INFORMATION ANCHORS: UNDERSTANDING AND USING FORECAST-RELEVANT INFORMATION IN TIME-SERIES FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Time-series foundation models (TSFMs) achieve strong forecasting performance, but output accuracy alone reveals little about what internal information their forecasts rely on, how that information is organized, or whether it is stable and functionally meaningful. We introduce INFORMATION ANCHORS, a framework for identifying hidden representations that carry strong information about the future. Using mutual information as a common measure, our framework reveals where forecast-relevant information concentrates, what forecast properties it represents, and how strongly model predictions respond to it. Across six heterogeneous TSFMs and seven data regimes, we find that forecast-relevant information is concentrated in a limited set of temporal regions rather than uniformly distributed across model representations. Compared with representations carrying less future information, information anchors expose more forecast properties, are associated with distinctive patterns in the observed history, and produce stronger prediction responses under controlled intervention. We further demonstrate the practical value of this information through ANCHOR-RAG, which uses information-anchor signals to guide the retrieval and integration of historical evidence. Overall, INFORMATION ANCHORS makes forecast-relevant information in TSFMs identifiable, interpretable, and practically useful, providing a general framework that connects model understanding with improved forecasting.

1 INTRODUCTION

Time-series foundation models (TSFMs) have shown strong forecasting ability across variables and domains (Ansari et al., 2024; 2025; Liu et al., 2025). However, their predictions provide limited insight into the information learned inside the models. A basic question therefore remains: *what forecast-relevant information do TSFMs learn, how is it represented internally, and is this information useful for understanding and improving forecasting?* Answering this question is important for both interpreting TSFMs and making better use of their learned representations.

The main difficulty is that forecast-relevant information has no explicit boundary inside a TSFM. Hidden representations mix useful temporal structure with contextual, redundant, and model-specific information, making it difficult to determine which parts are closely related to the future. Existing studies have examined influential inputs, decodable properties, or model responses to internal changes (Tonekaboni et al., 2020; Crabbé & Van Der Schaar, 2021; Alain & Bengio, 2016; Hewitt & Liang, 2019; Meng et al., 2022). However, these analyses do not provide a common view of forecast-relevant information itself: where it is concentrated, what it represents, and whether it can serve as a useful signal outside the original prediction process.

We introduce INFORMATION ANCHORS, a framework for identifying and studying forecast-relevant information in the hidden representations of frozen TSFMs. The key idea is to use mutual information (MI) (Kraskov et al., 2004) to measure how strongly an internal representation is related to the future. Representations carrying stronger future information are identified as information anchors. We then study what these anchors encode, how model predictions respond when they are changed, and what historical structures are associated with them. This provides a unified view of the location, content, and functional value of forecast-relevant information without modifying the underlying TSFM.

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.

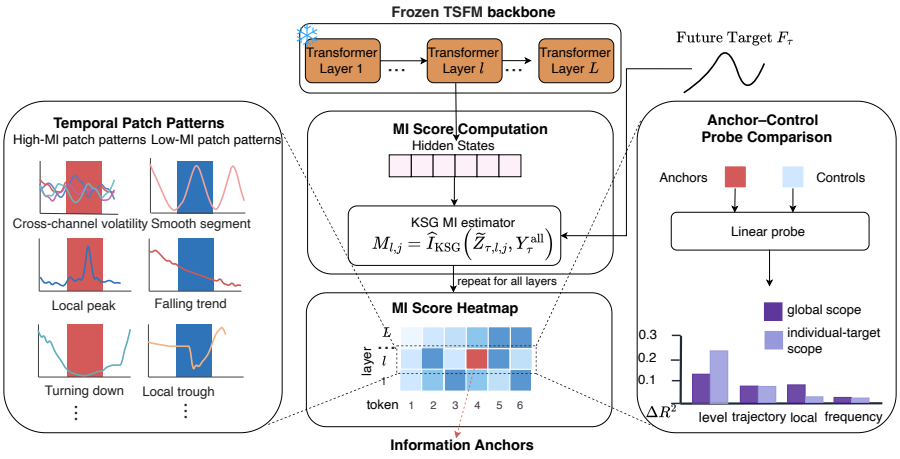

Figure 1: **Overview of INFORMATION ANCHORS.** A frozen TSFM yields a null-calibrated MI atlas; held-out probes, state replacement, and history-structure contrasts characterize the nominated temporal region; ANCHOR-RAG reuses the resulting history-only signal in a matched TS-RAG forecasting pipeline.

We evaluate INFORMATION ANCHORSon six TSFMs covering encoder, decoder-style, encoder–decoder, and mixer designs across seven data regimes.We find that forecast-relevant information is not uniformly distributed inside these models, but concentrates in a limited set of temporal regions. Compared with representations carrying less future information, information anchors expose more forecast properties, are associated with distinctive structures in the observed history, and produce stronger prediction responses under controlled intervention. These findings are broadly observed across the evaluated models, although their detailed information patterns differ.

We further examine whether the identified information can guide forecasting. Standard retrieval-augmented forecasting typically compares and uses historical representations without distinguishing how much forecast-relevant information they carry(Ning et al., 2026; Tire et al., 2024). Based on the information-anchor view, we develop ANCHOR-RAG, which uses forecast-relevant information to guide the retrieval and integration of historical evidence. ANCHOR-RAG achieves superior forecasting performance over state-of-the-art baseline, which demonstrate that the internal information revealed by INFORMATION ANCHORSis not only interpretable, but also useful for improving forecasting. Our contributions are:

- We introduce INFORMATION ANCHORS, a mutual-information-based framework that makes forecast-relevant information in TSFM representations an explicit object of analysis.
- We systematically characterize and validate information anchors across six heterogeneous TSFMs and seven data regimes, revealing where forecast-relevant information concentrates, what it represents, and how strongly predictions respond to it.
- We demonstrate the practical value of the identified information through ANCHOR-RAG, which uses information-anchor signals to guide retrieval and forecast integration and consistently improves over TS-RAG under matched settings.

2 SETUP

2.1 FORECASTING OBJECTS

For a fixed panel of C channels, let $x_t = (x_{t,1}, \ldots, x_{t,C}) \in \mathbb{R}^C$ be the observation at time t. At origin τ, history length L, and horizon P, the input, realized future, and frozen-model forecast are

$$X_\tau = [x_{\tau-L+1}, \ldots, x_\tau] \in \mathbb{R}^{L\times C}, \qquad F_\tau = [x_{\tau+1}, \ldots, x_{\tau+P}] \in \mathbb{R}^{P\times C}. \tag{1}$$

The frozen model m then produces

$$\widehat{F}_\tau = f_m(X_\tau) \in \mathbb{R}^{P\times C}. \tag{2}$$

All-variable analyses use this same fixed channel panel for the input, future summary, and prediction error. A designated target channel $c^\star$ supplies an additional, focused readout; it does not replace the multivariate analysis. For the downstream retrieval task, H_q, Y_q, and $\widehat{Y}_q$ denote the history, future, and forecast of one channel-wise query extracted from such an origin.

2.2 Native temporal states

Let $H^{(m)}_{\tau,l,j}$ be the frozen model's native state at layer or stage l and history patch j. Its shape may contain packed variables, separate channel states, or a channel-independent state. We analyze the vector $Z^{(m)}_{\tau,l,j} \in \mathbb{R}^{d_{m,l}}$: it is the native vector when one is available, or the concatenation of the fixed panel's channel states at patch j when channels are explicit. This grouping occurs after the forward pass and does not change the forecast. Normalized patch position and layer depth permit comparison across different native grids.

2.3 Evidence for an information anchor

An information anchor is a selected temporal region, assessed at a model-specific layer. We link three measurements: dependence between its state and a structured future summary; future properties recoverable from that state on held-out origins; and the change in prediction error after replacing the state with a matched donor. These measurements answer where future-related information concentrates, what it carries, and whether the forecast responds to changing it. They are evaluated on chronologically separated records.

3 Information-Anchor Framework

3.1 Locating future-related states

To locate information that persists beyond chance temporal alignment, we compare native states with a fixed summary Y^{all}_τ of the complete multivariate future. The summary combines level, trajectory, local-change, and spectral properties; projections of both future summaries and states are fitted on discovery data only. At each layer–patch position, a KSG estimator (Kraskov et al., 2004) measures $M_{l,j} = \widehat{I}_{\mathrm{KSG}}(\widetilde{Z}_{\tau,l,j}, Y^{\mathrm{all}}_\tau)$. This fixed nonparametric estimator provides one scoring rule across checkpoints; neural MI estimation offers a related approach (Belghazi et al., 2018). Legal temporal shifts give a position-specific null, so a large score reflects dependence relative to the same position under broken history–future alignment:

$$z_{l,j} = \frac{M_{l,j} - \mu^0_{l,j}}{\sigma^0_{l,j}}, \qquad p_{l,j} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}[M^{(b)}_{l,j} \geq M_{l,j}]}{B+1}. \tag{3}$$

We control the within-map false-discovery rate by the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995). Among layers on the forecast path, we select the one with the greatest mean calibrated information mass and identify its highest- and lowest-scoring patches:

$$l^\star = \arg\max_{l\in\mathcal{R}_m} a_l, \qquad j^+ = \arg\max_j z_{l^\star,j}, \qquad j^- = \arg\min_j z_{l^\star,j}. \tag{4}$$

Here $\mathcal{R}_m$ is the set of reachable layers and a_l is their mean calibrated mass. Selecting temporal position and model-specific layer separately accommodates different architectures. The future summary, estimator settings, shift count, and tie rules are fixed in Appendix C.

3.2 Testing content and prediction response

Dependence alone does not say which forecast properties a state contains. Chronologically held-out ridge probes therefore compare the selected High-MI state with a Low-MI state at the same layer. They read out level, change, local dynamics, and frequency for both the full channel panel and

target $c^\star$. Signed test R^2 retains weak or negative readouts, allowing the comparison to reveal which properties are actually recoverable. The precise targets and regularization protocol appear in Appendix C.

To test whether the frozen predictor responds to the selected state, we replace complete native states with states from temporally separated donor histories, using the same donors for High-MI, Low-MI, and random patch orders. The primary response is $\Delta\,\mathrm{MSE} = \mathrm{MSE}_{\mathrm{all}}(\widetilde{F}_{\tau,l,S}, F_\tau) - \mathrm{MSE}_{\mathrm{all}}(\widehat{F}_\tau, F_\tau)$, where $\widetilde{F}$ is the forecast after replacement of patch set S. Positive values mean worse prediction after replacement. Varying patch budgets, replacement dose, and donor matching tests whether the ordering persists beyond perturbation size or donor choice. Forecast drift provides a complementary measure of output displacement.

Finally, a cross-fitted pointwise MI scorer ranks patches within each observed history. Equal-size High- and Low-MI sets are compared on deterministic labels computed only from the observed history. This identifies the historical contexts in which future-related states occur, while keeping future outcomes out of those labels. Sampling, donor controls, and inference are specified in Appendices C and H.

3.3 ANCHOR-GUIDED FORECASTING

To reuse information anchors, ANCHOR-RAG keeps the Chronos-Bolt TS-RAG predictor (Ning et al., 2026) frozen (Figure 2). For channel-wise query q, H_q, Y_q, and $\widehat{Y}_q$ denote its observed length-L history, unknown P-step future, and forecast. Candidate i provides (H_i, Y_i); official history distance d_{qi}^{off} forms pool $\mathcal{P}_q$. On discovery history–future pairs, cross-fitted pointwise MI labels patches, and ridge predictor $\widehat{s}_j$ learns each label from the hidden token at $j \in \{1, \ldots, J\}$. At validation and test, the sidecar sees only observed-history tokens Z_{xj} for $x \in \{q\} \cup \mathcal{P}_q$. Write $\widehat{m}_{xj} = \widehat{s}_j(Z_{xj})$ and $u_{xj} = \mathrm{softmax}_j(\mathrm{zscore}_j\, \widehat{m}_{xj})$, with both operations across the J patches of history x. The history-only distance is

$$d_{qi}^{\mathrm{MI}} = 1 - \frac{\sum_j \sqrt{u_{qj}u_{ij}}\, \cos(Z_{qj}, Z_{ij})}{\sum_j \sqrt{u_{qj}u_{ij}}}. \tag{5}$$

Here cos is token cosine similarity. The sidecar selects the K smallest MI distances in $\mathcal{P}_q$, using stable order for ties:

$$\mathcal{I}_q = \mathrm{TopK}^{\min}_{i \in \mathcal{P}_q}\; d_{qi}^{\mathrm{MI}}. \tag{6}$$

Stored continuations are gathered only after selection. With $F_\theta^{(0.5)}$ the frozen TS-RAG/Chronos-Bolt median, the retrieval forecast is

$$\widehat{Y}_q^{\mathrm{ret}} = F_\theta^{(0.5)}(H_q, \{(H_i, Y_i) : i \in \mathcal{I}_q\})\,. \tag{7}$$

The frozen branch uses selected sequences without a distance bias. Official distance forms the pool; MI distance selects neighbors and weights the outer correction. Both distances use observed histories, never Y_q.

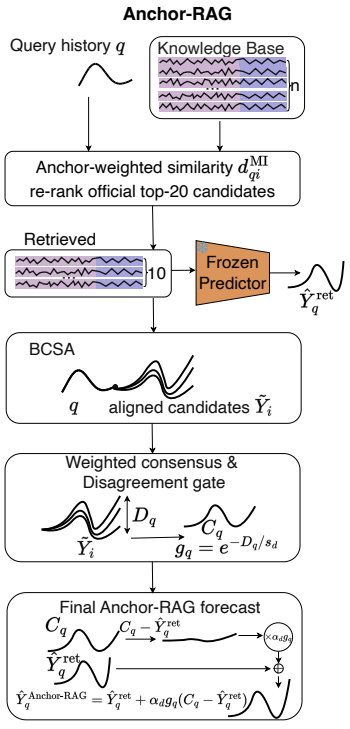

Figure 2: **ANCHOR-RAG pipeline.** MI selects ten of the official 20 candidates; BCSA and the disagreement gate correct the frozen forecast.

