CODA: TIME-SERIES FORECASTING THROUGH AUDITABLE CONCEPT-STATE TRANSITIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate long-horizon forecasting requires explanations tied to the prediction mechanism. We propose CODA, a Consistency-Driven Concept-State Architecture that learns five softly anchored temporal concepts and predicts through a direct, horizon-specific affine transition. The history encoder remains expressive, while the final head exposes the exact contribution of each concept and the last observation. Analytic targets stabilize concept meaning, and local compatibility penalties regularize the explanatory state. Across six public benchmarks, CODA achieves the lowest error in 10 of 12 primary-horizon MSE/MAE comparisons among the evaluated methods. In a separate multi-horizon evaluation against N-HiTS and ModernTCN, it attains the lowest MSE in 19 of 24 dataset–horizon settings and ties in one more. Concept-target R^2 ranges from 0.62 to 0.94, and removing semantic anchoring increases average MSE from 0.268 to 0.741. Controlled concept and residual ablations, an explicit Electricity coefficient analysis, and input-noise tests connect predictive performance with inspectable model behavior. CODA thus offers accurate forecasting through a compact semantic interface whose downstream effects are analytically verifiable.

1 INTRODUCTION

Forecasting systems in electricity, traffic, public health, and finance need accurate predictions and an inspectable account of how the observed history influences them. Deep forecasters capture nonlinear temporal structure through attention, decomposition, patching, and temporal convolutions (Zhou et al., 2021; Wu et al., 2021; Zhou et al., 2022; Nie et al., 2023; Luo & Wang, 2024), but predictive accuracy alone does not reveal which properties of the observed history determine a forecast. For an unexpectedly high demand prediction, for example, an analyst may want to distinguish a high local level from recent growth or strong oscillation. This calls for a forecasting interface that exposes both what its intermediate variables represent and how they affect future values.

Concept bottleneck models place named variables on the predictive path (Koh et al., 2020; Vandenhirtz et al., 2024). Two requirements become central in forecasting. Concepts must be computed from observations available at the forecast origin, and their downstream use must remain inspectable: meaningful intermediate variables do not by themselves make an opaque decoder interpretable (Rudin et al., 2022). Structured forecasters such as N-BEATS and N-HiTS expose trend/seasonality bases or multiresolution forecast components (Oreshkin et al., 2020; Challu et al., 2023). A complementary approach is to expose a small set of temporal coordinates whose meanings and horizon-specific effects can be evaluated separately. The challenge is to retain enough information for accurate forecasting while making these two forms of inspection concrete.

We introduce **CODA**, a Consistency-Driven Concept-State Architecture (Figure 1). A Transformer encodes a strictly past window into five learned concepts, softly supervised by analytic window statistics. A direct affine head predicts all horizons from these concepts and the last observation. The five coordinates are the only learned state passed to the head; the observed anchor has a separate, explicit contribution. Concept edits have exact coefficient-based effects when the head and anchor are fixed. Training combines forecasting error, semantic alignment, and compatibility penalties expressing an interaction relation and local regularity preferences. Exact sliding-window alternatives are also evaluated.

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.

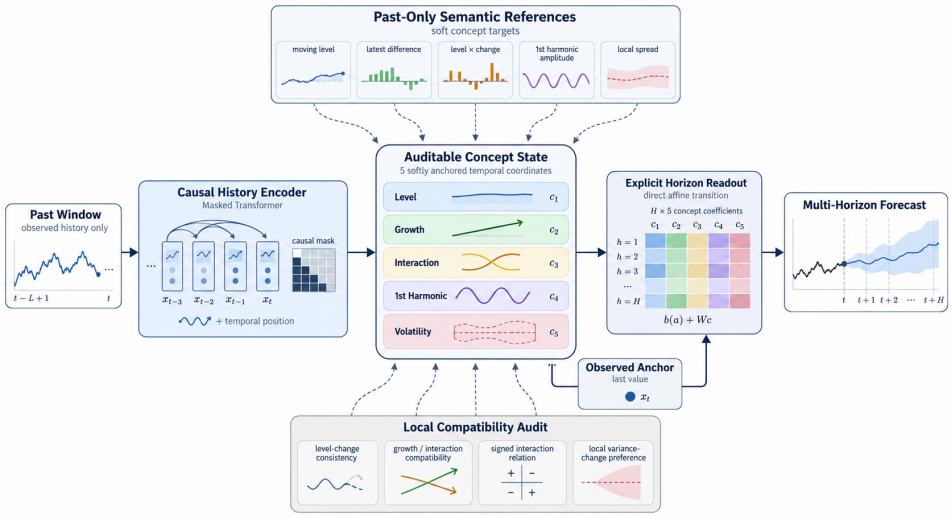


Figure 1: CODA architecture.

Why these five concepts? A central design question is how to choose concepts that are informative without becoming arbitrary. Our choice of *level, growth, signed level–change interaction, first-harmonic amplitude, and local volatility* follows three criteria. First, each concept must be *computable* from a strictly past window, so that it can serve as a legitimate forecasting variable rather than a retrospective summary. Second, each concept captures a distinct and recurrent aspect of local temporal behavior that appears across heterogeneous domains: central tendency, directional change, signal–change interaction, oscillatory strength, and local uncertainty. Third, the set must be *compact but expressive*: rich enough to parameterize useful forecast updates at the evaluated horizons, yet small enough to admit stable semantics and cross-dataset comparison. In this sense, the five concepts are not hand-crafted for any single benchmark, nor are they intended as a unique, minimal, or exhaustive ontology of time-series behavior. They are a deliberately chosen basis for a reusable concept–dynamics interface.

These criteria motivate the design; concept-removal and extension experiments test its predictive value (Section 4.2). Removing any one concept increases mean error, while adding acceleration improves Exchange but worsens the six-dataset mean. Soft supervision also makes the choice distinct from directly forecasting with the five engineered statistics: the learned coordinates can adapt to the task, and their agreement with the targets remains measurable. The corresponding analytic-statistic-plus-affine control tests the benefit of that adaptation.

The interface supports three questions: does a coordinate match its intended statistic, does the head use it, and can its effect be distinguished from other coordinates? Our theory and empirical analyses address these questions through semantic errors, forecast coefficients, horizon-wise cancellation, and concept-level examples. Sensitivity describes the fitted model; interventions on the underlying system require additional causal assumptions.

Our contributions are:

An inspectable forecasting architecture. CODA couples a nonlinear history encoder to softly anchored temporal concepts and a direct horizon-specific affine head. The formulation exposes each concept's contribution and the observed anchor, with explicit compatibility terms and controlled alternatives (Section 3).

A characterization of semantic error and predictive distinguishability. Proposition 1 separates analytic-interface error from weighted semantic mismatch. Proposition 2 characterizes forecast-preserving concept changes, the closest semantic state with the same forecast, and a conditional stability bound. Both results concern the actual prediction head.

Accuracy and concept-level evidence across six benchmarks. CODA attains the lowest error in 10 of 12 primary-horizon MSE/MAE comparisons among the evaluated methods. In a separate multi-horizon comparison with N-HiTS and ModernTCN, it wins 19 of 24 settings and ties once.

Concept-target R^2 ranges from 0.62 to 0.94; ablations, residual alternatives, and the analyses in Section 4.3 assess the roles of semantic supervision, concept choice, and downstream use.

2 Related Work

Long-horizon forecasting combines sparse attention (Li et al., 2019; Zhou et al., 2021), decomposition and frequency structure (Wu et al., 2021; Zhou et al., 2022), patching (Nie et al., 2023), linear models (Zeng et al., 2023), multiscale mixing (Wang et al., 2024), inverted attention (Liu et al., 2024), and temporal convolutions (Luo & Wang, 2024). N-BEATS (Oreshkin et al., 2020) offers an interpretable configuration with explicit trend and seasonality bases; N-HiTS (Challu et al., 2023) assembles multiresolution forecasts by hierarchical interpolation. These structured forecasters are close comparators. CODA instead exposes individually named, softly supervised coordinates with exact concept-to-horizon effects.

Concept Bottleneck Models (Koh et al., 2020; Vandenhirtz et al., 2024) place semantic variables on the predictive path. The forecasting framework of van Sprang et al. (2024) aligns Transformer representations with time features and an autoregressive surrogate. CODA couples five analytic temporal targets to an affine decoder: both alignment and downstream use can be inspected, addressing the distinction between meaningful intermediate variables and a transparent prediction rule (Rudin et al., 2022). Physics-informed networks (Raissi et al., 2019) and dynamics-guided forecasting (Wu et al., 2024; Lei et al., 2025; Ren et al., 2024) incorporate domain structure, whereas CODA uses a shared local regularization template. Appendix A gives the full Literature Review, including CNNs/RNNs, interpretable basis methods, physics-guided hybrids, and foundation models.

3 Concept-State Forecasting

3.1 Past-only encoding and semantic targets

At forecast origin t, let $x_t = y_{t-L:t-1}$ and predict $(y_t, \dots, y_{t+H-1})$. One model is trained per numeric channel, with batch shapes $x \in \mathbb{R}^{B\times L\times 1}$, $z \in \mathbb{R}^{B\times d}$, $c \in \mathbb{R}^{B\times 5}$, and $\hat{y} \in \mathbb{R}^{B\times H}$. Multivariate benchmarks are evaluated by aggregating these channel-wise forecasts. Scalar embeddings and sinusoidal positions feed a masked Transformer. Each pre-norm block contains attention and a feed-forward sublayer:

$$\widetilde{U} = U + \mathrm{MHA}(\mathrm{LN}(U)), \qquad U' = \widetilde{U} + \mathrm{FFN}(\mathrm{LN}(\widetilde{U})).$$

Attention uses per-head scale $\sqrt{d_h}$ and mask $M_{ij} = -\infty$ for $j > i$. The final token gives $z_t = E_\theta(x_t)$; a two-layer MLP gives $c_t = g_\phi(z_t)$. All tokens belong to observed history, and no future target enters the encoder.

For the final $\tau \le L$ observations, the five semantic targets are

$$\begin{aligned}
c^\star_{1,t} &= \frac{1}{\tau}\sum_{s=t-\tau}^{t-1} y_s && \text{level,}\\
c^\star_{2,t} &= y_{t-1} - y_{t-2} && \text{growth,}\\
c^\star_{3,t} &= y_{t-1}c^\star_{2,t} && \text{signed level–change interaction,}\\
c^\star_{4,t} &= 2\sqrt{a_1^2 + b_1^2} && \text{first-harmonic amplitude,}\\
c^\star_{5,t} &= \sqrt{\frac{1}{\tau}\sum_{s=t-\tau}^{t-1}(y_s - c^\star_{1,t})^2} && \text{local volatility,}
\end{aligned} \tag{1}$$

where $(a_1, b_1) = \mathrm{DFT}_1(y_{t-\tau:t-1} - c^\star_{1,t})$ are the cosine and sine components of the first Fourier coefficient. This amplitude describes the first window harmonic; it need not be the dominant frequency or encode phase. Growth is the latest raw difference, whereas level is a moving average. The analytic targets supervise the learned state through $\mathcal{L}_{\mathrm{con}} = \mathbb{E}_t\|c_t - c^\star_t\|_2^2$. They are soft labels, allowing deviations that improve forecasting while retaining measurable semantic alignment.

