NON-STATIONARITY IN THE EMBEDDING SPACE OF TIME SERIES FOUNDATION MODELS

Jinmyeong Choi, Brad Shook, Artur Dubrawski
Auton Lab
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{jinmyeoc, bshook, awd}@andrew.cmu.edu

ABSTRACT

Time series foundation models (TSFMs) are widely used as generic feature extractors, yet the notion of non-stationarity in their embedding spaces remains poorly understood. Recent work often conflates non-stationarity with distribution shift, blurring distinctions fundamental to classical time-series analysis and long-standing methodologies such as statistical process control (SPC). In SPC, non-stationarity signals a process leaving a stable regime—via shifts in mean, variance, or emerging trends—and detecting such departures is central to quality monitoring and change-point analysis. Motivated by this diagnostic tradition, we study how different forms of distributional non-stationarity—mean shifts, variance changes, and linear trends—become linearly accessible in TSFM embedding spaces under controlled conditions. We further examine temporal non-stationarity arising from persistence, which reflects violations of weak stationarity due to long-memory or near-unit-root behavior rather than explicit distributional shifts. By sweeping shift strength and probing multiple TSFMs, we find that embedding-space detectability of non-stationarity degrades smoothly and that different models exhibit distinct, model-specific failure modes.

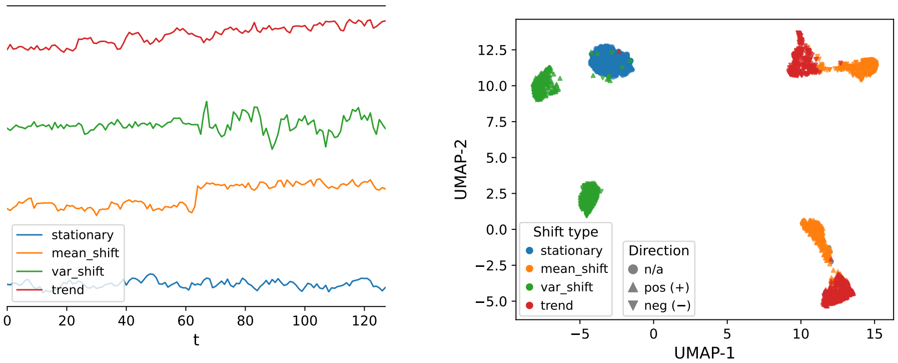
Track: Research

1 INTRODUCTION

Time series foundation models (TSFMs) are increasingly used as generic feature extractors in various downstream tasks (Ansari et al., 2024; 2025; Goswami et al., 2024; Auer et al., 2025; Talukder et al., 2024; Liang et al., 2024). Despite their growing adoption, the interpretation of *non–stationarity* within their embedding spaces remains unclear. In much of the recent literature, non–stationarity is treated synonymously with *distribution shift*, typically referring to changes in mean, variance, or the presence of trends. However, classical time-series theory distinguishes these changes from violations of weak stationarity arising from persistent temporal dependence (Kim et al., 2021; Liu et al., 2022; Fan et al., 2023).

This distinction is grounded historically in Statistical Process Control (SPC), where the goal is to detect departures from an in-control, stable process. In SPC, *mean shifts*, *variance inflation*, and *trends* are treated as distinct structural changes with different operational implications, and canonical SPC methods were explicitly designed to trade off sensitivity profiles and average run lengths (Montgomery, 2012).

In classical statistics, non-stationarity is defined through violations of weak stationarity, such as the presence of unit roots or time-varying autocovariance. In contrast, many modern approaches treat non-stationarity as a nuisance factor associated with changing marginal statistics, motivating normalization and stabilization



(a) Representative AR(1) windows under stationary, mean shift, variance shift, and trend.

(b) Two-dimensional UMAP projection of Chronos2 embeddings for the same windows.

Figure 1: Distributional non-stationarity in raw and embedding space (Chronos2).

techniques. These differing perspectives raise a fundamental question: *what aspects of non-stationarity are actually preserved or suppressed in the embedding spaces of TSFMs?*

We adopt a diagnostic approach to find the answer: under controlled autoregressive conditions, we probe how *distributional* non-stationarity—mean and variance shifts, linear trends—becomes accessible in TSFM embeddings as shift magnitude varies, and we contrast these with non-stationarity induced by *temporal persistence*. By systematically reducing the extent of shift and comparing multiple foundation models, we show that the appearance of non-stationarity in embedding spaces diminishes smoothly rather than abruptly.

2 Probing Non-Stationarity in TSFM Embeddings

We distinguish two notions of non-stationarity that are often conflated in embedding-based analyses: distributional non-stationarity and temporal non-stationarity. The first refers to explicit changes in marginal statistics over time, such as shifts in mean, changes in variance, or the presence of deterministic trends. These effects are commonly treated as distribution shift. In embedding space, a central diagnostic question is whether such deviations remain linearly accessible from learned representations.

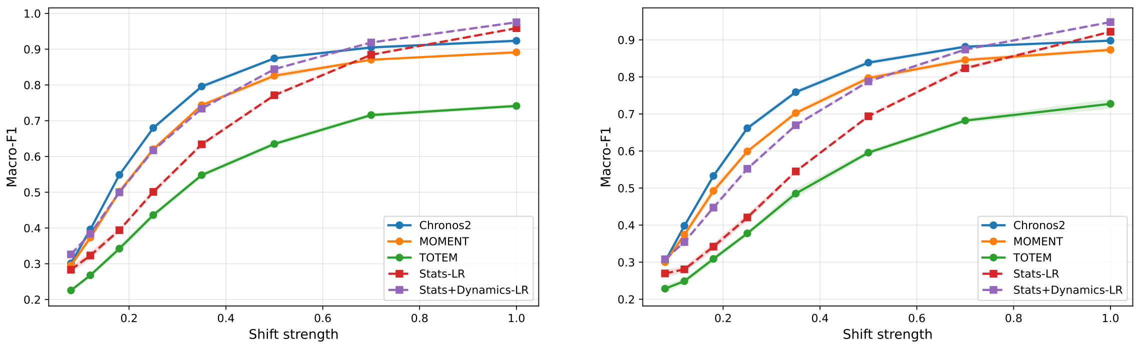
On the other hand, temporal non-stationarity is defined in classical time series analysis as a violation of weak stationarity. For the AR(1) process

$$x_t = \mu + \phi(x_{t-1} - \mu) + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2),$$

weak stationarity holds for $|\phi| < 1$ and breaks down at the unit-root boundary $\phi = 1$ (Cryer & Chan, 2008). Importantly, increasing persistence does not imply large distributional shifts within finite windows.

