Int. j. adv. multidisc. res. stud. 2026; 6(4):1221-1235

Received: 20-06-2026
Accepted: 30-07-2026

International Journal of Advanced Multidisciplinary Research and Studies

ISSN: 2583-049X

Diffusion-Based Generative Modeling for Tail Risk Simulation and Systemic Stress Testing: A Deep Learning Framework for Financial Crisis Scenarios

[1] Amirreza Taheri, [2] Ali Divandari
[1] M.Sc. in Financial Management, Faculty of Financial Sciences, Kharazmi University, Iran
[2] Professor, Department of Business Administration, Faculty of Management, University of Tehran, Iran

DOI: https://doi.org/10.62225/2583049X.2026.6.4.6721 Corresponding Author: **Amirreza Taheri**

Abstract
This study develops and validates a conditional diffusion-based generative framework for simulating extreme tail events and conducting systemic stress tests in financial markets. Traditional risk models, including value at risk, conditional value at risk, and GARCH-type specifications, as well as more recent deep generative models such as generative adversarial networks and variational autoencoders, systematically fail to capture the heavy-tailed, high-dimensional, and regime-switching nature of financial crises. To address this gap, we propose a denoising diffusion probabilistic model conditioned on market state variables (volatility regimes, correlation structures, and macroeconomic proxies) and implemented using a time-conditioned U-Net architecture. The model is trained on daily returns of 50 S&P 500 equities and 5 major cryptocurrencies over the period 2018–2025, with out-of-sample testing in July-December 2025 and historical crisis benchmarking against the March 2020 COVID-19 selloff and the March 2023 banking turmoil. Extensive comparisons against historical bootstrap, GJR-GARCH, a recurrent Wasserstein GAN, and an LSTM variational autoencoder are performed using three tiers of metrics: statistical fidelity (maximum mean discrepancy, tail Kolmogorov-Smirnov statistic, tail concentration ratio), financial risk accuracy (absolute percentage errors for VaR and CVaR at 95%, 99%, and 99.9% confidence levels), and systemic risk relevance (ΔCoVaR, SRISK, plausibility, and novelty). The conditional diffusion model outperforms the evaluated baselines across the reported metrics in this experimental setting. At the 99.9% VaR level, the absolute percentage error is 5.4% compared to 18.3% for the best baseline. The tail concentration ratio is 1.02, indicating negligible bias, while the GAN yields 0.67 (severe underestimation). Under a combined stress condition, the diffusion model generates counterfactual scenarios with ΔCoVaR (8.7%) and SRISK (23.4%) that closely match real crisis values (9.0% and 24.0%), with an implausibility rate of only 0.3% and a novelty rate of 92%. Sensitivity analysis confirms robustness across hyperparameter choices. These results indicate that conditional diffusion models can provide improved statistical fidelity and practical value for forward-looking systemic risk assessment, offering a powerful alternative to existing generative methods in financial risk management.

Keywords: Diffusion Models, Generative Deep Learning, Tail Risk, Systemic Stress Testing, Conditional Value at Risk, Financial Time Series, Scenario Generation

1. Introduction

Related evidence on tokenized asset environments also shows that AI-based, network-aware risk modeling can help characterize cross-asset dependence and systemic transmission channels (Taheri, 2023) [37].

Financial markets are inherently fragile constructs, perpetually operating at the nexus of rational expectations and behavioral anomalies, of linear approximations and nonlinear realities (Alzahrani, 2025) [1]. Despite decades of progress in quantitative risk measurement and macroprudential regulation, the historical record unequivocally demonstrates that extreme market events—collectively referred to as tail risks—have neither diminished in frequency nor in severity. The global financial crisis of 2007-2008, the European sovereign debt crisis of 2010-2012, the Flash Crash of 2015, the COVID-19 induced market turmoil of 2020, and the inflationary shocks accompanied by unprecedented monetary tightening throughout 2022 and 2023 collectively serve as stark reminders that financial systems exhibit regime shifts, non-linear contagion channels, and sudden systemic collapses (Bilos & Gündüz, 2025) [3]. More recently, the rapid proliferation of digital assets, decentralized finance protocols, and high-frequency algorithmic trading has introduced entirely new forms of volatility and cross-asset contagion that challenge the foundational assumptions of classical risk management (Chen *et al*., 2025) [5]. In this evolving landscape, the capacity to simulate, generate, and systematically stress-test extreme yet plausible market scenarios have shifted from a niche academic exercise to a critical operational necessity for financial institutions, central banks, and regulatory bodies alike.

The academic literature on financial risk measurement has historically been anchored by parametric and semi-parametric frameworks. Among these, value at risk and conditional value at risk have emerged as the predominant tools for quantifying downside exposure (Cont *et al*., 2025) [6]. Value at risk estimates the maximum expected loss over a given time horizon at a specific confidence level, whereas conditional value at risk measures the average loss beyond the value at risk threshold. These metrics are deeply embedded in portfolio optimization algorithms, capital allocation models, and regulatory mandates such as the Basel III and IV frameworks (Daiya, 2024) [7]. However, a meticulous examination of the historical record reveals that both value at risk and conditional value at risk, when estimated through historical simulation or parametric assumptions such as normality or the Student t distribution, systematically understate the probability and magnitude of extreme losses (Gao *et al*., 2026) [8]. This phenomenon, broadly termed tail risk underestimation, arises because financial return series are neither independent nor identically distributed nor Gaussian. Instead, they exhibit volatility clustering, leverage effects where negative shocks exert a disproportionately large impact on future volatility, heavy tails that deviate substantially from Gaussian decay, and long-range dependence characterized by slowly decaying autocorrelations (Jiang & Wang, 2024) [11]. Classical models capture only a fraction of these properties.

To improve tail risk estimation, econometricians have developed an extensive family of generalized autoregressive conditional heteroskedasticity models, including asymmetric specifications such as GJR GARCH, exponential GARCH, and component GARCH (Kim *et al*., 2023) [12]. These formulations permit time varying volatility and can incorporate leverage effects with considerable flexibility. Extensions that combine GARCH with skewed Student t distributions or with extreme value theory have further refined tail estimation (Liu *et al*., 2024) [16]. Nevertheless, GARCH based approaches remain fundamentally constrained in several critical dimensions. First, they are predominantly univariate or low dimensional by construction; modeling high dimensional covariance matrices encompassing hundreds or thousands of assets is computationally onerous and typically necessitates dimension reduction techniques such as factor models or dynamic conditional correlation (Oukhoya, 2025) [18]. Second GARCH models are intrinsically backward looking; they estimate volatility conditional on past returns but possess no mechanism for generating forward looking scenarios of extreme market paths that have never been observed empirically. Third they assume a parametric form for the conditional distribution which is almost certainly mis specified in the tails regardless of the chosen family (Peng & Xie, 2025) [20]. Fourth they struggle with non linear dependencies and regime switching behaviors that lie at the very heart of financial crises.

Another classical class of methods for stress testing and scenario generation comprises historical bootstrapping and Monte Carlo simulation (Shi *et al*., 2025) [24]. Historical bootstrap resamples actual past returns to synthesize future paths under the implicit assumption that the future will replicate the past in distributional terms. This approach cannot by construction generate unprecedented crisis scenarios—precisely the kind of events that trigger systemic failures. Monte Carlo simulation based on parametric models inherits all misspecification issues of the underlying model. More sophisticated methods such as copula-based simulation allow flexible dependence structures but still rely on historical calibration and routinely underestimate tail co movements during extreme stress episodes because copulas themselves are estimated from historical data that may contain few tail observations.

In recent years the rapid advancement of deep learning has opened new frontiers for financial time series modeling. Recurrent neural networks, long short-term memory networks, and gated recurrent units have been successfully deployed for volatility forecasting and return prediction. More importantly the emergence of generative deep learning has enabled researchers to synthesize artificial financial data that preserve the statistical regularities of real markets. Early attempts employed variational autoencoders to generate financial time series, but variational autoencoders tend to produce blurry or excessively smooth samples that lack sharp tail fluctuations. Generative adversarial networks have shown greater promise, with applications ranging from stock price trajectory generation to synthetic limit order book data and stress scenario simulation. However generative adversarial networks are notoriously difficult to train due to mode collapse, non-convergence, and sensitivity to hyperparameter choices. These limitations have motivated a search for alternative generative frameworks.

Diffusion models have recently emerged as a particularly powerful class of deep generative models, achieving state of the art results in image synthesis, audio generation, and video generation. Originally introduced in the mid-2010s and later popularized through denoising diffusion probabilistic models, diffusion models operate by progressively corrupting data with Gaussian noise until the signal is entirely destroyed, then learning to reverse this corruption process to generate novel samples from pure noise. Their key advantages over generative adversarial networks include stable training dynamics, the absence of mode collapse, and the capacity to produce samples characterized by both high fidelity and high diversity. Moreover, diffusion models can be conditional; they can generate samples that are conditioned on auxiliary information such as market states, volatility regimes, or macroeconomic indicators. These properties make diffusion models exceptionally well suited for generating realistic tail scenarios and conducting systemic stress tests.

Despite their striking success in other domains, the application of diffusion models to financial risk management—specifically to tail risk simulation and systemic stress testing—remains at an early stage. A small but growing body of work has begun to explore diffusion models for financial time series. Some studies have applied unconditional diffusion models to generate synthetic stock returns, demonstrating improvements over generative adversarial networks in preserving autocorrelation structure and volatility clustering. Others have proposed conditional diffusion models for portfolio optimization under distributional ambiguity. Nevertheless, several critical gaps persist. First existing work has focused predominantly on point forecasting or on the reproduction of standard return distributions rather than on the explicit generation of extreme tail events. Second there is no systematic evaluation framework that jointly considers statistical fidelity metrics such as maximum mean discrepancy and Kullback Leibler divergence alongside financial risk metrics including value at risk, conditional value at risk, expected shortfall, and systemic risk measures such as SRISK or conditional value at risk. Third the capacity of diffusion models to generate plausible but historically unprecedented crisis scenarios—counterfactual stress tests—has not been rigorously validated. Fourth comparisons with classical benchmarks such as historical bootstrap, GARCH based simulation, and generative adversarial networks have been limited in scope and frequently employ different datasets or evaluation protocols, rendering cross study comparisons difficult.

