EXCESS SPECTRAL SUPPRESSION: WHY DEEP FORECASTERS SACRIFICE PREDICTABLE TAIL MODES

Anonymous authors
under review

ABSTRACT

Deep forecasters are widely reported to over-smooth their targets. Yet any MSE-optimal predictor must discard unpredictable variance, so comparing output spectra with data spectra cannot separate legitimate shrinkage from genuine bias, and no existing definition of that bias is anchored to a Bayes reference. We supply one: *excess suppression* (ES), the per-mode skill gap to the Bayes reference, measured in the SVD mode space of the target matrix. We measure it on a synthetic family with an analytic Bayes predictor and on eight real benchmarks under a conservative attainable-skill proxy, across rank and time budgets, four intervention families (one applied at inference time), and three zero-shot foundation forecasters. The sacrifice is all-or-nothing: under a tight rank budget, models realize *none* of the attainable tail skill (median unrealized-to-attainable ratio 1.000 over all 29 skill-bearing channels), and tail-ES falls monotonically as the bottleneck widens. The effective rank of the *prediction operator* is the only axis that tracks the sacrifice across every intervention (Spearman -0.89); feature-spectrum entropy dissociates from it. Per-mode loss reweighting closes 31–64% of the tail gap while *improving* overall MSE on all four datasets with attainable tail skill, and the foundation forecasters show tail-ES within ± 0.05 of zero on all eight benchmarks. Tail sacrifice is thus an artifact of narrow budgets and the MSE objective, not of the data: it is measurable, diagnosable at inference time, and correctable in the loss. Code and run outputs are available at `https://anonymous.4open.science/r/excess-spectral-suppression-6980`.

1 INTRODUCTION

Point forecasts of noisy time series *should* be smoother than their targets: the MSE-optimal predictor is the conditional mean $\mu(x) = \mathbb{E}[y \mid x]$, and by the law of total variance

$$\mathrm{Var}(\mu(x)) \;=\; \mathrm{Var}(y) \;-\; \mathbb{E}\left[\mathrm{Var}(y \mid x)\right], \tag{1}$$

so the optimal predictor discards exactly the unpredictable part of the target's variance. The widely reported "over-smoothing" or "spectral blurring" of deep forecasters — from long-horizon benchmarks (Zhou et al., 2021; Wu et al., 2021; Zeng et al., 2023) to AI weather models (Lam et al., 2023; Bonavita, 2024) — is therefore not by itself a bias: comparing model output spectra with data spectra conflates legitimate shrinkage with harmful suppression. **Bias must be defined relative to a Bayes reference.**

This paper supplies that definition, a measurement protocol that makes it practical, a mechanistic account of where the bias comes from, and a loss-level correction. Our starting point is a structural analogy with popularity bias in recommendation (Lin et al., 2025), where models memorize popular items in the leading spectrum of their score matrix and dimensional collapse amplifies the bias. In forecasting, *mode energy* plays the role of item popularity, *low-energy yet predictable* structure that of niche-but-relevant items, and the mode spectrum of the prediction operator that of the score-matrix spectrum.

Excess suppression. Let $Y \in \mathbb{R}^{N \times H}$ collect N future windows of horizon H, and let $v_1, \ldots, v_K$ be its right singular vectors in decreasing order of energy (the *data modes*). For a forecaster f

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.

with predictions $\hat{Y}$, we define the per-mode skill and the excess suppression relative to the Bayes predictor $f^\star$:

$$R_k^2(f) \;=\; 1 - \frac{\|(\hat{Y} - Y)v_k\|^2}{\|Y v_k\|^2}, \qquad \mathrm{ES}_k(f) \;=\; R_k^2(f^\star) - R_k^2(f). \tag{2}$$

ES_k is invariant to the legitimate shrinkage of equation 1: an ideal learner has $\mathrm{ES}_k = 0$ at every k regardless of how noisy mode k is.

Energy order versus predictability order. Each data mode has two independent attributes: its energy σ_k and its predictability ρ_k, the fraction of the mode projection's variance explainable from the input. Two classical results imply that constrained learners allocate capacity mainly by energy. Reduced-rank regression keeps the directions with the largest singular values of the input–output cross-covariance ($\approx \sigma_k\sqrt{\rho_k}$) and zeroes out the rest (Izenman, 1975; Reinsel & Velu, 1998). Deep linear networks learn modes sequentially, at times $\propto 1/(\sigma_k\sqrt{\rho_k})$ (Saxe et al., 2014), so early stopping acts as an implicit rank constraint — the same low-rank bias that gradient descent exhibits in deep matrix factorization (Arora et al., 2019). The corollary is the central prediction of this paper: *under a rank or time budget, low-σ/high-ρ modes are sacrificed first*. They are cheap to ignore in MSE yet carry real predictable structure.

Collapse as an amplifier. Embedding width sets an explicit rank ceiling; representation collapse lowers the *effective* rank further. Collapse is documented across deep architectures — rank collapse in pure attention stacks (Dong et al., 2021), neural collapse in the terminal phase of training (Papyan et al., 2020) — and is severe in low-SNR time-series settings (Qiu et al., 2025; Ilbert et al., 2024). We quantify it throughout by the spectral-entropy effective rank of Roy & Vetterli (2007) and test how it modulates tail-ES under interventions on width, weight decay, dropout and training time, including spontaneous collapse in unconstrained models.

Correction. Because the pathology is energy-ordered allocation under a budget, the natural correction reorders the budget by predictability; indiscriminately amplifying the tail would also boost pure-noise directions. PredReg (§5) weights per-mode errors by $w_k \propto \hat{\rho}_k^{\gamma}/\sigma_k^{\alpha}$ with probe-estimated $\hat{\rho}_k$. Its $\gamma{=}0$ ablation is exactly the indiscriminate variant, and contrasting the two keeps the energy and predictability axes apart.

Contributions. (1) A bias definition and measurement protocol — mode basis from training targets only, raw-scale measurement, conservative attainable-skill reference — that separates excess suppression from legitimate shrinkage. (2) Evidence: complete suppression of predictable tail modes under budget constraints on synthetic data with an analytic Bayes reference, and universal rising-tail ES curves across eight real benchmarks, monotone in bottleneck rank. (3) Mechanism: the effective rank of the *prediction operator* is the only axis that tracks the sacrifice across four intervention families, one of them applied at inference time. (4) Correction: PredReg, its ablations, and a diagnostic that tells which variant a dataset needs.