The continuations form a second signal: MI weights them, BCSA aligns them to the query endpoint, and disagreement gates correction. Model weights, pool, K, origins, and normalization stay fixed. Let $\bar{d}_q^{\mathrm{MI}}$ and $\mathrm{sd}(d_q^{\mathrm{MI}})$ be the mean and standard deviation of selected distances, and let $\epsilon > 0$ stabilize denominators. For $i \in \mathcal{I}_q$, standardized distance z_{qi} and normalized weight w_{qi} are

$$z_{qi} = \frac{d_{qi}^{\mathrm{MI}} - \bar{d}_q^{\mathrm{MI}}}{\max\{\mathrm{sd}(d_q^{\mathrm{MI}}), \epsilon\}}, \qquad w_{qi} = \frac{\exp(-z_{qi}/T)}{\sum_{k \in \mathcal{I}_q} \exp(-z_{qk}/T)}. \tag{8}$$

These weights balance proximity with entropy and preserve a distributed evidence set.

To align candidate futures, let $H_{i,L}$ and $H_{q,L}$ be observed endpoints, ΔH first differences, and MAD their median absolute deviation. For step $h \in \{1, \dots, P\}$, Boundary-Conditioned Shape Alignment (BCSA) sets

$$a_q = \mathrm{MAD}(\Delta H_q) + \epsilon, \qquad a_i = \mathrm{MAD}(\Delta H_i) + \epsilon,$$
$$r_i = (a_q/a_i)^\eta, \qquad \widetilde{Y}_{i,h} = H_{q,L} + r_i(Y_{i,h} - H_{i,L}). \tag{9}$$

BCSA anchors displacements at the query endpoint and moderates scale through η; Appendix I gives the evaluated K, T, and η.

For dataset d, validation selects disagreement scale $s_d > 0$. Aligned consensus C_q, normalized disagreement D_q, and confidence g_q are

$$C_q = \sum_{i \in \mathcal{I}_q} w_{qi} \widetilde{Y}_i, \qquad D_q = \frac{P^{-1} \sum_{h=1}^{P} \sum_{i \in \mathcal{I}_q} w_{qi} (\widetilde{Y}_{i,h} - C_{q,h})^2}{\max\{\mathrm{Var}(H_q), \epsilon\}}, \qquad g_q = e^{-D_q/s_d}. \tag{10}$$

Here $C_{q,h}$ is consensus step h, $\mathrm{Var}(H_q)$ uses observed history, and $\alpha_d \in [0, 1]$ is the validation-selected gain. The final forecast is

$$\widehat{Y}_q^{\text{Anchor-RAG}} = \widehat{Y}_q^{\text{ret}} + \alpha_d g_q (C_q - \widehat{Y}_q^{\text{ret}}). \tag{11}$$

Since $0 \le \alpha_d g_q \le 1$, the convex update is bounded by $\alpha_d \|C_q - \widehat{Y}_q^{\text{ret}}\|$ and attenuated by disagreement. The validation-selected α_d, s_d are reused at test and in ablations unless their operation is disabled; Appendix I gives the grid and switches.

4 Experiments

We ask four linked questions: **RQ1**, where does future-related information concentrate, and at what resolution is its location reproducible? **RQ2**, which future properties can be read from the selected states across model architectures? **RQ3**, do changes to those states affect forecasts, and which observed histories contain them? **RQ4**, can this information improve matched retrieval forecasting, and which operations account for the gain?

We analyze six frozen checkpoints spanning encoder-only, decoder-style, encoder–decoder, and mixer computation paths (Appendix B). The common seven data regimes are ETTh1, ETTh2, ETTm1, ETTm2, Weather, Electricity, and Traffic (Zhou et al., 2021; Wu et al., 2022). ETT and Weather use all channels; Electricity and Traffic use fixed 16-channel panels shared by all models (Appendix A). Discovery fits the information maps and selections, validation selects probe regularization, and chronologically held-out origins support evaluation. Guard bands separate splits and moving-block inference addresses overlapping origins (Appendix C). All model weights remain frozen; primary errors average the fixed multivariate channel panel. The 42 maps yield 84 global and target probe runs and 42 intervention curves. Full-variable Electricity and Traffic analyses assess scaling beyond the fixed panels (Appendix K).

4.1 RQ1: Location and stability

To establish the spatial resolution of an anchor, we map MI across native layer–patch grids and resample chronological discovery blocks. Across 42 maps, temporal MI mass is concentrated (Figure 3a): the top quarter of patches holds 0.411 of the mass on average, versus 0.25 under a uniform allocation. The mean maximum-to-uniform anchor ratio is 1.879. Layer profiles differ among models (Figure 3b), so a common absolute layer index cannot represent every architecture. These are within-map summaries; raw MI magnitudes are not compared across checkpoints. Model-mean top-quarter mass lies between 0.401 and 0.417 across the six checkpoints. Temporal concentration therefore recurs across architectures even as their layer profiles take different forms.

The future summary emphasizes coarse structure: 68.6% of absolute projection loading mass falls on relative-level bins, versus 14.8% on trajectory, 6.0% on local dynamics, and 10.6% on frequency. The location results measure dependence on this structured future. Its component properties also give the subsequent probes interpretable targets.

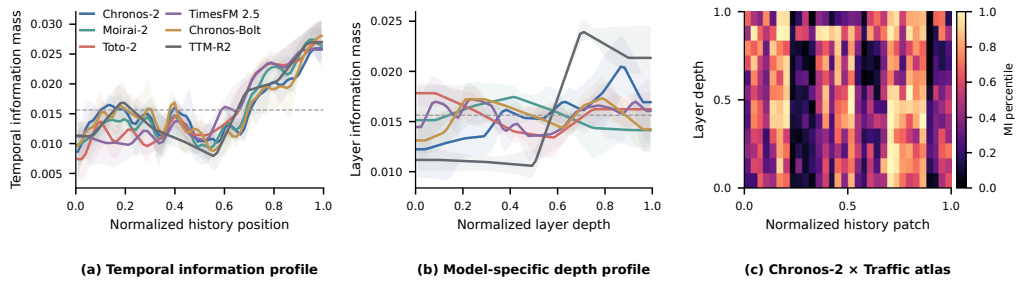

Figure 3: **Where future-related information concentrates.** **(a)** Temporal MI mass, with interquartile shading and a uniform reference. **(b)** Mass over normalized depth. **(c)** Chronos-2 on Traffic, colored by within-map MI percentile.

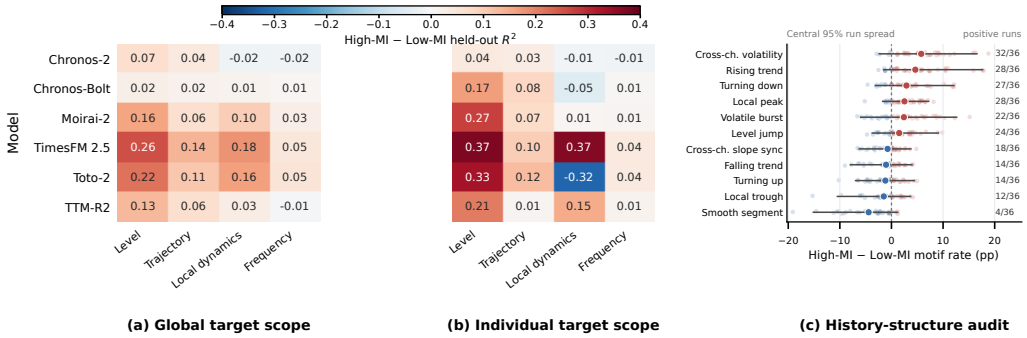

Figure 4: **Future readouts and observed histories.** **(a,b)** Held-out High-minus-Low probe R^2 for all channels and the selected target. **(c)** Equal-budget differences in history-pattern rates across 36 valid maps; bars span the central 95% range.

Across 50 block-resampling replicates per map, mean reselection is 0.484 for the selected patch, 0.333 for its layer, and 0.188 for the exact layer–patch pair (Appendix K). Patch recovery is 0.483–0.793 on ETT and Weather, compared with 0.103 on Electricity and 0.093 on Traffic's fixed channel panels; the globally top-ranked patch aggregated across layers has 0.725 mean recovery. Temporal position is thus the more reproducible location, while the layer supplies a fixed, model-specific site for the following analyses. The recovery gap between patch and exact pair guides interpretation: probes and interventions use a fixed selected layer, while cross-model localization concerns temporal position rather than a universal hidden layer.

4.2 RQ2: Recoverable future properties

To identify what the selected states carry, we fit chronological ridge probes on High-MI states and same-layer Low-MI controls, reading out future properties for all channels and for the selected target separately (Figure 4a,b). This tests whether high dependence corresponds to recoverable forecast structure, rather than a location score alone. Across 42 maps, High-minus-Low R^2 for level and trajectory averages $0.143/0.071$ across channels and $0.233/0.070$ for the target; frequency differences are smaller. All four computation paths therefore expose more coarse future information at their selected states, with a particularly strong target-level component. The relatively small frequency contrast agrees with the future summary's lower spectral loading, so the strongest readout concerns level and broad trajectory. The same hidden state is tested against both scopes. The stronger target-level contrast shows that the information map can locate states relevant to a particular forecast even when the model processes several channels together.

The form of that readout follows the computation path. The three decoder-style models have the strongest level contrast ($0.211/0.327$ globally/for the target), reaching $0.257/0.374$ in TimesFM 2.5.

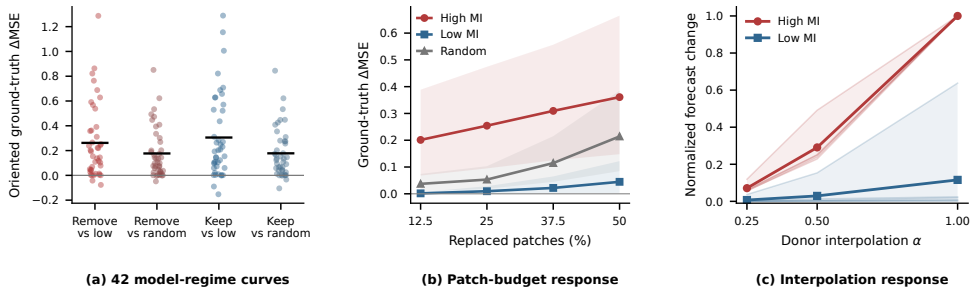

Figure 5: **Forecast response to state replacement. (a)** Error change across 42 model–regime curves. **(b)** Response by patch budget, with 95% bootstrap bands. **(c)** Normalized response by interpolation dose.

Their analyzed states lie on causal forecast paths, where coarse future structure remains recoverable by the held-out probes. Encoder-only Chronos-2 is more selective ($0.068/0.040$), showing that strong dependence does not imply an equally strong linear readout. Chronos-Bolt shifts from 0.023 across channels to 0.166 for the selected target before decoder cross-attention, concentrating accessibility at that interface. TTM-R2 occupies an intermediate regime ($0.131/0.209$) that differs from both encoder-only and decoder-style profiles. Comparable temporal localization therefore coexists with distinct recoverable future properties in native states. Thus shared temporal concentration does not require identical representations: the future properties exposed at an anchor depend on the model's computation path. MI establishes dependence on the structured future, while the held-out linear readout shows how directly a specific forecast property is available from the state. Their different profiles expose architecture-related differences in representation while keeping the location criterion fixed.

4.3 RQ3: Prediction response and history context

We test functional response by replacing selected native states with matched donor states, then comparing the change in ground-truth MSE with Low-MI and random patch orders. Separately, a history-only audit tests which observed patterns coincide with high pointwise MI. Across four nested budgets, High-MI replacement exceeds Low-MI and Random in 39/42 model–regime curves, with mean oriented effects 0.262 and 0.177 and crossed bootstrap intervals $[0.094, 0.488]$ and $[0.067, 0.329]$ (Figure 5). Complementary keep sets preserve the ordering. Mean ground-truth ΔMSE rises from 0.201 to 0.361 for High-MI, versus 0.002 to 0.045 for Low-MI and 0.037 to 0.214 for Random; within-state interpolation is monotone in 98.2% of evaluated pairs.

RMS-normalized contrasts remain positive in 37–40/42 High-versus-Low curves and 34–38/42 High-versus-Random curves. Outcome-blind matched donors retain a positive High-minus-Low response in 18/18 available curves (mean 0.165, 95% CI $[0.056, 0.307]$). Together, these controls address scale and donor mismatch while retaining position-specific response. The matched-donor effect reflects information at selected patches together with the forecast path's sensitivity to their coordinates; it is evidence of prediction response, not an isolated causal effect of MI itself. Increasing replacement budgets strengthen the High-MI response while the Low-MI trajectory remains flatter. The dose control shows a graded response at the same coordinate, connecting the full replacement effect to continuous perturbation.

Among the 36 maps with valid pointwise MI scorers, High-MI patches are enriched for cross-channel volatility ($+5.76$ percentage points; 32/36 positive runs), rising trends ($+4.62$; 28/36), turning down ($+2.91$; 27/36), and local peaks ($+2.52$; 28/36), while smooth segments are depleted (-4.42; 4/36 positive). These equal-budget contrasts describe target-aligned co-occurrence in observed history; Appendix H reports all 11 labels. Every checkpoint contributes positive mean cross-channel-volatility enrichment and negative mean smoothness enrichment, linking the pooled pattern to each computation path. Appendix G reports the dose, RMS, donor, and coordinate audits. Together, replacement connects selected states to the frozen forecast's behavior, while the observed-history comparison locates recurring contexts such as changing cross-channel volatility and turns. The two tests characterize complementary aspects of an anchor.

Table 1: **Matched downstream forecasts.** Raw MSE/MAE (lower in bold, other underlined); relative reductions from TS-RAG include paired moving-block 95% intervals. † marks unadjusted block sign-flip $p < 0.05$. Full precision and exact p values appear in Appendix I.