3.2 A DIRECT TRANSITION WITH EXACT CONCEPT EFFECTS

Let $a_t = y_{t-1}$ be the known anchor. For $h = 1, \dots, H$, the head predicts

$$\begin{aligned}\hat{y}_{t+h-1} &= a_t + \beta_{0,h} + \sum_{k=1}^{5} \beta_{h,k} c_{k,t} - \gamma_h (a_t - c_{1,t}) \\ &= b_h(a_t) + w_h^\top c_t, \end{aligned} \tag{2}$$

with $b_h(a) = (1 - \gamma_h)a + \beta_{0,h}$, $w_{h,1} = \beta_{h,1} + \gamma_h$, and $w_{h,k} = \beta_{h,k}$ for $k > 1$. The relaxation coefficient is $\gamma_h = \mathrm{softplus}(\widetilde{\gamma}_h) + 10^{-6}$. The affine drive and relaxation motivate an Euler-style transition, but the model generates all horizons directly from one concept vector. It neither integrates an ODE nor recomputes concepts from predicted observations.

The head exposes additive contributions $b_h(a_t), w_{h,1}c_{1,t}, \dots, w_{h,5}c_{5,t}$. For a model intervention that changes concepts by δ while fixing the anchor,

$$\hat{y}_h(c_t + \delta) - \hat{y}_h(c_t) = w_h^\top \delta, \qquad \frac{\partial \hat{y}_h}{\partial c_{k,t}} = w_{h,k}. \tag{3}$$

This equality gives the complete sign, magnitude, and horizon profile of an edit. The anchor also has an explicit coefficient $1 - \gamma_h$, so its contribution is accounted for. These are effects within the fitted model; their interpretation as changes to a real system would require additional causal assumptions.

3.3 COMPATIBILITY REGULARIZATION

The default regularizer couples the current learned state to the previous window's analytic reference. Write $D_k = c_{k,t} - c^\star_{k,t-1}$ and $D_{5^2} = c^2_{5,t} - (c^\star_{5,t-1})^2$. Its four terms are

$$\begin{aligned} R_{1,t} &= D_1 - c_{2,t}, & R_{2,t} &= D_2 - \frac{c_{3,t}}{a_t + \varepsilon}, \\ R_{3,t} &= c_{3,t} - a_t c_{2,t}, & R_{4,t} &= D_{5^2} - 2(a_t - c_{1,t})c_{2,t}, \end{aligned} \tag{4}$$

where $\varepsilon = 10^{-6}$. R_3 checks the interaction definition. R_1, R_2, and R_4 are soft local priors on level change, proportional growth, and variance change, respectively. They are not identities of the sliding-window targets. In particular, adding ε does not bound R_2 near zero; Appendix C gives a multiplicative alternative and exact window recurrences, with controlled results. Appendix I derives the exact recurrences and the interactions among residual terms. The first-harmonic concept is supervised through alignment and has no separate dynamical residual. A residual of the head's defining identity is zero and is excluded without changing the other weights.

With $\mathcal{L}_{\mathrm{data}} = \mathbb{E}_{t,h}(\hat{y}_{t+h-1} - y_{t+h-1})^2$ and $\mathcal{L}_{\mathrm{cmp}} = \mathbb{E}_t \sum_{j=1}^{4} \lambda_j R^2_{j,t}$, training minimizes

$$\mathcal{L}^{(e)} = \mathcal{L}_{\mathrm{data}} + \lambda_{\mathrm{con}}\mathcal{L}_{\mathrm{con}} + \lambda^{(e)}_{\mathrm{cmp}}\mathcal{L}_{\mathrm{cmp}} + \lambda_{\mathrm{reg}}\|\Theta\|_2^2, \quad \lambda^{(e)}_{\mathrm{cmp}} = \lambda_{\max}(1 - e^{-\rho e}). \tag{5}$$

The bounded warm-up increases regularization as the representation is learned. The default and alternative residual formulations are evaluated separately, keeping their associated numerical results distinct.

3.4 FROM SEMANTIC ERROR TO FORECAST ERROR

Proposition 1 (Concept-faithful error decomposition)**.** *For any reference forecaster $f^\star_h(x)$, define its error under the analytic concept interface as $\eta_h(x) = |b_h(a(x)) + w_h^\top c^\star(x) - f^\star_h(x)|$. Then the head in Eq.* (2) *satisfies*

$$|\hat{f}_h(x) - f^\star_h(x)| \le \eta_h(x) + \sum_{k=1}^{5} |w_{h,k}|\,|c_k(x) - c^\star_k(x)|. \tag{6}$$

If $\eta_h(x) \le \epsilon_h$ and $\|c(x) - c^\star(x)\|_2 \le \delta$, the bound is $\epsilon_h + \|w_h\|_2\delta$.

The proof is in Appendix F. The first term measures the adequacy of the specified concepts and affine head; the second weights semantic errors by their actual predictive influence. High alignment alone cannot ensure low forecast error, and equal alignment errors can matter differently at different horizons. This decomposition motivates evaluating semantic alignment, coefficient use, and forecast accuracy together.

3.5 WHICH CONCEPT DIRECTIONS ARE VISIBLE IN A FORECAST?

The affine head reveals individual coefficients, but two concept changes can still cancel across the entire forecast. Let $W \in \mathbb{R}^{H\times 5}$ have rows $w_h^\top$, and write the horizon profile as $F_a(c) = b(a) + Wc$. The following result characterizes precisely which parts of the semantic state the prediction can distinguish.

Proposition 2 (Horizon observability and semantic projection)**.** *Fix a fitted head, an observed anchor a, a learned state c, and its analytic target $c^\star$. Let $W^\dagger$ be the Moore–Penrose pseudoinverse and $P = W^\dagger W$. Then*

$$F_a(u) = F_a(c) \quad \Longleftrightarrow \quad u - c \in \ker W. \tag{7}$$

Among all states that preserve the full forecast, the unique closest state to the analytic target is

$$c^\circ = \arg\min_{u:\, F_a(u)=F_a(c)} \|u - c^\star\|_2^2 = c^\star + P(c - c^\star). \tag{8}$$

If W has column rank five, the forecast distinguishes every concept direction and, for any $u, v \in \mathbb{R}^5$,

$$\sigma_{\min}(W)\|u - v\|_2 \le \|F_a(u) - F_a(v)\|_2 \le \sigma_{\max}(W)\|u - v\|_2. \tag{9}$$

Appendix G gives the proof and its computational form. The projection separates prediction-relevant semantic error, $P(c - c^\star)$, from directions that leave every horizon unchanged. Full column rank eliminates this ambiguity; a small minimum singular value still makes some directions weakly visible. This is a property of the fitted coefficient matrix, which can be inspected directly. It complements target alignment: meaningful names and nonzero individual coefficients alone do not establish that five concepts play distinct predictive roles. The result keeps the anchor fixed and describes model-level distinguishability, while Proposition 1 connects semantic error to predictive error.

4 EXPERIMENTS

4.1 PROTOCOL AND FORECASTING ACCURACY

We evaluate Electricity, Traffic, Weather, Illness, Exchange, and ETT. CODA trains each numeric channel independently using chronological 70/10/20 splits, scalar standardization fitted on the training prefix, and validation-MSE early stopping. Results average over forecast origins and horizons, then channels and seeds 0–9. The default fixes $L = 120$, $\tau = 50$, $d = 64$, four attention heads, two encoder layers, $\lambda_{\text{con}} = 0.7$, $\lambda_{\max} = 1$, and $\rho = 0.05$. AdamW uses learning rate 10^{-3}, weight decay 10^{-4}, batch size 64, at most 50 epochs, patience 10, and gradient clipping 1.0. Appendix B gives data and implementation details; Appendix D reports full numerical comparisons.

Primary horizons. CODA achieves the lowest MSE on five of six datasets and the lowest MAE on five of six, giving 10 of 12 best results in Table 1. Against the primary ModernTCN evaluation, MSE decreases by 4.9% on Traffic, 6.6% on Weather, 1.7% on Illness, 42.7% on Exchange, and 10.5% on ETT. Its Electricity MSE is higher (0.138 versus 0.132), although its MAE is lowest. Against N-HiTS, MSE is lower on all six datasets; the six-dataset mean decreases from 0.287 to 0.268. Thus the semantic interface combines leading accuracy with a transparent decoder across diverse forecasting tasks.

Multiple horizons. The common-horizon evaluation is a separate run set, including a new validation-selected ModernTCN run; its $H = 96$ values therefore differ from Table 1. Across $H = 96, 192, 336, 720$, CODA has the lowest five-dataset mean MSE (Table 2). Including Illness at $H = 24, 36, 48, 60$, it wins 19 of 24 dataset–horizon comparisons and ties once; ModernTCN is stronger on Electricity at all four horizons. Table 3 and Appendix D.2 give the per-dataset results, making the aggregate advantage and its domain dependence visible. Across all 24 settings, mean ranks are 1.19 for CODA, 2.00 for ModernTCN, and 2.81 for N-HiTS (Appendix K).

4.2 WHAT MAKES THE INTERFACE EFFECTIVE?

Removing any one concept increases mean MSE (Table 4). Level has the largest effect; the signed interaction is the least costly to remove. The roles also vary across domains: dropping first-harmonic amplitude raises Electricity MSE from 0.138 to 0.159, while dropping volatility raises Exchange MSE

Table 1: Primary-horizon MSE/MAE ($H = 96$; Illness $H = 24$). Bold marks the minimum in each metric–dataset comparison. Values summarize the primary evaluation; baseline provenance is detailed in Appendix B.

Method	Electricity	Traffic	Weather	Illness	Exchange	ETT
DLinear	.141/.236	.410/.283	.172/.221	1.032/.812	.123/.249	.151/.246
TimesNet	.136/.226	.352/.246	.151/.199	.918/.776	.106/.238	.143/.242
PatchTST	**.129**/.222	.360/.249	.149/.198	.952/.793	.146/.276	.166/.256
iTransformer	.134/.221	.338/.239	.139/.183	.902/.765	.094/.218	.132/.231
TimeMixer	**.129**/.224	.360/.249	.147/.197	.877/.763	.117/.258	.164/.254
ModernTCN	.132/.218	.329/.235	.136/.179	.884/.754	.089/.211	.124/**.224**
N-BEATS	.151/.231	.356/.252	.148/.174	.981/.788	.079/.201	.121/.262
N-HiTS	.145/.224	.338/.239	.139/.164	.917/.759	.071/.190	.112/.248
CODA	.138/**.214**	**.313/.226**	**.127/.153**	**.869/.740**	**.051/.168**	**.111**/.236
CODA seed SD	.003/.002	.008/.005	.003/.003	.030/.024	.0018/.0035	.003/.004

Table 2: Mean MSE across Electricity, Traffic, Weather, Exchange, and ETT in the common-horizon evaluation (ten-seed means). Illness is reported separately because its horizons differ.

Method	$H = 96$	$H = 192$	$H = 336$	$H = 720$
N-HiTS	.161	.196	.229	.281
ModernTCN	.155	.189	.224	.276
CODA	**.148**	**.182**	**.216**	**.267**

Table 3: Per-dataset common-horizon MSE (ten-seed means). These are separate from the primary evaluation in Table 1.

