $70\%/10\%/20\%$ partitions for Weather and Electricity, and $12/4/4$-month blocks for ETTm1, applying a one-timestep purge at interior boundaries (exact index cutoffs and containment rules appear in Table 11). Across all runs, histories span $H = 512$ steps, forecast horizons cover $S \in \{96, 192, 336, 720\}$, and series are processed channel-independently. The linear bridge examines six allocation arms across budgets $K \in \{8, 16, 32\}$ (Table 1), while neural transfer focuses on $K = 16$ using three arms over three random seeds. Across every comparison, the placement of the K measurement intervals is the only factor that differs between arms.

Table 1: Allocation arms: input information, geometric formulation, and partition derivation method.

Arm	Geometric formulation	Derivation protocol
uniform	Uniform temporal spacing (fixed-resolution baseline)	Analytical construction
log	Geometric spacing across observation lags	Analytical construction
input_only	Input variance quantiles via $A_{\text{input}}^{1/3}$	Analytical construction
task_only	$(I, \widehat{M})$: task geometry with identity covariance	Surrogate DP + exact-J refinement
cov_only	$(\widehat{\Sigma}, I)$: history covariance with isotropic task weight	Surrogate DP + exact-J refinement
task_aware	$(\widehat{\Sigma}, \widehat{M})$: joint history and task geometry (TATC)	Surrogate DP + exact-J refinement

So that a comparison between objectives does not become a comparison between search procedures, task_only, cov_only, and task_aware receive the same procedure — surrogate DP proposal followed by deterministic local descent on exact J — each on its own geometry. This prevents task_aware from receiving a stronger search than its comparators; it does not equalize the optimization gap each geometry leaves, which remains a confound (§6). The remaining three arms, uniform, log, and input_only, are analytical reference partitions that do not involve numerical optimization.

All empirical quantities adhere to strict split separation. We fit the covariance estimator $\widehat{\Sigma}$ and the linear pilot $\widehat{C}$ exclusively on the training data, selecting the pilot ridge scalar on the validation data using the same grid for all arms. Once determined, temporal partitions are completely frozen. The test split is held out solely for final evaluation and never informs partition design, regularization penalties, epoch counts, reader architecture, or comparison arms. The arms, budgets, horizons, reader protocol, and the primary ETTm1 comparison were fixed before TEST was evaluated, and the TEST evaluation covers all 324 resulting cells (216 linear-reader and 108 neural-reader); the substitution and path audits of Appendices C.1 and C.2 were added afterward and use VAL only. To prevent silent leakage or software errors, 15 pre-outcome integrity checks are run on synthetic and benchmark data before reading the results, including 3 checks against theoretically proven invariants (Appendix C).

5.1 SURROGATE FIDELITY AND EXACT-RISK REFINEMENT

To assess whether optimizing the surrogate introduces substantial error, we compare the block-additive surrogate $\widetilde{J}$ directly against the exact risk J over the fixed development geometry $(\widehat{\Sigma}, \widehat{M})$. Because this comparison evaluates two analytical functions on identical matrices, it operates independently of downstream predictors and test observations. All geometry matrices were estimated on the training data, with ridge penalties tuned on the validation data prior to partition selection. In each cell, we sample a contiguous K-Partition uniformly over all valid compositions of H. Comparing objective values across these neutral candidates reveals how closely surrogate rankings track exact risk under realistic covariance structures without confounding from search heuristics. This is a post-hoc diagnostic, fully specified before any audit partition was drawn, and it uses no TEST-dependent quantity.

Surrogate fidelity varies substantially across benchmarks (Figure 1). At a budget of $K = 16$, rank agreement across all four forecast horizons stays consistently high on Weather ($+0.953$ to $+0.976$), moderate on ETTm1 ($+0.618$ to $+0.886$), and virtually non-existent on Electricity (-0.081 to $+0.015$, Table 10). The absence of signal persists in the low-risk tail where an optimizer operates: among the 100 sampled partitions with the lowest exact risk in each of the twelve Electricity cells

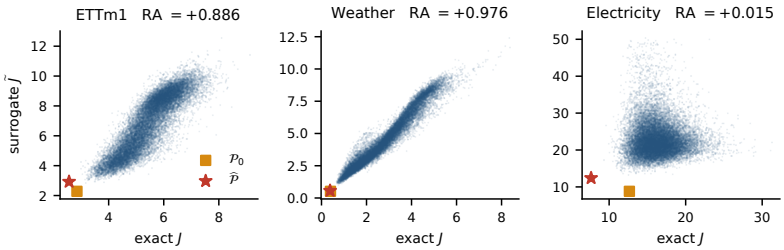

Figure 1: Exact J against the block-additive surrogate on 20,000 uniformly drawn partitions, $K = 16$, $S = 96$; no reader and no TEST. The same substitution that is nearly exact on Weather is uninformative on Electricity. Squares mark the surrogate's DP optimum $\mathcal{P}_0$, stars the refined $\widehat{\mathcal{P}}$.

(all three budgets), $\widetilde{J}$ reverses the exact ordering on 44.6% to 54.2% of the $\binom{100}{2}$ pairs, against 50% for an uninformative ordering.

Within each cell's 23,041-candidate audit pool (Appendix C), let $P_\star$ denote the partition with the lowest observed exact risk, serving as an empirical benchmark rather than a certified global optimum. Across all 36 evaluation cells, the initial surrogate optimum $\mathcal{P}_0$ exhibits substantial excess exact risk, $G = [J(\mathcal{P}_0) - J(P_\star)]/|J(P_\star)|$, with a median gap of 27.5% and peaks reaching 149.6%. Because dynamic programming solves $\widetilde{J}$ to global optimality, this gap reflects the discrepancy between the surrogate objective and the true risk function. Local exact-risk refinement successfully bridges this gap, recovering a median fraction $R = [J(\mathcal{P}_0) - J(\widehat{\mathcal{P}})]/[J(\mathcal{P}_0) - J(P_\star)]$ of 89.7%, providing empirical support for the two-stage optimization architecture.

Refinement lowers exact risk in all 36 cells, but $\widehat{\mathcal{P}}$ is the best of 32 random-start local minima in only 17 of them: 1 of 12 at $K = 8$, 4 of 12 at $K = 16$ and all 12 at $K = 32$. The shortfall thus concentrates on small budgets, where wide bands require long runs of improving unit moves, and any non-monotone stretch halts the descent. Lower exact risk also does not guarantee better forecasts: in a post-hoc VAL audit, substituting lower-J partitions into the unchanged neural reader improves validation error in six of thirteen cases and degrades it in seven, and only two of the thirteen changes exceed the three-seed range of both partitions, both of them increase (Appendix C.1).

5.2 Held-out validation of exact and surrogate risk

To verify whether the analytical risk accurately reflects generalization performance, we evaluate how well the exact risk J orders representations relative to the held-out test error. We fit a linear ridge head directly on the K coarsened measurements, evaluating performance within the linear model family that $\widehat{M}$ describes. We quantify ranking consistency using tie-aware rank agreement, defined as the Spearman rank correlation $\mathrm{RA}(x, y) = \mathrm{Corr}(r(x), r(y))$ where $r(\cdot)$ assigns average ranks to tied values. Across the twelve dataset-horizon configurations at each budget K (detailed in Table 9), rank agreement between exact risk J and held-out test error averages $+0.897$, $+0.980$, and $+0.775$ for $K = 8, 16, 32$. In contrast, the surrogate objective $\widetilde{J}$ evaluated on the identical representations and geometries achieves rank agreements of only $+0.066$, $+0.362$, and $+0.553$. These statistics measure agreement over the designed candidate set rather than population correlation.

Comparing objectives within each budget reveals that the exact risk J and the surrogate $\widetilde{J}$ diverge substantially in how accurately they order representations relative to held-out test error. Both objectives share the identical estimated covariance and task geometry, evaluating the same candidate partitions. They differ solely because the surrogate truncates cross-block covariance. This truncation can materially alter partition rankings: at $K = 16$ the two objectives order the same six partitions very differently in most ETTm1 and Electricity cells, while on Weather they agree exactly (Table 9). This is why the algorithm of §4 refines on exact J rather than stopping at the dynamic program.

Because exact risk J serves as an in-class linear proxy, this candidate family is not fully neutral: task_aware is directly optimized under the objective being ranked, whereas the surrogate proposal $\mathcal{P}_0$ is not included as an independent arm. Consequently, this evaluation provides a controlled

Table 2: Neural TEST MSE, $K = 16$, averaged over four horizons and three seeds. Only columns are comparable within the table (not against full-resolution forecasters); details in Appendix C.

dataset	uniform	log	task_aware	TA vs uniform	TA vs log
ETTm1	0.5532	**0.4624**	0.4748	-14.2%	$+2.7\%$
Weather	0.2632	**0.2422**	0.2451	-6.9%	$+1.2\%$
Electricity	0.3193	0.2414	**0.2195**	-31.3%	-9.1%

comparison on a fixed candidate set, confirming with an independent reader what §5.1 establishes across neutral partition distributions. We view these ranking agreements as empirical confirmation rather than primary structural evidence.

5.3 Transfer to a neural reader

To test whether partitions derived from linear pilot models generalize to nonlinear predictors, we evaluate coarsened representations using an independently trained Transformer reader. At a budget of $K = 16$, each token represents one interval, encoding its mean observation value, normalized temporal midpoint, and log relative width. These tokens pass into a two-block pre-LayerNorm Transformer encoder (Appendix C). Downstream models operate exclusively on the K compressed tokens without access to the full sequence, ensuring forecasting reflects representation quality rather than architectural workarounds. This experiment evaluates representation transfer to an independent family of nonlinear models. Because the reader sees only K band tokens and never the full-resolution history, its absolute errors are higher than those of full-resolution forecasters by design; only comparisons between arms are meaningful (Table 2).