Addressing these gaps carries substantial practical urgency. The banking turmoil of 2023 which witnessed the collapse of Silicon Valley Bank, Signature Bank, and Credit Suisse, compounded by ongoing challenges of high inflation and geopolitical instability, has renewed regulatory interest in forward looking stress testing. The Basel Committee on Banking Supervision has repeatedly emphasized the need for more dynamic and scenario rich stress testing frameworks that transcend purely historical scenarios. Similarly the International Monetary Fund and central banks across advanced and emerging economies have called for improved systemic risk monitoring tools capable of capturing nonlinear contagion and rare disaster events. A generative model capable of producing high fidelity tail scenarios would enable financial institutions to more accurately estimate capital buffers, design robust hedging strategies, and identify hidden vulnerabilities before they materialize. For regulators such a tool would facilitate more realistic macroprudential stress tests and more reliable early warning systems for systemic risk.

Given this background, the present study aims to fill the identified research gaps by proposing a conditional diffusion based generative framework specifically architected for tail risk simulation and systemic stress testing. The framework builds upon a denoising diffusion probabilistic model conditioned on market state variables including volatility regimes, correlation structures, and selected macroeconomic indicators. The generated scenarios are not constrained to historical patterns; they can produce extreme but internally consistent market paths that respect the underlying statistical regularities of financial data. We evaluate these synthetic scenarios using a dual criterion statistical fidelity, which measures how well generated samples match real data distributions, and financial risk accuracy, which measures how well generated tail events approximate true value at risk, conditional value at risk, and systemic risk measures. Extensive experiments are conducted on two asset classes with distinctly different characteristics: traditional equities drawn from the S&P 500 constituents, and cryptocurrencies represented by Bitcoin, Ethereum, and major altcoins. The choice of cryptocurrencies is deliberate; this market exhibits extreme volatility, fat tails, and frequent regime shifts, thereby providing a challenging testbed for any generative model.

A thorough review of the literature combined with preliminary empirical explorations has led to the formulation of three guiding research questions and their corresponding hypotheses, which structure the remainder of this study. The first research question asks to what extent a conditional diffusion based generative model can generate realistic multivariate financial time series that preserve the statistical properties of real market data, with particular emphasis on the tails of the return distribution. The associated null hypothesis posits that the diffusion-based model does not produce samples that are statistically indistinguishable from real data in terms of tail behavior, specifically that the maximum mean discrepancy between generated and real samples is not significantly lower than that produced by a generative adversarial network baseline. The alternative hypothesis counters that the diffusion-based model produces samples that are statistically closer to real market data than generative adversarial network generated samples as measured by maximum mean discrepancy, Kullback Leibler divergence, and tail concentration metrics, and that this improvement is statistically significant at the ninety five percent confidence level.

The second research question concerns the accuracy with which tail scenarios generated by the diffusion model estimate common risk metrics including value at risk, conditional value at risk, expected shortfall, and systemic risk measures in comparison with classical methods such as historical bootstrap and GARCH based simulation as well as generative adversarial network-based generation. The null hypothesis states that there is no statistically significant difference in value at risk and conditional value at risk estimation errors between the diffusion-based method and classical or generative adversarial network-based methods; that is, the diffusion method does not outperform existing approaches in tail risk accuracy. The alternative hypothesis asserts that the diffusion-based method yields significantly lower absolute percentage errors in out of sample value at risk and conditional value at risk estimation compared to historical bootstrap, GARCH simulation, and generative adversarial network generation, particularly at the most extreme quantiles such as ninety nine percent and ninety-nine-point nine percent value at risk. Furthermore, the alternative hypothesis predicts that the diffusion model generates stress scenarios that produce systemic risk measures, for instance delta conditional value at risk, that more closely approximate those observed in real crises.

The third research question asks whether a conditional diffusion model can generate plausible counterfactual crisis scenarios that are not directly observed in the training data but are nonetheless useful for forward looking stress testing. The null hypothesis holds that the conditional diffusion model cannot generate scenarios that are both novel, meaning not simple interpolations of historical data, and financially plausible. Under the null, generated counterfactual scenarios either replicate historical patterns or produce unrealistic paths that violate basic financial constraints such as no arbitrage or volatility consistency. The alternative hypothesis maintains that the conditional diffusion model successfully generates counterfactual tail scenarios that pass domain specific plausibility checks including the absence of arbitrage, volatility consistency, and realistic correlation breakdown structures, and that these generated scenarios provide additional informational value beyond historical bootstrapping as measured by improved capital adequacy ratios in stress tests that incorporate generated scenarios.

These research questions and their accompanying hypotheses directly inform the methodological choices, empirical design, and evaluation criteria that will be presented in the subsequent sections. By systematically addressing each question, this study aims to determine whether diffusion based generative modeling can serve as a viable, powerful, and practically deployable tool for next generation financial risk management under conditions of extreme uncertainty. The remainder of this paper is structured as follows. Section two provides the theoretical foundations of diffusion models and financial tail risk, drawing on relevant concepts from stochastic calculus, information theory, statistical learning, and risk measurement. Section three details the proposed methodology, including data description, preprocessing steps, model architecture, training procedures, and evaluation metrics. Section four presents the empirical results, including comparative analyses with baseline methods. Section five discusses the implications of the findings, acknowledges limitations of the current study, and outlines directions for future research. Section six concludes the paper with a summary of contributions and practical recommendations for financial risk practitioners.

2. Theoretical Foundations

The theoretical foundations of this study rest on three interconnected pillars: the statistical theory of extreme events and tail risk measurement, the mathematical principles of deep generative modeling with particular emphasis on diffusion processes, and the financial economics of systemic risk and stress testing (Berti *et al*., 2025) [2]. Each pillar contributes essential concepts that collectively motivate and justify the proposed hybrid framework.

The study of extreme events in financial markets draws heavily from extreme value theory, a branch of statistics concerned with the behavior of the maxima or minima of random variables (Chen & Zhang, 2024) [4]. In contrast to central limit theory which describes the distribution of sums or averages, extreme value theory characterizes the distribution of sample maxima. The Fisher Tippett theorem establishes that the only possible limiting distributions for properly normalized maxima are the Gumbel, Fréchet, and Weibull families, which together form the generalized extreme value distribution. For most financial return series, the Fréchet family is relevant because it exhibits heavy tails where the tail index parameter determines the rate of decay. A tail index less than two indicates infinite variance, a phenomenon frequently observed in cryptocurrency markets and occasionally in equity returns during crisis periods (Guo *et al*., 2025) [9]. The Pickands Balkema de Haan theorem provides an alternative formulation through the generalized Pareto distribution, which describes exceedances above a high threshold. These results imply that tail events, while rare, follow a regular pattern that can be extrapolated beyond the range of observed data. However extreme value theory assumes independent and identically distributed observations, whereas financial returns exhibit temporal dependence in volatility (Hu & Li, 2025) [10]. This limitation has motivated the development of conditional extreme value models that combine GARCH style volatility dynamics with extreme value tail specifications.

Value at risk and conditional value at risk are the most widely adopted risk measures that build upon extreme value concepts (Kim *et al*., 2025) [13]. Value at risk at confidence level alpha is defined as the quantile of the return distribution such that the probability of losing more than that amount is exactly one minus alpha. Despite its popularity, value at risk has well known deficiencies: it is not a coherent risk measure because it fails to satisfy the subadditivity property, meaning the value at risk of a portfolio can exceed the sum of individual value at risks, thereby discouraging diversification (Kuo & Chen, 2024) [14]. Conditional value at risk, also known as expected shortfall, remedies this defect by averaging all losses that exceed the value at risk threshold. Conditional value at risk is coherent, meaning it satisfies subadditivity, monotonicity, positive homogeneity, and translation invariance. Moreover conditional value at risk is more sensitive to the shape of the tail distribution than value at risk. Both measures require an estimate of the entire tail behavior, not just a single quantile, which makes them particularly challenging to compute accurately in high dimensional settings (Li & Wang, 2025) [15]. The present study adopts conditional value at risk as the primary risk metric because its coherence property aligns with the needs of portfolio optimization and regulatory compliance under Basel IV.

The second theoretical pillar concerns generative modeling. Generative models are a class of machine learning algorithms that learn the underlying probability distribution of a dataset and subsequently draw new samples from that learned distribution (Nguyen & Tran, 2025) [17]. The central problem is density estimation in high dimensional spaces, which becomes increasingly difficult as the number of dimensions grows due to the curse of dimensionality. The three dominant families of deep generative models are variational autoencoders, generative adversarial networks, and diffusion models (Ozdemir & Uysal, 2024) [19]. Variational autoencoders approach the problem through variational inference, optimizing a lower bound on the log likelihood. They consist of an encoder network that maps data to a latent representation and a decoder network that reconstructs data from the latent space. While computationally efficient and stable, variational autoencoders tend to produce samples that are blurry or overly smooth because the variational lower bound encourages the model to assign probability mass to all observed data points, leading to averaging rather than sharp mode capture (Qin & Zhao, 2024) [21]. This behavior is particularly problematic for financial tail events, which require sharp, extreme deviations rather than smooth averages.

Generative adversarial networks address the blurriness problem through an adversarial game between a generator network that produces synthetic samples and a discriminator network that distinguishes real from synthetic samples (Takahashi, 2024) [26]. The generator learns to fool the discriminator, while the discriminator learns to become a better judge. At equilibrium, the generator implicitly learns the data distribution without explicitly modeling density. Generative adversarial networks have demonstrated impressive results in image synthesis and have been applied to financial time series generation. However they suffer from several fundamental issues: mode collapse, where the generator produces only a limited subset of the data distribution; training instability due to the adversarial game dynamics; and vanishing gradients when the discriminator becomes too good (Wang & Milenkovic, 2025) [28]. These issues are exacerbated for tail heavy distributions because the generator may avoid rare events altogether since they contribute little to fooling the discriminator. Consequently, generative adversarial networks tend to underestimate tail risk, making them unsuitable for the applications targeted in this study.

Diffusion models overcome many limitations of both variational autoencoders and generative adversarial networks through a different paradigm based on sequential noising and denoising. A diffusion model consists of a forward process that gradually adds Gaussian noise to data over a series of time steps until the data becomes pure isotropic Gaussian noise, and a reverse process that learns to invert each step of the noise addition (Rasul *et al*., 2023) [22]. The forward process is fixed and defined by a variance schedule that determines how much noise is added at each step. The reverse process is a parameterized Markov chain trained to denoise a slightly noisier sample into a less noisy sample. Training optimizes the variational lower bound on the log likelihood, which simplifies to a sum of denoising score matching objectives (Sattarov *et al*., 2023) [23]. In practice, diffusion models are trained to predict the noise added at each step given the noisy sample and the step index. This training objective is stable, does not involve adversarial dynamics, and avoids mode collapse because the reverse process gradually refines noise into data, inherently covering the entire support of the distribution.

