WHEN DOES COARSE-TO-FINE FORECASTING HELP? AN EXACT ERROR DECOMPOSITION AND A CHECKABLE CRITERION

Anonymous authors
Paper under double-blind review

ABSTRACT

Coarse-grained future structure is often easier to predict than fine-grained dynamics, motivating a broad class of coarse-to-fine and multiscale forecasting methods. Yet **predictability does not necessarily imply predictive value**: under the same forecasting protocol, conditioning on a coarse-grained future can substantially improve a matched backbone, have little effect, or even degrade the final prediction. This raises a more fundamental question than whether a particular hierarchy works: **when can predictable coarse-grained information actually be converted into better fine-grained forecasts?** We study this question by separating three quantities that are often conflated by standard forecasting metrics: the predictability of future coarse-grained structure, the value of the resulting coarse anchor relative to a matched unconditional forecaster, and the additional error that can be recovered through fine-grained refinement. This decomposition yields a simple necessary condition that can be evaluated before training the refinement model. We partition the final error into a component determined by coarse prediction and a component that refinement can potentially recover. If the former already exceeds the error of the target direct forecaster, then no refinement can make the hierarchical predictor outperform that baseline. Across 28 dataset–horizon settings from seven multivariate forecasting benchmarks and four prediction horizons, this condition identifies eight such cases, and the hierarchical method loses in all eight. The remaining cases reveal qualitatively different success and failure regimes. In some settings, the coarse anchor itself accounts for nearly all of the improvement; in others, substantial refinement is required. More importantly, a harmful anchor does not necessarily imply failure, nor does improving an anchor guarantee a net gain. On Electricity, refinement fully compensates for an initially harmful anchor and ultimately surpasses the baseline. On long-horizon Traffic, refinement improves the anchor but remains insufficient to outperform the direct forecaster. Targeted interventions further separate **information availability** from **model realization**. On ETTm1, coarse future structure remains predictable at long horizons even when the learned hierarchy fails to exploit it. Improving only coarse extraction or only refinement provides partial recovery, whereas improving both produces a strongly coupled rescue effect. On Traffic, increasing coarse resolution reduces unresolved within-block structure and anchor error, turning $H = 336$ from a net loss into a net gain and substantially narrowing the deficit at $H = 720$. These results show that **coarse-grained predictability and realized predictive value are distinct properties**. The success of coarse-to-fine forecasting depends not only on whether coarse future structure exists and can be predicted, but also on whether it can be converted into a useful anchor relative to the target baseline, and whether refinement can recover the information that remains unresolved after anchoring.

1 INTRODUCTION

Long-horizon time-series forecasting maps a finite history to an extended future trajectory. Modern systems use direct regression, decomposition, patching, temporal mixing, and multiscale representa-

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

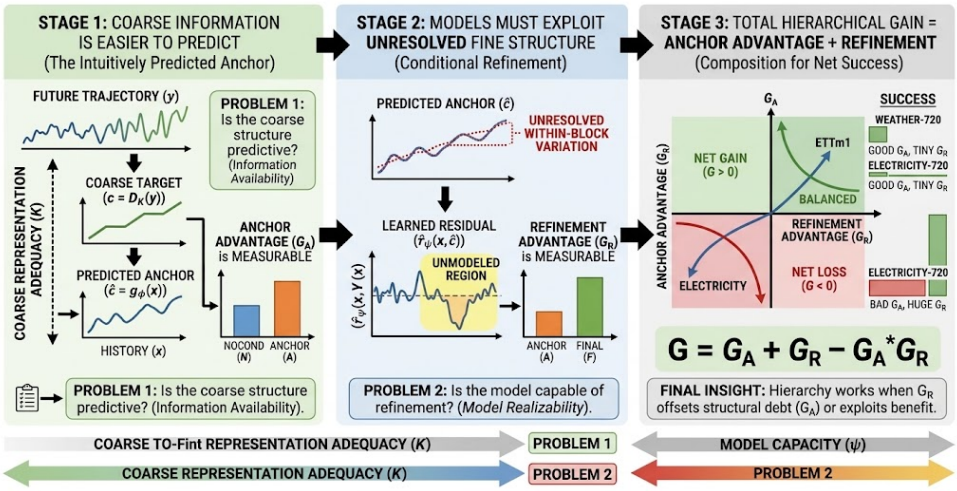

Figure 1: Core thesis and measurement framework. Coarse predictability is an information-availability question: can the future coarse structure be inferred from history? Anchor advantage measures whether the resulting coarse forecast improves over NoCond, while refinement advantage measures whether the model can recover the fine structure left unresolved by that anchor. The final gain depends on both terms.

tions (Zhou et al., 2021; Wu et al., 2021; Zhou et al., 2022; Zeng et al., 2023; Nie et al., 2023; Liu et al., 2024; Wang et al., 2024). A recurring motivation for the multiscale family is that slow structure can remain predictable when fine-scale fluctuations become difficult to extrapolate. A model may therefore first estimate a coarse future and then refine it.

That intuition is incomplete because a coarse-to-fine predictor contains two different sources of behavior. It first constructs a coarse *anchor*; it then tries to recover information that remains unresolved after anchoring. A strong anchor may explain almost all of the benefit, making refinement redundant. Conversely, refinement may improve a poor anchor while the complete system still performs worse than a matched direct predictor. Aggregate MSE cannot distinguish these cases.

We ask:

> *When can predictable coarse-grained information actually be converted into better fine-grained forecasts?*

Let N denote the test MSE of a matched unconditional forecaster, A the MSE of the expanded coarse anchor, and F the MSE after refinement. We define

$$G_A = \frac{N - A}{N}, \qquad G_R = \frac{A - F}{A}, \qquad G = \frac{N - F}{N}. \tag{1}$$

G_A measures the structural advantage or debt introduced by the anchor, while G_R measures recovery relative to that anchor. Since $A = N(1 - G_A)$ and $F = A(1 - G_R)$,

$$\boxed{G = G_A + G_R - G_A G_R.} \tag{2}$$

This identity turns "does hierarchy help?" into two measurable questions: is the anchor useful relative to not anchoring, and how much remaining error can refinement recover?

This framework yields four descriptive regimes, which we define and illustrate in Section 5.

The key contrast is between Electricity, where refinement overcompensates for a harmful anchor, and long-horizon Traffic, where refinement improves the anchor but does not recover enough. Thus, neither a good anchor nor a positive refinement gain is individually sufficient for final success.

The four regimes describe how the two parts contribute, not how much the method is worth, because a gain over a matched backbone says nothing about whether it beats a strong direct forecaster. To

make that dependence explicit we add a target direct forecaster: a plain linear model trained under the identical protocol. This changes the picture. Eq. 8 implies that the hierarchical predictor can beat a target forecaster of error E only when the part determined by coarse prediction is already smaller than E, and in the settings where that fails no refinement can help. We therefore report every gain twice, against the matched backbone and against the target forecaster, and we find settings where the method gains more than 60% over its own backbone while losing to the target forecaster on the same data.

We validate this interpretation with targeted interventions. On ETTm1, a ten-seed 2×2 experiment independently replaces the learned coarse extractor and the refinement rule. At $H = 720$, changing either component alone is insufficient, while changing both produces a strongly super-additive rescue. On Traffic, increasing the coarse length K reduces within-block unresolved energy and anchor error. The final gain changes from negative to positive at $H = 336$, while at $H = 720$ it approaches but does not cross zero. Finally, repeating the whole grid with a patch-tokenized second backbone preserves the sign of the final gain in 27 of 28 settings.

Our contributions are:

- an exact decomposition of coarse-to-fine forecasting into a coarse prediction error and a within-block term, together with a bound that shows how much refinement can and cannot recover;
- a necessary condition that can be checked before the refinement model is trained; on 28 dataset–horizon settings it identifies eight in which the method cannot win, and the method loses in all eight;
- evidence that gains over a matched backbone can invert the conclusion, so that a 63% gain over the backbone coexists with a 26% loss against the target forecaster;
- controlled interventions that separate information availability from model realization;
- a coarse-resolution intervention that identifies representation adequacy as a contributor to Traffic failure; and
- an auditable protocol with matched controls, train-only coarse predictors, seed-level results, data hashes, and negative-result reporting.

The appendix is organized to follow the main text. Appendix A gives the algebraic identity of Eq. 2, Appendix A.1 the per-setting decomposition and the check of Eq. 10, and Appendix B the datasets, training details, and statistical conventions. Appendices C and D give the full grid and the cross-dataset decomposition, Appendix E the published baselines, and Appendix F the same comparison in MAE. The interventions and robustness checks are in Appendices G, H, I, J, K, and L. Appendix M follows one setting from the history to the verdict, Appendices N.1 and N.2 give instance-level cases, Appendix 5 shows the regime map, and Appendix P maps questions to evidence.

2 Related Work

Long-horizon forecasting has progressed from recurrent models (Hochreiter & Schmidhuber, 1997; Salinas et al., 2020) to efficient Transformers (Zhou et al., 2021), decomposition (Wu et al., 2021; Zhou et al., 2022), linear baselines (Zeng et al., 2023), patch-based models (Nie et al., 2023), inverted tokenization (Liu et al., 2024), and multiscale mixing (Wang et al., 2024). These methods are strong predictors, but their usual comparison is an aggregate error table; our analysis is complementary, and we use the linear baseline as an explicit reference because it is trained under the same protocol and is competitive on these benchmarks.