2 Related work

Long-horizon forecasting architectures. We probe the benchmarks and model families of this line: recurrent–convolutional hybrids (Lai et al., 2018), basis-expansion networks (Oreshkin et al., 2020), long-sequence Transformers (Zhou et al., 2021; Wu et al., 2021), patch- and variate-tokenized Transformers (Nie et al., 2023; Liu et al., 2024), and the linear baselines that questioned whether any of it was necessary (Zeng et al., 2023). These models are scored almost exclusively by aggregate MSE, which cannot see mode-level allocation. Reversible instance normalization (Kim et al., 2022) is near-universal, which is why we measure on the raw scale. **Spectral bias, in networks and in forecasters.** Fredformer (Piao et al., 2024) architecturally corrects a Transformer bias toward high-energy frequency bands, building on standard frequency decomposition (Zhou et al., 2022; Wu et al., 2023). The spectral-bias line (Rahaman et al., 2019; Xu et al., 2020) and its input-encoding remedies (Tancik et al., 2020) concern pointwise function fitting, and damped kinetic-energy spectra in AI weather models (Bonavita, 2024) record the phenomenon without adjudicating it. We instead work in the SVD mode space of the target matrix (cross-window structure rather than

per-frequency bins), define bias *relative to a Bayes reference*, and correct it with a model-agnostic loss. **Rank/entropy collapse.** DropPatch and SAMformer (Qiu et al., 2025; Ilbert et al., 2024) — the latter importing sharpness-aware minimization (Foret et al., 2021) to escape it — establish that collapse hurts aggregate accuracy. We show that collapse *selects which modes* are hurt, and quantify the collapse–ES relation. **Imbalance, spectral and otherwise.** ReSN (Lin et al., 2025) regularizes the leading spectrum of a recommender's score matrix, and reweighting so that rare-but-informative components are not drowned out is standard in classification (Lin et al., 2017; Cui et al., 2019). PredReg is the mode-spectrum analogue, with one crucial difference: the "rarity" axis (energy) and the "value" axis (predictability) must be measured separately. Our ablations test this: variant B uses neither axis and fails everywhere, while $\gamma{=}0$ uses energy alone (§5). For pretrained forecasters (Ansari et al., 2024; Das et al., 2024; Woo et al., 2024), data-level treatments rebalance sampling; we instead ask *where* imbalance is memorized, and §6 tests whether these models inherit the sacrifice at all. **Classical theory.** We assemble, rather than re-derive, reduced-rank regression (Reinsel & Velu, 1998), low-rank approximation (Eckart & Young, 1936) and deep linear dynamics (Saxe et al., 2014) into a forecasting-specific suppression law with a new measurement protocol.

3 Measuring excess suppression

3.1 Protocol

The measurement pipeline addresses four confounders. (i) *Normalization*: instance normalization (RevIN) alters mode structure, so all spectra are measured on the raw (denormalized) scale. (ii) *Leakage*: the mode basis $\{v_k\}$ is estimated from training targets only; test targets are only projected onto it. (iii) *Multivariate structure*: the analysis is per channel; joint channel$\times$time mode bases are deferred (Appendix B). (iv) *Estimation noise*: all curves carry bootstrap (Efron, 1979) confidence bands over test windows.

On real data the Bayes reference is unavailable, so we use a *conservative attainable-skill proxy* $R^2_\star$: the per-mode maximum of R^2_k over a deep ensemble of high-capacity models, a full-rank closed-form linear predictor (OLS), and a strong simple baseline (DLinear). Reported ES is therefore a lower bound; strengthening the reference can only raise it.

3.2 Synthetic family with analytic reference

We construct $y = V\,\mathrm{diag}(\sigma)\big(\sqrt{\rho}\odot z + \sqrt{1-\rho}\odot \xi\big)$, $x = Az$ with orthonormal smooth A, V and latents $z, \xi \sim \mathcal{N}(0, I_K)$. Mode k then has exact energy σ_k (geometrically spaced) and predictability $\rho_k \in \{0.9, 0.05\}$ alternating along the spectrum, so all four (energy, predictability) buckets — HH, HL, **LH**, LL — are represented and energy is decorrelated from predictability. The Bayes predictor is analytic, $R^2_k(f^\star) = \rho_k$, and the empirical pipeline recovers it to within 0.012 (minimum mode-basis alignment $|\cos| = 0.995$), which validates the protocol.

3.3 Phenomenon

Synthetic: complete suppression of the LH bucket. Table 1 and Figure 1 summarize five seeds of two architectures (a two-layer linear network and a bottlenecked MLP) across rank budgets. At $r{=}8$ (of $K{=}32$ modes) the LH bucket reaches $\mathrm{ES} = 0.900$: the Bayes ceiling, i.e. *total* sacrifice of predictable tail structure. The other three buckets sit at ≈ 0.05, and model curves closely track the closed-form reduced-rank solution. Checkpoint trajectories reproduce the deep-linear staircase — modes are learned in $\sigma_k\sqrt{\rho_k}$ order — making training time a second budget axis with the same suppression pattern. At full width the linear model reaches the Bayes reference. The MLP does not: it retains a significant LH residual ($\mathrm{ES} = 0.062$, $t = 23.5$) and its representation collapses spontaneously to effective rank ≈ 11 of 32. Collapse amplifies suppression even without an explicit constraint.

Real data: universal rising tails, monotone in rank. Our real-data measurement study (*Phase A*) covers eight benchmarks (ETTh1/h2/m1/m2, Electricity, Traffic, Weather, Solar) with lookback 336, horizon $H{=}96$ and $K{=}48$ modes, split by energy into a *head* half and a *tail* half. Per channel, we train DLinear plus a bottlenecked MLP and a deep linear network at $r \in \{8, 32, 96\}$. Mean

Table 1: Mean ES per (energy, predictability) bucket on the synthetic family (5 seeds). LH = low-energy/high-predictability; the Bayes ceiling is $\rho = 0.9$.

Configuration	HH	HL	**LH**	LL
deep linear, r=8	0.043	0.048	**0.900**	0.051
MLP, r=8	0.122	0.054	**0.902**	0.051
deep linear, r=16	−0.000	0.023	**0.501**	0.052
MLP, r=16	0.010	0.045	**0.572**	0.074
deep linear, r=32	−0.000	0.001	0.001	0.014
MLP, r=32	0.011	0.032	**0.062**	0.037

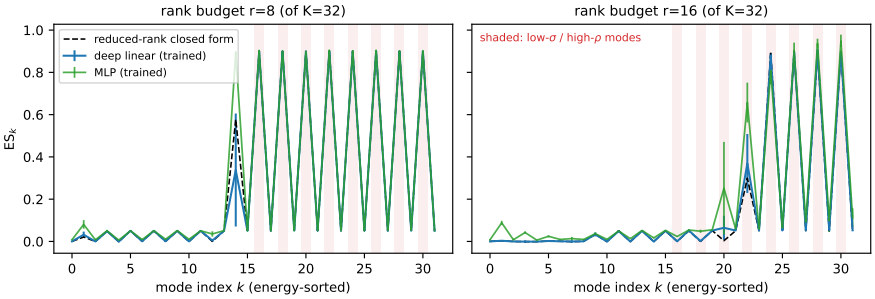

Figure 1: Synthetic ES_k curves at two rank budgets (5 seeds, 95% CIs). Trained networks (deep linear, MLP) coincide with the reduced-rank closed form. Shading marks the low-σ/high-ρ (LH) modes.

