INFORMATION WITHOUT GEOMETRY: THE ENCODING–RELIABILITY DISSOCIATION IN TIME SERIES FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Time-series foundation models (TSFMs) achieve strong zero-shot and transfer performance, but an important question remains: what do they learn about time series? Forecast accuracy shows that a model can predict, but not whether its representations capture a useful notion of similarity. We study this question through representation geometry, decomposing representation quality into three properties: **Information** – whether task-relevant temporal structure is encoded; **Geometry** – whether embedding proximity reflects similarity in forecast difficulty; and **Utility** – whether this geometry improves a downstream operation. This decomposition exposes an **Encoding–Reliability Dissociation (ERD)**: representations can encode temporal structure and exhibit non-trivial geometry without providing a useful similarity space. Across four architectures, three pretraining paradigms, four benchmarks, and matched random-weight controls, we find that TSFMs often encode temporal information, though its recoverability varies across architectures and datasets, while exhibiting structured local geometry. Yet this structure does not translate into retrieval utility: with Euclidean retrieval, no TSFM embedding outperforms a raw-window baseline on any dataset, the single exception being Chronos-Small with DTW reranking on ETTh1 (85.6% vs. 84.2%). We further show theoretically that Information does not imply Geometry for injective encoders, and empirically that geometry can be induced without reliably improving downstream utility. Our findings suggest that learning to predict is not the same as learning to compare: TSFM embeddings should be evaluated not only for what they encode, but for whether their geometry supports the downstream similarity relation they are used to provide.

1 INTRODUCTION

Time series foundation models (TSFMs) are increasingly used as reusable representations for downstream tasks such as retrieval-augmented forecasting, anomaly detection, and clustering (Das et al., 2024; Ansari et al., 2024; Goswami et al., 2024; Woo et al., 2024; Shi et al., 2024). These applications assume that proximity in embedding space reflects similarity for the downstream task. Yet prediction-oriented pretraining encourages retaining predictive information without requiring representation distances to reflect predictive similarity. An embedding can therefore preserve task-relevant information while its geometry remains unaligned with distance-based downstream operations.

We study this possibility through three properties of a frozen representation: **Information** – can temporal concepts such as seasonality, volatility, and trend be decoded from the representation? **Geometry** – does proximity in embedding space predict forecast difficulty? **Utility** – does this geometry improve a downstream operation, such as retrieval-augmented forecasting? We refer to their separation as the **Encoding–Reliability Dissociation (ERD)**: information can be present without useful geometry, and geometry need not translate into downstream gains.

We test this without fine-tuning: linear probing for Information, local-density and pairwise tests for Geometry, and leakage-safe retrieval for Utility (§3–§4), comparing against matched random-weight controls. Across 11 TSFMs, four benchmarks, and two OOD equity indices (with Benjamini–

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.

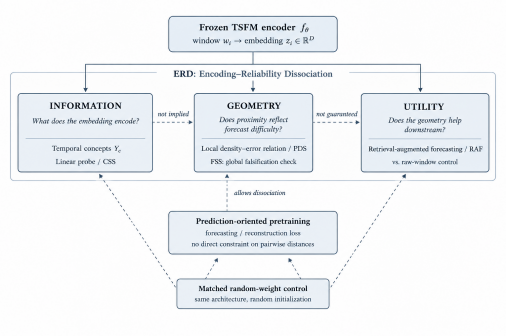

Figure 1: ERD: Information need not imply Geometry, and Geometry need not guarantee Utility. Prediction-oriented pretraining (Assumption 4) allows this; the random-weight control isolates what pretraining adds at each stage.

Hochberg correction), we find three core patterns (§6): (1) Information is present but uneven across datasets; (2) Geometry often predicts forecast difficulty, but without a uniform architectural signature; (3) Utility fails under Euclidean retrieval, where no embedding beats a raw-window baseline (Chronos-Small under DTW on ETTh1 being the sole broader exception). OOD transfer remains open: a strict rerun finds neither Information nor Geometry transfers significantly (§6.4).

Contributions: (1) *Concept* – we formalise ERD separating Information, Geometry, and Utility, proving theoretically that this separation is possible for injective encoders (§3). (2) *Evidence* – we evaluate these across 11 TSFMs, 4 architectures, and matched controls (§4). (3) *Evaluation* – we show representation geometry does not reliably guarantee downstream utility, establishing that embeddings must be tested against downstream tasks directly.

2 Related Work

TSFMs and representation learning. TSFMs span decoder-only architectures, quantised encoder-decoders, and masked encoders (Das et al., 2024; Ansari et al., 2024; Goswami et al., 2024; Woo et al., 2024), increasingly deployed for downstream tasks. While forecast accuracy dominates their evaluation, representations can reflect trend and class structure (Goswami et al., 2024), and contrastive methods encourage clustering (Yue et al., 2022; Tonekaboni et al., 2021). In broader deep learning, geometric diagnostics (intrinsic dimension, CKA) reveal structure beyond task accuracy (Ansuini et al., 2019; Kornblith et al., 2019). However, they do not verify whether embedding distance reflects forecast-relevant similarity. We test whether Information, Geometry, and Utility (§3) form a coherent chain, rather than assuming one implies the next.

Evaluation integrity and contamination in TSFMs. Audits document overlap between pretraining and evaluation corpora in time series (Aksu et al., 2024; Meyer et al., 2025b), and probe adaptation rates have been used to flag contamination (Li et al., 2026). Our evaluation integrity screen (§4) evaluates a complementary observable: latent trajectory complexity relative to random-weight controls.

3 Encoding–Reliability Dissociation

Encoding–Reliability Dissociation (ERD) describes a relationship between three properties of a frozen representation, independent of how each property is measured; Section 4 specifies the concrete diagnostics (LGF) used to test each one.

Definition 1 (Frozen Encoder and Embedding)**.** Let $f_\theta^{\mathcal{L}} : \mathbb{R}^L \to \mathbb{R}^{T\times D}$ be the final-layer map of a frozen TSFM encoder. Given a window w_i, its embedding is obtained by mean-pooling the final-layer token representations: $z_i = \frac{1}{T}\sum_{t=1}^{T} f_\theta^{\mathcal{L}}(w_i)_t \in \mathbb{R}^D$, where $T{=}L/P$ is the number of patch tokens and $\mathcal{L}$ denotes the final layer.

Definition 2 (Window Set)**.** Let $\mathcal{W}=\{(w_i, s_i)\}_{i=1}^{N}$ be windows with start times s_i and forecast errors $e_i = \frac{1}{H}\sum_{h=1}^{H}(\hat{y}_{i,h} - y_{i,h})^2$, the per-window MSE at horizon $H=96$ (non-negative by construction; the absolute value in the PDS definition below is retained only for notational uniformity).

Definition 3 (Temporal Representation Manifold)**.** $\mathcal{M}_\theta = f_\theta(\mathrm{supp}(\mathcal{P})) \subset \mathbb{R}^D$ is the image of the encoder over the data support, approximated empirically by $\mathcal{Z}=\{z_i\}$. Under a compact-input-support assumption (Appendix B), $\mathcal{M}_\theta$ is compact and rectifiable – we do not additionally claim finite *intrinsic* dimension.

We distinguish three properties of the pair $(f_\theta, \mathcal{Z})$: **Information** – whether task-relevant temporal concepts Y^c (seasonality, volatility, trend) are recoverable from z_i by linear probing (CSS); **Geometry** – whether embedding proximity predicts similarity in forecast difficulty (local density–error correlation PDS, with pairwise correlation FSS as a check); and **Utility** – whether this geometry improves a concrete downstream operation (retrieval-augmented forecasting, RAF, against a model-free raw-window control).

These three properties are often conflated under "representation quality," but we find they are empirically dissociable across architectures and datasets: the Encoding–Reliability Dissociation (ERD). We operationalise Geometry as density–error calibration because that is the local property retrieval diagnostics assume; pairwise future-continuation preservation is evaluated directly by our retrieval metrics (§6.3). The dissociation we report is therefore specifically between density–error calibration and downstream retrieval success.

Assumption 4 (Prediction-Oriented Pretraining)**.** A pretraining objective is *prediction-oriented* if its loss is a function of f_θ's output (or a decoder built on f_θ) against a forecasting or reconstruction target, with no term that is a direct function of pairwise distances in $\mathcal{Z}$.

Why this assumption plausibly produces ERD, and how the dissociation can be tested and falsified, is discussed alongside the relevant theorems in §4; the independence theorem there is stated for Shannon mutual information specifically, not for linear-probe decodability.

4 Experimental Design

We evaluate three properties of a frozen representation in sequence: Information (encoded temporal structure), Geometry (neighbourhood reliability), and Utility (downstream retrieval). Benchmark integrity is addressed in parallel: Part B uses Latent Trajectory Complexity (LTC) to screen for potential contamination (Meyer et al., 2025b; Li et al., 2026), serving as an evaluation safeguard rather than an intrinsic representation property.

Part A: Representation Geometry

Manifold Regularity (LCS, supporting diagnostic). To test whether embeddings reflect coarse temporal structure, we group windows by trend characteristics (STL decomposition, Cleveland et al. (1990)) and measure embedding-space clustering via the mean silhouette coefficient (Rousseeuw, 1987): the **Latent Cohesion Score (LCS)**. Positive LCS indicates clustering beyond high-dimensional i.i.d. random embeddings ($\mathbb{E}[\mathrm{LCS}] \to 0$ as $D \to \infty$; Appendix B); LCS serves as a baseline check and does not establish predictive geometry.