Model	Dataset	96	192	336	720
N-HiTS	Electricity	.145	.162	.181	.213
	Traffic	.338	.394	.447	.529
	Weather	.139	.172	.202	.249
	Exchange	.071	.104	.132	.184
	ETT	.112	.148	.183	.230
ModernTCN	Electricity	.133	.149	.169	.201
	Traffic	.323	.378	.432	.515
	Weather	.131	.163	.194	.242
	Exchange	.077	.101	.132	.183
	ETT	.111	.154	.193	.239
CODA	Electricity	.138	.152	.171	.202
	Traffic	.313	.371	.424	.502
	Weather	.127	.159	.190	.235
	Exchange	.051	.087	.120	.177
	ETT	.111	.141	.175	.219

from 0.051 to 0.062. An acceleration coordinate, $c^{\star}_{6,t} = y_{t-1} - 2y_{t-2} + y_{t-3}$, improves Exchange to 0.048 but raises mean MSE to 0.290. Thus the five-concept choice is a useful cross-domain default, with explicit trade-offs under extension; it is not a sufficiency assumption.

Component ablations separate these effects from merely imposing a small bottleneck. Removing compatibility regularization raises mean MSE by 33.6%; removing semantic alignment raises it by 176.4%; replacing the shared encoder with five independent concept heads raises it by 127.8%. An analytic-statistic-plus-affine control obtains mean MSE 0.331, versus 0.268 for learned soft concepts, a 19.0% reduction relative to directly using the target statistics as features. Direct loss-weight sweeps have a broad stable region: $\lambda_{\mathrm{cmp,max}} = 0.25, 0.5, 1, 2$ gives mean MSE 0.279, 0.271, 0.268, 0.276; $\lambda_{\mathrm{con}} = 0.25, 0.5, 0.7, 1.2$ gives 0.297, 0.272, 0.268, 0.274. Definition-consistent residual alternatives remain within 0.002 mean MSE of the default (Table 5). The interface remains effective across these distinct regularization choices.

Table 4: Primary-horizon mean MSE under concept removals and structural controls. Appendix D.3 gives the available per-dataset results.

Concept interface	MSE	Component configuration	MSE
All five	**.268**	Full CODA	**.268**
Without level	.315	No compatibility loss	.358
Without growth	.297	No semantic alignment	.741
Without interaction	.280	Five independent heads	.611
Without first harmonic	.293	Analytic targets + affine	.331
Without volatility	.289		
Five + acceleration	.290		

What soft supervision changes. The analytic-statistic control and alignment ablation address different alternatives to the proposed interface. Writing $r_t = c_t - c_t^\star$, the fitted head decomposes as

$$\hat{y}_{t+h-1} = b_h(a_t) + w_h^\top c_t^\star + w_h^\top r_t.$$

The first two terms predict from the specified statistics; the last permits task-dependent departures, penalized through $\|r_t\|_2^2$. Fixing $r_t = 0$ removes this adaptive channel, whereas removing alignment eliminates the direct penalty on current target mismatch. The two controls perform worse than full CODA, supporting the combination of learned adaptation and semantic supervision. This does not make the five targets statistically independent: for example, the signed interaction is a product of the last observation and growth. Such a product can still expand an affine head's predictive structure. Appendix J develops this distinction and the role of target supervision in constraining arbitrary coordinate rotations.

4.3 Interpreting forecasts through the concept interface

Semantic alignment and predictive influence. Concept-target R^2 spans 0.62–0.94 across the six datasets (Table 6). Level is consistently well aligned (0.88–0.94). The first harmonic aligns better on Electricity (0.90) than Exchange (0.62), matching its stronger standardized coefficient on Electricity (0.63 versus 0.09). Exchange instead assigns greater weight to volatility (0.57 versus 0.19 on Electricity). Alignment establishes which semantics the encoder captures; coefficients establish how strongly the head uses them. Appendix E gives all coefficient summaries and example concept trajectories.

Concept–increment correlations. Figure 2 compares concept associations with observed and predicted increments. In the `HUFL` panel, growth is positively associated and the signed interaction negatively associated with both increment series. In `ILI`, the level and first-harmonic associations instead change sign between the two rows. These patterns expose agreement and mismatch in the model's dependence structure beyond a forecast overlay.

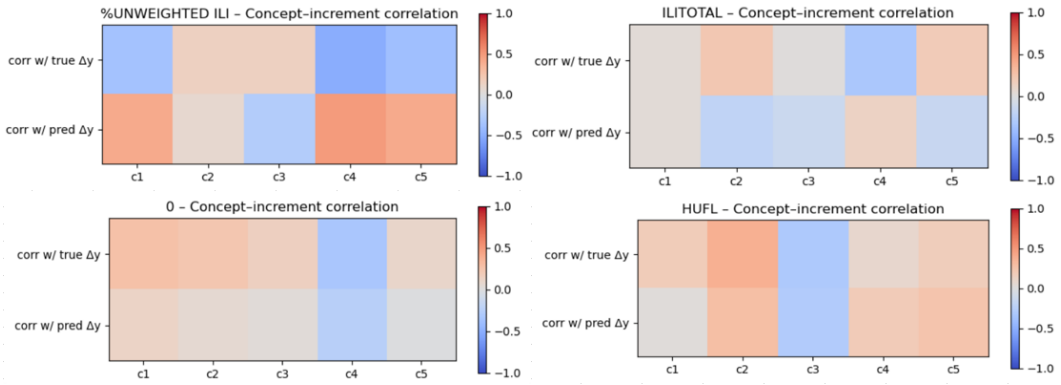


Figure 2: Concept–increment Pearson correlations for four representative channels, retaining their native identifiers. The upper row of each panel uses observed increments and the lower row predicted increments. Columns c_1–c_5 denote level, growth, signed interaction, first-harmonic amplitude, and volatility. Color denotes correlation, with a common range of $[-1, 1]$.

Table 5: Residual alternatives at the primary horizons. Mean MSE is recomputed from the per-dataset results in Appendix C.

Compatibility formulation	Mean MSE	Mean semantic R^2
Default	.268	.82
Exact-window mean	.270	.83
Multiplicative growth prior	.269	.82
Exact sliding variance	.267	.82
Remove head identity	.268	.82
All three alternatives	.269	.83

Table 6: Concept-target R^2 across datasets.

Dataset	Level	Growth	Interaction	First harmonic	Volatility
Electricity	.94	.83	.81	.90	.88
Traffic	.92	.79	.78	.87	.84
Weather	.91	.74	.72	.86	.81
ETT	.93	.77	.75	.88	.83
Illness	.89	.82	.80	.69	.76
Exchange	.88	.78	.77	.62	.90

Marginal correlations summarize co-variation across observations, whereas Eq. (3) isolates a concept's effect with the other coordinates fixed. For an anchored forecast increment $\hat{y}_{t+h-1} - a_t$, its covariance with c_k is

$$\mathrm{Cov}(c_k, \hat{y}_h - a) = \sum_{j=1}^{5} w_{h,j} \,\mathrm{Cov}(c_k, c_j) - \gamma_h \,\mathrm{Cov}(c_k, a). \tag{10}$$

Thus a weak or reversed marginal correlation can arise from correlated concepts and the observed anchor even though the decoder is affine. Figure 2 therefore complements coefficient inspection and semantic alignment. Appendix H derives this relationship, covers adjacent-horizon increments, and explains the channel-level patterns without treating correlation as a causal effect.

Table 7: Electricity coefficient analysis. $\bar{w}_k$ is the standardized dataset-level coefficient summary. The final column is its analytically implied response to a $+0.5$-SD concept edit, calculated using Eq. (3).

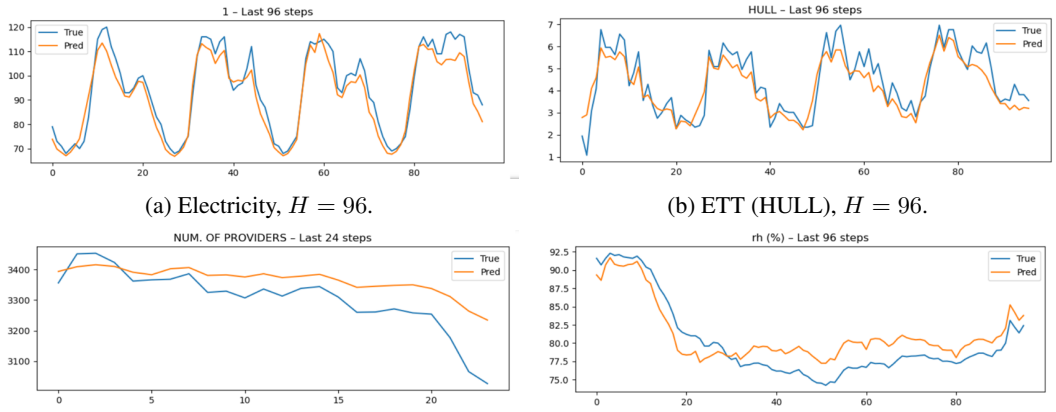
Concept	Target R^2	$\bar{w}_k$	Implied output change (SD)
Level	.94	$.91 + .52 = 1.43$	+.715
Growth	.83	.28	+.140
Interaction	.81	.07	+.035
First-harmonic amplitude	.90	.63	+.315
Volatility	.88	.19	+.095

Electricity walkthrough. Table 7 turns the interface into numerical reasoning about a forecast. Holding the observed anchor and other coordinates fixed, a half-standard-deviation increase in level implies a summary response of $0.5(0.91 + 0.52) = 0.715$ output SD. The corresponding first-harmonic edit gives $0.5(0.63) = 0.315$ output SD. Level and oscillatory strength therefore have distinct, inspectable contributions even when both can raise a predicted peak. For any individual forecast, the actual h-specific $w_{h,k}$ in Eq. (3) determines the response; dataset summaries do not recover that full path. An input perturbation can change several concepts and the anchor together, so a single-coordinate edit answers a model-sensitivity question rather than specifying an achievable intervention on electricity demand.

Diagnosing a semantic mismatch. The interface separates semantic error from predictive influence. For a window, $e_{k,t} = c_{k,t} - c^{\star}_{k,t}$ contributes exactly $w_{h,k} e_{k,t}$ to the difference from the analytic-reference forecast. These signed terms identify which coordinates drive that difference; the reference forecast itself can still be inaccurate. Residuals complement this decomposition by measuring agreement with the selected local regularity preferences.

The dataset-level evidence makes this distinction concrete. Electricity combines first-harmonic alignment of 0.90, a standardized coefficient of 0.63, and a 15.2% MSE increase when that concept is removed. On Exchange, the corresponding alignment and coefficient are 0.62 and 0.09, and removing it raises MSE by 3.9%; removing volatility instead raises MSE by 21.6%. Alignment, use, and removal cost therefore reveal distinct domain roles for the same named coordinates. The removal experiments retrain the model, complementing the fixed-model sensitivity in Eq. (3).

Forecast behavior. Figure 3 shows forecasts in the observed units. Electricity tracks the daily peaks while missing some extremes, consistent with the aggregate MAE/MSE trade-off. ETT follows recurring structure, with local under- and overestimation. The Illness provider-count example has a positive bias late in the window, while the Weather humidity forecast contracts the trough. These cases motivate inspecting the learned state and its coefficient contributions.