ES over the tail exceeds that over the head in 54 of the 56 (model, dataset) configurations; the two exceptions are ETTm1's wider MLPs (r=32, 96). Tail-ES also decreases monotonically in bottleneck rank (r=8 > 32 > 96) for every dataset and both architectures (Table 2). The table's last column is a noise floor: the tail-ES of a re-trained duplicate of the weakest reference member (Appendix B). At r=8, tail-ES sits 2.2–37× above this floor on seven of the eight datasets, the exception being ETTm1 at 0.9–1.0×. Several r=96 entries do fall to floor level, so we anchor universality claims at r=8 and read floor-level entries as upper bounds. The protocol also passes three robustness probes (Appendix B): a random-walk negative control, on which it measures *no* unrealized skill, and horizon and lookback sweeps, under which the complete-sacrifice law is stable.

Decomposing real-data ES: unrealized skill versus injected error. Real-data ES has two distinct components, which we report separately:

$$\mathrm{ES}_k \;=\; \underbrace{\big(R_k^2(f^\star) - \max(R_k^2(f), 0)\big)}_{\text{unrealized attainable skill}} \;+\; \underbrace{\max(-R_k^2(f), 0)}_{\text{injected error}}. \tag{3}$$

The second term is nonzero when the model does *worse than predicting the mode mean*: a rank-constrained prediction operator leaks error into low-energy directions it cannot control. The benchmarks split in two according to which term dominates. On ETT, Weather and Solar, injected error dominates the tail (49–105% of tail-ES). Traffic and Electricity instead carry substantial attainable tail skill (mean $R^2_\star \approx 0.26$ over the tail; single channels up to 0.82), and there unrealized skill dominates for high-capacity models; we call these two datasets, and any channel with attainable tail skill above 0.02, *skill-bearing*. Both components are strictly monotone in the rank budget, so the capacity-allocation mechanism holds in both regimes. The pathologies differ, however: the synthetic setting isolates pure sacrifice of predictable structure, while real benchmarks additionally exhibit spectral leakage.

The sacrifice is complete. The all-or-nothing prediction of Proposition 1 transfers to real data exactly. At r=8, the per-channel ratio of unrealized to attainable tail skill has median 1.000 (IQR $[1.000, 1.000]$) over all 29 skill-bearing channels (Fig. 2, left). Equivalently, r=8 models attain positive tail-mode R^2 in only 13 of 1440 (channel, mode) pairs, with a maximum of 0.07 against

Table 2: Phase A across the eight benchmarks (ranges over architectures; 7–8 channels each). Bold: the two datasets with substantial attainable tail skill. The last column is the twin re-run noise floor for that dataset. "≈ 1" reads "injection-dominated" (Appendix B).

	attainable	tail-ES		injected share	twin
dataset	tail $R^2_\star$	r=8	r=96	(r=96)	floor
ETTh1	0.102	0.92–1.21	0.19–0.28	0.49–0.71	0.19
ETTh2	0.037	1.48–2.92	0.27–0.45	0.86–0.92	0.44
ETTm1	0.019	0.35–0.40	0.12–0.13	0.84–0.88	0.39
ETTm2	0.000	6.32–13.84	0.73–0.99	≈ 1	1.46
Solar	-0.008	2.89–2.91	0.14–0.17	≈ 1	0.08
Traffic	**0.258**	0.56–0.64	0.17–0.19	**0.13–0.22**	0.25
Weather	-0.049	4.20–10.2	0.97–2.12	≈ 1	1.80
Electricity	**0.260**	0.63–0.70	0.22–0.24	**0.07–0.21**	0.22

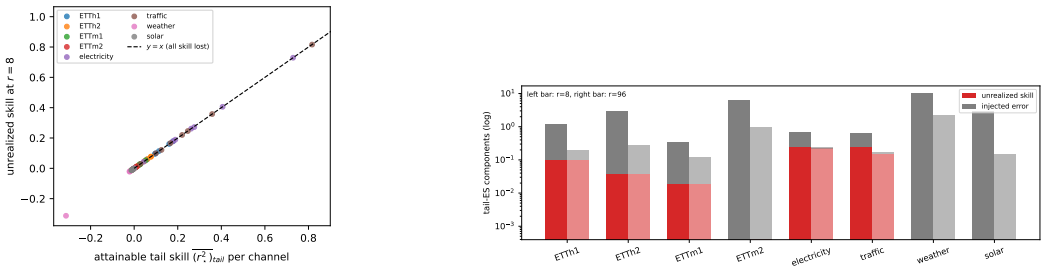

Figure 2: Left: per-channel attainable tail skill vs. unrealized skill at r=8, over 60 channels of 8 benchmarks; every channel sits on the y=x line. Right: tail-ES decomposition per dataset (unrealized vs. injected; left bars r=8, right bars r=96, log scale).

attainable values of up to 0.82. Under a tight rank budget, deep forecasters do not partially learn predictable tail structure; they discard it entirely.

Two caveats bound this claim. First, every model in the reference proxy is itself MSE-trained and subject to the same suppression law, so the proxy is circular in the tail. §5 turns this into a test: a reweighted loss that attained tail R^2 *above* $R^2_\star$ would show that the reference family underestimates attainable skill. The test comes back negative on all four datasets tried (Table 3), so the max-proxy stands. Second, a twin-seed control (a duplicate reference member, excluded from the pointwise max) calibrates how much apparent ES arises from seed noise and winner's-curse selection alone. The resulting floors are substantial, so we make no sub-floor claims: apparent DLinear tail suppression on ETT falls below its floor and is not reported.