Dataset	ANCHOR-RAG (Ours)	TS-RAG	ΔMSE (%) [95% CI]	ΔMAE (%) [95% CI]
ETTh1	**0.3484/0.3615**	0.3556/0.3623	$+2.027\,[1.151, 2.902]^{\dagger}$	$+0.239\,[-0.089, 0.557]$
ETTh2	**0.2443/0.2973**	0.2451/0.2981	$+0.312\,[-0.118, 0.707]$	$+0.257\,[0.047, 0.479]^{\dagger}$
ETTm1	**0.2768/0.3085**	0.2904/0.3113	$+4.661\,[3.572, 5.595]^{\dagger}$	$+0.884\,[0.398, 1.342]^{\dagger}$
ETTm2	**0.1437/0.2217**	0.1465/0.2230	$+1.853\,[1.242, 2.461]^{\dagger}$	$+0.584\,[0.312, 0.863]^{\dagger}$
Weather	**0.1349/0.1703**	0.1453/0.1770	$+7.166\,[5.657, 8.732]^{\dagger}$	$+3.790\,[2.855, 4.671]^{\dagger}$
Electricity	**0.1110**/0.2004	0.1120/**0.2003**	$+0.901\,[0.757, 1.048]^{\dagger}$	$-0.060\,[-0.120, 0.002]$
Exchange Rate	**0.0598/0.1678**	0.0624/0.1716	$+4.247\,[0.362, 8.101]$	$+2.229\,[0.165, 4.465]$
Wins / macro mean	MSE: 7/7; MAE: 6/7		$+3.024\,[2.405, 3.657]$	$+1.132\,[0.782, 1.484]$

4.4 RQ4: FORECASTING WITH INFORMATION ANCHORS

To test whether the anchor signal improves forecasting, we compare ANCHOR-RAG with matched TS-RAG and disable individual operations using the same frozen Chronos-Bolt backbone. MI has two direct roles: it selects ten histories from the official 20-candidate pool for the frozen predictor, then weights their aligned continuations in the outer correction. Candidate disagreement, computed with those MI weights, controls the correction gate. The downstream suite uses Exchange Rate in place of the mapping study's Traffic regime. Every run uses $L = 512$, $P = 64$, identical origins and normalization, and ten neighbors. The pure-MI selector and validation-selected α_d, s_d are fixed before test evaluation. Appendix I records the protocol and selected values, while Table 1 reports all seven test sets.

ANCHOR-RAG reduces MSE on all seven datasets and MAE on six, with equal-dataset mean reductions of 3.02% and 1.13%. MSE gains are largest on Weather (7.17%), ETTm1 (4.66%), and Exchange Rate (4.25%); paired block sign-flip tests reject zero change in five MSE and four MAE comparisons. Electricity has an MSE reduction of 0.90% with nearly equal MAE (0.2004 versus 0.2003), consistent with a stronger effect on larger errors. Exchange Rate has a positive bootstrap interval, while its block sign-flip test gives $p = 0.0698$. The two inferential summaries therefore capture different aspects of the paired test distribution.

All ablations reuse Full's validation-selected (α_d, s_d) and test origins. The w/o MI dist. and w/o MI-cond. gate arms retain the MI-selected frozen forecast and change only correction distance weights or the disagreement gate. The joint w/o MI arm restores official Top-10 selection for the frozen branch and official distances for correction; $\alpha_d = 0$ exposes the raw MI-selected frozen forecast. BCSA is disabled separately. Appendix I specifies every switch and the paired tests

Table 2: **Locked-parameter system ablation.** Raw MSE/MAE on seven datasets; the best and second distinct values among the displayed variants are bold and underlined. Each arm retains Full's validation settings except for the disabled operation.

Dataset	ANCHOR-RAG (Ours)	TS-RAG	w/o MI dist.	w/o MI-cond. gate	w/o BCSA	w/o MI	w/o MI + BCSA	w/o residual correction
ETTh1	0.3484/0.3615	0.3556/0.3623	0.3484/**0.3614**	**0.3459**/0.3618	0.3477/0.3616	0.3490/0.3615	0.3485/0.3617	0.3554/0.3622
ETTh2	0.2443/**0.2973**	0.2451/0.2981	0.2442/**0.2973**	0.2588/0.3109	0.2462/0.3019	**0.2433**/0.2975	0.2447/0.2989	0.2460/0.2978
ETTm1	**0.2768/0.3085**	0.2904/0.3113	0.2772/0.3087	0.3029/0.3161	0.2782/0.3127	0.2784/0.3093	0.2794/0.3129	0.2897/0.3110
ETTm2	**0.1437/0.2217**	0.1465/0.2230	0.1438/**0.2217**	8.8388/0.5583	0.1546/0.2355	0.1450/0.2226	0.1539/0.2340	0.1458/0.2224
Weather	**0.1349/0.1703**	0.1453/0.1770	0.1370/0.1727	0.1564/0.1785	0.1448/0.1887	0.1409/0.1769	0.1497/0.1933	0.1432/0.1747
Electricity	**0.1110**/0.2004	0.1120/0.2003	0.1111/0.2006	0.3074/0.2122	0.1112/0.2011	0.1113/0.2011	0.1116/0.2019	0.1119/**0.2001**
Exchange Rate	**0.0598/0.1678**	0.0624/0.1716	0.0600/0.1680	38.3487/0.9793	0.3081/0.3299	0.0599/0.1684	0.2316/0.2814	0.0627/0.1718

MI contribution. Against the arm that jointly replaces MI selection and correction distances with official choices, Full gains 0.846%/0.743% in macro MSE/MAE, with paired bootstrap intervals $[0.465, 1.231]$/$[0.536, 0.960]$ (Table 2; Appendix Table 22). Holding the MI-selected frozen forecast fixed, replacing MI distances in the correction changes its weights and weighted disagreement; Full then improves macro MSE/MAE by 0.308%/0.238% with intervals $[0.140, 0.483]$/$[0.129, 0.344]$. Correction-side MI distance therefore contributes even when the frozen predictor receives the same ten sequences.

The raw MI-selected forecast improves on TS-RAG for Weather (0.1432 versus 0.1453) and ETTm1 (0.2897 versus 0.2904), while Exchange Rate moves from 0.0624 to 0.0627 MSE. Alignment and gated correction lower these three results to 0.1349, 0.2768, and 0.0598; Full improves MSE over the raw MI-selected forecast on all seven datasets. MI identifies and weights the retrieved evidence, while alignment and confidence control make its contribution reliable across the matched suite.

Boundary alignment and disagreement gating have larger effects. Without BCSA, Exchange Rate MSE rises from 0.0598 to 0.3081 and median per-origin MSE from 0.051 to 0.235, indicating broad endpoint and scale mismatch across ordinary as well as difficult origins. Setting $g_q = 1$ applies Full's locked correction regardless of candidate disagreement: ETTm2 and Exchange Rate MSE jump to 8.8388 and 38.3487, while their 99th-percentile per-origin losses rise from 0.927/0.172 to 220.5/542.4. Exchange Rate's $\alpha_d = 1$ magnifies this unattenuated update. When candidate continuations disagree, an ungated consensus can move far from the frozen forecast; squared error then amplifies the few largest misses. Even ETTm2's smaller $\alpha_d = 0.10$ permits large errors when the consensus itself is far away. The quantiles show why these large mean MSE values are tail-driven, whereas the BCSA shift also affects the typical Exchange Rate origin. All variants retain Full's validation-selected parameters.

5 Related Work

Time-series explanations. FIT, Dynamask, and ContraLSP locate influential input regions (Tonekaboni et al., 2020; Crabbé & Van Der Schaar, 2021; Liu et al., 2024c). Information Anchors studies future dependence in frozen hidden states.

Representation probing and intervention. Probes test recoverability (Alain & Bengio, 2016; Hewitt & Liang, 2019); causal tracing and Patchscopes study hidden-state effects and content (Meng et al., 2022; Ghandeharioun et al., 2024). Gradient-based latent attribution (Shu et al., 2025) and massive query/key values (Jin et al., 2025) motivate functional tests. TSFM studies examine representations and interventions (Wiliński et al., 2024) or concept structure (Pandey et al., 2025); our MI, probe, and donor tests connect selected states to future content and forecast response.

Time-series foundation models. Patch, variate-token, mixer, and linear forecasting designs (Nie et al., 2022; Liu et al., 2024a; Wang et al., 2024; Zeng et al., 2023) sit alongside pretrained TSFMs such as TIMER and MOMENT (Liu et al., 2024b; Goswami et al., 2024). Our six checkpoints cover encoder-only, decoder-style, encoder–decoder, and mixer paths (Ansari et al., 2025; Liu et al., 2025; Khwaja et al., 2026; Das et al., 2023; Ansari et al., 2024; Ekambaram et al., 2024); normalized temporal location permits comparison across their native states.

Time-series retrieval. Retrieval methods select historical candidates by task-specific or stationarity-aware criteria (Han et al., 2025; Ning et al., 2026; Zhou et al., 2026); temporal alignment and learned representations provide related similarity tools (Cuturi & Blondel, 2017; Yue et al., 2022). RAF and TimeRAG retrieve examples for foundation or language models (Tire et al., 2024; Yang et al., 2025); we test MI-guided selection and correction with fixed TS-RAG and retrieval budget.

6 Conclusion

Across 42 frozen-TSFM information maps, Information Anchors locates temporal regions where dependence on a structured future concentrates. Block resampling identifies temporal localization as the reproducible unit; probes and donor replacement show how accessible content and prediction response vary across computation paths. The history comparison links these regions to recurring observed patterns. Reusing them in frozen Chronos-Bolt retrieval, Anchor-RAG improves matched TS-RAG MSE on 7/7 datasets and MAE on 6/7. Paired ablations clarify how MI guidance, boundary alignment, and disagreement gating shape this transfer. These findings establish a practical route from locating future-related information in frozen states to using it as a retrieval signal, while retaining the original forecaster.

AI USE STATEMENT

Generative AI tools assisted with code development, figure preparation, and manuscript editing. The authors take responsibility for the final content.

REPRODUCIBILITY STATEMENT

The anonymous code repository[1] contains the experimental code, registered configurations, and locked downstream protocol. Appendices A–K specify the datasets, chronological splits, preprocessing, checkpoints, estimators, donor rules, ANCHOR-RAG validation, paired inference, and resource costs. GPU-backed reruns require the external frozen checkpoints, benchmark data, and retrieval databases described in the repository.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pp. 531–540. PMLR, 2018.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Jonathan Crabbé and Mihaela Van Der Schaar. Explaining time series predictions with dynamic masks. In *International conference on machine learning*, pp. 2166–2177. PMLR, 2021.

Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series. In *International conference on machine learning*, pp. 894–903. PMLR, 2017.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2023.

Vijay Ekambaram, Arindam Jati, Pankaj Dayama, Sumanta Mukherjee, Nam H Nguyen, Wesley M Gifford, Chandra Reddy, and Jayant Kalagnanam. Tiny time mixers (ttms): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series. *Advances in Neural Information Processing Systems*, 37:74147–74181, 2024.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. *arXiv preprint arXiv:2401.06102*, 2024.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024.

Sungwon Han, Seungeon Lee, Meeyoung Cha, Sercan O Arik, and Jinsung Yoon. Retrieval augmented time series forecasting. *arXiv preprint arXiv:2505.04163*, 2025.

[1] https://anonymous.4open.science/r/Information-Anchors

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, pp. 2733–2743, 2019.

Mingyu Jin, Kai Mei, Wujiang Xu, Mingjie Sun, Ruixiang Tang, Mengnan Du, Zirui Liu, and Yongfeng Zhang. Massive values in self-attention modules are the key to contextual knowledge understanding. *arXiv preprint arXiv:2502.01563*, 2025.

Emaad Khwaja, Chris Lettieri, Gerald Woo, Eden Belouadah, Marc Cenac, Guillaume Jarry, Enguerrand Paquin, Xunyi Zhao, Viktoriya Zhukov, Othmane Abou-Amal, et al. Toto 2.0: Time series forecasting enters the scaling era. *arXiv preprint arXiv:2605.20119*, 2026.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International conference on learning representations*, 2021.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6):066138, 2004.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. Moirai 2.0: When less is more for time series forecasting. *arXiv preprint arXiv:2511.11698*, 2025.

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*, volume 2024, pp. 11116–11140, 2024a.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. *arXiv preprint arXiv:2402.02368*, 2024b.

Zichuan Liu, Yingying Zhang, Tianchun Wang, Zefan Wang, Dongsheng Luo, Mengnan Du, Min Wu, Yi Wang, Chunlin Chen, Lunting Fan, et al. Explaining time series via contrastive and locally sparse perturbations. In *International Conference on Learning Representations*, volume 2024, pp. 43401–43421, 2024c.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Kanghui Ning, Zijie Pan, Yu Liu, Yushan Jiang, James Zhang, Kashif Rasul, Anderson Schneider, Lintao Ma, Yuriy Nevmyvaka, and Dongjin Song. Ts-rag: Retrieval-augmented generation based time series foundation models are stronger zero-shot forecaster. *Advances in Neural Information Processing Systems*, 38:163170–163199, 2026.

Atharva Pandey, Abhilash Neog, and Gautam Jajoo. On the internal semantics of time-series foundation models. *arXiv preprint arXiv:2511.15324*, 2025.

Dong Shu, Xuansheng Wu, Haiyan Zhao, Mengnan Du, and Ninghao Liu. Beyond input activations: Identifying influential latents by gradient sparse autoencoders. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1673–1682, 2025.

Kutay Tire, Ege Onur Taga, Muhammed Emrullah Ildiz, and Samet Oymak. Retrieval augmented time series forecasting. *arXiv preprint arXiv:2411.08249*, 2024.

Sana Tonekaboni, Shalmali Joshi, Kieran Campbell, David K Duvenaud, and Anna Goldenberg. What went wrong and when? instance-wise feature importance for time-series black-box models. *Advances in Neural Information Processing Systems*, 33:799–809, 2020.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 38626–38652, 2024.

