time2time: Causal Intervention in Hidden States to Simulate Rare Events in Time Series Foundation Models

Debdeep Sanyal[1], Aaryan Nagpal[1], Dhruv Kumar[1, 2], Murari Mandal[1, 3], Saurabh Deshpande[1,*]
[1] Birla AI Labs, Office of Ananya Birla, [2] BITS Pilani, [3] KIIT Bhubaneswar

Abstract

While transformer-based foundation models excel at forecasting routine patterns, two questions remain: do they internalize semantic concepts such as market regimes, or merely fit curves? And can their internal representations be leveraged to simulate rare, high-stakes events such as market crashes? To investigate this, we introduce activation transplantation, a causal intervention that manipulates hidden states by imposing the statistical moments of one event (e.g., a historical crash) onto another (e.g., a calm period) during the forward pass. This procedure deterministically steers forecasts: injecting crash semantics induces downturn predictions, while injecting calm semantics suppresses crashes and restores stability. Beyond binary control, we find that models encode a graded notion of event severity, with the latent vector norm directly correlating with the magnitude of systemic shocks. Validated across two architecturally distinct TSFMs, Toto (decoder only) and Chronos (encoder-decoder), our results demonstrate that steerable, semantically grounded representations are a robust property of large time series transformers. Our findings provide evidence for a *latent concept space* that governs model predictions, shifting interpretability from post-hoc attribution to direct causal intervention, and enabling semantic "what-if" analysis for strategic stress-testing.

1 Introduction

Transformer-based time series foundation models (TSFM) currently define the state-of-the-art in time series forecasting Ansari et al. [2024], Woo et al. [2024], Das et al. [2024], Cohen et al. [2024], Nie et al. [2023], yet their black-box nature poses critical risks and limits their utility in various analysis Liu et al. [2024], Goel et al. [2024], Ahmed et al. [2022]. In high stakes domains such as finance, healthcare, and energy, trustworthy forecasts are critical, as a single misinterpreted prediction can lead to catastrophic consequences. This raises a fundamental question: Do these models develop a genuine understanding of the processes they model, or are they merely sophisticated curve-fitters, parroting complex patterns without comprehending their meaning? Answering this is essential for building reliable systems in critical applications.

Financial markets represent one of the most consequential domains, and thus serve as an ideal testbed for our study. This context leads us to a pivotal question: does a TSFM, pretrained on decades of market data, learn an internal, manipulable concept of a "market crash"? While prior work shows that TSFMs can represent low-level mathematical primitives like trends and seasonality and that they can be manipulated Wiliński et al. [2025], Queen et al. [2023], Ozyegen et al. [2023], it remains unknown

Code available at `https://github.com/birla-ai-labs/time2time`
*Correspondence: saurabh.deshpande-c@adityabirla.com

NeurIPS 2025 Workshop on Recent Advances in Time Series Foundation Models (BERT^{2}S).

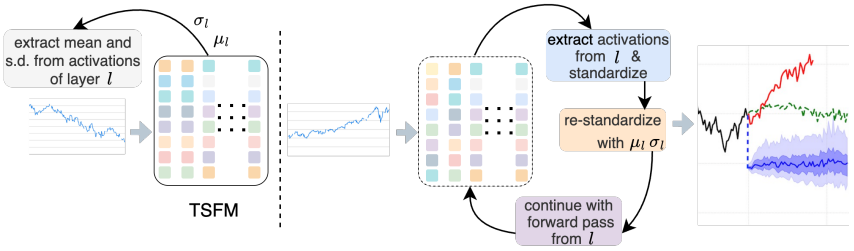


Figure 1: Overview of the proposed *time2time* intervention. We extract the statistical moments (mean and standard deviation) of hidden activations at layer $\updownarrow$ from a style event, standardize the target activations, and re-standardize them with the style statistics. This activation transplantation implants the dynamics of one event (e.g., a market crash) into another (e.g., a calm period), steering the model's forecast accordingly.

if they can grasp holistic events defined not by simple functions, but by complex real-world dynamics like panic-driven volatility and cascading price declines.

To test for causation beyond mere correlation, we introduce a direct intervention. We probe our hypothesis that market regimes are encoded in population-level activation statistics (mean and std), and by transplanting these statistics from a periods with differing dynamics (calm or crash) mid-forward pass. This allows us to "implant" the signature of a crash, which, if our hypothesis holds, should deterministically steer the model's forecast to predict a change in it's forecast scale.

