MacroVAE: Counterfactual Financial Scenario Generation via Macroeconomic Conditioning

Szymon Kubiak
City St George's, University of London
London, United Kingdom
szymon.kubiak@city.ac.uk

Tillman Weyde
City St George's, University of London
London, United Kingdom
T.E.Weyde@citystgeorges.ac.uk

Oleksandr Galkin
City St George's, University of London
London, United Kingdom
Oleksandr.Galkin@citystgeorges.ac.uk

Daniel Philps
University of Warwick
Coventry, United Kingdom
Daniel.Philps@wbs.ac.uk

Ram Gopal
University of Warwick
Coventry, United Kingdom
Ram.Gopal@wbs.ac.uk

Abstract

How would a portfolio perform under alternative macroeconomic conditions? Traditional scenario analysis in finance relies heavily on historical data, thus limiting risk assessment under rare or unobserved macroeconomic environments. We introduce MacroVAE, a variational autoencoder that generates realistic return sequences based on macroeconomic indicators to enable counterfactual scenario analysis. Trained on historical futures returns and macroeconomic data from global economies, MacroVAE generates return sequences conditioned on specified macroeconomic scenarios. The model uses convolutional ResNet blocks with Feature-wise Linear Modulation for stable macroeconomics-driven generation. In rolling out-of-sample evaluation, MacroVAE outperforms state-of-the-art generative baselines in replicating empirical distributions and financial stylized facts. We demonstrate two applications: counterfactual scenario analysis under alternative macroeconomic conditions, and forward-looking stress testing across diverse inflation-growth combinations, including ones rarely or not observed historically. MacroVAE enables systematic exploration of macroeconomic conditions, expanding the toolkit for portfolio risk management.

Keywords

Generative machine learning models, synthetic data, time-series, asset allocation, financial risk management

ACM Reference Format:
Szymon Kubiak, Tillman Weyde, Oleksandr Galkin, Daniel Philps, and Ram Gopal. 2025. MacroVAE: Counterfactual Financial Scenario Generation via Macroeconomic Conditioning. In *6th ACM International Conference on AI in Finance (ICAIF '25), November 15–18, 2025, Singapore, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3768292.3770360

1 Introduction

In 2022, surging inflation and aggressive monetary tightening reshaped global markets, triggering simultaneous losses in equities and bonds and challenging long-held assumptions about diversification [22]. These shocks illustrated how macroeconomic conditions can drive regime-dependent shifts in asset return distributions, altering means, volatilities, and correlations across portfolios [5, 6, 12]. Understanding these dynamics is essential for forward-looking risk management and scenario analysis.

Yet despite their importance, macroeconomic variables remain underused in data-driven financial models. This stems from modeling challenges: macro indicators are noisy, infrequent, and often lag market expectations; they operate at monthly or quarterly frequencies while asset returns evolve daily [13]; and their effects are nonlinear, time-varying, and difficult to disentangle from other market dynamics.

Most machine learning approaches in finance sidestep these challenges by either forecasting point estimates [15] or generating synthetic paths using recent price history [19, 21, 32]. Existing generative models that incorporate macroeconomic inputs do so alongside past returns, entangling macro conditions with recent price behavior and limiting purely counterfactual generation [26].

We introduce **MacroVAE**, a conditional variational autoencoder (VAE) [28] that learns how multivariate return distributions evolve with macroeconomic context, enabling generation conditioned solely on macro indicators. Trained on historical futures returns and macroeconomic data from global economies, MacroVAE generates return sequences for specified macroeconomic scenarios while preserving realistic multivariate structure.

Our contributions are:

- A generative model that captures multivariate return distributions as a function of macroeconomic context, evaluated using a rolling, out-of-sample framework
- Practical applications to macro-aware stress testing and counterfactual analysis, including scenarios rarely observed historically

In rolling out-of-sample evaluation, MacroVAE outperforms both historical benchmarks and leading generative baselines in replicating return distributions and generating realistic data that exhibits key financial stylized facts. To support adoption and further research, we release the full model and training code at https://github.com/szymkubiak/macrovae.

ICAIF '25, Singapore, Singapore

ACM ISBN 979-8-4007-2220-2/25/11
https://doi.org/10.1145/3768292.3770360

2 Related Work

Our work lies at the intersection of macro-finance modeling and deep generative models for financial data.

Macroeconomic Conditioning in Return Modeling. Macroeconomic variables such as inflation, interest rates, and economic growth significantly affect asset return dynamics. Empirical studies highlight state-dependent shifts in volatility, correlation, and tail behavior [5, 6, 12].

Traditional approaches that can incorporate macroeconomic variables include regime-switching models [16], threshold autoregressions [14], DCC-GARCH frameworks [10], and conditional copulas [7]. These models rely on strong parametric assumptions or fixed regime structures, limiting flexibility across varying macro environments.

Recent machine learning methods offer greater flexibility [15], but typically focus on point forecasts rather than full distributional modeling, restricting their use for scenario generation. We address this gap using a conditional generative model that learns how entire return distributions evolve with macro conditions.

Deep Generative Models for Financial Datasets. Generative Adversarial Networks (GANs) have been widely applied to financial data synthesis, with time series models such as TimeGAN [32], QuantGAN [31], and FIN-GAN [29] capturing stylized facts like volatility clustering and fat tails. Recent variants improve performance: CTS-GAN [19] uses a five-network architecture with conditional information for out-of-sample generation, while SigCWGAN [21] leverages path signatures for enhanced temporal consistency. Despite these advances, GANs remain prone to training instability and mode collapse [27].

Diffusion models offer improved sample quality and stability [18, 30], but lack explicit latent representations, limiting their ability to perform the controlled counterfactual analysis demonstrated in Section 6.1.

Beyond time series, generative models have been applied to other financial data modalities such as volatility surfaces [2], correlation matrices [20], and order books [8], highlighting the growing importance of synthetic data generation across quantitative finance.

Macroeconomic conditioning remains rare in generative modeling of financial data. Recent approaches [25, 26] require both macroeconomic variables and historical return sequences as inputs, but their dependence on observed return sequences constrains simulation of outcomes under novel macroeconomic scenarios.

