# PRISM: Patch-based Regression with an Interpretable Shape–Scale–Level Model for Time-Series Forecasting

**Anonymous authors**
Paper under double-blind review

## Abstract

Long-horizon time-series forecasting requires representations that preserve multi-scale temporal structure while adapting to changes in amplitude and level across windows. We study PRISM, a lightweight, channel-independent forecaster that composes established components, a multi-scale patch encoder with a shared U-Net, a learned-query direct (non-autoregressive) decoder, and an explicit output head $\hat{\mathbf{y}} = \text{shape} \cdot \text{scale} + \text{level}$. None of these components is individually novel. Rather than claim state-of-the-art accuracy, we ask a set of answerable questions: does the explicit decomposition mean what its names say, does multi-scale patching help, where does the model help on real data, and does controlled synthetic evidence support the mechanism.

Our load-bearing result is a controlled, three-seed mechanism study: under a gauge-aware analysis on synthetic data with known generative factors, the explicit shape · scale + level *head* recovers the generative factors, the scale head tracks the true amplitude where amplitude varies (across-sample $R^2 = 0.73$, mean of 3 seeds), the level head tracks the true offset where offset varies ($R^2 = 0.99$; 0.94 under joint regime shifts), and both stay near zero where their factor is absent. This recovery is a property of the *head*, not the multi-scale encoder: a single-patch control sharing the same head attains near-identical recovery ($R^2 = 0.67/0.98/0.93$), so it is backbone-robust. This is our only multi-seed result and we treat it as the headline. The *per-head raw* recovery is weaker than the gauge-aware $R^2$ (e.g. S2 scale regression slopes are 0.36–0.46, not $\approx 1$); we report both metrics. We also run the control this claim requires. Because the scalar heads are fed raw trailing-context statistics, we regress each true factor on the single corresponding statistic with no model at all: that trivial baseline *beats* the trained head on recovery magnitude wherever the factor varies (0.91 vs 0.73 on S2), so magnitude is not evidence of learned structure. What does not reduce to the statistic is selectivity — the trivial read-out fires indiscriminately where amplitude is absent (mean $R^2 = 0.46$) whereas the head is near-silent there (mean 0.04), a tenfold gap. We therefore state the mechanism claim as conditional selectivity rather than discovery of the factors, and this same pass-through structure explains why a faithful head turns out not to be a necessary one.

On real data the picture is more nuanced. Point estimates are seed 1, with standard deviations over 2–3 seeds per cell (3 seeds for 51 of 68 cells, 2 for 16, 1 for one cell), so we read WIN/LOSS verdicts as indicative rather than significance-tested. PRISM's strength here is *targeted rather than universal*: it matches or beats the best of six strong baselines on a characterizable high-trend, low-volatility niche (most clearly ETTm2, 3.5–13.9% better), while remaining competitive-or-better on a minority of cells overall and not surpassing the field in aggregate; the ETT wins additionally carry a train-split asterisk (0.8 vs 0.75) we flag but do not fully resolve. We regard the negative results as equally informative: the synthetic multi-scale advantage does not transfer, added capacity yields little, and gating, learnable-Fourier, period-folding, and channel-mixing prove inert or harmful; a necessity probe finds the level head necessary where level varies but the scale head



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.

not demonstrably necessary, so faithfulness does not imply necessity. As supporting evidence that the design is not tied to long horizons, a mask-aware member of the same family transfers to the short-series M4 setting, reaching 0.8125 weighted OWA over three seeds (ahead of the competition winner ES-RNN at 0.821, and within roughly 0.015 of N-BEATS at 0.797); we report this as a generalization check rather than a competition result. The contribution is new *knowledge*, not a new *component*: a verifiably faithful decomposition under controlled factors, an account of where a small patch forecaster helps and where it does not, and negative results that constrain what it buys.

## 1 Introduction

**The problem.** Long-horizon time-series forecasting (LTSF) maps a context window of length $L$ to a horizon of length $H$, with $H$ often as large as $L$. Two difficulties dominate. First, useful signal lives at several temporal scales at once (a daily cycle rides on a weekly trend), so a forecaster must represent multi-scale structure. Second, the amplitude and mean of a series drift between windows, so a model tuned to one scale/level regime can fail on another. The prevailing recipe addresses these with *direct* (non-autoregressive) prediction of the whole horizon, *patch* tokenization of the context, *cross-attention* decoding, and instance normalization or explicit decomposition of the target (Zhou et al., 2021; Wu et al., 2021; Nie et al., 2023; Zeng et al., 2023; Kim et al., 2022; Liu et al., 2024).

**The gap.** These components are effective and well studied, but two questions are rarely answered directly. (i) When a model outputs an explicit decomposition (a shape modulated by a scale and offset by a level), *does the decomposition mean what it says*, in the sense that each part recovers the generative factor it is named for? (ii) For a small, transparent composition of these components, *where does it actually help*, and do the mechanisms that explain it in a controlled setting transfer to real data? We take these as the object of study rather than pursuing a benchmark ranking.

**What PRISM is.** PRISM (Patch-based Regression with an Interpretable Shape–Scale–Level Model) is a channel-independent univariate core applied to every variate. It patches the context at scales $\mathcal{P} = \{4, 8, 16\}$, runs each scale through a *shared* two-level U-Net with a small learned-query attention bottleneck, concatenates the resulting tokens as memory, and decodes with $H$ learned future queries through a single transformer decoder layer. Its output is $\hat{\mathbf{y}} = \text{shape} \cdot \text{scale} + \text{level}$, where *shape* is a horizon-length waveform from the decoder, and *scale* (positive) and *level* are scalars-per-window conditioned on pooled memory plus five trailing-context statistics. Section 3 specifies the model exactly against the released reference implementation.

**Novelty, stated up front.** We claim *no component novelty*. Patching and multi-scale patching (Nie et al., 2023), U-Net encoders, rotary position embeddings (Su et al., 2024), query-based cross-attention decoding (Zhang & Yan, 2023; Liu et al., 2024), channel independence, and direct multi-horizon forecasting are all prior art. Our contribution is *new knowledge*: (i) a verifiably faithful shape–scale–level decomposition demonstrated under controlled generative factors (three seeds); (ii) the finding that the synthetic multi-scale advantage does not transfer to real data; (iii) a correlational characterization of where the model helps (a high-trend niche); and (iv) negative results on gating, learnable-Fourier, period-folding, and channel-mixing.

**Claim hierarchy.** Our evidence differs in strength across research questions, and we state which claims are multi-seed. The *only multi-seed* evidence supports **RQ4** (controlled-synthetic faithfulness) and is the load-bearing claim. Every real-data result (RQ1, RQ2, RQ3, RQ5) uses seed-1 point estimates with three-seed spread; the necessity probe (RQ2) is additionally single-horizon. The canonical statement of this caveat is in Section 4.

**Research questions.** **RQ1 (Does multi-patch help?)** On real LTSF, does the multi-scale patch encoder beat a matched single-patch encoder at equal capacity, look-back, and protocol? **RQ2 (Does shape–scale–level help?)** Do the scale and level heads recover, and are they necessary for, their named factors? **RQ3 (On what processes?)** Where is PRISM competitive-or-better on real data, and can that region be described by dataset characteristics? **RQ4 (Controlled mechanism.)** On synthetic data with *known* generative factors, do the heads recover exactly the factor their condition varies, and stay inert otherwise? **RQ5 (Efficiency.)** What are PRISM's measured parameters and throughput?

**Contributions.** (1) A precise, code-faithful description of a lightweight multi-scale patch forecaster with an explicit shape–scale–level head, framed as a *composition* of prior components (Section 3). (2) A controlled, three-seed, gauge-aware mechanism study establishing that the decomposition is faithful to known generative factors with raw-head recovery limits disclosed numerically (Section 7). (3) A study on 44 real LTSF cells (single seed): PRISM's advantage is narrow rather than broad, it matches or beats the best of six baselines only on a high-trend, low-volatility niche (most clearly ETTm2, 3.5–13.9% better) and is not SOTA in aggregate; the full per-cell accounting is in Section 5. Alongside this we report that multi-patch does not transfer, capacity barely helps, and several popular components are inert or harmful (Sections 5, 6). (4) A correlational characterization of PRISM's niche and a single-seed necessity probe whose result partly *contradicts* the decomposition hypothesis (Sections 8, 9). (5) A generalization check, not a separate contribution: a mask-aware member of the same family transfers to the short-series M4 setting (Makridakis et al., 2020), reaching 0.8125 weighted OWA over three seeds (ahead of ES-RNN's 0.821, REPORTED; within roughly 0.015 of N-BEATS's 0.797), which indicates the architecture is not specific to long horizons (Appendix L).

## 2 Related Work and Positioning

**Direct forecasting, patching, and query-based decoding.** A dominant line predicts the full horizon in one pass, avoiding iterative error accumulation (Zhou et al., 2021; Wu et al., 2021); lightweight models such as SparseTSF (Lin et al., 2024) and TimeMixer (Wang et al., 2024) show simple supervised structures can rival large attention models. Patch tokenization groups consecutive points into embedded segments (Nie et al., 2023) and enables channel independence, and several models decode with query tokens attending to context (Zhang & Yan, 2023; Liu et al., 2024). PRISM adopts this setting unchanged (patch lengths $\{4, 8, 16\}$, shared U-Net, one query per horizon step, RoPE (Su et al., 2024) on keys and queries); we *study* rather than assume whether multi-scale helps on real data (RQ1), reporting that it does not.

**Decomposition, normalization, and level–scale conditioning.** Making the target tractable by decomposition or input normalization is a recurring theme: DLinear (Zeng et al., 2023), N-BEATS (Oreshkin et al., 2020), N-HiTS (Challu et al., 2023), Autoformer (Wu et al., 2021), FEDformer (Zhou et al., 2022), and RevIN (Kim et al., 2022). PRISM connects at two points: per-window standardization (a form of instance normalization, no learned denormalization head) and the explicit shape · scale + level head, in which *level* plays a trend/offset role and *scale* an amplitude-adaptation role. This head is inspired by, not proven superior to, the decomposition literature; a direct single-head-vs-decomposition ablation on real data (shape-only vs full) is reported in Section 6, and finds the decomposition does not improve real-data accuracy at $n = 1$. We also distinguish our aim from the interpretable basis-decomposition line: N-BEATS (Oreshkin et al., 2020) and N-HiTS (Challu et al., 2023) offer interpretable trend/seasonality bases but, to our knowledge, do not measure whether those bases are *faithful* to known generative factors. Our gauge-aware recovery test (Section 7) is a methodology to *check* such claims, not a superiority claim. All building blocks are prior art; PRISM's contribution is the composition plus the mechanism study and the empirical map.

## 3 Method

PRISM is a direct, single-pass univariate core $f_\theta : \mathbb{R}^L \to \mathbb{R}^H$ that predicts the entire horizon at once (not step-by-step), applied independently to every channel (no cross-channel attention, mixing, or shared statistic). Each context window is standardized to zero mean and unit variance using only that window's own statistics, and predictions are denormalized before evaluation, so no future information enters the input scaling. This section is exact to the released reference implementation; full per-dimension details and complexity are in Appendix A.

**Multi-patch encoder and shared U-Net.** For each patch length $p \in \mathcal{P} = \{4, 8, 16\}$ with stride $\lfloor p/2 \rfloor$ (50% overlap), the replicate-padded context is unfolded into patches, each embedded by a *dedicated* linear map $E_p : \mathbb{R}^p \to \mathbb{R}^d$. The three scales' token sequences are concatenated along the token axis to form memory $Z \in \mathbb{R}^{B \times P \times d}$. A single, shared two-level U-Net (Ronneberger et al., 2015) (`_run_one`) is reused across all three scales; each level is a residual conv block (Conv1d–GroupNorm–GELU–Conv1d, kernel 3) with width fixed at $d$ throughout, no channel doubling, and

pooling mean-averages adjacent token pairs. A learned-query bottleneck ($n_q = 4$, 4 attention heads) cross-attends to the deepest tokens. The bottleneck is part of the default configuration; its individual contribution is outside the scope of the ablations we report here.

**Direct decoder (single layer, RoPE both sides).** The horizon is emitted in one pass from $H$ learned future queries. RoPE is applied to both the context keys and the future queries, with context keys using relative positions (last key at index 0) and future queries absolute positions. The decoder is a single `nn.TransformerDecoderLayer` (width $d_h$, 8 heads); the `nlayers=2` kwarg does *not* stack it. It executes self-attention over the queries, cross-attention to memory, and an FFN (not pure cross-attention).

