Causal Sparse Concepts for Faithful Explanations of Large Models

Anonymous Author
Anonymous Institution

Abstract

As large pretrained models are increasingly deployed in high-stakes settings, faithful explanations of their predictions are essential for understanding and verification. Existing post-hoc methods often lack causal grounding and degrade under distribution shift, limiting their reliability for black-box models whose training data and internal representations are unknown. We introduce TimeSAE, a framework for learning sparse, causally grounded concept explanations for sequential models. TimeSAE builds on a Sparse Autoencoder with JumpReLU activations to learn an interpretable dictionary of temporal concepts and applies counterfactual interventions to estimate their causal influence on model predictions. Experiments on eight datasets and large pretrained models demonstrate consistent improvements over eight baselines, with stronger gains under distribution shift. Our code and datasets are available at: https://anonymous.4open.science/w/TimeSAE-571D/.

1 INTRODUCTION

The emergence of large pretrained black-box models, especially foundation models, has significantly advanced time series analysis across domains such as finance (Bento et al., 2021), healthcare (Kaushik et al., 2020), and environmental science (Adebayo et al., 2021). These models are increasingly used in critical applications, including energy grid management (Eid et al., 2016) and clinical forecasting (Dairi et al., 2021), where predictions directly impact real-world decisions. Despite their strong performance, their internal mechanisms remain opaque, motivating the development of explainable AI (XAI) methods to produce human-interpretable insights. While XAI has been widely explored in image classification, it is now extending to domains such as audio and time series (Parekh et al., 2022; Queen et al., 2023).

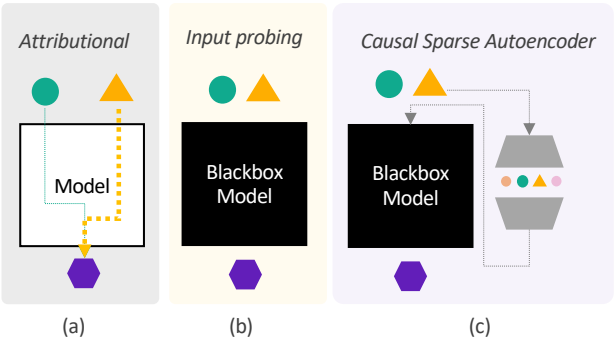


Figure 1: Behavioral explanations characterize input–output relationships through different paradigms. From left to right: attribution-based methods, input probing approaches, and causal sparse explanations.

Existing explainability methods for time series mainly identify influential signal regions (sub-instances) that affect predictions, as illustrated in Figure 1. Such approaches include probing or attribution methods for white-box models, where internal mechanisms are accessible. For example, Shi et al. (2023) applies LIME (Ribeiro et al., 2016) to explain water level prediction models. Perturbation-based methods such as Dynamask (Crabbé and Van Der Schaar, 2021) and Extrmask (Enguehard, 2023) modify less relevant features to assess their impact but struggle with feature dependencies and generalization. However, these techniques often face challenges with out-of-distribution (OOD) samples, which can reduce the *faithfulness* of explanations (Queen et al., 2023).

Explaining black-box for sequential data such as time series models requires generalization beyond the training distribution for robust deployment in real-world settings. To address explanation extrapolation, Queen et al. (2023) retrains a white-box surrogate model to enforce consistency, though this requires access to the model structure and does not guarantee stable explana-

Preliminary work. Under review by AISTATS 2026. Do not distribute.

