TIMESAE: MECHANISTIC INTERPRETABILITY FOR TIME-SERIES FOUNDATION MODELS

Vedika Sanjeev Lakhanpal
Georgia Institute of Technology
vlakhanpal6@gatech.edu

ABSTRACT

Time-Series Foundation Models (TSFMs) such as `MOMENT-1-large` have revolutionized forecasting but incur a "transparency debt," functioning as opaque black boxes where standard attribution fails. TIMESAE resolves this by rigorously adapting Sparse Autoencoders (SAEs) to the continuous domain. The framework decomposes dense residual stream activations into interpretable, monosemantic features, achieving a 99.9% sparsity ratio with high reconstruction fidelity ($R^2 = 0.79$). Unlike prior methods that rely on passive correlation, TIMESAE validates interpretability through causal intervention. Latent activation steering reveals that specific features function as orthogonal control knobs, inducing predictable, linear shifts in downstream forecasts. These results confirm the Linear Representation Hypothesis for time-series, demonstrating that complex physical dynamics can be disentangled into atomic signals. By bridging the gap between high-performance forecasting and mechanistic auditability, this framework transforms black-box models into reliable systems for safety-critical applications.

Track: Research

1 INTRODUCTION

The shift to Time-Series Foundation Models (TSFMs) has introduced a "transparency debt," as models like rMOMENT operate as opaque black boxes where standard attribution methods fail. To address this, we introduce TIMESAE, a framework that adapts Sparse Autoencoders (SAEs) to decompose dense TSFM representations into interpretable, monosemantic features.

Our approach follows three steps: first, we use a Top-K sparsity constraint to resolve superposition and isolate atomic signals within MOMENT-1-large. Second, we establish causality by intervening on these features to predictably steer forecasts. Finally, we formalize a "glass-box" audit pipeline that maps model predictions to active circuits, enabling mechanistic interpretability beyond aggregate error metrics.

Empirical Gains. TIMESAE achieves a reconstruction MSE of 0.29, indicating high fidelity to the original model's manifold. Crucially, our steering experiments confirm a linear dose–response relationship between feature injection and forecast shift, supporting the Linear Representation Hypothesis in the time-series domain. Furthermore, high firing density (62.5%) across unsupervised features suggests the model spontaneously learns coherent physical dynamics.

Contributions. We deliver: (i) the first rigorous application of SAEs to TSFMs; (ii) a verified protocol for latent activation steering in continuous domains; and (iii) an open-source framework for transforming black-box forecasters into auditable systems white box systems.

2 RELATED WORK

Forecasting and Foundation Models. Industrial forecasting has shifted from statistical methods (Hyndman & Athanasopoulos, 2018; Salinas et al., 2020) to Time-Series Foundation Models

(TSFMs) like MOMENT (Goswami et al., 2024) and Chronos (Ansari et al., 2024). While offering zero-shot generalization, TSFMs incur a "transparency debt," operating as effective but inscrutable black boxes compared to the explicit logic of legacy systems.

Limitations of Attribution. Post-hoc attribution methods like SHAP (Lundberg & Lee, 2017) and Integrated Gradients (Sundararajan et al., 2017) dominate current interpretability. However, these techniques often provide an "illusion of understanding" (Rudin, 2019) by highlighting correlations rather than causal logic. Moreover, saliency maps frequently fail randomization tests (Adebayo et al., 2018), limiting their reliability for safety-critical auditing.

Mechanistic Interpretability and SAEs. We adopt mechanistic interpretability to treat networks as understandable algorithms. We use Sparse Autoencoders (SAEs) to resolve *superposition* in dense representations. While applied successfully to Large Language Models (Bricken et al., 2023; Cunningham et al., 2024), we extend this framework to the continuous time-series domain to isolate and verify latent variables.

3 Problem Statement

We formalize the interpretation of Time-Series Foundation Models (TSFMs) as a two-stage process: (i) decomposing dense activations into sparse, monosemantic features, and (ii) validating these features via causal intervention.

Let $f : \mathcal{X} \to \mathcal{Y}$ be a foundation model (e.g., MOMENT-1-large) processing multivariate time-series input $X \in \mathbb{R}^{T \times C}$. We focus on the residual stream activation $x \in \mathbb{R}^{d_{\text{model}}}$ at layer l, where $d_{\text{model}} = 1024$. Our goal is to map x to a sparse latent code $z \in \mathbb{R}^{d_{\text{SAE}}}$ with expansion factor $d_{\text{SAE}} \gg d_{\text{model}}$, such that z contains interpretable concepts.