VAEs for Time Series. VAEs [28] provide a stable, likelihood-based framework for generation and have been applied to financial tasks such as Value-at-Risk estimation [4] and causally-structured time series modeling [1]. TimeVAE [9] demonstrates VAEs' effectiveness for multivariate time series generation with interpretable temporal components, though it does not condition generation on external variables.

MacroVAE extends this line of research by enabling flexible return generation across diverse market regimes while maintaining the stability advantages of VAE training over adversarial approaches.

3 Model Description

MacroVAE generates multivariate return sequences under observed or counterfactual macroeconomic states using a conditional VAE architecture. The model incorporates macroeconomic context through Feature-wise Linear Modulation (FiLM) [24] applied to convolutional ResNet blocks [17] throughout the decoder.

Let $x \in \mathbb{R}^{T \times A}$ denote a matrix of daily returns with $T = 22$ trading days and $A = 16$ assets. Each sample is associated with a macroeconomic vector $c \in \mathbb{R}^{36}$ of indicators (described in Section 4.1). The model consists of an encoder, a condition encoder, and three decoder heads. Below, we describe each component in detail.

3.1 Return Encoder

The encoder maps each return matrix $x \in \mathbb{R}^{T \times A}$ into the parameters of a latent distribution over a low-dimensional variable $z \in \mathbb{R}^{d_z}$, capturing temporal and cross-sectional structure. It consists of two initial 2D convolutional layers, followed by four ResNet-style blocks [17] (each with GroupNorm, GELU activations, and skip connections). The output is flattened and mapped to the mean $\mu_z \in \mathbb{R}^{d_z}$ and log-variance $\log \sigma_z^2 \in \mathbb{R}^{d_z}$ of the approximate posterior $q(z \mid x)$.

We follow the standard variational autoencoder formulation, where the approximate posterior is defined as:

$$q_\phi(z \mid x) = \mathcal{N}(z; \mu_z, \mathrm{diag}(\sigma_z^2)), \quad \text{with} \quad \sigma_z = \exp(0.5 \cdot \log \sigma_z^2),$$

and latent samples are drawn via the reparameterization trick:

$$z = \mu_z + \sigma_z \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I).$$

where $\odot$ denotes element-wise multiplication.

Latent parameters μ_z, σ_z define posterior uncertainty and are unrelated to predicted return means or volatilities ($\hat{\mu}_x$, $\hat{\ell}_x$), which are modeled separately.

3.2 Condition Encoder

The condition encoder transforms the macroeconomic input vector $c \in \mathbb{R}^{36}$ into a lower-dimensional embedding $e_c \in \mathbb{R}^8$ that captures relationships among macro variables. The architecture consists of a linear projection followed by two residual blocks with 1D convolutions, GroupNorm, and GELU activations.

3.3 Decoder

The decoder reconstructs asset return sequences from the latent representation z, conditioned on the macroeconomic embedding e_c. It consists of three parallel output heads that generate:

- **Normalized returns** $\hat{x}_{\text{norm}} \in \mathbb{R}^{T \times A}$: return sequences normalized to unit standard deviation across the 22-day time dimension for each asset.
- **Mean returns** $\hat{\mu}_x \in \mathbb{R}^{1 \times A}$: average daily returns per asset in the normalized space over the 22-day sequence.
- **Log-volatilities** $\hat{\ell}_x \in \mathbb{R}^{1 \times A}$: natural logarithm of the standard deviation of daily returns per asset over the 22-day sequence.

Each decoder head takes z as input and consists of multiple ResNet blocks [17]. The normalized return decoder uses 4 ResNet blocks with 2D convolutions, while the mean and volatility decoders

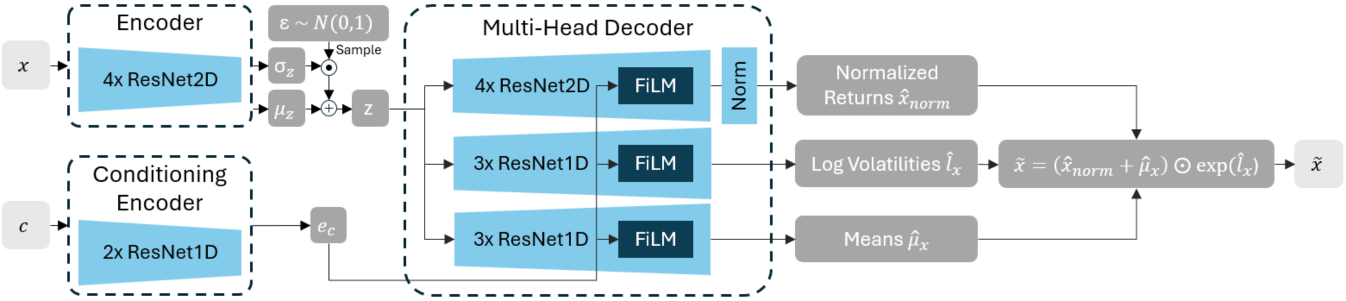

Figure 1: MacroVAE architecture. The model encodes 22-day return sequences into a latent variable z, and maps macroeconomic variables into embeddings e_c. The decoder generates normalized returns, average normalized returns, and return log-volatilities, with FiLM layers applying macro-driven modulation throughout. Final return sequences are reconstructed by combining all three outputs.

each use 3 blocks with 1D convolutions. At each layer, the macroeconomic embedding e_c modulates the features via Feature-wise Linear Modulation (FiLM) [24]:

$$\mathrm{FiLM}(h, e_c) = \gamma_{\mathrm{FiLM}}(e_c) \odot h + \beta_{\mathrm{FiLM}}(e_c)$$

where $\gamma_{\mathrm{FiLM}}(\cdot)$ and $\beta_{\mathrm{FiLM}}(\cdot)$ are learned linear transformations from the macro embedding e_c that apply elementwise scaling and shifting to the decoder's intermediate features h.