(a) Electricity, $H = 96$.

(b) ETT (HULL), $H = 96$.

(c) Illness (providers), $H = 24$.

(d) Weather (humidity), $H = 96$.

Figure 3: Forecast examples from four domains. Blue denotes observations and orange denotes CODA predictions. Quantitative comparisons use the full test evaluation rather than these individual windows.

Stability under input noise. Adding independent Gaussian noise with standard deviation $0.3\hat{\sigma}_y$ to test inputs increases CODA MSE by factors of 1.05–1.17 across datasets, with mean 1.12. FEDformer and LogTrans have mean ratios 1.47 and 1.51 under the same perturbation protocol (Appendix E.1). This controlled perturbation test complements the semantic and coefficient analyses.

5 Discussion and Conclusion

CODA makes the final prediction rule explicit while retaining a nonlinear history encoder. Its strongest empirical result is the combination of leading errors within the evaluated model set, stable semantic alignment, and coefficient-level inspection. The five targets constrain what can be explained: the first harmonic omits phase and higher harmonics, independent channel models omit cross-variable mechanisms, and the affine head has no cross-horizon trajectory constraint. Soft alignment also permits information beyond each target statistic to enter a concept, making held-out audits essential. Human decision benefit and real-world causal interventions remain separate evaluation questions.

The concept interface provides a practical basis for extending the model: a new past-computable statistic adds one supervised coordinate and one coefficient per horizon. The acceleration experiment shows why added complexity should be judged by both accuracy and semantic value; Appendix J develops the extension and its relation to semantic anchoring. Across the present benchmarks, five concepts give a compact and effective default. The resulting forecasting system exposes what its learned state means, how that state affects predictions, and where its modeling assumptions can be tested.

AI USE STATEMENT

In this work, we used generative AI tools during manuscript preparation for editing and polishing. We have not used generative AI tools for dataset processing in this process, and translation and qualitative analysis are not applicable. Additionally, we used generative AI tools for manuscript organization and figure-label editing. We take responsibility for the final content of this work, including text or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 3 defines the forecasting setup and concept targets. Appendix B gives the evaluation protocol and provenance, Appendix D the disaggregated results, and Appendices F–G the proposition proofs. Appendices H–K develop the correlation, residual, concept-extension, and comparative analyses.

REFERENCES

George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time Series Analysis: Forecasting and Control*. Wiley, 5 edition, 2015.

Cristian Challu, Kin G. Olivares, Boris N. Oreshkin, Federico Garza Ramirez, Max Mergenthaler Canseco, and Artur Dubrawski. NHITS: Neural Hierarchical Interpolation for Time Series Forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 6989–6997, 2023. doi: 10.1609/aaai.v37i6.25854. URL https://ojs.aaai.org/index.php/AAAI/article/view/25854.

Hengyu Fu, Zehao Dou, Jiawei Guo, Mengdi Wang, and Minshuo Chen. Diffusion Transformer Captures Spatial-Temporal Dependencies: A Theory for Gaussian Process Data. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=MbM1BqGpZu.

Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. TimeGPT-1, 2024. URL https://arxiv.org/abs/2310.03589.

Pradeep Hewage, Ardhendu Behera, Marcello Trovati, Ella Pereira, Morteza Ghahremani, Francesco Palmieri, and Yonghuai Liu. Temporal convolutional neural (TCN) network for an effective weather forecasting using time-series data from the local weather station. *Soft Computing*, 24(21): 16453–16482, 2020. doi: 10.1007/s00500-020-04954-0.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The Efficient Transformer. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkgNKkHtvB.

Jacek Klimek, Jakub Klimek, Witold Kraśkiewicz, and Mateusz Topolewski. Query Selector–Efficient transformer with sparse attention. *Software Impacts*, 11:100187, 2022. doi: 10.1016/j.simpa.2021.100187.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept Bottleneck Models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long-and short-term temporal patterns with deep neural networks. In *SIGIR*, 2018.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677): 1416–1421, 2023. doi: 10.1126/science.adi2336. URL https://www.science.org/doi/abs/10.1126/science.adi2336.

Leixiao Lei, Yigang He, Zhikai Xing, Zihao Li, and Yazhong Zhou. Physics-Informed LSTM-Based Time-Series Forecasting Model for Power Transformers. *IEEE Transactions on Industrial Informatics*, 21(7):5411–5419, 2025. doi: 10.1109/tii.2025.3555981.

Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/6775a0635c302542da2c32aa19d86be0-Paper.pdf.

Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021. doi: 10.1016/j.ijforecast.2021.03.012. URL https://www.sciencedirect.com/science/article/pii/S0169207021000637.

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. URL https://openreview.net/forum?id=JePfAI8fah.

Donghao Luo and Xue Wang. ModernTCN: A Modern Pure Convolution Structure for General Time Series Analysis. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vpJMJerXHU.

Lihua Mi, Yan Han, Lizhi Long, Hui Chen, and C.S. Cai. A physics-informed temporal convolutional network-temporal fusion transformer hybrid model for probabilistic wind speed predictions with quantile regression. *Energy*, 326:136302, 2025. doi: 10.1016/j.energy.2025.136302.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1ecqn4YwB.

Yao Qin, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, and Garrison Cottrell. A dual-stage attention-based recurrent neural network for time series prediction. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2017.

M. Raissi, P. Perdikaris, and G.E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045. URL https://www.sciencedirect.com/science/article/pii/S0021999118307125.

Zeqi Ren, Jianbo Yu, Jian Huang, Xiaofeng Yang, Siyang Leng, Yuping Liu, and Shifu Yan. Physically-guided temporal diffusion transformer for long-term time series forecasting. *Knowledge-Based Systems*, 304:112508, 2024. doi: 10.1016/j.knosys.2024.112508. URL https://www.sciencedirect.com/science/article/pii/S0950705124011420.

Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges. *Statistics Surveys*, 16:1–85, 2022. doi: 10.1214/21-SS133.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 2020.

Angela van Sprang, Erman Acar, and Willem Zuidema. Enforcing Interpretability in Time Series Transformers: A Concept Bottleneck Framework, 2024. URL https://arxiv.org/abs/2410.06070v1.

Moritz Vandenhirtz, Sonia Laguna, Ričards Marcinkevičs, and Julia E. Vogt. Stochastic Concept Bottleneck Models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 51787–51810. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/5c7894ac8788555f1cecf536f1e0fd35-Paper-Conference.pdf.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Huiqiang Wang, Jian Peng, Feihu Huang, Jince Wang, Junhui Chen, and Yifei Xiao. MICN: Multi-scale Local and Global Context Modeling for Long-term Series Forecasting. In *International Conference on Learning Representations*, 2023.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7oLshfEIC2.

Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in Time Series: A Survey. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pp. 6778–6786. International Joint Conferences on Artificial Intelligence Organization, 2023. doi: 10.24963/ijcai.2023/759. URL https://doi.org/10.24963/ijcai.2023/759. Survey Track.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22419–22430. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/bcc0d400288793e8bdcd7c19a8ac0c2b-Paper.pdf.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *International Conference on Learning Representations*, 2023.

Sifan Wu, Xi Xiao, Qianggang Ding, Peilin Zhao, Ying Wei, and Junzhou Huang. Adversarial Sparse Transformer for Time Series Forecasting. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17105–17115. Curran Associates, Inc., 2020.

Song Wu, Senliang Bao, Wei Dong, Senzhang Wang, Xiaojiang Zhang, Chengcheng Shao, Junxing Zhu, and Xiaoyong Li. PGTransNet: a physics-guided transformer network for 3D ocean temperature and salinity predicting in tropical Pacific. *Frontiers in Marine Science*, 11, 2024. doi: 10.3389/fmars.2024.1477710.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are Transformers Effective for Time Series Forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37:11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317.

Tianping Zhang, Yizhuo Zhang, Wei Cao, Jiang Bian, Xiaohan Yi, Shun Zheng, and Jian Li. Less is more: Fast multivariate time series forecasting with light sampling-oriented mlp structures. *arXiv preprint arXiv:2207.01186*, 2022.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:11106–11115, 2021. doi: 10.1609/aaai.v35i12.17325.

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*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27268–27286. PMLR, 2022. URL https://proceedings.mlr.press/v162/zhou22g.html.

Zhenyi Zhu, Yuchen Huang, and Liu Liu. PhysicsSolver: Transformer-enhanced Physics-Informed Neural Networks for forward and forecasting problems in partial differential equations. *Journal of Computational and Applied Mathematics*, 473:116900, 2026. doi: 10.1016/j.cam.2025.116900.

A LITERATURE REVIEW

Early time-series forecasting relies on stochastic linear models such as the Box–Jenkins ARIMA (Box et al., 2015), whose identification–estimation–diagnosis cycle remains influential but often fails under the non-stationarity of modern telemetry.

N-BEATS (Oreshkin et al., 2020) has a configuration with explicit trend and seasonality bases, and N-HiTS (Challu et al., 2023) decomposes forecasts across scales through hierarchical interpolation. Although both use MLP components, their structured prediction mechanisms distinguish them from generic opaque decoders.

Deep learning models capture time-series patterns with specifically-designed architectures spanning a wide range of foundation backbones, including CNNs (Wang et al., 2023; Wu et al., 2023; Hewage et al., 2020), RNNs (Lai et al., 2018; Qin et al., 2017; Salinas et al., 2020), Transformers (Vaswani et al., 2017) and MLPs (Zhang et al., 2022; Oreshkin et al., 2020; Challu et al., 2023). The quadratic cost of vanilla attention spurred efficient variants, Reformer (Kitaev et al., 2020) and LogTrans (Li et al., 2019), and domain-specific advances: Informer's ProbSparse attention (Zhou et al., 2021), Autoformer's trend–seasonality decomposition (Wu et al., 2021), FEDformer's Fourier sparsity (Zhou et al., 2022), PatchTST's patch-token strategy (Nie et al., 2023), and TimeMixer (Wang et al., 2024). A 2023 survey (Wen et al., 2023) reviews Transformer methods for time-series analysis.

Despite these advances, interpretability remains a challenge, prompting growing interest in concept-level supervision. Concept Bottleneck Models (Koh et al., 2020) demonstrate that forcing predictions through human-meaningful variables can maintain accuracy while enabling user interventions. Stochastic CBMs (Vandenhirtz et al., 2024) generalized this idea by modelling concept dependencies probabilistically. FF Bottleneck (van Sprang et al., 2024) used an AR model as a surrogate concept within a Transformer architecture; however, empirical results revealed that the standalone AR model outperformed FF Bottleneck on four out of the six datasets.

In parallel, physics-informed learning has emerged as a remedy for physically implausible outputs. Raissi's PINNs framework (Raissi et al., 2019) pioneered the joint optimization of data and governing equations; PGTransNet (Wu et al., 2024) integrated dynamics-guided self-attention for Pacific-Ocean temperature forecasting, while physics-informed LSTM variants have been applied to power-transformer health monitoring (Lei et al., 2025). The PhysicsSolver (Zhu et al., 2026) further attests to the potential of incorporating scientific priors into deep learning architectures.