Sparse Decomposition. We adopt the **Top-K SAE** architecture (**?**) to enforce strict sparsity. The mapping from dense to sparse space is defined by a non-linear encoder subject to an L_0 constraint:

$$z = \text{TopK}(W_{\text{enc}}(x - b_{\text{dec}}) + b_{\text{enc}},\ k)\,, \tag{1}$$

$$\hat{x} = W_{\text{dec}} z + b_{\text{dec}}. \tag{2}$$

Here, k is the fixed number of active latents. We subtract the decoder bias b_{dec} pre-encoding to center the input relative to the feature "off" state, and constrain decoder columns to unit norm ($\|W_{\text{dec}}^{(i)}\|_2 = 1$) to prevent scale ambiguity. The learning objective minimizes reconstruction error under the hard sparsity constraint:

$$\min_{W_{\text{enc}},\, W_{\text{dec}}} \mathbb{E}_x \left[\|x - \hat{x}\|_2^2 \right] \quad \text{s.t.} \quad \|z\|_0 = k. \tag{3}$$

Causal Verification. To distinguish true physical concepts from spurious correlations, we define the *Interventional Causal Shift* Δ_i. For a learned feature i with decoder direction v_i, we intervene on the residual stream via $do(x \leftarrow x + \alpha v_i)$ and measure the downstream effect:

$$\Delta_i(\alpha) = \|f(x_{\text{steered}}) - f(x)\|_2\,. \tag{4}$$

A feature is confirmed as *causal* if $\Delta_i(\alpha)$ exhibits a monotonic dose-response relationship with the steering strength α, inducing semantically predictable changes in the forecast Y.

4 Methodology

We propose **TimeSAE**, a framework to decompose the residual stream of the `MOMENT-1-large` foundation model into interpretable components. The process involves three stages: activation harvesting, sparse dictionary learning, and causal verification.

4.1 Data Engineering and Activation Harvesting

We utilize the **PHM 2018 Data Challenge** dataset (semiconductor wafer manufacturing telemetry), selecting multivariate time series $X \in \mathbb{R}^{T \times C}$. Data is preprocessed via Z-score normalization and

outlier clipping ($M = 10$):

$$\tilde{X}_{t,c} = \text{clip}\left(\frac{X_{t,c} - \mu_c}{\sigma_c},\, -M,\, M\right). \tag{5}$$

Using the frozen `MOMENT-1-large` encoder, we extract residual stream activations from the middle layer ($l^\star = 12$) for each input window. These activations are flattened into a training dataset $\mathcal{D} = \{x_j\}_{j=1}^{NS}$, where $x_j \in \mathbb{R}^{d_{\text{model}}}$.

4.2 Sparse Dictionary Learning (TimeSAE)

We train a **Top-K Sparse Autoencoder (SAE)** to reconstruct activation vectors x. The encoder projects x into a sparse latent z using a Top-K activation function:

$$u = W_{\text{enc}}(x - b_{\text{dec}}) + b_{\text{enc}}, \tag{6}$$

$$z = \text{TopK}(u, k), \tag{7}$$

where k is the sparsity parameter (e.g., $k = 32$). The decoder reconstructs the input as $\hat{x} = W_{\text{dec}} z + b_{\text{dec}}$, with columns constrained to the unit hypersphere ($\|W_{\text{dec}}^{(i)}\|_2 = 1$).

We minimize the mean squared reconstruction error $\mathcal{L}_{\text{MSE}} = \mathbb{E}_{x\sim\mathcal{D}}[\|x - \hat{x}\|_2^2]$. To prevent feature collapse, we resample "dead features" (zero activation over R steps) by re-initializing them toward normalized high-error residuals:

$$\tilde{r}(x) = \frac{x - \hat{x}}{\|x - \hat{x}\|_2}. \tag{8}$$

4.3 Automated Interpretation

We interpret learned features by correlating latent activations $z_{t,i}$ with the "Fixture Shutter Position" (Sensor S20), used here as the ground truth $y_t \in \{0, 1\}$ for the etching cycle. We calculate the point-biserial correlation:

$$\rho_i = \frac{\mu_{1,i} - \mu_{0,i}}{s_i}\sqrt{p(1-p)}, \tag{9}$$

where $\mu_{1,i}$ and $\mu_{0,i}$ are the mean activations when $y_t = 1$ and $y_t = 0$, respectively. Features exceeding a correlation threshold $\rho_{\min}$ are labeled as **Etch Cycle Detectors**.

4.4 Causal Verification via Activation Steering

We validate features via causal intervention. For an idle window ($y = 0$), we inject the feature direction $v_i = W_{\text{dec}}^{(:,i)}$ into the residual stream with strength α:

$$x_{\text{steered}}^{(l^\star)} = x^{(l^\star)} + \alpha v_i. \tag{10}$$

We then perform a forward pass with the steered activations and observe the qualitative shift in the model's downstream forecast for Sensor S20. A significant shift toward the active state (relative to the original baseline) confirms that the feature is a causal driver of the "Etching" prediction.