Multiscale and hierarchical models exploit several scales for reconciliation (Athanasopoulos et al., 2017; Wickramasuriya et al., 2019; Rangapuram et al., 2021), interpolation (Challu et al., 2023), decomposition (Wu et al., 2021; Zhou et al., 2022), or scale mixing (Wang et al., 2024), and multi-resolution computation is standard elsewhere, from feature pyramids and shifted-window hierarchies in vision (Lin et al., 2017; Ronneberger et al., 2015; Liu et al., 2021) to diffusion-style refinement in forecasting (Rasul et al., 2021; Tashiro et al., 2021). We ask a different question: when is a *predicted future coarse representation* a useful anchor, and when does a downstream refiner contribute beyond it? A multiscale model can improve because the coarse forecast is accurate, because refinement recovers fine structure, or because a strong refiner compensates for a poor anchor, and only the last two are refinement effects.

Standard ablations remove modules and compare final scores, which can conflate information content, estimation quality, and downstream usage. We therefore combine ANCHORONLY controls, train-only predictability probes, explicit residual anchoring, coarse-resolution interventions, second-backbone checks, checkpoint perturbations, and seed-level lineage, and we report MSE with paired seed effects and bootstrap summaries rather than a single favourable run (Hyndman & Koehler, 2006; Efron & Tibshirani, 1993).

3 PROBLEM FORMULATION

Let $x \in \mathbb{R}^{L\times C}$ be an observed multivariate history and $y \in \mathbb{R}^{H\times C}$ the future target. A matched direct forecaster produces

$$\hat{y}_N = f_\theta(x). \tag{3}$$

Let D_K map the future to K coarse blocks:

$$c = D_K(y), \qquad c_k = \frac{1}{b}\sum_{t=kb}^{(k+1)b-1} y_t, \quad b = H/K, \tag{4}$$

and let U_H expand a coarse trajectory to the original horizon. A coarse predictor estimates $\hat{c} = g_\phi(x)$ and yields

$$\hat{y}_A = U_H(\hat{c}). \tag{5}$$

We separately measure coarse predictability as the out-of-sample coefficient of determination

$$P_C = R_c^2 = 1 - \frac{\mathrm{MSE}(\hat{c}, c)}{\mathrm{Var}(c)}, \tag{6}$$

using a predictor fitted on training windows only. P_C measures information availability, not predictive value: a predictable coarse target can still produce a poor anchor if the learned condition is not transmitted or used effectively by the refinement model. For residual anchoring, the final forecast is

$$\hat{y}_F = U_H(\hat{c}) + r_\psi(x, \hat{c}), \tag{7}$$

where the predicted residual is centered within each coarse block. This constraint forces the refinement branch to model unresolved within-block structure rather than freely rewriting the anchor.

3.1 EXACT ERROR DECOMPOSITION AND A NECESSARY CONDITION

We partition the final error into a part determined by coarse prediction and a part that refinement can potentially recover. Write the within-block component of the target as $w = y - U_H D_K(y)$, so that $y = U_H c + w$ and w is block-mean zero. Define the coarse prediction error $\varepsilon_C^2 = \mathrm{MSE}(\hat{c}, c)$, the within-block energy $\sigma_W^2 = \mathrm{MSE}(w, 0)$, and the residual error $\rho = \mathrm{MSE}(w, r_\psi)$. Since $U_H(c - \hat{c})$ is constant within each block while both w and the centered residual average to zero within each block, the two parts do not overlap under the reported metric. The anchor and final errors therefore decompose exactly:

$$A = \varepsilon_C^2 + \sigma_W^2, \qquad F = \varepsilon_C^2 + \rho. \tag{8}$$

Eq. 8 is an identity, not an approximation; we verify that $A - \varepsilon_C^2 - \sigma_W^2$ is zero to floating-point tolerance in every reported run. It is diagnostic in a way that Eq. 2 is not, because it distinguishes an anchor that is wrong at the block level (large ε_C^2) from a refiner that fails to recover structure the anchor cannot express (large ρ). Two consequences follow.

First, refinement cannot repair anchor error, because the part determined by coarse prediction, ε_C^2, appears in both A and F. Thus

$$G_R = \frac{\sigma_W^2 - \rho}{\sigma_W^2 + \varepsilon_C^2} \le \frac{\sigma_W^2}{A}. \tag{9}$$

The refinement advantage is bounded by the within-block share of the anchor error. No refiner can compensate for an anchor error that is already larger than the target it is compared against.

Second, comparing the hierarchy with a direct forecaster of error E gives a criterion that can be evaluated before the refiner is trained. Because $F \ge \varepsilon_C^2$, the hierarchy can beat that forecaster only if

$$\varepsilon_C^2 < E, \tag{10}$$

and, when Eq. 10 holds, it does beat it exactly when $\rho < E - \varepsilon_C^2$. Both quantities ε_C^2 and σ_W^2 are properties of the data and of the anchor alone, so it can be evaluated without training a refinement model.

The three errors in Eq. 1 are computed under the same split, normalization, horizon, coarse length, backbone, optimizer family, checkpoint rule, and seed. ANCHORONLY is not an oracle: its coarse predictor is fitted using training data only. RIDGERESIDUAL is a deliberately engineered intervention that supplies a deterministic train-only Ridge coarse trajectory; it is used to test mechanism, not to claim a universal deployable forecaster.

4 EXPERIMENTAL PROTOCOL

We evaluate seven multivariate benchmarks (ETTh1, ETTh2, ETTm1, ETTm2, Electricity, Traffic, Weather) over four horizons $H \in \{96, 192, 336, 720\}$, with input length $L = 336$, fixed coarse length $K = 12$, and channel-wise normalization fitted on the training segment only. ETT datasets follow the official chronological split; Weather, Electricity, and Traffic use $7/1/2$ ratio splits. Dataset provenance, training details, and the full variant definitions are in Appendix B.

4.1 MATCHED VARIANTS

The controlled variants are NOCOND (the backbone without a coarse condition), TRAJPRED and RIDGEPRED (a learned, respectively train-only Ridge, coarse trajectory supplied as a condition), TRAJRESIDUAL and RIDGERESIDUAL (the same two supplies with explicit residual anchoring), and ANCHORONLY (the expanded coarse trajectory evaluated without neural refinement). All variants share the split, normalization, optimizer, checkpoint rule, and seed.

4.2 TARGET DIRECT FORECASTER

A matched backbone answers the mechanism question but not the value question, because a coarse-to-fine predictor can beat a weak backbone and still lose to a strong direct forecaster. We therefore add a target direct forecaster: DLinear (Zeng et al., 2023), a two-branch linear model trained under the identical protocol, so that architecture is the only difference.

4.3 TRAINING-BUDGET CONTROL

We hold everything fixed and raise the epoch budget five-fold (100 epochs, patience 10) for NOCOND and RIDGERESIDUAL on all seven datasets at $H = 96$ and $H = 720$. The control lowers the matched error by at most 8% and leaves the final error within 3% of its short-schedule value, while the anchor advantage moves by at most 0.06: the gap to the target forecaster is a property of the backbone rather than of the schedule. Appendix L reports the settings and shows why the necessary condition itself is unaffected.

4.4 METRICS AND CAVEATS

We report test MSE averaged over test windows, future positions, and channels, with the same comparison in MAE in Appendix F. Paired effects use bootstrap intervals over seed-level differences and exact sign-flip tests; with five seeds the smallest attainable two-sided sign-flip probability is 0.0625, so we report directional consistency rather than strong significance. ETTm1 $H = 96$ and $H = 720$ in the initial grid include three earlier-commit runs; a same-commit re-run moves the final gain by at most 0.019 and leaves every regime label unchanged (Appendix I). Ratio splits can differ in difficulty between validation and test segments, so we include checkpoint sensitivity and phrase the cross-dataset findings as mechanism evidence.

5 RESULTS: REGIMES OF THE MATCHED COMPARISON

This section describes how the hierarchy behaves relative to its own matched backbone; Section 6 then asks whether that behaviour amounts to predictive value. Table 4 reports the final gains over

the matched backbone, and Appendix C gives the full grid with the (G_A, G_R) decomposition. The boundary at $G_R = 0.05$ is a descriptive convention, not a natural phase transition, and the sign of G is always computed from N and F.

The grid contains 13 anchor-dominant settings, 7 anchor-plus-refinement settings, 6 bad-anchor but overcompensated settings, and 2 bad-anchor net-loss settings. This distribution rules out a binary "hierarchy helps or hurts" interpretation within the matched comparison; Section 6 then asks what that comparison is worth against an external forecaster. Across the 28 settings, RIDGERESIDUAL is the best variant in 26 settings and loses to NOCOND only on Traffic-$H = 336$ and Traffic-$H = 720$. Its gain over the matched backbone ranges from -30.68% to $+86.46\%$, but that range hides two different situations: on ETTh2, ETTm2, and Weather almost all of the gain comes from the anchor, whereas on ETTm1, ETTh1, and Electricity a measurable part comes from refinement.

5.1 THE FOUR REGIMES

Anchor-dominant settings ($G_A > 0$ with G_R near zero: ETTh2, ETTm2, Weather) have a useful anchor and almost no refinement gain, because their within-block energy is small; ETTm2-$H720$ is the limit case, where refinement leaves the anchor error unchanged (0.8525 against 0.8573). *Anchor + Refinement* settings ($G_A > 0$ and $G_R > 0$: ETTh1 at every horizon, ETTm1 at $H \geq 192$) carry a useful anchor and a substantial second stage. *Bad anchor, overcompensated* settings ($G_A < 0$ but $G > 0$; Electricity-$H720$: $G_A = -1.433$, $G_R = 0.717$) contradict the assumption that the anchor must be good. *Bad anchor, net loss* settings ($G_A < 0$ and $G < 0$; Traffic at $H \geq 336$) combine a large anchor debt with incomplete recovery. The labels are descriptive regions that we read continuously, and ETTm1-$H96$ ($G_R = 0.044$) sits on the boundary.

