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

Debdeep Sanyal[1,2], Aaryan Nagpal[1], Dhruv Kumar[1,3], Murari Mandal[1,2], Saurabh Deshpande[1]

[1]Birla AI Labs, Office of Ananya Birla, [2]KIIT Bhubaneshwar, [3]BITS Pilani

Paper GitHub

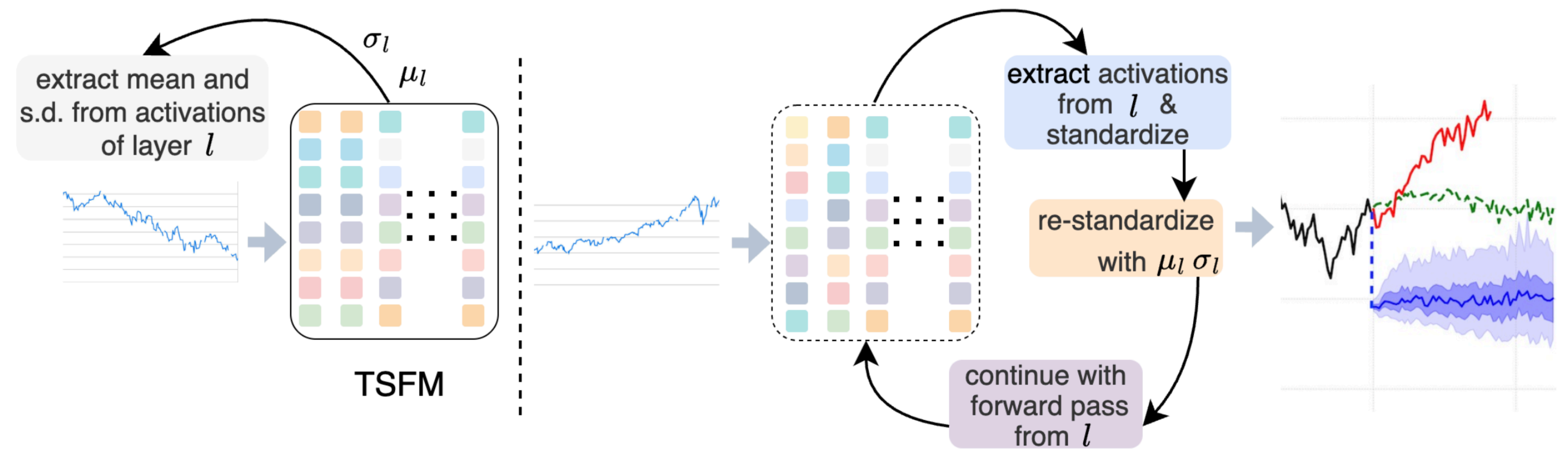
NEURAL INFORMATION PROCESSING SYSTEMS

BIRLA AI LABS

Motivation

Do Time Series Foundation Models understand market crashes?

Yes. We prove this via causal intervention.



Methodology

1. Extract Semantic Signature

We compute the mean and standard deviation vectors across the sequence length T_{in} of the input style, capturing the global dynamics.

$$\mu_l = \frac{1}{T_{in}} \sum A_l(X_{style}) \quad ; \quad \sigma_l = \sqrt{\frac{1}{T_{in}} \sum (A_l - \mu_l)^2}$$

2. Activation Transplantation

We standardize the target activations to remove their original context, then re-scale and shift them using the extracted style signature.

$$\tilde{A}_l = \left(\frac{A_l(X_{target}) - \mu_{target}}{\sigma_{target}} \right) \odot \sigma_{style} + \mu_{style}$$

3. Conditioned Forecast

The forward pass resumes from layer $l+1$ using the modified activation tensor $\tilde{A}_l$, producing a forecast conditioned on the implanted concept.