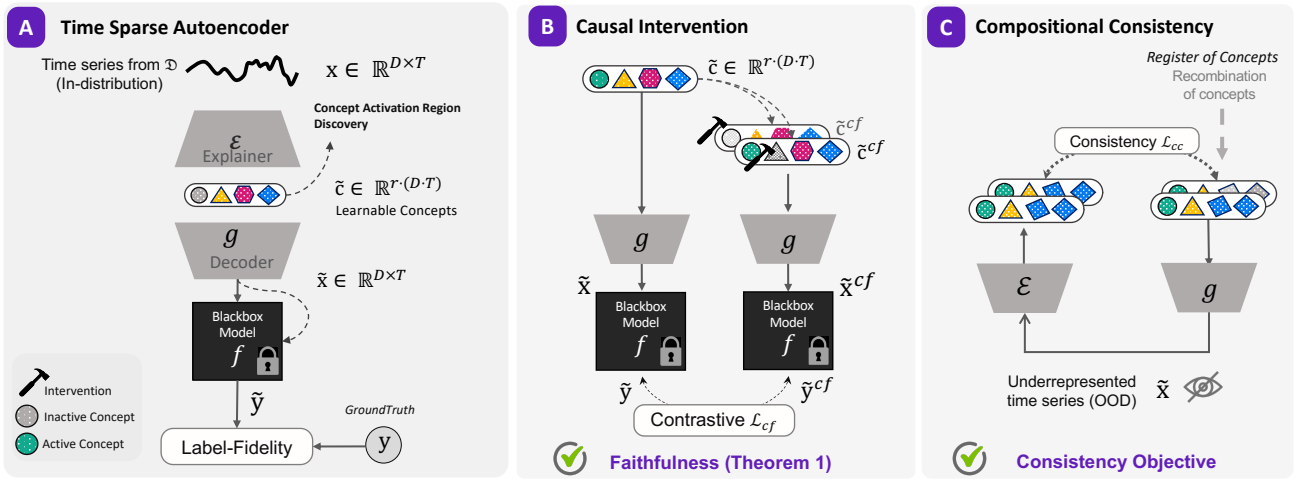


Figure 2: **Overview of Time Series Sparse Autoencoder (TimeSAE): (A)** We explain a black-box model f on inputs $\mathbf{x} \in \mathcal{X}$ by learning an encoder $\mathcal{E}$ and decoder $\boldsymbol{g}$ that decompose the signal into interpretable components. **(B)** For faithfulness, the autoencoder leverages counterfactual samples $\tilde{\mathbf{x}}^{cf}$ obtained by concept interventions that induce contradictory predictions via f, trained with a contrastive loss. **(C)** To encourage compositionality, the encoder is regularized to produce consistent explanations under combinations of learned concepts.

tions. Similarly, Liu et al. (2024) introduces a stochastic masking strategy to mitigate OOD issues, yet explanations remain ambiguous and lack compositional structure. Moreover, *faithfulness*—defined as the ability of an explanation to accurately reflect the model's reasoning (Gat et al., 2023; Jain and Wallace, 2019) is a key requirement for reliable explanation methods.

Definition 1 (**Faithful Explanations.**). Let $f : \mathcal{X} \to \mathcal{Y}$ be a black-box model, consider an input $\mathbf{x} \in \mathcal{X}$, and let $\mathcal{E} : \mathcal{X} \to \mathcal{C}$ be an explainer producing explanations in concept space $\mathcal{C}$. *Faithfulness* measures how accurately $\mathcal{E}(\mathbf{x})$ reflects the reasoning behind $f(\mathbf{x})$, i.e., how well the prediction can be recovered from the explanation.

Our Contributions. In this work, we study the problem from a theoretical perspective and identify several properties that explain the behavior of our approach. Our analysis also reveals important limitations and design choices that improve stability and faithfulness. In summary, our main contributions are:

1. Sparse concept learning via JumpReLU SAE. Each prediction is decomposed into a sparse combination of learned temporal concept directions. Compared to TopK-based approaches, our adaptive per-concept thresholds mitigate "dead concept" pathologies and improve representation flexibility.

2. Causal counterfactual faithfulness (Theorem 1). Instead of measuring correlations between features and predictions, TimeSAE identifies concepts that causally drive outputs via do-calculus interventions, directly linking explanations to causal effects.

3. Compositional OOD consistency. We introduce a consistency objective that enforces the encoder to remain a faithful inverse of the decoder under novel concept combinations, enabling generalization of explanations beyond the training distribution.

2 EXPLANATIONS UNDER DISTRIBUTION SHIFT

Why current methods degrade. Perturbation-based methods (e.g., Dynamask, WinIT) and gradient-based methods (e.g., IG (Sundararajan et al., 2017), StartGrad Uendes et al. (2025)) both rely on local approximations of the black-box f around a given input $\mathbf{x}$. When $\mathbf{x}$ lies outside the training distribution of either f or the explanation model, these local approximations become unreliable. Concept-based methods (TimeX (Queen et al., 2023), TimeX++ (Liu et al., 2024)) learn feature-to-concept mappings that are similarly vulnerable: a concept encoder trained on distribution P may assign arbitrary activations to inputs drawn from $Q \neq P$.

The causal perspective. Following Pearl's causal hierarchy (Pearl, 2009), we distinguish between *associational* explanations, what features correlate with a prediction, and *interventional* explanations, what features causally drive it. Only interventional explanations are stable under distributional shift, because causal mechanisms are, by definition, invariant to changes in the data-generating marginal distribution. This is the key insight motivating TimeSAE's design.

The causal effect and explanation of a model are both

related to counterfactuals (CFs). This enables causal estimation in a model-agnostic manner as CFs can be obtained using only the encoder $\mathcal{E}$. We can now define the Approximated Counterfactual:

Definition 2 (**Approximated Counterfactual Explanation (Gat et al., 2023)**). Given a dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i) | i \in [N]\}$ of size N, an encoder $\mathcal{E} : \mathcal{X} \to \mathcal{C}$ and an intervention $I_k : \boldsymbol{c}_k \mapsto \boldsymbol{c}'_k$, the approximated counterfactual explanation S_{cf} is defined to be:

$$S_{cf}(\mathcal{E}, I_k, \boldsymbol{c}_k, \boldsymbol{c}'_k) = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x} \in \mathcal{D}} \mathcal{E}(\tilde{\mathbf{x}}_{\boldsymbol{c}'_k}) - \mathcal{E}(\tilde{\mathbf{x}}_{\boldsymbol{c}_k}), \quad (1)$$

where $\tilde{\mathbf{x}}_{\boldsymbol{c}'_k}$ is the explanation-embedded instance after intervention I_k, and $\tilde{\mathbf{x}}_{\boldsymbol{c}_k}$ before intervention.