Michał Wiliński, Mononito Goswami, Willa Potosnak, Nina Żukowska, and Artur Dubrawski. Exploring representations and interventions in time series foundation models. *arXiv preprint arXiv:2409.12915*, 2024.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022.

Silin Yang, Dong Wang, Haoqi Zheng, and Ruochun Jin. Timerag: Boosting llm time series forecasting via retrieval-augmented generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 8980–8987, 2022.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11106–11115, 2021.

Shiqiao Zhou, Holger Schöner, Zipeng Wu, Edouard Fouché, IAG Wilson, and Shuo Wang. Stationarity-aware retrieval-augmented time series forecasting. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 6875–6886, 2026.

A Data and Channel Panels

Table 3: **Registered multivariate data regimes.** Registered C gives the channel count presented to every model in the common matrix. Native channels give the full dataset width.

Regime	Registered C	Native channels	L	P	Diagnostic target $c^\star$
ETTh1	7	7	512	96	OT
ETTh2	7	7	512	96	OT
ETTm1	7	7	512	96	OT
ETTm2	7	7	512	96	OT
Weather	21	21	384	96	T (degC)
Weather / TTM-R2	21	21	512	96	T (degC)
Electricity	16	321	512	96	channel 0
Traffic	16	862	512	96	channel 0

Why a fixed 16-channel panel? The Electricity and Traffic panels were fixed before atlas construction. Both include channel 0 and span the original column order. These registered multivariate panels provide the same input object to all six architectures. Full 321/862-channel registration provides a complementary native-width analysis, while the fixed panels define the matched six-model regime. Moirai-2 forms a packed sequence proportional to $C \times J$ and exceeds its fixed token limit. Chronos-2 group attention and Toto-2 variate attention also incur high joint-channel costs. Processing channels in independent batches would instead change the native multivariate computation under study. We therefore use two analysis tracks. The common matrix uses fixed 16-channel panels, whereas matched panel/full runs evaluate TimesFM 2.5, Chronos-Bolt, and TTM-R2. Their native context construction scales without joint-token expansion. The full-variable experiment therefore characterizes the scaling boundary alongside the matched-panel analysis.

The Electricity panel is $\{0, 1, 24, 46, 69, 92, 115, 137, 160, 183, 205, 228, 251, 274, 296, 319\}$. The Traffic panel is $\{0, 1, 62, 124, 185, 246, 308, 369, 431, 492, 553, 615, 676, 737, 799, 860\}$. Indices follow the original dataset column order and are not selected using MI, forecast error, or test statistics.

B Models and States

Table 4: **Frozen TSFMs and analyzed history states.** Computation paths, channel handling, and recorded history units in the evaluated configurations.

Model	Compute path	Channel handling	Analyzed history unit
Chronos-2	encoder-only	temporal and grouped cross-series attention	encoder-block history patch
Moirai-2	decoder-only	packed channel tokens with causal temporal attention	post-block packed history patch
Toto-2	decoder-only	alternating causal temporal and full variate attention	decoder-block history patch
TimesFM 2.5	decoder-only	channel-independent shared decoder	causal prefill history patch
Chronos-Bolt	encoder–decoder	channel-independent shared encoder	encoder history patch read by decoder cross-attention
TTM-R2	mixer	channel-independent shared mixer	backbone and forecast-decoder patch stage

Chronos-2. We record history-patch states after each encoder block. Context and future-token embeddings are encoded jointly, allowing intermediate history states to influence forecast tokens through later attention layers. The atlas retains all recorded layers; anchor selection for probing and donor interventions excludes final-layer history states because the forecast head reads final-layer future-token states without a subsequent history-to-future mixing step.

Moirai-2. The frozen backbone is decoder-only. In the evaluated packed multivariate configuration, channel-specific history patches share one token sequence, with a causal temporal mask and a variate-ID attention bias. We record post-block history states and aggregate same-time channel states for analysis.

Toto-2. The decoder alternates causal temporal attention and full variate attention. We record the history-patch states after each registered decoder block and align channel states by temporal patch for analysis.

TimesFM 2.5. The shared decoder processes each channel independently. We record history-patch states from the causal prefill pass and concatenate same-time channel states only for the analysis representation.

Chronos-Bolt. Each channel is patched and encoded independently by the shared encoder. We record history-patch states after each encoder block; the decoder reads the final encoder states through cross-attention, giving the recorded history states a downstream path to the forecast.

TTM-R2. The evaluated common-channel configuration applies shared mixer parameters independently to each channel. We record backbone and forecast-decoder states on the common temporal patch grid and group same-time channel states only for the analysis representation.

C Analysis Protocol

C.1 Chronological partitions and normalization

Every configuration records train, validation, discovery, and test boundaries. It also records an origin hash, checkpoint revision, adapter hook, cache signature, deterministic seed, and no-op intervention check. Forecast error uses TSLib train-fitted standardization, distinct from per-instance reversible normalization (Kim et al., 2021). The MI target first undergoes robust history-window normalization, followed by discovery-fitted feature standardization and PCA. Probe targets use the same future window and their own discovery-fitted target scaling. No test statistic selects an atlas cell, layer, probe regularization parameter, donor, or random order.

The discovery segment also supplies probe-training origins. For a boundary b, the final origin in the preceding segment is at most $b - P$, and the first origin in the following segment is $b + L$; hence the two complete history–future windows share no observations. This is a strict window-disjoint guard, not an additional empty $L + P$ embargo beyond the boundary. Origins within a split may overlap because the registered stride is four; moving-block resampling uses blocks based on $L + P$ to account for that dependence. Table 5 gives every raw boundary used by the 42 configurations.

Table 5: **Chronological split ledger.** Discovery origins are $[L, d - P]$, validation origins are $[d + L, v - P]$, and test origins are $[v + L, t - P]$, sampled at stride four before any origin cap. Brackets denote allowable inclusive bounds; the realized origins lie on the stride-four grid. Weather uses $L = 384$ except for TTM-R2, which uses $L = 512$; all regimes use $P = 96$.

Regime	Discovery end d	Validation end v	Test end t	L
ETTh1 / ETTh2	8,640	11,520	14,400	512
ETTm1 / ETTm2	34,560	46,080	57,600	512
Weather	36,887	42,157	52,696	384 / 512
Electricity	18,412	21,042	26,304	512
Traffic	12,280	14,034	17,544	512

C.2 Future statistics

The 26 per-channel MI features are 16 equal relative-time bin means, future mean, future standard deviation, least-squares slope, final value, total change, first-difference standard deviation, and four log spectral-band energies. The 12 interpretable probe targets are four future-quarter levels,

total change, linear trend, mean absolute first difference, first-difference standard deviation, second-difference roughness, and low/mid/high normalized spectral energy. The global probe uses the channel macro mean of each property; the target probe uses the preregistered diagnostic channel. These two scopes do not alter the multivariate forecasting task or the MI target.

C.3 Atlas construction and calibration

For each model–regime pair:

1. run the frozen model on 1,024 discovery origins and cache every registered native history unit;
2. fit one eight-dimensional hidden PCA per layer on all discovery origins and temporal patches jointly, reuse that layer-shared projection at every patch, and fit one separate eight-dimensional PCA view for the structured all-variable future;
3. estimate KSG MI using $k = 5$, float64 Chebyshev distances, and deterministic 10^{-8} jitter;
4. generate 199 legal temporal circular shifts, estimate a cellwise null, and retain raw MI, z, empirical p, and BH-adjusted q;
5. select the maximum-mean reachable layer and its High-MI/Low-MI patches using discovery scores; and
6. export normalized temporal and layer profiles together with the registered descriptors.

C.4 Target-aligned probes

For each selected High-MI/Low-MI unit, hidden PCA, target standardization, and ridge weights are fitted on the chronological probe-training partition. Validation chooses α from a fixed grid and test R^2 is computed once. We retain negative values because they distinguish a difficult property from an omitted result. The primary comparison is High-MI versus same-layer Low-MI, not whether every absolute R^2 is positive.

C.5 Donor interventions and statistics

Formal interventions use 128 test origins and eight donor shifts. Donors are observed origins separated from each recipient by at least $L + P$. Their states replace recipient states only at the same layer and patch coordinate. High-MI, Low-MI, and five random patch orders are nested across fractions and share identical donors. A no-op hook must reproduce the unperturbed prediction before a run is accepted. The hierarchical bootstrap resamples donors, then origins within each donor. The reanalysis also uses moving blocks over ordered origins and Benjamini–Hochberg correction over each complete 168-condition contrast family. Sign-flip tests operate at donor level. The complement experiment keeps S and replaces its complement. The resulting response is an operational sufficiency diagnostic under the donor distribution, with the estimand defined by the matched replacement protocol.

D Information Maps

Figure 7 shows all 42 native layer–patch atlases in one common plate. Each cell is percentile-normalized only within its own atlas; native patch and layer grids are not resampled. The plate deliberately omits candidate markers so the full topology remains visible. It supports within-run spatial comparison and qualitative shape comparison, not cross-run raw-MI ranking.

Table 7: **All 42 within-atlas topology descriptors.** $Q_{.25}$: top-quarter temporal concentration; A: anchor ratio; N_{eff}: effective temporal support; d_b: normalized boundary distance; D: expected normalized layer depth; H_l: normalized layer entropy. All quantities are normalized within each atlas; raw MI is never compared across models.

Model	Regime	$Q_{.25}$	A	N_{eff}	d_b	D	H_l
Chronos-2	ETTh1	.420	1.964	28.4	.410	.459	.987
	ETTh2	.375	1.932	29.9	.467	.560	.963
	ETTm1	.417	1.969	28.6	.414	.540	.992
	ETTm2	.392	1.958	29.5	.447	.594	.976

Model	Regime	$Q_{.25}$	A	N_{eff}	d_b	D	H_l
	Weather	.393	1.882	21.9	.432	.600	.971
	Electricity	.405	1.745	27.6	.479	.504	.999
	Traffic	.408	1.844	27.9	.490	.508	.997
Moirai-2	ETTh1	.428	1.958	27.9	.418	.421	.973
	ETTh2	.417	1.969	28.6	.391	.504	.986
	ETTm1	.422	1.969	29.2	.396	.479	.997
	ETTm2	.427	1.969	28.3	.411	.570	.976
	Weather	.436	1.958	21.4	.384	.532	.995
	Electricity	.422	1.844	27.6	.458	.471	.995
	Traffic	.370	1.792	29.9	.497	.443	.985
Toto-2	ETTh1	.435	1.938	14.1	.374	.478	.973
	ETTh2	.431	1.938	14.3	.385	.456	.987
	ETTm1	.436	1.938	14.0	.376	.506	.998
	ETTm2	.428	1.938	13.8	.379	.488	.992
	Weather	.438	1.917	10.1	.341	.508	.982
	Electricity	.366	1.562	15.1	.482	.493	.993
	Traffic	.367	1.708	14.9	.494	.520	.987
TimesFM 2.5	ETTh1	.416	1.938	14.1	.377	.499	.993
	ETTh2	.423	1.938	14.3	.405	.489	.988
	ETTm1	.437	1.938	13.8	.365	.497	.998
	ETTm2	.432	1.938	13.8	.370	.498	.997
	Weather	.438	1.917	10.1	.345	.497	.996
	Electricity	.375	1.631	15.0	.447	.544	.962
	Traffic	.395	1.738	14.5	.517	.484	.993
Chronos-Bolt	ETTh1	.433	1.969	27.6	.384	.491	.998
	ETTh2	.385	1.964	29.7	.476	.494	.989
	ETTm1	.424	1.969	28.7	.418	.517	.998
	ETTm2	.424	1.969	28.4	.453	.546	.991
	Weather	.396	1.951	21.9	.443	.524	.978
	Electricity	.410	1.750	27.6	.470	.478	.997
	Traffic	.400	1.917	28.2	.470	.480	.997
TTM-R2	ETTh1	.394	1.875	7.0	.435	.588	.951
	ETTh2	.419	1.825	7.2	.441	.615	.929
	ETTm1	.400	1.875	7.4	.446	.614	.937
	ETTm2	.388	1.875	7.1	.418	.609	.935
	Weather	.419	1.875	7.3	.423	.594	.960
	Electricity	.425	1.775	7.3	.388	.540	.961
	Traffic	.400	1.625	7.0	.449	.539	.988

E PROBE RESULTS

Eleven properties characterize components of the structured MI target; they are not independent discovery outcomes. Second-difference roughness is excluded from that target and supplies the following held-out audit.

For completeness, the aggregate target-aligned accessibility results are given below. Values average the six models and seven regimes; negative chronological R^2 values are retained.

Figure 8 reports the High-minus-Low difference for every model, regime, property, and scope in a single plate. The complete selected-unit table below retains absolute High- and Low-MI R^2, negative values, selected coordinates, and MI–probe alignment. Post-hoc best-cell values remain tabulated as diagnostic oracles and do not support primary claims.

Table 10: **Selected-unit probe audit for both target scopes.** Negative held-out R^2 values are retained. Selected coordinates are shared across Global and Target scopes.