**Shape–scale–level head.** A per-step linear head plus a 1-D conv refine stack produce the shape waveform $S \in \mathbb{R}^{B\times H}$. The scalar heads are conditioned not on the decoder hidden state but on the encoder memory mean-pooled over context tokens, $z_{\text{pool}} = \frac{1}{P}\sum_{i=1}^{P} Z_i$, concatenated with five trailing-context statistics (means, last value, std, two trend deltas; Appendix A). Two MLPs give the level $B(X)$ and, when `scale_aware=True`, the positive amplitude $A(X) = \text{clamp}_{\geq 10^{-3}}(\text{softplus}(\cdot))$. Broadcasting across the horizon,

$$\hat{Y} = S(X) \odot A(X) + B(X), \qquad S \in \mathbb{R}^{B\times H},\ A, B \in \mathbb{R}^{B\times 1}. \tag{1}$$

When `scale_aware=False` this reduces to $\hat{Y} = S(X) + B(X)$. Because the scalar heads read pooled memory and raw trailing statistics, not the per-step decoder state that drives $S$; the head is *not* a trivial affine rescaling of a single regressor; we nonetheless do not claim it beats a single head on real data.

**Identifiability caveat.** The product $S \odot A$ is not uniquely identifiable: for any $c > 0$, $(cS, A/c)$ leaves $S \odot A$ unchanged. We keep three notions separate: (1) forecast accuracy (gauge-invariant); (2) contribution-level recovery (does a head's across-sample variation track its factor's, after removing the gauge); (3) raw per-head correlation (meaningful only up to the gauge). All interpretability claims in Section 7 are stated at level (2), gauge-aware, with level (3) disclosed numerically.

Figure 1 (Appendix A) shows the complete four-panel architecture; the description below is exact to the released reference implementation.

## 4 Experimental Setup

**Datasets and protocol.** We evaluate on 11 datasets spanning four domains energy/industrial (Wind, ZafNoo, CzeLan), weather (Weather), finance (Exchange), temperature (ETTh1, ETTh2, ETTm1, ETTm2), and air quality (AQShunyi, AQWan), under the TFB rolling-forecast protocol (Qiu et al., 2024) (stride = 1), each at $H \in \{96, 192, 336, 720\}$ (44 cells). We scope the study to these 11 series and leave the large-scale Electricity/Traffic/Solar/PEMS benchmarks to future work. Windows are built over a strictly chronological split; each window is standardized using only its own context statistics (no future leakage); predictions are denormalized to original units before computing MSE/MAE; the best-validation checkpoint is used (no test-set model selection).

**Model and training.** The main model, PRISM-mp-247k, uses the default configuration: `plens=[4,8,16]`, $d_{\text{model}} = 48$, $d_{\text{hid}} = 96$, 8 heads, dropout 0.1, `scale_aware=true`. Optimization: AdamW (lr $10^{-4}$, wd $10^{-4}$), cosine schedule, Huber loss, batch 16 with grad-accum 2, bf16 AMP, early stopping (patience 5, max 200 epochs). Real-LTSF point estimates use seed = 1, with three-seed runs (seeds 1–3) complete for most cells, with per-seed standard deviations reported in Appendix C; no significance is claimed anywhere on real data. We use `train_ratio_in_tv=` 0.8 for all datasets, a deliberate anti-overfitting choice that differs from the OpenTS baselines' 0.75 (ETT)/0.875 (default); see caveats below.

**Baselines and provenance.** We compare against six reported baselines: DLinear, SparseTSF, iTransformer, PatchTST, TimeMixer, and TimesNet (Wu et al., 2023). All baseline MSE/MAE are the published OpenTS-Bench multivariate LTSF results (raw/denormalized), retrieved from the OpenTS query endpoint (leaderboard last-modified 2025-11-26; see the baseline-provenance record accompanying the release). *These baselines are REPORTED, not reproduced by us; PRISM numbers are NEW.* We never mix provenances in a single verdict. Comparability caveats (Appendix B): look-back is matched to each cell's winning baseline (44/44) but not iso-look-back across all six; PRISM's train split (0.8) differs from the baselines' (0.75/0.875) where PRISM's wins concentrate (ETT), so those wins carry a split asterisk, though the ETTh2/ETTm2 advantage is stable across many variants

Table 1: **Main results: head-to-head vs. six reported baselines.** Each cell is mean %-vs-best-baseline across the four horizons (per-dataset; 0.0 marks the row's best baseline, bold; lower is better; raw MSE spans $\sim 10^3\times$ across datasets so we compare only relative to the row best, never raw means). PRISM's last column: underlined where PRISM is competitive-or-better ($\leq 5\%$ of the best baseline). PRISM leads on ETTm2 and is level with the best on ETTh2/Wind; losses concentrate on high-volatility and large-multivariate series. Full per-horizon MSE/MAE in Appendix C.

| Dataset | DLinear | SparseTSF | iTrans. | PatchTST | TimeMixer | TimesNet | **PRISM** |
|---|---|---|---|---|---|---|---|
| ETTm2 | **0.0** | +0.9 | +6.7 | +1.3 | +2.7 | +18.3 | −9.3 |
| ETTh2 | +12.8 | **0.0** | +9.0 | +1.6 | +4.7 | +14.3 | +0.9 |
| Wind | **0.0** | +10.0 | +7.8 | +6.0 | +8.7 | +13.6 | +1.1 |
| Weather | +25.7 | +7.8 | +10.7 | **0.0** | +2.5 | +62.4 | +16.8 |
| ETTh1 | +17.5 | **0.0** | +10.9 | +8.3 | +9.4 | +22.0 | +54.7 |
| ETTm1 | +5.5 | +4.3 | +2.5 | **0.0** | +1.7 | +16.4 | +25.5 |
| Exchange | **0.0** | +8.3 | +8.3 | +8.3 | +8.3 | +16.7 | +88.0 |
| AQShunyi | **0.0** | +3.4 | +0.4 | +2.3 | +0.1 | +5.0 | +12.1 |
| AQWan | +2.6 | +6.7 | +1.5 | +4.6 | **0.0** | +1.9 | +17.3 |
| CzeLan | +16.0 | +10.4 | **0.0** | +4.0 | +1.0 | +2.7 | +34.6 |
| ZafNoo | +6.9 | +3.1 | +15.1 | **0.0** | +2.4 | +18.9 | +12.0 |
| **Competitive-or-better on 12/44 cells (27%); strong on a characterizable high-trend niche, not SOTA in aggregate.** | | | | | | | |

of the same core under an identical 0.8 protocol. Raw MSE spans $\sim 10^3\times$ across datasets, so we aggregate only via per-cell %-vs-best and verdict counts, never raw-MSE means.

## 5 Where PRISM Helps on Real Data

PRISM's advantage on real data is narrow and structured rather than broad: the cells where it is competitive-or-better cluster on a recognizable high-trend, low-volatility niche rather than scattering at random. Against the per-cell best-of-six baseline over all 44 cells, PRISM-mp-247k is competitive-or-better on 12/44 = 27% (7 WIN / 5 COMP at $\leq 5\%$ / 32 LOSS); **PRISM is not state of the art.** Table 1 summarizes; the full per-horizon table is in Appendix C. One caveat applies to every verdict: each baseline is compared at *its own* tuned look-back ($L \in \{96, 336, 512, 528\}$) while PRISM uses a *fixed* $L$ per cell (the winning baseline's), so this is not iso-look-back, and a fixed-$L$ head-to-head across all six is not yet run.

**Where the wins sit.** The wins are not scattered: ETTm2 is 4/4 WIN ($-3.5\%$ to $-13.9\%$ vs the best baseline, DLinear); ETTh2 is 1 WIN + 3 COMP (essentially tied at every horizon); Wind is competitive-or-better on 3/4 horizons. The winning datasets are *associated with* high trend and low volatility; the loss datasets are high-volatility, low-trend, or large-multivariate air-quality/hydrology series. We treat this as a correlational observation (quantified in Section 9; full characterization in Appendix J), not a causal mechanism, and reiterate the ETT-family split caveat.

**Horizon sensitivity.** Like all direct multi-horizon forecasters, PRISM's error grows as the horizon lengthens. Measured against each dataset's own $H = 96$ MSE (single seed; Appendix D), PRISM-mp-247k degrades by +14/+22/+40% on ETTh1, +21/+31/+50% on ETTh2, +21/+28/+43% on ETTm1, and +33/+59/+125% on ETTm2 at $H = 192/336/720$, with the steepest degradation on ETTm2 (+125% at $H = 720$). The "long-horizon" label thus denotes the *task* (horizons up to 720), not graceful scaling of PRISM's accuracy with $H$.

### 5.1 Generalization Check: Short-Horizon M4

To test whether the design is specific to long horizons, we apply the same PRISM core at the opposite end of the horizon spectrum. M4 series are short and variable-length, so a forecaster must ingest front-padded contexts; we therefore run PRISM-Masked, the PRISM architecture with a mask-aware encoder (masked U-Net convolutions, pooling, attention, RevIN) so padded positions carry no signal. It is the same family: it keeps PRISM's shape · scale + level head, RoPE, and U-Net topology, and runs PRISM's *single-patch* configuration ($p = 8$, stride 4; `plens=[8]` in our code) rather than the multi-scale $\{4, 8, 16\}$ used for LTSF. On a fixed window the masking reduces to identity. That the strongest short-horizon result comes from the single-patch configuration is consistent with RQ1, where multi-scale patching also fails to beat matched single-patch on real LTSF data.

**The same design transfers to short series.** We report M4 as a generalization check on the architecture, not as a competition entry. On OWA (lower is better; Naive2 = 1.000), PRISM-Masked reaches 0.719 ± 0.007 on Yearly and 0.785 ± 0.002 on Quarterly as a single model averaged over three seeds (Table 13 in Appendix L), below *both* the M4 winner ES-RNN (Smyl, 2020) (0.778/0.847, REPORTED) *and* N-BEATS (0.752/0.814; Oreshkin et al., 2019). Yearly and Quarterly are the two most heavily weighted frequencies (23,000 + 24,000 series, 47% of M4), so this is where the strength matters most. PRISM-Masked also beats ES-RNN on Weekly (0.734 vs 0.851), Daily (0.972 vs 1.046), and Hourly (0.416 vs 0.440), improving on the competition winner in five of six groups.

**Where it falls short.** The one group PRISM-Masked loses is Monthly (0.861 ± 0.003 vs N-BEATS's 0.819), the largest by count (48,000 series), which pulls the three-seed weighted total to 0.8125 ± 0.002: ahead of ES-RNN's 0.821 but behind N-BEATS's 0.797 by roughly 0.015. The reference figures are REPORTED, not reproduced in our harness. The same decomposition we probe on long horizons thus transfers to short, heterogeneous series once padding is handled.

## 6 Ablations: Multi-Patch, Capacity, and Cost

All ablation cells are single seed. Full tables (patch, capacity) are in Appendix E.

**RQ1: Multi-patch does NOT beat single-patch on real data (negative).** On controlled synthetic scale-varying data (condition S2, only amplitude varies) the multi-scale encoder improves forecast MSE by ≈29% over a *matched* single-patch model (same core, matched to within 0.43% on parameters, identical look-back/protocol/training, **only `plens` differing**; 3-seed mean). On real LTSF this advantage does not transfer: over the 17 cells where both multi-patch (mp) and single-patch-8 (sp8) exist at equal 247k capacity, mp is better in only 6/17; sp8 wins 11/17 (fair comparison: shared budget, look-back, protocol, seed; only `plens` differs). Whatever value PRISM has on real data comes from elsewhere in the model, since the multi-patch mechanism is where the synthetic-to-real gap opens.

**Capacity: small, mixed (247k is the main model).** Over the 24 cells where a genuine mp-423k run exists, the 423k model is better in 14/24 cells, mean improvement only −1.63%. Larger gains appear on harder/longer horizons (Exchange H720 −29.8%, H336 −17.9%; Weather H336 −12.6%) while 423k *hurts* on others (Exchange H96 +23.3%, ETTh2 H720 +5.3%). We report 247k as the main model; comparison is fair (only width differs). Context length is inconclusive: PRISM's look-back was matched to the winning baseline per cell (44/44), and a controlled $L$-sweep (which would be needed to attribute gains to look-back) is left to future work.

**Optional components: inert or harmful.** Under the default configuration, gating, learnable-Fourier, and period-folding produced no reliable improvement and were disabled; channel-mixing degraded markedly relative to channel independence (e.g. ETTh1 16.2/15.5 vs CI 9.98, H96 diagnostic; full grid in Appendix E). We therefore describe PRISM as a *deliberately simple composition*. The U-Net bottleneck and encoder are retained as default components; isolating their individual contributions is beyond the ablations reported here.

**Single-head vs decomposition on real data (single seed).** We ran the real-data control Section 2 framed as open: a *shape-only* head (decomposition removed, `scale_aware=false` *and* `use_level=false`, so both scalar heads drop and $\hat{Y} = S(X)$) vs the full shape · scale + level head (seed = 1, $H = 96$, raw MSE; Appendix F). Removing the decomposition does not hurt and slightly *improves* MSE: ETTh1 −7.1%, ETTh2 −7.4%, ETTm2 tied (+0.02%); ETTm1 shape-only better (−1.6% *pending*). At $n = 1$ the decomposition does not improve real accuracy over a shape-only head, consistent with the "faithful ≠ necessary" theme and the GATE-3 finding that the scale head is not demonstrably necessary (Section 8).