6 PREDICTIVE VALUE AGAINST A TARGET DIRECT FORECASTER

The regimes above are computed relative to the matched backbone, which answers the mechanism question but not whether the method is worth using. This section tests the necessary condition of Eq. 10 against a target direct forecaster: we compare RIDGERESIDUAL, the strongest variant in the grid, with DLinear trained under the identical protocol (Section 4).

Table 1 gives both readings for every setting: the gain G over the matched backbone and the ratio F/E of the final error to the target forecaster. The predictor beats it only on Weather (0.85–0.96) and comes within 14% on ETTm1. It loses on the other five datasets, from ETTh1 (1.03–1.31) and Electricity (1.17–1.32) to ETTm2 (1.29–2.17), Traffic (1.54–2.00), and ETTh2 (1.57–1.71). Twenty-four of the 28 settings lose to a linear model with 1.3×10^5 parameters trained under the identical protocol; the four exceptions are the Weather settings.

The criterion separates two different reasons for losing. On all four ETTh2 and all four ETTm2 settings the coarse prediction error alone exceeds the error of the target forecaster, and the refiner recovers essentially nothing. Those eight settings are out of reach for any refiner. The remaining twenty satisfy $\varepsilon_C^2 < E$, and there the outcome is decided by how much of the budget $E - \varepsilon_C^2$ the refiner spends. Weather wins at every horizon ($\rho = 0.029$ against a budget of 0.055 at $H = 96$) and ETTm1 is decided by 0.004 MSE; ETTh1 spends 0.212 against a budget of 0.114, Electricity spends 0.143 against 0.107, and Traffic spends 0.866 against 0.397. Eq. 9 states the same limit from the other side: on ETTh2-$H96$ the within-block share of the anchor error is 0.14, so refinement could remove at most 14% of an anchor error that is already larger than the reference error.

Two readings of the same model therefore disagree. On ETTh1-$H96$ the matched comparison reports $G = +0.631$, while the same trained model is 26% worse than the target forecaster. Since the anchor and the target forecaster use identical splits and budgets, the disagreement is a difference of denominator, not of protocol. We therefore report both numbers for every setting and treat a gain over the matched backbone as evidence about mechanism only. Originally reported PatchTST results are lower still on several of these settings (ETTh2-$H96$: 0.274 against our 0.288 target forecaster and our 0.453 final), so the target forecaster is a floor rather than a ceiling. The necessary condition should therefore be read against whichever direct forecaster an application actually uses: in practice the strongest one available. Here that is DLinear, and Appendix E lists stronger published baselines for reference.

Table 1: Two readings of the same runs. For every setting, G is the final gain of RIDGERESIDUAL over its matched backbone and F/E is the ratio of its final error to that of the target direct forecaster, so $F/E > 1$ means the method loses. All entries are five-seed means at fixed $K = 12$. † marks the datasets whose coarse prediction error alone already exceeds the target forecaster at every horizon, so that no refiner can win there (Eq. 10); the per-setting check is in Appendix A.1.

	$H = 96$		$H = 192$		$H = 336$		$H = 720$	
Dataset	G	F/E	G	F/E	G	F/E	G	F/E
ETTh1	$+.631$	1.26	$+.613$	1.31	$+.601$	1.27	$+.637$	1.03
ETTh2†	$+.872$	1.57	$+.796$	1.68	$+.748$	1.65	$+.605$	1.71
ETTm1	$+.644$	1.01	$+.673$	1.04	$+.628$	1.10	$+.533$	1.14
ETTm2†	$+.852$	1.29	$+.817$	1.50	$+.782$	1.83	$+.760$	2.17
Electricity	$+.384$	1.32	$+.392$	1.28	$+.330$	1.30	$+.313$	1.17
Traffic	$+.028$	1.65	$+.124$	1.54	$-.105$	1.77	$-.307$	2.00
Weather	$+.294$	0.85	$+.401$	0.86	$+.382$	0.91	$+.311$	0.96

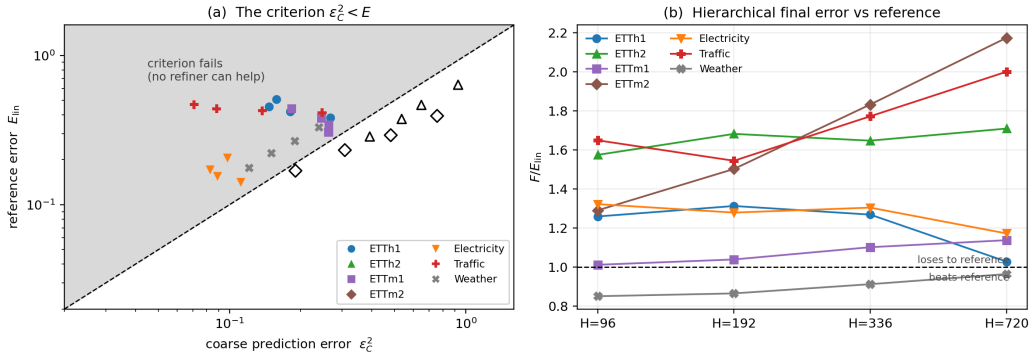

Figure 2: Criterion check on the 28 setting grid. Panel (a): each marker is one dataset–horizon setting, positioned by its coarse prediction error ε_C^2 and the DLinear reference error E_{lin}; filled markers satisfy $\varepsilon_C^2 < E$, open markers violate it and lie in the shaded region where $F \geq \varepsilon_C^2 > E$ makes a win impossible. Panel (b): the ratio of the hierarchical final error to the reference error; only Weather stays below the dashed line, and only ETTm1 stays close to it.

7 CONTROLLED INTERVENTION ON ETTM1

The grid is descriptive; to test whether the two terms correspond to separable mechanisms we intervene on coarse extraction and refinement independently, using the ten-seed ETTm1 factorial (NOCOND, TRAJPRED, RIDGEPRED, TRAJRESIDUAL, RIDGERESIDUAL) at $H \in \{96, 720\}$. At $H = 720$ the learned coarse probe has $P_C = -0.403$ while the train-only Ridge probe has $P_C = 0.704$, so coarse structure is available but the learned extractor does not realize it. Mean test MSE is 1.028 (NOCOND), 1.164 (TRAJPRED), 1.012 (RIDGEPRED), 1.167 (TRAJRESIDUAL), and 0.494 (RIDGERESIDUAL). The learned condition is nominally harmful under the paired sign-flip test ($+0.136$, one of ten wins, $p = 0.0332$); replacing only the extractor is near-neutral ($+0.48\%$) and changing only the refinement rule remains harmful, whereas the joint intervention gains 51.2%. The failure is systematic: the learned probe has non-positive P_C in 12 of 28 settings (all ETTh1 and ETTh2), while the Ridge probe is positive everywhere, minimum 0.341 (Appendix D).

We quantify non-additivity with

$$I = \text{MSE}(\text{RIDGERESIDUAL}) - \text{MSE}(\text{RIDGEPRED}) - \text{MSE}(\text{TRAJRESIDUAL}) + \text{MSE}(\text{TRAJPRED}). \tag{11}$$

The interaction is $I = -0.520$ at $H = 720$ and $I = -0.226$ at $H = 96$, with all ten seeds agreeing on the negative direction and exact sign-flip $p = 0.00195$ in both cases. Thus, better coarse extraction and explicit residual anchoring are Eq. 8 explains the negative sign instead of merely reporting it: the part determined by coarse prediction enters both errors alike, so a poor coarse extractor cannot be

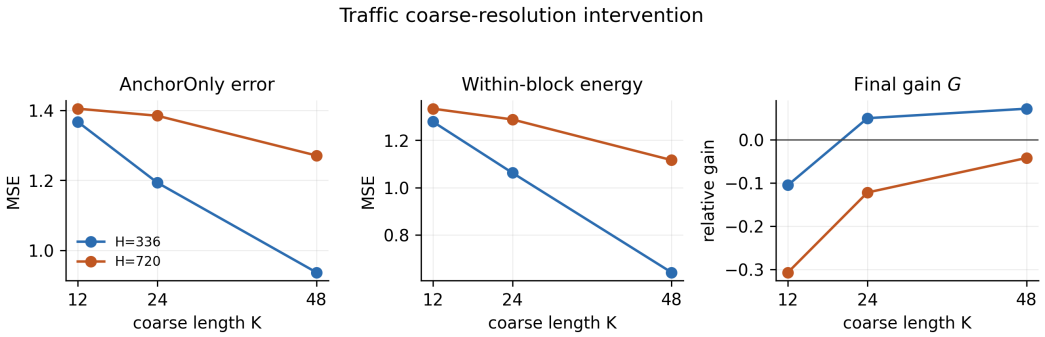

Figure 3: Traffic coarse-resolution intervention. Increasing K reduces AnchorOnly error and within-block energy. At $H = 336$ the final gain crosses zero; at $H = 720$ the deficit narrows but remains negative at $K = 48$.

repaired downstream and the two interventions must be judged jointly. The interaction is negative in all 28 grid settings, 24 of them at the five-seed sign-flip floor (Appendix G).

At ETTm1-$H96$ the Ridge anchor has MSE 0.3232 against a final 0.3093 (coarse share 85%, recovery 4.31%), while at $H = 720$ the anchor is 0.6689, the final is 0.4926, and recovery is 26.36%. ETTm1 is therefore not representative: ETTh2, ETTm2, and Weather stay anchor-dominated even at $H = 720$, and Electricity shows the largest measured recovery.

8 Traffic: Coarse Resolution as an Intervention

Traffic long horizons combine a large anchor debt with substantial within-block variation. We vary K with the rest of the protocol fixed; Figure 3 reports anchor MSE, within-block energy, and final gain.