		Selected		Global				Target			
Model	Regime	$l^\star$	j^+	High-MI	Low-MI	ΔR^2	ρ	High-MI	Low-MI	ΔR^2	ρ
Chronos-2	ETTh1	7	31	0.020	-0.073	0.094	0.019	-4.806	-5.038	0.232	0.042
	ETTh2	10	13	0.038	0.029	0.009	0.420	-0.114	-0.137	0.024	0.444
	ETTm1	10	31	0.153	0.106	0.047	0.277	-1.546	-1.401	-0.145	0.378
	ETTm2	10	19	0.107	0.070	0.036	0.751	-0.062	-0.071	0.009	-0.038
	Weather	10	6	-0.219	-0.175	-0.044	0.473	0.136	0.151	-0.016	0.489
	Electricity	3	24	-0.010	-0.000	-0.010	0.055	-0.078	-0.094	0.017	0.086
	Traffic	10	5	0.209	0.193	0.016	0.103	0.085	0.119	-0.035	0.072
Moirai-2	ETTh1	0	31	0.005	-0.034	0.038	0.464	-4.822	-4.896	0.074	0.495
	ETTh2	2	31	0.044	-0.163	0.207	0.390	-0.119	-0.289	0.169	0.139
	ETTm1	1	31	0.142	0.091	0.051	0.301	-1.345	-1.636	0.290	0.484
	ETTm2	4	31	0.130	-0.024	0.154	0.596	-0.709	-0.595	-0.114	-0.353
	Weather	5	23	-0.365	-0.521	0.156	0.550	0.052	-0.253	0.305	0.564
	Electricity	0	25	-0.021	-0.047	0.027	0.021	-0.049	-0.073	0.024	0.108
	Traffic	1	22	0.167	0.135	0.032	-0.421	0.076	0.066	0.009	-0.342
Toto-2	ETTh1	1	15	0.011	-0.027	0.037	0.232	-5.068	-4.992	-0.076	-0.381
	ETTh2	0	15	-0.006	-0.070	0.064	0.217	-0.178	-0.182	0.004	-0.016
	ETTm1	5	15	0.180	0.091	0.088	0.442	-1.447	-1.608	0.161	0.619
	ETTm2	4	15	0.124	-0.042	0.166	0.731	-0.331	-0.046	-0.285	-0.470
	Weather	4	11	-0.066	-0.474	0.408	0.466	0.219	-0.219	0.438	0.631

		Selected		Global				Target			
Model	Regime	$l^\star$	j^+	High-MI	Low-MI	ΔR^2	ρ	High-MI	Low-MI	ΔR^2	ρ
	Electricity	0	15	0.158	-0.000	0.158	0.112	0.067	-0.068	0.135	0.172
	Traffic	5	15	0.174	0.083	0.091	-0.251	0.094	0.038	0.056	-0.257
TimesFM 2.5	ETTh1	19	15	0.038	-0.030	0.069	0.265	-4.640	-5.290	0.650	0.534
	ETTh2	1	15	-0.006	-0.146	0.139	0.464	0.011	-0.180	0.190	0.470
	ETTm1	19	15	0.199	0.080	0.119	0.475	-1.610	-1.737	0.127	0.234
	ETTm2	4	15	0.126	-0.039	0.165	0.515	-0.043	-0.135	0.092	-0.132
	Weather	9	11	-0.030	-0.432	0.402	0.746	0.210	-0.225	0.434	0.824
	Electricity	18	15	0.154	-0.047	0.202	0.143	0.075	-0.102	0.176	0.143
	Traffic	1	11	0.161	0.104	0.057	-0.340	0.089	0.059	0.030	-0.362
Chronos-Bolt	ETTh1	0	31	-0.010	0.002	-0.012	0.100	-4.543	-4.605	0.062	0.058
	ETTh2	2	31	-0.014	0.017	-0.032	0.149	0.012	-0.127	0.139	0.241
	ETTm1	9	31	0.173	0.132	0.041	0.368	-1.385	-1.381	-0.004	0.058
	ETTm2	9	31	0.152	0.120	0.032	0.432	-0.259	-0.264	0.005	-0.120
	Weather	3	23	-0.369	-0.365	-0.003	0.371	0.047	0.016	0.031	0.379
	Electricity	5	30	0.141	0.072	0.068	0.081	0.089	-0.038	0.127	0.176
	Traffic	6	22	0.248	0.229	0.019	-0.122	0.143	0.097	0.046	-0.035
TTM-R2	ETTh1	3	7	-0.035	-0.042	0.007	0.508	-4.511	-4.755	0.243	0.780
	ETTh2	3	7	-0.014	-0.072	0.058	0.623	-0.234	-0.203	-0.031	0.420
	ETTm1	3	7	0.153	0.134	0.019	0.800	-1.399	-1.672	0.273	0.859
	ETTm2	3	7	0.139	0.055	0.083	0.789	-0.137	-0.215	0.078	0.798
	Weather	4	7	-0.269	-0.349	0.080	0.680	0.034	-0.021	0.056	0.716
	Electricity	3	7	0.146	0.001	0.145	0.611	0.091	-0.031	0.122	0.562
	Traffic	4	7	0.244	0.213	0.031	0.261	0.167	0.126	0.041	0.530

F INTERVENTION RESULTS

Figure 9 condenses all curve-level ground-truth effects. Remove panels replace the selected set; Keep panels preserve it and replace the complement. High-minus-control effects are oriented so positive values favor High-MI. The plate is a compact index of the complete numerical tables below; donor replacement remains operational and does not identify a mediator.

Table 13: **All-variable forecast drift at 25% replacement for all 42 model–regime pairs.** Δ_{low} and Δ_{rand} denote High-MI minus Low-MI and mean Random drift, respectively.

			vs. Low-MI		vs. Random	
Model	Regime	High-MI	Δ_{low}	p	Δ_{rand}	p
Chronos-2	ETTh1	.103	.094	.0039	.084	.0039
	ETTh2	.010	.009	.0039	.007	.0039
	ETTm1	.006	.004	.0039	.003	.0039
	ETTm2	.003	.000	.0078	.001	.0039
	Weather	.002	-.003	1.000	-.002	1.000
	Electricity	.038	.023	.0039	.011	.0039
	Traffic	.003	-.003	1.000	-.000	1.000
Moirai-2	ETTh1	.178	.169	.0039	.137	.0039
	ETTh2	.326	.324	.0039	.264	.0039
	ETTm1	.244	.230	.0039	.180	.0039
	ETTm2	.183	.182	.0039	.148	.0039
	Weather	.223	.223	.0039	.179	.0039
	Electricity	.653	.649	.0039	.575	.0039
	Traffic	.069	.056	.0039	-.091	1.000
Toto-2	ETTh1	.494	.461	.0039	.373	.0039
	ETTh2	.355	.342	.0039	.265	.0039
	ETTm1	.330	.330	.0039	.264	.0039
	ETTm2	.121	.121	.0039	.097	.0039
	Weather	.173	.173	.0039	.138	.0039
	Electricity	.378	.334	.0039	.247	.0039
	Traffic	.623	.623	.0039	.499	.0039
TimesFM 2.5	ETTh1	.826	.826	.0039	.661	.0039
	ETTh2	.330	.301	.0039	.221	.0039
	ETTm1	.959	.959	.0039	.767	.0039
	ETTm2	.256	.248	.0039	.193	.0039
	Weather	.330	.327	.0039	.257	.0039
	Electricity	1.289	1.288	.0039	1.031	.0039
	Traffic	1.049	1.003	.0039	.746	.0039
Chronos-Bolt	ETTh1	.782	.763	.0039	.688	.0039
	ETTh2	.207	.187	.0039	.156	.0039

			vs. Low-MI		vs. Random	
Model	Regime	High-MI	Δ_{low}	p	Δ_{rand}	p
	ETTm1	.549	.536	.0039	.515	.0039
	ETTm2	.101	.092	.0039	.082	.0039
	Weather	.167	.141	.0039	.112	.0039
	Electricity	.652	.643	.0039	.624	.0039
	Traffic	.163	.123	.0039	.098	.0039
	ETTh1	.041	-.006	.9883	-.008	1.000
	ETTh2	.050	.044	.0039	.033	.0039
	ETTm1	.092	.016	.0039	.044	.0039
TTM-R2	ETTm2	.041	.033	.0039	.030	.0039
	Weather	.059	.055	.0039	.045	.0039
	Electricity	.062	-.040	1.000	-.015	1.000
	Traffic	.084	-.025	1.000	-.023	1.000

F.1 Forecast error and donor-replacement effects are different metrics

Baseline error compares the unperturbed forecast with ground truth. Forecast drift compares a donor-replaced forecast with the same unperturbed forecast. Signed ground-truth delta-MSE is the primary prediction-response endpoint. Forecast drift is the complementary sensitivity endpoint. Ground-truth delta-MSE evaluates the change in predictive correctness. Forecast drift can be large when donor replacement perturbs a forecast that was already inaccurate.

Table 15: **TSLib train-standardized unperturbed forecast-error audit for all 42 model–regime pairs.** Values are reported for the preregistered diagnostic target and for the macro-average over all registered variables.

		Target		All-variable	
Model	Regime	MSE	MAE	MSE	MAE
	ETTh1	.059	.184	.378	.380
	ETTh2	.147	.296	.323	.352
	ETTm1	.030	.128	.311	.325
Chronos-2	ETTm2	.086	.206	.184	.248
	Weather	.073	.194	.216	.228
	Electricity	.414	.275	.117	.198
	Traffic	.410	.207	.458	.280
	ETTh1	.064	.189	.492	.466
	ETTh2	.152	.305	.355	.389
	ETTm1	.037	.145	.631	.499
Moirai-2	ETTm2	.148	.278	.239	.309
	Weather	.134	.250	.263	.289
	Electricity	.729	.546	.413	.416
	Traffic	.615	.448	.693	.472
	ETTh1	.057	.182	.392	.385
	ETTh2	.139	.291	.335	.357
	ETTm1	.029	.126	.322	.332
Toto-2	ETTm2	.080	.200	.195	.256
	Weather	.117	.236	.224	.243
	Electricity	.404	.269	.123	.206
	Traffic	.523	.265	.507	.320
	ETTh1	.058	.180	.413	.394
	ETTh2	.136	.286	.379	.382
	ETTm1	.030	.128	.328	.336
TimesFM 2.5	ETTm2	.087	.210	.208	.265
	Weather	.115	.228	.223	.246
	Electricity	.482	.312	.126	.208
	Traffic	.430	.239	.472	.302
	ETTh1	.069	.198	.394	.387
	ETTh2	.116	.260	.333	.358
	ETTm1	.033	.132	.334	.335
Chronos-Bolt	ETTm2	.088	.212	.188	.254
	Weather	.124	.245	.247	.269
	Electricity	.415	.278	.115	.197
	Traffic	.580	.273	.502	.307
	ETTh1	.052	.172	.382	.400
	ETTh2	.138	.293	.315	.369
	ETTm1	.030	.132	.333	.352
TTM-R2	ETTm2	.081	.212	.190	.268
	Weather	.080	.200	.197	.243
	Electricity	.370	.309	.145	.246

		Target		All-variable	
Model	Regime	MSE	MAE	MSE	MAE
	Traffic	.721	.395	.579	.406

G Intervention Controls

The curve-level, quantity-dose, and magnitude-dose summaries are consolidated in Figure 5; the tables and panels below retain the model-resolved values and additional perturbation controls without duplicating that main-text figure.

This table and the main text use the same signed ground-truth ΔMSE endpoint and the same curve unit: the mean over four registered budgets. The previously reported forecast-drift summary is retained as a complementary endpoint but is not mixed into this table.

Matching reduces history-trajectory and baseline-forecast distances to 0.868 and 0.854 of their unmatched means. Exact seasonal phase is achieved for 24.2% of retained pairs. The saved perturbation RMS is the donor–recipient displacement used by the response estimand.

H History Patterns

Motifs are deterministic labels computed from each observed history patch. They are not probe targets, and they do not participate in atlas selection. The cross-fitted critic selects the top 25% pointwise-MI patches per origin. The confirmatory audit uses runs with held-out critic accuracy at least 0.60; six ETTh2 runs fall outside this prespecified validity region and are retained as quality-control entries. All reported effect estimates pair selections within chronological origins and use the same 25% budget.

H.1 Pointwise protocol

Population atlas cells do not resolve individual origins. At frozen $l^\star$, a cross-fitted density-ratio critic scores every origin–patch pair. We assign deterministic history-only motifs covering trends, turns, extrema, level changes, smoothness, and cross-channel dynamics. Within each origin, the top and bottom quartiles of critic scores define equal-budget High-MI and Low-MI patch sets. Their paired motif-rate difference quantifies target-aligned co-occurrence in observed history.

The critic is fully specified as follows. Five contiguous folds are used, and a separate linear critic is fitted for each native patch. Positive examples are joint projected pairs (z, y); negative examples pair the same z with a fixed-point-free permutation of y. The feature map is $[z, y, \mathrm{vec}(zy^\top)]$, standardized on the training fold, followed by balanced logistic regression (L-BFGS, $C = 1$, 2,000 iterations, seed 2021). On the held-out fold, the positive logit is centered and scaled by logits from eight independent target derangements. These calibrated scores approximate a pointwise log density ratio up to critic error and are used only for within-origin patch ranking; they are not single-sample KSG MI estimates. Critic accuracy averages joint-versus-product classification over held-out folds and patches; the run-level threshold is 0.60.

Motifs are assigned from observed history only, while the future enters through the pointwise ranker. Equal patch budgets prevent either set from receiving more motif labels by construction. The audit therefore quantifies whether High-MI patches co-occur with recognizable history structure more often than Low-MI patches from the same origin.

We report the mean, central 95% spread, and positive-run count across the 36 critic-valid atlases to summarize heterogeneity across registered runs. Because motif labels and pointwise ranks share the same observed origins, these statistics provide a target-aligned co-occurrence audit that complements the held-out probe and intervention analyses.