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

Data

Real World Data

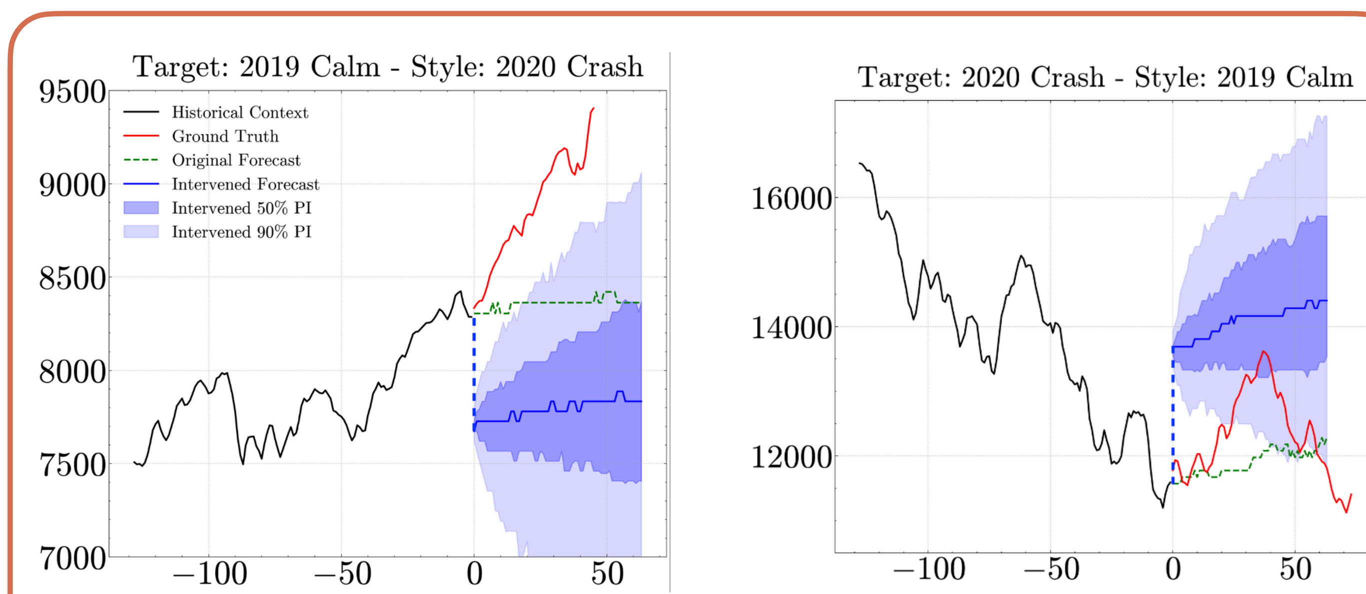
NASDAQ-100, containing periods of 2000 (Dot-com), 2008 (GFC), 2020 (COVID).

Synthetic Data

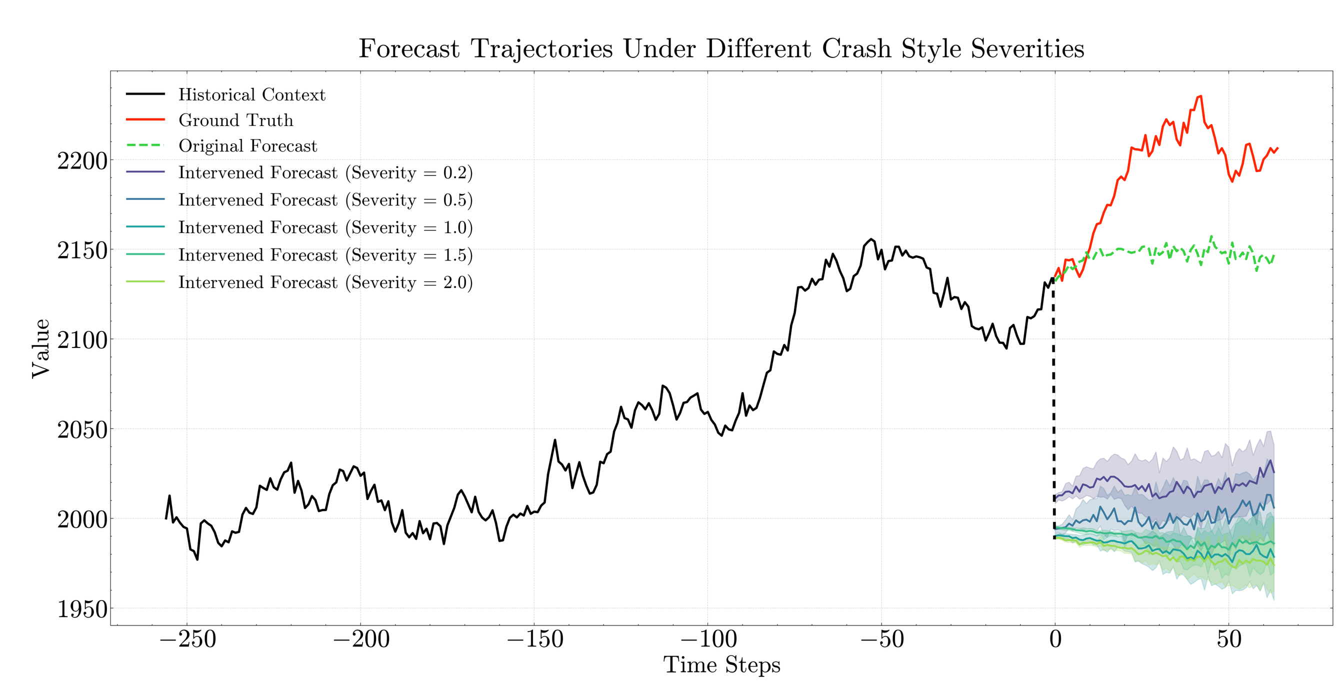
Discrete-time **Merton Jump-Diffusion** to isolate crash severity.