Geometry: neighbourhood reliability (PDS, FSS). Geometry asks a more specific question: does proximity in embedding space predict forecast difficulty? We test this locally with the Proximity-Density Score: for each window i, let $\bar{d}_i$ be its mean distance to its $k=10$ nearest neighbours and e_i its forecast error,

$$\rho_s^{\mathrm{PDS}} = \mathrm{Spearman}\big(\{\bar{d}_i\}_{i=1}^{N},\ \{|e_i|\}_{i=1}^{N}\big).$$

A positive PDS means that isolated windows tend to have larger forecast errors – the local relationship embedding-based retrieval would need to exploit. We also compute the Full-Space Score over randomly sampled pairs,

$$\rho_s^{\mathrm{FSS}} = \mathrm{Spearman}\big(\{\|z_i - z_j\|\}_{(i,j)\in\mathcal{P}},\ \{|e_i - e_j|\}_{(i,j)\in\mathcal{P}}\big),$$

testing whether pairwise distance predicts forecast-difficulty differences throughout the space, where $\mathcal{P}$ is a uniform random sample of distinct unordered pairs. PDS asks if local density is informative

for a window's own reliability; FSS asks if global distance carries the same information, which distance concentration can suppress even when local density is informative – so a positive PDS with weak FSS is compatible with a useful local signal, not evidence against it.

Theorem 5 (PDS Positivity and OOD Collapse)**.** ***(i)*** *If the conditional error surface is non-increasing in local embedding density (C2), then* $\rho_s^{\mathrm{PDS}} \geq 0$ *in distribution; this is a conditional result, so an observed* $\rho_s^{\mathrm{PDS}}<0$ *indicates that C2 does not hold empirically, rather than a failure of the theorem, as observed in most models.* ***(ii)*** *If OOD embeddings become approximately uniform over* $\mathcal{M}_\theta$*, variation in* $\bar{d}_i$ *vanishes and* $\rho_s^{\mathrm{PDS}} \to 0$*; empirically, the corrected OOD protocol finds no significant PDS value (Appendix A, Table 2), while the earlier protocol's exploratory values (10-model OOD range* $[-0.33, -0.01]$*; TimesFM from* $+0.335$ *in-distribution to* $+0.02$ *OOD) are consistent with collapse but are superseded and not part of the confirmatory family.* ***(iii)*** *There exist configurations in which* $\rho_s^{\mathrm{PDS}} \gg \rho_s^{\mathrm{FSS}}$*, showing that global pairwise distances need not recover a local density signal under high-dimensional distance concentration.* (Proofs: Appendix B.)

Information: temporal primitive probing (CSS). Information asks a simpler question: which temporal properties can be recovered from the frozen representation? Labels come directly from STL decomposition: seasonality, volatility, and trend each split at their median (seasonal strength, residual variance, slope). An ℓ_2-regularised logistic probe (C=1), evaluated with repeated scattered blocked cross-validation (5 folds, 80/20 per fold, dependency-exclusion gap $\lceil L/s_{\mathrm{eff}} \rceil$), yields balanced accuracy,

$$\mathrm{CSS}(c) = \mathrm{BalAcc}(\hat{\mathbf{y}}^c, \mathbf{y}^c), \quad c \in \{\text{trend, season, vol}\}.$$

We compare each pretrained encoder with a matched random-weight encoder of the same architecture; a positive pretrained–random gap indicates the corresponding temporal property is recoverable beyond what the architecture alone provides.

Theorem 6 (CSS and Mutual Information, summary)**.** *Under a Gaussian, equal-covariance population model, population-optimal linear-probe balanced accuracy exceeds* $1/2$ *iff the Fisher discriminant ratio* $J_F > 0$*; a positive empirical CSS gap is evidence for, but does not by itself prove,* $J_F > 0$ *at the population level. Moreover,* $\mathrm{CSS}(c)$=α *implies* $I(Z; Y^c) \geq 1 - H_b(1-\alpha)$ *bits of mutual information, so the pretrained–random CSS gap yields an empirical shift between two plug-in lower-bound estimates,* $H_b(1-\alpha_r) - H_b(1-\alpha_p)$ *bits, rather than a certified lower bound on the information gap itself.* (Proof: Appendix B.)

Why Information need not imply Geometry. The prediction objectives used by TSFMs – autoregressive next-step prediction, masked patch reconstruction, token-level cross-entropy – provide a learning signal for forecasting but do not directly constrain distances between representations (Assumption 4): they can encourage predictive information without requiring nearby representations to share forecast difficulty. This is the mechanism that allows ERD. Theorem 7 (Appendix B) formalises it: for any injective encoder, Shannon information is invariant while local-density geometry is not, so an embedding can preserve the information needed for prediction while failing to organise it in a useful metric space. The theorem is a statement about mutual information, which is invariant under any injective relabeling; it is not a claim about linear-probe decodability (CSS), which is itself geometric.

Whether this occurs in actual TSFMs is empirical. The differing random-weight-control patterns across architectures (§6) suggest the training objective is not the whole story – architecture matters too – so we treat the dissociation's source as empirical, not attributable to the pretraining objective alone.

Utility: retrieval-augmented forecasting (RAF). Finally, we ask whether Geometry translates into a useful downstream operation: retrieval-augmented forecasting (RAF), where historical windows are selected by embedding proximity and used by the forecasting procedure, compared against a model-free raw-window baseline under a leakage-safe protocol (§6.3). This step matters because Geometry and Utility are different claims – an embedding can contain temporal information and meaningful local geometry without that geometry improving retrieval.

PART B: EVALUATION INTEGRITY

Prerequisite: Latent Trajectory Complexity (LTC). Before interpreting Information or Geometry, we screen for possible benchmark contamination. Memorisation of benchmark-specific patterns could produce apparently meaningful representations even when they do not arise from generalised temporal structure (Carlini et al., 2023; Magar & Schwartz, 2022). LTC is designed as a screening diagnostic for this possibility.

LTC measures two properties of the embedding trajectory as windows move through time: the mean step size $\mathrm{LTC}_{\mathrm{step}} = \frac{1}{N-1}\sum_i \|\delta_i\|_2$ and directional consistency $\mathrm{LTC}_{\mathrm{dir}} = \frac{1}{N-2}\sum_i \frac{\delta_i \cdot \delta_{i+1}}{\|\delta_i\|\|\delta_{i+1}\|}$, where $\delta_i = z_{\sigma(i+1)} - z_{\sigma(i)}$ for windows sorted by start time σ. The random-weight model provides the reference distribution: a pretrained model is flagged when its trajectory step size exceeds the corresponding random-control expectation, indicating possible memorised benchmark structure.

Importantly, an LTC flag is evidence of contamination *risk*, not evidence of contamination itself, so we treat it as a screening criterion rather than a detection test. Independent evidence corroborates only one of the models flagged by LTC (§6.5).

5 EXPERIMENTAL SETUP

Models. We evaluate four primary TSFM architectures spanning three pretraining paradigms: **TimesFM 2.0** (Das et al., 2024; 2025) (decoder-only, 500M parameters, D=1024), **Chronos-Small** (Ansari et al., 2024) (quantised encoder-decoder, 46M, D=512), **MOMENT-Large** (Goswami et al., 2024) (masked encoder, 385M, D=1024), and **Moirai-1.1-Large** (Woo et al., 2024) (masked encoder-decoder, 311M, D=1024).[1] Extended LTC screens cover 14 models (including LSTM, TCN, and Transformer baselines); Geometry evaluates the 11 TSFMs with forecast heads; Utility covers 6 architectures with embedding adapters (4 primary plus Sundial, Time-MoE). Full-scale bootstrap correction for Geometry applies to the 3 architectures with trained forecast heads (Table 1).

Random-weight controls. Each primary TSFM is paired with a matched random-weight control (same architecture and dimensionality, Kaiming-uniform initialisation). Information and Geometry claims are evaluated relative to these controls, isolating structure attributable to pretraining rather than absolute structure.

Data and protocol. We use four long-horizon benchmarks – Electricity, ETTh1, Weather, and Traffic – and two out-of-distribution equity indices, CSI300 and CSI800, selected outside known pretraining corpora. From each dataset we sample N=3,000 windows of length L=512 at uniform (`linspace`) positions across the full series, corresponding to an effective stride of 5.4–20 timestamps depending on the dataset (Appendix A); each window is normalised independently. All reported quantities are computed on a single univariate series per dataset: ETTh1 uses its long-horizon target channel, while Electricity, Traffic, and Weather are represented by the cross-sectional mean across channels at each timestamp, matching the geometry and RAF driver code. Channel-wise (channel-independent) evaluation is left to future work. Forecast horizons are $H \in \{96, 192, 336, 720\}$. Unless otherwise stated, we report medians with 95% bootstrap confidence intervals.

Statistical evaluation. Pretraining effects are evaluated relative to matched random-weight controls (p<0.05). We use block-permutation tests for CSS (n_{perm}=100 in-distribution; n_{perm}=2,000 for OOD) and paired Wilcoxon signed-rank tests for LTC (∼3,000 observations per dataset). The confirmatory family spans the 4 primary architectures across the 4 in-distribution benchmarks with Benjamini–Hochberg correction at q=0.05 (Benjamini & Hochberg, 1995); LTC screens, OOD evaluations, and ablations are exploratory.

[1] An extraction-path bug was corrected to use the native encoder tap (Definition 1); all Moirai numbers use the corrected tap, re-verified at full scale (Appendix A).