Our results provide a striking confirmation. Injecting the statistical signature of a major crash (e.g., 2000 [†], 2008 [‡], or 2020 [§]) forces a calm period's forecast into a sharp downturn. The converse intervention; imposing calm statistics onto a crash, suppresses the downturn forecast. Beyond this binary steering, we uncover a deeper nuance in the model's understanding: the model quantitatively encodes the relative severity of different crashes, as measured by the norm of its latent representation. The model has learned not just *what* a crash is, but that the Dot-com crash of 2000 constitutes a more severe *representational event* than the 2008 crisis, revealing a rich, semantically grounded internal world. Building on these results, this work makes the following contributions:

- **First direct, causal evidence of semantic concepts:** We provide the first causal evidence that TSFMs learn high-level semantic concepts of real-world events, moving beyond simple primitives to holistic concepts like market crashes.
- **The discovery of nuanced, interpretable representations:** We discover that these learned concepts are remarkably nuanced, with event severity encoded as a continuous and interpretable latent variable (i.e., the activation vector norm).
- **A path to controllable, risk-aware simulation:** Our findings establish a practical path toward a new class of controllable simulations, enabling practitioners to conduct risk-aware stress-testing by simulating the impact of historic systemic shocks.

2 Causal Intervention via Activation Transplantation

The Semantic Hypothesis in Activations Let M be a decoder-only TSFM with parameters θ, consisting of L layers. For a given input time series $X \in \mathbb{R}^{N \times T_{\text{in}}}$, where N is the number of variates and T_{in} is the length of the historical context, the model generates a forecast $\hat{Y} \in \mathbb{R}^{N \times T_{\text{out}}}$. The activation tensor at the output of any layer $l \in \{1, ..., L\}$ is denoted as:

$$A_l(X) = f_l(A_{l-1}(X); \theta_l) \in \mathbb{R}^{N \times T_{\text{in}} \times D} \tag{1}$$

where f_l is the function of the l-th Transformer block, D is the hidden dimensionality, and $A_0(X)$ is the initial embedding of the input X. (For clarity, we omit the batch dimension from notation). Our

[†]https://en.wikipedia.org/wiki/Dot-com_bubble
[‡]https://en.wikipedia.org/wiki/2008_financial_crisis
[§]https://en.wikipedia.org/wiki/2020_stock_market_crash

guiding hypothesis is that the semantic signature of X is encoded within the statistics of $A_l(X)$ for some layer l.

The Transplantation Mechanism Our procedure requires two distinct time series: a **style** series, X_{style}, and a **target** series, X_{target}.

Step 1: Extracting the Semantic Signature. We first extract the semantic signature from the *Style* data's activations at layer l. This signature is defined as the mean and standard deviation vectors computed **across the sequence length (T_{in})**, capturing the characteristic neural statistics of the entire style period.

$$\mu_l(X_{\text{style}}) = \frac{1}{T_{\text{in}}} \sum_{t=1}^{T_{\text{in}}} A_l(X_{\text{style}})_{:,t,:} \quad ; \quad \sigma_l(X_{\text{style}}) = \sqrt{\frac{1}{T_{\text{in}}} \sum_{t=1}^{T_{\text{in}}} (A_l(X_{\text{style}})_{:,t,:} - \mu_l)^2} \tag{2}$$

where the statistics are computed over the time-step index t, resulting in $\mu_l, \sigma_l \in \mathbb{R}^{N \times D}$. This operation directly corresponds to taking the mean and standard deviation along the sequence length dimension of the activation tensor.

Step 2: Intervention on target Activations. Next, in a separate forward pass with the *target* data, we halt the computation after layer l. We then perform the core transplantation. We first standardize the target activations by their own statistics along sequence dimension to strip them of their original semantic signature. Then, we re-scale and shift them using the stored signature from the style data.

$$\tilde{A}_l(X_{\text{target}}) = \left(\frac{A_l(X_{\text{target}}) - \mu_l(X_{\text{target}})}{\sigma_l(X_{\text{target}}) + \epsilon} \right) \odot \sigma_l(X_{\text{style}}) + \mu_l(X_{\text{style}}) \tag{3}$$

where $\odot$ denotes element-wise multiplication (with broadcasting) and ϵ is a small constant (e.g., 10^{-5}) for numerical stability. This equation precisely replaces the time-averaged statistics of the target with those of the style, while preserving the time-step-varying structure of the normalized target.

Step 3: Generating the Intervened Forecast. Finally, we resume the forward pass from layer $l+1$, feeding the modified activation tensor $\tilde{A}_l$ into the rest of the network to produce a forecast conditioned on the implanted semantic concept.

$$\hat{Y}_{\text{intervened}} = M_{l+1 \to L}(\tilde{A}_l(X_{\text{target}})) \tag{4}$$

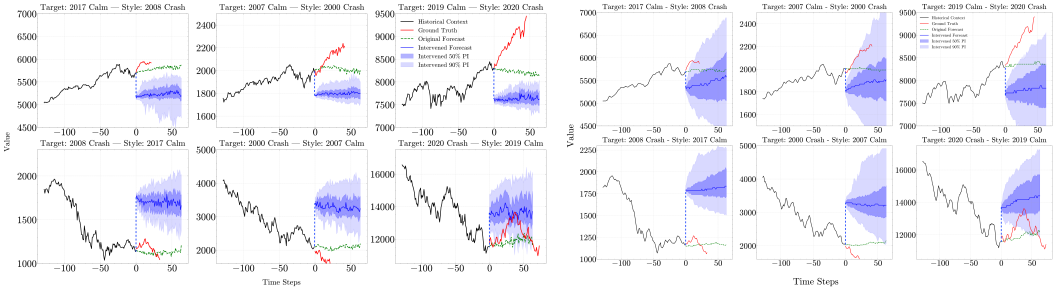
This framework provides a general and powerful tool for causally testing and steering the conceptual representations within a TSFM model.