Before any TEST value was computed, we fixed a single pass/fail criterion on ETTm1 at $K = 16$: averaged over four horizons and three seeds, task_aware $\leq \min(\text{uniform}, \text{log})$ in TEST MSE. As Table 2 shows, logarithmic allocation attains the lowest error, so this criterion, fixed before TEST evaluation, is not met. We report the result without changing the representation budget, splits, estimator, reader, or arms. While task_aware outperforms uniform allocation across all four individual horizons, it trails logarithmic allocation on the horizon average, achieving lower error at $S = 96$ and $S = 720$ but falling behind by larger margins at $S = 192$ and $S = 336$ (detailed per-horizon standard deviations appear in Appendix C).

task_aware has a lower mean TEST MSE than uniform in all twelve cells: all three matched seeds are lower in eleven, while at Weather $S = 720$ it is lower in two of three seeds and only 0.046% lower in mean. However, relative to logarithmic spacing, performance is dataset-dependent. Interpreting Table 2 requires separating two distinct effects: the transition from uniform to a multiscale prior, and the subsequent transition to an estimated task-aware allocation. Replacing uniform spacing with a fixed logarithmic grid reduces forecasting error by 16.4% on ETTm1, 8.0% on Weather, and 24.4% on Electricity, showing that much of the margin over uniform allocation stems from multiscale coarsening itself without requiring task data. Moving from this logarithmic baseline to TATC yields an additional 9.1% error reduction on Electricity, but slight increases on ETTm1 ($+2.7\%$) and Weather ($+1.2\%$). This comparison evaluates final representations rather than a mechanistic decomposition into prior and adaptive terms, underscoring the empirical strength of heuristic logarithmic grids. The ordering between allocations can also depend on the reader. On Weather at $K = 16$, logarithmic allocation has a higher held-out TEST error than uniform under the linear reader at all four horizons ($+1.68\%$ to $+4.06\%$) but lower error under the neural reader at all four (-4.79% to -11.12%), with identical splits and partitions. Furthermore, along a controlled interpolation path between logarithmic and task-aware allocations on validation data, exact risk tracks neural reader error closely on Electricity, remains largely decoupled on Weather, and shows irregular behavior on ETTm1 (Appendix C.2). Across all neural training runs, checkpoints were selected strictly by validation error over 15 complete epochs without early stopping (Appendix C).

5.4 High-resolution theory and finite-budget logarithmic spacing

Corollary 1 establishes that logarithmic spacing achieves asymptotic optimality if and only if task roughness decays cubically with lag, $A(x) \propto x^{-3}$. Evaluating the empirical ratio $R_{\log}$ across

the twelve task geometries $(\widehat{M})$ reveals that this condition fails decisively across all benchmarks (Appendix C). Values of $R_{\log,\mathrm{HR}}$ range from 9.65 to 16.03 rather than the theoretical target of 1, with fitted log-log slopes between $+0.08$ and -0.73 rather than -3. A flat roughness profile yields a reference value of 15.52. Empirical roughness profiles also diverge across domains: ETTm1 matches a flat profile, while Electricity exhibits a distinct structure. Asymptotic theory, therefore, cannot explain the finite-budget effectiveness of logarithmic spacing, leaving its empirical strength open. Internal boundary locations are also weakly identified by downstream loss: re-estimating (Sh, Mh) on contiguous sub-blocks of ETTm1 yields large boundary shifts with negligible change in linear reader validation error, confirming predictive non-uniqueness within a partition family (Appendix C.3).

6 Discussion

Tensions among these empirical findings are resolved by decomposing the pipeline from surrogate estimation to downstream forecasting into three stages: an approximation gap, an optimization gap, and a transfer gap. Approximation appears twice: in the reduction to $\widetilde{J}$, and in the continuum reduction $\rho^\star \propto A^{1/3}$, whose condition fails on all three benchmarks and so cannot explain where logarithmic spacing is strong. Optimization introduces a search gap governed by coordinate descent, which degrades at smaller budgets K. Transfer represents the gap between exact linear risk and downstream reader loss, where lower analytical risk does not always yield lower neural forecasting error (§5.1, §5.4, §5.3, §C.3). Rather than aligning on a single axis, benchmarks differ across these stages. On Weather, $\widetilde{J} \approx J$ and at $K = 16$ the refined partition sits a median 0.3% above best-found, yet logarithmic spacing still wins at two horizons. On Electricity, $\widetilde{J}$ carries almost no ordering signal, yet the refined allocation improves clearly on both uniform and logarithmic spacing. ETTm1 lies between surrogate fidelity; at $K = 16$, its refined partition sits at a median $\approx 7\%$ above best-found, as on Electricity (median 7.0%). Because these three stages can fail independently, evaluating coarsened representations requires auditing each isolation gap.

Several methodological boundaries delineate the scope of this study. First, optimization fidelity remains an active factor: the refined partition $\widehat{\mathcal{P}}$ stays a median of 3.0% above the best-found candidate, meaning local search limitations represent an alternative explanation wherever downstream margins are small (Appendix C.1). Second, the empirical evaluation covers three benchmarks, one Transformer architecture, three random seeds, and a designed six-arm family, providing descriptive comparisons rather than population inferences. Third, the task metric $\widehat{M}$ is derived from a linear pilot model, so the exact analytical risk inherits the limitations of this linear proxy, with partition stability (§C.3) evaluated exclusively on ETTm1. We also deliberately do not evaluate integration into complex end-to-end patch architectures, as patching simultaneously alters token span, content overlap, and model capacity. Finally, all models employ channel-shared pilots and readers, meaning $\widehat{M}$ reflects a single shared task geometry across channels rather than channel-specific dynamics.

7 Conclusion

Task-aware temporal coarsening allocates a fixed budget of contiguous temporal measurements before any downstream model is trained, using history covariance and task geometry estimated on development data only. Under the linear reader whose risk J describes, the task-aware allocation beats both uniform and logarithmic spacing in all 36 cells. Under a Transformer reader, it beats uniform spacing in all twelve datasets–horizon cells, one negligibly, but does not dominate a fixed logarithmic grid, and the ETTm1 criterion fixed before TEST evaluation was not met; the high-resolution theory does not explain the finite-budget strength of the logarithmic grid. Separating the approximation, optimization, and transfer gaps shows that each can fail on its own, so attributing a downstream outcome to any one of them requires checking the others. When an estimated allocation is worthwhile remains the central open finite-budget question.

AI USE STATEMENT

The authors directed the work and made all final scientific decisions. Generative AI tools assisted with developing and stress-testing the conceptual framing and hypotheses; formulating and checking mathematical claims and derivations; designing experiments and protocol safeguards; implementing, testing, and auditing code; analyzing and interpreting results; searching and synthesizing literature; and drafting and editing the manuscript. All AI-assisted material was reviewed, tested where applicable, and corrected by the authors. The authors take full responsibility for the final text, claims, code, and artifacts.

ETHICS STATEMENT

This research focuses on foundational mathematical and algorithmic principles of temporal representation and coarsening in time series forecasting. The empirical evaluations rely exclusively on standard, publicly available benchmark datasets (ETTm1, Weather, Electricity) that contain no personally identifiable information, human subjects, or proprietary data, and comply with their original licenses. Temporal coarsening reduces the number of tokens a downstream model reads, which can lower its computational and memory cost; we do not measure compute or energy use in this work. In real-world infrastructure deployments such as power grid operations or severe weather prediction, practitioners should ensure that temporal coarsening does not discard high-frequency anomalies or extreme transient events critical to operational safety.

REPRODUCIBILITY STATEMENT

Appendix A states the assumptions and provides complete proofs; Appendix C records the estimator, data splits, model, and training configuration, audit pools, integrity checks, and per-cell results. Geometry, penalties, partitions, epochs, and checkpoints were fixed using TRAIN and VAL only; TEST was used only for final scoring over the 324 held-out cells. An integrity check detected a development split-off-by-one before TEST evaluation; after correction, every affected experiment was rerun, and only corrected results are reported. The same code snapshot is provided in the submitted supplementary archive and at `https://anonymous.4open.science/r/review-code-9422/`; it contains the implementation, environment specification, data-acquisition and checksum instructions, frozen partitions, canonical result tables, and scripts to verify the reported tables and figures or rerun the principal empirical pipelines. The criterion fixed before the TEST evaluation failed, and other negative findings are reported in the main text.

REFERENCES

H. Fareed Ahmed, Toktam Samiei, and Erfan Nozari. On the optimal temporal resolution for information representation in neural activity: a theoretical analysis. *Frontiers in Computational Neuroscience*, 20, 2026. doi: 10.3389/fncom.2026.1885975.

Sravan Kumar Ankireddy, Nikita Seleznev, Nam H. Nguyen, Yulun Wu, Senthil Kumar, Furong Huang, and C. Bayan Bruss. TimeSqueeze: Dynamic patching for efficient time series forecasting. *arXiv:2603.11352*, 2026.

Ahmed Attia, Alen Alexanderian, and Arvind K. Saibaba. Goal-oriented optimal design of experiments for large-scale Bayesian linear inverse problems. *Inverse Problems*, 34(9):095009, 2018. doi: 10.1088/1361-6420/aad210.

D. L. Barrow and P. W. Smith. Asymptotic properties of best $L_2[0, 1]$ approximation by splines with variable knots. *Quarterly of Applied Mathematics*, 36(3):293–304, 1978. doi: 10.1090/qam/508773.

Richard Bellman. On the approximation of curves by line segments using dynamic programming. *Communications of the ACM*, 4(6):284, 1961. doi: 10.1145/366573.366611.

W. R. Bennett. Spectra of quantized signals. *Bell System Technical Journal*, 27(3):446–472, 1948. doi: 10.1002/j.1538-7305.1948.tb01340.x.

Jun Chen, Yong Fang, Ashish Khisti, Ayfer Özgür, and Nir Shlezinger. Information compression in the AI era: recent advances and future challenges. *IEEE Journal on Selected Areas in Communications*, 43(7):2333–2348, 2025. doi: 10.1109/jsac.2025.3560359.

Peng Chen, Yingying Zhang, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang, and Chenjuan Guo. Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2402.05956.

John S. Chipman and Peter Winker. Optimal aggregation by threshold accepting: An application to the German industrial classification system. Diskussionsbeiträge, Serie II 180, Universität Konstanz, Sonderforschungsbereich 178, 1992. URL https://hdl.handle.net/10419/101634.

John S. Chipman and Peter Winker. Optimal aggregation of linear time series models. *Computational Statistics & Data Analysis*, 49(2):311–331, 2005. doi: 10.1016/j.csda.2004.05.015.