## 7 What Each Head Represents: A Controlled Mechanism Study

This is the paper's load-bearing, only multi-seed result. The question is mechanistic, not predictive: *does each head of $\hat{y}$ = shape · scale + level recover the ground-truth generative factor it is named for, and stay inert where that factor is absent?* No SOTA or accuracy claim is made. Each synthetic window is built additively and exactly as $y = \text{level}_\star + \text{scale}_\star \cdot \text{shape}_\star + \text{noise}_\star$, storing the ground-truth parts, across seven conditions that isolate each factor (S1 gentle; S2 scale-varies; S3 level-varies; S4

multi-frequency; APERIODIC off-distribution; HIGH_NOISE; REGIME_SHIFT level *and* scale). Full condition specification, generator self-checks, and dataset content hash are in Appendix G.

**Faithfulness (gauge-aware, 3 seeds).** For each condition we regress, across samples, each recovered head against the true factor after resolving the gauge, and average over 3 seeds (Section 3 identifiability caveat, level 2).

**Headline finding.** Each head recovers exactly the factor its condition varies and stays near zero otherwise: S2 `scale_R`$^2$= 0.73 (level $\approx$ 0); S3 `level_R`$^2$= 0.99 and REGIME_SHIFT `level_R`$^2$= 0.94 (scale lower, 0.41); S1/S4 both scalar $R^2 \approx 0$; APERIODIC `shape_corr`= 0.56 (graceful, not catastrophic). The decomposition is therefore faithful when measured gauge-aware: the level head recovers strongly ($R^2 \geq 0.94$ where level varies) and the scale head recovers only moderately (S2 $R^2 = 0.73$, raw slopes 0.36–0.46). Being the only 3-seed result, this is our most defensible mechanistic claim (RQ4). We use "faithful" at the gauge-aware contribution level and avoid claiming a clean per-head disentanglement.

**Faithfulness is a property of the head, not the encoder.** The recovery is a property of the explicit shape · scale + level head rather than of the multi-scale encoder: a single-patch variant of the same core (v1, `plens=[8]`) attains near-identical gauge-aware recovery (S2 scale $R^2$ 0.67 vs 0.73; S3 level $R^2$ 0.98 vs 0.99; REGIME level $R^2$ 0.93 vs 0.94; Appendix G). This is the natural control (a simpler backbone with the same head), and it shows the recovery is head-driven and backbone-robust, not an artifact of PRISM's encoder. The complementary control that isolates the head against a shape-only (decomposition-removed) model on real data is reported in Section 6 (MC-3; Appendix F), with the single-seed caveat noted in Section 10.

**Per-seed spread on the headline numbers.** The gauge-aware $R^2$ is stable on the two strong recoveries and noisier on the off-distribution APERIODIC condition; per-seed values are in Appendix G.

**Raw recovery.** The gauge-aware $R^2$ is a chosen metric; the *raw* per-head recovery is weaker. In S2 the raw scale regression slopes are 0.36/0.38/0.46 across seeds (not $\approx$ 1), and in S3 the raw `shape_product_corr` swings 0.66 $\rightarrow$ 0.90 $\rightarrow$ 0.78. Faithfulness is thus a *contribution-level* property with per-seed noise in the raw slopes; it should not be read as exact per-head identifiability.

**Trivial-baseline control.** The scalar heads are not conditioned on the encoder alone. Following the locked configuration, `head_level` and `head_scale` each receive `[zpool, tf]`, where `tf` is five trailing-context statistics computed directly from the raw context: the mean and standard deviation of the last 48 steps, the last observed value, and two drift terms (`src/model.py`, `_trailing_feats`). Two of these are near-direct proxies for the generative factors we are testing recovery against, because the synthetic context is supplied unnormalized to the probe. A high $R^2$ is therefore *not by itself* evidence of a learned disentanglement: it could be a linear read-out of a supplied input feature. We test this explicitly. Table 7 regresses each true factor against the single corresponding raw statistic, with no model and no training, using the identical OLS $R^2$.

(Table 7 is in Appendix G.)

Two conclusions follow, and they cut in opposite directions. First, the *magnitude* of recovery is not the contribution: where scale varies, the trivial statistic is a *better* scale tracker than the trained head (0.908 vs 0.733 on S2; 0.856 vs 0.413 on REGIME_SHIFT), and it is also better calibrated (S2 slope 0.96 vs the head's 0.36–0.46). The level result is likewise matched by mean($x_{-48:}$) (0.968 vs 0.994 on S3). We therefore withdraw any reading of the headline $R^2$ as showing that PRISM *discovers* amplitude or offset; on this benchmark those quantities are recoverable by a two-line statistic, and the level head's recovery in particular is largely attributable to a supplied input feature.

Second, what does *not* reduce to the statistic is selectivity. The trivial scale read-out is indiscriminate: it reports $R^2$ = 0.34–0.69 in the five conditions where amplitude does *not* vary (mean 0.455), because std($x_{-48:}$) responds to any source of dispersion, including shape and noise. The trained scale head does not: it reports mean $R^2 = 0.043$ there, a tenfold reduction, while still reaching 0.733 in the one condition where amplitude genuinely varies. The head has learned *when* the scale pathway applies, which the input feature cannot express. This conditional selectivity, not the peak $R^2$, is the claim the mechanism study supports. We note the control is less favourable for the level head, which is marginally *less* selective than the raw statistic (0.100 vs 0.056); the level channel is where the pass-through concern is strongest, and we do not claim it as evidence of learned structure.

Third, this control also explains the otherwise puzzling result of Section 8, that a faithful head need not be a necessary one. If the scale and level pathways largely relay statistics that are already available to the network through other routes, then removing them should be close to harmless, which is what the necessity probe finds. Faithfulness and necessity come apart here for a concrete, mechanical reason rather than a mysterious one.

## 8 Are the Heads Necessary? A Single-Seed Probe

Faithfulness (Section 7) shows a head *tracks* its factor; it does not show the head is *necessary*. The necessity probe tests this by training per-condition specialist arms (`full`, `no_scale`, `no_level`) and measuring $\Delta$ = MSE(ablated) − MSE(full), so **positive $\Delta$ = head necessary**. This is a **single-seed (seed 0), single-horizon ($L = H = 96$)** probe; we treat $|\Delta\%| < 5\%$ as noise (full delta table and figure in Appendix H).

**Hypothesis vs finding.** The level → S3 link is the one clean necessity: removing the level head worsens S3 MSE by +21.3%, the largest trustworthy effect, matching the faithfulness result. The scale → S2 link is not supported (−1.1%), and level → REGIME_SHIFT is not trustworthy (+3.8%). Two large effects run *opposite* to the hypothesis: the level head is harmful in APERIODIC (−13.2%) and the scale head harmful in S3 (−12.7%, the level head's own regime). A head can faithfully track its factor yet not be necessary, or even harmful; S3 is the only regime where both axes agree. (Scope: $n = 1$, per-condition specialists, single $L/H = 96/96$, no SOTA claim.) A pooled model trained jointly across all seven conditions matches-or-beats the specialists on 6/7 synthetic conditions, with negative interference on the off-distribution APERIODIC condition; see Appendix I.

## 9 Discussion

**In order of evidential strength.** Our strongest result concerns *what the model represents*: under a gauge-aware, three-seed analysis each head recovers the generative factor its condition varies and stays inert otherwise (RQ4). The necessity probe (RQ2) reinforces this only partially and *contradicts* it in places (level faithful and necessary in S3; scale faithful yet not demonstrably necessary). A real-data control confirms the axes are distinct: a shape-only head ties-or-beats the full decomposition, so faithful does not mean necessary.

**On real data, the mechanism behaves differently than the synthetic story predicts.** PRISM is competitive-or-better on ∼27% of cells, associated with high-trend, low-volatility datasets (RQ3). Joining per-cell verdicts to TFB-provided characteristics (correlational only; 6/11 datasets, 24/44 cells; Appendix J), the trend signal holds (0.87 WIN vs 0.29 LOSS) but the volatility signal inverts (PRISM wins the *lower*-volatility datasets, 0.47 WIN vs 0.62 LOSS), and the synthetic multi-scale advantage does not reproduce. The synthetic→real mechanism therefore transfers on 1/3 of its axes (trend holds; periodicity is null; volatility inverts): a mechanism validated in a controlled setting can mislead if assumed to explain real-world performance.

**Small, simple model.** Capacity barely helps, context length is inconclusive, and the optional components are inert or harmful, so PRISM is a *deliberately simple composition*. At ∼0.25M parameters and ∼2040 samples/s (Appendix K) it is markedly smaller than several heavier baselines (RQ5).

## 10 Limitations

**No significance testing on real-LTSF results.** Point estimates are seed = 1 with standard deviations over 2–3 seeds per cell (3 seeds for 51 of 68 cells, 2 for 16, and a single seed for one cell; Appendix C); we report no p-values or CIs, so the 27% niche is *indicative* and small-margin verdicts can lie within seed spread. Ablation cells remain single seed, and M4 seed ensembling shifts OWA by ≈0.012 (Appendix L). The scalar heads receive factor-proxy input features. The level and scale heads are conditioned on five raw trailing-context statistics in addition to pooled encoder memory, and the synthetic probe supplies context unnormalized. A trivial one-feature regression matches or exceeds the heads' recovery *magnitude* in every condition where the factor varies (Table 7). Our faithfulness claim is consequently restricted to conditional *selectivity*, and we have not run the decisive ablation (retraining with the trailing statistics removed) that would fully separate learned rep-

resentation from feature pass-through. We regard this as the most important missing experiment in the paper. **Faithfulness is gauge-aware and synthetic.** The headline $R^2$ is measured after removing the identifiability gauge; raw per-head slopes are weaker (0.36–0.46 in S2). It shows the heads *can* recover known factors, not that real series decompose into them. **Decomposition value is unresolved on real data.** At seed = 1, $H = 96$, a shape-only head ties-or-beats the full decomposition on all four datasets (−7.1/−7.4/−1.6%, ETTm2 tied; Appendix F), and the necessity probe ($n = 1$, single-horizon) partly contradicts the faithfulness hypothesis; the pooled variant (Appendix I) remains synthetic-only. **Characterization is correlational and partial.** The trend/volatility story rests on 6/11 datasets (24/44 cells), splits by dataset family rather than a smooth gradient, and may be a post-hoc fit. **Baseline comparability.** Baselines are REPORTED (not reproduced); look-back is matched to the winning baseline, not iso-look-back across all six; and the ETT train split differs (0.8 vs 0.75) *where PRISM's wins concentrate*, a confound to be resolved by re-running ≥ 1 beaten baseline at 0.8. **Channel independence, coverage, efficiency.** PRISM cannot exploit genuine cross-channel dependence (likely contributing to large-multivariate losses); Electricity/Traffic/Solar/PEMS are outside the current scope; and efficiency is a single operating point (no FLOPs, latency, or memory curves).

## 11 Conclusion

PRISM is a small, channel-independent patch forecaster whose contribution is knowledge rather than a benchmark ranking, and three findings survive the study. First, an explicit shape–scale–level head can be made *verifiably faithful* to known generative factors under a gauge-aware, three-seed analysis (our load-bearing result, with its raw-recovery limits disclosed), and this faithfulness is a property of the head, not of the multi-scale encoder. Second, faithful does not mean useful: PRISM's real-data advantage is real but narrow, concentrated on a high-trend, low-volatility niche rather than broad SOTA; and the synthetic multi-scale advantage does not carry over, with a shape-only head matching the full decomposition. Third, the tested optional components (gating, learnable-Fourier, period-folding, channel-mixing) do not improve accuracy, capacity barely helps, context length is inconclusive, and channel-mixing is harmful so a deliberately simple composition is the accurate description. We hope this combination (a verifiable faithfulness claim alongside a transparent account of where a lightweight forecaster helps and fails) is more useful than a narrow benchmark win.

## Reproducibility Statement

The model is specified exactly against the released reference implementation in Section 3; every equation matches an implementation line, and parameter counts (247,172/423,140) are by direct enumeration. The full protocol (splits, per-window standardization, AdamW lr $10^{-4}$/cosine/wd $10^{-4}$, Huber, patience 5, bf16, batch 16, grad-accum 2, seed = 1) is in Section 4 and the released configuration. Baseline provenance (OpenTS query endpoint, leaderboard last-modified 2025-11-26) and comparability caveats are in Appendix B. The synthetic generator ships with self-checks and a recorded dataset content hash; the faithfulness, head-necessity, and efficiency results are reproduced by the released code and evaluation artifacts, which underlie the reported tables. All reported PRISM results are post the adapter loss/gradient-accounting fix; pre-fix numbers are not reported. All real-data point estimates are seed 1, with standard deviations over 2–3 seeds reported per cell (3 seeds for 51 of 68 cells, 2 seeds for 16, and a single seed for PRISM-mp-423k on ETTm2 $L$=512/$H$=720; Appendix C). Seed coverage per cell is listed in `prism_master_scores_multiseed.csv`. Code and evaluation artifacts are released to reproduce every reported number.