Hybrid work now marries physics with deep networks: a TCN–TFT ensemble improves probabilistic wind forecasts (Mi et al., 2025), while physically guided diffusion Transformers address long-horizon forecasting (Ren et al., 2024). Interpretability has advanced through Temporal Fusion Transformer's gating and variable-selection (Lim et al., 2021) and through sparse-attention schemes such as Adversarial Sparse Transformer (Wu et al., 2020) and Query-Selector attention (Klimek et al., 2022). Complementing advances in model architecture, evaluation paradigms have matured: GraphCast, a graph neural weather forecaster (Lam et al., 2023), demonstrates that ML systems can rival numerical solvers at global scale. Meanwhile, foundation models like TimeGPT (Garza et al., 2024) deliver strong zero-shot generalization capabilities across diverse time-series domains. Nevertheless, critical evaluations have shown that simple linear baselines can outperform many Transformer variants when evaluated rigorously, as highlighted in the "Are Transformers Effective for Time-Series Forecasting?" paper (Zeng et al., 2023). Additionally, the rise of diffusion-based Transformers (Fu et al., 2025) reflects the field's continued pursuit of the right balance between model complexity and effective inductive biases.

In summary, although current approaches span efficiency improvements, frequency-domain finesse, multiresolution structure and dynamics-guided regularization, these approaches expose different kinds of structure. CODA combines a softly supervised five-concept interface with an explicit affine forecast map and local compatibility regularization, and evaluates this combination on six heterogeneous benchmarks.

B Datasets and Implementation

Table 8: Dataset characteristics and primary horizons. Each numeric channel is modeled independently.

Dataset	H	Description
Electricity	96	Hourly consumption, 321 customers, 2012–2014.
Traffic	96	Hourly San Francisco Bay Area freeway occupancy.
Weather	96	21 meteorological indicators at 10-minute intervals.
Illness	24	Weekly US influenza surveillance, 2002–2021.
Exchange	96	Daily exchange rates for eight countries, 1990–2016.
ETT	96	15-minute transformer load and temperature series.

Training and preprocessing. CODA uses two pre-norm Transformer layers, four heads, width $d = 64$, lookback $L = 120$, and target window $\tau = 50$. The final token is mapped to five coordinates by $g_\phi(z) = W_2\,\mathrm{ReLU}(W_1 z + b_1) + b_2$. The training configuration uses AdamW, learning rate 10^{-3}, weight decay 10^{-4}, batch size 64, maximum 50 epochs, validation-MSE early stopping with patience 10, gradient clipping at 1.0, and no learning-rate scheduler. The concept weight is 0.7 and the compatibility weight approaches 1 with $\rho = 0.05$. The loader includes interpolation and backfill/forward-fill fallback for missing entries; the training prefix alone determines standardization. Forecast windows contain only past observations, and each target sequence spans t through $t + H - 1$.

Aggregation. For seed s, channel j, and its test origins $\mathcal{T}_j$, let

$$M_{s,j} = \frac{1}{|\mathcal{T}_j|H} \sum_{t \in \mathcal{T}_j} \sum_{h=1}^{H} (\hat{y}_{s,j,t+h-1} - y_{j,t+h-1})^2.$$

The reported MSE averages $M_{s,j}$ over channels and ten seeds. MAE replaces the square by absolute value. Seed SD describes variation of the channel-averaged metric across runs; it does not quantify temporal uncertainty from overlapping forecast windows. Reported cross-dataset means average the displayed per-dataset scores without weighting by channel count.

Baseline provenance. The primary ModernTCN results and common-horizon ModernTCN results come from two distinct local evaluations of its official implementation (Luo & Wang, 2024). The latter is a separately validation-selected multi-horizon run set, with ten-seed means (0–9). N-BEATS and N-HiTS comparisons use matched splits, normalization, validation rules, primary horizons, and seeds. DLinear/NLinear were evaluated locally; the remaining broad comparison uses official-repository or released results collected in the evaluation record. These sources do not imply identical hyperparameter search spaces. Table 10 preserves the available primary-evaluation standard deviations; multi-horizon means are reported without unrecorded uncertainty estimates. Comparisons are descriptive and carry no paired significance claim.

Computational structure. Attention costs $O(L^2 d)$ per encoder layer, in addition to token projections and feed-forward computation. A concept MLP of hidden width d_1 adds $O(dd_1 + 5d_1)$, and the direct head adds $O(5H)$ work. Model interventions can evaluate $W\delta$ without rerunning the Transformer. These operation counts characterize the interface overhead; no wall-clock speedup is inferred from them.

C Residual Definitions and Alternatives

The default terms in Eq. (4) use current learned concepts and previous-window analytic targets. This distinction matters: D_1 is not a difference of two learned states, and the raw growth target is not the finite difference of the moving mean. R_3 has a definitional target relation; R_1, R_2, and R_4 express local regularization preferences. Their magnitudes audit agreement with those preferences, not validity of a physical law.

Let $m_t^\star = c_{1,t}^\star$ and $q_t^\star = \tau^{-1} \sum_{s=t-\tau}^{t-1} y_s^2$. The moving-window statistics satisfy

$$\begin{aligned} m_t^\star &= m_{t-1}^\star + (y_{t-1} - y_{t-\tau-1})/\tau, \\ q_t^\star &= q_{t-1}^\star + (y_{t-1}^2 - y_{t-\tau-1}^2)/\tau, \qquad (c_{5,t}^\star)^2 = q_t^\star - (m_t^\star)^2. \end{aligned} \tag{11}$$

A definition-consistent mean residual compares $c_{1,t}$ with the first right-hand side. A variance residual compares $c_{5,t}^2$ with the second-moment update minus the square of the updated mean. Both include the outgoing observation and preserve the same previous analytic reference. The multiplicative alternative $R'_{2,t} = a_t D_2 - c_{3,t}$ removes division by a possibly small $a_t + \varepsilon$; it remains a proportional-growth prior rather than an analytic identity. These alternatives are compared individually and jointly in Table 9.

Table 9: Residual-formulation controls: per-dataset MSE and its arithmetic mean. The last row replaces the mean, proportional-growth, and variance priors together.

Formulation	Elec.	Traffic	Weather	Illness	Exchange	ETT	Mean
Default	.138	.313	.127	.869	.051	.111	.268
Exact-window mean	.139	.316	.128	.874	.052	.112	.270
Multiplicative growth	.139	.315	.127	.872	.051	.111	.269
Exact sliding variance	.137	.311	.126	.868	.050	.110	.267
Remove head identity	.138	.313	.127	.869	.051	.111	.268
All three alternatives	.139	.314	.127	.871	.051	.111	.269

The head identity is $(\hat{y}_{t+h-1} - a_t) - [\beta_{0,h} + \sum_k \beta_{h,k} c_{k,t} - \gamma_h (a_t - c_{1,t})] = 0$. Removing its squared residual leaves gradients unchanged when the four nontrivial weights are kept fixed. Table 9 evaluates alternative priors; comparable accuracy does not make the original priors exact identities. The first Fourier amplitude has no phase information and no separate temporal residual.

D Disaggregated Forecasting and Ablation Results

D.1 Primary evaluation

Table 10: Full primary evaluation: recorded mean ± seed SD. Primary horizons are 96 except Illness (24). For N-BEATS/N-HiTS, the available means are given in Table 1.

Method	Metric	Electricity	Traffic	Weather	Illness	Exchange rate	ETT
LogTrans	MSE	0.258±0.006	0.684±0.014	0.458±0.010	4.480±0.120	0.968±0.021	0.768±0.016
	MAE	0.357±0.005	0.384±0.007	0.490±0.008	1.444±0.045	0.812±0.013	0.642±0.012
Informer	MSE	0.274±0.007	0.719±0.015	0.300±0.008	5.764±0.160	0.847±0.019	0.365±0.010
	MAE	0.368±0.006	0.391±0.008	0.384±0.007	1.677±0.052	0.752±0.012	0.453±0.011
Autoformer	MSE	0.201±0.005	0.613±0.013	0.266±0.007	3.483±0.110	0.197±0.006	0.255±0.008
	MAE	0.317±0.004	0.388±0.007	0.336±0.006	1.287±0.041	0.323±0.009	0.339±0.009
FEDformer	MSE	0.183±0.004	0.562±0.012	0.217±0.006	2.203±0.085	0.183±0.005	0.203±0.006
	MAE	0.297±0.004	0.349±0.006	0.296±0.005	0.963±0.030	0.297±0.007	0.287±0.007
DLinear	MSE	0.141±0.004	0.410±0.011	0.172±0.005	1.032±0.037	0.123±0.004	0.151±0.004
	MAE	0.236±0.004	0.283±0.007	0.221±0.004	0.812±0.027	0.249±0.005	0.246±0.004
TimesNet	MSE	0.136±0.004	0.352±0.009	0.151±0.004	0.918±0.032	0.106±0.003	0.143±0.004
	MAE	0.226±0.003	0.246±0.006	0.199±0.003	0.776±0.026	0.238±0.004	0.242±0.004
PatchTST	MSE	**0.129**±0.003	0.360±0.009	0.149±0.004	0.952±0.033	0.146±0.004	0.166±0.004
	MAE	0.222±0.003	0.249±0.006	0.198±0.003	0.793±0.028	0.276±0.005	0.256±0.004
iTransformer	MSE	0.134±0.003	0.338±0.008	0.139±0.004	0.902±0.031	0.094±0.003	0.132±0.004
	MAE	0.221±0.003	0.239±0.006	0.183±0.003	0.765±0.025	0.218±0.004	0.231±0.004
TimeMixer	MSE	**0.129**±0.003	0.360±0.009	0.147±0.004	0.877±0.031	0.117±0.003	0.164±0.004
	MAE	0.224±0.003	0.249±0.006	0.197±0.003	0.763±0.026	0.258±0.004	0.254±0.004
ModernTCN	MSE	0.132±0.003	0.329±0.008	0.136±0.004	0.884±0.030	0.089±0.003	0.124±0.003
	MAE	0.218±0.003	0.235±0.005	0.179±0.003	0.754±0.024	0.211±0.004	**0.224**±0.004
CODA	MSE	0.138±0.003	**0.313**±0.008	**0.127**±0.003	**0.869**±0.030	**0.051**±0.0018	**0.111**±0.003
	MAE	**0.214**±0.002	**0.226**±0.005	**0.153**±0.003	**0.740**±0.024	**0.168**±0.0035	0.236±0.004

D.2 MULTIPLE FORECAST HORIZONS

Table 3 gives the five datasets with common horizons; Table 11 gives the four Illness horizons.

Table 11: Illness MSE in the separate multi-horizon evaluation.

Method	$H = 24$	$H = 36$	$H = 48$	$H = 60$
N-HiTS	.917	.941	.966	.992
ModernTCN	.907	.931	.956	.982
CODA	**.869**	**.894**	**.918**	**.943**

D.3 CONCEPT-SET AND COMPONENT ABLATIONS

Table 12: Per-dataset primary-horizon MSE for concept-set sensitivity.

Concept set	Electricity	Traffic	Weather	Illness	Exchange	ETT	Avg.
All five	.138	.313	.127	.869	.051	.111	.268
Without level	.168	.362	.152	1.015	.059	.134	.315
Without growth	.154	.348	.142	.957	.057	.125	.297
Without interaction	.145	.329	.133	.902	.054	.117	.280
Without first-harmonic amplitude	.159	.356	.149	.914	.053	.128	.293
Without volatility	.147	.339	.139	.925	.062	.121	.289
Five + acceleration	.153	.336	.147	.928	.048	.129	.290

Table 13: Component ablations at the primary horizons: per-dataset MSE and arithmetic mean.