3 Experiments & Results

To rigorously test the generality of our claims, we perform our experiments on two TSFMs, deliberately chosen for their architectural and scale diversity: Toto Open-Base-1.0 Cohen et al. [2024], a 103M parameter decoder-only architecture, and four variants of Chronos, 8M-710M parameter encoder-decoder model.

Activation transplantation provides direct, causal control over model forecasts. Figure 2 demonstrates that forecast reversals follow our interventions deterministically, moving beyond correlation to causal evidence. This control is bidirectional: imposing a "crash" signature on a "calm" context reliably induces sharp downturn forecasts, while injecting "calm" statistics into a "crash" offsets the forecast toward stability, counteracting the downturn implied by Toto and Chronos baselines (green). *Notably, the offsets produced by both Toto and Chronos are of comparable magnitude for identical intervention cases.* Hence, these effects generalize across two architecturally distinct TSFMs, demonstrating that steerable crash and calm representations are not architecture-specific, but a robust property of large TSFMs. (Appendix C.1 for ablations with synthetic data).

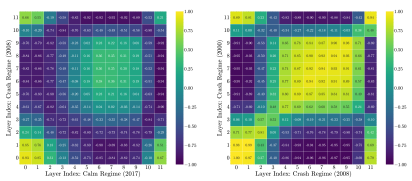
A unified "crash" concept emerges and solidifies across model depth. To further validate our findings, we project latent activation vectors from each layer onto the top 20 principal components



(a) Toto-Open-Base-1.0 (decoder-only) (b) Chronos-small (encoder–decoder)

Figure 2: **Forecast interventions via activation transplantation.** We intervene on model forecasts at $l = 8$ for both models by transferring statistical moments of hidden activations between regimes. **X-axis**: Time step in days **Y-axis**: NASDAQ 100 Index. **Top rows**: calm periods transplanted with crash statistics, which deterministically induce downturn forecasts simulating stress tests. **Bottom rows**: crash periods transplanted with calm statistics, which suppress downturns and restore stability. Shaded regions show 50% and 90% prediction intervals for the intervened forecasts, while green line indicates median forecasts by Toto and Chronos respectively. (Chronos Ablations in Section D)

via Principal Component Analysis (PCA) (Refer to Appendix F for ablations), thereby isolating the dominant low-dimensional structure of the representations. Within this reduced subspace, we compute cosine similarities for both within-regime and cross-regime inputs. As shown in Figure 3, latent layers exhibit consistently high similarity in this principal subspace for similar regime, indicating that distinct market regimes are encoded in a shared, lower-dimensional core representation (Appendix C.2 for ablations). This provides evidence that high-level event concepts, such as market crashes and calm periods, emerge as compact and geometrically coherent directions in the latent space. *Because regime information concentrates in a low-dimensional core, transplanting activation statistics perturbs shared concept directions and steers forecasts to the implanted regime.*



(a) Crash–Calm (b) Crash–Crash

Figure 3: Cross-regime similarity in reduced latent subspace. (a) crash–calm pairs are strongly anti-correlated in early layers but gradually align. (b) crash–crash pairs rapidly converge into a coherent latent subspace by mid layers.

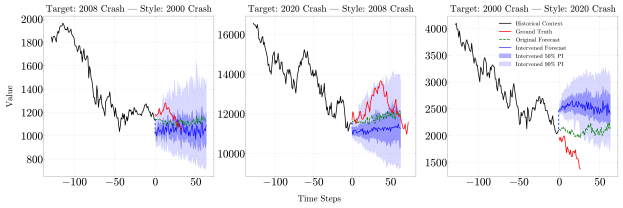


Figure 4: Cross-crash interventions reveal graded severity. Forecasts generated by transplanting crash signatures show that the forecast trajectory systematically deepens under severe signatures and is mitigated under milder ones, demonstrating that TSFMs encode crash events along a continuous latent severity axis.

Learned concepts are nuanced and quantitatively encode event severity. Having established that models distinguish calm and crash regimes in latent space, we next ask whether the notion of a "crash" is monolithic or admits fine-grained distinctions. To probe this, we transplant activation signatures across historical crashes (Figure 4). These cross-crash interventions reveal that the Dot-com (2000) signature induces sharper declines than the 2008 crisis, while the milder 2020 signature mitigates downturns. This demonstrates that TSFMs organize market regimes within a continuous semantic space, where each crash corresponds to a distinct point with interpretable magnitude.