## Use of AI Assistants Statement

AI-based coding and writing assistants were used to help implement experiments, analyze results, generate figures, and draft and edit this manuscript. All reported numbers were computed by the described code and verified against source artifacts by the authors; all scientific claims, framing, and the decision to report negative and single-seed results are the authors' own.

## REFERENCES


Cristian Challu, Kin G. Olivares, Boris N. Oreshkin, Federico Garza, Max Mergenthaler-Canseco, and Artur Dubrawski. N-hits: Neural hierarchical interpolation for time series forecasting. In *AAAI Conference on Artificial Intelligence*, 2023.

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 (ICLR)*, 2022.

Shengsheng Lin, Weiwei Lin, Wentai Wu, Haojun Chen, and Junjie Yang. Sparsetsf: Modeling long-term time series forecasting with 1k parameters. In *International Conference on Machine Learning (ICML)*, 2024.

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 (ICLR)*, 2024.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1): 54–74, 2020.

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.

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 (ICLR)*, 2020.

Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, Zhenli Sheng, and Bin Yang. Tfb: Towards comprehensive and fair benchmarking of time series forecasting methods. *Proceedings of the VLDB Endowment*, 17(9), 2024.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

Slawek Smyl. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, 36(1):75–85, 2020.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 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 (NeurIPS)*, 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 (ICLR)*, 2023.

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

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations (ICLR)*, 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*, 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 (ICML)*, 2022.

# A Architecture Details, Statistics, and Complexity

**Problem formulation and capacity.** Let $\mathbf{X} \in \mathbb{R}^{C\times L}$ be a multivariate context of $C$ variates over $L$ steps. PRISM applies a shared univariate core to each channel independently and stacks the results; there is no parameter coupling channels. The core maps univariate batches $X \in \mathbb{R}^{B\times L} \to \hat{Y} \in \mathbb{R}^{B\times H}$. The main model uses encoder width $d = 48$ and decoder width $d_h = 96$; a higher-capacity variant uses $d = 64$, $d_h = 128$ (Section E).

**Multi-patch temporal encoder.** For each patch length $p \in \mathcal{P} = \{4, 8, 16\}$ with stride $s_p = \lfloor p/2 \rfloor$ (i.e. $\{2, 4, 8\}$, 50% overlap), the replicate-padded context is unfolded into $n_p = \lfloor (L - p)/s_p \rfloor + 1$ patches (a final partial patch is appended if the strided windows do not cover $L$). Each scale owns a **dedicated** linear embedding $E_p : \mathbb{R}^p \to \mathbb{R}^d$, giving tokens $Z^{(p)} \in \mathbb{R}^{B\times n_p\times d}$; only the per-scale embeddings differ across scales. The three scales' token sequences are concatenated along the token axis to form the memory $Z \in \mathbb{R}^{B\times P\times d}$ with $P = \sum_p n_p$.

**Shared U-Net.** A single U-Net (`_run_one`) is reused across all three scales. Given a scale's tokens $Z^{(p)}$, we transpose to channels-first $z \in \mathbb{R}^{B\times d\times n_p}$ and run a two-level encoder–decoder over the token axis. Each level is a residual convolutional block

$$\texttt{Block}(z) = z + \big(\text{Dropout} \circ \text{Conv1d}_3 \circ \text{GELU} \circ \text{GroupNorm}_8 \circ \text{Conv1d}_3\big)(z), \tag{2}$$

with kernel size 3, padding 1 (length-preserving), GroupNorm of 8 groups, and **width fixed at $d$ throughout, there is no channel doubling**. After each down-block a length-halving pool **averages adjacent token pairs**, $\texttt{Pool} : [B, d, n] \mapsto [B, d, n/2]$ (zero-padding when $n$ is odd), and the pre-pool activations are retained as skips. Up-sampling uses nearest-neighbor interpolation to each skip length, a skip-add, and a residual up-block. The U-Net returns $Z^{(p)} + \tilde{z}$, preserving the token count $n_p$.

**Learned-query bottleneck.** At the deepest pooled level of each scale, a parameter $Q_{\text{bn}} \in \mathbb{R}^{n_q\times d}$ with $n_q = 4$ is broadcast over the batch and cross-attends (multi-head attention) to the deepest tokens as keys/values. RoPE is applied to the **deepest keys only**; the learned queries are position-free. The output is post-processed by `LayerNorm` → `Linear` → `GELU` → `LayerNorm` and feeds the up-sampling stream. This bottleneck is empirically inert (Section E); it is retained as part of the default configuration.

**Direct forecasting decoder.** The horizon is emitted in one pass from $H$ learned future queries $Q_{\text{fut}} \in \mathbb{R}^{H\times d_h}$. The memory $Z$ is projected to the decoder width, $M = \texttt{ctx_proj}(Z) \in \mathbb{R}^{B\times P\times d_h}$. Rotary embeddings are applied to **both** sides, $K = \text{RoPE}(M)$ and $q = \text{RoPE}(Q_{\text{fut}}, \text{future=True})$. The two sides use different conventions: by default (`abs_rotary=False`) the context keys use *relative* positions with the last key at index 0 ($\text{arange}(P) - (P-1)$), while the future queries use *absolute* positions ($\text{arange}(H)$). The decoder is a **single** `nn.TransformerDecoderLayer` (width $d_h$, 8 heads, feed-forward dimension $4d_h$, `batch_first`); the `nlayers=2` constructor kwarg does **not** stack the decoder. It executes, in order, (i) self-attention among the $H$ future queries, (ii) cross-attention from queries to $K$, and (iii) a position-wise FFN, each with residual and layer normalization.

**Shape head.** A per-step linear head $\texttt{head_shape} : \mathbb{R}^{d_h} \to \mathbb{R}$ produces a raw waveform $s_0 \in \mathbb{R}^{B\times H}$, refined by a 1-D convolutional stack over the horizon axis (`refine`): four kernel-5 (padding-2) convolutions with channel widths $1 \to 32 \to 32 \to 32 \to 1$ and GELU between them, $S = \texttt{refine}(s_0) \in \mathbb{R}^{B\times H}$.

**Exact five trailing statistics.** The two scalar heads are conditioned on the encoder memory pooled over the context-token axis, $z_{\text{pool}} = \frac{1}{P}\sum_{i=1}^{P} Z_i \in \mathbb{R}^d$, concatenated with $n_{\text{feat}} = 5$ trailing statistics of the raw window $X$ (`_trailing_feats`). With $w_k$ the last $k$ observations, the five statistics are, exactly: (1) $\text{mean}(w_{48})$; (2) $X_L$ (last value); (3) $\text{std}(w_{48}) + 10^{-6}$; (4) $\text{mean}(w_{48}) - \text{mean}(w_{48}[:1])$; (5) $\text{mean}(w_{24}) - \text{mean}(w_{48}[:24])$. These five are fixed. The context is $b_{\text{ctx}} = [\, z_{\text{pool}} \,\|\, \phi(X) \,] \in \mathbb{R}^{d+5}$. Two MLPs $\mathbb{R}^{d+5} \to 32 \to 1$ (GELU hidden) produce the additive level $B(X)$ and, when `scale_aware=True`, the amplitude $A(X) = \text{clamp}_{\geq 10^{-3}}(\text{softplus}(\texttt{head_scale}(b_{\text{ctx}})))$, strictly positive.

**Complexity and parameter counts.** With $P = \sum_p n_p$ memory tokens, the convolutional U-Net is $O(Pd^2)$; the bottleneck attention is cheap ($n_q = 4$); the decoder is dominated by self-attention $O(H^2 d_h)$, cross-attention $O(HPd_h)$, and FFN $O(Hd_h^2)$. Because $P \ll L$ and decoding is single-pass, PRISM avoids both quadratic-in-$L$ point attention and $H$-step autoregressive unrolling. At

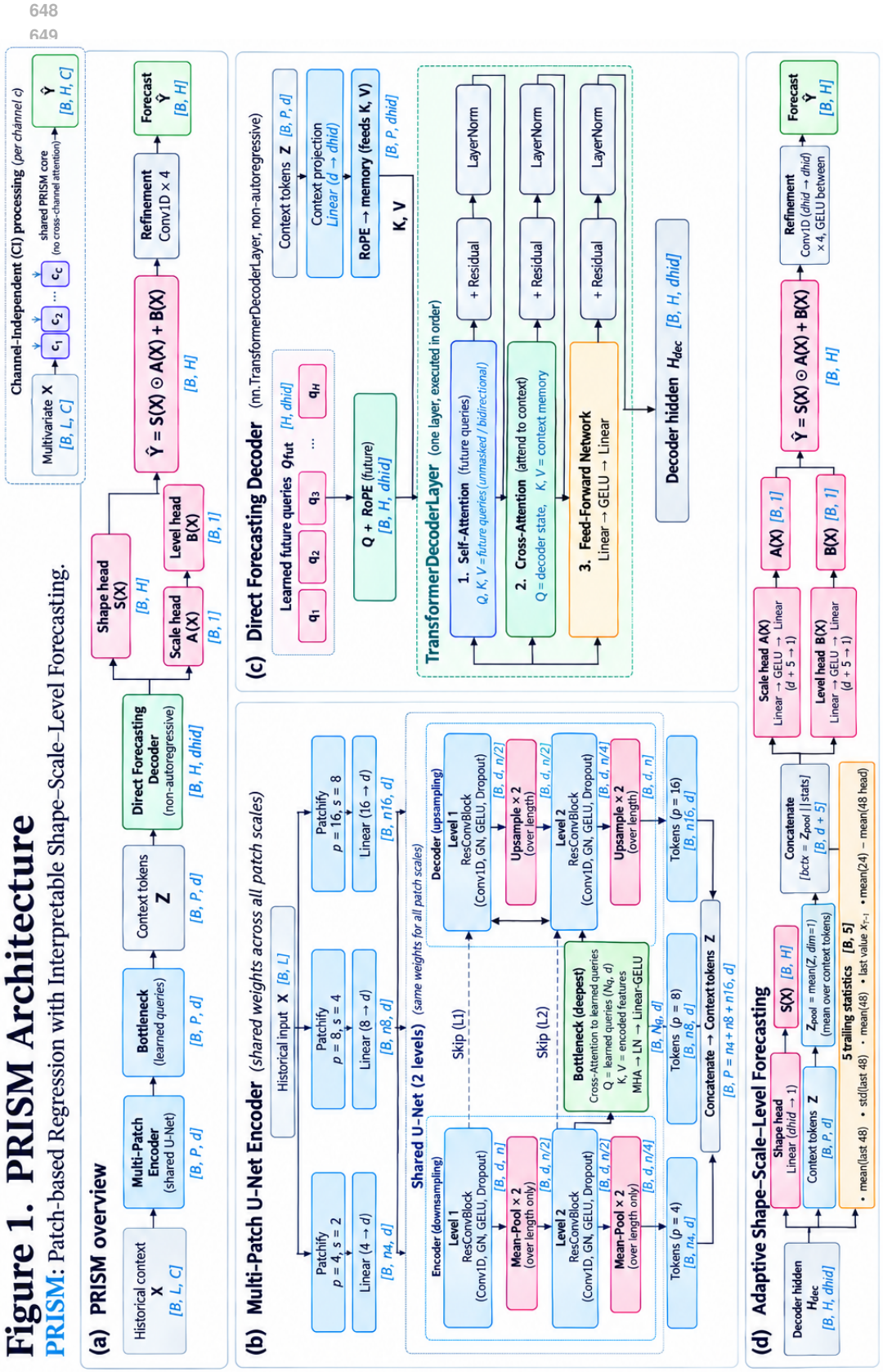


Figure 1: **PRISM architecture (four panels).** (a) End-to-end overview: a channel-independent core maps a standardized context window to the full horizon in a single (non-autoregressive) pass. (b) Multi-patch U-Net encoder: three patch scales (4/8/16, 50% overlap) share a two-level U-Net and a learned-query bottleneck; per-scale tokens are concatenated into the context memory. (c) Direct forecasting decoder: one Transformer decoder layer in which the future queries self-attend, then cross-attend to the context memory (rotary positions on both sides), followed by a feed-forward block. (d) Adaptive shape–scale–level output stage: the scale and level heads read the pooled context memory plus five trailing-context statistics, and modulate the decoded shape waveform, followed by a four-layer Conv1D refinement. The exact head equation is Eq. (1).

the locked LTSF setting ($L = H = 96$, $\mathcal{P} = \{4, 8, 16\}$) the main model ($d = 48, d_h = 96$) has **247,172** parameters and the higher-capacity variant ($d = 64, d_h = 128$) has **423,140** (both by direct enumeration). Parameter count is independent of $C$.