Zihang Dai, Guokun Lai, Yiming Yang, and Quoc V. Le. Funnel-transformer: Filtering out sequential redundancy for efficient language processing. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pp. 4271–4282, 2020.

Kuiye Ding, Fanda Fan, Chunyi Hou, Zheya Wang, Lei Wang, Zhengxin Yang, and Jianfeng Zhan. TimeMosaic: Temporal heterogeneity guided time series forecasting via adaptive granularity patch and segment-wise decoding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40 (25):20790–20798, 2026. doi: 10.1609/aaai.v40i25.39218. URL https://ojs.aaai.org/index.php/AAAI/article/view/39218.

Luca-Andrei Fechete, Mohamed Sana, Fadhel Ayed, Nicola Piovesan, Wenjie Li, Antonio De Domenico, and Tareq Si Salem. Goal-oriented time-series forecasting: Foundation framework design. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(25):21065–21073, 2026. doi: 10.1609/aaai.v40i25.39249. URL https://ojs.aaai.org/index.php/AAAI/article/view/39249.

Walter D. Fisher. On grouping for maximum homogeneity. *Journal of the American Statistical Association*, 53(284):789–798, 1958. doi: 10.1080/01621459.1958.10501479.

Walter D. Fisher. Optimal aggregation in multi-equation prediction models. *Econometrica*, 30(4): 744–769, 1962. doi: 10.2307/1909323.

Leon Götz, Marcel Kollovieh, Stephan Günnemann, and Leo Schwinn. Efficient time series processing for transformers and state-space models through token merging. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 20226–20246, 2025. arXiv:2405.17951.

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. HiPPO: Recurrent memory with optimal polynomial projections. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Brad Jackson, Jeffrey D. Scargle, David Barnes, Sundararajan Arabhi, Alina Alt, Peter Gioumousis, Elyus Gwin, Paungkaew Sangtrakulcharoen, Linda Tan, and Tun Tao Tsai. An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters*, 12(2):105–108, 2005. doi: 10.1109/LSP.2001.838216.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 4651–4664, 2021.

Hao Ji and Hans-Georg Müller. Optimal designs for longitudinal and functional data. *Journal of the Royal Statistical Society: Series B*, 79(3):859–876, 2017. doi: 10.1111/rssb.12192.

F. Laisney. Theory and practice in optimal aggregation of linear models. *Economics Letters*, 15(3–4): 315–324, 1984. doi: 10.1016/0165-1765(84)90119-8.

Xvyuan Liu, Xiangfei Qiu, Xingjian Wu, Zhengyu Li, Chenjuan Guo, Jilin Hu, and Bin Yang. Rethinking irregular time series forecasting: A simple yet effective baseline. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(28):23873–23881, 2026. doi: 10.1609/aaai.v40i28.39563. URL https://ojs.aaai.org/index.php/AAAI/article/view/39563.

Helmuth Naumer and Farzad Kamalabadi. Dimensionality collapse: Optimal measurement selection for low-error infinite-horizon forecasting. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 206 of *Proceedings of Machine Learning Research*, pp. 6166–6198, 2023.

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

Artidoro Pagnoni, Ram Pasunuru, Pedro Rodriguez, John Nguyen, Benjamin Muller, Margaret Li, Chunting Zhou, Lili Yu, Jason Weston, Luke Zettlemoyer, Gargi Ghosh, Mike Lewis, Ari Holtzman, and Srinivasan Iyer. Byte latent transformer: Patches scale better than tokens. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 9238–9258, 2025. doi: 10.18653/v1/2025.acl-long.453.

P. F. Panter and W. Dite. Quantization distortion in pulse-count modulation with nonuniform spacing of levels. *Proceedings of the IRE*, 39(1):44–48, 1951. doi: 10.1109/JRPROC.1951.230419.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:1911.05507.

Hyungmin Rha, Ming-Hung Kao, and Rong Pan. Design optimal sampling plans for functional regression models. *Computational Statistics & Data Analysis*, 146:106925, 2020. doi: 10.1016/j.csda.2020.106925.

Nir Shlezinger and Yonina C. Eldar. Deep task-based quantization. *Entropy*, 23(1):104, 2021. arXiv:1908.06845.

Nir Shlezinger, Yonina C. Eldar, and Miguel R. D. Rodrigues. Hardware-limited task-based quantization. *arXiv preprint arXiv:1807.08305*, 2019.

Andrea Silvestrini and David Veredas. Temporal aggregation of univariate and multivariate time series models: a survey. *Journal of Economic Surveys*, 22(3):458–497, 2008. doi: 10.1111/j.1467-6419.2007.00538.x.

René Victor Valqui Vidal. Optimal partition of an interval — the discrete version. In *Applied Simulated Annealing*, Lecture Notes in Economics and Mathematical Systems, pp. 291–312. Springer, 1993. doi: 10.1007/978-3-642-46787-5_15.

Yihang Wang, Yuying Qiu, Peng Chen, Yang Shu, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. LightGTS: A lightweight general time series forecasting model. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 64109–64126, 2025. arXiv:2506.06005.

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

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021. doi: 10.1609/aaai.v35i12.17325.

Federico Zucchi, Yi Xie, Chao Zhang, Keyuan Luo, Thomas Lampert, and Ziyue Li. Adaptive patching is harder than it looks for time-series forecasting. *arXiv:2606.04074*, 2026.

A PROOFS

Every result referenced in §2–§3 is stated and proved here, with its assumptions intact. **This appendix is self-contained**: no proof defers to another document. Results are referred to by the numbers they carry in the main text.

Throughout, $\mathcal{P} = \{I_1, \dots, I_K\}$ is a contiguous partition of the H ordered lags, $P_{\mathcal{P}}$ its aggregation matrix with rows u_k supported on I_k, $\mathcal{R}(\mathcal{P}; \Sigma) = \Sigma - \Sigma P_{\mathcal{P}}^\top (P_{\mathcal{P}} \Sigma P_{\mathcal{P}}^\top)^{-1} P_{\mathcal{P}} \Sigma$, $B_{\mathcal{P}}(\Sigma)$ is Σ with every entry joining two different bands set to zero, $R_I = \Sigma_{II} - \Sigma_{II} u_I u_I^\top \Sigma_{II} / (u_I^\top \Sigma_{II} u_I)$, and $\Delta_{\mathcal{P}} = J(\mathcal{P}; \Sigma, M) - \widetilde{J}(\mathcal{P}; \Sigma, M)$.

A.1 FORMULATION

Lemma 1 (Trace form and reparameterization invariance)**.** *The exact task-weighted residual risk admits the trace representation:*

$$J(\mathcal{P}; \Sigma, M) \;=\; \operatorname{tr}(M\Sigma) - \operatorname{tr}\left[(P_{\mathcal{P}} \Sigma P_{\mathcal{P}}^\top)^{-1} P_{\mathcal{P}} \Sigma M \Sigma P_{\mathcal{P}}^\top\right],$$

and for any invertible $B \in \mathbb{R}^{K \times K}$, $\mathcal{R}(BP_{\mathcal{P}}; \Sigma) = \mathcal{R}(P_{\mathcal{P}}; \Sigma)$, hence $J(BP_{\mathcal{P}}; \Sigma, M) = J(P_{\mathcal{P}}; \Sigma, M)$.

Proof. The first identity follows from the covariance of the optimal linear projection residual and the cyclic invariance of the trace. For the second, substitute $BP_{\mathcal{P}}$ and use $(BP_{\mathcal{P}} \Sigma P_{\mathcal{P}}^\top B^\top)^{-1} = B^{-\top} (P_{\mathcal{P}} \Sigma P_{\mathcal{P}}^\top)^{-1} B^{-1}$; the B and B^{-1} factors cancel. □

This is what licenses reading band sums and band means as the same representation, and it is used wherever the paper switches between them.

Lemma 2 (Exact additive interval cost on a diagonal history)**.** *If $\Sigma = D$ is diagonal with entries $d_j > 0$ then $J(\mathcal{P}; D, M) = \sum_k C_D(I_k)$ with $C_D(I) = \operatorname{tr}(M_{II} D_I) - d_I^\top M_{II} d_I / \sum_{j \in I} d_j$, where d_I is the vector of d_j on I.*

Proof. The bands are disjoint, so $P_{\mathcal{P}} D P_{\mathcal{P}}^\top$ is diagonal with entries $\sum_{j \in I} d_j$. Hence the matrix $D P_{\mathcal{P}}^\top (P_{\mathcal{P}} D P_{\mathcal{P}}^\top)^{-1} P_{\mathcal{P}} D$ is block diagonal with blocks $d_I d_I^\top / \sum_{j \in I} d_j$, and the remaining trace against M splits over bands. □

Additivity is what makes the problem exactly solvable: with $D(k, j)$ the optimal cost of covering the first j coordinates with k bands, $D(k, j) = \min_{i<j} \{D(k-1, i) + C_D(\{i+1, \dots, j\})\}$, evaluated in $O(KH^2)$ once the interval costs are tabulated. The recurrence is classical (Fisher, 1958; Bellman, 1961); what Lemma 2 supplies the block cost. For a scalar task $M = ww^\top$, the band cost specializes to $\sum_{j \in I} d_j (w_j - \bar{w}_{I,d})^2 \geq 0$ with $\bar{w}_{I,d} = \sum_{j \in I} d_j w_j / \sum_{j \in I} d_j$, representing the variance-weighted within-band spread of the task loadings the band fails to resolve, and positivity is immediate. For general $M \succeq 0$, the band cost remains nonnegative, $C_D(I) = \operatorname{tr}(M_{II} R_I) \geq 0$, but need not admit this scalar within-group-variance form.