For $H = 336$, increasing K from 12 to 24 to 48 reduces AnchorOnly MSE from 1.3667 to 1.1930 to 0.9367, reduces within-block energy from 1.2785 to 1.0621 to 0.6411, and changes final gain from -0.105 to $+0.055$ to $+0.073$. For $H = 720$, the same intervention reduces AnchorOnly MSE from 1.4044 to 1.3842 to 1.2706 and within-block energy from 1.3335 to 1.2879 to 1.1162; final gain improves from -0.307 to -0.122 to -0.042.

These changes support coarse under-resolution as a contributor to the Traffic failure, not as its only cause: $H = 720$ is still negative at $K = 48$. More resolution reduces the anchor debt, reverses the tested $H = 336$ setting, and only partly repairs the hardest horizon.

9 Robustness and Boundary Checks

9.1 Second Backbone

We repeat the entire grid with a patch-tokenized second backbone on the same channel-mixing trunk; it is deliberately not a faithful reproduction of official PatchTST, whose channel-independent design would add a confound. Final-gain signs agree with the main grid in 27 of the 28 settings. The exception is Traffic-$H336$, where the second backbone's NoCond baseline is worse by more than the hierarchical result, while G_A stays negative under both. Anchor quality is therefore more stable than refinement recoverability under this change.

9.2 Checkpoint Sensitivity

On Weather-$H720$, Electricity-$H720$, and Traffic-$H720$ we compare validation-selected checkpoints with longer patience and with last-epoch evaluation. No regime label flips and absolute values move by a few percent at most, which confirms that ratio-split datasets need cautious interpretation.

9.3 SCOPE

We make no leaderboard claim: the predictor loses to a plain linear target forecaster on 24 of the 28 settings, and published PatchTST results are lower still. The contribution is the decomposition, the necessary condition, and the demonstration that gains over a matched backbone can invert the conclusion.

10 DISCUSSION

10.1 WHAT THE DECOMPOSITION MEASURES

Eq. 8 separates two quantities that a single error score conflates: how accurately the anchor predicts the block structure, and how much of the remaining within-block variation the refiner recovers. Because the part determined by coarse prediction is shared by the anchor and final errors, a positive refinement gain is compatible with an anchor that is worse than doing nothing, and a large final gain is compatible with losing to a strong direct forecaster. Reporting the three errors together prevents both misreadings. Reporting the three errors together prevents both misreadings.

10.2 GAINS OVER A MATCHED BACKBONE ARE NOT PREDICTIVE VALUE

The comparison in Section 6 is the practical consequence. A matched backbone isolates the mechanism, which is why we use it, but it also inherits whatever weakness that backbone has. The same trained model gains 63% against its own backbone and loses 26% against the reference, and in eight settings the loss is structural because the coarse prediction error alone exceeds the reference error. Any claim that hierarchy improves long-horizon forecasting therefore needs a reference forecaster, not only a matched control.

10.3 PRACTICAL IMPLICATION

For a coarse-to-fine model, report the anchor error, the within-block energy, the final error, and at least one strong direct reference. Evaluate Eq. 10 before training the refiner: if ε_C^2 exceeds the reference error, no refiner can rescue the setting. If the criterion holds, the budget $E - \varepsilon_C^2$ says how much the refinement stage still has to recover, and Eq. 9 bounds what it can recover at all.

11 LIMITATIONS

First, the broad grid uses one controlled timestep backbone, and the second backbone is patch-tokenized rather than an official PatchTST replication. Second, RIDGERESIDUAL is a diagnostic intervention, not a deployable method. Third, the ratio splits give validation and test segments of different difficulty, so absolute scores and checkpoint selection are protocol-sensitive. Fourth, the regime labels are descriptive regions, not natural phase transitions, and settings near $G_R = 0.05$ should be read continuously. Finally, settings and overlapping windows are dependent and the five-seed sign-flip floor limits formal significance; the comparison uses one target forecaster under one budget, which is why Eq. 10 is stated for a general E.

12 CONCLUSION

We asked when a coarse-to-fine forecaster earns its complexity. The anchor and final errors decompose exactly into a coarse prediction error and a within-block term. Two things can then be checked before the refinement model is trained: refinement can never repair anchor error, and the hierarchical predictor can beat a target forecaster of error E only when $\varepsilon_C^2 < E$. Across 28 settings and a matched target forecaster, this condition identifies the eight settings in which the method cannot win, and the method loses in all eight. Gains over a matched backbone are not evidence of predictive value: on ETTh1-$H96$ the method gains 63% over its own backbone and loses 26% to the target forecaster. Hierarchy is a conditional inductive bias whose value should be established against the forecaster it is meant to replace.

Ethics Statement

This work uses public time-series benchmarks and does not involve human subjects, private records, or interventions on individuals. Potential operational use in energy or traffic systems should retain human oversight because forecast errors can affect resource allocation and safety decisions.

Reproducibility Statement

Every reported result is linked to a run-level record containing dataset path and SHA256, split, normalization rule, model variant, horizon, coarse length, seed, code snapshot, checkpoint selection rule, execution device, and final metrics. The appendix gives the full definitions of the decomposition, the matched variants, the intervention tables, and the principal caveats. All empirical figures and tables in the main paper are generated from frozen CSV result ledgers; the conceptual overview is an explanatory figure.

AI Use Statement

Generative AI tools were used as research assistants for brainstorming, experiment organization, draft code scaffolding, paper outlining, and language editing. All empirical results were produced by the authors' computational experiments. The authors reviewed the code, data protocol, mathematical statements, statistical summaries, references, figures, and final claims, and take responsibility for the final content.

REFERENCES

George Athanasopoulos, Rob J. Hyndman, Nikolaos Kourentzes, and Fotios Petropoulos. Forecasting with temporal hierarchies. *European Journal of Operational Research*, 262(1):60–74, 2017.

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

Bradley Efron and Robert J. Tibshirani. An introduction to the bootstrap. *Monographs on Statistics and Applied Probability*, 57, 1993.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997.

Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006.

Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 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.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

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.

Syama Sundar Rangapuram, Lucien D. Werner, Konstantinos Benidis, Pedro Mercado, Jan Gasthaus, and Tim Januschowski. End-to-end learning of coherent probabilistic forecasts for hierarchical time series. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

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

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3): 1181–1191, 2020.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. In *Advances in Neural Information Processing Systems*, 2021.

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*, 2024.

Shanika L. Wickramasuriya, George Athanasopoulos, and Rob J. Hyndman. Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, 114(526):804–819, 2019.

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.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, 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 *Proceedings of the 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 *Proceedings of the 39th International Conference on Machine Learning*, 2022.

A Decomposition and the Necessary Condition

From Eq. 1, $A = N(1 - G_A)$ and $F = A(1 - G_R)$. Therefore

$$F = N(1 - G_A)(1 - G_R) = N\left(1 - G_A - G_R + G_A G_R\right).$$

Rearranging gives $G = (N - F)/N = G_A + G_R - G_A G_R$. This is an identity, not an empirical approximation. It requires only that $N > 0$ and $A > 0$, which hold for the reported MSE values. The identity also explains why G_R must be paired with G_A: a large recovery fraction can be measured relative to an anchor whose absolute error is already unacceptable.

A.1 Exact decomposition and a criterion check

Table 2 reports the coarse prediction error ε_C^2, the within-block energy σ_W^2, the residual error ρ, the upper bound σ_W^2/A on what refinement can recover, and the reference error E_{lin} for every setting where the decomposition was reconstructed. The column labelled "recon." is the residual of Eq. 8; its magnitude is at floating-point level in every row, which verifies the identity numerically rather than assuming it. The criterion column evaluates Eq. 10. The same evidence is plotted in Figure 2.

Table 2: Exact error decomposition and criterion check on all available settings. ε_C^2 is the coarse prediction error, σ_W^2 the within-block energy, $\rho = F - \varepsilon_C^2$ the residual error, E_{lin} the DLinear reference. The reconstruction residual $A - \varepsilon_C^2 - \sigma_W^2$ is reported to show that the identity of Eq. 8 holds numerically.

Dataset	H	A	ε_C^2	σ_W^2	ρ	σ_W^2/A	E_{lin}	F/E_{lin}	recon.	Eq. 10
ETTh1	96	0.564	0.268	0.296	0.212	0.525	0.381	1.26	+3e-09	holds
ETTh1	192	0.683	0.181	0.502	0.369	0.735	0.419	1.31	-5e-09	holds
ETTh1	336	0.716	0.147	0.569	0.426	0.794	0.452	1.27	-1e-08	holds
ETTh1	720	0.739	0.158	0.581	0.362	0.786	0.506	1.03	-2e-08	holds
ETTh2	96	0.457	0.391	0.066	0.061	0.144	0.288	1.57	+6e-09	**fails**
ETTh2	192	0.635	0.535	0.100	0.099	0.157	0.377	1.68	+5e-09	**fails**
ETTh2	336	0.771	0.646	0.125	0.125	0.162	0.468	1.65	-8e-08	**fails**
ETTh2	720	1.087	0.924	0.163	0.164	0.150	0.637	1.71	-6e-08	**fails**
ETTm1	96	0.323	0.263	0.060	0.046	0.187	0.305	1.01	-3e-09	holds
ETTm1	192	0.400	0.264	0.136	0.092	0.341	0.342	1.04	-2e-11	holds
ETTm1	336	0.507	0.245	0.261	0.172	0.516	0.379	1.10	-2e-08	holds
ETTm1	720	0.669	0.184	0.485	0.316	0.726	0.439	1.14	+8e-09	holds
ETTm2	96	0.219	0.190	0.029	0.028	0.131	0.169	1.29	+1e-10	**fails**
ETTm2	192	0.351	0.307	0.044	0.042	0.124	0.232	1.50	+2e-09	**fails**
ETTm2	336	0.541	0.480	0.061	0.058	0.113	0.293	1.83	-6e-08	**fails**
ETTm2	720	0.852	0.756	0.097	0.102	0.114	0.394	2.17	-1e-07	**fails**
Electricity	96	0.437	0.112	0.326	0.075	0.744	0.141	1.32	+2e-09	holds
Electricity	192	0.717	0.089	0.628	0.109	0.876	0.155	1.28	-2e-10	holds
Electricity	336	0.811	0.083	0.729	0.140	0.898	0.171	1.30	-6e-09	holds
Electricity	720	0.852	0.098	0.754	0.143	0.885	0.205	1.17	-2e-08	holds
Traffic	96	0.950	0.247	0.703	0.435	0.740	0.413	1.65	-1e-08	holds
Traffic	192	1.264	0.138	1.126	0.520	0.891	0.426	1.54	+4e-09	holds
Traffic	336	1.367	0.088	1.278	0.690	0.935	0.439	1.77	-4e-09	holds
Traffic	720	1.404	0.071	1.334	0.866	0.950	0.468	2.00	-4e-09	holds
Weather	96	0.150	0.121	0.029	0.029	0.194	0.176	0.85	-1e-09	holds
Weather	192	0.193	0.151	0.042	0.041	0.219	0.221	0.86	-2e-09	holds
Weather	336	0.246	0.189	0.057	0.054	0.231	0.267	0.91	-5e-10	holds
Weather	720	0.325	0.239	0.086	0.078	0.263	0.330	0.96	-3e-09	holds