## B Baseline Provenance and Comparability Caveats

We compare against six reported baselines: DLinear, SparseTSF, iTransformer, PatchTST, TimeMixer, TimesNet. All baseline MSE/MAE are the published OpenTS-Bench multivariate LTSF results (raw/denormalized), retrieved from the OpenTS backend query endpoint (`opents.xyz/forecasting/multi/query`), leaderboard last-modified 2025-11-26 (recorded in the accompanying baseline-provenance record). **These baselines are REPORTED, not reproduced by us; PRISM numbers are NEW.** We never mix provenances in a single verdict. Four comparability caveats apply, stated as limitations of the baseline join:

1. **Look-back differs.** PRISM's per-cell look-back $L$ is set equal to the winning baseline's `seq_len` for that cell (equal in 44/44 cells), so PRISM is not disadvantaged on context relative to the winner; but the join is not iso-look-back across all six baselines.
2. **Train/validation split (intentional).** PRISM uses 0.8 vs the baselines' 0.75/0.875. This is deliberate anti-overfitting, but it means the ETT-family comparison is not on an identical split, and, notably, the ETT family is exactly where PRISM's wins concentrate (ETTh2/ETTm2), so those wins carry a split-comparability asterisk.
3. **No cross-dataset raw-MSE averaging.** Raw MSE spans $\sim 10^3\times$ across datasets, so we aggregate only via per-cell percentage-vs-best-baseline and win/competitive/loss counts, never raw-MSE means.
4. **Per-dataset endpoint provenance for the additional datasets.** The baseline values for the five datasets beyond the primary eight (AQShunyi, AQWan, CzeLan, Wind, and ZafNoo) are drawn from the same OpenTS source as the primary eight, but their per-dataset endpoint provenance was not separately re-verified; these should be confirmed for camera-ready.

## C Full 44-Cell Main Results

Table 2: **Full main results, raw denormalized MSE (point estimates: seed 1).** Baselines are REPORTED (OpenTS-Bench); per-cell best baseline in **bold** with its name. PRISM-mp-247k MSE/MAE are the seed-1 point estimates (NEW). Δ% = PRISM vs best baseline (negative = PRISM better). Verdict: WIN (Δ ≤ 0), COMP (0 < Δ ≤ 5%), LOSS (Δ > 5%). The ±std column gives the standard deviation of PRISM's MSE across three seeds (seeds 1,2,3); cells marked ($n$=2) currently have two seeds pending the third. Not evaluated for PRISM (out of scope): Electricity, Traffic, Solar, PEMS04, PEMS08.

| Dataset | H | Best baseline (MSE) | PRISM MSE | PRISM MAE | Δ% | Verdict | ±std (seeds 1–3) |
|---|---|---|---|---|---|---|---|
| ETTh1 | 96 | SparseTSF **7.573** | 10.899 | 1.914 | +43.9 | LOSS | ±0.2432 |
| ETTh1 | 192 | SparseTSF **8.469** | 12.465 | 2.104 | +47.2 | LOSS | ±0.1120 |
| ETTh1 | 336 | SparseTSF **8.947** | 13.323 | 2.225 | +48.9 | LOSS | ±0.2091 |
| ETTh1 | 720 | SparseTSF **8.621** | 15.291 | 2.429 | +77.4 | LOSS | ±0.1413 |
| ETTh2 | 96 | PatchTST **17.169** | 17.725 | 2.681 | +3.2 | COMP | ±0.2543 |
| ETTh2 | 192 | PatchTST **20.864** | 21.512 | 3.068 | +3.1 | COMP | ±0.1407 |
| ETTh2 | 336 | SparseTSF **23.762** | 23.266 | 3.250 | −2.1 | WIN | ±0.2828 |
| ETTh2 | 720 | SparseTSF **26.009** | 26.539 | 3.523 | +2.0 | COMP | ±2.01 |
| ETTm1 | 96 | PatchTST **5.844** | 7.591 | 1.473 | +29.9 | LOSS | ±0.0287 |
| ETTm1 | 192 | PatchTST **6.867** | 9.195 | 1.647 | +33.9 | LOSS | ±0.0252 ($n$=2) |
| ETTm1 | 336 | PatchTST **7.877** | 9.701 | 1.763 | +23.2 | LOSS | ±0.0480 |
| ETTm1 | 720 | PatchTST **9.196** | 10.883 | 1.908 | +18.3 | LOSS | ±0.0243 ($n$=2) |
| ETTm2 | 96 | DLinear **10.690** | 9.490 | 1.936 | −11.2 | WIN | ±0.0107 |
| ETTm2 | 192 | DLinear **14.222** | 12.645 | 2.241 | −11.1 | WIN | ±0.1640 |
| ETTm2 | 336 | DLinear **17.574** | 15.126 | 2.442 | −13.9 | WIN | ±0.0433 |
| ETTm2 | 720 | DLinear **22.131** | 21.351 | 2.957 | −3.5 | WIN | ±1.14 |

*Table 2 continued*

| Dataset | H | Best baseline (MSE) | PRISM MSE | PRISM MAE | Δ% | Verdict | ±std (seeds 1–3) |
|---|---|---|---|---|---|---|---|
| Weather | 96 | TimeMixer **1715.985** | 1758.058 | 12.889 | +2.5 | COMP | ±54.13 |
| Weather | 192 | PatchTST **1719.136** | 1964.069 | 14.306 | +14.3 | LOSS | ±50.81 ($n$=2) |
| Weather | 336 | PatchTST **1739.622** | 2298.086 | 17.024 | +32.1 | LOSS | ±149 |
| Weather | 720 | PatchTST **1858.973** | 2290.621 | 17.401 | +23.2 | LOSS | ±198 |
| Exchange | 96 | DLinear **0.001** | 0.0026 | 0.0338 | +162.9 | LOSS | ±0.0003 |
| Exchange | 192 | DLinear **0.002** | 0.0028 | 0.0359 | +41.7 | LOSS | ±0.0001 |
| Exchange | 336 | DLinear **0.003** | 0.0047 | 0.0462 | +58.3 | LOSS | ±0.0001 |
| Exchange | 720 | DLinear **0.006** | 0.0124 | 0.0813 | +105.8 | LOSS | ±0.0022 |
| Wind | 96 | DLinear **1009.923** | 1034.648 | 13.849 | +2.5 | COMP | ±11.56 |
| Wind | 192 | DLinear **1134.947** | 1195.205 | 15.716 | +5.3 | LOSS | ±4.88 |
| Wind | 336 | DLinear **1389.389** | 1386.865 | 17.650 | −0.2 | WIN | ±25.66 |
| Wind | 720 | DLinear **1456.767** | 1426.919 | 18.369 | −2.1 | WIN | ±11.31 |
| AQShunyi | 96 | iTransformer **130096.355** | 148184.229 | 85.084 | +13.9 | LOSS | ±749 |
| AQShunyi | 192 | TimesNet **141350.772** | 161487.200 | 87.188 | +14.3 | LOSS | ±1,670 |
| AQShunyi | 336 | DLinear **146874.030** | 169390.287 | 89.230 | +15.3 | LOSS | ±2,325 ($n$=2) |
| AQShunyi | 720 | TimeMixer **159674.347** | 174383.690 | 92.188 | +9.2 | LOSS | ±1,034 ($n$=2) |
| AQWan | 96 | TimeMixer **126740.228** | 161332.660 | 88.233 | +27.3 | LOSS | ±5,691 |
| AQWan | 192 | TimesNet **129902.387** | 174671.523 | 91.349 | +34.5 | LOSS | ±8,799 ($n$=2) |
| AQWan | 336 | TimeMixer **143069.326** | 166959.821 | 92.115 | +16.7 | LOSS | ±3,425 |
| AQWan | 720 | TimeMixer **163520.195** | 174363.873 | 93.123 | +6.6 | LOSS | ±3,439 |
| CzeLan | 96 | PatchTST **299509.104** | 347122.603 | 146.087 | +15.9 | LOSS | ±12,306 |
| CzeLan | 192 | TimesNet **328314.377** | 458679.170 | 171.840 | +39.7 | LOSS | ±26,249 ($n$=2) |
| CzeLan | 336 | TimeMixer **364145.105** | 547458.370 | 189.483 | +50.3 | LOSS | ±49,100 ($n$=2) |
| CzeLan | 720 | TimeMixer **373857.916** | 518735.180 | 194.297 | +38.8 | LOSS | ±42,998 |
| ZafNoo | 96 | PatchTST **14258.722** | 16214.446 | 36.274 | +13.7 | LOSS | ±268 |
| ZafNoo | 192 | SparseTSF **14771.508** | 16315.034 | 37.374 | +10.5 | LOSS | ±650 ($n$=2) |
| ZafNoo | 336 | SparseTSF **15400.368** | 17147.246 | 39.828 | +11.3 | LOSS | ±319 |
| ZafNoo | 720 | PatchTST **15831.538** | 18600.935 | 46.483 | +17.5 | LOSS | ±402 ($n$=2) |

## D Horizon Sensitivity

PRISM predicts the full horizon directly (non-autoregressive), so error does not accumulate step-by-step; nevertheless its raw MSE grows substantially with the horizon length $H$, as it does for direct multi-horizon forecasters generally. We quantify this within PRISM by normalizing each dataset's raw denormalized MSE (single seed, from Table 2) to that dataset's own $H = 96$ value. Table 3 reports the resulting %-degradation. We make **no claim of horizon-robustness**: the steepest degradation is on ETTm2 (+125% at $H = 720$), and even the mildest ETT family (ETTh1) reaches +40% at $H = 720$. The "long-horizon" label in this paper denotes the task, not a property of graceful error scaling.

Table 3: **Horizon sensitivity of PRISM-mp-247k, %-increase in raw MSE vs the same dataset's own $H = 96$ (single seed).** Computed directly from Table 2 (e.g. ETTm2 $H = 720$: 21.351/9.490 − 1 = +125%). $H = 96$ is the 0% reference by construction. PRISM's error grows steeply with horizon; the steepest degradation is ETTm2 at $H = 720$ (+125%).

| Dataset | $H$=96 (ref) | $H$=192 | $H$=336 | $H$=720 |
|---|---|---|---|---|
| ETTh1 | 0% | +14% | +22% | +40% |
| ETTh2 | 0% | +21% | +31% | +50% |
| ETTm1 | 0% | +21% | +28% | +43% |
| ETTm2 | 0% | +33% | +59% | +125% |

For context, the ETTm2 wins over the per-cell best-of-six baseline remain WINs at every horizon while the margin shrinks with $H$: −11.2% ($H$=96), −11.1% ($H$=192), −13.9% ($H$=336), −3.5% ($H$=720) (Table 2). The advantage narrows sharply at $H = 720$ but does not invert.

## E Full Patch and Capacity Ablations

**Multi-patch does NOT beat single-patch on real data.** On controlled synthetic scale-varying data (S2) the multi-scale encoder improves forecast MSE by ≈29% over a matched single-patch model (same PRISM core, matched to within 0.43%: 246,116 vs 247,172 parameters; identical look-back, protocol, and training; only `plens=[8]` vs `[4,8,16]` differing; 3-seed mean; computed from the synthetic-probe analysis in the accompanying code). On real LTSF this does not transfer: over the 17 cells where both mp and sp8 exist at equal 247k capacity, mp is better in only 6/17; sp8 wins 11/17 (notable sp8 gains: CzeLan H96 −12.5%, Exchange H192, ETTh1 H192).

Table 4: **Patch ablation, raw MSE (PRISM, seed = 1, 247k), rows where a genuine single-patch 247k run exists.** Best per row in **bold**. All cells are genuine 247k-capacity, seed-1 values from our released benchmark score table; rows without a single-patch 247k run are omitted rather than back-filled. The mp-vs-sp8 count (6/17 vs 11/17) is over the full set of 17 cells with both mp-247k and sp8-247k runs.

| Dataset | H | mp [4,8,16] | sp4 | sp8 | sp16 |
|---|---|---|---|---|---|
| AQShunyi | 192 | 161487.200 | 161250.855 | **157524.212** | 159784.963 |
| AQShunyi | 720 | 174383.690 | 168602.583 | **167430.549** | 178665.836 |
| AQWan | 336 | 166959.821 | **162181.556** | 170555.153 | 171652.127 |
| CzeLan | 96 | 347122.603 | 342352.137 | **303604.002** | 322928.445 |
| CzeLan | 336 | 547458.370 | **462756.268** | 527673.535 | 528534.225 |
| ETTh1 | 192 | 12.465 | 11.926 | **11.917** | 24.915 |
| ETTh1 | 336 | 13.323 | 12.952 | 13.064 | **12.686** |
| ETTh2 | 336 | 23.266 | 23.211 | 24.449 | **22.151** |
| ETTm1 | 96 | 7.591 | 7.474 | **7.347** | 7.373 |
| ETTm1 | 720 | 10.883 | 11.201 | **10.811** | 10.844 |
| ETTm2 | 96 | 9.490 | 9.729 | 9.461 | **9.433** |
| ETTm2 | 336 | **15.126** | 15.196 | 15.290 | 16.527 |
| Exchange | 96 | **0.0026** | 0.0027 | 0.0028 | 0.0043 |
| Exchange | 192 | 0.0028 | 0.0037 | **0.0023** | 0.0032 |
| Weather | 96 | 1758.058 | **1742.348** | 1800.837 | 1833.686 |
| Weather | 720 | 2290.621 | 2473.440 | 2949.809 | **2289.545** |
| Wind | 192 | 1195.205 | 1200.082 | **1193.028** | 1196.231 |

**Capacity: small, mixed.** Over the 24 cells where a genuine mp-423k run exists (ETT/Exchange/Weather families), the 423k model is better in 14/24 cells, mean improvement only −1.63%. Under the final locked configuration the capacity gap largely closes, so we report 247k as the main model. Comparison is fair: 247k and 423k share look-back, protocol, and seed; only width differs.