We use controlled AR(1) settings to analyze how these two forms of non-stationarity manifest in TSFM embeddings. We first examine distributional deviations and then contrast them with persistence-based effects.

Distributional Non-Stationarity. Figure 1 visualizes representative AR(1) windows under stationary, mean shift, variance shift, and trend conditions, together with their Chronos2 embeddings. In the raw data space, the shift types are visually distinct, and they are detectable by standard SPC techniques. In the embedding space, the data observed before and after the shift form distinct clusters, indicating that distributional deviations are preserved. However, visualization alone does not determine whether such information is linearly accessible. To quantify this, we conduct linear probing under progressively weaker shift magnitudes.



(a) Fixed persistence ($\phi = 0.6$).

(b) Random persistence ($\phi \sim U(0.3, 0.9)$).

Figure 2: Macro-F1 as a function of shift strength. Strong shifts are easily detected by both TSFM embeddings and statistics-based baselines. As shift magnitude decreases, separability degrades smoothly, with statistics-based baselines degrading more rapidly than Chronos2 and MOMENT. The qualitative behavior remains unchanged under random persistence, indicating that shift-type decodability is driven primarily by distributional structure rather than autoregressive persistence.

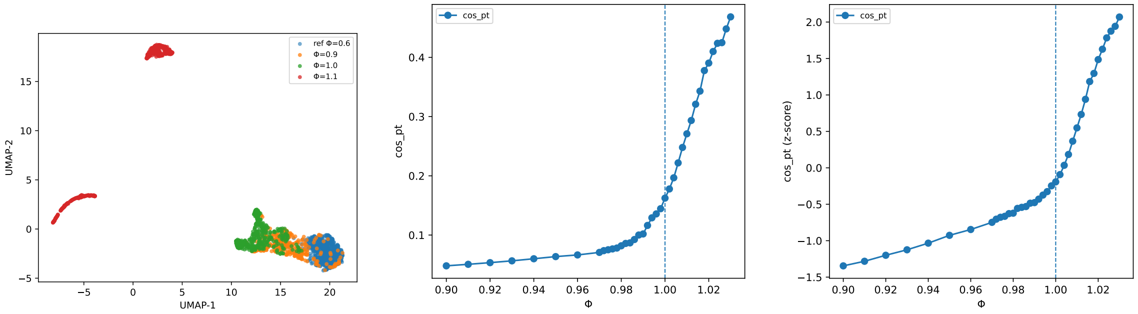
We generate window-level AR(1) sequences ($L = 128$) under four shift types: stationary (no shift), mean shift, variance shift, and trend. To control task difficulty, we introduce a shift-strength parameter $s \in (0, 1]$ that scales the magnitude of effect. For each data window, we obtain embeddings from Chronos2, MOMENT, TOTEM and train a multinomial linear classifier to predict the type of shift (Ansari et al., 2025; Goswami et al., 2024; Talukder et al., 2024). For comparison, we include two logistic regression baselines: *Stats-LR* uses only summary statistics as input features, while *Stats+Dynamics-LR* includes additional dynamical characteristics. The results are reported using Macro-F1 averaged over 5 random seeds. Detailed data generation parameters are described in Appendix A, further modeling details are in Appendix B, and full numerical results are provided in Appendix C.

As expected, overall decodability diminishes as the extent of shift is reduced. Figure 2 shows Macro-F1 as a function of shift strength. When shifts are strong ($s = 1.0$), all models—including statistics-based baselines—achieve high separability across shift types. In particular, *Stats-LR* attains near-perfect performance, indicating that large distributional shifts are easily detectable using marginal statistics alone.

As the shift magnitude decreases, its detection performance degrades gradually at rates that vary across models. While statistics-based baselines degrade rapidly, Chronos2 retains the highest separability in weaker regimes, followed by MOMENT, whereas TOTEM exhibits a markedly earlier collapse. Table 5 in the appendix quantifies these gaps in detail. These results suggest that pretrained TSFM representations capture weak distributional deviations that are not equally reflected in the handcrafted statistical features.

Temporal Non-Stationarity Figure 3a visualizes AR(1) windows with increasing ϕ and their corresponding Chronos2 embeddings. As ϕ increases, the raw signals exhibit stronger temporal dependence and longer memory. In contrast, the embedding representation evolves smoothly, without a sharp transition at the unit-root boundary ($\phi = 1$).

To quantify this behavior, we continuously vary ϕ and measure embedding discrepancies using cosine distance (see Figure 3b and 3c). Across all models, embedding distances increase monotonically with ϕ, indicating that persistence is encoded as a graded factor. This interpretation is further supported by the linear probing results (Appendix C.1), which show that ϕ can be predicted from embeddings with high accuracy,



(a) Chronos2 Embeddings (UMAP)

(b) ϕ-sweep (cosine, raw)

(c) ϕ-sweep (cosine, z-score)

Figure 3: Temporal non-stationarity in TSFM embeddings under an AR(1) persistence sweep. (a) Chronos2 embeddings shift smoothly with increasing ϕ, without a sharp transition at $\phi = 1$. (b) Cosine discrepancy (raw) increases monotonically with persistence. (c) The same trend persists after z-score normalization, indicating that the TSFM embeddings capture temporal dependence beyond scale effects.

confirming that persistence information is explicitly encoded rather than merely reflected in distance metrics. In particular, no discontinuity is observed at the unit-root boundary, although classical definitions treat $\phi = 1$ as a qualitative change in stationarity. These results suggest that TSFM embeddings do not reflect the classical stationarity boundary in a discrete manner. Instead, persistence is represented as a continuous dimension of the temporal structure. Model-specific failure modes are further analyzed via confusion matrices (Appendix C.3).