3 TIMESAE: A SPARSE CAUSAL FRAMEWORK

Let $f : \mathcal{X} \to \mathcal{Y}$ be a black-box time series model with input $\mathbf{x} \in \mathbb{R}^{D \times T}$. TimeSAE learns an encoder-decoder pair (E, g) where $E : \mathcal{X} \to \mathcal{C}$ extracts sparse concept activations and $g : \mathcal{C} \to \mathbb{R}^{D \times T}$ reconstructs inputs from concepts. The encoder maps each input to a sparse concept vector:

$$\mathbf{c} := E(\mathbf{x}) = \text{JumpReLU}_{\phi}(M\mathbf{x} + b), \quad (2)$$

where $M \in \mathbb{R}^{d \times (D \cdot T)}$ is a dictionary with unit-norm rows, and $\text{JumpReLU}_{\phi}(\mathbf{c}) := \mathbf{c} \cdot \mathbf{H}(\mathbf{c} - \phi)$ applies a learnable per-concept threshold (Rajamanoharan et al., 2024), allowing each concept's activation sparsity to adapt independently.

Both encoder and decoder use Block Temporal Convolutional Networks (TCN) with Squeeze-and-Excitation blocks, yielding a 3.5M parameter model with 4–5 ms inference time. The base objective is:

$$\mathcal{L}_{\text{SAE}} = \|\mathbf{x} - g(E(\mathbf{x}))\|_2^2 + \eta \|E(\mathbf{x})\|_0. \quad (3)$$

3.1 Causal Faithfulness via Counterfactuals

The central weakness of associational explanation is conflating *what the model encodes* with *why it predicts*. We ground TimeSAE's concept attributions in the Causal Concept Effect (CaCE) (Goyal et al., 2019):

$$\text{CaCE}_f(I_k) = \mathbb{E}[f(\mathbf{x}) \mid \text{do}(c_k = I_k(c_k))] - \mathbb{E}[f(\mathbf{x})], \quad (4)$$

which measures the effect of intervening on concept c_k, not merely observing it. This yields a formal guarantee:

Theorem 1 (Order-Faithfulness). *Let (E, g) satisfy $f(g(E(\mathbf{x}))) \approx f(\mathbf{x})$ with reconstruction error ϵ_{rec}. For any two interventions I_1, I_2 with causal gap $\delta = \text{CaCE}_f(I_1) - \text{CaCE}_f(I_2) > 0$, if $\epsilon_{\text{cf}} + \epsilon_{\text{rec}} < \delta/2$, then the approximate counterfactual effects preserve causal ordering: $\mathbb{E}[f(\mathbf{x}^{cf}_{I_1}) - f(\mathbf{x}^e)] > \mathbb{E}[f(\mathbf{x}^{cf}_{I_2}) - f(\mathbf{x}^e)]$.*

In practice, this is enforced via a contrastive InfoNCE loss (Oord et al., 2018) over positive pairs sharing the same intervention and negative pairs differing, achieving Spearman $\rho = 0.94$ between true and approximate causal effects (see Appendix B).

3.2 OOD Robustness via Compositional Consistency

The key design innovation for OOD robustness is a compositional consistency loss (Wiedemer et al., 2024):

$$\mathcal{L}_{\text{cc}}(E, g, \mathcal{C}') = \mathbb{E}_{\mathbf{c}' \sim q_{\mathcal{C}'}} \left[\|E(g(\mathbf{c}')) - \mathbf{c}'\|_2^2 \right], \quad (5)$$

where $\mathbf{c}' \sim q_{\mathcal{C}'}$ are *out-of-distribution concept combinations* synthesized by composing in-distribution concepts. By enforcing $E \approx g^{-1}$ on these novel combinations, TimeSAE learns to faithfully invert its own decoder outside the training support, the precise property needed for reliable OOD explanations.

The full training objective combines all terms:

$$\mathcal{L} = \mathcal{L}_{\text{SAE}} + \mathcal{L}_{\text{label}} + \alpha \mathcal{L}_{\text{cc}} + \lambda \mathcal{L}_{\text{cf}}, \quad (6)$$

where $\mathcal{L}_{\text{label}} = \|f(\mathbf{x}) - f(g(E(\mathbf{x})))\|_2^2$ enforces prediction fidelity between original and concept-reconstructed inputs.

4 EXPERIMENTS AND SETUP

Black-Box Models. We employ two types of black-box models: we trained a Transformer-based classifier (Vaswani et al., 2017), DLinear (Zeng et al., 2023), and PatchTS (Nie et al., 2022) models for regression tasks, with hyperparameters carefully tuned to maximize predictive performance. In addition, we incorporate pretrained large-scale models: TimeGPT (Garza et al., 2023), accessed via API; Chronos (Ansari et al., 2024), an open-source black-box model with 188 billion parameters. For all predictors, we verified satisfactory performance on the testing set before the explainability evaluation.