4 Conclusion

This study provides the first causal evidence that TSFMs encode abstract, steerable market regime concepts. Activation transplantation deterministically controls forecasts across architectures, with crash signatures inducing downturns and calm signatures restoring stability. Cross-crash interventions re-

veal a continuous severity axis, embedding historical events as addressable latent directions.Together, these findings move beyond post-hoc interpretation toward causal manipulation of hidden states, enabling steerable, risk-aware "what-if"" forecasting.

References

Sabeen Ahmed, Ian E. Nielsen, A. Tripathi, Shamoon Siddiqui, R. Ramachandran, and G. Rasool. Transformers in time-series analysis: A tutorial. *Circuits, Systems, and Signal Processing*, 42:7433 – 7466, 2022. URL https://api.semanticscholar.org/CorpusId:248505796.

A. Alexandrov, Konstantinos Benidis, Michael Bohlke-Schneider, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Danielle C. Maddix, Syama Sundar Rangapuram, David Salinas, J. Schulz, Lorenzo Stella, Ali Caner Türkmen, and Bernie Wang. Gluonts: Probabilistic time series models in python. *ArXiv*, abs/1906.05264, 2019. URL https://api.semanticscholar.org/CorpusId:186206975.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series, 2024. URL https://arxiv.org/abs/2403.07815.

Ben Cohen, Emaad Khwaja, Kan Wang, Charles Masson, Elise Ramé, Youssef Doubli, and Othmane Abou-Amal. Toto: Time series optimized transformer for observability, 2024. URL https://arxiv.org/abs/2407.07874.

Rama Cont and Peter Tankov. *Financial Modelling with Jump Processes*. Chapman & Hall/CRC Financial Mathematics Series. Chapman and Hall/CRC, 2004.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting, 2024. URL https://arxiv.org/abs/2310.10688.

Anubha Goel, P. Pasricha, and J. Kanniainen. Time-series foundation ai model for value-at-risk forecasting. In *unknown*, 2024. URL https://api.semanticscholar.org/CorpusId:273351247.

Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting, 2020. URL https://arxiv.org/abs/1912.09363.

Fuqiang Liu, Sicong Jiang, Luis F. Miranda-Moreno, Seongjin Choi, and Lijun Sun. Adversarial vulnerabilities in large language models for time series forecasting. In *International Conference on Artificial Intelligence and Statistics*, 2024. URL https://api.semanticscholar.org/CorpusId:274638157.

Robert C. Merton. Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 3(1):125–144, 1976. ISSN 0304-405X. doi: https://doi.org/10.1016/0304-405X(76)90022-2. URL https://www.sciencedirect.com/science/article/pii/0304405X76900222.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Ozan Ozyegen, Juyoung Wang, and Mucahit Cevik. Danlip: Deep autoregressive networks for locally interpretable probabilistic forecasting, 2023. URL https://arxiv.org/abs/2301.02332.

Owen Queen, Thomas Hartvigsen, Teddy Koker, Huan He, Theodoros Tsiligkaridis, and Marinka Zitnik. Encoding time-series explanations through self-supervised model behavior consistency, 2023. URL https://arxiv.org/abs/2306.02109.

Syama Sundar Rangapuram, M. Seeger, Jan Gasthaus, Lorenzo Stella, Bernie Wang, and Tim Januschowski. Deep state space models for time series forecasting. In *Neural Information Processing Systems*, 2018. URL http://papers.nips.cc/paper/8004-deep-state-space-models-for-time-series-forecasting.

Sima Siami-Namini, Neda Tavakoli, and A. Namin. The performance of lstm and bilstm in forecasting time series. *2019 IEEE International Conference on Big Data (Big Data)*, pages 3285–3292, 2019. URL http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9005997.

Xianyun Wen and Weibang Li. Time series prediction based on lstm-attention-lstm model. *IEEE Access*, 11: 48322–48331, 2023. URL https://api.semanticscholar.org/CorpusId:258731343.

Michał Wiliński, Mononito Goswami, Willa Potosnak, Nina Żukowska, and Artur Dubrawski. Exploring representations and interventions in time series foundation models, 2025. URL https://arxiv.org/abs/2409.12915.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers, 2024. URL https://arxiv.org/abs/2402.02592.

A Limitations and Future Work

While our work provides a foundational proof-of-concept for semantic control, it represents the first step into an exciting research landscape. Researchers can investigate whether these steerable semantic subspaces are a universal property of sequence models, extending beyond Transformers to architectures like State Space Models Rangapuram et al. [2018], Alexandrov et al. [2019] or LSTM based time series models Wen and Li [2023], Siami-Namini et al. [2019], Lim et al. [2020]. Furthermore, our intervention technique opens the door to discovering a broader "semantic vocabulary" within TSFMs, probing for concepts like earnings surprises in finance or seizure onsets in EEG data. Our findings invite a new research program focused not just on what these models can predict, but on the rich, internal worlds they build to do so.