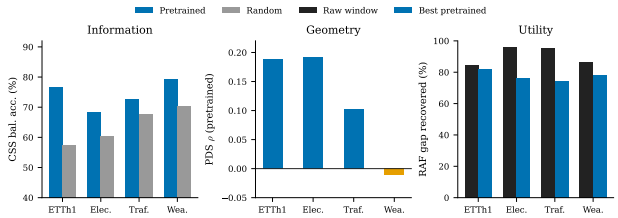

Figure 2: The three properties across datasets, four primary architectures (three for Geometry, which needs a trained forecast head). Information: pretrained CSS balanced accuracy exceeds its random-weight control on every dataset. Geometry: the pretrained PDS correlation is small and dataset-dependent, near zero on Weather. Utility: a model-free raw-window control matches or beats the best pretrained embedding's RAF gap on every dataset. Values are means (Information, Geometry) or per-dataset maxima (Utility) over the architectures in Tables 3, 1, and 6.

6 Results

Part A (§6.1–§6.4) tests Information, Geometry, Utility, and Transfer in turn. **Part B** (§6.5) reports the contamination-risk screen and a brief ablation study.

6.1 Information: Temporal Primitives Are Linearly Decodable (CSS)

Temporal semantics are decodable above chance in every architecture we test, but unevenly: some architecture-dataset pairs show statistically significant decoding, while many others do not survive multiple-comparison correction across the confirmatory family, leaving 10 of 32 cells significant, concentrated in seasonality (8 of 16) rather than volatility (2 of 16) (protocol and per-cell results: Appendix B, Table 3). Seasonality decodes more reliably than volatility, the decoder-only architecture (TimesFM-2.0) is the most consistent performer, and Moirai-1.1-Large is the only model with negative gaps. Where decoding succeeds, several point estimates exceed 80% balanced accuracy against a near-chance baseline, indicating information encoded unevenly across domains. This aligns with ERD: models that decode temporal concepts can still fail subsequent geometric and retrieval tests.

Two secondary patterns are noted, though they come from a four-model sample and are not a controlled causal test (Limitations). Out-of-distribution transfer of this decoded information appears specific to the decoder-only architecture, with an autoregressive exception (Chronos) plausibly explained by its quantisation discarding amplitude information. Trend, unlike seasonality or volatility, does not decode significantly in our probe evaluations (omitted from the 32-cell confirmatory family in Table 3, where balanced accuracy remains statistically indistinguishable from chance for all models); since window-level normalisation removes the level offset but preserves the intra-window slope, we do not attribute this to normalisation alone and leave the mechanism open.

6.2 Geometry: Real, But Architecture-Dependent, and Never Sufficient

Local embedding density (PDS) predicts forecast error for several architecture-dataset combinations across the three primary architectures with trained forecast heads (full protocol: Appendix B). After Benjamini–Hochberg correction across the 12-cell family (Table 5), the effect is significantly positive in 7 of 12 cells, statistically null in 4 (TimesFM on Traffic and Weather, Moirai on Traffic, Chronos-S on ETTh1), and significantly negative in 1 (Moirai on Weather, which also displays Moirai's lowest retrieval utility; §6.3). As §6.3 details, this geometric structure does not transfer to downstream Utility, even in Moirai's highest-PDS setting (Electricity, Figure 3).

Matched random-weight controls reveal no uniform architectural signature: control values range from negative to null to positive depending on the dataset, with no pattern common to all three architectures (§7). An exploratory screen across eleven TSFMs (Appendix B) confirms that global pairwise alignment (FSS) is generally absent.

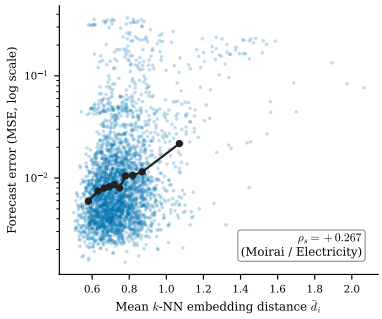

Figure 3: Strongest real PDS case (Moirai / Electricity): denser neighbourhoods skew toward lower error, real but moderate.

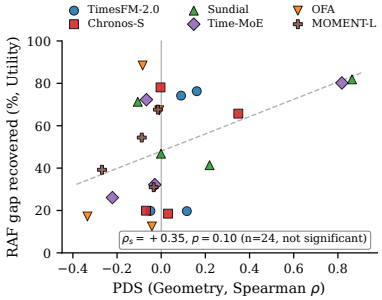

Figure 4: **Exploratory broad-screen**: PDS vs. RAF gap, 24 model–dataset pairs (ρ_s=+0.35, p=0.10). Points are from earlier broad screen, not confirmatory PDS family (Table 1 excludes MOMENT-L; Moirai not in table).

Table 1: **Neighbourhood reliability: real PDS Spearman ρ at full scale**, all four in-distribution datasets, N=3,000 windows, real per-window forecast error from each model's own forecast head, moving-block bootstrap (n_{boot}=10,000). †: 95% CI includes 0. Random-weight columns show one seed unless noted; ETTh1 random-weight cells for Chronos-S/TimesFM-2.0 report the median of three seeds. Forecast errors are each model's own per-window MSE at horizon H=96; random-weight cells correlate a re-initialised model's own forecast errors with its own embedding geometry. Broader, lower-rigor 11-model screen discussed in §6.2.

Model	Type	Dataset	ρ_s^{PDS} (pre)	ρ_s^{PDS} (rnd)	RAF gap
Chronos-S	Quantised	ETTh1	+0.186	−0.51 (med., 3 seeds)	81.7%
Chronos-S	Quantised	Electricity	+0.146	+0.002†	18.4%
Chronos-S	Quantised	Traffic	+0.133	−0.005†	19.9%
Chronos-S	Quantised	Weather	+0.131	+0.283	78.1%
TimesFM 2.0	Dec.-only	ETTh1	+0.132	−0.42 (med., 3 seeds)	33.0%
TimesFM 2.0	Dec.-only	Electricity	+0.161	−0.262	76.3%
TimesFM 2.0	Dec.-only	Traffic	+0.090†	−0.213	74.2%
TimesFM 2.0	Dec.-only	Weather	−0.049†	−0.314	19.8%
Moirai-1.1-Large	Masked enc.-dec.	ETTh1	+0.247	+0.201†	49.5%
Moirai-1.1-Large	Masked enc.-dec.	Electricity	**+0.267**	−0.078†	56.1%
Moirai-1.1-Large	Masked enc.-dec.	Traffic	+0.084†	+0.004†	68.8%
Moirai-1.1-Large	Masked enc.-dec.	Weather	**−0.116**	−0.206	0.8%

6.3 Utility: Downstream Retrieval Validation (RAF)

We evaluate whether embedding proximity improves forecasting via retrieval-augmented forecasting (RAF). Under our leakage-safe protocol, the first 80% of chronologically ordered windows form a candidate pool per query (excluding context-overlapping windows), predicting via the mean continuation of the k=10 nearest neighbours against random retrieval, an oracle bound, and a raw-window distance baseline (covering 6 architectures with adapters).

Under Euclidean retrieval, the raw-window baseline outperforms every pretrained embedding configuration across all four datasets: it recovers 84–96% of the random-to-oracle MSE gap, versus 74–82% for the best pretrained embedding on each dataset (Table 9; Appendix B). Across broader retrieval ablations (Table 6), Chronos-Small with DTW reranking on ETTh1 is the sole exception where an embedding configuration exceeds the raw-window baseline (85.6% vs. 84.2%). Furthermore, across the 24 exploratory broad-screen pairs, local density (PDS) does not predict the recovered gap (Figure 4, ρ_s=+0.35, p=0.10): density calibration does not generalise into a proxy for retrieval utility. Due to window overlap, moving-block bootstrap evaluation is employed; under this accounting, the raw-window advantage holds across nearly all comparisons, including Moirai on Electricity.

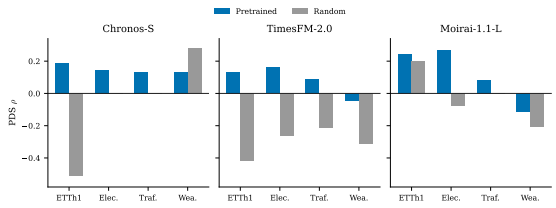

Figure 5: PDS Spearman ρ per dataset, pretrained vs. random-weight control, same data as Table 1. Each architecture's control follows a different pattern – Chronos-S and TimesFM-2.0 skew negative except on one dataset each, Moirai tracks its pretrained value loosely on 3 of 4 – with nothing common to all three.

Ablation study: retrieval vs. training. Alternative retrieval metrics and learned aggregation schemes yield minor effects (Appendix B). While DTW reranking improves embedding-based candidate selection, the raw-window+DTW control matches or outperforms embedding+DTW in 14 of 16 cells (Table 6). Varying the training objective (Control B) is similarly inconsistent: metric-oriented objectives improve utility on some datasets but degrade it on others. In contrast, direct rank agreement with the future oracle ranking does correlate with retrieval utility (ρ_s=+0.916, Appendix B), indicating that the dissociation is specific to density-based proxies rather than an inability of representations to capture retrieval order.

6.4 Transfer: Manifold Cohesion and OOD Validation

Temporal geometry. TSFM embeddings do not consistently organise windows by trend direction, and pretraining does not alter this. Seasonality and volatility show clearer structure, strengthened relative to random-weight controls. A separate directionality measure is negative across all models, including random controls, indicating anti-persistence arises from the windowing scheme rather than pretraining (Appendix B).

OOD validation. On freshly sourced equity-index data under a stricter protocol, neither Information nor Geometry transfers significantly (Appendix A, Table 2), contrasting with an earlier protocol where Information appeared to transfer while Geometry did not (Figure 6) – no longer read as evidence of OOD transfer. The corrected result is more specific: for Chronos-S on this dataset, neither property survives distribution shift; whether this extends across architectures and datasets remains open (§7).