Configuration	Elec.	Traffic	Weather	Illness	Exchange	ETT	Mean
CODA	.138	.313	.127	.869	.051	.111	.268
No compatibility loss	.171	.442	.168	1.122	.081	.165	.358
No semantic alignment	.360	.830	.291	2.380	.302	.285	.741
Five independent heads	.286	.705	.223	1.930	.227	.294	.611

The analytic-statistic-plus-affine control has mean MSE 0.331. The component-ablation mean-MSE increases are 33.6%, 176.4%, and 127.8%, calculated from the unrounded arithmetic means of the displayed per-dataset values. Removing concepts changes the interface dimension; it is distinct from retaining five learned coordinates but removing their semantic supervision. The acceleration extension changes only the final concept dimension and the associated per-horizon coefficients.

Table 14: Loss-weight sensitivity: mean primary-horizon MSE.

$\lambda_{\mathrm{cmp,max}}$	0	.25	.5	1.0	2.0
MSE	.358	.279	.271	**.268**	.276
λ_{con}	0	.25	.5	.7	1.2
MSE	.741	.297	.272	**.268**	.274

E SEMANTIC AND ROBUSTNESS ANALYSIS

Table 6 reports the complete concept-target alignment matrix.

For each concept, R^2 measures agreement between its learned trajectory and analytic target, aggregated over time and channels. An accurate forecast and a high target R^2 answer different questions. For example, Exchange has low forecast MSE but a comparatively weak first-harmonic alignment; level and volatility remain strongly aligned. The bottleneck therefore supports coordinate-specific interpretation rather than an all-or-nothing transparency label.

Table 15: Standardized dataset-level transition coefficients. Columns after γ give $\beta_1, \ldots, \beta_5$.

Dataset	γ	Level	Growth	Interaction	First harmonic	Volatility
Electricity	.52	.91	.28	.07	.63	.19
Traffic	.49	.88	.31	.05	.58	.22
Weather	.47	.84	.26	.09	.41	.27
ETT	.54	.89	.29	.11	.45	.24
Illness	.61	.73	.37	.33	.22	.41
Exchange	.58	.69	.42	.18	.09	.57

The coefficients in Table 15 are standardized dataset-level summaries. For level, the effective forecast weight is $\beta_{h,1} + \gamma_h$; for other concepts it is $\beta_{h,k}$. The observed-anchor coefficient is $1 - \gamma_h$. Eq. (3) is exact for the actual horizon-specific coefficients, but a table of aggregate coefficients cannot certify per-window edit directions or their empirical support.

E.1 Input-noise robustness

Table 16: MSE ratio under additive input noise with standard deviation $0.3\hat{\sigma}_y$, relative to clean evaluation.

Model	Elec.	Traffic	Weather	Illness	Exchange	ETT	Mean
CODA	1.12	1.15	1.05	1.17	1.09	1.11	1.12
FEDformer	1.47	1.51	1.39	1.44	1.53	1.46	1.47
LogTrans	1.52	1.58	1.42	1.55	1.49	1.50	1.51

Noise is injected independently into test inputs and the target remains unchanged. CODA shows a modest relative increase across all six datasets. This test isolates sensitivity to this perturbation distribution; it does not evaluate every kind of temporal shift.

E.2 Concept trajectories

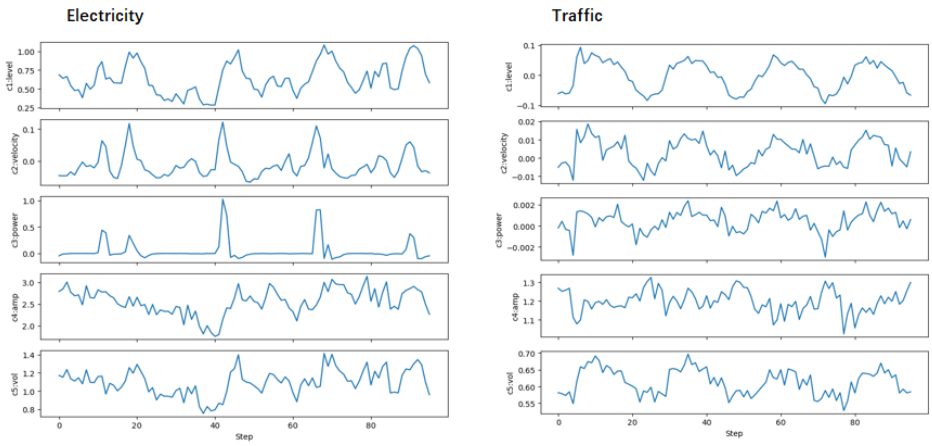


Figure 4: Example concept trajectories for Electricity and Traffic. From top to bottom: level, growth, signed interaction, first-harmonic amplitude, and volatility.

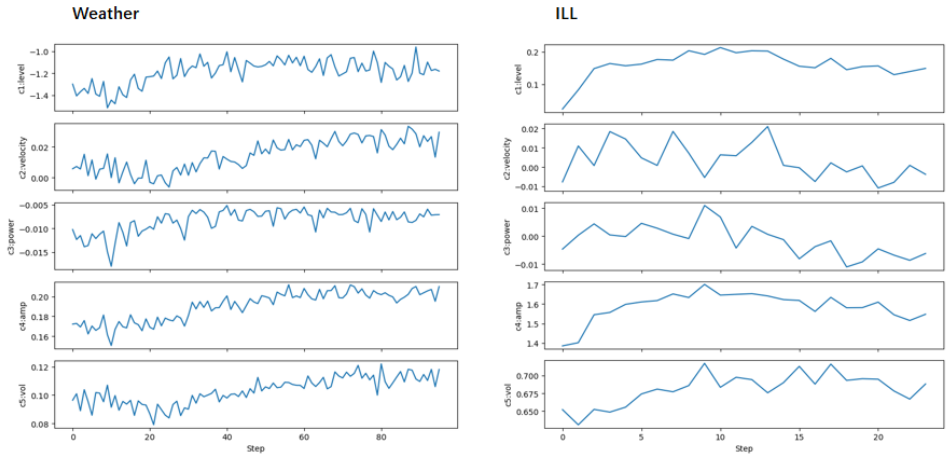


Figure 5: Example concept trajectories for Weather and Illness, in the same coordinate order as Figure 4.

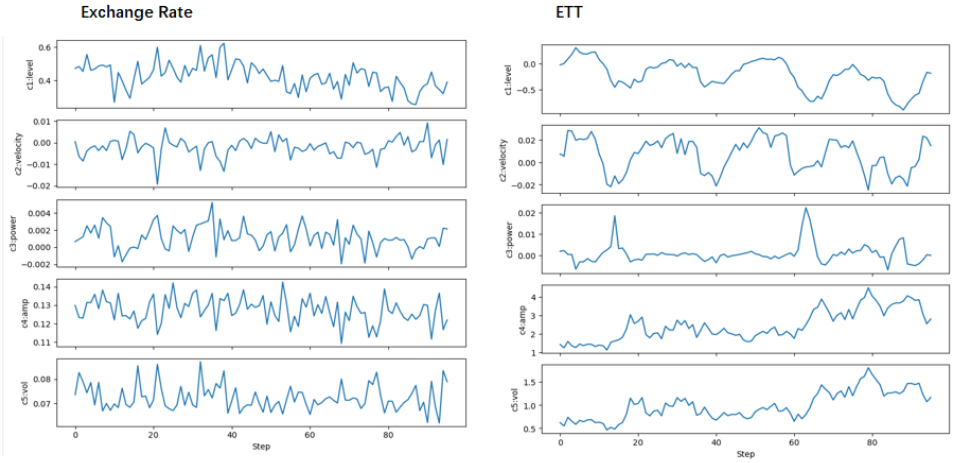


Figure 6: Example concept trajectories for Exchange and ETT, in the same coordinate order as Figure 4.

F Proof and Interpretation of Proposition 1

Fix a history x, its observed anchor $a(x)$, and a forecast horizon h. By Eq. (2),

$$\hat{f}_h(x) - f_h^\star(x) = b_h(a(x)) + w_h^\top c^\star(x) - f_h^\star(x) + w_h^\top (c(x) - c^\star(x)).$$

Taking absolute values and applying the triangle inequality gives

$$|\hat{f}_h(x) - f_h^\star(x)| \le \eta_h(x) + \sum_{k=1}^{5} |w_{h,k}|\, |c_k(x) - c_k^\star(x)|.$$

Cauchy–Schwarz bounds the sum by $\|w_h\|_2 \|c(x) - c^\star(x)\|_2$, yielding the stated conditional bound.

The analytic target map is fixed throughout; no coordinate is redefined to contain the desired forecast. The anchor is explicit on both sides of the decomposition. The term η_h includes both any insufficiency of the chosen statistics and the approximation error of the affine decoder. It can be large even under perfect alignment. Conversely, semantic errors aligned with small coefficients have limited direct effect on a forecast. The proposition is therefore an error accounting identity and bound for the actual architecture, not a universal approximation or semantic-identification theorem.

For a model edit with anchor fixed, the stronger equality $\hat{f}_h(c+\delta) - \hat{f}_h(c) = w_h^\top \delta$ follows directly. Its check requires the same channel, seed, horizon, and standardization on both sides. An independently chosen input perturbation need not realize such a concept edit: it can move several concepts jointly or leave the learned concept support. This distinction separates transparent model behavior from a causal interpretation of the data-generating process.

G Proof and Consequences of Horizon Observability

G.1 Forecast equivalence and the identifiable subspace

All statements concern one fitted channel model and one seed. The anchor is fixed, and the same concept vector produces every horizon. Thus $F_a(u) - F_a(v) = W(u - v)$, which immediately proves Eq. (7). In particular, the set of states with the same forecast as c is the affine set $c + \ker W$. This set concerns equality of the complete H-dimensional forecast, a stronger condition than equality at one horizon.

Let $r = \operatorname{rank}(W) > 0$ and use a reduced singular value decomposition

$$W = U_r \Sigma_r V_r^\top, \qquad W^\dagger = V_r \Sigma_r^{-1} U_r^\top, \qquad P = W^\dagger W = V_r V_r^\top.$$

The columns of U_r and V_r are orthonormal, and the diagonal entries of Σ_r are positive. Therefore $P = P^\top = P^2$ is the orthogonal projection onto the row space of W, while $I - P$ projects onto $\ker W$. Every concept discrepancy $e = c - c^\star$ has the unique orthogonal decomposition

$$e = Pe + (I - P)e, \qquad We = WPe, \qquad \|e\|_2^2 = \|Pe\|_2^2 + \|(I - P)e\|_2^2. \tag{12}$$

The first component changes at least one horizon unless it is zero. The second is invisible to the entire prediction profile. This distinguishes the number of named concept coordinates from the number of linearly distinguishable directions through which the fitted head uses them.