B DATA, TRAINING, AND STATISTICAL DETAILS

The six controlled variants are defined as follows: NOCOND is the matched backbone without a coarse condition; TRAJPRED is a learned coarse trajectory supplied as a condition; RIDGEPRED replaces that learned trajectory with a train-only channel-wise Ridge prediction; TRAJRESIDUAL and RIDGERESIDUAL add explicit residual anchoring to the learned and the Ridge trajectory respectively; and ANCHORONLY expands the coarse trajectory without any neural refinement. The Ridge conditioner uses 24 history bins, $\alpha = 1.0$, at most 4096 evenly sampled training windows, and no validation or test tuning. The Traffic intervention varies $K \in \{12, 24, 48\}$ and the ETTm1 factorial uses ten predeclared seeds at fixed $K = 12$.

The evidence is layered: the broad grid contains 700 run configurations (7 datasets $\times$ 4 horizons $\times$ 5 variants $\times$ 5 seeds), the residual-anchor audit adds 280 run-level decompositions, and the ETTm1 factorial adds 100. The reference and robustness layers add a DLinear target forecaster on all 28 settings with five seeds (140 configurations), a patch-tokenized second backbone on all 28 settings with three seeds (168), a ten-seed Electricity factorial (50), and the training-budget control (84). Every run keeps the SHA256 of its dataset file, split boundaries, seed, checkpoint rule, device, and final metrics.

We use the official ETT-small files for ETTh1, ETTh2, ETTm1, and ETTm2, and the standard benchmark files for Weather, Electricity, and Traffic. The following table records the row count, number of numeric channels, split rule, and SHA256 prefix for the exact files consumed by the broad grid. The date column is excluded from the numeric target channels.

Table 3: Dataset provenance and split protocol.

Dataset	Rows	Channels	Split	SHA256 prefix
ETTh1	17,420	7	official ETT	f18de3ad
ETTh2	17,420	7	official ETT	a3dc2c59
ETTm1	69,680	7	official ETT	6ce1759b
ETTm2	69,680	7	official ETT	db973ca2
Weather	52,696	21	train/val/test=.7/.1/.2	34ee981d
Electricity	26,304	321	train/val/test=.7/.1/.2	7e45845d
Traffic	17,544	862	train/val/test=.7/.1/.2	cb06463d

All experiments use input length $L = 336$, batch size 32, a two-layer timestep Transformer encoder and two-layer Transformer decoder with $d_{\text{model}} = 128$, eight attention heads, feed-forward width 256, and dropout 0.1. Training uses Adam with learning rate 5×10^{-4}, zero weight decay, a 20-epoch budget, patience 3, and $\lambda_{\text{coarse}} = 1$. The broad grid uses the timestep backbone and fixed $K = 12$; therefore the coarse ratio is H/K, not one shared temporal ratio across horizons. The patch-tokenized check uses patch length 16 and stride 8.

For Ridge interventions, the conditioner is fit on training windows only, uses 24 history bins, $\alpha = 1.0$, and at most 4096 evenly sampled training windows. The saved Ridge tensors are identical across seeds at a fixed horizon. For residual variants, the predicted residual is centered within each coarse block before being added to the repeated coarse anchor. The reconstruction identity and maximum residual block mean are checked for every reported run.

For clarity, the residual target is the within-block component of the normalized future, $y - U_H(D_K(y))$. The residual branch receives the history and the predicted coarse trajectory, predicts a full-resolution residual, and applies block-wise centering before reconstruction; the final output is the repeated predicted anchor plus this centered residual. TRAJPRED and RIDGEPRED use the same coarse-conditioned forecasting objective without this residual constraint, whereas TRAJRESIDUAL and RIDGERESIDUAL use the residual-anchored objective. The coarse loss weight is $\lambda_{\text{coarse}} = 1$ for learned coarse prediction; train-only Ridge variants do not fit Ridge weights using validation or test targets.

B.1 ADDITIONAL STATISTICAL AND PROTOCOL DETAILS

The broad grid uses five seeds per setting. We report means and sample standard deviations, paired seed-level effects, and exact sign-flip tests. Because five seeds give a two-sided sign-flip floor of 0.0625, the broad grid is used to establish regime structure and directional consistency, not to claim

Table 4: Final gain G over the matched backbone for RIDGERESIDUAL relative to NOCOND across the 28-setting broad grid. Values are means over five seeds at fixed $K = 12$; negative entries are retained as failure cases. These gains are computed against the matched backbone only, and Section 6 shows that they do not imply a gain over the target forecaster.

Dataset	$H = 96$	$H = 192$	$H = 336$	$H = 720$	broad pattern
ETTh1	+.631	+.613	+.601	+.637	refinement grows
ETTh2	+.872	+.796	+.748	+.605	anchor-dominant
ETTm1	+.644	+.673	+.628	+.533	anchor → refinement
ETTm2	+.852	+.817	+.782	+.760	anchor-dominant
Electricity	+.384	+.392	+.330	+.313	overcompensated
Traffic	+.028	+.124	-.105	-.307	long-horizon boundary
Weather	+.294	+.401	+.382	+.311	anchor-dominant

Table 5: Representative settings from the four regimes. Values are fixed-$K = 12$ means from the broad grid.

Setting	N	A	F	G_A	G_R	G
Weather-720	0.461	0.325	0.318	+0.294	+0.022	+0.310
ETTm1-720	1.069	0.669	0.500	+0.375	+0.253	+0.533
Electricity-720	0.350	0.852	0.241	-1.433	+0.717	+0.313
Traffic-720	0.717	1.404	0.937	-0.961	+0.333	-0.307

that every setting independently reaches conventional statistical significance. The ETTm1 mechanism intervention uses ten seeds, giving a minimum exact sign-flip probability of 0.00195; the reported interaction effects are computed before aggregation using matched seed rows.

All test-window metrics average squared error over future positions and channels. Forecast windows overlap, so window-level observations are not treated as independent replicates for the mechanism claims. The train-only Ridge audit reconstructs the saved weight tensors from training windows and verifies zero maximum difference. Residual decomposition additionally checks that the cross term is numerically zero up to floating-point tolerance and that the final block mean equals the selected anchor.

C FULL 28-SETTING SUMMARY

D CROSS-DATASET MECHANISM DECOMPOSITION

E PUBLISHED REFERENCE NUMBERS

Table 8 lists originally reported results for PatchTST and DLinear at $H = 96$ together with our own numbers on the same datasets. Two things should be read from it. First, our DLinear reproduces the originally reported DLinear values to within a few percent, which is the evidence that our protocol is comparable rather than merely matched internally. Second, PatchTST is stronger than our target forecaster on every dataset listed, so the criterion of Eq. 10 is conservative: with a stronger reference the threshold E falls and more settings fail it. These numbers come from the original publications and were produced under different schedules and data splits, so they are context rather than a controlled comparison.

F MEAN ABSOLUTE ERROR

The main text reports MSE; Table 9 reports the same comparison in MAE. The ordering of settings is unchanged, which indicates that the reference comparison is not an artefact of the squared-error metric.

Table 6: Full fixed-$K = 12$ phase grid. Values are mean MSE over five seeds; G is final gain relative to NoCond.