Table 5: **Capacity ablation, raw MSE (PRISM-mp, seed = 1), 24 overlapping cells.** Better per row in **bold**.

| Dataset | H | 247k (main) | 423k | Δ% (423k vs 247k) |
|---|---|---|---|---|
| ETTh1 | 96 | 10.899 | **10.691** | −1.9 |
| ETTh1 | 192 | 12.465 | **12.142** | −2.6 |
| ETTh1 | 336 | 13.323 | **13.227** | −0.7 |
| ETTh1 | 720 | **15.291** | 15.587 | +1.9 |
| ETTh2 | 96 | 17.725 | **17.466** | −1.5 |
| ETTh2 | 192 | 21.512 | **21.476** | −0.2 |
| ETTh2 | 336 | **23.266** | 23.864 | +2.6 |
| ETTh2 | 720 | **26.539** | 27.947 | +5.3 |
| ETTm1 | 96 | 7.591 | **7.430** | −2.1 |
| ETTm1 | 192 | 9.195 | **9.118** | −0.8 |
| ETTm1 | 336 | **9.701** | 9.764 | +0.7 |
| ETTm1 | 720 | **10.883** | 10.982 | +0.9 |
| ETTm2 | 96 | 9.490 | **9.419** | −0.8 |
| ETTm2 | 192 | **12.645** | 13.305 | +5.2 |
| ETTm2 | 336 | **15.126** | 15.744 | +4.1 |
| ETTm2 | 720 | 21.351 | **19.510** | −8.6 |
| Exchange | 96 | **0.0026** | 0.0032 | +23.3 |
| Exchange | 192 | 0.0028 | **0.0027** | −4.0 |
| Exchange | 336 | 0.0047 | **0.0039** | −17.9 |
| Exchange | 720 | 0.0124 | **0.0087** | −29.8 |
| Weather | 96 | **1758.058** | 1817.040 | +3.4 |
| Weather | 192 | **1964.069** | 1969.107 | +0.3 |
| Weather | 336 | 2298.086 | **2007.773** | −12.6 |
| Weather | 720 | 2290.621 | **2215.629** | −3.3 |

**Optional components (full context).** Under the locked configuration, gating, learnable-Fourier, and period-folding produced no reliable improvement and were disabled. As reference points (H96, seed = 1, in-harness, raw MSE): on ETTh1 the plain channel-independent core scores ≈9.98, multipatch ≈9.84, a gate variant ≈10.07, and a learnable-Fourier variant ≈10.05, all at or above the plain core; gate activations drift (0.25 → 0.72) without saturating (an inertness signature). Channel-mixing variants blow up (ETTh1 16.2/15.5 vs CI 9.98; ETTh2 45.5/55.8 vs 16.6), empirically justifying channel independence. These are drawn from an internal H96 diagnostic grid on ETTh1/ETTh2 (partial ETTm2/Exchange); a full per-cell component-ablation table is a stated missing result.

## F Single-Head vs Decomposition on Real Data (MC-3)

We ran the direct real-data control comparing a **shape-only** head (decomposition removed) against the full shape · scale + level head. In the shape-only arm both scalar heads are dropped (`scale_aware=false` *and* `use_level=false`), so $\hat{Y} = S(X)$; the full arm is the default configuration. Removing the decomposition ties-or-beats the full head on all four datasets, so at $n = 1$ the decomposition does not improve real-data accuracy over a shape-only head.

Table 6: **Head ablation, shape-only vs full decomposition, raw MSE (seed = 1, $H = 96$).** shape-only = decomposition removed (`scale_aware=false`, `use_level=false`; $\hat{Y}$ = shape). $\Delta\%$ = shape-only vs full (negative = shape-only better). ETTm1 `full` run was is included (shape-only ties-or-beats full on all four datasets). Single seed, single horizon, indicative only.

| Dataset | Full MSE | Shape-only MSE | $\Delta\%$ (shape-only vs full) |
|---|---|---|---|
| ETTh1 | 11.828 | 10.991 | −7.1 |
| ETTh2 | 20.705 | 19.164 | −7.4 |
| ETTm2 | 9.601 | 9.603 | +0.02 |
| ETTm1 | 7.597 | 7.472 | −1.6 |

## G Controlled Synthetic Generator (full specification)

Table 7: **Trivial-baseline control ($L/H = 96/96$, same test split and same OLS $R^2$ as Appendix G).** "Trivial" regresses the true factor against one raw trailing statistic that is fed to the head as an input feature: std($x_{-48:}$) for scale, mean($x_{-48:}$) for level. No model is involved. Rows marked ✓ are conditions in which that factor actually varies.

| | **scale $R^2$** | | | **level $R^2$** | | |
|---|---|---|---|---|---|---|
| Condition | varies | PRISM | Trivial | varies | PRISM | Trivial |
| S1 | | 0.003 | 0.419 | | 0.054 | 0.016 |
| S2 | ✓ | 0.733 | **0.908** | | 0.005 | 0.014 |
| S3 | | 0.000 | 0.378 | ✓ | 0.994 | 0.968 |
| S4 | | 0.020 | 0.336 | | 0.047 | 0.017 |
| APERIODIC | | 0.174 | 0.686 | | 0.330 | 0.206 |
| HIGH_NOISE | | 0.015 | 0.454 | | 0.065 | 0.026 |
| REGIME_SHIFT | ✓ | 0.413 | **0.856** | ✓ | 0.937 | 0.808 |
| *mean where factor absent* | | **0.043** | 0.455 | | 0.100 | **0.056** |

Each window is built additively and exactly, storing the ground-truth parts, $y = \text{level}_\star + \text{scale}_\star \cdot \text{shape}_\star + \text{noise}_\star$, where per window $\text{level}_\star$ is a scalar DC offset, $\text{scale}_\star > 0$ a positive amplitude, $\text{shape}_\star$ a length-$H$ waveform normalized to zero mean and unit variance, and $\text{noise}_\star$ additive Gaussian noise at a target SNR. Because parts are stored (not re-derived from the noisy signal), the true factor values are available for every sample.

Table 8: **Synthetic conditions.** S2 makes only scale vary, S3 only level; S1/S4 hold both nearly constant so correct scale/level recovery is near zero. APERIODIC is an off-distribution generalization test. Periods avoid clean divisors of the architecture's patch/statistic windows ($\{4, 8, 16, 24, 48, 96, 336, 512, 528\}$) with $\pm 0.3$ jitter, preventing a trivial arithmetic shortcut.

| Condition | What varies (design intent) | level $(\mu, \sigma)$ | scale (lo,hi) | shape | SNR (dB) |
|---|---|---|---|---|---|
| **S1** GENTLE | nothing varies (both low) | $(0, 0.05)$ | $(0.8, 1.2)$ | sine | 15 |
| **S2** SCALE-HIGH | **scale varies a lot** | $(0, 0.05)$ | $(0.2, 3.0)$ | sine | 15 |
| **S3** LEVEL-HIGH | **level varies a lot** | $(-2.0, 2.0)$ | $(0.8, 1.2)$ | sine | 15 |
| **S4** MULTIFREQ | shape complexity (2–3 harmonics) | $(0, 0.05)$ | $(0.8, 1.2)$ | harmonics | 15 |
| APERIODIC | off-dist.: no dominant period | $(0, 0.05)$ | $(0.8, 1.2)$ | random walk | 15 |
| HIGH_NOISE | robustness (low SNR) | $(0, 0.05)$ | $(0.8, 1.2)$ | sine | 3 |
| REGIME_SHIFT | level **and** scale both vary | $(-1.0, 1.0)$ | $(0.5, 2.5)$ | sine | 15 |

**Generator self-checks (verified in the accompanying code).** Over 50 samples per condition across all seven conditions: exact reconstruction ($\max |y - (\text{level}_\star + \text{scale}_\star \cdot \text{shape}_\star + \text{noise}_\star)| < 10^{-9}$); shape standardization ($|\text{mean}| < 10^{-4}$, $|\text{std} - 1| < 10^{-3}$); positive scale; periodicity label integrity against an FFT peak-to-mean detector; and byte-identical output under a fixed seed. The dataset is materialized once with a recorded content hash so every variant/seed trains on identical bytes.

**Reconstruction residual is a code-consistency check, not semantic faithfulness.** Across all 546 synthetic evaluation rows the internal reconstruction residual $\max |\hat{y} - (\text{shape} \cdot \text{scale} + \text{level})|$ lies in

$1.2 \times 10^{-7}$ to $9.5 \times 10^{-7}$ (mean $5.8 \times 10^{-7}$), floating-point roundoff. Because the model computes $\hat{y} = \text{shape} \cdot \text{scale} + \text{level}$ *by construction*, this residual is a **code-consistency check** confirming the reported shape, scale, and level are literally the pieces the model sums and multiplies to produce $\hat{y}$; it is *not* evidence that the pieces are semantically faithful to generative factors (that claim rests on the gauge-aware recovery of Table **??**). It makes the word "Interpretable" mechanically checkable rather than aspirational.

**Identifiability (three separate questions).** Because $S \odot A$ has the gauge freedom of the head, naively regressing each raw head against its factor conflates: (1) predictive accuracy (an MSE question); (2) contribution recovery, does a head's across-sample variation track its factor's, after removing the arbitrary gauge (our faithfulness metric, gauge-aware $R^2$); and (3) raw per-head correlation (meaningful only up to the gauge). We report (2) as the headline and also disclose (3) numerically. *Disclosed raw recovery:* in S2 the raw scale regression slopes are 0.36/0.38/0.46 across the three seeds (not $\approx 1$), and in S3 the raw `shape_product_corr` swings 0.66 → 0.90 → 0.78 across seeds.

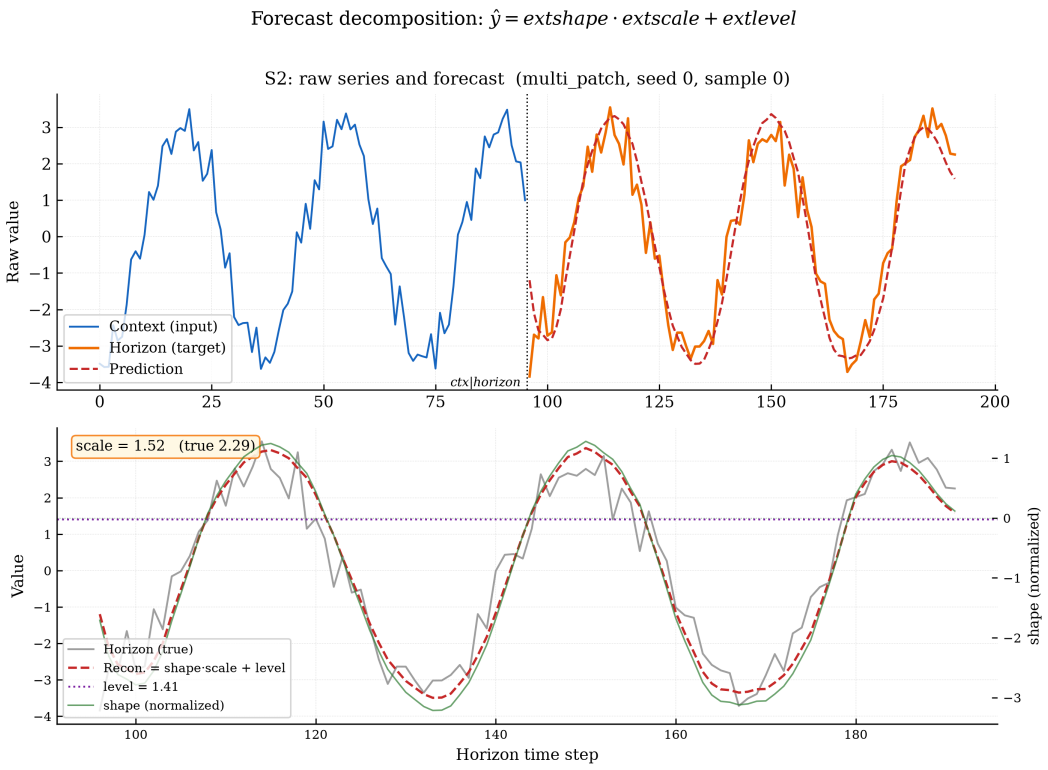


Figure 2: **Per-condition gauge-aware head recovery**: grouped bars of `scale_R`$^2$, `level_R`$^2$, `shape_corr` across the seven conditions (3-seed mean, per-seed points overlaid), showing each head lighting up in exactly the condition its factor varies.