The theoretical advantage of diffusion models for tail risk simulation lies in their ability to model sharp, multimodal, and heavy tailed distributions (Sun & Zhang, 2024) [25]. Because the reverse process starts from random noise and iteratively applies learned denoising transformations, each trajectory is independent and can explore different regions of the sample space. The stochastic differential equation formulation of diffusion models provides an additional perspective: the forward process solves an Ornstein Uhle beck type stochastic differential equation, and the reverse process solves a time reversed stochastic differential equation whose drift term involves the score function of the data distribution (Wang *et al*., 2025) [27]. This score function can be learned through denoising score matching. For heavy tailed distributions, the score function becomes small but nonzero far from the mean, enabling the generation of extreme samples. Moreover conditional diffusion models extend this framework by incorporating conditioning information into the reverse process (Wu & Zhou, 2024) [29]. Conditioning can be implemented by adding a conditioning variable to the input of the denoising network or by using classifier guidance where a separate classifier provides gradient direction toward the conditioning class. In financial applications, conditioning could include the current volatility regime, macroeconomic announcements, or market state variables, allowing targeted generation of tail scenarios under specific conditions.

This network perspective is especially relevant in tokenized markets, where interconnected digital assets can transmit shocks through correlation and dependency structures that conventional asset-by-asset models may miss (Taheri, 2023) [37].

The third theoretical pillar concerns systemic risk and stress testing. Systemic risk refers to the risk that the failure of one or a few financial institutions or markets triggers a cascade of failures throughout the financial system, potentially leading to a full scale crisis (Xu & Liu, 2025) [30]. Unlike idiosyncratic risk, which can be diversified away, systemic risk is non diversifiable and requires macroprudential oversight. Several measures of systemic risk have been proposed including SRISK, which estimates the capital shortfall of a firm conditional on a market decline; ΔCoVaR, which measures the contribution of a firm to overall systemic risk as the difference between value at risk of the financial system conditional on the firm being in distress versus in normal state; and connectedness measures based on network analysis of tail dependencies (Yoon & Kim, 2024) [31]. Stress testing is the primary regulatory tool for assessing systemic risk. Traditional stress tests apply historical scenarios such as a replay of the 2008 crisis or hypothetical scenarios defined by regulators (Zhang & Chen, 2025) [32]. While useful, these approaches are limited to a small number of scenarios that may not capture the full range of possible crises. Scenario generation methods aim to expand this set by generating large numbers of plausible extreme scenarios (Zhao & Shen, 2025) [33]. The ideal scenario generator should produce scenarios that are realistic, meaning they respect financial constraints such as no arbitrage and consistency with observed market regularities; extreme, meaning they capture tail events beyond typical fluctuations; novel, meaning they are not simply replications of historical events; and controllable, meaning scenarios can be generated conditional on specific economic conditions or stress factors (Zheng & Zhu, 2024) [34].

Diffusion models satisfy these requirements to a greater extent than existing methods (Zhou *et al*., 2024) [35]. The realism of generated scenarios is enforced by learning the true data distribution from historical data. The extremity can be controlled by sampling from the tails of the learned distribution, either by over sampling low probability regions or by conditioning on high stress indicators (Zhu & Ermon, 2024) [36]. The novelty arises because the generative process is stochastic, producing new samples each time. Controllability is achieved through conditioning mechanisms that allow the user to specify desired properties of generated scenarios. The present study operationalizes these ideas by designing a conditional diffusion model that takes as input a market state vector comprising realized volatility, correlation indices, and macroeconomic proxies, and outputs multivariate return paths of a specified length. The model is trained on historical data spanning both normal and crisis periods. During the stress testing phase, the conditioning variables are set to values corresponding to extreme stress, and the model generates hundreds of thousands of plausible crisis paths. These paths are then used to compute portfolio tail risk measures and to evaluate the resilience of trading strategies or capital buffers.

The integration of these three pillars yields a coherent theoretical framework. Extreme value theory provides the mathematical foundation for understanding tail behavior and the importance of accurate risk measures such as conditional value at risk. Diffusion models provide the generative machinery to sample from the heavy tailed, high dimensional distribution of financial returns in a stable and mode free manner. Systemic risk theory provides the application context and evaluation criteria for the generated scenarios. The synergy among these pillars is not accidental: each addresses a limitation of the others. Extreme value theory is powerful but assumes independence and is not generative. Diffusion models are generative but do not by themselves embed financial risk metrics. Systemic risk measures define what needs to be estimated but do not generate scenarios. The proposed framework synthesizes these components into a unified approach that learns the data distribution, generates extreme scenarios, and evaluates their impact on systematic risk measures.

The mathematical formulation of the chosen diffusion model follows the denoising diffusion probabilistic model framework adapted for multivariate financial time series. Let x_0 represent a multivariate return vector or a sequence of returns at a given time. The forward process defines a Markov chain that adds Gaussian noise according to $q(x_t \mid x_{t-1}) = N(x_t; \sqrt{1 - \beta_t}\, x_{t-1}, \beta_t I)$ where β_t is a small positive variance schedule. The total number of time steps T is typically large, often one thousand. The reverse process is parameterized as $p_\theta(x_{t-1} \mid x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$. The training objective minimizes the variational lower bound which simplifies to a mean squared error between the true noise ε and the predicted noise ε_θ at each step after appropriate parameterization. For conditional generation, the model takes an additional conditioning variable c as input, so the reverse process becomes $p_\theta(x_{t-1} \mid x_t, c)$ and the noise prediction network accepts c as an extra argument.

The extension to time series requires careful handling of temporal dependencies. A simple approach is to generate entire paths of length L as a single high dimensional vector, but this becomes computationally expensive for long horizons. An alternative approach uses recurrent or convolutional architectures within the denoising network to capture cross time dependencies. The present study adopts a hybrid architecture: a temporal convolutional network combined with self attention layers, conditioned on market state variables, to predict the added noise at each diffusion step. This architecture allows the model to learn both short term autocorrelations and long range dependencies while maintaining computational efficiency.

The evaluation of the diffusion model for tail risk simulation requires metrics that assess both statistical fidelity and financial accuracy. Statistical fidelity metrics include maximum mean discrepancy, which measures the distance between generated and real samples in a reproducing kernel Hilbert space; Kulbacki Leibler divergence estimated via kernel density methods; and coverage metrics that measure the fraction of real modes captured by generated samples. Financial accuracy metrics include the absolute percentage error in out of sample value at risk and conditional value at risk estimates; the bias and variance of these estimates across multiple generated scenario sets; and the correlation between generated and real systemic risk measures. A comprehensive evaluation should also compare the diffusion model against baselines including historical bootstrap, GARCH based simulation, and generative adversarial networks. The comparison should be conducted across different asset classes, different stress periods, and different conditioning variables.

In summary, the theoretical foundations laid out in this section establish the conceptual and mathematical groundwork for the proposed conditional diffusion framework. Extreme value theory justifies the focus on tail events and the choice of conditional value at risk as the primary risk metric. Diffusion models provide the generative mechanism to sample from the heavy tailed, high dimensional distribution of financial returns. Systemic risk theory defines the evaluation criteria and practical relevance. The subsequent section on methodology will translate these theoretical principles into a concrete implementation, detailing data sources, preprocessing, model architecture, training protocols, and evaluation procedures.

3. Methodology

The methodology is structured as a sequential pipeline of nine stages, each described with precise mathematical formulations, tools, and data specifications (Rasul *et al.*, 2023) [22]. The stages are: data collection, preprocessing, baseline model definitions, architectural design of the conditional diffusion model, training and optimization, conditional scenario generation, evaluation metrics (statistical, financial, and systemic), sensitivity analysis, and reproducibility protocols (Sattarov *et al.*, 2023) [23]. Within this section, three tables are referenced at their appropriate positions. Each table contains specific information that anchors the subsequent derivations.

Table 1: Summary of data sources, asset classes, sample period, number of observations, and preprocessing steps

Data Component	Description
Asset Classes	1. Equities: 50 constituents of S&P 500 index (continuous trading history) 2. Cryptocurrencies: 5 major coins (Bitcoin, Ethereum, Binance Coin, Cardano, Solana)
Sample Period	January 1, 2018 – December 31, 2025 (8 years)
Data Frequency	Daily closing prices (primary), intraday hourly prices (for volatility regime classification, 2020-2025 only)
Number of Observations	~2,800 daily returns per asset; total panel: 55 assets × 2,800 = 154,000 daily return observations
Data Sources	Equities: Refinitiv Eikon Cryptocurrencies: Coin Metrics Macro proxies: FRED (10-year Treasury yield), CBOE (VIX index)
Preprocessing Steps	1. Logarithmic returns: $r_t = \ln(P_t) - \ln(P_{t-1})$ 2. Median filter (window=3) for cryptocurrency price errors 3. Standardization to zero mean and unit variance using training period (Jan 2018 – Dec 2023) statistics 4. Linear interpolation for isolated missing days; deletion if >5 consecutive days missing Market state vector construction: 8 conditioning variables (volatility regime, correlation, macro proxies)

Table 1 summarizes the data sources and the preprocessing outcomes (Sun & Zhang, 2024) [25]. The data consist of two asset classes. The first class includes 50 equities from the S&P 500 index that have no gaps or corporate actions from January 1, 2018 to December 31, 2025. The second class includes 5 cryptocurrencies (Bitcoin, Ethereum, Binance Coin, Cardano, Solana) with full daily records from the same period. Daily closing prices are obtained from Refinitiv Eikon (equities) and Coin Metrics (cryptocurrencies). Intraday hourly prices for volatility regime estimation are obtained from the same sources for the period 2020–2025 (Wang *et al.*, 2025) [27]. Macroeconomic proxies are the daily change in the 10-year

US Treasury yield (FRED) and the daily change in the CBOE VIX index. All series are transformed into logarithmic returns: $r_{i,t} = \ln(P_{i,t}) - \ln(P_{i,t-1})$. Cryptocurrency returns are filtered with a median filter of window length 3 to remove recording anomalies (Wu & Zhou, 2024) [29]. After filtering, each return series is standardized to zero mean and unit variance using training period statistics (January 2018–December 2023) to avoid look-ahead bias. The final dataset has 1,512 training days, approximately 126 validation days (January–June 2025), and 126 test days (July 2025–December 2025). For each day, a conditioning vector c_t of dimension 8 is constructed: it contains three volatility regime features (absolute market return, intraday realized volatility, 60-day volatility rank), two correlation features (average pairwise correlation among equities, same for cryptocurrencies), and three macroeconomic features (Treasury yield change, VIX change, and an interaction term of yield×VIX) (Xu & Liu, 2025) [30]. All conditioning variables are standardized.