The final generated return samples are computed as:

$$\tilde{x} = (\hat{x}_{\mathrm{norm}} + \hat{\mu}_x) \odot \exp(\hat{\ell}_x),$$

producing return paths under given macro conditions.

3.4 Training Objective

The model is trained to minimize the following composite loss:

- **Normalized return reconstruction loss**: MSE between normalized empirical returns $x_{\mathrm{norm}}^{\mathrm{emp}}$ and the predicted normalized returns $\hat{x}_{\mathrm{norm}}$,
- **Log-volatility loss**: MSE between the empirical log-volatility $\ell_x^{\mathrm{emp}} = \log \sigma_x^{\mathrm{emp}}$ and the predicted log-volatility $\hat{\ell}_x$,
- **Mean loss**: MSE between empirical normalized mean returns μ_x^{emp} and predicted normalized means $\hat{\mu}_x$,
- **KL divergence**: Divergence between the approximate posterior $q(z \mid x)$ and the standard normal prior $p(z) = \mathcal{N}(0, I)$, scaled by an annealed weight β_{KL}. This regularization encourages a smooth and structured latent space, which supports stable sampling and enables the decoder to respond coherently to varying macroeconomic inputs.

The total loss is:

$$\mathcal{L} = \mathrm{MSE}(x_{\mathrm{norm}}^{\mathrm{emp}}, \hat{x}_{\mathrm{norm}}) + \mathrm{MSE}(\ell_x^{\mathrm{emp}}, \hat{\ell}_x) + \mathrm{MSE}(\mu_x^{\mathrm{emp}}, \hat{\mu}_x) + \beta_{\mathrm{KL}} \cdot D_{\mathrm{KL}}(q(z \mid x) \parallel p(z)).$$

We use the Adam optimizer (learning rate 0.0004), batch size 64, and train for 2000 epochs with gradient clipping (max norm 1.0). Uniform noise is added to macro inputs during training for regularization. The latent dimension is set to $d_z = 32$; all ResNet blocks use 128 hidden channels.

3.5 Design Rationale and Modeling Choices

Key design choices include:

Why a VAE?. We adopt a variational autoencoder to model return distributions in a stable, likelihood-based framework. Unlike GANs, VAEs offer tractable training and enable posterior sampling conditioned on macro inputs. This supports disentanglement between macroeconomic drivers and latent return dynamics, allowing diverse scenarios to be generated under fixed macro conditions by varying latent draws.

22-day return blocks. We model 22-day return sequences to match a typical monthly rebalancing horizon used by asset allocators.

ResNet architecture. We adopt convolutional ResNet blocks [17] to enable deeper architectures. Residual connections prevent gradient degradation and support learning of complex dependencies.

Macro conditioning via FiLM. FiLM layers [24] enable macroeconomic embeddings to modulate decoder activations throughout the network. This provides more effective conditioning than concatenation, as observed in our experiments.

Decomposed return modeling. The decoder separately predicts normalized returns, normalized mean levels, and log-volatilities, preventing high-volatility assets from dominating the loss and improving training stability. While only a single log-volatility is generated per return sequence, the normalized returns themselves are free to vary in volatility over time. This avoids imposing a constant-volatility assumption and allows the model to capture heteroskedastic behavior within each 22-day window.

4 Evaluation Framework

4.1 Dataset

We use daily futures returns covering 16 contracts across commodities, fixed income, and equity indices, calculated as percentage price changes. The asset selection follows [23], except for the XW1 contract, which we exclude due to differences in Bloomberg's pricing convention. The return data spans from May 2, 2000, to December 31, 2024. From this, we construct overlapping 22-day sequences of daily returns, which serve as inputs for training and evaluation.

To condition the generative model on macroeconomic context, we use 36 indicators from the US, EU, and China. We focus on these three economic blocs as they represent the majority of global GDP and are the primary drivers of international financial markets. The indicators include year-over-year changes in GDP, industrial production, retail sales, CPI, PPI, and M2 money supply for each region, as well as 6-month changes in their growth rates to capture macro momentum or deceleration. This vector provides comprehensive coverage of growth, inflation, and monetary policy conditions but limits dimensionality.

All macro data is aligned to actual release dates, and each 22-day return sequence is assigned the most recent macro values available before the first day of the sequence, ensuring that conditioning is based solely on information observable at the time. Indicators are standardized using in-sample means and standard deviations.

A full list of the futures contracts and macroeconomic indicators used is provided in the GitHub repository[1].

4.2 Training and Evaluation Procedure

We implement a rolling, out-of-sample evaluation framework to assess the realism and quality of generated return sequences in a manner consistent with practical investment applications.

The evaluation proceeds sequentially: for each calendar quarter, the model is trained on all historical data up to that point and then evaluated by generating return sequences for the following three months. The first model is trained on data through December 2014 and evaluated on January - March 2015, continuing quarterly through December 2024, providing 10 years of out-of-sample testing.

At each test step, the model generates 256 independent 22-day return sequences per month, conditioned on the macroeconomic vector observed the day before each sequence starts. All generated returns are strictly out-of-sample, ensuring no information leakage from training periods into evaluation.

We assess whether the conditional model can generate returns that reflect the statistical structure of held-out test data using Wasserstein distances and financial stylized facts. All baseline models and ablations follow the same evaluation procedure.

4.3 Model Comparison

To assess MacroVAE, we benchmark it against three groups of models: (1) baseline generative models adapted for macro-only conditioning, (2) historical return benchmarks, and (3) ablation variants of MacroVAE.

4.3.1 Baseline models adapted from prior work. Since most existing generative models condition on past returns rather than external variables, we adapt all baselines to rely solely on macroeconomic inputs for a consistent comparison.

- **CTS-GAN** [19]: Conditional extension of TimeGAN with a five-network architecture. We modified the model to remove historical return inputs. The generator receives concatenated macro features and random noise.
- **MC-TE-GAN** [26]: Transformer-based GAN. Although designed with macro conditioning in mind, the original implementation uses both macro variables and historical returns. We removed all historical return pathways and modified the generator to process only concatenated macro features and noise inputs.
- **TimeVAE** [9]: VAE with convolutional layers for time series generation. We adapted it by adding macro conditioning to the decoder through concatenation with latent variables. This serves as a simpler counterpart to MacroVAE, without FiLM-based conditioning or the decomposed return heads.