3 Discussion and Conclusion

Our results suggest that the appearance of non-stationarity in TSFM embedding spaces is neither abrupt nor model-agnostic. Instead, its linear accessibility depends on the type and magnitude of the underlying shift, as well as on the inductive biases of the representation. Detectability degrades smoothly with reduced magnitude of shift rather than collapsing at a particular boundary. Importantly, different shift types fail in distinct ways. Chronos2 and MOMENT retain access to variance-based deviations even under weak perturbations, while TOTEM exhibits an early collapse of mean shifts into the stationary class. These structured failure modes indicate that embedding spaces do not uniformly preserve all forms of relevant information. Rather, they selectively suppress or retain specific statistical attributes, such as absolute level or scale.

The persistence-based experiments further suggest that non-stationarity cannot be fully characterized in the embedding spaces by unit-root notions alone. Even when autoregressive coefficients are randomized, qualitative failure patterns persist. This highlights a gap between classical definitions of non-stationarity and how modern foundation models encode temporal structure that should be investigated further and perhaps pragmatically exploited.

The key takeaway is that TSFMs offer a unified representational interface in which diverse forms of non-stationarity—mean shifts, variance changes, trends, and persistence-based deviations—become jointly linearly accessible, enabling lightweight linear probes to recover sensitivity to a broad range of departures from stationarity without relying on bespoke, shift-specific detectors. This reframes non-stationarity detection as a problem of learned representation rather than a collection of manually engineered statistical tests. While this universality does not replace the rigorous guarantees of SPC, it complements them: TSFM embeddings can front-end classical detectors by surfacing heterogeneous change types through a shared representation, reducing the operational overhead of maintaining multiple specialized charts. Overall, this positions TSFMs as promising, domain-agnostic detectors of structured non-stationarity, capable of consolidating multiple monitoring tasks into a single embedding space and enabling more scalable approaches to forecasting and change detection in complex real-world systems.

ACKNOWLEDGMENTS

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2022-00143911, AI Excellence Global Innovative Leader Education Program). This work has been partially supported by the National Science Foundation (awards 2406231 and 2427948).

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Syndar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=gerNCVqqtR.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025. URL https://arxiv.org/abs/2510.15821.

Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. Tirex: Zero-shot forecasting across long and short horizons with enhanced in-context learning. (arXiv:2505.23719), May 2025. doi: 10.48550/arXiv.2505.23719. URL http://arxiv.org/abs/2505.23719. arXiv:2505.23719 [cs].

J.D. Cryer and K.S. Chan. *Time Series Analysis: With Applications in R*. Springer Texts in Statistics. Springer New York, 2008. ISBN 978-0-387-75959-3. URL https://books.google.com/books?id=bHke2k-QYP4C.

David A. Dickey and Wayne A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366):427–431, 1979. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2286348.

Wei Fan, Pengyang Wang, Dongkun Wang, Dongjie Wang, Yuanchun Zhou, and Yanjie Fu. Dish-ts: A general paradigm for alleviating distribution shift in time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7522–7529, 2023.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. In *International Conference on Machine Learning*, 2024.

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. URL https://openreview.net/forum?id=cGDAkQo1C0p.

Denis Kwiatkowski, Peter CB Phillips, Peter Schmidt, and Yongcheol Shin. Testing the null hypothesis of stationarity against the alternative of a unit root. how sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1-3):159–178, 1992. ISSN 0304-4076. doi: 10.1016/0304-4076(92)90104-Y.

Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 6555–6565, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671451. URL `https://doi.org/10.1145/3637528.3671451`.

Peiyuan Liu, Beiliang Wu, Yifan Hu, Naiqi Li, Tao Dai, Jigang Bao, and Shu-Tao Xia. Timebridge: Non-stationarity matters for long-term time series forecasting. *International Conference on Machine Learning*, 2025.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. 2022.

Douglas C. Montgomery. *Introduction to Statistical Quality Control*. Wiley, Hoboken, NJ, 7 edition, 2012. ISBN 9781118146811.

Nikolaos Passalis, Anastasios Tefas, Juho Kanniainen, Moncef Gabbouj, and Alexandros Iosifidis. Deep adaptive input normalization for price forecasting using limit order book data. *IEEE Transactions on Neural Networks and Learning Systems*, 2019.

Peter C. B. Phillips and Pierre Perron. Testing for a unit root in time series regression. *Biometrika*, 75(2): 335–346, 1988. ISSN 00063444. URL `http://www.jstor.org/stable/2336182`.

Lina Sjösten. A comparative study of the kpss and adf tests in terms of size and power, 2022.

Sabera J Talukder, Yisong Yue, and Georgia Gkioxari. TOTEM: TOkenized time series EMbeddings for general time series analysis. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=QlTLkH6xRC`.

Shiyu Wang, Jiawei Li, Xiaoming Shi, Zhou Ye, Baichuan Mo, Wenze Lin, Shengtong Ju, Zhixuan Chu, and Ming Jin. Timemixer++: A general time series pattern machine for universal predictive analysis. *arXiv preprint arXiv:2410.16032*, 2024.

A Data Generating Process

We generate synthetic time series using a controlled AR(1) process to study how distributional and temporal non-stationarity manifest in embedding space. All experiments are conducted at the window level with sequence length $L = 128$. Visualization of baseline and shifts are in Figure 5.

A.1 Baseline AR(1) Process

The stationary baseline is defined as an AR(1) process

$$x_t = \mu + \phi(x_{t-1} - \mu) + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2),$$

where $\mu = 0.5$, $\sigma = 0.06$, and $|\phi| < 1$ ensures weak stationarity. Unless otherwise specified, we use $\phi = 0.6$.

This baseline defines the **stationary** class.

A.2 Distributional Shift Types

We consider three forms of distributional non-stationarity applied at the window level.

Mean shift. The window is split into two halves. The first half follows the baseline process, while the second half uses a shifted mean:

$$\mu' = \mu + \Delta_\mu.$$

This produces a piecewise-constant mean within the window.