4 Mechanism: Collapse Modulates Suppression

4.1 The Suppression Law in the Linear-Gaussian Family

Proposition 1 (Rank budget)**.** *In the linear-Gaussian family of §3.2 (clean inputs), the population rank-r regression retains the r modes with the largest cross-covariance strengths $s_k = \sigma_k\sqrt{\rho_k}$ and predicts 0 in the others; hence* $\mathrm{ES}_k = \rho_k \cdot \mathbb{1}[\, s_k$ *not in top-*$r\,]$*. Excess suppression is thus all-or-nothing, and the sacrificed modes are exactly those with the smallest s_k; since $\sqrt{\rho_k} \le 1$ while σ_k spans orders of magnitude, these are in practice the low-σ modes,* however predictable.

Proposition 2 (Time budget)**.** *Under gradient flow in a depth-2 linear network with whitened inputs and decoupled initialization, mode k's skill follows a sigmoidal trajectory with time constant $\tau_k \propto 1/s_k$; stopping at time t therefore implements the soft rank constraint "keep modes with $s_k \gtrsim 1/t$", with the same sacrifice ordering as Proposition 1.*

Both statements assemble classical results (Reinsel & Velu, 1998; Saxe et al., 2014); proofs and precise conditions are in Appendix A. The empirical content is that trained networks agree with these curves (Fig. 1) and that the ordering transfers to nonlinear models and real data.

4.2 Interventions on real data

We intervene on per-channel PatchTST-style forecasters along four axes — embedding width $d_{\text{model}} \in \{8, \ldots, 128\}$, weight decay, dropout, and training time (checkpoint trajectories) — and measure tail-ES, its two components, and the spectral-entropy effective rank of the pre-head representation (*feature effective rank*). The decomposition of §3.3 is essential here, because there are two regimes. Where the tails carry *no* proxy-attainable skill (ETT), tail-ES is dominated by injected error, which tracks overfitting; regularization then lowers effective rank and tail-ES *together*, superficially contradicting the collapse hypothesis. On the skill-bearing datasets, the hypothesis concerns the *unrealized-skill* component, and there it holds — with one instructive exception.

Budget-like interventions trace the collapse curve. On all six skill-bearing channels, shrinking the width raises unrealized tail skill as feature effective rank falls (Table 4, Fig. 3). Strong weight decay (10^{-2}) collapses feature effective rank by 1.8–$6.5\times$ and multiplies unrealized skill by 4–$11\times$ on the four channels with non-negligible baseline sacrifice; the increase is larger still, though ratio-unstable, on the two channels with near-zero baselines. Injected error stays small throughout ($\leq 10\%$ of unrealized skill), so this is sacrifice of attainable skill rather than overfitting noise; mild weight decay (10^{-4}) even helps slightly.

Dropout dissociates rank from sacrifice. Dropout lowers feature effective rank comparably (2.2–$7\times$), yet unrealized skill *rises* on Traffic ch300 and *falls* on Electricity ch150, where dropout also improves overall MSE by 20%. A matched-rank contrast is the cleanest illustration: on Traffic ch300, width $d{=}8$ and dropout 0.5 reach the *same* feature effective rank (14.3 vs. 14.5) but differ $2.8\times$ in unrealized skill. Both findings replicate over three seeds with non-overlapping ranges (Table 4). The width-series collapse also persists when effective rank is measured at token level (ceiling d_{model}, hence comparable across widths), which rules out an ambient-dimension artifact. Feature-spectrum entropy is therefore a proxy, not the mechanism.

4.3 Measuring the operator: the axis that survives every intervention

We add two operator-side measurements to every intervention run: the effective rank of the prediction matrix $\hat{Y}$ itself, and that of a ridge surrogate of the input→prediction operator, $G = (X^\top X + \lambda I)^{-1} X^\top \hat{Y}$, which we call the *operator effective rank* (op-erank). We also add a fourth intervention family that can touch only the operator: *head truncation*, an inference-time SVD truncation of the trained head to rank $r \in \{2, \ldots, 96\}$, applied to the widest model without retraining. Four results follow, pooled over both skill-bearing datasets (3 seeds).

(i) *Op-erank is the surviving axis.* Pooled across *all* families, it tracks unrealized tail skill at Spearman -0.89 (-0.80 after per-channel standardization; within a dataset–channel–seed cell, -0.69 along width and -0.82 along truncation). Feature effective rank drops to -0.40, from -0.72 on the trained families alone: head truncation moves ES while leaving features untouched, which breaks the feature axis but not the operator axis. (ii) *Within every single family*, op-erank is the strongest or tied-strongest axis (-0.52 to -0.88). (iii) *Dropout gives the sharpest dissociation.* Moving dropout from its minimum to its maximum lowers feature effective rank by 53 (token-level: 11) yet changes op-erank by only 1.9 and unrealized tail skill by -0.008: dropout crushes the measured feature spectrum *without constraining the operator*, and the sacrifice does not move. (iv) *Truncation partly replays training.* On the shared op-erank range, truncation reproduces the dependence seen when training at reduced width qualitatively but not point-for-point (interpolated mean gap 0.056 at a mean ES level of 0.110): roughly half of what training at small width does to the tail can be replayed on a trained model by constraining its operator alone.

The mechanism claim thus rests on direct measurement rather than inference: the operator's effective rank can be measured, can be intervened on at inference time, and is the only axis that tracks the sacrifice across every intervention — including the two families (dropout and head truncation) that dissociate it from the feature spectrum.

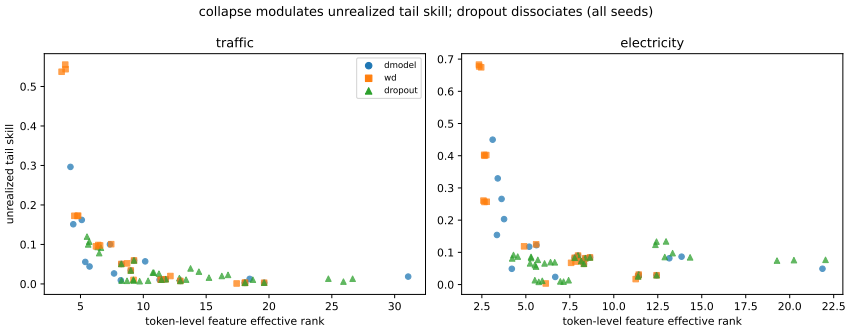

Figure 3: Unrealized tail skill vs. *token-level* feature effective rank (ceiling d_{model}, comparable across widths) on the skill-bearing channels of Traffic and Electricity. Markers denote intervention type; all three seeds are shown for replicated cells. Width and weight decay trace the collapse relation; dropout does not.