B Experimental Setup

The experiments are designed to show the robustness and generalizability of our findings. This is reflected in our choice of models, data, and parameters.

B.1 Models and Hardware

To validate our claims across different architectural paradigms and scales, we employ two distinct foundation models. The primary model for our main figures is **Toto-Open-Base-1.0** Cohen et al. [2024], a 103M parameter decoder-only Transformer. To confirm our findings are not an architectural artifact, we replicate key experiments on four variants of **Chronos-T5** Ansari et al. [2024], an encoder-decoder architecture, ranging from the 8M parameter `tiny` variant to the 710M parameter `large` model. All experiments were conducted on a single NVIDIA A6000 GPU to ensure a consistent and reproducible hardware environment. We performed all our intervtions

B.2 Real-World Data

For our primary experiments, we use daily closing values of the NASDAQ-100 index. We extract multiple historical periods to serve as both `Target` and `Style` inputs, detailed in Table 1. In Figure 2, the point "0" marks the end of the input and the start of the prediction. **The** 0 **points for all subplots in Figure 2 correspond to the End Date in Table 1.**

- **Input Context Length:** To create a continuous time series suitable for the model and account for non-trading days (weekends, holidays), we fill frequency gaps by inserting missing dates and imputing their values using the previous day's index. From this continuous, imputed series, we use a fixed input length of **128 time steps** (days). This choice is deliberate; the Toto model utilizes a patch size of 64, and an input length of 128 ensures the model processes exactly two full, non-overlapping patches, providing a clean and consistent representational structure for our interventions.

Table 1: Historical periods from the NASDAQ-100 index used for `Target` and `Style` inputs in our experiments.

Regime Name	Semantic Type	Start Date	End Date
2017 Calm	Calm	2017-01-12	2017-05-20
2007 Calm	Calm	2007-03-12	2007-07-18
2019 Calm	Calm	2019-06-01	2019-10-07
2008 Crash	Crash	2008-07-25	2008-11-30
2000 Crash	Crash	2000-08-31	2001-01-06
2020 Crash	Crash	2020-01-30	2020-06-06

Note: The end dates in the above Table reflect the final day of the *input context* fed to the model. For visualization purposes in our figures, we naturally use an extended segment of the time series, as can be seen in our code, to plot the subsequent ground truth against which our forecasts are evaluated.

B.3 Controlled Synthetic Data

To provide a controlled validation of our severity-encoding hypothesis (Section 3), we generate synthetic crash signals.

- **Input Context Length:** For these experiments, we use a longer input context of **256 time steps**. This allows us to observe how the model's behavior adapts when presented with a longer historical context than in the primary experiments.
- **Sampling:** For each forecast, we generate **256 samples** from the model's output head. This ensures a rich, well-defined predictive distribution, allowing for a robust analysis of both the median forecast and its associated uncertainty.

Synthetic Crash and Calm Regimes Generation

To systematically probe model behavior under controlled conditions, we generate synthetic financial time series using a discrete-time jump–diffusion process, a discretized analogue of the Merton model [Merton, 1976, Cont and Tankov, 2004].

Let S_t denote the price and $X_t = \log S_t$ the log-price. The dynamics are given by

$$X_{t+1} \;=\; X_t \;+\; \left(\mu - \tfrac{1}{2}\sigma^2\right) \;+\; \sigma\varepsilon_t \;+\; J_t, \qquad \varepsilon_t \sim \mathcal{N}(0,1), \tag{5}$$

where the jump term is defined as

$$N_t \sim \text{Poisson}(\lambda), \qquad J_t \;=\; \sum_{k=1}^{N_t} Z_{t,k}, \qquad Z_{t,k} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_J, \sigma_J^2). \tag{6}$$

Here N_t is the random number of jumps that occur between t and $t+1$. If $N_t = 0$, then $J_t = 0$ and the update is purely drift–diffusion; if $N_t = 1$, one Gaussian jump $Z_{t,1}$ is added; if $N_t = 2$, two independent shocks occur and $J_t = Z_{t,1} + Z_{t,2}$, and so on. Thus N_t controls *how many jumps occur*, while $Z_{t,k}$ controls *how large each jump is*.

Calm regimes: We define calm periods as stable markets with small positive drift, very low volatility, and no jumps:

$$\mu,\, \sigma,\, \lambda,\, \mu_J,\, \sigma_J = 2 \times 10^{-4},\; 3 \times 10^{-3},\; 0,\; 0,\; 0. \tag{7}$$

Crash regimes: By contrast, crash periods are instantiated by scaling parameters with a severity factor s: increasing volatility, amplifying negative drift, and introducing rare but negative jumps. Formally,

$$\begin{aligned} \mu(s) &= -8 \times 10^{-4}\, s, \quad & \sigma(s) &= 8 \times 10^{-3}\, s, \\ \mu_J(s) &= -2 \times 10^{-2}\, s, \quad & \sigma_J(s) &= 10^{-2}\sqrt{s}, \\ \lambda(s) &= 5 \times 10^{-2}\, s. \end{aligned} \tag{8}$$