A.2 WHAT THE SURROGATE IS

Proof of Proposition 1. Since $B_{\mathcal{P}}(\Sigma)$ has no entry joining two different blocks, $P_{\mathcal{P}} B_{\mathcal{P}}(\Sigma) P_{\mathcal{P}}^\top$ is diagonal with entries $u_k^\top \Sigma_{I_k I_k} u_k$, and column k of $B_{\mathcal{P}}(\Sigma) P_{\mathcal{P}}^\top$ is supported on I_k. Hence $B_{\mathcal{P}}(\Sigma) P_{\mathcal{P}}^\top (P_{\mathcal{P}} B_{\mathcal{P}}(\Sigma) P_{\mathcal{P}}^\top)^{-1} P_{\mathcal{P}} B_{\mathcal{P}}(\Sigma)$ is a sum of outer products each supported on $I_k \times I_k$, so $\mathcal{R}(\mathcal{P}; B_{\mathcal{P}}(\Sigma))$ is block diagonal with blocks R_{I_k}. A trace against a block-diagonal matrix reads only the diagonal blocks $M_{I_k I_k}$, giving $\sum_k \operatorname{tr}(M_{I_k I_k} R_{I_k}) = \widetilde{J}$. □

The argument goes through the residual and never through a task vector, which is why an arbitrary $M \succeq 0$ is available. The hypothesis actually used is $u_k^\top \Sigma_{I_k I_k} u_k > 0$ for each k, weaker than $\Sigma_{I_k I_k} \succ 0$. Neither contiguity nor uniform averaging enters.

Theorem 1 (Exact surrogate invariance under compound symmetry). *Let $\Sigma = \sigma^2 I + \lambda \mathbf{1}\mathbf{1}^\top \in \mathbb{R}^{H\times H}$ with variance parameter $\sigma^2 > 0$ and covariance parameter $\lambda \geq 0$, where I is the identity matrix and $\mathbf{1}$ is the all-ones vector. Let $P_\mathcal{P}$ denote uniform block averaging across intervals of width $w_k = |I_k|$. Then the residual error covariance satisfies*

$$\mathcal{R}(\mathcal{P}; \Sigma) = \mathrm{blockdiag}_{k=1}^{K} \left[\sigma^2 (I_{w_k} - w_k^{-1} \mathbf{1}\mathbf{1}^\top)\right], \tag{8}$$

where I_{w_k} is the $w_k \times w_k$ identity matrix. Consequently, $\Delta_\mathcal{P} = 0$ for every partition $\mathcal{P}$, every budget K, and every task metric $M \succeq 0$.

Proof of Theorem 1. Let $D_w = \mathrm{diag}(w_1, \ldots, w_K)$. Within a block of width w, uniform averaging gives $u_I = \mathbf{1}/w$, so $\Sigma_{II} u_I = (\sigma^2/w + \lambda)\mathbf{1}$ and $u_I^\top \Sigma_{II} u_I = \sigma^2/w + \lambda$; hence $R_I = \sigma^2(I_w - \mathbf{1}\mathbf{1}^\top/w)$ and λ *cancels outright*. For the full criterion, $P_\mathcal{P}\mathbf{1} = \mathbf{1}$ and $P_\mathcal{P}^\top w = \mathbf{1}$ give $P_\mathcal{P} \Sigma P_\mathcal{P}^\top = \sigma^2 D_w^{-1} + \lambda \mathbf{1}\mathbf{1}^\top$, and Sherman–Morrison cancels every $\mathbf{1}\mathbf{1}^\top$ term, leaving $\mathcal{R}(\mathcal{P}; \Sigma) = \mathrm{blockdiag}_k R_{I_k}$, which equals $\mathcal{R}(\mathcal{P}; B_\mathcal{P}(\Sigma))$. □

The cancellation uses $P_\mathcal{P}\mathbf{1} = \mathbf{1}$, $P_\mathcal{P}^\top w = \mathbf{1}$ and the within-block row being proportional to $\mathbf{1}_I$ (specifically uniform block averaging), so Theorem 1 does not inherit Proposition 1's generality. With a heterogeneous diagonal, $\Sigma_{II} u_I$ is no longer proportional to $\mathbf{1}$ and the argument is unavailable.

Proposition 2 (Fixed-coefficient decomposition of the gap). *For a scalar task $M = ww^\top$ write $U = P_\mathcal{P}^\top$ and $G = U^\top \Sigma U$; let z_0 minimize $(w - Uz)^\top B_\mathcal{P}(\Sigma)(w - Uz)$, let $z_1 = G^{-1} U^\top \Sigma w$, and set $e_0 = w - Uz_0$. Then $\Delta_\mathcal{P} = S_\mathcal{P} - R_\mathcal{P}$ with $S_\mathcal{P} = e_0^\top (\Sigma - B_\mathcal{P}(\Sigma)) e_0$ and $R_\mathcal{P} = (z_1 - z_0)^\top G (z_1 - z_0) \geq 0$. $S_\mathcal{P}$ **is signed.***

Proof of Proposition 2. Write $L(z) = (w - Uz)^\top \Sigma (w - Uz)$, a quadratic in z with Hessian $2G$ and minimizer z_1, so $L(z_0) - L(z_1) = (z_1 - z_0)^\top G (z_1 - z_0) = R_\mathcal{P} \geq 0$ since $G \succ 0$. By definition $L(z_1) = J(\mathcal{P}; \Sigma, ww^\top)$, while $L(z_0) = e_0^\top \Sigma e_0 = e_0^\top B_\mathcal{P}(\Sigma) e_0 + S_\mathcal{P} = \widetilde{J}(\mathcal{P}; \Sigma, ww^\top) + S_\mathcal{P}$, the middle equality using Proposition 1 and the optimality of z_0 under $B_\mathcal{P}(\Sigma)$. Subtracting gives $\Delta_\mathcal{P} = S_\mathcal{P} - R_\mathcal{P}$. □

$S_\mathcal{P}$ carries no sign, and neither does the gap. $R_\mathcal{P}$ is a quadratic form in $G \succ 0$ and is nonnegative. $S_\mathcal{P}$ is not: $\Sigma - B_\mathcal{P}(\Sigma)$ is the discarded cross-block part; its diagonal blocks are zero, so its trace is zero and it is indefinite unless it vanishes identically. A four-step counterexample settles it. Take $H = 4$, $\Sigma_{ij} = 2^{-|i-j|}$, $w = (1, -1, 1, -1)^\top$ and the two bands $\{1, 2\}, \{3, 4\}$. Then $S_\mathcal{P} = -\frac{1}{4}$, $R_\mathcal{P} = \frac{3}{20}$ and $\Delta_\mathcal{P} = -\frac{2}{5}$, which is $J = \frac{8}{5}$ against $\widetilde{J} = 2$: **truncation can make the surrogate *over*state the exact risk, not only understate it**. Neither the empirical audit nor any gate depends on a sign for $\Delta_\mathcal{P}$, and §5.1 samples partitions on both sides of zero.

The argument uses a scalar quadratic risk. For general $M \succeq 0$, both J and $\widetilde{J}$ are linear in M, so writing $M = \sum_r w_r w_r^\top$ gives $\Delta_\mathcal{P} = \sum_r S_\mathcal{P}^{(r)} - \sum_r R_\mathcal{P}^{(r)}$ with each $R_\mathcal{P}^{(r)} \geq 0$, which yields a sum of scalar-task gaps because z_0, z_1 and e_0 are per-component. No single-vector form is claimed for arbitrary M.

What none of these licenses. These are statements about the *value* of the two objectives at a given $\mathcal{P}$. Selecting a representation compares values across partitions, where $\Delta_\mathcal{P}$ enters only through $\Delta_\mathcal{P} - \Delta_\mathcal{Q}$: a gap constant across the partitions being compared, or affine in $\widetilde{J}(\mathcal{P})$ with positive slope, changes no ordering whatever its size. **Absolute approximation error is therefore the wrong quantity for ranking representations**, and we prove neither that Δ is near-constant nor that it is near-affine for any covariance class beyond Theorem 1, where it is identically zero. §5.2 reports what the ordering actually does on held-out data.

Theorem 2 (Asymptotic optimal partition density). *Let $A(x) > 0$ on $[a, c]$. For partitions generated by a positive continuously differentiable probability density ρ on $[a, c]$ with endpoints at its k/K quantiles,*

$$J_{\mathrm{partition}}(\rho, K) = \frac{1}{12K^2} \int_a^c A(x) \rho(x)^{-2}\, dx + o(K^{-2}), \tag{9}$$

which is globally minimized over such densities by

$$\rho^\star(x) = \frac{A(x)^{1/3}}{\int_a^c A(u)^{1/3}\,du}. \tag{10}$$

Proof of Theorem 2. On an interval of width Δ centered at x, Taylor expansion about the center and subtraction of the interval mean give, uniformly on $[a, c]$, $\sum_r \pi_r \int_I (f_r - \bar{f}_{r,I})^2 = A(x)\Delta^3/12 + O(\Delta^4)$. For density-quantile partitions $\Delta = [K\rho(x)]^{-1} + O(K^{-2})$ uniformly, so summing over K intervals convert the leading sum to the stated Riemann integral with accumulated remainder $O(K^{-3})$. Hölder's inequality gives $\left(\int A^{1/3}\right)^3 \le \int A/\rho^2$ for any normalized ρ, with equality exactly when $\rho^3 \propto A$. □

Two nearby boundary cases lie outside the stated theorem, which assumes $A > 0$ on a compact $[a, c]$ with $a > 0$. If A vanishes somewhere, the optimal density is interpreted on its active support, zero-roughness regions may receive vanishing density asymptotically, and extra conditions are needed to control maximum interval width there. If the domain is allowed to include the origin, the cube-root law remains local whenever the integrals exist, but a candidate density singular at zero requires a separate cutoff analysis.

Proof of Corollary 1. On $[a, c]$ with $a > 0$ the normalized logarithmic boundary density is $\rho_{\log}(x) = 1/(x \log(c/a))$. By Theorem 2 the optimal density is $\rho^\star \propto A^{1/3}$, so $\rho_{\log}$ and $\rho^\star$ agree almost everywhere iff $A^{1/3} \propto x^{-1}$, i.e. $A(x) = Cx^{-3}$. The inefficiency $R_{\log} = \int_a^c A\rho_{\log}^{-2} / \left(\int_a^c A^{1/3}\right)^3 \ge 1$ is the same Hölder inequality, with equality under the same condition. □

Remark 1 (Log grids are not scale-equivariant under dilation). Corollary 1 does *not* say that a fixed-K log grid survives physical dilation of the history. If $H_r \approx cr$ with geometric endpoints $h_m = H_r^{m/K}$, then for every $m < K$, $h_m/r \approx c^{m/K} r^{m/K-1} \to 0$, so **all internal rescaled endpoints collapse at the origin.** A fixed-log-resolution policy avoids that collapse by increasing K with $\log r$, a distinct policy from the fixed-K setting.