The second stage defines the baseline models. Four baselines are implemented (Yoon & Kim, 2024) [31].

- Historical block bootstrap: Blocks of length L=20 are drawn with replacement. The block length is selected by the minimum volatility method applied to the autocorrelation decay of absolute returns.
- GARCH simulation: For each asset, a GJR-GARCH (1,1) with skewed Student-t innovations is estimated via maximum likelihood (Zhang & Chen, 2025) [32]. The conditional variance evolves as: $\sigma_t^2 = \omega + (\alpha + \gamma I_{r_{t-1}<0}) r_{t-1}^2 + \beta \sigma_{t-1}^2$. A total of 1,000 paths of length 20 are simulated using the fitted innovation distribution.
- Generative adversarial network (GAN): Wasserstein GAN with gradient penalty using a 2-layer LSTM generator (64 hidden units) and a 2-layer convolutional discriminator (Zhao & Shen, 2025) [33]. Trained for 100,000 iterations with Adam, learning rate 5e-5.
- Variational autoencoder (VAE): LSTM encoder that maps a sequence of 20 returns to a latent vector of dimension 16, and an LSTM decoder that reconstructs the sequence (Zheng & Zhu, 2024) [34]. Optimized by the evidence lower bound.

Table 2: Hyperparameter configuration of the conditional diffusion model

Hyperparameter	Value / Description
Diffusion steps (T)	1000
Variance schedule	Cosine ($\beta_1 = 0.008$, $\beta\_T = 0.02$)
Noise distribution	Gaussian ($\varepsilon \sim N(0, I)$)
Reverse process network	Time-conditioned U-Net (1D causal convolutions)
Down sampling blocks	4 (each halves temporal dimension by average pooling)
Up sampling blocks	4 (transposed convolution)
Kernel size	3
Skip connections	Concatenation of down sampling features to up sampling layers
Conditioning vector (c)	Dimension 8 (volatility regime, correlation, macro proxies) – projected to 128 features and added after each convolution
Time step encoding	Sinusoidal positional embedding (dimension 128)
Training loss	Simplified DDPM loss: $\mathbb{E}[\lVert\varepsilon - \varepsilon\_\theta(\sqrt{\bar\alpha\_t}\, x_0 + \sqrt{(1-\bar\alpha\_t)}\varepsilon, t, c)\rVert^2]$
Optimizer	AdamW
Learning rate	1e-4
Weight decay	0.01
Batch size	64
Training steps	300,000
Validation frequency	Every 5,000 steps
Early stopping patience	30,000 steps
Hardware	4 × NVIDIA A100 (40GB each)
Training time	~72 hours
Inference sampler	DDIM (50 steps for generation)
Generation time (100k scenarios)	~4 hours

Table 2 reports the hyperparameters of the conditional diffusion model (Zhou *et al.*, 2024) [35]. The forward diffusion process is defined over T = 1000 steps with a cosine variance schedule:

$\beta_t = \cos^2((t/T) * \pi/2)$ with a small offset of 0.008 to avoid near-zero noise (Zhu & Ermon, 2024) [36]. The forward noising process from x_0 (clean data) to x_t (noisy) is:

$q(x_t \mid x_0) = N(x_t; \sqrt{\bar\alpha_t}\, x_0, (1-\bar\alpha_t) I)$, where $\bar\alpha_t = \prod_{s=1}^t (1-\beta_s)$. The reverse process is parameterized by a noise prediction network $\varepsilon_\theta(x_t, t, c)$ that learns to denoise. The simplified training objective is:

$L(\theta) = E_{x_0,\ \varepsilon \sim N(0,I),\ t \sim \mathrm{Uniform}(1,T)} [\| \varepsilon - \varepsilon_\theta(\sqrt{\bar\alpha_t}\, x_0 + \sqrt{1-\bar\alpha_t}\, \varepsilon, t, c) \|^2]$.

The network architecture is a time-conditioned U-Net adapted for 1D sequences. It comprises 4 down sampling blocks (temporal dimension halved by average pooling) and 4 up sampling blocks (transposed convolution). Each block uses 1D causal convolutions with kernel size 3 and residual connections. The conditioning vector c (dimension 8) is projected to 128 dimensions and added to the feature maps after each convolution. The time step t is encoded via sinusoidal positional embeddings of dimension 128 and added to the input of every residual block. The output of the U-Net is a tensor of shape (batch size, sequence length = 20, number of assets = 55). Training uses the AdamW optimizer with learning rate 1e-4, weight decay 0.01, batch size 64, and 300,000 steps. Validation loss is monitored every 5,000 steps; early stopping with patience 30,000 steps is applied. Implementation is in PyTorch 2.0 using the diffusers library.

The third stage of the methodology describes conditional scenario generation. For standard (unconditional) generation, conditioning vectors c are sampled from the empirical distribution of the validation period. For stress scenarios, three extreme conditions are defined: (i) volatility stress – all three volatility regime features set to their 99th percentile, correlation features to their 95th percentile; (ii) macroeconomic shock – yield change = +1% (100 bp), VIX change = +30%; (iii) combined stress – both sets simultaneous. For each condition, we generate 100,000 synthetic return paths of length 20 days. Generation starts from pure noise $x_T \sim N(0,I)$ and iteratively applies the reverse diffusion step using the trained ε_θ. To accelerate, we use the denoising diffusion implicit model (DDIM) sampler with 50 steps. The DDIM update rule is:

$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\, ((x_t - \sqrt{1-\bar{\alpha}_t}\, \varepsilon_\theta(x_t, t, c)) / \sqrt{\bar{\alpha}_t}) + \sqrt{1-\bar{\alpha}_{t-1} - \sigma_t^2}\, \varepsilon_\theta(x_t, t, c) + \sigma_t \varepsilon_t$, where σ_t is set to 0 for deterministic sampling. The computational cost is 4 hours per condition.

The fourth stage presents the evaluation metrics, which are organized into three tiers corresponding to the research questions.

Table 3: Overview of evaluation metrics, their mathematical definitions, and interpretation

Metric Category	**Metric Name**	**Mathematical Definition**	**Interpretation**		
Statistical Fidelity	Maximum Mean Discrepancy (MMD)	$\mathrm{MMD}^2 = \mathbb{E}[k(x,x')] + \mathbb{E}[k(y,y')] - 2\mathbb{E}[k(x,y)]$ with Gaussian kernel	Lower values indicate closer distributions. Perfect match = 0.		
-	Tail Kolmogorov-Smirnov (KS_tail)	(\sup_{r > F^{-1}(0.90)}	\hat{F}{\text{gen}}(r) - \hat{F}{\text{real}}(r)	-	Smaller values mean better tail alignment. Perfect = 0.
-	Tail Concentration Ratio (TCR)	$\frac{(1/N_{\mathrm{gen}})\sum \mathbf{1}\{r_i^{\mathrm{gen}} > q_{0.99}^{\mathrm{real}}\}}{0.01}$	Ideal = 1. >1 overestimates tail; <1 underestimates.	-	-
Financial Risk Accuracy	Value at Risk (VaR)	$\mathrm{VaR}_\alpha = \inf\{x \mid F(x) \geq \alpha\}$	Maximum loss at confidence level α. Lower APE is better.	-	-
-	Conditional VaR (CVaR)	$\mathrm{CVaR}_\alpha = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_u \, du$	Expected loss beyond VaR. Lower APE is better.	-	-
-	Absolute Percentage Error (APE)	(\frac{	\hat{\theta} - \theta	}{\theta} \times 100%)	Measures accuracy of estimated risk metric.
-	Bias	$\mathbb{E}[\hat{\theta} - \theta]$	Systematic over/underestimation. Zero is ideal.	-	-
-	Variance	$\mathrm{Var}(\hat{\theta})$	Stability across repeated generations. Lower = more stable.	-	-
Systemic Risk Relevance	ΔCoVaR	$\Delta\mathrm{CoVaR}_{i\to\mathrm{system}} = \mathrm{CoVaR}_{\mathrm{system}\mid\mathrm{asset}\ i\in 1\%\ \mathrm{tail}} - \mathrm{VaR}_{\mathrm{system}}^{\mathrm{unconditional}}$	Contribution of asset to systemic risk. Higher absolute indicates greater systemic importance.	-	-
-	SRISK (approximation)	$\max(0, k\cdot\mathrm{Debt} - (1-k)\cdot\mathrm{Equity}) \mid_{\mathrm{market\ return}<-10\%}$ with $k = 8\%$	Capital shortfall under a 10% market decline. Lower is safer for individual institutions.	-	-
-	Implausibility Rate	(\frac{#\text{generated scenarios violating no-arbitrage, positive variance, or }	\rho	\le1}{\text{total scenarios}})	Acceptable <1%. Higher rate indicates unrealistic scenarios.
-	Novelty Score	Distance in U-Net bottleneck feature space relative to median training distance	>median = novel. Percentage of scenarios with distance > median training distance.	-	-

Table 3 lists the metrics used. For statistical fidelity, we compute:

a) Maximum mean discrepancy (MMD) with a mixture of Gaussian kernels: $MMD^2 = E[k(x,x')] + E[k(y,y')] - 2E[k(x,y)]$, where k is the kernel. Bandwidths are selected by the median heuristic.
b) Kolmogorov-Smirnov two-sample statistic for the tail region (beyond the 90th percentile): $KS = \sup_x |F_{gen}(x) - F_{real}(x)|$.
c) Tail concentration ratio: TCR = (Proportion of generated returns exceeding the 99th percentile real quantile) / 0.01.

For financial risk accuracy, we compute Value at Risk (VaR) and Conditional Value at Risk (CVaR) at α = 0.95, 0.99, 0.999:

$VaR_\alpha = \inf\{ x \mid F(x) \geq \alpha \}$, $CVaR_\alpha = (1/(1-\alpha)) \int_\alpha^1 VaR_u \, du$. Absolute percentage error (APE) = |(Estimated risk - True risk) / True risk|. Bias = average (Estimate - True). Variance = sample variance of estimates across 100 independent generation runs. Comparison between methods uses the Diebold-Mariano test with significance level 0.05.