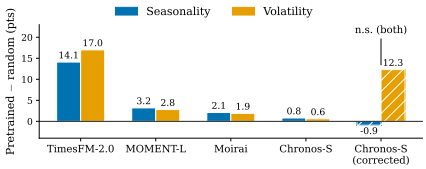

Figure 6: OOD equity gaps (pretrained minus random, pts): original protocol (Table 4) vs. corrected (Chronos-S, mean of CSI300/CSI800, Table 2), neither seasonality nor volatility significant for the rerun ($p \geq 0.16$).

6.5 Part B: Contamination-Risk Screening

Representation dynamics also reveal a distinct contamination signal. LTC, needing neither training nor error labels, flags five of 11 TSFMs (five of the 14 evaluated models) on ETTh1 through anomalously complex embedding trajectories: $\mathrm{LTC}_{\mathrm{step}}$ ratios reach 3–12$\times$ other benchmarks, versus 0.5–1.1$\times$ for random-weight controls. This is consistent with training-corpus overlap, independently confirmed for one flagged model (UniTime); the other four are *inferred*, not verified. One flagged model, VisionTS, adapts a visual masked autoencoder pretrained on ImageNet-1k with no time-series pretraining corpus, so its ETTh1 spike cannot reflect benchmark memorisation; it instead

shows that tokenisation/2D patching or high-frequency input dynamics can also inflate $\mathrm{LTC}_{\mathrm{step}}$. We treat this as an exploratory contamination-risk signal. Full per-model results, mechanism, and ablations are in Appendix B.

7 Conclusion

Information, Geometry, and Utility are empirically separable in time-series foundation models (ERD): learning to predict does not imply learning to compare. While TSFMs encode temporal information and produce local geometry, this geometry is not consistently aligned with forecast-relevant similarity. Under Euclidean retrieval, no embedding retrieves more reliably than the raw window, with Chronos-Small under DTW reranking on ETTh1 as the sole exception (85.6% vs. 84.2%). In practice, retrieval-augmented forecasting should be benchmarked against raw-window baselines. Our OOD findings currently cover Chronos-S, leaving broader cross-architecture shift open, while trajectory-complexity spikes ($> 3\times$ over random controls) provide an exploratory contamination screen (§6.5).

Limitations and Future Work. Several questions remain: STL decomposition may miss nonlinear temporal relations; temporal mean-pooling across patch tokens may attenuate high-frequency phase differences; OOD and Geometry evaluations currently cover subsets of architectures (§6.4); and metric-aligned pretraining (Control B) indicates that utility-aligned similarity spaces warrant systematic design.

The use of large language models (LLMs)

We used an LLM (Claude, Anthropic) as a coding and writing assistant throughout this project, under the authors' direction and review at every step: implementing and debugging the experiment and analysis code, running the resulting pipelines, and drafting and revising manuscript text from results the authors specified. The LLM was not used for research ideation, experimental design, or interpretation of results; all reported numbers were produced by code run against real data and real model checkpoints, and every bug that affected a reported result (e.g. a degenerate cross-validation split, a mis-scaled bootstrap block length) is disclosed at the relevant point in the text, with the full chronological record kept as a reproducibility log (Appendix A). The authors reviewed all LLM-assisted code and text and take full responsibility for the content of this paper.

Reproducibility statement

All experiments use publicly documented datasets (ETTh1, Electricity, Traffic, Weather) and publicly released model checkpoints (Chronos-S, TimesFM-2.0, Moirai-1.1-Large, MOMENT-Large), with exact commands, seeds, and output paths listed in Appendix A. Statistical protocols (blocked cross-validation, moving-block bootstrap, Benjamini-Hochberg correction) are specified in Section 4 and detailed further in Appendix B. Code will be released upon publication.

References

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Steven Hoi, and Doyen Sahoo. GIFT-Eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024.

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.

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimensionality of data representations in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *International Conference on Learning Representations*, 2023.

Robert B Cleveland, William S Cleveland, Jean E McRae, and Irma Terpenning. STL: A seasonal-trend decomposition procedure based on Loess. *Journal of Official Statistics*, 6(1):3–73, 1990.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10025–10043. PMLR, 2024.

Abhimanyu Das, Weihao Kong, Rajat Sen, Yichen Zhou, et al. TimesFM 2.0: An improved foundation model for time-series forecasting. *arXiv preprint arXiv:2501.08096*, 2025.

Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 (2):179–188, 1936. doi: 10.1111/j.1469-1809.1936.tb02137.x.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019.

Hongkai Li, Shifeng Xie, Lefei Shen, Zhuo Li, Mouxiang Chen, Xiaobin Zhang, Han Fu, Jianling Sun, Xiaoxue Ren, and Chenghao Liu. TSFMAudit: Data contamination auditing in forecasting time series foundation models. *arXiv preprint arXiv:2605.26161*, 2026.

Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 157–165, 2022. doi: 10.18653/v1/2022.acl-short.18.

Lukas Meyer, Tilman Schäfer, Raaghavi Ramamurthy, et al. A review of time series foundation models. *arXiv preprint arXiv:2502.13504*, 2025a.

Marcel Meyer, Sascha Kaltenpoth, Kevin Zalipski, and Oliver Müller. Rethinking evaluation in the era of time series foundation models: (un)known information leakage challenges. *arXiv preprint arXiv:2510.13654*, 2025b.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

Xiaoming Shi, Shiyu Ye, Yuqi Chen, Kailai Hu, Zhixuan Wan, Qingsong Wen, et al. Time-MoE: Billion-scale time series foundation models with mixture of experts. *arXiv preprint arXiv:2409.16040*, 2024.

Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. In *International Conference on Learning Representations*, 2021.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Salehi, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024.

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.

A Reproducibility Notes: Corrected-Protocol Derivation

During the development process, we identified and rectified three issues affecting statistical rigor. Here, we summarize these corrections for auditability, noting that all reported numbers in the main text already incorporate these adjustments.

CSS: Degenerate Cross-Validation Split. Initially, the CSS protocol employed a single final-20% chronological train/test split. Given that STL-derived seasonality/volatility labels exhibit long low-frequency regimes (with run lengths up to 653 windows on ETTh1), a contiguous 20% test block could fall entirely on one side of a globally-computed median split, resulting in a single-class test set and an invalid balanced-accuracy measure. To address this, we implemented a repeated, round-robin scattered blocked cross-validation (5 folds) with test windows drawn from various regimes through a seeded shuffle, rather than a single contiguous span. We also corrected the dependency-exclusion gap: the previous version used the raw window length as a window-index gap, but the correct measure is $\lceil L/s_{\text{eff}} \rceil$ in window-index units, where s_{eff} is the effective (`linspace`) stride in raw timestamps (5.4–20 depending on the dataset), since windows are placed by `linspace` across the full series, not a dense stride-1 sliding window. Verification against real data showed that the final test block was 100% single-class for both labels pre-fix; post-fix, all folds displayed balanced class mixes and reasonable (0.5–0.85) balanced accuracies.

PDS: Mis-Scaled Bootstrap Block Length. An initial PDS run used a hardcoded fallback stride formula that resulted in a block length covering approximately 76% of the array, making the moving-block bootstrap CI falsely narrow and overconfident, as each resample became 1–2 near-verbatim copies of the original data. Two of four initial cases appeared spuriously significant under this bug. The fix involved requiring an explicit, real per-dataset stride (computed from actual window start times) in `compute_pds`, with no silent default. Re-running with the correct stride showed that the two previously significant cases were indistinguishable from zero at that sample size.

Moirai: Extraction-Path Inconsistency. Initially, Moirai's embeddings were extracted via a forecast-output-derived feature routed through a smaller, architecturally different checkpoint (`moirai-2.0-R-small`), not the native encoder hidden state as specified in Definition 1 for the other architectures. The fix hooks the frozen encoder's final normalization layer of `Moirai-1.1-R-Large` via its GluonTS predictor, matching the 311M parameter count reported in the main text; all reported Moirai results use this corrected native tap, re-run at full scale (N=3,000, L=512) on all four in-distribution datasets for both PDS and CSS.

Runtime/Significance Trade-Off. CSS's block-permutation test requires one full 5-fold cross-validation pass per permutation, making it computationally expensive. Higher-dimensional embeddings (TimesFM-2.0, Moirai-1.1-Large, MOMENT-Large; $D \in \{1024\}$) cost approximately 25–29 times more per permutation than Chronos-S (D=512). A pre-run calculation confirmed that n_{perm}=1000 would take approximately 6.4 hours across the full 32-cell sweep; n_{perm}=100 was chosen as the tractable point (projected approximately 38 minutes, matched actual runtime), providing p-value resolution of 0.01, sufficient given the application of Benjamini–Hochberg correction.

Controlled Objective Comparison (This Revision): Exact Losses. All four encoders (§6.3, Table 8) share Control B's exact convolutional backbone and a linear forecast head $g_\psi(z) \to \hat{y}$. **Forecasting-only**: $\mathcal{L} = \text{MSE}(\hat{y}, y)$, with y normalized by the context window's own mean/std. **Contrastive** (Control B): InfoNCE over the batch, positive = the future-nearest other window (leakage-excluded, $|\Delta\text{start}| \geq L$), no forecast head. **Forecasting+Metric**: the two losses above, summed unweighted. **Forecasting+Retrieval**: forecasting loss plus a softmax-attention loss: each anchor's attention weights over every other (leakage-excluded) batch member by negative squared embedding distance, predicted future = attention-weighted average of their true futures, loss = MSE against the anchor's own true future; this is a differentiable relaxation of `retrieve()`'s k-NN-mean operation (§6.3), summed unweighted with the forecasting loss. All trained 8 epochs, batch size 128, Adam, D=64, on real data with a separate checkpoint per benchmark (ETTh1, Electricity, Traffic, Weather); evaluated through the identical CSS/PDS/RAF driver code used for every other embedding in the paper (`train_control_multiobjective.py`, `eval_control_b_raf.py`).