A.3 Correlated histories

The result is stated here in full, with its proof.

Microscopic fixed-lattice correlation. Take the stationary $\mathrm{AR}(1)$ lattice covariance $\Sigma_\phi(m, n) = \phi^{|m-n|}$ at fixed ϕ, so that the correlation spans $O(1)$ lattice sites and its physical length vanishes as resolution grows. Let g_0 be the block-mean residual of a task loading, $J_0 = \|g_0\|^2$ the isotropic objective on the same partition, $n_{\min}$ the minimum band occupancy, and for a finite vector g, let $\widetilde{g}$ denote its zero extension to $\mathbb{Z}$, and $\nabla\widetilde{g}$ is its first difference. We call a family *residually regular* if $\bar{\nu} = \sup_K n_{\min}\|\nabla\widetilde{g}_0\|^2/\|g_0\|^2 < \infty$ whenever $J_0 > 0$, which constrains the projection residual rather than the covariance.

Theorem 3 (Microscopic correlation multiplies a constant)**.** *Fix $\phi \in (-1, 1)$, let $\Lambda_\phi = (1+\phi)/(1-\phi)$, and consider a scalar task $M = ww^\top$ on the fixed lattice covariance Σ_ϕ. For any partition family with $J_0 > 0$, $n_{\min} \to \infty$ and $\bar{\nu} < \infty$,*

$$\left|\frac{J_\phi(w;\mathcal{P})}{\Lambda_\phi J_0(w;\mathcal{P})} - 1\right| \;\le\; \frac{1}{n_{\min}}\Big[\frac{2\bar{\nu}|\phi|}{(1-|\phi|)^2} + \frac{40\phi^2}{(1-|\phi|)^4}\Big] \;\xrightarrow[n_{\min}\to\infty]{}\; 0.$$

Consequently, on any class of partitions over which this convergence is uniform (requiring a common $\bar{\nu}$ bound and a common $n_{\min} \to \infty$), the leading-order problem is the isotropic problem times the constant Λ_ϕ, and the leading-order ordering of partitions is therefore unchanged.

Two scope conditions are part of the statement. **Uniformity is required**: the bound is a per-family asymptotic, and concluding anything about the *minimizer* over a class needs the remainder to vanish uniformly over that class, which is why it is hypothesized rather than inferred. Moreover, $\bar{\nu}$ is undefined when $J_0 = 0$; that case is the exactly representable one, where the partition already resolves the task loading, and both objectives vanish.

The budget K does not appear: the asymptotic parameter is band occupancy. The constant degrades like $(1-|\phi|)^{-4}$, so **the $|\phi| \to 1$ regime is not covered and no sharpness is claimed**; and $\Lambda_\phi < 1$ for $\phi < 0$, so microscopic correlation can lower distortion; stronger correlation does not necessarily degrade compression quality.

Proof of Theorem 3. (1) A symbol bound. Extended by zero, $Q_\phi(g) = (2\pi)^{-1} \int_{-\pi}^{\pi} s_\phi(\theta)|\hat{g}(\theta)|^2 d\theta$. With $D(\theta) = 1 + \phi^2 - 2\phi\cos\theta = (1-\phi)^2 + \phi|1-e^{i\theta}|^2 \geq (1-|\phi|)^2$,

$$s_\phi(\theta) - \Lambda_\phi = -\Lambda_\phi \frac{\phi|1-e^{i\theta}|^2}{D(\theta)}, \qquad \Lambda_\phi = s_\phi(0),$$

and since $|1-e^{i\theta}|$ is the symbol of the first difference, Parseval gives

$$\left|Q_\phi(g) - \Lambda_\phi \|g\|^2\right| \leq \Lambda_\phi C_\phi \|\nabla \tilde{g}\|^2, \qquad C_\phi = \frac{|\phi|}{(1-|\phi|)^2}. \tag{11}$$

(2) Upper bound. Testing the correlated problem with the isotropic minimizer c_0 and applying equation 11, $J_\phi \leq Q_\phi(g_0) \leq \Lambda_\phi J_0[1 + C_\phi \nu / n_{\min}]$, by the definition $\nu = n_{\min}\|\nabla \tilde{g}_0\|^2 / \|g_0\|^2$.

(3) Two elementary facts. A band-constant vector is rough only at $K+1$ places: if c takes value v_k on a block of size $n_k \geq n_{\min}$, then $\nabla \tilde{c}$ vanishes except at the $K-1$ interior boundaries and the two outer edges, so $\|\nabla \tilde{c}\|^2 \leq \sum_{k<K} 2(v_k^2 + v_{k+1}^2) + v_1^2 + v_K^2 \leq 5\sum_k v_k^2 \leq 5\|c\|^2 / n_{\min}$, using $\|c\|^2 = \sum_k n_k v_k^2 \geq n_{\min} \sum_k v_k^2$. And the correlated minimizer cannot move far: with $m_\phi = \min_\theta s_\phi = (1-|\phi|)/(1+|\phi|)$ and $M_\phi = \max_\theta s_\phi = (1+|\phi|)/(1-|\phi|)$ the form obeys $m_\phi \|g\|^2 \leq Q_\phi(g) \leq M_\phi \|g\|^2$. Let c_ϕ attain J_ϕ, put $c = c_\phi - c_0$ and $g_\phi = w - c_\phi = g_0 - c$. Orthogonality of g_0 to band constants gives $\|g_\phi\|^2 = J_0 + \|c\|^2$, so $m_\phi(J_0 + \|c\|^2) \leq J_\phi \leq Q_\phi(g_0) \leq M_\phi J_0$ and therefore $\|c\|^2 \leq (M_\phi/m_\phi - 1)J_0 = 4|\phi|(1-|\phi|)^{-2} J_0$.

(4) Lower bound. Apply equation 11 to g_ϕ and $\|\nabla \tilde{g}_\phi\|^2 \leq 2\|\nabla \tilde{g}_0\|^2 + 2\|\nabla \tilde{c}\|^2$: $J_\phi \geq \Lambda_\phi(J_0 + \|c\|^2) - 2\Lambda_\phi C_\phi \|\nabla \tilde{g}_0\|^2 - 2\Lambda_\phi C_\phi \|\nabla \tilde{c}\|^2$. Discard the nonnegative $\Lambda_\phi \|c\|^2$ and bound the last term by the two facts of (3) in turn: $2\Lambda_\phi C_\phi \|\nabla \tilde{c}\|^2 \leq 10 \Lambda_\phi C_\phi \|c\|^2 / n_{\min} \leq \Lambda_\phi J_0 \cdot 40\phi^2 (1-|\phi|)^{-4} n_{\min}^{-1}$. This is the second bracket term exactly; with (2) it gives the two-sided statement, the factor 2 on the first bracket term being the elementary $(a+b)^2 \leq 2a^2 + 2b^2$ of step (4). □

The constant degrades like $(1-|\phi|)^{-4}$ because the Fourier bound charges the *whole* high-frequency enhancement of s_ϕ against the residual roughness, whereas near $|\phi| = 1$ only a narrow low-frequency window matters. **The $|\phi| \to 1$ regime is therefore not covered, no rate is claimed there, and no sharpness is claimed anywhere.**

B Related work and positioning

Fixed and adaptive patching. Patch-based forecasters set one temporal resolution in advance (Nie et al., 2023); multi-scale and adaptive variants vary it by layer, by observed frequency content, or through a learned boundary predictor (Chen et al., 2024; Ankireddy et al., 2026; Wang et al., 2025; Zucchi et al., 2026). Liu et al. (2026) learn dynamically adjustable patch boundaries together with time-aware weighted aggregation for irregularly sampled series, where the boundaries must also account for the sampling itself; the setting differs from ours, but the mechanism is adaptive patching. Learned boundary predictors can be trained end-to-end against forecasting loss and are therefore task-sensitive, but their allocations are coupled to the particular trained system; Zucchi et al. (2026), for example, ask when content-adaptive patching inside a forecasting pipeline beats a tuned fixed baseline. TATC asks a complementary question: a *pre-reader* task-weighted allocation is fixed before the downstream learner exists, so transfer to a different reader can be tested directly in §5.3.

Compressive and latent-bottleneck memory. Compressive transformers, latent bottlenecks and state-space memories budget *stored state* rather than input resolution (Rae et al., 2020; Jaegle et al., 2021; Dai et al., 2020; Gu et al., 2020). The compression there happens inside the model and is trained jointly with it; here, it happens before the model and is fixed in advance, which is what makes the representation an experimental variable.

Token merging and dynamic tokenization. Token-merging and entropy-driven tokenization compress by input similarity or predictability (Götz et al., 2025; Pagnoni et al., 2025). These consult input or content statistics, or model-internal criteria, rather than the explicit pre-reader M-weighted criterion studied here. Where such a rule is a function of the history distribution alone, the insufficiency argument of §1 applies to it: two tasks sharing that distribution but differing in M receive the same allocation.

Aggregation, design, and quantization. The criterion itself is inherited. Optimal aggregation of linear models asks which pre-model aggregation map preserves a downstream linear objective (Fisher, 1962; Laisney, 1984; Chipman & Winker, 1992; 2005; Silvestrini & Veredas, 2008); goal-oriented optimal design asks the same question with a task-weighted criterion (Attia et al., 2018; Naumer & Kamalabadi, 2023); Naumer & Kamalabadi (2023) also seek forecasting measurements independent of a particular eventual forecaster, but for a different measurement-selection criterion and infinite-horizon setting. The contiguous-partition optimization here is Fisher's grouping problem under a block-additive cost (Fisher, 1958; Bellman, 1961; Vidal, 1993; Jackson et al., 2005). The asymptotic allocation law is the order-one case of sharp variable-knot spline asymptotics (Barrow & Smith, 1978) and the Panter–Dite point density (Panter & Dite, 1951; Bennett, 1948). **We take all of this as a starting point.** What is specific here is that the roughness field arises from *task* geometry, and that the finite-budget behavior of the two tractable reductions (the block-additive surrogate and the high-resolution density) is empirically tested rather than assumed.