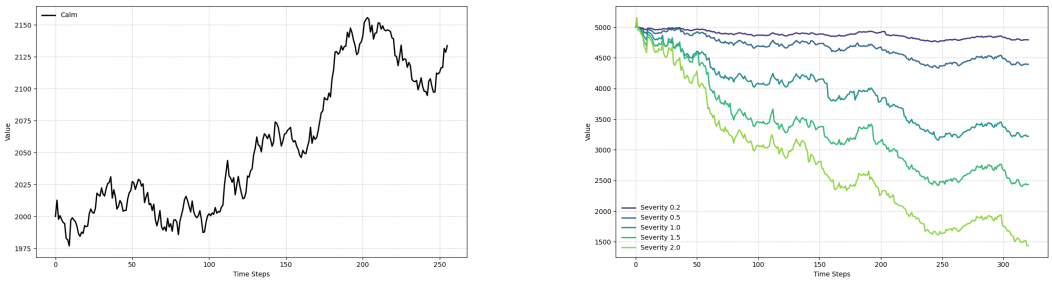
This synthetic data framework provides a principled method for generating realistic calm and crash regimes, with a tunable severity parameter s that controls the magnitude of systemic shocks. Figure 5 illustrates the synthetic calm and crash regimes used for the experiments in this work. By varying s, we obtain diverse univariate log-price time series that capture crashes of different intensities.

C Analysis on Synthetic data

To further support our claims, we conduct detailed experiments using synthetic data, since real crash–calm data is not readily available. In this context, and to remain aligned with the financial domain, we generate synthetic datasets based on established techniques.

C.1 Controlled Validation of Severity Encoding

Synthetic experiments confirm a continuous, severity-dependent encoding of crashes. To isolate and rigorously test the model's nuanced understanding of crash severity observed in the main paper,



(a) Synthetic Calm Series (b) Synthetic Crash Series

Figure 5: **Synthetic series generation.** (a) Calm trajectory initialized at $X_0 = 2000$ and (b) crash trajectory initialized at $X_0 = 5000$, both generated using Eq. 5 with parameters specified in Eq. 7 and Eq. 8, respectively.

we designed a controlled experiment using synthetic data. This allows us to move beyond the discrete, historical examples of the 2000, 2008, and 2020 crashes and probe the model's response to a continuous spectrum of event magnitudes.

Figure 6 shows the result of intervening on a calm historical context by transplanting activation statistics from synthetically generated crash signals of varying intensity, from a mild Severity = 0.2 to an extreme Severity = 2.0. The intervened forecasts exhibit a perfect dose-response relationship: a low severity of 0.2 induces only a mild, stabilizing downturn relative to the baseline, whereas increasing the severity to 1.0 and 2.0 produces progressively steeper and more dramatic forecasted declines. The clear visual ordering of the forecasts further highlights the model's ability to interpret this quantitative information.

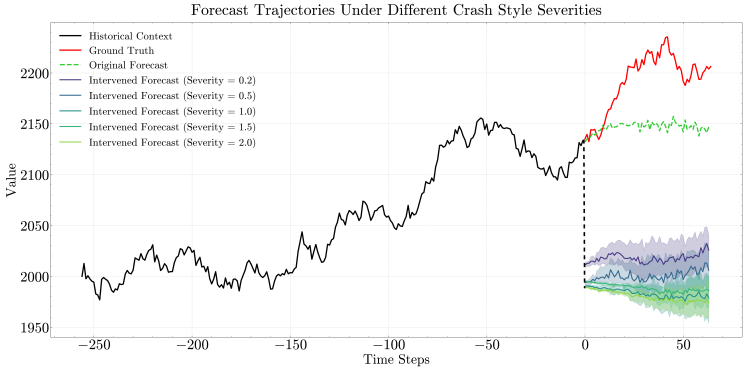


Figure 6: **Synthetic severity interventions for Toto-Open-Base-1.0**. Forecasts generated by transplanting activation statistics from synthetic crash signals of increasing intensity (Severity = 0.2–2.0) into a calm context. The forecasts exhibit a clear severity–response relationship, with downturn magnitude and predictive uncertainty both scaling with severity.

This experiment provides two crucial insights.

1. Model's internal representation of a "crash" is not a binary switch but a continuous, quantifiable concept. The intervention's magnitude directly and proportionally controls the forecasted outcome, confirming that our hypothesis of the model learning from a continuous space of severity is correct.

2. the prediction intervals also widen with increasing severity. This mirrors our finding with real-world data and reveals a sophisticated and consistent model behavior: more extreme interventions, which push the internal state further from its original context, correctly result in higher aleatoric uncertainty about the future path.