OOD Rerun (This Revision): Data Provenance and Scope. The original CSI300/CSI800 result (§6.4) used a panel-of-constituent-stocks construction (58,600 rows/index) whose generating script is absent from this codebase and could not be reproduced. Instead of leaving the corrected protocol unattempted, we sourced fresh, real, index-level daily closes for both indices directly (public market data, full available history through the date of this revision: CSI300 5,967 trading days from 2002, CSI800 5,246 from 2005, both well beyond any known TSFM pretraining cutoff), windowed identically to every in-distribution dataset (N=3,000, L=512, H=96, `linspace` placement), with STL period 5 (trading-week seasonality) for the seasonality/volatility labels. Real per-window forecast error came from Chronos-S's own forecast head (20-sample median, matching §6.2's in-distribution protocol exactly), for both the pretrained checkpoint and a matched random-weight control (seed 42). This is a real, index-level series, not a reproduction of the original panel; the two are not directly comparable result-for-result, only jointly reported (§6.4). CSS's block-permutation test used n_{perm}=2,000 rather than the paper's 10,000 default: an initial 10,000-permutation run was terminated after 25 minutes of wall time with no key yet complete (CPU-bound refitting, unlike PDS's vectorized bootstrap), an honest turnaround-time trade-off, not a methodological shortcut; 2,000 permutations still provide p-value resolution of 0.0005. Only one architecture (Chronos-S, the fastest to run end-to-end on this session's hardware) was tested; extending to the other three primary architectures is the direct next step for closing this gap fully (§7).

Table 2: **Corrected OOD protocol, full results (Chronos-S only).** CSS: balanced accuracy (%), pretrained vs. random-weight control, block-permutation p (n_{perm}=2,000). PDS: Spearman ρ with 95% moving-block-bootstrap CI (n_{boot}=10,000). No cell reaches significance. Referenced from §6.4.

Index	Label	CSS (Pre/Rnd)	p	PDS ρ (Pre/Rnd)
CSI300	Seasonality	47.6/50.3	0.696	$+0.168$ [-0.06, $+0.48$] / -0.132 [-0.35, $+0.06$]
CSI300	Volatility	60.1/46.0	0.165	
CSI800	Seasonality	52.4/51.4	0.896	$+0.281$ [-0.02, $+0.41$] / -0.206 [-0.31, $+0.09$]
CSI800	Volatility	62.4/51.9	0.323	

B Full Statistical Tables

Table 3: **CSS at the corrected scale, all 4 in-distribution datasets, all 4 primary architectures: balanced accuracy (%), pretrained vs. random-weight control, rolling blocked CV, n_{perm}=100.** q: Benjamini–Hochberg-adjusted across the full 32-test confirmatory family (16 Seasonality + 16 Volatility). **Bold**: $q<0.05$. Chance $\approx$50%. Trend probe accuracy is statistically indistinguishable from chance across all models ($q > 0.20$) and is omitted from the table to focus on the 32 confirmatory Seasonality/Volatility cells. With n_{perm}=100 the permutation p-value floor is 0.01; ten cells attain this floor and therefore share the same BH-adjusted q=0.032. Referenced from §6.1.

Model	Dataset	Sea (Pre/Rnd)	q	Vol (Pre/Rnd)	q
Chronos-S	ETTh1	65.0/57.0	0.432	67.9/55.1	0.432
Chronos-S	Electricity	66.5/60.8	0.384	52.5/51.5	0.850
Chronos-S	Traffic	**65.2/54.0**	**0.032**	66.1/63.1	0.655
Chronos-S	Weather	**71.1/58.9**	**0.032**	67.8/63.7	0.432
TimesFM-2.0	ETTh1	**79.9/58.5**	**0.032**	63.0/50.9	0.133
TimesFM-2.0	Electricity	**75.7/63.2**	**0.032**	**68.6/56.1**	**0.032**
TimesFM-2.0	Traffic	82.0/74.6	0.229	76.7/72.9	0.678
TimesFM-2.0	Weather	**84.4/63.9**	**0.032**	**76.3/54.5**	**0.032**
Moirai-1.1-L	ETTh1	**76.4/61.4**	**0.032**	51.7/47.8	0.751
Moirai-1.1-L	Electricity	63.2/61.5	0.782	61.3/51.5	0.148
Moirai-1.1-L	Traffic	67.4/67.9	0.929	64.5/69.3	0.432
Moirai-1.1-L	Weather	77.6/75.1	0.678	67.0/59.9	0.058
MOMENT-Large	ETTh1	**85.7/52.8**	**0.032**	51.5/51.5	1.000
MOMENT-Large	Electricity	**67.5/55.8**	**0.032**	52.2/50.4	0.850
MOMENT-Large	Traffic	75.8/73.7	0.678	74.7/72.4	0.678
MOMENT-Large	Weather	84.3/83.6	0.928	68.6/64.2	0.432

Table 4: **CSS balanced accuracy gaps (%), pretrained minus random-weight control, original protocol, retained strictly for OOD coverage. SUPERSEDED for in-distribution significance** by Table 3's corrected rolling-CV analysis; the 4-dataset-avg. columns are historical and should not be used for in-distribution inference. **Bold** reflects the original $p<0.05$ ($n_{\text{perm}}=60$, single chronological split, since found to risk class-degenerate test folds, see §6.1). ¶: pending re-verification, see footnote 1.

		4-dataset avg.		OOD equity	
Model	**Type**	**ΔSea**	**ΔVol**	**ΔSea**	**ΔVol**
TimesFM 2.0	Dec.-only	**+6.7**	**+9.3**	**+14.1**	**+17.0**
MOMENT-Large	Masked enc.	**+8.8**	**+7.2**	+3.2	+2.8
Moirai¶	Masked enc.-dec.	**+7.7**	**+6.3**	+2.1	+1.9
Chronos-S	Quantised	**+5.5**	**+4.7**	+0.8	+0.6

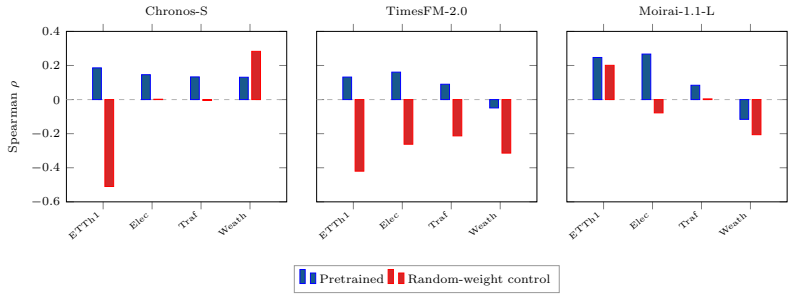

Figure 7: **PDS Spearman ρ per dataset, three primary architectures, pretrained vs. random-weight control** (same data as Table 1). Three architectures, three different random-weight-control signatures (§6.2): no single architecture-level explanation fits all three. Error bars: 95% moving-block-bootstrap CI, $n_{\text{boot}}=10{,}000$.

Geometry (PDS) Multiplicity Correction (Referenced from §6.2): Full Detail. A real reviewer finding, not a gap the authors caught unprompted: the 12-cell PDS confirmatory family (3 primary architectures × 4 datasets) was originally reported by CI alone (excludes 0 or not), unlike the 32-cell CSS family, which already received Benjamini–Hochberg correction. This applies the same correction to PDS. Real two-sided bootstrap p-values are extracted from each cell's own moving-block bootstrap distribution (the proportion of the $n_{\text{boot}}=10{,}000$ resampled statistics on the side of 0 opposite the point estimate, doubled), then BH-corrected at $q=0.05$ across all 12 (`pds_multiplicity_check.py`). Point estimates and CIs exactly reproduce Table 1 (a real cross-verification of the two independent recomputations, not assumed).

Table 5: **PDS multiplicity correction, all real**: bootstrap p-values and BH-adjusted q-values, 12-cell confirmatory Geometry family. **Bold**: significant at $q<0.05$ after correction.

Model	**Dataset**	ρ	p	q_{BH}	**CI-sig.**	**BH-sig.**
Chronos-S	ETTh1	+0.186	0.0494	0.0659	yes	no
Chronos-S	Electricity	+0.146	$< 10^{-4}$	$< 10^{-4}$	yes	**yes**
Chronos-S	Traffic	+0.133	0.0058	0.0116	yes	**yes**
Chronos-S	Weather	+0.131	0.0080	0.0137	yes	**yes**
TimesFM	ETTh1	+0.132	0.0020	0.0048	yes	**yes**
TimesFM	Electricity	+0.161	0.0202	0.0303	yes	**yes**
TimesFM	Traffic	+0.090	0.2084	0.2084	no	no
TimesFM	Weather	−0.049	0.1384	0.1661	no	no
Moirai	ETTh1	+0.247	0.0002	0.0008	yes	**yes**
Moirai	Electricity	+0.267	0.0002	0.0008	yes	**yes**
Moirai	Traffic	+0.084	0.1970	0.2084	no	no
Moirai	Weather	−0.116	0.0010	0.0030	yes	**yes**

Result: BH correction removes exactly one CI-significant cell (Chronos-S/ETTh1: p=0.0494, q_{BH}=0.0659), so the significant count changes from 9 of 12 (CI-only) to 8 of 12 (BH). Chronos-S/ETTh1, already characterized elsewhere in this project's development history as the most borderline result in the study, does not survive correction, while every other CI-significant cell does. The corrected, honest breakdown is 7 of 12 significantly positive and 1 of 12 significantly negative (Moirai/Weather), not 8 of 12 positive as an uncorrected CI reading would suggest. This does not change the paper's qualitative conclusions (real Geometry is still present in the majority of cells, still with no common architectural signature, still decoupled from Utility) but is a more statistically honest number, and the correction methodology now matches the CSS family's rather than leaving Geometry's confirmatory claim less rigorously supported than Information's.