Task-oriented compression. Compression evaluated against a downstream objective rather than reconstruction fidelity is an established framing (Chen et al., 2025; Ji & Müller, 2017; Rha et al., 2020; Ahmed et al., 2026). This paper is a temporally ordered, contiguous-budget instance of it, with the additional constraint that the allocation is fixed before any downstream learner is trained.

Exact finite constructions used in the comparison arms. The three non-J arms are fully deterministic at finite H and K. For uniform, raw endpoints are $\mathrm{round}(kH/K)$ for $k = 1, \ldots, K$. For log, they are the K points of a geometric sweep from 1 to H, rounded: $h_k = \mathrm{round}\big(H^{(k-1)/(K-1)}\big)$ for $k = 1, \ldots, K$, followed by the same deterministic repair: rounding causes early endpoints to collide, and at $H = 512$, $K = 16$, the raw sweep yields a repeated 2. **The exponent is $(k-1)/(K-1)$, not k/K**: the two differ from the third endpoint onward, and only the former is the frozen arm. For input_only, TRAIN defines $A_{\mathrm{input},j} = \mathbb{E}(q_{j+1} - q_j)^2$ for $j < H$, with the final entry copied from its predecessor; endpoints are the k/K quantiles of $A_{\mathrm{input}}^{1/3}$. In all three constructions, a deterministic forward-then-backward repair enforces strictly increasing integer endpoints, K nonempty bands, and a final endpoint H. Thus "input only" contains no target or task geometry.

C Full experimental results

This appendix reports every cell behind the main text's summaries, the audit pool behind §5.1, the full configuration, and the integrity gates.

The Weather $S = 720$ cell is the one referred to in §5.3: task_aware is lower than uniform by 0.000155, against seed standard deviations of 0.0080 and 0.0033.

C.1 Does better exact-risk optimization help the reader? A direct substitution

§5.1 and the Limitations treat optimization quality as a live confound for every close comparison. To separate optimization from transfer, we evaluate two alternative partitions for every $K = 16$ cell under the unchanged neural reader: the best of 32 multistart local minima, and the minimum of the $23{,}041$-partition audit pool. **Both are selected by exact J, never by reader performance.**

Standing. Post-hoc and VAL-only. These runs were specified after the TEST results in Table 2 existed, so these are descriptive diagnostics on the reused development validation set, not independent generalization evidence. Each score below is a VAL model-selection score, using the same best-epoch VAL metric as in §5.3, since TEST is evaluated only once in this study and this audit does not reopen it. **No TEST value is recomputed, and none is used to select any partition or score here.**

Table 3: Neural TEST MSE, mean (seed standard deviation) over three seeds, $K = 16$. Horizon averages appear as Table 2 in the main text.

dataset	S	uniform	log	task_aware
ETTm1	96	0.5269 (0.0154)	0.4147 (0.0091)	**0.4038** (0.0089)
	192	0.5134 (0.0264)	**0.4450** (0.0156)	0.4726 (0.0072)
	336	0.5546 (0.0410)	**0.4640** (0.0072)	0.5017 (0.0089)
	720	0.6179 (0.0215)	0.5258 (0.0176)	**0.5210** (0.0018)
Weather	96	0.1907 (0.0102)	0.1695 (0.0053)	**0.1656** (0.0004)
	192	0.2365 (0.0019)	**0.2127** (0.0006)	0.2135 (0.0026)
	336	0.2851 (0.0024)	0.2626 (0.0050)	**0.2611** (0.0022)
	720	0.3404 (0.0033)	**0.3241** (0.0034)	0.3402 (0.0080)
Electricity	96	0.2956 (0.0012)	0.2201 (0.0002)	**0.1884** (0.0003)
	192	0.3102 (0.0004)	0.2280 (0.0007)	**0.2083** (0.0009)
	336	0.3215 (0.0018)	0.2430 (0.0018)	**0.2241** (0.0014)
	720	0.3499 (0.0018)	0.2744 (0.0008)	**0.2573** (0.0011)

Table 4: Effect of exact-J refinement on the task-aware surrogate-DP proposal $\mathcal{P}_0$. Exact-J reduction is $(J(\mathcal{P}_0) - J(\widehat{\mathcal{P}}))/J(\mathcal{P}_0)$ under the common task-aware geometry; displacement is the summed absolute movement of the 15 internal boundaries. These are existing development quantities, not a new TEST analysis.

dataset	S	exact-J reduction	displacement
ETTm1	96	10.0%	104
	192	12.0%	102
	336	25.1%	151
	720	17.5%	121
Weather	96	4.3%	44
	192	3.0%	35
	336	17.2%	95
	720	15.9%	94
Electricity	96	39.6%	81
	192	30.0%	51
	336	28.2%	61
	720	33.0%	135

References to TEST only restate Table 2 in order to distinguish its ordering from the VAL ordering reported below. The VAL and TEST orderings are not identical: on Weather, log achieves lower mean TEST MSE at $S = 192$ and $S = 720$, while on VAL task_aware wins at all four Weather horizons. The cells where log wins *on VAL* are ETTm1 $S = 192$ and $S = 336$.

Design. P_{MS} is the lowest-exact-J of the 32 random-start local minima and $P_\star$ the lowest-exact-J of the pool of §5.1. In seven of the twelve cells, the two coincide, so there are 17 **unique substitutions** over 51 reader runs. The reader configuration is held constant (sharing architecture, optimizer, batch size, epoch budget, token features, and seeds $0/1/2$), so that arms differ solely in their interval endpoints. Writing L for seed-mean VAL MSE,

$$\Delta J_a = \frac{J(P_a) - J(\widehat{\mathcal{P}})}{|J(\widehat{\mathcal{P}})|}, \qquad \Delta L_a = \frac{L(P_a) - L(\widehat{\mathcal{P}})}{L(\widehat{\mathcal{P}})}, \qquad a \in \{\mathrm{MS}, \star\},$$

both signed, with negative better in each.

Lower exact risk does not consistently transfer. Over the 17 unique substitutions:

Sign agreement is 0.588 and tie-aware rank agreement between ΔJ and ΔL is $+0.40$. **Thirteen substitutions lowered exact J; six lowered VAL MSE and seven raised it.** The largest single case runs the wrong way: at ETTm1 $S = 96$ a partition with 5.77% lower exact task risk reads 8.27% worse, and the cell falls from beating log on VAL to losing to it. This is direct evidence

Table 5: Observed signs of exact-risk and reader-error changes over the 17 unique substitutions.

	reader improves, $\Delta L < 0$	reader worsens, $\Delta L > 0$
exact risk improves, $\Delta J < 0$	6	7
exact risk worsens, $\Delta J > 0$	0	4

of **non-calibration within the evaluated alternatives**: lower exact risk is neither sufficient for, nor consistently predictive of, lower reader error. It is the direct-substitution counterpart of §C.3's rank-correlation finding.

Table 6: Exact-risk and seed-mean VAL-MSE changes for every substitution in this audit. Values are percentages; daggered rows are those where the best multistart solution and the pool minimum is the same partition, so the two pairs of columns report one substitution.

dataset	S	ΔJ_{MS}	ΔL_{MS}	$\Delta J_\star$	$\Delta L_\star$
ETTm1	96†	-5.77	$+8.27$	-5.77	$+8.27$
	192†	-3.15	-1.30	-3.15	-1.30
	336†	-7.58	$+0.38$	-7.58	$+0.38$
	720†	-7.59	-1.65	-7.59	-1.65
Weather	96	-0.01	$+0.15$	-0.51	-0.08
	192†	-0.0034	-0.67	-0.0034	-0.67
	336†	-2.57	$+0.91$	-2.57	$+0.91$
	720†	-0.04	$+0.72$	-0.04	$+0.72$
Electricity	96	$+15.83$	$+5.09$	-4.09	-0.39
	192	$+10.94$	$+4.14$	-9.54	$+0.04$
	336	$+3.37$	$+2.76$	-8.83	$+2.32$
	720	$+11.88$	$+2.83$	-2.75	-0.21

Every contrast, signed.

ETTm1: optimization nearly resolves one horizon and fails at the other. At $S = 192$, where log wins on VAL by 0.008741 in seed mean, the pool-minimum partition closes 97.3% of that deficit, leaving 0.000235. **That residual is 0.7% of the pool-minimum arm's own three-seed range of 0.03413, and the partition beats log in one of the three seeds, not three**, so the remaining ordering is below the resolution of this experiment. At $S = 336$, the lower-J partition instead moves the reader the wrong way and widens the deficit. **Stronger observed optimization does not consistently rescue the ETTm1 comparison.**

It also buys no additional wins. $\widehat{\mathcal{P}}$ achieves lower seed-mean VAL MSE than log in 10 of the 12 cells; the pool-minimum partition does so in 8.

A control, and its scope. All four substitutions that *raised* exact J also raised reader error. They are the four Electricity multistart arms, and within these twelve $K = 16$ cells, the Electricity cells are the only ones where the frozen refinement outperforms the best multistart solution in exact J; across the wider 36-cell audit of §5.1, that is not the general pattern. **This is therefore recorded as a dataset-and-family-specific control, not a general monotonicity claim**, and four observations from one dataset and one family support nothing broader.

Scope. This is not a causal decomposition and carries no TEST or population-level claim, and $P_\star$ is a pool minimum rather than a certified optimum, so headroom outside the pool is not excluded. Per-dataset rank agreement is available from the supplementary result table and is not interpreted here: at five unique substitutions, Weather has essentially no resolution.