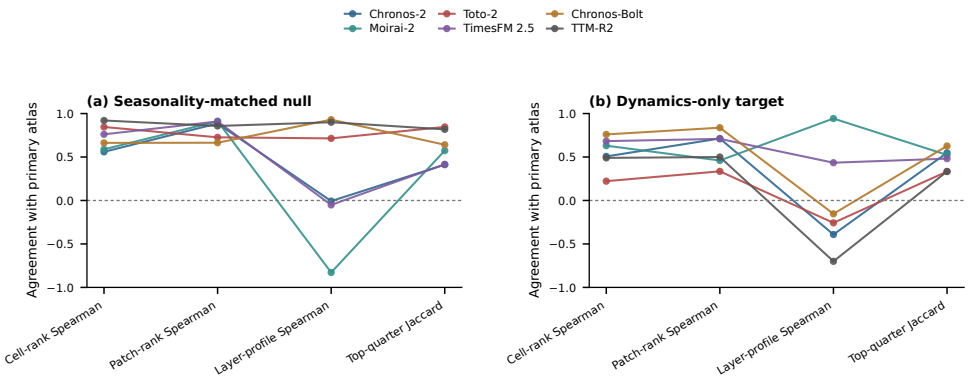

Figure 6: **Model-resolved target and null sensitivities.** Each line is one checkpoint. **(a)** Seasonality-matched temporal null. **(b)** Dynamics-only future target. Spearman metrics use their complete $[-1, 1]$ range; negative layer-profile agreement is valid and retained. Jaccard values occupy $[0, 1]$. All metrics compare with the primary ETTh1 atlas.

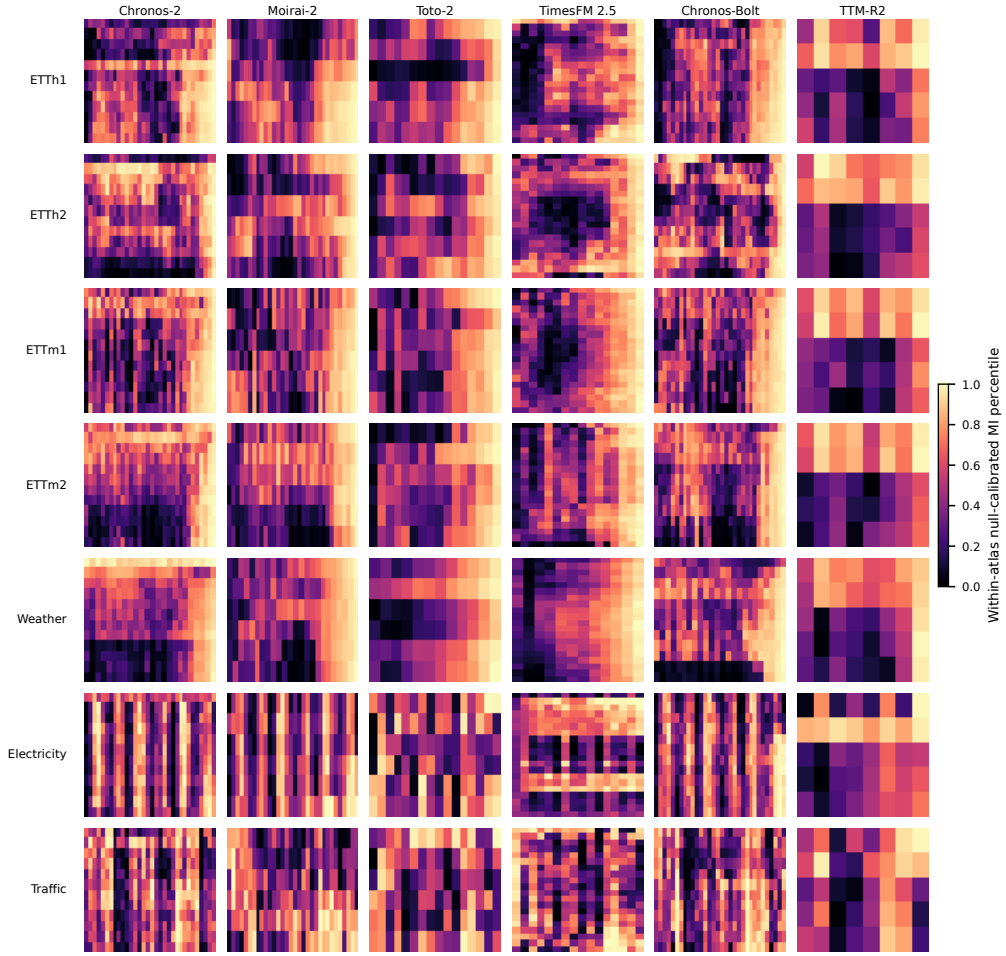

Figure 7: **Complete 6-model by 7-regime information atlas.** Columns are frozen models and rows are registered regimes. Color gives the within-atlas null-calibrated MI percentile; no candidate symbols are overlaid. Electricity and Traffic use the registered 16-channel panels.

Table 6: **Cross-dataset topology and target-aligned accessibility.** $Q_{0.25}$ is top-quarter temporal concentration, d_b is mean normalized boundary distance, and layer depth is expected normalized depth. ΔR^2 averages global and target High-MI-minus-Low-MI probe differences.

Model	$Q_{0.25}$	Anchor ratio	d_b	Layer depth	$\Delta R^2_{\text{probe}}$	$P(\Delta R^2 > 0)$
Chronos-2	0.401	1.899	0.449	0.538	0.017	0.643
Moirai-2	0.417	1.923	0.422	0.488	0.102	0.929
Toto-2	0.414	1.848	0.405	0.493	0.103	0.857
TimesFM 2.5	0.417	1.862	0.404	0.501	0.204	1.000
Chronos-Bolt	0.410	1.927	0.445	0.504	0.037	0.714
TTM-R2	0.406	1.818	0.429	0.586	0.086	0.929

Table 8: **Held-out target-aligned accessibility by scope.** High-MI and Low-MI values are reported together with their paired differences for the global and diagnostic-target scopes.

	Global			Target		
Semantic	High	Low	Δ	High	Low	Δ
Level	.272	.130	.143	.350	.117	.233
Trajectory	.259	.188	.071	.173	.103	.070
Local dynamics	$-.239$	$-.317$	.078	-4.060	-4.085	.025
Frequency	$-.078$	$-.096$	.019	$-.116$	$-.131$	.016

Table 9: **Target-independent roughness accessibility.** Positive counts refer to paired High-MI minus Low-MI held-out R^2 over 42 model–regime atlases.

Scope	High-MI R^2	Low-MI R^2	Difference	Positive
Global	-0.228	-0.287	0.058	25/42
Target	-3.965	-3.787	-0.178	24/42

H.2 History-structure summary

Thirty-six of 42 model–regime combinations satisfy the preregistered critic gate; the six ETTh2 runs outside the gate are shown as quality-control entries. Pointwise ranking and enrichment estimates use the 36 valid combinations. Relative to equal-budget Low-MI patches (Figure 11), the largest mean differences are cross-channel volatility ($+5.76$ points; 32/36 positive), rising trend ($+4.62$; 28/36), turning down ($+2.91$; 27/36), local peak ($+2.52$; 28/36), and volatile burst ($+2.41$; 22/36). Smooth segments are depleted (-4.42; 4/36 positive).

The complete eleven-label vocabulary is reported in the accompanying tables and figures, keeping both positive and contrasting co-occurrence patterns visible. Together, these descriptors form a history-structure map rather than a fixed dictionary of latent concepts.

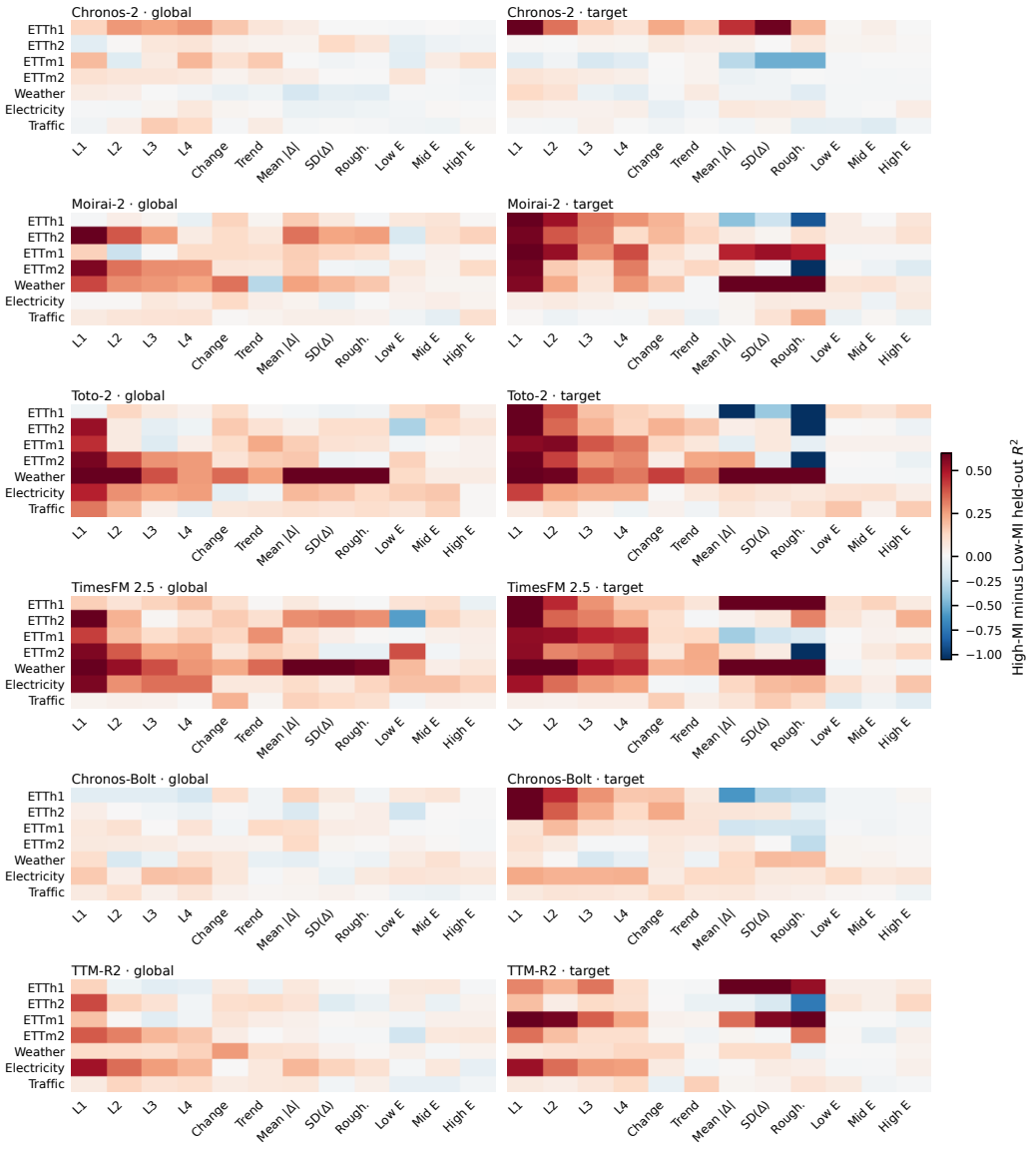

Figure 8: **Complete dataset-resolved target-aligned accessibility.** Rows pair global and target scopes for each model. Every heatmap contains seven regimes by twelve future properties. The shared diverging scale reports held-out High-MI minus Low-MI R^2 without clipping negative effects.

Table 11: **Positive and significant removal contrasts by patch fraction.** Each denominator is 42 model–regime pairs.

Control	12.5%	25%	37.5%	50%
Low-MI	39/42	37/42	37/42	39/42
Random mean	33/42	36/42	38/42	38/42



Figure 9: **Complete model–regime prediction-response comparison.** Rows separate Remove and Keep; columns compare High-MI with the registered Low-MI and Random controls. Each cell averages the four patch budgets and reports oriented all-variable ground-truth ΔMSE.

Table 12: **Prediction-response curve effects for all 42 model–regime pairs.** Each entry averages the four removal fractions. Positive values indicate a larger all-variable ground-truth delta-MSE under High-MI replacement than under the named control.

Control	Model	ETTh1	ETTh2	ETTm1	ETTm2	Weather	Electricity	Traffic
Low-MI	Chronos-2	0.027	−0.000	0.001	0.001	0.004	0.042	0.004
	Moirai-2	0.223	0.239	0.117	0.145	0.193	0.538	0.591
	Toto-2	0.362	0.252	0.311	0.109	0.154	0.358	0.629
	TimesFM 2.5	0.688	0.219	0.822	0.203	0.265	1.287	0.863
	Chronos-Bolt	0.568	0.138	0.391	0.078	0.103	0.764	0.243
	TTM-R2	0.014	0.050	0.013	0.045	0.069	−0.043	−0.077
Random	Chronos-2	0.028	−0.001	0.001	0.000	0.002	0.013	0.002
	Moirai-2	0.144	0.165	0.076	0.098	0.135	0.394	0.302
	Toto-2	0.269	0.173	0.202	0.071	0.108	0.201	0.409
	TimesFM 2.5	0.447	0.144	0.534	0.134	0.188	0.851	0.493
	Chronos-Bolt	0.473	0.098	0.335	0.058	0.075	0.622	0.074
	TTM-R2	0.013	0.037	0.024	0.033	0.048	−0.004	−0.049

Table 14: **Train-standardized unperturbed forecast error.** Rows are macro means over all seven registered regimes and provide metric-scale context for the registered comparisons.

Model	Regimes	Target MSE	Target MAE	All-variable MSE	All-variable MAE
Chronos-2	7	0.174	0.213	0.284	0.287
Moirai-2	7	0.269	0.309	0.441	0.406
Toto-2	7	0.193	0.224	0.300	0.300
TimesFM 2.5	7	0.191	0.226	0.307	0.305
Chronos-Bolt	7	0.203	0.228	0.302	0.301
TTM-R2	7	0.210	0.245	0.306	0.326

Table 16: **Hierarchical prediction-response effects.** Each unit is one model–regime curve averaged over four patch fractions. CIs resample both models and regimes.