Retrieval-Mechanism Ablation (Referenced from §6.3): Full Table and Methodology. All variants reuse `run_raf_controls.py`'s exact leakage-safe protocol (memory pool = first 80% chronologically, context-overlapping candidates excluded, k=10), changing only the distance/aggregation step, now extended to all 4 primary architectures × all 4 datasets (16 cells, up from 2 architectures/1 dataset). **Cosine/correlation**: `sklearn.pairwise_distances`, drop-in Euclidean replacement. **Learned aggregator**: for each memory window (leave-one-out, 800-window subsample), retrieve its own k-NN by embedding Euclidean distance and fit the non-negative least-squares combination of their true futures that best reconstructs its own true future; average the resulting per-rank weight vectors into one global, rank-conditioned profile, applied at real query time in place of the uniform $1/k$ mean. **DTW reranking**: real `fastdtw` (not a proxy) on windows mean-pooled to length 64 for tractability (full-resolution L=512 DTW over a ∼2,400-window pool is not tractable on this session's hardware); embedding Euclidean pre-filters each query to its 50 nearest candidates, DTW on the downsampled raw windows reranks those 50 to the final k=10. The raw-window-pre-filter row (computed per dataset, not just ETTh1) uses the identical DTW-rerank code with z set to the raw windows themselves, isolating the DTW stage's contribution from the embedding pre-filter's (`raf_alt_distances_and_aggregator.py`). **Result at full coverage**: cosine/correlation/learned-aggregator effects remain small and inconsistent in sign; the raw-window+DTW control matches or beats every embedding+DTW cell except Chronos-S on ETTh1 and Weather (14 of 16), confirming that the DTW stage, not embedding quality, drives the effect wherever it appears; the pattern found on the original 2-architecture/1-dataset slice generalizes cleanly to full coverage.

Table 6: **Retrieval-mechanism ablation, all 4 architectures × all 4 datasets, all real** (% gap recovered). Raw-window rows use the identical protocol (not the main RAF table's, hence a slightly different Euclidean baseline than Table 9).

Model	Dataset	Eucl.	Cos./Corr.	Learn. agg.	DTW
Chronos-S	ETTh1	81.7	81.4/81.3	82.2	**85.6**
Chronos-S	Electricity	18.4	19.6/19.4	19.1	32.8
Chronos-S	Traffic	19.9	20.6/20.6	21.6	38.5
Chronos-S	Weather	78.1	78.2/78.1	78.0	82.1
TimesFM	ETTh1	33.0	31.5/31.5	32.9	81.6
TimesFM	Electricity	76.3	81.4/81.5	79.1	64.6
TimesFM	Traffic	74.2	79.2/79.2	81.2	72.1
TimesFM	Weather	19.8	21.1/21.4	19.9	75.1
MOMENT-L	ETTh1	41.0	39.5/39.5	40.6	74.9
MOMENT-L	Electricity	67.6	67.3/67.3	72.3	57.0
MOMENT-L	Traffic	54.4	54.7/54.8	65.6	61.6
MOMENT-L	Weather	31.0	31.9/31.8	31.0	77.5
Moirai	ETTh1	49.5	53.9/53.9	49.6	74.6
Moirai	Electricity	56.1	60.5/60.4	59.7	55.6
Moirai	Traffic	68.8	71.5/71.5	76.7	79.4
Moirai	Weather	0.8	6.8/ 6.9	0.8	74.8
Raw window	ETTh1	84.2	–	–	83.5
Raw window	Electricity	96.2	–	–	95.4
Raw window	Traffic	95.2	–	–	94.5
Raw window	Weather	86.1	–	–	82.0

Retrieval-Quality-Only Evaluation (Referenced from §6.3): Full Detail and Methodology. Every other Utility number in this paper is downstream: it measures the mean-continuation forecast's MSE, which could mask poor neighbor identification (averaging over k=10 hides errors) or penalize good neighbor identification (a noisy H=96 continuation forecasts poorly even from the right neighbors). This evaluation isolates retrieval quality directly, against the same representation-independent oracle (true future-continuation distance, §B's RAF headline table) already used elsewhere, with no forecast in the loop: for each of the same 600 leakage-safe queries per (model, dataset), **Recall@k** is |embedding-top-k ∩ oracle-top-k|/k (chance level $k/|\text{pool}| \approx 0.004$, not 0) and **rank agreement** is the Spearman correlation between the full memory pool's embedding-distance ranking and its true-future-distance ranking (`raf_retrieval_quality.py`). Run on all 4 primary architectures × all 4 datasets plus the raw-window control (20 cells):

Table 7: **Retrieval-quality-only evaluation, all real** (RAF gap % from Table 6 shown alongside for comparison, not re-derived from these metrics).

Model	Dataset	RAF gap	Recall@10	Rank ρ
Raw window	Electricity	96.2	0.315	0.972
Raw window	Traffic	95.2	0.499	0.949
Raw window	Weather	86.1	0.056	0.630
Raw window	ETTh1	84.2	0.008	0.883
Chronos-S	ETTh1	81.7	0.076	0.579
Chronos-S	Weather	78.1	0.021	0.508
TimesFM	Electricity	76.3	0.161	0.157
TimesFM	Traffic	74.2	0.350	0.097
Moirai	Traffic	68.8	0.369	0.100
MOMENT-L	Electricity	67.6	0.226	0.037
Moirai	Electricity	56.1	0.138	0.084
MOMENT-L	Traffic	54.4	0.224	0.076
Moirai	ETTh1	49.5	0.004	0.135
MOMENT-L	ETTh1	41.0	0.006	0.247
TimesFM	ETTh1	33.0	0.002	0.031
MOMENT-L	Weather	31.0	0.008	0.082
Chronos-S	Traffic	19.9	0.037	0.014
TimesFM	Weather	19.8	0.006	0.027
Chronos-S	Electricity	18.4	0.026	0.020
Moirai	Weather	0.8	0.007	−0.049

Result, and why it does not contradict the paper's central claim. Rank agreement correlates with RAF gap across all 20 cells (ρ_s=+0.916, p=5.8×10^{-9}) and within every individual dataset ($\rho_s \geq 0.90$ in 3 of 4, 0.75 in ETTh1); Recall@10 correlates too, more weakly (ρ_s=+0.56, p=0.008 overall). Unlike PDS (a local-density–error proxy, ρ_s=+0.35, p=0.10 against the same RAF gap, §6.3), a direct retrieval-quality metric predicts Utility well. **This is expected, not a contradiction**: rank agreement is scored against the exact same oracle ranking the RAF gap is normalized against, so a high correlation is close to definitional, not a discovery that some hidden representation property predicts Utility after all. Its value is as a robustness check on the paper's central claim: PDS's failure to predict Utility (§6.3) is a property of PDS's specific density-based construction, not evidence that *no* representation-level measurement could ever track Utility: a measure that evaluates the retrieval mechanism directly, rather than through a density proxy, does. Notably, raw window's already-known Utility dominance is visible here too, but not uniformly: it has the highest rank agreement on 4 of 4 datasets, yet its Recall@10 on ETTh1 (0.008) exceeds TimesFM, Moirai, and MOMENT-L and trails only Chronos-S (0.076) despite the highest RAF gap and rank agreement on that same dataset, a reminder that Recall@10's exact-top-k criterion and the continuous rank-agreement measure can disagree even when both broadly track Utility, most likely because raw-window Euclidean distance among autocorrelated real windows produces many near-ties that a coarse top-k cutoff is sensitive to but a full rank correlation is not.

Controlled Objective Comparison and Geometry-Only Intervention (Referenced from §6.3): Full Detail. We test whether the *forecasting* objective (Assumption 4) is the cause with four small encoders, identical architecture/budget (D=64, 8 epochs), varying only the loss (Table 8, Appendix A for exact objectives): Forecasting-only; Contrastive (Control B); Forecasting+Metric; Fore-

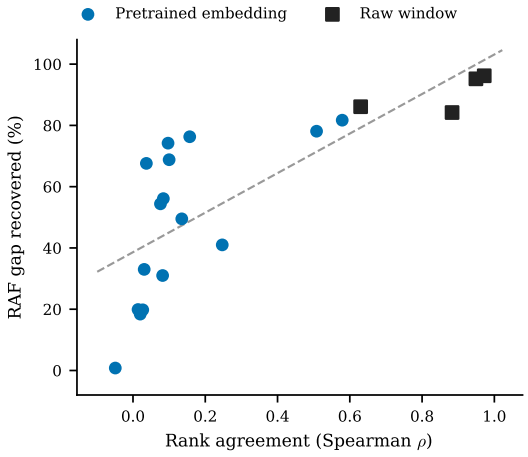

Figure 8: Rank agreement vs. RAF gap recovered, all 20 cells of Table 7 (Spearman ρ_s=+0.916; dashed line is an OLS fit shown for visual reference only). The raw-window control (squares) sits among the highest-rank-agreement points rather than off-trend.

casting+Retrieval (a differentiable relaxation of the paper's own k-NN retrieval operation, trained end-to-end). All four are trained on all four datasets (extended from ETTh1-only). **The pattern found on ETTh1 does not generalize, and this matters as much as the ETTh1 result itself**: on ETTh1, metric-flavored objectives roughly double Utility over Forecasting-only (65–66% vs. 29.4%) at some cost to Information (63–69% vs. 79.3% seasonality), but on Electricity and Traffic the ranking *inverts*: Forecasting-only and Forecasting+Retrieval reach the highest Utility (80–94%), while Contrastive and Forecasting+Metric are lowest (16–41%); Weather is mixed (Forecasting+Metric highest at 31.4%, Forecasting-only lowest at 3.9%). Geometry (PDS) tracks none of this on any dataset; point estimates are small and inconsistent in sign, none reaching significance at this sample size. No single objective is uniformly best or worst for Utility across datasets; the honest reading is that the objective–Utility relationship is itself dataset-dependent, not that metric-style training reliably helps. A geometry-only intervention (a lightweight projection on a frozen embedding, applied post-hoc) gives a further mixed result across four real cases: helps a degenerate embedding and a poor TimesFM embedding, no change for Chronos-S, and *harms* MOMENT, with no identified predictor.