C.2 CONTROLLED PATH AUDIT OF TRANSFER FIDELITY

Scope. Post-hoc and VAL-only, complete and reported regardless of direction. For each of the twelve cells, $e_k(\alpha) = \text{repair}(\text{rint}[(1-\alpha)e_k^{\log} + \alpha e_k^{\text{task_aware}}])$ at $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$, under the same reader and seeds as §5.3. **The two endpoints and their reader results were already known**: $\alpha = 0$ and $\alpha = 1$ reuse the published log and task_aware VAL runs, and the relationship between them (Table 2's TEST-measured task_aware-vs-log difference of $+2.7\%$ on ETTm1, $+1.2\%$ on Weather, and -9.1% on Electricity) motivated this audit. **Only the three interior arms,** 108 **new reader runs, are prospective**; every execution and integrity gate (endpoint hashes, validity, pairwise distinctness, frozen run identities, bitwise poison invariance, exact reference run replication, and exact reproduction of path endpoints) passed before any interior reader was trained. **The motivating comparison above is TEST; this audit is VAL only**, and on VAL task_aware already beats log in 10 of the 12 cells, so the ordering below need not match Table 2. Because VAL selects both the pilot penalty and each neural checkpoint, this is a development-set diagnostic rather than independent held-out evidence. With $d(\mathcal{P}, \mathcal{Q}) = 15^{-1} \sum_{k=1}^{15} |e_k^{\mathcal{P}} - e_k^{\mathcal{Q}}|$, the mean absolute displacement of the internal boundaries, the interior partitions closely track their nominal α: displacement from log covers 0.246–0.252, 0.498–0.501 and 0.748–0.754 of the log-to-task-aware distance at $\alpha = 0.25, 0.5, 0.75$, so rounding and repair introduce negligible path distortion. **This is one path and one reader**: it identifies no reader geometry and supports no causal claim about distance from log; with an interior arm beating both endpoints in only two of twelve cells, by less than typical seed variation (below), **it does not support a log-anchored objective**.

Table 7: The complete five-point path, all twelve cells. Relative change from $\alpha = 0$ (log) in percent; negative is lower (better) in both exact J and VAL MSE L. $\alpha = 0$ is omitted as the 0% reference by construction; $\alpha = 1$ reproduces task_aware on VAL rather than the TEST value in Table 2. Per-seed values are in the supplementary result table; paired differences follow directly because seeds are matched across α.

		ΔJ from log (%)				ΔL from log (%)			
dataset	S	.25	.5	.75	1	.25	.5	.75	1
ETTm1	96	−30.5	−45.2	−50.8	−61.3	−0.10	−2.31	+0.30	−4.79
	192	−29.5	−46.5	−48.6	−53.7	+1.78	+4.54	+1.13	+1.36
	336	−32.7	−50.4	−45.3	−54.6	+1.78	+5.97	+2.68	+1.09
	720	−31.5	−47.8	−45.8	−50.5	+4.84	+4.59	−0.07	−0.38
Weather	96	−40.4	−65.2	−78.0	−89.6	−1.46	−1.38	−2.53	−2.00
	192	−43.4	−67.0	−78.9	−87.6	−1.43	−1.77	−1.02	−1.01
	336	−42.7	−68.3	−81.3	−87.9	−0.49	+0.09	+0.17	−1.20
	720	−39.4	−66.2	−82.5	−89.8	−0.05	+0.13	+0.38	−0.68
Electricity	96	+6.8	−15.6	−26.0	−41.0	+2.39	−7.13	−11.16	−15.14
	192	−20.8	−12.7	−0.9	−28.9	−6.05	−5.26	−3.00	−9.74
	336	−18.6	−14.3	−10.6	−27.1	−4.03	−4.37	−4.31	−7.89
	720	−22.6	−11.9	−13.7	−33.0	−4.40	−3.49	−3.92	−8.13

Reading the curves. Transfer fidelity along this path is **dataset- dependent**. On Electricity, exact J closely tracks reader error: the prespecified tie-aware rank association between them is $+0.70$ to $+1.00$ across the four horizons, task_aware has the lowest VAL MSE in every cell and every seed, and every interior partition is worse than task_aware in all 36 (cell, α, seed) comparisons: moving toward log does not recover any performance advantage from proximity to it. Descriptively, not a prespecified statistic: the reader's error moves in the same direction as exact J at 15 of 16 adjacent steps, including where J itself is non-monotone, e.g. at $S = 192$ where J falls, rises twice and falls again and L does the same at every step. On Weather, transfer is nearly decoupled from exact J: rank association is $+0.10$ at three of the four horizons ($+0.80$ at $S = 96$) while exact J falls 88–90% along the path and mean VAL MSE moves by at most 2.5%. An interior arm beats both endpoints in only two of the twelve cells, both on Weather ($\alpha = 0.75$ at $S = 96$, $\alpha = 0.5$ at $S = 192$), by 0.53% and 0.77%, which fall below Weather's typical three-seed range of 1.19% and is not interpreted as a meaningful gain. On ETTm1, the association is irregularly signed (-0.40 to $+0.50$ across horizons), and the curves are seed-unstable. **No dataset shows the interior-optimum pattern that would motivate a log-anchored objective**, and no result or dataset is selectively omitted: Table 7 accounts for all 60 curve points, with the twelve $\alpha = 0$ values stated as zero references in its caption.

Table 8: The high-resolution log condition on all three development-estimated task geometries; no TEST quantities. **The condition is $R_{\log,\mathrm{HR}} = 1$ and a log–log slope of -3.** A perfectly flat A gives $R_{\log,\mathrm{HR}} = 15.52$.

	$S = 96$	$S = 192$	$S = 336$	$S = 720$	the condition
$R_{\log,\mathrm{HR}}$, ETTm1	15.909	14.638	15.282	14.765	1
$R_{\log,\mathrm{HR}}$, Weather	16.033	13.531	12.855	13.668	1
$R_{\log,\mathrm{HR}}$, Electricity	12.601	9.686	9.651	10.638	1
log–log slope, ETTm1	$+0.078$	-0.128	-0.072	-0.104	-3
log–log slope, Weather	-0.134	-0.278	-0.308	-0.295	-3
log–log slope, Electricity	-0.729	-0.651	-0.612	-0.517	-3

Table 9: Tie-aware rank agreement in every one of the thirty-six (dataset, K, S) cells, against held-out linear-reader TEST MSE. §5.2 quotes the column means at each K. Each cell ranks the same six constructions; agreement over a designed set, not an estimate of a population correlation.

		$K = 8$		$K = 16$		$K = 32$	
	S	J	$\widetilde{J}$	J	$\widetilde{J}$	J	$\widetilde{J}$
ETTm1	96	$+0.824$	$+1.000$	$+0.941$	$+0.941$	$+0.412$	$+0.941$
	192	$+0.941$	$+0.353$	$+0.882$	-0.059	$+0.412$	-0.412
	336	$+0.941$	$+0.235$	$+0.941$	-0.353	$+0.647$	$+0.118$
	720	$+0.941$	-0.353	$+1.000$	-0.353	$+0.647$	$+0.412$
Weather	96	$+1.000$	$+0.647$	$+1.000$	$+1.000$	$+0.588$	$+0.588$
	192	$+0.824$	-0.059	$+1.000$	$+1.000$	$+0.647$	$+0.647$
	336	$+0.647$	$+0.235$	$+1.000$	$+1.000$	$+1.000$	$+1.000$
	720	$+0.647$	-0.294	$+1.000$	$+1.000$	$+1.000$	$+1.000$
Electricity	96	$+1.000$	$+0.143$	$+1.000$	$+0.600$	$+0.943$	$+0.657$
	192	$+1.000$	-0.371	$+1.000$	-0.143	$+1.000$	$+0.829$
	336	$+1.000$	-0.371	$+1.000$	-0.143	$+1.000$	$+0.771$
	720	$+1.000$	-0.371	$+1.000$	-0.143	$+1.000$	$+0.086$

The statistic is defined in §5.2. Average ranks matter here because uniform and input_only are the *same* partition on ETTm1 and Weather, so they carry identical J, identical $\widetilde{J}$, and identical error; gate G12 checks the tie behavior on a known tied answer. The cells are not homogeneous, and exposing them shows where the averages come from. On Weather at $K = 16$, the two objectives rank the six arms identically in all four horizons, so that the dataset contributes nothing to the headline separation at the primary budget; ETTm1 and Electricity carry the $K = 16$ gap. The surrogate's failures are not spread evenly either: it is negative in thirteen of the thirty-six cells, and every one of those is at $S \geq 192$, while the exact objective is never negative anywhere. Averaged over cells, the surrogate improves with budget, and the exact objective does not, which is the budget dependence §5.2 reports.

Table 10: Tie-aware rank agreement between J and $\widetilde{J}$ on the uniform-random family, 20,000 partitions per cell, all 36 cells of the audit. **Fidelity is dataset and budget-dependent**, not a property of the surrogate alone: it is consistently strongest on Weather and near-perfect there at $K \geq 16$, degrades with horizon on ETTm1, and is absent on Electricity at $K \leq 16$. A second family of 3,000 perturbations of the six arms per cell (moving $r \in \{1, 2, 4, 8\}$ internal boundaries by up to $d \in \{1, 4, 16, 64\}$) yields the same separation at $K = 16$: $+0.978$ to $+0.991$ on Weather against -0.106 to $+0.573$ on ETTm1 and -0.150 to $+0.622$ on Electricity.

dataset	K	$S = 96$	$S = 192$	$S = 336$	$S = 720$
ETTm1	8	$+0.418$	$+0.206$	-0.034	-0.129
	16	$+0.886$	$+0.810$	$+0.678$	$+0.618$
	32	$+0.977$	$+0.961$	$+0.939$	$+0.915$
Weather	8	$+0.785$	$+0.787$	$+0.675$	$+0.569$
	16	$+0.976$	$+0.969$	$+0.964$	$+0.953$
	32	$+0.967$	$+0.961$	$+0.958$	$+0.955$
Electricity	8	-0.119	-0.111	-0.144	-0.180
	16	$+0.015$	$+0.006$	-0.017	-0.081
	32	$+0.256$	$+0.230$	$+0.156$	$+0.056$

C.3 Predictive non-uniqueness and calibration audit

Re-estimating $(\widehat{\Sigma}, \widehat{M})$ on contiguous blocks holding the most recent or the earliest 10%, 25% and 50% of ETTm1 training origins, and comparing each task_aware partition with the full-split one, shows that internal boundary locations are weakly identified by downstream loss. Across the 24 primary cells (three fractions, two anchors, four horizons), at most 3 of 15 internal boundaries keep their original positions, with summed boundary displacement between 149 and 939 lag positions, yet the VAL MSE of a linear ridge reader degrades by less than 1% in 19 cells and by less than 2.2% in 21. Substantially different allocations of the same K measurements can therefore have near-equivalent predictive utility.