Dataset	H	N	A	F	G_A	G_R	G
ETTh1	96	1.300	0.564	0.480	+0.566	+0.148	+0.631
ETTh1	192	1.421	0.683	0.550	+0.519	+0.195	+0.613
ETTh1	336	1.438	0.716	0.574	+0.502	+0.199	+0.601
ETTh1	720	1.432	0.739	0.521	+0.484	+0.296	+0.637
ETTh2	96	3.539	0.457	0.453	+0.871	+0.009	+0.872
ETTh2	192	3.112	0.635	0.634	+0.796	+0.001	+0.796
ETTh2	336	3.062	0.771	0.771	+0.748	-0.000	+0.748
ETTh2	720	2.758	1.087	1.089	+0.606	-0.001	+0.605
ETTm1	96	0.867	0.323	0.309	+0.627	+0.044	+0.644
ETTm1	192	1.086	0.400	0.355	+0.632	+0.111	+0.673
ETTm1	336	1.122	0.507	0.418	+0.548	+0.176	+0.628
ETTm1	720	1.069	0.669	0.500	+0.375	+0.253	+0.533
ETTm2	96	1.470	0.219	0.218	+0.851	+0.002	+0.852
ETTm2	192	1.904	0.351	0.349	+0.816	+0.005	+0.817
ETTm2	336	2.465	0.541	0.537	+0.781	+0.007	+0.782
ETTm2	720	3.574	0.852	0.857	+0.761	-0.006	+0.760
Electricity	96	0.303	0.437	0.187	-0.442	+0.573	+0.384
Electricity	192	0.326	0.717	0.198	-1.203	+0.724	+0.392
Electricity	336	0.333	0.811	0.223	-1.435	+0.725	+0.330
Electricity	720	0.350	0.852	0.241	-1.433	+0.717	+0.313
Traffic	96	0.701	0.950	0.681	-0.355	+0.283	+0.028
Traffic	192	0.750	1.264	0.658	-0.684	+0.479	+0.124
Traffic	336	0.704	1.367	0.778	-0.940	+0.431	-0.105
Traffic	720	0.717	1.404	0.937	-0.960	+0.333	-0.307
Weather	96	0.212	0.150	0.150	+0.292	+0.003	+0.294
Weather	192	0.320	0.193	0.191	+0.397	+0.007	+0.401
Weather	336	0.394	0.246	0.243	+0.374	+0.012	+0.382
Weather	720	0.461	0.325	0.318	+0.295	+0.022	+0.311

Table 7: Cross-dataset residual decomposition for RidgeResidual at the shortest and longest tested horizons. R is refinement gain relative to the matched piecewise-constant anchor; S is the coarse-component fraction of final MSE. Values are means over five seeds.

Dataset	R_{96}	R_{720}	S_{96}	S_{720}
ETTh1	+.148	+.296	.558	.305
ETTh2	+.009	-.001	.864	.849
ETTm1	+.044	+.253	.851	.368
ETTm2	+.002	-.006	.871	.882
Electricity	+.573	+.717	.598	.408
Traffic	+.283	+.333	.362	.076
Weather	+.003	+.022	.808	.753

G ETTm1 Ten-Seed Intervention

Table 12 extends the interaction check of Eq. 11 to the full grid. Every setting has a negative interaction and 24 of 28 reach the five-seed sign-flip floor, so the complementarity of coarse extraction and refinement holds across all seven datasets rather than only on ETTm1.

Table 13 reports the coarse predictability of both probes on the full grid.

H Electricity Ten-Seed Intervention

The ETTm1 factorial of Appendix G is repeated on Electricity, a 321-channel dataset with a ratio split, to check that the two interventions behave the same way outside the ETT family. Ten seeds are used per setting (five from the broad grid and five new ones), with the same coarse length $K = 12$ and the same variant definitions.

RidgeResidual beats NoCond on all ten seeds at both horizons, by 39.8% at $H = 96$ and by 31.4% at $H = 720$ in paired terms, and the interaction of Eq. 11 is -0.074 at $H = 96$, the same negative sign as in every setting of Table 12. Electricity is also the setting with the largest measured recovery in the 28-setting grid (Table 2), so the construction succeeds here for the reason the decomposition predicts: the within-block part is large but recoverable, not the anchor error small.

Table 8: Originally reported PatchTST and DLinear results at $H = 96$ next to our own numbers on the same datasets. Published values are those reported by Nie et al. (2023) for PatchTST and by Zeng et al. (2023) for DLinear, measured under their own protocols.

Dataset	PatchTST (reported)	DLinear (reported)	our DLinear	our RIDGERESIDUAL
ETTh1	0.370	0.375	0.381	0.480
ETTh2	0.274	0.289	0.288	0.453
ETTm1	0.293	0.299	0.305	0.309
ETTm2	0.166	0.167	0.169	0.218
Electricity	0.130	0.140	0.141	0.187
Traffic	0.367	0.410	0.413	0.681
Weather	0.152	0.176	0.176	0.150

Table 9: Mean absolute error of RIDGERESIDUAL relative to NOCOND on the broad grid, $\mathrm{MAE}(\text{RIDGERESIDUAL})/\mathrm{MAE}(\text{NOCOND})$. Values below 1 favour the hierarchical model.

Dataset	$H = 96$	$H = 192$	$H = 336$	$H = 720$
ETTh1	0.50	0.51	0.53	0.52
ETTh2	0.30	0.39	0.44	0.54
ETTm1	0.51	0.47	0.50	0.58
ETTm2	0.32	0.36	0.38	0.39
Electricity	0.77	0.77	0.82	0.82
Traffic	1.18	1.02	1.23	1.43
Weather	0.78	0.74	0.73	0.80

I LINEAGE CHECK AND TEN-SEED EXTENSIONS

Two robustness questions remain for the factorial: whether the regimes survive a different code state, and whether they survive ten seeds rather than five.

Table 15 re-runs all four ETTm1 horizons under a single code snapshot, which removes the one group of settings that mixed earlier commits in the broad grid. The final gain moves by at most 0.019 and no regime label changes. Table 16 extends six further settings from five to ten seeds. The same three-way pattern appears in each of them: a useful anchor with almost no refinement gain (ETTm2-$H720$, Weather-$H720$), a useful anchor with a substantial second stage (ETTm1-$H720$), a negative anchor that refinement overcompensates (Electricity-$H720$), and a negative anchor that stays negative (Traffic-$H720$). On Traffic-$H336$ one seed stopped at epoch 1, where its test error is twice the setting median; removing that seed leaves the final gain negative, which is why the main text treats the setting as a net loss.

J TRAFFIC RESOLUTION AND CHECKPOINT TABLES

K SECOND BACKBONE

The patch-tokenized second backbone uses patch length 16 and stride 8 on the same channel-mixing trunk. It is deliberately not a faithful reproduction of official PatchTST, whose channel-independent design would introduce an additional confound. Table 19 reports all 28 settings with three seeds each. Final-gain signs agree in 27 of them. The only sign change is Traffic-$H336$, where the reference baseline of the second backbone is worse by more than the hierarchical result improves; the bad-anchor diagnosis is stable because $G_A < 0$ under both backbones.

L TRAINING-BUDGET CONTROL

A short schedule is the most obvious alternative explanation for the regime structure, because the matched backbone is weaker than the target forecaster on most settings. The criterion is however insensitive to it by construction. ε_C^2 and σ_W^2 are computed from a train-only Ridge anchor, so they

Table 10: Ten-seed ETTm1 intervention. Values are mean ± sample standard deviation of test MSE.

Variant	$H = 96$	$H = 720$
NoCond	0.855 ± 0.070	1.028 ± 0.125
TrajPred	0.803 ± 0.117	1.164 ± 0.124
RidgePred	0.655 ± 0.085	1.012 ± 0.058
TrajResidual	0.683 ± 0.078	1.167 ± 0.182
RidgeResidual	0.309 ± 0.001	0.494 ± 0.015

Table 11: Paired effects and interactions for the ten-seed ETTm1 intervention.

H	Comparison	Δ MSE	Gain	Wins	exact p
96	TrajPred-NoCond	-0.052	+5.22%	7/10	.285
96	RidgePred-NoCond	-0.199	+22.70%	10/10	.00195
96	TrajResidual-NoCond	-0.172	+19.42%	10/10	.00195
96	RidgeResidual-NoCond	-0.546	+63.64%	10/10	.00195
96	interaction I	-0.226	–	10/10	.00195
720	TrajPred-NoCond	+0.136	-14.80%	1/10	.0332
720	RidgePred-NoCond	-0.016	+0.48%	7/10	.658
720	TrajResidual-NoCond	+0.139	-15.41%	1/10	.109
720	RidgeResidual-NoCond	-0.534	+51.22%	10/10	.00195
720	interaction I	-0.520	–	10/10	.00195

are functions of the data and the split alone: across the five seeds the anchor error has zero sample standard deviation in every one of the 28 settings (largest observed value 1.2×10^{-16}). A longer schedule can therefore move ρ and the matched error N, but it cannot move which settings satisfy $\varepsilon_C^2 < E$. We treat the criterion as a necessary condition and the schedule as a separate quantity, and we run a companion control that holds the data splits, normalization, optimizer, learning rate, coarse length, and checkpoint rule fixed while raising the epoch budget five-fold for NoCond and RidgeResidual on all seven datasets at $H = 96$ and $H = 720$. Table 20 reports the settings where all three seeds completed at both budgets. Five-fold more epochs lowers the matched error by at most 8% and leaves the final error within 3% of its short-schedule value, while the anchor advantage moves by at most 0.06. The gap between the matched backbone and the reference is therefore a property of this backbone rather than of the schedule, and the eight structural settings of Section 6 remain out of reach at any budget.

M A Worked Example from Start to Finish

To make the measurement concrete, we follow two settings from the raw history to the final verdict: ETTh1 at $H = 96$, where the method gains over its backbone and still loses, and Weather at $H = 96$, where it wins. Both use the same protocol, the same coarse length $K = 12$, and the same Ridge-anchored variant.

The steps are the same in both cases. (1) Build the coarse target by splitting the $H = 96$ future into $K = 12$ blocks of eight steps and averaging each block (Eq. 4). (2) Predict the anchor from history alone with the train-only Ridge conditioner and expand it back to full resolution (Eq. 7). (3) Split the anchor error into the part determined by coarse prediction, ε_C^2, and the within-block part, σ_W^2, using Eq. 8; the two add up to the anchor error exactly. (4) Train the refinement branch on the within-block component and read off the residual error ρ. (5) Check the necessary condition of Eq. 10 against the target forecaster, and compare the budget $E - \varepsilon_C^2$ with ρ.