Variance shift. The window is split into two halves with different innovation variances:

$$\varepsilon_t \sim \begin{cases} \mathcal{N}(0, \sigma_1^2), & t \leq L/2, \\ \mathcal{N}(0, \sigma_2^2), & t > L/2. \end{cases}$$

The second half continues from the last value of the first half to preserve temporal continuity.

Trend. A deterministic linear trend is added to the baseline process:

$$x_t^{\text{trend}} = x_t + \text{linspace}(0, \alpha, L),$$

where α controls the slope of the trend.

A.3 Shift Strength Scaling

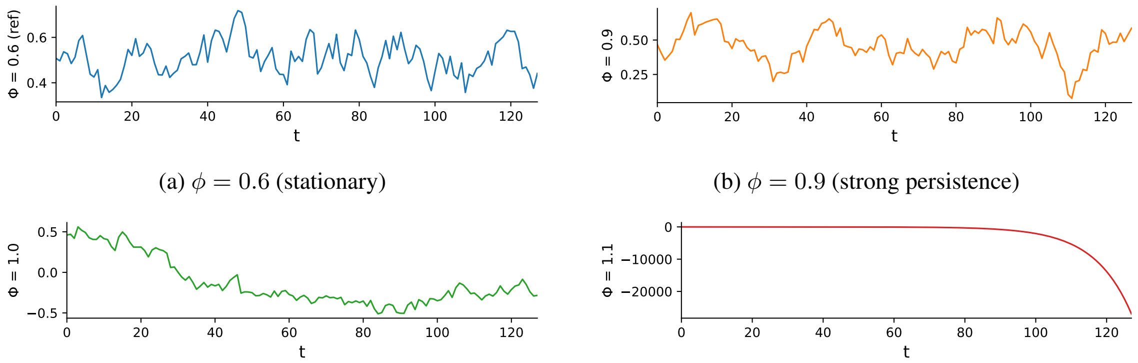
To control task difficulty, we introduce a shift-strength parameter $s \in (0, 1]$ that scales the magnitude of all distributional shifts.

Mean shift scaling.

$$\Delta_\mu \sim \text{Uniform}(0.2s,\ 0.6s) \cdot \text{sign}.$$

Trend scaling.

$$\alpha \sim \text{Uniform}(0.3s,\ 0.6s) \cdot \text{sign}.$$



(a) $\phi = 0.6$ (stationary)

(b) $\phi = 0.9$ (strong persistence)

(c) $\phi = 1.0$ (unit-root)

(d) $\phi = 1.1$ (explosive)

Figure 4: Representative AR(1) windows with increasing persistence. As ϕ increases from 0.6 to 1.1, the process transitions from a mean-reverting regime to a unit-root boundary and eventually to an explosive regime. Unlike distributional shifts, these changes modify temporal dependence rather than introducing explicit within-window shifts.

Variance shift scaling. Let $\sigma_0 = 0.06$ denote the baseline standard deviation. We sample

$$\sigma_{\text{low}} \sim \text{Uniform}(0.03, 0.06), \quad \sigma_{\text{high}} \sim \text{Uniform}(0.12, 0.20),$$

and interpolate toward σ_0 using strength s:

$$\sigma_1 = \sigma_0 + s(\sigma_{\text{low}} - \sigma_0), \quad \sigma_2 = \sigma_0 + s(\sigma_{\text{high}} - \sigma_0).$$

As $s \to 0$, both variances converge to σ_0, making the shift indistinguishable from the stationary class.

A.4 Temporal Persistence as a Nuisance Factor

To assess whether persistence drives shift separability, we also consider a nuisance setting where

$$\phi \sim \text{Uniform}(0.3, 0.9)$$

is sampled independently for each window while keeping the same distribution across all classes. This prevents label leakage while testing whether shift-type decodability depends on autoregressive persistence.

A.5 Illustration of Temporal Persistence

To illustrate temporal non-stationarity in the raw signal space, Figure 4 shows representative AR(1) windows generated with increasing values of ϕ. When $\phi = 0.6$, the process remains in the weakly stationary regime and fluctuates around the mean. As ϕ increases to 0.9, temporal dependence becomes stronger and the trajectory evolves more slowly. At the unit-root boundary $\phi = 1.0$, the process no longer reverts to the mean and instead exhibits random-walk-like behavior. For $\phi = 1.1$, the process becomes explosive, producing rapidly diverging trajectories.

These examples highlight that temporal non-stationarity differs qualitatively from the distributional shifts considered above. Rather than introducing an explicit change in mean, variance, or trend within a window, varying ϕ changes the persistence structure of the process itself.

Table 1: Parameter ranges used for synthetic data generation. Shift magnitudes are scaled by strength $s \in (0, 1]$ to control task difficulty.

Category	Parameter	Value / Range
Baseline AR(1)		
	Window length L	128
	Mean μ	0.5
	Innovation std. σ	0.06
	Persistence ϕ (fixed)	0.6
	Persistence ϕ (nuisance)	Uniform$(0.3, 0.9)$
Mean Shift		
	Mean change Δ_μ	Uniform$(0.2s, 0.6s) \cdot$ sign
	Shift structure	Half-window change
Variance Shift		
	Baseline std. σ_0	0.06
	Low variance	Uniform$(0.03, 0.06)$
	High variance	Uniform$(0.12, 0.20)$
	Strength interpolation	$\sigma = \sigma_0 + s(\sigma_{\text{raw}} - \sigma_0)$
	Shift structure	Half-window change (continuous)
Trend		
	Slope α	Uniform$(0.3s, 0.6s) \cdot$ sign
	Trend form	Additive linear ramp
Shift Strength		
	Strength levels	$\{1.0,\ 0.7,\ 0.5,\ 0.35,\ 0.25,\ 0.18,\ 0.12,\ 0.08\}$
	Interpretation	$s = 1$: strongest shift, $s \to 0$: indistinguishable

B MODELS

We evaluate three representative time series foundation models (TSFMs): Chronos2 (Ansari et al., 2025), MOMENT (Goswami et al., 2024), and TOTEM (Talukder et al., 2024). These models were selected to span diverse architectural paradigms and training objectives while enabling consistent extraction of window-level embeddings. Additionally, we evaluate two non-TSFM baselines *Stats-LR* and *Stats+Dynamics-LR*.