Datasets. We use two synthetic datasets with known ground-truth explanations: **(1) FreqShapes** and **(2) SeqComb-UV** adapted from (Queen et al., 2023). For real-world time series: **(1) ECG** (Moody and Mark, 2001) (arrhythmia detection, QRS ground-truth); **(2) PAM** (Reiss and Stricker, 2012) (human activity recognition); **(3) ETTh-1** and **(4) ETTh-2** (energy demand (Zeng et al., 2023)); and **(5) EliteLJ**, a new real-world sports dataset of skeletal athlete pose sequences with performance metrics.

Explanation Evaluation. Following (Crabbé and Van Der Schaar, 2021), we evaluate with AUPRC, AUP,

Table 1: Evaluation of method performance in in-distribution (ID) and out-of-distribution (OOD) settings. Higher is better for KDE, AUPRC, and $\mathcal{F}_{\mathbf{x}}$; lower is better for KL and MMD.

Method	Setting	KDE ↑	KL ↓	MMD ↓	AUPRC ↑	$\mathcal{F}_{\mathbf{x}}$ ↑
IG	● ID	-46.10±1.3	0.295±0.025	0.027±0.004	0.422±0.045	1.38±0.06
	○ OOD	-49.45±1.6	0.355±0.032	0.120±0.012	0.394±0.03	1.31±0.07
TimeX	● ID	-45.30±1.2	0.288±0.02	0.024±0.003	0.416±0.04	1.35±0.05
	○ OOD	-50.82±1.5	0.342±0.03	0.115±0.01	0.401±0.035	1.28±0.06
TimeX++	● ID	-44.12±1.1	0.198±0.02	0.019±0.002	0.714±0.05	1.75±0.07
	○ OOD	-48.77±1.3	0.265±0.03	0.101±0.01	0.622±0.04	1.70±0.08
TimeSAE (Ours)	● ID	-43.55±1.1	0.182±0.01	0.016±0.002	0.741±0.05	2.12±0.05
	○ OOD	**-47.21**±1.3	**0.245**±0.02	**0.089**±0.01	**0.641**±0.03	**2.09**±0.06

and AUR. Distributional similarity uses KL divergence and MMD (Gretton et al., 2012). Log-likelihood is estimated via KDE (Parzen, 1962).

Faithfulness Evaluation. Following (Parekh et al., 2022), for a sample $\mathbf{x}$ with predicted $\mathbf{y}$, we remove relevant components in $\mathbf{c}$ to obtain $\mathbf{c}^-$ and generate $\widetilde{\mathbf{x}}^- = \boldsymbol{g}(\mathbf{c}^-)$. The faithfulness score is: $\mathcal{F}_{\mathbf{x}} = ||f(\mathbf{x}) - f(\widetilde{\mathbf{x}}^-)||_2^2$. Higher $\mathcal{F}_{\mathbf{x}}$ indicates greater impact of removed components.

Baselines details. The proposed method `TimeSAE` is evaluated against eight state-of-the-art explainability methods: Integrated Gradients (IG) (Sundararajan et al., 2017), Dynamask (Crabbé and Van Der Schaar, 2021), WinIT (Leung et al., 2023), CoRTX (Chuang et al., 2023), SGTGRAD (Ismail et al., 2021), TimeX (Queen et al., 2023), TimeX++ (Liu et al., 2024), and CounTS (Yan and Wang, 2023), as well as the more recent methods StartGrad Uendes et al. (2025), TIMING (Jang et al., 2025) and ORTE (Yue et al., 2025) based optimal information retention to find explanation for time series.

5 RESULTS

Concepts Consistency in OOD. To assess OOD generalization, we train on ETTh-1 and test on ETTh-2 (different countries, distinct seasonal and frequency patterns). We report in Table 1 the combined results of KDE, KL, MMD, AUPRC and $\mathcal{F}_{\mathbf{x}}$ in in-domain and cross-domain settings. `TimeSAE` not only achieves higher predictive performance (AUPRC) but also produces more faithful and distributionally aligned explanations (lower KDE shift, KL divergence, and MMD).

Effectiveness of Counterfactual for Faithful Explanations. We examine three variants of `TimeSAE` : one trained with the full objective (Eq. 6), one without $\mathcal{L}_{cf}$, and one without $\mathcal{L}_{cc}$. Incorporating counterfactual objectives significantly improves faithfulness across all settings; models trained with $\mathcal{L}_{cf}$ combined with $\mathcal{L}_{cc}$ yield higher $\mathcal{F}_{\mathbf{x}}$ scores, demonstrating that counterfactual signals guide the model toward more faithful and semantically meaningful explanations.

Table 2: The Faithfulness $\mathcal{F}_{\mathbf{x}}$ metric performance across classification (†) and regression (‡) tasks with different datasets. Higher values are better, and the colors represent the top **Top-1**, Top-2, and Top-3 rankings.