C.2 Synthetic Validation of Cross-Regime Representational Geometry

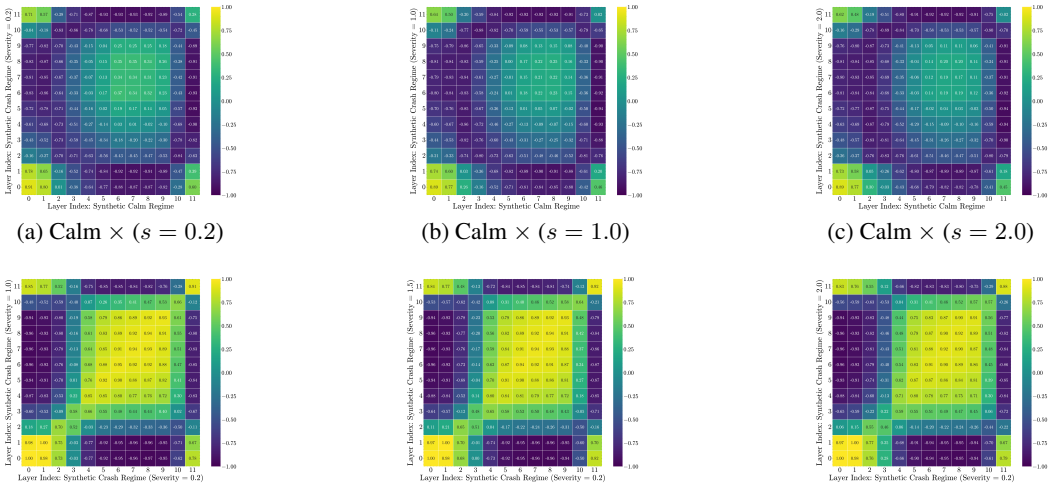
Cosine Similarity: Cosine similarity measures the cosine of the angle between two vectors, providing a simple yet effective way to assess similarity. Given two layer activation matrices $A^{(l)}$ and $A^{(m)}$, representing activations from layers l and m, the cosine similarity is computed as:

$$\text{cosine\_similarity}(\mathbf{A}^{(l)}, \mathbf{A}^{(m)}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{a}_i^{(l)} \cdot \mathbf{a}_i^{(m)}}{\|\mathbf{a}_i^{(l)}\| \|\mathbf{a}_i^{(m)}\|} \tag{9}$$

where $\mathbf{a}_i^{(l)}$ and $\mathbf{a}_i^{(m)}$ are the i-th activation vectors from layers l and m, and n is the number of samples.

We extend the controlled experiment by pairing the previously generated synthetic crash sequences with corresponding synthetic calm data. The goal is to test whether the representational geometry observed with real events generalizes beyond specific historical contexts.

Cross- and inter-regime representational similarity. Figure 7 shows cosine similarity matrices between synthetic crash regimes of varying severity and a calm baseline, projected into the PCA-reduced latent subspace of the Toto-Open-Base-1.0 model. As severity (s) increases, cross-regime similarity decreases consistently across depth, indicating that the model separates mild from extreme crashes more distinctly in latent space. This mirrors our real-world findings: just as historical crashes align along a continuous severity axis, synthetic crashes exhibit progressively lower similarity to calm regimes as intensity grows. These results confirm that severity encoding is not tied to idiosyncratic market episodes, but reflects a generalizable and quantifiable mechanism for representing systemic shocks.



(a) Calm × ($s = 0.2$) (b) Calm × ($s = 1.0$) (c) Calm × ($s = 2.0$)

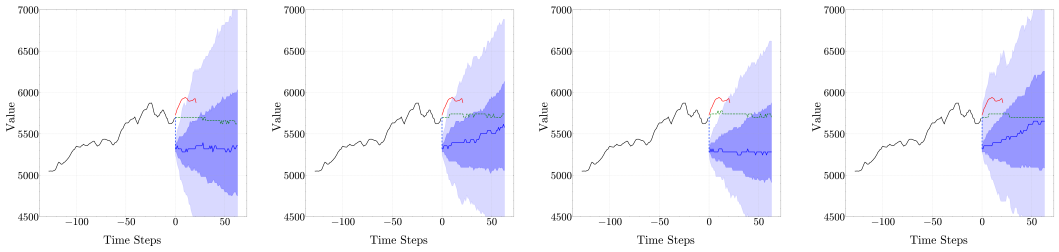
(d) ($s = 0.2$)×($s = 1.0$) (e) ($s = 0.2$) × ($s = 1.5$) (f) ($s = 0.2$) × ($s = 2.0$)

Figure 7: Synthetic regime similarity. Cosine similarity matrices in the PCA-reduced latent subspace (20 components) of Toto-Open-Base-1.0. Top row: cross-regime comparisons between synthetic crash and calm sequences at increasing severity levels ($s = 0.2, 1.0, 2.0$). Bottom row: within-regime comparisons across synthetic crash sequences of different severity.