B.1 CHRONOS2

Chronos2 (Ansari et al., 2025) is a pretrained time series foundation model designed for universal forecasting across univariate and multivariate settings.[1] It extends the Chronos family with group-attention mechanisms that enable cross-series information sharing and in-context learning. The model processes normalized input

[1] https://github.com/amazon-science/chronos-forecasting

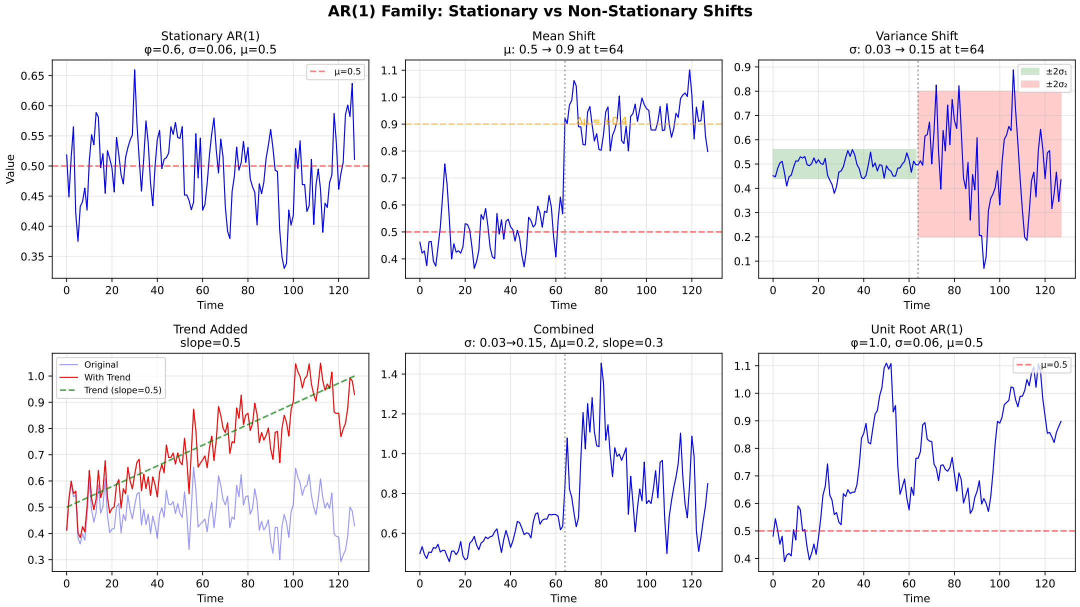


Figure 5: Representative AR(1) windows used for distributional non-stationarity experiments ($L = 128$): stationary, mean shift, variance shift, and trend.

sequences by segmenting them into patches, mapping them to embeddings, and applying a transformer stack with alternating time and group attention layers.

This design supports strong zero-shot forecasting and cross-domain generalization by leveraging shared temporal patterns across related series. Because Chronos2 provides fixed-length representations for input windows, it is well suited for embedding-based analysis.

Usage in this work. We use Chronos2 as an encoder to obtain window embeddings via mean pooling over token representations.

B.2 MOMENT

MOMENT (Goswami et al., 2024) is a family of open-source foundation models for general-purpose time series analysis, trained via large-scale multi-dataset pretraining.[2] It learns representations through self-supervised objectives such as masked reconstruction, enabling a single model to support forecasting, classification, anomaly detection, and imputation tasks.

By reconstructing masked segments, MOMENT learns embeddings that preserve salient temporal structure while promoting invariances to nuisance variations. This reconstruction-oriented objective makes MOMENT particularly suitable for studying which aspects of non-stationarity are preserved or suppressed in representation space.

[2] `https://github.com/moment-timeseries-foundation-model/moment`

Usage in this work. We use MOMENT as an embedder only, removing the reconstruction head and extracting latent representations directly from the encoder.

B.3 TOTEM

TOTEM (TOkenized Time Series EMbeddings) (Talukder et al., 2024) is a generalist time series foundation model based on discrete tokenization of time series data.[3] It employs a VQ-VAE-style architecture with convolutional encoders and a learned codebook to produce discrete token representations that can be used across forecasting, anomaly detection, and imputation tasks.

By representing time series as discrete tokens, TOTEM enables cross-domain training and strong zero-shot performance across diverse tasks. Its forecasting-oriented training encourages representations that emphasize temporal dynamics and structural patterns.

Usage in this work. We use the forecasting variant of TOTEM and extract embeddings from its encoder to obtain window-level representations.

B.4 Baselines

Our two baselines derive features from the time series without using TSFMs. The first baseline, *Stats-LR*, includes the following features: quantiles, mean, standard deviation, minimum, maximum, range, interquartile range, absolute mean, root mean square, mean square root, skewness, and kurtosis. The second baseline, *Stats+Dynamics-LR*, includes those same features in addition to: mean of first differences, standard deviation of first differences, slope of linear regression, and autocorrelations at lags 1, 2, and 3. These features train a logistic regression model to perform shift type classification.

B.5 Why These TSFMs?

We selected these models for three primary reasons.

(1) Comparable embedding extraction. All three models provide encoder outputs that can be converted into fixed-length window embeddings without task-specific fine-tuning, enabling consistent representation-level diagnostics.

(2) Diverse inductive biases. Chronos2 emphasizes cross-series forecasting and probabilistic modeling, MOMENT focuses on reconstruction-driven representation learning, and TOTEM leverages discrete tokenization and forecasting objectives. This diversity allows us to examine how training objectives shape the encoding of non-stationarity.

(3) Relevance to modern TSFMs. These models are representative of current design trends in time series foundation models and are widely used across forecasting and representation-learning tasks.

[3] `https://github.com/SaberaTalukder/TOTEM`

Table 2: Regression performance for predicting AR(1) persistence ϕ from embeddings. Lower MAE and higher r and R^2 indicate better preservation of temporal dependence.