5 PredReg: predictability-aware mode balancing

Stage 1 (one-off). Estimate the modes $\{v_k, \sigma_k\}$ from training targets, and estimate $\hat{\rho}_k$ with a ridge probe from inputs to mode projections, validated on a held-out *chronological* split of the training set (an iid split overestimates test-time predictability under distribution shift). **Stage 2 (loss).** On each batch B, minimize

$$\mathcal{L} = \sum_k w_k \, \|(\hat{Y}_B - Y_B)v_k\|^2, \qquad w_k \propto \frac{\hat{\rho}_k^{\gamma}}{\sigma_k^{\alpha}} \text{ clipped to } [w_{\min}, w_{\max}], \tag{4}$$

with weights energy-normalized so that $\sum_k w_k \sigma_k^2 = \sum_k \sigma_k^2$. Setting $\alpha{=}\gamma{=}0$ recovers MSE, and because w_k increases with $\hat{\rho}_k$, pure-noise directions are not amplified. **Variant B (ablation).** MSE plus a ReSN-style penalty on the spectral ratio $\hat{\sigma}_1^2 / \|\hat{Y}_B\|_F^2$ of the batch prediction matrix (computed by power iteration): spectrum-level suppression with no per-mode information.

The probe inherits the circularity — and the fix. The two components of real-data ES (§3.3) call for different treatments, and the weighting must respect both. Where a tail mode carries no probe-detectable predictability, $\hat{\rho}_k \to 0$ drives w_k to its lower clip. This *down*-weights the mode relative to MSE and removes the very penalty that keeps injected error out. A linear MSE-family probe is thus blind in exactly the modes the method is meant to protect: the reference circularity of §3.3 reappears in stage 1. We therefore use $w_k \propto (\hat{\rho}_k + \rho_0)^{\gamma}/\sigma_k^{\alpha}$ with a small floor ρ_0 (*PredReg-blend*). Where $\hat{\rho}_k \approx 0$ it maintains inverse-energy anti-injection pressure; where $\hat{\rho}_k$ is genuinely large it upweights predictable tail structure. The ablations map onto the two pathologies: $\gamma{=}0$ (inverse-energy only) is a pure anti-injection regularizer, floorless $\gamma{=}1$ is exposed to probe blindness, and variant B carries no per-mode information at all.

Synthetic validation: soft vs. hard budgets. On the linear-Gaussian family (5 seeds, deep linear networks), the nature of the rank budget determines what reweighting can do. At $r{=}16$ the rank suffices and the sacrifice is induced by training dynamics: MSE leaves $\text{ES}_{\text{LH}} = 0.506$, and PredReg *eliminates* it ($\text{ES}_{\text{LH}} = 0.001$) at a 3.4% cost in overall MSE (4.99 vs. 4.82). Inverse-energy weighting alone recovers most of it (0.077). Predictability weighting alone recovers LH but, lacking energy normalization, pays elsewhere (ES_{LL} $0.052 \to 0.266$). Variant B recovers *nothing* (0.505, identical to MSE), because a spectrum-level penalty cannot redirect a mode-level allocation. At $r{=}8$ the rank is strictly insufficient for the LH bucket, and every loss is pinned at $\text{ES}_{\text{LH}} \approx 0.87$–$0.90$: no reweighting escapes a hard architectural budget. This is exactly the complete-sacrifice regime seen at $r{=}8$ on real data (§3.3). The reversal on real data below is informative rather than contradictory: with a noiseless probe, full PredReg beats $\gamma{=}0$; with an MSE-family probe on real tails, $\gamma{=}0$ is safer. Predictability weighting helps only to the extent that $\hat{\rho}$ can be estimated.

Real data (Phase C). We retrain the bottlenecked MLPs under each loss. Four findings hold on all four datasets with attainable tail skill above 0.02 (Table 3); a fifth comes from the other four.

Table 3: Phase C: tail R^2 by training loss, averaged over the 24 tail modes and all measured channels (7 on ETTh1/ETTh2, 4 on Traffic/Electricity); MLP at r=96, and the r=32 ordering is identical. Bold: best per dataset. Twin noise floors on these channel subsets: $0.19/0.44/0.41/0.33$.

	ETTh1	ETTh2	Traffic	Electricity
proxy tail $R^2_\star$	0.102	0.037	0.363	0.418
probe-blind tail frac. ($\hat{\rho} \leq 10^{-3}$)	.46–.71	.08–.75	0–.08	0
MSE	-0.117	-0.317	$+0.016$	$+0.030$
PredReg (γ=1, floorless)	-0.156	-0.391	$+0.152$	$+0.145$
PredReg-blend (ρ_0=0.05)	-0.031	$\mathbf{-0.133}$	$+0.182$	$+0.152$
inverse-energy only (γ=0)	$\mathbf{-0.017}$	-0.207	$\mathbf{+0.239}$	$\mathbf{+0.182}$
variant B (spectral penalty)	-0.193	-1.611	-0.481	-0.898

(i) *The strong circularity claim fails.* No loss pushes tail R^2 above the pointwise-max proxy on more than 5.2% of tail (channel, mode) pairs on any dataset: at most 5 of 96 on Traffic and Electricity, and 5 of 168 on ETTh1/ETTh2. The max-proxy therefore remains a valid conservative reference under reweighted training.

(ii) *The sacrifice is nonetheless an artifact of the objective.* At fixed capacity, switching from MSE to the inverse-energy loss closes 31–64% of the tail gap (e.g. Traffic $-0.347 \rightarrow -0.124$) while *also* improving overall MSE on all four datasets (ETTh1 $9.60 \rightarrow 8.78$, ETTh2 $26.5 \rightarrow 23.3$, Electricity $2336 \rightarrow 2165$). Tail skill was not traded away for average accuracy; it was simply never asked for.

(iii) *Reweighting lifts a weak reference member past its own re-run.* On ETTh1, ETTh2 and Traffic, the inverse-energy model's tail gap is smaller than the twin floor, per channel as well as on average: one reweighted MLP lands closer to the pointwise max than a re-trained duplicate of the proxy's DLinear member does. Since the tail max is typically set by the OLS and ensemble members, the claim is limited to "above the weakest member's re-run".

(iv) *The probe-blindness diagnostic predicts which weighting works.* Floorless $\hat{\rho}$-weighting is *worse* than MSE exactly where the ridge probe is tail-blind (ETTh1: .46–.71 of tail modes at $\hat{\rho} \approx 0$; ETTh2: .08–.75) and *better* where it is not (Traffic and Electricity: 0–.08). The ρ_0 floor repairs the blind regime (best on ETTh2), and γ=0 is the robust default. Variant B, which has no per-mode information, degrades tail and overall accuracy everywhere, with up to $2.7\times$ worse MSE on Electricity.