Pooling the primary analysis with a variant that re-selects the pilot penalty on each reduced block gives 48 task_aware cells. Exact-J regret has median $+4.26\%$ (maximum $+21.62\%$), while VAL MSE changes by a median of only $+0.04\%$; the two disagree in sign in 10 cells, and the largest regret occurs in a cell whose VAL MSE improved by 0.44%. Differences in J within one partition family, therefore do not track downstream error, in contrast with the across-family ranking agreement of §5.2.

The surrogate-fidelity audit pool. Per cell, it is the union of 20,000 partitions drawn uniformly over compositions of H into K positive parts; 3,000 perturbations, 500 around each frozen arm, moving $r \in \{1, 2, 4, 8\}$ internal boundaries by up to $d \in \{1, 4, 16, 64\}$ with any draw that empties a band rejected rather than repaired; the DP optima of the three J-arm geometries; 32 exact-J local minima, started from further uniform draws and refined under the *identical* ± 1 steepest-descent rule, strict decrease and tie-breaks that produce $\widehat{\mathcal{P}}$; and the six frozen arms. That is 23,041 partitions. Sampling is seeded and deterministic, and two independent processes were verified to reproduce a byte-identical result file.

Checkpoint selection. Every neural arm trains all 15 epochs under one protocol; the reported checkpoint is the one minimizing VAL MSE, which is often reached early. There is no patience or early-stopping mechanism, the rule is identical across arms, and it was frozen before any sweep is run, ensuring that model selection remains strictly independent of test evaluation.

The complete configuration. Table 11 is the register of every knob, so that reproduction does not require opening the code. Values are transcribed from the released implementation and configuration.

The selected pilot penalties are an output of that frozen rule rather than a choice: 10^2 at every horizon on ETTm1, 10^4 at every horizon on Electricity, and 10^{-6} at $S \in \{96, 192, 336\}$ with 10^4 at $S = 720$ on Weather. The spread is the reason a penalty fitted on one dataset is never inherited by another.

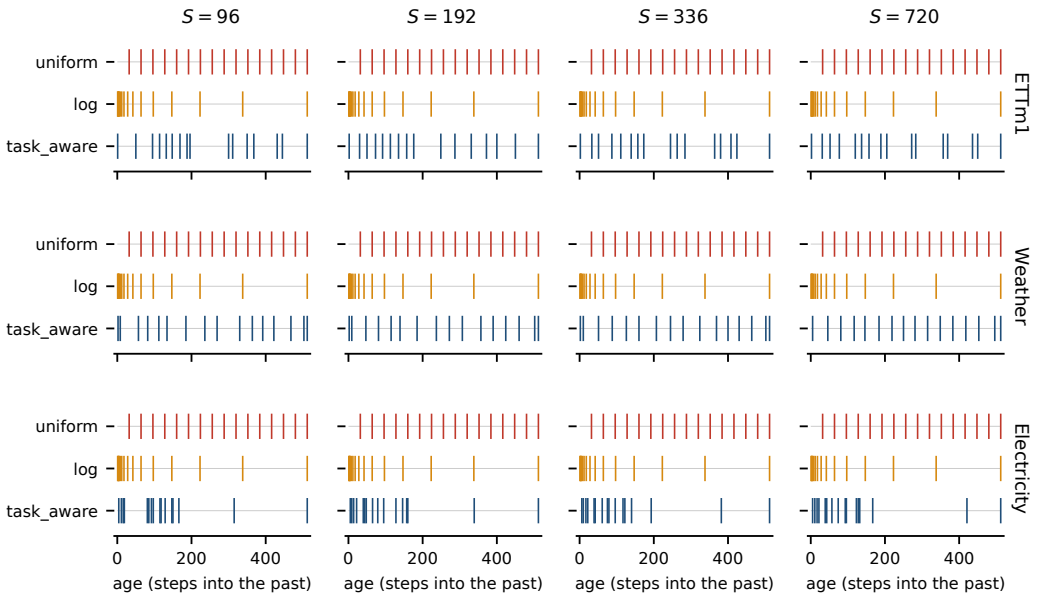

Figure 2: Where the $K = 16$ measurements are placed, in *all twelve* cells: every dataset, every horizon, no selection. Each tick is a band boundary on the age axis, 0 the most recent observation, and 512 the oldest. These are the development-estimated partitions as they stood before any TEST scoring; they are identical, endpoint for endpoint, to the partitions the TEST runs recorded, and no TEST quantity appears in the figure. uniform and log are fixed constructions and do not vary with the dataset; task_aware does, which is the point. Its allocations are neither uniform nor logarithmic: they are irregular, often clustered, and the clusters sit at different ages on different datasets.

Table 11: Every experimental choice, with the component that fixes it: the data and the estimator here, what consumes them in Table 12. Nothing in either table differs between allocation arms: no arm has a tuned hyperparameter of its own.

Geometry	
history H	512
budget K	16 primary; $\{8, 16, 32\}$ for the budget sensitivity
horizons S	$96, 192, 336, 720$
channels	channel-independent; one channel $\times$ one origin is one example
Splits, by index, TRAIN / VAL / TEST	
ETTm1	34,560 / 46,080 / 57,600, end index (cumulative months) $12/16/20$
Weather	36,887 / 42,157 / 52,696
Electricity	18,412 / 21,044 / 26,304
rule (Weather, Electricity)	$n_{\text{train}} = \lfloor 0.7L \rfloor$, $n_{\text{test}} = \lfloor 0.2L \rfloor$, VAL ends at $L - n_{\text{test}}$
containment	every target satisfies $o + S \leq b_i - 1$; context may reach back across a boundary, targets never cross one
standardization	per-channel z-score from the TRAIN-region mean and standard deviation; the pilot then centers lag coordinates on TRAIN means but does not rescale individual ages. A channel that is constant on TRAIN is refused rather than handled
Pilot (estimates $\widehat{\Sigma}$ and $\widehat{M}$)	
$\widehat{\Sigma}$	TRAIN covariance of the history window, streamed sufficient statistics
$\widehat{C}$, $\widehat{M}$	full-resolution ridge fitted on TRAIN; $\widehat{M} = \widehat{C}^{\top}\widehat{C}$, symmetrized
ridge grid	$\{10^{-6}, 10^{-4}, 10^{-2}, 1, 10^{2}, 10^{4}\}$, fixed before any dataset was loaded
penalty selection	lowest VAL MSE; declared VAL contact, identical for every arm

Table 12: The configuration, continued: what consumes the estimated geometry. No arm has a tuned hyperparameter of its own here either.

Solver	
proposal	exact dynamic program on $\widetilde{J}$, earliest-cut tie-break
complexity	The Bellman recurrence costs $O(KH^2)$ after interval costs are tabulated; the current dense general-covariance tabulator uses $O(H^4)$ arithmetic and $O(H^2)$ storage.
refinement	one internal boundary by ± 1, steepest descent on exact J, strict decrease, no restarts and no tunable search budget
downstream error in the loop	none, anywhere
Linear reader	
head	ridge on the K band means and nothing else
ridge grid, selection	the same six values; lowest VAL MSE, identical for every arm
seeds	none; closed form
Neural reader, frozen before the first sweep	
tokens	$K \times [$ band mean, normalized midpoint, log relative width $]$
architecture	$\mathrm{Linear}(3, 64) \rightarrow$ two pre-LN encoder blocks $\rightarrow$ mean pool $\rightarrow \mathrm{Linear}(64, S)$
d_{model} / heads / layers / feedforward	64 / 4 / 2 / 128
dropout / activation / norm-first	0.1 / GELU / yes
optimizer	Adam, learning rate 10^{-3}, weight decay 0
batch / epochs / loss	512 / 15, no schedule / MSE
checkpoint	the epoch of lowest VAL MSE; strict $<$, so a tie keeps the earlier epoch
seeds	$0, 1, 2$
parameters	$67{,}328 + 65S$, independent of K, which is what makes the budget comparison a capacity-matched one
Software; a record of what ran, not a cross-machine guarantee	
	Python 3.13.3, NumPy 2.2.6, pandas 2.2.3, PyTorch 2.7.1+cu118, CUDA 11.8, one consumer GPU. No determinism flag is set: setting one would change which kernels run and therefore invalidate every recorded result.

Table 13: The fifteen pre-outcome integrity gates, run on synthetic data and on each benchmark before any result is read.

gate	what must hold
G1	partition validity: all six arms contiguous, nonempty, ending at H
G2	the dynamic program against exhaustive enumeration at small H
G3	Proposition 1: $\widetilde{J} = J$ on the cross-block-truncated covariance
G4	reparameterization invariance: band sums and band means agree
G5	the additive interval cost reduces correctly at $\Sigma = I$
G6	every DP-returned partition is no worse than the competing arms
G7	no test information: poison the test region; every estimate and endpoint bit-identical
G8	the pilot is informative: VAL MSE below the TRAIN-mean predictor
G9	capacity parity: every arm presents K measurements to an identical head
G10	the $O(K^3)$ exact-J evaluator against the direct trace form
G11	refinement never raises J
G12	the rank statistic is tie-aware on a known tied answer
G13	target containment: every forecast target lies inside its declared split
G14	next-boundary poison: the timestep at each split boundary is inert
G15	split boundaries re-derive from the loaded file's own length

Pre-outcome gates. Fifteen gates run on synthetic data and on each benchmark before any result is read. Three of them check theorems that the paper proves, so an implementation error surfaces as a false theorem rather than as a plausible number. The synthetic qualification series is a sum of three Ornstein–Uhlenbeck components with timescales 2, 8 and 40 steps and weights 0.2, 0.3 and 0.5 (length $12{,}000$; $H = S = 96$, $K = 16$); the theorem-checking gates additionally use small AR(1) covariances with random positive semidefinite task matrices at $H \leq 18$. All synthetic draws use fixed seeds recorded in the released code. These runs qualify the implementation and contribute no reported results.

Gates G3, G4, and G5 are the three that check proved statements. G7 and G14 are the leakage gates; G15 prevents the release of a dataset whose length does not match the declared split rule.