All baseline models use the same standardized macroeconomic vector and generate 22-day return sequences. Each model is trained with its recommended hyperparameters, modified only to accommodate the macro-only setup, and evaluated under the same rolling out-of-sample framework as MacroVAE.

4.3.2 Historical baselines. Simple baselines that draw from historical return sequences within fixed lookback windows (1 month, 1 year, and all available history). Unlike generative models that produce 22-day sequences, these baselines use the actual historical data structure to contextualize generative model performance relative to backward-looking strategies.

4.3.3 Ablation models. Variants isolating MacroVAE components:

- **Unconditional MacroVAE**: No macro conditioning.
- **MacroVAE without FiLM**: Macro input concatenated with latent z instead of being applied via FiLM modulation.
- **MacroVAE with LSTM decoder**: LSTM blocks replace convolutional layers in the normalized returns decoder, with concatenation-based macro conditioning (no FiLM).

4.4 Evaluation Metrics

We assess generative performance using two complementary approaches: distributional similarity measured via Wasserstein distance and statistical realism based on financial stylized facts.

4.4.1 Wasserstein Distance. Our primary metric is the Wasserstein distance ($\mathcal{W}_1$) between generated and realized return sequences, a standard evaluation metric for generative models [1, 11, 21]. For each test month, we compute this distance over flattened 22-day return matrices across all assets, measuring how closely synthetic distributions align with out-of-sample empirical data. To analyze performance across market conditions, we report average Wasserstein distances by deciles of realized S&P 500 volatility, as well as for the top 5% and top 1% most volatile days.

4.4.2 Distributional Moments and Stylized Facts. To assess realism, we compute distributional moments and stylized facts for the generated return sequences and compare them to those observed in the corresponding out-of-sample test periods. All metrics are computed separately for each rolling evaluation month, then averaged across all evaluation periods. Our evaluation metrics are inspired by [29].

Cross-sectional metrics across assets.

- Distributional moments: mean, volatility, skewness, kurtosis
- Average pairwise correlations
- Heavy tails: power-law tail exponents (α) estimated via maximum likelihood

Temporal stylized facts. (computed within each 22-day sequence, averaged across assets and lags):

- Linear unpredictability: autocorrelation of returns (lags 1–10)

[1] https://github.com/szymkubiak/macrovae

Table 1: Mean Wasserstein-1 distances (basis points) between generated and realized return distributions across S&P 500 volatility deciles and extreme periods (top 5% and 1%). Lower values indicate better performance, with best results in bold. MacroVAE achieves the lowest overall distance (22.4 bps), demonstrating superior distributional accuracy.

	All Test	S&P500 Volatility Deciles										Top 5%	Top 1%
	Data	1	2	3	4	5	6	7	8	9	10	Volatility	Volatility
MacroVAE (ours)	**22.4**	18.5	**13.7**	**14.7**	18.8	**15.5**	**14.7**	21.5	18.3	29.2	71.0	196.4	263.8
Baseline Models													
CTS-GAN	43.7	47.0	31.5	24.4	45.3	44.1	27.2	42.0	34.4	43.7	104.5	167.6	235.2
MC-TE-GAN	168.4	174.7	183.5	185.5	161.1	168.3	199.5	172.8	165.2	133.0	150.4	**162.0**	**170.4**
TimeVAE	68.2	35.1	33.4	82.9	59.6	68.6	36.3	109.5	74.3	84.1	117.2	214.9	280.6
Historical Benchmarks													
All History	26.6	33.2	25.5	24.5	20.5	19.6	18.5	19.8	18.5	**29.0**	**61.3**	178.1	242.1
Last 1 Year	25.3	21.0	18.5	14.8	18.7	18.2	19.3	30.8	21.3	33.3	72.4	186.9	263.3
Last Month	30.7	**16.8**	15.3	20.9	20.9	24.5	21.8	28.8	25.6	49.8	103.6	225.6	271.6
Ablation Models													
MacroVAE w/o FiLM	22.9	21.8	14.7	15.0	**17.6**	17.4	15.0	20.9	17.9	30.2	69.6	193.8	261.0
MacroVAE w/ LSTM	24.2	22.5	15.5	15.4	19.4	19.3	15.9	24.5	18.6	31.5	69.6	201.0	266.1
MacroVAE Uncond.	25.3	30.2	24.1	23.3	18.2	16.9	15.7	**19.3**	**17.2**	30.1	63.3	183.0	245.9

- Volatility clustering: autocorrelation of absolute returns (lags 1–10)
- Coarse-fine volatility: ratio of coarse ($|\sum r|$) to fine ($\sum |r|$) volatility over $\tau = 5$ days (lags 1–10)
- Leverage effect: covariance between returns and future squared returns, normalized by return variance (lags 1–10)

5 Evaluation Results

5.1 Distributional Performance

MacroVAE achieves the lowest average Wasserstein-1 distance across all models (22.4 basis points, bps), outperforming both baseline generative models and historical benchmarks (Table 1). Performance is strongest in low-to-moderate volatility regimes (deciles 1-6) while remaining competitive during high-stress periods. Notably, MacroVAE outperforms the best historical benchmark (Last 1 Year: 25.3 bps) and substantially beats other generative models, with CTS-GAN achieving 43.7 bps and TimeVAE 68.2 bps.

Ablation results show that removing macro conditioning increases Wasserstein distance by 13% (from 22.4 to 25.3 bps), while architectural changes - omitting FiLM or replacing convolutional layers with LSTMs - have smaller effects (22.9 and 24.2 bps respectively). This suggests macroeconomic information contributes more to distributional accuracy than the specific incorporation mechanism.

5.2 Financial Stylized Facts

MacroVAE generates samples that closely reflect key statistical features of financial time series (Table 2). It achieves exact matches on critical properties including skewness (-0.1), correlations (0.14), and temporal stylized facts such as linear unpredictability (-0.03) and volatility clustering (-0.03), while producing reasonable approximations for other characteristics. Tail behavior is somewhat underestimated, with generated returns exhibiting lighter tails than observed in empirical data.