G.2 Proof of the nearest semantic state

Write $e = c - c^\star$. A state u preserves the forecast precisely when $W(u - c^\star) = We$, or equivalently $P(u - c^\star) = Pe$. Consequently every feasible state can be written uniquely as

$$u = c^\star + Pe + z, \qquad z \in \ker W.$$

Orthogonality gives

$$\|u - c^\star\|_2^2 = \|Pe\|_2^2 + \|z\|_2^2.$$

The unique minimum occurs at $z = 0$, proving Eq. (8). Uniqueness follows from strict convexity of the squared Euclidean distance on an affine feasible set. The resulting semantic discrepancy is $c^\circ - c^\star = Pe$, and the change from the learned state is $c^\circ - c = -(I - P)e$. Hence

$$F_a(c^\circ) = F_a(c), \qquad \|c^\circ - c^\star\|_2 \le \|c - c^\star\|_2, \qquad \|c^\circ - c\|_2 = \|(I - P)e\|_2. \tag{13}$$

If $W = 0$, the entire concept space is invisible, $P = 0$, and the same formulas reduce to $c^\circ = c^\star$. Thus the projection claim also covers the rank-zero case.

The projection is an algebraic reference for inspecting a fitted model. It does not assume that c° is the encoder output of a realizable history. If feasible concept states were additionally restricted to a set $\mathcal{C}$, the relevant optimization would intersect the affine constraint with $\mathcal{C}$; Eq. (8) would solve that constrained problem only when its solution belongs to $\mathcal{C}$. The reported forecasting results use the learned encoder state, with no projection inserted at inference.

G.3 Proof of the stability bounds

For any $d \in \mathbb{R}^5$, orthonormality yields

$$\|Wd\|_2^2 = \|\Sigma_r V_r^\top d\|_2^2 = \sum_{i=1}^{r} \sigma_i^2 (v_i^\top d)^2.$$

Consequently, even in the rank-deficient case,

$$\sigma_r \|Pd\|_2 \le \|Wd\|_2 \le \sigma_1 \|Pd\|_2. \tag{14}$$

When $r = 5$, $P = I$ and $\sigma_5 > 0$. Substituting $d = u - v$ proves Eq. (9). In this case the forecast map is injective in the concept state at fixed anchor, and its inverse on its image is Lipschitz with constant $1/\sigma_5$. The projection reduces to $c^\circ = c$: any nonzero correction of a semantic discrepancy must change at least one forecast horizon.

Rank alone does not quantify practical visibility. If σ_5 is positive but small, a unit change along its right singular vector produces a forecast change of only σ_5. Conversely, a direction associated with σ_1 is strongly amplified. The condition number σ_1/σ_5 describes the imbalance between these directions. These statements hold for a fixed fitted matrix; the dataset-level coefficient averages in Table 15 do not provide its singular values.

G.4 Cancellation, anchors, and coordinate units

As an algebraic illustration, suppose the first two columns of W coincide. Then $d = (1, -1, 0, 0, 0)^\top$ lies in $\ker W$, even if both columns have large entries. Increasing level and decreasing growth by the same numerical amount leaves every horizon unchanged in these coordinates. The example shows why inspecting each column separately cannot establish independent concept use. It is not an assertion about the rank of the trained models.

The fixed-anchor condition is equally essential. With $\alpha_h = 1 - \gamma_h$, two histories with different anchors and concepts satisfy

$$\widehat{\boldsymbol{y}}' - \widehat{\boldsymbol{y}} = \alpha\, \Delta a + W \Delta c.$$

An anchor change may compensate for a concept change. Joint distinguishability is governed by the augmented matrix $[\alpha\ W]$, whereas Proposition 2 isolates the learned concept interface. The explicitly observed anchor is therefore included in the explanatory accounting even though it is not a learned concept.

Singular values also depend on units. For $\widetilde{c} = S^{-1}(c - \mu)$ with positive diagonal S, the same forecast uses $\widetilde{W} = WS$ and an adjusted intercept. Rank is unchanged, but singular values and Euclidean distances generally change. Comparisons of conditioning require a common convention, such as concept scales estimated on the training split. The proposition holds in either coordinate system when the matrix and distances are transformed together. Computing a thin SVD costs $O(HK^2)$ for $K = 5$; it is performed on the head, without repeated calls to the Transformer.

H Correlation, Coefficients, and Forecast Increments

H.1 What the heatmap measures

Figure 2 provides a descriptive view of concept–increment association. Its native labels identify two Illness channels (`%UNWEIGHTED ILI` and `ILITOTAL`), a numerically named channel `0`, and the

transformer-load channel HUFL. For a concept C_k and an increment variable Z, Pearson correlation is

$$\mathrm{Corr}(C_k, Z) = \frac{\mathrm{Cov}(C_k, Z)}{\sqrt{\mathrm{Var}(C_k)\,\mathrm{Var}(Z)}},$$

provided both variances are positive. It summarizes centered linear association over the observations included in a panel. The figure conveys signs and relative strengths; it does not supply a numerical correlation table, uncertainty intervals, or a hypothesis test.

The two rows answer different questions. The upper row describes how the concept co-varies with observed change, whereas the lower row describes how it co-varies with the model's predicted change. Similar patterns suggest that the prediction preserves the displayed association. A mismatch identifies a relationship to examine using the explicit readout and the semantic targets. It does not, by itself, determine which relationship causes forecast error, since several concepts and the anchor vary together.

H.2 Covariance decomposition of an affine forecast

Fix a horizon and channel. Treat the state and anchor as random variables over forecast origins, while keeping fitted coefficients fixed. Assume finite second moments for population covariances; the same algebra holds for sample covariances. Let

$$\widehat{Z}_h = \hat{y}_h - a = \beta_{0,h} + w_h^\top C - \gamma_h A.$$

Covariance is linear in its second argument and vanishes against constants, giving

$$\mathrm{Cov}(C_k, \widehat{Z}_h) = w_{h,k}\,\mathrm{Var}(C_k) + \sum_{j \neq k} w_{h,j}\,\mathrm{Cov}(C_k, C_j) - \gamma_h\,\mathrm{Cov}(C_k, A).$$

This proves Eq. (10). With $\Sigma_C = \mathrm{Cov}(C, C)$ and $s_{CA} = \mathrm{Cov}(C, A)$, the vector of concept–increment covariances is

$$s_{C\widehat{Z}_h} = \Sigma_C w_h - \gamma_h s_{CA}. \tag{15}$$

The direct coefficient is only one contribution to a marginal association. Positive co-variation with another concept carrying a negative coefficient can offset a positive direct effect; the anchor term can also offset it. Cancellation can yield nearly zero marginal correlation while $w_{h,k}$ remains nonzero. No nonlinear decoder is needed for this behavior. The nonlinear component of CODA lies in the history encoder and the learned joint distribution of its concepts.

Conversely, a concept can correlate with the predicted increment even when its own coefficient is zero, through the off-diagonal covariance terms or the anchor. Thus neither a large heatmap entry nor a weak entry determines the concept's direct contribution. The coefficient, the target-alignment error, and the removal experiment provide complementary evidence.

H.3 Adjacent horizons and pooling

For $h \geq 2$, the difference between two adjacent predicted horizons is

$$\hat{y}_h - \hat{y}_{h-1} = (\beta_{0,h} - \beta_{0,h-1}) + (w_h - w_{h-1})^\top C - (\gamma_h - \gamma_{h-1})A. \tag{16}$$

Its covariance decomposition uses coefficient differences in place of w_h and γ_h. More generally, for any fixed linear contrast $d \in \mathbb{R}^H$, $d^\top \widehat{\boldsymbol{y}} = d^\top b(A) + (W^\top d)^\top C$. This covers an individual horizon, adjacent-horizon increments, and averages over a specified set of horizons.

If horizons or channels are pooled, the law of total covariance introduces an additional between-group term. With a grouping variable G,

$$\mathrm{Cov}(C_k, Z) = \mathbb{E}[\mathrm{Cov}(C_k, Z \mid G)] + \mathrm{Cov}(\mathbb{E}[C_k \mid G], \mathbb{E}[Z \mid G]).$$

Therefore a pooled heatmap cannot be substituted for the coefficient matrix of an individual channel and horizon. Figure 2 is interpreted at the level of the displayed representative channels; the exact intervention rule always uses the actual fitted head.

H.4 READING THE FOUR PANELS

The HUFL panel has positive growth associations and negative signed-interaction associations in both rows. These signs are visually consistent across observed and predicted increments. In the numerically named channel, the first-harmonic associations are negative in both rows, illustrating that greater oscillatory amplitude need not co-vary with a positive local increment. Amplitude describes oscillation strength, while the sign of an increment also depends on where the series lies in its cycle.

In %UNWEIGHTED ILI, level and first-harmonic correlations change sign between the rows. This motivates examining their joint covariance with growth, volatility, and the anchor, together with their direct coefficients. ILITOTAL also exhibits different association patterns between observed and predicted increments. These channel-level differences are compatible with a shared semantic vocabulary and separately fitted channel models.

An additional identity connects the two rows without assigning a causal meaning. For prediction error $E_Z = \widehat{Z} - Z$,

$$\mathrm{Cov}(C_k, E_Z) = \mathrm{Cov}(C_k, \widehat{Z}) - \mathrm{Cov}(C_k, Z).$$

Differences in covariance expose concept-associated error. Differences in correlation also depend on the standard deviations of Z and $\widehat{Z}$, so their magnitudes cannot be read directly as an error covariance. This is why the heatmap is combined with quantitative errors and disclosed coefficients, rather than interpreted as a stand-alone explanation of forecasting success.

I WINDOW ALGEBRA AND THE MEANING OF COMPATIBILITY

I.1 THE INCOMING AND OUTGOING OBSERVATIONS

The distinction between a raw increment and a moving-window increment is central to interpreting the regularizer. Set $u = y_{t-1}$, $v = y_{t-\tau-1}$, and $m = m^\star_{t-1}$. Moving the target window forward adds u and removes v. Hence its mean changes by $d = (u - v)/\tau$, giving $m^\star_t = m + d$. This proves the mean recurrence in Eq. (11) directly by cancellation of the $\tau - 1$ shared summands.

For the second moment, the same argument gives $q^\star_t = q^\star_{t-1} + (u^2 - v^2)/\tau$. Writing the previous variance as $V = q^\star_{t-1} - m^2$, the updated variance is

$$\begin{aligned} V^\star_t &= q^\star_{t-1} + (u^2 - v^2)/\tau - (m + d)^2 \\ &= V + (u^2 - v^2)/\tau - 2md - d^2. \end{aligned} \tag{17}$$

Both the outgoing observation and the change in the mean contribute. The exact-window variance alternative therefore compares $c^2_{5,t}$ to the entire right-hand side. It is equivalent to first updating the second moment and then subtracting the updated squared mean. No independence or stationarity assumption is required for these finite-window identities.

The default raw growth target, $g^\star_t = y_{t-1} - y_{t-2}$, has a different meaning. In general $d \neq g^\star_t$. At the analytic target, the level residual becomes

$$R_{1,t}(c^\star) = \frac{y_{t-1} - y_{t-\tau-1}}{\tau} - (y_{t-1} - y_{t-2}),$$

which is generally nonzero. Thus semantic supervision and the default level penalty can prefer different states. Their combination is a soft optimization trade-off, whose practical effect is assessed by the loss sweeps and the exact-window alternative.