Model	**MAE** $\downarrow$	**Pearson** $r \uparrow$	$R^2 \uparrow$
Chronos2	0.041	0.982	0.964
MOMENT	0.058	0.965	0.931
TOTEM	0.121	0.812	0.659
Stats-LR	0.089	0.901	0.811
Stats+Dynamics-LR	0.061	0.954	0.910

C Experiment Results

C.1 Temporal Persistence Regression

To complement the distributional shift analysis, we evaluate how well TSFM embeddings preserve temporal persistence. Specifically, we regress the AR(1) coefficient ϕ from window-level embeddings and measure prediction accuracy.

For each window, we generate AR(1) sequences with $\phi \in [0.3, 1.1]$ and train a linear regressor to predict ϕ from the embedding. We report mean absolute error (MAE), Pearson correlation (r), and coefficient of determination (R^2).

C.1.1 Results

Interpretation. Chronos2 embeddings preserve temporal persistence most faithfully, achieving the lowest MAE and highest correlation. MOMENT also captures persistence well, though with slightly reduced accuracy. In contrast, TOTEM exhibits substantially weaker alignment with ϕ, indicating that its embeddings encode persistence less explicitly.

Interestingly, the statistics-based baselines achieve competitive performance, suggesting that persistence can be partially recovered from low-order dynamics. However, TSFMs—particularly Chronos2—provide a more precise and stable encoding of temporal dependence.

C.2 Sequence Length Ablation

To test whether shift-type decodability depends on the choice of window length, we repeat the shift-type linear probing experiment across multiple sequence lengths $L \in \{64, 128, 256, 512\}$. We report Macro-F1 for representative shift strengths $s \in \{1.0, 0.25, 0.12\}$. The default setting used in the main paper is $L = 128$.

C.2.1 Fixed Persistence ($\phi = 0.6$)

Table 3 summarizes Macro-F1 across sequence lengths under fixed persistence ($\phi = 0.6$). Longer windows consistently improve separability for all methods, reflecting the benefit of additional temporal context. Importantly, the qualitative model ranking is unchanged across L, and the weak-shift regime ($s = 0.12$) continues to reveal substantial robustness differences across representations.

Table 3: Sequence-length ablation for shift-type probing under fixed persistence ($\phi = 0.6$). Values are Macro-F1 (mean±std over 5 seeds).

Strength	**Model**	$\mathbf{L = 64}$	$\mathbf{L = 128}$	$\mathbf{L = 256}$	$\mathbf{L = 512}$
Strong Shifts ($s = 1.0$)					
TSFMs	Chronos2	0.891±0.007	0.924±0.005	0.948±0.005	0.969±0.002
	MOMENT	0.859±0.007	0.891±0.004	0.914±0.004	0.933±0.003
	TOTEM	0.711±0.005	0.741±0.005	0.760±0.007	0.780±0.008
Baselines	Stats-LR	0.927±0.002	0.959±0.001	0.978±0.001	0.990±0.001
	Stats+Dynamics-LR	0.962±0.002	0.975±0.002	0.983±0.001	0.990±0.000
Moderate Shifts ($s = 0.25$)					
TSFMs	Chronos2	0.560±0.007	0.680±0.005	0.801±0.005	0.891±0.003
	MOMENT	0.523±0.006	0.620±0.006	0.713±0.006	0.789±0.009
	TOTEM	0.379±0.008	0.436±0.007	0.501±0.011	0.577±0.003
Baselines	Stats-LR	0.435±0.007	0.501±0.006	0.569±0.007	0.635±0.005
	Stats+Dynamics-LR	0.530±0.005	0.617±0.003	0.692±0.002	0.757±0.005
Weak Shifts ($s = 0.12$)					
TSFMs	Chronos2	0.347±0.005	0.396±0.007	0.500±0.007	0.643±0.003
	MOMENT	0.339±0.008	0.373±0.011	0.442±0.003	0.501±0.005
	TOTEM	0.251±0.007	0.268±0.005	0.297±0.006	0.349±0.007
Baselines	Stats-LR	0.293±0.007	0.323±0.012	0.356±0.013	0.377±0.006
	Stats+Dynamics-LR	0.345±0.005	0.384±0.005	0.442±0.006	0.503±0.008

C.2.2 RANDOM PERSISTENCE ($\phi \sim U(0.3, 0.9)$)

Table 4 repeats the same evaluation under random persistence. The same qualitative trends hold: longer windows improve decodability, while the relative robustness ordering across representations remains stable. This supports the conclusion that the observed failure modes are not artifacts of a particular window length or a fixed choice of ϕ.

Interpretation. Chronos2 embeddings preserve temporal persistence most faithfully, achieving the lowest MAE and highest correlation. MOMENT also captures persistence well, though with slightly reduced accuracy. In contrast, TOTEM exhibits substantially weaker alignment with ϕ, indicating that its embeddings encode persistence less explicitly.

C.3 CONFUSION MATRIX ANALYSIS

To better understand model-specific failure modes, we examine confusion matrices for shift-type classification at representative shift strengths. Rows correspond to true labels and columns to predicted labels.

Figures 6 and 7 show results for fixed persistence ($\phi = 0.6$) and random persistence ($\phi \sim U(0.3, 0.9)$), respectively.

Table 4: Sequence-length ablation for shift-type probing under random persistence ($\phi \sim U(0.3, 0.9)$). Values are Macro-F1 (mean±std over 5 seeds).

Strength	Model	$\mathbf{L = 64}$	$\mathbf{L = 128}$	$\mathbf{L = 256}$	$\mathbf{L = 512}$
Strong Shifts ($s = 1.0$)					
TSFMs	Chronos2	0.866±0.005	0.898±0.003	0.927±0.003	0.953±0.003
	MOMENT	0.836±0.005	0.873±0.005	0.903±0.004	0.927±0.004
	TOTEM	0.702±0.006	0.727±0.013	0.752±0.006	0.767±0.008
Baselines	Stats-LR	0.899±0.003	0.922±0.002	0.940±0.002	0.952±0.002
	Stats+Dynamics-LR	0.929±0.002	0.948±0.002	0.961±0.001	0.969±0.001
Moderate Shifts ($s = 0.25$)					
TSFMs	Chronos2	0.541±0.007	0.662±0.007	0.788±0.004	0.871±0.004
	MOMENT	0.508±0.006	0.599±0.006	0.687±0.004	0.760±0.006
	TOTEM	0.346±0.006	0.378±0.006	0.430±0.005	0.484±0.007
Baselines	Stats-LR	0.386±0.010	0.421±0.010	0.448±0.008	0.470±0.010
	Stats+Dynamics-LR	0.476±0.009	0.552±0.006	0.616±0.004	0.668±0.006
Weak Shifts ($s = 0.12$)					
TSFMs	Chronos2	0.345±0.006	0.398±0.007	0.493±0.007	0.630±0.003
	MOMENT	0.340±0.006	0.374±0.009	0.431±0.004	0.479±0.009
	TOTEM	0.244±0.004	0.249±0.007	0.268±0.005	0.285±0.004
Baselines	Stats-LR	0.270±0.007	0.281±0.008	0.302±0.011	0.321±0.015
	Stats+Dynamics-LR	0.331±0.007	0.354±0.005	0.394±0.004	0.430±0.009