Table 21 reports every quantity for the two settings. On ETTh1-$H96$ the anchor is useful — $A = 0.564$ against $N = 1.300$, so $G_A = +0.566$ — and the method gains $G = +0.631$ over its own backbone. The decomposition shows where that gain stops being enough: the coarse prediction part is 0.268 and the within-block part is 0.296, so the necessary condition holds against a target forecaster at $E = 0.381$, but the refinement branch leaves $\rho = 0.212$ of the recoverable part, more than the budget of 0.114. The final error is therefore 0.480, and the method loses to the target forecaster by 26%. On Weather-$H96$ the reverse holds: the recoverable part is small ($\sigma_W^2 = 0.029$) and the refiner leaves only $\rho = 0.029$, well inside the budget of 0.055, so the final error of 0.150 is 15% below the target forecaster.

Table 12: The 2×2 interaction of Eq. 11 on the full grid, five seeds. I compares the joint change of coarse extraction and refinement with the sum of the two isolated changes, so a negative value means the two interventions are complementary rather than additive. Wins counts seeds agreeing on the negative direction; p is the exact sign-flip probability, whose five-seed floor is 0.0625.

Dataset	$H = 96$	$H = 192$	$H = 336$	$H = 720$	sign-flip floor settings
ETTh1	-0.48 (5/5)	-0.69 (5/5)	-0.53 (5/5)	-0.76 (5/5)	4/4
ETTh2	-0.74 (4/5)	-2.70 (5/5)	-1.59 (5/5)	-1.24 (5/5)	3/4
ETTm1	-0.26 (5/5)	-0.41 (5/5)	-0.76 (5/5)	-0.57 (5/5)	4/4
ETTm2	-0.25 (5/5)	-0.78 (5/5)	-1.21 (5/5)	-2.31 (5/5)	4/4
Electricity	-0.07 (5/5)	-0.16 (5/5)	-0.10 (5/5)	-0.11 (5/5)	4/4
Traffic	-0.03 (4/5)	-0.03 (3/5)	-0.13 (5/5)	-0.11 (4/5)	1/4
Weather	-0.02 (5/5)	-0.05 (5/5)	-0.10 (5/5)	-0.14 (5/5)	4/4

Table 13: Coarse predictability P_C of the learned extractor and of the train-only Ridge probe on the full grid. A non-positive P_C means the probe is worse than predicting the mean, so the learned extractor does not realize coarse structure that the deterministic probe recovers from the same history.

Dataset	$H = 96$	$H = 192$	$H = 336$	$H = 720$	Ridge minimum
ETTh1	-0.09	-0.53	-0.61	-0.60	0.67
ETTh2	-0.17	-1.27	-0.69	-0.56	0.34
ETTm1	+0.35	+0.16	-0.06	-0.40	0.70
ETTm2	+0.53	+0.20	-0.20	-3.23	0.48
Electricity	+0.69	+0.46	+0.34	+0.29	0.61
Traffic	+0.56	+0.36	+0.20	+0.15	0.56
Weather	+0.67	+0.51	+0.40	+0.24	0.51

The pair makes the paper's two points at once. The necessary condition is only a floor: it holds in both settings, yet one of them loses, because holding it says nothing about how much of the recoverable part the refiner will actually recover. And a gain over the matched backbone is not a verdict: the same ETTh1-$H96$ run reports $+63\%$ against its own backbone and -26% against the target forecaster, so the two numbers have to be read together.

N Case Studies

N.1 A setting where the anchor carries the forecast: ETTm1 at $H = 720$

We provide an instance-level case study to complement setting-level averages. It uses the ETTm1 test set and the seed-42 checkpoints, and contains two selected test windows (indices 15 and 125), each with 720 future steps for the OT channel: the ground truth, the direct NoCond prediction, the refined prediction, and the upsampled coarse sketch. The windows were selected from the first 256 test windows by the improvement gap between the direct and refined predictions; this selection rule is disclosed so that the figure is interpreted as an illustrative case rather than an unbiased average.

For these windows, the visual distinction is most pronounced after the early part of the horizon: the direct prediction contracts toward a smoother trajectory, whereas the refined forecast follows the low-frequency envelope while retaining more of the target's changing amplitude. The sketch is intentionally smoother than the target; its role is not to reproduce every fluctuation, but to provide a usable anchor for the residual branch. This example therefore answers a narrower question than the grid: what does a favorable anchor and refinement pair look like in one forecast?

N.2 A setting where refinement is not enough: Traffic at $H = 720$

Traffic-$H720$ supplies the complementary negative case. Under $K = 12$, the train-only Ridge anchor has MSE 1.4044, compared with 0.7166 for NoCond. Residual refinement reduces the final error to 0.9369, so $G_R = 0.333$ is positive, but the final gain remains $G = -0.307$. The block-level anchor is therefore improved without becoming competitive with the direct forecaster.

The resolution intervention makes the failure interpretable rather than anecdotal. Increasing K from 12 to 24 to 48 reduces within-block energy from 1.3335 to 1.2879 to 1.1162 and anchor MSE from 1.4044 to 1.3842 to 1.2706. The final gain correspondingly moves from -0.307 to -0.122 to

Table 14: Electricity ten-seed factorial. Values are mean ± sample standard deviation of test MSE over ten seeds; the last column is the paired difference of NoCond minus RidgeResidual.

Variant	$H = 96$	$H = 720$	paired Δ ($H = 96$ / $H = 720$)
NoCond	0.3100 ± 0.0120	0.3517 ± 0.0091	–
TrajPred	0.2910 ± 0.0073	0.3304 ± 0.0152	–
RidgePred	0.2756 ± 0.0084	0.3375 ± 0.0284	–
TrajResidual	0.2757 ± 0.0045	0.3364 ± 0.0148	–
RidgeResidual	0.1865 ± 0.0011	0.2412 ± 0.0046	$+0.1235$ / $+0.1104$

Table 15: ETTm1 same-commit lineage check. All four horizons are re-run under one code snapshot with five seeds; ΔG is the change of the final gain relative to the broad grid.

H	N	A	F	G_A	G_R	G	ΔG
96	0.838	0.323	0.309	$+0.614$	$+0.045$	$+0.632$	-0.012
192	1.086	0.400	0.355	$+0.632$	$+0.111$	$+0.673$	0.000
336	1.122	0.507	0.418	$+0.548$	$+0.176$	$+0.628$	0.000
720	1.084	0.669	0.486	$+0.383$	$+0.273$	$+0.551$	$+0.019$

-0.042. Thus, the failure is consistent with insufficient coarse representation, but the residual deficit at $K = 48$ prevents the stronger claim that resolution alone explains the entire long-horizon failure.

O Figure Provenance

Figure 5 uses the same 28-setting ledger as Table 4. Figure 3 is generated directly from the Traffic K values reported in Table 17; no empirical plot is manually redrawn.

P Questions and Evidence

This appendix is organized around the questions the proposed measurement framework is intended to answer.

Table 16: Ten-seed extensions. Five additional seeds are added to the five of the broad grid for six settings; N, A, and F are means over ten seeds.

Setting	N	A	F	G_A	G_R	G
ETTm1-$H720$	1.002	0.669	0.489	$+0.333$	$+0.270$	$+0.513$
ETTm2-$H720$	4.149	0.853	0.861	$+0.794$	-0.010	$+0.793$
Weather-$H720$	0.477	0.325	0.317	$+0.319$	$+0.025$	$+0.336$
Electricity-$H720$	0.353	0.852	0.242	-1.413	$+0.716$	$+0.315$
Traffic-$H336$	0.866	1.367	0.785	-0.579	$+0.426$	$+0.093$
Traffic-$H720$	0.695	1.404	1.019	-1.020	$+0.274$	-0.465

Table 17: Traffic resolution intervention.

H	K	N	A	F	G	within-block
336	12	.7044	1.3667	.7781	-.105	1.2785
336	24	.7239	1.1930	.6838	+.055	1.0621
336	48	.6875	.9367	.6376	+.072	.6411
720	12	.7166	1.4044	.9369	-.307	1.3335
720	24	.6970	1.3842	.7815	-.122	1.2879
720	48	.7141	1.2706	.7439	-.042	1.1162

Table 18: Checkpoint sensitivity on representative ratio-split settings.

Setting	Selection	N	F	regime
Weather-720	val MSE	.4609	.3177	anchor-dominant
Weather-720	last epoch	.4673	.3249	anchor-dominant
Electricity-720	val MSE	.3502	.2407	overcompensated
Electricity-720	last epoch	.3206	.2344	overcompensated
Traffic-720	val MSE	.7166	.9369	bad anchor, net loss
Traffic-720	last epoch	.7243	.9380	bad anchor, net loss

Table 19: Patch-tokenized second backbone on the full grid (three seeds). Δ marks whether the final-gain sign agrees with the main grid. The Ridge anchor is identical across backbones, so G_A reflects the anchor and G the full pipeline.