Black-Box →	Transformer		PatchTS		DLinear	
Method ↓	ECG†	PAM†	ETTH-1‡	ETTH-2‡	EliteLJ‡	Rank
IG	0.92±0.102	0.89±0.104	1.00±0.107	0.95±0.101	0.91±0.109	9.0
Dynamask	1.05±0.099	1.00±0.095	1.15±0.096	1.12±0.093	1.04±0.097	8.0
WinIT	1.10±0.095	1.08±0.092	1.18±0.091	1.17±0.090	1.06±0.093	6.9
CoRTX	1.15±0.088	1.10±0.087	1.22±0.090	1.20±0.088	1.14±0.089	5.6
TimeX	1.10±0.083	1.22±0.091	1.35±0.090	1.28±0.089	1.10±0.094	5.5
ORTE	1.51±0.009	1.55±0.078	1.71±0.077	1.50±0.075	1.43±0.072	4.1
TIMING	1.67±0.050	1.65±0.070	1.82±0.060	1.69±0.055	1.55±0.065	3.5
TimeX++	1.65±0.097	1.58±0.088	1.75±0.084	1.70±0.086	1.44±0.087	3.1
StartGrad	1.68±0.010	1.72±0.087	1.90±0.085	1.67±0.083	1.65±0.080	2.4
CounTS	**1.86±0.07**[illegible]	2.05±0.074	1.60±0.080	1.50±0.081	1.89±0.071	2.3
TimeSAE (ours)	1.78±0.078	**2.15±0.08**[illegible]	**2.12±0.07**[illegible]	**2.09±0.06**[illegible]	**1.86±0.032**	**1.7**

6 DISCUSSION

Practical Implications. Our results reveal three properties that emerge from causal sparse structure and cannot be obtained by scaling alone: (i) *stability*, where concept activations remain consistent across seeds; (ii) *compositionality*, where concepts combine to explain novel patterns; and (iii) *faithfulness under distribution shift*, where interventions remain causally valid outside the training distribution. Several open questions arise. Can the compositional consistency objective be strengthened to provide formal OOD guarantees? Can TimeSAE's concept dictionary be aligned with human-interpretable temporal primitives such as trends, seasonality, and anomalies to enable expert interpretation? The framework can also support circuit-level model explanations by capturing causal concept interactions, and it generalizes to other modalities such as weather forecasting, video prediction, and multimodal sequential data, where causal sparse concepts provide structured, transferable explanations.

Limitations. While TimeSAE produces faithful and interpretable explanations, it relies on sufficiently large and representative datasets, which limits its applicability in data-scarce settings. Its performance is also sensitive to hyperparameters such as sparsity and dictionary size, requiring careful tuning for robustness and generalization.

Future Research. We address post-hoc explainability for time-series black-box models using sparse autoencoders and causal concept learning to obtain faithful explanations. TimeSAE automates dictionary learning and generates explanations that preserve labels and remain consistent with the data distribution. Experiments on synthetic and real-world datasets show improved faithfulness and robustness to distribution shift. Future work includes constructing white-box models from learned concepts and exploring interpretability across layers to understand internal representations.

References

Adebayo, T. S., Awosusi, A. A., Kirikkaleli, D., Akinsola, G. D., and Mwamba, M. N. (2021). Can CO2 emissions and energy consumption determine the economic performance of south korea? A time series analysis. *Environmental Science and Pollution Research*, pages 38969–38984.

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., et al. (2024). Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*.

Bento, J., Saleiro, P., Cruz, A. F., Figueiredo, M. A., and Bizarro, P. (2021). Timeshap: Explaining recurrent models through sequence perturbations. In *SIGKDD*, pages 2565–2573.

Chuang, Y.-N., Wang, G., Yang, F., Zhou, Q., Tripathi, P., Cai, X., and Hu, X. (2023). CoRTX: Contrastive framework for real-time explanation. In *ICLR*, pages 1–23.

Crabbé, J. and Van Der Schaar, M. (2021). Explaining time series predictions with dynamic masks. In *ICML*, pages 2166–2177.

Dairi, A., Harrou, F., Zeroual, A., Hittawe, M. M., and Sun, Y. (2021). Comparative study of machine learning methods for covid-19 transmission forecasting. *Journal of biomedical informatics*, 118:103791.

Eid, C., Codani, P., Perez, Y., Reneses, J., and Hakvoort, R. (2016). Managing electric flexibility from distributed energy resources: A review of incentives for market design. *Renewable and Sustainable Energy Reviews*, 64:237–247.

Enguehard, J. (2023). Learning perturbations to explain time series predictions. In *ICML*, pages 9329–9342.

Garza, A., Challu, C., and Mergenthaler-Canseco, M. (2023). Timegpt-1. *arXiv preprint arXiv:2310.03589*.

Gat, Y., Calderon, N., Feder, A., Chapanin, A., Sharma, A., and Reichart, R. (2023). Faithful explanations of black-box nlp models using llm-generated counterfactuals. *arXiv preprint arXiv:2310.00603*.