D Related Work: Perspectives on Non-Stationarity

Non-stationarity has long been recognized as a central challenge in time series analysis, yet its interpretation and treatment vary substantially across the literature. In classical statistics, non-stationarity is formalized through precise definitions, such as violations of weak stationarity or the presence of unit roots (Dickey & Fuller, 1979; Kwiatkowski et al., 1992; Phillips & Perron, 1988; Sjösten, 2022). In contrast, many modern deep learning approaches adopt a broader and often less explicit view, frequently using non-stationarity as an umbrella term for diverse forms of temporal variation. This section reviews how non-stationarity has been conceptualized and addressed in recent time series modeling work, with a focus on the assumptions implicit in these approaches.

Non-stationarity as distribution shift. A large body of recent work implicitly equates non-stationarity with distributional change over time. For example, several studies characterize non-stationary time series as those exhibiting continuously changing statistical properties or joint distributions, which are argued to hinder predictability. Within this perspective, non-stationarity is treated primarily as a nuisance factor that should be mitigated to simplify learning. Normalization-based techniques are commonly proposed to stabilize time series by removing shifts in mean, variance, or scale (Passalis et al., 2019; Kim et al., 2021). More structured approaches explicitly model time-varying normalization and denormalization processes to compensate for distributional drift during forecasting (Liu et al., 2022).

Table 5: Shift-type linear probing under varying shift strengths ($\phi = 0.6$ and $\phi \sim U(0.3, 0.9)$).

Type	**Model**	$\mathbf{s = 1.0}$	$\mathbf{s = 0.25}$	$\mathbf{s = 0.12}$
Fixed Shift *($\phi = 0.6$)*				
TSFMs	Chronos2	0.924 ± 0.005	**0.680 ± 0.005**	**0.396 ± 0.007**
	MOMENT	0.891 ± 0.004	0.620 ± 0.006	0.373 ± 0.011
	TOTEM	0.741 ± 0.005	0.436 ± 0.007	0.268 ± 0.005
Baselines	Stats-LR	0.959 ± 0.001	0.501 ± 0.006	0.323 ± 0.012
	Stats+Dynamics-LR	**0.975 ± 0.002**	0.617 ± 0.003	0.384 ± 0.005
Random Shift *($\phi \sim U(0.3, 0.9)$)*				
TSFMs	Chronos2	0.898 ± 0.003	**0.662 ± 0.007**	**0.398 ± 0.007**
	MOMENT	0.873 ± 0.005	0.599 ± 0.006	0.374 ± 0.009
	TOTEM	0.727 ± 0.013	0.378 ± 0.006	0.249 ± 0.007
Baselines	Stats-LR	0.922 ± 0.002	0.421 ± 0.010	0.281 ± 0.008
	Stats+Dynamics-LR	**0.948 ± 0.002**	0.552 ± 0.006	0.354 ± 0.005

Related ideas have also been explored in representation-space analyses. For instance, Dish-TS (Fan et al., 2023) distinguishes between *intra-space shift*, referring to temporal changes within a single representation space, and *inter-space shift*, describing misalignment across representations learned under different temporal regimes. These formulations further reinforce the view of non-stationarity as a form of distribution shift that disrupts stable representation learning.

Questioning over-stabilization. While normalization-based approaches have shown empirical success, a growing line of work challenges the assumption that non-stationarity should always be removed. Several recent studies argue that aggressive stabilization may inadvertently discard informative temporal structure (Liu et al., 2022; Wang et al., 2024; Liu et al., 2025). From this perspective, non-stationarity is not merely a source of noise but may encode meaningful dynamics that are essential for predictive tasks.

Positioning of this work. Our work differs from the above lines of research in both scope and intent. Rather than advocating for or against normalization, or proposing a new architectural solution, we adopt a diagnostic perspective. We ask how different meanings of non-stationarity—distributional shifts versus temporal dependence violations—manifest in the embedding spaces of time series foundation models. By disentangling these notions and analyzing their representation-level effects under multiple metrics, we aim to clarify what is preserved, suppressed, or entangled in learned embeddings, and how prior assumptions about non-stationarity shape model behavior.