Baseline generative models show deviations from empirical behavior. CTS-GAN exhibits unrealistic mean returns (-5.5 vs. 2.7 bps) and extreme tail exponent values (151.4 and 143.8 vs. 3.7 and 3.5), while MC-TE-GAN produces excessive volatility (219 vs. 98 bps) and mean returns (48.8 vs. 2.7 bps). TimeVAE performs better on volatility (72 bps) but overestimates correlations and tail exponents.

Historical benchmarks show mixed performance across metrics. The "All History" benchmark exhibits higher kurtosis (19.4 vs. 0.9) and severe volatility clustering artifacts (0.19 vs. -0.03), while "Last Month" achieves the best tail exponent match but has weaker overall performance.

Ablation results show that all MacroVAE variants perform well on stylized facts, with macro conditioning, FiLM layers, and architectural choices having relatively modest effects on statistical realism.

6 Macro-Conditional Scenario Generation

MacroVAE enables macro-conditioned scenario analysis for portfolio and risk management. We demonstrate two use cases: (1) **Counterfactuals** that simulate how returns would have evolved under alternative macro conditions, and (2) **Forward-looking stress tests** that assess portfolio behavior across hypothetical macro regimes, including those rarely observed in historical data.

Both applications rely on a macro shock propagation method. Given a non-standardized macro vector $c^{\text{orig}} \in \mathbb{R}^{36}$, we apply additive shocks δ to a selected subset $\mathcal{I}$, yielding $c^{\text{shock}}_{\mathcal{I}} = c^{\text{orig}}_{\mathcal{I}} + \delta$. The remaining variables are adjusted using the historical covariance matrix of the macro variables Σ: $c^{\text{shock}}_{-\mathcal{I}} = c^{\text{orig}}_{-\mathcal{I}} + \Sigma_{-\mathcal{I},\mathcal{I}} \Sigma^{-1}_{\mathcal{I},\mathcal{I}} \delta$. The resulting vector is then standardized using training-set statistics.

This approach helps generate plausible and internally consistent macro scenarios by leveraging historical co-movements. While we use this covariance-based adjustment for simplicity and interpretability, alternative propagation methods could be employed.

Table 2: Distributional characteristics and temporal stylized facts for test data and generated return sequences. Values closer to the test data indicate better realism, with the closest matches highlighted in bold (including ties). All values are computed across assets. Mean and Std values are in basis points.

	Distribution Characteristics					Heavy Tails		Temporal Stylized Facts			
	Mean	Std	Skew	Kurt	Corr	α_{pos}	α_{neg}	Lin. Unpred.	Vol Clust.	Coarse-Fine	Leverage
Test Data	2.7	98	-0.1	0.9	0.14	3.7	3.5	-0.03	-0.03	0.03	-0.8
MacroVAE (ours)	2.9	94	**-0.1**	0.2	**0.14**	6.6	6.5	**-0.03**	**-0.03**	**0.01**	-0.7
Baseline Models											
CTS-GAN	-5.5	106	0.0	**0.7**	0.10	151.4	143.8	-0.02	0.00	0.06	0.4
MC-TE-GAN	48.8	219	0.9	2.0	0.07	5.2	9.2	-0.04	-0.04	**0.01**	-0.2
TimeVAE	5.4	72	0.0	0.0	0.18	16.8	17.2	-0.01	-0.02	**0.01**	**-0.8**
Historical Benchmarks											
All History	2.2	122	-0.3	19.4	0.13	4.6	4.0	-0.01	0.19	0.35	-8.4
Last 1 Year	2.3	108	-0.2	4.1	**0.14**	**3.9**	4.0	-0.01	0.08	0.19	-4.0
Last Month	-1.2	106	**-0.1**	1.3	0.13	**3.9**	**3.7**	-0.04	**-0.03**	0.00	-1.0
Ablation Models											
MacroVAE w/o FiLM	3.1	**97**	**-0.1**	0.3	0.15	6.0	5.3	**-0.03**	**-0.03**	**0.01**	-1.4
MacroVAE w/ LSTM	3.1	94	**-0.1**	0.0	**0.14**	6.2	6.2	-0.04	**-0.03**	**0.01**	-0.6
MacroVAE Uncond.	**2.7**	106	**-0.1**	1.3	0.15	6.1	5.6	**-0.03**	**-0.03**	**0.01**	-0.5

6.1 Counterfactual Return Reconstruction via Macro Conditioning

Traditional scenario analysis replays past returns with fixed macro conditions, making it difficult to isolate macroeconomic effects. MacroVAE enables a more flexible approach by holding latent dynamics constant while varying macro inputs to reconstruct how returns might have evolved under alternative scenarios.

We demonstrate this using 2022, when US CPI inflation peaked at 9.1%, the highest since 1981, reflecting pandemic-driven stimulus, supply shocks, and tight labor markets [3]. Subsequent monetary tightening led to simultaneous equity and bond declines, disrupting their usual negative correlation and making 2022 an ideal test case.

6.1.1 Methodology: Latent-Preserving Resimulation. We simulate how 2022 return sequences would have evolved under lower inflation while holding all other latent return drivers fixed. The procedure involves:

1. **Segment returns**: Divide daily returns over the target period (e.g., 2022) into non-overlapping 22-day blocks.
2. **Encode**: Map each block x to its latent mean representation $z = \mu_z$ from the approximate posterior $q_\phi(z \mid x)$.
3. **Apply shocks**: For each 22-day block, take the macro vector observed prior to its first day and reduce its CPI YoY value by 1 to 6 percentage points in 1-point steps. Remaining variables are adjusted using the historical macro covariance matrix.
4. **Decode counterfactuals**: Pass the encoded z and shocked macro input to the decoder to generate a return block $\tilde{x}$.
5. **Aggregate**: Concatenate resimulated blocks to reconstruct the full return sequence for performance analysis.

6.1.2 Results and Insights. Figure 2 shows how return trajectories for S&P 500 futures (ES1), 10-year US Treasury futures (TY1), and an equal-weighted portfolio evolve under varying CPI shocks.