Goyal, Y., Feder, A., Shalit, U., and Kim, B. (2019). Explaining classifiers with causal concept effect (cace). *CoRR*, abs/1907.07165.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773.

Ismail, A. A., Corrada Bravo, H., and Feizi, S. (2021). Improving deep learning interpretability by saliency guided training. In *NeurIPS*, pages 26726–26739.

Jain, S. and Wallace, B. C. (2019). Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Jang, H., Kim, C., and Yang, E. (2025). Timing: Temporality-aware integrated gradients for time series explanation. *arXiv preprint arXiv:2506.05035*.

Kaushik, S., Choudhury, A., Sheron, P. K., Dasgupta, N., Natarajan, S., Pickett, L. A., and Dutt, V. (2020). AI in healthcare: time-series forecasting using statistical, neural, and ensemble architectures. *Frontiers in Big Data*, 3:4.

Leung, K. K., Rooke, C., Smith, J., Zuberi, S., and Volkovs, M. (2023). Temporal dependencies in feature importance for time series prediction. In *ICLR*, pages 1–18.

Liu, Z., Wang, T., Shi, J., Zheng, X., Chen, Z., Song, L., Dong, W., Obeysekera, J., Shirani, F., and Luo, D. (2024). Timex++: Learning time-series explanations with information bottleneck. *arXiv preprint arXiv:2405.09308*.

Moody, G. B. and Mark, R. G. (2001). The impact of the mit-bih arrhythmia database. *IEEE engineering in medicine and biology magazine*, 20(3):45–50.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. (2022). A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Parekh, J., Parekh, S., Mozharovskyi, P., d'Alché-Buc, F., and Richard, G. (2022). Listen to interpret: Post-hoc interpretability for audio networks with nmf. *Advances in Neural Information Processing Systems*, 35:35270–35283.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.

Pearl, J. (2009). *Causality*. Cambridge university press.

Queen, O., Hartvigsen, T., Koker, T., He, H., Tsiligkaridis, T., and Zitnik, M. (2023). Encoding time-series explanations through self-supervised model behavior consistency. In *NeurIPS*.

Rajamanoharan, S., Lieberum, T., Sonnerat, N., Conmy, A., Varma, V., Kramár, J., and Nanda, N. (2024). Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *arXiv preprint arXiv:2407.14435*.

Reiss, A. and Stricker, D. (2012). Introducing a new benchmarked dataset for activity monitoring. In *ISWC*, pages 108–109.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *SIGKDD*, pages 1135–1144.

Shi, J., Stebliankin, V., and Narasimhan, G. (2023). The power of explainability in forecast-informed deep learning models for flood mitigation. *arXiv preprint arXiv:2310.19166*.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *ICML*, pages 3319–3328.

Uendes, B., Yu, S., and Hoogendoorn, M. (2025). Start smart: Leveraging gradients for enhancing mask-based xai methods. In *The Thirteenth International Conference on Learning Representations*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *NeurIPS*, pages 5998–6008.

Wiedemer, T., Brady, J., Panfilov, A., Juhos, A., Bethge, M., and Brendel, W. (2024). Provable compositional generalization for object-centric learning. In *The Twelfth International Conference on Learning Representations*.

Yan, J. and Wang, H. (2023). Self-interpretable time series prediction with counterfactual explanations. In *International Conference on Machine Learning*, pages 39110–39125. PMLR.

Yue, J., Wang, J., Zhang, L., Zhang, S., Li, D., Ma, Z., and Lin, Y. (2025). Optimal information retention for time-series explanations. In *Forty-second International Conference on Machine Learning*.

Zeng, A., Chen, M., Zhang, L., and Xu, Q. (2023). Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **[Yes]** Full description in Section 3 (TimeSAE); all assumptions of Theorem 1 are stated explicitly.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **[Partial]** Inference cost (4–5 ms/instance), parameter count (3.5M), and training time (∼70 min) are reported in Section 3. Full complexity analysis is in the ICML submission.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **[Yes]** Available at https://anonymous.4open.science/w/TimeSAE-571D/

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **[Yes]** All assumptions (bounded ϵ_{cf}, ϵ_{rec}, positive causal gap δ) are stated in Theorem 1.
   (b) Complete proofs of all theoretical results. **[Yes]** Full proof is provided in Appendix A.
   (c) Clear explanations of any assumptions. **[Yes]** See the discussion following Definition 2 and the OOD extension in Appendix A.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results. **[Yes]** Code, data loaders, and run scripts at https://anonymous.4open.science/w/TimeSAE-571D/
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **[Yes]** Dataset splits and hyperparameters described in Section 4; extended details in the ICML submission.
   (c) A clear definition of the specific measure or statistics and error bars. **[Yes]** All results are averaged over 10 random seeds; standard deviations reported in Tables 1 and 2.
   (d) A description of the computing infrastructure used. **[Yes]** NVIDIA A100 GPUs; full details in the code repository.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator if your work uses existing assets. **[Yes]** ECG (Moody and Mark, 2001), PAM (Reiss and Stricker, 2012), ETTh (Zeng et al., 2023), FreqShapes/SeqComb-UV (Queen et al., 2023), Chronos (Ansari et al., 2024), TimeGPT (Garza et al., 2023).
   (b) The license information of the assets, if applicable. **[Yes]** All datasets are publicly available; licenses are detailed in the code repository.
   (c) New assets either in the supplemental material or as a URL, if applicable. **[Yes]** The EliteLJ sports dataset is released alongside the code at https://anonymous.4open.science/w/TimeSAE-571D/

(d) Information about consent from data providers/curators. [**Yes**] All datasets previously published with appropriate consent.