$$X_{t+1} = X_t + \underbrace{\left(\mu - \frac{1}{2}\sigma^2\right)}_{\text{Drift}} + \underbrace{\sigma\epsilon_t}_{\text{Diffusion}} + \underbrace{J_t}_{\text{Jump}}$$

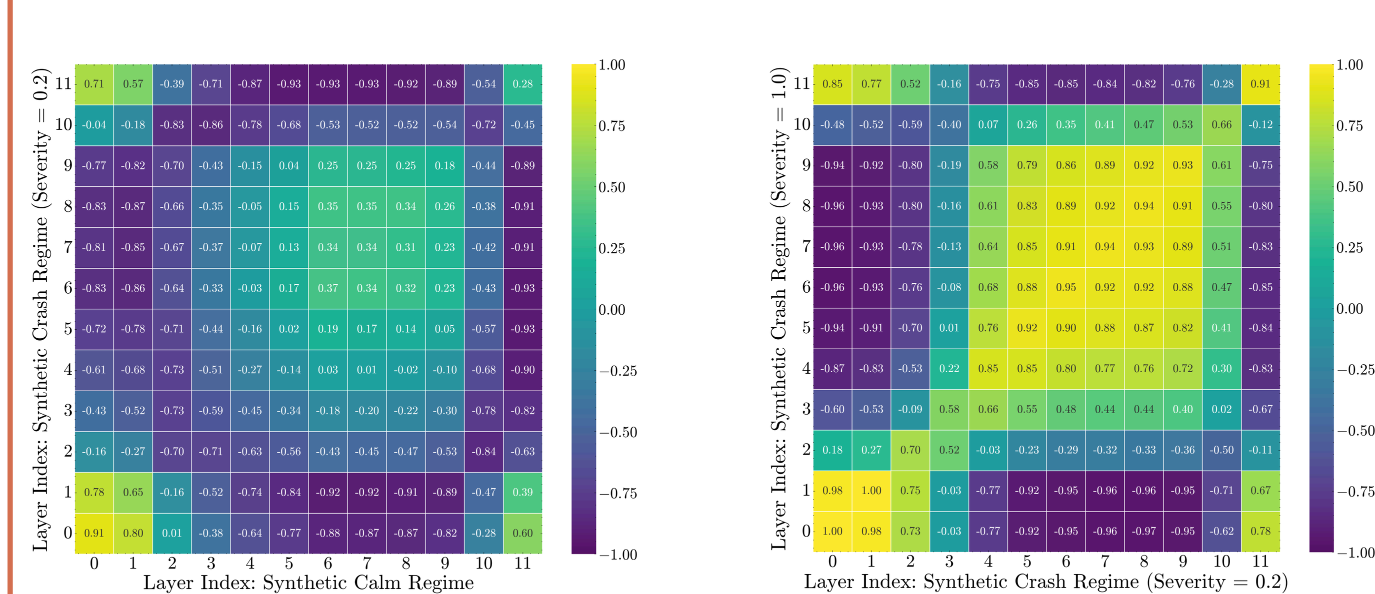
Experiments



Transplanting statistical moments (μ, σ) from a crash event into a calm context forces a sharp downturn prediction, reversing the model's original stable forecast. The converse (calm to crash) stands true as well.



Model exhibits a graded response to injection intensity. Increasing the crash severity scalar from mild to severe results in deeper forecasted downturns, confirming that "Crash" is encoded as a continuous manifold in the latent space.



Layer-wise cosine similarity reveals that opposing regimes become increasingly **orthogonal** in deeper layers, while distinct crash severities converge into a highly **aligned subspace**.