**Per-seed spread on the headline numbers.** The gauge-aware $R^2$ is stable on the two strong recoveries: S2 `scale_R`$^2$ across seeds is 0.70/0.72/0.79 (std ≈0.04) and S3 `level_R`$^2$ has std ≈0.00. APERIODIC `level_R`$^2$ is by contrast noisy across seeds (0.00/0.46/0.53, std ≈0.23), consistent with its off-distribution status.

## H Head-Necessity Probe: Full Delta Table

Table 9: **Head-necessity deltas** (bold = trustworthy, $|\Delta\%| \geq 5\%$). Single-seed (seed 0), single-horizon ($L = H = 96$), per-condition. Positive $\Delta$ = removing the head hurt = head necessary. Values are computed from our released evaluation artifacts.

| Condition | ΔMSE no-scale | Δ% no-scale | ΔMSE no-level | Δ% no-level |
|---|---|---|---|---|
| S1 | −0.0012 | −1.7% | −0.0040 | −5.7% |
| S2 | −0.0025 | −1.1% | −0.0061 | −2.6% |
| **S3** | −0.0255 | −12.7% | +0.0429 | +21.3% |
| S4 | −0.0065 | −3.5% | −0.0164 | −8.8% |
| **APERIODIC** | −0.0035 | −0.5% | −0.0953 | −13.2% |
| HIGH_NOISE | −0.0039 | −0.6% | −0.0089 | −1.5% |
| REGIME_SHIFT | +0.0111 | +4.3% | +0.0098 | +3.8% |

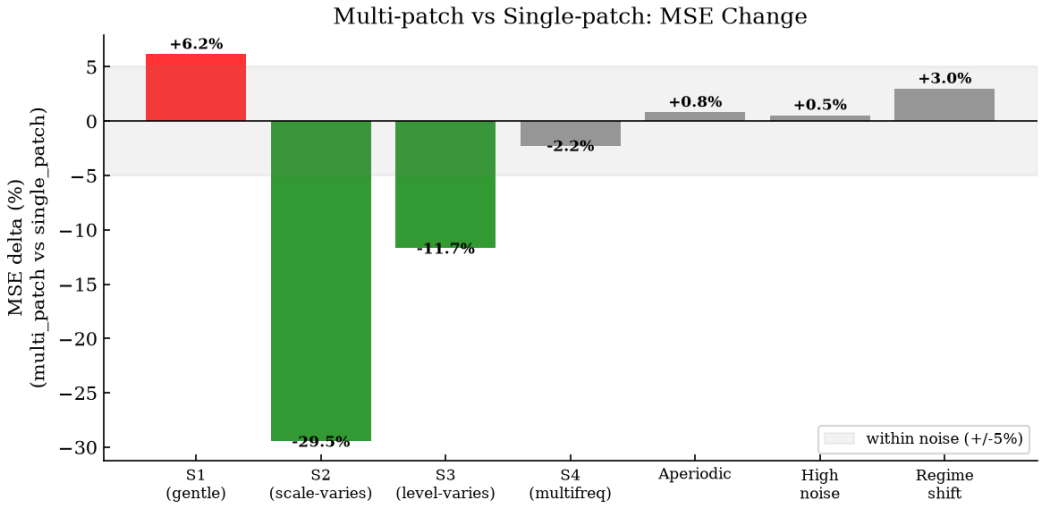


Figure 3: **Head-necessity deltas**: per-condition Δ% for no-scale and no-level arms with the ±5% noise band shaded, highlighting the level/S3 +21.3% necessity and the two harmful effects (level/APERIODIC −13.2%, scale/S3 −12.7%).

# I Pooled vs Per-condition Specialist (full)

Table 10: **Pooled vs per-condition specialist, synthetic raw MSE, $L = H = 96$.** Pooled = 3-seed mean (seeds 0,1,2); specialist = necessity-probe `full`, seed 0 (single). Δ% = pooled vs specialist (negative = pooled better). `level_r2` is *not* reported for the low-level-variance conditions where it is an uninterpretable metric artifact (near-constant true level makes the $R^2$ denominator degenerate). Indicative, not a matched comparison: the −42% to −76% gains compare a *3-seed pooled mean* against a *single seed-0* specialist, so they are **not seed-matched**.

| Condition | Pooled (3-seed mean) | Specialist (seed-0) | Δ% | Reading |
|---|---|---|---|---|
| S1 | 0.0406 | 0.0699 | −42% | pooled better |
| S2 | 0.1289 | 0.2316 | −44% | pooled better |
| S3 | 0.0478 | 0.2013 | −76% | pooled better |
| S4 | 0.1240 | 0.1850 | −33% | pooled better |
| REGIME_SHIFT | 0.1198 | 0.2604 | −54% | pooled better |
| HIGH_NOISE | 0.5649 | 0.5960 | −5% | marginal |
| APERIODIC | 0.8836 | 0.7231 | +22% | **pooled worse (negative interference)** |

A single jointly-trained model matches-or-beats the per-condition specialists on 6/7 conditions (large gains on S1–S4 and REGIME_SHIFT, marginal on HIGH_NOISE), with clear negative interference on APERIODIC (+22%), the off-distribution condition. This partially addresses the previously-open joint/pooled-necessity question; it is scoped to $L = H = 96$ and synthetic data only, and is indicative (the pooled figure is a 3-seed mean while the specialist reference is a single seed-0 run).

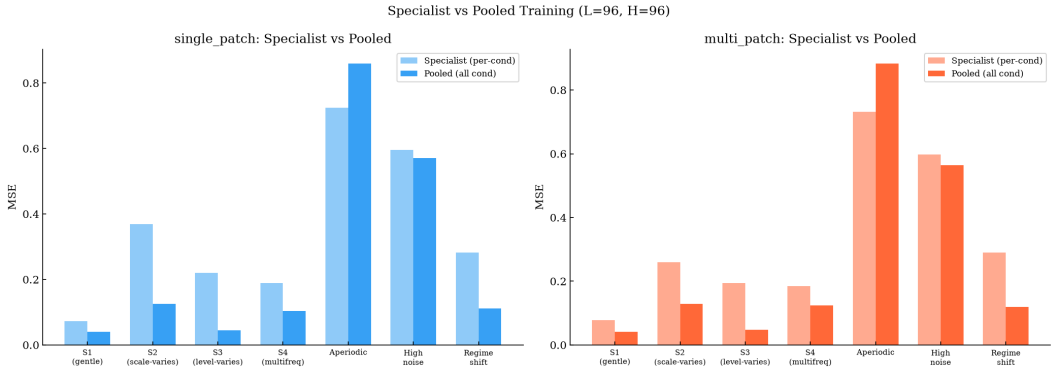


Figure 4: **Pooled vs per-condition specialist**: synthetic raw MSE at $L = H = 96$ for the jointly-trained pooled model (3-seed mean) against the per-condition `full` specialists (seed 0), showing pooled matches-or-beats on 6/7 conditions with negative interference on APERIODIC.

# J Dataset Characterization and Failure Analysis (full)

We joined the per-cell verdicts (vs best-of-six) to **TFB-provided** dataset characteristics, comparing via %-vs-best (raw MSE is not comparable across datasets). This analysis is **correlational only**: no causal claim. *Trend signal HOLDS:* wins associate with high-trend datasets, mean trend strength 0.87 (WIN) vs 0.29 (LOSS). *Volatility signal INVERTS:* the synthetic prediction was that wins have *higher* volatility, but PRISM wins the *lower*-volatility datasets (WIN 0.47 vs LOSS 0.62). *Periodicity INCONCLUSIVE* (∼ 16% margin). Results split by dataset *family*, not along a smooth gradient: PRISM wins the ETTh2/ETTm2 family (low volatility ∼ 0.47, high trend ∼ 0.87) and loses the ETTh1/ETTm1 family (high volatility ∼ 0.80, low trend ∼ 0.31), plus Exchange and most of Weather. The synthetic→real mechanism therefore transfers on **1/3** of its axes (trend holds; periodicity null; volatility inverts). **Major limitation:** only 6 of 11 datasets carry characterization rows (ETTh1, ETTh2, ETTm1, ETTm2, Exchange, Weather); the transfer test rests on 24 of 44 cells. AQShunyi, AQWan, CzeLan, Wind, ZafNoo are uncharacterized.

**The eight TFB characteristic definitions (provided, not computed by us).** The descriptors are the standard TFB series characteristics, computed by TFB's extractor (R `catch22`/`tsfeatures` plus `statsmodels`): **Correlation** (inter-variate correlation across channels); **Transition** (catch22 `SB_TransitionMatrix_3ac_sumdiagcov`); **Shifting** (catch22 `DN_OutlierInclude_p_001_mdrmd`); **Seasonality** (STL `seasonal_strength` = max(0, 1 − var(resid)/var(detrend)) at the top FFT period); **Trend** (STL `trend_strength` = max(0, 1 − var(resid)/var(deseasonal))); **Stationarity** (ADF test $p$-value); **Short_term_jsd** (Jensen–Shannon divergence of the empirical vs a fitted Gaussian over 30-step windows); and **Long_term_jsd** (the same JSD over 336-step windows). These eight are TFB-provided. The only layer PRISM adds is a thin per-cell verdict layer the WIN/COMP/LOSS mapping of each dataset's %-vs-best outcome onto the TFB descriptors; it computes no series characteristics of its own.

Table 11: **Characteristics and verdict layer (TFB-provided vs PRISM-layer).** Correlational only; no causal claim. Five uncharacterized datasets are excluded.

| Dataset | Trend | Volatility | Periodicity | Stationarity | Dist. shift | PRISM verdict layer |
|---|---|---|---|---|---|---|
| ETTh1 | 0.316 | 0.806 | 0.036 | 0.163 | 0.212 | LOSS 4/4 (high-vol/low-trend) |
| ETTm1 | 0.309 | 0.801 | 0.036 | 0.049 | 0.212 | LOSS 4/4 (high-vol/low-trend) |
| ETTh2 | **0.872** | 0.469 | 0.033 | 0.048 | 0.276 | 1 WIN + 3 COMP (high-trend) |
| ETTm2 | **0.870** | 0.468 | 0.034 | 0.018 | 0.275 | WIN 4/4 (high-trend/low-vol) |
| Exchange | 0.274 | 0.162 | 0.011 | 0.002 | 0.076 | LOSS 4/4 |
| Weather | 0.242 | 0.730 | 0.090 | 0.004 | 0.150 | 1 COMP + 3 LOSS |

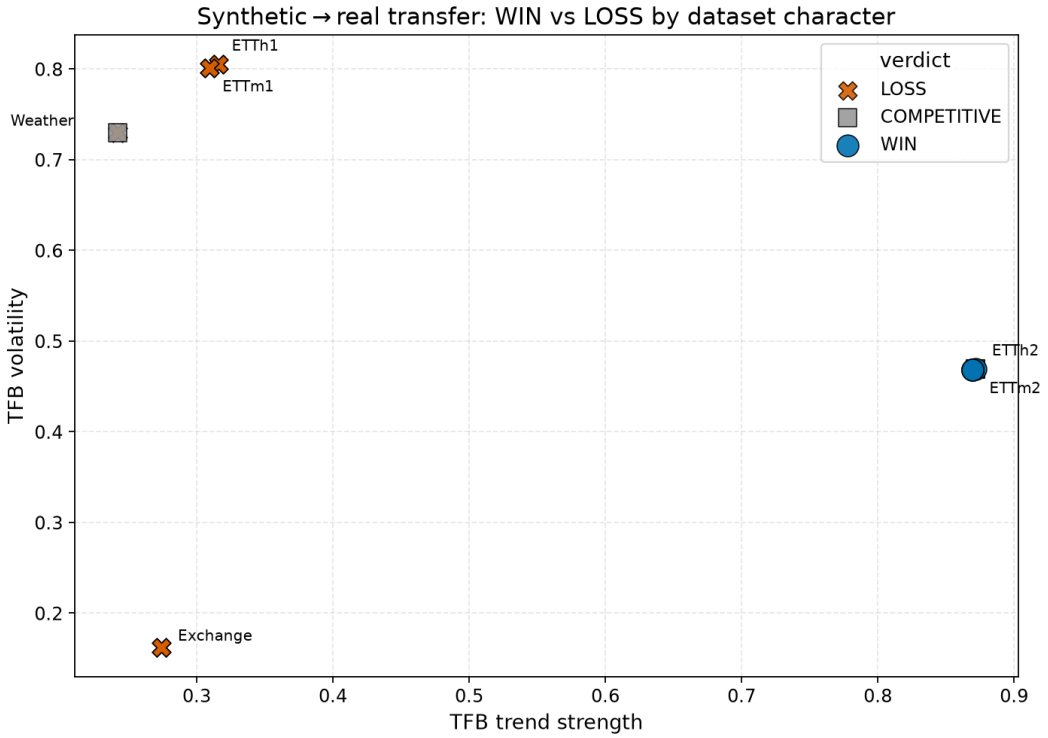


Figure 5: **Synthetic→real transfer**: WIN vs LOSS cells by TFB trend and volatility, showing trend separates (WIN 0.87 vs LOSS 0.29) while volatility inverts (WIN 0.47 vs LOSS 0.62).