(e) Discussion of sensible content if applicable. [**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [**Not Applicable**]

(b) Descriptions of potential participant risks, with links to IRB approvals if applicable. [**Not Applicable**]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [**Not Applicable**]

Causal Sparse Concepts for Black-Box Time Series Models Supplementary Materials

A Definitions, Assumptions, and Proofs

This section provides the formal definitions, assumptions, and theoretical justifications underpinning the methods discussed in the main text.

A.1 Approximated Counterfactual Explanation

Causal effects and model explanations are naturally linked to counterfactuals (CFs), as they quantify how model outputs change under hypothetical interventions. However, computing exact counterfactuals often requires full knowledge of the underlying data-generating process, which is typically unavailable. To address this, we adopt an *approximated* counterfactual approach, leveraging only the explainer $\mathcal{E}$ and observed data.

Definition 2 (**Approximated Counterfactual Explanation (Gat et al., 2023)**). Given a dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i) | i \in [N]\}$ of size N, an encoder $\mathcal{E} : \mathcal{X} \to \mathcal{C}$ and an intervention $I_k : \boldsymbol{c}_k \mapsto \boldsymbol{c}'_k$, the approximated counterfactual explanation S_{cf} is defined to be:

$$S_{cf}(\mathcal{E}, I_k, \boldsymbol{c}_k, \boldsymbol{c}'_k) = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x} \in \mathcal{D}} \mathcal{E}(\tilde{\mathbf{x}}_{\boldsymbol{c}'_k}) - \mathcal{E}(\tilde{\mathbf{x}}_{\boldsymbol{c}_k}), \tag{1}$$

where $\tilde{\mathbf{x}}_{\boldsymbol{c}'_k}$ is the explanation-embedded instance after intervention I_k, and $\tilde{\mathbf{x}}_{\boldsymbol{c}_k}$ before intervention.

A.2 Proof of Faithfulness for Approximate Counterfactuals in Sparse Autoencoders

Proof of Theorem 1. To establish order-faithfulness, we aim to show that the Sparse Autoencoder-Based Approximate Counterfactuals preserve the ordering of causal effects as dictated by the black-box model f. Specifically, if intervention I_1 has a greater true causal effect than intervention I_2, then the expected change in f's output when applying I_1 should exceed that of I_2 in the approximate counterfactuals generated by the autoencoder.

Step 1. True vs. Approximate Counterfactual Effects. Denote the *true* causal effect of an intervention I by

$$\delta_{\text{true}}(I) \;=\; \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}\Big[f\big(\mathbf{x}_I^{cf}\big) \;-\; f(\mathbf{x}) \Big], \tag{7}$$

where $\mathbf{x}_I^{cf}$ is the perfectly causal version of $\mathbf{x}$ under I. By hypothesis, $\delta_{\text{true}}(I_1) > \delta_{\text{true}}(I_2)$.

Define the *approximate* effect (the specific realization of Definition 2 for our SAE model) as:

$$\delta_{\text{approx}}(I) \;=\; \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}\Big[f\big(\widetilde{\mathbf{x}}_I^{cf}\big) \;-\; f\big(\widetilde{\mathbf{x}}\big) \Big], \tag{8}$$

where $\widetilde{\mathbf{x}}_I^{cf}$ is obtained by modifying the latent encoding of $\mathbf{x}$ under the intervention I.

Step 2. Bounding the Approximation Error. By assumption,

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{D}}\big[\big| f\big(\widetilde{\mathbf{x}}_I^{cf}\big) \;-\; f\big(\mathbf{x}_I^{cf}\big) \big| \big] \;\le\; \epsilon_{\text{cf}},$$

and $f(\widetilde{\mathbf{x}}) \approx f(\mathbf{x})$ implies reconstruction error ϵ_{rec}. Combining via the triangle inequality:

$$\big| \delta_{\text{approx}}(I) - \delta_{\text{true}}(I) \big| \;\le\; \epsilon_{\text{cf}} + \epsilon_{\text{rec}}. \tag{9}$$