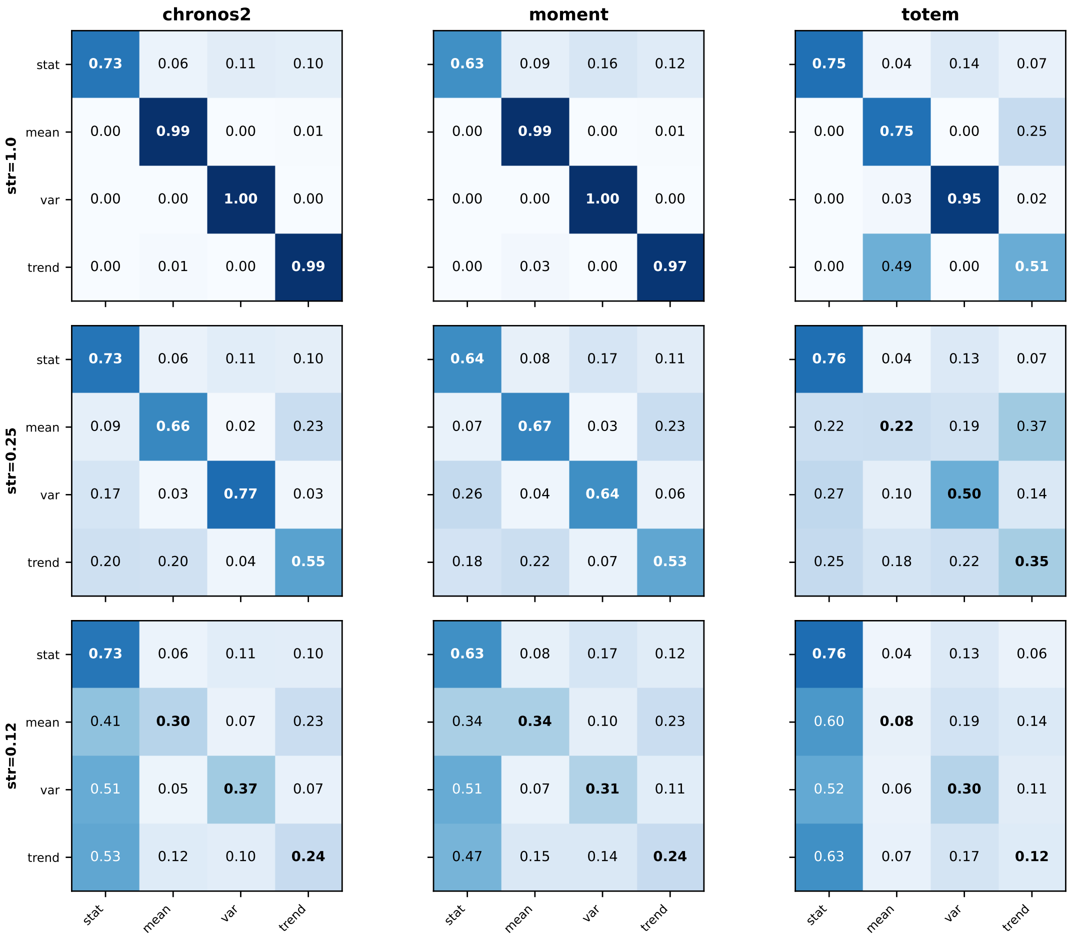


Figure 6: Confusion matrices for shift-type probing at sequence length $L = 128$ with fixed persistence ($\phi = 0.6$). As shift strength decreases (top to bottom), errors increase and reveal model-specific failure modes. Chronos2 and MOMENT retain strong diagonal structure for variance shifts, whereas TOTEM exhibits substantial collapse of mean shifts into the stationary class.

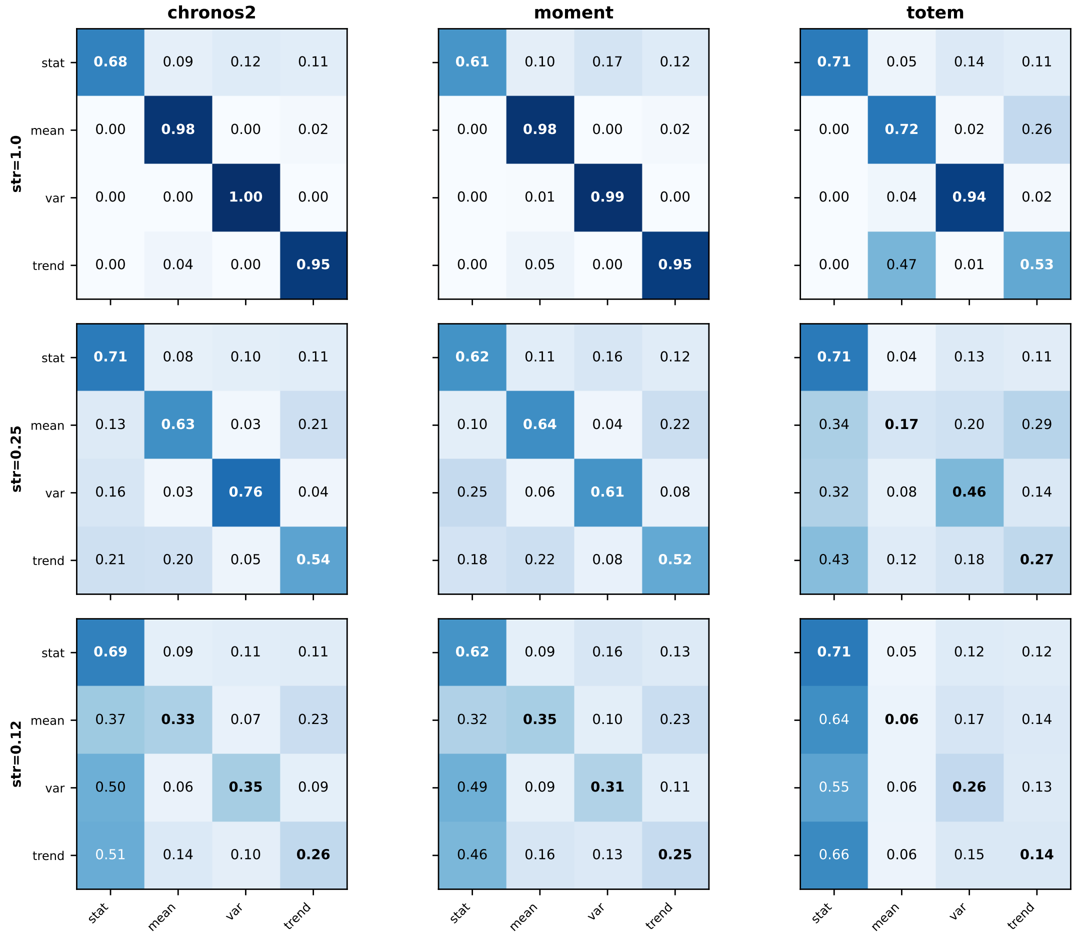


Figure 7: Confusion matrices under random persistence ($\phi \sim U(0.3, 0.9)$). Qualitative failure patterns remain consistent with the fixed-ϕ setting, indicating that shift-type confusion is driven by distributional structure rather than persistence.