Within-regime representational similarity. We next compare synthetic crash regimes of increasing severity against a mild-severity baseline ($s = 0.2$). Although similarity gradually decreases as the severity gap widens, the correlations remain consistently high across layers. This indicates that while the model is sensitive to graded differences in severity, all crash regimes are still anchored within a coherent latent subspace. In other words, the notion of a "crash" is stable and abstract, with severity expressed as a continuous modulation rather than a categorical separation.

D Causal Interventions are Invariant of Model-Size

Since the Toto-Open-Base-1.0 model is released only in a single configuration, we perform model size ablations using Chronos, which is available in multiple variants. As shown in Figure 8, activation transplantation yields consistent interventions across all Chronos variants, demonstrating that causal controllability is invariant to model size and thus reflects a fundamental property of time series foundation models, rather than a capacity artifact.



(a) tiny (8M) (b) small (46M) (c) base (200M) (d) large (710M)

Figure 8: **Chronos model interventions across scales:** Intervened forecasts for tiny, small, base, and large Chronos variants with 2008 crash semantics transplanted into 2017 calm data. The number of model parameters is indicated in parentheses for each variant.

E Activation-Space Similarity Across Model Depth

A unified "crash" concept emerges and solidifies across model depth. In addition to analyzing representations in the PCA-reduced subspace, **we also evaluate similarity directly in the original activation layers**. To probe the mechanism that enables our interventions, we quantify the representational geometry of market regimes using cosine similarity across layers (Table 2). Early layers show strong negative correlation between crash and calm regimes, capturing their antagonistic nature at the level of local features. By mid-depth (e.g., Layer 5), distinct crash events begin to collapse into a shared subspace, and in the final layers they converge into a highly aligned representation. This progression indicates the emergence of a stable and abstract "crash" concept, which provides a consistent semantic anchor for causal interventions to operate upon.

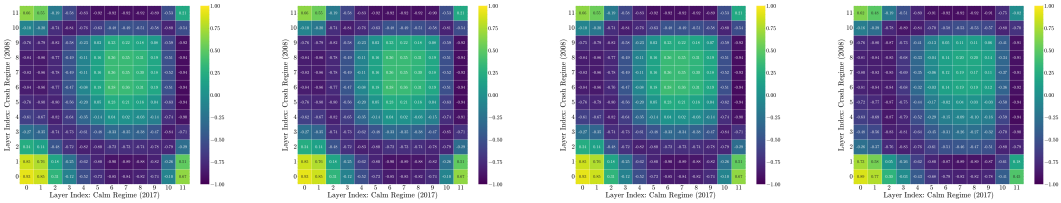
Table 2: Representational similarity: Cosine similarities between latent activation vectors across layers for cross-regime inputs for Toto model.

Events	L1	L2	L5	L9	L11
2008 (Crash) × 2017 (Calm)	-0.311	-0.194	0.327	0.562	0.741
2000 (Crash) × 2019 (Calm)	-0.416	-0.228	0.240	0.533	0.684
2000 (Crash) × 2008 (Crash)	0.193	0.471	**0.915**	**0.932**	**0.968**
2008 (Crash) × 2020 (Crash)	0.152	0.439	**0.882**	**0.938**	**0.957**

F Ablation on Dimensionality Reduction for Similarity Analysis

The choice of PCA components is critical for isolating the robust semantic signal from high-frequency noise. A core component of our analysis is measuring similarity between high-dimensional activation vectors. To do this robustly, we first project the activations onto their top k principal components. The choice of k is a crucial hyperparameter: too low, and we risk losing the core semantic information; too high, and we risk including instance-specific noise that obscures the underlying conceptual structure. Figure 9 provides the justification for our choice of k=20 in the main paper by showing the results for higher values: k=30, k=40, and k=50. The analysis reveals a clear and consistent trend: as we include more principal components beyond our chosen threshold, the clean, interpretable structure of the similarity matrix begins to diffuse.

We observed in Figures 3 and 7 that early layers show anti-correlation, mid-layers transition, and late layers achieve stable orthogonality. This clear pattern becomes progressively corrupted with



(a) $k = 20$ (b) $k = 30$ (c) $k = 40$ (d) $k = 50$

Figure 9: **PCA ablation on representational similarity.** Cosine similarity matrices for crash–calm regimes under different choices of principal components. Increasing k beyond 20 progressively diffuses the structured separation between regimes, supporting our choice of $k = 20$ in the main analysis.

noisy artifacts. This is not because the higher components are meaningless, but because they capture finer-grained, time-step-specific variance that is irrelevant to the high-level, abstract concept of "crash" or "calm".

This ablation demonstrates that the core semantic difference between market regimes is a relatively low-rank signal, effectively captured within the top 20 principal components. Our choice of k=20 is therefore a principled one, designed to maximize signal fidelity by capturing the entirety of the shared semantic concept while filtering out the high-frequency noise that varies from one specific market day to the next. This ensures that our main results reflect the model's understanding of the abstract concept itself, not the quirks of a particular historical instance.