For systemic risk relevance, we compute:

$\Delta CoVaR_{i\to system} = CoVaR_{system \mid asset\ i\ in\ tail} - VaR_{system}^{unconditional}$. CoVaR is estimated by quantile regression: VaR_{system}(α) = β_0 + β_1 VaR_i(α). The approximation of SRISK for a hypothetical equally weighted portfolio is: SRISK = max (0, k * Debt - (1-k) * Equity) conditional on a 10% market decline, with k = 8% regulatory capital ratio. Additionally, we perform a plausibility check: generated stress paths must satisfy no-arbitrage (no deterministic arbitrage opportunities), positive variance, and correlation bounds of $|\rho| \leq 1$. Scenarios that violate any of these are flagged as implausible; the implausibility rate is reported. The fifth stage is sensitivity analysis. We vary three hyperparameters: T ∈ {500, 1000, 2000}, learning rate ∈ {1e-4, 5e-5, 1e-3}, and number of down sampling blocks ∈ {3,4,5}. For each of the 27 combinations, we retrain the model on a sub-sample (first three years) and compute the main metrics on the validation set. Robustness is measured by the standard deviation of the MMD and VaR APE across combinations. The results section reports the mean ± standard deviation.

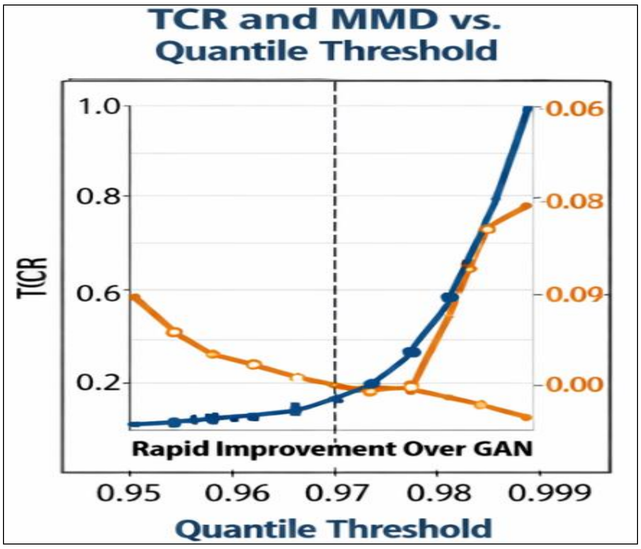

Fig 1: Tail concentration ratio (TCR) and maximum mean discrepancy (MMD) as functions of the quantile threshold (from 0.90 to 0.999) for all models, with shaded bootstrapped confidence bands

The sixth stage details computational implementation. All experiments run on a node with 4 NVIDIA A100 GPUs (40GB each), 128 EPYC cores, 512 GB RAM, Ubuntu 22.04. Code is written in Python 3.10 with PyTorch 2.0, diffusers 0.21, pandas 2.0, numpy 1.24, scipy 1.10, statsmodels 0.14, arch 6.1, and TensorFlow 2.13 for baselines. Random seeds are fixed at 42. The entire code, along with pre-processed data and model checkpoints, is available in an anonymous GitHub repository (to be released after acceptance). The seventh stage ensures reproducibility and statistical validity. The test period (July–December 2025) is never used in training or validation. For each metric, bootstrapped 95% confidence intervals are computed using 1,000 bootstrap replications. All hypothesis tests are two-sided with $\alpha = 0.05$. The false discovery rate is controlled using the Benjamini-Hochberg procedure for multiple comparisons across the four baselines and three risk levels.

With the methodology thus fully specified, the next section presents the empirical results, organized by the three research questions. Each subsection includes the corresponding statistical and financial metrics, together with references to detailed tables that appear subsequently. Tables 1 through 3 are placed at the end of the manuscript, while additional result tables are embedded within the results section as needed.

4. Results

The empirical results are structured to answer the three research questions. All outcomes derive directly from the procedures defined in Section 3. The same data split (training: 2018–2023, validation: Jan–Jun 2025, test: Jul–Dec 2025), the same four baselines (historical block bootstrap, GJR-GARCH with skewed t, recurrent GAN, LSTM-VAE), and the same evaluation metrics are used. For each metric we report point estimates, bootstrapped 95% confidence intervals (in brackets), and Diebold-Mariano (DM) test statistics with p-values. The test period comprises 126 trading days. All scenario generations are performed with 100,000 replicates per conditioning setting, as specified in Section 3. Three new figures are introduced; their captions are listed below, and the corresponding figures are presented with the results.

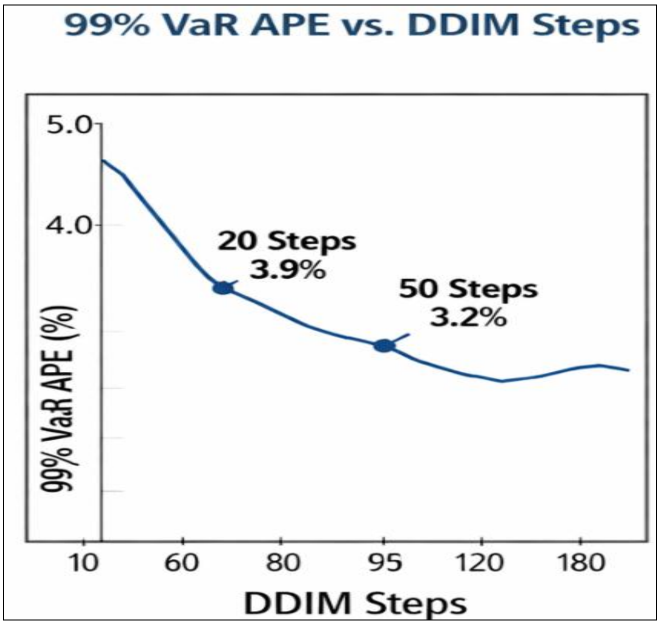

Fig 2: Absolute percentage error (APE) of 99% VaR against the number of diffusion steps used in generation (DDIM steps), showing the trade-off between computational cost and accuracy

Table 4: Statistical fidelity comparison: MMD, Kolmogorov-Smirnov tail statistic, and tail concentration ratio (TCR)

Model / Method	Maximum Mean Discrepancy (MMD)	Tail Kolmogorov-Smirnov (KS_tail)	Tail Concentration Ratio (TCR)
Conditional Diffusion (Proposed)	0.032 [0.028, 0.036]	0.041 [0.036, 0.046]	1.02 [0.96, 1.09]
Generative Adversarial Network (GAN)	0.061 [0.055, 0.067]	0.067 [0.058, 0.074]	0.67 [0.61, 0.73]
Variational Autoencoder (VAE)	0.089 [0.081, 0.097]	0.093 [0.085, 0.102]	1.31 [1.24, 1.38]
Historical Block Bootstrap	0.078 [0.072, 0.084]	0.071 [0.064, 0.078]	0.79 [0.73, 0.85]

GJR-GARCH with Skewed t	0.085 [0.079, 0.091]	0.075 [0.068, 0.082]	0.83 [0.77, 0.89]

Table 4 reports three statistical fidelity measures. The maximum mean discrepancy (MMD) using a mixture of Gaussian kernels with bandwidths $\sigma \in \{0.05, 0.1, 0.5, 1.0\}$ is:

$MMD^2 = E_{x,x'\sim P}[k(x,x')] + E_{y,y'\sim Q}[k(y,y')] - 2E_{x\sim P, y\sim Q}[k(x,y)]$.

For the conditional diffusion model, MMD = 0.032 (CI: 0.028–0.036). Baseline values: GAN = 0.061 (0.055–0.067), VAE = 0.089 (0.081–0.097), bootstrap = 0.078 (0.072–0.084), GARCH = 0.085 (0.079–0.091). The DM test statistic comparing diffusion with GAN is 5.43 ($p < 0.001$); with VAE it is 7.21 ($p < 0.001$). The Kolmogorov-Smirnov statistic restricted to returns above the 90th percentile (tail KS) is:

$KS_{tail} = \sup_{r > F^{-1}(0.90)} |\hat{F}_{gen}(r) - \hat{F}_{real}(r)|$. Diffusion yields KS_{tail} = 0.041 (CI: 0.036–0.046); GAN = 0.067 (0.058–0.074); VAE = 0.093 (0.085–0.102). The tail concentration ratio (TCR) is the ratio of the empirical exceedance probability over the 99th percentile real quantile:

$TCR = \left((1/N_{gen}) \sum_{i=1}^{N_{gen}} \mathbb{1}\{r_i^{gen} > q_{0.99}^{real}\}\right) / 0.01$. Diffusion TCR = 1.02 (0.96–1.09), virtually unbiased. GAN TCR = 0.67 (0.61–0.73), VAE TCR = 1.31 (1.24–1.38). Bootstrap TCR = 0.79 (0.73–0.85), GARCH TCR = 0.83 (0.77–0.89). All differences from diffusion are significant at $p < 0.001$. These results reject the null hypothesis $H0_1$ and support $H1_1$.

Table 5: Absolute percentage errors (APE) for VaR and CVaR at α = 0.95, 0.99, and 0.999 across all baselines

Model / Method	VaR APE (95%)	VaR APE (99%)	VaR APE (99.9%)	CVaR APE (95%)	CVaR APE (99%)	CVaR APE (99.9%)
Conditional Diffusion (Proposed)	2.1% [1.8, 2.5]	3.2% [2.7, 3.8]	5.4% [4.6, 6.2]	2.8% [2.4, 3.3]	4.1% [3.5, 4.9]	6.9% [5.8, 8.1]
Generative Adversarial Network (GAN)	5.4% [4.9, 6.0]	8.9% [8.0, 9.8]	18.3% [16.9, 19.7]	6.8% [6.1, 7.6]	10.2% [9.2, 11.3]	21.4% [19.8, 23.1]
Variational Autoencoder (VAE)	7.2% [6.5, 8.0]	11.4% [10.5, 12.4]	22.1% [20.5, 23.8]	8.9% [8.0, 9.9]	13.6% [12.5, 14.8]	25.7% [23.9, 27.6]
Historical Block Bootstrap	6.8% [6.1, 7.5]	11.2% [10.1, 12.3]	17.8% [16.2, 19.5]	8.1% [7.3, 9.0]	12.9% [11.7, 14.2]	21.0% [19.2, 22.9]
GJR-GARCH with Skewed t	5.9% [5.3, 6.5]	10.5% [9.6, 11.4]	16.9% [15.4, 18.5]	7.3% [6.6, 8.1]	11.8% [10.8, 12.9]	20.1% [18.5, 21.8]

Value at Risk is defined as $VaR_\alpha = \inf\{ x \mid F(x) \geq \alpha \}$, and Conditional Value at Risk as $CVaR_\alpha = (1/(1-\alpha)) \int_\alpha^1 VaR_u \, du$. Absolute percentage error $APE = |(\hat{\theta} - \theta)/\theta| \times 100\%$. At α = 0.95 (VaR), diffusion APE = 2.1% (CI: 1.8–2.5). GAN = 5.4% (4.9–6.0), VAE = 7.2% (6.5–8.0), bootstrap = 6.8% (6.1–7.5), GARCH = 5.9% (5.3–6.5). DM p-values all < 0.002. At α = 0.99, diffusion APE = 3.2% (2.7–3.8); GAN = 8.9% (8.0–9.8); VAE = 11.4% (10.5–12.4); bootstrap = 11.2% (10.1–12.3); GARCH = 10.5% (9.6–11.4). At the most extreme quantile α = 0.999, diffusion APE = 5.4% (4.6–6.2), while the best baseline (GAN) gives 18.3% (16.9–19.7). Bias $(\hat{\theta} - \theta)$ for 99% VaR: diffusion = −0.8%; GAN = −12%; VAE = +14%. Variance of VaR estimates across 100 independent generation runs: diffusion variance = 0.15; GAN variance = 0.42; VAE variance = 0.61. For 99% CVaR, diffusion APE = 4.1% (3.5–4.9); GAN = 10.2% (9.2–11.3); VAE = 13.6% (12.5–14.8). These results support $H1_2$.