Setting	H	N	F	G	G_A	G_R	sign
ETTh1	96	1.299	0.459	+0.647	+0.566	+0.186	✓
ETTh1	192	1.199	0.510	+0.575	+0.430	+0.254	✓
ETTh1	336	1.300	0.490	+0.623	+0.449	+0.316	✓
ETTh1	720	1.313	0.500	+0.619	+0.437	+0.323	✓
ETTh2	96	3.786	0.453	+0.880	+0.879	+0.008	✓
ETTh2	192	3.986	0.632	+0.842	+0.841	+0.006	✓
ETTh2	336	3.054	0.771	+0.748	+0.748	+0.000	✓
ETTh2	720	3.283	1.089	+0.668	+0.669	-0.001	✓
ETTm1	96	0.792	0.309	+0.610	+0.592	+0.044	✓
ETTm1	192	0.902	0.356	+0.605	+0.556	+0.110	✓
ETTm1	336	0.912	0.420	+0.539	+0.444	+0.170	✓
ETTm1	720	1.000	0.493	+0.507	+0.331	+0.263	✓
ETTm2	96	1.785	0.218	+0.878	+0.877	+0.002	✓
ETTm2	192	1.687	0.349	+0.793	+0.792	+0.004	✓
ETTm2	336	2.067	0.537	+0.740	+0.738	+0.007	✓
ETTm2	720	4.385	0.865	+0.803	+0.806	-0.014	✓
Electricity	96	0.294	0.181	+0.384	-0.487	+0.586	✓
Electricity	192	0.325	0.196	+0.397	-1.208	+0.727	✓
Electricity	336	0.343	0.212	+0.380	-1.367	+0.738	✓
Electricity	720	0.340	0.242	+0.290	-1.503	+0.716	✓
Traffic	96	1.382	0.623	+0.549	+0.313	+0.344	✓
Traffic	192	0.913	0.602	+0.341	-0.384	+0.524	✓
Traffic	336	1.013	0.790	+0.220	-0.349	+0.422	×
Traffic	720	1.019	1.127	-0.107	-0.379	+0.197	✓
Weather	96	0.240	0.150	+0.376	+0.374	+0.003	✓
Weather	192	0.339	0.191	+0.435	+0.431	+0.007	✓
Weather	336	0.394	0.244	+0.382	+0.375	+0.011	✓
Weather	720	0.477	0.319	+0.331	+0.318	+0.018	✓

Table 20: Training-budget control. NoCond and RidgeResidual re-run with a five-fold epoch budget at the two extreme horizons; rows are the settings where all three seeds completed at both budgets. ΔG_A is the change in anchor advantage relative to the main grid.

Dataset	H	N_{20}	N_{100}	F_{20}	F_{100}	G_{100}	ΔG_A
ETTh1	96	1.300	1.230	0.480	0.475	+0.614	+0.025
ETTh1	720	1.432	1.411	0.521	0.524	+0.628	+0.008
ETTh2	96	3.539	3.269	0.453	0.455	+0.861	+0.011
ETTh2	720	2.758	2.621	1.089	1.088	+0.585	+0.021
ETTm1	96	0.867	0.850	0.309	0.309	+0.637	+0.008
ETTm1	720	1.069	1.010	0.500	0.487	+0.518	+0.037
Electricity	96	0.303	0.292	0.187	0.188	+0.358	+0.054

Table 21: Worked example: every quantity of the measurement, from the history to the verdict, for one setting that loses and one that wins. Values are five-seed means at $K = 12$; E is the target direct forecaster.

Quantity	ETTh1-$H96$	Weather-$H96$
Matched backbone N	1.300	0.212
Anchor A	0.564	0.150
Final F	0.480	0.150
Gain over the backbone G	+0.631	+0.294
Coarse prediction part ε_C^2	0.268	0.121
Within-block part σ_W^2	0.296	0.029
Recoverable share σ_W^2/A	0.525	0.194
Residual error ρ	0.212	0.029
Target forecaster E	0.381	0.176
Budget $E - \varepsilon_C^2$	0.114	0.055
Necessary condition (Eq. 10)	holds	holds
Verdict F/E	1.26 (loses)	0.85 (wins)

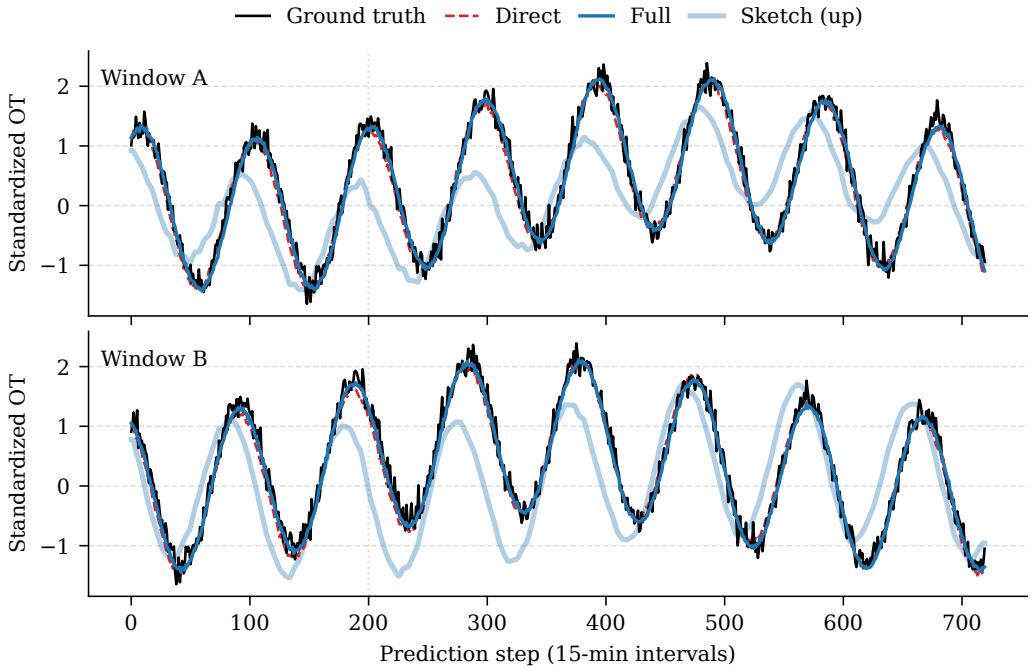

Figure 4: Concrete ETTm1-$H720$ test-window case study for the OT channel. Each panel shows the ground truth, the matched direct forecast, the refined hierarchical forecast, and the upsampled coarse sketch. The figure illustrates the intended mechanism: the sketch supplies a long-range envelope while the residual branch restores within-block variation. It is an instance-level illustration, not a replacement for the seed-level aggregate results.

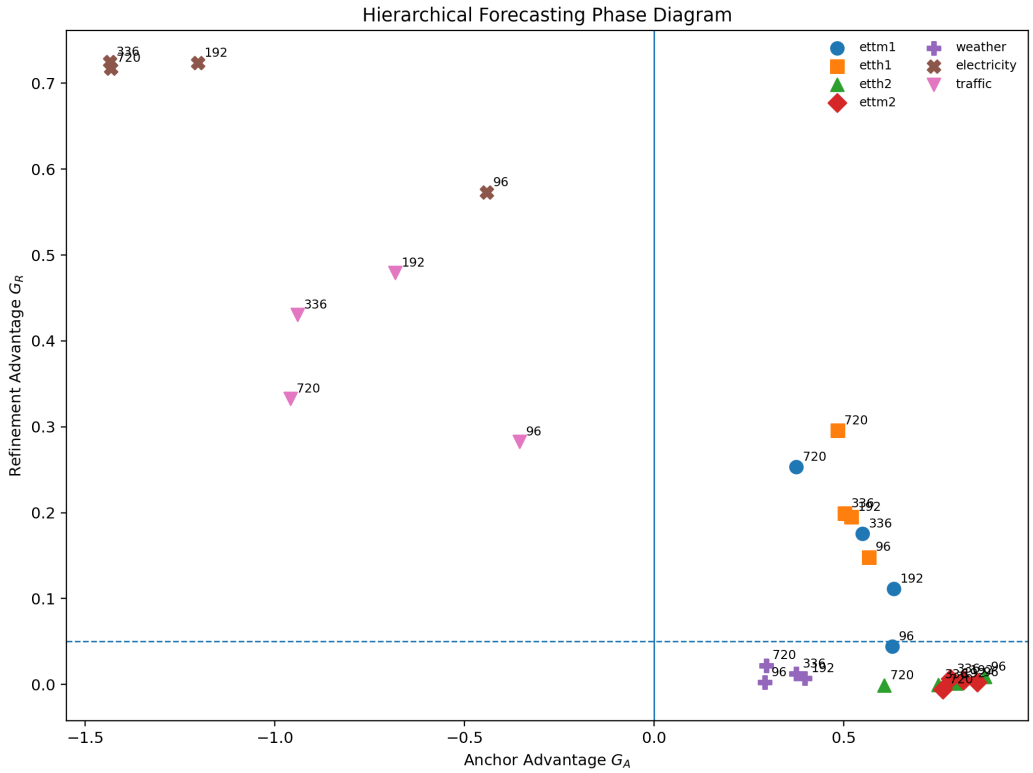

Figure 5: Regime map of anchor advantage versus refinement advantage across the 28 dataset–horizon settings. The horizontal boundary is descriptive only.

Table 22: Questions and the evidence that answers them.

Question	Evidence in this paper
What does the anchor error consist of?	Eq. 8: a coarse prediction error plus a within-block term, verified numerically in Appendix A.1.
Does refinement repair a poor anchor?	No: Eq. 9 bounds the refinement advantage by the within-block share of the anchor error.
Can the pipeline beat a direct forecaster?	Only if $\varepsilon_C^2 < E$; Appendix A.1 checks the criterion on every setting.
Is the matched backbone enough as a control?	No: Section 6 shows a 63% matched gain alongside a 26% loss against a linear reference.
Is the regime structure an artefact of under-training?	Appendix L re-runs the two decisive variants with a five-fold epoch budget.
Can a bad anchor still lead to a good final forecast?	Electricity-$H720$: $G_A < 0$, $G_R > 0$, and $G > 0$.
Can positive refinement still end in failure?	Traffic-$H720$: $G_R > 0$ but $G < 0$; the K intervention explains a partial repair.
Are these effects architecture-specific?	Patch-tokenized second backbone and checkpoint perturbation in Appendix K.
Can a reader inspect an individual forecast?	The ETTm1-$H720$ case study in Appendix N.1.