(v) *The prescription validates on the injection-dominated half.* On Solar, Weather, ETTm1 and ETTm2 the tail is injected error rather than unrealized skill (§3.3), so the diagnostic prescribes the inverse-energy loss. It is right on both counts: predictability weighting backfires and amplifies injection, while the prescribed γ=0 loss cuts injected tail error by 2–3$\times$ and improves overall MSE in 7 of 8 (dataset, width) cells. Across all eight benchmarks, the diagnostic thus assigns the correct tool. A one-axis-at-a-time hyperparameter sweep around the blend's defaults leaves this intact in all 14 configurations: each beats MSE on tail R^2, $w_{\max}$ and ρ_0 sit on flat plateaus, and the energy exponent α is the one sensitive axis, with more anti-injection pressure helping — the same direction as finding (ii) (Appendix B).

Architecture dependence: the correction is budget-conditional. Repeating Phase C with PatchTST (d=64) in place of the bottlenecked MLPs completes the budget story from the correction side. MSE-trained PatchTST already realizes nearly all attainable tail skill on the skill-bearing datasets (tail gap to proxy -0.026 on Traffic and -0.077 on Electricity, versus -0.35 and -0.39 for the MSE MLP), and reweighting moves its tail R^2 by at most 0.024. When the operator is not budget-constrained, the loss has nothing to reallocate — the mirror image of the synthetic r=8 endpoint. The correction thus operates strictly between the two regimes, and the op-erank of §4.3 tells a practitioner which regime a model is in. Anti-injection still pays for the unconstrained Transformer: on ETTh1/ETTh2 the reweighted losses improve its *overall* MSE by 14–34% at flat-to-slightly-worse tail R^2, consistent with the probe-blindness diagnostic there.

6 DO MODERN AND FOUNDATION FORECASTERS SHARE THE SACRIFICE?

Modern supervised baselines under the same lens. The Phase-A families are deliberately simple, so one objection remains: is the sacrifice an artifact of a weak model zoo? Under the identical protocol, neither PatchTST in its standard per-channel configuration (3 encoder layers, d=128, patching, RevIN) nor iTransformer (variate-token attention across all channels) exceeds the twin noise floor in dataset-mean tail-ES on *any* of the eight benchmarks. Modern architectures at standard capacity thus show no *claimable* sacrifice, which supports the thesis: sacrifice is created by budgets, not by depth or attention per se, and Phase A shows it emerging monotonically as the budget tightens. Two contrasts do not depend on the floor, because comparing models under the *same* proxy cancels the winner's-curse term. First, the per-channel pipeline, which has the narrower budget, retains more tail-ES than the cross-channel one on both skill-bearing datasets. Second, the RevIN ablation shows no consistent sign, so instance normalization is not the driver (Appendix B).

Foundation forecasters. If tail sacrifice is an artifact of budget and objective rather than a property of the data, forecasters trained with far larger capacity, on cross-series corpora, and with non-MSE objectives should not exhibit it. We evaluate three zero-shot foundation forecasters from independent families — Chronos-Bolt (small) and Chronos-2 (Ansari et al., 2024), and TimesFM-2.5 (200M) (Das et al., 2024) — under the *exact* Phase-A protocol on all eight benchmarks: same windows, same mode basis, same conservative reference $R^2_\star$, using the median (Chronos) or point (TimesFM) forecast.

Tail-ES lies in $[-0.028, +0.046]$ in all 24 (model, dataset) cells, i.e. within ± 0.05 of zero (Table 5, Appendix B). It is 4–71$\times$ below even the *least-suppressed* budgeted supervised configuration on every dataset, below the twin noise floor on all eight, and frequently *negative*: the zero-shot models beat the pointwise-max proxy on 5–89% of tail (channel, mode) pairs. Neither component accumulates anywhere; the largest is an injected error of 0.049 on Weather, where supervised models inject 0.97–10.2. The residual on Electricity (Bolt $+0.046$) concentrates on the two channels with the largest attainable tail skill, where a dedicated unconstrained PatchTST (0.08) still leads: the claim is the *absence of systematic sacrifice*, not zero-shot dominance. Two caveats cut in opposite directions. Variants of Traffic and Electricity appear in the pretraining corpora, so the ETT results carry the weight for unseen data. Conversely, our horizon (96) exceeds Chronos-Bolt's native prediction length (64); this biases the comparison *against* that model, yet its numbers agree with those of its two full-horizon siblings. This is the paper's thesis seen from the other side: remove the per-channel rank bottleneck and the single-series MSE objective together, and the predictable tail modes are no longer sacrificed.

The foundation-augmented reference: the strongest circularity test. Because they are trained outside the MSE family, the three foundation models test the reference circularity of §3.3 from a direction no reweighted supervised model can. We recompute the reference as $R^2_{\star,\mathrm{aug}}$, the pointwise max over the Phase-A proxy *and* the per-mode R^2 of all three foundation models (plus the modern supervised baselines, where available). Had the MSE-family proxy materially underestimated attainable skill, ES built on it would deflate under this augmentation. It does not: the tail reference moves by $+0.006$ to $+0.049$ across the eight datasets (head: $+0.004$ to $+0.082$), against supervised tail-ES values of 0.12–13.84, so every Phase-A number changes by at most 0.05. The augmentation instead *sharpens* the decomposition. On Weather and ETTm2 the unrealized component collapses to ≈ 0, leaving tail-ES there $\approx 100\%$ injected error: leakage, not sacrifice. Traffic and Electricity retain unrealized components of 0.19–0.31 for every supervised model, and this skill is now certified as *realizable*, because the foundation models realize it on the same windows. The sacrifice claim thus rests on exactly the datasets whose tails carry externally verified predictable structure.

7 DISCUSSION AND LIMITATIONS

The real-data reference proxy is conservative by construction, so reported ES underestimates the true bias. The analysis is per channel; joint multivariate mode bases are left to future work. Generative probabilistic forecasters raise the same allocation question at the level of sample pools, where the per-mode skill diagnostic transfers. Corpus-level rebalancing for foundation models and per-mode reweighting for supervised ones are complementary levers on the same failure.

REPRODUCIBILITY STATEMENT

Three choices decide every number, and §3.1 with Appendix B fixes all three: the mode basis is fitted on training targets only, spectra are measured on the raw scale, and the attainable-skill reference is a conservative pointwise-max proxy calibrated against a twin re-run noise floor. Both propositions carry their assumptions and proofs in Appendix A. The supplement at `https://anonymous.4open.science/r/excess-spectral-suppression-6980` ships the run directories behind every table and figure, so each headline number can be re-derived on a CPU without retraining; `REPRODUCTION.md` gives the command for each. Dataset preparation is scripted, but the benchmark CSVs are not redistributed.