Return responses are nonlinear, with mild CPI reductions (up to 2ppt) having limited effect while stronger shocks lead to marked improvements. Under a -6ppt CPI shock, the equal-weighted portfolio gains +17%, 27ppt above the realized outcome, with ES1 returns rising more than TY1. Volatility generally falls as inflation declines, but these effects plateau beyond -4ppt, and ES1's volatility slightly increases under the strongest shock. Correlations also shift, with the ES1-TY1 correlation declining from +0.17 to -0.15 under the largest shock, restoring the negative relationship typical of stable regimes.

The abrupt shift in ES1's behavior between -3ppt and -4ppt shocks suggests MacroVAE has learned that markets transition between distinct inflation regimes - perhaps reflecting the difference between modest inflation that markets can accept versus more extreme conditions that fundamentally alter market dynamics, triggering different investor behaviors.

This case study demonstrates MacroVAE's ability to uncover directional sensitivities, nonlinear responses, and regime effects under controlled macro variation. The same methodology can be applied to different time periods, asset classes, and macroeconomic variables.

6.2 Forward-Looking Scenario Analysis under Sparse Macro Conditions

Assessing portfolio behavior across diverse macroeconomic environments is essential for risk management, but macro-driven scenario analysis is challenging due to indicator release frequencies, lags, nonlinear effects [13]. Moreover, many plausible macro configurations (such as high inflation with low growth) are rarely observed.

MacroVAE addresses these limitations by generating return distributions under arbitrary macro scenarios, including sparse or unobserved regimes. This allows risk managers to explore a full grid of macroeconomic inputs, rather than relying solely on historical episodes.

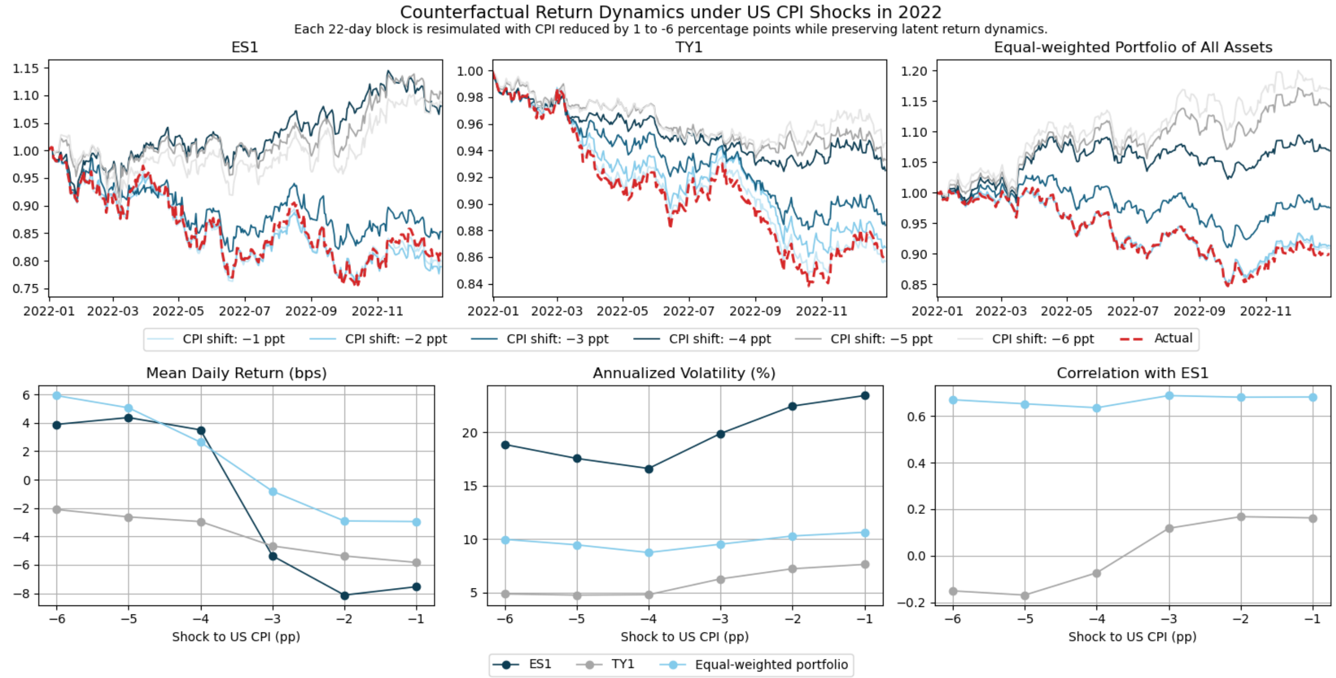

Figure 2: Counterfactual return dynamics under US CPI shocks in 2022. Each 22-day block is resimulated using its latent encoding with CPI YoY reduced by 1-6 percentage points. Other macro variables are adjusted via historical covariances. Top: cumulative returns for S&P 500 futures (ES1), 10-year Treasury futures (TY1), and equal-weighted portfolio. Bottom: changes in mean daily returns, volatility, and correlations.

6.2.1 Forward-Looking Stress Testing via Macro Grids. To assess portfolio performance across diverse macro environments, we construct a two-dimensional grid varying US CPI and GDP year-over-year over symmetric ranges for an equally-weighted portfolio of the 16 futures contracts. Figure 3 compares historical data (through December 2024) with synthetic scenarios:

- Top row: Historical statistics - observation counts, mean daily returns, and volatility for training data grouped into each grid cell.
- Bottom row: MacroVAE-generated statistics using 10,000 synthetic samples per cell, created by shocking the two variables and using the covariance-based propagation method to adjust remaining macro indicators.