Contrast	Positive curves	Mean effect	95% CI
Remove: high − low	39/42	0.262	[0.094, 0.488]
Remove: high − random	39/42	0.177	[0.067, 0.329]
Keep: high − low	39/42	0.305	[0.111, 0.563]
Keep: high − random	39/42	0.178	[0.065, 0.329]

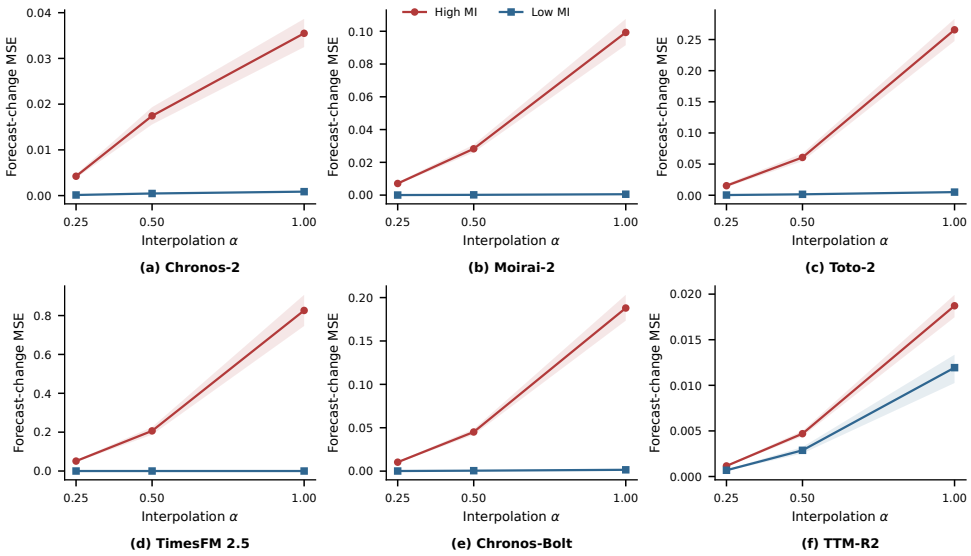

Figure 10: **Complete ETTh1 donor-interpolation dose response.** Curves show absolute all-variable forecast drift with paired hierarchical-bootstrap 95% intervals. Panel scales are model specific, so cross-model magnitudes are not compared.

Table 17: **High-minus-low donor interpolation effects on ETTh1.** Every model-specific 95% CI excludes zero in the high-MI direction at all three doses.

α	Mean all effect	Median all effect	All wins	Mean target effect	Target wins	Propagation-gain wins
0.25	0.01466	0.00861	6/6	0.00268	6/6	6/6
0.50	0.05955	0.03636	6/6	0.01120	6/6	6/6
1.00	0.23560	0.14260	6/6	0.04141	6/6	6/6

Table 18: **Outcome-blind matched-donor prediction response.** Within each donor shift and fraction, the audit retains the nearest 25% of recipients under seasonal phase, history trajectory, baseline forecast, perturbation magnitude, and high/low perturbation balance. Effects are high-minus-low signed ground-truth delta-MSE values, averaged over four fractions.

Dataset	Positive curves	Mean effect	Minimum	Maximum
ETTh1	6/6	0.210	0.025	0.448
ETTm1	6/6	0.188	0.005	0.583
Weather	6/6	0.098	0.002	0.208
All	18/18	0.165	two-way 95% CI [0.056, 0.307]	

Table 19: **Exact deterministic history-structure rules.** All quantiles are computed within each origin across its native patches. μ_j, b_j, and v_j denote target-channel patch mean, least-squares slope on $[-1, 1]$, and standard deviation of first differences; $b_{j,c}$ and $v_{j,c}$ are their channel-wise counterparts.

Family	Label	Boolean rule at patch j
Trend	Rising / falling	$b_j \geq Q_{.75}(b)$ / $b_j \leq Q_{.25}(b)$
Local variation	Volatile / smooth	$v_j \geq Q_{.75}(v)$ / $v_j \leq Q_{.25}(v)$
Turn	Turning up / down	$(b_j < 0, b_{j+1} > 0)$ / $(b_j > 0, b_{j+1} < 0)$
Level	Level jump	$j > 1$ and $\lvert\mu_j - \mu_{j-1}\rvert \geq Q_{.75}(\lvert\Delta\mu\rvert) + 10^{-8}$
Extremum	Local peak	$\mu_j \geq \mu_{j-1},\ \mu_j \geq \mu_{j+1},\ \mu_j > Q_{.75}(\mu)$
Extremum	Local trough	$\mu_j \leq \mu_{j-1},\ \mu_j \leq \mu_{j+1},\ \mu_j < Q_{.25}(\mu)$
Cross-channel	Slope synchrony	$\lvert C^{-1} \sum_c \mathrm{sign}(b_{j,c})\rvert \geq 0.75$
Cross-channel	Volatility	$\mathrm{median}_c(v_{j,c}) \geq Q_{.75}(\mathrm{median}_c\, v_{\cdot,c})$

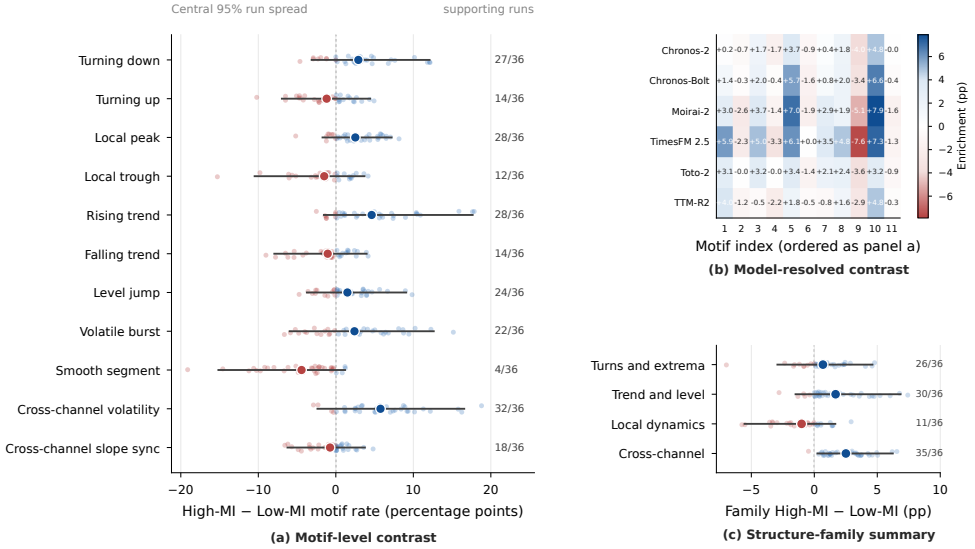

Figure 11: **High-MI–Low-MI contrasts across the complete 11-motif vocabulary. (a)** Per-run paired motif-rate differences for 36 critic-valid model–regime atlases; bars span the central 95% run range, large markers are means, and right labels give positive-run counts. **(b)** Model-resolved mean contrasts, with motifs ordered as in (a). **(c)** Family-level summaries. Motifs use history only; the future is used only by the pointwise MI ranker.

Table 20: **High-MI–Low-MI motif contrasts over 36 critic-valid atlases.** Rates are paired within origin and reported in percentage points. The central 95% run spread summarizes heterogeneity across the registered atlases.

Motif	Mean (pp)	2.5th pct.	97.5th pct.	Positive runs
Cross-channel volatility	+5.76	-2.37	+16.57	32/36
Rising trend	+4.62	-1.53	+17.66	28/36
Turning down	+2.91	-3.12	+12.09	27/36
Local peak	+2.52	-1.72	+7.19	28/36
Volatile burst	+2.41	-5.96	+12.63	22/36
Level jump	+1.50	-3.73	+9.08	24/36
Cross-channel slope sync	-0.75	-6.24	+3.75	18/36
Falling trend	-1.05	-7.93	+3.97	14/36
Turning up	-1.18	-6.96	+4.42	14/36
Local trough	-1.50	-10.48	+3.70	12/36
Smooth segment	-4.42	-15.15	+1.19	4/36

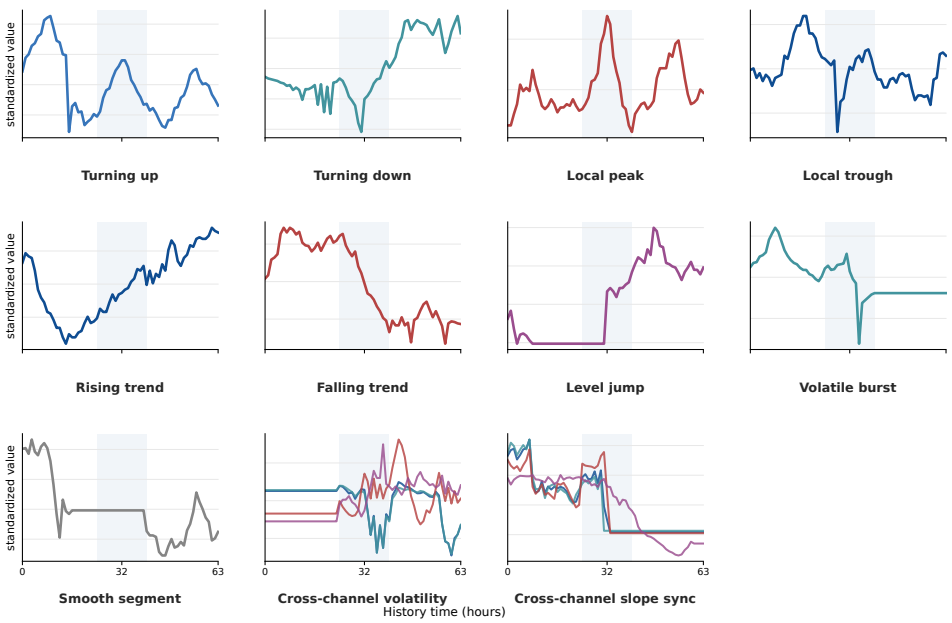

Figure 12: **Observed examples of the complete history-structure vocabulary.** All 11 registered motifs are illustrated with observed 64-hour ETTh1 histories. Single-series panels show the OT target; cross-channel panels show four individually standardized channels. The pale band marks the central 16-step local structure. These examples visualize deterministic history-only labels and do not enter MI selection or enrichment estimation.

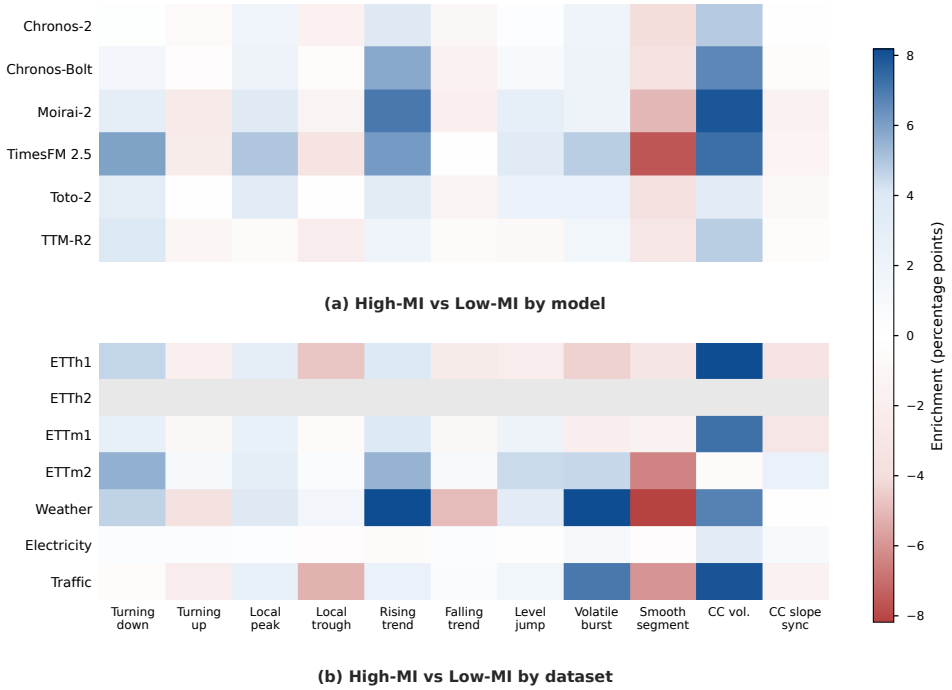

Figure 13: **Complete High-MI–Low-MI history-structure atlas.** Each cell reports a mean paired motif-rate difference in percentage points. **(a)** Model-resolved contrasts averaged over critic-valid datasets. **(b)** Dataset-resolved contrasts averaged over critic-valid models; ETTh2 is grey because all six critics fall below the preregistered gate. The vocabulary is descriptive and history-only.

I ANCHOR-RAG PROTOCOL

History-only MI sidecar and inference-time inputs. Within each domain, cross-fitted pointwise MI labels from discovery history–future pairs train per-patch ridge models on hidden history tokens. The hidden PCA, MI critic, standardizers, and ridge models are frozen before validation and test. At inference, Eq. 5 uses only the observed histories of the query and pool candidates. Stored candidate continuations are loaded after MI Top-10 selection; the query future is used only to score the forecast.

Matched TS-RAG protocol. All runs share the frozen TS-RAG/Chronos-Bolt model, $L = 512$, $P = 64$, the official Top-20 pool, and $K = 10$ neighbors. ANCHOR-RAG selects its Top-10 by stable ascending MI distance within that pool. The frozen predictor uses the selected histories and continuations without a distance bias; the outer correction also uses MI distances.

Locked-parameter ablations. All variants retain Full's validation-selected (α_d, s_d), $T = 1$, and frozen branch gain 1; none is retuned. "w/o MI rank (corr.)", "w/o MI dist.", and "w/o MI-cond. gate" keep Full's MI-selected frozen forecast fixed. They respectively use official Top-10 neighbors with MI weights in the correction, replace correction distances with official distances, or set $g_q = 1$. "w/o BCSA" removes endpoint and scale alignment. "w/o MI" selects official Top-10 neighbors for the frozen forecast and uses official distances for correction; "w/o MI & BCSA" also removes alignment. "w/o residual correction" sets $\alpha_d = 0$ but retains MI retrieval, yielding Eq. 7.