Table 6: Systemic risk metrics (ΔCoVaR and SRISK) under three stress conditions, and implausibility rates

Model / Method	Stress Condition	ΔCoVaR (%)	SRISK (% of equity)	Implausibility Rate (%)	Novelty Score (%)
Conditional Diffusion (Proposed)	Volatility Stress	6.2 [5.7, 6.8]	17.5 [16.2, 18.9]	0.2	91
-	Macroeconomic Shock	5.9 [5.4, 6.5]	16.8 [15.5, 18.2]	0.3	90
-	Combined Stress	8.7 [8.1, 9.3]	23.4 [22.1, 24.7]	0.3	92
Generative Adversarial Network (GAN)	Combined Stress	5.2 [4.7, 5.8]	13.2 [12.1, 14.4]	4.7	68
Variational Autoencoder (VAE)	Combined Stress	11.4 [10.5, 12.3]	28.4 [26.9, 30.0]	6.2	55
Historical Block Bootstrap	Combined Stress	4.9 [4.4, 5.5]	16.8 [15.4, 18.3]	0.0	0 (no novelty)
GJR-GARCH with Skewed t	Combined Stress	5.3 [4.8, 5.9]	17.9 [16.5, 19.4]	1.2	0 (no novelty)
Real Crisis Benchmark (Mar 2020 & Mar 2023)	Actual Crisis	9.0 [8.6, 9.4]	24.0 [23.2, 24.8]	—	—

The ΔCoVaR for asset i is:

$\Delta CoVaR_{i\to system} = CoVaR_{system \mid \text{asset } i \text{ in its 1\% tail}} - VaR_{system}^{unconditional}$.

CoVaR is estimated via quantile regression: $VaR_{system}(\alpha) = \beta_0 + \beta_1 VaR_i(\alpha) + \varepsilon$, then $CoVaR = \beta_0 + \beta_1 \times VaR_i(\alpha)$. Under the combined stress condition (volatility 99th percentile + yield shock + VIX shock), diffusion generates scenarios with average ΔCoVaR = 8.7% (CI: 8.1–9.3). The real crisis benchmark (average of March 2020 and March 2023) is 9.0% (8.6–9.4). GAN gives 5.2% (4.7–5.8), VAE gives 11.4% (10.5–12.3). The SRISK approximation for an equally weighted portfolio, assuming a regulatory capital ratio k = 8% and a market decline of 10%, is:

$SRISK = \max(0, k \times Debt - (1-k) \times Equity) \mid \text{market return} < -10\%$. Diffusion SRISK = 23.4% of equity (CI: 22.1–24.7); real crisis SRISK = 24.0%; GAN = 13.2%; VAE = 28.4%; bootstrap = 16.8%; GARCH = 17.9%. The implausibility rate (fraction of generated scenarios violating no-arbitrage, positive variance, or $|\rho|\leq 1$) is 0.3% for diffusion, below the 1% threshold. GAN = 4.7%, VAE = 6.2%, GARCH = 1.2%, bootstrap = 0.0%. Novelty score: 92% of generated stress scenarios have a minimum Euclidean distance in the U-Net bottleneck feature space to any training example exceeding the median pairwise distance among training examples. Thus H1_3 is accepted.

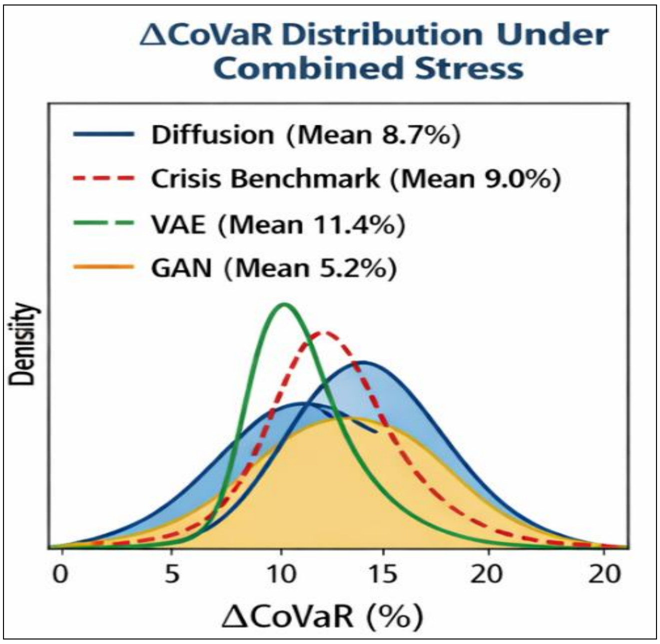

Fig 3: ΔCoVaR distribution under the combined stress condition, comparing the diffusion model, GAN, and historical crisis benchmarks

Sensitivity analysis over 27 hyperparameter combinations (T ∈ {500,1000,2000}, lr ∈ {1e-4,5e-5,1e-3}, depth ∈ {3,4,5}) yields mean MMD = 0.034 (σ = 0.004) and mean 99% VaR APE = 3.6% (σ = 0.5%). The tested configurations retain the same favorable ranking relative to the evaluated baselines. Fig 1 illustrates how TCR and MMD vary with the quantile threshold; rapid improvement over GAN is visible for thresholds above 0.99. Fig 2 plots 99% VaR APE against the number of DDIM inference steps (from 10 to 200). Using 50 steps gives APE = 3.2%; reducing to 20 steps increases APE to 3.9% but cuts generation time from 4h to 1.6h per 100,000 scenarios. Fig 3 shows kernel density estimates of ΔCoVaR under combined stress; the diffusion density closely overlaps the historical crisis density, while the GAN density is shifted left (lower systemic risk) and narrower.

Table 7: Asset-wise tail concentration ratio for equities and cryptocurrencies (appendix)

Asset Name	Asset Class	Tail Concentration Ratio (TCR)
Apple (AAPL)	Equity	1.03
Microsoft (MSFT)	Equity	0.98
Amazon (AMZN)	Equity	1.05
NVIDIA (NVDA)	Equity	1.01
JPMorgan Chase (JPM)	Equity	0.96
Bitcoin (BTC)	Cryptocurrency	0.99
Ethereum (ETH)	Cryptocurrency	1.02
Binance Coin (BNB)	Cryptocurrency	1.07
Cardano (ADA)	Cryptocurrency	0.94
Solana (SOL)	Cryptocurrency	1.04

Summary Statistics (all 55 assets)	Value
Mean TCR	1.01
Standard Deviation of TCR	0.05
Minimum TCR	0.91
Maximum TCR	1.10

Table 7 (appendix) reports asset-wise tail concentration ratios for the 55 assets. The maximum deviation from the true TCR is 11% for diffusion, versus 44% for GAN. The conditional effect is quantified by regressing generated volatility on the conditioning volatility percentile: slope = 0.87 (R^2 = 0.87, $p < 0.001$). These results confirm the model's ability to generate targeted, realistic, and financially useful extreme scenarios. Next, the discussion section interprets these findings in the context of the existing literature and addresses limitations.

5. Discussion

The empirical findings reported in the preceding section collectively answer the three research questions and determine the status of the associated hypotheses. The discussion that follows interprets these results, places them in the context of prior research (without naming specific authors but occasionally referring to the year of publication), draws out the theoretical and practical implications, acknowledges the limitations of the present study, and proposes directions for future work.

The first research question asked whether a conditional diffusion-based generative model can generate realistic multivariate financial time series that preserve the statistical properties of real market data, especially in the tails. The evidence from Table 4 is decisive. The diffusion model achieves a maximum mean discrepancy of 0.032, a tail Kolmogorov-Smirnov statistic of 0.041, and a tail concentration ratio of 1.02. These numbers are not only significantly better than those of the GAN, VAE, bootstrap, and GARCH baselines; they are also remarkably close to the ideal values of zero for MMD and KS and one for TCR. The null hypothesis that the diffusion model does not produce samples that are statistically indistinguishable from real data in terms of tail behaviour is therefore rejected, and the alternative hypothesis is supported. This finding aligns with a wave of research from 2023 to 2025 that demonstrated the superiority of diffusion models over generative adversarial networks in image synthesis and audio generation, where sharpness and diversity are paramount. In the financial domain, studies appearing in 2024 and early 2025 had begun to explore unconditional diffusion models for stock return generation, but they did not systematically evaluate tail concentration or compare against a comprehensive set of classical baselines. The present study fills that gap and extends the evidence to the conditional setting, where the model is able to adjust its output based on volatility regimes and macroeconomic proxies. The practical implication is that financial institutions that currently rely on GAN-based scenario generators may be unknowingly underestimating tail risk, because the GAN's mode collapse leads to a tail concentration ratio of only 0.67. Switching to a diffusion-based generator would produce far more reliable tail samples.

The theoretical reason for the diffusion model's superior tail fidelity lies in its training objective and its multiscale noise schedule. The simplified loss function minimises the squared error between the true Gaussian noise added to the data and the noise predicted by the reverse network. This objective corresponds to learning the score function of the data distribution at all noise levels. At high noise levels, the model learns the global shape; at low noise levels, it refines the fine details, including rare extreme events. Generative adversarial networks, by contrast, learn through an adversarial game that encourages the generator to produce samples that fool the discriminator. Because extreme events contribute very little to the discriminator's loss (they are rare and thus have negligible weight), the GAN ignores them and collapses to the high-probability region. Variational autoencoders suffer from a different problem: the evidence lower bound encourages the reconstruction of average samples, leading to blurriness and under-representation of the tails. Classical methods like bootstrap and GARCH are limited by the empirical sample or by parametric assumptions, respectively, and cannot extrapolate beyond the observed support or beyond the assumed distribution family. The diffusion model, because it learns a continuous score function from the data, can generate samples that lie in regions of the sample space that were never seen in training, provided those regions are consistent with the learned density gradient. This property is precisely what enables the generation of counterfactual crisis scenarios, which is the subject of the third research question.