6.2.2 Insights and Interpretation. Key insights from the macro grid analysis include:

- **Sparse historical coverage:** Many macro regimes, especially low-growth, low-inflation scenarios, have few or no historical observations, limiting empirical reliability.
- **Smooth and complete surfaces:** MacroVAE produces coherent surfaces across the full grid. Unlike the empirical heatmaps, the synthetic outputs are well-structured and densely populated, even in rarely observed regimes.
- **Pattern exploration:** The grid enables systematic exploration of macro-return relationships, such as how expected returns decline and volatility increases under weaker growth and lower inflation, along with other complex nonlinear sensitivities.
- **Rich scenario toolkit:** Each cell contains 10,000 simulated return paths, allowing computation of additional metrics (such as skewness, drawdowns, or factor exposures) for in-depth stress testing and portfolio analysis.

This shows how MacroVAE enables scenario analysis beyond what historical data alone can provide.

7 Limitations

While MacroVAE enables novel counterfactual scenario generation, several limitations should be acknowledged.

MacroVAE learns correlational rather than causal relationships between macroeconomic conditions and asset returns. The model treats macroeconomic conditions as exogenous, ignoring potential feedback effects between asset prices and macroeconomic variables. In reality, extreme conditions often coincide with policy interventions, central bank actions, or regulatory changes that could fundamentally alter return dynamics.

An implicit assumption in MacroVAE is that macroeconomic conditions influence returns without altering the underlying latent structure. In practice, however, the model may not fully disentangle macro-driven variation from residual latent factors, and some macro effects could be absorbed into the latent space. This could limit the model's ability to capture how macro shocks reshape return-generating processes.

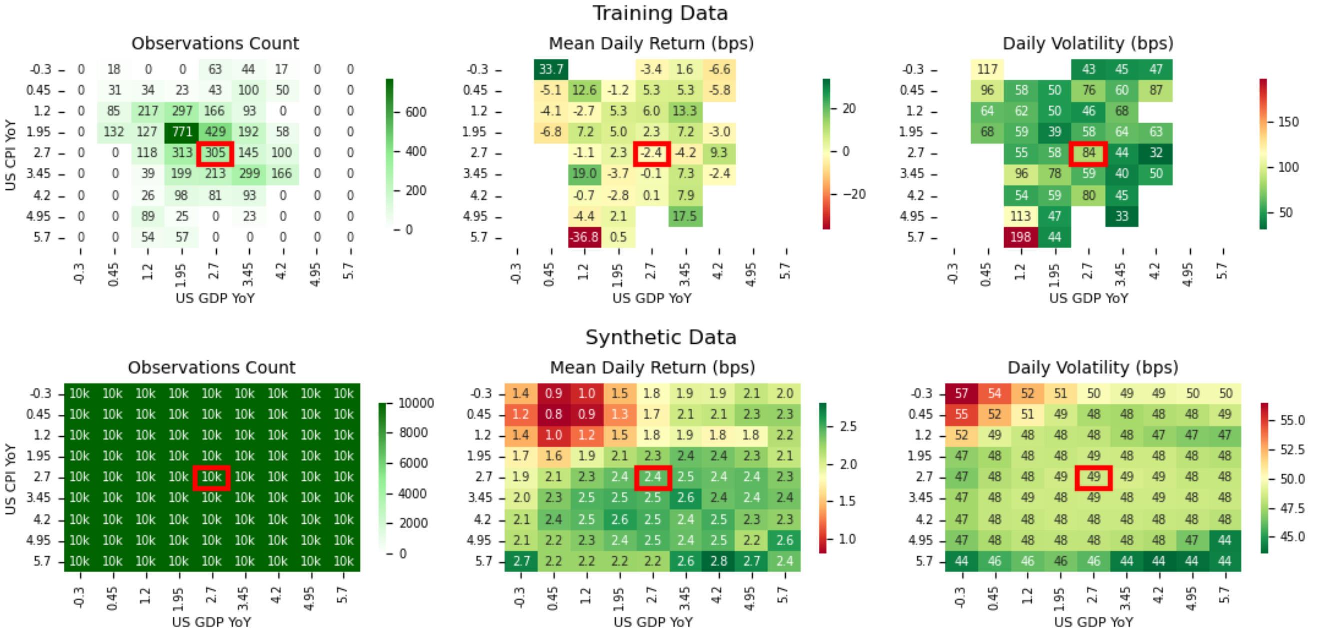

Figure 3: Macro-conditional scenario grid for an equally weighted portfolio of 16 futures contracts, conditioning on US CPI YoY (vertical) and US GDP YoY (horizontal). Top row: empirical results from training data (through 2024), showing observation counts, mean daily returns, and volatility (basis points) per macro configuration. Bottom row: MacroVAE-generated results using 10,000 simulations per cell. Red squares indicate current macroeconomic conditions. The synthetic surfaces provide smooth, complete coverage across the grid, including scenarios that are sparsely or never observed in the historical data.

The model's reliability is further limited when analyzing extreme or unprecedented macro conditions due to the absence of comparable historical data for validation. Additionally, structural changes in financial markets since the training period could reduce model relevance for scenarios that differ significantly from historical experience.

The model also relies on a specific set of 36 macroeconomic indicators from three regions, which may not capture all relevant macro dynamics or regional variations. Finally, the 22-day horizon may not capture how markets respond to macro shocks over longer periods.

8 Conclusion

We introduced MacroVAE, a conditional variational autoencoder that generates realistic asset return sequences conditioned on macroeconomic indicators. By disentangling macroeconomic context from latent return dynamics, MacroVAE enables counterfactual scenario generation and forward-looking stress testing beyond traditional historical approaches.

MacroVAE demonstrates superior empirical performance, achieving the lowest distributional distances to realized out-of-sample returns while reproducing financial stylized facts. Our applications illustrate practical value: counterfactual analysis reveals how alternative inflation scenarios would have altered 2022 portfolio outcomes, while forward-looking stress testing generates coherent scenarios across sparse macro regimes rarely observed historically.

Future work could extend MacroVAE by incorporating additional data modalities such as textual information from financial news, developing explainability mechanisms, incorporating causal inference techniques to enhance reliability under unprecedented macroeconomic conditions, and explicitly structuring the latent space to better disentangle macro-driven variation from residual factors.

References

[1] Beatrice Acciaio, Stephan Eckstein, and Songyan Hou. 2024. Time-Causal VAE: Robust Financial Time Series Generator. *arXiv preprint arXiv:2411.02947* (2024).