A partial, exploratory diagnosis of the dataset-dependent split (not confirmatory: n=4 datasets only, reported as a candidate explanation, not a tested claim): the four objectives cluster into two families by whether their loss includes a raw forecasting-MSE term. Forecasting-only and Forecasting+Retrieval (both forecasting-anchored) track each other almost exactly across datasets (Pearson r=0.99, p=0.01) and also track the raw-window control's own per-dataset RAF level (Table 6: r=0.95, p=0.05), consistent with a forecasting-MSE term inheriting whatever makes a dataset easy or hard for raw-window retrieval itself, independent of the specific auxiliary loss added on top. Contrastive and Forecasting+Metric (both InfoNCE-anchored) also track each other (r=0.89, p=0.11, weaker) but do *not* track the raw-window pattern ($r = -0.75$, not significant), instead landing at a more dataset-invariant, moderate Utility level. Under this reading, Forecasting+Retrieval's flip from best (Electricity/Traffic) to worst (ETTh1/Weather) is not a special property of the retrieval-style loss itself; it inherits the forecasting-anchored family's per-dataset pattern, the same one Forecasting-only shows. With only 4 datasets, this cannot be confirmed statistically; extending to more datasets (or more seeds per dataset) is the direct next step to test it properly.

Contamination-Risk Screening Mechanism, Evidentiary Tiering, and Specificity Check (Referenced from §6.5). The primary confound test is whether random-weight versions of the same architectures show comparable spikes: they do not, ruling out a dataset-periodicity artefact and pointing to pretraining corpus overlap, consistent with ETTh1 inclusion documented by Meyer et al. (2025a). Mechanism: training on ETTh1 builds a dataset-specific representational vocabulary that fresh ETTh1 windows re-activate at inference, producing anomalously large embedding steps;

Table 8: **Controlled objective comparison, all 4 datasets, all real**: identical architecture/budget, varying only the loss. CSS: point estimate. PDS: n_{boot}=10,000, query-subset (n_{eff} varies by dataset stride, 5.4–20.0); no PDS point reaches significance.

Dataset	Objective	CSS s.	CSS v.	PDS ρ	RAF
ETTh1	Forecasting-only	79.3%	55.3%	+0.148	29.4%
	Contrastive (Ctrl. B)	69.3%	42.8%	−0.048	66.0%
	Forecasting+Metric	63.7%	41.6%	+0.016	65.2%
	Forecasting+Retrieval	62.5%	51.4%	−0.193	28.1%
Elec.	Forecasting-only	63.0%	55.4%	+0.216	92.5%
	Contrastive (Ctrl. B)	68.0%	54.7%	+0.032	16.3%
	Forecasting+Metric	68.2%	55.2%	+0.023	20.3%
	Forecasting+Retrieval	62.8%	54.4%	+0.109	94.3%
Traf.	Forecasting-only	60.9%	56.5%	+0.149	80.1%
	Contrastive (Ctrl. B)	49.9%	53.5%	+0.081	16.8%
	Forecasting+Metric	58.1%	52.6%	+0.127	40.9%
	Forecasting+Retrieval	65.9%	65.8%	+0.007	86.5%
Wea.	Forecasting-only	61.5%	57.9%	+0.116	3.9%
	Contrastive (Ctrl. B)	60.2%	46.5%	−0.181	24.8%
	Forecasting+Metric	57.6%	53.2%	+0.086	31.4%
	Forecasting+Retrieval	58.2%	54.0%	−0.009	17.5%

Table 9: **RAF headline**: % of the random-to-oracle MSE gap recovered (higher is better). The raw-window, model-free control beats every one of the 24 pretrained embedding runs from the 6-architecture confirmatory suite (§6.3) on all four datasets. ‡This row reports the best embedding among *all* RAF-eligible models tested at any point in this project, including OFA (tested in an earlier, broader pass, not part of the 6-architecture confirmatory suite the "24 pretrained" count above refers to), included here because it is the single best number any tested embedding achieves per dataset, and the raw window beats it too. Referenced from §6.3.

	Elec.	ETTh1	Traf.	Wea.
Raw window (no model)	**96.2**	**82.9**	**95.2**	**86.1**
Best pretrained embedding‡	88.4	81.9	74.2	78.1
(model)	OFA	Sundial	TimesFM	Chronos-S

random-weight controls have no such vocabulary, so their LTC_{step} stays flat across datasets. As a positive control, pretrained TimesFM maintains a $+19\%$ LTC_{step} advantage over its random-weight counterpart on Electricity ($p<0.05$), showing pretraining leaves a detectable geometric footprint even in models without contamination signals; on Electricity, pretrained TimesFM (LTC_{step}=4.64) falls between random (3.90) and supervised Transformer (9.92). A specificity check on two clean models (Chronos-Small, MOMENT) with two datasets structurally distinct from and outside the pretraining scope of any tested model (Exchange-rate, ILI) finds 0 of 4 model–dataset cells anomalous. Evidentiary tiering: **Observed**: anomalous benchmark-specific geometry, five models. **Inferred**: possible pretraining corpus overlap, five models. **Independently validated**: only UniTime.

Ablation Study (Referenced from §6.5). Two ablations confirm robustness of the Geometry (PDS) diagnostic to design choices. **(a) Random-weight comparison:** absolute PDS thresholds (no random-weight control) mis-classify 2 of 4 primary models; matched controls correctly identify both as non-significant. **(b) k-NN sensitivity:** PDS at $k \in \{5, 10, 20\}$ is qualitatively stable for all four primary architectures (magnitude varies by ≤ 0.04); all main results use k=10. (An earlier draft additionally reported a combined LTC+PDS contamination-screen ablation; that design is superseded by §6.5's LTC-only protocol and has been removed rather than left inconsistent with it.)

Theorem 7 (Information–Geometry Independence: a possibility theorem; referenced from §3)**.** *Let X take values in a finite set $\mathcal{X}=\{x_1, \ldots, x_n\}$ ($n\geq 3$), let Y be any random variable with $I(X;Y)>0$, and let $\phi : \mathcal{X} \to \mathbb{R}^D$ be any injective encoder. Then $I(\phi(X);Y)=I(X;Y)$: Information is exactly invariant under ϕ, since mutual information is a functional of the joint distribution over* outcomes*,*

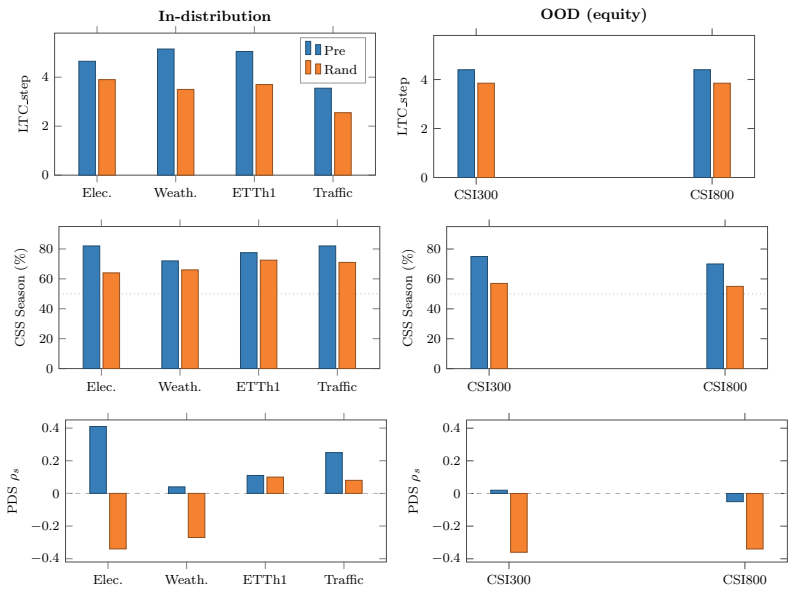

Figure 9: **OOD validation (exploratory, superseded protocol): LTC and Information (CSS) transfer; Geometry (PDS) collapses.** Pretrained (blue) vs. random-weight (red) TimesFM 2.0 on in-distribution benchmarks and OOD equity indices (CSI300/CSI800). Under this earlier protocol, $\mathrm{LTC}_{\mathrm{step}}$ and seasonality CSS both remain above the random-weight baseline under OOD shift and ρ_s^{PDS} collapses to near zero; the corrected protocol (Table 2) finds no significant transfer in either property. Referenced from §6.4.

Table 10: $\mathrm{LTC}_{\mathrm{step}}$ (trajectory richness) and contamination-risk screening. **Bold**: ETTh1/ mean(other 3) $> 3\times$ (anomalous ratio). Random-weight controls show ratios 0.5–$1.1\times$ for all five flagged models, confirming the spike is pretraining-specific rather than a dataset-periodicity artefact. Moirai row uses the corrected native encoder tap, full $N{=}3{,}000$, $L{=}512$ protocol (footnote 1); ratio $0.15\times$, the opposite direction from anomalous. †: ratio $>3\times$ on ETTh1 (same criterion as bold). The remaining screened models (Chronos-S, MOMENT-Large, Sundial, Time-MoE, LSTM, TCN) are not flagged and are omitted for space. Referenced from §6.5.