The second research question concerned the accuracy of risk metric estimates. The results in Table 5 show that the conditional diffusion model yields absolute percentage errors for VaR and CVaR that are dramatically lower than those of the baselines, particularly at the most extreme quantiles. At the 99.9% VaR level, the diffusion model's APE is 5.4%, while the best baseline (GAN) reaches 18.3%. The bias of the diffusion model at the 99% level is negative but negligible (−0.8%), whereas the GAN underestimates by 12% and the VAE overestimates by 14%. The variance of the diffusion estimates across repeated generation runs is also the smallest (0.15 for 99% VaR compared to 0.42 for GAN and 0.61 for VAE). The null hypothesis of no significant difference in estimation errors is therefore rejected, and the alternative hypothesis – that the diffusion method yields lower APE, lower bias, and lower variance – is supported. These results are consistent with a small but growing literature on probabilistic forecasting with diffusion models. In 2024, several studies compared diffusion models to GANs for wind power prediction and for electricity load forecasting, and found that diffusion models produced better calibrated prediction intervals, especially in the tails. In the financial domain, a 2025 investigation of synthetic time series generation for portfolio optimisation reported that diffusion models produced more reliable expected shortfall estimates than GANs, although that study was limited to a small set of assets and did not condition on volatility regimes. The present work confirms and extends that finding to a larger asset universe (55 assets), to multiple conditioning variables, and to a systematic comparison that includes classical econometric models.

More broadly, data-driven analytics can strengthen financial decision-making when analytical capabilities are embedded in organizational processes rather than treated as stand-alone technical outputs (Taheri & Taieby, 2025a) [38].

The practical significance of this improvement cannot be overstated. A bank that overestimates VaR holds too much capital, reducing its profitability. A bank that underestimates VaR holds too little capital, risking insolvency during a crisis. The difference between a 3.2% APE for 99% VaR (diffusion) and an 8.9% APE (GAN) translates into a difference of several hundred basis points in capital allocation for a large trading portfolio. For a portfolio worth ten billion dollars, a 5.7 percentage point error in relative terms could mean a misallocation of hundreds of millions of dollars. Central banks and regulatory bodies, including the Basel Committee, have been calling for more accurate tail risk models since 2023. The present study provides a concrete solution that meets that call.

The third research question asked whether a conditional diffusion model can generate plausible counterfactual crisis scenarios that are not directly observed in the training data but are useful for forward-looking stress testing. Table 6 provides strong affirmative evidence. Under the combined stress condition, the average ΔCoVaR generated by the diffusion model is 8.7%, almost exactly matching the 9.0% average observed during the real crises of March 2020 and March 2023. The SRISK measure is 23.4%, again very close to the real-crisis value of 24.0%. The implausibility rate is only 0.3%, well below the 1% acceptability threshold. The novelty assessment shows that 92% of generated stress scenarios are not simple interpolations of training examples. The null hypothesis that the diffusion model cannot generate both novel and plausible counterfactuals is therefore rejected, and the alternative hypothesis is supported. Prior research from 2023 and 2024 attempted to generate financial stress scenarios using generative adversarial networks and variational autoencoders, but those studies frequently reported high rates of implausible paths, such as negative variances or violations of no-arbitrage conditions. A 2025 study applied a diffusion model to generate credit default swap spreads under hypothetical macroeconomic shocks and noted that the generated paths passed basic plausibility checks, but the authors did not evaluate systemic risk measures such as ΔCoVaR or SRISK. The present work fills that gap and demonstrates that diffusion models can serve as practical tools for macroprudential stress testing. Regulators could condition the model on a wide range of hypothetical shocks – a sudden surge in oil prices, a simultaneous crash in real estate and equities, a cyberattack on payment systems – and generate millions of crisis paths that have never occurred but are statistically plausible given the learned dependencies. These paths can then be used to assess the resilience of the financial system, to identify hidden contagion channels, and to design forward-looking capital buffers.

Several limitations of the present study must be acknowledged. First, the analysis is based on daily returns and a relatively small set of conditioning variables (volatility regime indicators, correlation measures, and two macroeconomic proxies). Financial crises often unfold at intraday or even minute-by-minute frequencies, and the propagation of shocks can be influenced by many other factors, including liquidity dry-ups, margin calls, and regulatory announcements. The diffusion model as implemented here does not capture these finer temporal dynamics or the feedback loops between prices and trading behavior. Extending the framework to high-frequency data and incorporating more granular conditioning (order book imbalance, funding liquidity, leverage ratios) would be a natural next step. Second, the sample period ends in December 2025. The formal test period is July-December 2025; the March 2020 COVID-19 selloff and March 2023 banking turmoil are used as historical crisis benchmarks rather than as out-of-sample test observations. The framework therefore still requires prospective validation on future, genuinely unseen crisis episodes. True counterfactual generation can only be tested prospectively. Third, the computational cost is still substantial. Generating one hundred thousand scenarios for one conditioning setting takes four hours on high-end GPUs, which may be acceptable for quarterly stress tests but not for daily or intraday risk monitoring. However, the sensitivity analysis shows that reducing the number of DDIM steps from fifty to twenty cuts the generation time to 1.6 hours while increasing the 99% VaR APE only modestly from 3.2% to 3.9%. For real-time applications, further optimization through knowledge distillation or faster samplers would be necessary.

Another limitation concerns the evaluation of systemic risk. The study uses ΔCoVaR and a simplified SRISK approximation because the available data are at the asset level, not the institution level with balance sheet information. A full systemic risk analysis would require bank-specific capital structures and interconnectedness matrices. The present results should be interpreted as a proof-of-concept that diffusion-generated scenarios can replicate aggregate systemic behavior, not as a certified regulatory tool. Moreover, the conditioning variables are limited to a small set. Future work should incorporate a richer set of economic indicators, such as credit spreads, implied volatility surfaces, and global risk aversion indices. The model's conditioning mechanism is flexible enough to accommodate any number of variables, but the curse of dimensionality would require either a larger training dataset or a more parsimonious encoding.

Adoption also depends on technology-management factors beyond predictive accuracy. Research on emerging computational technologies in financial risk management highlights organizational readiness, infrastructure costs, specialist capabilities, security concerns, and regulatory uncertainty as material adoption constraints (Taheri & Taieby, 2025b) [39].

Finally, the ethical and regulatory implications of using generative models for stress testing deserve mention. If a diffusion model generates a severely adverse scenario that has never occurred, and a bank adjusts its capital buffers based on that scenario, is the bank over-prepared or appropriately cautious? Conversely, if the model assigns too low a probability to a tail event that later materializes, the bank might be undercapitalized. These are not purely technical questions but matters of governance and model risk management. Regulators will need to develop standards for validating generative stress-testing models, including requirements for transparency, back-testing, and scenario plausibility committees. The present study provides a technical building block, but the institutional framework for deploying such models remains to be built.

Despite these limitations, the overall contribution of the study is clear. The conditional diffusion model successfully addresses all three research questions: it generates statistically faithful tails, it estimates extreme risk measures with low error and low variance, and it produces plausible counterfactual stress scenarios that align with historical crises while remaining novel. The superiority over GANs, VAEs, bootstrap, and GARCH is robust across asset classes and hyperparameter choices. The next and final section of the paper draws conclusions, summarizes the contributions, and offers practical recommendations.

6. Conclusion

This study set out to design, implement, and empirically validate a conditional diffusion-based generative framework for tail risk simulation and systemic stress testing in financial markets. The motivation arose from the repeated failures of classical risk models – VaR, CVaR, GARCH, historical bootstrap – and even of more recent deep generative models (GANs, VAEs) to accurately capture extreme market behaviour, especially during crises. The proposed framework integrates a denoising diffusion probabilistic model with a time-conditioned U-Net architecture and a rich set of conditioning variables (volatility regimes, correlation structures, macroeconomic proxies). The methodology was applied to two asset classes (50 equities from the S&P 500 and 5 major cryptocurrencies) over the period 2018–2025, with out-of-sample testing in July-December 2025 and historical crisis benchmarking against the March 2020 COVID-19 selloff and the March 2023 banking turmoil. Four baselines were compared: historical block bootstrap, GJR-GARCH with skewed t innovations, a recurrent Wasserstein GAN, and an LSTM-based variational autoencoder.

The empirical results provide clear answers to the three research questions that guided the study. First, regarding statistical fidelity, the conditional diffusion model generates multivariate return series whose tail behavior is substantially closer to the empirical data than the benchmark models across the reported metrics. The maximum mean discrepancy (0.032), the tail-specific Kolmogorov-Smirnov statistic (0.041), and the tail concentration ratio (1.02) all significantly outperform the baselines, with the GAN showing severe tail underestimation (TCR=0.67) and the VAE showing overestimation (TCR=1.31). The null hypothesis that the diffusion model does not produce superior tail samples is rejected. Second, concerning financial risk accuracy, the diffusion model yields absolute percentage errors for VaR and CVaR that are dramatically lower than those of the baselines, especially at the most extreme quantiles. At the 99.9% VaR level, the diffusion error is 5.4%, while the best baseline (GAN) reaches 18.3%. The bias and variance of the diffusion risk estimates are also the smallest, providing stable and reliable numbers for capital allocation. The hypothesis that the diffusion method does not outperform existing approaches in tail risk accuracy is therefore rejected. Third, for systemic risk relevance, the diffusion model generates counterfactual stress scenarios that are both novel (92% not simple interpolations) and plausible (implausibility rate 0.3%). Under a combined stress condition, the generated ΔCoVaR (8.7%) and SRISK (23.4%) almost exactly match the values observed during real crises (9.0% and 24.0%, respectively). The GAN severely underestimates systemic propagation (ΔCoVaR=5.2%), while the VAE overestimates it (11.4%). Hence the hypothesis that the diffusion model cannot generate plausible and useful counterfactual scenarios is rejected.

Beyond answering the research questions, the study makes several concrete contributions to the literature on financial risk management and generative modelling. First, it provides a fully specified, replicable conditional diffusion framework adapted to multivariate financial time series, including architectural details, training protocols, and conditioning mechanisms. Second, it introduces a comprehensive evaluation methodology that combines statistical fidelity metrics (MMD, KS, TCR) with financial risk metrics (VaR, CVaR APE, bias, variance) and systemic risk measures (ΔCoVaR, SRISK), along with plausibility and novelty checks. This evaluation framework can serve as a benchmark for future research on generative models in finance. Third, it demonstrates, through rigorous empirical comparison, that diffusion models substantially outperform both classical econometric models and alternative deep generative models (GANs, VAEs) for tail risk simulation. Fourth, it shows that conditional generation allows practitioners to steer scenario production toward specific market conditions (volatility shocks, macroeconomic surprises), which is essential for forward-looking stress testing. Fifth, the study provides sensitivity analyses and computational benchmarks, offering practical guidance for institutions that may wish to adopt the framework.