Paired test-origin inference. Losses are averaged over the 64-step horizon and across channels to form one paired loss per forecast origin. Each metric uses 5,000 circular moving-block bootstrap replicates with 64-origin blocks. The two-sided sign-flip test uses 50,000 random sign assignments to non-overlapping block means, forming consecutive blocks of up to 64 origins and retaining the final partial block. For Exchange Rate (1,454 origins), this gives 22 complete blocks and one final block of 46 origins; block means receive equal weight. Macro intervals average dataset-level relative reductions within each replicate, weighting the seven datasets equally.

Table 21: **Paired test-origin uncertainty for ANCHOR-RAG relative to TS-RAG.** Effects are relative error reductions in percent, followed by moving-block bootstrap 95% intervals. Positive values favor ANCHOR-RAG. MSE/MAE inputs are the unrounded paired per-origin losses underlying Table 1.

Dataset	Origins	MSE reduction		MAE reduction	
		Effect [95% CI]	p_{block}	Effect [95% CI]	p_{block}
ETTh1	2,817	$+2.027\ [1.151, 2.902]$	$< 10^{-4}$	$+0.239\ [-0.089, 0.557]$	0.0915
ETTh2	2,817	$+0.312\ [-0.118, 0.707]$	0.1607	$+0.257\ [0.047, 0.479]$	0.0132
ETTm1	11,457	$+4.661\ [3.572, 5.595]$	$< 10^{-4}$	$+0.884\ [0.398, 1.342]$	0.0016
ETTm2	11,457	$+1.853\ [1.242, 2.461]$	$< 10^{-4}$	$+0.584\ [0.312, 0.863]$	0.0004
Weather	10,476	$+7.166\ [5.657, 8.732]$	$< 10^{-4}$	$+3.790\ [2.855, 4.671]$	$< 10^{-4}$
Electricity	5,197	$+0.901\ [0.757, 1.048]$	$< 10^{-4}$	$-0.060\ [-0.120, 0.002]$	0.0823
Exchange Rate	1,454	$+4.247\ [0.362, 8.101]$	0.0698	$+2.229\ [0.165, 4.465]$	0.0801
Macro mean	–	$+3.024\ [2.405, 3.657]$	–	$+1.132\ [0.782, 1.484]$	–

Table 22: **Paired ablation inference against Full.** Effects are equal-dataset mean relative error reductions with 95% moving-block bootstrap intervals; positive values favor Full. Wins count datasets with positive reduction. Significance counts use two-sided 64-origin block sign-flip tests with Holm adjustment across seven ablations and seven datasets within each metric.

Ablation	MSE			MAE		
	Effect [95% CI]	Wins	Holm $p < .05$	Effect [95% CI]	Wins	Holm $p < .05$
w/o MI rank (corr.)	$+0.055\ [-0.160, 0.257]$	5/7	2/7	$+0.067\ [-0.036, 0.171]$	3/7	1/7
w/o MI dist.	$+0.308\ [0.140, 0.483]$	6/7	1/7	$+0.238\ [0.129, 0.344]$	5/7	2/7
w/o MI-cond. gate	$+41.338\ [39.676, 42.951]$	6/7	7/7	$+22.876\ [21.068, 23.993]$	7/7	6/7
w/o BCSA	$+13.674\ [12.739, 14.532]$	6/7	3/7	$+9.712\ [8.755, 10.573]$	7/7	5/7
w/o MI	$+0.846\ [0.465, 1.231]$	6/7	3/7	$+0.743\ [0.536, 0.960]$	7/7	3/7
w/o MI + BCSA	$+13.188\ [11.851, 14.283]$	7/7	3/7	$+8.606\ [7.430, 9.671]$	7/7	5/7
w/o residual correction	$+2.820\ [2.148, 3.461]$	7/7	6/7	$+0.882\ [0.538, 1.235]$	6/7	2/7

Validation-only robust parameter selection. For each dataset, (α_d, s_d) is selected exclusively on three chronological pure-MI Top-10 validation blocks using a fixed grid and a prespecified robust blockwise criterion normalized by the matched validation baseline. The selected parameters are frozen before test evaluation and reused unchanged for the matched test and all ablations; no test-set metric or ablation-specific retuning is used.

Table 23: **Validation-selected ANCHOR-RAG parameters** Parameters are selected on pure-MI Top-10 validation sidecars and frozen before test evaluation. BCSA uses fixed $\eta = 0.5$ and $T = 1$.

Dataset	α_d	s_d	Dataset	α_d	s_d
ETTh1	0.10	0.55	ETTm2	0.10	2.90
ETTh2	0.05	3.45	Exchange Rate	1.00	0.20
ETTm1	0.10	50.00	Weather	0.35	2.75
Electricity	0.10	1.25			

J STATISTICS AND COST

The primary intervention family contains 42 model–regime configurations at each of four nested patch fractions. Benjamini–Hochberg correction covers all 168 cells within each mode, endpoint, and contrast family. This family definition accounts for shared data partitions, donors, and patch orders. The table reports positive effects, raw sign-flip results, corrected discoveries, and the independent RMS-normalized block-bootstrap audit. Holm correction serves as a deliberately stringent power diagnostic. With eight donor shifts, the exact sign-flip grid has a minimum non-zero one-sided p value of 0.0039; the prespecified BH family remains the primary multiplicity-controlled analysis.

Table 24: **Complete-family BH audit for prediction response.** Counts use signed all-variable ground-truth delta-MSE over the same 168-condition removal families. Favorable and adverse discoveries have positive and negative high-minus-control effects, respectively.

Contrast	Fraction	Positive	BH favorable	BH adverse
High-MI − Low-MI	12.5%	41	38	1
High-MI − Low-MI	25%	39	37	1
High-MI − Low-MI	37.5%	38	36	2
High-MI − Low-MI	50%	39	36	2
High-MI − random	12.5%	37	36	4
High-MI − random	25%	37	36	4
High-MI − random	37.5%	39	37	3
High-MI − random	50%	39	37	2

Table 25: **Corrected progressive intervention audit.** "BH" counts cells with complete-family $q < 0.05$ over 168 conditions. "RMS +" counts positive 95% moving-block bootstrap intervals for forecast drift divided by perturbation RMS.

Contrast	Fraction	Positive	Raw $p < .05$	BH $q < .05$	Holm $p < .05$	RMS +	Mean Δ_{RMS}
High-MI − Low-MI	12.5%	39	39	39	0	40	0.388
High-MI − Low-MI	25%	37	37	37	0	37	0.445
High-MI − Low-MI	37.5%	38	37	37	0	38	0.484
High-MI − Low-MI	50%	39	39	39	0	39	0.542
High-MI − random	12.5%	33	33	33	0	34	0.325
High-MI − random	25%	36	36	36	0	37	0.379
High-MI − random	37.5%	38	38	38	0	38	0.342
High-MI − random	50%	41	38	38	0	37	0.269

The widespread discoveries show that hidden-state dependence is detectable against the registered temporal null. They do not validate the relative ordering of neighboring cells. Fine ranking is therefore evaluated separately through estimator, projection, sample-size, and target-sensitivity analyses.

Table 26: **Atlas significance and offline analysis cost.** Significance fractions are the within-atlas BH-adjusted cell fractions already used for atlas selection. Analysis time is logged wall-clock time for the discovery atlas build and cache, not inference latency.

Model	Atlases	Mean significant fraction	Median	Minimum	Mean atlas seconds
Chronos-2	7	0.991	1.000	0.953	142.65
Moirai-2	7	0.985	1.000	0.922	60.69
Toto-2	7	0.972	1.000	0.885	53.60
TimesFM 2.5	7	0.972	1.000	0.822	241.19
Chronos-Bolt	7	0.999	1.000	0.995	159.85
TTM-R2	7	0.861	1.000	0.475	12.38
All 42	42	0.963	1.000	0.475	111.73

K Robustness and Scaling

K.1 Origin-selection stability across all 42 atlases

Table 27: **ETT-family length/horizon robustness.** Mean Spearman correlations from $(512, 96)$ to $(336, 192)$; TTM-R2 requires native $(512, 96)$.

Model	Patch profile	Layer profile	Full atlas
Chronos-2	0.450	0.404	0.509
Moirai-2	0.636	0.302	0.649
Toto-2	0.734	0.470	0.615
TimesFM 2.5	0.549	0.742	0.597
Chronos-Bolt	0.446	0.337	0.469

Table 28: **ETTm1 estimator sensitivity.** Six-model means relative to registered $k = 5$, PCA=8, and 199 null shifts.

Variant	Patch profile	Layer profile	Full atlas
$k = 3$	0.931	0.904	0.928
$k = 10$	0.934	0.828	0.931
PCA=4	0.760	0.418	0.742
PCA=16	0.805	0.711	0.766
99 nulls	0.839	0.896	0.917
399 nulls	0.944	0.932	0.964

Table 29: **ETTm1 discovery-sample stability.** Nested evenly spaced subsets lock the registered future and layer-shared hidden projections, then recompute KSG MI and 199 null shifts. Values average six models relative to the 1,024-origin atlas.

Origins	Patch ρ	Layer ρ	Cell ρ	Top-quarter Jaccard	Same selected patch
512	0.694	0.528	0.703	0.609	6/6
768	0.701	0.650	0.697	0.586	6/6

Table 30: **Full 42-atlas origin-selection audit.** Each model–regime atlas contributes 50 replicates retaining 12 of 16 chronological discovery blocks. Registered future/hidden projections are fixed; every replicate recomputes KSG MI and cell null scores with 49 legal shifts. Columns are reselection probabilities for the primary selected-patch index (ignoring layer), global top-patch index (a different object), selected layer, and exact layer–patch pair; J_{25} is mean top-quarter Jaccard. All values are proportions.

Model	Regime	Selected patch	Global patch	Layer	Exact pair	J_{25}
Chronos-2	Electricity	0.080	0.080	0.140	0.020	0.237
	ETTh1	0.620	0.880	0.520	0.340	0.491
	ETTh2	0.080	1.000	0.360	0.080	0.278
	ETTm1	0.700	1.000	0.160	0.080	0.218
	ETTm2	0.040	0.980	0.440	0.040	0.183
	Traffic	0.000	0.140	0.120	0.000	0.192
	Weather	0.000	0.940	0.360	0.000	0.171
Chronos-Bolt	Electricity	0.100	0.180	0.080	0.020	0.274
	ETTh1	0.300	0.640	0.160	0.100	0.498
	ETTh2	0.660	0.980	0.300	0.180	0.356
	ETTm1	0.700	1.000	0.160	0.160	0.390
	ETTm2	0.700	1.000	0.260	0.240	0.382
	Traffic	0.080	0.060	0.080	0.020	0.155

Model	Regime	Selected patch	Global patch	Layer	Exact pair	J_{25}
	Weather	0.660	0.940	0.120	0.120	0.398
Moirai-2	Electricity	0.020	0.040	0.200	0.020	0.230
	ETTh1	0.440	0.720	0.600	0.260	0.500
	ETTh2	0.940	1.000	0.500	0.480	0.411
	ETTm1	0.840	0.980	0.220	0.180	0.559
	ETTm2	0.700	0.980	0.200	0.100	0.537
	Traffic	0.040	0.080	0.120	0.000	0.164
	Weather	0.880	1.000	0.260	0.240	0.653
TimesFM 2.5	Electricity	0.100	0.340	0.080	0.020	0.242
	ETTh1	0.600	1.000	0.600	0.300	0.347
	ETTh2	0.820	0.960	0.420	0.400	0.274
	ETTm1	0.880	1.000	0.140	0.140	0.752
	ETTm2	0.800	1.000	0.180	0.180	0.616
	Traffic	0.140	0.320	0.140	0.020	0.183
	Weather	0.940	1.000	0.100	0.080	0.850
Toto-2	Electricity	0.160	0.060	0.240	0.000	0.210
	ETTh1	0.700	0.960	0.240	0.200	0.666
	ETTh2	0.760	1.000	0.620	0.500	0.438
	ETTm1	0.960	1.000	0.500	0.500	0.764
	ETTm2	0.880	1.000	0.240	0.220	0.583
	Traffic	0.060	0.160	0.360	0.040	0.183
	Weather	0.880	0.980	0.640	0.560	0.800
TTM-R2	Electricity	0.160	0.260	0.620	0.120	0.220
	ETTh1	0.240	0.580	0.620	0.120	0.353
	ETTh2	0.500	1.000	0.720	0.400	0.480
	ETTm1	0.680	0.980	0.680	0.540	0.667
	ETTm2	0.620	1.000	0.480	0.300	0.673
	Traffic	0.240	0.240	0.380	0.140	0.220
	Weather	0.620	0.980	0.640	0.420	0.633
Macro mean	42 atlases	0.484	0.725	0.333	0.188	0.415

The selected patch is more reproducible than the selected layer on average (0.484 versus 0.333); this ordering appears in 23/42 individual atlases (four tie). Exact-coordinate reselection is 0.188 and mean J_{25} is 0.415. ETTm1 alone gives the previously reported 0.793 versus 0.310 contrast. Electricity and Traffic selected-patch means are 0.103 and 0.093, respectively, on the registered 16-channel panels. The global top-patch index is more repeatable overall (0.725), especially on ETT and Weather, while the reachable-layer coordinate remains the preregistered site for held-out probes and interventions. With PCA projections fixed, this audit measures origin-selection stability and defines temporal patch identity as the primary reproducible unit, with layer depth serving as a model-conditional coordinate.

Table 31: **Matched registered-panel/full-variable stress test.** Spearman correlations from mean+standard-deviation channel views and 128 discovery origins.

Model	Dataset	Channels	Patch profile	Layer profile	Full atlas
TimesFM 2.5	Electricity	16 / 321	0.065	-0.060	0.068
TimesFM 2.5	Traffic	16 / 862	0.947	0.743	0.743
Chronos-Bolt	Electricity	16 / 321	0.314	-0.444	0.053
Chronos-Bolt	Traffic	16 / 862	0.765	0.697	0.805
TTM-R2	Electricity	16 / 321	0.603	0.749	0.252
TTM-R2	Traffic	16 / 862	0.780	0.955	0.671