[2] Maxime Bergeron, Nicholas Fung, John Hull, and Zissis Poulos. 2021. Variational autoencoders: A hands-off approach to volatility. *arXiv preprint arXiv:2102.03945* (2021).

[3] Ben Bernanke and Olivier Blanchard. 2023. *What caused the US pandemic-era inflation?* Vol. 86. PIIE, Peterson Institute for International Economics.

[4] Robert Buch, Stefanie Grimm, Ralf Korn, and Ivo Richert. 2023. Estimating the value-at-risk by Temporal VAE. *Risks* 11, 5 (2023), 79.

[5] Tolga Cenesizoglu and Allan Timmermann. 2008. Is the distribution of stock returns predictable? *Available at SSRN 1107185* (2008).

[6] Nai-Fu Chen, Richard Roll, and Stephen A Ross. 1986. Economic forces and the stock market. *Journal of business* (1986), 383–403.

[7] Lorán Chollete, Andréas Heinen, and Alfonso Valdesogo. 2009. Modeling international financial returns with a multivariate regime-switching copula. *Journal of financial econometrics* 7, 4 (2009), 437–480.

[8] Rama Cont, Mihai Cucuringu, Jonathan Kochems, and Felix Prenzel. 2023. Limit order book simulation with generative adversarial networks. *Available at SSRN 4512356* (2023).

[9] Abhyuday Desai, Cynthia Freeman, Zuhui Wang, and Ian Beaver. 2021. Timevae: A variational auto-encoder for multivariate time series generation. *arXiv preprint arXiv:2111.08095* (2021).

[10] Robert Engle. 2002. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of business & economic statistics* 20, 3 (2002), 339–350.

[11] Lars Ericson, Xuejun Zhu, Xusi Han, Rao Fu, Shuang Li, Steve Guo, and Ping Hu. 2024. Deep generative modeling for financial time series with application in VaR: A comparative review. *arXiv preprint arXiv:2401.10370* (2024).

[12] Eugene F Fama and Kenneth R French. 1989. Business conditions and expected returns on stocks and bonds. *Journal of financial economics* 25, 1 (1989), 23–49.

[13] Eric Ghysels, Arthur Sinko, and Rossen Valkanov. 2007. MIDAS regressions: Further results and new directions. *Econometric reviews* 26, 1 (2007), 53–90.

[14] Clive WJ Granger and Timo Teräsvirta. 1993. *Modelling nonlinear economic relationships.* oxford university Press.

[15] Shihao Gu, Bryan Kelly, and Dacheng Xiu. 2020. Empirical asset pricing via machine learning. *The Review of Financial Studies* 33, 5 (2020), 2223–2273.

[16] Massimo Guidolin and Allan Timmermann. 2007. Asset allocation under multivariate regime switching. *Journal of Economic Dynamics and Control* 31, 11 (2007), 3503–3544.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770–778.

[18] Hongbin Huang, Minghua Chen, and Xiao Qiao. 2024. Generative learning for financial time series with irregular and scale-invariant patterns. In *The Twelfth International Conference on Learning Representations.*

[19] Riasat Ali Istiaque, Chi Seng Pun, and Yuli Song. 2024. Simulating Asset Prices using Conditional Time-Series GAN. In *Proceedings of the 5th ACM International Conference on AI in Finance.* 770–778.

[20] Szymon Kubiak, Tillman Weyde, Oleksandr Galkin, Daniel Philps, and Ram Gopal. 2024. Denoising Diffusion Probabilistic Model for Realistic Financial Correlation Matrices. In *Proceedings of the 5th ACM International Conference on AI in Finance.* 1–9.

[21] Shujian Liao, Hao Ni, Marc Sabate-Vidales, Lukasz Szpruch, Magnus Wiese, and Baoren Xiao. 2024. Sig-Wasserstein GANs for conditional time series generation. *Mathematical Finance* 34, 2 (2024), 622–670.

[22] Roderick Molenaar, Edouard Sénéchal, Laurens Swinkels, and Zhenping Wang. 2024. Empirical Evidence on the Stock–Bond Correlation. *Financial Analysts Journal* 80, 3 (2024), 17–36.

[23] Jochen Papenbrock, Peter Schwendner, Markus Jaeger, and Stephan Krügel. 2021. Matrix Evolutions: Synthetic Correlations and Explainable Machine Learning for Constructing Robust Investment Portfolios. *The Journal of Financial Data Science* 3, 2 (2021), 51–69.

[24] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[25] Matteo Rizzato, Julien Wallart, Christophe Geissler, Nicolas Morizet, and Noureddine Boumlaik. 2023. Generative Adversarial Networks applied to synthetic financial scenarios generation. *Physica A: Statistical Mechanics and its Applications* 623 (2023), 128899.

[26] Alexander Michael Rusnak and Stéphane Daul. 2024. Macroeconomic Conditioned Synthetic Financial Markets. In *Proceedings of the 5th ACM International Conference on AI in Finance.* 150–158.

[27] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems* 29 (2016).

[28] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems* 28 (2015).

[29] Shuntaro Takahashi, Yu Chen, and Kumiko Tanaka-Ishii. 2019. Modeling financial time-series with generative adversarial networks. *Physica A: Statistical Mechanics and its Applications* 527 (2019), 121261.

[30] Yuki Tanaka, Ryuji Hashimoto, Takehiro Takayanagi, Zhe Piao, Yuri Murayama, and Kiyoshi Izumi. 2025. CoFinDiff: Controllable Financial Diffusion Model for Time Series Generation. *arXiv preprint arXiv:2503.04164* (2025).

[31] Magnus Wiese, Robert Knobloch, Ralf Korn, and Peter Kretschmer. 2020. Quant GANs: deep generation of financial time series. *Quantitative Finance* 20, 9 (2020), 1419–1440.

[32] Jinsung Yoon, Daniel Jarrett, and Mihaela Van der Schaar. 2019. Time-series generative adversarial networks. *Advances in neural information processing systems* 32 (2019).