AI USE STATEMENT

We have not used generative AI tools for any of the tasks with required disclosure: the hypotheses, the theoretical claims and their proofs, the experimental design, the implementation of the methods, the data preparation and the interpretation of results are the authors' own; synthetic data generation, translation and qualitative data analysis are not applicable to this work. We used generative AI tools for two tasks with recommended disclosure only: reviewing our experiment code, and editing the prose to improve readability. We have reviewed all AI-assisted work: each code-review suggestion was checked and tested by the authors before being adopted, every number reported here was produced by the released code, and all edited text was re-read against the results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, 2019.

Massimo Bonavita. On some limitations of current machine learning weather prediction models. *Geophysical Research Letters*, 51(12), 2024. e2023GL107377.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *International Conference on Machine Learning*, 2024.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*, 2021.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Romain Ilbert, Ambroise Odonnat, Vasilii Feofanov, Aladin Virmaux, Giuseppe Paolo, Themis Palpanas, and Ievgen Redko. Samformer: Unlocking the potential of transformers in time series forecasting with sharpness-aware minimization and channel-wise attention. In *International Conference on Machine Learning*, 2024.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2022.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 95–104, 2018.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Siyi Lin, Chongming Gao, Jiawei Chen, Sheng Zhou, Binbin Hu, Yan Feng, Chun Chen, and Can Wang. How do recommendation models amplify popularity bias? an analysis from the spectral perspective. In *ACM International Conference on Web Search and Data Mining*, 2025.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

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

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

Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Xihao Piao, Zheng Chen, Taichi Murayama, Yasuko Matsubara, and Yasushi Sakurai. Fredformer: Frequency debiased transformer for time series forecasting. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024.

Tianyu Qiu, Yi Xie, Yun Xiong, Hao Niu, and Xiaofeng Gao. Enhancing masked time-series modeling via dropping patches. In *AAAI Conference on Artificial Intelligence*, 2025. arXiv:2412.15315.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, 2019.

Gregory C. Reinsel and Raja P. Velu. *Multivariate Reduced-Rank Regression: Theory and Applications*, volume 136 of *Lecture Notes in Statistics*. Springer, 1998.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *European Signal Processing Conference (EUSIPCO)*, pp. 606–610, 2007.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, 2020.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *International Conference on Machine Learning*, 2024.

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*, 2021.

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

Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *Communications in Computational Physics*, 28(5):1746–1767, 2020.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *AAAI Conference on Artificial Intelligence*, pp. 11121–11128, 2023.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *AAAI Conference on Artificial Intelligence*, pp. 11106–11115, 2021.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*, 2022.

A THEORY

Proposition 1. With $x = Az$, $y = V\mathrm{diag}(\sigma)(\sqrt{\rho} \odot z + \sqrt{1-\rho} \odot \xi)$ and orthonormal A, V, the input–output cross-covariance is $\Sigma_{yx} = V\,\mathrm{diag}(\sigma\sqrt{\rho})\,A^\top$ and $\Sigma_{xx} = I$. The population reduced-rank regression of rank r (Reinsel & Velu, 1998) is $B_r = P_r B_{\mathrm{ols}}$ where $B_{\mathrm{ols}} = \Sigma_{yx}\Sigma_{xx}^{-1}$ and P_r projects onto the top-r left singular subspace of $B_{\mathrm{ols}}\Sigma_{xx}^{1/2}$; here that subspace is spanned by the columns v_k with the r largest $s_k = \sigma_k\sqrt{\rho_k}$. For retained modes the prediction equals the Bayes conditional mean in that mode ($R_k^2 = \rho_k$, $\mathrm{ES}_k = 0$); for dropped modes the prediction is 0 ($R_k^2 = 0$, $\mathrm{ES}_k = \rho_k$). □

Proposition 2. Under the conditions of Saxe et al. (2014) (depth-2 linear net $\hat{y} = W_2W_1x$, whitened inputs, initialization aligned with the task singular vectors and small), each input–output singular mode evolves independently with $s_k(t)$ following a logistic curve from ϵ to s_k with time constant $\tau_k = O(1/s_k)$. Truncating training at t leaves modes with $\tau_k \gg t$ at their (near-zero) initial coupling, reproducing the rank-budget selection with threshold $s_k \sim 1/t$. □

B PROTOCOL DETAILS AND ADDITIONAL RESULTS

Negative control: a dataset with nothing to sacrifice. Exchange-rate data (8 daily FX series, near-random-walk) has proxy tail $R_\star^2 = -0.05$: no attainable tail skill. Measured tail-ES there is enormous for constrained models (225 at r=8; 1.6 at r=96), yet its unrealized-skill component sits at -0.055 at every rank. That is exactly the (negative) proxy level: there is no attainable tail skill

Table 4: Interventions on the two representative skill-bearing channels (§4.2): unrealized tail skill before and after each intervention, with three-seed ranges where the cell was replicated. Width and weight decay raise sacrifice on both channels; dropout moves it in opposite directions.

		unrealized tail skill		3-seed range	
channel	intervention	base	after	base	after
Traffic ch300	width d=64 $\to$ 8	0.05	0.30	—	—
	weight decay 10^{-2}	0.051	0.555	—	0.537–0.555
	dropout 0.5	0.039	0.107	0.02–0.04	0.10–0.12
Electricity ch150	width d=64 $\to$ 8	0.08	0.45	—	—
	weight decay 10^{-2}	0.082	0.683	—	0.675–0.683
	dropout 0.5	0.133	0.060	0.123–0.134	0.056–0.066

to leave unrealized, so all of the measured ES is injected error and none of it reads as sacrificed predictable structure. The protocol therefore does not manufacture "sacrifice" where nothing is predictable. The unrealized/injected decomposition, not raw ES, is what carries the claim — which is why every headline number in this paper is a decomposed one.

Horizon stability. We repeat Phase A on ETTh1 at $H \in \{192, 336, 720\}$, with a K=48 mode basis at each horizon. The complete-sacrifice law at r=8 holds at every horizon: unrealized tail skill tracks attainable tail skill to the third decimal ($0.102/0.110/0.066/0.058$ against attainable $0.102/0.110/0.067/0.058$ at H=96/192/336/720; median per-channel ratio $1.000/1.000/0.999/0.996$). Rank-monotonicity of tail-ES (r=8 $>$ 32 $>$ 96) also holds at every horizon. Attainable tail skill itself thins with horizon, and the injected component falls by an order of magnitude ($1.11 \to 0.12$ at r=8) as tail-mode energies grow, so the tail-vs-head asymmetry attenuates at H=720; the budget mechanism, as measured by the decomposition, does not.