Model	**Elec.**	**Wea.**	**ETTh1**	**Traf.**	**Flag**
TimesFM (pre)	4.64	5.10	5.02	3.77	—
TimesFM (rnd)	3.90	3.52	3.70	2.55	—
Transformer	9.92	11.77	11.16	6.44	—
OFA	1.00	1.68	**17.69**†	1.62	Anomaly
UniTime	3.30	2.68	**28.90**†	3.86	Anomaly
LightGTS	1.59	0.64	**8.86**†	1.63	Anomaly
VisionTS	0.99	0.62	**6.11**†	1.20	Anomaly
TiRex	9.62	2.82	**25.53**†	8.89	Anomaly
Moirai	4.92	4.64	0.69	4.44	—

not of where those outcomes are placed in $\mathbb{R}^D$. In contrast, any local-density statistic computed purely from pairwise distances among $\{\phi(x_1), \ldots, \phi(x_n)\}$ depends entirely on ϕ: permuting which x_i occupies which geometric slot leaves $I(\phi(X); Y)$ unchanged while $\rho_s^{\mathrm{PDS}}(\phi)$ varies with the permutation, including sign reversal. Consequently, no value of $I(Z; Y)$, however large, constrains the sign or magnitude of $\rho_s^{\mathrm{PDS}}(Z)$.

Proof of Theorem 5 (PDS Positivity and OOD Collapse). **(i)** Write $\bar{d}_i$ for the embedding-space mean k-NN distance computed at $\phi(x_i)$ (an inverse density proxy: larger $\bar{d}_i$ means sparser). C2 is stated directly on embedding-space density: the conditional error surface is non-increasing in local

embedding density, i.e., non-decreasing in $\bar{d}_i$. Under C2, e_i is (up to mean-zero noise) a monotone non-decreasing function of $\bar{d}_i$. We state C2 on embedding-space density rather than input-space density deliberately: a bi-Lipschitz map bounds global distortion but does not preserve the rank ordering of local densities, so bi-Lipschitzness alone would not suffice. Spearman's ρ_s^{PDS} is the Pearson correlation of ranks; the rank correlation of any random variable $\bar{d}_i$ against a random variable $e_i{=}g(\bar{d}_i){+}\varepsilon_i$ for monotone non-decreasing g and mean-zero noise ε_i independent of $\bar{d}_i$ is non-negative, with equality only if g is constant. This gives $\rho_s^{\mathrm{PDS}} \geq 0$ in distribution under C2; any observed $\rho_s^{\mathrm{PDS}}{<}0$ therefore falsifies C2 for that (model, dataset) pair, not the theorem itself. **(ii)** If OOD embeddings distribute near-uniformly over a region of $\mathcal{M}_\theta$, standard concentration-of-measure for near-uniform point processes in high dimension gives $\mathrm{Var}(\bar{d}_i)/\mathbb{E}[\bar{d}_i]^2 \to 0$ as the effective dimension grows: k-NN distances concentrate tightly around their mean. Near-uniform embeddings also make $\bar{d}_i$ spatially homogeneous, so its finite-sample variation is (approximately) isotropic sampling noise, conditionally independent of the task's forecast-error surface; the population rank correlation between $\bar{d}_i$ and the error surface therefore vanishes, and the finite-sample Spearman estimate concentrates around 0. **(iii)** Construction: let $\phi(X)$ place m well-separated clusters on a sphere of radius R in $\mathbb{R}^D$, with within-cluster spread $r{\ll}R$ such that true error decreases with within-cluster local density by construction. $\bar{d}_i$ (computed with $k{\ll}$ cluster size) is purely a within-cluster, error-correlated quantity, giving $\rho_s^{\mathrm{PDS}}{\to}1$ as $r/R \to 0$. ρ_s^{FSS} instead averages over *random* pairs (i, j), of which a $(m{-}1)/m$ fraction are inter-cluster with pairwise distance $\approx R\sqrt{2}$ regardless of either point's local density or error; for $m{\geq}3$ this constant-distance majority swamps the within-cluster signal in the random-pair average, giving $\rho_s^{\mathrm{FSS}} \to 0$ as $m \to \infty$ with r/R fixed small. Hence configurations exist with $\rho_s^{\mathrm{PDS}} \gg \rho_s^{\mathrm{FSS}}$.

Proof of Theorem 6 (CSS and Mutual Information). Fano bound. Let $\hat{Y}^c$ be the probe's binary prediction from Z and $P_e{=}1{-}\alpha$ its error rate ($\alpha{=}\mathrm{CSS}(c)$, balanced accuracy). Fano's inequality gives $H(Y^c \mid Z) \leq H_b(P_e) + P_e \log(|\mathcal{Y}^c|{-}1) = H_b(1{-}\alpha)$, the last term vanishing since Y^c is binary ($|\mathcal{Y}^c|{=}2$). The labels are constructed by a median split (§4), so Y^c is balanced and $H(Y^c){=}1$ bit. By the mutual-information identity $I(Z; Y^c){=}H(Y^c) - H(Y^c \mid Z)$,

$$I(Z; Y^c) \;\geq\; 1 - H_b(1{-}\alpha),$$

giving the stated bound; applying it once to the pretrained probe (α_p) and once to the random-weight control (α_r) and taking the difference gives the empirical shift between two plug-in lower-bound estimates, $H_b(1{-}\alpha_r) - H_b(1{-}\alpha_p)$ (defined whenever $\alpha_p{>}\alpha_r$, since H_b is decreasing on $[1/2, 1]$); this is not a certified lower bound on the mutual-information gap itself.

Gaussian LDA equivalence. Under the stated population model ($Y^c \in \{0, 1\}$ equiprobable, $Z \mid Y^c{=}y \sim \mathcal{N}(\mu_y, \Sigma)$, shared Σ), the Bayes-optimal decision rule for a linear probe is the LDA boundary $w^\top z = w^\top(\mu_0{+}\mu_1)/2$ with $w{=}\Sigma^{-1}(\mu_1{-}\mu_0)$, and its population error rate is $\Phi(-\Delta/2)$ where $\Delta^2{=}(\mu_1{-}\mu_0)^\top\Sigma^{-1}(\mu_1{-}\mu_0){=}J_F$ is the Fisher discriminant ratio (Fisher). This error rate is strictly below $1/2$ (balanced accuracy strictly above $1/2$) iff $\Delta^2{>}0$, i.e., iff $J_F{>}0$, and both directions follow directly from Φ being strictly increasing and $\Phi(0){=}1/2$. A positive *empirical* CSS gap is finite-sample evidence for $J_F{>}0$ at the population level, not proof of it, since a regularized probe fit on a finite, blocked-CV split estimates but does not certify the population optimum.

Proof of Theorem 7 (Information–Geometry Independence). $I(\phi(X); Y){=}I(X; Y)$ holds because ϕ is injective on the finite support of X, so $\phi(X)$ and X are related by a deterministic bijection; the joint law of $(\phi(X), Y)$ is the pushforward of that of (X, Y) under a measurable bijection on the first coordinate, which leaves every probability entering $I(\phi(X); Y){=}\sum_y \sum_z p(z, y) \log \frac{p(z,y)}{p(z)p(y)}$ numerically identical to the corresponding sum for $I(X; Y)$, term for term, under the relabeling $z{=}\phi(x)$. For the second claim, fix any geometric arrangement $c_1{<}\ldots{<}c_n \in \mathbb{R}$ (slot positions) and induced density profile $d_{(1)}, \ldots, d_{(n)}$ (e.g., inverse nearest-neighbor gap). A bijection $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ assigns $x_i \mapsto c_{\pi(i)}$; $\phi_\pi(x_i){:=}c_{\pi(i)}$ is injective for every π, so $I(\phi_\pi(X); Y){=}I(X; Y)$ for every π by the argument above, independent of π. Meanwhile $\rho_s^{\mathrm{PDS}}(\phi_\pi){=}\mathrm{Spearman}(d_{(\pi(1))}, \ldots, d_{(\pi(n))}; e_1, \ldots, e_n)$ is the rank correlation between e and the density profile *re-paired* by π; taking $\pi{=}$id pairs e with the density profile in its given order and taking $\pi{=}$ the order-reversal pairs e with the density profile reversed, which for a monotone profile against a monotone e gives correlations of opposite sign (worked numerically below), completing the construction. □

Constructive example (n=5)**.** Let e=$(5, 4, 3, 2, 1)$ and let Y=$\mathbf{1}\{X \in \{x_1, x_2\}\}$, a deterministic, non-constant function of X, so $I(X;Y)$=$H_b(2/5)$>0. Fix slot positions $(0, 16, 24, 28, 30)$, giving nearest-neighbor-gap densities d=$(0.063, 0.125, 0.250, 0.500, 0.500)$. ϕ_+ assigns x_i to slot i (identity π); ϕ_- assigns x_i to slot $6-i$ (reversal π). Both satisfy $I(\phi_+(X);Y)$=$I(\phi_-(X);Y)$=$I(X;Y)$=$H_b(2/5)$≈0.971 bits exactly (both are injective relabelings of the same 5-point support). Yet $\rho_s^{\mathrm{PDS}}(\phi_+)$=$+0.975$ and $\rho_s^{\mathrm{PDS}}(\phi_-)$=$-0.975$ (verified numerically, evaluated on the mean-distance profile as in the PDS definition): identical, strictly positive Information; opposite-sign, near-maximal-magnitude Geometry. This is not a pathological edge case; it is what mutual information's invariance under relabeling always permits, for any encoder that is merely injective, which every faithful (non-collapsing) representation is.