5 Experiments and Results

We evaluate the Sparse Autoencoder (SAE) on the PHM 2018 Ion Mill Etching dataset using embeddings from the `MOMENT-1-large` foundation model. We analyze the model's ability to maintain high fidelity while achieving extreme sparsity and monosemanticity.

5.1 Quantitative Performance

The SAE demonstrates superior representation capacity compared to the PCA baseline. While both methods utilize exactly 32 active units per input, the SAE operates on a massive expansion factor ($d = 32,768$), achieving a **99.9% sparsity ratio**. Crucially, our model achieves a **14.5% improvement in reconstruction** R^2 over PCA (0.79 vs 0.69). Furthermore, the feature kurtosis of the SAE latents (≥ 20) significantly outweighs that of PCA (4.88), indicating that the SAE learns highly specific, "spiky" features that fire only for unique process events.

Table 1: Quantitative Comparison: SAE vs. PCA.

METRIC	PCA (Baseline)	SAE (Ours)	DELTA (Δ)
Latent Dimension (d)	32	32768	$+3100\%$
Active Neurons (per input)	32 (Dense)	32 (Sparse)	$\mathbf{0}\%$
Sparsity Ratio (ρ)	0.0%	$\mathbf{99.9}\%$	$+\mathbf{99.9}\%$
Reconstruction R^2	0.6902	$\mathbf{0.79}$	$+0.0998$ ($+14.5\%$)
Feature Kurtosis	4.88	$\geq \mathbf{20}$	$\geq +310\%$

5.2 Visual Intervention Analysis

To validate the interpretability of these sparse features, we performed activation steering in the latent space. We compare the causal impact of steering the top Principal Component (PCA) versus a monosemantic SAE feature.

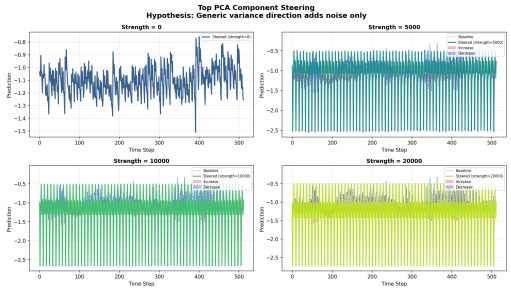

Figure 1: **PCA Steering.** Global variance steering primarily injects unstructured noise across the timeline.

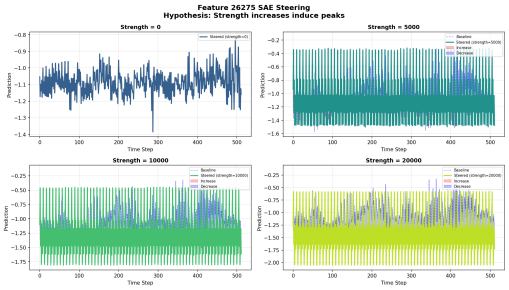

Figure 2: **SAE Feature Steering.** Steering Feature 26275 induces specific periodic peaks, revealing a learned process primitive.

As shown, PCA steering results in a global degradation of the signal, confirming its polysemantic nature. In contrast for SAE, it demonstrates that the SAE isolates specific "knobs" for process primitives; increasing steering strength amplifies distinct periodic transients without destroying the base signal structure.

6 Limitations

TimeSAE faces three primary limitations. First, the absence of a semantic dictionary in time series complicates auto-labeling abstract features. Second, the $32\times$ latent expansion restricts the method to offline auditing. Finally, the *Linear Representation Hypothesis* may underrepresent chaotic dynamics, necessitating future work on non-linear probing or Gated SAEs.

7 Conclusion

TimeSAE confirms the *Linear Representation Hypothesis* in TSFMs, delivering three contributions:

- **Feature Discovery:** Isolates sparse physical primitives (e.g., cycles) rather than memorized patterns.
- **Causal Steerability:** Demonstrates predictable forecast control via latent intervention.
- **High Fidelity:** Achieves 99.9% sparsity, resolving transparency debt in safety-critical systems.

By enabling *mechanistic monitoring*, TimeSAE allows for the diagnosis of specific neural circuits behind anomalies. Future research will scale this to multi-layer circuits and automate feature labeling to realize fully trustworthy, glass-box autonomous systems.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Trenton Bricken, Adly Templeton, Joshua Batson, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL `https://transformer-circuits.pub/2023/monosemantic-features/index.html`.

Hoagy Cunningham, Aidan Ewart, Lee Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, et al. Moment: A family of open time-series foundation models. In *International Conference on Machine Learning*, 2024.

Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 2018.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

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

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3319–3328, 2017.