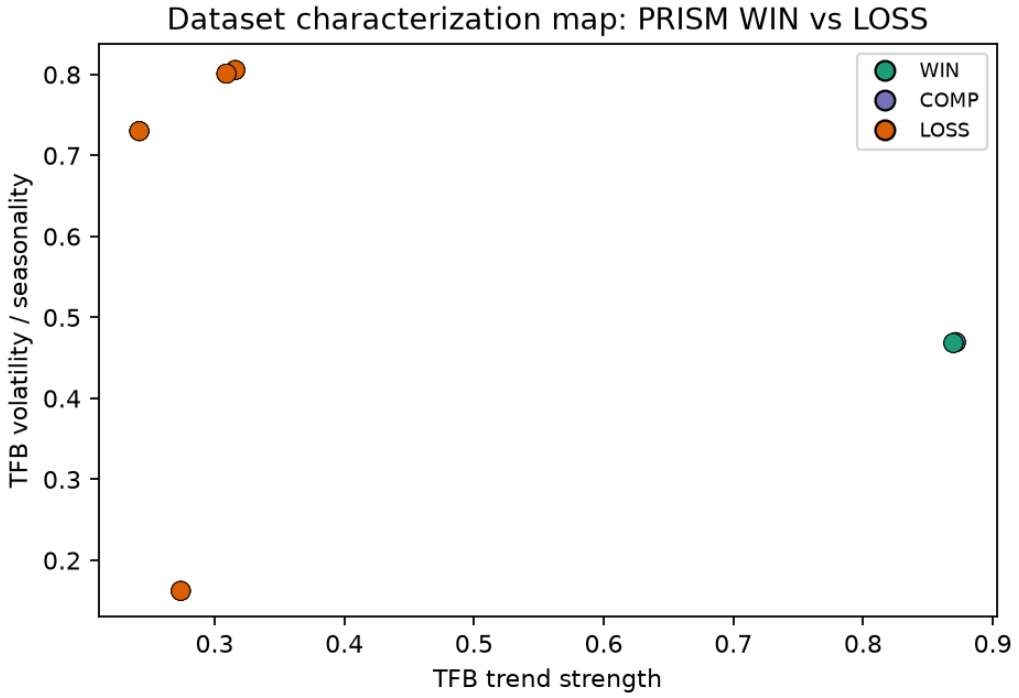


Figure 6: **Dataset characterization map**: the six characterized datasets on the trend × volatility plane, colored by PRISM verdict.

## K Efficiency (full)

Table 12: **Efficiency (batch 16, $L = 512$, $H = 96$, CUDA).**

| Model | Params (M) | Fwd (ms) | Throughput (samples/s) |
|---|---|---|---|
| PRISM-247k | 0.247 | 7.835 | 2042.2 |
| PRISM-423k | 0.423 | 7.855 | 2036.9 |

The two variants have near-identical latency and throughput (~2040 samples/s), so the capacity ablation is essentially free in wall-clock. At ~0.25M parameters PRISM is markedly smaller than several heavier baselines. Each PRISM cell is trained with early stopping (patience 5, min 1/max 200 epochs; cosine schedule; AdamW; Huber loss; batch 16, grad-accum 2; bf16). Convergence is governed by validation loss, not a fixed epoch budget. The real-data runs executed on a shared, contended GPU cluster (concurrent unrelated jobs, mixed accelerator types, preemption), so raw wall-clock time would reflect scheduling contention rather than intrinsic cost and we do not report it. This measurement is a single controlled operating point; FLOPs, per-horizon latency curves, peak GPU memory, the specific GPU model, and baseline efficiency numbers are evidence gaps. The inference chunking used to fit long series is a memory implementation detail, not a modeling contribution.

## L M4 Study: PRISM-Masked Beats the Competition Winner on 5/6 Groups

We include a short-horizon M4 study to show that the PRISM design generalizes beyond long-horizon forecasting when equipped with the mechanism M4's data demands. M4 series are short and variable-length, so a forecaster must handle front-padded contexts; we therefore evaluate **PRISM-Masked**, a member of the PRISM family that shares PRISM's shape·scale+level head, rotary position encoding, and U-Net topology, and uses PRISM's single-patch configuration ($p = 8$, stride 4) rather

than the multi-scale {4, 8, 16} used for LTSF, but replaces the plain encoder layers with mask-aware counterparts (masked U-Net convolutions, masked pooling, masked attention, and masked RevIN) so that padded positions carry no signal. The LTSF PRISM used elsewhere in this paper is the unmasked special case of the same family: on a fixed sliding window every position is real, so the masking reduces to identity.

PRISM-Masked transfers to M4. On OWA (lower is better; Naive2 = 1.000 by definition), PRISM-Masked as a single model averaged over three seeds attains a series-count-weighted total of **0.8125**, **beating the official M4 competition winner ES-RNN (Smyl, 2020) (0.821, REPORTED) on five of six frequency groups** (Yearly, Quarterly, Weekly, Daily, and Hourly), losing only on Monthly (0.861 vs. 0.836). We separate the two provenances: the ES-RNN comparison is REPORTED (taken from the M4 competition (Makridakis et al., 2020) record, not reproduced in our harness), and 0.8125 does not establish state of the art, the strongest published pure-DL result on M4, N-BEATS (Oreshkin et al., 2020), reports a lower overall OWA of 0.797 (its interpretable+generic ensemble; Table 2 therein, per-group Yearly 0.752, Quarterly 0.814, Monthly 0.819) and remains ahead on the weighted total by roughly 0.015. The claim is thus "PRISM-Masked beats the M4 winner on 5/6 groups and on the weighted total, and is competitive with the best published results," not "SOTA." We keep the two provenances separate throughout: ES-RNN (REPORTED official winner) and PRISM-Masked (NEW, reproduced from our own code; checksum-verified; re-running reproduces published OWA to within 0.6%). The masking is what makes M4 tractable for this family: on M4's front-padded, variable-length contexts the mask-aware encoder is required, whereas on fixed LTSF windows it reduces to the identity used elsewhere.

Table 13: **M4 OWA (lower better).** ES-RNN = REPORTED official winner; PRISM-Masked = NEW (single model, three-seed mean ± std). Bold = PRISM-Masked beats ES-RNN.

| Group | ES-RNN (REPORTED) | PRISM-Masked (3-seed mean ± std) | masked − ES-RNN |
|---|---|---|---|
| Yearly | 0.778 | 0.7192 ± 0.0054 | −0.059 |
| Quarterly | 0.847 | 0.7854 ± 0.0014 | −0.062 |
| Monthly | 0.836 | 0.8607 ± 0.0023 | +0.025 |
| Weekly | 0.851 | 0.7339 ± 0.0211 | −0.117 |
| Daily | 1.046 | 0.9719 ± 0.0053 | −0.074 |
| Hourly | 0.440 | 0.4161 ± 0.0031 | −0.024 |
| **Weighted** | 0.821 | 0.8125 ± 0.002 | −0.009 |

The study also yielded four training lessons that carry over to LTSF and that we adopted in the locked configuration. First, setting dropout to 0 in the attention and decoder helped on M4 (Hourly −0.047, Weekly −0.018). Second, RevIN / per-series instance normalization was preferable to a global scaler for heavy-tailed series, since a global scaler failed catastrophically on Hourly. Third, early stopping with best-checkpoint selection improved stability. Fourth, seed variance was non-trivial on the small-count frequencies (Weekly std 0.021), which is why we report three-seed means and flag multi-seed averaging as a reporting requirement for the real-LTSF results.

## M ILI Companion (PRISM-only)

20 PRISM results (seed = 1, $L = 96$, `target_channel=0`); ILI is outside the modern TFB 7-dataset protocol, so no comparable baseline join is made (protocol-incompatibility caveat). We report ILI as a companion track only and make no comparative claim.

## N Summary of Claims and Evidence

For transparency we summarize each principal claim alongside the evidence that supports it and the nature of that claim (fact, empirical result, interpretation, limitation, or hypothesis). The mapping is intended to help readers locate the supporting analysis within the paper.

Table 14: Summary of claims and their supporting evidence. Tag legend: [FACT], [EMPIRICAL RESULT], [INTERPRETATION], [LIMITATION], [HYPOTHESIS].

| # | Claim | Evidence (experiment → metric → result) | Tag |
|---|---|---|---|
| 1 | No component is individually novel | Section 2 positioning against prior art | [FACT] |
| 2 | Model equations match the released reference implementation | Section 3, per-equation correspondence | [FACT] |
| 3 | Params = 247,172/423,140 | Direct enumeration; Appendix K | [FACT] |
| 4 | Heads recover their factor gauge-aware (3-seed) | Section 7: S2 0.73, S3 0.99, REGIME 0.94 | [EMPIRICAL] |
| 5 | Heads stay inert where factor absent | S1/S4 scalar $R^2 \approx 0$; HIGH_NOISE scalars $\approx 0$ | [EMPIRICAL] |
| 5b | Decomposition holds numerically (recon $1.2\text{–}9.5 \times 10^{-7}$, 546 rows) | Appendix G | [FACT] |
| 6 | Raw per-head recovery weaker than gauge-aware $R^2$ | S2 slopes 0.36–0.46; S3 shape_prod_corr 0.66–0.90 | [LIMITATION] |
| 7 | Decomposition "faithful" (gauge-aware; scale recovery moderate, level strong) | Interpretation of claims 4–5, gauge-aware only; raw scale slopes 0.36–0.46 | [INTERPRETATION] |
| 8 | Multi-patch does not beat single-patch on real data | Section 6: mp 6/17 vs sp8 | [EMPIRICAL] |
| 9 | PRISM competitive-or-better on 27% of cells | Table 1: 7 WIN/5 COMP/32 LOSS, seed 1 | [EMPIRICAL] |
| 10 | PRISM is not SOTA | Table 1 (32 losses) | [EMPIRICAL] |
| 11 | Wins associate with high trend | Appendix J: WIN 0.87 vs LOSS 0.29 (24 cells) | [EMPIRICAL] |
| 12 | Trend association correlational, not causal | Discussion framing, 6/11 datasets | [INTERPRETATION] |
| 13 | Synthetic scale/volatility mechanism inverts on real data | WIN vol 0.47 vs LOSS 0.62 | [EMPIRICAL] |
| 14 | Capacity effect small/mixed | 423k better 14/24, mean −1.63% | [EMPIRICAL] |
| 15 | Context length inconclusive | No controlled sweep | [LIMITATION] |
| 16 | Gate/Fourier/fold inert or harmful; CI beats mixing | Section 6 (H96 diagnostic) | [EMPIRICAL] |
| 17 | Level head necessary in S3 (+21.3%) | Appendix H | [EMPIRICAL] |
| 18 | Scale head not demonstrably necessary anywhere | Appendix H, $n = 1$ | [EMPIRICAL] |
| 19 | Level head harmful in APERIODIC (−13.2%) | Appendix H | [EMPIRICAL] |
| 20 | Faithfulness ≠ necessity | Sections 7–8 comparison | [INTERPRETATION] |

| # | Claim | Evidence | Tag |
|---|---|---|---|
| 20b | Shape-only ties-or-beats full decomposition on real data ($n = 1$) | Appendix F: ETTh1 −7.1%, ETTh2 −7.4%, ETTm1 −1.6%, ETTm2 tied | [EMPIRICAL] |
| 21 | PRISM-Masked beats ES-RNN (M4 winner) on 5/6 groups | Appendix L: 3-seed weighted OWA 0.8125 vs 0.821 (REPORTED) | [EMPIRICAL] |
| 22 | Masking is load-bearing on M4, identity on fixed-window LTSF | Reasoning: padding exists on M4, not LTSF | [INTERPRETATION] |
| 23 | ES-RNN is the official M4 winner | REPORTED (M4-methods) | [FACT] |
| 24 | PRISM is lightweight (∼0.25M, ∼2040 samp/s) | Appendix K | [EMPIRICAL] |
| 25 | A simpler PRISM would lose little | Inference from claims 8, 14, 16 | [HYPOTHESIS] |
| 26 | Multi-seed averaging affects point estimates | M4 per-seed std up to 0.021 (Weekly) | [EMPIRICAL] |

**Outstanding methodological confounds.** Two comparability confounds remain open. First, the ETT train-split difference (0.8 vs 0.75) coincides with the cells where PRISM wins; resolving it would require re-running at least one beaten baseline at 0.8, or PRISM at 0.75. Second, the direct single-head-versus-decomposition ablation on real data is now run at a single seed (shape-only ties-or-beats full on all four completed datasets, ETTh1 −7.1%, ETTh2 −7.4%, ETTm1 −1.6%, ETTm2 tied; Appendix F), but remains single-seed/single-horizon, so the empirical value of the decomposition on real LTSF is only partially established. Both are stated as limitations in the main text (Section 10).