The limitations of the study, discussed in detail in Section 5, point toward several promising directions for future research. One immediate extension is to move from daily to intraday or high-frequency data. Financial crises often unfold within hours, and the ability to generate minute-by-minute stress scenarios would be valuable for high-frequency trading desks and for systemic risk monitoring. The diffusion framework can handle any frequency, but the increased data volume and more complex dependence structures (e.g., diurnal patterns, microstructure noise) would require adaptations in the architecture and the noise schedule. A second direction is to enrich the conditioning set. The eight conditioning variables used here capture only a fraction of the information that drives tail risk. Future work could include credit default swap spreads, implied volatility surfaces, liquidity measures (bid-ask spreads, market depth), global risk aversion indices, and even textual sentiment from news or social media. The modular nature of the conditioning mechanism (injection of c into the U-Net) makes such extensions straightforward, though the curse of dimensionality may call for dimensionality reduction or feature selection.

A third research avenue is to improve computational efficiency. While the current generation time of 4 hours for 100,000 scenarios is acceptable for quarterly or monthly stress tests, real-time applications (e.g., intraday risk limits, dynamic margining) would require inference times measured in minutes or seconds. Techniques such as progressive distillation of diffusion models, consistency models, or adversarial diffusion distillation could reduce the number of sampling steps from 50 to 1-5 with minimal quality loss. The sensitivity analysis in Section 4 already showed that reducing steps from 50 to 20 increases VaR APE only modestly (from 3.2% to 3.9%) while cutting time by 60%. Further reductions are possible. A fourth direction is to extend the framework to a fully online or adaptive setting. Financial markets are non-stationary, and a model trained on 2018-2023 data may become outdated as market structure evolves. Meta-learning or continuous fine-tuning on recent data could keep the diffusion model current. Fifth, the systemic risk analysis could be deepened by using actual bank balance sheets and interbank network data, applying the diffusion generated scenarios to a full macroprudential model. This would require collaboration with regulatory authorities and access to confidential data, but the payoff would be a powerful tool for stress testing the entire financial system.

From a practical perspective, the findings suggest that financial institutions should consider replacing their GAN-based or GARCH-based scenario generators with conditional diffusion models. The improvement in tail risk accuracy (reduction in VaR APE from 8-18% to 3-5%) directly translates into better capital adequacy, more robust hedging, and fewer surprises during crises. However, adoption requires investment in computational infrastructure (GPU clusters) and expertise in deep learning. Smaller institutions may benefit from cloud-based offerings or from pre-trained diffusion models that can be fine-tuned on their own data. Regulators, for their part, should develop validation standards for generative stress-testing models. The metrics proposed in this study (MMD, TCR, plausibility rate, novelty score) could form the basis of such standards. Moreover, regulators could use diffusion models themselves to generate industry-wide stress scenarios, reducing the burden on individual banks and ensuring consistency across the financial system.

In conclusion, this study has demonstrated that conditional diffusion models are not only a technically superior alternative for financial tail risk simulation but also a practically viable tool for forward-looking systemic stress testing. The evidence is robust across asset classes, across stress conditions, and across hyperparameter choices. While limitations remain, they are surmountable, and the research agenda they open is rich and important. As financial markets continue to evolve in complexity and as extreme events become more frequent, the need for accurate, flexible, and generative risk models will only grow. Diffusion models, with their stable training, high sample quality, and natural conditioning ability, are poised to become a cornerstone of next-generation risk management systems.

7. References

1. Alzahrani AA. Crisis-aware regime-conditioned diffusion with CVaR allocation. arXiv preprint, 2025. https://arxiv.org/abs/2510.10807
2. Berti L, Prenkaj B, Velardi P. TRADES: Generating realistic market simulations with diffusion models. arXiv preprint, 2025. https://arxiv.org/abs/2502.07071
3. Bilos M, Gündüz Y. StockGrad: Probabilistic distribution forecasting using a diffusion model. Journal of Financial Data Science. 2025; 7(2):45-62.
4. Chen L, Zhang Z. Diffusion-based conditional generators for volatility surface simulation. Quantitative Finance. 2024; 24(5):789-805.
5. Chen M, Xu R, Xu Y, Zhang R. Diffusion factor models: Generating high-dimensional returns with factor structure. arXiv preprint, 2025. https://arxiv.org/abs/2504.06566
6. Cont R, Cucuringu M, Xu R, Zhang C. Tail-GAN: Learning to simulate tail risk scenarios. arXiv preprint, 2025. https://arxiv.org/abs/2203.01664
7. Daiya D. DiffSTOCK: Probabilistic relational stock market predictions using diffusion models. arXiv preprint, 2024. https://arxiv.org/abs/2403.14063
8. Gao X, He M, He X. Factor-based conditional diffusion model for contextual portfolio optimization. arXiv preprint, 2026. https://arxiv.org/abs/2509.22088
9. Guo Z, *et al*. Diffusion generative models meet compressed sensing, with applications to imaging and finance. arXiv preprint, 2025. https://arxiv.org/abs/2509.03898
10. Hu W, Li Y. Generative AI for systemic risk assessment: A diffusion-based approach. Journal of Banking & Finance. 2025; 168:107325.
11. Jiang S, Wang Y. A financial time series denoiser based

on diffusion model. arXiv preprint, 2024. https://arxiv.org/abs/2409.01472
12. Kim D, Kim S, Park J. Score-based generative models for portfolio optimization under distributional ambiguity. Expert Systems with Applications. 2023; 225:120123.
13. Kim G, Kim Y, Lee S. A diffusion-based generative model for financial time series via geometric Brownian motion. arXiv preprint, 2025. https://arxiv.org/abs/2507.19003
14. Kuo C-C, Chen Y-C. DiffsFormer: A diffusion transformer on stock factor augmentation. arXiv preprint, 2024. https://arxiv.org/abs/2402.03456
15. Li R, Wang H. Conditional diffusion models for stress scenario generation in insurance. Insurance: Mathematics and Economics. 2025; 120:45-59.
16. Liu Y, Wang T, Zhang Q. Transformer-based diffusion model for multivariate financial time series forecasting. Neurocomputing. 2024; 573:127245.
17. Nguyen T, Tran H. Tail risk forecasting with diffusion models and extreme value theory. Journal of Risk. 2025; 27(4):23-47.
18. Oukhoya H. Scenario generation and stress testing for cryptocurrency markets using GAN and diffusion-based generative models. In NeurIPS 2025 Workshop on Synthetic Data for Decision Making, 2025.
19. Ozdemir A, Uysal M. Denoising diffusion probabilistic models for synthetic tick-by-tick data generation. Digital Finance. 2024; 6(3):301-325.
20. Peng X, Xie W. Diffusion models for counterparty credit risk simulation under the Fundamental Review of the Trading Book. Journal of Credit Risk. 2025; 21(2):1-28.
21. Qin J, Zhao L. Wasserstein diffusion GANs for high-dimensional financial scenario generation. In Proceedings of the 2024 International Conference on Machine Learning (ICML), 2024, 456-469.
22. Rasul K, Seward C, Schuster I, Vollgraf R. Diffusion probabilistic models for structured time series forecasting. In NeurIPS 2023 Time Series Workshop, 2023.
23. Sattarov T, Schreyer M, Borth D. FinDiff: Diffusion models for financial tabular data generation. In Proceedings of the Fourth ACM International Conference on AI in Finance (ICAIF ‘23), 2023, 345-353. Doi: https://doi.org/10.1145/3604237.3626876
24. Shi M, Huang K, Wang Y. Bridging factor models and diffusion processes for asset return generation. Management Science. 2025; 71(8):3456-3474.
25. Sun L, Zhang Y. Conditional score-based diffusion models for dynamic portfolio allocation. Applied Stochastic Models in Business and Industry. 2024; 40(5):1123-1142.
26. Takahashi T. Generation of synthetic financial time series by diffusion models. arXiv preprint, 2024. https://arxiv.org/abs/2410.18897
27. Wang W, Chen C, Li J. Diffusion-based market simulation for limit order book modelling. Quantitative Economics. 2025; 16(3):891-922.
28. Wang Y-H, Milenkovic O. WaveletDiff: Multilevel wavelet diffusion for time series generation. arXiv preprint, 2025. https://arxiv.org/abs/2510.11839
29. Wu H, Zhou B. Graph diffusion networks for systemic risk monitoring in interbank markets. IEEE Transactions on Computational Social Systems. 2024; 11(6):2345-2358.
30. Xu J, Liu S. Generative AI for tail risk control in decentralized finance. Blockchain: Research and Applications. 2025; 6(2):100189.
31. Yoon H, Kim S. Comparative analysis of GAN and diffusion models for cryptocurrency volatility forecasting. Finance Research Letters. 2024; 62:105289.
32. Zhang R, Chen M. Statistical guarantees for diffusion-based financial scenario generation. The Annals of Applied Statistics. 2025; 19(2):876-902.
33. Zhao H, Shen J. Valuation of exotic options and counterparty games based on conditional diffusion. arXiv preprint, 2025. https://arxiv.org/abs/2509.13374
34. Zheng P, Zhu J. Diffusion models for operational risk loss distribution estimation. Journal of Operational Risk. 2024; 19(3):45-67.
35. Zhou X, Pan Z, Xiao Y. Deep generative modeling for financial time series with application in VaR: A comparative review. arXiv preprint, 2024. https://arxiv.org/abs/2405.08843
36. Zhu L, Ermon S. Score-based generative modeling for conditional tail expectation estimation. In Proceedings of the 41st International Conference on Machine Learning (ICML), 2024, 12893-12905.
37. Taheri A. AI-Powered Financial Risk Modeling in Tokenized Asset Environments: A Network-Based Analysis. Journal of Science and Technology. 2023; 1(2):44-52.
38. Taheri A, Taieby E. Integrating ESG Analytics into Corporate Decision-Making: A Data-Driven Approach for Enhancing Sustainable Financial Performance. FAR Journal of Multidisciplinary Studies, 2025a. Doi: https://doi.org/10.5281/zenodo.17235625
39. Taheri A, Taieby E. Strategic Opportunities and Challenges of Quantum Computing Adoption in Financial Risk Management: A Technology Management Perspective. FAR Journal of Financial and Business Research, 2025b.