Lookback stability. We repeat Phase A on ETTh1 and Traffic at lookbacks $L \in \{96, 720\}$, with the protocol otherwise unchanged. The complete-sacrifice law is unaffected: under the per-channel statistic of §3.3, the r=8 unrealized-to-attainable ratio has median 1.000 in all four (dataset, L) cells (IQR $[1.000, 1.000]$ in three of them and $[0.998, 1.000]$ on Traffic at L=720; per-channel minimum 0.98). Rank-monotonicity of both tail-ES and its unrealized component holds at every lookback. Longer context *raises* the attainable tail skill on Traffic (mean $R^2_\star$ over the tail $0.18 \to 0.27$ from L=96 to 720), while constrained models realize essentially none of it at any L. The sacrifice is therefore not an artifact of context length: more context only raises what the budget throws away.

Noise-floor calibration. The twin control retrains the reference family's DLinear member with a different seed (901 vs. 900), excludes it from the pointwise max, and reports its tail-mean ES as the floor: how much apparent suppression a model *from the reference family itself* exhibits, from seed noise plus the winner's-curse selection implicit in a pointwise max. Floors (tail means): ETTh1 0.19, ETTh2 0.44, ETTm1 0.39, ETTm2 1.46, Solar 0.08, Traffic 0.25, Weather 1.80, Electricity 0.22. Two reading rules follow. (1) Any claimed per-model tail-ES below its dataset's floor is treated as an upper bound, not evidence of suppression — this retracts, e.g., an apparent DLinear tail-suppression reading on ETT. (2) Because DLinear is typically the *weakest* tail member of the family (on Traffic ch300: DLinear tail R^2 0.22 vs OLS 0.80, ensemble 0.67), the floor is dominated by the member-vs-max architecture gap rather than pure seed noise; comparisons against the floor therefore certify "above the weakest member's re-run", a deliberately weak but unambiguous bar.

Decomposition conventions. ES_k = unrealized + injected is a pointwise identity, so when $R^2_k(f^\star) < 0$ the unrealized term is negative and the injected share can exceed 1 — Solar 1.05 and Weather 1.02, both entered as "≈ 1" in Table 2, where the intended reading is "injection-dominated". Tail/head halves split the mode index range at $K/2$ under the energy ordering.

Modern supervised baselines. Detail behind §6. Neither PatchTST nor iTransformer exceeds its dataset's twin noise floor on dataset means on any of the eight benchmarks (Traffic: 0.099–0.109 and 0.003, floor 0.25; ETTh1: 0.107–0.134 and 0.035, floor 0.19; the largest values anywhere are

Table 5: Tail-ES of three zero-shot foundation forecasters vs. the budgeted supervised families of Phase A (range over architectures and ranks r=8–96), identical protocol. Negative entries beat the pointwise-max proxy. Discussed in §6.

dataset	Chronos-Bolt	Chronos-2	TimesFM-2.5	supervised (Phase-A range)
ETTh1	$+0.023$	$+0.036$	$+0.017$	0.19–1.21
ETTh2	$+0.005$	$+0.002$	$+0.005$	0.27–2.92
ETTm1	$+0.007$	$+0.011$	$+0.010$	0.12–0.40
ETTm2	$+0.028$	$+0.040$	$+0.013$	0.73–13.84
Solar	$+0.033$	-0.026	-0.009	0.14–2.91
Traffic	$+0.018$	-0.006	-0.027	0.17–0.64
Weather	-0.008	-0.003	-0.014	0.97–10.2
Electricity	$+0.046$	-0.028	$+0.013$	0.22–0.70

Weather's $0.69/0.43$ against a floor of 1.80). On both skill-bearing datasets the per-channel model retains more tail-ES than the cross-channel one on the same channels (Traffic $+0.106$, Electricity $+0.029$); the ordering inverts only on injection-dominated ETTm2, where cross-channel attention injects more, and architecture is confounded with training-data volume in this contrast. RevIN *raises* mean tail-ES on Traffic (0.109 vs 0.100) and *lowers* it on ETTh1 (0.107 vs 0.134), with strong per-channel heterogeneity. iTransformer's per-mode profile matches that of the foundation models, and its Traffic head-ES is negative.

Phase C on the injection-dominated benchmarks. Detail behind finding (v) of §5. On Solar, Weather, ETTm1 and ETTm2, predictability weighting backfires exactly as the probe-blindness diagnostic predicts, amplifying injection (tail R^2 at r=32, MSE $\rightarrow$ $\hat{\rho}$-weighting: Solar $-0.32 \rightarrow -1.96$, ETTm2 $-1.49 \rightarrow -3.56$). The prescribed γ=0 loss instead cuts injected tail error 2–3$\times$ (tail R^2 at r=32: Solar -0.13, Weather -1.91, ETTm1 -0.07, ETTm2 -0.71) *and* improves overall MSE in 7 of 8 (dataset, width) cells, the exception being a 0.1% tie on Weather at r=32. No loss pushes tail R^2 above the proxy on more than 2% of tail (channel, mode) pairs, as attainable ≈ 0 requires.

Hyperparameter sensitivity. The sweep of §5 covers $\alpha \in \{0.5, 1, 2\}$, $w_{\max} \in \{4, 10, 30\}$ and $\rho_0 \in \{0.02, 0.05, 0.1\}$, varied one axis at a time around the blend's defaults, on Traffic and ETTh2 with the MLP at r=96. All 14 configurations beat MSE on tail R^2 (Traffic $+0.211$ to $+0.335$ vs. $+0.149$ for MSE; ETTh2 -0.127 to -0.269 vs. -0.317), and α=2 is best on both datasets. The sweep trains on every window (stride 1), whereas the Traffic column of Table 3 uses stride 2. Absolute tail R^2 on Traffic is therefore higher here and comparable only within the sweep, against its own MSE baseline. ETTh2, run at stride 1 in both, agrees exactly (-0.317 in both).

Probe details. The ridge probe fits $x \mapsto Y v_k$ per mode with λ chosen on the validation split; $\hat{\rho}_k = \max(0, R^2_{\text{val},k})$, with the validation split taken as the chronologically last fraction of the training range (an iid split inflates $\hat{\rho}$ under distribution shift; kept as an ablation).

Two protocol variants are deliberately deferred: the normalized-vs-raw (RevIN-style) contrast, which changes which modes count as low-energy without changing the sacrifice mechanism, and joint multivariate mode bases, which entangle cross-channel correlation with the per-channel allocation question studied here.