I.2 INTERACTION AND PROPORTIONAL-GROWTH TERMS

The interaction target has the exact relation $c^\star_{3,t} = a_t c^\star_{2,t}$. This explains the algebraic form of R_3, while allowing a learned state to violate the relation under soft supervision. It also clarifies the behavior of R_2. If the current state exactly satisfies $R_3 = 0$, then

$$R_{2,t} = -c^\star_{2,t-1} + \frac{\varepsilon}{a_t + \varepsilon} c_{2,t}. \tag{18}$$

Away from zero and for small ε, this is dominated by the previous analytic growth. The current learned/previous analytic construction therefore should not be read as an autonomous differential equation for a learned velocity trajectory. A low residual value can reflect both the current representation and the supplied previous reference.

The multiplicative alternative removes the division:

$$R'_{2,t} = a_t(c_{2,t} - c^\star_{2,t-1}) - c_{3,t}.$$

Under $R_3 = 0$, it becomes $-a_t c^\star_{2,t-1}$. It improves numerical behavior near a small denominator but does not turn the proportional-growth preference into a target identity. Table 9 separates this change from the exact mean and variance replacements, so their predictive effects can be compared individually.

Similarly, R_4 uses a local level–growth product as a variance-change prior. Its target value need not vanish because Eq. (17) involves outgoing-window information. The near-matching errors under the exact alternatives support the usefulness of the concept interface across regularizers, rather than a physical-law interpretation of every default term.

I.3 The head identity and loss contributions

Define $G_h(a, c) = a + \beta_{0,h} + \sum_k \beta_{h,k} c_k - \gamma_h(a - c_1)$. Since the implemented head is $\hat{y}_h = G_h(a, c)$, the function $R_{5,h}(\Theta) = \hat{y}_h - G_h(a, c)$ is identically zero in exact arithmetic. Consequently

$$R^2_{5,h} = 0, \qquad \nabla_\Theta R^2_{5,h} = 2R_{5,h}\nabla_\Theta R_{5,h} = 0.$$

Removing this term while preserving all nonzero loss coefficients leaves the mathematical objective and its gradients unchanged. It differs from rescaling a mean over five residuals to a mean over four, which would change the weights of the remaining terms. This is the reason for specifying unchanged weights in the head-identity control.

These distinctions also separate diagnostics from objectives. Concept-target error measures distance from semantic labels; R_3 checks an interaction relation; the other default residuals express chosen local preferences. They need not agree on whether a state is desirable. Forecast MSE evaluates the resulting predictor, and the exact decomposition in Proposition 1 specifies how semantic discrepancies enter that predictor.

J Semantic Anchoring, Dependence, and Concept Extension

J.1 Why a product feature can add predictive structure

The five targets are interpretable coordinates, not statistically independent latent factors. In particular, the interaction is determined by the observed anchor and raw growth. This dependence does not make it redundant for an affine decoder. If concepts equal their analytic targets, its contribution is

$$w_{h,3} c^\star_{3,t} = w_{h,3} a_t c^\star_{2,t}.$$

It introduces a bilinear level–growth term in the underlying variables. An affine function of a_t and $c^\star_{2,t}$ alone cannot represent that product on a two-dimensional region. For example, the mixed derivative of ag with respect to a and g is one, while the mixed derivative of any affine function of a and g is zero. The feature therefore enlarges the analytic-reference function class while keeping its downstream coefficient explicit.

Its empirical importance can nevertheless be small or domain dependent. Removing the interaction raises six-dataset mean MSE from 0.268 to 0.280, the smallest increase among the five individual removals. This supports treating it as a useful part of this compact interface, without inferring that every dataset requires the same concept set.

The first-harmonic amplitude and volatility also summarize different aspects of a window. Volatility uses the total centered second moment; first-harmonic amplitude concentrates on one Fourier component. Strong variation at higher harmonics can increase volatility without proportionally increasing the first-harmonic amplitude. Neither target carries phase. This explains both their potential complementarity and a limitation when forecasting a phase-sensitive event.

J.2 How target supervision constrains semantic relabeling

Without target supervision, a bottleneck followed by an affine head has a reparameterization freedom: for any invertible Q, replacing c by Qc and W by WQ^{-1} preserves every forecast. Prediction loss alone cannot select which rotated or mixed coordinate should be called level or volatility.

Soft anchoring supplies a fixed reference map. In the idealized case $c = c^\star$, applying a transformation Q while keeping semantic labels fixed incurs

$$\mathbb{E}\|Qc^\star - c^\star\|_2^2 = \operatorname{tr}\big((Q-I)M_\star(Q-I)^\top\big), \qquad M_\star = \mathbb{E}[c^\star(c^\star)^\top].$$

If $M_\star$ is positive definite, this quantity is zero only for $Q = I$. The calculation illustrates how fixed targets can remove an arbitrary change of basis. Positive definiteness and perfect alignment are conditions of this limiting argument, not assumptions established for all benchmarks. Under finite alignment error, the learned coordinates retain flexibility; held-out target agreement remains necessary to assess their semantics.

This separates two sources of ambiguity. Target supervision constrains how the encoder labels its coordinates. Proposition 2 characterizes whether the fitted head distinguishes changes in those coordinates. Accurate forecasting, semantic anchoring, and distinguishable downstream use are three related but separate properties.

J.3 Extending the interface with acceleration

For a new past-computable statistic $s_t^\star$, extending the interface adds a learned coordinate $c_{6,t}$, a soft alignment term, and one coefficient per forecast horizon:

$$\widehat{y}_h^{(6)} = b_h(a) + w_h^\top c_{1:5} + w_{h,6}c_6, \qquad \mathcal{L}_{\text{con}}^{(6)} = \mathcal{L}_{\text{con}}^{(5)} + \mathbb{E}(c_6 - s^\star)^2.$$

The acceleration example uses $s_t^\star = y_{t-1} - 2y_{t-2} + y_{t-3}$. It needs three past observations and measures the change between the two latest raw growth increments. A sixth output adds one row and bias to the final concept layer and one column to the horizon readout. The history encoder and the original five target definitions remain intact.

Table 12 shows the observed trade-off. Exchange MSE decreases from 0.051 to 0.048, while every other displayed dataset becomes worse and the six-dataset mean rises to 0.290. The extension mechanics are straightforward, but additional semantic coordinates have a predictive cost as well as a possible benefit.

A useful extension has an interpretable target available at forecast time, a role distinct from the existing coordinates, and a benefit assessed on validation data. For a fixed fitted head, its coefficient column reveals its horizon profile; joint dependence with existing columns is examined through the same observability analysis, now with six coordinates. Selection using validation data precedes final test reporting. The acceleration results illustrate one such candidate, rather than an exhaustive search over possible concepts.

K Comparative Structure of the Forecasting Results

K.1 Primary-horizon margins against the strongest displayed alternative

The primary table supports two complementary views: absolute MSE/MAE and the margin over the strongest competing value in each column. For a metric M and dataset d, define

$$G_{d,M} = 100\frac{\min_{b\neq\text{CODA}} M_{b,d} - M_{\text{CODA},d}}{\min_{b\neq\text{CODA}} M_{b,d}}.$$

A positive value favors CODA. The comparator can differ by dataset and metric. Table 17 calculates these margins from Table 1; these are transformations of the reported scores, with no additional forecast runs.

Table 17: Primary-horizon margins calculated against the best competing score in Table 1. Positive percentages favor CODA.

Dataset	Best other MSE	MSE reduction	Best other MAE	MAE reduction
Electricity	0.129	-7.0%	0.218	+1.8%
Traffic	0.329	+4.9%	0.235	+3.8%
Weather	0.136	+6.6%	0.164	+6.7%
Illness	0.877	+0.9%	0.754	+1.9%
Exchange	0.071	+28.2%	0.190	+11.6%
ETT	0.112	+0.9%	0.224	-5.4%

The largest primary-horizon MSE margin is on Exchange, where N-HiTS is the strongest displayed alternative at 0.071 and CODA reaches 0.051. The 28.2% reduction against that alternative differs from the 42.7% reduction against ModernTCN; both use the same CODA score but answer different comparative questions. On Weather, the best alternative for MAE is N-HiTS (0.164), while ModernTCN is the best alternative for MSE (0.136).

Illness and ETT have smaller primary MSE margins over the closest alternative. Electricity favors CODA in MAE but not MSE, and ETT favors it in MSE but not MAE. These metric-dependent rankings are retained alongside the ten winning metric–dataset cells. MSE emphasizes larger errors more strongly than MAE; the rankings alone do not identify which particular events account for the difference. All primary comparisons retain the baseline provenance described in Appendix B.

K.2 Ranks across the common multi-horizon comparisons

Raw mean MSE combines error magnitudes across datasets. A rank summary instead gives equal importance to ordering within each dataset–horizon setting. With three methods and lower scores preferred, the midrank assigned to method m is

$$r_m(d,h) = 1 + \sum_{b \neq m} \mathbf{1}\{M_{b,d,h} < M_{m,d,h}\} + \frac{1}{2} \sum_{b \neq m} \mathbf{1}\{M_{b,d,h} = M_{m,d,h}\}.$$

Tied displayed scores receive the same midrank. The ranks in Table 18 are calculated from Table 3 and Table 11. The four common-horizon columns average over five datasets. The final column also includes the four Illness settings, for 24 comparisons in total.

Table 18: Mean ranks calculated from the separate multi-horizon evaluation. Lower is better; ties use midranks. The final column includes Illness at its own four horizons.

Method	$H = 96$	$H = 192$	$H = 336$	$H = 720$	All 24 settings
N-HiTS	2.80	2.80	2.70	2.80	2.81
ModernTCN	1.90	2.00	2.10	2.00	2.00
CODA	1.30	1.20	1.20	1.20	1.19

CODA has 19 strict wins, one tie for lowest error, and four losses. The losses are all on Electricity, while the tie is ETT at $H = 96$. The overall mean ranks are 1.19 for CODA, 2.00 for ModernTCN, and 2.81 for N-HiTS. The favorable ordering therefore persists when comparisons emphasize within-dataset rank rather than cross-dataset error magnitude.

Ranks intentionally discard effect size. A narrow win and a large win receive the same rank, so the per-dataset numerical tables remain necessary. The 24 settings also share datasets and overlapping forecast horizons; the count is a descriptive summary rather than 24 independent statistical trials. Likewise, a tie at three displayed decimals does not establish equality of the underlying unrounded means. This treatment makes the scope of the comparative advantage explicit while preserving all reported errors.

K.3 Interpreting ablation magnitudes and noise ratios

Component-ablation percentages use the arithmetic means of their six displayed dataset scores before rounding. For the compatibility-off, alignment-off, and independent-head controls, the means are

0.358167, 0.741333, and 0.610833, compared with 0.268167 for CODA. This gives relative increases of 33.6%, 176.4%, and 127.8%. Computing percentages from rounded means can change the last decimal, so the same aggregation rule is used throughout the text and tables.

Concept removal, coefficient editing, and input-noise injection operate at different levels. Removal retrains a model with a changed interface; coefficient editing holds the model fixed; noise injection changes the observed input and can change several concepts and the anchor jointly. Agreement among these views supports an interpretation, but one result cannot be substituted for another. In particular, the noise ratio is the perturbed MSE divided by clean MSE for the same model; a smaller ratio measures less relative degradation, while absolute forecasting quality still depends on the clean error. The main results therefore report both clean benchmark errors and the separate robustness ratios.