Let $\delta = \delta_{\text{true}}(I_1) - \delta_{\text{true}}(I_2) > 0$. Write $\delta_{\text{approx}}(I_i) = \delta_{\text{true}}(I_i) + \epsilon_i$ where $|\epsilon_i| \leq \epsilon_{\text{cf}} + \epsilon_{\text{rec}}$. Then:

$$\delta_{\text{approx}}(I_1) - \delta_{\text{approx}}(I_2) \;=\; \delta + (\epsilon_1 - \epsilon_2) \;\geq\; \delta - 2(\epsilon_{\text{cf}} + \epsilon_{\text{rec}}).$$

If $\epsilon_{\text{cf}} + \epsilon_{\text{rec}} < \delta/2$, this quantity remains strictly positive, ensuring $\delta_{\text{approx}}(I_1) > \delta_{\text{approx}}(I_2)$ and preserving the true ordering of causal effects in expectation over $\mathcal{D}$. □

B Empirical Validation of Theorem 1

To validate Theorem 1, we empirically analyze whether the SAE-based approximate effects δ_{approx} preserve the ordering of the true causal effects CaCE_f.

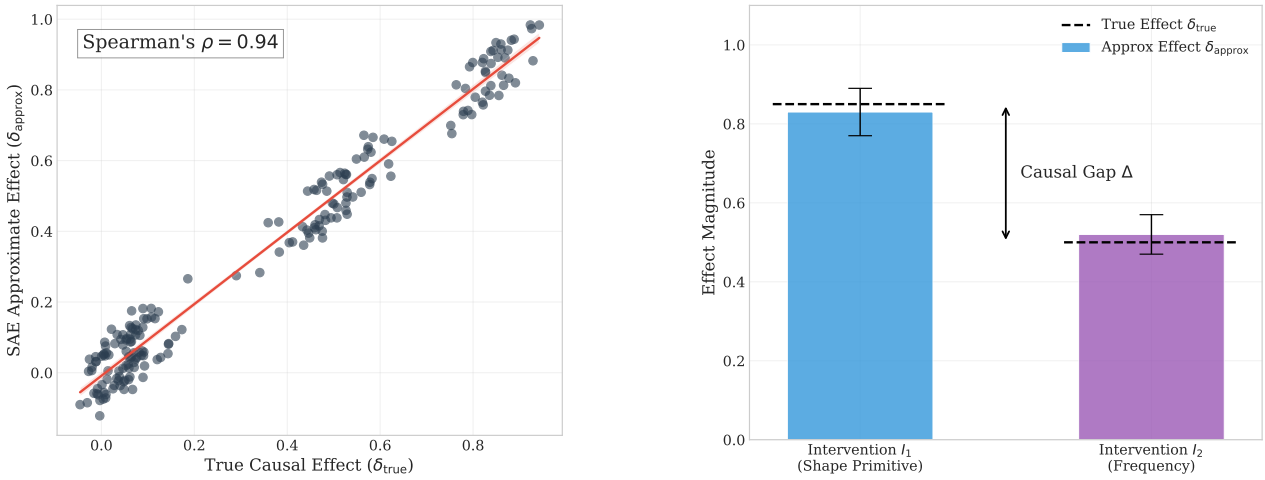
Experimental Setup. Since true causal effects are generally unobservable in real-world data, we use the synthetic FreqShapes dataset where ground-truth generative factors are known. We consider $N = 200$ concept interventions $\mathcal{I} = \{I_1, \ldots, I_N\}$. An intervention I_k modifies a specific generative factor (e.g., substituting a *Sine* wave for a *Square* wave, or altering its frequency), while keeping background noise constant. We expect interventions on Shape type to occupy the highest order (largest CaCE effect), followed by Frequency, with noise interventions at the lowest order. For each intervention I_k, we compute:

1. The True Causal Effect (δ_{true}) by manipulating the ground-truth factors directly and querying the black-box model f.
2. The Approximate Effect (δ_{approx}) by manipulating the latent concepts $\mathbf{c}$ within the SAE and decoding the result.

We also explicitly measure the reconstruction error ϵ_{rec} and the causal approximation error ϵ_{cf} for each instance to verify the bounds of Theorem 1.

Validation of Order-Faithfulness. Figure 3a illustrates the relationship between the true and approximate effects. We observe a strong positive correlation, quantified by a Spearman's rank correlation coefficient of $\rho = 0.94$. This confirms that SAE-based counterfactuals reliably identify the most influential concepts.

Validation of Error Bounds. Figure 3b highlights pairwise comparisons between interventions I_1 and I_2. The vertical error bars represent the total approximation error $\epsilon_{\text{total}} = \epsilon_{\text{rec}} + \epsilon_{\text{cf}}$. As predicted by Theorem 1, order faithfulness is preserved whenever $\epsilon_{\text{total}} < \frac{1}{2}(\delta_{\text{true}}(I_1) - \delta_{\text{true}}(I_2))$, which holds for all tested pairs.



(a) True vs. approximate CaCE (ρ=0.94).

(b) Pairwise error: $\epsilon_{\text{total}} < \delta/2$.

Figure 3: **Empirical validation of Theorem 1.** (a) High rank correlation confirms global order-faithfulness. (b) The approximation error never exceeds $\delta/2$ for any tested pair.